REAL-TIME MULTIPLE SPEAKER DOA ESTIMATION IN A CIRCULAR MICROPHONE ARRAY BASED ON MATCHING PURSUIT

Anthony Griffin[⋆]*, Despoina Pavlidi*[⋆†]*, Matthieu Puigt*[⋆]*, and Athanasios Mouchtaris*[⋆†]

[⋆]FORTH-ICS, Heraklion, Crete, Greece, GR-70013
[†]University of Crete, Department of Computer Science, Heraklion, Crete, Greece, GR-71409
Email: {agriffin, pavlidi, mpuigt, mouchtar}@ics.forth.gr

ABSTRACT

We recently proposed an approach inspired by Sparse Component Analysis for real-time localisation of multiple sound sources using a circular microphone array. The method was based on identifying time-frequency zones where only one source is active, reducing the problem to single-source localisation in these zones. A histogram of estimated Directions of Arrival (DOAs) was formed and then processed to obtain improved DOA estimates, assuming that the number of sources was known. In this paper, we extend our previous work by proposing a new method for the final DOA estimations, that outperforms our previous method at lower SNRs and in the case of six simultaneous speakers. In keeping with the spirit of our previous work, the new method is very computationally efficient, facilitating its use in real-time systems.

***Index Terms*—** Array signal processing, direction of arrival estimation, multiple source localisation

1. INTRODUCTION

The localisation of audio sources is something that we do very instinctively as human beings, using our in-built array of microphones—our ears. Thus it is a natural area of research for array signal processing, and one that has had a lot of interest over recent decades [1].

A common application of direction of arrival (DOA) estimation is in teleconferencing, where the knowledge of the location of a source can be used to steer a camera, or to enhance the capture of the desired source with beamforming, for example, thus avoiding the need for each speaker to wear a microphone.

Although most conversations or meetings are dominated by periods during which only one person is speaking, there are often times when multiple people are talking at once. Localising the sources at these times is a much more difficult problem. Indeed, even the smallest overlap of speakers—caused by a brief interjection, for example—can disrupt the localisation of the original source. A system that is designed to handle the localisation of multiple sources just sees the interjection as another source that can be simultaneously captured or rejected as desired.

A great deal of source localisation methods are based on the using the time difference of arrival (TDOA) [2] at different microphone pairs to estimate the DOA of the speaker. Many of them use the Generalised Cross-Correlation PHAse Transform (GCC-PHAT), which has significant limitations in the case of multiple sources and/or reverberant environments. Such limitations have been alleviated by considering ratios of the GCC-PHAT peaks in [3] and by using the redundant information contained in more than two microphones in [4].

As an alternative to the above classical approaches, Sparse Component Analysis (SCA) methods [5, ch. 10] may be seen as natural extensions of multiple-sensor single-source localisation methods to multiple source localisation. They basically assume that one source is dominant over the others in some time-frequency windows or "zones". Using this assumption, the multiple source propagation estimation problem may be rewritten as a single-source one in these windows or zones, and the above methods estimate a mixing/propagation matrix, and then try to recover the sources. Their main advantage is their flexibility to deal with the situations when the number of sources is lower or higher than the number of sensors. If we estimate this mixing matrix and if we know the geometry of the microphone array, we may then localise the sources, as proposed in [6, 7, 8], for example.

Most of the SCA approaches require the sources to be W-disjoint orthogonal (WDO) [9]—in each time-frequency window, at most one source is active—which is approximately satisfied by speech in anechoic environments but not in reverberant conditions. On the contrary, other methods assume that the sources may overlap in the time-frequency domain, except in some tiny "time-frequency analysis zones" where only one of them is active (e.g. [5, p. 395],[10]). Unfortunately, most of the SCA methods and their DOA extensions are off-line methods (e.g. [7] and the references within). However, [6, 8] are frame-based methods: [6] requires WDO sources, while our previous proposed method [8] used single-source zones

This work is funded by the Marie Curie IAPP "AVID MODE" grant within the 7th European Commission's Framework Program.

as in [10]. Note that concepts involved in [7, 8] look quite similar. However, our proposed approach [8] is real-time and uses a circular array of microphones while [7] works off-line and processes two-microphone only configurations.

