EFFICIENT TIME DELAY ESTIMATION BASED ON CROSS-POWER SPECTRUM PHASE

M. Matassoni and P. Svaizer

ITC-irst, via Sommarive 18 - Povo (TN), Italy
phone: + (39) 0461 314559, fax: + (39) 0461 314591,
email: {matasso,svaizer}@itc.it
web: http://shine.itc.it

ABSTRACT

Accurate Time Delay Estimation for acoustic signals acquired in noisy and reverberant environments is an important task in many speech processing applications. The Cross-power Spectrum Phase analysis is a popular method that has been demonstrated to perform well even in moderately adverse conditions. This paper describes an efficient approach to apply it in the case of static sources. It exploits the linearity of the generalized cross-correlation to accumulate information from a plurality of frames in the time domain. This translates into a reduced computational load and a more robust estimation. Several examples drawn from real and simulated data in typical applications are discussed.

1. INTRODUCTION

Time Delay Estimation (TDE) between different replicas of a signal, or between signals acquired by different sensors, is intrinsic in many signal processing problems [1]. According to the characteristics of the signals involved in the various applicative fields (e.g. radar, sonar, geophysical and seismic exploration, medical imaging, acoustics) and the corresponding appropriate propagation modeling, many algorithms have been described in the literature [2] to achieve an optimized time delay estimation accuracy.

Sound propagation in enclosures is governed by the laws of room acoustics, which in the most common cases of interest for speech processing (e.g. a talker in the car or in a office environment) lead to a linear, time-invariant model for the signals acquired by microphones. With the spreading employment of microphone arrays for beamforming, speech enhancement, distant talking vocal interfaces and acoustic source localization, an accurate estimation of time difference of arrival of speech wavefronts at the various microphones is of primary concern. The problem of TDE in presence of reverberation and of both diffuse and spatially correlated noise components has been addressed by many authors, who have pointed out the efficiency and good performance of the Phase Transform (PHAT) or Cross-power Spectrum Phase (CSP) [3, 4, 5]. This is a generalized cross-correlation [6] that does not require any a priori modeling of the noise statistics, and is particularly suitable for wide band signals as speech.

This paper describes an efficient method to implement the CSP in order to estimate the mutual delay between two signals. Thanks to the linearity of the CSP, in the case of constant delay, it is possible to accumulate signal frames (in time or frequency domains) before applying the transform. This way it is possible to both reduce the computational load and to enhance the coherence information in the case of low SNR.

The paper is organized as follows: after a brief review of the considered technique and its evolution in Section 2, Section 3 presents the experimental setup that analyzes some possible applications and shows results in terms of MSE. Finally in Section 4 the conclusions are drawn, presenting some issues for future work.

2. TIME DELAY ESTIMATION

The identification of the direction of wavefront arrival due to a single acoustic source is usually performed by means of delay estimation between a pair of microphones, although extensions to multiple microphones have been recently proposed [7]. The original waveform $s(t)$ emitted by a source S impinges on the microphones 0 and 1 after having been transformed by the convolution with the impulse responses between the source and the sensors:

$$x_i(t) = h_i(t) * s(t) + n_i(t), \quad i = 0,1 \tag{1}$$

where $x_i(t)$ are the microphone signals, $h_i(t)$ the impulse responses and $n_i(t)$ the additive noise sequences. In an ideal scenario the impulse responses between a far-field source and the microphones simply reduces to a propagation delay between the source and the microphones and a scaling factor α_i:

$$x_i(t) = \alpha_i s(t - T_i), \quad i = 0,1 \tag{2}$$

The Time Difference Of Arrival (TDOA) is defined by the relation: $\tau = T_1 - T_0$ assuming microphone 0 as reference.

The easiest approach to estimating the TDOA between the two microphones is the maximization of the value assumed by the cross-correlation as a function of the time lag. The correlation can be calculated as inverse Fourier transform of the cross-power spectrum $G(f) = X_0(f)X_1(f)^*$. In literature a multiplicity of variants of generalized cross-correlation have been presented, basically introducing a weighting factor in order to take into account the statistics of source signal and noise in a Maximum Likelihood scheme.

