Audio-Visual Speaker Localization via Weighted Clustering

Israel-Dejene Gebru, Xavier Alameda-Pineda, Radu Horaud, Florence Forbes

AUDIO-VISUAL SPEAKER LOCALIZATION VIA WEIGHTED CLUSTERING

Israel D. Gebru, Xavier Alameda-Pineda, Radu Horaud and Florence Forbes

INRIA Grenoble Rhone-Alpes, FRANCE

ABSTRACT

In this paper we address the problem of detecting and locating speakers using audiovisual data. We address this problem in the framework of clustering. We propose a novel weighted clustering method based on a finite mixture model which explores the idea of non-uniform weighting of observations. Weighted-data clustering techniques have already been proposed, but not in a generative setting as presented here. We introduce a weighted-data mixture model and we formally devise the associated EM procedure. The clustering algorithm is applied to the problem of detecting and localizing a speaker over time using both visual and auditory observations gathered with a single camera and two microphones. Audiovisual fusion is enforced by introducing a cross-modal weighting scheme. We test the robustness of the method with experiments in two challenging scenarios: disambiguate between an active and a non-active speaker, and associate a speech signal with a person.

Index Terms— Mixture models, audiovisual fusion, multimodal signal processing, weighted-data clustering.

1 Introduction

The problem of detecting and localizing active speakers from audiovisual data arises in many applications, e.g, human-computer interaction and human-robot interaction. A robust solution to this problem is likely to provide rich spatiotemporal information that can be exploited in complex situations, e.g., multi-party dialog between a robot and a group of people. In this paper we emphasize the role of audiovisual fusion in human-to-human, human-to-computer, and human-to-robot interactions and we show that multimodal data processing compensates for the weaknesses of visual-only or audio-only data analysis.

In this context, we focus on the development of a general-purpose active-speaker localization algorithm based on cross-modal clustering. In other words we would like to retrieve the spatiotemporal status of speakers in a group of people engaged in a social interplay. More importantly, we present a clustering methodology in which the two modalities (visual

and auditory) are weighted accordingly to their relevance. The original contribution of this paper is a weighted-data clustering algorithm that robustly localizes people that are both seen and heard.

A number of authors addressed speaker localization based on audio-visual fusion. [1] locates a sound source in an image, based on quantifying the temporal synchrony between the auditory and visual data flows. Subsequently [2] proposes a statistical framework to measure the amount of mutual information between an image region of interest and the auditory signal. [3–5] follow a similar approach to determine the active speaker among a few candidate faces. Both [6] and [7] achieve audio-visual alignment based on correlating the audio signal and the video content. The main advantage of these approaches is the versatility, since they are not constrained to a particular kind of objects and they only require one camera and one microphone. However, they require high-resolution images acquired with speaker-dedicated cameras. Therefore, their use is restricted mostly to static scenarios when the number of speakers is constant and known in advance.

Audio-visual speaker localization and detection has also been cast as a clustering task [8–10]. In [8,9] two Gaussian mixture models (GMM) are used, one GMM per modality. The parameters of the two GMM are constrained via a subset of tying parameters. The resulting EM algorithm has a computationally expensive M-step, involving non-linear optimization, due to the constraints on the parameters. In [10] the visual observation are mapped onto the one dimensional space of auditory observations and hence a single GMM is used to cluster both the audio and the visual data.

In this paper we introduce a novel cross-modal clustering algorithm based on weighting the data. A weight associated with each observation (auditory or visual) indicates the relevance of that observation. Intuitively, higher the weight, stronger the observation. As it will be described in detail below, the weights can be modeled as random variables and incorporated into a maximum-likelihood formulation. This will give rise to a weighted-data Gaussian mixture model (WD-GMM) that is naturally solved via an EM procedure. We show that the latter has an elegant closed-form solution. We explain in detail how the proposed formulation can be applied to cluster auditory and visual data. To the best of our knowledge only a few weighted-data clustering methods were proposed

in the past, either in conjunction with K-means [11] or for segmentation [12]. To the best of our knowledge weighted clustering has not yet been applied to audio-visual data.

The reminder of the paper[1] is structured as follows. Section 2 describes in detail the proposed weighted-data mixture model and the associated EM algorithm. Section 3 describes how we apply the proposed mixture model for speaker localization. Section 4 describes experiments with realistic scenarios. Section 5 concludes the paper.