Thus our previous work presented a method for real-time multiple source localisation using a circular microphone array [8] that was based on finding single-source zones [10], and performing single source DOA estimation on these zones using the method of [11]. These DOA estimations were then assembled into a histogram to enable the localisation of the multiple sources through peak-picking.

This work considers a similar framework but the localisation of the multiple sources is performed using a method inspired by Matching Pursuit (MP) [12], which provides a more accurate DOA estimation, particularly at lower values of signal-to-noise ratios (SNRs), and configurations of more sources. Similar to [8], we assume that the number of sources is known, using the source counting method of [13], for example.

2. PROBLEM STATEMENT

Similar to [8][13], we assume that M microphones of an equispaced circular array receive an anechoic mixture of P sources:

$$x_i(t) = \sum_{g=1}^{P} a_{ig} s_g\left(t - t_i(\theta_g)\right) + n_i(t), \qquad i = 1, \cdots, M \tag{1}$$

where $x_i(t)$ is the signal received by microphone m_i, a_{ig} are attenuation factors, $t_i(\theta_g)$ is the delay from source s_g to microphone m_i, θ_g is the DOA of the source s_g, and $n_i(t)$ is the noise at m_i. For one given source, the relative delay between signals at adjacent microphones, hereafter referred to as microphone pair $\{m_i m_{i+1}\}$, with the last pair being $\{m_M m_1\}$, is given by

$$\tau_{m_i m_{i+1}}(\theta_g) \triangleq t_{i+1}(\theta_g) - t_i(\theta_g) = l \sin(A - \theta_g + (i-1)\alpha/c), \tag{2}$$

where l is the distance between adjacent microphones, A is the obtuse angle formed by the chord $m_1 m_2$ and the x-axis (with m_1 placed on the x-axis [8]), and c is the speed of sound. We aim to estimate the DOAs, θ_g, of the P sources.

3. CONFIDENCE MEASURES AND LOCALISATION

3.1. Definitions and assumptions

We locate "constant-time analysis zones" in the time-frequency (TF) representation of the incoming data. Each of them is a set of adjacent TF points, denoted as (Ω). We assume that for each source there exists (at least) one zone (Ω), which we call a "single source analysis zone", where the source is dominant over the others, an assumption which is satisfied when working with speech signals [5, p. 395].

For a pair of signals (x_i, x_j), we define the cross-correlation over analysis zones of the moduli of their TF transform as

$$R'_{i,j}(\Omega) = \sum_{\omega \in \Omega} \left| X_i(\omega) \cdot X_j(\omega)^* \right|, \tag{3}$$

where $X_i(\omega)$ is the TF transform of $x_i(t)$ and $*$ stands for the complex conjugate. The associated correlation coefficient is

$$r'_{i,j}(\Omega) = R'_{i,j}(\Omega) / \sqrt{R'_{i,i}(\Omega) \cdot R'_{j,j}(\Omega)}. \tag{4}$$

3.2. Single-source confidence measures

We detect all constant-time analysis zones that satisfy the following inequality as single-source analysis zones:

$$\overline{r'}(\Omega) \geq 1 - \epsilon, \tag{5}$$

where $\overline{r'}(\Omega)$ is the average correlation coefficient between adjacent pairs of observations [10] and ϵ is a small user-defined threshold.

3.3. DOA estimation in a single-source zone

After the single-source analysis zones detection stage, we apply a modified version [8] of the algorithm in [11], in order to estimate the DOA of a speaker in each detected zone.

