Multimed Tools Appl
https://doi.org/10.1007/s11042-018-5944-2

Multimodal speaker diarization for meetings using volume-evaluated SRP-PHAT and video analysis

P. Cabañas-Molero[1] · M. Lucena[2] · J. M. Fuertes[2] · P. Vera-Candeas[1] · N. Ruiz-Reyes[1]

Received: 24 July 2017 / Revised: 26 January 2018 / Accepted: 26 March 2018

Abstract Speaker diarization is traditionally defined as the problem of determining "who speaks when" given an audio or video stream. This is an important task in many applications for meeting rooms, including automatic transcription of conversations, camera steering or content summarization. When the room is equipped with microphone arrays and cameras, speakers can be distinguished according to their location and the problem can be addressed through localization techniques. This article proposes a multimodal speaker diarization system for meeting environments based on a modified SRP-PHAT function evaluated on space volumes rather than discrete points. In our system, this function is used in combination with a circular array, enabling audio-based localization based on the selection of local maxima. Voicing detection is used to detect speech frames, whereas video analysis is introduced to aid in the decision when users move or simultaneously speak. The approach is evaluated on the well-known AMI dataset with approximately 100 hours of realistic meeting recordings and shows an average diarization error rate of 21% – 25%.

Keywords Speaker diarization · Meeting rooms · SRP-PHAT · Multimodal processing

✉ P. Cabañas-Molero
pcabanas@ujaen.es

M. Lucena
mlucena@ujaen.es

J. M. Fuertes
jmf@ujaen.es

P. Vera-Candeas
pvera@ujaen.es

N. Ruiz-Reyes
nicolas@ujaen.es

[1] Department of Telecommunication Engineering, University of Jaén, Linares, Jaén, Spain

[2] Department of Computer Science, University of Jaén, Jaén, Spain

Published online: 11 April 2018

1 Introduction

Automatic monitoring of multi-participant meetings has become an extensively studied research topic in the field of multimedia processing. The basic idea is to track the activity of the participants using the information provided by a set of sensors installed in the room, with the aim of developing advanced applications such as automatic transcription of conversations [20], content-based retrieval [40], audiovisual human-computer interaction (HCI) [30] and intelligent camera steering and beamforming [44]. In the most common scenario, the meeting room employs one or more microphone arrays and cameras situated at a convenient location (typically the centre of a table) in order to capture properly the scene. A concrete setup extracted from the Augmented Multi-party Interaction (AMI) corpus [9, 27] is shown in Fig. 1, where the meeting is recorded with an 8-element microphone array and 4 cameras.

One of the most important front-end tasks in many meeting monitoring systems is speaker diarization (SD). SD can be defined as the process of detecting "who speaks when" (or which participants are speaking at each moment) [2]. The problem is similar to voice activity detection (VAD), with the additional difficulty that the system must assign each voice segment to the correct participant (it can be viewed as performing VAD for each participant). In applications involving speech recognition, SD enables the association of each transcription with its corresponding speaker, which is useful for automatic content summarization. SD can even facilitate the transcription process in certain scenarios. For instance, in systems based on microphone arrays, a beamformer can be used to capture audio from the active speaker location. This way, the system can attenuate interfering sounds and discard non-speech segments that detrimentally affect speech recognition. SD is also useful in distributed meetings, enabling participants in remote locations to hear and view a signal focused on the active speaker. Often, these applications rely on the robustness of the SD algorithm, which may be difficult to achieve in realistic scenarios. In typical meeting rooms, the speech signal captured by distant microphones is affected by severe reverberation and background noise, and the video data include natural movements and poses in which the faces are partially occluded or obscured. Furthermore, spontaneous conversations usually

Fig. 1 **a** A meeting room included in the AMI dataset, equipped with an 8-channel microphone array and 4 cameras located at the centre of the table. **b** Video signals recorded with the cameras

produce considerable speaker overlap, thus increasing the complexity of the problem. The task is even more challenging if the system must operate in real-time.

Several approaches for SD have been proposed in the literature. These approaches primarily use audio features or a combination of audio and video cues. Depending on the employed audio techniques, we can distinguish two types of strategies: *speaker modelling* (SM) and *sound source localization* (SSL). SM approaches are based on the fact that each voice has a different timbre, such that a different model can be learned for each speaker. Commonly, the models of each speaker are constructed using Gaussian mixture models (GMMs) of features such as mel-frequency cepstral coefficients (MFCCs) [41]. In online systems, these models are often pre-trained for the target speakers [7, 21] and are not applicable to unknown participants. Other online algorithms are capable of adapting to new speakers by using generic models and clustering techniques [15, 36, 38], in a similar way to offline systems [13, 14, 33]. However, these online generic systems require a certain amount of initialization data (or latency) and usually do not deal with overlapping speakers. SSL approaches exploit the fact that each speaker occupies a different position during the recording, so users may be distinguished by determining the location of the sound source [1, 3, 20, 26]. Recent works on SD suggest that systems incorporating SSL provide better performance than SM methods [32], especially those involving the steered response power with the phase transform (SRP-PHAT) technique [11]. Since SSL algorithms such as SRP-PHAT provide the position of only the most prominent source, the common approach is to extend these algorithms to develop a diarization system. This task involves three main challenges. First, VAD must be used to determine whether the localized sound event is speech, so that no common noise sources are detected as speakers. Second, when there are multiple active speakers, the system must detect all of them and not only the most prominent one. Finally, if people move during the meeting, the system must assign them the same identity independently of their current position, using some type of tracking procedure.

To overcome some of these problems, many SSL-based systems use a multimodal approach in which the audio stream is processed in conjunction with other modalities of data, such as video. Multimodal processing consists of fusing information from different types of sensors to resolve complex problems. The advantage of multimodal approaches is that the correlation among different types of information can be exploited, improving the results of methods based on only one type of data. Usually, problems involving detection [23] or prediction [24, 25] of activities are addressed through multi-sensor techniques. For SD purposes, visual information provides useful features that can compensate for certain flaws of the audio analysis and can be extracted with efficient and reliable algorithms. In recent years, several SD algorithms combining SSL and video processing have offered interesting results and have outperformed methods based exclusively on audio. In [42], a pool of features derived from SSL and visual motion are selected by a boosting algorithm to train a decision tree to detect active speakers. In [32], a hybrid system combining SM and SSL is proposed. Location observations are obtained through visual analysis and SRP-PHAT, and identity likelihoods are computed from pre-trained GMMs for each speaker and combination of speakers. The fusion of both types of observations is performed with a hidden Markov model (HMM), where each state represents a unique combination of speakers. A similar strategy is followed in [34], but does not handle overlapping speech. SM is based on pre-trained models of acoustic and facial features, whereas SSL is used to aid in the detection of faces and to detect speaker changes. In [30], an SD system targeted for HCI applications is proposed. First, a camera and a depth sensor are employed to determine the 3D location of faces and to extract features from the inner lip contours. Then, the faces' locations are used to evaluate locally the SRP-PHAT in regions around each user

and to derive SSL-based features. All features are finally merged through a Support Vector Machine (SVM) classifier, achieving a high accuracy rate. However, the algorithm does not work if the users do not face the camera.

In this paper, we present a multimodal SD system for meeting rooms based on a modified SRP-PHAT function. This modification was originally proposed in [10, 26] and consists of evaluating the function on spatial volumes rather than discrete points. The advantage of this approach is that it achieves similar performance to the standard SRP-PHAT with fewer evaluations, and it is very appropriate for applications where the exact coordinates of the sources are not needed. In our system, unlike [10, 26], we use the function with a circular array configuration that divides the search space into equally-sized sectors around the array (assumed to be placed in the centre of the meeting table). We experimentally demonstrate that with this particular setup, active speakers can be detected as local maxima in the function, even in situations with overlapping speakers. The system is completed with a visual analysis module that measures motion and lip movements to increase its robustness for simultaneous speech or moving speakers (particularly when a user leaves his or her seat to make a presentation). A voicing detector is used to detect speech frames. The decision algorithm is completely unsupervised, and the system can be tuned with two threshold parameters. The evaluation with the AMI dataset demonstrates the robustness of the system in realistic meeting rooms with spontaneous conversations.

