Tracking the Active Speaker Based on a Joint Audio-Visual Observation Model

Israel Dejene Gebru, Silèye Ba, Georgios Evangelidis, Radu Horaud

▶ **To cite this version:**

Israel Dejene Gebru, Silèye Ba, Georgios Evangelidis, Radu Horaud. Tracking the Active Speaker Based on a Joint Audio-Visual Observation Model. ICCV Workshop on 3D Reconstruction and Understanding with Video and Sound , Dec 2015, Santiago, Chile. pp.702 - 708, 10.1109/ICCVW.2015.96 . hal-01220956

HAL Id: hal-01220956

https://hal.science/hal-01220956v1

Submitted on 27 Oct 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

Tracking the Active Speaker Based on a Joint Audio-Visual Observation Model

Israel D. Gebru , Silèye Ba, Georgios Evangelidis and Radu Horaud
INRIA Grenoble Rhône-Alpes, Montbonnot Saint-Martin, France

Abstract

Any multi-party conversation system benefits from speaker diarization, that is, the assignment of speech signals among the participants. We here cast the diarization problem into a tracking formulation whereby the active speaker is detected and tracked over time. A probabilistic tracker exploits the on-image (spatial) coincidence of visual and auditory observations and infers a single latent variable which represents the identity of the active speaker. Both visual and auditory observations are explained by a recently proposed weighted-data mixture model, while several options for the speaking turns dynamics are fulfilled by a multi-case transition model. The modules that translate raw audio and visual data into on-image observations are also described in detail. The performance of the proposed tracker is tested on challenging data-sets that are available from recent contributions which are used as baselines for comparison.

1. Introduction

In human-computer interaction (HCI) and human-robot interaction (HRI) it is often necessary to solve multi-party dialog problems. For example, if two or more persons are engaged in a conversation, one important task to be solved, prior to automatic speech recognition (ASR) and natural language processing (NLP), is to correctly assign speech segments to corresponding speakers. This problem is often referred to as speaker diarization in the speech/language processing literature and a number of solutions has been recently proposed, e.g., [2]. 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 combining 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.

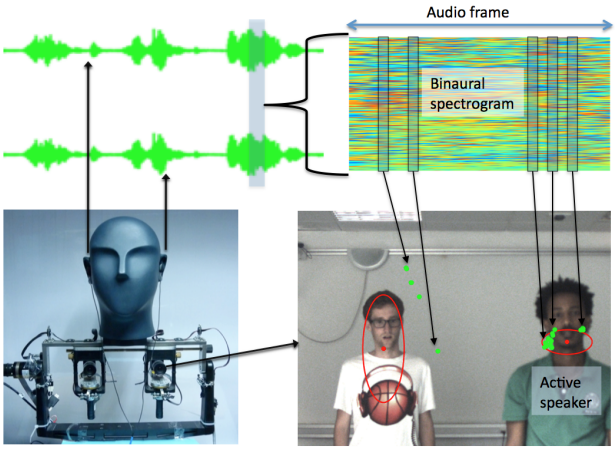

Figure 1: The auditory and visual data are recorded with two microphones and one camera. The audio signals are segmented into frames and each frame (vertical grey rectangle) is transformed into a binaural spectrogram. This spectrogram is composed of a sequence of binaural vectors (vertical rectangles) and each binaural vector is mapped onto a sound-source direction which corresponds to a point in the image plane (green dots). The proposed audio-visual tracker associates people detected in the image sequence with these sound directions via audio-visual clustering that is combined with an active-speaker transition model.

Several audio-visual diarization methods were recently proposed, e.g., [14, 2, 12]. Noulas et al. [14] proposed a graphical model, where latent discrete variables represent speaker identities and speaker visibilities over time. The main limitation of [14] as well as of other audio-visual approaches reviewed in [2] is that these methods require the detection of frontal faces and of mouth/lip motions. Indeed, audio-visual association is often solved using the temporal correlation, over several seconds, between facial features and audio features [15]. Minotto et al. [12] learn an SVM classifier using labeled audio-visual features, which is dependent on the acoustic properties of the training data. They combine voice activity detection with sound-source local-

ization using a linear microphone array. The latter can only provide the azimuth (horizontal) sound direction. Their method relies on mouth tracking, hence frontal views of the speakers are required as well.

