Universidad de Jaén

Escuela de Doctorado

TESIS DOCTORAL

Algoritmos de procesado de señal basados en Non-negative Matrix Factorization aplicados a la separación, detección y clasificación de sibilancias en señales de audio respiratorias monocanal

Presentada por:

Juan De La Torre Cruz

Dirigida por:

Pedro Vera Candeas

Francisco Jesús Cañadas Quesada

Jaén, enero de 2021

Esta Tesis doctoral ha sido realizada bajo la supervisión de

Dr. Pedro Vera Candeas
Departamento de Ingeniería de Telecomunicación
Escuela Politécnica Superior de Linares
Universidad de Jaén
Linares, Jaén, España

Dr. Francisco Jesús Cañadas Quesada
Departamento de Ingeniería de Telecomunicación
Escuela Politécnica Superior de Linares
Universidad de Jaén
Linares, Jaén, España

A mi familia, por enseñarme a volar.
A mi princesa, por no dejarme caer.

Agradecimientos

Esta Tesis doctoral, desarrollada en el Departamento de Ingeniería de Telecomunicación de la Universidad de Jaén y orientada en el diseño de algoritmos de tratamiento de señales sibilantes para mejorar el diagnóstico del sistema respiratorio humano, ha sido posible gracias al Programa de "Ayudas predoctorales para la formación del personal investigador" financiado por la Universidad de Jaén.

Comencé mi travesía académica realizando el Grado en Ingeniería de Tecnologías de Telecomunicación y posteriormente el Máster en Ingeniería de Telecomunicación, ambos impartidos en la Escuela Politécnica Superior de Linares. Actualmente me complace decir que volvería a confiar mi formación académica, sin ninguna duda, a esta excelente Escuela que destaca por la calidad de los conocimientos impartidos y por la profesionalidad y cercanía del profesorado. Por ello, me gustaría agradecer a todos los profesores y profesoras que me acompañaron en esa etapa de mi vida formándome, aconsejándome y preparándome para comenzar en el arduo y competitivo mundo profesional.

Esta Tesis ha supuesto un reto y una gran motivación para mí en el ámbito profesional. Especialmente por la posibilidad de contribuir en el campo de la neumología, realizando aportaciones científicas que proporcionen una vía de información complementaria al neumólogo para mejorar el origen de la posible enfermedad pulmonar obstructiva y con ello incrementar la fiabilidad del primer diagnóstico efectuado. Además, esta Tesis significa el inicio de una línea de investigación centrada en el campo del procesado de señales biomédicas con un amplio camino por abordar.

El desarrollo de esta Tesis ha sido posible gracias al esfuerzo, la colaboración y el apoyo de varias personas que han estado presentes en todo el trascurso del trabajo. En primer lugar, me gustaría comenzar mostrando mi más emotivo agradecimiento a mis directores de Tesis, Dr. Pedro Vera Candeas y Dr. Francisco Jesús Cañadas Quesada. Gracias a ellos pude iniciar mi camino en el mundo de la investigación. Comenzaron siendo mis profesores en varias de las

asignaturas del Grado y del Máster que curse en la Escuela Politécnica Superior de Linares, fueron mis tutores en mi Trabajo Fin de Grado y mi Trabajo Fin de Máster, y ahora los considero mis compañeros y mis amigos. He tenido la oportunidad de conocer a dos magníficos profesores, tutores, directores, investigadores, compañeros, amigos .... En definitiva, a dos personas excepcionales que se han convertido en dos claros referentes a seguir. Quisiera agradecerles que se hayan convertido en mis guías durante esta etapa, académica y profesional, de mi vida.

Además, nuevamente quisiera mostrar mi más sincero agradecimiento al Dr. Francisco Jesús Cañadas Quesada, por el enorme tiempo invertido en esta Tesis. Sin sus conocimientos científicos, su implicación, sus consejos, sus brillantes ideas, su apoyo, su tiempo, su pasión por la investigación ... no hubiese sido posible la realización de esta Tesis. Además, me gustaría mostrarle mi agradecimiento por ser el responsable de abrir una línea de investigación que significó el origen de esta Tesis Doctoral. En definitiva, gracias a todo su esfuerzo y a las muchas horas de trabajo que hemos tenido que desempeñar para conseguir que los frutos de esta Tesis salgan a la luz. Gracias por todo y sobre todo gracias por el apoyo profesional y personal que siempre he recibido por tu parte, te has convertido en una fuente de inspiración para mí y siempre podrás contar conmigo.

Del mismo modo, quisiera mostrar mi agradecimiento al Dr. Nicolás Ruiz Reyes, director del grupo de investigación "Tratamiento de Señales en Sistemas de Telecomunicación (TIC-188)", por la ayuda recibida en temas de procesado de señal y discusiones científicas.

Mi agradecimiento sincero al Dr. Damián Martínez Muñoz, por su colaboración en los procesos de optimización de los algoritmos y en la búsqueda de bases de datos de interés que han hecho posible alcanzar algunos de los objetivos de la Tesis Doctoral.

Igualmente, quiero hacer constar mi agradecimiento al Dr. Julio José Carabias Orti, por ser mi guía en mis primeros pasos en la docencia. Sin duda, tener a Julio como compañero me ha facilitado mucho el proceso de adaptación en el mundo docente.

No quiero dejar de expresar mis más sinceros agradecimientos a mis compañeros Antonio y Pablo por concederme su ayuda y colaboración siempre que lo he necesitado.

Quisiera también mostrar mi agradecimiento al Dr. Gerardo Pérez Chica, neumólogo del Hospital Universitario de Jaén, por haber compartido parte de sus valiosos conocimientos en el campo del sistema respiratorio humano. Sus aportes han sido de gran importancia para afrontar la totalidad de los objetivos propuestos en el plan de investigación desempeñado.

Me gustaría mostrar mi más profundo agradecimiento a mi familia. A mis padres, Juan y Josefa, por ese apoyo incondicional, por haberme ayudado en todo, por haber confiado siempre en mí y por demostrarme que siempre caminarán a mi lado sin importar las dificultades que se presenten. A mi hermana, Fátima, por el cariño y la fuerza que nos une, por apoyarme bajo cualquier circunstancia y por su valiosa incondicionalidad. No existen palabras de agradecimiento que puedan expresar mi gratitud. Pero, muchas gracias por todo lo que habéis hecho por mí y por permitirme disfrutar de esta familia.

Finalmente, y para mí el agradecimiento más importante, me gustaría agradecer el apoyo incondicional de Raquel. Gracias por acompañarme en esta dura travesía, animándome y apoyándome en cada día. Gracias por ser capaz de motivarme cuando más lo he necesitado y sobre

todo de conseguir iluminar mis días grises. Gracias por todo y recuerda que eres y serás la pieza clave de mi puzle.

Juan De La Torre Cruz
Enero de 2021

Resumen

A nivel mundial, las enfermedades pulmonares obstructivas constituyen un problema de salud pública de enorme y creciente importancia por su elevada prevalencia, alta morbimortalidad y coste socioeconómico. Actualmente, el proceso de auscultación sigue siendo el primer examen clínico que un neumólogo emplea para evaluar el estado del aparato respiratorio, debido a que se trata de un método no invasivo, de bajo coste, fácil de utilizar y especialmente seguro para el paciente. Sin embargo, el diagnóstico que se deriva de la auscultación sigue siendo un diagnóstico altamente subjetivo que se encuentra condicionado a la habilidad, experiencia y entrenamiento de cada médico en la escucha e interpretación de las señales de audio respiratorias. En consecuencia, se producen un alto porcentaje de diagnósticos erróneos que ponen en riesgo la salud de los pacientes e incrementan el coste asociado a los centros de salud.

Una de las principales tareas que un médico realiza cuando ausculta a un paciente es la búsqueda y el análisis de sonidos adventicios que puedan producirse durante la respiración. Los sonidos adventicios son producidos por la obstrucción de las vías respiratorias y su presencia es uno de los síntomas más comunes en las enfermedades pulmonares obstructivas. Concretamente, las sibilancias (wheezes o wheezing) son consideradas uno de los sonidos adventicios de mayor importancia, ya que alertan de la posible presencia de importantes enfermedades pulmonares obstructivas, tales como, asma, bronquiolitis, bronquiectasia o enfermedad pulmonar obstructiva crónica (EPOC).

Sin embargo, uno de los principales inconvenientes presentes en las líneas de investigación relacionadas con el procesado de señales sonoras biomédicas es la escasez de bases de datos estandarizadas. Por ese motivo, los métodos y algoritmos desarrollados en esta Tesis se han alejado de los populares enfoques de aprendizaje automático (Machine Learning o Neural Network), cuya fiabilidad depende de una base de datos de entrenamiento lo suficientemente versátil, para centrarse en explotar todas las posibilidades que ofrece uno de los enfoques de descomposición matricial más utilizado en el campo de procesado de señal, conocido como

Non-negative Matrix Factorization (NMF), considerando la escasa literatura que existe de esta técnica en este campo de investigación.

Esta Tesis tiene como objetivo principal el desarrollo de nuevos métodos y algoritmos NMF aplicados a la separación, detección y clasificación de sonidos respiratorios (sonidos emitidos por los pulmones humanos en un paciente sano) y sonidos sibilantes (sonidos adventicios continuos emitidos por el sistema respiratorio en un paciente enfermo) para proporcionar una vía de información complementaria al médico que ayude a mejorar el origen de la enfermedad y fiabilidad del primer diagnóstico emitido por el médico al analizar las señales sonoras capturadas mediante la auscultación. Con el propósito de abordar las tareas de mayor importancia para los neumólogos en el análisis y la caracterización de los sonidos sibilantes, varios trabajos de investigación han sido desarrollados.

En primer lugar, los sonidos respiratorios normales suponen una grave interferencia acústica que dificulta el proceso de escucha de las sibilancias durante la auscultación, ya que ambos sonidos se encuentran solapados en tiempo y frecuencia. En este sentido, se proponen dos sistemas que permiten aislar los sonidos sibilantes de la respiración normal. El primer sistema se basa en un enfoque NMF con regularizaciones espectro-temporales que permiten modelar el comportamiento de ambos sonidos. El segundo sistema propone una versión extendida del enfoque de cofactorización matricial, conocido como Non-negative Matrix Partial Co-Factorization (NMPCF), añadiendo información entre segmentos (fases respiratorias) en el proceso de compartición de bases espectrales. Asumiendo que los sonidos respiratorios se consideran eventos sonoros repetitivos durante la respiración y que las sibilancias podrían no estar presentes en todos los segmentos, la principal contribución consiste en añadir mayor importancia a los segmentos libres de sibilancias para modelar los patrones respiratorios repetitivos.

Por otro lado, una necesidad solicitada por los neumólogos es la existencia de herramientas fiables que permitan detectar los sonidos sibilantes. A este respecto, por un lado, se propone un sistema que permite detectar la presencia o ausencia de sibilancias en señales sonoras respiratorias. Específicamente, el método propuesto consiste en un enfoque NMF semi-supervisado, basado en el comportamiento tonal de las sibilancias, que permite extraer las trayectorias espectrales que las caracterizan para que posteriormente analizando la suavidad de las trayectorias se determine la presencia o ausencia de sibilancias. Por otro lado, se proponen dos sistemas que permiten realizar una localización temporal de las sibilancias. El primer sistema utiliza como etapa previa un enfoque NMF con regularizaciones espectro-temporales que permite obtener una separación entre sonidos respiratorios y sibilantes. En tal caso, la contribución en detección se basa en aplicar la divergencia Kullback-Leibler para discriminar entre áreas sibilantes y respiratorias. El segundo sistema consiste en un algoritmo recursivo que combina un enfoque NMF ortogonal con el uso de uno de los descriptores más fiables de dispersión espectral, conocido como Gini index, para localizar los intervalos temporales sibilantes en los sonidos respiratorios.

Además, para los especialistas resulta complejo clasificar el tipo de sibilancia (monofónica o polifónica) utilizando únicamente la información acústica del estetoscopio. A pesar de ello, los trabajos propuestos en la literatura para abordar esta tarea son escasos. Por ello, se propone un método basado en un enfoque NMF con regularizaciones para clasificar el tipo de sibilancia

atendiendo a su estructura armónica, aportando información relevante al médico para determinar el grado de obstrucción pulmonar y por tanto, del nivel de gravedad.

Por último, una de las principales limitaciones actuales de la auscultación sigue siendo la gran interferencia acústica causada por el ruido ambiental que rodea al paciente. Para este fin, se propone un sistema multicanal, basado en un enfoque de cofactorización matricial NMPCF, que permite modelar los sonidos interferentes repetitivos encontrados en ambos dispositivos monocanal utilizados para realizar las grabaciones (estetoscopio digital y micrófono externo), y así aislarlos de las señales biomédicas de interés.

Palabras clave: Non-negative Matrix Factorization (NMF), procesado de señal biomédica, auscultación, diagnóstico, asma, enfermedad pulmonar obstructiva crónica (EPOC), sibilancias, sonidos respiratorios normales, ruido ambiente, separación, clasificación, detección, divergencia, bases, activaciones, regularizaciones, suavidad, dispersión, tonalidad, ortogonalidad, recursividad, monofonía, polifonía, trayectorias espectrales, SDR, SIR, SAR.

Abstract

Globally, obstructive pulmonary diseases are a huge and growing public health problem due to their high prevalence, high morbidity and mortality, and socio-economic cost. Today, the auscultation process is still the first clinical examination used by physicians to assess the condition of the respiratory system, because it is a non-invasive, low-cost, easy-to-use and particularly safe method for the patient. However, the diagnosis derived from auscultation remains a highly subjective one that is conditioned by each physician's skill, experience and training in listening and interpreting respiratory audio signals. As a consequence, a high percentage of misdiagnoses occur that endanger the health of patients and increase the cost associated with health centres.

One of the main tasks, performed by doctors, in the auscultation is the detection and analysis of adventitious sounds that may occur during breathing. Adventitious sounds are produced by airway obstruction and their presence is one of the most common symptoms in obstructive pulmonary diseases. In particular, wheezing is widely considered one of the most important adventitious sound, as it alerts to the presence of relevant obstructive lung diseases such as asthma, bronchiolitis, bronchiectasis or chronic obstructive pulmonary disease (COPD).

However, one of the main drawbacks in the research field of biomedical signal processing is the limited availability of standardized databases. For this reason, all methods and algorithms developed in this Thesis have deviated from the popular machine learning approaches (Machine Learning or Neural Network), whose reliability and robustness depends on a sufficiently versatile training database, to focus on Non-negative Matrix Factorization (NMF) since NMF has not been previously applied in the analysis of wheezing sounds as far as the authors knowledge extends.

The main objective of this Thesis is the development of new NMF methods and algorithms applied to the separation, detection and classification of biomedical sounds, specifically, respiratory sounds (sounds emitted by the human lungs in a healthy patient) and wheezing sounds (continuous adventitious sounds emitted by the respiratory system in a unhealthy patient), in

order to provide a complementary information pathway to the physician to improve the reliability of the diagnosis provided by physicians when analyzing the sound signals captured by auscultation. In order to perform the most important tasks for pneumologists in the analysis and characterisation of wheezing sounds, several research projects have been developed.

Firstly, normal respiratory sounds are a serious acoustic interference that complicates the process of listening to wheezing during auscultation, as both sounds are overlapped in time and frequency. In this respect, two systems are proposed to isolate wheezing sounds from normal breathing with the aim of improving the sound quality of the wheezes to be interpreted by physicians. The first system is based on an NMF approach with spectro-temporal constraints that allow the behaviour of both sounds to be modelled. The second system proposes an extended version of the matrix co-factorization approach, known as Non-negative Matrix Partial Co-Factorization (NMPCF), adding information between segments (respiratory stages) in the process of sharing spectral bases. Assuming that respiratory sounds are considered repetitive sound events during breathing and that wheezing may not be present in all segments, the main contribution is to add more importance to the wheezing-free segments in order to model repetitive breathing patterns.

On the other hand, a need demanded by pulmonologists is to have reliable tools to detect wheezing sounds. In this respect, a system is proposed that allows the detection of the presence or absence of wheezing in respiratory sound signals. Specifically, the proposed method is based on a semi-supervised NMF approach, based on the tonal behaviour of the wheezes, which allows the extraction of the spectral trajectories that characterise them so that later, by analysing the smoothness of the trajectories, the presence or absence of wheezes can be determined. In addition, two systems are proposed which provide a temporal localization of the wheezing. The first system uses an NMF approach with spectral-temporal constraints that separates wheezes from normal respiratory sounds. In this case, the contribution is based on applying the Kullback-Leibler divergence to discriminate between wheezing and respiratory areas. The second system consists of a recursive algorithm that combines an orthogonal NMF approach with the use of one of the most reliable spectral sparse descriptors, Gini index, to locate the wheezing time intervals in the respiratory sounds.

Furthermore, the classification of the type of wheezing sound (monophonic or polyphonic) can be considered as a critical and difficult medical task using only the acoustic information from the stethoscope. Moreover, there is little work in the literature to address this research topic. Therefore, a constrained NMF approach is proposed to discriminate the type of wheezing according to its harmonic structure, providing the physician with relevant information to diagnose the degree of lung obstruction and therefore the level of severity.

Finally, one of the main limitations of auscultation today is still the large amount of acoustic interference caused by the ambient noise surrounding the patient. To this end, a multi-channel system is proposed, based on a NMPCF approach, which allows the repetitive interfering sounds found in both single-channel devices used to perform the recordings (digital stethoscope and external microphone), to be modelled and isolated from the biomedical signals of interest.

Keywords: Non-negative Matrix Factorization (NMF), biomedical signal processing, auscultation, diagnosis, asthma, chronic obstructive pulmonary disease (COPD), wheezing, normal respiratory sounds, ambient noise, separation, classification, detection, divergence, bases, activations, constraints, smoothness, sparseness, tonality, orthogonality, recursion, monophonic, polyphonic, spectral trajectories, SDR, SIR, SAR.

Índice general

Índice de Figuras

Índice de Tablas

Lista de publicaciones

Esta Tesis se presenta como un conjunto de trabajos publicados, acogiéndose a lo establecido en el artículo 25.2 del Reglamento de los Estudios de Doctorado de la Universidad de Jaén, aprobado en febrero de 2012 y modificado en febrero de 2019. En total esta memoria se compone de 7 publicaciones, las cuales están referenciadas en el texto como [P1], [P2], [P3], [P4], [P5], [P6] y [P7].

P1 **J. Torre-Cruz, F. Canadas-Quesada, P. Vera-Candeas, V. Montiel-Zafra and N. Ruiz-Reyes**, "Wheezing Sound Separation Based on Constrained Non-negative Matrix Factorization", in *Proceedings of the 10th International Conference on Bioinformatics and Biomedical Technology (ICBBT)*, Amsterdam, The Netherlands, pp. 18–24, May 2018. DOI: `https://doi.org/10.1145/3232059.3232072` **(Premio a la mejor ponencia de la sesión 2 del congreso)**.

P2 **J. Torre-Cruz, F. Canadas-Quesada, J. Carabias-Orti, P. Vera-Candeas and N. Ruiz-Reyes**, "A novel wheezing detection approach based on constrained non-negative matrix factorization", in *Applied Acoustics*, Volume 148, May 2019, pp. 276-288. DOI: `https://doi.org/10.1016/j.apacoust.2018.12.035`

P3 **J. Torre-Cruz, F. Canadas-Quesada, S. García-Galán, N. Ruiz-Reyes, P. Vera-Candeas and J. Carabias-Orti**, "A constrained tonal semi-supervised non-negative matrix factorization to classify presence/absence of wheezing in respiratory sounds", in *Applied Acoustics*, Volume 161, April 2020, pp. 107-188. DOI: `https://doi.org/10.1016/j.apacoust.2019.107188`

P4 **J. De La Torre Cruz, F.J. Cañadas Quesada, N. Ruiz Reyes, P. Vera Candeas and J.J. Carabias Orti**, "Wheezing Sound Separation Based on Informed Inter-Segment Non-

Negative Matrix Partial Co-Factorizatio", in *Sensors*, Volume 20, May 2020, pp. 26-79. DOI: `https://doi.org/10.3390/s20092679`

P5 **J. Torre-Cruz, F. Canadas-Quesada, J. Carabias-Orti, P. Vera-Candeas and N. Ruiz-Reyes**, "Combining a recursive approach via non-negative matrix factorization and Gini index sparsity to improve reliable detection of wheezing sounds", in *Expert Systems with Applications*, Volume 147, June 2020, pp. 113-212. DOI: `https://doi.org/10.1016/j.eswa.2020.113212`

P6 **J. De La Torre Cruz, F.J. Cañadas Quesada, N. Ruiz Reyes, S. García Galán, J.J. Carabias Orti and G. Peréz Chica**, "Monophonic and polyphonic wheezing classification based on constrained low-rank Non-negative Matrix Factorization", in *Sensors*. Status: under review.

P7 **J. Torre-Cruz, F. Canadas-Quesada, D. Martínez-Muñoz, N. Ruiz-Reyes, S. García-Galán and J. Carabias-Orti**, "An incremental algorithm based on multichannel non-negative matrix partial co-factorization for ambient denoising in auscultation", in *Applied Acoustics*. Status: under review.

CAPÍTULO 1

Introducción

El objetivo de este capítulo es enfatizar la motivación que ha originado esta línea de investigación. Para ello, inicialmente se expone la problemática que existe en la actualidad sobre las enfermedades pulmonares obstructivas, y cómo el análisis de los sonidos adventicios (en especial los sonidos sibilantes) puede ayudar en la mejora del primer diagnóstico derivado de la auscultación. En segundo lugar, se describen y justifican los objetivos de la línea de investigación, considerando las tareas y dificultades de mayor relevancia para los neumólogos en el análisis de los sonidos sibilantes. Finalmente, se resumen las publicaciones que han sido incluidas en esta Tesis doctoral y se presentan los diferentes capítulos en los que el libro está dividido.

1.1. Contexto y motivación de la investigación

El reto más importante que se presenta a los investigadores de cualquier ámbito científico es conseguir que los avances de sus investigaciones y sus contribuciones científicas mejoren la calidad de vida de los ciudadanos. El concepto ”esalud” (eHealth) representa esta intención en el ámbito sanitario. La Organización Mundial de la Salud (OMS) define este concepto como el uso de Tecnologías de la Información y de la Comunicación (TIC) para la salud. Por otro lado, la Fundación Tecnología y Salud lo define como el conjunto de TIC que se emplea en el entorno sanitario en materia de prevención, diagnóstico, tratamiento, seguimiento y gestión de la salud, actuando como una palanca de cambio en los sistemas sanitarios que permite el ahorro de costes y mejora de su eficiencia. En esta línea, el desafío que se persigue conseguir con esta Tesis es realizar aportaciones científicas que, en materia de diagnóstico, aumenten la eficiencia

de los diagnósticos para evitar poner en riesgo la salud de los pacientes y, en materia de gestión de la salud, disminuyan el coste asociado a los centros de salud propiciado por los diagnósticos erróneos.

Las enfermedades del aparato respiratorio se encuentran en continuo crecimiento y actualmente ocupan el tercer puesto en el ranking de causas de mortalidad, por detrás de las enfermedades cardiovasculares y el cáncer. La OMS informa que los daños causados por el tabaco son la principal causa de las muertes relacionadas con la salud pulmonar [25] y que la contaminación atmosférica urbana ha agravado considerablemente la situación aumentando el riesgo de padecer enfermedades respiratorias agudas [26] . En un estudio [14] presentado por La Oficina Estadística de la Unión Europea (Eurostat), sobre las enfermedades del sistema respiratorio, se manifestó que en 2016 se produjeron 339.000 muertes en la UE como consecuencia de las enfermedades del sistema respiratorio, lo que equivale al 7.5 % de todas las muertes de la UE. Además, añade que en 2017 las enfermedades respiratorias fueron la causa de al menos 1 de cada 10 muertes en España, Portugal, Bélgica, Grecia y Malta, entre otros. Considerando los datos extraídos del Instituto Nacional de Estadística (INE) [17], en las últimas cuatro décadas se ha producido un aumento considerable de las muertes causadas por enfermedades respiratorias en España. Como se puede observar en la Figura 1.1, en el año 2018 se produjeron alrededor de 25.000 muertes más que en el año 1980.

Considerando algunas de las enfermedades pulmonares más comunes, como el asma o la EPOC (Enfermedad Pulmonar Obstructiva Crónica). La OMS calcula que en la actualidad hay más de 339 millones de personas con asma en todo el mundo, siendo la enfermedad crónica

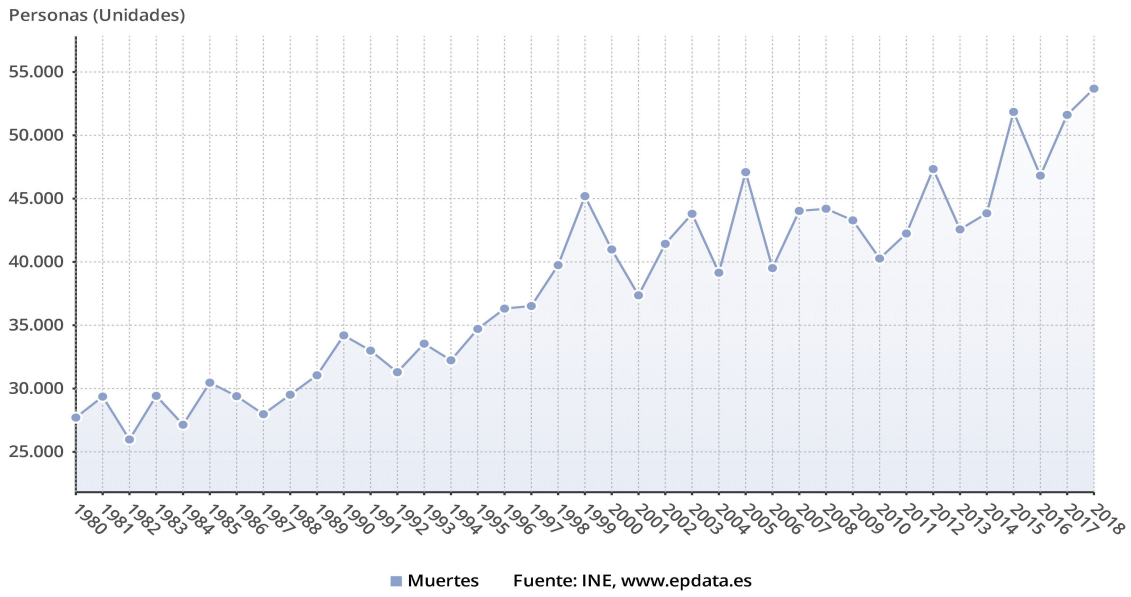

Figura 1.1: Estudio del Instituto Nacional de Estadística sobre las muertes causadas por enfermedades respiratorias en España. Esta figura ha sido obtenida a través del siguiente enlace `https://kutt.it/doQeQJ`.

más frecuente en los niños, afectando al 14 % de los niños a nivel global. Concretamente, la OMS calcula que en el año 2016 se registraron aproximadamente 420.000 muertes por asma produciéndose la mayoría de ellas en personas ancianas al ser este sector de población junto con el sector infantil los sectores más vulnerables ante dicha enfermedad [23]. Por otro lado, la OMS estima que actualmente unos 64 millones de personas sufren EPOC y que alrededor de 3 millones de personas mueren cada año, convirtiéndola en la tercera causa principal de muerte en todo el mundo [24].

Analizando precisamente las enfermedades del aparato respiratorio, uno de los principales inconvenientes es que un gran número de dichas enfermedades no se suelen diagnosticar correctamente por parte del médico en el primer diagnóstico y este error provoca que el regreso del paciente, con la enfermedad que inicialmente no fue detectada, sea más grave. Es en este punto donde toma relevancia el concepto eSalud (eHealth) que trata de aplicar nuevas herramientas derivadas de las Tecnologías de la Información y la Comunicación (TICs) para conseguir, entre otros logros, mejorar el primer diagnóstico del médico, evitando los denominados diagnósticos precoces, con el consecuente aumento de la calidad de vida de los pacientes y con ello, minimizar los costes del sistema sanitario al incrementar su eficacia. En este aspecto, la sociedad española de médicos de atención primaria informa que los diagnósticos erróneos cuestan al sistema sanitario español 2.000 € por paciente al año, comparado con los 400 € que cuesta un paciente correctamente diagnosticado [28].

Hoy día a nivel mundial, el primer diagnóstico que realiza un médico en consulta para evaluar el estado de salud del sistema respiratorio de un paciente se lleva a cabo mediante el análisis de señales sonoras a través de la auscultación, ya que dicho proceso es seguro, no invasivo, de bajo coste, fácil de aplicar en el paciente independientemente de la edad (patient-friendly) y rápido de realizar a través de sencilla instrumentación médica, p.e.: estetoscopio. Sin embargo, el proceso de auscultación cuenta con ciertas limitaciones o desventajas que ponen en peligro la eficacia del diagnóstico derivado. En primer lugar, el diagnóstico que se deriva de la auscultación sigue siendo un diagnóstico altamente subjetivo que se encuentra condicionado a la habilidad, experiencia y entrenamiento de cada médico en la escucha de dichas señales sonoras. Esta habilidad subjetiva para el análisis de las señales sonoras auscultadas implica que en numerosas ocasiones el médico no determine correctamente el origen de la posible enfermedad del paciente. Por otro lado, el hecho de que los sonidos respiratorios normales y los sonidos adventicios (anómalos e indicadores de un desorden pulmonar) se encuentren mezclados simultáneamente en tiempo y frecuencia hace que no sea posible su separación mediante técnicas clásicas de filtrado [247]. Este hecho origina que la aparición de los sonidos respiratorios normales, junto con el ruido ambiental que rodea al paciente, interfiera en la escucha de los sonidos adventicios de interés, entorpeciendo la capacidad cognitiva del médico al ser distraído por dichos sonidos interferentes lo que probablemente causará un aumento de falsos negativos en los diagnósticos realizados. Incluso es común que la capacidad cognitiva del médico se reduzca a lo largo del día, ya que el número de horas dedicadas a analizar los sonidos respiratorios aumenta, un hecho que se ve exacerbado por el estrés que el médico sufre con ciertos casos médicos [324, 154].

Los sonidos adventicios o anómalos que existen cuando el aparato respiratorio presenta un desorden son diversos. Entre ellos se pueden encontrar las sibilancias (wheezing), las crepitaciones (crackles), el estridor (stridor), el roncus (rhonchi) y los roces o frotes pleurales (pleural rub), entre otros [240, 212, 251]. Específicamente, los sonidos sibilantes son considerados un indicador fiable del grado de obstrucción bronquial relacionado con varias enfermedades pulmonares, como el asma, la bronquitis aguda, la bronquiolitis, la bronquiectasia y la EPOC [20, 119, 182, 192, 227, 221, 252, 102]. En consecuencia, esto ha propiciado que la detección, análisis, caracterización y mejora de las sibilancias se defina como una línea de investigación desafiante en el campo del procesado de señales biomédicas. En las últimas dos décadas, han aparecido un conjunto de contribuciones destinadas a mejorar el diagnóstico derivado de las sibilancias. Concretamente, se pueden encontrar trabajos destinados a la eliminación del ruido ambiente que rodea al paciente durante la auscultación [293, 68, 99], la detección y localización temporal de los sonidos sibilantes [192, 174, 233], y la clasificación del tipo de sibilancia [305, 135, 280]. Sin embargo, esta línea de investigación se puede considerar abierta debido, por un lado, a la necesidad de mejorar el rendimiento de los algoritmos propuestos para así incrementar su grado de fiabilidad y poder instaurar las TIC como ayuda para mejorar la eficacia de los diagnósticos y, por otro lado, a la aparición de nuevas tareas que podrían ayudar al médico en el diagnóstico derivado de la auscultación. En este sentido, hasta donde conoce el autor de esta Tesis, no se ha encontrado ninguna aportación científica encaminada a realizar una separación entre los sonidos respiratorios normales y los sonidos sibilantes de interés, que mejore la calidad acústica de las sibilancias para que el médico pueda extraer toda la información que las caracterizan sin que ninguna interferencia acústica pueda entorpecer la capacidad cognitiva del médico. En este sentido, un estudio reciente ha demostrado que los médicos no consiguen detectar parte de los sonidos adventicios debido al solapamiento de los sonidos respiratorios durante la inspiración y la espiración [42].

Aprovechando, por un lado, los estetoscopios digitales y los sensores especializados en la captura de señales respiratorias en el proceso de auscultación, y por otro lado, las capacidades que el procesado de señal puede tener en el campo de las señales acústicas biomédicas, en esta Tesis se aborda el desarrollo de métodos y algoritmos que mejoren la fiabilidad del primer diagnóstico efectuado sobre el estado del sistema respiratorio y ayuden a identificar patologías derivadas de los sonidos sibilantes. Específicamente, se propone explotar todas las posibilidades que ofrece el modelo de descomposición matricial NMF para la creación de métodos y algoritmos destinados a la separación y mejora de los sonidos sibilantes, la detección y localización temporal de las sibilancias, la clasificación del tipo de sonido sibilante, y la eliminación del ruido de fondo que rodea al paciente durante la auscultación. Con ello, se persigue evaluar el potencial del enfoque NMF en este ámbito científico, debido a la escasa literatura existente sobre la aplicación del enfoque NMF en este campo de investigación. Considerando que, en la actualidad, el rendimiento de la mayor parte de los métodos y algoritmos que existen en esta línea de investigación depende de una base de datos de entrenamiento (machine learning, neural networks, etc.), las contribuciones aportadas en esta Tesis proponen algoritmos que no depen-

den de ninguna base de datos externa. Por el contrario, se centran en modelar las características tiempo-frecuencia que diferencian a los distintos sonidos presentes durante la auscultación.

1.2. Justificación y objetivos de la investigación

La importancia de esta Tesis radica en que el diagnóstico derivado del proceso de auscultación sigue siendo un diagnóstico subjetivo que se encuentra condicionado a la habilidad, experiencia y entrenamiento de cada médico para interpretar las señales sonoras escuchadas mediante el estetoscopio. Concretamente, esta Tesis tiene como objetivo general el desarrollo de nuevos métodos y algoritmos basados en Non-negative Matrix Factorization aplicados al procesado de la señal de audio del aparato respiratorio para proporcionar una vía de información complementaria al médico que ayude a identificar patologías derivadas de las sibilancias (asma, bronquiolitis, bronquitis, bronquiectasia, EPOC, etc.) y aumente la fiabilidad del diagnóstico emitido al analizar las señales sonoras capturadas mediante el proceso de auscultación. El cumplimiento de este objetivo general trata de reducir la tasa de falsos negativos en la detección de posibles enfermedades pulmonares derivadas de las sibilancias, evitando poner en riesgo la salud de los pacientes y disminuyendo el coste asociado a los centros de salud.

Teniendo en cuenta las tareas y dificultades de mayor importancia para los neumólogos en el análisis y caracterización de los sonidos sibilantes, es preciso plantear una serie de objetivos específicos, cuya consecución permita alcanzar el objetivo general:

- Desarrollo de sistemas para la separación y mejora de la calidad sonora de las sibilancias en señales respiratorias monocanal [P1][P4]. Con la consecución de este objetivo se pretende aislar los sonidos sibilantes de la respiración normal para que el médico pueda centrarse en escuchar únicamente los sonidos sibilantes y extraer así la información relevante con mayor facilidad.

- Desarrollo de sistemas para la detección de señales sibilantes en señales respiratorias monocanal [P2][P3][P5]. Con la consecución de este objetivo se busca, por un lado, detectar la presencia o ausencia de sibilancias en señales sonoras respiratorias y, por otro lado, localizar el intervalo temporal en el cual las sibilancias se encuentran activas.

- Desarrollo de sistemas para la clasificación y el análisis del tipo de sibilancia en señales respiratorias monocanal [P6]. Con la consecución de este objetivo se persigue clasificar el tipo de sibilancia entre monofónicas y polifónicas, atendiendo a su estructura armónica, aportando información relevante al médico para determinar el nivel de gravedad de la enfermedad pulmonar.

- Desarrollo de sistemas para la eliminación del ruido ambiente que rodea al paciente durante el proceso de auscultación [P7]. Con la consecución de este objetivo se pretende eliminar/atenuar cualquier sonido interferente (personas hablando, llanto de los niños, ruido de la sirena de una ambulancia, ruidos típicos de la calle, etc.) a los sonidos biomédicos de interés para el médico.

Los métodos y algoritmos desarrollados en esta Tesis han sido diseñados siguiendo un enfoque no supervisado (ciego) para que su fiabilidad no dependa de una base de datos de entrenamiento, ya que en esta línea de investigación es común la escasez de bases de datos públicas. Concretamente, todos los sistemas propuestos se basan en uno de los enfoques de descomposición matricial más utilizados en el campo de procesado de señal, conocido como Non-negative Matrix Factorization (NMF). Teniendo en cuenta el conocimiento a priori que se tiene sobre la naturaleza de los distintos sonidos de interés que intervienen en los diferentes objetivos específicos de esta Tesis doctoral (sonidos sibilantes, sonidos respiratorios normales y ruido ambiental que rodea al paciente), todos los métodos propuestos aprovechan las características tiempo-frecuencia de dichos sonidos de interés para conseguir que el enfoque NMF modele el comportamiento diferencial de todos los sonidos presentes en la señal de entrada.

1.3. Contribuciones científicas

En esta Tesis se incluyen siete publicaciones, que se resumen en la siguiente lista. Se utiliza el orden cronológico de publicación.

[P1] Wheezing Sound Separation Based on Constrained Non-negative Matrix Factorization

Como se ha comentado anteriormente uno de los principales problemas a los que se enfrentan los médicos cuando auscultan a un paciente es el solapamiento que existe entre los sonidos respiratorios normales y las sibilancias. Como resultado, la capacidad cognitiva del médico se reduce pudiendo causar un diagnóstico erróneo debido a no escuchar los sonidos adventicios sibilantes con claridad. Este trabajo inicial presenta un enfoque NMF ciego con regularizaciones, para separar los sonidos sibilantes de los sonidos respiratorios en señales monocanal con ambos sonidos solapados en tiempo y frecuencia. Se proponen las regularizaciones, suavidad y dispersión, para modelar las características tiempo-frecuencia de los sonidos sibilantes y respiratorios. Específicamente, el espectrograma de las sibilancias puede ser modelado aplicando dispersión en frecuencia (espectro de banda estrecha o picos espectrales). Sin embargo, el espectrograma de los sonidos respiratorios normales puede ser modelado aplicando suavidad en frecuencia (espectro de banda ancha) y tiempo (la energía varía lentamente en el tiempo). Como resultado, se propone un método no supervisado (ciego) porque no requiere ningún entrenamiento sobre los sonidos activos presentes en la mezcla de entrada. Los resultados experimentales confirman que el método propuesto mejora la calidad de audio de las sibilancias eliminando la mayoría de los sonidos respiratorios, siendo una forma novedosa y pionera de aplicar con éxito un enfoque NMF a un sistema de separación entre sonidos sibilantes y respiratorios.

[P2] A novel wheezing detection approach based on constrained non-negative matrix factorization

Localizar el intervalo temporal en el que las sibilancias se encuentran activas durante la respiración del paciente es de vital importancia para diagnosticar correctamente la enfermedad pulmonar y disminuir la fatiga mental del médico. Sin embargo, actualmente se producen un alto porcentaje de diagnósticos erróneos, especialmente en escenarios ruidosos en los cuales la señal sibilante resulta casi inaudible debido al solapamiento de la señal respiratoria normal. Este trabajo propone un novedoso algoritmo de detección de sibilancias, basado en un enfoque NMF ciego con regularizaciones, que permite modelar y detectar la presencia de las sibilancias, localizando los intervalos temporales en los cuales las sibilancias se encuentran activas cuando se encuentran mezcladas con los sonidos respiratorios normales en una mezcla de audio monocanal. Concretamente este trabajo se compone de dos etapas en cascada: separación y detección. Siguiendo como base el trabajo previo [P1], el objetivo de la etapa de separación es realizar una estimación de los sonidos sibilantes y respiratorios introduciendo una serie de regularizaciones espectro-temporales (suavidad y dispersión) al modelo de factorización NMF para caracterizar el comportamiento de ambos sonidos. Considerando que los sonidos respiratorios pueden encontrarse aislados en algunas áreas de la mezcla (particularmente, en los intervalos del ciclo respiratorio donde las sibilancias se encuentran inactivas), la etapa de detección propone utilizar la divergencia Kullback-Leibler sobre el espectrograma de la mezcla de entrada y el espectrograma estimado respiratorio para distinguir las áreas sibilantes y las respiratorias. Los resultados experimentales señalan que el rendimiento del método propuesto en la localización temporal de las sibilancias supera a los métodos más relevantes del estado del arte mostrándose robusto en la evaluación de escenarios ruidosos.

[P3] A constrained tonal semi-supervised non-negative matrix factorization to classify presence/absence of wheezing in respiratory sounds

Desde el punto de vista clínico, detectar la presencia de sibilancias en los sonidos respiratorios es una tarea desafiante para la identificación temprana de enfermedades pulmonares. En este estudio, se propone un método para detectar automáticamente la presencia de sibilancias en señales sonoras respiratorias monocanal evitando que el sujeto regrese al centro de salud con un empeoramiento de la obstrucción de las vías respiratorias, manifestada por los sonidos sibilantes, que no se detectó en el primer examen clínico realizado por auscultación. El método propuesto está formado por tres etapas. A diferencia de la mayoría de algoritmos de detección de sibilancias en los que el rango espectral de las sibilancias se establece a priori, la primera etapa consiste en un algoritmo que estima el rango espectral en el que la probabilidad de que se produzcan sonidos sibilantes es máxima, dicho rango espectral es definido como “Band Of Interest (BOI)”. La segunda etapa, propone un enfoque NMF tonal semi-supervisado con regularizaciones para obtener los patrones espectrales que modelan la naturaleza tonal (picos espectrales) mostrada por las sibilancias en la BOI estimada, y así realizar una separación entre los sonidos sibilantes y respiratorios. A partir del espectrograma estimado sibilante, la tercera

etapa propone un método para clasificar la condición del sujeto como sano (ausencia de sibilancias) o enfermo (presencia de sibilancias) analizando la suavidad temporal de las trayectorias espectrales definidas por la energía más significativa factorizada en la estimación de la BOI. El sistema propuesto ha sido evaluado y comparado con algunos de los métodos más robustos del estado del arte, dando resultados competitivos en la detección de la presencia de sibilancias en sonidos respiratorios monocanal.

[P4] Wheezing Sound Separation Based on Informed Inter-Segment Non-Negative Matrix Partial Co-Factorizatio

Siguiendo la misma línea de investigación que en [P1], el método propuesto en este trabajo presenta una versión extendida del enfoque NMPCF, llamada Inter-Segment NMPCF (IIS-NMPCF), que elimina la mayor parte de la interferencia acústica causada por los sonidos respiratorios normales mientras se mantiene el contenido de los sonidos sibilantes que necesita el médico para hacer un diagnóstico fiable del estado de las vías respiratorias del sujeto. En este trabajo se asume la hipótesis de que los sonidos respiratorios normales pueden ser considerados eventos sonoros que se repiten durante la respiración del sujeto, así que, los sonidos respiratorios normales se pueden modelar compartiendo los patrones espectrales encontrados en cada etapa respiratoria (segmento), inspiración o espiración, con una señal de entrenamiento compuesta solo de sonidos respiratorios (donde las sibilancias no se encuentran activas). Sin embargo, esta compartición de patrones no puede aplicarse a los sonidos sibilantes, ya que las sibilancias podrían no estar activas en cada segmento debido a su naturaleza impredecible en el tiempo motivada por el trastorno pulmonar. Para mejorar el rendimiento de separación del enfoque NMPCF convencional que trata por igual todos los segmentos del espectrograma, la principal contribución de la propuesta consiste en añadir mayor importancia a los segmentos clasificados como no-sibilantes, para aumentar la fiabilidad del modelado de los sonidos respiratorios repetitivos en el proceso de cofactorización, utilizando información del tipo de segmento (sibilante o no-sibilante) proporcionada por un sistema de detección de presencia o ausencia de sibilancias previamente desarrollado [P3]. Los resultados del trabajo demuestran una mejora significativa en la calidad de audio de las sibilancias obtenida por la propuesta IIS-NMPCF en comparación con enfoques clásicos de separación, NMF y NMPCF.

[P5] Combining a recursive approach via non-negative matrix factorization and Gini index sparsity to improve reliable detection of wheezing sounds

Este trabajo sigue la misma línea de investigación que una de las publicaciones previas [P2]. Específicamente, este trabajo presenta un sistema experto e inteligente basado en un algoritmo recursivo que combina el modelo de descomposición matricial "Ortogonal Non-negative Matrix Factorization (ONMF)" y el uso del descriptor de dispersión espectral "Gini Index" para localizar el intervalo temporal de las sibilancias en sonidos respiratorios monocanal. El algoritmo recursivo está compuesto por cuatro etapas. La primera etapa se basa en un modelo ONMF que permite factorizar patrones espectrales (bases) que son más fieles a la forma en la que ocurren en

el mundo real, al minimizar la redundancia entre ellos. La segunda etapa clasifica la naturaleza periódica de las bases ONMF obtenidas anteriormente analizando la dispersión espectral que proporciona el descriptor Gini Index. La tercera etapa propone un novedoso criterio de parada que permite refinar, a lo largo de las iteraciones recursivas, la estimación del espectrograma sibilante. Concretamente, el criterio de parada permite realizar una nueva iteración del algoritmo recursivo siempre y cuando se elimine una cantidad significativa de sonidos respiratorios normales y se minimice la pérdida del contenido sibilante a lo largo del algoritmo recursivo. Finalmente, la cuarta etapa discrimina entre paciente sano (ausencia de sibilancias) y paciente enfermo (presencia de sibilancias) ateniendo a la dispersión mostrada por la distribución de la energía espectral del espectrograma estimado sibilante, indicando los intervalos temporales en los que las sibilancias están activas para los pacientes enfermos. Los resultados experimentales indican que la propuesta obtiene un mejor rendimiento de detección en comparación con los métodos más relevantes del estado del arte, incluyendo el método propuesto en [P2].

[P6] Monophonic and polyphonic wheezing classification based on constrained low-rank Non-negative Matrix Factorization

Los médicos manifiestan la complejidad de clasificar el tipo de sibilancia utilizando únicamente la información acústica proporcionada por el estetoscopio. Por ello este estudio presenta un método para clasificar el tipo de sibilancia, monofónica o polifónica, atendiendo a su estructura armónica, aportando información relevante al médico para determinar el grado de obstrucción pulmonar y por tanto, del nivel de gravedad. Específicamente, el método propuesto se basa en un enfoque NMF con regularizaciones que permite extraer las componentes en frecuencia que caracterizan a las sibilancias en señales respiratorias monocanal. Para conseguir compactar las componentes espectrales sibilantes, se propone una configuración low-rank que utiliza un número reducido de bases para modelar los sonidos sibilantes. Además, se propone utilizar las regularizaciones de suavidad y dispersión para modelar el comportamiento de los sonidos respiratorios y sibilantes. Estas regularizaciones han sido configuradas para modelar patrones espectrales sibilantes con la menor interferencia respiratoria posible. Posteriormente, se realiza un análisis de las diferentes componentes espectrales extraídas de las sibilancias y, basándose en la estructura armónica, se realiza una clasificación entres sibilancias monofónicas y polifónicas. Los resultados experimentales reportan que la propuesta ofrece un rendimiento superior que el método más relevante a día de hoy. Entre otras ventajas, el método propuesto destaca por ser no supervisado (no necesita bases de datos para entrenamiento) y por ser un método robusto contra la señal respiratoria normal que interfiere.

[P7] An incremental algorithm based on multichannel non-negative matrix partial co-factorization for ambient denoising in auscultation

Como se ha mencionado anteriormente, una de las principales limitaciones actuales en el diagnóstico derivado de la auscultación sigue siendo la gran interferencia acústica causada por el ruido ambiental que rodea al sujeto. Este trabajo presenta un algoritmo incremental, denotado

2C-NMPCF, que mejora la calidad de los sonidos biomédicos capturados durante la auscultación eliminando el ruido ambiental que interfiere. Concretamente, el modelo propuesto adapta el enfoque NMPCF convencional a un escenario multicanal compuesto por dos canales de entrada monocanal que capturan audio simultáneamente. Por un lado, la grabación interna simula el audio capturado con un estetoscopio en el que se pueden escuchar tanto los sonidos biomédicos del interior del cuerpo humano como los ruidos ambientales. Por otro lado, la grabación externa simula el audio capturado con un micrófono externo en el que sólo se capta el ruido ambiental que rodea al sujeto. Así pues, la primera contribución adapta el enfoque NMPCF a un punto de vista multicanal en el ámbito de las señales sonoras biomédicas, asumiendo que los ruidos ambientales pueden ser modelados como sonidos repetitivos que pueden encontrarse simultáneamente en ambos canales de entrada (estetoscopio y micrófono externo). La segunda contribución propone un algoritmo incremental, basado en el anterior enfoque NMPCF multicanal, que renueva el espectrograma biomédico estimado, a lo largo de las etapas incrementales, eliminando la mayor parte del ruido ambiental que no se eliminó en la etapa incremental anterior mientras se mantiene la mayor parte del contenido espectral biomédico de interés. Los resultados experimentales muestran que el método propuesto supera notablemente al método más relevante de referencia. Una ventaja notable del método propuesto es su robustez ante el retardo que puede existir entre las señales de ambos canales de entrada.

1.4. Estructura de la Tesis

En este apartado se presenta la estructura de la Tesis que recoge el trabajo de investigación desarrollado. Se estructura en cinco capítulos descritos a continuación:

- **Capítulo 1: Introducción.** El primer capítulo, que es donde nos encontramos en este momento, se centra fundamentalmente en realzar la motivación que ha originado esta línea de investigación, así como presentar los objetivos y organización de la Tesis doctoral.

- **Capítulo 2: Fundamentos del audio biomédico respiratorio.** El segundo capítulo, se encarga de presentan los conceptos básicos relacionados con el análisis de los sonidos respiratorios. En primer lugar, se describe el sistema respiratorio encargado de generar los sonidos biomédicos respiratorios, desde un punto de vista anatómico, fisiológico y patológico. En segundo lugar, se realiza una introducción al proceso de auscultación, destacando sus principios básicos, ventajas y limitaciones, así como las diferentes opciones disponibles para realizar la grabación de los sonidos auscultados. Para finalizar el capítulo, se introduce una clasificación completa de los tipos de sonidos respiratorios producidos por el aparato respiratorio, haciendo hincapié en los sonidos adventicios (entre ellos las sibilancias).

- **Capítulo 3: Factorización de matrices no negativas.** El tercer capítulo trata la temática de los modelos de señal basados en factorización de matrices no negativas. Se inicia presentando la motivación de utilizar estos enfoques en el campo del procesado de audio.

Además, se describen los principios en los que se basan los modelos de factorización de matrices no negativas y se presenta una clasificación de los diferentes enfoques desarrollados para la separación de fuentes sonoras. Por otro lado, se describen las principales regularizaciones y restricciones que pueden ser incorporadas a los modelos de descomposición para caracterizar el comportamiento tiempo-frecuencia de los sonidos presentes. Para finalizar, se presentan un conjunto de descriptores ampliamente utilizados para aplicar clustering de bases espectrales.

- **Capítulo 4: Revisión del estado del arte.** El cuarto capítulo presenta un estudio del estado del arte destinado al análisis de los sonidos sibilantes y respiratorios para la mejora del diagnóstico derivado de las patologías pulmonares obstructivas. En términos más específicos, se realiza una revisión del estado del arte atendiendo a los objetivos específicos planteados en esta Tesis: detección de sonidos sibilantes, clasificación del tipo de sibilancia (monofónica/polifónica) y eliminación del ruido ambiente que rodea al sujeto. Sin embargo, el objetivo específico relacionado con la mejora de audio de los sonidos sibilantes, ha sido una tarea pionera que, hasta donde conoce el autor de esta Tesis, ningún otro autor había desarrollado. Además, se realiza una recopilación de las principales bases de datos, repositorios online y bibliografía de sonidos sibilantes. Por último, se hace una descripción de las principales métricas utilizadas para medir el rendimiento de los algoritmos propuestos, en función de las diferentes tareas específicas abordadas en esta Tesis.

- **Capítulo 5: Resultados y conclusiones.** El quinto capítulo, se encarga de resumir los resultados y conclusiones más relevantes de las diferentes publicaciones científicas derivadas de esta Tesis doctoral. Además, se incluye una lista de posibles trabajos futuros que podrían ser desarrollados.

Finalmente, los artículos publicados y enviados (status: under review) durante el desarrollo de esta Tesis se presentan.

CAPÍTULO 2

Fundamentos del audio biomédico respiratorio

El audio biomédico respiratorio es la señal de entrada para los diferentes algoritmos o métodos desarrollados en esta Tesis. Por ello, es necesario comprender la naturaleza de estos sonidos, cómo son generados, cómo son grabados durante la auscultación y cómo pueden ser clasificados. En este sentido, este capítulo inicia con la descripción del sistema respiratorio humano encargado de generar estos sonidos, desde un punto de vista anatómico, fisiológico y patológico. En segundo lugar, se describen los principios básicos relacionados con el proceso de auscultación que permite capturar los sonidos del sistema respiratorio. Por último, se realiza una clasificación completa sobre los distintos tipos de sonidos biomédicos producidos por el aparato respiratorio.

2.1. Sistema respiratorio humano

El sistema respiratorio es de vital importancia para el ser humano. Una persona puede vivir varias semanas sin alimento y varios días sin agua, pero solo unos pocos minutos sin oxígeno. En realidad, la mayoría de las personas no podrían sobrevivir sin respirar durante más de 3 minutos, y aunque quisieran aguantar la respiración durante más tiempo, su sistema nervioso autónomo tomaría el control. Esto se debe a que cada célula del cuerpo necesita un suministro continuo de oxígeno para producir energía y crecer, repararse o reconstruirse, así como para mantener las funciones vitales del organismo. Sin embargo, aunque el oxígeno es una necesidad crítica para las células, en realidad es la acumulación de dióxido de carbono lo que impulsa

principalmente la necesidad de respirar. A continuación, se presentan los principios básicos del sistema respiratorio humano, en términos de anatomía, fisiología y patologías obstructivas que puede sufrir. El objetivo no es realizar un estudio exhaustivo sobre un tema tan extenso y complejo como el que nos ocupa, sino exponer los conocimientos base sobre el aparato que genera los sonidos biomédicos respiratorios objeto de estudio en esta Tesis. Para profundizar más en este tema se indica la bibliografía utilizada para desarrollar la información descrita en este sección [239, 166, 152, 322, 95, 190, 141].

2.1.1. Anatomía del sistema respiratorio humano

El sistema respiratorio humano se define como la red de órganos y tejidos que intervienen durante la respiración. De forma genérica, incluye las vías respiratorias, los pulmones, los vasos sanguíneos y los músculos que impulsan los pulmones. Dichas partes trabajan juntas con el principal objetivo de mover el oxígeno por todas las células del cuerpo y de eliminar los gases nocivos del organismo, como el dióxido de carbono. Concretamente, el sistema respiratorio humano comienza su funcionamiento en la cavidad nasal, continúa en la faringe y laringe, los cuales conducen a la tráquea que se ramifica para crear los bronquios, y finalmente bajan por los bronquiolos hasta llegar a los sacos alveolares. Este árbol termina en unas estructuras hinchadas, denominadas alvéolos, que están compuestas por una sola capa de células escamosas, rodeada por una red de capilares. Dentro de los alvéolos se lleva a cabo la principal función del sistema respiratorio, el intercambio de gases (oxígeno y dióxido de carbono). En la Figura 2.1 se puede visualizar las diferentes partes que componen el sistema respiratorio, comenzando por la cavidad nasal y finalizando en los alvéolos, situados en el interior de los pulmones.

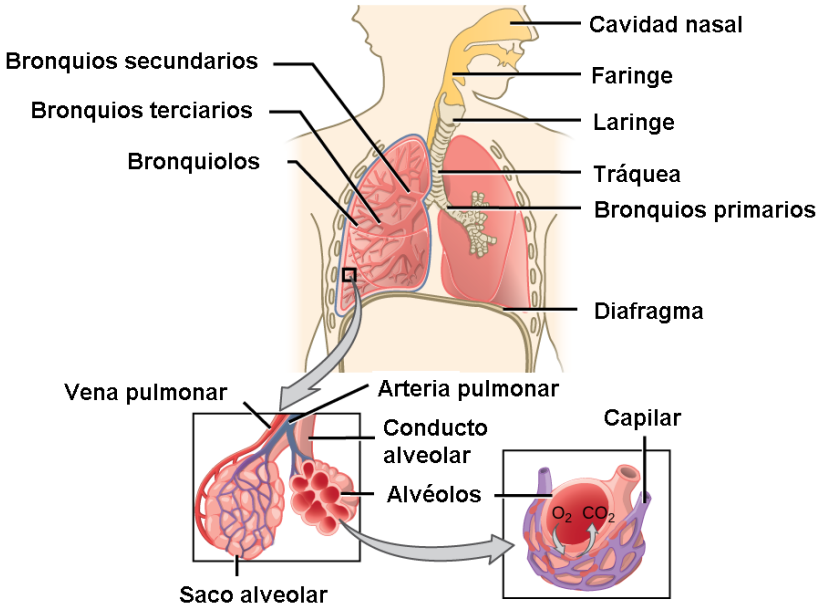

Figura 2.1: Esquema del sistema respiratorio humano. Imagen extraída del enlace `https://bit.ly/3oGV2ZJ`.

Los órganos del sistema respiratorio forman un sistema continuo de cavidades o pasajes, llamado tracto respiratorio, a través del cual el aire entra y sale del cuerpo. Las vías respiratorias se pueden dividir en dos grupos genéricos: las vías respiratorias superiores (tracto respiratorio superior) y las vías respiratorias inferiores (tracto respiratorio inferior). Los principales órganos que componen a cada grupo se muestran en la Figura 2.2.

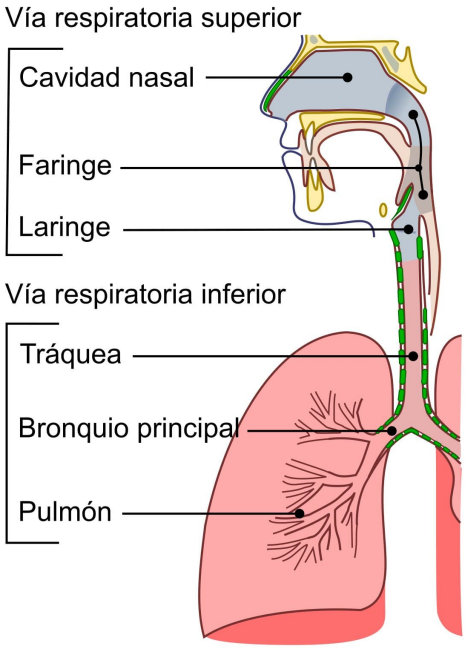

Figura 2.2: División del sistema respiratorio humano en tracto respiratorio superior e inferior. Imagen extraída de la plataforma de dominio publico Wikimedia Commons [35].

Tracto respiratorio superior [152, 322]:

El tracto respiratorio superior consta de órganos ubicados fuera de la caja torácica: cavidad nasal, faringe y laringe. Todos los órganos y estructuras que lo componen participan en la conducción o el movimiento del aire dentro y fuera del cuerpo. Específicamente, los órganos del tracto respiratorio superior proporcionan una ruta para que el aire se mueva entre la atmósfera exterior y los pulmones. También limpian, humedecen y calientan el aire entrante. Sin embargo, en estos órganos no se produce ningún intercambio de gases. Las distintas partes que componen al tracto respiratorio superior son descritas a continuación [152, 322]:

Cavidad nasal: es un espacio lleno de aire situado detrás de la nariz. Específicamente se encuentra ubicado encima del hueso que forma el paladar y se curva hacia abajo y atrás hasta unirse con la garganta. Se divide en dos secciones denominadas fosas nasales. A medida que el aire inhalado fluye a través de la cavidad nasal, se calienta y humedece. Los pelos de la nariz ayudan a atrapar las partículas más grandes que contiene el aire antes de que se adentren más en el tracto respiratorio. Además de sus funciones respiratorias, la cavidad nasal también contiene

quimiorreceptores que son necesarios para el sentido del olfato y que contribuyen de manera importante al sentido del gusto.

Faringe: es una estructura tubular que conecta la cavidad nasal y la parte posterior de la boca con otras estructuras más bajas en la garganta, incluyendo la laringe. La faringe tiene una doble función: por ella pasan tanto el aire como los alimentos (u otras sustancias ingeridas), por lo tanto, es un conducto común que forma parte de los sistemas respiratorio y digestivo. El aire pasa de la cavidad nasal a través de la faringe a la laringe (así como en dirección opuesta). Por otro lado, los alimentos pasan de la boca a través de la faringe al esófago. Este conducto, de unos 12.5 cm de longitud, se divide en tres regiones: nasofaringe, bucofaringe y laringofaringe.

Laringe: es una estructura móvil, que conecta la faringe y la tráquea y ayuda a conducir el aire a través de las vías respiratorias actuando normalmente como una válvula que impide el paso de los elementos deglutidos y cuerpos extraños hacia el tracto respiratorio inferior evitando el atragantamiento. La laringe también se llama órgano fonador porque contiene las cuerdas vocales, que vibran cuando el aire fluye sobre ellas, produciendo así el sonido. Concretamente, está formada por 9 cartílagos (epiglotis, tiroides, cricoides y tres pares de cartílagos más pequeños). Algunos de ellos permiten separar las cuerdas vocales para permitir el proceso de la respiración, mientras que otros permiten mover las cuerdas vocales para producir los diferentes tipos de sonidos que emite el ser humano.

Tracto respiratorio inferior [166, 141, 239]:

El tracto respiratorio inferior consta de una serie de órganos ubicados en la cavidad torácica. La tráquea y otros conductos del tracto respiratorio inferior que conducen el aire procedente del tracto respiratorio superior hasta los pulmones. Estos conductos forman un árbol invertido, conocido como árbol bronquial, con repetidas ramificaciones que surgen a medida que se adentran en los pulmones (bronquios, bronquiolos y conductos alveolares) para conectar finalmente con los sacos alveolares que contienen los alvéolos. En total, hay unos asombrosos 2.000 km de vías respiratorias que conducen el aire a través del tracto respiratorio humano. Sin embargo, sólo en los pulmones (concretamente en los alvéolos) se produce el intercambio de gases entre el aire y la corriente sanguínea. Las distintas partes que componen al tracto respiratorio inferior son descritas a continuación [166, 141, 239]:

Tráquea: es el conducto más amplio de las vías respiratorias. Sus dimensiones suelen estar comprendidas entre unos 10 y 15 cm de largo y 2,5 cm de ancho. Se trata de un tubo cartilaginoso que conecta la laringe con los bronquios primarios de los pulmones, permitiendo el paso del aire. La tráquea se extiende desde la laringe y se ramifica en los dos bronquios principales izquierdo y derecho. Se compone de una serie de cartílagos hialinos (entre 16 y 20) en forma de C. La parte trasera de cada anillo (la parte abierta de la C) está formada por musculo y tejido conectivo. Un tejido húmedo y liso llamado mucosa recubre el interior de la tráquea. La tráquea se ensancha y alarga ligeramente con cada inspiración, volviendo a su tamaño de reposo con cada espiración, por lo que forma un papel fundamental en el proceso de la respiración. El punto

inicial de la tráquea conecta con la laringe mediante un cartílago llamado cricoides y su punto final se ramifica en los bronquios primarios que conectan con cada pulmón.

Los bronquios y las vías respiratorias terminales: los primeros bronquios que se ramifican desde la tráquea son los bronquios principales izquierdo y derecho, también conocidos como bronquios primarios. Su estructura es similar a la de la tráquea, aunque en lugar de anillos presentan laminas cartilaginosas superpuestas. El bronquio principal derecho es corto y ancho, mientras que el izquierdo es largo y estrecho. Los bronquios principales conectan la tráquea con los pulmones e ingresan en su interior a través de una región llamada HILO. Al llegar a los pulmones correspondientes, los bronquios principales se subdividen en bronquios secundarios o lobulares y estos se ramifican en bronquios terciarios más estrechos o bronquios segmentarios. Posteriormente existen otras divisiones de los bronquios segmentarios que se agrupan para formar los denominados bronquios subsegmentarios. Cuando estas cavidades son demasiado estrechas para ser soportadas por el tejido cartilaginoso se denominan bronquiolos cuyas paredes contienen musculo liso y carecen de cartílago. Los bronquiolos continúan ramificándose y dan lugar a los conductos alveolares, los cuales conducen a los sacos alveolares o conjuntos de alvéolos microscópicos. Los alvéolos están constituidos principalmente de epitelio escamoso simple, lo que permite una rápida difusión del oxígeno y el dióxido de carbono. Estos diminutos sacos de aire son las unidades funcionales de los pulmones donde se produce el intercambio de gases entre el aire de los pulmones y la sangre de los capilares. Los dos pulmones pueden contener hasta 700 millones de alvéolos, proporcionando una enorme superficie total para que se produzca el intercambio de gases.

Pulmones: contienen todos los componentes del árbol bronquial más allá de los bronquios primarios y son los órganos más grandes del sistema respiratorio, llegando a ocupar la mayor parte del espacio en la cavidad torácica. Son dos órganos ligeros, elásticos y esponjosos, suspendidos dentro de la cavidad pleural del tórax, y separados por un espacio denominado mediastino, donde se ubica el corazón. El pulmón derecho es más corto que el izquierdo, debido al empuje hacia arriba del hígado sobre el diafragma, aunque tiene menor volumen el izquierdo a causa de la ubicación del corazón. Concretamente el pulmón derecho está dividido en tres lóbulos y el pulmón izquierdo se divide en dos, donde cada lóbulo es alimentado por uno de los bronquios secundarios. En la superficie mediastínica de cada pulmón se localiza el HILO pulmonar, el único punto de unión por donde pasan los vasos sanguíneos, los linfáticos, los bronquios y las fibras nerviosas hacia cada pulmón. Cada pulmón está cerrado por una membrana de doble capa, llamada pleura. La capa interna (pleura visceral) está adherida a la superficie del pulmón. La capa externa (pleura parietal) se adhiere a la pared torácica, al mediastino y al diafragma. El pequeño espacio comprendido entre la pleura visceral y parietal se denomina cavidad pleural. Dicha cavidad contiene una fina película de fluido que actúa como lubricante para reducir la fricción mientras las dos capas se deslizan una contra la otra, y ayuda a mantener las dos capas juntas mientras los pulmones se expanden y se contraen durante la respiración.

Músculos respiratorios: además de la gran cantidad de conductos que componen el árbol bronquial se suele incluir como parte del tracto respiratorio inferior a los músculos que intervienen en la respiración. El más importante es un gran músculo llamado diafragma con forma

de cúpula que se curva hacia los pulmones y separa el tórax del abdomen. Cuando se contrae, se aplana y por lo tanto aumenta el volumen de la cavidad torácica. Del mismo modo, la contracción de los músculos intercostales externos mueve las costillas hacia arriba y hacia fuera. Este aumento de volumen lleva a una caída de la presión dentro de los pulmones, permitiendo que el aire fluya pasivamente hacia las vías respiratorias. Aunque el diafragma y los músculos intercostales toman el papel clave dentro del proceso de la respiración, su acción respiratoria es asistida y aumentada por un complejo conjunto de otros grupos de músculos (abdominales, escalenos, esternocleidomastoideo, etc.). Además, los músculos de la laringe, la faringe y la cavidad nasal ajustan la resistencia del movimiento de los gases a través de las vías respiratorias superiores durante la inspiración y la espiración.

2.1.2. Fisiología del sistema respiratorio humano

Aunque el número de funciones que desempeña el sistema respiratorio humano es amplio, la función principal es proporcionar al cuerpo un intercambio de gases con el aire atmosférico asegurando, por un lado, la concentración permanente de oxígeno en la sangre necesaria para las reacciones metabólicas y, por otro lado, actuando como medio para la eliminación de los gases residuales del organismo que resultan de ellas. Esto se consigue gracias al desempeño de otras funciones más específicas como son: regulación y control de la respiración, mecanismo de la respiración (inspiración y espiración) e intercambio de gases.

Regulación y control de la respiración [95, 190]:

La respiración es un acto automático y rítmico regulado por el sistema nervioso. Los centros que controlan la respiración están situados en el tallo cerebral. Concretamente existen dos, uno situado en el bulbo raquídeo (neumotáxico) y otro situado en la protuberancia (apnéustico). Por un lado, el centro neumotáxico es el encargado de provocar la inspiración, mientras que el centro apnéustico se encarga de producir la espiración. Las redes neuronales dirigen los músculos que forman las paredes del tórax y el abdomen para producir un gradiente de presión que mueve el aire dentro y fuera de los pulmones. El ritmo respiratorio se regula mediante la interconexión recíproca de estímulos e inhibiciones de las neuronas del tronco encefálico. Además, existen unos sensores distribuidos por todo el cuerpo encargados de enviar señales a los centros que controlan la respiración. Por un lado, los quimiorreceptores detectan los cambios en los niveles de oxígeno en la sangre y, por otro lado, los mecanorreceptores controlan la expansión del pulmón, el tamaño de las vías respiratorias y la fuerza de contracción de los músculos respiratorios.

El sistema respiratorio humano cuenta con la capacidad de regular la frecuencia y la profundidad de la respiración en función de los cambios producidos en el medio interno o externo. Estas variaciones se producen principalmente cuando los niveles de dióxido de carbono son altos o cuando los niveles de oxígeno son bajos. Considerando las modificaciones de la profundidad y frecuencia respiratoria que se pueden dar, se definen tres tipos de regulaciones:

hiperpnea, hipopnea y apnea. La hiperpnea consiste en el aumento de la profundidad y frecuencia respiratoria, la hipopnea consiste en la disminución de la frecuencia respiratoria y la apnea se define como la suspensión completa de la respiración. En definitiva, la frecuencia respiratoria se adapta a la demanda de oxígeno por las células y a la necesaria eliminación de dióxido de carbono. Por otro lado, el sistema respiratorio tiene la capacidad de regularse cuando se produce alguna perturbación en las vías respiratorias, como por ejemplo un ataque de asma. Por último, el proceso de la respiración es regulado cuando la mecánica de los músculos respiratorios es alterada, como por ejemplo por la realización de ejercicio físico.

Aunque, los músculos respiratorios comentados anteriormente aumentan considerablemente la flexibilidad del acto de respirar, también complican la regulación de la respiración. Realmente los músculos respiratorios desempeñan otras funciones (como mantener la postura o incluso hablar), por ello el proceso de la respiración puede verse afectado por los centros cerebrales superiores llegando a provocar un control voluntario sobre la mecánica de la respiración. Por ello, el ser humano puede aguantar la respiración o respirar con mayor o menor frecuencia voluntariamente.

Mecánica de la respiración (inspiración y espiración) [141, 95]:

El intercambio de aire entre los pulmones y el medio ambiente es un proceso cíclico en el que se repiten dos movimientos: inspiración (inhalación) y espiración (exhalación). Estos procesos son vitales para proporcionar oxígeno a las células y eliminar el dióxido de carbono del cuerpo. La inspiración es un movimiento activo en el que se produce un ensanchamiento de la caja torácica y la consiguiente inhalación de aire. Por otro lado, la espiración es un movimiento pasivo en el que la reducción de la caja torácica provoca la expulsión de aire desde los pulmones hacia el exterior. La Figura 2.3 muestra el procedimiento de la mecánica de la respiración en la fase de inspiración y espiración.

En la mecánica de la respiración, la contracción y relajación de los músculos respiratorios actúa para cambiar el volumen de la cavidad torácica. A medida que la cavidad torácica y los pulmones se contraen o se expanden se produce un cambio en el volumen de los pulmones modificando a su vez la presión del interior de los pulmones. Desde un punto de vista físico, la ley de Boyle establece que, en un espacio cerrado, el volumen del gas es inversamente proporcional a la presión del gas (cuando la temperatura es constante) [112]. En este sentido, cuando el volumen de la cavidad torácica aumenta, el volumen del gas (aire) contenido en los pulmones también aumenta y la presión del aire dentro de los pulmones disminuye. Sin embargo, cuando el volumen de la cavidad torácica disminuye, el volumen del aire contenido en los pulmones también disminuye y la presión del aire dentro de los pulmones aumenta. El gradiente de presión que existe entre el medio ambiente (exterior) y el interior de los pulmones hace que el aire fluya desde la zona de mayor presión a la de menor presión.

Proceso de inspiración o inhalación: la inspiración es la fase del ciclo respiratorio en la que el aire entra en los pulmones. Durante este proceso el volumen de la caja torácica aumenta en tres direcciones:

Figura 2.3: Mecánica de la respiración (fase de inspiración y espiración). Esta figura ha sido obtenida a través del siguiente enlace `https://n9.cl/as6i`.

- Aumento del diámetro vertical: producido por la contracción del diafragma que desciende hacia el abdomen.
- Aumento del diámetro anteroposterior: producido por la elevación de la caja torácica, provocada por la contracción de varios pares de músculos, intercostal externo, esternocleidomastoideo, pectoral menor, escaleno y serrato anterior.
- Aumento del diámetro izquierda-derecha: producido por la elevación lateral de las costillas al contraerse los músculos intercostales.

La dilatación de la caja torácica implica una expansión de la pleura parietal que se transmite a la pleura visceral y a los pulmones. De este modo, la presión del gas contenido en los pulmones disminuye, por lo que se produce una succión de aire hasta conseguir una presión equivalente a la atmosférica. Tal como se ha comentado anteriormente, un aumento del volumen pulmonar resulta en una disminución de la presión dentro de los pulmones. La presión del ambiente externo a los pulmones es ahora mayor que la presión del aire dentro de los pulmones, lo que significa que el aire se mueve dentro de los pulmones debido al gradiente de presión.

Proceso de espiración o exhalación: la espiración es la fase de ventilación en la que el aire es expulsado de los pulmones. La relajación de los músculos respiratorios implica una disminución del volumen de la caja torácica, suficiente para permitir la salida del aire hacia el exterior. La contracción de la caja torácica se transmite a los pulmones. En ese momento la presión del gas contenido en los pulmones aumenta, por lo que el aire es expulsado hasta conseguir igualarse con la presión atmosférica. Según la ley de Boyle, una disminución del volumen pulmonar resulta en un aumento de la presión dentro de los pulmones. La presión

dentro de los pulmones es ahora mayor que en el ambiente externo, lo que significa que el aire sale de los pulmones debido al gradiente de presión.

Intercambio de gases y transporte de gases en la sangre [190, 141]:

El intercambio de gases en los pulmones tiene lugar entre el aire que llega a los alvéolos y la sangre que fluye por los capilares. El proceso de intercambio de gases pulmonares elimina el dióxido de carbono de la sangre y repone el suministro de oxígeno en la sangre. El sistema circulatorio es el encargado de transportar los gases desde los pulmones a los tejidos de todo el cuerpo y viceversa.

El intercambio de gases se produce gracias a un gradiente de presión existente entre los alvéolos y los capilares, a través de un proceso conocido como difusión. Cuando se produce el mecanismo de la respiración, el proceso de inspiración introduce a los pulmones un aire compuesto por una mezcla de gases que incluyen oxígeno y dióxido de carbono. Cada uno de estos gases tiene una presión distinta, relacionada con su concentración dentro de la mezcla de gases. Estas presiones individuales se denominan presiones parciales. La diferencia de presiones parciales entre los gases de los alvéolos y los capilares crean un gradiente de presión a través de la membrana respiratoria (membrana que separa los alvéolos y los capilares sanguíneos). Si la presión a cada lado de la membrana fuera la misma, no habría intercambio de gases. Es la variación en las presiones parciales de oxígeno y dióxido de carbono la que da a lugar a este proceso.

Los gases se mueven en ambas direcciones (hacia los capilares de la sangre y hacia los alvéolos de los pulmones). Concretamente, los gases se mueven de un área de presión alta (alta concentración) a un área de presión baja (baja concentración). Durante la etapa de inspiración, la concentración de oxígeno en los alvéolos es alta y en los capilares sanguíneos es baja, por ello el oxígeno viaja de los alvéolos a los capilares gracias al gradiente de oxígeno. Por el contrario, existe un gradiente inverso para el dióxido de carbono que se difunde desde la sangre a los alvéolos para ser expulsado posteriormente al exterior por el proceso de espiración. La Figura 2.4 muestra una imagen representativa del procedimiento que se lleva a cabo durante el intercambio de gases en los alvéolos.

Una vez que el oxígeno ha atravesado la membrana respiratoria y llega a la sangre pulmonar, es transportada hasta los capilares de los tejidos para que se difunda al interior de las células. Por otro lado, el dióxido de carbono se difunde desde las células del organismo hacia la sangre. El transporte del oxígeno y del dióxido de carbono por la sangre se realiza principalmente por la hemoglobina, una proteína situada en el interior de los glóbulos rojos. La sangre con un alto nivel de oxígeno que viaja desde los pulmones hacia todas las células del cuerpo se suele denominar sangre oxigenada y es de color rojo brillante debido a la unión de la hemoglobina y el oxígeno. En cambio, la sangre con un alto nivel de dióxido de carbono que viaja desde las células del organismo hasta los pulmones se suele denominar sangre desoxigenada y es de color rojo mucho más oscuro debido a la falta de oxígeno disponible para unirse a la hemoglobina.

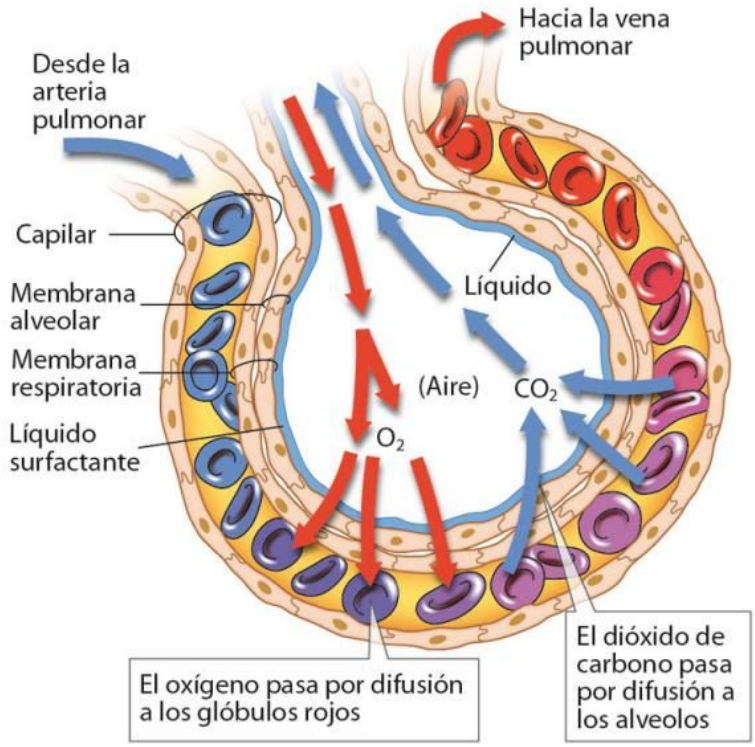

Figura 2.4: Intercambio de gases en los alvéolos. Esta figura ha sido obtenida a través del siguiente enlace `https://bit.ly/2JfRahT`.

2.1.3. Patologías obstructivas del sistema respiratorio humano

Existe un amplio abanico de patologías relacionadas con el sistema respiratorio, las cuales se suelen clasificar en función de los órganos, músculos y tejidos a los que afecta. El objetivo de esta sección es presentar una introducción a las patologías que afectan a las vías respiratorias del árbol bronquial, las cuales son las responsables de producir los denominados sonidos adventicios (como por ejemplo, las sibilancias). La presencia de estos sonidos, generados por la obstrucción de las vías respiratorias, indica una anomalía en el correcto funcionamiento de la mecánica de la respiración, aportando información relevante al médico para determinar el diagnóstico del paciente. A continuación, se muestra una descripción de las patologías más comunes y relevantes que afectan a las vías respiratorias y que están relacionadas principalmente con la presencia de sonidos sibilantes. Dichas patologías son: asma, enfermedad pulmonar obstructiva crónica (EPOC), bronquiolitis y bronquiectasia.

Asma [50]:

El asma es una enfermedad respiratoria común que afecta a más de 339 millones de personas a nivel mundial [23], siendo la enfermedad crónica más frecuente en los niños a nivel global. El asma se caracteriza principalmente por síntomas como: la presencia de sibilancias, falta de

aliento, opresión en el pecho y/o tos, y por una limitación variable del flujo de aire respiratorio. Estos síntomas pueden variar sus características en tiempo e intensidad. Principalmente estas variaciones están provocadas por factores como: ejercicio, exposición a alérgenos o irritantes, contaminación atmosférica, cambio del clima o las infecciones respiratorias virales. Los síntomas y la limitación del flujo de aire pueden resolverse espontáneamente o en respuesta a la medicación, y a veces pueden estar ausentes durante semanas o meses. Por otra parte, los pacientes pueden experimentar brotes episódicos de asma (los conocidos ataques de asma), en los cuales el revestimiento de los conductos bronquiales se hincha, lo que provoca el estrechamiento de las vías respiratorias y reduce el flujo de aire que entra y sale de los pulmones. El asma suele estar asociado con la hipersensibilidad de las vías respiratorias a los estímulos directos o indirectos, y con la inflamación crónica de las vías respiratorias. Estas características suelen persistir, incluso cuando no hay síntomas o la función pulmonar es normal, pero pueden normalizarse con el tratamiento. Los síntomas recurrentes de asma frecuentemente pueden causar insomnio, fatiga diurna, reducción de los niveles de actividad y ausentismo escolar y laboral. El asma es una enfermedad con una tasa de mortalidad relativamente baja, ya que con una correcta medicación se pueden evitar los desencadenantes del asma y la gravedad de la misma puede ser controlada. Sin embargo, la OMS estima que en el año 2016 se produjeron cerca de 420.000 muertes por asma en todo el mundo [23], siendo los países con sistemas sanitarios débiles los más afectados. Por lo tanto, es fundamental llevar a cabo una detección temprana de la enfermedad para poder aplicar un tratamiento adecuado evitando posibles demoras y el agravamiento en la calidad de vida de los pacientes.

EPOC [22]:

La enfermedad pulmonar obstructiva crónica (EPOC) es un término genérico que abarca varias enfermedades respiratorias que provocan dificultades respiratorias y empeoran a lo largo del tiempo. La OMS [24] destaca que en 2016 la prevalencia de las EPOC era superior a los 250 millones de casos en todo el mundo y que en el año 2015 murieron aproximadamente 3 millones de personas a causa de una EPOC, lo cual representa el 5 % de todas las muertes registradas durante ese año. En el caso de una persona sana, las vías respiratorias y los sacos de aire de los pulmones (sacos alveolares) son elásticos. Cuando se inicia la mecánica de la respiración, durante la fase de inspiración, las vías respiratorias llevan el aire a los sacos alveolares para que se inicie el proceso de intercambio de gases y se pueda aportar oxígeno a todas las partes del organismo. Después, durante la fase de espiración, el dióxido de carbono extraído de la sangre es expulsado al exterior. Sin embargo, en el caso de las personas con EPOC, el proceso descrito anteriormente se ve deteriorado debido principalmente a las siguientes razones: las vías respiratorias y las paredes de los sacos alveolares se vuelven menos elásticas y rígidas, las paredes de gran parte de los sacos alveolares se destruyen, las paredes de las vías respiratorias se vuelven más gruesas e inflamadas y las vías respiratorias producen más moco de lo habitual llegando incluso a obstruirse. Los síntomas más comunes que suelen experimentar las personas con EPOC son: tos crónica, falta de aliento al realizar actividades cotidianas (disnea), infec-

ciones respiratorias, aumento de fatiga y cansancio, aumento de mucosidad (flemas o esputos), presión en el pecho y aparición de sonidos sibilantes durante la mecánica de la respiración. Las principales causas que motivan la aparición de la EPOC son: exposición al humo del tabaco (fumadores activos y pasivos), contaminación del aire en interiores/exteriores, exposición a polvos y productos químicos e infecciones repetidas de las vías respiratorias inferiores durante la infancia. La EPOC que afecta a la mayoría de las personas se traduce en la aparición de enfisema y bronquitis crónica en función de las estructuras o tejidos dañados:

- Enfisema [298]: afecta a los sacos alveolares de los pulmones, así como a las paredes entre ellos. Pierden elasticidad y pueden llegar a dañarse.
- Bronquitis crónica [21]: afecta al revestimiento de las vías respiratorias, las cuales se irritan e inflaman constantemente. Esto hace que las vías respiratorias se hinchen, dificultando la entrada y salida de aire, y se produzca una mayor mucosidad.

Bronquiolitis [80]:

La bronquiolitis es una infección pulmonar de gran importancia a nivel mundial. Según el boletín de la OMS [270] se estima que cada año se producen 150 millones de nuevos casos, de los cuales entre los 11-20 millones (7-13 %) son lo suficientemente graves como para requerir el ingreso en el hospital. En todo el mundo, el 95 % de los casos se producen en países en desarrollo. Específicamente, la bronquiolitis es una enfermedad respiratoria causada principalmente por un virus que afecta a las vías respiratorias de menor tamaño (bronquiolos). La función de los bronquiolos es controlar el flujo de aire en los pulmones. Cuando estas vías presentan una infección o son dañadas, pueden inflamarse o congestionarse. Esto disminuye o bloquea el flujo de oxígeno que debe llegar a los alvéolos para producir el proceso de intercambio de gases con normalidad. Aunque generalmente esta enfermedad suele ser común en niños pequeños o bebés, la bronquiolitis también puede afectar a los adultos. El virus más común asociado con la bronquiolitis es el virus respiratorio sincitial (VRS). Sin embargo, a lo largo de los años, se ha descubierto que muchos otros virus causan la misma infección, entre los que se incluyen los siguientes: rinovirus humano, coronavirus, metapneumovirus humano, adenovirus, los virus parainfluenza y bocavirus humano. Los principales síntomas asociados a esta patología son: falta de aliento y fatiga, presencia de sonidos sibilantes audibles durante la respiración, aparición de sonidos crepitantes, respiración muy rápida, respiración dificultosa, tos, congestión nasal y catarro. Existen dos tipos principales de bronquiolitis, los cuales son:

- La bronquiolitis viral [114]: suele aparecer principalmente en los bebés y se debe a la presencia de un virus en los bronquiolos.
- La bronquiolitis obliterante o bronquiolitis constrictiva [171]: es una condición rara y peligrosa que se observa en los adultos. Esta enfermedad causa cicatrices en los bronquiolos. Esto bloquea los conductos de aire creando una obstrucción de las vías respiratorias que no se puede revertir.

Bronquiectasia [308]:

La bronquiectasia tiene un perfil creciente dentro de las enfermedades del aparato respiratorio. Actualmente los científicos la definen como una "epidemia mundial" emergente y la califican como un problema clínico en evolución, debido a la falta de terapia y la falta de comprensión de su inherente heterogeneidad [75]. Datos recientes sugieren que esta patología tiene una prevalencia de hasta 566/100.000 sujetos de toda la población con un aumento de aproximadamente el 40 % en la última década [258]. Concretamente, la bronquiectasia es una patología en la que los conductos bronquiales que componen el árbol bronquial sufren un ensanchamiento anómalo. En un paciente sano, las pequeñas glándulas, situadas en el revestimiento de las vías respiratorias, producen una pequeña cantidad de moco que permite mantener estos conductos húmedos y atrapar el polvo y la suciedad del aire inhalado. Sin embargo, debido al ensanchamiento producido por esta patología, la mucosidad y las bacterias tienden a acumularse en estas zonas ensanchadas. Esto puede provocar infecciones frecuentes y el bloqueo de las vías respiratorias. Los síntomas de la bronquiectasia pueden tardar meses o incluso años en desarrollarse. Algunos de los síntomas típicos que se pueden presentar son: tos crónica diaria, toser sangre, sonidos anormales o sibilancias en el pecho al respirar, falta de aliento, dolor en el pecho, toser grandes cantidades de moco espeso todos los días, fatiga o cansancio diario e infecciones respiratorias frecuentes. Aunque, no existe cura para esta enfermedad, es manejable. Con un tratamiento efectivo, se puede llevar una vida dentro de la normalidad. Sin embargo, los brotes deben ser tratados rápidamente para mantener el flujo de oxígeno que debe llegar al resto del cuerpo y prevenir un mayor daño de las vías respiratorias. Se pueden definir dos categorías principales para esta patología, las cuales son:

- Bronquiectasia por fibrosis quística. La fibrosis quística (FQ) [86] es una condición genética que causa una producción anormal de mucosidad.
- Bronquiectasia no relacionada con la fibrosis quística [308]. Las afecciones conocidas más comunes que pueden conducir a esta patología son: un sistema inmunológico con un funcionamiento anormal, enfermedad inflamatoria del intestino, enfermedades autoinmunes, EPOC, virus de la inmunodeficiencia humana (VIH), aspergilosis (una reacción pulmonar alérgica a los hongos) e infecciones pulmonares (como por ejemplo la tos ferina o la tuberculosis).

2.2. Proceso de auscultación

Los médicos han estado escuchando el cuerpo de los pacientes para efectuar su diagnóstico probablemente desde que comenzó la conocida práctica de la curación. En concreto, fue Hipócrates quien inició el concepto de auscultación aplicando el oído sobre el pecho del paciente para oír los sonidos respiratorios transmitidos en el interior y llamó a este procedimiento "auscultación inmediata o directa". Sin embargo, el concepto del estetoscopio no surgió hasta 1816. El médico francés René Laënnec necesitaba escuchar los sonidos que se producían en el pecho

de un paciente, así que enrolló un largo trozo de papel en forma de tubo y observó que con aquel dispositivo podía oír mucho mejor que colocando su oído directamente sobre el pecho del paciente. Laënnec publicó su obra maestra [180] en 1819 y fue en ese momento cuando comenzó el denominado arte de la auscultación que rápidamente se hizo popular en todo el mundo. Laënnec acuñó el nombre estetoscopio a partir de dos palabras griegas: stethos (pecho) y skopein (observar). También denomino al proceso como auscultación a partir de la palabra en latín auscultare (escuchar). Veinticinco años después de que Laënnec inventara el estetoscopio, George P. Camman desarrolló un diseño que incluía un auricular para cada oreja. Los profesionales médicos continuaron usando este diseño con pocos cambios durante casi un siglo. No fue hasta principios de los años 60 cuando el Dr. David Littmann patentó un nuevo diseño que mejoró significativamente el rendimiento acústico del estetoscopio. Unos años después la empresa 3M adquirió el negocio de estetoscopios del Dr. Littmann. Finalmente, los diseños del Dr. Littmann se convirtieron en el nuevo estándar de los estetoscopios, y ahora 3M Littmann es la marca de mayor confianza en el negocio [15].

Aunque el estetoscopio surgió hace más de 200 años, el proceso de auscultación sigue siendo el primer examen clínico que un médico utiliza para realizar un diagnóstico de las posibles patologías obstructivas presentes en el sistema respiratorio humano. Esto es debido a que la auscultación del sistema respiratorio es una técnica de diagnóstico de bajo coste, no invasiva, segura y fácil de realizar [274]. Además, la capacidad de diferenciar entre los sonidos normales y anormales (sonidos adventicios) sigue siendo esencial en la práctica clínica para conseguir un diagnóstico eficiente. Gracias a la invención del estetoscopio electrónico y de sensores y micrófonos especializados, en las últimas dos décadas, han aparecido contribuciones relevantes que atienden al análisis de los sonidos del sistema respiratorio para ayudar en la mejora del diagnóstico, así como en la educación y la pedagogía [43]. El objetivo de esta sección es presentar la información más relevante para comprender los principios de la auscultación respiratoria, sus ventajas y limitaciones, las partes y los tipos de estetoscopios, los estetoscopios electrónicos comerciales más importantes (utilizados por médicos), los micrófonos o sensores que se utilizan como alternativa para capturar los sonidos del interior del paciente y las técnicas alternativas que existen para analizar y diagnosticar el sistema respiratorio humano. Para ello se han utilizado algunos de los artículos y bibliografía más relevantes en el campo de la auscultación [104, 274, 43, 83, 224].

2.2.1. Principios de la auscultación respiratoria

Para llevar a cabo el examen clínico sobre el estado del sistema respiratorio es habitual el uso de un enfoque de evaluación ampliamente conocido y utilizado en medicina, en el cual la auscultación desempeña un papel fundamental. Este enfoque de evaluación es conocido por sus siglas IPPA [104, 255], ya que se puede dividir en las siguientes cuatro etapas:

- **I**nspección: el uso de los sentidos de la vista, el olfato y el oído para observar la condición normal, o cualquier desviación de la normalidad, de la persona en su conjunto, así como de una zona en particular.

- **P**alpación: tocar y sentir partes del cuerpo con las manos para determinar la temperatura, la textura, la humedad, el movimiento y la consistencia de las estructuras. En cuanto a la evaluación respiratoria se puede detectar la simetría de los pulmones.

- **P**ercusión: la transmisión del sonido que se produce cuando el médico golpea con un dedo con trazos cortos y agudos contra otro dedo que está colocado firmemente sobre un órgano determinado. Esto permite determinar la densidad de las estructuras que se encuentran dentro de una cavidad, como por ejemplo la cavidad torácica. Los pulmones deben sonar huecos a la percusión porque están llenos de aire.

- **A**uscultación: escuchar el movimiento del aire que se produce en los conductos de aire que componen el tracto respiratorio superior e inferior para determinar la presencia de los diversos sonidos normales y anormales que ocurren durante la mecánica de la respiración.

En términos generales, la auscultación consiste en escuchar los sonidos internos del cuerpo, generalmente utilizando un estetoscopio. La auscultación se realiza con el fin de examinar el sistema circulatorio y el sistema respiratorio (sonidos cardíacos y respiratorios), así como el sistema gastrointestinal (sonidos intestinales). Este proceso forma parte del examen clínico de un paciente y se utiliza habitualmente para aportar pruebas sólidas que incluyan o excluyan diferentes condiciones patológicas que se manifiestan clínicamente en el paciente. En el caso del sistema circulatorio, el médico examina los cuatro focos principales donde los sonidos de las válvulas cardíacas son más fuertes (aórtico, pulmonar, mitral y tricúspide). Durante la escucha el médico pone especial atención a cómo suena el corazón, la frecuencia con la que se produce cada sonido y la intensidad con la que suena. Los sonidos cardíacos tradicionalmente son rítmicos a corto plazo, entre 60-100 pulsaciones/min en un adulto sano, por lo que una cierta variación puede indicar que algunas zonas pueden no recibir suficiente sangre o que algunas válvulas pueden presentar una fuga. En el caso del sistema gastrointestinal, el médico examina diferentes regiones del abdomen para escuchar los diferentes sonidos presentes. En condiciones normales se deben de escuchar los sonidos en cualquier región del abdomen. La ausencia de sonido en alguna región puede indicar que el material digerido puede estar atascado o el intestino puede estar torcido. Por último, en el caso del sistema respiratorio, el médico examina los focos principales del árbol bronquial. El flujo de aire suena diferente cuando las vías respiratorias están bloqueadas, estrechadas o llenas de líquido y mucosidad. La detección de estos sonidos anormales es fundamental para detectar la patología respiratoria. Centrándonos en el objeto de estudio de esta Tesis, esta sección describe los principios básicos de la auscultación considerando los sonidos biomédicos producidos por el sistema respiratorio, los cuales son considerados sonidos rítmicos a largo plazo, entre 12-20 respiraciones/min en un adulto sano.

La auscultación tradicional de los sonidos respiratorios suele caracterizarse por una serie de consideraciones que permiten conseguir una interpretación eficiente de los sonidos respiratorios [274, 46], las cuales se resumen en la siguiente lista:

- Para comenzar, el ambiente óptimo [104] durante la auscultación respiratoria debe estar caracterizado por un entorno: (i) silencioso, ya que el ruido ambiente puede interferir en

la escucha de los sonidos de interés, por ello el paciente debe permanecer en silencio; (ii) caliente, para que el paciente se sienta cómodo y para evitar el escalofrío que puede ser añadido como ruido interferente; y (iii) bien iluminado, para detectar correctamente las posiciones que deben ser auscultadas.

- Durante el examen, el paciente deberá respirar un poco más profundo de lo normal por la boca. Esto permite aumentar la intensidad de los sonidos respiratorios de interés, facilitando su detección.

- El proceso de auscultación debe realizarse alrededor de la superficie torácica, la cual se divide en tres caras: cara posterior del tórax, cara anterior del tórax y cara lateral del tórax.

- La auscultación debe realizarse de forma sistemática y comparativa entre el lado izquierdo y derecho, para detectar cualquier asimetría entre los sonidos de ambos pulmones. Para ello se suele utilizar un enfoque denominado "stepladder", el cual consiste en auscultar las diferentes regiones siguiendo un patrón en zig-zag (ver Figura 2.5).

- El procedimiento se inicia auscultando los 9 focos que componen la cara anterior y las caras laterales del tórax, como se muestra en la Figura 2.5A (los puntos 5 y 9 corresponden a las caras laterales). Para ello se utiliza un patrón en zig-zag que comienza en la región supraclavicular derecha y desciende comparando el lado izquierdo y el derecho sucesivamente. Siguiendo el mismo enfoque se auscultan los 10 focos que componen la cara posterior y las caras laterales del tórax, como se muestra en la Figura 2.5B (los puntos 5, 7 y 10 corresponden a las caras laterales). Es importante evitar la escapula, ya que los sonidos pulmonares no pueden escucharse a través del hueso.

- Durante la auscultación debe escucharse al menos un ciclo de respiración en cada posición auscultada para apreciar los sonidos generados tanto en la fase de inspiración como en la fase de espiración.

- Para poder discriminar entre los sonidos respiratorios adventicios y los sonidos respiratorios normales es fundamental poder identificar las propiedades o atributos que caracterizan los sonidos respiratorios auscultados: frecuencia, tono, intensidad, sonoridad, timbre, duración, etc. En la Sección 2.3 se detallan los atributos anteriormente indicados.

2.2.2. Ventajas y limitaciones

El proceso de auscultación cuenta con una serie de ventajas, en comparación con el resto de técnicas para el diagnóstico del sistema respiratorio, que lo convierten en el primer examen clínico que se lleva a cabo en cualquier centro de salud para realizar el diagnóstico del aparato respiratorio [274]. Las principales ventajas que se definen son:

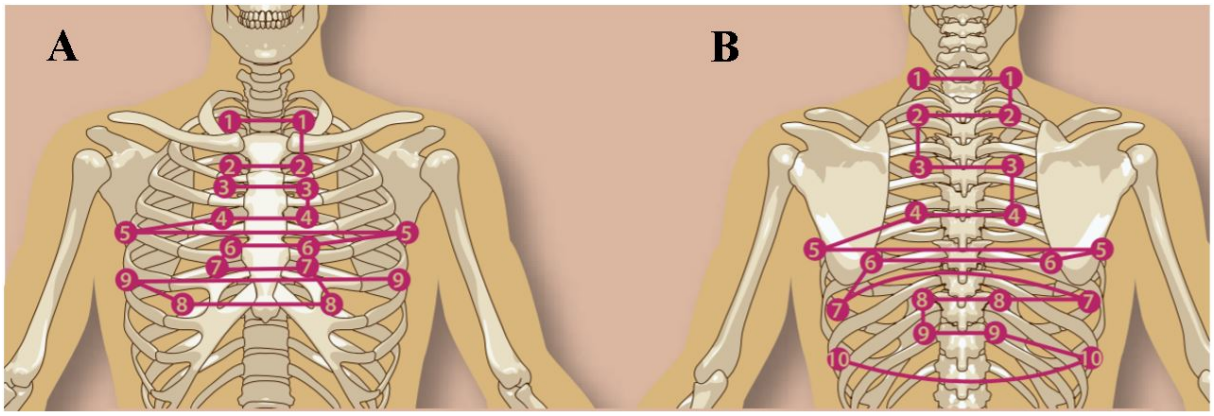

Figura 2.5: Focos de la caja torácica para la auscultación del sistema respiratorio. A) Cara anterior de la caja torácica. B) Cara posterior de la caja torácica. Imagen extraída de la referencia [46].

- Es una técnica fácil de utilizar que cualquier médico puede llevar cabo. En cambio, en el caso de otras técnicas como las radiografías se necesita a un especialista (radiólogo) para llevar a cabo el examen clínico y la instrumentación requerida es de mayor coste.
- Es una técnica rápida de realizar que permite evitar la aglomeración en los centros de salud.
- Es una técnica de bajo coste, esto permite que el método pueda ser utilizado en cualquier parte del mundo independientemente de su nivel económico. Cualquier centro sanitario puede contar con un estetoscopio para aplicar el examen, sin embargo, no en todos los centros sanitarios ni en todos los países se pueden permitir disponer de una máquina de rayos X.
- Es una técnica no invasiva y segura tanto para el paciente como para el especialista que realiza el examen. En contraposición, los rayos X, por ejemplo, contemplan una serie de efectos que resultan especialmente nocivos, entre los que destacan la posibilidad de desarrollar cáncer, posibles quemaduras en la piel, la aparición de cataratas, etc.
- Es una técnica que permite detectar los sonidos adventicios, los cuales indican la presencia de patologías respiratorias obstructivas, y por consiguiente permite elaborar un diagnóstico completo del sistema respiratorio. En este sentido, la precisión de los sonidos adventicios para la identificación de las patologías respiratorias, así como su estado de gravedad ha sido ampliamente confirmado [88, 217, 205, 104].

Por el contrario, el proceso de auscultación cuenta con una serie de limitaciones que componen la principal motivación de esta Tesis. Los inconvenientes que presenta esta técnica giran en torno a la capacidad cognitiva que tienen los médicos para interpretar correctamente los diferentes sonidos biomédicos auscultados. Las principales limitaciones o inconvenientes que sufre la auscultación, son las siguientes:

- En primer lugar, el diagnóstico que se deriva de la auscultación sigue siendo un diagnóstico altamente subjetivo que se encuentra condicionado a la habilidad, experiencia y entrenamiento de cada médico en la escucha de dichas señales sonoras. Esta habilidad subjetiva para el análisis de las señales sonoras auscultadas implica que en numerosas ocasiones el médico no determine correctamente el origen de la posible enfermedad del paciente. Por ello la fiabilidad de la auscultación ha sido legítimamente cuestionada, debido a la alta variabilidad de las observaciones entre especialistas [207, 104].

- Por otro lado, los sonidos respiratorios normales y los sonidos adventicios (anómalos e indicadores de un desorden pulmonar) se encuentren mezclados simultáneamente en tiempo y frecuencia. Esto origina que la aparición de los sonidos respiratorios normales, interfiera en la escucha de los sonidos adventicios de interés, entorpeciendo la capacidad cognitiva del médico al ser distraído por dichos sonidos interferentes lo que probablemente causará un aumento de falsos negativos en los diagnósticos realizados. En un estudio reciente se ha demostrado que los médicos no consiguen detectar parte de los sonidos adventicios debido al solapamiento de los sonidos respiratorios durante la inspiración y la espiración [42].

- Del mismo modo, el ruido ambiente que rodea al sujeto durante la auscultación provoca una interferencia en la escucha de los sonidos adventicios de interés que limita la capacidad cognitiva del médico [179].

- La capacidad del sistema auditivo humano está limitada. El médico especialista debe ser capaz de analizar las propiedades que definen al sonido (frecuencia, amplitud, timbre, etc) para poder determinar el tipo de sonido adventicio escuchado. Sin embargo, discriminar entre sonidos adventicios con características similares, como es el caso de sonidos sibilantes monofónicos y polifónicos, es una tarea bastante compleja de realizar mediante auscultación [305].

- Por último, es común que la capacidad cognitiva del médico se reduzca a lo largo del día, ya que el número de horas dedicadas a analizar los sonidos respiratorios aumenta, un hecho que se ve exacerbado por el estrés que el médico sufre con ciertos casos médicos [324, 154].

Aprovechando las ventajas de la auscultación y considerando las limitaciones que contempla se han definido los objetivos de esta Tesis. A causa de estas limitaciones surge la necesidad de diseñar algoritmos y métodos que permitan solventar los inconvenientes de la auscultación para aumentar la eficiencia de los diagnósticos, evitando diagnósticos erróneos que pueden poner en riesgo la salud de los pacientes. La necesidad de un sistema complementario a la decisión clínica para diagnosticar los posibles trastornos pulmonares del paciente, aplicando eSalud (eHealth), se ha vuelto vital en los últimos años para evitar poner en riesgo la salud de los pacientes, con la necesidad adicional de hacer sostenibles y eficientes los sistemas sanitarios [189, 279].

2.2.3. Tipos de estetoscopios

Un estetoscopio está formado por una variedad de componentes importantes que le permiten transferir los sonidos internos del cuerpo de un paciente a los oídos de los profesionales médicos para que puedan diagnosticar y tratar la condición médica del paciente. Independientemente, del tipo de estetoscopio, todos se caracterizan por el mismo diseño básico y las piezas que lo componen. Utilizando como base la empresa más relevante en la fabricación de estetoscopios (Littmann [19]), en la Figura 2.6 se muestra los principales componentes de un estetoscopio, los cuales son:

- **Arco metálico (Headset):** es la parte superior del estetoscopio, la cual entra en contacto con la cabeza del médico especialista y está compuesta por:
  - **Olivas (Eartips):** es la parte que entra en contacto con el oído del médico, donde recibe los sonidos biomédicos auscultados. Las olivas están generalmente hechas de goma o silicona y están diseñadas para crear un sello de ajuste dentro de los oídos para que los sonidos no deseados del exterior puedan aislarse. Estas almohadillas deben ser cómodas para evitar que los médicos que pasan mucho tiempo auscultando a diversos pacientes sufran de un estrés añadido que perjudique el diagnóstico.
  - **Ojivas o auriculares (Eartube):** son las partes de metal/acero del estetoscopio que se conectan a las olivas y a la manguera. Las ojivas están diseñadas para aislar y transferir el sonido a los oídos del médico con una mínima pérdida en la calidad de los sonidos biomédicos auscultados. Estos tubos ayudan a separar los sonidos en los canales izquierdo y derecho para proporcionar una mejor experiencia sonora, lo que permite al usuario diagnosticar más fácilmente la condición médica de sus pacientes.
- **Manguera o tubo (Tubing):** son tubos flexibles generalmente fabricados con PVC o neopreno. El propósito de la manguera es transferir y retransmitir los sonidos que son capturados por el diafragma o la campana para después enviarlos a las ojivas. Las paredes del tubo están hechas para evitar que el ruido se mezcle con otros sonidos. Además, algunos modelos disponen de un tubo doble, que permite que los sonidos se propaguen mejor y lleguen más nítidos a los oídos del especialista.
- **Base o vástago (Stem):** es la parte del estetoscopio que conecta la manguera con el receptor auscultador. Además, de conectar los dos componentes del estetoscopio, en algunos modelos también permite al usuario cambiar entre el diafragma y la campana que componen al receptor auscultador.
- **Pieza del pecho o receptor auscultador (Chest-piece):** es la parte inferior del estetoscopio, la cual entra en contacto con la piel del paciente y es la encargada de captar los sonidos del interior del cuerpo. Teniendo en cuenta un diseño estándar, está compuesta por:

- **Diafragma o membrana (Diaphragm):** es el extremo circular de mayor diámetro de la pieza del pecho. El diafragma está diseñado para captar sonidos con frecuencias medias y altas en comparación con la campana, y se suele utilizar para examinar los intestinos y los pulmones.
- **Campana (Bell):** es el extremo circular de menor diámetro de la pieza del pecho. La campana está diseñada para captar sonidos con frecuencias más bajas, los cuales son difíciles de percibir con el diafragma, y se suele utilizar para examinar los sonidos cardíacos y vasculares.

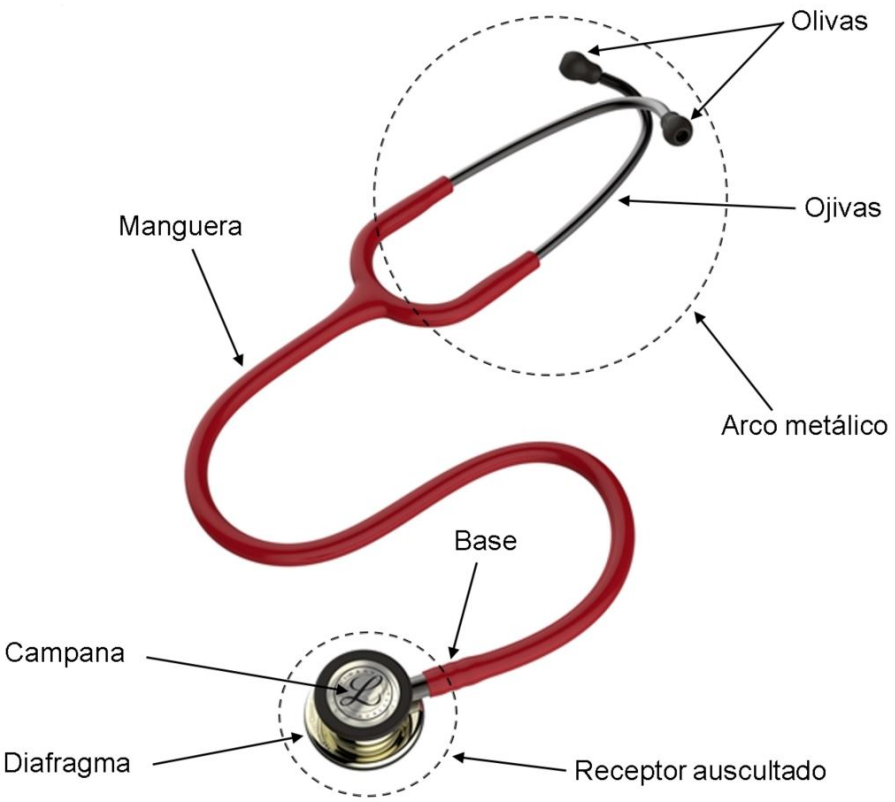

Figura 2.6: Partes del estetoscopio utilizando como modelo el estetoscopio Littman Classic III.

Tras presentar las partes que componen un estetoscopio, el resto de la sección se ocupa de describir los aspectos más relevantes sobre los tipos de estetoscopios. Actualmente, en el mercado se incluyen un amplio abanico de estetoscopios que, aunque siguen la estructura estándar descrita anteriormente, pueden ser clasificados en diversos tipos atendiendo a determinados factores. Por ejemplo, algunos estetoscopios se adaptan a ciertos tipos de pacientes o especialidades sanitarias, entre las que se encuentran las siguientes modalidades: fetal, neonatal, pediatría, enfermería, neumología, cardiología, veterinaria, etc. En otros casos se pueden clasificar en función de los componentes que se incluyen en la pieza del pecho (receptor auscultador): cabeza simple (diafragma o campana), cabeza doble (es el estándar y se compone del diafragma y la campana) y cabeza triple (es el menos utilizado y se compone de tres piezas que cubren un

mayor rango de frecuencias). Sin embargo, la clasificación que se muestra a continuación tiene un carácter más genérico, compuesta por dos tipos de estetoscopios: estetoscopios acústicos y estetoscopios digitales.

Estetoscopios acústicos tradicionales o convencionales:

Dentro de este tipo se incluyen todos aquellos estetoscopios que no están compuestos por un sistema digital. Son los tradicionales, los que han existido desde la invención del primer estetoscopio y los que presentan mayores limitaciones funcionales en la calidad del sonido biomédico auscultado. Un estetoscopio acústico funciona canalizando y dirigiendo la mayor cantidad de ondas sonoras de interés hacia los oídos. Para que podamos oír un sonido, las ondas sonoras deben causar vibraciones en las moléculas de aire, provocando cambios en la presión del aire que hacen que nuestros tímpanos vibren a su vez. Los sonidos internos del cuerpo como los latidos del corazón o el flujo de aire en las vías respiratorias causan ondas sonoras que golpean la pieza metálica (diafragma o campana) del pecho del estetoscopio cuando se coloca en un paciente. Posteriormente, la manguera de goma canaliza estas ondas sonoras en una dirección específica hasta que golpean las ojivas del estetoscopio y luego, finalmente, llegan a los oídos por las olivas. Debido a que las ondas sonoras son transmitidas por la manguera, llegan a los oídos con una mayor intensidad. Por ello al escuchar el corazón de un paciente con un estetoscopio suena más fuerte que colocando el oído justo al lado de su pecho. Las principales limitaciones de este tipo de estetoscopios giran en torno a la calidad del sonido capturado y a la limitada amplificación que sufren los sonidos. Incluso se ha demostrado que tanto la campana como el diafragma tiene una atenuación importante por encima de los 200 Hz, lo que limita la capacidad de discernir los sonidos en este rango de frecuencias [38]. Aunque la sensibilidad del oído humano oscila entre los 20 y los 20.000 Hz, el oído sigue una sensibilidad logarítmica a la frecuencia, por lo tanto, se requieren mayores cambios en las frecuencias más altas para discernirlas como diferentes [300]. Por ello, la limitación de los estetoscopios acústicos en la amplificación de los sonidos captados supone la principal desventaja. Por otro lado, y en comparación con los estetoscopios digitales, se puede destacar otra desventaja, la incapacidad de capturar la señal sonora en formato digital para su posterior análisis. Por último y de forma breve, las principales ventajas de este tipo de estetoscopios son: i) bajo precio en comparación con los digitales; y ii) no requieren baterías para funcionar.

Estetoscopios electrónicos o digitales:

El avance de la tecnología propició la aparición de los estetoscopios electrónicos o digitales. A diferencia de los estetoscopios acústicos, los estetoscopios electrónicos captan las vibraciones físicas del sonido, las traducen a una señal eléctrica y la optimizan para mejorar la calidad de audio de los sonidos auscultados y con ello el diagnóstico derivado. Los estetoscopios electrónicos utilizan una variedad de sensores, incluidos micrófonos de condensador y sensores piezoeléctricos, a fin de convertir las ondas acústicas en señales eléctricas para filtrarlas y pro-

cesarlas [209]. Entre las ventajas que soportan los estetoscopios electrónicos se pueden destacar [128]:

- Amplificación de sonidos de interés.
- Rango de frecuencia ajustable.
- Reducción del ruido ambiental.
- Grabación y reproducción.
- Conectividad con dispositivos externos (smartphone, tablet u ordenador) vía Wifi o Bluetooth.
- Algunos modelos incluyen una pantalla para visualizar el espectrograma de lo que se está capturando.

Por otro lado, las principales desventajas que pueden destacarse son:

- Mayor peso.
- Alimentación externa (baterías).
- Mayor coste que los estetoscopios acústicos.
- Los componentes electrónicos pueden dañarse si no se les presta atención.
- Posibilidad de interferencia de otros dispositivos electrónicos.

Destacar que la mayoría de estetoscopios electrónicos están diseñados como un estetoscopio tradicional, pero existen algunos modelos (como por ejemplo el estetoscopio digital Thinklabs One [33]) que se componen únicamente por un receptor auscultador conectado directamente a unos auriculares. Tras todo lo comentado anteriormente y centrándonos en los objetivos de esta Tesis, cabe destacar que gracias a este tipo de estetoscopios se hace posible el desarrollo de algoritmos y métodos que permitan ayudar en el diagnóstico de las patologías respiratorias, ya que transforman las ondas acústicas del interior del cuerpo en señales sonoras que pueden ser procesadas digitalmente.

2.2.4. Estetoscopios electrónicos comerciales

Como se ha mencionado anteriormente, los estetoscopios electrónicos, a diferencia de los estetoscopios acústicos, permiten capturar las señales sonoras del interior del cuerpo para su posterior procesado. Es por ello que los estetoscopios electrónicos toman especial atención para el desarrollo de algoritmos o métodos que ayuden a mejorar el diagnóstico respiratorio. El objetivo de esta sección es identificar los estetoscopios electrónicos más relevantes de la actualidad. Para ello se ha realizado un análisis exhaustivo sobre las empresas de mayor potencial y las más

aconsejadas por los especialistas en el campo de la auscultación. Entre las compañías de mayor prestigio destacan las siguientes: 3M Littmann [1], Thinklabs [33], Eko [10] y Ekuore [11]. A continuación, se describe para cada compañía el modelo de estetoscopio electrónico que ofrece mejores características y rendimiento.

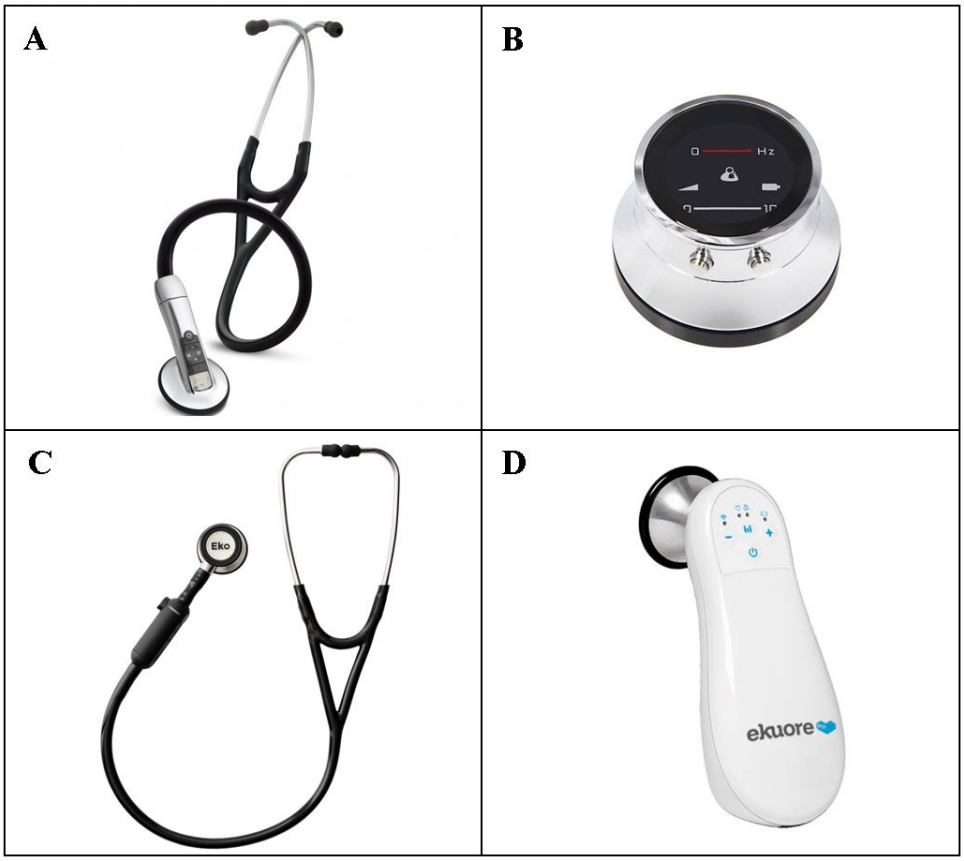

Figura 2.7: Estetoscopios electrónicos comerciales más relevantes: A) 3M Littmann Electronic Stethoscope Model 3200 [2]; B) Thinklabs ONE [33]; C) CORE Digital Stethoscope [6]; y D) Electronic stethoscope eKuore Pro [12].

3M Littmann [1]:

3M Littmann es considerada la compañía pionera en el campo de la auscultación, ofreciendo a lo largo de su trayectoria los estetoscopios de mayor potencial y relevancia, y actualmente, es la empresa con mayor trayectoria y la más valorada por los profesionales de la medicina. El último modelo desarrollado se denomina **3M Littmann Electronic Stethoscope Model 3200 [2]** (ver Figura 2.7A) y sus características principales son:

- Grabación y almacenamiento de hasta 12 pistas de 30 segundos de duración, para su posterior análisis o procesado.
- Eliminación del 85 % del ruido ambiente.

- Amplificación de los sonidos auscultados hasta un factor de x24.
- Conectividad con dispositivos externos vía Bluetooth. Se incluye un dispositivo que debe ser conectado a un PC con el sistema operativo Windows.
- Incluye modo campana o diafragma, para cambiar entre una u otra opción.
- Encendido automático al detectar la presión en el diafragma o la campana.
- Incluye un Software gratuito para la visualización de la señal auscultada en tiempo real.
- Catalogado como uno de los dispositivos más fiables y duraderos, debido a su composición.

Thinklabs [33]:

Sin embargo, en la actualidad otras compañías han conseguido posicionarse con propuestas innovadoras dentro del mercado de la auscultación electrónica. Thinklabs fue fundada en 1991 por Clive Smith, un graduado en Ingeniería Eléctrica de Caltech, apasionado de la electrónica médica, el sonido, la música y el procesado de señal. En honor a su lema "pensar profundamente en los problemas que importan y desarrollar soluciones imaginativas", sus novedosos diseños y la eficiencia de sus productos han convertido a Thinklabs en un rival digno ante la pionera 3M Littmann. Su estetoscopio electrónico, considerado el mejor del mercado, es conocido como **Thinklabs ONE [33]** (ver Figura 2.7B) y sus características principales son:

- Diseño innovador que rompe con la conocida estética de los estetoscopios tradicionales. Específicamente se compone solo del receptor auscultador (pieza del pecho), el cual se conecta directamente a unos auriculares.
- Considerado el estetoscopio de menor tamaño (cabe en la palma de la mano) y con mayor potencia del mercado.
- Amplificación de los sonidos auscultados por encima de un factor de x100. Es el estetoscopio electrónico del mercado que consigue la mayor amplificación de los sonidos.
- Grabación y almacenamiento de los sonidos para su posterior análisis y procesado utilizando la aplicación propia de Thinklabs.
- No ofrece conectividad inalámbrica incorporada, pero cuenta con un kit móvil que permite conectarlo a cualquier dispositivo móvil con iOS, Android o Windows.
- No dispone de un sistema de reducción de ruido ambiental. Esto ocasiona que se amplifiquen tanto los sonidos auscultados, como el ruido ambiental que rodea al paciente.

Eko [10]:

Por otro lado, merece la pena incluir a la compañía Eko por la diversidad de los productos ofrecidos. Eko destacó en el mercado gracias al diseño de un dispositivo, denominado **CORE Digital Attachment [5]**, que permite transformar un estetoscopio acústico en un estetoscopio electrónico, simplemente colocándolo entre el receptor auscultador y la manguera. Además, disponen de varios modelos de estetoscopio electrónicos. El más versátil y potente se denomina **CORE Digital Stethoscope [6]** (ver Figura 2.7C) y sus características principales son:

- Novedosa tecnología para la cancelación activa del ruido de fondo.
- Amplificación de los sonidos auscultados hasta un factor de x40.
- Conectividad con dispositivos externos vía Bluetooth. Compatible con cualquier dispositivo con iOS, Android o Windows.
- Incorpora inteligencia artificial aplicada a los sonidos cardíacos. Se ha convertido en el primer estetoscopio electrónico capaz de detectar soplos cardíacos (heart murmurs).
- Incluye un software que, entre otras funciones, permite mostrar el fonocardiograma en tiempo real.
- Grabación y almacenamiento de los sonidos auscultados para su posterior análisis y procesado, en pistas de hasta 120 segundos de duración.

Ekuore [11]:

Por último y sin duda la apuesta más arriesgada e innovadora viene de la mano de la compañía Ekuore. Su apuesta, supone una evolución significativa en el campo de la auscultación. Específicamente, han diseñado un dispositivo que lo definen como el primer estetoscopio electrónico inteligente (smart electronic stethoscope) que permite realizar un proceso de auscultación remota. Este dispositivo permite escuchar los sonidos auscultados a distancia, sin la necesidad de que el médico se encuentre en la misma sala que el paciente. Concretamente, el dispositivo transmite los sonidos auscultados a cualquier dispositivo (smartphone, tableta u ordenador) vía WiFi, para escucharlos y visualizarlos sin la necesidad de estar en contacto con el paciente. Este dispositivo se conoce como **Electronic stethoscope eKuore Pro [12]** (ver Figura 2.7D), y sus principales características se describen a continuación:

- Compuesto únicamente por un dispositivo encargado de capturar y transmitir los sonidos auscultados.
- Considerado el primer estetoscopio destinado a la telemedicina que da paso a la era de la auscultación remota. Mediante una APP desarrolla por eKuore el médico puede escuchar a distancia al paciente. Por ejemplo, supone una ventaja en el seguimiento o la monitorización continua de los pacientes crónicos, ya que el propio paciente puede utilizar el

dispositivo para auscultarse y después enviar los resultados al médico para su análisis, de esta forma se evitan las visitas innecesarias al hospital. Por otro lado, en el caso de examinar a pacientes con alguna enfermedad contagiosa (COVID-19), se puede evitar el contacto directo para disminuir las posibilidades de contagio.

- La pieza que entra en contacto con el paciente es desechable. Una característica que disminuye los posibles contagios.
- Incorpora filtros predefinidos para mejorar la señal cardiaca y pulmonar.
- Amplificación de los sonidos auscultados hasta un factor de x20.
- Conectividad con dispositivos externos vía WiFi. Además, permite conectar unos auriculares por cable o Bluetooth.
- Incluye una APP para controlar las diversas opciones de almacenamiento y grabación. Además, la APP ofrece otras posibilidades, como por ejemplo visualizar el fonocardiograma o incluso editar las señales sonoras grabadas.

Para finalizar, a modo de resumen, la Tabla 2.1 muestra una comparación entre los estetoscopios anteriormente descritos, en base a sus características más importantes.

Modelo estetoscopio	**3M Littmann Electronic Stethoscope Model 3200**	**Thinklabs ONE**	**CORE Digital Stethoscope**	**Electronic stethoscope eKuore Pro**
Peso (gramos)	98	50	87	150
Estructura	Similar al estetoscopio tradicional	Solo dispositivo receptor (chest-piece)	Similar al estetoscopio tradicional	Solo dispositivo receptor (chest-piece)
Conectividad	Bluetooth	kit móvil alámbrico	Bluetooth	WiFi
Duración Batería (horas)	50-60	5	10	6
Conexión Auriculares	-	Conector jack 3.5mm	-	Bluetooth
Duración máxima de grabación por pista (segundos)	30	Grabación ilimitada, solo disponible con APP (dispositivo externo)	120	120
Factor de amplificación	x 24	x 100	x 40	x 20
Atenuación ruido ambiental	✓	-	✓	✓
Software de visualización	✓	✓	✓	✓

Tabla 2.1: Caracterización y comparativa de los estetoscopios comerciales más relevantes del mercado. El símbolo ✓ indica que el estetoscopio incluye esa opción, mientras que el símbolo - indica lo contrario.

2.2.5. Sensores o micrófonos para la auscultación

Los estetoscopios electrónicos no son los únicos dispositivos que permiten transformar las ondas acústicas producidas en el interior del cuerpo en señales eléctricas sonoras. En las últimas dos décadas, muchos estudios, centrados en la mejora del diagnóstico, han optado por utilizar determinados sensores para capturar las señales sonoras del interior del cuerpo en lugar de los denominados estetoscopios digitales. El objetivo de esta sección es realizar un estudio sobre los sensores que se suelen utilizar como métodos de medición para digitalizar las señales sonoras biomédicas respiratorias. Específicamente, los sensores se analizan en el contexto de las recomendaciones especificadas para el Análisis del Sonido Respiratorio Computarizado, denotado en inglés como "Computerized Adventitious Respiratory Sounds Analysis (CORSA)", las cuales están basadas en un proyecto de la Sociedad Respiratoria Europea [289]. Las vibraciones de la pared torácica pueden registrarse mediante varios métodos diferentes de transducción, como el cambio de capacitancia y la piezoelectricidad [289]. Un sensor ideal debería ser pequeño, ligero, cómodo, barato, inalámbrico, capaz de producir mediciones fiables y tener una respuesta en frecuencia reproducible [254]. Los dos tipos principales de micrófonos utilizados para el análisis de los sonidos pulmonares son los micrófonos de condensador basados en el cambio de capacitancia, y los micrófonos de contacto basados en la piezoelectricidad. Los cuales se presentan a continuación:

- **Micrófonos de condensador:** son un tipo de micrófono acoplado al aire. Los micrófonos de condensador, como los micrófonos de electrodos, utilizan una transducción basada en el cambio de capacitancia para detectar los cambios en la presión acústica que modifican los valores de capacidad nominal [254, 189, 344]. Estos micrófonos tienen una respuesta en frecuencia casi plana en el rango de audio, lo que produce una distorsión mínima [254, 344]. Sin embargo, requieren un acoplamiento acústico a la pared del pecho con una cavidad de aire [247, 306]. Los micrófonos de sistemas microelectromecánicos (MEMS) se basan en los principios del condensador y proporcionan un rango de frecuencia y una SNR similares en comparación con los micrófonos de condensador convencionales, a la vez que proporcionan un factor de forma más pequeño [189, 213]. Debido a su gran ancho de banda, su alta sensibilidad, sus métodos de acoplamiento establecidos, su alta SNR (según las recomendaciones de CORSA) y su bajo coste, los micrófonos de condensador son utilizados ampliamente en este campo [344, 264, 103, 71, 262, 208, 254]. Entre los micrófonos de condensador más utilizados en el análisis de señales pulmonares, pueden ser destacados los modelos Shure Beta 98H/C y SM35 [200].

- **Micrófonos de contacto:** utilizan típicamente principios de transducción piezoeléctrica para crear un voltaje de salida proporcional al desplazamiento del sensor colocado directamente sobre la piel, sin el uso de una cámara de aire [306]. Estos sensores pueden caracterizarse por una sensibilidad extremadamente alta y tienen la ventaja de no captar tanto ruido ambiental como los micrófonos de condensador, pero son, por el contrario, muy sensibles a los artefactos de movimiento [344, 306]. En una revisión del estado del

arte se ha comprobado que, aunque los micrófonos de condensador son los más utilizados, también se pueden encontrar estudios que optan por el uso de este tipo de micrófonos para digitalizar los sonidos biomédicos respiratorios [110, 143]. Los principales micrófonos de contacto utilizados en el análisis de señales pulmonares son ECM-77B, ECM-T140, ECMT150 y ECM-KEC-2738 [240].

Además de la utilización individual de estos sensores, se han encontrado sistemas multi-sensor que utilizan un conjunto o un "array" de sensores para capturar los sonidos respiratorios. Las patologías pulmonares obstructivas producen una alteración en las vías respiratorias que transmiten el sonido y tienen efectos tanto espectrales como regionales (en función de la vía respiratoria afectada) que pueden ser analizados midiendo simultáneamente (array de sensores) en múltiples puntos de la superficie de la caja torácica [263]. Comprender la ubicación de los sonidos pulmonares anormales puede ayudar a identificar las áreas afectadas por la patología, así como a evaluar la gravedad de la patología en base a su distribución espacial. Para evaluar mejor la ubicación, se han desarrollado métodos de auscultación simultánea con múltiples sensores para "mapear" los sonidos en la superficie torácica [247, 52, 54, 69, 140]. Específicamente se pueden encontrar trabajados donde utilizan un array de micrófonos de condensador [177, 224, 41, 316] y un array de micrófonos de contacto [222, 133, 57]. Concretamente, en [52] se propone un sistema multi-sensor denominado "Vibration Response Imaging (VRI)" que ha sido objeto de estudio por su eficacia. El modelo VRI consiste en crear una representación en 2D de los sonidos de la respiración utilizando un conjunto de estetoscopios electrónicos que captan los sonidos del pecho mediante 18-40 sensores piezo-acústicos distribuidos en la espalda del sujeto para crear una imagen, en escala de grises y en tiempo real, que puede seguir dinámicamente la variación acústica a lo largo del ciclo respiratorio [52]. Por otro lado, en [177] intentaron formar una imagen acústica tridimensional (3D) de las probables ubicaciones de las fuentes sonoras utilizando múltiples sensores para saber cómo se propagaba el sonido lejos de esas fuentes.

2.2.6. Alternativas al proceso de auscultación para el diagnóstico de patologías respiratorias

Existe una amplia gama de procedimientos alternativos a la auscultación para ayudar a determinar el estado del sistema respiratorio humano y el diagnóstico de las posibles patologías presentes. Sin menospreciar el proceso de auscultación, ya que es la base de esta Tesis, se ha considerado interesante introducir las técnicas alternativas, al análisis de los sonidos biomédicos respiratorios, que existen para diagnosticar las patologías respiratorias. Se ha realizado una clasificación de los diferentes procedimientos alternativos que se pueden seguir para examinar al paciente. Los procedimientos pueden ser agrupados en tres grupos: métodos de laboratorio, pruebas de la función respiratoria y técnicas de imagen.

Métodos de laboratorio [127]:

Además de los análisis rutinarios de sangre y orina de laboratorio, se dispone de varias pruebas específicas para ayudar a determinar posibles patologías respiratorias específicas (por ejemplo, el asma puede ser detectado con el test que obtiene los niveles de inmunoglobulina E, denotados como lgE). En este grupo se incluyen:

- **Pruebas microbiológicas:** desempeñan un papel esencial en la investigación de las enfermedades respiratorias infecciosas causadas por virus, bacterias, hongos o parásitos.

- **Exámenes histológicos y citológicos:** desempeñan un papel fundamental en el diagnóstico de muchas enfermedades respiratorias malignas y benignas, incluidas las infecciones. Aparte del esputo expectante, que puede examinarse citológicamente, algunas muestras son adquiridas mediante diversas técnicas de biopsia, que se examinan más adelante y se envían para su evaluación histológica y/o citológica.

Pruebas de la función respiratoria [127]:

Este tipo de pruebas se centran en valorar el correcto funcionamiento del sistema respiratorio durante la mecánica de la respiración y el intercambio de gases que se lleva a cabo en los pulmones. En este grupo se incluyen:

- **Espirometría:** consiste en solicitar al paciente que, tras una inspiración máxima, expulse todo el aire de sus pulmones, durante el tiempo que necesite para ello, sobre un dispositivo específico. La espirometría permite evaluar la capacidad pulmonar del paciente y es una prueba clave para identificar la existencia de episodios de crisis asmáticos. Específicamente, estas pruebas analizan el volumen de aire exhalado y se utilizan para medir el efecto de los medicamentos broncodilatadores en la reversibilidad de la obstrucción, así como para determinar la capacidad de respuesta a las pruebas de provocación bronquial.

- **Capacidad pulmonar y resistencia de las vías respiratorias:** la capacidad pulmonar total puede determinarse mediante técnicas de dilución de gases o pletismografía corporal. Este último método también permite medir la resistencia de las vías respiratorias. La técnica de oscilación forzada, que mide la resistencia del sistema respiratorio total, tiene la ventaja de que el paciente no necesita realizar maniobras respiratorias específicas.

- **Capacidad de difusión:** la capacidad de difusión del monóxido de carbono en el pulmón (también conocido como factor de transferencia) suele realizarse con una prueba sobre un único ciclo respiratorio y mide la función general de intercambio de gases del pulmón.

- **Gasometría arterial:** es una de las pruebas de diagnóstico más útiles y consiste en medir la cantidad de oxígeno y de dióxido de carbono presente en la sangre. Este examen también determina la acidez (pH) de la sangre.

- **Pruebas de ejercicio cardiopulmonar (PECP):** evalúa el funcionamiento del corazón, circulación, respiración y metabolismo muscular en reposo y bajo creciente esfuerzo físico, hasta la máxima carga posible. La medición simultánea de concentraciones de oxígeno y dióxido de carbono en inspiración y espiración permite determinar cuánto oxígeno se inspira (VO2) y cuanto dióxido de carbono (VCO2) se espira.
- **Medición de la función de los músculos respiratorios:** se evalúa comúnmente midiendo las presiones máximas generadas en la boca durante la inspiración y espiración, mientras la vía aérea está ocluida.
- **Diagnóstico de los trastornos respiratorios del sueño:** se suele realizar mediante una polisomnografía. Consiste en el registro de la actividad cerebral, de la respiración, del ritmo cardíaco, de la actividad muscular y de los niveles de oxígeno en la sangre mientras se duerme.

Técnicas de imagen [127]:

Son las técnicas que se basan en el análisis de las imágenes que representan el estado del sistema respiratorio humano. En este grupo se incluyen:

- **Radiografía del tórax (Rayos X):** es una parte esencial para el diagnóstico y seguimiento de las patologías respiratorias. Es el primer paso en la evaluación radiológica de los pacientes con sospecha de enfermedades respiratorias. La radiografía digital moderna ofrece una alta calidad de imagen y la posibilidad de reducir la dosis de radiación.
- **Tomografía computarizada (TC) del tórax:** es la segunda modalidad radiológica más importante en medicina respiratoria, permitiendo una visualización mucho más detallada de las estructuras torácicas en comparación con la radiografía.
- **Angiografía pulmonar y bronquial:** son técnicas invasivas para la obtención de imágenes de los vasos sanguíneos y sólo se utilizan si las técnicas menos invasivas (TAC/resonancia magnética con contraste) fallan o necesitan ser confirmadas.
- **Fluoroscopia:** es una técnica de rayos X mediante la cual se visualiza directamente el movimiento respiratorio. Se utiliza principalmente para guiar la biopsia en lesiones pulmonares periféricas y para el diagnóstico diferencial de un diafragma elevado.
- **Resonancia magnética:** tiene la ventaja de que se evita la radiación. Es principalmente útil cuando se sospecha de una invasión tumoral del mediastino y la pared torácica.
- **Ultrasonografía:** se ha convertido en una importante técnica de imagen. Sus ventajas son la ausencia de radiación, el bajo coste y la movilidad. Se utiliza principalmente en la investigación de los derrames pleurales, pero también en el engrosamiento pleural, las anomalías de la pared torácica, para el diagnóstico del neumotórax y para las biopsias de las lesiones adyacentes a la pared torácica.

- **Técnicas de medicina nuclear:** se incluyen principalmente la Gammagrafía pulmonar de ventilación/perfusión para medir la respiración y la circulación en todas las áreas de los pulmones. Se utiliza principalmente en el diagnóstico de la embolia pulmonar.

2.3. Clasificación de los sonidos respiratorios

Se denominan sonidos respiratorios, al conjunto de sonidos generados por el flujo de aire que fluye a lo largo de los distintos conductos o vías que componen al sistema respiratorio humano, durante la mecánica de la respiración (proceso de inspiración y espiración). En esta sección se detallará la clasificación de los diferentes tipos de sonidos respiratorios que pueden ser producidos por el aparato respiratorio. Como se puede ver en la Figura 2.8, los diferentes tipos de sonidos respiratorios pueden ser agrupados en dos conjuntos diferenciados por la condición del paciente (paciente sano o paciente con una determinada patología respiratoria). Por un lado, los sonidos respiratorios normales hacen referencia a los sonidos producidos por el sistema respiratorio durante el ciclo respiratorio. Estos sonidos están siempre presentes durante la respiración del sujeto, independientemente de la presencia o ausencia de una determinada patología. Como veremos en la Sección 2.3.2, los sonidos respiratorios normales se pueden clasificar en función de dónde son generados dentro de la red de vías que componen el sistema respiratorio. Por otro lado, los sonidos adventicios o también denominados accidentales, aparecen en aquellos pacientes que presentan una patología respiratoria. Por lo tanto, la identificación de estos sonidos es una tarea de vital importancia para identificar anomalías en el correcto funcionamiento de la mecánica de la respiración, causadas por posibles trastornos respiratorios. Aunque se realiza una presentación de todos los sonidos adventicios, como se puede ver en la Figura 2.8, el objetivo de esta Tesis doctoral se centra en el estudio y el análisis de los sonidos sibilantes.

En 1816, cuando Laënnec dio origen al denominado proceso de auscultación y con ello se popularizo la escucha de los sonidos respiratorios, no existía una clasificación estándar de los diferentes tipos de sonidos respiratorios. Aunque en las últimas décadas se ha consolidado una clasificación más estandarizada, donde se diferencian claramente los diferentes sonidos respiratorios (como en la Figura 2.8), todavía surge la necesidad de definir una única clasificación que atienda a recoger todos los tipos de sonidos presentes durante la respiración del sujeto y sus correspondientes características. El principal problema radica en la caracterización tiempo-frecuencia de los distintos tipos de sonidos respiratorios. Por ejemplo, centrándonos en los sonidos sibilantes, American Thoracic Society (ATS) los define como un tono superior a 400 Hz cuya duración es superior a 250 ms [156]. Sin embargo, de acuerdo a las directrices establecidas por CORSA, las sibilancias se definen como un tono superior a 100 Hz cuya duración es superior a 100 ms [288]. Para solucionar esta problemática, en esta sección se mostrarán las características tiempo-frecuencia menos restrictivas. Por ejemplo, en el caso de los sonidos sibilantes, los cuales serán descritos en la Sección 2.3.3, el tono deberá ser superior a 100 Hz con una duración mínima de 100 ms.

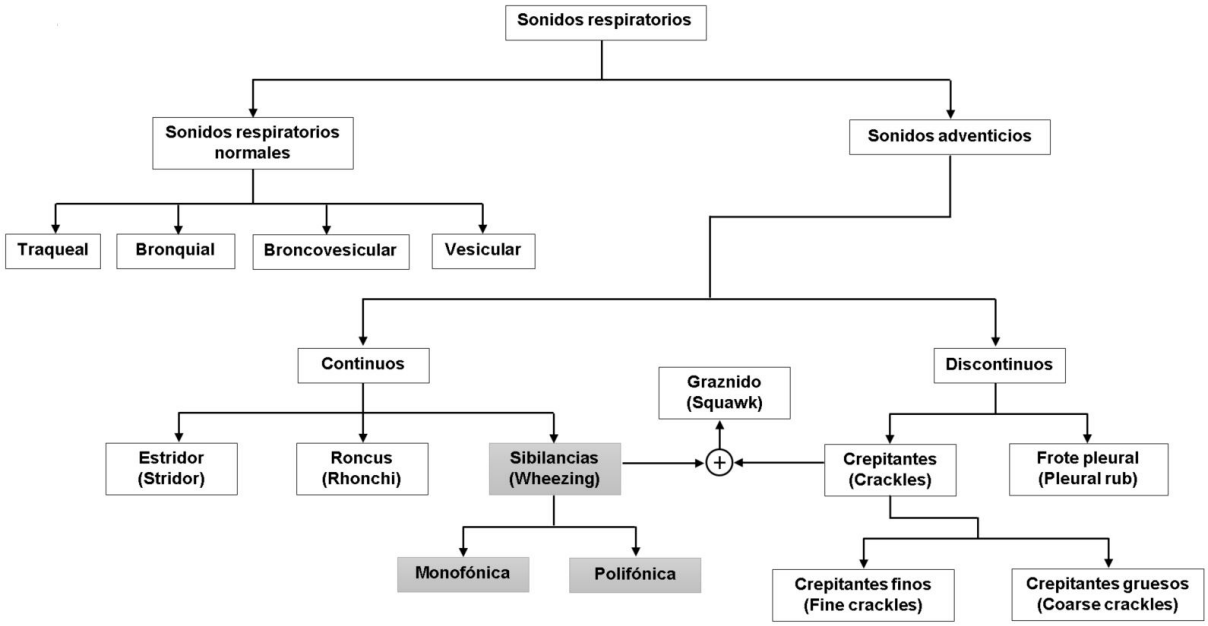

Figura 2.8: Clasificación de los sonidos respiratorios. Note que los sonidos adventicios denotados como "Squawk" son una combinación de sibilancias y crepitaciones. Los sonidos sibilantes han sido sombreados para destacar su relevancia en esta Tesis.

