ACTIVE SPEAKER DETECTION AND LOCALIZATION WITH A WEIGHTED-DATA MIXTURE MODEL

Israel D. Gebru [1,2] *Xavier Alameda-Pineda* [1] *Radu Horaud* [1] *Florence Forbes* [1]

[1]INRIA Grenoble Rhone-Alpes
655 Avenue de l'Europe
38334 Montbonnot, France

[2]Université Grenoble Alpes
621 Avenue Centrale
38041 Saint-Martin-d'Heres, France

ABSTRACT

In this paper we address the problem of detecting and locating speakers using audiovisual data. We propose to address this problem in the framework of data clustering. We propose a novel cross-modal clustering method based on finite mixture models and 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 devise the associated EM procedure which we justify from a theoretical point of view. 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 simulated data and we show experiments in two challenging scenarios: disambiguate between an active and a non-active speaker, and associate a voice with a person.

Index Terms— Mixture models, audiovisual fusion, multimodal signal processing, cross-modal clustering.

1. INTRODUCTION

The problem of detecting and localizing an active speaker arises in many applications, e.g, human-computer interaction, human-robot interaction, diarization, etc. A robust solution to this problem is likely to provide rich spatiotemporal information that can be exploited in complex situations, e.g., multiparty 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 in the development of a general-purpose robust instantaneous active speaker localization algorithm based the fusion of audio and video data. In other words we would like to retrieve the active speakers in a group of people engaged in a natural social interplay, by means of auditory and visual information. More importantly, we present a methodology in which these pieces of information are weighted accordingly to their relevance. Our most revealing contribution to the field is that the proper use of these weights can notably increase the performance of speaker localization task.

Among the different methods that perform speaker localization, only a few are performing the fusion of both audio and video modalities. [**?**] presents a method to locate sound source in the image, based on quantifying the synchrony between the auditory and the visual modalities. This work inspired a series of information-theory based papers. [**?**] proposes a statistical framework to measure the amount of mutual information between a region of interest on the image and the audio track. [**?**,**?**,**?**] follow a similar approach to determine the active speaker among a few candidate faces. The main advantage of these approaches is the versatility, since they are not constrained to a particular kind of objects. 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.

Recently, a series of papers [**?**, **?**, **?**] dealing with the instantaneous localization of speakers has been published. The common point of these studies is that they cast the speaker localization problem into a multimodal clustering task. In that sense we inspired from them. More precisely, [**?**, **?**] use two Gaussian Mixture Models (GMM) one 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 subroutines, due to the parameters' constraints. In [**?**] a single GMM is used to cluster multimodal data. The main contribution is that the mixture parameters are estimated relaying more on visual than on auditory data.

None of the methods above addresses the problem of audio-visual fusion with weighted data. Indeed, most of them are able to trust one of the two modalities, but none is able to give a different weight to the observations coming from one modality. Even if weighted-data clustering has been

addressed in the recent past [**?**, **?**], up to the authors' knowledge this is the very first study on how to use weights on multimodal data clustering.

The two main contributions of this paper are the following: 1) we propose a new clustering model based on finite mixture model named Weighted-Data Gaussian Mixture Model (WD-GMM),that explores the idea of non-uniform weighting of samples, 2) we develop a robust and instantaneously method for active speaker localizations by fusion information from audio and visual data using the propose mixture model.

The reminder of the paper is structured as follows. In Section 2 we present the proposed mixture model formulation, while the EM algorithm is presented in Section 3. Section 4 describes how we apply the proposed mixture model for speaker localization task. In Section 5 we illustrated experimental result on synthetic data and real audio-visual recording. Section 6 ends the paper by presenting some remark and suggestion for further work.

