Joint Detection and Localization of Multiple Speakers Using a Probabilistic Interpretation of the Steered Response Power

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

Detection and localization of multiple speakers in a noisy and reverberant environment is a fundamental and difficult task. In the literature, steered response power (SRP) based techniques are typically used to accomplish this task which can be computationally intensive. Nonetheless, the localization of multiple speakers remains a challenging in practice. In this paper, we present a novel approach based on a probabilistic interpretation of the SRP. The proposed method replaces the discrete search techniques by proposing an approximate analytical form of the SRP, which can adequately detect and localize multiple speakers. In addition to reliable detection and localization, the potential advantage of this approach is that it provides a probability density function (pdf) of the individual speaker positions rather than point estimates. Experiments on the AV16.3 corpus show the efficacy of the proposed approach.

Index Terms: Steered response power, Multiple speaker localization, Gaussian mixture.

1. Introduction

Microphone arrays have become an essential tool for a large number of signal processing problems. Their area of application includes speech separation/enhancement, acoustic source localization and tracking, but also more advanced approaches such as camera steering for teleconference systems and audio-visual tracking. Among these applications, the detection and localization of multiple concurrent speakers from a short segment of speech remains a difficult and open task; and that although an abundance of localization methods has been proposed in the literature: multi-channel cross correlation (MCCC) [1, 2], adaptive eigenvalue decomposition (ED) [3], time difference of arrival (TDOA) based techniques [4, 5, 6], just to name a few. In this work, we concentrate on the most widely used approach, namely, the steered response power [7]. Despite being reliable and robust, this technique has a few drawbacks: 1) a higher localization precision needs a finer search grid over the 3-D or 2-D space, which can greatly increase the computation cost, 2) in the case of multiple active speakers, detecting the number of speakers and estimating their locations is generally a difficult task, and finally, 3) the localization becomes difficult when a dominant source suppresses the rest of the speakers.

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).

While the first problem can be addressed by reducing the search space through inverse mapping of relative delays [8] or through the method of region contraction [9], the second issue remains difficult in practice. In [10], a sector based approach was introduced. Although this method has a low computation cost, it can only detect active regions, i.e. "sectors", which are more likely to contain the speakers. More accurate estimates require an additional search step inside each active sector. The authors of [11] proposed to combine agglomerative clustering (AC) with Gaussian mixture models (GMMs) and region zeroing (RZ). A similar approach has been used in [12], where the GMM is obtained with the Expectation-Maximization (EM) algorithm. The main difficulty of these approaches consists in determining the number of clusters or mixture components, which represent the active speakers.

In this paper, we propose an alternative solution based on a probabilistic interpretation of the SRP. It mainly deals with the first and second drawbacks that were mentioned earlier. More precisely, an approximate pdf of the source locations is obtained by 1) interpreting the generalized cross correlation function (GCC) [13] as a pdf of the TDOAs, 2) approximating it by a Gaussian mixture (GM) and, then, 3) using the fact that the SRP can be expressed as a sum of the GCC functions [7]. This gives a probabilistic version of the SRP, which can then be used to detect and localize multiple speakers. The advantage of this approach is that it avoids the discrete search step. The number of speakers can reliably be detected with a static threshold. The speaker locations are obtained as pdfs rather than as point estimates. The performance of the proposed approach is demonstrated on the AV16.3 corpus [14].

The paper is organized as follows. Section 2 reviews the classical SRP based acoustic source localization problem. Section 3 introduces the proposed approach. Section 4 presents the experimental results. Finally, in Section 5 we conclude.

2. The Conventional Steered Response Power Approach

The arrival of sound waves at a microphone array introduces time differences between the individual sensor/microphone pairs. The time difference depends on the positions of the microphones $\mathbf{m}_i$, $i = 1, \ldots, M$, as well as the source location $\mathbf{p}$, which is typically specified in spherical coordinates, i.e. the radius r, azimuth θ and elevation ϕ. More precisely, the TDOA introduced at the sensor pair $q = \{\mathbf{m}_i, \mathbf{m}_k\}$ is given by

$$\tau_q(l) = \frac{\|\mathbf{p} - \mathbf{m}_i\| - \|\mathbf{p} - \mathbf{m}_k\|}{c} \quad (1)$$

where $\|\cdot\|$ is the Euclidean norm and where c denotes the speed of sound.

