UNIVERSIDAD DE JAÉN

ESCUELA POLITÉCNICA SUPERIOR DE LINARES
DEPARTAMENTO DE INGENIERÍA DE TELECOMUNICACIÓN

TESIS DOCTORAL

SEPARACIÓN DE FUENTES SONORAS EN SEÑALES MUSICALES

PRESENTADA POR:
FRANCISCO JOSÉ RODRÍGUEZ SERRANO

DIRIGIDA POR:
DR. D. PEDRO VERA CANDEAS
DR. D. NICOLÁS RUIZ REYES

JAÉN, 17 DE JULIO DE 2014

ISBN 978-84-8439-878-3

A mamá, por creer siempre en mí.
A Pilar, por su apoyo incondicional.

Agradecimientos

Esta tesis doctoral, centrada en la separación de fuentes sonoras en señales musicales monocanal, ha sido posible gracias dos proyectos de investigación. El primero de ellos es el proyecto TEC2009-14414-C03-02, denominado *ANálisis, CLAsificación y Síntesis para la Separación de Sonidos. AnClaS*^{3}*v*2, financiado por el Ministerio de Ciencia e Innovación de España, siendo el Dr. D. Nicolás Ruiz Reyes el investigador principal del proyecto. El segundo es el proyecto TEC2012-38142- C04-03, denominado *Procesado distribuido y colaborativo de señales sonoras algoritmos, herramientas y aplicaciones: interpretación de música en directo - DisCoSound:Live*, financiado por el Ministerio de Economía y Competitividad de España, siendo el Dr. D. Pedro Vera Candeas el investigador principal del proyecto.

El desarrollo de este trabajo de investigación ha sido un gran reto y, a la vez, una gran motivación. Mis escasos conocimientos musicales me hacen ser profano en el campo en el que se enmarca la aplicación fundamental de la tesis. Sin embargo, la teoría de señal, y en concreto el procesado de señal de audio, es uno de los aspectos que más interés me ha suscitado a lo largo de mis estudios universitarios. Una motivación más es el hecho que los resultados de esta tesis puedan ser punto de partida para continuar en la investigación en escenarios más cercanos a las necesidades musicales actuales, como es la separación de fuentes en señales multicanal o el alineamiento música/partitura, e incluso que el conocimiento que puedan generar sea vertido sobre la industria musical.

Quiero agradecer al Dr. Nicolás Ruiz Reyes, Catedrático de universidad, Vicerrector de Infraestructuras de la Universidad de Jaén y Director del grupo de investigación *Tratamiento de Señales en Sistemas de Tele-*

comunicación del Departamento de Ingeniería de Telecomunicación de la Universidad de Jaén, por darme dado una doble oportunidad de trabajar el en seno de un grupo humano excepcional, por su implicación en esta tesis, supervisando, aconsejando y aportando sus ideas, así como por su siempre valorada opinión personal sobre el devenir laboral y personal en todas las circunstancias de la vida.

Igualmente quiero hacer constar mi sincero agradecimiento al Dr. Pedro Vera Candeas, Subdirector de la Escuela Politécnica Superior de Linares, con quien he trabajado a diario durante varios años. Quiero agradecer su implicación, interés y la cantidad de tiempo invertido para que esta tesis saliera a delante, así como haberme permitido trabajar junto a él a lo largo de todos mis estudios universitarios. Sin su claridad mental, sus brillantes ideas y su capacidad de abstracción, nada de esta tesis hubiera sido posible. Así mismo, le agradezco su opinión neutral sobre cualquier asunto debatido, exponiendo claramente las circunstancias, algo que me ha ayudado y de lo que también he aprendido.

Quiero también agradecer al Dr. Francisco Jesús Cañadas Quesada su apoyo, consejos y colaboración siempre que lo he necesitado. Igualmente hago extensible mi agradecimiento a todo el grupo de investigación *Tratamiento de Señales en Sistemas de Telecomunicación*, en especial al Dr. Damián Martínez Muñoz y el Dr. Raúl Mata Campos por el apoyo y colaboración que me han ofrecido siempre. También agradezco a los doctores Sebastián García Galán, José Enrique Muñoz Expósito y Rocío Pérez de Prado el apoyo e interés mostrados en todo momento.

Mi agradecimiento profundo y sincero al Dr. Julio José Carabias Orti, la interconexión entre sus líneas de investigación y las mías han enriquecido enormemente esta tesis, nuestras líneas de trabajo paralelas se han realimentado de manera muy fructífera. Sus consejos, ideas y colaboración han sido fundamentales para el desarrollo y consecución de este trabajo de investigación. También quiero agradecer a Pablo Cabañas Molero sus aportes técnicos y opinión cuando los he necesitado.

Todo trabajo tiene un reflejo del ambiente de trabajo en el que se ha llevado a cabo. Por ello debo agradecer a todos mis compañeros de la sala de investigadores: Diego, Antonio, Julio, Pablo, Rocío, Amparo, Juan, Pedro, David, Casto, Piedad, Paco, Salah y José Guadalupe, por sus consejos y

ánimo cuando se han necesitado. Entre todos nosotros se han forjado muchas amistades y vivido muy buenos momentos. Muchos de ellos ya son doctores, y otros lo serán en breve, ojalá nuestra querida sala vuelva a llenarse de conocimiento e investigación hasta volver a los momentos en los que se quedaba pequeña.

Quisiera agradecer al todo el Departamento de Ingeniería de Telecomunicación de la Universidad de Jaén, el apoyo y las facilidades prestadas. He sido alumno de la mayoría de los docentes que lo componen, a todos ellos les agradezco la formación y capacidades que me han aportado para ser Ingeniero de Telecomunicación y poder desarrollar mi labor investigadora.

A todos ellos, mi gratitud y mi mano tendida.

Francisco José Rodríguez Serrano
Linares, Junio 2014

Resumen

La separación de fuentes sonoras es, a día de hoy, un campo abierto en el ámbito de la investigación del procesado de señal. Se puede considerar como fuente sonora toda aquella que genere sonido susceptible de ser captado por un sensor de audio, es decir, un micrófono. Por tanto, son fuentes sonoras los instrumentos musicales, la voz, o cualquier fuente de ruido natural o artificial. La separación de fuentes viene motivada de distinta manera en función de las fuentes que se desean separar. Existe una gran corriente científica interesada en la separación de la voz y el ruido en señales ruidosas, por tanto su objetivo es mejorar la calidad de la señal de voz. De manera similar hay trabajos que pretenden separar una señal musical del ruido de fondo o del ruido propio del instrumento de grabación. En esta tesis el enfoque es distinto a estos. Se pretende separar la señal de varias fuentes musicales armónicas, es decir, instrumentos armónicos.

Este tipo de separación de fuentes viene motivada por el interés en contar con señales separadas de instrumentos de una misma composición para un gran abanico de potenciales aplicaciones a las que puede dar lugar. Sin embargo, en este trabajo de investigación el objetivo principal no es obtener una aplicación final, sino desarrollar técnicas de separación y evaluar qué tipo de información de entrada al sistema puede ser interesante emplear para que la separación obtenga los mejores resultados posibles. Se ha trabajado con señales monocanal, por ser el escenario más sencillo y versátil, entendiendo que las técnicas y conclusiones que se alcancen pueden ser aplicadas en el escenario multicanal de manera trivial.

El uso de fuentes de información adicionales a la señal, convierte este tipo de separación en separación de fuentes informada. La separación de

fuentes a ciegas, es decir, sin ningún tipo de información adicional, no obtiene resultados competitivos. Cuando se trabaja en el escenario multicanal, aunque no se aporte más información que las señales de entrada, éstas ya cuentan con la información espacial de manera intrínseca para trabajar la separación. Por tanto, en el escenario monocanal es necesario trabajar con otras fuentes de información.

Esta tesis se basa en la hipótesis de que el uso de modelos espectrales de instrumento, que ya han demostrado ser útiles en otros campos del procesado de señal musical, pueden ser una buena herramienta para discriminar la pertenencia de parte de la energía de la señal a un instrumento u otro, obteniendo resultados más competitivos que los actuales. Además otra hipótesis fundamenta esta tesis de manera que si se consigue solventar el problema de los parciales armónicos solapados de notas concurrentes, el nivel de aislamiento de las fuentes será mayor, obteniendo así mejores resultados en la calidad de la separación de las fuentes.

Siguiendo la línea marcada por estas hipótesis se han desarrollado varios trabajos que han ido aportando conocimiento y conclusiones sobre el uso de información adicional y la separación de parciales solapados en la separación de fuentes musicales.

En el primer trabajo se establece el marco de desarrollo para la separación de fuentes con la factorización de la señal mediante *(Non-Negative Matrix Factorization (NMF)* y se valoran varios modelos espectrales de instrumento para su uso como información previa de las fuentes. Estos modelos de instrumento se entrenan previamente mediante una factorización informada de unas señales de entrenamiento. A continuación quedan congelados y pasan a formar parte de unas de las matrices de factorización que no es modificada en tiempo de factorización de la señal de análisis. Se valoran los resultados de separación con tres modelos de instrumento y un sistema de separación de fuentes, del estado del arte, que no cuenta con modelos de instrumento. Estos resultados muestran los beneficios del uso de modelos que describan el comportamiento espectral de cada fuente, así como conclusiones interesantes sobre cada uno de los modelos analizados.

En el segundo de los trabajos se implementa una factorización de señal con una restricción de monofonía para cada fuente. Se trabaja sobre señales polifónicas compuestas por fuentes monofónicas. Esta restricción se

implementa en dos esquemas de factorización, uno iterativo con NMF y otro esquema *Sparse Coding (SC)* más apropiado para aplicaciones que requieran baja complejidad de cálculo.

En el tercer trabajo se pretende conseguir la adaptación de los modelos de instrumento inicialmente entrenados para que represente fielmente al instrumento real de la señal de entrada. Para ello es necesario realizar una adaptación controlada, esta adaptación debe contar con información temporal de *score* para conocer los parciales de cada nota que se encuentran solapados con los de otras notas de otro instrumento. Esta información no es válida para la adaptación, puesto que su uso llevaría el modelo hacia representaciones erróneas de cada nota. Una vez que se considera precisa la información de *score*, esta misma información se emplea como inicialización de la matriz de ganancias del algoritmo NMF, lo cual beneficia positivamente a los resultados de separación. Con todo ello, en este trabajo se propone un sistema de separación de fuentes informada con modelos de instrumento adaptativos. Los resultados obtenidos avalan la inclusión de ambas fuentes de información al proceso de separación. Además se propone una versión algoritmo de separación y adaptación que puede ejecutarse de manera *online*.

Por último se aborda el problema de los parciales solapados en el último trabajo. Cuando dos parciales de notas distintas coinciden en la misma posición tiempo/frecuencia, se produce una interferencia entre ambos fasores. Esta interferencia provoca que en el análisis de la señal sólo se conozca el valor de amplitud y fase del fasor suma, desconociéndose el valor de estos parámetros para cada fasor independiente. Gracias a los modelos de instrumento se puede estimar el valor de amplitud para cada fasor en correlación con los demás parciales no solapados de la misma nota. Posteriormente, con un proceso de minimización se estiman los valores de fase para cada fasor. De esta manera, los fasores originales pueden ser sintetizados sobre las señales de salida. Esta solución, con los resultados obtenidos, demuestra obtener un aislamiento importante entre las fuentes.

Palabras clave: separación de fuentes musicales, transcripción musical, frecuencia fundamental, polifonía, solapamiento de parciales, descomposición de señales, factorización, parcial armónico, modelo de instrumento,

información espectral, non-negative matrix factorization (NMF), modelo filtro-fuente, restricción monofónica, adaptación de modelos, modelo multi-excitación, fasor, amplitud, fase, máscaras de Wiener, análisis espectral.

Abstract

Nowadays, the sound source separation is an open field in the signal processing research. It can be considered as a sound source each element which generates any sound that can be detecter by a sound sensor, it is a microphone. Then a musical instrument, the human voice or any noise generator are sound sources. The sound source separation is motivated by different ways depending on the sources that are going to be segregated. There is a interest of the scientific community about the separation of voice and noise, it is the enhancement of a voice noisy signal. In a similar way, there are some works over the enhancement of musical signals mixed with recording or ambiental noise. This thesis has a different focus. It is pretended to separate some harmonic musical sources, which are musical instruments.

This kind of source separación is motivated by the interest of arranging with instrument separated signals from the same composition for a great amount of potential applications. However this research work is not focused on getting to a final application. Here, the main objective is to develop separation techniques and to assess which kind of additional information, apart from the input signal, could help to obtain better results. The separation has been developed over one channel signals because of being a simple and diverse scene. The techniques developed at this thesis are suitable for multichannel signal source separation.

When the source separation is made with some extra information, apart from the input signal, it is known as Informed Source Separation (ISS). The Blind Source Separation (BSS), which is made without any extra information, does not get competitive separation results over mono channel signals. In the multichannel case, despite considering only the input signals, the-

se signals has extra intrinsic information, which is the spatial information of the recording. So that, the mono channel scene requires other kind of information in order to obtain competitive results.

This thesis is based on the hypothesis that the use of spectral instrument models, which have been proven as useful information at other signal processing tasks, could be a good tool to discriminate what instrument is the owner of a concrete piece of energy from the input signal. Then better results than the actual ones could be obtained. Besides, other basic hypothesis of this thesis is that a higher isolation of the notes of each instrument could be achieved if the overlapped harmonic partial problem were solved.

Following this two hypothesis, several works have been developed in order to generate knowledge, results and conclusions about the use of extra information at the mono channel source separation and the separation of the overlapped harmonic partials.

In the first work a NMF factorization framework for the sound source separation is developed. Also, several spectral instrument models have been analyzed as extra information for the separation process. These instrument models are previously trained by an informed factorization of several training signals. Then, they are frozen and they are used to initialize one of the NMF matrixes which is not changed during the factorization of the input signal. The separation results with each instrument model have been valued and compared with a non-spectral informed state-of-the-art separation method. These results show the benefits of the uses of the instrument models which describes the different spectral behavior of each source. Also, some interesting conclusions about each kind of instrument model are presented.

In the second work, a monophonic-source constrained factorization signal system is implemented. It works over polyphonic signals composed by monophonic sources. This constraint is implemented over two frameworks: an iterative NMF and a Sparse Coding (SC) framework, which is more suitable for low complexity applications.

The third work claims for the adaptation of the instrument models, which are initially trained, to the real played source from the input signal. To do that, it is necessary to implement a controlled adaptation process while factorizing the input signal. This requires to use the score information in order to select which partial of each note are overlapped with other ones

from other instrument. This overlapped information is not valid to adapt the models because if it were used, the models would become up to a wrong representation of the notes. Once the score information is considered to select the overlapped partials, this information can be used to initialize the other matrix of the NMF framework. This initialization would benefit the separation results because of the simplification of the factorizing process. All this ideas are summarized to compose an informed source separation system with adaptive instrument models. The obtained results support the use of both information (instrument models and score) for the separation process. Besides an online algorithm for the separation and adaptation process is proposed.

Finally, the overlapped harmonic partials problem is addressed. When two partials from two, or more, notes from different instruments are placed at the same time/frequency region, both phasors are interfered. This interference causes that only the amplitude and the phase from the resulting added phasor is obtained while the amplitude and the phase of the two independent phasors are unknown. The instrument models allow to obtain an estimation of the amplitude of each overlapped partial with the aid of the non-overlapped ones from the same note. Then the phase of each phasor is estimated with a minimization process that takes into account the resulting added phasor parameters (amplitude and phase) and the estimated amplitudes for each one. After estimating all the parameters, each phasor can be synthesized over each output signals. This solution shows to obtain a important isolation of between the sources.

Key words: Sound Source Separation, Automatic Music Transcription, fundamental frequency, polyphony, overlapped partial, signal factorization, harmonic partial, instrument model, spectral information, non-negative matrix factorization (NMF), source-filter model, monophonic constraint, model adaptation, multi-excitation model, phasor, amplitude, phase, Wiener mask, spectral analysis.

Acrónimos

AMT	Automatic Music Transcription
ANLS	Alternative Non-Negative Least Squares Algorithm
AQO	Audio Quality Oriented
APS	Artifacts-related Perceptual Score
ASA	Auditory Scene Analysis
BHC	Basic harmonic constrained
BPM	Beats por minuto
BSS	Blind Source Separation
CASA	Computational Auditory Scene Analysis
DTW	Dynamic Time Warping
EM	Expectation Maximization
EUC	Euclidean (distancia)
ERB	Equivalent Rectangular Bandwidth
FFT	Fast Fourier Transform
FS-HMM	Factorial Scaled Hidden Markov Model
GEM	Generalized Expetation Maximization
GMM	Gaussian Mixture Model
GSMM	Gaussian Scaled Mixture Model
HCE	Harmonic Comb Excitation
HMM	Hidden Markov Model
ICA	Independent Component Analysis
IFFT	Inverse Fast Fourier Transform
IS	Itakura Saito (divergencia)
IPS	Interference-related Perceptual Score
ISA	Independent Subspace Analysis
ISR	source Image to Spatial Ratio

ISS	Informed Source Separation
KL	Kullback-Leibler (divergencia)
LTI	Linear Time Invariant
MAP	Maximum a Posteriori
MBHC-MS	Monophonic Basic Harmonic Constrained Model - Monophonic Signals
MBHC- PM	Monophonic Basic Harmonic Constrained Model - Polyphonic Signals
MEI	Multi-Excitation per Instrument
MIDI	Musical Instrument Digital Interface
ML	Maximum Likelihood
MU	Multiplicative Update
NMF	Non-negative Matrix Factorization
NNLS	Non-negative Least Squares
NNSC	Non-negative Sparse Coding
OPS	Overall Perceptual Score
PCA	Principal component analysis
PLCA	Probabilistic Latent Component Analysis
PSM	Perceptual Similarity Measure
QERB	Quadratic Equivalent Rectangular Bandwidth
SAGE	Space-Alternating Generalized Expectation-Maximization
SAR	Signal to Artifact Ratio
SC	Sparse Coding
SDR	Signal to Distortion Ratio
SIR	Signal to Interference Ratio
SNR	Signal to Noise Ratio
SO	Significance Oriented
SSS	Sound Source Separation
STFT	Short Time Fourier Transform
SVD	Single Value Decomposition
TPS	Target-related Perceptual Score

Índice general

Índice de figuras

Índice de tablas

Parte I

Planteamiento y Contexto de la Investigación

Capítulo 1

Introducción

1.1. Contexto y localización de la investigación

El sistema auditivo humano es capaz de procesar y discriminar los sonidos presentes a su alrededor. Si entendemos la escena auditiva como el conjunto de sonidos concurrentes en el tiempo, que están presentes en el entorno de un punto del espacio, y que pueden ser percibidos por el oído humano, el ser humano puede realizar un análisis de esta escena y establecer una prioridad en función del sonido que más interés le cause. Una persona que escucha un concierto de una orquesta puede identificar al instrumento que lleva la melodía y seguirla mentalmente, o bien, puede elegir los instrumentos acompañantes obviando la melodía, incluso puede centrar su atención sólo en un tipo de instrumento y seguir las notas tocadas, por ejemplo, por los clarinetes de la orquesta.

Esta capacidad de distinguir y privilegiar uno o varios de los sonidos que llegan a nuestro oído se denomina *Análisis de la escena auditiva* y se comenzó a estudiar en [BregmanBook90], donde se estudia la psicología del análisis de la escena auditiva (Auditory Scene Analysis, ASA) en humanos. Este documento significó el inicio del interés de la comunidad científica en las líneas del procesado digital de señales musicales polifónicas y multitímbricas. Se han estado tratando de automatizar ciertas acciones que el ser humano es capaz de realizar en relación a los sonidos musicales, como la separación de fuentes, relacionada con la discriminación auditiva ante-

riormente expuesta, e incluso la transcripción musical para personas mas ilustradas en el campo de la composición musical. A día de hoy, los avances realizados y los resultados obtenidos son prometedores pero aún no alcanzan un nivel para considerar resuelta esta problemática. Por tanto la separación de fuentes musicales aún representa un reto para los investigadores de hoy, dejando un gran campo abierto para desarrollar trabajos como el que se va a exponer en esta tesis doctoral.

La separación de fuentes musicales se ha abordado en los últimos años mediante métodos de descomposición de señal, junto con una clasificación de dichas componentes y finalmente la síntesis de las fuentes por separado. La mayoría de los algoritmos de descomposición de señal se basan en un modelo de señal lineal e instantáneo, en el que cada trama de la señal mezclada puede descomponerse en una suma ponderada de un conjunto de funciones base.

En la actualidad, la principal limitación en la separación de fuentes musicales está causada por la relación armónica existente entre los sonidos que concurren en el tiempo. Esta situación es muy común en la música occidental, debido a su riqueza en polifonía (por ejemplo en una orquesta de música). En estas situaciones se produce un solapamiento frecuencial de cierta parte de los distintos sonidos presentes en la escena. Este fenómeno se le denomina solapamiento de parciales armónicos, es el principal problema que se pretende solventar en esta tesis y será tratado progresivamente a lo largo de ella.

El problema de la separación de fuentes se tratará sobre señales monoaurales, siendo éste el escenario más complejo que existe si se atiende al número de canales de información para la separación de fuentes. Se ha trabajado sobre este tipo de señales porque no se pretende llegar a un sistema o aplicación final, sino valorar las diferentes técnicas y métodos empleados en el sistema, atendiendo a los tipos de información que son útiles para que la separación de fuentes sea más exitosa. Estas mismas técnicas pueden ser extrapoladas y aplicadas sobre sistemas de separación multicanal y de esa manera obtener una separación de mayor calidad sonora. No obstante, la mayoría de señales de audio en distribución se encuentran en estéreo (dos canales), siendo este un número insuficiente de canales para que los esquemas de separación multicanal puedan obtener un beneficio sustancial por el

uso de más de un canal. El uso de dos canales sería suficiente únicamente para unas combinaciones de fuentes muy concretas (dos o tres fuentes).

Una revisión del estado del arte revela que los algoritmos de separación de fuentes existentes se basan en el *Análisis de Componentes Independientes (Independent Component Analysis, ICA)*, *(Sparse Coding, SC)* y *Factorización de matrices no negativas (Non-negative Matrix Factorization, NMF)* como herramientas para la obtención de funciones base y, así mismo, para la descomposición de la señal de entrada mediante la suma ponderada de dichas funciones base. Tanto la fase de entrenamiento para la obtención de las funciones base, como la de descomposición de la señal pueden ser informadas (por ejemplo, si se les aporta cierta información de la ocurrencia de notas en cada instante temporal), o bien no informada (siguiendo el ejemplo, si no hay información temporal sobre la combinación de notas presente en la señal de entrada para cada instante). Incluso, se puede prescindir de la fase de entrenamiento, quedando pendiente, de esa manera, el aprendizaje de las funciones base para ser realizado durante la misma fase de descomposición, en cuyo caso se le denomina separación de fuentes a ciegas, *(Blind Source Separación, BSS)*. Sin embargo, en el caso de la separación de fuentes desde señales monofónicas, que prescinde de la información espacial, la ausencia de información temporal y/o un entrenamiento previo que permita generar modelos de las fuentes a separar, hace que los algoritmos obtengan resultados lejos de poder ser útiles para un potencial usuario.

En esta tesis doctoral se investiga sobre separación monoaural de fuentes informada de modelos de timbre de los instrumentos musicales y de información de partitura. Se demuestra la mejora que aporta este tipo de información para realizar una buena descomposición de la señal de entrada, así como la importancia y la manera de obtener modelos que se ajusten fielmente al instrumento real que ha generado el sonido, que en muchas ocasiones pueden diferir parcialmente del modelo correspondiente a otro instrumento del mismo tipo. Así mismo, este modelado del timbre de los instrumentos presentes en la composición, se configura como una información útil para la resolución del problema de la separación de los parciales armónicos solapados, que será abordado como fase final de esta tesis.

1.2. Hipótesis y Objetivos

La investigación propuesta en este capítulo se basa en dos hipótesis, las cuales se presentan a continuación. En esta tesis se pretende desarrollar la investigación que lleve a la demostración de la veracidad de estas hipótesis.

1.2.1. Hipótesis

HIPÓTESIS 1:

Si toda señal musical puede descomponerse en funciones básicas combinadas linealmente, el conocimiento de modelos físicos que caractericen [FletcherBook98] cada instrumento puede aportar información importante a la separación de fuentes si dichos modelos se adaptan para ser empleados como bases armónicas. Se espera demostrar que el uso de modelos de instrumento en la separación de fuentes permite obtener resultados más competitivos que los actuales [Klapuri10a]. Además se espera demostrar que la adaptación de dichos modelos a la escena e instrumento empleados en la generación del sonido puede optimizar los resultados obtenidos.

HIPÓTESIS 2:

Si dos tonos coinciden en la misma localización tiempo-frecuencia, el resultado de su suma es la suma de sus fasores asociados. Dado que los modelos de instrumento permiten estimar los módulos de cada uno de los fasores, se pretende diseñar una solución para obtener una aproximación a cada una de las fases de los fasores coincidentes. De esta manera la separación de parciales solapados (situación muy común en señales musicales de instrumentos armónicos) podría llevarse a cabo mediante la estimación de módulo y fase de cada uno de los parciales. Esta separación de parciales solapados reduciría en gran medida la interferencia de la señal de cada instrumento sobre las estimaciones de los demás instrumentos.

Con todo ello, es momento de plasmar los objetivos que se plantean al comienzo del desarrollo de la investigación.

1.2.2. Objetivos

1. *Adaptación del sistema de descomposición de señal con modelo paramétricos de instrumento para su uso en separación de instrumentos musicales de audio.* En el seno del grupo de investigación se ha desarrollado un modelo de señal que haciendo uso de modelos de instrumento ha sido aplicado a transcripción automática de señales polifónicas. El primer objetivo de la tesis es demostrar que el uso de modelos de instrumento supone una mejora sustancial en la separación de instrumentos musicales. Además, los modelos de instrumento, bajo determinadas condiciones podrán ser mejorados empleando información sobre las condiciones de la generación de la señal de audio. Los modelos de instrumento contienen información del sonido producido por el instrumento. Estas características pueden ser empleadas en las bases de separación para diferenciar la energía aportada por cada instrumento en la señal mezclada.

2. *Desarrollo de un método que permita estimar la fase de dos o más fasores que se suman al mezclarse las señales de audio.* En la mezcla de dos señales armónicas, algunos de los parciales de las notas que son tocadas simultáneamente pueden coincidir, total o parcialmente, en la misma localización tiempo-frecuencia. Cuando esto ocurre los parciales (con naturaleza de fasores) se suman vectorialmente. Los modelos de instrumento permiten estimar el módulo de los fasores originales. Teniendo en cuenta dicha estimación, se pretende estimar las fases que dan lugar al fasor suma.

1.3. Estructura de la tesis

La estructura de esta tesis se divide en cuatro bloques. Cada bloque está compuesto por un conjunto de capítulos tal y como se describe a continuación:

- **Planteamiento y Contexto de la Investigación** Este bloque está compuesto únicamente por el capítulo actual. Aquí se presentan los

objetivos e hipótesis, la estructura y las principales contribuciones de esta tesis.

- **Revisión de conocimientos**

  Este bloque contiene tres capítulos:

  En el capítulo 2 se describen los conceptos básicos de teoría musical y los principales atributos perceptuales que permiten caracterizar los sonidos, en especial los sonidos musicales. Además, se realiza una clasificación de los sonidos siguiendo varios criterios (número de canales, periodicidad, número de notas simultáneas y numero de timbres). Finalmente, se define e introduce el concepto de separación de fuentes musicales junto con algunas potenciales aplicaciones.

  En el capítulo 3 se desciñen algunos de los métodos de descomposición de señal mas populares, prestando especial atención a los métodos basados en NMF.

  En el capítulo 4 se describen las bases de datos comúnmente utilizadas para la evaluación de las prestaciones de los sistemas de separación de fuentes y las medidas utilizadas para cuantificar la calidad de la separación obtenida. Además se introducen los primeros sistemas de separación de fuentes y finalmente se describen los trabajos más importantes del estado del arte en la separación de fuentes musicales.

- **Contribuciones**

  Este bloque se compone de cuatro capítulos y se puede considerar el núcleo de la tesis, puesto que en ellos se describen las contribuciones científicas de la misma.

  En el capítulo 5 se presenta la estructura básica de un sistema de separación de fuentes musicales basado en NMF con modelos espectrales de instrumento. Se realiza una comparativa de varias propuestas de modelos de instrumento adecuados para el esquema NMF, basándose en la medida de la calidad de separación que se obtiene con cada uno de ellos. La principal contribución de este capítulo es el esquema de separación basado en NMF con modelos espectrales de instrumento, así como la selección de un modelo adecuado para este fin.

En el capítulo 6 se presenta un modelo de descomposición de señal con restricción de monofonía y armonicidad para señales polifónicas compuestas por fuentes monofónicas. Se evalúa la propuesta en dos aplicaciones típicas del procesado de señal musical como son: transcripción musical automática y separación de fuentes musicales. Las principales contribuciones son la restricción de monofonía para un sistema de descomposición de señal basado en NMF y la obtención de transcripción independiente para cada instrumento presente en la composición musical.

En el capítulo 7 se propone un sistema online de separación de fuentes con información temporal y modelos de instrumento adaptativo sobre un esquema NMF de descomposición de señal. La información temporal y espectral (modelos de instrumento) se consideran de gran utilidad a la hora de separa fuentes en señales monofónicas, y así se demuestra en este capítulo. Además se demuestra que la fidelidad de los modelos de instrumento es un punto crucial para la mejora de la calidad de separación. Las principales contribuciones de este capítulo son el diseño de un algoritmo de separación de fuentes informada, la adaptación, simultánea a la factorización, de los modelos de instrumento para representar con mayor fidelidad los instrumentos reales de la composición y el diseño de un algoritmo *online* de separación y adaptación de modelos de instrumento que utilice sólo información de tiempo pasado.

En el capítulo 8 se aborda el problema de los parciales solapados en frecuencia en notas que están activas de manera simultánea. Se propone realizar una estimación de la fase original de cada fasor solapado, que junto con la estimación de amplitud obtenida con el modelo de instrumento, permita sintetizar un parcial que minimize la distorsión con la señal mezclada original. Esta estimación de fase y amplitud, con la correspondiente síntesis de las zonas solapadas, pretetende disminuir la interferencia que se produce entre las pistas de cada instrumento por el fenómeno de los parciales solapados, constituyéndose esto como la principal contribución del capítulo.

- **Conclusiones y líneas futuras**

En el capítulo 9 se resumen las principales conclusiones de esta tesis y se plantean varias líneas futuras para continuar la línea de investigación.

1.4. Principales contribuciones de la tesis

Para finalizar este capítulo se exponen las principales contribuciones derivadas de esta tesis.

1. Desarrollo de un sistema informado de separación monoaural de fuentes instrumentales basado en el algoritmo de descomposición de señal NMF (Non-negative Matrix Factorization) [Rodriguez12].

2. Implementación de una restricción de monofonía por instrumento sobre un modelo NMF para la separación y transcripción independiente de señales polifónicas compuestas por fuentes monofónicas [Rodriguez13]

3. Adaptación del sistema de separación basado en NMF para el uso de información de partitura e información espectral (modelos de instrumento). Actualización de modelos de instrumento durante la factorización. Modificación del algoritmo de separación y actualización de modelos para poder ser ejecutado de manera *online*, sólo con información pasada[Rodriguez14].

4. Desarrollo de un método, que trata el problema de los parciales solapados en frecuencia. Estimación de la amplitud, con la ayuda de los modelos de instrumento, y la fase de cada parcial para reducir el nivel de interferencia entre los instrumentos presentes en cada composición. [Rodriguez14b]

Parte II

Revisión de conocimientos

Capítulo 2

Introducción a las señales musicales

2.1. Teoría Musical

Los sonidos que componen cualquier melodía o composición musical se pueden representar en una partitura (ver figura 2.1). La partitura está formada por una notación simbólica de la misma manera que las letras representan el habla humana. La representación de los sonidos con las notas no es suficiente para describir una composición, es necesario incluir información sobre otros aspectos como la velocidad de interpretación, la intensidad y matices de interpretación para que la melodía quede completamente descrita.

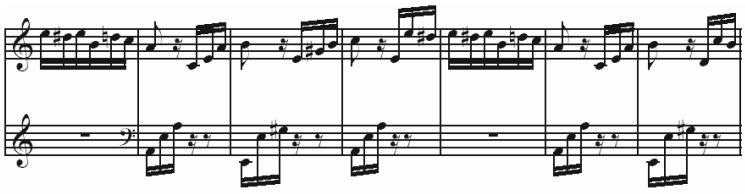


Figura 2.1: *Notación simbólica sobre una partitura de la obra "Fur Elise" compuesta por Ludwig van Beethoven.*

Los elementos básicos que conforman la notación simbólica tradicional

son los que se describen a continuación:

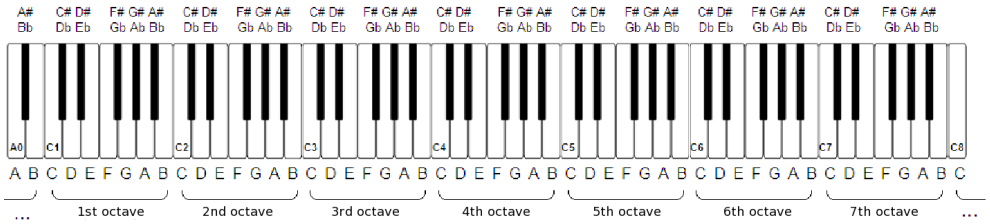


Figura 2.2: *Teclado de piano con 88 teclas incluyendo desde la nota A0 hasta la C8. Cada escala se compone de 7 teclas blancas (C, D, E, F, G, A, B) y 5 teclas negras (C# o Db, D# o Eb, F# o Gb, G# o Ab, A# o Bb).*

- **Evento** : unidad básica en música que representa un sonido con un determinado *pitch* o frecuencia fundamental (f_0) que se encuentra activo durante un determinado intervalo temporal. Las notas musicales se ordenan dentro de las escalas musicales. En la teoría musical occidental se suele dividir cada escala en doce notas (ver figura 2.2): siete notas sin alteración y cinco notas con alteración (símbolos que modifican la entonación), todas ellas ordenadas de grave a agudo. Estas divisiones de doce notas también se denomina escala cromática. En ellas el intervalo entre dos notas adyacentes se denomina semitono .

  - Según el sistema de notación latina, las siete notas sin alteración son: *Do - Re - Mi - Fa - Sol - La - Si*. Las notas con alteración son: (*Do# ó Reb, Re# ó Mib, Fa# ó Solb, Sol# ó Lab, La# ó Sib*) y se generan incrementando o disminuyendo un semitono las anteriores.
  - Según el sistema de notación inglesa o denominación literal (ver figura 2.2), las siete notas musicales sin alteración se representan mediante una letra desde la A hasta la G (*A,B,C,D,E,F, y G*). La nota latina *Do* es la equivalente a la nota ingles *C*, siguiendo esa equivalencia, la relación entre los dos sistemas se puede obtener de manera directa. Por analogía, en el sistema inglés hay cinco notas con alteración que son las siguientes: *o Ab, A# o Bb o C# o Db, D# o Eb, F# o Gb, G#*

- La **octava** a la que pertenece una nota se representa mediante un numero que sigue a la letra correspondiente (la nota *A4* es la nota A de la escala musical 4) (ver figura 2.2).

- En la música occidental, la escala musical más importante es la **escala diatónica** . Esta escala está compuesta por un conjunto de notas que difieren $2, 2, 1, 2, 2, 2, 1$ semitonos entre cada una. Estas separaciones dan lugar a las siete notas sin alteración.

- El **pentagrama** son cinco lineas horizontales y cuatro espacios que se usa para escribir sobre él los símbolos musicales. Por convenio se sigue la misma dirección de izquierda a derecha que en la escritura del lenguaje natural, por tanto, el pentagrama es una línea de tiempo, las notas situadas más a la izquierda se tocan antes que las que se encuentren a su derecha. Cuando el pentagrama se usa para instrumentos armónicos la posición del símbolo indica el pitch o frecuencia fundamental del sonido, cuando se trata de pentagramas para percusión, cada posición (en el eje vertical) corresponde a un instrumento distinto.

- La **clave** es un símbolo musical que se usa para indicar el pitch de las notas escritas sobre el pentagrama. Se sitúa sobre una de las líneas al comienzo del pentagrama e indica el nombre y pitch de las notas que se sitúen sobre dicha línea. Existen tres claves, clave de Sol, la clave de Fa y la clave de Do. Su nombre designa a la nota que se sitúe en la línea del pentagrama sobre la que se escriba cada clave.

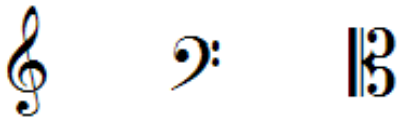


Figura 2.3: *Tipos de claves musicales (a) Clave de Sol; (b) clave de Fa; (c) clave de Do.*

- La **duración** de cada nota se representa mediante figuras musicales. Cada figura tiene una duración relativa respecto a las demás, por ejemplo, si como referencia de duración se establece la figura negra, el resto de figuras tienen las siguientes relaciones: redonda (4), blanca (2), negra (1), corchea (1/2), semicorchea (1/4), fusa (1/8) y semifusa(1/16).

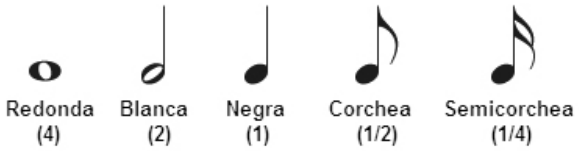


Figura 2.4: *Distintos tipos de figures musicales. Hay que indicar que las duraciones son relativas, en el caso estándar, la referencia de duración es la negra. De esa manera una redonda dura cuatro negras, una blanca dura dos negras, una corchea dura media negra y una semicorchea dura un cuarto de negra*

- El pentagrama se encuentra dividido, en la línea temporal, por unas líneas verticales que dividen la pieza en fragmentos, cada uno de ellos se denomina **compás** y se usan para organizar y hacer más cómoda la lectura de la partitura. Una división más gruesa y doble indica el comienzo y el final de la pieza musical. En ocasiones los compases se marcan con números para facilitar su localización, comenzando con el número 1 para el primer compás y continuar de manera creciente. Los compases son una herramienta para marcar el ritmo de la música. Al principio del pentagrama se indica mediante dos números en forma de fracción el ritmo del compás (ver figura 2.5). El numerador indica el número de tiempos que tendrá el compás y el denominador indica la unidad de tiempo, es decir, la figura que ocupa un tiempo completo del compás. Por ejemplo un compás $\frac{4}{4}$ indica que la unidad de tiempo es la negra y que cada compás se compone de 4 negras, si el compás fuese $\frac{2}{4}$ habría 2 negras en cada compás.

En el contexto del análisis musical cada evento o nota musical puede caracterizarse por tres parámetros básicos:

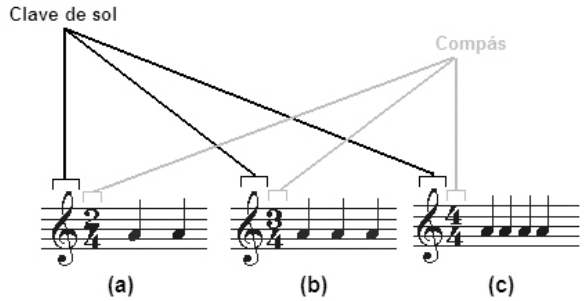


Figura 2.5: *Tipos de compás musical (a) Dos tiempos; (b) Tres tiempos; (c) Cuatro tiempos.*

- El **Pitch** es un atributo perceptual que permite ordenar los sonidos en una escala de frecuencia logarítmica, el pitch representa la nota asociada a un evento musical. Según [LindsayBook77], un sonido presenta un determinado pitch *"si la frecuencia de una sinusoide de amplitud arbitraria puede ser relacionada con el sonido percibido"*. Esto indica que un sonido tendrá un pitch sólo si ambos son relacionados entre sí por el sistema auditivo humano. El concepto físico asociado al pitch se denomina **frecuencia fundamental** f_0 que se define únicamente para señales periódicas o cuasi-periódicas [KlapuriBook06]. En esta tesis los términos pitch y frecuencia fundamental son considerados similares y se emplearán indistintamente a pesar de la leve diferencia de concepto que no afecta a su uso.

- **Onset** se define como el instante de tiempo en el que se inicia una nota musical [Klapuri99a].

- **Duración** es el intervalo temporal durante el cual una nota musical se encuentra activa en la señal de audio [Bello06]. Comienza en el instante de onset.

En la música occidental, los eventos musicales (notas) se ordenan en una escala logarítmica de manera similar al comportamiento logarítmico del oído humano [ZwickerBook90]. Por ello, la frecuencia fundamental, f_0 (Hz) de cada evento se puede representar como en la ecuación (2.1) mediante el uso

de la escala musical temperada (equal-tempered), suponiendo un teclado de piano estándar con 88 teclas [KlapuriBook06] (ver figura 2.2).

$$f_0\ (Hz) = f_{0ref}\ \cdot\ 2^{\frac{n}{12}},\quad n = -48, ..., 0, ..., 39 \tag{2.1}$$

donde n es el número de semitonos entre la frecuencia de referencia (f_{0ref}), y la frecuencia deseada f_0. Consideremos $f_{0_{ref}}$ como la frecuencia fundamental del evento $A4$ ($f_{0A4} = 440\ Hz$) subir un semitono nos lleva al evento $A4\#$ cuya frecuencia fundamental se obtiene con $f_{0A4\#} = f_{0A4} \cdot 2^{1/12} = 466{,}2\ Hz$.

El protocolo **MIDI** *(Musical Instrument Digital Interface)* es un estándar comúnmente usado para representación de información musical en dispositivos de audio. Este protocolo ofrece una relación entre Hertzios y la numeración MIDI (números naturales). Para traducir una frecuencia f_0 a notación MIDI se emplea la ecuación (2.2):

$$N_{MIDI} = [69 + 12 \cdot log_2\Big(\frac{f_0(Hz)}{440}\Big)], \tag{2.2}$$

en sentido opuesto, para convertir de la notación MIDI a hercios se emplea la ecuación (2.3):

$$f_0\ (Hz) = 440 \cdot 2^{\frac{N_{MIDI}-69}{12}}, \tag{2.3}$$

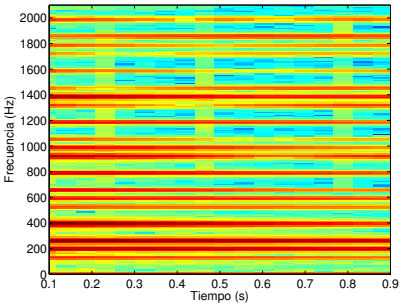
Por convenio, la nota MIDI 69 se ha asignado a la nota $A4$ con $f_0 = 440\ Hz$. En la tabla 2.1, se relacionan las frecuencias f_0 en hercios a cada nota MIDI para todas las notas musicales de las nueve octavas. Se puede ver que cubren un rango de 7900 Hz. Cada columna representa una escala musical y cada fila es un evento de la escala. En cada celda se muestran, en primer lugar la frecuencia fundamental (Hz) y en segundo lugar, entre paréntesis el numero de nota MIDI correspondiente.

En la teoría musical, se define como **intervalo musical** a la distancia en frecuencia de dos notas que comparten el tiempo de activación, es decir, son concurrentes [KlapuriPhD04]. Los intervalos son elementos muy comunes en música y representan uno de los escenarios más complejos para el procesado de señal musical por su relación con el problema de los parciales solapados, sonidos que cuentan con varios de sus parciales situados en

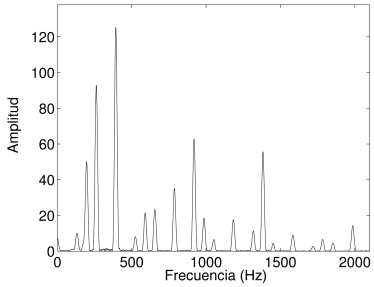
Tabla 2.1: *Relación entre todas las notas musicales y escalas con la nomenclatura MIDI*

Notas musicales (Hz - MIDI)									
Nota / Octava	0	1	2	3	4	5	6	7	8
C	16.35 (12)	32.70 (24)	65.41 (36)	130.8 (48)	261.6 (60)	523.3 (72)	1047.0 (84)	2093.0 (96)	4186.0 (108)
C#	17.32 (13)	34.65 (25)	69.30 (37)	138.6 (49)	277.2 (61)	554.4 (73)	1109.0 (85)	2217.0 (97)	4435.0 (109)
D	18.35 (14)	36.71 (26)	73.42 (38)	146.8 (50)	293.7 (62)	587.3 (74)	1175.0 (86)	2349.0 (98)	4699.0 (110)
D#	19.45 (15)	38.89 (27)	77.78 (39)	155.6 (51)	311.1 (63)	622.3 (75)	1245.0 (87)	2489.0 (99)	4978.0 (111)
E	20.60 (16)	41.20 (28)	82.41 (40)	164.8 (52)	329.6 (64)	659.3 (76)	1319.0 (88)	2637.0 (100)	5274.0 (112)
F	21.83 (17)	43.65 (29)	87.31 (41)	174.6 (53)	349.2 (65)	698.5 (77)	1397.0 (89)	2794.0 (101)	5588.0 (113)
F#	23.12 (18)	46.25 (30)	92.50 (42)	185.0 (54)	370.0 (66)	740.0 (78)	1480.0 (90)	2960.0 (102)	5920.0 (114)
G	24.50 (19)	49.00 (31)	98.00 (43)	196.0 (55)	392.0 (67)	784.0 (79)	1568.0 (91)	3136.0 (103)	6272.0 (115)
G#	25.96 (20)	51.91 (32)	103.80 (44)	207.7 (56)	415.3 (68)	830.6 (80)	1661.0 (92)	3322.0 (104)	6645.0 (116)
A	27.50 (21)	55.00 (33)	110.0 (45)	220.0 (57)	440.0 (69)	880.0 (81)	1760.0 (93)	3520.0 (105)	7040.0 (117)
A#	29.14 (22)	58.27 (34)	116.50 (46)	233.1 (58)	466.2 (70)	932.3 (82)	1865.0 (94)	3729.0 (106)	7459.0 (118)
B	30.87 (23)	61.74 (35)	123.50 (47)	246.9 (59)	493.9 (71)	987.8 (83)	1976.0 (95)	3951.0 (107)	7902.0 (119)

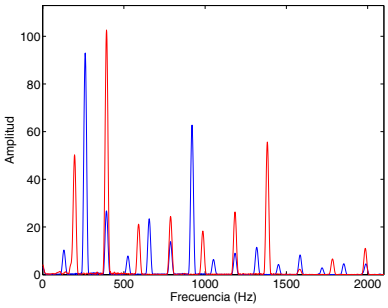
las mismas frecuencias que otros sonidos concurrentes. En la figura 2.6 se muestra un ejemplo del problema de parciales solapados en el intervalo conocido como quinta justa. En este intervalo, la relación entre sus frecuencias fundamentales es $2:3$, esto quiere decir que cada múltiplo de 2 del evento de mayor frecuenta ($G3$) está solapado con cada múltiplo de 3 del evento de menor frecuencia ($C3$)



(a) Espectrograma de la señal



(b) Espectro de la señal spectrogram



(c) Espectros individuales

Figura 2.6: *Señal polifónica perteneciente al intervalo musical denominado quinta justa, el cual está compuesto po los eventos (C3 y G3) de piano, donde aparece el problema de los parciales solapados. (a) Espectrograma de la señal. Cuanto más cálido es el color, mayor concentración de energía existe en la componente espectral. (b) Espectro de la señal completa. (c) Espectro de las señales individuales (el evento C3 se representa en color azul y el evento G3 en color rojo)*

Un **acorde** en música es una combinación de dos o más notas sonando de manera simultánea o casi simultánea [KlapuriBook06]. Los acordes más frecuentes son las triadas, que son acordes de tres notas. Un acorde puede ser consonante o disonante dependiendo de la relación entre las frecuencias de los parciales que componen dichos sonidos armónicos. Concretamente, un acorde consonante es la combinación de notas que producen un sonido agradable para la mayoría de las personas, mientras que uno disonante provocará una sensación molesta o desagradable.

El concepto de **armonía** musical esta relacionado con el uso de varios

pitchs simultáneos o acordes. La armonía analiza la relación existente entre las frecuencias fundamentales que componen un acorde y a la vez la relación entre el acorde y sus adyacentes.

La **melodía** se define como una sucesión temporal de notas musicales que se perciben como una entidad única. Habitualmente, las melodías consisten en uno o más pasajes que son cantados por un vocalista o por solos instrumentales, estos pasajes son repetidos durante la canción de varías formas distintas [KlapuriPhD04] [Paiva05] [Ryynanen08a].

2.2. Caracterización de los sonidos

Hay cuatro atributos perceptuales subjetivos que resultan muy útiles a la hora de caracterizar sonidos: frecuencia fundamental (f_0), sonoridad, duración y timbre [RossingBook90]. Hay que aclarar que la frecuencia fundamental no es un atributo perceptual pero sí que es un parámetro físico objetivo que se asocia al pitch.

2.2.1. Frecuencia fundamental (f_0)

Una señal $x(t)$ es periódica si se repite continuamente cada cierto tiempo. El periodo fundamental T_0 es el menor valor de T, en segundos, que satisface $x(t) = x(t + T0),\ \forall t$.

$$x(t) = x(t + nT0), \quad n \in (-\infty, \infty), \quad n \in \mathcal{Z} \tag{2.4}$$

Sin embargo, puesto que las señales periódicas presentes en el mundo que nos rodea son señales causales y reales y con duración finita, si existe un periodo fundamental T_0, entonces existen infinitos periodos para los cuales la señal $x(t)$ es periódica (ver ecuación (2.4)). La frecuencia fundamental f_0, en hercios, se define entonces para una señal periódica $x(t)$ como la inversa del periodo fundamental T_0 (ver ecuación (2.5)).

$$f_0\ (Hz) = \frac{1}{T0\ (s)} \tag{2.5}$$

Una señal periódica $x(t)$ se puede descomponer (ver figura 2.7) por una suma de sinusoides, dónde cada sinusoide $s_k(t)$ (ver ecuación (2.6)) se define

como la parte real de una exponencial compleja. Por tanto, cada exponencial compleja puede ser representada por una función coseno con tres parámetros (amplitud $|a_k|$, frequencia f_0 y fase ϕ_k).

$$s_k(t) = Re\big(a_k \cdot e^{jk2\pi f_0 t} Re\big) \tag{2.6}$$

$$x(t) = \sum_{k=-\infty}^{\infty} s_k(t) = \sum_{k=-\infty}^{\infty} Re\big(a_k \cdot e^{jk2\pi f_0 t} Re\big), \tag{2.7}$$

Puesto que las señales exponenciales complejas son autofunciones de los sistemas lineales e invariantes en el tiempo (LTI), toda combinación lineal de sinusoides a la entrada del sistema LTI ofrece, a su salida, la misma combinación lineal de sinusoides con un escalado en amplitud compleja. Esta propiedad de las sinusoides hace que estas señales sean idóneas para el modelado de señales musicales. Según el Teorema de Fourier [OppenheimBook97], toda señal periódica $x(t)$ se puede descomponer en infinitas sinusoides relacionadas armónicamente entre sí (ver ecuación (2.7)).

Considerando que la señal $x(t)$ es real, con valor medio nulo ($a_0 = 0$) y $a_k{=}|a_k| \cdot e^{j\phi_k}$, entonces $x(t){=}x^*(t) \rightarrow a_k{=}a^*_{-k}$,

$$x(t) = \sum_{k=-\infty}^{\infty} a_k \cdot e^{jk2\pi f_0 t} = \sum_{k=-\infty}^{-1} a_k \cdot e^{jk2\pi f_0 t} + \sum_{k=1}^{\infty} a_k \cdot e^{jk2\pi f_0 t} = \tag{2.8}$$

$$= \sum_{k=1}^{\infty} a_{-k} \cdot e^{-jk2\pi f_0 t} + a_0 + \sum_{k=1}^{\infty} a_k \cdot e^{jk2\pi f_0 t} = \tag{2.9}$$

$$= \sum_{k=1}^{\infty} \left(a_k \cdot e^{jk2\pi f_0 t} + a_{-k} \cdot e^{-jk2\pi f_0 t} \right) = \tag{2.10}$$

$$= \sum_{k=1}^{\infty} \left(a_k \cdot e^{jk2\pi f_0 t} + a^*_k \cdot e^{-jk2\pi f_0 t} \right) = \tag{2.11}$$

$$= \sum_{k=1}^{\infty} 2Re\left(a_k \cdot e^{jk2\pi f_0 t} \right) = \sum_{k=1}^{\infty} 2Re\left(|a_k| \cdot e^{j\phi_k} \cdot e^{jk2\pi f_0 t} \right) = \tag{2.12}$$

$$= 2\sum_{k=1}^{\infty} |a_k| \cos(2\pi f_k t + \phi_k) \tag{2.13}$$

donde cada componente $f_k = kf_0$, $k \in \mathbb{N}$ se denomina armónico y tiene una frecuencia múltiplo de la frecuencia fundamental [DeLiangBook06]. Sin embargo, las señales musicales del mundo real no son estacionarias a largo plazo, es decir, sus propiedades estadísticas (media μ_x y autocorrelación R_x) varían en función del tiempo. Entonces, se consideran porciones pequeñas del tiempo (decenas de milisegundos), denominadas *frames* o *tramas*, dentro de las cuales las propiedades estadísticas no varían y se pueden considerar señales estacionarias o cuasi-periódicas. De esta manera, según el Teorema de Fourier [OppenheimBook97], una señal cuasi-periódica $c(t)$ puede descomponerse en M sinusoides relacionadas armónicamente entre si (ver ecuación (2.14)),

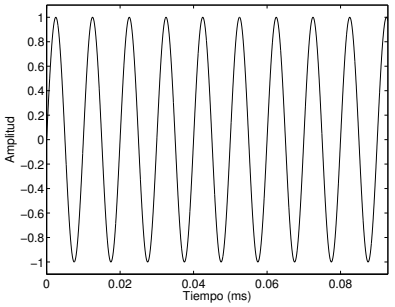
$$c(t) = \sum_{p=1}^{M} A_p \cos(2\pi f_p t + \phi_p) \tag{2.14}$$

En señales cuasi-periódicas, el concepto de armónico se amplia por el de **parcial**. Un parcial define componentes f_p cuyas frecuencias pueden desviarse levemente de los múltiplos de la frecuencia fundamental f_0, por tanto $f_p \approx pf_0$, $p \in \mathcal{N}$. En la figura 2.7 se muestran tres ejemplos de señales periódicas y cuasi-periódicas en los dominios temporal y frecuencial. En la figura 2.7(f), se muestra el espectro típico de señales musicales. Este espectro se caracteriza por los picos espectrales en cada $pf_0\ Hz$.