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 multi-party conversations, diarization is facilitated by participants that talk sequentially, presence of a short silence between speech turns, and participants facing the cameras while remaining seated or static. In these cases, audio-visual association based on temporal coincidence seems to provide satisfactory results, e.g., [9]. 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 persons may occasionally speak simultaneously. Moreover, people may wander around, turn their heads away from the sensors, be occluded by other people, suddenly disappear from the camera field of view, and appear again later on.

These problems were addressed by several authors in different ways. For example, [5] proposed a multi-speaker tracker using approximate inference implemented with a Markov chain Monte Carlo particle filter (MCMC-PF). In [13] a 3D visual tracker is proposed, based on MCMC-PF as well, to estimate the positions and velocities of the participants which are then passed to blind source separation based on beamforming [19]. Reported experiments of both [5, 13] require a network of distributed cameras to guarantee that frontal views of the speakers are always available. More recently, [10] proposed to use audio information to assist the particle propagation process and to weight the observation model. This implies that audio data are always available and that they are reliable enough to properly relocate the particles. While audio-visual multiple persons tracking methods provide an interesting methodology, they do not address the challenging speaker diarization problem.

In this paper we propose to enforce audio-visual spatial coincidence, e.g., [1, 8, 10], rather than temporal coincidence, e.g., correlation [9, 16], into diarization. We consider a setup consisting of people that are engaged in a multi-party conversation while they are free to move and to turn their attention away from the cameras. We propose to combine an online multi-person visual tracker [3], with a voice activity detector [17], and a sound-source localizer [4], e.g., Fig. 1. Assuming that the image and audio sequences are synchronized, we propose to group auditory features and visual features based on the premise that they share a common location if they are generated by the same speaker. We introduce a latent variable representing the active-speaker, and we devise an on-line tracker such that the identity and location of the active speaker is estimated over time. We propose a generative observation model, based on the recently proposed weighted-data Gaussian mixture [6], that evaluates the posterior probability of an observed person to be the active speaker, conditioned by the output 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 active speaker using temporal transition probabilities modeling speaking activity transition priors from frame $t-1$ to frame t. The proposed on-line 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 generative observation model; section 2.2 describes the proposed transition probabilities model. Section 3 describes implementation details and experiments. Finally, section 4 draws some conclusions. Videos, Matlab code and additional examples are available online.[1]

2. Tracking the Active Speaker

We start by introducing some notations and definitions. Upper-case letters denote random variables while lower-case letters denote their realizations. We consider an image sequence that is synchronized with an audio sequence and let t denote the frame index of both visual and audio modalities (without loss of generality, one can assume that audio and visual frames have the same temporal length). Let N be the maximum number of visual observations at any time. Hence at frame t we have $\mathbf{X}_t = (\boldsymbol{X}_{t1}, \ldots, \boldsymbol{X}_{tn}, \ldots, \boldsymbol{X}_{tN}) \in \mathbb{R}^{2\times N}$, where the random variable $\boldsymbol{X}_{tn}$ corresponds to the location of person n at t. We also introduce the binary variables $\boldsymbol{V}_t = (V_{t1}, \ldots, V_{tN})$ such that $V_{tn} = 1$ if person n is detected *visible* in frame t and $V_{tn} = 0$ if the person is not detected. The time series $\mathbf{X}_{1:t} = \{\mathbf{X}_1, \ldots, \mathbf{X}_t\}$ and associated visual presence masks $\mathbf{V}_{1:t} = \{V_1, \ldots, V_t\}$ can be estimated using a multi-person tracker. We perform multi-person tracking using [3] (see section 3 below). Let $N_t = \sum_n V_{tn}$ denote the number of persons that are visible at t.