The steered response power (SRP) approach now uses these TDOAs in order 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 [7] by using the generalized cross correlation (GCC).

2.1. Generalized Cross Correlation

Let $s_i(t)$ denote signal received at microphone m_i and let $s_k(t)$ denote the signal received at microphone m_k. Then, the generalized cross correlation $\mathcal{R}_q$ of the two signals is given by

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

where, $S_i(\omega)$ and $S_k(\omega)$ denote the short-time Fourier transforms of $s_i(t)$ and $s_k(t)$, respectively, and $\psi(\omega)$ denotes a pre-filter. A common choice of $\psi(\omega)$ is the Phase Transform (PHAT) weighting [13]: $\psi_{PHAT}(\omega) = \frac{1}{|S_i(\omega)S_k^*(\omega)|}$.

2.2. Single Speaker Localization based on SRP

After the definition of the GCC, the power returned from a particular location $\mathbf{p}\,(r, \theta, \phi)$ can now be estimated as [7]:

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

where, Q is the number of microphone pairs, $\mathcal{R}_q\,, q = 1, \ldots, Q$ are the corresponding GCCs and τ_q denotes the TDOA introduced at the q-th microphone pair. $\mathcal{K}$ is a constant offset, which is introduced by the auto-correlation of each microphone (see [7] for more details) and which is ignored in the rest of the paper. Once the SRP has been calculated for each position $\mathbf{p}$, the source location estimate $\hat{\mathbf{p}}(\hat{r}, \hat{\theta}, \hat{\phi})$ is determined according to [7]:

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

Scanning all possible source locations is computationally expensive. Hence, a variety of approaches have been proposed to reduce the computational burden, such as splitting the 2-D or 3-D space into sectors [10] or adopting space reduction strategies in which only some regions of interest of the space are considered (see e.g. [8, 9]). These approaches are, however, suffering from a poor resolution or from estimation errors due to the inherent discontinuity of the SRP (the latter may partly be avoided by interpolating the GCC functions). Another problem is that conventional SRP based acoustic source localization becomes more difficult in the multi-speaker case, which requires advanced techniques for jointly detecting the number and locations of the speakers.

As mentioned earlier in Section 1, sector based approaches (see e.g. [10]) divide the space into discrete sectors and locate speakers by detecting sectors that have an activity higher than a given threshold. Although, these approaches are computationally efficient, they perform only on the sector level. Thus, another step is required to obtain accurate estimates within each sector. Another category of approaches based on GMMs combined with clustering have been proposed as a solution to this problem. In [11], a GMM and a RZ approach have been combined, respectively, with agglomerative clustering. As a first step, this approach estimates the potential locations using the region contraction technique proposed in [9], and then clusters these locations to obtain the number and location of the speakers. A similar approach has been proposed in [12], where a GMM is obtained with the Expectation-Maximization algorithm (EM) after a clustering step. The components of the GM are then merged iteratively, until a minimal inter-component distance is reached. The difficulty of these approaches consists in 1) determining the number of clusters or GM components, which actually represents the number of active speakers, as well as 2) the poor precision which may result from the discrete search approaches and clustering techniques.

3. Proposed Approach

In this section, we present a novel approach which is based on a probabilistic interpretation of the SRP. It first approximates the pdf of the source locations (i.e. the SRP function) and then jointly detects the number of speakers as well as their locations. This is achieved by

1. interpreting each GCC as a pdf of the TDOA and then approximating this pdf using a GMM (Section 3.1)
2. obtaining a probabilistic approximation of the SRP by using the TDOA GMMs (from step 1) and the deterministic mapping between TDOAs and source locations as given by (1) (Section 3.2)
3. extracting the speaker locations from high power regions of the SRP pdf by using a numerical optimization algorithm (Section 3.3)

In doing so, this approach incorporates the information introduced by each GCC function. Thus, each speaker is characterized by a pdf rather than a point-based estimate. Moreover, this method does not require any discrete search method to estimate the optimal location, and thereby, can be expected to not suffer from the poor accuracy of the estimates. The probabilistic aspect of this approach allows us to efficiently estimate the number of speakers by defining a threshold which characterizes the uncertainty introduced by the noise. Contrary to previous methods which define environment dependent thresholds [10, 11], this probabilistic threshold is less dependent on the environment.