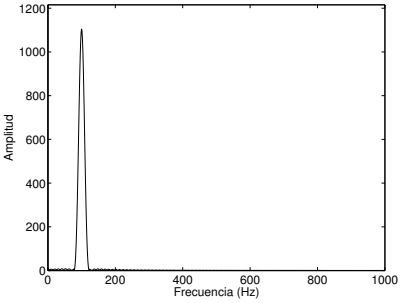
En la tabla 2.2 se puede ver el rango de frecuencias fundamentales de los instrumentos musicales más comunes para todas sus posibles notas.

2.2.2. Sonoridad, duración y timbre

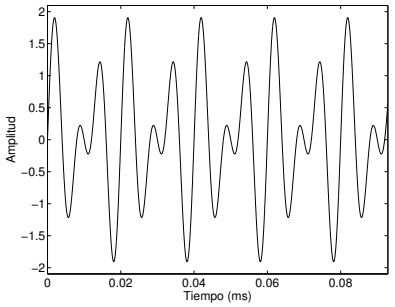
La **sonoridad** es un atributo de percepción auditiva que permite ordenar los sonidos en una escala desde los más bajos a los más altos en intensidad. La sonoridad no sólo depende de la potencia de un sonido, sino que también depende de su duración y la estructura en tiempo y frecuencia del mismo.



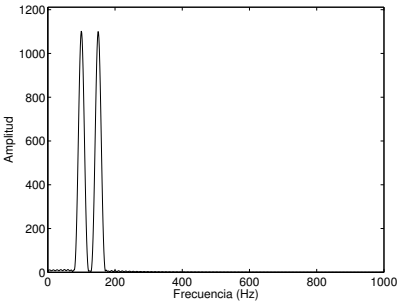
(a) Señal s_1 en el dominio del tiempo



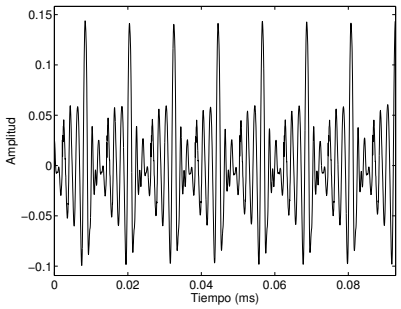
(b) Señal s_1 en el dominio de la frecuencia



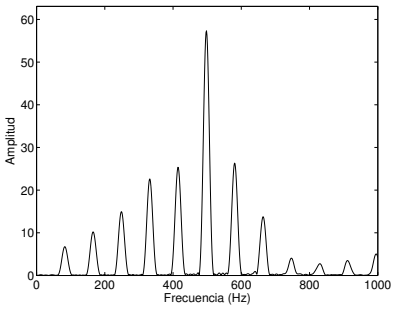
(c) Señal s_2 en el dominio del tiempo



(d) Señal s_2 en el dominio de la frecuencia



(e) Señal s_3 en el dominio del tiempo



(f) Señal s_3 en el dominio de la frecuencia

Figura 2.7: *Ejemplos de señales periódicas y cuasi-periódicas. Las gráficas (a) y (b) muestran una señal periódica s_1 de un sinusoide con $T_0 = 10ms$ y $f_0 = 100\,Hz$ en el dominio del tiempo y la frecuencia respectivamente. Las gráficas (c) y (d) muestran una señal periódica s_2 que está compuesta por la suma de dos sinusoides. La primera de ellas con $T_0 = 10ms$ y $f_0 = 100\,Hz$, y la segunda con $T_0 = 6{,}67ms$ y $f_0 = 150\,Hz$. La señal cuasi-periódica s_3 está generada por la nota $E2$ con $T_0 = 12{,}1ms$ y $f_0 = 82{,}41\,Hz$ de un fagot de la base de datos [Iowa06].*

Tabla 2.2: *Rangos de frecuencia de las posibles notas generadas por los instrumentos más comunes. [Brown09].*

Familia musical	Instrumento musical	Rango de frecuencias (Hz)
Vocal	Soprano	250-1000
	Contralto	200-700
	Baritono	110-425
	Bajo	80-350
Viento	Flautín	630-5000
	Flauta	250-2500
	Oboe	250-1500
	Clarinete (Bb)	125-2000
	Clarinete (Eb)	200-2000
	Fagot	55-575
	Trompeta	90-1000
	Saxofón Alto	125-900
	Saxofón Tenor	110-630
	Saxofón Soprano	225-1000
Metal	Trompeta	170-1000
	Trombon Tenor	80-600
	Trombon Bajo	63-400
	Tuba	45-375
Cuerda	Violin	200-3500
	Viola	125-1000
	Chelo	63-630
	Guitarra	80-630
	Piano	28-4100
Organo	Organo	20-7000
Percusión	Celesta	260-3500
	Timbales	90-180
	Carrillón	63-180
	Xilófono	700-3500

Para establecer la sonoridad en frecuencia se establecen unas curvas, denominadas isofónicas, de igual sonoridad con la referencia en $1KHz$. Estas curvas son variantes con la frecuencia, por ello dos sonidos con igual potencia pero distinta frecuencia no tendrán la misma sonoridad, puesto que es un atributo perceptual y no físico. La unidad de medida de sonoridad es el $fonio$ o fon. El umbral de silencio es un claro ejemplo de curva isofónica,

debido a que la sonoridad para $1KHz$ en el umbral del silencio equivale a 3 fonios. En la figura 2.8 se muestran las curvas isofónicas a partir del umbral del silencio.

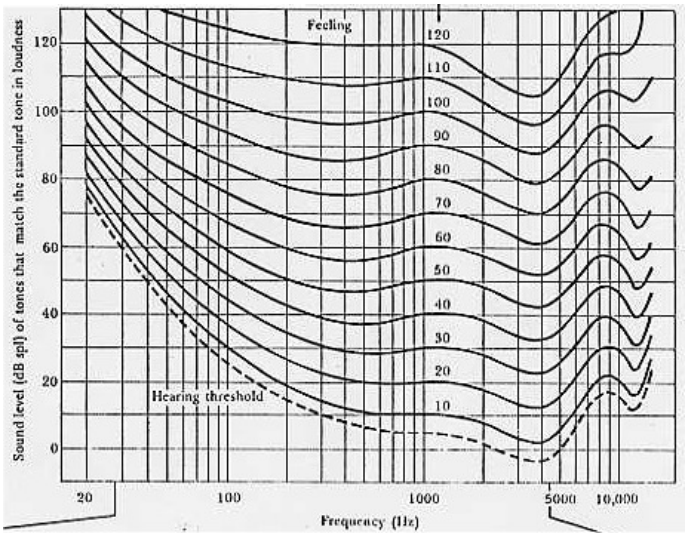


Figura 2.8: *Curvas isofónicas par tonos puros. Figura obtenida de: http://www.offbeatband.com/2009/08/the-difference-between-gain-volume-level-and-loudness/*

Un concepto musical muy relacionado con la sonoridad es la **dinámica musical**. La dinámica hace referencia al volumen de un sonido o nota. Sin embargo, también puede referirse a otros aspectos de la ejecución de una obra, como el estilo(staccato, legajo, etc.) o velocidad de interpretación. En la tabla 2.3 se muestra la relación entre sonoridad y dinámica.

Como se explicó en la sección 2.1, la **duración** es el intervalo de tiempo en el que un evento o nota 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 o nota comienza y deja de estar activo respectivamente.

El **timbre** es un atributo perceptual asociado al concepto de color del sonido [Wold96] y está muy relacionado con el reconocimiento de fuentes sonoras [Bregman90][Han95]. Este atributo permite distinguir eventos musicales con el mismo pitch, sonoridad y duración [Ans73]. A diferencia de otros atributos del sonido, el timbre no se puede definir mediante una única

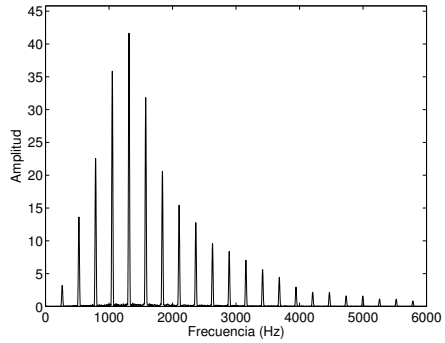
Tabla 2.3: *Relación entre sonoridad y los distintos tipos de dinámicas musicales*

Sonoridad de interpretación	Notation	Indicaciones de dinámica
Muy suave	*pp*	pianissimo
Suave	*p*	piano
Medio suave	*mp*	mezzopiano
Medio fuerte	*mf*	mezzoforte
Fuerte	*f*	forte
Muy fuerte	*ff*	fortissimo
Forzando	*sfz*	sforzando
Incremento gradual	$<$	crescendo
Decremento gradual	$>$	decrescendo

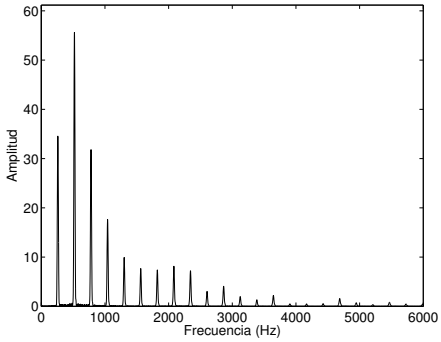
característica [PaivaPhD06], ya que hay múltiples características que afectan a la percepción auditiva, como es el grado de inarmonicidad de los parciales, el material de construcción del instrumento musical (por ejemplo, madera o metal), el tipo de instrumento musical (ej. cuerda o viento) o la excitación que produce el sonido (ej. pulsada o frotada) .

Entre las características que más afectan al timbre están el número de parciales, la envolvente espectral (relación relativa entre las amplitudes de los parciales), la envolvente temporal, el tiempo de ataque, onset asíncrono o cierta irregularidades espectrales [EronenMSc01] [Eronen01] [Zhang03].

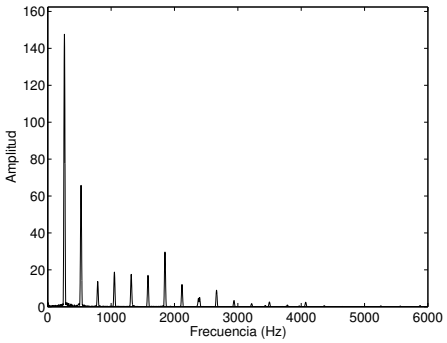
En la figura 2.9, se muestra el espectro de la nota musical $C4$ para 5 instrumentos distintos. Como se puede apreciar, el espectro presenta el mismo pitch, sonoridad y duración, pero el timbre depende del instrumento que genere la nota. Se puede apreciar también que los instrumentos de viento (figuras 2.9(a) y 2.9(b)) tienen mayor regularidad espectral que los de cuerda (figuras 2.9(c) y 2.9(d)). Otros instrumentos presentan espectros muy característicos, como en el caso del clarinete (figura 2.9(e)) que presenta nulos en los primeros parciales pares para notas de baja frecuencia [FletcherBook98].



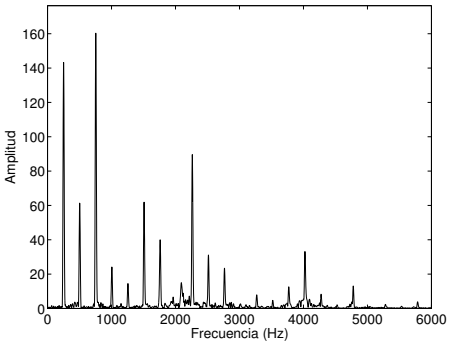
(a) Espectro de trompeta spectrum



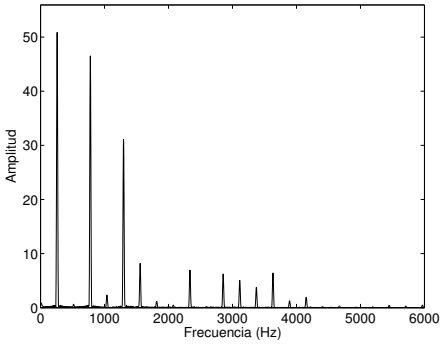
(b) Espectro de trompa



(c) Espectro de piano



(d) Espectro de violín



(e) Espectro de la nota E♭ de violín

Figura 2.9: *Espectro de la nota C4, con* $F_0 = 261{,}6\,Hz$*, tocada con la misma sonoridad y duración por diferentes instrumentos de la base de datos [Iowa06]. Cada espectro se caracteriza por el número de parciales y la relación que existe entre sus amplitudes (envolvente espectral).*

2.3. Clasificación de los sonidos

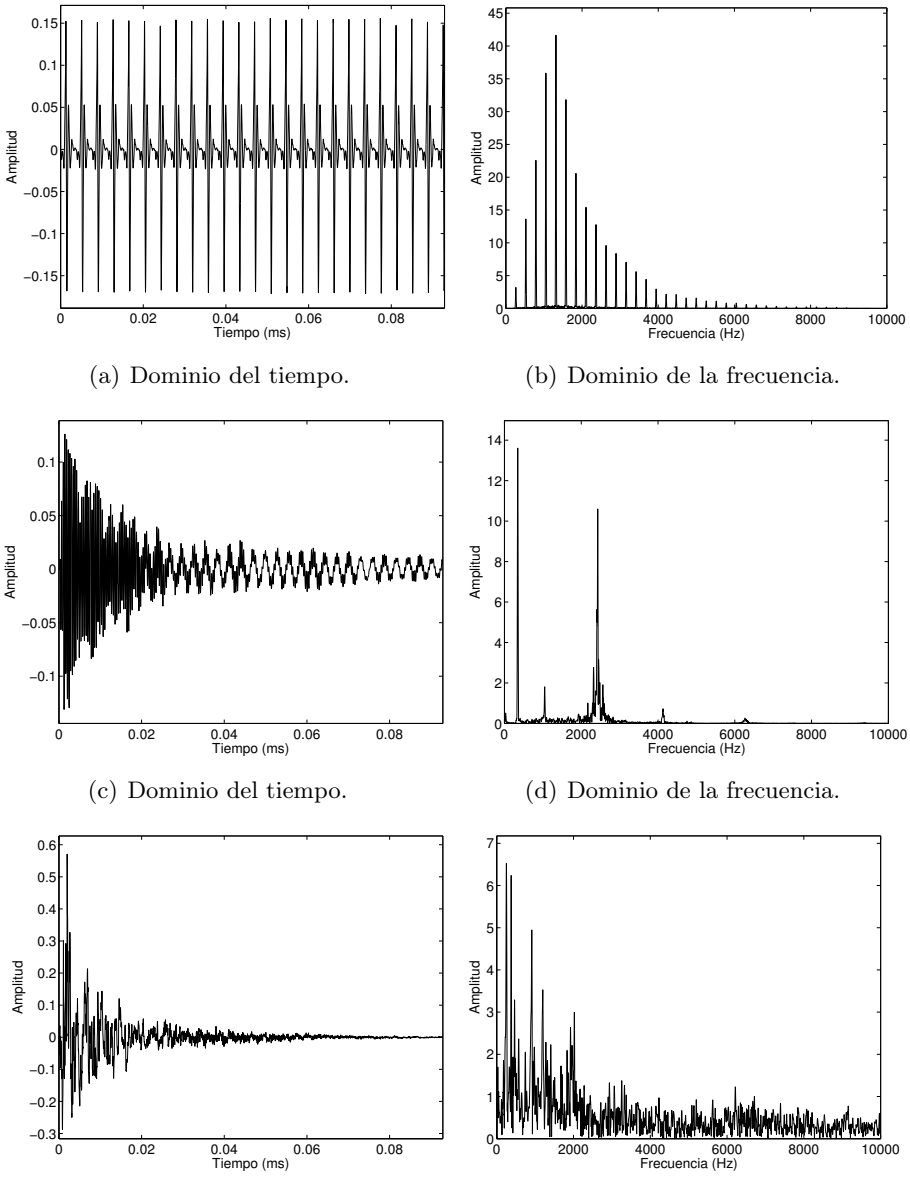
Los distintos sonidos presentes en el mundo real se pueden clasificar según diversos criterios, como son el número de canales, grado de periodicidad, número de eventos simultáneos o número de timbres musicales.

2.3.1. Sonidos armónicos

Un **sonido armónico**, (periódico o cuasi-periódico), está compuesto por una serie de sinusoides de frecuencias armónicamente relacionadas entre sí [PaivaPhD06]. Esto implica que los sonidos armónicos presenten una estructura espectral donde las componentes frecuenciales están espaciadas regularmente cada $f_0(Hz)$ [KlapuriPhD04]. Las figuras 2.10(a) y 2.10(b) muestran un sonido armónico de la nota $C4$ tocada por una trompeta en los dominios temporal y frecuencial. En la primera figura, se muestra una señal cuasi-periódica con periodo fundamental $T_0 = 3{,}8ms$, mientras que en la segunda se muestra el espectro donde los picos en frecuencia asociados a los parciales están separados $f_0 \approx 261{,}6\,Hz$ unos de otros.

Un **sonido inarmónico** no es periódico ni cuasi-periódico. En música hay dos grandes grupos de sonidos inarmónicos: los sonidos producidos con maza (ej. xilófono, vibráfono) y los producidos con batería (ej. tambor, bombo, caja). El primer grupo se caracteriza por presentar una pequeña periodicidad en el dominio del tiempo, la cual es suficiente para generar claramente un pitch en frecuencia. Por otro lado, los instrumentos del segundo grupo generan sonidos sin ninguna periodicidad en el tiempo, por tanto no presentan ningún pitch en frecuencia [KlapuriPhD04] [PaivaPhD06]. Las figuras 2.10(c) y 2.10(d) muestran un sonido producido por un instrumento de percusión con maza (xilófono) en ambos dominios, temporal y frecuencia, respectivamente. Las figuras 2.10(e) y 2.10(f) muestran un sonido producido por un instrumento de batería. Se puede ver que la percusión con maza genera algo de periodicidad en el tiempo y un pitch marcado, por el contrario la percusión de batería no presenta periodicidad y, por tanto, tampoco presenta pitch.

La tabla 2.4 incluye los instrumentos musicales occidentales que producen, o no, sonidos armónicos.



(a) Dominio del tiempo.

(b) Dominio de la frecuencia.

(c) Dominio del tiempo.

(d) Dominio de la frecuencia.

(e) Dominio del tiempo.

(f) Dominio de la frecuencia.

Figura 2.10: *Señales de sonidos armónicos e inarmónicos con una duración de* $92ms$*. Las figuras (a) y (b) muestran una señal armónica correspondiente a la nota* $C4$ *tocada por una trompeta. Las figuras (c) y (d) muestran una señal inarmónica de la nota* $F4$ *tocada por un xilófono. Las figuras (e) y (f) muestran un sonido percusivo tocado por una caja.*

Tabla 2.4: *Instrumentos de música occidental que producen, o no, sonidos armónicos[KlapuriPhD04]*

Sonido producido	**Familia de instrumentos**	**Instrumentos**
Armónicos o cuasi-armónicos	Instrumentos de cuerda	Piano, guitarra, violin, chelo, etc.
	Lengüeta	Clarinete, saxofón, oboe, fagot
	Viento-metal	Trompeta, trombón, tuba, trompa
	Flautas	Flauta, flautín, etc.
	Órganos de tubos	Órganos de tubos
	Voz humana (cantada)	Cuerdas vocales
No armónicos	Percusión de maza	Marimba, xilófono, vibráfono, carrillón
	Baterías	Baterías y platillos

Propiedades físicas de instrumentos armónicos.

Aunque la representación en series de Fourier (ecuación (2.14)) aporta una aproximación del modelo de señal para obtener la frecuencia fundamental de un instrumento armónico, surgen varios inconvenientes. La generación de sonidos por los instrumentos musicales supone la aparición de fenómenos físicos complejos, los cuales aportan diversas características espectrales a los sonidos. Además la interacción del sonido generado con la sala donde se produce incrementa el número y la complejidad de estos fenómenos físicos.

Según [YehPhD08], la generación de sonidos de instrumentos tiene cuatro partes: generador de excitación, resonador, radiación y acústica de sala (ver figura 2.11). Un resumen de las propiedades físicas del sonido instrumental para cada parte se detalla a continuación

Figura 2.11: Generación de sonidos musicales de instrumentos [YehPhD08]

1. **Excitación**

La excitación de un instrumento musical puede ser mecánica o con vibradores acústicos. La excitación, por sí misma o junto con el resonador, caracterizan el espectro del sonido, en ese punto se denomina modo normal de vibración. En los instrumentos armónicos, el modo normal genera frecuencias muy cercanas a los múltiplos de f_0.

La excitación de instrumentos de viento-madera se produce al hacer pasar una corriente de aire por unas lengüetas simples o dobles. Los instrumentos de viento-metal generan el sonido de manera similar, pero los elementos vibrantes son los labios.

La excitación de instrumentos de cuerda se produce golpeando, frotando o pulsando las cuerdas. Si una cuerda se excita a $1/h$ de su longitud desde uno de los extremos, el h-ésimo armónico en el modo normal se elimina. Cuando una cuerda es golpeada por un martillo, como en el caso del piano, los pulsos reflejados desde ambos extremos de la cuerda interactúan con el martillo generando efectos complejos que hacen que el espectro de vibración decaiga más rápidamente que en los casos de cuerda punteada.

En el caso de las cuerdas pulsadas, las frecuencias de los parciales resultantes se pueden calcular con el metido propuesto en [FletcherBook98] con la ecuación (2.15),

$$f_k = m f_0 \sqrt{1 + \varepsilon(h^2 - 1)} \tag{2.15}$$

donde m es el número de armónico y ε es el coeficiente de inarmonicidad. Los valores típicos para este coeficiente se encuentran en el rango desde 10^{-3} a 10^{-4} en las notas graves de piano [Conklin99].

El fenómeno de **inarmonicidad** está causado por la rigidez de las cuerdas reales, la cual se hace presente como una fuerza de recuperación de la tensión de la cuerda [OrtizPhD02]. En el caso ideal de una cuerda vibrante, cuando la longitud de onda es mucho mayor que el grosor de la cuerda, la velocidad de propagación de la onda es constante y los parciales se sitúan en los armónicos. Sin embargo para parciales de alta frecuencia, con una longitud de onda muy pequeña,

una cuerda fina se comporta como una barra gruesa de metal. La resistencia mecánica de la cuerda a curvarse se convierte en una fuerza adicional. A menos que la resistencia a la curvatura sea mucho menor que la tensión de la cuerda, se produce un incremento de la velocidad de propagación. Este fenómeno provoca que las frecuencias de los parciales se sitúe por encima de la de los armónicos ideales, generando de esa manera el efecto de la inarmonicidad. La figura 2.12 ilustra el fenómeno de la inarmonicidad para el caso del piano (instrumento de cuerda pulsada). Se aprecia en la figura que las frecuencias de los parciales se separan de los múltiplos de la frecuencia fundamental.

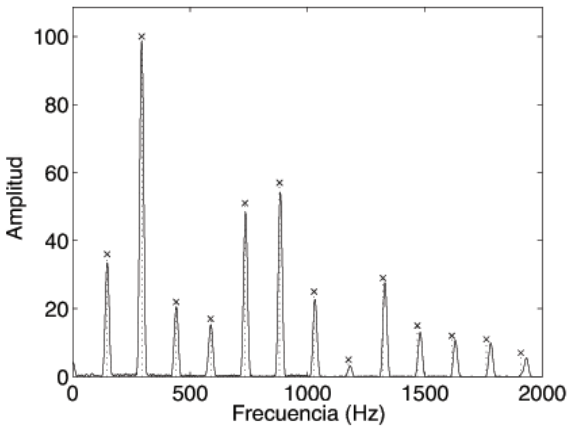


Figura 2.12: *Nota D3 tocada con piano de la base de datos [Iowa06]. La línea continua representa el espectro real de la nota afectada por la inarmonicidad. Los símbolos 'x' indican la posición ideal de cada armónico. Se aprecia como se produce una desviación de los parciales hacia frecuencias más altas conforme se va subiendo en frecuencia.*

2. **Resonancia**

   El efecto del generador de excitación provoca una vibración en el resonador. Este conjunto vibratorio es el que establece los regímenes de vibración que tendrán las ondas. Un **régimen oscilatorio** es el esta-

do en el que el conjunto vibratorio mantiene una oscilación constante de componentes armónicamente relacionados [YehPhD08].

En los instrumentos que usan los labios del músico como elemento vibrante, las frecuencias de resonancia y las amplitudes de los parciales dependen del conjunto formado por la boquilla y el cuerpo del instrumento [FletcherBook98]. El cuerpo del instrumento determina las frecuencias resonantes, mientras que la boquilla determina la amplitud máxima de la envolvente. El rango de notas que pueden ser producidas por el instrumento puede ser determinado mediante el análisis de la impedancia del cuerpo del instrumento. De esta manera, las posibles notas son aquellas cuya frecuencia fundamental f_0 no supera la impedancia máxima del cuerpo del instrumento. En una situación normal, la frecuencia fundamental f_0 y algunos de los armónicos coinciden con las frecuencias de resonancia más significativas del cuerpo del instrumento, pero el resto de los parciales que no coincidan con un máximo de la impedancia tendrán una amplitud de oscilación muy pequeña. A pesar de la baja amplitud para estos parciales, y de la aproximación no lineal, se puede considerar, en general, la amplitud del h-ésimo parcial como la h-ésima potencia inversa de la amplitud del primer parcial. En análisis de señales musicales de alto nivel, los parciales altos tendrán más importancia en el modelo cuanto más alta sea la amplitud relativa a la del parcial fundamental.

Los instrumentos de viento-madera tienen como característica que el cambio de frecuencia fundamental, y por tanto de nota tocada, lo realizan abriendo o cerrando determinados agujeros en el cuerpo del instrumento para modificar la longitud del tubo resonante [FletcherBook98][YehPhD08]. Las curvas de impedancia de los registros (notas) bajos pueden variar respecto de los de alta frecuencia, por el cambio físico que se produce en el cuerpo resonante. Los clarinetes, por ejemplo, presentan un espectro muy característico en las notas de baja frecuencia. Para estas notas se produce una ausencia casi completa del segundo y cuarto parcial [FletcherBook98]. Este efecto se produce porque los picos de resonancia máxima del instrumento en notas de baja frecuencia coinciden con la serie de armónicos impares

y sólo los primeros armónicos coinciden bien con los picos de impedancia máxima del tubo. Por ello, los parciales impares de la nota más baja tienen una caída de 3 dB/octava, mientras que los parciales pares caen con una pendiente de 6 dB/octava.

Normalmente la flauta produce notas en los picos de resonancia del cuerpo del instrumento más altos en frecuencia. Cuando se toca la flauta la corriente de aire oscila a una frecuencia particular. En concreto, si la vibración es grande, como cuando se toca una nota con fuerza, la flauta genera una gran cantidad de parciales. Para notas bajas los primeros parciales se generan por ondas estacionarias. Por el contrario, para notas altas, la resonancia en la flauta no es armónica, por lo que sólo un pequeño número de parciales (sólo un parcial en la tercera y cuarta octava) son soportados por la resonancia del cuerpo de la flauta.

3. **Radiación**

   Las ondas generadas por el conjunto de excitador y resonador están compuestas por una gran variedad de modos normales. Los instrumentos musicales tienen un modo complicado de radiar los sonidos para que puedan llegar a los receptores de audio (oídos, micrófonos, etc.). La investigación en la radiación de los instrumentos musicales se basa en los modelos mas simples de fuentes generadoras de onda: monopolios, dipolos y fuentes multipunto. Por tanto los modos de radiación de los instrumentos se pueden aproximar mediante la combinación de estas fuentes simples.

   Cuando los cantantes cambian su estilo de barroco a romántico, lo que hacen es modificar la cantidad de energía que es radiada desde el cuerpo y no sólo desde la boca. La naturaleza y la calidad de una guitarra depende mucho de las áreas de radiación de su superficie [FletcherBook98]. La geometría compleja del grand piano permite obtener unos patrones de radiación muy diferenciados y de ahí se genera la gran fascinación de los músicos por este instrumento. Para simplificar la radiación de los instrumentos desde los agujeros abiertos de un instrumento de viento o el cuerpo de resonancia de un piano, se puede tratar como un conjunto de fuentes puntuales. A la hora de radiar

un sonido, los patrones de directividad dependen directamente de la frecuencia de la nota, cada nota presenta una eficiencia de radiación distinta.

En general, el sonido radiado se comporta de manera más direccional para altas frecuencias [YehPhD08]. Los parciales bajos de instrumentos de metal se distribuyen de manera uniforme alrededor de la campana del instrumento en todas direcciones, mientras que los parciales de alta frecuencia forman un haz más directivo que sigue la dirección del eje del cuerpo del instrumento. Aunque la curva de resonancia de los instrumentos de viento es progresivamente más débil para los parciales más altos, la alta eficiencia de radiación (en la dirección del eje del instrumento) en estos parciales compensa la pérdida de energía generada en los parciales altos. Los parciales pares de una nota de clarinete se radian de manera más intensa que los parciales impares, lo cual compensa la débil resonancia en los parciales impares.

4. **Acústica de sala**

   La acústica de la sala describe el comportamiento del sonido en un espacio cerrado donde determinados rangos de frecuencias pueden ser reforzados por los modos normales de resonancia de la sala. Las frecuencias no atenuadas de las ondas vibratorias en un prisma rectangular de aire rodeado de superficies reflectantes se pueden describir mediante la siguiente ecuación [YehPhD08][Rayleigh45]

$$f_r = \frac{c}{2}\sqrt{\frac{n_x^2}{L_x^2} + \frac{n_y^2}{L_y^2} + \frac{n_z^2}{L_z^2}} \tag{2.16}$$

   donde c es la velocidad del sonido en el aire, L_x, L_y y L_z son las dimensiones del prisma rectangular y n_x, n_y y n_z determinan los modos de vibración: modo axial (dos superficies), modo tangencial (cuatros superficies) y modo oblicuo (seis superficies).

   La **reverberación** es la persistencia de un sonido en un espacio cerrado después de que la fuente de sonido termine de producirlo. El sonido principal es seguido de reflexiones rápidas y posteriormente una densa serie de reflexiones llamadas reverberantes. Este fenómeno desaparece

lentamente conforme la energía del sonido es absorbida por los muros y el aire. La duración de esta absorción del sonido o tiempo de reverberación, es un parámetro a tener en cuenta en el diseño de grandes salas que necesitan tener tiempos de reverberación determinados en función de la actividad que vaya a desarrollarse en su interior (para salas de conciertos u ópera se necesitan tiempos de reverberación entre 1,5 y 2 segundos).

2.3.2. Señal monoaural, estéreo y multicanal

Los **sonidos monoaurales** (también llamados mono) derivan de señales mono-canal. Estas señales son las más complicadas de analizar desde el punto de vista del procesado de señal, puesto que no aportan ningún tipo de información espacial de las fuentes [LiPhD08]. Las señales mono están compuestas directamente por la suma de todas las fuentes de sonido presentes en la grabación.

Los **sonidos estereofónicos**, derivan en señales de audio de dos canales, son aquellos que son grabados con, al menos, dos micrófonos independientes. Las señales estéreo aportan información espacial de la localización relativa de las fuentes de sonido. La música comercial se distribuye en formato de señales estéreo [Ryynanen08a].

Los **sonidos multicanal**, también conocidos como sonido *surround*, son aquellos generados a partir de, al menos, cuatro canales de audio independientes. Por ejemplo, los sistemas de reproducción surround 5.1 están compuestos por cinco altavoces. Los sistemas 5.1 tienen canales izquierdos y derechos, como en el caso estéteo, y además un canal central para conversaciones de las películas o el vocalista en el caso de la música y para efectos especiales y sonido surround. Un canal adicional *subwoofer* (el .1 o canal LFE, Low Frequency Effects) añade sonido de muy baja frecuencia para fuentes musicales y efectos especiales en películas.

2.3.3. Sonidos monofónicos y polifónicos

Las piezas musicales pueden ser monofónicas, teniendo un solo instrumento que toque sólo una nota en cada instante temporal (ej. solo de trompeta), o polifónicas si hay uno o más instrumentos que toquen más de una

nota en cada instante temporal (ej. piano o una orquesta) [Plumbley02].

Los **sonidos monofónicos** son aquellos en los que solamente aparece una frecuencia fundamental f_0 activa en cada instante [Plumbley02]. La figura 2.13(b) muestra el espectro de un sonido con $f_0 \approx 261{,}6\,Hz$.

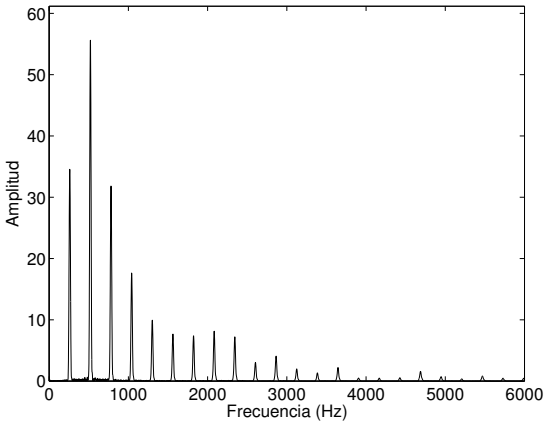
Los **sonidos polifónicos** son sonidos con dos o más frecuencias fundamentales f_0 activas al mismo tiempo. Estas f_0 pueden ser producidas por uno o más instrumentos [Plumbley02]. En la figura 2.13(b) se muestra el espectro de un sonido polifónico compuestao por cuatro frecuencias fundamentales $f_{01} = 155{,}6\,Hz$, $f_{02} = 329{,}6\,Hz$, $f_{03} = 392{,}0\,Hz$ and $f_{04} = 554{,}4\,Hz$. Se puede apreciar una clara dificultad para distinguir los pitches de cada nota cuando el nivel de polifonía crece. Un caso muy común en la música occidental es la concurrencia de varias notas con una relación racional en su frecuencia, lo que provoca que algunos de sus parciales tengan posiciones en frecuencia coincidentes, se solapen e interfieran en amplitud y fase. Esta compleja suma de parciales puede provocar interferencias constructivas o destructivas en función de la relación de las fases de los parciales. En el caso de interferencias constructivas, el pico espectral, resultante de la suma de ambos, tendrá más energía que la de cada uno de los parciales de manera independiente. En el caso de interferencia destructiva, la energía del pico espectral resultante puede llegar a ser nula.

2.3.4. Sonidos monotímbricos y multitímbricos

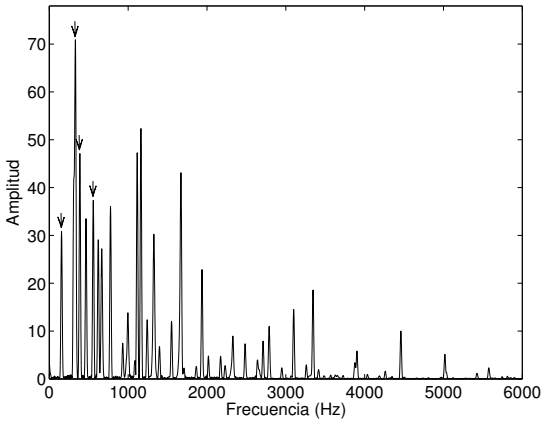
Dependiendo del número de instrumentos, una interpretación musical puede ser considerada monotímbrica o multitímbrica.

El concepto de **monotímbrico** se relaciona con las piezas musicales que son interpretadas por un instrumento con un único timbre o envolvente espectral. Por otro lado el concepto de **multitímbrico** se relaciona con las piezas musicales en las que participan dos o más instrumentos, cada uno de ellos con un timbre distinto.

En las figuras 2.14(a) y 2.14(b) se muestra el espectro de un sonido monotímbrico y otro multitímbrico respectivamente. Se puede ver que el espectro multitímbrico es más complejo de analizar puesto que no se puede identificar el número de parciales representativos de cada nota ni las relaciones e amplitud entre parciales adyacentes para componer la envolvente



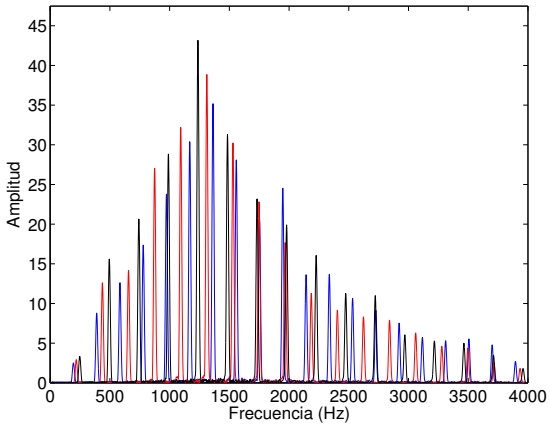
(a) Espectro de sonido monofónico



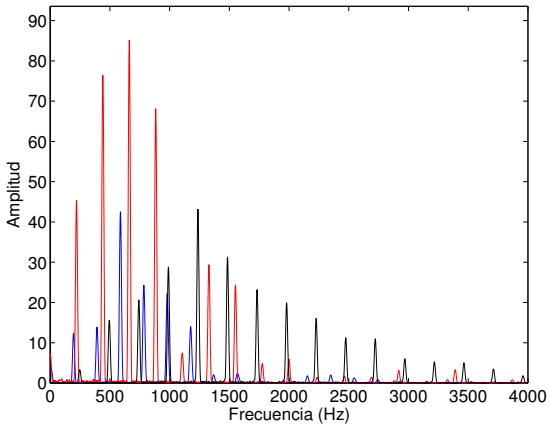
(b) Espectro de sonido polifónico

Figura 2.13: *Espectro de sonidos en función del número de pitches activos: (a) Sonido monofónico en el que está presente la nota C4 tocada por un fagot de la base de datos [Iowa06]; (b) Sonido polifónico compuesto por cuatro notas musicales (D#3 (Saxo Alto), E4 (Flauta), G4 (Clarinete E♭) and C#5 (Oboe) de la base de datos [Iowa06]). Las frecuencias fundamentales se indican con asteríscos '*'.*

espectral



(a) Espectro monotímbrico.



(b) Espectro multitímbrico.

Figura 2.14: *Espectro de sonido en función del número de instrumentos presentes: (a) Sonido monotímbrico en el que se muestran notas C3 (azul) y G3 (rojo) de piano. Ambas notas tienen el mismo patrón espectral que define el timbre del instrumento. (b) Sonido multitímbrico en el que la nota G3 (azul) B3 (rojo) y A3 (negro) son tocadas por una flauta, un piano y una trompeta, respectivamente, de la base de datos [Iowa06]. Cada nota tiene un patrón espectral distinto por ser producido por distintos instrumentos.*

2.4. Estimación *multi-pitch*

La **estimación *multi-pitch*** de diferentes sonidos concurrentes en señales polifónicas es uno de los primeros y principales problemas abordados

por el procesado digital de señal, por estar presente en muchas de las aplicaciones de dicho procesado, como la transcripción automática, la recopilación de información musical y el contenido musical de las señales.

2.4.1. Definición

La estimación *multi-pitch* en señales musicales polifónicas es un problema complicado a la vez que interesante que ha atraído la atención de muchos investigadores a lo largo de las últimas décadas y aún no ha sido resuelto de manera definitiva. Debido a las altas tasas de error producidas cuando el nivel de polifonía se incrementa, es un campo de investigación abierto a nuevas propuestas y mejoras de las soluciones actuales.

Las señales más apropiadas para evaluar la estimación *multi-pitch* son las señales polifónicas, de la misma manera que las señales de voz son las más apropiadas para estimadores mono-pitch [Christensen07]. En una señal polifónica, la estimación *multi-pitch* supone estimar el número de pitches activos en cada instante, deduciendo con ello el número de fuentes de audio y estimando la f_0 de cada fuente y la amplitud de cada componente armónica [YehPhD08].

La estimación *multi-pitch* es la tarea principal en un sistema de transcripción automática de música, sin embargo, la transcripción automática supone incluir otras tareas de más alto nivel, como por ejemplo: estimación de tempo, métrica y clave de cada sección de la pieza; identificación y etiquetado de la ornamentación musical; reconocimiento de los instrumentos que están presentes; y la segregación de las voces (ej. la melodía y el acompañamiento).

2.5. Transcripción Automática de Música

A lo largo de la historia, la transcripción de composiciones musicales ha sido llevada a cabo por músicos experimentados. De hecho, una persona sin ningún tipo de educación musical no está capacitado para transcribir música polifónica, en la que diferentes sonidos están activos a la vez [KlapuriMSc97]. Cuanto más rica es la polifonía de una señal musical, más complicada es su transcripción y más experimentada debe ser la persona que la lleve a

cabo. Esto lleva a definir la necesidad de un sistema que permita realizar de manera automática la transcripción de una composición musical.

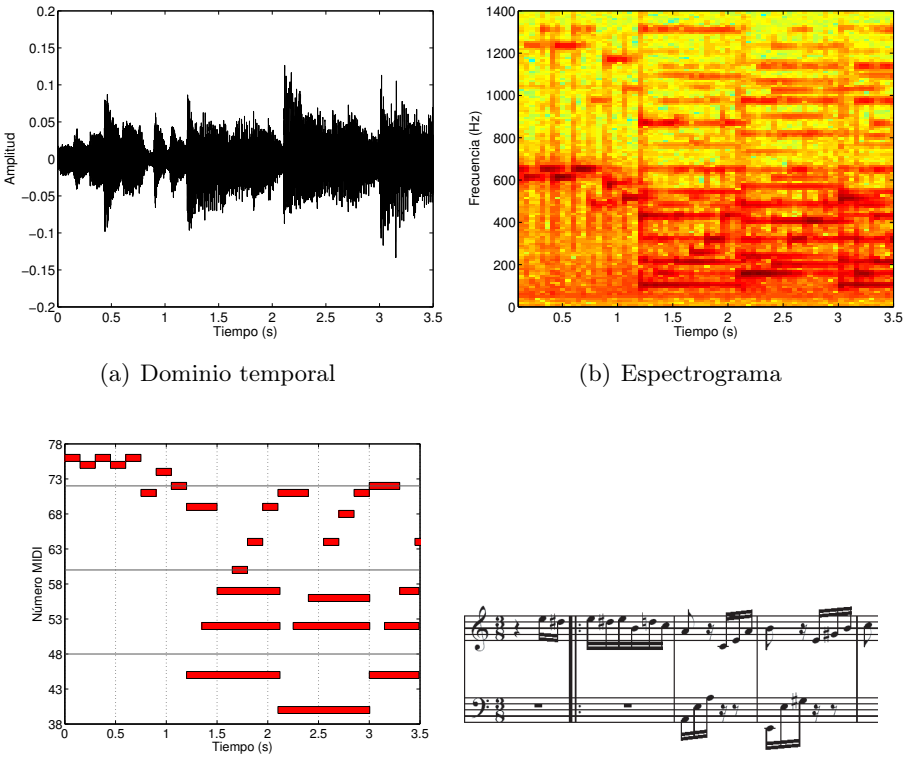
2.5.1. Definición

La **transcripción musical** *(Automatic Music Transcription, AMT)* se define como la acción de escuchar una pieza musical y escribir en notación musical las notas que la constituyen [Martin96a]. En otras palabras, consiste en transformar una señal acústica en una representación simbólica, indicando notas, pitches, tiempos y clasificación de los instrumentos presentes.

En el significado tradicional, transcribir una pieza conlleva un conjunto de tareas de alto nivel como: estimar el pitch y tiempos de activación de cada nota, estimación del tempo, métrica y clave de la interpretación o identificación de instrumentos. Estas tareas son muy complejas desde el punto de vista de la acústica y el procesado de señal, además de incrementarse la complejidad con el incremento del nivel de polifonía de la composición

El concepto de transcripción automática de música tiene múltiples definiciones. En [KlapuriBook06] y [Ryynanen08a], la AMT se determina por el análisis de una señal acústica para anotar el *pitch*, tiempo de *onset*, duración y la fuente de cada sonido que tenga lugar en ella. La estimación de estos parámetros se representa mediante símbolos que indican información de *pitch*, *onset* y duración de cada nota. Para evaluar los sistemas de transcripción se emplea el formato MIDI, un protocolo estándar que permite la comunicación entre elementos electrónicos musicales. Otra definición, dada en [Plumbley02], establece que el objetivo de la AMT es analizar una señal de audio para encontrar los instrumentos y las notas que son tocadas, a la vez que produce una escritura simbólica de la pieza musical. Una tercera definición se da en [Bello06], donde dice que la AMT es el proceso de convertir una grabación musical en una notación simbólica u otra representación similar de la interpretación. La figura 2.15 muestra algunas de las representaciones de la AMT, dependiendo del dominio usado, de la misma señal de audio.

Un sistema de transcripción perfecto debe ser capaz de estimar, para cada nota, no sólo los parámetros anteriormente comentados (*pitch*, *onset* y duración), sino también la sonoridad y los instrumentos que han generado



(a) Dominio temporal

(b) Espectrograma

(c) Piano-roll

(d) Notación simbólica

Figura 2.15: *Señal musical de* 3,5*s de la pieza "Fur Elise" de Ludwing van Beethoven. (a) Representación de la señal en el dominio temporal. (b) Representación tiempo-frecuencia en la que la energía de una frecuencia se representa en mediante la intensidad del color (más intensidad en colores más cálidos). (c) Representación MIDI. Cada nota se indica por un rectángulo rojo. El eje horizontal representa el tiempo, el eje vertical indica el número MIDI de la nota tocada. (d) Notación musical clásica de la pieza interpretada.*

cada nota.

2.6. Separación de fuentes sonoras

El ser humano, en una situación en la que recibe ondas sonoras generadas por diferentes fuentes, como puede ser una conversación en un ambiente

ruidoso, o un concierto de música, es capaz de discriminar la fuente sonora a la que quiere prestar atención aislando el resto a pesar de continuar escuchándolas. Esta tarea de discriminar, siguiendo determinados criterios o patrones, es la que pretende llegar a conseguir la separación de fuentes de audio. La capacidad del oído humano de realizar esta discriminación se denomina *Análisis de la escena auditiva* (Auditory Scene Analisys, ASA) [ZwickerBook90]. Los primeros estudios sobre ASA se encuentran en [BregmanBook90]. Conseguir automatizar ASA no es trivial, a día de hoy, es tema de estudio en diversos grupos de investigación a nivel mundial y ha dado lugar al concepto de *Análisis computacional de la escena auditiva* (Computational Auditory Scene Analisys, CASA).

2.6.1. Definición

La **separación de fuentes sonoras** *(Sound Source Separation, SSS)* pretende recuperar las señales de audio que han sido generadas por varias fuentes y se encuentran mezcladas en una (señal monoaural) o varias señales (señal multicanal). La separación de fuentes no sólo consta de la detección de las notas y su asignación a una de las fuentes, también hay que sintetizar la señal correspondiente a cada una de ellas.

La SSS se puede dividir en dos grandes grupos: la separación de fuentes instrumentales [Every06] [Burred07] y la extracción de la señal de voz (singing voice) [LiPhD08]. En la actualidad hay menos trabajos en el campo de la extracción de la señal de voz, aunque hay muchas posibles aplicaciones como el reconocimiento y alineamiento de la letra [Wang04] [Mesaros10] o la identificación del cantante [Mesaros07]. Sin embargo, esta tesis se centra en el otro grupo, en la separación de fuentes instrumentales [Morita06] [Virtanen07b].

Atendiendo a otro criterio, la SSS puede clasificarse en función de la información previa con la que cuente, siendo posibles múltiples combinaciones con distintos tipos de información. La separación que no cuenta con ningún tipo de información se denomina Separación de Fuentes a Ciegas *(Blind Source Separation, BSS)*. A día de hoy, esta separación aporta resultados prometedores en el caso de separación de señales multicanal [Ozerov10], pero no es así en el caso de señales monoaurales [Virtanen07b] [ComonBook10],

como son las que se tratan en esta tesis. La BSS ha sido estudiada durante años para las señales monoaurales [Parra98b], pero nunca ha conseguido dar resultados convincentes. La comunidad científica ha apostado por agregar distintas fuentes de información al sistema de separación para que sea capaz de obtener señales separadas de mayor calidad. Ésto hace que ya no se trate de BSS, sino de separación de fuentes informada *(Informed Source Separation, ISS)*. La ISS es muy amplia en cuanto a la tipología y morfología de la información que se aporta al sistema. La información puede ser temporal, como en el caso de información de simbólica de *score (Score-Informed Source Separation)*[Ewert12], o espectral, como en el caso de uso de modelos espectrales de las fuentes [Fritsch13]. Se pueden usar ambos tipos de información en el mismo sistema, como se ha comentado anteriormente las señales monoaurales precisan de información adicional para poder obtener una separación de calidad aceptable.

2.6.2. Aplicaciones de la SSS

Hay muchas aplicaciones potenciales para la separación de fuentes. Por ejemplo, puede ser una herramienta muy útil para estudiantes de música. La capacidad de segregar las fuentes instrumentales, de una señal que contenga varias de ellas, puede simplificar enormemente el estudio y aprendizaje de las piezas musicales por parte de los músicos. Así mismo, puede resultar de gran utilidad la supresión de uno de los instrumentos de una grabación para que el alumno pueda tocar en conjunto con los demás y de esa manera perfeccionar la ejecución de cada pieza.

La SSS puede considerarse igualmente como una potente herramienta de preprocesado para la transcripción automática de música. La AMT tiene serias complicaciones para realizar la transcripción por instrumento en el caso de señales polifónicas y multitimbricas. Una etapa de preprocesado que separe las fuentes presentes en la señal para una transcripción posterior de cada una de manera independiente puede potenciar enormemente los resultados obtenidos si se comparan con la transcripción directa de la señal mezclada.

Las tecnologías de comunicaciones, también pueden beneficiarse del uso de SSS. En señales de conversaciones telefónicas, la separación de la voz y

el ruido ambiental añadido por el entorno, es un procesado muy necesario para conservar la inteligibilidad de la conversación y, por tanto, la utilidad de la llamada telefónica. En otras ocasiones, además del ruido, pueden presentarte otros fenómenos como el eco y la reverberación, los cuales pueden afectar a la inteligibilidad o procesos de reconocimiento de voz y transcripción musical. Sería deseable, para dichas aplicaciones la separación de la fuente principal de los efectos generados, que pueden considerarse como otras fuentes secundarias presentes en la señal.

Otra aplicación de la SSS es la adaptación de señales monoaurales para ser reproducidas en sistemas 5.1 sin necesidad de acudir a las fuentes originalmente grabadas. Igualmente, la codificación y compresión de fuentes individuales puede demostrar mayor potencia que en el caso de una señal de distintas fuentes mezcladas, siguiendo la filosofía de objetos audiovisuales descrita por el estándar MPEG-4 [Mitianoudis02].

2.7. Conclusiones

En este capítulo se han introducido los conceptos musicales básicos para comprender el funcionamiento de un sistema de separación de fuentes y las consideraciones que lleva consigo. Se han presentado las características perceptuales de cualquier sonido, así mismo se han expuesto varias clasificaciones para ellos, atendiendo a diversos criterios de clasificación. A continuación se han descrito algunos sistemas de procesado de señal musical relacionados con la separación de fuentes, como son la estimación *multi-pitch* y la transcripción automática de música. Finalmente se ha definido el concepto de Separación de Fuentes Sonoras (SSS) y algunas de sus potenciales aplicaciones.

Capítulo 3

Modelos de descomposición de señal

3.1. Introducción

El espectrograma de una señal de audio puede descomponerse como una combinación lineal de unas determinadas funciones espectrales básicas. Siguiendo esta afirmación, se puede decir que, el espectro de amplitud de la señal $X(f,t)$ en la trama t y frecuencia f se modela como la suma ponderada de funciones base, de manera que:

$$\hat{X}(f,t) = \sum_{n=1}^{N} b_n(f) g_n(t) \tag{3.1}$$

donde $g_n(t)$ es la ganancia de la función base n en la trama t, y $b_n(f)$, $n = 1, ..., N$ son las funciones base. Desde otro punto de vista, se puede decir que la señal de entrada se modela como la suma de unas componentes con bases fijas y amplitudes variantes en el tiempo. Así mismo, la descomposición de la señal se puede describir con notación matricial de la siguiente manera:

$$\hat{\mathbf{X}} = \mathbf{BG} \tag{3.2}$$

o bien, si se considera un término de error residual r:

$$X(f,t) = \sum_{n=1}^{N} b_n(f) g_n(t) + r(f,t) \tag{3.3}$$

La interpretación práctica de los parámetros que describen el modelo depende de la aplicación. Por ejemplo, cuando se trata de instrumentos armónicos, en el contexto de transcripción musical automática, cada función base representa de manera ideal un único *pitch*, y las ganancias correspondientes contienen la información de *onset* y *offset* para cada *pitch*. En el caso de la separación de fuentes musicales, una función base y su ganancia asociada pueden representar, por ejemplo, la contribución en energía de todos los tonos con un *pitch* fundamental en concreto, o bien, un patrón espectral de algún instrumento percusivo.

En la bibliografía se pueden encontrar diversos métodos que usan este tipo de modelos, y son aplicados en: separación de fuentes[Klapuri10b], extracción de melodía [Durrieu10], transcripción musical [Vincent10, Carabias11] y reconocimiento de instrumentos musicales [Heittola09].

Algunos de los métodos de descomposición, encontrados en la bibliografía, y que usan estos modelos de señal se describen en este capítulo con distinta profundidad. Por ejemplo, *Independent Component Analysis* (ICA) [Plumbley03], *Sparse Coding* (SC) [Abdallah04] y *Non-Negative Matrix Factorization* (NMF) [Lee99]. Se realizará una descripción más profunda de las versiones basadas en NMF, por ser las empleadas por los métodos propuestos en esta tesis.

3.2. *Independent Component Analysis (ICA)*

El análisis de componentes independientes (ICA) es un método de descomposición de señal usado, generalmente, en separación de fuentes a ciegas con señales multicanal (BSS) [Nishikawa03]. Una escena típica para este tipo de problema es el comúnmente conocido como *cocktail party problem*, dónde se pretenden separar diferentes voces de distintas personas hablando al mismo tiempo y, por tanto, se encuentran mezcladas en las señales. Si se sitúan varios micrófonos en distintos puntos de la sala, cada uno recogerá una señal distinta, aunque en todas ellas estarán presentes todas las

voces, con ciertas particularidades en función de las posiciones de los hablantes y los micrófonos. Este problema se enmarca en BSS, puesto que no se conoce *a priori* ninguna información de los parámetros de la mezcla, ni información espacial de los sujetos, ni características espectrales de sus voces. Un problema análogo puede presentarse con la separación de los instrumentos musicales en una orquesta, cuando no se tiene ningún otro tipo de información adicional.

Los algoritmos ICA suponen que cada trama de cada canal, $\mathbf{x_c}(t)$, es una mezcla lineal de las señales emitidas por cada fuente, $\mathbf{g}(t)$, y lo expresan de manera que,

$$\mathbf{x_c}(t) = \mathbf{B}\mathbf{g}(t), \tag{3.4}$$

donde $\mathbf{B}$ es la matriz de mezcla con coeficientes desconocidos.

