IDIAP RESEARCH REPORT

# A Mixed-State I-Particle Filter for Multi-Camera Speaker Tracking

Daniel Gatica-Perez [1] Guillaume Lathoud [1]
Iain McCowan [1] Jean-Marc Odobez [1]

IDIAP–RR 03-25

June 2003

PUBLISHED IN
IEEE ICCV Workshop on Multimedia Technologies
in E-Learning and Collaboration (WOMTEC)
Nice, Oct. 2003

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[1] IDIAP, Martigny, Switzerland

IDIAP Research Report 03-25

# A MIXED-STATE I-PARTICLE FILTER FOR MULTI-CAMERA SPEAKER TRACKING

Daniel Gatica-Perez Guillaume Lathoud Iain McCowan Jean-Marc Odobez

JUNE 2003

PUBLISHED IN
IEEE ICCV Workshop on Multimedia Technologies
in E-Learning and Collaboration (WOMTEC)
Nice, Oct. 2003

**Abstract.** Tracking speakers in multi-party conversations represents an important step towards automatic analysis of meetings. In this paper, we present a probabilistic method for audio-visual (AV) speaker tracking in a multi-sensor meeting room. The algorithm fuses information coming from three uncalibrated cameras and a microphone array via a mixed-state importance particle filter, allowing for the integration of AV streams to exploit the complementary features of each modality. Our method relies on several principles. First, a mixed state space formulation is used to define a generative model for camera switching. Second, AV localization information is used to define an importance sampling function, which guides the search process of a particle filter towards regions of the configuration space likely to contain the true configuration (a speaker). Finally, the measurement process integrates shape, color, and audio observations. We show that the principled combination of imperfect modalities results in an algorithm that automatically initializes and tracks speakers engaged in real conversations, reliably switching across cameras and between participants.

# 1 Introduction

Speaker detection and tracking constitute relevant tasks for applications that include automatic meeting analysis [26, 19, 4] and remote conferencing [27]. In the context of meetings, speaker turn patterns convey a rich amount of information about the dynamics of a group and the individual behaviour of its members, including trends of influence, dominance and level of interest, as documented by a solid body of literature in social psychology [17, 20].

The use of audio and video as separate cues for tracking are classic problems in signal processing and computer vision. However, although audio-based speaker localization offers very valuable information about speaker turns [15], sound and visual information are jointly generated when people speak, and provide complementary advantages for speaker tracking if their dependencies are jointly modeled [25]. Initialization and recovery from failures are bottlenecks in visual tracking that can be robustly addressed with audio. However, precise object localization is better suited to visual processing. There exists substantial evidence about the role that non-verbal behaviour plays in meetings in general, and in turn-taking in particular [20]. Automatically analyzing this behaviour, expressed in the form of gaze, facial expressions, or body postures, requires reliable localization and tracking of human body parts. AV tracking therefore represents a valuable step towards the understanding of rich multimodal behaviours.

Single-camera AV speaker tracking has attracted considerable attention [3, 22, 25, 1]. Among the multiple approaches, generative models that pose tracking as a statistical inference problem, and use either exact [22, 1] or approximate [25, 28] methods for inference, have shown encouraging performance. In contrast, tracking speakers in multi-camera scenarios has been less commonly studied [27, 28, 4]. While single-camera AV tracking algorithms are useful for remote conferencing, meeting rooms usually call for the use of several cameras to cover the different areas where meetings unfold (table, whiteboards, and projector screen). Furthermore, cameras in meeting rooms often have little or no overlapping fields of view (FOVs). In this sense, AV tracking shares some features with other cases of multi-camera surveillance [12, 21, 13].

In particular, Sequential Monte Carlo (SMC) or particle filters (PFs) represent a principled methodology for data fusion [6] that has been recently used for AV tracking in single-camera [25] and multi-camera [28] setups. For a state-space model, a PF recursively approximates the filtering distribution of states given observations using a dynamical model and random sampling by (i) predicting candidate configurations, and (ii) measuring their likelihood, in a process that amounts to random search in a configuration space.

Current SMC formulations for AV speaker tracking usually fuse audio and video only at the measurement level, thus leading to symmetrical models in which each modality accounts for the same relevance, and depending on the dynamical model to generate candidate configurations. Furthermore, cameras and microphones are independently (and carefully) calibrated for state modeling and measuring in 2-D or 3-D. Such formulations tend to overlook several important features of AV data. First, audio is a strong cue to model discontinuities that clearly violate usual assumptions in dynamics (including speaker turns across cameras), and (re)initialization. Its use for prediction would therefore bring benefits to modeling realistic situations. Second, audio can be inaccurate at times, but provides a good initial localization guess that could be enriched by extra visual localization information, and integrated in a principled framework. Third, although audio might be imprecise, and visual calibration can be erroneous due to distortion in wide-angle cameras, their joint occurrence tends to be more consistent, and can be robustly learned from data.

This paper presents a mixed-state particle filter for multi-camera AV speaker tracking, which addresses the points discussed above, and exploits the complementary features of the audio and video modalities. In the first place, a mixed-state space (with discrete and continuous components) allows for the definition of a generative model for camera switching [11], [23]. In the second place, we advocate for the asymmetrical use of modalities in the particle filter formulation. Audio and color information are first used for sampling, and introduced via importance sampling (IS) [8], [10], by defining an IS function that emphasizes the most informative regions of the space. Additionally, audio, color and shape information are jointly used to compute the likelihood of candidate configurations. In the third place, we present a simple yet robust AV calibration procedure that estimates a direct 3-D to 2-D+camera-index mapping from audio localization estimates onto the image planes. The procedure does not requiring precise geometric calibration of cameras and microphones. The result is a principled method that can initialize and track moving speakers, and switch between multiple meeting participants across cameras in a real setting.

The paper is organized as follows. Section 2 presents our algorithm. Section 3 describes the experimental setup. Section 4 presents results. Section 5 provides final remarks.

