A SCALABLE FRAMEWORK FOR MULTIPLE SPEAKER LOCALIZATION AND TRACKING

Nilesh Madhu and Rainer Martin

{firstname}.{lastname}@rub.de
Institute of Communication Acoustics (IKA), Ruhr-Universität Bochum
44780 Bochum, Germany.

ABSTRACT

In this paper we present a novel, scalable approach to the localization and tracking of multiple speakers using microphone arrays. The approach is capable of localizing sources both in non-competing and in concurrent situations, and is based on the disjointness of speech in the short-time discrete frequency domain (STFD). The algorithm operates on a narrowband localization cost function in the STFD and yields an estimate of the speaker activity per time frame b by applying a Mixture of Gaussians (MoG) fit to the narrowband localization estimates. The advantages of the proposed method are manifold: it allows us to use a coarser search grid for the cost function evaluation, without compromising on the location accuracy; it allows for a *soft-decision* on the number and position of sources on the fly; the framework is scalable to multi-array systems; and it can also serve as a base framework for enhancement algorithms. In principle, this approach is not specific to speakers and will work for any source combination provided they exhibit some temporal and spectral disjointness.

Index Terms— Localization, Mixture of Gaussians, Tracking, Multi-talker localization, SRP

1. INTRODUCTION

In multichannel acoustic signal processing, source localization is an important component, commensurate with its number of applications such as speech processing, industrial acoustics and sonar to cite a few. These applications are different in that the sources in each case may be broadband or narrowband, may evince spectral or temporal disjointness or both, may show sparsity in the temporal or spectral domain, etc. Depending upon the source characteristics, which are usually known *a priori*, localization algorithms need to be appropriately tailored for the purpose. The contribution here focusses on speaker localization. Under certain conditions, which shall become clear subsequently, the proposed approach may be extended to other source combinations as well.

Localization is done either by the direct approaches (cost function computed over a set of preselected candidate locations) or the indirect approaches (estimate the inter-sensor time delays of arrival (TDOA) and find the source locations by non-linear optimization and parameter fitting). When more than two microphones are available, the approach of choice is frequently the steered response power (SRP) [1] algorithm (and its variants) due to its ease of implementation and scalability in terms of addition of new sensors and selection of the candidate locations (see e.g., [2]). These algorithms are usually implemented in the short-time discrete frequency (STFD) domain.

An important problem in this context is the adaptive detection of the number of active speakers. Detection can be done using Akaike's Information Criterion (AIC), Rissanen's Minimum Description Length (MDL), or the Bayesian Information Criterion (BIC) (see, e.g, [3] and references therein). But the formulation of these criteria is difficult for the broadband case, especially where disjoint sources like speech are concerned. Moreover, the detection problem is coupled with the localization, requiring a multidimensional non-linear maximum likelihood optimization, which adds to the complexity. Furthermore, speech signals in a natural scenario are dynamic: a speaker may start, be active for a while, fall silent, and then start again. Even *within* active speaker segments, we have speech pauses. For these reasons, most applications either assume the number of concurrently active speakers to be known or implicitly assume a single dominant speaker.

The imperative questions handled in this contribution are:

- deciding *when* and *where* (in an appropriately defined reference system) a speaker was active,
- determination of the number of active sources in any time frame,
- deciding when a new source has become active ('birth' of a source), and
- deciding when a source has fallen silent ('death' of a source).

Note that we shall impose *neither* the constraint of constant multi-speaker activity (competing situation) *nor* that of single source dominance.

The document is ordered as follows: in the next section, the signal model and the assumptions the approach is based on are described. The proposed approach is then detailed. Finally, the effectiveness of the proposed approach is illustrated on single- and multiple- speaker recordings made in a reverberant and noisy room.

2. SIGNAL AND LOCALIZATION MODEL

We consider the case of an M sensor array, with the localization being done along the azimuth: $\theta \in [0, 180^\circ]$, where θ is measured with respect to the array axis. The signals input to the array are segmented, windowed and transformed into the discrete frequency domain using the discrete Fourier transform (DFT). Next, for each transformed frame b, we select the subset $\{k : k_{\text{low}} < k \leq K/2\}$ of K' bins from the K bins available to us, where K is the length of the DFT. The upper bound is due to the symmetric nature of the DFT, which makes the upper half of the spectrum ($k > K/2$) redundant and the lower bound is because very low frequencies do not yield good directional estimates.

