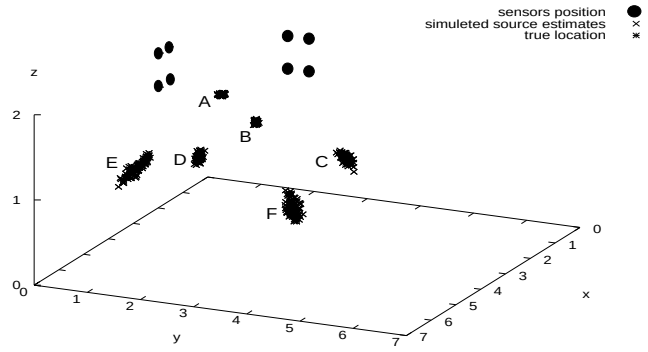

Figure 3. *Distribution of source location estimates around the true position for sources at the six points evidenced in Figure 1, when TDOA estimates are all affected by gaussian noise with equal distribution.*

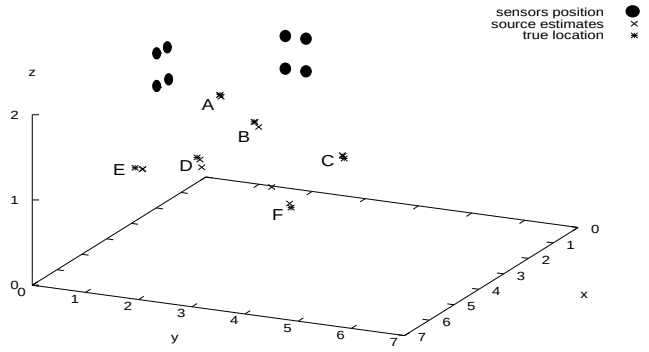

Figure 4. *Graphical representation of the results of a source location experiment. Source position estimates and true source positions (partially overlapped) are reported.*

positions using a loudspeaker.

It must be observed, however, that the assumptions about the error covariance $\mathbf{R}$ are not very realistic in a real environment, where the statistics of the TDOA estimates are affected by the noise and reverberation conditions, by distance and orientation of the microphone pairs and by directivity and frequency characteristics of the acoustic source.

Due to all these elements an assessment of performance obtained in real experiments would require a statistical analysis on a large amount of data. To give an idea of the performance obtained with our preliminary data, Table 1 reports on the average location error measured using data collected by the two array configurations using the loudspeaker as acoustic source. When using all the microphone pairs available in each square subarray (a total of 12 pairs) and an outlier detection algorithm, average performance is only slightly better than that obtained with only the four diagonal pairs. Besides, the two sensor configurations do not show a meaningful difference of location accuracy.

	$Config1$	$Config2$
12 mic pairs	0.17	0.19
4 mic pairs	0.20	0.20

Table 1. *Average location error (in meters) obtained using 12 microphone pairs (and an outlier detection algorithm) and 4 microphones pairs (diagonals) by the two array configurations.*

In the experiments with a real talker, performance was similar, although it could not be compared by means of an average error, because, in practice, the actual position of the talker's mouth is hardly determined with sufficient accuracy. Anyway it was noticed that at the most distant positions an effective location was possible only with utterances pronounced with loudness beyond a certain level and while facing the microphone array.

6. CONCLUSIONS

In this paper a preliminary work on the use of the CSP technique in three-dimensional acoustic source location has been described. The mentioned results and those obtained by simulation need to be confirmed after a large corpus collection to better characterize the statistics of the TDOA estimate errors. Nevertheless, experimental results are encouraging and show that accurate location estimates can be obtained in a real noisy and reverberant environment up to a distance of several meters.

REFERENCES

[1] M.S. Brandstein, "A Framework for Speech Source Localization Using Sensor Arrays", *Ph.D. thesis*, Brown University, Providence, RI, 1995.

[2] D.V. Rabinkin, R.J. Renomeron, A. Dahl, J.C. French, J.L. Flanagan, M.H. Bianchi, "A DSP Implementation of Source Location Using Microphone Arrays", J. Acoustic. Soc. Am., vol. 99(4), Pt. 2, p. 2503, April 1996.

[3] M. Omologo, P. Svaizer, "Acoustic Event Localization using a Crosspower-Spectrum Phase based Technique", *Proc. IEEE ICASSP*, Adelaide 1994, vol. 2, pp. 273-276.