# 2 Our approach for AV tracking

Given a discriminative object representation and a Markov state-space model, with hidden states $\{\mathbf{x}_t\}$ that represent object configurations, and observations $\{\mathbf{y}_t\}$ extracted from an AV sequence composed of multiple camera and microphone data streams, the distribution $p(\mathbf{x}_t|\mathbf{y}_{1:t})$ can be recursively computed by

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) \propto p(\mathbf{y}_t|\mathbf{x}_t) \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}) d\mathbf{x}_{t-1}, \tag{1}$$

where $\mathbf{y}_{1:t} = \{\mathbf{y}_1, ..., \mathbf{y}_t\}$. The integral in Eq. 1 represents the prediction step, in which the dynamical model $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ and the previous distribution $p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})$ are used to compute a prediction distribution, which is then used as prior for the update step, and multiplied by the likelihood $p(\mathbf{y}_t|\mathbf{x}_t)$ to generate the current filtering distribution. Except for a few special cases, exact inference in this model is intractable. SMC methods are usually employed to approximate Eq. 1 for non-linear, non-Gaussian problems as follows. The filtering distribution is defined by a set of weighted samples or particles $\{(\mathbf{x}_t^{(i)}, \pi_t^{(i)}), i = 1, ..., N\}$, where $\mathbf{x}_t^{(i)}$ and $\pi_t^{(i)}$ denote the i-th sample and its importance weight at the current time. The point-mass approximation is given by $\hat{p}_N(\mathbf{x}_t|\mathbf{y}_{1:t}) = \sum_{i=1}^{N} \pi_t^{(i)} \delta(\mathbf{x}_t - \mathbf{x}_t^{(i)})$. The prediction step propagates each particle according to the dynamics, and the updating step reweights them using their likelihood, $\pi_t^{(i)} \propto \pi_{t-1}^{(i)} p(\mathbf{y}_t|\mathbf{x}_t^{(i)})$. A resampling step using the new weights is necessary to avoid degradation of the particle set [6].

A PF for AV speaker tracking involves the definition of the state-space, the speaker model, the dynamical process, the sampling strategy, the AV calibration procedure, and the probability models for observations. Each of these issues are discussed in the following subsections.

## 2.1 Mixed-state space for multi-camera tracking

State-spaces defined on the image plane or in 3-D are sensible choices. However, 3-D modeling usually requires precise camera calibration and the computation of non-trivial features [28]. In this paper, we define a mixed-state model in which (i) human heads in the image plane are modeled as elements of a template-space, allowing for the description of a template and a set of valid transformations [2], and (ii) cameras depicting people are indexed by a discrete variable. Specifically, a state is defined by

$$X_t = (k_t, x_t), k \in \{0, .., N_K - 1\}, x_t \in \mathbb{R}^{N_x},$$

where $k_t$ is a discrete $N_K$-valued camera index, and $x_t$ is a continuous vector in the space of transformations $\mathbb{R}^{N_x}$. Furthermore, the dynamical model can be factorized as follows,

$$p(X_t|X_{t-1}) = p(k_t|X_{t-1}) p(x_t|k_t, X_{t-1}). \tag{2}$$

The first factor in the right side of Eq. 2 constitutes a generative model for switching cameras: for any given geometric transformation at the previous time, $p(k_t = n|k_{t-1} = m, x_{t-1}) = \mathbf{T}_{mn}(x_{t-1})$ represents a transition probability matrix (TPM) to switch between cameras. The second factor, $p(x_t|x_{t-1}, k_{t-1} = m, k_t = n) = p_{mn}(x_t|x_{t-1})$ denotes elements of a set of $N_K^2$ continuous dynamical models, one for each possible camera transition. Additionally, the observation process depends on the complete configuration (camera index + transformation),

$$p(\mathbf{y}_t|X_t) = p(\mathbf{y}_t|k_t, x_t).$$

The corresponding graphical model is described in Fig. 1. Note that it differs from other switching models proposed in the visual tracking literature [11], [23].

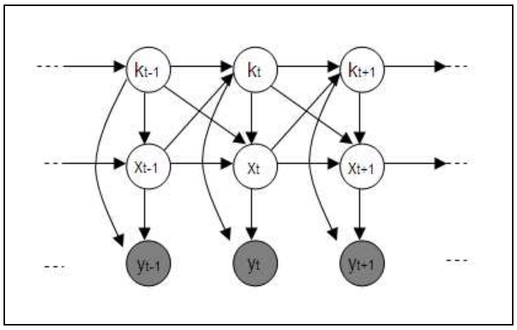


Figure 1: Graphical model for multi-camera tracking. Observed (resp. hidden) variables are denoted by gray (resp. white) nodes.

For the current implementation, we used three cameras ($N_K = 3$), and the space $\mathbb{R}^{N_x}$ has been chosen to be a subspace of the affine transformations comprising translation $T^x, T^y$ and scaling $\theta$, i.e., $x_t = (T_t^x, T_t^y, \theta_t)$.

The estimated tracked configuration is computed as usual in mixed-state models [11]. The MAP estimate of the camera index $\hat{k}_t$, and the weighted mean of the continuous component $\hat{x}_t$ given the MAP discrete estimate are computed by

$$\hat{k}_t = \arg\max_j \sum_{i \in \mathcal{I}_j} \pi_t^{(i)}; \qquad \hat{x}_t = \frac{\sum_{i \in \mathcal{I}_{\hat{k}_t}} \pi_t^{(i)} x_t^{(i)}}{\sum_{i \in \mathcal{I}_{\hat{k}_t}} \pi_t^{(i)}}, \tag{3}$$

where $\mathcal{I}_j = \{i | k_t^{(i)} = j\}$.

## 2.2 Person model

Speaker heads are represented by their silhouettes (contours) in the image plane [2]. In particular, we used a parameterized vertical ellipse to represent the basic shape.

## 2.3 Dynamical models

The uncountable set of TPMs $\{\mathbf{T}_{mn}(x_{t-1})\}$ is coarsely quantized based on the value of $(T_{t-1}^x, T_{t-1}^y)$ to allow for camera switching based on the speaker location at the previous time. The image planes depicted by each of the cameras are divided into a set of likely and unlikely regions for camera switching $\{\mathcal{R}_k^{sw}\}, \{\mathcal{R}_k^{\overline{sw}}\}$. A likely region is for instance the one occupied by speakers when they stand up from their seats to go to the whiteboard or projector screen (note that in practice, a speaker might be viewed by more than one camera based on their overlapping FOVs). Two TPMs are then defined, $\mathbf{T}_{mn}(\mathcal{R}^{sw})$ and $\mathbf{T}_{mn}(\mathcal{R}^{\overline{sw}})$.