2 Weighted-Data GMM

In this Section, we present the intuition and the formal definition of the proposed weighted-data model. Let $\boldsymbol{x} \in \mathbb{R}^d$ be a random vector following a multivariate Gaussian distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$, namely $p(\boldsymbol{x}|\theta) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, with the notation $\theta = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$. Let $w > 0$ be a weight indicating the relevance of the observation $\boldsymbol{x}$. Intuitively, higher the weight w, stronger the impact of $\boldsymbol{x}$. The weight can therefore be incorporated into the model by "*observing $\boldsymbol{x}$ w times*". In terms of the likelihood function, this is equivalent to raise $p(\boldsymbol{x}|\theta)$ to the power w. However, $(\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}))^w$ is not a probability distribution because it does not integrate to one. It is straightforward to notice that

$$(\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}))^w \propto \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}, \frac{1}{w}\boldsymbol{\Sigma}\right).$$

Therefore, w plays the role of a precision. Nevertheless, this model is not a standard Gaussian because there is weight w associated with each sample. Subsequently, we write:

$$\hat{p}(\boldsymbol{x}|\theta, w) = \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}, \frac{1}{w}\boldsymbol{\Sigma}\right), \tag{1}$$

from which we write a mixture model with K components:

$$\tilde{p}(\boldsymbol{x}|\Theta, w) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}_k, \frac{1}{w}\boldsymbol{\Sigma}_k\right), \tag{2}$$

where $\Theta = \{\pi_1, \ldots, \pi_K, \theta_1, \ldots, \theta_K\}$ are the model parameters, $\pi_1, \ldots, \pi_K$ are the mixture coefficients, satisfying $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$, and $\theta_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$ are the parameters of the k-th component. We will refer to the model in (2) as the *weighted-data Gaussian mixture model* (WD-GMM).

The difference between the standard GMM and WD-GMM is the *data* weight w. This raises the question on how to define w. We already remarked that in the proposed model the data weights play the role of precisions. Therefore, notable difference between standard Gaussian mixtures and our model is that given n observations, $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$, there is a different weight w_i, hence a different precision, associated with *each* observation $\boldsymbol{x}_i$. Therefore the weights w_i could be considered as hidden variable characterizing the mixture, in addition to

the standard mixture model parameters. Within a Bayesian formalism, the weights will be treated as random variables. Since (1) is a Gaussian, a convenient choice for the conjugate prior $p(w)$ is the gamma distribution. This ensures that the weight posteriors are gamma distributions as well. Hence we have:

$$p(w) = \mathcal{G}(w; \alpha, \beta) \tag{3}$$

$$\mathcal{G}(w; \alpha, \beta) = \frac{1}{\Gamma(\alpha)} \beta^\alpha w^{\alpha-1} \exp(-\beta w) \tag{4}$$

where Γ is the gamma function. The mean and variance of the gamma distribution are given by:

$$E[w] = \frac{\alpha}{\beta}, \quad \text{var}[w] = \frac{\alpha}{\beta^2}. \tag{5}$$

We now formulate the maximum likelihood problem and an associated EM algorithm to estimate the model parameters. Given the observed data, $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, we assume $\boldsymbol{X}$ to be independent and drawn from (2). Let $\boldsymbol{W} = \{w_i\}_{i=1}^n$ be the set of associated weights, i.e., w_i is associated with $\boldsymbol{x}_i$ and follows a gamma distribution with parameters α_i, β_i. We denote with $\phi_i = \{\alpha_i, \beta_i\}$ the parameters of the prior distribution on w_i, and with $\Phi = \cup_{i=1}^n \phi_i$. The observed-data log-likelihood writes:

$$\ln \tilde{p}(\boldsymbol{X}|\Theta, \boldsymbol{W}) = \sum_{i=1}^{n} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}\left(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \frac{1}{w_i}\boldsymbol{\Sigma}_k\right) \right). \tag{6}$$

It is well known that direct maximization of the log-likelihood function is problematic in case of mixtures and that the expected complete-data log-likelihood must be considered instead of (6). Hence, we introduce a set of n hidden (or assignment) variables $\boldsymbol{Z} = \{z_1, \ldots, z_n\}$ associated with the observed variables $\boldsymbol{X}$ and such that $\boldsymbol{z}_i = k$, $k \in \{1, \ldots, K\}$ if and only if $\boldsymbol{x}_i$ is generated by the k's Gaussian component of the mixture.