We also consider auditory information. Audio activity is described by the binary variable $A_t \in \{0, 1\}$ that is estimated using voice activity detection (VAD) and which is equal to 1 if audio activity is detected at t and 0 otherwise. Whenever a frame has audio activity, a binaural (two microphones) sound-source localization (SSL) algorithm provides spatial audio information: a sound-source direction (azimuth and elevation) is mapped onto the image plane,

[1] https://team.inria.fr/perception/avdiarization/

e.g., [4], Fig. 1, and section 3 below. Let K be the number of sound-source directions estimated at frame t when $A_t = 1$. Let $\mathbf{Y}_t = (\boldsymbol{Y}_{t1}, \ldots, \boldsymbol{Y}_{tk}, \ldots, \boldsymbol{Y}_{tK}) \in \mathbb{R}^{2\times K}$ denote the K sound-source directions at t. Hence, VAD combined with SSL estimate a time series of sound locations $\mathbf{Y}_{1:t} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_t\}$ and associated *audio-activity binary masks* $\mathbf{A}_{1:t} = \{A_1, \ldots, A_t\}$.

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., [14] 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 $S_t = n, n \in \{1, \ldots, N\}$ if person n is both observed and speaks at t, and $S_t = 0$ if none of the visible persons speaks at t. Notice that $S_t = 0$ encompasses two cases, namely (i) there is audio activity at t ($A_t = 1$) but sound-source locations cannot be associated with one of the visible persons, and (ii) there is no audio activity at t ($A_t = 0$). The active-speaker tracking can be formulated as a maximum a posteriori (MAP) estimation problem:

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

The posterior probability (1) can be written as:

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

where we used the notation $\boldsymbol{u}_t = (\boldsymbol{x}_t, \boldsymbol{v}_t, \boldsymbol{y}_t, a_t)$. The numerator of (2) expands as:

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

The denominator of (2) expands as:

$$\sum_{j=0}^{N} \Big(P(\boldsymbol{u}_t | S_t = j) \Big(\sum_{i=0}^{N} P(S_t = j | S_{t-1} = i) \times P(S_{t-1} = i | \boldsymbol{u}_{1:t-1}) \Big) \Big).$$

The evaluation of this recursive relationship requires (i) the joint audio-visual likelihood $P(\boldsymbol{u}_t | S_t = s_t)$, (ii) the transition probabilities $P(S_t = j | S_{t-1} = i)$, and (iii) the initial posteriors $P(S_1 = s_1 | \boldsymbol{u}_1), s_1 \in \{0, 1, \ldots, n, \ldots, N\}$. The exact evaluation of (1) is tractable and hence solving the MAP problem (2) is straightforward.

2.1. Audio-Visual Association

In this section we derive an expression for the joint audio-visual likelihood. One crucial feature of the proposed model is its ability to robustly associate the acoustic activity at frame 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, let Z_{tk} be the (audio) observation-to-person assignment variable in our mixture model. The case $A_t = 1$ is first considered, namely there is audio activity at t. The source location observed variables $\boldsymbol{Y}_{tk}$ are assumed to be drawn from the following WD-GMM (weighted-data Gaussian mixture model) [6]:

$$P(\boldsymbol{y}_{tk} | \boldsymbol{x}_t, \boldsymbol{v}_t, A_t = 1; \boldsymbol{\theta}_t, \boldsymbol{\phi}_{tk}) = \sum_{n=1}^{N} \pi_{tn} v_{tn} \mathcal{N}(\boldsymbol{y}_{tk} | \boldsymbol{x}_{tn}, \frac{1}{w_{tk}} \boldsymbol{\Sigma}_{tn}), \quad (3)$$

where the parameters of the posterior gamma distribution are estimated with $\boldsymbol{\theta}_t = (\{\pi_{tn}\}_{n=1}^N, \{\boldsymbol{\Sigma}_{tn}\}_{n=1}^N)$ denotes the GMM free parameters, namely the priors $\pi_{tn} = P(S_t = n)$, $\sum_{n=1}^N v_{tn}\pi_{tn} = 1$ and the 2×2 covariance matrices $\boldsymbol{\Sigma}_{tn}$. In the proposed formulation, the mixture mean vectors, $\{\boldsymbol{x}_{tn}\}_{n=1}^N$ are observed and they correspond to image locations of people heads, while the visibility variables $\{v_{tn}\}_{n=1}^N$ allow to consider only those that are visible at t. For convenience we only address the case $N_t \geq 1$. Notice that this model comprises a weight variable $w_{tk} > 0$ drawn from a gamma distribution $\mathcal{G}(w; \alpha, \beta) = \Gamma^{-1}(\alpha)\beta^\alpha w^{\alpha-1} e^{-\beta w}$ with parameters $\boldsymbol{\phi} = (\alpha, \beta)$. There is a weight *associated with each audio observation* $\boldsymbol{y}_{tk}$ and one may notice that the weight acts as a precision, higher the weight more relevant the observation, and that the observed data are independent but not identically distributed.