2. WEIGHTED-DATA GMM

In this section we formally define the proposed mixture model. Let $\boldsymbol{x}$ be a random vector following a normal distribution parameterized by $\theta = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ (mean and covariance), i.e., $p(\boldsymbol{x}|\theta) = \mathcal{N}(\boldsymbol{x}; \theta)$. Let $w > 0$ be an indicator of the relevance of the observed sample $\boldsymbol{x}$. We refer to w as the *weight* of $\boldsymbol{x}$. Intuitively, higher the weight, more important the observation. In maximum likelihood formulations one can enforce this importance by *observing* $\boldsymbol{x}$ w *times*. Regarding the likelihood, 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 show that $(\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}))^w \propto \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}, \frac{1}{w}\boldsymbol{\Sigma}\right)$. Subsequently, we write:

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

This equation can be used to write a K-component GMM:

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

where $\Theta = \{\pi_1, \ldots, \pi_K, \theta_1, \ldots, \theta_K\}$ are the mixture parameters, with $\sum_{k=1}^{K} \pi_k = 1$, and $\theta_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$. We will refer to (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 first remark that w plays the role of a precision. From a statistical point of view it is natural to choose w as a random variable with a gamma distribution prior. Indeed, the gamma distribution is the conjugate prior of the precision matrix. The use of a conjugate prior ensures that the posterior distribution will also be a gamma distribution. This is convenient since in maximum likelihood with hidden variables the posterior distribution is often needed. Hence we write $w \sim \mathcal{G}(\alpha, \gamma)$ to express that w follows the gamma distribution with parameters $\alpha > 0$ (shape) and $\gamma > 0$ (inverse scale or rate). The probability density function is given by

$$\mathcal{G}(w; \alpha, \gamma) = \frac{\gamma^\alpha}{\Gamma(\alpha)} w^{\alpha-1} \exp(-w\gamma), \tag{3}$$

where $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}\mathrm{e}^{-t}\mathrm{d}t$ is the gamma function.

3. MAXIMUM LIKELIHOOD ESTIMATION

We now formulate the maximum likelihood problem and an associated EM algorithm to estimate the model parameters. Let $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$ be the set of i.i.d. observations drawn from a K-component WD-GMM (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, γ_i. We denote with $\phi_i = \{\alpha_i, \gamma_i\}$ the parameters of the prior distribution on w_i, and with $\Phi = \cup_{i=1}^n \phi_i$. In addition to $\boldsymbol{W}$, we consider the set $\boldsymbol{Z} = \{\boldsymbol{z}_i\}_{i=1}^n$ of observation-to-component assignment latent variables. As in standard GMM, $\boldsymbol{z}_i = [z_{i1}, \ldots, z_{iK}] \in \{0, 1\}^K$ with $\sum_{k=1}^K z_{ik} = 1$, and $z_{ik} = 1$ if and only if $\boldsymbol{x}_i$ was generated by the k^{th} component.

Summarizing, we are given a set of observations $\boldsymbol{X}$, and a set of corresponding weight priors Φ. Both the weights $\boldsymbol{W}$ and the observation-to-component assignments $\boldsymbol{Z}$ are hidden variables. Finally, the model is parametrized by Θ, which encompasses the mixture proportions $\{\pi_k\}_{k=1}^K$ and the parameters $\{\theta_k\}_{k=1}^K = \{\boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^K$.

Maximum likelihood problems with hidden variables are usually solved by the Expectation-Maximization (EM) algorithm [**?**], which iteratively maximizes the expected complete-data log-likelihood:

$$\mathcal{Q}\left(\Theta, \Theta^{(r)}\right) = \mathrm{E}_{q(\boldsymbol{Z},\boldsymbol{W})}\{\ln P(\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{X}|\Theta, \Phi)\}, \tag{4}$$

where $q(\cdot) = P\left(\cdot|\boldsymbol{X}, \Phi, \Theta^{(r)}\right)$ denotes the posterior distribution given the observations and the parameters at the r^{th} 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). One iteration of EM is detailed below.