Maximum likelihood problems with hidden variables are usually solved with EM, which iteratively maximizes the expected complete-data log-likelihood:

$$\mathcal{Q}\left(\Theta, \Theta^{(r)}\right) = \mathrm{E}_{q(\boldsymbol{Z}, \boldsymbol{W})}[\ln P(\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{X}|\Theta, \Phi)], \tag{7}$$

where $q(\boldsymbol{Z}, \boldsymbol{W}) = P(\boldsymbol{Z}, \boldsymbol{W}|\boldsymbol{X}, \Phi, \Theta^{(r)})$ denotes the posterior distribution given the observations and the parameters at the rth iteration, namely $\Theta^{(r)}$. Indeed, the EM algorithm iterates between computing the posterior distribution $q(\boldsymbol{Z}, \boldsymbol{W})$ using the current parameter set $\Theta^{(r)}$ (E-step) and use this posterior to maximize $\mathcal{Q}$ over the model parameters, thus yielding $\Theta^{(r+1)}$ (M-step). We notice that, for the posterior distribution, we can always write:

$$P\left(\boldsymbol{Z}, \boldsymbol{W}|\boldsymbol{X}, \Phi, \Theta^{(r)}\right) = \prod_{i=1}^{n} p\left(\boldsymbol{z}_i, w_i|\boldsymbol{x}_i, \phi_i, \Theta^{(r)}\right). \tag{8}$$

This posterior can be farther factorized as:

$$p\left(\boldsymbol{z}_i, w_i|\boldsymbol{x}_i, \phi_i, \Theta^{(r)}\right) = P\left(w_i|\boldsymbol{z}_i, \boldsymbol{x}_i, \phi_i, \Theta^{(r)}\right) P\left(\boldsymbol{z}_i|\boldsymbol{x}_i, \Theta^{(r)}\right),$$

---

[1]Supplementary materials for this paper are available online at https://team.inria.fr/perception/research/wdgmm/.

where both quantities on the right-hand side have closed-form expressions. The computation of each of these expressions can be seen respectively as a E-W step and as a E-Z step, although it would be more correct to talk about a E-ZW step.

E-Z step. The marginal posterior distribution for z_i is obtained by integrating over the weight variable in the expression of $p\left(z_i, w_i | \boldsymbol{x}_i, \phi_i, \Theta^{(r)}\right)$. As above, we denote the responsibilities with $\eta_{ik}^{(r+1)} = p(z_i = k | \boldsymbol{x}_i, \phi_i, \Theta^{(r)})$. We successively obtain:

$$
\begin{aligned}
\eta_{ik}^{(r+1)} &= \int p\left(z_i = k, w_i | \boldsymbol{x}_i, \phi_i, \Theta^{(r)}\right) dw_i \\
&\propto \int \pi_k^{(r)} \, p\left(\boldsymbol{x}_i | z_i = k, w_i, \phi_i, \Theta^{(r)}\right) \, p(w_i | \phi_i) \, dw_i \\
&\propto \pi_k^{(r)} \, \mathcal{P}(\boldsymbol{x}_i | \boldsymbol{\mu}_k^{(r)}, \phi_i, \boldsymbol{\Sigma}_k^{(r)}),
\end{aligned}
$$

where $\mathcal{P}(\boldsymbol{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \phi_i)$ denotes the probability distribution function of the Pearson type VII distribution which can be seen as a generalization of the Student's t-distribution:

$$
\begin{aligned}
\mathcal{P}(\boldsymbol{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \phi_i) &= \frac{\Gamma(\alpha_i + d/2)}{|\boldsymbol{\Sigma}_k|^{1/2} \, \Gamma(\alpha_i) \, (2\pi \beta_i)^{d/2}} \\
&\times \left(1 + \frac{\left\| \boldsymbol{x}_i - \boldsymbol{\mu}_k^{(r)} \right\|_{\boldsymbol{\Sigma}_k^{(r)}}^2}{2\beta_i} \right)^{-(\alpha_i + d/2)}
\end{aligned}
$$