Next, for each selected discrete frequency bin k of each frame b, we compute the SRP functional $\mathcal{J}_{\mathrm{SRP}}(\theta, k, b)$ [4, 1] over a pre-selected grid of search locations. From this, we may compute an estimate of the source azimuth as:

$$\widehat{\theta}(k, b) = \underset{\theta}{\operatorname{argmax}}\, \mathcal{J}_{\mathrm{SRP}}(\theta, k, b). \tag{1}$$

Under the assumption of speech disjointness and given the sparsity of speech, each time-frequency point (k, b) can be attributed to a single dominant speaker. Thus, the $\widehat{\theta}(k, b)$ indicate the dominant source location at that bin and frame[1]. Consequently, over multiple bins of a single frame, we should have enough data to approximate the speaker locations. This is done by clustering the $\widehat{\theta}(k, b)$ estimates obtained for each frame b.

3. MOG MODELLING

For the clustering, we model the vector sequence

$$\widehat{\boldsymbol{\theta}}\,(b) = \left(\widehat{\theta}(1, b),\, \dots,\, \widehat{\theta}(K', b)\right)^T \tag{2}$$

as a set of K' realizations of a MoG process and estimate the parameters of this process using the Expectation Maximization (EM) [7] approach. As this estimation is done on a per-frame basis, we shall subsequently drop the frame index for convenience and reintroduce it when necessary. Further, as the number of sources is not known *a priori*, we start with a predefined model order $\mathcal{I}$, where $\mathcal{I}$ is selected to be an over-estimation. The EM clustering on the K' values $\widehat{\theta}$ yields
the means : $\boldsymbol{\theta} = (\theta_1,\, \dots,\, \theta_{\mathcal{I}})^T$,
the variances: $\boldsymbol{\Xi} = \left(\sigma_1^2,\, \dots,\, \sigma_{\mathcal{I}}^2\right)^T$, and
the weights/probabilities: $\mathbf{P} = (P_1,\, \dots,\, P_{\mathcal{I}})^T$
of the $\mathcal{I}$ components.

These initial values may overdetermine the underlying process. Thus, we shall 'shrink' our model if necessary. To this end we define a shrink threshold Υ and :

if $\exists i, i'$, such that $|\theta_i - \theta_{i'}| \leq \Upsilon$

$$\begin{aligned} \theta_i &\leftarrow \frac{P_i\theta_i + P_{i'}\theta_{i'}}{P_i + P_{i'}} \\ \sigma_i^2 &\leftarrow \frac{P_i\sigma_i^2 + P_{i'}\sigma_{i'}^2}{P_i + P_{i'}} \\ P_i &\leftarrow P_i + P_{i'} \\ \mathcal{I} &\leftarrow \mathcal{I} - 1 \end{aligned} \tag{3}$$

The rationale behind the selected shrinkage operation is to remove the θ_i that are very close together (in practical situations, we do not have point sources and thus, the sources always have a minimum separation, which is indicated by Υ). Following the sequence of steps in (3), the $\mathcal{I} - 1$ parameters are re-estimated *using the newly averaged values* as initial seeding for the EM:

$$\begin{aligned} \boldsymbol{\theta}_{\mathrm{init}} &= (\theta_1,\, \dots,\, \theta_i,\, \dots,\, \theta_{i'-1},\, \theta_{i'+1},\, \dots, \theta_{\mathcal{I}})^T \\ \boldsymbol{\Xi}_{\mathrm{init}} &= \left(\sigma_1^2,\, \dots,\, \sigma_i^2,\, \sigma_{i'-1}^2,\, \sigma_{i'+1}^2,\, \dots, \sigma_{\mathcal{I}}^2\right)^T \\ \mathbf{P}_{\mathrm{init}} &= (P_1,\, \dots,\, P_i,\, P_{i'-1},\, P_{i'+1},\, \dots, P_{\mathcal{I}})^T \end{aligned} \tag{4}$$

This process is repeated until all cluster centroids have a minimum separation of Υ. The number of clusters so obtained indicate the number of sources in that frame, with their respective probability of activity and their variance. Note that further shrinkage can be obtained by setting thresholds on the weights P_i or the variances σ_i^2. This would better help eliminate transient sources. However, MoG components with low weights could also indicate the onset of a source. Therefore, in our implementation, we shall not impose such thresholds, and rely instead on the tracking algorithm for the purpose of discarding transient sources.