The proposed algorithm is able to perform robustly in realistic meeting rooms, handle overlapping speech and execute in real-time with low latency. The main novelty of the system is the combination of an efficient volume-evaluated version of SRP-PHAT with several signal processing techniques to overcome many of the adversities found in these scenarios. A voicing detector is used because voiced speech frames have very robust features against background noise, and location estimates are often more reliable on them. Detection of lip movement and motion filtering are used to refine the SD decisions and detect whether a user is giving a presentation at the whiteboard. Furthermore, the proposed division of the search space allows to locate sound sources by selecting local maxima above a fixed threshold, thus providing also an efficient way to detect simultaneous sources.

This paper is organized as follows. In Section 2, we present a general overview of the system and explain the required setup. The detailed description of the proposed approach is provided in Sections 3 (audio analysis, including the SSL and voicing detection modules) and 4 (video analysis and decision algorithm). In Section 5, we evaluate the system with the AMI database and discuss the results. Conclusions and future work lines are presented in Section 6.

2 System overview

Our system requires a meeting room and table-top hardware similar to the ones used in the AMI corpus, as depicted in Fig. 1. In the AMI dataset, audio and video signals are captured with a microphone array and a set of cameras placed at the centre of the table. The audio equipment consists of an 8-element circular equi-spaced microphone array with a diameter of 20 cm that is composed of omnidirectional microphones that record at 16 kHz. Visual information is captured with several table-top cameras covering the space occupied by the users and operating at 25 fps. These cameras are not required to obtain high-resolution images (in our experiments, the image size of the recordings is 288x352 pixels), but their field of view (FoV) must be wide enough to capture the joint space where the participants are seated.

As detailed in [27], to create the AMI dataset, 8 Sennheiser MK2E-P-C miniature omni-directional electret microphones were used for the array. The audio was digitalized at 16 kHz, 16 bit resolution with an eight-channel pre-amplifier device connected to the computer through a single ADAT Lightpipe fibre optic cable that carried all eight channels. To record the video, four Sony XC555 subminiature cameras with 6 mm lenses were mounted under the central microphone array, providing close-up views of each of the meeting's participants. Four digital video recorders were used to record directly the output of the cameras to Mini-DV cassettes. The video was encoded at a bitrate of 2300 kbps using the DivX AVI codec 5.2.1 with a maximum interval of 25 frames between two consecutive MPEG keyframes. The dataset includes other signals acquired with additional microphones and cameras, but they are not required for our algorithm. This configuration is a fine example of the typical hardware setup employed in smart meeting rooms. In fact, several works in the literature have used similar configurations to develop actual prototypes of practical applications based on SD [20, 40]. Other setups employ special purpose devices having multiple microphones and cameras, for instance, to support distributed meetings [44] or, more recently, for entertainment applications [30] and informal meetings [17].

A schematic representation of the meeting room is illustrated in Fig. 2. The participants are expected to be located at certain predefined regions around the centred table, according to the available seats in the room. Consequently, assuming that users do not change their seats, an active speaker can be identified according to his/her predefined region. With this assumption, the problem of SD is then reduced to a combination of VAD and multi-speaker SSL. In our required setup, each camera must be arranged to capture exclusively the region corresponding to one seat/user. If the number of cameras is less than the number of seats, some cameras can be used to capture more than one region as long as the portion of the FoV corresponding to each seat is known in advance. In our experiments, we assume that the number of seats is $N = 4$ with a single camera per position. As shown in Fig. 2, one portion of the workspace corresponds to the whiteboard area. During a meeting, users may temporarily leave their seats and stand at the whiteboard to make presentations. In this case,

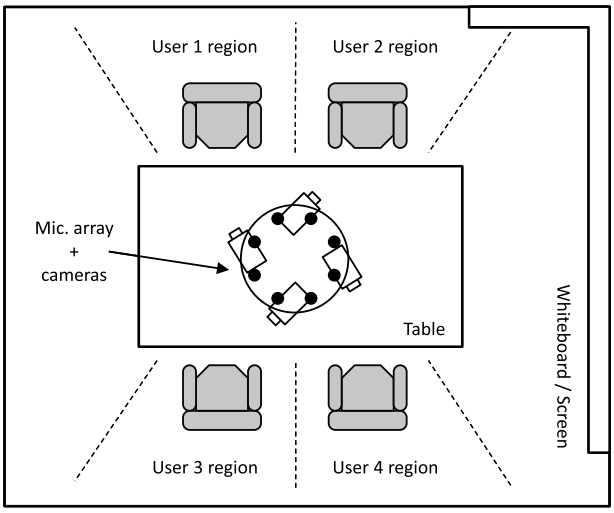

Fig. 2 Schematic representation of the meeting room

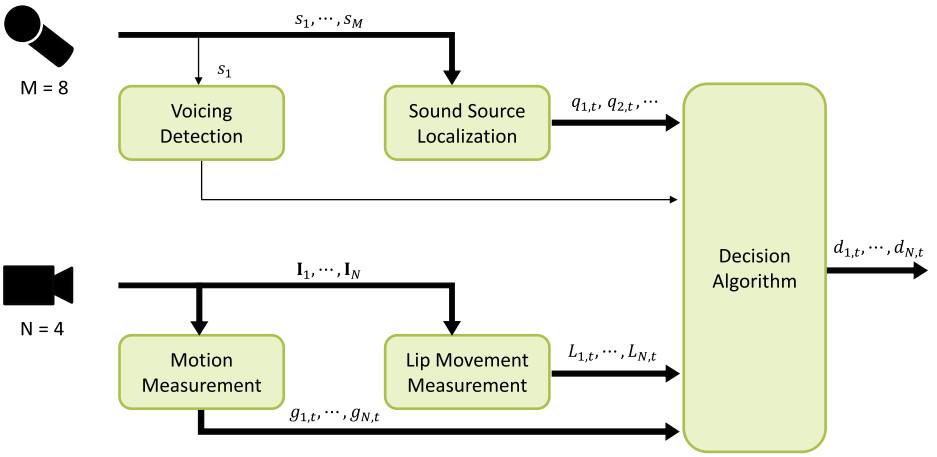

Fig. 3 Block diagram of the proposed multimodal SD approach

the speaker using the whiteboard must be assigned the same identity by the system even if he/she is outside of his/her corresponding region.

The architecture of the proposed multimodal approach is shown in Fig. 3. The audio processing module receives the eight-channel microphone signals s_m $(m = 1, \ldots, M,\ M = 8)$ as inputs, and performs voicing detection and SSL. The voicing detection module processes one of the channels, s_1, to obtain a binary decision $v_t \in \{0, 1\}$ indicating that voiced speech has been detected at frame t. The SSL module discretizes the meeting room space into Q sectors and provides the set of sectors $\hat{q}_{i,t}$ $(i = 1, 2, \ldots)$ containing sound at the tth frame. At this point, the audio module alone is able to determine the identity of the active speakers by mapping the sectors $\hat{q}_{i,t}$ to the pre-defined regions occupied by the participants. However, to offer a more robust diarization result (especially in situations with overlapping speech or when one of the participants is using the whiteboard), we also include visual information in the decision.

The video processing part receives the signals captured by the N cameras $\mathbf{I}_n$ $(n = 1, \ldots, N,\ N = 4)$ and performs motion and lip detection on each stream. At each frame t, the visual tracking algorithm returns two values: a motion measure $g_{n,t}$ indicating the movement detected in camera n and a measure of the lip movement $L_{n,t}$ in the detected face (in the case of affirmative facial detection). The latter is a representative indicator of speech activity. Finally, the information provided by both the audio and video modules is combined by the decision algorithm to provide binary speaker diarization results $d_{n,t} \in \{0, 1\}$ for each individual speaker n.

