# Audio-Visual Clustering for Multiple Speaker Localization

Vasil Khalidov[1], Florence Forbes[1], Miles Hansard[1], Elise Arnaud[1,2] & Radu Horaud[1]

[1] INRIA Grenoble Rhône-Alpes, 655 avenue de l'Europe, 38334 Montbonnot, France
[2] Université Joseph Fourier, BP 53, 38041 Grenoble Cedex 9, France

**Abstract.** We address the issue of identifying and localizing individuals in a scene that contains several people engaged in conversation. We use a human-like configuration of sensors (binaural and binocular) to gather both auditory and visual observations. We show that the identification and localization problem can be recast as the task of clustering the audio-visual observations into coherent groups. We propose a probabilistic generative model that captures the relations between audio and visual observations. This model maps the data to a representation of the common 3D scene-space, via a pair of Gaussian mixture models. Inference is performed by a version of the Expectation Maximization algorithm, which provides cooperative estimates of both the activity and the 3D position of each speaker.

**Key words:** multiple speaker localization, audio-visual integration, unsupervised clustering

## 1 Introduction

In most systems that handle multi-modal data, audio and visual inputs are first processed by modality-specific subsystems, whose outputs are subsequently combined. The performance of such procedures in realistic situations is limited. Confusion may arise from factors such as background acoustic and visual noise, acoustic reverberation, visual occlusions. The different attempts that have been made to increase robustness are based on the observation that improved localization and recognition can be achieved by integrating acoustic and visual information. The reason is that each modality may compensate for weaknesses of the other one, especially in noisy conditions. This raises the question of how to efficiently combine the two modalities in different natural conditions and according to the task at hand.

The first question to be addressed is *where* the fusion of the data should take place. There are several possibilities. In contrast to the fusion of previous independent processing of each modality [1], the integration could occur at the feature level. In this case audio and video features are concatenated into a larger feature vector which is then used to perform the task of interest. However, owing to the very different physical natures of audio and visual stimuli, direct integration is not straightforward. There is no obvious way to associate dense visual maps with sparse sound sources. The approach that we propose lies between these two extremes. The input features are first transformed into

a common representation and the processing is then based on the combined features in this representation. Within this strategy, we identify two major directions depending on the type of *synchrony* being used. The first one focuses on *spatial synchrony* and implies combining those signals that were observed at a given time, or through a short period of time, and correspond to the same source (e.g. speaker). Generative probabilistic models in [2] and [3] for single speaker tracking achieve this by introducing dependencies of both auditory and visual observations on locations in the image plane. Although authors in [2] suggested an enhancement of the model that would tackle the multi-speaker case, it has not been implemented yet. Explicit dependency on the source location that is used in generative models can be generalized using particle filters. Such approaches were used for the task of single speaker tracking [4],[5],[6],[7],[8] and multiple speaker tracking [9],[10],[7]. In the latter case the parameter space grows exponentially as the number of speakers increases, so efficient sampling procedures were suggested [10], [7] to keep the problem tractable.

The second direction focuses on *temporal synchrony*. It efficiently generalizes the previous approach by making no a priori assumption on audio-visual object location. Signals from different modalities are grouped if their evolution is correlated through time. The work in [11] shows how principles of information theory can be used to select those features from different modalities that correspond to the same object. Although the setup consists of a single camera and a single microphone and no special signal processing is used, the model is capable of selecting the speaker among several persons that were visible. Another example of this strategy is offered in [12]. Matching is performed there based on audio and video onsets (times at which sound/motion begins). This model is successfully tested even on the case with multiple sound sources. Most of these approaches are however non-parametric and highly dependent on the choice of appropriate features. Moreover they usually require learning or ad hoc tuning of quantities such as window sizes, temporal resolution, etc. They appear relatively sensitive to artifacts and may require careful implementation.

The second question to be addressed is *which* features to select in order to best account for the individual and combined modalities. Some methods rely on complex audio-visual hardware such as an array of microphones that are calibrated mutually and with respect to one or more cameras [6],[10],[8]. A microphone array can provide an estimate 3D location of each audio source. By the use of a microphone pair, certain characteristics, such as interaural time difference (ITD) and interaural level difference (ILD), can be computed as indicators of the 3D position of the source. This localization plays important role in some algorithms, such as partitioned sampling [6] and is mostly considered as the core fusion strategy component. A single microphone is simpler to set up, but it cannot, on its own, provide audio spatial localization. The advantage of using two or more cameras is twofold. First, one may use as many cameras as needed in order to make all parts of a room observable ("smart room" concept). This increases the reliability of visual feature detection because it helps to solve both the occlusion problem and the non fronto-parallel projection problem. Nevertheless, selecting the appropriate camera to be used in conjunction with a moving target can be quite problematic, environment changes require partial or total recalibration. Second, the use of a stereo pair allows the extraction of depth information through the computation of binocular dis-

parities, though so far there has been no attempt to use such a setup. Typically "smart room" models do not consider the problem of speaker localization as a 3D problem, although speakers move and speak in a 3D environment and therefore generate observations that retain in their nature the characteristics of this 3D environment. Projecting on a 2D video frame it is impossible to deal with occlusion of speakers located at the same 2D position but at different depths, so such models are obliged to suppose that there are no occlusions or to consider them as a special case [10].