The frequency where the magnitude of the cross-power spectrum (defined as $R_{i,i+1}(\omega) = X_i(\omega) \cdot X_{i+1}(\omega)^*$, over the frequency range of a zone (Ω)) reaches its maximum is denoted as $\omega_i^{\max}$ [8].

Then, using (2), with $1 \leq i \leq M$ and $0 \leq \phi < 2\pi$, we evaluate the Phase Rotation Factors [11],

$$G_{m_i \to m_1}^{(\omega_i^{\max})}(\phi) \triangleq e^{-j\omega_i^{\max} \tau_{m_i \to m_1}(\phi)}, \tag{6}$$

where $\tau_{m_i \to m_1}(\phi) \triangleq \tau_{m_1 m_2}(\phi) - \tau_{m_i m_{i+1}}(\phi)$ is the difference in the relative delay between the signals received at pairs $\{m_1 m_2\}$ and $\{m_i m_{i+1}\}$. We estimate the Circular Integrated Cross Spectrum, defined in [11] as

$$\mathrm{CICS}(\phi) \triangleq \sum_{i=1}^{M} G_{m_i \to m_1}^{(\omega_i^{\max})}(\phi) \angle R_{i,i+1}(\omega_i^{\max}). \tag{7}$$

The estimated DOA of a speaker in the considered zone is then given by

$$\hat{\theta} = \arg \max_{0 \leq \phi < 2\pi} \mathrm{CICS}(\phi). \tag{8}$$

3.4. Block-based histogram

Once we have estimated all the local DOAs in the single-source zones (Sections 3.2 & 3.3), we form a histogram from the set of estimations in a block of B consecutive frames and we smooth it by applying an averaging filter with a window of length h_N [6]. Using B consecutive frames increases the accuracy of the final DOA estimations.

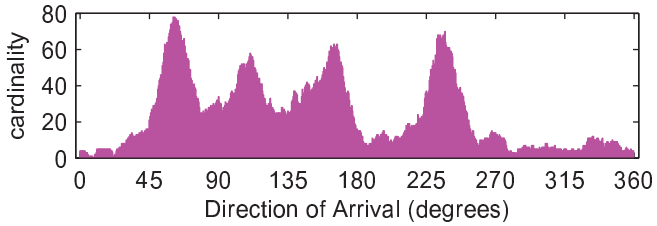


Fig. 1. Example of a smoothed histogram of four speakers in a simulated reverberant environment at 20dB SNR.

4. DOA ESTIMATION OF MULTIPLE SOURCES

Given $\boldsymbol{y}_n$, the length-L smoothed histogram in the n-th frame, and an estimate of the number of active sources $\hat{P}_n$, our goal is then to find the DOA of each source.

Figure 1 shows an example histogram with four active sources at $60°, 105°, 165°$, and $240°$ and with 20dB SNR. The four sources are clearly visible and similarly shaped, which inspired us to approach the DOA estimation problem as one of sparse approximation using source atoms. Thus the idea—proceeding along similar lines to MP—is to find the DOA of a source by correlation with a source atom, remove its contribution, and repeat the process until $\hat{P}_n$ sources have been found.

We chose to model each source atom as a smooth pulse, such as that of a Blackman window, although the choice of the window did not prove to be critical. The choice of the width is key, and reasoning and experiments showed that a high *accuracy* of the method requires wide source atoms at lower SNRs and narrow source atoms at higher SNRs. Furthermore, the *resolution* of the method—the ability to discriminate between two closely spaced sources—is adversely affected as the width of the source atom increases. This suggests making the width a parameter in the estimation process, however this would come at the cost of an increase in computational complexity—which we wish to avoid—thus, we chose to use fixed-width source atoms.

Further investigation revealed that a two-width method provided a good compromise between these constraints, where a narrower width is used to accurately pick the location of each peak, but a wider width is used to account for its contribution to the overall histogram and provide better performance at lower SNRs.