3.1. TDOA Gaussian Mixture Model

Interpreting the normalized GCC as a pdf of the TDOA allows for a probabilistic approach to the source localization problem. More precisely, we propose to approximate the pdf of the TDOA (for each microphone pair) by a GMM. This approximation is based on the assumption (A-1) that each peak in the GCC function (and thereby, each Gaussian component in the approximation) corresponds, at the very most, to one source. We also assume (A-2) that the error introduced in the TDOA detection is a Gaussian process. While the latter assumption has been found to achieve good results in practice [6], the former does not always hold. That is so as the set of locations which have the same TDOA can be approximated as a cone [4] (under the far field assumption). As a result, all the sources lying on this cone correspond to the same GGC peak (and thereby, to the same Gaussian component in the approximation). Using more than one microphone pair, however, reduces the number of possible locations to the intersection of the cones, which allows us to differentiate the sources. Hence, we can conclude that for any pair of locations there might (in the worst case) be a few microphone pairs, for which the Gaussian components are the same. But for the other pairs, (A-1) is still valid.

The most popular approach to estimate the maximum likelihood parameters of a GMM from a given data is the EM algo-

rithm. Using this approach to approximate a TDOA GMM for each microphone pair and each time t, however, would be computationally expensive. Thus, we use a computationally less expensive method which assigns a Gaussian distribution to each peak (assumption (A-1)).

Let K_t^q be the number of GCC peaks of the q-th microphone pair at time t. Furthermore, let $\{\tau_q^1, \ldots, \tau_q^{C_q}\} = [-\tau_q^{max}, \tau_q^{max}]$ be the set of possible TDOAs, with cardinality C_q, and let $\{w_q^1, \ldots, w_q^{C_q}\} = \{\mathcal{R}_q(\tau_q^1), \ldots, \mathcal{R}_q(\tau_q^{C_q})\}$ denote their corresponding GCC values. Negative GCC values are set to 0 (if there are any). For ease of notation, the time index t is dropped in the rest of paper. Subsequently, the TDOA GMM is constructed as follows:

1. Determine the K^q peaks of the GCC.
2. Determine the K^q blocks $\{B_1^q, ..., B_K^q\}$ corresponding to the different peaks. By block we mean the peak interval, which starts at its left foot and ends at the right foot (e.g., see Figure 1).
3. Calculate the Gaussian pdf associated to each block.
4. Calculate and normalize the GMM weights.

The Gaussian pdf $\mathcal{N}(\tau_q; \mu_k^q, (\sigma_k^q)^2)$ corresponding to the k^{th} block B_k^q and its mixture weight $\hat{w}_k^q$ are given by :

$$\mu_k^q = \frac{\sum_{\{l|\tau_q^l \in B_k^q\}} w_q^l \cdot \tau_q^l}{\sum_{\{l|\tau_q^l \in B_k^q\}} w_q^l} \quad (5)$$

$$(\sigma_k^q)^2 = \frac{\sum_{\{l|\tau_q^l \in B_k^q\}} w_q^l \cdot (\tau_q^l - \mu_k^q)^2}{\sum_{\{l|\tau_q^l \in B_k^q\}} w_q^l} \quad (6)$$

$$\hat{w}_k^q = \frac{\sum_{\{l|\tau_q^l \in B_k^q\}} w_q^l}{\sum_{l=1}^{C_q} w_q^l} \quad (7)$$

The means and the weights of the GMM are based on the assumption (A-1) discussed in Section 3.1. Therefore, the cross energy introduced by each block (in the GCC function) is counted for the same Gaussian, assuming that it has been generated by the same speaker. This assumption will help us to extract the different sources in a future step, by assigning each Gaussian component in the GMM, at the very most, to a single speaker.