E-step: We notice that the posterior distribution is separable on i, thus we write $q(\boldsymbol{Z}, \boldsymbol{W}) = \prod_{i=1}^n q(\boldsymbol{z}_i, w_i)$. We can develop this further and write:

$$q(\boldsymbol{z}_i, w_i) = P\left(w_i|\boldsymbol{z}_i, \boldsymbol{x}_i, \phi_i, \Theta^{(r)}\right) P\left(\boldsymbol{z}_i|\boldsymbol{x}_i, \phi_i, \Theta^{(r)}\right), \tag{5}$$

where both quantities on the right-hand side have closed-form expressions.

E-W step: Because we use the conjugate prior for w_i, we know beforehand that the posterior is a gamma distribution too. Subsequently we write:

$$P\Big(w_i|z_{ik}=1,\boldsymbol{x}_i,\phi_i,\Theta^{(r)}\Big)=\mathcal{G}(w_i|\alpha_i^{(r+1)},\gamma_{ik}^{(r+1)}),$$

where $\alpha_i^{(r+1)}$ and $\gamma_{ik}^{(r+1)}$ denote the parameters of the posterior distribution and are given by the following expressions:

$$\alpha_i^{(r+1)}=\alpha_i+\frac{d}{2} \tag{6}$$

$$\gamma_{ik}^{(r+1)}=\gamma_i+\frac{1}{2}\left\|\boldsymbol{x}_i-\boldsymbol{\mu}_k^{(r)}\right\|_{\boldsymbol{\Sigma}_k^{(r)}}, \tag{7}$$

collectively denoted by: $\phi_i^{(r+1)}=\left\{\alpha_i^{(r+1)},\left\{\gamma_{ik}^{(r+1)}\right\}_{k=1}^{K}\right\}$.

E-Z step: The posterior distribution of $\boldsymbol{z}_i$, denoted by $\eta_{ik}^{(r+1)}=P\big(z_{ik}=1|\boldsymbol{x}_i,\phi_i,\Theta^{(r)}\big)$, is computed by marginalization:

$$\begin{aligned}\eta_{ik}^{(r+1)}&=\int P\Big(z_{ik}=1,w_i|\boldsymbol{x}_i,\phi_i,\Theta^{(r)}\Big)\,\mathrm{d}w_i\\&\overset{k}{\propto}\int P\Big(\boldsymbol{x}_i|z_{ik}=1,w_i,\phi_i,\Theta^{(r)}\Big)\,\pi_k^{(r)}P(w_i|\phi_i)\,\mathrm{d}w_i\\&\overset{k}{\propto}\pi_k^{(r)}\mathcal{P}_{\text{VII}}\left(\boldsymbol{x}_i;\boldsymbol{\mu}_k^{(r)},\boldsymbol{\Sigma}_k^{(r)},\phi_i\right),\end{aligned}$$

where $\mathcal{P}_{\text{VII}}$ denotes the Pearson type VII distribution:

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

M-step: The parameter updates are obtained by maximizing the expected complete-data log-likelihood (4). This maximization yields closed-form expressions for all the model parameters. While the expression for the mixture proportions correspond to the standard GMM, i.e., $\pi_k^{(r+1)}=\frac{1}{n}\sum_{i=1}^{n}\eta_{ik}^{(r+1)}$, the updates for the means and covariances are provided in (8), with the notation $\overline{w}_{ik}^{(r+1)}=\frac{\alpha_i^{(r+1)}}{\gamma_{ik}^{(r+1)}}$.

4. SPEAKER LOCALIZATION AND DIARIZATION

In this Section we apply the WD-GMM to the problem of active speaker localization using audio-visual fusion. More precisely, we develop an instantaneous active speaker detector, localizer using auditory and visual data. Moreover, cross-modal weights are used to systematically provide a relevance measure for each observation point in a data driven fashion. That is to say the auditory observation (Section 4.1) are used to weight the visual observations (Section 4.2) and vise-versa, as explained in Section 4.3. Determining the number of active speakers is statistical solved in Section 4.4 and the prominent speaker diarization in Section 4.5.

