IDIAP RESEARCH REPORT

# Further Applications of Sector-Based Detection and Short-Term Clustering

Guillaume Lathoud [a,b]

IDIAP–RR 06-26

May 2006

[a] IDIAP Research Institute, CH-1920 Martigny, Switzerland
[b] Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland

**IDIAP Research Institute** **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 06-26

# Further Applications of Sector-Based Detection and Short-Term Clustering

Guillaume Lathoud

May 2006

**Abstract.** This paper presents an effective implementation of detection-localization of multiple speech sources with microphone arrays. In particular, the Scaled Conjugate Gradient descent is used for fast and precise localization, within a pre-detected volume of space, and a close to real-time implementation is provided. An unsupervised approach to speech/non-speech discrimination is also proposed. The integrated system is then successfully applied to segmentation of spontaneous multi-party speech, as found in meetings. Based on this system, the unsupervised speaker clustering task is then investigated, using distant microphones only. This task is challenging due to the poor quality of the signal and the fast-changing speaker turns encountered in spontaneous speech. An extension of the BIC criterion to multiple modalities is proposed, allowing to combine the strengths of speaker location information – useful in the short term – and acoustic speaker information, i.e. MFCCs – useful in the longer term. A dramatic improvement in speaker clustering results is obtained by the combined approach, as compared with the acoustic-alone approach, and results are close to those obtained with close-talking microphones. Finally, an initial investigation on automatic audio-visual calibration is exposed.

# 1 Introduction

This report investigates the integration of two previous works [1, 2], with application to the meeting room domain. On one hand, in [1], instantaneous detection of multiple speakers was investigated, using a microphone array. It was proposed to discretize the space around a microphone array into sectors, and to detect, for each sector, whether it contains active speakers or not. It was shown to effectively detect multiple speakers, while at the same time reducing the search space for subsequent speaker localization – but precise localization was not implemented. On the other hand, in [2], a short-term clustering approach was proposed to group and denoise instantaneous (precise) location estimates, with an application to segmentation of multi-party speech, as found in meetings. The present paper examines the missing step between the two works: instantaneous (precise) speaker localization within each active sector. A complete, fully online implementation for multispeaker detection-localization is proposed, that achieves near real-time detection *and* localization of multiple speech sources. After evaluating the quality of the proposed implementation, we apply it to several tasks, namely speech segmentation, speech/non-speech classification and speaker clustering, including an extension of the BIC criterion [3, 4] to multiple modalities. Finally, we present an initial study on automatic audio-visual calibration. Overall, the idea is *not* to build a perfect system in a given condition, but rather a robust system, which requires little manual intervention from the user. A particular emphasis is thus given to self-tuning/automatic approaches. The rest of this section gives a brief overview of each aspect of the report.

First, an implementation of multisource detection-localization is detailed. It uses sector-based detection-localization [1, 5, 6] as a first step, followed by classical "point-based" audio speaker localization [7]. Both detection and localization require virtually no tuning. For each time frame, sector-based detection-localization simply reduces the search space to a few active regions of space or "sectors" (Fig. 1b), for a limited computational cost. An automatic threshold selection is used [6], that relies on unsupervised fitting of a probabilistic model. For localization, within each active sector, the SRP-PHAT metric [8] is optimized, using the Scaled Conjugate Gradient (SCG) algorithm [9]. Compared to other gradient descent algorithms, the SCG has the advantage of requiring virtually no parameter tuning. Second-order information (the Hessian) is approximated using the gradient only. This is particularly friendly in the localization case, where the dimensionality is quite large, e.g. 512 frequency bins. In practice, it appeared to require only 5 or 6 iterations to converge. Tests on the AV16.3 corpus [10] validate the approach, including a variety of sequences, with static and moving speakers. Up to 3 simultaneous speakers are correctly detected and located, including with an alternate low-cost implementation (GCC-PHAT time-delay estimation [11, 12] instead of SRP-PHAT).

Second, the instantaneous location estimates are clustered locally in space and time, to form many short utterances (speech segmentation). This is done applying the Short-Term Clustering (STC) approach [2]. One issue addressed by this paper is the presence of erroneous location estimates due to machine interference, e.g. a beamer. Location-dependent MFCC features are proposed, that are derived from the sector-based detection. We propose two ways of exploiting these sector-based MFCCs: by comparing to a threshold the standard deviation over an audio segment of the sector-based C0, at virtually no cost, or by EM fitting of a full covariance Gaussian Mixture Model (GMM), which is more costly. Evaluation on a multispeaker speech/silence segmentation task, on the M4 Corpus [13], shows that overall results compare well with those of close-talking microphones (lapels), with an improvement on overlapped speech.

Third, the speaker clustering task is considered, on dynamic multi-party speech in meetings, using distant microphones only (microphone array). The classical single channel "acoustic only" MFCC-based approach is considered, starting with many clusters (one per speech segment), and using the Bayesian Information Criterion (BIC) for merging [3, 4]. As it is difficult to build reliable speaker models from the very short speech utterances encountered in spontaneous speech (e.g. 1 or 2 seconds), this paper proposes to merge acoustic cues (MFCCs) and location cues (speaker direction) in a principled manner. This is done at the model selection level, by merging the BIC scores of the two modalities. Intuitively, location information helps to disambiguate in the short-term between, for example, two speakers having fast alternating speaker turns, since they are located differently. Results show that the merged approach provides a consistent improvement over the "acoustic only" speaker clustering, in terms of Diarization Error Rate (DER) [14]. Results are superior to those of a state-of-the-art GMM/HMM approach [15].

However, at the signal processing level, it appears that none of the audio-only speaker clustering approaches benefit from delay-sum multichannel filtering, which confirms conclusions from another work [16]. Here, best results were always obtained extracting "acoustic" information (MFCCs) from a single channel of the microphone array, while using the whole microphone array to extract location information only. Furthermore, it was observed that a sharp change of location (standing up, moving away to a presentation screen), *consistently* led to several different clusters for the same speaker (e.g. far and close locations). One interpretation is that room distortion varies, depending on speaker and microphone locations. Although standard processing techniques were tried to dereverberate [17] and/or denoise [18] the signal prior to MFCC extraction, this "artificial difference" remained the same, *consistently* leading to several clusters for the same person. On the other hand, cepstral normalization techniques (mean or variance normalization) cannot be used, because they destroy important speaker-specific information. To conclude, it appears that signal processing techniques developed to enhance Automatic Speech Recognition in adverse environments, such as beamforming [19], dereverberation [17] and denoising [18], *cannot* be used to enhance speaker identity cues. This accumulation of evidence suggests a major shift of paradigm for modelling speaker identity with audio cues.

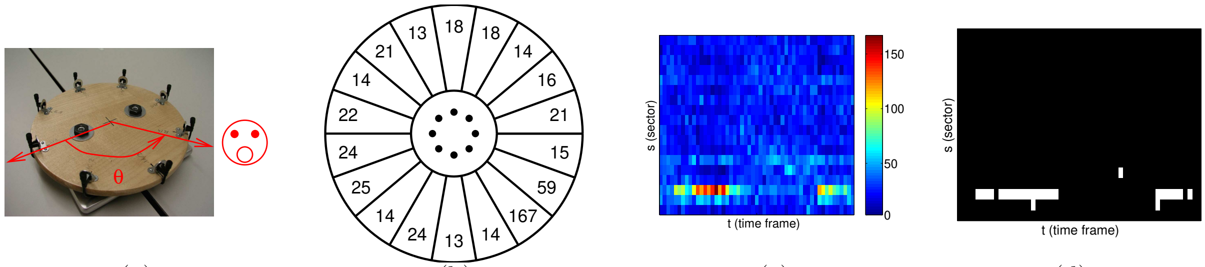
Finally, considering these limitations of audio-only speaker clustering, a practical alternative is to add the visual modality. In particular, this paper considers the need of audio-visual methods for a prior audio-visual calibration, as in [20]. Ideally, no technical knowledge should be required from the user of a meeting videoconferencing system, with automatic segmentation/summarization/clustering features, as those discussed in this article. A preliminary experiment investigates *automatic* audio-visual calibration, only requiring one user to do a short walk around the room once, while speaking.

To summarize, the main contributions of this paper can be listed as follows:

- Section 2: A practical implementation for *joint* detection and localization of multiple speakers, with microphone arrays. The search space is drastically reduced in the first step [6], for a limited cost. This in turn allows for easy use of gradient descent techniques, such as the robust SCG [9], as well as multiple microphone arrays. The approach is validated on real, moving, multispeaker data [10].
- Section 2: A simple methodology to jointly evaluate detection and localization without an arbitrary precision threshold. It is based on EM fitting [21] of a Gaussian + Uniform model on direction estimation errors.
- Section 3: An unsupervised speech/non-speech classification approach, that uses "location-dependent" MFCCs. It is shown to be useful to remove extraneous audio sources (e.g. beamer), as well as some body motion noises.
- Section 4: A generic approach to merge multiple modalities, within the BIC model selection framework [3]. It is successfully applied to the merge of essentially long-term "acoustic" speaker identity information (MFCCs), with short-term "location" cues (speaker direction from the array).
- Section 4: Speaker clustering experiments on the M4 Meeting Corpus [13] that validate this merging approach, while highlighting important signal processing issues with respect to speaker identification with distant microphones.
- Section: 5: A simple approach to automatic audio-visual calibration in an indoor environment.

A complete MATLAB/C implementation is fully available, along with test data, for Sections 2 and 3. In particular, optimized C code is available for GCC-PHAT, the sector-based detection-localization, the SCG descent and the Time-Delay Estimation (TDE). Please refer to the following link:

`http://mmm.idiap.ch/Lathoud/2006-multidetloc`




(a) Microphone array

(b) Discretization into sectors, and activeness values (one time frame)

(c) Activeness values (many time frames)

(d) Final binary decisions (inactive or active)

Figure 1: Detection-localization, step 1: around a microphone array (a), the space is divided into sectors (b). For a given time frame, speech "activeness" is estimated for each sector (b), using SAM-SPARSE-MEAN [1, 5]. Repeating this over time produces a spatio-temporal "activeness" pattern (c). The final "inactive/active" decision (d) is taken by comparing "activeness" to a threshold. The threshold is determined automatically, as in [6].

# 2 Multisource Detection-Localization

## 2.1 Proposed Approach

Many existing localization methods can benefit from a prior step that reduces the search space to a limited volume of space, as explained in the review in [22]. A fully integrated multisource detection-localization system also needs to determine active and silent frames. As a first step, we thus propose to restrict the search in space and time *jointly*, using the recently proposed SAM-SPARSE-MEAN approach [1, 5]. With respect to localization, it appears to be advantageous over classical detection features such as energy or SNR estimation. Indeed, in the experimental study reported in Annex A, we determine whether a more conservative threshold on the detection feature implies a smaller spatial localization error. It turns out to be the case of SAM-SPARSE-MEAN, but not energy or SNR. Following is a description of the proposed two-step approach for multisource detection-localization:

**Step 1: Sector-based detection-localization (Fig. 1).** The space around a microphone array (Fig. 1a) is discretized into volumes of space called "sectors" (Fig. 1b). For each time frame and each sector, the SAM-SPARSE-MEAN "activeness" value [5] is computed from the multiple microphone signals (Figs. 1b and 1c). Based on a given activeness value, deciding whether or not there is at least one active source in the corresponding sector of space is done by comparing the activeness value to a threshold (Fig. 1d). The threshold is obtained automatically, from unsupervised statistical modelling and a user-defined performance target [6], such as False Alarm Rate (FAR) or False Rejection Rate (FRR). See [23] for a formal definition of FAR and FRR. As a result, for each time frame we will obtain a pattern of active sectors: zero, one or multiple sectors (Fig. 1).

After trying various performance targets including FAR and FRR, we selected the following compromise. First, "conservative" detection is realized by selecting a threshold corresponding to the target FAR = 0.5%. Second, for each active (sector, time frame) detected so far, within the same sector and within a window of time frames, e.g. T = ±0.5 sec, a second threshold is applied, that corresponds to the "less conservative" target FRR = 0.5%.

**Step 2: Point-based localization.** Within each active sector of space (white squares in Fig. 1d), run a "point-based" search, i.e. determine the most likely point of origin of speech. For localization, we chose to use a parametric approach [24], which leads to optimize spatial location parameters with respect to a cost function such as SRP-PHAT [8]. By repeating this process within each active sector, we can potentially achieve multisource localization.

Note that based on the studies in [8] and [6], it is strictly equivalent to:

1. Maximize SRP-PHAT [8],
2. Minimize the Phase Domain Metric (PDM) [6],
3. Maximize the delay-sum power of multiple signals from which magnitude was normalized (unit magnitude, phase only is kept) [8, 6].

We selected the PDM to express a cost function $\mathcal{C}$ to be minimized, because it leads to relatively simple mathematical expressions, as shown further below. $\mathcal{C}$ is a function of (1) the *hypothesized* spatial location of the source, (2) the *observed* phase values for each microphone and each frequency of a Digital Fourier Transform (DFT) decomposition. The location estimate should minimize $\mathcal{C}$.

The issue of parametric methods that try to find parameters that minimize a cost $\mathcal{C}$ is the cost of exploring the entire search space. Typically, a gradient descent approach could require many steps to converge, depending on its initialization point. Our approach reduces the cost in two ways: first, the search is limited to the active sector. Second, minimization is done through the Scaled Conjugate Gradient (SCG) algorithm [9]. SCG was chosen because of its speed efficiency, relative to other descent methods, due to its efficient approximation of second order information (Hessian) using first order derivatives only (gradient). In our case, it requires only a few iterations to converge (typically 5 to 10). Moreover, although SCG does have numerical parameters, in practice they do not require tuning. The paper [9] contains very clear step-by-step instructions describing its implementation.

