Real-time multiple sound source localization using a circular microphone array based on single-source confidence measures

Despoina Pavlidi, Matthieu Puigt, Anthony Griffin, Athanasios Mouchtaris

REAL-TIME MULTIPLE SOUND SOURCE LOCALIZATION USING A CIRCULAR MICROPHONE ARRAY BASED ON SINGLE-SOURCE CONFIDENCE MEASURES

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

[⋆] FORTH-ICS, Heraklion, Crete, Greece, GR-70013
[†] University of Crete, Department of Computer Science, Heraklion, Crete, Greece, GR-71409

ABSTRACT

We propose a novel real-time adaptative localization approach for multiple sources using a circular array, in order to suppress the localization ambiguities faced with linear arrays, and assuming a weak sound source sparsity which is derived from blind source separation methods. Our proposed method performs very well both in simulations and in real conditions at 50% real-time.

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

1. INTRODUCTION

Audio source localization using an array of sensors is a rich topic which has interested many signal processing researchers for more than 30 years [1]. Applications e.g. include speaker location discovering in a teleconference, event detection and tracking, robot movement in an unknown environment, etc. Among all the approaches proposed in the literature e.g. beamforming [2], using grids of possible locations [3] or a probabilistic framework [4], numerous ones are based on Time Difference Of Arrival (TDOA) [5] at different microphone pairs for estimating the Direction of Arrival (DOA). Many of them use the Generalized Cross-Correlation PHAse Transform (GCC-PHAT), which has significant limitations in the case of multiple sources and/or reverberant environments. Such limitations have been partially solved by considering ratios of the GCC-PHAT peaks [6] and by using the redundant information contained in more than two microphones [7]. Further work proposed to change the geometry of the array of sensors in order to suppress some localization ambiguities due to linear arrays [8, 9].

As an alternative to the above classical approaches, Sparse Component Analysis (SCA) methods [10] may be seen as natural extensions of multiple sensor single source localization methods to multiple source localization. They basically assume that sources are sparse in an analysis domain obtained after a sparsifying transform (usually a short-time Fourier transform) and that, as a consequence, 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 (i.e containing for each source columns of gains due to attenuation during the propagation to the sensors, and of TDOAs), and then try to recover the sources. Their main advantage is their flexibility to deal with both (over-)determined and underdetermined configurations, i.e. the cases when the number of sources is resp. (strictly) lower or higher than the number of sensors. By only considering the estimation of this mixing matrix, and by taking advantage of the known geometry of microphones in the array, it is then possible to localize the sources, as e.g. proposed in [11].

SCA approaches are mainly divided in two families. Most of them require a strong source sparsity assumption named W-Disjoint Orthogonality (WDO) [12]: in each time-frequency window, at most one source is active. From a signal processing point of view, WDO is a nice assumption which is almost fulfilled by speech signals in *anechoic* environments. However, this assumption does not hold in reverberant conditions [13] and/or when source signals are musical. Moreover, SCA methods based on this assumption are usually derived from DUET [12] which is unable to estimate "large" time shifts. 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 (see e.g. [14] and the references within). They particularly use "constant-time single-source analysis zones", i.e. a set of frequency-adjacent time-frequency windows in order to estimate TDOAs, and are able to accurately estimate a large range of time shifts (typically up to 200 samples in [14]).

Unfortunately, most of the SCA approaches are off-line methods, except a few ones [15, 16]. The work in [15] assumes the WDO assumption and is thus not well-suited to reverberant configurations. Such an approach has then been considered for a localization problem in [11]. Furthermore, [16] looks for single-source zones, but does not estimate the TDOAs and has thus never been considered in a localization problem. In this paper, we propose a new adaptative multiple-source localization approach, using the relaxed sparsity assumption of [14, 16], but which additionally estimates DOAs. We thus assume a much weaker and much more realistic sparsity assumption than [11, 12, 15]. Moreover, and contrary to [14], we take into consideration the known geometry of the microphone array in order to perform a better estimation of DOAs. In particular, we use a circular array of sensors which reduces the location indeterminacies inherent to linear arrays [8].