The posterior probability of a sound-source direction to be associated with the n-th visible person writes [6]:

$$\eta_{tkn} = P(Z_{tk} = n | \boldsymbol{y}_{tk}, \boldsymbol{x}_t, \boldsymbol{v}_t, A_t = 1; \boldsymbol{\theta}_t, \boldsymbol{\phi}_{tk}) \propto \pi_{tn} v_{tn} \mathcal{P}(\boldsymbol{y}_{tk} | \boldsymbol{x}_{tn}, \boldsymbol{\Sigma}_{tn}, \alpha_{tk}, \beta_{tk}), \quad (4)$$

where $\mathcal{P}$ denotes the Pearson type VII probability distribution function (the reader is referred to [18] for a recent discussion regarding this distribution, also called the Arellano-Valle and Bolfarine generalized t-distribution [11]):

$$\mathcal{P}(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{\Sigma}, \alpha, \beta) = \frac{\Gamma(\alpha + d/2)}{|\boldsymbol{\Sigma}|^{1/2}\, \Gamma(\alpha)\, (2\pi\beta)^{d/2}} \left(1 + \frac{\|\boldsymbol{y} - \boldsymbol{x}\|^2_{\boldsymbol{\Sigma}}}{2\beta}\right)^{-(\alpha + \frac{d}{2})}. \quad (5)$$

The WD-GMM formulation allows one to write the posterior distribution of w_{tk}, which is a gamma distribution because it is the conjugate prior of the precision of the Gaussian distribution:

$$P(w_{tk} | Z_{tk} = n, \boldsymbol{y}_{tk}, \boldsymbol{x}_{tn}; \boldsymbol{\theta}_t, \gamma_{tk}, \delta_{tkn}) \propto \mathcal{G}(w_{tk}; \gamma_{tk}, \delta_{tkn}), \quad (6)$$

where $\gamma_{tk} = \alpha_{tk} + d/2$ and $\delta_{tkn} = \beta_{tk} + 1/2 \left\| \boldsymbol{y}_{tk} - \boldsymbol{x}_{tn} \right\|^2_{\boldsymbol{\Sigma}_{tn}}$. This allows to evaluate the posterior mean of w_{tk}, namely:

$$\overline{w}_{tk} = \sum_{n=1}^{N} v_{tn} \eta_{tkn} \overline{w}_{tkn}, \tag{7}$$

where $\overline{w}_{tkn} = \gamma_{tk}/\delta_{tkn}$ is the conditional mean, which is needed to update the mixture parameters (proportions and covariances in our case) during the maximization step (please consult Section 5 in [6] for more details). By inspection of the above equations it is easily seen that the value of $\overline{w}_{tk}$ is small if the distances between an audio observation $\boldsymbol{y}_{tk}$ and the cluster centers $\boldsymbol{x}_{tn}$ are large. In other words, the weight associated with an observed sound location that is far away from the observed persons is small compared with the weight of an observed sound location that coincides with a person location. Hence, the estimated value of w_{tk}, namely $\overline{w}_{tk}$, reduces the influence of outliers. Notice that the weights w_{tk} play a different role than the responsibilities η_{tkn}. Indeed, the responsibilities are normalized, $\sum_{n=1}^{N} \eta_{tkn} = 1$, hence they can only account for a relative measure of the data relevance. Therefore, we use the estimated weights $\{\overline{w}_{tk}\}_{k=1}^{K}$ and an inlier/outlier threshold w_s to classify the audio observations into an inlier set $\mathcal{Y}_{\text{in}}$ and an outlier set $\mathcal{Y}_{\text{out}}$.

