# Audio-Visual Speech-Turn Detection and Tracking*

Israel D. Gebru, Silèye Ba, Georgios Evangelidis, and Radu Horaud

INRIA Grenoble Rhône-Alpes, Montbonnot Saint-Martin, FRANCE

**Abstract.** Speaker diarization is an important component of multi-party dialog systems in order to assign speech-signal segments among participants. Diarization may well be viewed as the problem of detecting and tracking speech turns. It is proposed to address this problem by modeling the spatial coincidence of visual and auditory observations and by combining this coincidence model with a dynamic Bayesian formulation that tracks the identity of the active speaker. Speech-turn tracking is formulated as a latent-variable temporal graphical model and an exact inference algorithm is proposed. We describe in detail an audio-visual discriminative observation model as well as a state-transition model. We also describe an implementation of a full system composed of multi-person visual tracking, sound-source localization and the proposed online diarization technique. Finally we show that the proposed method yields promising results with two challenging scenarios that were carefully recorded and annotated.

**Keywords:** Speaker diarization, audio-visual fusion, sound-source localization, multi-person tracking, temporal graphical models.

## 1 Introduction

In human-computer interaction (HCI) and human-robot interaction (HRI) it is often necessary to solve the multi-party dialog problem. For example, if two or more persons are engaged in a discussion, one important task to be solved, prior to automatic speech recognition (ASR) and natural language processing (NLP), is to correctly assign speech segments among the participants. In the speech and language processing literatures, this problem is often referred to as speaker diarization and a number of methods has been recently proposed to solve this problem, e.g., [1]. When only auditory data are available, the task is very difficult because of the inherent ambiguity of mixed acoustic signals captured by the microphones. An interesting alternative consists in fusing auditory and visual data. The two modalities provide complementary information and hence audio-visual approaches to speaker diarization are likely to be more robust than audio-only approaches.

An audio-visual diarization method was recently proposed [7] where the hidden (latent) discrete variables represent the speaker identity and the speaker visibility at time

* Support from EU-FP7 ERC AdG VHIA (#340113) and STREP EARS (#609645) is greatly acknowledged.

$t$. The main limitation of [7] as well as of other audio-visual approaches reviewed in [1] is that these methods require the recognition of frontal faces and characterization of lip motions. Indeed, audio-visual association is often solved using the temporal correlation between facial features and audio features [8].

More generally, audio-visual association for speaker diarization can be achieved on the premise that a speech signal *coincides* with a person that is visible and that emits a sound. This coincidence must occur both in space and time. In formal dialogs, diarization is facilitated by the fact that the participants talk sequentially, that there is a short silence between speech turns, and that the participants face the cameras and are static or remain seated. In these cases, audio-visual association based on temporal coincidence seems to provide satisfactory results, e.g., [5]. In informal settings, which are very common particularly in HRI, the situation is much more complex. The perceived audio signals are corrupted by environmental noise, reverberations, and several people may occasionally speak simultaneously. People wander around, turn their heads away from the sensors, and come in and out of the fields of view of the cameras.

These problems were addressed by several authors in different ways. For example, [4] proposes a multi-speaker tracker using approximate inference implemented with a Markov chain Monte Carlo particle filter (MCMC-PF). [6] uses a 3D visual tracker, based on MCMC-PF as well, to estimate the positions and velocities of the participants which are then passed to a blind source separation method. This provides a proof of concept benchmark for moving speakers. MCMC-PF tracking cannot easily handle a varying number of speakers. Moreover, the reported experiments in both [4] and [6] are carried out with a microphone array and several cameras to guarantee that frontal views of the speakers are permanently visible. They do not specifically address speaker diarization which is a difficult problem in its own right.

In this paper it is proposed to enforce spatial coincidence into diarization. We consider a setup consisting of participants that are engaged in a multi-party dialog while they are allowed to move and to turn their attention away from the cameras. We propose to combine an online multi-person visual tracker [2], with a voice activity detector [9], and a sound-source localizer [3] . Assuming that the image and audio sequences are synchronized, we propose to group auditory features and visual features on the premise that they share a common location if they are generated by the same speaker. We define a speech-turn latent variable and we devise an online tracker such that at each frame $t$ the identity of the active-speaker is estimated. We propose a discriminative observation model that evaluates the posterior probability of speech turns, conditioned by the outputs of a multi-person visual tracker, a sound-source localizer, and a voice activity detector. We also propose a dynamic model that allows to estimate the transition probabilities, from $t-1$ to $t$, of the speech-turn variable. The proposed online speech-turn tracking method uses an efficient exact inference algorithm.