Basándose, por tanto, en la información de la señal de entrada, $x(t)$, los algoritmos ICA intentan invertir el proceso de mezcla buscando una matriz de separación óptima $\mathbf{W}$, con la que la estimación de las fuentes separadas sería

$$\hat{\mathbf{g}}(t) = \mathbf{W}\mathbf{x}(t) \tag{3.5}$$

Idealmente, la matriz de separación, $\mathbf{W}$, sería la inversa de la matriz de mezcla, $\mathbf{B}$, y las fuentes reconstruidas idénticas a las señales independientes de cada fuente. Por consiguiente, los métodos ICA pretenden buscar una matriz $\mathbf{W}$ que reconstruya fuentes que sean entre ellas lo más independientes posible. Según el teorema central del límite, la suma de dos o mas variables aleatorias independientes tiene una distribución que se puede considerar más Gaussiana que cualquiera de las variables aleatorias iniciales por separado [Hyvarinen00]. Por tanto, si se reconstruyen las fuentes orginales con distribuciones que sean lo menos Gaussianas posible, un algoritmo ICA podría, potencialmente, recuperar las fuentes originales. La entropía negativa y la curtosis son unas medidas ampliamente utilizadas de la no Gaussianidad. Por ejemplo, los valores de curtosis serán positivos (para distribuciones con un gran pico en torno a cero y amplias colas), negativos (para distribuciones estrechas que son bastante constantes en torno a cero y, a la vez, tienen largas colas con pequeña amplitud), o cero para el ca-

so de distribuciones Gaussianas. De este modo, los algoritmos ICA pueden diseñarse para modificar **W** de manera que las fuentes reconstruidas obtengan un valor absoluto de curtosis alto hasta que obtengan un conjunto de valores óptimos.

Algunos algoritmos propuestos, que siguen la filosofía ICA son: [BarryMSc03], que usa una variante probabilistica basada en la estimación de máxima similitud; Plumbey propone varios algoritmos para implementar un ICA no negativo. En [Plumbley01] se propone un algoritmo basado en una red neuronal, mientras que en [Plumbley03] se proponen algoritmos basados en un modelo de Análisis de Componente Principal (PCA) no lineal.

3.3. *Sparse Coding*

Los algoritmos de *Sparse coding* representan la señal mezclada de entrada escogiendo un número reducido de elementos básicos de entre un gran conjunto de ellos [Olshausen97, KlapuriBook06]. Esta opción es usada en la bibliografía para el aprendizaje de estructuras o bases y para la separación de datos con distinto origen. En el modelo lineal de señal representado en la ec.(3.1), la restricción de dispersión se suele aplicar sobre las ganancias **G**, lo que implica una probabilidad alta de que un elemento de **G** sea cero. Consecuentemente, solo unas pocas componentes de **G** se activan en cada instante y durante periodos cortos de tiempo. Esta restricción encaja bien con la idea de que, en música, es habitual que sólo un pequeño conjunto de las posibles notas suenan simultáneamente.

En esta sección se presenta una breve descripción de un encuadre probabilístico para este tipo de algoritmos. Se puede encontrar una descripción extendida en [KlapuriBook06].

En este esquema de descomposición de señales, la fuentes y las matrices de mezcla se estiman maximizando las distribuciones de probabilidad resultantes de cada una de las fuentes estimadas. En concreto, la distribución de probabilidad de **B** y **G** dado un espectrograma de la señal de entrada **X** se define en [KlapuriBook06] como:

$$\max_{\mathbf{B},\mathbf{G}} p(\mathbf{B},\mathbf{G}|\mathbf{X}) \propto \max_{\mathbf{B},\mathbf{G}} p(\mathbf{X}|\mathbf{B},\mathbf{G})p(\mathbf{B},\mathbf{G}) \tag{3.6}$$

donde $p(\mathbf{X}|\mathbf{B},\mathbf{G})$ es la probabilidad de observación de $\mathbf{X}$ dado $\mathbf{B}$ y $\mathbf{G}$, y $p(\mathbf{B},\mathbf{G})$ es la probabilidad conjunta de $\mathbf{B}$ y $\mathbf{G}$.

Para facilitar el cálculo matemático, se puede asumir el término de ruido de la ec.(3.3), que es independiente e uniformemente distribuido (i.i.d.), en el modelo $\mathbf{BG}$, y que tiene una distribución normal con varianza σ^2 y media cero. Bajo esta suposición, la probabilidad de $\mathbf{B}$ y $\mathbf{G}$ se puede estimar como:

$$p(\mathbf{X}|\mathbf{B},\mathbf{G}) = \prod_{f,t} \frac{1}{\alpha\sqrt{2\pi}} \exp\left(-\frac{([X]_{f,t} - [BG]_{f,t})^2}{2\sigma^2}\right) \tag{3.7}$$

Suponiendo que $p(\mathbf{B},\mathbf{G}) \propto p(\mathbf{B})$, se puede imponer una función de probabilidad exponencial dispersa en cada trama $[\mathbf{G}]_{n,t}$ [KlapuriBook06].

$$p([\mathbf{G}]_{f,t}) = \prod_{f,t} \frac{1}{Z} \exp(-\lambda([G]_{n,t})) \tag{3.8}$$

donde se supone que todos los valores de $\mathbf{G}$ son independientes entre si. Z es un factor de normalización y λ se usa para controlar la envolvente de la distribución de manera que penalice los valores no nulos (activos) de $\mathbf{G}$. Valores típicos de λ vienen dados por las funciones $f(x) = \log(1+x^2)$, $f(x) = |x|$ and $f(x) = x^2$ [Olshausen97, KlapuriBook06, Virtanen03].

Con todo ello, la ec.(3.6) queda como sigue,

$$p(\mathbf{X}|\mathbf{B},\mathbf{G}) = \prod_{f,t} \frac{1}{\alpha\sqrt{2\pi}} \exp\left(-\frac{([X]_{f,t} - [BG]_{f,t})^2}{2\sigma^2}\right) \times \prod_{f,t} \frac{1}{Z} \exp(-\lambda([G]_{n,t})) \tag{3.9}$$

Si se aplican logaritmos, los productos se convierten en sumas y los operadores exponenciales y escalares se pueden descartar para la minimización, por tanto,

$$\min_{\mathbf{B},\mathbf{G}} \frac{1}{2\sigma^2} \|\mathbf{X} - \mathbf{BG}\|_F^2 + \sum_{n,t} \lambda([G]_{n,t}) \tag{3.10}$$

donde F representa la norma de Frobenius.

Con esta formulación, la representación dispersa de las ganancias se puede abordar como un problema de minimización, donde la función de coste es el sumatorio ponderado del error de reconstrucción $||\mathbf{X} - \mathbf{BG}||_F^2$

y el factor de penalización aplicado sobre G. La varianza σ^2 se usa para balancear ambos términos.

La condición de no negatividad es una restricción apropiada cuando se trabaja con funciones base en el dominio de la frecuencia, dado que los espectrogramas de potencia y magnitud sólo contienen valores no negativos por definición. Con esta limitación es posible usar el algoritmo del gradiente o incluso combinar NMF y *sparse coding* [Hoyer04, Abdallah04].

Los algoritmos de *sparse coding* se han usado para diferentes aplicaciones sobre señales musicales. Por ejemplo, Abdallah y Plumbley [Abdallah04] aplican esta solución a transcripción automática de señales sintéticas de piano. Virtanen [Virtanen03], sin embargo, la aplica para transcripción de instrumentos percusivos desde señales MIDI sintetizadas.

3.4. *Non-negative Matrix Factorization (NMF)*

3.4.1. Introducción

Non-negative Matrix Factorization (NMF) es un conocido método de descomposición no supervisada que permite representar datos en dos dimensiones como una combinación lineal de un conjunto de elementos básicos representativos. Dada la matriz de datos $\mathbf{X}$ con dimensión $F \times T$ y valores no negativos, NMF es el problema de encontrar una factorización de la forma

$$\mathbf{X} \approx \mathbf{BG} = \hat{\mathbf{X}} \tag{3.11}$$

donde $\mathbf{B}$ y $\mathbf{G}$ son matrices de valores no negativos de dimensiones $F \times N$ and $N \times T$, respectivamente. Habitualmente, N toma valores tal que $FN + NT << FT$, reduciendo, por tanto, la dimensión de las matrices a tratar.

Esta herramienta se ha aplicado en multitud de aplicaciones y campos, tales como el aprendizaje de patrones de rostros, patrones semánticos de texto [Lee99], transcripción automática de música polifónica [Smaragdis03], separación de fuentes de sonido [Virtanen07b] o clasificación de whiskies Escoceses [Young06].

En las aplicaciones sobre señales de audio, la matriz $\mathbf{X}$ se emplea como una representación tiempo-frecuencia (por ejemplo, el espectro de magnitud

o potencia), en la que se encuadran los *bins* frecuenciales f de cada trama temporal t.

El objetivo del algoritmo NMF es encontrar las matrices no negativas $\mathbf{B}$ and $\mathbf{G}$ que cumplan

$$(\mathbf{B}, \mathbf{G}) = \arg \min_{B,G>0} D(\mathbf{X}|\mathbf{BG}) \tag{3.12}$$

donde $D(\mathbf{X}|\mathbf{BG})$ es una función de coste definida como

$$D(\mathbf{X}|\mathbf{BG}) = D\left(\mathbf{X}|\hat{\mathbf{X}}\right) = \sum_{f=1}^{F}\sum_{t=1}^{T} d\left(X(f,t)|\hat{X}(f,t)\right) \tag{3.13}$$

y donde $d(x|\hat{x})$ es una función de coste escalar. Algunas de las funciones de coste más conocidas y usadas son: la distancia Euclidea, que se define como

$$d_{EUC}(x|y) = \frac{1}{2}(x-y)^2 \tag{3.14}$$

y la divergencia (generalizada) Kullback-Leibler (KL), que se define como

$$d_{KL}(x|y) = x\log\frac{x}{y} - x + y \tag{3.15}$$

Ambas funciones de coste son positivas, toman valor cero si y sólo si $x = y$. Lee y Seung [Lee01] proponen un algoritmo de gradiente descendente para resolver el problema de minimización (ec.(3.13)) usando las funciones de coste descritas. Si en el algoritmo del gradiente se usa un paso adecuado, las ecuaciones de actualización resultantes son multiplicativas, por lo que la función de coste no crecerá. Indudablemente, la simplicidad de las reglas de actualización han hecho que NMF sea una técnica muy popular, y la mayoria de las aplicaciones, antes mencionadas, usen el agoritmo de Lee y Seung para minimizar la distancia Euclidea o la divergencia de KL.

Otra función de coste ampliamente conocida es la de Itakura-Saito (IS), cuyas características son apropiadas para el espectro de la señal de voz. La divergencia IS se define como

$$d_{IS}(x|y) = \frac{x}{y} - \log\frac{x}{y} - 1 \tag{3.16}$$

Esta función de coste fue definida por Itakura y Saito [Itakura68] a partir de la estimación de máxima probabilidad *(Maximum Likelihood, ML)* del espectro de voz con modelado autoregresivo. Se presentó como "una medida de parecido entre dos espectros" y se convirtió en una función de coste muy usada entre los investigadores en el campo de la voz durante los años 70. Una de las características más interesantes de la divergencia IS es la invarianza de escala, es decir, a las componentes de baja energía en $\mathbf{X}$ se les asigna la misma relevancia relativa que a las de alta energía. Esta peculiaridad es de gran importancia cuando se tratan espectros de señales de audio, en los que los coeficientes de $\mathbf{X}$ se engloban en un gran rango dinámico de valores.

En resumen, la elección de la función de coste para el algoritmo NMF debe tener en cuenta el tipo de datos que se van a analizar. En la bibliografía, hay muchos trabajos cuya finalidad es mejorar los resultados de un algoritmo con determinada función de coste, sin embargo, sólo unos pocos tratan de seleccionar la mejor función de coste para el tipo de datos que se están analizando.

3.4.2. Modelos NMF

En el modelo NMF básico, la única restricción es la no negatividad de los elementos de todas las matrices intervinientes. El resto de propiedades de la descomposición son efectos adicionales e intrínsecos al algoritmo. Por tanto, el hecho de que el algoritmo sea capaz de obtener información sobre la señal de entrada, desarrolle una separación u obtenga datos interpretables y con significado, se puede considerar como una "buena noticia". Aún así, no es una herramienta mágica, la mayoría de los autores contrarrestan esta falta de control en la descomposición con ciertas restricciones que dan lugar a variantes del NMF básico. Tales restricciones, como por ejemplo, la dispersión, localización espacial o continuidad temporal, mejoran dichos efectos adicionales en los que los autores se apoyan para lograr datos que se adecúen a su propósito. Algunas de esas restricciones se describen en este capítulo. Una variante típica consiste en agregar un término de penalización,

también llamado de regularización, a la función de coste y minimizarla, ver en [Virtanen07b, Li01, Hoyer04].

Esta variante del término de regularización tienes ciertos inconvenientes [Bertin10]. En primer lugar, se debe establecer un criterio que cuantifique la propiedad que se desea limitar o penalizar. En segundo lugar, no se asegura la convergencia para el esquema evolutivo del modelo. Además, el término de penalización se debe seleccionar empíricamente. Por estas razones, algunos autores han adoptado la idea de la restricción de no negatividad a otros esquemas. Dentro de algunos de los esquemas NMF es interesante contar con una ordenación de las bases. Este orden puede ser útil para aplicarse en un sentido práctico. Si se aplica una restricción de armonicidad a cada base, se puede relacionar cada una de ellas con un *pitch* determinado, dotando de sentido tanto a la base como a sus ganancias correspondientes.

Existen varios esquemas de desarrollo del método de descomposición NMF. En cada esquema se pueden implementar algunas de las restricciones mencionadas anteriormente. Dependiendo del esquema que se use para desarrollar el método NMF, se pueden clasificar en dos grupos: modelos deterministas o estocásticos.

Modelos deterministas

A. Modelo armónico básico *(Basic Harmonic Constrained Model, BHC)*

   Cuando se trabaja con instrumentos tonales, las notas musicales (obviando los tiempos transitorios) son cuasi-periodicas, y su espectro se compone de picos espectrales regularmente espaciados. Por tanto, se puede pensar que los elementos en las bases $b_{n,j}(f)$ deberían tener una forma armónica. Esta característica armónica, propia de los espectros de notas musicales, se les puede imponer a las bases de manera que:

$$b_{n,j}(f) = \sum_{m=1}^{M} c_{m,n,j} G\left(f - m f_0(n)\right) \tag{3.17}$$

   donde $m = 1, ..., M$ representa el número de armónico, $c_{m,n,j}$ representa la amplitud para m-ésimo parcial del *pitch* n y el instrumento j, $f_0(n)$ es la frecuencia fundamental para el *pitch* n, $G(f)$ es el espectro

de amplitud de la ventana de análisis, y el espectro de un componente armónico en la frecuencia $mf_0(n)$ se aproxima ubicando dicha trasnsformada en es la frecuencia armónica correspondiente $G(f - mf_0(n))$.

Esta condición de armonicidad mejora el modelado, puesto que se puede relacionar, de antemano, una función base a un *pitch* mediante la frecuencia fundamental del *pitch* $f_0(n)$. Por el contrario, si se usa el modelo básico de NMF que describe la ec.(3.1), la frecuencia fundamental sólo se puede estimar una vez aprendidos los parámetros del modelo. En [Vincent10], se demuestra que el uso de la restricción de armonicidad mejora la fidelidad de un sistema de transcripción.

Para el modelo NMF-BHC de la ec.(3.17), el espectro de magnitud de la señal $X(f,t)$ en la trama t se estima como

$$\hat{X}(f,t) = \sum_{n,j} g_{n,j}(t) \sum_{m} c_{m,n,j} G\left(f - mf_0(n)\right) \tag{3.18}$$

donde $n = 1, ..., N$, siendo N el número de *pitches* y $j = 1, ..., J$, siendo J el número total de instrumentos. Los parámetros libres del modelo BHC, que serán estimados en la ejecución NMF, son las ganancias $g_n(t)$ y las amplitudes $c_{m,n,j}$.

B. Modelo filtro-fuente con excitación armónica plana *(Source-filter Model with Harmonic-Comb Excitation, HCE)*

El principal problema del modelo NMF básico (ec.(3.1)) y del modelo NMF-BHC (ec.(3.18)) es que requieren una función base para representar cada *pitch* de cada instrumento. Esto conlleva que se tengan que ajustar un gran número de parámetros para cada *pitch*, que además no están relacionados entre sí, por tanto la estimación o adaptación del modelo resulta poco aproximada. Para reducir la complejidad del aprendizaje, en cuanto al número de parámetros, Virtanten y Klapuri [Virtanen06] propusieron modelar cada base como el producto del espectro de una excitación $e_n(f)$ y un filtro $h_j(f)$. A cada función base resultante se le asigna un par de indices, de excitación n y filtro j, de manera que:

$$b_{n,j}(f) = h_j(f)e_n(f), \quad n = 1, ..., N, j = 1, ..., J, \tag{3.19}$$

donde N es el número total de excitaciones y J es el número total de filtros. Normalmente a cada instrumento se le asigna un único filtro que se corresponde con la estructura resonante del cuerpo del instrumento físico.

Este modelo de filtro-fuente con una excitación por *pitch* tiene origen en el procesado de voz y la síntesis de sonido. En el caso del procesado de voz, la excitación modela el sonido producido por las cuerdas vocales, mientras que el filtro modela el efecto de resonancia del tracto vocal [RabinerBook78]. Para el caso de la síntesis de sonidos, el filtro colorea una señal con gran contenido espectral para obtener el sonido deseado.

A pesar de sus ventajas, este tipo de modelos tienen un problema asociado: las excitaciones dependen del picth, es decir, hay una excitación para cada pitch. Por tanto, el problema del número de parámetros que deben ser ajustados no está resuelto. Diversos estudios, como por ejemplo [Badeau09, Heittola09, Klapuri10b], consideran la excitación como un conjunto de elementos unitarios, ubicados en las posiciones armónicas de la frecuencia fundamental. Esta propuesta determina un modelo de excitación de *peine armónico*, que consiste en la suma de componentes armónicas tal que:

$$e_n(f) = \sum_{m=1}^{M} G(f - mf_0(n)) \tag{3.20}$$

donde $m = 1, ..., M$ es el índice de los armónicos, $f_0(n)$ es la frecuencia fundamental de la excitación n, $G(f)$ es una réplica espectral de la ventana de análisis, y el espectro de una componente armónica en la frecuencia $mf_0(n)$ se aproxima con $G(f - mf_0(n))$.

Cuando se usa un filtro fuente con excitación de peine armónico (HCE) para componer las funciones base, la representación tiempo-frecuencia de la señal se puede describir como:

$$\hat{X}(f,t) = \sum_{n,j} g_{n,j}(t) h_j(f) \sum_{m=1}^{M} G(f - m f_0(n)) \tag{3.21}$$

donde $n = 1, ..., N$, siendo N el número de *pitches* y $j = 1, ..., J$, siendo J el número de instrumentos. En la figura 3.1 se muestran los dos parámetros del modelo HCE para un clarinete: el filtro y los valores unitarios del vector de excitación que se trasladan a las posiciones armónicas correspondientes para cada *pitch* y muestrean al filtro.

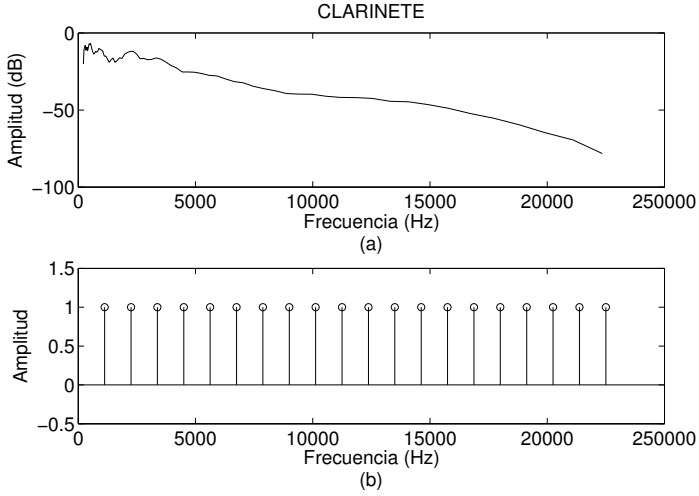


Figura 3.1: *Parámetros de un modelo Harmonic Comb Excitation para un clarinete (a) Filtro estimado. (b) Excitación plana (20 parciales) situados en las posiciones armónicas para el* pitch *con* $f_0 = 1100\,Hz$.

Los parámetros libres del modelo, que deben estimarse, son las ganancias temporales $g_{n,j}(t)$ y el filtro $h_j(f)$, los cuales pueden ajustarse con un esquema NMF con ecuaciones de actualización multiplicativas*(Multiplicative Update Rules, MU)*, que está descrito en el apartado 3.4.3.

B.1 Limitaciones del modelo HCE

El modelo HCE es capaz de representar el espectro de notas con una

distribución suave de energía entre los distintos picos espectrales. En la figura 3.2 se puede ver un ejemplo en el caso de la flauta.

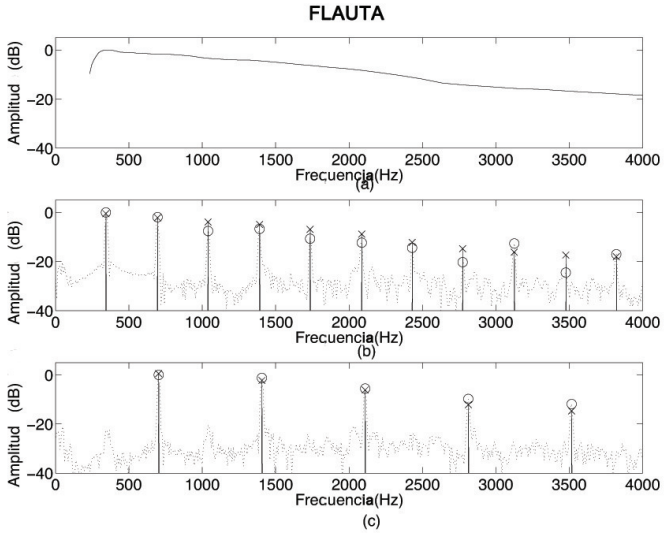


Figura 3.2: *(a) Filtro estimado para una flauta, usando el modelo Harmonic Comb Excitation. (b) Comparación entre el espectro original (círculos) y la estimación de espectro (cruces) para una nota con frecuencia fundamental* $349,2$ *Hz. (c) Idem para una nota con frecuencia fundamental* $698,5$ *Hz (una octava superior)*

Sin embargo, no todos los instrumentos generan una distribución espectral de energía suave. Por ejemplo, en el caso del clarinete, debido a ciertas características físicas, las notas de baja frecuencia presentan una ausencia casi completa de los segundos y cuartos parciales [FletcherBook98]. En el espectro de la nota de frecuencia más baja que puede producir un clarinete, la amplitud del los picos espectrales correspondientes a los parciales impares cae aproximadamente -3 dB/octava, mientras que en el caso de los parciales pares la caída alcanza los -6 dB/octava. El rango de frecuencias fundamentales de un clarinete es desde 145 Hz hasta 1500 Hz, que en escala MIDI se corresponde con un rango desde la nota 50 hasta la 90.

En la figura 3.3 se muestra un ejemplo de una estimación poco fiel, realizada con un modelo HCE sobre una nota de clarinete. En el apartado

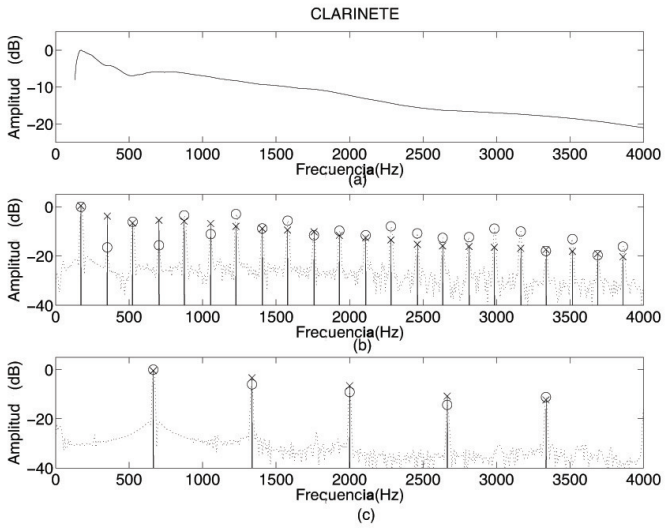


Figura 3.3: *(a) Filtro estimado para un clarinete, usando el modelo Harmonic Comb Excitation. (b) Comparación entre el espectro original (círculos) y la estimación de espectro (cruces) para una nota con frecuencia fundamental* $174,61\,Hz$. *(c) Idem para una nota con frecuencia fundamental* $659,26\,Hz$ *(una octava superior)*

(a) se muestra el filtro de instrumento estimado y en los apartados (b) y (c) se muestran los espectros de funciones base para las notas con frecuencias fundamentales $174,61\,Hz$ (nota MIDI 53) and $659,26\,Hz$ (nota MIDI 76), respectivamente. Se puede apreciar, que no todos los espectros de nota se pueden representar fielmente con el modelo HCE, debido a las diferencias de comportamiento entre las notas de baja y alta frecuencia, dentro de su rango dinámico. El modelo HCE trata de compensar el defecto del modelado de las excitaciones con un filtro que se adapte a todas las notas, sin embargo no es capaz de obtener un filtro que se adecue bien a los espectros tan distintos que se le presentan. Por consiguiente, cuando se trate con instrumentos que tengan un espectro irregular (no suave), como es el caso del clarinete, será muy complicado obtener una representación buena del espectro para los distintos *pitches* si se modela únicamente el filtro de instrumento.

C. Modelo armónico con excitación múltiple *(Multi-Excitation Model (MEI)*

Para superar las anteriores limitaciones, en [Carabias11] se describe un modelo con múltiples excitaciones, las cuales se combinan para adaptarse a las características del instrumento. Este modelo es una extensión del modelo de filtro-fuente y consigue generar una excitación compuesta por la combinación lineal de un conjunto de excitaciones base propias de cada instrumento. Con este modelo, el espectro de una nota es generado por la excitación armónica correspondiente a esa nota, la cual multiplica al filtro de instrumento. Por consiguiente, la excitación $e_{n,j}(f)$, además de tener naturaleza armónica, es diferente para cada *pitch* del instrumento. Las excitaciones de *pitch*, se construyen como una combinación lineal de unas excitaciones base, que son propias de cada instrumento, y la ponderación de ellas se realiza con unos factores, o pesos, que varían en función del *pitch*.

Formulando este modelo, la excitación por *pitch* quedaría de la siguiente manera,

$$e_{n,j}(f) = \sum_{i=1}^{I} w_{i,n,j} \sum_{m=1}^{M} v_{i,m,j} G\left(f - m f_0(n)\right) \tag{3.22}$$

donde $w_{i,n,j}$ es el peso del vector de excitación i-ésimo para el *pitch* n y el instrumento j, $v_{i,m,j}$ es el parcial m-ésimo del vector de excitación i-ésimo para el instrumento j. Finalmente, el espectro estimado de la señal completa en una trama se define de la siguiente manera,

$$\hat{X}(f,t) = \sum_{n,j} g_{n,j}(t) h_j(f) \sum_{i=1}^{I} w_{i,n,j} \sum_{m=1}^{M} v_{i,m,j} G\left(f - m f_0(n)\right) \tag{3.23}$$

donde $n = 1, ..., N$, siendo N el número de *pitches* y $j = 1, ..., J$, siendo J el número de instrumentos. M indica el número total de armónicos considerados y I indica el número de excitaciones consideradas, siendo $I << N$. El hecho de usar un número bajo de excitaciones base I, reduce significativamente el número de parámetros del modelo, lo que

ayuda a realizar un proceso de aprendizaje más efectivo. En la figura 3.4 se muestran los parámetros estimados para un clarinete con el modelo MEI.

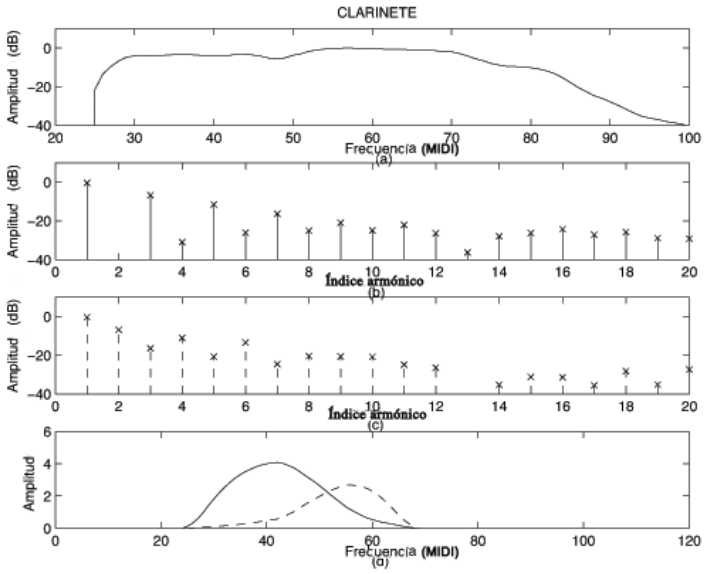


Figura 3.4: *Parámetros estimados para el modelo multiexcitación propuesto para un fichero de clarinete (fichero RWC 311CLNOM). (a) Filtro estimado. (b) Primera excitación base (v1;m;j). (c) Segunda excitación base (v2;m;j). (d) Pesos de ponderación de las excitaciones base en función del* pitch *en escala MIDI (w1;p;j el línea continua y w2;p;j en línea discontinua).*

En la figura 3.5 se muestra la predicción de espectro para el clarinete usando el modelo MEI. Se puede ver, que este modelo es capaz de generar funciones base irregulares (no suaves) que encajen mejor con el espectro original del instrumento que en el caso del modelo HCE, mostrado en la figura 3.3, especialmente para los primeros parciales de las notas de baja frecuencia.

Modelos estadísticos

En los modelos estadísticos, la elección previa y adecuada de las funciones de distribución para $p(\mathbf{B})$ y $p(\mathbf{G})$, es la manera de inducir las pro-

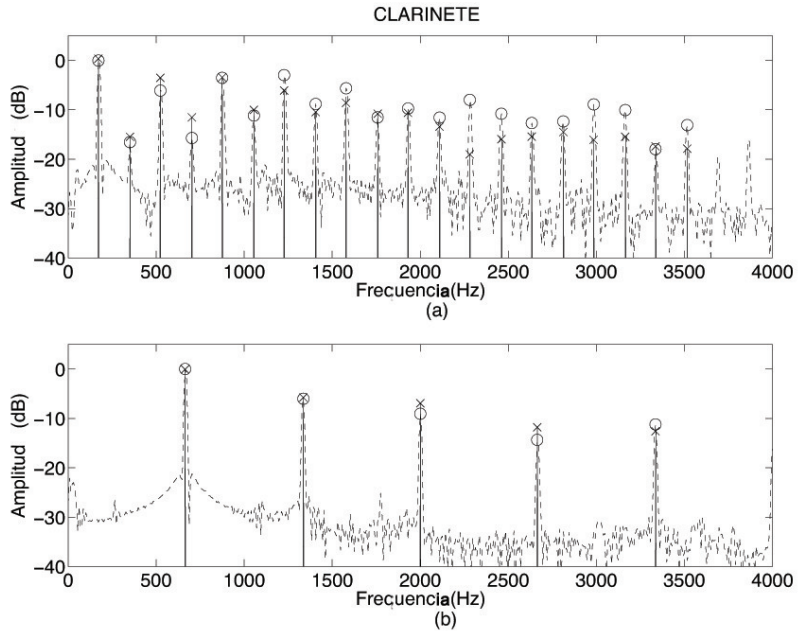


Figura 3.5: *Comparación entre un espectro de nota de clarinete (linea discontinua) y el espectro modelado con el modelo de excitación múltiple MEI (línea sólida) para el* pitch 174,61 Hz *(a) y el pitch* 659,26 Hz *(b).*

piedades deseables en la descomposición. Además, el uso de un esquema estadístico aporta una gran base teórica y un gran conjunto de algoritmos eficientes con convergencia demostrada, como el caso del algoritmo de *Expectation-Maximization* (EM) y sus variantes, para estimar los parámetros de NMF. De hecho, una gran ventaja del punto de vista estadístico es la posibilidad de elegir desde la estimación ML hasta la estimación del máximo a posteriori (MAP), gracias a la regla de Bayes:

$$p(\mathbf{B},\mathbf{G}|\mathbf{X}) = \frac{p(\mathbf{X},\mathbf{B}|\mathbf{G})p(\mathbf{B})p(\mathbf{G})}{p(\mathbf{X})} \tag{3.24}$$

La elección de cierta función de coste para medir el parecido entre $X(f,t)$ y $\hat{X}(f,t)$, implica ciertas suposiciones estadísticas sobre cómo se genera $X(f,t)$ a partir de $\hat{X}(f,t)$. Ya se ha señalado en algunas publicaciones, como por ejemplo [Virtanen08, Cemgil08, Fevotte09a, Fevotte09b], que la

distancia Euclidea, y las divergencias KL e IS se basan en los siguientes modelos generativos:

$$X(f,t) \sim \mathcal{N}(X(f,t); \hat{X}(f,t), \sigma^2) \quad EUC - NMF \tag{3.25}$$
$$X(f,t) \sim \mathcal{P}(X(f,t); \hat{X}(f,t)) \quad KL - NMF \tag{3.26}$$
$$X(f,t) \sim \mathcal{G}(X(f,t); \hat{X}(f,t)) \quad IS - NMF \tag{3.27}$$
$$\tag{3.28}$$

donde $\mathcal{N}$, $\mathcal{P}$ y $\mathcal{G}$ indican las distribuciones Gaussiana, de Poisson y Gamma, respectivamente, definidas en la tabla 3.1, y donde $X(f,t)$ obedece a la parametrización $\hat{X}(f,t) = \sum_n b_n(f) g_t(t)$. La probabilidad de los parámetros **B** y **G** en los modelos anteriores puede corresponderse con la función de coste (3.13), de manera que la versión determinista de NMF es equivalente a la estimación de máxima probabilidad (ML) [Fevotte09b].

Gaussiana multivariable	c=1/2 (caso real) or 1 (caso complejo) $\mathcal{N}_c(x\|\mu,\Sigma) = \|\pi\Sigma\|^{-1} \exp -(x-\mu)^G \Sigma^{-1}(x-\mu)$
Poisson	$\mathcal{P}(x\|\lambda) = \exp(-\lambda)\frac{\lambda^x}{x!}$
Binomial	$\mathcal{B}(x\|n,p) = \begin{pmatrix} n \\ x \end{pmatrix} p^x (1-p)^{n-x}$
Gamma	$\mathcal{G}(u\|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} u^{-1} \exp(-\beta u), u \geq 0$

Tabla 3.1: *Distribuciones de probabilidad comunes*

Modelos combinados

Una característica importante de las distribuciones Gaussiana y de Poisson es que se pueden sumar sin perder sus propiedades intrínsecas. Gracias a ello, cuando $v = \sum_n c_n$, y c_n son variables Gaussianas (o de Poisson), v también es Gaussiana (o de Poisson). Inversamente, cualquier variable v puede descomponerse en $\sum_n c_n$ sin ninguna modificación del modelo asociado. A continuación, se presentan los modelos con mayor profundidad en cuanto a su estructura de combinación. Consideremos el modelo generativo:

$$v_{ft} = \sum_n c_{n,ft} \tag{3.29}$$

$$c_{n,ft} \sim p(c_{n,ft}|\theta_n) \tag{3.30}$$

donde $\theta_n = \{b_{:,n}, g_{n,:}\}$. Durante los siguientes párrafos se describe cómo los modelos Euclidean-NMF, KL-NMF e IS-NMF son equivalentes al modelo de estimación ML $\theta = \{\theta_1, ..., \theta_N\}$ en determinados casos particulares. $\mathbf{C_n}$ y $\mathbf{V}$ denotan las matrices, con tamaño $F \times T$ que contienen elementos $\{c_{n,ft}\}_{ft}$ y $\{v_{ft}\}_{ft}$, respectivamente.

- NMF con distancia Euclidea (EUC-NMF)

  Se considera el siguiente modelo generativo:

$$c_{n,ft} \sim \mathcal{N}\left(c_{n,ft}; b_{fn}g_{nt}, \frac{\sigma^2}{N}\right) \tag{3.31}$$

  De forma sencilla se puede ver que [Fevotte09a]

$$-\log p(\mathbf{X}|\mathbf{B}, \mathbf{G}, \sigma^2) = \frac{1}{\sigma^2} D_{EUC}(\mathbf{X}|\mathbf{BG}) + \frac{NF}{2}\log(2\pi\sigma^2) \tag{3.32}$$

  Por tanto, la estimación ML de $\mathbf{B}$ y $\mathbf{G}$ es equivalente a la descomposición NMF de $\mathbf{X} = \mathbf{V}$ en $\mathbf{BG}$ usando la distancia Euclidea.

  Aún así, existe una interpretación ambigua con el modelo generativo definido en las ecuaciones (3.29), (3.30), (3.31), puesto que pueden producir datos negativos, mientras que NMF no los genera. Aunque el problema final de optimización resultante es el mismo cuando los valores de $\mathbf{X}$ son no negativos, hay una diferencia semántica entre los dos puntos de vista, de EUC-NMF y la estimación ML con el modelo generativo Gaussiano. Una propuesta mas razonable podría ser adoptar componentes generados por una función normal truncada [Fevotte09b], sin embargo, esto rompe la correspondencia formal entre las dos propuestas, por el hecho de necesitar una normalización de las distribuciones.

- NMF con la divergencia generalizada KL (KL-NMF)

  En este caso comenzamos suponiendo el siguiente modelo generativo:

$$c_{n,ft} \sim \mathcal{P}(c_{n,ft}; b_{fn}g_{nt}) \tag{3.33}$$

  De forma sencilla se puede ver que [Fevotte09a]

$$-\log p(\mathbf{X}|\mathbf{B},\mathbf{G}) \stackrel{c}{=} D_{KL}(\mathbf{X}|\mathbf{BG}) \tag{3.34}$$

  donde $\stackrel{c}{=}$ indica igualdad en condiciones constantes. Por tanto, la estimación ML de $\mathbf{B}$ y $\mathbf{G}$ es equivalente a la descomposición NMF de $\mathbf{X} = \mathbf{V}$ en $\mathbf{BG}$ usando la divergencia KL. Los valores de $\mathbf{V}$ generados por el modelo generativo de las ecuaciones (3.29), (3.30), (3.33) son no negativos, pero aún persiste una interpretación ambigua al tratar datos reales, puesto que el proceso de Poisson produce números enteros [Fevotte09b].

- NMF con la divergencia IS (IS-NMF)

  Suponemos el siguiente modelo generativo:

$$c_{n,ft} \sim \mathcal{N}_c(c_{n,ft}; 0, b_{fn}g_{nt}) \tag{3.35}$$

  donde $\mathcal{N}_c$ representa la distribución Gaussiana compleja, que se encuentra definida en el Apéndice. Los valores de $\mathbf{V}$ obtenidos por el modelo son complejos (aunque también se podrían usar distribuciones Gaussianas reales). Por tanto, se puede demostrar que [Fevotte09a]:

$$-\log p(\mathbf{V}|\mathbf{B},\mathbf{G}) \stackrel{c}{=} D_{IS}(|\mathbf{V}|^{,2}|\mathbf{BG}) \tag{3.36}$$

  donde $|\mathbf{V}|^{,2}$ es la matriz con valores $|v_{ft}|^2$. Por consiguiente, la estimación ML de $\mathbf{B}$ and $\mathbf{G}$ es equivalente a la descomposición NMF de $\mathbf{X} = |\mathbf{V}|^{,2}$ en $\mathbf{BG}$, usando la divergencia IS.

3.4.3. Algoritmos

Multiplicative update rules (MU)

[Lee01] propone un algoritmo iterativo, basado en ecuaciones, o reglas, de actualización multiplicativas (MU), para el cálculo de los valores de los parámetros libres del modelo en cada iteración de la descomposición. Siguiendo estas reglas, se ha demostrado que la distorsión $D(X(f,t)|\hat{X}(f,t))$ decrece en cada iteración y además asegura la no negatividad de los valores de las bases y ganancias. Dichas reglas multiplicativas se obtienen aplicando un escalado diagonal al incremento del algoritmo del gradiente (ver [Lee01] para más detalles).

En resumen, la ecuación de actualización para cada parámetro θ_l viene dada por la derivada parcial de la función de coste $\nabla_{\theta_l} D$, como el cociente de dos términos positivos $\nabla^+_{\theta_l} D$ y $\nabla^-_{\theta_l} D$, de manera que:

$$\theta_l \leftarrow \theta_l \frac{\nabla^-_{\theta_l} D(x|\hat{x})}{\nabla^+_{\theta_l} D(x|\hat{x})} \tag{3.37}$$

suponiendo que $\nabla_{\theta_l} D = \nabla^+_{\theta_l} D - \nabla^-_{\theta_l} D$. En conclusión, este método es equivalente a actualizar cada parámetro, multiplicándolo su valor en la iteración anterior por el ratio de las partes negativa y positiva de la derivada de la distorsión respecto a ese parámetro.

Tal y como se dice en [Lee01] la principal ventaja del uso de las reglas de actualización multiplicativas en la ec.(3.37) es que aseguran la no negatividad de las bases y las ganancias. En [Lee01] se demuestra que, para la distancia Euclídea y la divergencia de *Kullback-Leibler*, no se puede asegurar la obtención de un mínimo global por no ser un problema convexo, pero si que existen varias técnicas que aseguran la evolución decreciente de la distorsión y por tanto permiten encontrar mínimos locales de distorsión para las matrices de bases y ganancias.

El algoritmo del gradiente es la técnica más simple para poder encontrar estos mínimos locales, pero su convergencia puede ser excesivamente lenta. Otros métodos, como el del gradiente conjugado, convergen más rápidamente pero su implementación se complica en gran medida. En [Lee01] se proponen las ecuaciones de actualización multiplicativas como una técnica

con un buen compromiso entre la velocidad de convergencia y complejidad de implementación.

Por tanto, para asegurar que la distancia Euclidea sea decreciente $\|X - BG\|$, se usan las siguientes ecuaciones de actualización [Lee01]:

$$\begin{aligned} G_{nt} &\leftarrow G_{nt} \frac{(B^T X)_{nt}}{(B^T BG)_{nt}} \\ B_{fn} &\leftarrow B_{fn} \frac{(XG^T)_{fn}}{(BGG^T)_{fn}} \end{aligned} \tag{3.38}$$

En el caso de la divergencia KL $D_{KL}(X|BG)$, las ecuaciones de actualización, para que la divergencia sea decreciente, quedan de la siguiente manera[Lee01]:

$$\begin{aligned} G_{nt} &\leftarrow G_{nt} \frac{\sum_f B_{fn} X_{ft}/(BG)_{ft}}{\sum_f B_{nt}} \\ B_{fn} &\leftarrow B_{fn} \frac{\sum_n G_{nt} X_{ft}/(BG)_{ft}}{\sum_n G_{nt}} \end{aligned} \tag{3.39}$$

Para la divergencia IS $D_{IS}(X|BG)$, las ecuaciones son las siguientes [Fevotte09a]:

$$\begin{aligned} G_{nt} &\leftarrow G_{nt} \frac{\sum_f B_{fn} X_{ft}/(BG)_{ft}^2}{\sum_f B_{ft}/(BG)_{ft}} \\ B_{fn} &\leftarrow B_{fn} \frac{\sum_n G_{nt} X_{ft}/(BG)_{ft}^2}{\sum_n G_{nt}/(BG)_{ft}} \end{aligned} \tag{3.40}$$

Sin embargo, en la práctica se puede observar que la convergencia del algoritmo con a divergencia IS es, aún, un problema abierto.

Expectation Maximization (EM)

En la sección 3.4.2 se describe cómo los modelos EUC-NMF, KL-NMF e IS-NMF subyacen bajo los modelos de combinación estadísticos. Cada componente actúa como una variable latente que puede usarse como un

dato completo dentro del algoritmo EM. En esta configuración, la siguiente función se debe minimizar iterativamente:

$$Q(\theta|\theta') \stackrel{def}{=} -\int_C \log p(C|\theta)p(C|V,\theta')dC$$

donde $\theta = \{\mathbf{B}, \mathbf{G}\}$ y $\mathbf{C}$ es el tensor con porciones C_n y elementos $c_{n,ft}$. La convergencia de este algoritmo a un punto estacionario está garantizada. Usando independencia condicional,

$$p(C|\theta) = \prod_n p(C_n|\theta_n)$$

la función EM se puede expresar como:

$$Q(\theta|\theta') = \sum_n Q_n(\theta_n|\theta')$$

$$Q_n(\theta_n|\theta') \stackrel{def}{=} -\int_{C_n} \log p(C_n|\theta_n)p(C_n|V,\theta')dC_n \tag{3.41}$$

Si se supone independencia y distribución uniforme (i.i.d.) [Fevotte09a, Fevotte09b], la función se puede simplificar a

$$Q_n(\theta_n|\theta') = -\sum_{ft}\int_{c_{n,ft}} \log p(c_{n,ft}|\theta_n)p(c_{n,ft}|v_{ft},\theta')dC_{n,ft} \tag{3.42}$$

A continuación, se detalla el algoritmo EM para los casos de distancia Euclidea, divergencia KL y divergencia IS.

- EUC-NMF

$$\begin{aligned} -\log p(c_{n,ft}|\theta_n) &\stackrel{c}{=} \frac{1}{2\sigma^2}(c_{n,ft} - b_{fn}g_{nt})^2 \\ p(c_{n,ft}|v_{ft},\theta) &= \mathcal{N}(c_{n,ft}|\mu^{post}_{n,ft}, \lambda^{post}_{n,ft}) \end{aligned}$$

con

$$\mu^{post}_{n,ft} = b_{fn}g_{nt} + \frac{1}{N}(v_{ft} - \hat{v}_{ft}) \quad , \quad \lambda^{post}_{n,ft} = \frac{N-1}{N^2}\sigma^2 \tag{3.43}$$

donde $\hat{v}_{ft} = \hat{X}(f,t) = \sum_n b_{fn} g_{nt}$. Por tanto, la minimización de la función (3.42) bajo la condición de no negatividad de las componentes lleva a

$$g_{nt} = \left[\frac{\sum_f b_{fn} \left(\frac{1}{N}(v_{ft} - \hat{v}'_{ft}) + b'_{fn} g'_{nt} \right)}{\sum_f b_{fn}^2} \right]_+ \tag{3.44}$$

$$b_{fn} = \left[\frac{\sum_n g_{nt} \left(\frac{1}{N}(v_{ft} - \hat{v}'_{ft}) + b'_{fn} g'_{nt} \right)}{\sum_n g_{nt}^2} \right]_+ \tag{3.45}$$

donde $\lfloor v \rfloor_+ = \max\{v, 0\}$. Estas ecuaciones de actualización difieren de las habituales ecuaciones multiplicativas dadas en la ec.(3.38).

- KL-NMF

$$\begin{aligned} -\log p(c_{n,ft}|\theta_n) &\stackrel{c}{=} -b_{fn} g_{nt} + c_{n,ft} log(b_{fn} g_{nt}) \\ p(c_{n,ft}|v_{ft}, \theta) &= \mathcal{B}(c_{n,ft}|v_{n,ft}, \pi_{n,ft}) \end{aligned}$$

donde $\mathcal{B}$ representa una distribución binomial, definida en el Apéndice, $\pi_{n,ft} = b_{ft} g_{nt} / \hat{v}_{ft}$ y para esta ocasión $\hat{v}_{ft} = \hat{X}(f,t) = \sum_n b_{ft} g_{nt}$. Con todo ello se llega a las siguientes ecuaciones de actualización:

$$g_{nt} = g'_{nt} \frac{\sum_f b'_{fn} \left(\frac{v_{ft}}{\hat{v}'_{ft}} \right)}{\sum_f b_{fn}} \tag{3.46}$$

$$b_{fn} = b'_{fn} \frac{\sum_n g'_{nt} \left(\frac{v_{ft}}{\hat{v}'_{ft}} \right)}{\sum_n g_{nt}} \tag{3.47}$$

las cuales coinciden con el caso de las ecuaciones multiplicativas dadas en la ec.(3.39).

- IS-NMF

$$-\log p(c_{n,ft}|\theta_n) \stackrel{c}{=} \log(b_{fn}g_{nt}) + \frac{|c_{n,ft}|^2}{b_{fn}g_{nt}} \quad (3.48)$$

$$p(c_{n,ft}|v_{ft},\theta) = N(c_{n,ft}|\mu^{post}_{n,ft}\lambda^{post}_{n,ft}) \quad (3.49)$$

con

$$\mu^{post}_{n,ft} = \frac{b_{fn}g_{nt}}{\sum_l b_{fl}g_{\ln}}v_{ft} \quad , \quad \lambda^{post}_{n,ft} = \frac{b_{fn}g_{nt}}{\sum_l b_{fl}g_{\ln}}\sum_{l\neq n} b_{fl}g_{\ln} \quad (3.50)$$

Por tanto,

$$g_{nt} = \frac{1}{F}\sum_f \frac{v'_{n,ft}}{b_{fn}} \quad (3.51)$$

$$b_{fn} = \frac{1}{N}\sum_n \frac{v'_{n,ft}}{g_{nt}} \quad (3.52)$$

con $v'_{n,ft} = \left|mu^{post'}_{n,ft}\right|^2 + \lambda^{post'}_{n,ft}$. En este caso, las ecuaciones también difieren de las que se dieron en la ec.(3.40), y son equivalentes a las del algoritmo SAGE descrito en [Fevotte09a].

Space-Alternating Generalized Expectation-Maximization, (SAGE)

Este algoritmo es una extensión del algoritmo EM para modelos de datos con ciertas estructuras particulares. Se sabe que es capaz de tener una convergencia más rápida que el algoritmo EM, aunque una iteración de SAGE es computacionalmente más costosa que la del EM.

El algoritmo SAGE consiste en elegir, para cada subconjunto de parámetros θ_n un espacio de datos completo y específico. Este espacio de datos para θ_n se puede seleccionar de manera sencilla de manera que sea $C_n \stackrel{def}{=} [c_{n,1},...,c_{n,N}]$ [Fevotte09a]. De todo ello resulta una función del tipo EM para cada subconjunto θ_n como:

$$Q_n^{ML}(\theta_n|\theta') \stackrel{def}{=} -\int_{C_n} \log p(C_n|V,\theta')dC_n \tag{3.53}$$

Una iteración i del algoritmo SAGE supone una fase de cálculo *(E-step)* y otra de minimización *(M-step)* $Q_n^{ML}(\theta_n|\theta')$ for $n = 1, ..., N$. Hay que destacar que en este caso $'$ contiene los valores de los parámetros más actualizados, y no sólo los valores de la iteración $i-1$, como en el caso del algoritmo EM estándar. Esta es la razón fundamental por la que el algoritmo SAGE tiene un coste computacional mayor.

En [Fevotte09a, Bertin10] se pueden encontrar más detalles sobre la estimación de las ecuaciones de actualización de IS-NMF usando el algoritmo SAGE. Las actualizaciones de los parámetros se calculan siguiendo las siguientes reglas:

$$g_{nt}^{(i+1)} = \frac{1}{F}\sum_f \frac{v'_{n,ft}}{b_{ft}^{(i)}} \tag{3.54}$$

$$b_{fn}^{(i+1)} = \frac{1}{F}\sum_n \frac{v'_{n,ft}}{g_{nt}^{(i+1)}} \tag{3.55}$$

Alternative Non Negative Least Squares Algorithm, (ANLS)

Este algoritmo es una propuesta ampliamente utilizada para la solución del problema NMF [Berry07, Kim08b] que busca la solución que se describe en la ecuación (3.56).

$$D\left(X \,\|B_0G_0\right) = \min_{B\geq 0, G\geq 0} D\left(X\,\|BG\right) \tag{3.56}$$

Tal y como se dice en [Lee01], la minimización de la distorsión de las matrices de ganancias y bases ($\mathbf{B}$ y $\mathbf{G}$) es un problema no convexo. Sin embargo la minimización del problema para cada una de ellas por separado sí es un problema convexo. La idea del algoritmo ALS es convertir el problema no convexo en dos problemas convexos fijando una de las matrices y optimizando la otra de manera alternativa, tal y como se muestra en el algoritmo 1.

Algoritmo 1 Descripción del algoritmo ANLS

1 Inicialización de $\mathbf{W}^{(0)} \in \mathbb{R}, \mathbf{H}^{(0)} \in \mathbb{R}, \mathbf{W}(i,j), \mathbf{H}(i,j) \geq 0$
2 **for** $i = 1, 2, \ldots \rightarrow$ convergencia **do**
3 $\quad \mathbf{G}^{(i)} = \arg \min_{\mathbf{G} \geq 0} D\left(\mathbf{X} \| \mathbf{BG}\right)$
4 $\quad \mathbf{B}^{(i)} = \arg \min_{\mathbf{B} \geq 0} D\left(\mathbf{X} \| \mathbf{BG}\right)$
5 **end for**

En [Grippo00] se resuelve la optimización de cada una de las matrices de manera alternativa (líneas 2: y 3: del algoritmo 1) mediante *Non Negative Least Squares (NNLS)* con un método adecuado, como por ejemplo el método de los conjuntos activos *(Active Sets)* descrito en [Lawson95, Bjork96]. La solución del problema NNLS para cada matriz, se debe realizar de manera independiente para cada columna de la matriz $\mathbf{G}$ y cada fila de la matriz $\mathbf{B}$, una implementación eficiente de este tipo se propone en [Alonso14].

En [Kim11] también se propone un algoritmo, llamado *Block Pivoting*, basado en esta idea. Sin embargo, en este caso los autores proponen realizar el proceso de optimización sobre grupos de columnas que tengan los mismos conjuntos de restricciones activas al aplicar el método de los conjuntos activos. Este algoritmo proporciona buenos resultados en tiempo de ejecución y una buena precisión en la estimación de las matrices $\mathbf{G}$ y $\mathbf{B}$.

3.4.4. Restricciones sobre los modelos

Como se ha comentado anteriormente, en la versión estándar de NMF, la única restricción es la no negatividad de los valores de todas matrices del modelo. El resto de las propiedades de la descomposición no son controladas, y todas ellas se producen como "efectos secundarios". A pesar de ser efectos no controlados, la descomposición adquiere ciertas propiedades de la señal original que pueden ser muy útiles para llevar a cabo una separación o la extracción de información relativa a la interpretación. El siguiente paso natural, es tratar de potenciar y controlar estos efectos añadiendo más restricciones, de manera explícita, al problema de factorización.

Restricciones sobre modelos deterministas

La manera de introducir restricciones al modelo es usando un término de penalización. De esta forma, además de minimizar el error de reconstrucción D_r, la función de coste incluye un término D_c que cuantifica la característica que se desea restringir. Por tanto, el problema NMF con restricción se puede formular de la siguiente manera:

$$\min_{\mathbf{B},\mathbf{G}} D_r(\mathbf{X}|\mathbf{BG}) + \lambda D_c(\mathbf{X}|\mathbf{BG}) \tag{3.57}$$

donde λ es un factor que se puede ajustar para regular la influencia de la restricción sobre el procedimiento de minimización NMF.

