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:

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

(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 $m_i$, $\theta_g$ is the DOA of the source $s_g$, and $n_i(t)$
is the noise at $m_1$. For one given source, the relative delay between signals at adjacent microphones, hereafter referred to as microphone pair $\{m_im_{i+1}\}$, with the last pair being $\{m_Mm_1\}$, is given by

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

where $l$ is the distance between adjacent microphones, $A$ is the angle formed by the chord $m_1m_2$ 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, $\theta_g$.

### 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 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 ‘single source analysis zone’, where that source is dominant in each source there exists (at least) one zone $\theta$ of active sources, along with

where $\tau_{m_im_{i+1}}(\phi) \triangleq \tau_{m_1m_2}(\phi) - \tau_{m_{i+1}m_1}(\phi)$ is the difference in the relative delay between the signals received at pairs $\{m_1m_2\}$ and $\{m_{i+1}m_1\}$. We estimate the Circular Integrated Cross Spectrum, defined in [12] as

$$\text{CICS}(\phi) \triangleq \sum_{i=1}^M G_{m_1m_i}(\phi) \cdot R_{m_i+1m_1}(\phi).$$

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

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

#### 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, $P(v)$, $0 \leq v < 2\pi$.

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

$$\hat{\theta}_i = \frac{h_N \sum_{j=1}^{h_i} j \cdot P(j)}{\sum_{j=1}^{h_i} P(j)}, \quad \left\{ l_i = k - h_N/2 \right\}$$

where $i = 1, \cdots, \hat{P}$. The index $k$ is one of the $\hat{P}$ highest local peaks of $P(v)$ 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 $P_{\text{MAX}}$.

#### 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 $i_1 = 1$, where $i_1$ corresponds to a counter of the peaks assigned to sources so far.

b. We also set $u_{i_1} = u_1 = \arg \max_{0 \leq \phi < 2\pi} P(u)$, i.e. the
The maximum value of \( q \) circular shift to the left. \( m \) either end wrap around, and a negative value of LPC-smoothed counterpart from which the total number of areas. We represent the envelope of the histogram with its peaks of the smoothed histogram and to suppress any noisy mates have been proposed in literature \([11]\). Here, we give we note that peak-search approaches on histograms of esti-
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, is an algorithm inspired by Matching Pursuit. The idea is to note that all these parameters were fixed, and 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 \( r_{\text{coeff}} = 0.5 \), and reverberation time \( T_{60} = 0.25s \). 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 \( \Omega \) was 344Hz with 50% overlapping in frequency, and with \( f_{\text{max}} = 4kHz \) 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), \( z_{\text{static}} = 0.05 \sum_{j=0}^{360} P(\hat{j}) \) and \( \delta = 20° \), and the LPC order used was 16. The thresholds for the Matching Pursuit-based method (MP) were \( \gamma = \{0.15, 0.14, 0.12, 0.11\} \). 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 (\( B = 1s \) in this example). We refer to these periods as “transition periods”, which we define as the time interval
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 $10^5$ 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.

ACKNOWLEDGMENT

This work is funded by the Marie Curie IAPP “AVID MODE” grant within the European Commission’s FP7.

REFERENCES