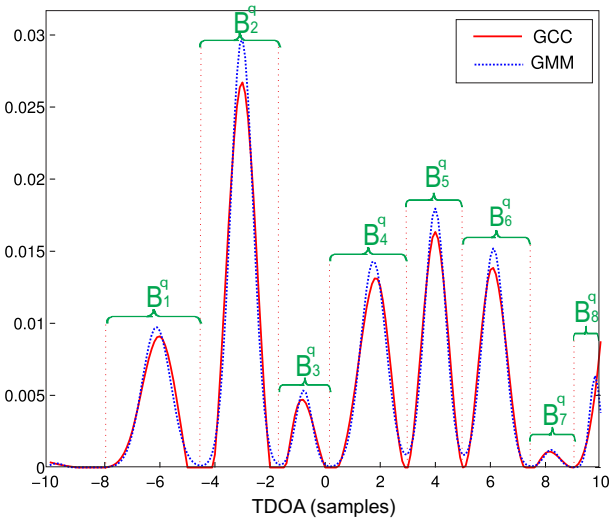


Figure 1: *Illustration of the TDOA GMM ($K^q = 8$).*

3.2. Probabilistic Steered Response Power

After having approximated the GCC functions by GMMs, a probabilistic approximation of the SRP can be obtained by replacing the GCCs in (3) by their probabilistic counterparts. This yields:

$$SRP_{prob}(\mathbf{p}) = \frac{1}{Z} \sum_{q=1}^{Q} \sum_{k=1}^{K^q} \hat{w}_k^q \cdot \mathcal{N}_k^q(\tau_q(\mathbf{p}), \mu_k^q, (\sigma_k^q)^2) \quad (8)$$

where Z is a normalization term.

3.3. Multiple Speaker Localization Algorithm

Given SRP_{prob}, multiple speaker localization is relatively easy. The general idea here is to find in each GCC approximation the Gaussian component which has generated the speaker and then sum these individual (normalized) components to obtain a pdf of the speaker location. The main issue is to find the components representing the *same* speaker (in each GMM approximation). But this can be sorted out by employing the assumption (assumption (A-1) in Section 3.1) that each component in the GCC approximation maps, at the very most, to a unique single speaker. In addition to that, the spatial region in the neighborhood of a given speaker is characterized by high SRP values. Thus, extracting the speakers can be solved by simply finding locations from the high power regions.

This can be done by calculating a coarse grid (e.g., 10° to 20°) and then taking the location with maximum energy. Let $\mathbf{p} = \mathbf{p}_0^s$ be the initial estimate of s-th speaker location. We then extract for each microphone pair q, $q = 1, \ldots, Q$, the Gaussian $\mathcal{N}\left(\tau_q; \mu_{k_{s,q}}^q, (\sigma_{k_{s,q}}^q)^2\right)$ which has generated the speaker. That is done according to:

$$k_{s,q} = \underset{k}{\operatorname{argmax}}\, \mathcal{N}_k^q\left(\tau_q(\mathbf{p}_0^s), \mu_k^q, (\sigma_k^q)^2\right) \quad (9)$$