[4] M. Omologo, P. Svaizer, "Use of the Crosspower-Spectrum Phase in Acoustic Event Location", to appear in *IEEE Trans. on Speech and Audio Processing.*

[5] G. C. Carter (ed.) "*Coherence and Time Delay Estimation*", IEEE Press, New York (1993)

[6] C. H. Knapp, G. C. Carter, "The Generalized Correlation Method for Estimation of Time Delay", IEEE Trans. on Acoustics, Speech and Signal Processing, vol. ASSP-24, n. 4, August 1976.

[7] D. J. Torrieri, "Statistical Theory of Passive Location Systems" IEEE Trans. on Aerospace and Electronic Systems, vol. AES-20, n. 2, March 1984.

[8] Y. T. Chan, K. C. Ho, "A Simple and Efficient Estimator for Hyperbolic Location", IEEE Trans. on Signal Processing, vol. SP-42, n. 8, February 1994.

[9] M. Omologo, P. Svaizer, "Acoustic Source Location in Noisy and Reverberant Environment using CSP Analysis", *Proc. IEEE ICASSP*, Atlanta 1996, vol. 2, pp. 921-924.

[10] E.E. Jan, "Parallel Processing of Large Scale Microphone Arrays for Sound Capture" *Ph.D. thesis*, Rutgers University, New Brunswick, NJ, 1995.

Simulation experiments [9] showed that there are threshold values for the allowable SNR and reverberation time T_{60} of a given environment. Over these values, the TDOA estimation procedure tends to produce unreliable delay estimates. The use of a redundant set of microphone pairs from which to derive TDOA estimates, together with the strategy of discarding the least reliable data, or those that do not agree with the majority (outliers) [10], improves the performance in such critical conditions.

A large separation between two sensors provides accurate resolution in estimating a wavefront angle of arrival. However, TDOA estimation becomes unreliable when the signals are too dissimilar. Dissimilarities are induced both by noise and by reverberation phenomena. Also source directionality may introduce further disparity at two distant sensor positions. For this reason, the sensors of a microphone pair, for which TDOA has to be estimated, should not be too far apart. This is especially true in enclosures with significant levels of noise and reverberation. Therefore delays should be estimated only locally and not all relative to a single microphone.

5. 3-D LOCATION EXPERIMENTS

Experiments were performed in a $7m$ by $10m$ by $3m$ room characterized by a reverberation time $T_{60} \simeq 0.35\ s$ and in which ventilation noise was produced by some workstations. The microphone array consisted in 8 PZM omnidirectional sensors organized into two orthogonal squares and connected to a DT3818 Data Translation acquisition board installed on a PC under the Linux operating system. The objective was to assess the performance of the location system when a talker was active in several positions of the room. The evaluation of the accuracy of the system is a critical task because a talker is not a point source and its actual "position" is hardly determined and controllable. For this reason a preliminary data collection was carried out by using a loudspeaker as acoustic source. Apart from its physical dimension, the positioning of this type of source can be much more accurate, stable and reproducible. Furthermore it is possible to reproduce exactly the same acoustic stimuli in different room positions. Six representative positions were chosen inside the room and for each of them the loudspeaker was positioned at two different heights (i.e. at 110 cm and at 160 cm above the floor). Figure 1 illustrates a map of the room used in the experimentations and shows the positions where the acustic stimuli were produced, and the positions of microphones and of noise generators (i.e. computers).

Besides the reproduction of a short utterance pronounced by a male speaker, also a sequence of white noise was generated in each position. This allowed to determine an upper bound of performance, since white noise is the "ideal" signal for estimating the TDOA with an approach based on phase information.

Acoustic data were collected using two slightly different microphone configurations (see Figure 2). Configuration 1 was composed of two squares of microphones having side of 0.4 m and centered on coordinates (1.7,0,1.6) and (0,1.7,1.6). Configuration 2 was composed of a square of 0.3 m centered on (1.71,0,0.95) and a square of 0.5 m centered on (0,1.75,1.6).

A second data corpus was collected with a stationary

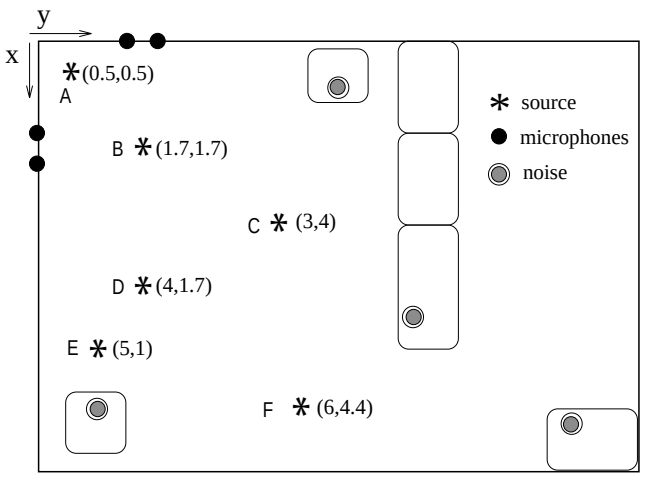

Figure 1. *Overhead view of the experimental room (7m x 10m x 3m). Positions of microphones, signal sources and noise sources are evidenced.*