A continuación se presentan algunas de las restricciones más conocidas:

- Dispersión [Hoyer04, Virtanen07b]: NMF ha demostrado ser una herramienta de análisis útil para tratar diversos tipos de datos. Una de las propiedades más útiles que presenta es que sus descomposiciones son fácilmente interpretables de manera intuitiva por ser dispersas. Sin embargo, en algunas ocasiones, la dispersión de los datos que obtiene NMF, por sí sólo, no es suficiente; en estos casos resulta muy útil poder controlar en grado de dispersión que se produce en el proceso de factorización. Por ejemplo, cuando el espectro de una fuente cubre parcialmente el espectro de otra, la última fuente se puede modelar como la suma del primer sonido más un residuo. El uso de ganancias dispersas puede ayudar a poder identificar la segunda fuente con espectro propio.

- Monofonía [Carabias13]: Cuando se está trabajando con señales monofónicas, o bien instrumentos monofónicos, se puede usar esta propiedad como restricción a la hora de descomponer la señal. Si se conoce que el instrumento sólo es capaz de generar una nota en cada instante, no tendría sentido permitir que más de un *pitch* se activara a la vez. Este es un caso extremo de dispersión, que se puede aplicar a toda la señal, o bien, a cada instrumento si la señal se compone de varios instrumentos monofónicos.

- Localización espacial [Li01]: esta restricción es particularmente interesante para muchas aplicaciones de procesado de imagen. La localización de rasgos es de utilidad en el reconocimiento de objetos, incluyendo estabilidad frente a deformaciones, variaciones de luminosidad y oclusiones parciales. Básicamente, la función de coste se modifica para imponer el área de las características que se buscan y adaptando la representación para que la identificación de elementos sea más sencilla, como en el caso del reconocimiento facial.

- Continuidad temporal [Chen06, Virtanen07b, Bertin10]: en el algoritmo NMF estándar, y en la mayoría de sus variantes, las tramas temporales se consideran independientes entre sí, lo cual no es del todo cierto en las señales de audio por ejemplo. Un rasgo distintivo inherente al sonido de instrumentos musicales es la continuidad temporal. De hecho, como consecuencia de la presencia de una nota, con un *pitch* constante, las ganancias de tramas adyacentes g_{nt} tienden a ser similares unas a otras. En [Virtanen07b] y [Chen06] se introduce un término de penalización en la función de coste NMF para tener en cuenta esta continuidad temporal. En [Virtanen07b], el término esta relacionado directamente con la diferencia $g_{nt} - g_{n(t-1)}$, mientras que en [Chen06] el término depende de una ratio entre la variación de g_n a corto y largo plazo. Ambos términos han demostrado aportar suavidad en las líneas temporales de $\mathbf{G}$.

En la Tabla 3.2 se muestran algunos de los términos de penalización más comunes en la bibliografía:

Estas restricciones se integran en el proceso de factorización mediante los términos penalización como se indica en la ecuación (3.57). Seguidamente las ecuaciones de actualización, incluyendo la restricción correspondiente, se obtienen de la misma manera que cuando no existe término de penalización en el cálculo de la función de coste, de manera que se minimizan ambos términos $D_r(\mathbf{X}|\mathbf{BG})$ y $D_c(\mathbf{X}|\mathbf{BG})$ (ecuación (3.57)). Esta minimización se lleva a cabo con la derivada parcial de la función de coste $\nabla_{\theta_l} D$, quedando la actualización de cada parámetro θ_l según la ecuación (3.37).

Tabla 3.2: *Restricciones comunes en el problema NMF mediante término de penalización* D_c[Bertin10]

Dispersión	$D_c(\mathbf{X}\|\mathbf{BG}) = \sum_{n=1}^{N} \frac{1}{\sqrt{N}-1} \left(\sqrt{N} - \sum_{n=1}^{N} \|g_{nt}\| \Big/ \sqrt{\sum_{n=1}^{N} g_{nt}^2}\right)$	[Hoyer04]
	$D_c(\mathbf{X}\|\mathbf{BG}) = \sum_{n=1}^{N}\sum_{t=1}^{T} f\left(g_{nt}/\sigma_n\right); f(x)\log\left(x^2+1\right)$	[Virtanen07b]
	$D_c(\mathbf{X}\|\mathbf{BG}) = \left\|B^p\right\|, p < 2$	[Raczynski08]
Activación simple	$D_c(\mathbf{X}\|\mathbf{BG}) = \left\|W \odot \left(BB^T\right)\right\|, W_{i,i} = 0$	[Raczynski08]
Localización espacial	$D_c(\mathbf{X}\|\mathbf{BG}) = \lambda_1 \sum_{n=1}^{N}\sum_{n'=1}^{N} \left[B^T B\right]_{nn'} - \lambda_2 \sum_{n=1}^{N} \left[GG^T\right]_{nn}$	[Li01]
Continuidad temporal	$D_c(\mathbf{X}\|\mathbf{BG}) = \sum_{n=1}^{N}\sum_{n=1}^{N} \left\|g_{nt} - g_{n(n-1)}\right\|^2$	[Virtanen07b]
	$D_c(\mathbf{X}\|\mathbf{BG}) = \sum_{n=1}^{N} \log\left(p\left(g_{nt}\right)\right)$	[Bertin10]

Restricciones sobre modelos estadísticos

Otra forma de inducir ciertas propiedades en el proceso de factorización NMF es elegir un esquema estadístico, como se explicó en la sección 3.4.2, y realizar una adecuada selección de los tipos de distribuciones de probabilidad de las variables.

En [Virtanen08], la restricción se implementa mediante el uso de una cadena de Markov con unas variables auxiliares $z_{k,n}$. Mediante una cadena gamma, se puede seleccionar una distribución de probabilidad a priori, que garantice que las ganancias sean estrictamente no negativas, y positivamente correladas (variación lenta en el tiempo). Entonces, bajo ciertas suposiciones, el máximo a posteriori (MAP) nos lleva a una función que incorpora dos términos: la discordancia acústica, medida mediante la divergencia KL más un término que penaliza los cambios grandes en las ganancias de una trama a otra.

$$\sum_{n,j}\sum_{t=1}^{T} d(a, g_{n,t,j} z_{n,t,j} a) + d(a, g_{n,t,j} z_{n,t+1,j} a), \tag{3.58}$$

donde a es un término que relaciona los parámetros de una trama con los de la siguiente y d es la divergencia KL. Las tramas adyacentes se relacionan de manera más fuerte con valores altos de a. $z_{k,n}$ son variables auxiliares que se usan para un cálculo más rápido y sencillo de las ecuaciones de actualización de las ganancias. Estas variabes $z_{k,n}$ dependen de las ganancias, de manera que:

$$z_{n,t,j} = \begin{cases} 1/(g_{n,j}(1) + b) & t = 1 \\ 2/(g_{n,j}(t) + g_{n,j}(t-1)) & 1 < t < T+1 \\ 1/g_{n,j}(t) & t = T+1 \end{cases} \tag{3.59}$$

Por tanto, este esquema permite realizar un cálculo rápido y sencillo de las actualizaciones de las ganancias, con unas ecuaciones que quedan de la siguiente manera:

$$g_{n,j}(t) \leftarrow g_{n,j}(t) \frac{2/g_{n,j}(t) + \nabla^{-}_{g_{n,j}(t)} D(X(f,t)|\hat{X}(f,t))}{a\,(z_{n,t,j} + z_{n,t+1,j}) + \nabla^{+}_{g_{n,j}(t)} D(X(f,t)|\hat{X}(f,t))} \tag{3.60}$$

Algunas propuestas similares, para reforzar la continuidad temporal en las ganancias se desarrollan en [Fevotte09a, Bertin10]. Sin embargo, en dichas propuestas, se plantea una cadena gamma inversa a priori y la actualización de parámetros en un esquema SAGE usando la divergencia IS.

3.5. Información previa sobre las fuentes

Los algoritmos que se han descrito en este capítulo contienen algunas suposiciones generales, que tienen relación con el tipo de señal que van a analizar, como es la independencia de las fuentes o la no negatividad. Estas suposiciones son generales y se imponen desde el núcleo del algoritmo, pudiendo ser aptos para una separación de fuentes a ciegas (BSS). A pesar de esto, tal como se comentó en el Capítulo 1, ese tipo de separación

no es adecuada para señales monoaurales y se necesitan ciertas fuentes de información que guíen al algoritmo para obtener una calidad de separación aceptable.

Siguiendo la estructura de los algoritmos descritos, se puede incluir información temporal y/o espectral que sirva como guía, o punto de partida para ejecutar la separación de las fuentes presentes en la señal de entrada [Fritsch13].

3.5.1. Información temporal

La información temporal viene dada por datos que indiquen la presencia de notas musicales en determinados instantes temporales. La variabilidad de esta posible fuente de información va desde indicar los tiempos de onset, pasando por incluir, así mismo, tiempos de offset, hasta indicar el *pitch* de cada nota y su duración [Duan11]. Este tipo de información se suele representar sobre las ganancias, $\mathbf{G}$ ó $g_n(t)$ por su interpretación práctica, mediante una inicialización previa a la ejecución del algoritmo, como es el caso de [Ewert12] y [Hennequin11]. La información espectral la constituyen datos que describan el comportamiento espectral de las fuentes, sobre las que se suelen particularizar esos datos. La naturaleza armónica de las mismas puede constituir una información básica y muy útil para una buena descomposición de la señal.

3.5.2. Información espectral

A esta información puede agregarse otra, además de conocer la posición en frecuencia de los parciales, se les puede dar un valor de amplitud a cada uno. De esta manera, si se estructuran bien las matrices de funciones base $\mathbf{B}$ ó $b_{n,j}(f)$, éstas pueden representar una envolvente espectral para cada *pitch* considerado y para cada instrumento de la señal de entrada, como se hace en [Carabias11].

A la información espectral por instrumento se les suele denominar modelos de fuente, o modelos de instrumento, y es habitual el uso de esta información se para la inicialización de las funciones base. Los datos de inicialización suelen obtenerse de un proceso previo de entrenamiento que permita obtener información del espectro generado por cada *pitch*. Los mo-

delos descritos en el apartado 3.4.2 encajan perfectamente con esta idea de la modelización de las fuentes. Una vez inicializadas las funciones base, éstas se usan para representar la señal polifónica mediante la suma ponderada de dichas funciones, se estiman las ganancias mientras las bases permanecen fijas. Sin embargo, durante el proceso de descomposición las funciones base podrían ser actualizadas bajo determinadas condiciones, con el objetivo de obtener una mejor descomposición de la señal de entrada.

3.5.3. Aplicaciones con información previa de las fuentes

En la bibliografía, es habitual usar para el entrenamiento de las funciones base el mismo método que se usará para descomponer la señal. Por ejemplo en [Jang03] se usa ICA para ambos procedimientos, el aprendizaje de las funciones base y la descomposición de la señal. En [Carabias11], se usa NMF para un proceso de entrenamiento de modelos de instrumento parametrizados que generan una función base para cada pitch, las cuales, son datos de entrada para el proceso de descomposición. Benaroya et al. en [Benaroya03] comparan dos métodos de entrenamiento, uno de ellos es el mismo que usan para la descomposición, *Non-negative Sparse Coding* y por otro lado, tratan de utilizar el espectro de tramas aleatorias de la señal de análisis como funciones base para la descomposición. Schmidt y Olsson [Schmidt06] proponen el uso de *Non-negative Sparse Coding* para entrenamiento y factorización de la señal de entrada. Estos métodos realizan el entrenamiento sobre la misma señal que se desea descomponer, sin embargo, se puede realizar un entrenamiento, mucho más supervisado, sobre un conjunto de señales independientes y específicas para la obtención de la función base de cada *pitch*. Este es el caso del sistema de transcripción percusiva de FitzGerald et al. [FitzGerald03], donde la función base de cada elemento percusivo se obtiene de una señal aislada de dicho elemento. Otros trabajos de la bibliografía describen como se pueden generar las funciones base iniciales de manera manual, como en el caso de [Lepain99].

En la práctica, es complicado obtener unas funciones base adecuadas al instrumento real de la señal mezclada y realizar una buena descomposición de la señal de entrada. Por tanto, una alternativa es inicializar las funciones base con las que se obtengan de una fase de entrenamiento previa, o dise-

ñadas de manera manual, y posteriormente permitir al algoritmo que las adapte y perfeccione con la información de la señal a descomponer, como en el caso de [Abdallah04]. *Sparse Coding* y *NMF* son métodos muy apropiados para este tipo de aprendizaje, puesto que el error de reconstrucción se minimiza cuando se activan un número pequeño de funciones base (restricción de dispersión). Esta posible mejora de los modelos de instrumento en tiempo de descomposición motiva y constituye una de las propuestas de esta tesis, desarrollada en el Capítulo 7.

3.6. Conclusiones

En este capítulo se han introducido los conceptos básicos de descomposición de señal (apartado 3.1), así mismo se ha descrito el método de descomposición de señal más usado en los últimos tiempos, NMF (apartado 3.4). Este método de descomposición de señal se ha utilizado en diversas aplicaciones de procesado de señal, como la transcripción musical con modelos de instrumento [Carabias11] o separación de fuentes [Fritsch13]. Se han planteado sus conceptos principales y algunas propuestas de algoritmos flexibles para perfilar y mejorar el modelo de cara a la aplicación para la que se desea desarrollar. Finalmente, en el parrado 3.5 se indican, brevemente, las posibles fuentes de información de las que se puede nutrir el sistema de separación. Algunos trabajos de la bibliografía, como en [Fritsch13] y en [Ewert12], hacen uso de estos tipos de información para la mejora de los resultados en la factorización y separación de fuentes musicalesEstas fuentes son fundamentales para obtener resultados aceptables en la separación de señales monoaurales.

Capítulo 4

Estado del arte en separación de fuentes musicales

Este capítulo se divide en cuatro partes. En la primera de ellas (apartado 4.1) se describen las bases de datos más usadas para la evaluación de los sistemas de separación de fuentes, así como las usadas en fases de entrenamiento de modelos de instrumento. En el apartado 4.2 se detallan las medidas usadas para evaluar los resultados obtenidos con los métodos de separación y, de esa manera, poder compararlos con otros métodos del estado del arte. En el apartado 4.3 se muestran las primeras propuestas para la separación de fuentes musicales. Finalmente, en el apartado 4.4 se hace una revisión de los métodos de separación de fuentes de la bibliografía.

4.1. Bases de datos musicales

4.1.1. Introducción

No existe una base datos estándar para evaluar los sistemas de separación de fuentes musicales. Sin embargo, sí que existen un conjunto de bases de datos que son ampliamente utilizadas para este objeto. Su establecimiento como bases de datos de evaluación viene avalado por su rigurosidad y la aceptación por parte de la comunidad científica.

Las bases de datos musicales pueden dividirse en dos grupos: bases de datos de sonidos sintéticos y bases de datos de sonidos reales. Las primeras

son bases de datos compuestas por señales cuyas notas han sido generadas por un sintetizador de ficheros simbólicos MIDI (sintetizadores profesionales, que son capaces de generar sonidos muy similares a los de un instrumento real). Por otro lado, las bases de datos de sonidos reales (generados por instrumentos físicos reales, tocados por músicos profesionales) contienen señales grabadas por músicos profesionales. Además, estas señales reales, contienen las particularidades de la interpretación del músico, las características acústicas de la sala y del propio instrumento. Ambos tipos de bases de datos son usados en la bibliografía, sin embargo, es conveniente trabajar con bases de datos de instrumentos reales, por ser más cercanos a un escenario real. Así mismo, estas bases de datos son convenientes cuando el sistema trabaja con modelos de instrumento, puesto que la información de los modelos se ajustará más a un instrumento real, con todas sus particularidades.

Las bases de datos de sonidos reales, a su vez, pueden dividirse en dos subgrupos:

- Bases de datos de sonidos individuales. Estas bases de datos se componen de un conjunto de notas musicales tocadas de manera aislada por distintos instrumentos (también de distintas familias), músicos, estilos, ... Las notas se pueden encontrar en un orden aleatorio o siguiendo determinados criterios (numero máximo de notas concurrentes, número de instrumentos,...) para obtener diferentes tipos de mezclas polifónicas. Así mismo, las notas se encuentran aisladas, de manera que se puede tener todas las notas del rango dinámico de un determinado instrumento para analizarlas por separado. Algunas de las bases de datos más conocidas son *Musical Instrument Sound Database* de *Real World Computing (RWC)*, [Goto02] [Goto05], *University of Iowa Musical Instrument Samples* [Iowa06], y *McGill University's Master Samples (MUMS)* [McGill92]. Estas bases de datos se usan para realizar entrenamiento de modelos de instrumentos, o bien, para construir señales de entrada con determinadas características deseadas.

- Bases de datos compuestas de fragmentos de composiciones musicales. Este tipo de bases de datos son muy útiles para la evaluación de

los sistemas de separación de fuentes en diferentes escenas con interpretaciones de distinto tipo (distintos instrumentos, acústicas de sala diferentes, distinto nivel de polifonía,...). Además de ello, permiten comparar los resultados obtenidos con los de otros métodos del estado del arte, por el hecho de haber usado las mismas bases de datos y conjuntos de ficheros. Las bases de datos más utilizadas son: *Classical Music Database and the Jazz Music Database* de *Real World Computing (RWC)* [Goto02] [Goto05] y *The Bach Chorals Dataset* propuesta en [Duan11].

4.1.2. *RWC Musical Instrument Sound Database*

Esta base de datos incluye 50 instrumentos musicales. Para ofrecer una amplia gama de sonidos, de cada uno de ellos se ofrecen:

- Variaciones (3 fabricantes de instrumentos, 3 músicos): Cada variación se distingue por ser interpretada con un instrumento de distinto fabricante y un intérprete diferente.
- Técnica interpretativa: Dentro de las posibilidades del instrumento del que se trate, contiene grabaciones con diferentes técnicas interpretativas(normal, staccato, vibrato, ...).
- *Pitch* (todo el rango): Para cada estilo interpretativo, el músico toca, de forma individual, todo el rango de notas que el instrumento es capaz de generar con intervalos de un semitono.
- Matiz dinámico (3 niveles dinámicos): La base de datos también contiene grabaciones con cada uno de los tres niveles dinámicos (forte, mezzo, piano), abarcando, igualmente, todo el rango de notas del instrumento.

Por ejemplo, para el elemento de la base de datos RWC-MDB-I-2001 No. 01 (Piano) las 88 notas del piano están grabadas usando tres pianos de tres fabricantes distintos (Yamaha, Bösendorfer, and Steinway), cuatro estilos interpretativos diferentes (normal, staccato, pedal, y repetición del mismo sonido) con tres niveles dinámicos (forte, mezzo y piano). En total, hay 3168 ($3 \times 88 \times 4 \times 3$) sonidos individuales para este elemento.

El sonido de estos 50 instrumentos está grabado con 16 bits a $44100Hz$ en 3544 ficheros de audio monoaural, lo que supone un total de 29,1 Gbytes y un tiempo de grabación de 91,6 horas (incluyendo los intervalos de silencio entre las notas). Cada fichero contiene una secuencia de sonidos individuales ordenados ascendentemente según su *pitch* para completar todo el rango de posibles notas del instrumento.

En esta tesis se han usado notas individuales de clarinete, fagot, saxofón y violin, para los experimentos finales, además de algunos otros instrumentos para experimentos parciales e intermedios.

4.1.3. *McGill University's Master Samples (MUMS)*

La base de datos MUMS [McGill92] contiene más de 6000 muestras de sonido que representan la mayoría de los instrumentos musicales clásicos y populares con una gran variedad de estilos de interpretación. En concreto, contiene 6546 notas grabadas, que se dividen en 2204 de cuerda, 1595 de teclado, 1197 de viento madera 1087 de percusión y 463 de viento metal. La duración de las notas se encuentra entre 2 y 10 segundos, según el instrumento que la interprete. Cada nota de cada instrumento está grabada de manera independiente en un fichero a $44100Hz$ de frecuencia de muestreo y 24 bits de cuantificación.

4.1.4. *RWC Classical Music Database database*

Esta base de datos puede ser útil para una evaluación de separación de fuentes si se mezclan varios ficheros, puesto que cada grabación corresponde a un único instrumento. Ha sido ampliamente usada en la evaluación de sistemas de transcripción musical automática, principalmente por las siguientes razones:

1. Las grabaciones de esta base de datos son de instrumentos polifónicos reales (no sintéticos) en diferentes condiciones y escenas musicales.

2. Durante los últimos años, la base de datos de RWC ha tenido una atención especial por la comunidad científica y ha sido usada para la evaluación de múltiples sistemas de transcripción musical, separación

de fuentes y extracción de melodía, entre otros. Este hecho facilita la comparación con otros métodos del estado del arte.

3. La calidad de las grabaciones es alta, mayor que otras bases de datos que han caído en desuso por esta razón.

4. Se tienen los ficheros MIDI alineados de todas las interpretaciones. Esta información puede ser usada como referencia a la hora de medir la calidad de los resultados de transcripción musical, o incluso para ser usada como información de entrada a un posible sistema de separación de fuentes.

4.1.5. *Bach Chorals Dataset*

Esta base de datos ha sido diseñada para que pueda ser usada en la evaluación de diversos campos de investigación activos, como la estimación *multi-pitch*, el *Audio-score Alignment*, separación de fuentes, etc. Esta base de datos contiene grabaciones de cada instrumento del grupo de diez piezas corales de J.S. Bach. Así mismo, contiene los ficheros MIDI correspondientes, el alineamiento ideal del audio con el fichero simbólico y los valores de *pitch* reales para cada instrumento. Las grabaciones de las cuatro partes de cada pieza (soprano, alto, tenor y bajo) se han realizado usando violín, clarinete, saxofón y fagot, respectivamente. La grabación de cada parte se ha realizado de manera independiente, mientras que el músico escucha por unos auriculares las grabaciones de los demás instrumentos.

Los datos de pitch y notas de cada parte han sido obtenidos de manera independiente y posteriormente han sido combinadas todas las fuentes de información. Para su análisis han usado tramas de $46ms$ y salto de $10ms$. Por otro lado, para la obtención del valor de *pitch* de cada parte (instrumento) han empleado un algoritmo robusto de detección de *pitch* llamado YIN [Cheveigne02]. Han procesado con este algoritmo la grabación de cada instrumento de manera independiente para detectar el *pitch* generado en cada trama, posteriormente han repasado manualmente para resolver pequeños errores obvios.

El alineamiento entre la señal de audio y el fichero MIDI ha sido llevado a cabo de manera manual, mediante la marcación del *beat* (ritmo) de la

interpretación por un músico experto mientras escucha la grabación. De esa manera han asociado el *beat* de la grabación y del fichero MIDI para obtener el alineamiento por interpolación entre cada golpe de *beat*.

4.1.6. Base de datos de instrumentos polifónicos de viento madera

Para evaluar los modelos descritos en el capítulo 5, se ha usado la base de datos de instrumentos de viento madera también usados en [Vincent10]. Las señales polifónicas se han generado con la mezcla de grabaciones de instrumentos individuales procedentes de la prueba de algoritmos de estimación *multi-pitch* de *Third Music Information Retrieval Evaluation Exchange (MIREX2007)* [MIREX2007]. Esta base de datos se compone de un quinteto de la quinta variación de L. van Beethoven para cuarteto de cuerda Op.18 No. 5. Cada parte (flauta, oboe, clarinete, trompa y fagot) se a grabado de manera independiente, mientras el intérprete escuchaba el resto de instrumentos con auriculares. La mezcla de las señales se ha realizado sobre los primeros 30 segundos de interpretación de las piezas monofónicas, como se hace en [Vincent10]. Con esta configuración se pueden obtener niveles de polifonía desde 2 hasta 5, por tanto un total de 26 señales polifónicas.

4.2. Medidas de evaluación

En todos los campos de investigación es necesario obtener unas medidas que permitan evaluar la calidad objetiva de los resultados de cualquier método propuesto. Para la separación de fuentes se han propuesto varios métodos de evaluación, aunque no todos han sido ampliamente usados por la comunidad científica. En este apartado se describen las medidas objetivas que se han usado para medir la calidad de los sistema de separación de fuentes en esta tesis.

Lambert [Lambert99] y Schobben y Torkkola [Schobben99] fueron los primeros en proponer una medida de evaluación genérica de la separación BSS para poder comparar los resultados de diversos sistemas de mejora y separación de voz. Posteriormente se han desarrollado otros métodos de evaluación específicos para cada trabajo con evaluaciones basadas en medidas

indirectas de los resultados o test subjetivos. Con este conjunto de medidas no es posible realizar una buena comparación de los resultados obtenidos por los distintos métodos de separación.

Schobben y Torkkola [Schobben99] proponen una métodolología de evaluación de la separación de fuentes y un conjunto de datos para que sean, ambos, empleados para la comparación de los algoritmos. Para ello proponen dos casos diferenciados. Uno de ellos radica en la separación controlada de un conjunto de señales sintéticas, que resulta muy útil para conocer los límites del algoritmo, al poder escoger las señales con las que se va a evaluar. El otro consta de un conjunto de señales reales, con los correspondientes sonidos aislados, y que se encuentran disponibles en la red para la comunidad científica. Los autores proponen una serie de parámetros que definen la dificultad del problema que se pretende evaluar, así como los procedimientos para realizar la mezcla de los sonidos grabados. Es importante indicar que durante la grabación de las fuentes individuales también se produce la mezcla de las demás fuentes. En el desarrollo, $v_{o,s_j}(t)$ representa la mezcla $v_o(t)$ en la grabación de la fuente $s_j(t)$ y $\hat{s}_j(t)$ representa la fuente separada estimada. La primera medida propuesta es la de distorsión, una medida de cuán distorsionada está la fuente estimada respecto de la fuente real. Esta medida se define como:

$$D_j^{distortion} = 10\log_{10}\left(\frac{\sum_t \left| v_{j,x_j}(t) - \frac{v_{j,x_j}(t)}{\hat{s}_j(t)} \right|^2}{\sum_t \left| v_{j,x_j}(t) \right|^2}\right) \tag{4.1}$$

donde el índice j indica la fuente $j-sima$. La segunda de las medidas evalúa la cantidad de señal que se han conseguido separar, de manera que:

$$D_j^{separation} = 10\log_{10}\left(\frac{\sum_t \left| \hat{s}_{j,x_j}(t) \right|^2}{\sum_t \left| \sum_{i \neq q} \hat{s}_{q,x_i}(t) \right|^2}\right) \tag{4.2}$$

donde $\hat{x}_{q,x_i}$ es la $q-sima$ salida del sistema de mezcla en cascada cuando sólo está activa x_i.

Vincent et al. [Vincent03a] proponen dos grupos de medidas de evaluación de BSS. La primera de ellas *Audio Quality Oriented (AQO)*, y la segunda *Significance Oriented (SO)*. El objetivo del primer grupo es obtener una buena relación señal/ruido (SNR) y reducir los artefactos en las señales separadas. Por otro lado, el segundo grupo se centra en la obtención de características de la escena de audio mediante la estimación de las fuentes y/o los parámetros del proceso de mezcla.

Siguiendo esta misma filosofía, Vicent et al. [Emiya11] proponen dividir el término de error total , entre la señal estimada por el sistema $\hat{s}_j(t)$ y la señal real $s_j(t)$, en tres términos que se relacionan con tres tipos de error:

$$\hat{s}_j(t) - s_j(t) = e_j^{target}(t) + e_j^{interf}(t) + e_j^{artif}(t) \tag{4.3}$$

donde $e_j^{target}(t)$ es el término de error de objetivo, $e_j^{interf}(t)$ es el término de error por interferencia del resto de fuentes sobre la fuente j y $e_j^{artif}(t)$ es el término de error atribuido a los artefactos generados por el algoritmo de separación.

La distorsión total entre la fuente estimada y la fuente real se define como:

$$D_j^{total} = \frac{\sum_t |\hat{s}_j(t) - s_j(t)|^2}{\sum_t |s_j(t)|^2} \tag{4.4}$$

Y usando dichos términos de error, se establecen las distintas distorsiones relativas a cada uno de ellos. La distorsión causada por el error de objetivo:

$$D_j^{\text{target}} = \frac{\sum_t \left| e_j^{target}(t) \right|^2}{\sum_t |s_j(t)|^2} \tag{4.5}$$

La distorsión asociada a la interferencia proveniente de otras fuentes:

$$D_j^{\text{interf}} = \frac{\sum_t \left| e_j^{interf}(t) \right|^2}{\sum_t \left| s_j(t) + e_j^{target}(t) \right|^2} \tag{4.6}$$

Y la distorsión asociada a los artefactos generados por el algoritmo de separación:

$$D_j^{\text{artif}} = \frac{\sum_t \left| e_j^{artif}(t) \right|^2}{\sum_t \left| s_j(t) + e_j^{target}(t) + e_j^{interf}(t) \right|^2} \tag{4.7}$$

De esta manera, se pueden definir los valores de *Signal to Distortion Ratio (SDR)*, *source Image to Spatial Ratio (ISR)*, *Signal to Interference Ratio (SIR)*y *Signal to Artifact Ratio (SAR)* como:

$$SDR_j = 10\log_{10}\left(\frac{1}{D_j^{total}}\right) \tag{4.8}$$

$$ISR_j = 10\log_{10}\left(\frac{1}{D_j^{target}}\right) \tag{4.9}$$

$$SIR_j = 10\log_{10}\left(\frac{1}{D_j^{\text{interf}}}\right) \tag{4.10}$$

$$SAR_j = 10\log_{10}\left(\frac{1}{D_j^{artif}}\right) \tag{4.11}$$

Este esquema de medidas no representa la calidad perceptual de la separación de fuentes, lo que se considera como su principal inconveniente. Las medidas anteriores no tienen en cuenta algunos fenómenos como el enmascaramiento espectral o la intensidad percepción del sonido por el oído humano. Este inconveniente se intenta solventar en [Emiya11], donde se propone un protocolo de pruebas subjetivas para la evaluación de la calidad de percepción en separación de fuentes de audio. Las nuevas medidas se basan en la evaluación de los términos de error en un conjunto de bandas, relacionadas con las particularidades de la percepción del sonido. Estas medídas se llaman *Overall Perceptual Score (OPS)*, *Target-related Perceptual Score (TPS)*, *Interference-related Perceptual Score (IPS)* and *Artifacts-related Perceptual Score (APS)*. Todas ellas componen el paquete *Perceptual Evaluation of Audio Source Separation (PEASS)*, que se encuentra disponible en la red[1].

[1]http://bass-db.gforge.inria.fr/peass/

Estas medidas perceptuales, propuestas en [Emiya11], se implementan en dos fases. En la primera de ellas se emplea el modelo perceptual PEMO-Q de [Huber06]. Con este banco de filtros se obtiene una medida de similaridad perceptual *Perceptual Similarity Measure (PSM)*. La importancia perceptual de la distorsión media de cada componente ($q_j^{total}, q_j^{target}, q_j^{interf}, q_j^{artif}$) se evalúa comparando cada señal estimada con ella misma menos el término de error considerado, de manera que:

$$q_j^{total} = PSM\left(\hat{s}_j, s_j\right) \tag{4.12}$$

$$q_j^{target} = PSM\left(\hat{s}_j, \hat{s}_j - e_j^{target}\right) \tag{4.13}$$

$$q_j^{interf} = PSM\left(\hat{s}_j, \hat{s}_j - e_j^{interf}\right) \tag{4.14}$$

$$q_j^{artif} = PSM\left(\hat{s}_j, \hat{s}_j - e_j^{artif}\right) \tag{4.15}$$

En una segunda fase se combinan las cuatro medidas con un mapeado no lineal y se adaptan a una escala de medidas subjetivas. En el caso de la propuesta de [Emiya11], se usa una red neuronal de una capa oculta, que ha sido entrenada previamente, como esquema no lineal de mezcla de todas la medidas $q_j^{total}, q_j^{target}, q_j^{interf}, q_j^{artif}$, para obtener los valores de OPS, TPS, IPS y APS.

4.3. Primeros pasos en la separación de fuentes monoaural

Los primeros pasos en la separación de fuentes de sonido se dieron en el campo de la separación de señales de voz [Lee88, Parsons76, Quatieri90, Weintraub84]. Sin embargo, en los últimos años los grandes avances en el procesado digital de señal de audio han hecho que las técnicas de separación de fuentes musicales avancen notablemente.

El caso de la separación musical, en algunos aspectos, se puede ver como un reto mayor que el de la señal de voz. Los instrumentos musicales

tienen una mayor variedad de mecanismos de producción de sonido distintos, mientras que la generación de voz cuenta con el mismo mecanismo en todos los individuos. Además, las características espectrales y temporales del sonido de los instrumentos musicales presenta, también, una gran variabilidad. En general, cuando dos personas están hablando, se intercambian los turnos, de manera que en pocas ocasiones se producen solapamientos de ambas fuentes, caso contrario al de los instrumentos musicales, los cuales producen sonidos simultáneos en la mayor parte del tiempo.

Las propuestas relativas a la separación de fuentes de audio en señales monoaurales se pueden clasificar en tres categorías: métodos basados en modelos, métodos de aprendizaje no supervisado y métodos psicoacústicos. A pesar de esta clasificación, la mayoría de los algoritmos implementan características de varios de estos tipos de métodos.

- **Métodos basados en modelos**

  Estos métodos usan modelos parámetros que describan el comportamiento espectral y/o temporal de las fuentes presentes en la señal. Los parámetros pueden se pueden aprender desde señales específicas de entrenamiento, o bien, de la propia señal mezclada. Algunos algoritmos que usan este tipo de información previa sobre las fuentes son [Every06, Maher90, Quatieri90].

- **Métodos de aprendizaje no supervisado**

  Los métodos de aprendizaje no supervisados usan modelos simples no paramétricos, y con menos información previa sobre las fuentes presentes en la señal. Para suplir esta carencia de información, tratan de aprenderla desde la misma señal que se pretende separar. Estos métodos suelen usar información de principios teóricos, como la suposición de independencia entre las fuentes. La reducción de la redundancia de información produce representaciones de las señales aisladas. Algunos algoritmos que siguen estos métodos son ICA, NMF y SC, como en [Jutten91, Saruwatari01].

- **Métodos psicoacústicos**

  Las habilidades cognitivas humanas permiten pervivir y reconocer las fuentes independientes de un sonido mezclado. Generalmente, los mé-

todos de este tipo cuentan con dos fases. En la primera la señal de entrada se descompone en componentes tiempo/frecuencia básicas, siendo éstas clasificadas entre las diferentes fuentes en la segunda fase, mediante el seguimiento de determinadas reglas de clasificación. Con ellas, algunos trabajos han implementado algoritmos de separación de fuentes [Cooke93, EllisPhD96]. En los últimos tiempos, las reglas usadas por estos métodos han sido consideradas insuficientes para problemas de separación complejos [Smaragdis97]. Es probable que el sistema auditivo humano use principios tanto innatos como aprendidos para la separación mental de fuentes [BregmanBook90] y que sea la fisiología del sistema auditivo periférico aplique los mecanismos innatos de bajo nivel [BregmanBook90]. A pesar de que algunos de los mecanismos de separación de alto nivel también se suponen innatos, no se conoce mucho el efecto real del aprendizaje [BregmanBook90]. Aunque el sistema auditivo humano no es capaz de sintetizar las fuentes por separado, es un sistema de referencia para el desarrollo de los sistemas automatizados de separación de fuentes desde señales monoaurales, puesto que es el único sistema robusto que lo consigue en un amplio conjunto de escenarios.

4.4. Técnicas de separación de fuentes musicales

Aunque la clasificación de métodos de factorización de señal sigue un patrón bien definido, la clasificación de los algoritmos de separación de fuentes no es una tarea fácil, puesto que su diseño es muy flexible y admite multitud de variantes. En este apartado se realiza una clasificación de los algoritmos en función de la información con la que cuenten para llevar a cabo la tarea de la separación de fuentes. Aún así, la clasificación no es estricta, puesto que algunos algoritmos pueden combinar información de distintos tipos y su estructura puede variar para adaptarse al problema. En concreto, se han establecido tres categorías para clasificar los algoritmos: separación sin información previa, separación con información temporal y separación con información espectral.

Esta tesis pretende valorar el uso de varios tipos de información en el núcleo de factorización sobre un esquema de separación de fuentes mono-

aural. En concreto se valorarán el uso de información de instrumento con modelos parametrizados que describan sus características físico-acústicas e información temporal sobre la activación de las notas.

4.4.1. Algoritmos de separación a ciegas

Se consideran algoritmos de separación a ciegas, aquellos sistemas de separación de fuentes que no cuentan con información temporal de la secuencia de notas que toca cada instrumento, ni con información espectral específica de los instrumentos presentes en la composición. Un algoritmo de separación que no cuente con un mínimo de información sobre la señal que se pretende separar, no obtendrá resultados aceptables a menos que se le agreguen ciertas restricciones que ayuden a explotar las posibilidades de los métodos de factorización de manera ordenada. Por tanto, también se pueden considerar como algoritmos de separación a ciegas aquellos que cuentan con cierta información respecto al comportamiento general de la señal. Por ejemplo, restricciones de suavidad temporal, dispersión, armonicidad, independencia estadística y ortogonalidad de las fuentes, etc.

Algoritmos de no correlación

Cuando no se cuenta con mucha información sobre ellas presentes en la señal de entrada, una de las suposiciones más simples e inmediatas que se puede aplicar es la de no correlación de las fuentes.

Uno de los primeros algoritmos de separación de fuentes de este tipo, fue propuesto por Weinstein et al. en [Weinstein93b]. Este trabajo se basa en la no correlación de las fuentes combinado con cierta información previa sobre las fuentes que permita ofrecer mejores resultados. En [Belouchrani97] se propone un algoritmo basado en la coherencia temporal de las fuentes. En él se usan un conjunto de matrices de covarianza que son diagonalizadas para obtener unas variables estadísticas estacionarias de segundo orden, en base a las cuales realizan la clasificación de las fuentes. Ziehe y Muller [Ziehe98] proponen otro algoritmo BSS llamado TDSEP, que se basa en un conjunto de matrices de correlación de segundo orden. La matriz de separación se obtiene mediante una decorrelación de las matrices anteriores y seguida de una diagonalización de las matrices resultantes. Un sistema similar fue

propuesto por Parra y Spence [Parra00], en el que convierten el problema en una descomposición en valores singulares (SVD). En [Parra98b] y [Parra04] también se usa una decorrelación entre las fuentes en un esquema de BSS por descomposición en valores singulares.

Independent Component Analysis (ICA)

ICA es una de las técnicas de BSS mas populares y se basa en la independencia estadística entre las fuentes. A diferencia de las técnicas PCA *(Principal Component Analysis)*, la medida de independencia de las fuentes no es la varianza (decorrelación), sino que se usan algunas momentos estadísticos de orden superior. Cuando se usan momentos de tercer o cuarto orden, hay que suponer la no gaussianidad de las fuentes.

Algunos de los momentos usados para medir la independencia de las fuentes, son: sesgo, curtosis, entropía o medida de maxima probabilidad. Habitualmente el proceso de separación se basa en un algoritmo del gradiente sobre estas medidas estadísticas dada una determinada matriz de separación.

La suposición de la independencia de las fuentes permite explotar otras propiedades estadísticas para realizar la separación. Por ejemplo, la consideración de las fuentes como variables no Gaussianas, dirigió los pasos hacia el método ICA [Jutten91, Comon94]. Esta consideración de la no gaussianidad de las fuentes es necesaria, puesto que el método ICA supone que la mezcla de las fuentes será más Gaussiana que las fuentes por el teorema central del límite.

Otros métodos emplean soluciones de maximización de la probabilidad sobre las varianzas no estacionarias y suaves que presentan las fuentes [Pham01].

Finalmente, Parra y Saida [Parra04], demostraron que cuando se realizan suposiciones profundas sobre la naturaleza de las fuentes, la separación lineal de fuentes a ciegas es equivalente a la descomposición en valores singulares generalizada. Estas suposiciones son la independencia de fuentes no Gaussianas y la consideración de las fuentes como no estacionarias o dispersas.

Cuando se tratan señales de audio, estos métodos se pueden aplicar

en el dominio del tiempo *(Time Domain-ICA, TD-ICA)* considerando la propia forma de onda de las fuentes como variables aleatorias. Sin embargo, hay una restricción importante en su uso, requieren, al menos, un canal por cada fuente presente en la mezcla. En las señales musicales no es muy común, dado que el formato habitual es el estéreo de dos canales. Para evitar esta limitación se puede suponer cierto grado de desunión en tiempo/frecuencia, y aplicar estos métodos en el dominio de la frecuencia *(Frequency Domain-ICA, FD-ICA)*. En esta variante se usan los *bins* tiempo/frecuencia como variables aleatorias independientes y además, se necesita un paso adicional, en el que se agrupen los *bins* que pertenecen a cada instrumento. En [Nishikawa02] y [Nishikawa03], se desarrollan sistemas de separación de fuentes que usan, en cascada, métodos TD-ICA y FD-ICA, y en [Nishikawa02b] se realiza una comparación entre ambos métodos.

Algoritmos deterministas sin inicialización

Al igual que ocurre con los algoritmos estadísticos, los algoritmos deterministas necesitan un mínimo de información para poder aplicar ciertos criterios a la hora de separar las fuentes de sonido. Se pueden considerar algoritmos a ciegas aquellos algoritmos que aplican ciertas restricciones a la estructura de las matrices de descomposición, pero que no cuentan con una inicialización previa de éstas.

El esquema determinista más utilizado es NMF, el cual será la base sobre la que se sustentarán todas las propuestas de estas tesis. Este esquema ofrece mayor flexibilidad a la hora de agregarle restricciones o inicializaciones a la descomposición de la señal. Aunque este apartado sólo se refiere a las restricciones, dejando el uso de información inicial para los siguientes.

En general, el uso del algoritmo básico NMF, por si sólo, no ofrece unos resultados interpretables para una determinada aplicación. Un parámetro crítico en una descomposición NMF es la elección del número de bases. Conforme el número de bases se eleva, el error de reconstrucción se reduce, sin embargo las componentes van siendo cada vez menos significativas y contienen menos valor de interpretación. Existe un compromiso entre la información que obtienen las bases y la calidad de la reconstrucción de la señal con dichas bases [MaxerPhD13].

En el caso de las señales de audio, los espectros tienen una naturaleza cambiante en el tiempo para una determinada base. Si se estudia el caso de una nota de piano, desde su activación hasta su desactivación puede presentar distintas envolventes espectrales, correspondientes a sus fases (ataque, sostenimiento y decadencia) [Kameoka12]. Para analizar este tipo de espectros con NMF clásico, se debería usar una gran cantidad de bases, que requerirían un postprocesado para agruparlas en función de la nota a la que pertenezcan. Sin embargo, este incremento de bases hace que decrezca la información práctica que contienen de cada una de ellas, dejando de ser útiles para la aplicación. Para mantener su nivel de importancia simbólica se deben aplicar ciertas restricciones o regularizaciones que disminuyan la libertad del modelo [Virtanen07b].

Siguiendo la nomenclatura del capítulo 2, las propuestas de restricciones para los modelos NMF se realizan sobre las matrices $\mathbf{B}$ y $\mathbf{G}$ de la ecuación 3.12, o bien, sobre la función de coste de la ecuación 3.13.

En [Kameoka12], se asigna un conjunto de bases para cada posible nota en la señal de entrada. Este incremento en las bases se restringe en dos sentidos: sólo una base de cada conjunto puede activarse en cada instante, y por otro lado, las bases se establecen como una serie que se irán activando siguiendo una cadena de Markov. En este caso, ambas restricciones se efectúan directamente sobre las ganancias $\mathbf{G}$, aunque requieren de una ordenación adecuada de las bases.

En el mismo trabajo [Kameoka12], proponen otras restricciones relacionadas directamente con la naturaleza de la señal musical. En dicha propuesta, incluyen información de *tempo*, *beat* y *onset*. De esta manera las activaciones de las notas deben producirse en torno a determinados instantes descritos por estos parámetros y siguiendo un patrón de activación rápida y sostenimiento.

Virtanen en [Virtanen06] aplica dos tipos de restricciones a un modelo NMF básico. Estas restricciones son de continuidad temporal y dispersión, ambas sobre la matriz de ganancias $\mathbf{G}$. El algoritmo NMF propuesto usa como dato de entrada el espectrograma de la señal mezclada. En general NMF trata de manera independiente cada columna de $\mathbf{G}$, que representa una trama temporal. Sin embargo, en la señales musicales esto no es del todo cierto, cada trama temporal tiene una vinculación fuerte con las adyacentes,

puesto que cada evento nota perdura durante un conjunto de tramas, por tanto la activación de una determinada base en una trama t está ligada a provenir de la activación de la misma en la trama $t-1$, su activación en la trama $t+1$, o ambas cosas. Esta restricción se implementa asignando un término de coste adicional para los grandes cambios en las ganancias. Se demuestra en dicho trabajo que esta restricción hace más robusto el sistema de separación. En el caso de la dispersión de las ganancias, se implementan con otro término de coste adicional que penaliza las ganancias no nulas. Sin embargo esta restricción no muestra resultados significativos en cuanto a la calidad de la separación las fuentes.

4.4.2. Algoritmos de separación con información temporal

En este apartado se engloban los algoritmos de separación de fuentes que cuentan, a priori, con la secuencia de notas que interpretará cada instrumento. Esta información puede estar, o no, alineada. Es decir, se puede conocer únicamente la secuencia de notas, o bien, la secuencia junto con el tiempo y duración de cada una de ellas. Generalmente, los algoritmos que no cuentan con la secuencia alineada, realizan un proceso de alineamiento, o usan información simbólica idealmente alineada, por lo que a partir de ese momento tienen un funcionamiento muy similar. Esta información temporal se usa, generalmente, para inicializar la matriz de ganancias o activaciones $\mathbf{G}$. De esta manera se acotan bastante las ganancias que es necesario estimar, limitando la libertad del sistema de factorización y obteniendo unos resultados de superior calidad frente a los que permiten libertad total en esta matriz de activaciones. Grosso modo, puede considerarse como una restricción de dispersión estricta, se están indicando exactamente los *pitches* que se van a activar en cada momento, con la única necesidad de la estimación de las amplitudes correspondientes. El resto de *pitches* inicializados a cero, nunca se activarán. Sin embargo, no todos los algoritmos usan la información temporal de las notas para inicializar la matriz de ganancias.

Se describen en este apartado algunos de los algoritmos con información temporal, también llamados *Score-informed*, que se encuentran en la bibliografía. Algunos de ellos sólo cuentan con este tipo de información, aunque otros, la combinan con otras informaciones, como es el caso de los algorit-

mos propuestos en esta tesis, que además de la información temporal usan modelos espectrales de instrumento.

Algoritmos sin inicialización de la información temporal

Un pequeño grupo de los algoritmos que usan información temporal, no la usan directamente para la inicialización de las matrices de ganancias de los modelos deterministas. Estos algoritmos, que habitualmente usan modelos de separación probabilisticos, usan información simbólica por instrumento en formato MIDI, que se sintetiza y se usa de entrenamiento para obtener distribuciones a priori para cada una de las posibles notas tocadas por cada instrumento.

Ganseman et al. [Ganseman10] usan la información simbólica en formato MIDI por cada uno de los instrumentos. En una primera fase realizan un alineamiento de dicha información con la señal de entrada, de manera que obtienen una estimación de la posición temporal de cada una de las notas. El alineamiento lo realizan con una técnica ampliamente utilizada *Dynamic Time Warping (DTW)* [Ellis03, Turetsky03]. Posteriormente sintetizan cada instrumento en una señal de audio, las cuales serán funciones de probabilidad a priori que asignan a cada instrumento para el proceso de separación. El proceso de separación lo llevan a cabo con un algoritmo *Probabilistic Lantent Component Analysis (PLCA)* [Priestley74, Ganseman12], que sigue una filosofía similar a la de NMF pero desde un punto de vista probabilistico.

La calidad de la separación de la propuesta de Ganseman et al. [Ganseman10] depende, en gran medida, de la calidad del sintetizador que se utilice, no llegando a ser nunca la señal de un instrumento real. Así mismo, con el uso de la información simbólica de las notas que son tocadas, se restringe la posibilidad de activar notas erróneas en el algoritmo NMF. Se considera este aspecto como fundamental cuando se trata de separación monocanal. Sin la información temporal, los resultados de la separación a ciegas de señales monocanal dista mucho de poder ser empleada en sistemas comerciales.

Algoritmos con inicialización de la información temporal

La mayoría de propuestas que usan información temporal, además de otras posibilidades, la usan en la inicialización de ciertas matrices de activación de notas para el algoritmo NMF.

En el caso de [Ewert12] la propuesta consiste en inicializar tanto las matrices temporales de activaciones y las matrices de bases espectrales. Por tanto esta propuesta también puede enmarcarse en el siguiente apartado. En esta propuesta emplean ficheros MIDI que suponen idealmente alineados. De estos ficheros simbólicos extráen información de *onset, offset y pitch* para cada nota. La inicialización temporal que realizan es en dos matrices, una que contiene la activación de las notas en su fase sostenida y de decaimiento, y otra matriz que indica las zonas de onset de cada nota. Realizan esta separación porque en las matrices de bases contarán con dos tipos de bases para una misma nota, una que representará las fases de ataque y onset, y otro tipo de base para la misma nota en fase de sostenimiento y decaimiento. Las matriz de ganancias $\mathbf{G}$ se inicializa con valor 1 en aquellas posiciones tiempo/frecuencia que se consideran activadas y valor 0 en el resto. La matriz de bases $\mathbf{B}$ se iniciativa con valor 1 en las posiciones para la f_0 de cada base y sus correspondientes posiciones armónicas. Una vez inicializadas las matrices, la separación se lleva a cabo con un algoritmo NMF, filtrado de Wienner y transformada IFFT sobre los espectrogramas estimados correspondientes.

En [Hennequin11], se usa una inicialización aún más sencilla de la matriz de activaciones de NMF, puesto que no se discrimina entre las distintas zonas de las notas. La matriz de activación $\mathbf{G}$ del modelo NMF se iniciativa con valor 1 en las zonas en las que la información del fichero MIDI indica la activación de cada una de las notas, y valor 0 en el resto. La matriz de bases $\mathbf{B}$ se inicializa de manera aleatoria. A continuación, se ejecuta el algoritmo NMF de manera iterativa, minimizando una medida de divergencia entre el espectrograma de entrada y los estimados. De esta manera se obtiene una factorización de la señal de entrada que es sintetizada y evaluada.

Duan y Pardo en [Duan11], desarrollan un sistema de separación de fuentes musicales con información temporal. Esta propuesta incluye, como punto fuerte un bloque de alineamiento entre la información simbólica y

la señal de audio, que es capaz de ejecutarse en tiempo real. El bloque de separación de la fuentes se basa en la información alineada, generando unas máscaras de separación con la información de *picth* y suponiendo que todas las notas y todos los instrumentos tienen una envolvente espectral con caída exponencial. El bloque de alineamiento se analizará con más detalle en el apartado 7.2.1 de esta tesis. El bloque de separación no tiene un diseño profundo y realiza dos suposiciones que no son ciertas en la mayoría de las ocasiones, no todos los instrumentos tienen el mismo timbre, ni todas las notas de un instrumento generan la misma envolvente espectral. A pesar de esto, el sistema completo muestra resultados aceptables gracias a los buenos resultados de la fase de alineamiento.

4.4.3. Algoritmos de separación con información espectral

Los algoritmos de separación de fuentes con información espectral son menos frecuentes, pero a pesar de ello existen algunos sistemas que, tras un proceso de aprendizaje de cierta información espectral, la usan como punto de partida para la obtención de una mejor separación de las fuentes.

Hasta ahora, los modelos de fuente, o modelos de instrumento, han sido propuestos en trabajos que pretendían mejorar la factorización de las señales en modelos de descomposición como NMF, por ejemplo en [Virtanen06]. En este trabajo se propone un modelo paramétrico de cada fuente con una excitación armónica de amplitud constante y un filtro de instrumento que modele la respuesta del cuerpo resonante del mismo. De esta manera todas las bases espectrales, de cada pitch del mismo instrumento estarán relacionadas por la estructura del filtro. Esta propuesta, a pesar de no ser un sistema de separación en sí mismo, es una de las bases iniciales para esta tesis. Se considera que un modelo paramétrico de fuente puede ser beneficioso para obtener una buena separación de las mismas. Además su estructura paramétrica permite su actualización, de manera controlada, durante el proceso de factorización. En la sección 3.4.2 se describe este modelo de filtro-fuente.

Siguiendo la misma línea, en [Carabias11], se propone un modelo paramétrico que amplía al anterior. En este caso, la excitación armónica se divide en multiples vectores de excitación, los cuales se combinan linealmente para cada *pitch*, de esta manera pueden adaptarse a cambios de comportamien-

to de la excitación en función del *pitch*. La excitación armónica resultante, cuyos parciales ya contienen información sobre la envolvente espectral de la nota, es multiplicada por el filtro de resonancia del instrumento como en el caso anterior. En [Carabias11] no se aplica este modelo de instrumento para la separación de fuentes, pero este será el modelo de instrumento empleado en las propuestas de esta tesis para el desarrollo de un sistema de separación de fuentes flexible y adaptable a la señal de entrada. En la sección 3.4.2 se describe este modelo de multiexcitación con filtro-fuente.

En [Fritsch13], se usa la la información simbólica en formato MIDI como dato de entrada, por tanto, podría ser enmarcado en el apartado anterior. Se ha incluido en este apartado porque su objetivo final, con el uso de la información temporal en MIDI, es la inicialización de matrices con información espectral para el proceso de factorización. En este caso no aplican algoritmo de alineamiento, usan la información simbólica idealmente alineada para sintetizar las señales. Una vez sintetizadas, ejecutan una fase de entrenamiento, donde un algoritmo NMF obtiene una base por cada nota generando así un modelo de instrumento no paramétrico (una base, independiente de las demás, para cada nota). Los parámetros de este modelo se usan como inicialización para una segunda fase de factorización con NMF, en este caso determinista, para finalmente separar las señales estimadas mediante filtrado de *Wienner* [Fevotte11b] sintetizar las señales separadas estimadas mediante

FASST *Flexible Audio Source Separation Toolkit* [Ozerov11] es un sistema de separación de fuentes flexible y configurable para realizar diferentes tipos de separación con distintos métodos de factorización. Este sistema está inicialmente concebido para trabajar con señales multicanal, sin embargo, su estructura configurable permite limitarlo para el uso con señales monoaurales. Esta herramienta de separación podría encuadrarse en cualquiera de los apartados, puesto que su estructura permite incluirle información temporal, o no, así como realizar, o no, un proceso de entrenamiento de bases espectrales para inicializar el proceso de factorización. Este esquema permite seleccionar el modelo de descomposición basado en HMM o NMF, y sobre ambos se pueden establecer otras restricciones como la monofonía o gaussianidad de las fuentes. La descomposición la lleva a cabo mediante el algoritmo *Generalized Expectation Maximization (GEM)*. Además de

la flexibilidad en los modelos de factorización, es flexible en los modelos de representación de las señales. Se puede usar tanto la representación FFT clásica como la QERB *(Quadratic Equivalent Rectangular Bandwidth)*, siendo esta última más adecuada para las señales musicales, por su resolución en escala logarítmica.