Regarding the individual dynamical models, when $m = n$ each of the distributions $p_{nn}(x_t|x_{t-1})$ is defined by a second-order auto-regressive dynamical model on $x_t$. With an augmented continuous state component denoted by $\hat{x}_t = (x_t, x_{t-1})^T$, each switching dynamical model is defined by $\hat{x}_t = A_{nn}\hat{x}_{t-1} + B_{nn}(w_t, 0)^T$, where $A_{nn}, B_{nn}$ are the parameters of each model, and $w_t$ is a white noise process. This set of pdfs allows to handle camera-specific motion models, potentially useful for tracking objects viewed from rather different perspectives. When $m \neq n$, the pdf is substituted by a prior distribution $p_n^0(\hat{x}_t)$ that draws samples in the current image plane in region $\{\mathcal{R}_n^{sw}\}$. A full state in the augmented model is therefore defined by

$$\mathbf{x}_t = (X_t, X_{t-1}).$$

## 2.4 Mixed-state i-particle filters

The basic PF relies only on the dynamical model to generate candidate configurations, which as discussed earlier has limitations due to imperfect motion models and the need for reinitialization, due for instance to speaker turns. Additional knowledge about the true configurations can be extracted from other AV cues, and modeled via importance sampling [8], [10], by using an IS function $I_t(\mathbf{x}_t)$ that emphasizes the most informative regions of the space. The technique first draws samples from $I_t(\cdot)$ rather than from the filtering distribution, concentrating particles in better proposal regions. It then introduces a correction mechanism in order to keep the particle set as a faithful representation of the original distribution, defined by an importance ratio,

$$w_t^{(i)} = \frac{\hat{p}_N(\mathbf{x}_t^{(i)}|\mathbf{y}_{1:t-1})}{I_t(\mathbf{x}_t^{(i)})} = \frac{\sum_{j=1}^{N} \pi_{t-1}^{(j)} p(\mathbf{x}_t^{(i)}|\mathbf{x}_{t-1}^{(j)})}{I_t(\mathbf{x}_t^{(i)})}, \tag{4}$$

and applied to the particle weights, $\pi_t^{(i)} \propto w_t^{(i)} p(\mathbf{y}_t|\mathbf{x}_t^{(i)})$. Furthemore, a reinitialization prior can introduced via a two-component mixture,

$$\tilde{p}(\mathbf{x}_t|\mathbf{y}_{1:t-1}) = \alpha q_t(\mathbf{x}_t) + (1-\alpha)\hat{p}_N(\mathbf{x}_t|\mathbf{y}_{1:t-1}), \tag{5}$$

where $q_t(\mathbf{x}_t)$ denotes a reinitialization prior, and $\{\alpha, 1-\alpha\}$ is the prior on the mixture. Another variation in the model can be further introduced, in which samples are drawn from the original dynamics, the dynamics with IS, and the reinitialization prior with probabilities $\alpha_d, \alpha_i$ and $\alpha_r$, respectively [10].

We extend the previous use of I-PFs to multimodal fusion. Audio tends to be imprecise for localization, due to discontinuities during periods of non-speech, as well as effects of reverberation and other noise. Audio does have some important advantages however, such as the ability to provide instantaneous localization at reasonable computational expense. Additionally, even though audio can be inaccurate, it can still provide reasonable proposals that could be enriched by the use of extra visual localization information, and integrated in the IS function. We propose an asymmetrical use of modalities, where audio and skin color are used for localization via sampling (as part of the IS function and the reinitialization prior), and shape, color and audio are further used as observations in the measurement process. For the reinitialization prior, the IS function is directly used, $I_t \propto q_t$.

Fig. 2 summarizes the particle filter algorithm. The definition of the IS function and observation models is described in detail in following subsections.

---

Generate $\{k_t^{(i)}, x_t^{(i)}, \pi_t^{(i)}\}$ from $\{k_{t-1}^{(i)}, x_{t-1}^{(i)}, \pi_{t-1}^{(i)}\}$.

1. `Compute IS function` $I_t(\cdot)$.
2. `Resampling`. Resample $\{k_{t-1}^{(i)}, x_{t-1}^{(i)}\}$ to generate $\{\tilde{k}_{t-1}^{(i)}, \tilde{x}_{t-1}^{(i)}\}$ based on $\{\pi_{t-1}^{(i)}\}$.
3. `Prediction`. For each $\{\tilde{k}_{t-1}^{(i)}, \tilde{x}_{t-1}^{(i)}\}$:
   (a) generate a uniformly distributed number $\beta \in [0, 1]$.
   (b) if $\beta < \alpha_r$, sample from $q_t(\mathbf{x}_t)$ to produce $(k_t^{(i)}, x_t^{(i)})$, and set $w_t^{(i)} = 1$.
   (c) if $\alpha_r < \beta < \alpha_r + \alpha_i$, sample from $I_t(\mathbf{x}_t)$ to produce $(k_t^{(i)}, x_t^{(i)})$, and set $w_t^{(i)}$ as in Eq. 4.
   (d) if $\alpha_r + \alpha_i < \beta$, sample from $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ to produce $(k_t^{(i)}, x_t^{(i)})$ as follows and set $w_t^{(i)} = 1$,
      i. sample from $\mathbf{T}_{mn}(x_{t-1})$ to generate $k_t^{(i)}$.
      ii. sample from $p_{mn}(x_t|x_{t-1})$ to generate $x_t^{(i)}$.
4. `Measurement`. Re-weight each particle by computing the observation likelihood and weighting by the importance weight, $\pi_t^{(i)} = w_t^{(i)} p(\mathbf{y}_t|k_t^{(i)}, x_t^{(i)})$. Normalize all weights such that $\sum_i \pi_t^{(i)} = 1$.

---

Figure 2: Mixed-state i-particle filter algorithm.

