A PROBABILISTIC FRAMEWORK FOR MULTIPLE SPEAKER LOCALIZATION

Youssef Oualil Mathew Magimai.-Doss
Friedrich Faubel Dietrich Klakow

Idiap-RR-37-2012

Version of JANUARY 15, 2013

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

A PROBABILISTIC FRAMEWORK FOR MULTIPLE SPEAKER LOCALIZATION

Youssef Oualil[1,2], *Mathew Magimai.-Doss*[2], *Friedrich Faubel*[1], *Dietrich Klakow*[1]

[1] Spoken Language Systems, Saarland University, Saarbrücken, Germany
[2] Idiap Research Institute, CH-1920 Martigny, Switzerland
youssef.oualil@lsv.uni-saarland.de

ABSTRACT

This paper presents a novel probabilistic framework for localizing multiple speakers with a microphone array. In this framework, the generalized cross correlation function (GCC) of each microphone pair is interpreted as a probability distribution of the time difference of arrival (TDOA) and subsequently approximated as a Gaussian mixture. The distribution parameters are estimated with a weighted expectation maximization algorithm. Then, the joint distribution of the TDOA Gaussian mixtures is mapped to a multimodal distribution in the location space, where each mode represents a potential source location. The approach taken here performs the localization by 1) reducing the search space to some regions that are likely to contain a source and then 2) extracting the actual speaker locations with a numerical optimization algorithm. The effectiveness of the proposed approach is shown using the AV16.3 corpus.

Index Terms— Microphone arrays, localization, multiple speakers, Gaussian mixture, steered response power.

1 Introduction

Acoustic source localization using microphone arrays has become an essential tool for developing more robust and accurate solutions to a large number of signal processing problems, such as speech separation/enhancement and speaker diarization/tracking. Acoustic source localization approaches can be divided into two main categories: two-step approaches, where the source location is extracted by virtue of geometrical intersection [1, 2, 3, 4] and single-step approaches, which aim at inferring the source location directly from the signals, such as multi-channel cross correlation (MCCC) [5], adaptive eigenvalue decomposition [6, 7, 8], and steered response power (SRP) based techniques (e.g. [9, 10, 11, 12]).

These approaches were originally developed for single speaker localization and have then be extended to multiple speakers by using agglomerative clustering techniques [13], Gaussian mixture (GM) approximations in the location space [14, 15] as well as a sector based approach [16, 17]. Despite their relative success, these extensions have some inherent shortcomings such as a high computation cost due to the discretization of the entire space [9], the discrete search dilemma (computation cost versus precision) [10, 11, 12], and a general difficulty with jointly estimating of the number and locations of multiple simultaneous speakers [14, 15, 16, 17].

In recent work [18], we proposed a probabilistic interpretation of the SRP (denoted as pSRP) that addressed the aforementioned problems. This approach however, as most SRP-based localization methods [14, 15], does not address the problem of local maxima resulting from the multivalued TDOA-location function, which can strongly affect multiple speaker detection, especially when the number of microphones is low. Following a line of thought similar to our previous work [18], we present a novel probabilistic framework which consists of i) approximating each GCC function by a GM model using a weighted expectation maximization (WEM) algorithm (Section 2), ii) using the GM models in order to obtain a joint probability density function (pdf) that describes the entire multidimensional space of the TDOAs, iii) mapping this distribution to the source location space (Section 3) and, then, iv) using the resulting source location pdf to identify regions that are likely to contain active sources. The actual location estimates are subsequently obtained with a numerical optimization algorithm (Section 4).

The advantage of this new approach is that it preserves the computational efficiency and accuracy of our previous work [18] while additionally solving the problem of multiple local maxima. The extension to multiple speakers is straight-forward (Section 5). The effectiveness of the proposed approach is finally demonstrated through an experimental study in Section 6, including comparisons to the SRP, pSRP, and MCCC on a single speaker localization task, and to the pSRP on a multiple speaker localization task.

This work was supported by the European Union through the Marie- Curie Initial Training Network (ITN) SCALE (Speech Communication with Adaptive LEarning, FP7 grant agreement number 213850); by the Federal Republic of Germany, through the Cluster of Excellence for Multimodal Computing and Interaction (MMCI); and by the Swiss National Science Foundation through the Swiss National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM2).

