A FAST CUMULATIVE STEERED RESPONSE POWER FOR MULTIPLE SPEAKER DETECTION AND LOCALIZATION

Youssef Oualil, Friedrich Faubel, Dietrich Klakow

Spoken Language Systems, Saarland University, Saarbrücken, Germany
youssef.oualil@lsv.uni-saarland.de

ABSTRACT

This paper presents a novel approach for detecting and localizing multiple speakers using a microphone array. In this framework, the classical Steered Response Power (SRP) technique is combined with a novel two-step search strategy to reduce the computation cost. The approach taken here performs the localization by 1) using the spatial information provided by each Generalized Cross Correlation (GCC) function to reduce the search space to a few subspaces that are likely to contain a source. From these, the most likely region is extracted as the subspace that maximizes the Cumulative SRP. Then, 2) the optimal source location is estimated using the classical search approach in the reduced space. The source/noise detection is further improved using an unsupervised Bayesian classifier. Experiments on the AV16.3 corpus show that the proposed method is approximately 47 times faster than the classical SRP, without any noticeable degradation of the localization performance.

Index Terms— Steered response power, multiple speaker localization, microphone arrays.

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] and single-step approaches, which aim at inferring the source location directly from the signals, such as multi-channel cross correlation (MCCC) [3], adaptive eigenvalue decomposition [4], and the well-known SRP based techniques (e.g. [5, 6, 7]). Although the SRP approach is robust and reliable, it is computationally expensive as it requires a fine discretization of the space for a better localization precision. Dmochowski et *al.* [6] proposed to overcome this issue by reducing the search space through inverse mapping of the Time Difference Of Arrival (TDOA), whereas Do et *al.* [7] used iterative reduction search strategies to estimate the optimal source location. Other improvements of the SRP made use of spatial averaging techniques. This idea was investigated in [8] using a sector-based approach. A similar method was proposed in [9] based on mapping compact volumes in the location space to closed intervals in the TDOA space.

Following a line of thought similar to [8, 9], we propose a novel framework. It combines the advantages of search space reduction strategies [6, 7] and spatial averaging techniques [8] by i) using the spatial information introduced by each microphone pair GCC function to partition the TDOA space into a set of intervals of dominance (Section 3.1), ii) using all the resulting partitions and the array geometry to reduce the location space to few regions, which are likely to contain a source (Section 3.2). This is followed by iii) extracting the speaker subspace as the region which maximizes the cumulative SRP (Section 3.3), and iv) performing the classical SRP search in the reduced space.

In doing so, the proposed approach drastically decreases the computation cost by reducing the search space. On top of that, it improves the multiple speaker localization performance through use of the cumulative SRP. The extension to multiple speakers is straight-forward (Section 3.4). Finally, the effectiveness of the proposed method is demonstrated by means of an experimental study in Section 5, including comparisons to the conventional SRP, and MCCC approaches on a single speaker localization task, and to the probabilistic SRP [10] on a multiple speaker localization task.

2. THE CONVENTIONAL SRP APPROACH

The arrival of sound waves at a microphone array introduces TDOAs between the individual microphone pairs. This TDOA depends on the source location $\mathbf{s}$ as well as the positions $\mathbf{m}_h$, $h = 1, \ldots, M$, of the microphones where M denotes the number of microphones. More precisely, the TDOA introduced at the microphone pair $q = \{\mathbf{m}_g, \mathbf{m}_h\}$ is given by

$$\tau_q(\mathbf{s}) = (\|\mathbf{s} - \mathbf{m}_h\| - \|\mathbf{s} - \mathbf{m}_g\|) \cdot c^{-1} \qquad (1)$$

where c denotes the speed of sound in the air. The SRP approach uses these TDOAs to construct a spatial filter (delay-and-sum beamformer) which scans all possible source locations. The speaker position is subsequently extracted as that position where the signal energy is maximized. These steps can be implemented efficiently using the GCC function [5].