The remainder of this paper is organized as follows. Section 2 formally describes the proposed exact inference method; section 2.1 describes the audio-visual discriminative observation model, section 2.2 describes the proposed transition probabilities model. Section 3 describes implementation details and experiments. Finally, section 4 draws some conclusions.

## 2 A Graphical Model for Tracking Speech Turns

We start by introducing some notations and definitions. Upper-case letters denote random variables while lower-case letters denote realizations of random variables. We consider an image sequence that is synchronized with an audio sequence and let $t$ denote the temporal index of both visual and audio frames. There are at most $N$ visual observations at frame $t$, $\mathbf{X}_t = (\boldsymbol{X}_{t1}, \ldots, \boldsymbol{X}_{tn}, \ldots, \boldsymbol{X}_{tN}) \in \mathbb{R}^{2\times N}$, where the random variable $\boldsymbol{X}_{tn}$ corresponds to the location of person $n$ in image $t$. Then, a multi-person tracker, e.g., [2] (section 3) provides a time series of $N$ image locations, namely $\mathbf{X}_{1:t} = \{\mathbf{X}_1, \ldots, \mathbf{X}_t\}$ and associated *visual-presence binary masks* $\mathbf{V}_{1:t}$, namely variable $V_{tn}$ associated with $\boldsymbol{X}_{tn}$ such that $V_{tn} = 1$ if person $n$ is present in image $t$ and 0 otherwise. Hence $N_t = \sum_n V_{tn}$ denotes the number of persons that are present (observed) at $t$. We also consider an audio-source localizer that provides the azimuth and elevation of the dominant sound source at each audio frame $t$, e.g., [3] (section 3). The sound-source location can then be mapped onto the image plane, such that an azimuth-elevation pair of observations is transformed into an image location modeled by a random variable $\boldsymbol{Y}_t \in \mathbb{R}^2$. Sound-source localization (SSL) together with voice activity detection (VAD) provide a time series of sound locations $\mathbf{Y}_{1:t} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_t\}$ and associated *speech-activity binary masks* $\mathbf{A}_{1:t} = \{A_1, \ldots, A_t\}$, such that $A_t = 1$ if there is an active audio source at frame $t$ and 0 otherwise.

The objective is to track the active speaker which amounts to associate over time the audio activity (if any) with one of the tracked persons. This is also referred to as audio-visual speaker diarization, e.g., [7] which is addressed below in the framework of temporal graphical models; A time-series of discrete latent variables is introduced, $\mathbf{S}_{1:t} = \{S_1, \ldots, S_t\}$ such that $S_t = n, n \in \{1, 2, \ldots, N\}$ if person $n$ is both observed and speaks at frame $t$, and $S_t = 0$ if none of the visible persons speaks at frame $t$. Notice that $S_t = 0$ encompasses two cases: first, there is an active sound-source at $t$ ($A_t = 1$) but its location cannot be associated with one of the visible persons and second, there is no active sound-source at $t$ ($A_t = 0$). The active-speaker or, equivalently, speech-turn tracker can be formulated as a maximum a posteriori (MAP) estimation problem:

$$\hat{s}_t = \underset{s_t}{\operatorname{argmax}}\, P(S_t = s_t | \boldsymbol{x}_{1:t}, \boldsymbol{v}_{1:t}, \boldsymbol{y}_{1:t}, \boldsymbol{a}_{1:t}). \tag{1}$$

The posterior probability (1) can be written as:

$$P(S_t = s_t | \boldsymbol{u}_{1:t}) = \frac{P(\boldsymbol{u}_t | S_t = s_t, \boldsymbol{u}_{1:t-1}) P(S_t = s_t | \boldsymbol{u}_{1:t-1})}{P(\boldsymbol{u}_t | \boldsymbol{u}_{1:t-1})}, \tag{2}$$

where we used the notation $\boldsymbol{u}_t = (\boldsymbol{x}_t, \boldsymbol{v}_t, \boldsymbol{y}_t, a_t)$. This can be further developed as:

$$P(S_t = s_t | \boldsymbol{u}_{1:t}) = \frac{P(\boldsymbol{u}_t | S_t = s_t) \sum_{i=0}^{N} P(S_t = s_t | S_{t-1} = i) P(S_{t-1} = i | \boldsymbol{u}_{1:t-1})}{\sum_{j=0}^{N} (P(\boldsymbol{u}_t | S_t = j) (\sum_{i=0}^{N} P(S_t = j | S_{t-1} = i) P(S_{t-1} = i | \boldsymbol{u}_{1:t-1})))} \tag{3}$$