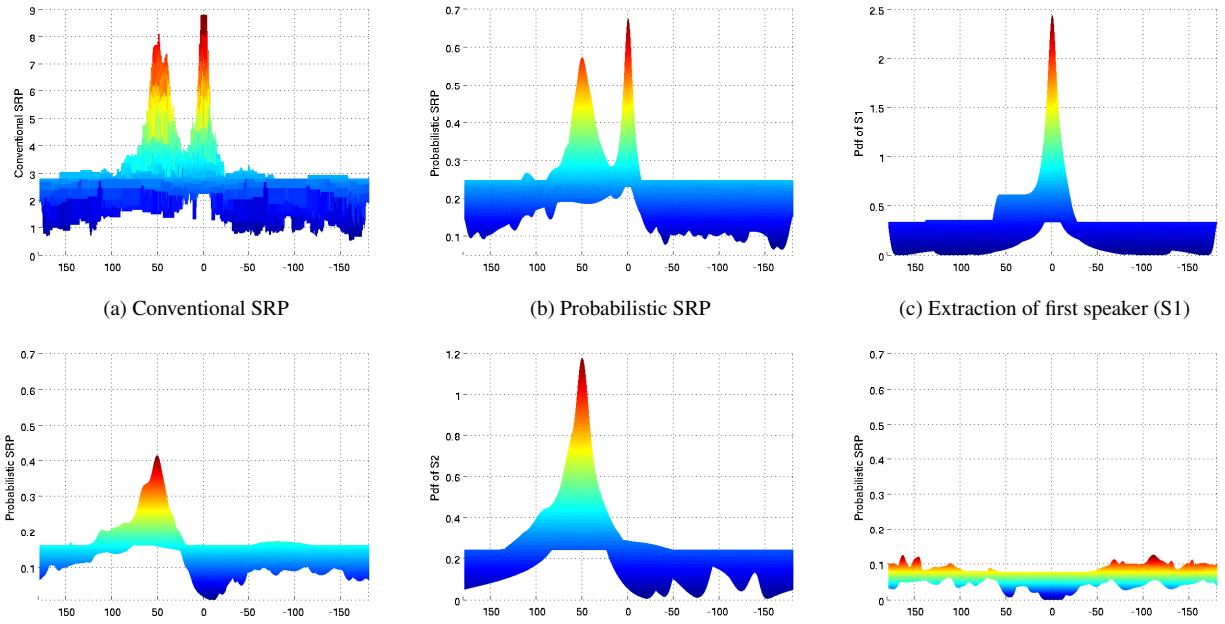
The pdf $SRP_{prob}^s(\mathbf{p})$ of the speaker s is then given as:

$$SRP_{prob}^s(\mathbf{p}) = \frac{1}{Z_s} \sum_{q=1}^{Q} \hat{w}_{k_{s,q}}^q \mathcal{N}\left(\tau_q(\mathbf{p}), \mu_{k_{s,q}}^q, (\sigma_{k_{s,q}}^q)^2\right) \quad (10)$$

where Z_s is a normalization term to obtain a pdf. This pdf can be interpreted as the restriction of the SRP_{prob} on the spatial region in the neighborhood of speaker s.

Now, having extracted the pdf of the source s, we can proceed to estimate the optimal location. This is non-trivial as the relationship between TDOA and location, i.e. (1), is not linear. Hence, a numerical optimization algorithm is required. We here use the *Broyden-Fletcher-Goldfarb-Shanno* algorithm (BFGS) which is a popular choice for a quasi-Newton algorithm [15]. In principle, however, any other numerical optimization algorithm could be used.

It can be observed that the approximation in (10) does not require a grid search to perform the localization. Furthermore, it can be easily extended to the multiple speaker case. The pseudo-code in Algorithm 1 presents an iterative algorithm to detect and localize multiple speakers. Note that the number of speakers may be overestimated in the case where the maximum number of concurrent speakers N_{max} is not known. This may increase the computation complexity, but it does not affect the localization performance. In order to ensure that we do not miss any speaker, the **BREAK** instruction in Step 10 of Algorithm 1 can be removed.



(a) Conventional SRP (b) Probabilistic SRP (c) Extraction of first speaker (S1)

(d) Probabilistic SRP after extraction of S1 (e) Extraction of second speaker (S2) (f) Probabilistic SRP after extraction of S2

Figure 2: *A frontal view illustrating the probabilistic SRP and the multiple speaker localization algorithm. The horizontal coordinates are the azimuth [-180°,180°] and the elevation [0°,90°].*