2.1. Generalized Cross Correlation

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

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

where $S_{g/h}(\omega)$ denotes the short-time Fourier transforms of $s_{g/h}(t)$ and where $\psi(\omega)$ denotes a pre-filter. A common choice of $\psi(\omega)$ is the phase transform (PHAT) weighting [11].

2.2. SRP-based Single Speaker Localization

The steered response power returned from a particular location $\mathbf{s}$ can be calculated as [5]:

$$SRP(\mathbf{s}) = 4\pi \sum_{q=1}^{Q} \mathcal{R}_q(\tau_q(\mathbf{s})) + \mathcal{K} \qquad (3)$$

where Q denotes the number of microphone pairs. $\mathcal{K}$ is a constant introduced by the auto-correlation of each microphone (see [5] for more details). Therefore, $\mathcal{K}$ is ignored in the rest of the paper. Once the SRP has been calculated for each position $\mathbf{s}$, the source location estimate $\hat{\mathbf{s}}$ is determined according to [5]:

$$\hat{\mathbf{s}} = \underset{\mathbf{s}}{\operatorname{argmax}}\, SRP(\mathbf{s}). \qquad (4)$$

Scanning all possible source locations on a discrete grid over the 3-D/2-D space is computationally expensive. Section 3 introduces a novel approach to overcome this problem.

3. PROPOSED APPROACH

The GCC function has been widely used to estimate the TDOA introduced by a source at the microphone pairs. Under ideal conditions – more precisely, in noise-free/reverberation-free environments and under the assumption of signals originated by *point* sources – the GCC function is proportional to a shifted delta function, where the shift is given by the TDOA generated by the source at the microphone pair. In practice, however, the presence of noise and reverberation introduce secondary peaks. Furthermore, diffuse sound sources may flatten the peaks, causing high GCC values to span over TDOA intervals, which map to connected regions instead of point locations. Hence, we propose to characterize each acoustic event in the room by an interval of TDOA values, which is centered at a GCC peak. In particular, we assume that all the GCC values in this interval were generated by the same source.

3.1. Acoustic Dominance-based TDOA Space Partition

In contrast to classical TDOA-based source localization approaches [1, 2], which obtain the source location by mapping GCC peaks to the location space, we propose to associate each acoustic event with the TDOA interval where the source is assumed to be dominant. The reseulting intervals are subsequently called the *intervals of dominance*. An acoustic event can be generated by actual sources (speech, coughs, laughs, etc.) or by noise sources (projector, door slams, etc.). Multipaths reflections from reverberation are considered acoustic events of "virtual" noise sources.

Formally, let K_q be the number of GCC peaks of the q-th microphone pair at time t and let $\{\tau_q^1, \ldots, \tau_q^{K_q}\}$ be the corresponding TDOA values. For ease of notation, the time index t is dropped in the rest of the paper. Then the TDOA observation space $[-\tau_q^{max}, \tau_q^{max}]$ with $\tau_q^{max} = \|\mathbf{m}_h - \mathbf{m}_g\| \cdot c^{-1}$ can be expressed as the union of the intervals of dominance I_q^k, $k = 1, \ldots, K_q$:

$$]-\tau_q^{max}, \tau_q^{max}] = \bigcup_{k=1}^{K_q} I_q^k \qquad (5)$$

The k-th interval of dominance I_q^k associated to the k-th peak/acoustic event is given by

$$I_q^1 = \left[-\tau_q^{max}, \tau_q^{1,max}\right] \text{ and } \; I_q^k = \left]\tau_q^{k,min}, \tau_q^{k,max}\right] \qquad (6)$$

