Multiple Acoustic Sources Localization Using Incident Signal Power Comparison

Daniele Salvati\textsuperscript{1} \quad Antonio Rodà \textsuperscript{1} \quad Sergio Canazza\textsuperscript{2} \quad Gian Luca Foresti\textsuperscript{1}

\textsuperscript{1} AVIRES Lab.
Dep. of Mathematics and Computer Science, University of Udine, Italy
\{daniele.salvati, antonio.roda, gianluca.foresti\}@uniud.it

\textsuperscript{2} Sound and Music Computing Group
Dep. of Information Engineering, University of Padova, Italy
canazza@dei.unipd.it

Abstract

We present a novel approach to locate multiple acoustic sources in far-field environments, in order to solve an interesting problem in different application domains, such as: audio surveillance systems and soundscape analysis frameworks. This approach aims at finding a solution to the ambiguities in Direction Of Arrivals (DOAs) combination caused by simultaneous multiple sources. The algorithm is based on two steps: the separation of the sources by means of beamforming techniques and the comparison of the Incident Signal Power (ISP) spectrum by means of a spectral distance measure. We implemented a prototype, composed by two linear arrays, that has been successfully tested in a real noisy environment.

1. Introduction

The use of microphone sensors in advanced surveillance systems is fairly recent. Microphone arrays allow to extract information about the space location of one or more sources, by means of signal processing techniques. Moreover, machine learning methods can be exploited to recognize sound events such as screams or gunshots [24]. The Acoustic Source Localization (ASL) systems are applicable in various contexts such as, for example, conferencing systems for tracking the speakers [22], beamforming applications for the reduction of noise coming from concurrent sources [8], acoustical analysis of a mechanical device [25]. A widely used approach to estimate source positions consists in two steps: in the first step, a set of Time Difference Of Arrivals (TDOAs) are estimated using measurements across various combinations of microphones; in the second step, knowing the position of sensors and the velocity of sound, the source positions can be estimated by geometric considerations and using approximation methods such as least-square techniques [21]. However, this two steps approach has been tested in many single source scenarios, but in multiple sources case it requires new considerations to correctly associate the TDOAs to the same source.

This article proposes a novel approach to solve the ambiguities caused by simultaneous multiple sources. It is based on source separation and the check of similarity among sounds. The first step consists of source separation by means of beamforming techniques, maximizing the SNR of the sources and minimizing the sounds coming from other directions. The second one involves the comparison of ISP spectrum by means of a spectral distance measure. It allows to identify sounds so that the spectrum power distance minimizes an error criterion.

In the literature, several works address the problem of multiple sources using an approach based on the tracking of the sources: in [11, 26] by means of a particle filter, also known as sequential Monte Carlo method, and in [9, 23] by means of the Kalman filter theory. Methods based on movement tracking can fail in some specific situations: i) during the initialization phase of the filter, ii) in the presence of sources whose trajectories are unpredictable (e.g. in the case of rapid changes of the velocity vector), iii) when two sources have intersecting trajectories. On the contrary, some works consider an approach without tracking. In [15, 19] are proposed solutions of multiple sources problem in near-field and reverberant environments.

Our work considers to locate multiple acoustic sources in far-field and noisy environments, in order to solve a problem that is of significant interest in applications such as audio surveillance systems and soundscape analysis frameworks. In far-field condition, from the TDOA among the microphones we can estimate the DOA of sound, but not its distance. Thus, a localization system requires a network of arrays (at least two arrays for two-dimensional space).

The paper is organized as follows. After presenting the signal model in Section 2, we describe the multiple sources
localization problem in Section 3. In Section 4 we illustrate our proposed approach. Finally, Section 5 illustrates the prototype system and some experimental results, obtained in a real-world scenario.

2. Signal model

We assume $N$ acoustic sources and $R$ arrays, each one composed by $M$ microphones, considering omnidirectional characteristics of both the sources and the microphones. The discrete-time signal received by the $m_{th}$ microphone of $r_{th}$ array in a free-field environment can be modeled as

$$x_{r,m}(k) = \sum_{n=1}^{N} \alpha_{n,r,m} s_n(k - t_{n,r} - \tau_{n,r,m}) + v_{r,m}(k)$$

(1)

where $\alpha_{n,r,m}$ is the attenuation of the sound propagation (inversely proportional to the distance from source $n$ to microphone $m$ of array $r$), $s_n(k)$ are the unknown uncorrelated source signals, $t_{n,r}$ is the propagation time from the unknown source $n$ to the reference sensor of array $r$, $\tau_{n,r,m}$ is the TDOA of the $n_{th}$ signal between the $m_{th}$ microphone and the reference of $r_{th}$ array, $v_{r,m}(k)$ is additive noise signal at the $m_{th}$ sensor, which is assumed to be uncorrelated with not only all the source signals but also with the noise observed at the other sensors.