The remainder of the paper then reads as follows. We describe the considered localization problem in Section 2. We then introduce our proposed method in Section 3. Section 4 provides an experimental validation of the approach while we conclude and discuss future directions of the incoming work in Section 5.

2. PROBLEM STATEMENT

For an equispaced circular array of M microphones, the signal received at each microphone m_i is

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

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

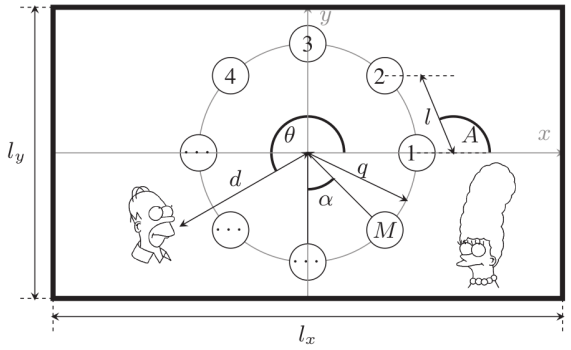

Fig. 1. Circular sensor array configuration.

where P is the known number of sources s_g, assumed to be far-field, a_{ig} and $t_i(\theta_g)$ are respectively the attenuation factor and the delay from source s_g to microphone m_i, θ_g is the DOA of the source s_g observed with respect to the x-axis (Fig.1), and $n_i(t)$ is the noise at m_i. For one given source, the relative delay between signals received 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

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

where l and α are resp. the distance and angle between $\{m_i m_{i+1}\}$, A is the obtuse angle formed by the chord $m_1 m_2$ and the x-axis, and c is the speed of sound. We aim to estimate the DOAs θ_g.

3. PROPOSED METHOD

3.1. Definitions and assumptions

Before describing our proposed method, we first introduce the definitions and assumptions of the proposed method. We follow the framework of [14] that we recall hereafter for the sake of clarity. We consider a short-time Fourier transform as a sparsifying transform. In practice, we partition the incoming data in overlapping time frames on which we compute a Fourier transform, hence providing a time-frequency (TF) representation of observations. We then define a "constant-time analysis zone", (t, Ω), as a series of frequency-adjacent TF points (t, ω) . In the remainder of the paper, we ommit t in the (t, Ω) for simplicity. We assume the existence, for each source, of (at least) one constant-time analysis zone, said to be "single-source", where one source is "isolated", i.e. it is dominant over the others. Note that this assumption is much weaker than WDO since sources can overlap in the TF domain except in these few single-source analysis zones. We further assume that when several sources are active in the same analysis zone, they should vary so that the moduli of at least two observations are linearly dependent. This last assumption, satisfied in practice by audio signals, allows us to process correlated sources, contrary to classical statistic-based DOA methods. For any pair of signals (x_i, x_j), we respectively define the cross-correlation over analysis zones of their TF transform and of their moduli as

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

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

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

We now introduce the proposed method whose core stages are:

1. The application of a joint-sparsifying transform to the observations, using the above TF transform.
2. The single-source analysis zones detection (see Section 3.2).
3. The DOA estimation (see Sections 3.3 and 3.4).

3.2. Single-source confidence measures

In this section, we describe how to find single-source analysis zones. Our approach is based on the following theorem [14]:

Theorem 1 *A necessary and sufficient condition for a source s_k to be isolated in an analysis zone (Ω) is*

$$r'_{i,j}(\Omega) = 1 \qquad \forall i, j \in \{1, \ldots, M\}. \quad (5)$$

In practice, we do not consider the correlation between all the pairs (i, j) of observations, but the average correlation between pairs $(i, i+1)$ of observations [14], denoted $\overline{r'}(\Omega)$ hereafter. Moreover, in practice, we consider that an analysis zone is single-source iff

$$\overline{r'}(\Omega) \geq 1 - \epsilon, \quad (6)$$

where ϵ is a small user-defined threshold.