E-W step. The posterior distribution for w_i, namely $p(w_i | z_i = k, \boldsymbol{x}_i, \phi_i, \Theta^{(r)})$ is a Gamma distribution since the Gamma distribution is the conjugate prior for the precision of a Gaussian distribution. Therefore, we just need to compute the parameters defining the Gamma distribution:

$$
\begin{aligned}
p(w_i | z_i = k, \boldsymbol{x}_i, \phi_i, \Theta^{(r)}) &\propto p(\boldsymbol{x}_i | z_i = k, w_i, \phi_i, \Theta^{(r)}) p(w_i) \\
&\propto \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_k^{(r)}, \boldsymbol{\Sigma}_k^{(r)} / w_i) \, \mathcal{G}(w_i; \alpha_i, \beta_i) \\
&= \mathcal{G}(w_i; \alpha_i^{(r+1)}, \beta_{ik}^{(r+1)}) \quad (9)
\end{aligned}
$$

with $\alpha_i^{(r+1)} = \alpha_i + \frac{d}{2}$, and $\beta_{ik}^{(r+1)} = \beta_i + \frac{1}{2} \left\| \boldsymbol{x}_i - \boldsymbol{\mu}_k^{(r)} \right\|_{\boldsymbol{\Sigma}_k^{(r)}}^2$. We denote by $\overline{w}_{ik}$ the conditional expectation of w_i:

$$
\overline{w}_{ik} = E[w_i | Z_i = k, \boldsymbol{x}_i]. \quad (10)
$$

Using (5) we obtain the following update rule for the weights:

$$
\overline{w}_{ik}^{(r+1)} = \frac{\alpha_i^{(r+1)}}{\beta_{ik}^{(r+1)}}. \quad (11)
$$

Although estimates of the weights w_i are needed neither by the E-step nor by the M-step of the algorithm, one may want to update the weights through the marginal posteriors:

$$
\begin{aligned}
p(w_i | \boldsymbol{x}_i, \phi_i, \Theta^{(r)}) &= \sum_{k=1}^{K} p(w_i | z_i = k, \boldsymbol{x}_i, \Theta^{(r)}, \phi_i) p(z_i = k | \boldsymbol{x}_i) \\
&= \sum_{k=1}^{K} \mathcal{G}(w_i; \alpha_i^{(r+1)}, \beta_{ik}^{(r+1)}) \, \eta_{ik}^{(r+1)},
\end{aligned}
$$

and from (11) we obtain $\overline{w}_i^{(r+1)} = E[w_i] = \sum_{k=i}^{K} \eta_{ik}^{(r+1)} \overline{w}_{ik}^{(r+1)}$.

M-step. This step maximizes the expected complete-data log-likelihood over the mixture parameters. By expanding (7) and by omitting terms that do not depend on the parameters π_k, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, we have:

$$
\begin{aligned}
\mathcal{Q}_{\text{R}}\left(\Theta, \Theta^{(r)}\right) &= \sum_{i=1}^{n} \sum_{k=1}^{K} \int_{w_i} \eta_{ik}^{(r+1)} \ln \mathcal{N}\left(\boldsymbol{x}_i; \boldsymbol{\mu}_k; \frac{1}{w_i} \boldsymbol{\Sigma}_k\right) \\
&\times p(w_i | \boldsymbol{x}_i, z_i = k, \Theta^{(r)}) \mathrm{d}w_i \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} \eta_{ik}^{(r+1)} \left(\ln \pi_k - \ln |\boldsymbol{\Sigma}_k|^{1/2} \right. \\
&\left. - \frac{\overline{w}_{ik}^{(r+1)}}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^{\top} \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right).
\end{aligned}
$$

The parameters updates come from canceling out the derivatives of the expected complete-data log-likelihood (7). All updates are closed-form expressions:

$$
\pi_k^{(r+1)} = \frac{1}{n} \sum_{i=1}^{n} \eta_{ik}^{(r+1)}, \quad (12)
$$

$$
\boldsymbol{\mu}_k^{(r+1)} = \frac{\sum_{i=1}^{n} \overline{w}_{ik}^{(r+1)} \eta_{ik}^{(r+1)} \boldsymbol{x}_i}{\sum_{i=1}^{n} \overline{w}_{ik}^{(r+1)} \eta_{ik}^{(r+1)}}, \quad (13)
$$

$$
\boldsymbol{\Sigma}_k^{(r+1)} = \frac{\sum_{i=1}^{n} \eta_{ik}^{(r+1)} \overline{w}_{ik}^{(r+1)} \left(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(r+1)} \right) \left(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(r+1)} \right)^{\top}}{\sum_{i=1}^{n} \eta_{ik}^{(r+1)}}.
$$

$$ \quad (14) $$

3 Active Speaker Localization

We now show how the WD-GMM algorithm can robustly and accurately solve the speaker localization problem. We briefly describe an audio source localization method that is able to map natural sounds, such as speech, onto an image. Once the audio sources are represented in the 2D image plane, it is possible to combine them with a face detection and localization algorithm. The auditory and visual observations thus obtained are grouped into clusters in order to associated a sound source with the face of a speaker.