We propose to use a human-like sensor setup that has both, binaural hearing and stereo vision. We noticed that so far there has been no attempt to use visual depth cues in combination with 3D auditory cues. The advantages of this system include augmented field of view, preservation of 3D spatial information and intrinsic calibration (invariant to environment changes). Another benefit is a potential possibility of a more symmetric integration, in which none of the streams is assumed to be dominant and weighting of the modalities is based on statistical properties of the observed data.

The originality of our proposal is to embed the problem in the physical 3D space, which is not only natural but has more discriminative power in terms of speakers identification and localization. Typically, it is possible to discriminate between visually adjacent speakers, provided that we consider them in 3D space. We try to combine benefits from both types of synchronies. Our approach makes use of spatial synchrony, but unlike the majority of existing models, performing the binding in 3D space fully preserves localization information so that the integration is reinforced. At the same time we do not rely on high-level feature detectors such as structural templates [10], colour models [6] or face detectors, so that the model becomes more general, flexible and stable. Then our approach resembles those based on temporal synchrony in the sense that we recast the problem of how to best combine audio and visual data for speaker identification and localization as the task of finding coherent groups of observations in data. The statistical method for solving this problem is cluster analysis. The 3D positions are chosen as a common representation to which both audio and video features are mapped, through two Gaussian mixture models.

Our contribution is then to propose a unified framework in which we define a probabilistic generative model that links audio and visual data by mapping them to a common 3D representation. Our approach has the following main features: 1) the number of speakers can be determined in accordance with the observed data using statistically well based model selection criteria; 2) a joint probabilistic model, specified through mixture models which share common parameters, captures the relations between audio and video observations; 3) 3D speaker localization within this framework is defined as a maximum likelihood estimation problem in the presence of missing data, and is carried out by adopting a version of the Expectation Maximization (EM) algorithm; 4) we show that such a setting can adapt well to our model formulation and results into cooperative estimation of both speaker 3D positions and speaker activity (speaking or not speaking) using procedures for standard mixture models.

## 2 A Missing Data Model for Clustering Audio-Visual Data

Given a number of audio and visual observations, we address the problem of localizing speakers in a 3D scene as well as determining their speaking state. We will first assume

that the number of speakers is known and fixed to $N$. Section 4 addresses the question of how to estimate this number when it is unknown. We consider then a time interval $[t_1, t_2]$ during which the speakers are assumed to be static. Each speaker can then be described by its 3D location $\boldsymbol{s} = (x, y, z)^T$ in space. We then denote by $\mathbf{S}$ the set of the $N$ speakers' locations, $\mathbf{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n, \ldots, \boldsymbol{s}_N\}$, which are the unknown parameters to be determined.

Our setup consists of a stereo pair of cameras and a pair of microphones from which we gather visual and auditory observations over $[t_1, t_2]$. Let $\mathbf{f} = \{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_m, \ldots, \boldsymbol{f}_M\}$ be the set of $M$ visual observations. Each of them has *binocular coordinates*, namely a 3D vector $\boldsymbol{f}_m = (u_m, v_m, d_m)^\top$, where $u$ and $v$ denote the 2D location in the Cyclopean image. This corresponds to a viewpoint halfway between the left and right cameras, and is easily computed from the original image coordinates. The scalar $d$ denotes the binocular disparity at $(u, v)^T$. Hence, Cyclopean coordinates $(u, v, d)^T$ are associated with each point $\boldsymbol{s} = (x, y, z)^T$ in the visible scene. We define a function $\mathcal{F} : I\!R^3 \to I\!R^3$ that describes this one-to-one relation:

$$\mathcal{F}(\boldsymbol{s}) = (u; v; d) = z^{-1}\,(x; y; B)^T \qquad \mathcal{F}^{-1}(\boldsymbol{f}) = (x; y; z) = Bd^{-1}\,(u; v; 1)^T \quad (1)$$

where $B$ is the length of the inter-camera baseline.

Similarly, let $\boldsymbol{g} = \{g_1, \ldots, g_k, \ldots, g_K\}$ be the set of $K$ auditory observations, each represented by an auditory disparity, namely the *interaural time difference*, or ITD. To relate a location to an ITD value we define a function $\mathcal{G} : I\!R^3 \to I\!R$:

$$\mathcal{G}(\boldsymbol{s}) = c^{-1}\big(\|\boldsymbol{s} - \boldsymbol{s}_{M_1}\| - \|\boldsymbol{s} - \boldsymbol{s}_{M_2}\|\big) \quad (2)$$

Here $c \approx 330\mathrm{ms}^{-1}$ is the speed of sound and $\boldsymbol{s}_{M_1}$ and $\boldsymbol{s}_{M_2}$ are microphone locations in camera coordinates. We notice that isosurfaces defined by (2) are represented by one sheet of a two sheet hyperboloid in 3D. So given an observation we can deduce the surface that should contain the source.