The crucial point is to express the gradient of $\mathcal{C}$ [6] with respect to location parameters. We are using a Uniform Circular Array (UCA), see Fig. 1a, which is known to realize most spatial discrimination in terms of direction, especially azimuth (direction angle in the horizontal plane), while having very poor resolution in terms of radius. Therefore, spherical coordinates are preferred. Moreover, in order to enforce the $r > 0$ constraint without adding any specific constraint to the gradient descent framework, we introduce “logspherical” coordinates:

$$(\ \alpha,\ \beta,\ \gamma\ ) \quad \in \quad \mathbb{R}^3 \tag{1}$$

Where

- $\alpha$ is the azimuth angle in radians,
- $\beta$ is the elevation angle in radians,
- $\gamma \stackrel{\text{def}}{=} \log r$, and $r > 0$ is the radius in meters,

of the hypothesized source location, relative to the center of the microphone array, whose geometry is known. On the other hand, the expression of the PDM cost $\mathcal{C}$ [6] and its gradient, is more natural in terms of Euclidean coordinates $(x,\ y,\ z)$, since Euclidean distances need to be expressed and derivated. We therefore express the gradient of $\mathcal{C}$ in Euclidean space, then convert it to logspherical space using the following conversion:

$$\begin{bmatrix} \frac{\partial\mathcal{C}}{\partial\alpha} \\ \frac{\partial\mathcal{C}}{\partial\beta} \\ \frac{\partial\mathcal{C}}{\partial\gamma} \end{bmatrix} = \begin{bmatrix} -y & x & 0 \\ -z\cos\alpha & -z\sin\alpha & r\cos\beta \\ x & y & z \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial\mathcal{C}}{\partial x} \\ \frac{\partial\mathcal{C}}{\partial y} \\ \frac{\partial\mathcal{C}}{\partial z} \end{bmatrix} \tag{2}$$

which is obtained from the decomposition:

$$\frac{\partial\mathcal{C}}{\partial\alpha} = \frac{\partial\mathcal{C}}{\partial x}\frac{\partial x}{\partial\alpha} + \frac{\partial\mathcal{C}}{\partial y}\frac{\partial y}{\partial\alpha} + \frac{\partial\mathcal{C}}{\partial z}\frac{\partial z}{\partial\alpha} \tag{3}$$

and similar decompositions for $\beta$ and $\gamma$. From Eq. 2, we can see that the additional cost of using logspherical coordinates is very small, as compared to Euclidean coordinates. Moreover, in a sanity check experiment, we verified that the SCG takes less iterations to converge in logspherical coordinates, than in Euclidean coordinates.

## 2.2 The Cost Function and its Gradient in Euclidean coordinates

This subsection reminds the mathematical definition of $\mathcal{C}$ [6], then expresses its gradient in Euclidean coordinates. It is presented in the most general case where (1) there can be multiple sources within each sector, and (2) only a subset of the frequency bins defined by the DFT are used to define $\mathcal{C}$.

Let $\mathbf{v} \stackrel{\text{def}}{=} (x, y, z)^{\mathrm{T}}$ a vector of Euclidean coordinates, an hypothesized source location. For a given pair $p$ of microphones $\mathbf{m}_1^{(p)}$ and $\mathbf{m}_2^{(p)}$ $(1 \leq p \leq P)$, the time delay in samples is:

$$\tau_p(\mathbf{v}) \stackrel{\text{def}}{=} \frac{f_s}{c}\left(||\mathbf{v} - \mathbf{m}_2^{(p)}|| - ||\mathbf{v} - \mathbf{m}_1^{(p)}||\right) \tag{4}$$

where $f_s$ is the sampling frequency in Hz and $c$ the speed of sound in the air in m/s (typically 342 m/s), and the coordinates of the two microphones are written:

$$\mathbf{m}_1^{(p)} \stackrel{\text{def}}{=} \left[x\left(\mathbf{m}_1^{(p)}\right), y\left(\mathbf{m}_1^{(p)}\right), z\left(\mathbf{m}_1^{(p)}\right)\right]^{\mathrm{T}} \tag{5}$$

$$\mathbf{m}_2^{(p)} \stackrel{\text{def}}{=} \left[x\left(\mathbf{m}_2^{(p)}\right), y\left(\mathbf{m}_2^{(p)}\right), z\left(\mathbf{m}_2^{(p)}\right)\right]^{\mathrm{T}} \tag{6}$$

Note that microphone pairs $p = 1 \cdots \mathrm{P}$ can be placed anywhere in the room. In particular, this allows for the use of multiple microphone arrays with exactly the same mathematical development.

Let $\{\mathbf{v}_n\} \stackrel{\text{def}}{=} (\mathbf{v}_1 \cdots \mathbf{v}_{\mathrm{N}})$ a set of N 3-D locations (e.g. multiple source locations in the same sector):

$$\mathbf{v}_n \stackrel{\text{def}}{=} [x_n, y_n, z_n]^{\mathrm{T}} \tag{7}$$

Let $(i_1, \cdots i_{\mathrm{N_f}})$ a list of $\mathrm{N_f}$ frequency bins, subset of $\{1, 2, \cdots, \mathrm{N_{BINS}}\}$, where $\mathrm{N_{BINS}}$ is the number of DFT positive frequency bins and $\{\omega_1, \omega_2, \cdots, \omega_{\mathrm{N_{BINS}}}\}$ are the associated digital frequencies $(0 \leq \omega_i \leq \pi)$.

We propose to minimize the following cost function [6], with respect to $\{\mathbf{v}_n\}$:

$$\mathcal{C}\left(\{\mathbf{v}_n\}, \{\omega_i\}, \left\{\hat{\theta}_i\right\}\right) \stackrel{\text{def}}{=} \frac{1}{\mathrm{N}} \sum_{n=1}^{\mathrm{N}} \frac{1}{\mathrm{N_f}} \sum_{j=1}^{\mathrm{N_f}} \frac{1}{\mathrm{P}} \sum_{p=1}^{\mathrm{P}} \sin^2\left(\frac{\hat{\theta}_p\left(\omega_{i_j}\right) - \gamma_p\left(\mathbf{v}_n, \omega_{i_j}\right)}{2}\right) \tag{8}$$

where:

- $\hat{\theta}_p(\omega_i)$ is the measured relative phase between the two microphones of pair $p$,
- $\gamma_p(\mathbf{v}_n, \omega_i) \stackrel{\text{def}}{=} \omega_i \cdot \tau_p\left(\mathbf{v}_n\right)$ is the theoretical relative phase between the two microphones of pair $p$.

In the following, to simplify the notation we assume $i_j = j$, as if the frequency bins where ordered $(i = 1, 2, \cdots, \mathrm{N_f})$, without loss of generality. Eq. 8 can be rewritten using $\sin^2 u = \frac{1}{2}(1 - \cos 2u)$:

$$\mathcal{C} = \frac{1}{2} - \frac{1}{2\ \mathrm{N}\ \mathrm{N_f}\ \mathrm{P}} \sum_{n,i,p} \cos\left(\hat{\theta}_p(\omega_i) - \gamma_p(\mathbf{v}_n, \omega_i)\right) \tag{9}$$

Let $\Delta_{n,i,p} \stackrel{\text{def}}{=} \cos\left(\hat{\theta}_p(\omega_i) - \gamma_p(\mathbf{v}_n, \omega_i)\right)$. We express the derivative of $\mathcal{C}$ with respect to one parameter $x_k$, where $1 \leq k \leq \mathrm{N}$:

$$\frac{\partial \mathcal{C}}{\partial x_k} = -\frac{1}{2\ \mathrm{N}\ \mathrm{N_f}\ \mathrm{P}} \sum_{n,i,p} \frac{\partial \Delta_{n,i,p}}{\partial x_k} = -\frac{1}{2\ \mathrm{N}\ \mathrm{N_f}\ \mathrm{P}} \sum_{i,p} \frac{\partial \Delta_{k,i,p}}{\partial x_k} \tag{10}$$

Since $\cos' u = -\sin u$, each term of the sum in Eq. 10 develops into:

$$\frac{\partial \Delta_{k,i,p}}{\partial x_k} = \frac{\partial}{\partial x_k}\left[\gamma_p\left(\mathbf{v}_k, \omega_i\right)\right] \cdot \sin\left(\hat{\theta}_p\left(\omega_i\right) - \gamma_p\left(\mathbf{v}_k, \omega_i\right)\right) \tag{11}$$

$$= \omega_i \frac{f_s}{c} \cdot \sin\left(\hat{\theta}_p\left(\omega_i\right) - \gamma_p\left(\mathbf{v}_k, \omega_i\right)\right) \cdot \frac{\partial}{\partial x_k}\left[||\mathbf{v}_k - \mathbf{m}_2^{(p)}|| - ||\mathbf{v}_k - \mathbf{m}_1^{(p)}||\right] \tag{12}$$

Using the relation $\frac{\partial b}{\partial a} = \frac{1}{2b}\frac{\partial}{\partial a}\left[b^2\right]$, we can write, for $q = 1$ or 2:

$$\frac{\partial}{\partial x_k}\left[||\mathbf{v}_k - \mathbf{m}_q^{(p)}||\right] \quad = \quad \frac{1}{2||\mathbf{v}_k - \mathbf{m}_q^{(p)}||}\cdot\frac{\partial}{\partial x_k}\left[||\mathbf{v}_k - \mathbf{m}_q^{(p)}||^2\right] \quad = \quad \frac{x_k - x(\mathbf{m}_q^{(p)})}{||\mathbf{v}_k - \mathbf{m}_q^{(p)}||} \tag{13}$$

Finally, for each source $k$:

$$\frac{\partial \mathcal{C}}{\partial x_k} \quad = \quad -\frac{f_s}{2\ c\ \mathrm{N}\ \mathrm{N_f}\ \mathrm{P}}\sum_{i,p}\left\{\omega_i \cdot \sin\left(\hat{\theta}_p(\omega_i) - \gamma_p(\mathbf{v}_k, \omega_i)\right)\cdot\left[\frac{x_k - x(\mathbf{m}_2^{(p)})}{||\mathbf{v}_k - \mathbf{m}_2^{(p)}||} - \frac{x_k - x(\mathbf{m}_1^{(p)})}{||\mathbf{v}_k - \mathbf{m}_1^{(p)}||}\right]\right\} \tag{14}$$

and a similar expression is obtained for $y_k$ and $z_k$.

### 2.3 Computational Cost

The computational cost of computing $\mathcal{C}$ (Eq. 8) and the all coordinates of its gradient (Eq. 14) is directly proportional to the product $\mathrm{N}\cdot\mathrm{N_{BINS}}\cdot\mathrm{P}$. In this subsection, we examine each of these factors, then define an implementation of Sections 2.1 and 2.2 that has potentially a lower cost. The purpose here is to achieve real-time detection-localization.

N: In the UCA setup considered here (Fig. 1a), we defined 20-degree sectors in terms of azimuth angle $\alpha$. In a meeting room environment, it is reasonable to assume at most 1 source per sector: $\mathrm{N} = 1$. In fact, we did try higher values of N on data where two persons are known to get much closer than 20 degrees (`seq18` in the AV16.3 corpus [10]). However, it did not provide any advantage over the $\mathrm{N} = 1$ case, possibly because the localization error of parametric methods such as SRP-PHAT/PDM is on the order of 2.5 degrees, in terms of standard deviation, which is too large to distinguish between sources within the same sector. All results in this paper use $\mathrm{N} = 1$.

$\mathrm{N_{BINS}}$: The number of frequency bins used in the cost and gradient can be reduced in two ways. First by reducing the total number of bins $\mathrm{N_{BINS}}$. For a time window of $\mathrm{N_{SAMPLES}}$ samples, any value of $\mathrm{N_{BINS}}$ below $\mathrm{N_{SAMPLES}}/2$ degrades the localization precision. However, it may still be acceptable for applications where the maximum precision is not required, for example speaker detection-localization in an environment such as a meeting, where people are seated and relatively well separated from each other. An example of study with various number of FFT bins is shown in Tab. 1.

| $\mathrm{N_{BINS}} =$ | $\mathrm{N_{SAMPLES}}/4$ | $\mathrm{N_{SAMPLES}}/2$ | $\mathrm{N_{SAMPLES}}$ |
|---|---|---|---|
| Number of correct location estimates | 700 | 826 | 814 |
| Std dev (degrees) | 2.959 | 2.862 | 2.767 |

Table 1: Effect of reducing the number of FFT bins, on the detection-localization result (single speaker sequence `seq01` from the AV16.3 corpus [10]). In this case the time-domain window size is $\mathrm{N_{SAMPLES}} = 512$, and $\mathrm{N_{BINS}} = \mathrm{N_{SAMPLES}}$ corresponds to a zero-padded FFT.

The second way to reduce the number of frequency bins is to select only "active" frequencies – typically spectral peaks – to define the subset $\{i_j\}$ (see Section 2.2). Based on [25], we propose the following restriction (illustrated by Fig. 2).

- For each time frame $t$, for each frequency bin $f$, the geometrical mean $m_{f,t}$ of the magnitudes is computed, across all microphones of an array.
- Frequency bins $\{i_j\}$ used to compute the cost $\mathcal{C}$ must be either a peak of magnitude ($m_{f,t} > \max(m_{f+1,t},\ m_{f-1,t})$), or right next to a peak frequency bin.
- Frequency bins $\{i_j\}$ must also be above the geometrical mean of all magnitudes $\{m_{1,t}\cdots m_{f,t}\cdots m_{\mathrm{N_{BINS}},t}\}$.

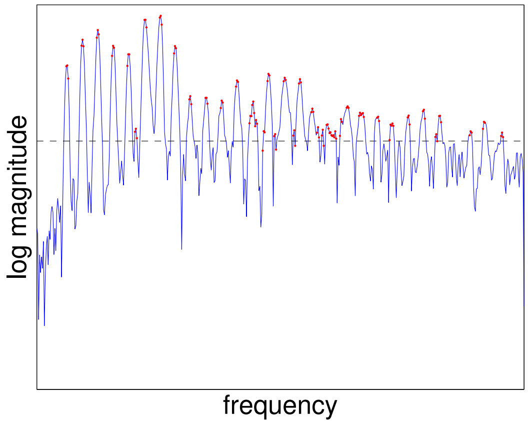



Figure 2: Example of bin selection: only the bins with magnitude above the geometric mean (horizontal dashed line), and at or next to a magnitude peak are selected (red dots).

P: since we need to work on short time frames (e.g. 32 ms), where both speech signal and location are stationary, a large enough number of microphone pairs is required to achieve a decent spatial resolution. This is well explained in [8]. Still, we did try to reduce the number of microphone in the array (and thus the total number of pairs). In our setup, using less than 6 microphones did not provide usable localization results. All results in this paper use, for each microphone array, the full number of microphone pairs $\mathrm{P} = M\,(M-1)\,/2$ where $M = 8$ is the number of microphones.

**Active sectors:** Finally, it is also possible to limit the maximum number of sectors to be searched, to a "reasonable" value. For example, with 20-degree sectors, the whole 360-degree range is spanned with 18 sectors. Thus, one can limit the search to the $\mathrm{N_{MAX}} = 6$ most active sectors.

**SCG iterations:** In practice, we found that 5 to 6 iterations are enough for the SCG descent to converge, when using the proposed cost function $\mathcal{C}$, i.e. in logspherical coordinates.

## 2.4 "FULL", "FAST" and "FASTTDE" Implementations

This section describes 3 different implementations of the 2-step approach for detection-localization, presented in Section 2.1. The goal is to compare the performance of a full-search implementation with that of low-cost, near real-time implementations. As previously mentionned, for the SCG descent we use only one location per active sector. All implementations are done in a fully online manner: the data is processed by blocks (e.g. 10 seconds), and model parameters for sector-based detection-localization [26] and short-term-clustering [2] are updated *at the end* of each block. Finally, concerning sector-based detection-localization, the Shifted Rice [26] was replaced with with a Shifted Erlang [18] for greater stability, because each block contains less data than tested previously.