Here, $\tau_q^{k,min}$ and $\tau_q^{k,max}$ are given by

$$\tau_q^{k,min} = \max\left\{\tau_q \mid \tau_q \leq \tau_q^k, \partial\mathcal{R}_q(\tau_q) = 0\right\} \qquad (7)$$

$$\tau_q^{k,max} = \min\left\{\tau_q \mid \tau_q \geq \tau_q^k, \partial\mathcal{R}_q(\tau_q) = 0\right\} \qquad (8)$$

where τ_q^k is the TDOA corresponding to the k-th GCC peak and where $\partial\mathcal{R}_q$ denotes the first derivative of $\mathcal{R}_q$. In words, $\tau_q^{k,min}$ and $\tau_q^{k,max}$ represent the left and right feet of the k-th peak τ_q^k of the GCC function (see example Fig. 1-b). The intervals of dominance $\{I_q^k\}_{k=1}^{K_q}$ are mutually disjoint. Therefore, these intervals map to mutually disjoint sets of locations. Furthermore, mapping each microphone pair TDOA space partition leads to a new partition of the location space. This important property is very useful to extract the location subspaces which are likely to contain a source (Section 3.2).

3.2. From the TDOA Space to the Location Space

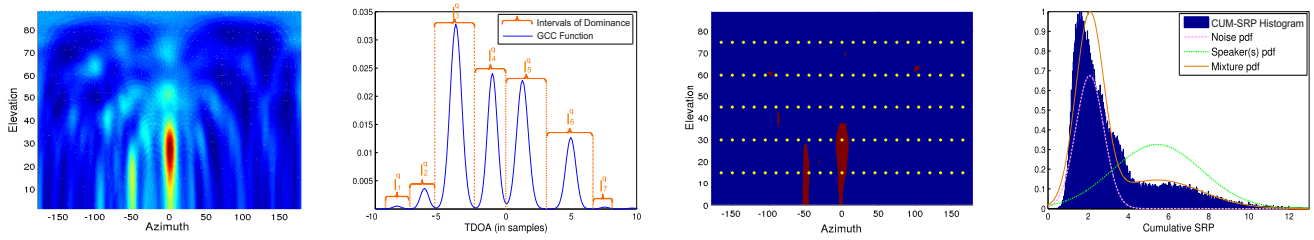
The search space reduction is obtained by mapping all TDOA space partitions to the location space, followed by the intersections of the resulting location space partitions. Considering only non-empty intersections yields a few likely regions of the location space.

Formally, let $\mathcal{I}_q = \{I_q^k\}_{k=1}^{K_q}$ be the TDOA space partition of the q-th microphone pair, and let $\mathcal{S}$ denote the location space. Then each interval I_q^k maps to a subspace of locations given by

$$\mathcal{S}_q^k = \{\mathbf{s} \in \mathcal{S} \mid \tau_q(\mathbf{s}) \in I_q^k\} \qquad (9)$$

Mapping all the intervals $\{I_q^k\}_{k=1}^{K_q}$ leads to a partitioning $\mathcal{S}_q = \{\mathcal{S}_q^k\}_{k=1}^{K_q}$ of the location space $\mathcal{S}$, with

$$\mathcal{S} = \bigcup_{k=1}^{K_q} \mathcal{S}_q^k \qquad (10)$$



(a) Conventional SRP : Top view (b) GCC-based TDOA space partition (c) Search space reduction (d) Noise/speaker classification

Fig. 1: Figure 2: *The graphs in (a) exemplifies the SRP approach for a frame with two speakers. The figure (b) illustrates the GCC-based TDOA space partition to intervals of dominance. The graph in (c) presents the subspaces of dominance resulting from mapping all the TDOA spaces partitions. Finally, the graph in (d) illustrates the classification approach used in Section 4.*

The localization of an acoustic source $\mathcal{A}$ requires the extraction of the intervals of dominance $\{\mathcal{I}_q^{\mathcal{A}}\}_{q=1}^{Q}$ where $\mathcal{A}$ is dominant. Each of these intervals is then mapped to a location subspace $\mathcal{S}_q^{\mathcal{A}}$ according to eq (9). The *region of dominance* $\mathcal{S}^{\mathcal{A}}$ associated with the source $\mathcal{A}$ is defined as follows :

$$\mathcal{S}^{\mathcal{A}} = \bigcap_{q=1}^{Q} \mathcal{S}_q^{\mathcal{A}} = \{\mathbf{s} \in \mathcal{S} \mid \forall q \in \{1,\dots,Q\} : \tau_q(\mathbf{s}) \in I_q^{\mathcal{A}}\} \quad (11)$$