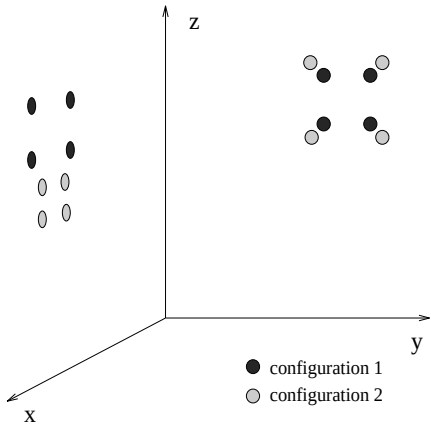

Figure 2. *Microphone array arrangement for the two configuration used in the experiments.*

talker standing on the same x and y coordinates of the previous six positions, while pronouncing a short utterance.

A region of theoretical location uncertainty was estimated for each source position by computing the covariance matrix $\mathbf{Q}$ defined by equation (10). The error covariance matrix $\mathbf{R}$ was assumed to be a diagonal matrix with equal diagonal elements, i.e. the delay estimation errors of all the microphone pairs are assumed to be independent and equally distributed with standard deviation equal to one sampling interval.

The diagonal elements of the matrix $\mathbf{Q}$ correspond to the variance of the estimate along the coordinate axes x, y and z, while its eigenvectors and eigenvalues represent orientation and amplitude of the axes of an ellipsoid centered on the exact source position.

Figure 3 is the result of a simulation of configuration 1 and shows how source location estimates are distributed around the actual source positions, in the case of the coordinates evidenced in Figure 1 and for a height $z = 110$ cm. Figure 4 illustrates graphically the result of one of the experiments of source location performed at the same

nals. This corresponds to using only the phase information in the Crosspower Spectrum of the two signals. However, phase information is correct only at those frequencies where the desired signal is dominant over interferences. Therefore in absence of specific knowledge, a suitable compromise consists in giving higher weight to the cross-spectral component exhibiting higher energy. The phase difference amount (i.e. the frequency domain counterpart of time delay) can be extracted directly from frequency domain with a best linear fitting of phase slope [1] or after reverting again into time domain [3]. In the experiments described in the following we have applied the latter method. Denoting with $s_i(n)$ and $s_k(n)$ the discrete-time sequences obtained by sampling the signals acquired by microphones i and k, the TDOA estimate between the two channels is obtained as the index at which the sequence $phc_{ik}(l)$ of *phase correlation*:

$$phc_{ik}(l) = DFT^{-1}\left\{\frac{DFT\{s_i(n)\}DFT\{s_k(n)\}^*}{|DFT\{s_i(n)\}|\,|DFT\{s_k(n)\}|}\right\} \quad (1)$$

assumes its maximum value. A local interpolation is then used to achieve sub-sample resolution.

Starting from a set of delay estimates, each characterized by its own figure of reliability, the Maximum Likelihood (ML) estimation of the source position is accomplished by solving a set of nonlinear equations relating microphone positions and observed delays [7]. Iterative methods (e.g. gradient search) are then applied to derive a least square estimate. A considerable saving in computations may be obtained by the use of sub-optimal closed-form approximations [8][1].

3. AN ITERATIVE SOLUTION

Let us consider a microphone array composed of M sensors and select a set of N microphone pairs MP=$\{(m_i', m_i'')\}$ $i = 1..N$. Let the coordinates of the generic microphone m be indicated by the vector $\mathbf{p}_m = [x_m, y_m, z_m]^T$ and the coordinates of the acoustic source s by the vector $\mathbf{p}_s = [x_s, y_s, z_s]^T$. The $N \times 1$ vector of TDOA estimations $\hat{\mathbf{d}}$ associated to the set MP and to the source s can be expressed as:

$$\hat{\mathbf{d}} = \mathbf{d}(\mathbf{p}_s) + \mathbf{n}. \quad (2)$$

Here $\mathbf{d}$ is the $N \times 1$ vector representing the theoretical delays, and $\mathbf{n}$ is the vector of estimation errors. The i-th component of $\mathbf{d}$ is given by:

$$d_i(\mathbf{p}_s; \mathbf{p}_{m_i'}, \mathbf{p}_{m_i''}) = \left[\frac{|\mathbf{p}_s - \mathbf{p}_{m_i'}|}{c} - \frac{|\mathbf{p}_s - \mathbf{p}_{m_i''}|}{c}\right] \quad (3)$$

where c is the speed of sound.