2 GCC function as PDF of the TDOA

Let M denote the number of microphones, and let $s_g(t)$ denote the signal received at microphone $\mathbf{m}_g$, $g = 1, \dots, M$. Then the generalized cross correlation (GCC) function $\mathcal{R}_q$ of the microphone pair $q = \{\mathbf{m}_g, \mathbf{m}_h\}$ is given by

$$\mathcal{R}_q(\tau) = \frac{1}{2\pi} \int_0^{2\pi} \psi(\omega) S_g(\omega) S_h^*(\omega) e^{j\omega\tau} \mathrm{d}\omega \tag{1}$$

where $S_.(\omega)$ denotes the short-time Fourier transforms of $s_.(g)$ and where $\psi(\omega)$ denotes a pre-filter. A common choice of $\psi(\omega)$ is the phase transform (PHAT) weighting [19].

The TDOA a source introduces at a microphone pair is estimated as the time alignment which maximizes the GCC function of the signals. Hence, the higher the GCC value the more likely it is that the alignment is the "true" TDOA. From this point of view, the *normalized* cross-correlation of two signals can be interpreted as a pdf of the TDOA. Alternatively, the GCC function could be regarded as a set of observations from a hidden distribution. In this work, the hidden distribution is a GM model. This choice is justified by the multi-modality of the GCC function in noisy and/or reverberant environments. Furthermore, the Gaussianity assumption of the TDOA error has been proven to be a valid assumption in speaker tracking approaches [20, 21]. For more details, the reader is referred to [18, 22].

Let $\{\tau_i^q\}_{i=1}^{C^q}$ be the set of TDOA values of the q-th microphone pair and let $\{w_i^q\}_{i=1}^{C^q}$ be the corresponding *normalized* GCC values. Negative GCC values, if there is any are set to zero. Our goal now is to estimate the parameters $\Theta = \{\mu_k^q, \sigma_k^q, p_k^q\}_{k=1}^{K^q}$ of the GM distribution ($K^q \leq C^q$), which optimally approximates the GCC function at $\{\tau_i^q\}_{i=1}^{C^q}$. The μ_k^q, σ_k^q and p_k^q here denote the mean, standard deviation and mixture weight of the k-th component. We assume that each peak in the GCC function results from an acoustic event in the room, which could have been either generated by a desired source or by noise sources (the reverberation paths are assumed to be generated by virtual noise sources). Therefore, the number of mixture components K^q is given by the number of GCC peaks. For the sake of readability, the microphone pair index q is dropped in the remaining part of this section. We propose to estimate the hidden paramters Θ using a *weighted expectation maximization* (WEM) algorithm, which iterates between two steps:

An **Expectation Step** which performs a soft assignment of the observations to the mixture components. This is given by the probabilities p_{ik}, $i = 1, \dots, C$ and $k = 1, \dots, K$, which indicate how likely it is that the observation τ_i was generated by component k. Formally: $p_{ik} = \mathcal{P}(l_i = k | \tau_i, \Theta)$ where l_i is the sample label in the mixture.

A **Weighted Maximization Step** in which the mixture parameters Θ are re-estimated using the new soft assignment:

$$C_k = \sum_{i=1}^{C} w_i \cdot p_{ik}, \qquad p_k = \frac{C_k}{\sum_{k=1}^{K} C_k} \tag{2}$$

$$\mu_k = \frac{1}{C_k} \sum_{i=1}^{C} w_i \cdot p_{ik} \cdot \tau_i \tag{3}$$

$$\sigma_k^2 = \frac{1}{C_k} \sum_{i=1}^{C} w_i \cdot p_{ik} \cdot (\tau_i - \mu_k)^2 \tag{4}$$

Here, C_k can be interpreted as the soft proportion of samples that are labeled with k. The observation weights w_i incorporate the individual information about the peaks and valleys of the GCC function. As a result, the approximation relies more on the regions with a high likelihood, whereas it ignores the ones with low likelihood. This can be regarded as a "pseudo-regression" of the GCC function using a GM distribution (see example in Figure 1).