Given eq (11), we can conclude that the acoustic source localization problem can be reduced to extracting the space regions of dominance, which are expressed as intersections of $\{\mathcal{S}_q^k\}_{k=1}^{K_q}$, $q = 1,\dots,Q$. Theoretically, the number of all possible intersections is large and equal to $\prod_{q=1}^{Q} K_q$. In practice however, most of these intersections are empty. This is due to the physical constraints introduced by the microphone pairs. More precisely, if $\mathcal{S}^{\mathcal{A},P}$ represents the sub-intersection of the first P microphone pairs ($P \leq Q$) then the volume of $\mathcal{S}^{\mathcal{A},P}$ decreases when P is increased. For all true sources, it can be expected for a given number P that

$$\forall q \in \{P+1,\dots,Q\}, \exists \mathcal{S}_q^{k_p} \in \mathcal{S}_q : \mathcal{S}^{\mathcal{A},P} \subset \mathcal{S}_q^{k_p} \quad (12)$$

The intersection of $\mathcal{S}^{\mathcal{A},P}$ with the remaining sets of the partition $\mathcal{S}_q$ are mostly empty (when P is large enough). This drastically decreases the number of intersections that need to be performed. The experiments conducted in this paper have shown that such a property occurs when $P \geq 4$.

The extraction of all intersections is analytically intractable. Hence, we propose an alternative iterative solution (Algorithm 1). This is done using eq (11), which shows that each region of dominance $\mathcal{S}^d$ is defined by the set of intervals of dominance which map to it. Therefore, the extraction of dominant subspaces reduces to finding all possible combinations of the intervals of dominance. Formally, this can be done using a *coarse grid* (15° to 30° or 50 to 100 cm). The grid resolution is chosen such that at least one location falls into each $\mathcal{S}^d$. Then, for each location $\mathbf{s}_0$ in this grid (dots in Fig. 1-c), the associated intervals of dominance $I_q^{\mathbf{s}_0}$ are extracted such that $\tau_q(\mathbf{s}_0) \in I_q^{\mathbf{s}_0}$.

Algorithm 1 : Extraction of the Subspaces of Dominance

Let $\mathcal{G}$ be the coarse grid.
Let $\mathcal{D}_{\mathcal{S}}$ be the set of the subspaces of dominance.
$\forall q \in \{1,\dots,Q\}$ calculate the TDOA partition $\{I_q^k\}_{k=1}^{K_q}$
for each $\mathbf{s}_0 \in \mathcal{G}$ **do**
  $\forall q \in \{1,\dots,Q\}$ find $k_{\mathbf{s}_0,q}$ such that $\tau_q(\mathbf{s}_0) \in I_q^{k_{\mathbf{s}_0,q}}$
  if $\{\mathcal{S}_q^{k_{\mathbf{s}_0,q}}\}_{q=1}^{Q} \notin \mathcal{D}_{\mathcal{S}}$ **then**
    Add $\{\mathcal{S}_q^{k_{\mathbf{s}_0,q}}\}_{q=1}^{Q}$ to $\mathcal{D}_{\mathcal{S}}$.
  end if
end for