If a normalization factor is applied in order to preserve only the phase information:

$$G_{PH}(f) = \frac{X_0(f)X_1(f)^*}{\|X_0(f)\|\|X_1(f)\|} \tag{3}$$

the CSP or PHAT is obtained as:

$$CSP(t) = IDFT[G_{PH}(f)] \tag{4}$$

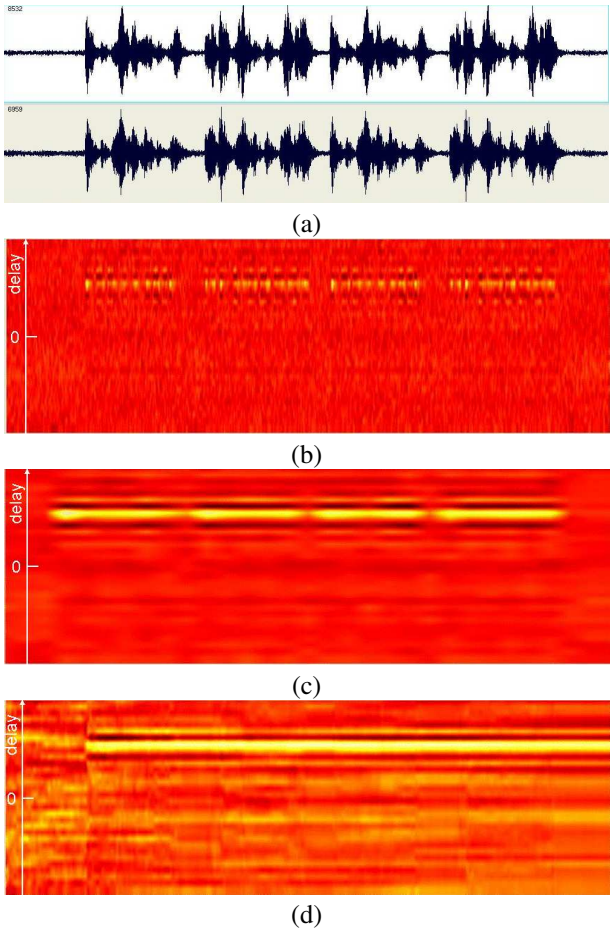

Figure 1: Example of a CSP analysis with the signal pair (a): in (b) the window length is equal to 1024, in (c) equal to 16384 and in (d) the CSP obtained with the proposed processing applied progressively over frames (window length = 1024).

Using the formulation (2) and considering that the delay in time domain corresponds to a phase rotation in frequency domain, it turns out that the IDFT of the function (3) presents a delta pulse centered on the delay τ. The delay estimate is derived from:

$$\tilde{\tau} = \arg\max_t CSP(t) \tag{5}$$

As it is common use for the extraction of other acoustic features, the signal is divided and analyzed in frames. Thus an estimate $\tilde{\tau}_k$ is generated at each frame k starting from the normalized cross-spectrum:

$$G_{PH}(f)_k = \frac{X_{0,k}(f)X_{1,k}(f)^*}{\|X_{0,k}(f)\|\|X_{1,k}(f)\|} \tag{6}$$

This allows in principle to correctly track a moving speaker or to detect different speakers, but the method is not robust against disturbances, i.e. false estimates can be obtained for some frames. Both reverberation and noise, accounted for in the more realistic model (1), contribute to increase the variance of the delay estimates and can produce spurious peaks in the CSP function. Under these circumstances, a larger analysis window helps in reducing the instability of the correct peak, provided that the speaker does not change his/her position within the considered time interval. This observation suggests to enhance the estimation based on a single-frame basis by averaging the CSP over multiple frames. In frequency domain this corresponds to an averaged cross-power spectrum:

$$G_{PH}(f) = \sum_{k=1}^{K} \frac{X_{0,k}(f)X_{1,k}(f)^*}{\|X_{0,k}(f)\|\|X_{1,k}(f)\|} \tag{7}$$

The idea proposed here derives from the observation that the sum of the DFT obtained by different frames, thanks to the linearity of the transform, is equivalent to a single DFT of a wrapped version of the input signal obtained by accumulation of the signal over the analysis window:

$$x_w(n) = \sum_{k=1}^{K} x(n+kL) \tag{8}$$

where k indicates the frame index and L the number of points of the window. A single CSP computation, after having accumulated the two signals, is then sufficient to estimate the required delay. The corresponding transforms are applied on a window of lenght L instead of $K \cdot L$.

