A REAL-TIME SYSTEM FOR MULTIPLE ACOUSTIC SOURCES LOCALIZATION BASED ON ISP COMPARISON

Daniele Salvati
AVIRES Lab.
Dep. of Mathematics and Computer Science
University of Udine, Italy
daniele.salvati@dimi.uniud.it

Antonio Rodà
AVIRES Lab.
Dep. of Mathematics and Computer Science
University of Udine, Italy
antonio.roda@uniud.it

Sergio Canazza
Sound and Music Computing Group
Dep. of Information Engineering
University of Padova, Italy
canazza@dei.unipd.it

Gian Luca Foresti
AVIRES Lab.
Dep. of Mathematics and Computer Science
University of Udine, Italy
foresti@dimi.uniud.it

ABSTRACT

The growing demand for automatic surveillance systems that integrates different types of sensors, including microphones, requires to adapt and optimize the already studied techniques of Acoustic Source Localization to meet the constraints imposed by the new application scenario. In this paper, we present a real-time prototype for multiple acoustic sources localization in a far-filed and free-field environment. The prototype is composed by two linear arrays and utilizes an innovative approach for the localization of multiple sources. The algorithm is based on two steps: i) the separation of the sources by means of beamforming techniques and ii) the comparison of the power spectrum by means of a spectral distance measure. The prototype was successfully tested in a real environment.

1. INTRODUCTION

Acoustic Source Localization (ASL) allows to extract information about the space location of one or more sources using microphone arrays and signal processing techniques. The ASL techniques are applicable in various contexts as, for example, the tracking of the speaker during a conference [1], the reduction of noise coming from concurrent sources [2] or the acoustical analysis of a mechanical device [3]. Recently, the growing demand for automatic surveillance systems (networks of video cameras integrated with other types of sensors) has sparked interest in systems capable of monitoring the presence and movement of sound sources in public places [4]. It is therefore necessary to adapt and optimize the already studied ASL techniques to meet the constraints imposed by the new application scenario: i) the far-field condition (it is often necessary to locate sources at a distance of tens of meters), in which the acoustic pressure wave can be approximated to a plane wave; ii) the need to monitor sources that are moving on a two-dimensional space (the plane of a square, a street or a monitored park); iii) the need to place sensors on a plane different from that monitored, in order to avoid damage by pedestrians or vehicles; iv) the need to have a reduced number of arrays, not to invade the public spaces in an excessive way. In fact, whereas in the near-filed case a linear array of at least three microphones should be sufficient to locate the sources position in a two-dimensional space, in the far-field case the estimation of the source position is extremely difficult, if not almost impossible, using a single array: from the Time Difference Of Arrival (TDOA) among the microphones we can estimate the Direction Of Arrival (DOA) of the sound, but not its distance. Therefore, the two-dimensional position of the source can be estimated using at least two linear arrays, by means of the triangulation of the DOA estimations (see Figure 4). As our aim is to test a network configuration with a minimum number of arrays (see the point iv above), in the next we will refer to a capture system composed by only two arrays. This system works efficiently in the case of a single source, but if there are more than one source, we have the problem of what are the correct angles to link with each others (see Figure 4).

In the literature, several works address similar problems using an approach based on the tracking of the sources: in [5] [6] [7] by means of a particle filter, also known as sequential Monte Carlo method, and in [8] [9] 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 with unpredictable trajectory (e.g. in the case of rapid changes of the velocity vector), iii) when two sources have intersecting trajectories (see Figure 1).

This article explores a new approach, which can be applied in a manner complementary to the filter-based methods, enhancing the performance of the system when the movement tracking is difficult. Our approach is based on the source separation and the comparison of similarity among sounds.

Assume that n sound sources have been identified for each array, coming from n different DOAs. Since the localization of any source in the two-dimensional plane requires the triangulation of information measured by the two arrays, we must associate each DOA estimated by the first array with one corresponding to the same source estimated by the second array. In total we have n^2 possible combinations, but we don't know a-priori which are the right n pairs. To this end, the sources estimated by each array are separated by means of beamforming techniques, maximizing the SNR of the sources and minimizing the interferences and sounds

coming from other directions. Then, we proceed to compare the signal coming from all the n directions identified by the first array with those identified by the second one. Finally, among the n^2 possible pairs, the n pairs whose signals have a greater similarity are selected.

To check the similarity of acoustic sources, we use a method based on the spectral distance measure, that allow to identify sounds with the comparison of spectral power in such a way that their difference spectrum power minimizes an error criterion. We describe the spectral distance function in Section 4.2.

The rest of this paper is organized as follows: after presenting the system architecture in Section 2, we briefly summarize the adopted algorithm for the Time Delay Estimation in Section 3. In Section 4 we illustrate how the two-dimensional position of the source can be evaluated starting from the TDOAs estimated by the two arrays. Finally, Section 5 illustrates the prototype of the real-time system and some preliminary experimental results, obtained in a real-world scenario.