We address the problem of speaker localization within an unsupervised clustering framework. The rationale is that there should exist groups in the observed data that correspond to the different audio-visual objects of the scene. We will consider mixtures of Gaussians in which each component corresponds to a group or class. Each class is associated to a speaker and the problem is recast as the assignment of each observation to one of the class as well as the estimation of each class center. The centers of the classes are linked to the quantities of interest namely the speakers 3D localizations. More specifically, the standard Gaussian mixture model has to be extended in order to account for the presence of observations that are not related to any speakers. We introduce an additional background (outlier) class modelled as a uniform distribution, which increases robustness. The resulting classes are indexed as $1, \ldots, N, N+1$, the final class being reserved for outliers. Also, due to their different nature, the same mixture model cannot be used for both audio and visual data. We used two mixture models, in two different observations spaces (our audio features are 1D while visual features are 3D) with the same number of components corresponding to the number of speakers and an additional outlier class. The class centres of the respective mixtures are linked through common but unknown speaker positions. In this framework, the observed data are naturally augmented with as many unobserved or missing data. Each missing data

point is associated to an observed data point and represents the memberships of this observed data point to one of the $N+1$ groups. The complete data are then considered as specific realizations of random variables. Capital letters are used for random variables whereas small letters are used for their specific realizations. The additional assignment variables, one for each individual observation, take their values in $\{1,\ldots,N,N+1\}$. Let $\boldsymbol{A}=\{A_1,\ldots,A_M\}$ denote the set of assignment variables for visual observations and $\boldsymbol{A'}=\{A'_1,\ldots,A'_K\}$ the set of assignment variables for auditory observations. The notation $\{A_m=n\}$, for $n\in\{1,\ldots,N,N+1\}$, means that the $m^{\text{th}}$ observed visual disparity $\boldsymbol{f}_m$ corresponds to speaker $n$ if $n\neq N+1$ or to the outlier class otherwise. Values of assignment variables for auditory observations have the same meaning.

Perceptual studies have shown that, in human speech perception, audio and video data are treated as class conditional independent [13, 14]. We will further assume that the individual audio and visual observations are also independent given assignment variables. Under this hypothesis, the joint conditional likelihood can be written as:

$$P(\mathbf{f},\boldsymbol{g}\,|\,\mathbf{a},\mathbf{a}')=\prod_{m=1}^{M}P(\boldsymbol{f}_m|a_m)\prod_{k=1}^{K}P(g_k|a'_k). \tag{3}$$

The different probability distributions to model the speakers on one side and the outliers on the other side are the following. The likelihoods of visual/auditory observations, given that they belong to a speaker, are Gaussian distributions whose means respectively $\mathcal{F}(\boldsymbol{s}_n)$ and $\mathcal{G}(\boldsymbol{s}_n)$ depend on the corresponding speaker positions through functions $\mathcal{F}$ and $\mathcal{G}$ defined in (2) and (1). The (co)variances are respectively denoted by $\boldsymbol{\Sigma}_n$ and $\sigma_n^2$,

$$P(\boldsymbol{f}_m|A_m=n)=\mathcal{N}\big(\boldsymbol{f}_m|\mathcal{F}(\boldsymbol{s}_n),\boldsymbol{\Sigma}_n\big), \tag{4}$$

$$P(g_k|A'_k=n)=\mathcal{N}\big(g_k|\mathcal{G}(\boldsymbol{s}_n),\sigma_n^2\big). \tag{5}$$

Similarly, we define the likelihoods for an visual/auditory observation to belong to an outlier cluster as uniform distributions:

$$P(\boldsymbol{f}_m|A_m=N+1)=1/V\quad\text{ and }\quad P(g_k|A'_k=N+1)=1/U, \tag{6}$$

where $V$ and $U$ represent the respective 3D and 1D observed data *volumes* (see Sect.4).

For simplicity, we then assume that the assignment variables are independent. More complex choices would be interesting such as defining some Markov random field distribution to account for more structure between the classes. Following [15] the implementation of such models can then be reduced to adaptive implementations of the independent case making it natural to start with

$$P(\mathbf{a},\mathbf{a}')=\prod_{m=1}^{M}P(a_m)\prod_{k=1}^{K}P(a'_k)\;. \tag{7}$$

The prior probabilities are denoted by, for all $n=1,\ldots,N+1$, $\pi_n=P(A_m=n)$ and $\pi'_n=P(A'_k=n)$. The posterior probabilities, denoted by $\alpha_{mn}=P(A_m=n|\boldsymbol{f}_m)$ and $\alpha'_{kn}=P(A'_k=n|g_k)$, can then be calculated, for all $n=1,\ldots,N+1$, using