**"FULL" implementation:** In the following, the abbreviation "FULL" refers to the original, unconstrained implementation described in Sections 2.1 and 2.2, that is:

- SCG is applied within all active sectors.
- For each SCG descent, at most 30 iterations. The search is initialized in the middle of the corresponding active sector.
- DFT is implemented using $\mathrm{N_{BINS}} = \mathrm{N_{SAMPLES}}$, which means zero-padding.
- All frequency bins are used.
- The frame shift is 10 ms, the frame length is 32 ms.

**"FAST" implementation:** On the contrary, in the following, "FAST" refers to a low-cost implementation using the following constraints:

- SCG is applied within, at most, the $\mathrm{N}_{\mathrm{MAX}} = 6$ most active sectors, according to the posterior probability of activity defined in [26].
- For each SCG descent, at most 10 iterations.
- DFT is implemented using $\mathrm{N}_{\mathrm{BINS}} = \mathrm{N}_{\mathrm{SAMPLES}}/2$, which, due to the hermitian symmetry, means as many Fourier coefficients $(2 \cdot \mathrm{N}_{\mathrm{BINS}})$ as there are time-domain samples.
- The frequency bin selection strategy described in Section 2.3.
- The frame shift is 16 ms, the frame length is 32 ms.

**"FASTTDE":** Finally, we also implemented a variant called "FASTTDE", where the SCG descent in "FAST" is replaced with a direct method based on time-delay estimation:

- From the 8-microphone circular array, two square subarrays are defined (microphones #1 #3 #5 #7 and microphones #2 #4 #6 #8).
- The time-domain GCC-PHAT function [11] is estimated through inverse FFT, for the two diagonal pairs of each subarray (#1-#5 and #3-#7, #2-#6 and #4-#8). It is typically upsampled (e.g. 10 or 20 times).
- For each active sector of space, and for each microphone pair, the Time-Delay Estimation (TDE) is implemented by finding the maximum of the time-domain GCC-PHAT function *within the range of time-delays corresponding to this sector.*[1]
- For each active sector of space and for each subarray, the direction of the source is estimated as an azimuth[2], from the two time-delays, as in [12], Section 7.2. It is considered as valid only if within the sector. Since there are two subarrays, there may be two valid direction estimates, in which case we average them. We thus end up with zero or one azimuth direction estimate per active sector.

It is important to note that the sector-dependent range of permitted time-delays implicitly allows to have different time-delay values for different sectors. It can be seen as a principled way to apply the single-source GCC-PHAT method to a multisource problem.

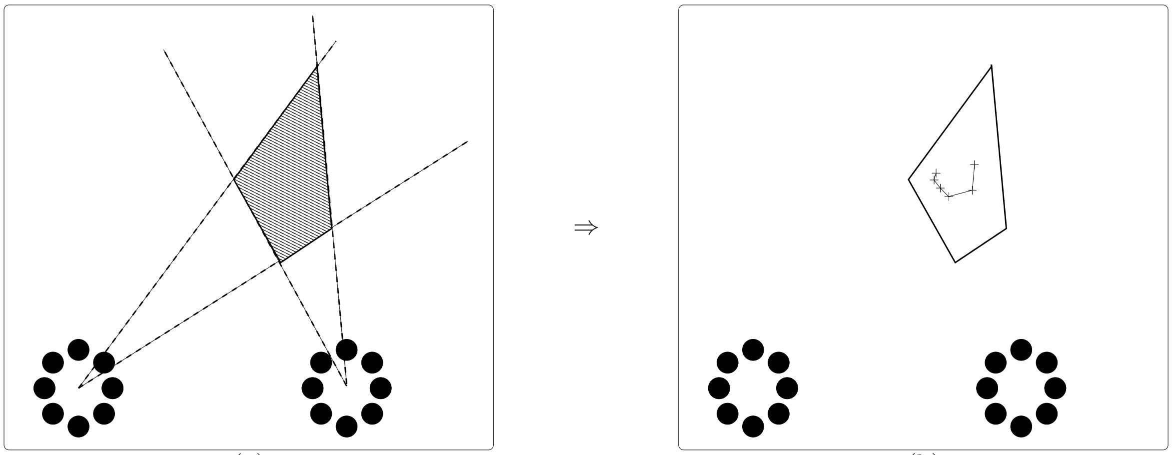
**Code optimization:** all approaches are available in a Matlab implementation that includes C routines for GCC-PHAT, SAM-SPARSE-MEAN, SCG descent and TDE, through the MEX interface:
`http://mmm.idiap.ch/Lathoud/2006-multidetloc`

## 2.5 Multiple Microphone Arrays

The PDM cost function $\mathcal{C}$ defined by Eq. 8 puts no constraint on the placement of microphone pairs. Thus, the SCG descent can be applied to multiple microphone arrays. Spatial resolution is then much finer than with one array (at least in the Fresnel area defined by the multiple microphone arrays), so $\mathcal{C}$ and its gradient are better expressed using Euclidean coordinates $[x, y, z]^{\mathrm{T}} \in \mathbb{R}^3$. However, without prior information, a complete search through the entire $\mathbb{R}^3$ space would be intractable in real-time. We thus propose to apply the same 2-step approach as in Section 2.1 to the case of multiple arrays. In the first step, for each array independently, each sector is determined to be active or inactive. The intersections of active sectors are then limited volumes of space in which to search for the location(s) of the source(s), through SCG descent in terms of 3-D location. See Fig. 3 for an example.

[1]This range can be estimated once, offline, from the geometry of the array and the sectors. It is then stored for all further computations.

[2]Note that elevation is also estimated during this process, but it is not very precise with UCAs.



(a)
Sector-based
detection-localization

$\Rightarrow$

(b)
Gradient descent
(SCG)

Figure 3: Proposed 2-step approach, with two microphone arrays.

One advantage of this sector-based approach is that these volumes, determined by the intersection of any pair of sectors, can be precomputed once for all, and stored in memory. As for the second step (SCG descent), the implementation for multiple arrays is exactly the same as for a single array, as already mentioned in Section 2.2.

## 2.6 Evaluation Method

Since we have proposed an integrated multisource detection-localization system, we need to jointly evaluate:

- Localization: The spatial precision resulting from the 2-step approach. The goal is to check whether the system is providing decent localization or not, across various test cases.
- Detection: The number of *correctly* localized speakers at each time frame. The goal is to check whether the detection part is able to (1) detect an active speaker when he can be correctly localized (this differs from the single channel speech/silence discrimination task), (2) detect multiple active speakers at the same time, (3) correctly detect silences.

While spatial precision can be defined in terms of bias and standard deviation, it is not clear what "*correctly* localized" means. Usually a threshold is arbitrarily defined on the localization error (e.g. 5 degrees). This is subject to caution, since the localization error may vary across test cases, so defining a single threshold for all test cases may not be the best choice. On the other hand, defining a separate threshold for each test case does not permit to compare results between test cases.

Thus, we propose to avoid the use of a threshold, replacing it with a statistical approach. Instead of first estimating the localization precision, then estimating the number of correctly localized speakers, both are estimated jointly, as mathematical expectation quantities, based on a simple Gaussian + Uniform model $\mathcal{M}_{\mathrm{G+U}}$, which can be fitted using the EM algorithm [21]. An example of fit of this model is shown in Fig. 4. The data used for the fit is the localization error $\delta$.

Formally, let $\delta$ be the angle error, i.e. the difference between a given result location estimate and its closest location in the ground-truth (e.g. in degrees). The probability density function (pdf) of the $\mathcal{M}_{\mathrm{G+U}}$ model is defined as a mixture of two components:

$$p\left(\delta \middle| \mathcal{M}_{\mathrm{G+U}}\right) \stackrel{\mathrm{def}}{=} P_{\mathrm{G}} \cdot f_{\mathrm{G}}\left(\delta\right) + \frac{P_{\mathrm{U}}}{180}. \tag{15}$$

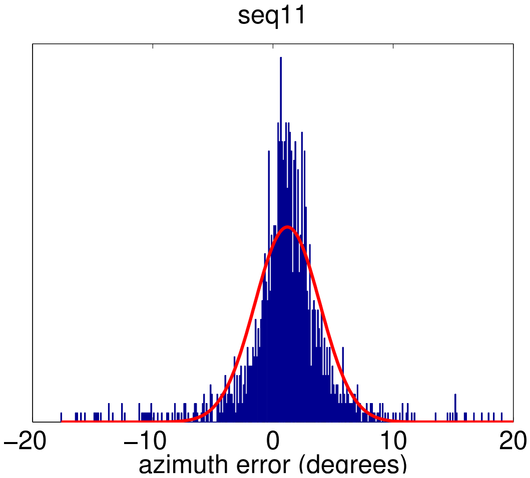



Figure 4: Example of fit of the Gaussian + Uniform model $\mathcal{M}_{\mathrm{G+U}}$ on the localization error $\delta$. $\mathcal{M}_{\mathrm{G+U}}$ is used for evaluation of the detection-localization. `seq11` is a recording with a single moving speaker, from [10]. The blue histogram represents the distribution of localization errors. The red curve represents the Gaussian pdf, modelling "correct" estimates. The uniform pdf is not represented.

where $P_{\mathrm{G}}$ and $P_{\mathrm{U}}$ are the priors of "correctly localized" and "incorrectly localized", and $f_{\mathrm{G}}$ is assumed to be a Gaussian pdf with parameters $\mu_{\mathrm{G}}$ and $\sigma_{\mathrm{G}}$:

$$f_{\mathrm{G}} \quad \sim \quad \mathcal{N}\left(\mu_{\mathrm{G}}, \sigma_{\mathrm{G}}\right) \tag{16}$$

Then the desired quantities for performance evaluation of localization and detection can all be directly estimated from the $\mathcal{M}_{\mathrm{G+U}}$ model:

- $\mu_{\mathrm{G}}$ and $\sigma_{\mathrm{G}}$ are the proposed estimates of the bias and standard deviation of the localization error, for an audio source that was detected and "correctly located".
- $100 \cdot P_{\mathrm{G}}$ is the percentage of location estimates that are correct.
- In a given time frame $t$, the number of correctly located speakers $\hat{\mathrm{N}}_{\mathrm{C}}(t)$ is estimated as a conditional expectation:

$$\begin{aligned} \hat{\mathrm{N}}_{\mathrm{C}}(t) &\overset{\mathrm{def}}{=} \mathbf{E}\left\{\mathrm{N}_{\mathrm{C}}(t) \mid \mathcal{M}_{\mathrm{G+U}}\right\} && (17) \\ &= \sum_{n=1}^{\mathrm{N}(t)} p\left(\mathrm{G} \mid \delta_n(t),\ \mathcal{M}_{\mathrm{G+U}}\right) && (18) \\ &= \sum_{n=1}^{\mathrm{N}(t)} \frac{P_{\mathrm{G}} \cdot f_{\mathrm{G}}\left(\delta_n(t)\right)}{P_{\mathrm{G}} \cdot f_{\mathrm{G}}\left(\delta_n(t)\right) + \frac{P_{\mathrm{U}}}{180}} && (19) \end{aligned}$$

where $\mathrm{N}(t)$ is the number of location estimates at time frame $t$, given by the detection-localization system (0, 1 or more), $\delta_1(t) \cdots \delta_n(t) \cdots \delta_{\mathrm{N}(t)}(t)$ the corresponding angle errors. Each angle error $\delta_n(t)$ is defined as the difference between a location estimate and the closest location in the ground-truth.

An example of histogram of all $\hat{\mathrm{N}}_{\mathrm{C}}(t)$ values for all time frames $t$ is shown in Fig. 5. In order to summarize this result, we discretize $\hat{\mathrm{N}}_{\mathrm{C}}(t)$ into integer bins, summing all the values $\hat{\mathrm{N}}_{\mathrm{C}}(t)$ such that: $0 \leq \hat{\mathrm{N}}_{\mathrm{C}}(t) < 0.5$, $0.5 \leq \hat{\mathrm{N}}_{\mathrm{C}}(t) < 1.5$, $1.5 \leq \hat{\mathrm{N}}_{\mathrm{C}}(t) < 2.5$, etc.

A complete Matlab code to conduct the evaluation, along with examples, are available at:

`http://mmm.idiap.ch/Lathoud/2006-multidetloc`

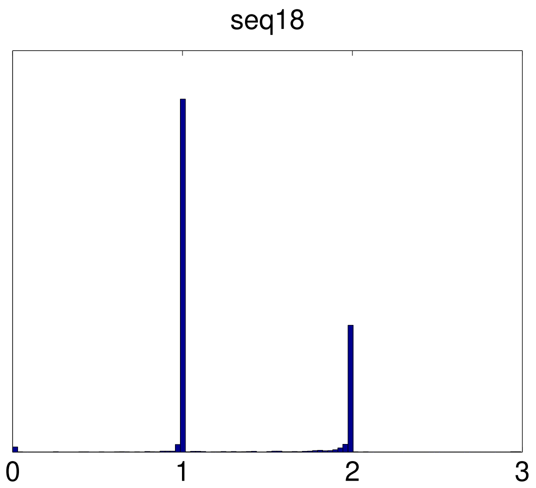



Figure 5: Histogram of $\hat{\mathrm{N}}_{\mathrm{C}}(t)$, the estimated number of correctly localized speakers in a 2-speaker sequence. Note that $\hat{\mathrm{N}}_{\mathrm{C}}$ is a continuous value.

### 2.7 Experimental Protocol

The three implementations FULL, FAST and FASTTDE were run on 8 different recordings of the freely available AV16.3 database [10]. We present azimuth results obtained with one 8-microphone circular array. 3 cameras were used to reconstruct the "true" location of each speaker, relative to the microphones, as described in [10].

Two recordings were made with static speakers:

- `seq01`: 1 speaker at 16 different locations, facing the arrays,
- `seq37`: 3 simultaneous speakers, 2 seated and 1 standing at 5 different locations, facing the arrays.

Six recordings were made with moving speakers:

- `seq11`: 1 moving speaker, speaking continuously, facing the arrays.
- `seq15`: 1 moving speaker, speaking discontinuously, with long silences.
- `seq18`: separation test: 2 moving speakers getting as close to each other as possible, facing the arrays.
- `seq24`: crossing test: 2 moving speakers passing in front of each other, facing the arrays.
- `seq40`: partial occlusion: test similar to `seq37`, except that the standing speaker is continuously moving.
- `seq45`: motion & full occlusion: 3 moving speakers, walking around while speaking continuously.