3 Audio analysis

3.1 Sound source localization with modified SRP-PHAT

As noted above, we evaluate the potential positions of the sound sources using a modified version of the SRP-PHAT function. The classical SRP-PHAT algorithm is well-known as one of the most robust SSL approaches in the presence of acoustic corruptions, such as

reverberation and background noise. This algorithm can be interpreted as a beamforming-based technique in which the microphone array is electronically steered (with a delay-and-sum beamformer) to each candidate location to find the one with maximum power. The algorithm can be summarized with the following operation:

$$\hat{\mathbf{x}}_t = \arg\max_{\mathbf{x}\in\mathcal{G}} P_t(\mathbf{x}), \tag{1}$$

where $\mathcal{G}$ denotes the set of all candidate spatial points and $P_t(\mathbf{x})$ is the power of the tth audio frame at the output of a weighted delay-and-sum beamformer aimed at point $\mathbf{x} = [x, y, z]$.

DeBiase [11] demonstrated that $P_t(\mathbf{x})$ can be computed from the generalized cross correlation (GCC) functions for all microphone pair combinations. The GCC function for a microphone pair (k, l) is computed as

$$R_t^{kl}(\tau) = \int_{-\infty}^{\infty} \frac{S_k(\omega, t) S_l^*(\omega, t)}{|S_k(\omega, t) S_l^*(\omega, t)|} e^{j\omega\tau} d\omega, \tag{2}$$

where τ is the time delay, $*$ denotes the complex conjugate and $S_k(\omega, t)$ is the Fourier transform of the signal at the kth microphone. The denominator term in (2) is the PHAT weighting function, whose purpose is to remove amplitude information in order to emphasize equally the phase differences in all frequency components. Although there are other weighting methods, the PHAT transform has been reported to be very effective in reverberant scenarios for all kinds of sources, including speech. Strictly speaking, when using the PHAT transform, $R_t^{kl}(\tau)$ is the GCC-PHAT function and $P_t(\mathbf{x})$ is the SRP-PHAT function.

By removing terms that are not dependent on $\mathbf{x}$, the function $P_t(\mathbf{x})$ can be computed by summing all GCCs evaluated at the correct delay:

$$P_t(\mathbf{x}) = \sum_{k=1}^{M} \sum_{l=k+1}^{M} R_t^{kl}\left(\tau_{kl}(\mathbf{x})\right), \tag{3}$$

where $\tau_{kl}(\mathbf{x})$ is the inter-microphone time-delay function (IMTDF). This function yields the time difference of arrival between microphones k and l for a source located at $\mathbf{x}$. Mathematically, it can be computed with a simple geometric operation:

$$\tau_{kl}(\mathbf{x}) = \frac{\|\mathbf{x} - \mathbf{x}_k\| - \|\mathbf{x} - \mathbf{x}_l\|}{c}, \tag{4}$$

where c is the sound propagation speed, $\mathbf{x}_k$ and $\mathbf{x}_l$ are the microphone positions, and $\|\cdot\|$ denotes the Euclidean norm.

Despite its robustness, the main disadvantage of the SRP-PHAT algorithm is the large number of required computations. To successfully locate the source, the spatial resolution of the grid $\mathcal{G}$ must be relatively high, which requires one to evaluate $P_t(\mathbf{x})$ for a large number of points. If the spatial resolution is reduced, the actual source location may land too far from the adjacent sampled positions, increasing the risk of choosing a very distant global maximum.

Authors have proposed modifications to the SRP-PHAT method to alleviate the computational cost without sacrificing precision [12, 43]. For our purposes, the most interesting strategy is that presented by Cobos et al. [10]. Instead of evaluating the SRP-PHAT function at discrete points, the authors integrate the information regarding the volume surrounding each candidate position. To this end, they formulate an alternative expression of (3), where

the GCCs are accumulated over a range of lags corresponding to a certain volume. This modified function for a region/volume q is expressed as

$$P'_t(q) = \sum_{k=1}^{M} \sum_{l=k+1}^{M} \sum_{\tau=\ell_1^{kl}(q)}^{\ell_2^{kl}(q)} R_t^{kl}(\tau), \tag{5}$$

where $\ell_1^{kl}(q)$ and $\ell_2^{kl}(q)$ are the accumulation limits. These limits must be determined for each microphone pair (k, l) and region q because they depend on the microphones' locations and the geometry of the region. These regions can be arbitrarily large, which may be sufficient for applications in which it is not necessary to determine the exact coordinates of the source but only its area. We can reasonably adopt this approach for our problem.

The first step of our approach consists of discretizing the search space into Q equally-sized sectors. The geometry of each sector is illustrated in Fig. 4. Assuming a cylindrical coordinate system $\mathbf{x} = [\rho, \psi, z]$ with its origin at the centre of the array, the boundaries of each sector q are $\rho \in [\rho_{\min}, \rho_{\max}]$, $z \in [0, z_{\max}]$ and $\psi \in \left[(q-1)\frac{2\pi}{Q}, q\frac{2\pi}{Q}\right]$, where $q = 1, \ldots, Q$. This division is different from the speaker-based division illustrated in Fig. 2, and it covers a large amount of space around the array. Since the region of each potential speaker is known for a given room, a simple mapping can be used to relate each sector q to its corresponding participant. The motivation of this division is to ensure that all volumes have the same size and equal importance when evaluating the modified SRP-PHAT function. Furthermore, it globally characterizes the localization space with a reasonable value of Q, that should be a trade-off between angular resolution and computational cost.

Given this geometry, the problem is to determine correctly the accumulation limits $\ell_1^{kl}(q)$ and $\ell_2^{kl}(q)$. As demonstrated in [10], when the IMTDF $\tau_{kl}(\mathbf{x})$ is evaluated inside a volume, its maximum and minimum values always occur on the surface that delimits the volume. Consequently, for each sector q, the accumulation limits can be determined by finding the maximum and minimum values of $\tau_{kl}(\mathbf{x})$ on its six boundary surfaces. That is,

$$\ell_1^{kl}(q) = \min_{\mathbf{x} \in \mathcal{G}_q} \tau_{kl}(\mathbf{x}), \tag{6}$$

$$\ell_2^{kl}(q) = \max_{\mathbf{x} \in \mathcal{G}_q} \tau_{kl}(\mathbf{x}), \tag{7}$$

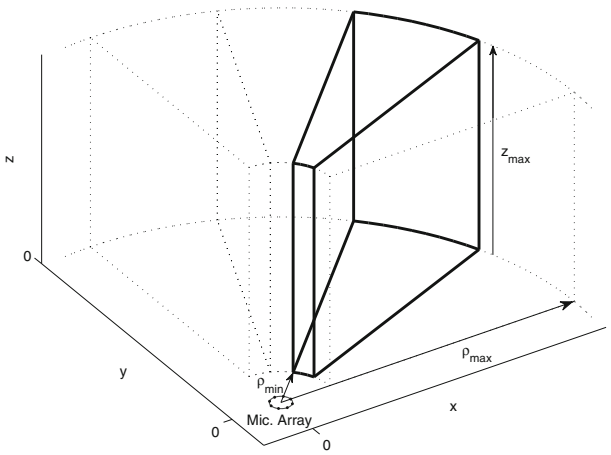

Fig. 4 Spatial division of the search space into equally-sized 3D sectors

where $\mathcal{G}_q$ is the set of points within the surfaces that enclose sector q. Observe that these limits can be pre-computed before running the algorithm once a certain configuration for the parameters Q, $\rho_{\min}$, $\rho_{\max}$ and $z_{\max}$ has been set. In our case, we numerically determined these limits by evaluating the IMTDF on a fine grid of boundary points and searching for the global maximum and minimum.

The resulting SRP-PHAT function, $P'_t(q)$, can be interpreted as the proportion of frequency bins pointing at sector q. In silent situations, all values of $P'_t(q)$ tend to be similar because the phase differences are uniformly distributed. When a sound event with a certain intensity occurs, certain frequency bins will point to the sector where the source is located. This process will increase the value of $P'_t(q)$ in the closest sectors and reduce the value in certain other sectors. An example is illustrated in Fig. 5 for the cases of a silent frame (Fig. 5a) and a single active speaker (Fig. 5b). In the presence of simultaneous speakers, the functional peaks may not be particularly noticeable, but they still appear as local maxima with large values (see Fig. 5c and Fig. 5d).