3.3. DOA estimation in a single-source zone

At this point, by considering all the single-source analysis zones satisfying (6), we re-examine the single source multi-sensor DOA problem in these zones, hence the interest in such sparsity assumption. In order to estimate the DOA of a speaker in a single-source constant-time analysis zone, we propose a modified version of the algorithm in [8], which is designed exclusively for circular arrays. We selected this algorithm because of its anti-reverberation characteristics, in conjunction with the robust behaviour in noisy environments and the computational efficiency.

We consider the circular array geometry (Fig.1) introduced in Section 2. Since the estimation of the DOA takes place in a constant-time analysis zone, the phase of the Cross-Power Spectrum of a microphone pair is evaluated over the frequency range of the specific zone as $\angle R_{i,i+1}(\Omega) = \frac{R_{i,i+1}(\Omega)}{|R_{i,i+1}(\Omega)|}$ where $R_{i,i+1}(\Omega)$ is defined in (3). We denote as $\omega_i^{\max}$ the frequency where the magnitude of the cross-power spectrum reaches its maximum, given by,

$$\omega_i^{\max} = \arg\max_{\Omega} |R_{i,i+1}(\omega)|. \quad (7)$$

At this point, in [8], the harmonics selection module selects only the indices of the peaks of the Cross Power Spectrum for the localization. Instead, since we aim at a real-time application, we use only the $\omega_i^{\max}$ frequency, which corresponds to the strongest component of the cross-power spectrum in a single-source zone. Experimentally this introduced inaccuracy was found to result in acceptable performance.

Using (2) and (7), with $1 \leq i \leq M$ and $0 \leq \phi < 2\pi$, we evaluate the Phase Rotation Factors [8],

$$G_{m_i\to m_1}^{(\omega_i^{\max})}(\phi) \triangleq e^{-j\omega_i^{\max}\tau_{m_i\to m_1}(\phi)}, \quad (8)$$

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 proceed with the estimation of the Circular Integrated Cross Spectrum, defined in [8] as

$$\text{CICS}(\phi) \triangleq \sum_{i=1}^{M} G_{m_i \to m_1}^{(\omega_i^{\max})}(\phi) \angle R_{i,i+1}(\omega_i^{\max}). \quad (9)$$

The DOA of the speaker in the specific single source zone is, then, obtained as,

$$\hat{\theta} = \arg \max_{0 \le \phi < 2\pi} \text{CICS}(\phi). \quad (10)$$

3.4. Improved block-based decision

In the above analysis, several single-source zones may lead to the same DOA, as the isolated source is the same in each of them. Deriving the DOA for each sound source involves clustering the estimated DOAs, which can be done by finding peaks in their histogram for a particular time segment. This motivated us to apply an approach based on Parzen windows for obtaining a density function from the estimated DOAs [17]. For every time frame of incoming data, we evaluate the confidence measures (5) for all analysis zones and we discard those zones that do not satisfy (6). In each single source zone we apply the algorithm described in Section 3.3 and we get an estimate of the DOA at the specific single-source analysis zone. Then, from the set of estimations in a block of B consecutive frames, we estimate the density function of the estimations, by applying a rectangular window over the estimations of this block.

If we denote as v the independent variable, the probability density function of v according to [17] is:

$$\mathbb{P}(v) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{h_N} w\left(\frac{v - v_i}{h_N}\right), 0 \le v < 2\pi \quad (11)$$

where N is the total number of estimates in a block, h_N is the length of the window and $w(.)$ is the rectangular window. The DOA of each of the P sound sources is estimated as

$$\hat{\theta}_i = \frac{h_N N \sum_{j=l_l}^{l_h} j \cdot \mathbb{P}(j)}{\sum_{j=l_l}^{l_h} \mathbb{P}(j)}, \qquad \left\{ \begin{array}{l} l_l = k - h_N/2 \\ l_h = k + h_N/2 \end{array} \right\} \quad (12)$$

where $i = 1, \cdots, P$. The index k is one of the P highest local peaks of $\mathbb{P}(v)$ and there is a 1 to 1 correspondence between i and k. The P highest local peaks are selected under the constraint that they are "distant-enough", i.e. separated by a user-defined threshold δ. The block of estimates slides with each new time frame.