In practice, the WEM algorithm divides the TDOA space into TDOA-intervals $\{I_k\}_{k=1}^{K}$, subsequently called *intervals of dominance*, where a single component is assumed to be dominant in each interval. Formally, the interval I_k associated with the k-th component is given by: $I_k = [\tau_k^{min}, \tau_k^{max}]$ where τ_k^{min} and τ_k^{max} are respectively the minimum and maximum values of $\{\tau_i | l_i = k\}_{i=1}^{C}$. The approach proposed in [18] can be regarded as a practical approaximation of the WEM algorithm.

3 Joint PDF of the Source Location

After estimating the GM distribution for all microphone pairs, we create a joint pdf of the TDOAs $p(\tau^1, \dots, \tau^Q)$ under the assumption that the τ^q , $q = 1, \dots, Q$ are independent random variables:

$$p(\tau^1, \dots, \tau^Q) = \prod_{q=1}^{Q} p(\tau^q) = \prod_{q=1}^{Q} \sum_{k=1}^{K^q} p_k^q \cdot \mathcal{N}^q(\tau^q; \mu_k^q, (\sigma_k^q)^2) \tag{5}$$

Theoretically, this joint pdf describes the entire multidimensional joint space of the TDOAs. In practice however, given that the TDOA distributions were generated by the same mixture of acoustic events in the room, and captured by the same set of

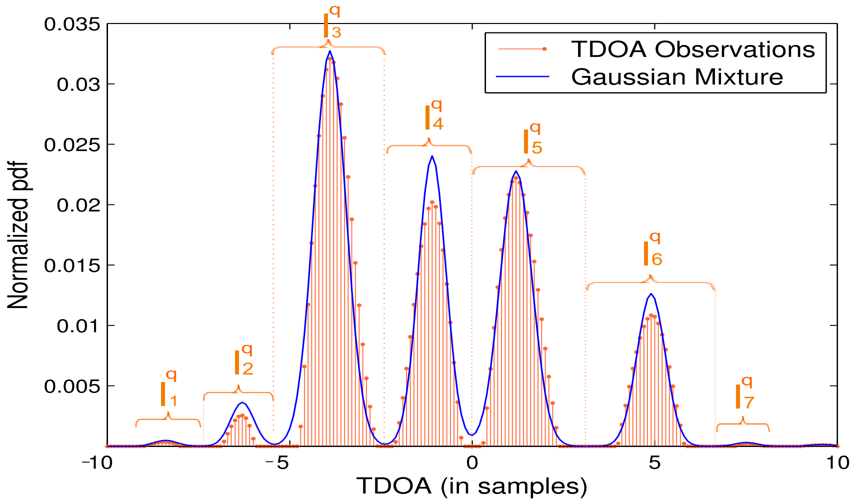


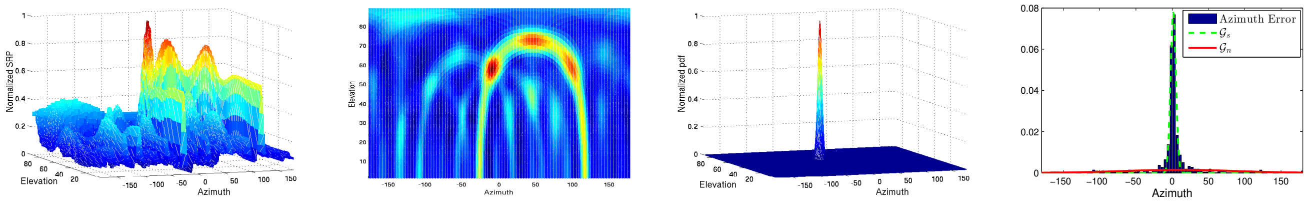
Figure 1: *The TDOA Gaussian mixture distribution estimated with the weighted expectation maximization algorithm ($K^q = 7$). The curly brackets (on top) indicate the intervals of dominance, I_k.*

microphones, we can conclude that not every TDOA combination is physically possible. The subspace of physically possible combinations is obtained by simply mapping the joint pdf to the location space using the TDOA-location function:

$$\tau^q(\mathbf{s}) = \frac{\|\mathbf{s} - \mathbf{m}_g\| - \|\mathbf{s} - \mathbf{m}_h\|}{c}. \tag{6}$$