3.3. The Cumulative SRP

The space reduction approach is based on extracting those subspaces where each acoustic event is dominant. Hence, in the absence of spacial aliasing, we can assume that the contribution of other sources is negligible in each of the subspaces. As a consequence, all the signal power coming from that region is assumed to be generated by the same acoustic source. Formally, let $\mathcal{A}$ be an acoustic source. The SRP$^{\mathcal{A}}$ associated with $\mathcal{A}$ is given by the restriction of eq (3) on the subspace of dominance $\mathcal{S}^{\mathcal{A}}$. That is

$$SRP^{\mathcal{A}}(\mathbf{s}) = SRP(\mathbf{s}) \cdot \mathbb{1}_{\mathcal{S}^{\mathcal{A}}}(\mathbf{s}) \quad (13)$$

where $\mathbb{1}_{\mathcal{S}^{\mathcal{A}}}(\mathbf{s})$ is the indicator function, which is 1 if $\mathbf{s} \in \mathcal{S}^{\mathcal{A}}$ and 0 otherwise. Given the definition in eq (11), we can further simplify (13) to

$$SRP^{\mathcal{A}}(\mathbf{s}) \quad \propto \quad \sum_{q=1}^{Q} \mathcal{R}_q(\tau_q(\mathbf{s})) \cdot \prod_{q=1}^{Q} \mathbb{1}_{I_q^{\mathcal{A}}}(\mathbf{s}) \quad (14)$$

Now, we define the cumulative SRP (C-SRP) of the source $\mathcal{A}$, denoted by $SRP^c(\mathcal{A})$, as the sum of steered power originating from all locations $\mathbf{s}$ in the region of dominance $\mathcal{S}^{\mathcal{A}}$. More precisely, $SRP^c(\mathcal{A})$ is calculated according to

$$SRP^c(\mathcal{A}) = \int_{\mathcal{S}} SRP^{\mathcal{A}}(\mathbf{s}) \cdot \mathrm{d}\mathbf{s} = \int_{\mathcal{S}^{\mathcal{A}}} SRP(\mathbf{s}) \cdot \mathrm{d}\mathbf{s} \quad (15)$$

$$\approx \sum_{q=1}^{Q} \int_{I_q^{\mathcal{A}}} \mathcal{R}_q(\tau_q) \cdot \mathrm{d}\tau_q \approx \sum_{q=1}^{Q} \sum_{\tau_q \in I_q^{\mathcal{A}}} \mathcal{R}_q(\tau_q) \quad (16)$$

Table 1 : Single Speaker Localization Results

Approaches	seq01-1p-0000				seq02-1p-0100				seq03-1p-0100			
	d_r	$\sigma_{s,\theta}$	$\sigma_{s,\phi}$	t	d_r	$\sigma_{s,\theta}$	$\sigma_{s,\phi}$	t	d_r	$\sigma_{s,\theta}$	$\sigma_{s,\phi}$	t
MCCC	31.81	**1.87**	11.64	∞	77.85	1.81	8.54	∞	69.67	1.49	**5.42**	∞
SRP	**33.79**	2.09	13.57	55.58	**78.64**	1.74	9.67	55.77	**69.88**	**1.46**	6.31	55.74
PA	30.08	1.90	**10.83**	**1.16**	76.52	**1.71**	**7.92**	**1.17**	69.41	1.47	6.76	**1.16**

Table 2 : Multiple Speaker Detection Rate d_r (%)

	seq18-2p-0101		seq40-3p-0111		seq37-3p	
	PA	pSRP	PA	pSRP	PA	pSRP
S 1	**54.19**	51.72	**27.28**	23.79	31.25	**32.59**
S 2	45.78	**45.92**	**32.25**	25.72	**59.65**	28.52
S 3	—	—	47.44	**56.32**	**40.29**	9.74

Table 3 : Multiple Speaker Localization Results

	seq18-2p-0101		seq40-3p		seq37-3p	
	PA	pSRP	PA	pSRP	PA	pSRP
$\sigma_{s,\theta}$	**1.78**	2.22	2.67	**1.95**	**2.44**	3.0
$\sigma_{s,\phi}$	**4.50**	8.93	8.92	**6.59**	8.25	**8.20**
p_s	**0.87**	0.86	**0.77**	0.74	**0.79**	0.53

The region of dominance $\mathcal{S}^{\mathcal{A}}$ is extracted as the one with the highest cumulative SRP. Then, the optimal location estimate $\mathbf{s}_{opt}^{\mathcal{A}}$ is obtained using the classical approach in the reduced space $\mathcal{S}^{\mathcal{A}}$. This is done by maximizing the SRP output on a sub-grid of locations, centered on the initial location $\mathbf{s}_0 (\in \mathcal{S}^{\mathcal{A}})$ given by the coarse grid (from Algorithm 1). All the sub-grids are calculated offline.