Since the number of active speakers in frame t is unknown, we establish a threshold $P_{\min}$ on the functional peaks to determine the presence of source activity. As a result, the algorithm returns the sectors $\hat{q}_{i,t}$ $(i = 1, \ldots, N_t)$ that produce local maxima above $P_{\min}$,

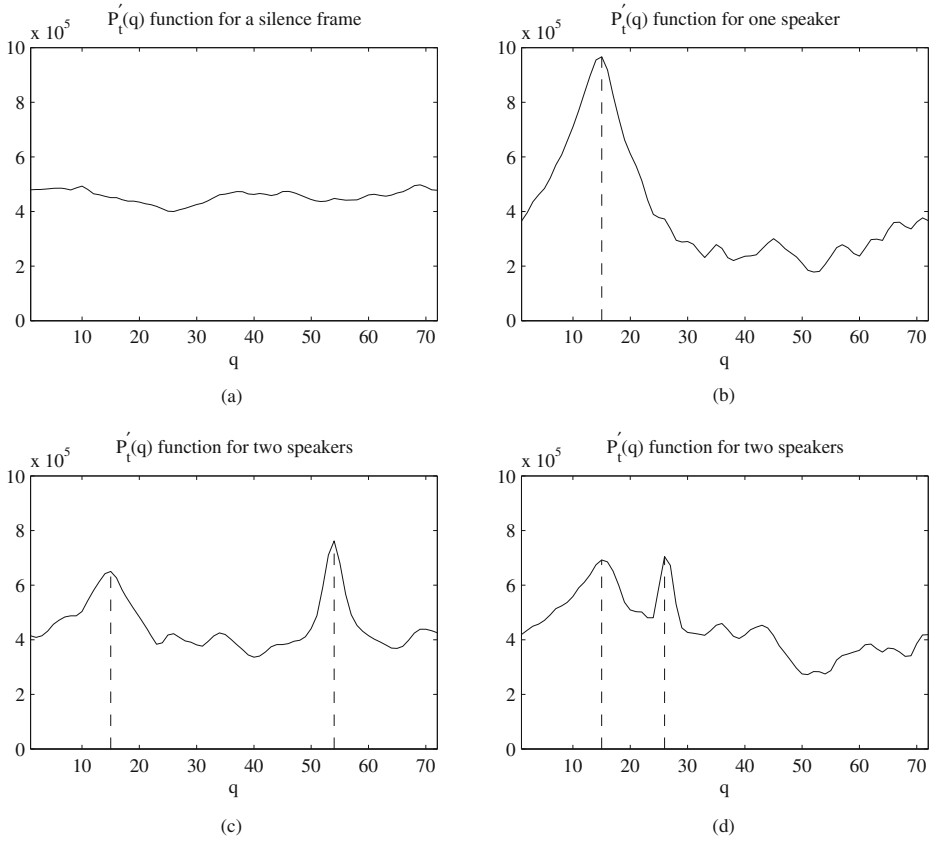

Fig. 5 Modified SRP-PHAT function $P'_t(q)$ for **a** a silent frame, **b** one single speaker, **c** two distant speakers and **d** two close speakers

where N_t is the number of sectors returned in frame t and $\hat{q}_{1,t}$ is the sector corresponding to the global maximum. It is important to note that just looking at the SRP-PHAT response for one sector is not sufficient for determining the presence of a sound source because the value can be affected by adjacent sources. Moreover, to provide local maxima, the angular resolution must be high enough to ensure that two speakers are not located in consecutive sectors.

3.2 Voicing detection

The localization module provides information about the most probable locations of potential active sound sources. However, for SD purposes, it is also necessary to discriminate whether the incoming sound is speech. Meeting environments are particularly challenging for this problem because the signal is affected by multiple stationary and non-stationary noises (air conditioning, computer fans, door slams, table hits, etc.) and speech is captured by distant-range microphones, which often causes mismatches in conventional close-range VAD models.

In the context of meetings and HCI applications, certain approaches also employ location estimates to perform VAD. The common idea is that when there is an active speaker, the location estimates across time will be concentrated around the true speaker location. Therefore, their spatio-temporal consistency can be modelled and exploited for VAD [8, 26]. However, this strategy may be rather sensitive to noise and isolated localization errors, especially in spontaneous meetings with multiple noise sources. In [30], the same idea is more successfully applied in combination with a visual face tracker. However, the method requires the users to face continuously the camera, which may not occur during meetings. Another attempt to combine SSL and visual features for VAD is discussed in [42], but the performance for overlapping speech is not evaluated. Other algorithms address VAD and SSL as separate problems. In this case, more conventional VAD techniques are applied, often based on statistical modelling of audio features [16, 20]. To increase performance, these techniques are usually combined with beamforming to isolate the located sources [4, 6].

In the proposed approach, speech detection is performed independently of SSL. Unlike conventional VAD-based systems, which aim to detect complete speech segments, we use a voicing detector to discriminate only voiced speech. The motivation for this approach is twofold. On the one hand, voiced phonemes exhibit specific properties (periodicity, harmonicity) that can be used to discriminate quite well between speech and non-speech frames, avoiding many of the false positives arising in VAD. On the other hand, voiced frames are sparse, meaning that most of their energy is concentrated in a few bins. This property reduces overlap in the frequency domain when simultaneous sources are active and makes speech bins to dominate over background noise levels. For this reason, location estimates obtained in voiced frames are often more reliable than those obtained in unvoiced speech.

In our system, we use the voicing detector included in the PEFAC pitch estimator, described in [19] and briefly summarized in the following. First, the input one-channel signal s_1 is transformed to the log-frequency domain by computing its power spectrum $S_1(f, t)$. The power spectrum is then normalized with respect to the deviation between a smoothed version of the spectrum and a fixed universal long-term average speech spectrum. This process tends to remove the singularities of the speech signal and attenuate noise components. Finally, the normalized spectrum $\overline{S}_1(f, t)$ is convolved with a harmonic filter to produce a pitch salience function $Z(f, t)$. From both $\overline{S}_1(f, t)$ and $Z(f, t)$, a 2-element feature vector is calculated at each frame for voicing detection as follows:

– the log-mean power of the normalized spectrum $E_t = \log\left(\frac{1}{F}\sum_f \overline{S}_1(f,t)\right)$, where F is the number of frequency bins in the log-frequency domain. Because voiced speech contains more energy, this feature is typically higher in voiced frames.
– the ratio of the sum of the highest three peaks in $Z(f,t)$ to E_t

$$h_t = \frac{\sum_{i=1}^{3} Z(f_i,t)}{E_t}, \tag{8}$$

where f_i indicates the frequencies corresponding to the three highest peaks. This ratio measures the fraction of the total power that is harmonically related and is thus much higher in voiced frames.

Voiced and unvoiced classes are modelled by GMMs using the 2-element feature vector $[E_t, h_t]$ as an input. The voicing state is discriminated by using a likelihood ratio test as follows:

$$v_t = \begin{cases} \text{Voiced}, & 1/(1+p_{u,t}/p_{v,t}) > 0.5 \\ \text{Unvoiced}, & \text{otherwise} \end{cases} \tag{9}$$

where $p_{u,t}$ and $p_{v,t}$ are the output probabilities returned by the unvoiced and voiced GMMs, respectively. In our experiments, we observed that this voicing detector performs relatively well even with simultaneous speakers. We must stress that we did not perform any training of the GMMs on our own, and therefore, the algorithm was run with the same parameters as in [19].

4 Video analysis and decision algorithm

Given the results of the SSL and voicing detection modules, the system is technically able to determine the activity of the speakers at each frame. However, in practice, our audio analysis faces two important problems. First, given a voiced frame, there is no guarantee that the secondary peaks in the modified SRP-PHAT function correspond to speech activity. Whereas the main peak is due to an actual active speaker on most occasions, the secondary peaks are often associated with noise produced by the users (table hits, object movements, etc.), which causes numerous false positives. Second, when a user is giving a presentation at the whiteboard, the algorithm can detect that someone is speaking at the whiteboard area but it cannot determine who is speaking. Both problems can be addressed by analysing the video signals provided by the close-view cameras. In the first case, visual techniques to detect speech activity can be used to validate secondary peaks. In the second case, the problem is solved by measuring motion in the image.