Altogether, this formulation allows one to characterize the audio activity of each observed person. Assuming that the audio observations are independent, one obtains the likelihood of person n to be the active speaker:

$$P(\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{v}_t, A_t = 1 | S_t = n) \propto \begin{cases} \sum\limits_{k \in \mathcal{Y}_{\text{in}}} \eta_{tkn}, 1 \leq n \leq N \\ \sum\limits_{k \in \mathcal{Y}_{\text{out}}} \eta_{tkn}, n = 0 \end{cases} \tag{8}$$

If there is no audio activity at time t, $A_t = 0$, then $S_t = 0$ (there is no active speaker) and the likelihood of an active speaker is a uniform distribution:

$$\begin{aligned} &P(\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{v}_t, A_t = 0 | S_t = n) \propto \\ &P(S_t = n | \boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{v}_t, A_t = 0) = \begin{cases} r & n = 0 \\ \frac{1-r}{N_t} & 1 \leq n \leq N \end{cases} \end{aligned} \tag{9}$$

where $r \in [0, 1]$ describes the probability that there is no audio activity t, i.e., either there is no visible person or none of the visible persons speaks.

2.2. State Transition Model

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

$$\begin{aligned} &p(\boldsymbol{S}_t = j | \boldsymbol{S}_{t-1} = i) \\ &= \begin{cases} p_s & \text{if} \quad i = j \text{ and } v_{t-1i} = v_{ti} = 1 \\ (1 - p_s)/N_t & \text{if} \quad i \neq j \text{ and } v_{t-1i} = v_{tj} = 1 \\ 0 & \text{if} \quad v_{t-1i} = v_{t-1j} = 1 \text{ and } v_{tj} = 0 \\ 1/N_t & \text{if} \quad v_{t-1i} = 1, v_{ti} = 0 \text{ and } v_{tj} = 1 \\ 1/N & \text{if} \quad v_{t-1i} = 0 \text{ and } v_{tj} = 0. \end{cases} \end{aligned} \tag{10}$$

The first case of (10) 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 latter transition probabilities are only defined for completeness as transition between non-visible persons are forbidden by the observation model. 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):

$$\begin{aligned} p(\boldsymbol{S}_t = j | \boldsymbol{S}_{t-1} = i) &= \frac{1 - v_{ti}}{N_t} + v_{t-1i} v_{tj} \\ &\times \left(p_s \delta_{ij} + \frac{(1 - p_s)(1 - \delta_{ij})}{N_t} + \frac{1 - v_{ti}}{N_t} \right) \end{aligned} \tag{11}$$

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

3. Implementation and Experiments

As already outlined, the proposed active-speaker tracker may well be viewed as a diarization process summarized as follows: track multiple persons based on visual information, estimate the auditory activity, and associate this activity to one of the tracked persons. Unlike existing audio-visual diarization approaches, which assume that the participants are always facing the cameras, the proposed model can deal with participants that are temporarily occluded, or who come in and out of the field of view of the camera. Unfortunately there are no publicly available datasets that include participants that take speech turns while they wander around, occlude each other and move in and out of the camera field of view.

Therefore we recorded our own data,[2] gathered with two microphones and one camera e.g., Fig. 1. The audio data

[2] https://team.inria.fr/perception/avtrack1/

are delivered by two microphones plugged into the ears of an acoustic dummy head; the visual data are delivered by a video camera. The two modalities are synchronized such that the video frames are temporally aligned with the audio samples. The videos are recorded at 25 FPS while the audio signals are sampled at 48000 Hz. With this setup, we gathered two scenarios, the *counting* scenario, Fig. 2 and the *chat* scenario, Fig. 3. The *counting* sequence has 500 video frames (20 seconds) while the *chat* sequence has 850 video frames (34 seconds).

We briefly describe the multi-person tracking and sound-source localization techniques used to obtain estimates of our observed auditory and visual variables (Sec. 2.1). Among the visual tracking methods that are currently available, we chose the multi-person tracker of [3]. 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 realizations of the visual observation variables $\mathbf{X}_{1:t}$ and associated *visual-presence binary masks* $\mathbf{V}_{1:t}$, as explained in detail in Sec. 2.