In far-field environment the relationship between TDOA and DOA can be solved easily with geometrical considerations. Therefore, for a generic pair of microphones with TDOA $\tau_{n,r}$, DOA estimate is obtained as

$$\theta_{n,r} = \arcsin \left( \frac{\tau_{n,r} c}{d} \right)$$

(2)

where $c$ is the speed of sound and $d$ the distance between microphones.

For each source $n$ we can define the vector

$$\Theta_n = [\theta_{1,n}, \theta_{2,n}, \ldots, \theta_{R,n}]^T.$$  

(3)

that contains the DOAs estimated by each array.

In the case of $N$ sources, we can write the matrix $(R \times N)$, that contains all DOAs as

$$\Theta = \begin{bmatrix} \theta_{1,1} & \theta_{1,2} & \ldots & \theta_{1,N} \\ \theta_{2,1} & \theta_{2,2} & \ldots & \theta_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{R,1} & \theta_{R,2} & \ldots & \theta_{R,N} \end{bmatrix}.$$  

(4)

3. Multiple sources localization

DOA estimation is a crucial step of ASL systems. It can be calculated by means of three classes of algorithms: i) TDOA based method (e.g. the Generalized Cross-Correlation (GCC) [10, 14, 16], the Steered Response Power Phase Transform (SRP-PHAT) [4, 17, 27], the Multichannel Cross-Correlation Coefficient (MCCC) [3]); ii) techniques based on steered beamformer energy response [7]; iii) high-resolution spectral estimation algorithms (e.g. MUltiple SIgnal Classification (MUSIC) [1, 20]).

In presence of multiple simultaneously active sources $N$, we can estimate the DOAs angles, obtaining for each array $r$ the following vector

$$\hat{\Theta}_r = [\theta_{r,1}, \theta_{r,2}, \ldots, \theta_{r,N}]^T.$$  

(5)

where we consider the angle values in ascending order ($\theta_{r,1} < \theta_{r,2} < \theta_{r,3}$ and so on).

Then, the estimated ordered matrix $\hat{\Theta}$ is defined as

$$\hat{\Theta} = \begin{bmatrix} \theta_{1,1} & \theta_{1,2} & \ldots & \theta_{1,N} \\ \theta_{2,1} & \theta_{2,2} & \ldots & \theta_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{R,1} & \theta_{R,2} & \ldots & \theta_{R,N} \end{bmatrix}.$$  

(6)

The position of the $n_{th}$ source can be calculated by combining the DOAs estimated by the $R$ arrays for that source. The problem to face is to assign those $R$ DOAs values correctly to the $n_{th}$ source.

For sake of clarity, we consider the case with two sources and two arrays. Figure 1 shows that the two DOAs calculated by the two arrays can be combined following two different configurations: i) $\theta_{1,1} - \theta_{1,2}, \theta_{1,1} - \theta_{2,2}$; ii) $\theta_{1,2} - \theta_{2,1}, \theta_{1,1} - \theta_{2,2}$. The first configuration implies the correct localization of the sound sources, whereas the second one leads to a wrong localization of both the sources.

In general, our goal is to get the matrix $\Theta$, ordering properly the values of Eq. (6). Considering $\theta_{i,r}$ as the $i_{th}$ DOA of array $r$, the assignment to correct value of angle for unknown sources can be ambiguous, namely the exact position of element in the matrix of Eq. (4) cannot be uniquely determined

$$\theta_{i,r} \rightarrow \theta_{r,n}.$$  

(7)

The possible combinations of DOAs of matrix (6) are $(N)!^R\cdot R^{-1}$.

Figure 1. The problem of multiple sources localization.
4. Incident signal power comparison

Our approach to correctly combine the DOAs of the multiple sources is based on a similarity criterion among sources. To check this similarity we estimate for each array the ISP referring to all estimated DOAs by means of beamforming techniques. Once obtained the ISPs, we define an efficient error criterion for comparing the different possible combinations of the ISPs by means of a spectral distance measure.