3.4. Multiple Speaker Localization Algorithm

The proposed acoustic source localization approach can be easily extended to the multiple speaker case. Algorithm 2 presents one possible extension using an iterative approach. The algorithm is iterative in order to overcome the one-to-many aspect of the TDOA-location mapping (eq (1)), which causes each interval $\mathcal{I}_q^k$ to map to more than one subspace. This idea is implemented by successively zeroing the restriction of the GCC function on $\mathcal{I}_q^{\mathbf{s}_n^{opt}}$ (step 6). The sub-grid used in the second search step (step 4) is calculated offline by associating each location $\mathbf{s}_0$ in the coarse grid $\mathcal{G}$ to a small grid centered on $\mathbf{s}_0$. In the case where N_{max} is unknown, it can be simply overestimated.

Algorithm 2 : Multiple Speaker Localization Algorithm

Let N_{max} be the maximum number of speakers.
Extract the set of regions of dominance $\mathcal{D}_{\mathcal{S}}$ (Algorithm 1)
for $n = 1 : N_{max}$ **do**
1. $\forall \mathcal{S} \in \mathcal{D}_{\mathcal{S}}$: calculate $\mathcal{C}(\mathcal{S}) = SRP^c(\mathcal{S})$
2. Find $\mathcal{S}_n^{max} = \text{argmax}_{\mathcal{S}}\, \mathcal{C}(\mathcal{S})$
3. Define $\mathcal{C}_n^{opt} = \mathcal{C}(\mathcal{S}_n^{max})$
4. Find $\mathbf{s}_n^{opt} = \text{argmax}_{\mathbf{s}}\, SRP^{\mathcal{S}_n^{max}}(\mathbf{s})$ on a sub-grid
5. Add $(\mathbf{s}_n^{opt}, \mathcal{C}_n^{opt})$ to the set of *potential* speakers
6. Set the restriction of $\mathcal{R}_q$ on $I_q^{\mathbf{s}_n^{opt}}$ to 0

end for

4. NOISE/SOURCE CLASSIFICATION

The proposed method extracts the source location as the one with the highest cumulative SRP, but it does not consider whether this location has been generated by an actual source or by secondary peaks. This problem becomes more difficult in the multiple speaker scenario, where the secondary peaks, resulting from the one-to-many mapping of the TDOA-location relationship, become comparable to the low-energy speakers. In this work, we propose to accomplish this task using an unsupervised Bayesian classifier. The proposed approach uses the cumulative SRP values $\mathcal{C}_n^{opt}$, $n = 1, \ldots, N_e$ ($N_e = N_{max} \times$ *number of frames*), as a classification feature. Then, a 2-component Gaussian mixture fit is calculated using the Expectation-Maximization (EM) algorithm (Fig. 1-d). More precisely, the 2-Gaussian mixture fit is given by

$$f(\mathcal{C}) = w_n \cdot f_n(\mathcal{C}|noise) + w_s \cdot f_s(\mathcal{C}|source) \qquad (17)$$

where $f_n(.)$ and $f_s(.)$ represent the likelihood distributions of the noise and speaker estimates respectively. w_n and w_s denote the corresponding priors. The posterior probability of source/noise given an estimate $\mathbf{s}$, with a cumulative SRP equal to $\mathcal{C}$, is calculated according to

$$p(source|\mathbf{s}) = \frac{w_s \cdot f_s(\mathcal{C}|source)}{w_n \cdot f_n(\mathcal{C}|noise) + w_s \cdot f_s(\mathcal{C}|source)} \qquad (18)$$

$$p(noise|\mathbf{s}) = 1 - p(source|\mathbf{s}) \qquad (19)$$

The location estimate $\mathbf{s}$ is considered to be an actual source if $p(source|\mathbf{s}) > p(noise|\mathbf{s})$. The classification task can be performed at the end of the localization, as it can be done online, by updating the Gaussian mixture parameters after each T frames.