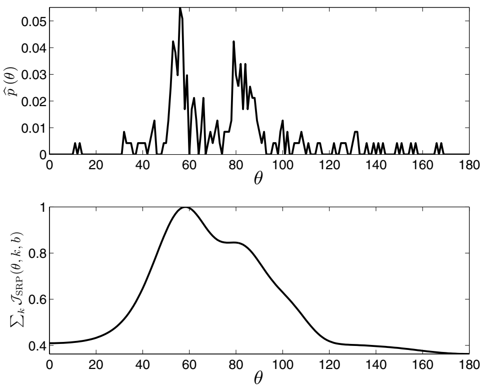


Fig. 1. Note that the histogram clearly indicates contributions from two major azimuths. The frequency averaged SRP cost function is not as clear.

As a further point of interest, note that simply averaging the SRP functional over the frequencies and extracting their maxima to find the source location and number is not necessarily a viable alternative to the clustering of the individual azimuth values. This is illustrated in Fig. 1.

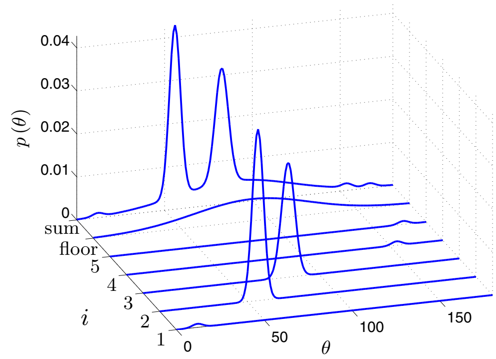


Fig. 2. MoG decomposition of the histogram from Fig. 1 using $\mathcal{I} = 5$ components. The two sources are clearly visible as the largest components, 'floor' is the noise floor component, and 'sum' is the net MoG model of the histogram from Fig. 1.

3.1. Modelling the noise floor

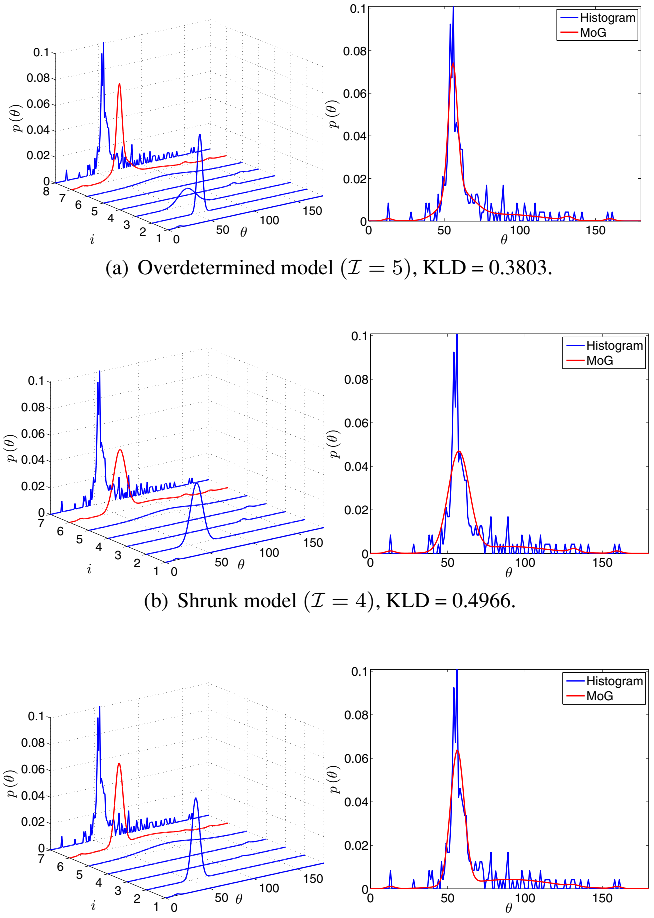
In bins where no source is active, the estimated $\widehat{\theta}$ is randomly distributed over the azimuth space. This noise floor is modelled in the MoG as a 'hidden' component $(\mathcal{I} + 1)$, with constant mean $(\theta_{(\mathcal{I}+1)} = 90°)$ and a large standard deviation $\sigma_{(\mathcal{I}+1)} > \Upsilon_v$. For