Summarizing, the steps of the algorithm are: i) TDOAs and DOAs estimation; ii) source separation by means of beamforming techniques; iii) ISP comparison by means of spectral distance measure; iv) check of the most consistent target combination minimizing an error criterion; v) sources localization.

We summarize here known background information about beamforming and spectral distance measure. In particular, we briefly describe the three source separation methods to estimate the ISP and the four spectral distance functions, i.e. the tools we use to compare our experimental results. Finally we introduce the spectral distance estimation and the error criterion to estimate the more consistent target combination.

4.1. Source separation

The beamforming can be seen as a combination of the delayed signals from each microphone, so that an expected pattern of radiation is preferentially observed. In the frequency domain, the output of beamformer can be written as linear combination of the data at the \( M \) microphones of array

\[
Y(w) = wX
\]

where \( w \) is the vector of beamformer weights for steering and filtering the data, and \( X \) is the vector of \( M \) signal received by microphones. The output power spectral density of beamformer is given by

\[
E\{ |Y(w)|^2 \} = w^H E\{ XX^H \} w = w^H R_x w
\]

where \( R_x \) is the spatial correlation matrix. We define, then, the a steering vector to direct the beamformer on desire direction as

\[
a = [e^{-jw f_1(\theta)} e^{-jw f_2(\theta)} \ldots e^{-jw f_M(\theta)}]^T \]

where \( f_i(\theta) \) is a function of DOA depending on array geometry. Hence, the three beamforming techniques we use are: the Delay and Sum (DS) beamforming

\[
ISP_{DS}(w) = a^H R_x a,
\]

the Filter and Sum (FS) beamforming with Dolph-Chebyshev windows \( h \) [5]

\[
ISP_{FS}(w) = (ha)^H R_x (ha),
\]

the Minimum Variance Distortionless Response (MVDR) beamforming [2]

\[
ISP_{MVDR}(w) = \frac{1}{a^H R_x^{-1} a}.
\]

4.2. Spectral distance function

Distance measures produce measurements of dissimilarity of two sound spectra. We use the most four common Spectral Distance Functions (SDF) in order to verify how our system performance varies as a function of different equations. A classic spectral estimation method is Linear Prediction (LP) [12], where we insert the minus one to standardize the minimum to zero as all functions

\[
LP_{ij} = \frac{1}{L} \sum_{w=0}^{L-1} \left( \frac{ISP_i(w)}{ISP_j(w)} - 1 \right)
\]

where \( L \) is the number of samples of the observation time. The other functions are the Itakura-Saito (IS) distance measure [13]

\[
IS_{ij} = \frac{1}{L} \sum_{w=0}^{L-1} \left( \log \frac{ISP_i(w)}{ISP_j(w)} - 1 \right),
\]

the Root Mean Square (RMS) log [18]

\[
RMS_{ij} = \frac{1}{L} \sum_{w=0}^{L-1} \left( \log \frac{ISP_i(w)}{ISP_j(w)} \right)^2,
\]

and the COSH measure [6]

\[
COSH_{ij} = \frac{1}{L} \sum_{w=0}^{L-1} \left( \log \frac{ISP_i(w)}{ISP_j(w)} - \log \frac{ISP_i(w)}{ISP_j(w)} - 2 \right).
\]

4.3. Spectral distance estimation

Let us represent the sorted matrix of DOAs by the graph theory. Then we can express the matrix (6) and all its combinations as composed of nodes and edges, that connect pairs of vertices. We define the spectral distance estimation of two DOAs of different arrays as

\[
E_{ijkl} = |SDF[ISP_{ij}(w), ISP_{kl}(w)]|
\]

where \( i \) and \( k \) are index label of array, \( i \neq k \), \( j \) and \( l \) are index label of ordered DOAs of array. We have \( N^2R(R-1)/2 \) spectral distance measures. Then, we define these spectral distance function values as weights of edges of graph. An example of three arrays and three sources is showed in Figure 2.
For each source, identified by R nodes, we have \( R(R-1)/2 \) edges, then the number of edges for a combination of angles is \( Q = NR(R-1)/2 \). We can define the spectral distance estimation of the generic combination \( z \) as the sum of all weights of edges

\[
D(z) = \sum_{q=1}^{Q} E_{ijkl}(q).
\]

(19)

Now we define the vector of all combinations

\[
\mathbf{D} = [D(1) \ D(2) \ldots D((N!)^{R-1})]
\]

(20)

Finally, the index of the minimum value of the vector \( \mathbf{D} \) identifies the target combination

\[
\hat{z} = \arg\min_{\text{index}} \mathbf{D}.
\]

(21)

### 5. Experimental results

In the context of audio surveillance the already known ASL techniques need to be adapted to the constraints imposed by the new application scenario: i) the need to monitor sources that are moving on a two-dimensional space (the plane of a square, a street or a monitored park); ii) the need to place sensors on a different plane from that monitored, in order to avoid damage by pedestrians or vehicles; iii) the need to have a reduced number of arrays, not to invade too much the public spaces.