5. EXPERIMENTS AND RESULTS

We evaluate the proposed approach using the AV16.3 corpus [12], where human speakers have been recorded in a smart meeting room (approximately 30m^2 in size) with a 20cm 8-channel circular microphone array. The sampling rate is 16 kHz and the real mouth position is known with an error $\leq$ 5cm [12]. 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 [11] weighting; and a voice activity detector was used in order to suppress silence frames. The localization task is performed in the entire 3D space but, due to the far-field assumption in which the range is ignored, the results are limited to the direction of arrival (DOA). More precisely, the results are reported in terms of the detection rate d_r and the standard deviations of the azimuth $\sigma_{s,\theta}$, and elevation $\sigma_{s,\phi}$. These measures are obtained by fitting a 2-component Gaussian mixture to the estimates error. We also report the real-time factor t on a standard Pentium(R) Dual-Core CPU clocked at 2.50GHz. In the multiple speaker scenario, we also report the percentage of correct estimates p_s. The detection threshold of the probabilistic SRP (pSRP) [10] is chosen such that the resulting false alarm rate is equal to that of the proposed approach.

Table 1 presents the performance of the proposed approach (PA) on single source sequences, and compares it to two well-known approaches, namely the SRP [5] and the MCCC [3]. Note that in these experiments the detection approach from Section. 4 was not used, and N_{max} was set to 1. The coarse grid resolution used in the pSRP and the PA is $20^\circ \times 20^\circ \times 30cm$ for the azimuth, elevation and range, respectively, whereas the resolution of the SRP, MCCC and the reduced search grid (second step of the approach) is $1^\circ \times 1^\circ \times 10cm$. The latter has a size of $30^\circ \times 40^\circ \times 4m$. 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 $N_{max} = 4$.

The results in Table 1 show that the performance of the proposed approach is comparable to the other approaches. More precisely, the standard deviation of the azimuth $\sigma_{s,\theta}$ and elevation $\sigma_{s,\phi}$ as well as the detection rate d_r are comparable, whereas the proposed approach (PA) is approximately 47 times faster than the classical SRP, with an almost-real time performance on a standard machine. That is without any noticeable degradation of the performance. This result illustrates the efficiency of the proposed approach. The MCCC approach however is very slow (noted ∞ in the Table 1) 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 C-SRP performs slightly better than the pSRP approach. This improvement appears clearly in the increased percentage of correct estimates p_s and the average detection rate d_r of each speaker. This improvement is due to the C-SRP, which locates the most likely regions to contain the speakers. It is also worth mentioning that the proposed unsupervised classification approach leads to a FAR $\approx 10\%$ for all experiments. Whereas the detection approach used in the pSRP approach leads to different FARs when the threshold is fixed. This result makes the proposed unsupervised classification technique more attractive. Regarding the real-time factor, we have also found that the C-SRP is 3 times faster than the pSRP.

6. CONCLUSION

We have proposed a novel framework to the multiple speaker localization problem. This approach proposes a two-step search strategy to reduce the computation cost of the classical SRP, without any noticeable degradation of the performance. The proposed framework also presents a cumulative SRP, which improves the multiple speaker detection rate. This approach however does not address the problem of suppressed sources, that occurs in the multiple speaker case. This is part of our future work.

7. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

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

[9] M. Cobos, A. Marti, and J.J. Lopez, "A modified srp-phat functional for robust real-time sound source localization with scalable spatial sampling," *Signal Processing Letters, IEEE*, vol. 18, no. 1, pp. 71 –74, 2011.

[10] 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.

[11] 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.

[12] 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.