4.1 Video analysis

Video Vocal Activity Detection (VVAD) is a task that has been addressed in the literature in different ways. The methods in [29, 37] compute the optical flow of the mouth region, while [28, 35] use deformable models to track lip contours. In both cases, the video data must have sufficient resolution to provide a reliable VVAD measure, otherwise their applicability is reduced. The works in [5, 22] propose an approach based on measuring mouth movement using a model of the subject's skin appearance extracted from two patches located below the speaker's eyes and by segmenting the mouth area into skin and not skin pixels. This approach does not work very well when the subject does not expose his/her skin, such as when the user has a beard or wears glasses.

To perform voice activity detection in video data, we exploit the fact that a person moves his/her lips when he/she talks. We employ a simple and efficient method to detect lip movement. First, we use the classic Viola and Jones algorithm [39] (based on a cascade of Haar-like features) to perform face detection on each frame of the sequence. When a face is detected, we use a second detector of the same type on the lower third of its bounding box to determine the exact location of the subject's mouth. Then, we use the distance and overlapping of the bounding boxes to determine whether they correspond to the same face. When we detect the same face on two consecutive frames, we align the areas corresponding to the mouth and compute a similarity measure between them by using a normalized square difference:

$$L_{n,t} = \frac{\sum_{a,b}(I^m_{n,t}(a,b) - I^m_{n,t-1}(a,b))^2}{\sqrt{\sum_{a,b} I^m_{n,t}(a,b)^2 \cdot \sum_{a,b} I^m_{n,t-1}(a,b)^2}}, \tag{10}$$

where $I^m_{n,t}(a,b)$ and $I^m_{n,t-1}(a,b)$ are the sub-images corresponding to the mouth of subject n in two consecutive frames and (a,b) are the pixel coordinates. Lower values of $L_{n,t}$ correspond to high similarity (i.e., low activity) for the subject's mouth. Then, according to this measure, a speaker can be considered active if $L_{n,t} > L_{\min}$, where $L_{\min}$ is a decision threshold.

Our voice activity measure is defined only when we find the same face and its corresponding mouth in two consecutive frames. In other cases, we simply cannot detect any voice activity or its absence. A value near zero means that the subject is not talking (i.e., the mouth is still). A relatively high value means that the subject is moving his/her lips, and we can assume that he/she is talking.

Our voice activity measure is extremely fast to compute, achieves reasonable results with relatively low resolution video, and does not impose restrictions beyond the face detection itself. However, we must stress that we measure only the activity of the mouth area, which does not imply that the subject is talking. Moreover, if the face of the subject is not detected (for example, due to occlusions, face turning, etc.), our technique cannot determine the vocal activity of the subject. As we can see in Fig. 6, a correlation exists between the actual voice activity of the subject and our voice activity measure.

To address the problem of moving speakers, the motion cue can be used to determine whether a user has left his/her seat. The simplest motion filter computes the absolute difference between two subsequent frames. To introduce robustness against small movements, we average this difference over the last T frames, which results in the following motion measure:

$$g_{n,t} = \frac{1}{T}\sum_{i=0}^{T-1} |\mathbf{I}_{n,t-i} - \mathbf{I}_{n,t-i-1}|, \tag{11}$$

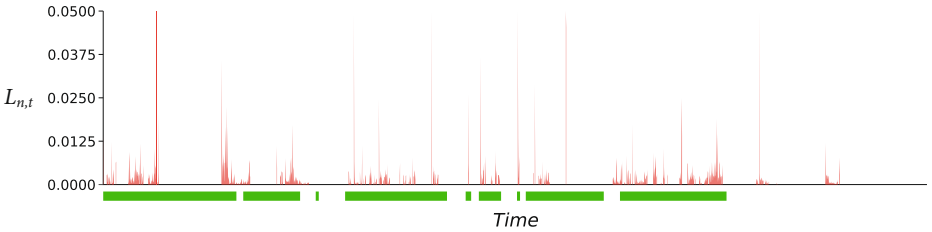

Fig. 6 Voice activity measure $L_{n,t}$ for a 120 second video fragment. The green bar below the time axis represents the ground truth of the sequence

where $\mathbf{I}_{n,t}$ is the image of user n at frame t. A value close to 0 indicates the absence of movement in the stream and hence that the user is not present in the image.

4.2 Decision algorithm

Once the values of v_t, $q_{i,t}$, $g_{n,t}$ and $L_{n,t}$ have been determined, it is possible to indicate the voice activity $d_{n,t}$ of each user. We denote the set of sectors corresponding to user n as $\mathcal{Q}_n$ and the sectors corresponding to the whiteboard area as $\mathcal{Q}_w$. The flowchart of the decision process performed for each user n is illustrated in Fig. 7. Essentially, for each detected voiced frame, speaker n is considered to be active if the maximum peak above $P_{\min}$ of the modified SRP-PHAT function (that is, $q_{1,t}$) falls within $\mathcal{Q}_n$. If $\mathcal{Q}_n$ contains only secondary peaks above $P_{\min}$, the system also checks the lip movement measure of user n against threshold $L_{\min}$ to filter out possible noises. Clearly, mouth movements do not always represent speech, and the system can be tricked if noise and lip movements coincide. The same occurs for non-speech vocal sounds, such as laughing or yawning. If one of the peaks falls within $\mathcal{Q}_w$, the active status is assigned to user n only if its close-view camera does not perceive any movement.

Since the method gives results for only voiced speech frames, it is necessary to estimate unvoiced frames to complete the diarization process. According to previous studies [18], voiced speech is typically preceded by 300 ms and followed by 500 ms of unvoiced speech.

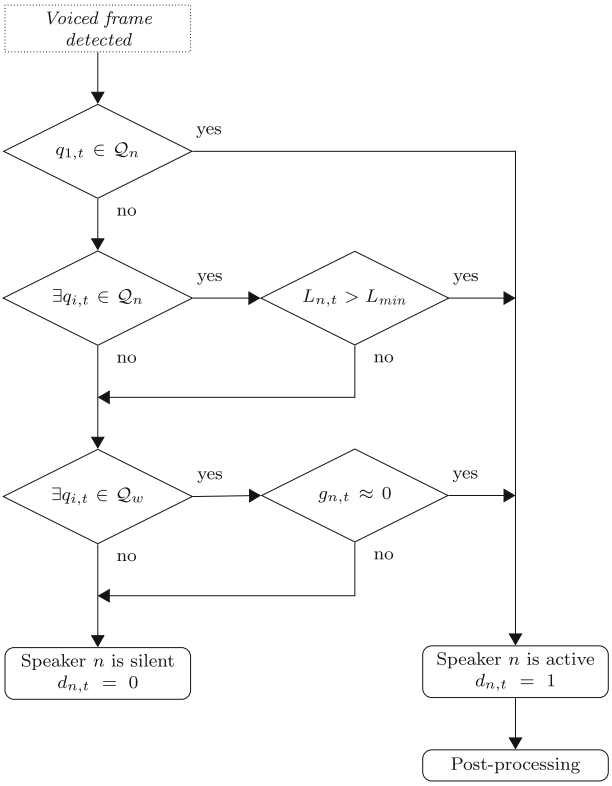

Fig. 7 Flowchart of the decision process for each user n

Therefore, for each active frame of user n, we extend the activation status to the immediately preceding and following frames. A benefit of this temporal extension is that the active status will not be affected by short speech pauses. In fact, for many practical applications, it is often desirable to detect speech hiatuses between words or sentences as active speech. Therefore, a post-processing stage can be included to filter out even longer speech pauses.

Except for the face and voicing detection modules, that are trained externally, the system is completely unsupervised and can be tuned by the parameters $P_{\min}$ and $L_{\min}$. To establish these parameters, a range of values were experimented with using the AMI dataset, and the combination that produced the minimum error was chosen. Given the size of the database, that contains three different rooms and multiple speakers, we assume that the final setting is generically applicable to other scenarios. In the literature, supervised approaches are often used for the fusion of audiovisual features using classifiers such as SVMs [30], decision trees [42] or HMMs [32]. Although these techniques can compete with unsupervised approaches in terms of classification speed, their performance heavily depends on the size and variety of the training data.