## 2.5 AV calibration

Single-camera AV calibration works have usually assumed simplified configurations [25], [1]. For multi-camera settings, authors have resorted to rigorous camera calibration procedures [28]. However, camera calibration models become more complex for wide-angle lenses (a usual requirement in video conferencing/meeting). Furthermore,

Although audio localization estimates are usually noisy, and visual calibration is affected by geometric distortion, their joint occurrence tends to be more consistent. We have therefore opted for a rough AV calibration procedure, which estimates a mapping from audio configurations in 3-D onto the corresponding camera image plane (or planes if there are overlapping FOVs) using a training sequence, but without requiring precise geometric calibration of audio and video. For this purpose, we collected a sequence with a person speaking while performing activities in the room in typical locations (walking, sitting, moving while seated, standing at the whiteboard and projector screen areas). The audio localization procedure described in Section 2.8 was used to compute 3-D points $Z_t$ for each frame, and a visual (shape-based) PF tracker, hand-initialized in the proper image plane, was used to compute the corresponding 2-D+camera-index points. The set of correspondences obtained for the training set was used to define a mapping between discrete sets $C : \mathbb{R}^3 \rightarrow \{0, .., N_K - 1\} \times \mathbb{R}^2$, such that 3-D positions are mapped into vectors containing camera index and image position, $C(Z_t) = (k_t, T_t^x, T_t^y)$. The mapping for new data is computed via nearest neighbor search.

## 2.6 AV fusion for measurement

We propose to combine shape and localization (audio and color) information in the measurement process. The sole usage of shape information is clearly limited to discriminate between two different human heads. In presence of multiple people or visual clutter, the shape likelihood is multimodal, and particles with large weights would be generated for each person, and likely remain there even after a speaker turn. Furthermore, the mean configuration (Eq. 3) would be a bad representation of the posterior, as it would lie somewhere between the peaks of the distribution without corresponding to any object. Fusing shape and localization information (e.g. audio) in the observation process would solve the above ambiguity, tracking speaker turns with lower latency, and locking only onto the current speaker. We propose to fuse modalities as follows,

$$p(\mathbf{y}_t|\mathbf{x}_t) = p(\mathbf{y}_t^{sh}|\mathbf{x}_t)p(\mathbf{y}_t^{loc}|\mathbf{x}_t), \tag{6}$$

where $p(\mathbf{y}_t^{sh}|\mathbf{x}_t)$ denotes a shape-based observation likelihood, and $p(\mathbf{y}_t^{loc}|\mathbf{x}_t)$ represents a localization likelihood, that uses audio and color information, as described in the following subsections.

## 2.7 Shape observations model

The observation model assumes that shapes are embedded in clutter [2]. Edge-based measurements are computed along $L$ normal lines to a hypothesized contour, resulting in a vector of candidate positions for each line, $\mathbf{y}_t^l = \{\nu_m^l\}$ relative to the point lying on the contour $\nu_0^l$. With some usual assumptions, the shape-based observation likelihood for $L$ normal lines can be expressed as

$$p(\mathbf{y}_t^{sh}|\mathbf{x}_t) \propto \prod_{l=1}^{L} p(\mathbf{y}_t^l|\mathbf{x}_t) \propto \prod_{l=1}^{L} \max\left(K, \exp(-\frac{\|\hat{\nu}_m^l - \nu_0^l\|^2}{2\sigma^2})\right), \tag{7}$$

where $\hat{\nu}_m^l$ is the nearest edge detected on the $l^{th}$ line, and $K$ is a constant introduced when no edges are detected.

## 2.8 Audio observation model

The audio speaker localization approach used in our work consists of two steps: finding candidate source locations $\hat{Z}_t$, and classifying them as speech or non-speech. Details are presented in the following.

### 2.8.1 Source localization

To locate sources, a simple single source localization technique based on Time Delay of Arrival (TDOA) is used. In particular, we use the SRP-PHAT technique described in [5], due to its low computational requirements and suitability for reverberant environments.

We define a vector of theoretical time-delays associated with a 3-D location $Z \in \mathbb{R}^3$ as

$$\boldsymbol{\tau}^Z \triangleq (\tau^{1,Z}, \ldots, \tau^{p,Z}, \ldots \tau^{P,Z}), \tag{8}$$

where $P$ is the number of pairs and $\tau^{p,Z}$ is the delay (in samples) between the microphones in pair $p$:

$$\tau^{p,Z} = \frac{(||Z - M_1^p|| - ||Z - M_2^p||)\, f_s}{c}, \tag{9}$$

where $M_1^p, M_2^p \in \mathbb{R}^3$ are the locations of the microphones in pair $p$, $||.||$ is the Euclidean norm, $f_s$ the sampling frequency, and $c$ the speed of sound. Note that for a given time-delay $\tau_0$ and a given pair $p$ there exists a hyperboloid of locations $Z$ satisfying $\tau^{p,Z} = \tau_0$.

From two signals $s_1^p(t)$ and $s_2^p(t)$ of a given microphone pair $p$, the frequency-domain GCC-PHAT [14] is defined as:

$$\hat{G}^p_{PHAT}(f) \quad \triangleq \quad \frac{S_1^p(f) \cdot [S_2^p(f)]^*}{\left|S_1^p(f) \cdot [S_2^p(f)]^*\right|}, \tag{10}$$

where $S_1^p(f)$ and $S_2^p(f)$ are Fourier transforms of the two signals and $[\cdot]^*$ denotes the complex conjugate. Typically the two Fourier transforms are estimated on Hamming-windowed segments of 20-30 ms. By performing an Inverse Fourier Transform, and summing the time-domain GCC-PHAT $\hat{R}^p_{PHAT}(\tau)$ across pairs, we obtain the SRP-PHAT measure,

$$P_{SRP-PHAT}(Z) \quad \triangleq \quad \sum_{p=1}^{P} \hat{R}^p_{PHAT}(\tau^{p,Z}), \tag{11}$$

From this, the source location is estimated as

$$\hat{Z} = \arg\max_{Z \in \mathbb{R}^3} [P_{SRP-PHAT}(Z)], \tag{12}$$

Based on geometrical considerations, at least 3 microphone pairs ($P \geq 3$) are required to obtain a unique peak.