Aunque los sonidos cardíacos no son objeto de estudio en esta Tesis, es importante considerar la interferencia que pueden sufrir los sonidos respiratorios a causa de estos sonidos. Esta interferencia es debida a que las áreas de auscultación de los sonidos respiratorios y cardíacos se encuentran muy próximas, por ello se suele producir un solapamiento espectral de ambos tipos de sonidos. Específicamente, la frecuencia dominante de todos los sonidos respiratorios oscila entre 60 y 1.000 Hz [273], como veremos a lo largo de la actual sección. Sin embargo, la frecuencia dominante de los sonidos cardíacos suele estar por debajo de los 100 Hz [247, 263, 288, 131, 241]. La diferencia entre el rango de frecuencias hace que sea más fácil filtrar los sonidos cardíacos de los sonidos respiratorios. En general, se suele aplicar un filtro paso-alto por encima de 100 Hz para eliminar el ruido significativo del corazón, así como las interferencias eléctricas [247, 263, 290, 131, 241]. Aunque este procedimiento suele ser habitual para eliminar los sonidos cardíacos interferentes, en otros trabajos se proponen otros métodos más robustos que asumen que la frecuencia dominante de los sonidos cardíacos suele estar comprendida por debajo de los 320 Hz y por lo tanto los sonidos cardíacos y respiratorios sufren un solapamiento espectral mayor entre 60 y 320 Hz [62, 70, 87, 246, 193, 335].

Específicamente, en esta sección se detallará en primer lugar, la naturaleza de los sonidos respiratorios y cómo estos son generados, así como las características o propiedades del sonido necesarias para poder modelar y distinguir los distintos tipos de sonidos respiratorios. En segundo lugar, se presentarán los diferentes tipos de sonidos respiratorios normales. Por último, se detallarán los diferentes tipos de sonidos adventicios, haciendo hincapié en los sonidos sibi-

lantes, objeto de estudio en esta Tesis. Para afrontar esta tarea se utilizarán como base algunos de los trabajos de mayor relevancia en el campo de la clasificación de los sonidos respiratorios [288, 263, 290, 241, 247, 58, 104, 83, 145, 274, 42, 265, 84, 216].

2.3.1. Características del sonido respiratorio

Antes de comenzar a discutir los sonidos respiratorios, es conveniente elaborar la naturaleza del sonido en sí mismo y cómo este es generado. El flujo de aire a través de las vías respiratorias del árbol bronquial causa turbulencias, que originan las vibraciones que percibimos como ruidos o sonidos respiratorios. Las turbulencias se producen en las zonas donde la velocidad del aire es mayor y en aquellas con condiciones geométricas que dificultan un flujo laminar. Específicamente, el flujo turbulento es desorganizado y caótico por naturaleza, y se produce cuando la alta velocidad del flujo pasa a través de una vía respiratoria de gran diámetro, especialmente a través de una vía respiratoria con paredes irregulares. Esto ocurre principalmente en la tráquea y en las bifurcaciones de los bronquios primarios, lobulares y segmentarios. Sin embargo, solo el flujo de aire turbulento es responsable de la producción de los sonidos respiratorios [116]. A diferencia del flujo de aire turbulento, el flujo laminar se produce en situaciones de bajo flujo y es silencioso. Específicamente, el flujo laminar tiene forma parabólica, ya que el aire de las capas centrales se mueve más rápido que el de las capas periféricas, con poco o ningún flujo transversal. En las vías respiratorias más cercanas a los alvéolos, el flujo de aire es laminar y por lo tanto no hay turbulencias y tampoco se originan ruidos respiratorios. Para una explicación más minuciosa, Sarkar [274] ofrece una útil descripción de la producción de los sonidos respiratorios e incluye que el patrón del flujo laminar se puede modelar con la ecuación de Poiseuille.

Por otro lado, las variaciones o deformaciones sufridas durante la propagación de los sonidos respiratorios están vinculadas a varios factores [247]. En primer lugar, entra en juego la respuesta acústica del estetoscopio o los sensores utilizados para realizar el proceso de auscultación. En segundo lugar, la asimetría de los sonidos presentes en la auscultación que podrían indicar la presencia de alguna patología y se producen principalmente por la obstrucción de las vías respiratorias. Por último, la composición heterogénea del cuerpo puede actuar como un filtro ante estos sonidos. Específicamente, el tórax humano está compuesto por cuatro tipos diferentes de materiales con propiedades acústicas significativamente diferentes: tejido duro (hueso), tejido blando (músculo, grasa, etc.), aire en las principales vías respiratorias conductoras del árbol bronquial, y tejido parenquimatoso, que es una mezcla heterogénea de tejido blando y aire que se encuentra en los sacos alveolares y en los bronquiolos más pequeños. Las características de estos diferentes componentes afectan a la forma en la que los sonidos se transmiten a través del tórax durante la auscultación. Específicamente, el sonido en el interior de las vías respiratorias experimenta una absorción, dependiente de la frecuencia, en las paredes de las vías respiratorias y el tejido parenquimatoso circundante. Se ha demostrado que los sonidos de alta frecuencia se propagan más lejos dentro de la estructura de ramificación de las vías respiratorias, mientras que los sonidos de baja frecuencia tienden a acoplarse antes a las paredes

de las vías respiratorias [177, 266, 331]. Sin embargo, debido a la atenuación de los sonidos de alta frecuencia en el tejido parenquimatoso circundante, la mayor parte de la energía de la señal de los sonidos respiratorios registrados en la superficie del torso se concentra en las frecuencias más bajas [177, 332]. Por ello, el análisis de la transmisión del sonido en la cavidad torácica sugiere que el tórax, en general, actúa como un filtro paso bajo, absorbiendo frecuencias más altas a medida que el sonido viaja a través del árbol bronquial [266, 123, 79, 163, 204]. En consecuencia, dependiendo de los puntos de análisis donde se lleve a cabo la auscultación, el rango de frecuencias y la intensidad de los sonidos variará. Como se verá en el siguiente punto, los sonidos respiratorios traqueales y bronquiales presentan mayor frecuencia e intensidad que los sonidos respiratorios broncovesiculares y vesiculares.

La física del sonido toma un papel crucial para discriminar o identificar los diferentes tipos de sonidos respiratorios que pueden aparecer durante la auscultación. Además, las propiedades o atributos del sonido pueden describir lo que un ser humano experimenta al escuchar un determinado sonido. Específicamente, las características o atributos objetivos de un sonido son: frecuencia, intensidad y duración. Por otro lado, las principales propiedades subjetivas de un sonido son: tono, sonoridad y timbre. A continuación, se definen estas características de acuerdo a los sonidos respiratorios [145, 274] y se establece una relación entre las características objetivas (definen a los sonidos con valores objetivos) y las características subjetivas (las que describen cómo el sonido es escuchado por el médico).

Frecuencia y tono (Pitch) [145, 274]: por un lado la frecuencia es una característica objetiva que mide el número de oscilaciones o vibraciones por unidad de tiempo, en ciclos por segundo, y se expresa en hercios (Hz). Por lo tanto, a mayor frecuencia mayor número de vibraciones y viceversa. Como se ha mencionado anteriormente, en las vías respiratorias de mayor diámetro (tráquea y bronquios principales) se produce una mayor cantidad de vibraciones (debido a un mayor flujo de aire turbulento), que en las vías de menor diámetro (bronquios secundarios y bronquiolos). Por ello, como se verá en la siguiente sección los sonidos respiratorios traqueales y bronquiales se distribuyen en un rango de frecuencias mayor que los sonidos respiratorios broncovesiculares y vesiculares. Por otro lado, el tono es una característica subjetiva que representa como de “alto (agudo)” o “bajo (grave)” se escucha un sonido. Está directamente relacionado con la frecuencia, cuanto mayor sea la frecuencia, más alto se escuchará el sonido y viceversa. Es por ello que los sonidos respiratorios traqueales o bronquiales se escuchan más alto que el resto. Realmente esta característica es considerada una de las más importantes para distinguir entre todos los tipos de sonidos respiratorios, normales y adventicios. Ya que, por ejemplo, cada tipo de sonido adventicio tiene un rango de frecuencias característico. En el caso particular de las sibilancias, que se pueden caracterizan por un tono (Pitch) situado entre los 100 y los 1.000 Hz, resulta más sencillo comprender la relación entre frecuencia y tono. Notar que, aunque no todos los sonidos respiratorios se pueden modelar como un tono, es habitual el uso de esta métrica subjetiva para entender como de alto o bajo percibe el oído humano el sonido, dependiendo de la frecuencia. En concreto, el oído humano puede percibir las ondas sonoras en un amplio rango de frecuencias, que van desde los 20 a los 20.000 Hz.

Intensidad y sonoridad (Loudness) [145, 274]: la intensidad es una característica objetiva que está relacionada con la energía de las ondas sonoras y mide el flujo medio de energía por unidad de área perpendicular a la dirección de propagación. Por otro lado, la sonoridad es una característica subjetiva que permite ordenar los sonidos en una escala desde los más bajos a los más altos en intensidad, es decir, desde los sonidos que suenan más suave a los sonidos que suenan más fuerte. La sonoridad no sólo depende de la potencia de un sonido, sino que también depende de su duración y la estructura espectro-temporal del mismo. Realmente, la sonoridad es una característica más complicada de estandarizar, ya que está determinada por la fuente inicial que produce el sonido, la amplitud de las vibraciones, la distancia que recorren las vibraciones y el material por el que estas se desplazan posteriormente. Esto explica por qué algunos sonidos pulmonares se perciben con fuerza, como en un pulmón consolidado o con suavidad, como en un pulmón lleno de bulas enfisematosas. En términos generales, los sonidos con mayor intensidad se perciben con mayor fuerza que los sonidos de baja intensidad. Sin embargo, dos sonidos con la misma intensidad no sonaran con la misma fuerza si su frecuencia es distinta. En 1933, Fletcher y Munsen [113] determinaron que el oído humano tiene su máxima sensibilidad entre los 2.000 y los 5.000 Hz. Cuanto más alejada se encuentra la frecuencia del rango espectral anterior mayor intensidad será necesaria para escuchar los sonidos respiratorios.

Duración [145, 274]: la duración es una característica objetiva y se refiere al intervalo de tiempo en el que un determinado evento está activo en la señal de audio. Este parámetro está asociado a los términos onset y offset, que indica el instante de tiempo en el que el evento comienza y deja de estar activo respectivamente. La duración de las vibraciones determina si el oído del médico discierne el sonido como de larga o corta duración. Esta métrica permite por ejemplo distinguir las prolongadas sibilancias espiratorias de un paciente que sufre una EPOC. La duración también es considerada una de las características más relevantes para distinguir entre los diferentes tipos de sonidos adventicios que se pueden producir. Por ejemplo, los sonidos sibilantes se caracterizan por tener una duración mayor que los sonidos crepitantes.

Timbre [145, 274]: el timbre es una característica subjetiva que hace referencia a la calidad del sonido escuchado por el médico y depende principalmente de las componentes en frecuencia que caracterizan a un determinado sonido. Esta propiedad permite diferenciar dos sonidos con el mismo tono y la misma intensidad. Particularmente, los sonidos sibilantes suelen estar compuestos de varias componentes en frecuencia. La frecuencia fundamental o frecuencia primaria es la más baja de la onda sonora y determina el tono del sonido. Las frecuencias superiores a la fundamental se denominan frecuencias secundarias o armónicos. Atendiendo a la localización de las frecuencias secundarias, las sibilancias pueden ser clasificadas como monofónicas o polifónicas. Si estas frecuencias están relacionadas armónicamente entre sí (múltiplos enteros de la frecuencia fundamental) las sibilancias son monofónicas y en caso contrario son polifónicas. Por lo tanto, dependiendo de la localización de estas frecuencias secundarias (también denominadas parciales), el timbre variará permitiendo diferenciar un sonido con el mismo tono e intensidad. Por ello, se puede pensar en el timbre como las características "musicales" de los sonidos de la respiración. En general los sonidos respiratorios, se consideran mezclas muy complejas de sonidos de diferentes frecuencias, lo que les da un timbre característico que pue-

de ser modificado por diferentes condiciones patológicas. Por ejemplo, esto permite al médico distinguir entre los diferentes sonidos producidos en el tórax, como los sonidos de la respiración vesicular generados en el tejido pulmonar normal, y los sonidos de la respiración bronquial transmitidos a través de un pulmón consolidado.

En el procesado de los sonidos biomédicos respiratorios, existen varios métodos de análisis para la extracción de las características tiempo-frecuencia de los sonidos respiratorios grabados durante la auscultación. Estas técnicas de análisis son fundamentales para la clasificación y detección de los diferentes sonidos adventicios (sibilancias, crepitaciones, etc.). Específicamente, el espectrograma en magnitud de los sonidos respiratorios puede ser obtenido utilizando la magnitud de una variante de la transformada de Fourier, denominada "Short-Time Fourier Transform (STFT)", aplicando una determinada función de ventana (Hamming, Hann, etc). Este tipo de representación tiempo-frecuencia ha sido ampliamente utilizado en el campo de los sonidos biomédicos respiratorios [263, 317, 267, 268, 62] ya que permite extraer el tiempo, la frecuencia y la intensidad de los diferentes sonidos de entrada. Destacar que este tipo de representación ha sido utilizado a lo largo de la actual sección para mostrar los diferentes tipos de sonidos respiratorios. Específicamente, los parámetros de la STFT, así como la frecuencia de muestreo, han sido ajustados en cada representación para mejorar la visualización de cada tipo de sonido. Por otro lado, destacar que todas las contribuciones presentadas durante el transcurso de esta Tesis utilizan este tipo de representación tiempo-frecuencia para extraer las características de los sonidos sibilantes y respiratorios normales. A modo de ejemplo, la Figura 2.9 muestra el espectrograma obtenido a partir de una señal de entrada compuesta por sonidos respiratorios normales y varias sibilancias. Para obtener el espectrograma representado en esta figura, se ha utilizado una frecuencia de muestreo de 2.048 Hz, una ventana Hamming de 256 muestras de longitud con un solapamiento del 25 % (resolución temporal igual a 31.3 ms) y el orden de la transformada discreta de Fourier aplicada, denotada como Discrete Fourier Transform (DFT), se establece en 512 bins de frecuencia (el doble del tamaño de la ventana). Las siguientes observaciones pueden ser destacadas: (i) los sonidos sibilantes se encuentran marcados con rectángulos rojos. Las sibilancias mostradas tienen un comportamiento tonal, cuya frecuencia fundamental (tono) oscila en torno a los 320 Hz. Sin embargo, los sonidos respiratorios normales presentan un espectro en banda ancha con una variación suave de la energía a lo largo del rango de frecuencias; (ii) la intensidad de los sonidos sibilantes mostrados en el ejemplo es mayor que la de los sonidos respiratorios normales. Sin embargo, la distribución tiempo-frecuencia de los sonidos respiratorios normales produce una sonoridad similar entre ambos sonidos, lo que dificulta la escucha de las sibilancias; y (iii) la duración de los sonidos sibilantes mostrados oscila en torno a 1 segundo. Esta continuidad en el tiempo permite caracterizar a los sonidos sibilantes como trayectorias espectrales.

Además de las características del sonido comentadas anteriormente, el desarrollo de un estudio experimental y empírico de los diferentes tipos de sonidos respiratorios ha propiciado la definición de una serie de propiedades más específicas cuyo objetivo es facilitar la distinción entre los diferentes sonidos presentes:

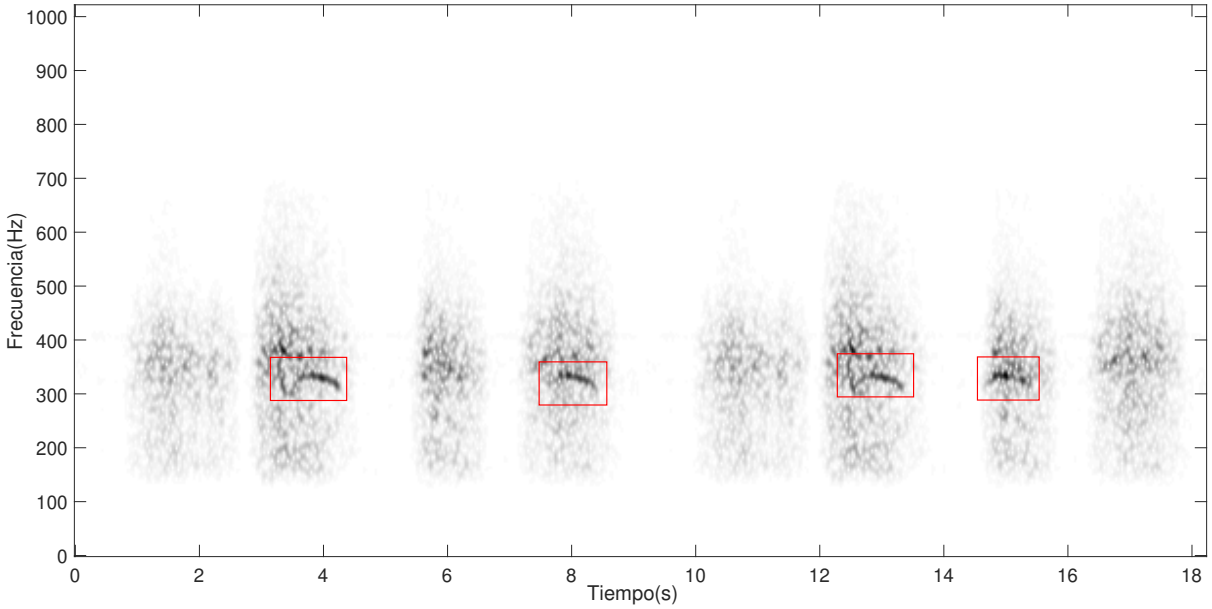

Figura 2.9: Representación tiempo-frecuencia (espectrograma) de una señal de audio respiratoria en la que se pueden observar cuatro sibilancias (rectángulos rojos), mezcladas con sonidos respiratorios normales. Las energías más altas se indican con un color más oscuro.

Continuidad/Discontinuidad temporal: esta propiedad está directamente relacionada con la duración de un determinado evento. La continuidad temporal permite modelar aquellos sonidos cuya energía varia lentamente en el tiempo (este concepto se denomina también suavidad temporal), como en el caso de los sonidos adventicios continuos (por ejemplo, las sibilancias). En la Figura 2.9 se puede observar como las sibilancias son continuas en el tiempo. Por otro lado, la discontinuidad temporal hace referencia a aquellos sonidos que ocurren de forma intermitente o cuya duración temporal es inmediata (este concepto se denomina también dispersión temporal). Esta propiedad modela el comportamiento de los sonidos adventicios discontinuos (como por ejemplo los crepitantes). Por lo tanto, conseguir diferenciar ambas propiedades permite modelar los sonidos sibilantes de manera diferente a los crepitantes.

Suavidad/Dispersión espectral: esta propiedad está directamente relacionada con la distribución de la energía a lo largo del rango espectral (rango de la frecuencia). La suavidad espectral permite caracterizar aquellos sonidos que son suaves en frecuencia, es decir, cuya energía varía lentamente en el rango espectral. Este comportamiento ocurre de forma general en los sonidos respiratorios normales, los cuales pueden ser modelados como un espectro en banda ancha, como se puede ver en la Figura 2.9. Por otro lado, la dispersión espectral permite caracterizar aquellos sonidos que son dispersos en frecuencia, es decir, los que pueden ser modelados por picos espectrales en banda estrecha, como ocurre en el caso de los sonidos sibilantes en particular. Por lo tanto, conseguir diferenciar ambas propiedades permite distinguir entre los sonidos respiratorios normales y los sonidos sibilantes.

Repetitividad temporal: esta propiedad permite diferenciar aquellos eventos que se repiten a lo largo del tiempo. Por ejemplo, los sonidos respiratorios normales se pueden consideran

patrones repetitivos en el tiempo. Durante la mecánica de la respiración, los procesos de inspiración y espiración son producidos a lo largo del tiempo. Sin embargo, los sonidos adventicios en general y los sonidos sibilantes en particular no se pueden modelar por esta propiedad. Específicamente, las sibilancias pueden estar presentes en ciertas etapas de la señal respiratoria y ausentes en otras debido a la naturaleza impredecible del desorden pulmonar. En la Figura 2.9 se puede ver el comportamiento de los sonidos respiratorios normales y los sonidos sibilantes considerando esta propiedad.

2.3.2. Sonidos respiratorios normales

Como se ha mencionado anteriormente, los ruidos o sonidos respiratorios se producen como resultado del aire que fluye por los pulmones y se clasifican como normales o anormales (adventicios). Los sonidos respiratorios normales se definen como aquellos que se producen en las vías respiratorias sanas por una respiración fisiológica no forzada. En términos generales, los sonidos respiratorios normales pueden extenderse en frecuencia hasta los 5.000 Hz [123], sin embargo, la mayor parte de la energía está comprendida entre los 60 y 1.000 Hz [273]. En concreto, su comportamiento en frecuencia se puede caracterizar como un espectro en banda ancha, donde el cambio de la energía es suave. Los sonidos respiratorios normales se pueden dividir en ciclos respiratorios, y cada ciclo respiratorio está formado por la etapa de inspiración y espiración. La duración de las etapas varía en función de las vías respiratorias auscultadas. Además, la intensidad y el rango de frecuencias de los sonidos respiratorios normales varía a lo largo de la auscultación del árbol bronquial. Como se puede ver en la Figura 2.10, los sonidos de mayor intensidad y mayor rango de frecuencia se producen en la tráquea. Sin embargo, estos parámetros van disminuyendo a lo largo del árbol bronquial, hasta que finalmente en niveles muy próximos a los alvéolos existe un flujo laminar que no genera sonido.

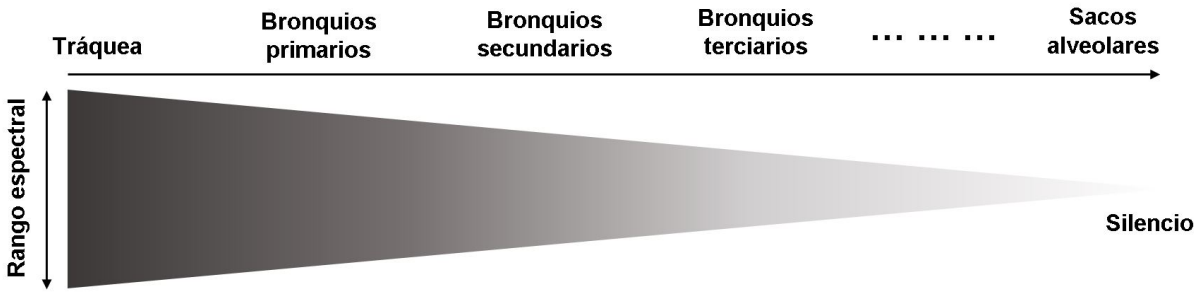

Figura 2.10: Variación de la intensidad y del rango de frecuencias del sonido respiratorio producido a lo largo de las vías respiratorias del árbol bronquial. La anchura de la barra indica cómo varía el rango espectral y el color de su interior cómo varía la intensidad. A mayor anchura mayor rango espectral y un color oscuro indica mayor intensidad.

Los sonidos respiratorios normales suelen clasificarse en cuatro tipos de acuerdo al área de auscultación [274, 255, 104, 145, 83, 241, 84, 58, 247]: traqueales, bronquiales, broncovesiculares y vesiculares. Como se muestra a continuación las características de frecuencia, intensidad

y duración de cada etapa (ratio de Inspiración-Espiración) varían a lo largo de los cuatro tipos de sonidos respiratorios normales. A modo de presentación, la Figura 2.11 muestra una clasificación visual de los distintos tipos de sonidos respiratorios normales, considerando las áreas de auscultación, así como la intensidad y la duración de cada etapa respiratoria.

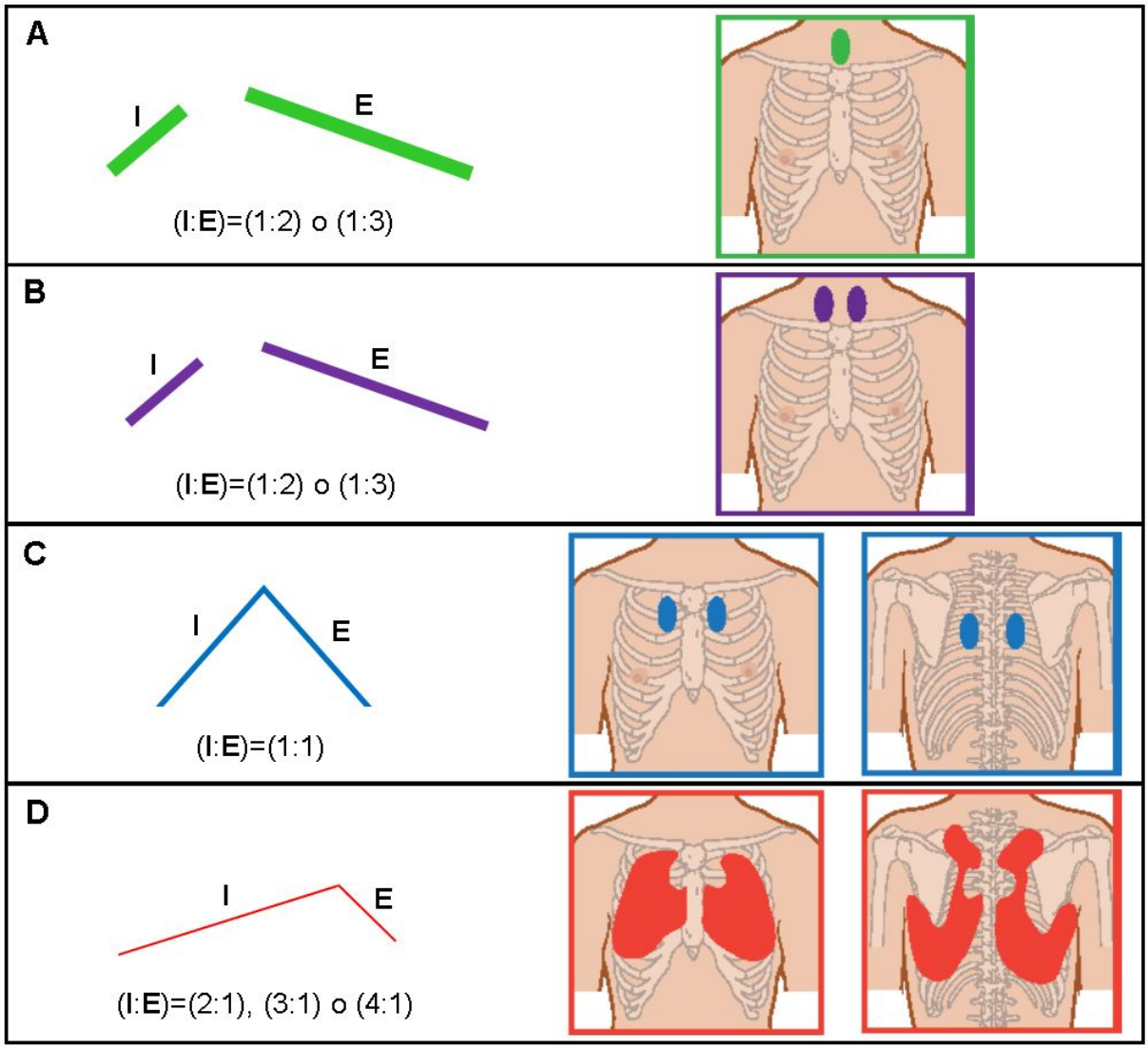

Figura 2.11: Áreas de auscultación y variación de las características (intensidad y duración de cada etapa) para los cuatro tipos de sonidos respiratorios normales: A) Sonido respiratorio traqueal; B) Sonido respiratorio bronquial; C) Sonido respiratorio broncovesicular; y D) Sonido respiratorio vesicular o pulmonar. En cada subfigura, la representación derecha corresponde a las áreas de auscultación y la representación izquierda a las características del ciclo respiratorio. El grosor de la línea indica la intensidad del sonido y la longitud indica la duración de cada etapa. La notación (I:E) hace referencia al ratio de Inspiración-Espiración. La discontinuidad en las líneas de las subfiguras A y B indican una pausa clara entre ambas fases.

El sonido respiratorio traqueal es un sonido áspero, muy fuerte y agudo que se escucha sobre la tráquea (por encima de la muesca subclavicular). La fase espiratoria es de mayor duración que la fase inspiratoria. Específicamente, el ratio de Inspiración-Espiración (I:E) suele ser (1:2) o (1:3). Además, existe una pausa entre ambas etapas. La frecuencia típica del sonido de la respiración traqueal varía de 100 a 1.500 Hz, con una caída brusca de la potencia por encima de una frecuencia de corte de aproximadamente 800 Hz [123]. El rango de frecuencia del sonido traqueal es mucho más amplio que el del sonido vesicular o pulmonar, con frecuencias que pueden llegar a los 5.000 Hz [123]. La auscultación sobre la tráquea no se realiza de forma rutinaria, pero puede ser útil en ciertas condiciones específicas. En primer lugar, tiene una calidad o timbre tubular y hueco, por lo que es un buen modelo para estudiar el sonido de la respiración bronquial. En segundo lugar, algunos sonidos adventicios, como el estridor, se encuentran mezclados con la respiración traqueal normal [337]. Por último, el análisis del sonido traqueal también es útil para el seguimiento de los pacientes con el síndrome de apneas-hipopneas del sueño (SAH) [228].

El sonido respiratorio bronquial es fuerte y agudo, similar al sonido traqueal, pero con ligeras variaciones. Estos sonidos se escuchan sobre el manubrio (espacio en el que la tráquea se divide en los bronquios primarios). Al igual que el sonido traqueal existe una pausa entre la fase de inspiración y espiración, y el ratio (I:E) suele ser (1:2) o (1:3). Son más agudos y más fuertes que los sonidos respiratorios que se oyen sobre otras partes de los pulmones (broncovesiculares o vesiculares), pero más silenciosos y de sonido más hueco (tubular) en comparación con los sonidos respiratorios traqueales. Es común agrupar a los sonidos traqueales con los sonidos bronquiales [58, 247], ya que tienen características muy similares al estar generados por las vías respiratorias de mayor diámetro o caudal. Por ello, el sonido respiratorio bronquial, al igual que el sonido respiratorio traqueal, suele producirse entre los 100 y los 1.500 Hz. Sin embargo, la intensidad de estos sonidos sufre una caída más notable cuando la frecuencia es superior a los 800 Hz. Diferentes desordenes pulmonares como, neumonía, tumores pulmonares, atelectasia (colapso de parte de un pulmón) o neumotórax, pueden provocar que los sonidos respiratorios bronquiales puedan ser escuchados en otras regiones de los pulmones, siendo un claro indicador de las patologías anteriormente mencionadas. Estos sonidos son denominados sonidos respiratorios bronquiales anormales, cuando son escuchados fuera de su área de auscultación.

El sonido respiratorio broncovesicular es un sonido con una sonoridad, intensidad y tono intermedio. Se pueden escuchar sobre el primer y segundo espacio intercostal junto al esternón y entre las escápulas. La fase inspiratoria y espiratoria de estos sonidos tiene la misma longitud, siguiendo un ratio (I:E) de (1:1). En [145] se describe que este sonido es similar al de soplar a través de una pajita. Los sonidos broncovesiculares son intermedios entre los bronquiales y los vesiculares, y son generados principalmente por los bronquios secundarios en ambos pulmones. Por ello el rango de frecuencias y la intensidad de estos sonidos presentan valores intermedios entre los sonidos bronquiales y vesiculares (pulmonares).

El sonido respiratorio vesicular (pulmonar) es un sonido suave de tono bajo. Se escucha en la mayor parte de la periferia del pulmón, donde se encuentra la red compuesta de bronquiolos cada vez más estrechos. La fase inspiratorio es de mayor duración que la fase espiratorio,

siguiendo un ratio (I:E) de (2:1), (3:1) o (4:1). Al igual que los sonidos broncovesiculares, no existe pausa entre inspiración y espiración. Específicamente, los sonidos respiratorios pulmonares se transmiten a través del tejido pulmonar y la pared del pecho. Estos sonidos son más silenciosos que los traqueales y bronquiales. La inspiración es más sonora que la espiración, ya que se desvanece rápidamente en el tiempo. Esto se debe a que el flujo de aire turbulento durante la espiración se distribuye rápidamente hacia las vías respiratorias de mayor diámetro (tráquea y bronquios primarios). En el análisis del sonido, el rango de frecuencia de los sonidos pulmonares normales parece ser más estrecho que el de los sonidos traqueales, extendiéndose desde menos de 100 Hz hasta 1.000 Hz [58]. Sin embargo, la mayor parte de la energía suele estar distribuida en un rango espectral comprendido entre los 200 y 600 Hz [83].

En términos generales, suele considerarse que los sonidos respiratorios normales contienen la mayor parte de la energía en el rango espectral comprendido entre 60-1.000 Hz [273]. La Figura 2.12 muestra un ejemplo del espectrograma obtenido para los diferentes sonidos respiratorios normales descritos anteriormente. Todas las señales de audio respiratorias mostradas en esta figura han sido obtenidas del repositorio adjunto al libro [83]. La pausa clara entre la inspiración y la espiración, en el caso de los sonidos respiratorios traqueales y bronquiales, se debe a la ausencia de la fase alveolar en las vías respiratorios de mayor diámetro (tráquea y bronquios primarios) [274].

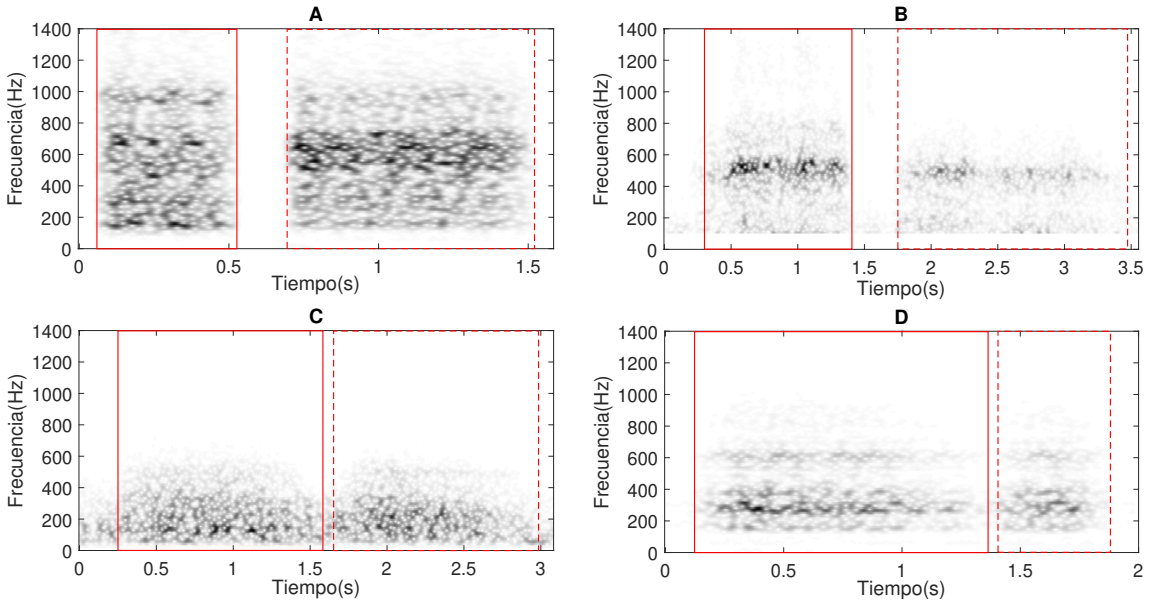

Figura 2.12: Representación tiempo-frecuencia (espectrograma) de un ciclo respiratorio completo (Inspiración y Espiración) para los cuatro tipos de sonidos respiratorios normales: A) Sonido respiratorio traqueal; B) Sonido respiratorio bronquial; C) Sonido respiratorio broncovesicular; y D) Sonido respiratorio vesicular o pulmonar. Los rectángulos continuos indican la fase de inspiración y los discontinuos la fase de espiración.

Para finalizar es preciso indicar que en algunos casos también se suelen incluir los sonidos respiratorios normales que se escuchan en la boca [274, 83]. Sin embargo, en la mayoría

de trabajos estos sonidos no se clasifican como sonidos respiratorios normales [255, 104, 145, 241, 84, 58, 247], ya que no son examinados durante el proceso de auscultación realizado para evaluar el estado del sistema respiratorio, o simplemente porque no se consideran sonidos auscultados. Los sonidos respiratorios que se escuchan en la boca contienen una frecuencia distribuida entre 200 y 2.000 Hz como el ruido blanco normal [118]. Su intensidad es tan fuerte y áspera como la de los ruidos respiratorios de la tráquea y los bronquios principales, y presentan un tono moderadamente alto [274]. En una persona sana, la respiración es silenciosa en la boca, pero es fácilmente audible incluso a distancia en pacientes con bronquitis crónica y asma. Sin embargo, este signo se utiliza con menos frecuencia en la actualidad. Una razón puede ser que el estridor y las sibilancias se suelen confundir a menudo con una respiración ruidosa [118], pero el simple método de escuchar una respiración ruidosa en la boca sin la necesidad de un dispositivo de auscultación puede ser un signo clínico importante. Generalmente el ruido respiratorio que se escucha en la boca se debe al aumento de las turbulencias causadas por las irregularidades de la superficie de las vías respiratorias, los cambios abruptos en la dirección del flujo, o el estrechamiento de las vías respiratorias que generan un flujo más rápido [118].

2.3.3. Sonidos respiratorios adventicios

Los sonidos respiratorios adventicios o accidentales se definen como aquellos sonidos respiratorios adicionales que se superponen a los sonidos respiratorios normales [287, 288]. La presencia de estos sonidos durante la mecánica de la respiración (inspiración y espiración) suele indicar la presencia de un trastorno pulmonar. Diversas patologías y lesiones pulmonares provocan alteraciones en las vías respiratorias que transmiten el sonido y dan lugar a sonidos respiratorios adventicios que, si se analizan adecuadamente, pueden proporcionar información adicional sobre la gravedad y la ubicación de la enfermedad. En esta sección se examinan los diferentes tipos de sonidos respiratorios adventicios y sus características espectrales, haciendo especial hincapié en los sonidos sibilantes, objeto de estudio en esta Tesis. La clasificación de los sonidos adventicios ha sido elaborada a partir de las definiciones y características espectrales establecidas por “European Respiratory Society (ERS)” [289]. Los sonidos adventicios pueden ser clasificados en dos grupos: sonidos adventicios continuos (como las sibilancias) y sonidos adventicios discontinuos (como los crepitantes) [274, 42, 224, 145, 286, 117, 197]. Por otro lado, existen sonidos adventicios que comparten características comunes entre los sonidos adventicios continuos y discontinuos, y se denominan graznidos (squawks).

Sonidos adventicios continuos (musicales):

Los sonidos adventicios continuos, denotados en inglés como Continuous Adventitious Sounds (CAS), son definidos principalmente por su continuidad temporal, con una duración superior a los 100 ms. Además, estos sonidos se suelen definir como sonidos “musicales”, ya que describen trayectorias espectrales similares a las de las notas de un instrumento (este comportamiento puede ser visto en el espectrograma correspondiente a estos sonidos). Dependiendo

de la localización en el rango espectral, la duración y el tono (pitch) de estas trayectorias espectrales, los sonidos adventicios continuos pueden ser clasificados como: sibilancia (wheezing), roncus (rhonchi) y estridor (stridor).

Las sibilancias (wheezing) son producidas por la obstrucción localizada de las vías respiratorias, causada por un cuerpo extraño, un tapón mucoso o un posible tumor [274]. Las sibilancias son un hallazgo inespecífico e incluso pueden detectarse en una persona sana hacia el final de la espiración, cuando esta espiración es forzada. Sin embargo, se ha demostrado que las sibilancias patológicas, las que son consecuencia de una patología respiratoria, pueden producirse incluso cuando la mecánica de la respiración es suave [58]. Diferentes estudios han intentado profundizar en el mecanismo de producción de las sibilancias. Inicialmente, Forgacs, en 1967, propuso que las sibilancias son generadas por las oscilaciones de las paredes bronquiales iniciadas por el flujo de aire que las atraviesa, y el tono de las sibilancias depende de las propiedades mecánicas de las paredes bronquiales [115]. En su estudio, Forgacs define a las sibilancias como un sonido musical y las compara con el sonido producido por una trompeta de juguete, cuyo sonido es producido por la vibración de una lengüeta. Posteriormente, Grotberg y Gavriely, propusieron un modelo matemático basado en la dinámica de fluidos para intentar explicar la mecánica de la producción de las sibilancias [132]. Las oscilaciones que generan el sonido sibilante comienzan cuando la velocidad del flujo de aire alcanza un valor crítico. Este modelo muestra que las sibilancias siempre vienen acompañadas por una limitación del flujo de aire, pero la limitación del flujo de aire no necesariamente está acompañada por sibilancias.

En términos generales, las sibilancias se definen como sonidos adventicios continuos generados por la obstrucción de la tráquea o los bronquios que componen el árbol bronquial. Pueden localizarse durante la inspiración, la espiración o en ambas fases del ciclo respiratorio, como se muestra en la Figura 2.13. Las sibilancias siempre aparecen solapadas con los sonidos respiratorios normales, ya que ambos sonidos son producidos por el mismo flujo de aire que recorre el árbol bronquial. Por lo que pueden aparecer junto con uno de los cuatro tipos de sonidos respiratorios normales (traqueal, bronquial, broncovesicular y vesicular o pulmonar). Además, las sibilancias se caracterizan por tener un tono (pitch) alto, lo que hace que en ciertas ocasiones sean más sonoras que los sonidos respiratorios normales. Esto ocasiona que a menudo las sibilancias sean audibles en la boca del paciente, sin la necesidad de realizar el proceso de auscultación, cuando el sujeto realiza la mecánica de respiración, inspirando y espirando el flujo de aire a través de la boca [287]. Como se ha mencionado anteriormente, no existe una caracterización tiempo-frecuencia única para los sonidos sibilantes. Por un lado, ATS define estos sonidos como un tono superior a 400 Hz cuya duración es mayor de 250 ms [156]. Sin embargo, de acuerdo a las directrices establecidas por CORSA, las sibilancias se definen como un tono superior a 100 Hz cuya duración es mayor de 100 ms [288]. En esta Tesis se ha optado por conservar las características tiempo-frecuencia menos restrictivas. En este sentido, los sonidos sibilantes se caracterizan por tener una frecuencia fundamental (pitch) comprendida entre los 100 y los 1.000 Hz con una duración superior a los 100 ms [289]. Por otro lado, la forma de onda de un sonido sibilante en el dominio del tiempo se asemeja a la de un sonido sinusoidal.

Por lo tanto, los sonidos sibilantes se caracterizan por mostrar trayectorias espectrales de banda estrecha (picos espectrales) [119] como se puede observar en la Figura 2.13.

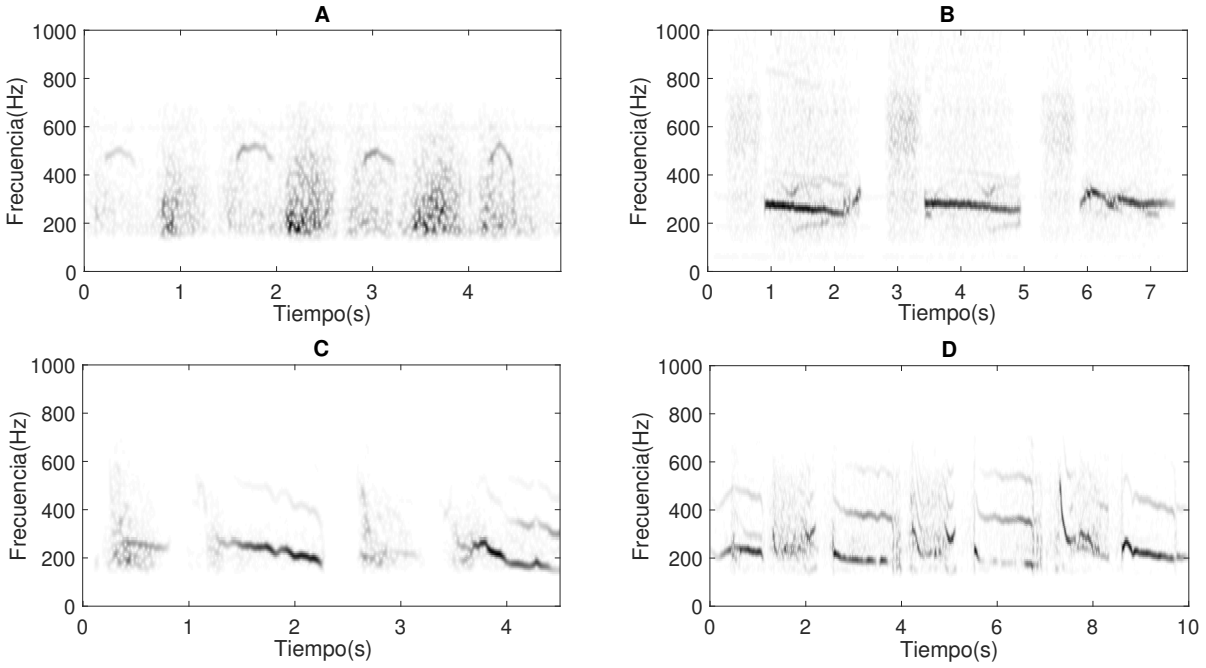

Figura 2.13: Representación tiempo-frecuencia (espectrograma) de varias señales respiratorias con sibilancias (wheezing) presentes durante la mecánica de la respiración (inspiración y espiración): A) Sibilancias durante la inspiración; B) Sibilancias durante la espiración; C) y D) sibilancias durante la inspiración y la espiración.

La sibilancia es probablemente el término acústico más utilizado en el campo de la medicina del sistema respiratorio humano [247]. Esto se debe a que en cientos de publicaciones cada año, las sibilancias son referenciadas como un indicador de la obstrucción de las vías respiratorias, como un parámetro para medir la gravedad del asma, o como un clasificador para las encuestas epidemiológicas, por nombrar algunos ejemplos [247]. La identificación de estos sonidos durante el ciclo respiratorio normal, es de gran importancia en el diagnóstico de las patologías obstructivas de las vías respiratorias [296]. De hecho, Sovijarvi [287] indica que las sibilancias pueden mostrar características acústicas sintomáticas, no solo de la presencia de anomalías en el sistema respiratorio, sino también de la gravedad y la ubicación de la obstrucción de las vías respiratorias en algunas de las patologías respiratorias obstructivas más frecuentes. Aunque las sibilancias suelen estar asociadas a los episodios de asma, numerosos estudios demuestran que dichos sonidos son también indicadores de otras enfermedades como EPOC, bronquiolitis o bronquiectasia entre las más comunes [58, 320, 247, 295, 109, 39, 219], y que el nivel de gravedad de estas enfermedades podría estar relacionado con la duración, número, frecuencia fundamental y localización temporal de las sibilancias dentro del intervalo de inspiración o espiración en un ciclo respiratorio. Concretamente, las sibilancias se suelen definir como un signo clínico común en pacientes con enfermedades obstructivas de las vías respiratorias, y en particular durante episodios agudos de asma.

Las sibilancias pueden ser clasificadas, como **sibilancias monofónicas (monophonic wheezing)** o **sibilancias polifónicas (polyphonic wheezing)**, de acuerdo a la estructura armónica que existe entre las diferentes trayectorias espectrales o componentes en frecuencia que las componen [227, 156, 245, 59, 305, 153, 160, 294]. Por un lado, las sibilancias monofónicas están compuestas por un único pico espectral de banda estrecha (frecuencia fundamental) o por la frecuencia fundamental junto a sus armónicos. Por lo tanto, las sibilancias monofónicas se caracterizan por definir, a lo largo del tiempo, una única trayectoria espectral (ver Figura 2.14A) o varias trayectorias espectrales relacionadas armónicamente entre sí (ver Figura 2.14B). Por otro lado, las sibilancias polifónicas están compuestas por un conjunto de picos espectrales de banda estrecha (tonos) no relacionados armónicamente entre sí. Por lo tanto, las sibilancias polifónicas, a lo largo del tiempo, se caracterizan por un conjunto de trayectorias espectrales sin relación armónica (ver Figura 2.15). Concretamente las sibilancias monofónicas se originan por la obstrucción de una de las vías respiratorias de mayor diámetro (tráquea, bronquios primarios y secundarios), y están relacionadas con el asma [294, 117, 274]. Sin embargo, las sibilancias polifónicas son originadas por una obstrucción múltiple en las vías respiratorias más centrales y de menor diámetro en los pulmones (bronquiolos), y están comúnmente relacionadas con la EPOC [294, 117, 155].

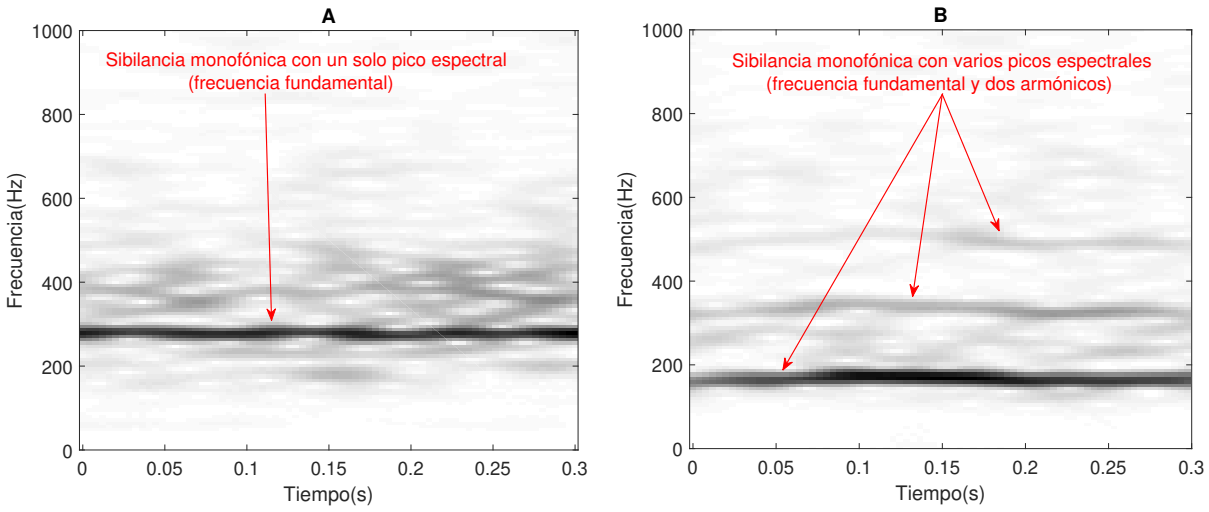

Figura 2.14: Representación tiempo-frecuencia (espectrograma) de dos ejemplos de sibilancias monofónicas: A) Sibilancia monofónica compuesta por un único pico espectral (una trayectoria espectral continua en el tiempo); y B) Sibilancia monofónica compuesta por varios picos espectrales, la componente de la frecuencia fundamental y sus armónicos (varias trayectorias espectrales relacionadas armónicamente).

El estridor (stridor) es un sonido musical fuerte, agudo (tono alto) y continuo que se produce por el rápido flujo de aire que atraviesa un segmento obstruido en las vías respiratorias extratorácicas (tráquea y laringe) [51, 58], por ello suele estar solapado con los sonidos respiratorios traqueales. Tiene un mecanismo de producción análogo al de la vibración de la lengüeta de un instrumento musical, como ocurre con las sibilancias. En el análisis de señal se caracteriza

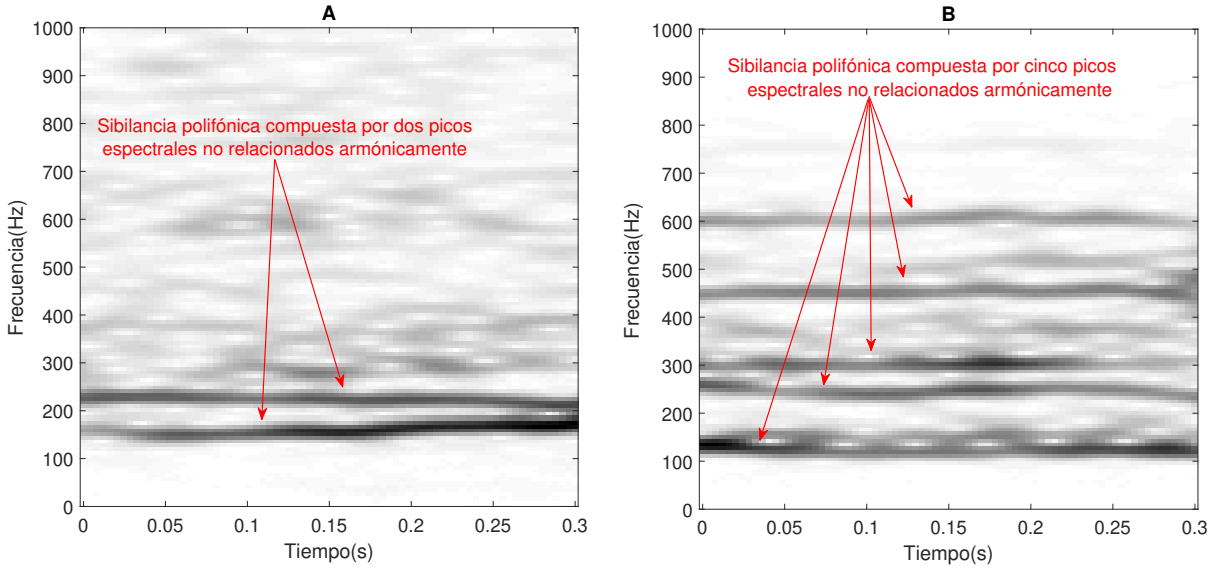

Figura 2.15: Representación tiempo-frecuencia (espectrograma) de dos ejemplos de sibilancias polifónicas. En ambos casos las sibilancias polifónicas se componen de varios picos espectrales sin relación armónica (varias trayectorias espectrales no relacionadas armónicamente).

por tener una forma de onda sinusoidal, con una frecuencia fundamental generalmente superior a 500 Hz (pudiendo alcanzar los 2.000 Hz), a menudo acompañada de varios armónicos, y una duración superior a los 250 ms [58, 265] (ver Figura 2.16). Considerando las características del estridor, es habitual definirlo como un tipo de sonido sibilante monofónico. Sin embargo, el estridor puede distinguirse de las sibilancias porque es un sonido más prominente en el cuello que en el interior de la pared torácica, generalmente es un sonido más intenso que las sibilancias y suele aparecer principalmente solo durante la inspiración. Como se ha mencionado, el estridor suele estar relacionado con la obstrucción de la tráquea o la laringe, la cual puede estar producida por: epiglotitis, laringotraqueitis, traqueomalacia, laringomalacia, estenosis, anafilaxia, carcinoma traqueal, disfunción de las cuerdas vocales o inhalación de un cuerpo extraño en la tráquea o la laringe, entre otras posibles patologías o anomalías [274, 151, 58]. Además, la evaluación del estridor es especialmente útil en los pacientes de la unidad de cuidados intensivos que han sido sometidos a una extubación, porque su aparición puede ser un signo de obstrucción de las vías respiratorias extratorácicas que requiere una intervención rápida [145].

El roncus (rhonchi) es considerado una variante del sonido sibilante, diferenciado principalmente por tener un tono más grave. En el análisis de señal se caracteriza por tener una forma de onda sinusoidal, con una frecuencia fundamental menor a los 200 Hz, generalmente cercana a los 150 Hz (responsable de su parecido con el sonido de los ronquidos), y una duración superior a los 100 ms [58, 247] (ver Figura 2.17). El roncus, al ser un sonido de tono bajo, se escucha mejor sobre la pared torácica, por ello se encuentra solapado con los sonidos respiratorios broncovesiculares o pulmonares. Puede ocurrir tanto en la inspiración como en la espiración y es alterado al toser. Es un síntoma típico en la bronquitis aguda o crónica (EPOC) y suele estar

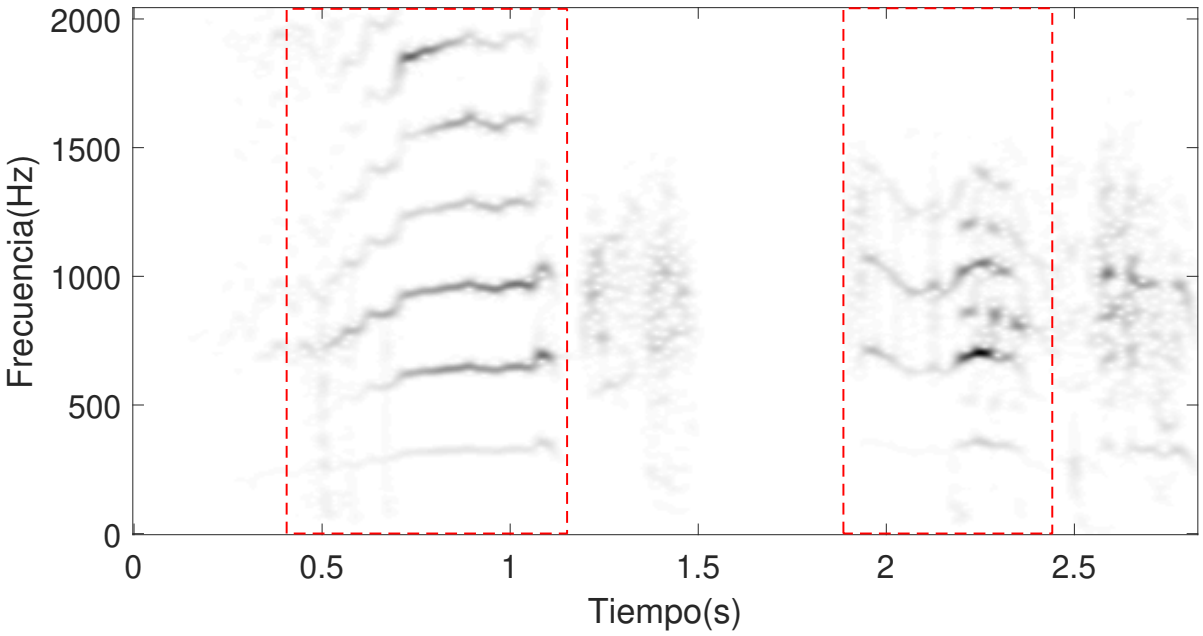

Figura 2.16: Representación tiempo-frecuencia (espectrograma) de una señal de audio respiratoria con varios sonidos estridor (stridor). Note que los rectángulos rojos señalan la fase de inspiración donde estos sonidos están activos.

acompañado de una hipersecreción bronquial. Normalmente desaparece con la tos, excepto en los denominados casos de roncus fijo, en los que la tos no lo elimina, lo que suele indicar la obstrucción de las vías respiratorias por cuerpos extraños. El roncus y las sibilancias probablemente comparten el mismo mecanismo de generación, sin embargo el roncus, a diferencia de las sibilancias, puede desaparecer después de la tos, lo que sugiere que las secreciones juegan un papel fundamental en la generación de estos sonidos. Por ello, este sonido suele considerarse un indicador de la constricción de las paredes de las vías respiratorias asociada al engrosamiento de la mucosa, el edema o el broncoespasmo [58].

Sonidos adventicios discontinuos (no musicales):

Los sonidos adventicios discontinuos, denotados en inglés como Discontinuous Adventitious Sounds (DAS), son definidos principalmente por su discontinuidad temporal. A diferencia de los CAS, los DAS no tienen propiedades musicales, ya que no se pueden caracterizar como notas que describen una trayectoria espectral a lo largo del tiempo. Los DAS suelen definirse como sonidos explosivos o burbujeantes con una duración inferior a los 25 ms y con un carácter repetitivo [162]. Por ello, en lugar de describir trayectorias espectrales a lo largo del tiempo, se pueden representar con pulsos o patrones espectrales de banda ancha que aparecen de forma intermitente a lo largo del tiempo (este comportamiento puede ser visto en el espectrograma de estos sonidos). Entre los principales sonidos adventicios discontinuos se encuentran los sonidos crepitantes (crackles) y el frote pleural (pleural rub).

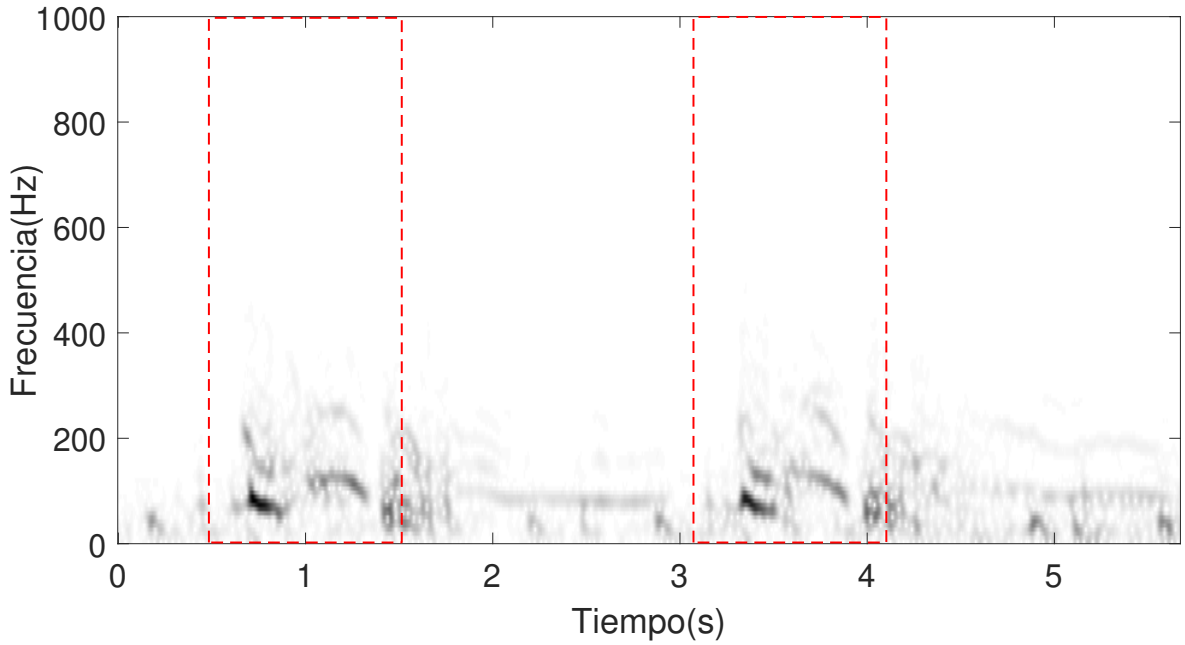

Figura 2.17: Representación tiempo-frecuencia (espectrograma) de una señal de audio respiratoria con varios sonidos roncus (rhonchi). Note que los rectángulos rojos delimitan las zonas dentro del ciclo respiratorio donde estos sonidos se encuentran activos.

Los sonidos crepitantes (crackles) son sonidos adventicios discontinuos, explosivos y no musicales que se escuchan normalmente en la inspiración y a veces durante la espiración. Los sonidos crepitantes suelen indicar que existe una anomalía patológica en el tejido pulmonar o en las vías respiratorias. En términos generales, el rango de frecuencia de estos sonidos es de 60-2.000 Hz [288], encontrándose la mayor proporción de la potencia en el rango de 60-1.200 Hz [37]. Además, los sonidos crepitantes, al ser discontinuos tienen una duración inferior a 20 ms, y aparecen de forma intermitente durante el ciclo respiratorio en forma de patrones espectrales en banda ancha, como se puede ver en la Figura 2.18. Las condiciones clínicas en las que pueden presentarse crepitaciones incluyen neumonía, fibrosis pulmonar, EPOC, bronquiectasia e insuficiencia cardíaca, entre otras [58, 287]. En la detección de sonidos crepitantes, el número de crepitantes en un mismo ciclo respiratorio es importante porque permite indicar la gravedad de los trastornos pulmonares y de las vías respiratorias [249, 292]. Sin embargo, más que la cantidad de crepitantes, su posición dentro del ciclo respiratorio y la variación en la forma de onda que los genera, son las características clave que permiten determinar el tipo de patología pulmonar [287].

Los crepitantes suelen clasificarse como **crepitantes finos (fine crackles)** o **crepitantes gruesos (coarse crackles)** en función de una serie de consideraciones:

- El mecanismo para la generación de los crepitantes finos es la repentina apertura inspiratoria de las vías respiratorias pequeñas, mantenidas cerradas por las fuerzas de la superficie durante la anterior espiración [315, 162]. Por otro lado, los crepitantes gruesos

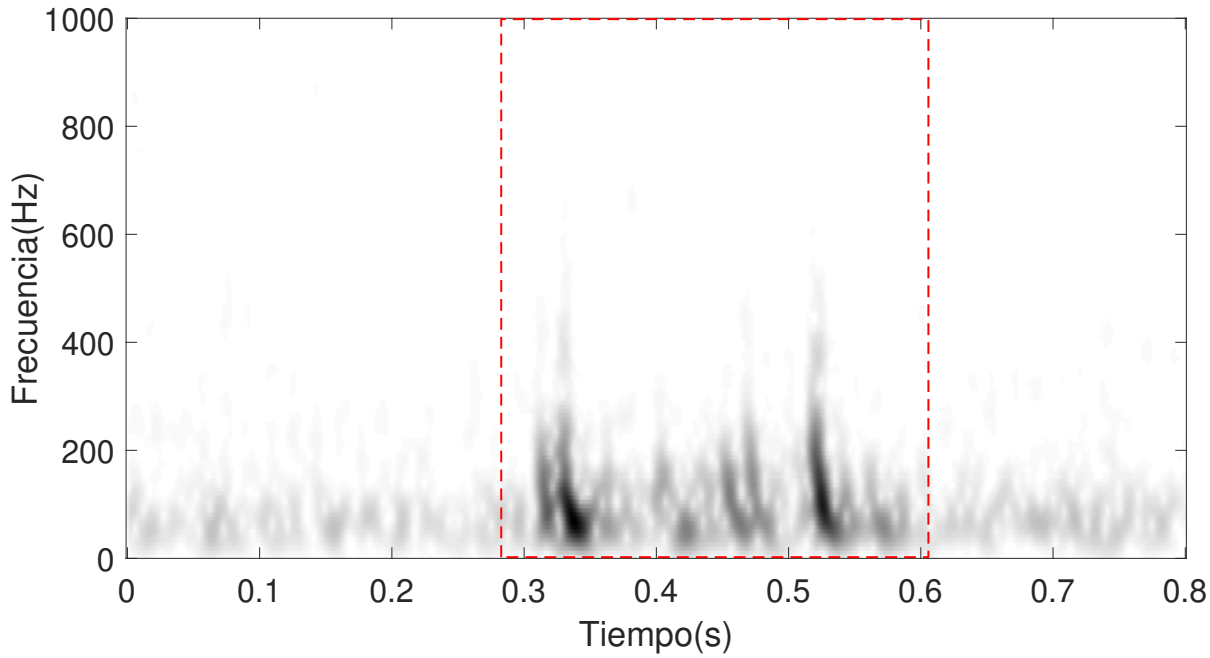

Figura 2.18: Representación tiempo-frecuencia (espectrograma) de una señal de audio respiratoria con sonidos crepitantes. Note que el rectángulo rojo marca la zona del ciclo respiratorio donde están presentes los sonidos crepitantes. Los sonidos crepitantes aparecen de forma intermitente como patrones espectrales de banda ancha.

son producidos por burbujas de aire que pasan a través de bronquios largos o segmentos bronquiectásicos, cuando se abren y cierran de forma intermitente [117, 162].

- En auscultación, los crepitantes finos se suelen escuchar durante la mitad o el final de la inspiración, en regiones pulmonares dependientes (donde se localizan las vías respiratorias de menor tamaño), y no son transmitidos a la boca. Sin embargo, los crepitantes gruesos se pueden escuchar en el inicio de la inspiración y a lo largo de la espiración. Además, pueden ser auscultados en cualquier región pulmonar y pueden ser transmitidos a la boca [58].

- Por otro lado, los crepitantes finos no se ven alterados por la tos, aunque pueden cambiar o desaparecer por cambios en la posición del cuerpo (por ejemplo, inclinándose hacia delante). En cambio, los crepitantes gruesos, pueden cambiar o desaparecer con la tos y no se ven influenciados por cambios en la posición del cuerpo [58].

- Los crepitantes gruesos son fuertes, de tono bajo y menos numerosos por respiración, mientras que los crepitantes finos son suaves, de tono más alto y más numerosos por respiración [274].

- En términos de sonoridad, los crepitantes gruesos suenan como la sal vertida en una sartén caliente, mientras que los crepitantes finos suenan más como tiras de velcro que se separan lentamente o una botella de agua con gas que se abre [274].

- En el análisis del sonido, los crepitantes finos tienen una duración más corta (en torno a 5 ms) en comparación con los gruesos (en torno a 15 ms). Además, los crepitantes finos se caracterizan por una frecuencia mayor (en torno a 650 Hz), en comparación con los gruesos (en torno a 350 Hz) [58].
- En el análisis patológico, los crepitantes finos están relacionados con neumonía, fibrosis o insuficiencia cardiaca, entre otras. Mientras que los crepitantes gruesos están relacionados con bronquiectasia, edema pulmonar grave, bronquitis crónica y EPOC, entre otras [251, 223].

Por otro lado, Murphy [225] propuso clasificar objetivamente el tipo de sonido crepitante mediante el análisis de su forma de onda característica a lo largo del tiempo. Como se puede observar en la Figura 2.19, la forma de onda de los sonidos crepitantes generalmente se puede representar como una onda sinusoidal larga y amortiguada [336, 307]. Concretamente, el parámetro IDW se refiere a la duración (ms) entre el comienzo y la primera intersección con la línea temporal (ya sea por encima o por debajo); el parámetro 2CD indica la duración de los dos primeros ciclos del sonido crepitante; y el parámetro TDW corresponde a la duración total del sonido crepitante. Considerando estos parámetros, la ATS clasifica los sonidos crepitantes como gruesos cuando las duraciones medias de IDW y 2CD están en torno a 1.5 y 10 ms, respectivamente, y como finos cuando están en torno a 0.7 y 5 ms [286]. Por otro lado, CORSA establece que los crepitantes gruesos ocurren cuando 2CD>10 ms, y finos cuando 2CD<10 ms [288].

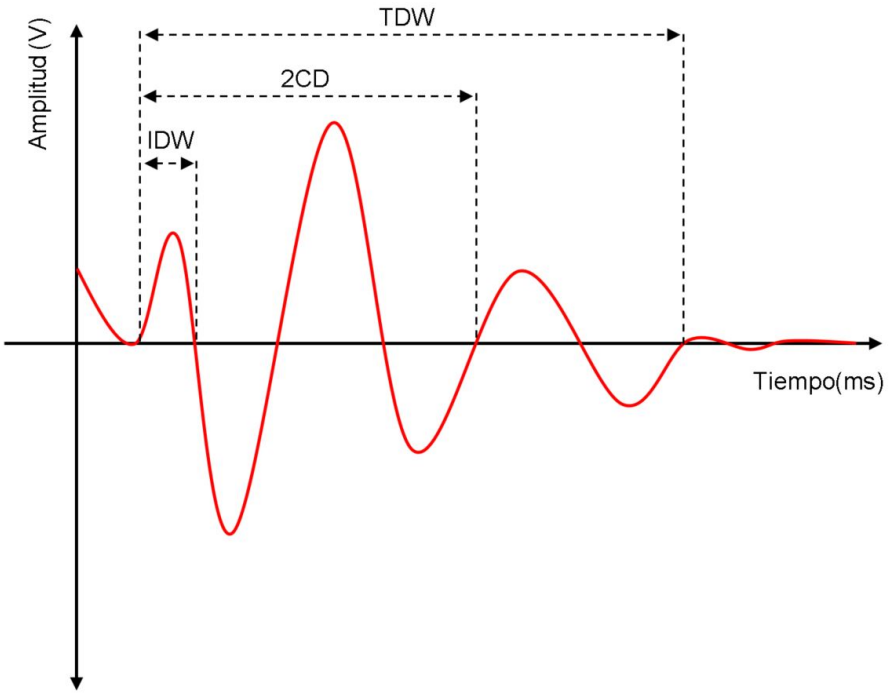

Figura 2.19: Forma de onda genérica de un sonido crepitante.

El frote pleural (pleural rub): es un sonido adventicio discontinuo, explosivo, de corta duración, de carácter rítmico y chirriante que se puede escuchar durante la inspiración y la espiración (ver Figura 2.20). Típicamente el componente inspiratorio es mostrado en el componente espiratorio [117]. Estos sonidos son causados por el roce de las membranas pleurales durante la respiración. En personas sanas, la pleura parietal y visceral se deslizan uno sobre la otra silenciosamente. Sin embargo, en personas con ciertas patologías respiratorias, la pleura visceral se inflama y se solidifica. Por lo que el deslizamiento entre ambas capas pleurales genera un rozamiento que produce el sonido denominado como frote pleural [58]. En términos de sonoridad, suele compararse con el sonido de andar por la nieve o como el crujido del cuero nuevo [42]. En el análisis del sonido, el frote pleural se caracteriza por tener una duración mayor de 15 ms y una frecuencia inferior a los 350 Hz [251, 42]. En términos patológicos, son sonidos causados por la pleuresía o por un tumor pulmonar, y suelen ser un indicador de enfermedades como la pleuritis o el mesotelioma [58]. El diagnóstico diferencial entre el frote pleural y los crepitantes gruesos suele ser difícil, porque la forma de onda de ambos sonidos es muy similar. Sin embargo, el frote pleural tiene una mayor duración y una frecuencia menor. Además, el frote pleural suele ser bifásico (presente en ambas fases del ciclo respiratorio) y no se ve alterado al toser. En cambio, los crepitantes gruesos pueden ocurrir independientemente en ambas fases y se ven alterados al toser [274].

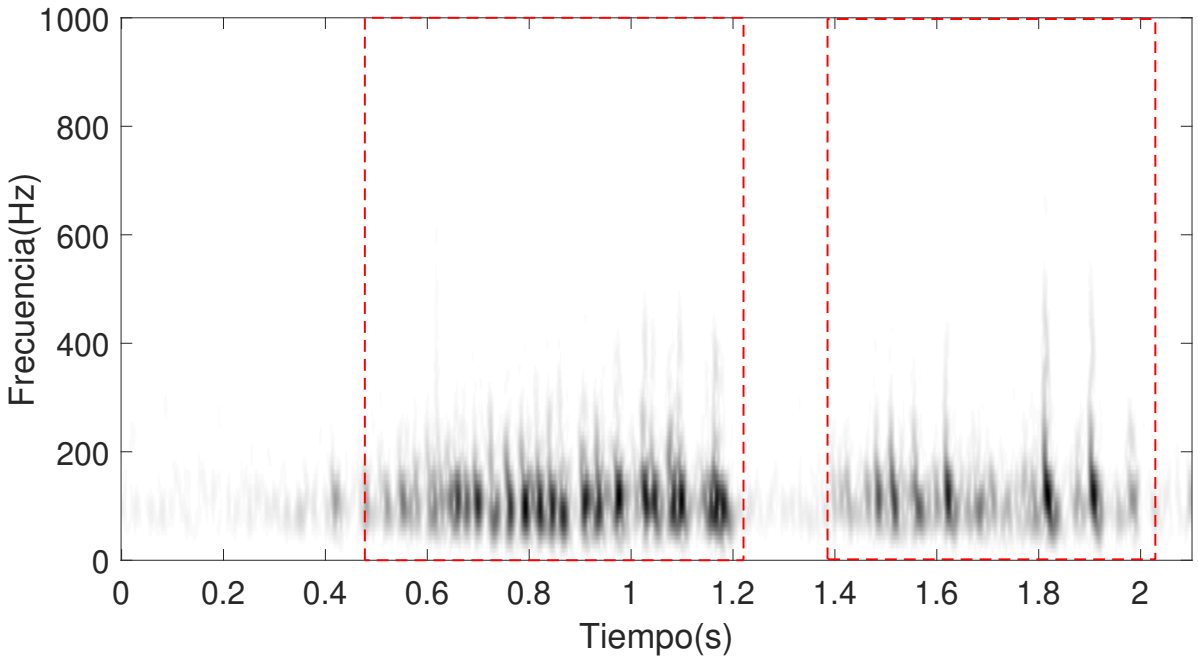

Figura 2.20: Representación tiempo-frecuencia (espectrograma) de una señal de audio respiratoria con el denominado sonido adventicio frote pleural. Note que los rectángulos rojos señalan las etapas de inspiración y espiración donde se genera el frote pleural. El frote pleural aparece de forma rítmica como patrones espectrales de banda ancha.

Sonidos adventicios mixtos (continuos y discontinuos):

Dentro de esta categoría podemos encontrar **los graznidos (squawks)** que contienen tanto componentes musicales, como no musicales. Los graznidos son sonidos que aparecen durante el final de la inspiración, y se denominan mixtos, porque están compuestos por una sibilancia de corta duración y sonidos crepitantes (ver Figura 2.21). Estos sonidos son generados por las oscilaciones producidas en las vías respiratorias periféricas (las de menor diámetro), localizadas en zonas pulmonares desinfladas, cuando sus paredes permanecen en contacto durante un periodo de tiempo largo y seguidamente se abren al final de la inspiración [58, 117]. En el análisis del sonido, los graznidos se caracterizan por contener una sibilancia corta con una duración inferior a los 200 ms y una frecuencia fundamental situada entre 200 y 300 Hz, junto con algunos sonidos crepitantes [58]. Además, La frecuencia fundamental de la sibilancia suele venir acompañada por un conjunto de armónicos, como se puede observar en la Figura 2.21. En términos patológicos, estos sonidos pueden ser escuchados en pacientes con fibrosis pulmonar a causa de una neumonitis por hipersensibilidad [97], en pacientes con trastornos pulmonares intersticiales [287, 288] y en pacientes con neumonía o bronquiolitis obliterante [237, 124].

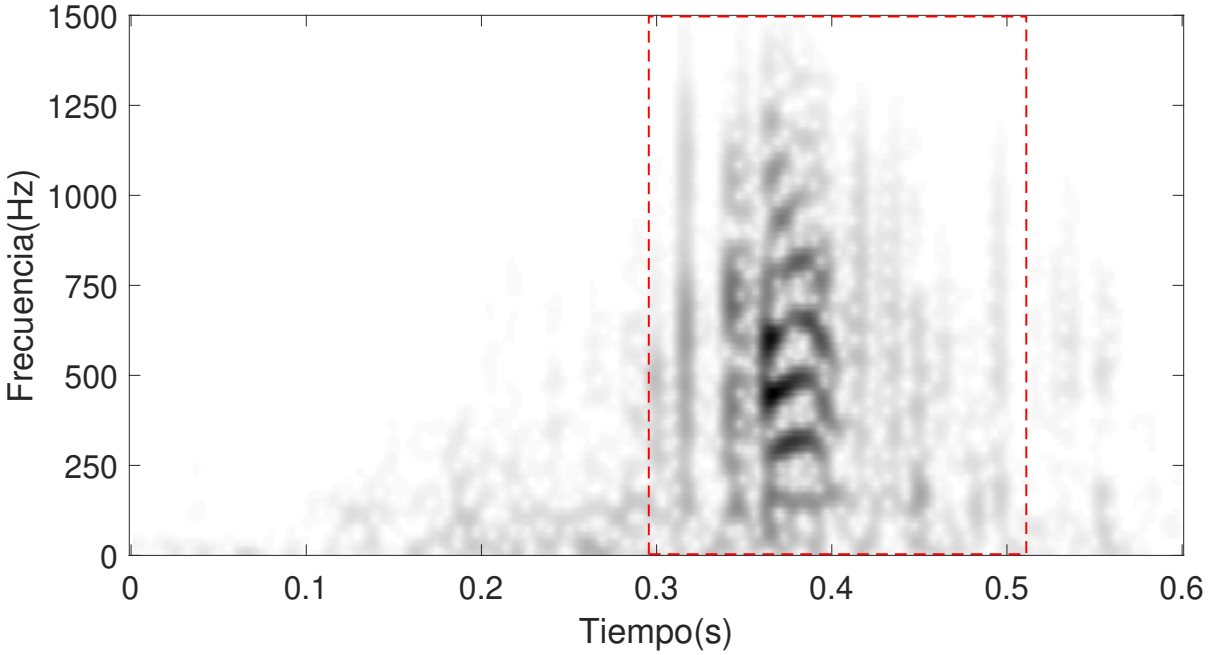

Figura 2.21: Representación tiempo-frecuencia (espectrograma) de una señal de audio respiratoria con graznidos. Note que el rectángulo rojo delimita la zona dentro del ciclo respiratorio donde está presente este sonido. Los graznidos se componen de una sibilancia de corta duración y varios sonidos crepitantes.

Para finalizar con esta sección, a modo de resumen, la Tabla 2.2 muestra una clasificación completa de los diferentes tipos de sonidos adventicios, en función de sus características, causas y patologías asociadas.

Sonido Adventicio	Tipo	Duración (ms)	Frecuencia (Hz)	Timbre	Fase Respiratoria	Causa	Patología Asociada
Sibilancia	Continuo	>100	(100-1.000)	Musical	Inspiración, espiración o bifásico	Estrechamiento y obstrucción de las vías respiratorias, limitación del flujo de aire	Asma, EPOC, Bronquiolitis, Bronquiectasia
Estridor	Continuo	>250	(500-2.000)	Musical	Inspiración	Flujo de aire turbulento en la laringe o la tráquea, obstrucción de estas vías	Epiglotitis, Laringotraqueitis, Traqueomalacia, Laringomalacia, Anafilaxia, Estenosis
Roncus	Continuo	>100	<200	Musical	Inspiración, espiración o bifásico	Secreción en los bronquios, engrosamiento de la mucosa	Bronquitis aguda o crónica (EPOC)
Crepitantes Finos	Discontinuo	±5	±650	No musical, explosivo	Inspiración (al inicio)	Apertura repentina de las pequeñas vías respiratorias	Neumonía, Fibrosis, Insuficiencia cardiaca,
Crepitantes Gruesos	Discontinuo	±15	±350	No musical, explosivo	Inspiración (al final), espiración o bifásico	Burbujas de aire en los grandes bronquios o segmentos bronquiectásicos	Bronquiectasia, Edema pulmonar grave, Bronquitis crónica (EPOC)
Frote Pleural	Discontinuo	>15	<350	No musical, Rítmico	Bifásico	Roce de las membranas pleurales entre sí	Pleuresía, Tumor pulmonar
Graznido	Mixto	<200	(200-300)	Musical y No musical	Inspiración (al final)	Oscilaciones en las vías respiratorias periféricas	Fibrosis pulmonar, Trastornos Intersticiales, Neumonía, Blonquiolitis obliterante

Tabla 2.2: Clasificación de los distintos tipos de sonidos respiratorios adventicios.

2.4. Conclusiones

En este segundo capítulo se han presentado conceptos básicos relacionados con el audio biomédico respiratorio necesarios para entender cómo los sonidos respiratorios son generados y la importancia de los sonidos adventicios para identificar posibles patologías respiratorias en el sujeto. En primer lugar, se ha realizado una descripción del sistema respiratorio humano, atendiendo a su anatomía, fisiología y patología, la cual permite comprender la naturaleza de los sonidos respiratorios y las diferentes partes implicadas en su generación, así como las patologías pulmonares obstructivas de mayor relevancia en relación a los sonidos sibilantes. En segundo lugar, se han descritos los principios básicos sobre el procedimiento de auscultación, identificando sus ventajas y limitaciones, así como las diferentes opciones comerciales disponibles para realizar la grabación de los sonidos auscultados. Por último, se ha realizado una descripción de los distintos tipos de sonidos respiratorios que pueden producirse durante la mecánica de la respiración, categorizando entre los sonidos respiratorios normales y adventicios, y distinguiendo los distintos sonidos adventicios que pueden ser producidos durante la auscultación de un sujeto que sufre alguna patología respiratoria obstructiva.

Para concluir este capítulo es preciso indicar que se ha realizado una presentación completa de los distintos tipos de sonidos adventicios, con el objetivo de mostrar las diferencias y similitudes entre ellos. Sin embargo, esta Tesis está centrada únicamente en el análisis de los sonidos sibilantes.

CAPÍTULO 3

Factorización de matrices no negativas

ESTE capítulo trata la temática de los modelos de señal basados en factorización de matrices no negativas. Se inicia presentando la motivación de utilizar estos enfoques en el campo del procesado de audio. Además, se describen los principios en los que se basan los modelos de factorización de matrices no negativas y se presenta una clasificación de los diferentes enfoques desarrollados para la separación de fuentes sonoras. Por otro lado, se describen las principales regularizaciones y restricciones que pueden ser incorporadas a los modelos de descomposición para caracterizar el comportamiento tiempo-frecuencia de los sonidos presentes. Para finalizar, se presentan un conjunto de descriptores ampliamente aplicados en estrategias de clustering.

Notación matemática

Las letras mayúsculas en negrita denotan matrices, como por ejemplo $\mathbf{X}$. Las letras minúsculas en negrita denotan vectores, como por ejemplo $\mathbf{x}$. Las letras mayúsculas y minúsculas simples, como n o N, denotan escalares. $\mathbb{R}_+$ denota el conjunto de escalares positivos. El vector de entrada i^{th} de una matriz $\mathbf{A}$ se indica con letras minúsculas en negrita como $\boldsymbol{a}_i$. El elemento de entrada $(i,j)^{th}$ de una matriz $\mathbf{A}$ se indica con letras minúsculas como a_{ij}. La multiplicación elemento a elemento entre dos matrices $\mathbf{A}$ y $\mathbf{B}$ se indica como $\mathbf{A} \odot \mathbf{B}$. La división elemento a elemento entre dos matrices $\mathbf{A}$ y $\mathbf{B}$ se indica como $\mathbf{A} \oslash \mathbf{B}$. El operador traspuesto de una matriz $\mathbf{A}$ está representado como $\mathbf{A}^{\mathrm{T}}$. La estimación de un parámetro se indica con el símbolo $\hat{}$, como $\hat{\mathbf{A}}$, $\hat{\boldsymbol{a}}$, $\hat{a}$.

3.1. Introducción

Muchos tipos de datos pueden representarse como combinaciones constructivas de partes, es decir, como combinaciones estrictamente aditivas, donde ninguna de las partes produce sustracción. Estos datos se denominan a menudo datos de composición, y los modelos matemáticos utilizados para representarlos se denominan modelos de composición. Estos modelos adoptan la forma de combinaciones lineales no negativas de partes que también son no negativas, asegurando que la combinación sea puramente constructiva. Este concepto dio lugar a la aparición del denominado enfoque "Factorización de matrices no negativas", denotado en inglés como Non-negative Matriz Factorization (NMF). El enfoque NMF fue iniciado por Paatero y Tapper [236, 235], y posteriormente por Lee y Seung [184, 185]. En términos generales, el enfoque NMF es una técnica de factorización que se utiliza para la representación lineal de datos bidimensionales no negativos y su principal ventaja es que puede utilizarse para reducir la dimensionalidad de una gran cantidad de datos con el fin de encontrar estructuras ocultas mediante la representación basada en partes con patrones de datos no negativos. Específicamente, y como veremos más adelante, los modelos NMF [105, 312] tratan las representaciones tiempo-frecuencia (espectrograma) no negativas de la señal de entrada como matrices, que se descomponen en productos de matrices, cuyas componentes son no negativas. Así, algunas de estas matrices representan los patrones espectrales (características espectrales) de los eventos sonoros existentes en la señal de entrada y otras matrices representan su respectiva activación temporal (características temporales), es decir, aquellos intervalos temporales en los cuales los eventos sonoros anteriormente mencionados se encuentran activos. En definitiva, debido a la propiedad de no negatividad, el enfoque NMF genera un modelo de interpolación lineal aditiva que da como resultado la llamada representación basada en partes u objetos [146, 184].

El enfoque NMF se convirtió en una herramienta de análisis de datos cada vez más popular y, durante las últimas dos décadas, ha sido exitosamente aplicado en muchos campos, como en procesado de imágenes [340, 244, 158, 195, 319], procesado de audio [260, 105, 345, 64, 243], procesado de textos [142] o biomedicina [73, 85, 301, 62]. La razón de que el modelo NMF se haya convertido en uno de los enfoques más populares y versátiles, a diferencia de muchas otras representaciones lineales como el Análisis de Componentes Independientes, denotado en inglés como Independent Component Analysis (ICA) [150, 194, 220, 238], y el Análisis de Componentes Principales, denotado en inglés como Principal Component Analysis (PCA) [56, 137, 334], es su capacidad para obtener una representación basada en partes de los objetos más representativos (por ejemplo, notas musicales y acordes en audio) imponiendo restricciones y regularizaciones no negativas que permiten sólo combinaciones aditivas, no sustractivas de los datos de entrada.

Las descomposiciones de las señales de audio basadas en el enfoque NMF han dado lugar a novedosas soluciones para afrontar diversos problemas en el procesado de audio. Específicamente, Smaragdis y Brown [283] fueron los encargados de iniciar una importante línea de investigación basada en el modelo NMF para la transcripción musical, separación de fuentes sonoras, mejora del habla, etc. Desde el comienzo de esta línea han surgido numerosas con-

tribuciones dedicadas a extraer o suprimir instrumentos específicos en pistas de audio mixtas [312, 170, 139], eliminar las fuentes interferentes al habla para mejorar su calidad e inteligibilidad [49, 161, 261, 291, 327, 282], reconocimiento de los instrumentos presentes en una pista musical [271], clasificación del género musical [242], transcripción automática de la partitura de los instrumentos [55, 66, 138, 183], codificación [232, 250], reconocimiento de voz [125, 126] o identificación del orador [302]. La característica común de todas estas propuestas es la descomposición no negativa del espectrograma de la señal observada (señal de entrada) en un diccionario de componentes espectrales elementales, que modelan el comportamiento espectro-temporal de los sonidos de interés (notas, acordes, sonidos percusivos, ruido interferente u otras estructuras adaptativas más complejas).

Sin embargo, tras una revisión exhaustiva, se comprobó por el autor de esta Tesis que el enfoque NMF nunca antes había sido aplicado en el ámbito del procesado de los sonidos adventicios, en general, y en el ámbito del procesado de los sonidos sibilantes, en particular. La motivación para aplicar estos modelos en el análisis de los sonidos sibilantes se debe principalmente a que estos sonidos pueden ser modelados con patrones espectrales, de forma similar o cercana a como se modelan las notas de un instrumento armónico y más similar a la voz del cantante (singing-voice). Esto se debe a que las sibilancias se componen de uno o varios tonos que definen trayectorias espectrales a lo largo del tiempo, de la misma forma a como ocurre con las notas musicales. Además, los sonidos respiratorios normales, los cuales siempre están solapados con los sonidos sibilantes, pueden ser modelados por patrones espectrales en banda ancha que los diferencian claramente del comportamiento de las sibilancias. Todo ello ha propiciado que todos los métodos o algoritmos desarrollados y aportados en esta Tesis, aprovechen y exploten las capacidades del enfoque NMF para contribuir en las principales tareas de interés relacionadas con el análisis de los sonidos sibilantes: eliminación del ruido ambiente que rodea al sujeto durante la auscultación, mejora de audio de los sonidos sibilantes, detección temporal de las sibilancias y clasificación del tipo de sibilancia considerando su estructura armónica.

En las siguientes secciones de este capítulo se presentaran los principios básicos del enfoque NMF estándar, la clasificación de los modelos NMF/NMPCF para la separación de fuentes sonoras, las restricciones y regularizaciones que suelen ser utilizadas para mejorar el modelado de las características de los sonidos, los descriptores más utilizados para la clasificación (clustering) de las bases espectrales que componen el diccionario de bases espectrales y por último se resumirán los aspectos clave de los enfoques NMF/NMPCF utilizados en las contribuciones aportadas durante el desarrollo de esta Tesis doctoral. Considerar que este capítulo únicamente se centra en el procesado de señales de audio.

3.2. Modelo estándar

Como se ha mencionado anteriormente, todo lo explicado en relación al enfoque de descomposición NMF se centra en el procesado de la señal de audio. Por lo tanto, partiendo de una señal sonora de entrada $\mathbf{s}(n)$, donde $n = 1, ..., N$ y N denota el número total de muestras

considerando la frecuencia de muestreo aplicada. El espectrograma en magnitud $\mathbf{X} \in \mathbb{R}_{+}^{F \times T}$ de la señal de entrada $\mathbf{s}(n)$ puede ser obtenido a partir de la magnitud de la transformada STFT, aplicando una función ventana (como por ejemplo, Hamming o Hann) y un solapamiento entre las ventanas para aumentar la resolución temporal. En este sentido, el espectrograma en magnitud de entrada $\mathbf{X}$ está compuesto por T tramas temporales (frames), F intervalos de frecuencia (bins) y un conjunto de unidades tiempo-frecuencia x_{ft}, donde $f = 1, ..., F$ y $t = 1, ..., T$. Varios espectrogramas, relacionados con la representación de sonidos respiratorios normales y adventicios, han sido previamente mostrados en la Sección 2.3.

En términos generales, el modelo NMF estándar representa el espectrograma en magnitud de entrada $\mathbf{X}$ como una combinación lineal no negativa de elementos espectrales (unidades atómicas) no negativos, denominados bases o componentes espectrales $\mathbf{b}_k$, donde $k = 1, ..., K$ y K denota el número total de componentes o bases espectrales de la combinación lineal. En su forma más simple, estas bases se consideran vectores espectrales, que representan sonidos de estado estacionario, de tal forma que cualquier vector espectral del espectrograma de entrada $\mathbf{x}_t$ puede ser descompuesto en una combinación lineal no negativa de estas denominadas bases espectrales $\mathbf{b}_k$, como se muestra a continuación:

$$\mathbf{x}_t = \sum_{k=1}^{K} \mathbf{b}_k a_{kt} \tag{3.1}$$

donde a_{kt} define la activación (también denominada peso o ganancia) para cada componente espectral $\mathbf{b}_k$ en la trama temporal t. En consecuencia, el espectrograma de entrada $\mathbf{X}$, en su totalidad, puede ser modelado por una combinación lineal de bases espectrales $\mathbf{b}_k$ que varía a lo largo de cada una de sus tramas t.

Una vez comentado el principio básico en el que se fundamenta este modelo, a continuación se describe el enfoque NMF estándar desde el propio punto de vista matricial. En este sentido, el enfoque NMF estándar, permite descomponer, aproximar o factorizar el espectrograma $\mathbf{X}$ como el producto de dos matrices: la matriz de bases (patrones) espectrales no negativas $\mathbf{B} \in \mathbb{R}_{+}^{F \times K}$ y la matriz de activaciones (pesos) temporales no negativas $\mathbf{A} \in \mathbb{R}_{+}^{K \times T}$, como se muestra a continuación:

$$\mathbf{X} \approx \hat{\mathbf{X}} = \mathbf{BA} \tag{3.2}$$

donde $\hat{\mathbf{X}} \in \mathbb{R}_{+}^{F \times T}$ es el espectrograma estimado o reconstruido. En este sentido, las columnas de la matriz de bases $\mathbf{B}$, denotadas como $\mathbf{b}_k$, definen las componentes o patrones espectrales que pueden describir el comportamiento espectral de eventos sonoros activos en el espectrograma de entrada $\mathbf{X}$, y las filas de la matriz de activaciones $\mathbf{A}$, denotadas como $\boldsymbol{a}_k$, representan la actividad temporal a_{kt} para cada componente espectral k en la trama temporal t. Dicho de otra forma, la matriz $\mathbf{B}$ forma un diccionario compuesto por K bases espectrales $\mathbf{b}_k$ y la matriz $\mathbf{A}$ define el peso a_{kt} con el que las diferentes componentes espectrales $\mathbf{b}_k$ aparecen a lo largo de las tramas temporales t. Una representación del modelo de descomposición NMF descrito en la Ec. (3.2) es mostrada en la Figura 3.1.

Antes de continuar con la explicación del modelo NMF, es necesario destacar que la selección del número de componentes K es fundamental para mejorar el rendimiento del proceso de

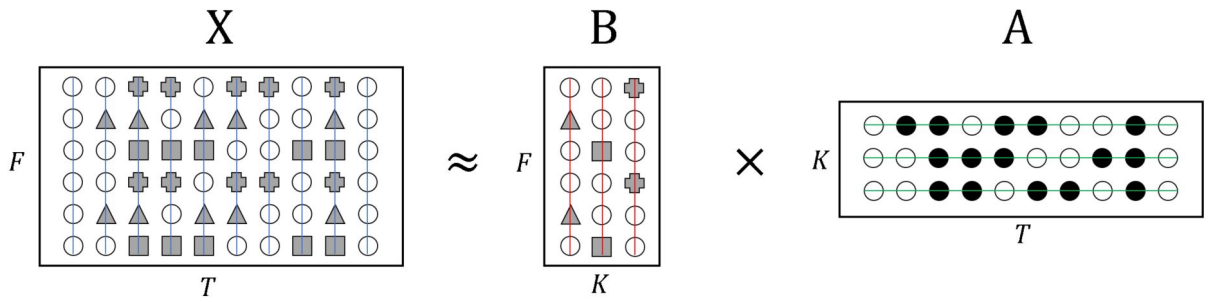

Figura 3.1: Ilustración del modelo de descomposición NMF estándar. El modelo descompone el espectrograma en magnitud de entrada $\mathbf{X} \in \mathbb{R}_+^{F \times T}$ en el producto de dos matrices $\mathbf{B} \in \mathbb{R}_+^{F \times K}$ y $\mathbf{A} \in \mathbb{R}_+^{K \times T}$. Las líneas rojas marcan las distintas componentes espectrales $\mathbf{b}_k$ que componen al diccionario $\mathbf{B}$. Las líneas verdes indican las filas $\boldsymbol{a}_k$ de la matriz $\mathbf{A}$, las cuales definen el comportamiento temporal a_{kt} para cada componente espectral $\mathbf{b}_k$ en la trama temporal t. Las líneas azules señalan los vectores espectrales $\mathbf{x}_t$ del espectrograma $\mathbf{X}$. Cada $\mathbf{x}_t$ puede ser descompuesto por una combinación lineal de $\mathbf{b}_k$ considerando su activación en cada trama $\boldsymbol{a}_k$.

descomposición. Un número reducido de componentes puede considerarse perjudicial, ya que varios sonidos presentes en la señal de entrada pueden factorizarse en una misma base espectral $\mathbf{b}_k$ del diccionario $\mathbf{B}$, sin que sea posible separarlos. Sin embargo, un número mayor de componentes no puede considerarse perjudicial, ya que el valor absoluto de las bases espectrales no significativas será casi cero o muy pequeño en comparación con las bases espectrales significativas [276]. Generalmente, se elige K de tal manera que $FK + KT \ll FT$ para reducir la dimensión de los datos.

Dicho esto, la descomposición o factorización del espectrograma de la señal de entrada $\mathbf{X}$ en el producto $\mathbf{BA}$ suele ser obtenida minimizando una función de divergencia,

$$\arg\min_{\mathbf{B},\mathbf{A}} D(\mathbf{X}|\mathbf{BA}) \quad \mathbf{B}, \mathbf{A} \geq 0 \tag{3.3}$$

Esta función de divergencia mide el error cometido entre el espectrograma de entrada $\mathbf{X}$ y el espectrograma estimado $\hat{\mathbf{X}} = \mathbf{BA}$. En el caso ideal, donde la estimación es óptima, es decir cuando $\mathbf{X} = \hat{\mathbf{X}} = \mathbf{BA}$, el valor de la divergencia es cero. Típicamente, la función de divergencia suele ser calculada a nivel escalar, es decir,

$$D(\mathbf{X}|\hat{\mathbf{X}}) = D(\mathbf{X}|\mathbf{BA}) = \sum_{f=1}^{F} \sum_{t=1}^{T} d(x_{ft}, \hat{x}_{ft}) \tag{3.4}$$

donde $d(x, \hat{x})$ es una función de dos variables escalares $x, \hat{x}$. A menudo denominada función de coste.

La función de coste o divergencia $d(x, \hat{x})$ más común en los problemas de descomposición de matrices es la distancia Euclídea (EUC) [105], expresada como

$$d_{\text{EUC}}(x, \hat{x}) = (x - \hat{x})^2 \tag{3.5}$$

Sin embargo, en el contexto del modelado de los patrones sonoros, se han encontrado otras medidas de divergencia más apropiadas [66, 312, 323]. La principal limitación de la distancia Euclídea es que tiende a enfatizar los errores producidos en las energías más altas, lo que da lugar a soluciones en las que sólo se representan con precisión los intervalos en frecuencia (bins) de mayor energía. Esto genera un problema importante en el análisis del audio, ya que las señales de audio suelen tener un gran rango dinámico, y algunas de las componentes de menor energía (a menudo en frecuencias más altas) son perceptiblemente tan importantes como los componentes de alta energía. Por ello, en el procesado de audio se requieren medidas de divergencia que asignen mayor énfasis a las componentes de baja energía. Dos alternativas ampliamente utilizadas son: la divergencia Kullback-Leibler (KL)

$$d_{\mathrm{KL}}(x, \hat{x}) = x \log(x/\hat{x}) - x + \hat{x} \tag{3.6}$$

y la divergencia Itakura-Saito (IS)

$$d_{\mathrm{IS}}(x, \hat{x}) = x/\hat{x} - \log(x/\hat{x}) - 1 \tag{3.7}$$

A diferencia de la distancia EUC, la divergencia IS asigna la misma importancia a las componentes de alta y baja energía, ya que se considera una métrica invariante en la escala. La divergencia KL proporciona un buen compromiso entre ambos [66, 312, 323] y es por ello que haya sido utilizada en el desarrollo de todos los modelos NMF propuestos en esta Tesis doctoral. En el procesado de audio, es popular generalizar las funciones de divergencia descritas anteriormente en la función β-divergencia [107, 78, 176], la cual es definida en función del parámetro β, como:

$$d_\beta(x, \hat{x}) = \begin{cases} \frac{1}{\beta(\beta-1)}(x^\beta + (\beta-1)\hat{x}^\beta - \beta x \hat{x}^{\beta-1}), & \beta \in \mathbb{R}_+ \setminus \{0, 1\} \\ x \log \frac{x}{\hat{x}} - x + \hat{x}, & \beta = 1 \\ \frac{x}{\hat{x}} - \log \frac{x}{\hat{x}} - 1, & \beta = 0. \end{cases} \tag{3.8}$$

Como puede comprobarse, la distancia EUC ($\beta = 2$), la divergencia KL ($\beta = 1$) y la divergencia IS ($\beta = 0$) son casos particulares de esta medida genérica. Además, también existen divergencias que tienen por objetivo optimizar la calidad perceptiva de la representación [232], lo que resulta útil en las aplicaciones de codificación de audio. Sin embargo, en la mayoría de las demás aplicaciones basadas en modelos de descomposición, como la separación de fuentes sonoras y el análisis de señales de audio, la calidad de la representación se ve más afectada por su capacidad de aislar las unidades de composición latentes de una señal mixta, que por la capacidad de representar con precisión las observaciones. Por lo tanto, las divergencias simples como KL o IS son las más utilizadas incluso en aplicaciones en las que una mezcla se separa en partes para escucharlas por separado.

Como se ha mencionado anteriormente, el problema descrito en la Ec. (3.3) requiere minimizar una función de divergencia, indicada en la Ec. (3.4), para estimar las matrices de bases $\mathbf{B}$ y activaciones $\mathbf{A}$, y así obtener el espectrograma estimado $\hat{\mathbf{X}}$ con el menor error de reconstrucción en comparación con el espectrograma de entrada $\mathbf{X}$. Los métodos de minimización más

comunes para resolver el problema descrito son denominados: Actualizaciones Multiplicativas, denotado en inglés como Multiplicative Update (MU) [185], y Maximización de la Esperanza, denotado en inglés como Expectation Maximization (EM) [91]. En este sentido, la técnica más popular para estimar los parámetros del modelo NMF fue propuesta inicialmente por Lee y Seung [185], y consiste en un algoritmo iterativo basado en las denominadas reglas de actualización multiplicativas. Primero, los parámetros del modelo (específicamente las matrices $\mathbf{B}$ y $\mathbf{A}$) son inicializados con valores positivos aleatorios, y luego son actualizados iterativamente aplicando un gradiente multiplicador. En contraste con otros métodos basados en el gradiente, las reglas de actualización multiplicativas aseguran la no negatividad de los parámetros, ya que tanto el gradiente como los parámetros son no negativos. Por lo tanto, a partir de la Ec. (3.4) (considerando la función de coste general β-divergencia), las matrices $\mathbf{B}$ y $\mathbf{A}$ pueden ser estimadas aplicando un algoritmo de gradiente descendiente basado en las denominadas reglas de actualización multiplicativas. Específicamente, estas reglas pueden ser obtenidas seleccionado los términos negativos y positivos de la derivada parcial de la función divergencia $D(\mathbf{X}|\hat{\mathbf{X}})$ con respecto a los parámetros $\mathbf{B}$ y $\mathbf{A}$, como se muestra a continuación:

$$\mathbf{B} \leftarrow \mathbf{B} \odot \frac{\left[\frac{\partial D\left(\mathbf{X}|\hat{\mathbf{X}}\right)}{\partial \mathbf{B}}\right]^{-}}{\left[\frac{\partial D\left(\mathbf{X}|\hat{\mathbf{X}}\right)}{\partial \mathbf{B}}\right]^{+}} = \mathbf{B} \odot \frac{\left(\hat{\mathbf{X}}^{\beta-2} \odot \mathbf{X}\right)\mathbf{A}^{\mathrm{T}}}{\hat{\mathbf{X}}^{\beta-1}\mathbf{A}^{\mathrm{T}}} \tag{3.9}$$

$$\mathbf{A} \leftarrow \mathbf{A} \odot \frac{\left[\frac{\partial D\left(\mathbf{X}|\hat{\mathbf{X}}\right)}{\partial \mathbf{A}}\right]^{-}}{\left[\frac{\partial D\left(\mathbf{X}|\hat{\mathbf{X}}\right)}{\partial \mathbf{A}}\right]^{+}} = \mathbf{A} \odot \frac{\mathbf{B}^{\mathrm{T}}\left(\mathbf{X} \odot \hat{\mathbf{X}}^{\beta-2}\right)}{\mathbf{B}^{\mathrm{T}}\hat{\mathbf{X}}^{\beta-1}} \tag{3.10}$$

donde $[\ ]^{-}$ y $[\ ]^{+}$ se refieren al conjunto de términos negativos y positivos de la derivada parcial, el símbolo $\odot$ denota al producto matricial elemento a elemento, la división mostrada es elemento a elemento, y el símbolo $^{\mathrm{T}}$ es el operador traspuesto de una matriz. Este procedimiento siempre mantiene la no negatividad de ambas matrices estimadas ($\mathbf{B}$ y $\mathbf{A}$), ya que todos los parámetros que intervienen en el proceso de actualización son no negativos. Además, este esquema de optimización no es decreciente con respecto a la función de coste β-divergencia, y es capaz de alcanzar un mínimo local después de unas pocas iteraciones [107, 185].

Además de las reglas de actualización multiplicativas, se han desarrollado varios enfoques alternativos basados en métodos de segundo orden [341], gradiente proyectado [77], etc. Asimismo, se han propuesto alternativas para mejorar el rendimiento computacional disminuyendo el tiempo de ejecución del algoritmo, como el método de conjunto activo (active-set) de Newton [168, 278, 313]. Este método, puede producir una convergencia significativamente más rápida en comparación con el método MU y EM, especialmente cuando se utiliza un gran número de componentes espectrales K.

Finalmente, algo importante a considerar sobre el modelo NMF estándar es que la única restricción que aplica es la no negatividad de todos los elementos que componen las matrices

del modelo. Esta restricción solo asegura la convergencia a un mínimo local para obtener el espectrograma estimado o reconstruido de la mezcla con el menor error de reconstrucción. Sin embargo, esta reconstrucción no garantiza que la descomposición obtenida este compuesta por objetos basados en partes con significado físico como ocurre en el mundo real (ver Figura 3.2). Es decir, el modelo NMF estándar tiene el objetivo de minimizar el error de reconstrucción, pero no garantiza que los patrones espectrales estimados y almacenados en el diccionario de bases $\mathbf{B}$ representen o modelen fielmente el comportamiento de un determinado sonido que se encuentre activo en la señal factorizada. Incluso, aunque se consiga generar un diccionario de bases $\mathbf{B}$ lo suficientemente fiable, pueden existir múltiples matrices de activaciones $\mathbf{A}$ que produzcan una solución de mínima divergencia para los eventos sonoros existentes en la señal de entrada. Realmente, el modelo NMF estándar puede generar infinitas descomposiciones diferentes que puedan aproximarse al espectrograma de entrada, con un error mínimo, ya que las matrices $\mathbf{B}$ y $\mathbf{A}$ son inicializadas con valores positivos aleatorios. Para resolver esta problemática y conseguir resultados realistas, es necesario incorporar alguna información previa al modelo de descomposición para caracterizar el comportamiento de los sonidos presentes en la señal de entrada. En este sentido la Sección 3.3 describe los diferentes enfoques de factorización NMF y cofactorización NMPCF utilizados para la separación de fuentes sonoras. Como se verá en la Sección 3.3.2, el enfoque NMPCF se considera una extensión del enfoque NMF, sustituyendo el principio de factorización de matrices no negativas por el de cofactorización de matrices no negativas. La Sección 3.4 describe las principales regularizaciones y restricciones que suelen ser añadidas al modelo de descomposición para aportar características tiempo-frecuencia de los sonidos presentes en la mezcla a las matrices de bases y activaciones. Por último, la Sección 3.5 define los principales descriptores utilizados para aplicar clustering a las bases espectrales que componen el diccionario $\mathbf{B}$ en función de su distribución espectral.

3.3. Modelos aplicados a la separación de fuentes sonoras

La separación de las señales de audio en sus fuentes sonoras individuales es probablemente la aplicación más inmediata de los modelos de descomposición matricial [172, 321, 284, 63]. Esencialmente, se supone que cualquier fuente sonora, que compone a la señal mezcla de entrada, tiene sus funciones base (patrones espectrales) característicos. En este sentido, la mezcla se compone entonces de átomos de las fuentes sonoras individuales, de manera que la separación de cualquier fuente presente en la mezcla sólo requiere la segregación de la contribución de sus funciones base.

Con el objetivo de facilitar la explicación de los diferentes enfoques, vamos a considerar que la señal mezcla de entrada se compone únicamente de dos fuentes sonoras. Sin embargo, los enfoques tratados pueden ser extendidos para utilizarlos con mezclas compuestas por más fuentes. Específicamente, vamos a considerar que la señal mezcla de entrada $\mathbf{s}(n)$ está compuestas por dos fuentes sonoras (por ejemplo, la fuente sonora de interés y la fuente sonora interferente), $\mathbf{s}_W(n)$ y $\mathbf{s}_R(n)$, cuya mezcla se supone que es aditiva, es decir, $\mathbf{s}(n) = \mathbf{s}_W(n) + \mathbf{s}_R(n)$. Como

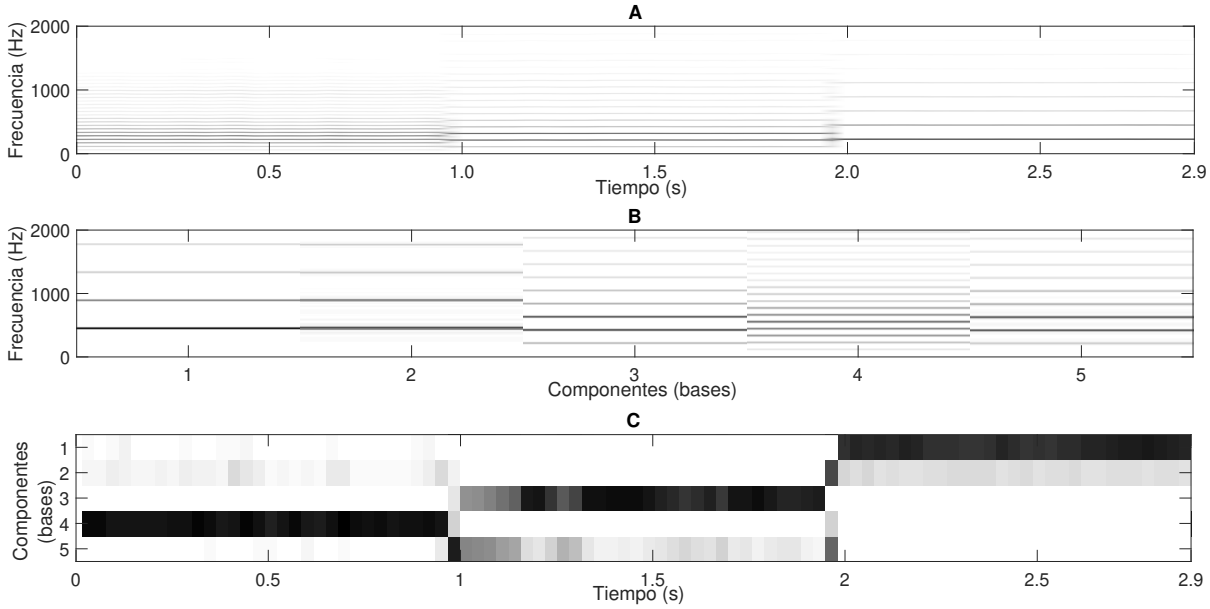

Figura 3.2: Ejemplo de factorización aplicando el modelo NMF estándar, utilizando $K = 5$ componentes, sobre un fragmento de señales musicales monofónicas, con una duración aproximada de 2.9 segundos, compuesto por tres sonidos armónicos con frecuencias fundamentales $f_{01} = 107Hz$, $f_{02} = 217Hz$ y $f_{03} = 440Hz$ tocados por un trombón. El color más oscuro indica mayor amplitud. A) Espectrograma en magnitud de la señal de entrada $\mathbf{X}$; B) Diccionario de bases $\mathbf{B}$; y C) Matriz de activaciones $\mathbf{A}$. Figura extraída de la referencia [65].

se ha descrito anteriormente, NMF factoriza el espectrograma de entrada $\mathbf{X}$ en el producto de dos matrices no negativas: matriz de bases $\mathbf{B}$ y matriz de activaciones $\mathbf{A}$. Ajustando el problema a la dinámica de separación de fuentes, el espectrograma mezcla $\mathbf{X}$ puede expresarse aproximadamente como la aditividad lineal entre los espectrograma en magnitud de las dos fuentes presentes: $\mathbf{X}_W \in \mathbb{R}_+^{F\times T}$ y $\mathbf{X}_R \in \mathbb{R}_+^{F\times T}$. Asumiendo esto, los espectrogramas estimados $\hat{\mathbf{X}} \in \mathbb{R}_+^{F\times T}$, $\hat{\mathbf{X}}_W \in \mathbb{R}_+^{F\times T}$ y $\hat{\mathbf{X}}_R \in \mathbb{R}_+^{F\times T}$ pueden ser obtenidos como se muestra a continuación:

$$\begin{aligned}\mathbf{X} = \mathbf{X}_W + \mathbf{X}_R \approx \hat{\mathbf{X}} = \hat{\mathbf{X}}_W + \hat{\mathbf{X}}_R = \mathbf{BA} = \begin{bmatrix}\mathbf{B}_W & \mathbf{B}_R\end{bmatrix}\begin{bmatrix}\mathbf{A}_W \\ \mathbf{A}_R\end{bmatrix} \\ = \mathbf{B}_W\mathbf{A}_W + \mathbf{B}_R\mathbf{A}_R\end{aligned} \tag{3.11}$$

donde los subíndices W y R se utilizan para diferenciar las dos fuentes presentes. Por un lado, $\mathbf{B}_W \in \mathbb{R}_+^{F\times K_W}$ y $\mathbf{A}_W \in \mathbb{R}_+^{K_W\times T}$, son las matrices de bases y de activaciones de la fuente denotada como W. Mientras que, $\mathbf{B}_R \in \mathbb{R}_+^{F\times K_R}$ y $\mathbf{A}_R \in \mathbb{R}_+^{K_R\times T}$, son las matrices de bases y de activaciones de la fuente denotada como R. El número de componentes para cada fuente es denotado como K_W y K_R. Una representación del enfoque NMF descrito en la Ec. (3.11), para la separación de dos fuentes sonoras, es mostrada en la Figura 3.3.

Tomando como referencia el modelo descrito anteriormente, a continuación se describen los diferentes enfoques de descomposición matricial (NMF/NMPCF) aplicados a la separación de fuentes sonoras. Como veremos ambos enfoques se basan en los mismos principios de des-

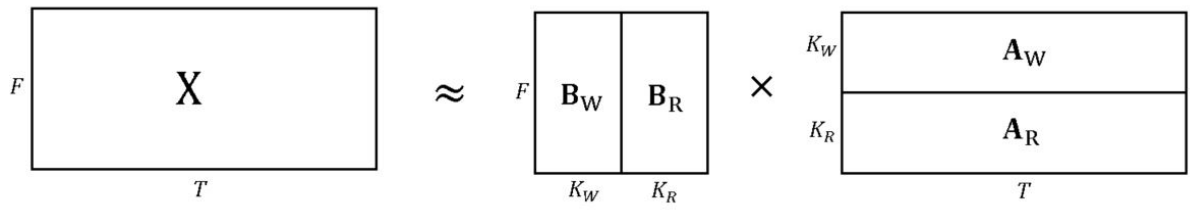

Figura 3.3: Ilustración del modelo de descomposición NMF estándar aplicado a la separación de dos fuentes sonoras, denotadas como W y R.

composición de matrices no negativas. La principal diferencia entre los enfoques NMF y los NMPCF es: (i) los modelos basados en NMF realizan la factorización de una única matriz (como se puede ver en la Figura 3.3); sin embargo, (ii) los modelos basados en NMPCF realizan la factorización conjunta de varias matrices al mismo tiempo. Este concepto es conocido como cofactorización y el enfoque NMPCF es considerado una extensión del enfoque NMF.

3.3.1. Modelos basados en el enfoque NMF

Como se ha mencionado anteriormente los enfoques NMF se basan en la factorización de una única matriz. Siguiendo el modelo de descomposición descrito en la Ec. (3.11), para la separación de dos fuentes sonoras, la factorización del espectrograma de entrada $\mathbf{X}$ en el producto matricial $\begin{bmatrix}\mathbf{B}_W & \mathbf{B}_R\end{bmatrix}\begin{bmatrix}\mathbf{A}_W \\ \mathbf{A}_R\end{bmatrix}$ puede ser obtenida minimizando una función objetivo de divergencia, al igual que en el modelo NMF estándar (ver Ec. (3.3)), expresada como,

$$\underset{\mathbf{B}_W,\mathbf{B}_R,\mathbf{A}_W,\mathbf{A}_R}{\arg\min} D(\mathbf{X}|\begin{bmatrix}\mathbf{B}_W & \mathbf{B}_R\end{bmatrix}\begin{bmatrix}\mathbf{A}_W \\ \mathbf{A}_R\end{bmatrix}) \quad \mathbf{B}_W,\mathbf{B}_R,\mathbf{A}_W,\mathbf{A}_R \geq 0 \tag{3.12}$$

En este sentido, siguiendo las misma mecánica descrita en la sección anterior, las matrices de bases ($\mathbf{B}_W$, $\mathbf{B}_R$) y de activaciones ($\mathbf{A}_W$, $\mathbf{A}_R$), correspondientes a cada fuente sonora (W, R) de la mezcla, pueden ser estimadas aplicando un algoritmo de gradiente descendiente basado en las denominadas reglas de actualización multiplicativas (ver Ec. (3.9) y Ec. (3.10)). Suponiendo la utilización de la función de divergencia KL ($\beta = 1$) para medir el error de reconstrucción entre el espectrograma original $\mathbf{X}$ y estimado $\hat{\mathbf{X}}$, las reglas multiplicativas de actualización pueden ser expresadas como,

$$\mathbf{B}_z \leftarrow \mathbf{B}_z \odot \left(\Big(\mathbf{X} \oslash \mathbf{BA}\Big){\mathbf{A}_z}^{\mathrm{T}} \oslash \Big(\mathbf{1}\mathbf{A}_z^{\mathrm{T}}\Big)\right), \quad z = W, R \tag{3.13}$$

$$\mathbf{A}_z \leftarrow \mathbf{A}_z \odot \left({\mathbf{B}_z}^{\mathrm{T}}\Big(\mathbf{X} \oslash \mathbf{BA}\Big) \oslash \Big({\mathbf{B}_z}^{\mathrm{T}}\mathbf{1}\Big)\right), \quad z = W, R \tag{3.14}$$

donde $\mathbf{1} \in \mathbb{R}_+^{F\times T}$ representa una matriz compuesta por todo unos, el símbolo $\odot$ denota al producto matricial elemento a elemento, el símbolo $\oslash$ denota la división matricial elemento a

elemento, y el símbolo $^{\mathrm{T}}$ es el operador traspuesto de una matriz. De la misma forma que en el modelo NMF estándar, las matrices de bases y activaciones se inicializan con números positivos aleatorios, y posteriormente se dejan evolucionar las reglas definidas hasta la convergencia del algoritmo.

En general las técnicas de separación de fuentes, basadas en NMF, suelen ser clasificadas en tres categorías, considerando el entrenamiento previo que se realiza para aprender las matrices de bases ($\mathbf{B}_W$, $\mathbf{B}_R$) de las diferentes fuentes presentes en la mezcla: no supervisado, semi-supervisado y supervisado.

Enfoque NMF no supervisado (ciego)

Los modelos de separación basados en un enfoque NMF ciego no necesitan ninguna etapa previa de entrenamiento para aprender las bases de los sonidos presentes en la mezcla. Por lo tanto, en el proceso de actualización de las reglas multiplicativas se dejarán evolucionar todas las matrices de bases y activaciones. Sin embargo, este tipo de enfoques necesitan técnicas adicionales para conseguir separar las fuentes presentes en la mezcla, ya que el modelo NMF ciego no garantiza que la descomposición obtenida esté compuesta por objetos basados en partes con significado físico como ocurre en la vida real. En otras palabras, el modelo NMF ciego carece de información para caracterizar los sonidos presentes en la mezcla. Por ello, es habitual que los métodos de separación basados en este enfoque utilicen regularizaciones o restricciones (ver Sección 3.4) que permitan modelar el comportamiento tiempo-frecuencia de las fuentes sonoras [202, 63, 64]. Estos modelos suelen ser denominados NMF regularizados, denotados en inglés como Constrained NMF (CNMF), y su principal ventaja es su robustez si se consigue modelar correctamente las fuentes sonoras a separar al no depender de ninguna base de datos de entrenamiento. Por lo tanto, este tipo de modelos son interesantes de utilizar en aquellos ámbitos donde exista escasez de bases de datos para ser aplicadas en etapas previas de entrenamiento (hecho habitual en el ámbito científico del procesado de señales sonoras respiratorias). Entre los trabajos publicados en esta Tesis, cinco de ellos [P1][P2][P3][P5][P6] están basados en un enfoque NMF regularizado.

Enfoque NMF semi-supervisado

Por otro lado, los modelos de separación basados en un enfoque NMF semi-supervisado [284, 96, 186, 61, 181] son utilizados cuando se dispone de bases de datos fiables para el entrenamiento de alguna de las matrices de bases $\mathbf{B}_W$ o $\mathbf{B}_R$. Es crucial que esas bases de entrenamiento únicamente estén compuestas por los sonidos aislados que se quieren entrenar, sin interferencia de otras fuentes sonoras presentes en la mezcla. En los enfoques NMF semi-supervisados, solo una de las matrices de bases es entrenada. Suponiendo que la matriz de bases $\mathbf{B}_R$ ha sido previamente aprendida (en una etapa anterior), la función objetivo de divergencia se puede expresar, como,

$$\underset{\mathbf{B}_W,\mathbf{A}_W,\mathbf{A}_R}{\arg\min}\ D(\mathbf{X}|\begin{bmatrix}\mathbf{B}_W & \mathbf{B}_R\end{bmatrix}\begin{bmatrix}\mathbf{A}_W\\ \mathbf{A}_R\end{bmatrix}) \quad \mathbf{B}_W, \mathbf{A}_W, \mathbf{A}_R \geq 0 \tag{3.15}$$

Por lo tanto, la matriz de bases $\mathbf{B}_R$, que previamente había sido aprendida, permanecerá fija durante el proceso de actualización. Sin embargo, las matrices $\mathbf{B}_W$, $\mathbf{A}_W$ y $\mathbf{A}_R$ son inicializadas con valores aleatorios positivos, y actualizadas mediante las reglas de actualización multiplicativas (ver Ec. (3.13) y Ec. (3.14)) hasta la convergencia del algoritmo.

Enfoque NMF supervisado

Por último los modelos de separación basados en un enfoque NMF supervisado [321, 76, 311, 282] son utilizados cuando se dispone de bases de datos fiables para el entrenamiento de todas las matrices de bases $\mathbf{B}_W$ y $\mathbf{B}_R$. En los enfoques NMF supervisados, es necesario realizar un entrenamiento a priori de todas las fuentes sonoras presentes en la mezcla, para aprender las matrices de bases que almacenan sus patrones espectrales. En este caso, la función objetivo de divergencia se puede expresar, como,

$$\underset{\mathbf{A}_W,\mathbf{A}_R}{\arg\min}\, D(\mathbf{X}|\begin{bmatrix}\mathbf{B}_W & \mathbf{B}_R\end{bmatrix}\begin{bmatrix}\mathbf{A}_W \\ \mathbf{A}_R\end{bmatrix}) \quad \mathbf{A}_W, \mathbf{A}_R \geq 0 \tag{3.16}$$

Por lo tanto, las matrices de bases $\mathbf{B}_W$ y $\mathbf{B}_R$, que previamente habían sido aprendidas, permanecerán fijas durante el proceso de actualización. Sin embargo, las matrices de activaciones $\mathbf{A}_W$ y $\mathbf{A}_R$ son inicializadas con valores aleatorios positivos, y actualizadas mediante las reglas de actualización multiplicativas (ver Ec. (3.14)) hasta la convergencia del algoritmo.

Como se ha mencionado, tanto el enfoque NMF semi-supervisado, como el supervisado, necesitan una etapa previa para realizar el entrenamiento de alguna o todas las matrices de bases que componen al modelo de separación. Actualmente, existen diversas técnicas para realizar el aprendizaje de un diccionario $\mathbf{B}$ a partir de bases de entrenamiento compuestas por sonidos aislados de la fuente de interés [311, 313, 285, 282, 126, 272]. Principalmente existen dos enfoques para el aprendizaje de diccionarios: (i) el primer enfoque, "decomposition based learning", trata de aprender las bases del diccionario mediante la factorización de los datos de entrenamiento, en los cuales la fuente sonora a entrenar está aislada; (ii) el segundo enfoque, "exemplar-based approach", utiliza muestras de los datos de entrenamiento para construir el diccionario, sin realizar ninguna factorización.

A diferencia del modelo NMF ciego, los modelos semi-supervisado y supervisado, cuentan con información relevante de las características de alguna o todas las fuentes sonoras presentes en la mezcla, ya que se ha realizado un entrenamiento previo para aprender los patrones espectrales de algunas o todas las fuentes. Sin embargo, el rendimiento de separación de estos modelos está fuertemente ligado a la disposición de diccionarios adecuados para cada fuente. Por lo tanto, los resultados de separación de los algoritmos basados en un enfoque NMF semi-supervisado o supervisado, dependerán de la fiabilidad de las bases de datos de entrenamiento ya que dichas bases de datos deberán ser lo suficientemente robustas, en el sentido de que contengan una diversidad significativa de señales, para que las características espectro-temporales extraídas puedan ser generalizadas para poder ser encontradas en la mayoría de fuentes sonoras del tipo que se pretende aprender.

3.3.2. Modelos basados en el enfoque NMPCF

La cofactorización parcial de matrices no negativas, expresada en inglés como Non-negative Matrix Partial Co-Factorization (NMPCF), es considerada otro tipo de enfoque utilizado en la separación de fuentes sonoras [338, 170, 147, 169]. La idea principal en la que se basan este tipo de enfoques, es realizar una factorización conjunta (cofactorización) usando varias matrices de entrada para obtener un conjunto de bases espectrales compartidas (activas simultáneamente en dichas matrices). Por lo tanto, los enfoques NMPCF, a diferencia de los enfoques NMF utilizan varias matrices de entrada en el proceso de descomposición matricial.

A continuación se muestra una clasificación general de los principales enfoques NMPCF encontrados en la literatura para la separación de fuentes sonoras, considerando las matrices que intervienen en el proceso de cofactorización:

Enfoque NMPCF semi-supervisado

Partiendo del modelo de separación descrito en la Ec. (3.11), el cual tiene como objetivo separar la fuente sonora W y R, el enfoque NMPCF semi-supervisado [338] utiliza una matriz adicional $\mathbf{Y} \in \mathbb{R}_{+}^{F \times T}$ que consiste en un espectrograma compuesto solo por sonidos de una de las fuentes que componen a la mezcla (denominado espectrograma de entrenamiento). Por ejemplo, si consideramos que el espectrograma de entrenamiento $\mathbf{Y}$ solo contiene sonidos de la fuente R, este puede ser descompuesto o estimado como,

$$\mathbf{Y} = \hat{\mathbf{Y}} \approx \mathbf{B}_R \mathbf{H}_R \tag{3.17}$$

donde $\hat{\mathbf{Y}} \in \mathbb{R}_{+}^{F \times T}$ es la matriz estimada del espectrograma de entrenamiento para la fuente R, la matriz de bases $\mathbf{B}_R$ puede ser tratada como la misma matriz $\mathbf{B}_R$ del modelo descrito en la Ec. (3.11), y la matriz de activaciones $\mathbf{H}_R \in \mathbb{R}_{+}^{K_R \times T}$ representa las activaciones temporales de cada base del diccionario $\mathbf{B}_R$ en el espectrograma de entrenamiento $\mathbf{Y}$.

Considerando, el modelo correspondiente a la matriz mezcla de entrada $\mathbf{X}$ (ver Ec. (3.11)) y a la matriz de entrenamiento $\mathbf{Y}$ (ver Ec. (3.17)), el enfoque NMPCF semi-supervisado realiza la factorización conjunta del espectrograma $\mathbf{X}$ e $\mathbf{Y}$ compartiendo la matriz de bases $\mathbf{B}_R$ (ver Figura 3.4). Por lo tanto, las características espectrales de la fuente sonora R estarán recogidas en la matriz $\mathbf{B}_R$, mientras que la matriz $\mathbf{B}_W$ representara la parte restante de los sonidos incluidos en el espectrograma de entrada $\mathbf{X}$, los cuales se suponen que pertenecen a la fuente W.

Como ocurre con los enfoques NMF, la factorización conjunta (cofactorización) de ambos espectrogramas $\mathbf{X}$ e $\mathbf{Y}$, puede ser obtenida minimizando una función objetivo de divergencia. En este caso, la función objetivo estará compuesta por la suma de dos funciones de divergencia para minimizar el error de reconstrucción en ambos espectrogramas, como se muestra a continuación:

$$\underset{\mathbf{B}_W, \mathbf{B}_R, \mathbf{A}_W, \mathbf{A}_R, \mathbf{H}_R}{\arg\min} \quad D(\mathbf{X}|\hat{\mathbf{X}}) + \lambda_R D(\mathbf{Y}|\hat{\mathbf{Y}}) \quad \mathbf{B}_W, \mathbf{B}_R, \mathbf{A}_W, \mathbf{A}_R, \mathbf{H}_R \geq 0 \tag{3.18}$$

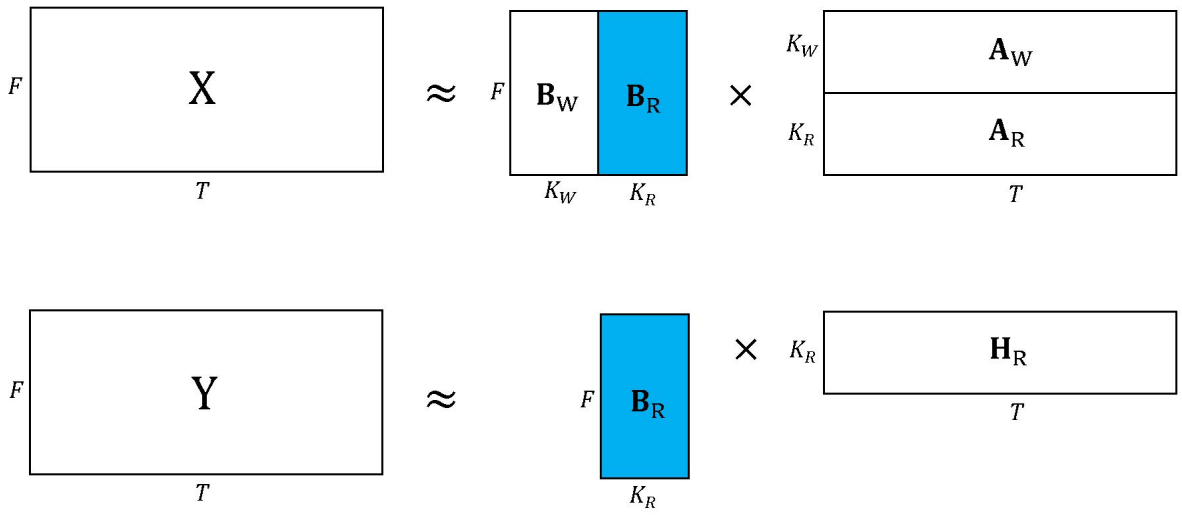

Figura 3.4: Ilustración del modelo de descomposición NMPCF semi-supervisado aplicado a la separación de dos fuentes sonoras, denotadas como W y R. El espectrograma de entrenamiento $\mathbf{Y}$ está compuesto por sonidos del tipo R.

donde el parámetro λ_R controla la importancia relativa entre los espectrogramas $\mathbf{X}$ e $\mathbf{Y}$ para estimar el diccionario $\mathbf{B}_R$. En este sentido, cuanto mayor sea el valor de λ_R mayor será la contribución del espectrograma $\mathbf{Y}$ en la cofactorización de las bases $\mathbf{B}_R$. En los enfoques NMPCF también es habitual utilizar las denominadas reglas de actualización multiplicativas para estimar las matrices de bases $\mathbf{B}_R$, $\mathbf{B}_W$ y activaciones $\mathbf{A}_R$, $\mathbf{A}_W$, $\mathbf{H}_R$ que intervienen en el proceso. Suponiendo el uso de la función divergencia KL para medir el error de reconstrucción entre los espectrogramas originales $\mathbf{X}$, $\mathbf{Y}$ y estimados $\hat{\mathbf{X}}$, $\hat{\mathbf{Y}}$, las reglas de actualización pueden ser expresadas como,

$$\mathbf{B}_R \leftarrow \mathbf{B}_R \odot \frac{\left(\mathbf{X} \oslash \hat{\mathbf{X}}\right)\left({\mathbf{A}_R}^{\mathrm{T}}\right) + \lambda_R \left(\mathbf{Y} \oslash \hat{\mathbf{Y}}\right)\left({\mathbf{H}_R}^{\mathrm{T}}\right)}{\left(\mathbf{1}{\mathbf{A}_R}^{\mathrm{T}}\right) + \lambda_R \left(\mathbf{1}{\mathbf{H}_R}^{\mathrm{T}}\right)} \tag{3.19}$$

$$\mathbf{B}_W \leftarrow \mathbf{B}_W \odot \frac{\left(\mathbf{X} \oslash \hat{\mathbf{X}}\right)\left({\mathbf{A}_W}^{\mathrm{T}}\right)}{\left(\mathbf{1}{\mathbf{A}_W}^{\mathrm{T}}\right)} \tag{3.20}$$

$$\mathbf{A}_R \leftarrow \mathbf{A}_R \odot \frac{\left({\mathbf{B}_R}^{\mathrm{T}}\right)\left(\mathbf{X} \oslash \hat{\mathbf{X}}\right)}{\left({\mathbf{B}_R}^{\mathrm{T}}\mathbf{1}\right)} \tag{3.21}$$

$$\mathbf{A}_W \leftarrow \mathbf{A}_W \odot \frac{\left({\mathbf{B}_W}^{\mathrm{T}}\right)\left(\mathbf{X} \oslash \hat{\mathbf{X}}\right)}{\left({\mathbf{B}_W}^{\mathrm{T}}\mathbf{1}\right)} \tag{3.22}$$

$$\mathbf{H}_R \leftarrow \mathbf{H}_R \odot \frac{\left({\mathbf{B}_R}^{\mathrm{T}}\right)\left(\mathbf{Y} \oslash \hat{\mathbf{Y}}\right)}{\left({\mathbf{B}_R}^{\mathrm{T}}\mathbf{1}\right)} \tag{3.23}$$

donde, al igual que en los modelos NMF, las matrices son inicializadas con números aleatorios positivos y se dejan evolucionar hasta la convergencia del algoritmo. Como puede comprobar las reglas de actualización para las matrices $\mathbf{B}_W$, $\mathbf{A}_R$, $\mathbf{A}_W$ y $\mathbf{H}_R$ son las mismas que las del enfoque de separación NMF (ver Ec. (3.13) y Ec. (3.14)), porque estas matrices no son compartidas en el proceso de cofactorización. Por lo tanto, el enfoque NMPCF semi-supervisado realiza la cofactorización de ambos espectrogramas para modelar las características espectrales de los sonidos de la fuente R.

En relación a los trabajos publicados durante la Tesis, la publicación [P7] propone adaptar el enfoque NMPCF semi-supervisado a un escenario multicanal compuesto por dos canales de entrada monocanal que capturan audio simultáneamente. El primer canal captura el audio de un estetoscopio, el cual se compone de sonidos biomédicos y ruido ambiental que los interfiere. Por otro lado, el segundo canal captura el ruido ambiental que rodea al paciente, mediante un micrófono. El objetivo de esta publicación es mejorar la calidad de los sonidos biomédicos capturados por el estetoscopio, eliminando el ruido ambiente que rodea al paciente.

Enfoque NMPCF supervisado

Partiendo del enfoque NMPCF semi-supervisado descrito anteriormente, el enfoque NMPCF supervisado [147] utiliza un espectrograma de entrenamiento adicional $\mathbf{Z} \in \mathbb{R}_+^{F\times T}$ compuesto solo por sonidos de la fuente W, el cual puede ser descompuesto o estimado como,

$$\mathbf{Z} = \hat{\mathbf{Z}} \approx \mathbf{B}_W\mathbf{H}_W \tag{3.24}$$

donde $\hat{\mathbf{Z}} \in \mathbb{R}_+^{F\times T}$ es la matriz estimada del espectrograma de entrenamiento para la fuente W, la matriz de bases $\mathbf{B}_W$ puede ser tratada como la misma matriz $\mathbf{B}_W$ del modelo descrito en la Ec. (3.11), y la matriz de activaciones $\mathbf{H}_W \in \mathbb{R}_+^{K_W\times T}$ representa las activaciones temporales de cada base del diccionario $\mathbf{B}_W$ en el espectrograma de entrenamiento $\mathbf{Z}$.

Específicamente, el enfoque supervisado realiza el proceso de factorización conjunta considerando los modelos correspondientes al: espectrograma mezcla de entrada $\mathbf{X}$ (ver Ec. 3.11), espectrograma de entrenamiento $\mathbf{Y}$ para la fuente R (ver Ec. 3.17) y espectrograma de entrenamiento $\mathbf{Z}$ para la fuente W (ver Ec. 3.24). De forma más precisa, este enfoque realiza una doble cofactorización. Por un lado, realiza la factorización conjunta del espectrograma $\mathbf{X}$ e $\mathbf{Y}$ compartiendo la matriz de bases $\mathbf{B}_R$, y por otro lado, realiza la factorización conjunta del espectrograma $\mathbf{X}$ y $\mathbf{Z}$ compartiendo la matriz de bases $\mathbf{B}_W$ (ver Figura 3.5). Por lo tanto, se utiliza información adicional para modelar el comportamiento de ambas fuentes presentes en la mezcla.