The evaluation of this recursive relationship requires the observed likelihood $P(\boldsymbol{u}_t|S_t = s_t)$ and the transition probabilities $P(S_t = j|S_{t-1} = i)$. Because the number of persons that are simultaneously present is small (3-5 persons), the exact evaluation of (3) is tractable and hence the MAP estimator (1) is straightforward.

### 2.1 Audio-Visual Observation Model

One crucial feature of the proposed model is its ability to robustly associate the acoustic activity at time $t$ with a person. The generative model that is proposed below assigns the audio activity, if any, to a person, or to nobody. In this context, the state variable $S_t$ plays the role of an assignment variable in a mixture model. If a sound-source is active at time $t$, ($A_t = 1$) its location $\boldsymbol{y}_t$ is assumed to be drawn from the following Gaussian/uniform mixture:

$$P(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{v}_t, A_t = 1; \boldsymbol{\theta}_t) = \sum_{n=1}^{N} \pi_{tn} v_{tn} \mathcal{N}(\boldsymbol{y}_t|\boldsymbol{x}_{tn}, \boldsymbol{\Sigma}_{tn}) + \pi_{t0}\mathcal{U}(\beta_t), \tag{4}$$

where $\boldsymbol{\theta}_t = (\{\pi_{tn}\}_{n=0}^{N}, \{\boldsymbol{\Sigma}_{tn}\}_{n=1}^{N}, \beta_t)$ denotes the set of model parameters, namely the priors $\pi_{tn} = P(S_t = n), \pi_{t0} + \sum_{n=1}^{N} v_{tn}\pi_{tn} = 1$, the $2 \times 2$ covariance matrices $\boldsymbol{\Sigma}_{tn}$, and a parameter $\beta_t$ that characterizes the outlier component of the mixture, namely a uniform distribution. The posterior probability of a sound-source to be associated with the $n$-th visible person writes:

$$P(S_t = n|\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{v}_t, A_t = 1; \boldsymbol{\theta}_t) = \frac{\pi_{tn} v_{tn} \mathcal{N}(\boldsymbol{y}_t|\boldsymbol{x}_{tn}, \boldsymbol{\Sigma}_{tn})}{\sum_{k=1}^{N} \pi_{tk} v_{tk} \mathcal{N}(\boldsymbol{y}_t|\boldsymbol{x}_{tk}, \boldsymbol{\Sigma}_{tk}) + \pi_{t0}\mathcal{U}(\beta_t)}. \tag{5}$$

We can also write the posterior probability that a sound source is not associated with a visible person, either because it corresponds to a sound emitted by a non visible person or emitted by another type of source, i.e., the posterior of the uniform component of the mixture:

$$P(S_t = 0|\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{v}_t, A_t = 1; \boldsymbol{\theta}_t) = \frac{\pi_{t0}\mathcal{U}(\beta_t)}{\sum_{k=1}^{N} \pi_{tk} v_{tk} \mathcal{N}(\boldsymbol{y}_t|\boldsymbol{x}_{tk}, \boldsymbol{\Sigma}_{tk}) + \pi_{t0}\mathcal{U}(\beta_t)}. \tag{6}$$

If there is no audio activity at time $t$ ($A_t = 0$), the posterior can be evaluated with the following formula, where $r$ is a small positive scalar, e.g., $r = 0.2$:

$$P(S_t = 0|\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{v}_t, A_t = 0; r) = \begin{cases} r/N_t \text{ if } 1 \leq n \leq N \\ 1 - r \text{ if } n = 0. \end{cases} \tag{7}$$