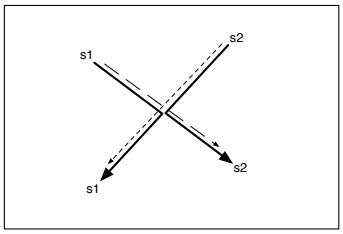

Figure 1: *Two sound sources (s1 and s2) with intersecting trajectories. Continuous lines are the real trajectories, dashed lines are the wrong tracked trajectories.*

2. SYSTEM ARCHITECTURE

To solve the problem of multi-source localization in a two dimensional space we propose the logical architecture showed in Figure 2. Basically, it 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. We use the Multichannel Cross-Correlation Coefficient (MCCC) method [10] to calculate the TDOA, because this method allows to take advantage of the redundant information provided by multiple sensors. Besides, to improve the resolution of the peaks for the TDOA estimations and minimize the influence of noise and interferences, we apply a Phase Transform (PHAT) filter [11], before calculating MCCC. Processing the TDOA information with the knowledge of the array geometry and the acoustic environment (far-field and free-field), we can calculate the DOA of the sound source. Finally, the two-dimensional coordinates of the source can be estimated combining the DOAs at the left and right arrays (see Figure 4). 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 array. In case of a stationary noisy environments, correlated among the array microphones, the ASL can be improved by means of a de-noise task [12].

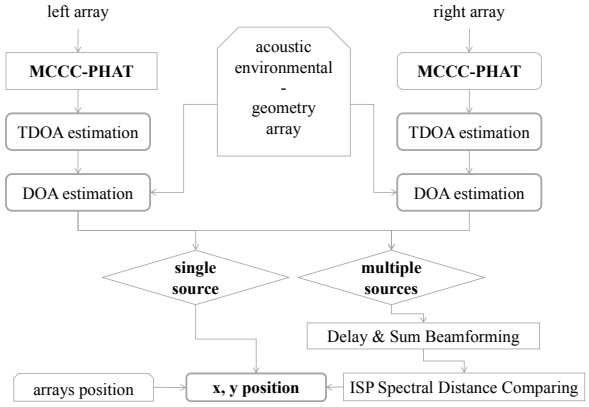

Figure 2: *The block diagram of the processor, showing the data flow of all the tasks of the experimental system implementation.*

3. TIME DELAY ESTIMATION

3.1. MCCC method

The MCCC algorithm is a spatial correlation-based method. According to [13], we consider a linear array on N ($N \geq 2$) microphones. The discrete-time signal received by the n_{th} microphone in a free-field environment with M multiple sources can be modeled as:

$$y_n[k] = \sum_{m=1}^{M} \alpha_{nm} s_m[k - t_m - F_n(\tau_m)] + v_n[k] \qquad (1)$$

where α_{nm} is the attenuation of the sound propagation (inversely proportional to the distance from source m to microphone n), $s_m[k]$ are the unknown source signals, t_m is the propagation time from the unknown source m to the reference sensor, $F_n(\tau_m)$ is the TDOA of the m_{th} signal between the n_{th} microphone and the reference, $v[k]$ is an additive noise signal at the n_{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. The function $F_n(\tau_m)$ depends on the microphone array geometry. In our case, for a linear and equispaced arrays, i.e. Uniform Linear Array (ULA), and for a single source we have:

$$F_n(\tau) = (N-1)\tau, \quad n = 2, \ldots, N \qquad (2)$$

If we consider a single source and neglect the noise terms, we have:

$$y_n[k + F_n(\tau)] = \alpha_n s[k - t] \qquad (3)$$

Therefore, $y_1[k]$ is aligned with $y_n[k + F_n(\tau_n)]$, and the new signal vector can be written:

$$\mathbf{y}[k, p] = \begin{bmatrix} y_1[k] & y_2[k+\tau] \ldots y_n[k + (N-1)\tau] \end{bmatrix}^T \qquad (4)$$

where p is a dummy variable for the hypothesized TDOA τ. The spatial correlation matrix of N microphones array is:

$$\mathbf{R}[p] = \begin{bmatrix} \sigma^2_{y_1} & r_{y_1 y_2}[p] & \ldots & r_{y_1 y_N}[p] \\ r_{y_2 y_1}[p] & \sigma^2_{y_2} & \ldots & r_{y_2 y_N}[p] \\ \vdots & \vdots & \ddots & \vdots \\ r_{y_N y_1}[p] & r_{y_N y_2}[p] & \ddots & \sigma^2_{y_N} \end{bmatrix} \qquad (5)$$

where $\sigma_{y_i}^2 = E[y_i]^2$ is the variance of signal y_i and $r_{y_i y_j}[p]$ is the cross-correlation between y_i and y_j. The spatial correlation matrix can be factored as:

$$\mathbf{R}[p] = \mathbf{\Sigma}\widetilde{\mathbf{R}}[p]\mathbf{\Sigma} \tag{6}$$

where

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{y_1} & 0 & \dots & 0) \\ 0 & \sigma_{y_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ddots & \sigma_{y_N} \end{bmatrix} \tag{7}$$

and

$$\widetilde{\mathbf{R}}[p] = \begin{bmatrix} 1 & \rho_{y_1 y_2}[p] & \dots & \rho_{y_1 y_N}[p] \\ \rho_{y_2 y_1}[p] & 1 & \dots & \rho_{y_2 y_N}[p] \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{y_N y_1}[p] & \rho_{y_N y_2}[p] & \dots & 1 \end{bmatrix} \tag{8}$$

is a symmetric matrix, and

$$\rho_{y_i y_j}[p] = \frac{r_{y_i y_j}[p]}{\sigma_{y_i}\sigma_{y_j}} \tag{9}$$

is the Pearson Correlation Coefficient (PCC) between the i_{th} and j_{th} aligned microphone signals. The MCCC algorithm can be used to estimate the TDOA between the first two microphone signals as:

$$\widehat{\tau}^{\mathrm{MCCC}} = \arg(\mathrm{local})\min_{p} \det[\widetilde{\mathbf{R}}[p]] \tag{10}$$

