IDIAP RESEARCH REPORT

A Sector-Based Approach for Localization of Multiple Speakers with Microphone Arrays

Guillaume Lathoud [a,b]   Iain A. McCowan [a]

IDIAP–RR 04-15

May 2004

submitted for publication

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

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

[a] IDIAP Research Institute, CH-1920 Martigny, Switzerland
[b] Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland

IDIAP Research Report 04-15

A Sector-Based Approach for Localization of Multiple Speakers with Microphone Arrays

Guillaume Lathoud Iain A. McCowan

May 2004

submitted for publication

Abstract. Microphone arrays are useful in meeting rooms, where speech needs to be acquired and segmented. For example, automatic speech segmentation allows enhanced browsing experience, and facilitates automatic analysis of large amounts of data. Spontaneous multi-party speech includes many overlaps between speakers; moreover other audio sources such as laptops and projectors can be active. For these reasons, locating multiple wideband sources in a reasonable amount of time is highly desirable. In existing multisource localization approaches, search initialization is very often an issue left open. We propose here a methodology for estimating speech activity in a given sector of the space rather than at a particular point. In experiments on more than one hour of speech from real meeting room multisource recordings, we show that the sector-based greatly reduces the search space. At the same time, it achieves effective localization of multiple concurrent speakers.

1 Introduction

Microphone arrays are useful to find the points of origin of multiple incoming acoustic signals. In this paper we focus on human speech, which is a wideband signal. In spontaneous multi-party speech, overlaps occur often [1], and indoor environments are usually highly reverberant. Thus, there is a need to localize multiple concurrent sources. We chose to use Uniform Circular Arrays (UCAs) because in the horizontal plane, their characteristics are almost invariant with direction [2], therefore imposing no constraint on the location of the source.

Existing approaches for source localization can be divided in two groups: parametric and non-parametric. Parametric approaches [3], also known as beamforming or maximum likelihood approaches, define a spatial likelihood function for each point of the space. Such a function can have multiple local maxima. Searching the entire space for all local maxima of this function is an expensive process.

Non-parametric approaches [4], also known as signal subspace, high-resolution or eigenanalysis methods, do not rely on such a function. Examples are the well-known MUSIC [5] and ESPRIT [6] algorithms, which typically achieve higher resolution than parametric methods. However, these methods were originally designed for narrowband signals and Uniform Linear Arrays (ULAs). Previous work extended non-parametric approaches from ULAs to UCAs [7], from narrowband to wideband signals [8], and both [9]. Only the latter [9] is relevant to our problem. Globally, coherent signals such as speech and its reverberations still seem to be a problem with these methods, since reverberations have to be modeled explicitly. Also, steering matrices have to be defined for each sector of the space. Finding which sector(s) of the space contain active acoustic source(s) is an open issue.

From this review we can see that finding the active sector(s) is an issue for both parametric and non-parametric approaches, as already mentioned in [2]. There is a need for a method that allows detection of acoustic waves from a sector of the space, rather than from a specific point or from a specific direction. Achieving sector-based source localization with a low computational cost would allow fast localization of the sound sources, by quickly restricting the search space to a small number of sectors.

One successful work in this direction is [10]. It is a multi-level approach that relies on prior knowledge of room metrics, among other things. On the contrary, this paper explicitly defines a generic Sector Activity Measure, without need for prior knowledge other than the microphone array's geometry. Since high resolution is not needed for sector-based localization, our approach is based on parametric methods. An implementation called SAM-PHAT is proposed and extensively tested on multiple speaker cases, including more than one hour of real meeting room recordings. We show that the proposed sector-based approach greatly reduces the search space for a low computational cost.

Section 2 presents the sector-based approach. Section 3 presents and justifies the experimental protocol. Section 4 gives and discusses the results, and Section 5 concludes.

2 The Sector-Based Approach

Searching the entire space for multiple local maxima of a point-based likelihood function leads to an infinite number of possibilities. Even a discretized space or grid would include a very large number of points, in order to localize sources that could be in any locations.