5 Experimental results

5.1 Experiment setup

All our experiments were conducted on the AMI dataset, which is composed of approximately 100 hours of multimodal meeting recordings with 4 participants. As explained in [9, 27], the data were captured in three different instrumented meeting rooms corresponding to different project sites that were denoted by the characters I (for Idiap), E (for Edinburgh) and T (for TNO). Although the rooms are roughly similar (all include a table, four chairs, a screen/whiteboard and recording equipment), they differ in their acoustic properties due to differences in overall shape and construction. As explained in Section 2, the meetings were recorded with a far-field microphone array and four cameras mounted at the centre of the table. Some of the meetings contained in the dataset naturally occur (denoted by character N), whereas others are elicited (denoted by S for a scenario in which the participants play different roles in a design team and B for other elicited scenarios). The meetings are not scripted, and the participants behave naturally while discussing concepts around a table or when presenting on a whiteboard.

The experimental evaluation consists of comparing the proposed diarization result to the ground-truth annotations included in the database. Recording sessions where participants interchange their seats, where there were technical failures (such as microphone fails) or in which people used the whiteboard in groups were excluded from the evaluation (since the system cannot handle those scenarios).

To compute the modified SRP-PHAT function, a frame length of 1280 samples (80 ms at a sampling rate of 16,000 Hz) with a 50% frame overlap was used. This setting corresponds to a rate of 25 SSL functions per second so that video and audio features are synchronized. The search space was discretized into $Q = 72$ sectors, each covering an angle of 5 degrees, where the limits of the sectors were set to $z_{\max} = 1$ m, $\rho_{\min} = 30$ cm and $\rho_{\max} = 3$ m. For each room, the annotation of the set of sectors $\mathcal{Q}_n$ corresponding to each user was manually performed by inspecting the SRP-PHAT response and the video signals. The voicing detector was run with the same temporal resolution as in the SSL step, and the remaining parameters were set to default values.

With this setting, the search space is characterized with only 72 sectors, which involves making only 72 evaluations of the modified SRP-PHAT function. This is an important advantage over conventional SRP-PHAT, as it has been demonstrated that the performance of SRP-PHAT substantially degrades when a small number of candidates is used. For a comprehensive comparison between both approaches in terms of localization accuracy for different grid resolutions, we refer the reader to [10].

The effectiveness of our SD algorithm was assessed in terms of diarization error rate (DER). According to this measure, a frame is considered misclassified if the global speaking activity is not correctly estimated. Therefore, this measure corresponds to the fraction of time that the diarization result is not correctly attributed to one or more speakers (or to non-speech during silent periods) [2].

5.2 Results

Our DER results for both audio-only and multimodal approaches are summarized in Tables 1 and 2, providing the average error for each meeting type. In the first case, only the audio features are used to determine the active speakers, such that a speaker is considered to be active if one of his/her sectors yields a functional value above $P_{\min}$ and a speech frame has been detected. In the second case, the video features are used to refine the audio results and to detect whether a person has moved to the whiteboard, as explained in Section 4 and Fig. 7. As noted above, meetings are divided according to the room they were recorded in (first letter) and the type of scenario (second letter). For each meeting type, the subset tagged with "whiteboard" comprises all recordings in which the whiteboard/screen was used during the session, whereas the other subset contains recordings in which the users remained seated. These tables also present the error percentage for each type of frame, divided according to the number of active speakers (one or more than one) annotated in the ground-truth data. These results allow us to evaluate the performance of the method in situations with overlapping speech and to compare them to frames with a single speaker.

Table 1 DER (%) per meeting room for audio-only diarization

	Audio-only		
	1 speaker	>1 speaker	all frames
ES	22.58	52.68	24.87
ES (whiteboard)	42.65	65.45	45.68
EN	21.41	48.15	27.16
IS	29.14	43.13	29.91
IS (whiteboard)	55.77	69.32	58.53
IN	14.58	45.35	20.04
IN (whiteboard)	27.30	49.05	33.19
IB	22.54	48.63	26.49
IB (whiteboard)	39.44	55.01	39.24
TS (whiteboard)	63.63	70.50	64.45
All	22.05	47.58	25.69
All (whiteboard)	45.75	61.86	48.21

The results for the meetings where the whiteboard is used are provided separately

Table 2 DER (%) per meeting room for multimodal diarization

	Multimodal		
	1 speaker	>1 speaker	all frames
ES	17.00	51.87	20.60
ES (whiteboard)	22.11	55.63	27.24
EN	14.63	47.14	21.88
IS	18.66	40.02	21.68
IS (whiteboard)	25.96	56.19	32.26
IN	11.32	43.24	17.13
IN (whiteboard)	14.62	40.37	21.63
IB	16.08	47.98	22.01
IB (whiteboard)	22.17	46.36	23.91
TS (whiteboard)	17.98	44.37	21.68
All	15.53	46.05	20.66
All (whiteboard)	20.56	48.58	25.34

The results for the meetings where the whiteboard is used are provided separately

As shown in Table 1, when a single user is speaking, the audio module is able to provide a correct diarization result in almost 78% of the frames. During overlapping speech moments, the audio algorithm correctly detects all active speakers in approximately half of the frames (47.58% of DER). Globally, the DER achieved by the audio-only method in meetings where the participants remained seated is equal to 25.69%. As expected, the result is much worse on the "whiteboard" subset, since the audio module cannot handle situations where users leave their positions.

As shown in Table 2, the percentage of errors is improved when the video information is employed to filter erroneous secondary peaks. This effect especially occurs in single-speaker frames, where the DER is 15.53% on the non-whiteboard subset. This improvement is not particularly noticeable in frames with overlapping speakers, but the results demonstrate that the filtering performed by the video module does not degrade the results. Overall, in recordings where the participants remain seated, the achieved multimodal DER is 20.6%. In sequences where the whiteboard/screen is used, the result is slightly worse, being approximately equal to 25%. In some "whiteboard" sequences, the presenting person occupies the sectors of other users in certain frames, which may explain the loss of performance in these sequences. As expected, the multimodal method clearly outperforms the audio-only approach in "whiteboard" meetings, demonstrating the benefits of multimodal analysis.

5.3 Computational cost

For the audio module, the majority of the cost comes from the localization algorithm. Let F be the number of frequency bins of the Fourier transform of a frame and $L = M(M-1)/2$ the number of microphone pairs. The computational cost of the SRP-PHAT function (in number of operations) is given by the following expression [11]:

$$C_a \approx \left(6.125L^2 + 3.75L\right) F \log_2 F + 15FL(1.5L - 1) + (45L^2 - 30L)v, \qquad (12)$$

where ν is the number of evaluations of the SRP-PHAT function. Since the number of operations added by the modified version is negligible, the above formula is valid for both approaches [10]. In our experiment, $F = 2048$ bins, $L = 28$ pairs, and $\nu = Q = 72$ evaluations, resulting in a cost of $C_a = 148.3$ millions of operations per frame. Since the method processes 25 frames per second, the number of operations is approximately 3707 million each second. The voicing detection algorithm does not dramatically increase complexity. As indicated in [19], for a Matlab implementation of the algorithm, the processing time is close to 0.2 s for each second of audio in an Intel Xeon CPU.

For the video module, we use the Viola-Jones face detection technique, followed by a mouth detection stage with the same algorithm, and a similarity computation. The Viola-Jones face detection algorithm [39] uses a cascade of classifiers with Haar-like features. By using the so-called *integral images*, whose computation is linear in the number of pixels of the image, this method can evaluate each filter with a small number of additions and subtractions, so its complexity can be expressed as:

$$C_v \approx NP + T \cdot NC$$

where NP is the number of pixels of the image, T is the number of image search windows, and NC is the average number of classifiers evaluated for each window. We start with 40×40 pixels search windows, increasing its size by a factor of 1.1. Giving that our images are of 352×288 pixels, $NP = 100000$ and $T = 900000$ approximately. NC is tipically 10, with each classifier requiring on average 8 single operations. Thus, face detection takes on average 72 million operations each frame (7200 million each second, assuming 4 cameras). Mouth localisation applied to detected faces and similarity computation takes on average 1 million operations per frame, giving that a single face of $60x60$ pixels is detected. The motion filter does not notably increase complexity, because it is based on a simple subtraction of frames.