3.2. MCCC-PHAT method

In our system we use a filtered cross-correlation function, the Generalized Cross-Correlation (GCC) [11], the most common technique employed for TDOA estimation of microphone pairs. The GCC in the frequency domain is:

$$r_{y_1 i y_j}^{\mathrm{GCC}}[p] = \sum_{f=0}^{L-1} \Psi[f] S_{y_i y_j}[f] e^{\frac{j2\pi pf}{L}} \tag{11}$$

where L is the number of samples of the observation time, $\Psi[f]$ is the frequency domain weighting function, and the cross-spectrum of the two signals is defined as:

$$S_{y_i y_j}[f] = E\{Y_i[f] Y_j^*[f]\} \tag{12}$$

where $Y_i[f]$ and $Y_j[f]$ are the Discrete Fourier Transform (DFT) of the signals and * denotes the complex conjugate. GCC is used for minimizing the influence of uncorrelated noise and interferences, and maximizing the peak in correspondence of the time delay. In order to improve their robustness to additive noise, see [14]. The PHAT weighting function normalizes the amplitude of the spectral density of the two signals and uses only the phase information to compute the GCC:

$$\Psi_{\mathrm{PHAT}}[f] = \frac{1}{|S_{y_i y_j}[f]|} \tag{13}$$

It is widely acknowledged that GCC is able to provide consistent performance when the characteristics of the source signal change over time. Thus its performance is dramatically reduced in case of pseudo-periodic sounds.

Now we can write the new spatial correlation matrix for MCCC-PHAT:

$$\mathbf{R}^{\mathrm{PHAT}}[p] = \begin{bmatrix} 1 & r_{y_1 y_2}^{\mathrm{GCC}}[p] & \dots & r_{y_1 y_N}^{\mathrm{GCC}}[p] \\ r_{y_2 y_1}^{\mathrm{GCC}}[p] & 1 & \dots & r_{y_2 y_N}^{\mathrm{GCC}}[p] \\ \vdots & \vdots & \ddots & \vdots \\ r_{y_N y_1}^{\mathrm{GCC}}[p] & r_{y_N y_2}^{\mathrm{GCC}}[p] & \ddots & 1 \end{bmatrix} \tag{14}$$

The TDOA estimation between the first two microphones in the general case of multiple sources is:

$$\widehat{\tau}^{\mathrm{MCCC-PHAT}} = \arg(\mathrm{local})\min_{p} \det[\mathbf{R}^{\mathrm{PHAT}}[p]] \tag{15}$$

4. SOURCES LOCALIZATION

4.1. Single source localization

The location of one sound source depends on the estimation of its DOA at the microphone arrays. In a far-field condition, the DOA value θ can be calculated as:

$$\theta = \arcsin\left(\frac{\tau c}{d}\right) \tag{16}$$

where c is the speed of sound and d the distance between microphones. The assumed DOA range is: -90° +90°, where zero is in front of the array, and the microphone reference is the first from left. As showed in Figure 3, the calculation of the two-dimensional

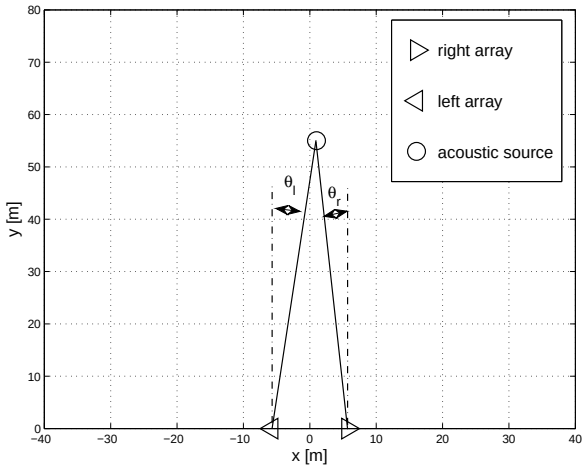

Figure 3: *Single source localization; x, y axes reference.*

position of the source is a simple trigonometric problem. However, we must consider that the two arrays are not coincident with the plane of interest, but they are placed at a certain height. Referring to Figure 3, we have to consider that the possible points identified by DOA are located on a cone surface, whose vertex is placed in the array and whose axis is the straight line joining the two arrays. Every array presents a cone: the intersection of the two cones is represented by a circumference. The intersection point between the circumference and the plane of interest is the estimation of the source distance from arrays. Hence, we consider d_a the distance

of the arrays, h the height of arrays above the plane of interest, θ_r and θ_l the DOA estimated on right and left array:

$$x = \frac{d_a}{2}\Big(\frac{\tan\theta_l + \tan\theta_r}{\tan\theta_l - \tan\theta_r}\Big) \tag{17}$$

$$y = \sqrt{\Big(\frac{d_a}{\tan\theta_l - \tan\theta_r}\Big)^2 - h^2} \tag{18}$$

4.2. Innovative approach for multiple source localization

In case of multiple sources, the wavefront of each source will arrive at each of the two arrays with a different angle. We have therefore the problem of how to correctly associate the DOA angles relative to the same source, otherwise we may incur in the location of false sources (Figure 4), i.e. in a wrong estimate of the sound sources position.