We adopted the sound-source localization method of [4] to estimate sound directions with two degrees of freedom (azimuth and elevation). A prominent advantage of this method, in the context of audio-visual analysis, is that it provides a built-in mechanism for mapping sound directions onto image locations. Hence sound-source directions are eventually expressed in pixel coordinates. In practice, the signals delivered by the two microphones are transformed in the Fourier domain in the following way: the short-time Fourier transform (STFT) is applied to a 0.064 s window of the two signals and this window is shifted along the time axis with 0.008 s hops (or 0.056 s overlap between successive windows). With this Fourier domain sampling, there are 5 feature vectors associated with each video frame. In order to increase the number of audio observations that are associated with a video frame, we consider a longer audio frame while we allow a large overlap between audio frames: this yields 30 feature vectors for each video frame.

A complex-valued feature vector is thus built from each window, whose module and argument describing the ILD (interaural level difference) and IPD (interaural phase difference) respectively. It is well known that these binaural cues contain sound direction information. Each feature vector is then mapped onto the image plane using the piecewise-affine high-dimensional to low-dimensional regression method of [4]. In combination with voice activity detection (VAD), this process provides a time series of realizations of both the sound direction variables $\mathbf{Y}_{1:t}$ and the associated *speech-activity binary masks* $\mathbf{A}_{1:t}$, as detailed in section 2.

In addition to our own data, we also tested our method on the dataset used in [12]. These recordings contain one to three *static* persons *facing* the camera and the microphones, i.e., a Kinect. It is important to note that this dataset often contains persons that speak simultaneously and that speaker diarization is quite challenging in this case. Within this dataset, the *Two10* sequence is a representative example and hence we applied our method to this sequence. The audio recordings in this dataset used a microphone configuration quite different than ours, namely a linear microphone array with 8 microphones. For this reason we applied the SRP-PHAT sound-source localization method to the audio data available with the *Two10* sequence, which only provides the sound's azimuth; this direction is then mapped onto an image column using the microphone-to-camera transformation parameters of the Kinect, hence there is a large vertical sound-direction uncertainty.

We compared the proposed method with [7] and with [12]. The main difference between the current work and [7] is the audio-visual association model. In [7] a GMM with a uniform component (GMM+U) is used while here we propose to use the weighted-data GMM (WD-GMM). Moreover, [7] considers a single audio observation for each video frame and the parameters of the GMM+U mixture are manually defined. The parameters of the proposed WD-GMM observation model are learned on-line by gathering audio observations within a 0.4 s window centered on each video frame. This robustly clusters audio observations generated by the same person. The diarization method proposed in [12] uses a supervised classifier (SVM), trained using sequences from the same dataset (same acoustic environment), to discriminate between speaking and non-speaking persons. This contrasts with our on-line joint audio-visual observation model which is completely unsupervised.

Table 1 quantitatively compares the methods in terms of the speaker diarization performance. The proposed model outperforms the one proposed in [7] for *counting* sequence, while it competes the state-of-the-art method of [12], although the latter benefits from training on data from the same experimental setting.

Figures 2, 3 and 4 display our diarization results on the *counting*, *chat* and *Two10* sequences, respectively. The proposed method obtains very good results over the *counting* sequence (see Figure 2) even if the sequence exhibits large portions where the two speakers speak at the same time. The performance over the challenging case of the *chat* sequence is lower than for the *counting* sequence. This drop can be explained by the fact that one speaker is mostly facing away both the camera and the microphones, thus his localization from audio data is much more challenging because of reverberations. Finally, the results on sequence *Two10* (Fig. 4) should be interpreted on the premise that