Let $\boldsymbol{q}$ be a length-Q row vector containing a length-Q Blackman window, then let $\boldsymbol{u}$ be a length-L row vector whose first Q values are populated with $\boldsymbol{q}$ and then padded with $L - Q$ zeros. Now let $\boldsymbol{u}^{(k)}$ denote a version of $\boldsymbol{u}$ that has been "circularly" shifted to the right by k elements, the circular shift means that the elements at either end wrap around, and a negative value of k implies a circular shift to the left.

Now choose $Q = 2Q_0 + 1$ where Q_0 is a positive integer. The maximum value of $\boldsymbol{q}$ (or equivalently $\boldsymbol{u}$) will occur at (Q_0+1)-th position. Now define $\boldsymbol{r} = \boldsymbol{u}^{(-Q_0)}$. The maximum value of the length-L row vector $\boldsymbol{r}$ occurs at its first element. Let the elements of $\boldsymbol{r}$ be denoted r_i, and its energy be given by $E_r = \sum r_i^2$. Now form the matrix R, which consists of circularly shifted versions of $\boldsymbol{r}$. Specifically, the k-th row of R is given by $\boldsymbol{r}^{(k-1)}$.

As previously discussed, we need two widths of source atoms, so let $\mathrm{R_N}$ and $\mathrm{R_W}$ be matrices for the peak detection (denoted by "N" for narrow) and the masking operation (denoted by "W" for wide), respectively, with corresponding source atom widths Q_N and Q_W.

Our algorithm proceeds as follows:

i. Set the loop index $j = 1$
ii. Form the product $\boldsymbol{a} = \mathrm{R_N}\boldsymbol{y}_{n,j}$
iii. Let the elements of $\boldsymbol{a}$ be given by a_i, find $i^* = \arg\max_i a_i$
iv. The DOA of this source is given by $(i^* - 1) \times 360°/L$
v. Remove the contribution of this source as

$$\boldsymbol{y}_{n,j+1} = \boldsymbol{y}_{n,j} - (\boldsymbol{r}_\mathrm{W}^{(i^*-1)})^T \frac{a_{i^*}}{E_{r_\mathrm{N}}}$$

vi. If $j < \hat{P}_n$, increment j and go to step ii.

It should be noted that this method was developed with the goal of being computationally-efficient so that the DOA estimation could be done in real-time. In particular, the matrix $\mathrm{R_N}$ was found to be an efficient way of dealing with the inherent circularity of the histogram due to its measuring direction modulo $360°$. It should be clear that $\mathrm{R_N}$ is a circulant matrix and will contain $L - Q$ zeros on each row, and both of these properties may be exploited to provide a reduced computational load.

5. RESULTS

In order to investigate the performance of our proposed method, we conducted various simulations in a reverberant room. We used the fast image-source method (ISM)[14] to simulate a room of $6 \times 4 \times 3$ meters. The boundaries were assumed to be plane reflective walls, characterised by uniform reflection coefficient $r_\mathrm{coef} = 0.5$, and reverberation time $T_{60} = 0.25$s. A circular array with 8 omnidirectional microphones and a radius of 5cm was placed in the centre of the room, coinciding with the origin of the x and y-axis. All the point sources were speech signals located 1.5m from the array, sampled at 44.1kHz, processed in frames of 2048 samples, with 50% overlapping in time. The FFT size was 2048 and the width of the TF analysis zones Ω was 344Hz with 50% overlapping in frequency, and with $f_\mathrm{max} = 4$kHz as the highest frequency of interest. The sound velocity was $c = 343$ m/s. The single-source confidence measure threshold was $\epsilon = 0.2$, histogram bin size was $0.5°$, and $h_N = 5°$ was the averaging filter window length. We used $B = 43$ frames in the histogram, equating to a "history length" of one