3.1 2D Audio Source Localization

Extracting meaningful features from the auditory signals acquired with one or several microphones is a difficult task for several reasons. First, the auditory data are contaminated by background and microphone noise and by reverberations, which highly perturb the signal. Second, the information we need for our task, i.e., the position of sound source, is embedded in the different auditory channels in a complex and environment-dependent fashion. Third, the information is

sparsely distributed in the spectrograms of the auditory signals, and it is only meaningful when the sound sources are active.

In this work we use the 2D sound source localization method proposed in [13]. This method uses a binaural acoustic head attached to a camera. It is a supervised sound localization method that starts by learning a regression model between spectral binaural featuers and the associated sound position in the image plane (please consult [13] for more details). Once the regression model is trained, it is possible to estimate the position of an observed sound. Let $\boldsymbol{A} = \{\boldsymbol{a}_j\}_{j=1}^{n_a} \in \mathbb{R}^2$ denote the set auditory observations.

3.2   Face Detection and Localization

In a human-computer or human-robot interactive scenario, active speakers are likely to face the recording device. Hence, one can rely on face detection. However, this has shown to be a limitation. e.g., [2, 5, 14], in which non-frontal detection was not possible. Instead, we first detect human upper body using [15]. This detector provides an approximate location of the head. In order to refine this localization, we use the facial-landmark detector of [16]. One of the prominent features of this method is that it provides position of the lips. Therefore, once a face is detected and located in the image, the approximate position of the lips can be easily estimated. In this way a general-purpose visual-based face localizer, that is robust to light changes and to head orientation, can be built. Let $\boldsymbol{V} = \{\boldsymbol{v}_l\}_{l=1}^{n_v} \in \mathbb{R}^2$ denote the set of visual observations, namely lip positions in the image plane.

3.3   Cross-Modal Weighting

In this section we explain how the WD-GMM algorithm is applied to the problem at hand. Let $\boldsymbol{X}$ denote the joint set of audio and visual data: $\boldsymbol{X} = \boldsymbol{A} \cup \boldsymbol{V}$ with $\boldsymbol{x}_i = \boldsymbol{a}_i$ for $i = 1, \ldots, n_a$ and $\boldsymbol{x}_i = \boldsymbol{v}_{i-n_a}$ for $i = n_a + 1, \ldots, n = n_a + n_v$. In other words, the first n_a are auditory observations and the remaining n_v are the visual observations. As with any EM algorithm we must provide initial values for the model parameters, in particular we must provide a set of initial weights $\boldsymbol{W}^{(0)} = \{w_i^{(0)}\}_{i=1}^{i=n} \in \mathbb{R}$. Intuitively, we would like auditory observations that are close to visual observations to have higher relevance that those auditory observation lying far away from all visual observations. The same intuition hold for visual observations that are close/far from auditory observations. The rationale behind this choice is that one auditory observation far away from all visual observations is probably an outlier. However, when an auditory observation is close to many visual observations, there is a bigger chance that it corresponds to an underlying audio-visual cluster (a speaker). Therefore, the latter kind of observations should have larger weights than the former kind of observations. In order to implement this, we initialize the each weight in the following way:

$$w_i^{(0)} = \sum_{s \in \mathcal{S}_i} \exp\left(-\frac{D^2(\boldsymbol{x}_i, \boldsymbol{x}_s)}{\sigma}\right),$$

where D is a distance function and σ is a positive scalar. In the previous formula, $\mathcal{S}_i = \{1, \ldots, n_a\}$ if $i > n_a$ and $\mathcal{S} = \{n_a + 1, \ldots, n_v\}$ if $i \leq n_a$. That is to say that we use the visual observations to compute the weight for the $\boldsymbol{a}_j$'s and the auditory observations to compute the weight for the $\boldsymbol{v}_l$'s. The parameters of the prior gamma distribution are set to $\alpha_i = w_i^{(0)2}$ and $\beta_i = w_i^{(0)}$. In this way, the mean and variance of the prior distribution for w_i are $w_i^{(0)}$ and 1 respectively.