In this equation, $\mathbf{s}$ denotes the location and and c denotes the speed of sound in the air. The incorporation of this mapping into the joint TDOA pdf from (5) leads to a pdf $p(\mathbf{s})$ that describes the mixture of acoustic events in the location space:

$$p(\mathbf{s}) \approx \prod_{q=1}^{Q} \sum_{k=1}^{K^q} p_k^q \cdot \mathcal{N}^q(\tau^q(\mathbf{s}); \mu_k^q, (\sigma_k^q)^2) \tag{7}$$



(a) Conventional SRP : Side view (b) Conventional SRP : Top view (c) Proposed Approach : Side view (d) $\mathcal{G}_s + \mathcal{G}_n$ Estimation

Figure 2: *The graphs in (a) and (b) exemplify the SRP approach for a frame with a single speaker. Here the dominant GCC peak generates multiple local maxima. The graph in (c) shows that the proposed approach cancels the secondary local maxima using the information provided by the other GCC functions. Finally, the graph in (d) shows the evaluation method used in Section 6.*

4 Acoustic Source Extraction/Localization

The source location pdf $p(\mathbf{s})$ is a prior multimodal distribution, where each "mode" represents a potential acoustic source. The key idea behind the extraction/localization of each source consists of calculating the restriction of $p(\mathbf{s})$ on that region of the location space where the source is dominant.

4.1 Mapping the Dominance from the TDOA-Space to the Location Space

The notion of *interval of dominance* in the TDOA space maps to the *region of dominance* notion in the location space through the location-TDOA function, given by (6). Formally, Let $\mathcal{A}$ be an acoustic source and let us assume that we can extract, for each microphone pair $q \in \{1, \ldots, Q\}$, the interval of dominance $I^q_{\mathcal{A}}$ where $\mathcal{A}$ is dominant. Then the corresponding region of dominance $\mathcal{D}_{\mathcal{A}}$ in the location space is defined as follows:

$$\mathcal{D}_{\mathcal{A}} = \{\mathbf{s} \in \text{Space} \mid \forall q \in \{1, \ldots, Q\} : \tau^q(\mathbf{s}) \in I^q_{\mathcal{A}}\} \tag{8}$$

The corresponding indicator function $\mathbb{1}_{\mathcal{D}_{\mathcal{A}}}$ of $\mathcal{D}_{\mathcal{A}}$ is given by $\mathbb{1}_{\mathcal{D}_{\mathcal{A}}}(\mathbf{s}) = \prod_{q=1}^{Q} \mathbb{1}_{I^q_{\mathcal{A}}}(\mathbf{s})$ with $\mathbb{1}_{I^q_{\mathcal{A}}}(\mathbf{s}) = \left\{ \begin{array}{cc} 1 & \text{if } \tau^q(\mathbf{s}) \in I^q_{\mathcal{A}} \\ 0 & \text{otherwise.} \end{array} \right.$

The extraction of $\mathcal{A}$ is subsequently obtained by calculating the restriction $p_{\mathcal{A}}(\mathbf{s})$ of $p(\mathbf{s})$ on $\mathcal{D}_{\mathcal{A}}$. This is achieved by calculating the product of $p(\mathbf{s})$ and $\mathbb{1}_{\mathcal{D}_{\mathcal{A}}}$, which simplifies to:

$$p_{\mathcal{A}}(\mathbf{s}) \propto \prod_{q=1}^{Q} \left(\mathbb{1}_{I^q_{\mathcal{A}}} \cdot \sum_{k=1}^{K^q} p^q_k \cdot \mathcal{N}^q(\tau^q(\mathbf{s}), \mu^q_k, (\sigma^q_k)^2) \right). \tag{9}$$

Knowing that on each interval $I^q_{\mathcal{A}}$, the GM contribution mainly comes from the component $k(\mathcal{A})$ associated to this interval, we can further simplify $p_{\mathcal{A}}(\mathbf{s})$ to obtain a more practical approximation:

$$p_{\mathcal{A}}(\mathbf{s}) \approx \prod_{q=1}^{Q} p^q_{k(\mathcal{A})} \cdot \mathcal{N}^q(\tau^q(\mathbf{s}), \mu^q_{k(\mathcal{A})}, (\sigma^q_{k(\mathcal{A})})^2). \tag{10}$$

