Source counting in real-time sound source localization using a circular microphone array

Despoina Pavlidi*†, Anthony Griffin*, 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: {pavlidi, agriffin, mpuigt, mouchtar}@ics.forth.gr

***Abstract*—Recently, we proposed an approach inspired by Sparse Component Analysis for real-time localization 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 localization for 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 three different methods for counting the number of sources by looking for prominent peaks in the derived histogram based on: (a) performing a peak search, (b) processing an LPC-smoothed version of the histogram, (c) employing a matching pursuit-based approach. The third approach is shown to perform very accurately in simulated reverberant conditions and additive noise, and its computational requirements are very small.**

I. Introduction

For more than 30 years, audio source localization using an array of sensors has generated wide interest in the signal processing community [1]. Indeed, applications are numerous, including speaker location discovering in a teleconference, event detection and tracking, and robot movement in an unknown environment.

Among all the approaches proposed in the literature, numerous ones are based on Time Difference Of Arrival (TDOA) [2] at different microphone pairs for estimating the Direction of Arrival (DOA). Many of them are based on the Generalized Cross-Correlation PHAse Transform (GCC-PHAT).

As an alternative to the above classical approaches, Sparse Component Analysis (SCA) methods [3, ch. 10] may be seen as natural extensions of multiple sensor single source localization methods to multiple source localization. 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 both the situations when the number of sources is respectively (strictly) 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 localize the sources, as proposed in [4]–[6], for example.

Most of the SCA approaches require the sources to be W-disjoint orthogonal (WDO) [7]—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. [3, p. 395], [8]). Unfortunately, most of the SCA methods and their DOA extensions are off-line methods (e.g. [5] and the references within). However, [4] and [6] are frame-based methods: [4] requires WDO sources while our previous proposed method [6] used single-source zones as in [8]. Note that concepts involved in [5] and [6] look quite similar. However, our proposed approach [6] is real-time and uses a circular array of microphones while [5] works off-line and processes two-microphone only configurations.

A second issue in source localization consists of estimating the number of sources, known as source counting. Many methods of the literature propose estimating the intrinsic dimension of the recorded data, i.e. for an acoustic problem, they estimate the number of active sources at each time instant. Most of them are based on information theory (see [9] and the references within). In our considered problem, the estimation of the number of sources is different. Indeed, the different single-source zones may lead to a set of DOAs that we need to cluster. In classification, some approaches for estimating both the clusters and their numbers have been proposed (e.g. [10]), while several solutions specially dedicated to DOAs have been tackled in [3, p. 388] and [11].

In this paper, we propose an extension to our previous work in [6] which both counts the number of sources and locates them in real time. For that purpose, we propose three approaches working on the histogram of estimated DOAs and based on the amplitude of the histogram, its linear predictive coding analysis, and matching pursuit.

II. Problem statement

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

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

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 [6]), and c is the speed of sound. We aim to estimate the number P of active sources, along with the corresponding DOAs, θ_g.

III. Confidence Measures and Localization

A. Definitions and assumptions

In this paper we focus our attention on the estimation of the number of sound sources, impinging on an array of sensors. This comes as a natural extension to our previous work [6], where the number of sources was assumed as known a priori and we recall it here for the sake of clarity. 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 "single source analysis zone", where that source is dominant over the others. For any pair of signals (x_i, x_j), we define the cross-correlation over analysis zones of the moduli of their TF transform as

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

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

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

B. Single-source confidence measures

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

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

where $\overline{r'}(\Omega)$ is the average correlation coefficient between adjacent pairs of observations [8] and ϵ is a small user-defined threshold.

C. DOA estimation in a single-source zone

After the single-source analysis zones detection stage, we apply a modified version [6] of the algorithm in [12], in order to estimate the DOA of a speaker in each detected zone.

We consider the circular array geometry introduced in Section II. We denote as $\omega_i^{\max}$ 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 [6].