En este caso, la función objetivo estará compuesta por la suma de tres funciones de divergencia para minimizar el error de reconstrucción en los tres espectrogramas que intervienen en el proceso, como se muestra a continuación:

$$\begin{aligned}\underset{\mathbf{B}_W,\mathbf{B}_R,\mathbf{A}_W,\mathbf{A}_R,\mathbf{H}_R,\mathbf{H}_W}{\arg\min} \quad & D(\mathbf{X}|\hat{\mathbf{X}}) + \lambda_R D(\mathbf{Y}|\hat{\mathbf{Y}}) + \lambda_W D(\mathbf{Z}|\hat{\mathbf{Z}}) \\ & \mathbf{B}_W, \mathbf{B}_R, \mathbf{A}_W, \mathbf{A}_R, \mathbf{H}_R, \mathbf{H}_W \geq 0\end{aligned} \tag{3.25}$$

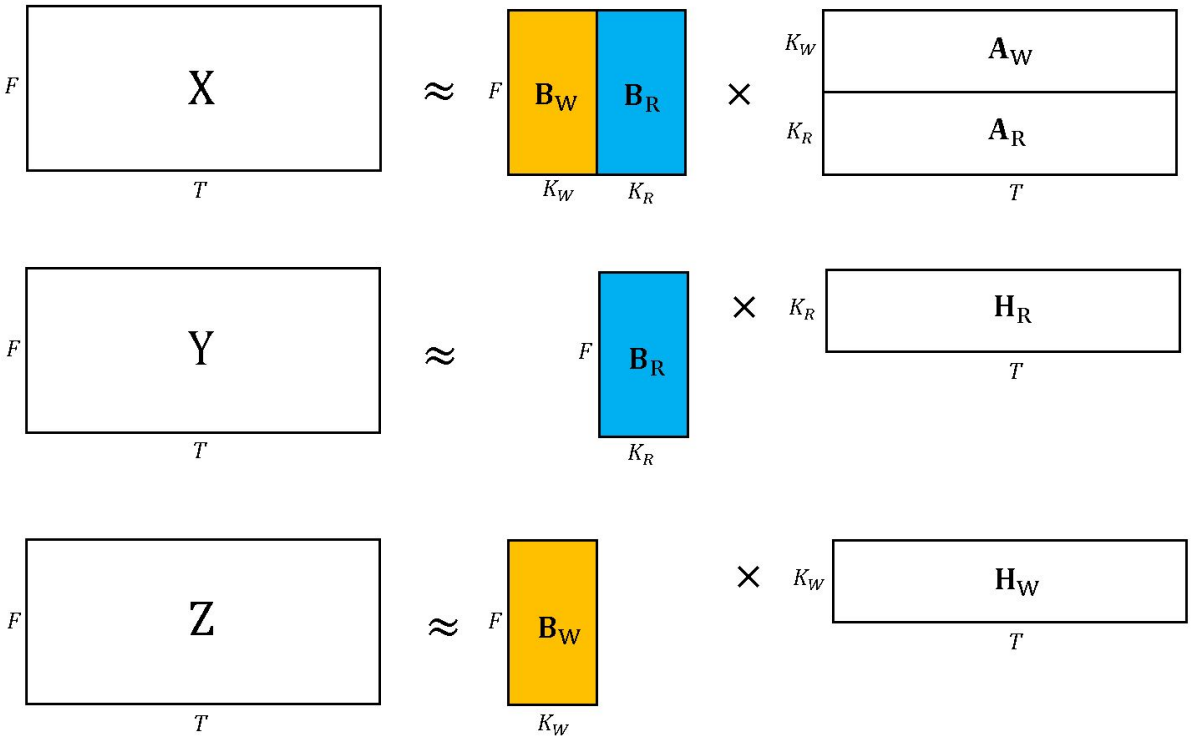

Figura 3.5: Ilustración del modelo de descomposición NMPCF supervisado aplicado a la separación de dos fuentes sonoras, denotadas como W y R. El espectrograma de entrenamiento $\mathbf{Y}$ está compuesto por sonidos del tipo R, mientras que el espectrograma de entrenamiento $\mathbf{Z}$ está compuesto por sonidos del tipo W.

donde el parámetro λ_W controla la importancia relativa entre los espectrogramas $\mathbf{X}$ y $\mathbf{Z}$ para estimar el diccionario $\mathbf{B}_W$. Al igual que en el caso anterior, si se utilizada la función divergencia KL para medir el error de reconstrucción entre los espectrogramas originales $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ y estimados $\hat{\mathbf{X}}$, $\hat{\mathbf{Y}}$, $\hat{\mathbf{Z}}$, las reglas de actualización pueden ser expresadas como,

$$\mathbf{B}_R \leftarrow \mathbf{B}_R \odot \frac{\left(\mathbf{X} \oslash \hat{\mathbf{X}}\right)\left(\mathbf{A}_R{}^{\mathrm{T}}\right) + \lambda_R \left(\mathbf{Y} \oslash \hat{\mathbf{Y}}\right)\left(\mathbf{H}_R{}^{\mathrm{T}}\right)}{\left(\mathbf{1}\mathbf{A}_R{}^{\mathrm{T}}\right) + \lambda_R \left(\mathbf{1}\mathbf{H}_R{}^{\mathrm{T}}\right)} \tag{3.26}$$

$$\mathbf{B}_W \leftarrow \mathbf{B}_W \odot \frac{\left(\mathbf{X} \oslash \hat{\mathbf{X}}\right)\left(\mathbf{A}_W{}^{\mathrm{T}}\right) + \lambda_W \left(\mathbf{Z} \oslash \hat{\mathbf{Z}}\right)\left(\mathbf{H}_W{}^{\mathrm{T}}\right)}{\left(\mathbf{1}\mathbf{A}_W{}^{\mathrm{T}}\right) + \lambda_W \left(\mathbf{1}\mathbf{H}_W{}^{\mathrm{T}}\right)} \tag{3.27}$$

$$\mathbf{A}_R \leftarrow \mathbf{A}_R \odot \frac{\left(\mathbf{B}_R{}^{\mathrm{T}}\right)\left(\mathbf{X} \oslash \hat{\mathbf{X}}\right)}{\left(\mathbf{B}_R{}^{\mathrm{T}}\mathbf{1}\right)} \tag{3.28}$$

$$\mathbf{A}_W \leftarrow \mathbf{A}_W \odot \frac{\left(\mathbf{B}_W{}^{\mathrm{T}}\right)\left(\mathbf{X} \oslash \hat{\mathbf{X}}\right)}{\left(\mathbf{B}_W{}^{\mathrm{T}}\mathbf{1}\right)} \tag{3.29}$$

$$\mathbf{H}_R \leftarrow \mathbf{H}_R \odot \frac{\left(\mathbf{B}_R{}^{\mathrm{T}}\right)\left(\mathbf{Y} \oslash \hat{\mathbf{Y}}\right)}{\left(\mathbf{B}_R{}^{\mathrm{T}}\mathbf{1}\right)} \tag{3.30}$$

$$\mathbf{H}_W \leftarrow \mathbf{H}_W \odot \frac{\left(\mathbf{B}_W{}^{\mathrm{T}}\right)\left(\mathbf{Z} \oslash \hat{\mathbf{Z}}\right)}{\left(\mathbf{B}_W{}^{\mathrm{T}}\mathbf{1}\right)} \tag{3.31}$$

Al igual que en el caso anterior, las reglas de actualización para las matrices que no intervienen en la cofactorización $\mathbf{A}_R$, $\mathbf{A}_W$, $\mathbf{H}_R$ y $\mathbf{H}_W$ son las mismas que las del enfoque de separación NMF (ver Ec. (3.14)). Por lo tanto, el enfoque NMPCF supervisado, por un lado, realiza la cofactorización de los espectrogramas $\mathbf{X}$ e $\mathbf{Y}$ para modelar las características espectrales de la fuente R que se repiten en ambos espectrogramas, y por otro lado, la cofactorización de los espectrogramas $\mathbf{X}$ e $\mathbf{Z}$ para modelar las características espectrales de la fuente W que se repiten en ambos espectrogramas.

La principal diferencia que existe entre los enfoques semi-supervisado y supervisado del modelo NMPCF, con respecto a los del modelo NMF se muestra a continuación: (i) los enfoques NMF utilizan un diccionario fijo, el cual ha sido obtenido en una etapa previa, para modelar el comportamiento de los sonidos presentes en la mezcla; (ii) sin embargo, los enfoques NMPCF, no necesitan una etapa de entrenamiento previa, ya que se aplica una factorización matricial conjunta entre el espectrograma mezcla de entrada y los espectrogramas de entrenamiento, con el fin de obtener un diccionario de bases dinámico para cada fuente sonora. Por lo tanto, en el enfoque NMF las bases obtenidas de la etapa de entrenamiento podrían ser muy diferentes a aquellas bases activas en la señal de entrada (gran dependencia de la base de datos de entrenamiento), mientras que en el enfoque NMPCF las bases obtenidas de las señales de entrenamiento no son fijas, sino que se adaptan dinámicamente a las bases factorizadas de la señal de entrada al ser una cofactorización simultánea entre ambos espectrogramas (entrenamiento y entrada).

Enfoque NMPCF no supervisado (ciego)

Este enfoque es denominado no supervisado o ciego porque no necesita ninguna matriz adicional para modelar los sonidos presentes en el espectrograma mezcla de entrada $\mathbf{X}$. Los modelos basados en este enfoque pueden ser utilizados para la separación de fuentes sonoras, cuando una de las fuentes presenta patrones espectrales que se repiten a lo largo del tiempo (por ejemplo: la señal rítmica de los tambores, los latidos del corazón o incluso los sonidos respiratorios normales) [169]. En este sentido, el espectrograma mezcla de entrada $\mathbf{X}$ es dividido en L segmentos $\mathbf{X}^{(1)}$,$\mathbf{X}^{(2)}$, $\cdots$,$\mathbf{X}^{(L)}$. El objetivo de este enfoque es realizar una factorización conjunta de todos los segmentos para extraer aquellos patrones espectrales que se encuentran activos a lo largo del tiempo (repetición temporal) en los segmentos analizados. Si consideramos que la fuente sonora R es la que presenta un comportamiento repetitivo, el modelo de separación basado en este enfoque se expresa como,

$$\mathbf{X}^{(l)} \approx \hat{\mathbf{X}}^{(l)} = \hat{\mathbf{X}}_R^{(l)} + \hat{\mathbf{X}}_W^{(l)} = \mathbf{B}_R\mathbf{A}_R^{(l)} + \mathbf{B}_W^{(l)}\mathbf{A}_W^{(l)} \tag{3.32}$$

donde el término $^{(l)}$ permite identificar al segmento $l = 1, ..., L$ del espectrograma mezcla de entrada $\mathbf{X}$.

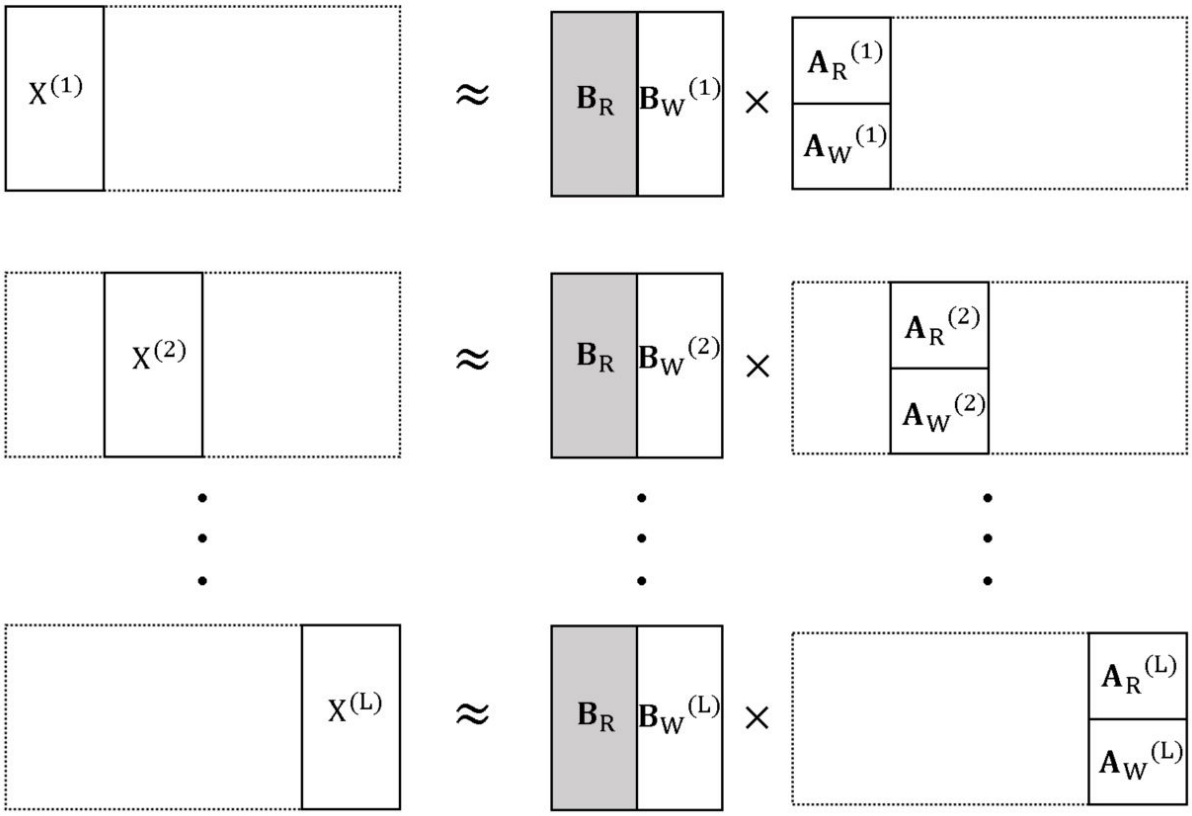

Figura 3.6: Ilustración del modelo de descomposición NMPCF no supervisado (ciego) aplicado a la separación de dos fuentes sonoras, denotadas como W y R. Este modelo realiza la cofactorización de los L segmentos en los que el espectrograma de entrada $\mathbf{X}$ ha sido dividido.

Específicamente, este enfoque realiza la factorización conjunta entre todos los segmentos $\mathbf{X}^{(1)}$,$\mathbf{X}^{(2)}$, $\cdots$,$\mathbf{X}^{(L)}$ que componen al espectrograma $\mathbf{X}$, compartiendo la matriz de bases $\mathbf{B}_R$ (ver Figura 3.6). Por lo tanto los patrones espectrales repetitivos de la fuente R estarán recogidos en la matriz $\mathbf{B}_R$, mientras que las matrices $\mathbf{B}_W^{(l)}$ representarán los sonidos restantes en cada segmento l. En este caso, la función objetivo estará compuesta por la suma de la función divergencia correspondiente a cada segmento (L segmentos en total), como se muestra a continuación:

$$\underset{\mathbf{B}_R,\mathbf{B}_W^{(l)},\mathbf{A}_W^{(l)},\mathbf{A}_R^{(l)}}{\arg\min} \sum_{l=1}^{L} D(\mathbf{X}^{(l)}|\hat{\mathbf{X}}^{(l)}) \quad \mathbf{B}_R, \mathbf{B}_W^{(l)}, \mathbf{A}_R^{(l)}, \mathbf{A}_W^{(l)} \geq 0 \tag{3.33}$$

donde la importancia relativa entre los L segmentos para estimar el diccionario $\mathbf{B}_R$ es la misma independientemente del segmento. Si se utilizada la función divergencia KL para medir el error de reconstrucción entre los espectrogramas originales $\mathbf{X}^{(l)}$ y estimados $\hat{\mathbf{X}}^{(l)}$, las reglas de

actualización pueden ser expresadas como,

$$\mathbf{B}_R \leftarrow \mathbf{B}_R \odot \frac{\sum_{l=1}^{L} \left[\left(\mathbf{X}^{(l)} \oslash \hat{\mathbf{X}}^{(l)} \right) \left(\mathbf{A}_R^{(l)} \right)^{\mathrm{T}} \right]}{\sum_{l=1}^{L} \left[\mathbf{1} \left(\mathbf{A}_R^{(l)} \right)^{\mathrm{T}} \right]} \tag{3.34}$$

$$\mathbf{B}_W^{(l)} \leftarrow \mathbf{B}_W^{(l)} \odot \frac{\left(\mathbf{X}^{(l)} \oslash \hat{\mathbf{X}}^{(l)} \right) \left(\mathbf{A}_W^{(l)} \right)^{\mathrm{T}}}{\mathbf{1} \left(\mathbf{A}_W^{(l)} \right)^{\mathrm{T}}} \tag{3.35}$$

$$\mathbf{A}_R^{(l)} \leftarrow \mathbf{A}_R^{(l)} \odot \frac{\left(\mathbf{B}_R^{(l)} \right)^{\mathrm{T}} \left(\mathbf{X}^{(l)} \oslash \hat{\mathbf{X}}^{(l)} \right)}{\left(\mathbf{B}_R^{(l)} \right)^{\mathrm{T}} \mathbf{1}} \tag{3.36}$$

$$\mathbf{A}_W^{(l)} \leftarrow \mathbf{A}_W^{(l)} \odot \frac{\left(\mathbf{B}_R^{(l)} \right)^{\mathrm{T}} \left(\mathbf{X}^{(l)} \oslash \hat{\mathbf{X}}^{(l)} \right)}{\left(\mathbf{B}_W^{(l)} \right)^{\mathrm{T}} \mathbf{1}} \tag{3.37}$$

como en cualquier enfoque basado en NMF, las matrices son inicializadas con números aleatorios positivos y se dejan evolucionar hasta la convergencia del algoritmo. En conclusión, este enfoque se basa en la repetitividad temporal de una fuente sonora, para obtener los patrones espectrales que se repiten a lo largo de los diferentes segmentos que componen a la señal mezcla de entrada. Al ser un enfoque ciego, su principal ventaja es que no depende de ninguna base de datos de entrenamiento (al igual que el enfoque NMF ciego). Sin embargo, a diferencia del enfoque NMF ciego, este modelo permite obtener una separación atendiendo al criterio de repetitividad temporal o fuentes sonoras rítmicas. Además, cabe destacar que este enfoque puede ser combinado con un enfoque NMPCF semi-supervisado, utilizando una matriz adicional correspondiente a un espectrograma de entrenamiento de la fuente sonora repetitiva, para reforzar y mejorar el modelado de estos sonidos [170].

En relación a los trabajos publicados durante la Tesis, la publicación [P4] propone una versión extendida del modelo NMPCF definido en [170]. El objetivo de esta publicación es eliminar la interferencia acústica de los sonidos respiratorios normales, aislando de esta forma los sonidos sibilantes de interés para el diagnóstico de las patologías respiratorias. Para ello, se propone modelar los sonidos respiratorios normales, como patrones espectrales que se repiten temporalmente en cada ciclo respiratorio. Para mejorar el rendimiento del enfoque NMPCF propuesto en [170], que trata por igual todos los segmentos del espectrograma, la principal contribución de la propuesta consiste en variar la importancia de cada segmento en el modelado de los patrones respiratorios, distinguiendo entre segmentos no-sibilantes (sonidos sibilantes inactivos) y segmentos sibilantes (sonidos sibilantes activos). Además, esta publicación contiene una evaluación exhaustiva entre los diferentes enfoques de separación NMF/NMPCF comentados en la sección actual, mostrando las ventajas y desventajas de cada enfoque.

3.4. Regularizaciones y restricciones

Como se ha descrito anteriormente, el enfoque NMF estándar modela el espectrograma en magnitud de una señal mezcla de entrada como el producto de una matriz de bases y una matriz de activaciones, con la única restricción de que todos los elementos que componen a cada matriz sean no negativos. Bajo esta restricción de no negatividad, el objetivo es minimizar la función de coste que mide el error de reconstrucción entre el espectrograma de entrada y el espectrograma estimado. Sin embargo, esta restricción no asegura una representación significativa y fiable, en las que las bases o activaciones tengan significado físico para una interpretación coherente en el ámbito científico determinado (sonidos musicales o incluso los sonidos biomédicos auscultados). En este sentido, varias propiedades pueden ser utilizadas para mejorar la singularidad de los mínimos locales obtenidos por el enfoque NMF estándar, incorporando un significado físico a las funciones de bases y activaciones, para converger a una mejor solución desde el criterio de calidad acústica. En particular, estas propiedades pueden aplicarse de dos maneras: i) utilizando regularizaciones (también conocidas como términos de penalización) que se añaden a la función de coste global en la factorización para modelar las propiedades deseadas de los eventos sonoros presentes en la mezcla (ver Sección 3.4.1); y ii) imponiendo restricciones al modelo de señal del enfoque NMF para fijar algunos de los elementos de las matrices y reducir el número de parámetros libres (ver Sección 3.4.2).

3.4.1. Regularizaciones

Para obtener soluciones óptimas y converger en un menor mínimo local añadiendo significado a las bases o activaciones factorizadas tal y como aparecen dichas bases o activaciones en las fuentes sonoras del mundo real, es habitual incluir regularizaciones adicionales en el modelo de descomposición, a fin de mejorar y controlar las propiedades deseadas de los sonidos presentes en la señal de entrada. La forma de introducir estas regularizaciones en el modelo NMF es aplicando un enfoque basado en "términos de penalización". Es decir, en lugar de minimizar sólo el error de reconstrucción definido por la función de divergencia (por ejemplo: EUC, KL o IS), la función de coste objetivo incluye uno o más términos que cuantifican o regulan las propiedades deseadas en las matrices de bases y activaciones. En este sentido, la función de coste regularizada que debe ser minimizada, puede expresarse como:

$$\arg\min_{\mathbf{B},\mathbf{A}} D(\mathbf{X}|\mathbf{BA}) + \mu D_B(\mathbf{B}) + \lambda D_A(\mathbf{A}) \quad \mathbf{B}, \mathbf{A} \geq 0 \tag{3.38}$$

donde las funciones $D_B(\cdot)$ y $D_A(\cdot)$ definen las regularizaciones para modelar las propiedades deseadas de las matrices $\mathbf{B}$ y $\mathbf{A}$, mientras que μ y λ son parámetros de peso que pueden ser ajustados para aumentar o disminuir la influencia de las regularizaciones sobre el procedimiento de minimización del enfoque NMF. Las regularizaciones también pueden ser definidas individualmente para un determinado subconjunto de bases o activaciones. Por ejemplo, en los enfoques de separación de fuentes sonoras descritos en la Sección 3.3, se pueden aplicar diferentes regularizaciones a las matrices de bases, $\mathbf{B}_W$ y $\mathbf{B}_R$, considerando las características espectrales de

los sonidos que componen la mezcla. De la misma forma, se pueden aplicar diferentes regularizaciones a las matrices de activaciones, $\mathbf{A}_W$ y $\mathbf{A}_R$, considerando las características temporales.

Muchos autores han propuesto regularizaciones específicas junto con sus correspondientes reglas multiplicativas de actualización. Virtanen [312] propuso una solución genérica, basada en un enfoque heurístico, para obtener las ecuaciones de actualización multiplicativas que permiten minimizar la función de coste objetivo compuesta por una o más regularizaciones (ver Ec. (3.38)). Este enfoque permite obtener ecuaciones de actualización multiplicativas similares a las presentadas por Lee y Seung [185] (ver Ec. (3.9) y Ec. (3.10)). Este método se basa en el cálculo de las derivadas parciales de la función objetivo $\Gamma(\cdot)$ (considerando que $\Gamma(\cdot)$ es la función de coste definida en la Ec. (3.38) y puede estar compuesta por una o más regularizaciones) con respecto a los parámetros $\mathbf{B}$ y $\mathbf{A}$. Las reglas de actualización multiplicativas se formulan como

$$\mathbf{B} \leftarrow \mathbf{B} \odot \frac{\left[\frac{\partial \Gamma(\cdot)}{\partial \mathbf{B}}\right]^{-}}{\left[\frac{\partial \Gamma(\cdot)}{\partial \mathbf{B}}\right]^{+}} \tag{3.39}$$

$$\mathbf{A} \leftarrow \mathbf{A} \odot \frac{\left[\frac{\partial \Gamma(\cdot)}{\partial \mathbf{A}}\right]^{-}}{\left[\frac{\partial \Gamma(\cdot)}{\partial \mathbf{A}}\right]^{+}} \tag{3.40}$$

donde $[\ ]^{-}$ y $[\ ]^{+}$ se refieren al conjunto de términos negativos y positivos de la derivada parcial, el símbolo $\odot$ denota al producto matricial elemento a elemento y la división matricial es elemento a elemento. Observe que las normas de actualización introducidas en la Ec. (3.9) y la Ec. (3.10) son un caso particular de estas expresiones sin que ninguna regularización sea aplicada. Por consiguiente, cualquier modelo NMF regularizado puede resolverse encontrando los gradientes (las derivadas parciales) de la función objetivo con respecto a cada parámetro.

A continuación se presentan algunas de las regularizaciones más comunes centradas en el procesado de audio para la separación de fuentes sonoras.

Dispersión (Sparsity o Sparseness):

En términos generales, la regularización de dispersión "sparsity o sparseness" $D_{sp}(\cdot)$ denota que una fuente sonora puede considerarse inactiva la mayor parte del tiempo o de la frecuencia [98, 312]. Esta regularización puede aplicarse a las bases espectrales $\mathbf{B}$ y a las activaciones temporales $\mathbf{A}$ del modelo NMF. Por un lado, la dispersión temporal $D_{sp}(\mathbf{A})$, aplicada a la matriz de activaciones $\mathbf{A}$, supone que las fuentes están inactivas la mayor parte del tiempo, por lo que se asigna un alto coste a las activaciones no nulas. Por otro lado, la dispersión espectral $D_{sp}(\mathbf{B})$, aplicada a la matriz de bases $\mathbf{B}$, supone que las fuentes están inactivas la mayor parte de la frecuencia, por lo que se asigna un alto coste a las bases no nulas. Como resultado, la regularización de dispersión a menudo obtiene mejores mínimos locales en la descomposición del NMF, minimizando la presencia de falsas activaciones o la aparición de energía espuria en frecuencia (bases espectrales). Esta regularización puede ser añadida a la función de coste

objetivo como un término de penalización aplicado a cada matriz regularizada,

$$\arg\min_{\mathbf{B},\mathbf{A}} D(\mathbf{X}|\mathbf{BA}) + \lambda_1 D_{sp}(\mathbf{B}) + \lambda_2 D_{sp}(\mathbf{A}) \quad \mathbf{B}, \mathbf{A} \geq 0 \tag{3.41}$$

donde λ_1 y λ_2 controlan el efecto de la regularización de dispersión aplicada a cada matriz. Como se propone en [312], uno de los términos de penalización más utilizados para aplicar la regularización de dispersión a la matriz $\mathbf{B}$ y $\mathbf{A}$ es la norma l_1, ya que se ha demostrado que es menos sensible a los cambios del parámetro de pesado que controla la importancia de la regularización en el proceso de factorización NMF (λ_1 para $D_{sp}(\mathbf{B})$ y λ_2 para $D_{sp}(\mathbf{A})$). Por lo tanto, $D_{sp}(\mathbf{B}) = \|\mathbf{B}\|_1$ y $D_{sp}(\mathbf{A}) = \|\mathbf{A}\|_1$.

La Figura 3.7 muestra un ejemplo del efecto de aplicar el término de penalización de dispersión temporal $D_{sp}(\mathbf{A})$ al modelo NMF. Comparando la Figura 3.2 (Modelo NMF estándar) y la Figura 3.7 (Modelo NMF con la regularización de dispersión temporal), se puede observar que la regularización de dispersión proporciona una mejor solución, ya que esta regularización consigue representar de forma más fiable las activaciones en los intervalos temporales reales en los que cada sonido está activo. Este hecho minimiza la presencia de activaciones erróneas en aquellos intervalos en los que estas componentes están inactivas.

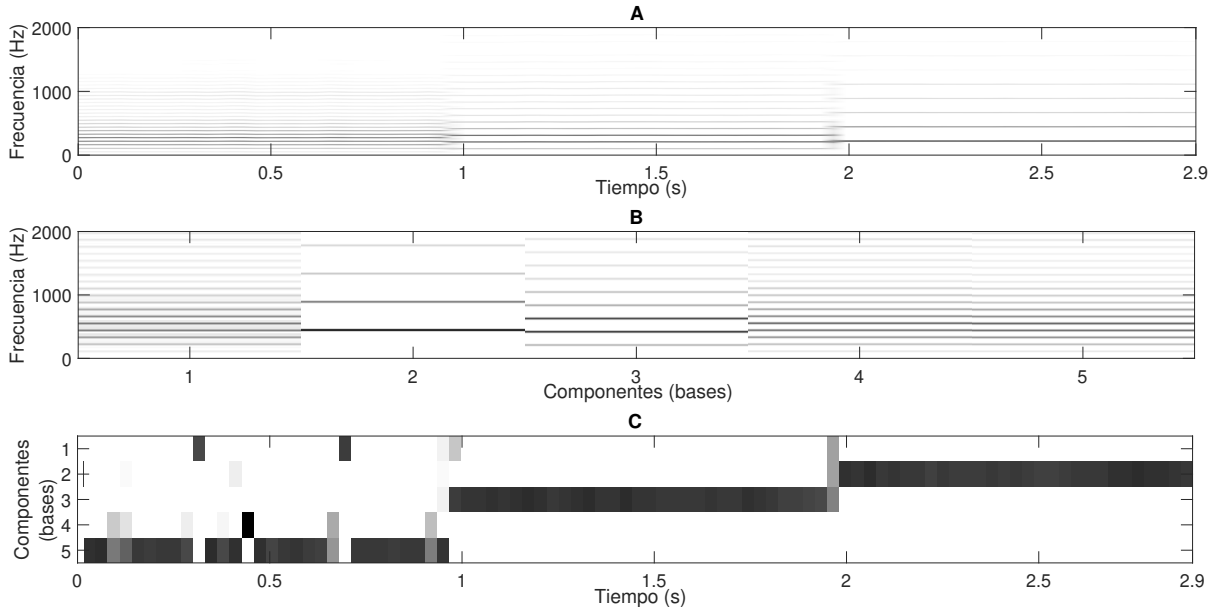

Figura 3.7: Ejemplo de factorización aplicando el modelo NMF regularizado (dispersión temporal $D_{sp}(\mathbf{A})$), utilizando $K = 5$ componentes, sobre un fragmento de señales musicales monofónicas, con una duración aproximada de 2.9 segundos, compuesto por tres sonidos armónicos con frecuencias fundamentales $f_{01} = 107Hz$, $f_{02} = 217Hz$ y $f_{03} = 440Hz$ tocados por un trombón. El color más oscuro indica mayor amplitud. A) Espectrograma en magnitud de la señal de entrada $\mathbf{X}$; B) Diccionario de bases $\mathbf{B}$; y C) Matriz de activaciones $\mathbf{A}$. Figura extraída de la referencia [65].

Suavidad (Smoothness):

Generalmente, la regularización de suavidad "smoothness" $D_{sm}(\cdot)$ controla como de continuos o suaves son los cambios espectrales o temporales relacionados con una fuente sonora [312]. Esta regularización puede aplicarse a las bases espectrales $\mathbf{B}$ y a las activaciones temporales $\mathbf{A}$ del modelo NMF. Por un lado, la suavidad temporal $D_{sm}(\mathbf{A})$, aplicada a la matriz de activaciones $\mathbf{A}$, indica como de lentas son las variaciones de la amplitud a lo largo del tiempo. Por otro lado, la suavidad espectral $D_{sm}(\mathbf{B})$, aplicada a la matriz de bases $\mathbf{B}$, indica como de lentas son las variaciones de amplitud a lo largo del rango de frecuencias. Esta regularización puede ser añadida a la función de coste objetivo como un término de penalización aplicado a cada matriz regularizada,

$$\arg\min_{\mathbf{B},\mathbf{A}} D(\mathbf{X}|\mathbf{BA}) + \lambda_1 D_{sm}(\mathbf{B}) + \lambda_2 D_{sm}(\mathbf{A}) \quad \mathbf{B}, \mathbf{A} \geq 0 \tag{3.42}$$

donde λ_1 y λ_2 controlan el efecto de la regularización de suavidad aplicada a cada matriz. Como propuso Virtanen en [312], la suavidad temporal $D_{sm}(\mathbf{A})$ de las componentes se refuerza asignando un alto coste a los grandes cambios en el tiempo producidos entre las activaciones a_{kt} y $a_{k(t-1)}$ en tramas adyacentes,

$$D_{sm}(\mathbf{A}) = \sum_{k=1}^{K} \frac{1}{\sigma_k^2} \sum_{t=2}^{T} \left(a_{kt} - a_{k(t-1)}\right)^2 \tag{3.43}$$

donde las activaciones son normalizadas por su desviación estándar, aplicando $\sigma_k = \sqrt{\frac{1}{T}\sum_{t=1}^{T} a_{kt}^2}$, para evitar que la escala numérica de las activaciones afecte a la función de coste [312, 64]. La Figura 3.8 muestra un ejemplo del efecto de aplicar el término de penalización de suavidad temporal $D_{sm}(\mathbf{A})$ al modelo NMF.

Por otro lado, en [64, 210], la regularización de suavidad temporal propuesta en [312] fue adaptada para evaluar la suavidad espectral de las bases, la regularización de suavidad espectral resultante se define como,

$$D_{sm}(\mathbf{B}) = \sum_{k=1}^{K} \frac{1}{\sigma_k^2} \sum_{f=2}^{F} \left(b_{fk} - b_{(f-1)k}\right)^2 \tag{3.44}$$

donde $\sigma_k = \sqrt{\frac{1}{F}\sum_{f=1}^{F} b_{fk}^2}$ representa la desviación estándar utilizada para normalizar las funciones base como en la Ec. (3.43).

Ortogonalidad (Orthogonality):

La ortogonalidad "orthogonality" $D_{or}(\cdot)$, propuesta inicialmente por Li y otros [191], es una regularización que puede aplicarse a la matriz de bases $\mathbf{B}$ y de activaciones $\mathbf{A}$. Ding y otros [92], propusieron aplicar la ortogonalidad en el enfoque NMF estándar para demostrar su capacidad de conseguir una clasificación de las bases espectrales con interpretaciones más

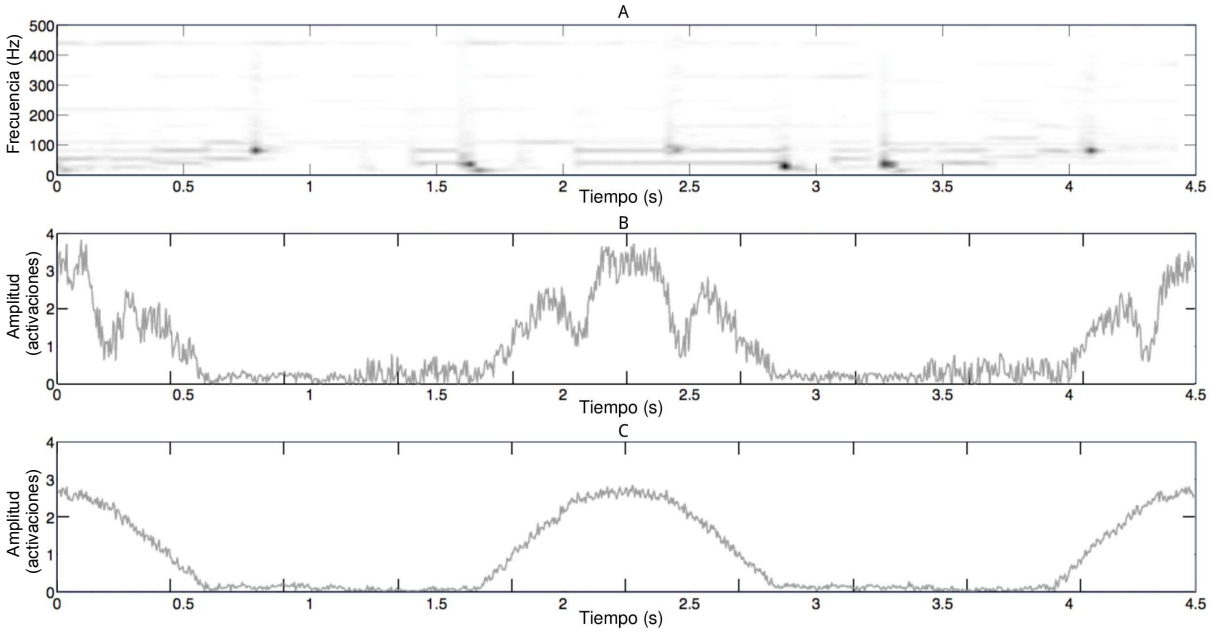

Figura 3.8: Ejemplo del efecto de aplicar suavidad temporal al modelo NMF utilizando como señal de entrada un fragmento musical monofónico, con una duración aproximada de 4.5 segundos. A) Espectrograma en magnitud de la señal de entrada $\mathbf{X}$; B) Amplitud de las activaciones utilizando el modelo NMF estándar; y C) Amplitud de las activaciones utilizando el modelo NMF regularizado (suavidad temporal $D_{sm}(\mathbf{A})$). Figura extraída de la referencia [65].

rigurosas. Centrándonos, en la ortogonalidad espectral $D_{or}(\mathbf{B})$, aplicada a la matriz de bases $\mathbf{B}$, se debería cumplir que $\mathbf{B}^{\mathrm{T}}\mathbf{B} = I$, siendo I la matriz de identidad y $^{\mathrm{T}}$ el operador traspuesto. De forma similar, en el caso de la ortogonalidad temporal $D_{or}(\mathbf{A})$, aplicada a la matriz de activaciones $\mathbf{A}$, se debería cumplir que $\mathbf{A}\mathbf{A}^{\mathrm{T}} = I$. Considerando el caso de la ortogonalidad espectral, esta regularización puede ser añadida a la función de coste objetivo como un término de penalización,

$$\arg\min_{\mathbf{B},\mathbf{A}} D(\mathbf{X}|\mathbf{B}\mathbf{A}) + \lambda D_{or}(\mathbf{B}) \quad \mathbf{B}, \mathbf{A} \geq 0 \tag{3.45}$$

$$D_{or}(\mathbf{B}) = \frac{1}{2} Trace\big(\mathbf{B}^T\mathbf{B} - I\big) \tag{3.46}$$

donde λ controla el efecto de la regularización de ortogonalidad espectral aplicada a la matriz $\mathbf{B}$ y el operador $Trace$ calcula la suma de los elementos diagonales de la matriz cuadrada $\mathbf{B}^T\mathbf{B} - I$. Como resultado, $D_{or}(\mathbf{B})$ factoriza las bases espectrales del diccionario $\mathbf{B}$ tan disímiles (ortogonales) como sea posible con el objetivo de minimizar la redundancia entre ellas. Por lo tanto, un enfoque NMF basado en la ortogonalidad tiene un mejor rendimiento en la agrupación o clasificación de fuentes sonoras en comparación con el enfoque NMF estándar, ya que la ortogonalidad espectral permite factorizar los patrones espectrales más distintivos que componen la mezcla de entrada. La Figura 3.9 muestra un ejemplo del efecto de aplicar el término de penalización de ortogonalidad espectral $D_{or}(\mathbf{B})$ al modelo NMF. Además, la Figura

3.10 muestra la correlación entre las bases espectrales mostradas en la Figura 3.2 (Modelo NMF estándar), Figura 3.7 (Modelo NMF con la regularización de dispersión temporal) y Figura 3.9 (Modelo NMF con la regularización de ortogonalidad espectral). La Figura 3.10 confirma que la regularización de ortogonalidad espectral obliga a representar la señal de entrada utilizando bases espectrales más diferenciadas. Este comportamiento puede ser observado comparando las tres subfiguras de la Figura 3.10, donde la regularización de ortogonalidad consigue mostrar con mayor claridad la matriz de identidad $\mathbf{I}$ al comparar las bases espectrales del diccionario $\mathbf{B}$.

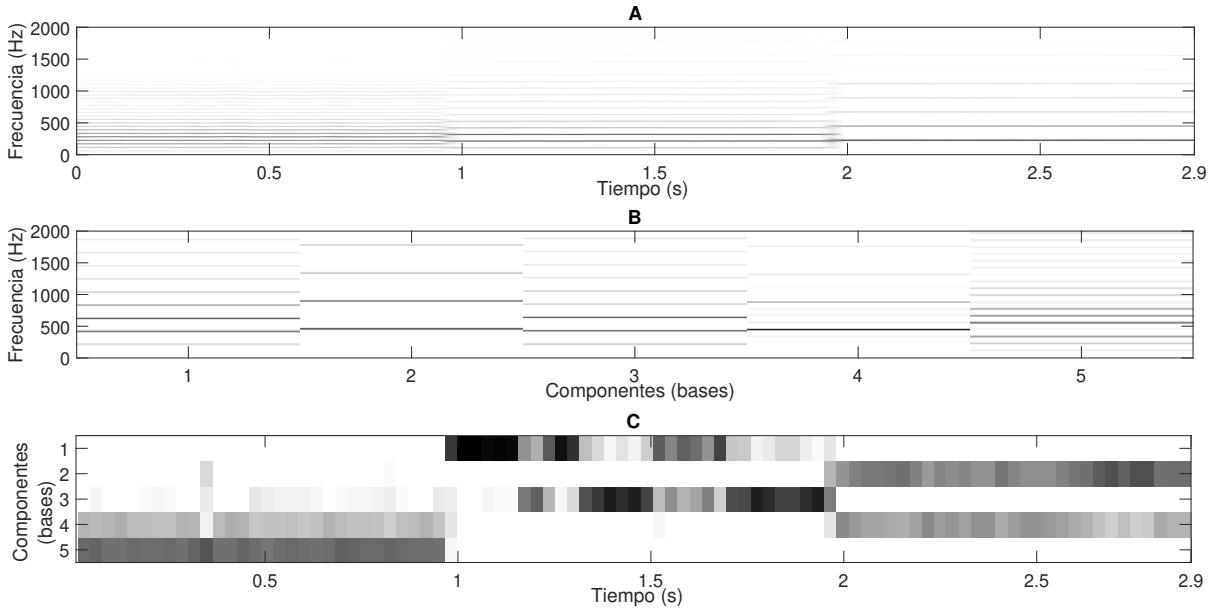

Figura 3.9: Ejemplo de factorización aplicando el modelo NMF regularizado (ortogonalidad espectral $D_{or}(\mathbf{B})$), utilizando $K = 5$ componentes, sobre un fragmento de señales musicales monofónicas, con una duración aproximada de 2.9 segundos, compuesto por tres sonidos armónicos con frecuencias fundamentales $f_{01} = 107Hz$, $f_{02} = 217Hz$ y $f_{03} = 440Hz$ tocados por un trombón. El color más oscuro indica mayor amplitud. A) Espectrograma en magnitud de la señal de entrada $\mathbf{X}$; B) Diccionario de bases $\mathbf{B}$; y C) Matriz de activaciones $\mathbf{A}$. Figura extraída de la referencia [65].

Las siguientes dos regularizaciones (Correlación cruzada y Monofonía) se basan en esta regularización de ortogonalidad y entre otras aplicaciones, han sido aplicadas a los ámbitos de transcripción musical y restauración de audio.

Correlación cruzada (Cross-Correlation):

En un escenario musical, donde varios instrumentos pueden ser tocados al mismo tiempo, es habitual encontrar que las bases espectrales que componen el diccionario $\mathbf{B}$, a veces, puedan representar patrones espectrales correspondientes a más de una fuente sonora. En otras palabras, una misma base puede mezclar los patrones espectrales de diferentes fuentes sonoras. Cuando esta situación ocurre, la estimación de cada fuente sonora contendrá residuos de las otras fuentes presentes, lo que degrada el rendimiento de la separación. Grais y Erdogan [129]

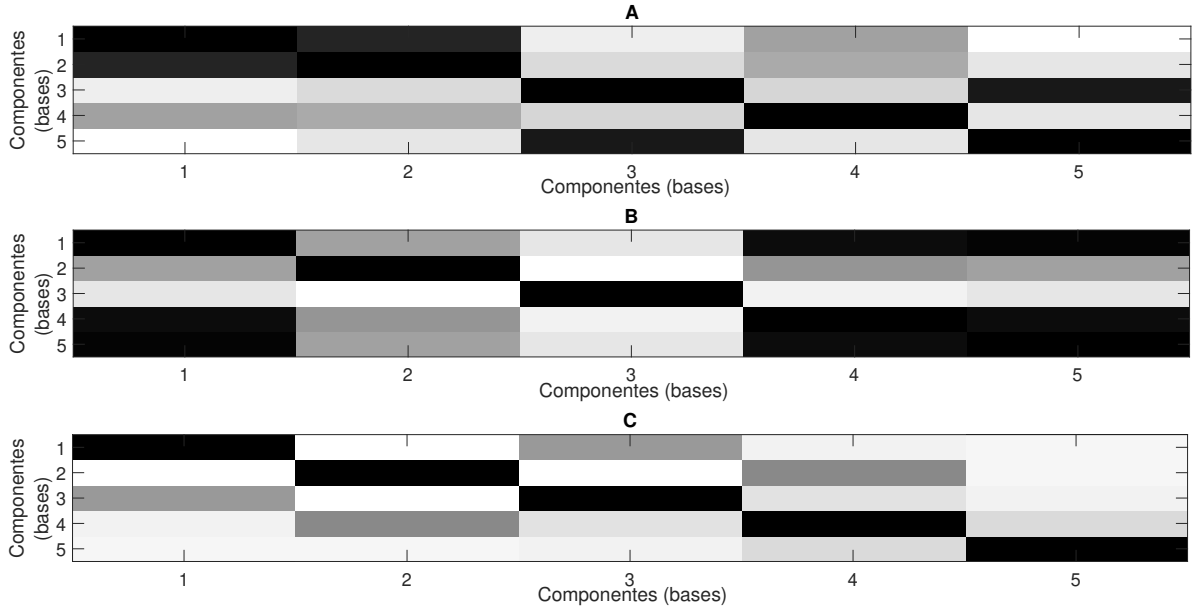

Figura 3.10: Correlación entre las distintas bases del diccionario $\mathbf{B}$ (en total $K = 5$ componentes o bases espectrales), obtenidas a partir de un fragmento de señales musicales monofónicas, con una duración aproximada de 2.9 segundos, compuesto por tres sonidos armónicos con frecuencias fundamentales $f_{01} = 107Hz$, $f_{02} = 217Hz$ y $f_{03} = 440Hz$ tocados por un trombón. A) Aplicando el modelo NMF estándar (ver Figura 3.2); B) Aplicando el modelo NMF con la regularización de dispersión temporal (ver Figura 3.7); y C) Aplicando el modelo NMF con la regularización de ortogonalidad espectral (ver Figura 3.9). Figura extraída de la referencia [65].

proponen aplicar la correlación cruzada "Cross-Correlation" $D_{cc}(\cdot)$ a las bases de la matriz $\mathbf{B}$ para minimizar la redundancia entre ellas, a partir de dos diccionarios para dos fuentes distintas. Los resultados muestran que la correlación cruzada reduce la redundancia entre el primer diccionario con respecto al segundo y viceversa, evitando que ambos diccionarios puedan contener patrones espectrales de ambas fuentes y mejorando el rendimiento de separación entre las fuentes.

Raczynski y otros [259], propusieron un término de penalización $D_{cc}(\mathbf{A})$, basado en la correlación cruzada de la matriz de activaciones $\mathbf{A}$, para mejorar el rendimiento en el contexto de la transcripción musical,

$$\underset{\mathbf{B},\mathbf{A}}{\arg\min}\, D(\mathbf{X}|\mathbf{BA}) + D_{cc}(\mathbf{A}) \quad \mathbf{B},\mathbf{A} \geq 0 \tag{3.47}$$

$$D_{cc}(\mathbf{A}) = \big(\lambda_1 \varphi_1(\mathbf{A}) + \lambda_2 \varphi_2(\mathbf{A})\big) \tag{3.48}$$

donde λ_1 y λ_2 son los parámetros que controlan la importancia de los términos regularizados. En primer lugar, se introduce el término de penalización $\varphi_1(\mathbf{A})$, basado en la correlación cruzada de las activaciones $\mathbf{A}$, para disminuir el cruce espectral (división de los patrones espectrales) entre las posibles notas musicales, de la siguiente forma:

$$\varphi_1(\mathbf{A}) = \big|\mathbf{C} \odot (\mathbf{A}\mathbf{A}^{\mathrm{T}})\big| \tag{3.49}$$

donde $\mathbf{C}$ es una matriz de ponderación que selecciona los términos cruzados a penalizar y por cuánto son penalizados. Los pesos se establecen de tal manera que los términos no cruzados (elementos en la diagonal) no se penalizan, es decir $c_{ii} = 0$, y de tal manera que las penalizaciones sólo dependan de los intervalos entre las notas, por lo tanto la matriz $\mathbf{C}$ es circulante. Específicamente, suele ocurrir que la matriz $\mathbf{C}$ penalice los intervalos de octava y quinta. En este sentido, aplicar las regularizaciones de dispersión y correlación cruzada es útil para reducir al mínimo las tasas de error, tales como los errores de octava y quinta [259].

En segundo lugar, se introduce el término de penalización $\varphi_2(\mathbf{A})$, basado en la correlación cruzada en el dominio del tiempo, para fomentar la suavidad temporal de las activaciones, de la siguiente forma:

$$\varphi_2(\mathbf{A}) = -\left|\mathbf{D} \odot (\mathbf{A}^{\mathrm{T}}\mathbf{A})\right| \tag{3.50}$$

donde $\mathbf{D}$ se utiliza para penalizar la discontinuidad temporal entre tramas consecutivas. Como en la Ec. (3.49), $\mathbf{D}$ es forzada a ser circulante y tener una diagonal compuesta por ceros ($d_{ii} = 0$). Específicamente, la matriz $\mathbf{D}$ propuesta en [259] muestra un perfil exponencial decreciente, de tal forma que los elementos que se acercan a la diagonal tienen valores altos y los que están más lejos de la diagonal tienden a 0.

Monofonía (Single-Activation):

La regularización de monofonía "Single-Activation" $D_{sa}(\cdot)$, fue propuesta por Canadas y otros [63] para mejorar el modelado de las características temporales mostradas en la matriz de activaciones $\mathbf{A}$, en relación al comportamiento temporal típico del ruido de vinilo. Específicamente la motivación de su propuesta se debe a que el ruido de vinilo revela una naturaleza no estacionaria, la cual implica que las activaciones relacionadas con la misma base no deberían ser consecutivas, ya que este ruido debe ser distinto en cada trama temporal. Esta regularización puede considerarse una ampliación de la correlación cruzada y por lo tanto también se basa en la propiedad de ortogonalidad. Concretamente, la monofonía temporal $D_{sa}(\mathbf{A})$, aplicada a la matriz de activaciones $\mathbf{A}$, permite lograr que cada componente espectral del diccionario $\mathbf{B}$ represente un patrón espectral completo de una determinada fuente sonora (por ejemplo, un patrón espectral de un ruido de vinilo que se encuentre activo, o el de una nota musical). Es decir, esta regularización tiene como objetivo evitar que los patrones espectrales que modelan una parte representativa de un sonido (una nota tocada por un instrumento, por ejemplo) no se dividan en varías bases espectrales, sino en una única base. En este sentido, la monofonía temporal minimiza el número de bases espectrales del diccionario $\mathbf{B}$ que pueden ser activadas simultáneamente en cada trama temporal. Esta regularización puede ser añadida a la función de coste objetivo como un término de penalización,

$$\underset{\mathbf{B},\mathbf{A}}{\arg\min}\, D(\mathbf{X}|\mathbf{B}\mathbf{A}) + \lambda D_{sa}(\mathbf{A}) \quad \mathbf{B}, \mathbf{A} \geq 0 \tag{3.51}$$

$$D_{sa}(\mathbf{A}) = \frac{1}{K\,(K-1)} \sum_{k=1}^{K} \sum_{t=1}^{\mathrm{T}} \mathbf{A}\mathbf{A}^{\mathrm{T}} - Trace\left(\mathbf{A}\mathbf{A}^{\mathrm{T}}\right) \tag{3.52}$$

donde K es el número de componentes, λ controlan el efecto de la regularización de monofonía temporal aplicada a la matriz $\mathbf{A}$ y el operador $Trace$ calcula la suma de los elementos diagonales de la matriz cuadrada $\mathbf{A}\mathbf{A}^{\mathrm{T}}$. Para equilibrar la importancia de la regularización monofónica en el proceso de descomposición, $D_{sa}(\mathbf{A})$ es ponderado por un factor de $\frac{1}{K(K-1)}$. En consecuencia, el coste $D_{sa}(\mathbf{A})$ será igual a 1.0 en el peor de los casos.

3.4.2. Restricciones

Otro método para mejorar la interpretación física de los patrones espectrales en la matriz de bases $\mathbf{B}$ o de activaciones $\mathbf{A}$ es la imposición de restricciones al modelo de señal del enfoque NMF. Las restricciones pueden permitir guiar el proceso de factorización, fijando algunos de los elementos de las matrices. Esto equivale a reducir el número de parámetros libres.

En esta sección se presentan una serie de técnicas no supervisadas en las cuales la matriz de bases $\mathbf{B}$ o un subconjunto de ella es fijada. Aunque la idea de fijar algunos o todos los componentes de la matriz $\mathbf{B}$ es simple, se ha investigado mucho sobre qué valores utilizar para fijar estos componentes [65]. A continuación se presentan algunas de las restricciones más comunes centradas en el procesado de audio para la separación de fuentes sonoras musicales. Para profundizar más en este tema se indica la bibliografía utilizada para desarrollar la información descrita en esta sección [65].

Modelo Excitación/Filtro para la voz (Source/Filter Model to Speech):

En [96], Durrieu y otros proponen un modelo (Source/Filter Model) para representar la señal de interés (normalmente el instrumento musical principal o la voz del cantante) y aislarla de las fuentes sonoras restantes. Esta representación es especialmente interesante para los sonidos vocales, ya que se basa en la aproximación de las características esenciales, con las que los sonidos vocales son producidos. Según el modelo, cada vector espectral de la voz $\mathbf{v}_t$ se descompone en una parte de excitación $\mathbf{v}_t^{F_0}$ multiplicada por otra parte de filtrado $\mathbf{v}_t^{\Phi}$, las cuales están respectivamente compuestas por una combinación lineal de P bases de excitación elemental $\mathbf{b}_p^{F_0}$ y E bases de filtro elemental $\mathbf{b}_e^{\Phi}$, como se muestra a continuación:

$$\mathbf{v}_t = \mathbf{v}_t^{\Phi} \odot \mathbf{v}_t^{F_0} = \left(\sum_{e=1}^{E} a_{et}^{\Phi} \mathbf{b}_e^{\Phi} \right) \odot \left(\sum_{p=1}^{P} a_{pt}^{F_0} \mathbf{b}_p^{F_0} \right) \tag{3.53}$$

donde a_{et}^{Φ} y $a_{pt}^{F_0}$ son ganancias o activaciones no negativas. Las bases de excitación $\mathbf{b}_p^{F_0}$ representan el conjunto discreto de sonidos a partir de los cuales se puede construir la señal, y los cuales son modulados a su vez por una combinación de bases de filtro $\mathbf{b}_e^{\Phi}$. Dado que el modelo está diseñado para representar los sonidos vocales, es conveniente que cada vector $\mathbf{b}_p^{F_0}$ represente la señal glotal correspondiente a una frecuencia fundamental o tono. En [96], las bases $\mathbf{b}_p^{F_0}$ son generadas utilizando el modelo de fuente glotal denominado KLGLOTT88 [173], lo que resulta en un diccionario fijo de excitaciones relacionadas con el tono. Si se utiliza un número suficiente de excitaciones P, es posible disponer de una resolución adecuada para cubrir todo el

rango tonal de la voz cantada o el habla de una persona. Por otro lado, las bases de filtro $\mathbf{b}_e^{\Phi}$ deben ser capaces de representar la envolvente suave de la señal. En [96], estas bases se generan a partir de una familia de funciones de suavizado, lo que da como resultado un diccionario fijo de componentes de suavizado que pueden combinarse para representar cualquier envolvente suave arbitraria.

El modelo Excitación/Filtro tiene dos ventajas interesantes. En primer lugar, describe un modelo genérico que es característico de la voz cantada o del habla, y significativamente discriminativo ante los típicos sonidos del ruido. En segundo lugar, considerando que la información correspondiente al tono y al timbre es representada individualmente, el modelo proporciona una descripción más estructurada de los sonidos vocales.

Modelo Armónico para la música (Harmonic Model to Music):

Cuando se trata de sonidos de instrumentos musicales, cada función base del diccionario $\mathbf{B}$ puede representarse idealmente por un único tono (pitch) y sus correspondientes activaciones pueden contener información sobre sus tiempos de inicio (onset) y de fin (offset). Varios trabajos [314, 260, 108, 309, 55, 66, 67] propusieron restringir el modelo NMF (ver Ec. (3.2)) forzándolo a ser armónico. El modelo armónico es particularmente útil para el análisis y la separación de las señales de audio musicales ya que, al utilizar esta restricción, cada base puede definir una única frecuencia fundamental (tono).

En [309, 55] las bases de la matriz $\mathbf{B}$ del modelo NMF estándar (ver Ec. (3.2)) son definidas como una combinación ponderada de patrones espectrales armónicos de banda estrecha $\mathbf{P} \in \mathbb{R}_+^{F0\times K\times U}$, las cuales son arbitrariamente fijadas,

$$\mathbf{b}_k = \sum_{u=1}^{U} e_{ku}\mathbf{p}_{ku} \tag{3.54}$$

donde cada base espectral $\mathbf{b}_k$ está asociada a un único tono (con su correspondiente frecuencia fundamental $f0$). Cada base $\mathbf{b}_k$ es obtenida a partir de una combinación lineal de u patrones armónicos $\mathbf{p}_{ku}$ con diferente forma pero compartiendo la misma frecuencia fundamental del tono $f0$. Estos patrones armónicos son escalados mediante un conjunto de coeficientes $\mathbf{E} \in \mathbb{R}_+^{K\times U}$, los cuales definen la representación espectral real de cada componente $\mathbf{b}_k$.

Algunos enfoques [67, 120] utilizan una extensión del modelo anterior empleando una única excitación armónica plana, donde las bases de cada nota e instrumento se reducen a un conjunto de coeficientes que definen la forma del patrón espectral armónico,

$$\mathbf{b}_{nj} = \sum_{h=1}^{H} r_{njh}\mathbf{g}(f - hf_0^n) \tag{3.55}$$

donde $\mathbf{b}_{nj}$ se refiere a la base espectral de la nota n del instrumento j, H define el número de armónicos por cada nota, r_{njh} es la amplitud del armónico h para la nota n y el instrumento j, f_0^n es la frecuencia fundamental de la nota n, $\mathbf{g}(f)$ es el espectro en magnitud de la función

ventana, y el espectro de una componente armónica situada en la frecuencia hf_0^n es aproximado por $\mathbf{g}(f - hf_0^n)$.

Aunque el modelo Excitación/Filtro (Source/Filter Model) tiene su origen en el procesado del habla y la síntesis del sonido, un modelo similar puede extrapolarse a los instrumentos musicales [314, 45]. De hecho, cada instrumento puede ser representado usando un solo filtro que corresponde a la estructura resonante del cuerpo del instrumento, mientras que las excitaciones pueden representarse como componentes espectrales en magnitud de múltiplos enteros de una cierta frecuencia fundamental, como se muestra a continuación:

$$\mathbf{b}_{nj} = \mathbf{q}_j \underbrace{\sum_{h=1}^{H} \mathbf{g}(f - hf_0^n)}_{\mathbf{e}_n} \tag{3.56}$$

donde cada base $\mathbf{b}_{nj}$ es modelada como el producto entre el espectrograma en magnitud de la excitación $\mathbf{e}_n$, correspondiente a la nota n almacenada en la matriz $\mathbf{E} \in \mathbb{R}_+^{F \times N}$, y el filtro $\mathbf{q}_j$, correspondiente al instrumento j almacenado en la matriz $\mathbf{Q} \in \mathbb{R}_+^{F \times J}$.

Algunas extensiones del modelo Excitación-Filtro para señales de música pueden ser encontradas en la literatura. Por ejemplo, Heittola y otros [138] propusieron un modelo donde en lugar de definir una excitación para cada tono posible, estas excitaciones son dadas por un estimador multipitch. Además, los filtros $\mathbf{Q}$ se representan como una combinación lineal de respuestas elementales fijas. En particular, los autores definieron las respuestas elementales como respuestas en magnitud paso banda, triangulares y uniformemente espaciadas sobre una escala de frecuencias Mel. Finalmente, en [67], los autores propusieron descomponer la única excitación plana por una combinación de unos pocos patrones de excitación, para mejorar el modelado de los cambios en el timbre entre notas a través de la frecuencia.

3.5. Clustering de bases espectrales

Otra tarea en la que varios autores [243, 149] se han centrado, es realizar una clasificación o agrupación de las diferentes bases espectrales $\mathbf{b}_k$, que componen al diccionario $\mathbf{B}$ descrito en el modelo NMF (ver Ec. (3.2)). La clasificación de las bases espectrales permite diferenciar los patrones espectrales que caracterizan a una fuente sonora con respecto a las demás fuentes.

Aunque existen muchos descriptores para clustering tanto de bases como activaciones, esta Tesis presenta aquellos descriptores que analizan la distribución espectral de las bases $\mathbf{b}_k$ para clasificarlas en función de su carácter tonal. En este sentido, se han propuesto diferentes descriptores [243, 149] que permiten medir el grado de tonalidad que caracteriza a cada base espectral. Realmente la tonalidad de una base espectral, define como de dispersa (sparsity) o suave (smoothness) es su distribución espectral. Específicamente, una base espectral tendrá un alto grado de tonalidad cuando la mayor parte de la energía se encuentre localizada en torno a una frecuencia, es decir, cuando su distribución espectral describa un patrón en banda estrecha (sparsity en frecuencia). Por otro lado, una base espectral tendrá un bajo grado de tonalidad cuando la mayor parte de la energía esté uniformemente distribuida a lo largo del rango

de frecuencias, es decir, cuando su distribución espectral describa un patrón en banda ancha (smoothness en frecuencia). En este sentido, Park y otros [243] propusieron utilizar algunos de estos descriptores para realizar la separación entre los sonidos armónicos y percusivos. La motivación de su propuesta es que los sonidos percusivos se pueden modelar con patrones espectrales en banda ancha y los sonidos armónicos con patrones espectrales en banda estrecha. Por lo tanto, midiendo el grado de tonalidad de las distintas bases espectrales $\mathbf{b}_k$ del diccionario $\mathbf{B}$, estas pueden ser agrupadas en dos subconjuntos: las de mayor tonalidad (armónicas) y las de menor tonalidad (percusivas).

A continuación se muestran los principales descriptores propuestos [243, 149, 44, 328] que ofrecen mejores resultados a la hora de medir el grado de tonalidad (dispersión) de una distribución espectral. Note que en todas las ecuaciones mostradas a continuación, $\mathbf{b}_k$ denota la distribución de energía espectral de la base k del diccionario $\mathbf{B}$, y b_{fk} se refiere al valor de energía correspondiente a dicha base k en la frecuencia (bins) f. Cada descriptor devuelve un valor escalar ψ que determina la tonalidad del vector espectral $\mathbf{b}_k$ analizado. Dicho esto, los descriptores más relevantes son:

- **Shannon entropy** (ψ_E): mide el grado de incertidumbre presente en el vector espectral.

$$\psi_E = -\sum_{f=1}^{F} \left(b_{fk}^2 \log b_{fk}^2\right) \tag{3.57}$$

- **Spectral flatness** (ψ_F): relación entre los valores medios geométricos y aritméticos del contenido espectral. En concreto, proporciona información sobre si la señal es más parecida a un tono o al ruido.

$$\psi_F = \frac{\sqrt[F]{\prod_{f=0}^{F-1} b_{fk}}}{\frac{\sum_{f=0}^{F-1} b_{fk}}{F}} \tag{3.58}$$

- **Gini index** (ψ_G): mide la desigualdad de una distribución espectral.

$$\psi_G = \frac{F+1}{F} - \frac{2\sum_{f=1}^{F}(F+1-f)\, b_{fk}^{(sorted)}}{F\sum_{f=1}^{F} b_{fk}^{(sorted)}} \tag{3.59}$$

  donde $b_{fk}^{(sorted)}$ indica el vector $\mathbf{b}_k$ ordenado ascendentemente.

- **Kurtosis** (ψ_K): mide el grado de pico de una distribución espectral. En términos coloquiales mide como de picuda (peakedness) es la distribución.

$$\psi_K = \frac{\sum_{f=1}^{F} b_{fk}^4}{\left(\sum_{f=1}^{F} b_{fk}^2\right)^2} \tag{3.60}$$

- $\frac{\|\mathbf{b}_k\|_1}{\|\mathbf{b}_k\|_2}$ (ψ_L): norma l_1 de la distribución espectral, normalizada por su norma l_2. En concreto, proporciona información sobre el grado de dispersión (sparsity) del vector espectral.

$$\psi_L = \frac{\sum_{f=1}^{F} b_{fk}}{\sqrt{\sum_{f=1}^{F} b_{fk}^2}} \tag{3.61}$$

- **Hoyer** (ψ_H): versión normalizada del descriptor ψ_L y está directamente relacionado con el grado de dispersión del vector espectral.

$$\psi_H = \left(\sqrt{F} - \frac{\sum_{f=1}^{F} b_{fk}}{\sqrt{\sum_{f=1}^{F} b_{fk}^2}}\right)\left(\sqrt{F}-1\right)^{-1} \tag{3.62}$$

- **Spectral peaks entropy** (ψ_S): mide la entropía en base al ratio entre cada pico espectral y la suma total de estos picos.

$$\psi_S = -\sum_{p=1}^{P} \frac{c_p}{\sum_{p=1}^{P} c_p} \log_{10}\left(\sum_{p=1}^{P} c_p\right) \tag{3.63}$$

  donde c_p denota el valor de cada pico espectral $(c_1, c_1, ..., c_P)$ del vector espectral $\mathbf{b}_k$.

Note que el vector espectral $\mathbf{b}_k$ define una distribución espectral tonal (sparsity) cuando el valor de los descriptores ψ_E, ψ_F y ψ_L es bajo o cuando el valor de los descriptores ψ_G, ψ_H, ψ_K y ψ_S es alto. Considerando los descriptores anteriormente descritos, Hurley y Rickard [149] demostraron que el descriptor Gini index se puede considerar la métrica de dispersión que mejor rendimiento ofrece para describir y clasificar la tonalidad de un conjunto de vectores espectrales.

Finalmente, la Figura 3.11 muestra un ejemplo de clasificación de bases espectrales NMF, atendiendo al grado de tonalidad. Específicamente, este ejemplo ha sido obtenido de uno de los trabajos publicados [P5] en esta Tesis. En este caso, se realiza la clasificación de una matriz de bases $\mathbf{B}$ compuesta por $K = 80$ bases espectrales utilizando el descriptor ψ_G. A partir de esta clasificación se obtienen dos subconjuntos de bases: las de mayor dispersión (correspondientes a los sonidos sibilantes) y las de menor dispersión (correspondientes a los sonidos respiratorios). Tenga en cuenta que el modelo NMF que se propuso en este trabajo utiliza la regularización de ortogonalidad para reducir la similitud entre bases y hacerlas lo más distintas posibles.

3.6. Conclusiones

En este capítulo se ha realizado una revisión sobre los modelos de señal basados en factorización de matrices no negativas. Se ha realizado una introducción sobre la motivación de utilizar estos modelos en el campo del procesado de audio, considerando su rendimiento para modelar las características tiempo-frecuencia de los diferentes sonidos que pueden componer a

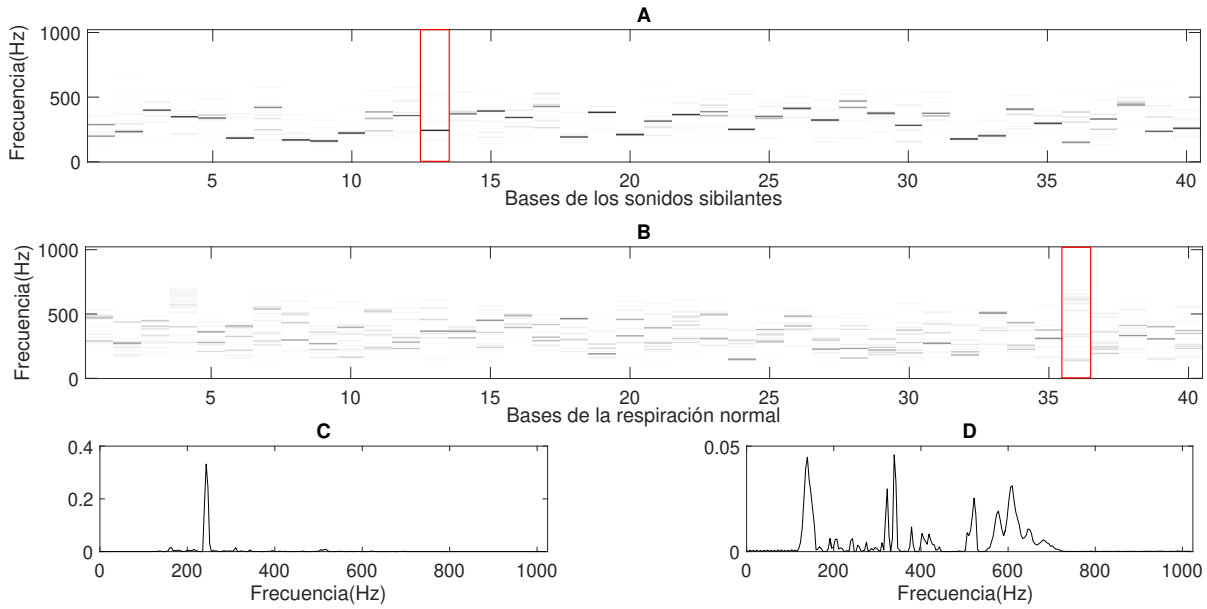

Figura 3.11: Ejemplo de clasificación de bases espectrales correspondiente al trabajo publicado [P5]. A) Conjunto de bases con mayor grado de tonalidad, asociadas a los sonidos sibilantes. B) Conjunto de bases con menor grado de tonalidad, asociadas a los sonidos de la respiración normal. C) Distribución de la energía espectral de la base más tonal $\mathbf{b}_{13}$ (rectángulo rojo). D) Distribución de la energía espectral de la base menos tonal $\mathbf{b}_{36}$ (rectángulo rojo).

una señal mezcla. En segundo lugar, se han detallado los principios en los que se basa el modelo NMF estándar para la factorización de señales de audio, incluyendo una explicación exhaustiva sobre el funcionamiento del modelo. En tercer lugar, se han presentados los diferentes enfoques, basados en descomposición de matrices no negativas (NMF y NMPCF), utilizados para la separación de fuentes sonoras. En cuarto lugar, se han descrito las diferentes regularizaciones y restricciones que pueden ser añadidas al modelo de descomposición para incorporar una descripción de las características tiempo-frecuencia de los sonidos presentes en la mezcla, y así dar un sentido más realista al modelo de descomposición. Por último, se han presentado un conjunto de descriptores de tonalidad (dispersión) comúnmente utilizados para realizar un clustering de bases espectrales, en función de su distribución espectral, con el objetivo de mejorar el rendimiento de separación entre fuentes sonoras.

La motivación de utilizar un enfoque de descomposición matricial en el análisis de los sonidos sibilantes se debe principalmente a que, en esta Tesis doctoral, se ha demostrado que sus características tiempo-frecuencia pueden ser modeladas con patrones espectrales. Además, los sonidos respiratorios normales, que siempre se encuentran solapados con los sonidos sibilantes, presentan características distintivas con respecto a las sibilancias, permitiendo así la posibilidad de discriminar ambos tipos de sonidos utilizando regularizaciones. Considerando, que el enfoque NMF nunca antes había sido utilizado en el análisis de los sonidos adventicios, y las posibilidades que este enfoque ofrece para el modelado de señales sonoras, se decidió explotar y evaluar la potencialidad de técnicas NMF en este campo de investigación para contribuir en las

Publicación	Enfoque (NMF/NMPCF)	Regularizaciones	Descriptores (tonalidad)
[P1]	NMF	Dispersión y Suavidad	
[P2]	NMF	Dispersión y Suavidad	
[P3]	NMF	Monofonía	Gini index
[P4]	NMPCF		
[P5]	NMF	Ortogonalidad	Gini index
[P6]	NMF	Dispersión y Suavidad	
[P7]	NMPCF		

Tabla 3.1: Fundamentos en los que se basan los modelos de descomposición matricial propuestos en los trabajos publicados. Las celdas sin texto indican la ausencia de regularizaciones o descriptores en las propuestas.

principales tareas de interés relacionadas con el análisis de los sonidos sibilantes. Es por ello, que todos los trabajos publicados en esta Tesis se fundamentan en este tipo de enfoque. Aunque, cada trabajo publicado puede ser detenidamente analizado en su correspondiente sección, se ha considerado interesante resumir los modelos de descomposición matricial propuestos en cada contribución. Note que únicamente se comentaran los aspectos más relevantes en cuanto al enfoque NMF o NMPCF propuesto. En este sentido, la Tabla 3.1 resume los fundamentos en los que se basa cada modelo, atendiendo al tipo de enfoque de descomposición matricial (NMF o NMPCF), las regularizaciones aplicadas y los descriptores de tonalidad utilizados.

A continuación se describen los aspectos clave de los modelos NMF/NMPCF propuestos en cada trabajo publicado:

- La publicación [P1] propone un modelo de separación basado en un enfoque NMF no supervisado y regularizado. En concreto se aplican las regularizaciones de suavidad y dispersión, para modelar las características tiempo-frecuencia de los sonidos respiratorios normales y los sonidos sibilantes. Esta propuesta inicial confirmó: i) la hipótesis de que ambas fuentes sonoras podían ser separadas, aprovechando las propiedades tiempo-frecuencia que las diferencian; y ii) el potencial del enfoque NMF en el análisis de este tipo de señales sonoras. Por ello, se siguió trabajando con esta técnica en esta línea de investigación.

- La publicación [P2] realiza un proceso de optimización sobre el enfoque NMF no supervisado y regularizado propuesto en [P1] para mejorar el rendimiento de separación de la propuesta inicial. Además, el modelo de separación optimizado es utilizado en la propuesta [P2] para obtener los intervalos temporales en los cuales las sibilancias se encuentran activas.

- El objetivo de la publicación [P3] es detectar la presencia de sonidos sibilantes en señales sonoras respiratorias monocanal. Dentro del esquema del método propuesto se pueden encontrar dos enfoques basados en NMF. En primer lugar, partiendo de un enfoque NMF no supervisado se propone utilizar el descriptor Gini index para clasificar las bases espectrales NMF en dos grupos: bases sibilantes y bases respiratorias. Este modelo en particular, permite estimar el rango espectral, denotado como "Band Of Interest (BOI)", en el cual la probabilidad de que se produzcan sonidos sibilantes es máxima. En segundo lugar, se propone un enfoque semi-supervisado con la regularización de monofonía. Este enfoque utiliza la BOI, obtenida anteriormente, para crear un diccionario de bases que modele el comportamiento tonal de los sonidos sibilantes. Además, la regularización de monofonía es aplicada para minimizar el número de bases espectrales que pueden activarse simultáneamente en el tiempo. Este modelo, permite realizar la separación entre los sonidos sibilantes y respiratorios normales. Note que, aunque el modelo ha sido denominado como semi-supervisado, no depende de ninguna base de datos de entrenamiento (la información necesaria es obtenida de la BOI estimada). Por lo tanto, se podría considerar un enfoque NMF no supervisado (ciego).

- La publicación [P4] presenta una versión extendida del enfoque NMPCF [170] para realizar una separación entre los sonidos sibilantes y respiratorios normales. Específicamente, en [P4] se propone modelar los sonidos respiratorios normales, como patrones espectrales que se repiten a lo largo del tiempo. La principal contribución de la propuesta consiste en variar la importancia de cada segmento en el modelado de los patrones respiratorios, distinguiendo entre segmentos no-sibilantes (sonidos sibilantes inactivos) y segmentos sibilantes (sonidos sibilantes activos), con el objetivo de modelar de manera más precisa los patrones espectrales de sonidos respiratorios normales que son los que se consideran repetitivos temporalmente al aparecer en cada ciclo respiratorio.

- El objetivo de la publicación [P5] es detectar los intervalos temporales en los que las sibilancias se encuentran activas. Dentro del esquema del método propuesto se ha diseñado un modelo de separación (entre sonidos sibilantes y sonidos respiratorios normales) basado en un enfoque NMF no supervisado que aplica la regularización de ortogonalidad para minimizar la redundancia que existe entre las bases espectrales. Además se aplica el descriptor Gini index para distinguir las bases espectrales con mayor probabilidad de ser sibilantes. Considerando lo anterior, se propone un algoritmo recursivo que permite refinar, a lo largo de las iteraciones recursivas, la estimación del espectrograma sibilante. Finalmente, se propone un novedoso criterio de parada que permite minimizar la perdida de contenido de interés sibilante mientras se elimina gran cantidad de contenido respiratorio interferente.

- El objetivo de la publicación [P6] es clasificar el tipo de sibilancia considerando su estructura armónica. Para ello se propone un enfoque NMF regularizado que permite extraer las componentes en frecuencia que caracterizan a las sibilancias. En este sentido, se utilizan

las regularizaciones de suavidad y dispersión para modelar el comportamiento distintivo de los sonidos respiratorios y sibilantes, y un número reducido (low-rank) de bases para compactar las componentes espectrales sibilantes.

- La publicación [P7] propone adaptar el enfoque NMPCF semi-supervisado a un escenario multicanal compuesto por dos canales de entrada monocanal que capturan audio simultáneamente. El primer canal captura el audio de un estetoscopio, el cual se compone de sonidos biomédicos y ruido ambiental que los interfiere. Por otro lado, el segundo canal captura el ruido ambiental que rodea al paciente, mediante un micrófono. El objetivo de esta publicación es mejorar la calidad de los sonidos biomédicos capturados por el estetoscopio, eliminando el ruido ambiente que rodea al paciente.

CAPÍTULO 4

Revisión del estado del arte

El cuarto capítulo presenta un estudio del estado del arte destinado al análisis de los sonidos sibilantes y respiratorios para la mejora del diagnóstico derivado de las patologías pulmonares obstructivas. En términos más específicos, se realiza una revisión del estado del arte atendiendo a los objetivos específicos planteados en esta Tesis: detección de sonidos sibilantes, clasificación del tipo de sibilancia (monofónica/polifónica) y eliminación del ruido ambiente que rodea al sujeto auscultado. Sin embargo, el objetivo específico relacionado con la mejora de audio de los sonidos sibilantes, ha sido una tarea pionera que, hasta donde conoce el autor de esta Tesis, ningún otro autor había desarrollado como tal. Además, se realiza una recopilación de las principales bases de datos, repositorios online y bibliografía que recogen señales sibilantes. Por último, se hace una descripción de las principales métricas utilizadas para medir el rendimiento de los algoritmos propuestos, en función de las diferentes tareas específicas abordadas en esta Tesis.

4.1. Separación de señales sonoras sibilantes

Los sonidos respiratorios normales y los sonidos sibilantes se encuentren mezclados simultáneamente en tiempo y frecuencia. Esto se debe a que ambos sonidos son producidos por las vías respiratorias que componen al árbol bronquial. Por lo tanto, es fisiológicamente imposible escuchar ambos sonidos por separado. Esto origina que los sonidos respiratorios normales, interfieran en la escucha de los sonidos sibilantes de interés, entorpeciendo la capacidad cognitiva del médico al ser confundido por dichos sonidos interferentes, lo que probablemente causará un aumento del número de diagnósticos erróneos, fundamentalmente, del número de falsos ne-

gativos estimados. En este sentido, un estudio reciente [42] ha demostrado que los médicos no consiguen detectar parte de los sonidos adventicios activos en la señal analizada (incluidos los sonidos sibilantes) debido al solapamiento de los sonidos respiratorios durante la inspiración y la espiración. Por lo tanto, la tarea de separar y mejorar la calidad de audio de los sonidos sibilantes en señales respiratorias monocanal, se puede considerar relevante y de interés para los neumólogos.

Pese a la necesidad de abordar esta tarea, hasta donde conoce el autor de esta Tesis, no se ha desarrollado ninguna aportación científica encaminada a realizar una separación entre los sonidos respiratorios normales y los sonidos sibilantes de interés, que mejore la calidad acústica de las sibilancias para que el médico pueda extraer toda la información que las caracterizan sin que ninguna interferencia acústica pueda entorpecer la capacidad cognitiva del médico. En este sentido, las aportaciones científicas derivadas de las publicaciones [P1] y [P4] son consideradas pioneras, respondiendo a la tarea de interés destinada a la separación y mejora sonora de los sonidos sibilantes solapados con los sonidos respiratorios normales.

4.2. Detección de señales sonoras sibilantes

La detección de sonidos sibilantes en señales sonoras respiratorias monocanal, capturadas durante el proceso de auscultación, es sin duda la principal tarea en la que la mayoría de los investigadores en este ámbito científico han centrado su esfuerzo. Esto es debido, a que los sonidos sibilantes son considerados un indicador fiable del grado de obstrucción del árbol bronquial relacionado con algunas de las enfermedades pulmonares de mayor relevancia, como asma, bronquitis aguda, bronquiolitis, bronquiectasia y EPOC [20, 119, 182, 192, 227, 221, 252, 102]. Por lo tanto, una detección rigurosa de las sibilancias puede resultar útil durante el transcurso del primer diagnóstico, resultante del proceso de auscultación, para proporcionar una vía de información complementaria al médico que facilite la decisión tomada en el diagnóstico y evite la aparición de falsos positivos o falsos negativos.

En las últimas décadas se han publicado diferentes algoritmos dedicados a la detección de la presencia o ausencia de sibilancias en señales sonoras respiratorias, o incluso la detección del intervalo temporal en el cual las sibilancias se encuentran activas. A continuación, se presenta un conjunto de los algoritmos más relevantes de detección de sibilancias encontrados en la literatura, los cuales se basan en diferentes técnicas de procesado de señal y diferentes características tiempo-frecuencia (descriptores) aplicadas para distinguir entre sonidos respiratorios normales y sonidos sibilantes.

Los trabajos iniciales, relacionados con la detección de sibilancias, se centraron en el análisis de los picos espectrales y en la aplicación de un proceso de umbralización para distinguir entre ambos sonidos [102, 143, 40, 295, 156]. En este sentido, Taplidou y Hadjileontiadis [295], propusieron un detector espectro-temporal de sibilancias que localiza e identifica automáticamente los sonidos sibilantes basándose en la eliminación de la tendencia espectral (grandes variaciones de amplitud entre muestras son eliminadas y se genera una señal en frecuencia mu-

cho más suave), la separación del espectro en bandas de frecuencia y la detección/clasificación de los picos espectrales. Por otro lado, Jain y Vepa [156], proponen un algoritmo automático para detectar y cuantificar los episodios sibilantes (intervalos temporales), basado en el análisis del espectrograma y una serie de condiciones tiempo-frecuencia, establecidas por la ATS (ver Figura 4.1).

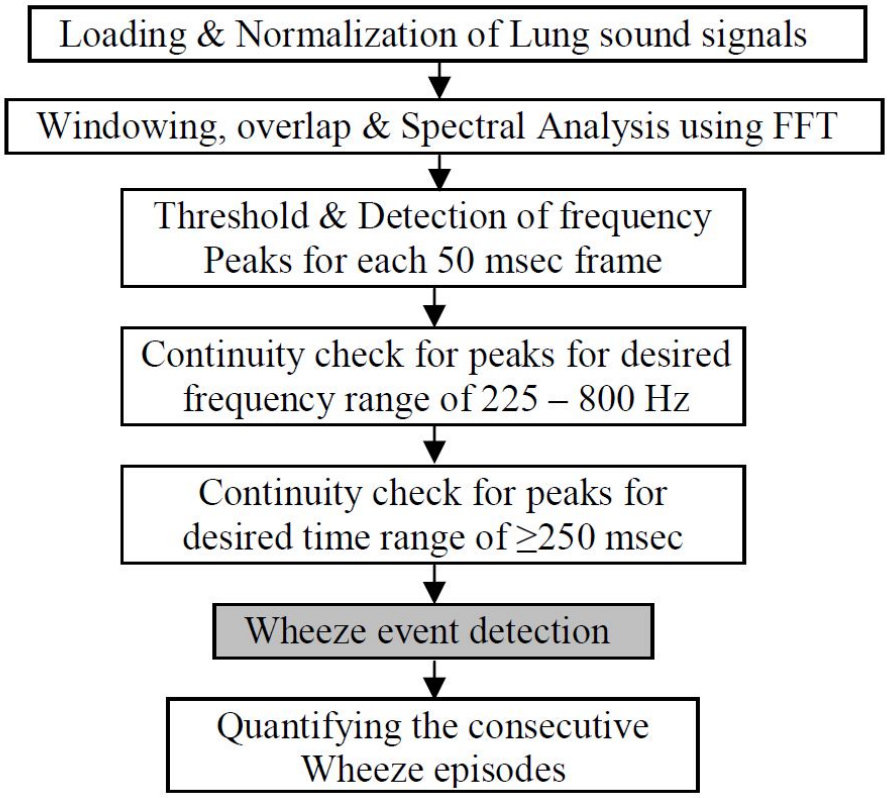

Figura 4.1: Diagrama de bloques del método propuesto en [156].

Sin embargo, la mayoría de enfoques propuestos por los autores en este campo se basan en la extracción de características y la aplicación de clasificadores para identificar los sonidos sibilantes. En concreto, el clasificador "Support Vector Machine (SVM)" ha sido ampliamente utilizado considerando diferentes características espectrales, como por ejemplo: densidad espectral de potencia media y localización de armónicos [59]; intensidad, frecuencia media y frecuencia de la desviación estándar [199]; potencia de la banda espectral [226]; índice de tonalidad [330, 329]; o frecuencia instantánea [198]. En este sentido, Nabi y otros [226], proponen analizar estadísticamente el comportamiento de las sibilancias obtenidas de un conjunto de pacientes asmáticos de diferente gravedad. En concreto, proponen dividir la banda espectral en 20 subbandas para calcular la potencia de cada una de ellas y así caracterizar los sonidos sibilantes. En [330], los autores presentan un método de gran eficacia para la detección automática de sibilancias asmáticas en los sonidos de la respiración. Específicamente, proponen utilizar la fluctuación de la envolvente espectral de audio, denotada en inglés como Audio Spectral

Envelope (ASE), del estándar MPEG-7 y el valor del índice de tonalidad, denotado en inglés como Tonality Index (TI), del estándar MPEG-2, como características discriminatorias para los sonidos sibilantes, junto con un clasificador SVM. Por otro lado, Mendes y otros [218], proponen utilizar características musicales y el modelo de regresión logística, denotado en inglés como Logistic Regression Model (LRM). En concreto, se evaluaron treinta características para identificar al conjunto de mayor rendimiento en la detección de sibilancias. Veintinueve características musicales (centroid, flatness, zerocross, etc.) ampliamente utilizadas en el campo de la restauración musical y una característica relacionada con la forma de la sibilancia detectada.

En particular, los coeficientes cepstrales en las frecuencias de Mel, denotados en inglés como Mel Frequency Cepstral Coefficients (MFCC), han resultado ser la forma más utilizada en la literatura para extraer las características de las componentes de la señal de audio respiratoria con el objetivo de identificar el contenido relevante de los sonidos sibilantes. Los MFCC han sido propuestos en combinación de diferentes enfoques de clasificación: K-Nearest Neighbors (KNN) [281], Gaussian Mixture Models (GMM) [214, 48, 74, 47], Logistic Regression Model(LRM) [253] y Support Vector Machine (SVM) [215]. Shaharum y otros [281], propusieron un detector de sibilancias para clasificar los distintos niveles de gravedad del asma. La propuesta se basa en la extracción de las características MFCC de la señal auscultada en combinación con un método de clasificación basado en el algoritmo KNN. En [215], los autores proponen un sistema de reconocimiento de patrones, compuesto por dos capas, para la detección de sibilancias en pacientes con asma (ver Figura 4.2). La primera capa consiste en el diseño de dos clasificadores SVM en cascada (actuando en paralelo) que utilizan características basadas en los MFCC. La segunda capa aplica un proceso de umbralización para mejorar la fiabilidad y el rendimiento de detección de sibilancias.

Otros estudios han aplicado diferentes enfoques basados en redes neuronales, denotadas en inglés como Neural Networks (NN), en el análisis de los sonidos sibilantes obteniendo un rendimiento fiable [119, 192, 174, 267]. Lin y otros [192], introducen un método que busca los bordes horizontales o casi horizontales del espectrograma y aplica un clasificador NN de retropropagación, denotado en inglés como Back-Propagation Neural Networks (BPNN), utilizando características como, el rango de frecuencia y la pendiente de la sibilancia. En [267], proponen un método basado en la extracción y el procesado de la información espectral del ciclo respiratorio, y en la utilización de esos datos para el reconocimiento automático de las sibilancias (ver Figura 4.3). Primero se preprocesa el ciclo respiratorio, para normalizar su información espectral, y luego se obtiene su espectrograma. Después, la imagen del espectrograma es procesada mediante un filtro de convolución bidimensional y un proceso de umbralización para aumentar el contraste y aislar sus componentes de mayor amplitud, respectivamente. Además, para generar datos más comprimidos, de cara al posterior reconocimiento automático, la proyección espectral del espectrograma procesado es calculada y almacenada en una matriz. Finalmente, Los valores de mayor magnitud de la matriz y sus respectivos espectros se localizan y se utilizan como entradas a una red neuronal artificial de perceptrón multicapa obteniendo como resultado un indicador de la presencia de sibilancias.

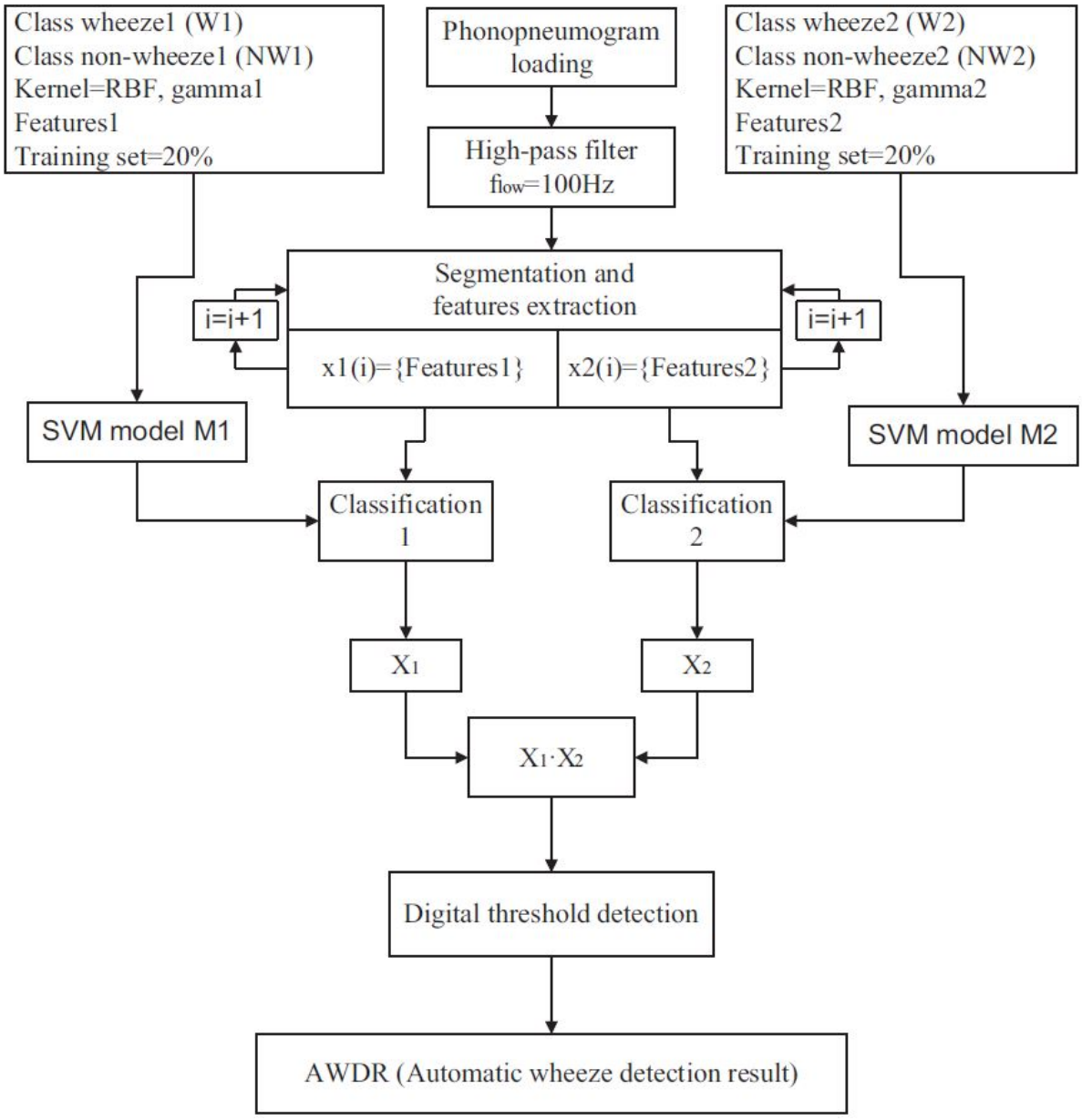

Figura 4.2: Diagrama de bloques del método propuesto en [215].

Otros trabajos se han centrado en aprovechar las posibilidades del dominio de la transformada Wavelet para detectar los sonidos sibilantes [182, 165]. Le y otros [182] proponen un método, para la detección de sibilancias en sonidos respiratorios. Para ello sugieren modelar los sonidos sibilantes y respiratorios normales en el dominio de la transformada Wavelet utilizando la función densidad de probabilidad gaussiana. En particular, proponen realizar la detección (segmentación) de sibilancias utilizando un modelo multimodal Markoviano denominado Hidden Markov Chain (HMC). En conclusión, los autores consiguen demostrar la eficacia del parámetro de forma (shape parameter) de la función densidad de probabilidad gaussiana para discriminar entre sonidos respiratorios normales y sibilantes.

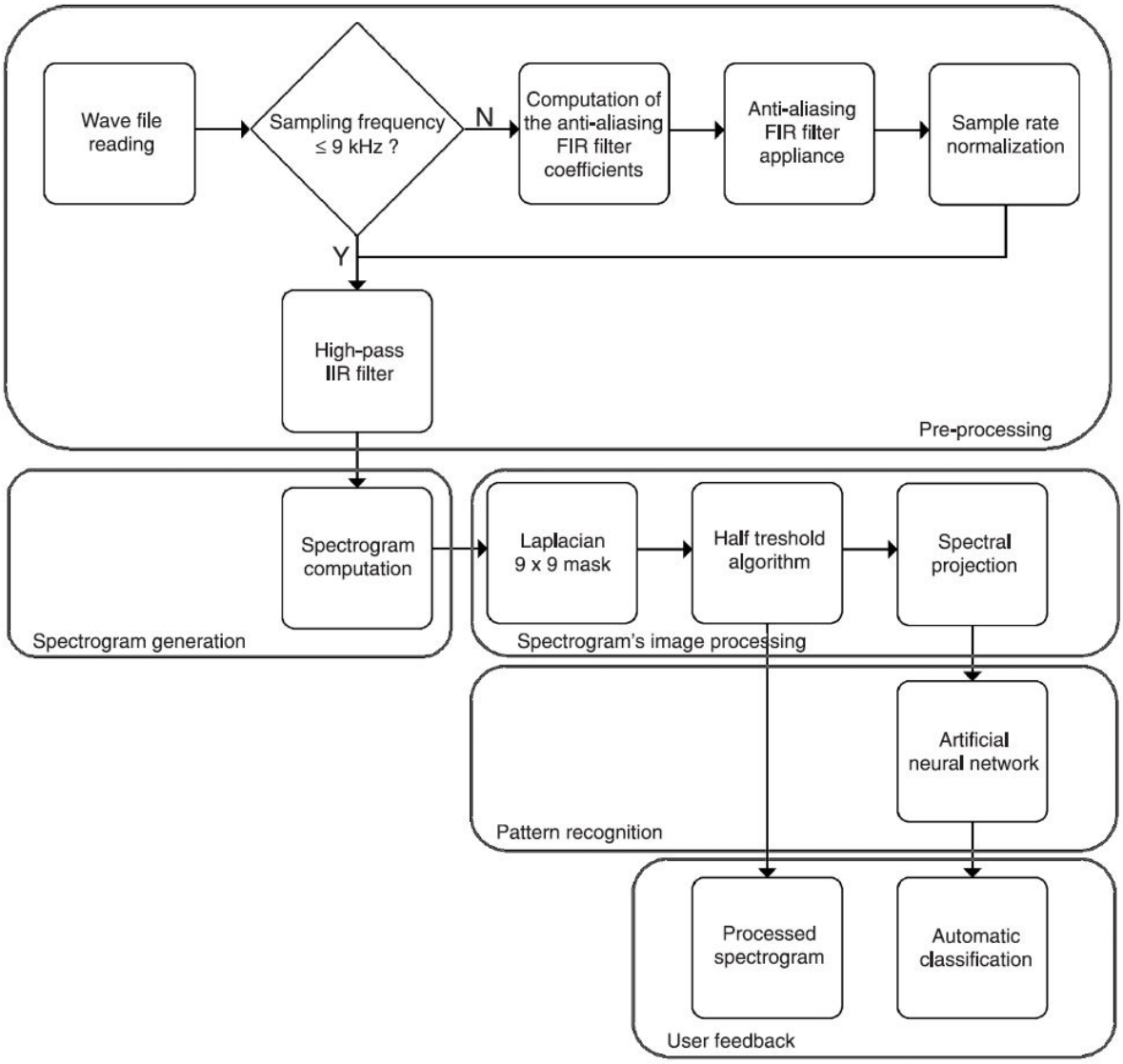

Figura 4.3: Diagrama de bloques del método propuesto en [267].

Para finalizar, también se han encontrado otros métodos para la detección de sibilancias basados en diferentes enfoques, tales como: modelo autorregresivo, denotado en inglés como Autoregressive Model [157, 82]; modelado auditivo, denotado en inglés como Auditory Modelling [256]; principio de entropía [159, 343]; modelo oculto de Márkov, denotado en inglés como Hidden Markov Model (HMM) [233]; análisis discriminativo lineal, denotado en inglés como Linear discriminant analysis (LDA) [44]; o función de longitud de autocorrelación, denotada en inglés como AutoCorrelation Length (ACL) [101]. Qiu y otros [256], presentan un algoritmo para la detección automática de sibilancias basado en la modelización auditiva, denominado algoritmo de umbral dependiente de la frecuencia y la duración. Primero, la frecuencia y duración media de cada componente es obtenida automáticamente. Posteriormente se introduce el concepto de umbral dependiente de la frecuencia y la duración para la detección de

sibilancias. Como mencionan en el estudio, una novedad con respecto a los trabajos basados en umbralización es que el umbral está basado en la potencia correspondiente a una gama de frecuencias determinada, en lugar de a la potencia global de toda la banda. Jin y otros [159], proponen un método de detección de sibilancias utilizando el histograma de la entropía de cada muestra obtenido a partir de señales respiratorias filtradas en banda estrecha. Además, proponen utilizar la distorsión media de los histogramas obtenidos, como rasgo discriminatorio, para detectar los segmentos sibilantes. Oletic y Vilas [233] presentaron un detector de sibilancias basado en la detección iterativa de las trayectorias espectrales que caracterizan a cada sibilancia, mediante la modelización de las mismas utilizando HMM (ver Figura 4.4). La fortaleza de la propuesta se debe a la robustez del rastreo de las líneas de frecuencia, el método de localización temporal de las líneas de frecuencia y la aplicación de la sustracción de líneas.

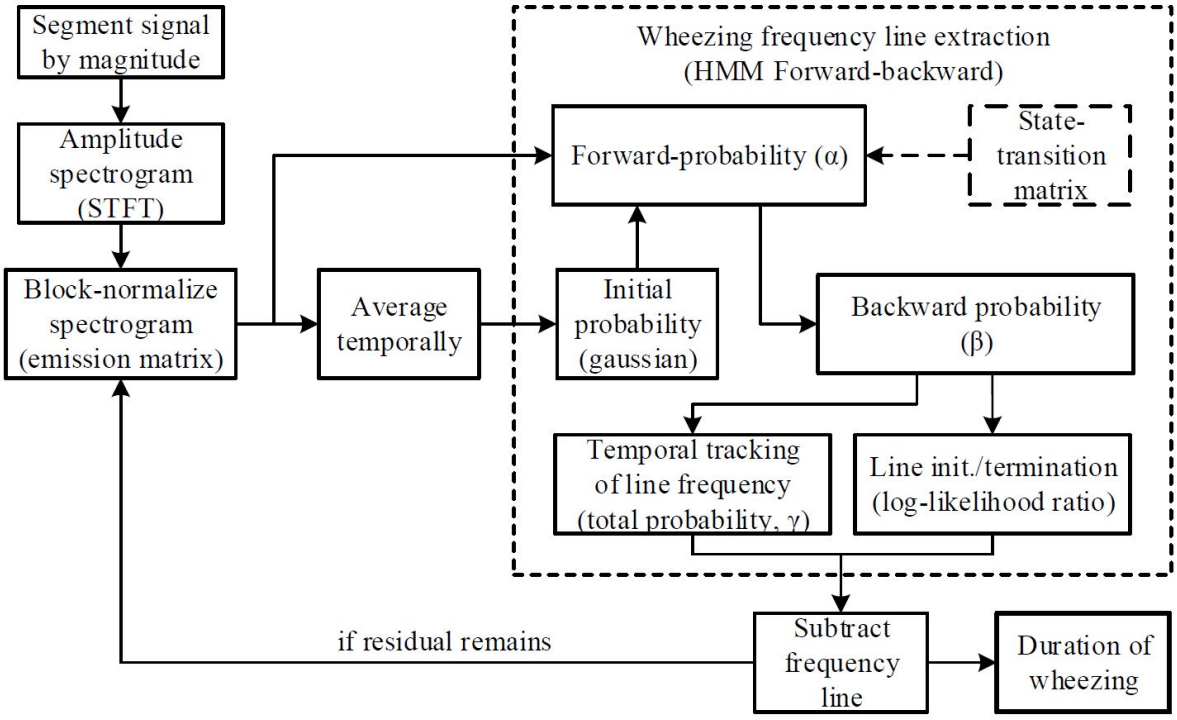

Figura 4.4: Diagrama de bloques del método propuesto en [233].

En general los principales aspectos a mejorar (limitaciones) que se han observado en los métodos de detección de la literatura son los siguientes: i) la mayoría se basan en un enfoque de clasificación que requiere una etapa previa de entrenamiento y cuyo rendimiento depende en gran medida de esta etapa previa; ii) por otro lado, en la mayoría de los métodos suponen que los sonidos sibilantes son de mayor intensidad y sonoridad que los sonidos respiratorios normales con los que acústicamente están mezclados. Con el objetivo de solventar las limitaciones anteriores y de mejorar el rendimiento de los algoritmos propuestos, para así, incrementar su grado de fiabilidad y poder instaurar las TIC como ayuda a los diagnósticos derivados de la auscultación, las publicaciones [P2][P3][P5] proponen novedosos enfoques no supervisados para

detectar la presencia o ausencia de sonidos sibilantes y su localización temporal. Además de no depender de una etapa de entrenamiento previa, los métodos propuestos asumen que los sonidos respiratorios normales pueden ser más audibles que las sibilancias analizadas.

4.3. Clasificación de señales sonoras sibilantes

Como se ha descrito en la Sección 2.3.3, los sonidos sibilantes pueden ser clasificados en dos categorías principales, monofónicos y polifónicos, considerando su comportamiento espectral: i) las sibilancias monofónicas están compuestas por un único pico espectral en banda estrecha (frecuencia fundamental) o por varios picos espectrales en banda estrecha relacionados armónicamente entre sí (frecuencia fundamental junto a sus correspondientes armónicos); y ii) las sibilancias polifónicas están compuestas por un conjunto de picos espectrales (tonos) de banda estrecha no relacionados armónicamente [227]. Con respecto a la fuente que los genera, las sibilancias monofónicas se originan por la obstrucción de una de las vías respiratorias de mayor diámetro, y están relacionadas con el asma [294, 117, 274]. Sin embargo, las sibilancias polifónicas son originadas por una obstrucción múltiple de las vías respiratorias más centrales y de menor diámetro en los pulmones, y están relacionadas con la EPOC [294, 117, 155]. Por lo tanto, distinguir entre sibilancias monofónicas y polifónicas es una tarea crítica en el diagnóstico diferencial del asma [294, 117, 274] y la EPOC [294, 117, 155].

A pesar de los avances en el análisis de los sonidos del sistema respiratorio, la tarea relacionada con la clasificación del tipo de sibilancia resulta ser un problema novedoso, interesante desde el punto de vista médico y no resuelto [227, 305]. En realidad, existen un conjunto relativamente pequeño de trabajos [135, 156, 160, 229, 304, 303, 305] en los cuales el análisis de sibilancias monofónicas y polifónicas es tratado. Sin embargo, únicamente los trabajos propuestos por Ulukaya y otros [304, 303, 305] y Hashemi y otros [135] están directamente relacionados con la clasificación monofónica/polifónica de sibilancias como tal. En este sentido, todos los enfoques de clasificación hasta la fecha se basan en la extracción de características y la aplicación de clasificadores típicos para distinguir entre sibilancias monofónicas y polifónicas.

El trabajo más reciente y relevante [305], propone extraer una única característica, denotada como Peak Energy Ratio (PER), a partir de un tipo de transformada wavelet, denotada como RAtional Dilation Wavelet Transform (RADWT), para discriminar entre sibilancias monofónicas y polifónicas (ver Figura 4.5). Una particularidad interesante de la transformada RADWT es que permite variar la resolución en frecuencia, como también ocurre con la transformada wavelet-packet. En este sentido, la característica PER es obtenida a partir de la energía de las subbandas de los coeficientes wavelet. En concreto, el valor PER se obtiene mediante el primer pico y el segundo pico con mayor energía (considerando que el segundo pico no es consecutivo al primero) de todas las subbandas de los coeficientes wavelet. Además, en el mismo artículo [305], todos los métodos de extracción de características propuestos en la literatura [135, 303, 304] fueron evaluados aplicando clasificadores SVM, KNN y Extreme Learning Machine (ELM). Los resultados de este trabajo concluyeron que el método propuesto en [305]

obtenía las siguientes ventajas: i) solo utiliza una característica (denotada como PER) para realizar la clasificación del tipo de sibilancia, por lo tanto el tiempo computacional en la etapa de clasificación es inferior al del resto de métodos del estado del arte, los cuales utilizan más de una característica; y ii) obtuvo el mejor rendimiento de clasificación para distinguir entre sibilancias monofónicas y polifónicas (con un valor de precisión del 86 %).

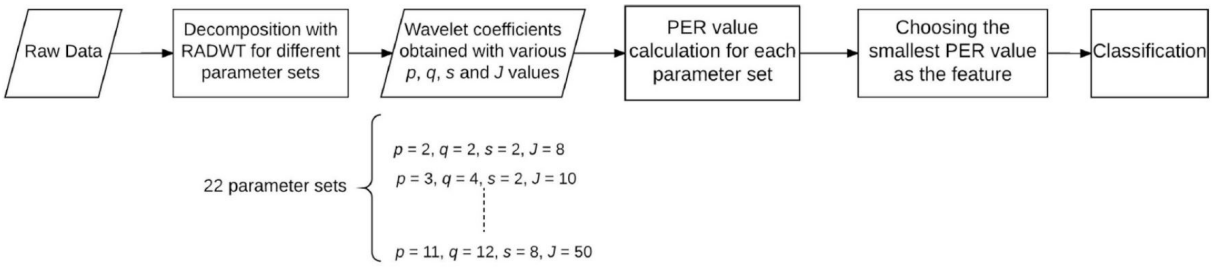

Figura 4.5: Diagrama de bloques del método propuesto en [305].

Considerando la notable escasez de avances en esta línea de investigación, la propuesta [P6] desarrollada en esta Tesis presenta un método eficiente para clasificar el tipo de sonido sibilante (monofónico/polifónico) eliminando el sonido interferente causado por los sonidos respiratorios normales que dificultan la tarea de clasificación.

4.4. Eliminación del ruido ambiente en el proceso de auscultación

Durante el proceso de auscultación aparecen sonidos (ruidos) que interfieren la correcta escucha de los sonidos biomédicos de interés procedentes del interior del cuerpo del sujeto. Estos ruidos varían en función del lugar donde se realice el proceso de auscultación. Por ejemplo, en una consulta es típico que ruidos como, el murmullo de las personas, el llanto de los bebés, el tono del teléfono móvil o el sonido procedente de la calle (coches, camiones, animales, generadores de energía, etc.) interfieran en la escucha de los sonidos biomédicos [99]. Por otro lado, existen espacios donde se generan ruidos que complican considerablemente la labor del médico, por ejemplo el sonido de la sirena de ambulancias [60], si la auscultación se realiza en salas de consulta de urgencias o en la propia ambulancia, o el sonido de las palas en los helicópteros de rescate [148]. En este sentido, varios estudios han demostrado la dificultad que genera la interferencia de estos sonidos durante el proceso de auscultación, disminuyendo drásticamente la capacidad cognitiva del médico para examinar los sonidos biomédicos de interés del paciente [179, 342, 346, 130, 60, 148]. Hasta tal punto que, en escenarios altamente ruidosos, como en el interior de la cabina de un helicóptero “MBB BO-105”, se ha demostrado una precisión del 0 % en la detección de sonidos respiratorios [148].

La primera contribución bien documentada en la que se propuso un sistema para eliminar los efectos del ruido ambiental fue realizado por el ejército de los EE.UU en 1993 [339]. En este estudio intentaron detectar los sonidos cardíacos y pulmonares, utilizando un diseño compuesto

por dos transductores (estetoscopios), en un helicóptero "Bell Jet Ranger 206B" en marcha. Para ello uno de los transductores se colocó en una de las zonas típicas de auscultación para obtener los sonidos de interés (sonidos cardíacos y pulmonares), y el otro transductor se colocó en otro lugar del pecho del paciente para obtener una señal de referencia del ruido ambiente con una alta fidelidad. Utilizando un algoritmo clásico de filtrado adaptativo, en concreto el enfoque "Least-Mean-Squares (LMS)", consiguieron una reducción del ruido ambiental de unos 30 dB entre los 20 y los 100 Hz, y una reducción de 10 a 20 dB entre los 100 y 300 Hz. Tras esta contribución muchos investigadores encontraron una línea atractiva y necesaria en la que participar.

Una revisión del estado del arte ha mostrado que los investigadores en esta línea de investigación se centran principalmente en optimizar el diseño de un estetoscopio (Hardware) que permita mejorar la eliminación del ruido ambiental. En este sentido, en la literatura se pueden encontrar un amplio conjunto de patentes en las que se proponen diferentes diseños para hacer frente a esta problemática [134, 234, 275, 72, 81, 94, 144, 231]. Por ejemplo, Orten [234] patentó un estetoscopio electrónico que incluye un micrófono, un equipo de amplificación y un altavoz para el usuario. Además, comprende un circuito de filtrado ajustable, el cual tiene un efecto paso banda, donde la frecuencia central y el ancho de banda pueden ser ajustados a voluntad por el usuario en todo el rango audible. Por otro lado, Houtsma [144] patentó un estetoscopio con características de supresión de ruido que lo hacen adecuado para su uso en ambientes altamente ruidosos. La principal contribución de esta patente, fue la inclusión de una barrera de supresión en el receptor auscultador (Chest-piece) para evitar que las ondas superficiales resultantes del ruido ambiental lleguen al transductor, que recoge los sonidos biomédicos auscultados, proporcionando así una señal de mayor calidad para ser procesada y a su vez escuchada por el profesional médico. Recientemente, Nieminen [231] ha presentado una patente que proporciona dispositivos y métodos en los que las señales de auscultación, las señales de ultrasonido y las señales de ruido ambiental pueden ser adquiridas simultáneamente por un mismo dispositivo portátil para proporcionar la cancelación del ruido de fondo. Desde un punto de vista comercial, la necesidad de actuar ante el ruido ambiental, ha desencadenado una batalla en la que las distintas compañías dedicadas a la fabricación de estetoscopios (3M Littmann, Eko, Ekuore, etc.) compiten para ofrecer el mejor rendimiento en la eliminación del ruido ambiental que interfiere durante la auscultación. Como se ha comentado en la Sección 2.2.4, algunos de los estetoscopios digitales que ofrecen mejores prestaciones, incluyendo la cancelación del ruido de fondo, son "3M Littmann Electronic Stethoscope Model 3200" de la compañía 3M Littmann, y "CORE Digital Stethoscope" de la compañía Eko.

Aunque en este ámbito se han publicado diferentes enfoques y métodos para eliminar el ruido ambiental activo durante el proceso de auscultación, la mayoría de dichos métodos [111, 293, 248, 89, 230, 53, 318, 201] se basan en la técnica clásica de filtrado adaptativo [136]. Así, Patel y otros [248], propusieron un esquema de filtrado adaptativo basado en los enfoques "Least-Mean-Squares (LMS)" y "Normalized Least-Mean-Squares (NLMS)" para extraer los sonidos respiratorios de la señal biomédica auscultada afectada por el ruido ambiental. En su propuesta se lograron aproximadamente 15 dB de reducción de ruido en el rango de frecuencia de 100-600 Hz con el algoritmo LMS, mientras que el algoritmo NLMS, más complejo, propor-

cionó una convergencia más rápida y hasta 5 dB de reducción adicional de ruido. Por otro lado, Wang y otros [318], propusieron una extensión a los sistemas tradicionales de cancelación de ruido basados en filtrado adaptativo. Para ello, diseñaron un sistema multi-sensor que permite detectar los ciclos respiratorios para adaptar el sistema de cancelación de ruido a cada uno de los ciclos respiratorios. Aparte de los algoritmos basados en filtrado adaptativo, Emmanouilidou y otros, presentaron dos métodos [99, 100], basados en sustracción espectral multibanda, denotado en inglés como Multiband Spectral Subtraction (MSS), con el objetivo de mejorar la calidad acústica de los sonidos pulmonares durante la auscultación pediátrica en entornos ruidosos y maximizar la fiabilidad del diagnóstico emitido. En el trabajo [99], los autores proponen un esquema multibanda, basado en la configuración de dos micrófonos, para suprimir el ruido de fondo preservando al mismo tiempo el contenido sonoro procedente del sistema respiratorio. Concretamente, el algoritmo analiza cada banda espectral de manera no uniforme y utiliza información acerca de la probabilidad de la presencia de los sonidos de interés (sonidos respiratorios) en dichas bandas para aplicar una penalización en aquellas bandas espectrales donde la probabilidad de aparición de sonidos de interés es menor.

Se ha observado que esta línea de investigación no cuenta con un objetivo común en las diferentes publicaciones encontradas en la literatura por el autor de esta Tesis, ya que mientras que algunos trabajos se centran en mejorar la calidad acústica de los sonidos respiratorios [293, 318], otros trabajos se centran en mejorar la calidad de los sonidos cardíacos [89, 53]. Incluso algunos trabajos se centran en escenarios ruidosos específicos, como la auscultación dentro de una ambulancia [201] o helicóptero de rescate [230]. Además, otros trabajos se centran en ramas específicas de la medicina pulmonar, como es el caso de los trabajos presentados por Emmanouilidou y otros [99, 100], los cuales están orientados a la auscultación pediátrica. Esto supone una dificultad para comparar las diferentes propuestas de forma equitativa, ya que cada trabajo está particularizado para una tarea específica, aunque todos tengan el mismo objetivo general (eliminar el ruido ambiente que rodea al paciente).

Considerando el interés encontrado en la tarea de eliminar el ruido ambiental que rodea al paciente tanto desde el punto de vista médico como comercial, la propuesta [P7] desarrollada en esta Tesis, presenta un modelo multicanal (micrófono externo y estetoscopio) basado en un enfoque NMPCF para eliminar el ruido ambiental que interfiere a los sonidos biomédicos auscultados. A diferencia de la mayoría de propuestas de la literatura, las cuales se basan en el análisis diferencial de las bandas espectrales que distinguen los sonidos de interés del ruido ambiental, la propuesta [P7] trata de modelar los patrones espectrales interferentes que se detectan simultáneamente en ambos canales de entrada para eliminar los sonidos que interfieren en la escucha de los sonidos biomédicos procedentes del interior del cuerpo humano del sujeto. Una fortaleza destacable de la propuesta presentada en esta Tesis es que aunque el método elimina la mayor parte del ruido ambiental, es capaz de preservar el contenido sonoro de los sonidos que fueron emitidos en el interior del cuerpo del sujeto, como por ejemplo, sonidos respiratorios normales, sonidos respiratorios adventicios (sibilancias), sonidos cardíacos, etc. En otras palabras, el método propuesto no está particularizado para ningún tipo de señal biomédica o escenario ruidoso específico.

4.5. Bases de datos

El campo de investigación relacionado con el análisis de los sonidos sibilantes presenta una importante problemática relacionada con la carencia de bases de datos de sonidos sibilantes estandarizadas. Realmente, esta problemática puede ser extendida al análisis de cualquier sonido adventicio en general. Esto supone que cada autor utilice una base de datos propia de sonidos sibilantes, y por ello en cada trabajo se utiliza una base de datos distinta a la del resto de propuestas. En consecuencia, la evaluación del rendimiento de diferentes métodos del estado del arte es alcanzada gracias a la implementación de los diferentes algoritmos y a su posterior evaluación con una base de datos propia.

Asumiendo esta problemática, y tras realizar una revisión exhaustiva de la literatura se ha concluido que los autores, en el campo del análisis de los sonidos sibilantes, diseñan las bases de datos siguiendo tres metodologías posibles:

- Por un lado, algunos autores crean sus bases de datos a partir de señales sonoras respiratorias obtenidas de diferentes pacientes con patologías obstructivas pulmonares y que presentan sonidos sibilantes. Considerando el aspecto médico, es la forma más efectiva, fiable y robusta de construir una base de datos de sonidos sibilantes. Sin embargo, no todos los autores cuentan con un equipo médico que les ofrezca esta posibilidad. Además, la creación de bases de datos a partir de pacientes reales es un proceso lento, ya que por un lado se debe contar con el consentimiento del paciente, por otro lado, es necesario realizar múltiples grabaciones hasta que los sonidos sibilantes son generados y, por último, cabe destacar el enorme tiempo que el médico debe invertir para analizar y detectar los sonidos sibilantes.

- Por otro lado, los autores que no pueden optar a la posibilidad anterior, crean sus propias bases de datos a partir de repositorios online de sonidos sibilantes y libros que incluyen sonidos sibilantes como recurso adicional. En este sentido, el resto de la sección presenta los principales repositorios online y la bibliografía más utilizada por los autores en este campo de investigación para la creación de bases de datos sibilantes. Además, se describe una base de datos de sonidos sibilantes y crepitantes que ha surgido recientemente, de la cual algunos autores han extraído algunas sibilancias para componer sus propias bases de datos.

- Por último, la tercera alternativa es la compartición de bases de datos entre autores, incluyendo en la publicación la correspondiente referencia al trabajo que define la base de datos en cuestión y los agradecimientos a los autores que han facilitado el uso de dicha base de datos. En este sentido, los autores Dr. Dinko Oletic y Dr. Vedran Bilas nos dieron la posibilidad de emplear la base de datos que utilizaron en [233] para incluirla en la evaluación del rendimiento de los algoritmos propuestos en esta Tesis doctoral.

4.5.1. Base de datos ICBHI

Recientemente ha sido publicada [269] una base de datos de sonidos respiratorios compuesta principalmente de sonidos respiratorios normales, sonidos sibilantes y sonidos crepitantes. Esta base de datos fue creada originalmente para apoyar el desafío científico organizado por la conferencia internacional "Int. Conf. on Biomedical Health Informatics (ICBHI) 2017". Desde el año 2019, el conjunto de datos públicos y privados del desafío ICBHI está disponible de forma gratuita [16] y cada vez hay más investigadores que lo están utilizando en el análisis de sonidos respiratorios en sus trabajos más recientes (y no sólo en el ámbito del análisis de sibilancias) [121, 122, 175, 90]. Como se describe en [269], la base de datos ICBHI se construyó con el objetivo de dar soporte a las contribuciones científicas dedicadas a la clasificación de sonidos respiratorios (sonidos respiratorios normales, sonidos sibilantes y sonidos crepitantes) y de eliminar la carencia de una base de datos estandarizada.

La base de datos ICBHI contiene muestras de audio, recogidas de forma independiente por dos equipos de investigación, a lo largo de varios años. La mayor parte de la base de datos está compuesta de audios grabados por el equipo de investigación de la Facultad de Ciencias de la Salud de la Universidad Aveiro (ESSUA), recogidos en el Laboratorio de Investigación y Rehabilitación Respiratoria (Lab3R) de la ESSUA y en el Hospital Infante D. Pedro, Aveiro (Portugal). El segundo equipo de investigación, compuesto por la Universidad Aristóteles de Salónica (AUTH) y la Universidad de Coímbra (UC), adquirió sonidos respiratorios en el Hospital General de Papanikolaou, Salónica (Grecia) y en el Hospital General de Imathia (Unidad de Salud de Naousa), Grecia.

La base de datos está compuesta por 5,5 horas de sonidos respiratorios, con un total de 6898 ciclos respiratorios (un ciclo respiratorio consta de la etapa de inspiración y espiración), de los cuales 1864 contienen crepitaciones, 886 contienen sibilancias y 506 contienen tanto crepitaciones como sibilancias. Dichos sonidos fueron obtenidos a partir de 920 grabaciones de audio de 126 sujetos diferentes. Las grabaciones fueron recogidas utilizando un equipamiento variado compuesto por tres estetoscopios (3M Littmann Classic II SE Stethoscope, 3M Litmmann 3200 Electronic Stethoscope y WelchAllyn Meditron Master Elite Electronic Stethoscope) y un micrófono de condensador (AKG C417L Microphone), y su duración oscila entre los 10 y los 90 segundos. Esta base de datos también indica las ubicaciones torácicas de las que se adquirieron las grabaciones. Indicar que los ciclos respiratorios fueron etiquetados por médicos, especialistas en este ámbito, los cuales determinaron, para cada ciclo respiratorio, si existían sibilancias, crepitaciones, una combinación de ellas o ningún sonido respiratorio adventicio. Además de la información demográfica de los pacientes (edad, sexo, peso, altura e IMC), la base de datos también incluye el diagnóstico para cada uno de ellos.

Esta base de datos ha sido creada para evaluar escenarios acústicos típicos que aparecen en el mundo real, por ello en las señales sonoras respiratorias se puede escuchar diferentes tipos de sonidos que actúan como ruido ambiental, tales como murmullo de las personas, llanto de los niños, etc.

4.5.2. Repositorios online de sonidos respiratorios

Existe un conjunto de repositorios online que proporcionan acceso a diferentes tipos de sonidos respiratorios, ya sean sonidos respiratorios normales o sonidos respiratorios adventicios, entre ellos sibilancias. Sin duda, esta es la opción más rápida que cualquier investigador, centrado en la línea del análisis de los sonidos sibilantes, puede llevar a cabo para construir su propia base de datos. La Tabla 4.1 presenta los principales repositorios de sonidos respiratorios que los investigadores suelen utilizar en esta línea. Entre estos repositorios podemos encontrar empresas dedicadas al diseño de estetoscopios (Thinklabs, Littmann o Stethographics), software dedicados al procesado de señales respiratorias (RALE), escuelas de enfermería y medicina (NursingCenter o Colorado State University), etc.

Repositorios online de sonidos respiratorios
R.A.L.E repository [31]
Stethographics lung sound samples [30]
3m Littmann stethoscopes [3]
East tennessee state university pulmonary breath sounds [8]
Lippincott Nursing Center [18]
Thinklabs Digital Stethoscope [32]
Thinklabs youtube [34]
Emedicine/Medscape [13]
E-learning resources [7]
Respiratory wiki [27]
Easy Auscultation [9]
SoundCloud. Lung Sounds [29]
Colorado State University [4]

Tabla 4.1: Repositorios online de sonidos respiratorios. En estos repositorios se pueden encontrar tanto sonidos respiratorios normales, como sonidos respiratorios adventicios (sibilancias, crepitaciones, etc).

4.5.3. Bibliografía especializada en sonidos adventicios

Por último, podemos encontrar un conjunto de libros diseñados para la docencia de los sonidos que emite el sistema respiratorio humano. Estos libros vienen acompañados con un CD-ROM o en su defecto, dan acceso a un espacio virtual en el que se incluyen, a modo de ejemplo, diferentes señales de sonidos respiratorios normales y adventicios. Aunque estos libros no son gratuitos, puede ser otra opción para el diseño de bases de datos compuestas por sonidos sibilantes. La Tabla 4.2 muestra una lista de los principales libros que, entre los diferentes tipos de sonidos respiratorios, contienen ejemplos de sonidos sibilantes.

Bibliografía especializada en sonidos adventicios
Understanding Lung Sounds, the 2^{nd} edition [187]
Understanding Lung Sounds, the 3^{rd} edition [188]
Understanding Heart Sounds and Murmurs: With An Introduction to Lung Sounds [299]
Auscultation Skills: Breath & Heart Sounds [83]
Fundamentals of Lung and Heart Sounds [326]
Heart and Lung Sounds Reference Library [333]
Lung Sounds: A Practical Guide [325]
Secrets Heart & Lung Sounds Workshops [206]
Respiratory Physiology: A Clinical Approach [277]
Lung Sounds: An Introduction to the Interpretation of the Auscultatory Finding [178]
The Chest: Its Signs and Sounds [93]

Tabla 4.2: Bibliografía especializada en sonidos adventicios. En los repositorios de estos libros se pueden encontrar tanto sonidos respiratorios normales, como sonidos respiratorios adventicios (sibilancias, crepitaciones, etc).

Las bases de datos utilizadas en los diferentes trabajos propuestos en esta Tesis han sido construidas utilizando sibilancias de la base de datos ICBHI, de los repositorios online descritos anteriormente y de la base de datos [233] compartida por los autores Dr. Dinko Oletic y Dr. Vedran Bilas. En cada una de las publicaciones podrá encontrar una descripción detallada de las bases de datos construidas para la evaluación de los métodos propuestos.

4.6. Métricas de evaluación

Como en cualquier campo de investigación, el establecimiento de una metodología de evaluación adecuada es esencial para evaluar la fiabilidad de las soluciones propuestas. En esta sección se describen las principales métricas objetivas utilizadas para medir el rendimiento de los algoritmos dedicados a afrontar las principales tareas relacionadas con el análisis de sonidos sibilantes: separación, detección y clasificación de sibilancias, y eliminación del ruido ambiente que rodea al paciente durante el proceso de auscultación.

4.6.1. Métricas de separación

Como se ha mencionado anteriormente, no se han encontrado trabajos centrados exclusivamente en la separación entre sonidos sibilantes y sonidos respiratorios normales, para mejorar la calidad acústica de las sibilancias. Sin embargo, la tarea relacionada con la separación de fuentes sonoras ha sido ampliamente abordada en diferentes campos (por ejemplo, separación de fuentes musicales [64, 312]). En este sentido, Vicent y otros [310], presentaron una serie de métricas objetivas para medir el rendimiento de los algoritmos dedicados a la separación de

fuentes sonoras. Incluso han desarrollado una Toolbox en MATLAB, denominada "BSS EVAL toolbox" [106], para medir el rendimiento de las propuestas en términos de separación. Esto ha generado que estas métricas se hayan estandarizado en el campo de la separación de fuentes sonoras [62, 64, 312].

Con el objetivo de centrarnos en los algoritmos de separación de fuentes sonoras sibilantes y respiratorias normales, objeto de esta Tesis, es importante clarificar que las métricas, que se van a definir a continuación, son obtenidas a partir de las señales de audio originales $s_W(n)$, $s_R(n)$ y de las señales de audio estimadas $\hat{s}_W(n)$, $\hat{s}_R(n)$ por el algoritmo de separación. En concreto, $s_W(n)$, $\hat{s}_W(n)$ corresponden a la señal original y estimada de los sonidos sibilantes, respectivamente, y $s_R(n)$, $\hat{s}_R(n)$ corresponden a la señal original y estimada de los sonidos respiratorios normales, respectivamente.

Dicho esto, las métricas propuestas para medir el rendimiento de separación son: 1) Source-to-distortion ratio (SDR), que proporciona información sobre la calidad general del proceso de separación; 2) Source-to-interferences ratio (SIR), que informa de la presencia de interferencias contenidas en la fuente de interés. Por ejemplo, en el caso de la separación de sonidos sibilantes y respiratorios normales, esta métrica indica la presencia de sonidos sibilantes contenidos en la señal respiratoria estimada y viceversa; y 3) Source-to-artifacts ratio (SAR), que proporciona información sobre los artefactos en las señales estimadas procedentes del algoritmo de separación o de la resíntesis de la señal estimada. El principio para obtener el valor de estas métricas es descomponer el error total, entre la señal estimada $\hat{s}_i(n)$ y la señal original $s_i(n)$, en tres términos relacionados con tres tipos de error, como se muestra a continuación:

$$\hat{s}_i(n) - s_i(n) = {e_i}^{interf}(n) + {e_i}^{artifacts}(n) + {e_i}^{spatial}(n) \tag{4.1}$$

donde ${e_i}^{interf}(n)$ es el término de error relacionado con la interferencia producida por la fuente sonora no deseada; ${e_i}^{artifacts}(n)$ es el término de error atribuido a los artefactos generados por el algoritmo de separación; y ${e_i}^{spatial}(n)$ es el término de error correspondiente a la distorsión espacial. Además, el término i identifica la fuente sonora deseada, en el caso de la separación de los sonidos sibilantes y respiratorios normales, $i = W$ para referirse a la fuente sonora sibilante o $i = R$ para referirse a la fuente sonora respiratoria. Utilizando estos términos de error las métricas SDR_i, SIR_i y SAR_i, para ambas fuentes sonoras $i = W$ o $i = R$, pueden ser obtenidas, expresadas en decibelios (dB), como se muestra a continuación:

$$SDR_i(dB) = 10\log_{10}\left(\frac{\|s_i(n)\|^2}{\|{e_i}^{interf}(n) + {e_i}^{artifacts}(n) + {e_i}^{spatial}(n)\|^2}\right) \tag{4.2}$$

$$SAR_i(dB) = 10\log_{10}\left(\frac{\left\|s_i(n) + {e_i}^{interf}(n) + {e_i}^{spatial}(n)\right\|^2}{\|{e_i}^{artifacts}(n)\|^2}\right) \tag{4.3}$$

$$SIR_i(dB) = 10\log_{10}\left(\frac{\|s_i(n)\|^2}{\|{e_i}^{interf}(n)\|^2}\right) \tag{4.4}$$

donde, cuanto mayor sea el valor de estas métricas mejor será el rendimiento de separación del algoritmo evaluado. Sin embargo, es fundamental que cada una de las métricas presente valores

apropiados. Por ejemplo, en el caso de que los valores de las métricas SDR_W, SAR_W sean altos ($SDR_W = 10$ dB y $SAR_W = 12$ dB) no se garantiza que el rendimiento de separación del algoritmo sea robusto. En concreto, estos valores indican una buena calidad de la señal sibilante estimada, preservando el contenido de las sibilancias, y una presencia mínima de artefactos procedentes del algoritmo de separación. Sin embargo, a pesar de que las métricas SDR_W, SAR_W hayan obtenido buenos resultados, puede ocurrir que el valor de la métrica SIR_W sea bajo (por ejemplo, $SIR_W = -2{,}5$ dB), indicando la presencia de sonidos interferentes procedentes de la señal respiratoria.

En concreto, estas métricas han sido utilizadas para evaluar los algoritmos de separación propuestos en las publicaciones [P1][P4] y en la etapa de optimización del algoritmo de separación propuesto en la publicación [P2].

4.6.2. Métricas de detección

Esta sección se centra en describir las métricas de detección más utilizadas en la línea del análisis de los sonidos sibilantes. Como se ha mencionado anteriormente, la línea dedicada a detectar los sonidos sibilantes puede ser dividida en dos tareas más específicas: i) detectar la presencia o ausencia de sonidos sibilantes; y ii) detectar (localizar) los intervalos temporales en los cuales las sibilancias están presentes. En este sentido, la publicación [P3] propone un método para detectar la presencia o ausencia de sonidos sibilantes. Por otro lado, las publicaciones [P2][P5], proponen un método para detectar o localizar los intervalos temporales en los cuales las sibilancias se encuentran activas dentro del ciclo respiratorio.

Con el objetivo de presentar una descripción de las métricas más utilizadas en ambas tareas de detección, esta sección se ha dividido en dos partes:

Métricas para los algoritmos que detectan la presencia o ausencia de sonidos sibilantes

Estas métricas permiten evaluar el rendimiento de detección de aquellos algoritmos destinados a detectar la presencia o ausencia de sibilancias en señales sonoras respiratorias. En otras palabras, aquellos algoritmos que proporcionan una discriminación o una clasificación entre señales respiratorias de sujetos sanos (sin sonidos sibilantes) y enfermos (con sonidos sibilantes). En concreto, las métricas más relevantes [198, 252] para medir la capacidad de discriminar entre sujetos sanos y enfermos son las siguientes: 1) Sensitivity (SE), la capacidad de clasificar correctamente sujetos enfermos analizando solo señales correspondientes a sujetos enfermos; 2) Specificity (SP), la capacidad de clasificar correctamente sujetos sanos analizando solo señales correspondientes a sujetos sanos; 3) Positive Predictive Value (PPV), la probabilidad de que un sujeto clasificado como enfermo sea enfermo evaluando las señales de sujetos sanos y enfermos; 4) Negative Predictive Value (NPV), la probabilidad de que un sujeto clasificado como sano sea sano evaluando las señales de sujetos sanos y enfermos; y 5) Accuracy (ACC), la capacidad de clasificar correctamente a un paciente como sano o enfermo.

$$SE(\%) = \frac{TP}{TP + FN} \tag{4.5}$$

$$SP(\%) = \frac{TN}{TN + FP} \tag{4.6}$$

$$PPV(\%) = \frac{TP}{TP + FP} \tag{4.7}$$

$$NPV(\%) = \frac{TN}{TN + FN} \tag{4.8}$$

$$ACC(\%) = \frac{(TP + TN)}{(TP + FP + FN + TN)} \tag{4.9}$$

donde, TP (True Positive) representa el número de sujetos enfermos correctamente clasificados, TN (True Negative) representa el número de sujetos sanos correctamente clasificados, FP (False Positive) representa el número de sujetos sanos erróneamente clasificados como sujetos enfermos, y FN (False Negative) representa el número de sujetos enfermos erróneamente clasificados como sujetos sanos.

Métricas para los algoritmos que detectan los intervalos temporales sibilantes

Estas métricas permiten evaluar el rendimiento de detección de aquellos algoritmos destinados a detectar los instantes temporales en los cuales las sibilancias se encuentran activas, dentro del ciclo respiratorio. En este sentido las métricas ACC, SE y SP, han sido ampliamente utilizadas en el campo de la detección temporal de las sibilancias [233, 297, 59, 281, 215]. Sin embargo, a diferencia de las métricas definidas en la sección anterior, las cuales son obtenidas a nivel de señal respiratoria, las métricas de esta sección se obtienen a nivel de trama (frame). Es decir, estas métricas analizan las tramas temporales que se han detectado correctamente o erróneamente. En este sentido, el conjunto de métricas (SE, SP y ACC) se define como: 1) Sensitivity (SE), la probabilidad de detectar tramas temporales sibilantes correctamente (ver Ec. (4.5)); 2) Specificity (SP), la probabilidad de detectar tramas temporales respiratorias normales (sin sibilancias) correctamente (ver Ec. (4.6)); y 3) Accuracy (ACC), la probabilidad de detectar tanto las tramas temporales sibilantes como las respiratorias normales correctamente (ver Ec. (4.9)).

Es importante considerar que, a diferencia de las métricas de la sección anterior, TP (True Positive) representa el número de tramas sibilantes detectadas correctamente, TN (True Negative) representa el número de tramas respiratorias normales detectadas correctamente, FP (False Positive) representa el número de tramas respiratorias normales erróneamente detectadas como sibilantes, y FN (False Negative) representa el número de tramas sibilantes erróneamente detectadas como respiratorias normales.

4.6.3. Métricas de clasificación

Como se ha mencionado anteriormente, muy pocos trabajos han propuesto algoritmos destinados a la clasificación de sibilancias entre monofónicos y polifónicos. El trabajo más reciente y relevante [305], propone utilizar la tasa de precisión ACC para medir el rendimiento de los

algoritmos de clasificación de sibilancias. En la publicación [P6], se propone utilizar diferentes tasas de ACC para proporcionar una evaluación más rigurosa, considerando tanto las dos posibles formas de sibilancias monofónicas que se pueden presentar (sibilancia monofónica con un solo pico espectral o sibilancia monofónica compuesta por la frecuencia fundamental y sus armónicos correspondientes), como las sibilancias polifónicas (varios picos espectrales no relacionados armónicamente). En este sentido, las siguientes tasas de precisión ACC se proponen en [P6]: 1) ACC_G, la capacidad de clasificar correctamente sibilancias monofónicas y polifónicas; 2) ACC_M, la capacidad de clasificar correctamente sibilancias monofónicas; 3) ACC_P, la capacidad de clasificar correctamente sibilancias polifónicas; 4) ACC_{M1}, la capacidad de clasificar correctamente sibilancias monofónicas tipo 1 (sibilancia monofónica con un solo pico espectral); y ACC_{M2}, la capacidad de clasificar correctamente sibilancias monofónicas tipo 2 (sibilancia monofónica compuesta por la frecuencia fundamental y sus armónicos correspondientes).

$$ACC_G(\,\%) = \frac{(TP + TM)}{(TP + TM + FP + FM))} \tag{4.10}$$

$$ACC_P(\,\%) = \frac{TP}{(TP + FP)} \tag{4.11}$$

$$ACC_M(\,\%) = \frac{TM}{(TM + FM)} \tag{4.12}$$

$$ACC_{M1}(\,\%) = \frac{TM1}{(TM1 + FM1)} \tag{4.13}$$

$$ACC_{M2}(\,\%) = \frac{TM2}{(TM2 + FM2)} \tag{4.14}$$

donde, TP (True Polyphonic) representa el número de sibilancias polifónicas correctamente clasificadas, TM (True Monophonic) representa el número de sibilancias monofónicas correctamente clasificadas, FP (False Polyphonic) representa el número de sibilancias polifónicas erróneamente clasificadas como monofónicas, FM (False Monophonic) representa el número de sibilancias monofónicas erróneamente clasificadas como polifónicas, $TM1$ (True Monophonic type 1) representa el número de sibilancias monofónicas tipo 1 correctamente clasificadas, $TM2$ (True Monophonic type 2) representa el número de sibilancias monofónicas tipo 2 correctamente clasificadas, $FM1$ (False Monophonic type 1) representa el número de sibilancias monofónicas tipo 1 erróneamente clasificadas como polifónicas, y $FM2$ (False Monophonic type 2) representa el número de sibilancias monofónicas tipo 2 erróneamente clasificadas como polifónicas.

4.6.4. Métricas de eliminación del ruido ambiente

Las métricas SDR, SIR y SAR, descritas en la Sección 4.6.1, para medir el rendimiento de separación de los algoritmos de separación de fuentes, pueden ser utilizadas para medir el rendimiento de los algoritmos orientados a la eliminación de ruido [211]. Esto es posible,

tratando a la señal de interés como una fuente sonora y al ruido ambiental como otra fuente sonora distinta. De esta forma, al igual que en la Sección 4.6.1, los algoritmos centrados en la eliminación del ruido ambiente pueden ser vistos como algoritmos de separación de dos fuentes sonoras (la fuente sonora de interés y la fuente sonora que interfiere). Como consecuencia, en la publicación [P7] se propone utilizar las métricas SDR (ver Ec. (4.2)) y SIR (ver Ec. (4.4)) para medir la calidad acústica de los sonidos biomédicos estimados una vez que el ruido ambiente fue eliminado. Sin embargo, la métrica SAR (ver Ec. (4.3)) no fue utilizada en esta publicación ya que sus resultados no mostraron conclusiones relevantes.

No obstante, otras métricas utilizadas en trabajos recientes, que abordan la eliminación del ruido ambiental en la auscultación pulmonar, son detalladas a continuación [99, 100, 164]:

1) Segmental Signal-to-Noise Ratio ($fSNRseg$): es una medida objetiva de la calidad de la señal estimada en ventanas de tiempo corto (short-time), considerando la dinámica de la señal y la no estacionariedad del ruido [196]. Puede ser expresada como:

$$fSNRseg = \frac{10}{T}\sum_{\tau=1}^{T}\frac{\sum_{k=1}^{K} w_k SNR^F}{\sum_{k=1}^{K} w_k} \tag{4.15}$$

$$SNR^F = \log_{10}\left(\frac{|X(k,\tau)|^2}{\left(|X(k,\tau)| - \left|\hat{X}(k,\tau)\right|\right)^2}\right) \tag{4.16}$$

donde X es la representación espectral de la señal limpia (sin ruido interferente), $\hat{X}$ es la representación espectral de la señal estimada y w_k representa el peso para la banda de frecuencia k, donde $k = 1, ..., K$ y K denota el número total de bandas de frecuencia. Como se indica en [99], SNR^F se calcula aplicando ventanas de corta duración (30 ms) para tener en cuenta la dinámica de la señal y la no estacionalidad del ruido, utilizando una ventana de Hanning. Para cada trama, denotada como τ, la representación espectral de la señal original limpia (sin ruido) $X(k,\tau)$ y de la señal estimada $\hat{X}(k,\tau)$ es calculada mediante un filtrado de banda crítica. Siguiendo las indicaciones de [99], el ancho de banda, las frecuencias centrales de los 25 filtros que utilizan y los pesos perceptivos (Articulation Index) w_k siguen los propuestos en [257, 196]. Utilizando las indicaciones anteriormente descritas, la métrica $fSNRseg$ puede alcanzar un valor máximo igual a 35 cuando las señales comparadas (original y estimada) son idénticas. Por el contrario, $fSNRseg$ alcanza un valor mínimo, por debajo de -8 cuando una de las señales procede de un proceso blanco Gaussiano.

2) Normalized-Covariance Measure (NCM): es una métrica utilizada específicamente para medir la inteligibilidad del habla, denotada en inglés como "speech intelligibility (SI)", considerando la calidad auditiva de la señal estimada en varias bandas de frecuencia. En concreto, es una medida del índice de transmisión vocal, basada en la voz, que captura un promedio ponderado de una señal a la cantidad de ruido SNR^N. Donde, SNR^N se calcula a partir de la covarianza de la envolvente de las dos señales en diferentes bandas

de frecuencia k, y se normaliza para que tenga valores entre 0 y 1 [203]. En el caso de esta métrica, y como indica [99], los valores de pesado de cada banda w_k, pueden ser establecidos siguiendo las normas ANSI-1997 [36]. Considerando lo anterior, esta métrica puede expresarse como:

$$NCM = \frac{\sum_{k=1}^{K} w_k SNR^N(k)}{\sum_{k=1}^{K} w_k} \tag{4.17}$$

Aunque esta métrica está centrada en el habla (como muchas otras métricas relacionadas con la calidad auditiva), está construida para dar cuenta de las características auditivas del oído humano, reflejando así de forma general cómo una persona percibe la mejora de calidad acústica conseguida por la señal estimada (tras eliminar el ruido ambiental).

3) Three-Level Coherence Speech Intelligibility Index ($CSII$): según el estándar ANSI-1997 [36], es una métrica basada en la inteligibilidad del habla (SI), como ocurre en el caso de la métrica anterior NCM, que mide el índice de inteligibilidad del habla, denotado en inglés como "speech intelligibility index (SII)". A diferencia de la métrica NCM, la métrica $CSII$ utiliza una estimación de la SNR en el dominio espectral, para cada trama $\tau = 1, ..., T$, y la SNR_{ESI}^N de la señal residual, la cual es calculada utilizando filtros "roex" y la coherencia cuadrada de magnitud (magnitude-squared coherence) seguida de un proceso de normalización (valores entre 0 y 1) [99]. La métrica $CSII$, puede ser divida siguiendo el enfoque de tres niveles $CSII$ ($CSII_{low}$, $CSII_{med}$ y $CSII_{high}$) [196, 167]. Para ello es necesario dividir la señal en regiones de amplitud baja, media y alta, utilizando la raíz de la media cuadrática (root-mean-square) para cada trama τ. Considerando lo anterior, esta métrica puede expresarse como:

$$CSII = \frac{1}{T} \sum_{\tau=1}^{T} \frac{\sum_{k=1}^{K} w_k SNR_{ESI}^N(k,\tau)}{\sum_{k=1}^{K} w_k} \tag{4.18}$$

El conjunto de métricas presentado anteriormente es calculado utilizando la señal biomédica original y la señal biomédica estimada por el algoritmo de eliminación del ruido ambiental. Para finalizar, indicar que cuanto mayor sea el valor de estas métricas mejor calidad acústica tendrá la señal biomédica estimada.