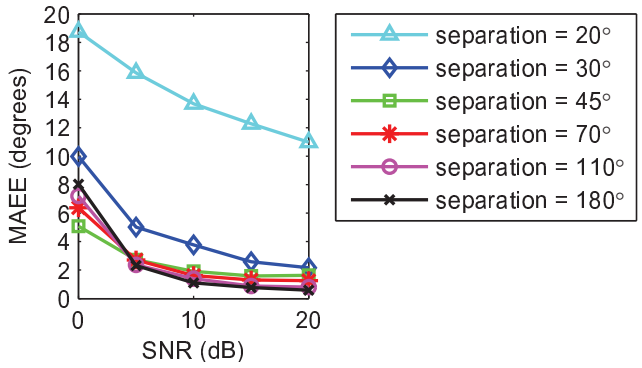


Fig. 2. DOA estimation error vs SNR for pairs of simultaneous speakers in a simulated reverberant environment.

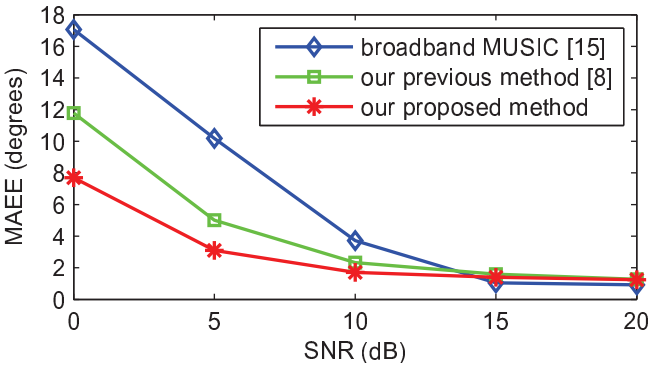


Fig. 3. DOA estimation error vs SNR for four intermittant speakers in a simulated reverberant environment.

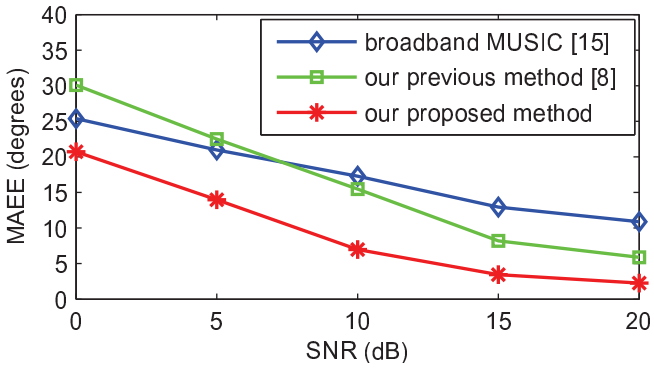


Fig. 4. DOA estimation error vs SNR for six simultaneous speakers in a simulated reverberant environment.

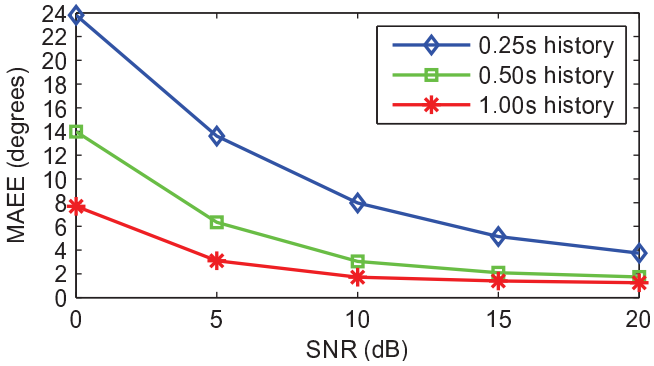


Fig. 5. DOA estimation error vs SNR for four intermittant speakers in a simulated reverberant environment.

second unless otherwise stated. The final values chosen for the source widths were $Q_{\mathrm{P}} = 81$ and $Q_{\mathrm{M}} = 161$, or $40°$ and $80°$ respectively. However, due to the shape of the Blackman window, the effective widths are closer to $20°$ and $40°$. It should also be noted that we simulated each orientation of sources in $10°$ steps around the array in order to capture a more accurate performance all around the array.