The acoustic source location is then obtained using numerical optimization algorithms [23]. Theoretically, $p_{\mathcal{A}}(\mathbf{s})$ is not a convex function. However, in practice this function has a sharp peak and very flat tails (Figure 2-c). Furthermore, any initial guess from $\mathcal{D}_{\mathcal{A}}$ ensures the convergence to the optimal location. In the following section, we describe how the region of dominance $\mathcal{D}_{\mathcal{A}}$ can be determined.

4.2 Extraction of Regions with High Likelihood

Knowing that the $\{I^q_k\}_{k=1}^{K^q}$, $q = 1, \ldots, Q$, form a "partition" of the TDOA space, and given the definition of the region of dominance $\mathcal{D}_{\mathcal{A}}$ in (8), we can conclude that all the locations in $\mathcal{D}_{\mathcal{A}}$ will map to the same combination of intervals. Hence, the extraction of $\mathcal{D}_{\mathcal{A}}$ can be reduced to finding a single location from it. Formally, this can be done using a *coarse grid* (15° to 30° or 50 to 150 cm). The grid resolution is chosen such that at least one location falls into $\mathcal{D}_{\mathcal{A}}$. Then, for each location $\mathbf{s}_0$ in this grid, and for each microphone pair q, 1) the associated interval of dominance $I^q_{\mathbf{s}_0}$ is extracted such that the corresponding Gaussian component $\mathcal{N}^q(\tau^q(\mathbf{s}), \mu^q_k, (\sigma^q_k)^2)$ maximizes $\tau^q(\mathbf{s_0})$, and 2) the cumulative distribution $\mathcal{C}(\mathcal{D}_{\mathbf{s}_0})$ over the associated region of dominance $\mathcal{D}_{\mathbf{s}_0}$ (given by (8)) is calculated according to:

$$\begin{aligned} \mathcal{C}(\mathcal{D}_{\mathbf{s}_0}) &= \int_{D_{\mathbf{s}_0}} p(\mathbf{s}) \cdot d\mathbf{s} = \prod_{q=1}^{Q} \int_{I^q_{\mathbf{s}_0}} p(\tau^q) \cdot d\tau^q && (11) \\ &\approx \prod_{q=1}^{Q} \sum_{\{\tau^q \in I^q_{\mathbf{s}_0}\}} \mathcal{R}_q(\tau^q). && (12) \end{aligned}$$

The region of dominance $\mathcal{D}_{\mathcal{A}}$ is extracted as the one with the highest cumulative distribution. The restriction of $p(\mathbf{s})$ on $\mathcal{D}_{\mathcal{A}}$ is then calculated according to (10), and the corresponding $\mathbf{s}_0$ ($\in \mathcal{D}_{\mathcal{A}}$) is used as an initial guess to run the numerical optimization. The experiments reported below used the gradient descent algorithm to perform this task [23].

4.3 Acoustic Source Detection

The proposed method extracts the source location as the one with the highest likelihood but it does not consider whether this location has been generated by a dominant GCC peak or by a consistent set of peaks from different GCCs. This problem becomes more difficult in the multiple speaker scenario, as the number of current sources is unknown. Furthermore, when only few microphones are available a dominant GCC peak may generate many local maxima.

Finding the optimal location requires the optimization of $p_{\mathcal{A}}(\mathbf{s})$, which is equivalent to the minimization of the Maximum Likelihood (ML) criterion [4, 20] given by $\arg\min_{\mathbf{s}} \epsilon(\mathbf{s})$ with

$$\epsilon(\mathbf{s}) = \sum_{q=1}^{Q} \frac{1}{(\sigma^q_{k(\mathcal{A})})^2} \cdot \left[\tau^q(\mathbf{s}) - \mu^q_{k(\mathcal{A})}\right]^2 \tag{13}$$

The error function $\epsilon(\mathbf{s})$ characterizes the consistency of the optimal location $\mathbf{s}_{opt}$. More precisely, a high value of $\epsilon(\mathbf{s}_{opt})$ means that $\mathbf{s}_{opt}$ has been generated by a dominant peak rather than a combination of GCC peaks and vice versa. Hence, $\epsilon(\mathbf{s})$ can be used as a validation criterion. Formally, s_{opt} is assumed to be generated by an actual source if $\frac{1}{Q} \cdot \epsilon(\mathbf{s}_{opt}) \leq \Gamma$, where Γ is a predefined threshold.