We therefore propose to transform a given point-based spatial likelihood function (such as SRP-PHAT [11]) into a generic sector-based activity measure. This new measure will allow to detect active sound sources within a volume of the physical space, rather than at a particular point in space. First, the search space is partitioned into a small number of volumes, called "sectors" hereafter. Each sector is then evaluated by a Sector Activity Measure (SAM). The SAM values can be used for localizing active sectors: for a given sector, a higher SAM value indicates a higher likelihood of having at least one active source within the sector. This in turn can be used to reduce the search space of point-based methods.

2.1 Partition of the Search Space into Sectors

A sector is a connected volume $\mathbf{S} \subset \mathbb{R}^3$ of physical space. By "connected volume" we mean that for any two points $\mathbf{x}_1$ and $\mathbf{x}_2$ in volume $\mathbf{S}$, we can define a continuous contour $\mathcal{C}_{\mathbf{x}_1,\mathbf{x}_2} \subset \mathbf{S}$. For example, the space around a horizontal planar microphone array can be partitioned in "vertical slices":

$$\begin{array}{l} for\ i = 1 \dots N_{sectors} : \\ \mathbf{S}_i = \left\{ (r,\theta,\phi) \in \mathbb{R}^3 \ \middle|\ r \geq r_{min},\ \theta_{i-1} \leq \theta < \theta_i,\ 0 \leq \phi \leq \frac{\pi}{2} \right\} \end{array} \tag{1}$$

where r, θ, ϕ designate radius, azimuth and elevation with respect to the microphone array center, $\theta_i = i \frac{2\pi}{N_{sector}}$ and microphones are all in the sphere $r < r_{min}$. More generally, any partition along radius, azimuth and elevation can be defined, depending on the microphone array's geometry.

2.2 Definition of a Sector Activity Measure (SAM)

Section 4 will give evaluation in terms of azimuth θ. However in this Section we use Cartesian coordinates, in order to keep equations simple.

Assuming that a spatial likelihood function $\mathbb{L}(\mathbf{x})$ is available for any point $\mathbf{x}$ in the search space (see [3] for a review of such functions), we simply propose to evaluate sound activity within a given sector $\mathbf{S}$ as:

$$SAM(\mathbf{S}) \triangleq \frac{1}{V(\mathbf{S})} \int\int\int_{\mathbf{S}} \mathbb{L}\left([x\ y\ z]^T \right)\ dx\ dy\ dz \tag{2}$$

where $V(\mathbf{S}) = \int\int\int_{\mathbf{S}} dx\ dy\ dz$ is the volume of sector $\mathbf{S}$, and x, y, z are Cartesian coordinates.

2.3 Definition of SAM-PHAT

We propose to define SAM-PHAT as the Sector Activity Measure that integrates the point-based SRP-PHAT measure [11]. For each location $\mathbf{x}$, SRP-PHAT is defined as:

$$\mathbb{L}_{SRP-PHAT}(\mathbf{x}) \triangleq \frac{1}{P} \sum_{p=1}^{P} R_{PHAT}^{(p)} \left(\mu^{(p)}(\mathbf{x}) \right) \tag{3}$$

where $\mathbf{x} = [x, y, z]^T \in \mathbb{R}^3$ is a point in space expressed in Cartesian coordinates, and P is the number of microphone pairs. For example, with 4 microphones, there are P=6 pairs. $R_{PHAT}^{(p)}(\mu)$ is the time domain GCC-PHAT [12] for microphone pair p. $\mu^{(p)}(\mathbf{x})$ is the vector of theoretical time-delays associated with location $\mathbf{x}$:

$$\mu(\mathbf{x}) \quad \triangleq \quad \left[\mu^{(1)}(\mathbf{x}) \cdots \mu^{(p)}(\mathbf{x}) \cdots \mu^{(P)}(\mathbf{x}) \right]^T \tag{4}$$