4.5. Conclusiones

En el primer apartado de este capítulo se describen las bases de datos que se han usado para evaluar los algoritmos propuestos en esta tesis. Las bases de datos se han seleccionado en función de las necesidades de entrenamiento y evaluación de cada algoritmo. Para el entrenamiento se necesitan bases de datos bien anotadas para poder tener certeza de lo que el algoritmo va a encontrar en el análisis de la señal. Así mismo, para la evaluación se necesitan bases de datos de señales reales que cuenten con las señales totalmente aisladas, para poder compararlas con las que estima el algoritmo y establecer medidas de la calidad de la separación realizada.

En el segundo apartado se describen las medidas usadas para evaluar la separación de fuentes de audio. Estas medidas son ampliamente utilizadas en las propuestas de la bibliografía relacionada. Por tanto la comparación entre los resultados de otros algoritmos y los propuestos, se puede realizar utilizando las mismas bases de datos y las mismas medidas de calidad de separación.

En el tercer apartado se describen los primeros trabajos que se llevaron a cabo el el campo de la separación de fuentes de audio, así como su evolución hasta llegar a lo que hoy en día se encuentra en la bibliografía.

Finalmente, en el cuarto apartado, se describe una revisión de las propuestas de algoritmos de separación de fuentes musicales mas importantes del estado del arte. Se han tratado de clasificar en función del tipo de información a priori sobre las fuentes que utilizan. A pesar de ello, no es una clasificación cerrada, puesto que algunos de ellos se pueden incluir en varios apartados.

Parte III

Contribuciones

Capítulo 5

Separación de fuentes con modelos de instrumento mediante factorización NMF

En este capítulo se propone un sistema de separación de fuentes mediante factorización con NMF. Se hace uso de modelos de instrumento, previamente entrenados, para inicializar la matriz del algoritmo que contiene las funciones base para la descomposición, mientras que la matriz de activaciones se deja libre para ser estimada por el sistema. La obtención de los modelos de instrumento se realiza con varios modelos paramétricos, de esta manera se evalúa el rendimiento de la separación con cada uno de ellos. Los modelos estudiados son: 1. Modelo NMF básico con restricción armónica, 2. Modelo filtro-fuente con excitación armónica unitaria y 3. Modelo filtro-fuente con excitación múltiple. Los parámetros de los modelos se obtienen con un algoritmo NMF ampliado en una fase previa de entrenamiento o aprendizaje sobre señales aisladas de cada instrumento. Una vez realizado el entrenamiento se tiene una función base para cada *pitch*. Estas funciones base quedan fijas en el proceso de factorización y el algoritmo NMF se encarga de estimar las ganancias para cada una de ellas a lo largo del tiempo. La obtención de una matriz de ganancias para cada instrumento, permite utilizar la reconstrucción del modelo para generar unas máscaras de Wiener para realizar la separación. Finalmente se realiza una comparación de la

separación realizada con los distintos modelos de instrumento y el nivel de polifonía

5.1. Introducción

Un espectrograma de una señal de audio se puede descomponer como la combinación lineal de un conjunto de funciones base. Siguiendo este modelo, el espectrograma de amplitud de la señal $X(f,t)$ en la trama t para cada frecuencia f se modela como la suma de funciones base, de manera que:

$$\hat{X}(f,t) = \sum_{n=1}^{N} g_n(t) b_n(f) \tag{5.1}$$

donde $g_{n,t}$ es la ganancia de la función base n en la trama t, y $b_n(f)$, $n = 1, ..., N$ son las funciones base. Desde otro punto de vista, la señal se modela como la suma de unas componentes con bases fijas y amplitudes variables en el tiempo. Cuando se trata con instrumentos armónicos, cada función base se puede asociar a un *pitch*, y sus ganancias correspondientes contienen la información a cerca de su comportamiento en las fases de *onset*, sostenimiento y *offset*, de las notas con el *pitch* correspondiente.

En la bibliografía se pueden encontrar algunas propuestas que usan este tipo de modelos en aplicaciones como la separación de fuentes [Klapuri10b][Jaiswal11], la extracción de melodía [Durrieu10], transcripción musical [Vincent10, Carabias11] y el reconocimiento de fuentes sonoras [Heittola09].

Aunque existen varios algoritmos de descomposición de señal, en la actualidad NMF [Lee99] es el algoritmo de descomposición mas utilizado para aplicaciones que requieren la descomposición de señales del audio. En este esquema de descomposición NMF, las ganancias y las bases se restringen para que sean siempre no negativas, a la vez que se minimiza la distorsión entre el espectro de la señal de entrada y la reconstruida con los parámetros del modelo.

En las aplicaciones de separación de fuentes es conveniente realizar un aprendizaje previo de las funciones base que se van a emplear, sobre todo cuando la señal de entrada es monoaural. En la fase de descomposición las bases espectrales tienen un buen comportamiento cuando la escena

musical (construcción de los instrumentos, interpretación del músico, características de la sala, etc.) no difieren mucho entre el entrenamiento y la señal para descomponer. Algunas propuestas incluyen bases adaptativas [Smaragdis03, Virtanen07b, Vincent10, Carabias11], pero los resultados que obtienen no son fiables por la cantidad de parámetros libres que se le presentan al algoritmo NMF a la hora de realizar la descomposición.

5.2. Modelos NMF

5.2.1. Modelo armónico básico

Cuando se trata con instrumentos tonales, las notas musicales (excluyendo los transitorios) se pueden considerar cuasi periódicas, y su espectro se compone de picos espectrales regularmente espaciados. Entonces, las funciones base $b_{n,j}(f)$ presentan esta forma armónica, de manera que se les impone dicha restricción armónica y resulta:

$$b_{n,j}(f) = \sum_{m=1}^{M} a_{n,m,j} G(f - m f_0(n)) \tag{5.2}$$

donde $m = 1, ..., M$ es el número de parciales armónicos considerados, $a_{n,m,j}$ es la amplitud del parcial m-ésimo del *pitch* n y el instrumento j, $f_0(n)$ es la frecuencia fundamental para el *pitch* n, $G(f)$ es la transformada de la ventana de análisis y el espectro de un parcial armónico en la frecuencia $mf_0(n)$ se aproxima por $G(f - mf_0(n))$.

Esta restricción armónica hace que el modelo represente mejor la señal por la clara correspondencia entre cada función base y el *pitch* por medio de la frecuencia fundamental $f_0(n)$. Esta mejora, frente al modelo básico NMF de la ecuación 5.1, se demuestra en [Vincent10] con la aplicación de transcripción musical, y se justifica por el hecho de que, en el modelo básico, es necesario estimar la frecuencia fundamental después de haber estimado los parámetros del modelo, mientras en el modelo de la ecuación 5.2 las funciones base y la frecuencia fundamental están relacionados *a priori*.

En modelo armónico básico (BHC) de la ecuación 5.2, la representación completa de la señal $X(f, t)$ para cada trama t se describe como:

$$\hat{X}(f,t) = \sum_{j=1}^{J} \sum_{n=1}^{N} g_{n,t,j} \sum_{m=1}^{M} a_{n,m,j} G(f - m f_0(n)) \tag{5.3}$$

donde J es el número de instrumentos. Los parámetros libres, que deben ser estimados por el algoritmo NMF, son las ganancias $g_{n,t}$ y las amplitudes de los parciales $a_{n,m,j}$.

where J is the number of instruments. The free parameters of the BHC model are the time gains $g_{n,t}$ and the pitch amplitudes $a_{n,m,j}$.

5.2.2. Modelo filtro-fuente con excitación armónica plana

En el modelo que se describe en la ecuación 5.2, cada *pitch* de cada instrumento se representa con una función base diferente e independiente de las demás. Esto implica tener que estimar una gran cantidad de parámetros que hacen que sea compleja la obtención de un modelo fiable para el instrumento. Virtanen y Klapuri [Virtanen06] proponen relacionar todos los *pitches* modelando cada base como el producto de una excitación armónica $e_n(f)$ y un filtro-fuente $h_j(f)$. Cada función base se indexa con n y el filtro con j, de manera que:

$$b_{n,j}(f) = h_j(f) e_n(f), \quad n = 1, ..., N, j = 1, ..., J, \tag{5.4}$$

donde N es el número de excitaciones y J es el número de filtros-fuente. Es común que cada instrumento se represente con un único filtro cuya función se interprete como la representación del cuerpo de resonancia del instrumento.

Este modelo de filtro-fuente con una excitación por pitch tiene su origen en el procesado de señal de voz y síntesis de audio. En el procesado de la señal de voz, las excitaciones modelan el sonido producido por las cuerdas vocales, y el filtro representa el tracto vocal y su efecto de resonancia[RabinerBook78]. La propuesta del filtro-fuente se ha aplicado a otras tareas relacionadas con el procesado de audio, como la extracción de melodía en [Durrieu10]. En la síntesis de sonido el filtro-fuente se utiliza para colorear una señal rica en componentes espectrales y obtener el sonido deseado en [Valimaki06]. Sin embargo, la producción de sonidos por parte

de instrumentos musicales es, en general, más compleja que la propuesta del filtro-fuente, que no es capaz de modelar completamente dichos sonidos.

Además, este tipo de filtro-fuente tiene un problema asociado: las excitaciones dependen del *pitch*, lo cual no resuelve el problema de la gran cantidad de parámetros que se deben estimar. El siguiente paso consiste en restringir el modelo de la ecuación 5.4 de manera que las excitaciones sólo impongan armonicidad. En algunos trabajos [Badeau09, Heittola09, Klapuri10b] se introduce el uso de excitaciones $e_n(f)$ compuestas por elementos unitarios situados en posiciones múltiplos de la frecuencia fundamental del *pitch* n. Esta propuesta lleva a modelar las bases con unos excitaciones con forma de peine de manera que:

$$b_{n,j}(f) = \sum_{m=1}^{M} h_j(mf_0(n))G(f - mf_0(n)) \tag{5.5}$$

donde $h_j(mf_0(n))$ es el muestreo del filtro-fuente $h_j(f)$ en la frecuencia $mf_0(n)$.

Cuando se usa el modelo de filtro-fuente con excitaciones armónicas unitarias *(source-filter model with Harmonic-Comb Excitation (HCE)* para la definición de las funciones base, la representación completa de la señal se describe de la siguiente manera:

$$\hat{X}(f,t) = \sum_{j=1}^{J}\sum_{n=1}^{N} g_{n,t,j} \sum_{m=1}^{M} h_j(mf_0(n))G(f - mf_0(n)) \tag{5.6}$$

Los parámetros libres del modelo, que debe estimar el algoritmo de factorización son las ganancias $g_{n,t,j}$ y el filtro-fuente $h_j(f)$ para cada instrumento j.

5.2.3. Modelo filtro fuente con excitación multiple

La excitación armónica plana es capaz de representar espectros de notas con una distribución suave. Sin embargo, algunos instrumentos no presentan estas características suaves en toda la frecuencia.

Una alternativa interesante es el uso de las excitaciones múltiples propuestas en [Carabias11]. Este modelo es una extensión del modelo de filtro-fuente en el que las excitaciones se componen de la suma ponderada de

un conjunto de excitaciones básicas para cada instrumento. Siguiendo este modelo, el espectro de la función base para una nota se genera mediante la excitación armónica múltiple de dicha nota, multiplicada por el filtro-fuente correspondiente. En este esquema, la excitación $e_{n,j}(f)$ es diferente para cada *pitch* y cada instrumento, y además tiene naturaleza armónica. Para cada instrumento, la excitación en cada *pitch* se obtienen mediante la suma ponderada de las excitaciones base, las cuales son únicas para cada instrumento, mientras que los pesos que las modulan varían en función del *pitch*

Siguiendo este modelo, la excitación para cada *pitch* se obtiene como:

$$e_{n,j}(f) = \sum_{m=1}^{M} \sum_{i=1}^{I} w_{i,n,j} v_{i,m,j} G\left(f - m f_0(n)\right) \tag{5.7}$$

donde $v_{i,m,j}$ es la i-ésima excitación base para el instrumento j (compuesta por M parciales), y $w_{i,n,j}$ son los pesos de la i-ésima excitación para el *pitch* n y el instrumento j. Finalmente, el modelo filtro-fuente con excitación múltiple *(Source-filter Model with Multi-Excitation per Instrument (MEI)* representa el espectro de amplitud de la señal $X(f,t)$ de manera que:

$$\hat{X}(f,t) = \sum_{n,j} g_{n,t,j} h_j(f) \sum_{m=1}^{M} \sum_{i=1}^{I} w_{i,n,j} v_{i,m,j} G\left(f - m f_0(n)\right) \tag{5.8}$$

donde $n = 1, ..., N$ (N es el número total de *pitches*) y $j = 1, ..., J$ (J es el número total de instrumentos), M representa el número de parciales armónicos considerados e I el número de excitaciones por instrumento, siendo $I << N$. El uso de un número pequeño de excitaciones base I reduce significativamente el número de parámetros del modelo, lo cual hace que el proceso de aprendizaje sea más efectivo. Los parámetros libres del modelo, que deben ser estimados por el algoritmo NMF son: las ganancias temporales $g_{n,t,j}$, el filtro de instrumento $h_j(f)$, las excitaciones base $v_{i,m,j}$ y los pesos de la mezcla de las excitaciones $w_{i,n,j}$.

El uso de un sólo filtro para cada instrumento supone la independencia entre el filtro y las excitaciones. Para instrumentos como la flauta, el filtro (el tubo) varía en función de la nota que se está tocando (la longitud del tubo).

En consecuencia, la reducción del número de parámetros libres (un filtro por instrument y unas pocas excitaciones base) supone una limitación en el potencial del modelo MEI, a cambio de poder tener una buena estimación por la reducción del número de parámetros.

5.2.4. NMF ampliado para la estimación de parámetros del modelo de señal

La restricción de no negatividad de los parámetros del modelo NMF ha mostrado ser una restricción eficiente en el aprendizaje de espectrogramas de audio [Smaragdis03, Virtanen06, Vincent10]. Por tanto, se aplica la restricción de no negatividad a todos los parámetros del modelo descrito en la ecuación 5.8. Bajo estas condiciones, se estiman los parámetros del modelo mediante la minimización del error de reconstrucción del espectrograma estimado de la señal $\hat{X}(f,t)$ y el espectrograma de la señal de entrada $X(f,t)$.

Para la obtención de los parámetros del modelo que minimizan la distorsión de Kullback-Leibler (KL), en [Lee01] se propone un algoritmo iterativo denominado *Multiplicative Update Rules (MU)*. Se ha demostrado que con este algoritmo la divergencia es decreciente con las iteraciones y además asegura la no negatividad de los parámetros por su naturaleza multiplicativa. Estas reglas de actualización de los parámetros se obtienen mediante la el escalado diagonal del paso del algoritmo del gradiente (ver [Lee01] para más detalles). Si se expresan las derivadas parciales de la función de coste como la diferencia entre los términos positivos y negativos $\nabla_{\theta_l} D = \nabla^+_{\theta_l} D - \nabla^-_{\theta_l} D$, la ecuación de actualización multiplicativa para cada parámetro escalar θ_l viene dada por el cociente de dos términos positivos, de la siguiente manera:

$$\theta_l \leftarrow \theta_l \frac{\nabla^-_{\theta_l} D(X(f,t)|\hat{X}(f,t))}{\nabla^+_{\theta_l} D(X(f,t)|\hat{X}(f,t))} \tag{5.9}$$

La principal ventaja de este tipo de ecuaciones de actualización es que aseguran la no negatividad de los parámetros en un algoritmo NMF ampliado.

El esquema propuesto en [Lee01] se puede usar para cada uno de los modelos anteriormente descritos: BHC, HCE y MEI. A continuación se presentan las ecuaciones de actualización derivadas para el modelo MEI. Para

no extender este apartado no se presentan las de los otros modelos, que se derivarían igualmente de la ecuación 5.9 para los parámetros correspondientes

$$g_{n,t,j} \leftarrow g_{n,t,j} \frac{\sum_{i,f,m} r_t(f) h_j(f) w_{i,n,j} v_{i,m,j} G(f - m f_0(n))}{\sum_{i,f,m} h_j(f) w_{i,n,j} v_{i,m,j} G(f - m f_0(n))} \tag{5.10}$$

$$h_j(f) \leftarrow h_j(f) \frac{\sum_{t,m,n} r_t(f) g_{n,t,j} w_{i,n,j} v_{i,m,j} G(f - m f_0(n))}{\sum_{t,m,n} g_{n,t,j} w_{i,n,j} v_{i,m,j} G(f - m f_0(n))} \tag{5.11}$$

$$v_{i,m,j} \leftarrow v_{i,m,j} \frac{\sum_{f,t,n} r_t(f) g_{n,t,j} h_j(f) w_{i,n,j} G(f - m f_0(n))}{\sum_{f,t,n} g_{n,t,j} h_j(f) w_{i,n,j} G(f - m f_0(n))} \tag{5.12}$$

$$w_{i,n,j} \leftarrow w_{i,n,j} \frac{\sum_{f,t,m} r_t(f) g_{n,t,j} h_j(f) v_{i,m,j} G(f - m f_0(n))}{\sum_{f,t,m} g_{n,t,j} h_j(f) v_{i,m,j} G(f - m f_0(n)} \tag{5.13}$$

donde $r_t(f) = \frac{X(f,t)}{\hat{X}(f,t)}$.

Se debe aclarar que todos los modelos descritos anteriormente se han usado en la bibliografía para otras aplicaciones (por ejemplo, transcripción musical) pero no han sido aplicados para la separación de señales monoaurales. Estos modelos tienen en común la restricción de armonicidad. El modelo BHC describe cada *pitch* de manera independiente entre ellos. Por el contrario, HCE y MEI usan el filtro-fuente para crear la dependencia entre todos los *pitches* de cada instrumento. El modelo MEI con $I = 1$ difiere del modelo HCE en que la excitación resultante no es plana (componentes unitarias del modelo HCE) y se modela con la excitación base $v_{1,m,j}$. En el apartado de experimentos de este capítulo se evalúa el rendimiento de cada uno de estos modelos para la separación de fuentes musicales.

5.3. Aplicación la separación de fuentes musicales.

La separación de fuentes es una aplicación que se puede llevar a cabo incluso sin información previa [Virtanen07b]. La propuesta de este capítulo

se basa en el entrenamiento de modelos NMF a lo largo de todo el rango de notas de cada instrumento. En una fase posterior de factorización, los parámetros del modelo quedan fijos, excepto las ganancias $g_{n,t,j}$, que son estimadas por el algoritmo para las funciones base entrenadas. La fase final de separación estima las fuentes con la ayuda de los espectrogramas de amplitud estimados por el algoritmo NMF $\hat{X}(f,t)$.

En este capítulo se realiza una separación de fuentes basada en descomposición NMF con tres modelos diferentes: 1) el modelo armónico básico (BHC) de la ecuación 5.3; 2) el modelo de fltro-fuente con excitación plana (HCE) de la ecuación 5.6; y 3) el modelo de filtro-fuente con excitación múltiple (MEI) de la ecuación 5.8.

Para la evaluación del rendimiento de cada modelo, se propone realizar la separación de fuentes con una base datos de cinco instrumentos y comparar los resultados que obtiene cada uno de los modelos descritos.

5.3.1. Configuración de los experimentos

Datos de entrenamiento

Para la fase de entrenamiento, se toma el rango dinámico completo de cada instrumento de la base de datos de instrumentos de RWC [Goto02]. Se han considerado los cinco instrumentos que estarán presentes en la fase de factorización (clarinete, flauta, oboe, trompa y fagot). Para cada instrumento, se dispone de las grabaciones individuales con un intervalo de un semi tono entre cada nota. De todas las opciones disponibles en la base de datos, se han seleccionado las grabaciones con un estilo interpretativo normal y nivel dinámico *mezzo*.

Datos de evaluación

Para evaluar el rendimiento de cada modelo descrito en este capítulo, en la aplicación de separación de fuentes, se ha usado la base de datos de instrumentos de viento madera descrita en el apartado 4.1.6 de esta tesis [MIREX2007]. Se han usado todos los instrumentos (fagot, clarinete, flauta, trompa y oboe). Se han usado los 30 primeros segundos de cada grabación, por lo que se obtienen un total de 26 mezclas polifónicas con niveles de

polifonía desde 2 hasta 5.

Representación tiempo/frecuencia de las señales.

Los parámetros de los modelos se entrenan con los datos de entrenamiento. Por tanto, dichos datos deben estar etiquetados con el instrumento j y el pitch n que está activo en cada trama t. En el sistema de separación propuesto en este capítulo se usa la resolución de un semi tono, al igual que en [Carabias11].

Para esta representación tiempo/frecuencia, la implementación más directa es la integración de los bins de la STFT correspondientes al mismo semitono. A pesar de ello, esta representación deja fuera de los rangos adoptados parte de la energía de las notas, que sale fuera del ámbito de cada semitono. Como consecuencia de ello, los modelos armónicos NMF pueden encontrar energía de una nota fuera del rango MIDI (y sus múltiplos) que le corresponden a cada *pitch*, lo cual limita la capacidad de separación del modelo.

El tamaño de trama y salto son 128 ms y 10 ms respectivamente. Más detalles sobre la representación tiempo/frecuenca usada se pueden ver en [Carabias11].

Proceso de separación basado en NMF

Con todos los modelos de descomposición, basados en NMF, se ha realizado la separación de fuentes de la siguiente manera:

- Se calcula la representación tiempo/frecuencia de la señal mezclada $X(f,t)$
- Se estima la factorización $\hat{X}(f,t) \approx \sum_n g_{n,t,j} b_{n,j}(f)$. Las funciones base quedan fijas para todos los modelos, mientras que las ganancias temporales se estiman usando el algoritmo de minimización NMF para cada señal mezclada.
- La factorización se puede particularizar para cada instrumento j, obteniendo $\hat{X}_j(f,t)$. El sistema cuenta con información de los instrumentos que están presentes en la señal. Otros sistemas, como en [Heittola09]

tienen que implementar una fase de clasificación de instrumentos antes de la separación.

- Las máscaras de separación se obtienen mediante el cálculo de la proporción de energía en cada celda tiempo/frecuencia (trama t y nota MIDI f) para el instrumento j. El uso de este tipo de máscaras de Wiener es común en la bibliografía de separación de fuentes [Every06]

- Una vez que se obtiene la máscara de separación para cada instrumento, el espectrograma de la señal de entrada se filtra con ellas de manera multiplicativa. En este sistema, por el tipo de representación tiempo/frecuencia usada, todos los bins correspondientes a un mismo intervalo MIDI, se tratan de la misma manera. El espectrograma de las señales se calcula con 8192 bien frecuenciales.

Comparación con el estado del arte

Para tener una comparación justa con el estado del arte en la descomposición de señales, se ha realizado la separación con la herramienta *Flexible Audio Source Separation Toolbox (FASST)* [Ozerov11]. Dada la posibilidad de configuración de esta herramienta, se han establecido descomposiciones NMF con $K = 114$ funciones base, que coincide con el rango de notas de todos los instrumentos considerados, y se ha configurado para tratar señales monoaurales. La separación de fuentes se ha realizado en dos fases: 1) En primer lugar se realiza una fase de entrenamiento. Para ello, se realiza la descomposición de cada señal del conjunto de datos de entrenamiento. Las activaciones del modelo se inicializan a cero para todas las bases k que no están activas en cada trama. Las funciones base obtenidas se almacenan para usarse en la segunda fase. 2) A continuación se implementa la fase de factorización. Se realiza la descomposición para cada pieza del conjunto de datos de evaluación (con niveles de polifonía desde 2 hasta 5) usando las funciones base entrenadas y dejándolas fijas durante este proceso. De esta manera se obtiene una estimación de las activaciones de cada función base, la cual es utilizada para la separación de la señal. Esta propuesta de dos fases (entrenamiento y factorización) se ha adaptado para poder tener una comparación con datos obtenidos en las mismas condiciones que la propues-

ta de este capítulo. En FASST la frecuencia de muestreo es de 44,100Hz, la transformada usada es la *Short-Time Fourier Transform (STFT)*, con una trama de análisis de 5,644 muestras y un santo del 50 % de la trama.

Evaluación

Para realizar una evaluación objetiva del rendimiento de la separación se usan las medidas implementadas en [Vincent07]. Suponiendo que la señal separada para cada instrumento se puede descomponer en:

$$\hat{x}_j(t) = x_j(t) + e_{interf}(t) + e_{artif}(t) \tag{5.14}$$

donde $\hat{x}_j(t)$ es la señal separada para el instrumento j, $x_j(t)$ es la señal original aislada del instrumento j, $e_{interf}(t)$ es el término de error asociado a la interferencia de las otras fuentes y $e_{artif}(t)$ es el término de error asociado a los artefactos generados por el procedimiento de separación. Las medidas para cada señal separada son *Source to Distortion Ratio* (SDR), *Source to Interference Ratio* (SIR), y *Source to Artifacts Ratio* (SAR) [Vincent07].

En el esquema NMF los parámetros que se van a estimar se inicializan de manera aleatoria. Esto provoca que, en la fase de factorización, los espectrogramas obtenidos difieran en cada ejecución, dando, por tanto, diferentes resultados de separación. Se han realizado 30 ejecuciones para cada algoritmo para demostrar la relevancia estadística de las medidas. El intervalo de confianza del 95 % es menor de 0,7dB para todos los algoritmos, lo que indica que las diferencias entre todos ellos son estadísticamente relevantes.

5.3.2. Resultados

Tabla 5.1: Medidas objetivas para la Separación de fuentes en señales con distintos instrumentos (dB)

Algoritmo / Polifonía		2			3			4			5		
(all values in dB)		SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
BHC		8.8	10.9	11.0	6.8	8.6	9.4	5.2	5.9	8.3	3.8	2.4	7.5
HCE		5.5	8.3	11.0	2.9	4.8	9.6	1.7	2.8	8.7	0.9	-8.2	8.2
MEI	I=1	9.0	10.9	11.4	6.5	8.3	9.8	4.2	5.7	8.9	0.9	1.3	7.9
	I=2	9.1	11.1	11.4	6.9	8.9	9.8	5.0	6.7	8.6	2.8	4.1	8.0
	I=4	9.4	11.3	11.5	7.6	9.4	10.0	6.3	7.8	9.0	5.2	6.2	8.2
FASST		2.2	4.3	6.8	0.3	0.7	4.2	-0.8	-2.1	3.0	-1.3	-4.3	2.4
Wiener ideal		12.4	14.0	12.8	11.4	13.2	11.9	10.8	12.7	11.3	10.3	12.4	10.8

En la tabla 5.1 se presentan los resultados para los modelos BHC, HCE y MEI. Se han comparado con la separación ideal con máscaras de Wiener y la herramienta FASST. Las máscaras ideales de Wiener se han obtenido con las señales originales. Esta es una situación poco factible en la realidad, pero se ha realizado para tener una referencia del máximo redimiendo del procedimiento de separación y sus limitaciones. La separación está limitada por la resolución tiempo/frecuencia empleada.

En relación a los resultados de separación obtenidos con la herramienta FASST, se puede ver que sólo el modelo HCE tiene un rendimiento similar a FASST, ocurriendo esto para niveles altos de polifonía. Las funciones base que se han entrenado con FASST no están controladas ni relacionadas mediante un filto-fuente como en el modelo BHC, pero no cuentan con la restricción armónica del modelo BHC. La principal diferencia entre la separación con FASST y el procedimiento basado en NMF propuesto se debe a la resolución tiempo/frecuencia empleada. La resolución lineal usada en la herramienta FASST no es apropiada para la separación de señales de audio, como puede verse en los resultados. En esas condiciones, las pequeñas variaciones en la frecuencia fundamental de cualquier nota (menores de un cuarto de tono) pueden producir variaciones mayores que el ancho de la ventana de análisis en altas frecuencias (parciales altos). En consecuencia, la capacidad de separación están limitadas porque las funciones base entrenadas no modelan bien las diferentes envolventes espectrales que se general de cada nota notas (por las pequeñas variaciones de la frecuencia fundamental). Otras propuestas de separación de fuentes [Vincent06] implementan, también, la resolución logarítmica para la representación de las señales, tal y como se hace en la propuesta de este capítulo. Además de todo ello, se ha observado que la restricción armónica es bastante beneficiosa para la obtención de buenos resultados de separación por la reducción de interferencia entre las fuentes (ver valores de SIR en la tabla 5.1).

El modelo de excitación múltiple (MEI) para $I = 4$ excitaciones obtiene los mejores resultados. De hecho, los resultados mejoran escalonadamente cuando se incrementa desde $I = 1$ hasta $I = 4$. Este comportamiento se justifica por el incremento en la capacidad de modelado de las fuentes que dan las múltiples excitaciones. El modelo HCE, que comparte con MEI el paradigma del filtro-fuente, obtiene peores resultados para todas las me-

didas. Esto quiere decir que el modelo HCE no tiene suficiente grado de libertad para adaptar sus parámetros a los espectros de las notas.

El caso del modelo BHC es particularmente interesare. Este modelo es capaz de modelar cada nota de cada instrumento de manera independiente. Esto implica que es el modelo con mayor grado de libertad, por no estar limitado por el filtro-fuente. Sin embargo, sus resultados son comparables con el modelo MEI para $I = 1, 2$, pero no alcanza los resultados del modelo MEI cuando $I = 4$. Las posibles diferencias entre los instrumentos de las bases de datos de entrenamiento y factorización pueden explicar este comportamiento, puesto que variaciones en la escena musical producen variaciones en el espectro de las notas. Visto de otra manera, el modelo BHC se entrena específicamente para la interpretación de cada nota, mientras que el modelo MEI cuenta con una representación más flexible gracias al uso del filtro-fuente. En cualquier caso el modelo BHC presenta un rendimiento competitivo para el restode propuestas, en especial cuando crece el nivel de polifonía.

Para conluir, hay que decir que el modelo MEI propuesto en [Carabias11] consigue los mejores resultados principalmente por las siguientes razones: 1) El uso de un conjunto de excitaciones proporcionan unas capacidades de modelado muy elevados, sobre todo si se compara con el modelo HCE; 2) este método se basa el el modelo filtro-fuente, que proporciona una gran tolerancia cuando las condiciones de la escena musical varían, en contraposición a lo que ocurre con el modelo BHC.

5.4. Conclusiones

En este capítulo se ha demostrado la viabilidad de la separación de fuentes de varios instrumentos con el uso de NMF con modelos de instrumento. Estos modelos definen ciertos parámetros del instrumento j, lo que permite estimar el espectro de amplitud independiente $\hat{X}_j(f,t)$ para cada instrumento j. La separación con máscaras de Wiener se puede definir mediante la asignación de una parte proporcional de la amplitud de cada celda tiempo/frecuencia a cada instrumento, en relación a la energía total de todos los instrumentos en esa celda. Las medidas de la separación obtenida indican que el modelo de excitación múltiple MEI, propuesto en [Carabias11], obtie-

ne los mejores resultados de separación entre todos los modelos comparados, usando un número bajo de excitaciones.

Como conclusiones de mejora, se puede decir que: 1) la resolución en frecuencia es muy limitada, y es conveniente ampliarla. A pesar de que los modelos de instrumento están anotados en la resolución de un semitono, es posible interpolarlos para usarlos con mayor resolución; y 2) En los modelos con filtro-fuente, los parámetros de las funciones base se podrían actualizar en la fase de factorización para adaptarlos a la nueva escena musical. Si se realiza una actualización controlada, los modelos actualizados pueden obtener mejores resultados en separación.

Capítulo 6

NMF-*Sparse Coding* con restricción monofónica y modelos de instrumento aplicado a SSS y AMT

En este capítulo se propone un modelo de descomposición de señal con restricción de monofonía y se aplica a señales polifónicas compuestas por fuentes de distintos instrumentos monofónicos. La restricción monofónica es particularmente efectiva para los instrumentos tonales, puesto que cada nota está únicamente relacionada a una base de descomposición. La restricción se implementa imponiendo la activación de una única en cada instante temporal durante el proceso de descomposición. El método propuesto utiliza unos modelos de instrumento armónicos que son entrenados previamente mediante un proceso supervisado. Ambas restricciónes (armonicidad de las bases y monofonía de las fuentes) se implementan de manera determinista. Se ha evaluado el sistema propuesto para dos aplicaciones relacionadas con el análisis de las señales musicales, separación de fuentes (SSS) y transcripción musical (AMT). Así mismo se han comparado los resultados obtenidos con otros sistemas del estado del arte, usando un conjunto de señales polifónicas que contienen interpretaciones de instrumentos monofónicos.

6.1. Introducción

La separación de fuentes de audio (SSS) y la transcripción musical automática (AMT) son dos tareas distintas de procesado de señal musical, pero comparten ciertos procesos, como puede ser el de descomposición de la señal, en ciertas implementaciones. De hecho, algunos autores ven la AMT como un proceso previo a la SSS [Jaiswal11], mientras que otros piensan que la SSS se puede considerar un proceso previo a la AMT [Gainza07].

Por un lado, SSS se puede aplicar a muchas señales disponibles en día a día, las cuales se componen de la mezcla de los sonidos de varios instrumentos. La separación de fuentes es el proceso por el cual una señal mezclada de audio se descompone en las distintas fuentes individuales que la forman. Los escenarios para la separación de fuentes se puede clasificar, según el número de sensores (micrófonos) y fuentes que existan. Los escenarios sobredeterminados, son aquellos en los que existen más sensores que fuentes[Hyvarinen00, Smaragdis98, Zibulevsky01, Reyes03]. Los escenarios determinados, son aquellos que cuentan con el mismo número de fuentes que de sensores. Y, por último, los infradeterminados, son los casos en los que el número de fuentes es mayor que el de fuentes. En este capítulo se desarrolla un sistema enmarcado en el último de ellos, se lleva a cabo la separación de fuentes con un único sensor [Namgook09, Sawada11], lo que supone ser el escenario más complejo de los posibles.

Por otro lado, AMT es el proceso por el cual se genera una información simbólica temporal de las notas que se tocan en una grabación musical. La transcripción musical de señales polifónicas es una tarea compleja, principalmente por el solapamiento en frecuencia de algunos parciales de las notas tocadas. La transcripción de instrumentos tonales, [KlapuriPhD04], es un tipo de transcripción muy extendida. En ella se estiman los tiempos de inicio y final, así como los *pitches* de cada una de las notas. Aún así, los sistemas de transcripción no suelen proporcionar la transcripción individual de cada instrumento que está presente en la interpretación, diferenciando entre cada uno de ellos.

En este capítulo se describe un método de descomposición de señal, que se puede aplicar tanto a SSS como a AMT de la misma señal polifónica con instrumentos monofónicos. Otros métodos similares se han usado, igual-

mente, para las mismas aplicaciones de audio (SSS y AMT) con resultados relevantes [Reyes03, Helen05, Wang05].

Estos métodos pretenden descomponer el espectrograma de la señal de audio en una combinación lineal de ciertas funciones base espectrales. El espectro de magnitud de la señal $X(f,t)$ en la trama t y en la frecuencia f se modela como la suma ponderada de las funciones base, tal que:

$$\hat{X}(f,t) = \sum_{n=1}^{N} g_n(t) b_n(f) \tag{6.1}$$

donde $g_n(t)$ es la ganancia de la función base n en la trama t, y $b_n(f)$, $n = 1, ..., N$ son las N funciones base. Cuando se trata con sonidos armónicos, en el ámbito de la AMT, es conveniente que cada función base represente idealmente a un *pitch*, de esta manera sus ganancias correspondientes contendrán de manera directa la información de *onset* y *offset* de la nota correspondiente a dicho *pitch*.

El proceso de aprendizaje de las funciones base puede ser supervisado o no supervisado, dependiendo de la información a priori que se tenga de la composición (como los instrumentos que se están tocando). En el caso supervisado, las funciones base se pueden dejar fijas, o adaptarlas a la escena musical real de la composición analizada. En el sistema propuesto en este capítulo, se usa un proceso de entrenamiento supervisado [Carabias11], que configura unas funciones base fijas para el proceso de descomposición.

Existen algunos métodos, en el estado del arte, para la descomposición de señales de audio, como son *Independent Component Analysis (ICA)* [Plumbley03], *Non-Negative Matrix Factorization* [Lee99], y *Sparse Coding* [Abdallah04]. Pueden verse más detalles en el Capítulo 3 de esta tesis.

Como se muestra en el Capítulo 3, la restricción de dispersión se puede aplicar sobre el proceso de descomposición de señales. En las aplicaciones de SSS y AMT, esta restricción ha recibido una importante atención[Abdallah04, Abdallah06, Virtanen07b, Ozerov10]. *Sparse Coding* trata de realizar una descomposición espectral dispersa con ganancias que presentan funciones de densidad de probabilidad centradas en cero y con largas colas [Hoyer04], de manera que la mayor parte de la energía se asocia a unas pocas funciones base con ganancias positivas. Esta suposición encaja con el concepto por el cual, en la música occidental, sólo unas pocas notas suenan activas al

mismo tiempo. Si se trabaja con espectrogramas de potencia o amplitud, se pueden combinar NMF y *Sparse Coding* en algoritmos *Non-Negative Sparse Coding (NNSC* [Abdallah06, Hoyer04] para la descomposición de señales.

El método propuesto en este capítulo, usando restricción de monofonía para cada instrumento, impone la dispersión de tal manera que sólo una ganancia correspondiente a cada instrumento puede estar activa en cada instante temporal. Esta dispersión extrema ha sido usada en otras propuestas para descomposición de señal en el estado del arte. Por ejemplo, en un esquema probabilístico, esta restricción se introduce en un *Gaussian Scaled Mixture Model (GSMM)* [Benaroya06], o en un *Factorial Scaled Hidden Markov Model (FS-HMM)* [Ozerov09], con suposiciones de Gaussianidad y divergencia Itakura Saito (IS).

En este capítulo, se propone un método de factorización determinista aplicado al procesado de señales monoaurales polifónicas compuestas por la mezcla de distintos instrumentos monofónicos. Algunas propuestas de la bibliografía usan este tipo de señales, como GSMM [Benaroya06] y FS-HMM [Ozerov09], pero en un esquema probabilístico. El método propuesto presenta la novedad de implementar las restricciones de monofonía y armonicidad de manera determinista. Se supone que cada instrumento presente en la señal es monofónico, es decir, solo una posible nota por cada instrumento estará activa en cada momento. Este tipo de señales son muy típicas con instrumentos de viento y algunos casos de cuerda frotada. Ciertas composiciones corales se han compuesto para este tipo de instrumentación (por ejemplo, las corales de Bach, que se emplean en la evaluación del sistema propuesto). Hasta donde se conoce, dentro de la bibliografía, ningún otro método hasta ahora proporciona una transcripción de una señal de estas características para cada instrumento por separado. Al final del capítulo, el método propuesto, y sus variantes, se evalúan y comparan con otros métodos del estado del arte, proporcionando unos resultados prometedores.

El resto del capítulo se estructura de la siguiente manera: en el apartado 6.2 se revisan los modelos de señal con restricciones armónicas y de dispersión propuestos en estudios previos, así como el esquema teórico de NMF y modelado de instrumentos que se empleará; en el apartado 6.3 se describe el método propuesto para restringir el modelo de descomposición de manera que sólo tenga una base activa para cada instrumento en cada

instante; en el apartado 6.4 se aplica el modelo propuesto para SSS y AMT con señales polifónicas conformadas por fuentes monofónicas, los resultados se comparan con otros métodos del estado del arte al final de dicha sección; finalmente, se proponen lineas de futuro y conclusiones en el apartado 6.5

The paper is structured as follows: Section 2 reviews the harmonic and sparsity constrained signal models from previous studies, as well as theoretical background on NMF and instrument modelling; Section 3 explains the proposed method for constraining a polyphonic signal model to have a single non-zero gain per instrument at each frame and provides the algorithm for signal spectral decomposition; the proposed approach is applied in Section 4 for SSS and AMT using polyphonic mixtures composed of several monophonic single-instrument sources, the results are compared with those obtained by other state-of-the-art methods; finally, we draw some conclusions and discuss future work in Section 5.

6.2. Base teórica

6.2.1. Modelo armónico básico *Basic Harmonic Constrained (BHC)*

Las notas musicales (a excepción de los transitorios) en los instrumentos tonales son cuasi-periódicas, con un espectro que presenta picos espectrales regularmente espaciados [Carabias11]. De hecho, en la bibliografía se puede ver, que la mayoría de los modelos de instrumento cuentan con esta restricción de armonicidad [Carabias11, Bertin10, Vincent10, Raczynski07]. Si esta peculiaridad del espectro de las notas se traduce en la restricción armónica sobre las funciones base, entonces, se puede asociar una función base a cada pitch n por medio de la frecuencia fundamental correspondiente f_0. La restricción se aplica sobre el modelo de la ecuación (6.1), de manera que:

$$b_{n,j}(f) = \sum_{m=1}^{M} a_{n,j}[m]G(f - mf_0(n)) \tag{6.2}$$

donde $b_{n,j}(f)$ son las funciones base para cada nota n y el instrumento j, m es el número de parcial seleccionado, M es el número total de parciales,

$a_{n,j}[m]$ representa la amplitud del parcial m para la nota n y el instrumento j, $G(f)$ es una réplica de la transformada de la ventana de análisis, y el espectro de una componente armónica unitaria en la frecuencia $mf_0(n)$ se puede representar como $G(f - mf_0(n))$.

Si se incluye la restricción en la ecuación (6.1), el modelo BHC del espectro de amplitud de una señal musical quedaría de la siguiente manera:

$$\hat{X}(f,t) = \sum_{j=1}^{J} \sum_{n=1}^{N(j)} \sum_{m=1}^{M} g_{n,j}(t) a_{n,j}[m] G(f - mf_0(n)) \tag{6.3}$$

donde J es el número total de instrumentos y $N(j)$ es el número total de notas que puede tocar el instrumento j. Las ganancias $g_{n,j}(t)$ y las amplitudes para cada parcial $a_{n,j}[m]$ son los parámetros del modelo que se necesitan estimar. Habitualmente, estos parámetros se estiman mediante la minimización del error de reconstrucción entre el espetrograma de la señal de entrada $X(f,t)$ y el estimado $\hat{X}(f,t)$.

Las funciones de coste, para medir el error de reconstrucción, más conocidas son la distancia Euclídea (EUC), la divergencia Kullback-Leibner (KL) y la divergencia Itakura-Saito (IS). La β-divergencia (ecuación 6.4) es otra función de coste conocida, que además agrupa las tres anteriores en tres casos particulares de su definición: EUC ($\beta = 2$), KL ($\beta = 1$) y IS ($\beta = 0$).

$$D_\beta(x|\hat{x}) = \begin{cases} \frac{1}{\beta(\beta-1)} \left(x^\beta + (\beta-1)\hat{x}^\beta - \beta x \hat{x}^{\beta-1} \right) & \beta \in (0,1) \cup (1,2] \\ x \log \frac{x}{\hat{x}} - x + \hat{x} & \beta = 1 \\ \frac{x}{\hat{x}} + \log \frac{x}{\hat{x}} - 1 & \beta = 0 \end{cases} \tag{6.4}$$

Algunos sistemas de la bibliografía ya usan la β-divergencia como función de coste, puede verse en [Vincent10, Fevotte09b, Fevotte11b].

6.2.2. Modelo BHC con restricción de dispersión

La dispersión es una restricción natural, que se aplica a las ganancias y obliga al modelo a tener sólo unas pocos valores no nulos en las ganancias $g_{n,j}(t)$ de cada trama t. La suposición de dispersión concuerda con la naturaleza de la música, en la que sólo unas pocas notas, de todas las posibles, se tocarán simultáneamente [Abdallah04]. Existen otros

trabajos en los que se aplica esta restricción en el procesado de señal de audio [Carabias11, Abdallah04, Virtanen07b, Hoyer04, Gemmeke11].

Una manera común de introducir la restricción de dispersión en el modelo de señal es incluyendo un término de penalización en la función de coste [Gemmeke11] que se desea minimizar. Este término descarta las soluciones en las que muchos valores de las ganancias tomen valores no nulos. De esta manera, la función de cote quedaría como:

$$D(X(f,t)|\hat{X}(f,t)) = D_\beta(X(f,t)|\hat{X}(f,t)) + \lambda \sum_{f,t} \phi(g_{n,j}(t)) \tag{6.5}$$

donde D_β representa la distorsión definida en la ecuación (6.4), λ es un parámetro que controla el peso del término de penalización, y ϕ es la función que penaliza las ganancias no nulas. En la bibliografía se encuentran varias opciones para la definición del término de penalización. Por ejemplo, Olshausen y Field [Olshausen97], proponen $\phi(x) = -exp(-x^2)$, $\phi(x) = log(x^2 - 1)$ y $\phi(x) = |x|$, como posibles funciones de penalización. En la propuesta de este capítulo se ha usado la tercera de ellas para la fase experimental. Se ha demostrado que es menos sensible a las variaciones del parámetro λ [Virtanen07b] y es un medio efectivo para la búsqueda de soluciones dispersas [Candes08].

6.2.3. Modelos con restricción monofónica

Como se ha comentado, la dispersión es una restricción adecuada para señales musicales. Sin embargo, se puede matizar para el caso de señales polifónicas compuestas por instrumentos monofónicos (que sólo tocan una nota en cada instante), de manera que la dispersión sea tal que, sólo una ganancia de cada instrumento se pueda activar en cada trama t. Esta restricción de dispersión extrema se ha propuesto para métodos de descomposición de señal con esquemas probabilísticos. Por ejemplo, Benaroya *et al.* [Benaroya06] proponen un método para SSS en el que la STFT de cada fuente se modela con un *Gaussian Mixture Model (GMM)*; el GMM se modula con un parámetro de seguimiento de amplitud para cada trama, lo que resulta en un *Gaussian Scaled Mixture Model (GSMM)*, donde se supone que las fuentes son variables implícitamente monofónicas con muchos

posibles estados. Ozerov *et al.* [Ozerov09] proponen un método, llamado *Factorial Scaled Hidden Markov Model (FS-HMM)* que generaliza los modelos GSMM y NMF usando la divergencia Itakura-Saito e incorpora la continuidad temporal mediante el modelado de Markov.

6.2.4. NMF ampliado para la estimación de parámetros

La restricción de no negatividad de los valores de los parámetros ha mostrado ser eficiente para el aprendizaje y la factorización de espectrogramas de señales de audio[Virtanen06]. De hecho, esta restricción ha sido ampliamente utilizada en transcripción musical [Carabias11, Bertin10, Vincent10] y la separación de fuentes [Virtanen06, Ozerov11].

Cuando los valores de los parámetros son no negativos, como en el caso de los espectros de amplitud, una manera común para factorizarlos es la minimización del error de reconstrucción entre el espectrograma de la señal de entrada $X(f,t)$ y la estimación $\hat{X}(f,t)$

Para obtener los parámetros del modelo que minimicen la función de coste, Lee *et al.* [Lee01] proponen un algoritmo iterativo basado en reglas de actualización multiplicativas. Con estas reglas, $D_\beta(X(f,t)|\hat{X}(f,t))$ es una función decreciente en cada iteración, y además, se asegura la no negatividad de los valores de los parámetros. Las reglas de actualización propuestas, se obtienen aplicando un escalado diagonal del tamaño de paso del algoritmo decreciente del gradiente (ver más detalles en [Lee01]. La regla de actualización para cada parámetro escalar θ_l viene dada por las derivadas parciales de la función de coste $\nabla_{\theta_l} D_\beta$, como el cociente de dos términos positivos $\nabla^-_{\theta_l} D_\beta$ y $\nabla^+_{\theta_l} D_\beta$:

$$\theta_l \leftarrow \theta_l \frac{\nabla^-_{\theta_l} D_\beta(X(f,t)|\hat{X}(f,t))}{\nabla^+_{\theta_l} D_\beta(X(f,t)|\hat{X}(f,t))} \tag{6.6}$$

La principal ventaja del uso de la regla multiplicativa de la ecuación (6.6) es la no negatividad asegurada de los parámetros que se actualicen con ella. Para el modelo con restricción armónica de la ecuación (6.3), las reglas de actualización multiplicativas que minimizan la β-divergence son las siguientes [Fevotte11b]:

$$a_{n,j}[m] \leftarrow a_{n,j}[m] \frac{\sum_{f,t} X(f,t)\hat{X}(f,t)^{\beta-2} g_{n,j}(t) G(f - mf_0(n))}{\sum_{f,t} \hat{X}(f,t)^{\beta-1} g_{n,j}(t) G(f - mf_0(n))} \tag{6.7}$$

Además, cuando se usa el término de penalización de la ecuación (6.5) con $\phi(x) = |x|$, las ganancias se estiman siguiendo la siguiente regla de actualización [Gemmeke11]:

$$g_{n,j}(t) \leftarrow g_{n,j}(t) \frac{\sum_{f,m} X(f,t)\hat{X}(f,t)^{\beta-2} a_{n,j}[m] G(f - mf_0(n))}{\lambda + \sum_{f,m} \hat{X}(f,t)^{\beta-1} a_{n,j}[m] G(f - mf_0(n))} \tag{6.8}$$

donde λ es el término de regularización. La restricción de dispersión se elimina cuando $\lambda = 0$.

6.2.5. Modelos de instrumento

Todos los modelos de este capítulo necesitan que las funciones base $b_{n,j}(f)$ sean aprendidas *a priori* para cada nota n y cada instrumento j. Como se comenta en el apartado (6.2), las funciones base se pueden estimar de los picos de amplitud espectral $a_{n,j}[m]$, siendo m el número de parcial seleccionado cuando se usa la restricción armónica.

Las amplitudes $a_{n,j}[m]$ se estiman *a priori* usando la base de datos RWC [Goto02, Goto04]. Se usa el subconjunto de notas aisladas por instrumento (mas detalles en el apartado 4.1). Se establece que $R_{n,j}(t)$ se constituya como una matriz binaria tiempo/frecuencia en la que se representa la transcripción real de los datos de entrenamiento. La dimensión t representa las tramas temporales y la dimensión f representa la escala musical en numeración MIDI. Como $R_{n,j}(t)$ es conocida, la matriz de ganancias en la fase de entrenamiento se inicializa de tal manera que sólo el valor de la ganancia asociada con el pitch activo n en la trama t y tocado por el instrumento j sea la unidad, mientras que el resto de ganancias quedan inicializadas a cero. Las ganancias que se inicializan a cero, quedan siempre con valor nulo por las reglas de actualización multiplicativas, por lo que no hay opción a errar en el pitch activo de cada trama. Con esta inicialización, la aplicación de la restricción de dispersión no es necesaria para la fase de entrenamiento.

El proceso de aprendizaje de los parámetros de los modelos de instrumentos se describe en el Algoritmo 2.

Algoritmo 2 Algoritmo de entrenamiento de los modelos de instrumento

1. Se calcula $X(f,t)$ de una señal que contienen la interpretación aislada de todas las notas de cada instrumento.
2. Se inicializan las ganancias $g_{n,j}(t)$ con la transcripción real $R_{n,j}(t)$ y las amplitudes $a_{n,j}[m]$ con valores positivos aleatorios.
3. Se actualizan las amplitudes $a_{n,j}[m]$ usando la ecuación (6.7).
4. Se actualizan las ganancias $g_{n,j}(t)$ usando la ecuación (6.8) con $\lambda = 0$
5. Se repiten los pasos 2-3 hasta que el algoritmo converja o se alcance el máximo de iteraciones permitidas.
6. Se calculan las funciones base $b_{n,j}(f)$ para cada instrumento j usando la ecuación (6.2).

El algoritmo 2 de entrenamiento calcula las funciones base $b_{n,j}(f)$ para cada instrumento, necesarias para la fase de factorización. Dichas bases $b_{n,j}(f)$ se dejan fijas, por tanto, la factorización de las nuevas señales se reduce a estimar las ganancias $g_{n,j}(t)$ del modelo. Este proceso de entrenamiento y las funciones base que obtiene son usados para todos los modelos de descomposición espectral de este capítulo.

6.3. Modelo de factorización propuesto

6.3.1. Modelo armónico básico con restricción de monofonía para señales monofónicas (MBHC-MS)

En primer lugar, se introduce la restricción de monofonía en el caso más simple, para señales monofónicas (el índice j se elimina de las ecuaciones). Una vez realizado el proceso de entrenamiento, descrito en el apartado anterior, se pueden estimar las ganancias para un determinado instrumento con la ecuación (6.9), usando las funciones base de la fase de entrenamiento. La función base $b_{n_{opt}}(f)$ y la ganancia $g_{n_{opt},t}$ correspondiente se seleccionan de manera que minimicen la función de β-divergencia en la trama t, suponiendo que sólo una de las ganancias puede tener valor no nulo. Dado esto, el

modelo de señal con la restricción de monofonía (implementada de manera determinista) se describe de la siguiente manera:

$$\hat{X}_n(f,t) = g_{n_{opt},t} b_{n_{opt}}(f) \tag{6.9}$$

donde $\hat{X}_n(f,t)$ es la señal modelada para la nota optima n_{opt} en la trama t.

$$n_{opt}(t) = \arg \min_{n=1,\dots,N} D_\beta \left(X(f,t) | g_{n,t} b_n(f) \right) \tag{6.10}$$

Estimación de ganancias con *Sparse Coding* para señales monofónicas

El modelo MBHC-MS de la ecuación (6.10) permite calcular las ganancias directamente desde la señal de entrada $X(f,t)$ y las amplitudes $a_n[m]$, sin la necesidad de un algoritmo NMF iterativo para señales monofónicas. En este método, la ganancia optima en cada trama $g_{n_{opt},t}$ es aquella ganancia que minimiza la función de coste respecto de las demás. La ganancia óptima se calcula mediante una búsqueda exhaustiva, sin algoritmo iterativo, sobre el conjunto de valores de distorsión generados por cada nota en cada trama. En términos prácticos, la nota que, suponiendo que es la única nota activa, consigue el menor valor de distorsión es la nota optima en cada trama.

Usando β-divergencia, la función de coste para la nota n en la trama t se puede calcular como:

$$D_\beta(X(f,t)|g_{n,t} b_n(f)) = \sum_f \frac{1}{\beta(\beta-1)} \cdot \tag{6.11}$$

$$\cdot \Big(X(f,t)^\beta + (\beta-1)(g_{n,t} b_n(f))^\beta - \beta X(f,t)(g_{n,t} b_n(f))^{\beta-1} \Big) \tag{6.12}$$

El valor de la ganancia para la nota n en la trama t se calcula minimizando la ecuación (6.11). Convenientemente, esta minimización tiene una única solución no nula, gracias a la naturaleza escalar de las ganancias.

$$g_{n,t} = \frac{\sum_f X(f,t) b_n(f)^{(\beta-1)}}{\sum_f b_n(f)^\beta} \tag{6.13}$$

Finalmente, la nota que minimiza la β-divergencia en cada trama se selecciona como la nota óptima

$$n_{opt}(t) = \arg \min_{n=1,...,N} D_\beta \left(X(f,t) | \frac{\sum_f X(f,t) b_n(f)^{(\beta-1)}}{\sum_f b_n(f)^\beta} b_n(f) \right) \quad (6.14)$$

La solución propuesta es válida para el rango $\beta \in [0, 2]$ y señales monofónicas.

La ecuación (6.14) describe la selección de la nota óptima en la trama t para el modelo MBHC-MS. Esta ecuación, representa la nota que minimiza la distorsión entre la señal de entrada y la reconstruida con las ganancias estimadas y las funciones base seleccionadas, que corresponden con cada nota óptima.