4.7. Conclusiones

En este capítulo se ha presentado una revisión del estado del arte centrada en aquellos trabajos que han tratado las principales tareas de interés en el análisis de los sonidos sibilantes: detección de sonidos sibilantes, clasificación del tipo de sibilancia y eliminación del ruido ambiente que rodea al paciente. Por otro lado, se ha mostrado la carencia de trabajos relacionados con la separación y mejora de audio de los sonidos sibilantes en señales respiratorias monocanal. Considerando esta laguna, dos de las publicaciones [P1][P4] incluidas en esta Tesis se han centrado en abordar esta tarea de interés en el análisis de los sonidos sibilantes.

A continuación, se ha mostrado la problemática relacionada con la escasez de bases de datos estandarizadas en este ámbito científico. Además, se han descrito las principales métricas utilizadas en la evaluación del rendimiento de los algoritmos, considerando las principales tareas relacionadas con el análisis de los sonidos sibilantes.

Para finalizar, es importante destacar que en la revisión del estado del arte en relación al análisis de los sonidos sibilantes, no se ha encontrado ningún trabajo basado en un enfoque de descomposición de matrices no negativas (NMF/NMPCF). En este sentido, los trabajos propuestos en esta Tesis doctoral demuestran la potencialidad de técnicas NMF/NMPCF en el análisis de señales sonoras respiratorias, concretamente, en las tareas de separación/detección/clasificación de sonidos sibilantes, así como denoising de ruido ambiental aplicado a señales sonoras biomédicas.

CAPÍTULO 5

Resultados y Conclusiones

ESTA Tesis doctoral aborda diferentes modelos de procesado de señal aplicados en audio biomédico, basados en factorización de matrices no negativas, que pueden ser empleados para mejorar el diagnóstico derivado del proceso de auscultación, proporcionando una vía de información complementaria al médico que ayude a identificar patologías asociadas a los sonidos sibilantes. En particular, esta Tesis se centra en el desarrollo de métodos y algoritmos novedosos que tienen el papel de afrontar algunas de las tareas o dificultades de mayor relevancia para los neumólogos en el análisis de los sonidos sibilantes, como son: la separación y mejora de audio de los sonidos sibilantes, la detección y localización temporal de las sibilancias, la clasificación del tipo de sibilancia, y la eliminación del ruido ambiente que rodea al paciente durante la auscultación. Todos los métodos propuestos han sido evaluados y comparados con los métodos más relevantes del estado del arte, mostrando sus fortalezas y debilidades específicas para cada tarea. En esta línea, este último capítulo resume los resultados, conclusiones y líneas futuras de las diferentes aportaciones científicas derivadas de esta Tesis. Además, se incluyen dos secciones en las que se exponen las conclusiones generales de la investigación y diferentes líneas futuras que pueden surgir gracias al desarrollo de este trabajo.

5.1. Separación de señales sonoras sibilantes

Como se ha comentado anteriormente uno de los principales problemas a los que se enfrentan los médicos cuando auscultan a un paciente es el solapamiento que existe entre los sonidos respiratorios normales y las sibilancias. Como resultado, la capacidad cognitiva del médico se reduce pudiendo causar un diagnóstico erróneo debido a no escuchar los sonidos adventicios

sibilantes con claridad. Las publicaciones [P1][P4] presentan métodos para hacer frente a esta problemática.

5.1.1. Publicación [P1]

En la publicación [P1] se propone un método basado en un enfoque NMF con regularizaciones, para separar los sonidos sibilantes de los sonidos respiratorios en señales monocanal donde ambos sonidos se encuentran solapados en tiempo y en frecuencia. En concreto, se proponen utilizar las regularizaciones de suavidad (smoothness) y dispersión (sparseness) para obtener un modelado diferencial de los patrones espectrales que caracterizan a ambos sonidos. Tras una revisión de la literatura y un estudio experimental se observó que: i) el espectrograma de los sonidos sibilantes puede ser modelado aplicando dispersión en frecuencia, ya que las sibilancias se caracterizan mediante la distribución de energía en picos espectrales de banda estrecha; y ii) el espectrograma de los sonidos respiratorios normales puede ser modelado aplicando suavidad espectral en frecuencia y en el tiempo, ya que los sonidos respiratorios se representan como patrones espectrales en banda ancha, cuya energía varía lentamente a lo largo del tiempo.

Para evaluar el rendimiento del método propuesto se crearon tres bases de datos (DORIG, D0dB y D5dB) compuestas cada una por 20 señales mezcladas (sonidos respiratorios y sonidos sibilantes), con una duración entre 5 y 20 segundos. Cada señal de entrada fue generada mezclando señales sibilantes manualmente separadas o aisladas (por medio de una máscara tiempo-frecuencia aplicada al espectrograma de la señal mezcla original para seleccionar solo las frecuencias de cada trama correspondientes a las sibilancias) y señales respiratorias normales (en las cuales las sibilancias están inactivas), extraídas de los repositorios online más utilizados en el estado del arte de este ámbito (ver Sección 4.5.2). Para cada mezcla, el número de sibilancias y la ubicación temporal de cada una de ellas se determinó mediante un proceso pseudoaleatorio utilizando una distribución uniforme estándar. Las tres bases de datos fueron creadas a partir de las mismas señales, pero variando la SNR entre ellas. Específicamente, la base de datos DORIG presenta la SNR original tras realizar la mezcla, entre 2 y 8 dB, la base de datos D0dB fue construida fijando la SNR a 0 dB y la base de datos D5dB con una SNR de -5 dB.

El método propuesto fue comparado con la versión estándar del enfoque NMF para evaluar la mejora que supone caracterizar los sonidos presentes en la mezcla con las regularizaciones propuestas (suavidad y dispersión). Los resultados permitieron extraer las siguientes conclusiones: i) La propuesta de modelar los sonidos respiratorios y sibilantes añadiendo significado físico (con regularizaciones) al enfoque NMF, mejora significativamente la calidad de audio de los sonidos sibilantes eliminando la mayoría de los sonidos respiratorios. En concreto la propuesta mejora al modelo NMF estándar sobre 9 dB en SDR y 10 dB en SIR, evaluando la base de datos DORIG, sobre 6.3 dB en SDR y 10 dB en SIR, evaluando la base de datos D0dB, y sobre 4.9 dB en SDR y 13 dB en SIR, evaluando la base de datos D5dB; ii) El método propuesto recupera la mayoría de los sonidos sibilantes que no pueden ser escuchados en la mezcla, incluso cuando se evalúan entornos ruidosos con una SNR<0 dB (donde las sibilancias

son apenas audibles en la señal mezcla debido a la alta interferencia respiratoria); iii) No se requiere ninguna etapa de entrenamiento ya que se propone un método no supervisado (ciego); y iv) Una de las desventajas observadas en el método propuesto es la presencia de espurios en el intervalo temporal en el que sólo están presentes los sonidos respiratorios.

En conclusión, este trabajo fue el responsable de iniciar la línea de investigación abordada en esta Tesis, ya que demostró que NMF era una herramienta potencialmente aplicable a este campo científico permitiendo modelar de forma diferencial los sonidos sibilantes y respiratorios normales. Adicionalmente, el método propuesto presentado en esta publicación puede considerarse pionero en este ámbito al ser el primer trabajo, hasta donde conoce el autor de esta Tesis, donde se aplica NMF a la separación de sonidos sibilantes y respiratorios, obteniendo resultados prometedores. Por último, en esta publicación se propusieron las siguientes líneas futuras: i) la eliminación de los sonidos respiratorios espurios activos aún en la señal sibilante estimada. En concreto esta línea fue abordada en la publicación [P2], realizando la optimización del método propuesto, y en la publicación [P4] presentando un enfoque novedoso basado en NMPCF; y ii) el desarrollo de un detector de sibilancias basado en el método propuesto. En este sentido, la publicación [P2] puede verse como la continuación de este trabajo.

5.1.2. Publicación [P4]

En [P4] se propone una versión extendida del enfoque NMPCF publicado en [170], denominada Inter-Segment NMPCF (IIS-NMPCF), para eliminar la mayor parte de la interferencia acústica causada por los sonidos respiratorios normales mientras se preserva el contenido de los sonidos sibilantes que el médico necesita para proporcionar un diagnóstico fiable del estado de las vías respiratorias del sujeto. En concreto, se parte de las siguientes hipótesis: i) los sonidos respiratorios normales pueden ser considerados eventos sonoros que se repiten durante la mecánica de la respiración del sujeto. Por ello, los sonidos respiratorios normales pueden ser modelados compartiendo patrones espectrales presentes en las diferentes etapas respiratorias (inspiración o espiración); y ii) los sonidos sibilantes no pueden ser modelados compartiendo patrones espectrales de cada segmento respiratorio, ya que las sibilancias podrían no estar activas en algunos segmentos debido a su naturaleza impredecible en el tiempo, motivada por el trastorno pulmonar derivado. Bajo estas hipótesis y con el objetivo de mejorar el rendimiento de separación del enfoque NMPCF convencional [170], el cual trata por igual todos los segmentos del espectrograma de entrada, la principal contribución de la propuesta consiste en dar mayor importancia (peso), en el modelado de los patrones respiratorios, a los segmentos clasificados como no-sibilantes utilizando información del tipo de segmento (sibilante o no-sibilante) proporcionada por un sistema de detección de presencia o ausencia de sibilancias previamente presentado en [P3].

Considerando la problemática derivada de la carencia de bases de datos estandarizadas de sonidos sibilantes (descrita en la sección anterior), dos bases de datos, P1 y T1, fueron creadas a partir de un conjunto de señales sonoras de diferentes sujetos obtenidas de los repositorios online más referenciados (ver Sección 4.5.2) y de la base de datos ICBHI (ver Sección 4.5.1),

para abordar la etapa experimental. En concreto, la base de datos P1 está compuesta por 48 señales (3/4 del total de señales utilizadas en los experimentos) y fue utilizada en el proceso de optimización hiperparamétrico del método propuesto. En cambio, la base de datos T1 está compuesta por 16 señales (1/4 del total de señales utilizadas en los experimentos) y fue utilizada para evaluar el rendimiento del método propuesto en términos de separación de fuentes sonoras (sonidos sibilantes y sonidos respiratorios normales). En este sentido, ambas bases de datos, P1 y T1, fueron creadas mezclando señales de sonidos sibilantes manualmente separadas (los sonidos respiratorios normales están inactivos) y señales de sonidos respiratorios normales (los sonidos sibilantes están inactivos), extraídas de los repositorios anteriormente mencionados y garantizando que siempre fuesen distintas para ambas bases de datos. Con el objetivo de evaluar el rendimiento del método propuesto en función de la SNR de la señal mezclada, tres bases de datos fueron generadas, variando la SNR, a partir de las señales correspondientes a la base de datos T1: T1H (SNR=5 dB), T1M (SNR=0 dB) y T1L (SNR=-5 dB). En conjunto, todas las bases de datos (P1, T1H, T1M y T1L) ofrecen 1.474 segundos de grabación, 96 pacientes enfermos (con sibilancias), 874 eventos respiratorios (inspiración o espiración) y 133 sibilancias.

El principal atractivo de la etapa experimental se debe a que el rendimiento del método propuesto IIS-NMPCF ha sido comparado con diferentes modelos de separación de fuentes sonoras (basados en el enfoque NMF o NMPCF) propuestos en el estado del arte, los cuales se definieron en la Sección 3.3 y son:

- **Modelo NMF estándar** sin ningún tipo de entrenamiento previo y sin regularizaciones.

- **Modelo NMF semi-supervisado (SSNMF)** [284, 186], en el cual las bases respiratorias son aprendidas previamente y permanecen fijas durante el proceso iterativo.

- **Modelo NMF supervisado (SNMF)** [321, 76], en el cual tanto las bases sibilantes como las respiratorias son aprendidas previamente y permanecen fijas durante el proceso iterativo.

- **Modelo NMF regularizado (CNMF)** [202, 63], en el cual se incluyen las regularizaciones de suavidad y dispersión para modelar los sonidos sibilantes y respiratorios. En concreto este modelo corresponde al presentado en la publicación [P1], el cual fue optimizado en la publicación [P2].

- **Modelo NMPCF semi-supervisado (1S-NMPCF)** [170], el cual realiza una factorización conjunta (cofactorización) entre el espectrograma mezcla de entrada y un espectrograma de entrenamiento respiratorio compartiendo las bases espectrales respiratorias.

- **Modelo NMPCF supervisado (2S-NMPCF)** [147], el cual realiza un doble proceso de cofactorización. Por un lado, realiza la factorización conjunta entre el espectrograma mezcla de entrada y un espectrograma de entrenamiento respiratorio compartiendo las bases

espectrales respiratorias, y por otro lado, realiza la factorización conjunta entre el espectrograma mezcla de entrada y un espectrograma de entrenamiento sibilante compartiendo las bases espectrales sibilantes.

- **Modelo NMPCF ciego (T-NMPCF)** [170], el cual realiza una factorización conjunta entre los diferentes segmentos en los que se divide el espectrograma mezcla de entrada para compartir los patrones espectrales respiratorios que se repiten a lo largo del tiempo
- **Modelo semi-supervisado (ST-NMPCF)** [170], el cual es una combinación del modelo T-NMPCF y el modelo 1S-NMPCF. En concreto, este modelo fue tomado como base en la versión extendida del método propuesto IIS-NMPCF en esta publicación [P4].

Además, con el objetivo de analizar la importancia del espectrograma respiratorio de entrenamiento en el proceso de cofactorización, para modelar las bases espectrales respiratorias, el método propuesto fue evaluado utilizando un espectrograma de entrenamiento respiratorio (en este caso el modelo propuesto fue denotado como **IIS-NMPCF**) y sin utilizarlo (**IIS$_0$-NMPCF**). Los resultados permitieron extraer las siguientes conclusiones: i) Los modelos semi-supervisados (SSNMF y 1S-NMPCF) obtienen mejores resultados que los modelos supervisados (SNMF y 2S-NMPCF), lo que indica que ambas señales de entrenamiento proporcionan una sobre-información que causa ambigüedad espectro-temporal en la factorización o cofactorización de los diccionarios sibilantes y respiratorios; ii) Los modelos basados en el enfoque NMPCF (por ejemplo, 1S-NMPCF) obtienen mejores resultados que los que se basan en el enfoque NMF (por ejemplo, SSNMF), ya que los modelos NMF utilizan un diccionario fijo, el cual ha sido previamente obtenido, y los modelos NMPCF construyen diccionarios dinámicos que se adaptan a cada espectrograma mezcla de entrada; iii) El modelo T-NMPCF mejora al 1S-NMPCF, lo que indica que modelar los patrones respiratorios repetitivos que ocurren a lo largo de la señal mezcla de entrada es más efectivo que utilizar una señal respiratoria de entrenamiento; iv) El modelo regularizado CNMF propuesto en [P1] y optimizado en [P2] obtiene uno de los mejores rendimientos en separación, ocupando la cuarta posición en el ranking de los modelos evaluados (por debajo del modelo ST-NMPCF y de los modelos propuestos IIS$_0$-NMPCF y IIS-NMPCF), lo que indica que las regularizaciones de suavidad y dispersión son eficientes para modelar los sonidos sibilantes y respiratorios normales presentes en la mezcla de entrada; v) El método propuesto, sin utilizar un espectrograma respiratorio de entrenamiento (IIS$_0$-NMPCF), obtiene mejores resultados que el modelo T-NMPCF y el modelo ST-NMPCF (combinación del modelo T-NMPCF y 1S-NMPCF), lo que indica que añadir mayor importancia a los segmentos no-sibilantes en el proceso de cofactorización (para modelar los patrones respiratorios repetitivos) mejora el rendimiento de separación, ya que esto evita que los patrones respiratorios modelados se vean contaminados por espurios sibilantes en los segmentos sibilantes; y vi) Por último, el modelo propuesto, utilizando un espectrograma respiratorio de entrenamiento (IIS-NMPCF), obtiene el mejor rendimiento en comparación con todos los modelos indicados y en todos los escenarios evaluados (variando la SNR). En concreto, la propuesta IIS-NMPCF mejora al modelo base ST-NMPCF sobre 5.8 dB en SDR, 4.9 dB en SIR y 7.5 dB en SAR evaluando el escenario más ruidoso (SNR=-5 dB).

En conclusión, el método propuesto permite mejorar la calidad sonora de los sonidos sibilantes maximizando la eliminación de las interferencias acústicas causadas por los sonidos respiratorios normales y preservando la mayor parte de la información determinante contenida en las sibilancias para el correcto diagnóstico. Además, considerando los resultados obtenidos, el método propuesto puede ser considerado un instrumento apropiado para ser aplicado en entornos sonoros en los que los sonidos sibilantes son apenas audibles (SNR<0 dB). Para finalizar, en esta publicación se propuso como trabajo futuro el desarrollo de un modelo NMF basado en regularizaciones para modelar los diferentes tipos de sonidos sibilantes, considerando su contenido espectral, a fin de clasificar automáticamente las sibilancias y poder proporcionar información del tipo de trastorno pulmonar ya que la distribución de energía en frecuencia de una sibilancia va asociada al tipo de vía respiratoria obstruida, lo cual puede revelar información acerca de la gravedad de la obstrucción pulmonar que sufre el sujeto. Esta línea fue abordada posteriormente en la publicación [P6].

5.2. Detección de señales sonoras sibilantes

Desde un punto de vista clínico, la detección de los sonidos sibilantes es una tarea desafiante para diagnosticar correctamente algunas de las patologías respiratorias más relevantes (asma, bronquiolitis, bronquiectasia o EPOC) y para alcanzar una identificación temprana de dichas patologías. Actualmente se producen un alto porcentaje de diagnósticos erróneos, especialmente en escenarios ruidosos en los cuales los sonidos sibilantes son casi inaudibles debido al solapamiento de la respiración normal. Por ello, se ha manifestado la necesidad de métodos o algoritmos que detecten sonidos sibilantes en señales sonoras respiratorias con el fin de evitar que el paciente regrese al centro de salud con un empeoramiento de la obstrucción de las vías respiratorias, indicada por los sonidos sibilantes, que no se detectó en el primer examen clínico realizado por auscultación. En este sentido, las publicaciones [P2][P5] presentan métodos centrados en localizar el intervalo temporal en el cual las sibilancias están activas, dentro de los ciclos respiratorios que componen a la señal respiratoria de entrada, y la publicación [P3] presenta un método para detectar la presencia o ausencia de sibilancias en señales respiratorias.

5.2.1. Publicación [P2]

En [P2] se propone un novedoso algoritmo, basado en un enfoque NMF ciego con regularizaciones, que permite modelar y detectar la presencia de los sonidos sibilantes en señales sonoras respiratorias monocanal, localizando los intervalos temporales en los cuales las sibilancias se encuentran activas. En general, el método propuesto permite separar y detectar los sonidos sibilantes y respiratorios procedentes de una señal mezclada. Por ello, el esquema de la propuesta se compone de dos etapas en cascada: separación y detección. La etapa de separación utiliza como base el trabajo previo de separación [P1]. En concreto, el objetivo de la etapa de separación es realizar una estimación de los sonidos sibilantes y respiratorios introduciendo una serie de regularizaciones espectro-temporales (suavidad y dispersión), al modelo de

factorización NMF, para caracterizar el comportamiento de ambos sonidos. Como se ha mencionado, esta etapa propone utilizar las mismas regularizaciones que demostraron su eficacia en el trabajo previo [P1]. Sin embargo, en el trabajo actual se realizó un proceso de optimización para obtener el conjunto de parámetros que ofrecía el mejor rendimiento en términos de separación. Posteriormente, considerando que los sonidos respiratorios pueden encontrarse aislados en algunas áreas de la mezcla (particularmente, en los intervalos del ciclo respiratorio donde las sibilancias se encuentran inactivas), la etapa de detección propone utilizar la divergencia Kullback-Leibler sobre el espectrograma de la mezcla de entrada y el espectrograma estimado respiratorio para distinguir las áreas sibilantes y las respiratorias. Dado que el espectrograma estimado respiratorio (obtenido en la etapa de separación) está libre de sonidos sibilantes, la divergencia Kullback-Leibler tendrá valores muy pequeños en áreas donde solo estén presentes los sonidos respiratorios. Sin embargo, en áreas donde los sonidos respiratorios y sibilantes estén solapados el valor de la divergencia Kullback-Leibler será mucho más alto.

Tres bases de datos (E1, T1 y T2) fueron utilizadas en la etapa experimental del método propuesto. La base de datos E1 fue utilizada en la optimización del modelo descrito en la etapa de separación. Las bases de datos T1 y T2 fueron utilizadas para evaluar el rendimiento de detección de sibilancias. En concreto, la base de datos E1 fue generada mezclando sonidos sibilantes manualmente separados (sin sonidos respiratorios normales) y sonidos respiratorios normales (sin sonidos sibilantes), obtenidos de los repositorios online más referenciados (ver Sección 4.5.2) y de la base de datos ICBHI (ver Sección 4.5.1). Esta base de datos está compuesta por 48 mezclas de audio con una SNR entre 0 dB y 9 dB, con una duración entre 4 y 25 segundos por mezcla, y con un total de 92 sibilancias y 154 ciclos respiratorios. Por otro lado, la base de datos T1 es la misma que se utilizó en [233], la cual fue compartida por los autores. Esta base de datos está compuesta por 16 señales sonoras (8 señales de pacientes sanos sin sibilancias y 8 señales de pacientes enfermos con sibilancias), con una duración entre 4 y 51 segundos por grabación, y con un total de 36 sibilancias y 168 ciclos respiratorios. Por último, la base de datos T2 fue generada siguiendo un procedimiento similar al utilizado en la base de datos E1 (garantizando que los sonidos de ambas bases de datos fuesen distintos) y está compuesta por 16 mezclas de audio, con una duración entre 7 y 22 segundos por mezcla, y con un total de 41 sibilancias y 63 ciclos respiratorios. Con el objetivo de evaluar el rendimiento del método propuesto en función de la SNR de la señal mezclada, tres bases de datos fueron generadas, variando la SNR, a partir de las señales correspondientes a la base de datos T2: T2H (SNR=5 dB), T2M (SNR=0 dB) y T2L (SNR=-5 dB).

El método propuesto fue comparado con tres de los algoritmos de detección de sibilancias más relevantes de la literatura [233, 215, 281]. En concreto, en [215, 281] proponen métodos basados en la extracción de características (MFCC) y clasificadores (SVM y KNN), y en [233] proponen extraer las trayectorias espectrales de las sibilancias utilizando HMM. Los resultados obtenidos permitieron extraer las siguientes conclusiones: i) El rendimiento de detección de sibilancias del método propuesto es competitivo en comparación con los métodos de la literatura, obteniendo los mejores resultados en términos de SE (95,71 %) y ACC (95,86 %); ii) Aunque la propuesta obtiene los mejores resultados en términos de SE, también obtiene los peores

resultados en términos de SP (93,02 %), lo que sugiere que el método propuesto tiende a proporcionar un mayor número de falsos positivos (tramas respiratorias detectadas erróneamente como tramas sibilantes) con el fin de detectar el intervalo temporal completo en el que ocurren las sibilancias; iii) la robustez del método propuesto es demostrada ya que todas las métricas de detección se reducen como máximo en un 3 % comparando las bases de datos de diferente SNR (T2H, T2M y T2L), lo que sugiere que el método propuesto puede ser una herramienta útil para ser aplicada en ambientes ruidosos, donde los sonidos sibilantes son apenas audibles debido a los sonidos respiratorios normales; y iv) El método propuesto detecta correctamente la ausencia de sibilancias en todas las señales sonoras correspondientes a pacientes sanos de la base de datos T1, confirmando la fiabilidad de la propuesta para discriminar entre pacientes sanos (sin sibilancias) y enfermos (con sibilancias).

Por último, en esta publicación se propusieron las siguientes líneas futuras: i) desarrollo de nuevas características espectro-temporales para discriminar correctamente entre bases espectrales sibilantes y no sibilantes (respiratorias) a partir de un enfoque NMF. En concreto, esta línea fue abordada posteriormente en la publicación [P5] y [P3], donde se propone utilizar uno de los descriptores más relevantes (Gini index) para medir el grado de tonalidad de las bases espectrales, y así clasificarlas entre bases sibilantes y respiratorias; y ii) diseñar enfoques alternativos basados en NMF para adaptar el modelo de detección a un escenario en tiempo real.

5.2.2. Publicación [P5]

En [P5] se propone un sistema experto e inteligente basado en el comportamiento diferencial mostrado entre los sonidos sibilantes y respiratorios normales en el dominio tiempo-frecuencia. En concreto, se define un algoritmo recursivo que combina el enfoque de descomposición matricial "Orthogonal Non-negative Matrix Factorization (ONMF)" y la utilización del descriptor de tonalidad (dispersión espectral) "Gini index" para localizar el intervalo temporal activo de las sibilancias en señales sonoras respiratorias monocanal. En general, el método se compone de cuatro etapas. La primera etapa consiste en aplicar el enfoque ONMF, en lugar del enfoque NMF estándar, ya que permite factorizar o descomponer patrones espectrales (bases) minimizando la redundancia entre ellos y presentando una mayor fidelidad a lo que ocurre en el mundo real. La segunda etapa clasifica la naturaleza periódica de las bases ONMF analizando el grado de tonalidad (dispersión espectral) obtenido por el descriptor Gini index en el dominio de la frecuencia. La tercera etapa propone un novedoso criterio de parada que permite refinar, a lo largo de las iteraciones recursivas, el espectrograma estimado sibilante. Específicamente, el criterio de parada propuesto permite aplicar una nueva iteración del algoritmo recursivo siempre y cuando se elimine una proporción significativa de contenido respiratorio normal a cambio de minimizar la pérdida de contenido significativo sibilante activo en el espectrograma sibilante estimado. Por último, la cuarta etapa discrimina entre paciente sano (sin sibilancias) y paciente enfermo (con sibilancias) calculando el grado de dispersión (Gini index) a partir de la distribución de la energía espectral del espectrograma sibilante estimado, localizando los intervalos temporales en los que las sibilancias están activas para los pacientes clasificados como enfermos.

Considerando que el método propuesto sigue la misma línea de investigación que la publicación previa [P2], las bases de datos que se utilizaron para medir el rendimiento del algoritmo propuesto fueron las mismas que las utilizadas en el trabajo previo (T1, T2H, T2M y T2L). Además, el método propuesto fue comparado con los mismos algoritmos descritos en [P2], incluido el método que se propuso en el anterior trabajo. Los resultados obtenidos permitieron extraer las siguientes conclusiones: i) El método propuesto proporciona los mejores resultados generales de detección en comparación con los otros métodos de la literatura, considerando todos los escenarios evaluados con diferente SNR. En concreto, comparando T2H y T2L, mientras que los resultados de SE, SP y ACC del método propuesto sólo bajan un 2 %, los resultados proporcionados por los otros métodos bajan un 9 %, esto sugiere que el método propuesto es más robusto que los métodos del estado del arte a medida que las condiciones de SNR empeoran (incluso cuando los sonidos respiratorios normales son más fuertes que los sibilantes); ii) A diferencia de los enfoques basados en Machine Learning, donde su rendimiento depende de una base de datos de entrenamiento, el rendimiento de ambos métodos propuestos (el trabajo previo [P2] y el actual) no está sujeto a ninguna base de datos de entrenamiento debido a que ambos se basan en un enfoque no supervisado (ciego). Este hecho podría hacer que ambos métodos sean atractivos para su uso en entornos clínicos, ya que a menudo no se dispone de bases de datos de sonidos sibilantes; iii) Al igual que en el trabajo previo [P2], la propuesta actual detecta correctamente la ausencia de sibilancias en todas las señales correspondientes a pacientes sanos de la base de datos T1, confirmando la fiabilidad de este método para discriminar entre pacientes sanos (sin sibilancias) y enfermos (con sibilancias); y iv) El método previamente propuesto [P2] obtiene mejores resultados SE que SP, lo que sugiere que este método es más eficiente para detectar el intervalo temporal sibilante completo a expensas de proporcionar un mayor número de falsos positivos (tramas respiratorias detectadas erróneamente como tramas sibilantes). Sin embargo, la propuesta actual obtiene mejores resultados SP que SE, lo que sugiere que este método es más eficiente para detectar exactamente el intervalo temporal sibilante a expensas de proporcionar un mayor número de falsos negativos (tramas sibilantes detectadas erróneamente como tramas respiratorias).

Por último, cabe destacar que tanto el trabajo propuesto previamente [P2] como la propuesta actual [P5], presentan una etapa en su esquema en la cual se determina la condición del paciente. Por ello, en el caso de que en esa etapa se produzca un falso negativo (paciente enfermo erróneamente diagnosticado como paciente sano) los intervalos temporales sibilantes no podrán ser localizados, ya que el algoritmo asume que no existen. Esto motivó el desarrollo de un método centrado únicamente en determinar la condición del sujeto (sano o enfermo), el cual fue abordado en la publicación [P3]. Además, considerando que la mayoría de algoritmos del estado del arte no consideran el ruido ambiente que rodea al paciente durante la auscultación, el cual podría afectar negativamente al rendimiento de detección, en esta publicación se propuso como trabajo futuro desarrollar un enfoque de factorización incluyendo regularizaciones que modelasen el comportamiento del ruido ambiental para cancelar el ruido de fondo típicamente encontrado en la consulta de un médico. En concreto, esta línea fue abordada posteriormente en la publicación [P7].

5.2.3. Publicación [P3]

En [P3] se propone un método para detectar automáticamente la presencia o ausencia de sonidos sibilantes en señales sonoras respiratorias monocanal, es decir, un método que determina si la señal de audio respiratoria corresponde a un paciente sano (ausencia de sibilancias) o enfermo (presencia de sibilancias). En general, el método propuesto se compone de tres etapas. A diferencia de la mayoría de algoritmos de detección de sibilancias, en los cuales el rango espectral de los sonidos sibilantes se establece a priori, la primera etapa presenta un novedoso algoritmo que estima el rango espectral en el que la probabilidad de que se produzcan los sonidos sibilantes para cada señal analizada es máxima, dicho rango espectral es definido como "Band Of Interest (BOI)". La segunda etapa, propone un enfoque NMF semi-supervisado con regularizaciones basado en el principio de tonalidad para obtener los patrones espectrales que modelan la naturaleza tonal (picos espectrales en banda estrecha) mostrada por las sibilancias en la BOI estimada, y así realizar una separación entre los sonidos sibilantes y los sonidos respiratorios normales. A partir del espectrograma estimado sibilante, la tercera etapa propone un método para clasificar la condición del sujeto como sano (sin sibilancias) o enfermo (con sibilancias) analizando la suavidad temporal de las trayectorias espectrales definidas por la energía más significativa presente en cada trama de la BOI estimada. Las hipótesis en las que se basa el método propuesto son: i) las trayectorias espectrales en el caso de los pacientes enfermos son distinguidas por la naturaleza continua de los sonidos sibilantes; y ii) las trayectorias espectrales de los sujetos sanos son distinguidas por la naturaleza aleatoria o ruidosa de los sonidos respiratorios normales. Note que, aunque el modelo ha sido denominado como semi-supervisado, no depende de ninguna base de datos de entrenamiento (la segunda etapa obtiene la información necesaria a partir de la BOI estimada). Por lo tanto, se podría considerar un enfoque NMF no supervisado (ciego).

Seis bases de datos (P1, T1, T2H, T2M, T2L y T3) fueron utilizadas en la etapa experimental de la publicación, las cuales se encuentran detalladas en [P3]. En total, estas bases de datos proporcionan 4.107 segundos de grabación, 114 sujetos sanos, 96 sujetos enfermos, 2.168 eventos respiratorios (inspiración o espiración) y 219 sibilancias. Algunos detalles de las mismas se comentan a continuación:

- La base de datos P1 fue utilizada en la etapa de entrenamiento de los algoritmos del estado del arte [198, 215, 281] con los que el método propuesto fue comparado. Está compuesta por 48 pacientes sanos y 48 pacientes enfermos, y las señales sonoras fueron obtenidas de los repositorios online más referenciados (ver Sección 4.5.2) y de la base de datos ICBHI (ver Sección 4.5.1).

- La base de datos T1 (solo pacientes sanos) fue utilizada para evaluar la robustez de los algoritmos comparados en diagnosticar correctamente a sujetos sanos. En concreto se compone de 48 señales de sujetos sanos obtenidas de los repositorios online más referenciados en la literatura y de la base de datos ICBHI.

- Las bases de datos T2H, T2M y T2L (solo pacientes enfermos) fueron utilizadas para evaluar la robustez de los algoritmos comparados cuando los sonidos sibilantes están enmascarados por los sonidos respiratorios normales. Son las mismas bases de datos utilizadas en las publicaciones [P2] y [P5]. En concreto, las tres bases de datos fueron generadas variando la SNR entre la señal sibilante y respiratoria mezclada: T2H (SNR=5 dB), T2M (SNR=0 dB) y T2L (SNR=-5 dB).

- La base de datos T3 (pacientes sanos y enfermos) fue utilizada para evaluar el rendimiento de detección de sibilancias de los algoritmos comparados en un escenario clínico real, donde los algoritmos tenían que decidir si el diagnóstico de un sujeto era sano o enfermo. Está compuesta por 18 señales de pacientes sanos y 30 de pacientes enfermos, y es la misma que se utilizó en [233], la cual fue compartida por los autores.

- Con el objeto de validar los resultados, es importante resaltar que cada fichero es único en su respectiva base de datos y que por lo tanto no ha sido utilizado en las demás.

El método propuesto fue comparado con tres de los algoritmos, para la detección de la presencia de sibilancias, más relevantes de la literatura [198, 215, 281]. En concreto, en [215, 281] proponen métodos basados en la extracción de características (MFCC) y clasificadores (SVM y KNN), y en [281] propone un método para detectar la presencia de sibilancias combinando un enfoque denominado "Empirical mode decomposition (EMD)", extracción de características a partir de la frecuencia instantánea y el clasificador SVM. Además, para evaluar el rendimiento de los métodos con respecto a la longitud temporal de la señal de entrada, se propuso una evaluación basada en tres niveles: primer nivel (señal respiratoria completa compuesta por varios ciclos respiratorios), segundo nivel (ciclo respiratorio compuesto por la fase de inspiración y espiración) y tercer nivel (fase respiratoria, es decir, inspiración o espiración). Los resultados obtenidos en la etapa experimental permitieron extraer las siguientes conclusiones: i) El método propuesto proporcionó los mejores resultados de clasificación de presencia/ausencia de sibilancias, en comparación con los otros métodos de la literatura, teniendo en cuenta todas las bases de datos evaluadas y todos los niveles de evaluación. En concreto, considerando la evaluación de la base de datos T1, la propuesta obtuvo un 100 % en términos de SP, lo que indica su efectividad para detectar correctamente pacientes sanos; ii) A diferencia del resto de métodos del estado del arte evaluados, el método propuesto obtuvo los mismos resultados en los tres niveles de evaluación, independientemente de la base de datos evaluada. Por ello, una de las principales ventajas de la propuesta es la robustez con respecto a la longitud temporal de la señal para clasificar correctamente la condición del sujeto; iii) Comparando la evaluación de las bases de datos T2H y T2L en el primer nivel de evaluación, los resultados de la métrica SE del método propuesto sólo bajan un 6,25 % mientras que los resultados obtenidos por los otros métodos bajan por encima del 18,75 %. Esto sugiere que el método propuesto es más robusto que los métodos de la literatura para evaluar escenarios ruidosos, específicamente, cuando SNR<0 dB; y iv) A diferencia de los métodos de la literatura (basados en machine learning), el método propuesto no depende de ninguna base de datos de entrenamiento (la base de datos P1

solo ha sido utilizada en el entrenamiento de los otros métodos), ya que se basa en modelar las características espectro-temporales de los sonidos sibilantes para discriminarlos de los sonidos respiratorios normales.

Por último, en esta publicación se propuso como trabajo futuro el desarrollo de un modelo NMF basado en regularizaciones para modelar los diferentes tipos de sonidos sibilantes (monofónicos/polifónicos), considerando su estructura espectral, a fin de clasificar automáticamente la gravedad del trastorno pulmonar. Esta línea fue abordada posteriormente en la publicación [P6].

5.3. Clasificación de señales sonoras sibilantes

Como se ha mencionado anteriormente, los médicos manifiestan la complejidad de clasificar el tipo de sibilancia (monofónica/polifónica) utilizando únicamente la información acústica proporcionada por el estetoscopio. Además, se ha demostrado que, distinguir entre sibilancias monofónicas y polifónicas es una tarea crítica en el diagnóstico diferencial del asma y la EPOC, ya que el asma está asociado a las sibilancias monofónicas y la EPOC a las sibilancias polifónicas. En este sentido, la publicación [P6] presenta un método que aborda este tema.

5.3.1. Publicación [P6]

En [P6] se propone un método para clasificar el tipo de sibilancia, monofónica o polifónica, considerando su estructura armónica, aportando información relevante al médico para determinar el nivel de gravedad de la enfermedad pulmonar atendiendo al tipo de vía respiratoria obstruida. Ninguno de los métodos de la literatura tiene en cuenta la interferencia que los sonidos respiratorios normales pueden causar en la clasificación del tipo de sibilancia. Por ello, se propone un algoritmo basado en un enfoque NMF regularizado que permite clasificar el tipo de sibilancia eliminando la interferencia causada por los sonidos respiratorios. En general, el método propuesto se compone de dos etapas. La primera etapa presenta un novedoso enfoque denotado como "Constrained Low-Rank Non-negative Matrix Factorization (CL-RNMF)", para extraer los patrones espectrales en banda estrecha que caracterizan a las sibilancias con la menor interferencia respiratoria posible. En concreto, se propone una configuración de bajo rango (low-rank) con un número muy reducido de bases sibilantes para compactar sus componentes en frecuencia en el menor número de bases posibles, para su posterior análisis armónico. Además, se propone aplicar un conjunto de regularizaciones tiempo-frecuencia (suavidad y dispersión) al modelo CL-RNMF para modelar el comportamiento diferencial entre los sonidos sibilantes y respiratorios normales. Partiendo de las componentes sibilantes aisladas en la etapa anterior, la segunda etapa analiza la estructura armónica de la sibilancia aplicando un conjunto de condiciones para determinar el tipo de sibilancia (monofónica o polifónica). A diferencia de los algoritmos del estado del arte, cuya clasificación se basa en analizar la energía de las componentes espectrales sibilantes, el método propuesto realiza la clasificación considerando

exclusivamente la localización de las componentes sibilantes en frecuencia (sin atender a ningún criterio de energía).

Debido a la inexistencia de una base de datos pública donde el tipo de sibilancia esté clasificado (monofónica o polifónica), se creó una base de datos etiquetada con la colaboración del neumólogo Dr. Gerardo Pérez Chica del Hospital Universitario de Jaén (España). En concreto, la base de datos está formada por 400 segmentos sibilantes con una duración superior a 100 ms, extraídos de señales sonoras respiratorias encontradas en los repositorios online más referenciados (ver Sección 4.5.2) y en la base de datos ICBHI (ver Sección 4.5.1). El tipo de sibilancia fue etiquetado a partir de una inspección acústica realizada por el neumólogo y una verificación visual del espectrograma considerando la estructura armónica que distingue a ambos tipos (monofónica/polifónica). Específicamente, la base de datos se compone de 200 segmentos monofónicos y 200 segmentos polifónicos. Considerando que las sibilancias monofónicas pueden mostrar dos estructuras armónicas diferentes, sibilancias con un único pico espectral en banda estrecha (tipo 1) y sibilancias con un pico espectral en banda estrecha más sus armónicos correspondientes (tipo 2), la parte correspondiente a los segmentos monofónicos se compuso de 100 monofónicos tipo 1 y 100 monofónicos tipo 2.

Con el objetivo de demostrar la efectividad del método propuesto, el rendimiento de clasificación de la propuesta fue comparado con el método más relevante y robusto de la literatura [305]. Específicamente, en [305] proponen extraer una única característica, denotada como "Peak Energy Ratio (PER)", a partir de un tipo de transformada wavelet, denotada como "RAtional Dilation Wavelet Transform (RADWT)", para discriminar entre sibilancias monofónicas y polifónicas aplicando diferentes clasificadores típicos de la literatura (SVM, KNN y ELM). El método propuesto fue comparado con los resultados obtenidos por cada uno de los clasificadores utilizando un esquema de validación cruzada, denotado como "Leave One-Out (LOO)". Sin embargo, el método propuesto, al no necesitar una fase de entrenamiento, fue evaluado a partir de la base de datos creada sin seguir ningún esquema de validación cruzada. Los resultados obtenidos en la etapa experimental permitieron extraer las siguientes conclusiones: i) El método propuesto proporciona los mejores resultados de clasificación de sibilancias en general en comparación con el método más relevante del estado del arte. En concreto, la propuesta consigue una mejora, en términos de ACC_G, de 8.25 % comparándolo con el clasificador SVM, de 12 % comparándolo con el clasificador KNN y de 10.5 % comparándolo con el clasificador ELM. Esto demuestra que una clasificación basada en la localización espectral de las componentes sibilantes es más fiable que la basada en la energía de las componentes; ii) La mejora obtenida por el enfoque propuesto sobre el método de referencia, considerando las tasas de precisión individuales (ACC_P y ACC_M), demuestra la capacidad de la propuesta para clasificar correctamente ambos tipos de sibilancia. En concreto, la propuesta obtiene unas tasas de 92.5 %, en términos de ACC_M, y 91.5 %, en términos de ACC_P; iii) Los resultados obtenidos, en términos de ACC_{M1} y ACC_{M2}, muestran la capacidad de la propuesta para distinguir correctamente ambos tipos de sibilancias monofónicas (sibilancia monofónica con un único pico espectral o sibilancia monofónica con un pico espectral más sus armónicos). Específicamente, la propuesta obtiene unas tasas de 91 %, en términos de ACC_{M1}, y 94 %, en términos de ACC_{M2}; iv) A

diferencia de los métodos basados en clasificadores, el método propuesto es un método no supervisado (ciego) que no requiere ningún tipo de entrenamiento; y v) El número de segmentos sibilantes (400 segmentos) utilizados en la etapa experimental de la publicación es significativamente mayor que el del resto de trabajos de la literatura. Por ejemplo, el método más relevante del estado del arte utilizó 300 segmentos [305].

En conclusión y tomando como referencia los resultados obtenidos, el método propuesto puede ser una herramienta útil en la clasificación de los sonidos sibilantes para diferenciar distintas patologías respiratorias (asma y EPOC) y determinar su grado de severidad. Por último, en esta publicación se propuso como trabajo futuro el diseño de nuevas regularizaciones para evaluar la potencialidad de NMF en el modelado de otros sonidos adventicios (crepitaciones, estridor, roncus y frote pleural).

5.4. Eliminación del ruido ambiente en el proceso de auscultación

El ruido ambiental que rodea al paciente durante la auscultación supone una interferencia acústica en la escucha de los sonidos biomédicos de interés. En este sentido, se ha manifestado que el ruido ambiental implica una de las principales limitaciones que disminuye drásticamente la capacidad cognitiva del médico para examinar los sonidos biomédicos de interés durante la auscultación, especialmente en escenarios con un alto carácter ruidoso, como ambulancias o helicópteros de rescate. En este sentido, la publicación [P7] presenta un método que aborda esta limitación.

5.4.1. Publicación [P7]

En [P7] se propone un algoritmo incremental, denominado 2C-NMPCF, para mejorar la calidad acústica de los sonidos biomédicos capturados durante el proceso de auscultación eliminando el ruido ambiental que los interfiere. En concreto, el modelo propuesto adapta el enfoque NMPCF convencional a un escenario multicanal compuesto por dos canales de entrada monocanal, los cuales capturan audio simultáneamente (grabación interna y externa). Por un lado, el primer canal corresponde a la grabación interna, la cual simula el audio capturado con un estetoscopio en el que se pueden escuchar tanto los sonidos biomédicos del interior del cuerpo humano, como los ruidos ambientales que rodean al paciente. Por otro lado, el segundo canal corresponde a la grabación externa, la cual simula el audio capturado con un micrófono externo en el que sólo se capta el ruido ambiental que rodea al sujeto. Considerando esto, la primera contribución adapta el enfoque NMPCF a un punto de vista multicanal asumiendo que los ruidos ambientales pueden ser modelados como sonidos repetitivos que pueden encontrarse simultáneamente en ambos canales de entrada (estetoscopio y micrófono externo). La segunda contribución propone un algoritmo incremental, basado en el anterior enfoque NMPCF multicanal, que mejora el espectrograma biomédico estimado, a lo largo de las etapas incrementales,

eliminando la mayor parte del ruido ambiental que no se eliminó en la etapa incremental anterior mientras se preserva la mayor parte del contenido espectral biomédico de interés.

Debido a la falta de bases de datos públicas compuestas de sonidos biomédicos mezclados con ruido ambiental, en [P7] se creó una base de datos D_C para simular las señales sonoras capturadas por un estetoscopio durante la auscultación. En concreto, la base de datos D_C fue creada a partir de una base de datos de ruido ambiental D_N y una base de datos de sonidos biomédicos D_B. Por un lado, D_N fue creada a partir de un conjunto amplio de 5 tipos de ruido ambiental: sirena de ambulancia, llanto de bebé, murmullo (personas hablando), coches (interior del vehículo para simular la auscultación en ambulancias) y ruidos de la calle (circulación de coches, motor del coche en marcha, autobuses, camiones, niños y niñas gritando, gente hablando o trabajos de obra). Según la información proporcionada por el personal médico del Hospital de Jaén (España), estos ruidos fueron definidos como los más molestos para un médico que pueden aparecer durante el proceso de auscultación. En total, D_N se compone de 150 señales sonoras, 30 para cada tipo de ruido. Por otro lado, D_B está compuesta por 150 señales biomédicas (sin ningún tipo de sonido ambiental presente) obtenidas de bases de datos públicas y privadas, específicamente 75 señales de sonidos cardíacos y 75 señales de sonidos pulmonares (donde se incluyen sonidos respiratorios normales con sibilancias ausentes o presentes). Considerando las bases de datos D_N y D_B, cada grabación correspondiente a D_C fue generada mezclando cada grabación biomédica de D_B con una grabación de cada tipo de ruido elegida aleatoriamente de D_N. Generando un total de 750 señales sonoras (150 señales biomédicas por 5 tipos de ruido). Teniendo en cuenta la base de datos definida D_C, dos bases de datos fueron creadas posteriormente:

- Una base de datos de optimización D_O, la cual se compone de dos tercios de todas las mezclas de la base de datos D_C (500 mezclas).

- Una base de datos de evaluación D_T, la cual se compone de un tercio de todas las mezclas de la base de datos D_C (250 mezclas). Además, se crearon diferentes bases de datos variando la SNR para evaluar la robustez de los métodos comparados. En este sentido, las bases de datos $D_{T_{-20}}$ (SNR=-20 dB), $D_{T_{-15}}$ (SNR=-15 dB), $D_{T_{-10}}$ (SNR=-10 dB) y $D_{T_{-5}}$ (SNR=-5 dB) se refieren a la misma base de datos D_T, pero usando diferentes SNR en la mezcla de la señal biomédica y del ruido ambiental.

El método propuesto fue comparado con uno de los algoritmos más relevantes del estado del arte, basado en sustracción espectral multibanda MSS [99]. Los resultados obtenidos en la etapa experimental permitieron extraer las siguientes conclusiones: i) En términos generales, el método propuesto proporciona el mejor rendimiento para eliminar el ruido ambiental en comparación con el algoritmo de referencia, considerando todos los escenarios evaluados (distintos tipos de ruido y distintas SNR); ii) El método propuesto muestra su mayor robustez en aquellos escenarios sonoros en los que el ruido ambiental y los sonidos biomédicos tienen un comportamiento espectral distinto. En concreto, el método propuesto obtiene el mejor rendimiento eliminando el ruido ambiental compuesto por la sirena de una ambulancia o el llanto

de un bebé, ya que las características tiempo-frecuencia de estos sonidos son muy disímiles en comparación con los sonidos biomédicos evaluados. iii) Por último se evaluó un escenario donde se consideraba el efecto de insertar un retardo temporal entre el primer canal (estetoscopio) y el segundo canal (micrófono externo). El propósito de este retardo es simular el tiempo que lleva a cabo un estetoscopio digital al aplicar un filtrado u otras operaciones de procesado. Este estudio reveló, que la ventaja más notable de la propuesta radica en su fortaleza ante la variación del retardo que puede existir entre los dos canales de entrada.

Finalmente, en esta publicación se propuso como trabajo futuro el desarrollo de algoritmos aplicados a la caracterización de algunos de los ruidos ambientales más importantes que pueden existir en determinadas situaciones de emergencia clínica, como el ruido encontrado dentro de los helicópteros de rescate.

5.5. Conclusiones generales

En términos generales, se han desarrollado diferentes métodos y algoritmos destinados a abordar las tareas más relevantes en el análisis de los sonidos sibilantes. En concreto, estos métodos permiten proporcionar una vía de información complementaria al médico que ayude a identificar patologías respiratorias derivadas de los sonidos sibilantes (por ejemplo: asma, bronquiolitis, bronquitis, bronquiectasia o EPOC) y aumente la fiabilidad del diagnóstico emitido al analizar las señales acústicas capturadas mediante el proceso de auscultación. En este sentido, las diferentes publicaciones científicas aportadas en esta Tesis presentan algoritmos que tratan de hacer frente a las principales dificultades a las que un especialista (neumólogo) tiene que enfrentarse a diario en consulta: i) eliminación del ruido ambiental que rodea al paciente durante la auscultación [P7]; ii) separación y mejora acústica de los sonidos sibilantes, solapados con los sonidos respiratorios normales [P1][P4]; iii) detección de la presencia o ausencia de sibilancias [P3]; iv) localización de los intervalos temporales en los cuales las sibilancias se encuentran activas durante el proceso de respiración [P2][P5]; y v) clasificación del tipo de sibilancia considerando su estructura armónica, para determinar el nivel de obstrucción de la vía respiratoria y la patología respiratoria asociada (asma o EPOC) [P6]. En conclusión, la unión o combinación de todas las propuestas permite crear un sistema completo para extraer toda la información redundante de los sonidos sibilantes y así reducir la tasa de falsos positivos o falsos negativos en la detección de posibles patologías respiratorias, evitando poner en riesgo la salud de los pacientes y disminuyendo el coste asociado a los centros de salud.

Teniendo en consideración, las posibilidades que los modelos de descomposición de matrices no negativas (NMF o NMPCF) pueden ofrecer para modelar las características tiempo-frecuencia de las señales sonoras y que estos enfoques matemáticos nunca antes habían sido utilizados en el procesado de los sonidos sibilantes, los diferentes métodos propuestos en las publicaciones se han basado en NMF o NMPCF demostrando la potencialidad de las técnicas de factorización de matrices no negativas en este ámbito científico. Por último, considerando la problemática relacionada con la carencia de bases de datos de sonidos sibilantes, las dife-

rentes propuestas presentadas en esta Tesis proponen enfoques no supervisados (ciegos) con el objeto de prescindir de una etapa de entrenamiento y ser herramientas adecuadas en problemas de este tipo en los cuales, existe escasez de datos disponibles para aprender a partir de grandes cantidades de datos a utilizar en etapas de entrenamiento.

5.6. Líneas futuras

En el contexto del procesado de señal aplicado a señales sonoras procedentes del proceso de la auscultación, la comunidad científica sigue mostrando un creciente interés en el desarrollo de nuevos métodos y algoritmos que ayuden a mejorar la fiabilidad de los diagnósticos orientados a la detección precoz de patologías respiratorias. Esta Tesis ha demostrado que la investigación centrada en los sonidos adventicios, en particular, los sonidos sibilantes, es un reto en el que aún queda mucho camino por recorrer tanto desde el punto de vista de la detección y clasificación de sonidos adventicios, como desde el punto de vista de la mejora del rendimiento y coste computacional de los métodos actuales. En esta sección se presentan las principales líneas de trabajo futuras que, aunque no han sido abordadas en el desarrollo de esta Tesis doctoral, podrían presentar contribuciones relevantes en el diagnóstico de enfermedades obstructivas respiratorias.

- Desarrollo de métodos aplicados al análisis de sonidos sibilantes en tiempo real. Para ello, se propone diseñar enfoques alternativos basados en NMF para adaptar los algoritmos y métodos propuestos en esta Tesis a un entorno en tiempo real. La finalidad que se persigue es desarrollar herramientas que aporten al neumólogo aquella información relevante asociada a los sonidos sibilantes durante la ejecución del examen clínico derivado de la auscultación.

- Desarrollo de un sistema de tele-monitorización sibilante para dispositivos móviles. Con ello, se pretende desarrollar una herramienta de monitorización fiable, no-invasiva, de bajo coste e individual que pueda ser realizada desde el hogar del sujeto, es decir, de fácil acceso a la gran masa de la población (fundamentalmente, al sector de personas más vulnerables como es el caso de ancianos y niños pequeños) y que pueda ser realizada tantas veces como el sujeto lo considere oportuno. Esto ayudará a capturar acústicamente los sonidos aparecidos durante las típicas crisis respiratorias que suelen ocurrir de madrugada, así como permitir un sistema sanitario sostenible desde el punto de vista de recursos humanos y materiales.

- Diseño de nuevas regularizaciones tiempo-frecuencia que permitan el modelado de diferentes sonidos adventicios, por ejemplo, crackles, con el fin de desarrollar una herramienta fiable para proporcionar una probabilidad acerca de la presencia de ciertas patologías respiratorias caracterizadas por la presencia de dichos sonidos adventicios.

- Desarrollo de técnicas de procesado de señal, basadas en algoritmos NMF complejos (Complex NMF), aplicadas a señales sonoras multicanal utilizando sistemas multi-sensor,

con varios sensores acústicos situados simultáneamente en diferentes zonas del cuerpo del sujeto, con el fin de localizar espacialmente (3D) la ubicación de los distintos tipos de sonidos adventicios capturados durante la auscultación. Este hecho implica un avance significativo en el diagnóstico proporcionado por la auscultación ya que no sólo se obtendría el sonido sino la zona espacial responsable de su aparición, lo cual está directamente relacionado con determinadas patologías que tienen su foco origen en zonas muy concretas del sistema respiratorio.

- Desarrollo de métodos robustos para cancelar el ruido ambiental en escenarios altamente ruidosos, por ejemplo, el interior de una ambulancia en estado de emergencia o el interior de un helicóptero de rescate. Se propone el diseño de algoritmos expertos que permitan modelar el comportamiento espectro-temporal de dichos sonidos ruidosos y eliminarlos de la señal biomédica de interés.

Bibliografía

[1] 3m Littmann. https://www.3m.com.es/3M/es_ES/Littmann-ES/.

[2] 3M Littmann Electronic Stethoscope Model 3200. https://www.littmann.com/3M/en_US/littmann-stethoscopes/products/~/3M-Littmann-Electronic-Stethoscope-Model-3200/?N=5932256+8711017+3293188392+3294857497&rt=rud.

[3] 3m littmann stethoscopes. http://www.3m.com/healthcare/littmann/lung.html.

[4] Colorado State University. http://www.cvmbs.colostate.edu/clinsci/callan/breath_sounds.htm.

[5] CORE Digital Attachment. https://shop.ekohealth.com/products/core-digital-attachment.

[6] CORE Digital Stethoscope. https://shop.ekohealth.com/products/3m-littmann-core-digital-stethoscope.

[7] E-learning resources. https://www.ers-education.org/e-learning/reference-database-of-respiratory-sounds.aspx.

[8] East tennessee state university pulmonary breath sounds. http://faculty.etsu.edu.

[9] Easy Auscultation. https://www.easyauscultation.com/lung-sounds-reference-guide.

[10] Eko. https://www.ekohealth.com.

[11] Ekuore. https://www.ekuore.com/es/.

[12] Electronic stethoscope eKuore Pro. https://www.ekuore.com/es/estetoscopio-electronico-pro/.

[13] Emedicine/Medscape. https://emedicine.medscape.com/article/1894146-

overview#a3.

[14] EUROSTAT. Respiratory diseases statistics. https://ec.europa.eu/eurostat/statistics-explained/index.php/Respiratory_diseases_statistics#Deaths_from_diseases_of_the_respiratory_system.

[15] Historia de los estetoscopios Littmann. https://www.littmann.com/3M/en_US/littmann-stethoscopes/education-center/history/.

[16] ICBHI 2017 Challenge. https://bhichallenge.med.auth.gr.

[17] Instituto Nacional de Estadística. Muertes por enfermedades respiratorias. https://www.epdata.es/muertes-enfermedades-respiratorias/80c722a8-043b-4fd2-a662-5ed70427e543.

[18] Lippincott NursingCenter. https://www.nursingcenter.com.

[19] Littmann. Anatomy of a Stethoscope. https://www.littmann.in/3M/en_IN/littmann-stethoscopes-in/education/how-to-choose/anatomy/.

[20] MERCK MANUAL. Wheezing. https://www.merckmanuals.com/home/lung-and-airway-disorders/symptoms-of-lung-disorders/wheezing.

[21] National Heart, Lung, and Blood Institute. Bronchitis. https://www.nhlbi.nih.gov/health-topics/bronchitis.

[22] National Heart, Lung, and Blood Institute. Chronic Obstructive Pulmonary Disease (COPD). https://www.nhlbi.nih.gov/health-topics/copd.

[23] Organización Mundial de la Salud. Asma. https://www.who.int/es/news-room/fact-sheets/detail/asthma.

[24] Organización Mundial de la Salud. Enfermedad pulmonar obstructiva crónica (EPOC). https://www.who.int/respiratory/copd/es/.

[25] Organización Mundial de la Salud. La OMS destaca la enorme magnitud de la mortalidad por enfermedades pulmonares relacionadas con el tabaco. https://www.who.int/es/news-room/detail/29-05-2019-who-highlights-huge-scale-of-tobacco-related-lung-disease-deaths.

[26] Organización Mundial de la Salud. ¿Qué consecuencias sanitarias acarrea la contaminación atmosférica urbana? https://www.who.int/phe/health_topics/outdoorair/databases/health_impacts/es/.

[27] Respiratory wiki. http://respwiki.com/Breath_sounds.

[28] SEMERGEN. Coste asociado a los diagnósticos erróneos. https://www.semergen.es/.

[29] SoundCloud. Lung Sounds. https://soundcloud.com/search?q=lung%20sounds.

[30] Stethographics lung sound samples. http://www.stethographics.com.

[31] The r.a.l.e. repository. http://www.rale.ca.

[32] Thinklabs Digital Stethoscope. https://www.thinklabs.com.

[33] Thinklabs ONE. https://www.thinklabs.com.

[34] Thinklabs youtube. https://www.youtube.com/channel/UCzEbKuIze4AI1523_AWiK4w.

[35] Wikimedia Commons, conducting passages.svg. https://commons.wikimedia.org/wiki/File:Illu_conducting_passages.svg?uselang=es.

[36] Methods for calculation of the speech intelligibility index. *American National Standard Institute*, 1997. ANSI-S3.5-1997-R2007.

[37] A. Abbas and A. Fahim. An automated computerized auscultation and diagnostic system for pulmonary diseases. *Journal of medical systems*, 34(6):1149–1155, 2010.

[38] M. Abella, J. Formolo, and D. G. Penney. Comparison of the acoustic properties of six popular stethoscopes. *The Journal of the Acoustical Society of America*, 91(4):2224–2228, 1992.

[39] M. Albers, T. Schermer, G. van den Boom, R. Akkermans, C. van Schayck, C. van Herwaarden, and C. van Weel. Efficacy of inhaled steroids in undiagnosed subjects at high risk for copd: results of the detection, intervention, and monitoring of copd and asthma program. *Chest*, 126(6):1815–1824, 2004.

[40] A. Alic, I. Lackovic, V. Bilas, D. Sersic, and R. Magjarevic. A novel approach to wheeze detection. In *World Congress on Medical Physics and Biomedical Engineering 2006*, pages 963–966. Springer, 2007.

[41] D. Anantham, F. J. Herth, A. Majid, G. Michaud, and A. Ernst. Vibration response imaging in the detection of pleural effusions: a feasibility study. *Respiration*, 77(2):166–172, 2009.

[42] E. Andrès, R. Gass, A. Charloux, C. Brandt, and A. Hentzler. Respiratory sound analysis in the era of evidence-based medicine and the world of medicine 2.0. *Journal of medicine and life*, 11(2):89, 2018.

[43] E. Andrès, A. Hajjam, and C. Brandt. Advances and innovations in the field of auscultation, with a special focus on the development of new intelligent communicating stethoscope systems. *Health and Technology*, 2(1):5–16, 2012.

[44] S. Aydore, I. Sen, Y. P. Kahya, and M. K. Mihcak. Classification of respiratory signals by linear analysis. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2617–2620. IEEE, 2009.

[45] R. Badeau, V. Emiya, and B. David. Expectation-maximization algorithm for multi-pitch estimation and separation of overlapping harmonic spectra. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3073–3076. IEEE, 2009.

[46] R. Báez Saldaña, S. Monraz Pérez, P. Castillo González, U. Rumbo Nava, R. García Torrentera, R. Ortíz Siordia, and T. I. Fortoul van der Goes. La exploración del tórax: una

guía para descifrar sus mensajes. *Revista de la Facultad de Medicina UNAM*, 59(6):43–57, 2016.

[47] M. Bahoura. Pattern recognition methods applied to respiratory sounds classification into normal and wheeze classes. *Computers in biology and medicine*, 39(9):824–843, 2009.

[48] M. Bahoura and C. Pelletier. Respiratory sounds classification using gaussian mixture models. In *Canadian Conference on Electrical and Computer Engineering 2004 (IEEE Cat. No. 04CH37513)*, volume 3, pages 1309–1312. IEEE, 2004.

[49] D. Bansal, B. Raj, and P. Smaragdis. Bandwidth expansion of narrowband speech using non-negative matrix factorization. In *Ninth European Conference on Speech Communication and Technology*, 2005.

[50] E. D. Bateman, S. Hurd, P. Barnes, J. Bousquet, J. Drazen, M. FitzGerald, P. Gibson, K. Ohta, P. O'byrne, S. Pedersen, et al. Global strategy for asthma management and prevention: Gina executive summary. *European Respiratory Journal*, 31(1):143–178, 2008.

[51] R. P. Baughman and R. G. Loudon. Stridor: Differentiation from asthma or upper airway noise1-3. *Am Rev Respir Dis*, 139:1407–1409, 1989.

[52] H. D. Becker. Vibration response imaging-finally a real stethoscope. *Respiration*, 77(2):236, 2009.

[53] F. Belloni, D. D. Giustina, M. Riva, and M. Malcangi. A new digital stethoscope with environmental noise cancellation. In *Proceedings of the 12th WSEAS International Conference on Mathematical and Computational Methods in Science and Engineering, Faro, Portugal*, pages 3–5. Citeseer, 2010.

[54] T. Bergstresser, D. Ofengeim, A. Vyshedskiy, J. Shane, and R. Murphy. Sound transmission in the lung as a function of lung volume. *Journal of Applied Physiology*, 93(2):667–674, 2002.

[55] N. Bertin, R. Badeau, and E. Vincent. Enforcing harmonicity and smoothness in bayesian non-negative matrix factorization applied to polyphonic music transcription. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):538–549, 2010.

[56] P. Besse and J. O. Ramsay. Principal components analysis of sampled functions. *Psychometrika*, 51(2):285–311, 1986.

[57] M. Blanco, R. Mor, A. Fraticelli, D. P. Breen, and H. Dutau. Distribution of breath sound images in patients with pneumothoraces compared to healthy subjects. *Respiration*, 77(2):173–178, 2009.

[58] A. Bohadana, G. Izbicki, and S. S. Kraman. Fundamentals of lung auscultation. *New England Journal of Medicine*, 370(8):744–751, 2014.

[59] P. Bokov, B. Mahut, P. Flaud, and C. Delclaux. Wheezing recognition algorithm using recordings of respiratory sounds at the mouth in a pediatric population. *Computers in*

Biology and Medicine, 70:40–50, 2016.

[60] L. H. Brown, J. E. Gough, D. M. Bryan-Berg, and R. C. Hunt. Assessment of breath sounds during ambulance transport. *Annals of emergency medicine*, 29(2):228–231, 1997.

[61] P. Cabañas-Molero, D. Martínez-Muñoz, P. Vera-Candeas, F. J. Cañadas-Quesada, and N. Ruiz-Reyes. Compositional model for speech denoising based on source/filter speech representation and smoothness/sparseness noise constraints. *Speech Communication*, 78:84–99, 2016.

[62] F. Canadas-Quesada, N. Ruiz-Reyes, J. Carabias-Orti, P. Vera-Candeas, and J. Fuertes-Garcia. A non-negative matrix factorization approach based on spectro-temporal clustering to extract heart sounds. *Applied Acoustics*, 125:7–19, 2017.

[63] F. J. Cañadas-Quesada, P. Vera-Candeas, D. Martinez-Munoz, N. Ruiz-Reyes, J. J. Carabias-Orti, and P. Cabanas-Molero. Constrained non-negative matrix factorization for score-informed piano music restoration. *Digital Signal Processing*, 50:240–257, 2016.

[64] F. J. Canadas-Quesada, P. Vera-Candeas, N. Ruiz-Reyes, J. Carabias-Orti, and P. Cabanas-Molero. Percussive/harmonic sound separation by non-negative matrix factorization with smoothness/sparseness constraints. *EURASIP Journal on Audio, Speech, and Music Processing*, 2014(1):26, 2014.

[65] J. Carabias-Orti, F. Canadas-Quesada, P. Vera-Candeas, and N. Ruiz-Reyes. Non-negative matrix factorization (nmf) applied to monaural audio signal processing. In *Independent Component Analysis (ICA): Algorithms, Applications and Ambiguities*, chapter 7, pages 247–327. Nova, 2018.

[66] J. J. Carabias-Orti, F. J. Rodríguez-Serrano, P. Vera-Candeas, F. J. Cañadas-Quesada, and N. Ruiz-Reyes. Constrained non-negative sparse coding using learnt instrument templates for realtime music transcription. *Engineering Applications of Artificial Intelligence*, 26(7):1671–1680, 2013.

[67] J. J. Carabias-Orti, T. Virtanen, P. Vera-Candeas, N. Ruiz-Reyes, and F. J. Canadas-Quesada. Musical instrument sound multi-excitation model for non-negative spectrogram factorization. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1144–1158, 2011.

[68] G.-C. Chang and Y.-F. Lai. Performance evaluation and enhancement of lung sound recognition system in two real noisy environments. *Computer methods and programs in biomedicine*, 97(2):141–150, 2010.

[69] S. Charleston-Villalobos, S. Cortes-Rubiano, R. González-Camerena, G. Chi-Lem, and T. Aljama-Corrales. Respiratory acoustic thoracic imaging (rathi): assessing deterministic interpolation techniques. *Medical and Biological Engineering and Computing*, 42(5):618–626, 2004.

[70] S. Charleston-Villalobos, L. Dominguez-Robert, R. Gonzalez-Camarena, and A. Aljama-Corrales. Heart sounds interference cancellation in lung sounds. In *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1694–1697. IEEE, 2006.

[71] C.-H. Chen, W.-T. Huang, T.-H. Tan, C.-C. Chang, and Y.-J. Chang. Using k-nearest neighbor classification to diagnose abnormal lung sounds. *Sensors*, 15(6):13132–13158, 2015.

[72] S.-F. Chen. Contact type electronic stethoscope with a noise interference resisting function for auscultation, Oct. 12 2006. US Patent App. 11/100,438.

[73] Z. Chen, A. Cichocki, and T. M. Rutkowski. Constrained non-negative matrix factorization method for eeg analysis in early detection of alzheimer disease. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 5, pages V–V. IEEE, 2006.

[74] J.-C. Chien, H.-D. Wu, F.-C. Chong, and C.-I. Li. Wheeze detection using cepstral analysis in gaussian mixture models. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3168–3171. IEEE, 2007.

[75] S. H. Chotirmall and J. D. Chalmers. Bronchiectasis: an emerging global epidemic, 2018.

[76] H. Chung, E. Plourde, and B. Champagne. Discriminative training of nmf model based on class probabilities for speech enhancement. *IEEE Signal Processing Letters*, 23(4):502–506, 2016.

[77] A. Cichocki and A.-H. Phan. Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, 92(3):708–721, 2009.

[78] A. Cichocki, R. Zdunek, and S.-i. Amari. Csiszar's divergences for non-negative matrix factorization: Family of new algorithms. In *International Conference on Independent Component Analysis and Signal Separation*, pages 32–39. Springer, 2006.

[79] A. Cohen and A. Berstein. Acoustic transmission of the respiratory system using speech stimulation. *IEEE transactions on biomedical engineering*, 38(2):126–132, 1991.

[80] T. V. Colby. Bronchiolitis: pathologic considerations. *American journal of clinical pathology*, 109(1):101–109, 1998.

[81] R. J. Copt, J. G. Butera III, R. J. Summers, et al. Noise reduction assembly for auscultation of a body, July 6 2017. US Patent App. 15/403,598.

[82] S. Cortes, R. Jane, J. Fiz, and J. Morera. Monitoring of wheeze duration during spontaneous respiration in asthmatic patients. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 6141–6144. IEEE, 2006.

[83] J. S. Coviello. *Auscultation skills: breath & heart sounds*. Lippincott Williams & Wilkins, 2013.

[84] F. Dalmay, M. Antonini, P. Marquet, and R. Menier. Acoustic properties of the normal chest. *European Respiratory Journal*, 8(10):1761–1769, 1995.

[85] C. Damon, A. Liutkus, A. Gramfort, and S. Essid. Non-negative matrix factorization for single-channel eeg artifact rejection. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1177–1181. IEEE, 2013.

[86] P. B. Davis. Cystic fibrosis since 1938. *American journal of respiratory and critical care medicine*, 173(5):475–482, 2006.

[87] S. Debbal and F. Bereksi-Reguig. Spectral analysis of the pcg signals. *The Internet journal of microbiology*, 2, 2007.

[88] L. Delaunois. Lung auscultation: back to basic medecine, 2005.

[89] D. Della Giustina, M. Riva, F. Belloni, and M. Malcangi. Embedding a multichannel environmental noise cancellation algorithm into an electronic stethoscope. *International Journal of Circuits/System and Signal Processing*, (2), 2011.

[90] F. Demir, A. Sengur, and V. Bajaj. Convolutional neural networks based efficient approach for classification of lung diseases. *Health Information Science and Systems*, 8(1):4, 2020.

[91] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[92] C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 126–135, 2006.

[93] G. Druger. *The chest: its signs and sounds*. Humetrics Corporation, 1973.

[94] T. E. Drummond, H. M. Carim, and C. D. Oster. Stethoscope with frictional noise reduction, Oct. 5 2010. US Patent 7,806,226.

[95] S. P. Dugar, M. Latifi, and E. Mireles-Cabodevila. Respiratory system physiology. In *Basic Sciences in Anesthesia*, pages 329–354. Springer, 2018.

[96] J.-L. Durrieu, B. David, and G. Richard. A musically motivated mid-level representation for pitch estimation and musical audio source separation. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1180–1191, 2011.

[97] J. Earis, K. Marsh, M. Pearson, and C. Ogilvie. The inspiratory "squawk" in extrinsic allergic alveolitis and other pulmonary fibroses. *Thorax*, 37(12):923–926, 1982.

[98] J. Eggert and E. Korner. Sparse coding and nmf. In *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541)*, volume 4, pages 2529–2533. IEEE, 2004.

[99] D. Emmanouilidou, E. D. McCollum, D. E. Park, and M. Elhilali. Adaptive noise suppression of pediatric lung auscultations with real applications to noisy clinical settings in

developing countries. *IEEE Transactions on Biomedical Engineering*, 62(9):2279–2288, 2015.

[100] D. Emmanouilidou, E. D. McCollum, D. E. Park, and M. Elhilali. Computerized lung sound screening for pediatric auscultation in noisy field environments. *IEEE Transactions on Biomedical Engineering*, 65(7):1564–1574, 2017.

[101] S. Emrani, T. Gentimis, and H. Krim. Persistent homology of delay embeddings and its application to wheeze detection. *IEEE Signal Processing Letters*, 21(4):459–463, 2014.

[102] T. R. Fenton, H. Pasterkamp, A. Tal, and V. Chernick. Automated spectral characterization of wheezing in asthmatic children. *IEEE transactions on biomedical engineering*, (1):50–55, 1985.

[103] M. A. Fernandez-Granero, D. Sanchez-Morillo, and A. Leon-Jimenez. Computerised analysis of telemonitored respiratory sounds for predicting acute exacerbations of copd. *Sensors*, 15(10):26978–26996, 2015.

[104] T. Ferns and S. West. The art of auscultation: evaluating a patient's respiratory pathology. *British Journal of Nursing*, 17(12):772–777, 2008.

[105] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural computation*, 21(3):793–830, 2009.

[106] C. Févotte, R. Gribonval, and E. Vincent. Bss_eval toolbox user guide–revision 2.0. 2005. `https://hal.inria.fr/inria-00564760`.

[107] C. Févotte and J. Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural computation*, 23(9):2421–2456, 2011.

[108] D. FitzGerald, M. Cranitch, and E. Coyle. Extended nonnegative tensor factorisation models for musical sound source separation. *Computational Intelligence and Neuroscience*, 2008, 2008.

[109] J. A. Fiz, R. Jané, A. Homs, J. Izquierdo, M. A. Garcia, and J. Morera. Detection of wheezing during maximal forced exhalation in patients with obstructed airways. *Chest*, 122(1):186–191, 2002.

[110] J. A. Fiz, R. Jané, D. Salvatella, J. Izquierdo, L. Lores, P. Caminal, and J. Morera. Analysis of tracheal sounds during forced exhalation in asthma patients and normal subjects: bronchodilator response effect. *Chest*, 116(3):633–638, 1999.

[111] J. S. Fleeter and G. R. Wodicka. Auscultation of heart and lung sounds in high-noise environments using adaptive filters. *The Journal of the Acoustical Society of America*, 104(3):1781–1781, 1998.

[112] J. D. Flesch and C. J. Dine. Lung volumes: measurement, clinical use, and coding. *Chest*, 142(2):506–510, 2012.

[113] H. Fletcher and W. A. Munson. Loudness, its definition, measurement and calculation. *Bell System Technical Journal*, 12(4):377–430, 1933.

[114] T. A. Florin, A. C. Plint, and J. J. Zorc. Viral bronchiolitis. *The Lancet*, 389(10065):211–224, 2017.

[115] P. Forgacs. Crackles and wheezes. *The Lancet*, 290(7508):203–205, 1967.

[116] P. Forgacs. Lung sounds. *British journal of diseases of the chest*, 63(1):1–12, 1969.

[117] P. Forgacs. The functional basis of pulmonary sounds. *Chest*, 73(3):399–405, 1978.

[118] P. Forgacs, A. Nathoo, and H. Richardson. Breath sounds. *Thorax*, 26(3):288–295, 1971.

[119] K. E. Forkheim, D. Scuse, and H. Pasterkamp. A comparison of neural network models for wheeze detection. In *IEEE WESCANEX 95. Communications, Power, and Computing. Conference Proceedings*, volume 1, pages 214–219. IEEE, 1995.

[120] J. Fritsch and M. D. Plumbley. Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 888–891. IEEE, 2013.

[121] S. Gairola, F. Tom, N. Kwatra, and M. Jain. Respirenet: A deep neural network for accurately detecting abnormal lung sounds in limited data setting. *arXiv preprint arXiv:2011.00196*, 2020.

[122] M. T. García-Ordás, J. A. Benítez-Andrades, I. García-Rodríguez, C. Benavides, and H. Alaiz-Moretón. Detecting respiratory pathologies using convolutional neural networks and variational autoencoders for unbalancing data. *Sensors*, 20(4):1214, 2020.

[123] N. Gavriely, Y. Palti, and G. Alroy. Spectral characteristics of normal breath sounds. *Journal of applied physiology*, 50(2):307–314, 1981.

[124] D. M. Gedde, B. Corrin, D. A. Brewerton, R. J. Davies, and M. Turner-Warwick. Progressive airway obliteration in adults and its association with rheumatoid disease. *QJM: An International Journal of Medicine*, 46(4):427–444, 1977.

[125] J. Geiger, F. Weninger, A. Hurmalainen, J. Gemmeke, M. Wöllmer, B. Schuller, G. Rigoll, and T. Virtanen. The tum+ tut+ kul approach to the chime challenge 2013: Multistream asr exploiting blstm networks and sparse nmf. *Proceedings CHiME 2013*, pages 25–30, 2013.

[126] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen. Exemplar-based sparse representations for noise robust automatic speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2067–2080, 2011.

[127] G. J. Gibson, R. Loddenkemper, Y. Sibille, and B. Lundbäck. *European lung white book*. European Respiratory Society, 2013.

[128] J. J. Goldberger and J. Ng. *Practical signal and image processing in clinical cardiology*. Springer, 2010.

[129] E. M. Grais and H. Erdogan. Discriminative nonnegative dictionary learning using cross-coherence penalties for single channel source separation. In *Interspeech*, pages 808–812, 2013.

[130] D. Groom. The effect of background noise on cardiac auscultation. *American Heart Journal*, 52(5):781–790, 1956.

[131] V. Gross, A. Dittmar, T. Penzel, F. Schuttler, and P. Von Wichert. The relationship between normal lung sounds, age, and gender. *American journal of respiratory and critical care medicine*, 162(3):905–909, 2000.

[132] J. B. Grotberg and N. Gavriely. Flutter in collapsible tubes: a theoretical model of wheezes. *Journal of Applied Physiology*, 66(5):2262–2273, 1989.

[133] K. K. Guntupalli, R. M. Reddy, R. H. Loutfi, P. M. Alapat, V. D. Bandi, and N. A. Hanania. Evaluation of obstructive lung disease with vibration response imaging. *Journal of Asthma*, 45(10):923–930, 2008.

[134] T. R. Harley. Active noise control stethoscope, July 23 1996. US Patent 5,539,831.

[135] A. Hashemi, H. Arabalibiek, and K. Agin. Classification of wheeze sounds using wavelets and neural networks. In *International Conference on Biomedical Engineering and Technology*, volume 11, pages 127–131. IACSIT Press, 2011.

[136] S. Haykin. *Adaptive filter theory*. Prentice-Hall, Inc., 1996.

[137] J. He and W.-S. Gan. Multi-shift principal component analysis based primary component extraction for spatial audio reproduction. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 350–354. IEEE, 2015.

[138] T. Heittola, A. Klapuri, and T. Virtanen. Musical instrument recognition in polyphonic audio using source-filter model for sound separation. In *ISMIR*, pages 327–332, 2009.

[139] M. Helen and T. Virtanen. Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. In *2005 13th European Signal Processing Conference*, pages 1–4. IEEE, 2005.

[140] B. Henry and T. J. Royston. Localization of adventitious respiratory sounds. *The Journal of the Acoustical Society of America*, 143(3):1297–1307, 2018.

[141] A. Hernando, C. Guillamas, E. Gutiérrez, G. Sánchez-Cascado, L. Tordesillas, and M. J. Méndez. *Técnicas básicas de enfermería. Novedad 2017*. Editex, 2017.

[142] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, 1999.

[143] A. Homs-Corbera, J. A. Fiz, J. Morera, and R. Jané. Time-frequency detection and analysis of wheezes during forced exhalation. *IEEE Transactions on Biomedical Engineering*, 51(1):182–186, 2004.

[144] A. Houtsma. High noise environment stethoscope, Dec. 21 2006. US Patent App. 11/425,312.

[145] D. Howell. Signs of respiratory disease: lung sounds. *Oxford: Academic Press*, pages 35–41, 2006.

[146] P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, 5:1457–1469, Dec. 2004.

[147] Y. Hu and G. Liu. Separation of singing voice using nonnegative matrix partial co-factorization for singer identification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(4):643–653, 2015.

[148] R. C. Hunt, D. M. Bryan, V. S. Brinkley, T. W. Whitley, and N. H. Benson. Inability to assess breath sounds during air medical transport by helicopter. *Jama*, 265(15):1982–1984, 1991.

[149] N. Hurley and S. Rickard. Comparing measures of sparsity. *IEEE Transactions on Information Theory*, 55(10):4723–4741, 2009.

[150] A. Hyvarinen, J. Karhunen, and E. Oja. Independent component analysis and blind source separation, 2001.

[151] J. B. Ida and D. M. Thompson. Pediatric stridor. *Otolaryngologic Clinics of North America*, 47(5):795–819, 2014.

[152] C. M. Ionescu. The human respiratory system. In *The Human Respiratory System*, pages 13–22. Springer, 2013.

[153] R. S. Irwin, P. J. Barnes, and H. Hollingsworth. Evaluation of wheezing illnesses other than asthma in adults. *UpToDate. Waltham: UpToDate*, 2013.

[154] M. Iskander. Burnout, cognitive overload, and metacognition in medicine. *Medical Science Educator*, 29(1):325–328, 2019.

[155] C. Jácome, A. Oliveira, and A. Marques. Computerized respiratory sounds: a comparison between patients with stable and exacerbated copd. *The clinical respiratory journal*, 11(5):612–620, 2017.

[156] A. Jain and J. Vepa. Lung sound analysis for wheeze episode detection. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2582–2585. IEEE, 2008.

[157] R. Jané, S. Cortes, J. Fiz, and J. Morera. Analysis of wheezes in asthmatic patients during spontaneous respiration. In *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 2, pages 3836–3839. IEEE, 2004.

[158] S. Jia and Y. Qian. Constrained nonnegative matrix factorization for hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 47(1):161–173, 2008.

[159] F. Jin, F. Sattar, and D. Y. Goh. Automatic wheeze detection using histograms of sample entropy. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1890–1893. IEEE, 2008.

[160] K. S. S. F. Jin, F. Adventitious sounds identification and extraction using temporal–spectral dominance-based features. *IEEE Transactions On Biomedical Engineering*, 58(11):3078–3087, 2011.

[161] C. Joder, F. Weninger, F. Eyben, D. Virette, and B. Schuller. Real-time speech separation by semi-supervised nonnegative matrix factorization. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 322–329. Springer, 2012.

[162] A. Jones. A brief overview of the analysis of lung sounds. *Physiotherapy*, 81(1):37–42, 1995.

[163] M. Jones and D. Thomas. Acoustic detection of physiological change within the thorax. In *International Conference on Acoustic Sensing and Imaging*, pages 201–205. IET, 1993.

[164] A. Kala, A. Husain, E. D. McCollum, and M. Elhilali. An objective measure of signal quality for pediatric lung auscultations. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 772–775. IEEE, 2020.

[165] A. Kandaswamy, C. S. Kumar, R. P. Ramanathan, S. Jayaraman, and N. Malmurugan. Neural classification of lung sounds using wavelet coefficients. *Computers in biology and medicine*, 34(6):523–537, 2004.

[166] B. Kara Rogers Senior Editor. *The Respiratory System*. The Human Body. Britannica Educational Pub., 2010.

[167] J. M. Kates and K. H. Arehart. Coherence and the speech intelligibility index. *The journal of the acoustical society of America*, 117(4):2224–2237, 2005.

[168] J. Kim and H. Park. Fast nonnegative matrix factorization: An active-set-like method and comparisons. *SIAM Journal on Scientific Computing*, 33(6):3261–3281, 2011.

[169] M. Kim, J. Yoo, K. Kang, and S. Choi. Blind rhythmic source separation: Nonnegativity and repeatability. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2006–2009. IEEE, 2010.

[170] M. Kim, J. Yoo, K. Kang, and S. Choi. Nonnegative matrix partial co-factorization for spectral and temporal drum source separation. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1192–1204, 2011.

[171] T. E. King. Bronchiolitis obliterans. *Lung*, 167(1):69–93, 1989.

[172] D. Kitamura, N. Ono, H. Saruwatari, Y. Takahashi, and K. Kondo. Discriminative and reconstructive basis training for audio source separation with semi-supervised nonnegative matrix factorization. In *2016 IEEE International Workshop on Acoustic Signal*

Enhancement (IWAENC), pages 1–5. IEEE, 2016.

[173] D. H. Klatt and L. C. Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America*, 87(2):820–857, 1990.

[174] K. Kochetov, E. Putin, S. Azizov, I. Skorobogatov, and A. Filchenkov. Wheeze detection using convolutional neural networks. In *EPIA Conference on Artificial Intelligence*, pages 162–173. Springer, 2017.

[175] X. H. Kok, S. A. Imtiaz, and E. Rodriguez-Villegas. A novel method for automatic identification of respiratory disease from acoustic recordings. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2589–2592. IEEE, 2019.

[176] R. Kompass. A generalized divergence measure for nonnegative matrix factorization. *Neural computation*, 19(3):780–791, 2007.

[177] M. Kompis, H. Pasterkamp, and G. R. Wodicka. Acoustic imaging of the human chest. *Chest*, 120(4):1309–1321, 2001.

[178] S. Kraman. *Lung Sounds: An Introduction to the Interpretation of Auscultatory Findings*. Association of American Medical Colleges, 2007.

[179] D. Kumar, P. d. Carvalho, M. Antunes, and J. Henriques. Noise detection during heart sound recording. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3119–3123. IEEE, 2009.

[180] R. T. H. Laennec. *De l'auscultation médiate: ou, Traité du diagnostic des maladies des poumons et du coeur; fondé principalement sur ce nouveau moyen d'exploration*, volume 2. Culture et civilisation, 1819.

[181] C. Laroche, H. Papadopoulos, M. Kowalski, and G. Richard. Genre specific dictionaries for harmonic/percussive source separation. 2016.

[182] S. Le Cam, A. Belghith, C. Collet, and F. Salzenstein. Wheezing sounds detection using multivariate generalized gaussian distributions. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 541–544. IEEE, 2009.

[183] C.-T. Lee, Y.-H. Yang, and H. H. Chen. Multipitch estimation of piano music by exemplar-based sparse representation. *IEEE Transactions on Multimedia*, 14(3):608–618, 2012.

[184] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[185] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[186] H. Lee, J. Yoo, and S. Choi. Semi-supervised nonnegative matrix factorization. *IEEE Signal Processing Letters*, 17(1):4–7, 2009.

[187] S. Lehrer. *Understanding Lung Sounds, the 2nd edition*. PHILADELPHIA W.B SAUNDERS COMPANY, 1993.

[188] S. Lehrer. *Understanding Lung Sounds, the 3rd edition*. PHILADELPHIA W.B SAUNDERS COMPANY, 2002.

[189] S. Leng, R. San Tan, K. T. C. Chai, C. Wang, D. Ghista, and L. Zhong. The electronic stethoscope. *Biomedical engineering online*, 14(1):1–37, 2015.

[190] M. G. Levitzky. *Pulmonary physiology*. Number 1. McGraw-Hill Education, 2018.

[191] S. Z. Li, X. W. Hou, H. J. Zhang, and Q. S. Cheng. Learning spatially localized, parts-based representation. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages 207–212. IEEE, 2001.

[192] B.-S. Lin, H.-D. Wu, and S.-J. Chen. Automatic wheezing detection based on signal processing of spectrogram and back-propagation neural network. *Journal of healthcare engineering*, 6, 2015.

[193] C. Lin and E. Hasting. Blind source separation of heart and lung sounds based on non-negative matrix factorization. In *2013 International Symposium on Intelligent Signal Processing and Communication Systems*, pages 731–736. IEEE, 2013.

[194] C. Liu and H. Wechsler. Independent component analysis of gabor features for face recognition. *IEEE transactions on Neural Networks*, 14(4):919–928, 2003.

[195] H. Liu, Z. Wu, X. Li, D. Cai, and T. S. Huang. Constrained nonnegative matrix factorization for image representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1299–1311, 2011.

[196] P. C. Loizou. *Speech enhancement: theory and practice*. CRC press, 2013.

[197] R. Loudon and R. L. Murphy Jr. Lung sounds. *American Review of Respiratory Disease*, 130(4):663–673, 1984.

[198] M. Lozano, J. A. Fiz, and R. Jané. Automatic differentiation of normal and continuous adventitious respiratory sounds using ensemble empirical mode decomposition and instantaneous frequency. *IEEE journal of biomedical and health informatics*, 20(2):486–497, 2015.

[199] M. Lozano-Garcia, J. A. Fiz, C. Martinez-Rivera, A. Torrents, J. Ruiz-Manzano, and R. Jane. Novel approach to continuous adventitious respiratory sound analysis for the assessment of bronchodilator response. *PloS one*, 12(2):e0171455, 2017.

[200] B.-Y. Lu. Unidirectional microphone based wireless recorder for the respiration sound. *Journal of Bioengineering and Biomedical Science*, 6(3):195, 2016.

[201] B.-Y. Lu, M.-L. Hsueh, and H.-D. Wu. Reducing the ambulance siren noise for distant auscultation of the lung sound. *Acoustics Australia*, 45(2):381–387, 2017.

[202] N. Lu, T. Li, J. Pan, X. Ren, Z. Feng, and H. Miao. Structure constrained semi-nonnegative matrix factorization for eeg-based motor imagery classification. *Computers in Biology and Medicine*, 60:32–39, 2015.

[203] J. Ma, Y. Hu, and P. C. Loizou. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *The Journal of the Acoustical Society of America*, 125(5):3387–3405, 2009.

[204] M. Mahagnah and N. Gavriely. Gas density does not affect pulmonary acoustic transmission in normal men. *Journal of Applied Physiology*, 78(3):928–937, 1995.

[205] S. Mangione. *Physical Diagnosis Secrets E-Book: With STUDENT CONSULT Online Access*. Elsevier Health Sciences, 2012.

[206] S. Mangione. *Secrets Heart & Lung Sounds Audio Workshop Access Code, 2nd Edition*. Hanley & Belfus, 2015.

[207] S. Mangione and L. Z. Nieman. Pulmonary auscultatory skills during training in internal medicine and family practice. *American journal of respiratory and critical care medicine*, 159(4):1119–1124, 1999.

[208] H. A. Mansy, R. A. Balk, W. H. Warren, T. J. Royston, Z. Dai, Y. Peng, and R. H. Sandler. Pneumothorax effects on pulmonary acoustic transmission. *Journal of applied Physiology*, 119(3):250–257, 2015.

[209] A. Marshall and S. Boussakta. Signal analysis of medical acoustic sounds with applications to chest medicine. *Journal of the Franklin Institute*, 344(3-4):230–242, 2007.

[210] R. Marxer and J. Janer. Study of regularizations and constraints in nmf-based drums monaural separation. In *International Conference on Digital Audio Effects Conference (DAFx-13)*, 2013.

[211] Y. Matsui, S. Makino, N. Ono, and T. Yamada. Multiple far noise suppression in a real environment using transfer-function-gain nmf. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 2314–2318. IEEE, 2017.

[212] S. Matsunaga, K. Yamauchi, M. Yamashita, and S. Miyahara. Classification between normal and abnormal respiratory sounds based on maximum likelihood approach. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 517–520. IEEE, 2009.

[213] U. Mayat, F. Qureshi, S. Ahmed, Y. Athavale, and S. Krishnan. Towards a low-cost point-of-care screening platform for electronic auscultation of vital body sounds. In *2017 IEEE Canada International Humanitarian Technology Conference (IHTC)*, pages 1–5. IEEE, 2017.

[214] P. Mayorga, C. Druzgalski, R. Morelos, O. Gonzalez, and J. Vidales. Acoustics based assessment of respiratory diseases using gmm classification. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pages 6312–6316. IEEE, 2010.

[215] I. Mazić, M. Bonković, and B. Džaja. Two-level coarse-to-fine classification algorithm for asthma wheezing recognition in children's respiratory sounds. *Biomedical Signal Processing and Control*, 21:105–118, 2015.

[216] S. McGee. Auscultation of the lungs. *Evidence-based physical diagnosis (3 rd ed.). Elsevier Saunders, Philadelphia*, 2012.

[217] S. McGee. *Evidence-based physical diagnosis e-book*. Elsevier Health Sciences, 2016.

[218] L. Mendes, I. Vogiatzis, E. Perantoni, E. Kaimakamis, I. Chouvarda, N. Maglaveras, V. Tsara, C. Teixeira, P. Carvalho, J. Henriques, et al. Detection of wheezes using their signature in the spectrogram space and musical features. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5581–5584. IEEE, 2015.

[219] N. Meslier, G. Charbonneau, and J. Racineux. Wheezes. *European respiratory journal*, 8(11):1942–1948, 1995.

[220] B. Mijović, M. De Vos, I. Gligorijević, J. Taelman, and S. Van Huffel. Source separation from single-channel recordings by combining empirical-mode decomposition and independent component analysis. *IEEE transactions on biomedical engineering*, 57(9):2188–2196, 2010.

[221] M. Milicevic, I. Mazic, and M. Bonkovic. Classification accuracy comparison of asthmatic wheezing sounds recorded under ideal and real-world conditions. In *15th International Conference on Artificial Intelligence, Knowledge Engineering and Databases (AIKED 2016), Venice*, 2016.

[222] R. Mor, I. Kushnir, J.-J. Meyer, J. Ekstein, and I. Ben-Dov. Breath sound distribution images of patients with pneumonia and pleural effusion. *Respiratory care*, 52(12):1753–1760, 2007.

[223] M. Munakata, H. Ukita, I. Doi, Y. Ohtsuka, Y. Masaki, Y. Homma, and Y. Kawakami. Spectral and waveform characteristics of fine and coarse crackles. *Thorax*, 46(9):651–657, 1991.

[224] R. L. Murphy. In defense of the stethoscope. *Respiratory Care*, 53(3):355–369, 2008.

[225] R. L. Murphy Jr, S. K. Holford, and W. C. Knowler. Visual lung-sound characterization by time-expanded wave-form analysis. *New England Journal of Medicine*, 296(17):968–971, 1977.

[226] F. G. Nabi, K. Sundaraj, and C. K. Lam. Identification of asthma severity levels through wheeze sound characterization and classification using integrated power features. *Biomedical Signal Processing and Control*, 52:302–311, 2019.

[227] Y. Nagasaka. Lung sounds in bronchial asthma. *Allergology International*, 61(3):353–363, 2012.

[228] H. Nakano, M. Hayashi, E. Ohshima, N. Nishikata, and T. Shinohara. Validation of a new system of tracheal sound analysis for the diagnosis of sleep apnea-hypopnea syndrome. *Sleep*, 27(5):951–957, 2004.

[229] R. Naves, B. H. Barbosa, and D. D. Ferreira. Classification of lung sounds using higher-order statistics: A divide-and-conquer approach. *Computer methods and programs in biomedicine*, 129:12–20, 2016.

[230] G. Nelson, R. Rajamani, and A. Erdman. Noise control challenges for auscultation on medical evacuation helicopters. *Applied Acoustics*, 80:68–78, 2014.

[231] G. Nieminen. Device including ultrasound, auscultation, and ambient noise sensors, Apr. 9 2020. US Patent App. 16/593,173.

[232] J. Nikunen and T. Virtanen. Object-based audio coding using non-negative matrix factorization for the spectrogram representation. In *Audio Engineering Society Convention 128*. Audio Engineering Society, 2010.

[233] D. Oletic and V. Bilas. Asthmatic wheeze detection from compressively sensed respiratory sound spectra. *IEEE journal of biomedical and health informatics*, 22(5):1406–1414, 2017.

[234] B. Orten. Auscultation apparatus, Jan. 28 2003. US Patent 6,512,830.

[235] P. Paatero. Least squares formulation of robust non-negative factor analysis. *Chemometrics and intelligent laboratory systems*, 37(1):23–35, 1997.

[236] P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

[237] R. Paciej, A. Vyshedskiy, D. Bana, and R. Murphy. Squawks in pneumonia. *Thorax*, 59(2):177–178, 2004.

[238] A. Painsky, S. Rosset, and M. Feder. Large alphabet source coding using independent component analysis. *IEEE Transactions on Information Theory*, 63(10):6514–6529, 2017.

[239] P. Palange and G. Rohde. *ERS handbook of respiratory medicine*. European Respiratory Society, 2019.

[240] R. Palaniappan, K. Sundaraj, and N. U. Ahamed. Machine learning in lung sound analysis: a systematic review. *Biocybernetics and Biomedical Engineering*, 33(3):129–135, 2013.

[241] R. Palaniappan, K. Sundaraj, and N. U. Ahamed. Machine learning in lung sound analysis: a systematic review. *Biocybernetics and Biomedical Engineering*, 33(3):129–135, 2013.

[242] Y. Panagakis, C. Kotropoulos, and G. R. Arce. Music genre classification via sparse representations of auditory temporal modulations. In *2009 17th European Signal Processing Conference*, pages 1–5. IEEE, 2009.

[243] J. Park, J. Shin, and K. Lee. Exploiting continuity/discontinuity of basis vectors in spectrogram decomposition for harmonic-percussive sound separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(5):1061–1074, 2017.

[244] A. Pascual-Montano, J. M. Carazo, K. Kochi, D. Lehmann, and R. D. Pascual-Marqui. Nonsmooth nonnegative matrix factorization (nsnmf). *IEEE transactions on pattern analysis and machine intelligence*, 28(3):403–415, 2006.

[245] H. Pasterkamp. The highs and lows of wheezing: A review of the most popular adventitious lung sound. *Pediatric Pulmonology*, 53(2):243–254, 2018.

[246] H. Pasterkamp, R. Fenton, A. Tal, and V. Chernick. Interference of cardiovascular sounds with phonopneumography in children. *American Review of Respiratory Disease*, 131(1):61–64, 1985.

[247] H. Pasterkamp, S. S. Kraman, and G. R. Wodicka. Respiratory sounds: advances beyond the stethoscope. *American journal of respiratory and critical care medicine*, 156(3):974–987, 1997.

[248] S. B. Patel, T. F. Callahan, M. G. Callahan, J. T. Jones, G. P. Graber, K. S. Foster, K. Glifort, and G. R. Wodicka. An adaptive noise reduction stethoscope for auscultation in high noise environments. *The Journal of the Acoustical Society of America*, 103(5):2483–2491, 1998.

[249] P. Piirila and A. Sovijarvi. Crackles: recording, analysis and clinical significance. *European Respiratory Journal*, 8(12):2139–2148, 1995.

[250] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies. Sparse representations in audio and music: from coding to source separation. *Proceedings of the IEEE*, 98(6):995–1005, 2009.

[251] R. X. A. Pramono, S. Bowyer, and E. Rodriguez-Villegas. Automatic adventitious respiratory sound analysis: A systematic review. *PloS one*, 12(5):e0177926, 2017.

[252] R. X. A. Pramono, S. A. Imtiaz, and E. Rodriguez-Villegas. Evaluation of features for classification of wheezes and normal respiratory sounds. *PloS one*, 14(3):e0213659, 2019.

[253] R. X. A. Pramono, S. A. Imtiaz, and E. Rodriguez-Villegas. Evaluation of mel-frequency cepstrum for wheeze analysis. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4686–4689. IEEE, 2019.

[254] K. N. Priftis, L. J. Hadjileontiadis, and M. L. Everard. Breath sounds. *From Basic Science to Clinical Practice. 1st Ed., Cham: Springer*, 2018.

[255] J. Proctor and E. Rickards. How to perform chest auscultation and interpret the findings. *Nursing Times*, 116(1):23–26, 2020.

[256] Y. Qiu, A. Whittaker, M. Lucas, and K. Anderson. Automatic wheeze detection based on auditory modelling. *Proceedings of the Institution of Mechanical Engineers, Part H:*

Journal of Engineering in Medicine, 219(3):219–227, 2005.

[257] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements. *Objective measures of speech quality*. Prentice Hall, 1988.

[258] J. K. Quint, E. R. Millett, M. Joshi, V. Navaratnam, S. L. Thomas, J. R. Hurst, L. Smeeth, and J. S. Brown. Changes in the incidence, prevalence and mortality of bronchiectasis in the uk from 2004 to 2013: a population-based cohort study. *European Respiratory Journal*, 47(1):186–193, 2016.

[259] S. Raczyński, N. Ono, and S. Sagayama. Extending nonnegative matrix factorization—a discussion in the context of multiple frequency estimation of musical signals. In *2009 17th European Signal Processing Conference*, pages 934–938. IEEE, 2009.

[260] S. A. Raczyński, N. Ono, and S. Sagayama. Multipitch analysis with harmonic nonnegative matrix approximation. In *in ISMIR 2007, 8th International Conference on Music Information Retrieval*. Citeseer, 2007.

[261] B. Raj, R. Singh, and T. Virtanen. Phoneme-dependent nmf for speech enhancement in monaural mixtures. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[262] A. Rao, S. Chu, N. Batlivala, S. Zetumer, and S. Roy. Improved detection of lung fluid with standardized acoustic stimulation of the chest. *IEEE journal of translational engineering in health and medicine*, 6:1–7, 2018.

[263] A. Rao, E. Huynh, T. J. Royston, A. Kornblith, and S. Roy. Acoustic methods for pulmonary diagnosis. *IEEE reviews in biomedical engineering*, 12:221–239, 2018.

[264] A. Rao, J. Ruiz, C. Bao, and S. Roy. Tabla: A proof-of-concept auscultatory percussion device for low-cost pneumonia detection. *Sensors*, 18(8):2689, 2018.

[265] S. Reichert, R. Gass, C. Brandt, and E. Andrès. Analysis of respiratory sounds: state of the art. *Clinical medicine. Circulatory, respiratory and pulmonary medicine*, 2:CCRPM–S530, 2008.

[266] D. A. Rice. Transmission of lung sounds. In *Seminars in respiratory medicine*, volume 6, pages 166–170. Copyright© 1985 by Thieme Medical Publishers, Inc., 1985.

[267] R. Riella, P. Nohama, and J. Maia. Method for automatic detection of wheezing in lung sounds. *Brazilian Journal of Medical and Biological Research*, 42(7):674–684, 2009.

[268] S. Rietveld, M. Oud, and E. H. Dooijes. Classification of asthmatic breath sounds: preliminary results of the classifying capacity of human examiners versus artificial neural networks. *Computers and Biomedical Research*, 32(5):440–448, 1999.

[269] B. M. Rocha, D. Filos, L. Mendes, G. Serbes, S. Ulukaya, Y. P. Kahya, N. Jakovljevic, T. L. Turukalo, I. M. Vogiatzis, E. Perantoni, et al. An open access database for the evaluation of respiratory sound classification algorithms. *Physiological Measurement*, 40(3):035001, 2019.

[270] I. Rudan, L. Tomaskovic, C. Boschi-Pinto, and H. Campbell. Global estimate of the incidence of clinical pneumonia among children under five years of age. *Bulletin of the World Health Organization*, 82:895–903, 2004.

[271] R. Rui and C. Bao. Projective non-negative matrix factorization with bregman divergence for musical instrument classification. In *2012 IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC 2012)*, pages 415–418, 2012.

[272] T. N. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, D. Van Compernolle, K. Demuynck, J. F. Gemmeke, J. R. Bellegarda, and S. Sundaram. Exemplar-based processing for speech recognition: An overview. *IEEE Signal Processing Magazine*, 29(6):98–113, 2012.

[273] A. J. Salazar, C. Alvarado, and F. E. Lozano. System of heart and lung sounds separation for store-and-forward telemedicine applications. *Revista Facultad de Ingeniería Universidad de Antioquia*, (64):175–181, 2012.

[274] M. Sarkar, I. Madabhavi, N. Niranjan, and M. Dogra. Auscultation of the respiratory system. *Annals of thoracic medicine*, 10(3):158, 2015.

[275] S. J. Scalise, A. S. Rainone, and D. W. Davis. Auscultation augmentation device, Sept. 22 1998. US Patent 5,812,678.

[276] B. Schuller, A. Lehmann, F. Weninger, F. Eyben, and G. Rigoll. Blind enhancement of the rhythmic and harmonic sections by nmf: Does it help? In *Proc. Intern. Conf. on Acoustics (NAG/Tagungsband Fortschritte der Akustik-DAGA 2009), Rotterdam, The Netherlands*, pages 361–364, 2009.

[277] R. M. Schwartzstein and M. J. Parker. *Respiratory Physiology: A Clinical Approach*. Lippincott Williams & Wilkins, 2006.

[278] P. S. J. Sebastián, T. Virtanen, V. M. Garcia-Molla, and A. M. Vidal. Analysis of an efficient parallel implementation of active-set newton algorithm. *The Journal of Supercomputing*, 75(3):1298–1309, 2019.

[279] I. Sen, M. Saraclar, and Y. P. Kahya. A comparison of svm and gmm-based classifier configurations for diagnostic classification of pulmonary sounds. *IEEE Transactions on Biomedical Engineering*, 62(7):1768–1776, 2015.

[280] N. Sengupta, M. Sahidullah, and G. Saha. Lung sound classification using cepstral-based statistical features. *Computers in biology and medicine*, 75:118–129, 2016.

[281] S. M. Shaharum, K. Sundaraj, S. Aniza, R. Palaniappan, and K. Helmy. Classification of asthma severity levels by wheeze sound analysis. In *2016 IEEE Conference on Systems, Process and Control (ICSPC)*, pages 172–176. IEEE, 2016.

[282] P. Smaragdis. Convolutive speech bases and their application to supervised speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):1–12, 2006.

[283] P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No. 03TH8684)*, pages 177–180. IEEE, 2003.

[284] P. Smaragdis, B. Raj, and M. Shashanka. Supervised and semi-supervised separation of sounds from single-channel mixtures. In *International Conference on Independent Component Analysis and Signal Separation*, pages 414–421. Springer, 2007.

[285] P. Smaragdis, M. Shashanka, and B. Raj. A sparse non-parametric approach for single channel separation of known sounds. In *Advances in neural information processing systems*, pages 1705–1713, 2009.

[286] A. T. Society et al. Updated nomenclature for membership reaction. *ATS NEWS*, 3:5–6, 1977.

[287] A. Sovijarvi. Characteristics of breath sounds and adventitious respiratory sounds. *Eur Respir Rev*, 10:591–596, 2000.

[288] A. Sovijarvi, F. Dalmasso, J. Vanderschoot, L. Malmberg, G. Righini, and S. Stoneman. Definition of terms for applications of respiratory sounds. *European Respiratory Review*, 10(77):597–610, 2000.

[289] A. Sovijärvi, J. Vanderschoot, and J. Earis. *Computerized Respiratory Sound Analysis (CORSA): Recommended Standards for Terms and Techniques: ERS Task Force Report*. Munksgaard, 2000.

[290] A. Sovijarvi, J. Vanderschoot, and J. Earis. Standardization of computerized respiratory sound analysis. *European Respiratory Review*, 10(77):585–585, 2000.

[291] P. Sprechmann, A. Bronstein, M. Bronstein, and G. Sapiro. Learnable low rank sparse models for speech denoising. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 136–140. IEEE, 2013.

[292] B. Suki, A. M. Alencar, Z. Hantos, and H. E. Stanley. Generation and propagation of crackle sound and its relation to lung structure. *ASME-PUBLICATIONS-BED*, 50:639–640, 2001.

[293] A. Suzuki, C. Sumi, K. Nakayama, and M. Mori. Real-time adaptive cancelling of ambient noise in lung sound measurement. *Medical and Biological Engineering and Computing*, 33(5):704–708, 1995.

[294] S. Taplidou and L. Hadjileontiadis. Analysis of wheezes using wavelet higher order spectral features. *IEEE Transactions On Biomedical Engineering*, 57(7):1596–1610, 2010.

[295] S. A. Taplidou and L. J. Hadjileontiadis. Wheeze detection based on time-frequency analysis of breath sounds. *Computers in biology and medicine*, 37(8):1073–1083, 2007.

[296] S. A. Taplidou, L. J. Hadjileontiadis, I. K. Kitsas, K. I. Panoulas, T. Penzel, V. Gross, and S. M. Panas. On applying continuous wavelet transform in wheeze analysis. In *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology*

Society, volume 2, pages 3832–3835. IEEE, 2004.

[297] S. Theodoridis, K. Koutroumbas, et al. Pattern recognition. *IEEE Transactions on Neural Networks*, 19(2):376, 2008.

[298] W. M. Thurlbeck and N. Müller. Emphysema: definition, imaging, and quantification. *AJR. American journal of roentgenology*, 163(5):1017–1025, 1994.

[299] A. G. Tilkian and M. B. Conover. *Understanding heart sounds and murmurs with an introduction to lung sounds*. W.B SAUNDERS COMPANY, 2001.

[300] A. Torres-Jimenez, S. Charleston-Villalobos, R. Gonzalez-Camarena, G. Chi-Lem, and T. Aljama-Corrales. Asymmetry in lung sound intensities detected by respiratory acoustic thoracic imaging (rathi) and clinical pulmonary auscultation. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4797–4800. IEEE, 2008.

[301] H. Tsubakida, T. Shiratori, A. Ishiyama, and Y. Ono. Nonnegative matrix factorization common spatial pattern in brain machine interface. In *The 3rd International Winter Conference on Brain-Computer Interface*, pages 1–4. IEEE, 2015.

[302] C. Tzagkarakis and A. Mouchtaris. Sparsity based robust speaker identification using a discriminative dictionary learning approach. In *21st European Signal Processing Conference (EUSIPCO 2013)*, pages 1–5. IEEE, 2013.

[303] S. Ulukaya, I. Sen, and Y. P. Kahya. Feature extraction using time-frequency analysis for monophonic-polyphonic wheeze discrimination. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5412–5415. IEEE, 2015.

[304] S. Ulukaya, I. Sen, and Y. P. Kahya. A novel method for determination of wheeze type. In *23nd Signal Processing and Communications Applications Conference (SIU)*, pages 2001–2004, 2015.

[305] S. Ulukaya, G. Serbes, and Y. P. Kahya. Wheeze type classification using non-dyadic wavelet transform based optimal energy ratio technique. *Computers in biology and medicine*, 104:175–182, 2019.

[306] L. Vannuccini, J. Earis, P. Helisto, B. Cheetham, M. Rossi, A. Sovijarvi, and J. Vanderschoot. Capturing and preprocessing of respiratory sounds. *European Respiratory Review*, 10(77):616–620, 2000.

[307] L. Vannuccini, M. Rossi, and G. Pasquali. A new method to detect crackles in respiratory sounds. *Technology and Health Care*, 6(1):75–79, 1998.

[308] M. Vendrell, J. de Gracia, C. Olveira, M. Á. Martínez, R. Girón, L. Máiz, R. Cantón, R. Coll, A. Escribano, and A. Solé. Diagnosis and treatment of bronchiectasis. *Archivos de Bronconeumología (English Edition)*, 44(11):629–640, 2008.

[309] E. Vincent, N. Bertin, and R. Badeau. Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):528–537, 2009.

[310] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing*, 14(4):1462–1469, 2006.

[311] E. Vincent and X. Rodet. Underdetermined source separation with structured source priors. In *International Conference on Independent Component Analysis and Signal Separation*, pages 327–334. Springer, 2004.

[312] T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE transactions on audio, speech, and language processing*, 15(3):1066–1074, 2007.

[313] T. Virtanen, J. F. Gemmeke, and B. Raj. Active-set newton algorithm for overcomplete non-negative representations of audio. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(11):2277–2289, 2013.

[314] T. Virtanen and A. Klapuri. Analysis of polyphonic audio using source-filter model and non-negative matrix factorization. In *Advances in models for acoustic processing, neural information processing systems workshop*, volume 18. Citeseer, 2006.

[315] A. Vyshedskiy, R. M. Alhashem, R. Paciej, M. Ebril, I. Rudman, J. J. Fredberg, and R. Murphy. Mechanism of inspiratory and expiratory crackles. *Chest*, 135(1):156–164, 2009.

[316] A. Vyshedskiy and R. Murphy. Acoustic biomarkers of chronic obstructive lung disease. *Research Ideas and Outcomes*, 2:e9173, 2016.

[317] L. R. Waitman, K. P. Clarkson, J. A. Barwise, and P. H. King. Representation and classification of breath sounds recorded in an intensive care setting using neural networks. *Journal of clinical monitoring and computing*, 16(2):95–105, 2000.

[318] H. Wang, H. Zheng, G. Yin, et al. Multi-sensor lung sound extraction via time-shared channel identification and adaptive noise cancellation. In *2004 43rd IEEE Conference on Decision and Control (CDC)(IEEE Cat. No. 04CH37601)*, volume 4, pages 3599–3604. IEEE, 2004.

[319] S. Wang, C. Deng, W. Lin, G.-B. Huang, and B. Zhao. Nmf-based image quality assessment using extreme learning machine. *IEEE transactions on cybernetics*, 47(1):232–243, 2016.

[320] Z. Wang, S. Jean, and T. Bartter. Lung sound analysis in the diagnosis of obstructive airway disease. *Respiration*, 77(2):134–138, 2009.

[321] Z. Wang and F. Sha. Discriminative non-negative matrix factorization for single-channel speech separation. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3749–3753. IEEE, 2014.

[322] S. E. Weinberger, B. A. Cockrill, and J. Mandel. *Principles of Pulmonary Medicine E-Book*. Elsevier Health Sciences, 2017.

[323] F. Weninger and B. Schuller. Optimization and parallelization of monaural source separation algorithms in the openblissart toolkit. *Journal of Signal Processing Systems*, 69(3):267–277, 2012.

[324] B. K. Wiederhold, P. Cipresso, D. Pizzioli, M. Wiederhold, and G. Riva. Intervention for physician burnout: A systematic review. *Open Medicine*, 13(1):253–263, 2018.

[325] R. L. Wilkins, J. E. Hodgkin, and B. Lopez. *Lung Sounds: A Practical Guide*. Mosby, 1996.

[326] R. L. Wilkins, J. E. Hodgkin, and B. Lopez. *Fundamentals of lung and heart sounds*. Mosby, 2004.

[327] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran. Speech denoising using nonnegative matrix factorization with priors. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4029–4032. IEEE, 2008.

[328] M. Wisniewski and T. P. Zielinski. Application of tonal index to pulmonary wheezes detection in asthma monitoring. In *2011 19th European Signal Processing Conference*, pages 1544–1548. IEEE, 2011.

[329] M. Wisniewski and T. P. Zielinski. Tonality detection methods for wheezes recognition system. In *2012 19th International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 472–475. IEEE, 2012.

[330] M. Wiśniewski and T. P. Zieliński. Joint application of audio spectral envelope and tonality index in an e-asthma monitoring system. *IEEE journal of biomedical and health informatics*, 19(3):1009–1018, 2014.

[331] G. Wodicka, A. Aguirre, P. DeFrain, and D. Shannon. Phase delay of pulmonary acoustic transmission from trachea to chest wall. *IEEE transactions on biomedical engineering*, 39(10):1053–1059, 1992.

[332] G. R. Wodicka, K. N. Stevens, H. L. Golub, and D. C. Shannon. Spectral characteristics of sound transmission in the human respiratory system. *IEEE transactions on biomedical engineering*, 37(12):1130–1135, 1990.

[333] D. Wrigley. *Heart and Lung Sounds Reference Library*. PESI Healthcare, 2011.

[334] J. Xie, W. Chen, D. Zhang, S. Zu, and Y. Chen. Application of principal component analysis in weighted stacking of seismic data. *IEEE Geoscience and Remote Sensing Letters*, 14(8):1213–1217, 2017.

[335] A. Yadollahi and Z. M. Moussavi. A robust method for heart sounds localization using lung sounds entropy. *IEEE transactions on biomedical engineering*, 53(3):497–502, 2006.

[336] M. Yeginer and Y. Kahya. Modeling of pulmonary crackles using wavelet networks. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 7560–7563. IEEE, 2006.

[337] M. Yonemaru, K. Kikuchi, M. Mori, A. Kawai, T. Abe, T. Kawashiro, T. Ishihara, and T. Yokoyama. Detection of tracheal stenosis by frequency analysis of tracheal sounds. *Journal of applied physiology*, 75(2):605–612, 1993.

[338] J. Yoo, M. Kim, K. Kang, and S. Choi. Nonnegative matrix partial co-factorization for drum source separation. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1942–1945. IEEE, 2010.

[339] G. L. Zacharias, A. X. Miao, J. A. Moore, R. D. Collier, and M. Asdigha. Active noise cancellation stethoscope. Technical report, CHARLES RIVER ANALYTICS INC CAMBRIDGE MA, 1993.

[340] S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas. Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification. *IEEE Transactions on Neural Networks*, 17(3):683–695, 2006.

[341] R. Zdunek and A. Cichocki. Nonnegative matrix factorization with constrained second-order optimization. *Signal Processing*, 87(8):1904–1916, 2007.

[342] G. Zenk. Stethoscopic detection of lung sounds in high noise environments. *Master's thesis, Purdue University*, 1994.

[343] J. Zhang, W. Ser, J. Yu, and T. Zhang. A novel wheeze detection method for wearable monitoring systems. In *2009 International Symposium on Intelligent Ubiquitous Computing and Education*, pages 331–334. IEEE, 2009.

[344] D. S. Zhdanov, A. S. Bureev, L. A. Khokhlova, A. I. Seleznev, and I. Y. Zemlyakov. Short review of devices for detection of human breath sounds and heart tones. *Biology and Medicine*, 6(3):1, 2014.

[345] B. Zhu, W. Li, R. Li, and X. Xue. Multi-stage non-negative matrix factorization for monaural singing voice separation. *IEEE Transactions on audio, speech, and language processing*, 21(10):2096–2107, 2013.

[346] L. S. Zun and L. Downey. The effect of noise in the emergency department. *Academic emergency medicine*, 12(7):663–666, 2005.

Paper 1

Wheezing Sound Separation Based on Constrained Non-negative Matrix Factorization

J. Torre-Cruz, F. Canadas-Quesada, P. Vera-Candeas, V. Montiel-Zafra and N. Ruiz-Reyes, "Wheezing Sound Separation Based on Constrained Non-negative Matrix Factorization", in *Proceedings of the 10th International Conference on Bioinformatics and Biomedical Technology (ICBBT)*, Amsterdam, The Netherlands, pp. 18–24, May 2018. DOI: `https://doi.org/10.1145/3232059.3232072`

- Congreso: The 10th International Conference on Bioinformatics and Biomedical Technology (ICBBT).
- Fecha y lugar de celebración: del 16 al 18 de Mayo de 2018, Amsterdam (The Netherlands).
- Páginas: 18-24.
- ISBN: 9781450363662.
- Premio a la mejor ponencia de la sesión 2 del congreso.

Wheezing Sound Separation Based on Constrained Non-Negative Matrix Factorization

J. Torre-Cruz
Telecommunication Engineering Department, University of Jaen Scientific-Technological Campus of Linares, Avda. de la Universidad s/n, 23700, Linares (Jaen), Spain

F. Canadas-Quesada
Telecommunication Engineering Department, University of Jaen Scientific-Technological Campus of Linares, Avda. de la Universidad s/n, 23700, Linares (Jaen), Spain
Phone: +34 953 648510
E-mail: fcanadas@ujaen.es

P. Vera-Candeas
Telecommunication Engineering Department, University of Jaen Scientific-Technological Campus of Linares, Avda. de la Universidad s/n, 23700, Linares (Jaen), Spain

V. Montiel-Zafra
Telecommunication Engineering Department University of Jaen, Scientific-Technological Campus of Linares, Avda. de la Universidad s/n, 23700 Linares (Jaen), Spain

N. Ruiz-Reyes
Telecommunication Engineering Department University of Jaen, Scientific-Technological Campus of Linares, Avda. de la Universidad s/n, 23700 Linares (Jaen), Spain

ABSTRACT
Auscultation remains the first clinical examination that a physician performs to detect respiratory diseases originated by wheezes, which are the most specific asthmatic symptoms. It is common that respiratory sounds (normal breath sounds) acoustically interfere wheezes with both frequency and time domain. As a result, the physician's cognitive ability is reduced causing a misdiagnosis or inability to clearly hear all significant sounds to detect a pulmonary disease. This paper presents a constrained non-negative matrix factorization (NMF) approach to separate wheezes from respiratory sounds applied to single-channel mixtures. The proposed constraints, smoothness and sparseness, attempts to model common spectral behaviour shown by wheezes and normal breath sounds. Specifically, the spectrogram of a wheeze can be modelled as a narrowband spectrum (sparseness in frequency). However, the spectrogram of a normal breath sound can be modelled as a wideband spectrum (smoothness in frequency) with a slow temporal variation (smoothness in time). Experimental results report that the proposed method improves the audio quality of the wheezes removing most of the respiratory sounds, being a novel way to successfully apply a NMF approach to a wheeze/respiratory sound separation.

CCS Concepts
•Applied computing → Health care information systems.

Keywords
Non-negative matrix factorization (NMF); Constraint; Wheeze; Respiratory; Smoothness; Sparseness.

ICBBT '18, May 16–18, 2018, Amsterdam, Netherlands

ACM ISBN 978-1-4503-6366-2/18/05…$15.00
DOI: https://doi.org/10.1145/3232059.3232072

1. INTRODUCTION
Two hundred and thirty-five million people suffer from asthma, being asthma the most common chronic disease among children according to the World Health Organization (WHO) [1].

Therefore, it is crucial that physicians are acoustically trained in the analysis of wheezing, which can be seen as spectral trajectories superimposed on respiratory sounds (normal breath sounds), because wheezes are the most specific asthmatic symptom [2] and provide useful information from different pulmonary diseases [3].

Auscultation can be defined as the process of listening sounds emitted in the human breathing process using instruments such as stethoscopes. In general terms, most of the first clinical examinations are carried out by means of auscultation since it is an inexpensive, non-invasive, safe, easy-to-perform and fast method to early detect any respiratory problem [4]. However, it is well known that a high percentage of diagnoses are highly dependent on the physician's experience and acoustic training.

Because of wheezes and respiratory sounds are simultaneously mixed both in frequency and time domain, one of the current challenging topics in biomedical engineering and bio-signal processing attempts to separate them in order to enhance the audio quality of wheezes that are listened by the physician. The reason is because some wheezes sounds that may be diagnostically significant may be masked by respiratory sounds or another background sounds. The temporal interference is caused because the respiratory sounds cannot cease when the wheezes are being listened since wheezes are usually associated with airway obstructions, so respiratory sounds and wheezes are generated by the same airflow through the bronchial tree. The spectral interference is due to the overlapping in frequency between the spectral bands in which the respiratory sounds and wheezes sounds are active.

The respiratory sounds, generated during the inspiration and expiration stages of the human breathing process, show a wideband spectrum where most of the energy is concentrated in the frequency band 60Hz-1000Hz [5]. Wheezes can be defined as pitched sound generated on inspiration, expiration or both stages of the human breathing process. Specifically, wheezes show sinusoidal oscillations whose pitch frequency is located in the spectral range of [100Hz-1000Hz] with duration longer than

100ms according to Computerized Respiratory Sound Analysis (CORSA) [3] [6] as can be seen in Figure 1. Basically, two types of wheezes are labelled: monophonic and polyphonic. A monophonic wheeze has a single pitch frequency but a polyphonic wheeze has more than one frequencies located at integer multiple of the pitch frequency. Besides, most of wheezes can be considered louder that the underlying respiratory sounds [6].

Several methods have been proposed to detect and classify wheezes using different approaches: spectral signature [7], [8], auditory modelling [9], entropy [10], SVM classifiers [11], wavelet transform [12], mel-frequency cepstral coefficients (MFCC) [13] and tonal index [14].

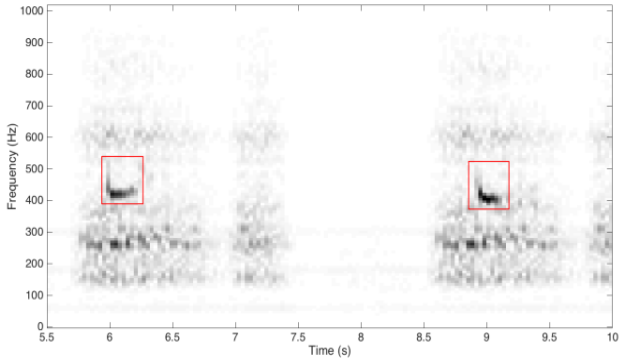

Figure 1. Magnitude spectrogram of a mixture composed of two wheezes (red squares) and respiratory sounds. A darker grey colour represents higher energy of each frequency.

However, there is a reduced number of works [15], [16] specifically focused on the sound separation of adventitious sounds. Non-negative matrix factorization (NMF) is a recent approach that has attracted the attention of the scientific community in the last years because it has been successfully applied to audio [17], [18] and image [19], [20] but NMF has not been applied to wheeze/respiratory sound separation to our knowledge.

This paper presents a constrained non-negative matrix factorization (NMF) approach applied to separate wheezes from respiratory sounds in single-channel mixtures. The mixture, composed of wheezes and respiratory sounds, is decomposed using a cost function to integrate spectral features, spectral smoothness and sparseness, into the NMF decomposition process. These features model the spectral behaviour of wheezes applying sparseness in frequency (spectral peaks) and the respiratory sounds are modelled applying smoothness in time (for magnitudes that vary slowly in time) and smoothness in frequency (the energy slowly is reduced in frequency). As a result, the proposed method is an unsupervised (blind) method because does not require any training about the sounds active in the mixture to analyse.

The remainder of this paper is organized as follows. In Section 2, a brief description about NMF is presented. In Section 3, the proposed method is detailed. Experimental results are shown in Section 4 and finally, conclusions and future work are reported in Section 5.

2. NON-NEGATIVE MATRIX FACTORIZATION

Non-negative matrix factorization (NMF) [21] or unconstrained NMF is a technique for linear representation of two-dimensional non-negative data that provides a parts-based representation of objects by imposing nonnegative constraints. NMF decomposes the spectrogram $X_{F,T}$ into the product of two non-negative matrices B and G using a linear combination of K elementary spectral patterns (basis functions or components) with time-varying gains,

$$X_{F,T} \approx \hat{X}_{F,T} = B_{F,K} G_{K,T} \quad (1)$$

where f=1,...,F and t=1,...,T indicate the frequency bin and the time frame, $\hat{X}_{F,T}$ is the reconstructed matrix, $B_{F,K}$ is the basis matrix or dictionary and $G_{K,T}$ is the time-varying gain matrix. In order to reduce the dimensionality of the data, the number of components K is selected to fulfill that $FK+KT<<FT$. Thus, the factorization is calculated by minimizing a cost function $D(X|\hat{X})$ as follows,

$$D(X|\hat{X}) = \sum_{f=1}^{F} \sum_{t=1}^{T} d(X_{f,f}|\hat{X}_{f,t}) \quad (2)$$

Applying an iterative algorithm based on multiplicative update rules, the cost function $D(X|\hat{X})$ is minimized and the non-negativity of the dictionary and the activations is ensured. In more detail, the multiplicative update rule for an arbitrary parameter Z is computed as follows,

$$Z = Z \odot \frac{\left[\frac{\partial D(X|\hat{X})}{\partial Z}\right]^{-}}{\left[\frac{\partial D(X|\hat{X})}{\partial Z}\right]^{+}} \quad (3)$$

, where the operator $\odot$ represents the element-wise multiplication and the operator division is the element-wise division.

3. PROPOSED METHOD

Unconstrained NMF can only ensure convergence to local minima that achieves the reconstruction. However, this reconstruction does not guarantee that the factorization is composed of parts-based objects with physical meaning as can be found in real life. To overcome this problem, it is useful to incorporate some prior information into the NMF decomposition by means of constraints. As a result, constrained NMF can find better solutions from those provided by unconstrained NMF, adding physical interpretation to the bases or gains. In this work, we add two constraints into the NMF factorization in order to model the common behaviour shown by wheezes or respiratory sounds found in spectrograms. The block diagram of the proposed method is shown in Figure 2

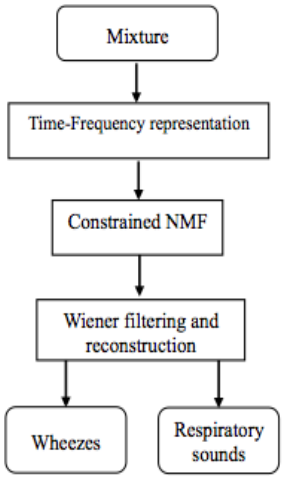

Figure 2. Flowchart of the proposed method.

3.1 Signal Factorization

Suppose the mixture $x(t)$, composed of wheezes $x_w(t)$ and respiratory $x_r(t)$ sounds, is additive, that is, $x(t) = x_w(t) + x_r(t)$. As a result, the magnitude spectrogram of the mixture $X = X_W + X_R$ (X_R represents the magnitude spectrogram of respiratory sounds and X_W represents the magnitude spectrogram of wheeze sounds). Each magnitude spectrogram, composed of T frames and F frequency bins, has been computed from the magnitude of the Short-Time Fourier Transform (STFT) using a Hamming window of size N with 50% overlap.

A normalization process is performed to achieve a NMF independent with respect to the norm of the signal X. Specifically, the normalized magnitude spectrogram X_{n_β} is computed,

$$X_{n_\beta} = \frac{X}{\left(\frac{\sum_{f=1}^{F}\sum_{t=1}^{T} x_{f,t}^{\beta}}{FT}\right)^{\frac{1}{\beta}}} \quad (4)$$

The proposed method attempts to separate the wheeze sounds discriminating between wheeze and respiratory bases in the NMF factorization. For this reason, we have used an objective function to factorize X_{n_β} into two spectrograms: $\hat{X}_R$ (estimated respiratory spectrogram) and $\hat{X}_W$ (estimated wheeze spectrogram). Our factorization model can be defined as follows,

$$X_{n_\beta} \approx \hat{X}_R + \hat{X}_W = B_{R_{F,P}} G_{R_{P,T}} + B_{W_{F,Q}} G_{W_{Q,T}} \quad (5)$$

, where B_R, B_W, G_R and G_W are the basis and gain matrices of the respiratory and the wheeze sounds. All of these matrices are non-negative matrices. The number of respiratory and wheeze components will be denoted as P and Q. The L^2-norm of each column of B_R or B_W is equal to 1.0. The L^2-norm of each row of G_R or G_W is equal to 1.0.

These parameters B_R, B_W, G_R and G_W are normally estimated by minimizing the reconstruction error between the input spectrogram X_{n_β} and the estimated spectrogram $(\hat{X}_R + \hat{X}_W)$. One of the most popular cost functions used in source separation is the β-divergence cost [29] shown in eq. (6),

$$d_\beta\left(X_{n_\beta}|(\hat{X}_R + \hat{X}_W)\right) = \sum_{f=1}^{F}\sum_{t=1}^{T} \frac{1}{\beta(\beta-1)}\Big(X_{n_\beta}^{\beta}{}_{f,t} + (\beta-1)(\hat{X}_R + \hat{X}_W)_{f,t}^{\beta} - \beta X_{n_\beta}{}_{f,t}(\hat{X}_R + \hat{X}_W)_{f,t}^{\beta-1}\Big) \quad (6)$$

In this work, a value of $\beta = 1.3$ has been used because a preliminary evaluation shows the best results. The β-divergence cost can ensure convergence to local minima that only achieves the signal reconstruction. However, this factorization cannot perform a correct discrimination between respiratory and wheezing bases. For this reason, we propose to include the sparseness and smoothness constrains into the NMF decomposition to model the spectral and temporal behaviour of respiratory sounds and wheezing.

3.2 Sparseness and Smoothness to Characterize Respiratory Sound and Wheezing

The main contribution of this paper is the development of a respiratory/wheezing separation based on unsupervised NMF using constraints that attempt to model common features observed in respiratory and wheezing sounds. These constraints are known as sparseness and smoothness, which has been successfully applied in music processing [30], [31] but not to the respiratory/wheezing scenario to our knowledge.

We propose the use of smoothness to model respiratory sounds as follows: respiratory sounds can be considered smooth in time (slow variation of the magnitude spectrogram along time) and frequency (wideband spectrum), as can be seen in Figure 1. As occurs in [30], we define the smoothness constraint ϕ and it is applied to the matrices B_R and G_R, denoted as ϕ_B and ϕ_G.

$$\phi_B = \frac{T}{P}\sum_{p=1}^{P}\frac{1}{\sigma_{B_R}{}^2}\sum_{f=2}^{F}\left(B_{R_{f-1,p}} - B_{R_{f,p}}\right)^2 \quad (7)$$

$$\phi_G = \frac{F}{P}\sum_{p=1}^{P}\frac{1}{\sigma_{G_R}{}^2}\sum_{t=2}^{T}\left(G_{R_{p,t-1}} - G_{R_{p,t}}\right)^2 \quad (8)$$

Normalization is used to make the global objective function independent of the signal norm. The bases B_R are normalized by $\sigma_{B_R} = \sqrt{\frac{1}{F}\sum_{f=1}^{F} B_{R_{f,p}}{}^2}$ and the activations G_R are normalized by $\sigma_{G_R} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} G_{R_{p,t}}{}^2}$.

We propose the use of sparseness to model wheezing as follows: wheezing can be considered sparse in frequency because a wheeze is characterized by one (monophonic) or more than one (polyphonic) spectral peaks as can be seen in Figure 1. As occurs in [30], we define sparseness constraint ψ, which is associated with the matrix B_W.

$$\psi_B = \frac{T}{Q}\sum_{q=1}^{Q}\sum_{f=1}^{F}\left|\frac{B_{W_{f,q}}}{\sigma_{B_W}}\right| \quad (9)$$

Normalization is used to make global objective function independent to the signal norm. The bases B_W are normalized by $\sigma_{B_W} = \sqrt{\frac{1}{F}\sum_{f=1}^{F} B_{W_{f,q}}{}^2}$.

Thus, we assure that each respiratory or wheezing cost has been normalized in order to obtain the same weight in the global objective function.

3.3 A Global Objective Function Based on the Proposed NMF Algorithm

The global objective function D including the β-divergence, smoothness and sparseness cost, is formulated as follows:

$$D = d_\beta\left(X_{n_\beta}|(\hat{X}_R + \hat{X}_W)\right) + \lambda_B\phi_B + \lambda_G\phi_G + \alpha_B\psi_B \quad (10)$$

, where λ_B defines the weight of the spectral smoothness applied to B_R, λ_G represents the weight of the temporal smoothness applied to G_R and α_B is the weight of the spectral sparseness applied to B_W. After an optimization process, results indicated no significant differences using $\lambda_B \neq \lambda_G$. As a result, it can be defined as unique general weight $\lambda = \lambda_B = \lambda_G$ (see eq. (11)). The best results have been obtained with $\lambda = 0.1$ and $\alpha_B = 0.2$.

$$D = d_\beta\left(X_{n_\beta}|(\hat{X}_R + \hat{X}_W)\right) + \lambda(\phi_B + \phi_G) + \alpha_B\psi_B \quad (11)$$

Applying a gradient descent algorithm [21] based on multiplicative update rules (see eq. (3)), the respiratory basis matrix B_R and the respiratory gain matrix G_R are formulated in eq. (12-13),

$$B_R = B_R \odot \frac{\left[\frac{\partial d_\beta}{\partial B_R}\right]^- + \lambda\left[\frac{\partial \phi_B}{\partial B_R}\right]^-}{\left[\frac{\partial d_\beta}{\partial B_R}\right]^+ + \lambda\left[\frac{\partial \phi_B}{\partial B_R}\right]^+} \quad (12)$$

$$G_R = G_R \odot \frac{\left[\frac{\partial d_\beta}{\partial G_R}\right]^- + \lambda\left[\frac{\partial \phi_G}{\partial G_R}\right]^-}{\left[\frac{\partial d_\beta}{\partial G_R}\right]^+ + \lambda\left[\frac{\partial \phi_G}{\partial G_R}\right]^+} \quad (13)$$

$$\left[\frac{\partial d_\beta}{\partial B_R}\right]^- = \left[\left(\hat{X}_R + \hat{X}_W\right)^{\beta-2} \odot X_{n_\beta}\right] G_R{}' \quad (14)$$

$$\left[\frac{\partial d_\beta}{\partial B_R}\right]^+ = \left[\left(\hat{X}_R + \hat{X}_W\right)^{\beta-1}\right] G_R{}' \quad (15)$$

$$\left[\frac{\partial d_\beta}{\partial G_R}\right]^- = B_R{}'\left[\left(\hat{X}_R + \hat{X}_W\right)^{\beta-2} \odot X_{n_\beta}\right] \quad (16)$$

$$\left[\frac{\partial d_\beta}{\partial G_R}\right]^+ = B_R{}'\left[\left(\hat{X}_R + \hat{X}_W\right)^{\beta-1}\right] \quad (17)$$

$$\left[\frac{\partial \phi_B}{\partial B_R}\right]^+_{f,p} = \frac{4FB_{R_{f,p}}}{\sum_{j=1}^{F} B_{R_{j,p}}^2} \quad (18)$$

$$\left[\frac{\partial \phi_G}{\partial G_R}\right]^+_{p,t} = \frac{4TG_{R_{p,t}}}{\sum_{i=1}^{T} G_{R_{p,i}}^2} \quad (19)$$

$$\left[\frac{\partial \phi_B}{\partial B_R}\right]^-_{f,p} = 2F\left[\frac{\left(B_{R_{f-1,p}} + B_{R_{f+1,p}}\right)}{\sum_{j=1}^{F} B_{R_{j,p}}^2}\right] + \frac{2FB_{R_{f,p}} \sum_{j=2}^{F}\left(B_{R_{j,p}} - B_{R_{j-1,p}}\right)^2}{\left(\sum_{j=1}^{F} B_{R_{j,p}}^2\right)^2} \quad (20)$$

$$\left[\frac{\partial \phi_G}{\partial G_R}\right]^-_{p,t} = 2T\left[\frac{\left(G_{R_{p,t-1}} + G_{R_{p,t+1}}\right)}{\sum_{i=1}^{T} G_{R_{p,i}}^2}\right] + \frac{2TG_{R_{p,t}} \sum_{i=2}^{T}\left(G_{R_{p,i}} - G_{R_{p,i-1}}\right)^2}{\left(\sum_{i=1}^{T} G_{R_{p,i}}^2\right)^2} \quad (21)$$

Substituting the wheezing basis B_W and the wheezing gain G_W into eq. (3), the multiplicative update rules of the wheezing are formulated in eq. (22) and (23),

$$B_W = B_W \odot \frac{\left[\frac{\partial d_\beta}{\partial B_W}\right]^- + \alpha_B\left[\frac{\partial \psi_B}{\partial B_W}\right]^-}{\left[\frac{\partial d_\beta}{\partial B_W}\right]^+ + \alpha_B\left[\frac{\partial \psi_B}{\partial B_W}\right]^+} \quad (22)$$

$$G_w = G_w \odot \frac{\left[\frac{\partial d_\beta}{\partial G_W}\right]^-}{\left[\frac{\partial d_\beta}{\partial G_W}\right]^+} \quad (23)$$

$$\left[\frac{\partial d_\beta}{\partial B_W}\right]^- = \left[\left(\hat{X}_R + \hat{X}_W\right)^{\beta-2} \odot X_{n_\beta}\right] G_W{}' \quad (24)$$

$$\left[\frac{\partial d_\beta}{\partial B_W}\right]^+ = \left[\left(\hat{X}_R + \hat{X}_W\right)^{\beta-1}\right] G_W{}' \quad (25)$$

$$\left[\frac{\partial d_\beta}{\partial G_W}\right]^- = B_W{}'\left[\left(\hat{X}_R + \hat{X}_W\right)^{\beta-2} \odot X_{n_\beta}\right] \quad (26)$$

$$\left[\frac{\partial d_\beta}{\partial G_W}\right]^+ = B_W{}'\left[\left(\hat{X}_R + \hat{X}_W\right)^{\beta-1}\right] \quad (27)$$

$$\left[\frac{\partial \psi_B}{\partial B_W}\right]^+_{f,q} = \frac{1}{\sqrt{\frac{1}{F}\sum_{j=1}^{F} B_{W_{j,q}}^2}} \quad (28)$$

$$\left[\frac{\partial \psi_B}{\partial B_W}\right]^-_{f,q} = \sqrt{F}\frac{B_{W_{f,q}} \sum_{j=1}^{F} B_{W_{j,q}}}{\left(\sum_{j=1}^{F} B_{W_{j,q}}^2\right)^{\frac{3}{2}}} \quad (29)$$

3.4 Signal Reconstruction

The estimated respiratory signal $\hat{x}_r(t)$ and the estimated wheezing signal $\hat{x}_w(t)$ can be synthesized from the magnitude estimated spectrograms $\hat{X}_R$ (see eq. (30)) and $\hat{X}_W$ (see eq. (31)). Wiener filtering [32] has been used to create the masks that guarantee a conservative reconstruction. The Wiener masks S_R (see eq. (32)) and S_w (see eq. (33)) represent the contribution of each signal in the spectrogram of mixture. In order to obtain the estimated complex spectrograms of the separated signals, each mask is multiplied by the complex spectrogram X_c of the mixture $x(t)$. Finally, the inverse STFT is applied to synthesize the respiratory signal and the wheezing signal in time domain.

$$\hat{X}_R = B_R G_R \quad (30)$$

$$\hat{X}_w = B_w G_w \quad (31)$$

$$S_R = \frac{\hat{X}_R^2}{\hat{X}_W^2 + \hat{X}_R^2} \quad (32)$$

$$S_W = \frac{\hat{X}_W^2}{\hat{X}_R^2 + \hat{X}_W^2} \quad (33)$$

$$\hat{x}_r(t) = IDFT(S_R X_c) \quad (34)$$

$$\hat{x}_w(t) = IDFT(S_W X_c) \quad (35)$$

4. EXPERIMENTAL RESULTS

4.1 Dataset, Setup and Metrics

The dataset (denoted DORIG) is composed of twenty mixtures. Each mixture, with duration between 5 and 20 seconds, is composed of wheezes signals (only wheezes) and respiratory signals (only normal breath sounds) from Internet pulmonary data sources [2], [22]-[26] that show an initial signal-to-noise ratio (SNR) between 2dB and 8dB. The wheezes used in the mixtures have been manually separated in order to evaluate the proposed method using objective metrics. For each mixture, the number of wheezes and the temporal location of each of them have been determined using a pseudo-random process that uses the standard uniform distribution.

Two new datasets have been created to evaluate the robustness of the proposed method. Specifically, each new dataset has been created mixing the original wheeze signals and respiratory signals to obtain SNR=0dB (denoted as D0dB) and SNR=-5dB (D5dB).