4. RESULTS

In order to evaluate the proposed algorithm, we performed speech localization simulations and real-time experiments. We denote F_s the sampling frequency, $f_{\max}$ the highest frequency of interest and q the radius of the circular array. The aforementioned parameters take values: $F_s = 44.1$ kHz, $f_{\max} = 4$ kHz, and $q = 0.05$ m, which guaranties the absence of spatial aliasing in a circular array [9]. The number of microphones was $M = 8$, the single-source confidence measure threshold was $\epsilon = 0.2$, the Parzen window length was $h_N = 5°$, the angular threshold was $\delta = 10°$, the frame size was equal to 2048 samples, whereas the Block size B was equal to 44100 samples. The FFT size was 2048 and the width of the TF analysis zones Ω was 344 Hz. The overlapping, both in time and frequency domain, was 50%. The sound velocity was $c = 343$ m/s.

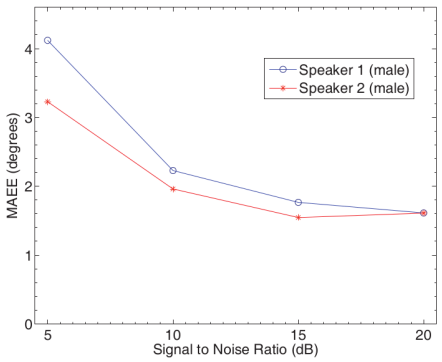

Fig. 2. Mean Absolute Estimation Error (MAEE) of the DOA in light reverberant simulated environment for various SNRs of white additive noise

In order to simulate a real reverberant room, we used the fast image-source method (ISM) described in [18]. The length, width and height of the room were respectively set to 6 m, 4 m, and 3 m. The boundaries were assumed plane reflective walls, characterized by the uniform reflection coefficient, $r_{\text{coef}} = 0.5$ and the reverberation time was set to $T_{60} = 0.25$ s. The circular array was placed in the centre of the room, which coincides with the origin of the x and y-axis. The coordinates of microphone m_1 were $(0.05, 0, 1)$ (as in Fig. 1) and the distance between adjacent microphones was $l = 38$ mm. The microphones are assumed to be ideal point receivers with omnidirectional directivity patterns. The sound sources are omnidirectional point transmitters and they are located 1.5 m away from the centre of the array. In Fig. 2, we present the Mean Absolute Estimation Error (MAEE) in white Gaussian noise conditions for SNR=$\{5, 10, 15, 20\}$ dB. The two sources are separated by $45°$. The MAEE is evaluated from $0°$ to $360°$ in $10°$ steps for all cases.

The real-time experiments were conducted in a typical office room with approximately the same dimensions and placement of the microphone array as in the simulations. The signal to noise ratio in the room was, on average, 15 dB, mainly because of the presence of A/C units. The algorithm was implemented in software executed on a standard PC (Intel 2.40 GHz Core 2 CPU, 2GB RAM). We used Shure SM93 microphones (omnidirectional) with a TASCAM US2000 8-channel USB soundcard. The execution time is 50 % real time (i.e. 50% of the available processing time). In the following results, some percentage of the estimated error can be attributed to the inaccuracy of the source positions. Fig. 3(a) shows the DOA estimation of two sources, where Speaker 2 is sitting at distance 1.5 m from the centre of the array and at $15°$, while Speaker 1 is following a circular motion from $75°$ to $345°$ with an almost steady speed and also at distance 1.5 m from the array. The performance achieved in this experiment shows the moving source is accurately traced. In Fig. 3(b) we show the estimation of DOA for a 3 sources scenario. Two male speakers are sitting at $240°$ and $0°$ while a female speaker is at $160°$. The maximum deviation from the real DOA is $4.5°$ for Speaker 1, $1.4°$ for Speaker 2 and $3.1°$ for Speaker 3. Even if the presence of three sources reduces the number of single-source zones, the performance achieved here is close to the one presented in the simulations with 2 sources. Due to lower spectral overlap, the best accuracy is obtained with the female speaker. The MAEE error of 2 static speakers, spaced by $45°$, for pair positioning from $0°$ to $360°$