En resumen, se ha presentado un método novedoso que impone las restricciones de armonicidad y monofonía de una manera determinista, así mismo, desarrolla la descomposición NNSC con β-divergencia [Fevotte09a] y usas información específica de instrumento, la cual ha sido obtenida de una fase de aprendizaje supervisado. Modelo armónico básico con restricción de monofonía para señales monofónicas

6.3.2. Modelo armónico básico con restricción de monofonía para mezclas polifónicas (MBHC-PM)

Las señales polifónicas se generan cuando se mezclan las señales de algunos instrumentos monofónicos que focal a la vez. Este tipo de señales son comunes en la música occidental, especialmente con instrumentos de viento. La restricción monofónica anterior se puede extender a este caso, con ciertas particularidades. El modelo de señal se define como:

$$\hat{X}(f,t) = \sum_{j=1}^{J} g_{n_j(t),j} b_{n,j}(f) \quad (6.15)$$

donde $j = 1, ..., J$ es el índice de instrumentos y $n_j(t)$ es la nota tocada por el instrumento j el el tiempo t. Ahora, el modelo de señal cuenta con diferentes conjuntos de funciones base $b_{n,j}(f)$ para cada instrumento. Se

debe puntualizar que a este modelo se le aplica una restricción monofónica porque sólo una nota $n_j(t)$ puede estar activa en cada trama t para cada instrumento j.

La ecuación (6.15) describe el modelo MBHC-PM de descomposición de señal. Al contrario que en la ecuación (6.9) (donde sólo una nota estaba presente en cada trama), hay más de una nota tocada al mismo tiempo (puede haber hasta una nota por cada instrumento). Por tanto, la señal se compone de la suma de las J contribuciones de los instrumentos. Cada contribución se describe como la multiplicación de la ganancia estimada para la nota seleccionada y su función base correspondiente.

Como en el caso del método MBHC-MS, las funciones base $b_{n,j}(f)$ para cada instrumento j se aprenden *a priori* y se mantienen fijas en la fase de factorización. Cada base modela el espectro de una única nota para un instrumento dado (ver ecuación (6.2)).

En este método se precisa conocer los instrumentos que están presentes en la señal de entrada para seleccionar los conjuntos de funciones base correspondientes. Se deben estimar, por tanto, las ganancias $g_{n_j(t),j}$ para los distintos instrumentos en cada trama.

Para este caso, del modelo armónico con restricción monofónica para mezclas polifónicas, la distorsión en la trama t usando β-divergencia se puede expresar como:

$$D_\beta(X(f,t)|\sum_{j=1}^{J} g_{n_j(t),j} b_{n,j}(f)) = \sum_f \frac{1}{\beta(\beta-1)} \cdot \tag{6.16}$$

$$\cdot \left(X(f,t)^\beta + (\beta-1)(\sum_{j=1}^{J} g_{n_j(t),j} b_{n,j}(f))^\beta - \beta X(f,t)(\sum_{j=1}^{J} g_{n_j(t),j} b_{n,j}(f))^{\beta-1} \right) \tag{6.17}$$

La ecuación (6.16) es, para el modelo MBHC-PM, la equivalente a la ecuación (6.11) para el modelo MBHC-MS. Esta ecuación representa la distorsión de la señal estimada con la nota seleccionada de cada instrumento. En el caso del modelo MBHC-MS (solo una nota estaba activa en cada instante) sólo tenía una solución no nula (ecuación 6.13). Sin embargo, en el caso del modelo MBHC-PM (más de una nota puede estar activa en

cada momento, una por instrumento), la solución se puede obtener por dos métodos: NMF (apartado 6.3.2) y *Sparse Coding* (apartado 6.3.2).

La nota óptima para cada instrumento j en cada trama t se calcula como la combinación de notas de todos los instrumentos que minimizan la distorsión en la trama t. Se calculan las ganancias $g_{n_j(t),j}$ con cada combinación de notas y posteriormente se calcula la distorsión asociada para seleccionar la combinación de notas óptima (una nota por instrumento).

Estimación de ganancias con NMF para mezclas polifónicas de fuentes monofónicas

La restricción de monoonía para mezclas polifónicas de señales monofónicas se impone mediante un esquema determinista, de manera que se le exige a las ganancias $g_{n_j(t),j}$, tener una sola ganancia activa en cada trama t para cada instrumento j. Esto implica que como máximo habrá J notas activas en cada trama. Las J notas activas (como máximo una por instrumento) serán aquellas que minimicen la distorsión entre el espectrograma de la señal de entrada y el espectrograma estimado por el modelo de descomposición. Esta combinación óptima de notas se busca sobre el rango dinámico de cada instrumento. El espacio de búsqueda combinatorio se representa a continuación:

$$\Psi = \{M_k, 1 \leqslant k \leqslant S\} \tag{6.18}$$

donde M_k es la k-ésima combinación compuesta por una única nota de cada instrumento y S es el número total de posibles combinaciones. Cada combinación M_k se puede formular como:

$$M_k = \left\{ n_j^k, j = 1, ..., J \right\} \tag{6.19}$$

donde n_j^k es la nota tocada por el instrumento j en la k-ésima combinación de Ψ

Para señales polifónicas, las ganancias no se pueden calcular de manera directa como en el caso del método MBHC-MS. Las ganancias, en este caso, deben ser calculadas con un algoritmo decreciente del gradiente. Este procedimiento se basa en la minimización de la distorsión entre el espetrograma

estimado y el espectrograma de entrada, mediante el uso del algoritmo NMF ampliado con reglas de actualización multiplicativas (MU), tal y como se describe en la ecuación (6.6), siguiendo [Lee01]. En este caso la distorsión que debe minimizarse se describe en la ecuación (6.16) y debe calcularse para cada combinación M_k de Ψ. En la práctica, la minimización se lleva a cabo con el cálculo de la derivada parcial de la distorsión. Para la nota $n_i^k(t)$ y el instrumento i de la ganancia $g_{n_i^k(t),i}$, se puede formular dicha distorsión como:

$$\frac{dD_\beta}{dg_{n_j^k(t),j}} = \sum_f \left(\sum_{j=1}^{J} g_{n_j^k(t),j} b_{n_j^k(t),j}(f) \right)^{\beta-1} b_{n_j^k(t),j}(f) - \tag{6.20}$$

$$- \sum_f X(f,t) \left(\sum_{j=1}^{J} g_{n_j^k(t),j} b_{n_j^k(t),j}(f) \right)^{\beta-2} b_{n_j^k(t),j}(f) \tag{6.21}$$

donde $n_j^k(t)$ y j indican la nota seleccionada y el instrumento, respectivamente, que deben ser minimizados para la combinación M_k. Entonces la regla MU para la ganancia $g_{n_j^k(t),j}$ se puede describir como:

$$g_{n_j^k(t),j} \leftarrow g_{n_j^k(t),j} \frac{\sum_f X(f,t) \left(\sum_{j=1}^{J} g_{n_j^k(t),j} b_{n_j^k(t),j}(f) \right)^{\beta-2} b_{n_j^k(t),j}(f)}{\sum_f \left(\sum_{j=1}^{J} g_{n_j^k(t),j} b_{n_j^k(t),j}(f) \right)^{\beta-1} b_{n_j^k(t),j}(f)} \tag{6.22}$$

La ganancia de la nota $n_j^k(t)$ y el instrumento j para la combinación M_k en cada trama t se estima usando el algoritmo del gradiente y aplicando la ecuación (6.22) durante unas pocas iteraciones. De hecho, se han usado sólo $\alpha = 5$ iteraciones. El cálculo de más iteraciones no genera mejores resultados en las pruebas realizadas. El cálculo de NMF se usa para factorizar la trama de análisis con sólo unas notas activas y evaluar la distorsión que genera su reconstrucción. Dado que el máximo número de posibles notas es 4 (se ha testado con mezclas de 4 instrumentos), y sólo se tienen que estimar las ganancias de esas 4 notas, con un número bajo de iteraciones es suficiente.

Además de todo ello, las ganancias se inicializan con la estimación directa del método MBHC-MS, suponiendo que sólo una nota está activa, por lo que la factorización únicamente debe matizar dicha inicialización.

Para justificar el uso de $\alpha = 5$ iteraciones, la tabla 6.1 muestra la distorsión que genera la factorización de 4 instrumentos con 5, 10, 15 y 20 iteraciones. La columna de 0 iteraciones representa la distorsión calculada con los valores de la inicialización.

Iteraciones	**0**	**5**	**10**	**15**	**20**
Distorsión	$2{,}85 * 10^6$	$2{,}59 * 10^6$	$2{,}59 * 10^6$	$2{,}59 * 10^6$	$2{,}59 * 10^6$

Tabla 6.1: *Distorsión causada por la factorización con el modelo MBHC-PS con* $[0, 5, 10, 15, 20]$ *iteraciones sobre un fichero con la mezcla de* 4 *instrumentos*

Tras estimar las ganancias $g_{n_i^k(t),j}$ para todas las combinaciones M_k, se aplica la ecuación (6.16) para el cálculo de la distorsión asociada. La solución óptima en cada trama se obtiene mediante la selección de la combinación M_k que genera menor distorsión, como se indica en la ecuación (6.23)

$$M_{k_{opt}} = \arg \min_{M_k \in \Psi} D_\beta \left(X(f,t) | \sum_{j=1}^{J} g_{n_j^k(t),j} b_{n_j^k(t),j}(f) \right) \tag{6.23}$$

En resumen, el método para la descomposición de mezclas polifónicas de instrumentos monofónicos usando β-divergencia se muestra en el algoritmo 3. El rendimiento de este algoritmo en las aplicaciones de SSS y AMT se muestra en las tablas 6.4 y 6.6, y se compara comparación con otros métodos del estado del arte.

Estimación de ganancias usando *Non Negative Sparse Coding (NNSC)* para mezclas polifónicas de fuentes monofónicas

A pesar de la reducción del número reducido de iteraciones que se precisan para ejecutar el algoritmo de factorización de la sección 6.3.2, el proceso debe repetirse para cada combinación M_k del conjunto Ψ. Es ampliamente conocido que la naturaleza iterativa del algoritmo NMF lo hace poco recomendable para aplicaciones en tiempo real.

Algoritmo 3 Algoritmo de estimación de ganancias para MBHC-PM

```
Se iniciativa b_{n,j}(f) con los modelos de instrumento correspondientes.
for t=1 hasta el número de tramas de análisis do
  for k=1 hasta S do
    Se inicializan las ganancias g_{n_j^k(t),j} con los valores del método directo MBHC-MS (suponiendo que sólo hay presente un instrumento en la señal) para las notas n_j^k de la combinación M_k y cero para el resto.
    for α iteraciones do
      for i=1 hasta J do
        Se actualizan las ganancias g_{n_i^k(t),i} usando la ecuación (6.22).
      end for
    end for
    Se calcula la β-divergencia con la ecuació (6.16).
  end for
  Se selecciona la combinación de notas M_k que generan el menor valor de β − divergencia usando la ecuación (6.20).
end for
```

El modelo MBHC-PM se puede adaptar para que, mediante *sparse coding*, se obtenga una solución directa (como en el modelo MBHC-MS), y de esta manera, evitar el proceso iterativo. Esta opción permitiría usar MBHC-PM en aplicaciones en tiempo real para señales con niveles de bajos de polifonía, tal y como se demuestra en la tabla 6.5.

Para $\beta = 2$ (distancia Euclieda), la ecuación (6.20) se puede simplificar de manera que las ganancias se calculen directamente con NNSC, sin ningún algoritmo iterativo. El mínimo global de la función de coste de la ecuación (6.20) se encuentra para $\beta = 2$ suponiendo que $D_\beta = 0$. La expresión resultante se puede particularizar para la combinación M_k en la trama t de manera que:

$$\sum_{j=1}^{J} g_{n_j^k(t),j} \sum_f b_{n_i^k(t),i}(f) b_{n_j^k(t),j}(f) = \sum_f b_{n_i^k(t),i}(f) X(f,t) \tag{6.24}$$

La ecuación (6.24) se puede expresar de manera matricial como:

$$\mathbf{gB} = \mathbf{c} \tag{6.25}$$

donde $\mathbf{g}$ es un vector de tamaño $1 \times J$, $\mathbf{B}$ es una matriz de tamaño $J \times J$ que depden de las funciones base y $\mathbf{c}$ es un vector de tamaño $1 \times J$ que depende de las ganancias y de la señal de audio. $g(j) = g_{n_j^k(t),j}$ es el vector de ganancias para la combinación seleccionada M_k en la trama T, $B(j,i) = \sum_f b_{n_i^k(t),i}(f) b_{n_j^k(t),j}(f)$ y $c(i) = \sum_f b_{n_i^k(t),i}(f) X(f,t)$. $\mathbf{B}$ se puede calcular *a priori*, porque es la matriz de correlación cruzada de las bases $b_{n_j^k(t),j}$. $\mathbf{B}$ toma valores altos cuando las notas están armónicamente relacionadas y valores bajos en cualquier otro caso. $\mathbf{c}$ debe calcularse de manera online, puesto que depende del espectrograma de la señal de audio.

Con todo ello, las ganancias se pueden calcular en un solo paso siguiendo la siguen ecuación:

$$\mathbf{g} = \mathbf{cB}^{-1} \tag{6.26}$$

donde $g(j) = g_{n_j^k(t),j}$. La ecuación (6.26) puede generar valores negativos, los cuales se sustituyen por valor cero, como se indica en [MaxerPhD13].

Una vez estimadas las ganancias para todas las combinaciones de Ψ, se usa la ecuación(6.23) para seleccionar la combinación óptima $M_{k_{opt}}$ que genera la menor distorsión en cada trama.

6.3.3. Selección de candidatos para mezclas polifónicas de fuentes monofónicas

Una búsqueda exhaustiva sobre Ψ conlleva un coste computacional elevado, puesto que Ψ contienen un gran número de combinaciones, lo cual incrementa drásticamente cuando crece el número de instrumentos (nivel de polifonía).

Una expresión general para calcular el número de combinaciones de los elementos de un conjunto con repetición de notas por instrumento es:

$$S = \left(\frac{N_t}{J}\right) = \frac{N_t!}{J!(N_t - J)!} \tag{6.27}$$

donde S es el número total de combinaciones, $N_t = \sum_{j=1}^{J} N(j)$ es el número total de notas de todos los instrumentos, $N(j)$ es el número de notas del instrumento j y J es el número de notas (elementos) de cada combinación, que coincide con el número de instrumentos, si éstos son monofónicos. La expresión anterior se debe modificar para eliminar las combinaciones que contienen más de una nota del mismo instrumento (sin repetir notas del mismo instrumento), quedando como sigue:

$$S = \frac{N_t!}{J!(N_t - J)!} - \sum_{j=1}^{J} \frac{N(j)!}{J!(N(j) - J)!} \tag{6.28}$$

donde J es el número de instrumentos y $N(j)$ es el número de posibles notas tocadas por el instrumento j.

Por ejemplo, un dueto de violin (46 posibles notas) y clarinete (40 posibles notas), producirían 1840 combinaciones según la ecuación (6.28). Es más, para una señal con nivel de polifonía 4 (con fagot, clarinete, violin y saxofón) el número de combinaciones alcanza los 23 millones. Este gran número de combinaciones posibles genera un coste computacional muy elevado, por lo que el espacio de búsqueda Ψ debe ser reducido. Esta reducción del espacio de búsqueda se puede llevar a cabo limitando las posibles notas para cada instrumento. En la ecuación (6.28), el número de posibles notas para cada instrumento $N(j)$ contiene todas las notas del rango dinámico que puede tocar cada instrumento j. En vez de dar todas esas opciones, el espacio de búsqueda se limita a C notas candidatas para ser la óptima para cada instrumento. Estas notas candidatas precisan ser seleccionadas previamente mediante un algoritmo de transcripción rápido.

Las notas candidatas se seleccionan usando información de los modelos de instrumentos presentes en la señal de entrada. La selección de candidatos debe ser rápida para que pueda presentarse como una alternativa que ahorre coste computacional y tiempo.

En este capítulo se propone la obtención de una subconjunto de notas candidatas mediante el uso del modelo MBCH-MS del apartado 6.3.1. Aunque el modelo MBCH-MS se ha diseñado para señales monofónicas, se ha adaptado para señales polifónicas, de manera que se realiza la estimación de ganancias suponiendo que sólo un instrumento está presente en la señal

de entrada (a pesar de conocer que es una señal polifónica). La distorsión que obtiene esta solución monofónica se calcula con la ecuación (6.13) y la ecuación (6.11). Las C notas que causan una menor distorsión con este cálculo son las notas seleccionadas como candidatas para el instrumento en cuestión, y de esa manera reducir el espacio de búsqueda del método MBHC-PM. Esta selección de candidatos tiene un coste computacional muy bajo, lo que repercute en una selección de candidatos muy rápida.

El algoritmo 4 describe el procedimiento de cálculo para la selección de notas candidatas.

Algoritmo 4 Algoritmo de selección de candidatos

```
1 Se inicializan las bases b_{n,j}(f) con los datos de entrenamiento de cada
  instrumento
2 for j=1 hasta J do
3   for t=1 hasta el número de tramas de análisis do
4     Se calcula el método MBHC-MS con la ecuación (6.13) y la ecuación
      (6.11)
5     Se seleccionan las C notas que producen los menores valores de
      β-divergenncia para el instrumento j en la trama t, siendo C el
      número de notas candidatas
6   end for
7 end for
```

% de notas perdidas	**No. de notas candidatas por instrumento**				
Polifonía	**5**	**10**	**15**	**20**	**25**
2	9 %	5 %	1,6 %	0,3 %	0,08 %
3	16 %	7 %	2,2 %	0,4 %	0,1 %
4	24 %	10 %	2,8 %	0,4 %	0,1 %

Tabla 6.2: *Porcentaje de notas perdidas por la selección de candidatos*

Llegados a este punto, la clave es determinar el número óptimo de candidatos C que reduce el coste computacional de manera suficiente, y que no sea tan restrictivo como para perder la nota correcta. Se ha evaluado el rendimiento usando la base de datos de corales de Bach [Duan11] para

determinar el número de candidatos que se seleccionarán por cada instrumento. Los resultados se muestran en la tabla 6.2. Si se seleccionan 15 candidatos por instrumento se pierden menos del 5 % de las notas correctas en el proceso de selección de candidatos, lo que se considera aceptable para esta aplicación.

Nivel de polifonía	**2**	**3**	**4**
Selección de candidatos ($C = 15$)	225	12.825	483.000
Rango dinámico completo	1560	197.000	23.987.000

Tabla 6.3: *Número de combinaciones S para la selección de candidatos* ($C = 15$) *y usando el rango dinámico completo de cada instrumento. Para polifonía de nivel 2 se han usado fagot y clarinete, para polifonía 3 se han usado fagot, clarinete y saxofón; y para polifonía 4 se han usado fagot, clarinete, saxofón y violín*

La tabla 6.3 compara el número de combinaciones con el proceso de selección de candidatos propuesto, y sin él. Se muestra el que el número de combinaciones se reduce enormemente si se seleccionan 15 candidatos por cada instrumento. Se debe indicar que el número de combinaciones para el proceso de selección de candidatos se realiza usando la ecuación (6.28), donde C sustituye a $N(j)$ como el número de posibles notas por instrumento. La consecuencia de la aplicación de este proceso de selección de candidatos será evaluado a continuación, aplicándolo a AMT y SSS.

6.4. Evaluación

En este apartado se evalúan los algoritmos propuestos en el apartado 6.3 aplicados en SSS y AMT con señales polifónicas compuestas por instrumentos monofónicos. Así mismo, los resultados se comparan con los obtenidos por otros algoritmos del estado del arte.

En el caso de la AMT se obtiene una transcripción independiente para cada instrumento. Hasta donde se conoce, en la bibliografía no hay otros trabajos que obtengan este tipo de información. Por estas razón, se ha adaptado un método de descomposición de señal del estado del arte, diseñado para señales monofónicas, para poder comparar el método propuesto.

6.4.1. Datos de entrenamiento y evaluación

En la fase de entrenamiento (ver apartado 6.2.5), las funciones base se estiman usando la base de datos *RWC Musical Instrument Sound Database* y todo el rango dinámico para cada instrumento. Se han utilizado cuatro instrumentos en la fase de evaluación (violín, clarinete, saxofón y fagot). Los sonidos de esta base de datos se encuentran disponibles con una resolución de un semitono a lo largo de todo el rango dinámico de cada instrumento. En la base de datos se encuentran interpretaciones de todas las notas con varios estilos interpretativos y matices dinámicos, como se describe en el apartado 4.1.2. Se han usado los ficheros con un estilo normal y un matiz dinámico *mezzo*. Se ha comprobado que el entrenamiento con diferentes estilos lleva a la obtención de distintos modelos de instrumento. Sin embargo, en [Carabias11] se demuestra que la configuración seleccionada (estilo normal y matiz *mezzo*) obtiene un modelo representativo de todos los demás.

Para la fase de evaluación, se usa la base de datos propuesta en [Duan11] (ver apartado 4.1.5). Esta base datos contiene 10 corales de cuatro partes de J.S. Bach con su información MIDI correspondiente. Los ficheros de audio tienen una duración aproximada de 30 segundos y las grabaciones reales están muestreadas a 44,1KHz. Cada coral de la base de datos está compuesta por un cuarteto de instrumentos (violín, clarinete, saxofón y fagot) y cada instrumento está grabado en una pista independiente. Estas pistas se mezclan para crear un total de 10 interpretaciones con nivel de polifonía 4, 60 duetos y 40 trios.

6.4.2. Configuración de los experimentos

Representación tiempo/frequencia de las señales

En la bibliografía se encuentran muchos sistemas de procesado de señal que usan una discretización logarítmica de frecuencia. Por ejemplo, en [Bertin10, Vincent10] se usa una división en bandas uniformemente espaciadas en la escala *Equivalent Rectangular Bandwidth (ERB)*. En este capítulo se usa la resolución de un semitono como en [Carabias11]. Además la base de datos de entrenamiento y la anotación real de los datos tienen esa misma resolución, con notas separadas un semitono, como se ha comenta-

do anteriormente. Por ello, se emplea una representación tiempo/frecuencia obtenida mediante la integración de los bins de la STFT que corresponden a cada intervalo de semitono.

El tamaño de trama de análisis y el salto empleado son $128ms$ y $32ms$ respectivamente. Otros valores de parámetros que se han configurado para la fase experimental son los siguientes: 1) 20 parciales para cada función base en los modelos armónicos ($M = 20$); y 2) máximo de 50 iteraciones de los algoritmos NMF, excepto para el caso del modelo MBHC-PM, cuyo valor ha sido 5, dicha elección se justifica en el apartado 6.3.2.

Separación de musical: métodos y medidas de calidad

- La separación de fuentes consiste en setimar la amplitud correspondiente de cada posición tiempo/frequencia (en las matrices de representación de señal empleadas) para cada fuente por separado. Algunos sistemas emplean la separación binaria, la cual entrega toda la energía de un punto tiempo/frecuencia de la matriz de la señal de entrada a uno de los instrumentos. Sin embargo, se ha demostrado que se obtienen mejores resultados cuando la decisión no es binaria, sino que se reparte la energía de manera proporcional entre todas las fuentes. En la práctica, este es el método más adecuado para las señales armónicas polifónicas, debido al solapamiento de algunos de los parciales de unas notas con los de otras. El uso de máscaras de Wiener es habitual en la bibliografía de la separación de fuentes [Every06]. En este capítulo se han usado modelos de instrumento que permiten realizar una buena estimación de la amplitud de los parciales solapados.

- Para una evaluación objetiva del rendimiento del método de separación se han usado las medidas de calidad de separación implementadas en [Emiya11, Vincent12]. Estas medidas están ampliamente aceptadas por la comunidad científica especializada en este ámbito, lo que facilita enormemente la comparación del rendimiento de diferentes algoritmos. Se supone que cada fuente separada produce un modelo de distorsión tal que:

$$\hat{s}_j(t) - s_j(t) = e_j^{target}(t) + e_j^{interf}(t) + e_j^{artif}(t) \tag{6.29}$$

donde $\hat{s}_j$ es la fuente estimada para el instrumento j, s_j es la fuente original del instrumento j,e^{target} es el término de error asociado con la distorsión respecto de la fuente j original, e^{interf} es el término de error causado por la interferencia de otras fuentes y e^{artif} es el término de error asociado a los artefactos generados por el propio sistema de separación.

Las métricas para cada señal separada son *Source to Distortion Ratio* (SDR), *Source to Interference Ratio* (SIR), y *Source to Artifacts Ratio* (SAR) [Emiya11, Vincent12].

$$SDR_j = 10\log_{10}\frac{\sum_t |s_j(t)|^2}{\sum_t |\hat{s}_j(t) - s_j(t)|^2} \tag{6.30}$$

$$SIR_j = 10\log_{10}\frac{\sum_t \left|s_i(t) + e_j^{target}(t)\right|^2}{\sum_t \left|e_j^{interf}(t)\right|^2} \tag{6.31}$$

$$SAR_j = 10\log_{10}\frac{\sum_t \left|s_i(t) + e_j^{target}(t) + e_j^{interf}(t)\right|^2}{\sum_t \left|e_j^{artif}(t)\right|^2} \tag{6.32}$$

Transcripción musical: método y medidas de calidad

- Dadas las amplitudes variantes en el tiempo de todas las funciones base $g_{n,j}(t)$, el método de transcripción empleado es el mismo que en los trabajos de [Carabias11, Bertin10, Vincent10]. Se determina si una nota está activa, o no, usando la siguiente ecuación para cada una de las funciones base:

$$\Omega(n,j,t) = g_{n,j}(t) \geq \left(10^{T/20} \max_{nt} g_{n,j}(t)\right) \tag{6.33}$$

donde $\Omega(n,j,t)$ es la transcripción binaria resultante y T es el umbral de detección en decibelios (dB) que se establece en función de los datos de entrenamiento.

En los métodos basados en el modelo BHC se precisa de un umbral para decidir cuales de las notas están activas en cada trama. Por el contrario, en los métodos basados en los modelos MBHC no es necesario, puesto que sólo una nota por instrumento está activa en cada instante. Sin embargo, si que se utiliza un umbral para decidir si existen notas activas, por ejemplo en los intervalos de silencio.

- Los métodos de transcripción se pueden evaluar con dos tipos de medidas: medidas a nivel de nota y medidas a nivel de trama. En la evaluación de las propuestas de este capítulo se han usado las medidas a nivel de trama, como en [Carabias11]. En la práctica, se usan las medidas a nivel de trama descritas en [Dixon00b] para la evaluación objetiva de los métodos de transcripción. La medida de precisión Acc(%) se define como:

$$Acc = \frac{TP}{FP + FN + TP} \tag{6.34}$$

donde TP *(true positives)* es el número de tramas con notas correctamente transcritas, FP *(false positives)* es el número de tramas con notas transcritas como activas cuando no lo están y FN *(false negatives)* es el número de tramas con notas activas que no se han transcrito. Acc puede indicar en un rango desde 0 hasta 1, considerándose el valor 1 como la transcripción perfecta.

6.4.3. Métodos para comparación

Los beneficios del uso de las propuestas del apartado 6.3 se destacan en comparación con los métodos del apartado 6.2 (BCH y BHC con restricción de dispersión). Además, los métodos propuestos se comparan también con dos métodos con restricción monofónica del estado del arte: *Gaussian Scaled Mixture Model (GSMM)* [Benaroya06] y *Factorial Scaled Hidden Markov Models (FS-HMM)* [Ozerov09]. Ambos están implementados en la herramienta *Flexible Audio Source Separation Toolbox (FASST)* [Ozerov11].

A pesar de que FASST se diseño originariamente para la separación de fuentes, se ha adaptado sus datos de salida para emplearlos también en

transcripción musical. FASST ofrece una matriz de ganancias a la salida del proceso de factorización. Se ha establecido un umbral sobre esta matriz para poder obtener una transcripción binaria de la señal de entrada. Este proceso de aplicación de umbral a las matrices de ganancias se ha usado de la misma manera para el método propuesto en el apartado 6.3 como se indica en el apartado 6.4.2.

Se han evaluado varias configuraciones de la herramienta FASST, se muestran los resultados más significativos para la comparación con los métodos de la propuesta. FASST permite usar tanto representación clásica de señal con FFT como con QERB, que es más adecuada para las señales musicales por su resolución frecuencial logarítmica, en contra de la resolución lineal de la FFT. En una resolución lineal, pequeñas variaciones de la frecuencia fundamental pueden producir variaciones mayores del ancho del lóbulo principal de la ventana, cuando se tratan las altas frecuencias. El mejor rendimiento se ha encontrado usando la escasa QERB para representar las señales y ejecutando la factorización con el algoritmo *Generalised Expectation Maximization (GEM)*, en el que se ha modificado el modelo generativo para que use una distribución de Poisson (en la versión original, FASST usa una distribución Gaussiana con divergencia IS). El uso de la distribución de Poisson como modelo generativo es equivalente a la factorización con la divergencia de KL ($\beta = 1$) (ver apartado 3.4.2). El número de bases se ha configurado en $K = 114$ (notas midi desde 24 hasta 137), lo que lo hace independiente del instrumento, puesto que todos los instrumentos usados en la evaluación tienen sus rangos dinámicos comprendidos en él.

6.4.4. Resultados

Como ya se ha indicado, se ha evaluado la fiabilidad del método propuesto sobre aplicaciones de SSS y AMT con mezclas polifónicas de fuentes monofónicas de la base de datos descrita en el apartado 4.1.5. Se ha estudiado el comportamiento de los modelos BHC y BHC con restricción de dispersión y los métodos MBHC-PM en función del parámetro β. En la práctica, el valor $\beta = 1{,}5$ obtiene los mejores resultados. En [Carabias11] se puede ver la optimización del parámetro β. Por tanto el método MBHC-PM usará el valor óptimo de β, mientras que la versión MBHC-PM factorizada

con *Sparse Coding* usará $\beta = 2$. Como se muestra a continuación, los resultados obtenidos con MBHC-PM factorizado con NNSC no distan mucho de la versión iterativa (MBHC-PM con NMF), mientras que si que obtiene un ahorro importante en coste computacional y tiempo de ejecución, por ello será la versión adecuada del algoritmo para aplicaciones en tiempo real.

Los resultados que se muestran son la media de todos los ficheros evaluados por separado en cada método y aplicación. Siguiendo lo descrito en [Carabias11], los parámetros a estimar del modelo NMF se inicialzan de manera aleatoria y las medidas para cada fichero se repiten en 30 ejecuciones. En los experimentos que se han llevado a cabo, el 95 % del intervalo de confianza se encuentra por devalúo del 1,6 % en todos los algoritmos. Esto quiere decir que las diferencias existentes entre todos los algoritmos son estadísticamente importantes. Un resultado similar general los experimentos de separación de fuentes, donde el 95 % del intervalo de confianza para las medidas de SDR supone menos de 1,4 dB para todos los métodos en comparación.

Resultados en Separación de fuentes (SSS)

	J=2			J=3			J=4		
Método	**SDR**	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**
NMF MBHC-PM	**7,94**	20,03	11,95	**3,14**	15,63	4,84	-	-	-
NMF MBHC-PM *con selección de candidatos*	**7,94**	20,01	11,95	**3,14**	15,56	4,82	**1.81**	14,33	1,43
NNSC MBHC-PM	6,15	17,83	9,38	2,51	14,77	3,6	-	-	-
NNSC MBHC-PM *con selección de candidatos*	6,31	18,32	9,6	2,51	14,77	3,6	1.38	13,64	0,51
BHC *con restricción de dispersión*	4,32	17,34	4,72	2,04	15,64	3,62	1.26	14,13	-2,01
BHC	4,19	18,24	4,56	1,9	14,41	1,5	1.13	23,9	0,77
FS-HMM	3,6	14,54	5,5	1,4	12,32	4,35	0,94	10,27	3,05
GSMM	3,72	14,7	5,9	1,56	13,01	4,56	0,96	10,58	3,12

Tabla 6.4: *Resultados de separación de fuentes (dB) usando polifonía de nivel 2, 3 y 4, para los métodos: MBHC-PM con factorización NMF (NMF MBHC-PM* $\beta = 1{,}5$*), MBHC-PM con factorización NMF y selección de candidatos (NMF MBHC-PM con selección de candidatos* $\beta = 1{,}5$*), MBHC-PM con factorización NNSC (NNSC MBHC-PM* $\beta = 2$*) y MBHC-PM con factorización NNSC y selección de candidatos (NNSC MBHC-PM con selección de candidatos* $\beta = 2$*). Se muestra también la comparación con métodos del estado del arte (BHC, BHC con restricción de dispersión (*$\lambda = 1$*), GSMM y FS-HMM).*

Se muestran en la tabla 6.4 los resultados numéricos en la aplicación de SSS en términos de SDR, SIR y SAR (en dB) para todos los métodos evaluados.

Los métodos MBHC-PM y MBHC-PM con selección de candidatos muestran resultados muy similares para todos los niveles de polifonía, lo cual demuestra que el uso de 15 candidatos por instrumentos es una buena elección. En el apartado 6.3.3, se justifica el uso de la selección de candidatos por el gran coste computacional que supone no usarlo. La tabla 6.2 muestra que se pierden menos del 5 % de las notas con el uso de 15 candidatos por instrumento. La tabla 6.4 confirma que el uso de esta reducción del espacio de búsqueda exhaustiva no tiene consecuencias negativas en los resultados de separación de fuentes.

El método NNSC MBHC-PM ($\beta = 2$) es levemente superado por la versión NMF MBHC-PM ($\beta = 1{,}5$) en todos los niveles de polifonía, pero cuenta con un ahorro importante en coste computacional. Esta ventaja lo hace muy adecuado para aplicaciones en tiempo real.

Teniendo en cuenta todas estas consideraciones, se puede decir que todos los algoritmos MBHC-PM rinden mejor que los otros algoritmos con los que se han comparado, alcanzando un valor de 7,94 dB en polifonía de nivel 2. El siguiente método con mejor resultado (BHC con restricción de dispersión) obtiene una media de, aproximadamente, 2,5 dB por debajo de los métodos MBHC-PM. La razón de los mejores resultados de los algoritmos MBHC-PM frente a las versiones de BHC es el uso de la restricción de monofonía. Además, los modelos con restricción monofónica eluden en mayor medida las interferencias con otras fuentes y los artefactos generados por la separación, como se puede ver en los valores de SIR y SAR de la tabla 6.4 en todos los niveles de polifonía. El método BHC sin restricción alcanza un valor similar a la versión con restricción de dispersión, sin embargo los métodos FS-HMM y GSMM obtienen peores resultados. Este menor rendimiento de los métodos de la herramienta FASST puede estar causado por el menor número de parámetros que tiene que se tienen que calcular en los modelos MBHC-PM, BHC y BHC con restricción de dispersión frente a los métodos FS-HMM y GSMM. Todos los modelos con restricción armónica tienen que estimar un número similar de parámetros, puesto que en ellos se definen las funciones base con sólo M valores de amplitud como se describe en la

ecuación (6.2). Sin embargo, los métodos de la herramienta FASST usan todos los *bins* en frecuencia para representar las funciones base.

La tabla 6.5 muestra las medidas de tiempo de ejecución para un fichero de 30 segundos de duración de un dueto y un terceto. Se puede ver que el uso de la etapa de selección de candidatos reduce considerablemente el tiempo de ejecución. Así mismo se puede ver que los algoritmos BHC, FS-HMM y GSMM no son adecuados para aplicaciones en tiempo real. El método NNSC MBHC-PM con selección de candidatos y $\beta = 2$ reduce el tiempo de ejecución en, aproximadamente, un 40 %, pero los resultados empeoran (ver tabla 6.4). Sin embargo, la reducción de tiempo de ejecución mas grande se consigue usando la fase de selección de candidatos. El hecho de seleccionar $C = 15$ notas candidatas por instrumento obtiene una calidad de separación similar a la vez que reduce el tiempo de ejecución en más de un 99 % para los ejemplos de la tabla 6.5. El algoritmo MBHC-PM sin selección de candidatos no se ha ejecutado con nivel de polifonía 4 por la gran cantidad de combinaciones que genera (ver tabla 6.3).

El método MBHC-PM (con ambos tipos factorización: NNSC y NMF) con la selección de candidatos y sin ella, obtiene unos resultados muy similares, como se puede ver en la tabla 6.4. Los resultados en la aplicación de AMT se ejecutarán, por tanto, sólo con la versión que cuenta con selección de candidatos.

Finalmente, una implementación en tiempo real sólo sería posible para los métodos NMF MBHC-PM y NNSC MBHC-PM, ambos con selección de candidatos para $J = 2$, como se puede ver en la tabla 6.5.

Todos los experimentos se han llevado a cabo usando Matlab sobre un procesador Intel Xeon de 2.0GHz.

Resultados en transcripción automática de música (AMT)

La tabla 6.6 muestra los resultados de AMT para los mismos métodos que han sido evaluados para SSS, siendo eliminados en esta tabla las versiones sin selección de candidatos. Los resultados de AMT concuerdan con los obtenidos en SSS para cada método.

El método MBHC-PM supera claramente a todos los demás métodos comparados, como ocurre en SSS. De esta manera se demuestra la fiabili-

Método	**J=2**	**J=3**
NMF MBHC-PM	18.074 s	1.026.342 s
NNSC MBHC-PM	3.157 s	179.949 s
NMF MBHC-PM con selección de candidatos	21 s	1.228 s
NNSC MBHC-PM con selección de candidatos	13 s	747 s
BHC con restricción de dispersión	356 s	20.295 s
BHC	356 s	20.295 s
FS-HMM	23.425 s	1.335.225 s
GSMM	24.362 s	1.388.634 s

Tabla 6.5: *Tiempo de ejecución para una pieza de* 30 *segundos de duración con niveles de polifonía* 2 *y* 3.

dad de la restricción de monofonía para lase señales polifónicas compuestas por fuentes monofónicas. De nuevo, los mejores resultados los obtiene el algoritmo NMF MBHC-PM ($\beta = 1{,}5$) frente al algoritmo NNSC MBHC-PM ($\beta = 2$). Por ello, se puede llegar a decir que, como en [Carabias11], la distancia Euclidea ($\beta = 2$) no es el valor óptimo del parámetro β, sin embargo el algoritmo NNSC, que sólo se puede ejecutar con $\beta = 2$, es mucho menos complejo que la versión NMF, hablando en términos computacionales.

La principal diferencia entre los resultados de AMT y SSS se da en los métodos BHC y BHC con restricción de dispersión. Se aprecia un incremento considerable, en la distancia entre ellos, si se miran los resultados de la tabla 6.6 y la tabla 6.4. Por tanto, se puede decir que la restricción de dispersión es más efectiva en AMT que en SSS, probablemente a causa de la decisión del umbral para tomar la decisión de la transcripción binaria (ver ecuación (6.33)). Por otro lado, la SSS con máscaras de Wiener y restricción de dispersión favorece la concentración de energía en ciertas posiciones tiempo/frecuencia, pero como la energía se distribuye proporcionalmente entre los instrumentos, todos ellos tienen cierta energía en todas las posiciones tiempo/frecuencia.

En general, las restricciones de dispersión y monofonía encajan bien con las fuentes monofónicas que con los métodos que no usan estas restricciones (modelo BHC). Así mismo, la restricción monofónica se muestra como una

Método	Instrumento	J=2	J=3	J=4
NMF MBHC-PM *con selección de candidatos*	*fagot*	0,55	0,4	0,32
	clarinete	0,65	0,44	0,39
	saxofón	0,64	0,43	0,38
	violín	0,55	0,39	0,33
	Media	**0,60**	**0,42**	**0,35**
NNSC MBHC-PM *con selección de candidatos*	*fagot*	0,38	0,3	0,23
	clarinete	0,6	0,41	0,28
	saxofón	0,56	0,36	0,27
	violín	0,43	0,31	0,22
	Media	**0,49**	**0,35**	**0,25**
BHC *con restricción de dispersión*	*fagot*	0,33	0,26	0,20
	clarinete	0,53	0,41	0,34
	saxofón	0,5	0,33	0,19
	violín	0,32	0,21	0,16
	Media	**0,42**	**0,3**	**0,22**
BHC	*fagot*	0,33	0,23	0,19
	clarinete	0,41	0,26	0,20
	saxofón	0,36	0,18	0,12
	violín	0,3	0,16	0,14
	Media	**0,35**	**0,21**	**0,16**
FS-HMM	*fagot*	0,27	0,15	0,1
	clarinete	0,33	0,16	0,12
	saxofón	0,22	0,09	0,09
	violín	0,25	0,14	0,11
	Media	**0,27**	**0,14**	**0,11**
GSMM	*fagot*	0,24	0,14	0,09
	clarinete	0,35	0,17	0,12
	saxofón	0,22	0,15	0,1
	violín	0,3	0,16	0,12
	Media	**0,28**	**0,15**	**0,1**

Tabla 6.6: *Resultados de transcripción automática musical (Acc) para niveles de polifonía* 2, 3 *y* 4 *para los métodos: NMF MBHC-PM con selección de candidatos y* $\beta = 1{,}5$ *y NNSC MBHC-PM con selección de candidatos y* $\beta = 2$*). Se muestra también la comparación con métodos del estado del arte (BHC, BHC con restricción de dispersión (*$\lambda = 1$*), GSMM y FS-HMM).*

mejor opción, para señales polifónicas compuestas por fuentes monofónicas, que la restricción de dispersión de la ecuación (6.5).

Todos los métodos empeoran la precisión de sus resultados cuando crece el nivel de polifonía, puesto que es más complicado distinguir cada nota de cada instrumento. Esto se produce porque, al incrementar el nivel de polifonía, se complica la tarea de seleccionar la función base que mejor encaja con el espectro, puesto que se producen solapamientos que corrompen la naturaleza de las notas. Se debe recalcar que, el sistema propuesto aporta una transcripción para cada instrumento, otros métodos para señales polifónicas, como los propuestos en [KlapuriPhD04, Vincent10], ejecutan una transcripción general de la pieza, sin segregar la información de cada instrumento. Este mismo efecto de pérdida de precisión por el incremento del nivel de polifonía se ve reflejado en la tabla 6.4, con los resultados de SSS.

Los métodos FS-HMM y GSMM sufren las mismas dificultades que en SSS: tienen que estimar más parámetros que los otros modelos por la ausencia de una restricción armónica, lo que resulta en un rendimiento peor en estas aplicaciones.

Si se estudian los resultados para cada tipo de instrumento, se puede ver que los resultados para saxofón y clarinete están un 10 % por encima de los de fagot y violín. La diferencia en el rendimiento de los resultados de cada instrumento se puede atribuir al parecido del modelo de instrumento entrenado con el que se ha tocado en la señal de evaluación. Hay que recordar que se han usado dos bases de datos diferentes para entrenamiento y para evaluación. Esta discordancia entre los modelos entrenados y la señal real puede estar causada por la forma que tiene el músico de interpretar, como por ejemplo la manera de frotar las cuerdas del violín, o bien, algunas diferencias físicas entre ambos instrumentos, como en el caso del fagot.

6.5. Conclusiones

En este capítulo, se ha propuesto un método de factorización de señal con restricción de monofonía (MBHC-PM) para mezclas polifónicas de fuentes monofónicas, en el cual las restricciones armónica y de monofonía se implementan de manera determinista. Se han presentado dos versiones del método para realizar la factorización: un algoritmo basado en NMF (adecua-

do para $\beta = [0, 2]$) y otro algoritmo basado en NNSC de baja complejidad (sólo válido para $\beta = 2$). El método MBHC-PM y otros métodos del estado del arte se han evaluado sobre una base de datos que contiene 40 ficheros de solos de instrumentos monofónicos (fagot, clarinete, violín y saxofón), 10 ficheros por instrumento. Se han ejecutado los métodos para las aplicaciones de SSS y AMT, y se han evaluado los resultados, obteniéndose, en ambos casos, los mejores resultados con el método propuesto MBHC-PM.

El método propuesto ofrece una transcripción por instrumento en los experimentos de AMT, gracias al uso de modelos de instrumento, previamente aprendidos, para poder distinguir entre las notas de unos y otros instrumentos.

Los métodos BHC y BHC con restricción de dispersión no tienen restricción de monofonía, y por tanto, son más adecuados para las señales polifónicas. A estos métodos les cuesta mucho acumular la energía en la activación de una sola base, por lo que sus resultados son peores para el escenario propuesto.

Los métodos FS-HMM y GSMM deben estimar una gran cantidad de parámetros, frente a los pocos parámetros de los métodos con restricción armónica. Esto hace que sus resultados no alcancen los mejores rendimientos.

Los resultados de SSS y AMT muestran que el incremento en el nivel de polifonía tiene un efecto negativo directo en los resultados obtenidos. Sin embargo, en niveles bajos de polifonía, usando modelos de instrumento, se alcanzan resultados prometedores.

En resumen, en este capítulo se remarca las ventajas del método MBHC-PM sore otros métodos del estado del arte, principalmente por las restricciones de monofonía y armonicidad. Se concluye también que es un método apto para aplicaciones en tiempo real, con señales cuyo nivel de polifonía sea bajo.

Capítulo 7

Separación online e informada de fuentes con modelos adaptativos

En este capítulo se propone un sistema *online* de separación de fuentes con información temporal y modelos de instrumento adaptativos. El sistema se desarrolla sobre un esquema de factorización NMF, donde cada instrumento se modela con un modelo de filtro fuente con excitación múltiple que aporta flexibilidad y parametrización para su adaptación. Se entrenan los modelos de instrumento iniciales con grabaciones de instrumentos del mismo tipo, posteriormente, durante el proceso de factorización, estos modelos iniciales se actualizan bajo ciertas condiciones para adaptarlos al instrumento real que se está separando. Para determinar las condiciones de la actualización es necesario conocer la información simbólica temporal de la composición, la cual viene dada por una fase previa de alineamiento entre dicha información y la señal de audio. Además esta fase de alineamiento ofrece los datos de manera *online*, combinando así con la misma característica del proceso de separación. Se demuestra que el uso de modelos adaptativos frente al uso de modelos fijos y la separación sin información de instrumento. Además se compara el rendimiento del sistema de manera *online* frente a la versión *offline*, valorando los resultados en una comparación con el estado del arte.

7.1. Introducción

El objetivo de la separación de fuentes sonoras (SSS) es segregar las fuentes constituyentes de una señal en la que se encuentran mezcladas. La separación de fuentes permite disponer a todo tipo de usuarios, desde amateur hasta profesionales, de las pistas separadas de cada instrumento de una composición, las cuales sólo están disponibles en los estudios de grabación para grabaciones modernas y no existen para grabaciones antiguas. Con ello, la separación de fuentes crece en interés por la cantidad de aplicaciones directas que pueden desarrollarse. Se pueden crear conciertos personalizados de manera que el usuario elija la instrumentación que desea escuchar en cada momento, modificando la elección en el momento que el usuario lo desee. Otro tipo de aplicaciones que se pueden desarrollar a partir de la SSS son las de educación musical. Con la separación de fuentes *offline* se puede eliminar un instrumento de una grabación, de manera que el músico pueda practicar su interpretación siendo acompañado por los demás. Así mismo una versión *online* podría proveer a cada instrumentista de un conjunto musical, su interpretación aislada al finalizar una sesión de ensayo. De esta manera podrían buscar errores interpretativos para perfeccionar su interpretación. Estas fuentes separadas estarían disponibles sin la necesidad de un sistema de grabación en el que cada instrumentista estuviese en una habitación insonorizada y complejos equipos para la grabación. Además de todo ello, la SSS es una fase de preprocesado muy potente para otras aplicaciones de procesado de señal musical, tales como la transcripción musical automática [Gainza07] o *beat tracking* [Chordia09]. El uso de las fuentes separadas simplifica enormemente estas tareas de análisis musical, incluso cuando la separación no tiene demasiada calidad

En función del número de fuentes (número de instrumentos) y sensores (número de micrófonos) que estén presentes en la grabación de la señal, o señales, mezcladas, el problema de la SSS se puede clasificar en tres casos. El caso sobredeterminado es aquel en el que el número de sensores es mayor que el de fuentes [Hyvarinen00]. El caso determinado es en el que el número de fuentes es el mismo que el de sensores. Por último, el caso infradeterminado es en el que el número de fuentes es mayor que el número de sensores. Habitualmente, los casos sobrederterminado y determinado se

suelen abordar con ciertos métodos (por ejemplo *Independent Component Analysis (ICA) or Independent Subspace Analysis (ISA))* que no se pueden aplicar al caso infradeterminado, puesto que su desempeño depende de terner, al menos, el mismo número de señales que de fuentes [Babaie06]. El caso infradeterminado es el caso más común para la mayor parte de las grabaciones musicales existentes (como las grabaciones monoaurales o grabaciones estéreo con una voz y dos o tres instrumentos). Un esquema de desarrollo importante, y ampliamente utilizado en el caso infradeterminado, es *Non-Negative Matrix Factorization (NMF)* [Virtanen06] [Bryan00].

En función del uso, o no, de información previa, la SSS se puede considerar informada *(Informed Source Separation, ISS)*, o no informada *(Blind Source Separation, BSS)* [ComonBook10]. Hasta la fecha, el rendimiento de la separación BSS depende demasiado de la naturaleza de la señal y no obtiene una calidad de separación suficiente para su uso práctico en aplicaciones musicales. En su lugar, para obtener separación de fuentes en el caso infraderminado de calidad es necesario recurrir a la ISS. Hay varios tipos de información que se pueden usar en ISS. Se puede introducir información espectral por medio de modelos de instrumento que sean entrenados *a priori* [Ewert12, Fritsch13, Rodriguez13, Simsekli12]. Así mismo, se puede introducir información simbólica temporal *(score information)* de las notas que se tocan en cada instante [Ewert12, Fritsch13, Duan11, Ganseman10, Hennequin11], la cual debe estar alineada con la señal, o ser alineada como paso previo a su uso en la separación. En este capítulo se trata con el problema de la separación *online* de fuentes armónicas con información temporal y modelos de instrumento de señales monoaurales. Se combina la propuesta de alineamiento de información de *score* de [Duan11] con la propuesta de factorización NMF con modelo de instrumento de filtro-fuente con excitación múltiple de [Carabias11] para la composición del modelo básico en la comparación de las propuestas de este capítulo. De este sistema inicial se parte para su ampliación mediante: 1) la adaptación de los modelos de instrumento, previamente entrenados, a modelos que representen el instrumento real que se toca en la composición que se está separando, 2) el diseño de un algoritmo *online* para la adaptación de modelos y la separación de fuentes. Con todo ello, el sistema final propuesto toma como entrada una señal monoaural con varios instrumentos presentes, unos modelos de instru-

mentos previamente entrenados (con una señal distinta a la de separación) e información de *score*; posteriormente alinea la información de *score* con la señal, adapta los modelos a los instrumentos presentes en la señal y realiza la separación de las fuentes, todo de manera *online*.

En la sección experimental de este capítulo, se demuestra que el uso de modelos de instrumento, y su adaptación a los instrumentos realmente tocados, mejora significativamente el rendimiento del sistema de separación de fuentes. Así mismo, se demuestra que el *online* propuesto obtiene una calidad de separación casi tan buena como la versión *offline*.

7.1.1. Trabajos relacionados

Ewert y Müller [Ewert12] proponen un sistema que iniciativa los modelos de instrumento (patrones espectrales) de distintos instrumentos con un patrón armónico con amplitud decreciente constante y adapta, de manera no controlada, estos patrones a los instrumentos tocados. Aunque este sistema no requiere de un entrenamiento previo de los modelos iniciales, la inicialización no es acorde a ciertos instrumentos que cuentan con fluctuaciones de amplitud importantes de ciertos instrumentos, como el clarinete o el fagot.

Fritsch y Plumbley [Fritsch13] presentan un método para la separación de fuentes musicales, que usa información de *score* para la descomposición basada en un esquema NMF. En esta propuesta los modelos de instrumento se inicializan con unos datos obtenidos de un entrenamiento previo sobre unas señales resultantes de la síntesis de ficheros MIDI. Estos modelos se adaptan, de manera no controlada, en el proceso de factorización. La inicialización es mas cercana que en la propuesta de [Ewert12] a la realidad del instrumento, pero tiene una dependencia importante en el sintetizador usado, además de realizar posteriormente una adaptación no controlada.

Las diferencias principales entre el método propuesto en este capítulo y los propuestos en [Ewert12] and [Fritsch13] son las siguientes: 1) los patrones espectrales iniciales son modelos de instrumento que se han entrenado usando señales de instrumentos reales, en lugar de modelos artificiales o modelos entrenados sobre señales sintéticas; 2) la adaptación del modelo propuesto en este capítulo es controlada, de manera que usa sólo la informa-

ción espectral no solapada, lo que hace el sistema más robusto en escenarios con gran nivel de polifonía; y 3) el método propuesto sólo necesita información de la trama actual y tramas pasadas para la separación de la trama actual (naturaleza *online*), mientras que ambas propuestas, [Fritsch13] y [Ewert12], necesitan información de tramas futuras para la separación de la trama actual (naturaleza *offline*).

En este capítulo se usa el término de latencia algorítmica como el retardo existente entre la recepción de la señal y el comienzo del proceso de separación. En el algoritmo propuesto, la latencia algorítmica es de media trama, puesto que se comienza a separar una trama justo al finalizar su recepción. En la práctica, la latencia real depende de la implementación. Sin embargo, la latencia real se puede reducir con el uso de ordenadores más avanzados o programación optimizada, mientras que la latencia algorítmica no puede ser reducida más de lo que está.

Existen algunas propuestas *online* para la separación de fuentes con un esquema NMF [Duan12, Joder12, Simon12]. Sin embargo, [Duan12, Joder12] se han evaluado en aplicaciones de mejora de la calidad de la señal de voz, y su diseño considera la adaptación de una de las fuentes (voz o ruido) mientras que mantiene el modelo de la otra fuente constante. En el método que se propone en este capítulo todos los modelos de las fuentes (instrumentos) se adaptan simultáneamente. El método propuesto en [Simon12] es adecuado para señales multicanal, pero no así para señales monoaurales. La información de mezcla (que representa la información espacial de las grabaciones multicanal) juega un papel muy importante en ese tipo de sistemas de separación. Así mismo, la inicialización aleatoria de los parámetros del modelo, sin ningún otro tipo de información haría muy compilada la distinción entre las distintas fuentes en las señales monoaurales. De hecho, todas esas propuestas [Duan12, Joder12, Simon12] no se han aplicado para la separación de fuentes informada.

Además de la propuesta de [Duan11], hay otros métodos de alineamiento de señal con información de *score* [Dixon05, Cont06, Cont10]. El motivo de la elección del método propuesto en [Duan11] para fusionarlo con la propuesta de separación de fuentes es su diseño para el alineamiento de señales polifónicas con varios instrumentos, entregando información de alineamiento independiente para cada uno de ellos. Otros métodos sólo están

diseñados [Dixon05, Cont06] o evaluados [Cont10] sobre señales polifónicas de un único instrumento, como el piano.