Using (2), with $1 \leq i \leq M$ and $0 \leq \phi < 2\pi$, we evaluate the Phase Rotation Factors [12],

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

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 [12] as

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

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

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

D. Block-based decision

Since we have estimated all the local DOAs in the above single-source zones (Sections III-B and III-C), we form the histogram from the set of estimations in a block of B consecutive frames and we smooth it by applying an average filter with a window of length h_N [6]. This way we estimate the probability density function of the estimations, $\mathbb{P}(v)$, $0 \leq v < 2\pi$.

We then proceed with the estimation of the number of sources, P. Given this estimation, $\hat{P}$, we estimate the final DOAs, as:

$$\hat{\theta}_i = \frac{h_N N \sum_{j=l_l}^{l_h} j \cdot \mathbb{P}(j)}{\sum_{j=l_l}^{l_h} \mathbb{P}(j)}, \qquad \left\{ \begin{matrix} l_l = k - h_N/2 \\ l_h = k + h_N/2 \end{matrix} \right\} \quad (9)$$

where $i = 1, \cdots, \hat{P}$. The index k is one of the $\hat{P}$ highest local peaks of $\mathbb{P}(v)$ and there is a 1-to-1 correspondence between i and k.

IV. Counting the Sources

Most of the approaches on the Source Counting problem are based on information theoretic criteria, with most dominant the Minimum Description Length (MDL) [9]. They depend on ordered eigenvalues of the estimated covariance matrix of the observation vectors, in the same spirit as it has been proposed in the MUSIC algorithm framework [13]. These methods are computationally intensive and have difficulty robustly estimating the number of active sources. Further to these drawbacks, in our considered problem the estimation of the number of sources is different as we are working with a histogram of the DOA estimations. Thus we investigate three different methods to estimate the number of sources: a Peak Search approach, a Linear Predictive Coding (LPC) approach and a Matching Pursuit approach under the constraint that the maximum number of sources cannot exceed a user defined upper threshold P_{MAX}.

A. Peak Search

In order to estimate the number of sources we perform a peak search of the Block-histogram in the following manner.

a. Since there is always at least one active source in a block of estimates, we set $i_s = 1$, where i_s corresponds to a counter of the peaks assigned to sources so far. We also set $u_{i_s} = u_1 = \arg\max_{0 \leq \phi < 2\pi} \mathbb{P}(u)$, i.e. the

histogram bin which corresponds to the highest peak of the smoothed histogram. Finally, we set the threshold $z_{i_s+1} = \max\{\mathbb{P}(u_{i_s})/2, z_{\text{static}}\}$, where z_{static} is a user-defined static threshold.

b. We locate the next highest peak in the smoothed histogram, $\mathbb{P}(u_{i_s+1})$. If $\mathbb{P}(u_{i_s+1}) \geq z_{i_s+1}$ and $u_{i_s+1} \notin [u_{j_s} - \delta, u_{j_s} + \delta], \forall u_{j_s}$ with $j_s < (i_s+1)$ then $i_s = i_s+1$, $z_{i_s+1} = \max\{\mathbb{P}(u_{i_s})/2, z_{\text{static}}\}$

c. We stop when a peak in the histogram fails to satisfy the threshold z_{i_s+1} or if the upper threshold P_{MAX} is reached. The estimated number of sources is $\hat{P} = i_s$.

We note that peak-search approaches on histograms of estimates have been proposed in literature [11]. Here, we give another perspective on these approaches by processing a smoothed histogram and by using a non-static peak threshold.

B. Linear Predictive Coding

Linear Predictive Coding (LPC) coefficients are widely used to provide an all-pole smoothed spectral envelope of speech and audio signals. We use this to more clearly point out the peaks of the smoothed histogram and to suppress any noisy areas. We represent the envelope of the histogram with its LPC-smoothed counterpart from which the total number of peaks is chosen as our estimate of $\hat{P}$.