Finally, by assuming a uniform distribution over the priors of visible person $n$ ($v_{tn} = 1$), i.e., $\pi_{t0} = \pi_{tn} = 1/(N_t + 1)$, and by remarking that the observed-data likelihood $P(\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{v}_t, a_t)$ does not depend on $S_t$, we obtain the following observation model:

$$P(\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{v}_t, a_t|S_t = n) \propto P(S_t = n|\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{v}_t, a_t). \tag{8}$$

### 2.2 State Transition Model

The state transition probabilities, $p(\boldsymbol{S}_t = j|\boldsymbol{S}_{t-1} = i)$, provide a temporal model for for tracking speech turns. Several cases need be considered based on the presence/absence of persons and on their speaking status (for convenience and without loss of generality we set $v_{t0} = 1$):

$$p(\boldsymbol{S}_t = j|\boldsymbol{S}_{t-1} = i) = \begin{cases} p_s & \text{if } i = j \text{ and } v_{t-1i} = v_{ti} = 1 \\ (1-p_s)/N_t & \text{if } i \neq j \text{ and } v_{t-1i} = v_{tj} = 1 \\ 0 & \text{if } v_{t-1i} = v_{t-1j} = 1 \text{ and } v_{tj} = 0 \\ 1/N_t & \text{if } v_{t-1i} = 1, v_{ti} = 0 \text{ and } v_{tj} = 1 \\ 1/N & \text{if } v_{t-1i} = 0 \text{ and } v_{ti} = 0. \end{cases} \quad (9)$$

The first case of (9) defines the self-transition probability, $p_s$, e.g., $p_s = 0.8$, of person $i$ present at both $t-1$ and $t$. The second case defines the transition probability from person $i$ present at $t-1$ to another person $j$ present at $t$. The third case simply forbids transitions from person $i$ present at $t-1$ to person $j$ present at $t-1$ but not present at $t$. The fourth case defines the transition probability from person $i$ present at $t-1$ but not present at $t$, to a person $j$ present at $t$. The fifth case defines the transition probability from person $i$ not present at $t-1$ to person $j$ that is not present at $t$. These five cases can be grouped in a compact way to yield the state transition probability matrix ($\delta_{ij} = 1$ if $i = j$ and 0 otherwise):

$$p(\boldsymbol{S}_t = j|\boldsymbol{S}_{t-1} = i) = v_{t-1i}v_{tj}\left(p_s\delta_{ij} + \frac{(1-p_s)(1-\delta_{ij})}{N_t} + \frac{1-v_{ti}}{N_t}\right) + \frac{1-v_{t-1i}}{N}. \quad (10)$$

One may easily verify that $\sum_{j=1}^{N} p(\boldsymbol{S}_t = j|\boldsymbol{S}_{t-1} = i) = 1$.

## 3 Implementation and Experiments

As already outlined, the speech-turn tracking method that we propose in this paper may well be viewed as a speaker diarization process – track several persons in parallel, estimate their auditory status, and assign a speech segment to the dominant speaker. Unlike existing audio-visual diarization approaches, which only consider the *temporal coincidence* of the two modalities and which assume that the participants are always visible by the cameras, the proposed method enforces *spatial coincidence* such that it can deal with participants that are temporarily occluded, or who come in and out of the field of view of the cameras. Unfortunately there are no publicly available datasets that would allow us to test the robustness of our method in the presence of moving/occluding participants and to compare it with other methods. The only datasets currently available correspond to formal meetings where the participants are seated and are permanently facing the cameras. Benchmarking against existing approaches was not possible because other methods do not cope with the audio-visual alignment issue.

Therefore we recorded our own data, gathered with two microphones and one camera [3]. The two modalities are synchronized such that video frames are temporally aligned with audio frames. Hence the frame index $t$ is shared by the two modalities. We gathered two scenarios, the *counting* scenario (fig. 1) and the *chat* scenario (fig. 2). The videos are recorded at 25 FPS while the audio signals are sampled at 48000 Hz. Hence a video frame is 40 ms long. To ensure temporal synchronization between the two modalities, we define 40 ms audio frames in the following way. An audio frame is composed of several 64 ms windows shifted by 8 ms. Hence, a 40 ms audio frame is composed of 5 consecutive windows that partially overlap. The *counting* sequence has 500 frames (20 seconds) while the *chat* sequence has 850 frames (34 seconds).