Bayes' theorem. For $n \neq N+1$, using (4) and (5) we obtain for each $m = 1, \dots M$

$$\alpha_{mn} = \frac{|\boldsymbol{\Sigma}_n|^{-1/2} \exp\left(-\frac{1}{2} \|\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s}_n)\|^2_{\boldsymbol{\Sigma}_n}\right) \pi_n}{\sum_{i=1}^{N} |\boldsymbol{\Sigma}_i|^{-1/2} \exp\left(-\frac{1}{2} \|\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s}_i)\|^2_{\boldsymbol{\Sigma}_i}\right) \pi_i + (2\pi)^{3/2} V^{-1} \pi_{N+1}} \quad (8)$$

and for each $k = 1, \dots K$

$$\alpha'_{kn} = \frac{|\sigma_n|^{-1} \exp\left(-\left(g_k - \mathcal{G}(\boldsymbol{s}_n)\right)^2 / \left(2\sigma_n^2\right)\right) \pi'_n}{\sum_{i=1}^{N} |\sigma_i|^{-1} \exp\left(-\left(g_k - \mathcal{G}(\boldsymbol{s}_i)\right)^2 / \left(2\sigma_i^2\right)\right) \pi'_i + (2\pi)^{1/2} U^{-1} \pi'_{N+1}}, \quad (9)$$

where we adopted the notation $\|\boldsymbol{x}\|^2_{\boldsymbol{\Sigma}} = \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x}$ for the Mahalanobis distance.

## 3 Estimation Using the Expectation Maximization Algorithm

Given the probabilistic model defined above, we wish to determine the speakers that generated the visual and auditory observations, that is to derive values of assignment vectors $\mathbf{a}$ and $\mathbf{a}'$, together with the speakers' position vectors $\mathbf{S}$. The speakers' positions are part of our model unknown parameters. Let $\boldsymbol{\Theta}$ denote the set of parameters in our model, $\boldsymbol{\Theta} = \{\boldsymbol{s}_1, \dots, \boldsymbol{s}_N, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_N, \sigma_1, \dots, \sigma_N, \pi_1, \dots, \pi_N, \pi'_1, \dots, \pi'_N\}$ . Direct maximum likelihood estimation of mixture models is usually difficult, due to the missing assignments. The Expectation Maximization (EM) algorithm [16] is a general and now standard approach to maximization of the likelihood in missing data problems. The algorithm iteratively maximizes the expected complete-data log-likelihood over values of the unknown parameters, conditional on the observed data and the current values of those parameters. In our clustering context, it provides unknown parameter estimation but also values for missing data by providing membership probabilities to each group. The algorithm consists of two steps. At iteration $q$, for current values $\boldsymbol{\Theta}^{(q)}$ of the parameters, the *E step* consists in computing the conditional expectation with respect to variables $\boldsymbol{A}$ and $\boldsymbol{A}'$,

$$Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(q)}) = \sum_{\mathbf{a},\mathbf{a}' \in \{1,N+1\}^{M+K}} \log P(\mathbf{f}, \boldsymbol{g}, \mathbf{a}, \mathbf{a}'; \boldsymbol{\Theta})\ P(\mathbf{a}, \mathbf{a}' | \mathbf{f}, \boldsymbol{g}, \boldsymbol{\Theta}^{(q)}) \quad (10)$$

The *M step* consists in updating $\boldsymbol{\Theta}^{(q)}$ by maximizing (10) with respect to $\boldsymbol{\Theta}$, i.e. in finding $\boldsymbol{\Theta}^{(q+1)}$ as $\boldsymbol{\Theta}^{(q+1)} = \arg\max_{\boldsymbol{\Theta}} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(q)})$. We now give detailed descriptions of the steps, based on our assumptions.

**E Step.** We first rewrite the conditional expectation (10) taking into account decompositions (3) and (7) that arise from independency assumptions. This leads to $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(q)}) = Q_{\mathcal{F}}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(q)}) + Q_{\mathcal{G}}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(q)})$ with

$$Q_{\mathcal{F}}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(q)}) = \sum_{m=1}^{M} \sum_{n=1}^{N+1} \alpha_{mn}^{(q)} \log(P(\boldsymbol{f}_m | A_m = n; \boldsymbol{\Theta})\ \pi_n)$$

$$\text{and} \quad Q_{\mathcal{G}}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(q)}) = \sum_{k=1}^{K} \sum_{n=1}^{N+1} \alpha_{kn}^{\prime(q)} \log(P(g_k | A'_k = n; \boldsymbol{\Theta})\ \pi'_n),$$

where $\alpha_{mn}^{(q)}$ and $\alpha_{kn}^{\prime(q)}$ are the expressions in (8) and (9) for $\boldsymbol{\Theta} = \boldsymbol{\Theta}^{(q)}$ the current parameter values. Substituting expressions for likelihoods (4) and (5) further leads to

$$Q_{\mathcal{F}}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(q)}) = -\frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N}\alpha_{mn}^{(q)}\left(\|\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s}_n)\|_{\boldsymbol{\Sigma}_n}^2 + \log\left((2\pi)^3|\boldsymbol{\Sigma}_n|\pi_n^{-2}\right)\right) - \frac{1}{2}\sum_{m=1}^{M}\alpha_{m,N+1}^{(q)}\log\left(V^2\pi_{N+1}^{-2}\right) \tag{11}$$

$$\text{and} \quad Q_{\mathcal{G}}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(q)}) = -\frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N}\alpha_{kn}^{\prime(q)}\left(\frac{(g_k - \mathcal{G}(\boldsymbol{s}_n))^2}{\sigma_n^2} + \log(2\pi\sigma_n^2\pi_n^{\prime -2})\right) - \frac{1}{2}\sum_{k=1}^{K}\alpha_{k,N+1}^{\prime(q)}\log(U^2\pi_{N+1}^{\prime -2})\ . \tag{12}$$