All mixtures were band-limited from 100Hz-1000Hz (as previously mentioned, we assume that wheezes are not active below 100Hz). The other processing parameters are the following ones: sampling rate fs = 2KHz, size of Hamming window N =128 samples, with 50% overlap (temporal resolution of 62.5ms). In our experiments, we have used the same respiratory and wheezing components $P = Q = 50$. In addition, the number of iterations used for each NMF decomposition is equal to 100 iterations because the convergence of NMF is reached. Because the NMF separation performance depends on the initial values of basis and

gain matrices, we have repeated five times each separation for each mixture and the results shown in this work are averaged values.

The evaluation of the proposed method has been performed using three metrics [27], [28], which are widely used in the field of sound source separation [17], [18], [32]: source-to-distortion ratio (SDR) reports information on the overall sound quality of the separation process; source-to-interferences ratio (SIR) indicates the presence of respiratory sounds in wheezes and vice versa; and source-to-artifacts ratio (SAR) reports information on the artifacts from separation and/or resynthesis.

4.2 Results

In order to compare the separation performance of the proposed method compared to standard NMF, Figures 3, 4 and 5 show SDR, SIR and SAR results evaluating the three datasets (DORIG, D0dB and D5dB) previously mentioned. Figures 3, 4 and 5 show that the proposed method obtains the best SDR, SIR and SAR results compared to standard NMF for all the datasets evaluated. The reason is because the proposed constraints included into the NMF factorization achieve to find a better local minimum that provides physical meaning to the estimated wheezing and respiratory bases and gains as can be observed in real life. Specifically, the proposed method improves the standard NMF about 9dB in SDR and 10dB in SIR evaluating the dataset DORIG, about 6.3dB in SDR and 10dB in SIR evaluating the dataset D0dB and about 4.9dB in SDR and 13dB in SIR evaluating the dataset D5dB.

Focusing on the datasets DORIG and D0dB, results indicate a promising respiratory/wheeze performance of the proposed method in SDR, SIR and SAR metrics. Focusing on the dataset D5dB that simulates a high noisy environment in which the wheezes are barely audible in the mixtures due to the high respiratory interference, the proposed method obtains satisfactory audio quality of the estimated wheezes. Specifically, the proposed method still recovers most of the wheezing sounds that cannot be heard in the mixture. This fact can be considered crucial since it could help to reduce the number of false negatives diagnosed by the physician. Focusing on standard NMF, it cannot successfully separate respiratory and wheezing sounds into any datasets evaluated being its separation performance considerably worse as noise increases. In other words, standard NMF is not a reliable method because it can only ensure convergence to local minima that enables the signal reconstruction without physical meaning unlike occurs with the proposed method. In particular, standard NMF is not able to recover any wheezing that cannot be heard in any mixture analysing the dataset D5dB.

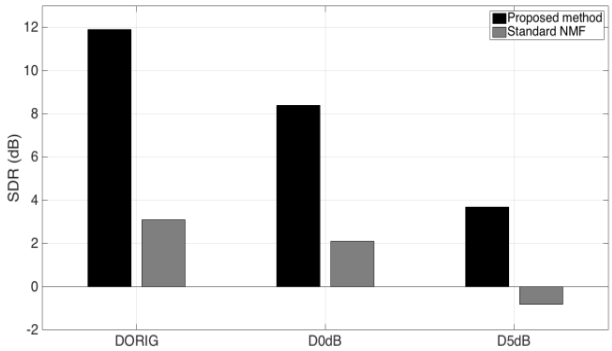

Figure 3. SDR results. Each bar represents the average value of SDR for each evaluated dataset.

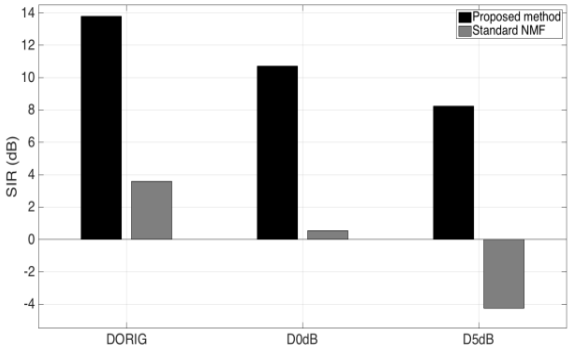

Figure 4. SIR results. Each bar represents the average value of SIR for each evaluated dataset.

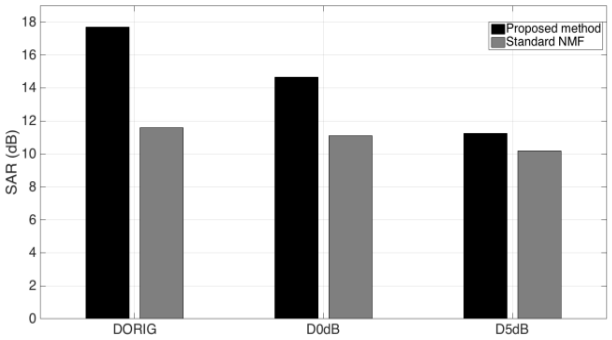

Figure 5. SAR results. Each bar represents the average value of SAR for each evaluated dataset.

To illustrate the separation performance of the proposed method, we display the spectrograms of the original respiratory and wheezing spectrograms that composed the input mixture (Figure 6), the estimated wheezing signals (Figure 7) and the estimated respiratory signals (Figure 8) provided by the standard NMF and the proposed method.

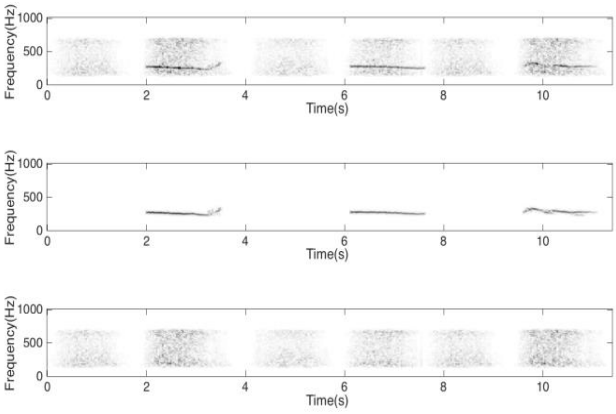

Figure 6. Magnitude spectrogram of a mixture (top) from dataset D0dB. The original wheezes (middle) and the original respiratory sounds (bottom) that compose the mixture.

Figure 7 (top) shows that the proposed method reconstructs a promising wheezing spectrogram with an audio quality comparable to the original sound used in the mixture (Figure 6 (middle)). However, standard NMF (Figure 7 (bottom)) does not obtain a satisfactory wheezing reconstruction because a large amount of respiratory sound still remains active and a high percentage of wheezing has been lost. Specifically, the third wheeze has been completely removed.

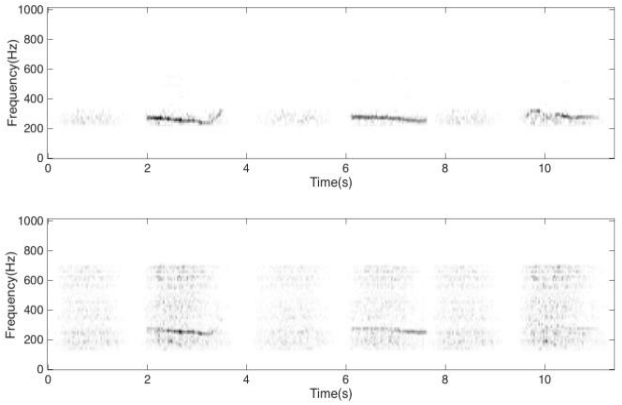

Figure 7. The estimated wheezing signals provided by the proposed method (top) and standard NMF (bottom) from the mixture shown in Figure 6.

As can be seen in Figure 8 (top), the proposed method removes most of the respiratory sounds from the mixture avoiding to capture wheeze sounds. Instead, standard NMF can only remove a small proportion of respiratory sounds but at the expense of eliminating a large amount of wheezes. It can be seen as some of the first and second wheezing, along with almost all of the third wheezing, are present in the spectrogram provided by standard NMF. This fact confirms that standard NMF does not work because a remarkable disadvantage of standard NMF is the lost of wheezing content when the objective is only to remove the respiratory sounds.

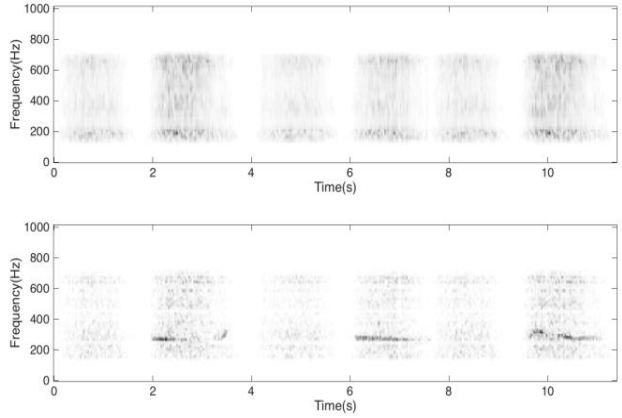

Figure 8. The estimated respiratory sounds provided by the proposed method (top) and standard NMF (bottom) from the mixture shown in Figure 6.

5. CONCLUSIONS AND FUTURE WORK

This paper presents a method for separating respiratory sounds and wheezing. Our proposal is based on unsupervised NMF approach that automatically distinguishes between respiratory and wheezing bases using the constraints smoothness or sparseness into the NMF decomposition process. The developed method includes the following advantages: (i) modelling of respiratory sounds and wheezing adding physical meaning to the NMF solution, and (ii) no supervised training is required to classify the bases. One of the disadvantages observed in the proposed method is the presence of spurious components in the temporal interval in which only the respiratory sounds are present.

Results show that the proposed method is robust compared to different signal-to-noise ratios and significantly outperforms standard NMF. The proposed method, adding smoothness and sparseness constrains, obtains better factorization solutions because the proposed constrained NMF obtains object-parts with physical meaning as can be found in the respiratory and wheeze sounds in real life. The proposed method still recovers most of the wheezing sounds that cannot be heard in the mixtures even when a high noisy environment is evaluated (the wheezes are barely audible in the mixtures due to the high respiratory interference). This fact can be considered crucial because it can improve the physician's diagnosis reducing the number of false negatives. However, standard NMF cannot successfully separate respiratory and wheezing sounds into any datasets evaluated.

Future works will focus on two directions: removal of respiratory spurious sounds active in the estimated wheeze signal and development of a wheezing detector based on the proposed method.

6. ACKNOWLEDGMENTS

This work was supported by the Spanish Ministry of Economy and Competitiveness under Project TEC2015-67387-C4-2-R.

7. REFERENCES

[1] World Health Organization, Chronic respiratory diseases, http://www.who.int/respiratory/asthma/en/.

[2] Oletic, D., Bilas, V. Asthmatic Wheeze Detection from Compressively Sensed Respiratory Sound Spectra, IEEE Journal of Biomedical and health informatics, vol. PP, no. 99, 2017 DOI: 10.1109/JBHI.2017.2781135.

[3] Sovijarvi, F., Dalmasso, F,. Vanderschoot, J., Malmberg, L., Righini, G., and Stoneman, S. Definition of terms for applications of respiratory sounds, European Respiratory Review, vol. 10, pp. 597-610, 2000.

[4] Sarkar, M., Madabhavi, I., Niranjan, N., and Dogra, M. Auscultation of the respiratory system, Annals of Thoracic Medicine, vol. 10, no. 3, pp. 158-168, 2015.

[5] Lozano F, Salazar A, Alvarado C. System of heart and lung sounds separation for store-and-forward telemedicine applications. Revista Facultad Ingennieria Univ. Antioquia, 2012.

[6] Lin, B., Lin, B., Wu, H., Chong, F., Chen, S. Wheeze recognition based on 2D bilateral filtering of spectrogram, Biomedical Engineering Applications, basis and communications, vol. 18, no. 3, 2006, DOI= https://doi.org/10.4015/S1016237206000221.

[7] Taplidou, S., and Hadjileontiadis, L. Wheeze detection based on time-frequency analysis of breath sounds, Comput. Biol. Med., vol. 37, no. 8, pp.1073-1083, 2007.

[8] Riella, R., Nohama, P., and Maia, J., Method for automatic detection of wheezing in lung sounds, Braz. J. Med. Biol. Res., vol. 42, no. 7, pp. 674-684, 2009.

[9] Qiu, Y., Whittaker, A., Lucas, M., and Anderson, K. Automatic wheeze detection based on auditory modeling, in Proc. Inst. Mech. Eng. H., vol. 219, no. 3, pp. 219-227, 2005.

[10] Feng, J., Farook, S., and Daniel, G. Automatic wheeze detection using histograms of sample entropy, in Conf. Proc. IEEE Eng. Med. Biol. Soc., pp. 1890-1893, 2008.

[11] Mazic, I., Bonkovic, M., Dzaja, B. Two-level coarse-to-fine classification algorithm for asthma wheezing recognition in children's respiratory sounds, Biomedical Signal Processing and Control, vol. 21, pp. 105-118, 2015.

[12] Kandaswamy, A., Kumar, C., Ramanathan, R., Jayaraman, S., and Malmurugan, N. Neural classification of lung sounds using wavelet coefficients, Comput.Biol. Med., vol. 34 pp. 523–537, 2004.

[13] Chien, J., Wu, H., Chong, F., Li, C., Wheeze detection using cepstral analysis in Gaussian Mixture Models, Conf. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf. pp. 3168–3171, 2007.

[14] Wisniewski, M., Zielinski, T., Application of tonal index to pulmonary wheezes detection in asthma monitoring, 19th European Signal Processing Conference (EUSIPCO), 2011.

[15] Hadjileontiadis, L., Panas, S. Separation of Discontinuous Adventitious Sounds from Vesicular Sounds Using a Wavelet-Based Filter, IEEE Transactions on biomedical engineering, vol. 44, no. 12, 1997.

[16] Bahoura, M., Lu, X. Separation of Crackles From Vesicular Sounds Using Wavelet Packet Transform, IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, 2006.

[17] Canadas-Quesada, F., Ruiz-Reyes, N., Carabias-Orti, J., Vera-Candeas, P., and Fuertes-Garcia, J. A non-negative matrix factorization approach based on spectro-temporal clustering to extract heart sounds, Applied Acoustics, vol. 125, pp. 7-19, 2017.

[18] Canadas-Quesada, F., Vera-Candeas, P., Martinez-Munoz, D., Ruiz-Reyes. N., Carabias-Orti, J., and Cabanas-Molero, P. Constrained non-negative matrix factorization for score-informed piano music restoration, Digital Signal Processing, vol. 50, pp. 240-257, 2016.

[19] Monga, V., and Mhcak, M. Robust and secure image Hashing via non-negative matrix factorizations. IEEE Trans. Inf. Forensics Secur., vol. 2, no. 3, pp. 376-390, 2007.

[20] Monga, V., and Mhcak, M. Robust and secure image Hashing via non-negative matrix factorizations. IEEE Trans. Inf. Forensics Secur., vol. 2, no. 3, pp. 376-390, 2007.

[21] Lee, D., and Seung, S. Algorithms for non-negative matrix factorization, in Proceedings of Advances in Neural Inf. Process. System, pp. 556–562, 2000.

[22] The r.a.l.e. repository. [Online]. Available: http://www.rale.ca/.

[23] Stethographics lung sound samples. [Online]. Available: http://www.stethographics.com/.

[24] 3m littmann stethoscopes. [Online]. Available: http://solutions.3m.com/.

[25] East tennessee state university pulmonary breath sounds. [Online]. Available: http://faculty.etsu.edu.

[26] ICBHI 2017 Challenge. [Online]. Available: https://bhichallenge.med.auth.gr/sites/default/.

[27] Vincent, E., Fevotte, C., Gribonval, R. Performance measurement in blind audio source separation. IEEE Trans. Audio Speech Lang. Process, vol. 14, no. 4, pp. 1462-1469, 2006.

[28] Fevotte, C., Gribonval, R., Vincent, E. BSS_EVAL toolbox user guide - Revision, 2.0, Technical Report 1706, IRISA (April 2005).

[29] Fevotte, C., Bertin, N., Durrieu, JL. Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. Neural Comput. 21(3), 793830 (2009).

[30] Virtanen, T. Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. IEEE Trans. Audio Speech Lang. Process. 15(3), 1066–1074 (2007).

[31] Eggert, J., Korner, E. Sparse coding and NMF, in Proceedings of the International Joint Conference on Neural Networks (IJCNN4) (Budapest, Hungary, 25–29 July 2004), pp. 2529–2533.

[32] Parras-Moral, J., Canadas-Quesada, F., Vera-Candeas, P., Ruiz-Reyes. N. Audio restoration of solo guitar excerpts using a excitation-filter instrument model, in Stockholm Music Acoustics Conference jointly with Sound And Music Computing Conference (Stockholm, Sweden, 30 July).

Paper 2

A novel wheezing detection approach based on constrained non-negative matrix factorization

J. Torre-Cruz, F. Canadas-Quesada, J. Carabias-Orti, P. Vera-Candeas and N. Ruiz-Reyes, "A novel wheezing detection approach based on constrained non-negative matrix factorization", in *Applied Acoustics*, Volume 148, May 2019, pp. 276-288. DOI: `https://doi.org/10.1016/j.apacoust.2018.12.035`

- Estado: Publicado.
- Revista: Applied Acoustics.
- ISSN: 0003-682X.
- Factor de impacto (JCR 2019): 2.440.
- Cuartiles por área de conocimiento:
  - Acoustics: Q2, 9/32.

Applied Acoustics 148 (2019) 276–288

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

A novel wheezing detection approach based on constrained non-negative matrix factorization

J. Torre-Cruz *, F. Canadas-Quesada, J. Carabias-Orti, P. Vera-Candeas, N. Ruiz-Reyes

Department of Telecommunication Engineering, University of Jaen, Campus Cientifico-Tecnologico de Linares, Avda. de la Universidad, s/n, 23700 Linares, Jaen, Spain

ARTICLE INFO

Article history:
Received 20 September 2018
Received in revised form 3 December 2018
Accepted 27 December 2018
Available online 4 January 2019

Keywords:
Detection
Non-negative matrix factorization (NMF)
Divergence
Wheezing
Smoothness
Sparseness

ABSTRACT

The early wheezing detection is still a challenging task in biomedical signal processing because the presence of wheeze sounds often indicate respiratory diseases from airway obstructions. Currently, most of the first clinical examinations to detect any airway obstructions are carried out using auscultation. However, a high percentage of diagnoses are misdiagnosed since they are highly dependent on the physician's training in the wheezing detection, especially in noisy environments in which weak wheeze sounds can be masked by louder respiratory sounds. In this work, we propose a novel wheezing detection approach, based on Constrained Non-negative Matrix Factorization, that uses two-stage cascade: separation and detection. The novelty of the separation stage is to model wheeze and respiratory sounds as reliably as possible that they can be observed in the nature incorporating constraints (sparseness and smoothness) into the NMF factorization. Once the estimated wheezing and respiratory signal are obtained from the separation stage, the detection contribution is based on the use of the Kullback-Leibler divergence to discriminate between wheezing and respiratory areas. The experiments have been conducted using three different datasets composed of healthy or unhealthy patients. First, an optimization process is applied to obtain the optimal parameters of the separation stage. Finally, the performance of the wheezing detection of the proposed method is evaluated taking into account other state-of-the-art methods. Experimental results report that i) the proposed method outperforms recent state-of-the-art wheezing detection approaches showing a robust wheezing detection performance even evaluating noisy environments and ii) the ability of the proposal to reliably detect healthy patients.

1. Introduction

Ambient Assisted Living (AAL) based on acoustic event detection (AED) is currently a challenging topic that has attracted the attention of the signal processing community in the last years [1–4]. In fact, the analysis of respiratory sounds to diagnose patients suffering from lung diseases can be considered as a specific AED task applied in the field of biomedical signal processing [5,6] because it attempts to maximize the reliability of diagnoses by reducing the degree of subjectivity provided by physicians.

Specifically, the appearance of wheezing is an indicator of breathing problems shown by several obstructive pulmonary diseases, such as asthma, bronchiolitis, bronchitis or bronchiectasis, which affect people of all ages worldwide. These respiratory diseases are mainly caused by airway obstructions when air moves through narrowed or swelling breathing tubes in the lungs, reducing the amount of air that can pass through the air passages [7]. As an example, in 2015, there were nearly 400,000 deaths from asthma, most of which occurred in low- and middle-income countries according to the World Health Organization (WHO) [8]. As a consequence, the wheezing detection is one of the most challenging tasks in biomedical engineering and bio-signal processing research community.

Focusing on wheezing context, the aim of auscultation is to listen wheeze sounds emitted in the human breathing process using simple medical instrumentation such as stethoscopes. Currently, most of the first clinical examinations to detect respiratory diseases from airway obstructions are carried out by means of auscultation since it is a non-invasive, low-cost, easy-to-perform, patient-friendly and fast method regardless of age [9]. However, a high percentage of diagnoses are misdiagnosed since they are highly dependent on the physician's experience and acoustic training in the auscultation process. Therefore, it would be crucial to develop a robust wheezing detection system in order to early prevent complications resulting from misdiagnosis or undetected airway obstruction symptoms.

* Corresponding author.
E-mail address: jtorre@ujaen.es (J. Torre-Cruz).

https://doi.org/10.1016/j.apacoust.2018.12.035
0003-682X/

It is well-known the temporal and spectral interference between wheeze and respiratory sounds during the inspiration and/or expiration of the breathing stages. The temporal interference is caused because wheeze sounds and respiratory sounds are simultaneously generated by the same airflow through the bronchial tree of the lungs. Instead, the spectral interference is due to the spectral overlapping problem that occurs in the spectral bands in which both types of sounds are active. Respiratory sounds (RS) are represented by a wide-band spectrum where most of the energy is concentrated in the frequency band 60 Hz-1000 Hz [10]. However, wheezing or wheeze sounds (WS) can be defined as continuous adventitious sounds that show a pitched sound generated on during the stages of the human breathing process. Specifically, WS show sinusoidal waveforms in time domain, a set of narrowband spectral peaks forming frequency lines over time (spectral trajectories) that superimpose on normal respiratory sounds in the frequency domain, whose pitch frequency is located in the spectral range of 100 Hz–1000 Hz with duration longer than 100 ms according to Computerized Respiratory Sound Analysis (CORSA) [11–13] as can be seen in Fig. 1. Specifically, WS can be classified into two main categories: monophonic and polyphonic. Monophonic WS are only composed of one fundamental frequency (pitch) but polyphonic WS are composed of one pitch and its harmonically-related frequencies that are located at approximately integer multiple of its associated pitch. In this paper, the term RS only refers to that sounds emitted by the lungs due to the breathing process without WS. The term WS only refers to that sounds composed of wheezing without RS. The term mixture refers to a signal composed of RS and WS.

Over the last two decades, many works have been reported to achieve wheezing detection system. Although several methods can be found in the state-of-the-art literature, which are based on different approaches such as auditory modelling [14], entropy [15], neural networks [16–18], wavelet transform [19,20], tonal index [21,22], mel-frequency cepstral coefficients (MFCC) [23–25] and classifiers [26–29], the most widely used methods to detect wheeze sounds are based on the extracted information provided by the analysis of the spectral peaks of the spectrum [30–35]. Alic et al. [31] modified the searching of spectral peaks using wavelet denoising in order to remove the noise in spectrum. In [32], an enhanced wheeze detector is developed which automatically locates and identifies wheezing-episodes during breath sound recordings based on the subtraction of the underlying breath sound, detection and classification of the detected peaks as wheezes and non-wheezes. Bahoura [24] reported pattern recognition methods to classify RS and WS using features based on Fourier transform, linear predictive coding, wavelet transform and Mel-frequency cepstral coefficients (MFCC) in combination with the classification methods based on vector quantization, Gaussian mixture models (GMM) and artificial neural networks. Wisniewski and Zielinski [21] analysed wheezes detection in normal breath sound as a problem of detection of multi-tones in colored noise using a set of robust descriptors. In [26], a two-layer pattern recognition system architecture for asthma wheezing detection is developed. The first layer consists of two SVM classifiers specifically designed as a cascade stacked in parallel using features based on MFCC. The second layer is realized using a digital detection threshold, with the aim of improving the process of wheezing detection. Shaharum et al. [28] detected wheezing to classify different levels of asthma severity using a feature extraction based on MFCC in combination with the classification method based on the K-nearest neighbour (KNN) algorithm. Recently, Oletic and Vilas [29] presented a wheezing detector by iteratively detecting individual wheezing frequency lines by modelling them using HMM. Its robustness was dominantly influenced by the robustness of frequency line tracking, the method of temporal localization of frequency line and the implementation of the line subtraction. Nevertheless, the main drawback of the most wheezing detection methods is the assumption that spectral peaks of WS are louder than the RS with which they are acoustically mixed [36,30,13,31,20,37,21,38,27].

Because non-negative matrix factorization has the ability to find hidden spectral patterns [39,43] by means of parts-based representation with non-negativity of the data, we propose a novel constrained non-negative matrix factorization approach to model and detect the presence of wheezing, locating the temporal intervals in which WS are active when they are mixed with RS in mono-channel audio mixtures. As far as the authors knowledge extends, non-negative matrix factorization approach has never been applied before to wheezing detection. The main contribution of this work is a robust wheezing detector based on the estimated respiratory and wheezing signals obtained by the proposed constrained NMF. Thus, the proposed method is able to detect and separate the wheezing and respiratory signals from the mixture. In fact, the mixture is decomposed using a cost function to reconstruct the mixture adding typical spectro-temporal behaviors observed in most WS and RS in real life. We assume that WS can be modeled using sparseness in frequency since WS usually show one or a few harmonically-related narrow peaks in the spectrum. We assume that RS are represented as a wide-band spectrum that

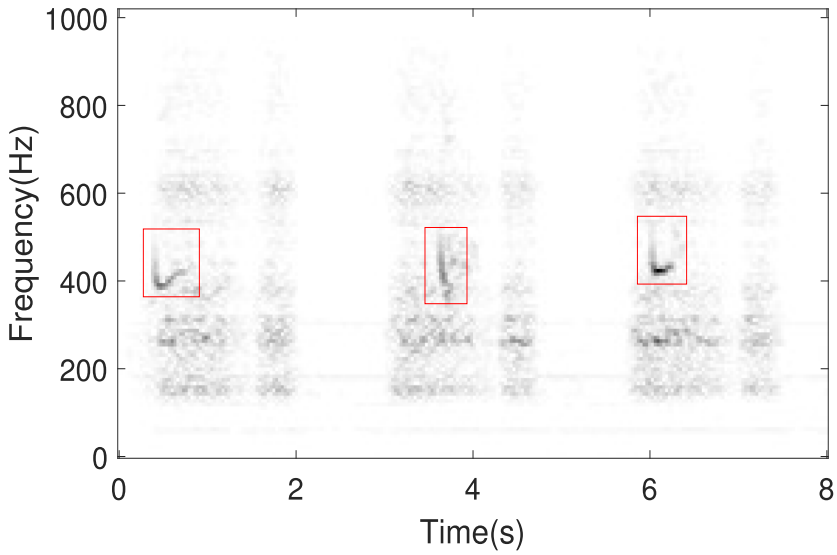

Fig. 1. Spectrogram of a mixture composed of three wheeze sounds (red rectangles) and respiratory sounds. A darker grey colour represents higher energy of each frequency.

shows smoothness (slow variation) in time and frequency. Specifically, RS are the only sounds of the mixture that is not overlapped in some areas (particularly, in the areas of the respiratory cycle where wheezing is inactive). Therefore, we propose the use of the Kullback-Leibler divergence applied to the input spectrogram and the estimated respiratory spectrogram to discriminate between wheezing and respiratory temporal intervals (areas). The Kullback-Leibler divergence will have a very small value in areas where only the RS are present. However, the Kullback-Leibler divergence will have a high value in areas where RS and WS are active. Our proposal is a completely blind method because does not require information about the number of sound sources neither training of the sounds to detect.

The remainder of this paper is organized as follows. In Section 2, a description about the fundamentals of non-negative matrix factorization is briefly presented. We subsequently propose a robust method to detect wheeze sounds in Section 3. In Section 4, an optimization of the main parameters of the separation stage is carried out in order to maximize the detection performance of the proposed method. Finally, we conclude in Section 5 and provide perspectives on further research.

2. Non-negative matrix factorization

Recently, non-negative Matrix Factorization (NMF) or unconstrained NMF [39,40] has attracted a lot of attention in the field of biomedical signal processing [41–44] because it provides parts-based representation of the most representative objects by imposing non-negative constraints that allow only additive combinations of the input data. Specifically, NMF factorizes the input magnitude spectrogram $\mathbf{X}_{F,T}$ of a mixture signal $x(t)$ into the product of two non-negative estimated matrices, basis matrix $\hat{\mathbf{B}}_{F,K}$ and activation matrix $\hat{\mathbf{A}}_{K,T}$ (see Eq. (1)),

$$\mathbf{X}_{F,T} \approx \hat{\mathbf{X}}_{F,T} = \hat{\mathbf{B}}_{F,K}\hat{\mathbf{A}}_{K,T} \tag{1}$$

where $\hat{\mathbf{X}}_{F,T}$ is the estimated or reconstructed spectrogram and the variables F, T and K represent the number of frequency bins, the number of time frames and the rank or the number of components (generally, $FK + KT \ll FT$ in order to reduce the dimensions of the data). The columns of the estimated basis matrix $\hat{\mathbf{B}}$ are basis functions (or spectral patterns) and the rows of the estimated activation matrix $\hat{\mathbf{A}}$ represent the temporal intervals in which the previous basis functions are active. The NMF factorization is performed by minimizing a cost function $D\left(\mathbf{X}|\hat{\mathbf{X}}\right)$,

$$D\left(\mathbf{X}|\hat{\mathbf{X}}\right) = \sum_{f=1}^{F}\sum_{t=1}^{T} d\left(X_{f,t}|\hat{X}_{f,t}\right) \tag{2}$$

where $d(i,j)$ is a function of two scalar variables i, j. Some of the most widely used cost functions applied to audio processing are the Euclidean distance, the generalized Kullback-Leibler divergence and the Itakura-Saito divergence [45]. The cost function $D\left(\mathbf{X}|\hat{\mathbf{X}}\right)$ is minimized, ensuring the non-negativity of the bases and the activations using an iterative algorithm based on multiplicative update rules [45]. Given a parameter $\mathbf{Z}$, its multiplicative update rules are obtained calculating the partial derivatives of the cost function as follows,

$$\mathbf{Z} = \mathbf{Z} \odot \frac{\left[\frac{\partial D\left(\mathbf{X}|\hat{\mathbf{X}}\right)}{\partial \mathbf{Z}}\right]^{-}}{\left[\frac{\partial D\left(\mathbf{X}|\hat{\mathbf{X}}\right)}{\partial \mathbf{Z}}\right]^{+}} \tag{3}$$

where $\odot$ is the element-wise multiplication.

However, the main problem of the NMF is the trade-off between signal reconstruction and physical interpretation of the factorized parts-based objects. In other words, the non-negativity can only ensure convergence to local minima that implies the reconstruction of the mixture but the solution may not make physical sense as these type of sounds can be found in real life [46,47]. The previous problem can be overcome adding some prior information into the factorization procedure. This prior information can be added by means of constraints. As result, constraints help to find better local minima incorporating physical sense to the basis or activations into the factorization procedure. Next, two specific constraints used in our proposal (detailed in Section 3) are briefly explained: sparseness and smoothness.

2.1. Sparseness

In general terms, sparseness ψ means that the sources can be considered inactive most of the time or frequency [48,49]. Therefore, sparseness can be applied to the estimated NMF basis or activation matrices. Temporal sparseness $\psi\left(\hat{\mathbf{A}}\right)$, applied to estimated activation matrix $\hat{\mathbf{A}}$, assumes that the sources are inactive most of the time so, a high cost is assigned to nonzero activations. Spectral sparseness $\psi\left(\hat{\mathbf{B}}\right)$, applied to estimated basis matrix $\hat{\mathbf{B}}$, assumes that the sources are inactive most of the frequency so, a high cost is assigned to nonzero bases. As a result, sparseness constraint ψ often obtains better local minima in the NMF decomposition, minimizing the presence of false activations or bases. For example, considering the spectral sparseness,

$$D\left(\mathbf{X}|\hat{\mathbf{X}}\right) = D_{*}\left(\mathbf{X}|\hat{\mathbf{X}}\right) + \alpha\psi\left(\hat{\mathbf{B}}\right) \tag{4}$$

where $D_{*}\left(\mathbf{X}|\hat{\mathbf{X}}\right)$ is the reconstruction cost function to be minimized (i.e. Euclidean, Kullback-Leibler, ...) and $\psi\left(\hat{\mathbf{B}}\right)$ penalizes nonzero bases as previously mentioned. Specifically, one of the most used penalty term is the L^1-norm $\psi\left(\hat{\mathbf{B}}\right) = \left\|\hat{\mathbf{B}}\right\|_1$ as proposed in [49] because it was demonstrated to be less sensitive to changes of the parameter α that controls the importance of the constraint in the factorization process. In a similar way, the temporal sparseness $\psi\left(\hat{\mathbf{A}}\right)$ can be calculated but now taking into account the estimated activations matrix $\hat{\mathbf{A}}$.

2.2. Smoothness

Generally, smoothness ϕ means how continuous or smooth are the spectral or temporal changes related to a source [49]. Therefore, smoothness can be applied to the estimated NMF basis or activation matrices. Temporal smoothness $\phi\left(\hat{\mathbf{A}}\right)$, applied to estimated activation matrix $\hat{\mathbf{A}}$, reports how slow the amplitude variations over time are. Spectral smoothness $\phi\left(\hat{\mathbf{B}}\right)$, applied to estimated basis matrix $\hat{\mathbf{B}}$, indicates how slow the amplitude variations over frequency are. For example, consider temporal smoothness $\phi\left(\hat{\mathbf{A}}\right)$,

$$D\left(\mathbf{X}|\hat{\mathbf{X}}\right) = D_{*}\left(\mathbf{X}|\hat{\mathbf{X}}\right) + \lambda\phi\left(\hat{\mathbf{A}}\right) \tag{5}$$

where $D_{*}\left(\mathbf{X}|\hat{\mathbf{X}}\right)$ is the cost function, $\phi\left(\hat{\mathbf{A}}\right)$ is the function that penalizes abrupt temporal changes and parameter λ controls the importance of the smoothness constraint. According [49], temporal smoothness $\phi\left(\hat{\mathbf{A}}\right)$ of the components is enforced by assigning a

high cost to large changes in time between the activations $\widehat{A}_{f,t}$ and $\widehat{A}_{f,t-1}$ in adjacent frames as follows,

$$\phi\left(\hat{\mathbf{A}}\right) = \sum_{f=1}^{K} \frac{1}{\sigma_f^2} \sum_{t=2}^{T} \left(\widehat{A}_{f,t} - \widehat{A}_{f,t-1}\right)^2 \tag{6}$$

where the activations are normalised by their standard deviation $\sigma_f = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\widehat{A}_{f,t}^2}$ to prevent the numerical scale of the activations from affecting the cost [49,46]. In a similar way, the spectral smoothness $\phi\left(\hat{\mathbf{B}}\right)$ is calculated but considering the spectral changes between the bases $\widehat{B}_{f,t}$ and $\widehat{B}_{f-1,t}$ in adjacent bins as follows,

$$\phi\left(\hat{\mathbf{B}}\right) = \sum_{t=1}^{K} \frac{1}{\sigma_t^2} \sum_{f=2}^{F} \left(\widehat{B}_{f,t} - \widehat{B}_{f-1,t}\right)^2 \tag{7}$$

where the bases are normalised by their standard deviation $\sigma_t = \sqrt{\frac{1}{F}\sum_{f=1}^{F}\widehat{B}_{f,t}^2}$ to prevent the numerical scale of the bases from affecting the cost [49,46].

3. The proposed method

The main problem to detect wheeze sounds from mixture is that both wheeze sounds and respiratory sounds occur simultaneously in frequency and time domain. Therefore, the goal of the proposed method is to improve the wheezing detection applying wheeze/respiratory sound separation. For this purpose, our proposed method consists of two-stage cascade: wheeze/respiratory sound separation (Stage I) and wheezing detection (Stage II). A block diagram of the proposed method is shown in Fig. 2.

3.1. Stage I. Wheeze/respiratory sound separation

The mixture $x(t)$ of unhealthy patients is composed of wheeze sounds $x_w(t)$ and respiratory sounds $x_r(t)$. We assume that the mixture of these sounds is additive $x(t) = x_r(t) + x_w(t)$. The input magnitude spectrogram $\mathbf{X}$ of the mixture can be represented as $\mathbf{X} = \mathbf{X}_R + \mathbf{X}_W$, being $\mathbf{X}_R$ the magnitude spectrogram of only respiratory sounds and $\mathbf{X}_W$ the magnitude spectrogram of only wheeze sounds. Each magnitude spectrogram, composed of T frames and F frequency bins, has been computed using the magnitude of the Short-Time Fourier Transform (STFT) using a Hamming window of size N with 25% overlap.

In order to make the proposed model independent of the size and scale of the input spectrogram $\mathbf{X}$, a normalization process is required. Thus, the normalized magnitude spectrogram $\mathbf{X}_n$ is computed as follows,

$$\mathbf{X}_n = \frac{\mathbf{X}}{\left(\frac{\sum_{f=1}^{F}\sum_{t=1}^{T} X_{f,t}}{FT}\right)} \tag{8}$$

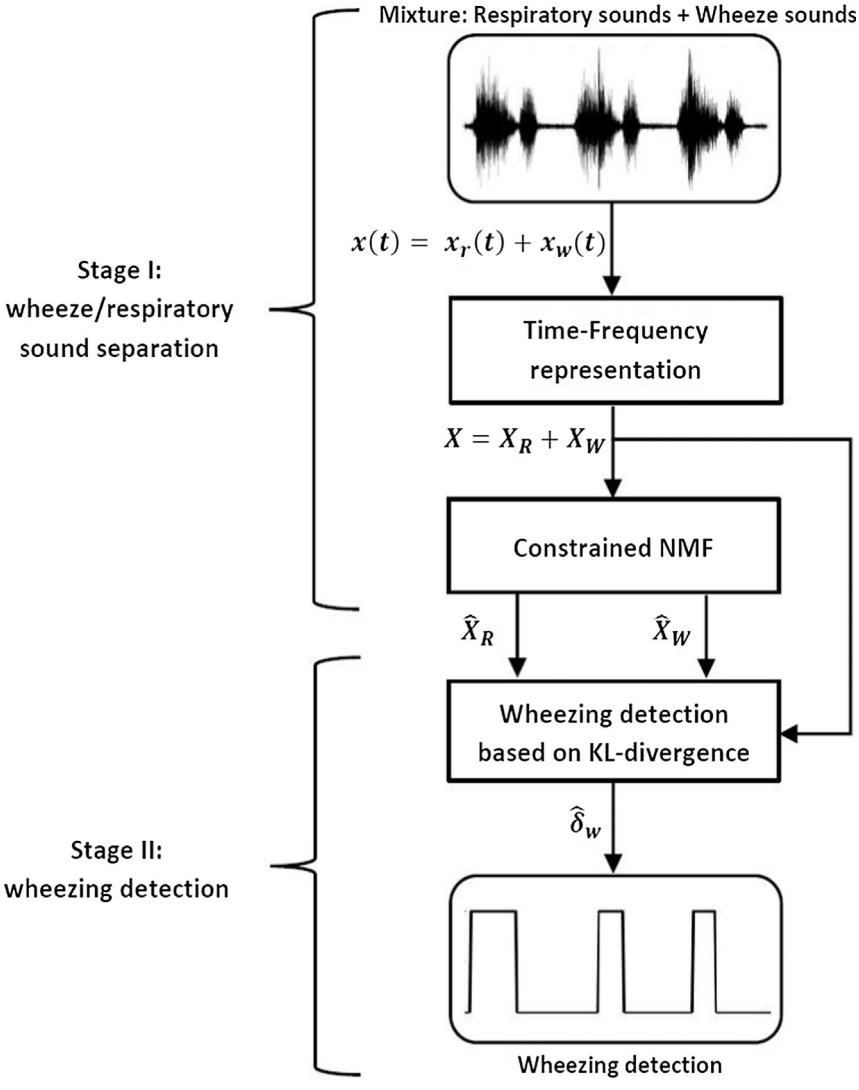

Fig. 2. Block-scheme of the proposed method.

The goal of this stage is to separate wheeze sounds and respiratory sounds. For this purpose, an objective function is defined to decompose a mixture normalized spectrogram $\mathbf{X}_n$ into two separated or estimated spectrograms, $\widehat{\mathbf{X}}_R$ (only respiratory sounds spectrogram) and $\widehat{\mathbf{X}}_W$ (only wheeze sounds spectrogram). The factorization of each spectrogram depends on the estimated basis matrix $\hat{\mathbf{B}}$ (frequency characteristics) and the estimated activation matrix $\hat{\mathbf{A}}$ (temporal characteristics).

$$\mathbf{X}_n = \mathbf{X}_R + \mathbf{X}_W = \mathbf{B}_R\mathbf{A}_R + \mathbf{B}_W\mathbf{A}_W \approx \hat{\mathbf{X}}_n = \hat{\mathbf{X}}_R + \hat{\mathbf{X}}_W = \hat{\mathbf{B}}_R\hat{\mathbf{A}}_R + \hat{\mathbf{B}}_W\hat{\mathbf{A}}_W \tag{9}$$

where $\mathbf{X}_R, \mathbf{X}_W$ are the original spectrograms of the respiratory and wheeze sounds; $\mathbf{B}_R, \mathbf{A}_R$ are the original basis and activation matrix of the respiratory sounds; $\mathbf{B}_W, \mathbf{A}_W$ are the original basis and activation matrix of the wheeze sounds; $\widehat{\mathbf{X}}_n$ is the estimated or reconstructed normalized spectrogram of the mixture; $\hat{\mathbf{B}}_R, \hat{\mathbf{A}}_R$ are the estimated basis and activation matrix of the respiratory sounds; $\hat{\mathbf{B}}_W, \hat{\mathbf{A}}_W$ are the estimated basis and activation matrix of the wheeze sounds. The number of respiratory and wheezing components will be denoted as K_r and K_w, respectively. All of these matrices are non-negative.

The goal of the NMF is to estimate the basis $\left(\hat{\mathbf{B}}_R, \hat{\mathbf{B}}_W\right)$ and activation matrices $\left(\hat{\mathbf{A}}_R, \hat{\mathbf{A}}_W\right)$ minimizing the reconstruction error between the input spectrogram $\mathbf{X}_n$ and the estimated spectrogram $\hat{\mathbf{X}}_n$. In this work, we propose to minimize the Kullback–Leibler divergence cost function $D_{KL}\left(\mathbf{X}_n|\hat{\mathbf{X}}_n\right)$ (see Eq. (10)), because it has been successfully applied in audio signal processing [45,49].

$$D_{KL}\left(\mathbf{X}_n|\hat{\mathbf{X}}_n\right) = \sum_{f=1}^{F}\sum_{t=1}^{T} X_{n_{f,t}} \log\frac{X_{n_{f,t}}}{\hat{X}_{n_{f,t}}} - X_{n_{f,t}} + \hat{X}_{n_{f,t}} \tag{10}$$

As previously mentioned, NMF can only reconstruct the spectrogram of the input mixture but ensuring the convergence of the function to local minima. However, these local minima cannot discriminate between wheeze and respiratory bases so, the wheeze/respiratory sound separation is not successfully performed. In order to find a better NMF decomposition that shows spectro-temporal features of the WS and RS as can be observed in real life, we propose to incorporate constraints, sparseness and smoothness, into the NMF decomposition. Thus, we assume that RS can be considered smooth in time (slow variation of the magnitude spectrogram along time) and frequency (wideband spectrum). However, WS can be considered sparse in frequency because monophonic WS or polyphonic WS is characterized by one or more than one narrowband spectral peaks.

The global objective function $D\left(\mathbf{X}_n|\hat{\mathbf{X}}_n\right)$ that must be minimized taking into account the signal reconstruction, based on the Kullback-Leibler divergence, and the smoothness and sparseness constraints is detailed as follows,

$$D\left(\mathbf{X}_n|\hat{\mathbf{X}}_n\right) = D_{KL}\left(\mathbf{X}_n|\hat{\mathbf{X}}_n\right) + \lambda_B\phi\left(\hat{\mathbf{B}}_R\right) + \lambda_A\phi\left(\hat{\mathbf{A}}_R\right) + \alpha_B\psi\left(\hat{\mathbf{B}}_W\right) \tag{11}$$

where λ_B defines the weight of the spectral smoothness $\phi\left(\hat{\mathbf{B}}_R\right)$ applied to only estimated respiratory basis matrix $\hat{\mathbf{B}}_R$, λ_A represents the weight of the temporal smoothness $\phi\left(\hat{\mathbf{A}}_R\right)$ applied to only estimated respiratory activation matrix $\hat{\mathbf{A}}_R$ and α_B is the weight of the spectral sparseness $\psi\left(\hat{\mathbf{B}}_R\right)$ applied to only estimated wheezing basis matrix $\hat{\mathbf{B}}_W$.

The estimated respiratory basis matrix $\hat{\mathbf{B}}_R$ (see Eq. (12)) and the estimated respiratory activation matrix $\hat{\mathbf{A}}_R$ (see Eq. (13)) can be obtained by applying a gradient descent algorithm [45] based on multiplicative update rules (see Eq. (3)). The equations of each term of the respiratory multiplicative update rules can be found in the Appendix A.

$$\hat{\mathbf{B}}_R = \hat{\mathbf{B}}_R \odot \frac{\left[\frac{\partial D_{KL}\left(\mathbf{X}_n|\hat{\mathbf{X}}_n\right)}{\partial \hat{\mathbf{B}}_R}\right]^- + \lambda_B\left[\frac{\partial\phi\left(\hat{\mathbf{B}}_R\right)}{\partial \hat{\mathbf{B}}_R}\right]^-}{\left[\frac{\partial D_{KL}\left(\mathbf{X}_n|\hat{\mathbf{X}}_n\right)}{\partial \hat{\mathbf{B}}_R}\right]^+ + \lambda_B\left[\frac{\partial\phi\left(\hat{\mathbf{B}}_R\right)}{\partial \hat{\mathbf{B}}_R}\right]^+} \tag{12}$$

$$\hat{\mathbf{A}}_R = \hat{\mathbf{A}}_R \odot \frac{\left[\frac{\partial D_{KL}\left(\mathbf{X}_n|\hat{\mathbf{X}}_n\right)}{\partial \hat{\mathbf{A}}_R}\right]^- + \lambda_A\left[\frac{\partial\phi\left(\hat{\mathbf{A}}_R\right)}{\partial \hat{\mathbf{A}}_R}\right]^-}{\left[\frac{\partial D_{KL}\left(\mathbf{X}_n|\hat{\mathbf{X}}_n\right)}{\partial \hat{\mathbf{A}}_R}\right]^+ + \lambda_A\left[\frac{\partial\phi\left(\hat{\mathbf{A}}_R\right)}{\partial \hat{\mathbf{A}}_R}\right]^+} \tag{13}$$

The estimated wheezing basis matrix $\hat{\mathbf{B}}_W$ (see Eq. (14)) and the estimated wheezing activation matrix $\hat{\mathbf{A}}_W$ (see Eq. (15)) can be obtained by applying a gradient descent algorithm [45] based on multiplicative update rules (see Eq. (3)). The equations of each term of the wheezing multiplicative update rules can be found in the Appendix A.

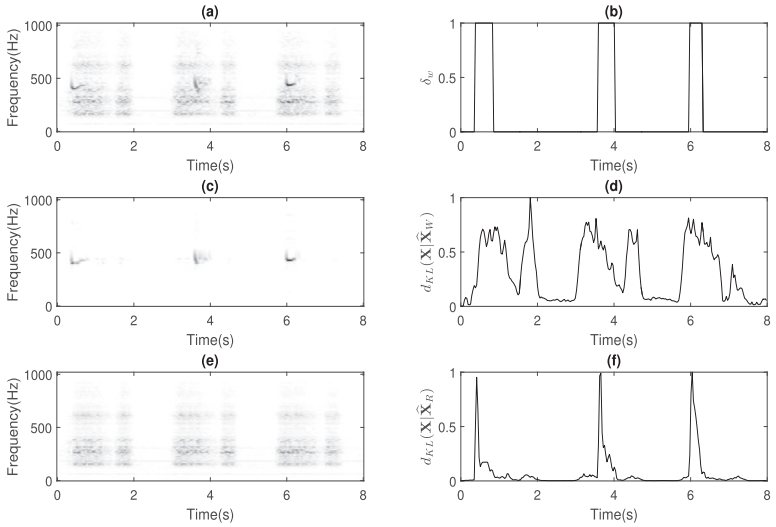

Fig. 3. **(a)** Magnitude spectrogram $\mathbf{X}$ of a mixture $x(t)$. **(b)** The ideal wheezing detection δ_w. **(c)** The estimated wheezing spectrogram $\hat{\mathbf{X}}_W$. **(d)** $d_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}_W\right)$. **(e)** The estimated respiratory spectrogram $\hat{\mathbf{X}}_R$. **(f)** $d_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}_R\right)$. Note that $d_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}_W\right)$ and $d_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}_R\right)$ have been normalized to adjust the values between 0 and 1.

$$\hat{\mathbf{B}}_W = \hat{\mathbf{B}}_W \odot \frac{\left[\frac{\partial D_{KL}(\mathbf{X}_n|\hat{\mathbf{X}}_n)}{\partial \hat{\mathbf{B}}_W}\right]^- + \alpha_B \left[\frac{\partial \psi(\hat{\mathbf{B}}_W)}{\partial \hat{\mathbf{B}}_W}\right]^-}{\left[\frac{\partial D_{KL}(\mathbf{X}_n|\hat{\mathbf{X}}_n)}{\partial \hat{\mathbf{B}}_W}\right]^+ + \alpha_B \left[\frac{\partial \psi(\hat{\mathbf{B}}_W)}{\partial \hat{\mathbf{B}}_W}\right]^+} \tag{14}$$

$$\hat{\mathbf{A}}_W = \hat{\mathbf{A}}_W \odot \frac{\left[\frac{\partial D_{KL}(\mathbf{X}_n|\hat{\mathbf{X}}_n)}{\partial \hat{\mathbf{A}}_W}\right]^-}{\left[\frac{\partial D_{KL}(\mathbf{X}_n|\hat{\mathbf{X}}_n)}{\partial \hat{\mathbf{A}}_W}\right]^+} \tag{15}$$

The estimated respiratory and wheezing basis and activation matrices are obtained updating the rules until the algorithm converges or reaches a maximum number of iterations M_{iter}. Next, the estimated magnitude spectrograms $\hat{\mathbf{X}}_R$ and $\hat{\mathbf{X}}_W$ can be obtained from the estimated basis and activation matrices (see Eq. (16) and (17)). Finally, the estimated magnitude spectrograms $\hat{\mathbf{X}}_R$ and $\hat{\mathbf{X}}_W$ are denormalized multiplying by the denominator of Eq. (8). The wheeze/respiratory sound separation procedure is summarized in Algorithm 1.

$$\hat{\mathbf{X}}_R = \hat{\mathbf{B}}_R \hat{\mathbf{A}}_R \tag{16}$$

$$\hat{\mathbf{X}}_W = \hat{\mathbf{B}}_W \hat{\mathbf{A}}_W \tag{17}$$

Algorithm 1 Stage I: wheeze/respiratory sound separation procedure

Require: $x(t), K_r, K_w, \lambda_B, \lambda_A, \alpha_B$ and M_{iter}.

1 Compute the normalized magnitude spectrogram $\mathbf{X}_n$ of the mixture signal $x(t)$ using Eq. (8).

2 Initialize $\hat{\mathbf{B}}_R, \hat{\mathbf{A}}_R, \hat{\mathbf{B}}_W$ and $\hat{\mathbf{A}}_W$ with random nonnegative values.

3 Update the estimated respiratory basis matrix $\hat{\mathbf{B}}_R$ using Eq. (12).

4 Update the estimated respiratory activation matrix $\hat{\mathbf{A}}_R$ using Eq. (13).

5 Update the estimated wheezing basis matrix $\hat{\mathbf{B}}_W$ using Eq. (14).

6 Update the estimated wheezing activation matrix $\hat{\mathbf{A}}_W$ using Eq. (15).

7 Repeat steps 3–6 until the algorithm converges (or until the maximum number of iterations M_{iter} is reached).

8 Compute magnitude estimated spectrograms $\hat{\mathbf{X}}_R$ and $\hat{\mathbf{X}}_W$ using Eq. (16) and (17).

9 Denormalize magnitude estimated spectrograms $\hat{\mathbf{X}}_R$ and $\hat{\mathbf{X}}_W$ multiplying by the denominator of Eq. (8).

return $\hat{\mathbf{X}}_R$ and $\hat{\mathbf{X}}_W$

In order to optimize the separation stage (subSection 4.5), the estimated respiratory signal $\hat{x}_r(t)$ and the estimated wheezing signal $\hat{x}_w(t)$ are synthesized using the estimated magnitude spectrograms $\hat{\mathbf{X}}_R$ and $\hat{\mathbf{X}}_W$ with a Wiener filtering [50] and the inverse STFT. Indicate that the Wiener filtering ensures that the reconstruction process is conservative by means of the use of a respiratory $\mathbf{M}_R$ and wheezing $\mathbf{M}_W$ mask. The Wiener masks $\mathbf{M}_R$ ans $\mathbf{M}_W$ represent the relative energy contribution of each source with respect to the energy of the input mixture. These masks are defined as,

$$\mathbf{M}_R = \frac{\left|\hat{\mathbf{X}}_R\right|^2}{\left(\left|\hat{\mathbf{X}}_R\right|^2 + \left|\hat{\mathbf{X}}_W\right|^2\right)} \tag{18}$$

$$\mathbf{M}_W = \frac{\left|\hat{\mathbf{X}}_W\right|^2}{\left(\left|\hat{\mathbf{X}}_R\right|^2 + \left|\hat{\mathbf{X}}_W\right|^2\right)} \tag{19}$$

In order to obtain the estimated complex spectrograms of the separated signals, each mask is multiplied by the complex spectrogram $\mathbf{X}_c$ of the mixture $x(t)$ as follows,

$$\hat{\mathbf{X}}_R = \mathbf{M}_R \odot \mathbf{X}_c \tag{20}$$

$$\hat{\mathbf{X}}_W = \mathbf{M}_W \odot \mathbf{X}_c \tag{21}$$

Finally, the inverse overlap-add STFT is applied to synthesize the estimated respiratory signal $\hat{x}_r(t)$ and the estimated wheezing signal $\hat{x}_w(t)$ in time domain as follows,

$$\hat{x}_r(t) = IDFT\left(\hat{\mathbf{X}}_R\right) \tag{22}$$

$$\hat{x}_w(t) = IDFT\left(\hat{\mathbf{X}}_W\right) \tag{23}$$

3.2. Stage II. Wheezing detection

The goal of this stage is to determine the presence of WS and the temporal intervals or areas in which wheezing is active from the estimated respiratory $\hat{\mathbf{X}}_R$ and wheezing $\hat{\mathbf{X}}_W$ spectrograms obtained in the stage I. Thus, the estimated wheezing detection $\hat{\delta}_w$ is defined (frame-by-frame) to assign the value 1 to the temporal frames where wheezing is active and the value 0 when the wheezing is inactive.

$$\hat{\delta}_w(t) = \begin{cases} 1 & \text{frame with wheezing active} \\ 0 & \text{frame with wheezing inactive} \end{cases} \tag{24}$$

where $t = 1,\ldots T$, being T the number of frames.

Initially, the proposed method performs a preliminary power analysis of the estimated wheezing spectrogram $\widehat{P}_W = |\hat{\mathbf{X}}_W|^2$ to classify whether the input spectrogram $\mathbf{X}$ corresponds to a healthy or unhealthy patient.

In the case of healthy patients, $\widehat{P}_W$ will be equal to 0 because the constrained NMF does not find wheeze bases that can be factorized using the spectral sparseness constraint. Therefore, the output of the wheezing detection stage assigns $\hat{\delta}_w = 0$ to all frames.

In the case of unhealthy patients, $\widehat{P}_W$ will be higher to 0. In this scenario, we propose the use of the Kullback-Leibler divergence $d_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}_R\right)$ frame-by-frame between input spectrogram $\mathbf{X}$ and estimated respiratory spectrogram $\hat{\mathbf{X}}_R$ to discriminate wheezing areas as can be observed in Eq. (25),

$$d_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}_R\right)_t = \sum_{f=1}^{F} X_{f,t} \log \frac{X_{f,t}}{\widehat{X}_{R_{f,t}}} - X_{f,t} + \widehat{X}_{R_{f,t}}, \tag{25}$$

where $t = 1,\ldots T$, being T the number of frames.

The Kullback-Leibler divergence has an interesting property. It is in particular scale-invariant, meaning that low energy components of $\mathbf{X}$ bear the same relative importance as high energy ones. So, although the separation stage provides satisfactory results both in the estimated wheezing (Fig. 2 and respiratory (Fig. 3e) spectrograms, it must be highlighted that the RS are the only sounds of the mixture that is not overlapped in some areas (particularly, in the areas of the respiratory cycle where wheezing is inactive), as can be seen in Fig. 3a. For this reason, $d_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}_R\right)$ is the appropriate way to discriminate the wheezing areas from respiratory areas.

As a result, $d_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}_R\right)$ will have a very small value in areas where only the RS are active. However, in the areas where RS and WS sounds are active, $d_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}_R\right)$ provides a high value (see Fig. 3f).

On the other hand, $d_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}_W\right)$ does not allow to discriminate the wheezing areas, because the WS are always overlapping in the mixture with the RS. For this reason $d_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}_W\right)$ tends to obtain high values, except in the breathing silences (when the airflow is inactive), as can be observed in Fig. 3d.

From $d_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}_R\right)$, a threshold is defined to determine the wheezing areas. In this paper, we have used the Otsu algorithm [51] to define the threshold ζ_{otsu}. As shown in Fig. 4, the Otsu algorithm allows to define an optimal threshold ζ_{otsu} that clearly differentiates two groups in a histogram, in this case, wheezing and respiratory (non-wheezing) areas. Finally, each frame of the unhealthy mixture $\mathbf{X}$ is labelled as wheezing (wheezing active) or not (wheezing inactive) by means of $\hat{\delta}_w$, combining $d_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}_R\right)$ and the threshold ζ_{otsu} as follows,

$$\hat{\delta}_w(t) = \begin{cases} 1 & \text{if } d_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}_R\right) \geqslant \zeta_{otsu} \\ 0 & \text{if } d_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}_R\right) < \zeta_{otsu} \end{cases} \tag{26}$$

where $t = 1,\ldots T$, being T the number of frames.

The wheezing detection procedure is summarized in Algorithm 2.

Algorithm 2 Stage II: wheezing detection procedure

Require: $\mathbf{X}, \widehat{\mathbf{X}}_W$ and $\widehat{\mathbf{X}}_R$.

Compute the power of the estimated wheezing spectrogram $\widehat{P}_W$.

———————— **Healthy patients** ————————

if $\widehat{P}_W = 0$ **then**

return $\hat{\delta}_w = 0, \forall t$.

end if

———————— **Unhealthy patients** ————————

if $\widehat{P}_W \neq 0$ **then**

1 Compute KL-divergence d_{KL} between input spectrogram $\mathbf{X}$ and estimated respiratory spectrogram $\widehat{\mathbf{X}}_R$ using Eq. (25).

2 Compute histogram of $d_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}_R\right)$.

3 Compute threshold ζ_{otsu}.

4 Combine $d_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}_R\right)$ and ζ_{otsu} to obtain the estimated wheezing detection $\hat{\delta}_w$ using Eq. (26).

return $\hat{\delta}_w$

end if

4. Experimental results

In this section, we asses the potential of the proposed method in the field of wheezing detection applied to mono-channel audio mixtures. As can be seen, an optimization process will be applied to obtain the optimal parameters of its separation stage. Finally, the performance of the wheezing detection of the proposed method is evaluated taking into account other state-of-the-art methods.

4.1. Dataset

Three datasets have been used in the experiments of the proposed method. The dataset E1 has been used in the optimization of the separation stage. The datasets T1 and T2 have been used in the wheezing detection testing. It must be highlighted that the optimization dataset is not a part of the test datasets (T1 and T2) in order to validate the results. The datasets E1, T1 and T2 can be seen in Table 1.

To optimize the parameters of the separation stage, the dataset E1 have been created mixing only WS manually separated (by means of a time–frequency mask applied to the mixture spectrogram to select only the bins of each frame corresponding to wheezing) and only RS (in which wheezing is inactive) obtained from widely used Internet pulmonary data sources [29,52–59]. This dataset is composed of 48 mixtures which show a signal-to-noise ratio (SNR) between 0 dB and 9 dB, with duration between 5 and 24 s, with a total of 92 wheezing and 154 respiratory cycles.

The dataset T1 is the same dataset evaluated in [29] which has been directly shared from its authors. This dataset is composed of 16 recordings of healthy and unhealthy patients, with duration between 4 and 51 s. Specifically, the first part of the dataset consisted of 8 recordings from unhealthy patients, containing a total of 36 intervals of intermittent wheezing dispersed in 31 respiratory cycles. The second part of the dataset consisted of 8 recording from healthy patients, composed of 40 respiratory cycles of normal respiratory sounds.

In order to evaluate the robustness of the proposed detection method, we have created three datasets T2H (SNR = 5 dB), T2M

Table 1
Characteristics of datasets.

Identifier	Type of patients	SNR(dB)
	Optimization dataset	
E1	Unhealthy	[0–9]
	Test datasets	
T1	Unhealthy/Healthy	[2–8]
T2H	Unhealthy	5
T2M	Unhealthy	0
T2L	Unhealthy	−5

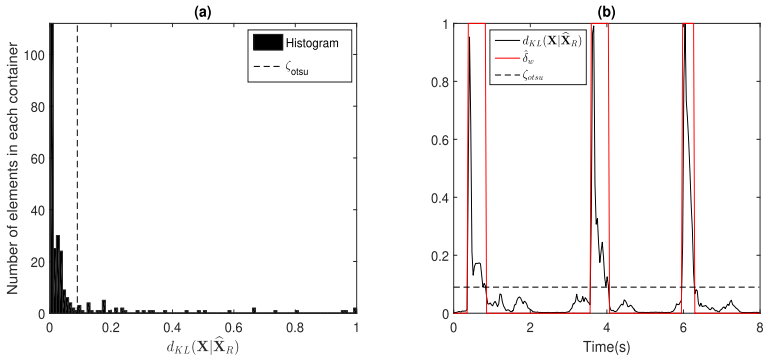

Fig. 4. **(a)** Histogram applied to $d_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}_R\right)$ and threshold ζ_{otsu}. **(b)** Estimated wheezing detection $\hat{\delta}_w$ using $d_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}_R\right)$ and ζ_{otsu}. Note that $d_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}_R\right)$ has been normalized to adjust the values between 0 and 1.

(SNR = 0 dB) and T2L (SNR = −5 dB) with different signal-to-noise ratio (SNR) from the dataset T2 as can be seen in Table 1. The dataset T2 has been generated using a similar procedure and the same Internet data sources used in the creation of the dataset E1. However, the mixtures created in this datasets are not the same as that used in dataset E1 to provide valid results. Specifically, the datasets T2H, T2M and T2L are composed of 16 mixtures with duration between 7 and 22 s, with a total of 41 wheezing and 63 respiratory cycles.

4.2. Initializations

All mixtures were band-limited from 100 Hz to 1000 Hz (as previously mentioned, we assume that wheezes are not active below 100 Hz and above 1000 Hz). The other processing parameters are the following ones: sampling rate $fs = 2048Hz$, size of Hamming window $N = 256$ samples, with 25% overlap (temporal resolution of 31.3 ms). In addition, the convergence was empirically achieved after 120 iterations in all proposed NMF factorization. For this reason, the parameter $M_{iter} = 120$ has been used in this work.

4.3. Metrics

4.3.1. Separation

Three metrics are used to optimize the parameters of the separation stage [60,61], which are widely used in the field of sound source separation [43,46,49]: (1) the source-to-distortion ratio (SDR), which provides information on the overall quality of the separation process; (2) the source-to-interferences ratio (SIR), which is a measure of the presence of WS in the respiratory signal and vice versa; and (3) the source-to-artifacts ratio (SAR), which provides information on the artifacts in the separated signal from separation and/or resynthesis. In this paper, the SDR, SIR and SAR metrics have been calculated using the estimated wheezing signal $\hat{x}_w(t)$ obtained in the stage I.

4.3.2. Detection

Three metrics are used to evaluate the performance of the proposed method for the wheezing detection [62], which are widely used in the field of wheezing detection [27–29]: (1) sensitivity (SE), the probability of detecting wheezing frames correctly; (2) specificity (SP), the probability of detecting respiratory (without wheezing) frames correctly; and (3) accuracy (ACC), the probability of detecting wheezing/respiratory frames correctly.

4.4. Algorithms for comparison of wheezing detection results

We have used three recent state-of-the-art wheezing detection methods to evaluate the proposed method: HMMFL [29], TSVM [26] and MKNN [28]. The methods HMMFL, TSVM and MKNN have been implemented in this study, whereas the wheezing detection results provided by HMMFL (applied to the dataset T1) have been directly taken from [29]. The training dataset used by TSVM and MKNN is the same dataset E1 used in the separation optimization by the proposed method.

4.5. Separation optimization

An initial study showed that the detection results depend directly on the separation results. Specifically, the detection results increased considerably when the separation metrics improve. Therefore, an optimization process was motivated by these results to obtain the optimal separation parameters ($K_r, K_w, \lambda_B, \lambda_A$ and α_B) in order to maximize the detection performance of the proposed method.

Preliminary results of the optimization process indicated no significant SDR differences (lower than 0.3 dB) evaluating different number of wheezing K_w and respiratory K_r components, specifically, from 50 to 250 components, but the number of wheezing and respiratory components must be greater than 50 components in order to correctly model the spectral diversity of the wheezing and respiratory spectral patterns. In this work, $K_w = K_r = 150$ have been selected because preliminary results maximized the separation performance, in SDR, SIR and SAR, using this size of components.

Fig. 5 shows the optimization of the parameters λ_B, λ_A and α_B in order to analyse the effect of the smoothness and sparseness constraints in the optimization dataset E1. NMF (specifically, unconstrained NMF in which smoothness and sparseness constraints are disable, i.e., $\lambda_B = \lambda_A = \alpha_B = 0$) achieves a SDR average value of the estimated wheezing signal approximately equal to 2 dB (see Fig. 5a). This fact indicates that NMF does not properly separate WS and RS because the factorization model is not able to correctly discriminate between wheeze and respiratory bases. Fig. 5a confirms the previous claim since the minimum SDR values are reached when the sparseness constraint are disable ($\alpha_B = 0$). For this reason, the sparseness constraint can be considered crucial since it plays a fundamental role in the wheezing spectral model.

Results show that the maximum SDR value is obtained by enabling the smoothness and sparseness constraints simultaneously because these constraints are capable of modeling typical spectral and temporal features observed in RS and WS as can be found in real life. It can be observed that the maximum SDR value, approximately equal to 14 dB in Fig. 5c, is provided by the proposed method enabling the three proposed constraints, that is, $\lambda_B = 0.5, \lambda_A = 1$ and $\alpha_B = 3$ respectively. Highlight that the constrained NMF that uses the proposed constraints, smoothness and sparseness, produces a significant SDR improvement of approximately 12 dB compared to the unconstrained NMF (in which all proposed constraints are inactive). In this manner, the constrained NMF provides that the estimated wheezing signal exhibit common spectro-temporal features shown by WS, attenuating RS and vice versa considering the estimated respiratory signal. As a consequence, the factorized spectrograms $\hat{\mathbf{X}}_W$ and $\hat{\mathbf{X}}_R$ exhibit time-frequency energy distributions in a more accurate way than it can be found in real-world WS or RS. Besides, Fig. 5 shows that the SDR value, independently of α_B, drops when the values of λ_B and λ_A are very low or very high. On the one hand, when λ_B and λ_A values are very low, the signal reconstruction is performed without adding the common wheezing and respiratory spectro-temporal features to the bases/activations obtained from NMF. On the other hand, when λ_B and λ_A values are very high, the separation performance is not appropriate because it does not allow that NMF prioritizes the reconstruction of the WS and RS with the help of the previous constraints.

As shown in Table 2, in order to analyze the importance of each constraint $\phi\left(\hat{\mathbf{B}}_R\right), \phi\left(\hat{\mathbf{A}}_R\right)$ and $\psi\left(\hat{\mathbf{B}}_W\right)$ in the proposed separation method, eight possible parameters setups (enabling or disabling smoothness and sparseness constraints) have been defined. For each setup, the values of the weights λ_B, λ_A and α_B that maximize the optimization results have been selected.

Fig. 6 shows SDR, SIR and SAR results of the estimated wheezing signal for each parameters setup defined in Table 2. Each box represents 48 data points, one for each signal of the test dataset E1: each blue box represents the analysis for SDR values; each red box represents the analysis for SIR values; and each black box represents the analysis for SAR values. The lower and upper lines of each box show the 25th and 75th percentiles for the dataset E1. The line in the middle of each box represents the median value of the dataset E1. The lines extending above and below each box

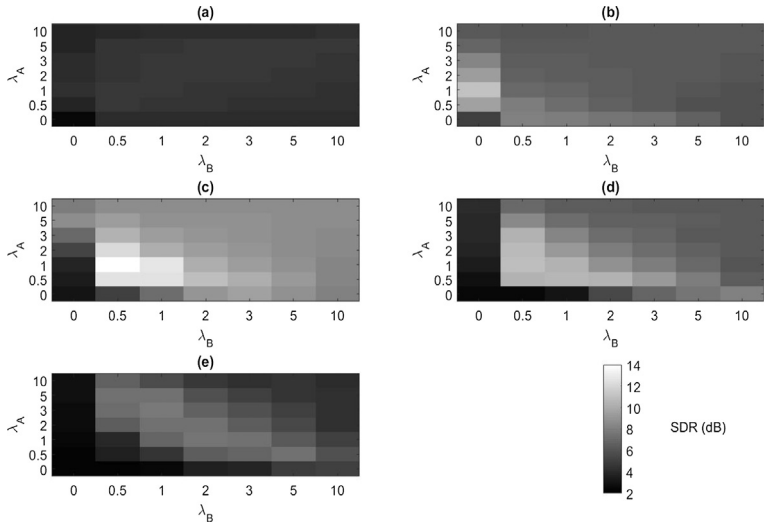

Fig. 5. SDR average results of the estimated wheezing signal provided by the hyperparametric analysis applied to the parameters optimization λ_B, λ_A and α_B in the dataset E1. **(a)** $\alpha_B = 0$, **(b)** $\alpha_B = 1$, **(c)** $\alpha_B = 3$, **(d)** $\alpha_B = 5$ and **(e)** $\alpha_B = 10$. The number of wheezing K_w and respiratory K_r components is equal to 150.

Table 2
Parameters setups in order to analyze the importance of the constraints $\phi\left(\hat{\mathbf{B}}_R\right)$, $\phi\left(\hat{\mathbf{A}}_R\right)$ and $\psi\left(\hat{\mathbf{B}}_W\right)$. the symbol – indicates constraint disabled. The symbol ✔indicates constraint enabled.

Identifier	$\phi\left(\hat{B}_R\right)$	$\phi\left(\hat{A}_R\right)$	$\psi\left(\hat{B}_W\right)$	λ_B	λ_A	α_B
S1	–	–	–	0	0	0
S2	✔	–	–	0.5	0	0
S3	–	✔	–	0	1	0
S4	–	–	✔	0	0	1
S5	✔	✔	–	1	1	0
S6	✔	–	✔	3	0	3
S7	–	✔	✔	0	1	1
S8	✔	✔	✔	0.5	1	3

show the extent of the rest of the samples, excluding outliers. Outliers are defined as points that are over 1.5 times the interquartile range from the sample median, which are shown as crosses. Results show, as mentioned above, that the worst results are obtained when all constraints are disabled (Fig. 6: S1). However, a significant improvement is obtained (between 1.5–2.9 dB in SDR, 2.2–6 dB in SIR and 0.5–1.4 dB in SAR) when one of the used constraint is enabled (Fig. 6: S2, S3 and S4). This fact indicates that the smoothness and sparseness constraints makes it possible to efficiently model the common spectro-temporal behavior of the sounds active in the mixture (WS and RS). Results suggest that the sparseness constraint is the constraint that seems to be more significant in the separation performance of the proposed method comparing the cases when only one constraint is active (S2, S3 or S4). Specifically, the setup S4 produces a significant improvement of 2.9 dB in SDR, 6 dB in SIR and 1.4 dB in SAR over the unconstrained NMF. In addition, this setup exceeds the optimization results in the case of enabling the two smoothness constrains S5 that characterize the RS. This seems to indicate that characterizing the WS, using the sparseness constraint, is more important than characterizing RS by means of smoothness constraints to separate both sounds in the proposed method. The best separation results are obtained when the smoothness and sparseness constraints are enabled in the setups S6, S7 and S8. Comparing the setups S7 and S8 in which the spectral sparseness constraint is enabled, results imply that the temporal smoothness constraint S7 is more significant that the spectral smoothness constraint S6. Finally, the optimal separation performance is achieved when all the constraints are enabled in S8. Thus, S8 produces a significant improvement of 11.9 dB in SDR, 14.4 dB in SIR and 7.6 dB in SAR over the unconstrained NMF. As can be observed, Fig. 6 shows the evolution

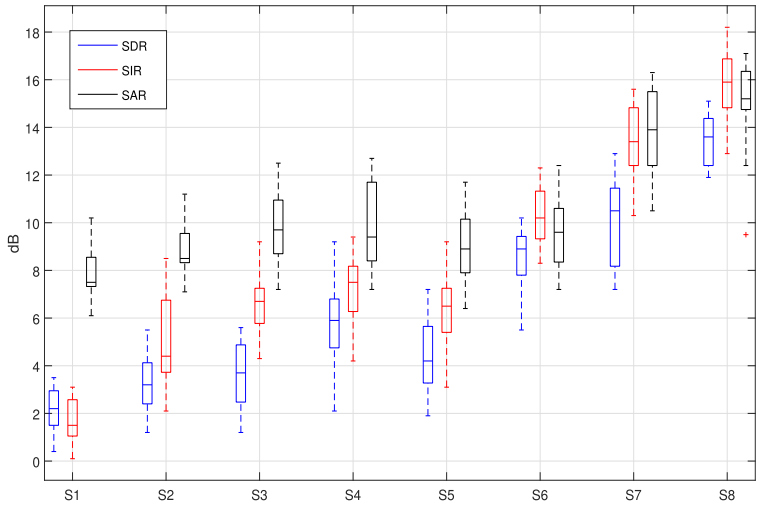

Fig. 6. Optimization results for the different parameters setups defined in Table 2.

of the optimization process from the worst results, obtained using the unconstrained NMF, until the best results achieved enabling all the proposed constraints. This fact confirms that the smoothness and sparseness constraints, used in the proposed method, are able to correctly model the WS and RS, incorporating physical meaning as can be found in nature and real life. Although SDR, SIR and SAR results exhibit an upward behavior, SIR results tend to overcome SDR results. Moreover, SAR results are often higher for all parameter configurations evaluated, indicating that the proposed method avoids introducing artifacts in the separation stage.

Finally, the optimal parameters that maximize the separation results have been used as a starting point of the detection stage to evaluate the wheezing detection performance of the proposed method: number of wheezing K_w and respiratory K_r components equal to 150; weight of the spectral smoothness $\lambda_B = 0.5$; weight of the temporal smoothness $\lambda_A = 1$; and weight of the spectral sparseness $\alpha_B = 3$.

4.6. Detection results

Table 3 shows sensitivity (SE), specificity (SP) and accuracy (ACC) results evaluating the dataset T1 between the proposed method and the previous state-of-the-art methods. Results report that the proposed method is competitive compared to the other evaluated methods. Specifically, the proposed method obtains the best SE and ACC results, 95.71% and 95.86%, respectively. Focusing on the SE metric, the proposed method achieves a significant improvement of approximately 6.4%, 10.4% and 14.85% compared to HMMFL, TSVM and MKNN. This fact indicates that the proposed method is more effective in correctly detecting wheezing frames compared to the state-of-the-art methods evaluated. However, the proposed method obtains the lowest SP result compared to the other methods. This fact suggests that the proposed method tends to provide a higher number of false positives (respiratory frames mistakenly detected as wheezing frames) in order to detect the whole wheezing temporal interval. Moreover, HMMFL obtains a better detection performance compared to TSVM taking into account SE, SP and ACC but TSVM outperforms the detection performance compared to MKNN. Highlight that the proposed method correctly detects all recordings of the dataset T1 corresponding to healthy patients. This fact confirms the reliability of the proposed method to discriminate healthy and unhealthy patients.

Table 4 shows sensitivity (SE), specificity (SP) and accuracy (ACC) results in order to evaluate the wheezing detection robustness of the proposed method and the state-of-the-art methods evaluated using three different SNR datasets (T2H, T2M and T2L). Evaluation indicates that the proposed method provides the best overall detection results compared to the other evaluated methods considering all SNR scenarios evaluated. Specifically, the proposed method outperforms the second best state-of-the-art method (TSVM), in terms of SE and ACC, about 9% and 3% in the dataset T2H, 9% and 4% in the dataset T2M and 14% and 8% in the dataset T2L. However, TSVM method obtains the highest SP results compared to the other methods. It can be observed that SE, SP and ACC detection results of the proposed method decreases an average of approximately 2.2% comparing T2H vs T2M, 1.0% considering T2M vs T2L and 3% considering T2H vs T2L. However, SE, SP and ACC detection results of the rest of the state-of-the-art methods decrease faster when the signal-to-noise ratio drops. Comparing the evaluated datasets with lower (T2H) and higher (T2L) noise: i) the SE reduction of the detection performance is about 3.9% (the proposed method), 8.8% (HMMFL), 8.5% (TSVM) and 8.6% (MKNN); ii) the SP reduction of the detection performance is about 2.8% (the proposed method), 9.7% (HMMFL), 5.6% (TSVM) and 5.7% (MKNN); iii) the ACC reduction of the detection performance is about 2.7% (the proposed method), 8.2% (HMMFL), 8.2% (TSVM) and 6.1% (MKNN). Results demonstrate the higher robustness of the proposed method in noisy environments compared to the other evaluated methods. This robustness shown by the proposed method suggests a greater ability in order to detect the presence of weak WS that can be masked by louder RS. Finally, a remarkable advantage of the proposed method is that it does not depend on any training data set due to its unsupervised nature.

Table 5 shows an analysis of the computational cost for each step of the proposed method, from the input of the mixture signal until the wheezing detection is provided. The number of elementary operations (Multiplications and Additions) and the number of elementary mathematical functions (Functions) such as divisions, square root, exponent, and logarithm are modeled. Note that the parameters on which the computational cost depends are the following: N, F, T, K_r, K_w and M_{iter}, previously defined, and the number of histogram bars H. Results indicate the majority of the cost comes from the Wheeze/respiratory sound separation stage, as can be observed in Table 5. In our experiments, $N = 256$ samples, $F = 256$ bins, $K_r = 150$ components, $K_w = 150$ components, $M_{iter} = 120$ iterations, $H = 100$ bars and $T = 32$ frames per second of mixture signal, obtaining a total computational cost of 966.25 millions of multiplications per second, 1233.5 millions of additions per second and 180.42 millions of elementary mathematical functions per second. The computation cost (in seconds) of the proposed method which has been computed using Matlab on a PC with Intel Core i7-7700HQ CPU of 2.8 GHz and 16 GB of RAM is shown in Fig. 7. It can be observed that the computation cost of the proposed method increases with the size of the temporal duration of the input mixture. However, the processing factor P_F defined as the ratio between the computation cost and the temporal duration of the input mixture decreases with the size of the mixture. This fact guarantees that the computation cost of the proposed method is less than the temporal duration of the input mixture. Specifically, the computation cost of the proposed method to detect wheezing is approximately 4.8 s when mixtures of 30 s, are analyzed. Thus, the proposed method adds complementary

Table 3
Wheezing detection comparison between the proposed method and reference state-of-the-art methods [29] [26] [28] evaluating the dataset T1.

Algorithm	SE (%)	SP (%)	ACC (%)
Proposed Method	**95.71**	93.02	**95.86**
HMMFL	89.34	**96.28**	94.91
TSVM	85.32	95.36	90.59
MKNN	80.86	93.27	88.48

Each value in bold indicates the highest value obtained in each column.

Table 4
Robustness wheezing detection performance comparison between the proposed method and reference state-of-the-art methods [29] [26] [28] evaluating the datasets T2H, T2M and T2L.

Algorithm	SE (%)	SP (%)	ACC (%)
Dataset T2H (SNR = 5 dB)			
Proposed Method	**99.48**	90.77	**97.41**
HMMFL	79.5	88.62	80.06
TSVM	90.38	**95.52**	94.66
MKNN	82.33	90.84	88.12
Dataset T2M (SNR = 0 dB)			
Proposed Method	**97.27**	88.60	**95.06**
HMMFL	73.83	85.86	76.86
TSVM	88.60	**92.46**	91.29
MKNN	78.68	87.62	86.94
Dataset T2L (SNR = −5 dB)			
Proposed Method	**95.57**	87.97	**94.70**
HMMFL	70.72	78.94	71.81
TSVM	81.93	**89.96**	86.45
MKNN	73.77	85.12	82.01

Table 5
Computational complexity of each step of the proposed method.

Steps	Multiplications	Additions	Functions
	Stage I: wheeze/respiratory sound separation procedure (Algorithm 1)		
$\mathbf{X}_n$	$T2N\log_2(N)$	$T3N\log_2(N)$	
$\hat{\mathbf{B}}_R$	$M_{iter}(9FK_r + FT(1+K_r) + K_r)$	$M_{iter}(FK_r(7+T))$	$M_{iter}(F(4K_r+T))$
$\hat{\mathbf{A}}_R$	$M_{iter}(9TK_r + FT(1+K_r) + K_r)$	$M_{iter}(TK_r(7+F))$	$M_{iter}(T(4K_r+F))$
$\hat{\mathbf{B}}_W$	$M_{iter}(6FK_w + FT(1+K_w) + 3K_w)$	$M_{iter}(FK_w(4+T))$	$M_{iter}(2K_w(1+F)+FT)$
$\hat{\mathbf{A}}_W$	$M_{iter}(K_wT(1+F)+FT)$	$M_{iter}(K_wT(F-1))$	$M_{iter}(K_wT)$
$\widehat{\mathbf{X}}_n$	$M_{iter}(FT(K_w+K_r))$	$M_{iter}(FT(2K_w+2K_r-3))$	
$\widehat{\mathbf{X}}_R$	FTK_r	FTK_r	
$\widehat{\mathbf{X}}_W$	FTK_w	FTK_w	
	Stage II: wheezing detection procedure (Algorithm 2)		
$\widehat{P}_W$	FT	FT	
$d_{KL}\left(X\vert\widehat{X}_R\right)$	FT	$2FT$	$2FT$
$hist_{d_{KL}}$		HT	
ζ_{otsu}	$2H+H^2$	$2H+H^2$	$H+H^2$
$\hat{\delta}_w$		T	

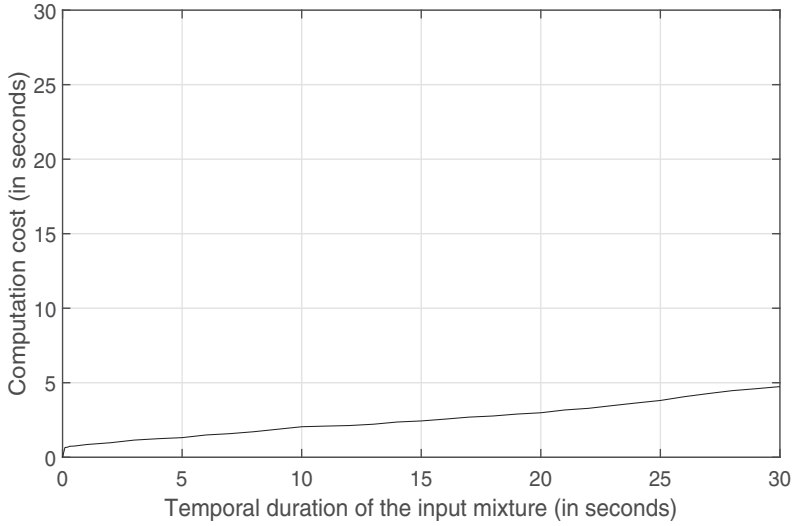

Fig. 7. Computation cost of the proposed method.

information for the physician diagnostic without slowing down the normal course of each physician consultation.

5. Conclusions and future work

In this paper, we propose a novel constrained non-negative matrix factorization approach to detect the presence of wheezing, locating the temporal intervals in which WS are active when they are mixed with RS in mono-channel audio mixtures. As far as the authors knowledge extends, non-negative matrix factorization approach has never been applied before to wheezing detection. The main contribution of the separation stage is to design a NMF framework adding typical spectro-temporal behaviors observed in most WS and RS in real life. The main contribution of the detection stage is the use of the Kullback-Leibler divergence applied to the estimated respiratory spectrogram obtained from constrained NMF to discriminate between wheezing and respiratory temporal intervals (areas).

The most relevant conclusions from the experimental results indicate the following: i) the wheezing detection performance of the proposed method is competitive compared to other state-of-the-art methods; ii) the robustness of the proposal is demonstrated because all detection metrics are reduced by a maximum of 3% comparing three datasets with a SNR difference of 5 dB between them. In this manner, results suggest that the proposed method can be an appropriate tool to be applied in noisy environments; and iii) the proposed method achieves a promising rate for detecting correctly between healthy and unhealthy patients due to wheezing.

Future work will focus on two directions: (i) novel spectro-temporal features to correctly discriminate between wheezing and non-wheezing (respiratory) bases/activations from NMF approaches, and (ii) alternative constrained non-negative matrix factorization approaches not applied to the whole input mixture spectrogram but in temporal segments for wheezing real-time detection.

Acknowledgment

The authors would like to thank Dr. Dinko Oletic and Dr. Vedran Bilas for sharing their test dataset. This work was supported by the Spanish Ministry of Economy and Competitiveness under Project TEC2015-67387-C4-2-R.

Appendix A. Terms of the multiplicative update rules

Here, each of the terms belonging to the respiratory multiplicative update rules are detailed:

$$\left[\frac{\partial D_{KL}\left(\mathbf{X}_n|\hat{\mathbf{X}}_n\right)}{\partial \hat{\mathbf{B}}_R}\right]^{-} = \left[\hat{\mathbf{X}}_n^{-1} \odot \mathbf{X}_n\right]\hat{\mathbf{A}}_R' \tag{A.1}$$

$$\left[\frac{\partial D_{KL}\left(\mathbf{X}_n|\hat{\mathbf{X}}_n\right)}{\partial \hat{\mathbf{B}}_R}\right]^{+} = \hat{\mathbf{A}}_R' \tag{A.2}$$

$$\left[\frac{\partial D_{KL}\left(\mathbf{X}_n|\hat{\mathbf{X}}_n\right)}{\partial \hat{\mathbf{A}}_R}\right]^{-} = \hat{\mathbf{B}}_R'\left[\hat{\mathbf{X}}_n^{-1} \odot \mathbf{X}_n\right] \tag{A.3}$$

$$\left[\frac{\partial D_{KL}\left(\mathbf{X}_n|\hat{\mathbf{X}}_n\right)}{\partial \hat{\mathbf{A}}_R}\right]^{+} = \hat{\mathbf{B}}_R' \tag{A.4}$$

$$\left[\frac{\partial \phi\left(\hat{\mathbf{B}}_R\right)}{\partial \hat{\mathbf{B}}_R}\right]^{+}_{f,k_r} = \frac{4F\widehat{B}_{R_{f,k_r}}}{\sum_{j=1}^{F}\widehat{B}^2_{R_{j,k_r}}} \tag{A.5}$$

$$\left[\frac{\partial \phi\left(\hat{\mathbf{A}}_R\right)}{\partial \hat{\mathbf{A}}_R}\right]^{+}_{k_r,t} = \frac{4T\widehat{A}_{R_{k_r,t}}}{\sum_{i=1}^{T}\widehat{A}^2_{R_{k_r,i}}} \tag{A.6}$$

$$\left[\frac{\partial \phi\left(\hat{\mathbf{B}}_R\right)}{\partial \hat{\mathbf{B}}_R}\right]^{-}_{f,k_r} = 2F\left[\frac{\left(\widehat{B}_{R_{f-1,k_r}} + \widehat{B}_{R_{f+1,k_r}}\right)}{\sum_{j=1}^{F}\widehat{B}^2_{R_{j,k_r}}}\right] + \frac{2F\widehat{B}_{R_{f,k_r}}\sum_{j=2}^{F}\left(\widehat{B}_{R_{j,k_r}} - \widehat{B}_{R_{j-1,k_r}}\right)^2}{\left(\sum_{j=1}^{F}\widehat{B}^2_{R_{j,k_r}}\right)^2} \tag{A.7}$$

$$\left[\frac{\partial \phi\left(\hat{\mathbf{A}}_R\right)}{\partial \hat{\mathbf{A}}_R}\right]^{-}_{k_r,t} = 2T\left[\frac{\left(\widehat{A}_{R_{k_r,t-1}} + \widehat{A}_{R_{k_r,t+1}}\right)}{\sum_{i=1}^{T}\widehat{A}^2_{R_{k_r,i}}}\right] + \frac{2T\widehat{A}_{R_{k_r,t}}\sum_{i=2}^{T}\left(\widehat{A}_{R_{k_r,i}} - \widehat{A}_{R_{k_r,i-1}}\right)^2}{\left(\sum_{i=1}^{T}\widehat{A}^2_{R_{k_r,i}}\right)^2} \tag{A.8}$$

Here, each of the terms belonging to the wheezing multiplicative update rules are detailed:

$$\left[\frac{\partial D_{KL}\left(\mathbf{X}_n|\hat{\mathbf{X}}_n\right)}{\partial \hat{\mathbf{B}}_W}\right]^{-} = \left[\hat{\mathbf{X}}_n^{-1} \odot \mathbf{X}_n\right]\hat{\mathbf{A}}_W' \tag{A.9}$$

$$\left[\frac{\partial D_{KL}\left(\mathbf{X}_n|\hat{\mathbf{X}}_n\right)}{\partial \hat{\mathbf{B}}_W}\right]^{+} = \hat{\mathbf{A}}_W' \tag{A.10}$$

$$\left[\frac{\partial D_{KL}\left(\mathbf{X}_n|\hat{\mathbf{X}}_n\right)}{\partial \hat{\mathbf{A}}_W}\right]^{-} = \hat{\mathbf{B}}_W'\left[\hat{\mathbf{X}}_n^{-1} \odot \mathbf{X}_n\right] \tag{A.11}$$

$$\left[\frac{\partial D_{KL}\left(\mathbf{X}_n|\hat{\mathbf{X}}_n\right)}{\partial \hat{\mathbf{A}}_W}\right]^{+} = \hat{\mathbf{B}}_W' \tag{A.12}$$

$$\left[\frac{\partial \psi\left(\hat{\mathbf{B}}_W\right)}{\partial \hat{\mathbf{B}}_W}\right]^{+}_{f,k_w} = \frac{1}{\sqrt{\frac{1}{F}\sum_{j=1}^{F}\widehat{B}^2_{W_{j,k_w}}}} \tag{A.13}$$

$$\left[\frac{\partial \psi\left(\hat{\mathbf{B}}_W\right)}{\partial \hat{\mathbf{B}}_W}\right]^{-}_{f,k_w} = \sqrt{F}\frac{\widehat{B}_{W_{f,k_w}}\sum_{j=1}^{F}\widehat{B}_{W_{j,k_w}}}{\left(\sum_{j=1}^{F}\widehat{B}^2_{W_{j,k_w}}\right)^{\frac{3}{2}}} \tag{A.14}$$

Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.apacoust.2018.12.035.

References

[1] Stowell D, Giannoulis D, Benetos E, Lagrange M, Plumbley MD. Detection and classification of acoustic scenes and events. IEEE Trans Multimedia 2015;17(10):1733–46.

[2] Cobos M, Perez-Solano JJ, Berger LT. Acoustic-Based Technologies for Ambient Assisted Living. In: Introduction to Smart eHealth and eCare Technologies. Boca Raton, USA: Taylor & Francis Group; 2016. p. 159–80.

[3] Alsina-Pages RM, Navarro J, Alias F, Hervas M. homeSound: Real-Time Audio Event Detection Based on High Performance Computing for Behaviour and Surveillance Remote Monitoring, Sensors, vol. 17, no. 4; 2017.

[4] Martin-Morato I, Cobos M, Ferri FJ. Analysis of data fusion techniques for multi-microphone audio event detection in adverse environments. In: 19th IEEE International Workshop on Multimedia Signal Processing (MMSP), Lutton, UK.

[5] Mondal A, Banerjee P, Tang H. A novel feature extraction technique for pulmonary sound analysis based on EMD. Comput Methods Programs Biomed 2018;159:199–209.

[6] Mondal A, Banerjee P, Somkuwar A. Enhancement of lung sounds based on empirical mode decomposition and Fourier transform algorithm. Comput Methods Programs Biomed 2017;139:119–36.

[7] MedlinePlus. Wheezing; 2018. [Online]. Available:https://medlineplus.gov/ency/article/003070.htm.

[8] World Health Organization, Chronic respiratory diseases; 2018. [Online]. Available:http://www.who.int/respiratory/asthma/en/.

[9] Sarkar M, Madabhavi I, Niranjan N, Dogra M. Auscultation of the respiratory system. Ann Thoracic Med 2015;10(3):158–68.

[10] Lozano F, Salazar A, Alvarado C. System of heart and lung sounds separation for store-and-forward telemedicine applications. Antioquia: Revista Facultad Ingennieria Univ; 2012.

[11] Pasterkamp H, Kraman SS, Wodicka GR. Respiratory sounds advances beyond stethoscope. Am J Respir Crit Care Med 1997;156:974–87.

[12] Sovijarvi F, Dalmasso F, Vanderschoot J, Malmberg L, Righini G, Stoneman S. "Definition of terms for applications of respiratory sounds. Eur Respir Rev 2000;10:597–610.

[13] Lin B, Lin B, Wu H, Chong F, Chen S. Wheeze recognition based on 2D bilateral filtering of spectrogram. Biomed Eng Appl Basis Commun 2006;18(3).

[14] Qiu Y, Whittaker A, Lucas M, Anderson K. Automatic wheeze detection based on auditory modeling. Proc Inst Mech Eng 2005;219(3):219–27.

[15] Jin F, Sattar F, Goh D. Automatic wheeze detection using histograms of sample entropy. In 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society; 2008. p. 1890–93.

[16] Forkheim K, Scuse D, Pasterkamp H. A comparison of neural network models for wheeze detection. In: WESCANEX 95. Communications, Power, and Computing. Conference Proceedings. IEEE, vol. 1; 1995. p.214–19.

[17] Lin B, Wu H, Chen S. Automatic wheezing detection based on signal processing of spectrogram and back-propagation neural network. J Healthcare Eng 2015;6(4):649–72.

[18] Kochetov K, Putin E, Azizov S, Skorobogatov I, Filchenkov A. Wheeze detection using convolutional neural networks. In: Progress in Artificial Intelligence. EPIA 2017. Lecture Notes in Computer Science, 10423. Springer; 2017.

[19] Kandaswamy A, Kumar C, Ramanathan R, Jayaraman S, Malmurugan N. Neural classification of lung sounds using wavelet coefficients. Comput Biol Med 2004;34:523–37.

[20] Le Cam S, Belghith A, Collet Ch, Salzenstein F. Wheezing sounds detection using multivariate generalized gaussian distributions. IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan; 2009.

[21] Wisniewski M, Zielinski T. Tonality detection methods for wheezes recognition system. 19th International Conference on Systems, Signals and Image Processing (IWSSIP); 2012. p 472–75.

[22] Wisniewski M, Zielinski T. Joint application of audio spectral envelope and tonality index in an e-asthma monitoring system. IEEE J Biomed Health Inf 2015;19(3):1009–18.

[23] Chien J, Wu H, Chong F, Li C. Wheeze detection using cepstral analysis in Gaussian Mixture Models. In: 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society Models. p. 3168–71.

[24] Bahoura M. Pattern recognition methods applied to respiratory sounds classification into normal and wheeze classes. Comput Biol Med 2009;39 (9):824–43.
[25] Mayorga P, Druzgalski C, Morelos R, Gonzalez O, Vidales J. In: IEEE Engineering in Medicine and Biology Annual International Conference of the. p. 6312–6.
[26] Mazic I, Bonkovic M, Dzaja B. Two-level coarse-to-fine classification algorithm for asthma wheezing recognition in children's respiratory sounds. Biomed Signal Process Control 2015;21:105–18.
[27] Bokov P, Mahut B, Flaud P, Delclaux C. Wheezing recognition algorithm using recordings of respiratory sounds at the mouth in a pediatric population. Comput Biol Med 2016;70:40–50.
[28] Shaharum S, Sundaraj K, Aniza S, Palaniappan R, Helmy K. Classification of asthma severity levels by wheeze sound analysis. IEEE Conference on Systems, Process and Control (ICSPC); 2016. p. 172–76.
[29] Oletic D, Bilas V. Asthmatic wheeze detection from compressively sensed respiratory sound spectra. IEEE J Biomed Health Inf 2017. https://doi.org/10.1109/JBHI.2017.2781135.
[30] Homs-Corbera A, Fiz J, Morera J, Jane R. Time-frequency detection and analysis of wheezes during forced exhalation. IEEE Trans Biomed Eng 2004;51 (1):182–6.
[31] Alic A, Lackovic I, Bilas V, Sersic D, Magjarevic R. A novel approach to wheeze detection. In: World Congress on Medical Physics and Biomedical Engineering 2006. Springer; 2007. p. 963–6.
[32] Taplidou SA, Hadjileontiadis LJ. Wheeze detection based on time-frequency analysis of breath sounds. Comput Biol Med 2007;37(8):1073–83.
[33] Jain A, Vepa J. "Lung sound analysis for wheeze episode detection", 30th. Ann Int IEEE EMBS Conference 2008:2582–5.
[34] Riella R, Nohama P, Maia J. Method for automatic detection of wheezing in lung sounds. Braz J Med Biol Res 2009;42(7):674–84.
[35] Mendes L, Vogiatzis I, Perantoni E, Kaimakamis E, Chouvarda I, Maglaveras N, et al. Detection of wheezes using their signature in the spectrogram space and musical features. In: Proceedings of the, 2015 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBC; 2015. p. 5581–84.
[36] Shreur H, Vanderschoot J, Zwinderman A, Dijkman JH J, Sterk P. The effect of methacholine-induced acute airway narrowing on lung sounds in normal and asthmatic subjects. Eur Respir J 1995;8:257–65.
[37] Nagasaka Y. Lung Sounds in Bronchial Asthma. Allergology Int 2012;61:353–63.
[38] Oletic D, Arsenali B, Bilas V. Low-power wearable respiratory sound sensing. Sensors 2014;14:6535–66.
[39] Lee D, Seung H. Learning the parts of objects by non-negative matrix factorization. Nature 1999;401(6755):788–91.
[40] Lee D, Seung S. Algorithms for non-negative matrix factorization. In: Proceedings of Advances in Neural Inf. Process. System; 2000. p. 556–62.
[41] Damon C, Liutkus A, Gramfort A, Essid S. Non-negative matrix factorization for single-channel EEG artifact rejection. In: IEEE International Conference on Acoustics, Speech and Signal Processing. p. 1177–81.
[42] Tsubakida H, Shiratori T, Ishiyama A, Ono Y. Nonnegative matrix factorization common spatial pattern in brain machine interface. In: 3rd International Winter Conference on Brain-Computer Interface. p. 1–4.
[43] Canadas-Quesada F, Ruiz-Reyes N, Carabias-Orti J, Vera-Candeas P, Fuertes-Garcia J. A non-negative matrix factorization approach based on spectro-temporal clustering to extract heart sounds. Appl Acoust 2017;125:7–19.
[44] Torre-Cruz J, Canadas-Quesada F, Vera-Candeas P, Montiel-Zafra V, Ruiz-Reyes N. Wheezing sound separation based on constrained non-negative matrix factorization. In: Proc ICBBT 10th International Conference on Bioinformatics and Biomedical Technology. p. 18–24. https://doi.org/10.1145/3232059.3232072.
[45] Févotte C, Bertin N, Durrieu J. Nonnegative matrix factorization with the Itakura-Saito divergence with application to music analysis. Neural Comput 2009;21(3):793–830.
[46] Canadas F, Vera P, Ruiz N, Carabias J, Cabanas P. Percussive harmonic sound separation by non-negative matrix factorization with smoothness-sparseness constraints. J Audio Speech Music Process 2014;2014(26):1–17.
[47] Laroche C, Kowalski M, Papadopoulos H, Richard G. A structured nonnegative matrix factorization for source separation. In: 23rd European Signal Processing Conference (EUSIPCO); 2015. p. 2033–37.
[48] Eggert J, Korner E. Sparse coding and nmf. Neural Networks 2004.
[49] Virtanen T. Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. IEEE Trans Audio Speech Lang Process 2007;15(3):1066–74.
[50] Parras J, Canadas F, Vera P, Ruiz N. Audio restoration of solo guitar excerpts using a excitation-filter instrument model. Stockholm Music Acoustics Conference jointly with Sound And Music Computing Conference; 2013.
[51] Yuan X, Martnez JF, Echert M, pez-Santidrin L. An improved Otsu threshold segmentation method for underwater simultaneous localization and mapping-based navigation. Sensors 2016;16(7):1148.
[52] The r.a.l.e. repository. [Online]. Available:http://www.rale.ca/.
[53] Stethographics lung sound samples. [Online]. Available:http://www.stethographics.com/.
[54] 3m littmann stethoscopes. [Online]. Available:http://solutions.3m.com/.
[55] East tennessee state university pulmonary breath sounds. [Online]. Available: http://faculty.etsu.edu.
[56] ICBHI 2017 Challenge. [Online]. Available:https://bhichallenge.med.auth.gr/sites/default/.
[57] Lippincott NursingCenter. [Online]. Available:https://www.nursingcenter.com/.
[58] Thinklabs Digital Stethoscope. [Online]. Available:https://www.thinklabs.com/.
[59] Emedicine/Medscape. [Online]. Available:https://emedicine.medscape.com/.
[60] Fevotte C, Gribonval R, Vincent E. BSS_EVAL toolbox user guide – Revision, 2.0, Technical Report 1706, IRISA; (April 2005).
[61] Vicent E, Fevotte C, Gribonval R. Performance measurement in blind audio source separation. IEEE Trans Audio Speech Language Process 2006;14 (4):1462–9.
[62] Theodoridis S, Koutroumbas K. Pattern Recognition. Third edition. Elsevier: Academic Press; 2006.

Paper 3

A constrained tonal semi-supervised non-negative matrix factorization to classify presence/absence of wheezing in respiratory sounds

J. Torre-Cruz, F. Canadas-Quesada, S. García-Galán, N. Ruiz-Reyes, P. Vera-Candeas and J. Carabias-Orti, "A constrained tonal semi-supervised non-negative matrix factorization to classify presence/absence of wheezing in respiratory sounds", in *Applied Acoustics*, Volume 161, April 2020, pp. 107-188. DOI: `https://doi.org/10.1016/j.apacoust.2019.107188`

- Estado: Publicado.
- Revista: Applied Acoustics.
- ISSN: 0003-682X.
- Factor de impacto (JCR 2019): 2.440.
- Cuartiles por área de conocimiento:
  - Acoustics: Q2, 9/32.

Applied Acoustics 161 (2020) 107188

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

A constrained tonal semi-supervised non-negative matrix factorization to classify presence/absence of wheezing in respiratory sounds

J. Torre-Cruz *, F. Canadas-Quesada, S. García-Galán, N. Ruiz-Reyes, P. Vera-Candeas, J. Carabias-Orti

Department of Telecommunication Engineering, University of Jaen, Campus Cientifico-Tecnologico de Linares, Avda. de la Universidad, s/n, 23700 Linares, Jaen, Spain

ARTICLE INFO

Article history:
Received 31 July 2019
Received in revised form 8 November 2019
Accepted 15 December 2019
Available online 27 December 2019

Keywords:
Non-negative matrix factorization (NMF)
Divergence
Wheezing
Smoothness
Monophonic constraint
Spectral trajectories

ABSTRACT

From a clinical point of view, the detection of wheezing presence in respiratory sounds is a challenging task for early identification of pulmonary diseases since wheezing is the main manifestation associated to airway obstruction. In this article, we propose a novel method to detect the presence or absence of wheeze sounds in breath recordings in order to increase the reliability of the subjective diagnosis provided by the physician in the auscultation process. Specifically, it is assumed an unhealthy subject when wheeze sounds can be detected during breathing. The proposed method consists of three stages. The first stage attempts to estimate the spectral interval, band of interest (BOI), that shows the highest probability to find wheeze sounds. In the second stage, a constrained tonal semi-supervised non-negative matrix factorization (NMF) approach is applied to obtain spectral patterns that models the periodic or tonal nature typically shown by wheeze sounds. The third stage analyzes the estimated wheezing spectrogram based on the smoothness of the spectral trajectories from the most significant energy previously factorized in the BOI. Our system has been evaluated and compared to other state-of-the-art methods, yielding competitive results in the wheezing presence detection in respiratory sounds.

1. Introduction

To this day, auscultation is still the main technique used in most of the health centres in order to obtain the first clinical diagnosis of the status of the respiratory system of the subjects. From a clinical point of view, this technique shows several advantages since it is non-invasive, low-cost, easy-to-perform and subject-friendly that avoids long-time subjects monitoring [1]. The more important auscultation diagnosis becomes in low-income countries, e.g. Sub-Saharan Africa, due to weak health centres in which there are often no more sophisticated and expensive diagnostic tools such as chest radiographs. However, a correct diagnosis derived from auscultation depends largely on the experience and acoustic training of the physician to interpret what is hearing because a conventional stethoscope tends to attenuate or amplify sounds within the spectral range in which wheeze sounds appear [2,3]. For this reason, many research efforts are applied in biomedical signal processing to develop a reliable method for the early wheezing occurrence detection since it is associated with respiratory obstruction observed in different pulmonary diseases such as asthma, chronic obstructive pulmonary disease (COPD) or even pneumonia, causing narrowing and spasms in the small airways of lungs [4,5]. World Health Organization (WHO) estimated 383,000 deaths due to asthma and 1.4 million deaths of children under five years old worldwide due to pneumonia during 2015 [6,7].

The sounds emitted during breathing are generally classified into normal breath sounds and adventitious sounds (AS) such as wheezing, crackles, pleural rubs and stridor. Normal breath sounds (RS), emitted by healthy lungs, are characterized by a wideband spectrum where most of the energy is concentrated in the frequency band 60 Hz–1600 Hz [8]. Wheeze sounds (WS), emitted by unhealthy lungs, are defined as continuous sinusoidal waveforms in time characterized by a dominant fundamental frequency (pitch). As a result, WS show a set of narrowband spectral peaks forming frequency lines over time (spectral trajectories) that superimpose on RS in the frequency domain but there is still no agreement on a single frequency and duration of wheezes in the literature. This fact increases the rate of misdiagnosis because the pitch of a wheeze can vary which implies that it can be located in another range of the spectrum [9]. According to American Thoracic Society (ATS), WS are defined as a pitch higher than 400 Hz whose duration is longer than 250 ms [10]. According to Computerized Respiratory Sound Analysis (CORSA) guidelines, WS are defined as a pitch higher than 100 Hz with duration longer than

* Corresponding author.
E-mail address: jtorre@ujaen.es (J. Torre-Cruz).

https://doi.org/10.1016/j.apacoust.2019.107188
0003-682X/

100 ms [11] as shown in the time-frequency representation of Fig. 1A. This paper assumes the following facts: i) the term mixture refers to any single-channel input signal which can be composed only of RS (healthy subject) or RS mixed with WS (unhealthy subject); ii) the term RS indicates the no presence of any type of AS (see Fig. 1B), as a result, RS is equivalent to non-wheeze sounds; iii) the term WS indicates the presence of WS (see Fig. 1A); iv) the minimum temporal duration for WS has been selected as 100 ms using the CORSA standard [2]; and finally, v) the pitch range associated to wheezing is not fixed a priori since it will be estimated from each mixture.

Several algorithms to detect the wheezing occurrence in respiratory sounds have been published in the last decades using various approaches: Autoregressive (AR) model [12], Auditory modelling [13], Entropy [14], Linear analysis [15], Autocorrelation [16], Neural networks (NN) [5,17], Wavelet transform [4], Instantaneous frequency (IF) analysis [18], Tonal index [19], Mel-frequency cepstral coefficients (MFCC) [20,21], Gaussian Mixture Models (GMM) [22,23], Image processing [24], Spectral peaks identification [25,26], Hidden Markov model (HMM) [27], Empirical mode decomposition (EMD) [3,28]and Non-negative matrix factorization (NMF) [29].

In [4], it is reported a modeling for wheezing and normal breath sounds in the wavelet packet domain using generalized gaussian distributions. Kochetov et al. [17] proposed wheeze detection using convolutional neural networks. Taplidou and Hadjileontiadis [25] located wheezing based on a smoothing procedure that estimates the trend of the frequency content at each time using mean filtering. Wisniewski and Zielinski [19] combined feature extraction from MPEG-7 and MPEG-2 standards using support vector machine (SVM). To discriminate wheeze and non-wheeze sounds, Mazic et al. [20] developed a two-layer pattern recognition using two SVM classifiers designed as a cascade stacked in parallel using MFCC features and thresholding. In [21], it applied a feature extraction based on MFCC and K-nearest neighbour (KNN) to classify different levels of asthma severity shown by wheeze sounds. Mendes et al. [26] identified wheezing sounds using a temporal Gaussian regularization and a reduction of the false positives based on the morphological opening by reconstruction operator. Lozano et al. [28] detected not only wheezing but any continuous adventitious sounds (CAS) using EMD, features extracted from IF sequences and SVM. Recently, our previous study [29] proposed a two-stage cascade system for wheezing detection. In the first stage, a constrained NMF is designed integrating smoothness and sparseness into the decomposition. In the second stage, the Kullback–Leibler divergence is applied to discriminate between wheeze and non-wheeze sounds frames.

The aim of the proposed method is to automatically detect the presence of wheezing sounds in breath recordings avoiding the subject's return to the health centre with a worsening of the airway obstruction, manifested by wheeze sounds, that was not detected early in the first clinical examination performed by auscultation. Three novel contributions are proposed by authors. Unlike most wheezing detection algorithms in which the pitch range of the wheeze is set a priori, our first contribution, band of interest (BOI), estimates the spectral range in which the probability to find wheeze sounds is maximum. As a second contribution, we present a constrained tonal semi-supervised NMF that factorizes into spectral patterns that correctly model the tonal nature, by narrowband peaks, shown by WS in the estimated BOI in order to separate WS and RS. From the separated wheezing spectrogram, we propose an intuitive method to classify the subject's condition as healthy (absence of wheezing) or unhealthy (presence of wheezing) analyzing the temporal smoothness of the spectral trajectories defined by most significant energy factorized in the estimated BOI.

The rest of this paper is structured as follows. Section 1.1 briefly reviews non-negative matrix factorization (NMF). The proposed method is described in Section 2. Section 3 details and discusses the experimental evaluation. Finally, Section 4 concludes the paper.

1.1. Non-negative matrix factorization

Non-negative matrix factorization (NMF) [30] is a decomposition technique that has attracted a lot of efforts in different fields of biomedical signal processing in the last years [31,32]because NMF is a powerful tool to learn additive part-based spectrograms of the most representative sources by imposing non-negative constraint. The aim of NMF can be defined as follows: given a single-channel mixture $x(t)$ with magnitude spectrogram $\mathbf{X} \in \mathbb{R}_{+}^{F\times T}$ with non-negative entries, NMF approximates $\mathbf{X}$ into the product of two non-negative matrices, basis matrix $\mathbf{B} \in \mathbb{R}_{+}^{F\times K}$ and activation matrix $\mathbf{A} \in \mathbb{R}_{+}^{K\times T}$ as shown in Eq. (1),

$$\mathbf{X} \approx \hat{\mathbf{X}} = \mathbf{BA} \tag{1}$$

obtaining the estimated spectrogram $\hat{\mathbf{X}}$, the number of frequency bins F, the number of time frames T and the number of components K (K is usually chosen such that $K(F+T) \ll FT$ in order to reduce the data dimension). Therefore, $\mathbf{B}$ can be interpreted as a dictionary or a set of spectral patterns that codes the frequency information associated to the sources active in the spectrogram and $\mathbf{A}$ as a matrix of activations that indicates the activity of each spectral basis in a given time frame.

The NMF factorization is generally calculated by minimizing a cost function $D\left(\mathbf{X}|\hat{\mathbf{X}}\right)$ that penalizes the error reconstruction between $\mathbf{X}$ and $\hat{\mathbf{X}}$. Considering the most used cost functions such

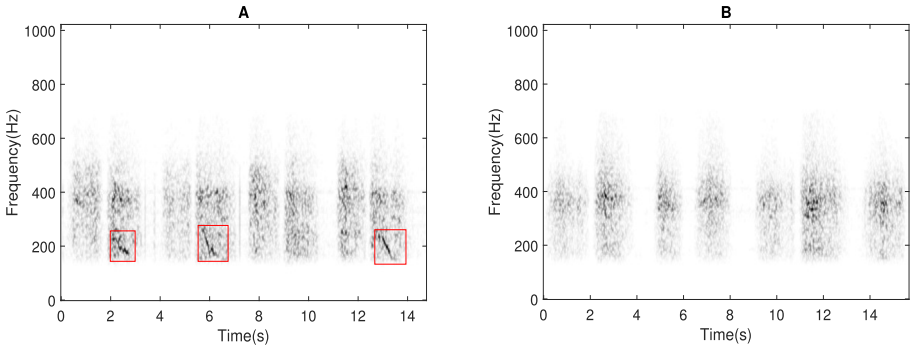

Fig. 1. Time-Frequency representation (spectrogram) of breathing recording. A) Spectrogram of a mixture (unhealthy subject) composed of three wheeze sounds (red rectangles) and normal breath sounds. B) Spectrogram of a mixture (healthy subject) composed only of normal breath sounds. A darker grey colour represents higher energy of each frequency.

as Euclidean distance, the generalized Kullback-Leibler divergence, the Itakura-Saito divergence and the Cauchy distribution [33,34], in this paper, we have used the generalized Kullback-Liebler divergence $D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}\right)$ because it is non-increasing and ensures the non-negativity of **B** and **A** using multiplicative update rules [30] and moreover, recent previous works [29,31]have obtained competitive results in biomedical signal processing,

$$D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}\right)=\sum_{f,t}\left(X_{f,t}\log\frac{X_{f,t}}{\widehat{X}_{f,t}}-X_{f,t}+\widehat{X}_{f,t}\right) \tag{2}$$

The main disadvantage of the NMF factorization is its inability to reconstruct each isolated source because NMF only ensures the reconstruction of the whole mixture but it does not guarantee parts-based objects with physical interpretation as occurs in real world [35]. In order to find an optimal NMF factorization, some of the most widely used solutions in audio processing to add physical meaning to the NMF bases and activations are the following: i) adding prior information of the sources into the NMF decomposition using additional constraints [36] and ii) train or characterize spectral patterns of target sounds to be separated [37].

2. Materials and methods

Our proposal attempts to classify the subject's condition, that is, healthy or unhealthy. We assume as a healthy subject that subject in which the absence of wheezing sounds has been detected. We assume as an unhealthy subject that subject in which the presence of wheezing sounds has been detected. For this purpose, the proposed method (see Fig. 2) is composed of three stages: Estimation of the Spectral Band Of Interest (Stage I), Wheezing/Normal breath Sound Separation (Stage II) and Wheezing presence/absence classification (Stage III).

2.1. Mixing model and time-frequency representation

As previously mentioned in Section 1, the term mixture $x(t)$ refers to any single-channel input signal: i) $x(t)=r(t)$ only composed of RS signal $r(t)$, that is, a healthy subject; or ii) $x(t)=r(t)+s(t)$ composed of RS signal $r(t)$ mixed in an approximately linear manner with WS signal $s(t)$, that is, an unhealthy subject.

The magnitude spectrogram **X** of a mixture $x(t)$ is composed of T frames, F frequency bins and a set of time–frequency units $X_{f,t}$, being $f=1,\ldots,F$ and $t=1,\ldots,T$. Each unit $X_{f,t}$ is defined by the f^{th} frequency bin at the t^{th} frame and is calculated from the magnitude of the Short-Time Fourier Transform (STFT) using a Hamming windows of N samples with 25% overlap. A normalization is applied to **X** in order to be independent of the size and scale where the normalized magnitude spectrogram $\overline{\mathbf{X}}$ is computed as follows,

$$\overline{\mathbf{X}}=\frac{\mathbf{X}}{\left(\frac{\sum_{f,t}X_{f,t}}{FT}\right)} \tag{3}$$

To avoid complex nomenclature throughout the paper, the variable **X** is hereinafter referred to the normalized magnitude spectrogram previously computed in Eq. (3).

2.2. Stage I: estimation of the spectral band of interest

The goal of this stage is to estimate the spectral interval, defined as band of interest (BOI), in which the probability to find wheeze sounds is maximum. The estimation of BOI is performed by modelling the tonal nature shown by wheeze sounds in time–frequency domain (e.g., it can be observed that BOI is delimited between 150–300 Hz in the magnitude spectrogram of an unhealthy subject shown in Fig. 1A).

We propose a method combining NMF and the periodicity principle to estimate the BOI assuming that a wheeze sound exhibits a strongly periodic or tonal nature characterized by the presence of narrowband spectral peaks. In order to estimate BOI, this stage can be divided into three steps that are detailed below.

2.2.1. Step I: factorization of the basis matrix applying NMF

The first step attempts to estimate the basis **B** and activations **A** matrix minimizing the reconstruction error defined by the cost function $D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}\right)$ (see Eq. (2)). Finally, the matrices **B** and **A** are obtained by applying a gradient descent algorithm based on multiplicative update rules [33].

2.2.2. Step II: clustering of bases based on the tonality principle

Here, the goal is to cluster the NMF bases **B** obtained in *Step I* into two groups: (i) NMF bases that show higher periodicity $\mathbf{B}_W$ (see Fig. 3A depicting bases that model the tonal nature typically shown by WS); and (ii) NMF bases that show lower periodicity $\mathbf{B}_R$ (see Fig. 3B representing bases that model the non-tonal nature shown by RS). The higher periodicity tends to concentrate the energy of the NMF bases in narrowband spectral peaks (see Fig. 3C) while the lower periodicity distributes such energy in wideband spectrum (see Fig. 3D). Therefore, we assume that tonality or periodicity implies a spectrum that shows a sparse distribution of energy.

Although a preliminary study related to several sparse descriptors [38] has been analyzed in this work, empirical results indicated that the sparse descriptor Gini index β provided the best classification results of the NMF bases in terms of the degree of periodicity. Besides, these results are supported by recent works [38,39]that have demonstrated the reliability and robustness shown by β to correctly cluster between sparse and non-sparse

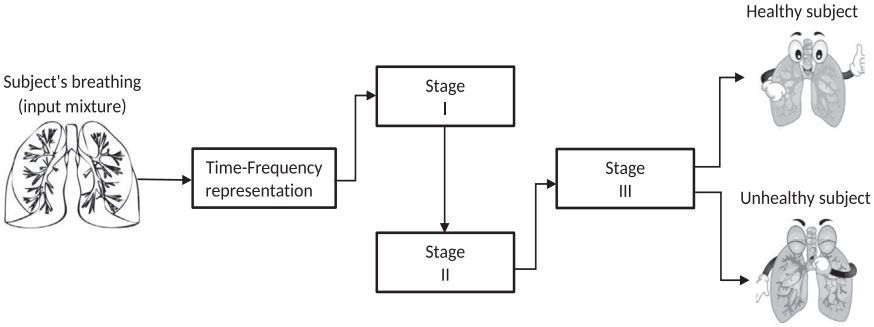

Fig. 2. Block-scheme of the proposed method.

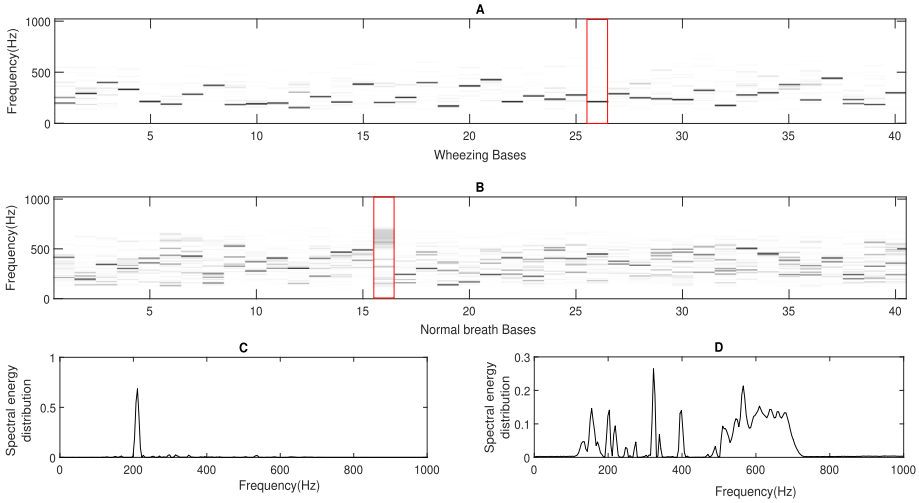

Fig. 3. Example of the NMF bases ($K = 80$ bases) classification between WS and RS considering the magnitude spectrogram belonging to the unhealthy subject shown in Fig. 1A. A) The matrix $\mathbf{B}_W$ is composed of the forty NMF bases with highest spectral tonality (narrowband spectral peaks). B) The matrix $\mathbf{B}_R$ is composed of the forty NMF bases with lowest spectral tonality behavior (wideband). C) The spectral energy distribution (sparse) associated to the most tonal NMF basis, specifically, $\mathbf{B}_W(26)$ (red rectangle displayed in Fig. 3A). D) The spectral energy distribution (non-sparse) of the least tonal NMF basis, specifically, $\mathbf{B}_R(16)$ (red rectangle displayed in Fig. 3B).

distributions. Therefore, we propose to apply the Gini index $\beta(k)$ to calculate the degree of tonality or periodicity for each k^{th} NMF basis of the dictionary $\mathbf{B}$. High values of $\beta(k)$ imply high periodic nature and vice versa. In addition, a remarkable advantage of $\beta(k)$ is that it is a normalized descriptor whose values are ranged between 0 and 1 for each NMF basis.

According to each value of $\beta(k)$, $\mathbf{B}_W$ and $\mathbf{B}_R$ are obtained applying a thresholding process. The aim of this thresholding is not to achieve the optimal separation between WS and RS since it will be obtained in Section 2.3 as shown in Fig. 6. Specifically, this thresholding attempts to locate the BOI identifying two sets of NMF bases, $\mathbf{B}_W$ and $\mathbf{B}_R$, that show different periodic behavior. For this purpose, we propose to apply a threshold ζ_m equal to the median of the sparse values provided by the sparse descriptor $\beta(k)$ for the following two reasons: (i) the median is both a robust threshold against the dispersion of periodicity values and resistant to gross periodicity error by avoiding the problem generated by outliers [40,41]; and (ii) the median achieves to cluster uniformly between two groups, providing half of the NMF bases for both groups $\mathbf{B}_W$ and $\mathbf{B}_R$ (see Eq. (4)). This allows to factorize a reliable set of NMF bases $\mathbf{B}_W$ to locate the BOI in which the probability to find WS is maximum due to the factorization of NMF bases with higher periodicity.

$$\mathbf{B}(k) \rightarrow \begin{cases} \mathbf{B}(k) \in \mathbf{B}_W & \text{if } \beta(k) \geqslant \zeta_m \\ \mathbf{B}(k) \in \mathbf{B}_R & \text{if } \beta(k) < \zeta_m \end{cases} \tag{4}$$

where $k = 1,\ldots,K$.

Next, the estimated wheezing spectrogram $\hat{\mathbf{X}}_W$ used to locate the BOI can be reconstructed as follows,

$$\hat{\mathbf{X}}_W = \mathbf{B}_W \mathbf{A}_W \tag{5}$$

where the wheezing activations matrix $\mathbf{A}_W$ has been factorized using the same set of components clustered in $\mathbf{B}_W$.

2.2.3. Step III: using energy distortion to select the spectral BOI

From the estimated wheezing spectrogram $\hat{\mathbf{X}}_W$, the third step selects the spectral range BOI in which the probability of finding wheeze sounds is maximum. As a result, BOI shows the least spectral energy loss when comparing $\hat{\mathbf{X}}_W$ and $\mathbf{X}$ since $\hat{\mathbf{X}}_W$ roughly models most of the wheeze sounds from the input spectrogram $\mathbf{X}$. To calculate BOI, we propose to use the information provided by the spectral energy distortion $\delta(f)$ (see Eq. (6)) measuring the degree of similarity $\upsilon(f)$ for each frequency bin along all frames between $\mathbf{X}$ and $\hat{\mathbf{X}}_W$ as shown in Eq. (7). A high value of $\upsilon(f_i)$ indicates that the frequency f_i exhibits a high tonal content as typically shown by WS because f_i shows low attenuation in the spectral energy distortion. A low value of $\upsilon(f_i)$ indicates that the frequency f_i exhibits a low tonal content as shown by RS because f_i shows high attenuation in the spectral energy distortion.

$$\delta(f) = \sqrt{\sum_{t=1}^{T} X_{f,t} - \sum_{t=1}^{T} \hat{X}_{W_{f,t}}}, \quad f = 1 \ldots F \tag{6}$$

$$\upsilon(f) = \frac{\sum_{t=1}^{T} \hat{X}_{W_{f,t}}}{\delta(f)}, \quad f = 1 \ldots F \tag{7}$$

Following a conservative strategy to ensure that all wheeze sounds are contained in the BOI, we propose to find the most prominent peak located in $\upsilon(f)$. Note that the prominence of a peak is defined as the minimum vertical distance that the signal must descend on either side of the peak before either climbing back to a level higher than the peak or reaching an endpoint [42]. As shown in Eq. (8), the boundaries (f_{min}, f_{max}) of the BOI can be calculated in terms of the width Δ of the mainlobe related to the most prominent peak previously mentioned. A preliminary analysis indicated that the use of an additional margin $\eta = 50$ Hz from the mainlobe of the most prominent peak achieved the best performance to contain most of the wheeze sounds determined by the estimated BOI.

$$BOI = [f_{min} - f_{max}] \quad f_{min} = min(\Delta) - \eta \quad f_{max} = max(\Delta) + \eta \tag{8}$$

where f_{min} and f_{max} represent the minimum and maximum frequency that defines the spectral interval of the BOI. Fig. 4 shows an example of the procedure described in this step when the magnitude spectrogram corresponds to the unhealthy subject shown in Fig. 1A.

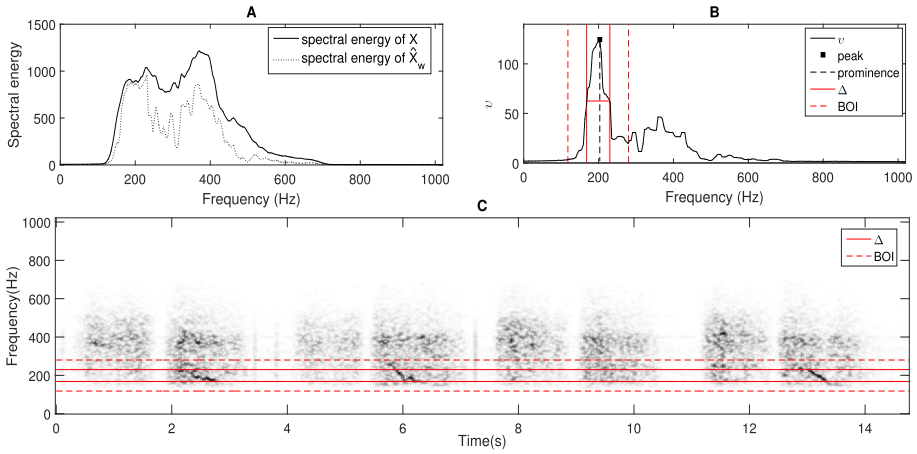

Fig. 4. BOI estimation when the magnitude spectrogram corresponds to the unhealthy subject shown in Fig. 1A. A) Spectral energy of $\mathbf{X}$ and $\hat{\mathbf{X}}_W$. B) Degree of similarity $\upsilon(f)$ and location of the most prominent peak in terms of Δ. C) Magnitude spectrogram shown in Fig. 1A including the frequency range limited by Δ and BOI.

2.3. Stage II: wheezing/normal breath sound separation

The main contribution of this stage is the development of a Wheezing/Normal breath sound separation based on a constrained tonal semi-supervised NMF approach using redundant information from the BOI obtained in the previous stage I. To that end, an objective function is defined to decompose a mixture spectrogram $\mathbf{X}$ into two separated or estimated spectrograms, $\hat{\mathbf{X}}_{R_I}$ (only normal breath sounds spectrogram) and $\hat{\mathbf{X}}_{W_I}$ (only wheeze sounds spectrogram). The factorization model can be defined as follows,

$$\mathbf{X} \approx \hat{\mathbf{X}}_I = \hat{\mathbf{X}}_{R_I} + \hat{\mathbf{X}}_{W_I} = \mathbf{B}_{\mathbf{R_I}} diag(D_R)\mathbf{A}_{\mathbf{R_I}} + \mathbf{B}_{\mathbf{W_I}} diag(D_W)\mathbf{A}_{\mathbf{W_I}} \tag{9}$$

where $\mathbf{B}_{\mathbf{R_I}}, \mathbf{B}_{\mathbf{W_I}}, \mathbf{A}_{\mathbf{R_I}}$ and $\mathbf{A}_{\mathbf{W_I}}$ are the estimated basis and activation matrices of the normal breath sounds and the wheeze sounds and $\hat{\mathbf{X}}_I$ is the estimated magnitude spectrogram estimated at this stage. All of these matrices are non-negative matrices. The number of wheezing and normal breath sounds components is the same and is denoted by Q. The L^2-norm of each column of $\mathbf{B}_{\mathbf{R_I}}$ or $\mathbf{B}_{\mathbf{W_I}}$ is equal to 1.0. The terms D_R and D_W represent vectors with the L^2-norm of each activation component of wheezing and normal breathing, respectively. Therefore, the L^2-norm of each row of $\mathbf{A}_{\mathbf{R_I}}$ or $\mathbf{A}_{\mathbf{W_I}}$ be equal to 1.0 due to the normalization procedure at each iteration. The $diag()$ operator is the diagonal matrix.

To refine and improve the estimated wheezing spectrogram $\hat{\mathbf{X}}_W$ of the stage I, this stage attempts to model the spectral periodic or tonal behavior typically shown by wheeze sounds. We propose to use a constrained tonal semi-supervised NMF approach described below.

2.3.1. A tonal semi-supervised NMF

The tonal semi-supervised NMF is defined as a semi-supervised NMF in which the target source (wheezing) is learned in advanced modelling the spectral tonal or periodic behavior typically shown by WS. The tonal or periodic behavior is modelled using the matrix $\mathbf{B}_{\mathbf{W_I}}$ that is composed of a set of simulated pitches (narrowband peaks) that model all possible tonal sounds active in the estimated BOI. As a result, each q^{th} basis of the dictionary $\mathbf{B}_{\mathbf{W_I}}$ represents a single pitch located at a specific frequency f_q within the BOI. In order to cover all the pitches that can be found in the BOI, each frequency f_q is separated from each other by a value given by the frequency resolution $\tau = 4$ Hz used throughout the model as follows,

$$\mathbf{B}_{\mathbf{W_I}}(q) = G(f - f_q), \quad f_q \in [f_{min} : \tau : f_{max}] \tag{10}$$

where $G(f)$ is the STFT of a sinusoidal signal multiplied by a Hamming window of N samples, f_q is the fundamental frequency of the pitch for the q^{th} NMF basis related to the dictionary $\mathbf{B}_{\mathbf{W_I}}$ and the number of components $Q = \left\lceil \frac{f_{max} - f_{min}}{\tau} \right\rceil$ is obtained in terms of the spectral resolution τ and the BOI (see Fig. 5).

2.3.2. Adding monophonic constraint

We propose to incorporate temporal monophonic constraint $\Omega(\mathbf{A}_{\mathbf{W_I}})$ [43] to the wheezing activation matrix $\mathbf{A}_{\mathbf{W_I}}$ in order to minimize the number of wheezing bases $\mathbf{B}_{\mathbf{W_I}}$ that can be simultaneously activated at each frame. This constraint forces to factorize a more reliable wheezing spectrogram as in typically found in the real-world. Specifically, this constraint $\Omega(\mathbf{A}_{\mathbf{W_I}})$ is computed taking into account the cross-correlations $\mathbf{A}_{\mathbf{W_I}}\mathbf{A}_{\mathbf{W_I}}^T$ detailed in Eq. (11),

$$\Omega(\mathbf{A}_{\mathbf{W_I}}) = \frac{1}{Q(Q-1)} \sum_{q=1}^{Q} \sum_{t=1}^{T} \mathbf{A}_{\mathbf{W_I}}\mathbf{A}_{\mathbf{W_I}}^T - Trace\left(\mathbf{A}_{\mathbf{W_I}}\mathbf{A}_{\mathbf{W_I}}^T\right) \tag{11}$$

where the *Trace* operator computes the sum of diagonal elements of a square matrix. To equilibrate the importance of the monophonic constraint in the decomposition process, $\Omega(\mathbf{A}_{\mathbf{W_I}})$ is weighted by a factor of $\frac{1}{Q(Q-1)}$. Consequently, the cost $\Omega(\mathbf{A}_{\mathbf{W_I}})$ will be equal to 1.0 in the worst case.

The global objective function $D\left(\mathbf{X}|\hat{\mathbf{X}}_I\right)$ that must be minimized taking into account the signal reconstruction, based on the Kullback-Leibler divergence, and the temporal monophonic constraint is detailed as follows,

$$D\left(\mathbf{X}|\hat{\mathbf{X}}_I\right) = D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}_I\right) + \lambda\Omega(\mathbf{A}_{\mathbf{W_I}}) \tag{12}$$

where λ defines the weight of the temporal monophonic constraint $\Omega(\mathbf{A}_{\mathbf{W_I}})$ applied to only estimated wheezing activations matrix $\mathbf{A}_{\mathbf{W_I}}$.

The estimated wheezing activation matrix $\mathbf{A}_{\mathbf{W_I}}$ (see Eq. (13)), the estimated normal breath basis matrix $\mathbf{B}_{\mathbf{R_I}}$ (see Eq. (14)) and the estimated normal breath activation matrix $\mathbf{A}_{\mathbf{R_I}}$ (see Eq. (15)) can be obtained by applying a gradient descent algorithm based on multiplicative update rules. The estimated wheezing basis matrix $\mathbf{B}_{\mathbf{W_I}}$ is fixed and not updated because it has been modeled based on the tonal nature of WS.

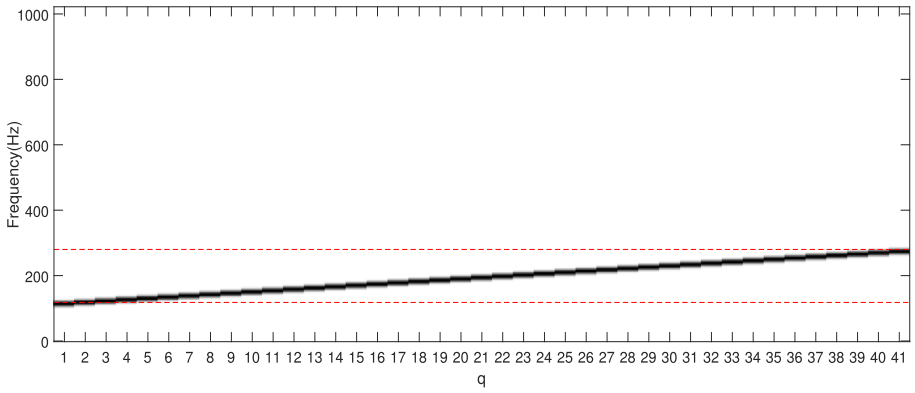

Fig. 5. Dictionary $\mathbf{B}_{\mathbf{W}_I}$ established for the BOI shown in Fig. 4C. The red dotted lines represent the range of the BOI. In this case, the matrix $\mathbf{B}_{\mathbf{W}_I}$ is composed by $Q = 41$ wheezing bases because the BOI, previously obtained in the stage I (Fig. 4C), is spectrally located between $f_{min} = 118$ Hz and $f_{max} = 280$ Hz and the spectral resolution $\tau = 4$ Hz.

$$\mathbf{A}_{\mathbf{W}_I} \leftarrow \mathbf{A}_{\mathbf{W}_I} \odot \frac{\left(\mathbf{B}_{\mathbf{W}_I} diag(D_W)\right)^T \left(\mathbf{X} \text{ø} \hat{\mathbf{X}}_I\right) + 2\lambda \frac{1}{Q(Q-1)} \mathbf{A}_{\mathbf{W}_I}}{\left(\mathbf{B}_{\mathbf{W}_I} diag(D_W)\right)^T \mathbf{1}_{F,T} + 2\lambda \frac{1}{Q(Q-1)} \mathbf{1}_{Q,Q} \mathbf{A}_{\mathbf{W}_I}} \tag{13}$$

$$\mathbf{B}_{\mathbf{R}_I} \leftarrow \mathbf{B}_{\mathbf{R}_I} \odot \frac{\left(\mathbf{X} \text{ø} \hat{\mathbf{X}}_I\right) \left(diag(D_R) \mathbf{A}_{\mathbf{R}_I}\right)^T}{\mathbf{1}_{F,T} \left(diag(D_R) \mathbf{A}_{\mathbf{R}_I}\right)^T} \tag{14}$$

$$\mathbf{A}_{\mathbf{R}_I} \leftarrow \mathbf{A}_{\mathbf{R}_I} \odot \frac{\left(\mathbf{B}_{\mathbf{R}_I} diag(D_R)\right)^T \left(\mathbf{X} \text{ø} \hat{\mathbf{X}}_I\right)}{\left(\mathbf{B}_{\mathbf{R}_I} diag(D_R)\right)^T \mathbf{1}_{F,T}} \tag{15}$$

where $\odot$ is the element-wise multiplication, ø is the element-wise division, $\mathbf{1}_{F,T}$ represents a matrix of all-ones composed of F rows and T columns and ^{T} is the transpose operator.

The set of matrices are obtained updating the rules until the algorithm converges or reaches a maximum number of iterations M. Note that the activation matrices $\mathbf{A}_{\mathbf{W}_I}, \mathbf{A}_{\mathbf{R}_I}$ and the basis matrix $\mathbf{B}_{\mathbf{R}_I}$ must be normalized at each iteration. In particular, the normalization process (see Eq. (16)) ensures that the sum of the square elements of each q^{th} column of the basis matrices $\mathbf{B}_{\mathbf{W}_I}, \mathbf{B}_{\mathbf{R}_I}$ equals 1.0, as does the sum of the square elements of each q^{th} row of the activation matrices $\mathbf{A}_{\mathbf{W}_I}, \mathbf{A}_{\mathbf{R}_I}$ [43].

$$\mathbf{G}(q) = \frac{\mathbf{G}(q)}{\sqrt{\sum \mathbf{G}^2(q)}} \tag{16}$$

$$D_J(q) = D_J(q) \sqrt{\sum \mathbf{G}^2(q)} \tag{17}$$

where $(\mathbf{G}, J) = \{(\mathbf{A}_{\mathbf{W}_I}, W), (\mathbf{A}_{\mathbf{R}_I}, R), (\mathbf{B}_{\mathbf{R}_I}, R)\}$ respectively. If we consider the activation matrix $\mathbf{G} = (\mathbf{A}_{\mathbf{W}_I}, \mathbf{A}_{\mathbf{R}_I})$, $\sqrt{\sum \mathbf{G}^2(q)} = \sqrt{\sum_{t=1}^{T} \mathbf{G}^2(q,t)}$. If we consider the basis matrix $\mathbf{G} = \mathbf{B}_{\mathbf{R}_I}$, $\sqrt{\sum \mathbf{G}^2(q)} = \sqrt{\sum_{f=1}^{F} \mathbf{G}^2(f,q)}$. Next, the estimated magnitude spectrograms $\hat{\mathbf{X}}_{R_I}$ and $\hat{\mathbf{X}}_{W_I}$ (see Eq. (18)) can be obtained from the estimated basis and activation matrices. In this work, we focus only on the estimated wheezing spectrogram $\hat{\mathbf{X}}_{W_I}$ to differentiate healthy and unhealthy subjects. The separation procedure is summarized in Algorithm 1.

$$\hat{\mathbf{X}}_{J_I} = \mathbf{B}_{\mathbf{J}_I} diag(D_J) \mathbf{A}_{\mathbf{J}_I} \tag{18}$$

where J = (W or R) as occurs in Eq. (17).

Algorithm 1 Wheezing/Normal breath sound separation procedure

Require: The input magnitude spectrogram $\mathbf{X}$ and its estimated BOI.