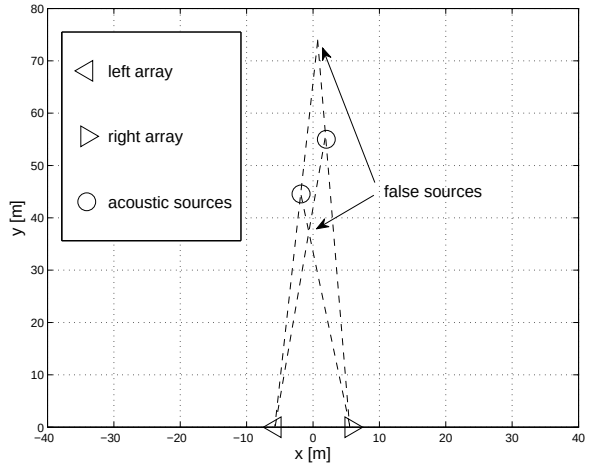

Figure 4: *The problem of multiple source localization.*

To solve this problem, we propose the use of a beamforming technique to separate sound sources, and the comparison among the power spectrum of the different sources to determine the more consistent pairs of angles. Suppose that left and right arrays detect the presence of M sources, and then M DOAs are calculated for each of the two arrays. Steering the beamformer to these M directions, M signals can be obtained for each of the arrays. The power spectrum of each signal calculated from one array is compared with that of the M signals related to the other array. Among the M^2 possibilities are then chosen the M couples with the less distance between their power spectrum.

The beamforming [15] can be seen as a combination of the delayed signals from each microphone in such a way that an expected pattern of radiation is preferentially observed. The process can be subdivided in two sub-tasks: synchronization and weight-and-sum. The synchronization task consists in delaying (or advancing) each sensor output of an adequate interval of time, so that the signal components coming from a desired direction are synchronized. The information required in this step is the TDOA estimation. The weight-and-sum task consists in weighting the aligned signals and then adding the results together to form a single output. The output signal of beamformer allows to enhance a desired signal from its detection corrupted by noise or competing sources. The Delay & Sum Beamforming (DSB) is the classical technique for realizing directional array systems. In general, the DSB output y is formed as:

$$y[k] = \frac{1}{N}\sum_{n=1}^{N} y_n[k + F_n(\tau)] \tag{19}$$

and power spectrum output, Incident Signal Power (ISP), of ULA on the desired direction is:

$$P[\theta, f] = \Big|\sum_{n=1}^{N} Y_n[f] e^{\frac{-j2\pi f(n-1)d\sin\theta}{c}}\Big|^2 \tag{20}$$

where N is the number of microphone signals, $Y_n[f]$ is the DFT of the signal, d is the distance between the microphones, c is the speed of sound, θ is the DOA of the interesting sound.

The comparison between the M^2 couples of ISPs obtained by beamforming requires the definition of a Spectral Distance Function (SDF), to be used as error criteria. Among several possibilities (for comparison see [16]), we considered five of the most used functions (presented below), in order to verify how our system performance varies as a function of SDF.

The first selected function is a simple Spectral Difference (SD):

$$SD(\theta_l, \theta_r) = \frac{1}{L}\sum_{f=0}^{L-1}(P[\theta_l, f] - P[\theta_r, f]) \tag{21}$$

where θ_l and θ_r are the DOA measured by the left and right arrays.

A classic spectral estimation method is Linear Prediction (LP) [17], where we insert the minus one to standardize the minimum to zero as the other SDF:

$$LP(\theta_l, \theta_r) = \frac{1}{L}\sum_{f=0}^{L-1}\Big(\frac{P[\theta_l, f]}{P[\theta_r, f]} - 1\Big). \tag{22}$$

The others functions are the Itakura-Saito (IS) distance measure [18]

$$IS(\theta_l, \theta_r) = \frac{1}{L}\sum_{f=0}^{L-1}\Big(\frac{P[\theta_l, f]}{P[\theta_r, f]} - \log\frac{P[\theta_l, f]}{P[\theta_r, f]} - 1\Big); \tag{23}$$

the Root Mean Square (RMS) log [19]

$$RMS(\theta_l, \theta_r) = \frac{1}{L}\sum_{f=0}^{L-1}\Big(\log\frac{P[\theta_l, f]}{P[\theta_r, f]}\Big)^2; \tag{24}$$

and the COSH measure [20]

$$\begin{aligned} COSH(\theta_l, \theta_r) = \frac{1}{L}\sum_{f=0}^{L-1}\Big(&\frac{P[\theta_l, f]}{P[\theta_r, f]} - \log\frac{P[\theta_l, f]}{P[\theta_r, f]} \\ &+ \frac{P[\theta_r, f]}{P[\theta_l, f]} - \log\frac{P[\theta_r, f]}{P[\theta_l, f]} - 2\Big). \end{aligned} \tag{25}$$

Assuming there are M sources, a list of M DOAs from each of the two arrays will be estimated; in particular, $\bar{\theta}_l = [\theta_{l_1}, ..., \theta_{l_M}]$ and $\bar{\theta}_r = [\theta_{r_1}, ..., \theta_{r_M}]$ are the vector containing the angles determined from the left and the right array respectively. To locate the sources is therefore necessary to find, among the $M!$ possibilities, the correct combination of M pairs between elements of $\bar{\theta}_l$ and $\bar{\theta}_r$. Let i one of these combinations and $i(m)$ the m_{th} pair of the

combination, we define the Spectral Distance Estimation (SDE) as the mean value of the SDFs calculated on the M pairs of the combination.