The advantage of the proposed method is twofold: on one hand the computational complexity is reduced and, on the other hand, the intrinsic integration effect contributes to enhance the estimation, provided that the acoustic source does not change its position during the time interval corresponding to the analysis window.

3. EXPERIMENTS

The approach can be advantageously applied for time delay estimation in critical conditions, signal alignment and rapid aiming for delay-and-sum beamforming. The proper setup of this method, in terms of window length and number of accumulated frames, depends on the actual application. In the following some examples are shown, evaluating the algorithm in different situations and applications.

3.1 Test data

The signals employed in the experimentations have been collected at our labs within the EU projects CHIL and HIWIRE. Three different contexts have been chosen to test the algorithms, trying to set up tasks with increasing difficulty in terms of time delay estimation.

The first set of signals (T1) is the result of a simulation in various rooms. With pairs of real impulse responses previously measured using chirp signals [8] in rooms of different sizes and acoustics, the signals received by two microphones have been generated by simulating a source placed in several positions (8 in this work). The resulting synthetic corpus aims at testing the algorithms in environments characterized by different noise and reverberation components.

The assumed model for the simulation of channel i is the following:

$$x_{ij}(t) = h_{ij}(t) * s(t) + \alpha_i \sum_{l \neq j} h_{il}(t) * n_l(t), \tag{9}$$

where $s(t)$ is the clean signal, $h_{ij}(t)$ is the measured impulse response between position j and sensor i, $n_l(t)$ are white

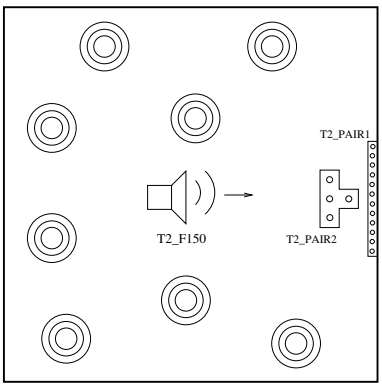

Figure 2: Map of the acoustically treated room (5m × 3.5m) used to collect testset T2, reporting on positions of microphones, simulated talker and noise sources.

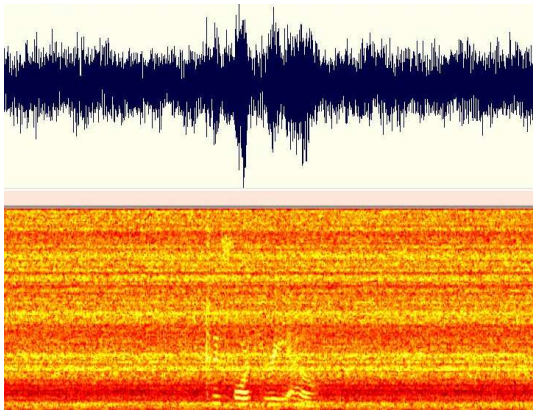

Figure 3: Waveform and spectrogram of a signal of testset T2.

noise sequences and α_i is a proper scale factor accounting for the selected resulting SNR. Moreover, noise sequences associated to the various location are uncorrelated. A long sequence of speech (about 140s) has been generated for 5 rooms and 7 SNR values (from 0 to 30dB). An additional virtual room was considered including a dummy condition in which the impulse response is simply a delayed delta, representing an ideal non-reverberant environment.

The second set (T2) is a realistic scenario where the dominant disturbances are represented by (almost) diffuse noise realized in an acoustically treated room by means of several loudspeakers (see Figure 2). Here the acoustic source was a high-quality loudspeaker (placed at position T2_F150 of Figure 2) reproducing utterances in front of two microphone arrays, arranged in linear ((T2_PAIR1) and 2D fashion (T2_PAIR2). The task consists in estimating, for each sentence, the time delay from the speech component received by pairs of microphones of the two arrays.