$\mu^{(p)}$ is the theoretical time delay (in samples) between the microphones in pair p, given by

$$\mu^{(p)}(\mathbf{x}) \quad \triangleq \quad \frac{\left(||\mathbf{x} - \mathbf{m}_1^{(p)}|| - ||\mathbf{x} - \mathbf{m}_2^{(p)}|| \right) f_s}{c} \tag{5}$$

where $\mathbf{m}_1^{(p)} \in \mathbb{R}^3$ and $\mathbf{m}_2^{(p)} \in \mathbb{R}^3$ are Cartesian coordinates of the microphone locations in pair p, f_s is the sampling frequency in Hz and c is the speed of sound in the air in m/s (usually 342 m/s). We note that $\mu^{(p)}$ are continuous, non-linear functions of $\mathbf{x}$.

From Eqs. (2) and (3), SAM-PHAT develops into:

$$\begin{array}{l} SAM_{PHAT}(\mathbf{S}) = \\ \frac{1}{P} \sum_{p=1}^{P} \frac{1}{V(\mathbf{S})} \int\int\int_{\mathbf{S}} R_{PHAT}^{(p)} \left(\mu^{(p)} \left([x\ y\ z]^T \right) \right)\ dx\ dy\ dz \end{array} \tag{6}$$

Computing each term involves an expensive 3-dimensional integration. A change of variable $\mathbf{y} = \mu^{(p)}(\mathbf{x})$ is difficult, because analytical inversion of the function $\mu^{(p)}(\mathbf{x})$ is not trivial: $\mu^{(p)}(\mathbf{x})$ is not bijective.

In the rest of this paper, we'll assume that each sector S is a connected volume. Since $\mu^{(p)}(\mathbf{x})$ is continuous and $\mathbf{S}$ is a connected volume, $\mathbf{S}$ is projected into a segment:

$$\mu^{(p)}(\mathbf{S}) = \left[\mu^{(p)}_{min}(\mathbf{S}),\, \mu^{(p)}_{max}(\mathbf{S})\right] \tag{7}$$

Lower and upper limits of this segment are respectively minimum and maximum time-delays across all points in sector $\mathbf{S}$, for microphone pair p.

In order to approximate SAM-PHAT with a simpler version, we simply average the time-domain GCC-PHAT function on each segment $\left[\mu^{(p)}_{min}(\mathbf{S}),\, \mu^{(p)}_{max}(\mathbf{S})\right]$. Hence the "simplified SAM-PHAT":

$$\begin{aligned} &\overline{SAM_{PHAT}}\,(\mathbf{S}) \triangleq \\ &\tfrac{1}{P}\textstyle\sum_{p=1}^{P} \tfrac{1}{\Delta\mu^{(p)}(\mathbf{S})} \int_{\mu^{(p)}_{min}(\mathbf{S})}^{\mu^{(p)}_{max}(\mathbf{S})} R^{(p)}_{PHAT}(\mu)\, d\mu \end{aligned} \tag{8}$$

with $\Delta\mu^{(p)}(\mathbf{S}) = \mu^{(p)}_{max}(\mathbf{S}) - \mu^{(p)}_{min}(\mathbf{S})$. The 3-dimensional integration in Eq. (6) is reduced to a 1-dimensional integration.

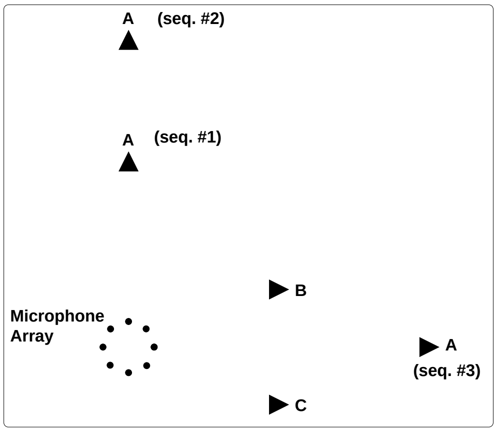


Figure 1: Top view of the experimental setup for Seq. #1, #2 and #3: 3 loudspeakers A,B,C.