1 Create the dictionary $\mathbf{B}_{\mathbf{W}_I}$ using Eq. (10).
2 Initialize $\mathbf{A}_{\mathbf{W}_I}, \mathbf{B}_{\mathbf{R}_I}$ and $\mathbf{A}_{\mathbf{R}_I}$ with random non-negative values.
3 Update $\mathbf{A}_{\mathbf{W}_I}$ using Eq. (13).
4 Normalize the activation matrix $\mathbf{A}_{\mathbf{W}_I}$ to obtain a L^2-norm of each row of the matrix equal to 1.0 using Eq. (16).
5 Update $diag(D_W)$ with the values that were used to normalize each row of the activation matrix $\mathbf{A}_{\mathbf{W}_I}$ using Eq. (17).
6 Update $\mathbf{B}_{\mathbf{R}_I}$ using Eq. (14).
7 Normalize the basis matrix $\mathbf{B}_{\mathbf{R}_I}$ to obtain a L^2-norm of each column of the matrix equal to 1.0 using Eq. (16).
8 Update $diag(D_R)$ with the values that were used to normalize each column of the basis matrix $\mathbf{B}_{\mathbf{R}_I}$ using Eq. (17).
9 Update $\mathbf{A}_{\mathbf{R}_I}$ using Eq. (15).
10 Normalize the activation matrix $\mathbf{A}_{\mathbf{R}_I}$ to obtain a L^2-norm of each row of the matrix equal to 1.0 using Eq. (16).
11 Update $diag(D_R)$ with the values that were used to normalize each row of the activation matrix $\mathbf{A}_{\mathbf{R}_I}$ using Eq. (17).
12 Repeat steps 3–11 until the algorithm converges (or until the maximum number of iterations M is reached).
13 Compute magnitude estimated spectrograms $\hat{\mathbf{X}}_{R_I}$ and $\hat{\mathbf{X}}_{W_I}$ using Eq. (18).

return $\hat{\mathbf{X}}_{W_I}$

Fig. 6 shows a comparison between the estimated wheezing spectrograms $\hat{\mathbf{X}}_W$ (stage I) and $\hat{\mathbf{X}}_{W_I}$ (stage II). It can be seen that $\hat{\mathbf{X}}_{W_I}$ is able to more reliably represent the spectral trajectories formed only by the energy of the wheeze sounds, almost completely eliminating the normal respiratory sounds compared to $\hat{\mathbf{X}}_W$. The facts that motivate this promising improvement shown by $\hat{\mathbf{X}}_{W_I}$ are due to: (i) the modeling of WS using information from

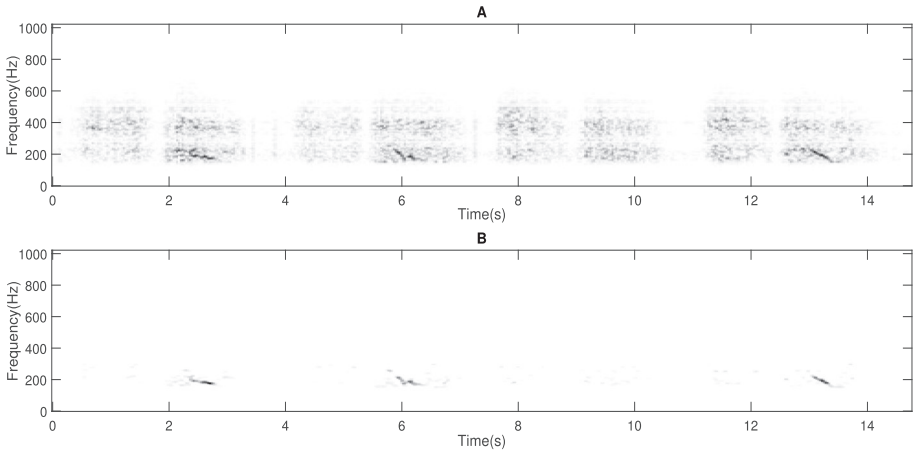

Fig. 6. The estimated wheezing spectrograms related to the unhealthy subject previously shown in Fig. 1A. A) $\hat{\mathbf{X}}_W$ from stage I. B) $\hat{\mathbf{X}}_{W_l}$ from stage II.

$\mathbf{B}_{\mathbf{W_l}}$ and BOI; (ii) the integration of the monophonic constraint into the tonal semi-supervised NMF approach that more accurately models the spectral trajectories generated by WS as they occur in the real world.

2.4. Stage III: wheezing presence/absence classification

As previously mentioned, WS can be considered pitched sounds that display spectral trajectories over time. The main goal of this stage is to define the subject's condition, that is, healthy or unhealthy motivated by the presence or absence of WS. For that purpose, we intend to take advantage of the temporal smoothness information extracted from the most significant spectral peaks forming the previous trajectories factorized in the estimated wheezing spectrogram $\hat{\mathbf{X}}_{W_l}$. The authors assume that the smoothness modelled by spectral trajectories from WS found in $\hat{\mathbf{X}}_{W_l}$ is smoother compared to RS. The reason is because the spectral trajectories from WS tend to show an organized structure in the time–frequency domain (see Fig. 7A) but the spectral peaks that compose the spectral trajectories from RS tend to be randomly distributed in the time–frequency domain without showing any organized structure due to the noisy nature shown by RS (see Fig. 7B). Next, this stage is detailed which is divided into two steps.

2.4.1. Step I: determining the optimal candidate for wheezing temporal interval

The proposed method is developed to analyze any mixture signal $x(t)$ independently of its temporal length. As a consequence to minimize its computational cost, it is only analyzed the optimal candidate OCW for wheezing temporal interval in which the probability of finding wheezing content is maximum. Considering all WS detected in $\hat{\mathbf{X}}_{W_l}$, OCW can be defined as the wheezing that shows the longest duration and the highest energy, with the time duration being more of a priority than energy. In the event that several candidates CW for wheezing intervals have the same temporal duration, the one with the highest energy is chosen as OCW. For this purpose, the spectral energy $\xi(t)$ from $\hat{\mathbf{X}}_{W_l}$ is calculated frame-by-frame as follows,

$$\xi(t) = \sum_{f=1}^{F} \hat{X}_{W_{l_{f,t}}} \ , t = 1 \ldots T \tag{19}$$

A threshold process is used to define all CW detected in $\hat{\mathbf{X}}_{W_l}$ that contain most of the spectral energy. Specifically, the thresholding ζ_t is based on the Otsu algorithm [44] since it allows to differentiates two groups in a histogram, in this case, intervals candidate for wheezing (CW) and intervals not candidates for wheezing. Next, the candidate CW with the longest duration will be selected as OCW. Once OCW has been chosen, the estimated wheezing spec-

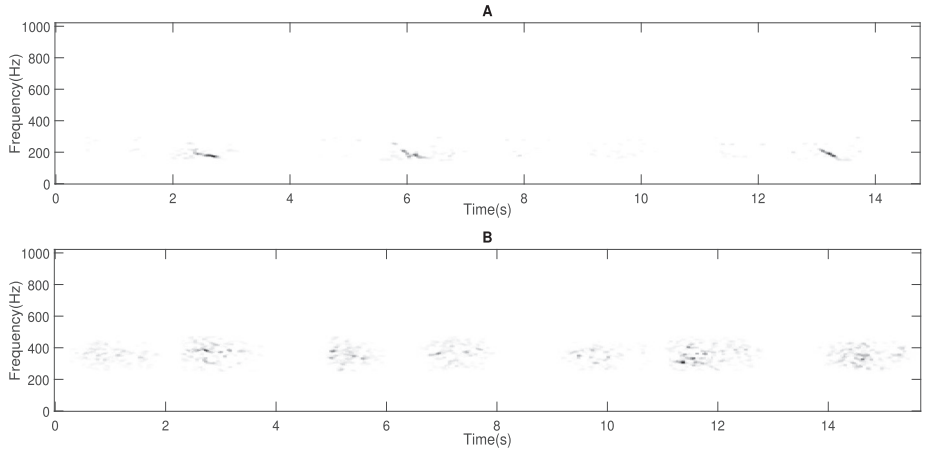

Fig. 7. Comparison between the estimated wheezing spectrograms $\hat{\mathbf{X}}_{W_l}$. A) $\hat{\mathbf{X}}_{W_l}$ related to the unhealthy subject shown in Fig. 1A. B) $\hat{\mathbf{X}}_{W_l}$ related to the healthy subject shown in Fig. 1B.

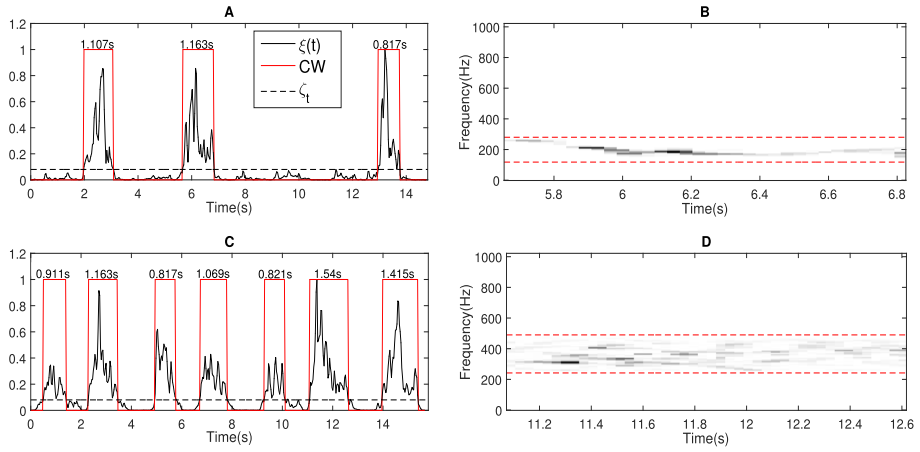

Fig. 8. Estimation of the candidate intervals CW, from the spectrogram $\hat{\mathbf{X}}_{W_l}$, that compete to be the optimal candidate interval OCW. A) CW obtained in the unhealthy subject previously shown in Fig. 1A. B) The spectrogram **Y** related to the OCW chosen from Fig. 8A that shows the highest probability to contain WS. It can be observed that OCW is located in the time interval [5.67s-6.81s]. C) CW obtained in the healthy subject previously shown in Fig. 1B. D) The spectrogram **Y** related to the OCW chosen from Fig. 8C that shows the highest probability to contain WS. It can be observed that OCW is located in the time interval [11.09s-12.61s]. Note that $\xi(t)$ has been normalized to adjust the values between 0 and 1. Both Fig. 8B and Fig. 8D, the red dotted lines display the BOI spectral boundaries associated to each selected interval. Specifically, BOI = [118 Hz–280 Hz] for Fig. 8B and BOI = [242 Hz–490 Hz] associated to Fig. 8D.

trogram $\hat{\mathbf{X}}_{W_l}$ will be limited by the temporal boundaries belonging to OCW in the time domain and the frequency limits belonging to BOI in the frequency domain to locate the portion **Y** of the estimated wheezing spectrogram $\hat{\mathbf{X}}_{W_l}$ where the probability of finding wheezing spectral trajectories is maximum (see Fig. 8).

2.4.2. Step II: smoothness related to the spectral trajectories contained in OCW

To measure the regularity of each spectral trajectory in the time-frequency domain, we propose to determine the spectral smoothness from the set of the most significant spectral peaks detected frame-by-frame. Specifically, the spectral smoothness is obtained calculating the degree of fluctuation σ_i of each spectral trajectory i normalized by the estimated BOI as follows,

$$\sigma_i = \frac{\sum_{z=1}^{Z}\left(S_{i,z} - S_{i,z+1}\right)^2}{BOI} \tag{20}$$

where Z represents the number of frames belonging to the spectrogram **Y** and S is the frequency of the peaks that describe the spectral trajectory i along frames. Empirical results report that the metric σ increases proportionally as the spectral trajectory is more irregular or less smooth. Otherwise, the metric σ decreases proportionally as the spectral trajectory is more regular or smoother in the time-frequency domain. Due to WS are characterized as pitched sounds, we assume that each wheeze sound can be composed by more than one frequency (harmonicity), specifically, by a maximum number of three spectral peaks [45] at each frame: S_1 (lowest frequency), S_2 (intermediate frequency) and S_3 (highest frequency). In order to include the harmonic nature of the WS in the proposed method, we propose to measure the smoothness of each spectral trajectory by means of the set of spectral peaks S_1, S_2 and S_3 as follows,

$$\sigma_G = \sigma_1 + \sigma_2 + \sigma_3 \tag{21}$$

where σ_1, σ_2 and σ_3 denote the degree of fluctuation of each spectral trajectory formed by the previous set of spectral peaks S_1, S_2 and S_3 that have been detected at each frame. When only a single peak is detected in a frame, we assume that this peak refers to the peak of the lowest frequency S_1 since it is common that WS concentrate most of the energy in the peak that show the minimum frequency in a multipitch structure. As a consequence, only frames with two and three spectral peaks detected will be considered for the calculation of σ_2 and σ_3, respectively. Fig. 9 shows the spectral trajectories related to the unhealthy and healthy subject in Fig. 1. In the case of the unhealthy subject (see Fig. 9A), only a single trajectory was detected composed of the spectral peaks S_1, typically found in WS. Besides, it is demonstrated that the continuity of the wheezing trajectories display high smoothness. However, in the case of the healthy subject (see Fig. 9B), three spectral trajectories motivated by the set of peaks of type S_1, S_2 and S_3 were found. It confirms that RS exhibit a high irregularity or discontinuity in time–frequency domain, in other words, a low smoothness because RS tend to locate randomly the peaks in the spectral domain due to the noisy nature of RS.

Our empirical analysis, reported the following facts: (i) the spectral trajectories in the case of unhealthy subjects are distinguished by the continuous nature of WS and the value of $\sigma_G \leq 1$; (ii) the spectral trajectories in the case of healthy subjects are distinguished by the noisy nature of RS and the value of $\sigma_G \gg 1$. For this reason, we set the threshold $\zeta_h = 1$ in order to determine the subject's condition as follows,

$$\text{subject's condition} = \begin{cases} unhealthy & \text{if } \sigma_G \leqslant \zeta_h \\ healthy & \text{if } \sigma_G > \zeta_h \end{cases} \tag{22}$$

In the case of WS compose by more than one spectral peak, the value of the metric σ_G is still significantly low compared to those ones provided by RS. This is because the spectral trajectories originated by the peaks S_1, S_2 and S_3 tend to show an organized structure in the time–frequency domain, i.e. each spectral trajectory will be smooth and continuous due to its wheezing nature and as a consequence, the general metric σ_G will be the sum of 3 small values (σ_1, σ_2 and σ_3).

3. Results and discussion

3.1. Datasets

To evaluate the classification performance of the proposed method, six databases P1, T1, T2H, T2M, T2L and T3 have been used which are detailed in Table 1. In total, these databases provide 4107 s of recording, 114 healthy subjects, 94 unhealthy subjects, 2168 respiratory events (a respiratory event is defined as an inspiration or expiration) and 219 wheezes. Databases P1, T1, T2H, T2M and T2L have been created by collecting a lot of recordings of different subjects of the most widely used Internet pulmonary

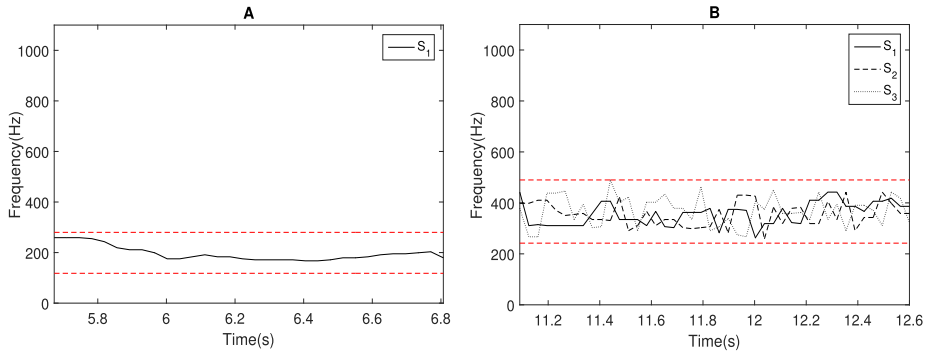

Fig. 9. Spectral trajectories defined by the spectral peaks S_1, S_2, S_3. A) The spectral trajectory ($\sigma_G = 0.45$), related to the unhealthy subject shown in Fig. 8B, that is composed by a single spectral peak S_1. B) The spectral trajectory ($\sigma_G = 98.5$), related to the healthy subject shown in Fig. 8D, that is composed by three types of spectral peaks S_1, S_2 and S_3. Although σ_G in Fig. 9B is significantly higher compared to σ_1 in Fig. 9A, highlight that each smoothness σ_1, σ_2 or σ_3 measured in Fig. 9B is significantly higher compared to the smoothness σ_1 measured in Fig. 9A.

Table 1
Characteristics of each database.

ID1	ID2	ID3	ID4	ID5	ID6	ID7	ID8	ID9	ID10
P1	96	48/48	5–24	1442	[0–9]	[4–16]	784	[1–8]	92
T1	48	48/0	8–49	1051	–	[6–30]	500	–	0
T2H	16	0/16	7–22	251	5	[6–14]	126	[1–5]	41
T2M	16	0/16	7–22	251	0	[6–14]	126	[1–5]	41
T2L	16	0/16	7–22	251	−5	[6–14]	126	[1–5]	41
T3	48	18/30	4–65	861	[2–8]	[4–20]	506	[3–11]	86

ID1: identifier; ID2: number of recordings; ID3: number of recordings captured from healthy/unhealthy subjects; ID4: the shortest and longest duration, in seconds, captured from recordings; ID5: total duration in seconds; ID6: the lowest and highest SNR, in dB, between WS and RS; ID7: the minimum and maximum number of respiratory events found in the recordings; ID8: the total number of respiratory events; ID9: the minimum and maximum number of wheezes found in the recordings; ID10: the total number of wheezes. The fifth column from T1 does not show SNR because it is only composed of healthy subjects. Analogously, the eighth column from T1 does not show the minimum and maximum number of wheezes found because it does not contain wheeze sounds.

repositories [46–58]. The databases collected consisted of sounds captured from the trachea, anterior, and posterior chest; using either a stethoscope or microphone. These were collected from subjects with different pathologies, including asthma, bronchitis, COPD, and croup. Indicate that the age was not considered as a significant variable because the variations caused by age difference in automatic auscultation is too small to be clinically relevant [59]. Database T3 has been directly shared by authors [27].

Using the visual inspection of spectrograms, datasets P1, T2H, T2M and T2L have been created mixing only WS recordings manually separated (by means of a time-frequency mask applied to the mixture spectrogram to select only the bins of each frame corresponding to wheezing) and only RS recordings (in which wheezing is inactive) obtained from [46–58]. The datasets T2H (SNR = 5 dB), T2M (SNR = 0 dB) and T2L (SNR=-5 dB) represent the same dataset T2 but each of them uses a different signal-to-noise ratio (SNR) between WS and RS. After each manual separation for each dataset recording, the power related to WS and RS are calculated and the signal with the highest power is left fixed while the signal with the lowest power is scaled to obtain the desired SNR in order to avoid audio saturation or distortion in the signal scaling process. However, datasets P1 and T3 preserve the original SNR. More details of dataset T3 can be found in [27].

The previous databases have been evaluated with the following purpose:

- Optimization. P1 optimizes the training of the state-of-the-art methods.
- Testing. T1 assesses the robustness of the proposed method in correctly diagnosing healthy subject.
- Testing. T2H, T2M and T2L assess the robustness of the proposed method when WS are masked by RS.
- Testing. T3 assesses the wheezing detection performance of the proposed method in real medical scenario to decide whether the diagnosis of subject is healthy or unhealthy.

In order to validate the results, it should be noted that each file in one of the following databases P1, T1, T2 and T3 has only been used in that database and has not been used again in the rest of the remaining databases (e.g., P1 is not a part of the rest of datasets T1, T2 and T3).

3.2. Experimental setup

A preliminary study show that the following parameters provide the best trade-off between the classification performance and the computational cost: sampling rate $f_s = 2048$ Hz, Hamming window with $N = 256$ samples length and 25% overlap (temporal resolution of 31.3 ms), a discrete Fourier transform using $2N$ points (frequency resolution of 4 Hz). Furthermore, the convergence of the NMF approaches was empirically achieved after 90 iterations for all signals, so $M = 90$ iterations has been selected. Moreover, a number of components $K = 80$ have been used in the NMF approach of the stage I. Finally the optimal weight of the temporal monophonic constraint is $\lambda = 0.05$.

3.3. State-of-the-art methods for comparison

Three recent state-of-the-art wheezing detection methods have been used to evaluate the performance of the proposed method: TSVM [20], MKNN [21] and EEMD [28]. The state-of-the-art methods have been implemented strictly following the instructions shown in their respective references. MKNN, TSVM and EEMD are supervised approaches using the database P1 for the training

of each of them. However, the proposed method does not use any training due to unsupervised approach. In this work, the subject will be diagnosed as unhealthy when MKNN or TSVM locates a wheezing temporal interval $\geqslant$100 ms. Moreover, EEMD has been implemented without limiting the duration of the input mixture.

3.4. Evaluation metrics

The ability to discriminate between healthy and unhealthy subjects has been evaluated by means of five metrics commonly used in the field of wheezing detection [27–29,60]: i) Sensitivity (SE), the ability to correctly classify a subject as unhealthy analyzing only the recordings corresponding to unhealthy subjects; ii) Specificity (SP), the ability to correctly classify a subject as healthy analyzing only the recordings corresponding to healthy subjects; iii) Positive Predictive Value (PPV), the probability that a subject classified as unhealthy truly has the disease derived from wheezing sounds. Here, recordings of both unhealthy and healthy subjects are analyzed; iv) Negative Predictive Value (NPV), the probability that a subject classified as healthy truly does not have the disease derived from wheezing sounds. Here, recordings of both unhealthy and healthy subjects are analyzed; and v) Accuracy (ACC), the ability to correctly classify a subject as healthy or unhealthy.

$$SE = \frac{TP}{TP + FN} \tag{23}$$

$$SP = \frac{TN}{TN + FP} \tag{24}$$

$$PPV = \frac{TP}{TP + FP} \tag{25}$$

$$NPV = \frac{TN}{TN + FN} \tag{26}$$

$$ACC = \frac{(TP + TN)}{(TP + FP + FN + TN)} \tag{27}$$

In these equations, TP (True Positive) represent the number of unhealthy subjects correctly diagnosed, TN (True Negative) represent the number of healthy subjects correctly diagnosed, FP (False Positive) represent the number of healthy subjects misdiagnosed as unhealthy subjects, and FN (False Negative) represent the number of unhealthy subjects misdiagnosed as healthy subjects.

3.5. Results

In order to assess the performance of the proposed method together with the state-of-the-art methods with respect to the temporal length of the input mixture $x(t)$, we propose to evaluate using three levels of analysis: breathing mixture (first level), respiratory cycle (second level) and respiratory stage (third level). Each next level of analysis is composed of a greater number of audio segments from the input mixture being each segment of shorter duration:

- First level: breathing mixture. The entire input mixture $x(t)$ is analyzed at once (see Fig. 10A). The subject will be diagnosed as unhealthy when at least one wheezing is detected anywhere on the input mixture.
- Second level: respiratory cycle. Each respiratory cycle that composes the input mixture $x(t)$ is analyzed as an independent unit (see Fig. 10B). In this case, the subject will be diagnosed as unhealthy when at least one wheezing is detected in one respiratory cycle among all those that compose the input mixture $x(t)$.
- Third level: respiratory stage. Each respiratory stage (inspiration or expiration) that composes each respiratory cycle is analyzed as an independent unit (see Fig. 10C). In this case, the subject will be diagnosed as unhealthy when at least one wheezing is detected in one respiratory stage among all those that compose the input mixture $x(t)$.

Fig. 11 shows SP results evaluating the database T1 between the proposed method and the aforementioned state-of-the-art methods for the three levels of analysis described above. The database T1 can only be evaluated with the metric SP due to the non-existence of TP and FN since T1 is only composed of audio recordings captured from healthy subjects. Results indicate that the proposed method outperforms the correct diagnosis performance of healthy subject compared to the state-of-the art methods for all levels of analysis. Focusing on the first level of analysis (Breathing mixture level), the proposed method accomplishes a significant improvement of approximately 14.58%, 16.67% and 20.83% versus TSVM, MKNN and EEMD, respectively. This fact seems to suggest that the proposed method is more reliable in correctly diagnosing healthy subjects compared to previous baseline methods, showing a promising robustness. The reason is that the proposed method is the only method of those evaluated that has correctly classified as healthy all recordings from healthy subjects belonging to the database T1 that are evaluated at a breathing mixture level, respiratory cycle level and respiratory stage level.

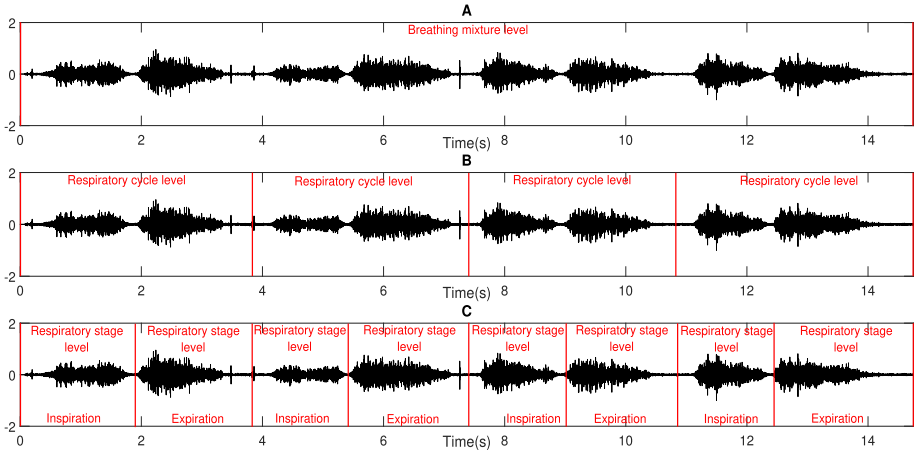

Fig. 10. Different levels of analysis of the input mixture associated to the spectrogram shown in Fig. 1A. A) Breathing mixture level. B) Respiratory cycle level. C) Respiratory stage level. It can be observed a single unit to a breathing mixture level, four units to a respiratory cycle level and finally, eight units to a respiratory stage level.

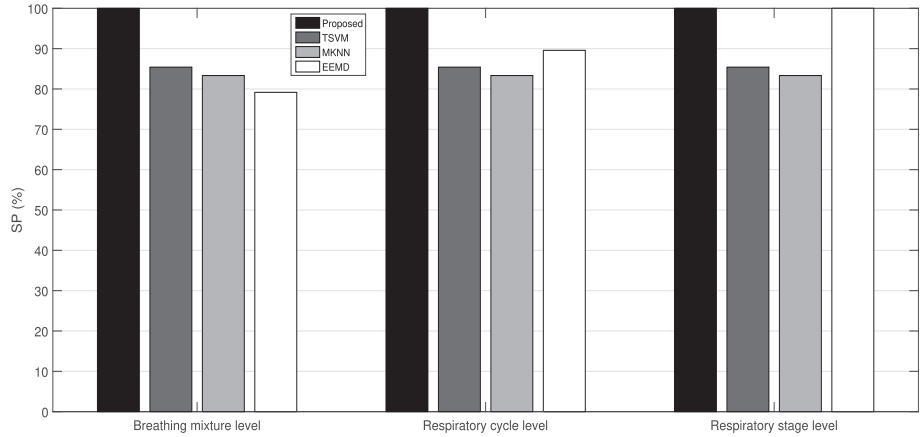

Fig. 11. Comparative SP results between the proposed method and the aforementioned state-of-the-art methods in the database T1 for the three levels of analysis.

It confirms the robustness of the proposed method with respect to the temporal length of the recording to correctly diagnose of healthy subjects. Both, TSVM and MKNN achieve a similar result in the three levels of analysis evaluated but TSVM obtains a slightly better classification performance compared to MKNN. Nevertheless, TSVM and MKNN can be considered robust with the recording duration since are based on the detection of the temporal intervals in which wheezing is active, i.e. MKNN and TSVM determine the wheezing frames and then analyze the duration of each temporal interval detected as wheezing to confirm the presence or absence of them in the recording. In contrast, the performance of the method EEMD highly depends on the length of the recording analyzed. Specifically, the performance of EEMD equals the performance of the proposed method analyzing at respiratory stage level. However, EEMD reduces its classification performance by 20.83% when analyzing at breathing mixture level and 10.42% when analyzing at respiratory cycle level. This behavior is due to the fact that EEMD is based on the extraction of segments that are candidates for being wheezing segments, which are obtained from a set of thresholds that directly depend on the input recording. Specifically, when the duration of the input recording is longer than the respiratory stage, EEMD begins to erroneously detect the wheezing candidate segments that impairs its performance in the classification between presence or absence of WS in breath recordings. While EEMD classifies all the wheeze segments it has extracted, the proposed method only analyzes the segment with the highest probability of finding WS, previously defined as OCW. For this reason, the results of the proposed method are the same in all three levels of analysis, while the EEMD results worsen at the respiratory cycle level and breathing mixture level. The wheezing classification performance against the duration of the recording obtained by the proposed method is a notable advantage, since a correct wheezing diagnosis can be independent of the performance in the detection of the segment at respiratory level, a critical stage for the operation of the EEMD to maximize its classification results.

Fig. 12 shows the SE results by evaluating three databases T2H (SNR = 5 dB), T2M (SNR = 0 dB) and T2L (SNR=−5 dB) with different SNR to compare the robustness of the proposed method in the diagnosis of unhealthy subjects when WS are increasingly masked by RS. The databases T2H, T2M and T2L can only be evaluated with this metric due to the non-existence of TN and FP because these databases are composed exclusively of audio recordings from unhealthy subjects, in other words, each recording contains at least one wheezing. Results show that the proposed method provides the best overall wheezing presence/absence classification results compared to the other evaluated state-of-the-art methods considering all SNR scenarios evaluated for the three levels of analysis. Focusing on the breathing mixture level, it can be observed the following:

- Database T2H: the SE improvement of the proposed method is about 12.5% (TSVM), 12.5% (MKNN) and 18.75% (EEMD).
- Database T2M: the SE improvement of the proposed method is about 18.75% (TSVM), 25% (MKNN) and 25% (EEMD).
- Database T2L: the SE improvement of the proposed method is about 25% (TSVM), 31.25% (MKNN) and 31.25% (EEMD).

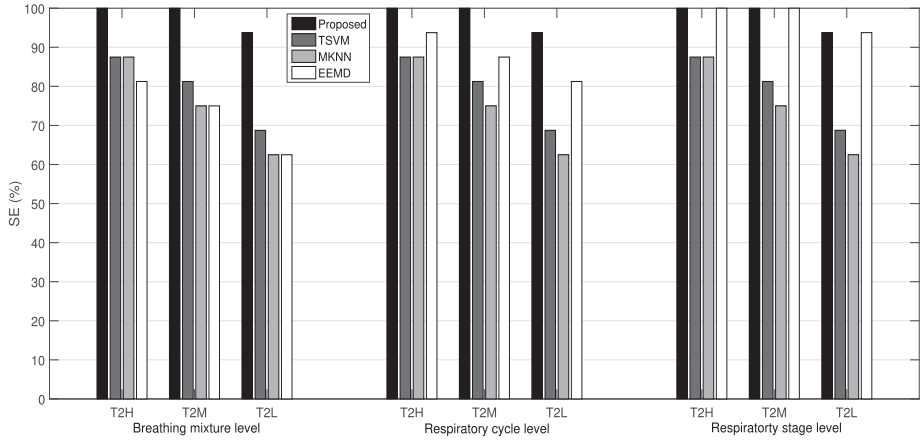

Fig. 12. Comparative SE results between the proposed method and the aforementioned state-of-the-art methods in the databases T2H, T2M and T2L for the three levels of analysis.

Results indicate that the proposed method is more effective and reliable to diagnose unhealthy subjects compared to the other state-of-the-art methods evaluated. It should be emphasized that the proposed method is the only method that correctly classifies, evaluating T2H and T2M, all recordings from unhealthy subjects considering the three levels of analysis. EEMD equals the performance of the proposed method in T2H and T2L but applied in a respiratory stage level. Unlike EEMD, the proposed method again does not need to analyze at the respiratory stage level to maximize its best classification performance. This fact avoids that the performance of the proposed method be dependent on the correct selection of the level of analysis. In addition, it can be clearly seen how EEMD improves its classification performance from Fig. 12 (left) to Fig. 12 (right) as the temporal length of the analyzed recording is reduced to its optimal level, in this case, the respiratory stage level. This is the reason why EEMD obtains worse results compared to the other methods at breathing mixture level but outperforms TSVM and MKNN both respiratory cycle level and respiratory stage level, being this improvement higher at the last level of those mentioned. TSVM achieves the same or better performance than MKNN in any evaluated scenario and analysis level. Highlight that the classification improvement between the proposed method and the rest of the evaluated methods increases as the WS are more difficult to detect because they are masked to a greater extent by RS. This fact seems to demonstrate the greater robustness of the proposed method compared to the assessed state-of-the-art methods. Considering T2L in which an acoustic scenario SNR= −5dB has been modeled, WS are barely audible due to the high interference caused by RS so the wheezing classification process is more complex. Focusing on the decrease in SE results by comparing T2H with T2L at all three levels of analysis, the following can be observed:

- Considering a breathing mixture level: The SE results drop approximately 6.25% (Proposed), 18.75% (TSVM), 25% (MKNN) and 18.75% (EEMD).
- Considering a respiratory cycle level: The SE results drop approximately 6.25% (Proposed), 18.75% (TSVM), 25% (MKNN) and 12.5% (EEMD).
- Considering a respiratory stage level: The SE results drop approximately 6.25% (Proposed), 18.75% (TSVM), 25% (MKNN) and 6.25% (EEMD).

Based on the previous results, it can be said that the wheezing presence/absence classification performance taking into account the proposed method, TSVM and MKNN do not depend on the temporal length of the recording. Although EEMD equals the performance of the proposed method evaluating in a respiratory stage level, only the performance of EEMD worsens significantly from the third level of analysis (respiratory stage level) to the first one (breathing mixture level).

In order to assess the performance of the proposed method in a real medical scenario to detect the presence (unhealthy subject) or absence (healthy subject) of WS, Table 2 shows detailed results evaluating the dataset T3. The proposed method outperforms the classification performance compared to the state-of-the-art methods in terms of SP, SE, ACC, PVP and NVP for the three levels of analysis. The better performance of the proposed method compared to TSVM, MKNN and EEMD suggests that our proposal is competitive to improve the reliability of the subjective diagnosis provided by the physician in the auscultation process to detect the presence of WS. The reason can be due to the fact that the proposed method models common spectro-temporal characteristics that attempt to describe the typical behavior of wheezing sounds as occurs in the nature of the real world. Specifically, a detailed analysis of the proposed method in terms of each evaluation metric indicates the following:

- The performance of the proposed method in terms of SP indicates that 100% of the healthy subjects are correctly classified.
- The performance of the proposed method in terms of SE indicates that 93.3% of the unhealthy subjects are correctly classified.
- The performance of the proposed method in terms of ACC indicates that 95.8% of the healthy and unhealthy subjects are correctly classified.
- The performance of the proposed method in terms of PPV indicates that 100% of the subjects classify as unhealthy subjects are correctly detected and that none of the healthy subjects have been erroneously classified as unhealthy.
- The performance of the proposed method in terms of NPV indicates that 90% of the subjects are correctly classified.

Similarly as occurs in the evaluation of the databases T1 and T2, TSVM and MKNN show their robustness with respect to the temporal length of the recording. TSVM and MKNN are machine learning methods based on the extraction of features and classifiers. However, TSVM obtains better performance compared to MKNN because TSVM consists of two SVM classifiers specifically designed to avoid false positives and false negatives as much as possible. The best performance of the evaluated method is achieved in terms of SP for any level of analysis (the proposed method), in terms of SE for any level of analysis (MKNN and TSVM) and in terms of SE at the first level of analysis and in terms of SP at the second and third levels of analysis (EEMD). The proposed method outperforms the second best state-of-the-art method EEMD in terms of ACC (16.6% in the first level of analysis, 10.4% in the second level of analysis and 2.1% in the third level of analysis) but this ACC improvement is significantly higher, approximately 25%, compared to MKNN and TSVM for the three levels of analysis. Although Table 2 confirms that EEMD can be considered an efficient method to classify wheezing presence/absence, EEMD results shows a significantly decrease of its classification performance when the level of analysis is not the respiratory stage.

4. Conclusions and future work

In this study, we propose a novel method to automatically detect the presence or absence of wheezing sounds in breath recordings. Three novel contributions are proposed by authors. The first contribution, band of interest (BOI), estimates the spectral

Table 2
Comparative results (%) in order to classify healthy/unhealthy subjects evaluating the database T3 for the three levels of analysis.

	Breathing mixture level					Respiratory cycle level					Respiratory stage level				
Method	SP	SE	ACC	PPV	NPV	SP	SE	ACC	PPV	NPV	SP	SE	ACC	PPV	NPV
Proposed	**100**	**93.3**	**95.8**	**100**	**90**	**100**	**93.3**	**95.8**	**100**	**90**	**100**	**93.3**	**95.8**	**100**	**90**
TSVM	55.5	76.6	68.7	74.2	58.8	55.5	76.6	68.7	74.2	58.8	55.5	76.6	68.7	74.2	58.8
MKNN	50	70	62.5	70	50	50	70	62.5	70	50	50	70	62.5	70	50
EEMD	77.8	80	79.2	85.7	70	88.9	83.3	85.4	92.6	76.2	94.4	**93.3**	93.7	96.5	89.5

Each value in bold indicates the highest value obtained in each column.

range in which the probability to find wheeze sounds is maximum. As a second contribution, a constrained tonal semi-supervised NMF is presented to factorize into spectral patterns that correctly model the tonal nature, by narrowband peaks, shown by WS in the estimated BOI. From the separated wheezing spectrogram, we propose an intuitive method to classify healthy or unhealthy subjects analyzing the temporal smoothness of the spectral trajectories defined by the most significant energy decomposed in the estimated BOI.

The main conclusions derived from the experimental results indicate that:

- The proposed method provides the best overall wheezing presence/absence classification results compared to the other state-of-the-art methods considering all databases evaluated. Results suggest that the proposed method obtains a promising performance to improve the reliability of the subjective diagnosis provided by the physician in the auscultation process.
- As the results confirm, one of the main advantages of the proposed method is the robustness with respect to the temporal length of the recording to correctly classify the subject's condition.
- Similar to what occurs in the real world, identifying the presence of wheezing is much more complex in those acoustic scenarios in which wheeze sounds are barely audible due to the high interference caused by normal breath sounds. Comparing the evaluation of the databases T2H and T2L in the first level of analysis, SE results of the proposed method only drop 6.25% while results obtained by the other state-of-the-art methods drop above 18.75%. It suggest that the proposed method is more robust than baseline methods in the assessment of noisy scenarios, specifically, SNR < 5 dB.
- Unlike the state-of-the-art methods evaluated based on machine learning, the proposed method does not depend on any training dataset but the modelling of common spectro-temporal characteristics typically shown in wheezing sounds to be discriminate from normal breath sounds.

Future work will focus on the design of new NMF constraints that model the time–frequency behaviour of different types of wheezing, specifically monophonic and polyphonic wheezing with the aim of assessing the severity of lung disease caused by the appearance of these wheezing sounds.

Conflict of interest

There is no conflict of interest.

Acknowledgments

The authors would like to thank Dr. Vedran Bilas and Dr. Dinko Oletic for sharing their dataset labelled T3 in this manuscript, and the anonymous reviewers for their helpful and constructive comments that greatly contributed to improving the final version of the paper.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.apacoust.2019.107188.

References

[1] Sarkar M, Madabhavi I, Niranjan N, Dogra M. Auscultation of the respiratory system. Ann Thoracic Med 2015;10(3):158.
[2] Pasterkamp H, Kraman SS, Wodicka GR. Respiratory sounds: advances beyond the stethoscope. Am J Respiratory Crit Med 1997;156(3):974–87.
[3] Lozano-García M, Fiz JA, Martínez-Rivera C, Torrents A, Ruiz-Manzano J, Jané R. Novel approach to continuous adventitious respiratory sound analysis for the assessment of bronchodilator response. PloS One 2017;12(2):e0171455.
[4] Le Cam S, Belghith A, Collet C, Salzenstein F. Wheezing sounds detection using multivariate generalized gaussian distributions. In: IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE; 2009. p. 541–4.
[5] Lin BS, Wu HD, Chen SJ. Automatic wheezing detection based on signal processing of spectrogram and back-propagation neural network. J Healthcare Eng 2015;6(4):649–72.
[6] World Health Organization. Asthma,https://www.who.int/news-room/fact-sheets/detail/asthma [Online. Accessed: 2019-25-07].
[7] World Health Organization, Pneumonia,https://www.who.int/maternal_child_adolescent/news_events/news/2011/pneumonia/en/ [Online. Accessed: 2019-25-07].
[8] Salazar AJ, Alvarado C, Lozano FE. System of heart and lung sounds separation for store-and-forward telemedicine applications. Revista Facultad de Ingeniería Universidad de Antioquia 2012;64:175–81.
[9] Lin BS, Lin BS, Wu HD, Chong FC, Chen SJ. Wheeze recognition based on 2d bilateral filtering of spectrogram. Biomed Eng: Appl, Basis Commun 2006;18(03):128–37.
[10] Jain A, Vepa J. Lung sound analysis for wheeze episode detection, in: 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE; 2008, pp. 2582–2585.
[11] Sovijarvi A, Dalmasso F, Vanderschoot J, Malmberg L, Righini G, Stoneman S. Definition of terms for applications of respiratory sounds. Eur Respiratory Rev 2000;10(77):597–610.
[12] Bokov P, Mahut B, Flaud P, Delclaux C. Wheezing recognition algorithm using recordings of respiratory sounds at the mouth in a pediatric population. Comput Biol Med 2016;70:40–50.
[13] Qiu Y, Whittaker A, Lucas M, Anderson K. Automatic wheeze detection based on auditory modelling. Proc Inst Mech Eng, Part H: J Eng Med 2005;219(3):219–27.
[14] Zhang J, Ser W, Yu J, Zhang T. A novel wheeze detection method for wearable monitoring systems. In: International Symposium on Intelligent Ubiquitous Computing and Education, IEEE; 2009. p. 331–34.
[15] Aydore S, Sen I, Kahya YP, Mihcak MK. Classification of respiratory signals by linear analysis. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE; 2009. p. 2617–620.
[16] Emrani S, Gentimis T, Krim H. Persistent homology of delay embeddings and its application to wheeze detection. IEEE Signal Process Lett 2014;21(4):459–63.
[17] Kochetov K, Putin E, Azizov S, Skorobogatov I, Filchenkov A. Wheeze detection using convolutional neural networks. In: EPIA Conference on Artificial Intelligence. Springer; 2017. p. 162–73.
[18] Jin F, Krishnan S, Sattar F. Adventitious sounds identification and extraction using temporal-spectral dominance-based features. IEEE Trans Biomed Eng 2011;58(11):3078–87.
[19] Wisniewski M, Zielinski TP. Joint application of audio spectral envelope and tonality index in an e-asthma monitoring system. IEEE J Biomed Health Inf 2015;19(3):1009–18.
[20] Mazić I, Bonković M, Džaja B. Two-level coarse-to-fine classification algorithm for asthma wheezing recognition in children's respiratory sounds. Biomed Signal Process Control 2015;21:105–18.
[21] Shaharum SM, Sundaraj K, Aniza S, Palaniappan R, Helmy K. Classification of asthma severity levels by wheeze sound analysis. In: IEEE Conference on Systems, Process and Control (ICSPC). IEEE; 2016. p. 172–6.
[22] Bahoura M, Pelletier C. Respiratory sounds classification using gaussian mixture models. In: Canadian Conference on Electrical and Computer Engineering, vol. 3, IEEE; 2004. p. 1309–312.
[23] Mayorga P, Druzgalski C, Morelos R, Gonzalez O, Vidales J. Acoustics based assessment of respiratory diseases using gmm classification. In: Annual International Conference of the IEEE Engineering in Medicine and Biology, IEEE; 2010. p. 6312–316.
[24] Riella R, Nohama P, Maia J. Method for automatic detection of wheezing in lung sounds. Braz J Med Biol Res 2009;42(7):674–84.
[25] Taplidou SA, Hadjileontiadis LJ. Wheeze detection based on time-frequency analysis of breath sounds. Comput Biol Med 2007;37(8):1073–83.
[26] Mendes L, Vogiatzis I, Perantoni E, Kaimakamis E, Chouvarda I, Maglaveras N, Tsara V, Teixeira C, Carvalho P, Henriques J, et al. Detection of wheezes using their signature in the spectrogram space and musical features. In: 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE; 2015. pp. 5581–584.
[27] Oletic D, Bilas V. Asthmatic wheeze detection from compressively sensed respiratory sound spectra. IEEE J Biomed Health Inf 2018;22(5):1406–14.
[28] Lozano M, Fiz JA, Jané R. Automatic differentiation of normal and continuous adventitious respiratory sounds using ensemble empirical mode decomposition and instantaneous frequency. IEEE J Biomed Health Inf 2015;20(2):486–97.
[29] Torre-Cruz J, Canadas-Quesada F, Carabias-Orti J, Vera-Candeas P, Ruiz-Reyes N. A novel wheezing detection approach based on constrained non-negative matrix factorization. Appl Acoust 2019;148:276–88.
[30] Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature 1999;401(6755):788–91.
[31] Canadas-Quesada F, Ruiz-Reyes N, Carabias-Orti J, Vera-Candeas P, Fuertes-Garcia J. A non-negative matrix factorization approach based on spectro-temporal clustering to extract heart sounds. Appl Acoust 2017;125:7–19.

[32] Dia N, Fontecave Jallon J, Gumery PY, Rivet B. Denoising phonocardiogram signals with non-negative matrix factorization informed by synchronous electrocardiogram. In: 26th European Signal Processing Conference (EUSIPCO). IEEE; 2018. p. 51–5.
[33] Févotte C, Bertin N, Durrieu JL. Nonnegative matrix factorization with the itakura-saito divergence: with application to music analysis. Neural Comput 2009;21(3):793–830.
[34] Liutkus A, Fitzgerald D, Badeau R. Cauchy nonnegative matrix factorization. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), IEEE; 2015. p. 1–5.
[35] Laroche C, Kowalski M, Papadopoulos H, Richard G. A structured nonnegative matrix factorization for source separation. In: 23rd European Signal Processing Conference (EUSIPCO), IEEE; 2015. p. 2033–37.
[36] Canadas-Quesada FJ, Vera-Candeas P, Ruiz-Reyes N, Carabias-Orti J, Cabanas-Molero P. Percussive/harmonic sound separation by non-negative matrix factorization with smoothness/sparseness constraints. J Audio, Speech, Music Process 2014;2014(26):1–17.
[37] Kitamura D, Ono N, Saruwatari H, Takahashi Y, Kondo K. Discriminative and reconstructive basis training for audio source separation with semi-supervised nonnegative matrix factorization. In: IEEE International Workshop on Acoustic Signal Enhancement (IWAENC). IEEE; 2016. p. 1–5.
[38] Hurley N, Rickard S. Comparing measures of sparsity. IEEE Trans Inf Theory 2009;55(10):4723–41.
[39] Feng C, Xiao L, Wei Z. Compressive sensing isar imaging with stepped frequency continuous wave via gini sparsity. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS). IEEE; 2013. p. 2063–6.
[40] Toh KKV, Isa NAM. Noise adaptive fuzzy switching median filter for salt-and-pepper noise reduction. IEEE Signal Process Lett 2010;17(3):281–4.
[41] Rafii Z, Pardo B. Repeating pattern extraction technique (repet): a simple method for music/voice separation. IEEE Trans Audio, Speech, Language Process 2013;21(1):73–84.
[42] Prominence criterion of a peak according to the MATLAB software,https://es.mathworks.com/help/signal/ref/findpeaks.html?searchHighlight=findpeak&s_tid=doc_srchtitle#buff2uu [Online. Accessed: 2019-25-07].
[43] Cañadas-Quesada FJ, Vera-Candeas P, Martinez-Munoz D, Ruiz-Reyes N, Carabias-Orti JJ, Cabanas-Molero P. Constrained non-negative matrix factorization for score-informed piano music restoration. Digital Signal Process 2016;50:240–57.
[44] Yuan X, Martínez JF, Eckert M, López-Santidrián L. An improved otsu threshold segmentation method for underwater simultaneous localization and mapping-based navigation. Sensors 2016;16(7):1148.
[45] Pramono RXA, Bowyer S, Rodriguez-Villegas E. Automatic adventitious respiratory sound analysis: a systematic review. PloS One 2017;12(5):e0177926.
[46] The r.a.l.e. repository.http://www.rale.ca [Online. Accessed: 2019-25-07].
[47] Stethographics lung sound samples.http://www.stethographics.com [Online. Accessed: 2019-25-07).
[48] 3m littmann stethoscopes.https://www.3m.com [Online. Accessed: 2019-25-07].
[49] East tennessee state university pulmonary breath sounds,http://faculty.etsu.edu [Online. Accessed: 2019-25-07].
[50] ICBHI 2017 Challenge.https://bhichallenge.med.auth.gr [Online. Accessed: 2019-25-07].
[51] Lippincott NursingCenter.https://www.nursingcenter.com [Online. Accessed: 2019-25-07].
[52] Thinklabs Digital Stethoscope.https://www.thinklabs.com [Online. Accessed: 2019-25-07].
[53] Thinklabs youtube.https://www.youtube.com/channel/UCzEbKuIze4AI1523_AWiK4w [Online. Accessed: 2019-25-07].
[54] Emedicine/Medscape.https://emedicine.medscape.com/article/1894146-overview#a3 [Online. Accessed: 2019-25-07].
[55] E-learning resources.https://www.ers-education.org/e-learning/reference-database-of-respiratory-sounds.aspx [Online. Accessed: 2019-25-07].
[56] Respiratory wiki.http://respwiki.com/Breath_sounds [Online. Accessed: 2019-25-07].
[57] Easy Auscultation.https://www.easyauscultation.com/lung-sounds-reference-guide [Online. Accessed: 2019-25-07].
[58] Colorado State University.http://www.cvmbs.colostate.edu/clinsci/callan/breath_sounds.htm [Online. Accessed: 2019-25-07].
[59] Gross V, Dittmar A, Penzel T, Schuttler F, Von Wichert P. The relationship between normal lung sounds, age, and gender. Am J Respiratory Crit Care Med 2000;162(3):905–9.
[60] Pramono RXA, Imtiaz SA, Rodriguez-Villegas E. Evaluation of features for classification of wheezes and normal respiratory sounds. PloS One 2019;14(3):e0213659.

Paper 4

Wheezing Sound Separation Based on Informed Inter-Segment Non-Negative Matrix Partial Co-Factorizatio

J. De La Torre Cruz, F.J. Cañadas Quesada, N. Ruiz Reyes, P. Vera Candeas and J.J. Carabias Orti, "Wheezing Sound Separation Based on Informed Inter-Segment Non-Negative Matrix Partial Co-Factorizatio", in *Sensors*, Volume 20, May 2020, pp. 26-79. DOI: `https://doi.org/10.3390/s20092679`

- Estado: Publicado.
- Revista: Sensors.
- ISSN: 1424-8220.
- Factor de impacto (JCR 2019): 3.275.
- Cuartiles por área de conocimiento:
  - Engineering, electrical and electronic: Q2, 77/266.
  - Instruments and instrumentation: Q1, 15/66.

sensors

MDPI

Article

Wheezing Sound Separation Based on Informed Inter-Segment Non-Negative Matrix Partial Co-Factorization

Juan De La Torre Cruz *, Francisco Jesús Cañadas Quesada, Nicolás Ruiz Reyes, Pedro Vera Candeas and Julio José Carabias Orti

Departament of Telecommunication Engineering, University of Jaen, Campus Cientifico-Tecnologico de Linares, Avda. de la Universidad, s/n, 23700 Linares, Jaen, Spain; fcanadas@ujaen.es (F.J.C.Q.); nicolas@ujaen.es (N.R.R.); pvera@ujaen.es (P.V.C.); carabias@ujaen.es (J.J.C.O.)

* Correspondence: jtorre@ujaen.es

Received: 13 March 2020; Accepted: 5 May 2020; Published: 8 May 2020

Abstract: Wheezing reveals important cues that can be useful in alerting about respiratory disorders, such as Chronic Obstructive Pulmonary Disease. Early detection of wheezing through auscultation will allow the physician to be aware of the existence of the respiratory disorder in its early stage, thus minimizing the damage the disorder can cause to the subject, especially in low-income and middle-income countries. The proposed method presents an extended version of Non-negative Matrix Partial Co-Factorization (NMPCF) that eliminates most of the acoustic interference caused by normal respiratory sounds while preserving the wheezing content needed by the physician to make a reliable diagnosis of the subject's airway status. This extension, called Informed Inter-Segment NMPCF (IIS-NMPCF), attempts to overcome the drawback of the conventional NMPCF that treats all segments of the spectrogram equally, adding greater importance for signal reconstruction of repetitive sound events to those segments where wheezing sounds have not been detected. Specifically, IIS-NMPCF is based on a bases sharing process in which inter-segment information, informed by a wheezing detection system, is incorporated into the factorization to reconstruct a more accurate modelling of normal respiratory sounds. Results demonstrate the significant improvement obtained in the wheezing sound quality by IIS-NMPCF compared to the conventional NMPCF for all the Signal-to-Noise Ratio (SNR) scenarios evaluated, specifically, an SDR, SIR and SAR improvement equals 5.8 dB, 4.9 dB and 7.5 dB evaluating a noisy scenario with SNR = -5 dB.

Keywords: sound separation; non-negative matrix partial co-factorization; bases; repetitive; sharing; wheezing; normal respiratory sounds; informed; inter-segment

1. Introduction

Chronic Respiratory Diseases (CRDs) can be defined as disorders of the airways and other physiological structures of the respiratory system. One of the most common CRDs is Chronic Obstructive Pulmonary Disease (COPD) that is responsible for more than 3 million deaths of people each year which is equivalent to 6% of all deaths worldwide [1]. COPD is often characterized by the presence of wheeze sounds since wheezes provide relevant clues that alert about a respiratory disorder [2,3]. Although CRDs currently have no medical cure, early detection of wheezing from auscultation can lead to treatment when the disease is in its early stage, thus improving people's quality of life. Although there are other clinical alternatives, such as chest radiography and laboratory analysis, auscultation remains the main technique used in most of the health centers in low-income and middle-income countries to provide the first medical diagnosis of the status of the lung due to its

low cost, safety and non-invasive nature. Nevertheless, this early detection by the physician depends largely on the subjective diagnosis based on both the training and expertise in interpreting what hears with the stethoscope and the vulnerability to normal respiratory sounds that can mask the presence of sounds of interest, such as wheezing [4]. Today, many researchers continue to investigate in biomedical signal processing to enhance the clarity of the wheezing sounds with the aim that all useful medical information contained in the wheezing sound signal is heard in the process of auscultation.

In general terms, the respiratory sounds can be classified into two main categories: normal and abnormal (adventitious, such as wheezes), according to the Computerized Respiratory Sound Analysis (CORSA) guidelines [5]. Although wheeze and normal respiratory sounds appear simultaneously since both of them are generated by the same air flow through the lungs, normal respiratory sounds are always present in each respiratory cycle since they are automatically generated by the breathing process. However, the occurrence of wheezing sounds is random because of the respiratory disorder so they do not have to be present in all breathing cycles. So, normal respiratory sounds (RS) are generated by healthy lungs and they are represented by broadband spectrum where most of the energy is concentrated in the spectral band 60 Hz–1000 Hz [6]. Wheeze sounds (WS) are abnormal sounds, generated by unhealthy lungs that suffer narrowing of airways, superimposed onto the RS. Therefore, WS can be described as pitched and continuous sounds which usually have a fundamental frequency (pitch) located between 100 Hz–1000 Hz with duration longer than 100 ms, displaying spectral trajectories of narrowband spectral peaks [7] as shown in Figure 1. In this work, any single-channel signal composed of both RS and WS will be referred as mixture.

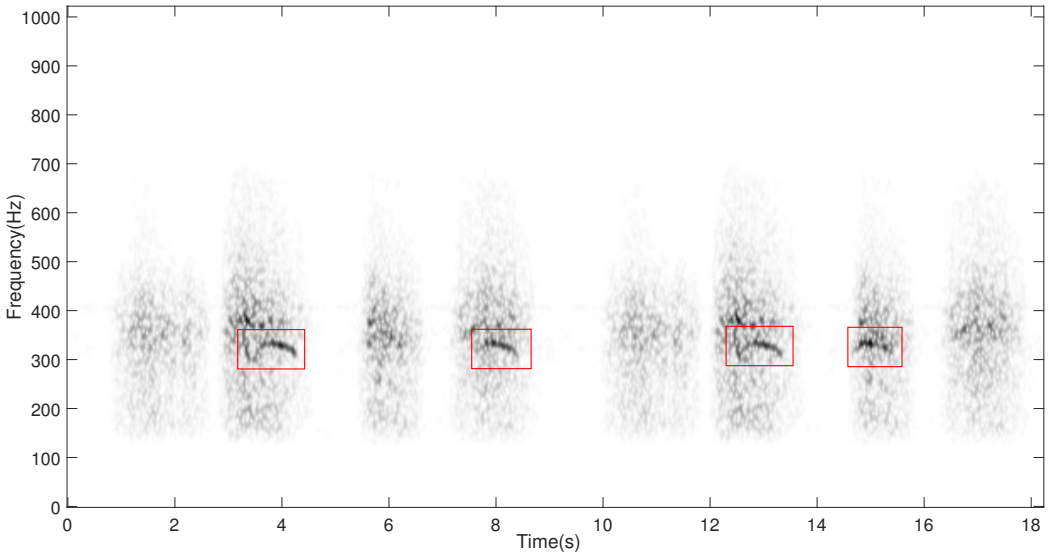

Figure 1. Time-frequency representation of a breathing recording from an unhealthy subject in which four wheezes (red rectangles), mixed with normal respiratory sounds, can be observed. Higher energies are indicated by darker colour.

It is common that the cognitive capacity of the physician is reduced throughout the day as the number of hours spent analyzing respiratory sounds increases, a fact that is exacerbated by the stress to which the physician is subjected to certain medical cases [8,9]. The presence of WS is often associated with obstructions of the airways. However, the interference caused by RS causes the loss of relevant wheezing content in WS which makes it difficult to provide a reliable diagnosis of the status of the lung according to what is being heard through the stethoscope. Sound source separation approaches have been widely applied to overcome this problem by isolating the sounds of interest (target) from those that act as acoustic interference (non-target) [10].

Many biomedical signal processing challenges, such as ambient denoising [11], wheezing detection and classification are still open to the machine learning research community. In [11], a denoising

approach is proposed to remove ambient noise from lung sound recordings by means of an adaptive subtraction method that operates in the spectral domain. Focusing on both wheezing detection and classification tasks, the initial works are based on spectral peaks analysis applying thresholding [2,12–15] that obtain sensitivity/specificity results from 71% to 98%. Like this, Taplidou and Hadjileontiadis [14] proposed a spectro-temporal wheeze detector that automatically locates and identifies wheeze sounds based on spectral trend elimination, separation of the spectrum into frequency bands and peak detection/classification. Most of the wheezing detection and classification approaches are based on the feature extraction and classifier configuration: (i) Musical features and Logistic Regression Model (LRM) [16]; (ii) Spectral features and Support Vector Machine (SVM) such as Power spectral density mean and harmonics [17], Intensity, mean frequency and standard deviation frequency [18], Power spectral band [19], Tonality index [20] and Ensemble Empirical Mode Decomposition (EEMD) [21]; and finally, (iii) Mel Frequency Cepstral Coefficients (MFCC) using K-nearest neighbour (KNN) [22], LRM [23] and Gaussian Mixture Model (GMM) [24], that obtain sensitivity/specificity results from 90% to 99%. Thus, a wheezing detection [20] was developed at the segment level by means of a SVM classifier whose features are the spectral envelope variation and a tonality index. Other works have been focused on the wavelet domain [25,26]. In this context, Ulukaya et al. [26] presented a tunable RAtional Dilation Wavelet Transform (RADWT) based method to discriminate monophonic and polyphonic wheeze sounds by means of localized energy peaks which are calculated from wavelet coefficients. Other studies have applied different types of neural networks (NN) to wheezing sound analysis [27–30] obtaining the best promising performance in terms of sensitivity and specificity results, specifically, from 86% and 100%. Thus, Lin et al. [27] introduce a method that searches for horizontal or nearly horizontal edges of the spectrogram and a back-propagation neural network (BPNN) classifier is applied using features such as, frequency range and the slope of the potential wheeze. However, wheezing detection and classification tasks could be improved applying sound source separation techniques as a preliminary step since these techniques can increase the clarity of the wheezing content hidden in the signal being auscultated. Although very few works [31,32] have addressed in depth the separation of wheezing sound sources to the best of our knowledge, all of them are based on Non-negative Matrix Factorization (NMF) since NMF is a recent and promising tool that can extract hidden sound events with physical interpretation in nature. Specifically, Torre et al. [31] present a constrained NMF approach to separate wheezes from respiratory sounds applied to single-channel mixtures. The proposed constraints, smoothness and sparseness, model common spectral behavior shown by wheezes and normal breath sounds. Results report that the proposed method improves the acoustic quality of the wheezes removing most of the respiratory sounds.

In this paper, an extended version of Non-negative Matrix Partial Co-Factorization (NMPCF) is proposed to suppress RS while preserving the wheezing acoustic content. Here, we assume that RS can be considered as repetitive sound events during breathing so, RS can be modeled by sharing together the spectral patterns found in each respiratory stage (segment), inspiration or expiration, with a respiratory training signal. However, this sharing of patterns can not be applied to wheezes since WS could not be present at each segment due to their unpredictable nature in time motivated by the pulmonary disorder. To improve the sound separation performance of the conventional NMPCF that treats equally all segments of the spectrogram, the main contribution of the proposed method adds higher importance to those segments classified as non-wheezing using inter-segment information informed by a wheezing detection system. As a result, our proposal is able to characterize RS more accurately by forcing to model more on those non-wheezing segments in the bases sharing process into the NMPCF decomposition.

The rest of this paper is structured as follows. First, Section 2 briefly reviews the background of the most relevant approaches based on Non-negative Matrix Factorization and Non-negative Matrix Partial Co-Factorization. Section 3 details the proposed method. Section 4 discusses the evaluation and the experimental results. Finally, conclusions and further research are presented in Section 5.

2. Background

2.1. Non-Negative Matrix Factorization

Non-negative Matrix Factorization (NMF) [33,34] is a rank-reduction method that has been widely applied to learning images [35] and audio [36]. NMF includes the non-negativity constraint to recover hidden patterns of the input data using basis and activation matrices. Considering a monaural input mixture $x(t)$, composed of sources of interest (target) $x_W(t)$ and non-target sources $x_R(t)$, NMF factorizes the input spectrogram $\mathbf{X}$ into the product of two non-negative matrices: basis matrix $\mathbf{U} \in \mathbb{R}_+^{F\times K}$ and activation matrix $\mathbf{V} \in \mathbb{R}_+^{K\times T}$ as shown in Equation (1). We assume an approximate linear additivity between the input spectrograms $\mathbf{X}_W \in \mathbb{R}_+^{F\times T}$ and $\mathbf{X}_R \in \mathbb{R}_+^{F\times T}$. The subscript W is often used to refer the sounds of interest and the subscript R is applied to the sounds that act as acoustic interference,

$$\mathbf{X} = \mathbf{X}_W + \mathbf{X}_R \approx \hat{\mathbf{X}} = \hat{\mathbf{X}}_W + \hat{\mathbf{X}}_R = \mathbf{UV} = \begin{bmatrix}\mathbf{U}_W & \mathbf{U}_R\end{bmatrix}\begin{bmatrix}\mathbf{V}_W \\ \mathbf{V}_R\end{bmatrix} = \mathbf{U}_W\mathbf{V}_W + \mathbf{U}_R\mathbf{V}_R \tag{1}$$

obtaining the estimated spectrograms $\hat{\mathbf{X}} \in \mathbb{R}_+^{F\times T}$, $\hat{\mathbf{X}}_W \in \mathbb{R}_+^{F\times T}$, $\hat{\mathbf{X}}_R \in \mathbb{R}_+^{F\times T}$ with F frequency bins and T frames using K bases and the corresponding time-varying activations. Therefore, $\mathbf{U}$ can be interpreted as a dictionary of spectral bases or patterns that represents the frequency information associated to the target and non-target sources active in the input spectrogram. Instead, $\mathbf{V}$ represents a matrix of activations that indicates the activity of each spectral basis in a given frame.

NMF is often calculated using an iterative algorithm, based on multiplicative update rules [33], to obtain those parameters that reduce the cost function $D(\mathbf{X}|\hat{\mathbf{X}})$ based on penalizing the error reconstruction between $\mathbf{X}$ and $\hat{\mathbf{X}}$. In this paper, the generalized Kullback-Liebler divergence $D_{KL}(\mathbf{X}|\hat{\mathbf{X}})$ [37] has been applied because it confirms the non-negativity of $\mathbf{U}$ and $\mathbf{V}$ as can be observed in Equations (3) and (4). In addition, recent works [32,38] report that $D_{KL}(\mathbf{X}|\hat{\mathbf{X}})$ can be used in biomedical signal processing to achieve promising results,

$$D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}\right) = \mathbf{X}\log\frac{\mathbf{X}}{\hat{\mathbf{X}}} - \mathbf{X} + \hat{\mathbf{X}} \tag{2}$$

$$\mathbf{U}_z \leftarrow \mathbf{U}_z \odot \left(\left(\mathbf{X} \oslash \mathbf{UV}\right){\mathbf{V}_z}^{\mathrm{T}} \oslash \left(\mathbf{1}\mathbf{V}_z^{\mathrm{T}}\right)\right), \quad z = W, R \tag{3}$$

$$\mathbf{V}_z \leftarrow \mathbf{V}_z \odot \left({\mathbf{U}_z}^{\mathrm{T}}\left(\mathbf{X} \oslash \mathbf{UV}\right) \oslash \left({\mathbf{U}_z}^{\mathrm{T}}\mathbf{1}\right)\right), \quad z = W, R \tag{4}$$

where $\mathbf{U}_W \in \mathbb{R}_+^{F\times K_W}$, $\mathbf{U}_R \in \mathbb{R}_+^{F\times K_R}$, $\mathbf{V}_W \in \mathbb{R}_+^{K_W\times T}$ and $\mathbf{V}_R \in \mathbb{R}_+^{K_R\times T}$ are initialized as random positive matrices, $\mathbf{1} \in \mathbb{R}_+^{F\times T}$ represents an all-ones matrix, $^{\mathrm{T}}$ is the transpose operator, $\odot$ is the element-wise multiplication, $\oslash$ is the element-wise division and $K = K_W + K_R$ indicates the number of bases, being K_W the number of bases related to the sounds of interest and K_R the number of bases related to the acoustic interference.

The main drawbacks shown by NMF can be summarized in the following three points: (i) poor signal quality when the iterative algorithm reaches a poor local minimum; (ii) NMF can not reconstruct each source because it does not have enough information to cluster all the bases generated by the same source; (iii) NMF does not guarantee a parts-based objects reconstruction with physical meaning as occurs in nature [39]. To overcome this problem, three approaches have been widely proposed in literature [40]: (i) supervised NMF (SNMF) [41,42] in which $\mathbf{U}_W$ and $\mathbf{U}_R$ are learned in advanced by means of training and fixed during the iterative process. As a result, only the activations matrices $\mathbf{V}_W$ and $\mathbf{V}_R$ are updated; (ii) semi-supervised NMF (SSNMF) [43,44] in which $\mathbf{U}_R$ is learned in advanced by means of training and fixed during the iterative process. As a result, $\mathbf{V}_W$, $\mathbf{V}_R$ and $\mathbf{U}_W$ are updated;

and (iii) constrained NMF (CNMF) in which no training is used because different constraints are included into the factorization procedure to model the specific time-frequency characteristics of the sources to extract [45,46].

To sum up, SNMF, SSNMF and CNMF find better solutions compared to NMF since all of them model, into the bases or activations obtained from the factorization, temporal or spectral behaviors shown by the sounds, that are intended to be recovered, in nature. Nevertheless, the main disadvantages observed in both SNMF and SSNMF are the following: (i) highly dependent of the training data so, the separation performance is limited to the spectral similarity between the training and sounds contained in the input mixture and; (ii) there may not be public training databases available. On the other hand, the main disadvantage observed in constrained NMF approaches, such as CNMF is the difficulty of mathematically defining both the constraints that correctly model the temporal and spectral behaviors shown by the target sources and their incorporation into the cost function on which the factorization is based [47].

2.2. Non-Negative Matrix Partial Co-Factorization

Non-negative Matrix Partial Co-Factorization (NMPCF) has been used in several audio processing tasks, such as extraction of rhythmic sources [48–50], singing-voice separation [51] or speaker diarization [52]. The main idea of NMPCF is to apply a joint matrix factorization using multiple input matrices to obtain a set of shared spectral bases or temporal activations.

In general, NMPCF-based methods can be classified into four approaches: (i) semi-supervised factorization (1S-NMPCF) [50] in which a joint decomposition, considering the input mixture and a training matrix related to repetitive sounds, is performed by sharing some bases active in both of them [48]; (ii) supervised factorization (2S-NMPCF) in which a joint decomposition, considering the input mixture and two training matrices related to repetitive and non-repetitive sounds, is performed by sharing some bases active between each training matrix and the input mixture [51]; (iii) unsupervised factorization (T-NMPCF) [50] in which a joint decomposition using multiple shorter segments from the input mixture is obtained factorizing them into repetitive sound events by finding common bases across segments [49]; and (iv) semi-supervised factorization (ST-NMPCF) [50] in which a joint decomposition of the input mixture is performed using a training matrix associated to repetitive sound events and multiple shorter segments to make advantage of both spectral and temporal modelling of repetitive sounds.

However, NMPCF-based approaches treat all segments of the input mixture decomposition together equally, ignoring the importance of each specific segment in the modelling of the repetitive and non-repetitive sounds. As a result, it could be interesting to investigate how to include the importance of different segments according their spectral content to weight the spectral modelling of the repetitive sounds in the joint factorization and as a consequence, to improve the separation quality of the sounds of interest.

3. Proposed Method

The aim of the proposed method is to enhance the quality of the WS by removing the RS that implicitly appear in the human breathing process. In order to improve the separation performance between WS and RS of the NMF-based and NMPCF-based baseline methods, we propose a modified NMPCF approach denominated Informed Inter-Segment Non-negative Matrix Partial Co-Factorization (IIS-NMPCF) that adds higher importance into the NMPCF factorization to those segments in which WS are not present. For this purpose, IIS-NMPCF consists of three stages: (i) Segmentation; (ii) Classification between presence/absence of WS and finally (iii) Adding weighting into the NMPCF decomposition. The flowchart of the proposed method is shown in Figure 2, and details are depicted in the following Sections 3.1 and 3.2.

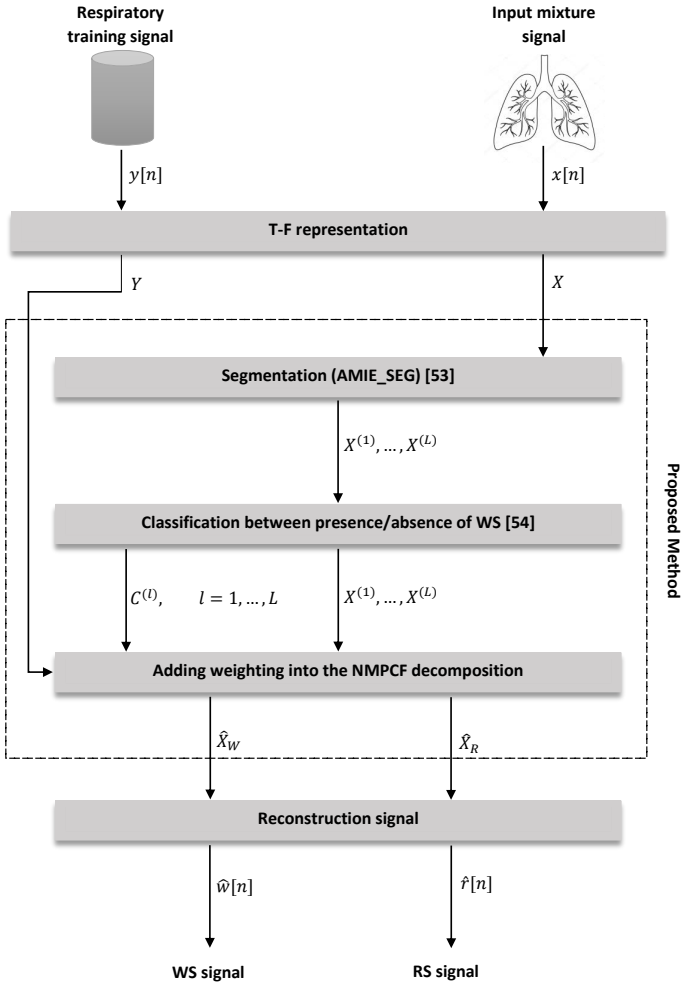

Figure 2. Flowchart of the proposed method IIS-NMPCF.

3.1. Time-Frequency Signal Representation

Let $x[n]$ denote the n-th sample of a mixture signal, which consists of the sum of wheezing $w[n]$ and normal respiratory sounds $r[n]$. The magnitude spectrogram $\mathbf{X}$ of a mixture signal $x[n]$ can be represented as $\mathbf{X} = \mathbf{X}_W + \mathbf{X}_R$, being $\mathbf{X}_W$ the magnitude spectrogram of only WS and $\mathbf{X}_R$ the magnitude spectrogram of only RS. Each unit $X_{f,t}$ is defined by the f-th frequency bin at the t-th frame and is calculated from the magnitude of the Short-Time Fourier Transform (STFT) using a Hamming window of N samples with 25% overlap. A normalization process is applied in order to ensure that the proposed method can be independent of the size and scale of the magnitude spectrogram $\mathbf{X}$. To avoid complex nomenclature throughout the paper, the variable $\mathbf{X}$ is hereinafter referred to the normalized magnitude spectrogram $\overline{\mathbf{X}}$ computed as follows,

$$\overline{\mathbf{X}} = \frac{\mathbf{X}}{\left(\frac{\sum_{f,t} X_{f,t}}{FT}\right)} \tag{5}$$

Besides, $y[n]$ denote the n-th sample of the respiratory training signal, which consists of a concatenation of different respiratory stages composed only of RS (for more details see Section 4.3). The magnitude spectrogram $\mathbf{Y}$ of the respiratory training signal $y[n]$ has been calculated following the same procedure used with the previous magnitude spectrogram $\mathbf{X}$.

3.2. Wheezing Sound Separation Using Informed Inter-Segment NMPCF

The key assumptions behind the proposed method IIS-NMPCF to apply WS and RS source sound separation are the following:

(i) RS are often characterized by similar spectral patterns that represent a wideband noise spectrum showing time and frequency smoothness [32]. In this way, $\mathbf{Y}$ can be useful to replicate these similar RS spectro-temporal behaviors observed in most of the subjects.

(ii) In addition, RS can be considered as repetitive events in human breathing so, RS can be modeled sharing common spectral patterns that can be found throughout all breathing stages (segments), that is, some basis vectors can be shared during the inter-segment analysis due to the repeatability of RS. If we divide the input mixture spectrogram $\mathbf{X}$ into segments $\mathbf{X}^{(1)}$,$\mathbf{X}^{(2)}$, $\cdots$,$\mathbf{X}^{(L)}$, we can get L-segments from the given mixture $x[n]$ that share common spectral patterns. For this purpose, we have used AMIE_SEG [53] that automatically allows to segment the mixture spectrogram $\mathbf{X}$ into inspiratory and expiratory stages.

(iii) However, WS can be present or absent in the respiratory stages due to the pulmonary disorder. Therefore, we can define an indicator $C^{(l)}$ to distinguish between non-wheezing ($C^{(l)} = 0$) and wheezing ($C^{(l)} = 1$) segments. Note that the term $^{(l)}$ refers to the segment identifier $l = 1, \ldots, L$ of the mixture spectrogram $\mathbf{X}$. In the case of wheezing segments, the spectral patterns of both RS and WS are present. For this reason, we propose to weight the importance of wheezing and non-wheezing segments into the conventional NMPCF decomposition to improve the wheezing sound separation performance. The classification between non-wheezing and wheezing segments is provided by a wheezing detection algorithm previously developed by authors [54].

Considering two input spectrograms $\mathbf{X}$ and $\mathbf{Y}$, the factorization of the conventional ST-NMPCF lets the common respiratory basis vectors $\mathbf{U}_R$ be shared jointly between the spectrogram $\mathbf{Y}$ and L-segments $\mathbf{X}^{(1)}$,$\mathbf{X}^{(2)}$, $\cdots$,$\mathbf{X}^{(L)}$ of the input spectrogram $\mathbf{X}$ (see Figure 3),

$$\mathbf{X}^{(l)} \approx \hat{\mathbf{X}}^{(l)} = \hat{\mathbf{X}}_R^{(l)} + \hat{\mathbf{X}}_W^{(l)} = \mathbf{U}_R diag(Dx_R^{(l)})\mathbf{V}_R^{(l)} + \mathbf{U}_W^{(l)} diag(Dx_W^{(l)})\mathbf{V}_W^{(l)} \tag{6}$$

$$\mathbf{Y} \approx \hat{\mathbf{Y}} = \mathbf{U}_R diag(Dy_R)\mathbf{H}_R \tag{7}$$

where $\hat{\mathbf{X}}$, $\hat{\mathbf{Y}}$ are the estimated or reconstructed spectrograms of the input mixture and the respiratory training signal; $\hat{\mathbf{X}}_R$, $\hat{\mathbf{X}}_W$ are the estimated spectrograms of the RS and WS; $\mathbf{U}_R$, $\mathbf{U}_W$ are the estimated basis matrices of the RS and WS; $\mathbf{V}_R$, $\mathbf{V}_W$ are the estimated activation matrices of the RS and WS for the mixture; $\mathbf{H}_R$ is the estimated activation matrix of the RS for the respiratory training signal. All of these matrices are non-negative matrices. The number of respiratory and wheezing components will be denoted as K_R and K_W, respectively. The L^2-norm of each column of $\mathbf{U}_R$ or $\mathbf{U}_W$ is equal to 1.0. The terms Dx_R and Dx_W represent vectors with the L^2-norm of each activation component of RS and WS, respectively. Similarly, the term Dy_R represents a vector with the L^2-norm of each activation component of RS. Therefore, the L^2-norm of each row of $\mathbf{V}_R$, $\mathbf{V}_W$ or $\mathbf{H}_R$ be equal to 1.0 due to the normalization procedure at each iteration. The operator $diag()$ is the diagonal matrix.

Figure 3 depicts those models with L-segments of the mixture spectrogram $\mathbf{X}$ and the respiratory training spectrogram $\mathbf{Y}$. As mentioned in the key assumption (i), Equation (7) models the respiratory training reconstruction by letting the estimated basis matrix $\mathbf{U}_R$ to contain spectral patterns that define the common behavior of RS. As mentioned in the key assumption (ii), Equation (6) aims to learn the common basis vectors $\mathbf{U}_R$ of L-segments $\mathbf{X}^{(1)}$,$\mathbf{X}^{(2)}$, $\cdots$,$\mathbf{X}^{(L)}$ to model repetitive spectral components throughout the segments, since RS can be considered as repetitive sound events in human breathing. On the other hand, $\mathbf{U}_W^{(l)}$ is responsible for recovering WS that can be contained in each segment. Combining the two previous factorization models, $\mathbf{U}_R$ can model both spectral characteristics of the respiratory training $\mathbf{Y}$ and temporally repeating components belonging to the segments $\mathbf{X}^{(1)}$,$\mathbf{X}^{(2)}$, $\cdots$,$\mathbf{X}^{(L)}$. Considering the previous assumption (iii), the main contribution of

the proposed method is to give greater importance, by means of weighting, to those segments classified as non-wheezing ($C^{(l)} = 0$) in the NMPCF decomposition to learn more accurate the common basis vectors $\mathbf{U}_R$ since these segments will not be interfered by WS so, the spectral modelling of RS will be more acoustically reliable. In Figure 3, the segments $\mathbf{X}^{(2)}$ and $\mathbf{X}^{(L)}$ are classified as non-wheezing segments.

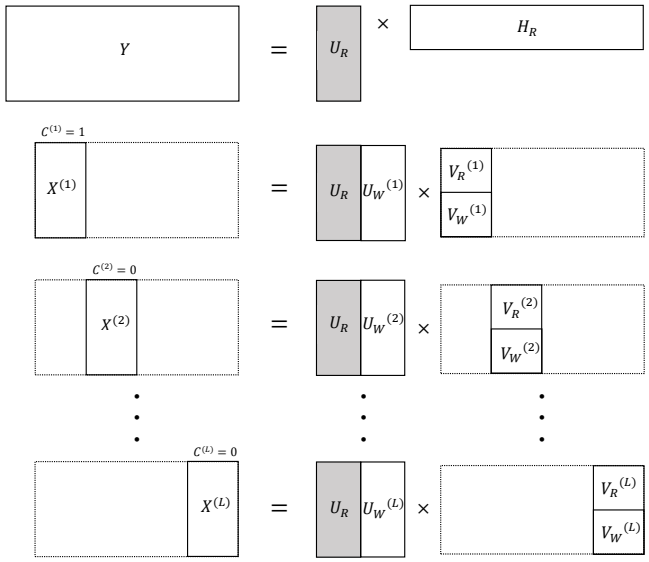

Figure 3. Pictorial illustration of the matrix decomposition based on IIS-NMPCF.

The objective function of the proposed method IIS-NMPCF can be constructed to minimize the residuals of the models (6) and (7),

$$\Gamma_{IIS-NMPCF} = \overbrace{\sum_{l=1}^{L}\left[\lambda^{C^{(l)}} D_{KL}\left(\mathbf{X}^{(l)}|\hat{\mathbf{X}}^{(l)}\right)\right] + LD_F\left(\mathbf{U}_R|\mathbf{0}\right) + \sum_{l=1}^{L}\left[D_F\left(\mathbf{U}_W^{(l)}|\mathbf{0}\right)\right]}^{\text{Objective function applied to the set of } L\text{-segments of the input mixture}} + \overbrace{\alpha D_{KL}\left(\mathbf{Y}|\hat{\mathbf{Y}}\right) + D_F\left(\mathbf{U}_R|\mathbf{0}\right)}^{\text{Objective function applied to the respiratory training}} \quad (8)$$

where $D_{KL}\left(\right)$ is the Kullback–Leibler divergence used to calculate the signal reconstruction error for each segment $D_{KL}\left(\mathbf{X}^{(l)}|\hat{\mathbf{X}}^{(l)}\right)$ and the respiratory training spectrogram $D_{KL}\left(\mathbf{Y}|\hat{\mathbf{Y}}\right)$. The penalization term $D_F\left(\right)$ represents the Frobenius norm applied to each basis matrix in order to prevent basis vectors from convergence to too small values [50]. The weighting factor $\lambda^{C^{(l)}}$ controls the relative importance of each segment matrix $\mathbf{X}^{(l)}$ depending on the type of segment, wheezing ($C^{(l)} = 1$) or non-wheezing ($C^{(l)} = 0$), in the factorization model. The weighting factor α controls the relative importance of the respiratory training matrix $\mathbf{Y}$ in the factorization model.

Highlight that the weighting factor $\lambda^{C^{(l)}}$ plays a crucial role in the proposed method. The reason is because $\lambda^{C^{(l)}}$ controls the importance of which segments are more relevant in the modelling of the spectral patterns related to RS, specifically, those segments in which WS are not detected. Therefore, the following considerations about the parameter $\lambda^{C^{(l)}}$ must be taken into account:

(a) According to the estimated basis matrix $\mathbf{U}_R$ or $\mathbf{U}_W^{(l)}$, the weighting factor $\lambda^{C^{(l)}}$ can be classified as $\lambda_R^{C^{(l)}}$ or $\lambda_W^{C^{(l)}}$, respectively. As mentioned above, WS are always overlapped with RS so, we assume that none of the segments will model the behaviour of WS better than another. However, RS can be found isolated in some segments of human breathing due to the unpredictable nature

of the pulmonary disorder. In this case, those segments in which WS are not contained will be more relevant to model the behaviour of RS. In this manner, $\lambda_W^{C^{(l)}}$ will set the same value for all segments, that is, $\lambda_W^{C^{(l)}} = \lambda_W, l = 1, \ldots, L$ and $\lambda_R^{C^{(l)}}$ will be variable depending on the type of segment, wheezing ($C^{(l)} = 1$) or non-wheezing ($C^{(l)} = 0$), is analyzed. In addition, the value assigned to the weighing factors must satisfy $\lambda_R^{C^{(l)}} > \lambda_W$ (see Section 4.4) since RS are always present in all segments of the input mixture and WS may not be.

(b) Focusing on the type of segment indicated by the parameter $C^{(l)}$, the weighting factor $\lambda_R^{C^{(l)}}$ can be classified as λ_R^0 or λ_R^1. The parameter λ_R^0 is associated with the non-wheezing segments ($C^{(l)} = 0$) and λ_R^1 is associated with the wheezing segments ($C^{(l)} = 1$). This allows to give greater importance to non-wheezing segments for the modeling of respiratory basis $\mathbf{U}_R$. As consequence, the value assigned to the weighing factors must satisfy $\lambda_R^0 > \lambda_R^1$ (see Section 4.4).

Given the above, the estimated basis matrices $\mathbf{U}_R$, $\mathbf{U}_W^{(l)}$ and activations matrices $\mathbf{V}_R^{(l)}$, $\mathbf{V}_W^{(l)}$, $\mathbf{H}_R$ can be obtained by applying a gradient descent algorithm based on multiplicative update rules as follows,

$$\mathbf{U}_R \leftarrow \mathbf{U}_R \odot \frac{\sum_{l=1}^{L} \left[\lambda_R^{C^{(l)}} \left(\mathbf{X}^{(l)} \oslash \hat{\mathbf{X}}^{(l)} \right) \left(diag \left(Dx_R^{(l)} \right) \mathbf{V}_R^{(l)} \right)^T \right] + \alpha \left(\mathbf{Y} \oslash \hat{\mathbf{Y}} \right) \left(diag \left(Dy_R \right) \mathbf{H}_R \right)^T}{\sum_{l=1}^{L} \left[\lambda_R^{C^{(l)}} \mathbf{1}_{F,T} \left(diag \left(Dx_R^{(l)} \right) \mathbf{V}_R^{(l)} \right)^T \right] + \alpha \mathbf{1}_{F,T} \left(diag \left(Dy_R \right) \mathbf{H}_R \right)^T + 2 \left(L + 1 \right) \mathbf{U}_R} \tag{9}$$

$$\mathbf{U}_W^{(l)} \leftarrow \mathbf{U}_W^{(l)} \odot \frac{\lambda_W \left(\mathbf{X}^{(l)} \oslash \hat{\mathbf{X}}^{(l)} \right) \left(diag \left(Dx_W^{(l)} \right) \mathbf{V}_W^{(l)} \right)^T}{\lambda_W \mathbf{1}_{F,T} \left(diag \left(Dx_W^{(l)} \right) \mathbf{V}_W^{(l)} \right)^T + 2\mathbf{U}_W^{(l)}} \tag{10}$$

$$\mathbf{V}_R^{(l)} \leftarrow \mathbf{V}_R^{(l)} \odot \frac{\left(\mathbf{U}_R diag \left(Dx_R^{(l)} \right) \right)^T \left(\mathbf{X}^{(l)} \oslash \hat{\mathbf{X}}^{(l)} \right)}{\left(\mathbf{U}_R diag \left(Dx_R^{(l)} \right) \right)^T \mathbf{1}_{F,T}} \tag{11}$$

$$\mathbf{V}_W^{(l)} \leftarrow \mathbf{V}_W^{(l)} \odot \frac{\left(\mathbf{U}_W^{(l)} diag \left(Dx_W^{(l)} \right) \right)^T \left(\mathbf{X}^{(l)} \oslash \hat{\mathbf{X}}^{(l)} \right)}{\left(\mathbf{U}_W^{(l)} diag \left(Dx_W^{(l)} \right) \right)^T \mathbf{1}_{F,T}} \tag{12}$$

$$\mathbf{H}_R \leftarrow \mathbf{H}_R \odot \frac{\left(\mathbf{U}_R diag \left(Dy_R \right) \right)^T \left(\mathbf{Y} \oslash \hat{\mathbf{Y}} \right)}{\left(\mathbf{U}_R diag \left(Dy_R \right) \right)^T \mathbf{1}_{F,T}} \tag{13}$$

The set of matrices $\mathbf{U}_R$, $\mathbf{U}_W^{(l)}$, $\mathbf{V}_R^{(l)}$, $\mathbf{V}_W^{(l)}$, $\mathbf{H}_R$ are obtained updating the rules (9)–(13) until the algorithm converges or reaches a maximum number of iterations M. At each iteration, the activation matrices $\mathbf{V}_R^{(l)}$, $\mathbf{V}_W^{(l)}$, $\mathbf{H}_R$ and the basis matrices $\mathbf{U}_R$, $\mathbf{U}_W^{(l)}$ must be normalized applying the L^2-norm (see Equation (14)). As a result, Dx_R, Dx_W, Dy_R must be updated multiplying by the L^2-norm obtained at each previous normalization (see Equation (15)). The normalization process ensures that both the sum of the square elements of each k-th column of the basis matrices $\mathbf{U}_R$, $\mathbf{U}_W^{(l)}$ and the sum of the square elements of each k-th row of the activation matrices $\mathbf{V}_R^{(l)}$, $\mathbf{V}_W^{(l)}$, $\mathbf{H}_R$ equals 1.0 [46].

$$\mathbf{G}(k) = \frac{\mathbf{G}(k)}{\sqrt{\sum \mathbf{G}^2(k)}} \tag{14}$$

$$D_J(k) = D_J(k) \sqrt{\sum \mathbf{G}^2(k)} \tag{15}$$

where $(\mathbf{G}, J, k) = \{(\mathbf{U}_R, R, k_R), (\mathbf{U}_W^{(l)}, W, k_W), (\mathbf{V}_R^{(l)}, R, k_R), (\mathbf{V}_W^{(l)}, W, k_W), (\mathbf{H}_R, R, k_R)\}$ respectively. If we consider the basis matrix $\mathbf{G} = (\mathbf{U}_R, \mathbf{U}_W^{(l)}) \rightarrow \sqrt{\sum \mathbf{G}^2(k)} = \sqrt{\sum_{f=1}^{F} \mathbf{G}^2(f,k)}$. If we consider the activation matrix $\mathbf{G} = (\mathbf{V}_R^{(l)}, \mathbf{V}_W^{(l)}, \mathbf{H}_R) \rightarrow \sqrt{\sum \mathbf{G}^2(k)} = \sqrt{\sum_{t=1}^{T} \mathbf{G}^2(k,t)}$.

After the updating process, the estimated spectrograms $\hat{\mathbf{X}}_R^{(l)}$ and $\hat{\mathbf{X}}_W^{(l)}$ for each segment can be reconstructed as follows:

$$\hat{\mathbf{X}}_R^{(l)} = \mathbf{U}_R diag(Dx_R^{(l)}) \mathbf{V}_R^{(l)} \tag{16}$$

$$\hat{\mathbf{X}}_W^{(l)} = \mathbf{U}_W^{(l)} diag(Dx_W^{(l)}) \mathbf{V}_W^{(l)} \tag{17}$$

Note that $\hat{\mathbf{X}}_R^{(l)}$ and $\hat{\mathbf{X}}_W^{(l)}$ must be denormalized by multiplying by the denominator of Equation (5). A Wiener filtering [32,55] has been applied in order to ensure a conservative signal reconstruction and to obtain the estimated complex wheezing and respiratory spectrogram of each segment. $\hat{\mathbf{X}}_R$ and $\hat{\mathbf{X}}_W$ are obtained by concatenating the estimated complex spectrograms of each segment, $\hat{\mathbf{X}}_R = \left[\hat{\mathbf{X}}_R^{(1)}, \hat{\mathbf{X}}_R^{(2)}, \ldots, \hat{\mathbf{X}}_R^{(L)}\right]$ and $\hat{\mathbf{X}}_W = \left[\hat{\mathbf{X}}_W^{(1)}, \hat{\mathbf{X}}_W^{(2)}, \ldots, \hat{\mathbf{X}}_W^{(L)}\right]$, respectively. Finally, the inverse overlap-add STFT is applied to synthesize the estimated RS signal $\hat{r}[n]$ and the estimated WS signal $\hat{w}[n]$ in time domain using the phase of the input mixture. The wheezing/normal respiratory sound separation procedure is summarized in Algorithm 1.

Algorithm 1 Wheezing sound separation using IIS-NMPCF.

Require: $x[n]$, $y[n]$, K_R, K_W, λ_R^0, λ_R^1, λ_W, α and M.

1) Compute the normalized magnitude spectrogram $\mathbf{X}$ of the mixture $x[n]$.
2) Compute the normalized magnitude spectrogram $\mathbf{Y}$ of the training $y[n]$.
3) Divide the spectrogram $\mathbf{X}$ into L-segments $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \cdots, \mathbf{X}^{(L)}$ using AMIE_SEG [53].
4) Classify the L-segments into wheezing ($C^{(l)} = 1$) and non-wheezing ($C^{(l)} = 0$) using a wheezing detection algorithm [54].
5) Initialize each activation and basis matrix $\mathbf{U}_R$, $\mathbf{U}_W^{(l)}$, $\mathbf{V}_R^{(l)}$, $\mathbf{V}_W^{(l)}$, $\mathbf{H}_R$ with random non-negative values.
6) Update each activation and basis matrix $\mathbf{U}_R$, $\mathbf{U}_W^{(l)}$, $\mathbf{V}_R^{(l)}$, $\mathbf{V}_W^{(l)}$, $\mathbf{H}_R$ using Equations (9)–(13) for the predefined number of iterations M. At each iteration, normalize each activation and basis matrix $\mathbf{U}_R$, $\mathbf{U}_W^{(l)}$, $\mathbf{V}_R^{(l)}$, $\mathbf{V}_W^{(l)}$, $\mathbf{H}_R$ and update the terms Dx_R, Dx_W and Dy_R using Equations (14) and (15).
7) Compute the estimated magnitude spectrograms $\hat{\mathbf{X}}_R^{(l)}$ and $\hat{\mathbf{X}}_W^{(l)}$.
8) Denormalize the estimated magnitude spectrograms $\hat{\mathbf{X}}_R^{(l)}$ and $\hat{\mathbf{X}}_W^{(l)}$.
9) Apply a Wiener filtering [32] on $\hat{\mathbf{X}}_R^{(l)}$ and $\hat{\mathbf{X}}_W^{(l)}$.
10) Concatenate all the estimated complex respiratory spectrograms: $\hat{\mathbf{X}}_R = \left[\hat{\mathbf{X}}_R^{(1)}, \hat{\mathbf{X}}_R^{(2)}, \ldots, \hat{\mathbf{X}}_R^{(L)}\right]$.
11) Concatenate all the estimated complex wheezing spectrograms: $\hat{\mathbf{X}}_W = \left[\hat{\mathbf{X}}_W^{(1)}, \hat{\mathbf{X}}_W^{(2)}, \ldots, \hat{\mathbf{X}}_W^{(L)}\right]$.
12) Synthesize $\hat{r}[n]$.
13) Synthesize $\hat{w}[n]$.

return $\hat{r}[n]$ and $\hat{w}[n]$

4. Experimental Results

4.1. Dataset and Metric

Because there is no public database where only wheeze sounds can be found to the best of our knowledge, two datasets P1 and T1 (T1H, T1M and T1L), detailed in Table 1, have been used in the evaluation of the proposed method with a total of 64 recordings considering the two databases. Specifically, the database P1 consists of 48 recordings (that is, 3/4 of the total recordings used in the

experiments) and the database T1 consists of 16 recordings (that is, 1/4 of the total recordings used in the experiments). The dataset P1 has been used in the hyperparametric optimization process (see Section 4.4) while the dataset T1 has been used in the separation testing (see Section 4.5). The databases P1 and T1 have been created by collecting a set of recordings from different subjects of the most widely used Internet pulmonary repositories [56–68]. These recordings, captured from the trachea, anterior, and posterior chest using either a stethoscope or microphone, were collected from subjects with different pathologies, including asthma, bronchitis or COPD. The databases P1 and T1 have been created by randomly selecting recordings from the above-mentioned repositories. It must be highlighted that P1 is not a part of T1 in order to validate the results. Therefore, the recordings selected for the database P1 are not the same as the recordings selected for the database T1. In total, these databases provide 1474 s of recording, 96 unhealthy subjects, 874 respiratory events (a respiratory event is defined as inspiration or expiration) and 133 wheezes. Note that each recording has been created using single-channel configuration, a sampling rate equals 2048 Hz and a bit resolution of 16 bits.

Specifically, the datasets P1 and T1 (T1H, T1M and T1L) have been created mixing only WS recordings manually separated $w[n]$, in which respiratory sounds are inactive, and only RS recordings $r[n]$, in which wheezing sounds are inactive, obtained from the above-mentioned repositories. Highlight that wheezing sounds cannot be recorded isolated since WS are always overlapped with RS, that is, both sounds are produced by the same bronchial tree in the lungs. To do this, a MATLAB tool, designed by the authors, has been used to visually modify the spectrogram values. Specifically, this tool behaves as an eraser that allows us, by means of the mouse, to set to zero those bins of the spectrogram that we observe that do not belong to a wheeze sound, a fact that is also verified by a listening inspection of the resulting signal. Therefore, only the bins corresponding to WS have been kept active for each signal $w[n]$. Both the fundamental component of WS and its corresponding harmonics have been considered. Note that the recordings used to create the database P1 are different from those used to create the database T1.

The datasets T1H (SNR = 5 dB), T1M (SNR = 0 dB) and T1L (SNR = -5 dB) are composed of the same set of signals $w[n]$ and $r[n]$ but they have been mixed using a different Signal-to-Noise Ratio (SNR). Specifically, T1H is composed of mixtures in which the power of $w[n]$ is 5 dB greater compared to $r[n]$ so, WS are louder than RS. The dataset T1M is composed of mixtures in which the power of both $w[n]$ and $r[n]$ is the same so, both type of sounds is similarly audible. Finally, the dataset T1L is composed of mixtures in which the power of $w[n]$ is 5 dB lower compared to $r[n]$ so, RS are louder than WS. Note that in each mixture process, the power related to $w[n]$ and $r[n]$ are calculated and the signal with the highest power is left fixed while the signal with the lowest power is scaled to obtain the desired SNR in order to avoid audio saturation or distortion in the signal scaling process.

Table 1. Characteristics of each database.

ID1	ID2	ID3	ID4	ID5	ID6	ID7	ID8	ID9
P1	48	5–24	721	[0–9]	[4–16]	496	[1–8]	92
T1H	16	7–22	251	5	[6–14]	126	[1–5]	41
T1M	16	7–22	251	0	[6–14]	126	[1–5]	41
T1L	16	7–22	251	-5	[6–14]	126	[1–5]	41

ID1: identifier; ID2: number of recordings captured from unhealthy subjects; ID3: the shortest and longest duration, in seconds, captured from recordings; ID4: total duration in seconds; ID5: the lowest and highest SNR, in dB, between WS and RS; ID6: the minimum and maximum number of respiratory events found in the recordings; ID7: the total number of respiratory events; ID8: the minimum and maximum number of wheezes found in the recordings; ID9: the total number of wheezes.

To assess the sound separation performance of the proposed method, the BSS EVAL toolbox [69,70] has been applied because it is widely used in the field of sound source separation. The metrics used are the following: (1) Source-to-distortion ratio (SDR), which provides information on the overall quality of the separation process; (2) Source-to-interferences ratio (SIR), which reports the presence of WS

contained in RS and vice versa; and (3) Source-to-artifacts ratio (SAR), which provides information on the artifacts in the separated signal from separation and/or resynthesis. The principle to obtain the value of these metrics is to decompose the total error, between the estimated target signal $\hat{s}[n]$ and the original target signal $s[n]$, in three terms related to three types of error, as follows [70]:

$$\hat{s}[n] - s[n] = {e_s}^{interf}[n] + {e_s}^{artifacts}[n] + {e_s}^{spatial}[n] \tag{18}$$

where ${e_s}^{interf}[n]$ is the error term related to the interference produced by the unwanted sources; ${e_s}^{artifacts}[n]$ is the error term attributed to the artifacts generated by the separation algorithm; and ${e_s}^{spatial}[n]$ is the error term attributed to spatial distortion. We can now define the SDR, SIR and SAR values, expressed in dB, as follows:

$$SDR = 10\log_{10}\frac{\|s[n]\|^2}{\left\|{e_s}^{interf}[n] + {e_s}^{artifacts}[n] + {e_s}^{spatial}[n]\right\|^2} \tag{19}$$

$$SIR = 10\log_{10}\frac{\|s[n]\|^2}{\left\|{e_s}^{interf}[n]\right\|^2} \tag{20}$$

$$SAR = 10\log_{10}\frac{\left\|s[n] + {e_s}^{interf}[n] + {e_s}^{spatial}[n]\right\|^2}{\left\|{e_s}^{artifacts}[n]\right\|^2} \tag{21}$$

Note that the term s indicates the target signal to be analyzed. In this article s could be the wheezing signals ($s = w$) and the respiratory signals ($s = r$). Therefore, in the case of the wheezing signals ($\hat{s}[n]$, $s[n]$, ${e_s}^{interf}[n]$, ${e_s}^{artifacts}[n]$, ${e_s}^{spatial}[n]$) = ($\hat{w}[n]$, $w[n]$, ${e_w}^{interf}[n]$, ${e_w}^{artifacts}[n]$, ${e_w}^{spatial}[n]$) and in the case of the respiratory signals ($\hat{s}[n]$, $s[n]$, ${e_s}^{interf}[n]$, ${e_s}^{artifacts}[n]$, ${e_s}^{spatial}[n]$) = ($\hat{r}[n]$, $r[n]$, ${e_r}^{interf}[n]$, ${e_r}^{artifacts}[n]$, ${e_r}^{spatial}[n]$). The estimated signals $\hat{w}[n]$, $\hat{r}[n]$ are obtained by the separation algorithm, the original signals $w[n]$, $r[n]$ are obtained from the original separated signals used in the creation of the mixtures of the databases and the error terms are obtained using the BSS EVAL toolbox. We refer the reader to [70] for more details.

In this article, three different sets of SDR, SIR and SAR metrics will be analyzed as follows: (i) SDR_w, SIR_w and SAR_w are referred to WS, (ii) SDR_r, SIR_r and SAR_r are referred to RS; and (iii) SDR_m is associated to the average considering SDR_w and SDR_r, SIR_m is associated to the average considering SIR_w and SIR_r, and SAR_m is associated to the average considering SAR_w and SAR_r.

4.2. Experiments Setup

According to the results obtained in similar works [32,54] related to wheezing sound analysis, the following parameters provided the best trade-off between the separation performance and the computational cost: sampling rate $f_s = 2048$ Hz, Hamming window with $N = 256$ samples length and 25% overlap (temporal resolution of 31.3 ms), and a discrete Fourier transform using 2N points.

The performance of the proposed method depends on the initial values with which each activation and basis matrix is initialized. For this reason, we have evaluated four times each input mixture with the proposed method and therefore, the results are averaged values. Furthermore, the convergence of the proposed method was empirically achieved after 50 iterations for all mixtures, so M = 50 iterations.

4.3. Comparison Methods

A set of reference baseline sound source separation methods have been compared to assess the sound separation performance achieved by the proposed method (IIS-NMPCF). As mentioned in Section 2, these methods can be divided into two groups: (i) NMF-based methods (NMF, SNMF, SSNMF and CNMF); and (ii) NMPCF-based methods (1S-NMPCF, 2S-NMPCF, T-NMPCF and ST-NMPCF).

Highlight that the main parameters of the previous baseline methods have been optimized using the database P1. However, the following considerations must be taken into account to a fair comparison:

- A training signal $y[n]$, created to simulate the behavior of RS, is used in the baseline methods SNMF, SSNMF, 1S-NMPCF, 2S-NMPCF, ST-NMPCF and the proposed method IIS-NMPCF. The training signal $y[n]$ has been created by concatenating randomly a set of normal respiratory stages only composed of RS obtained from the previously mentioned Internet pulmonary repositories [56–68]. Specifically, the signal $y[n]$ has a temporal duration of 128 s and 54 respiratory stages (inspiration or expiration). Note that the normal respiratory stages used to construct y[n] do not correspond to any of the respiratory stages used in the databases P1 or T1.
- SNMF and 2S-NMPCF must use a training signal to simulate the behaviour of wheezing sounds. Taking into account that WS can be defined as continuous adventitious sounds that show a pitched sound (see Section 1), a signal $z[n]$ has been created by concatenating a set of single pitches located along the frequency band 100 Hz–1000 Hz in which WS are typically present. Each pitch is represented by a sinusoidal signal multiplied by a Hamming window of N samples. The distance between the frequencies of each pitch is equal to the value provided by the spectral spacing of the model. Considering that all evaluated methods have used the same parameters previously mentioned in Section 4.2, the spectral spacing equals to 4 Hz.
- T-NMPCF and ST-NMPCF as well as IIS-NMPCF has been implemented using AMIE_SEG [53] to divide the input spectrogram $\mathbf{X}$ into the L-segments $\mathbf{X}^{(1)}$,$\mathbf{X}^{(2)}$, $\cdots$,$\mathbf{X}^{(L)}$.
- CNMF has been evaluated using its optimal parameters found in [32].

4.4. Optimization

The proposed method employs a wide range of parameters K_R, K_W, α, λ_W, λ_R^0 and λ_R^1 that can affect significantly the separation performance and the reconstructed sound quality. A hyperparametric optimization procedure has been applied to the main parameters of the proposed method IIS-NMPCF to obtain the optimal parameters that maximize the audio quality of the estimated wheezing signal $\hat{w}[n]$. In this work, a preliminary evaluation using visual inspection reduced the parameter space as follows: K_R = (8, 16, 32, 64, 128, 256, 512), K_W = (8, 16, 32, 64, 128, 256, 512), α = (0, 0.01, 0.1, 1, 10, 100), λ_W = (0.001, 0.01, 0.1, 1, 10), λ_R^0 = (0.001, 0.01, 0.1, 1, 10, 100) and λ_R^1 = (0.001, 0.01, 0.1, 1, 10, 100).

The hyperparametric procedure is performed for each mixture of the dataset P1 in order to obtain the audio quality of the estimated wheeze signal $\hat{w}[n]$ in terms of SDR_w, SIR_w and SAR_w. This procedure has been computed by evaluating all the possible combinations of the parameters K_R, K_W, α, λ_W, λ_R^0 and λ_R^1 that can be found within the parameter space defined above, providing the SDR_w, SIR_w and SAR_w average values for each combination of parameters. Table 2 shows the optimal combination of the previous parameters that provides the best separation performance in terms of SDR_w. Specifically, the optimal parameters corroborate our previous assumptions described in Section 3.2: (i) the highest weighting factor $\lambda_R^0 = 10$ is due to the high importance of the non-wheezing segments in the factorization of the respiratory bases since RS can be modeled by sharing spectral patterns that can be found in all non-wheezing segments during the breathing process; (ii) the second highest weighting factor $\alpha = 1$ is associated with the training signal since RS typically show common spectral behavior; (iii) the low weighting factor $\lambda_R^1 = 0.1$ is associated with the wheezing segments in the factorization of the respiratory bases since WS can interfere in the RS reconstruction; and (iv) the lowest weighting factor $\lambda_W = 0.01$ is due to none of the L-segments is only composed by isolated WS.

Table 2. The optimal parameters of the proposed method that obtain the best wheezing audio quality evaluating the dataset P1.

IIS-NMPCF approach parameters	K_W	K_R	λ_R^0	α	λ_R^1	λ_W
Optimal values	64	32	10	1	0.1	0.01

Focusing on the parameter space defined above and keeping the optimal parameters shown in Table 2, the aim of the rest of the section is to analyze the stability and efficiency of the proposed method when its main parameters K_W, K_R, α, λ_W, λ_R^0 and λ_R^1 are distanced from the optimal values.

Figure 4 shows the SDR_w results varying the number of respiratory K_R and wheezing K_W components. Figure 4 shows that the difference, in terms of SDR_w, between the configuration of the parameters K_R and K_W that provides the best performance (SDR_w = 16.99 dB) and the worst performance (SDR_w = 14.01 dB) is approximately 3 dB. Therefore, the proposed method is stable within the defined parameter space K_W and K_R since the maximum loss that the algorithm can suffer is less than 3 dB regardless of the number of wheezing K_W and respiratory K_R components evaluated. Besides, the difference in SDR_w results is marginal (less than 0.2 dB) either using $K_W \geq 256$ and $K_R \geq 256$ or (less than 0.3 dB) using $K_W \leq 16$ and $K_R \leq 16$. Highlight that the proposed factorization model needs a minimum of respiratory and wheezing components so that WS and RS can be modelled correctly. An empirical analysis showed that the SDR_w results start to drop significantly when $K_W < 16$ and $K_R < 16$. Figure 4 shows that SDR_w results increase when the number of wheezing components is greater than the number of respiratory components ($K_W > K_R$). Specifically, comparing the parameter space $K_W \in [32 - 512]$ and $K_R \in [8 - 16]$ with $K_W \in [8 - 16]$ and $K_R \in [32 - 512]$, the performance of the method, in terms of SDR_w, improves by about 1.7 dB. As a result, RS seem to be modelled with a lower number of bases than WS. Finally, the best performance of the proposed method IIS-NMPCF can be found in the parameter space comprised by $K_W \in [32 - 128]$ and $K_R \in [32 - 128]$ with SDR_w results above 16.5 dB. As previously indicated in Table 2, the proposed method provides its highest wheezing separation performance, SDR_w = 16.99 dB, using $K_W = 64$ and $K_R = 32$.

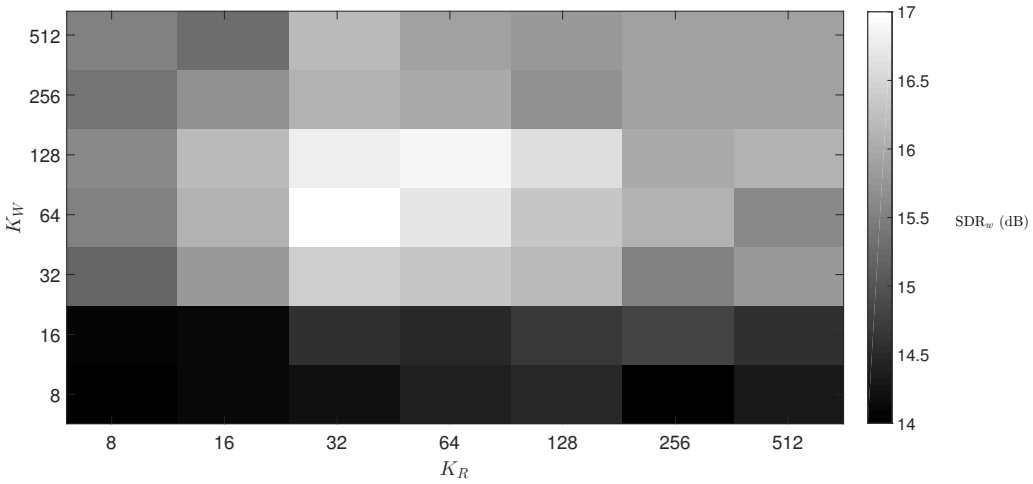

Figure 4. SDR_w average results from the hyperparametric optimization of the proposed method varying the parameters K_W and K_R. The rest of parameters are the following: $\lambda_R^0 = 10$, $\alpha = 1$, $\lambda_R^1 = 0.1$ and $\lambda_W = 0.01$.

Figure 5 shows the optimization of the parameters λ_W, λ_R^0 and λ_R^1 of the proposed method in terms of SDR_w results, of the proposed method. Figure 5E shows a poor wheezing separation when the proposed method uses a $\lambda_W = 10$ since the performance of the proposed method decreases exponentially (below 2 dB) in this scenario. The reason seems to indicate that WS are always overlapped with RS since both are produced by the same airflow through the bronchial tree of the lungs. Therefore, the proposed method wrongly models the wheeze bases when $\lambda_W \geq 10$ since it assumes that the L-segments of the input mixture are composed mostly of prominent WS. Figure 5A shows that SDR_w results decrease significantly when $\lambda_W = 0.001$. In this case, the use of an excessively low weighting factor makes WS less important in the factorization process, causing that the separation process is not performed correctly since the estimated respiratory signal $\hat{r}[n]$ contains both WS and RS. Figure 5B,D show the lower and upper limit of the weighting factor λ_W so that the performance of

the method is not drastically affected. Figure 5 shows an improvement of the wheeze separation performance of the proposed method when $\lambda_R^0 > \lambda_R^1$. Results suggest that, unlike the wheezing segments, the non-wheezing segments improve the modeling of the RS bases since these segment do not contain wheeze content so, they are not interfered by WS. As a result, λ_R^0 must be greater than λ_R^1 to increase the quality of the reconstructed respiratory signal $\hat{r}[n]$. In the parameter space comprised by $\lambda_R^0 \in [0.001 - 100]$ and $\lambda_R^1 \in [10 - 100]$, the SDR_w results are reduced significantly as can be seen in Figure 5. Therefore, a remarkable increase of λ_R^1 causes that the factorization model inserts a large proportion of wheezing interferences into the reconstructed respiratory signal. This fact produces more of the WS to be present in the reconstructed respiratory signal $\hat{r}[n]$ rather than in the reconstructed wheezing signal $\hat{w}[n]$. It can be observed that the maximum SDR_w value, approximately equal to 17 dB in Figure 5B, is provided by the proposed method for the set of parameters $\lambda_W = 0.01$, $\lambda_R^1 = 0.1$ and $\lambda_R^0 = 10$. This optimization process confirms the assumptions introduced in Section 3.2. Firstly, the proposed method provides the greatest importance, with a weighting factor of $\lambda_R^0 = 10$, to the non-wheezing segments for the factorization of the basis matrix related to RS. Secondly, the proposed method provides less importance, with a weighing factor of $\lambda_R^1 = 0.1$, to the wheezing segments for the factorization of the basis matrix of the RS. Finally, the proposed method provides the least importance, with a weighting factor of $\lambda_W = 0.01$, to the L-segments that composes the input mixture signal for the factorization of the basis matrix of WS, as in none of these segments are WS isolated.

Note that when $\lambda_W = \lambda_R^0 = \lambda_R^1$, the proposed method works similarly to the conventional NMPCF approach, that is, ST-NMPCF. In particular, Figure 5B shows that ST-NMPCF obtains a SDR_w result equal to 13 dB (4 dB less than the optimal value obtained with the proposed method) using $\lambda_W = \lambda_R^0 = \lambda_R^1 = 0.01$. This improvement provided by the proposed method confirms that adding different weighting factors to different segments of the input mixture into the NMPCF factorization enhances the acoustic fidelity of the spectral content of both RS and WS in the sound separation.

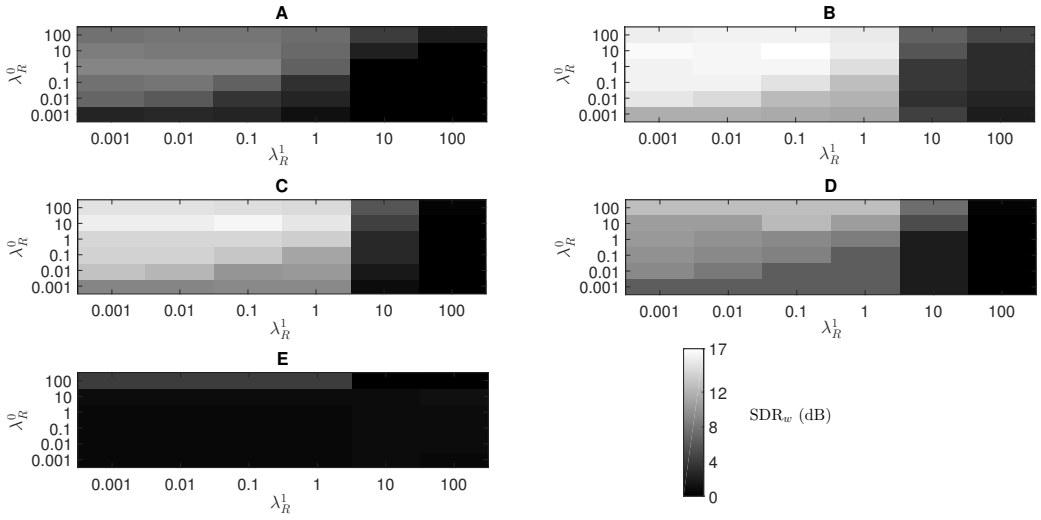

Figure 5. SDR_w average results from the hyperparametric optimization of the proposed method varying the parameters λ_W, λ_R^0 and λ_R^1. The rest of parameters are the following: $K_W = 64$, $K_R = 32$ and $\alpha = 1$. (**A**) $\lambda_W = 0.001$, (**B**) $\lambda_W = 0.01$, (**C**) $\lambda_W = 0.1$, (**D**) $\lambda_W = 1$ and (**E**) $\lambda_W = 10$.

Focusing on the importance of the respiratory training signal $y[n]$ in the proposed IIS-NMPCF approach, Figure 6 shows SDR_w, SIR_w and SAR_w results of the estimated wheezing signal evaluating the parameter space of the weighting factor α. Each box represents 48 data points, one for each mixture of the optimization dataset P1: each blue box represents the analysis for SDR_w values; each red box represents the analysis for SIR_w values; and each black box represents the analysis for SAR_w values. The lower and upper lines of each box show the 25th and 75th percentiles. The line in the middle of each box represents the median value. The diamond in the center of each box represents the

average value. The lines extending above and below each box show the extent of the rest of the samples, excluding outliers. Outliers are defined as points that are over 1.5 times the interquartile range from the sample median, which are shown as crosses. The proposed method using $\alpha = 0$, henceforth called IIS_0-NMPCF, does not use any training to model the respiratory bases. IIS_0-NMPCF shows an efficient performance with an average separation results of SDR_w = 14 dB, SIR_w = 18 dB and SAR_w = 15 dB. Based on these results, it can be confirmed that IIS_0-NMPCF maintains a remarkable performance in the quality of the estimated wheezing signal $\hat{w}[n]$. However, the best average separation results, SDR_w = 17 dB, SIR_w = 22 dB and SAR_w = 20 dB, are obtained using $\alpha = 1$. The optimal configuration of the proposed method IIS-NMPCF ($\alpha = 1$) produces a significant improvement of 3 dB in SDR_w, 4 dB in SIR_w and 5 dB in SAR_w compared to IIS_0-NMPCF. As a result, two conclusions are stated: (i) the performance of IIS-NMPCF is mainly due to the importance of the different segments depending on the presence or absence of WS so, not using any respiratory training signal the method maintains good separation results; and (ii) the use of a respiratory training signal significantly improves the performance of the proposed method IIS-NMPCF since it is combined both the information provided by the spectral patterns found at inter-segments with the information provided by the spectral patterns found in the respiratory training signal. This fact implies that the probability of finding wheezing interferences in the factorized respiratory bases decreases considerably.

Moreover, SDR_w, SIR_w and SAR_w results, obtained using $\alpha > 10$, suffer a significant decrease compared to the best performance provided by the proposed method ($\alpha = 1$) as shown in Figure 6. In this case ($\alpha > \lambda_R^0$), the factorization gives more importance to the spectral patterns obtained from the respiratory training signal instead of the spectral patterns shared between the different segments, that is, the proposed method IIS-NMPCF performs similarly to 1S-NMPCF.

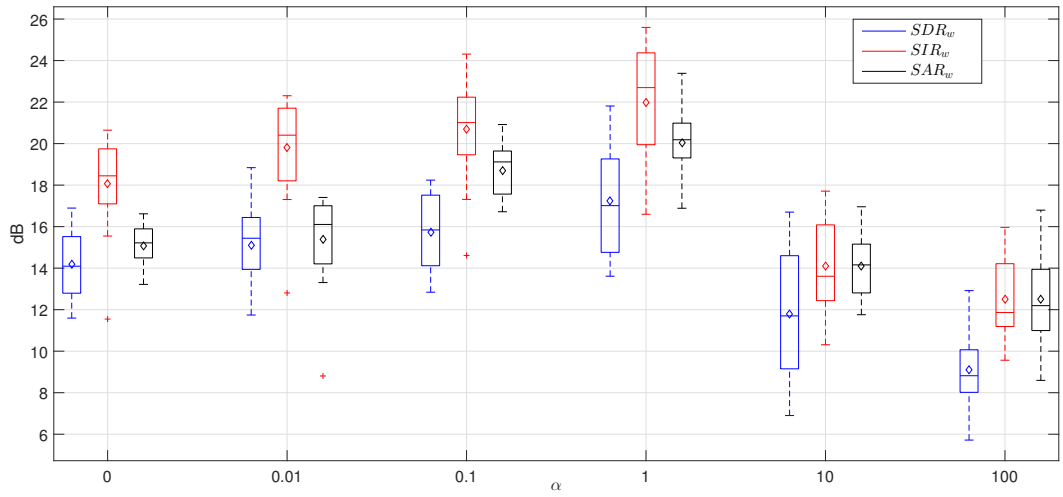

Figure 6. SDR_w, SIR_w and SAR_w average results from the hyperparametric optimization of the proposed method varying the parameter α. The rest of parameters are the following: $K_W = 64$, $K_R = 32$, $\lambda_R^0 = 10$, $\lambda_R^1 = 0.1$ and $\lambda_W = 0.01$.

4.5. Results and Discussion

This section assesses the sound quality of the estimated or reconstructed WS and RS obtained by the proposed method (IIS_0-NMPCF and IIS-NMPCF) and the baseline separation NMF-based and NMPCF-based methods described in Section 2. Table 3 describes the methods evaluated, indicating the approach on which they are based and the spectro-temporal information used in the modelling of WS and RS.

Table 3. Characteristics of the methods evaluated.

Method	Approach	Modelling Associated to WS and RS
NMF	NMF	
SSNMF	NMF	$y[n]$
SNMF	NMF	$y[n]$ and $z[n]$
CNMF	NMF	Sparseness and Smoothness constraints
1S-NMPCF	NMPCF	$y[n]$
2S-NMPCF	NMPCF	$y[n]$ and $z[n]$
T-NMPCF	NMPCF	L-segments
ST-NMPCF	NMPCF	L-segments and $y[n]$
IIS_0-NMPCF	NMPCF	L-segments and $C^{(l)}$
IIS-NMPCF	NMPCF	L-segments, $C^{(l)}$ and $y[n]$

Next, SDR, SIR and SAR results of the estimated wheezing signal $\hat{w}[n]$ and the estimated respiratory signal $\hat{r}[n]$ obtained by the proposed method and the aforementioned baseline methods evaluating the testing datasets T1H (see Figure 7), T1M (see Figure 8) and T1L (see Figure 9) are analyzed to extract interesting information about the sound separation performance of the methods evaluated. Each blue box corresponds to the SDR_w, SIR_w and SAR_w results of the estimated wheezing signal while each red box corresponds to the SDR_r, SIR_r and SAR_r results of the estimated respiratory signal. Note that the methods have been shown sorted from lowest to highest separation performance to represent results as a ranking. The following information can be derived from the analysis of results from Figures 7–9:

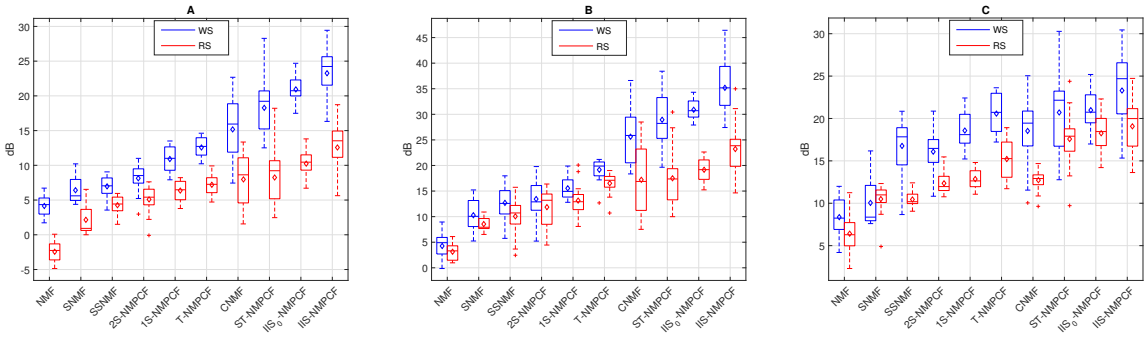

Figure 7. SDR_w and SDR_r results (**A**), SIR_w and SIR_r results (**B**) and SAR_w and SAR_r results (**C**) evaluating the dataset T1H (SNR = 5 dB). Note that SDR_w, SIR_w and SAR_w are represented by blue boxes while SDR_r, SIR_r and SAR_r are represented by red boxes.

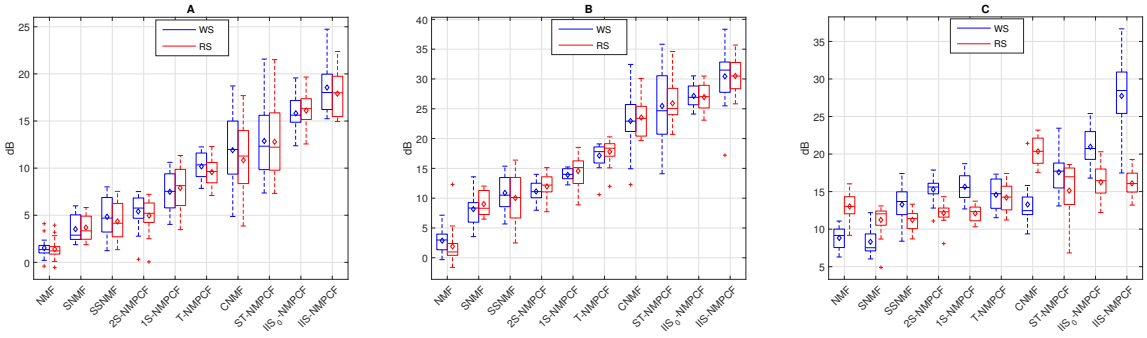

Figure 8. SDR_w and SDR_r results (**A**), SIR_w and SIR_r results (**B**) and SAR_w and SAR_r results (**C**) evaluating the dataset T1M (SNR = 0 dB). Note that SDR_w, SIR_w and SAR_w are represented by blue boxes while SDR_r, SIR_r and SAR_r are represented by red boxes.

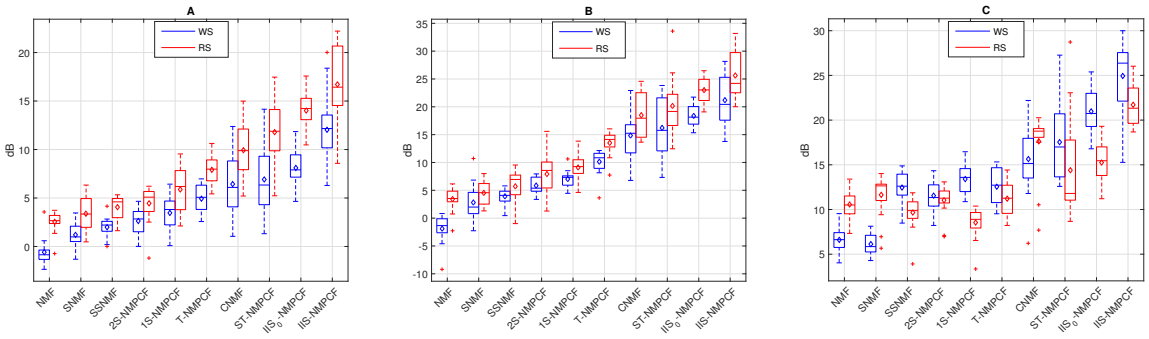

Figure 9. SDR_w and SDR_r results (**A**), SIR_w and SIR_r results (**B**) and SAR_w and SAR_r results (**C**) evaluating the dataset T1L (SNR = −5 dB). Note that SDR_w, SIR_w and SAR_w are represented by blue boxes while SDR_r, SIR_r and SAR_r are represented by red boxes.

- The decrease in SNR affects significantly the SDR and SIR results for both WS and RS. Focusing on Figure 7 in which SNR = 5 dB, results tend to be higher for reconstructed WS compared to the reconstructed RS because WS are louder than RS, so the sound separation benefits the audio quality of the reconstructed WS. Focusing on Figure 8 in which SNR = 0 dB, results for both WS and RS tend to remain stable because both WS and RS are similarly audible, so the performance of the sound separation seems to work equally between WS and RS. However, in Figure 9 in which SNR = −5 dB, results tend to be better for reconstructed RS since RS are louder than WS. This decrease in SNR implies that SDR_m and SIR_m results are worse in T1L compared to T1H. The reason is because RS are louder than WS when SNR < 0 dB (T1L) and as a consequence, WS be inaudible in this acoustic scenario so, the reduction of the SNR implies a greater time-frequency overlapping from RS to WS than the opposite.
- The standard NMF is ranked at the bottom, obtaining the worst sound separation performance since it achieves the signal reconstruction but not a factorization composed of audio events with physical meaning. The standard NMF cannot group the factorized bases to the sound source that generated them unlike the other methods because the standard NMF does not incorporate any type of information into the factorization process to model the spectro-temporal characteristics shown by WS and RS.
- Semi-supervised approaches (SSNMF and 1S-NMPCF) obtain better performance compared to supervised approaches (SNMF and 2S-NMPCF). Regardless of the approach, NMF or NMPCF, the use of the RS training signal is more effective that the use of both RS and WS training signals. It indicates that both training signals provide over-information that causes spectro-temporal ambiguity in the factorization of both WS and RS dictionaries.
- NMPCF-based methods (1S-NMPCF) obtain better separation performance than NMF-based methods (SSNMF). This fact seems to be because SSNMF uses a fixed dictionary composed of respiratory bases previously trained. However, 1S-NMPCF does not need a previous training stage, since it applies a joint matrix factorization using the input mixture and the respiratory training to obtain a dynamic dictionary of respiratory bases shared between both signals, obtaining a different dictionary of bases for each input mixture.
- Comparing NMPCF-based methods, T-NMPCF improves the separation performance compared to 1S-NMPCF. Results suggest that the dictionary of respiratory bases is more efficient when the input mixture is divided into segments in order to find repetitive patterns of RS.
- ST-NMPCF, the combination of the approaches 1S-NMPCF and T-NMPCF, obtains a significant improvement of the wheezing separation performance. Specifically, SDR_w = 5.96 dB and SIR_w = 9.73 dB evaluating T1H (Figure 7). It indicates that a more reliable modelling of RS can be achieved using jointly the shared respiratory spectral patterns along the segments and a prior knowledge of the respiratory spectral content by means of the respiratory training signal.

- CNMF [32] obtains competitive SDR SIR and SAR results compared to the methods above, ranking fourth. In some cases, WS and RS are modelled efficiently by applying its proposed constraints, but in other cases in which WS and RS are uncommon, CNMF does not model properly the spectro-temporal behavior of the target sounds.

Focusing on the main contribution proposed in this work, the incorporation of higher importance to those segments classified as non-wheezing in the co-factorization process, Figures 7–9 reveal the following information:

- A significant separation performance improvement over the conventional T-NMPCF and ST-NMPCF is achieved adding greater importance to the non-wheezing segments in the co-factorization process. The SDR_w improvement of IIS_0-NMPCF over T-NMPCF is about 8.31 dB (T1H), 5.18 dB (T1M) and 4.85 dB (T1L). The SIR_w improvement of IIS_0-NMPCF over T-NMPCF is about 11.09 dB (T1H), 10.18 dB (T1M) and 8.33 dB (T1L). The SDR_w improvement of IIS_0-NMPCF over ST-NMPCF is about 2.67 dB (T1H), 3.03 dB (T1M) and 1.69 dB (T1L). The SIR_w improvement of IIS_0-NMPCF over ST-NMPCF is about 1.98 dB (T1H), 2.25 dB (T1M) and 1.87 dB (T1L). Results suggest that the inclusion of inter-segment information into the co-factorization process for modeling repetitive RS improves significantly the separation performance because it avoids that the respiratory spectral patterns obtained from the factorization remaining uncontaminated in wheezing segments.
- Adding prior knowledge of RS to IIS_0-NMPCF improves significantly the sound separation performance. The SDR_w improvement of IIS-NMPCF over IIS_0-NMPCF is about 3.07 dB (T1H), 2.89 dB (T1M) and 4.12 dB (T1L). The SIR_w improvement of IIS-NMPCF over IIS_0-NMPCF is about 4.96 dB (T1H), 3.23 dB (T1M) and 3.02 dB (T1L). However, the dispersion between SDR and SIR results increases when the respiratory training signal is incorporated into the co-factorization process.

Focusing on the SAR results observed in Figure 7C, Figure 8C and Figure 9C: (i) NMPCF-based methods produce fewer artifacts than NMF-based methods; (ii) the spectro-temporal information used in the modelling of WS and RS allows to reduce the ambiguity that NMPCF-based methods are affected by decreasing the amount of artifacts. For this reason, the proposed method IIS-NMPCF, which uses more spectro-temporal information to model RS compared to the other NMPCF-based methods, obtains the best separation performance in terms of SAR.

In order to guarantee the relevance of the respiratory and wheezing SDR, SIR and SAR results shown in Figures 7–9, an analysis of the statistical significance, using an one-side paired *t*-test, has been performed comparing the proposed method (IIS-NMPCF) with the rest of the evaluated methods as shown in Tables 4–6. It can be observed that results confirm the significant improvement obtained by IIS-NMPCF compared to the other evaluated methods.

Table 4. Analysis of the statistical significance of the respiratory/wheezing SDR, SIR and SAR results comparing the proposed method (IIS-NMPCF) with the other evaluated methods using an one-sided paired *t*-test in the databases T1H (see Figure 7).

Method	SDR_r	SIR_r	SAR_r	SDR_w	SIR_w	SAR_w
NMF	6.1×10^{-10}	4.1×10^{-10}	5.4×10^{-3}	1.9×10^{-11}	1.8×10^{-11}	4.8×10^{-7}
SSNMF	1.4×10^{-7}	5.5×10^{-8}	4.8×10^{-3}	3.2×10^{-10}	4.4×10^{-12}	8.9×10^{-8}
SNMF	1×10^{-7}	3.1×10^{-7}	4.5×10^{-8}	7.9×10^{-12}	1.4×10^{-10}	1.2×10^{-2}
2S-NMPCF	3.3×10^{-7}	5.5×10^{-6}	4.9×10^{-7}	2.6×10^{-11}	1.7×10^{-10}	1.8×10^{-3}
1S-NMPCF	6×10^{-6}	4×10^{-7}	2.7×10^{-6}	5.2×10^{-10}	9.2×10^{-11}	4.9×10^{-3}
T-NMPCF	3.8×10^{-5}	5.7×10^{-5}	3.3×10^{-4}	4.4×10^{-9}	8.9×10^{-9}	2.9×10^{-2}
CNMF	1.6×10^{-4}	1.7×10^{-3}	1.8×10^{-7}	2.6×10^{-6}	1.3×10^{-6}	7.2×10^{-4}
ST-NMPCF	1.5×10^{-4}	9.5×10^{-6}	5.2×10^{-2}	3.9×10^{-5}	5.3×10^{-7}	2.2×10^{-2}
IIS_0-NMPCF	4×10^{-2}	2.2×10^{-3}	1×10^{-1}	4.2×10^{-2}	8.2×10^{-3}	1.1×10^{-1}

Each cell shows the parameter ρ that represents the probability of setting a statistically significant result. Considering a confidence interval of 95%, small values of $\rho < 0.05$ indicate that there exists statistical significance of the results evaluated.

Table 5. Analysis of the statistical significance of the respiratory/wheezing SDR, SIR and SAR results comparing the proposed method (IIS-NMPCF) with the other evaluated methods using an one-sided paired *t*-test in the databases T1M (see Figure 8).

Method	$\mathbf{SDR}_r$	$\mathbf{SIR}_r$	$\mathbf{SAR}_r$	$\mathbf{SDR}_w$	$\mathbf{SIR}_w$	$\mathbf{SAR}_w$
NMF	4×10^{-13}	4.1×10^{-14}	9.4×10^{-2}	6.2×10^{-13}	2×10^{-12}	2.7×10^{-9}
SNMF	4.3×10^{-13}	7.3×10^{-13}	4.3×10^{-2}	2.8×10^{-13}	1×10^{-10}	1.3×10^{-10}
SSNMF	9.7×10^{-11}	2.3×10^{-10}	2.5×10^{-6}	5.6×10^{-11}	4.3×10^{-10}	2.9×10^{-7}
2S-NMPCF	6.9×10^{-11}	1.1×10^{-11}	4.9×10^{-6}	5.1×10^{-11}	5.8×10^{-11}	3.5×10^{-8}
1S-NMPCF	8.7×10^{-8}	4.9×10^{-10}	1.3×10^{-6}	1.7×10^{-8}	1.4×10^{-9}	3.3×10^{-7}
T-NMPCF	9.7×10^{-9}	3.7×10^{-9}	1.1×10^{-7}	6.9×10^{-9}	1.6×10^{-7}	1.2×10^{-9}
CNMF	9.6×10^{-7}	1.1×10^{-5}	5.7×10^{-5}	9.4×10^{-6}	7.4×10^{-5}	5.2×10^{-7}
ST-NMPCF	1.9×10^{-4}	1.6×10^{-4}	4.4×10^{-2}	8.4×10^{-5}	4.3×10^{-4}	1.3×10^{-9}
IIS_0-NMPCF	4.1×10^{-2}	6.3×10^{-4}	4×10^{-1}	2×10^{-2}	3.1×10^{-2}	3.3×10^{-4}

Each cell shows the parameter ρ that represents the probability of setting a statistically significant result. Considering a confidence interval of 95%, small values of $\rho < 0.05$ indicate that there exists statistical significance of the results evaluated.

Table 6. Analysis of the statistical significance of the respiratory/wheezing SDR, SIR and SAR results comparing the proposed method (IIS-NMPCF) with the other evaluated methods using an one-sided paired *t*-test in the databases T1L (see Figure 9).

Method	$\mathbf{SDR}_r$	$\mathbf{SIR}_r$	$\mathbf{SAR}_r$	$\mathbf{SDR}_w$	$\mathbf{SIR}_w$	$\mathbf{SAR}_w$
NMF	2.1×10^{-9}	6.8×10^{-12}	3.9×10^{-8}	2.2×10^{-9}	1.5×10^{-12}	8.6×10^{-10}
SNMF	5.5×10^{-10}	3×10^{-11}	1.9×10^{-5}	1.7×10^{-9}	2.7×10^{-12}	8×10^{-12}
SSNMF	5.3×10^{-10}	1.1×10^{-13}	6.5×10^{-10}	1.2×10^{-9}	6.2×10^{-11}	3.6×10^{-10}
2S-NMPCF	2.8×10^{-10}	3.1×10^{-9}	4.7×10^{-10}	1.1×10^{-8}	1.2×10^{-10}	9.3×10^{-10}
1S-NMPCF	1.8×10^{-9}	9.9×10^{-12}	5.5×10^{-12}	2.1×10^{-8}	3.3×10^{-9}	2.9×10^{-6}
T-NMPCF	1.5×10^{-7}	1.7×10^{-8}	5.4×10^{-12}	1.8×10^{-7}	1.2×10^{-7}	2.5×10^{-8}
CNMF	2×10^{-5}	4.4×10^{-4}	3.9×10^{-2}	4.7×10^{-9}	3.4×10^{-4}	5.6×10^{-6}
ST-NMPCF	5.6×10^{-4}	4×10^{-6}	3.7×10^{-4}	1.9×10^{-6}	1.1×10^{-3}	5.2×10^{-6}
IIS_0-NMPCF	3.6×10^{-2}	9.6×10^{-3}	3×10^{-6}	2.5×10^{-3}	2.7×10^{-2}	4.3×10^{-3}

Each cell shows the parameter ρ that represents the probability of setting a statistically significant result. Considering a confidence interval of 95%, small values of $\rho < 0.05$ indicate that there exists statistical significance of the results evaluated.

Finally, a set of spectrograms are presented in Figures 10 and 11 in order to display the sound separation performance obtained by each of the assessed methods. Unlike the other evaluated methods, it can be observed that the proposed method IIS-NMPCF removes most of the RS in the estimated wheezing spectrogram $\hat{\mathbf{X}}_W$ keeping most of the wheezing spectral content. This fact confirms the advantage of the proposed method since most of the clinical useful information contained in the estimated spectrogram $\hat{\mathbf{X}}_W$ will be available to the physician to maximize the reliability of medical diagnosis. The MATLAB implementation of the proposed method is shared by the authors and can be downloaded from GitHub (https://github.com/JTORRECRUZ/Sensors_IIS-NMPCF).

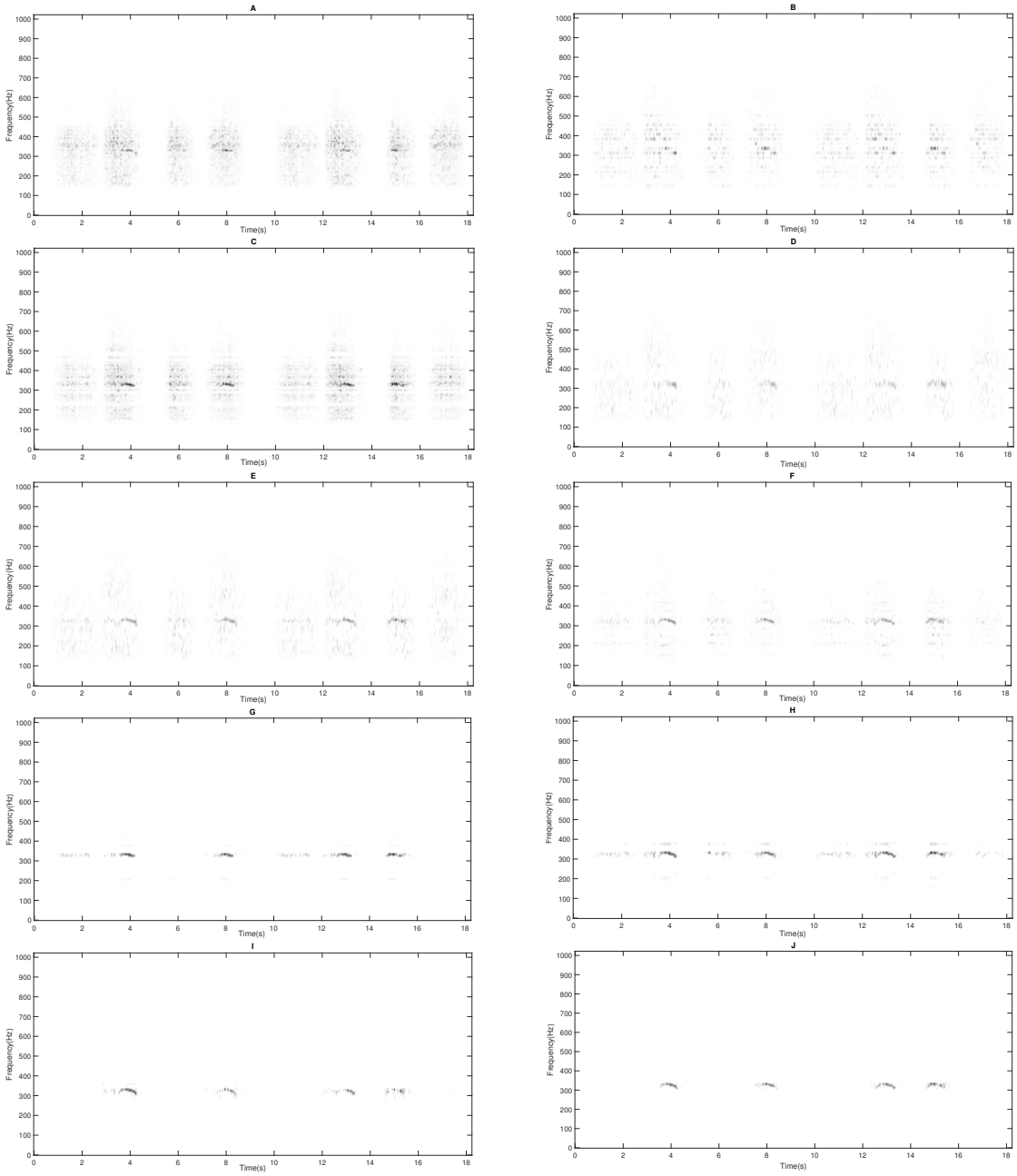

Figure 10. The estimated wheezing spectrogram $\hat{\mathbf{X}}_W$ obtained from the input spectrogram $\mathbf{X}$ shown in Figure 1 for the different methods evaluated. (**A**) NMF, (**B**) SNMF, (**C**) SSNMF, (**D**) 2S-NMPCF, (**E**) 1S-NMPCF, (**F**) T-NMPCF, (**G**) CNMF, (**H**) ST-NMPCF, (**I**) IIS_0-NMPCF and (**J**) IIS-NMPCF.