Thus, to test in a real noisy scenario the algorithms for the multiple sources localization, we made a prototype that has been installed on the roof of the building that houses our department at University. The prototype includes two linear arrays, each one composed by four omnidirectional microphones. The arrays are located at a distance of 11.4 m between them and a height of 12.1 m above the plane. The sample rate of digital system is 48 kHz and the microphone distance is \( d = 25 \) cm.

Basically, the prototype consists of two parallel processing lines, corresponding to the left and right arrays. The first processing step is the TDOA estimation, based on the measurement of the time difference between the signals received by different microphones. The MCCC method [3] is used to compute the TDOA, because this method allows to take advantage of the redundant information provided by multiple sensors. Besides, in order to improve the resolution of the peaks for the TDOA estimations and to minimize the influence of noise and interferences, we apply a Phase Transform (PHAT) filter [10], before calculating the MCCC. Then, knowing the array geometry, we can compute the DOAs value for each sound source. Finally, the two-dimensional coordinates of the sources can be estimated combining the DOAs at the left and right arrays with simple geometric considerations. If more than one source is identified, a beamformer and a spectral distance comparison provide a guide to solve the problem of associating the DOAs of the left array with those of the right one.

We carried out two types of experiments. The first one uses sounds with different spectral content, named test \( P \). The second one, on the contrary, uses sounds with similar spectral content, named test \( C \).

Test \( P \) is composed by eight parts (\( P1, P2, \ldots P8 \)), each one with three sources positioned in different placements. Figure 3 shows the map of the monitored area (parking lots), with the position of arrays and sources. The three sound sources are: a human voice (\( S1 \)), a hammer that strikes an iron bar (\( S2 \)) and a car (\( S3 \)).
In the various parts of test $P$ sources are positioned at increasing distances, along the y axis ($P1 − P4$) and along the x axis ($P5 − P8$); e.g., in the part $P1$ we placed $S1$ in 1, $S2$ in 2 and $S3$ in 3 (see Figure 3).

Table 1 summarizes the results, comparing the localization success rate (in percentage) with different beamforming algorithms and spectral distance functions. The localization success rate is the ratio between the number of correct combinations and the Number Of Ambiguities (NOA) for that part of the test. The audio signal was divided into 50% overlapping, Hanning-windowed frames with a length of 140 ms. NOA is the number of frames in which we have ambiguity to properly associate the DOAs to the sources, i.e. the associations are in practice wrong (see Section 3).

We consider three Frequency Ranges (FR) for the spectral distance estimation. Table 1 summarizes the results of all tests.

In test $C$, two car sounds have been used. The test was carried out by placing two car sources in 1 and 7, referring of Figure 3. In Figure 4 we compare the total localization success rate of the complex acoustic scene ($T$) with the results of the case of two cars, in order to test if our algorithm works even with similar spectral content sounds. We notice that the accuracy decreases, especially with regard to RMS and COSH functions, and this highlights the limitation of the proposed approach in case of spectrally similar sources.

![Figure 4. Comparison of the summary results T with the result C of car-car test with NOA=39. FR=[20, 675] Hz.](image)
6. Conclusions
We presented an innovative approach that addresses the problem of multiple sources localization separating the sources by means of beamforming techniques and comparing the ISP of the beamformer output by means of a spectral distance function. We evaluated the system in a real scenario, installing a hardware/software prototype on the roof of the university building and analyzing the results comparing three types of beamforming and four functions for the spectral distance estimation. The best performances are obtained with Dolph-Chebyshev beamforming and RMS spectral distance measure on frequency range between 20 Hz and the spatial aliasing frequency limit. We achieved a success rate of 92.8%. We showed the limitation of the proposed algorithm in case of sources that have a similar spectral content. As future work: in order to improve the performance in these cases, we intend to consider other audio features, as well as the ISP, to describe the evolution over time of the observed signal; we will also provide to compare our approach with a filter tracking system.

References