3 Experimental Protocol

We use the "simplified SAM-PHAT" measure (defined in Eq. (8)), abbreviated hereafter as "SAM-PHAT". **We report sector-based experiments in two directions:**

First, we demonstrate that by using the SAM-PHAT measure, it is possible to accurately detect multiple concurrent sources. To do so, we use *all* sectors that are local maxima of SAM-PHAT, and assess whether or not each of the multiple active sources was correctly found. A sector is a local maximum when it has a higher SAM-PHAT value than all neighbouring sectors.

Second, we demonstrate that it is possible to limit the search space without losing accuracy. To do so, the same tests are repeated, using the N-best local maxima only.

For all studies reported here, the data comes from real recordings made in an instrumented meeting room [13] with a horizontal circular 8-microphone array (10 cm radius) set on a table. Section 3.1

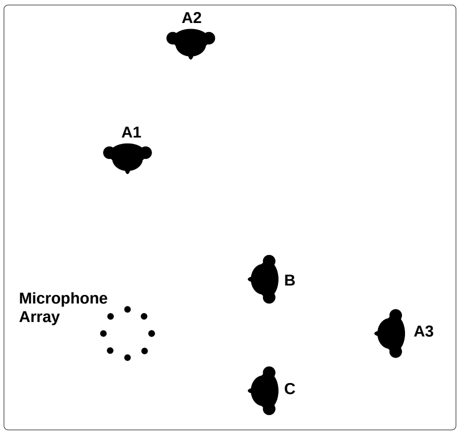


Figure 2: Top view of experimental setup for Seq. #5: 3 persons A,B,C. Person A speaks successively from 3 different locations A1, A2 and A3.

details the data. Section 3.2 gives a preliminary analysis of the results given by a simple SRP-PHAT [11] point-based grid search. Based on this analysis, Section 3.3 describes the protocol for sector-based experiments. The results are given and discussed in Section 4.

All results are expressed in terms of azimuth of the source relative to the microphone array. For all recordings the time frames are 32 ms long, with 16 ms overlap.

3.1 Data

Simultaneous speech was recorded from multiple non-moving acoustic sources. We recorded sequences #1, #2, #3 with loudspeakers in order to obtain absolute performance values, as explained in Section 3.2, while testing various loudspeaker locations. Seq. #4 and #5 are then used to show that the proposed approach also works on real human speech.

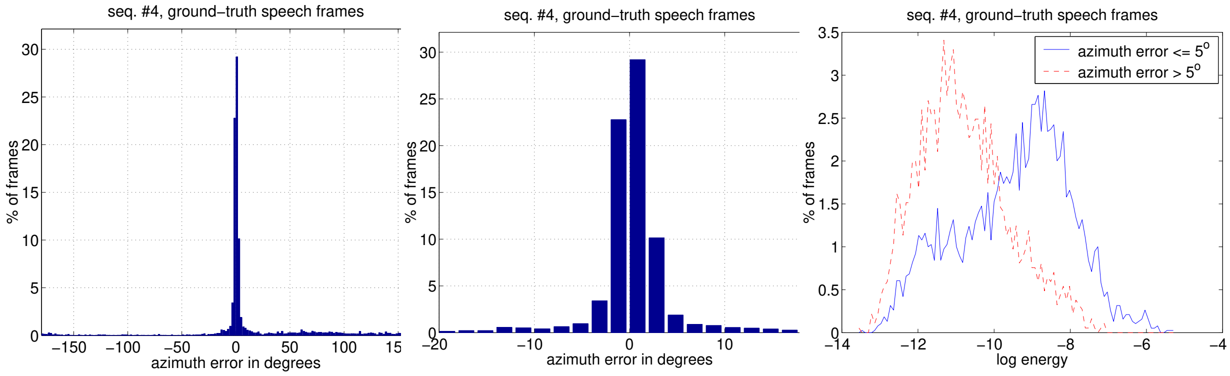
Seq. #1, #2, #3 each contain 20 minutes of synthetic speech, as an alternation of 4 seconds of stationary vowel sound followed by 2 seconds of silence. All possible combinations of 2 and 3 active loudspeakers and 5 different vowels are played sequentially. Vowels are synthesized using a LPC vocoder[1] and constant LPC coefficients, estimated from real speech. Fig. 1 shows the physical setup of the three loudspeakers. In Seq. #1, all three loudspeakers are placed at 0.8 m from the array, to test whether the proposed approach allows detection of sources with equal power. In Seq. #2, loudspeaker A is placed at 1.8 m from the array, to test if the proposed method works with one source being much further than the others. In Seq. #3, loudspeaker A is placed at 1.45 m from the array, in the middle direction between B and C. This tests whether the proposed approach can deal with a larger distance for A *and* lower angular separation.