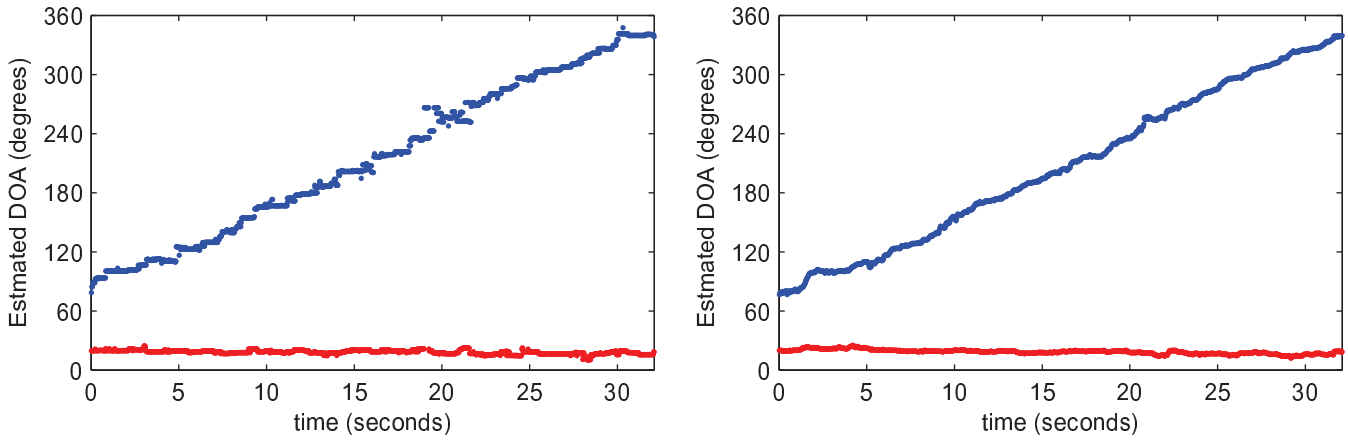
The performance of our system was measured by the mean absolute estimated error (MAEE) which measures the difference between the true DOA and the estimated DOA over all speakers, all 36 orientations and all the frames of the source signals. Figure 2 shows the MAEE against SNR for pairs of speech signals at various separations. Our method performs well for larger separations, but the effective resolution with the chosen parameters is somewhere around $25°$.

Figure 3 shows the MAEE against SNR for four speakers originally located at $0°$, $45°$, $105°$ and $180°$. The speakers were intermittant, but there was a significant part of the signals where all four were active simultaneously. The results of our previous work [8] are also presented, along with our implementation of the broadband extension of the well-known MUSIC algorithm as described in [15]. For the estimation of the covariance matrix of the received signals needed by MUSIC, we used the observation vectors of $B = 43$ consecutive frames, in order to compare the systems fairly. The rest of the parameters are the same as in the other methods. The three methods perform similarly at higher SNRs, but the method proposed in this work outperforms the other two for SNRs below 15dB.

Six people speaking simultaneously might be considered a stress test for a localisation system, pushing the limits of what is achievable. Figure 4 presents the results for just that with speakers originally located at $0°$, $60°$, $105°$, $180°$, $250°$ and $330°$. The proposed method clearly peforms the best for all SNRs but is probably only usable at an SNR of 10dB or higher in this particularly taxing case.

Next, in Figure 5 we return again to the four intermittant speaker simulation and explore the effect of differing history lengths using our proposed method. There is an obvious performance degradation in the DOA estimation as the history length decreases, as the algorithm has less data to work with in the histogram. However decreasing the history also decreases the latency of the system, in turn increasing responsiveness. At higher SNRs there is very little degradation, suggesting that in a system with reasonable SNR, a history of 0.5 seconds could provide an accurate and responsive system.