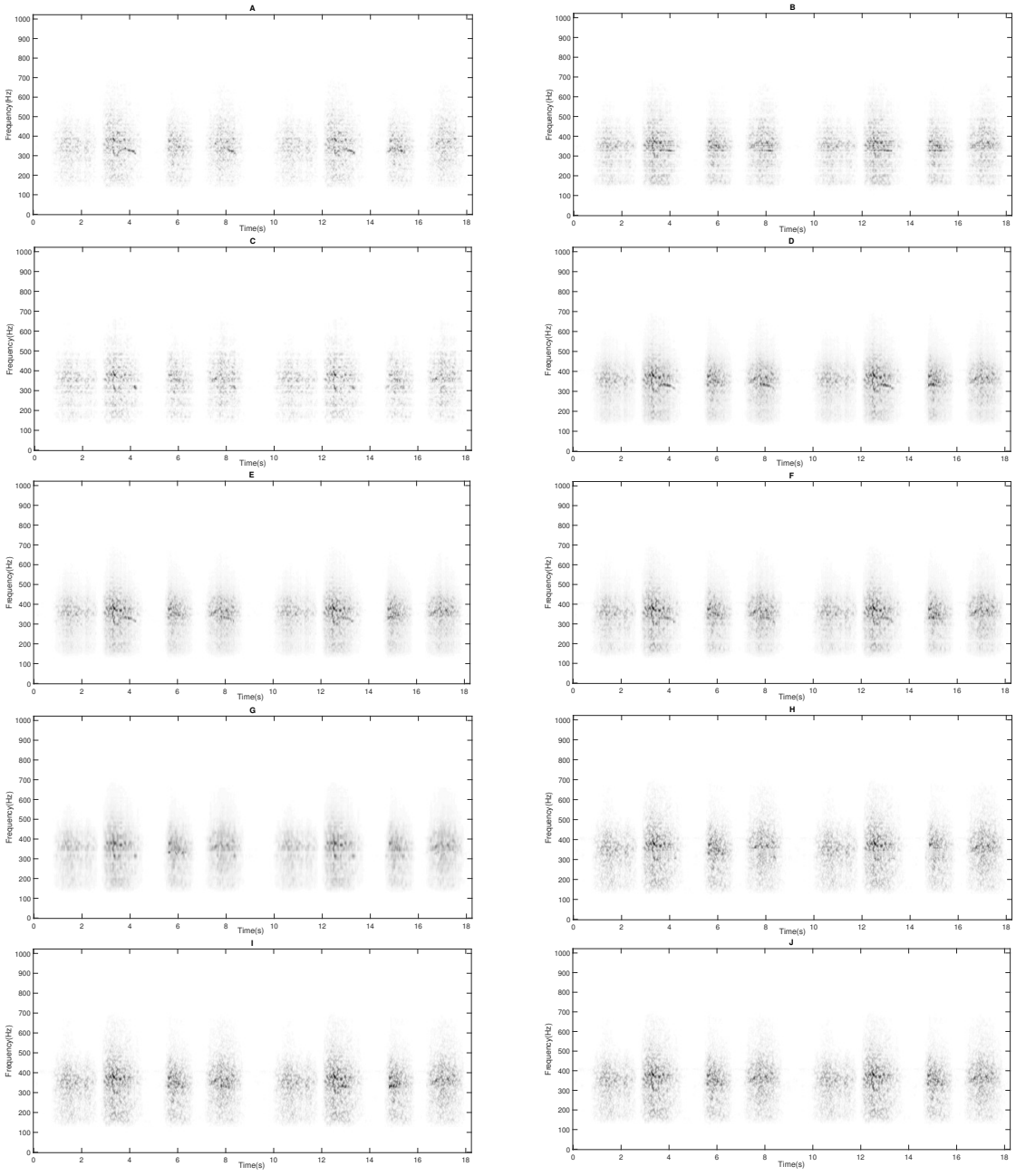

Figure 11. The estimated respiratory spectrogram $\hat{\mathbf{X}}_R$ obtained from the input spectrogram $\mathbf{X}$ shown in Figure 1 for the different methods evaluated. (**A**) NMF, (**B**) SNMF, (**C**) SSNMF, (**D**) 2S-NMPCF, (**E**) 1S-NMPCF, (**F**) T-NMPCF, (**G**) CNMF, (**H**) ST-NMPCF, (**I**) IIS_0-NMPCF and (**J**) IIS-NMPCF.

5. Conclusions

We propose an extended version of Non-negative Matrix Partial Co-Factorization (NMPCF) approach to separate wheezing and respiratory sounds improving their acoustic quality. We assume that RS can be considered as sound events that are repeated during the human breathing process. However, WS may or may not be present along the segments due to the unpredictable nature of the pulmonary disorder. The main contribution of the proposed method is to add importance to the segments classified as non-wheezing to improve the sound separation performance of the conventional NMPCF which treats all segments of the input spectrogram equally. As a result, our proposal (IIS_0-NMPCF/IIS-NMPCF) is able to characterize RS more accurately by forcing to model more on those non-wheezing segments in the bases sharing process into the NMPCF approach.

The main conclusions from the experimental results indicate that adding more importance to the non-wheezing segments into the decomposition procedure (NMPCF) models more accurate the spectro-temporal characteristics related to repetitive sound events of the mixture. In this work, these repetitive sound events are represented by RS that are present in all cycles of the breathing. Experimental SDR, SIR and SAR results report that the proposed method IIS-NMPCF outperforms significantly all evaluated methods providing competitive and promising results in the wheezing sound separation. This fact confirms the ability of the proposed method to improve the sound quality of WS maximizing both the removal of the acoustic interference caused by RS and that as much wheezing content is maintained. As a result, all useful medical information contained in the estimated wheezing can be clearly preserved.

It can be observed that the separation performance for the different evaluated methods drops when the SNR decreases. Considering the acoustic scenario in which RS are louder than WS (SNR < 0 dB), WS are barely audible due to the high interference produced by RS. Although in this case the reduction of the SNR implies a greater time-frequency overlapping from RS to WS, our proposal still achieves the best performance compared to the other baseline methods evaluating. Therefore, the proposed method can be considered an useful tool to be applied in sound environments in which WS are barely audible.

Future work will focus on the development of new constraints to be incorporated into NMF-based approaches for modelling different types of WS according to their spectral content in order to automatically classify the severity of the lung disorder.

Author Contributions: Conceptualization, J.D.L.T.C., F.J.C.Q., N.R.R. and P.V.C.; methodology, J.D.L.T.C., F.J.C.Q., N.R.R. and P.V.C.; software, J.D.L.T.C., F.J.C.Q. and J.J.C.O.; validation, J.D.L.T.C. and J.J.C.O.; writing–original draft, J.D.L.T.C., F.J.C.Q. and J.J.C.O.; writing–review and editing, N.R.R. and P.V.C.; supervision, N.R.R. and P.V.C. All authors have read and agreed to the submitted version of the manuscript.

Funding: This work was supported by the Programa Operativo FEDER Andalucía 2014-2020 under project with reference 1257914.

Acknowledgments: The authors would like to thank the pulmonologist Gerardo Pérez Chica from the University Hospital of Jaén (Spain) for all the constructive discussions about the sound wheezing in the auscultation process.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. World Health Organization. Chronic Respiratory Diseases. Available online: https://www.who.int/health-topics/chronic-respiratory-diseases#tab=tab_1 (accessed on 6 February 2020).
2. Fenton, T.R.; Pasterkamp, H.; Tal, A.; Chernick, V. Automated spectral characterization of wheezing in asthmatic children. *IEEE Trans. Biomed. Eng.* **1985**, *32*, 50–55. [CrossRef] [PubMed]
3. Pramono, R.X.A.; Imtiaz, S.A.; Rodriguez-Villegas, E. Evaluation of features for classification of wheezes and normal respiratory sounds. *PLoS ONE* **2019**, *14*, e0213659. [CrossRef] [PubMed]
4. Pasterkamp, H.; Kraman, S.S.; Wodicka, G.R. Respiratory sounds: Advances beyond the stethoscope. *Am. J. Respir. Crit. Care Med.* **1997**, *156*, 974–987. [CrossRef] [PubMed]
5. Sovijarvi, A.; Dalmasso, F.; Vanderschoot, J.; Malmberg, L.; Righini, G.; Stoneman, S. Definition of terms for applications of respiratory sounds. *Eur. Respir. Rev.* **2000**, *10*, 597–610.
6. Salazar, A.J.; Alvarado, C.; Lozano, F.E. System of heart and lung sounds separation for store-and-forward telemedicine applications. *Rev. Fac. Ing. Univ. Antioq.* **2012**, *64*, 175–181.
7. Forkheim, K.E.; Scuse, D.; Pasterkamp, H. A comparison of neural network models for wheeze detection. In Proceedings of the IEEE WESCANEX 95 Communications, Power, and Computing, Winnipeg, MB, Canada, 15–16 May 1995; Volume 1, pp. 214–219.
8. Wiederhold, B.K.; Cipresso, P.; Pizzioli, D.; Wiederhold, M.; Riva, G. Intervention for physician burnout: A systematic review. *Open Med.* **2018**, *13*, 253–263. [CrossRef]
9. Iskander, M. Burnout, cognitive overload, and metacognition in medicine. *Med. Sci. Educ.* **2019**, *29*, 325–328. [CrossRef]
10. Zhou, Q.; Feng, Z.; Benetos, E. Adaptive Noise Reduction for Sound Event Detection Using Subband-Weighted NMF. *Sensors* **2019**, *19*, 3206. [CrossRef]

11. Emmanouilidou, D.; McCollum, E.D.; Park, D.E.; Elhilali, M. Adaptive noise suppression of pediatric lung auscultations with real applications to noisy clinical settings in developing countries. *IEEE Trans. Biomed. Eng.* **2015**, *62*, 2279–2288. [CrossRef]
12. Homs-Corbera, A.; Fiz, J.A.; Morera, J.; Jané, R. Time-frequency detection and analysis of wheezes during forced exhalation. *IEEE Trans. Biomed. Eng.* **2004**, *51*, 182–186. [CrossRef]
13. Alic, A.; Lackovic, I.; Bilas, V.; Sersic, D.; Magjarevic, R. A novel approach to wheeze detection. In *World Congress on Medical Physics and Biomedical Engineering*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 963–966.
14. Taplidou, S.A.; Hadjileontiadis, L.J. Wheeze detection based on time-frequency analysis of breath sounds. *Comput. Biol. Med.* **2007**, *37*, 1073–1083. [CrossRef] [PubMed]
15. Emrani, S.; Gentimis, T.; Krim, H. Persistent homology of delay embeddings and its application to wheeze detection. *IEEE Signal Process. Lett.* **2014**, *21*, 459–463. [CrossRef]
16. Mendes, L.; Vogiatzis, I.; Perantoni, E.; Kaimakamis, E.; Chouvarda, I.; Maglaveras, N.; Tsara, V.; Teixeira, C.; Carvalho, P.; Henriques, J.; et al. Detection of wheezes using their signature in the spectrogram space and musical features. In Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, Italy, 25–29 August 2015; pp. 5581–5584.
17. Bokov, P.; Mahut, B.; Flaud, P.; Delclaux, C. Wheezing recognition algorithm using recordings of respiratory sounds at the mouth in a pediatric population. *Comput. Biol. Med.* **2016**, *70*, 40–50. [CrossRef]
18. Lozano-García, M.; Fiz, J.A.; Martínez-Rivera, C.; Torrents, A.; Ruiz-Manzano, J.; Jané, R. Novel approach to continuous adventitious respiratory sound analysis for the assessment of bronchodilator response. *PLoS ONE* **2017**, *12*, e0171455. [CrossRef] [PubMed]
19. Nabi, F.G.; Sundaraj, K.; Lam, C.K. Identification of asthma severity levels through wheeze sound characterization and classification using integrated power features. *Biomed. Signal Process. Control* **2019**, *52*, 302–311. [CrossRef]
20. Wisniewski, M.; Zielinski, T.P. Joint application of audio spectral envelope and tonality index in an e-asthma monitoring system. *IEEE J. Biomed. Health Inform.* **2015**, *19*, 1009–1018. [CrossRef] [PubMed]
21. Lozano, M.; Fiz, J.A.; Jané, R. Automatic differentiation of normal and continuous adventitious respiratory sounds using ensemble empirical mode decomposition and instantaneous frequency. *IEEE J. Biomed. Health Inform.* **2015**, *20*, 486–497. [CrossRef]
22. Shaharum, S.M.; Sundaraj, K.; Aniza, S.; Palaniappan, R.; Helmy, K. Classification of asthma severity levels by wheeze sound analysis. In Proceedings of the IEEE Conference on Systems, Process and Control (ICSPC), Bandar Hilir, Malaysia, 16–18 December 2016; pp. 172–176.
23. Pramono, R.X.A.; Imtiaz, S.A.; Rodriguez-Villegas, E. Evaluation of Mel-Frequency Cepstrum for Wheeze Analysis. In Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019; pp. 4686–4689.
24. Mayorga, P.; Druzgalski, C.; Morelos, R.; Gonzalez, O.; Vidales, J. Acoustics based assessment of respiratory diseases using GMM classification. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology, Buenos Aires, Argentina, 31 August–4 September 2010; pp. 6312–6316.
25. Le Cam, S.; Belghith, A.; Collet, C.; Salzenstein, F. Wheezing sounds detection using multivariate generalized Gaussian distributions. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 19–24 April 2009; pp. 541–544.
26. Ulukaya, S.; Serbes, G.; Kahya, Y.P. Wheeze type classification using non-dyadic wavelet transform based optimal energy ratio technique. *Comput. Biol. Med.* **2019**, *104*, 175–182. [CrossRef]
27. Lin, B.S.; Wu, H.D.; Chen, S.J. Automatic wheezing detection based on signal processing of spectrogram and back-propagation neural network. *J. Healthc. Eng.* **2015**, *6*, 649–672. [CrossRef]
28. Kochetov, K.; Putin, E.; Azizov, S.; Skorobogatov, I.; Filchenkov, A. Wheeze detection using convolutional neural networks. In *EPIA Conference on Artificial Intelligence*; Springer: Cham, Switzerland, 2017; pp. 162–173.
29. Jin, F.; Krishnan, S.; Sattar, F. Adventitious sounds identification and extraction using temporal-spectral dominance-based features. *IEEE Trans. Biomed. Eng.* **2011**, *58*, 3078–3087.
30. Riella, R.; Nohama, P.; Maia, J. Method for automatic detection of wheezing in lung sounds. *Braz. J. Med. Biol. Res.* **2009**, *42*, 674–684. [CrossRef] [PubMed]

31. Torre-Cruz, J.; Canadas-Quesada, F.; Vera-Candeas, P.; Montiel-Zafra, V.; Ruiz-Reyes, N. Wheezing Sound Separation Based on Constrained Non-Negative Matrix Factorization. In Proceedings of the 10th International Conference on Bioinformatics and Biomedical Technology (ICBBT), Amsterdam, The Netherlands, 16–18 May 2018; pp. 18–24.
32. Torre-Cruz, J.; Canadas-Quesada, F.; Carabias-Orti, J.; Vera-Candeas, P.; Ruiz-Reyes, N. A novel wheezing detection approach based on constrained non-negative matrix factorization. *Appl. Acoust.* **2019**, *148*, 276–288. [CrossRef]
33. Lee, D.D.; Seung, H.S. Learning the parts of objects by non-negative matrix factorization. *Nature* **1999**, *401*, 788–791. [CrossRef] [PubMed]
34. Lee, D.D.; Seung, H.S. Algorithms for non-negative matrix factorization. In Proceedings of the Advances in Neural Information Processing Systems, Denver, CO, USA, 3–8 December 2001; pp. 556–562.
35. Zafeiriou, S.; Tefas, A.; Buciu, I.; Pitas, I. Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification. *IEEE Trans. Neural Netw.* **2006**, *17*, 683–695. [CrossRef]
36. Benetos, E.; Kotropoulos, C. Non-negative tensor factorization applied to music genre classification. *IEEE Trans. Audio Speech Lang. Process.* **2010**, *18*, 1955–1967. [CrossRef]
37. Févotte, C.; Bertin, N.; Durrieu, J.L. Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. *Neural Comput.* **2009**, *21*, 793–830. [CrossRef]
38. Canadas-Quesada, F.; Ruiz-Reyes, N.; Carabias-Orti, J.; Vera-Candeas, P.; Fuertes-Garcia, J. A non-negative matrix factorization approach based on spectro-temporal clustering to extract heart sounds. *Appl. Acoust.* **2017**, *125*, 7–19. [CrossRef]
39. Laroche, C.; Kowalski, M.; Papadopoulos, H.; Richard, G. A structured nonnegative matrix factorization for source separation. In Proceedings of the 23rd European Signal Processing Conference (EUSIPCO), Nice, France, 31 August–4 September 2015; pp. 2033–2037.
40. Kitamura, D.; Ono, N.; Saruwatari, H.; Takahashi, Y.; Kondo, K. Discriminative and reconstructive basis training for audio source separation with semi-supervised nonnegative matrix factorization. In Proceedings of the 2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC), Xi'an, China, 13–16 September 2016; pp. 1–5.
41. Wang, Z.; Sha, F. Discriminative non-negative matrix factorization for single-channel speech separation. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 3749–3753.
42. Chung, H.; Plourde, E.; Champagne, B. Discriminative training of NMF model based on class probabilities for speech enhancement. *IEEE Signal Process. Lett.* **2016**, *23*, 502–506. [CrossRef]
43. Smaragdis, P.; Raj, B.; Shashanka, M. Supervised and semi-supervised separation of sounds from single-channel mixtures. In *International Conference on Independent Component Analysis and Signal Separation*; Springer: Berlin, Germany, 2007; pp. 414–421.
44. Lee, H.; Yoo, J.; Choi, S. Semi-supervised nonnegative matrix factorization. *IEEE Signal Process. Lett.* **2009**, *17*, 4–7.
45. Lu, N.; Li, T.; Pan, J.; Ren, X.; Feng, Z.; Miao, H. Structure constrained semi-nonnegative matrix factorization for EEG-based motor imagery classification. *Comput. Biol. Med.* **2015**, *60*, 32–39. [CrossRef]
46. Cañadas-Quesada, F.J.; Vera-Candeas, P.; Martinez-Munoz, D.; Ruiz-Reyes, N.; Carabias-Orti, J.J.; Cabanas-Molero, P. Constrained non-negative matrix factorization for score-informed piano music restoration. *Digit. Signal Process.* **2016**, *50*, 240–257. [CrossRef]
47. Carabias-Orti, J.; Canadas-Quesada, F.; Vera-Candeas, P.; Ruiz-Reyes, N. Non-Negative Matrix Factorization (NMF) Applied to Monaural Audio Signal Processing. In *Independent Component Analysis (ICA): Algorithms, Applications and Ambiguities*; Salazar, A., Vergara, L., Eds.; Nova Science Publisher's: Hauppauge, NY, USA, 2018; Chapter 7.
48. Yoo, J.; Kim, M.; Kang, K.; Choi, S. Nonnegative matrix partial co-factorization for drum source separation. In Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 14–19 March 2010; pp. 1942–1945.
49. Kim, M.; Yoo, J.; Kang, K.; Choi, S. Blind rhythmic source separation: Nonnegativity and repeatability. In Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 14–19 March 2010; pp. 2006–2009.

50. Kim, M.; Yoo, J.; Kang, K.; Choi, S. Nonnegative matrix partial co-factorization for spectral and temporal drum source separation. *IEEE J. Sel. Top. Signal Process.* **2011**, *5*, 1192–1204. [CrossRef]
51. Hu, Y.; Liu, G. Separation of singing voice using nonnegative matrix partial co-factorization for singer identification. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *23*, 643–653. [CrossRef]
52. Seichepine, N.; Essid, S.; Févotte, C.; Cappé, O. Soft nonnegative matrix co-factorization. *IEEE Trans. Signal Process.* **2014**, *62*, 5940–5949. [CrossRef]
53. Chen, H.; Yuan, X.; Li, J.; Pei, Z.; Zheng, X. Automatic Multi-Level In-Exhale Segmentation and Enhanced Generalized S-Transform for wheezing detection. *Comput. Methods Progr. Biomed.* **2019**, *178*, 163–173. [CrossRef] [PubMed]
54. Torre-Cruz, J.; Canadas-Quesada, F.; García-Galán, S.; Ruiz-Reyes, N.; Vera-Candeas, P.; Carabias-Orti, J. A constrained tonal semi-supervised non-negative matrix factorization to classify presence/absence of wheezing in respiratory sounds. *Appl. Acoust.* **2020**, *161*, 107–188. [CrossRef]
55. Grais, E.M.; Erdogan, H. Single channel speech music separation using nonnegative matrix factorization and spectral masks. In Proceedings of the 2011 17th International Conference on Digital Signal Processing (DSP), Corfu, Greece, 6–8 July 2011; pp. 1–6.
56. The r.a.l.e. Repository. Available online: http://www.rale.ca (accessed on 6 February 2020).
57. Stethographics Lung Sound Samples. Available online: http://www.stethographics.com (accessed on 6 February 2020).
58. 3 m Littmann Stethoscopes. Available online: https://www.3m.com (accessed on 6 February 2020).
59. East Tennessee State University Pulmonary Breath Sounds. Available online: http://faculty.etsu.edu (accessed on 6 February 2020).
60. ICBHI 2017 Challenge. Available online: https://bhichallenge.med.auth.gr (accessed on 6 February 2020).
61. Lippincott NursingCenter. Available online: https://www.nursingcenter.com (accessed on 6 February 2020).
62. Thinklabs Digital Stethoscope. Available online: https://www.thinklabs.com (accessed on 6 February 2020).
63. Thinklabs Youtube. Available online: https://www.youtube.com/channel/UCzEbKuIze4AI1523_AWiK4w (accessed on 6 February 2020).
64. Emedicine/Medscape. Available online: https://emedicine.medscape.com/article/1894146-overview#a3 (accessed on 6 February 2020).
65. E-learning Resources. Available online: https://www.ers-education.org/e-learning/reference-database-of-respiratory-sounds.aspx (accessed on 6 February 2020).
66. Respiratory Wiki. Available online: http://respwiki.com/Breath_sounds (accessed on 6 February 2020).
67. Easy Auscultation. Available online: https://www.easyauscultation.com/lung-sounds-reference-guide (accessed on 6 February 2020).
68. Colorado State University. Available online: http://www.cvmbs.colostate.edu/clinsci/callan/breath_sounds.htm (accessed on 6 February 2020).
69. Vincent, E.; Gribonval, R.; Févotte, C. Performance measurement in blind audio source separation. *IEEE Trans. Audio Speech Lang. Process.* **2006**, *14*, 1462–1469. [CrossRef]
70. Févotte, C.; Gribonval, R.; Vincent, E. BSS_EVAL Toolbox User Guide—Revision 2.0. 2005, p. 19. Available online: https://hal.inria.fr/inria-00564760 (accessed on 1 May 2020).

Paper 5

Combining a recursive approach via non-negative matrix factorization and Gini index sparsity to improve reliable detection of wheezing sounds

J. Torre-Cruz, F. Canadas-Quesada, J. Carabias-Orti, P. Vera-Candeas and N. Ruiz-Reyes, "Combining a recursive approach via non-negative matrix factorization and Gini index sparsity to improve reliable detection of wheezing sounds", in *Expert Systems with Applications*, Volume 147, June 2020, pp. 113-212. DOI: https://doi.org/10.1016/j.eswa.2020.113212

- Estado: Publicado.
- Revista: Expert Systems with Applications.
- ISSN: 0957-4174.
- Factor de impacto (JCR 2019): 5.452.
- Cuartiles por área de conocimiento:
  - Computer science, artificial intelligence: Q1, 21/137.
  - Engineering, electrical and electronic: Q1, 32/266.

Expert Systems With Applications 147 (2020) 113212

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Combining a recursive approach via non-negative matrix factorization and Gini index sparsity to improve reliable detection of wheezing sounds

Juan De La Torre Cruz*, Francisco Jesús Cañadas Quesada, Julio José Carabias Orti, Pedro Vera Candeas, Nicolás Ruiz Reyes

Department of Telecommunication Engineering, University of Jaen, Campus Cientifico-Tecnologico de Linares, Avda. de la Universidad, s/n, Linares 23700, Jaen, Spain

ARTICLE INFO

Article history:
Received 25 April 2019
Revised 15 January 2020
Accepted 16 January 2020
Available online 16 January 2020

Keywords:
Wheezing
Detection
Non-negative matrix factorization
Gini index
Sparsity
Clustering

ABSTRACT

Auscultation constitutes a fast, non-invasive and low-cost tool widely used to diagnose respiratory diseases in most of the health centres. However, the acoustic training and expertise acquired by the physician is still crucial to provide a reliable diagnosis of the status of the lung. Each wrong diagnosis increases the risk to the health of patients and the costs associated with the treatment of the disease detected. A wheezing detection system can be useful to the physician to minimize the subjectivity of the interpretation of the breathing sounds, misdiagnoses due to stress and elucidating complex acoustic scenes (such as louder background noises). Highlight that the presence of wheeze sounds is one of the main indicators of respiratory disorders from airway obstructions. This work presents an expert and intelligent system to detect wheeze sounds based on a recursive algorithm that combines orthogonal non-negative matrix factorization (ONMF) and the sparsity descriptor Gini index. The recursive algorithm is composed of four stages. The first stage is based on ONMF modelling to factorize the spectral bases as dissimilar as possible. The second stage clusters the ONMF bases into two categories: wheezing and normal breath. The third stage proposes a novel stopping criterion that controls the loss of wheezing spectral content at the expense of removing normal breath content in the recursive algorithm. Finally, the fourth stage determines the patient's condition to locate the temporal intervals in which wheeze sounds are active for unhealthy patients. Experimental results report that the proposed method: (i) provides the best detection performance compared to the recent state-of-the-art wheezing detection approaches, achieving the highest robustness in noisy environments; and (ii) reliably distinguishes the patient's condition (healthy/unhealthy). The strengths of the proposed method are the following: (i) its unsupervised nature since it does not depend on any training stage to learn in advanced the sounds of interest (wheezing). This fact could make this method attractive to be used in clinical settings because wheezing sound databases are often unavailable; and (ii) the modelling of the spectral behaviour by means of a common feature, the sparsity, that represents the typically energy distributions shown by most of the wheeze and normal breath sounds.

1. Introduction

It is well known that the analysis of the sounds, generated during the breathing, indicates the health of the lungs because any abnormal sound overlapped on the normal breath sounds can alert about a respiratory disorder. Wheezing[1] detection is still one of the most challenging research tasks in bio-signal processing (Shaharum, Sundaraj, & Palaniappan, 2012) because wheeze sounds have been considered a reliable indicator of the degree of the bronchial obstruction related to several pulmonary diseases, such as asthma, acute bronchitis, bronchiolitis and chronic obstructive pulmonary disease (COPD) over the last three decades

* Corresponding author.
E-mail addresses: jtorre@ujaen.es (J. De La Torre Cruz), fcanadas@ujaen.es (F.J. Cañadas Quesada), carabias@ujaen.es (J.J. Carabias Orti), pvera@ujaen.es (P. Vera Candeas), nicolas@ujaen.es (N. Ruiz Reyes).

https://doi.org/10.1016/j.eswa.2020.113212
0957-4174/

[1] https://www.merckmanuals.com/home/lung-and-airway-disorders/symptoms-of-lung-disorders/wheezing.

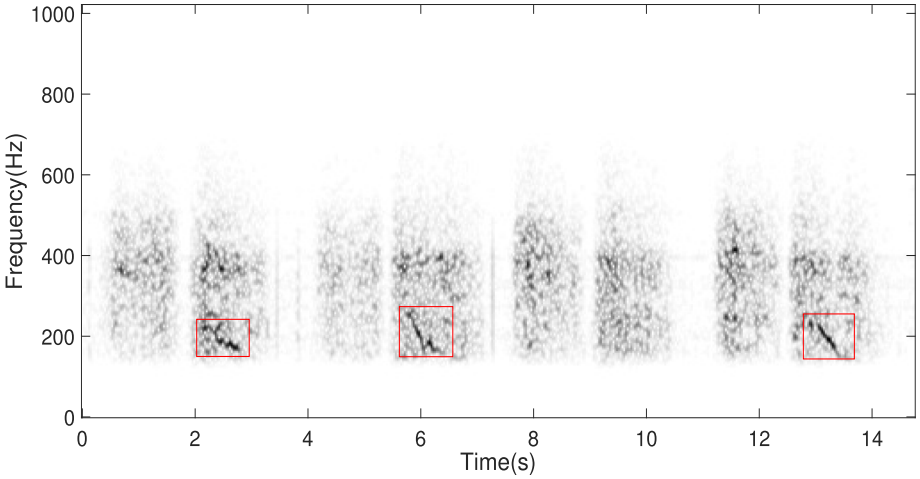

Fig. 1. Spectrogram of a mixture (unhealthy patient) composed of three wheeze sounds (red rectangles) and normal breath sounds. The wheeze sounds start at 2.2s, 5.8s and 12.9s.

(Fenton, Pasterkamp, Tal, & Chernick, 1985; Forkheim, Scuse, & Pasterkamp, 1995; Le Cam, Belghith, Collet, & Salzenstein, 2009; Lin, Wu, & Chen, 2015).

A large percentage of the wheezing detection diagnoses are still not correct today because the subjective evaluation of the physician by means of a stethoscope depends to a large extent on the sound quality analyzed and his/her acoustic training. This misdiagnosis is the main cause of the patient's return to the health center with a worsening of the disease that was not detected in the first medical examination performed by auscultation. For example, asthma affects around 3 million people in Spain and 30% of the diagnoses are not correct. This fact costs the Spanish health system € 2000 per patient and year, compared to € 400 for a correct diagnosed patient[2]. It is therefore essential to develop a robust wheezing detection method in order to early detect respiratory diseases, such as asthma or COPD, characterized all of them by the presence of wheeze sounds. This early detection will increase the patientâ;;s quality of life minimizing the cost of the health center in the treatment of the disease detected. In addition, a wheezing detection method can be useful to physicians to minimize the subjectivity of the interpretation of the breathing sounds and misdiagnoses due to stress and mental fatigue.

Wheezing is one of the continuous adventitious sounds present in lung that is clinically defined as abnormal (Forkheim et al., 1995). Wheeze sounds, associated with obstructions in the airways, are characterized as a set of spectral peaks over time that compose spectral trajectories superimposed on normal breath sounds during the inspiration or expiration. Wheeze and normal breath sounds are mixed together in the time-frequency domain since both of them are simultaneously generated by the same airflow through the bronchial tree of the lungs and share part of the spectral bands in which both types of sounds are active (Torre-Cruz, Canadas-Quesada, Carabias-Orti, Vera-Candeas, & Ruiz-Reyes, 2019). Normal breath sounds or respiratory sounds (RS) are represented by a wideband spectrum where most of the energy is concentrated in the frequency band 60Hz-1000Hz (Salazar, Alvarado, & Lozano, 2012). Wheeze sounds (WS) show musical and sinusoidal behavior which evidences their periodic nature in time-frequency domain. Therefore, WS are characterized as pitched sounds whose pitch frequencies are usually located between 100Hz-1000Hz with duration longer than 100ms according to Computerized Respiratory Sound Analysis (CORSA) (Lin, Lin, Wu, Chong, & Chen, 2006; Pasterkamp, Kraman, & Wodicka, 1997; Sovijarvi et al., 2000; Wisniewski & Zielinski, 2012a) as can be observed in Fig. 1. In this paper, the term RS refers to normal breath sounds in which WS are not active. The term WS refers to wheeze sounds in which RS are not active. The term mixture refers to any single-channel input signal that can be composed only of RS (healthy patient) or RS mixed with WS (unhealthy patient).

Many wheezing detection algorithms, based on different signal processing techniques and spectro-temporal features, have been proposed in the last three decades: Autoregressive (AR) model (Cortes, Jane, Fiz, & Morera, 2006; Jané, Cortes, Fiz, & Morera, 2004), Auditory modelling (Qiu, Whittaker, Lucas, & Anderson, 2005), Entropy (Jin, Sattar, & Goh, 2008; Zhang, Ser, Yu, & Zhang, 2009), Neural networks (NN) (Forkheim et al., 1995; Kochetov, Putin, Azizov, Skorobogatov, & Filchenkov, 2017; Lin et al., 2015), Wavelet transform (Kandaswamy, Kumar, Ramanathan, Jayaraman, & Malmurugan, 2004; Le Cam et al., 2009), Tonal index (Wisniewski & Zielinski, 2012b; 2015), Mel-frequency cepstral coefficients (MFCC) (Bahoura, 2009; Chien, Wu, Chong, & Li, 2007; Shaharum, Sundaraj, Aniza, Palaniappan, & Helmy, 2016), Gaussian Mixture Models (GMM) (Bahoura & Pelletier, 2004; Mayorga, Druzgalski, Morelos, Gonzalez, & Vidales, 2010), Classifiers (Bokov, Mahut, Flaud, & Delclaux, 2016; Jin, Krishnan, & Sattar, 2011; Mazić, Bonković, & Džaja, 2015; Mondal, Banerjee, & Tang, 2018; Ulukaya, Sen, & Kahya, 2015), Spectral peaks identification (Alic, Lackovic, Bilas, Sersic, & Magjarevic, 2007; Fenton et al., 1985; Homs-Corbera, Fiz, Morera, & Jané, 2004; Jain & Vepa, 2008; Mendes et al., 2015; Riella, Nohama, & Maia, 2009; Taplidou & Hadjileontiadis, 2007), Hidden Markov model (HMM) (Oletic & Bilas, 2018) and recently, Non-negative matrix factorization (NMF) (Torre-Cruz et al., 2019). In addition, deep learning for audio classification (Nanni, Costa, Aguiar, Silla Jr., & Brahnam, 2018; Shuvaev, Giaffar, & Alexei, 2017) could also be consider to be extended and adapted for the analysis of adventitious sounds in respiratory mixtures. Kochetov et al. (2017) proposed wheeze detection using convolutional neural networks. Taplidou and Hadjileontiadis (2007) located wheezing based on the subtraction of the underlying breath sound. Wisniewski and Zielinski (2012b) analysed wheeze sounds as a problem of multi-tone detection in colored noise using a set of robust descriptors. Shaharum et al. (2016) detected wheezing to classify different levels of asthma severity using feature extraction based on MFCC. Mazić et al. (2015) developed a two-layer pattern recognition system architecture for asthma wheezing detection. The first layer consists of two SVM classifiers specifically designed as a cascade stacked in parallel using features based on MFCC. The second layer is realized using a digital detection threshold, with the aim of improving the process of wheezing detection. Ulukaya et al. (2015) presented a

[2] https://www.semergen.es/.

monophonic-polyphonic wheeze discrimination system evaluating ratios of quartile frequencies and mean crossing irregularity to exploit the differences in the power spectrum and the periodicity in the time domain. Oletic and Bilas (2018) modeled the temporally-evolving instantaneous frequencies of individual wheezing spectral lines in the time-frequency plane as a random walk using HMM. Torre-Cruz et al. (2019) proposed a two-stage cascade system for wheezing detection. The first stage consists of the use of non-negative matrix factorization incorporating constraints (sparseness and smoothness) to model wheeze and respiratory sounds as reliably as possible that they can be observed in the real life. The second stage is based on the use of the Kullback–Leibler divergence to discriminate between wheezing and respiratory temporal areas. The results showed that the method described is competitive compared to the state-of-the-art methods evaluated.

This manuscript is a significantly extended work of our previous publication (Torre-Cruz et al., 2019). An expert and intelligent system is presented based on the different behaviour shown by WS and RS in the time-frequency domain. Specifically, we propose a recursive algorithm that combines orthogonal non-negative matrix factorization (ONMF) and the use of the Gini index as a spectral sparsity to locate wheezing temporal intervals in respiratory sounds. Since ONMF, compared to NMF, allows to find hidden spectral patterns (bases) that are more faithful to how they occur in the real world by minimizing the redundancy between them, our first contribution classifies the periodic nature of the previous ONMF bases analyzing the sparsity provided by the Gini index in the frequency domain. Our second contribution proposes a recursive algorithm that refines, along the recursive iterations, the initial estimated wheezing spectrogram by means of a novel stopping criterion. As far as the authors knowledge extends, the proposed stopping criterion has never been applied before to any non-negative matrix factorization approach since the proposed criterion analyzes how the wheezing and normal breath spectral content is being distributed in the estimated wheezing spectrogram at each iteration. Specifically, the proposed stopping criterion allows to perform a new iteration in the recursive algorithm when a significant amount of normal breath sounds is removed at the expense of minimizing the loss of significant wheezing sounds contained in the estimated wheezing spectrogram throughout the recursive algorithm. Finally, the third contribution discriminates between healthy and unhealthy patients according to the sparsity shown by the spectral energy distribution of the refined wheezing spectrogram, finding those temporal intervals in which wheezing is active for unhealthy patients. We assume that unhealthy patients show a narrowband spectral energy distribution but healthy patients exhibit a wider spreading energy distribution in the frequency domain.

Although most wheezing detection approaches consider that spectral peaks of WS are louder than RS (Alic et al., 2007; Bokov et al., 2016; Homs-Corbera et al., 2004; Le Cam et al., 2009; Lin et al., 2006; Nagasaka, 2012; Oletic, Arsenali, & Bilas, 2014; Schreur, Vanderschoot, Zwinderman, Dijkman, & Sterk, 1995; Wisniewski & Zielinski, 2012b), the proposed method does not follow this assumption so we assume that wheezing may be more or less loud than normal breath sounds.

The rest of the article is organized as follows. Section 2 briefly introduces non-negative matrix factorization (NMF), the constraint orthogonality and the sparsity based on the Gini index to classify the ONMF bases according to the periodic character shown by wheeze sounds. The proposed wheezing detection method is detailed in Section 3. Experimental datasets, initialization and metrics are described in Section 4. Performance evaluation and results are reported in Section 5, and finally, Section 6 concludes the paper.

2. Theoretical background

2.1. Non-negative matrix factorization

Nonnegative Matrix Factorization (NMF) or unconstrained NMF (Lee & Seung, 1999; 2001) is a widely used rank reduction method that has attracted a lot of attention in image, audio, biomedicine and machine learning in the last decade (Canadas-Quesada, Vera-Candeas, Ruiz-Reyes, Carabias-Orti, & Cabanas-Molero, 2014; Chen, Cichocki, & Rutkowski, 2006; Liu, Wu, Li, Cai, & Huang, 2012; Park, Shin, & Lee, 2017; Torre-Cruz, Canadas-Quesada, Vera-Candeas, Montiel-Zafra, & Ruiz-Reyes, 2018). Focusing on audio, NMF can successfully learn spectrograms due to its ability to obtain additive part-based decompositions of the most representative objects by imposing non-negative constraints. Given an input mixture $x(t)$ with magnitude spectrogram $\mathbf{X} \in \mathbb{R}_{+}^{F \times T}$, the goal of NMF is to approximate $\mathbf{X}$ into the product of two non-negative estimated matrices, basis matrix $\mathbf{B} \in \mathbb{R}_{+}^{F \times K}$ and activation matrix $\mathbf{A} \in \mathbb{R}_{+}^{K \times T}$ as shown in Eq. (1),

$$\mathbf{X} \approx \hat{\mathbf{X}} = \mathbf{BA} \tag{1}$$

where $\hat{\mathbf{X}}$ is the estimated spectrogram and F, T and K represents the number of frequency bins, the number of time frames and the rank or number of components (is generally chosen so that, $K(F+T) \ll FT$ in order to reduce the dimensionality of the data). Therefore, $\mathbf{B}$ can be interpreted as a dictionary of spectral templates (bases or patterns) that compose the input mixture and $\mathbf{A}$ as a matrix of activations (gains) that indicates the time intervals in which the previous spectral templates are active.

Generally, the NMF decomposition is performed by minimizing a cost function $D(\mathbf{X}|\hat{\mathbf{X}})$ that penalizes the error between $\mathbf{X}$ and $\hat{\mathbf{X}}$. This minimization ensures the nonnegativity of the bases and activations using an iterative algorithm based on multiplicative update rules. Popular cost functions are the Euclidean distance, the generalized Kullback–Leibler divergence, the Itakura–Saito divergence and the Cauchy distribution (Févotte, Bertin, & Durrieu, 2009; Liutkus, Fitzgerald, & Badeau, 2015). In this paper, we propose to minimize the generalized Kullback–Liebler divergence $D_{KL}(\mathbf{X}|\hat{\mathbf{X}})$ because previous works (Canadas-Quesada, Ruiz-Reyes, Carabias-Orti, Vera-Candeas, & Fuertes-Garcia, 2017; Torre-Cruz et al., 2018) have obtained promising results in audio processing,

$$D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}\right) = \sum_{f=1}^{F}\sum_{t=1}^{T} X_{f,t} \log \frac{X_{f,t}}{\hat{X}_{f,t}} - X_{f,t} + \hat{X}_{f,t} \tag{2}$$

However, unconstrained NMF cannot automatically assign each basis to its sound source since non-negativity ensures convergence to local minima that only enables the reconstruction of the mixture. Thus, this reconstruction does not guarantee that the obtained factorization is composed of parts-based objects with physical interpretation as occurs in real world (Laroche, Kowalski, Papadopoulos, & Richard, 2015). The following two ways are usually used to find better solutions from those provided by unconstrained NMF: i) adding prior information of the sound sources into the NMF factorization by means of additional constraints (see Section 2.2); and ii) using descriptors applied to NMF bases or activations in order to classify the components that satisfy a given criterion (see Section 2.3).

2.2. Orthogonality

The constraint orthogonality ϕ integrated into the NMF factorization procedure, orthogonal NMF (ONMF), has improved the separation performance in audio processing (Cañadas-Quesada et al., 2016; Li, Hou, Zhang, & Cheng, 2001) since it enforces to decorrelate the NMF components. It can be applied to both basis matrix

B and activations matrix **A**. If the orthogonality $\phi(\mathbf{B})$ is applied to NMF bases, $\mathbf{B}^T\mathbf{B} = I$ must be fulfilled, in other words, $\phi(\mathbf{B}) = \mathbf{B}^T\mathbf{B} - I$ must be minimized being ^{T} the transpose operator and I the identity matrix. Adding $\phi(\mathbf{B})$ into the NMF factorization procedure,

$$D(\mathbf{X}|\hat{\mathbf{X}}) = D_{KL}(\mathbf{X}|\hat{\mathbf{X}}) + \phi(\mathbf{B}) = D_{KL}(\mathbf{X}|\hat{\mathbf{X}}) + \frac{1}{2}Trace\left(\mathbf{B}^T\mathbf{B} - I\right) \quad (3)$$

The operator *Trace* computes the sum of diagonal elements of the square matrix $\mathbf{B}^T\mathbf{B}$. As a result, $\phi(\mathbf{B})$ factorizes bases as dissimilar (orthogonal) as possible, minimizing the redundancy between them. As a result, ONMF shows better sound source separation performance since it is able to factorize a wider set of true spectral templates active in the input mixture (Grais & Erdogan, 2013).

2.3. Sparsity

It is known that a sparse representation contains a large percentage of the signal energy in a small number of coefficients. Therefore, it can be considered that the most sparse distribution is that in which all the energy is contained in a single coefficient and the rest of coefficients are zero. On the other hand, the least sparse distribution would be the one in which all the energy is uniformly distributed among all the coefficients.

Unlike normal breath sounds, wheezing sounds are characterized by sinusoidal or tonal behavior. This periodicity can be used to classify between wheeze and normal breath sounds assuming that a wheezing spectrum shows a sparse distribution of energy. The reason is because periodicity implies spectral peaks in which energy are located in narrowband instead of normal breath sounds in which the energy are concentrated in wideband.

Many sparse descriptors have been proposed in audio processing to classify periodic and non-periodic behavior (Aydore, Sen, Kahya, & Mihcak, 2009; Canadas-Quesada et al., 2014; Hurley & Rickard, 2009; Park et al., 2017; Torre-Cruz et al., 2018; Wisniewski & Zielinski, 2011; 2012a; 2012b). In this work, we propose to use the Gini index β, which measures the inequality of a distribution, as sparse descriptor because it has been demonstrated its reliability and robustness to correctly cluster between sparse and non-sparse distributions (Feng, Xiao, & Wei, 2013; Hurley & Rickard, 2009; Park et al., 2017). Specifically, β is applied on the spectral ONMF bases in order to classify them between periodic and non-periodic bases. Thus, the descriptor is mathematically defined for a spectral vector b as follows,

$$\beta = \frac{F+1}{F} - \frac{2\sum_{f=1}^{F}(F+1-f)b_f^{(sorted)}}{F\sum_{f=1}^{F}b_f^{(sorted)}} \quad (4)$$

where b^{sorted} denotes b sorted in ascending order. An important advantage of β is that it is normalized, and assumes values between 0 and 1 for any vector. The value of the descriptor β increases as a distribution concentrates all its energy in a single coefficient, in other words, increase its sparse behavior ($\beta \to 1$). However, the value of the descriptor β decreases as the energy of a distribution is more evenly distributed, that is, reduce its sparse behavior ($\beta \to 0$).

3. Proposed method of wheezing detection

The aim of the proposed method is to detect wheeze sounds from single-channel audio mixtures. In this work, the term detection implies the time location of the intervals in which the wheezing sounds are active. The proposed method is a recursive method based on the periodicity principle to discriminate between wheezing and normal breath spectral templates assuming that a wheeze sound exhibits a strongly periodic or tonal nature which is characterized by the presence of narrowband spectral peaks. As can be observed in Fig. 2, the proposed method is composed of four stages: obtaining ONMF bases, clustering of the ONMF bases, stop criterion of the recursive method and determine the patient's condition (e.g., healthy/unhealthy).

3.1. Stage I: Obtaining ONMF bases

Using a time-frequency (T-F) representation, the spectrogram **X** of the input signal $x(t)$, obtained from unhealthy/healthy patient, has been computed by means of the magnitude of the Short-Time Fourier Transform (STFT) applying a Hamming window of size N with 25% overlap. First, a normalization process must be applied to make the proposed model independent of the size and scale of the input spectrogram **X**. Specifically, the normalized magnitude spectrogram $\mathbf{X}_n$ is computed as follows,

$$\mathbf{X}_n = \frac{\mathbf{X}}{\left(\frac{\sum_{f=1}^{F}\sum_{t=1}^{T}X_{f,t}}{FT}\right)} \quad (5)$$

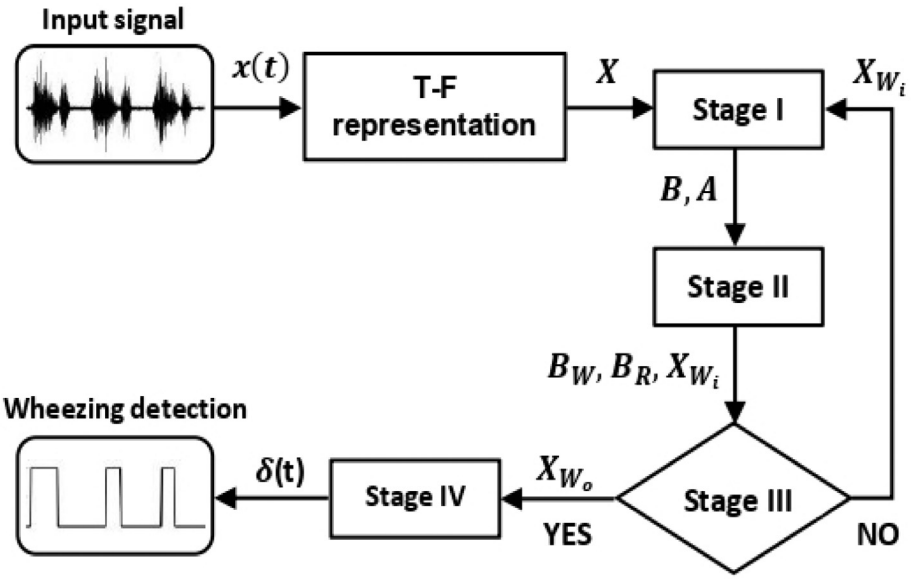

Fig. 2. The block diagram of the proposed method.

This stage attempts to estimate the basis **B** and activations **A** matrix minimizing the reconstruction error between the input spectrogram $\mathbf{X}_n$ and the estimated spectrogram $\hat{\mathbf{X}}_n$. We are interested in using the constraint orthogonality into the NMF, ONMF, because it enforces to decorrelate the bases of the dictionary improving the clustering interpretation. Specifically, the orthogonality constraint $\phi(\mathbf{B})$ factorizes bases (spectral templates) as dissimilar (orthogonal) as possible with the aim of minimizing redundancy between them. The previous constraint $\phi(\mathbf{B})$ (see Section 2.2) can be reformulated using the penalty term shown in Eq. (6),

$$\phi(\mathbf{B}) = \min_{(b_q,b_j)\in\mathbf{B}} = \sum_{b_q,b_j} < b_q, b_j > \tag{6}$$

being $< b_q, b_j >$ the dot product operator between the qth and jth bases of the dictionary **B**. It can be observed that the minimization of $\phi(\mathbf{B})$ provides a set of orthogonal bases, that is, each basis is orthogonal to the rest of them as is similarly the case in Li et al. (2001). ONMF obtains better sound source clustering performance compared to NMF because ONMF is able to factorize the most distinct true spectral templates that composed the input mixture.

According to these observations, the global objective function $D(\mathbf{X}_n|\hat{\mathbf{X}}_n)$ must be minimized using the Kullback-Leibler divergence $D_{KL}(\mathbf{X}_n|\hat{\mathbf{X}}_n)$ and the orthogonality constraint $\phi(\mathbf{B})$, as can be seen in Eq. (7).

$$\begin{aligned} D(\mathbf{X}_n|\hat{\mathbf{X}}_n) &= D_{KL}(\mathbf{X}_n|\hat{\mathbf{X}}_n) + \phi(\mathbf{B}) \\ &= D_{KL}(\mathbf{X}_n|\hat{\mathbf{X}}_n) + \frac{1}{2}Trace\left(\mathbf{B}^T\mathbf{B} - I\right) \end{aligned} \tag{7}$$

Finally, the basis matrix **B** (see Eq. (8)) and the activation matrix **A** (see Eq. (9)) can be obtained by applying a gradient descent algorithm based on multiplicative update rules (Ding, Li, Peng, & Park, 2006; Yoo & Choi, 2010) to the global objective function $D(\mathbf{X}_n|\hat{\mathbf{X}}_n)$ until the algorithm converges after M_i iterations.

$$\mathbf{B} \leftarrow \mathbf{B} \odot \sqrt{\frac{\mathbf{X}_n\mathbf{A}^T}{\mathbf{B}\mathbf{B}^T\mathbf{X}_n\mathbf{A}^T}} \tag{8}$$

$$\mathbf{A} \leftarrow \mathbf{A} \odot \frac{\mathbf{B}^T\mathbf{X}_n}{\mathbf{B}^T\mathbf{B}\mathbf{A}} \tag{9}$$

where **A** and **B** are initialized as random positive matrices, and $\odot$ is the Hadamard product (calculated element-wise) (Ding et al., 2006; Lee & Seung, 2001; Yoo & Choi, 2010).

3.2. Stage II: Clustering of the ONMF bases

Based on the periodic and non-periodic nature of a wheeze and normal breath sound, we propose to apply the sparse descriptor Gini index in the frequency domain in order to discriminate between wheezing and normal breath ONMF bases, denoted as $\mathbf{B}_W$ and $\mathbf{B}_R$, respectively. The periodic nature observed in the previous factorized bases allows them to be clustered into two sets: ONMF bases that show higher periodicity (see Fig. 3a) and lower periodicity (see Fig. 3b). This higher periodicity tends to concentrate the energy of the ONMF bases in narrowband spectral peaks (see Fig. 3c) while non-periodicity distributes such energy in wideband spectrum (see Fig. 3d).

The sparse descriptor Gini index β, defined in Section 2.3, calculates the degree of periodicity of each ONMF basis. High values of β imply high periodic nature (that is, WS) and small values of β indicates non-periodic nature (that is, RS). A thresholding process is used to classify the bases into two groups, ($\mathbf{B}_W$ and $\mathbf{B}_R$), according to the degree of periodicity calculated. Specifically, we have used a threshold ζ_m equal to the median of the sparse values provided by the sparse descriptor Gini index β for each basis kth of the dictionary **B** since this type of thresholding has been widely used in image processing (Toh & Isa, 2010) and audio processing (Rafii & Pardo, 2013) providing better discrimination performance between periodic and non-periodic bases.

Highlight that ONMF bases cannot be exactly factorized into normal breath bases or wheezing bases because all bases, to a greater or lesser extent, have information corresponding to WS and RS. However, these bases can be labelled as periodic or non-periodic bases according to the level of periodicity calculated by the sparse descriptor Gini index and the thresholding process. Preliminary results indicated that ζ_m obtains a promising bases classification because it has the advantage of discriminate uniformly between two groups, providing half of the input bases for both groups $\mathbf{B}_R$ and $\mathbf{B}_W$. Thus, the kth ONMF basis $\mathbf{B}(k)$ belongs to WS when $\beta(k) \geq \zeta_m$. However, $\mathbf{B}(k)$ belongs to RS when $\beta(k) < \zeta_m$ as Eq. (10).

$$\mathbf{B}(k) \rightarrow \begin{cases} \text{belongs to } \mathbf{B}_W(k_c) & \text{if} \quad \beta(k) \geq \zeta_m \\ \text{belongs to } \mathbf{B}_R(k_c) & \text{if} \quad \beta(k) < \zeta_m \end{cases} \tag{10}$$

where k=1,...,K and k_c=1,...,K/2.

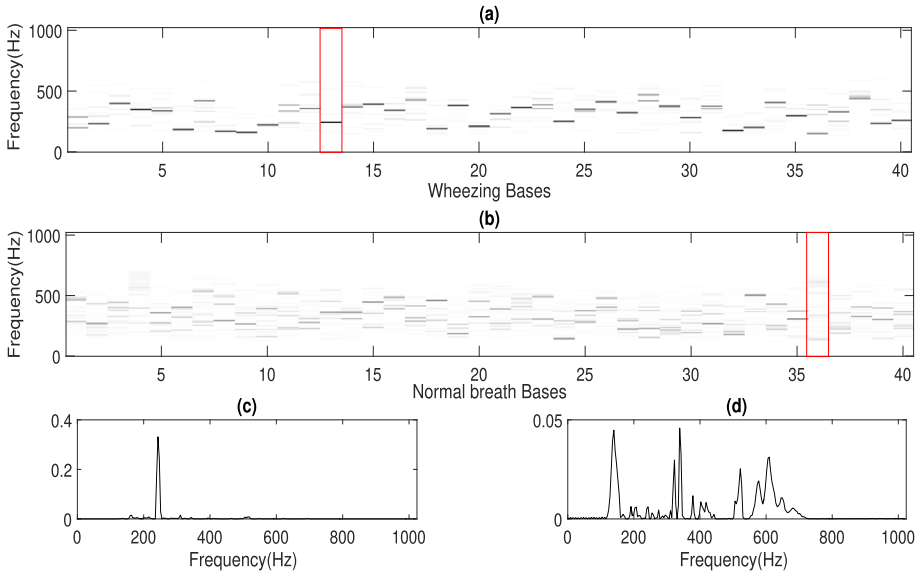

Fig. 3. **(a)** The set of ONMF bases that show higher periodicity $\mathbf{B}_W$ associated to WS. **(b)** The set of ONMF bases that show lower periodicity $\mathbf{B}_R$ associated to RS. **(c)** The spectral energy distribution of the most periodical ONMF basis B_{13} shown in Fig. 3a (red rectangle). **(d)** The spectral energy distribution of the least periodical ONMF basis B_{36} shown in Fig. 3b (red rectangle).

Once the ONMF bases matrix $\mathbf{B}$ has been clustered between $\mathbf{B}_W$ (see Fig. 3a) and $\mathbf{B}_R$ (see Fig. 3b), the estimated wheezing spectrogram $\mathbf{X}_W$ can be reconstructed, as follows:

$$\mathbf{X}_W = \mathbf{B}_W \mathbf{A}_W \tag{11}$$

where the wheezing activations matrix $\mathbf{A}_W$ has been created using the same set of components classified in the wheezing bases matrix $\mathbf{B}_W$. In the next section, the previous estimated wheezing spectrogram $\mathbf{X}_W$ is equal to estimated wheezing spectrogram $\mathbf{X}_{W_1}$ related to the first recursive iteration.

3.3. Stage III: Stop criterion of the recursive method

In this stage, a recursive method is proposed that consisting of the stage I, II and III to improve the estimated wheezing spectrogram $\mathbf{X}_{W_1}$. At each recursive iteration i, the spectrogram $\mathbf{X}_{W_i}$ is the input of the stage I in recursion $i+1$ in order to factorize a new set of ONMF bases to be recluster into wheezing bases or normal breath bases according to the discrimination performed by the sparse descriptor Gini index β. Like this, this process is repeated using the estimated wheezing spectrogram $\mathbf{X}_{W_i}$, output of the stage II, at each recursive iteration i until a novel proposed stop criterion is fulfill. The goal of this proposed criterion is to maintain the significant spectral content of wheezing along recursive iterations while the spectral content of normal breath sounds is reduced. Nevertheless, an initial evaluation confirmed that part of the remaining wheezing energy is lost at each recursive iteration. Based on this observation, this stage attempts to find the optimal iteration i_o that achieves the greatest amount of significant WS and the least amount of RS. Our proposal considers the worst case and measures the spectral difference, by means of γ_i, between the least periodic basis β_{B_W} clustered into the wheezing bases $\mathbf{B}_W$ and the most periodic normal breath basis β_{B_R} clustered into the normal breath bases $\mathbf{B}_R$ at the recursive iteration i, as follows:

$$\gamma_i = min\{\beta_{B_W}\} - max\{\beta_{B_R}\} \tag{12}$$

High values of γ_i indicate that the proposed method is able to cluster the ONMF bases $\mathbf{B}$ computed in the recursive iteration i into two groups ($\mathbf{B}_W$ and $\mathbf{B}_R$) whose periodic behavior is sufficiently different. As a result, the estimated wheezing spectrogram $\mathbf{X}_{W_i}$ will have reduced its RS energy at the expense of maintaining the WS energy active in the input mixture $x(t)$. On the other hand, low values of γ_i indicate the presence of RS bases $\mathbf{B}_R$ that show a strong periodic nature (that is, there is at least one RS basis with a periodicity similar to the minimum periodicity shown by a WS basis). In this case, the estimated wheezing spectrogram $\mathbf{X}_{W_i}$ loses significant wheezing energy making the wheezing detection task more difficult.

A threshold ζ_s is defined to determine the optimal iteration i_o that implies the stop of the recursive process. In this paper, the threshold ζ_s is calculated by applying a percentage ρ to the spectral difference between the least periodic wheezing basis and the most periodic normal breath basis for the first recursive iteration γ_1 (see Eq. (13)). This threshold allows the proposed method to be adjusted to the periodicity values of the ONMF bases according to each particular mixture $x(t)$. Specifically, it is confirmed that the best performance was found empirically (for more details about the set of recordings used in the analysis, see Section 4.1) to be $p = 0.1$.

$$\zeta_s = \rho\gamma_1 \tag{13}$$

Finally, the optimal iteration is equal to $i_o = i - 1$ which occurs when $\gamma_i \leq \zeta_s$ (see Eq. (14)). This fact demonstrates that the periodicity calculated to the normal breath basis that exhibits the highest periodic behavior can be considered similar to the periodicity calculated to the wheezing basis that exhibits the lowest periodic behavior. As a result, the estimated wheezing spectrogram at the recursive iteration i begins to lose significant wheezing spectral content. To correctly detect wheezing, the optimal iteration i_o corresponds with the previous iteration $i-1$.

$$i_o = \left(\underset{\gamma_i \leq \zeta_s}{\operatorname{argmin}}\, i \right) - 1 \tag{14}$$

As can be seen in Fig. 4a, the periodicity associated to $\mathbf{B}_W$ and $\mathbf{B}_R$ increases at each recursive iteration because part of the RS are eliminated along recursive iterations advance. However, as can be seen in Fig. 4b, the spectral difference decreases in each recursive iteration. When the recursive iteration $i = 5$, the reader can observe that the spectral difference begins to converge and the wheezing and normal breath bases show the same periodic character. Therefore in order to guarantee a discrimination based on the periodicity between the wheezing and normal breath bases, the proposed method must stop in the previous recursive iteration, in this case, $i_o = 4$. As shown in Fig. 5, the estimated wheezing spectrogram $\mathbf{X}_{W_i}$ improves along the recursive iterations until reaching the estimated wheezing spectrogram $\mathbf{X}_{W_o}$ at the optimal iteration i_o. Nevertheless, in the next recursive iteration $i = 5$ (see Fig. 5f), it can be clearly displayed that $\mathbf{X}_{W_i}$ has lost a significant part of the wheezing energy.

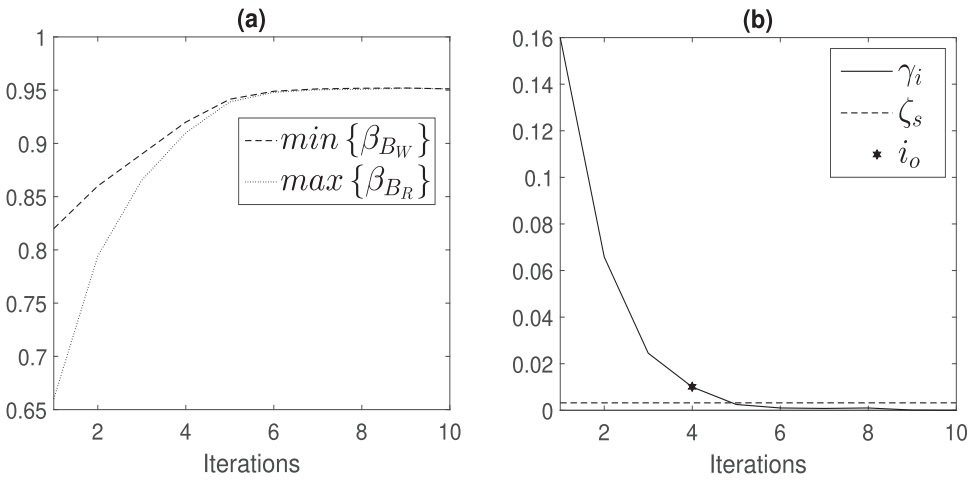

Fig. 4. **(a)** β value for the $min\{\beta_{B_W}\}$ and $max\{\beta_{B_R}\}$ along the recursive iteration applied to the input spectrogram $\mathbf{X}$ shown in Fig. 1. **(b)** The optimal iteration i_o based on γ_i and ζ_s.

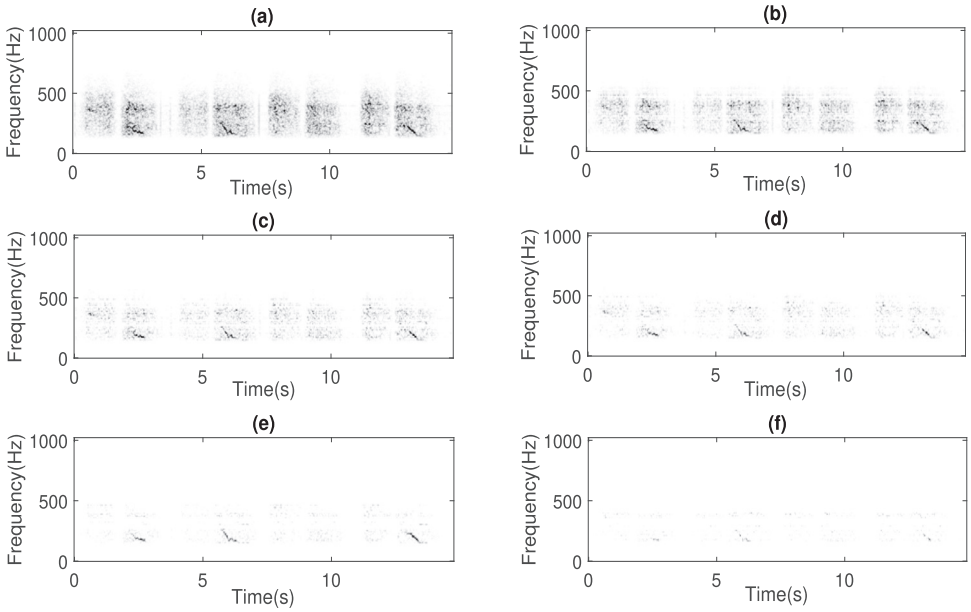

Fig. 5. **(a)** Magnitude spectrogram $\mathbf{X}$ shown in Fig. 1. The estimated wheezing spectrogram $\mathbf{X}_{W_i}$ output of the stage II at each recursive iteration i. **(b)** $i = 1$, **(c)** $i = 2$, **(d)** $i = 3$, **(e)** $i = 4$ and **(f)** $i = 5$. In this example, $i_o = 4$.

3.4. Stage IV: Determine the patient's condition

The main goal of this stage is to determine the patient's condition, that is, to decide whether the patient is a healthy or unhealthy patient. For this purpose, we propose to analyze the sparse behaviour of the spectral energy distribution ϵ for the estimated wheezing spectrogram $\mathbf{X}_{W_o}$ obtained from the optimal recursive iteration i_o as follows,

$$\epsilon(f) = \sum_{t=1}^{T} \mathbf{X}_{W_{o_{f,t}}} \tag{15}$$

where f=1,...F. In this manner, ϵ represents a vector compose of the spectral energy distribution along frames.

We assume that the spectral energy distribution ϵ is concentrated in narrowband in the case of an unhealthy patient, i.e. a typical spectral peak from a periodic signal. On the other hand, the spectral energy distribution ϵ is concentrated in wideband in the case of a healthy patient.

In a similar way as the sparse descriptor Gini index β was previously applied in stage II, the stage IV uses the sparse metric β_ϵ to determine whether the spectral energy distribution ϵ is related to an unhealthy or healthy patient. Specifically, β_ϵ is defined as the sparse descriptor β applied to the spectral energy distribution ϵ associated to the estimated wheeze spectrogram $\mathbf{X}_{W_o}$.

Based on the fact that β is normalized between 0 and 1, the value $\beta_\epsilon = 0$ corresponds to a perfectly healthy patient and $\beta_\epsilon = 1$ to a perfectly unhealthy patient. Our empirical observations, in randomly selected recordings from the most widely used Internet pulmonary repositories (for more details about the set of recordings used in the analysis, see Section 4.1), reported the following facts: (i) ϵ is often concentrated in narrowband in the case of unhealthy patients (see Fig. 6e) and its value of $\beta_\epsilon > 0.5$; (ii) ϵ is often concentrated in wideband in the case of healthy patients

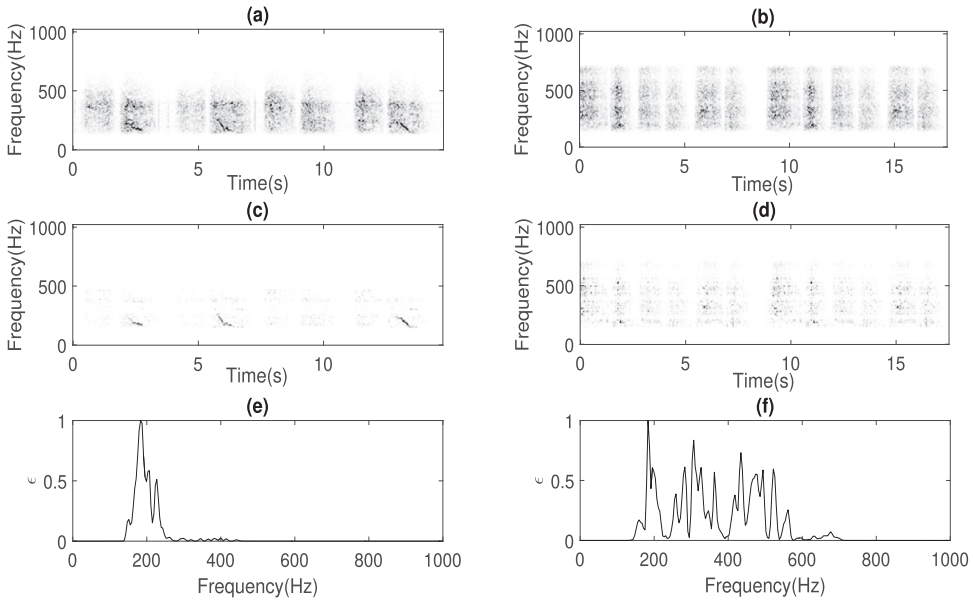

Fig. 6. **(a)** Magnitude spectrogram $\mathbf{X}$ from an unhealthy patient shown in Fig. 1. **(b)** Magnitude spectrogram $\mathbf{X}$ of a healthy patient. **(c)** The estimated wheezing magnitude spectrogram $\mathbf{X}_{W_o}$ for an unhealthy patient. **(d)** The estimated wheezing magnitude spectrogram $\mathbf{X}_{W_o}$ for a healthy patient. **(e)** The spectral energy distribution ϵ for an unhealthy patient. **(f)** The spectral energy distribution ϵ for a healthy patient.

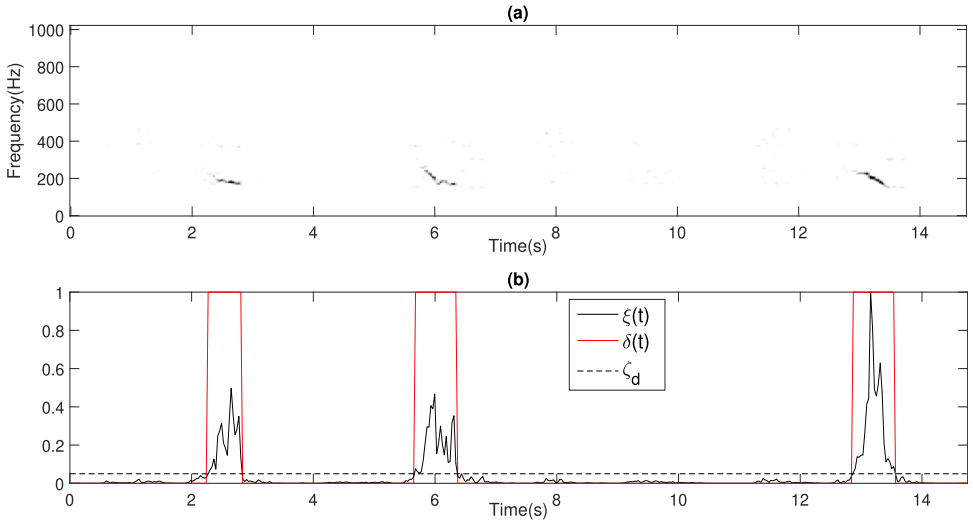

Fig. 7. **(a)** Prominent estimated wheezing spectrogram $\mathbf{X}_{W_p}$ for the input spectrogram $\mathbf{X}$ shown in Fig. 1. **(b)** Estimated wheezing detection $\delta(t)$ using $\xi(t)$ and ζ_d. Note that $\xi(t)$ has been normalized to adjust the values between 0 and 1.

(see Fig. 6f) and its value of $\beta_\epsilon < 0.5$. For this reason, we set the threshold $\zeta_h = 0.5$ in order to determine the patient's condition as follows,

$$\text{patient's condition} = \begin{cases} unhealthy & \text{if } \beta_\epsilon \geq \zeta_h \\ healthy & \text{if } \beta_\epsilon < \zeta_h \end{cases} \tag{16}$$

Taking into account the patient's condition previously obtained, the final goal of this stage is to locate the temporal intervals or areas (the term area denotes a group of continuous temporal frames) in which wheezing is active from the estimated wheezing spectrogram $\mathbf{X}_{W_o}$ obtained as the output of the stage III in the optimal recursive iteration i_o. A wheezing frame t, in which wheezing is active, provides a value $\delta(t) = 1$. On the other hand, a normal breath (non-wheezing) frame t, in which wheezing is inactive, provides a value $\delta(t) = 0$.

$$\delta(t) = \begin{cases} 1 & \text{frame with wheezing active} \\ 0 & \text{frame with wheezing inactive} \end{cases} \tag{17}$$

where t=1,...T.

In the case of healthy patient, the output of the wheezing detection is $\delta(t) = 0, \forall t$. In the case of unhealthy patient, we propose to emphasize the energy values corresponding to the wheezing frames using $\mathbf{X}_{W_p} = \mathbf{X}_{W_o}^2$, being $\mathbf{X}_{W_p}$ the prominent estimated wheezing spectrogram. As can be seen in Fig. 7a, the energy located in wheezing frames have been highlighted in order to improve the identification of areas in which wheezing is active.

Due to the fact that $\mathbf{X}_{W_p}$ presents almost the totality of the WS with the least presence of RS, we propose to calculate the energy $\xi(t)$ frame-by-frame of $\mathbf{X}_{W_p}$ to discriminate the wheezing areas, as can be observed in Eq. (18).

$$\xi(t) = \sum_{f=1}^{F} \mathbf{X}_{W_{p_{f,t}}} \tag{18}$$

where t=1,...T.

A threshold ζ_d is defined to determine the wheezing areas. A preliminary analysis (for more details about the set of recordings used in the analysis, see Section 4.1) showed that the best wheezing detection performance was achieved using a threshold $\zeta_d = 0.05$ to guarantee the maximum detection of wheezing frames losing the least amount of them. Finally, each frame of the unhealthy mixture $\mathbf{X}$ is labelled as wheezing (wheezing active) or not (wheezing inactive) by means of $\delta(t)$, combining $\xi(t)$ and the threshold ζ_d (see Fig. 7b) as follows,

$$\delta(t) = \begin{cases} 1 & \text{if } \xi(t) \geq \zeta_d \\ 0 & \text{if } \xi(t) < \zeta_d \end{cases} \tag{19}$$

where t=1,...T.

The pseudo code of the proposed method for the wheezing detection is detailed in the Algorithm 1. Specifically, the stage I covers the steps 1, 2 and 3; the stage II covers the steps 4, 5, 6 and 7; the stage III covers the steps 8, 9, 10 and 11; and the stage IV covers the steps 12, 13, 14, 15 and 16. Highlight that the recursive algorithm starts at step 2 and ends at step 8.

Algorithm 1 Proposed method for wheezing detection

Require: $x(t)$, M_i, K, p, ζ_h and ζ_d.
1: Compute the normalized magnitude spectrogram $\mathbf{X}_n$ using Eq. (5).
2: Initialize $\mathbf{B}$ and $\mathbf{A}$ with random nonnegative values.
3: Update $\mathbf{B}$ and $\mathbf{A}$ using Eq. (8)-(9) and M_i iterations.
4: Compute β for each basis k^{th} of the dictionary $\mathbf{B}$ using Eq. (4).
5: Compute the threshold ζ_m.
6: Cluster the ONMF bases $\mathbf{B}$ into $\mathbf{B}_W$ and $\mathbf{B}_R$ using Eq. (10).
7: Reconstruct the estimated wheezing spectrogram $\mathbf{X}_{W_i}$ using Eq. (11).
8: Compute the spectral difference γ_i using Eq. (12).
9: Compute the threshold ζ_s using Eq. (13).
10: Repeat steps 2-8 until $\gamma_i \leq \zeta_s$. Note that $\mathbf{X}_{W_i}$ is the input of the stage I in the recursive iteration $i+1$.
11: Select the estimated wheezing spectrogram $\mathbf{X}_{W_o}$ (see Eq. (14)).
12: Compute the spectral energy distribution ϵ using Eq. (15).
13: Compute β_ϵ for the spectral energy distribution ϵ usingEq. (4).
if $\beta_\epsilon < \zeta_h$ **then**
  return $\delta(t) = 0, \forall t$. ▷ Healthy patient
else
  14: Calculate $\mathbf{X}_{W_p} = \mathbf{X}_{W_o}^2$.
  15: Calculate $\xi(t)$ using Eq. (18).
  16: Obtain $\delta(t)$ using Eq. (19).
  return $\delta(t)$. ▷ Unhealthy patient
end if

Table 1
Characteristics of each dataset: identifier (ID1); number of recordings (ID2); number of recordings captured from healthy/unhealthy patients (ID3); the shortest and longest duration, in seconds, captured from recordings (ID4); total duration in seconds (ID5); the lowest and highest SNR (in dB) between WS and RS (ID6); the minimum and maximum number of respiratory events found in the recordings (ID7); total number of respiratory events (ID8); the minimum and maximum number of wheezes found in the recordings (ID9); total number of wheezes (ID10).

ID1	ID2	ID3	ID4	ID5	ID6	ID7	ID8	ID9	ID10
T1 (Oletic & Bilas, 2018)	16	8/8	4-51	230	[2-8]	[4-20]	168	[3-7]	36
T2H (Torre-Cruz et al., 2019)	16	0/16	7-22	251	5	[6-14]	126	[1-5]	41
T2M (Torre-Cruz et al., 2019)	16	0/16	7-22	251	0	[6-14]	126	[1-5]	41
T2L (Torre-Cruz et al., 2019)	16	0/16	7-22	251	-5	[6-14]	126	[1-5]	41

4. Experimental setup

4.1. Datasets

To determine the value of the parameters ρ, ζ_h and ζ_d, a set of preliminary analyses were performed using 96 randomly selected recordings from the most widely used Internet pulmonary repositories (R.A.L.E repository[3], Stethographics lung sound samples[4], 3m littmann stethoscopes[5], East tennessee state university pulmonary breath sounds[6], ICBHI 2017 Challenge[7], Lippincott NursingCenter[8], Thinklabs Digital Stethoscope[9], Thinklabs youtube[10], Emedicine/Medscape[11], E-learning resources[12], Respiratory wiki[13], Easy Auscultation[14], Colorado State University[15]). In total, the previous selected recordings provide 1442 s of respiratory sounds, 48 healthy patients, 48 unhealthy patients, 784 respiratory events (a respiratory event is defined as an inspiration or expiration) and 92 wheezes. Highlight that the recordings used for fixing the parameters ρ, ζ_h and ζ_d are not a part of any testing datasets in order to validate the results.

The testing datasets T1 and T2 (T2H, T2M and T2L) are used to assess the wheezing detection performance of the proposed method in order to provide a detection comparison with some of the most recent state-of-the-art algorithms.

The dataset T1 has been directly shared by authors (Oletic & Bilas, 2018). T1 is composed of eight recordings captured from unhealthy patients and eight recordings captured from healthy patients. The goal of this dataset is to evaluate the task of wheezing detection determining the patient's condition. More details of T1 can be found in the section Dataset in Oletic and Bilas (2018).

The datasets T2H, T2M and T2L (Torre-Cruz et al., 2019) have been directly used to evaluate the robustness of the proposed method taking into account different signal-to-noise (SNR) environments in wheezing detection. Specifically, the datasets T2H (SNR=5dB), T2M (SNR=0dB) and T2L (SNR=-5dB) represent the same dataset T2 but using different SNR between WS and RS. Using the visual inspection of spectrograms, each dataset T2H, T2M and T2L were created mixing only WS recordings manually separated (by means of a time-frequency mask applied to the mixture spectrogram to select only the bins of each frame corresponding to wheezing) and only RS recordings (in which wheezing is inactive). More details of T2H, T2M, T2L can be found in Torre-Cruz et al. (2019).

Considering all the previous databases (T1, T2H, T2M and T2L), it is provided 983 s of recording, 8 healthy patients, 24 unhealthy patients, 546 respiratory events and 77 wheezes. The characteristics of the datasets are detailed in Table 1. In order to validate the results, it is indicated that T1 is not a part of T2.

4.2. Initializations

As previously mentioned, we assume that WS are not active below 100Hz and above 1000Hz. Thus, all mixtures were band-limited from 100Hz-1000Hz. The experimental results show that the following parameters provide the best trade-off between the wheezing detection performance and the computational cost: sampling rate $f_s = 2048$Hz, Hamming window with $N = 256$ samples length and 25% overlap (temporal resolution of 31.3 ms), a discrete Fourier transform using $2N$ points (frequency resolution of 4Hz). Furthermore, the ONMF convergence was empirically achieved after 120 iterations for all signals, so $M_i = 120$ iterations has been used. Moreover, a number of components $K = 80$ have been selected since preliminary results indicated the best wheezing detection performance.

4.3. Evaluation Metrics

Three metrics are used to evaluate the performance of the proposed method, which are commonly used in the field of wheezing detection (Mazić et al., 2015; Oletic & Bilas, 2018; Shaharum et al., 2016; Theodoridis & Koutroumbas, 2006; Torre-Cruz et al., 2019): sensitivity $SE = \frac{TP}{TP+FN}$, the probability of detecting wheezing frames correctly; specificity $SP = \frac{TN}{TN+FP}$, the probability of detecting normal breath frames correctly; and accuracy $ACC = \frac{(TP+TN)}{(TP+FP+FN+TN)}$, the probability of detecting wheezing/normal breath frames correctly. The terms TP, FN, FP, and TN are the amount of the true positive, false negative, false positive and true negative test results, respectively.

4.4. State-of-the-art wheezing detection methods for comparison

Four recent state-of-the-art wheezing detection methods have been used to evaluate the performance of the proposed method: HMMFL (Oletic & Bilas, 2018), TSVM (Mazić et al., 2015), MKNN (Shaharum et al., 2016) and CNMF (Torre-Cruz et al., 2019). MKNN and TSVM are supervised approaches. However, CNMF, HMMFL and the proposed method do not use any type of training due to their unsupervised approach. The wheezing detection results shown by the previous state-of-the-art methods are been directly taken from (Torre-Cruz et al., 2019).

[3] http://www.rale.ca.
[4] http://www.stethographics.com.
[5] https://www.3m.com.
[6] http://faculty.etsu.edu.
[7] https://bhichallenge.med.auth.gr.
[8] https://www.nursingcenter.com.
[9] https://www.thinklabs.com.
[10] https://www.youtube.com/channel/UCzEbKuIze4AI1523_AWiK4w.
[11] https://emedicine.medscape.com/article/1894146-overview#a3.
[12] https://www.ers-education.org/e-learning/.
[13] http://respwiki.com/Breath_sounds.
[14] https://www.easyauscultation.com.
[15] http://www.cvmbs.colostate.edu/clinsci/callan/.

5. Results

In this section, the wheezing detection performance of the proposed method is evaluated. Table 2 shows SE, SP and ACC results evaluating the dataset T1 between the proposed method and the aforementioned state-of-the-art methods. The proposed method outperforms the detection performance compared to the baseline methods in terms of ACC and SP. Specifically, the proposed method achieves a significant improvement of approximately 0.26% (CNMF), 1.21% (HMMFL), 5.53% (TSVM) and 7.64% (MKNN), in terms of ACC. Focusing on the SP metric, the proposed method accomplishes a significant improvement of approximately 4.29%, 1.03%, 1.95% and 4.04% versus CNMF, HMMFL, TSVM and MKNN, respectively. However, the method CNMF improves the proposed method in 1.47% in terms of SE. This is due to the fact that the constraints used by CNMF have been optimized to detect the entire wheeze time interval at the expense of providing a greater number of false positives, in other words, frames related to normal breath sounds mistakenly detected as wheezing frames. Because the proposed method achieves better SP compared to the SE results shown in Table 2, it suggests that our proposal tends to increase the number of false negatives in exchange for detecting exactly the full range of frames in which wheezing is active. Nevertheless, the proposed method is still obtaining the best ACC results because it is based on a recursive process that avoids the loss of the spectral content of wheezing while eliminating most of the content of normal breath sounds that act as spurious sounds in the wheezing detection task. A remarkable strength shown by the method is that it is the only method of compared methods that has correctly detected as healthy all healthy patient recordings belonging to the dataset T1. This fact confirms that the proposed method can be considered a reliable wheezing detection method to determine the patient's condition.

Table 3 shows SE, SP and ACC results evaluating three datasets T2H (SNR=5dB), T2M (SNR=0dB) and T2L (SNR=-5dB), each of them with a different SNR to compare the robustness of the task of wheezing detection of the proposed method and the state-of-the-art methods. Results report that the proposed method provides the best overall detection results compared to the other evaluated methods considering all SNR scenarios evaluated. Focusing on the ACC results of the proposed method, the following can be observed:

- Dataset T2H: the ACC improvement of the proposed method is about 0.18% (CNMF), 18.06% (HMMFL), 3.46% (TSVM) and 10% (MKNN).
- Dataset T2M: the ACC improvement of the proposed method is about 2.07% (CNMF), 20.27%(HMMFL), 5.84% (TSVM) and 10.19% (MKNN).
- Dataset T2L: the ACC improvement of the proposed method is about 0.65% (CNMF), 23.54% (HMMFL), 8.9% (TSVM) and 13.34% (MKNN).

In the same way, comparing the SP results of the proposed method with the baseline methods:

Table 2
Wheezing detection comparison evaluating the dataset T1.

Method	SE (%)	SP (%)	ACC (%)
Proposed Method	94.24	**97.31**	**96.12**
CNMF (Torre-Cruz et al., 2019)	**95.71**	93.02	95.86
HMMFL (Oletic & Bilas, 2018)	89.34	96.28	94.91
TSVM (Mazić et al., 2015)	85.32	95.36	90.59
MKNN (Shaharum et al., 2016)	80.86	93.27	88.48

Each value in bold indicates the highest value obtained in each column.

Table 3
Wheezing detection performance comparison evaluating the datasets T2H, T2M and T2L.

Method	SE (%)	SP (%)	ACC (%)
Dataset T2H (SNR=5dB)			
Proposed Method	94.23	**97.22**	**98.12**
CNMF (Torre-Cruz et al., 2019)	**99.48**	90.77	97.41
HMMFL (Oletic & Bilas, 2018)	79.50	88.62	80.06
TSVM (Mazić et al., 2015)	90.38	95.52	94.66
MKNN (Shaharum et al., 2016)	82.33	90.84	88.12
Dataset T2M (SNR=0dB)			
Proposed Method	93.61	**96.29**	**97.13**
CNMF (Torre-Cruz et al., 2019)	**97.27**	88.60	95.06
HMMFL (Oletic & Bilas, 2018)	73.83	85.86	76.86
TSVM (Mazić et al., 2015)	88.60	92.46	91.29
MKNN (Shaharum et al., 2016)	78.68	87.62	86.94
Dataset T2L (SNR=-5dB)			
Proposed Method	92.74	**94.99**	**95.35**
CNMF (Torre-Cruz et al., 2019)	**95.57**	87.97	94.70
HMMFL (Oletic & Bilas, 2018)	70.72	78.94	71.81
TSVM (Mazić et al., 2015)	81.93	89.96	86.45
MKNN (Shaharum et al., 2016)	73.77	85.12	82.01

- Dataset T2H: the SP improvement of the proposed method is about 6.45% (CNMF), 8.6% (HMMFL), 1.7% (TSVM) and 6.38% (MKNN).
- Dataset T2M: the SP improvement of the proposed method is about 7.69% (CNMF), 10.43% (HMMFL), 4.83% (TSVM) and 8.67% (MKNN).
- Dataset T2L: the SP improvement of the proposed method is about 7.02% (CNMF), 16.05% (HMMFL), 5.03% (TSVM) and 9.87% (MKNN).

These results show that there is a tendency for the proposed method to obtain greater improvements and robustness in the wheezing detection compared to the rest of the methods evaluated as the SNR of the acoustic scenario is reduced, since this SNR reduction implies a greater sound masking of RS compared to WS (remind that SNR $<$ 0dB makes RS louder than WS). As occurs in Table 2, SE results of the proposed method is ranked in second position in Table 3, showing a value of 94.23% in T2H, 93.61% in T2M and 92.74% in T2L. Results show the robustness both the proposed method and CNMF in noisy environments compared to the other evaluated methods. While SE, SP and ACC results of the proposed method are reduced an average of approximately 2% comparing T2H vs T2L, the same results of HMMFL, TSVM and MKNN drops approximately 9% comparing the same datasets. Moreover, TSVM and MKNN obtains a better detection performance compared to HMMFL taking into account SE, SP and ACC. The reason seems to be that HMMFL is based on tracking multiple individual wheeze frequency lines using hidden Markov model (HMM). Therefore, the spectral peaks related to WS have to be louder than RS so that the tracking of the wheeze frequency lines are not interfered by RS. In other words, the detection performance of HMMFL decreases considerably in noisy environments where the energy of RS is greater than the energy of WS. Although, TSVM and MKNN are methods based on machine learning that employ features based on MFCC. TSVM obtains better detection performance compared to MKNN because TSVM consists of two SVM classifiers specifically designed as a cascade stacked in parallel to ensure wheezing detection avoiding false positives and false negatives as much as possible.

6. Conclusions andfuture work

In this paper, we propose a novel recursive method based on orthogonal non-negative matrix factorization (ONMF) applied

to wheezing detection in single-channel breath sound mixtures. The first contribution proposes to classify the periodic (wheezing sounds) and non-periodic (normal breath sounds) bases obtained ONMF decomposition using the sparse descriptor Gini index in the frequency domain in order to model the periodic behavior of wheeze sounds. The second contribution presents a novel recursive method to improve the estimated wheezing spectrogram from breath sounds. This recursive process is applied to the estimated wheezing spectrogram of each recursive iteration. A novel stop criterion is proposed to avoid the loss of the spectral content of wheezing while eliminating most of the content of normal breath sounds that act as spurious sounds in the wheezing detection task. The third contribution determines the condition of each patient evaluated (healthy or unhealthy) analyzing the sparse behaviour of the spectral energy distribution of the estimated wheezing spectrogram obtained from the optimal recursive iteration. Finally, the wheezing frames are located.

The main conclusions derived from the experimental results indicate that: (i) the wheezing detection results drop as the SNR decreases. In a similar way as occurs in real life, it is more complex to identify wheeze sounds in those acoustic scenarios in which wheeze sounds are barely audible due to the high interference caused by normal breath sounds; (ii) the proposed method provides the best overall detection results compared to the other state-of-the-art methods considering all SNR scenarios evaluated. Comparing T2H and T2L, while SE, SP and ACC results of the proposed method only drops 2%, results provided by the other state-of-the-art methods drops 9%; (iii) the proposed method is more robust than baseline methods as the SNR conditions worsen (SNR $<$ 5dB). Highlight that the proposed method is able to achieve a promising wheezing detection even when the normal breath sounds are louder than wheezing sounds; (iv) the reliable ability of the proposed method to determine the patient's condition: healthy or unhealthy; and (v) unlike other approaches based on machine learning, the proposed method does not depend on any training dataset due to its unsupervised approach.

There are some limitations of the proposed method for wheezing detection: (i) the proposal has been designed only to analyze recordings composed of RS and WS. In recordings with background noises (such as speech, cough, crying, a door opens or closes, etc) or other adventitious sounds (such as crackles, stridor, pleural rubs, rhonchi, etc), the detection results would be negatively affected; (ii) in the case that the stage IV of the proposed method results in a false negative (unhealthy patient misdiagnosed as healthy patient) the time intervals in which the wheezing sounds are active cannot be located; and (iii) better SP results compared to SE results since the recursive algorithm tends to detect exactly the full time intervals in which wheezing is active at the expense of increasing the number of false negatives (frames related to wheezing sounds mistakenly annotated as normal breath frames).

In future work, we will focus on improving the tasks of detection and sound quality related to the wheezing sounds as both tasks are implicitly linked. Better wheezing detection implies to reduce the physicians mental fatigue by more carefully analyzing only those time intervals in which the wheezing sounds have been detected. Better acoustic quality of the wheezing sounds implies to enhance the diagnosis by basing it on the acoustic fidelity of the spectral content that has been heard. Specifically, an extension of this work will be focused on the improvements in both detection and sound quality of the wheezing developing a factorization approach in which the temporal repetitive behaviour shown by normal breath sounds will be added as a novel constraint into the decomposition procedure. This new constraint could be applied in complex acoustic scenes in this research field to remove loud background noises active in the physicians room.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Credit authorship contribution statement

Juan De La Torre Cruz: Investigation, Software, Data curation, Writing - original draft, Writing - review & editing. **Francisco Jesús Cañadas Quesada:** Investigation, Methodology, Resources, Writing - original draft. **Julio José Carabias Orti:** Investigation, Conceptualization, Visualization, Writing - review & editing. **Pedro Vera Candeas:** Supervision, Formal analysis, Writing - review & editing. **Nicolás Ruiz Reyes:** Project administration, Supervision.

Acknowledgment

The authors would like to thank Dr. Dinko Oletic and Dr. Vedran Bilas for sharing their dataset called T1 in this manuscript.

References

Alic, A., Lackovic, I., Bilas, V., Sersic, D., & Magjarevic, R. (2007). A novel approach to wheeze detection. In *World congress on medical physics and biomedical engineering* (pp. 963–966). Springer.

Aydore, S., Sen, I., Kahya, Y. P., & Mihcak, M. K. (2009). Classification of respiratory signals by linear analysis. In *Annual international conference of the ieee engineering in medicine and biology society (embc)* (pp. 2617–2620). IEEE.

Bahoura, M. (2009). Pattern recognition methods applied to respiratory sounds classification into normal and wheeze classes. *Computers in biology and medicine, 39*(9), 824–843.

Bahoura, M., & Pelletier, C. (2004). Respiratory sounds classification using gaussian mixture models. In *Canadian conference on electrical and computer engineering: 3* (pp. 1309–1312). IEEE.

Bokov, P., Mahut, B., Flaud, P., & Delclaux, C. (2016). Wheezing recognition algorithm using recordings of respiratory sounds at the mouth in a pediatric population. *Computers in biology and medicine, 70*, 40–50.

Canadas-Quesada, F., Ruiz-Reyes, N., Carabias-Orti, J., Vera-Candeas, P., & Fuertes–Garcia, J. (2017). A non-negative matrix factorization approach based on spectro-temporal clustering to extract heart sounds. *Applied Acoustics, 125*, 7–19.

Cañadas-Quesada, F. J., Vera-Candeas, P., Martinez-Munoz, D., Ruiz-Reyes, N., Carabias-Orti, J. J., & Cabanas-Molero, P. (2016). Constrained non-negative matrix factorization for score-informed piano music restoration. *Digital Signal Processing, 50*, 240–257.

Canadas-Quesada, F. J., Vera-Candeas, P., Ruiz-Reyes, N., Carabias-Orti, J., & Cabanas–Molero, P. (2014). Percussive/harmonic sound separation by non-negative matrix factorization with smoothness/sparseness constraints. *Journal on Audio, Speech, and Music Processing, 2014*(26), 1–17.

Chen, Z., Cichocki, A., & Rutkowski, T. M. (2006). Constrained non-negative matrix factorization method for eeg analysis in early detection of alzheimer disease. In *Ieee international conference on acoustics speech and signal processing proceedings: 5*. IEEE.

Chien, J.-C., Wu, H.-D., Chong, F.-C., & Li, C.-I. (2007). Wheeze detection using cepstral analysis in gaussian mixture models. In *29th annual international conference of the ieee engineering in medicine and biology society (embc)* (pp. 3168–3171). IEEE.

Cortes, S., Jane, R., Fiz, J., & Morera, J. (2006). Monitoring of wheeze duration during spontaneous respiration in asthmatic patients. In *27th annual international conference of the ieee engineering in medicine and biology society (embc)* (pp. 6141–6144). IEEE.

Ding, C., Li, T., Peng, W., & Park, H. (2006). Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th acm sigkdd international conference on knowledge discovery and data mining* (pp. 126–135). ACM.

Feng, C., Xiao, L., & Wei, Z. (2013). Compressive sensing isar imaging with stepped frequency continuous wave via gini sparsity. In *Ieee international geoscience and remote sensing symposium (igarss)* (pp. 2063–2066). IEEE.

Fenton, T. R., Pasterkamp, H., Tal, A., & Chernick, V. (1985). Automated spectral characterization of wheezing in asthmatic children. *IEEE transactions on biomedical engineering, 32*(1), 50–55.

Févotte, C., Bertin, N., & Durrieu, J.-L. (2009). Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural computation, 21*(3), 793–830.

Forkheim, K. E., Scuse, D., & Pasterkamp, H. (1995). A comparison of neural network models for wheeze detection. In *Ieee wescanex 95. communications, power, and computing. conference proceedings: 1* (pp. 214–219). IEEE.

Grais, E. M., & Erdogan, H. (2013). Discriminative nonnegative dictionary learning using cross-coherence penalties for single channel source separation.. In *Interspeech* (pp. 808–812).

Homs-Corbera, A., Fiz, J. A., Morera, J., & Jané, R. (2004). Time-frequency detection and analysis of wheezes during forced exhalation. *IEEE Transactions on Biomedical Engineering, 51*(1), 182–186.
Hurley, N., & Rickard, S. (2009). Comparing measures of sparsity. *IEEE Transactions on Information Theory, 55*(10), 4723–4741.
Jain, A., & Vepa, J. (2008). Lung sound analysis for wheeze episode detection. In *30th annual international conference of the ieee engineering in medicine and biology society (embc)* (pp. 2582–2585). IEEE.
Jané, R., Cortes, S., Fiz, J., & Morera, J. (2004). Analysis of wheezes in asthmatic patients during spontaneous respiration. In *26th annual international conference of the ieee engineering in medicine and biology society (embc): 2* (pp. 3836–3839). IEEE.
Jin, F., Krishnan, S., & Sattar, F. (2011). Adventitious sounds identification and extraction using temporal-spectral dominance-based features. *IEEE Transactions on Biomedical Engineering, 58*(11), 3078–3087.
Jin, F., Sattar, F., & Goh, D. Y. (2008). Automatic wheeze detection using histograms of sample entropy. In *30th annual international conference of the ieee engineering in medicine and biology society (embc)* (pp. 1890–1893). IEEE.
Kandaswamy, A., Kumar, C. S., Ramanathan, R. P., Jayaraman, S., & Malmurugan, N. (2004). Neural classification of lung sounds using wavelet coefficients. *Computers in biology and medicine, 34*(6), 523–537.
Kochetov, K., Putin, E., Azizov, S., Skorobogatov, I., & Filchenkov, A. (2017). Wheeze detection using convolutional neural networks. In *Epia conference on artificial intelligence* (pp. 162–173). Springer.
Laroche, C., Kowalski, M., Papadopoulos, H., & Richard, G. (2015). A structured nonnegative matrix factorization for source separation. In *23rd european signal processing conference (eusipco)* (pp. 2033–2037). IEEE.
Le Cam, S., Belghith, A., Collet, C., & Salzenstein, F. (2009). Wheezing sounds detection using multivariate generalized gaussian distributions. In *Ieee international conference on acoustics, speech and signal processing* (pp. 541–544). IEEE.
Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature, 401*(6755), 788–791.
Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems* (pp. 556–562).
Li, S. Z., Hou, X., Zhang, H., & Cheng, Q. (2001). Learning spatially localized, parts-based representation. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1*, 207–212.
Lin, B.-S., Lin, B.-S., Wu, H.-D., Chong, F.-C., & Chen, S.-J. (2006). Wheeze recognition based on 2d bilateral filtering of spectrogram. *Biomedical Engineering: Applications, Basis and Communications, 18*(03), 128–137.
Lin, B.-S., Wu, H.-D., & Chen, S.-J. (2015). Automatic wheezing detection based on signal processing of spectrogram and back-propagation neural network. *Journal of healthcare engineering, 6*(4), 649–672.
Liu, H., Wu, Z., Li, X., Cai, D., & Huang, T. S. (2012). Constrained nonnegative matrix factorization for image representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*(7), 1299–1311.
Liutkus, A., Fitzgerald, D., & Badeau, R. (2015). Cauchy nonnegative matrix factorization. In *Ieee workshop on applications of signal processing to audio and acoustics (waspaa)* (pp. 1–5). IEEE.
Mayorga, P., Druzgalski, C., Morelos, R., Gonzalez, O., & Vidales, J. (2010). Acoustics based assessment of respiratory diseases using gmm classification. In *Annual international conference of the ieee engineering in medicine and biology* (pp. 6312–6316). IEEE.
Mazić, I., Bonković, M., & Džaja, B. (2015). Two-level coarse-to-fine classification algorithm for asthma wheezing recognition in children's respiratory sounds. *Biomedical Signal Processing and Control, 21*, 105–118.
Mendes, L., Vogiatzis, I., Perantoni, E., Kaimakamis, E., Chouvarda, I., Maglaveras, N., … Henriques, J., et al. (2015). Detection of wheezes using their signature in the spectrogram space and musical features. In *37th annual international conference of the ieee engineering in medicine and biology society (embc)* (pp. 5581–5584). IEEE.
Mondal, A., Banerjee, P., & Tang, H. (2018). A novel feature extraction technique for pulmonary sound analysis based on emd. *Computer methods and programs in biomedicine, 159*, 199–209.
Nagasaka, Y. (2012). Lung sounds in bronchial asthma. *Allergology International, 61*(3), 353–363.
Nanni, L., Costa, Y., Aguiar, R., Silla Jr. , C., & Brahnam, S. (2018). Ensemble of deep learning, visual and acoustic features for music genre classification. *Journal of New Music Research, 47*(4), 383–397.
Oletic, D., Arsenali, B., & Bilas, V. (2014). Low-power wearable respiratory sound sensing. *Sensors, 14*(4), 6535–6566.
Oletic, D., & Bilas, V. (2018). Asthmatic wheeze detection from compressively sensed respiratory sound spectra. *IEEE journal of biomedical and health informatics, 22*(5), 1406–1414.
Park, J., Shin, J., & Lee, K. (2017). Exploiting continuity/discontinuity of basis vectors in spectrogram decomposition for harmonic-percussive sound separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25*(5), 1061–1074.
Pasterkamp, H., Kraman, S. S., & Wodicka, G. R. (1997). Respiratory sounds: advances beyond the stethoscope. *American journal of respiratory and critical care medicine, 156*(3), 974–987.
Qiu, Y., Whittaker, A., Lucas, M., & Anderson, K. (2005). Automatic wheeze detection based on auditory modelling. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine, 219*(3), 219–227.
Rafii, Z., & Pardo, B. (2013). Repeating pattern extraction technique (repet): A simple method for music/voice separation. *IEEE transactions on audio, speech, and language processing, 21*(1), 73–84.
Riella, R., Nohama, P., & Maia, J. (2009). Method for automatic detection of wheezing in lung sounds. *Brazilian Journal of Medical and Biological Research, 42*(7), 674–684.
Salazar, A. J., Alvarado, C., & Lozano, F. E. (2012). System of heart and lung sounds separation for store-and-forward telemedicine applications. *Revista Facultad de Ingeniería Universidad de Antioquia*, (64), 175–181.
Schreur, H., Vanderschoot, J., Zwinderman, A., Dijkman, J., & Sterk, P. (1995). The effect of methacholine-induced acute airway narrowing on lung sounds in normal and asthmatic subjects. *European Respiratory Journal, 8*(2), 257–265.
Shaharum, S. M., Sundaraj, K., Aniza, S., Palaniappan, R., & Helmy, K. (2016). Classification of asthma severity levels by wheeze sound analysis. In *Ieee conference on systems, process and control (icspc)* (pp. 172–176). IEEE.
Shaharum, S. M., Sundaraj, K., & Palaniappan, R. (2012). A survey on automated wheeze detection systems for asthmatic patients. *Bosnian journal of basic medical sciences, 12*(4), 249–255.
Shuvaev, S., Giaffar, H., & Alexei, K. (2017). Representations of sound in deep learning of audio features from music. *ArXiv, abs/1712.02898*.
Sovijarvi, A., Dalmasso, F., Vanderschoot, J., Malmberg, L., Righini, G., & Stoneman, S. (2000). Definition of terms for applications of respiratory sounds. *European Respiratory Review, 10*(77), 597–610.
Taplidou, S. A., & Hadjileontiadis, L. J. (2007). Wheeze detection based on time-frequency analysis of breath sounds. *Computers in biology and medicine, 37*(8), 1073–1083.
Theodoridis, S., & Koutroumbas, K. (2006). *Pattern Recognition* (3rd). Academic Press.
Toh, K. K. V., & Isa, N. A. M. (2010). Noise adaptive fuzzy switching median filter for salt-and-pepper noise reduction. *IEEE signal processing letters, 17*(3), 281–284.
Torre-Cruz, J., Canadas-Quesada, F., Carabias-Orti, J., Vera-Candeas, P., & Ruiz-Reyes, N. (2019). A novel wheezing detection approach based on constrained non-negative matrix factorization. *Applied Acoustics, 148*, 276–288.
Torre-Cruz, J., Canadas-Quesada, F., Vera-Candeas, P., Montiel-Zafra, V., & Ruiz-Reyes, N. (2018). Wheezing sound separation based on constrained non-negative matrix factorization. In *Proceedings of the 10th international conference on bioinformatics and biomedical technology (icbbt)* (pp. 18–24). ACM.
Ulukaya, S., Sen, I., & Kahya, Y. P. (2015). Feature extraction using time-frequency analysis for monophonic-polyphonic wheeze discrimination. In *37th annual international conference of the ieee engineering in medicine and biology society (embc)* (pp. 5412–5415). IEEE.
Wisniewski, M., & Zielinski, T. P. (2011). Application of tonal index to pulmonary wheezes detection in asthma monitoring. In *19th european signal processing conference* (pp. 1544–1548). IEEE.
Wisniewski, M., & Zielinski, T. P. (2012a). Fast and robust method for wheezes recognition in remote asthma monitoring. In *Information technologies in biomedicine* (pp. 568–576). Springer.
Wisniewski, M., & Zielinski, T. P. (2012b). Tonality detection methods for wheezes recognition system. In *19th international conference on systems, signals and image processing (iwssip)* (pp. 472–475). IEEE.
Wisniewski, M., & Zielinski, T. P. (2015). Joint application of audio spectral envelope and tonality index in an e-asthma monitoring system. *IEEE journal of biomedical and health informatics, 19*(3), 1009–1018.
Yoo, J.-H., & Choi, S.-J. (2010). Nonnegative matrix factorization with orthogonality constraints. *Journal of computing science and engineering, 4*(2), 97–109.
Zhang, J., Ser, W., Yu, J., & Zhang, T. (2009). A novel wheeze detection method for wearable monitoring systems. In *International symposium on intelligent ubiquitous computing and education* (pp. 331–334). IEEE.

Paper 6

Monophonic and polyphonic wheezing classification based on constrained low-rank Non-negative Matrix Factorization

J. De La Torre Cruz, F.J. Cañadas Quesada, N. Ruiz Reyes, S. García Galán, J.J. Carabias Orti and G. Peréz Chica, "Monophonic and polyphonic wheezing classification based on constrained low-rank Non-negative Matrix Factorization", in *Sensors*. Status: under review.

- Estado: En revisión.
- Revista: Sensors.
- ISSN: 1424-8220.
- Factor de impacto (JCR 2019): 3.275.
- Cuartiles por área de conocimiento:
  - Engineering, electrical and electronic: Q2, 77/266.
  - Instruments and instrumentation: Q1, 15/66.

MDPI

Article

Monophonic and polyphonic wheezing classification based on constrained low-rank Non-negative Matrix Factorization

Juan De La Torre Cruz [1,*], Francisco Jesús Cañadas Quesada [1], Nicolás Ruiz Reyes [1], Sebastián García Galán [1], Julio José Carabias Orti [1] and Gerardo Peréz Chica [2]

[1] Department of Telecommunication Engineering. University of Jaen, Campus Cientifico-Tecnologico de Linares, Avda. de la Universidad, s/n, 23700 Linares, Jaen, Spain; jtorre@ujaen.es (J.D.L.T.C.); fcanadas@ujaen.es (F.J.C.Q.); nicolas@ujaen.es (N.R.R.); sgalan@ujaen.es (S.G.G.); carabias@ujaen.es (J.J.C.O.)

[2] Pneumology Clinical Management Unit of the University Hospital of Jaen, Av. del Ejercito Espanol, 10, 23007 Jaen, Spain; gerardo.perez.sspa@juntadeandalucia.es (G.P.C.)

* Correspondence: jtorre@ujaen.es

Version January 25, 2021 submitted to Sensors

Abstract: The appearance of wheezing sounds is widely considered by physicians as a key indicator to detect early pulmonary disorders or even the severity associated with respiratory diseases, as occurs in the case of asthma and chronic obstructive pulmonary disease. From a physician's point of view, the monophonic and polyphonic wheezing classification is still a challenging topic in biomedical signal processing since both types of wheezes are sinusoidal in nature. Unlike most of the classification algorithms in which interference caused by normal respiratory sounds is not addressed in depth, our first contribution proposes a novel Constrained Low-Rank Non-negative Matrix Factorization (CL-RNMF) approach, never applied to classification wheezing as far as the authors knowledge extends, which incorporates several constraints (sparseness and smoothness) and a low-rank configuration to extract the wheezing spectral content minimizing the acoustic interference from normal respiratory sounds. The second contribution automatically analyzes the harmonic structure of the energy distribution associated with the estimated wheezing spectrogram to classify the type of wheezing. Experimental results report that: i) the proposed method outperforms the most recent and relevant state-of-the-art wheezing classification method by approximately 8% accuracy; ii) unlike state-of-the-art methods based on classifiers, the proposed method uses an unsupervised approach that does not require any training.

Keywords: monophonic; polyphonic; wheezing; non-negative matrix factorization; spectral pattern; spectrogram; constraint; low-rank; asthma; chronic obstructive pulmonary disease

1. Introduction

Chronic Respiratory Diseases (CDRs) are increasingly a huge and growing public health problem due to their high prevalence, high morbidity and mortality, and socio-economic cost. CRDs can be defined as disorders of the airways and other physiological structures of the respiratory system [1]. Some of the most common and relevant CRDs are asthma and Chronic Obstructive Pulmonary Disease (COPD). According to World Health Organization (WHO), there were 417,918 deaths due to asthma at the global level in 2016 [2] and approximately 3 million people die from COPD every year, which is 6% of all deaths worldwide [3]. Although chronic diseases currently have no medical cure, early detection can lead to appropriate treatment when the disease is in its early stages, thus improving people's quality of life.

The auscultation examination is considered a widely used method of detecting CDRs because it is a non-invasive, inexpensive, easy, comfortable and fast method regardless of age [4]. However, the auscultation process has several limitations that reduce the reliability of the diagnosis: i) high subjectivity conditioned by the physician's training to recognize and interpret the sounds captured by the stethoscope [5,6]; ii) the discrimination between adventitious sounds with similar characteristics, such as monophonic and polyphonic wheezing sounds, is a harder task to perform by means of auscultation [7]; and iii) normal respiratory sounds and adventitious sounds (abnormal and indicative of a lung disorder) are simultaneously mixed in the time and frequency domain, complicating the physician's analysis of the valuable clinical information contained in adventitious sounds [5,8,9]. Considering the above, a misdiagnosis is the main cause of the patient returning to the health center with a worsening of the disease that was not detected in the first medical examination performed by auscultation, so in recent years it has become crucial to develop novel approaches to help physicians provide reliable diagnoses applied to lung disorders, with the implicit fact of reducing health care costs [10,11].

In general, the sounds generated during breathing can be classified into two main categories: normal respiratory sounds and adventitious sounds, such as wheezing. Both sounds are mixed in the time-frequency domain as they are simultaneously generated by the same air flow through the bronchial tree of the lungs and also share part of the spectral bands in which they are active [5,9]. Specifically, normal respiratory sounds are represented by a wideband spectrum where most of the energy is concentrated in the frequency band 60Hz - 1000Hz [12]. However, the guidelines established by Computerized Respiratory Sound Analysis (CORSA) [5,13] define wheezing or wheeze sounds as a pitch located between 100Hz – 1000Hz whose duration is greater than 100ms. In fact, wheezes are considered Continuous Adventitious Sounds (CAS), as they can be represented using trajectories of narrowband spectral peaks over time. The appearance of wheezing is widely considered by doctors as a clue to be able to detect early either respiratory diseases, or the severity associated with CRDs, as occurs in the case of asthma and COPD [14,15]. For this reason, many research efforts have been applied in biomedical signal processing in order to develop reliable methods for the early wheezing detection. In this sense, many wheezing detection algorithms, based on different approaches, can be found in the state-of-the-art literature: Autoregressive (AR) model [16], Auditory modelling [17], Entropy [18], Neural networks (NN) [19,20], Wavelet transform [21,22], Tonal index [23,24], Mel-frequency cepstral coefficients (MFCC) [25,26], Gaussian Mixture Models (GMM) [27,28], Spectral peaks identification [29–31], Hidden Markov model (HMM) [32] and recently, Non-negative Matrix Factorization (NMF) [9,33,34].

Wheezing can be classified into two main categories according to the spectral behaviour [35]: (i) wheezes that occur with a single peak or with the harmonics associated to that single basal peak are called Monophonic (MP) wheezes (as can be seen in Figure 1); and (ii) wheezes that occur with variable peaks that differ in harmonics are called Polyphonic (PP) wheezes (as can be seen in Figure 2). The scientific interest in the field of biomedical sound signal processing in automatically performing this classification lies in the fact that PP wheezes are usually caused by the pathology of small airways and MP wheezes are caused by the pathology of larger airways [36]. In fact, several studies [4,37–39] have shown that MP and PP wheezes exhibit distinctive physiological and pathological characteristics: i) in physiological analysis, MP wheezes are caused by a single bronchial narrowing while PP wheezes are caused by multiple central bronchial compression; and ii) in pathological analysis, MP wheezes are an indicator of the presence of asthma while PP wheezes can be considered as a sound marker of COPD.

Despite advances in the analysis of respiratory sounds, MP/PP wheezing classification is a critical step in the diagnosis of asthma [4,37,38] and COPD diseases [37–39] so, it is still a challenging topic in biomedical signal processing [7] since both types of wheezes are sinusoidal in nature. Although there are relatively few works [7,40–45] in which the analysis of MP/PP wheezing is treated, the only works focused on the task of classifying MP/PP wheezing in depth are [7,40,44,45] to our knowledge. All these MP/PP wheezing classification approaches are based on the feature extraction and classifier

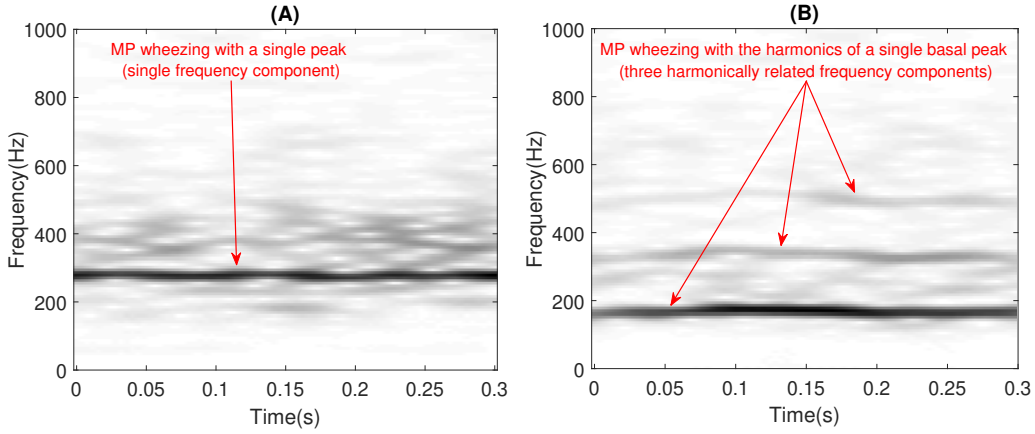

Figure 1. Time-frequency representation of two examples of MP wheezing: (A) with a single basal peak. (B) with the harmonics of a single basal peak. Note that the frequency components are harmonically related in (B).

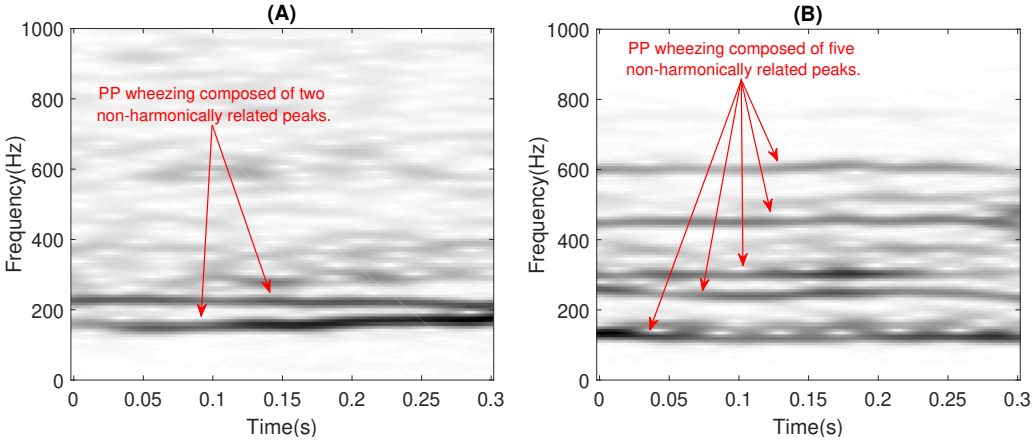

Figure 2. Time-frequency representation of two examples of PP wheezing. Note that the frequency components are not harmonically related in the case of PP wheezing.

configuration. Ulukaya et al. [7] proposed to extract a single feature, Peak Energy Ratio (PER), from a RAtional Dilation Wavelet Transform (RADWT) to discriminate between MP and PP wheezes. Specifically, PER is obtained from the first and second peak with the highest energy of all subbands of the wavelet coefficients (considering that the second peak is not consecutive to the first one). Moreover, the authors applied a robust evaluation methodology in which most of the relevant feature extraction methods [40,44,45] were evaluated using some of the most popular classifiers (SVM, KNN and ELM) and leave-one-out (LOO) cross validation schemes. The results reported that the proposed method, based on only one feature (PER), obtained the best MP/PP wheezing classification performance showing an accuracy equals to 86%.

However, none of the state-of-the-art methods consider the interference generated by normal respiratory sounds that can affect the MP/PP wheezing classification task. In this work, our proposal is based on Non-negative Matrix Factorization (NMF) approach in order to classify MP/PP wheezing sounds according to the harmonic structure shown by removing the sound interference caused by normal respiratory sounds. The first contribution of this work proposes a novel Constrained Low-Rank Non-negative Matrix Factorization (CL-RNMF) approach which allows the spectral patterns associated with wheezing sounds to be extracted with the least possible sound interference from normal breath sounds. Specifically, we propose a low-rank configuration using a reduced number of wheeze bases to compact the frequency components into the fewest possible bases for further analysis without loss of relevant wheeze content. In addition, the proposed CL-RNMF approach incorporates a set of

constraints to model the spectro-temporal behavior of wheezing and normal respiratory sounds. These constraints helps to acoustically isolate the wheezing spectral patterns from normal respiratory sounds. To classify between MP or PP wheezing sounds, the second contribution analyzes the harmonic structure of the previous reduced number of wheezing bases based on the spectral location of the wheezing components, rather than the energy of their components.

The structure of the paper is described. Section 2 briefly reviews the principles of Non-negative Matrix Factorization, focusing on the standard approach and some regularizations used to model the properties of the interest sounds. The proposed MP/PP wheezing classification method is presented in Section 3. Section 4 details and discusses the experimental evaluation. Finally, we conclude in Section 5 and provide perspectives on further research.

2. Theoretical Background

2.1. Non-negative Matrix Factorization

Non-negative Matrix Factorization (NMF) or standard NMF [46,47] is a decomposition technique that has attracted special attention in different fields of biomedical signal processing in the last years [48,49]. Previous works show the efficiency of the NMF approach to detecting [9,33,34] and improving the audio quality of wheezing [50,51]. In general terms, NMF can be defined as an unsupervised learning tool used for linear representation of non-negative two-dimensional (2D) data where its main advantage is to reduce the dimensionality of a large amount of data in order to find hidden structures by means of part-based representation with non-negative patterns. From a mixture signal $x(t)$, its magnitude spectrogram $\mathbf{X} \in \mathbb{R}_+^{F\times T}$ is obtained by means of the Short-Time Fourier Transform (STFT) applying a window function (e.g. Hamming or Hann) and inter-window overlap to increase the temporal resolution, being F the number of frequency bins and T the number of time frames. Here, standard NMF decomposes the magnitude spectrogram $\mathbf{X}$ into the product of two non-negative matrices: spectral basis matrix (patterns) $\mathbf{B} \in \mathbb{R}_+^{F\times K}$ and temporal activation matrix (weights) $\mathbf{A} \in \mathbb{R}_+^{K\times T}$, being K the rank or the number of components (spectral bases),

$$\mathbf{X} \approx \hat{\mathbf{X}} = \mathbf{BA} \tag{1}$$

where $\hat{\mathbf{X}} \in \mathbb{R}_+^{F\times T}$ is the estimated spectrogram. Each column of the basis matrix $\mathbf{B}$ defines a spectral pattern that describes the spectral behaviour of an active sound event in the input spectrogram $\mathbf{X}$. Each row of the activation matrix $\mathbf{A}$ represents a temporal gain for a spectral pattern. In other words, the matrix $\mathbf{B}$ provides a dictionary composed of K spectral bases and the matrix A defines the weight with which the different spectral bases appear along the temporal frames. Due to the nonnegativity property, NMF underlies an additive linear interpolation model which results in the so-called parts-based representation [46].

The decomposition or factorization of the input magnitude spectrogram $\mathbf{X}$ into the product $\mathbf{BA}$ is usually sought minimizing a defined scalar-valued divergence,

$$\arg\min_{\mathbf{B},\mathbf{A}} D(\mathbf{X}|\mathbf{BA}) \quad \mathbf{B},\mathbf{A} \geq 0 \tag{2}$$

This divergence function measures the error made in the approximation of the observed spectrogram $\mathbf{X}$ and the reconstruction $\mathbf{BA}$. Typically, the divergence is computed entry-wise, i.e.

$$D(\mathbf{X}|\hat{\mathbf{X}}) = D(\mathbf{X}|\mathbf{BA}) = \sum_{f=1}^{F}\sum_{t=1}^{T} d(X_{f,t}|\hat{X}_{f,t}) \tag{3}$$

where $d(i,j)$ is a function of two scalar variables i,j. It is often called cost function and is a positive function of $i \in \mathbb{R}_+$ given $j \in \mathbb{R}_+$ with a single minimum for $i = j$. Some of the most

popular cost functions are the Euclidean distance, the generalized Kullback-Leibler divergence, the Itakura-Saito divergence and the Cauchy distribution [52,53]. In this paper, we propose to minimize the generalized Kullback-Liebler divergence $D_{KL}(\mathbf{X}|\hat{\mathbf{X}})$ (see Equation (4)) because previous works [9,33,34,48,50,51] have obtained promising results in biomedical signal processing since $D_{KL}(\mathbf{X}|\hat{\mathbf{X}})$ provides a scale-invariant factorization, that is, low energy sound components of $\mathbf{X}$ bear the same relative importance as high energy ones into the decomposition process.

$$D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}\right) = \sum_{f=1}^{F}\sum_{t=1}^{T} X_{f,t}\log\frac{X_{f,t}}{\hat{X}_{f,t}} - X_{f,t} + \hat{X}_{f,t} \tag{4}$$

The most popular minimization method to solve the problem shown in Equation (2) is based on the so called multiplicative update rules, initially proposed by Lee and Seung [46]. This method obtains the basis and activation matrices, minimising the Kullback-Liebler divergence function $D_{KL}(\mathbf{X}|\hat{\mathbf{X}})$ and ensuring the non-negativity of the estimated matrices. These rules are obtained directly from the negative and positive terms of the partial derivative of the divergence function $D_{KL}(\mathbf{X}|\hat{\mathbf{X}})$ with respect to the parameters $\mathbf{B}$ and $\mathbf{A}$,

$$\mathbf{B} \leftarrow \mathbf{B} \odot \frac{\left[\frac{\partial D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}\right)}{\partial \mathbf{B}}\right]^{-}}{\left[\frac{\partial D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}\right)}{\partial \mathbf{B}}\right]^{+}} = \mathbf{B} \odot \left(\left(\mathbf{X} \oslash \mathbf{B}\mathbf{A}\right)\mathbf{A}^{\mathrm{T}} \oslash \left([\mathbf{1}]\mathbf{A}^{\mathrm{T}}\right)\right) \tag{5}$$

$$\mathbf{A} \leftarrow \mathbf{A} \odot \frac{\left[\frac{\partial D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}\right)}{\partial \mathbf{A}}\right]^{-}}{\left[\frac{\partial D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}\right)}{\partial \mathbf{A}}\right]^{+}} = \mathbf{A} \odot \left(\mathbf{B}^{\mathrm{T}}\left(\mathbf{X} \oslash \mathbf{B}\mathbf{A}\right) \oslash \left(\mathbf{B}^{\mathrm{T}}[\mathbf{1}]\right)\right) \tag{6}$$

where $[\mathbf{1}] \in \mathbb{R}_{+}^{F\times T}$ represents an all-ones matrix, $^{\mathrm{T}}$ is the transpose operator, $\odot$ is the element-wise multiplication and $\oslash$ is the element-wise division. This procedure always maintains the non-negativity of both parameters, since the used terms in the updating are also non-negative.

As previously described, NMF models the magnitude spectrogram of an input mixture signal as a product of a basis matrix and a activations matrix with the only constraint of the element-wise non-negativity of all matrices. Under this constraint, the aim is to minimize the cost function of the reconstruction error. However, the main problem of the NMF is the trade-off between signal reconstruction and physical interpretation of the factorized parts-based objects. In other words, this nonnegativity of the parameters does not guarantee a meaningful part-based representation when dealing with real-world mixture signals [54,55]. Several properties can be used to improve the uniqueness of the local minima obtained by NMF, incorporating physical meaning to the basis functions and activations. In particular, these properties can be implemented using regularizations which are added to the global cost function in the factorization model. The main constraints, sparseness and smoothness, used in this paper to model the spectro-temporal behaviour of wheezing and normal respiratory sounds are briefly described below.

2.2. Spectral Sparseness

Spectral Sparseness $\psi(\mathbf{B})$ denotes that, for each source, most of its frequencies are zero or close to zero [56,57]. This constraint enforces that only a few frequencies bins predominate in each spectral basis, whilst the other bins are cancelled. It is implemented by incorporating a penalty term to the NMF objective function. In practice, the L^1-norm is often used because it was demonstrated to be less

sensitive to changes of the parameter that controls the importance of the constraint in the factorization process. Then, the optimization problem can be expressed as,

$$\arg\min_{\mathbf{B},\mathbf{A}} D(\mathbf{X}|\mathbf{BA}) + \alpha \left\|\mathbf{B}\right\|_1 \quad \mathbf{B},\mathbf{A} \geq 0 \tag{7}$$

where α is the weight parameter that adjust the influence of the constraint.

2.3. Temporal/Spectral Smoothness

Generally, smoothness ϕ means how continuous or smooth are the spectral or temporal changes related to a source [57]. Smoothness constraints have been defined for both activation $\mathbf{A}$ and basis functions $\mathbf{B}$ and added to the global cost function as penalty terms as follows,

$$\arg\min_{\mathbf{B},\mathbf{A}} D(\mathbf{X}|\mathbf{BA}) + \lambda\phi(\mathbf{A}) + \beta\phi(\mathbf{B}) \quad \mathbf{B},\mathbf{A} \geq 0 \tag{8}$$

where $\phi(\mathbf{A})$ and $\phi(\mathbf{B})$ are the functions that penalize non-smooth temporal activations or spectral patterns and the parameters λ and β control the effect of the regularization into the decomposition procedure.

Temporal smoothness (a.k.a. smooth activations) $\phi(\mathbf{A})$, applied to the estimated activation matrix $\mathbf{A}$, reports how slow the amplitude variations over time are. In other words, temporal smoothness accounts for the fact that real-world sounds usually have a temporal structure, and their acoustic characteristics vary slowly as a function of time. In [57], the authors proposed to model the temporal smoothness regularization $\phi(\mathbf{A})$ by applying a high cost to large changes produced between adjacent frames in the activation matrix $\mathbf{A}$ as follows,

$$\phi(\mathbf{A}) = \sum_{k=1}^{K} \frac{1}{\sigma_k^2} \sum_{t=2}^{T} \left(A_{k,t} - A_{k,t-1}\right)^2 \tag{9}$$

where $\sigma_k = \sqrt{\frac{1}{T}\sum_{t=1}^{T} A_{k,t}^2}$ indicates the standard deviation used to normalize the activation functions. This normalization provides that the cost of regularization is independent of the numerical scale of activation [54,57].

Spectral smoothness (a.k.a. smooth basis) $\phi(\mathbf{B})$, applied to the estimated basis matrix $\mathbf{B}$, measures how fast the amplitude changes along the frequency axis, that is, it allows to model the behaviour of those sounds that are represented by wideband spectrum. In [54,58], the authors proposed to model the spectral smoothness regularization $\phi(\mathbf{B})$ by applying a high cost to large changes produced between adjacent bins in the basis matrix $\mathbf{B}$ as follows,

$$\phi(\mathbf{B}) = \sum_{k=1}^{K} \frac{1}{\sigma_k^2} \sum_{f=2}^{F} \left(B_{f,k} - B_{f-1,k}\right)^2 \tag{10}$$

where $\sigma_k = \sqrt{\frac{1}{F}\sum_{f=1}^{F} B_{f,k}^2}$ represents the standard deviation used to normalize the basis functions. This normalization achieves that the cost of regularization is independent of the numerical scale of basis [54,58].

3. Proposed method

The main problem in classifying wheezes from a mixture is that both wheezing sounds and normal respiratory sounds occur simultaneously in the time and frequency domain. Considering the acoustic interference caused by normal respiratory sounds, the proposed signal model is composed of two stages: Modelling and separation of wheezing spectral patterns from normal respiratory sounds based on CL-RNMF (stage I) and Classification between MP/PP wheezing according to its harmonic structure (stage II). In this manner, the goal of the stage I is to model the spectral patterns

that characterize wheezing sounds by isolating them from respiratory interference. The aim of stage II is to analyze the location of the frequency components extracted from the previous stage to determine the type, monophonic or polyphonic, of wheezing according how the wheezing energy is locating in the frequency domain. The flowchart of the proposed method is shown in Figure 3, and details are depicted in Sections 3.1, 3.2 and 3.3.

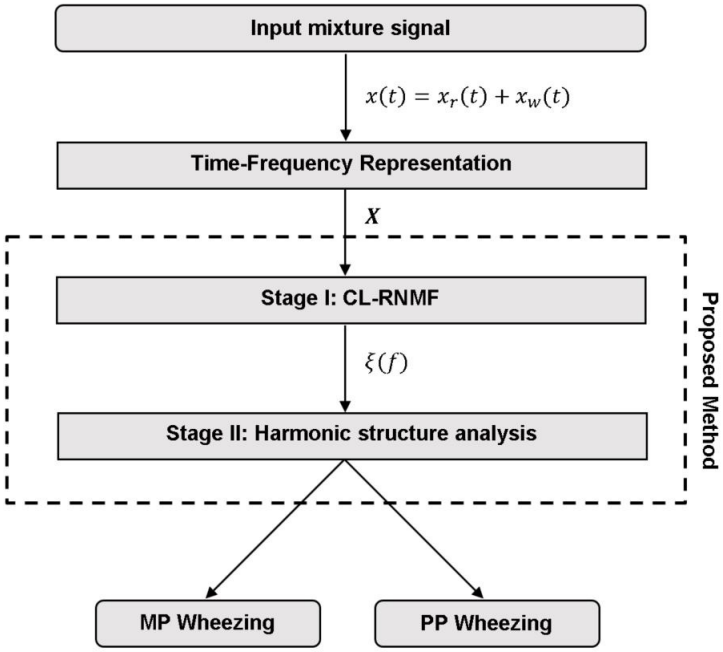

Figure 3. Flowchart of the proposed method.

3.1. Time-Frequency Signal Representation

Time-frequency representation by means of spectrograms has been demonstrated to be useful for visualizing the characteristics and behavior of both wheezing and normal respiratory sounds [9,33,34,50,51], . The input mixture signal $x(t)$ is composed of wheeze sounds $x_w(t)$ (MP or PP wheezing) and normal respiratory sounds $x_r(t)$ overlapping in the time and frequency domain. We assume that the mixture of these sounds is additive and can be expressed as $x(t) = x_r(t) + x_w(t)$. The input magnitude spectrogram $\mathbf{X} \in \mathbb{R}_+^{F \times T}$ of the input mixture signal can be represented as $\mathbf{X} = \mathbf{X}_R + \mathbf{X}_W$, being $\mathbf{X}_R \in \mathbb{R}_+^{F \times T}$ the magnitude spectrogram of only respiratory sounds and $\mathbf{X}_W \in \mathbb{R}_+^{F \times T}$ the magnitude spectrogram of only wheeze sounds. Specifically, each magnitude spectrogram is composed of T frames, F frequency bins and a set of time-frequency units $X_{f,t}$, being $f = 1, ..., F$ and $t = 1, ..., T$. Each unit $X_{f,t}$ is defined by the f^{th} frequency bin at the t^{th} frame and is calculated from the magnitude of the Short-Time Fourier Transform (STFT) using a Hamming windows of N samples with 10% overlap. In this work, a normalization process is applied in order to achieve independence regarding the size and scale of the input spectrogram $\mathbf{X}$. Thus, the normalized magnitude spectrogram $\overline{\mathbf{X}}$ is computed as follows,

$$\overline{\mathbf{X}} = \frac{\mathbf{X}}{\left(\frac{\sum_{f=1}^{F} \sum_{t=1}^{T} X_{f,t}}{FT} \right)} \tag{11}$$

To avoid complex nomenclature throughout the paper, the variable $\mathbf{X}$ is hereinafter referred to the normalized magnitude spectrogram previously computed in Equation (11).

3.2. Stage I: Constrained Low-Rank Non-negative Matrix Factorization (CL-RNMF)

As mentioned above, it is common that normal respiratory sounds mask the presence of the wheezing sounds. As a result, this sound mask makes the task of wheezing classification difficult since the spectral patterns associated to normal respiratory sounds can be confused with wheezing spectral content. Therefore, the aim of this stage is to provide a reliable modeling of the different frequency components (spectral patterns) that compose a wheeze, removing any sound interference from normal respiratory sounds. For this purpose, we propose a constrained low-rank non-negative matrix factorization (CL-RNMF) approach because as far as the authors knowledge extends, non-negative matrix factorization approach has never been applied before to MP/PP wheezing classification. In addition, our approach is an unsupervised method because does not require any training of the sounds to classify. Therefore, the proposed method decomposes a magnitude mixture spectrogram $\mathbf{X}$ into two estimated spectrograms: $\hat{\mathbf{X}}_R$ (only normal respiratory sounds without wheezing) and $\hat{\mathbf{X}}_W$ (only wheeze sounds without normal respiratory sounds). In this manner, each estimated spectrogram can be factorized into the product of its corresponding estimated basis and activation matrices: i) $\mathbf{B}_R \in \mathbb{R}_+^{F\times K_r}$ and $\mathbf{A}_R \in \mathbb{R}_+^{K_r\times T}$ to the factorization of $\hat{\mathbf{X}}_R$, being K_r the number of respiratory components; and ii) $\mathbf{B}_W \in \mathbb{R}_+^{F\times K_w}$ and $\mathbf{A}_W \in \mathbb{R}_+^{K_w\times T}$ to the factorization of $\hat{\mathbf{X}}_W$, being K_w the number of wheezing components. The proposed separation model can be formulated with the following objective function,

$$\mathbf{X} \approx \hat{\mathbf{X}}_R + \hat{\mathbf{X}}_W = \mathbf{B}_R\mathbf{A}_R + \mathbf{B}_W\mathbf{A}_W \tag{12}$$

where, considering the non-negative property that characterises the NMF approach, all the matrices that compose the previous model are non-negative.

As previously mentioned, this stage attempts to ensure that $\mathbf{B}_W$ contains reliable modeling of the wheezing spectral patterns by means of narrowband spectral peaks that typically characterize the wheeze content. The key assumptions behind the proposed CL-RNMF approach to model wheezing spectral patterns are the following:

- **Low-Rank**: the number of wheezing components should be much less than the number of normal respiratory components, that is, $K_w \ll K_r$. This assumption allows that the number of frequency components can be reduced in the least number of bases possible for their posterior analysis, while normal respiratory sounds are modeled using a higher range of components. Experimental results showed that the best classification performance was obtained when $2 \leq K_w \leq 6$ and $K_r \geq 32$. In particular, when $K_w = 1$, the proposed CL-RNMF approach tends to converge very quickly at the expense of losing relevant wheezing content. On the other hand, when $K_w > 6$, the spectral wheezing patterns tend to be splitted into different components of the matrix $\mathbf{B}_W$.
- **Constraints**: characterize wheezing sounds and normal respiratory sounds using opposite restrictions between both sounds. The use of constraints allows to isolate the spectral wheezing patterns from the spectral patterns of normal respiratory sounds. Therefore, in order to find a better NMF decomposition that shows spectro-temporal features of the wheezing and normal respiratory sounds as can be observed in real-world, we propose to incorporate sparseness and smoothness into the NMF decomposition process. As shown in Figure 1 and 2, wheezing sounds can be considered sparse in frequency because MP wheezing or PP wheezing is characterized by one or more than one narrowband spectral peaks. Moreover, wheezing sounds can be considered smooth or continuous events in time, that is, slow variation of the magnitude spectrogram along time. On the other hand, normal respiratory sounds can be considered smooth in frequency, that is, they can be modeled assuming wideband spectral patterns. Therefore, $\mathbf{B}_W$ should contain wheezing spectral patterns composed of one or more than one narrowband spectral peaks, depending on the spectral complexity of each wheezing, and $\mathbf{B}_R$ should be composed of a set of wideband spectral patterns that model the behavior of normal respiratory sounds.

Considering the key assumptions mentioned above, the global objective function $D\left(\mathbf{X}|\hat{\mathbf{X}}\right)$ that must be minimized in order to estimate the basis ($\mathbf{B}_R$, $\mathbf{B}_W$) and activation ($\mathbf{A}_R$, $\mathbf{A}_W$) matrices

is composed of: (i) the Kullback-Leibler divergence cost function $D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}\right)$ to minimize the reconstruction error between the input spectrogram $\mathbf{X}$ and estimated spectrogram $\hat{\mathbf{X}}$, (ii) the spectral sparseness $\psi\left(\mathbf{B}_W\right)$ and temporal smoothness $\phi\left(\mathbf{A}_W\right)$ restrictions applied to $\mathbf{B}_W$ and $\mathbf{A}_W$, respectively, to model the wheezing spectral patterns, and (iii) the spectral smoothness $\phi\left(\mathbf{B}_R\right)$ restriction applied to $\mathbf{B}_R$, to model the spectral patterns of normal respiratory sounds. The global objective function $D\left(\mathbf{X}|\hat{\mathbf{X}}\right)$ is detailed as follows,

$$D\left(\mathbf{X}|\hat{\mathbf{X}}\right) = D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}\right) + \alpha\psi\left(\mathbf{B}_W\right) + \lambda\phi\left(\mathbf{A}_W\right) + \beta\phi\left(\mathbf{B}_R\right) \tag{13}$$

where the equations of terms $D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}\right)$, $\psi\left(\mathbf{B}_W\right)$, $\phi\left(\mathbf{A}_W\right)$ and $\phi\left(\mathbf{B}_R\right)$ can be found in Section 2. The parameters α, λ and β define the weight to control the effect of the regularization. Experimental results showed that the best classification performance is obtained when all weights are equal $\alpha = \lambda = \beta$, being the optimal value $\alpha = \lambda = \beta = 0.5$. Analyzing the sound separation performance of the previous decomposition, we have observed empirically that the acoustic interference suffered by wheezing sounds from normal respiratory sounds is minimum and no significant loss of wheezing content occurs when $\alpha = \lambda = \beta$. However, significant losses of wheezing content appear when $\alpha = \lambda > \beta$ or significant sound interference by normal respiratory sounds can be observed when $\alpha = \lambda < \beta$.