4.1. The Audio Modality

Extracting meaningful features from the auditory signals acquired at the microphones is a difficult task for several reasons. First, the two auditory channels encompass valuable information about the position and content of the sound sources, which is combined in a complex and environment-dependent fashion. Second, both signals are contaminated by noise, coming from the microphones, and reverberations, which can highly perturb the signal. Third, the information is sparsely distributed in the auditory signal, both in time and frequency. On one side, auditory features will only be meaningful when the sound sources are active. On the other side, common sounds such as speech, are characterized by a sparse spectrum.

In order to provide to the EM algorithm reliable auditory features, we chose the sound source localization method proposed in [**?**] for its performance and robustness. The method uses spectral binaural cues to learn the effect of the environment on the acoustic signals and therefore it is able to accurately find the sound source position. More precisely, the method requires a training phase with white noise in which the position of the sound source is known during the extraction of the spectral binaural cues (Interaural Phase and Level Differences). These position-cue pairs are used to learn a probabilistic mapping from the source space to the spectral domain. Moreover, the probabilistic framework provides the inversion mapping, thus a sound source localization mapping from spectral binaural cues. Therefore, the algorithm is able to decouple the content of the sound source from its position. In addition, the probabilistic model is specifically designed to cope with the microphone noise. One may think that the use of white noise in the training phase limits the applicability of the algorithm. However, one prominent feature of the probabilistic model is the explicit modeling of *missing data* situations. That is to say, that the localization mapping does not need a spectral cue with meaningful information in all frequency bands. Instead, the mapping makes use of those frequency bands in which the source is emitting. For all these reasons, we find that the method proposed in [**?**] is extremely well adapted to the scenario of our current research.

In practice, we train the method with a loudspeaker emitting white noise and carrying an easy-to-detect target. This target provides the image location associated to the extracted binaural cues. Once the localization mapping is learnt, we use it to extract potential sound source locations that will be denoted by $\boldsymbol{A}=\{\boldsymbol{a}_j\}_{j=1}^{n_a}$.

$$\boldsymbol{\mu}_k^{(r+1)} = \frac{\sum_{i=1}^{n} \overline{w}_{ik}^{(r+1)} \eta_{ik}^{(r+1)} \boldsymbol{x}_i}{\sum_{i=1}^{n} \eta_{ik} \overline{w}_{ik}^{(r+1)}}, \qquad \boldsymbol{\Sigma}_k^{(r+1)} = \frac{\sum_{i=1}^{n} \eta_{ik}^{(r+1)} \overline{w}_{ik}^{(r+1)} \left(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(r+1)}\right) \left(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(r+1)}\right)^\top}{\sum_{i=1}^{n} \eta_{ik}^{(r+1)}}. \tag{8}$$

4.2. The Visual Modality

Together with the signals acquired at the microphones, we also use the image flows captured by a color camera. Visual information is less sparse than auditory information. As long as the speakers are in the field of view and they are not occluded, the light reflected on their surface will be captured by the camera sensors.

As in the previous section, we would like to provide to the EM algorithm features that robustly localize the people in the field of view. One may immediately think about detection the speakers' face. However, this has shown to be a limitation in many previous works [**?**, **?**, **?**], in which non-frontal detection was not possible. Instead, we first detect human upper body using [**?**]. This detector provides an approximate location of the head. In order to refine this localization, we run the landmark face detector presented [**?**]. One of the prominent features of this method is that it provides position of the lip landmarks. Therefore, if a face is found, the head position is replaced by the average position of the lip landmarks. In this way we build a general-purpose visual person localizer that is robust to light changes and to pose thanks to [**?**, **?**] and that always provides a localization, refined in the case of frontal detection. From now on, these localizations will be denoted by $\boldsymbol{V} = \{\boldsymbol{v}_l\}_{l=1}^{n_v}$, here after referred as visual observations.