$$SDE(i) = \frac{1}{M} \sum_{m=1}^{M} |SDF(i(m))|, \tag{26}$$

and the vector Spectral Difference Mean Estimation (SDME) as:

$$\mathbf{SDME} = \begin{bmatrix} SDE(1) & SDE(2) \dots SDE(M!) \end{bmatrix} \tag{27}$$

Finally, the index of the minimum value of the vector $\mathbf{SDME}$ identifies the target combination:

$$\widehat{i} = \underset{index}{\text{argmin}} \mathbf{SDME} \tag{28}$$

5. EXPERIMENTAL RESULTS

To test in a real scenario the algorithms for the multi-sources localization, we made a prototype that has been installed on the roof of the building that houses the computer science department in Udine (see Figure 5). The prototype includes two linear arrays, each one composed by four 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 25 cm. We use for our experimental the sound sources of human voice, scream, shot gun, car, bus, horn. Though the localization is theoretically possible with just two microphones, more sensors allow ASL to make a quite robust time delay estimation, using the redundant information coming from the six TDOAs calculated for each array; in general, $N(N-1)/2$ where N is the number of microphones. Moreover, using only four microphones in each array, the computational load is not too high, both for MCCC-PHAT algorithm and DSB, and the prototype works in real-time also with an entry level personal computer. The improvement due to the redundant information coming from the four microphones can be noticed by comparing Figures 6, 7 and 8. In particular, note Figure 6 that shows the GCC-PHAT calculated for each pair of microphones: M1M2 (GCC-PHAT of microphone 1 and 2) seem to present two sources, with two peaks, instead M1M3 has two peaks very close together. This ambiguity disappears analyzing the minimum peak of the MCCC-PHAT (see Figure 7), which considers all six combinations between the four microphones. Finally, Figure 8 shows the comparison among the MCCC-PHAT calculated with a different number of microphones, from which we can see that the location of the source is more evident as the number of microphones increases.

In case of multiple sources, Figure 9 shows the presence of two peaks for each array related to the DOAs of two sources. The real sources are positioned with respect to the xy coordinates (-3,30) and (5,20) in meters. By estimating the position of local minima we can calculate the TDOAs. In this application scenario the real sources are found matching angles $\theta_{r1}\theta_{l1}$ and $\theta_{r2}\theta_{l2}$, otherwise we fall into the error of false localizing sources (Figure 10). In Figure 11 we compare ISP of the beamformer output for each DOA array. It is visible that the proper pair of angles have a similar power spectrum. Applying the SDF, we get the vector $\mathbf{SDME}$ reported in Table 1:

All the spectral distance functions minimize the vector $\mathbf{SDME}$ with the correct angles combination. The limitation of this approach relates to cases of acoustic sources with a similar spectral

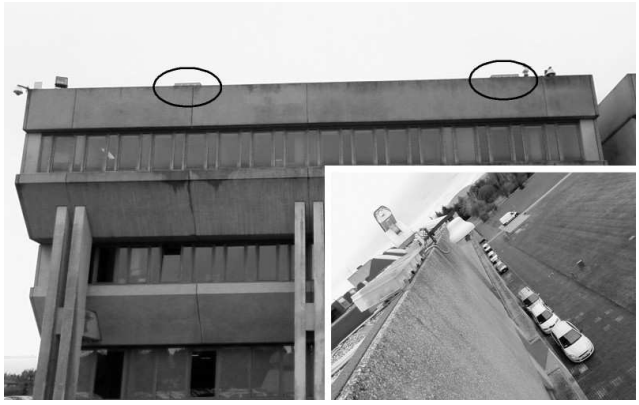

Figure 5: *The prototype installed on the roof of the university building. Inside the circles the two arrays. The microphones are protected from the weather by a waterproof box.*

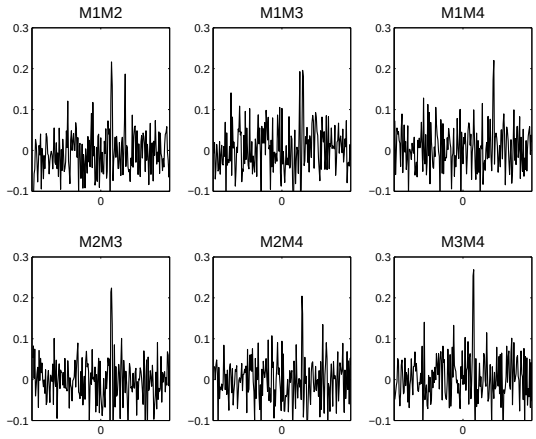

Figure 6: *Comparing the GCC-PHAT function of all microphone pairs of array.*

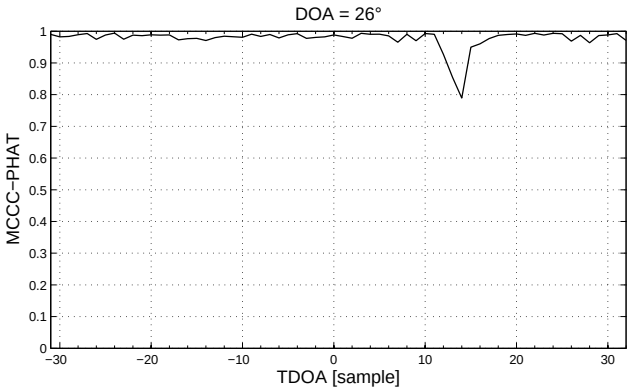

Figure 7: *MCCC-PHAT algorithm performance with a single source with reference of the same time window of Figure 6.*

