ACOUSTIC SOURCE LOCATION IN NOISY AND REVERBERANT ENVIRONMENT USING CSP ANALYSIS

Maurizio Omologo and Piergiorgio Svaizer

IRST-Istituto per la Ricerca Scientifica e Tecnologica
I-38050 Povo di Trento (Italy)

ABSTRACT

A linear four microphone array can be employed for acoustic event location in a real environment using an accurate time delay estimation.

This paper refers to the use of a specific technique, based on Crosspower Spectrum Phase (CSP) analysis, that yielded accurate location performance. The behavior of this technique is investigated under different noise and reverberation conditions.

Real experiments as well as simulations were conducted to analyze a wide variety of situations. Results show system robustness at quite critical environmental conditions.

1. INTRODUCTION

The development of microphone array technology [1] is a fundamental step for the advancement of hands-free speech recognition [2], teleconferencing and acoustic surveillance systems [3]. These applications require capabilities such as automatic talker location [4] and spatially selective pick-up of sound [5], which can be performed by the processing of acoustic signals supplied by a multichannel acquisition setup.

Acoustic source location by means of microphone arrays requires very accurate Time Delay Estimation (TDE) techniques in order to achieve a satisfactory precision. This is related to the high speed of sound pressure propagation when considered in relationship with the distance between microphones placed in a given room. Due to phenomena like reverberation and environmental additive noise, this distance cannot be made very large without reducing the coherence between the signals whose mutual delay has to be estimated.

The Crosspower-Spectrum Phase (CSP) based technique has been compared with other delay estimators such as cross-correlation and LMS filtering [3] and has proven to be the most effective for generic broad-band signal (e.g. speech), even when employed in noisy and reverberant [6] environments. This technique is a generalized cross-correlation method [7] that takes advantage from the information about phase differences between signals, contained in their cross-power spectrum.

This paper represents an extension of the subject discussed in [3] and addresses the issue of location accuracy yielded by a linear array of four microphones. Parameters like inter-sensor distance, signal to noise ratio, and reverberation time are taken into consideration in order to assess their impact on source location.

The remainder of the paper is organized as follows. Section 2 focuses on a formal definition of the source location problem; it introduces the Coherence Measure (CM) representation and its use in the location procedure. Section 3 describes the Crosspower-Spectrum Phase analysis as well as the corresponding CM representation. Section 4 reports on the location experiments conducted in a real environment and a consequent investigation on system capabilities evaluated by simulation under different noisy and reverberant conditions. Finally, Section 5 gives some conclusions and introduces issues that deserve to be investigated next.

2. PROBLEM DEFINITION

2.1. Signal Model

For a given source signal $r(t)$, propagated in a generic noisy and reverberant environment, the signal $s_i(t)$ acquired by the acoustic sensor "i", can be expressed as:

$$s_i(t) = h_i(t) * r(t) + n_i(t) \tag{1}$$

where $*$ denotes convolution, $h_i(t)$ is the acoustic impulse response between the source and the output of the i-th microphone and $n_i(t)$ is an additive noise.

We also indicate with δ_{ik} the relative delay of direct wavefront arrival between microphones "i" and "k".

2.2. Coherence Measure

Information on mutual delay between signals can be associated to a Coherence Measure function $C_{ik}(t, \tau)$ that expresses, for a hypothesized delay τ, the similarity between segments (centered at the time instant t) extracted from two generic signals s_i and s_k. During source emission this function is expected to have a prominent peak at the delay $\tau = \delta_{ik}$, corresponding to the direction of wavefront arrival. For each couple of microphones, a bidimensional representation of the CM function can be conceived (see Fig. 1 and Fig. 2). In this representation horizontal axis is referred to time, vertical axis is referred to delay and the coherence magnitude is represented through a grey scale. Both for moving and for stationary sources, the CM function can be exploited to derive the source position. The experiments described in the following, are performed using a sampling frequency of 48 kHz and a CM analysis step of 10.67 ms.