In all cases, after running multisource detection-localization (FULL, FAST or FASTTDE), we removed noisy location estimates, using short-term clustering [2] followed by the cheap SNSLOW Speech/Non-Speech discrimination (Section 3.2). The focus here is the quality of the detection-localization ; please refer to Section 3 for more details on SNS.

## 2.8 Results and Discussion

Fig. 6 depicts examples of localization results, along with the ground-truth locations of the various speakers. Tab. 2 presents localization results on all 8 recordings, in terms of bias, standard deviation and percentage correct, as given by the $\mathcal{M}_{\mathrm{G+U}}$ evaluation (Section 2.6). As for detection, Tab. 3 presents the distribution of $\hat{\mathrm{N}}_{\mathrm{C}}(t)$, the number of speakers correctly detected *and* located, estimated as described in Section 2.6. Finally, Tab. 4 shows the effective computational cost.

**Global results:** For all eight recordings except `seq40`, visual inspection of location estimates against the ground-truth confirms that FULL, FAST and FASTTDE (1) effectively detect and locate multiple sources, (2) exhibit a low number of spurious location estimates. (1) is confirmed by the distributions shown in Tab. 3, which have significant components with 2 or more speakers. (2) is confirmed by the "percentage correct" in Tab. 2, which is often between 95 % and 100 %. The failure on `seq40` (Fig. 6) is reflected by the high standard deviation values in Tab. 2. This may have two possible (non-exclusive) explanations: (1) strong interference between the three speech signals, due to *partial* occlusions, (2) low power received at the microphone array, due to the downward orientation of the speakers'heads [27], because they are reading books aloud. It contrasts with the success `seq45` (Fig. 6), which also contains three simultaneous speakers, but *full* occlusions. For all three methods, the standard deviation on `seq45` is within the range of other, "easier" recordings.

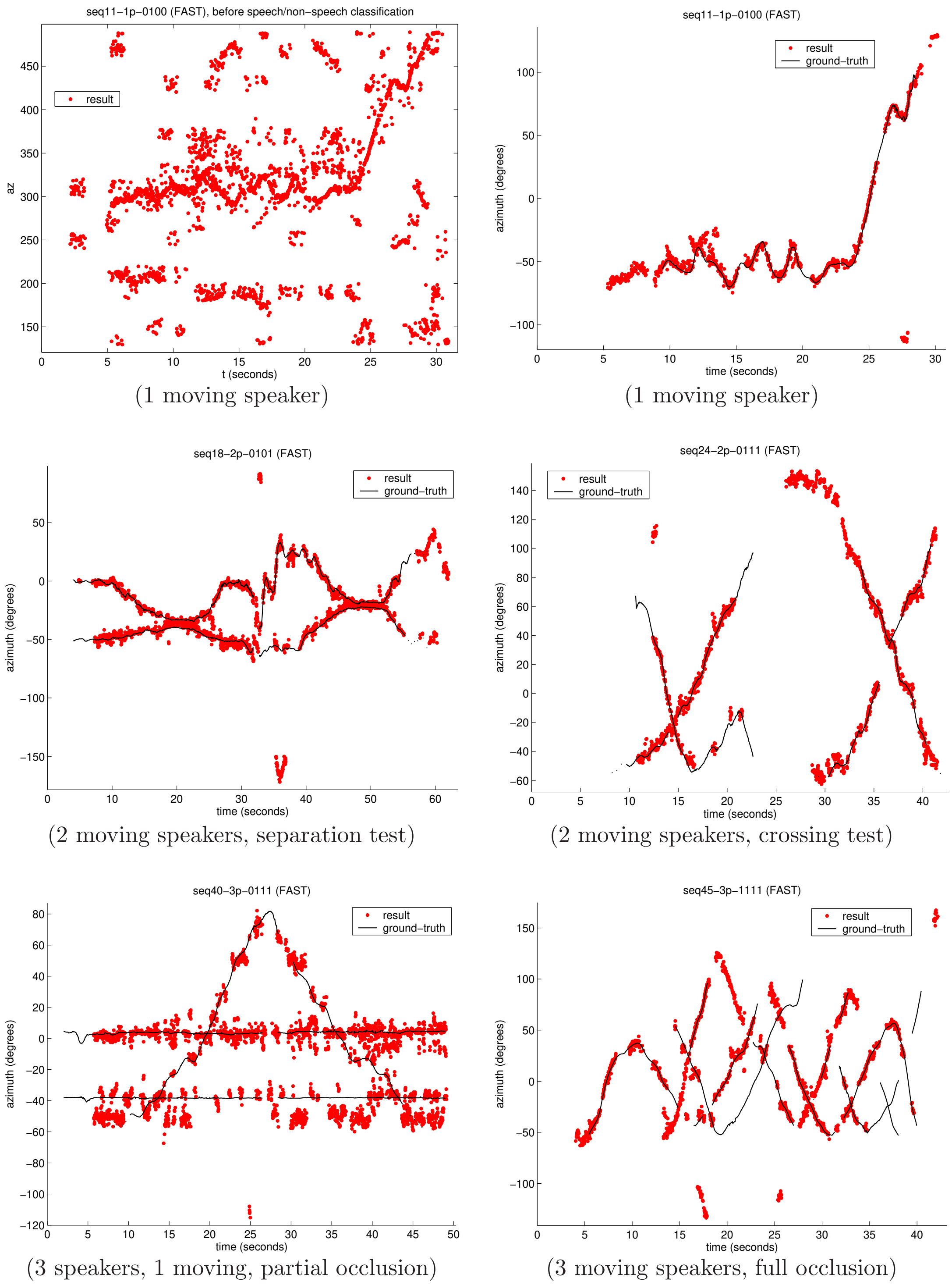
**Comparison between FULL and FAST:** Looking at the averages in Tab. 2, two observations can be made. The FAST implementation exhibits similar precision to the FULL implementation, but a higher percentage of correct estimates. This is possibly a positive impact of the frequency bin selection strategy used in FAST (Fig. 2). The total duration of correct estimates (not reported) is lower for FAST than for FULL, by 4.2 %. This can be explained by the limit on the number of active sectors, in the case of FAST. Finally, the computational cost of FAST is much lower, in fact close to real-time speed (Tab. 4), although part of the implementation is still in Matlab. Overall, we can state that FAST is an effective multisource implementation of a parametric search for *multiple* local maxima of SRP-PHAT. Various possibilities arise to achieve real-time implementation, including:

- Parallelization of the SCG descent over several dedicated CPUs (one per active sector).
- Integration within a tracking framework (use past knowledge through an update mechanism).
- Implementation in C of the remaining MATLAB code (see the "Input" column in Tab. 4).

**Comparison between FAST and FASTTDE:** In the case of FASTTDE, the localization cost becomes negligible ("TDE" column in Tab. 4), but the localization precision is degraded (standard deviation in Tab. 2). This degradation corresponds to the extensive study in [8], that compares GCC-PHAT and SRP-PHAT. Also, some noisy location estimates appear, as shown by the "2" column in Tab. 3, for the single-speaker sequences `seq01`, `seq11` and `seq15`. Since FASTTDE behaves decently in terms of detection, one could use it to greatly reduce the search space, then run FAST within the "speech" (sector, time frame) pairs.

**Specific multispeaker case:** `seq37` is a case with only simultaneous speakers (2 or 3), for extended periods of time (about one minute for each combination of speakers and locations). We compared visually the location estimates with the ground-truth. On the positive side, the 2 or 3 speakers are correctly detected and located over a long run (e.g. a minute), *as long as they are at comparable distances from the array.* On the other hand, whenever one speaker was standing about two times further from the array than the other two (seated) speakers, he was nearly completely missed. Further analysis showed that this is because the "missed" speaker is dominant in only few frequency bins because of his further distance from the array. This induces a low SAM-SPARSE-MEAN value [5], and therefore a low posterior probability of activity [26].

To conclude, we can state that the proposed detection-localization system was proved able to detect and correctly locate up to 3 simultaneous speakers, including on a highly dynamic 3-speaker case (e.g. FAST on `seq45`). On the other hand, the failure observed on the "partial occlusion" case suggests that further research should investigate joint spectrum-location estimation. Work going in that direction includes a signal subspace microphone array approach relying on a frequency-dependent model for each source [28], and a data-driven approach for binaural localization [29].




(1 moving speaker) (1 moving speaker)

(2 moving speakers, separation test) (2 moving speakers, crossing test)

(3 speakers, 1 moving, partial occlusion) (3 moving speakers, full occlusion)

Figure 6: Result (red dots) of the detection-localization ("FAST" implementation, followed by short-term clustering and SNSLOW). The ground-truth (black curves) is derived from the cameras. Gaps are due to the mouth of a person being occluded on at least one camera **(gaps are not related to silences)**.

| Recording | FULL (all active sectors) Bias | Std dev. | % corr. | FAST (up to 6 sectors) Bias | Std dev. | % corr. | FASTTDE (up to 6 sectors) Bias | Std dev. | % corr. |
|---|---|---|---|---|---|---|---|---|---|
| seq01 (1, static) | -0.47 | 2.65 | 96.4 | -0.33 | 2.60 | 97.6 | 0.38 | 3.46 | 98.7 |
| seq37 (3, static) | -0.05 | 2.63 | 90.3 | 0.63 | 2.68 | 95.8 | 2.75 | 6.57 | 97.4 |
| seq11 (1, moving) | 1.18 | 2.78 | 87.3 | 1.29 | 2.67 | 92.6 | 2.36 | 5.69 | 97.3 |
| seq15 (1, moving) | 0.30 | 1.76 | 79.1 | 0.17 | 1.77 | 89.3 | 1.19 | 5.30 | 88.0 |
| seq18 (2, moving, separation test) | 0.32 | 2.09 | 93.4 | 0.39 | 2.06 | 96.2 | 0.61 | 3.18 | 98.1 |
| seq24 (2, moving, crossing test) | 0.16 | 2.99 | 90.4 | 0.22 | 2.99 | 96.3 | -0.00 | 4.04 | 98.6 |
| seq40 (3, moving, partial occlusion) | -1.31 | **5.37** | 100 | -1.94 | **6.02** | 99.7 | -0.16 | **6.44** | 100 |
| seq45 (3, moving, full occlusion) | 0.36 | **3.30** | 91.3 | 0.38 | **2.46** | 88.3 | 0.16 | **3.65** | 93.7 |
| Average | 0.06 | 2.95 | 91.0 | **0.10** | **2.91** | **94.5** | 0.91 | 4.79 | 96.5 |

Table 2: Localization precision, in degrees, along with the percentage of correct location estimates.

| Recording | FULL 0 | 1 | 2 | 3 | 4 | FAST 0 | 1 | 2 | 3 | 4 | FASTTDE 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| seq01 | 1.4 | 98.3 | 0.4 | 0.0 | 0.0 | 1.2 | 98.5 | 0.3 | 0.0 | 0.0 | 0.1 | 89.3 | 10.5 | 0.0 | 0.0 |
| seq37 | 0.6 | 62.9 | **35.2** | **1.3** | 0.0 | 0.4 | 68.0 | **30.4** | **1.3** | 0.0 | 0.2 | 50.7 | **40.7** | **8.2** | 0.1 |
| seq11 | 3.7 | 95.3 | 1.1 | 0.0 | 0.0 | 2.1 | 97.1 | 0.8 | 0.0 | 0.0 | 0.6 | 82.2 | 17.2 | 0.0 | 0.0 |
| seq15 | 2.4 | 97.1 | 0.5 | 0.0 | 0.0 | 2.1 | 97.7 | 0.2 | 0.0 | 0.0 | 2.2 | 78.9 | 18.3 | 0.6 | 0.0 |
| seq18 | 0.6 | 65.7 | **33.5** | 0.2 | 0.0 | 0.6 | 72.2 | **27.1** | 0.1 | 0.0 | 0.5 | 56.2 | **34.2** | 9.1 | 0.0 |
| seq24 | 1.6 | 73.4 | **24.1** | 0.9 | 0.0 | 0.5 | 76.9 | **21.9** | 0.7 | 0.0 | 0.3 | 76.7 | **21.1** | 1.9 | 0.0 |
| seq40 | 0.0 | 51.7 | **40.4** | **7.5** | 0.5 | 0.0 | 54.5 | **38.7** | **6.4** | 0.5 | 0.0 | 41.1 | **41.8** | **15.6** | 1.5 |
| seq45 | 0.3 | 57.2 | **30.4** | **12.1** | 0.0 | 0.6 | 66.9 | **28.7** | **3.9** | 0.0 | 0.0 | 74.7 | **19.0** | **5.2** | 1.1 |

Table 3: Distribution of the number of correct simultaneous location estimates (percentage of frames). For recordings with multiple simultaneous speakers, the multispeaker cases are in bold face.

| Recording | FULL Input | SCG | Total | FAST Input | SCG | Total | FASTTDE Input | TDE | Total |
|---|---|---|---|---|---|---|---|---|---|
| seq01 | 0.70 | 9.56 | 13.29 | 0.28 | 0.42 | 1.54 | 0.43 | 0.07 | 1.29 |
| seq37 | 0.68 | 20.93 | 26.93 | 0.27 | 0.91 | 2.84 | 0.43 | 0.12 | 2.15 |
| seq11 | 0.69 | 19.04 | 24.39 | 0.27 | 0.73 | 2.21 | 0.41 | 0.10 | 1.76 |
| seq15 | 0.69 | 9.68 | 14.10 | 0.27 | 0.43 | 1.65 | 0.42 | 0.06 | 1.46 |
| seq18 | 0.70 | 25.50 | 31.52 | 0.25 | 1.02 | 2.82 | 0.42 | 0.12 | 1.94 |
| seq24 | 0.69 | 19.02 | 24.43 | 0.28 | 0.76 | 2.37 | 0.42 | 0.11 | 1.78 |
| seq40 | 0.67 | 27.19 | 33.55 | 0.28 | 0.97 | 2.76 | 0.43 | 0.11 | 1.96 |
| seq45 | 0.67 | 22.71 | 28.71 | 0.28 | 0.85 | 2.53 | 0.43 | 0.11 | 1.91 |
| Average | 0.69 | 19.20 | 24.62 | 0.27 | 0.76 | 2.34 | 0.42 | 0.10 | 1.78 |

Table 4: Effective computational cost: computation duration divided by recording duration. **(real time = 1)** We used a MATLAB/C implementation on a Pentium 4, with 3.2 GHz CPU speed and 1 GB of RAM. "SCG" is the time spent doing SCG descent only. "TDE" is the time spent doing TDE-based localization only. "Input" is the time spent reading and buffering wave files (variations due to MATLAB). The cost of FFT and GCC-PHAT is very small (around 0.003 real time duration).