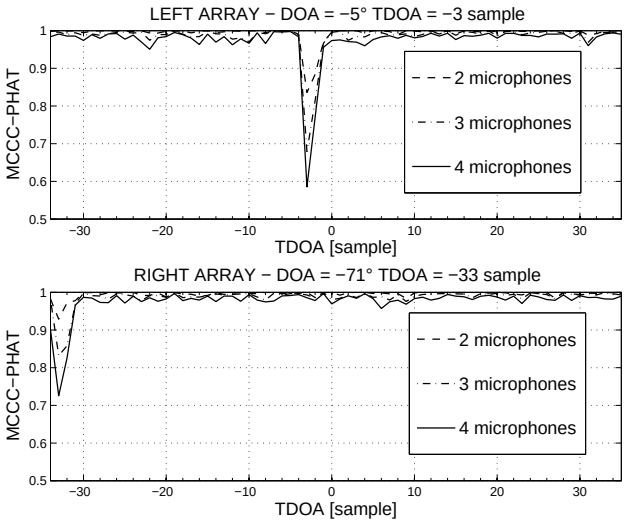

Figure 8: *Comparing the microphone number of arrays and the MCCC-PHAT algorithm performance with a single source.*

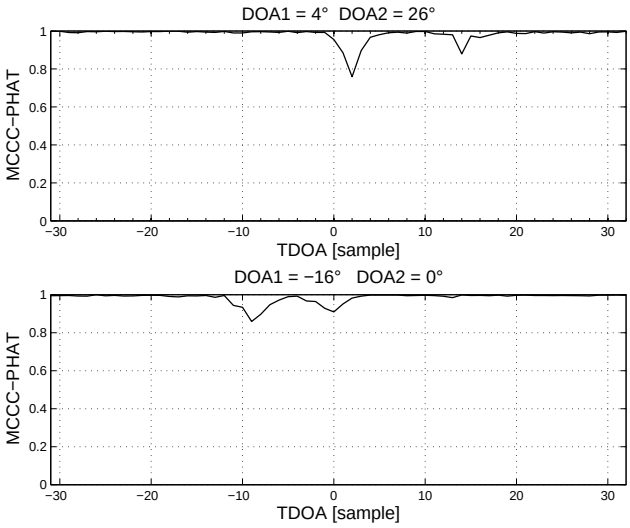

Figure 9: *MCCC-PHAT algorithm performance with two sources.*

Table 1: *SDME of multiple sources referring to Figure 9 (M=2). Each row represents the SDME calculated using a different distance function. Each column represents one of the different $M!$ angle combinations.*

SDF	$\theta_{l1}\theta_{r1} - \theta_{l2}\theta_{r2}$	$\theta_{l1}\theta_{r2} - \theta_{l2}\theta_{r1}$
SD	278	1019
LP	0.4	5.4
IS	0.6	5
RMS	7	189
COSH	7	10

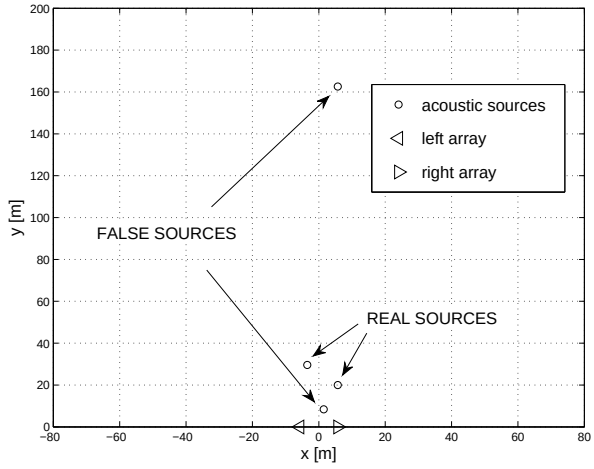

Figure 10: *Localization of two sources processing the DOA information of Figure 9: true and false sources.*

content, where it is very difficult to link the angles between arrays correctly. In Figure 12 it is shown the case of three sources, two of which have a similar spectral content, a car, located in (-5,47), and a bus, located in (0,6); the third is a human voice, positioned in (3,35). In this case, applying the task of identification, we obtain the ISP for each DOA (Figure 13) and the vector $\mathbf{SDME}$ of the six possible combinations (Figure 14). The estimated minimum value of the vector $\mathbf{SDME}$ actually identifies the correct angles combination, ($DOA1_l - DOA2_r$; $DOA2_l - DOA3_r$; $DOA3_l - DOA1_r$), but looking at Figure 13 it is well marked the similarity of human voice ISP and, on the contrary, it is difficult to check the matching pairs of the other two, so that we could make the mistake of locating the sources in a wrong position. We can see, even in Figure 13, that having index = 2, corresponding to the combination ($DOA1_l - DOA1_r$; $DOA2_l - DOA3_r$; $DOA3_l - DOA3_r$), there is a second peak due to the similarity of the two sounds, and angles ($DOA2_l - DOA3_r$) refers to the human voice. In this example the RMS log is the only spectral function that gives an incorrect result.