Seq. #4 lasts 3 minutes 40 seconds. A single human speaker is recorded at each of 16 locations, covering an area that includes the five locations depicted in Fig. 2. Precisely, this area spans 121 degrees of azimuth and radius 0.7 m to 2.36 m, relative to the array.

Seq. #5 lasts 8 minutes 30 seconds: three human speakers, static while speaking. Speaker A spoke at three different locations A1, A2, A3. Fig. 2 shows the persons' locations.

In the loudspeaker case, precise speech/silence ground-truth (GT) segmentations and true 3D locations are known by construction. In the human case, speech/silence GT segmentations were provided by a human listener. We took particular care *not* to miss any speech in the GT segmentation, therefore GT speech segments often include silences - e.g. a pause between two words. 3D location

[1] Available at `http://www.tcts.fpms.ac.be/cours/1005-08/speech/lpcvocoder.zip`



(a) (b) = zoom of (a) (c)

Figure 3: SRP-PHAT point-based search on Seq. #4 (single human speaker): (a) shows the histogram of azimuth errors; (b) shows a zoom of (a); (c) shows the histogram of log energy values.

truth was provided with a 3D error (1.2 cm) negligible compared to the mouth size, from calibrated sameras (using CalTech's software[2], process not detailed here).

3.2 Preliminary Experiment

Parametric methods [3] suffer from a low angular resolution. The goal of this Section is to evaluate the effective angular resolution of the SRP-PHAT point-based measure. The motivation is that a similar angular resolution for the proposed sector-based measure SAM-PHAT can be expected, since it is also built on the time-domain GCC-PHAT function.

We ran a simple SRP-PHAT point-based single source localization algorithm (detailed in [14]) on *all* time frames of Seq. #4 (single human speaker). Figs. 3a and Fig. 3b show the distribution of azimuth errors for frames marked as "speech" in the GT. These figures are interpreted as follows:

- On frames containing speech strong enough to be localized, a maximum error of about 5 degrees is achieved, as compared with the true azimuth of the source.
- On frames containing silence or weak speech, the error can be seen as the result of a uniform random process.

A commonly used strategy for evaluating localization is to select speech frames with high energy only, and ignore other frames. However, we can see on Fig. 3c that in terms of energy, there is a large overlap between the two groups "correctly localized" and "incorrectly localized". **Therefore, all results reported here were computed using *all* frames marked as "speech" in the GT.**

For a given source, some of the GT "speech" frames may contain in fact weak speech or silence. The activity/silence priors $(\alpha, 1-\alpha)$ are assumed the same for all sources. In multisource cases, on GT "speech" frames the probability of each possible number of simultaneously active sources is derived from α, taking into account all possible combinations. For example, in the 3-source case:

P(0 active source) $= (1-\alpha)^3$.
P(1 active source) $= 3\alpha(1-\alpha)^2$.
P(2 active sources) $= 3\alpha^2(1-\alpha)$.
P(3 active sources) $= \alpha^3$.

In the case of loudspeakers (Seq. #1, #2 and #3), $\alpha = 1$, so the target histogram is an exact target and performance measures have an absolute meaning, as given in Tables 2a and 2b.