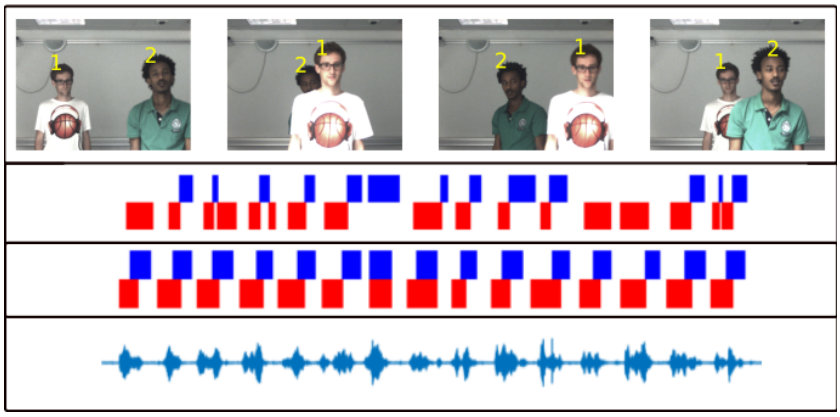

Figure 2: The *counting* sequence involves two moving persons that occasionally occlude each other. Visual tracking results (first row). Diarization results (second row) illustrated with a color diagram: each color corresponds to the audio activity of a person. Ground-truth diarization (third row); notice that there is a systematic overlap between the two speech signals. The raw audio signal delivered by the left microphone (fourth row).

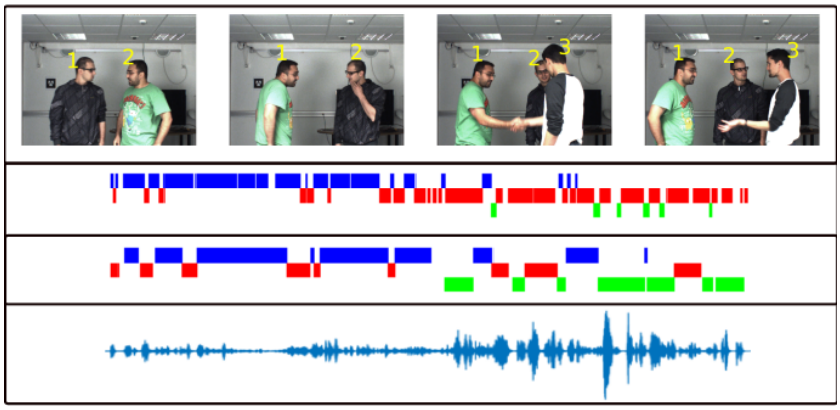

Figure 3: The *chat* sequence involves two then three moving persons that take speech turns and that occasionally occlude each other.

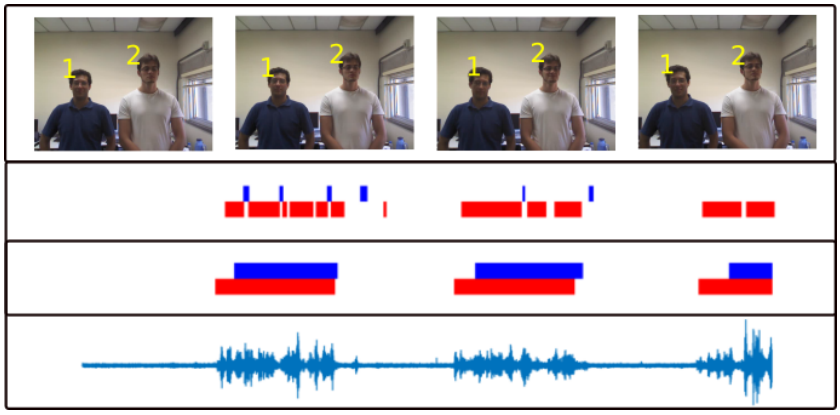

Figure 4: The *Two10* sequence from [12] involves two static persons that speak simultaneously and always face the camera and the microphones.

our method detects only one speaker at a time.

Table 1: Correct detection rates (CDR) obtained by the proposed method and two other methods. The *Chat* and *Two10* sequences contain overlapping speaking persons. The *Chat* sequence contains a varying number of persons that take speech turns.