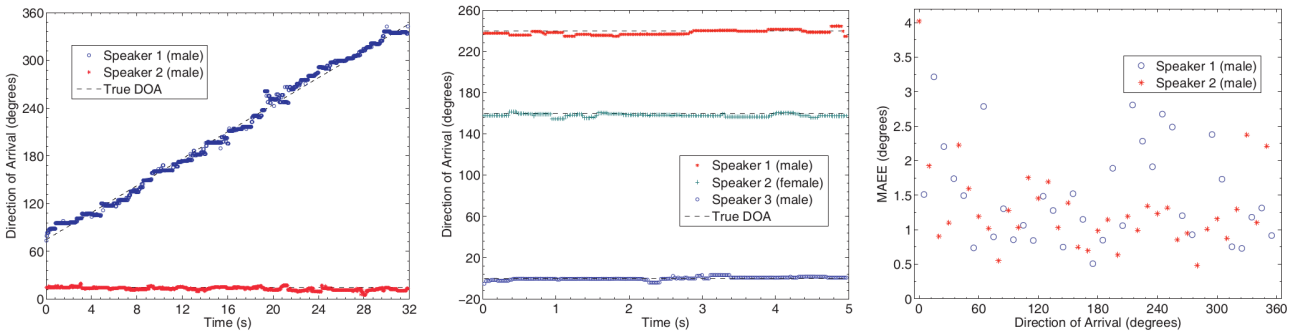

Fig. 3. Estimations in real environment: (a) DOA estimation of 2 sources. Speaker 1 is moving from 75° to 345°, while Speaker 2 is static at 15°. (b) Estimation of DOA of 3 static sources at $\{240, 160, 0\}^\circ$. (c)Mean Absolute Estimation Error (MAEE) of the DOA of 2 male static sources versus the true DOA. The sources are separated by 45°. The MAEE is evaluated from 0° to 360° in 10° steps.

in 10° steps around the array is shown in Fig. 3(c). The maximum MAEE is 3.4° for Speaker 1 and 4.0° for Speaker 2.

5. CONCLUSION AND FUTURE WORK

In this paper, we proposed a novel real-time adaptative source localization approach using a circular array, in order to suppress the localization ambiguities faced with linear arrays, and assuming a weak sparsity assumption which is derived from blind source separation methods. To the best of our knowledge, such a configuration has never been considered before. Our proposed method performs very well both in simulations and in real conditions at 50% of real-time. In future work, we will characterize the performance of the proposed method in various scenarios involving more sources with closer DOAs. We also plan to investigate the real-time estimation of the number of sources, which is here assumed to be known.

6. 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] W. Kellermann, "Beamforming for speech and audio signals," in *Handbook of Signal Processing in Acoustics*, D. Havelock, S. Kuwano, and M. Vorländer, Eds. Springer, New York, 2008.

[3] D. Malioutov, M. Cetin, and A.S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. on Sig. Proc.*, vol. 53, no. 8, pp. 3010–3022, August 2005.

[4] S.T. Birchfield and D.K. Gillmor, "Fast bayesian acoustic localization," in *Proc. of ICASSP*, 2002.

[5] 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.

[6] 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.

[7] 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.

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

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

[10] R. Gribonval and M. Zibulevsky, "Sparse component analysis," in *Handbook of Blind Source Separation, Independent Component Analysis and Applications*, P. Comon and C. Jutten, Eds., pp. 367–420. Academic Press, 2010.

[11] 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.

[12] 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.

[13] S. Schulz and T. Herfet, "On the window-disjoint-orthogonality of speech source in reverberant humanoid scenarios," in *Proc. of DAFx–08*, 2008, pp. 241–248.

[14] 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.

[15] S. Rickard, R. Balan, and J. Rosca, "Real-time time-frequency based blind source separation," in *Proc. of ICA*, 2001, pp. 651–656.

[16] D. Smith, J. Lukasiak, and I. Burnett, "Two channel, block adaptative audio separation using the cross correlation of time frequency information," in *Proc. of ICA*, 2004, vol. LNCS 3195.

[17] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern classification*, Wiley, 2001.

[18] 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, aug. 2010.