C. Matching Pursuit

The third method we propose to perform the source counting is an algorithm inspired by Matching Pursuit. The idea is to pick the peaks of the smoothed histogram by correlation, and then remove the contribution of each source. We choose a source atom to be approximated by a smooth pulse such as that of the Blackman window. 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 $\boldsymbol{u}^{(m)}$ denote a version of $\boldsymbol{u}$ that has been "circularly" shifted to the right by m elements, the circular shift means that the elements at either end wrap around, and a negative value of m 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 the energy in one row 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 m-th row of R is given by $\boldsymbol{r}^{(m-1)}$. Finally, let $\boldsymbol{\gamma}$ be a length-P_{MAX} vector whose elements γ_i are some predetermined thresholds, representing the relative energy of the i-th source.

Given $\boldsymbol{y}$, the smoothed histogram in the current frame, our algorithm proceeds as follows:

a. Set the initial value $\boldsymbol{y}_1 = \boldsymbol{y}$, and the loop index $j = 1$

b. Form the product $\boldsymbol{b} = \text{R}\boldsymbol{y}_j$

c. Let the elements of $\boldsymbol{b}$ be given by b_i, find $i^* = \arg\max_i b_i$

d. Remove the contribution of this source as

$$\boldsymbol{y}_{j+1} = \boldsymbol{y}_j - (\boldsymbol{r}^{(i^*-1)})^T \frac{b_{i^*}}{E_r}$$

e. Calculate the contribution of this source as

$$\delta_j = \sum_i \frac{y_{j,i} - y_{j+1,i}}{y_{1,i}}$$

where $y_{j,i}$ is the i-th element of $\boldsymbol{y}_j$.

f. If $\delta_j > \gamma_j$ increment j, else go to step h.

g. If $j \leq P_{\text{MAX}}$ go to step b.

h. The number of sources in the current frame is equal to $j - 1$.

It should be noted that this method was developed with the goal of being computationally-efficient so that the source counting could be done in real-time. In particular, the matrix R 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 R is a circulant matrix and will contain $L - Q$ zeros on each row, and both of which may be exploited to provide a reduced computational load.

V. Results and Discussion

In order to investigate the performance of our methods, we conducted simulations of 4 audio sources 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, characterized by uniform reflection coefficient $r_{\text{coef}} = 0.5$, 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. The four 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 $f_{\max} = 4$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 average filter window length. For the Peak Search method (PS), $z_{\text{static}} = 0.05 \sum_{j=0}^{360} \mathbb{P}(j)$ and $\delta = 20°$, and the LPC order used was 16. The thresholds for the Matching Pursuit-based method (MP) were $\boldsymbol{\gamma} = \{0.15, 0.14, 0.12, 0.1\}$. It is important to note that all these parameters were fixed, and in particular, independent of the signal-to-noise ratio (SNR).

We tested all three methods with $P_{\text{MAX}} = 4$ and with block sizes—referred to also as history lengths—equal to 0.25s, 0.5s and 1s. Fig. 1 shows an example DOA estimation of the four sources at $10°, 55°, 115°$, and $190°$. Note that the estimation of each source is prolonged for some period of time after he/she stops talking or respectively is delayed when he/she starts talking. This is due to the fact that the DOA estimation at each time instant is based on a block of estimates of length B seconds ($B = 1$s in this example). We refer to these periods as "transition periods", which we define as the time interval

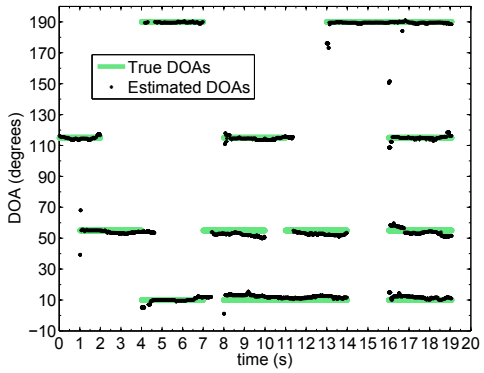

Fig. 1. Estimation of DOA of 4 speakers at $10°, 55°, 115°$, and $190°$ in a simulated reverberant environment with SNR=20 dB and a one second history.