Once the correct DOA pairs are identified, the xy coordinates can be estimated. Localization resolution depends, regarding the DOA estimation, on the accurate calculation of cross-correlation, which is linked to, referring to equation (16), the sampling rate and the distance between microphones. Of course, this also influences the minimum resolvable distance between two sources. It is important to highlight that the distance of the microphones determines the minimum frequency beyond which spatial aliasing can occur. It means that a sound, which does not contain spectral components below the minimum frequency, cannot be uniquely localized. Instead, the resolution which are calculated the xy coordinates is related also to the distance between the arrays (Figure 15).

6. CONCLUSIONS

We presented a real-time prototype for multiple acoustic sources localization in a far-filed and free-field environment. The system is based on two linear arrays and on an innovative algorithm that

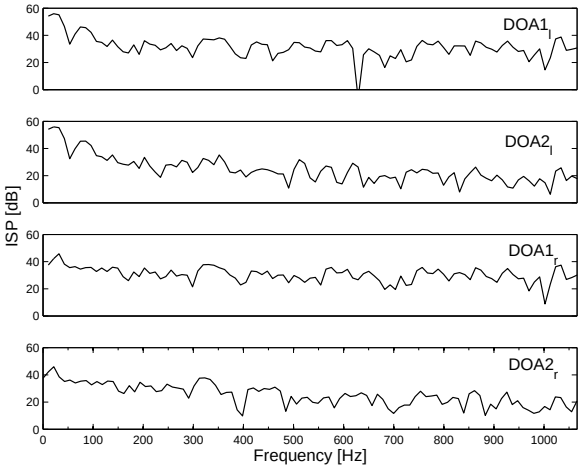

Figure 11: *Comparing the ISP on different DOA estimations of Figure 9.*

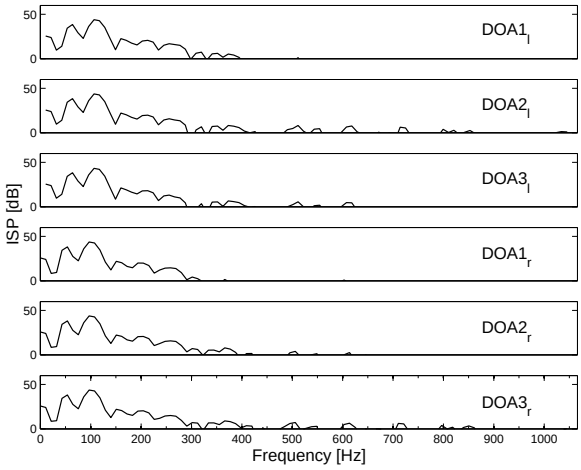

Figure 13: *Comparing the ISP on different DOA estimations of Figure 12.*

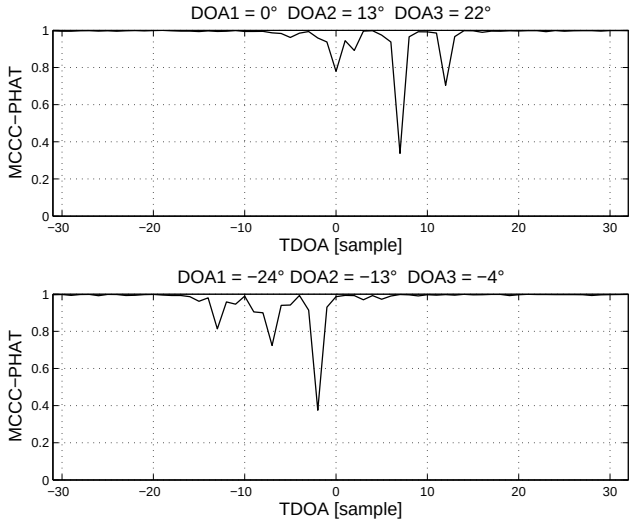

Figure 12: *MCCC-PHAT algorithm performance with three sources. Car sound* $(DOA1_l, DOA2_r)$*, human voice* $(DOA2_l, DOA3_r)$ *and bus sound* $(DOA3_l, DOA1_r)$.

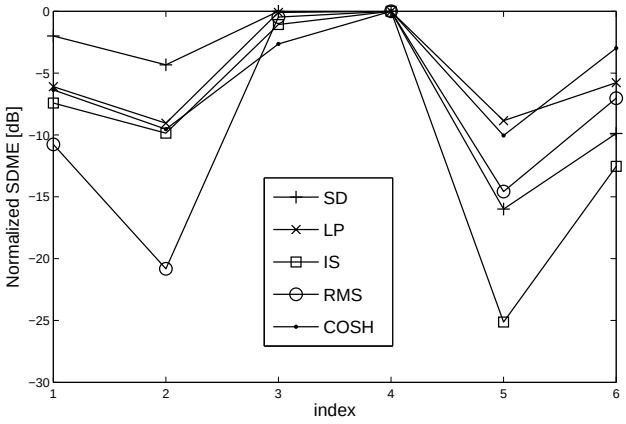

Figure 14: *Normalized SDME, comparing the SDF, referring to Figure 13.*

addresses the problem of multiple sources localization i) separating the sources by means of a Delay & Sum Beamforming and ii) comparing the Incident Signal Power of the beamformer output by means of a spectral distance function. We evaluated the system in a real scenario, installing the prototype on the roof of the university building and analyzing the sound events that happened in the parking in front of the building. For the moment, we successfully tested the functionality of the system with two and three sources. Since the localization algorithm is based on the spectrum distance, five different distance functions were assessed: all except the Root Mean Square log have correctly localized the sources both with two and three events.