Similarly to testset T2, a third set of signals (T3) has been acquired in a further more real scenario. A talker uttered some sentences in the CHIL room, a large room (see Figure 4), acoustically characterized by a reverberation time (T_{60}) of about 0.7s, and was simultaneously recorded by 24 microphones arranged in 6 T-shaped arrays, indicated by A, B, C, D, E, and F in Figure 4. The positions of the talker are identified by T3_POS1, T3_POS2 and T3_POS3; at each position the speaker uttered with different orientations, as indicated by the arrows of Figure 4. Note that with this setup in each recording the microphones in front of the speaker have picked up the direct waveform while those behind the speaker have received mainly the reflections.

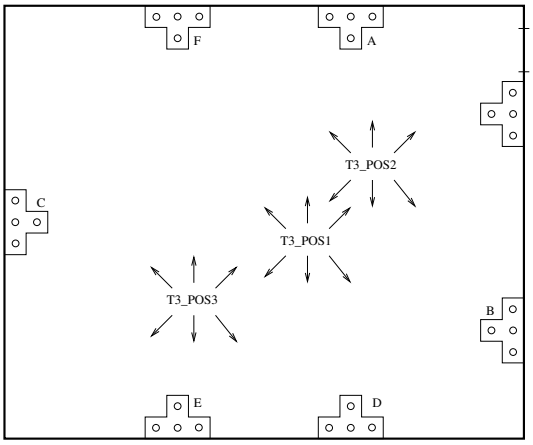

Figure 4: Map of the CHIL room (6m × 5m), used to collect T3, reporting on positions of microphones and acoustic sources.

Task	#pos	pairs	F (kHz)	SNR (dB)	$T_{60}(s)$
T1	1	1	16	0÷30	0.0÷0.7
T2	1	2	16	1÷5	0.15
T3	3	12	44.1	10÷15	0.7

Table 1: Corpora used in the experimental setup: number of speakers positions, numbers of microphones pairs, sampling frequency, SNR and T_{60}.

Table 1 summarizes the characteristics of the three sets adopted for the experimental evaluation of the algorithms.

3.2 Results

The performance of the algorithms are measured in terms of Mean Square Error (samples^2), with respect to the known correct source positions, and Anomaly Rate (AR), defined as the rate of estimates that differ from the known references more than half the correlation time (4 samples at 16kHz) [7].

For reference, besides the standard CSP algorithm, that provides an estimate every frame, we apply a *median filter* for selecting a reduced number of estimates, according to a threshold for the peak of the CSP function. So the hypothesized delay is derived only from the most reliable frames. The proposed method using CSP on accumulated signals is indicated by *accCSP* in the tables.

3.2.1 Testset T1

Figures 5, 6, 7, 8, and 9 report the performance in a synthetic but realistic scenario, where different room acoustics have been properly simulated. An average performance comparison is provided by Table 2.

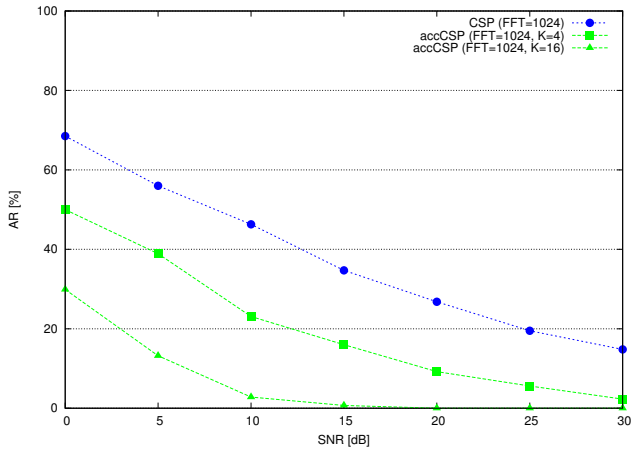

Figure 5: Anomaly Rate results for a low-reverberation environment ($T_{60} = 0.15s$).