[2]Available at `http://www.vision.caltech.edu/bouguetj/ calib_doc/`

In the case of speech from humans (Seq. #4 and #5), we estimated $\alpha = 0.674$, as the proportion of frames where the point-based search is below 5 degree error on Seq. #4. The target histogram is therefore an indicative target only, as given in Tables 3a and 3b.

3.3 Metrics for Sector-Based Results

We first ran tests on Seq. #1, #2 and #3 in order to obtain absolute performance measures. Three types of tests were conducted in order to determine whether 1) the precision of the sector-based method compares with the precision of the point-based method, 2) multiple concurrent speakers can be localized correctly with the sector-based method, 3) use of the N-best sectors only is sufficient to achieve good results in the multiple sources cases. Finally, we ran tests on Seq. #4 and #5 to check whether the conclusions hold when loudspeakers are replaced with humans. In more details:

- **Precision**: Results are presented as an average across all locations. For each location, the proportion of speech frames having a local maximum of SAM-PHAT within 5 degree "azimuth error" of the true direction is estimated. Azimuth error is the angle between the true direction and the boundary of the closest sector being a local maximum of SAM-PHAT. When the true direction is in that sector, azimuth error is zero.
- **Localization of concurrent sources**: The histogram of the number of sources found simultaneously is calculated. On each frame labeled as "speech" in the GT, the number of simultaneous sources correctly localized is counted. "Correctly localized" means within 5 degree azimuth error.
- **N-best** : The same two tests are repeated, using the N-best local maxima of SAM values only. We show how the precision and the localization of concurrent speakers vary with N.

4 Results

4.1 Performance Evaluation: Seq. #1, #2 and #3

The space around the microphone array is partitioned into sectors as in Eq. (1) (no overlap between neighbouring sectors). Two types of partitions are used: 5-degree wide sectors and 10-degree wide sectors, respectively. In the following, "simplified SAM-PHAT" is abbreviated as "SAM-PHAT".

Precision: Table 1 shows for each sequence, the proportion of frames where a loudspeaker is correctly localized. Correct localization is obtained in all cases with 5-degree sectors, i.e. in more than 93% of the frames. This is particularly significant since the data always contains multiple concurrent sources. Results for 10-degree sectors show that using sectors that are too large degrades the performance. In the following we present results for 5-degree sectors only.

	Seq. #1	Seq. #2	Seq. #3
5^o sectors	98.6	98.4	93.7
10^o sectors	97.3	94.7	82.0

Table 1: Precision on Seq. #1, #2 and #3 (loudspeakers) with 5-degree sectors and 10-degree sectors: percentage of frames within 5 degree error (average of the 3 locations)

Localization of concurrent sources: The histogram of the number of sources correctly found is reported for 2-source and 3-source cases in Tables 2a and 2b, respectively. The rightmost column shows that the SAM-PHAT approach performs very well: in all cases but one, *all* active sources are found more than 92% of the time. On the remaining case (3 concurrent sources in Seq. #3), the performance is 80.5%. These results validate the use of SAM-PHAT to localize concurrent sources.

N-best sectors: The variation of the precision with N, on multiple source recordings, is reported in Fig. 4. The worst case is the most distant source: Seq. #2, location A. A possible interpretation

Number of loudspeakers found	0	1	2
Target histogram	0	0	100
Seq. #1, 5^o sectors	0.0	1.7	**98.3**
Seq. #2, 5^o sectors	0.0	1.8	**98.2**
Seq. #3, 5^o sectors	0.0	7.5	**92.5**

(a) 2 concurrent loudspeakers

Number of loudspeakers found	0	1	2	3
Target histogram	0	0	0	100
Seq. #1, 5^o sectors	0.0	0.2	4.6	**95.2**
Seq. #2, 5^o sectors	0.0	0.2	5.2	**94.6**
Seq. #3, 5^o sectors	0.0	2.3	17.3	**80.5**

(b) 3 concurrent loudspeakers

Table 2: Localization of concurrent sources (loudspeakers): number of sources found within each time frame (within 5 degree error). Values are percentages of GT "speech" frames with (a) 2 active sources, (b) 3 active sources.

is that the corresponding GCC-PHAT peak is smaller with increasing distance, because the power received by the array is smaller for A than for B or C.