Assuming $\mathbf{n} \in \mathcal{N}(\mathbf{0}, \mathbf{R})$, where $\mathbf{R}$ is the error covariance matrix, the ML estimate of the source position based on the TDOA vector $\hat{\mathbf{d}}$ is obtained by deriving the coordinates $\mathbf{p}_s$ that maximize the probability:

$$Pr[\hat{\mathbf{d}}|\mathbf{p}_s] = \frac{1}{(2\pi)^{N/2}|\mathbf{R}|^{1/2}} e^{-\frac{1}{2}[\hat{\mathbf{d}}-\mathbf{d}(\mathbf{p}_s)]^T\mathbf{R}^{-1}[\hat{\mathbf{d}}-\mathbf{d}(\mathbf{p}_s)]} \quad (4)$$

This solution represent the minimum variance unbiased estimate. If we assume that $\mathbf{R}$ is independent of $\mathbf{p}_s$, then the solution is the position that minimizes the quantity:

$$H = [\hat{\mathbf{d}} - \mathbf{d}(\mathbf{p}_s)]^T\mathbf{R}^{-1}[\hat{\mathbf{d}} - \mathbf{d}(\mathbf{p}_s)]. \quad (5)$$

The delay d_i in equation (3) is a nonlinear function of the position $\mathbf{p}_s$, however the vector $\mathbf{d}$ can be expressed in a Taylor series in the neighborhood of a position $\mathbf{p}_0$ as:

$$\mathbf{d}(\mathbf{p}) \simeq \mathbf{d}(\mathbf{p}_0) + \mathbf{G} \cdot (\mathbf{p} - \mathbf{p}_0) \quad (6)$$

Here $\mathbf{G}$ is the $N \times 3$ gradient matrix having ∇d_i as i-th row:

$$\mathbf{G} = \begin{bmatrix} \frac{\partial d_1}{\partial x} & \frac{\partial d_1}{\partial y} & \frac{\partial d_1}{\partial z} \\ \vdots & \vdots & \vdots \\ \frac{\partial d_N}{\partial x} & \frac{\partial d_N}{\partial y} & \frac{\partial d_N}{\partial z} \end{bmatrix}. \quad (7)$$

At the generic position $\mathbf{p} = [x, y, z]^T$ we have:

$$\frac{\partial d_i}{\partial x} = \frac{1}{c}\left[\frac{(x - x_{m_i'})}{|\mathbf{p} - \mathbf{p}_{m_i'}|} - \frac{(x - x_{m_i''})}{|\mathbf{p} - \mathbf{p}_{m_i''}|}\right] \quad (8)$$

and similarly for $\frac{\partial d_i}{\partial y}$ and $\frac{\partial d_i}{\partial z}$.

The source location can be performed with a gradient search, as progressive approximation starting from an initial point $\mathbf{p}_0$ and applying the following update at the v-th iteration [7]:

$$\hat{\mathbf{p}}_{v+1} = \hat{\mathbf{p}}_v + (\mathbf{G}_v^T\mathbf{R}^{-1}\mathbf{G}_v)^{-1}\mathbf{G}_v^T\mathbf{R}^{-1}(\hat{\mathbf{d}} - \mathbf{d}(\hat{\mathbf{p}}_v)) \quad (9)$$

The covariance matrix $\mathbf{Q}$ of a location estimate is computed starting from its associated gradient matrix $\mathbf{G}$ as:

$$\mathbf{Q} = (\mathbf{G}^T\mathbf{R}^{-1}\mathbf{G})^{-1}. \quad (10)$$

The matrix $\mathbf{Q}$ can be used to predict the variance of the obtainable location estimate for a source placed at a given point and for a given microphone array arrangement. Its eigenvectors and eigenvalues determine the shape of the region of location uncertainty at every point of interest.

4. EFFECT OF NOISE AND REVERBERATION

Diffuse noise, i.e. noise with low spatial coherence, reduces the Signal-to-Noise Ratio (SNR) of the frequency bands where it is mainly concentrated. However, this does not bias significantly the component of the Crosspower Spectrum phase related to the dominant acoustic source, that, instead, is assumed to emit wavefronts with clear directional characteristics. Noise components with high spatial coherence (direct-path noise), on the contrary, may act as competitive sources and introduce ambiguity in the delay estimation procedure that assumes a single dominant acoustic source active at every instant.