0-7803-3192-3/96 $5.00©1996 IEEE

2.3. Source location procedure

For a stationary source, starting from Coherence Measure $C_{ik}(n,l)$ for all physically admissible delays l ($-l_{MAX} <= l <= l_{MAX}$) and signal frames n ($1 <= n <= N$) corresponding to an acoustic event, a lag can be estimated with one sample precision as follows:

$$\widehat{l}_{ik} = \arg\max_{l}[\sum_{n=1}^{N} C_{ik}(n,l)]. \quad (2)$$

A sub-sample refinement of this estimate is then provided by a suitable interpolation.

Once given two delay estimates, obtained by processing two microphone pairs, the source position is computed as crossing point between the directions associated to these delays. If three or more microphone pairs are available the location procedure is modified introducing a weighting of each delay estimate according to its estimation uncertainty.

3. CROSSPOWER-SPECTRUM PHASE ANALYSIS

The Generalized Cross Correlation between $s_i(t)$ and $s_k(t)$ is defined as:

$$R_{ik}^{(g)}(t,\tau) = \int_{-\infty}^{+\infty} \psi(t,f) G_{ik}(t,f) e^{j2\pi f\tau} df \quad (3)$$

where $G_{ik}(t,f)$ is the Crosspower Spectrum at instant t and $\psi(t,f)$ is a frequency weighting filter.

A way to sharpen the cross correlation peak in the generic broad-band case is to "whiten" the input signals: the choice

$$\psi(t,f) = \frac{1}{|G_{ik}(f)|} \quad (4)$$

leads to the so-called Phase Transform technique [7]. This corresponds to using only phase information to obtain the Coherence Measure:

$$R_{ik}^{(p)}(t,\tau) = \int_{-\infty}^{+\infty} \frac{G_{ik}(t,f)}{|G_{ik}(t,f)|} e^{j2\pi f\tau} df. \quad (5)$$

Unless signals are strictly narrow-band, and supposing there is no reverberation, this CM approaches at every instant t a delta function centered at the delay δ_{ik}. The presence of additive noise produces a distortion of phase information in the Crosspower Spectrum. As a consequence the CSP-CM is altered and exhibits a peak fading depending on the instantaneous SNR. Figure 1 shows the CSP-CM of two speech signals corrupted by an additive white gaussian noise (average SNR is 30 dB). The CM peak is centered on the interchannel delay of about 14 samples and fades in correspondence of low SNR frames.

When reverberation is present, many reflected wavefronts reach the sensors after the direct wavefront. This makes TDE for direct wavefront more difficult and in some cases impossible (e.g. when there are phenomena of substantial constructive interference between reflections). Nevertheless, the sharp peaks of the CSP-CM are an advantage over other TDE techniques even in the case of reverberant environment. Unless the amount of reverberation is too high, the accuracy of the CM peak associated to the relative delay of the direct wavefronts is fairly unaffected by the arrival of reflected wavefronts.

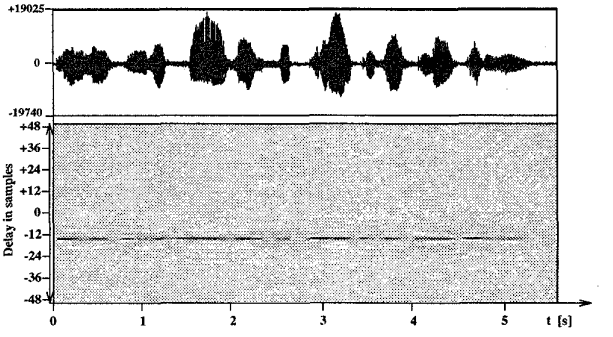

Figure 1: *Coherence Measure computed for a microphone pair with an average SNR of 30 dB. The upper part shows one of the signals.*