[1] In [5, 6], clustering is performed on the TDOA estimates $\tau_{mm'}$ over all microphone pairs (m, m') of the array, and for each bin and each frame. This leads to a vector clustering model in contrast to the simpler scalar clustering in the SRP case. However, our approach is applicable in both models.

this hidden component, only the weight and variance are adapted in the EM step, with the standard deviation constrained to be above Υ_v. This noise floor component is *not* assigned the status of a source.

3.2. Why re-estimate?

To justify the necessity of the re-estimation step after updating the parameters as in (3), consider Fig. 3. This presents the histograms and the estimated MoG fit (the penultimate curve) for the three cases: (a) initial estimation (overdetermined), (b) shrunk model (from (3)) and (c) re-estimated, more compact model ($\mathcal{I} - 1$ elements) with initial seeding from (4). The noise floor is indicated by the component before the estimated MoG fit and the histogram obtained from the SRP functional is the last component. We see that the MoG fit to the histogram improves after re-estimation, as compared to simply shrinking. This is clearly visible and is also indicated by the corresponding, lower, Kullback-Leibler distance (KLD).



(a) Overdetermined model ($\mathcal{I} = 5$), KLD = 0.3803.

(b) Shrunk model ($\mathcal{I} = 4$), KLD = 0.4966.

(c) Shrunk and re-estimated model ($\mathcal{I} = 4$), KLD = 0.4510.

Fig. 3. MoG fit illustrating the effect of shrinkage. (a) indicates the initial (overdetermined) estimate; (b), the estimate after shrinkage according to (3); and (c), the model obtained after shrinkage and re-estimation with initial seeding as in (4). The axis labelling is the same as in Fig. 2. One source was active.

4. SOURCE TRACKING

In general, the number of sources changes from frame to frame. This time variance of $\mathcal{I}$ is due to the following reasons:

- some means might be spurious (errors in the cost function, especially at lower frequencies and frequencies with low SNR),
- some means might indicate the onset of a new source,
- additionally, it could also happen that a source detected in the previous frame(s) is not present in the current frame (within the threshold window $[-\Upsilon, \Upsilon]$) – indicating a speech pause for that source or the sign that it is dying out.

Note that only a constantly active source would contribute to a MoG component (θ_i) that does not change significantly from frame to frame. The MoG model of Section 3 cannot account for this time variant nature of $\mathcal{I}$ and cannot distinguish between transient and active sources. Therefore, we shall extend our localization framework to include source tracking in order to preserve sources of interest and discard transient sources.

For this, we maintain a *frame-independent* record of *averaged* means, variances and probabilities, denoted as:

$$\begin{aligned}\overline{\boldsymbol{\theta}} &= \left(\overline{\theta_1}, \ldots, \overline{\theta_{\mathcal{I}_T}}\right)^T \\ \overline{\boldsymbol{\Xi}} &= \left(\overline{\sigma_1^2}, \ldots, \overline{\sigma_{\mathcal{I}_T}^2}\right)^T \\ \overline{\mathbf{P}} &= \left(\overline{P_1}, \ldots, \overline{P_{\mathcal{I}_T}}\right)^T\end{aligned} \tag{5}$$

where $\mathcal{I}_T$ indicates the number of elements currently present in the averaged mixture. Further, borrowing an idea from packet-switched networks, we associate with each source $\bar{i}$ in the averaged set a 'time to live', $\mathrm{TTL}_{\bar{i}}$. Now consider the $\mathcal{I}\,(b)$ elements of the current frame b, obtained as detailed in the previous section.