4.3. Cross-modal Weighting

As discussed in Section 3, we need to provide the prior parameters Φ of the weights $\boldsymbol{W}$ associated to the audio-visual observations $\boldsymbol{X} = \boldsymbol{A} \cup \boldsymbol{V}$. From now on we write $\boldsymbol{x}_i = \boldsymbol{a}_i$ 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.

In this context the following natural question arises: how can we systematically provide values for Φ in a data-driven fashion and that will help the EM algorithm group the observations? 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 observations should have larger weight than the former kind of observations. In order to make this intuition real we compute the following quantity for each observation $\boldsymbol{x}_i$:

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

where D is a distance function. 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 $\alpha_i = \gamma_i w_i^{(0)} + 1$ and $\gamma_i = \gamma$. In this way, the mode of the prior distribution for w_i is $w_i^{(0)}$. γ is a parameter to be set experimentally, and we observed that values in the range 0.5 to 10 did not have a noticeable effect on the results. We have chosen $\gamma = 0.5$ since it sets the prior variance of w_i to a big value. In this way, the EM algorithm is less constrained by the initialization.

4.4. Determining the Number of Speakers

One of the limitations of the EM algorithm is that, by itself, it is unable to choose the best model fitting the set of observations $\boldsymbol{X}$. In other words, the proposed EM runs for a given number of components K. However, in our particular application, we do not know the number of speakers beforehand. In order to overcome this issue, we use the Bayesian Information Criterion (BIC). BIC is a quantity that should be computed after with the maximum likelihood parameters. Most importantly BIC penalizes the models based on their dimensionality. This higher the number of free parameters of the model, the larger the penalization. This is meant to avoid over-fitting, in the particular case of GMM-like models, it has desirable statistical properties, see [**?**]. BIC has the following expression:

$$\text{BIC}(\boldsymbol{X}, \Theta_K) = \ln \tilde{p}(\boldsymbol{X}|\Theta_K, \Phi) - \frac{6K \ln(n)}{2}. \tag{9}$$

The model maximizing BIC will be chosen.

4.5. Post-Processing

The EM algorithm is the right methodology to solve for the ML problem formulated above. Together with the cross-modal weighting and the BIC, they set up a robust method to coherently group auditory and visual observations. However, the present probabilistic framework is application-blind. That is to say that the model best fitting the auditory and visual observations does not necessarily correspond to the best representation of the ongoing social interplay. In our particular

case, this translates into getting spurious groups of observations that do not correspond to a speaker in the scene. More precisely, we may have groups of auditory observations that do not contain any visual observations and groups of only visual observations. In the first case, the cluster should be discarded, since the probability of a systematic fail of the upper-body detector is very low. In the second case, we could keep the cluster and mark it as a potentially silent speaker.

We are mostly interested in clusters that contain both auditory and visual observations. With this aim, we classify all the observations into clusters using 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 cluster chosen by BIC. We found this value high enough to discard clusters containing auditory outliers and small enough to guarantee the good sensibility of the system.

5. EXPERIMENTAL RESULTS

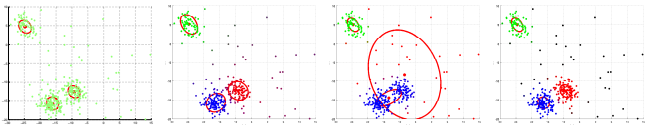
In this section, our proposed model is first tested on clustering a synthetic dataset, then applied to a real audio-visual data sequence acquired using a robotic setup.

5.1. Synthetic Data

We verify on a toy example that our proposed model and the EM algorithm behaves as expected. The experiment have been carried out as follows. First we generate a toy dataset from a GMM having 3 component, a number of samples from uniform distribution (UD) are added as noise. We compare the robustness to fit among different models: standard GMM (Std GMM), Std GMM plus a uniform component (Std GMM +U) and WD-GMM. One way in which the presence of atypical observations in the data has been handled when fitting GMM has been to include an additional component having a UD, i.e (Std GMM +U).