**M Step.** The goal is to maximize (10) with respect to the parameters $\boldsymbol{\Theta}$ to find $\boldsymbol{\Theta}^{(q+1)}$. Optimal values for priors $\pi_n$ and $\pi'_n$ are easily derived independently of the other parameters by setting the corresponding derivatives to zero and using the constraints $\sum_{n=1}^{N+1}\pi_n = 1$ and $\sum_{n=1}^{N+1}\pi'_n = 1$. The resulting expressions are

$$n = 1, \ldots, N+1, \quad \pi_n^{(q+1)} = \frac{1}{M}\sum_{m=1}^{M}\alpha_{mn}^{(q)} \quad \text{and} \quad \pi_n^{\prime\,(q+1)} = \frac{1}{K}\sum_{k=1}^{K}\alpha_{kn}^{\prime(q)}\ . \tag{13}$$

The optimization with respect to the other parameters is less straightforward. Using a coordinate system transformation, we substitute variables $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N$ with $\hat{\boldsymbol{f}}_1 = \mathcal{F}(\boldsymbol{s}_1), \ldots, \hat{\boldsymbol{f}}_N = \mathcal{F}(\boldsymbol{s}_N)$. For convenience we introduce the function $h = \mathcal{G} \circ \mathcal{F}^{-1}$ and the parameter-set $\tilde{\boldsymbol{\Theta}} = \left\{\hat{\boldsymbol{f}}_1, \ldots, \hat{\boldsymbol{f}}_N, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_N, \sigma_1, \ldots, \sigma_N\right\}$. Setting the derivatives with respect to the variance parameters to zero, we obtain the usual empirical variances formulas. Taking the derivative with respect to $\hat{\boldsymbol{f}}_n$ gives

$$\frac{\partial Q}{\partial \hat{\boldsymbol{f}}_n} = \sum_{m=1}^{M}\alpha_{mn}\left(\boldsymbol{f}_m - \hat{\boldsymbol{f}}_n\right)^T\boldsymbol{\Sigma}_n^{-1} + \frac{1}{\sigma_n^2}\sum_{k=1}^{K}\alpha'_{kn}\left(g_k - h(\hat{\boldsymbol{f}}_n)\right)\boldsymbol{\nabla}_n^{\top} \tag{14}$$

where the vector $\boldsymbol{\nabla}_n$ is the transposed product of Jacobians $\boldsymbol{\nabla}_n = \left(\frac{\partial \mathcal{G}}{\partial \boldsymbol{s}}\frac{\partial \mathcal{F}^{-1}}{\partial \boldsymbol{f}}\right)^T_{\boldsymbol{f}=\hat{\boldsymbol{f}}_n}$ which can be easily computed from definitions (1) and (2). The resulting derivation includes a division by $d$ and we note here that cases when $d$ is close to zero correspond to points on very distant objects (for fronto-parallel setup of cameras) from which no 3D structure can be recovered. So it is reasonable to set a threshold and disregard the observations that contain small values of $d$.

Difficulties now arise from the fact that it is necessary to perform simultaneous optimization in two different observation spaces, auditory and visual. It involves solving a

system of equations that contain derivatives of $Q_{\mathcal{F}}$ and $Q_{\mathcal{G}}$ whose dependency on $\boldsymbol{s}_n$ is expressed through $\mathcal{F}$ and $\mathcal{G}$ and is non-linear. In fact, this system does not yield a closed form solution and the traditional EM algorithm cannot be performed. However, setting the gradient (14) to zero leads to an equation of special form, namely the *fixed point equation* (FPE), where the location $\hat{\boldsymbol{f}}_n$ is expressed as a function of the variances and itself. Solution of this equation together with the empirical variances give the optimal parameter set. for this reason we tried the versions of the M-step that iterate through FPE to obtain $\hat{\boldsymbol{f}}_n$. But we observed that such solutions tend to make the EM algorithm converge to local maxima of the likelihood.

An alternative way to seek for the optimal parameter values is to use a gradient descent-based iteration, for example, the Newton-Raphson procedure. However, the limiting value $\tilde{\boldsymbol{\Theta}}^{(q+1)}$ is not necessarily a global optimizer. Provided that the value of $Q$ is improved on every iteration, the algorithm can be considered as an instance of the Generalized EM (GEM) algorithm. The updated value $\tilde{\boldsymbol{\Theta}}^{(q+1)}$ can be taken of the form

$$\tilde{\boldsymbol{\Theta}}^{(q+1)} = \tilde{\boldsymbol{\Theta}}^{(q)} + \gamma^{(q)}\boldsymbol{\Gamma}^{(q)}\left[\frac{\partial Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(q)})}{\partial \tilde{\boldsymbol{\Theta}}}\right]^T_{\boldsymbol{\Theta}=\boldsymbol{\Theta}^{(q)}} \tag{15}$$