The maximization is implemented using an exhaustive search over a fixed grid of points, $H \subset \mathbb{R}^3$ such that

$$\forall Z \in \mathbb{R}^3 \quad \exists Z_H \in H \quad \Gamma(Z, Z_H) \leq \gamma_0, \tag{13}$$

where $\Gamma(Z_1, Z_2)$ is the distance in time-delay space (in samples),

$$\Gamma(Z_1, Z_2) \quad \triangleq \quad \sqrt{\frac{1}{P} \sum_{p=1}^{P} \left(\tau^{p,Z_1} - \tau^{p,Z_2}\right)^2}, \tag{14}$$

and $\gamma_0$ is the desired precision in samples. Since we typically upsample the time-domain GCC-PHAT function $\hat{R}^p_{PHAT}(\tau)$ with a factor $\alpha_{up}$ (e.g. 20), the desired precision is set accordingly to $\gamma_0 = 1/\alpha_{up}$.

The grid $H$ is built by picking points heuristically on a few concentric spheres centered on the microphone array. The spheres' radii were also determined by $\gamma_0$. Conceptually this approach relates to [9].

For each time frame, our implementation therefore approximates Eq. 12 with

$$\hat{Z} \approx \arg\max_{Z \in H} [P_{SRP-PHAT}(Z)]. \tag{15}$$

#### 2.8.2 Speech/non-speech classification

Speech/non-speech classification is typically seen as a pre-processing step, often based on an energy threshold criterion. In the current work however, we propose basing the speech/non-speech decision purely on the localization information, performing it after the location estimate has been obtained.

Conventional single-channel speech/non-speech segmentation approaches are based upon energy, SNR estimation (as in [5]) or more complex estimators such as zero-crossing rate [16]. While relatively robust, techniques based on energy thresholding can often miss low-energy beginnings of words, or even entire speaker turns, when these are short.

Here we pose the problem of speech/non-speech classification in the framework of localization. We first run single source localization on each time frame, then *based on the localization results* classify each frame as speech or non-speech, relying on short-term clustering of location estimates, as explained in the following.

**Short-term Clustering Algorithm**. Our motivation for short-term clustering is that noisy location estimates feature high variations over time, while location estimates are consistent during speech periods. The proposed algorithm has three steps:

1. build short-term clusters of frames whose location estimates are close to each other.

2. retain only "significant clusters" by applying a duration constraint.

3. label those frames belonging to any significant cluster as speech, others as non-speech. The result can then be used by the particle filter.

In step 1, two frames $t_1$ and $t_2$ belong to the same cluster if the following condition is true:

$$d(\hat{Z}_{t_1}, \hat{Z}_{t_2}) < d_0 \quad and \quad |t_2 - t_1| \leq T_0, \tag{16}$$

where $d_0$ and $T_0$ are fixed thresholds in space and time respectively. $d(\hat{Z}_{t_1}, \hat{Z}_{t_2})$ is a distance defined according to the setup. For example, with a single, planar microphone array it is reasonable to use the difference in azimuth between $\hat{Z}_{t_1}$ and $\hat{Z}_{t_2}$. In our experiments $d_0$ was set to 5 degrees. For $T_0$, a typical value should be close to the length of a phoneme (e.g. 200 ms).

For step 2, within each cluster, we find the longest segment. If that segment lasts more than a threshold $T_{consec}$ (typical value: 100ms), the cluster is kept as "significant", otherwise it is dropped. Simpler criteria such as minimum cluster duration or the minimum number of frames within the cluster did not prove adequate. Additionally, to eliminate far-field noise sources (e.g. PC, projector), we also discard clusters whose average SRP-PHAT value is below a threshold.

In step 3, frames belonging to any significant cluster are labeled as speech, others as non-speech. In the usual case where the audio frame rate is higher than the video frame rate, we downsample the audio by grouping audio 3-D estimates between consecutive video frames. For example, with audio frame rate 62.5 fps and video frame rate 25 fps, there can be zero (non-speech), one, two or three (speech) audio 3-D estimates $\{\hat{Z}_t\}$ per video frame.

## 2.9 Importance sampling function

As stated before, the IS function is defined by audio and color information. For each frame, each of the audio estimates in 3-D is mapped onto their corresponding image planes and 2-D locations (Section 2.5). In the camera setup in the meeting room (Fig. 3), one camera has overlapping FOV with the other two, while these two cameras do not share FOVs with each other. We used a majority rule to keep all the proposals $C(Z_t) = (k_t, T_t^x, T^y)$ for the specific $k_t^*$ that has the largest number of audio estimates $N_t^a$.

While multiple audio estimates are beneficial for sampling, in some cases all of them are inaccurate (due to errors in the audio localization and the AV mapping), but roughly close to the true configuration. Therefore, such estimates can be used as initial proposals, and enriched by using additional visual localization information in the IS process. Specifically, skin color blobs are computed at each time for the camera $k_t^*$ chosen by audio as described before. A 20-component Gaussian Mixture Model (GMM) of skin color was estimated from a training set of people of different ethnicities participating in real meetings in the room, collected over several days. Skin pixels were classified based on thresholding on the skin likelihood, followed by morphological postprocessing to extract blobs. Then, the centroid positions of all $N_t^v$ skin blobs within a radius $r_k$ from any image-mapped audio estimate are also considered as proposal locations. The IS is then defined as $I_t(k_t, x_t) = \delta(k_t - k_t^*) I_t^{k^*}(x_t)$ where $I_t^{k^*}(x_t)$ is a GMM using all AV proposals as components,

$$I_t^{k^*}(T_t^x, T_t^y, \theta_t) = \sum_{j=1}^{N_t^a + N_t^v} \lambda_t^j \mathcal{N}(\mu_t^j, \Sigma_t^j) \tag{17}$$

where $\lambda_t^j$ denotes the prior on the mixture, the means $\mu_t^j = (\mu_t^{T^x}, \mu_t^{T^y}, \mu_t^{\theta})$ consist of either a projected audio estimate onto the image plane or a skin blob centroid position $(\mu_t^{T^x}, \mu_t^{T^y})$, and a camera-dependent, time-independent scale factor $\mu_t^{\theta}$, and the covariance matrices $\Sigma_t^j$ are diagonal, with translation components proportional to the (camera-dependent) mean head size in the training set, and with scaling component proportional to the variance in scale of head sizes. In case of non-speech, no IS function exists, so the filter draws samples only from the dynamics.