From Equation (13), the estimated basis matrices ($\mathbf{B}_W$ and $\mathbf{B}_R$) and activation matrices ($\mathbf{A}_W$ and $\mathbf{A}_R$) can be obtained by applying a gradient descent algorithm based on multiplicative update rules. Specifically, the multiplicative update rules to learn those matrices can be computed by taking negative and positive terms of the partial derivative of the global objective function $D\left(\mathbf{X}|\hat{\mathbf{X}}\right)$ with respect to $\mathbf{B}_W$, $\mathbf{B}_R$, $\mathbf{A}_W$ and $\mathbf{A}_R$, respectively,

$$\mathbf{B}_W \leftarrow \mathbf{B}_W \odot \frac{\left[\frac{\partial D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}\right)}{\partial \mathbf{B}_W}\right]^- + \alpha\left[\frac{\partial \psi(\mathbf{B}_W)}{\partial \mathbf{B}_W}\right]^-}{\left[\frac{\partial D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}\right)}{\partial \mathbf{B}_W}\right]^+ + \alpha\left[\frac{\partial \psi(\mathbf{B}_W)}{\partial \mathbf{B}_W}\right]^+} \tag{14}$$

$$\mathbf{B}_R \leftarrow \mathbf{B}_R \odot \frac{\left[\frac{\partial D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}\right)}{\partial \mathbf{B}_R}\right]^- + \beta\left[\frac{\partial \phi(\mathbf{B}_R)}{\partial \mathbf{B}_R}\right]^-}{\left[\frac{\partial D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}\right)}{\partial \mathbf{B}_R}\right]^+ + \beta\left[\frac{\partial \phi(\mathbf{B}_R)}{\partial \mathbf{B}_R}\right]^+} \tag{15}$$

$$\mathbf{A}_W \leftarrow \mathbf{A}_W \odot \frac{\left[\frac{\partial D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}\right)}{\partial \mathbf{A}_W}\right]^- + \lambda\left[\frac{\partial \phi(\mathbf{A}_W)}{\partial \mathbf{A}_W}\right]^-}{\left[\frac{\partial D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}\right)}{\partial \mathbf{A}_W}\right]^+ + \lambda\left[\frac{\partial \phi(\mathbf{A}_W)}{\partial \mathbf{A}_W}\right]^+} \tag{16}$$

$$\mathbf{A}_R \leftarrow \mathbf{A}_R \odot \frac{\left[\frac{\partial D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}\right)}{\partial \mathbf{A}_R}\right]^-}{\left[\frac{\partial D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}\right)}{\partial \mathbf{A}_R}\right]^+} \tag{17}$$

where, for each multiplicative update rule, the division between the negative and positive terms of the partial derivatives is a element-wise division. More details related to the equations of each partial derivative of the multiplication update rules can be found in Appendix A. Finally, the estimated respiratory and wheezing basis ($\mathbf{B}_W$ and $\mathbf{B}_R$) and activation matrices ($\mathbf{A}_W$ and $\mathbf{A}_R$) are obtained updating the previous rules until the algorithm converges using M iterations. Figure 4 shows the estimated matrices $\mathbf{B}_W$ and $\mathbf{B}_R$ decomposing the MP wheezing spectrogram shown in Figure 1B. As can be observed, the matrix $\mathbf{B}_W$ contains spectral patterns that characterize a typical MP wheezing which are represented by means of a set of narrowband spectral peaks (or frequency components). In contrast, the estimated matrix $\mathbf{B}_R$ is composed of a set of wideband spectral patterns that characterize

normal respiratory sounds. Therefore, the proposed CL-RNMF approach achieves to extract the wheezing spectral content at the expense of removing normal respiratory sounds.

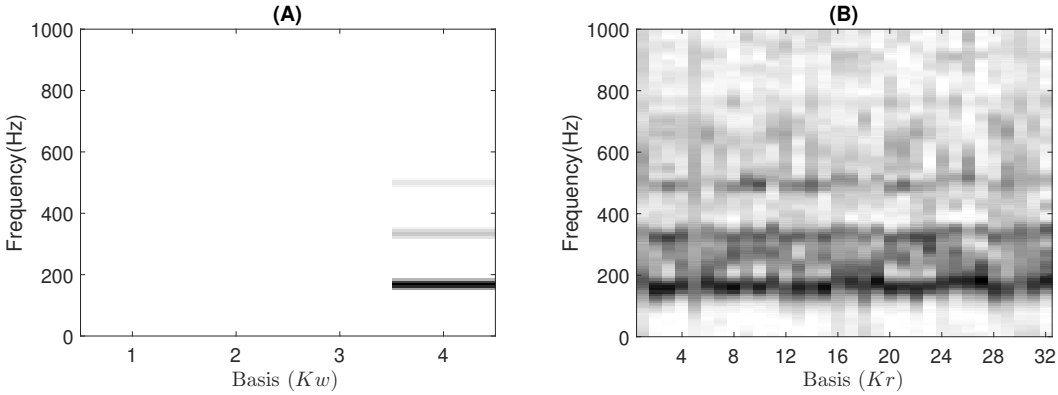

Figure 4. Example of the estimated matrices $\mathbf{B}_W$ and $\mathbf{B}_R$ obtained from the proposed CL-RNMF approach, analyzing the MP wheezing spectrogram previously shown in Figure 1B. (A) although the matrix $\mathbf{B}_W$ is composed of four spectral bases, the spectral wheezing patterns have been compacted into the fourth basis $\mathbf{B}_W(4)$. This spectral basis $\mathbf{B}_W(4)$ is composed of three narrowband spectral peaks; (B) the matrix $\mathbf{B}_R$ is composed of thirty-two wideband spectral bases.

Experimentally, we have found that the proposed CL-RNMF approach tends to compact all the narrowband spectral peaks into a single basis of the matrix $\mathbf{B}_W$, as shown in Figure 4A. However, considering that CL-RNMF uses a small set of wheezing components (K_w), in some cases the narrowband spectral peaks are divided into several bases of the same matrix $\mathbf{B}_W$. To clarify this issue, Figure 5 shows the matrix $\mathbf{B}_W$ obtained for the different examples of MP and PP wheezing described in Section 1. As shown in Figure 5D, the energy of the narrowband spectral patterns, that characterizes that PP wheezing, are divided into two bases $\mathbf{B}_W(1)$ and $\mathbf{B}_W(2)$ belonging to the matrix $\mathbf{B}_W$. In both bases, $\mathbf{B}_W(1)$ and $\mathbf{B}_W(2)$, all narrowband spectral peaks have been correctly modeled.

Finally, we propose to obtain the spectral energy distribution $\xi(f)$ (see Equation (18)) from the set of bases that compose the matrix $\mathbf{B}_W$. It makes it possible to compact the spectral distribution of all narrowband spectral peaks that make up the input MP or PP wheezes to analyze their harmonic structure in the stage II.

$$\xi(f) = \sum_{k_w=1}^{K_w} \mathbf{B}_{W_{f,k_w}} \quad , f = 1...F \tag{18}$$

Figure 6 shows the spectral energy distribution $\xi(f)$ obtained for the four examples of wheezing shown in Section 1. The pseudo code of this stage I for the modelling and separation of wheezing spectral patterns based on CL-RNMF is detailed in the Algorithm 1.

Algorithm 1 CL-RNMF

Require: $x(t)$, K_r, K_w, α, β, λ and M.

1: Compute the normalized magnitude spectrogram $\mathbf{X}$ using Equation (11).
2: Initialize $\mathbf{B}_W$, $\mathbf{B}_R$, $\mathbf{A}_W$ and $\mathbf{A}_R$ with random nonnegative values.
3: Update the estimated wheezing basis matrix $\mathbf{B}_W$ using Equation (14).
4: Update the estimated respiratory basis matrix $\mathbf{B}_R$ using Equation (15).
5: Update the estimated wheezing activations matrix $\mathbf{A}_W$ using Equation (16).
6: Update the estimated respiratory activations matrix $\mathbf{A}_R$ using Equation (17).
7: Repeat steps 3-6 until the algorithm converges (or until the maximum number of iterations M is reached).
8: Compute the spectral energy distribution $\xi(f)$ from $\mathbf{B}_W$ using Equation (18).

return $\xi(f)$

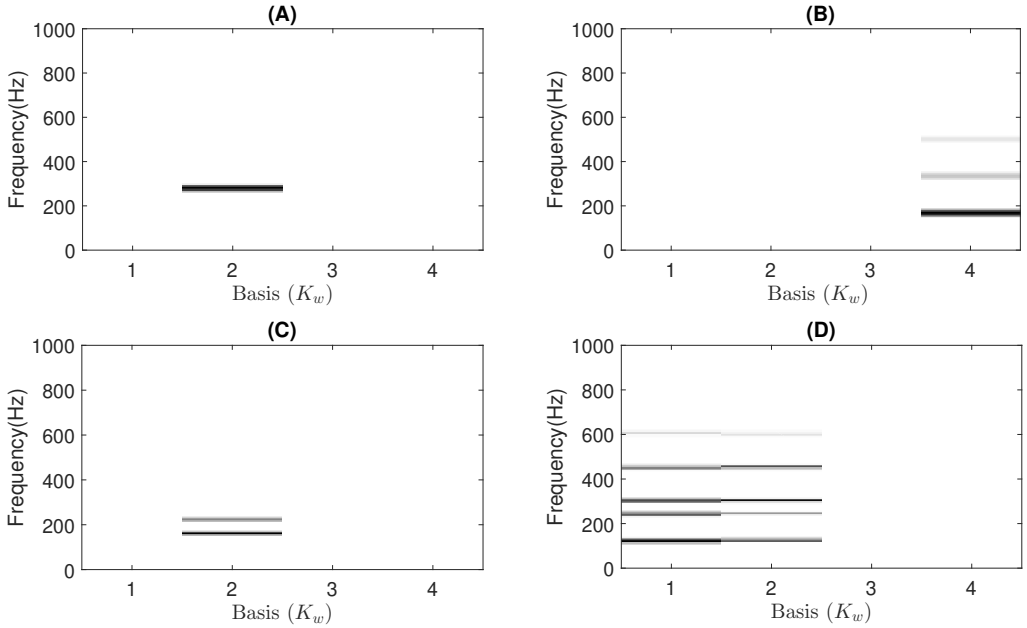

Figure 5. The estimated basis matrices $\mathbf{B}_W$ obtained from CL-RNMF in the examples shown in Section 1. (A) $\mathbf{B}_W$ for the MP wheezing shown in Figure 1A. (B) $\mathbf{B}_W$ for the MP wheezing shown in Figure 1B. (C) $\mathbf{B}_W$ for the PP wheezing shown in Figure 2A. (D) $\mathbf{B}_W$ for the PP wheezing shown in Figure 2B. The wheezing spectral patterns have been compacted into a single basis, $\mathbf{B}_W(2)$ (in cases (A)), $\mathbf{B}_W(4)$ (in cases (B)) and $\mathbf{B}_W(2)$ (in cases (C)). However, the energy of the narrowband spectral peaks has been divided into two bases $\mathbf{B}_W(1)$ and $\mathbf{B}_W(2)$ as can be seen in case (D).

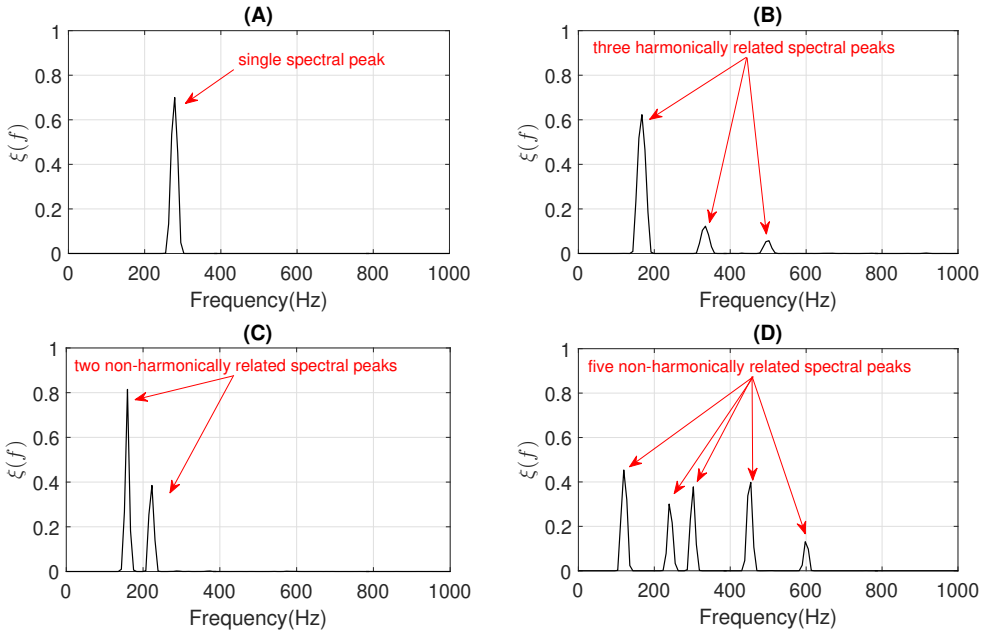

Figure 6. The spectral energy distribution $\xi(f)$ provided by CL-RNMF from the estimated basis matrix $\mathbf{B}_W$ shown in Figure 5: (A) Figure 5A. (B) Figure 5B. (C) Figure 5C. (D) Figure 5D.

3.3. Stage II: Harmonic structure analysis

The goal of this stage is to classify between MP and PP wheezing analyzing the spectral energy distribution $\xi(f)$ of the different narrowband spectral peaks obtained in the previous stage. Depending on the harmonic structure, wheezing can be classified as MP or PP. Specifically, MP wheezing is composed of a single narrowband spectral peak or with the set of harmonically related narrowband

spectral peaks. In contrast, PP wheezing is composed of several non-harmonically related narrowband spectral peaks. For this reason, we propose to obtain the number of narrowband spectral peaks η, that can be found from $\xi(f)$. Note that the procedure to detect the spectral peaks is a simple task since, as can be seen in Figure 6, the spectral energy distribution $\xi(f)$ from CL-RNMF clearly provides a set of narrowband spectral peaks typically found in wheezing sounds. Once the parameter η is obtained, a preliminary classification of the type of wheezing can be performed as follows,

$$\text{Wheezing category} = \begin{cases} \text{MP} & \text{if} \quad \eta = 1 \\ \text{MP or PP} & \text{if} \quad \eta > 1 \end{cases} \tag{19}$$

Wheezing can only be classified as MP when $\eta = 1$ since a wheezing is composed of a single narrowband spectral peak, as can be seen in Figure 6A. However, a wheezing can be classified as MP or PP when $\eta > 1$, depending on the harmonic structure that exists between the different narrowband spectral peaks. Specifically, the wheezing is classified as MP if the set of spectral peaks are harmonically related between them. The wheezing is classified as PP if the spectral peaks are not harmonically related between them. In order to perform the classification between MP and PP wheezing in the case of $\eta > 1$, we propose a two-step procedure, as follows:

- The objective of the first step is to locate, in terms of frequency, all the narrowband spectral peaks detected in the previous stage I. For this, we propose to locate the most prominent frequency $f_p(z)$ in each spectral peak $z = 1, ..., \eta$. Each value $f_p(z)$ has been calculated using the *findpeaks* function provided by MATLAB software [59] due to the satisfactory results obtained in several preliminary analyzes performed. Figure 7 shows the location $f_p(z)$, in terms of frequency, of each spectral peak for the MP example previously shown in Figure 1B.
- The objective of the second step is to check if the different spectral peaks $z = 1, ..., \eta$ are harmonically related or not. We assume that the first spectral peak ($z = 1$) represents the basal peak. So, the wheezing is classified as MP if the rest of spectral peaks ($z = 2, ..., \eta$) are located in the harmonic frequencies (integer multiple) of the basal peak. Otherwise, the wheezing is classified as PP. From the width Δ of the mainlobe of the basal peak ($z = 1$) and the value of its most prominent frequency $f_p(1)$, the spectral intervals where the possible harmonic frequencies should be located are calculated as follows,

$$\Lambda_z = \left[z f_p(1) - (\Delta/2), z f_p(1) + (\Delta/2) \right] \quad , z = 1...\eta \tag{20}$$

  where $[i, j]$ denotes the spectral interval comprised between the lower limit i and the upper limit j, in terms of frequency. Specifically, Λ_1 represents the spectral interval associated to the basal peak and Λ_z ($z = 2, ..., \eta$) corresponds to the spectral intervals where the harmonic frequencies should be located. Note that the width of the mainlobe Δ has been obtained by positioning the reference line beneath the peak at a vertical distance equal to half the peak prominence [59].

Considering the two-step procedure described above, wheezes that are composed of several narrowband spectral peaks ($\eta > 1$) can be classified as MP or PP as follows,

$$\text{Wheezing category} = \begin{cases} \text{MP} & \text{if} \quad f_p(z) \subseteq \Lambda_z \quad , z = 2...\eta \\ \text{PP} & \text{otherwise} \end{cases} \tag{21}$$

where $v \subseteq V$ denotes that element v is contained in the interval V. Therefore, when the frequency $f_p(z)$ of all possible harmonic spectral peaks $z = 2...\eta$ are located in their corresponding spectral intervals Λ_z, the wheezing is classified as MP. Otherwise, wheezing is classified as PP because the narrowband spectral peaks that characterize the wheezing are not harmonically related. This occurs when the frequency $f_p(z)$ at least one of the possible harmonic spectral peaks is not located in its corresponding spectral intervals Λ_z. Figure 7 shows an example of the procedure described for MP

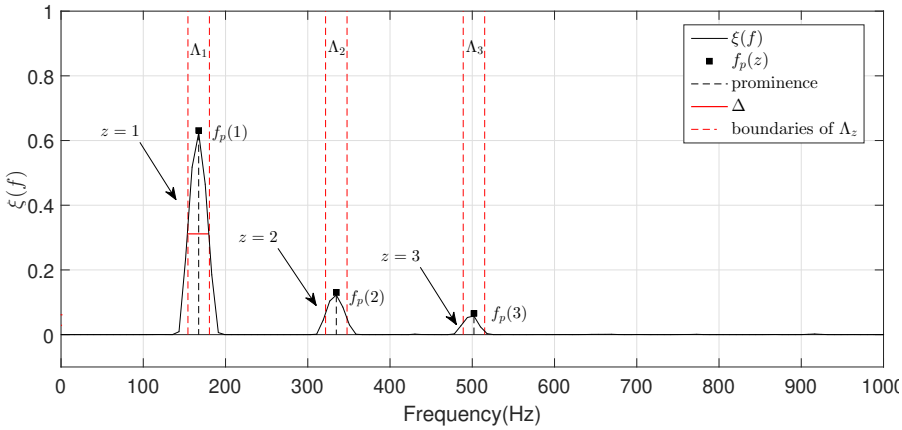

Figure 7. Example of the proposed two-step procedure to classify between MP and PP wheezing when $\eta > 1$ from the example of MP wheezing shown in Figure 1B. Note that the arrows indicate the narrowband spectral peaks that compose the wheezing. In this case, the wheezing is classified as MP because all spectral peaks are harmonically related.

wheezing composed of a basal peak and two harmonics. Figure 8 shows two examples of the procedure described for two PP wheezing with several non-harmonically related spectral peaks. Finally, the pseudo code of this stage for the classification between MP/PP wheezing according to its harmonic structure is detailed in the Algorithm 2.

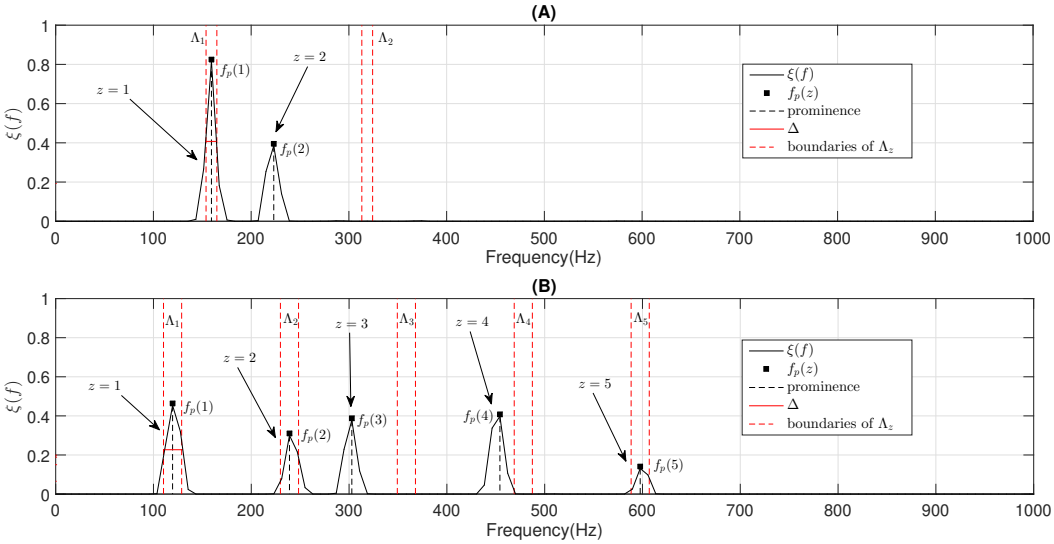

Figure 8. Example of the proposed two-step procedure to classify between MP and PP wheezing when $\eta > 1$, considering the two examples of PP wheezing shown in Figure 2. (A) Two-step procedure applied to the PP wheezing shown in Figure 2A. (B) Two-step procedure applied to the PP wheezing shown in Figure 2B. Note that the arrows indicate the narrowband spectral peaks that compose the wheezing. In this case both wheezing are classified as PP because not all spectral peaks are harmonically related.

4. Experimental Results and Discussion

4.1. Data collection

As far as the authors knowledge extends, there is no public wheeze database in which wheezing has been labeled as monophonic or polyphonic. For this reason, we have received the collaboration of a pneumologist from the University Hospital of Jaén (Spain) to create and label a database according to

Algorithm 2 Harmonic structure analysis

Require: $\xi(f)$.
1: From $\xi(f)$ detect the number η of narrowband spectral peaks.
if $\eta = 1$ **then**
  return Wheezing category = MP
else
  2: Locate the frequency $f_p(z)$ in each spectral peak $z = 1, ..., \eta$.
  3: Compute the spectral intervals Λ_z using Equation (20).
  if $f_p(z) \subseteq \Lambda_z$, $z = 2...\eta$ **then**
    return Wheezing category = MP
  else
    return Wheezing category = PP
  end if
end if

the wheezing harmonic structure. The database has been created by collecting and categorizing a set of recordings from different subjects of the most widely used Internet pulmonary repositories [60–72]. Specifically, all previous recordings were collected from subjects with CDRs (asthma or COPD). Note that the set of recordings selected for this assessment are only composed of normal respiratory sounds and wheezing sounds.

The type of wheezing (MP or PP) was labeled by the pneumologist by means of an acoustic inspection and a visual verification of the spectrogram considering the harmonic structure that distinguishes both types of wheezing. The database consisted of 200 MP and 200 PP wheezing segments, where the duration of each segment was at least 100 ms, to be consistent with literature. As mentioned above, MP wheezing can show two different harmonic structures: (type 1) wheezes with a single peak, that is, only the fundamental frequency component is active, and (type 2) wheezes with the harmonics of a single basal peak, that is, both the fundamental frequency component and its frequencies harmonically related are active. Therefore, to guarantee the maximum variability of the MP wheezing, the 200 MP wheezing segments are divided into 100 MP wheezing segments with a single peak and 100 MP wheezing segments with the harmonics of a single basal peak. Note that all segments are independent of each other, since each segment corresponds to a different wheezing from the rest. Finally, all segments in the database are sampled at 4096 Hz and have a length between 100 and 700 ms. Figure 9 shows the classification performed on the database created.

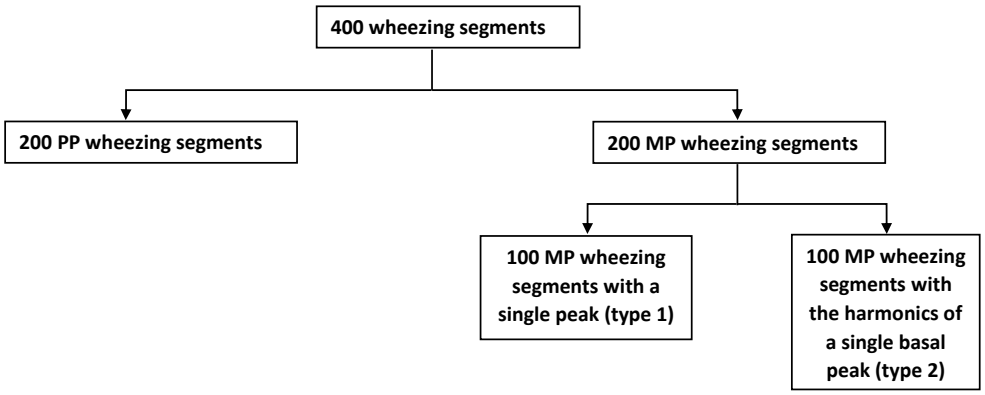

Figure 9. Scheme of the types of wheezing contained in the database.

4.2. Experimental Setup

To be consistent with the literature, we assume that wheezing sounds are not active below 100Hz and above 1000Hz. For this reason, all segments that compose the database were band-limited from 100Hz-1000Hz.

The length of the signal frames is set to $N = 256$ samples (62.5 ms). This frame size is considered large enough to assume a perfect spectral representation of all wheezing frequency components. The overlap between frames is set to 10% (6.25 ms). To obtain the time-frequency representation, windowing with a Hamming window is applied, and the order of the Discrete Fourier Transform (DFT) is set to $2N$ frequency bins similarly as occurs in [9,33]. This DFT size provides a high enough resolution for modeling the spectral patterns of wheezing sounds, and was chosen empirically as a trade-off between achieved quality and complexity. Besides, we have empirically observed that the reconstruction error converges after a 50 iterations so, the maximum number of iterations for the decomposition equal to $M = 50$.

Finally, note that the performance of the proposed method depends on the initial values with which the basis matrices $\mathbf{B}_W$, $\mathbf{B}_R$ and the activation matrices $\mathbf{A}_W$, $\mathbf{A}_R$ have been initialised. Although the obtained results are not dispersed and keep the same behavior, in order to overcome this issue, we have run the proposed method five times for each segment that composes the database and the results shown in this paper are averaged values.

4.3. Evaluation Metrics

The accuracy rates (ACC) are used to evaluate the performance of the proposed method, which are commonly used in the field of wheezing classification [7]. In order to provide a fair evaluation of the classification performance obtained by the proposed method and the state-of-the-art algorithms, the following accuracy rates are proposed: (i) ACC_G is the ability to correctly classify a wheezing segment as MP or PP; (ii) ACC_P represents the ability to correctly classify a wheezing segment as PP; (iii) ACC_M corresponds to the ability to correctly classify a wheezing segment as MP; (iv) ACC_{M1} indicates the ability to correctly classify a wheezing segment as MP type 1; and (v) ACC_{M2} reports the ability to correctly classify a wheezing segment as MP type 2. The terms used in Equations (22)-(26) are described in Table 1.

$$ACC_G = \frac{(TP + TM)}{(TP + TM + FP + FM))} \tag{22}$$

$$ACC_P = \frac{TP}{(TP + FP)} \tag{23}$$

$$ACC_M = \frac{TM}{(TM + FM)} \tag{24}$$

$$ACC_{M1} = \frac{TM1}{(TM1 + FM1)} \tag{25}$$

$$ACC_{M2} = \frac{TM2}{(TM2 + FM2)} \tag{26}$$

4.4. State-of-the-art method for comparison

In order to measure the MP/PP classification performance of the proposal, we have used the most recent and relevant state-of-the-art algorithm [7], denoted as UPER in this paper. The method UPER have been implemented strictly following the instructions specified by the authors in [7]. Firstly, the values of the metric PER have been obtained using the 19^{th} parameter set ($p = 10$, $q = 11$, $s = 7$ and $J = 45$) in the RADWT model. Then, three classifiers, support vector machine (SVM) with radial basis function kernel (RBF kernel), k-nearest neighbor (KNN) and extreme learning machine

Terms	Definitions
TP (True PP)	PP wheezing segments correctly classified
TM (True MP)	MP wheezing segments correctly classified
FP (False PP)	PP wheezing segments misclassified as MP
FM (False MP)	MP wheezing segments misclassified as PP
$TM1$ (True MP type 1)	MP type 1 wheezing segments correctly classified
$TM2$ (True MP type 2)	MP type 2 wheezing segments correctly classified
$FM1$ (False MP type 1)	MP type 1 wheezing segments misclassified as PP
$FM2$ (False MP type 2)	MP type 2 wheezing segments misclassified as PP

Table 1. Definition of the terms that appear in the metrics detailed in Equations (22)-(26)

(ELM) were applied to PER features. The classification performance of UPER has been obtained in leave-one-out (LOO) cross validation schemes with SVM, KNN and ELM classifiers. Specifically, LOO cross-validation is a particular case of leave-p-out (LPO) cross-validation with $p = 1$. So, LOO scheme involves using 1 observation as the validation set and the remaining observations as the training set. This is repeated in all ways to cut the database on a validation set of 1 observation and a training set. Considering the database evaluated in this work (400 segments in total), the LOO cross validation scheme has 400 possible combinations of validation in which the training set is composed of 399 segments and only one segment is tested, as can be observed in Figure 10). Results shown in this paper for all classifiers are the average values obtained from the 400 possible validation combinations.

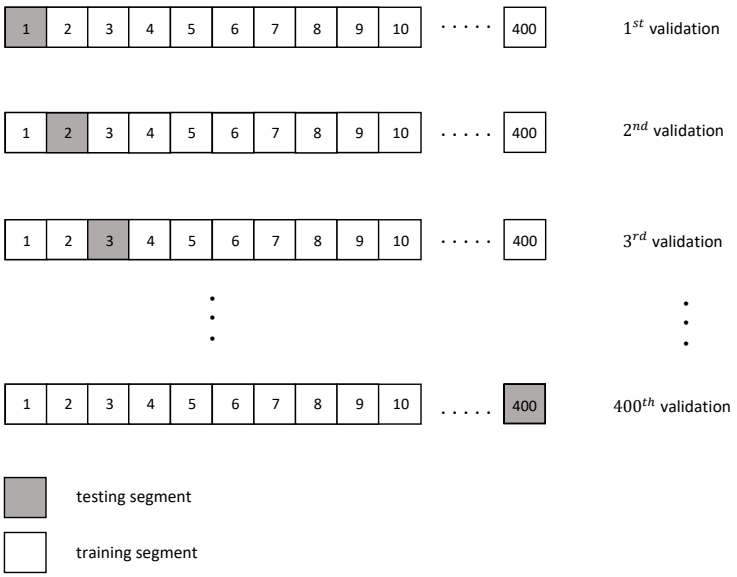

Figure 10. LOO cross-validation scheme for the database described in this paper.

4.5. Accuracy results

In this section, we have evaluated the MP/PP classification performance between the proposed method and UPER [7]. A remarkable distinction between the two methods is that the proposed method is completely unsupervised or blind (no training) but the method UPER depends on a training database.

Table 2 shows the MP/PP classification results, in terms of the accuracy rates, evaluating the database described in section 4.1. Results provided by UPER, considering the three classifier versions (SVM, KNN and ELM), have been obtained by applying a LOO cross validation scheme as was previously described in section 4.4. Results report that the proposed method provides the best overall MP/PP classification results compared to UPER considering all evaluated metrics. Focusing on the different accuracy rates, the following can be observed:

- the improvement, in terms of ACC_G, of the proposed method is about 8.25% UPER (SVM), 12% UPER (KNN) and 10.5% UPER (ELM).
- the improvement, in terms of ACC_P, of the proposed method is about 4% UPER (SVM), 7.1% UPER (KNN) and 5.5% UPER (ELM).
- the improvement, in terms of ACC_M, of the proposed method is about 12.5% UPER (SVM), 17% UPER (KNN) and 15.5% UPER (ELM).
- the improvement, in terms of ACC_{M1}, of the proposed method is about 5% UPER (SVM), 10% UPER (KNN) and 8% UPER (ELM).
- the improvement, in terms of ACC_{M2}, of the proposed method is about 20% UPER (SVM), 24% UPER (KNN) and 23% UPER (ELM).

Algorithm	$\mathbf{ACC_G}$	$\mathbf{ACC_P}$	$\mathbf{ACC_M}$	$\mathbf{ACC_{M1}}$	$\mathbf{ACC_{M2}}$
Proposed Method	**92%**	**91.5%**	**92.5%**	**91%**	**94%**
UPER (SVM) [7]	83.75%	87.5%	80%	86%	74%
UPER (KNN) [7]	80%	84.4%	75.5%	81%	70%
UPER (ELM) [7]	81.5%	86%	77%	83%	71%

Table 2. Comparative ACC results between the proposed method and UPER.

The main advantage of UPER is that it only uses one feature (PER value) to discriminate between MP and PP wheezing. As shown in Table 2, the SVM classifier obtains the best classification performance in the method UPER. Specifically, the classifier SVM achieves an improvement of 2.25% (KNN) and 1.5% (KNN), in terms of ACC_G. These results are consistent with those obtained by the authors in [7], confirming that the SVM classifier with RBF kernel obtains the best classification performance when the number of features (only one PER value) is small [73].

Performing an empirical analysis of the proposed method and UPER, the following observations were extracted:

(i) Due to the time-frequency overlapping problem, normal respiratory sounds often mask wheezing sounds hiding relevant medical information [5]. While the proposed method (based on CL-RNMF) allows to remove as much as possible the acoustic interference from normal respiratory sounds, the method UPER is based on a feature PER obtained from the sub-band energy of the wavelet coefficients so, the presence of normal respiratory sounds interferes in the selection of the optimal sub-bands that really belong to the wheezing components.

(ii) The method UPER has more difficulty to discriminate between PP and MP wheezing composed by a basal peak and its harmonics since it achieves the worst performance in terms of ACC_{M2}. The reason is because UPER is based on energy and ignoring the spectral location of the components that model the harmonic behavior of the MP wheezing. Results in Table 2 suggest that MP/PP classification based on the spectral location of the harmonic structure as occurs in the proposed method is more reliable than the use of the energy of the wheezing spectral components as occurs in UPER.

The LOO cross-validation scheme does not show the dependency that classifiers have with the size of the training segments set, since this scheme always uses 1 segment as the validation set and the remaining segments as the training set. For this reason, we propose to use an LPO cross-validation scheme by varying the size of the training segments set. LPO scheme requires training and validating the model C_p^n times, where n is the number of segments that compose the database, p is the number of validation segments, and C_p^n is the binomial coefficient. As a result, the associated computational cost can be excessive. In order to overcome this issue, we have limited the number of iterations of the LPO scheme to 500. Furthermore, the same number of MP and PP wheezes are selected for both training and validation sets in each iteration. Specifically, we have used four LPO schemes: (i) $p = 80$ uses 80% of the total segments as the training set in each iteration; (ii) $p = 160$ uses 60% of the total segments

as the training set in each iteration; (iii) $p = 240$ uses 40% of the total segments as the training set in each iteration; and (iv) $p = 320$ uses 20% of the total segments as the training set in each iteration. Considering all the instructions described above, Table 3 shows the MP/PP classification results, in terms of ACC_G, obtained by UPER using its three classifier versions (SVM, KNN and ELM) in order to assess its dependence on the training set size. Comparing the LOO scheme with the LPO scheme ($p = 320$), the ACC_G reduction of the classification performance is about 7.5% (SVM), 8.25% (KNN) and 6.25% (ELM). Results report that the PER feature allows to distinguish between MP and PP wheezing even when the training set size is reduced. In addition, ELM classifier shows less dependence on the training database size compared to SVM and KNN.

Scheme	Training set	Validate set	SVM	KNN	ELM
LOO	399 (99.75%)	1 (0.25%)	**83.75%**	80%	81.5%
LPO ($p = 80$)	320 (80%)	80 (20%)	**81.5%**	79.25%	80%
LPO ($p = 160$)	240 (60%)	160 (40%)	**80.5%**	77.75%	79.5%
LPO ($p = 240$)	160 (40%)	240 (60%)	**78.25%**	74.75%	77.25%
LPO ($p = 320$)	80 (20%)	320 (80%)	**76.25%**	71.75%	75.25%

Table 3. Comparative results, in term of ACC_G, between the three classifier versions (SVM, KNN and ELM) of the method UPER using four LPO cross-validation schemes.

5. Conclusions and Future Work

In this paper, we present a novel Constrained Low-rank Non-negative Matrix Factorization (CL-RNMF) approach to classify monophonic and polyphonic wheezing sounds according to its harmonic structure. The first contribution of this work proposes a CL-RNMF framework that allows extracting the spectral patterns that characterize wheezing sounds with the least possible interference from normal respiratory sounds. Specifically, a low-rank configuration with a reduced number of wheezing bases is presented to compact its frequency components in the least number of bases possible for their posterior analysis. In addition, CL-RNMF uses a set of constraints to model the spectro-temporal behavior of wheezing and normal respiratory sounds. As far as the authors knowledge extends, Non-negative Matrix Factorization approach has never been applied before to MP/PP wheezing classification. The second contribution analyzes the harmonic structure of the energy distribution from the estimated wheezing spectrogram provided by CL-RNMF to determine the type of wheezing, allowing a more efficient classification based on the location of the wheezing frequency components, rather than the energy of their components.

The most relevant conclusions from the experimental results indicate the following: (i) the proposed method provides the best overall performance related to MP/PP wheezing classification compared to the most relevant method of the state-of-the-art; (ii) unlike most state-of-the-art methods based on classifiers, the proposed method is an unsupervised (blind) approach that does not require any training from wheezing sounds; (iii) the proposed method achieves to remove most of the interference from normal respiratory sounds; (iv) specific accuracy rates, ACC_M and ACC_P, obtained by the proposed method seem to suggest the ability of the proposal to classify both monophonic and polyphonic wheezing sounds correctly.

Future work will be focused on the design of new constraints, to be applied in NMF approaches, that improve the modelling of time-frequency respiratory sound events analyzing different types of adventitious sounds, such as wheezes and crackles. The objective of this future research line is to perform an early detection and classification among the different types of adventitious sounds active in the auscultation process in order to maximize the reliability of the diagnosis issued by the physician in case of pathologies of lung diseases caused by the appearance of such adventitious sounds.

Author Contributions: conceptualization, J.D.L.T.C., F.J.C.Q., N.R.R. and G.P.C.; data curation, J.D.L.T.C. and F.J.C.Q.; formal analysis, J.D.L.T.C., F.J.C.Q., N.R.R., S.G.G. and J.J.C.O.; investigation, J.D.L.T.C. and F.J.C.Q.; methodology, J.D.L.T.C., F.J.C.Q. and N.R.R.; software, J.D.L.T.C., F.J.C.Q., S.G.G. and J.J.C.O.; supervision, F.J.C.Q., N.R.R., S.G.G. and J.J.C.O.; validation, J.D.L.T.C., F.J.C.Q. and G.P.C.; visualization, J.D.L.T.C.; writing–original

draft, J.D.L.T.C., F.J.C.Q., N.R.R., S.G.G., J.J.C.O. and G.P.C.; writing–review and editing, J.D.L.T.C., F.J.C.Q., N.R.R., S.G.G. and J.J.C.O. All authors have read and agreed to the submitted version of the manuscript.

Funding: This work was supported by the Programa Operativo FEDER Andalucia 2014-2020 under project with reference 1257914 and the Ministry of Economy, Knowledge and University, Junta de Andalucia under Project P18-RT-1994.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A Terms of the multiplicative update rules

Here, each of the terms belonging to the multiplicative update rule to obtain the basis matrix $\mathbf{B}_W$:

$$\left[\frac{\partial D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}\right)}{\partial \mathbf{B}_W}\right]^{-} = \left(\mathbf{X} \oslash \hat{\mathbf{X}}\right)\mathbf{A}_W^{\mathrm{T}} \tag{A1}$$

$$\left[\frac{\partial D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}\right)}{\partial \mathbf{B}_W}\right]^{+} = \left([\mathbf{1}]\mathbf{A}_W^{\mathrm{T}}\right) \tag{A2}$$

$$\left[\frac{\partial \psi\left(\mathbf{B}_W\right)}{\partial \mathbf{B}_W}\right]^{-}_{f,k_w} = \sqrt{F}\left(\frac{B_{W_{f,k_w}} \sum_{j=1}^{F} B_{W_{j,k_w}}}{(\sum_{j=1}^{F} B^2_{W_{j,k_w}})^{\frac{3}{2}}}\right) \tag{A3}$$

$$\left[\frac{\partial \psi\left(\mathbf{B}_W\right)}{\partial \mathbf{B}_W}\right]^{+}_{f,k_w} = \frac{1}{\sqrt{\frac{1}{F}\sum_{j=1}^{F} B^2_{W_{j,k_w}}}} \tag{A4}$$

Here, each of the terms belonging to the multiplicative update rule to obtain the basis matrix $\mathbf{B}_R$:

$$\left[\frac{\partial D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}\right)}{\partial \mathbf{B}_R}\right]^{-} = \left(\mathbf{X} \oslash \hat{\mathbf{X}}\right)\mathbf{A}_R^{\mathrm{T}} \tag{A5}$$

$$\left[\frac{\partial D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}\right)}{\partial \mathbf{B}_R}\right]^{+} = \left([\mathbf{1}]\mathbf{A}_R^{\mathrm{T}}\right) \tag{A6}$$

$$\begin{aligned}\left[\frac{\partial \phi\left(\mathbf{B}_R\right)}{\partial \mathbf{B}_R}\right]^{-}_{f,k_r} = 2F\left(\frac{(B_{R_{f-1,k_r}} + B_{R_{f+1,k_r}})}{\sum_{j=1}^{F} B^2_{R_{j,k_r}}}\right) + \\ + \frac{2FB_{R_{f,k_r}} \sum_{j=2}^{F} (B_{R_{j,k_r}} - B_{R_{j-1,k_r}})^2}{(\sum_{j=1}^{F} B^2_{R_{j,k_r}})^2}\end{aligned} \tag{A7}$$

$$\left[\frac{\partial \phi\left(\mathbf{B}_R\right)}{\partial \mathbf{B}_R}\right]^{+}_{f,k_r} = \frac{4FB_{R_{f,k_r}}}{\sum_{j=1}^{F} B^2_{R_{j,k_r}}} \tag{A8}$$

Here, each of the terms belonging to the multiplicative update rule to obtain the activations matrix $\mathbf{A}_W$:

$$\left[\frac{\partial D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}\right)}{\partial \mathbf{A}_W}\right]^{-} = \mathbf{B}_W^{\mathrm{T}}\left(\mathbf{X} \oslash \hat{\mathbf{X}}\right) \tag{A9}$$

$$\left[\frac{\partial D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}\right)}{\partial \mathbf{A}_W}\right]^{+} = \left(\mathbf{B}_W^{\mathrm{T}}[\mathbf{1}]\right) \tag{A10}$$

$$\left[\frac{\partial\phi\left(\hat{\mathbf{A}}_W\right)}{\partial\mathbf{A}_W}\right]^{-}_{k_w,t} = 2T\left(\frac{\left(A_{W_{k_w,t-1}} + A_{W_{k_w,t+1}}\right)}{\sum_{i=1}^{T} A^2_{W_{k_w,i}}}\right) + \\ + \frac{2TA_{W_{k_w,t}}\sum_{i=2}^{T}(A_{W_{k_w,i}} - A_{W_{k_w,i-1}})^2}{(\sum_{i=1}^{T} A^2_{W_{k_w,i}})^2} \tag{A11}$$

$$\left[\frac{\partial\phi\left(\mathbf{A}_W\right)}{\partial\mathbf{A}_W}\right]^{+}_{k_w,t} = \frac{4TA_{W_{k_w,t}}}{\sum_{i=1}^{T} A^2_{W_{k_w,i}}} \tag{A12}$$

Here, each of the terms belonging to the multiplicative update rule to obtain the activations matrix $\mathbf{A}_R$:

$$\left[\frac{\partial D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}\right)}{\partial\mathbf{A}_R}\right]^{-} = \mathbf{B}_R^{\mathrm{T}}\left(\mathbf{X} \oslash \hat{\mathbf{X}}\right) \tag{A13}$$

$$\left[\frac{\partial D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}\right)}{\partial\mathbf{A}_R}\right]^{+} = \left(\mathbf{B}_R^{\mathrm{T}}[\mathbf{1}]\right) \tag{A14}$$

References

1. World Health Organization, Chronic Respiratory Diseases. https://www.who.int/health-topics/chronic-respiratory-diseases#tab=tab_1, Online. Accessed: 2020-12-30.
2. World Health Organization, Asthma. https://www.who.int/news-room/fact-sheets/detail/asthma, Online. Accessed: 2020-12-30.
3. World Health Organization, Chronic obstructive pulmonary disease . http://www.emro.who.int/health-topics/chronic-obstructive-pulmonary-disease-copd/index.html, Online. Accessed: 2020-12-30.
4. Sarkar, M.; Madabhavi, I.; Niranjan, N.; Dogra, M. Auscultation of the respiratory system. *Annals of thoracic medicine* **2015**, *10*, 158.
5. Pasterkamp, H.; Kraman, S.S.; Wodicka, G.R. Respiratory sounds: advances beyond the stethoscope. *American journal of respiratory and critical care medicine* **1997**, *156*, 974–987.
6. Lozano-Garcia, M.; Fiz, J.A.; Martinez-Rivera, C.; Torrents, A.; Ruiz-Manzano, J.; Jane, R. Novel approach to continuous adventitious respiratory sound analysis for the assessment of bronchodilator response. *PloS one* **2017**, *12*, e0171455.
7. Ulukaya, S.; Serbes, G.; Kahya, Y.P. Wheeze type classification using non-dyadic wavelet transform based optimal energy ratio technique. *Computers in biology and medicine* **2019**, *104*, 175–182.
8. Andrès, E.; Gass, R.; Charloux, A.; Brandt, C.; Hentzler, A. Respiratory sound analysis in the era of evidence-based medicine and the world of medicine 2.0. *Journal of medicine and life* **2018**, *11*, 89.
9. Torre-Cruz, J.; Canadas-Quesada, F.; Carabias-Orti, J.; Vera-Candeas, P.; Ruiz-Reyes, N. A novel wheezing detection approach based on constrained non-negative matrix factorization. *Applied Acoustics* **2019**, *148*, 276–288.
10. Leng, S.; San Tan, R.; Chai, K.T.C.; Wang, C.; Ghista, D.; Zhong, L. The electronic stethoscope. *Biomedical engineering online* **2015**, *14*, 1–37.
11. Sen, I.; Saraclar, M.; Kahya, Y.P. A comparison of SVM and GMM-based classifier configurations for diagnostic classification of pulmonary sounds. *IEEE Transactions on Biomedical Engineering* **2015**, *62*, 1768–1776.
12. Salazar, A.J.; Alvarado, C.; Lozano, F.E. System of heart and lung sounds separation for store-and-forward telemedicine applications. *Revista Facultad de Ingeniería Universidad de Antioquia* **2012**, pp. 175–181.
13. Sovijarvi, A.; Dalmasso, F.; Vanderschoot, J.; Malmberg, L.; Righini, G.; Stoneman, S. Definition of terms for applications of respiratory sounds. *European Respiratory Review* **2000**, *10*, 597–610.
14. Meslier, N.; Charbonneau, G.; Racineux, J. Wheezes. *European respiratory journal* **1995**, *8*, 1942–1948.

15. Baughman, R.P.; Loudon, R.G. Lung sound analysis for continuous evaluation of airflow obstruction in asthma. *Chest* **1985**, *88*, 364–368.
16. Cortes, S.; Jane, R.; Fiz, J.; Morera, J. Monitoring of wheeze duration during spontaneous respiration in asthmatic patients. 27th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2006, pp. 6141–6144.
17. Qiu, Y.; Whittaker, A.; Lucas, M.; Anderson, K. Automatic wheeze detection based on auditory modelling. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine* **2005**, *219*, 219–227.
18. Zhang, J.; Ser, W.; Yu, J.; Zhang, T. A novel wheeze detection method for wearable monitoring systems. International Symposium on Intelligent Ubiquitous Computing and Education. IEEE, 2009, pp. 331–334.
19. Lin, B.S.; Wu, H.D.; Chen, S.J. Automatic wheezing detection based on signal processing of spectrogram and back-propagation neural network. *Journal of healthcare engineering* **2015**, *6*, 649–672.
20. Kochetov, K.; Putin, E.; Azizov, S.; Skorobogatov, I.; Filchenkov, A. Wheeze detection using convolutional neural networks. EPIA Conference on Artificial Intelligence. Springer, 2017, pp. 162–173.
21. Kandaswamy, A.; Kumar, C.S.; Ramanathan, R.P.; Jayaraman, S.; Malmurugan, N. Neural classification of lung sounds using wavelet coefficients. *Computers in biology and medicine* **2004**, *34*, 523–537.
22. Le Cam, S.; Belghith, A.; Collet, C.; Salzenstein, F. Wheezing sounds detection using multivariate generalized Gaussian distributions. IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2009, pp. 541–544.
23. Wisniewski, M.; Zielinski, T.P. Tonality detection methods for wheezes recognition system. 19th International Conference on Systems, Signals and Image Processing (IWSSIP). IEEE, 2012, pp. 472–475.
24. Wisniewski, M.; Zielinski, T.P. Joint application of audio spectral envelope and tonality index in an e-asthma monitoring system. *IEEE journal of biomedical and health informatics* **2015**, *19*, 1009–1018.
25. Chien, J.C.; Wu, H.D.; Chong, F.C.; Li, C.I. Wheeze detection using cepstral analysis in gaussian mixture models. 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2007, pp. 3168–3171.
26. Bahoura, M. Pattern recognition methods applied to respiratory sounds classification into normal and wheeze classes. *Computers in biology and medicine* **2009**, *39*, 824–843.
27. Bahoura, M.; Pelletier, C. Respiratory sounds classification using Gaussian mixture models. Canadian Conference on Electrical and Computer Engineering. IEEE, 2004, Vol. 3, pp. 1309–1312.
28. Mayorga, P.; Druzgalski, C.; Morelos, R.; Gonzalez, O.; Vidales, J. Acoustics based assessment of respiratory diseases using GMM classification. Annual International Conference of the IEEE Engineering in Medicine and Biology. IEEE, 2010, pp. 6312–6316.
29. Taplidou, S.A.; Hadjileontiadis, L.J. Wheeze detection based on time-frequency analysis of breath sounds. *Computers in biology and medicine* **2007**, *37*, 1073–1083.
30. Jain, A.; Vepa, J. Lung sound analysis for wheeze episode detection. 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2008, pp. 2582–2585.
31. Mendes, L.; Vogiatzis, I.; Perantoni, E.; Kaimakamis, E.; Chouvarda, I.; Maglaveras, N.; Tsara, V.; Teixeira, C.; Carvalho, P.; Henriques, J.; others. Detection of wheezes using their signature in the spectrogram space and musical features. 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2015, pp. 5581–5584.
32. Oletic, D.; Bilas, V. Asthmatic wheeze detection from compressively sensed respiratory sound spectra. *IEEE journal of biomedical and health informatics* **2018**, *22*, 1406–1414.
33. Torre-Cruz, J.; Canadas-Quesada, F.; García-Galán, S.; Ruiz-Reyes, N.; Vera-Candeas, P.; Carabias-Orti, J. A constrained tonal semi-supervised non-negative matrix factorization to classify presence/absence of wheezing in respiratory sounds. *Applied Acoustics* **2020**, *161*, 107188.
34. De La Torre Cruz, J.; Cañadas Quesada, F.J.; Carabias Orti, J.J.; Vera Candeas, P.; Ruiz Reyes, N. Combining a recursive approach via non-negative matrix factorization and Gini index sparsity to improve reliable detection of wheezing sounds. *Expert Systems with Applications* **2020**, *147*, 113212.
35. Nagasaka, Y. Lung Sounds in Bronchial Asthma. *Allergology International* **2012**, *61*, 353–363.
36. Mason, R.C.; Murray, J.F.; Nadel, J.A.; Gotway, M.B. *Murray & Nadel's Textbook of Respiratory Medicine E-Book*; Elsevier Health Sciences, 2015.

37. Taplidou, S.A.; Hadjileontiadis, L.J. Analysis of wheezes using wavelet higher order spectral features. *IEEE Transactions on biomedical engineering* **2010**, *57*, 1596–1610.
38. Forgacs, P. The functional basis of pulmonary sounds. *Chest* **1978**, *73*, 399–405.
39. Jácome, C.; Oliveira, A.; Marques, A. Computerized respiratory sounds: a comparison between patients with stable and exacerbated COPD. *The clinical respiratory journal* **2017**, *11*, 612–620.
40. Hashemi, A.; Arabalibiek, H.; Agin, K. Classification of wheeze sounds using wavelets and neural networks. International Conference on Biomedical Engineering and Technology. IACSIT Press, 2011, Vol. 11, pp. 127–131.
41. Jain, A.; Vepa, J. Lung sound analysis for wheeze episode detection. 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, 2008, pp. 2582–2585.
42. Jin, F.; Krishnan, S.; Sattar, F. Adventitious sounds identification and extraction using temporal–spectral dominance-based features. *IEEE Transactions on Biomedical Engineering* **2011**, *58*, 3078–3087.
43. Naves, R.; Barbosa, B.H.; Ferreira, D.D. Classification of lung sounds using higher-order statistics: A divide-and-conquer approach. *Computer methods and programs in biomedicine* **2016**, *129*, 12–20.
44. Ulukaya, S.; Sen, I.; Kahya, Y.P. A novel method for determination of wheeze type. 23nd Signal Processing and Communications Applications Conference (SIU), 2015, pp. 2001–2004. doi:10.1109/SIU.2015.7130257.
45. Ulukaya, S.; Sen, I.; Kahya, Y.P. Feature extraction using time-frequency analysis for monophonic-polyphonic wheeze discrimination. 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2015, pp. 5412–5415.
46. Lee, D.D.; Seung, H.S. Learning the parts of objects by non-negative matrix factorization. *Nature* **1999**, *401*, 788–791.
47. Lee, D.D.; Seung, H.S. Algorithms for non-negative matrix factorization. Advances in neural information processing systems, 2001, pp. 556–562.
48. Canadas-Quesada, F.; Ruiz-Reyes, N.; Carabias-Orti, J.; Vera-Candeas, P.; Fuertes-Garcia, J. A non-negative matrix factorization approach based on spectro-temporal clustering to extract heart sounds. *Applied Acoustics* **2017**, *125*, 7–19.
49. Dia, N.; Fontecave-Jallon, J.; Gumery, P.Y.; Rivet, B. Denoising Phonocardiogram signals with Non-negative Matrix Factorization informed by synchronous Electrocardiogram. 2018 26th European Signal Processing Conference (EUSIPCO). IEEE, 2018, pp. 51–55.
50. Torre-Cruz, J.; Canadas-Quesada, F.; Vera-Candeas, P.; Montiel-Zafra, V.; Ruiz-Reyes, N. Wheezing sound separation based on constrained non-negative matrix factorization. Proceedings of the 2018 10th International Conference on Bioinformatics and Biomedical Technology, 2018, pp. 18–24.
51. De La Torre Cruz, J.; Cañadas Quesada, F.J.; Ruiz Reyes, N.; Vera Candeas, P.; Carabias Orti, J.J. Wheezing Sound Separation Based on Informed Inter-Segment Non-Negative Matrix Partial Co-Factorization. *Sensors* **2020**, *20*, 2679.
52. Févotte, C.; Bertin, N.; Durrieu, J.L. Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. *Neural computation* **2009**, *21*, 793–830.
53. Liutkus, A.; Fitzgerald, D.; Badeau, R. Cauchy nonnegative matrix factorization. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). IEEE, 2015, pp. 1–5.
54. Canadas-Quesada, F.J.; Vera-Candeas, P.; Ruiz-Reyes, N.; Carabias-Orti, J.; Cabanas-Molero, P. Percussive/harmonic sound separation by non-negative matrix factorization with smoothness/sparseness constraints. *EURASIP Journal on Audio, Speech, and Music Processing* **2014**, *2014*, 26.
55. Laroche, C.; Kowalski, M.; Papadopoulos, H.; Richard, G. A structured nonnegative matrix factorization for source separation. 2015 23rd European Signal Processing Conference (EUSIPCO). IEEE, 2015, pp. 2033–2037.
56. Eggert, J.; Korner, E. Sparse coding and NMF. 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541). IEEE, 2004, Vol. 4, pp. 2529–2533.
57. Virtanen, T. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE transactions on audio, speech, and language processing* **2007**, *15*, 1066–1074.
58. Marxer, R.; Janer, J. Study of regularizations and constraints in NMF-based drums monaural separation. International Conference on Digital Audio Effects Conference (DAFx-13), 2013.

59. Prominence criterion of a peak according to the MATLAB software. https://es.mathworks.com/help/signal/ref/findpeaks.html?searchHighlight=findpeak&s_tid=doc_srchtitle#buff2uu, Online. Accessed: 2020-12-30.
60. The r.a.l.e. repository. http://www.rale.ca, Online. Accessed: 2020-12-30.
61. Stethographics lung sound samples. http://www.stethographics.com, Online. Accessed: 2020-12-30.
62. 3m littmann stethoscopes. https://www.3m.com, Online. Accessed: 2020-12-30.
63. East tennessee state university pulmonary breath sounds. http://faculty.etsu.edu, Online. Accessed: 2020-12-30.
64. ICBHI 2017 Challenge. https://bhichallenge.med.auth.gr, Online. Accessed: 2020-12-30.
65. Lippincott NursingCenter. https://www.nursingcenter.com, Online. Accessed: 2020-12-30.
66. Thinklabs Digital Stethoscope. https://www.thinklabs.com, Online. Accessed: 2020-12-30.
67. Thinklabs youtube. https://www.youtube.com/channel/UCzEbKuIze4AI1523_AWiK4w, Online. Accessed: 2020-12-30.
68. Emedicine/Medscape. https://emedicine.medscape.com/article/1894146-overview#a3, Online. Accessed: 2020-12-30.
69. E-learning resources. https://www.ers-education.org/e-learning/reference-database-of-respiratory-sounds.aspx, Online. Accessed: 2020-12-30.
70. Respiratory wiki. http://respwiki.com/Breath_sounds, Online. Accessed: 2020-12-30.
71. Easy Auscultation. https://www.easyauscultation.com/lung-sounds-reference-guide, Online. Accessed: 2020-12-30.
72. Colorado State University. http://www.cvmbs.colostate.edu/clinsci/callan/breath_sounds.htm, Online. Accessed: 2020-12-30.
73. Chang, C.C.; Lin, C.J. LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* **2011**, *2*, 1–27.

Paper 7

An incremental algorithm based on multichannel non-negative matrix partial co-factorization for ambient denoising in auscultation

J. Torre-Cruz, F. Canadas-Quesada, D. Martínez-Muñoz, N. Ruiz-Reyes, S. García-Galán and J. Carabias-Orti, "An incremental algorithm based on multichannel non-negative matrix partial co-factorization for ambient denoising in auscultation", in *Applied Acoustics*. Status: under review.

- Estado: En revisión.
- Revista: Applied Acoustics.
- ISSN: 0003-682X.
- Factor de impacto (JCR 2019): 2.440.
- Cuartiles por área de conocimiento:
  - Acoustics: Q2, 9/32.

An incremental algorithm based on multichannel non-negative matrix partial co-factorization for ambient denoising in auscultation

Juan De La Torre Cruz *, Francisco Jesús Cañadas Quesada, Damián Martínez-Muñoz, Nicolás Ruiz Reyes, Sebastián García Galán, Julio José Carabias Orti

Departament of Telecommunication Engineering. University of Jaen, Campus Cientifico-Tecnologico de Linares, Avda. de la Universidad, s/n, 23700 Linares, Jaen, Spain

Abstract

One of the major current limitations in the diagnosis derived from auscultation remains the ambient noise surrounding the subject, which prevents successful auscultation. Therefore, it is essential to develop robust signal processing algorithms that can extract relevant clinical information from auscultated recordings analyzing in depth the acoustic environment in order to help the decision-making process made by physicians. The aim of this study is to implement a method to remove ambient noise in biomedical sounds captured in auscultation. We propose an incremental approach based on multichannel non-negative matrix partial co-factorization (NMPCF) for ambient denoising focusing on high noisy environment with a Signal-to-Noise Ratio (SNR) $\leq$ -5 dB. The first contribution applies NMPCF assuming that ambient noise can be modelled as repetitive sound events simultaneously found in two single-channel inputs captured by means of different recording devices. The second contribution proposes an incremental algorithm, based on the previous multichannel NMPCF, that refines the estimated biomedical spectrogram throughout a set of incremental stages by eliminating most of the ambient noise that was not removed in the previous stage at the expense of preserving most of the biomedical spectral content. To evaluate the performance of the proposed method, recordings composed of biomedical sounds mixed with ambient noise that typically surrounds a medical consultation room have been used to simulate high noisy environments showing a SNR from -20 dB to -5 dB. Experimental results show that the proposed method significantly outperforms the baseline method indicating that SDR and SIR improvement are near-constant in most of the SNR scenarios evaluated. A remarkable advantage of the proposed method is its high robustness to provide promising denoising performance analyzing delayed input signals (e.g., a SDR and SIR improvement equals to 13 dB and 21 dB when the delay equals to 25 ms). The proposed approach obtains the best ambient denoising results compared to the baseline method considering all evaluated sound scenarios taking into account biomedical sounds, ambient noises, SNR and delay.

Keywords: Auscultation, Biomedical, Ambient noise, Non-negative matrix partial co-factorization, Multichannel, Incremental

1. Introduction

Auscultation is defined as the technique of listening the internal sounds produced by the human organs by means of a stethoscope. This technique is simple, non-invasive, safe and inexpensive that provides valuable clinical information in the diagnosis of the status of the heart, lung and airways [1, 2]. Although

*Corresponding author. Tlf.: (+34) 953648592

Email addresses: jtorre@ujaen.es (Juan De La Torre Cruz *), fcanadas@ujaen.es (Francisco Jesús Cañadas Quesada), damian@ujaen.es (Damián Martínez-Muñoz), nicolas@ujaen.es (Nicolás Ruiz Reyes), sgalan@ujaen.es (Sebastián García Galán), carabias@ujaen.es (Julio José Carabias Orti)

today there are more advanced medical tools to analyze the status of the heart and lung, such as chest radiography, electrocardiography (ECG), spirometry or laboratory analyses, auscultation is still one of the most widely used techniques to detect any cardiac or pulmonary disease. However, the diagnosis derived from auscultation shows two main limitations: i) high subjectivity due to the physician's expertise to recognize sounds that reveal any physiological disorder [3]; ii) high dependence on the ambient noise surrounding the subject to provide a reliable diagnosis [4].

Because the process of auscultation in a soundproof room is not possible in most cases, especially in low-income and middle-income countries supported by a resource-poor health system, ambient denoising performed in the examination room of the health center is still a challenging task in biomedical signal processing in order to maximize the reliability of a diagnosis. The main effects caused by ambient noise are the masking, distortion and weakness of the sound of interest that may provide relevant clues in the diagnosis as shown in Figure 1. As a result, the probability of making a medical error increases when auscultation is performed in a noisy environment since the physician is not able to correctly interpret the diagnostic information contained in the sound signal from auscultation. In this work, the term biomedical means that the sound sources that have generated the sounds of interest have been the human internal organs, and specifically, the heart and the lung.

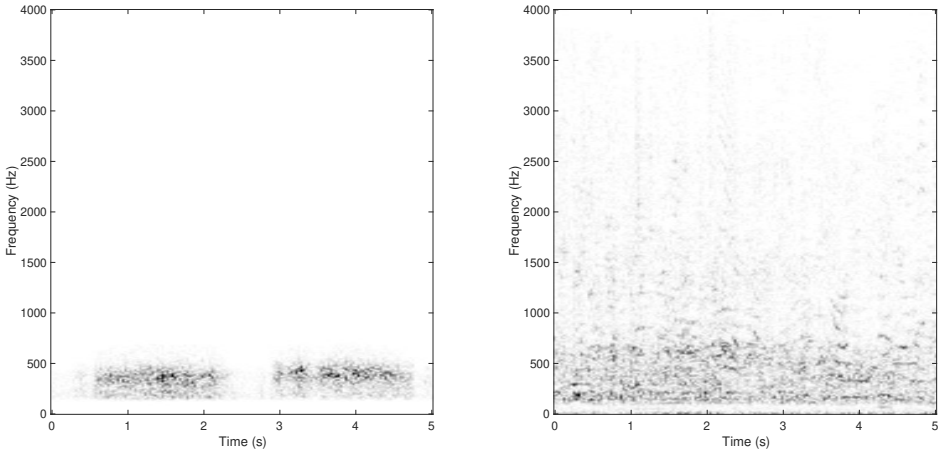

Figure 1: Spectrogram from a clean lung sound recording: (Left) under no ambient noise; (Right) mixed with ambient noise (babble) in a Signal-to-Noise Ratio (SNR) equals -10 dB. Comparing both figures, it can be observed that the spectral content from lung is completely masked by ambient noise. Higher energies are indicated by darker colour.

In recent years, several signal processing tasks have been applied in the field of biomedical information retrieval such as, sound source separation [5, 6, 7] as well as sound event detection [8, 9, 10, 11] and classification [12, 13, 14, 15, 16]. However, most of their experimental results have been obtained in environments in which the biomedical sounds are not acoustically contaminated by ambient noises. Therefore, the task of ambient denoising is still an open research topic in biomedical engineering being most of the approaches based on adaptive filtering [17, 18, 19, 20, 21] and spectral subtraction [22, 23, 24, 25]. Chang and Lai [23] proposed a two-channel spectral subtraction method, based on autoregressive (AR), mel-frequency cepstral coefficients (MFCC) and dynamic time warping (DTW), applied to the lung sound signals under noisy conditions before the extraction of lung sound features. Emmanouilidou et al. [24] developed a multiband spectral scheme, based on two-microphone setup, to suppress the background noise while successfully preserving the lung sound content to maximize the informative diagnostic value obtained from auscultation. The algorithm analyzes each frequency band in a nonuniform manner and uses prior knowledge of the target sounds to apply a penalty in the spectral domain. It follows from the above that it is crucial to develop robust signal processing algorithms that can extract relevant clinical information from auscultated recordings taking into consideration the acoustic environment that surrounds the subject in order to improve the decision making process made by physicians.

It is well known that the conventional Non-negative Matrix Partial Co-Factorization (NMPCF) enforces

a joint matrix decomposition using multiple matrices to recover a set of shared spectral patterns (bases) that model the spectral behavior of some of the sound sources contained in the single-channel input. Over the last decade, NMPCF has been successfully applied in several single-channel sound signal processing fields: i) Music Information Retrieval (MIR) such as, singing-voice separation [26], rhythmic extraction [27, 28, 29] and speaker diarization [30]; and ii) Enhancement of biomedical sounds such as, normal respiratory and wheezes sound separation [31]. In this work, we propose an incremental algorithm, called 2C-NMPCF, that improves the quality of biomedical sounds captured in auscultated recordings by applying the conventional NMPCF from a multi-channel scenario rather than a single-channel. In this paper, the term multichannel refers to the use of two single-channel audio inputs simultaneously captured by means of different recording devices. As occurs in [24], these two single-channel inputs are defined as: (i) the internal recording that comes from the audio captured using a stethoscope in which both biomedical sounds from inside the human body and ambient noises can be listened; (ii) the external recording that comes from the audio captured using an external microphone in which only the ambient noise that surrounds the subject is captured. Specifically, our first contribution applies NMPCF from a multichannel point of view assuming that ambient noises can be modelled as repetitive sounds that can be simultaneously found in both single-channel inputs. In other words, we implicitly assume that the spectral patterns that characterize the ambient noises are repeated sound events contained in both the spectrograms from the internal and external recordings. Our second contribution proposes an incremental algorithm, based on the previous multichannel NMPCF, that refines the estimated biomedical spectrogram through a set of incremental stages by eliminating a high amount of ambient noise that was not extracted in the previous stage, especially in the case of high noise environments. In this work, a high noisy environment provides a Signal-to-Noise Ratio (SNR) lower than 0 dB.

The paper is structured as follows: Section 2 details the datasets, the state-of-the-art method for comparison and the proposed method. The metrics, setup and results are shown in Section 3. Finally, Section 4 presents the conclusions and future work.

2. Materials and Methods

2.1. Data collection

Due to the lack of publicly available databases consisting of biomedical sounds mixed with ambient noises to the best of our knowledge, we have created the database D_C. The database D_C is composed by the ambient noise database D_N and the biomedical database D_B in order to simulate auscultation recordings captured from a stethoscope.

The database D_N has been created taking into account a wide range of ambient noises collected from databases widely used in the field of sound source separation [32] and sound event detection [33, 34]. Most of these ambient noises have been classified as some of the most disturbing noises that can appear in the auscultation performed in the hospital room according to information provided by medical personnel from the Hospital of Jaen (Spain). For this reason, the database D_N is composed of five types of ambient noise in order to assess the denoising performance of the proposed method considering common indoor and outdoor ambient noises that typically surround a medical consultation room: ambulance siren [35, 36], baby crying [37], babble (people speaking) [38, 39], car (inside the vehicle) [40] and street (car passing by, car engine running, car idling, bus, truck, children yelling, people talking, workers on the street) [41, 42]. The database D_N consists of a total of 150 single-channel recordings of ambient noises, of which each type of noise consists of 30 recordings. Each recording has a duration of 5 seconds that has been obtained applying a pseudo-random process, based on the standard uniform distribution, to select a starting time followed by a 5 seconds interval.

The database D_B consists of a total of 150 single-channel biomedical recordings from public and private biomedical databases, specifically, 75 heart recordings [43, 44] (typically in the range 10Hz-320 Hz [45, 46]) and 75 lung recordings [47] (typically in the range 50Hz-2500 Hz [48, 49]). Highlight that ambient noises are not listened on each recording. Each recording has a duration of 5 seconds which has been obtained applying a pseudo-random process similarly as used in the database D_N. Each mixture recording belonging to the database D_C has been generated mixing each recording from the database D_B with a recording of

each type of noise randomly chosen from the database D_N. Indicate that the recordings of noise used for the mixtures with the heart recordings are the same as those used for the mixing with the lung recordings. For each mixture recording from D_C, the ambient noise used in the internal recording is the same noise used in the external recording. As a result, two databases are created from D_C, the optimization database D_O and the testing database D_T,

- The optimization database D_O is generated randomly selecting two-thirds of all mixtures recordings from the database D_C.
- The testing database D_T is generated using the remainder one-third mixtures recordings that are not used in the database D_O.

The set of recordings used in the optimization database D_O is not the same as that used in the testing database D_T in order to validate the denoising results. Moreover, several SNR have been applied in the mixing process to create the database D_C in order to evaluate high noisy environments. In this way, the databases $D_{T_{-20}}$ (SNR=-20 dB), $D_{T_{-15}}$ (SNR=-15 dB), $D_{T_{-10}}$ (SNR=-10 dB) and $D_{T_{-5}}$ (SNR=-5 dB) refer to the same database D_T but using different SNR between biomedical and ambient noise recordings. For example, a value SNR=-20 dB indicates that the power of the ambient noise is 100 times greater compared to the power of the biomedical sounds used in the audio mixture. Table 1 describes the characteristics of the data, according to the databases used in the experimental evaluation.

ID_1	ID_2	ID_3	ID_4	ID_5	ID_6	ID_7	ID_8	ID_9	ID_{10}	ID_{11}
D_B	75	75	-	-	-	-	-	150	-	750
D_N	-	-	30	30	30	30	30	-	150	750
D_O	50	50	20	20	20	20	20	100	100	2500
D_T	25	25	10	10	10	10	10	50	50	1250

Table 1: Characteristics of the data. ID_1: database identifier; ID_2: number of clean heart recordings; ID_3: number of clean lung recordings; ID_4: number of ambulance siren noise recordings; ID_5: number of babycrying noise recordings; ID_6: number of babble noise recordings; ID_7: number of car noise recordings; ID_8: number of street noise recordings; ID_9: total number of biomedical (clean heart and lung) recordings; ID_{10}: total number of ambient noise recordings; ID_{11}: temporal duration for all recordings in seconds.

2.2. Baseline method for comparison

One of the most referenced method, based on Multiband Spectral Subtraction (MSS) [24], for suppressing the ambient noise has been implemented to evaluate the performance of the proposed method. Note that the best configuration (algorithm B) of MSS has been used to perform a fair comparison. The reader can refer to [24] for more details.

2.3. Proposed method for ambient denoising

The main problem that physicians point out when performing the auscultation process in high noisy environments is that the biomedical sounds are severely overlapped with ambient noises so, part of the valuable clinical information contained in the sounds of interest is masked. The aim of the proposed method is to improve the quality of the biomedical sounds captured by a stethoscope in high noisy environments applying Non-negative Matrix Partial Co-Factorization (NMPCF) in a multichannel (two distinct single-channel signals) scenario (2C-NMPCF). The flowchart of the proposed method 2C-NMPCF is shown in Figure 2.

2.3.1. Time-Frequency representation

The internal signal (first single-channel) $x(t)$ represents the sounds captured by a digital stethoscope that is composed of two types of additive sound sources: (i) the biomedical sounds $s(t)$ from the subject; (ii) the ambient noises $n(t)$ surrounding the subject that are still heard inside the human body. We assume that $s(t)$ and $n(t)$ can be considered independent sound sources, that is, $x(t) = s(t) + n(t)$. Moreover, an

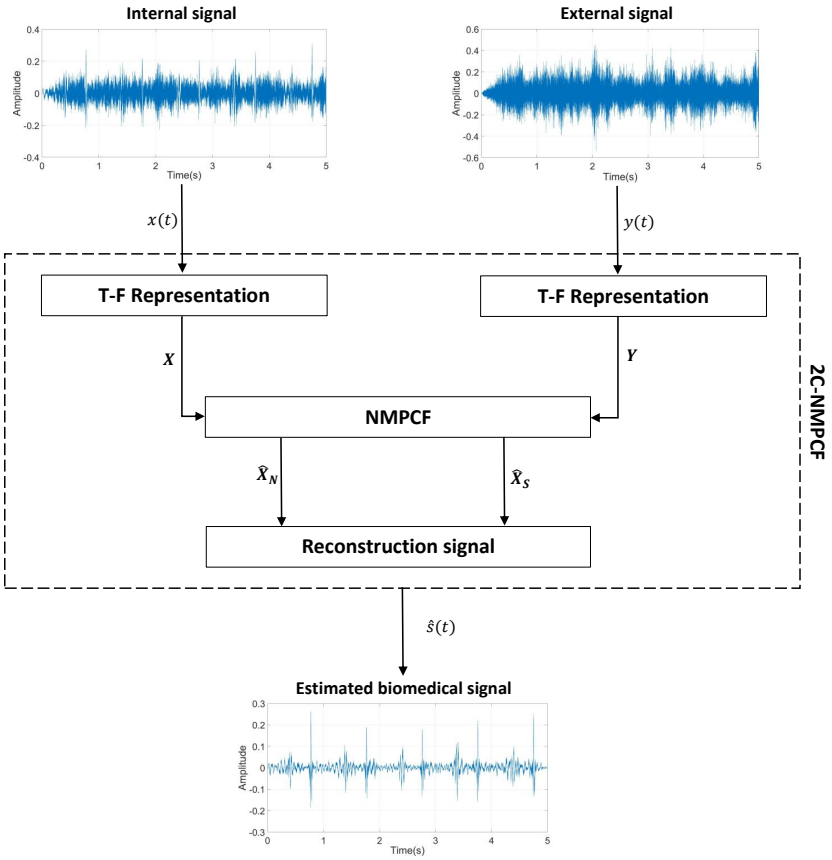

Figure 2: Flowchart of the proposed method 2C-NMPCF.

external microphone, located outside of the subject, captures all ambient noises that are represented by the external signal (second single-channel) $y(t)$.

The complex and magnitude spectrogram $\mathbf{X}_c \in \mathbb{C}_+^{F\times T}$, $\mathbf{X} \in \mathbb{R}_+^{F\times T}$ associated to the internal signal $x(t)$ and the magnitude spectrogram $\mathbf{Y} \in \mathbb{R}_+^{F\times T}$ associated to the external signal $y(t)$ are calculated using the Short-Time Fourier Transform (STFT) applying a Hamming window of size N with 50% overlap. Indicate that the complex values associated with $\mathbf{X}_c$, in which the phase information is included, are used later in the resynthesis process. The size and scale of the magnitude spectrograms depend on each input single-channel signal. Therefore, a normalization process is applied in order to ensure that the proposed method is independent of the size and scale of the input spectrograms. Thus, the normalized spectrograms $\overline{\mathbf{X}} \in \mathbb{R}_+^{F\times T}$ and $\overline{\mathbf{Y}} \in \mathbb{R}_+^{F\times T}$ are computed as follows,

$$\overline{\mathbf{Z}} = \frac{\mathbf{Z}}{\left(\frac{\sum_{f,t} Z_{f,t}}{FT}\right)} \tag{1}$$

where $\mathbf{Z} = \{\mathbf{X},\mathbf{Y}\}$ according to the input spectrogram. The variables F and T represent the number of frequency bins and the number of time frames. To avoid the complex nomenclature throughout the manuscript, the variables $\overline{\mathbf{X}}$ and $\overline{\mathbf{Y}}$ are hereinafter referred as $\mathbf{X}$ and $\mathbf{Y}$, respectively.

2.3.2. Multichannel Non-negative Matrix Partial Co-Factorization (2C-NMPCF)

The idea of the proposed method 2C-NMPCF is to enforce a joint matrix decomposition using multiple matrices $\mathbf{X}$, $\mathbf{Y}$ obtained from distinct single-channel spectrograms instead of the several excerpts of the same

single-channel spectrogram as occurs in the conventional NMPCF. The main contribution of the proposed method 2C-NMPCF is to exploit the spectral patterns that are shared in two distinct spectrograms since we assume that ambient noises can be modelled as repetitive sound events that can be simultaneously found in the spectrograms associated both the internal and external signal. This modeling allows to remove most of the ambient noises that are active in the internal signal improving the quality of the biomedical sounds from the auscultation process. The proposed method 2C-NMPCF is composed of two stages:

- Stage 1. This stage is applied to the internal signal $x(t)$. The input spectrogram $\mathbf{X}$ is decomposed into two separated or estimated spectrograms, the magnitude spectrogram only composed of biomedical sounds $\hat{\mathbf{X}}_S \in \mathbb{R}_+^{F\times T}$ and the magnitude spectrogram only composed of ambient noises $\hat{\mathbf{X}}_N \in \mathbb{R}_+^{F\times T}$. The factorization of each spectrogram depends on the estimated basis matrix $\mathbf{U} \in \mathbb{R}_+^{F\times K}$ (dictionary of spectral patterns) and the estimated activation matrix $\mathbf{V} \in \mathbb{R}_+^{K\times T}$ (temporal gains) as follows,

$$\mathbf{X} \approx \hat{\mathbf{X}} = \hat{\mathbf{X}}_N + \hat{\mathbf{X}}_S = \mathbf{UV} = \begin{bmatrix}\mathbf{U}_N & \mathbf{U}_S\end{bmatrix} \begin{bmatrix}\mathbf{V}_N \\ \mathbf{V}_S\end{bmatrix} = \mathbf{U}_N\mathbf{V}_N + \mathbf{U}_S\mathbf{V}_S \tag{2}$$

  where $\hat{\mathbf{X}} \in \mathbb{R}_+^{F\times T}$ is the estimated or reconstructed magnitude spectrogram of the first channel signal. $\mathbf{U}_N \in \mathbb{R}_+^{F\times K_N}$ and $\mathbf{V}_N \in \mathbb{R}_+^{K_N\times T}$ are the estimated basis and activations matrix of the ambient noises. The variables $\mathbf{U}_S \in \mathbb{R}_+^{F\times K_S}$ and $\mathbf{V}_S \in \mathbb{R}_+^{K_S\times T}$ are the estimated basis and activation matrix of the biomedical sounds. The parameter $K = K_N + K_S$ indicates the number of bases, being K_N the number of bases related to the ambient noises and K_S the number of bases related to the biomedical sounds. In this stage, the described decomposition model (see Equation (3)) does not obtain a parts-based objects reconstruction with physical meaning as occurs in real-world. Therefore, this stage cannot distinguish between spectral patterns belonging to biomedical sounds and ambient noises.

- Stage 2. This stage is applied to the external signal $y(t)$. We assume that the external signal is only composed of ambient noises, therefore the goal of this model is to reconstruct the external magnitude spectrogram $\mathbf{Y}$ by using the basis matrix $\mathbf{U}_N$ composed of the spectral patterns that characterize the ambient noises,

$$\mathbf{Y} \approx \hat{\mathbf{Y}} = \mathbf{U}_N\mathbf{H}_N \tag{3}$$

  where $\hat{\mathbf{Y}} \in \mathbb{R}_+^{F\times T}$ is the estimated or reconstructed magnitude spectrogram of the external signal. The variable $\mathbf{H}_N \in \mathbb{R}_+^{K_N\times T}$ is the estimated activations matrix of the ambient noises for the external signal. Note that $\mathbf{U}_N$ can be treated as the same matrix previously used in Equation (2).

Specifically, 2C-NMPCF extends the conventional NMPCF to a multichannel scenario sharing the frequency basis matrix $\mathbf{U}_N$ to factorize simultaneously two distinct single-channel magnitude spectrograms $\mathbf{X}$, $\mathbf{Y}$ as shown in Figure 3. The proposed method 2C-NMPCF allows to factorize jointly $\mathbf{X}$ and $\mathbf{Y}$ so the spectral patterns of the ambient noises, active in both spectrograms, are shared in the same dictionary $\mathbf{U}_N$ since we assume that ambient noises can be considered repetitive sounds that can be simultaneously active in both magnitude spectrograms $\mathbf{X}$ and $\mathbf{Y}$. Contrarily, the dictionary $\mathbf{U}_S$ represents the spectral patterns of the biomedical sounds that only can be found in the internal magnitude spectrogram $\mathbf{X}$.

2.3.3. Objective Function and Update rules

The objective function of the proposed method 2C-NMPCF that must be performed to minimize the residuals of the two previous models, see Equations (2)-(3), is detailed as follows,

$$\Gamma = D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}\right) + \lambda D_{KL}\left(\mathbf{Y}|\hat{\mathbf{Y}}\right) \tag{4}$$

where the parameter λ controls the relative importance between the internal and the external magnitude spectrogram. So, the contribution of the magnitude spectrogram $\mathbf{Y}$ is greater when the parameter λ increases. In this paper, the Kullback–Leibler divergence, see Equation (5), has been used to calculate the signal reconstruction error for the internal spectrogram $D_{KL}\left(\mathbf{X}|\hat{\mathbf{X}}\right)$ and the external spectrogram

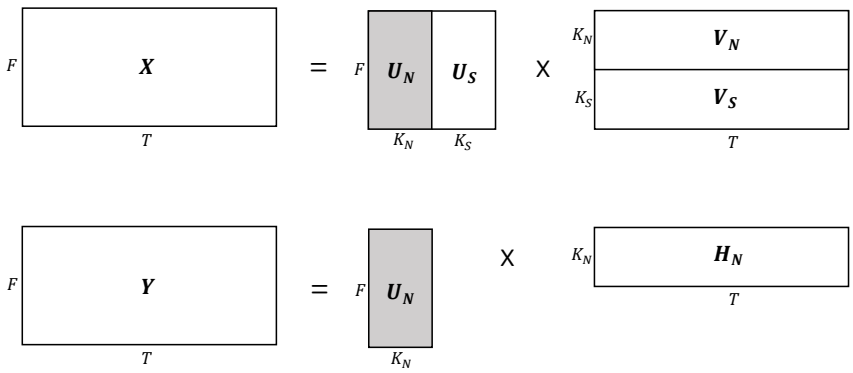

Figure 3: Matrix decomposition based on multichannel NMPCF (2C-NMPCF).

$D_{KL}\left(\mathbf{Y}|\hat{\mathbf{Y}}\right)$. The reason is because this cost function D_{KL} is non-increasing, ensuring the non-negativity of the estimated basis and activations matrices and moreover, several works have demonstrated promising results in the field of biomedical signal processing [7, 11, 50].

$$D_{KL}\left(\mathbf{Z}|\hat{\mathbf{Z}}\right) = \mathbf{Z}\log\frac{\mathbf{Z}}{\hat{\mathbf{Z}}} - \mathbf{Z} + \hat{\mathbf{Z}}, \quad \mathbf{Z} = \{\mathbf{X}, \mathbf{Y}\} \tag{5}$$

From Equation (4), the estimated basis matrices $\mathbf{U}_N$,$\mathbf{U}_S$ and activation matrices $\mathbf{V}_N$,$\mathbf{V}_S$,$\mathbf{H}_N$ can be obtained by applying a gradient descent algorithm based on multiplicative update rules. The multiplicative update rules to learn those matrices can be obtained by taking negative and positive terms of the partial derivative of the cost function Γ with respect to $\mathbf{U}_N$,$\mathbf{U}_S$, $\mathbf{V}_N$,$\mathbf{V}_S$ and $\mathbf{H}_N$, respectively,

$$\mathbf{U}_N \leftarrow \mathbf{U}_N \odot \frac{\left(\mathbf{X} \oslash \hat{\mathbf{X}}\right)(\mathbf{V}_N)^T + \lambda\left(\mathbf{Y} \oslash \hat{\mathbf{Y}}\right)(\mathbf{H}_N)^T}{(\mathbf{V}_N)^T + \lambda(\mathbf{H}_N)^T} \tag{6}$$

$$\mathbf{U}_S \leftarrow \mathbf{U}_S \odot \frac{\left(\mathbf{X} \oslash \hat{\mathbf{X}}\right)(\mathbf{V}_S)^T}{(\mathbf{V}_S)^T} \tag{7}$$

$$\mathbf{V}_N \leftarrow \mathbf{V}_N \odot \frac{(\mathbf{U}_N)^T\left(\mathbf{X} \oslash \hat{\mathbf{X}}\right)}{(\mathbf{U}_N)^T} \tag{8}$$

$$\mathbf{V}_S \leftarrow \mathbf{V}_S \odot \frac{(\mathbf{U}_S)^T\left(\mathbf{X} \oslash \hat{\mathbf{X}}\right)}{(\mathbf{U}_S)^T} \tag{9}$$

$$\mathbf{H}_N \leftarrow \mathbf{H}_N \odot \frac{(\mathbf{U}_N)^T\left(\mathbf{Y} \oslash \hat{\mathbf{Y}}\right)}{(\mathbf{U}_N)^T} \tag{10}$$

where $\odot$ is the element-wise multiplication, $\oslash$ is the element-wise division and $()^T$ is the transpose operator. The set of activation and basis matrices for both the internal and external magnitude spectrograms is obtained updating the rules detailed in Equations (6)-(10) using an iterative process until the algorithm converges or reaches a maximum number of iterations M.

Focusing on the separation process applied to the biomedical sounds and ambient noises present in the internal spectrogram, the estimated magnitude spectrograms $\hat{\mathbf{X}}_N$ and $\hat{\mathbf{X}}_S$ can be obtained from the estimated basis $\mathbf{U}_N$,$\mathbf{U}_S$ and activation matrices $\mathbf{V}_N$,$\mathbf{V}_S$ as follows:

$$\hat{\mathbf{X}}_N = \mathbf{U}_N\mathbf{V}_N \tag{11}$$

$$\hat{\mathbf{X}}_S = \mathbf{U}_S \mathbf{V}_S \tag{12}$$

In order to denormalize the estimated magnitude spectrograms of the internal spectrogram, the matrices $\hat{\mathbf{X}}_N$,$\hat{\mathbf{X}}_S$ are multiplied by the denominator of Equation (1) when $\mathbf{Z} = \mathbf{X}$. To guarantee a conservative strategy in the reconstruction process, the estimated biomedical signal $\hat{s}(t)$ (Equation (14)) is computed by the inverse overlap-add STFT of the element-wise multiplication between the complex spectrogram $\mathbf{X}_c$ and a Wiener mask [11, 7] that represents the relative energy contribution of the biomedical sounds to the energy of the internal signal $x(t)$. The estimated ambient noise signal $\hat{n}(t)$ (Equation (13)) is calculated in a similar way as explained above in Equation (14), but now taking into account that the Wiener mask explains the relative energy contribution of the ambient noise sounds to the energy of the internal signal $x(t)$.

$$\hat{n}(t) = IDFT\left(\mathbf{X}_c \odot \frac{\left|\hat{\mathbf{X}}_N\right|^2}{\left(\left|\hat{\mathbf{X}}_N\right|^2 + \left|\hat{\mathbf{X}}_S\right|^2\right)}\right) \tag{13}$$

$$\hat{s}(t) = IDFT\left(\mathbf{X}_c \odot \frac{\left|\hat{\mathbf{X}}_S\right|^2}{\left(\left|\hat{\mathbf{X}}_S\right|^2 + \left|\hat{\mathbf{X}}_N\right|^2\right)}\right) \tag{14}$$

The pseudo code of the proposed method 2C-NMPCF for the ambient denoising in auscultation is summarized in the Algorithm 1. Although the proposed method can return the estimated biomedical signal $\hat{s}(t)$ and the estimated ambient noise signal $\hat{n}(t)$, only the signal $\hat{s}(t)$ is required for evaluation purposes in this work.

Algorithm 1 Ambient denoising using 2C-NMPCF

Require: $y(t)$, $x(t)$, K_N, K_S, λ and M.

1: Compute the normalized magnitude spectrogram $\mathbf{X}$ of the internal signal $x(t)$ using Equation (1).
2: Compute the normalized magnitude spectrogram $\mathbf{Y}$ of the external signal $y(t)$ using Equation (1).
3: Initialize each activation and basis matrix $\mathbf{U}_N$,$\mathbf{U}_S$,$\mathbf{V}_N$,$\mathbf{V}_S$,$\mathbf{H}_N$ with random non-negative values.
4: Update each activation and basis matrix $\mathbf{U}_N$,$\mathbf{U}_S$,$\mathbf{V}_N$,$\mathbf{V}_S$,$\mathbf{H}_N$ using Equations (6)-(10) for the predefined number of iterations M.
5: Compute the estimated magnitude spectrograms $\hat{\mathbf{X}}_N$ using Equation (11).
6: Compute the estimated magnitude spectrograms $\hat{\mathbf{X}}_S$ using Equation (12).
7: Denormalize the estimated magnitude spectrograms $\hat{\mathbf{X}}_N$, $\hat{\mathbf{X}}_S$ multiplying by a factor equal to the denominator of Equation (1) when $\mathbf{Z} = \mathbf{X}$.
8: Synthesize the estimated ambient noise $\hat{n}(t)$ using the Equation (13).
9: Synthesize the estimated biomedical signal $\hat{s}(t)$ using the Equation (14).
return $\hat{s}(t)$

2.3.4. Improving the sound quality of biomedical signals by means of an incremental algorithm based on 2C-NMPCF

The main limitation of the proposal 2C-NMPCF is related to the objective function (see Equation (4)) used to minimize the residuals of the ambient noise. Specifically, the iterative process based on the multiplicative update rules that obtains the denoised biomedical basis and activation matrices is applied until the convergence of 2C-NMPCF into a local minimum after M iterations. For this reason, 2C-NMPCF by itself is not able to extract all spectral patterns associated to ambient noises. To overcome the previous limitation, we propose an incremental algorithm that runs 2C-NMPCF more than once improving the estimated biomedical signal $\hat{s}_i(t)$ obtained in the incremental iteration i by removing additional spectral

content associated to ambient noise that 2C-NMPCF was not able to remove in the previous incremental iteration $i-1$. In general considering the iteration i, the internal signal $x_{i+1}(t)$ of the next incremental iteration $i+1$ is the estimated biomedical signal $\hat{s}_i(t)$ obtained in the current incremental iteration i, that is, $x_{i+1}(t) = \hat{s}_i(t)$. Note that in the first incremental iteration $i = 1$, $x_1(t)$=$x(t)$. However, the external signal $y(t)$ is fixed for all incremental iterations since we assume that the signal $y(t)$ is only composed by ambient noises (no biomedical sounds are active). The flowchart of the proposed incremental algorithm based on 2C-NMPCF is shown in Figure 4.

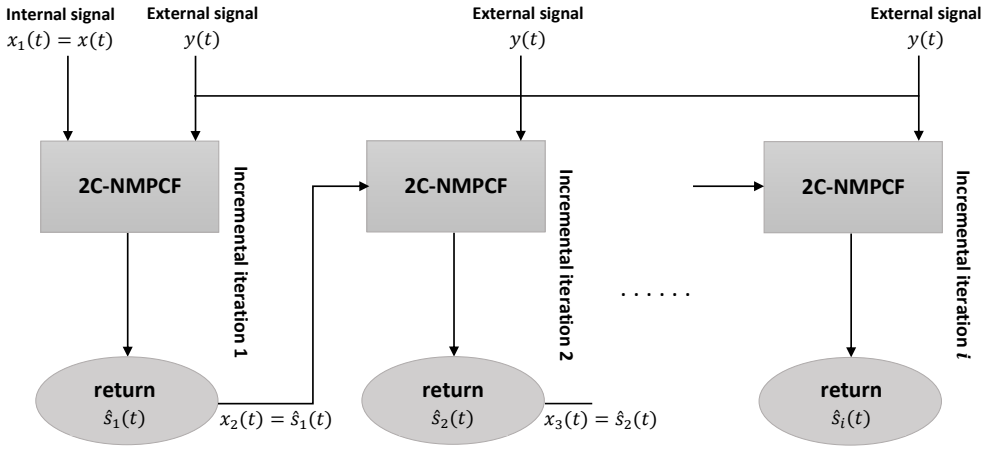

Figure 4: Flowchart of the proposed incremental algorithm based on 2C-NMPCF.

We assume the following assumptions in order to justify our incremental proposal based on 2C-NMPCF: (i) the objective function Γ would converge into a better local minimum at each incremental iteration since it would find remainder spectral patterns of ambient noise that have not been extracted in the previous iteration $i-1$ but they are still being repeated in both the internal $\mathbf{X}_i$ and the external $\mathbf{Y}$ magnitude spectrograms in the current incremental iteration i; (ii) 2C-NMPCF will remove most of the ambient noise while preserving the content of the biomedical sounds until an optimal number of incremental iterations $i = i_o$. From this optimal iteration $i = i_o$, our proposal can continue to eliminate hidden patterns of ambient noise that are still active at the expense of eliminating also spectral content related to biomedical signals. Summarizing, this incremental approach attempts to maintain most of the biomedical content $\hat{s}_{i_o}(t)$ removing most of the ambient noise through the incremental iterations. An illustrative example of the performance of the proposed incremental approach is shown in Figure 5.

3. Evaluation

3.1. Metrics

To evaluate the quality of the biomedical signals estimated by the proposed method, the BSS EVAL toolbox [51, 52] has been used because it proposes a set of metrics, widely applied in the field of sound source separation [11, 7] and background noise removal [53], that quantify the quality of the sound separation between the original biomedical signal and its estimation. Two metrics, measured in dB, are used as occurs in [54, 55]: (i) Source to Distortion ratio (SDR) measures the overall quality of the estimated biomedical signal; and (ii) Source to Interference ratio (SIR) measures the presence of ambient noise in the estimated biomedical signal. Higher values of these ratios indicate better sound quality of the estimated biomedical signal.

In this paper, the optimization and testing results have been obtained calculating both SDR and SIR median values [56]. These results do not show the absolute SDR and SIR values obtained from the estimated

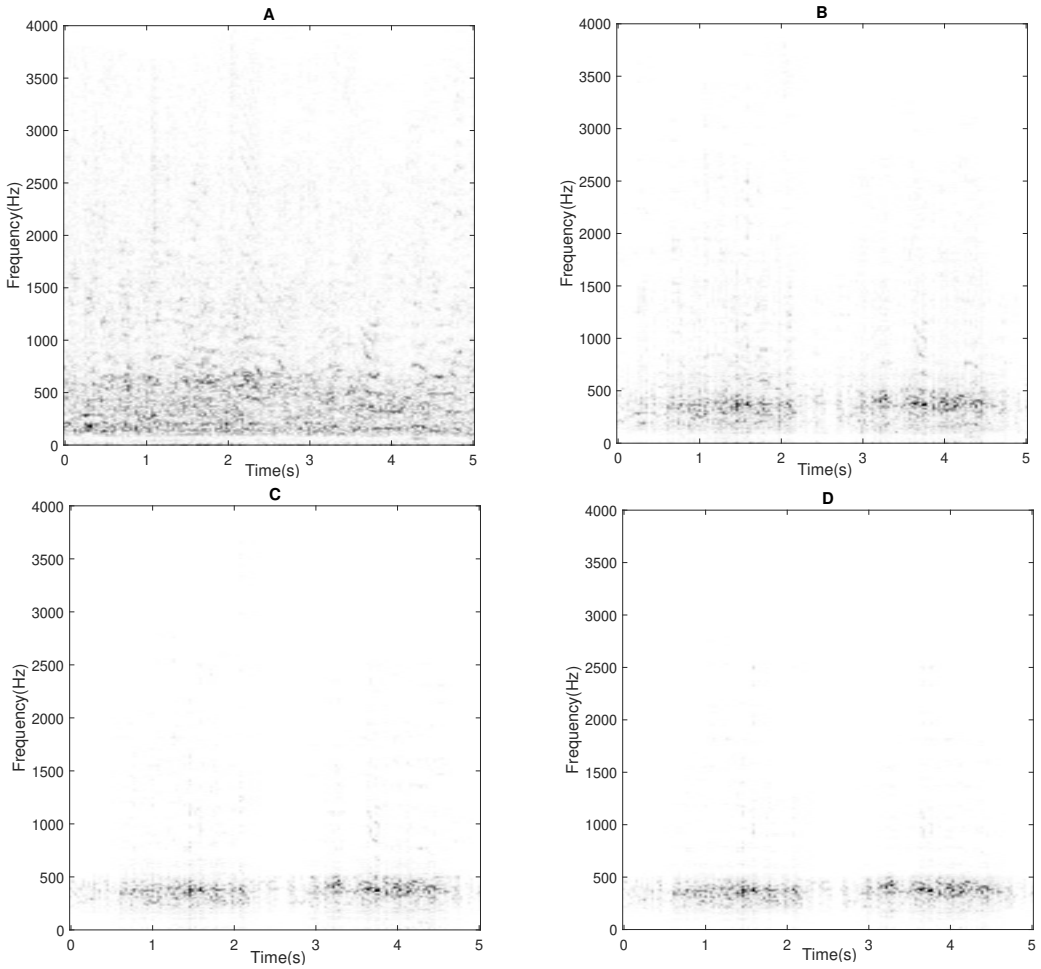

Figure 5: **A**) Magnitude spectrogram $\mathbf{X}$ of the internal signal previously shown in Fig. 1 (Right); Some of the estimated biomedical magnitude spectrograms $\hat{\mathbf{X}}_S$ provided by the incremental algorithm based on 2C-NMPCF through the incremental iterations i: **B**) $i = 1$, **C**) $i = 2$ and **D**) $i = 3$. Here, it can be observed that the estimated biomedical spectrogram is refined after each incremental iteration by removing spurious ambient noise that are still active in the previous incremental iteration maintaining most of the biomedical spectral content.

biomedical signal $\hat{s}_i(t)$ but the SDR and SIR improvement comparing $\hat{s}_i(t)$ and the original internal signal $x(t)$.

As occurs in [24], two metrics related to speech intelligibility (SI) are added to the objective assessment of the proposed method, which have been previously used in the field of ambient noise suppression in lung auscultation [24, 25]: (i) Normalized-Covariance Measure (NCM); and (ii) Coherence Speech Intelligibility Index (CSII). Specifically, a three-level CSII approach is used by dividing the signal into three amplitude regions: low (CSII_{low}), mid (CSII_{mid}) and high (CSII_{high}). In this work, each metric was computed between the original biomedical signal $s(t)$ and the estimated biomedical signal $\hat{s}(t)$. Note that higher values of these metrics indicate better sound quality of the estimated biomedical signal. Finally, more details of these metrics can be found by the reader in [24, 25, 57].

3.2. Setup

Because most of the spectral content both the biomedical signals [45, 46, 48, 49] and the ambient noise [23, 58] is concentrated below 4 KHz, in this work, a sampling rate equals $f_s = 8$ KHz has been used as occurs in [24].

A preliminary study showed that the following parameters for time-frequency representation provide the best trade-off between the separation performance and the computational cost: a Hamming window with $N = 512$ samples length (64ms) and 50% overlap; and a discrete Fourier transform using 2N points similarly as in [11, 50]. Furthermore, the convergence of the proposed method was empirically observed after 50 iterations, so the parameter M is fixed to $M = 50$.

Finally, note that the performance of the proposed method depends on the initial values with which the basis matrices $\mathbf{U}_S$, $\mathbf{U}_N$ and the activation matrices $\mathbf{V}_S$, $\mathbf{V}_N$, $\mathbf{H}_N$ have been initialised. Although the obtained results are not dispersed and keep the same behavior, in order to overcome this issue, we have run the proposed method three times for each mixture and the results shown in this paper have been calculated using the median values as previously mentioned.

3.3. Results

In this section, experimental results related to the optimization and testing are detailed.

3.3.1. Optimization results

Several parameters must be fitted to obtain the best performance of the proposed method in the removal of ambient noise. Four parameters have been evaluated using the database D_O: (i) The number of biomedical bases K_S used to characterise the spectral content of the biomedical signal $s(t)$, specifically, $K_S \in [16, 32, 64, 128, 256]$; (ii) The number of ambient noise bases K_N used to characterise the spectral content of the ambient noise $n(t)$, specifically, $K_N \in [16, 32, 64, 128, 256]$; (iii) The value λ to balance the importance of the internal $\mathbf{X}$ and external $\mathbf{Y}$ magnitude spectrograms in the co-factorization process. In this case, $\lambda \in [0.01, 0.1, 1, 10, 25, 50, 100, 250, 500, 1000]$; (iv) The number of incremental iterations i applied to 2C-NMPCF.

The optimization process is composed of two steps:

1. Step I. Optimize the three parameters K_S, K_N, λ in order to reach the greater median of the SDR improvement when applying 2C-NMPCF considering all the types of ambient noises and SNR previously mentioned.
2. Step II. Once the parameters of 2C-NMPCF have been optimized, it must be obtained the optimal number of incremental iterations $i = i_o$ to achieve the best performance of the proposed method (see Figure 4).

Figure 6 shows the median of the SDR improvement analyzing the search space derived from the parameters λ, K_S and K_N. Results indicate that the proposed method provides the best denoising performance, by means of the maximum median value of the SDR improvement, using the optimal parameters K_N=256 and K_S=16. These optimal values demonstrate that ambient noise requires a greater number of spectral patterns compared to biomedical sounds due to their greater spectral diversity. The analysis of different lung and heart signals indicates that the spectral modeling of these biomedical sounds is simpler and therefore needs a smaller dictionary of bases since lung sounds could be factorize by a low-rank decomposition using broadband spectral patterns that show temporal and spectral smoothness. However, heart sounds could be modeled as low-frequency pulses located in regular intervals in time.

Figure 7 shows the median of the SDR and SIR improvement using the previous optimal values of the parameters K_S and K_N (that is, K_S= 16 and K_N=256). It can be confirmed that giving importance to the sharing of spectral bases in the joint factorization process finds a better local minimum in the factorization process, since ambient noise clearly reveals its simultaneous presence both in the internal and external signal spectrogram. Results report a significant and stable SDR and SIR improvement equals 14 dB and 19.5 dB using $\lambda \geq 10$. For this reason, we have chosen the optimal parameter λ=10.

Figure 8 depicts the optimal number of incremental iterations i of the proposed method using the previous optimal parameters K_S, K_N and λ. It is observed that both SDR and SIR improvement increases sharply when applying 2C-NMPCF in the second incremental iteration (i =2), higher increase for SIR compared to SDR. In the third incremental iteration (i =3), the SDR improvement increases slightly and then starts to decrease gradually while the SIR improvement continues to grow. Experimental results indicate that

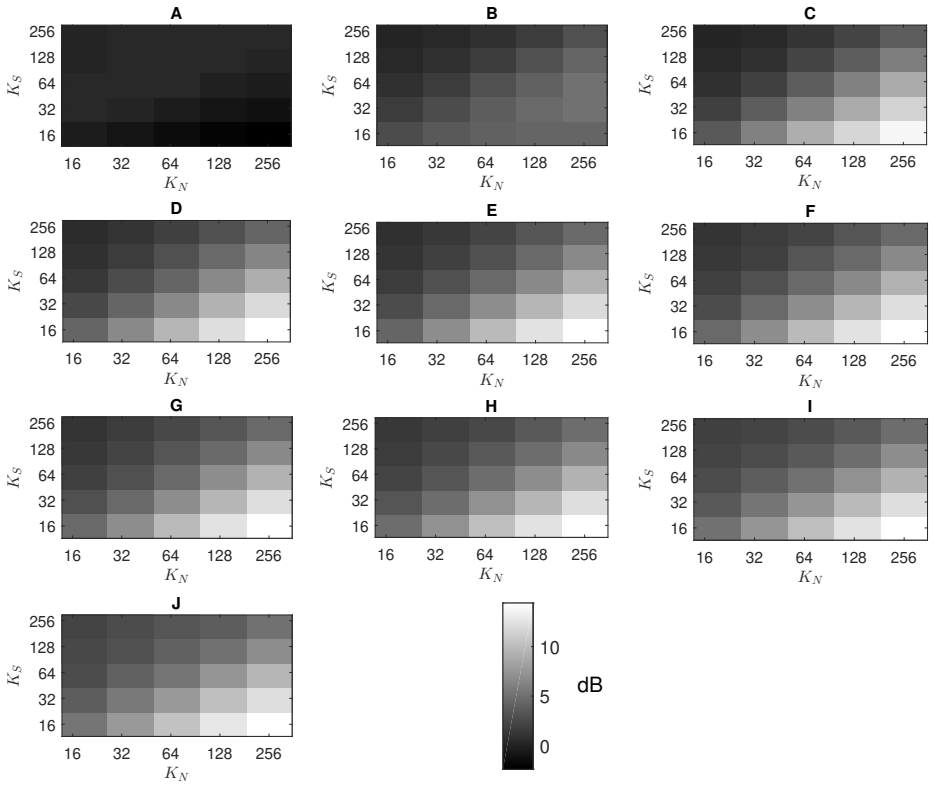

Figure 6: Median values of the SDR improvement (dB) evaluating D_O for the following λ values: λ =0.01 (A), λ =0.1 (B), λ =1 (C), λ =10 (D), λ =25 (E), λ =50 (F), λ =100 (G), λ =250 (H), λ =500 (I) and λ =1000 (J).

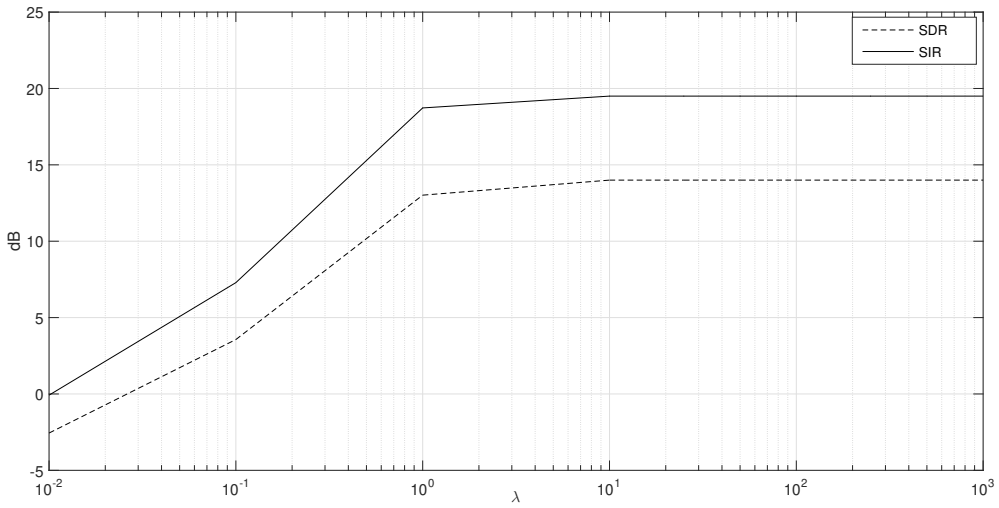

Figure 7: Median values of the SDR (dashed line) and SIR (solid line) improvement of the proposal algorithm evaluating D_O, keeping fixed K_S=16 and K_N=256 varying the parameter λ.

ambient noise continues to be suppressed at the expense of starting to remove biomedical spectral content when $i > 3$. For this reason, the optimal number of incremental iterations has been set at $i_o = 3$ with the aim of providing the greatest suppression of ambient noise while maintaining most of the biomedical content, being i_o the iteration in which the maximum SDR improvement is obtained.

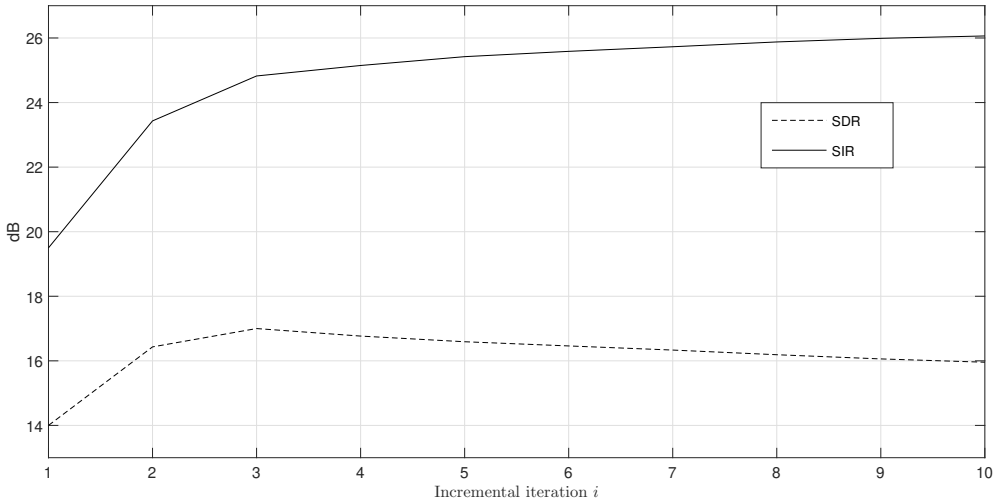

Figure 8: Median values of the SDR (dashed line) and SIR (solid line) improvement with the number of incremental iterations i evaluating D_O, keeping fixed (K_S, K_N, λ)=(16, 256, 10).

3.3.2. Objective results

Figure 9 shows SDR and SIR improvement comparing the performance of the removal ambient noise evaluating the database D_T for the proposed method (2C-NMPCF) and the baseline method MSS [24]. Hereinafter, the SDR and SIR improvement of the proposed method are represented by SDR_P and SIR_P while the SDR and SIR improvements of the method MSS are represented by SDR_M and SIR_M. Each box represents 250 data points, one for each recording of the database D_T. The lower and upper lines of each box show the 25th and 75th percentiles. The line in the middle of each box represents the median value. The diamond in the center of each box represents the average value. The lines extending above and below each box show the extent of the rest of the samples, excluding outliers. Outliers are defined as points that are over 1.5 times the interquartile range from the sample median, which are shown as crosses.

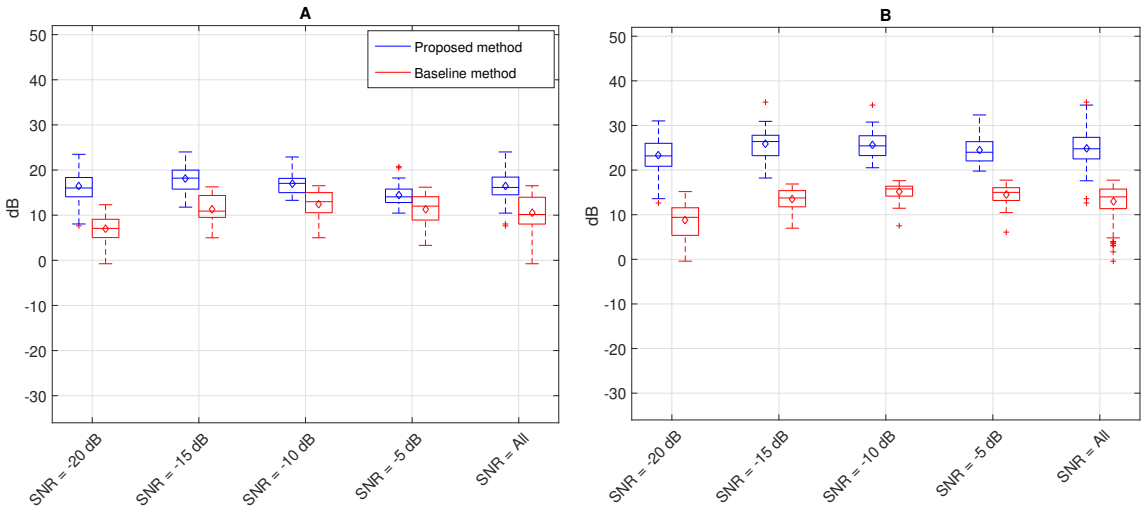

Figure 9: SDR (A) and SIR (B) improvement considering all the noises (Ambulance Siren, Baby Crying, Babble, Car and Street) for each SNR from database D_T. Each box represents the denoising performance results for each method evaluated. The color of the legend, associated to each evaluated method shown in the subfigure 9.A, refers to the same methods for all subfigures.

Figure 9A shows that the proposed method significantly outperforms the performance of the baseline method MSS for each SNR, which shows that the worse the SNR, the better the denoising behaviour shown by the proposed method as the biomedical sounds are masked to a greater extent and therefore the ambient noise existing in both spectrograms of the NMPCF becomes more prominent, achieving approximately 7 dB of average advantage of SDR_P over SDR_M taking into account all SNRs. Figure 9B reports a remarkable SIR_P results, specifically, an approximate average advantage of SIR_P over SIR_M of 11 dB. This fact indicates an interesting advantage of the proposed method because its denoising performance hardly depends on the SNR of the input signal, keeping a near-constant SIR_P around 25 dB. Moreover, SDR_P and SIR_P results suggest that our incremental approach, assuming ambient noise as repetitive sound event found in both the internal and external spectrograms, is a more reliable and robust feature to eliminate ambient noise compared to the penalty proposed by MSS that is based on the distortion minimization of the biomedical sounds in low frequencies and penalize noise occurrences with strong energy at high spectral bands.

Figure 10 shows a detailed analysis of the denoising performance of the proposed method considering each type of ambient noise evaluated previously in Figure 9. Each box represents 50 data points, one for each recording of the database D_T.

- Ambulance siren noise: Figures 10A and 10B show that the proposed method outperforms MSS in terms of SDR, obtaining an average improvement about 20 dB taking into account all SNRs. This promising denoising performance of the proposed method remains in terms of SIR_P whose average improvement is near-constant around 30 dB.

- Baby crying noise: Figures 10C and 10D show that the proposed method significantly improves the SDR and SIR denoising performance compared with MSS, showing an interesting advantage particularly in high noisy environments. The best denoising results are obtained from the suppression both the ambulance siren and baby crying noise. Results seem to suggest that the strong harmonic structure contained into these two types of ambient noise facilitates the sound separation between noise and biomedical sounds since the more dissimilar the spectral patterns of the noise and the biomedical signal, the better the noise suppression performance of the proposed method.

- Babble noise: Figures 10E and 10F indicate that there is no significant difference between the performance of the proposed method and MSS but the proposed method obtains a slightly worse performance in some cases, specifically, in SNR=-20 dB. Experimental results suggest that the spectrum of the Babble noise and the biomedical signal is more similar compared to the above ambient noises (ambulance siren and baby crying noise) so, the greater are the spectral differences between the target sounds and the noise, the better is the performance of the proposed method since the repetitive behaviour of the ambient noise in the co-factorization process is more easy to suppress.

- Car noise: Figures 10G and 10H indicate that the metric SDR_P exceeds the SDR_M for small SNR values but this improvement decreases as the SNR increases. However, the SIR_P is always significantly higher compared to SIR_M regardless of the SNR, achieving an average advantage over MSS from 16 dB (SNR=-20 dB) to 14 dB (SNR=-5 dB).

- Street noise: Figures 10I and 10J show that the SDR_P is competitive compared to SDR_M. Specifically, SIR_P is clearly better than SIR_M keeping their values near-constant, approximately equals 20 dB, with no dependence on the SNR.

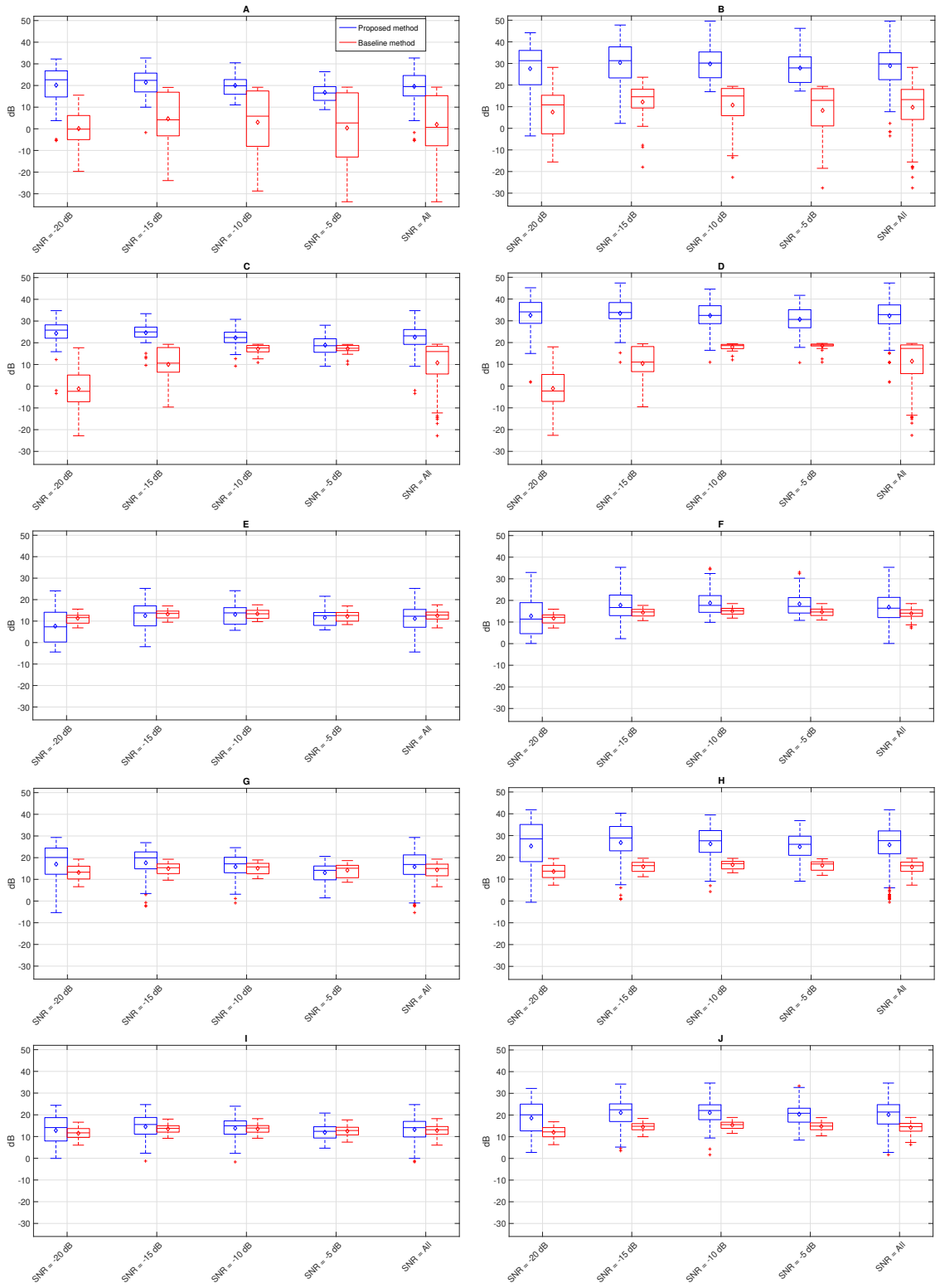

Figure 10: SDR and SIR improvement results provide by the proposed method and the baseline method (MSS) evaluating D_T and each type of ambient noise along SNR: Ambulance Siren (A and B), Baby crying (C and D), Babble (E and F), Car (G and H) and Street (I and J). Each box represents the denoising performance results for each method evaluated. The color of the legend, associated to each evaluated method shown in the subfigure 10.A, refers to the same methods for all subfigures.

To complete our study, it has been considered the effect of inserting a time delay between the internal and external spectrogram. The purpose of the delay is to simulate the time processing that takes the digital stethoscope to apply filtering, artifacts removal and other signal processing operations [59]. Figure 11 shows how the delay affects to the computation of the median value of the metrics SDR_P, SIR_P, SDR_M and SIR_M considering all the previous ambient noises and SNRs. Both SDR_P and SIR_P results confirm that the most remarkable advantage of the proposed method is its robustness with the variation of the delay. Considering the median SDR_P and SIR_P, it can be seen that the proposed method shows a stable behavior in relation to the delay variation between the internal and external signals used in the co-factorization, in contrast to the high dependence of the MSS method on the delay. Specifically, the proposed method outperforms MSS by about 4 dB in terms of SDR_P and 9 dB in terms of SIR_P when there is no delay. However, the denoising performance between the proposed method and MSS is accentuated when the delay is active. Specifically, SDR_M and SIR_M results decrease as the delay increases, reaching a improvement of 13 dB in terms of SDR_P and 21 dB in terms of SIR_P comparing the performance of both methods evaluated applying a delay equals to 25 milliseconds.

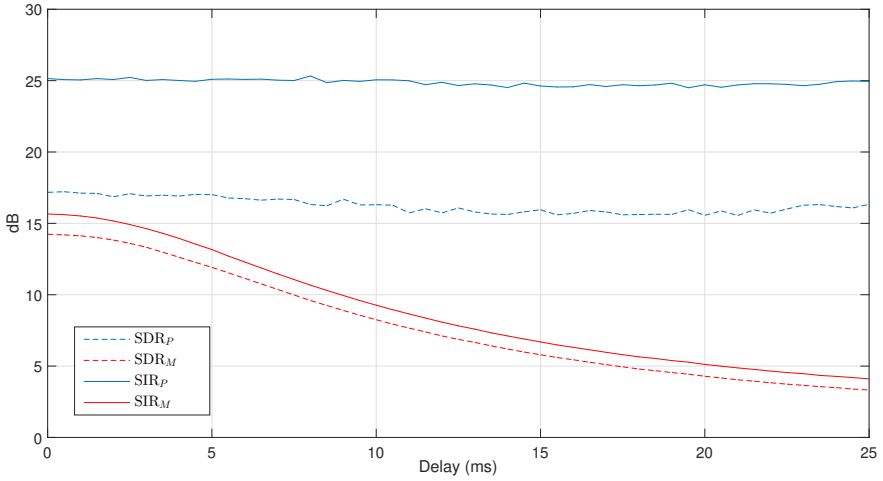

Figure 11: Median values of the SDR and SIR improvement analyzing all the noises (Ambulance Siren, Baby Crying, Babble, Car and Street) and SNRs (-20 dB, -15 dB, -10 dB, -5 dB) varying the delay between the internal and external spectrograms in the database D_T.

Figure 12 shows NCM, CSII_{low}, CSII_{med} and CSII_{high} results comparing the performance of the ambient noise removal evaluating the database D_T for the proposed method and MSS [24]. Note that the label "Original value" refers to the result computed between the original biomedical signal $s(t)$ and the original internal signal $x(t)$, that is, the original value indicates the sound quality of the internal signal with no improvement applied. Specifically, each box represents 250 data points, one for each recording of the database D_T. Figure 12 reports that ambient denoising results obtained by the proposed method are competitive compared to MSS evaluating all SNR scenarios. Moreover, these results obtained by both the proposed method and MSS are higher than the original values, which implies that both methods improve the acoustic quality of biomedical sounds by eliminating ambient noise that typically hinders clinical examination. Finally, it can be seen how the above results fall as the SNR decreases, similar to what happens in the real world because it is more difficult to hear biomedical sounds in those acoustic scenarios where biomedical sounds are barely audible due to high ambient noise levels.

Figure 13 shows the computational cost of the proposed method (2C-NMPCF) and the baseline method (MSS) which has been computed using Matlab 2020a on a PC with Intel Core i7-7700HQ CPU of 2.8 GHz and 16 GB of RAM. It can be observed that the computational cost of both methods increases with the size of the analyzed signal and both costs are less compared to the temporal duration of the input signal. The computational cost of the proposed method to eliminate ambient noise is approximately 60% lower compared to that obtained by MSS when the slopes of the curves associated with each method are analyzed.

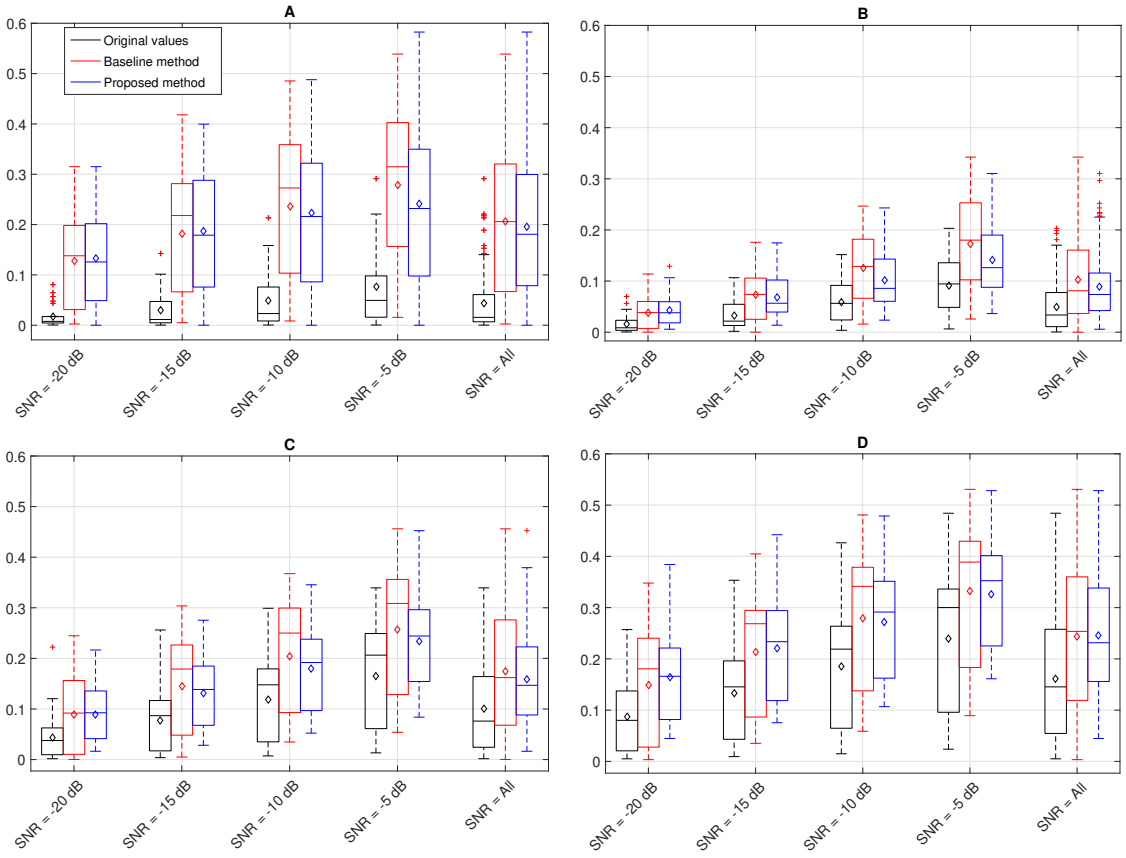

Figure 12: NCM (A), CSII_{low} (B), CSII_{med} (C) and CSII_{high} (D) results considering all the noises (Ambulance Siren, Baby Crying, Babble, Car and Street) for each SNR from database D_T. Each box represents the denoising performance results for each method evaluated. The color of the legend, associated to each evaluated method shown in the subfigure 12.A, refers to the same methods for all subfigures.

Specifically, for the case in which the duration of the input signal is 35 seconds, the proposed method used 5 seconds to provide the output while MSS took approximately 12 seconds. In general terms, the processing factor P_F defined as the ratio between the computational cost and the temporal duration of the input signal is lower for the proposed method.

4. Conclusions and Future Work

In this work, we propose an incremental algorithm based on multichannel non-negative matrix partial co-factorization (NMPCF) for ambient denoising in auscultation focusing on high noisy environment with a Signal-to-Noise Ratio (SNR) lower than 0 dB. The first contribution applies NMPCF from a multichannel point of view assuming that ambient noise can be modelled as repetitive sound events found in two single-channel audio inputs simultaneously captured by means of different recording devices. The second contribution proposes an incremental algorithm, based on the previous multichannel NMPCF, that refines the estimated biomedical spectrogram through a set of incremental stages by eliminating most of the ambient noises that was not removed in the previous stage.

The optimization process indicates that the best performance of the proposed method is obtained using a higher number of noise bases compared to the number of biomedical bases. It suggests that the energy distribution of the types of ambient noises analyzed is more complex compared to biomedical sounds due to the greater spectral diversity shown by the time-frequency structures of such noises.

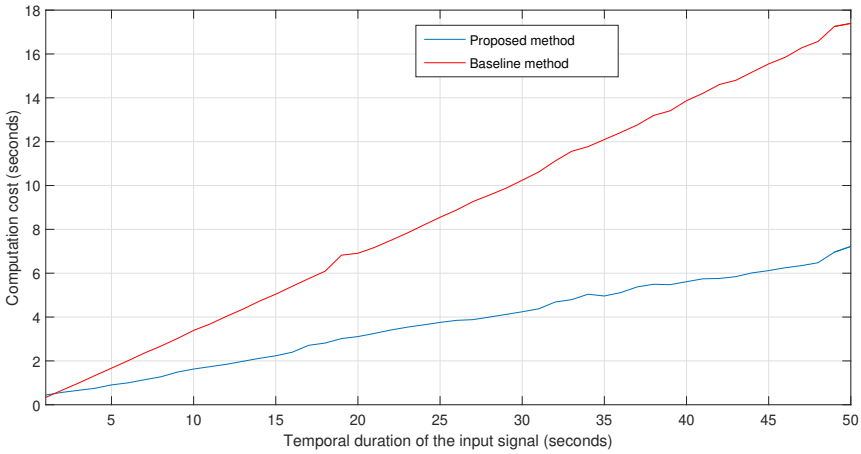

Figure 13: Comparison between the computational cost of the proposed method 2C-NMPCF and the baseline method MSS depending on the temporal duration of the input signal.

The most relevant conclusions from the experimental results indicate the following: (i) the proposed method provides the best overall ambient denoising performance compared to MSS considering all SNR scenarios evaluated; (ii) the proposed method shows greater robustness in those sound scenarios in which ambient noise and biomedical sounds exhibit distinct spectral behaviour. Specifically, the proposed method obtains the best performance removing the ambient noise composed of ambulance siren or baby crying since the time-frequency characteristics of these sounds are very dissimilar compared to the evaluated biomedical sounds; and (iii) the most remarkable advantage of the proposed method is its robustness with the variation of the delay, scenario simulated to replicate the process of recording the internal signal by a digital stethoscope.

Future work will focus mainly on developing algorithms applied to the characterization of some of the most acoustically disturbing ambient noises active in certain clinical emergency situations, such as noise inside a helicopter when urgent monitoring is required.

Funding

This work was supported by the Programa Operativo FEDER Andalucia 2014-2020 under project with reference 1257914 and the Ministry of Economy, Knowledge and University, Junta de Andalucia under Project P18-RT-1994.

Acknowledgment

The authors would like to thank Dr. Dinko Oletic and Dr. Vedran Bilas for sharing their lung recordings. The authors would like to thank the pulmonologist Gerardo Perez Chica from the University Hospital of Jaen (Spain) for his assistance related to ambient noises. Finally, we would like to thank the anonymous reviewers for their helpful and constructive comments that greatly contributed to improve the final version of the paper.

References

[1] A. K. Abbas, R. Bassam, Phonocardiography signal processing, Synthesis Lectures on Biomedical Engineering 4 (1) (2009) 1–194.

[2] M. Sarkar, I. Madabhavi, N. Niranjan, M. Dogra, Auscultation of the respiratory system, Annals of Thoracic Medicine 10 (3) (2015) 158–168.

[3] S. Taplidou, L. Hadjileontiadis, Wheeze detection based on time-frequency analysis of breath sounds, Computers in biology and medicine 37 (8) (2007) 1073–1083.

[4] D. Kumar, P. Carvalho, M. Antunes, J. Henriques, Noise detection during heart sound recording, in: 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2009, pp. 3119–3123.
[5] T. Tsalaile, S. Sanei, Separation of heart sound signal from lung sound signal by adaptive line enhancement, in: 15th European Signal Processing Conference, IEEE, 2007, pp. 1231–1235.
[6] C. Lin, E. Hasting, Blind source separation of heart and lung sounds based on nonnegative matrix factorization, in: International symposium on intelligent signal processing and communication systems (ISPACS), IEEE, 2013, pp. 731–736.
[7] F. Canadas-Quesada, N. Ruiz-Reyes, J. Carabias-Orti, P. Vera-Candeas, J. Fuertes-Garcia, A non-negative matrix factorization approach based on spectro-temporal clustering to extract heart sounds, Applied Acoustics 125 (2017) 7–19.
[8] G. Serbes, C. Sakar, Y. Kahya, N. Aydin, Pulmonary crackle detection using time–frequency and time–scale analysis, Digital Signal Processing 23 (3) (2013) 1012–1021.
[9] M. Zivanovic, M. Gonzalez-Izal, Quasi-periodic modeling for heart sound localization and suppression in lung sounds, Biomedical Signal Processing Control 8 (2013) 586–595.
[10] V. Varghees, K. Ramachandran, A novel heart sound activity detection framework for automated heart sound analysis, Biomed. Signal Process. Control. 13 (2014) 174–188.
[11] J. Torre-Cruz, F. Canadas-Quesada, J. Carabias-Orti, P. Vera-Candeas, N. Ruiz-Reyes, A novel wheezing detection approach based on constrained non-negative matrix factorization, Applied Acoustics 148 (2019) 276–288.
[12] F. Jin, F. Sattar, D. Goh, New approaches for spectro-temporal feature extraction with applications to respiratory sound classification, Neurocomputing 123 (2014) 362–271.
[13] S. Raj, K. Ray, O. Shankar, Cardiac arrhythmia beat classification using dost and pso tuned svm, Computer methods and programs in biomedicine 136 (2016) 163–77.
[14] P. Li, Y. Wang, J. He, L. Wang, Y. Tian, T.-s. Zhou, T. Li, J.-s. Li, High-performance personalized heartbeat classification model for long-term ecg signal, IEEE Transactions on Biomedical Engineering 64 (1) (2016) 78–86.
[15] D. Bardou, K. Zhang, S. Ahmad, Lung sounds classification using convolutional neural networks, Artificial Intelligence in Medicine 88 (2018) 58–69.
[16] R. X. A. Pramono, S. A. Imtiaz, E. Rodriguez-Villegas, Evaluation of features for classification of wheezes and normal respiratory sounds., PloS one 14 (3) (2019) e0213659–e0213659.
[17] A. Suzuki, C. Sumi, K. Nakayama, M. Mori, Real-time adaptive cancelling of ambient noise in lung sound measurement, Medical and Biological Engineering and Computing 33 (5) (1995) 704–708.
[18] S. B. Patel, T. F. Callahan, M. G. Callahan, J. T. Jones, G. P. Graber, K. S. Foster, K. Glifort, G. R. Wodicka, An adaptive noise reduction stethoscope for auscultation in high noise environments, The Journal of the Acoustical Society of America 103 (5) (1998) 2483–2491.
[19] J. S. Fleeter, G. R. Wodicka, Auscultation of heart and lung sounds in high-noise environments using adaptive filters, The Journal of the Acoustical Society of America 104 (3) (1998) 1781–1781.
[20] D. Della Giustina, M. Riva, F. Belloni, M. Malcangi, Embedding a multichannel environmental noise cancellation algorithm into an electronic stethoscope, International Journal of Circuits/System and Signal Processing (2) (2011).
[21] G. Nelson, R. Rajamani, A. Erdman, Noise control challenges for auscultation on medical evacuation helicopters, Applied Acoustics 80 (2014) 68–78.
[22] N. W. Evans, J. S. Mason, W.-M. Liu, B. Fauve, An assessment on the fundamental limitations of spectral subtraction, in: 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, Vol. 1, IEEE, 2006, pp. I–I.
[23] G. Chang, Y. Lai, Performance evaluation and enhancement of lung sound recognition system in two real noisy environments, Computer methods and programs in biomedicine 97 (2) (2015) 141–150.
[24] D. Emmanouilidou, E. McCollum, D. Park, M. Elhilali, Adaptive noise suppression of pediatric lung auscultations with real applications to noisy clinical settings in developing countries, IEEE Transactions on Biomedical Engineering 62 (9) (2015) 2279–2288.
[25] D. Emmanouilidou, E. D. McCollum, D. E. Park, M. Elhilali, Computerized lung sound screening for pediatric auscultation in noisy field environments, IEEE Transactions on Biomedical Engineering 65 (7) (2018) 1564–1574.
[26] Y. Hu, G. Liu, Separation of singing voice using nonnegative matrix partial co-factorization for singer identification, IEEE/ACM Transactions on Audio, Speech, and Language Processing 23 (4) (2015) 643–653.
[27] J. Yoo, M. Kim, K. Kang, S. Choi, Nonnegative matrix partial co-factorization for drum source separation, in: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2010, pp. 1942–1945.
[28] M. Kim, J. Yoo, K. Kang, S. Choi, Blind rhythmic source separation: Nonnegativity and repeatability, in: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2010, pp. 2006–2009.
[29] M. Kim, J. Yoo, K. Kang, S. Choi, Nonnegative matrix partial co-factorization for spec- tral and temporal drum source separation, IEEE Journal on Selected Topics in Signal Processing 5 (6) (2011) 1192–1204.
[30] N. Seichepine, S. Essid, C. Févotte, O. Cappé, Soft nonnegative matrix co-factorization, IEEE Transactions on Signal Processing 62 (22) (2014) 5940–5949.
[31] J. De La Torre Cruz, F. J. Cañadas Quesada, N. Ruiz Reyes, P. Vera Candeas, J. J. Carabias Orti, Wheezing sound separation based on informed inter-segment non-negative matrix partial co-factorization, Sensors 20 (9) (2020) 2679.
[32] D. Badawy, N. Duong, A. Ozerov, On-the-fly audio source separation—a novel user-friendly framework, IEEE/ACM Transactions on Audio, Speech, and Language 25 (2) (2016) 261–272.
[33] V. Bisot, R. Serizel, S. Essid, G. Richard, Leveraging deep neural networks with nonnegative representations for improved environmental sound classification, in: IEEE International Workshop on Machine Learning for Signal Processing (MLSP), IEEE, 2017, pp. 1–6.
[34] A. Mesaros, A. Diment, B. Elizalde, T. Heittola, E. Vincent, R. Bhiksha, T. Virtanen, Sound event detection in the dcase

2017 challenge, IEEE/ACM Transactions on Audio, Speech and Language Processing 27 (6) (2019) 992–1006.
[35] Freesound by Music Technology Group, Universitat Pompeu Fabra, https://freesound.org/, online. Accessed: 2020-04-27 (2005).
[36] Findsound by Comparisonics Corporation, https://www.findsounds.com/, online. Accessed: 2020-04-27 (2020).
[37] Detection and Classification of Acoustic Scenes and Events DCASE 2017 Challenge. Detection of rare sound events (Tampere University of Technology), http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/task-rare-sound-event-detection, online. Accessed: 2020-04-27 (2017).
[38] Signal Processing Information Base (SPIB). NOISEX database. Speech Babble, http://spib.linse.ufsc.br/noise.html, online. Accessed: 2020-04-27 (1990).
[39] ETSI TS 103 224 V1. Speech and multimedia Transmission Quality (STQ); A sound field reproduction method for terminal testing including a background noise database. Background Noise Database: cafeteria and pub, https://docbox.etsi.org/stq/Open/TS%20103%20224%20Background%20Noise%20Database/Binaural, online. Accessed: 2020-04-27 (2014).
[40] Detection and Classification of Acoustic Scenes and Events DCASE 2017 Challenge. Sound event detection in real life audio (Tampere University of Technology), http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/task-acoustic-scene-classification, online. Accessed: 2020-04-27 (2017).
[41] TUT Sound events 2017, Development dataset, https://zenodo.org/record/814831, online. Accessed: 2020-04-27 (2017).
[42] TUT Sound events 2017, Evaluation dataset, https://zenodo.org/record/1040179, online. Accessed: 2020-04-27 (2017).
[43] PASCAL Classifying heart sounds challenge, http://www.peterjbentley.com/heartchallenge/, online. Accessed: 2020-04-27 (2011).
[44] PhysioNet/CinC challenge. National Institute of General Medical Sciences and the National Institute of Biomedical Imaging and Bioengineering, https://www.physionet.org/physiobank/database/challenge/2016/, online. Accessed: 2020-04-27 (2013).
[45] S. Charleston-Villalobos, L. Dominguez-Robert, R. Gonzalez-Camarena, A. Aljama-Corrales, Heart sounds interference cancellation in lung sounds, in: 2006 International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2006, pp. 1694–1697.
[46] S. M. Debbal, F. Bereksi-Reguig, Spectral analysis of the pcg signals, The Internet journal of microbiology 2 (2006).
[47] D. Oletic, V. Bilas, Asthmatic wheeze detection from compressively sensed respiratory sound spectra, IEEE journal of biomedical and health informatics 22 (5) (2018) 1406–1414.
[48] A. Sovijarvi, J. Vanderschoot, J. Earis, Standardization of computerized respiratory sound analysis, European Respiratory Review 10 (77) (2000) 585–585.
[49] S. Reichert, R. Gass, C. Brandt, E. Andrès, Analysis of respiratory sounds: state of the art, Clinical medicine. Circulatory, respiratory and pulmonary medicine 2 (2008) CCRPM–S530.
[50] J. Torre-Cruz, F. Canadas-Quesada, S. García-Galán, N. Ruiz-Reyes, P. Vera-Candeas, J. Carabias-Orti, A constrained tonal semi-supervised non-negative matrix factorization to classify presence/absence of wheezing in respiratory sounds, Applied Acoustics 161 (2020) 107–188.
[51] E. Vincent, R. Gribonval, C. Févotte, Performance measurement in blind audio source separation, IEEE transactions on audio, speech, and language processing 14 (4) (2006) 1462–1469.
[52] C. Févotte, R. Gribonval, E. Vincent, Bss_eval toolbox user guide–revision 2.0 (2005).
[53] Y. Matsui, S. Makino, N. Ono, T. Yamada, Multiple far noise suppression in a real environment using transfer-function-gain nmf, in: 2017 25th European Signal Processing Conference (EUSIPCO), IEEE, 2017, pp. 2314–2318.
[54] A. Liutkus, D. Fitzgerald, Z. Rafii, Scalable audio separation with light kernel additive modelling, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2015, pp. 76–80.
[55] F. Li, M. Akagi, Blind monaural singing voice separation using rank-1 constraint robust principal component analysis and vocal activity detection, Neurocomputing 350 (2019) 44–52.
[56] S. Venkataramani, C. Subakan, P. Smaragdis, Neural network alternatives toconvolutive audio models for source separation, in: IEEE International Workshop on Machine Learning for Signal Processing, IEEE, 2017, pp. 1–6.
[57] P. C. Loizou, Speech enhancement: theory and practice, CRC press, 2013.
[58] G.-C. Chang, A comparative analysis of various respiratory sound denoising methods, in: 2016 International Conference on Machine Learning and Cybernetics (ICMLC), Vol. 2, IEEE, 2016, pp. 514–518.
[59] S. Leng, R. San Tan, K. Tshun, C. Chai, C. Wang, G. D., L. Zhong, The electronic stethoscope, Biomedical engineering online 66 (2015). doi:10.1186/s12938-015-0056-y.