Together, the SRP-PHAT function and the lip detection module take approximately 11000 millions of operations per second, which can be executed in real time by modern multi-core processors. The proposed algorithm was implemented in C and Matlab. Therefore, it is not fully optimized, although it is able to run faster than real-time in a machine with a Pentium i7 processor. Certain modules introduce a delay of a few frames (voicing detection and post-processing), but the system is still appropriate for real-time applications.

5.4 Comparison to other methods

Our SD system achieves results that are in line with other methods in the literature. However, providing a comparison in terms of SD performance is difficult since the results of a particular system can vary significantly among different meeting scenarios [2, 33]. Specifically, a number of important factors affect DER results, such as the acquisition setup, degree of speech overlap and spontaneity, number of speakers, available prior information, algorithm training conditions and real-time requirements. In practice, we can offer comparisons with methods evaluated on similar databases and analyse the strengths and weaknesses of each proposal. For example, Friedland et al. [14] used the IS subset of the AMI corpus to achieve a DER of 32.1% for the audio modality and 25.31% for the multimodal system (both working with one-channel audio). Their method is based on the clustering of acoustic and visual features. Therefore, it is robust to users changing their respective positions but cannot operate in real-time and does not handle overlapping speech. Based on similar principles, the ICSI audiovisual SD system described in [15] obtains a DER of 32.5% on the

challenging RT-09 dataset. Their multichannel audio-only system that incorporates beamforming and delay features achieves 17.2% for the offline version and 39.3% for the online version (which requires the first 1000 seconds for training). In the same dataset, the LIA-EURECOM system [13] that is also based on the clustering of acoustic features achieves a DER of 23.5%. Noulas et al. [31] used two meeting recordings from the AMI corpus (acquired in IDIAP and Edinburgh rooms) to achieve a DER of 33% and 20% for the audio modality and 16% and 11% for the multimodal fusion. Their probabilistic framework is able to work without making assumptions about the recording setup or the number/location of speakers, although it does not operate in real-time. The approach described in [20] performs audio-only SD in real-time by clustering the directions of arrival estimated with GCC-PHAT and using a statistical model-based VAD. The authors report a DER of 21.4% in a meeting session with 4 participants seated around a table.

Unlike the above methods based on timbral features, our localization-based algorithm has the advantage of working in real-time. Its main disadvantage is that it requires a priori knowledge of the area where each participant is expected to be seated. However, this requirement can be fulfilled in most practical scenarios for which the real-time processing is clearly more beneficial. Other localization-based methods, such as the one proposed in [42], do not have this disadvantage. However, unlike [42], our method does not require training and can handle simultaneous speakers. In contrast to the method in [20], our algorithm uses a volume-based localization that reduces computational complexity and employs the video to help in the diarization process. This approach allows one speaker to leave his/her seat to give a presentation on a whiteboard, unlike [20], where the users are required to remain seated.

6 Conclusion

This paper presents an approach for performing multimodal speaker diarization on meetings based on a modified SRP-PHAT function. Unlike conventional SRP-PHAT, this function is evaluated on space volumes, and when used in combination with a circular array topology, it is able to localize active speakers by extracting local maxima from the function. Visual lip movements are measured to discard potentially erroneous secondary peaks, and motion filtering is used to detect speakers walking to the whiteboard/screen. Voicing detection is used to discriminate speech frames from non-speech frames. The method requires a circular microphone array placed at the centre of the meeting space, cameras for capturing users' faces, and prior information about the candidate set of sectors where each user can be seated.

The experimental results were obtained with the challenging AMI dataset, obtaining a DER of approximately 21% – 25%. When using audio-only features, the DER is equal to 25% in sequences where the users remain seated. When using both video and audio cues, the DER is 21% in these same sequences because the system is able to correct certain decisions of the audio module. These results are in line with other SD algorithms in the literature applied in similar conditions. However, unlike many previous works, our system is able to run faster than real-time, is very efficient (thus requiring less computational resources) and correctly detects a significant portion of the frames that contain overlapping speech.

There are a number of ways in which our diarization system could be improved. For example, video could be used to track the speakers around the room, allowing higher freedom of movement to the participants and probably requiring additional cameras to capture the whole room. The audio module could be improved by using beamforming techniques

driven by the location information provided by both the SRP-PHAT and video modules, thus producing a separate signal for each user. The enhanced signals could be used to detect more precisely speech activity, at the cost of increasing computation time. Furthermore, the room could be equipped with several arrays distributed around the room that work in parallel to choose the best signal.

Acknowledgements This work was supported by the Andalusian Economy and Knowledge Council under project 2010-TIC6762, and the Spanish Ministry of Economy and Competitiveness under project TEC2015-67387-C4-2-R.

References

1. Ajmera J, Lathoud G, McCowan L (2004) Clustering and segmenting speakers and their locations in meetings. In: IEEE international conference on acoustics, speech, and signal processing (ICASSP), vol 1, pp 605–608
2. Anguera X, Bozonnet S, Evans N, Fredouille C, Friedland G, Vinyals O (2012) Speaker diarization: a review of recent research. IEEE Trans Audio Speech Lang Process 20(2):356–370
3. Araki S, Hori T, Fujimoto M, Watanabe S, Yoshioka T, Nakatani T, Nakamura A (2010) Online meeting recognizer with multichannel speaker diarization. In: 44th ASILOMAR conference on signals, systems and computers, pp 1697–1701
4. Araki S, Okada M, Higuchi T, Ogawa A, Nakatani T (2016) Spatial correlation model based observation vector clustering and MVDR beamforming for meeting recognition. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 385–389
5. Aubrey A, Rivet B, Hicks Y, Girin L, Chambers J, Jutten C (2007) Two novel visual voice activity detectors based on appearance models and retinal filtering. In: 15th european signal processing conference (EUSIPCO), pp 2409–2413
6. Bergh TF, Hafizovic I, Holm S (2016) Multi-speaker voice activity detection using a camera-assisted microphone array. In: 23rd international conference on systems, signals and image processing (IWSSIP), pp 1–4
7. Biagetti G, Crippa P, Falaschetti L, Orcioni S, Turchetti C (2016) Robust speaker identification in a meeting with short audio segments, pp 465–477. Springer International Publishing, Cham
8. Blauth DA, Minotto VP, Jung CR, Lee B, Kalker T (2012) Voice activity detection and speaker localization using audiovisual cues. Pattern Recogn Lett 33(4):373–380
9. Carletta J, Ashby S, Bourban S, Flynn M, Guillemot M, Hain T, Kadlec J, Karaiskos V, Kraaij W, Kronenthal M, Lathoud G, Lincoln M, Lisowska A, McCowan I, Post W, Reidsma D, Wellner P (2005) The AMI meeting corpus: a pre-announcement. In: International workshop on machine learning for multimodal interaction. Springer, pp 28–39
10. Cobos M, Marti A, Lopez JJ (2011) A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling. IEEE Signal Processing Letters 18(1):71–74
11. DiBiase JH (2000) A high-accuracy, low-latency technique for talker localization in reverberant environments. Ph.D. thesis, Brown University, Providence, RI
12. Do H, Silverman HF, Yu Y (2007) A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array. In: IEEE International conference on acoustics, speech and signal processing (ICASSP), vol 1, pp 121–124
13. Fredouille C, Bozonnet S, Evans N (2009) The LIA-EURECOM RT'09 speaker diarization system. In: RT'09 NIST Rich transcription workshop, vol 15, pp 17–23
14. Friedland G, Hung H, Yeo C (2009) Multi-modal speaker diarization of real-world meetings using compressed-domain video features. In: IEEE International conference on acoustics, speech and signal processing (ICASSP), pp 4069–4072
15. Friedland G, Janin A, Imseng D, Anguera X, Gottlieb L, Huijbregts M, Knox MT, Vinyals O (2012) The ICSI RT-09 speaker diarization system. IEEE Trans Audio Speech Lang Process 20(2):371–381
16. Fujimoto M, Ishizuka K, Nakatani T (2009) A study of mutual front-end processing method based on statistical model for noise robust speech recognition. In: 10Th annual conference of the international speech communication association (INTERSPEECH), pp 1235–1238