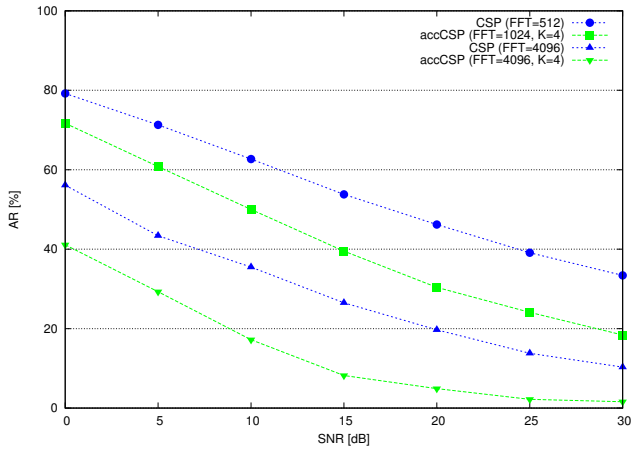

Figure 6: Anomaly Rate results as averages of 5 different rooms, characterized by a T_{60} ranging from 0.0 to $0.7s$.

The good results confirm the validity of the enhanced algorithm in this scenario. Note that the presence of reverberation, as expected, tends to reduce the performance of the algorithms; moreover, the gain of the proposed method progressively decreases with larger reverberation times, as shown by Figure 8.

T1	MSE	AR(%)
CSP	0.45	48.5
median filter (K=8)	0.24	32.1
accCSP (K=8)	0.31	27.2
median filter (K=16)	0.27	30.4
accCSP (K=16)	0.29	22.2

Table 2: Performance on testset T1 in terms of averages of Mean Squared Error and Anomaly Rate (%) for 6 rooms and 7 SNR conditions. As in equation (8), K is the number of accumulated frames.

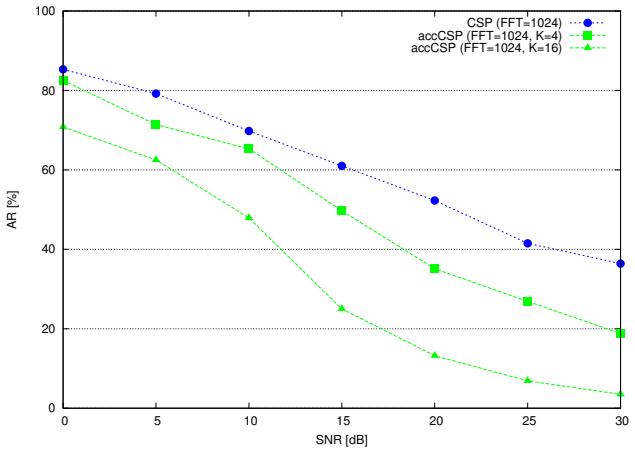

Figure 7: Anomaly Rate results for a medium-reverberation environment ($T_{60} = 0.65s$).

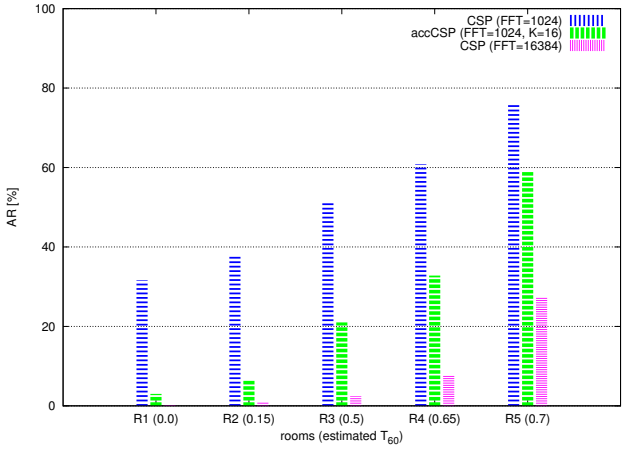

Figure 8: Anomaly Rate results as averages of the 7 SNR conditions for 5 (simulated) environments.

3.2.2 Testset T2

In this case, for comparison purposes, a *baseline* processing was considered, consisting in a *CSP* analysis extended on a window equal to the whole signal; note that it can be considered as a sort of upper bound for the method, although it is not suitable for a practical use.

T2	MSE	AR(%)
baseline	0.05	4.5
median filter	0.37	49.0
accCSP	0.20	24.0

Table 3: Performance on testset T2 in terms of Mean Squared Error and Anomaly Rate (%).