3.4   Determining the Number of Speakers

One of the limitations of EM algorithms is that, by itself, it is unable to choose the best model fitting the set of observations $\boldsymbol{X}$. In other words, the number of K of GMM components must be provided. In our particular application K corresponds to the number of speakers, but we do not know K beforehand. In order to overcome this issue, we use the Bayesian information criterion (BIC). BIC is a quantity that can be estimated from the maximum likelihood parameters. Most importantly, BIC penalizes the models based on their dimensionality: Higher the number of free model parameters, larger the penalization. This is meant to avoid over-fitting and in the particular case of GMM, it has desirable statistical properties [17].

3.5   Post Processing the Clusters

Together with the cross-modal weighting and with BIC, the proposed EM sets up a robust method to coherently group auditory and visual observations. However, so far the formulation is application-independent. The model best fitting the auditory and visual observations does not necessarily correspond to the best representation in terms of detecting active speakers. In our particular case, this translates into getting spurious groups of observations that do not correspond to a speaker that is both seen and heard. More precisely, three types of clusters may be present: (i) *auditory clusters* with weak visual content, (ii) *visual clusters* with weak auditory content and (iii) *audiovisual clusters* containing strong audio and visual content. In the first case, the cluster should be discarded, since the probability of a systematic fail of the upper-body detector is very low. Moreover, an auditory cluster may represent a non-speech audio source or a reverberant sound. In the second case, we could keep the cluster and mark it as a potentially silent speaker.

We are mostly interested in third type of clusters that contain both auditory and visual observations. With this aim, we classify all the observations into clusters based on maximum a posteriori (MAP). Clusters containing both video observations and a sufficient number of audio observations are marked as active speakers. By "sufficient" we mean no less than $\frac{n_a + n_v}{\hat{K}}$, where $\hat{K}$ is the number of clusters chosen with BIC. We found this value high enough to discard clusters containing auditory outliers and small enough to guarantee the good sensibility of the system.
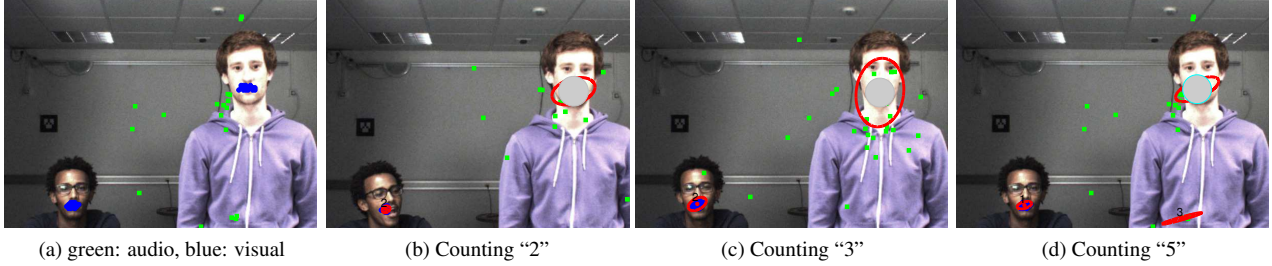
(a) green: audio, blue: visual	(b) Counting "2"	(c) Counting "3"	(d) Counting "5"

Fig. 1: Results obtained with *Scene-1*: the right person speaks while the left person makes lip movements: (a) audio (green) and visual (blue) observations, (b)-(d) the audio-visual cluster found with our method.

4 Experimental Results

The active speaker detection and localization method is applied to audio-visual sequences acquired using a setup composed of a binaural acoustic head and a color camera. Two audio-visual scenarios are used in our experiments. In the first scenario, *Scene-1*, we consider a scene involving an active speaker and a passive one. While the active speaker counts from "1" to "10", the passive one makes fake lip movements. In the second scenario, *Scene-2*, four persons are engaged in an informal conversation, i.e., occasionally two persons speak simultaneously instead of taking speech turns. Notice that in such a scenario, people do not always face the camera and may be partially occluded by another person. Aside from the speech, there are other acoustic events such as reverberations and non-speech sounds present in the room. In order to quantify the performance of the proposed framework, we have manually labeled the the speakers lips through all the sequences.