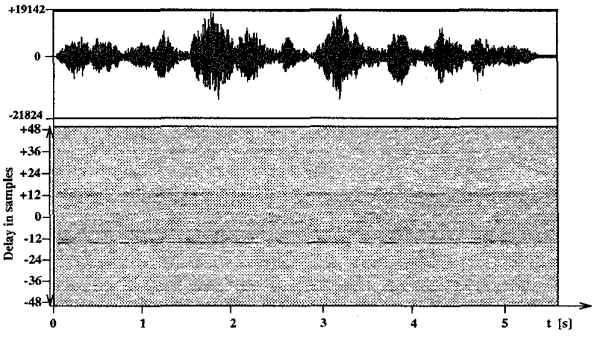

Figure 2: *Coherence Measure computed for a microphone pair with a reverberation time of 0.3 seconds. The upper part shows one of the signals.*

In Figure 2 speech signal is propagated in an environment having a reverberation time T_{60} of about 0.3 seconds. The consequent CSP-CM is characterized by a scattering of secondary peaks due to multipath reflections, but the primary peak is correctly centered on the direct wavefront delay (about 14 samples).

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

A set of real experiments was conducted to evaluate the location capabilities provided by the CSP-based technique. For this purpose, a real-time location system was used, that employs an array of 4 microphones organized into two pairs.

The system was installed in a room of $6m$ x $10m$ x $3m$. Location experiments were limited to a 2D subspace of $6m$ x $6m$ inside the room. The SNR due to background noise was estimated between 10 and 20 dB for the different stimuli and positions. A reverberation time T_{60} of about 0.3 seconds was estimated starting from impulsive response measurements performed using maximum-length pseudo random sequence as excitatory signal [8].

4.2. Array Geometry

Simulation Experiments

Simulation experiments were initially realized to establish theoretical location performance in a region of size equal to that used for real experiments, when an error-affected time delay estimation is imposed. Random errors were generated according to a Gaussian probability distribution, and added to the theoretical delays associated to uniformly distributed source positions in an area of 6m x 6m.

Figure 3 shows the consequent average location error obtained with equispaced arrays of different length and error standard deviation in the range from 0.01 to 0.5 samples (at 48 kHz). Curve asymptots for large microphone distances are due to limitation of the region where source positions are simulated. Figure 4 reports on analogous results when using two microphone pairs at different distances from each other (distance between microphones of each pair was set equal to 30 cm).

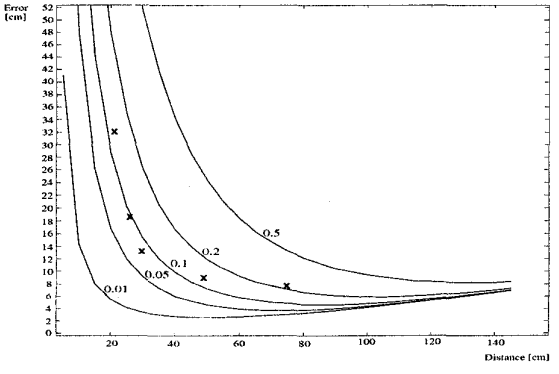

Figure 3: *Average location error obtained by simulation of an equispaced array of 4 microphones, given an error-affected TDE with standard deviation values of the gaussian error distribution in the range [0.01..0.5]. X axis represents the distance between microphones. Source positions are uniformly distributed in an area of 6m x 6m. Experimental system performance is reported as well and displayed as "x".*

Real Signal Experiments

In real experiments, a rigorous placement of the sensors turned out to be crucial to achieve unbiased results. The acoustic stimuli were generated by a loudspeaker installed on a suitable support, carefully positioned in order to have an accurate source placement as well (in this case a realistic tolerance could be 3cm x 3cm).

Each experiment consisted in generating a set of acoustic stimuli (speech, whistle, explosion) through the loudspeaker in the following 15 room positions: (-1.8, 0.5), (-1.8, 1.5), (0, 1.5), (1.5, 1), (2.5, 1.5), (3.5, 2.75), (2, 2.75), (0, 2.75), (-2.3, 2.75), (-1.5, 4), (0, 4), (2, 4), (2.5, 5), (0, 5.5), (-2, 5.5).