Finally, the importance function is also used for the AV localization-based observation likelihood,

$$p(\mathbf{y}_t^{loc} | \mathbf{x}_t) \propto I_t(\mathbf{x}_t), \tag{18}$$

in case there is audio, and it is a fixed constant otherwise.

# 3 Experimental setup

AV recordings were made in a 8.2m×3.6m×2.4m meeting room containing a 4.8m×1.2m rectangular meeting table, and equipped with fully synchronized video and audio capture devices [18]. The configuration is shown in Fig. 3. The video equipment includes three identical PAL-quality, CCTV cameras (SONY SSC-DC58AP), each with a wide-angle lens with adjustable FOV ($38° - 80°$), connected to a MiniDV tape recorder. Two cameras in opposite walls record frontal views of two participants at the table, including the workspace area. These cameras were set in order to avoid occlusion by participants seated on the opposite side, and have null overlapping FOVs. A third wide-view camera looked over the top of the participants towards the white-board and projector screen. The audio equipment consisted of an eight-element circular equi-spaced microphone array centered on the table, with diameter 20cm, and composed of high quality miniature electret microphones (Sennheiser MKE 2-5-C). Video was captured at 25 fps, while audio was recorded at 16kHz, with features estimated at 62.5 fps. Images are 288×360 pixels. In such setup, human heads are approximately 35×55 pixels in the close-views, and about 20×30 in the wide-angle view.

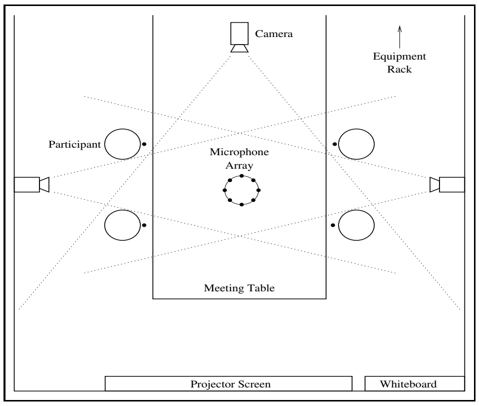


Figure 3: Meeting recording configuration.

# 4 Results and discussion

## 4.1 Audio speaker localization evaluation

This section presents an evaluation of both parts of the audio speaker localization system. Section 4.1.1 describes the test case, then Section 4.1.2 gives performance of the audio source localization system. Section 4.1.3 reports results of the speech/non-speech classification system, comparing with an energy-based approach. Global audio performance is reported in Section 4.1.4.

### 4.1.1 Test case

We recorded a human speaking the same utterance at nine known, fixed locations, including two seated positions and seven standing positions. These nine locations spanned an area of 67 degrees in azimuth, and from 0.9m to 2m in radius.

The recording was annotated in two manners:

- "Located Ground Truth (GT)": the true beginning and end of each of the nine segments was determined by a human listener (42 sec total). Each of these segments was thus annotated with a beginning, an end and a 3-D location.
- "Speech/non-speech GT": the entire recording (92 sec) was segmented in terms of speech and non-speech by a human listener. The speech segments included the nine located segments plus others of unknown location.

### 4.1.2 Audio source localization

Audio source localization was evaluated on the nine located segments. Cumulated error histograms are shown in Fig. 4. We observe that most frames (70.5%) yield an angle error below six degrees, most other frames having a much higher error. The latter may well correspond to short silences between words. In contrast, the radius estimates are not accurate, as expected with a single microphone array.

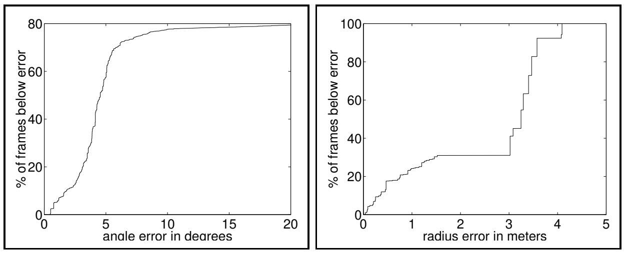


Figure 4: Source localization performance (cumulative histograms).

### 4.1.3 Speech/non-speech classification

After running source localization on all frames, we applied the speech/non-speech classification described in Section 2.8.2. We used azimuth only and chose thresholds $d_0$ = 5 degrees, $T_0$ = 200 ms and $T_{consec}$ = 100 ms. The result was measured with:

- $A$ = frame accuracy on ground truth speech frames.
- $B$ = frame accuracy on ground truth non-speech frames.

The obtained results are $A$ = 0.766, meaning that the system missed 23.4% of human-labeled speech frames, and $B$ = 0.984, indicating that false alarms happened in only 1.6% of non-speech frames. The pair $(A, B)$ is indicated by a cross in Fig. 5.

In order to compare our algorithm with a single-channel method, we applied a threshold on frame energy to determine an alternate speech/non-speech classification, and computed the corresponding $(A, B)$ pairs. By varying the energy threshold we obtained the continuous curve in Fig. 5. The energy-based system performed noticeably lower, especially if we consider that

- To achieve a similar performance on speech frames ($A$ = 0.766) the energy-based system induces many false alarms on non-speech frames ($B$ = 0.854).
- To achieve a similar performance on non-speech frames ($B$ = 0.984) the energy-based system induces a much lower proportion of correct speech frames ($A$ = 0.428).

Since the recording contained only single source segments and a small background noise, we can anticipate that the energy-based system would perform even worse on recordings with spontaneous, overlapping speech.

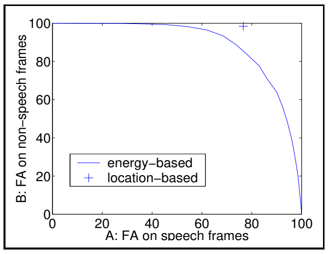


Figure 5: Speech/non-speech classification performance.

### 4.1.4 Global audio performance

On the located speech segments, we counted the number of frames falling below an azimuth error $\epsilon_{az}$ of $6^o$ before and after speech/non-speech classification. Results are reported in Table 1. For the energy-based, baseline system, we chose a SNR threshold such that the frame accuracy on GT non-speech frames would be the same as for the location-based system ($B$=0.984).