# 3 Speech/Non-Speech Classification

In an environment with human sound sources only, relying on the SAM-SPARSE-MEAN detection approach [6], as in the above sections, may be sufficient to discriminate speech from silence, as illustrated by the experiments reported above. However, the type of sound source may well be less constrained, even indoors: for example, machines such as a beamer and laptops may be used in a meeting. In such a case, even the wideband source = speech source assumption, implicitly made when using SAM-SPARSE-MEAN, is not enough to discriminate between speech and non-speech.

This section thus proposes a further extension of SAM-SPARSE-MEAN to the Speech/Non-Speech (SNS) classification task. The spectrum is filtered in a location-dependent manner, thus producing "Sector-Based MFCCs". Two SNS classifiers are proposed: SNSLOW and SNSGMM. SNSLOW uses a fixed threshold on the Sector-Based C0, for virtually no cost. SNSGMM models Sector-Based MFCCs in an unsupervised manner, with a full covariance matrix GMM. This model is then splitted in two, to discriminate between speech activity and machine activity. Multispeaker speech segmentation experiments are reported, that validate the approach on the M4 Corpus of meetings [13].

## 3.1 Sector-Based MFCCs

The SAM-SPARSE-MEAN approach introduced in [1, 5] relies on the following sparsity assumption. Within each (DFT) frequency band of the spectrum, only one sector of space is assumed dominant. For each active source, we propose to extract MFCC acoustic features that depend on the sector in which the source is located. For each sector, the frequency spectrum is filtered in a binary manner, setting to zero the magnitude at a frequency bin if a sector is not dominant in that frequency:

$$m'_{s,f,t} \stackrel{\text{def}}{=} \begin{cases} m_{f,t} & \text{if sector } s \text{ dominant at time t} \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

However, since setting magnitude to zero may introduce artificial dynamics in the cepstrum domain, the spectrum is floored to a non-zero value $\sigma$ corresponding to the average background noise level, as in Unsupervised Spectral Subtraction [18]:

$$m''_{s,f,t} \stackrel{\text{def}}{=} \max\left(1, \frac{m'_{s,f,t}}{\sigma}\right). \tag{21}$$

MFCC features, specific to each sector $s$, can then be extracted from the resulting spectrum $m''_{s,f,t}$. To conclude, the following steps need to be followed, in order to extract sector-based MFCCs:

1. Select one microphone in the array, compute the magnitude spectrogram $\{m_{f,t}\}$, where $\{\cdot\}$ denotes a set.

2. Estimate the average noise level $\sigma$ from the spectrogram as described in [18].

3. For each sector $s$, compute the binary-filtered spectrum $m'_{s,f,t}$ (Eq. 20).
   Floor it with $\sigma$ to obtain $m''_{s,f,t}$ (Eq. 21).

4. For each sector $s$, extract MFCC features from spectrum $m''_{s,f,t}$.

This way of "separating" spectra from different sources is very approximate for two reasons. First, it does not use the precise spatial location of each source, but only their sector index. Second, the sector-based binary filter is derived from an *average* delay-sum over a volume of space [5]. It is thus certainly suboptimal, as compared to approaches such as beamforming [24], that focus towards a point in space. Finally, the magnitude spectrum of only one microphone is extracted and modified. Nevertheless, it is sufficient to build an unsupervised speech/non-speech classifier, as described in the next section, and tested further below. The computational cost is negligible.

A complete Matlab implementation of sector-based MFCC extraction is available at:
`http://mmm.idiap.ch/Lathoud/2006-multidetloc`

## 3.2 Low-Cost Speech/Non-Speech Classifier (SNSLOW)

As a baseline system, we propose to use measures of wideband activeness and non-stationarity: SAM-SPARSE-MEAN and the Sector-Based C0 coefficient, respectively. All non-speech segments are discarded, only speech segments are kept. A speech segment is defined as a segment that has:

$$\text{standard deviation of Sector} - \text{Based } C_0 \quad > \quad 0.6 \tag{22}$$

*and* at least two "wideband frames". A wideband frame is defined as verifying the condition:

$$p\,(\text{wideband}) \quad > \quad \theta\,(\text{FAR} = 1\%) \tag{23}$$

where $p\,(\text{wideband})$ is given by a probabilistic model fitted on SAM-SPARSE-MEAN in an unsupervised manner [26]. $\theta\,(\text{FAR} = 0.1\%)$ is determined in an unsupervised manner, as in [6]. It corresponds to a target FAR of 0.1%.

The fixed threshold value of 0.6 on the sector-based C0 coefficient is justified by the fact that sector-based MFCCs are derived from a normalized spectrum (Eq. 21).

## 3.3 Full Covariance GMM Speech/Non-Speech Classifier (SNSGMM)

One approach for Speech/Non-Speech (SNS) classification is to train a model on a set of recordings, against a known ground-truth segmentation, and to test on another set of recordings, where the unknown segmentation is to be estimated [30]. However, this type of approach runs the risk of "overfitting" the data seen during training, thus not performing well on unseen data that widely differs from the training data, as for example, different microphone characteristics, different types of noises, different types of room reverberation, etc.

An alternative proposed here, is to treat each set of data separately, through an unsupervised approach. A general (simple) assumption is needed, that is hoped to be valid in many different environments. Although an unsupervised approach may not perform as well as a training/testing approach, in terms of maximum performance, it is expected to better adapt to a wider range of environments – to generalize – without need for manual intervention.

The proposed approach relies on the following assumption: the various dimensions of the MFCC features of speech signals are correlated, while for machines such as beamer, there is no correlation (in particular, if the machine noise is stationary). This leads to the following approach:

- Fit a full covariance matrix GMM on sector-based MFCC features of active sectors, using the Expectation-Maximization (EM) algorithm [21].

- While doing EM, some of the components of the GMM may need to be collapsed to diagonal covariance matrices. This typically happens when a full covariance matrix becomes badly conditioned, thus not invertible within the numerical limits of the computer. In such a case, it is constrained to be diagonal, and the EM algorithm continues with the new constrained model.

- After convergence, the GMM is separated into two GMMs: diagonal components in one GMM, to model noise sources, non-diagonal components in the other GMM, to model speech sources.

Finally, one question remains: how to choose the dimensionality of the model? In our case, this means how to choose the number of GMM components? This can be done by trying various number of GMM components, e.g. from 2 to 10, and picking the result with maximum BIC score [3].

To summarize, the unsupervised SNS classification is implemented as follows:

1. Extract sector-based MFCCs from active (sector, time frame) pairs. Reduce the size of the data to a fixed number of vectors, e.g. 10000, by ordering sector-based MFCCs along C0, and picking vectors at regular intervals.
2. For $c = 2 \cdots 10$, fit a full-covariance GMM on the reduced data, with $c$ components. Some of the components may collapse to diagonal-covariance Gaussians.
3. Pick the GMM with maximum BIC score [3].
4. Split it into two GMMs: one with diagonal components, the other with non-diagonal components.
5. Using these two GMMs, it is possible to evaluate how likely an audio segment (series of sector-based MFCC vectors, as in Section 3.1) is to contain speech rather than noise.

**Final decision:** All non-speech segments are discarded, only speech segments are kept. A speech segment is defined as a segment that has:

$$\text{standard deviation of Sector} - \text{Based } C_0 \quad > \quad 0.6 \tag{24}$$

*and* at least two "speech frames". A speech frame is defined as verifying the conditions:

$$p\,(\text{wideband and non} - \text{noisy}) \quad > \quad \theta\,(\text{FAR} = 1\%) \tag{25}$$

where a simplifying independence assumption gives the posterior probability $p\,(\text{wideband and non} - \text{noisy}) = p\,(\text{wideband}) \cdot p\,(\text{non} - \text{noisy})$, and $\theta\,(\text{FAR} = 1\%)$ is a threshold on the posterior. $\theta\,(\text{FAR} = 1\%)$ is determined in an unsupervised manner, as in [6]. It corresponds to a target FAR of 1%. Each of the three quantities used in the final decision is defined as follows:

- Standard deviation of Sector-Based $C_0$: For each segment, the standard deviation of the sector-based $C_0$ coefficient, obtained from the sector-based MFCCs (Section 3.1). If it is less than a threshold (0.6), then the segment is considered as non-speech. This fixed value of 0.6 is reasonable given that the sector-based MFCCs are derived from a normalized magnitude spectrum (Eq. 21).
- $p\,(\text{wideband})$: The posterior probability of speech activity in a (sector, time frame), as given by the model of SAM-SPARSE-MEAN [26]. A sound source emitting a wideband signal is characterized by a large SAM-SPARSE-MEAN value, which in turn implies a large posterior value $p\,(\text{wideband})$.
- $p\,(\text{non} - \text{noisy})$: The posterior probability of "non-noisiness", as given by the two-GMM SNS model described in Section 3.3.

A Matlab implementation of the final SNS decision is available at:
`http://mmm.idiap.ch/Lathoud/2006-multidetloc`

### 3.4 Application to Location-Based Speech Segmentation of Meetings

In a previous work [2], an 8-microphone UCA was used to segment spontaneous multi-party speech, using instantaneous single source localization on all time frames, followed by short-term clustering (Fig. 7, left side). In tests on the M4 Corpus of meetings [13], high accuracy was obtained in terms of precision (PRC) and recall (RCL). Indeed, this translated into a segmentation comparable to that obtained with close-talking microphones (lapels), with an important improvement on overlapped speech. However, in [2], the SNS discrimination task was not addressed, since only location was used. Prior knowledge (average location of humans, and machines) was thus necessary to eliminate noise sources such as a beamer.

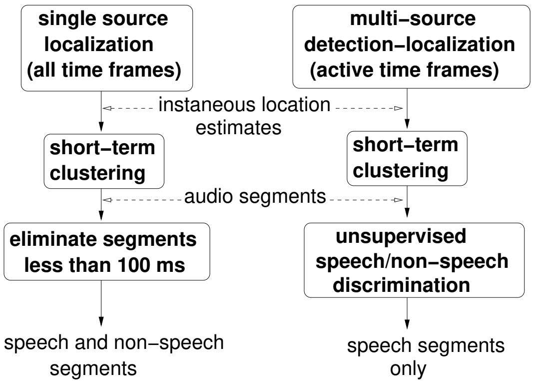



Figure 7: Speech segmentation with microphone arrays: the two implementations evaluated in Section 3.4. **Left:** the single source implementation that was proposed in [2], without any active frame detection or speech/non-speech discrimination. **Right:** the multi-source speech detection-localization approach proposed in this paper.

In this subsection, we propose to modify this implementation as follows (Fig. 7, right side):

- As a first step, multisource detection [6] and localization (Section 2) are included. Note that both were implemented in a fully online manner, by updating the detection model [26] at the end of every 10-second block. The result is, for each time frame, zero, one or more instantaneous location estimates.
- The short-term clustering [2] is also implemented in a fully online manner, where the probabilistic model is updated at the end of each 10-second block. The result is a cluster tag given to each location estimate. Note that, since multiple location estimates may be available at a given time frame, the "bounded computational load" described in [2] is ensured by defining $\mathrm{T_{short}}$, $\mathrm{T_{past}}$ and $\mathrm{T_{future}}$ as numbers of location estimates – as opposed to number of time frames in [2].
- As a final step, sector-based MFCCs (Section 3.1) are extracted for each active (location, time frame). Then each short-term cluster is classified between speech and non-speech, using either SNSLOW (Section 3.2) or SNSGMM (Section 3.3). Only speech clusters are kept.
- The first and last frames of each speech cluster define a speech segment.

## 3.5 Experimental Protocol

We evaluated the proposed multisource speech segmentation systems (Fig. 7) on the M4 Corpus [13], using 24-dimensional Sector-Based MFCC vectors. The test corpus includes 18 short meetings from a publicly available database (`http://mmm.idiap.ch`). The total amounts to 1h45 of multichannel speech data. In all meetings, an independent observer provided a very precise speech/silence segmentation. Because of this high precision, the ground-truth includes many very short segments. Indeed, 50% of the speech segments are shorter than 0.938 seconds.

For the both single source and multisource systems, we used $\mathrm{T_{short}} = \mathrm{T_{past}} = 7$. We found there was little difference between $\mathrm{T_{future}} = 1$ and the $\mathrm{T_{future}} = 7$ value used previously [2]. For the multisource system we used the FAST implementation, with either SNSLOW or SNSGMM.

To do the evaluation, each speech segment is attributed to the closest speaker in space, as given by a ground-truth location. The speech/silence segmentation of all speakers within a meeting is evaluated in terms of precision (PRC), recall (RCL) and F-measure ($\mathrm{F} = 2 \times \mathrm{PRC} \times \mathrm{RCL} / (\mathrm{PRC} + \mathrm{RCL})$), as in [2]. Note that only the segmentation is evaluated here. For speaker clustering investigations, please refer to the Section 4.

| | Single source, with manual SNS [2] | Multisource, with automatic SNS SNSLOW | SNSGMM |
|---|---|---|---|
| PRC | 79.7 ( 55.4 ) | 83.6 (66.6) | 83.8 ( 71.8 ) |
| RCL | 94.6 ( 84.8 ) | 89.5 (86.8) | 90.9 ( 82.0 ) |
| F | **86.5 ( 67.0 )** | **86.3 (74.6)** | **87.2 ( 75.7 )** |

Table 5: Segmentation results on the M4 Corpus (the higher, the better). Only distant microphones are used. Values are percentages, brackets indicate results on overlapped speech only. The multisource system uses the FAST implementation.

| Speech/non-speech decision granularity | Result without post-processing | Result with post-processing |
|---|---|---|
| Individual location estimate | 48.1 | 84.6 |
| Cluster of location estimates | 83.1 | 87.2 |

Table 6: F-measure on the M4 corpus, for two different granularities of the SNS classification. In all cases, the FAST multisource detection-localization system is used, with SNSGMM. "Post-processing" uses basic morphological operators on the speech segmentation, along the time axis: dilation, closure, opening and erosion (parameters tuned on a separate set of 3 meetings, to maximize the F-measure.)

### 3.6 Results

**Comparison single source/multisource:** Results are given in Tab. 5. In the single source case, speech/non-speech discrimination was done by hand. In the multisource case, it was automatic. Several observations can be made:

- The multisource system (SNSGMM) brings a slight improvement over the single source system, in terms of F-measure. This is already a success given that the multisource method includes an automatic SNS discrimination, which the single source method does not. PRC and RCL are slightly changed, most likely due to the elimination of non-speech segments in the multisource case.
- A major improvement is seen on overlaps (for both SNSLOW and SNSGMM), which validates the multisource approach, and corresponds to the fact that it can *effectively* detect and locate simultaneous speakers (Section 2.8).

Overall, we can conclude that the proposed multisource detection-localization approach effectively segments overlapping speech in an automatic manner, while rejecting background noise. Overlapping speech forms a non-negligible portion of spontaneous multi-party speech [31].