The first set of experiments was devoted to assess system performance in the case of equispaced microphone arrays. In Figure 3 location errors experimentally obtained using microphone separation ranging from 20cm to 75cm

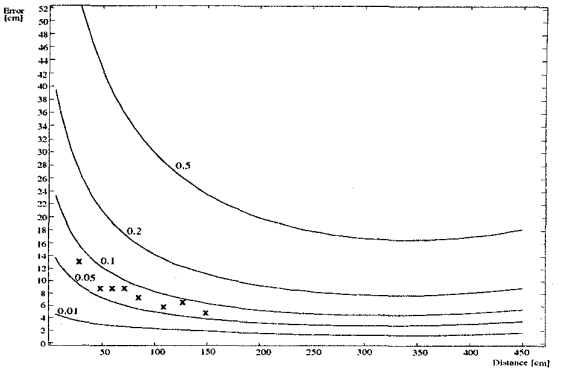

Figure 4: *Average location error obtained by simulation of a two-microphone-pair array, given standard deviation values of the gaussian error distribution in the range [0.01..0.5]. Intra-pair distance is set to 30cm. X axis represents inter-pair distance. Source positions are uniformly distributed in an area of 6m x 6m. Experimental system performance is reported as well and displayed as "x".*

are marked by "x". A comparison between experimental results and simulation curves in Figure 3 shows that the obtained performance corresponds to what is provided by a simulation characterized by a TDE error standard deviation proximate to 0.1.

A second set of experiments was conducted to further investigate the influence of the array geometry, maintaining a constant separation between the microphones of each pair, and changing the distance between pairs from 30cm to 150cm. Again Figure 4 shows that results correspond to those expected in the case of an error with a standard deviation of about 0.1.

4.3. Noise and Reverberation

A further set of simulations was realized to investigate the effects of noise and reverberation on TDE and location performance, using a speech sentence as source signal. Exploiting a geometrical model of the experimental room, the image method [9] (modified to take into account non-integer arrival times) was employed to derive the room impulse responses between any source position and each microphone of the array. In this experiment the distances between microphones were set to a 30-90-30 cm configuration and the reflection coefficient β of the walls was varied in the range 0-0.95, from anechoic to highly reverberant (T_{60}=1.25 s). An additional parameter taken into consideration in the experiment was the SNR, computed as ratio between energy of speech and energy of an additive white gaussian noise.

For each noise and reverberation condition, a uniform random distribution of 300 sources was considered in the area of interest.

TDE was performed on the corrupted signals and the delay estimates whose error exceeded 3 samples were classified as anomalous (complete failure of delay estimation). The percentage of anomalous estimates in the various conditions is reported in Table 1. The standard deviation (in

sample unit) of the delay estimation error in the remaining cases is presented in Figure 5.

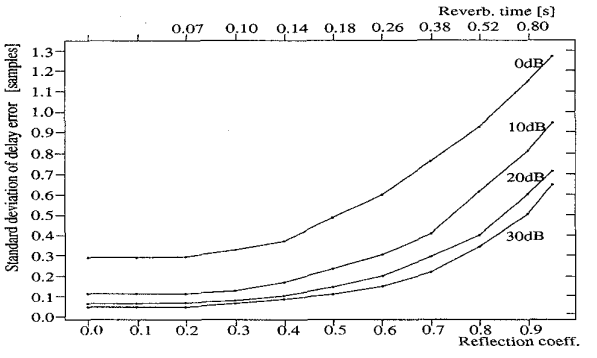

Figure 5: *Error standard deviation of the non-anomalous delay estimates under different wall reflection coefficients (yielding different reverberation time) and four SNR conditions.*

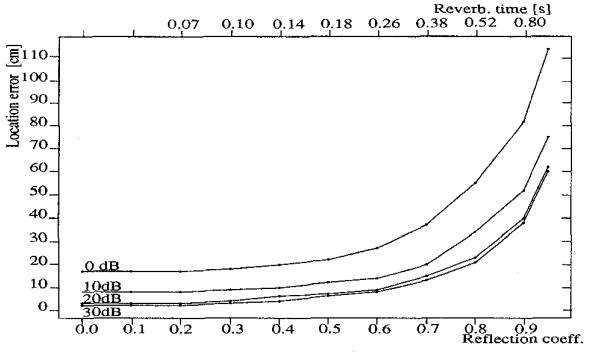

Figure 6: *Average location error for sources located in the allowed region as a function of T_{60} and SNR.*