Algorithm 1 : Multiple Speaker Localization Algorithm

```
Let $N_{max}$ be the maximum number of speakers.
Let $\mathcal{G}$ be the coarse grid.
1. Estimate the GM approximations and define $p(\mathbf{s})$.
2. Calculate $\mathcal{C}(\mathcal{D}_i)$ for each location $\mathbf{s}_i \in \mathcal{G}$.
for $n = 1 : N_{max}$ do
    3. Find $\mathcal{D}_n^{max}$ which maximizes $\mathcal{C}(\mathcal{D}_i)$.
    4. Calculate $p_{\mathbf{s}_n^{max}}(\mathbf{s})$.
    5. Run the optimization of $p_{\mathbf{s}_n^{max}}(\mathbf{s})$ to estimate $\mathbf{s}_n^{opt}$.
    if $p(\mathbf{s}_n^{opt}) > \Gamma$ then
        6. Add $\mathbf{s}_n^{opt}$ to the set of speakers $S$.
    end if
    7. Remove all $\mathcal{D}_i$ for which $\mathbf{s}_n^{opt} \in \mathcal{D}_i$.
    8. Go to step 1.
end for
9. Return the set of speakers $S$.
```

Table 1 : Single Speaker Localization Results

Approaches	seq01-1p-0000				seq03-1p-0100				seq11-1p-0100			
	p_s	$\sigma_{s,\theta}$	$\sigma_{s,\phi}$	t	p_s	$\sigma_{s,\theta}$	$\sigma_{s,\phi}$	t	p_s	$\sigma_{s,\theta}$	$\sigma_{s,\phi}$	t
MCCC	**0.28**	3.50	15.56	26.25	**0.66**	3.27	9.75	26.61	0.68	**3.17**	9.65	26.56
SRP	0.27	**3.02**	15.16	1.99	0.65	2.71	9.30	2.04	0.70	3.37	10.07	1.98
pSRP	0.27	3.73	13.38	**0.69**	0.63	3.39	8.72	**0.69**	**0.72**	4.61	10.51	**0.69**
PA	**0.28**	3.90	**8.40**	0.79	0.62	**2.52**	**4.68**	0.78	**0.72**	3.69	**6.76**	0.75

Table 2 : (%) of frames with N correct simultaneous estimates

# of Speakers	seq18-2p-0101		seq40-3p-0111		seq37-3p-0001	
	PA	pSRP	PA	pSRP	PA	pSRP
1S	23.53	**66.69**	21.41	**59.48**	25.16	**65.02**
2S	**63.05**	23.59	**44.95**	23.04	**41.02**	15.36
3S	—	—	**7.47**	1.72	**7.51**	0.96

Table 3 : Multiple Speaker Localization Results

	seq18-2p-0101		seq40-3p-0111		seq37-3p-0001	
	PA	pSRP	PA	pSRP	PA	pSRP
$\sigma_{s,\theta}$	**2.00**	2.79	**1.90**	4.75	**2.66**	5.0
$\sigma_{s,\phi}$	**4.51**	9.76	**7.93**	13.40	**9.47**	13.43
p_s	**0.60**	0.45	**0.50**	0.45	**0.54**	0.46

5 Multiple Speaker Localization Algorithm

The proposed acoustic source localization approach can be easily extended to the multiple speaker case. Algorithm 1 presents one possible extension using an iterative approach. The algorithm is iterative in order to avoid re-detecting the same region of dominance in the case where more than one location $\mathbf{s}_i$ belongs to it. This idea is implemented by successively discarding all the regions $\mathcal{D}_i$ that contain the optimal location $\mathbf{s}_n^{opt}$ from the previous iteration (step 7). In the case where N_{max} is unknown, it can be simply overestimated.