Algorithm 1 : Multiple Speakers Localization Algorithm

Let N_{max} be the maximum number of speakers.
for $i = 1 : N_{max}$ **do**
1. Calculate the SRP_{prob} for a coarse grid.
2. Use the location with the maximum energy as initialization $\mathbf{p}_0^s$.
for all microphone pairs **do**
3. Determine the Gaussian component which has generated $\mathbf{p}_0^s$.
end for
4. Define the pdf $SRP_{prob}^s(\mathbf{p})$ of the current potential speaker s.
5. Run an optimization algorithm on $SRP_{prob}^s(\mathbf{p})$ to estimate the optimal location $\mathbf{p}_{opt}^s$.
if $SRP_{prob}(\mathbf{p}_{opt}^s) > P_{noise}$ **then**
6. Add $\mathbf{p}_{opt}^s$ to the set S of speakers.
7. Discard the Gaussian components $k_{s,q}$ of speaker s from the SRP_{prob}.
8. Normalize the weights.
9. Go to step 1.
else
11. BREAK.
end if
end for
10. Return the set of speakers S.

It is also worth mentioning that the speaker locations from the previous time frame can be used as initialization for the current time frame. Thus, the coarse grid in Step 1 of Algorithm 1 will be required only if the maximum number of speakers has not been reached and if the current location has a probability SRP_{prob} higher than the confidence threshold (i.e. P_{noise}). However, as the energy of each speaker spans over a sector [10], a coarse grid (10° to 20°) can be used as initialization process. The probabilistic threshold P_{noise} characterizes the regions of confidence. In fact, assuming that the source location can be approximated by a Gaussian distribution, the region of confidence is represented then by an ellipse with equal likelihood. Therefore, defining a region of confidence is equivalent to defining a threshold P_{noise} over the pdf SRP_{prob}.

Figure 2 shows a localization example of two overlapping speakers using Algorithm 1. The confidence threshold is $P_{noise} = 0.3$. The proposed algorithm not only localizes the speakers, but it also provides an approximation of the pdf of each speaker. This approximation can be efficiently used to obtain more information about the sources such as, the uncertainty of the estimates given by the variance as well as the higher order moments. More precisely, this can be done using importance/rejection sampling techniques to approximate this pdf by a single Gaussian distribution (as suggested by Fig. 2c and Fig. 2e). This is a part of the future work.

4. Experiments and Results

4.1. Database and Experimental Setup

In order to evaluate the proposed approach, we performed a set of localization experiments on the AV16.3 corpus [14]. In this corpus, real 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 [14]. We present experiments for 4 different sequences with a varying number of speakers. The first sequence is `seq18-2p-0101` where two moving speakers talk simultaneously while getting as close as possible to each other and then slowly move apart. The purpose of this sequence is to test the separability and

Sequence seq18-2p-0101 / two speakers		
	S1	S2
Anomalies Rate (%)	21.51%	15.01%
Azimuth RMSE	2.01°	1.58°

Sequence seq37-3p-0001 / three speakers			
	S1	S2	S3
Anomalies Rate (%)	17.20%	17.77%	15.48%
Azimuth RMSE	1.31°	2.68°	1.70°

Sequence seq24-2p-0111 / two speakers		
	S1	S2
Anomalies Rate (%)	34.69%	25.81%
Azimuth RMSE	1.71°	1.84°

Sequence seq40-3p-0111 / three speakers			
	S1	S2	S3
Anomalies Rate (%)	29.64%	22.52%	23.77%
Azimuth RMSE	1.98°	1.94°	2.40°

Table 1: *Multiple source/speaker localization results of four different sequences from the AV13.6 corpus [14] with two/three real human concurrent speakers.*

localization of sources that are close. The second sequence seq24-2p-0111 shows two moving speakers crossing each other twice. The third experiment is performed on the three speakers sequence seq40-3p-0111. Two speakers are seated while the third speaker is initially standing and then walking back and forth behind the seated speakers. The motivation is both multi-source localization and separation. In the last sequence seq37-3p-0001 two speakers remain seated and a third one is standing at five different positions. The number of speakers talking simultaneously varies between two and three. The above sequences are 57, 47, 49 and 511 seconds in length, respectively.

In the experiments, which are reported below, the signal was divided into frames of 1024 samples (64ms). All the GCCs were calculated under use of PHAT [13] weighting. The probabilistic threshold P_{noise} was 0.3 and the maximum number of speakers N_{max} was 5. As there is no point in localizing an inactive speaker, we further used a voice activity detector for suppressing silence frames. Due to the planar array geometry and the far-field sources, the location space is limited to the set of azimuth angles in the range $[-180°, 180°]$.