The location procedure outlined above was then applied without distinction between anomalous and non-anomalous estimates, but excluding cases in which the directions provided by each microphone pair were divergent or intersecting outside a tolerance area around the considered region. Table 2 gives the percentage of failed locations in the various conditions. Figure 6 shows the average error associated to the located sources, depending on reverberation time and SNR.

SNR \ β	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95
30 dB	0	0	0	0	0	0	0	0.5	2.5	16.2	31.4
20 dB	0	0	0	0	0	0	0	1.0	3.9	19.7	35.4
10 dB	0	0	0	0	0	0	0.3	2.5	14.4	29.7	48.1
0 dB	0	0	0	0	0.6	2.9	8.0	21.0	35.7	55.9	65.4

Table 1: Percentage of anomalous TDE estimates as a function of reflection coefficient and SNR.

SNR \ β	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95
30 dB	0	0	0	0	0	0	0.6	1.4	0.6	19.0	30.3
20 dB	0	0	0	0	0	0.7	2.3	5.3	8.0	20.0	29.6
10 dB	0	0	0	0	1.3	3.0	4.3	5.7	12.7	25.0	38.9
0 dB	0.7	2.3	2.3	2.6	2.8	5.0	10.1	18.6	32.0	40.8	46.9

Table 2: Percentage of location failures as a function of reflection coefficient and SNR.

5. CONCLUSIONS

This paper shows that a simple array of four microphones combined with a very accurate TDE technique may represent the basis of a robust acoustic event location system.

Results evidence that there are critical values of SNR and reverberation time for this approach, while the configuration of the simple linear array does not represent a critical issue, provided that an accurate TDE technique is employed and that microphones are not too close to each other. Concerning the hostile conditions, further work should be devoted to investigate the employment of more microphones as well as 2D array geometries.

Both simulations and real experiments have been performed. In the latter case only a typical office condition has been investigated: further real experiments should be conducted in a varechoic chamber to assess the real performance.

Future work will be devoted to extend this activity as well as to apply this technique to a talker tracking task for the development of a hands-free dictation scenario.

6. REFERENCES

[1] J.L. Flanagan, H.F. Silverman, "Workshop on Microphone Arrays: Theory, Design & Application", CAIP Center - Rutgers University, October 1994.

[2] D. Giuliani, M. Matassoni, M. Omologo, P. Svaizer, "Hands Free Continuous Speech Recognition in Noisy Environment using a Four Microphone Array", *Proc. ICASSP*, Detroit 1995, vol. 2, pp. 860-863.

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

[4] H. F. Silverman, S. E. Kirtman, "A Two-stage Algorithm for Determining Talker Location from Linear Microphone Array Data", *Computer Speech and Language* (1992) 6, pp.129-152.

[5] E.E. Jan, P. Svaizer, J.L. Flanagan, "Matched-Filter Processing of Microphone Array for Spatial Volume Selectivity", *Proc. of IEEE ISCAS*, Seattle, May 1995, pp. 1460-1463.

[6] E.E. Jan, P. Svaizer, J.L. Flanagan, "A Database for Microphone Array Experimentation", *Proc. of Eurospeech'95*, Madrid, September 1995, pp. 813-816.

[7] 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, pp. 320-327.

[8] M.R. Schroeder, " Integrated-impulse method measuring sound decay without using impulses", *J. Acoust. Soc. Am.*, 66(2), August 1979, pp. 497-500.

[9] J.B. Allen, D.A. Berkley, "Image method for efficiently simulating small-room acoustics" *J. Acoust. Soc. Am.*, vol. JASA 65(4), April 1979, pp. 943-950.