6 Experiments and Results

We evaluate the proposed approach using the AV16.3 corpus [24], where human speakers have been recorded in a smart meeting room (approximately 30m^2 in size) with a 20cm 8-channel circular microphone array. In this work, only 4 of the microphones are used to highlight the ability of the proposed approach to cancel local maxima. The sampling rate is 16 kHz and the real mouth position is known with an error $\leq$ 5cm [24]. The AV16.3 corpus has a variety of scenarios, such as stationary or quickly moving speakers, varying number of simultaneous speakers, etc. In the experiments reported below, the signal was divided into frames of 512 samples (32ms); the GCCs were calculated using PHAT [19] weighting; and a voice activity detector was used in order to suppress silence frames.

We use a new evaluation method, based on a 2-components GM fitting of the estimates error, noted $\mathcal{G}_s + \mathcal{G}_n$ (see Figure 2-d), which aims at modeling the ”noise+source(s)“ estimates. Due to the far-field assumption, in which the range is ignored, the localization task is performed in the entire 3D space but the results are limited to the direction of arrival (DOA). More precisely, the results are reported in terms of the azimuth and elevation precision given by the standard deviations $\sigma_{s,\theta}$ and $\sigma_{s,\phi}$ of the Gaussian $\mathcal{G}_s$ (see Figure 2-d), respectively, in addition to the percentage of correct estimates p_s which is nothing but the mixture weight of $\mathcal{G}_s$. We also report the real-time factor t on a standard Intel i7-2600K CPU clocked at 3.4GHz. In the multiple speaker scenario, we report the percentage of frames with correct number of simultaneous speakers. p_s in this case represents the percentage of correct estimates for all speakers. The detection thresholds of the proposed approach (PA) and the pSRP method are chosen such that the resulting false alarm rate is the same for both methods and equal to 0.3.

Table 1 presents the performance of the proposed approach on single source sequences, and compares it to two well-known approaches, namely the SRP [9] and the MCCC [5]. In addition to this, results for the probabilistic SRP (pSRP) [18] are shown. Note that in these experiments the detection approach from Section. 4.3 was not used. Hence, N_{max} was set to 1. The coarse grid resolution is $20^\circ \times 30^\circ \times 50cm$ for the azimuth, elevation and range, respectively, whereas the resolution of the SRP and MCCC grid is $1^\circ \times 1^\circ \times 10cm$. The merits of applying the proposed approach to multiple speaker localization are shown in Tables 2 and 3, which present results for sequences with a varying number of simultaneous speakers (between zero and three). In these experiments $N_{max} = 5$.

The results in Table 1 show that the proposed approach performs better than the other approaches. More precisely, the azimuth precision $\sigma_{s,\theta}$ and the percentage p_s of correct estimates are comparable, whereas the elevation precision $\sigma_{s,\phi}$ is highly improved. This is mainly due to the reduction of the elevation variance around the peaks, which become sharper (see e.g. Figure 2-c). Regarding the computation cost, we can conclude that the proposed approach is comparable to the Probabilistic SRP but three times faster than the classical SRP. The MCCC approach however is very slow due to the calculation of the correlation matrix determinant for all locations at each frame. Regarding the multiple speaker scenarios in Tables 2 and 3, we can see that the proposed approach deals better with the problem of local maxima resulting from the multivalued TDOA-location function, where a dominant GCC peak may generate many peaks in the location space (e.g. Figure 2). This improvement appears in the increased number of correct estimates p_s, and the percentage of frames with correct simultaneous estimates. The latter shows that the proposed approach leads to a lower number of frames with one correct estimate (compared to the pSRP), whereas the percentage of frames with two or more correct estimates increases. This shows the capacity of the proposed approach to detect the sources in the location space. Regarding the localization precision, we can once again conclude that the proposed approach gives more accurate elevation estimates and comparable azimuth.

7 CONCLUSION

We have proposed a probabilistic framework to the multiple speaker localization problem. This approach presents a different method to combine the GCC functions in order to increase the localization precision, especially when only few microphones are available. The proposed method was also shown to be more effective in cancelling the local maxima resulting from the multivalued TDOA-location function, which highly affects the multiple speaker detection performance. The future work will focus on developing a better detection approach to improve the noise/source decision.