In a reverberant environment, many reflected wavefronts reach the sensor after the direct one. Even if the reverberation time is low, multipath effects may arise as effect of reflection on surrounding surfaces. This makes TDOA estimation for direct wavefront more difficult and in some cases unfeasible (e.g. if constructive interference of reflections overwhelms the direct-path energy).

ACOUSTIC SOURCE LOCATION IN A THREE-DIMENSIONAL SPACE USING CROSSPOWER SPECTRUM PHASE

Piergiorgio Svaizer *Marco Matassoni* *Maurizio Omologo*
IRST-Istituto per la Ricerca Scientifica e Tecnologica I-38050 Povo di Trento (Italy)

ABSTRACT

A microphone array can be used to locate a dominant acoustic source in a given environment. This capability is successfully employed to locate an active talker in teleconferencing or other multi-speaker applications. In this work the source location is obtained in two steps: 1) a Time Difference Of Arrival (TDOA) computation between the signals of the array; 2) an "optimal" source location based on the interchannel delay estimates and on a geometrical description of the sensor arrangement. The Crosspower Spectrum Phase technique was used for TDOA estimation, while a Maximum Likelihood approach was followed to derive the source coordinates. Source location experiments in a three-dimensional space were performed by means of an array of 8 microphones. For this purpose both a loudspeaker and a real talker were used to collect data in a large noisy and reverberant room.

1. INTRODUCTION

The spatial sampling of a sound field, performed by means of several properly distributed microphones, allows to derive information about the position of the acoustic sources emitting in a given environment. This principle, widely exploited in passive underwater sonar, has found successful employment in microphone array processing as well, contributing to the development of applications such as teleconferencing, acoustic surveillance and hearing aids.

Automatic location of the active talker is of particular interest in a teleconferencing system, where it would be desirable to use a self-aiming video camera. The information about the position of the active talker is also necessary to achieve a satisfactory spatial selectivity in the picking-up of the speech message by means of a beamformer.

Source location based on Time Delay Of Arrival (TDOA) has been shown to be effectively implementable and to provide good performance even in moderately reverberant environment and in noisy conditions [1] [2].

An important issue in this type of location algorithms is the TDOA estimation technique, that must produce highly accurate and rapidly updated inter-channel delay estimates. Besides, a suitable geometrical arrangement of the sensor is essential to accomplish an adequate "resolution power" in the three-dimensional space. Finally, the location criterion must take into account all the cues implied by a (possibly redundant) set of delay estimates in order to derive the most likely source coordinates.

This work reports on the activity carried out at IRST laboratories in order to develop a real time 3-D source location system, based on an 8-channel acquisition setup and the Crosspower Spectrum Phase (CSP) method for time delay estimation [3] [4].

2. SOURCE LOCATION

Various approaches can be considered to solve the problem of acoustic source location. Methods based on the eigenvector analysis of the spatial covariance matrix of the signals acquired by an array of sensors, have been deeply investigated and widely applied, in particular in the case of multiple narrow-band sources.

When the signals of interest are wide-band and no hypothesis can be assumed about the statistics of the desired source and of the interfering noise, the problem must be addressed with other approaches. In particular, two types of techniques have been applied to solve the problem of talker location in a room environment.

The first type includes the methods that look for the position from which a maximum acoustic power is received by an array of sensors (or a maximum of phase alignment between the input channels is detected). In order to find this position it is necessary to scan a grid of admissible positions by means of a steerable beamformer. The disadvantages of this approach consist in poor temporal and spatial resolutions and in a heavy computational load.

The second type of techniques comprises the TDOA based methods, that derive the source coordinates starting only from the mutual delays occurring between the direct wavefront arrivals at the microphones. A lot of literature exists about time delay estimation techniques [5]. The most immediate approach to computing the delay between the signals acquired by two microphones is to find a maximum of cross-correlation between the two signals. This only gives good results, in practice, if the source emits white noise and therefore the cross-correlation approximates a delta pulse centered on the delay of interest. In the general case, the autocorrelation of the source signal modifies the shape of the cross-correlation peak. This may be particularly unconvenient if the source signal is non-stationary and there are reverberation phenomena (as for a talker in an enclosure).

If the autocorrelation of the source signal and that of the interfering noise are known, they can be exploited to design an optimal filter that facilitates the delay estimation (Generalized Cross-Correlation approach [6]). If no a priori knowledge about the statistics of the involved signals is available, an effective approach is to whiten the input sig-