It can be seen that the location-based system classifies almost all correct location estimates as speech frames, as opposed to the energy-based system which misses many of them. The maximum azimuth error is also significantly reduced. This result is particularly important in the context of speaker tracking: we can expect the location-based system to detect speaker changes very well and with a very small delay.

| speech/non-speech classification technique | $\epsilon_{az}$ $\leq 6^o$ (frames) | $\epsilon_{az}$ $> 6^o$ (frames) | $\epsilon_{az}$ max (degrees) |
|---|---|---|---|
| none | 2081 | 626 | 179.6 |
| location-based | 2076 | 100 | 37.7 |
| energy-based | 1523 | 89 | 177.3 |

Table 1: Overall audio speaker localization performance.

## 4.2 AV tracking

Parameters in the model (dynamics and observations) were specified by hand, and kept fixed for all experiments. Regarding the dynamical models, the TPMs were defined by

$$\mathbf{T}_{mn}(\mathcal{R}^{sw}) = \begin{bmatrix} 0.90 & 0.05 & 0.05 \\ 0.05 & 0.90 & 0.05 \\ 0.05 & 0.05 & 0.90 \end{bmatrix},$$

and

$$\mathbf{T}_{mn}(\mathcal{R}^{\overline{sw}}) = \begin{bmatrix} 0.950 & 0.025 & 0.025 \\ 0.025 & 0.950 & 0.025 \\ 0.025 & 0.025 & 0.950 \end{bmatrix}.$$

Furthermore, in the current implementation we have used identical motion parameters for all models $p_{mn}(x_t|x_{t-1})$, $A_{nn} = \left[\begin{smallmatrix} 2 & -1 \\ 1 & 0 \end{smallmatrix}\right]$, $B_{nn} = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right]$, and the white noise process $w_t$ has standard deviations for translation and scaling equal to 4 and 0.0001, respectively. For the shape-based observations, the number of measurement lines $L = 16$, each with length 20 pixels, and a standard deviation $\sigma$ in Eq. 7 equal to 5 pixels. Finally, for the IS function, we used a uniform prior for all mixture components, ($\lambda_t^j = \frac{1}{N_t^a + N_t^v}$), and a radius $r_k = 70$ pixels.

The results should be fully appreciated by looking directly at the AV sequences accompanying this paper[1]. The sequences are encoded in AVI (using DIVX), and RealMedia formats.

Fig. 6 shows the results of tracking four speakers engaged in a two-minute conversation in the meeting room (3000 frames), using 500 particles (weighted mean of the posterior in red, estimated by Eq. 3, and standard deviation from the mean in yellow). Speakers talk at a natural pace, and one of the participants stands up and addresses the others from the projection screen and whiteboard areas (see sequence `demo–test–seq1.avi`). Audio data are non-continuous (1815 audio samples in 3000 frames), and there is a considerable amount of overlapping speech. The tracker is automatically instantiated when a person starts talking, and remains for the most part in accurate track across participants for the rest of the sequence with small latency. In case of overlapping speech, the tracker locks onto only one speaker. Tracking is more challenging for the objects observed by the wide-view camera due to distance from the array and object size.

Table 2 presents an objective evaluation of the results, using a semi-automatically generated ground-truth (GT) of speaker segments, which consists of the camera index and the approximate speaker's head centroid in the corresponding image plane for each speaker segment. Segments with overlapping speech were not considered for evaluation, as our tracker does not output results for multiple simultaneous speakers.

We define two performance measures, and present results averaged over ten runs of the particle filter. The first measure is the error on the estimated camera indices $\epsilon_k$ (with range [0,1]). Results are presented for each camera, and for all the results combined. Camera indices are very well estimated for cameras 1 and 2. Most errors arise from the wide-view camera: the tracker has a tendency to lock onto the speakers at the table, given their shorter distance to the microphone array, which captures small audio activity. Globally, the camera indices were correctly estimated in 88.73% of the frames labeled in the GT.

The second performance measure is the median over time of the error in the image plane $\epsilon_{(T^x,T^y)}$, between the GT and estimated mean 2-D positions, computed over all those frames for which the estimated camera index was correct. The main source of error for cameras 1 and 2 is the fitting of the contour template onto the neck contour rather than onto the chin. It is interesting to notice that for the wide-view camera, the camera index was more difficult to predict, but the error in 2-D for the cases for which the index was correct remained small.

| error type | modality | $\text{cam}_1$ | $\text{cam}_2$ | $\text{cam}_3$ | global |
|---|---|---|---|---|---|
| $\epsilon_k(\times 10^{-2})$ | AV | 1.91 | 0.31 | 25.00 | 11.27 |
| | | (0.09) | (0.09) | (0.39) | (0.18) |
| $\epsilon_{(T^x,T^y)}$ | AV | 1.88 | 1.69 | 0.40 | 1.00 |
| | | (0.08) | (0.18) | (0.01) | (0.03) |
| | A | 11.39 | 11.86 | 10.60 | 11.20 |
| | V | 4.57 | 4.88 | 2.19 | 3.52 |

Table 2: AV tracking results. The std of each measure is shown in parenthesis; the units of $\epsilon_{(T^x,T^y)}$ are pixels.

For comparison, the results of audio-only localization are also shown in Table 2. Figures have been computed only taking into account frames with detected speech. The errors reported correspond to the median over

[1] `www.idiap.ch/~gatica/av–tracking–multicam.html`.

time of the mininum error between the GT and all the audio estimates available at each frame. Such errors are the combined effect of 3-D localization and AV calibration. We have also included the results obtained by a visual-only, histogram-based tracker [24] initialized by hand at each speaker turn. Errors are slightly higher, although visually the performance is similar. For this sequence, AV fusion has shown better performance than each independent modality.

The benefit of using color and audio information compared to audio-only in the IS function (Eq. 17) and in the measurement process (Eq. 6) can be appreciated in the sequence `video2.avi`. In this video, the GMM defining the IS function consists only of audio estimates. It can be observed that tracking is less accurate than the observed in the previous example, due to the inherent limitations of single microphone-array estimates and the errors introduced by 3-D-to-2-D mapping that are not being improved by the use of color, as in `demo−test−seq1.avi`.