We evaluated the performance of the method by computing the speaker localization error and the percentage of correct detections. The localization error is computed as the distance between the center of an audio-visual cluster and the ground truth. We consider that an active speaker is correctly detected if there is an overlap between the audio-visual cluster and a circle of radius r centered at the "true" location of the speaker's lips. It is meaningless to perform localization if there is no auditory activity in the scene. Hence, localization is performed only if there is auditory activity. In this context, we introduce an analysis temporal window. A window of 0.8 seconds (or 20 video frames) is used to acquire audio-visual observations in *Scene-1*, and 0.4 seconds for *Scene-2*. We found that these values are long enough to collect enough audio-visual observations. Clearly, the value of the window length have to be chosen considering a trade-off between the scene dynamics and the system response time. In total we analysed ten test windows in *Scene-1* and 100 test windows in *Scene-2*. The values of r that we considered were 40 and 80 pixels. These values were chosen such as to correspond to a face bounding box way that resembles the face bounding box (80×80 pixels) as well as half of it (40×40 pixels). We compared the performance of the proposed framework with audio-only clustering,

Table 1: This table compares the localization error (in pixel). A "-" means that no active speaker is found based on the post-processing procedure, see Section 3.5.

Counting	1	2	3	4	5	6	7	8	9	10	Avg.
Audio-only	148.02	47.81	67.84	51.27	50.28	69.02	49.72	46.21	69.19	61.99	71.78
WD-GMM	3.91	4.04	5.75	0.46	4.70	5.03	5.14	1.81	2.38	3.93	4.20
GMM	12.24	6.72	3.86	17.48	20.79	16.97	16.30	16.97	22.49	22.49	15.13

(a) Results for *Scene-1*. Localization errors are computed on each time interval corresponding to audio-visual activity.

Time Interval	#64	#72	#156	#164	#180	#212	#220	#256	#288	#328	Total Avg.
Audio-only	64.55	61.65	76.67	131.70	44.86	56.92	21.48	88.44	115.82	49.68	73.17
WD-GMM	23.86	42.54	12.15	–	10.45	16.56	20.00	34.98	99.78	47.25	37.17
GMM	–	–	19.21	–	–	35.82	–	152.31	127.92	72.82	64.61

(b) Results for *Scene-2*. The first ten detections are given here. The time interval is indexed by the number of its first frame, namely #64. . .#72, etc.

Table 2: This table displays the number of times the active speaker is correctly detected, over all the sequences.

	Scene-1 (10 time intervals)		Scene-2 (100 time intervals)	
	$r = 80$	$r = 40$	$r = 80$	$r = 40$
WD-GMM	10	10	82	72
GMM	10	10	42	20

and with standard GMM-based clustering.

Table 1 summarizes the results in terms of localization error. It can be seen that audio-only clustering yields the lowest performance. The method with standard GMM gives improved performance since it uses additional visual information. However, the proposed framework outperforms the other two methods. The use of cross-modal weights increases the robustness of the proposed method in fitting the given auditory and visual data in the presence of outliers. These outliers are visual observations from non-speaking faces and noisy sound-source positions.

Table 2 summarizes the results in terms of the number of correct detections. Both GMM and WD-GMM work perfectly on Scene-1, i.e.,thet are able to correctly associate the sound track with the true speaker, all the time. This is as expected because there is only one active speaker in *Scene-1* and both methods are able to find audio-visual clusters. In a more realistic scenario, *Scene-2*, occasionally two people speak simultaneously, and when this happens the auditory observations contain many outliers. As a result, GMM only finds
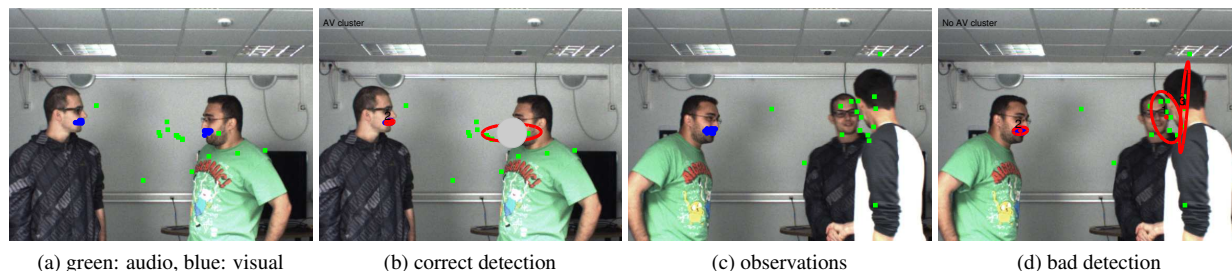
| (a) green: audio, blue: visual | (b) correct detection | (c) observations | (d) bad detection |