We also examined how well multiple concurrent sources are *simultaneously* localized, as a function of N. Fig. 5 shows results for the 2-source and 3-source cases. Each point of the curve has the same meaning as the rightmost column of the histograms in Table 2a and 2b.

On all results we can see that N=6 is sufficient to obtain near optimal results. This shows that the search space can be greatly reduced for a minimal cost of performance.

4.2 Results with Human Speakers: Seq. #4 and #5

Based on Section 4.1 we used the 6-best local maxima only, to determine whether a reduced search space still allows to localize the speaker(s) in practice.

Precision: On Seq. #4 (a single speaker) we found that the speaker was correctly localized 79.2% of the time (average across the 16 locations). The worst location gave 60.3%. This compares very well with the estimated activity prior $\alpha = 0.674$ (see Section 3.2).

Localization of concurrent speakers: Seq. #5. Tables 3a and 3b show histograms of the number of sources correctly found within each time frame, along with an indicative "target" histogram. The "target" was computed based on the estimated activity/silence priors (see Section 3.2). From the results we can conclude that multiple concurrent speakers are accurately detected with the SAM-PHAT measure.

5 Conclusion

This paper introduced a generic approach for estimating speech activity in a given sector of the space. The motivation is twofold: to reduce the search space for existing multisource localization techniques, and to achieve multisource localization in practice. We proposed a Sector Activity Measure, called SAM-PHAT, which relies on one-dimensional summation of the time-domain GCC-PHAT function. We showed on more than one hour of real meeting room recordings that both goals are attained, including cases with 3 concurrent speakers. Future work will investigate integration of the SAM-PHAT measure into applications for automatic meeting data analysis.

Number of speakers found	0	1	2
Target histogram	10.7	44.0	45.3
Seq. #5, 5^o sectors	3.2	50.1	46.8

(a) 2 concurrent human speakers

Number of speakers found	0	1	2	3
Target histogram	3.5	21.6	44.4	30.5
Seq. #5, 5^o sectors	1.1	26.0	55.8	17.1

(b) 3 concurrent human speakers

Table 3: 6-best localization of concurrent human speakers: number of speakers found within each time frame (within 5 degree error). Values are percentages of GT "speech" frames with (a) 2 active speakers, (b) 3 active speakers.

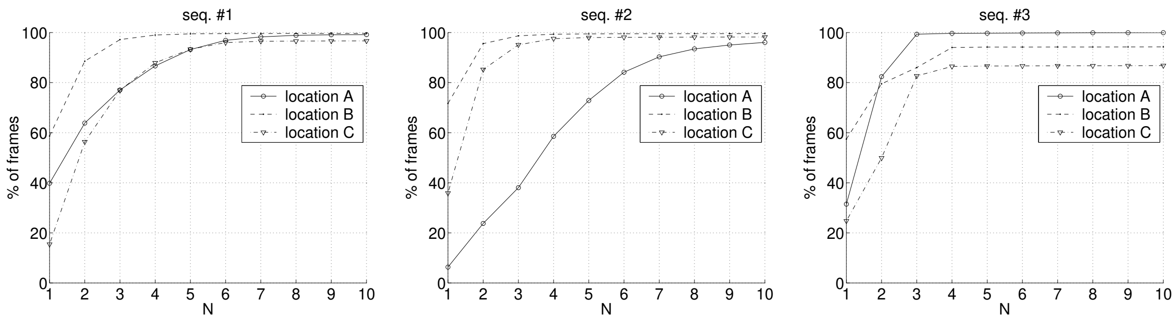


Figure 4: N-best 5-degree sectors: precision as a function of N.

6 Acknowledgments

The authors acknowledge the support of the European Union through the M4 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. This paper benefited from the valuable comments of Mathew Magimai.-Doss.