**Comparison SNSLOW/SNSGMM:** Although the results are slightly inferior for SNSLOW, they are pretty close to those of SNSGMM. This is remarkable, given the simplicity of SNSLOW. Inspection of several meetings showed that SNSGMM is better at removing noises such as body motion and paper shuffling. In order to evaluate a possible difference between the two, purely in terms of multisource detection-localization, we repeated the tests presented in Section 2, on the AV16.3 Corpus. Tab. 7 shows that there is very little difference between the two, on somewhat "clean" conditions.

**Comparison with/without short-term clustering:** We checked whether the short-term clustering is useful or not. An alternative speech/non-speech decision was implemented, which does not use the short-term clustering. For each location estimate independently, $p\,(\text{wideband and non}-\text{noisy})$ is compared to an automatic threshold, similarly to above. Based on the remaining location estimates, a meeting segmentation is produced. Results are reported in Tab. 6, where the advantage of using short-term clustering is clear. Short-term clustering not only leads to the best results, but it is also much less sensitive to post-processing of the speech segments.

| SNS method | SNSLOW | SNSGMM |
|---|---|---|
| % correct (duration) | 95.8 % (671.9 sec) | 97.3 % (664.9 sec) |
| bias | 0.39 | 0.38 |
| std. dev. | 2.46 | 2.54 |

Table 7: Comparison of two Speech/Non-Speech (SNS) discrimination methods on the detection-localization task (AV16.3 corpus, "FAST" implementation). `seq40` was not used to compute these averages because it produces large errors, both with and without GMM, as in Tab. 2.

# 4 Audio-Only Speaker Clustering

This section builds on the location-based short-term clustering method introduced in [2]. Audio segments are determined using location cues alone, irrespective of the audio content (speech or noise). Section 3 already presented a speech/non-speech (SNS) classification approach, that allows to drop audio segments containing noise. The present section addresses a remaining question: how to determine the identity of the speaker? More precisely, we investigate the speaker clustering task, where no enrollment data is available. While many works [4, 15] examined single, close-talking audio channel, constrained environments such as broadcast news data, this section investigates multichannel recordings of highly dynamic speech, as found in spontaneous multi-party speech.

The main contribution is a method to combine location cues from a microphone array and MFCC cues from one of the microphones in order to obtain a more robust speaker clustering result. This is particularly challenging given that only distant microphones are used. The idea is to provide an automatic system that does not force participants to wear any device at all.

## 4.1 Combining Two Modalities: Location Cues and Acoustic Cues

The purpose of speaker clustering is to separate a data set $X$ into a partition of "clusters" $X_1 \cdots X_R$. An optimization criterion is needed to choose the number $R$ of clusters and their boundaries. The ideal result is a one-to-one mapping, where each cluster $X_r$ $(1 \leq r \leq R)$ contains data from only one speaker, and the data from each speaker is contained in one cluster only.

In general, the Bayesian Information Criterion (BIC) [3] is defined for a set $X$ of $N$ observed data samples, and a model $M$ with $K$ free parameters:

$$\mathcal{BIC} \stackrel{\text{def}}{=} \log p\,(\,X|\,M) - \frac{\lambda}{2} K \cdot \log N \tag{26}$$

where $p\,(\,X|\,M)$ is the likelihood of the observed data $X$, given the model $M$. $\lambda$ is an adjusting parameter (in the original BIC definition $\lambda = 1$). BIC allows to compare various models with different number of free parameters $K$, by selecting the model with the maximum BIC value. If the values of the model parameters are selected in an optimal manner, the first term in Eq. 26 is supposed to increase when the number of free parameters $K$ increases. The second term in Eq. 26 is often called the "penalty term", as it penalizes models that have too many free parameters.

In the case of acoustic speaker clustering [4], a global model $M$ may not be available. In such a case, only local comparisons are made. In each comparison, the possibility of merging two clusters $X_1$ and $X_2$ into a merged cluster $X_{1+2}$ is evaluated by computing the difference in BIC scores:

$$\begin{aligned}\Delta_{1,2} &\stackrel{\text{def}}{=} \mathcal{BIC}_{1+2} - \mathcal{BIC}_1 - \mathcal{BIC}_2 &(27)\\ &= \log p\,(\,X_{1+2}|\,M_{1+2}) - \log p\,(\,X_1|\,M_1) - \log p\,(\,X_2|\,M_2)\\ &\quad - \frac{\lambda}{2}\left(K_{1+2} \cdot \log\left(N_1 + N_2\right) - K_1 \cdot \log N_1 - K_2 \cdot \log N_2\right) &(28)\end{aligned}$$

A typical iteration of agglomerative clustering is to test all possible pairs of clusters $(i, j)$ $(1 \leq i < j \leq R)$ and merge the pair of clusters that yields the maximum $\Delta_{i,j}$ value, as long as it is positive. This scheme aims to find the partition into clusters $X = \bigcup_{r=1}^{R} X_r$, where $R$ is the number of clusters, that maximizes the modified BIC score:

$$\mathcal{MBIC} \stackrel{\mathrm{def}}{=} \sum_{r=1}^{R} \mathcal{BIC}_r \tag{29}$$

$$= \sum_{r=1}^{R} \left[ \log p\left( X_r \middle| M_r \right) - \frac{\lambda}{2} K_r \cdot \log N_r \right] \tag{30}$$

Let us now consider the problem at hand: merging small speech segments using two possible modalities, namely location cues (azimuth direction estimates) and acoustic cues (MFCC vectors). In the case of location cues, a global model is available [2]. It has a fixed number of parameters $K_{\mathrm{loc}} = \mathrm{const}$ (means and standard deviations of the Gaussians), so that only the likelihood term is relevant:

$$\mathcal{BIC}_{\mathrm{loc}} = \log p\left( X_{\mathrm{loc}} \middle| M_{\mathrm{loc}} \right) + \mathrm{const} \tag{31}$$

As a practical implementation, we define one location observation per pair of speech segments given by the short-term clustering [2] and SNS classification (Section 3). This location observation is equal to the difference between the average azimuths of the two speech segments. The total number of such location observations is:

$$N_{\mathrm{loc}} \stackrel{\mathrm{def}}{=} \frac{R_0 \left( R_0 - 1 \right)}{2} \tag{32}$$

where $R_0$ is the total number of speech segments. When a pair of speech segments is separated by a small time duration (e.g. 5 seconds), we model the azimuth difference using a bi-Gaussian model, similarly to [2]. Otherwise, we assume a uniform distribution of azimuth differences. The optimization process amounts to select an optimal coherent graph that links any two speech segments (vertices) with a relationship “same” or “different” (edges), similarly to [2]. Given that the number of parameters is constant, maximizing $\mathcal{BIC}_{\mathrm{loc}}$ is strictly equivalent to the maximum likelihood scheme described in [2].

In the case of acoustic cues (MFCC vectors), a global model is not available so we define $\mathcal{MBIC}_{\mathrm{ac}}$ as in Eq. 30. We start with one cluster per speech segment: $R = R_0$ and merge them iteratively until $\mathcal{MBIC}_{\mathrm{ac}}$ has reached a maximum. The total number of acoustic observations is:

$$N_{\mathrm{ac}} \stackrel{\mathrm{def}}{=} \sum_{r=1}^{R} N_r \tag{33}$$

Note that since there can be a large number $R_0$ of initial clusters, for each cluster we use a single Gaussian model with a full covariance matrix (instead of a GMM). Indeed, the merge of two single Gaussians into one single Gaussian can be done analytically (as opposed to the somewhat time-consuming EM fitting [21] of a GMM). The analytical formulas are given in Annex B.

In practice, location cues are very useful to separate fast changing speaker turns, but carry no reliable long-term information (speakers can move while silent). On the other hand, acoustic cues such as MFCC do carry usable long-term speaker identity cues, but GMM modelling requires a minimum amount of speech for each cluster (at least 2 or 3 seconds). In spontaneous multi-party speech, many speech segments are short so it is not always possible to build reliable acoustic models. From these considerations, it would be desirable to combine the strengths of both location and acoustic modalities, in order to build a reliable speaker clustering system. Namely, the location cues could help to prevent merging two speech segments that are very close in time but very far in space, as observed in fast-changing speaker turns.

An existing combination approach [32], that was successful on a limited amount of data (6 meetings from the M4 Corpus [13]), was tried on a larger amount of data (12 meetings from the same corpus). Unfortunately, it failed producing meaningful speaker clusters, most likely because of the uniform initialization of the clusters (many speakers would fall into the same cluster). Alternatively, we propose here to initialize the clusters using the speech segments provided by Section 3: $R = R_0$. The proposed optimization criterion is:

$$\frac{\mathcal{BIC}_{\text{loc}}}{N_{\text{loc}}} + \frac{\mathcal{MBIC}_{\text{ac}}}{N_{\text{ac}}} \tag{34}$$

where the normalization by the total number of observations (respectively $N_{\text{loc}}$ and $N_{\text{ac}}$) aims at factoring out the number of terms in the log likelihood term of each BIC score. The idea is to add values that are somewhat comparable. The same iterative merging algorithm can be used, as in the original acoustic-only clustering algorithm. Note that the "merged" BIC criterion proposed in Eq. 34 can potentially be used for any number or types of modalities.

## 4.2 Experimental Results

Speaker clustering experiments were conducted on the M4 Corpus of meetings [13]. We used the 21 short meetings – about 5 minutes each – with ground-truth speech/silence segmentation for each speaker, with 4 participants in each meeting. For a given method, speaker clustering was applied on each meeting separately, then the overall performance metrics were evaluated, as in [14]. From the 21 meetings, 3 were used for tuning parameters for post-processing, and 18 were used for performance evaluation. In most cases "tuning parameters" only meant dropping short silences (e.g. less than 0.25 second). We tuned this duration parameter in order to achieve equal MISS and FA [14]. MISS is the percentage of time in the ground-truth that was labeled as speech in the ground-truth, and silence in the result. FA is the percentage of time in the ground-truth that was labeled as silence in the ground-truth, and speech in the result.

For each tested method, the Diarization Error Rate (DER) [14] was computed over the 18 meetings. DER = MISS + FA + SPKR, where SPKR is the percentage of time in the ground-truth that was labeled to different speakers in the ground-truth and the result. DER is expressed as a percentage. Finally, a Speaker Activity Detection (SAD) metric [14] is calculated, which is defined similarly to DER, but using only two classes (speech and silence), instead of speaker labels. Before calculating the DER and SAD, short silences are removed from both ground-truth and result. These short silences are defined as less than 0.3 second (as in the NIST specification [14]) or less than 2.0 seconds, thus defining two subtasks called "0.3 task" and "2.0 task", respectively.

Five clustering methods use distant microphones only. Except for "GMM/HMM", all the others start with $R = R_0$ clusters, one per speech segment given by the multispeaker segmentation scheme depicted in Fig. 7, right side (FAST + SNSGMM). Note that we used 24-dimensional MFCC vectors as acoustic cues.

**"GMM/HMM":** Acoustic cues only, using single channel MFCCs from one microphone in the array – not to be confused with Sector-Based MFCCs. The speaker clustering algorithm is the one described in [15]. For each meeting, we started the merging process with $R = 10$ clusters.

**"ac only":** Acoustic cues only, using single channel MFCCs from one microphone in the array – not to be confused with Sector-Based MFCCs. The speaker clustering algorithm uses BIC with a tunable $\lambda$, as described in Section 4.1.

**"ac + loc":** Acoustic + location (Eq. 34), also using single channel MFCCs.

**"ds, ac + loc":** Acoustic + location (Eq. 34), using MFCCs extracted from the delay-sum beamformed signal. The beamforming direction is the azimuth of the current active speaker.

**"loc. only":** Location only (Eq. 31), MFCCs are not used. "cheating" means that the speaker id of each speech segment is given by prior knowledge of speakers locations.

| Microphones | Method | 0.3 task | | | 2.0 task | | |
|---|---|---|---|---|---|---|---|
| | | DER | | SAD | DER | | SAD |
| Distant ( mic. array ) | GMM/HMM | 37.6 | (63.9) | 14.4 | 32.2 | (68.6) | 4.6 |
| | ac only | 51.7 | (77.5) | 11.4 | 48.9 | (76.7) | 4.4 |
| | ds, ac + loc | 27.8 | (40.4) | 11.1 | 24.8 | (35.6) | 4.3 |
| | ac + loc | 18.1 | (37.4) | 9.7 | 15.2 | (30.0) | 4.0 |
| | loc. only (cheating) | 12.5 | (19.7) | 8.2 | 10.4 | (17.9) | 4.1 |
| Close-talking | lapelmix-GMM/HMM | 29.7 | (66.0) | 14.4 | 24.0 | (65.1) | 4.5 |
| | lapels-seg (cheating) | 8.2 | (34.9) | 4.7 | 9.1 | (27.2) | 2.1 |

Table 8: Speaker clustering results on 18 meetings of the M4 Corpus [13] (the lower, the better). Brackets indicate results on overlapped speech only. "ds" stands for delay-sum beamforming. "GMM/HMM" is the speaker clustering algorithm described in [15].

For comparison, two other results are reported, that use close-talking microphones only (lapels):

**"lapels-seg":** The multichannel segmentation baseline described in [2]. The word "cheating" reminds that the number of speakers and the speaker ids are *de facto* known when lapels are used (one lapel per person).

**"lapelmix-GMM/HMM":** The single channel iterative speaker clustering scheme described in [15]. For each meeting, we started the merging process with 10 clusters. For each meeting, the single channel is formed by first adding the four lapel signals.

From the results reported in Tab. 8, several observations can be made. First, the methods rank in similar orders, whether the task is the "0.3 task" (precise segmentation) or the "2.0 task" (rough segmentation). Second, for all metrics, the proposed combination "ac + loc" significantly improves over both "GMM/HMM" and "ac only" results. The "ac + loc" results are in fact close to the "cheating" results, i.e. the best that can be obtained based on the underlying multispeaker segmentation method. Finally, delay-sum beamforming does not seem to help, as already noted in [16].

By comparing "cheating" with "lapels-seg", one can see that the microphone array-based methods can potentially achieve results that amount to a slight degradation in overall DER (3 or 4 % absolute), for a large improvement on overlapped speech (10 to 15 % absolute). So they can indeed be used for speaker clustering in meetings. The single channel speaker clustering scheme "lapelmix-cluster" does not seem to yield effective results. In fact, compensation heuristics such as the "purification" step proposed in [16] are necessary for this scheme to be effective on spontaneous multi-party speech.

## 4.3 Future Directions

Overall, the results presented above show that the proposed "ac + loc" combination brings significant improvement over "ac only". This validates the proposed criterion (Eq. 34), in so far as it combines complementary strengths from two modalities. A closer look at the results may suggest future directions of research. We tried "ac + loc" on a concatenation of three meetings, in which some people appear at different locations: different seats around the microphone array, and some standing locations when doing a presentation (further from the array). The same person seated at different locations tends to be clustered correctly (1 cluster). On the contrary, people who stand up and move away to do a presentation systematically end up being clustered into 2 different clusters, depending on the distance (close while seated, far while standing). We tried low-level signal processing methods that have proved useful to improve MFCC-based Automatic Speech Recognition (ASR) results, hoping that they would also improve speaker clustering results. Unfortunately, neither delay-sum beamforming nor dereverberation [17] produced any improvement at all, with respect to this issue.