Fig. 2: Results obtained *Scene-2* showing both a success and a failure of the method. (a) and (c): audio (green) and visual (blue) observations, (b): the active speaker is correctly detected, (d): the active speaker is not detected because the speaker's face is not detected and hence there is no audio-visual cluster in this case.

noisy clusters that do not corresponds to any of the speakers present in the scene. On the contrary, WD-GMM performs much better because of the weights; Moreover, it is able to find audio-visual clusters that belong to the dominant speaker when there are several active speakers. Fig. 1 shows some results from *Scene-1* while Fig. 2 shows both a success case and a failure in an extremely challenging situation.

5   Conclusion

In this paper, we presented a weighted-data mixture model and derived an EM algorithm that is theoretically well justified. The data weighting scheme provides a flexible tool to characterize the quality of the data. However, initial weight values must be informative with respect to the application at hand. We demonstrated the validity and usefulness of the model on the task of active speaker detection and localization and validated the fact that audio source localization and face detection are complementary and hence their union can substantially improve both the robustness and the accuracy of cross-modal clustering. Because of the usage of weights, the proposed clustering method is less affected by the presence of non-speaking faces, reverberations, background sounds, and so forth. The proposed method has the ability to locate an active speaker, and disambiguate between speaking and non-speaking people in a realistic scenario.

6   References

[1] J. Hershey and J. Movellan, "Audio-vision: Using audio-visual synchrony to locate sounds," in *NIPS*, 2000.

[2] J. W Fisher and T. Darrell, "Speaker association with signal-level audiovisual fusion," *IEEE TMM*, vol. 6, no. 3, pp. 406–413, 2004.

[3] T. Butz and J Thiran, "Feature space mutual information in speech-video sequences," in *ICME*, 2002.

[4] M. J. Beal, H. Attias, and N. Jojic, "Audio-video sensor fusion with probabilistic graphical models," in *ECCV*. 2002.

[5] P. Besson, V. Popovici, J-M. Vesin, J Thiran, and M. Kunt, "Extraction of audio features specific to speech production for multimodal speaker detection," *IEEE TMM*, vol. 10, no. 1, pp. 63–73, 2008.

[6] Z. Barzelay and Y. Schechner, "Onsets coincidence for cross-modal analysis," *IEEE TMM*, vol. 12, no. 2, pp. 108–120, 2010.

[7] A. Llagostera Casanovas, G. Monaci, P. Vandergheynst, and R. Gribonval, "Blind audiovisual source separation based on sparse redundant representations," *IEEE TMM*, vol. 12, no. 5, pp. 358 –371, 2010.

[8] V. Khalidov, F. Forbes, and R. Horaud, "Conjugate mixture models for clustering multimodal data," *Neural Computation*, vol. 23, no. 2, pp. 517–557, 2011.

[9] V. Khalidov, F. Forbes, M. Hansard, E. Arnaud, and R. Horaud, "Detection and localization of 3d audio-visual objects using unsupervised clustering," in *ICMI*, 2008.

[10] X. Alameda-Pineda, V. Khalidov, R. Horaud, and F. Forbes, "Finding audio-visual events in informal social gatherings," in *ICMI*, 2011.

[11] M. Ackerman, S. Ben-David, S. Branzei, and D. Loker, "Weighted clustering.," in *AAAI*, 2012.

[12] F. Forbes, S. Doyle, D. Garcia-Lorenzo, C. Barillot, and M. Dojat, "A weighted multi-sequence markov model for brain lesion segmentation," in *AISTATS*, 2010.

[13] A. Deleforge, V. Drouard, L. Girin, and R. Horaud, "Mapping sounds on images using binaural spectrograms," in *EUSIPCO*, 2014.

[14] A. Noulas, G. Englebienne, and B. J. A. Krose, "Multimodal speaker diarization," *IEEE TPAMI*, vol. 34, no. 1, pp. 79–93, 2012.

[15] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, "Progressive search space reduction for human pose estimation," in *CVPR*, 2008.

[16] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *CVPR*, 2012.

[17] C. Keribin, "Consistent estimation of the order of mixture models," *Sankhya Series A*, vol. 62, no. 1, pp. 49–66, 2000.