Regarding the computational complexity, the algorithm requires to calculate the $M!$ elements of the vector Spectral Difference Mean Estimation, where M is the number of sound sources. In reality not all the M^2 pairs of angles are geometrically consistent, but only those that meet the condition $\theta_l > \theta_r$. Thus, the vector will have $M!$ elements only in the worst case. Nevertheless, the order of complexity of the algorithm makes it suitable only in contexts where the number of sources to be localized is limited. Alternatively, if the number of sources to localize is high, the algorithm can be integrated with a traditional tracking system, based on filtering. In this case, the localization algorithm based on the comparison of the ISPs would come into action only for those sources where the tracking system is unable to respond with a sufficient likelihood of success. Another limitation of the proposed algorithm is the location of sources that have a similar spectral content. To improve the performance in these cases it is possible to consider other audio features, as well as the ISP, describing the evolution over time of the observed signal. This will be one of the things we address on the continuation of this work.

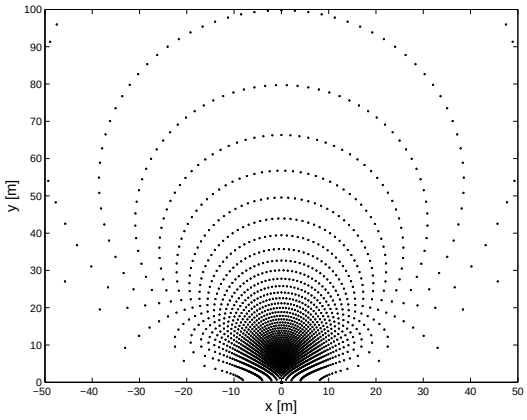

Figure 15: *The xy sample space position of plane of interest. Array distance 11.4 m and the height of arrays 12.1 m.*

7. ACKNOWLEDGMENTS

This work is partially supported by the Interreg IV Italy-Austria project n. 4697 "SRSNet - Intelligent Audio/Video Sensor Networks".

8. REFERENCES

[1] Norbert Strobel and Rudolf Rabenstein, "Robust speaker localization using a microphone array," in *In Proceedings of the X European Signal Processing Conference, volume III*, 2000, pp. 1409–1412.

[2] Yutaka Kaneda and Juro Ohga, "Adaptive microphone-array system for noise reduction," *The Journal of the Acoustical Society of America*, vol. 76, no. 1, pp. 84–84, 1984.

[3] Saligrama R. Venkatesh, David R. Polak, and Satish Narayanan, "Beamforming algorithm for distributed source localization and its application to jet noise," *AIAA journal*, vol. 41, no. 7, pp. 1238–1246, 2003.

[4] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *AVSS '07: Proceedings of the 2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, Washington, DC, USA, 2007, pp. 21–26, IEEE Computer Society.

[5] Darren B. Ward, Eric A. Lehmann, and Robert C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 11, pp. 826–836, 2003.

[6] D. Riva, D. Saiu, A. Sarti, M. Tagliasacchi, S. Tubaro, and F. Antonacci, "Tracking multiple acoustic sources using particle filtering," in *Proc. European Signal Processing Conference*, Florence, Italy, September 4-8, 2006.

[7] Wing-Kin Ma, Ba-Ngu Vo, S. Singh, and A. Baddeley, "Tracking an unknown time-varying number of speakers using tdoa measurements a random finite set approach," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 54, no. 9, pp. 3291–3304, 2006.

[8] U. Klee, T. Gehrig, and J. McDonough, "Kalman filters for time delay of arrival-based source localization," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 1–15, 2006.

[9] D.E. Sturim, M.S. Brandstein, and H.F. Silverman, "Tracking multiple talkers using microphone-array measurements," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, Munich, German, April 21-24, 1997, pp. 371–374.

[10] J. Chen, J. Benesty, and Y. Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 549–557, 2003.

[11] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, May 1976.

[12] D. Salvati and S. Canazza, "Improvement of acoustic localization using a short time spectral attenuation with a novel suppression rule," in *Proc. International Conference on Digital Audio Effect*, Como, Italy, September 1-4, 2009, pp. 150–156.

[13] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation using spatial correlation techniques," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Kyoto, Japan, September. 17-23, 2003, pp. 207–210.

[14] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," in *Proc. IEEE ICASSP*, 1994, vol. 2, pp. 273–276.

[15] H. Johnson and Dan E. Dudgeon, *Array Signal Processing: Concepts and Techniques*, Simon & Schuster, 1993.

[16] B. Wei and J. D. Gibson, "Comparison of distance measures in discrete spectral modeling," in *Proc. IEEE Digital Signal Processing Workshop*, Hunt, Texas, Oct. 15-18, 2000.

[17] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[18] R. J. McAulay, "Maximum likelihood spectral estimation and its application to narrow-band speech coding," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, no. 2, pp. 243–251, 1984.

[19] L. L. Pfeifer, "Inverse filter for speaker identification," Tech. Rep., RADCTR-74-214, Speech Communications Research Lab Inc Santa Barbara California, 1974.

[20] Jr. A. H. Gray and J. D. Markel, "Distance measures for speech processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-28 (4), pp. 380–391, 1976.