En el apartado 7.2 de este capítulo, se revisan los métodos del estado del arte sobre los que se sustentan las propuestas, que se desarrollan en el apartado 7.3. Los experimentos y la comparación con el estado del arte se describen en el apartado 7.4. Finalmente en el apartado 7.5 se llega ciertas conclusiones, una vez valorados los resultados.

7.2. Antecedentes

En esta apartado se resumen los métodos de la bibliografía que sirven de base sobre la que se construye la propuesta de este capítulo. El objetivo de este capítulo es implementar un sistema *online* de separación de fuentes con información de *score* y modelos adaptativos de instrumento. El sistema se desarrolla sobre un esquema NMF en el cual se inicializalizan las activaciones con la información de *score* y las funciones base con modelos de instrumento previamente entrenados. Este esquema, con las modificaciones precisas, será capaz de actualizar los modelos de instrumento mientras factoriza la señal de entrada, y todo ello de manera *online*.

7.2.1. Alineamiento de *score* y la señal

The audio-score alignment module of the proposed score-informed source separation method was proposed in [Duan11]. It is an online algorithm that aligns a piece of polyphonic music audio played by multiple instruments with its score. The basic idea is to El bloque de alineamiento del método de separación propuesto se describe en [Duan11]. Este bloque consiste en un algoritmo en tiempo real que alinea una interpretación musical polifónica con su información de *score*. La idea de la que parte es ver la interpretación de la música como un camino en un espacio de estados de dos dimensiones. Estas dimensiones son posición de *score* y *tempo*. El espacio de estados se considera continuo y el camino que sigue la interpretación se encuentra oculto. Se pretende deducir el camino correcto a partir de la señal de audio de manera *online*.

Matemáticamente, $\mathbf{y}_n$ es la n-ésima trama temporal de la señal de audio,

y se asocia a una variable bidimensional $\mathbf{s}_n = (x_n, v_n)^T$ en la que x_n es la posición en el *score* (en *beats*), v_n es su *tempo* (en *beats* por minuto, BPM) y T indica transposición de la matriz. Se pretende estimar la posición actual en el *score* x_n a partir de las observaciones actual y anterior $\mathbf{y}_1, \cdots, \mathbf{y}_n$. Se modela este problema con una cadena oculta de Markov (HMM).

Una cadena oculta de Markov tiene dos partes: un modelo de proceso $p(\mathbf{s}_n|\mathbf{s}_{n-1})$ que describe como son las transiciones de un estado a otro del modelo, y un modelo de observación $p(\mathbf{y}_n|\mathbf{s}_n)$ que describe la probabilidad de cada estado para la trama actual $\mathbf{y}_n$. La diferencia entre una cadena de Markov y un modelo de estados finitos es que los estados son variables continuas que pueden tomar infinitos valores.

El proceso $p(\mathbf{s}_n|\mathbf{s}_{n-1})$ se define con dos ecuaciones dinámicas. La posición en el *score* cambia de la trama anterior a la actual en función del *tempo*. El *tempo* avanza de manera aleatoria, o bien se mantiene estable, dependiendo de la posición en la que se encuentre. El modelo de observación $p(\mathbf{y}_n|\mathbf{s}_n)$ se define mediante el modelo de probabilidad *multi-pitch* propuesto en [Duan10]. Para cualquier conjunto de *pitches*, el modelo de probabilidad de estimación *multi-pitch* obtiene la probabilidad de que dicho conjunto cuadre con el espectrograma de una trama concreta. Entonces, dado un estado hipotético, el conjunto de *pitches* que se supone que están presentes en la n-ésima trama temporal viene dado por la información de *score*. Ofreciendo este conjunto de *pitches* al estimador *multi-pitch*, se puede conocer la probabilidad de la presencia de dicho conjunto en la trama de audio. Cuanto mayor sea la probabilidad dada por el estimador, mejor cuadrarán los *pitches* con el espectrogama de la trama de audio. Una vez definidos el modelo de proceso de Markov y el modelo de observación de los datos, [Duan11] usa *particle filtering* para estimar el camino de estados que se obtiene a partir de las observaciones (estimación *multi-pitch* sobre las tramas), de manera *online*

7.2.2. Modelo de múltiple excitación por instrumento (MEI)

El modelo MEI se aplica sobre un esquema NMF. Este esquema NMF pretende descomponer cada espectro de amplitud $\hat{X}(f,t)$ de la señal de mezclada de entrada en una combinación lineal de ciertas funciones base, o

patrones espectrales:

$$X(f,t) \approx \hat{X}(f,t) = \sum_{j=1}^{J} \sum_{n=1}^{N} g_{n,j}(t) b_{n,j}(f) \tag{7.1}$$

donde $b_{n,j}(f)$ es la n-ésima función base para el j-ésimo instrumento y $g_{n,j}(t)$ es la ganancia de dicha función base en la trama t. Cuando se trabaja con sonidos de instrumentos armónicos, cada función base se corresponde con un *pitch* de manera ideal, y la representación de las ganancias se corresponde con la fuerza de activación de cada *pitch*.

El modelo de excitación múltiple propuesto por Carabias et al. [Carabias11] es una extensión del modelo filtro-fuente propuesto en [Virtanen06]. Este modelo tiene como origen el procesado de voz y la síntesis de sonido. En el campo del procesado de voz, la excitación modela el sonido producido por las cuerdas vocales, mientras que el filtro modela el efecto de resonancia del tracto vocal [RabinerBook78]. En la síntesis de sonido, el filtro-fuente colorea [Valimaki06] una señal con gran riqueza espectral para obtener el sonido deseado.

Las funciones base $b_{n,j}(f)$ de la ecuación (7.1) dependen tanto del *pitch* como del instrumento, por lo que contienen gran cantidad de parámetros. Para reducir el número de parámetros de las funciones base, Virtanen y Klapuri [Virtanen06] proponen un modelo en el que cada función base es el producto de una excitación $e_n(f)$, dependiente del *pitch*, y un filtro $h_j(f)$:, dependiente del instrumento

$$b_{n,j}(f) = h_j(f) e_n(f), \quad n = 1, ..., N, j = 1, ..., J. \tag{7.2}$$

$e_n(f)$ codifica la información del *pitch* y $h_j(f)$ representa la respuesta en frecuencia del cuerpo de resonancia del instrumento. Esta propuesta reduce significativamente el número de parámetros del modelo. Sin embargo, dado que una pieza musical puede contener muchos *pitches* diferentes y se precisa de un espectro completo (excitación) para cada *pitch*, aún es grande el número de parámetros.

Para reducir aún más la complejidad del modelo, algunos trabajos [Badeau09, Heittola09, Klapuri10b] introducen el uso de las excitaciones $e_n(f)$ como componentes espectrales de amplitud unidad situadas en las posiciones en

frecuencias múltiplos enteros del *pitch* n. Finalmente se modelan las funciones base como el producto de un filtro, dependiente del instrumento, y una excitación armónica plana, donde cada componente de la excitación es una réplica de la transformada de la ventana de análisis trasladada a la frecuencia correspondiente

$$b_{n,j}(f) = \sum_{m=1}^{M} h_j(mf_0(n))G(f - mf_0(n)) \tag{7.3}$$

donde M es el número de componentes armónicas consideradas, $f_0(n)$ es la frecuencia fundamental del *pitch* n y $G(f)$ es el espectro de amplitud de la ventana de análisis. Se incluye el término $G(f - mf_0(n))$ para representar la restricción armónica del modelo, siendo ésta la representación del espectro de amplitud de la ventana de análisis trasladado a cada posición armónica de $f_0(n)$.

Este modelo de excitación armónica unitaria, junto con el filtro de instrumento (que normalmente tiene una respuesta frecuencial suave), es capaz de representar ciertos instrumentos que tienen una envolvente suave en frecuencia. Sin embargo, la envolvente de algunos instrumentos, como el clarinete, no son suaves y no pueden ser correctamente representados con una excitación plana y un filtro suave. Por ejemplo, el segundo parcial armónico del clarinete tiene, normalmente, muy poca amplitud. Ésto hace imposible la representación de las notas del clarinete con un solo filtro. En el apartado 3.4.2 del Capítulo 3 se pueden ver más detalles sobre esta limitación

Una alternativa interesante es el uso de un modelo con excitación múltiple, propuesto en [Carabias11]. Este modelo define la excitación como una combinación lineal de varios vectores de excitación. El modelo de filtro-fuente de la ecuación (7.2) precisa una excitación distinta para cada *pitch* (y cada instrumento) $e_n(f)$, mientras que el modelo de excitación múltiple solo necesita $I = 2$ excitaciones por instrumento para modelar correctamente los instrumentos de la base de datos de evaluación.

Siguiendo el modelo de excitación múltiple, cada excitación se define como:

$$e_{n,j}(f) = \sum_{m} \left(\sum_{i} w_{i,n,j} v_{i,m,j} \right) G(f - mf_0(n)) \tag{7.4}$$

donde $m = 1, ..., M$ es el número de armónico y $i = 1, ..., I$ es el índice del vector de excitación base siendo $I << N$. $v_{i,m,j}$ es el m-ésimo parcial del i-ésimo vector de excitación para el instrumento j y $w_{i,n,j}$ es el valor de ponderación del i-ésimo vector de excitación para el *pitch* n y el instrumento j. $G\left(f - mf_0(n)\right)$ es una réplica de la trasnformada de la ventana de análisis trasladada a la frecuencia el m-ésimo parcial armónico del *pitch* n.

La clave del modelo MEI is la división de las excitaciones $e_n(f)$ en dos partes: los vectores de excitación base $v_{i,m,j}$ y los pesos de ponderación $w_{i,n,j}$. Se llama vectores de excitación base $v_{i,m,j}$ porque consisten en I vectores de M valores para cada instrumento j. Los pesos son valores escalares que multiplican cada uno de los I vectores de excitación base para que se combinen linealmente y obtener los M parciales de la excitación para el *pitch* n. Todos los parámetros del modelo se resumen en la tabla 7.2.

La amplitud de cada parcial de las excitaciones finales $e_{n,j}(f)$ se obtiene como una combinación lineal de los I vectores de excitación $v_{i,m,j}$ ponderados por $w_{i,n,j}$. Los vectores de excitación son dependientes del instrumento, pero no dependen del *pitch*. Sin embargo, los pesos de ponderación dependen tanto del instrumento como del *pitch*. Finalmente, las funciones base para el modelo MEI se obtienen de la siguiente manera:

$$b_{n,j}(f) = \sum_{i,m} h_j(mf_0(n)) w_{i,n,j} v_{i,m,j} G(f - mf_0(n)) \tag{7.5}$$

Parámetro	**Size**	**Descripción**
$X(f,t)$	$F \times T$	Representación tiempo/frecuencia de la señal de entrada
$\hat{X}(f,t)$	$F \times T$	TRepresentación tiempo/frecuencia de la señal de reconstruida
$g_{n,j}(t)$	$N \times T \times J$	Ganancia para cada *pitch* de cada instrumento en cada trama
$h_j(f)$	$F \times J$	Filtro para cada instrumento
$v_{i,m,j}$	$I \times M \times J$	Vectores de excitación. I vectores con M valores para cada instrumento
$w_{i,n,j}$	$I \times N \times J$	Valor de ponderación de cada vector de excitación en función de *pitch*
$b_{n,j}(f)$	$N \times F \times J$	Funciones base compuestas conforme a la ecuación 7.5
$G\left(f - mf_0(n)\right)$	$F \times 1$	Transformada de la ventana de análisis trasladada a la frecuencia $mf_0(n)$
F	946	Número de índices de frecuencia con resolución de 1/8 de semitono
T	-	Número de tramas de análisis
N	-	Número de *pitches* para cada instrumento (resolución de 1/8 de semitono)
J	-	Número de fuentes (instrumentos)
M	20	Número de parciales considerados para cada excitación
I	2	Número de vectores de excitación para cada instrumento

Tabla 7.2: *Parámetros del modelo de señal MEI y sus tamaños.*

El número de índices de frecuencia se corresponde con las 114 notas MIDI en la resolución de 1/8 de semitono. Esta resolución conlleva 946 índices de frecuencia. Cada vector de excitación tiene 20 parciales y se usan 2 vectores de excitación por instrumento. Por tanto, hay N pares de pesos de ponderación, un par para cada índice de *pitch*. Los pesos de ponderación son valores escalares que multiplican cada a cada vector de excitación y se combinan linealmente para obtener los 20 parciales de la excitación para un *pitch* en concreto. Cada instrumento puede tocar un número distinto de notas. En el caso del piano, puede tocar 88 notas, el fagot puede tocar 37, el violín puede tocar 45, ... Por tanto el número de índices N depende del instrumento y resulta de la división del número de notas que puede tocar el instrumento en la resolución de 1/8 de semitono.

La reducción del número de parámetros por el uso del modelo MEI se puede demostrar con un ejemplo sencillo. Se va a calcular el número total de parámetros con el modelo MEI, con el modelo filtro-fuente básico y con el modelo de excitación armónica plana para el caso del clarinete. El clarinete puede tocar 37 *pitches* en la escala cromática, lo que corresponde a 296 *pitches* ($N = 296$) con la resolución de 1/8 de semitono para la propuesta de este capítulo. El modelo filtro-fuente necesita $946 \times 1 = 946$ parámetros ($F \times J$) para representar el filtro ($h_j(f)$) y $296 \times 20 = 5920$ parámetros ($N \times M$) para los vectores de excitación ($e_n(f)$), lo que supone un total de 6886 parámetros para representar el clarinete con el modelo de filtro fuente. En el caso del modelo con la excitación armónica plana, sólo se necesitan 946 parámetros para representar el filtro, puesto que las excitaciones carecen de información de amplitud. Finalmente, siguiendo la tabla 7.2, el modelo MEI necesita los mismos 946 parámetros para representar el filtro ($h_j(f)$), $2 \times 20 \times 1 = 40$ parámetros ($I \times M \times J$) para representar los vectores de excitación($v_{i,m,j}$) y $2 \times 296 \times 1 = 592$ parámetros ($I \times N \times J$) para representar los pesos de ponderación de los vectores de excitación ($w_{i,n,j}$). Todo esto significa que el modelo MEI precisa de 1578 parámetros para representar el modelo del clarinete.

El modelo más ligero (atendiendo al número de parámetros necesarios) es el de excitación armónica plana, sin embargo esta excitación no es capaz de representar las envolventes espectrales de ciertos instrumentos musicales, como en el caso del clarinete (ver apartado 3.4.2). Por otro lado, el modelo

filtro-fuente y el modelo MEI si son capaces dec representar correctamente el comportamiento espectral de distintos instrumentos. En el ejemplo mostrado, el modelo MEI reduce en un 77 % el número de parámetros para representar el modelo del clarinete en relación al número de parámetros que necesita el modelo filtro-fuente. En resumen, el modelo MEI conserva la flexibilidad del modelo filtro-fuente con un número de parámetros mucho más bajo.

Dado el modelo MEI, el espectro de amplitud de la señal de entrada se puede descomponer sustituyendo la ecuación (7.5) en la ecuación (7.1), de manera que:

$$\hat{X}(f,t) = \sum_{n,m,i,j} g_{n,j}(t) h_j(m f_0(n)) w_{i,n,j} v_{i,m,j} G(f - m f_0(n)). \tag{7.6}$$

Estimación de parámetros NMF

La restricción de no negatividad para los parámetros del modelo de representación de la señal, ha demostrado ser eficiente para las señales de audio [Virtanen06]. De hecho, esta restricción ha sido ampliamente aplicada en SSS [Virtanen06, Ozerov11].

Dado el modelo MEI de la ecuación (7.6), se desea estimar los parámetros de manera que el error de reconstrucción entre el espectrograma de entrada $X(f,t)$ y el espectrograma estimado $\hat{X}(f,t)$ se minimize. La función de β-divergencia [Vincent10, Fevotte11b] se usa en el sistema propuesto en este capítulo como función de coste del error.

En [Lee01], se propone un algoritmo iterativo basado en reglas de actualización iterativas para la obtención de los parámetros que minimicen la función de coste. Bajo esas reglas de actualización $D_\beta(X(f,t)\|\hat{X}(f,t))$ es decreciente en cada iteración y además aseguran la no negatividad de los parámetros (funciones base y ganancias). La regla de actualización multiplicativa para un parámetro θ_l viene dada por la expresión de la derivada parcial $\nabla_{\theta_l} D_\beta$ como el cociente de dos términos positivos $\nabla^-_{\theta_l} D_\beta$ and $\nabla^+_{\theta_l} D_\beta$:

$$\theta_l \leftarrow \theta_l \frac{\nabla^-_{\theta_l} D_\beta(X(f,t)||\hat{X}(f,t))}{\nabla^+_{\theta_l} D_\beta(X(f,t)||\hat{X}(f,t))}. \tag{7.7}$$

La principal ventaja del uso de las reglas multiplicativas derivadas de la ecuación (7.7) es la conservación de la no negatividad de los parámetros, lo que lleva a un algoritmo ampliado de factorización NMF.

7.3. Método propuesto de ISS

En este apartado se describe el método propuesto para la separación de fuentes musicales de manera *online* con información de *score* y modelos de instrumentos adaptativos. El método se presenta en cuatro etapas, de manera que cada una amplia la anterior. En el apartado 7.3.1 se describe el proceso de entrenamiento para la obtención de los modelos de instrumento iniciales. En el apartado 7.3.2 se presenta el algoritmo de separación básico con los modelos de instrumento entrenados previamente sin que éstos sean adaptados durante la factorización. En el apartado 7.3.3 se propone el método de adaptación de los modelos de instrumento a los instrumentos reales de la composición. Finalmente, en el apartado 7.3.4 se propone el algoritmo de separación y adaptación de manera *online*.

7.3.1. Modelado de instrumentos

El modelo descrito en el apartado 7.2.2 requiere estimar las funciones base $b_{n,j}(f)$ para cada nota n y cada instrumento j. Estas funciones base $b_{n,j}(f)$ se aprenden *a priori* usando grabaciones de notas aisladas de los instrumentos que componen cada composición. Para ello se utilizan las grabaciones de la base de datos de RWC descrita en el apartado 4.1.2. Para realizar el entrenamiento se inicializan las ganancias de un sistema NMF con la anotación real de las señales de entrenamiento. Se inicializan con la unidad el *pitch* n que se encuentre activo en cada instante t, y con cero el resto de ellos. El resto de parámetros del modelo MEI (Tabla 7.2) se inicializan con valores aleatorios positivos. A continuación se actualizan iterativamente todos los parámetros, como se describe en el algoritmo 5, hasta que el algoritmo converge. Las ecuaciones de actualización son el resultado de aplicar la ecuación (7.7) para cada uno de los parámetros del modelo, quedando de la siguiente manera:

$$g_{n,j}(t) \leftarrow g_{n,j}(t) \frac{\sum\limits_{f,m,i} w_{i,n,j} v_{i,m,j} h_j(f) X(f,t) \hat{X}(f,t)^{\beta-2} G(f - m f_0(n))}{\sum\limits_{f,m,i} w_{i,n,j} v_{i,m,j} h_j(f) \hat{X}(f,t)^{\beta-1} G(f - m f_0(n))}, \tag{7.8}$$

$$h_j(f) \leftarrow h_j(f) \frac{\sum\limits_{t,m,n,i} w_{i,n,j} v_{i,m,j} X(f,t) \hat{X}(f,t)^{\beta-2} G(f - m f_0(n))}{\sum\limits_{t,m,i} w_{i,n,j} v_{i,m,j} \hat{X}(f,t)^{\beta-1} G(f - m f_0(n))}, \tag{7.9}$$

$$v_{i,m,j} \leftarrow v_{i,m,j} \frac{\sum\limits_{t,f,n} h_j(f) w_{i,n,j} X(f,t) \hat{X}(f,t)^{\beta-2} G(f - m f_0(n))}{\sum\limits_{t,f,n} h_j(f) w_{i,n,j} \hat{X}(f,t)^{\beta-1} G(f - m f_0(n))}, \tag{7.10}$$

$$w_{i,n,j} \leftarrow w_{i,n,j} \frac{\sum\limits_{t,f,m} h_j(f) v_{i,m,j} X(f,t) \hat{X}(f,t)^{\beta-2} G(f - m f_0(n))}{\sum\limits_{t,f,m,i} h_j(f) v_{i,m,j} \hat{X}(f,t)^{\beta-1} G(f - m f_0(n))}. \tag{7.11}$$

Una vez que los parámetros del modelo MEI se estiman para cada instrumento con el esquema de factorización NMF, se calculan las funciones espectrales base $b_{n,j}(f)$ para cada *pitch* de cada instrumento. Dado que se usa una resolución frecuencial de 1/8 de semitono, cada índice f representa uno de los rangos de frecuencia resultantes. El proceso de entrenamiento se muestra resumido en el algoritmo 5.

Cada función base $b_{n,j}(f)$, necesarias en la fase de factorización, se obtiene en este proceso de entrenamiento. En la aplicación práctica, los instrumentos entrenados difieren ligeramente de los instrumentos reales de la composición (las señales de entrenamiento no son de los mismos instrumentos físicos que los de las señales de evaluación de la separación). En este capítulo se realiza la separación con los instrumentos entrenados y posteriormente se propone una manera de adaptar los modelos a los instrumentos reales, para mejorar la calidad de la separación.

Algoritmo 5 Descripción del algoritmo de entrenamiento

1 Se calcula $X(f,t)$ de una interpretación aislada de cada nota para cada instrumento de la base de datos.
2 Se inicializan las ganancias $g_{n,j}(t)$ con la información de transcripción real de la base de datos y el resto de parámetros del modelo, $h_j(f)$, $v_{i,m,j}$ and $w_{i,n,j}$, con valores aleatorios positivos.
3 Se actualiza el filtro-fuente $h_j(f)$ siguiendo la ecuación(7.9).
4 Se actualizan los vectores de excitación $v_{i,m,j}$ siguiendo la ecuación (7.10).
5 Se actualizan los pesos de ponderación $w_{i,n,j}$ siguiendo la ecuación(7.11).
6 Se actualizan las ganancias $g_{n,j}(t)$ siguiendo la ecuación (7.8).
7 Se repiten los pasos 3,4,5 y6 hasta que el algoritmo converge (o se alcanza el número máximo de iteraciones permitidas.
8 Se calculan las funciones $b_{n,j}(f)$ para cada instrumento j con la ecuación (7.5).

7.3.2. Separación con modelos de instrumento iniciales fijos

En este apartado se describe el algoritmo de separación básico. En este caso el esquema de factorización NMF se compone de dos parámetros, las ganancias $g_{n,j}(t)$ y las funciones base $b_{n,j}(f)$, como se describe en la ecuación (7.1). El algoritmo 6 muestra el proceso de separación con los modelos de instrumentos iniciales fijos.

Algoritmo 6 Algoritmo de separación de fuentes con modelos de instrumento fijos

1 Se calcula el espectrograma $X(f,t)$ de la señal de entrada
2 Se inicializan las ganancias $g_{n,j}(t)$ con la anotación real de la información de *score* y las funciones base $b_{n,j}(f)$ con las obtenidas del proceso de alineamiento del algoritmo 5.
3 **for** C iteraciones **do**
4 Se actualizan las ganancias $g_{n,j}(t)$ siguiendo la ecuación (7.8)
5 **end for**

Se usa la información MIDI (*score*) alineada para inicializar las ganancias$g_{n,j}(t)$

para la factorización NMF. Se da un valor aleatorio positivo a las posiciones tiempo/frecuencia en las que la información del fichero MIDI indica que determinados *pitches* n del instrumento j están activos en cada trama t. Las posiciones asociadas a los *pitches* no activos se inicializan a cero. El proceso de alineamiento de la información de *score* y la señal de entrada se realiza como se indica en el apartado 7.2.1, siguiendo la propuesta de [Duan11].

Una vez que las ganancias $g_{n,j}(t)$ se han inicializado, y usando los modelos de instrumento $b_{n,j}(f)$ previamente entrenados (algoritmo 5), se ejecuta un algoritmo NMF iterativo (algoritmo 6) para la factorización de la señal. Este algoritmo actualiza las ganancias $g_{n,j}(t)$ con la ecuación (7.8), mientras que las funciones base $b_{n,j}(f)$ se mantienen fijas. Una vez finalizado el proceso de factorización, se realiza la separación de las señales tal y como se describe en el apartado 7.3.5.

Aunque la factorización NMF es intrínescamente *offline*, cuando las funciones base $b_{n,j}(f)$ se mantienen fijas, el único parámetro que se actualiza son las ganancias $g_{n,j}(t)$. Para ello se usa la ecuación (7.8), en la que las ganancias de la trama t sólo dependen de la señal de entrada $X(f,t)$ en la trama t, por tanto el algoritmo se convierte en *online*. Otra interpretación es que el proceso de factorización sólo tiene que calcular las ganancias de la trama t, asociadas a cada función base fija, que minimizan la divergencia entre la señal de entrada y la señal reconstruida, ambas en la trama t. Desde otro punto de vista [Kim08], este caso deja de ser un problema NMF para convertirse en un problema NNLS, el cual permite implementar algoritmos de baja complejidad, pero sólo usando la distancia Euclidea como medida de distorsión ($\beta = 2$).

En este capítulo se usa la función de β-divergencia, donde β se mueve en el rango $\beta = [0, 2]$. En [Carabias11] las ecuaciones de actualización se desarrollan sólo para la divergencia de Kullback-Leibler, que se corresponde con el caso de $\beta = 1$. En [Fritsch13] y [Hennequin11] se usa la función de β-divergencia pero sólo usan el caso de $\beta = 1$, por lo que no hacen uso de su potencialidad. El uso de la β-divergencia como medida de distorsión aporta flexibilidad al sistema y permite estudiar el comportamiento del sistema para distintos valores del rango $\beta = [0, 2]$, incluyendo los casos particulares de distancia Euclidea o $\beta = 2$, divergencia Kullback-Leibler o $\beta = 1$ y divergencia Itakura-Saito o $\beta = 0$. En el apartado 7.4 se muestra

un estudio del rendimiento del sistema de factorización propuesto, aplicado a separación de fuentes, en función del parámetro β.

7.3.3. Modelos de instrumento adaptativos

Los modelos de instrumento previamente entrenados son una aproximación a los instrumentos reales de la señal de entrada que se pretende separar. Sin embargo, a pesar de ser el mismo tipo de instrumento, con la misma estructura y construcción, siempre existen diferencias que afectan directamente a las características espectrales de la señal que generan. Estas diferencias pueden deberse a la diferencia de fabricante, materiales o incluso al intérprete que ha realizado la grabación o la sala en la que se ha grabado.

La manera ideal de obtener un modelo de instrumento fiel al que se ha tocado, y se pretende separar, sería realizar el entrenamiento con la señal aislada de dicho instrumento y el mimo músico. Ésto es algo que muy pocas veces se podrá realizar en situaciones reales. Por ello, se propone en este apartado la adaptación de los modelos aproximados (aprendidos de señales de otro instrumento físico) a los instrumentos que realmente son interpretados en la señal que se desea separar, durante el mismo proceso de factorización. De esta manera los modelos se adecuan mejor a las características espectrales de los notas que se van a presentar en la factorización, las funciones base se parecerán más a los elementos que componen el espectrograma y los resultados de separación serán mejores. Esta es la propuesta principal de este capítulo y demuestra que mejor será la calidad de la separación cuanto más se ajuste el modelo de instrumento (funciones base $b_{n,j}(f)$) a la señal que se desea separar.

Para realizar la adaptación de los modelos de instrumento se inicializan las ganancias $g_{n,j}(t)$ con la información de *score* alineada, de la misma manera que en el apartado 7.3.2. La adaptación de modelos se realizó también en [Carabias11], sin el uso de información de *score*. En un escenario en el que están presentes varios instrumentos, la adaptación de parámetros no controlada sufre de las interferencias que se provocan en los parciales solapados entre ambos instrumentos. Este problema puede solventarse cuando se dispone de información de *score*. Adicionalmente en [Carabias11] se demuestra que los resultados varían en función de la combinación de parámetros que

tengan la libertad de adaptación. En concreto, cuando se de libertad de adaptación a los pesos de ponderación $w_{i,n,j}$ el modelo se degrada por la gran cantidad de parámetros que contiene $w_{i,n,j}$ para ser adaptados. Sin embargo, los vectores de excitación $v_{i,m,j}$ y el filtro de instrumento $h_j(f)$ contienen menor número de parámetros y su adaptación no degrada el modelo de instrumento, como se puede ver en [Carabias11].

En el sistema de separación propuesto, se cuenta con la información de *score* alineada por un bloque al efecto. Sin embargo, este bloque puede cometer posibles errores que tendrán consecuencias en los resultados de separación. Estos errores de alineamiento provocarán que en el proceso de adaptación del modelo, éste se actualice con datos erróneos de las notas que están siendo tocadas en ciertos instantes. Por ello, para evitar un gran efecto de los errores de alineamiento sobre la adaptación de los modelos, se propone mantener fijos los pesos de ponderación de las excitaciones $w_{i,n,j}$. Este parámetro ha demostrado ser el más sensible del modelo [Carabias11], por contener un mayor número de valores. En el ejemplo del modelo del clarinete (apartado 3.4.2) los pesos de ponderación tienen 592 valores, mientras que los vectores de excitación suponen sólo 40 valores. La actualización del filtro de instrumento $h_j(f)$ y los vectores de excitación $v_{i,m,j}$, dejando fijos los pesos de ponderación $w_{i,n,j}$ permite ajustar el modelo a la vez que se mantiene robusto frente a posibles errores de alineamiento. Además de esta justificación, los resultados preliminares han demostrado este mismo razonamiento.

Con todo ello, en cada iteración del algoritmo NMF, se actualizan tres parámetros del modelo de señal de la ecuación (7.6): las ganancias $g_{n,j}(t)$, los filtros de instrumento $h_j(f)$ y los vectores de excitación $v_{i,m,j}$. En cada iteración las ganancias se actualizan con la ecuación (7.8), de la misma manera que en el apartado 7.3.2. Sin embargo la actualización de los parámetros relacionados con el modelo de instrumento $h_j(f)$ y $v_{i,m,j}$ no pueden actualizarse de manera no controlada con las mimas ecuaciones que en la fase de entrenamiento. En dicha fase, las notas se presentan de manera aislada, y sólo una nota es tocada en cada instante, por lo que no existe el problema del solapamiento de algunos de sus parciales con los de otra nota. Este hecho sí que ocurre en el caso de la señal que se pretende separar, por lo que parte de la información espectral se corrompe por dichas interferencias

de parciales. Según sea la relación relativa entre las fases de dichos parciales solapados, la interferencia puede ser constructiva o destructiva. En consecuencia, la adaptación de los parámetros del modelo se debe implementar sin incluir ese tipo de información corrupta de los parciales solapados.

Para determinar cuales son los parciales que se consideran solapados, para cada instrumento se detectan las zonas tiempo/frecuencia $\{f', t', j'\}$ que consideran solapadas con alguno de los otros instrumentos con la ayuda de la información de *score* alineada. Inmediatamente después de inicializar las ganancias con la información de *score*, la señal estimada para cada instrumento j' se puede calcular como:

$$\hat{X}_j(f,t) = \sum_{n,m,i} g_{n,j}(t) h_j(m f_0(n)) w_{i,n,j} v_{i,m,j} G(f - m f_0(n)). \qquad (7.12)$$

Para seleccionar las zonas tiempo/frecuencia solapadas para cada instrumento $\{f', t', j'\}$, se realiza una estimación de energía por instrumento con la ecuación (7.12). Cuando la estimación de energía para el instrumento j' no predomina en una localización tiempo/frecuencia (t', f'), se agrega dicho punto al conjunto de puntos solapados $\{f', t', j'\}$. Las zonas tiempo/frecuencia solapadas $\{f', t', j'\}$ para el instrumento j' son aquellos puntos que cumplen la siguiente condición $\frac{\hat{X}_{j'}(f,t)}{\sum_{j=1, j\neq j'}^{J} \hat{X}_j(f,t)} < 10$. Se realiza la estimación para todos los instrumentos.

Una vez que se han estimado los puntos tiempo/frecuencia considerados como solapados, se puede llevar a cabo la adaptación de los modelos de instrumento. En la fase de entrenamiento se actualiza el filtro $h_j(f)$ y los vectores de excitación $v_{i,m,j}$ con todos los puntos de los ejes temporal y de frecuencia. En esta ocasión las ecuaciones de actualización (7.9) y (7.10) se calculan para conjuntos de puntos tiempo/frecuencia diferentes para cada instrumento, cada uno tiene zonas solapadas $\{f', t', j'\}$ en las que no se actualiza, puesto que la información espectral no es fiel a la propia del instrumento. En el algoritmo de actualización, para el instrumento j', la señal de entrada $X(f,t)$ y la señal reconstruida $\hat{X}(f,t)$ se establecen nulas en las posiciones estimadas como solapadas$\{f', t', j'\}$. De esta manera los modelos de instrumento se actualizan y adaptan sin la información corrupta

de los parciales solapados. Todos estos cálculos se resumen en el algoritmo 7

Algoritmo 7 Algoritmo de separación *offline* con modelos de instrumento adaptativos

1 Se inicializan las ganancias $g_{n,j}(t)$ con la información de *score*.
2 Se inicializan los parámetros $h_j(f)$, $v_{i,m,j}$ y $w_{i,n,j}$ con los parámetros entrenados en el algoritmo 1.
3 Se identifican las zonas tiempo/frecuencias solapdas $\{f', t', j'\}$ que cumplen la condición $\frac{\hat{X}_{j'}(f,t)}{\sum_{j=1,j\neq j'}^{J} \hat{X}_j(f,t)} < 10$.
4 **for** C iteraciones **do**
5 Se actualizan las ganancias $g_{n,j}(t)$ con la ecuación (7.8).
6 Se actualizan los filtros $h_j(f)$ y los vectores de excitación $v_{i,m,j}$ con la ecuación (7.9) y la ecuación (7.10) excluyendo las zonas tiempo/frecuencia solapadas $\{f', t', j'\}$ para cada instrumento (se anulan las zonas solapadas en $X(f,t)$ y $\hat{X}(f,t)$).
7 Se calculan las funciones base $b_{n,j}(f)$ para cada instrumento j con la ecuación (7.5).
8 **end for**

El algoritmo 7 actualiza iterativamente los modelos de instrumento para adaptarlos a los instrumentos reales a la vez que factoriza la señal de entrada, en cada iteración con modelos mas fieles a la realidad. Esta actualización conduce a la obtención de mejores resultados de separación de fuentes, frente a los resultados obtenidos con los modelos iniciales, como se muestra en el apartado 7.4.4. A pesar de estos resultados, el algoritmo descrito precisa de la señal de entrada completa para la actualización de los modelos, por lo que no puede ser ejecutado en aplicaciones *online*, que deberían obtener los datos de salida conforme llegan los datos de entrada. En el siguiente apartado se modifica este algoritmo para que pueda trabajar en escenarios *online*.

7.3.4. Adaptación *online* de los modelos de instrumento

El proceso de separación de fuentes se considera *online* cuando, en cada trama, las fuentes separadas se pueden estimar con información exclusivamente de la trama actual y las tramas anteriores. En consecuencia con esta manera de utilizar la información, los modelos de instrumento utilizados para la separación de cada trama se deben obtener sólo con la información desde la primera trama hasta la trama que se encuentra en proceso de separación. A causa de ello, los modelos de instrumento que se usan para las primeras tramas de la señal son los que se entrenan previamente, tal y como se describe en el apartado 7.3.1. Mientras el tiempo va transcurriendo, estos modelos iniciales se van adaptando con la información espectral y temporal que va llegando, de esta manera, las tramas de señal posteriores cuentan con unos modelos de instrumento mejor adaptados a los instrumentos que contiene la señal. Esta particularidad hace que las primeras tramas cuenten con modelos que difieren más de los instrumentos que los usados en las últimas tramas. Por tanto, el resultado de separación sufrirá una degradación si se compara con la separación con modelos adaptados de forma *offline*, dado que la versión *offline* cuenta con los modelos de instrumentos adaptados desde la primera trama.

Si se emplean todas las tramas anteriores a la actual (desde el comienzo de la señal) para adaptar el modelo de instrumento, la complejidad computacional de esta adaptación crece con el paso del tiempo. Se propone adaptar los modelos de instrumento sólo con la información espectral de las tramas $t-T_{update}$ hasta $t-1$. De esta manera se mantiene constante el coste computacional de la adaptación. Al finalizar la adaptación de los modelos con las tramas $t - T_{update}$ hasta $t - 1$, se almacena el modelo, que será el modelo inicial cuando se llegue a la trama $t + T_{update}$ para continuar su adaptación progresiva. Se han considerado ventanas de actualización de un segundo ($T_{update} = 1s$), es decir, la adaptación de modelos se realiza cada segundo, con la información que ha entrado en el sistema durante el último segundo y partiendo del modelo adaptado un segundo antes. La adaptación se realiza con las mismas consideraciones del apartado 7.3.3, los parciales solapados se evitan para que las interferencias no afecten en la adaptación de modelos. La propuesta del algoritmo de separación y adaptación *online* se describe

en el Algoritmo 8.

Algoritmo 8 Algoritmo *online* de separación de fuentes con modelos de instrumentos adaptativos

Se inicializan las ganancias $g_{n,j}(t)$ con la información de *score*
Se inicializan los parámetros del modelo $h_j(f)$, $v_{i,m,j}$ y $w_{i,n,j}$ con los modelos de instrumento entrenados previamente
Se identifican las zonas tiempo/frecuencia solapadas $\{f', t', j'\}$ que satisfacen $\frac{\hat{X}_{j'}(f,t)}{\sum_{j=1,j\neq j'}^{J} \hat{X}_j(f,t)} < 10$.
for trama $t = 0$ hasta el número de tramas de duración de la señal completa **do**
  for C iteraciones **do**
    Se actualizan las ganancias $g_{n,j}(t)$ con la ecuación (7.8) (en la trama t)
  end for
  if la trama t es múltiplo de T_{update} **then**
    for C iteraciones **do**
      Se actualizan los filtros $h_j(f)$ y los vectores de excitación $v_{i,m,j}$ con la ecuaciones (7.9) y (7.10) en las zonas tiempo/frecuencia no solapadas $\{f', t', j'\}$ para cada instrumento entre las tramas $[t - T_{update}, t - 1]$ (estableciendo $X(f,t)$ y $\hat{X}(f,t)$ con valor cero en las zonas solapadas).
    end for
    Se calculan las funciones base $b_{n,j}(f)$ para cada instrumento j usando la ecuación (7.5).
  end if
end for

En el algoritmo *online* propuesto, tanto la fase de alineamiento como la de factorización se obtienen sin información futura, por lo tanto el algoritmo es capaz de generar los datos de salida para la trama t tras su recepción, sin necesidad de más información posterior. Se ha usado el parámetro $C = 50$ para el número de iteraciones del algoritmo, se ha establecido este valor con una serie de datos de validación.

7.3.5. Obtención de las señales separadas

Mascaras ideales de Wiener

En este capítulo, las fuentes $s_j(t), j = 1...J$ que componen la señal mezclada $x(t)$ se han sumado de manera lineal, por tanto $x(t) = \sum_{j=1}^{J} s_j(t)$. Si la densidad espectral de potencia de la fuente j en el bin tiempo/frecuencia (f,t) se representa con $|X_j(f,t)|^2, j = 1...J$, entonces, cada fuente idealmente separada $s_j(t)$ se puede estimar de la señal mezclada $x(t)$ usando un filtrado generalizado de Wiener en el dominio de la STFT *(Short Time Fourier Transform)*. El filtro de Wiener $\alpha_{j'}$ representa la contribución relativa de energía de cada fuente respecto de la energía total de la señal $x(t)$. El filtro de Wiener $\alpha_{j'}$ para cada posición tiempo/frecuencia (t,f) se define como:

$$\alpha_{j'}(t,f) = \frac{|X_{j'}(f,t)|^2}{\sum_j |X_j(f,t)|^2} \tag{7.13}$$

donde la estimación del espectrograma de amplitud para cada instrumento $X_{j'}(f,t)$ se calcula siguiendo la ecuación (7.12). La suma de todos los espectrogramas de amplitud estimados $|\hat{X}_{j'}(f,t)|^2$ es el espectrograma de amplitud de la señal de entrada $|X(f,t)|^2$. Entonces, para obtener el espectrograma de amplitud estimado para cada fuente $\hat{X}_{j'}(f,t)$ se usa la ecuación (7.14)

$$\hat{X}_{j'}(f,t) = \sqrt{\alpha_{j'}(t,f)} \cdot X(f,t). \tag{7.14}$$

Finalmente, cada fuente estimada $\hat{s}_{j'}(t)$ se calcula con la suma solapada de la transformada STFT inversa de cada trama de los espectros de amplitud estimados $\hat{X}_{j'}(f,t)$ con la fase que obtiene el espectrograma de la señal de entrada.

Obtención de las señales separadas

Una vez que las ganancias se han estimado con cualquiera de los métodos propuestos, la manera de obtener de la señal estimada separada es siempre la

misma. En primer lugar, el espectrograma de amplitud de la fuente estimada $\hat{X}_j(f,t)$ se calcula de la siguiente manera:

$$\hat{X}_j(f,t) = g_{n,j}(t)b_{n,j}(f). \tag{7.15}$$

Entonces, se calculan las máscaras de Wiener con la ecuación (7.13). Estas máscaras de Wiener se aplican sobre el espectrograma de la señal de entrada $X(f,t)$ siguiendo la ecuación (7.14). Por últimos la fuente estimada $\hat{s}_j(t)$ se calcula con la transformada STFT inversa sobre $\hat{X}_j(f,t)$.

7.4. Experimentos

7.4.1. Datos de entregamiento y evaluación

En la fase de entrenamiento (ver apartado 6.2.5), las funciones base se estiman usando la base de datos *RWC Musical Instrument Sound Database* y todo el rango dinámico para cada instrumento. Se han utilizado cuatro instrumentos en la fase de evaluación (violín, clarinete, saxofón y fagot). Los sonidos de esta base de datos se encuentran disponibles con una resolución de un semitono a lo largo de todo el rango dinámico de cada instrumento. En la base de datos se encuentran interpretaciones de todas las notas con varios estilos interpretativos y matices dinámicos, como se describe en el apartado 4.1.2. Se han usado los ficheros con un estilo normal y un matiz dinámico *mezzo*. Se ha comprobado que el entrenamiento con diferentes estilos lleva a la obtención de distintos modelos de instrumento. Sin embargo, en [Carabias11] se demuestra que la configuración seleccionada (estilo normal y matiz *mezzo*) obtiene un modelo representativo de todos los demás.

Para la fase de evaluación, se usa la base de datos propuesta en [Duan10, Duan11] (ver apartado 4.1.5). Esta base datos contiene 10 corales de cuatro partes de J.S. Bach con su información MIDI correspondiente. Los ficheros de audio tienen una duración aproximada de 30 segundos y las grabaciones reales están muestreadas a $44{,}1KHz$. Cada coral de la base de datos está compuesta por un cuarteto de instrumentos (violín, clarinete, saxofón y fagot) y cada instrumento está grabado en una pista independiente. Estas pistas se mezclan para crear un total de 10 interpretaciones con nivel de polifonía 4, 60 duetos y 40 trios.

7.4.2. Configuración de los experimentos

Representación tiempo/frequencia de las señales

En la bibliografía se encuentran muchos sistemas de procesado de señal que usan una discretización logarítmica de frecuencia. Por ejemplo, en [Bertin10, Vincent10] se usa una división en bandas uniformemente espaciadas en la escala *Equivalent Rectangular Bandwidth (ERB)*. Cuando se usan modelos de instrumento con restricción armónica, la señal reconstruida se calcula con un término derivado de la traslación de la transformada de la ventana de análisis $G(f - mf_0)$ a la frecuencia correspondiente a cada *pitch* mf_0. Este término aparece en la ecuación (7.5) cuando se usa el modelo MEI. Con estas condiciones, el uso de una resolución en frecuencia relacionada con la resolución de *pitch* es aconsejable para facilitar el procesado de la señal. Además, la base de datos de entrenamiento este anotada con una precisión de un semitono, por lo que los datos de referencia ya marcan este tipo de resolución. En este capítulo se ha usado una resolución de 1/8 de semitono en frecuencia. Se ha implementado una representación tiempo/frecuencia con la suma de los bins de la STFT correspondientes a cada intervalo de 1/8 de semitono. Por otro lado, cuando se calcula la separación con las máscaras de Wiener, el mismo valor de máscara se aplica sobre todos los bins que pertenecen al mismo intervalo de 1/8 de semitono.

Model parameters

El tamaño de trama y de salto para el análisis con STFT se ha establecido en 128ms y 32ms respectivamente. Así mismo, el número de iteraciones para los algoritmos NMF es de $C = 50$.

El modelo MEI se calcula con los siguientes parámetros: 20 parciales armónicos para cada función base en los modelos armónicos ($M = 20$); $I = 2$ vectores de excitación; y $J = 4$ instrumentos.

En relación a la resolución de *pitch* hay que indicar que en la fase de entrenamiento se ha usado una resolución de un semitono (la misma que tiene la anotación de la base de datos), mientras que en la fase de separación, las funciones base $b_{n,j}(f)$ se han adaptado a la resolución de 1/8 de semitono mediante la réplica de la envolvente de cada función base a lo largo de

las 8 nuevas posiciones de *pitch* para cada una de la resolución anterior. Los instrumentos reales no sólo producen sonidos en las frecuencias MIDI exactas, pueden efectuar vibrato en torno a ellas. Con esta resolución de 1/8 de semitono se puede captar mejor la variación del *pitch* de los instrumentos reales.

El uso de la β-divergencia como medida de distorsión en el esquema NMF hace necesaria la elección del valor para el parámetro β para la obtención de los mejores resultados posibles. Para encontrar el valor óptimo de este parámetro, se ha realizado un estudio de los resultados de separación con diferentes valores en el rango$[0, 2]$. En la figura 7.1 se aprecia que los mejores resultados se dan con valores entre $1,2$ y $1,6$. El valor óptimo de β se encuentra en torno a $1,5$ y éste ha sido el valor usado para el parámetro en los experimentos realizados.

A pesar de que los valores bajos de β han sido considerados apropiados para otras aplicaciones de procesado de señal [Bertin10, Fevotte09b], en este caso el rendimiento es muy bajo. La razón de esta pérdida de rendimiento con este valor del parámetro β está relacionada con la restricción de armonicidad empleada. Los instrumentos musicales no son idealmente armónicos, y por tanto, generan valores pequeños de energía alrededor de los bins correspondientes a las frecuencias armónicas. Estas componentes no se modelan bien cuando se usan valores bajos de β, ocurriendo lo mismo con el ruido de fondo de la señal. Los valores de energía bajos y el ruido de fondo no afectan a las descomposiciones con valores altos de β porque en estos casos el algoritmo da mayor importancia a los valores grandes de energía. En el caso de $\beta = 0$ las diferencias de amplitud afectan al cálculo de la divergencia por la invarianza de escala [Carabias11].

Medidas de separación de fuentes

Para llevar a cabo una medida objetiva del rendimiento de los algoritmo de separación propuestos, se han usado las medidas implementadas en [Vincent06, BSSEVAL], que se describen con mayor profundidad en el Capítulo 4. Estas medidas han sido adoptadas por la comunidad científica, por tanto facilitan la comparación de resultados con otros algoritmos propuestos en la bibliografía. Se supone que cada señal estimada contiene un modelo

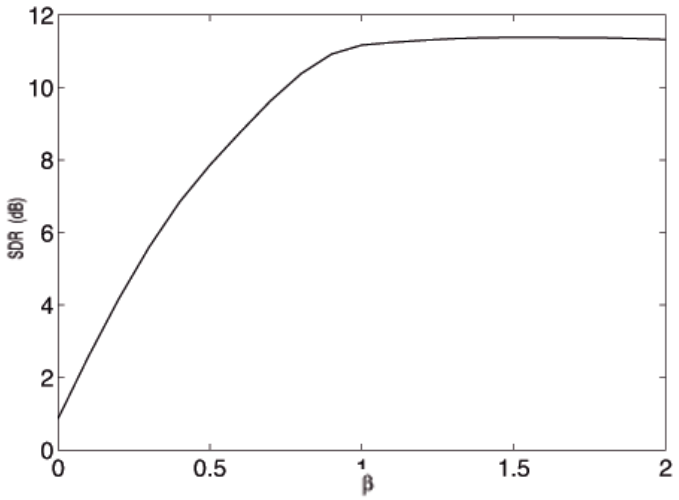


Figura 7.1: *Rendimiento de la separación de fuentes con nivel de polifonía* 2 *para diferentes valores del parámetro* β

de distorsión tal que:

$$\hat{s}_j(t) - s_j(t) = e_j^{target}(t) + e_j^{interf}(t) + e_j^{artif}(t) \tag{7.16}$$

donde $\hat{s}_j$ es la señal estimada para el instrumento j, s_j es la señal original del instrumento j, e^{target} es el término de error asociado a la componente de distorsión, e^{interf} es el término de error asociado a las interferencias de otras fuentes y e^{artif} es el término de error asociado a los artefactos generados por el proceso de separación. Las medidas asociadas a cada término de error son: *Source to Distortion Ratio* (SDR), *Source to Interference Ratio* (SIR), y *Source to Artifacts Ratio* (SAR).

$$SDR_j = 10\log_{10}\frac{\sum_t |s_j(t)|^2}{\sum_t |\hat{s}_j(t) - s_j(t)|^2} \tag{7.17}$$

$$SIR_j = 10\log_{10}\frac{\sum_t \left|s_i(t) + e_j^{target}(t)\right|^2}{\sum_t \left|e_j^{interf}(t)\right|^2} \tag{7.18}$$

$$SAR_j = 10\log_{10}\frac{\sum_t \left| s_i(t) + e_j^{target}(t) + e_j^{interf}(t) \right|^2}{\sum_t \left| e_j^{artif}(t) \right|^2} \tag{7.19}$$

7.4.3. Algoritmos para comparar

Se han comparado diferentes configuraciones del método propuesto y un método de separación con información temporal, propuesto en [Duan11], llamado *Soundprism* y que se ha establecido como base. Este método separa las fuentes usando enmascaramiento armónico y los valores relativos de energía para cada parcial se establecen según su número de parcial, es decir, se consideran siempre unas envolventes exponenciales decrecientes. Es un algoritmo *online* pero sin modelos de instrumento.

El método propuesto se evalúa con tres configuraciones. La configuración *Proposed fixed* se refiere a la versión *online* del método propuesto con modelos de instrumentos fijos (apartado 7.3.2). La configuración *Proposed adaptive offline* se refiere a la versión *offline* del método propuesto con modelos adaptativos de instrumento (apartado 7.3.3). Por último, la configuración *Proposed adaptive online* se refiere a la adaptación de modelos de instrumento de manera *online* (apartado 7.3.4).

Así mismo, se ha comparado con unos datos *Oracle*, que se considera la mejor separación teórica con los métodos de máscaras tiempo/frecuencia y el banco de filtros utilizado. El cálculo de estos valores requiere el uso de las señales de cada instrumento aisladas. Las señales mezcladas se hacen pasar por el banco de filtros con resolución de 1/8 de semitono. A continuación, las señales aisladas se filtran de la misma manera y se usan para obtener las máscaras ideales de Wiener, que son las mejores máscaras de separación con la resolución considerada. Dichas máscaras se aplican sobre la señal mezclada para obtener las señales separadas *Oracle*. Esta separación ofrece los mejores resultados posibles para la configuración que se ha establecido y se considera el techo de las medias que se pueden obtener con los métodos propuestos.

7.4.4. Resultados

Separación con *score* ideal

El método de separación propuesto esta destinado a trabajar en escenarios con información temporal de *score*. Sin embargo, los errores de alineamiento (en caso de no usar el alineamiento ideal) pueden afectar a la calidad de la separación de fuentes. Para separar los efectos producidos por los errores de alineamiento y centrar la atención en los métodos de separación, en primer lugar se usan los datos de alineamiento ideales. Esta información se obtiene con el algoritmo [Yin05] sobre las fuentes aisladas antes de la mezcla, con algunas correcciones manuales, que permiten conocer el *pitch* de cada fuente en cada instante temporal.

La figura 7.2 muestra la comparación de los resultados de 60 duetos. Se puede ver que todos los métodos, incluido el *Oracle* (barra 5), tienen una considerable desviación típica. Este hecho se produce por las distintas dificultades existentes para separar los distintos tipos de instrumentos. Si se comparan con el método base (*Soundprism*, barra 2), el método propuesto, con sus tres configuraciones (barras 3, 4 y 5) mejora considerablemente en las medidas SDR y SIR.

Se ha efectuado una prueba *t-test* sobre los resultados de separación obtenidos para evaluar su relevancia estadística. Los resultados de esta prueba muestran que las mejoras entre unas y otras configuraciones del sistema tienen relevancia estadística ($p < 10^{-6}$). En términos de SAR, el método propuesto con los modelos de instrumento iniciales (Barra 2) no es estadísticamente mejor que *Soundprism* (Barra 1), pero el método propuesto con los modelos adaptativos (Barra 3) si que es estadísticamente mejor que *Soundprism*. Además, el uso de modelos adaptativos de instrumento (Barra 3) mejora estadísticamente las tres medidas (SDR, SIR y SAR) respecto a la separación con los modelos iniciales ($p < 10^{-5}$). El algoritmo *online* con los modelos adaptativos (Barra 4) consigue un rendimiento levemente más bajo que la versión *offline* (Barra 3), sin embargo, aún es estadísticamente mejor que la versión *offline* con los modelos de instrumento iniciales (Barra 2) y el algoritmo *online* sin modelos de instrumento (*Soundprism*, barra 1). Si se compara con los resultados *Oracle* (Barra 5), se puede ver que el algoritmo *online* adaptativo (Barra 4) queda $2,5$ dB por debajo, lo que deja

abierto el camino para poder mejorar su rendimiento.

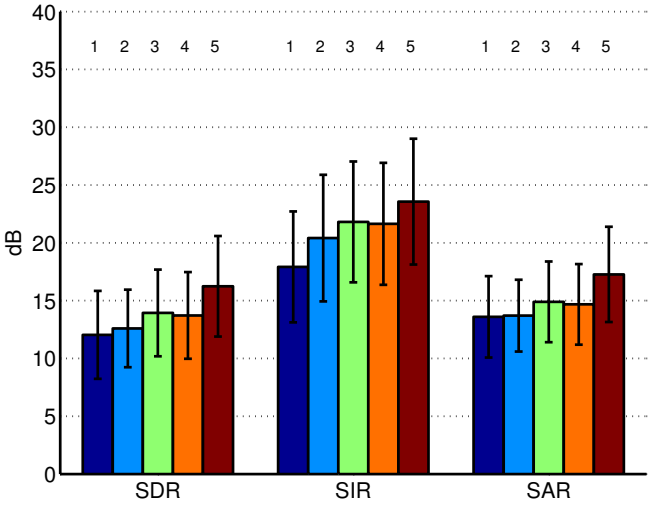


Figura 7.2: *Resultados de separación de fuentes sobre* 60 *duetos usando la información de* score *ideal. Cada barra muestra la media de las* 120 *medidas sobre las* 120 *pistas separadas. La línea vertical sobre cada barra indica el rango de desviación típica de la muestra. Los cinco métodos son: 1)* Soundprism*, 2) Método propuesto con modelos iniciales (apartado 7.3.2), 3) Método propuesto* offline *con modelos adaptativos (apartado 7.3.3), 4) Método propuesto* online *con modelos adaptativos (apartado 7.3.4), y 5)* Oracle

La figura 7.3 muestra la comparación de los resultados de separación en función del nivel de polifonía. Se muestran sólo los resultados de la medida SDR, puesto que la tendencia de las medidas SIR y SAR es similar. Con el incremento del nivel de polifonía, el rendimiento de todos los métodos (incluyendo *Oracle*) desciende significativamente. Al igual que ocurre con los resultados de los duetos en la figura 7.2, el método propuesto en sus tres configuraciones (Barras 2, 3 y 4) mejoran el rendimiento de *Soundprism* (Barra 1) en trios y cuartetos, datos que se confirman con la prueba *t-test* ($p < 10^{-4}$). Además de ello, la mejora del uso de modelos adaptativos (Barras 3 y 4), frente al uso de los modelos iniciales (Barra 2), sigue teniendo

relevancia estadística considerable ($p < 10^{-8}$).