The results are reported in terms of the anomaly rate (AR) [2] - i.e. the percentage of estimates that vary from the true azimuth by more than 5°- and by the *root-mean-square error* (RMSE) for the non-anomalous estimates. Although, the miss-detection rate is a good evaluation method for this approach, it is not possible to calculate it for this corpus. This is due to the lack of ground truth segmentation of each speaker which can show whether the speaker is active or not. Instead, the corresponding localization figures are shown.

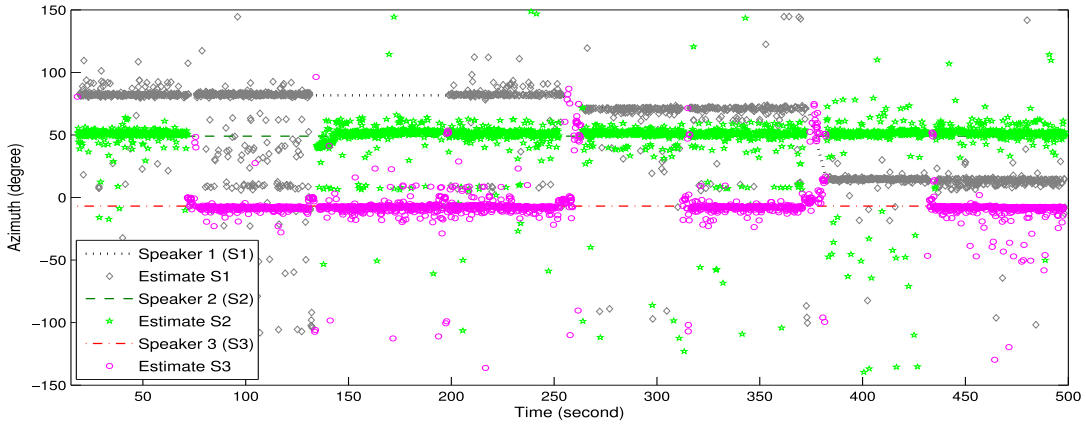
4.2. Results and Analysis

Although we proposed the probabilistic SRP approach as a solution to the multiple speaker localization problem, we, in a first step, ran it on a number of single speaker sequences. This allowed us to compare its performance to that of the conventional SRP (grid resolution $= 0.5°$). On sequence seq02-1p-0000, we obtained an AR of 35.24% and a RMSE of $1.75°$ without thresholding ($P_{noise} = 0$) compared to 36.04% as AR and $1.95°$ as RMSE for the conventional SRP. Similar results were obtained using other single speaker sequences of the corpus. These results show that, the Gaussianity assumption, which we made in the GCC approximation and the proposed GMM, do not affect the performance of the algorithm. Actually, these approximations help improving the accuracy (RMSE) of the estimates. Furthermore, setting the probabilistic threshold to $P_{noise} = 0.3$ discards 18% of the estimates, with a false rejection rate, i.e. percentage of estimates wrongly discarded, of 3.83%, leading to an AR of 22.19% with a RMSE of $1.65°$. This improvement is expected as the probabilistic threshold is used as an uncertainty criterion of the estimates. Therefore, the percentage of outliers is reduced. It is worth mentioning that a similar threshold could be defined on the conventional SRP. The main problem however resides in the dependance of the thresholding on the environment and the source location. As consequence, the threshold value should be adapted to (and within) each recording, whereas, the proposed threshold P_{noise} is defined over a pdf.