An importance aspect of the proposed model is the weight for observations. For the toy dataset we proceed as follows. We take into consideration the rational assumption that data points that are in a dense area should have a higher weight. Then, we generate the weight w_i as sum of the inverse exponential pairwise euclidean distance from a point to all other points as in (4.3) and initialize the gamma distribution parameter as $\alpha_i = \gamma_i w_i + 1$ and we fix the value for γ_i. It is known that the EM algorithm depend on initial parameters. Thus, the same initial parameters is given too all models. We initialize with a result from K-means.

The clustering result obtained are shown in Figure 1. The result obtained using WD-GMM (Fig. 1b) compares well with the true grouping (Fig.1a). The result obtained from Std GMM (Fig.1c) fails to adequately model the data. Obviously, the one of the component is attempting to model the background noise. In the other hand, the Std GMM +U (Fig. 1d) works well since it is the same model used to generate the data in the first instance. However, this model, unlike the WD-GMM model, cannot be expected to work as well in situations when the noise is not uniform. We argue that weighting data in GMM can substantially improve performance, especially when a weighting scheme is derived in a way that give lower weight value to outliers.



(a) True grouping (b) Fitting a WD-GMM (c) Fitting a Std GMM (d) Fitting a Std GMM +U

Fig. 1. Result of Clustering on a toy dataset using different models.

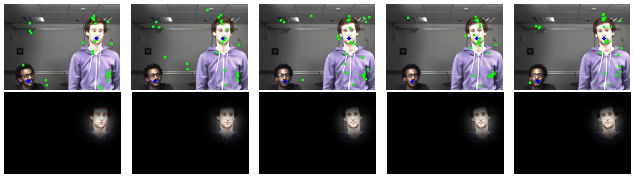


Fig. 2. The speaking-turn scenario. One persons count while the person on the left make false lip movement. 1st row shows observation from audio and video in green and blue color respective. 2nd row, shows the localized active speaker using our method

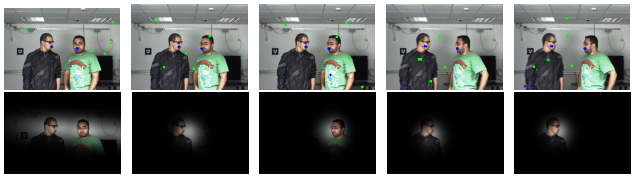


Fig. 3. The speaking-turn scenario. Two persons engage in a normal conversation take speech turns

5.2. Results with Real Audio-Visual Data

Audio-visual data is acquired on a motorized human dummy head - called Popeye. Popeye resides in a normal room and the scene is recorded via two microphone in Popeye ear's. Popeye also has two eyes (stereo camera). However, in our experiments we only used one eye. While using stereo can provide depth information and capture a larger part of the scene, this is left for future work.

Two sets of experiments are conducted. In the first experiments a person counts in front of the camera, the result is shown in Figure 2. In the second experiments, two persons engage in a normal conversation take speech turns, the

results is illustrated in Figure 3. From Figures 2 and 3, in the first row panel, we can see two things. First, some of the audio observations are scattered around the true source location i.e mouth. Second, there are some noisy observations due to voice of the active speaker is too weak or the audio segment corrupted by reverberation and background noise which result incorrect localization. The figures in the second row highlights the location of active speakers by our method.

6. CONCLUSION

In this paper, we presented a mixture model that weight samples differently and derived an EM algorithm that is theoretically well justified. The sample weighting scheme provides flexible tool to include external information for robust model parameters learning. However, the weight associate with each sample should be given to the model and it should be developed for each specific task at hand. We demonstrate the validity and usefulness of the model in simulated data modeling task on a toy dataset. The experiments conducted demonstrate the robustness of proposed method in fitting data in the presence of outliers. We demonstrate the capability of the model on active speaker localization task and validate that complementing audio source localization with visual information can substantially improve performance compared to the audio modality alone. Future research direction could be integrating audio-video tracking system on top of our result.