Finally, in Figure 6 we turn our attention to DOA estimations of real recordings of two male speakers speaking simultaneously and continuously in a large office room. The first speaker was static at about $20°$ and the other was walking



(a) Using the method of [8]. (b) Using the proposed method.

Fig. 6. DOA estimations for recordings of one moving and one static speaker, both speaking continuously and simultaneously.

slowly around the array from about $80°$ to $340°$, and the SNR was approximately 15dB. Figure 6(a) presents the DOA estimations using the method of [8]. The two sources are clearly localised, but there is a "stair-like" effect particularly visible on the moving source. This effect is much reduced in the results using the method proposed in this paper—shown in Figure 6(b)—implying that this method tracks the source much more responsively.

In keeping with the real-time spirit of our previous work [8] [13], we also implemented the algorithm described in Section 4 in C++ to measure its computational performance, which we found to be 5% of the available processing time, making it an excellent candidate to be included in our real-time multiple source localisation system.

6. CONCLUSIONS

In this paper we extended our previous work on real-time multiple sound source localisation using a circular microphone array [8][13], by proposing a new method which improves the DOA estimation accuracy of multiple sources. The method was tested and compared with other previously published methods in a simulated reverberant environment and on real data, and shown to perform very well in most conditions, requiring only 5% of the available processing time.

7. REFERENCES

[1] H. Krim and M. Viberg, "Two decades of array signal processing research - the parametric approach," *IEEE Sig. Proc. Mag.*, pp. 67–94, July 1996.

[2] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP Journal on Appl. Sig. Proc.*, vol. 2006, pp. 1–19, 2006.

[3] D. Bechler and K. Kroschel, "Considering the second peak in the GCC function for multi-source TDOA estimation with microphone array," in *Proc. of IWAENC*, 2003, pp. 315–318.

[4] J. Benesty, J. Chen, and Y. Huang, "Time-delay estimation via linear interpolation and cross correlation," *IEEE Trans. on Speech and Audio Proc.*, vol. 12, no. 5, September 2004.

[5] P. Comon and C. Jutten, *Handbook of blind source separation, independent component analysis and applications*, Academic Press, 2010.

[6] M. Swartling, N. Grbić, and I. Claesson, "Source localization for multiple speech sources using low complexity non-parametric source separation and clustering," *Sig. Proc.*, vol. 91, no. 8, pp. 1781–1788, 2011.

[7] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Sig. Proc.*, 2011.

[8] D. Pavlidi, M. Puigt, A. Griffin, and A. Mouchtaris, "Real-time multiple sound source localization using a circular microphone array based on single-source confidence measures," in *Proc. of ICASSP*, 2012, pp. 2625–2628.

[9] Ö. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Sig. Proc.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[10] M. Puigt and Y. Deville, "A new time-frequency correlation-based source separation method for attenuated and time shifted mixtures," in *Proc. of ECMS*, 2007, pp. 34–39.

[11] A. Karbasi and A. Sugiyama, "A new DOA estimation method using a circular microphone array," in *Proc. of EUSIPCO*, 2007, pp. 778–782.

[12] S. Mallat and Z. Zhang, "Matching pursuit with time frequency dictionaries," *IEEE Trans. on Sig. Proc.*, vol. 41, no. 12, pp. 3397–3415, December 1993.

[13] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Source counting in real-time sound source localization using a circular microphone array," in *Proc. of SAM*, 2012, pp. 529–532.

[14] E.A. Lehmann and A.M. Johansson, "Diffuse reverberation model for efficient image-source simulation of room impulse responses," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 18, no. 6, pp. 1429 –1439, August 2010.

[15] S. Argentieri and P. Danès, "Broadband variations of the MUSIC high-resolution method for sound source localization in robotics," in *Proc. of IROS*, 2007, pp. 2009–2014.