To conclude, it seems that feature variability between close and distant locations is a bottleneck to speaker clustering with distant microphones, which suggests some basic research. A previous study [33] already showed that the usual linear transmission model does not always hold: the coherence between two microphones placed at different directions from the mouth, but equal distance, is sometimes very low. In the present case, it seems that a similar study is needed, varying distance between microphone and mouth, within an indoor environment. It could shed some light on limits of the linear model for reverberant environments, and possible alternatives. A speaker recognition study proposed a location-dependent Cepstral Mean Normalization [34], that aims at removing transmission channel distortion, without removing speaker-specific characteristics. However, it requires training data with multiple speakers at multiple locations for each different room, which could limit the ease of use for end-users.

On the practical side, it could be interesting to modify existing, larger systems such as [16], by integrating the use of the joint criterion (Eq. 34).

# 5 Automatic AV Calibration in Discrete Space

The previous section gave some insights on successes and limits of audio-only speaker clustering, concerning spontaneous multi-party speech within an indoor environment. It could be beneficial to compensate shortcomings of the audio-only approach by using an additional modality: visual information from cameras. For example, in the audio-visual speaker tracking framework [20], it is possible to infer speech segments' boundaries while keeping track of speaker identities. The advantage of the visual modality is that identity cues such as faces are visible at almost all times, whether the person is speaking or not. Thus, this information could be highly useful to merge two speech segments from the same speaker that exhibit a high variability in audio cues, due a change in location (in particular, a change of distance relative to the microphone array, as mentioned in the previous section).

In order to fuse audio and video cues, some spatial information is needed that relates the locations of the microphones to the locations of the cameras: this spatial information forms the audio-visual calibration. For example, a short sequence can be recorded with a single speaker moving around the room, while speaking. One practical issue of schemes such as [20] is the need for manual initialization of the visual tracking, in the audio-visual calibration procedure. In this section, we propose an alternative scheme that does not require manual initialization.

Both audio and video spaces are discretized, as depicted in Fig. 8. The audio space, e.g. azimuth direction from a microphone array, is discretized into sectors, as in the SAM-SPARSE-MEAN approach [1, 5]. For example, 18 sectors, each spanning 20 degrees, cover the entire 360-degree space around a circular microphone array (Fig. 8, right side). For each camera, the visual space is discretized into blocks, e.g. 24 blocks vertically and 32 blocks horizontally (Fig. 8, left side). Having such a discretization of both audio and video spaces allows for exploring the correlation between *all* locations from each modality. For example, by finding peaks of covariance, we could determine that having a speaker in a given audio sector is highly correlated with having a speaker in a given video block of a camera. This type of information can in turn be used for initialization of more complex, model-based parametric audio-visual calibration schemes, that could be an extension of [35, 36].

As a proof of concept, we used a short training sequence `seq11` from the AV16.3 Corpus [10], that was recorded with an 8-microphone array in the middle of the room, and 3 cameras on the walls around the room (Fig. 9, top row). The "audio activity indicator" is defined as the posterior probability of having speech activity within a given sector and a given time frame. The probabilistic estimate (a value between 0 and 1) is derived from the SAM-SPARSE-MEAN metric [1], as described in [26]. The "video motion indicator" is defined as the average across all pixels in a given block, of the posterior probability of having motion in all three components R, G, B. For each component R, G, or B, this probabilistic estimate is derived by fitting a model with two components (moving or not moving) on a video motion feature (resp. $\mathrm{vmf_R}, \mathrm{vmf_G}, \mathrm{vmf_B}$), computed using 3 consecutive frames $t-1,\ t,\ t+1$:

$$\mathrm{vmf_R}\left(x,y,t\right) \stackrel{\mathrm{def}}{=} \sqrt{\left|\mathrm{c_R}\left(x,y,t+1\right)-\mathrm{c_R}\left(x,y,t\right)\right|\cdot\left|\mathrm{c_R}\left(x,y,t\right)-\mathrm{c_R}\left(x,y,t-1\right)\right|} \quad (35)$$

where $c_R(x,y,t)$ is the color value (e.g. from 0 to 255) for color R, pixel $(x,y)$ and time frame $t$. $\text{vmf}_G$ and $\text{vmf}_B$ are defined similarly. On each video motion feature $\text{vmf}_R, \text{vmf}_G, \text{vmf}_B$, a probabilistic model is fitted using EM [21]. It is a 2-component model, comprising a Gamma pdf and a Shifted Rice pdf, as in Annex B of [26]. Based on a simplifying independence assumption between R, G, and B, the posterior probability of motion at pixel $(x,y)$ and time frame $t$ is estimated as:

$$p\left(\text{motion}|\ \text{pixel}\ x,y,t\right) = p\left(\text{motion}|\ \text{vmf}_R\left(x,y,t\right)\right) \cdot p\left(\text{motion}|\ \text{vmf}_G\left(x,y,t\right)\right) \cdot p\left(\text{motion}|\ \text{vmf}_B\left(x,y,t\right)\right)$$

For each block of pixels, the "video motion indicator" is a value between 0 and 1, obtained by averaging $p\left(\text{motion}|\ \text{pixel}\ x,y,t\right)$ across all pixels $(x,y,t)$ in the block, at time frame $t$.

**Covariance analysis for calibration:** we propose to compute the covariance over time between each "audio activity indicator" and each "video motion indicator", as illustrated by Fig. 8. This approach is justified as long as the short calibration recording contains a single moving speaker, in a fixed indoor environment (as in `seq11`): there is no data association ambiguity between motion in a region of the video image and speech activity in a sector of space. Please report to Annex C for implementation details. One possible use of the estimated covariance is to determine, for each block of pixels of a camera, which audio sector corresponds the most. The result of this analysis is depicted in Fig. 9 (the same type of covariance analysis could be conducted between cameras). This audio-visual information can serve as an initialization to more complex, model-based audio-visual calibration techniques, that would for example extend video-only calibration [35, 36].

# 6 Conclusion

This paper investigated applications of microphone arrays to speaker detection, localization and clustering in an indoor environment. An effective approach for multisource joint detection-localization was proposed, that uses a fast sector-based predetection first step, followed by a Scaled Conjugate Gradient descent of the SRP-PHAT measure. Convergence of the location estimate typically happens between 5 and 10 iterations. Tests on real data show that up to 3 simultaneous speakers can be correctly detected and located, and that the approach is fit for real-time applications. In a second part, an effective, unsupervised approach for speech/non-speech discrimination was proposed, that builds on the multisource detection-localization, to determine speech segments while effectively eliminating most machine noises (e.g. beamer), along with some human body noises, as can be found in multi-party speech (meetings). This result is quite interesting, given the fact that we do not use any training data. Third, we examined the speaker clustering task using distant microphones only, where the speech segments that were detected and located, are clustered to form "speakers", without training data. A generalization of the BIC criterion to multiple modalities was proposed, and applied to the merging of long-term acoustic information (MFCCs) with short-term location information (speaker direction). Speaker clustering using the merged BIC criterion yielded a major improvement over acoustic-alone speaker clustering, including a state-of-the-art approach. Experiments on a corpus of meetings recorded with distant microphones show that the speaker clustering performance is close to the optimum that could be obtained with the underlying multispeaker detection-localization. Results also compare well with those of a speech segmentation technique using 1 close-talking microphone per person. An important signal processing issue seems to be the signal distortion variability that appears when the distance is changing, between a speaker and the microphone array. It systematically leads to two different clusters for the same person (close or far from the microphone array). Although a solution exists for distant speaker recognition, where training data is available, it is at present not clear how this can be done within the framework of unsupervised speaker clustering. Based on this analysis, we suggested to add another modality: visual information from cameras, which requires a calibration between microphones and cameras. An initial investigation on automatic audio-visual calibration of a microphone array and several cameras shows that a simple covariance analysis, based on systematic discretization of the space of each modality, can lead to meaningful informations in order to initialize more complex, model-based calibration procedures.

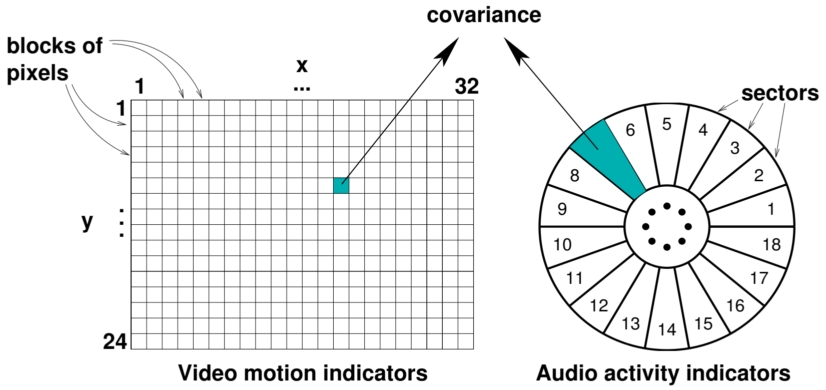



Figure 8: Audio-visual calibration: both audio and video spaces are discretized. Covariance is calculated between each video motion indicator (block of pixels) and each audio activity indicator (sector of space around the microphone array). Numbers represent indices of video blocks and audio sectors.

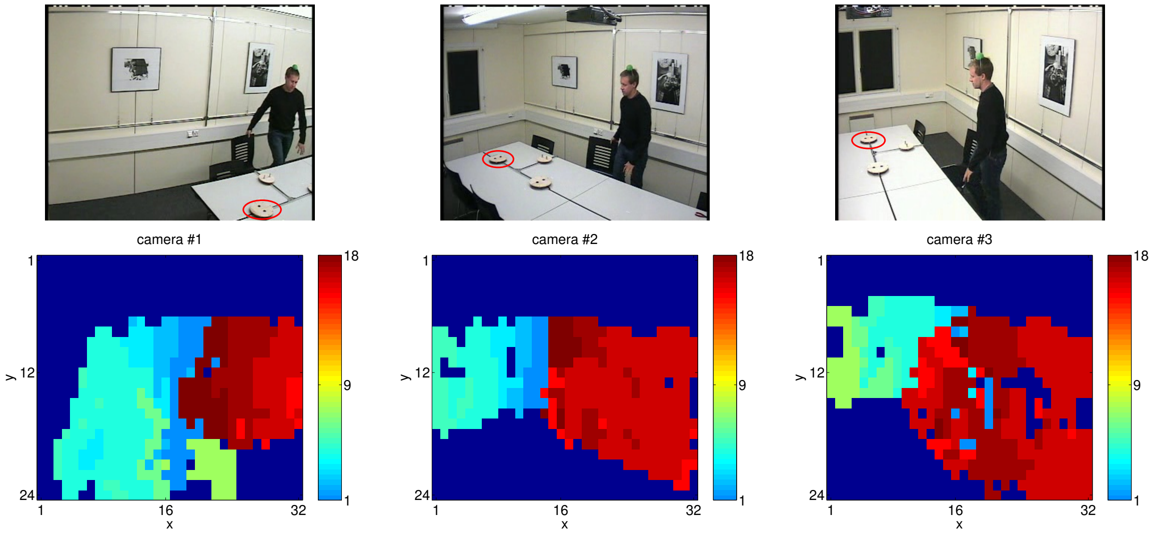



Figure 9: Result of the covariance analysis between audio activity and video motion. The image plane of each camera is depicted. Top row: snapshot of the `seq11` recording (microphone array indicated by a red ellipse). Bottom row: result of the covariance analysis. For each block of pixels, the index (1 to 18) of the audio sector with highest covariance is represented by a color (colorbar). The background color (dark blue) indicates that all audio sectors have a covariance inferior to $e^{-6}$.

# Annex A: Comparison of Detection Features for Localization

This annex presents an experimental comparison of four different features for speech detection. This task differs from the usual speech/silence classification task, because the purpose is to determine, for each time frame, whether an active speaker can be *correctly localized* or not. The evaluation is conducted using two different azimuth localization methods, GCC-PHAT [11] and SRP-PHAT [8]. Ideally, a "good" feature for localization-oriented speech detection would exhibit smaller localization errors when the detection threshold is more conservative. In the following, we first describe the four detection features and their usage, then the two localization methods GCC-PHAT and SRP-PHAT. The single source recording `seq01` from the AV 16.3 corpus [10] was used for evaluation, using only the first circular 8-microphone array. In all cases, we are not only evaluating each detection feature, but rather its complete integration within an automatic threshold selection system, similarly to [6]. The goal is to determine whether a feature that allows for good classification of *all* frames (speech and silence) also allows for good localization precision (azimuth error on speech frames only). The selected range of threshold values corresponds to target False Alarm Rates from 1e-14 % to 99.0%. As explained in [6], no training data is needed for this automatic threshold selection strategy.

**SNR estimate for detection:** The multimicrophone SNR estimate presented in [37] was implemented to evaluate the instantaneous SNR within each time frame (a non-negative value). A probabilistic model is fitted in an unsupervised manner on *all* SNR estimate values, that models silence with one component (mixture of a Dirac pdf and a Rice pdf [38] with parameters $\sigma_0$ and $V_0$) and speech with another component (shifted Rice pdf, where the shift is $\sqrt{V_0^2 + 2 \cdot \sigma_0^2}$, similarly to Annex B of [26]). The posterior probability of having an active frame is then estimated for each time frame, using the fitted model. Based on all estimated posteriors, a threshold on the posterior probability is then selected, corresponding to a given target False Alarm Rate value, as in [6]. Finally, each time frame is classified as "silence" or "speech" by comparing the posterior probability of activity with the threshold. See [6, 26] for theoretical and implementation details.