If $\exists i\,(b)\,, \bar{i},$ such that $|\theta_{i(b)} - \overline{\theta_{\bar{i}}}| \leq \Upsilon$

$$\begin{aligned}\overline{\theta_{\bar{i}}} &\leftarrow \frac{\overline{\theta_{\bar{i}}}\,\overline{P_{\bar{i}}} + \theta_{i(b)}P_{i(b)}}{\overline{P_{\bar{i}}} + P_{i(b)}} \\ \overline{\sigma_{\bar{i}}^2} &\leftarrow \frac{\overline{\sigma_{\bar{i}}^2}\,\overline{P_{\bar{i}}} + \sigma_{i(b)}^2 P_{i(b)}}{\overline{P_{\bar{i}}} + P_{i(b)}} \\ \overline{P_{\bar{i}}} &\leftarrow \overline{P_{\bar{i}}} + P_{i(b)} \\ \mathrm{TTL}_{\bar{i}} &\leftarrow \min\left(\mathrm{TTL}_{\max}, \mathrm{TTL}_{\bar{i}} + 1\right)\end{aligned} \tag{6a}$$

and, for each $i\,(b)$ such that $\nexists\,\bar{i}$, with $|\theta_{i(b)} - \overline{\theta_{\bar{i}}}| \leq \Upsilon$,

$$\begin{aligned}\overline{\boldsymbol{\theta}} &\leftarrow \overline{\boldsymbol{\theta}} \cup \theta_{i(b)} \\ \overline{\boldsymbol{\Xi}} &\leftarrow \overline{\boldsymbol{\Xi}} \cup \sigma_{i(b)}^2 \\ \overline{\mathbf{P}} &\leftarrow \overline{\mathbf{P}} \cup P_{i(b)} \\ \mathcal{I}_T &\leftarrow \mathcal{I}_T + 1 \\ \mathrm{TTL}_{\mathcal{I}_T} &\leftarrow \mathrm{TTL}_{\min}\end{aligned} \tag{6b}$$

and, for each $\bar{i}$ such that $\nexists\, i\,(b)\,$, with $|\theta_{i(b)} - \overline{\theta_{\bar{i}}}| \leq \Upsilon$,

$$\mathrm{TTL}_{\bar{i}} \leftarrow \mathrm{TTL}_{\bar{i}} - 1 \tag{6c}$$

Following this, $\overline{\mathbf{P}}$ is renormalized to guarantee $\sum_{\bar{i}} \overline{P_{\bar{i}}} = 1$. Equation (6a) indicates the update for a source that was already present in previous frames. Equation (6b) handles the situation where a possible new source has entered the system and (6c) indicates the case where an existing source was absent in the current frame. Within this framework, sources with a TTL ≤ 0 are considered to have 'died' and are removed from the averaged set. Note that the limitation $\mathrm{TTL}_{\max}$ is required in order to limit the source lifetime, giving us a reasonable period of 'aging' and 'death' for a source that is not active anymore.

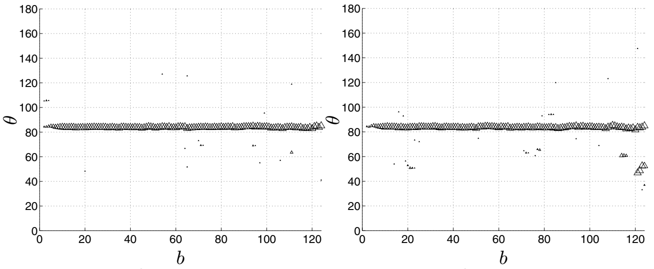
4.1. Source Number Estimation

After the updates in (6), the number of active speakers in each frame is determined, based on the parameters of the respective MoG components. Note that the MoG model provides us with a rich set of parameters to base our decision upon. For our implementation, we choose to denote a source as active if the *standard deviation* of the corresponding component is below Υ. This choice is based on the observation that MoG components that model active sources demonstrate a peaky distribution, with a low variance. The locations of the active speakers are then given by the means of the selected MoG components.