17. Gebru I, Ba S, Li X, Horaud R (2017) Audio-visual speaker diarization based on spatiotemporal bayesian fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2017.2648793
18. Ghaemmaghami H, Baker BJ, Vogt RJ, Sridharan S (2010) Noise robust voice activity detection using features extracted from the time-domain autocorrelation function. In: 11th annual conference of the international speech communication association (INTERSPEECH), pp 3118–3121
19. Gonzalez S, Brookes M (2014) PEFAC - a pitch estimation algorithm robust to high levels of noise. IEEE/ACM Trans Audio Speech Lang Process 22(2):518–530
20. Hori T, Araki S, Yoshioka T, Fujimoto M, Watanabe S, Oba T, Ogawa A, Otsuka K, Mikami D, Kinoshita K, Nakatani T, Nakamura A, Yamato J (2012) Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera. IEEE Trans Audio Speech Lang Process 20(2):499–513
21. Hung H, Friedland G (2008) Towards audio-visual on-line diarization of participants in group meetings. In: Workshop on multi-camera and multi-modal sensor fusion algorithms and applications
22. Liu Q, Wang W, Jackson P (2011) A visual voice activity detection method with adaboosting. In: Sensor signal processing for defence (SSPD), pp 1–5
23. Liu Y, Nie L, Han L, Zhang L, Rosenblum DS (2015) Action2activity: Recognizing complex activities from sensor data. In: International joint conference on artificial intelligence (IJCAI), pp 1617–1623
24. Liu Y, Zhang L, Nie L, Yan Y, Rosenblum DS (2016) Fortune teller: Predicting your career path. In: Proceedings of the AAAI conference on artificial intelligence, pp 201–207
25. Liu Y, Zheng Y, Liang Y, Liu S, Rosenblum DS (2016) Urban water quality prediction based on multi-task multi-view learning. In: International joint conference on artificial intelligence (IJCAI)
26. Marti A, Cobos M, Lopez JJ (2011) Real time speaker localization and detection system for camera steering in multiparticipant videoconferencing environments. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 2592–2595
27. McCowan I, Carletta J, Kraaij W, Ashby S, Bourban S, Flynn M, Guillemot M, Hain T, Kadlec J, Karaiskos V, Kronenthal M, Lathoud G, Lincoln M, Lisowska A, Post W, Reidsma D, Wellner P (2005) The AMI meeting corpus. In: 5th international conference on methods and techniques in behavioral research, pp 137–140
28. Minotto VP, Lopes CBO, Scharcanski J, Jung CR, Lee B (2013) Audiovisual voice activity detection based on microphone arrays and color information. IEEE Journal of Selected Topics in Signal Processing 7(1):147–156
29. Minotto VP, Jung CR, Lee B (2014) Simultaneous-speaker voice activity detection and localization using mid-fusion of svm and hmms. IEEE Trans Multimedia 16(4):1032–1044
30. Minotto VP, Jung CR, Lee B (2015) Multimodal multi-channel on-line speaker diarization using sensor fusion through SVM. IEEE Trans Multimedia 17(10):1694–1705
31. Noulas A, Englebienne G, Krose BJ (2012) Multimodal speaker diarization. IEEE Trans Pattern Anal Mach Intell 34(1):79–93
32. Rozgic V, Han KJ, Georgiou PG, Narayanan S (2010) Multimodal speaker segmentation and identification in presence of overlapped speech segments. Journal of Multimedia 5(4):322–331
33. Sarafianos N, Giannakopoulos T, Petridis S (2016) Audio-visual speaker diarization using fisher linear semi-discriminant analysis. Multimed Tools Appl 75(1):115–130
34. Schmalenstroeer J, Kelling M, Leutnant V, Haeb-Umbach R (2009) Fusing audio and video information for online speaker diarization. In: 10th annual conference of the international speech communication association (INTERSPEECH), pp 1163–1166
35. Scott D, Jung CR, Bins J, Said A, Kalker A (2009) Video based VAD using adaptive color information. In: 11Th IEEE international symposium on multimedia, pp 80–87
36. Soldi G, Beaugeant C, Evans N (2015) Adaptive and online speaker diarization for meeting data. In: 23Rd european signal processing conference (EUSIPCO), pp 2112–2116
37. Tiawongsombat P, Jeong MH, Yun JS, You BJ, Oh SR (2012) Robust visual speakingness detection using bi-level HMM. Pattern Recogn 45(2):783–793
38. Vaquero C, Vinyals O, Friedland G (2010) A hybrid approach to online speaker diarization. In: 11Th annual conference of the international speech communication association (INTERSPEECH), pp 2638–2641
39. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: Computer vision and pattern recognition (CVPR), vol 1, pp 511–518
40. Wellner P, Flynn M, Guillemot M (2004) Browsing recorded meetings with Ferret. In: International workshop on machine learning for multimodal interaction. Springer, pp 12–21

41. Wooters C, Huijbregts M (2008) The ICSI RT07s speaker diarization system. In: Multimodal technologies for perception of humans: International evaluation workshops CLEAR 2007 and RT 2007. Springer, pp 509–519
42. Zhang C, Yin P, Rui Y, Cutler R, Viola P (2006) Boosting-based multimodal speaker detection for distributed meetings. In: IEEE 8Th workshop on multimedia signal processing (MMSP), pp 86–91
43. Zhang C, Zhang Z, Florencio D (2007) Maximum likelihood sound source localization for multiple directional microphones. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), vol 1, pp 125–128
44. Zhang C, Florencio D, Ba DE, Zhang Z (2008) Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings. IEEE Trans Multimedia 10(3):538–548

P. Cabañas-Molero was born in Ciudad Real, Spain, in 1983. He received the M.S. and Ph.D. degrees in Telecommunication Engineering from the University of Jaén, Spain, in 2008 and 2016, respectively.

Currently, he is a post-doctoral Researcher at the Telecommunication Engineering Department, University of Jaén. His areas of research interest include Sound Source Separation, Automatic Sound Classification and Speech Processing.

M. Lucena received the MS and PhD degrees both in Informatics from the University of Granada in 1994 and 2003, respectively. Since 1994 he has been with the Informatics Department at University of Jaen (Spain). Currently he teaches at the Polytechnic Engineering School as an Assistant. He is member of the IAPR association.

His current interest includes Image Retrieval from Databases, Feature extraction from Colour Images, Tracking Algorithms and 3D data processing.

J. M. Fuertes received the MS and PhD degrees both in Informatics from the University of Granada in 1992 and 1999, respectively. Since 1992 he has been with the Informatics Technology Department at University of Jaen (Spain). Currently he teaches at the Polytechnic Engineering School as permanent associate professor in Computer Science and Artificial Intelligence. He is member of the IAPR association.

His current interest includes Image Retrieval from Databases, Feature extraction from Colour Images, Tracking Algorithms, Computerized Archaeology.

P. Vera-Candeas was born in Madrid, Spain, in 1976. He received the M.Sc. degree in Telecommunication Engineering from the University of Málaga (UMA), Málaga, Spain, in 2000 and the Ph.D. degree from the University of Alcalá, Alcalá de Henares, Spain, in 2006. Since 2000, he has been with the Telecommunication Engineering Department, University of Jaén. Currently, he is an Associate Professor in Signal Processing and Communications Area.

His areas of research interest are Signal Processing and its Applications to Audio Analysis and Ultrasonic NDT. He has been involved in research projects of the Spanish Ministry of Science and Education (MEC) and private companies.

N. Ruiz-Reyes was born in Linares (Jaén), Spain, in 1967. He received the M.Sc. degree in Telecommunication Engineering from the Technical University of Madrid (UPM), Madrid, Spain, in 1993 and Ph.D. degree in Telecommunication Engineering from the University of Alcalá, Alcalá de Henares, Spain, in 2001. Since 1998, he has been with the Telecommunication Engineering Department, University of Jaén. Currently, he is a Professor in the Signal Processing and Communications Area.

His areas of research interest are Signal Processing and its Applications to Communications, Speech and Audio Analysis, Electrical and Biomedical Engineering, and Ultrasonic NDT. He is coauthor of about 150 papers, and is involved in research projects of the Spanish Ministry of Science and Education, European Commission, and private companies.