where $\boldsymbol{\Gamma}^{(q)}$ is a linear operator that depends on $\tilde{\boldsymbol{\Theta}}^{(q)}$ and $\gamma^{(q)}$ is a scalar sequence of gains. For instance, for Newton-Raphson procedure one should use $\gamma^{(q)} \equiv 1$ and $\boldsymbol{\Gamma}^{(q)} = -\left[\frac{\partial^2 Q}{\partial \tilde{\boldsymbol{\Theta}}^2}\right]^{-1}_{\boldsymbol{\Theta}=\boldsymbol{\Theta}^{(q)}}$. The principle here is to choose $\boldsymbol{\Gamma}^{(q)}$ and $\gamma^{(q)}$ so that (15) defines a GEM sequence. In what follows we would concentrate on the latter algorithm as soon as it gives better results and potentially gives more flexibility.

**Clustering.** Besides providing parameter estimation, the EM algorithm can be used to determine assignments of each observation to one of the $N+1$ classes. Observation $\boldsymbol{f}_m$ (resp. $g_k$) is assigned to class $\eta_m$ (resp. $\eta'_k$) if $\eta_m = \underset{n=1,\dots N+1}{\operatorname{argmax}} \alpha_{mn}$ (resp. $\eta'_k = \underset{n=1,\dots N+1}{\operatorname{argmax}} \alpha'_{kn}$). We use this in particular to determine active speakers using the auditory observations assignments $\eta'_k$'s. For every person we can derive the speaking state by the number of associated observations. The case when all $\eta'_k$'s are equal to $N+1$ would mean that there is no active speaker.

## 4 Experimental Results

Within the task of multi-speaker localization there are three sub-tasks to be solved. First, the number of speakers should be determined. Second, the speakers should be localized and finally, those who are speaking should be selected. The proposed probabilistic model has the advantage of providing a means to solve all three sub-tasks at once. There is no need to develop separate models for every particular sub-task, and at the same time we formulate our approach within the Bayesian framework which is rich and flexible enough to suit the requirements.

To determine the number of speakers, we gather sufficient amount of audio observations and apply the Bayesian Information Criterion (BIC) [17]. This is a well-founded approach to the problem of model selection, given the observations. The task of localization in our framework is recast into the parameter estimation problem. This gives an

**Fig. 1.** Equipment setup for data recording

opportunity to efficiently use the EM algorithm to estimate the 3D positions. We note here that our model is defined so as to perform well in the single speaker case as well as in the multiple speakers case without any special reformulation. To obtain the speaking state of a person we use the posterior probabilities of the assignment variables calculated at the E step of the algorithm.

We evaluated the ability of our algorithms to estimate the 3D locations of persons and their speaking activity in a meeting situation. The audio-visual sequence that we used is a part of the scenario set that was acquired by the experimental setup shown in Fig.1. A mannequin with a pair of microphones built-in into its ears and a helmet with a pair of stereoscopic cameras attached to the front, served as the acquisition device. The reason for choosing this configuration was to record data from the perspective of a person, i.e. to try to capture what a person would hear and see while being in a certain natural environment. Each of the recorded scenarios comprised two audio tracks and two sequences of images, together with calibration information. The sequence of interest in our case is a meeting scenario (500 stereo-frames at 25fps), shown on Figure 2. There are 5 persons seating around a table, but only 3 persons are visible. The algorithm was applied to short time intervals that correspond to three video frames. Audio and visual observations were collected within each interval using the following techniques. A standard procedure was used to identify "interest points" in the left and right images [18]. These features were put into binocular correspondence by comparing the local image-structure at each of the candidate points, as described in [19]. The cameras were calibrated [20] in order to define the $(u, v, d)^T$ to $(x, y, z)^T$ mapping (1). Auditory disparities were obtained through the analysis of cross-correlogram of the filtered left and right microphone signals for every frequency band [21]. On an average, there were about 1200 visual and 9 auditory observations within each time interval.

We report here on the results obtained by the versions of the algorithm based on a gradient descent (GD) technique, with $\mathbf{\Gamma}$ being block diagonal. We used $\left[-\partial^2 Q/\partial \hat{\boldsymbol{f}}_n^2\right]^{-1}_{\boldsymbol{\Theta}=\boldsymbol{\Theta}^{(q)}}$ as a block for $\hat{\boldsymbol{f}}_n$, so that the descent direction is the same as in Newton-Raphson method. In the examples that we present we adopted the same video variance matrix $\mathbf{\Sigma}$ for all the clusters, thus there was one common block in $\mathbf{\Gamma}^{(q)}$ that performed linear mapping of the form $\mathbf{\Gamma}^{(q)}_{\mathbf{\Sigma}}(\cdot) = \left(\sum_{n=1}^{N}\sum_{m=1}^{M}\alpha_{mn}^{(q)}\right)^{-1}\mathbf{\Sigma}^{(q)}(\cdot)\,\mathbf{\Sigma}^{(q)}$. This direction