8 References

[1] R. O. Schmid, "A new approach to geometry of range difference location," *IEEE Trans. Aerospace and Electronic Systems*, vol. 8, no. 6, pp. 821–835, 1972.

[2] H. C. Schau and A. Z. Robinson, "Passive source localization employing intersecting spherical surfaces from time-of-arrival differences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 8, pp. 1223 – 1225, Aug. 1987.

[3] J. O. Smith and J. S. Abel, "Closed-form least-squares source location estimation from range-difference measurements," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 12, pp. 1661 – 1669, Dec. 1987.

[4] M. S. Brandstein, J. E. Adcock, and H. F. Silverman, "A closed-form location estimator for use with room environment microphone arrays," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 7, no. 1, pp. 45–50, Jan. 1997.

[5] J. Chen, J. Benesty, and Y. Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 11, no. 6, pp. 549–557, 2003.

[6] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 384–391, 2000.

[7] J. Dmochowski, J. Benesty, and S. Affes, "Direction of arrival estimation using the parameterized spatial correlation matrix," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1327 –1339, May 2007.

[8] J. Dmochowski, J. Benesty, and S. Affes, "The generalization of narrowband localization methods to broadband environments via parametrization of the spatial correlation matrix," in *Proc. EUSIPCO*, Sep. 2007, pp. 763–767.

[9] J. H. DiBiase, *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*, Ph.D. thesis, Brown University, 2000.

[10] J. P. Dmochowski, J. Benesty, and S. Affes, "Fast steered response power source localization using inverse mapping of relative delays," in *Proc. ICASSP*, 2008, pp. 289–292.

[11] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction(SRC) on a large-aperture microphone array," in *Proc. ICASSP*, 2007, pp. 121–124.

[12] H. Do and H. F. Silverman, "A fast microphone array SRP-PHAT source location implementation using coarse-to-fine region contraction(CFRC)," in *Proc. WASPAA*, 2007, pp. 295 –298.

[13] H. Do and H.F. Silverman, "A method for locating multiple sources from a frame of a large-aperture microphone array data without tracking," in *Proc. ICASSP*, Apr. 2008, pp. 301 –304.

[14] M. Nilesh and M. Rainer, "A scalable framework for multiple speaker localization and tracking," in *Proc. IWAENC*, 2008.

[15] H. Do and H. F. Silverman, "SRP-PHAT methods of locating simultaneous multiple talkers using a frame of microphone array data," in *Proc. ICASSP*, 2010, pp. 125–128.

[16] G. Lathoud and I. A. McCowan, "A sector-based approach for localization of multiple speakers with microphone arrays," in *Proc. SAPA Workshop*, Oct. 2004.

[17] G. Lathoud and M. Magimai.-Doss, "A sector-based, frequency-domain approach to detection and localization of multiple speakers," in *Proc. ICASSP*, Mar. 2005, vol. 3, pp. 265 –268.

[18] Youssef Oualil, Mathew Magimai.-Doss, Friedrich Faubel, and Dietrich Klakow, "Joint detection and localization of multiple speakers using a probabilistic interpretation of the steered response power," in *Proc. SAPA Workshop*, 2012.

[19] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, 1976.

[20] U. Klee, T. Gehrig, and J. McDonough, "Kalman filters for time delay of arrival-based source localization," *EURASIP Journal on Applied Signal Processing*, pp. 167–167, 2006.

[21] A. Levy, S. Gannot, and A. P. Habets, "Multiple-hypothesis extended particle filter for acoustic source localization in reverberant environments," *IEEE Trans. Acoust., Speech, Signal Process.*, 2010.

[22] Y. Oualil, F. Faubel, M. Magimai.-Doss, and D. Klakow, "A TDOA Gaussian mixture model for improving acoustic source tracking," in *Proc. EUSIPCO*, 2012, pp. 1339 –1343.

[23] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, 2nd edition, 2006.

[24] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "AV16.3: An audio-visual corpus for speaker localization and tracking," in *Proc. MLMI 04 Workshop*, May 2006, pp. 182–195.