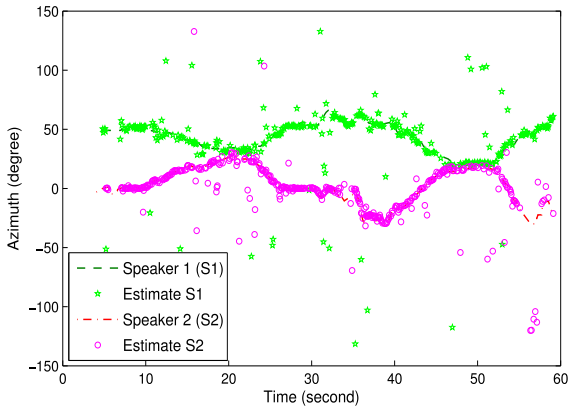
Table 1 shows the multiple speaker localization results for the above mentioned sequences. Given the single source localization results reported in [2] with different localization techniques (including the conventional SRP) but on the same corpus, we can conclude that anomaly rate is low and that the RMSE of the azimuth is very good. We can also see that the results are highly correlated to the speech energy level of each speaker as well as the distance between the speakers and the microphone array. The first sequence seq18-2p-0101 shows that the AR and the RMSE of the first speaker are worse than those of the second speaker. This degradation can be explained by the difference in the average distance which is about 0.97m for S2 and 1.20m for S1, We can also conclude from the recordings that S2 is louder than S1. A similar remark holds for sequence seq24-2p-0111 where the average distance of S1 and S2 is 1.81m and 1.69m, respectively, and for sequence seq37-3p-0001 where S1, S2 and S3 are 1.90m, 0.99m and 0.79m far from the center of the array. In the case of sequence seq40-3p-0111, S1, S2, and S3 are on average 1.01m, 1.84m, 1.21m away from the array, respectively. However, S1 has the worst AR. This result can be better understood by listening to the recording of this sequence, which shows that S1 is completely suppressed by S2 and S3. In this work, we have formulated the probabilistic SRP in the time domain. The issue of suppressed speakers however, can be better addressed in the frequency domain, where the separation of sources is more efficient [10, 12]. This aspect is currently under investigation.

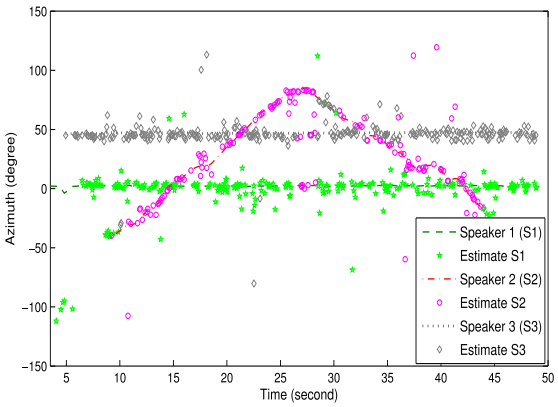
5. Conclusions

We proposed a multiple source localization approach based on a probabilistic approximation of the steered response power. In this approach, each speaker is characterized by a pdf, instead of a point estimate. The speakers are efficiently detected and located with a probabilistic confidence threshold. The potential of the approach has been demonstrated through experiments on the AV16.3 corpus. We are currently formulating the probabilistic SRP approach in the frequency domain in order to address the problem of suppressed speakers. Our future work will extend the probabilistic SRP approach to a more flexible probabilistic framework.



(a) Sequence "seq37-3p-0001" : Localization of three speakers where the number of active sources changes over time.





(b) Sequence "seq18-2p-0101" : Localization of two concurrent speakers. (c) Sequence "seq40-3p-0111" : Localization of three concurrent speakers.

Figure 3: *Illustration of the localization and separation performance of the proposed algorithm.*

6. References

[1] 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.

[2] 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.

[3] 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.

[4] M. Brandstein, J. Adcock, and H. Silverman, "A closed-form method for finding source locations from microphone-array time-decay estimates," *Proc. ICASSP*, vol. 5, pp. 3019–3022, 1995.

[5] 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, 1997.

[6] Y. Oualil, F. Faubel, and D. Klakow, "A multiple hypothesis Gaussian mixture filter for acoustic source localization and tracking," in *Proc. IWAENC*, Sep. 2012, pp. 233–236.

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

[8] 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.

[9] 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.

[10] G. Lathoud and M. Magimai.-Doss, "A sector-based, frequency-domain approach to detection and localization of multiple speakers," in *Proc. ICASSP*, 3 2005.

[11] 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.

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

[13] 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.

[14] 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.

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