5. EXPERIMENTAL EVALUATIONS

We shall illustrate the performance of the proposed approach on recordings made in a reverberant room ($T_{60} = 0.5\,\text{s}$, critical distance $\approx 0.8\,\text{m}$), using a 5 element linear microphone array, with the elements placed at distances of 3, 8, 15, and 25 cm, respectively, from the first element. The sources were positioned at a distance of around $1.0\,\text{m}$ from the array center, and were sampled at a rate of $f_\text{s} = 8\,\text{kHz}$. White noise at 5dB SNR was added to the recorded signals for the experiments. The DFT analysis was based on a $K = 512$ point DFT, with a 50% overlap and a von Hann window. The shrink threshold was set to $\Upsilon = 10°$ and the noise floor threshold was set to $\Upsilon_v = 20°$. For each frame, the MoG model was initialized with a fixed number of elements $\mathcal{I} = 5$, and the parameters estimated (with shrinkage if necessary). Due to space constraints, the evaluation here does not cover all aspects of the proposed algorithm. Further results may be found at `http://www.rub.de/ika/ika/forschung/gruppe_martin/hum_mach_interf/hum_mach_eng.htm`.

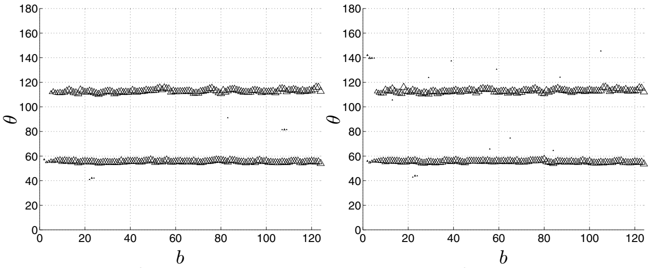
Fig. 4 indicates the performance of the proposed approach on a single source around broadside. Note that the performance is not degraded significantly when the azimuth resolution of the SRP search grid is decreased from 1° to 7°. The size of the marker is proportional to the TTL of the source. This allows us to see the birth-death process of the sources more clearly.



(a) 1° resolution (b) 7° resolution

Fig. 4. Localization results using the proposed model on SRP functionals with varying search grid resolutions. Signals were from one speaker, close to broadside ($\theta \approx 85°$) with a background SNR of 5dB.

Fig. 5 indicates the performance of the system for two simultaneously active speakers (approx ±30°) away from broadside. Again, we see that the system is able to track both speakers with low false positives, even when the resolution of the SRP search grid is lowered. This speaks both for the capability of the proposed approach and the parameter choice for source number detection.



(a) 1° resolution (b) 7° resolution

Fig. 5. Localization results for two competing speakers, approximately ±30° around the broadside. The background SNR was 5dB.

6. CONCLUSIONS

We have proposed a simple and elegant framework for the simultaneous localization of mutliple speakers, exploiting the sparsity and disjointness of speech for the purpose. We have further shown that our framework has the additional benefit of providing the same localization accuracy even when the SRP functional is computed over a coarser search grid. This reduces the computational load without compromising accuracy. This is particulary beneficial for arrays with a large number of sensors as in [2]. Additionally, we have presented a simple framework for tracking the sources. In order to more accurately model moving sources, the framework may be enhanced by suitable state-of-the-art algorithms (e.g., [8]). This is a subject for future research.

7. REFERENCES

[1] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, “Robust localization in reverberant rooms,” in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Springer-Verlag, 2001.

[2] H. F. Silverman, Y. Yu, J. M. Sachar, and W. R. Patterson, III, “Performance of real-time source-location estimators for a large-aperture microphone array,” *IEEE Trans. Speech and Audio*, vol. 13, no. 4, July 2005.

[3] M. Wax and T. Kailath, “Determining the number of signals by information theoretic criteria,” in *Proc. IEEE ICASSP*, 1984.

[4] N. Madhu and R. Martin, “Acoustic source localization with microphone arrays,” in *Advances in Digital Speech Transmission*, R. Martin, U. Heute, and C. Antweiler, Eds. John Wiley & Sons, Ltd., 2008.

[5] J. Cermak, S. Araki, H. Sawada, and S. Makino, “Blind speech separation by combining beamformers and a time frequency mask,” in *Proc. IWAENC*, 2006.

[6] S. Araki, H. Sawada, R. Mukai, and S. Makino, “A novel blind source separation method with observation vector clustering,” in *Proc. IWAENC*, 2005.

[7] J. A. Bilmes, “A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models,” Tech. Rep. TR-97-021, U.C. Berkeley, 1998.

[8] S. Gannot and T. G. Dvorkind, “Microphone array speaker localizers using spatial-temporal information,” *EURASIP Journal of Applied Signal Processing*, no. 1, 2006.