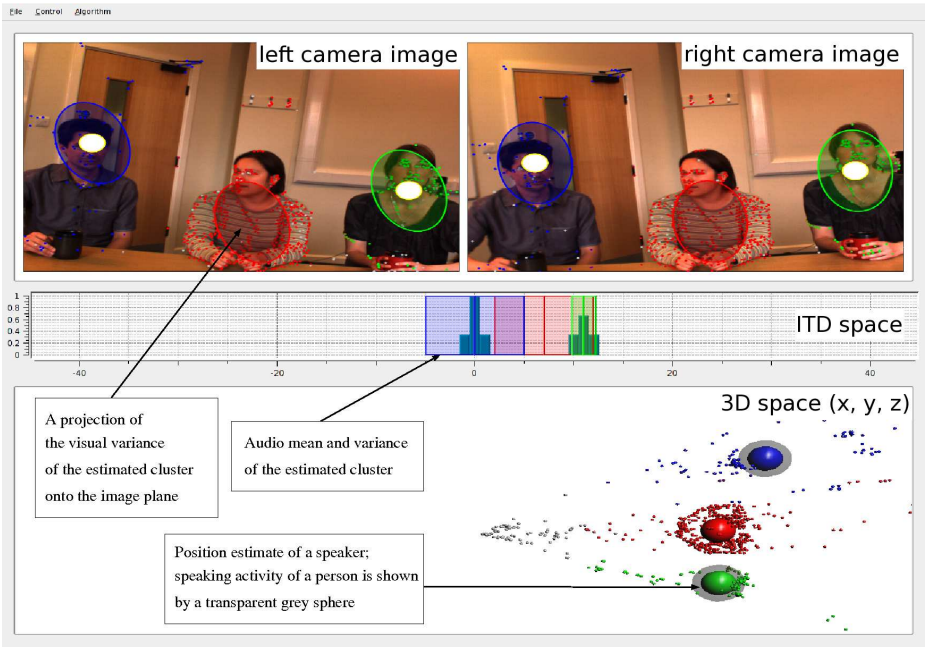


**Fig. 2.** A typical output of the algorithm: stereoscopic image pair, histogram of ITD observation space and 3D clustering (see text for details).

change corresponds to a step towards the empirical variance value. Analogous blocks (cells) were introduced for audio variances, though, unlike the visual variances, individual parameters were used. We performed 1 iteration per each M step, as further iterations did not yield significant improvements. The sequence of gains was chosen to be $\gamma^{(q)} \equiv 1$ (classical GD) and $\gamma^{(q)} = 0.5 + 1/(2(q+1))$ (relaxed GD). Relaxed GD showed moderate behaviour around the optimal point, which causes slower, but more stable convergence with respect to classical GD. This feature of the relaxed GD could prove to be useful in the case of strong noise. By adjusting $\gamma^{(q)}$ one can improve certain properties of the algorithm, such as convergence speed, accuracy of the solution as well as its dynamic properties in the case of parameters changing through time.

Currently we use the Viola-Jones face detector [22] to initialize the EM algorithm from visual disparities that lie within a face. But the results of application of BIC criterion to the observations show that it is capable of determining correctly the number of speakers. Hence we do not strongly rely on initial face detection. As we consider the dynamic evolution of the algorithm, the current estimates would provide good initializations for the next run of the algorithm.

Figure 2 shows a typical output of our algorithm applied to a time interval. The interest points are shown as dots in the left and right images. The 3D visual observations are shown in $x, y, z$ space, below the images. One may notice that after stereo reconstruction there are both inliers and outliers, as well as 3D points that belong to the background. The histogram representation of the ITD observation space is given in the

middle. Transparent ellipses in the images represent projections of the visual covariances corresponding to 3D clusters. The three 3D spheres (blue, red and green) show the locations of cluster centers. Transparent grey spheres surround the current speakers (there are two speakers in this example), also shown with white circles in the image pair. Clusters in the ITD space have a similar representation: the transparent coloured rectangles designate the variances of each cluster, while solid coloured lines drawn at their centres are the corresponding cluster centres. We would like to emphasize the fact that despite the majority of visual observations being located on the central speaker, the influence of the audio data helped to keep the location estimates distinct. At the same time, owing to fine spatial separation of the visual data, the auditory *variances* were adapted rather than the means. This shows the benefits of the combined generative model with respect to separate modality-specific models. The proposed model does not require any explicit modality weighting, as soon as the variances in (14) encode the "reliability" of the observations and the weighting occurs on parameter estimation.

The model was tested on 166 time intervals taken from the meeting scenario with 89 occurences of auditory activity. The soundtrack was labelled manually on the basis of detected onsets and offsets. In total 75 occurences were detected with error probabilities for "missed target" (speaking person detected as non-speaking) and "false alarm" (non-speaking person detected as speaking) being $P_1 = 0.16$ and $P_2 = 0.14$ respectively. Analysis showed that many errors of the first type are due to discretization (frames are processed independently) and proper "dynamic" version of the algorithm could potentially reduce $P_1$ to $0.08$. Currently the auditory observations are collected even when there is no prominent sound, which gives birth to the major part of "false alarm" errors. Such low-energy regions of spectrogram can be detected and suppressed leading to $P_2 = 0.07$. The location estimates for the persons in the middle (green) and on the right (red) lie within their bodies and do not change much. Being accumulated along all chosen time intervals, they form dense clouds of radius 2cm and 4cm respectively. For the person on the left (blue), 97% estimates lie within the body and form the cloud of radius 5cm, though the rest 3% are 10cm away. The reason for this behaviour is, again, the discretization and the problem can be easily resolved by means of tracking. These results show that the model demonstrates reliable 3D localization of the speaking and non-speaking persons present in the scene.