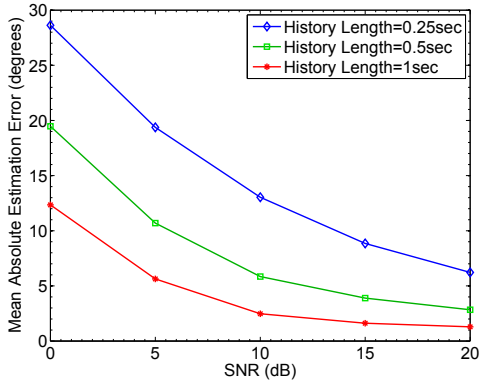

Fig. 2. MAEE of the DOA excluding the transition periods in a simulated reverberant environment for various history lengths and various SNRs

starting when a new or existing speaker starts or stops talking and ending B seconds later. In Fig. 2 we show the Mean Absolute Estimation Error (MAEE) of the four speakers for various SNR values. It should be noted that each point is an average of 36 simulations in which each speaker was shifted by $10°$ steps each simulation in order to capture a more accurate performance all around the array. In Table I, we give success rates of the source counting (percentage of frames correctly counting the number of sources) for the three methods under consideration with various history lengths and differing values of SNRs. For these results and the estimation of the MAEE, the transition periods were not taken into account.

There is an obvious performance improvement for both the DOA estimation and source counting as the history length increases, as the algorithms have more data to work with in the histogram. However increasing the history also increases the latency of the system, in turn decreasing responsiveness. The results in Fig. 2 and Table I suggest that a history length of 0.5 s might be a good compromise. The DOA algorithm runs in 50% of real-time [6], while all three proposed methods add only 5% to that computational time. The Matching Pursuit method is clearly the best performing source counting method.

VI. Conclusion

In this paper we extended our previous work on real-time multiple sound source localization using a circular microphone array [6], by proposing three different methods for counting the number of sources. All these methods identify prominent peaks in the smoothed histogram from the DOA estimation, and are simple and efficient to implement. The methods were tested in a simulated reverberant environment, with various additive noise conditions. In particular, the matching pursuit based method was found to perform very accurately in most conditions, requiring only 5% of the available processing time.

TABLE I
Source Counting success rates excluding transition periods

Method	History Length	SNR (dB) 0	5	10	15	20
PS	0.25s	33.5%	43.9%	58.3%	69.3%	76.0%
LPC	0.25s	25.5%	39.3%	55.2%	61.8%	63.5%
MP	0.25s	44.1%	60.2%	77.6%	85.0%	88.4%
PS	0.5s	47.8%	62.1%	75.5%	82.8%	86.0%
LPC	0.5s	35.3%	58.0%	72.8%	74.6%	74.4%
MP	0.5s	61.2%	81.7%	94.2%	96.0%	96.6%
PS	1s	51.5%	68.9%	81.6%	88.3%	90.6%
LPC	1s	46.3%	78.5%	83.7%	81.7%	79.4%
MP	1s	77.4%	97.5%	100.0%	100.0%	100.0%

Acknowledgment

This work is funded by the Marie Curie IAPP "AVID MODE" grant within the European Commission's FP7.

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

[4] 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.

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

[6] 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 *ICASSP*, 2012, to appear.

[7] 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.

[8] 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.

[9] E. Fishler, M. Grosmann, and H. Messer, "Detection of signals by information theoretic criteria: General asymptotic performance analysis," *IEEE Trans. on Sig. Proc.*, vol. 50, no. 5, pp. 1027–1036, May 2002.

[10] G. Hamerly and C. Elkan, "Learning the k in k-means," in *Proc. of NIPS*, 2003, pp. 281–288.

[11] B. Loesch and B. Yang, "Source number estimation and clustering for underdetermined blind source separation," in *Proc. IWAENC*, 2008.

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

[13] C.T. Ishi, O. Chatot, H. Ishiguro, and N. Hagita, "Evaluation of a music-based real-time sound localization of multiple sound sources in real noisy environments," in *Proc. of IROS*, 2009, pp. 2027 –2032.

[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, 2010.