Performance of the method on a cluttered background is shown in Fig. 7, and in videos `demo−test−seq2.avi` (1200 frames), and `demo−test−seq3.avi` (800 frames). The sequences display a four-party conversation with a fifth person walking in the room, and creating visual distractions by approaching the speakers. The tracker can get momentarily distracted by the walking person, or by the background visual clutter, but recovers in all cases. Although not shown here, work for a single-camera version of the system has shown that our formulation can also handle reinitialization in cases of total AV occlusion [7].


Figure 6: Tracking speakers in the meeting room. Frames 100, 500, 1100, 1500, 1900, and 2700.

Figure 7: Tracking speakers with visual distractions. Frames 313, 563, 564, 640, 645, and 731.

# 5 Conclusions

We have shown that AV fusion via mixed-state i-particle filters makes good use of the complementary advantages of individual modalities for speaker tracking in a multi-camera room. Our method can consistently track speakers in multi-party conversations. Current work concentrates in the generalization of the method to a multiple-object tracker, which involves the integration of person-dependent color object models, and the consistent labeling of tracked objects along time and across cameras.

# 6 Acknowlegments

This work was funded by the Swiss NCCR on Interactive Multimodal Information Management (IM)2, and the European IST project M4. We thank our colleagues at IDIAP (Thierry Collado, Florent Monay, and Darren Moore) for technical support.

# References

[1] M. Beal, H. Attias, and N. Jojic. Audio-video sensor fusion with probabilistic graphical models. In *Proc. ECCV*, May 2002.
[2] A. Blake and M. Isard. *Active Contours*. Springer-Verlag, 1998.
[3] R. Cutler and L. Davis. Look who's talking: Speaker detection using video and audio correlation. In *Proc. IEEE ICME*, New York, July 2000.
[4] R. Cutler, Y. Rui, A. Gupta, J. Cadiz, I. Tashev, L. He, A. Colburn, Z. Zhang, Z. Liu, and S. Silverberg. Distributed meetings: A meeting capture and broadcasting system. In *Proc. ACM Multimedia Conference*, 2002.
[5] J. DiBiase, H. Silverman, and M. Brandstein. Robust localization in reverberant rooms. In M. Brandstein and D. Ward, editors, *Microphone Arrays*, chapter 8, pages 157–180. Springer, 2001.
[6] A. Doucet, N. de Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001.
[7] D. Gatica-Perez, G. Lathoud, I. McCowan, J.-M. Odobez, and D. Moore, Audio-visual speaker tracking with importance particle filters". In *Proc. IEEE ICIP*, Barcelona, Sep. 2003.
[8] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis*. Chapman and Hall, 1995.
[9] S. M. Griebel and M. S. Brandstein. Microphone array source localization using realizable delay vectors. In *Proceedings of the 2001 IEEE Workshop on Applications of Signal Proce ssing to Audio and Acoustics*, pages 71–74, 2001.
[10] I. Isard and A. Blake. ICONDENSATION: Unifying low-level and high-level tracking in a stochastic framework. In *Proc. ECCV*, Freiburg, June 1998.
[11] I. Isard and A. Blake. A Mixed-State CONDENSATION Tracker with Automatic Model-Switching. In *Proc. IEEE ICCV*, 1998.
[12] V. Kettnaker and R. Zabih. Bayesian multi-camera surveillance. In *Proc. IEEE CVPR*, June 1999.
[13] S. Khan, O. Javed, Z. Rasheed, and M. Shah. Human tracking in multiple cameras. In *Proc. IEEE ICCV*, July 2001.
[14] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Trans. on ASSP*, 24(4):320–327, August 2000.
[15] G. Lathoud and I. McCowan. Location based speaker segmentation. In *Proc. IEEE ICASSP*, Hong Kong, April 2003.
[16] L. Lu and H. J. Zhang. Content analysis for audio classification and segmentation. *IEEE Transactions on Speech and Audio Processing*, 10(7), 2002.
[17] J. McGrath. *Groups: Interaction and Performance*. Prentice-Hall, 1984.
[18] D. Moore. The IDIAP Smart Meeting Room. *IDIAP Communication 02-07*, 2002.
[19] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke. The meeting project at ICSI. In *Proc. of the Human Language Technology Conference*, San Diego, CA, March 2001.
[20] K. Parker. Speaking turns in small group interaction: a context sensitive event sequence model. *Journal of Personality and Social Psychology*, 54(6):965–971, 1988.
[21] H. Pasula, S. Russell, M. Ostland, and Y. Ritov. Tracking many objects with many sensors. In *Proc. IJCAI-99*, Stockholm, 1999.
[22] V. Pavlovic, A. Garg, and J. Rehg. Multimodal speaker detection using error feedback dynamic bayesian networks. In *Proc. IEEE CVPR*, Hilton Head Island, SC, 2000.
[23] V. Pavlovic, J. M. Rehg, and J. MacCormick. Learning switching linear models of human motion. In *Proc. NIPS*, Denver, Dec. 2000.
[24] P. Perez, C. Hue, J. Vermaak, and M. Gagnet. Color-based probabilistic tracking. In *Proc. ECCV*, May 2002.
[25] J. Vermaak, M. Gagnet, A. Blake, and P. Perez. Sequential Monte Carlo fusion of sound and vision for speaker tracking. In *Proc. IEEE ICCV*, Vancouver, July 2001.
[26] A. Waibel, M. Bett, F. Metze, K. Ries, T. Schaaf, T. Schultz, H. Soltau, H. Yu, and K. Zechner. Advances in automatic meeting record creation and access. In *Proc. IEEE ICASSP*, Salt Lake City, UT, May 2001.
[27] C. Wang, S. Griebel, and M. Brandstein. Robust automatic video-conferencing with multiple cameras and microphones. In *Proc. IEEE ICME*, New York City, July 2000.
[28] D. Zotkin, R. Duraiswami, and L. Davis. Multimodal 3-D tracking and event detection via the particle filter. In *IEEE Work. on Det. and Rec. of Events in Video*, July 2001.