## 5 Conclusion

We presented a unified framework that captures the relationships between audio and visual observations, and makes full use of their combination to accurately estimate the number of speakers, their locations and speaking states. Our approach is based on unsupervised clustering and results in a very flexible model with further modelling capabilities. In particular, it appears to be a very promising way to address dynamic tracking tasks.

**Acknowledgements**. The authors would like to warmly thank Heidi Christensen for providing the ITD detection software, and also Martin Cooke, Jon Barker, Sue Harding, Yan-Chen Lu and other members of the Speech and Hearing Research Group (Department of Computer Science, University of Sheffield) for helpful discussions and comments. We would also like to express gratitude to anonymous reviewers for constructive

remarks. This work has been funded by the European Commission under the Project POP (Perception on Purpose, FP6-IST-2004-027268).

## References

1. M. Heckmann, F. Berthommier, and K. Kroschel. Noise adaptive stream weighting in audio-visual speech recognition. *EURASIP J. Applied Signal Proc.*, 11:1260–1273, 2002.
2. M. Beal, N. Jojic, and H. Attias. A graphical model for audiovisual object tracking. *IEEE Trans. PAMI*, 25(7):828–836, 2003.
3. A. Kushal, M. Rahurkar, L. Fei-Fei, J. Ponce, and T. Huang. Audio-visual speaker localization using graphical models. In *Proc.18th Int. Conf. Pat. Rec.*, pages 291–294, 2006.
4. D. N. Zotkin, R. Duraiswami, and L. S. Davis. Joint audio-visual tracking using particle filters. *EURASIP Journal on Applied Signal Processing*, 11:1154–1164, 2002.
5. J. Vermaak, M. Ganget, A. Blake, and P. Pérez. Sequential monte carlo fusion of sound and vision for speaker tracking. In *Proc. 8th Int. Conf. Comput. Vision*, pages 741–746, 2001.
6. P. Perez, J. Vermaak, and A. Blake. Data fusion for visual tracking with particles. In *Proc. of IEEE (spec. issue on Sequential State Estimation)*, volume 92, pages 495–513, 2004.
7. Y. Chen and Y. Rui. Real-time speaker tracking using particle filter sensor fusion. In *Proc. of IEEE (spec. issue on Sequential State Estimation)*, volume 92, pages 485–494, 2004.
8. K. Nickel, T. Gehrig, R. Stiefelhagen, and J. McDonough. A joint particle filter for audio-visual speaker tracking. In *ICMI '05*, pages 61–68, 2005.
9. N. Checka, K. Wilson, M. Siracusa, and T. Darrell. Multiple person and speaker activity tracking with a particle filter. In *IEEE Conf. Acou. Spee. Sign. Proc.*, pages 881–884, 2004.
10. D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan. Audiovisual probabilistic tracking of multiple speakers in meetings. *IEEE trans. Audi. Spee. Lang. Proc.*, 15(2):601–616, 2007.
11. J. Fisher and T. Darrell. Speaker association with signal-level audiovisual fusion. *IEEE Trans. on Multimedia*, 6(3):406–413, 2004.
12. Z. Barzelay and Y.Y. Schechner. Harmony in motion. In *IEEE Conf. Comput. Vision Pat. Rec. (CVPR)*, pages 1–8, 2007.
13. J.R. Movellan and G. Chadderdon. Channel separability in the audio-visual integration of speech: A bayesian approach. In D.G. Stork and M.E. Hennecke, editors, *Speechreading by Humans and Machines: Models, Systems and Applications*, NATO ASI Series, pages 473–487. Springer, Berlin, 1996.
14. D.W. Massaro and D.G. Stork. Speech recognition and sensory integration. *American Scientist*, 86(3):236–244, 1998.
15. G. Celeux, F. Forbes, and N. Peyrard. EM procedures using mean-field approximations for Markov model-based image segmentation. *Pattern Recognition*, 36:131–144, 2003.
16. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38, 1977.
17. G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, March 1978.
18. C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. 4th Alvey Vision Conference*, pages 147–151, 1988.
19. M. Hansard and R.P. Horaud. Patterns of binocular disparity for a fixating observer. In *Adv. Brain Vision Artif. Intel., 2nd Int. Symp.*, pages 308–317, 2007.
20. Intel OpenCV Computer Vision library. http://www.intel.com/technology/computing/opencv.
21. H. Christensen, N. Ma, S.N. Wrigley, and J. Barker. Integrating pitch and localisation cues at a speech fragment level. In *Proc. of Interspeech 2007*, pages 2769–2772, 2007.
22. P. Viola and M. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004.