We briefly describe the multi-person tracking and sound-source localization techniques used to gather values for the observed auditory and visual variables, i.e., section 2.1. Among the visual tracking methods that are currently available, we chose the multi-person tracker of [2]. This method has several advantages, namely (i) it robustly handles fragmented tracks, which are due to occlusions or to unreliable detections, and (ii) it performs online discriminative learning to handle similar appearances of different persons. The multi-person tracker provides values of the visual observation variables $\mathbf{X}_{1:t}$ and associated *visual-presence binary masks* $\mathbf{V}_{1:t}$, as explained in detail in section 2.

Sound localization consists in finding the direction of arrival (DOA) of an acoustic signal from multi-microphone recordings. We adopted the method of [3] that estimates the DOA of a sound with two degrees of freedom (azimuth and elevation) using a binaural acoustic dummy head. A prominent advantage of this method over other DOA

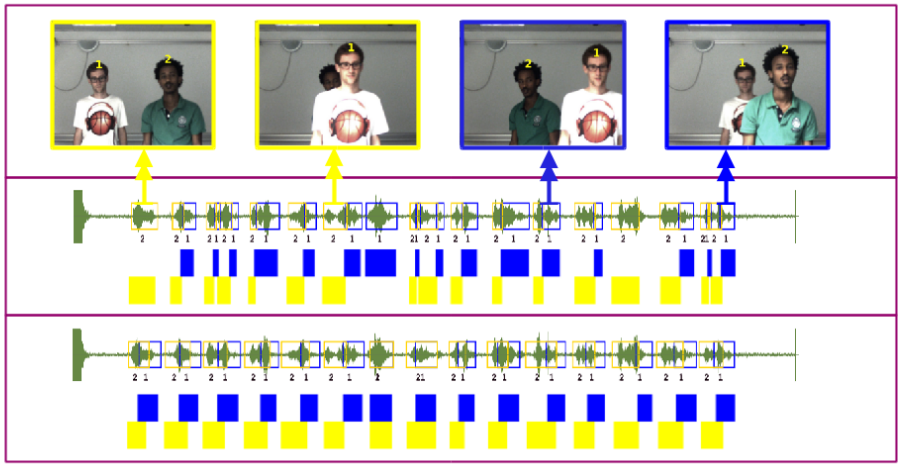

Fig. 1: The *counting* sequence involves two moving persons that occasionally occlude each other (top). Diarization results (middle) are also illustrated with a color diagram. Ground-truth diarization (bottom); notice that there is a systematic overlap between the two speech signals.

methods is that it provides a built-in mechanism for directly mapping a binaural feature vector associated with an audio frame, onto an image location. Hence a DOA associated with a sound source in expressed in pixel coordinates. In practice, the STFT is applied to 64 ms windows of the left and right microphone signals and a complex-valued binaural feature vector is built for each window using the ILD (interaural level difference) and IPD (interaural phase difference). Then we apply the method of [3] to a short spectrogram, composed of five consecutive windows (roughly corresponding to a frame), to estimate a DOA for each audio frame. In combination with a voice activity detector (VAD), this provides a time series of realizations of both the sound location variables $\mathbf{Y}_{1:t}$ and the associated *speech-activity binary masks* $\mathbf{A}_{1:t}$, as detailed in section 2.

These auditory and visual observations are used to evaluate the likelihoods (8) and transition probabilities (9) which in turn are plugged into (3) to estimate the posteriors of the speech-status variable given the observations, $P(S_t = s_t|\boldsymbol{u}_{1:t})$. In all our experiments we used the following numerical values for the model's free parameters: $r = 0.2, p_s = 0.8, \mathbf{\Sigma} = \text{Diag}\,[60, 120], \beta = 300000$. The proposed value of $p_s$ achieves a good compromise between either assigning speech to the current speaker or jumping to another person. The parameters $\mathbf{\Sigma}$ and $\beta$ are expressed in pixels. The value of $\beta$ corresponds to a uniform distribution over an image of $640 \times 480$ pixels. The method yields 75% correct results for the *counting* sequence and 64% correct results for the *chat* sequence.