References

[1] E. Shriberg, A. Stolcke, and D. Baron. Observations on overlap: findings and implications for automatic processing of multi-party conversation. In *Proceedings of Eurospeech 2001*, volume 2, pages 1359–1362, 2001.

[2] J.J. Fuchs. On the application of the global matched filter to doa estimation with uniform circular arrays. *IEEE Transactions on Signal Processing*, 49(4), April 2001.

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

[4] P. Stoica and R. Mose. *Introduction to Spectral Analysis.* Prentice-Hall, 1997.

[5] R.O. Schmidt. Multiple Emitter Location and Signal Parameter Estimation. *IEEE Transactions on Antennas and Propagation*, AP-34:276–280, March 1986.

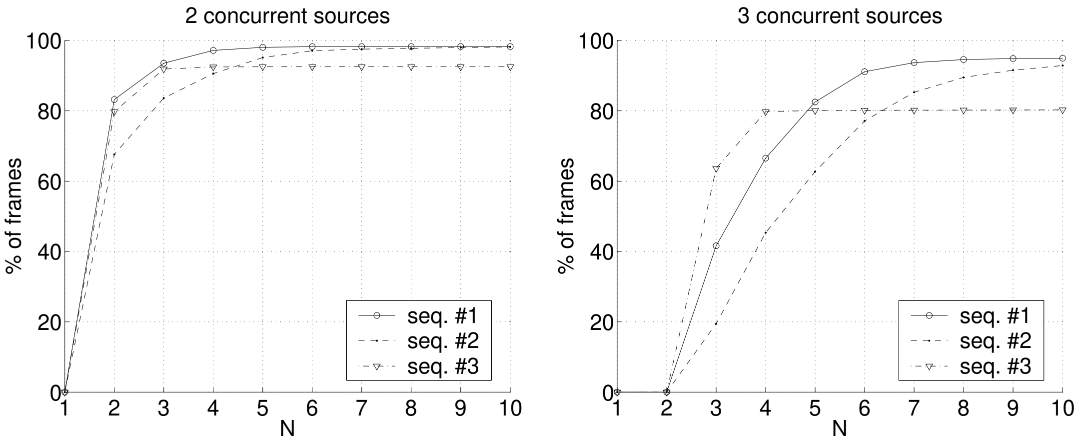


Figure 5: N-best 5-degree sectors: correct localization of *all* concurrent sources, as a function of N.

[6] R. Roy and K. Kailath. ESPRIT - Estimation of Signal Parameters via Rotational Invariance Techniques. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(7):984–995, July 1989.

[7] A.H. Tewfik and W. Hong. On the application of uniform linear array bearing estimation techniques to uniform circular arrays. *IEEE Transactions on Signal Processing*, 40(4), April 1992.

[8] G. Su and M. Morf. Signal subspace approach for multiple wide-band emitter location. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 31(12):1502–1522, December 1983.

[9] B. Friedlander and A.J. Weiss. Direction finding for wide-band signals using an interpolated array. *IEEE Transactions on Signal Processing*, 41(4), April 1993.

[10] R. Duraiswami, D. Zotkin, and L.S. Davis. Active speech source localization by a dual coarse-to-fine search. In *Proceedings of ICASSP 2001*, 2001.

[11] J. DiBiase, H. Silverman, and M. Brandstein. Robust localization in reverberant rooms. In M. Brandstein and D. Ward, editors, *Microphone Arrays*, chapter 8, pages 157–180. Springer, 2001.

[12] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustic, Speech and Signal Processing*, ASSP-24(4):320–327, August 1976.

[13] D. Moore. The IDIAP Smart Meeting Room. IDIAP-COM 07, IDIAP, 2002.

[14] Daniel Gatica-Perez, Guillaume Lathoud, Iain McCowan, and Jean-Marc Odobez. A Mixed-State I-Particle Filter for Multi-Camera Speaker Tracking. In *2003 IEEE Int. Conf. on Computer Vision Workshop on Multimedia Technologies for E-Learning and Collaboration (ICCV-WOMTEC)*, 2003.