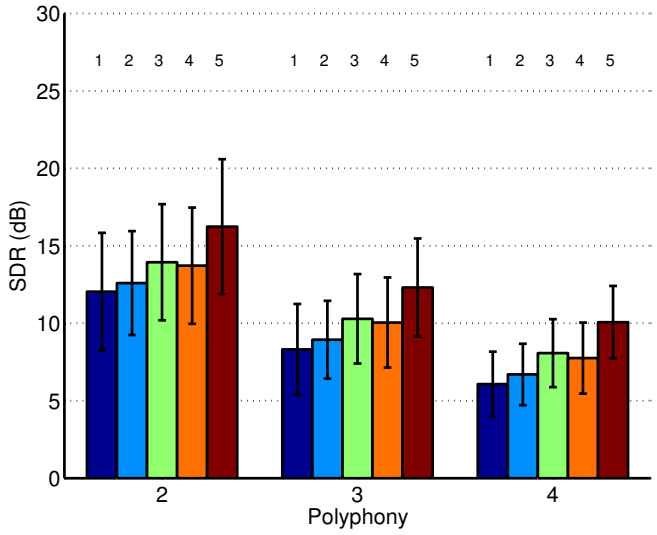


Figura 7.3: *Resultados de separación de fuentes frente al nivel de polifonía, con el uso de información de* score *ideal. Cada barra muestra la media de las* 120 *medidas para duetos,* 120 *medidas para trios, y* 40 *medidas para cuartetos, donde cada medida se calcula para cada pista separada.La línea vertical sobre cada barra indica el rango de desviación típica de la muestra. Los cinco métodos son: 1)* Soundprism*, 2) Método propuesto con modelos iniciales (apartado 7.3.2), 3) Método propuesto* offline *con modelos adaptativos (apartado 7.3.3), 4) Método propuesto* online *con modelos adaptativos (apartado 7.3.4), y 5)* Oracle

Separación con *score* alineado

En este apartado, se comparan los métodos de separación propuestos haciendo uso de la información de *score* alineada a la señal de audio como información de entrada al sistema. Estas pruebas evalúan el método propuesto en una situación más realista. La figura 7.4 muestra los resultados de separación de duetos. De manera similar a lo que ocurre en la figura 7.2, el método propuesto con los modelos adaptativos (Barras 3 y 4) tienen un rendimiento mayor a *Soundprism* (Barra 1) en SDR y SIR, este

dato esta contrastado con la prueba *t-test* ($p < 10^{-3}$). El método propuesto con los modelos iniciales (Barra 2) supera a *Soundprism* (Barra 1) en SIR ($p = 4{,}1 \times 10^{-5}$), pero no así en SDR ($p = 0{,}16$). Esto demuestra, de nuevo, el uso de modelos de instrumento adaptativos para la separación de fuentes musicales. La mejora entre el uso de los modelos iniciales (Barra 2) y los modelos adaptativos (Barras 3 y 4) vuelve a tener importancia estadística en todas las medidas ($p < 10^{-4}$), tanto en la versión *offline* como en la versión *online*, a pesar de que el algoritmo *online* (Barra 4) disminuye levemente el rendimiento respecto de la versión *offline* (Barra 3).

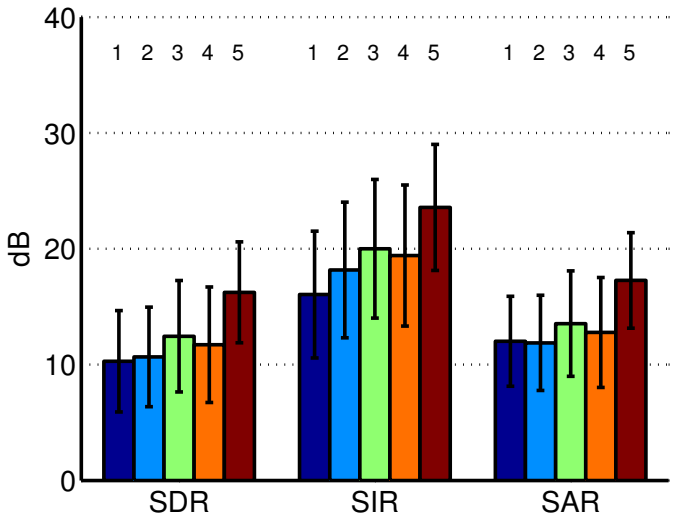


Figura 7.4: *Resultados de separación de fuentes sobre* 60 *duetos usando la información de* score *alineada. Cada barra muestra la media de las* 120 *medidas sobre las* 120 *pistas separadas. La línea vertical sobre cada barra indica el rango de desviación típica de la muestra. Los cinco métodos son: 1)* Soundprism*, 2) Método propuesto con modelos iniciales (apartado 7.3.2), 3) Método propuesto* offline *con modelos adaptativos (apartado 7.3.3), 4) Método propuesto* online *con modelos adaptativos (apartado 7.3.4), y 5)* Oracle

A su vez, la figura 7.5 muestra los resultados de separación con diferentes niveles de polifonía. De la misma manera que en la figura 7.3, sólo se

muestran los valores de SDR, puesto que los valores de SIR y SAR siguen la misma tendencia. De nuevo, el método propuesto, en sus tres configuraciones (Barras 2, 3 y 4) obtienen mejores resultados que *Soundprism* (Barra 1) para todos los niveles de polifonía, esta mejora está contrastada con el valor de la prueba *t-test* ($p < 10^{-4}$). La mejora de los resultados con modelos adaptativos (Barras 3 y 4) frente a los resultados con modelos iniciales (Barra 2) también es estadísticamente significativa para todos los niveles de polifonía ($p < 10^{-5}$). Estos resultados vuelven a mostrar los beneficios del uso de modelos de instrumento, preferentemente adaptativos, en los sistemas de separación de fuentes informada. Además muestran que la propuesta del algoritmo *online* también cumple con las expectativas y el método de separación es apto para una aplicación realista.

Si estos resultados se comparan los obtenidos con la información de *score* ideal de la figura 7.3, hay dos observaciones interesantes. En primer lugar, la media de la medida SDR de todos los métodos de la figura 7.5 (exceptuando *Oracle*) descienden, mientras que la desviación típica se incrementa. Este efecto se produce por los errores de alineamiento del sistema. En segundo lugar, con el incremento del nivel de polifonía, la degradación va siendo menos significativa para casi todos los métodos. Esto se puede explicar también por el rendimiento del sistema de alineamiento, el cual produce mejor alineamiento cuando el nivel de polifonía crece, por tener más información para alinear. Para el conjunto de señales de evaluación, el acierto del sistema de alineamiento es mejor en las piezas con mayor polifonía [Duan11].

7.5. Conclusiones

En este capítulo se propone un sistema de separación de fuentes informado. Este sistema hace uso de modelos que describen el comportamiento espectral de los instrumentos involucrados. Se han comparado distintas configuraciones de este sistema, un método del estado del arte (*Soundprism*) como base y la separación *Oracle*.

El método propuesto, al contrario que la mayoría de las propuestas de la bibliografía, hace uso de modelos de instrumento paramétricos que se aprenden de interpretaciones de instrumentos reales. Así mismo, estos modelos iniciales se actualizan y perfilan a los instrumentos que son tocados en la

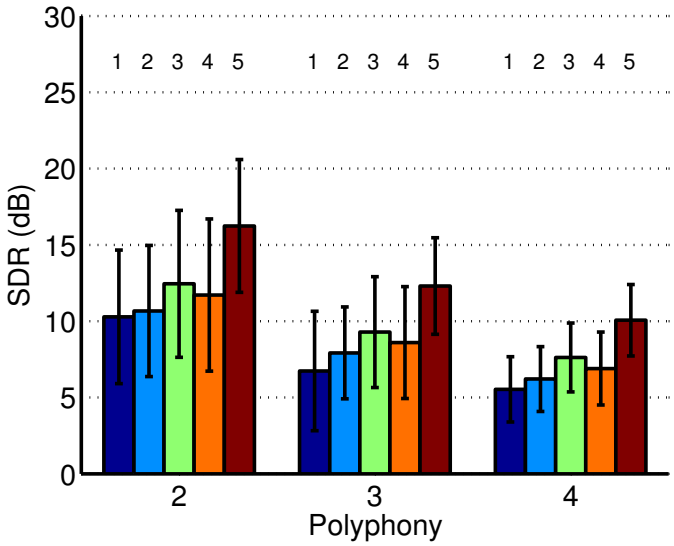


Figura 7.5: *Resultados de separación de fuentes en función del nivel de polifonía, usando la información de* score *alineada.Cada barra muestra la media de las* 120 *medidas para duetos,* 120 *medidas para trios, y* 40 *medidas para cuartetos, donde cada medida se calcula para cada pista separada.La línea vertical sobre cada barra indica el rango de desviación típica de la muestra. Los cinco métodos son: 1)* Soundprism*, 2) Método propuesto con modelos iniciales (apartado 7.3.2), 3) Método propuesto* offline *con modelos adaptativos (apartado 7.3.3), 4) Método propuesto* online *con modelos adaptativos (apartado 7.3.4), y 5)* Oracle

composición que se desea separar. Esta actualización es controlada, gracias a la información de *score* que permite obtener los parciales y tiempos en los que la información espectral no se corrompe a cause de solapamientos de parciales entre notas simultáneas. Además, esta actualización puede llevarse a cabo durante el proceso de separación, y todo ello sin información futura, es decir, de manera *online*.

Se ha evaluado el método propuesto con una base de datos musical del estado del arte y comparado con otro sistema que no hace uso de modelos de instrumento. Los experimentos muestran que el uso de modelos de instrumento mejora la calidad de la separación. También la actualización de

los modelos a los instrumentos empleados, mejora significativamente la separación, a la vista de los resultados anteriores. Así mismo, la versión *online* del algoritmo de separación y adaptación de los modelos obtiene unos resultados que no distan mucho de los obtenidos en la versión *offline*. Todos los resultados disminuyen cuando crece el nivel de polifonía, aunque con dicho crecimiento la pérdida de calidad es cada vez menor, debido a la mejora del proceso de alineamiento cuando hay mayor número de instrumentos.

Parte IV

Conclusiones y líneas futuras

Capítulo 8

Conclusiones y líneas futuras

8.1. Puntos relevantes

- Existen una gran variedad de aplicaciones de la separación de fuentes musicales. Por ejemplo, la generación de conciertos a medida para el oyente, pudiendo seleccionar qué parte de la orquesta desea escuchar en cada momento, o aplicaciones musicales educativas, que permitan examinar las interpretaciones de manera exhaustiva, permitiendo analizar el desempeño de uno o varios músicos en particular. Las grabaciones independientes de los instrumentos de una interpretación no suelen estar disponibles para el público en general, sin embargo, los sistemas de separación de fuentes permiten a los usuarios trabajar con las pistas separadas para cualquiera que sea su objetivo.

- La separación de fuentes a ciegas sobre señales monocanal sin información adicional no proporciona resultados útiles. La falta de información espacial debe suplirse con la aportación de información espectral y/o temporal, como pueden ser los modelos espectrales de instrumento e información de *score* o partitura. Esta información y otras restricciones hacen que los resultados de la separación de fuentes sobre señales monocanal obtenga resultados prometedores, a la vez que certifica ciertas estrategias que pueden ser extrapoladas a la separación multicanal.

- Existen una gran variedad de esquemas de factorización de señal. Una primera clasificación puede separar los métodos deterministas y los estadísticos. Resulta más sencillo aplicar restricciones al proceso de factorización cuando se emplea un método determinista que estadístico, siendo para el primero restricciones prácticas y en el segundo caso se pueden considerar restricciones más teóricas, en función del tipo de señal que se está tratando.

- Un método ampliamente utilizado para la factorización de señales es NMF. Su estructura y flexibilidad, en su versión determinista, permite agregarle las restricciones e información necesarias para la separación de fuentes musicales en señales monocanal. Por estas razones y por el conocimiento del método ha sido usado como herramienta básica de factorización en todos los capítulos de esta tesis.

- La cuestión básica de la separación de fuentes musicales es decidir a qué instrumento pertenece cierta parte de energía de la señal que se sitúa en una determinada posición tiempo/frecuencia. Si en dicha posición sólo existe energía de un instrumento, se precisan mecanismos que, en base a ciertos análisis de la señal, discriminen la fuente de la que procede. En cambio, si la energía en una posición tiempo frecuencia se debe a la presencia de más de un instrumento, hay que articular otros mecanismos que estimen la cantidad de energía que pertenece a cada uno de los instrumentos. Además, si se pretende ser riguroso en la separación de este último caso, denominado solapamiento de parciales armónicos, también es preciso estimar la fase instantánea del fasor que representa la aportación de energía de cada instrumento para su síntesis posterior.

- No existe una base de datos estandarizada para evaluar los sistemas de separación de fuentes musicales. Sin embargo, si que existen determinadas bases de datos que son usadas masivamente por la comunidad científica para la evaluación de este tipo de sistemas, por tanto son bases de datos estandarizadas de facto. En el Capítulo 4 se describen las bases de datos utilizadas en esta tesis para la evaluación y comparación de los sistemas propuestos. El uso de las mismas bases de datos

permite comparar de manera sencilla el rendimiento de unos sistemas con otros. Además, existen librerías para la medida de la calidad de la separación que vienen siendo empleadas por la mayoría de los investigadores de este campo, hecho que facilita también la comparación de los sistemas.

8.2. Contribuciones de la tesis

Se debe remarcar que esta tesis se centra en la separación de fuentes musicales sobre señales monocanal y hace un análisis de la información adicional que se puede utilizar, así como las mejoras que supone su uso. Con el uso de la información adicional se pretende obtener una separación de las fuentes que obtenga la menor interferencia posible entre ellas.

8.2.1. Modelos de instrumento en SSS

En esta tesis se han propuesto los modelos espectrales de instrumento como información importante a la hora de desarrollar la separación de fuentes. Estos modelos de instrumento tienen como objetivo poder discriminar la pertenencia de la energía de la señal a un instrumento en concreto. Además, estos modelos han sido empleados para la separación de parciales solapados, puesto que la amplitud del parcial solapado se ha estimado gracias a la información no solapada de otros parciales. Estos datos se relacionan gracias a los modelos de instrumento. En el capítulo 7 se ha comparado el sistema de separación propuesto con modelos de instrumento frente a otro sistema del estado del arte que no los considera. Se ha demostrado que el uso de modelos de instrumento produce mejoras sustanciales en la separación de fuentes.

En el capitulo 5 se propone el sistema inicial de fuentes con modelos de instrumento. Estos modelos se han entrenado previamente y permanecen constantes durante el proceso de factorización. Este capítulo tiene como objetivo analizar el rendimiento de varios modelos de instrumento en la separación de fuentes, obteniéndose varias conclusiones. El modelo BHC, el más flexible de todos por no tener la restricción del filtro-fuente. Tiene un rendimiento similar al modelo MEI para $I = 1, 2$, no siendo así cuando

$I > 2$. Su gran flexibilidad le hace alcanzar buenos valores de separación, pero el modelo MEI con varias excitaciones se hace más robusto ante las diferencias de los instrumentos de entrenamiento y evaluación. El uso de filtro-fuente le aporta robustez y su carácter paramétrico reduce mucho el número de parámetros del modelo, lo que se hace adecuado para una posible adaptación. El modelo HCE, que cuenta con un sólo parámetro, el filtro, no tiene la flexibilidad suficiente para representar de manera correcta todo el rango de notas de los instrumentos, por lo que sus resultados quedan lejos de los de los otros dos modelos (BHC y MEI).

Tradicionalmente los modelos de instrumento se obtienen de una fase previa de entrenamiento informado. En el capítulo 7 se propone la mejora de los modelos de instrumento iniciales mediante una actualización de los mismos en la fase de factorización de la señal de entrada. La adaptación de los modelos es controlada y se precisa de modelos paramétricos para evitar la inestabilidad de los mismos en el proceso de adaptación. Teniendo en cuenta las conclusiones a las que se llega en el capítulo 5, se ha usado el modelo MEI, un modelo paramétrico y flexible que puede ser adaptado. Se ha demostrado que la adaptación de los modelos de instrumento, desde los modelos inicialmente entrenados hasta unos modelos más cercanos a los ideales, es posible de manera controlada. Esta adaptación debe realizarse con información que no esté corrompida por interferencia de otros instrumentos y sólo sobre los parámetros del modelo menos cuantiosos y sensibles a errores (filtro y excitaciones). Esta adaptación controlada de los modelos produce una mejora apreciable en la calidad de separación de las señales y un mayor aislamiento de las fuentes, produciéndose menos interferencias entre ellas.

8.2.2. Información temporal en SSS

El uso de información previa sobre las fuentes presentes en la señal a separar ha demostrado ser útil en la separación de fuentes sobre señales monocanal. Se han tratado dos tipos de información, la información espectral (modelos de instrumento) y la información temporal de la interpretación (*score* o partitura). Al igual que la información espectral, la información temporal o de *score* se constituye como una información fundamental y de

gran ayuda para la separación de fuentes en señales monocanal así como para el proceso de adaptación de los modelos de instrumento.

Uno de los objetivos principales del capitulo 7 es la adaptación de los modelos de instrumento entrenados a los instrumentos reales de la señal de entrada. Esta adaptación debe ser controlada, puesto que la información de parciales solapados entre las notas de distintos instrumentos no debe ser tenida en cuenta. Para conocer cuando se producen estas situaciones de solapamiento de parciales es necesario conocer el *pitch* que está tocando cada instrumento en cada instante. Una vez que se cuenta con esta información temporal, la misma se puede emplear para la inicialización de la matriz de ganancias del sistema de factorización. Con esta inicialización sólo se tendrán que estimar las ganancias relacionadas con los *pitches* que se encuentren activos, lo que revierte en una mejora en la calidad de la separación realizada.

La información temporal de *score* suele encontrarse en formato MIDI y no necesariamente con la misma temporización que la señal de entrada al sistema. Este hecho hace que sea necesario un proceso de alineamiento que puede introducir error en el sistema. En el capitulo 7 se ha demostrado que si se cuenta con un alineador fiable, los posibles errores de sus datos de salida no afectan, en gran medida, a la robustez del sistema de adaptación de los modelos de instrumento ni a la separación de las fuentes. Este buen desempeño se ha demostrado comparando los resultados obtenidos al incluir la fase de alineamiento, frente a los resultados de la separación usando la información de *score* ideal.

8.2.3. Algoritmos de factorización

NMF es un modelo de factorización de señal muy flexible, ampliamente utilizado en el campo de aplicación y cuenta con diversos algoritmos para su implementación. En esta tesis se ha empleado el algoritmo NMF determinístico, el cual admite la inclusión de restricciones, información de inicialización y la modificación de sus matrices a lo largo de las iteraciones (adaptación de modelos de instrumento) de manera sencilla. En esta tesis se ha empleado este algoritmo NMF para el sistema de separación de fuentes con modelos de instrumento del capítulo 5, para el sistema de factoriza-

ción de señal con restricción monofónica MBHC-PM del capítulo 6 y para el sistema de separación de fuentes informado con modelos adaptativos de instrumento del capítulo 7. En ellos se han combinado distintas inicializaciones de las matrices, con restricciones tanto en la matriz de ganancias como en la de bases, así como la modificación en tiempo de ejecución iterativa de la matriz de bases. Todo ello certifica la flexibilidad y modularidad de este algoritmo NMF determinista.

En el capitulo 7 se ha modificado el algoritmo de separación y adaptación de modelos de instrumento, basado en un método NMF iterativo, para poder ejecutarse con una latencia algorítmica de media trama. Es decir, separación de fuentes en una trama concreta y la adaptación de los modelos de instrumento se puede llevar a cabo sólo con la información de las tramas pasadas, sin la necesidad de emplear información futura. Este algoritmo permite, si se ejecuta en una máquina con suficiente potencia de cálculo, realizar la separación y adaptación de modelos de manera *online*, permitiendo su desempeño en aplicaciones reales que requieran de esta característica.

La restricción de monofonía implementada en el capitulo 6 se puede considerar una información temporal, puesto que se conoce que en cada instante temporal sólo habrá una nota activa de cada instrumento, a pesar de no conocer previamente cuál es esa nota. Sin embargo se ha querido demostrar la implementación del método MBHC-PM sobre dos tipos algoritmos de factorización NMF y NNSC. El segundo de ellos diseñado como versión de baja complejidad para poder ser empleado en aplicaciones concretas que requieran de esta característica sacrificando cierta calidad en la factorización. La restricción de monofonía ha demostrado ser útil frente a algoritmos de factorización que no cuentan con ella gracias al mejor desempeño de todas las versiones del método propuesto con restricción monofónica.

8.3. Líneas futuras

Esta tesis ha hecho un recorrido sobre la linea de investigación en separación de fuentes en señales monocanal. Ha comenzado con la adaptación del algoritmo NMF determinístico para su uso con modelos de instrumento, se ha seleccionado el modelo de instrumento mas idóneo para los objetivos

de esta tesis, se han aplicado restricciones sobre el proceso de factorización, se ha integrado el uso de información temporal de *score*, se han adaptado los modelos de instrumento a los instrumentos reales de la señal de entrada, se ha adaptado el algoritmo para poder ejecutarse de manera online y se ha abordado la problemática de la separación de los parciales solapados.

Todo este recorrido lleva a una fase final de esta línea de investigación en concreto, pero todo el conocimiento generado se puede, y de hecho se está, trasladando a otros escenarios. Por ejemplo la separación de fuentes en señales multicanal o el alineamiento música-partitura.

El primero de los escenarios mencionados supone agregar una dimensión mas al problema a la vez que supone una fuente de información adicional, la información espacial. Esta información, por si misma ya permite realizar una separación de las fuentes presentes en cada uno de los sensores (micrófonos). Si además de ello, se cuenta con la información temporal y espectral que ha sido certificada como válida en esta tesis, se pueden llegar a obtener resultados de separación de muy alta calidad. En este tipo de escenarios multicanal, es habitual que las fuentes no estén constituidas por un sólo instrumento, sino por grupos de instrumentos del mismo tipo en una configuración orquestal. La separación que se puede obtener en una escenario de ese tipo con suficiente información puede ser bastante impactante.

El segundo de los escenarios, el alineamiento música-partitura, es un procedimiento ampliamente demandado por la industria musical actual, tanto por empresas del sector, como por músicos profesionales que demandan aplicaciones basadas en este procedimiento. Esta línea de procesado de señal puede tener ciertos puntos en común con el desarrollo de esta tesis. Los algoritmos de alineamiento suelen trabajar sobre matrices de datos que relacionan dos tipos de información. La confección de estas matrices se puede realizar de múltiples formas, entre ellas puede estar el cálculo de la distorsión generada en determinadas circunstancias con un modelo espectral de los instrumentos involucrados. Otras aplicaciones finales, como el acompañamiento automático a un instrumentista, pueden requerir una fase de síntesis musical, en la que pueden ser cruciales los modelos de instrumento y la síntesis de sonido que ha sido realiza en el último capítulo de la parte III de la tesis.

En esta tesis se ha abordado la separación de fuentes musicales, pero

estas técnicas de separación se pueden aplicar a la separación de fuentes sonoras en general, como pueden ser la voz o el ruido, pudiendo desarrollar sistemas de mejora de la calidad de la señal de voz o de música (frente al ruido) o la separación de distintas voces simultáneas.

Como se puede ver, hay algunas ideas interesantes para trabajar en el futuro en una investigación de calidad. Hemos planeado, no solo trabajar para generar publicaciones científicas, sino también para la colaboración con empresas del sector y de esta manera introducir el conocimiento generado en la economía.

Publicaciones

Revistas indexadas en JCR

1. **F.J. Rodríguez-Serrano**, J.J. Carabias-Orti, P. Vera-Candeas, F.J. Canadas-Quesada, y N. Ruiz-Reyes, "Monophonic constrained non-negative sparse coding using instrument models for audio separation and transcription of monophonic source-based polyphonic mixtures," *Journal on Multimedia Tools and Applications*, vol.63, no.2, Marzo 2013.

2. J.J. Carabias-Orti, **F.J. Rodriguez-Serrano**, P. Vera-Candeas, F.J. Canadas-Quesada, y N. Ruiz-Reyes, "Constrained non-negative sparse coding using learnt instrument templates for realtime music transcription," *Journal on Engineering Applications of Artificial Intelligence*, vol.26, no.7, pp. 1671-1680, Agosto 2013.

3. J.J. Carabias-Orti, M. Cobos, P. Vera-Candeas y **F.J. Rodríguez-Serrano**, "Nonnegative signal factorization with learnt instrument models for sound source separation in close-microphone recordings", *EURASIP Journal on Advances in Signal Processing*, Diciembre 2013.

4. **F.J. Rodriguez-Serrano**, Z. Duan, P. Vera-Candeas, B. Pardo y J.J. Carabias-Orti, "Online Score-Informed Source Separation with Adaptive Instrument Models," *Journal of New Music Research*, [Under review].

Congresos Internacionales

1. **F.J. Rodriguez-Serrano**, P. Vera-Candeas, P. Cabañas Molero, J.J. Carabias-Orti y N. Ruiz Reyes. "Amplitude Modulated Sinusoidal Modeling for Audio Onset Detection" *The 18th European Signal Processing Conference (EUSIPCO 2010)*. 23-27 Agosto 2010 Aalborg, Dinamarca.

2. **F.J. Rodriguez-Serrano**, J.J. Carabias-Orti, P. Vera-Candeas, T. Virtanen y N. Ruiz-Reyes, "Multiple Instrument Mixtures Source Separation Evaluation Using Instrument-Dependent NMF Models" *Latent Variable Analysis and Source Separation,10th International Conference on (LVA/ICA 2012)*. 12-15 de Marzo de 2012, Tel-Aviv, Israel.

Bibliografía

[Abdallah04] S Abdallah and M Plumbley. Polyphonic music transcription by non-negative sparse coding of power spectra. *in Proc. 5th Int. Society for Music Information Retrieval Conf. (ISMIR), Barcelona, Spain*, 2004.

[Abdallah06] S Abdallah and M Plumbley. Unsupervised analysis of polyphonic music by sparse coding. *Neural Networks, IEEE Transactions on*, 17(1):179–196, 2006.

[Alonso14] P. Alonso, V. M. Garcia, F.J. Martinez-Zaldivar, A. Salazar, L. Vergara, and A.M. Vidal. Parallel approach to NNMF on multicore architecture. *The Journal of Supercomputing*, 2014.

[Ans73] S Martin. American national standard psychoacoustical terminology. *American National Standards Institute*, 1973.

[BSSEVAL] C. Févotte, R Gribonval, and E. Vincent. BSS EVAL toolbox user guide - Revision 3.0. *http://bassdb. gforge.inria.fr/bss eval/.*

[Babaie06] Massoud Babaie-zadeh and Christian Jutten. Semi-Blind Approaches for Source Separation and Independent component Analysis. In *In proceeding of the 14th European Symposium on Artificial Neural Networks, ESANN*, 2006.

[Badeau09] R. Badeau, V. Emiya, and B David. Expectation-maximization algorithm for multi-pitch estimation and separation of overlapping harmonic spectra. *in Proc. Int.*

Conf. Acoust., Speech, Signal Process. (ICASSP), Taipei, Taiwan, 2009.

[BarryMSc03] J Barry. Polyphonic music transcription using independent component analysis. *MSc*, 2003.

[Bello06] JP Bello, L. Daudet, and MB Sandler. Automatic Piano Transcription Using Frequency and Time-Domain Information. *IEEE Transactions on Acoustic, Speech and Signal Processing*, 14(6):2242–2251, 2006.

[Belouchrani97] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique using second-order statistics. *Signal Processing, IEEE Transactions on*, 45(2):434–444, February 1997.

[Benaroya03] L Benaroya, L M Donagh, F Bimbot, and R Gribonval. Non negative sparse representation for Wiener based source separation with a single sensor. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP'03)*, pages VI–613–16. IEEE, 2003.

[Benaroya06] L Benaroya, F Bimbot, and R Gribonval. Audio source separation with a single sensor. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):191–199, January 2006.

[Berry07] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, 52(1):155–173, 2007.

[Bertin10] 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.

[Bjork96] A. Bjork. *Numerical Methods for Least Squares Problems.* SIAM, May 1996.

[BregmanBook90] A Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound.* MIT Press, 1990.

[Brown09] S Brown. listenhearsoundprojects. *http://wwwlistenhearcouk*, 2009.

[Bryan00] D. Bryan, D Lee, and H S Seung. Algorithms for non-negative matrix factorization. In *Proc. of the Neural Information Processing Systems (NIPS)*, Denver, April 2000.

[Burred07] J.J. Burred and T. Sikora. Monaural source separation from musical mixtures based on time-frequency timbre models. *Proceedings of International Conference on Music Information Retrieval (ISMIR), Vienna, Austria*, 2007.

[Candes08] E.J Candes and M.B Wakin. An Introduction To Compressive Sampling. *Signal Processing Magazine, IEEE*, 25(2):21–30, 2008.

[Carabias11] 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. *Selected Topics in Signal Processing, IEEE Journal of*, PP(99):1, 2011.

[Carabias13] J.J. Carabias-Orti, F.J. Rodriguez-Serrano, P. Vera-Candeas, F.J. Canadas-Quesada, and N. Ruiz-Reyes. Constrained non-negative sparse coding using learnt instrument templates for realtime music transcription. *Engineering Applications of Artificial Intelligence*, 2013.

[Cemgil08] A. Cemgil. Bayesian inference for nonnegative matrix factorisation models. *downloads.hindawi.com*, 2008.

[Chen06] Z Chen, A Cichocki, and T M Rutkowski. Constrained non-Negative Matrix Factorization Method for EEG Analysis

in Early Detection of Alzheimer Disease. *In Proc. IEEE International Confer- ence on Acoustics, Speech, and Signal Processing (ICASSP'06), Toulouse, France*, 2006.

[Cheveigne02] A deCheveigne and H Kawahara. YIN, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am*, 111(4):1917–1930, 2002.

[Chordia09] P. Chordia and A. Rae. Using source separation to improve tempo detection. In *Proc. of the 10th Int. Society for Music Information Retrieval*, April 2009.

[Christensen07] M. Christensen, P Stoica, A Jakobsson, and S Jensen. The Multi-Pitch Estimation Problem: Some New Solutions. *in Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007.

[Comon94] P Comon. Independent component analysis, a new concept? *Signal processing*, 1994.

[ComonBook10] P Comon and C Jutten. *Handbook of Blind Source Separation: Independent Component Analysis and ... - Pierre Comon, Christian Jutten - Google Books*. Academic Press, 2010.

[Conklin99] H Conklin Jr. Generation of partials due to nonlinear mixing in a stringed instrument. *J. Acoust. Soc. Am*, 1999.

[Cont06] A Cont. Realtime Audio to Score Alignment for Polyphonic Music Instruments, using Sparse Non-Negative Constraints and Hierarchical HMMS. *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, 5:V–V, 2006.

[Cont10] A Cont. A Coupled Duration-Focused Architecture for Real-Time Music-to-Score Alignment. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(6):974–987, June 2010.

[Cooke93] M P Cooke. *Modeling auditory processing and organisation.* Cambridge University Press, Cambridge, March 1993.

[DeLiangBook06] W DeLiang and G Brown. Computational Auditory Scene Analysis, 2006.

[Dixon00b] S. Dixon. On the computer recognition of solo piano music. *in Proc. of of Australasian Computer Music Conference*, 2000.

[Dixon05] S. Dixon. Live tracking of musical performances using online time warping. In *Proc. Int. Conf. on Digital Audio Effects (DAFx)*, Madrid, April 2005.

[Duan10] Zhiyao Duan, B. Pardo, Speech Changshui Zhang Audio, and Language Processing IEEE Transactions on. Multiple Fundamental Frequency Estimation by Modeling Spectral Peaks and Non-Peak Regions. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(8):2121–2133, November 2010.

[Duan11] Z. Duan and B. Pardo. Soundprism: An online system for score-informed source separation of music audio. *Selected Topics in Signal Processing, IEEE Journal of*, 5(6):1205–1215, 2011.

[Duan12] Z. Duan, G Mysore, and P Smaragdis. Online PLCA for real-time semi-supervised source separation. In *in Proc. Int. Conf on Latent Variable Analysis and Signal Separation (LVA/ICA 2012)*, April 2012.

[Durrieu10] JL Durrieu, G. Richard, B David, and C. Févotte. Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):564–575, 2010.

[Ellis03] D Ellis. Dynamic Time Warp (DTW) in Matlab. *http://www.ee.columbia.edu/ dpwe/resources/matlab/dtw/*, March 2003.

[EllisPhD96] D.P.W. Ellis. *Prediction-driven computational auditory scene analysis.* PhD thesis, Massachusetts Institute of Technology, March 1996.

[Emiya11] V. Emiya, E. Vincent, and N. Harlander. Subjective and objective quality assessment of audio source separation. *Audio*, 2011.

[Eronen01] A Eronen. Comparision of Features fo Musical Instrument Recognition. *In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2001.

[EronenMSc01] A Eronen. *Automatic Musical Instrument Recognition.* PhD thesis, Tampere University of Technology, 2001.

[Every06] M.R. Every and J.E. Szymanski. Separation of synchronous pitched notes by spectral filtering of harmonics. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1845–1856, 2006.

[Ewert12] S. Ewert and M. Muller. Using score-informed constraints for NMF-based source separation. *Acoustics*, 2012.

[Fevotte09a] C. Févotte, N. Bertin, and J Durrieu. Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural Computation*, 2009.

[Fevotte09b] C. Févotte and A.T. Cemgil. Nonnegative matrix factorizations as probabilistic inference in composite models. *17th European Signal Processing Conference (EUSIPCO 2009), Glasgow, Scotland*, 2009.

[Fevotte11b] C. Févotte and J. Idier. Algorithms for nonnegative matrix factorization with the beta-divergence. *Neural Computation*, 23, March 2011.

[FitzGerald03] D. FitzGerald, R Lawlor, and E. Coyle. Prior subspace analysis for drum transcription. *Audio Engineering Society Convention ...*, 2003.

[FletcherBook98] N H Fletcher and T D Rossing. *The Physics of Musical Instruments*. Springer, 1998.

[Fritsch13] J. Fritsch and M.D. Plumbley. Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, pages 888–891, May 2013.

[Gainza07] M. Gainza and E. Coyle. Automating Ornamentation Transcription. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, pages I–69–I–72, April 2007.

[Ganseman10] J. Ganseman, P. Scheunders, GJ Mysore, and JS Abel. Source separation by score synthesis. In *International Computer Music Conference Proceedings*, 2010.

[Ganseman12] J. Ganseman, P. Scheunders, and S. Dixon. Improving PLCA-based score-informed source separation with invertible Constant-Q Transforms. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 2634–2638, August 2012.

[Gemmeke11] J.F Gemmeke, T. Virtanen, and A. Hurmalainen. Exemplar-Based Sparse Representations for Noise Robust Automatic Speech Recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(7):2067–2080, 2011.

[Goto02] M. Goto, H Hashiguchi, T Nishimura, and R Oka. RWC music database: Popular, classical, and jazz music database. *in Proc. Int. Symp. Music Inf. Retrieval*, 2002.

[Goto04] M. Goto. A real-time music-scene-description system: Predominant-F0 estimation for detecting melody and bass lines in real-word audio signals. *Speech Communications*, 43(4):311–329, 2004.

[Goto05] M. Goto. RWC Music Database. *National Institute of Advanced Industrial Science and Technology (AIST)*, 2005.

[Grippo00] L Grippo and M Sciandrone. On the convergence of the block nonlinear Gauss–Seidel method under convex constraints. *Operations Research Letters*, 26(3):127–136, 2000.

[Heittola09] T. Heittola, A. Klapuri, and T. Virtanen. Musical instrument recognition in polyphonic audio using source-filter model for sound separation. *in Proc. 10th Int. Society for Music Information Retrieval Conf. (ISMIR), Kobe, Japan*, 2009.

[Helen05] M. Helen and T. Virtanen. Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. In *Proc. 13th European Signal Processing Conference (EUSIPCO), Antalaya, Turkey,*, March 2005.

[Hennequin11] Romain Hennequin, Bertrand David, and Roland Badeau. Score informed audio source separation using a parametric model of non-negative spectrogram. In *ICASSP 2011 - 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 45–48. IEEE, 2011.

[Hoyer04] P Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.

[Huber06] R. Huber and B. Kollmeier. PEMO-Q; A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):1902–1911, November 2006.

[Hyvarinen00] A Hyvärinen and E Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

[Iowa06] The University of Iowa. Iowa Music Database. *University of Iowa*, 2006.

[Itakura68] F Itakura and S. Saito. Analysis synthesis telephony based on the maximum likelihood method. *In Proc 6th International Congress on Acoustics*, pages C–17 – C–20, 1968.

[Jaiswal11] R. Jaiswal, D. FitzGerald, D. Barry, E. Coyle, and Scott Rickard. Clustering NMF basis functions using Shifted NMF for monaural sound source separation. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, pages 245–248, May 2011.

[Jang03] GJ Jang and TW Lee. A maximum likelihood approach to single-channel source separation. *The Journal of Machine Learning Research*, 2003.

[Joder12] C. Joder, F. Weninger, F. Eyben, D. Virette, and B. Schuller. Real-time speech separation by semi-supervised nonnegative matrix factorization. In *in Proc. Int. Conf on Latent Variable Analysis and Signal Separation (LVA/ICA 2012)*, pages 322–329, April 2012.

[Jutten91] Christian Jutten and Jeanny Herault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal processing*, 24(1):1–10, 1991.

[Kameoka12] H. Kameoka, M. Nakano, K. Ochiai, Y. Imoto, K. Kashino, and S. Sagayama. Constrained and regularized variants of non-negative matrix factorization incorporating music-specific constraints. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, pages 5365–5368, March 2012.

[Kim08] J. Kim, H. Park, and H. Nonnegative matrix factorization based on alternating nonnegativity constrained least squa-

res and active set method. *SIAM J. Matrix Anal. Appl.*, 30:713–730, April 2008.

[Kim08b] Dongmin Kim, Suvrit Sra, and Inderjit S Dhillon. Fast newton-type methods for the least squares nonnegative matrix approximation problem. *Statistical Analysis and Data Mining*, pages 38–51, 2008.

[Kim11] Jingu Kim and Haesun Park. Fast Nonnegative Matrix Factorization: An Active-Set-Like Method and Comparisons. *SIAM J. Sci. Comput.*, 33(6):3261–3281, 2011.

[Klapuri10a] A. Klapuri and T. Virtanen. Representing Musical Sounds With an Interpolating State Model. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):613–624, 2010.

[Klapuri10b] A. Klapuri, T. Virtanen, and T. Heittola. Sound source separation in monaural music signals using excitation-filter model and em algorithm. *in Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Dallas, USA*, pages 5510–5513, 2010.

[Klapuri99a] A. Klapuri. Sound onset detection by applying psychoacoustic knowledge. *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1999.

[KlapuriBook06] Perfecto Herrera-Boyer, Anssi Klapuri, and Manuel Davy. *Signal Processing Methods for Music Transcription.* Springer US, Boston, MA, 2006.

[KlapuriMSc97] A. Klapuri. Automatic Transcription of Music. *MSc*, pages 1–87, 1997.

[KlapuriPhD04] A. Klapuri. *Signal Processing Methods for the Automatic Transcription of Music.* PhD thesis, Tampere University of Technology, 2004.

[Lambert99] R.H. Lambert. Difficulty measures and figures of merit for source separation. *Independent Component Analysis*, February 1999.

[Lawson95] C.L. Lawson and R.J. Hansom. *Solving Least Squares Problems*. SIAM, May 1995.

[Lee01] D Lee and H Seung. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 2001.

[Lee88] C K Lee and D G Childers. Cochannel speech separation. *Journal of the Acoustical Society of America*, March 1988.

[Lee99] D D Lee and H S Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[Lepain99] P Lepain. Polyphonic pitch extraction from musical signals. *Journal of New Music Research*, 1999.

[Li01] Li. Learning spatially localized, parts-based representation. *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, 1:I–207 – I–212 vol.1, 2001.

[LiPhD08] Li Yipeng. *Monaural Musical Sound Separation*. PhD thesis, The Ohio State University, 2008.

[LindsayBook77] P Lindsay and D Norman. Human information processing: An introduction to psychology, 1977.

[MIREX2007] Music Information Retrieval Evaluation. Multiple Fundamental Frequency Estimation & Tracking Results. 2007.

[Maher90] R Maher. Evaluation of a method for separating digitized duet signals. *Journal Audio Engineering Society*, 38(12):956–979, 1990.

[Martin96a] K Martin. A Blackboard System for Automatic Transcription of Simple Polyphonic Music. *MIT Media Laboratory Perceptual Computing Section Technical Report No. 399*, 1996.

[MaxerPhD13] Ricard Marxer-Piñ on. *Audio Source Separation for Music in Low-latency and High-latency Scenarios.* PhD thesis, Universidad Pompeu Fabra, Barcelona, 2013.

[McGill92] F Opolko and J Wapnick. McGill University Master Samples. *McGill University*, 1992.

[Mesaros07] A. Mesaros, T. Virtanen, and A. Klapuri. Singer Identification in Polyphonic Music Using Vocal Separation and Pattern Recognition Methods. *ISMIR*, 2007.

[Mesaros10] A. Mesaros and T. Virtanen. Automatic recognition of lyrics in singing. *EURASIP Journal on Audio*, 2010.

[Mitianoudis02] Nikolaos Mitianoudis and Mike E Davies. Audio Source Separation: Solutions and Problems. *Internation Journal of Adaptive Control and Signal Processing*, pages 1–15, January 2002.

[Morita06] Satoru Morita and Yasuhito Nanri. Sound Source Separation of Trio using Stereo Musig Sound Signal Based on Independent Component Analysis. *Multimedia and Expo, 2006 IEEE International Conference on*, pages 185–188, 2006.

[Namgook09] Namgook Cho and C.-C.J. Kuo. Underdetermined audio source separation from anechoic mixtures with long time delay. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, pages 1557–1560, April 2009.

[Nishikawa02] T Nishikawa, H Saruwatari, K. Acoustics Speech Shikano, and Signal Processing ICASSP 2002 IEEE International Conference on. Bund source separation based

on Multi-Stage ICA combining frequency-domain ICA and time-domain ICA. *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, 1, March 2002.

[Nishikawa02b] T Nishikawa, H Saruwatari, and K Shikano. Comparison of time-domain ICA, frequency-domain ICA and multistage ICA. In *11th European Signal Processing Conference (EUSIPCO)*, March 2002.

[Nishikawa03] Nishikawa, T Abe, H Saruwatari, and H Shikano. Overdetermined blind source separation of real acoustic sounds based on multistage ICA using subarray processing. *Signal Processing and Information Technology, 2003. ISSPIT 2003. Proceedings of the 3rd IEEE International Symposium on*, March 2003.

[Olshausen97] BA Olshausen and DF Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision research*, pages 3311–3325, 1997.

[OppenheimBook97] A Oppenheim. Signals & Systems, 1997.

[OrtizPhD02] L. Ortiz-Berenguer. *Identificación automática de acordes musicales*. PhD thesis, Universidad Politécnica de Madrid, 2002.

[Ozerov09] A. Ozerov, C. Févotte, and M. Charbit. Factorial Scaled Hidden Markov Model for polyphonic audio representation and source separation. *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA '09. IEEE Workshop on*, pages 121–124, 2009.

[Ozerov10] A. Ozerov and C. Févotte. IEEE Xplore - Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(3):550–563, 2010.

[Ozerov11] A. Ozerov, E. Vincent, and F Bimbot. A general flexible framework for the handling of prior information in audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, January 2011.

[Paiva05] RP Paiva, T Mendes, and A Cardoso. An auditory model based approach for melody detection in polyphonic musical recordings. *Computer Music Modeling and Retrieval*, pages 21–40, 2005.

[PaivaPhD06] R Paiva. *Melody Detection in Polyphonic Audio*. PhD thesis, Computer Science Department, Science and Technology, University of Coimbra (Portugal), 2006.

[Parra00] L Parra and C Spence. Convolutive blind separation of non-stationary sources. *Speech and Audio Processing, IEEE Transactions on*, 8(3):320–327, May 2000.

[Parra04] L Parra and P. Sajda. Blind source separation via generalized eigenvalue decomposition. *J. Mach. Learn. Res.*, 4(7-8):1261–1269, 2004.

[Parra98b] L Parra, C Spence, and B De Vries. Convolutive blind source separation based on multiple decorrelation. In *Neural Networks for Signal Processing VIII, 1998. Proceedings of the 1998 IEEE Signal Processing Society Workshop*, pages 23–32, August 1998.

[Parsons76] T. Parsons. Separation of speech from interfering speech by means of harmonic selection. *Journal of the Acoustical Society of America*, March 1976.

[Pham01] Dinh-Tuan Pham and Jean-Francois Cardoso. Blind separation of instantaneous mixtures of nonstationary sources. *IEEE Transactions on Signal Processing*, 49(9), March 2001.

[Plumbley01] M Plumbley. Adaptive lateral inhibition for non-negative ICA. *Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation (ICA)*, 2001.

[Plumbley02] MD Plumbley, SA Abdallah, JP Bello, ME Davies, G Monti, and MB Sandler. Automatic music transcription and audio source separation. *Cybernetics and Systems*, 33(6):603–627, 2002.

[Plumbley03] M Plumbley. Algorithms for nonnegative independent component analysis. *IEEE Transactions on Neural Networks*, 14(3):534–543, 2003.

[Priestley74] M. Priestley, T. Rao, and H. Tong. Applications of principal component analysis and factor analysis in the identification of multivariable systems. *Automatic Control, IEEE Transactions on*, 19(6):730–734, December 1974.

[Quatieri90] T F Quatieri and R.G. Danisewicz. An approach to co-channel talker interference suppression using a sinusoidal model for speech. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 38(1):56–69, January 1990.

[RabinerBook78] R Rabiner and R W Schafer. Digital Processing of Speech Signals. *Prentice Hall, Upper Saddle River, New Jersey*, page 512, 1978.

[Raczynski07] S. Raczyński, N. Ono, and S. Sagayama. Multipitch analysis with harmonic nonnegative matrix approximation. *in Proc. 8th Int. Society for Music Information Retrieval Conf. (ISMIR), Vienna, Austria*, 2007.

[Raczynski08] S. Raczyński, Nobutaka Ono, and S. Sagayama. Extending nonnegative matrix factorization - A discussion in the Context of multiple frequency estimation of musical signals . In *16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland*, 2008.

[Rayleigh45] JWS Rayleigh. The Theory of Sound, volume 2, 2nd. edition. *Dover Publications, New York*, 2, 1945.

[Reyes03] M.J. Reyes-Gomez, B. Raj, and D.R.W. Ellis. Multichannel source separation by factorial HMMs. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, pages I–664–I–667 vol.1, April 2003.

[Rodriguez12] F.J. Rodriguez-Serrano, J.J. Carabias-Orti, P. Vera-Candeas, T. Virtanen, and N. Ruiz-Reyes. Multiple Instrument Mixtures Source Separation Evaluation Using Instrument-Dependent NMF Models. In Theis, Fabian, Cichocki, Andrzej, Yeredor, Michael, Zibulevsky, and Arie, editors, *in Proc. Int. Conf on Latent Variable Analysis and Signal Separation (LVA/ICA 2012)*, April 2012.

[Rodriguez13] F.J. Rodriguez-Serrano, J.J. Carabias-Orti, P. Vera-Candeas, F Canadas-Quesada, and N. Ruiz-Reyes. Monophonic constrained non-negative sparse coding using instrument models for audio separation and transcription of monophonic source-based polyphonic mixtures. *Multimedia Tools and Applications*, April 2013.

[Rodriguez14] F.J. Rodriguez-Serrano, Z. Duan, P. Vera-Candeas, B. Pardo, and J.J. Carabias-Orti. Online Score-Informed Source Separation with Adaptive Instrument Models. *Journal of New Music Research [Under Review]*.

[Rodriguez14b] F.J. Rodriguez-Serrano and P. Vera-Candeas. Overlapped harmonics separation for Music Source Separation. *On writting*.

[RossingBook90] T Rossing. The Science of Sound, 1990.

[Ryynanen08a] M. Ryynanen and A P Klapuri. Automatic transcription of melody, bass line, and chords in polyphonic music. *Computer Music Journal*, 32(3):72–86, 2008.

[Saruwatari01] H Saruwatari, Toshiya Kawamura, and K. Shikano. Fast-convergence algorithm for ICA-based blind source separation using array signal processing. In *Statistical Signal Processing, 2001. Proceedings of the 11th IEEE Signal Processing Workshop on*, pages 464–467, 2001.

[Sawada11] H. Sawada, S. Araki, and S. Makino. Underdetermined Convolutive Blind Source Separation via Frequency Bin-Wise Clustering and Permutation Alignment. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(3):516–527, March 2011.

[Schmidt06] M Schmidt and R Olsson. Single-channel speech separation using sparse non-negative matrix factorization. In *in International Conference on Spoken Language Processing (INTERSPEECH.*

[Schobben99] D Schobben, K Torkkola, and P Smaragdis. Evaluation of blind signal separation methods. *Proc of ICA and BSS*, 1999.

[Simon12] L. Simon and E. Vincent. A general framework for online audio source separation. In *in Proc. Int. Conf on Latent Variable Analysis and Signal Separation (LVA/ICA 2012)*, April 2012.

[Simsekli12] U. Simsekli and A.T. Cemgil. Score guided musical source separation using Generalized Coupled Tensor Factorization. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 2639–2643, August 2012.

[Smaragdis03] P Smaragdis and JC Brown. Non-negative matrix factorization for polyphonic music transcription. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, USA*, pages 177–180, 2003.

[Smaragdis97] P Smaragdis. *Redundancy Reduction for Computational Audition, a Unifying Approach.* PhD thesis, Massachusetts

Institute of Technology, Massachusetts Institute of Technology, March 1997.

[Smaragdis98] P Smaragdis. Blind separation of convolved mixtures in the frequency domain. *Neurocomputing*, March 1998.

[Turetsky03] Robert J Turetsky and Daniel P W Ellis. Ground-Truth Transcriptions of Real Music from Force-Aligned MIDI Syntheses. In *4th International Symposium on Music Information Retrieval ISMIR-03*, 2003.

[Valimaki06] Vesa Välimäki, Jyri Pakarinen, Cumhur Erkut, and Matti Karjalainen. Discrete-time modelling of musical instruments. *Rep. Prog. Phys.*, 69(1), 2006.

[Vincent03a] E. Vincent, C. Févotte, R Gribonval, and L Benaroya. A tentative typology of audio source separation tasks. *... Blind Signal Separation ...*, 2003.

[Vincent06] E. Vincent. Musical source separation using timefrequency source priors. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):91–98, 2006.

[Vincent07] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca. First Stereo Audio Source Separation Evaluation Campaign: Data, Algorithms and Results. *Independent Component Analysis and Signal Separation Lecture Notes in Computer Science*, 4666:552–559, March 2007.

[Vincent10] 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, 2010.

[Vincent12] E. Vincent. Improved perceptual metrics for the evaluation of audio source separation. In *10th int. conf. on latent variable analysis and signal separation*, Tel-Aviv, March 2012.

[Virtanen03] T. Virtanen. Sound source separation using sparse coding with temporal continuity objective. *IEEE International Conference on Acoustic, Speech, and Sinal Processing*, 2003.

[Virtanen06] T. Virtanen and A. Klapuri. Analysis of polyphonic audio using source-filter model and non-negative matrix factorization. *Advances in Models for Acoustic Processing, Neural Information Processing Systems Workshop*, 2006.

[Virtanen07b] T. Virtanen. Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. *Audio*, 2007.

[Virtanen08] T. Virtanen, A. Cemgil, and S Godsill. Bayesian extensions to non-negative matrix factorisation for audio signal modelling. *in Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Las Vegas, USA*, pages 1825–1828, 2008.

[Wang04] Y. Wang, M. Kan, T. Nwe, A. Shenoy, and J. Yin. LyricAlly: Automatic synchronization of acoustic musical signals and textual lyrics. *in ACM international conference on Multimedia*, 2004.

[Wang05] B. Wang, Q. Mary, and M.D. Plumbley. Musical audio stream separation by non-negative matrix factorization. In *in Proc. UK Digital Music Research Network (DMRN) Summer Conf*, 2005.

[Weinstein93b] E Weinstein, M Feder, and A.V. Oppenheim. Multi-channel signal separation by decorrelation. *Speech and Audio Processing, IEEE Transactions on*, 1(4):405–413, October 1993.

[Weintraub84] M Weintraub. The GRASP sound separation system. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '84.*, pages 69–72, March 1984.

[Wold96] E Wold, T Blum, D Keislar, and J Wheaton. Content-based Classification, Search and Retrieval of Audio. *IEEE Multimedia*, 3(3):27–36, 1996.

[YehPhD08] C. Yeh. *Multiple Fundamental Frequency Estimation of Polyphonic Recordings.* PhD thesis, University Paris VI, 2008.

[Yin05] J. Yin. Music transcription using an instrument model. *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, 3:iii–217–iii–220 Vol. 3, 2005.

[Young06] S Stanley Young, Paul Fogel, and Douglas M Hawkins. Clustering scotch whiskies using nonnegative matrix factorization. *Joint Newsletter for the Section on Physical and Engineering Sciences and the Quality and Productivity Section of the American Statistical Association*, 14(1):11–13, 2006.

[Zhang03] T Zhang. System and method for automatic singer identification. *Hp Labs, Tech. Rep. HPL-2003-8*, 2003.

[Zibulevsky01] P. Kisilev, M. Zibulevsky, and Y.Y. Zeevi. Blind source separation using multinode sparse representation. In *Image Processing, 2001. Proceedings. 2001 International Conference on*, pages 202–205 vol.3, 2001.

[Ziehe98] Andreas Ziehe and Klaus-Robert Müller. TDSEP - an efficient algorithm for blind separation using time structure. In *Proceedings of the 8th International Conference on Artificial Neural Networks, ICANN'98*, 1998.

[ZwickerBook90] E Zwicker and H Fastl. *Psychoacoustics, Facts and Models.* Springer, 1990.