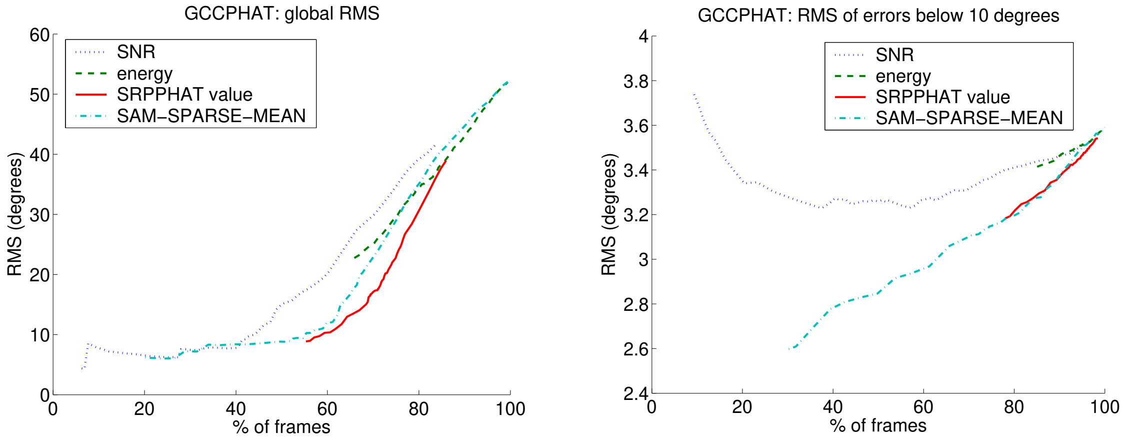
**Energy for detection:** Within each time frame, the instantaneous energy is averaged across microphones (a non-negative value). A bi-Gaussian model is fitted in log energy domain (one Gaussian for silence, one Gaussian for speech), and the same procedure is conducted as for the SNR estimate, using the fitted bi-Gaussian model and a given target False Alarm Rate value.

**SRP-PHAT value for detection:** Within each time frame, the multimicrophone location-dependent SRP-PHAT metric defined in [8] is maximized, by searching through locations in space. If negative, the obtained maximum value is replaced with zero, thus yielding a value between 0 and 1. A Dirac+Rice+Shifted Rice model is fitted, as for the SNR estimate. The same procedure is conducted as for the SNR estimate, using the fitted model and a given target False Alarm Rate value.

**SAM-SPARSE-MEAN for detection:** The physical space is discretized into a small number of sectors (e.g. 18 sectors, each sector spans 20 degree of azimuth range). As in [1, 5], the SAM-SPARSE-MEAN value is calculated for each time frame *and* each sector. An adequate probabilistic model is fitted on *all* the obtained values [6, 26], and for each time frame, the posterior probability of having an active time frame is estimated, as in [26]. The posterior probability is used as a detection feature (value between 0 and 1). A threshold is selected automatically, as for the SNR estimate.

**Localization methods:** In order to evaluate the four detection features, we test them as a prior detection step for two different localization methods: GCC-PHAT [11] and SRP-PHAT [8]. GCC-PHAT is implemented using the two squares of microphones defined by the array (microphones 1, 3, 5, 7 and microphones 2, 4, 6, 8). For each square, the azimuth direction is determined from pair-wise GCC-PHAT time-delay estimates [11], by solving equations analytically, as in [12]. The two direction estimates are then recombined by averaging estimated azimuths and elevations. Whenever none of the two squares produces solvable equations, or when they produce two resulting direction estimates differing by more than 90 degrees, the time frame is dropped. This happened in 13.4 % of the time frames.

The SRP-PHAT method [8] uses all 8 microphones to find the location in space that maximizes




(a) (b)

Figure 10: Variation of the RMS azimuth error (in degrees) when the detection threshold is varied. "% of frames" is the proportion of active frames that are above a given value of the detection threshold. The GCC-PHAT azimuth localization method is used [11]. In (b), only frames with a localization error below 10 degrees are considered.

the SRP-PHAT metric. It is always solvable, thus no frame is dropped.

Results are reported in Figs. 10 and 11. "RMS" stands for Root Mean Square azimuth error. From Figs. 10a and 11a, it appears that energy appears *not* to be an adequate measure for speech detection in the localization context, which confirms the study conducted in [22]. The other three measures exhibit decent behaviour: in both GCC-PHAT and SRP-PHAT cases, increasing the detection threshold permits to reach a precision below 10 degree in terms of RMS error.

A more detailed analysis is presented in Figs. 10b and 11b, where only results below 10 degree RMS error are considered. It appears clearly that only SAM-SPARSE-MEAN allows to reduce the RMS error, as the detection threshold is increased. A possible reason for this success is that a high SAM-SPARSE-MEAN value corresponds to a large bandwidth occupied by the speech source. This in turn directly impact on the precision of both GCC-PHAT and SRP-PHAT location estimates.

# Annex B: Some Analytical Formulas for Single Gaussians

All formulas are valid in any dimensionality $D$. Assume that $N$ data samples in $\mathbb{R}^D$ are modeled with a single Gaussian of mean $\mu$ and covariance matrix $\Sigma$. For the covariance matrix $\Sigma$, we use the best unbiased estimate (normalization by $N-1$). The log likelihood of the $N$ data samples, given the single Gaussian model $\mathcal{N}(\mu, \Sigma)$, can be computed *without the data*, by using the analytical formula:

$$\log p\left(X \mid \mathcal{N}(\mu, \Sigma)\right) = -\frac{ND}{2}\log 2\pi - \frac{N}{2}\log|\Sigma| - \frac{(N-1)\,D}{2} \tag{36}$$

Similarly, assume two data sets of $N_1$ and $N_2$ samples, each modelled with a single Gaussian of parameters $(\mu_1, \Sigma_1)$ and $(\mu_2, \Sigma_2)$, respectively. The merge of the two data sets can be modelled with a single Gaussian, which parameters can be calculated *without the data*, using the analytical formulas:

$$\mu = \frac{N_1 \cdot \mu_1 + N_2 \cdot \mu_2}{N_1 + N_2} \tag{37}$$

$$\Sigma = \frac{(N_1 - 1)\,\Sigma_1 + N_1\mu_1\mu_1^{\mathrm{T}} + (N_2 - 1)\,\Sigma_2 + \mu_2\mu_2^{\mathrm{T}} - (N_1 + N_2)\,\mu\mu^T}{N_1 + N_2 - 1} \tag{38}$$

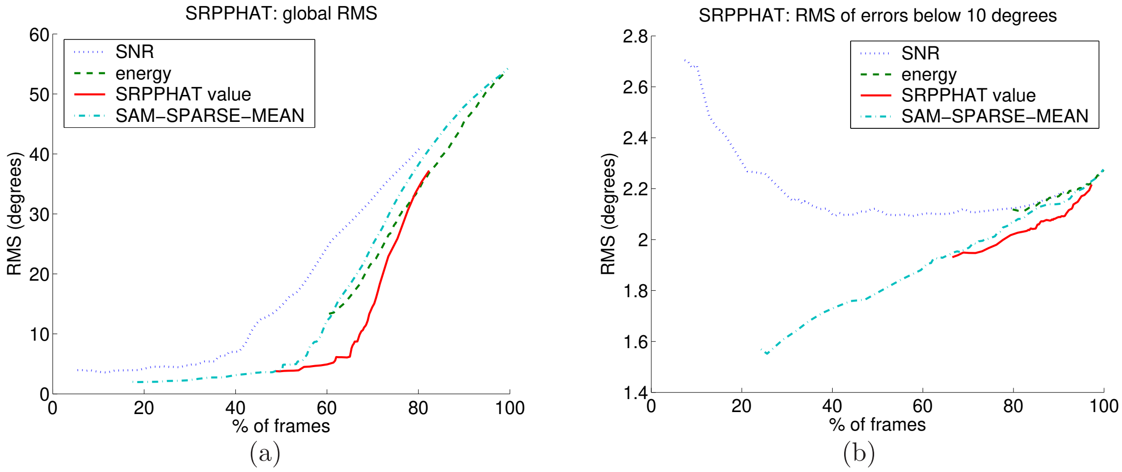



Figure 11: Variation of the RMS azimuth error (in degrees) when the detection threshold is varied. "% of frames" is the proportion of active frames that are above a given value of the detection threshold. The SRP-PHAT azimuth localization method is used [8]. In (b), only frames with a localization error below 10 degrees are considered.

# Annex C: Implementation Details for AV Calibration

This section contains a few notes concerning Section 5, about the estimation of the covariance between audio activity indicators and video motion indicators. Two issues were addressed, using the frame-level and sector-level audio activity posteriors (report to Annex C of [26] for their exact definition).

- First, audio and video modalities usually have different frame rates. Frame-level and sector-level posterior estimates of audio activity [26] were downsampled using the `max` and `mean` operators, respectively.

- Second, the covariance needs to be calculated "on speech only". We implemented this by weighting each video frame with the downsampled frame-level posterior estimate of audio activity.

# Acknowledgments

The author acknowledges the support of the European Union through the AMI and HOARSE projects. This work was also carried out in the framework of the Swiss National Center of Competence in Research (NCCR) on Interactive Multi-modal Information Management (IM)2. The author would like to thank Dr Fabio Valente for help with the speaker clustering performance evaluation.

# References

[1] G. Lathoud and M. Magimai.-Doss, "A Sector-Based, Frequency-Domain Approach to Detection and Localization of Multiple Speakers," in *Proc. of ICASSP*, 2005.

[2] G. Lathoud, I. McCowan, and J. Odobez, "Unsupervised location-based segmentation of multi-party speech," in *Proc. the 2004 NIST ICASSP Meeting Recognition Workshop*, 2004.

[3] G. Schwartz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[4] S. Chen and P. Gopalkrishnan, “Speaker, environment and channel change detection and clustering via the bayesian information criterion,” *IBM Technical Journal*, 1998.

[5] G. Lathoud, J. Bourgeois, and J. Freudenberger, “Sector-based detection for hands-free speech enhancement in cars,” *EURASIP Journal on Applied Signal Processing, Special Issue on Advances in Multimicrophone Speech Processing*, 2006.

[6] G. Lathoud, M. Magimai.-Doss, and H. Bourlard, “Threshold selection for unsupervised detection, with an application to microphone arrays,” in *Proc. ICASSP*, 2006.

[7] J. DiBiase, H. Silverman, and M. Brandstein, “Robust localization in reverberant rooms,” in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Springer, 2001, ch. 8, pp. 157–180.

[8] J. DiBiase, “A high-accuracy, low-latency technique for talker localization in reverberant environments,” Ph.D. dissertation, Brown University, Providence RI, USA, 2000.

[9] M. Moller, “A scaled conjugate gradient algorithm for fast supervised learning,” *Neural Networks*, vol. 6, pp. 525–533, 1993.

[10] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, “AV16.3: an Audio-Visual Corpus for Speaker Localization and Tracking,” in *Proc. the 2004 MLMI Workshop, S. Bengio and H. Bourlard Eds, Springer Verlag*, 2005.

[11] C. Knapp and G. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. ASSP-24, no. 4, pp. 320–327, August 1976.

[12] M. Brandstein, “A framework for speech source localization using sensor arrays,” Ph.D. dissertation, Brown University, 1995.

[13] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, “Automatic analysis of multimodal group actions in meetings,” *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 27, no. 3, pp. 305–317, March 2005.

[14] “The Rich Transcription Spring 2003 (RT-03S) Evaluation Plan,” Tech. Rep., 2003.

[15] J. Ajmera and C. Wooters, “A robust speaker clustering algorithm,” in *Proc. of the IEEE ASRU Workshop*, 2003.

[16] X. Anguera, C. Wooters, B. Peskin, and M. Aguilo, “Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system,” in *Proc. NIST RT05s*, 2005.

[17] D. Gelbart and N. Morgan, “Evaluating long-term spectral subtraction for reverberant asr,” in *Proc. the IEEE ASRU Workshop*, 2001.

[18] G. Lathoud, M. Magimai.-Doss, B. Mesot, and H. Bourlard, “Unsupervised Spectral Subtraction for Noise-Robust ASR,” in *Proc. the IEEE ASRU Workshop*, December 2005.

[19] M. Omologo, M. Matassoni, and P. Svaizer, “Speech recognition with microphone arrays,” in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Springer, 2001, ch. 15, pp. 331–353.

[20] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, “Audio-visual probabilistic tracking of multiple speakers in meetings,” *IEEE Trans. on Speech and Audio Process., accepted for publication with minor revisions*, December 2005.

[21] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society, Series B*, vol. 39, pp. 1–38, 1977.

[22] G. Lathoud and I. McCowan, “A Sector-Based Approach for Localization of Multiple Speakers with Microphone Arrays,” in *Proc. of the SAPA Workshop*, Oct. 2004.

[23] S. Bengio, J. Mariéthoz, and M. Keller, "The expected performance curve," in *Proceedings of the ICML 2005 workshop on ROC Analysis in Machine Learning*, Bonn, Germany, 2005.

[24] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *IEEE Signal Processing Magazine*, vol. 13, pp. 67 – 94, July 1996.

[25] G. Lathoud, M. Magimai.-Doss, and H. Bourlard, "Channel normalization for unsupervised spectral subtraction," IDIAP-RR 06-09, 2006.

[26] ——, "Threshold Selection for Unsupervised Detection, with an Application to Microphone Arrays," IDIAP-RR 05-52, 2005.

[27] W. Chu and A. Warnock, "Detailed directivity of sound fields around human talkers," IRC-CNRC, Canada, Tech. Rep. IRC-RR-104, December 2002.

[28] Y. Grenier, "Wideband source location through frequency-dependent modeling," *IEEE Trans. on Signal Processing*, vol. 42, no. 5, May 1994.

[29] J. Nix and V. Hohmann, "Sound source localization in real sound fields based on empirical statistics of interaural parameters," *J. Acoust. Soc. Am.*, no. 1, pp. 463–479, 2006.

[30] L. Lu and H. J. Zhang, "Content analysis for audio classification and segmentation," *IEEE Trans. on Speech and Audio Process.*, vol. 10, no. 7, 2002.

[31] E. Shriberg, A. Stolcke, and D. Baron, "Can prosody aid the automatic processing of multi-party meetings? Evidence from predicting punctuation and disfluencies, and overlapping speech," in *Proc. the ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding (Prosody-2001)*, 2001.

[32] J. Ajmera, G. Lathoud, and I. McCowan, "Segmenting and clustering speakers and their locations in meetings," in *Proceedings the 2004 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2004.

[33] I. Schwetz, G. Gruhler, and K. Obermayer, "Correlation and stationarity of speech radiation," *IEEE Trans. Speech Audio processing*, vol. 12, no. 5, September 2004.

[34] L. Wang, N. Kitaoka, and S. Nakagawa, "Robust distant speaker recognition based on position dependent cepstral mean normalization," in *Proc. Interspeech*, 2005.

[35] T. Svoboda, "Multi-Camera Self-Calibration," August 2003. [Online]. Available: http://cmp.felk.cvut.cz/ svoboda/SelfCal/index.html

[36] J. Y. Bouguet, "Camera Calibration Toolbox for Matlab," January 2004. [Online]. Available: http://www.vision.caltech.edu/bouguetj/calib_doc/

[37] J. Chen and W. Ser, "Speech detection using microphone array," *Electronic Letters*, vol. 36, no. 2, January 2000.

[38] S. Rice, "Mathematical analysis of random noise," in *Selected Papers on Noise and Stochastic Processes*, N. Wax, Ed., Dover, New York, 1954, pp. 133–254.