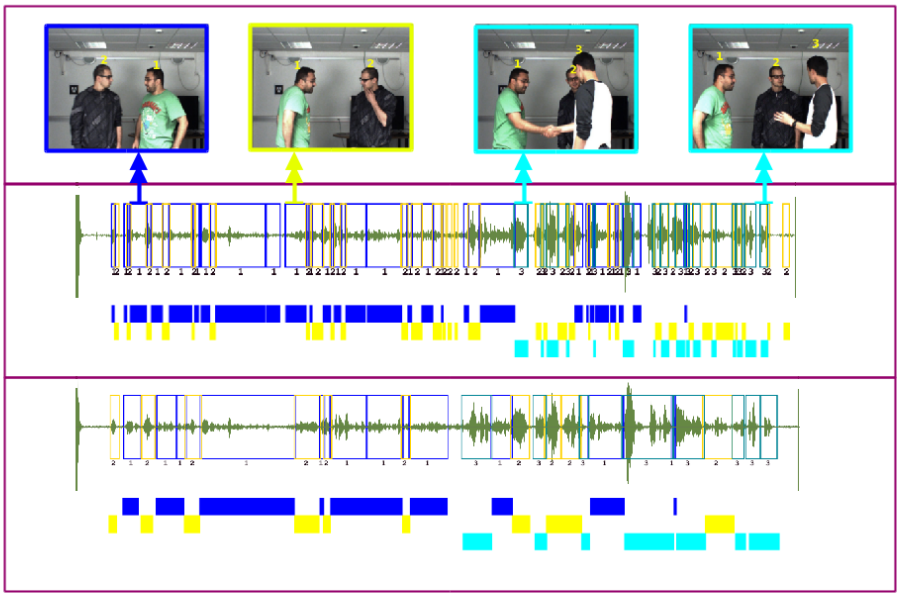

Fig. 2: The *chat* sequence involves two then three moving persons that take speech turns and that occasionally occlude each other (top). Diarization results (middle) and ground-truth (bottom); notice that in this case there is no speech overlap.

## 4 Conclusions

The paper addressed the problem of speaker diarization using auditory and visual data gathered with two microphones and one camera. Recent work in audio-visual diarization has capitalized on temporal coincidence of the two modalities, e.g., [1, 7]. In contrast, we propose a speech-turn detection and tracking method that enforces spatial coincidence, namely it materializes that a sound-source and associated visual-object should have the same spatial location. Consequently, it is possible to perform speaker localization by detecting persons in an image, localizing a sound source, mapping the sound-source location onto the image and associating the source location with one of the persons that are present in the image. Moreover, this process can be plugged into a latent-variable temporal graphical model that robustly tracks the identity of the active speaker. We described in detail the proposed method and illustrated its effectiveness with two challenging scenarios involving moving people, visual occlusions, and a reverberant room. In the future we plan to incorporate a more robust voice activity detector/tracker that is robust with respect to non-stationary acoustic event and to mixed speech signals.

## References

1. Anguera Miro, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., Vinyals, O.: Speaker diarization: A review of recent research. IEEE Transactions on Audio, Speech, and Language Processing 20(2), 356–370 (2012)
2. Bae, S.H., Yoon, K.J.: Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In: Computer Vision and Pattern Recognition. pp. 1218–1225 (2014)
3. Deleforge, A., Horaud, R., Schechner, Y.Y., Girin, L.: Co-localization of audio sources in images using binaural features and locally-linear regression. IEEE Transactions on Audio, Speech and Language Processing 23(4), 718–731 (2015)
4. Gatica-Perez, D., Lathoud, G., Odobez, J.M., McCowan, I.: Audiovisual probabilistic tracking of multiple speakers in meetings. IEEE Transactions on Audio, Speech and Language Processing 15(2), 601–616 (2007)
5. Kidron, E., Schechner, Y.Y., Elad, M.: Cross-modal localization via sparsity. IEEE Transactions on Signal Processing 55(4), 1390–1404 (2007)
6. Naqvi, S., Yu, M., Chambers, J.: A multimodal approach to blind source separation of moving sources. IEEE Journal of Selected Topics in Signal Processing 4(5), 895 –910 (2010)
7. Noulas, A., Englebienne, G., Krose, B.J.A.: Multimodal speaker diarization. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(1), 79–93 (2012)
8. Potamianos, G., Neti, C., Gravier, G., Garg, A., Senior, A.W.: Recent advances in the automatic recognition of audiovisual speech. Proceedings of the IEEE 91(9), 1306–1326 (2003)
9. Sohn, J., Kim, N.S., Sung, W.: A statistical model-based voice activity detection. IEEE Signal Processing Letters 6(1), 1–3 (1999)