Sequence	Proposed	[7]	[12]
Counting (Fig. 2)	84%	75%	n/a
Chat (Fig. 3)	55%	64%	n/a
Two10 (Fig. 4)	88%	n/a	92%

4. Conclusions

The paper addressed the problem of active speaker tracking 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., [2, 14]. In contrast, we propose a speech-turn detection and tracking method that enforces spatial coincidence: it exploits that a sound-source and associated visual-object should have the same spatial location. Consequently, it is possible to perform speaker localization by detecting and localizing persons in an image, estimating the directions of arrival of the active sound sources, mapping these sound directions onto the image, and associating the dominant sound source with one of the persons that are visible in the image. Moreover, this process is plugged into a dynamic Bayesian framework that robustly tracks the identity of the speakers and estimates a speech-turn latent variable. We described in detail the proposed method and illustrated its effectiveness with challenging scenarios involving moving people who speak inside a reverberant room and who may visually occlude each other. In the future, we plan to extend our method such that it can robustly deal with simultaneously speaking people. This could be addressed by incorporating rich characterization of the acoustic data and by making use of sound-source separation algorithms.

References

[1] X. Alameda-Pineda and R. Horaud. Vision-guided robot hearing. *The International Journal of Robotics Research*, 34(4-5):437–456, Apr. 2015.

[2] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):356–370, 2012.

[3] S.-H. Bae and K.-J. Yoon. Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In *Computer Vision and Pattern Recognition*, pages 1218–1225, 2014.

[4] A. Deleforge, R. Horaud, Y. Y. Schechner, and L. Girin. 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.

[5] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan. Audiovisual probabilistic tracking of multiple speakers in meetings. *IEEE Transactions on Audio, Speech and Language Processing*, 15(2):601–616, 2007.

[6] I. D. Gebru, X. Alameda-Pineda, F. Forbes, and R. Horaud. EM algorithms for weighted-data clustering with application to audio-visual scene analysis. *arXiv:1509.01509*, Sept. 2015.

[7] I. D. Gebru, S. Ba, G. Evangelidis, and R. Horaud. Audio-visual speech-turn detection and tracking. In *The Twelfth International Conference on Latent Variable Analysis and Signal Separation*, Liberec, Czech Republic, Aug. 2015.

[8] V. Khalidov, F. Forbes, and R. Horaud. Conjugate mixture models for clustering multimodal data. *Neural Computation*, 23(2):517–557, Feb. 2011.

[9] E. Kidron, Y. Y. Schechner, and M. Elad. Cross-modal localization via sparsity. *IEEE Transactions on Signal Processing*, 55(4):1390–1404, 2007.

[10] V. Kilic, M. Barnard, W. Wang, and J. Kittler. Audio assisted robust visual tracking with adaptive particle filtering. *IEEE Transactions on Multimedia*, 17(2):186–200, 2015.

[11] S. Kotz and S. Nadarajah. *Multivariate t Distributions and their Applications*. Cambridge University Press, 2004.

[12] V. P. Minotto, C. R. Jung, and B. Lee. Multimodal on-line speaker diarization using sensor fusion through SVM. *IEEE Transactions on Multimedia*, 2015.

[13] S. Naqvi, M. Yu, and J. Chambers. A multimodal approach to blind source separation of moving sources. *IEEE Journal of Selected Topics in Signal Processing*, 4(5):895 –910, 2010.

[14] A. Noulas, G. Englebienne, and B. J. A. Krose. Multimodal speaker diarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):79–93, 2012.

[15] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior. Recent advances in the automatic recognition of audio-visual speech. *Proceedings of the IEEE*, 91(9):1306–1326, 2003.

[16] M. E. Sargin, Y. Yemez, E. Erzin, and M. A. Tekalp. Audio-visual synchronization and fusion using canonical correlation analysis. *IEEE Transactions on Multimedia*, 9(7):1396–1403, 2007.

[17] J. Sohn, N. S. Kim, and W. Sung. A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, 6(1):1–3, 1999.

[18] J. Sun, A. Kabán, and J. M. Garibaldi. Robust mixture clustering using Pearson type VII distribution. *Pattern Recognition Letters*, 31(16):2447–2454, 2010.

[19] B. D. Van Veen and K. M. Buckley. Beamforming: A versatile approach to spatial filtering. *IEEE ASSP Magazine*, 5(2):4–24, 1988.