The T2 scenario is more realistic since the data are real (i.e. sounds are really propagating in the physical environment), although the speaker is simulated by means of an high-quality loudspeaker. The corresponding results are presented in Table 3. The window length was set to 1024 and the value of K depended on the length of the actual waveform,

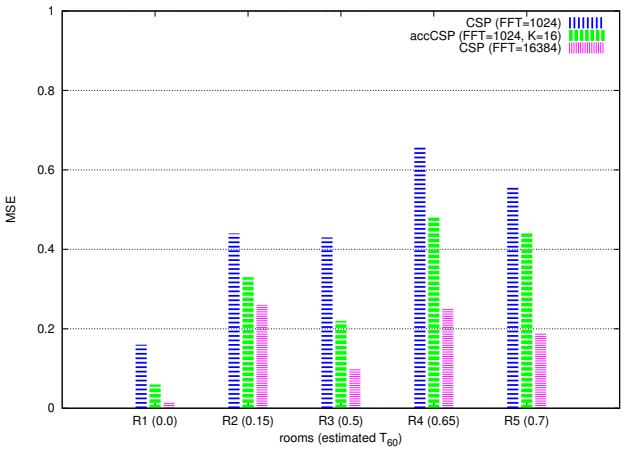

Figure 9: MSE results as averages of the 7 SNR conditions for 5 (simulated) environments.

since a single overall estimate was derived for each utterance.

In this context each recording presents a short speech segment with respect to the length of the background noise (see Figure 3), leading to a complex task without the use of a voice-activity detector. In these more critical conditions the new algorithm provides again a benefit in terms of reduced MSE and AR.

3.2.3 Testset T3

T3	MSE	AR(%)
median filter	3.4	43.8
accCSP	1.2	19.5

Table 4: Performance on testset T3 in terms of Mean Squared Error and Anomaly Rate (%).

The following experiment analyzes the situation of a talker not correctly oriented towards the microphones. See [9] for a more detailed description of the corpus. The references for this task are not precisely known because of the inherent uncertainty of the position of a human. Hence the performance is computed assuming as reference the *baseline* results: signal pairs for which this processing did not provide acceptable results (i.e. there was no or insufficient coherence between signals, as a consequence of unfavorable orientation of the speaker's head) were discarded. In this test the window length was set to 16384.

Even under the real acoustic conditions of testset T3, it is confirmed (see Table 4) that the accumulation of signal frames leads to a more robust time delay estimation.

3.3 Signal alignment

Another problem where this technique can be successfully applied is the estimation of a temporal shift between a pair of (long) signal recorded with different acquisition devices. In this situation the problem is represented not only by the presence of reverberation and noise but also by the unknown range of the primary delay between the recordings. This delay can be in principle far greater than the analysis window and the estimation with the standard algorithm can be influenced by local fluctuations of the delay, according to the current position of the main acoustic source. The *accCSP*, effectively taking into account the global waveform, has shown to provide a robust averaged estimate with a tractable analysis window.

4. CONCLUSIONS

In this paper we have proposed an enhancement of a popular method for time delay estimation and tested it in various contexts where the acoustic source position can be considered unknown but fixed, leading to a faster version of the algorithm that provides also better results in unfavorable conditions. From the experiments it turns out that, under the hypothesis of static source, the *accCSP* method provides better performance with respect to a popular algorithm and, at the same time, it reduces the computational requirements.

The proposed approach also provides a satisfactory method to realign signals with reduced processing. With an accurate design of the long analysis window it is possible to derive an extension for moving sources, by comparing the local information with the long-term one. Future activities regard the adoption of this approach for acoustic source localization and for speech enhancement by beamforming.

Acknowledgments: This research work was partially supported by the IST EU FP6 research program HIWIRE.

REFERENCES

[1] "Special Issue on Time-Delay Estimation,," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 29, 1981.

[2] G. C. Carter editor, *Coherence and Time Delay Estimation*, IEEE Press, 1983.

[3] M. Omologo and P. Svaizer, "Use of the cross-power-spectrum phase in acoustic event location," *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 3, pp. 288–292, 1997.

[4] D. Ward and M. Brandstein eds., *Microphone Arrays: Techniques and Applications*, Springer, Berlin, 2001.

[5] T. Gustafsson, B. D. Rao, and M. Trivedi, "Source Localization in Reverberant Environments: Modeling and Statistical Analysis," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, 2003.

[6] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-24, no. 4, pp. 320–327, April 1976.

[7] J. Chen, Y. Huang, and J. Benesty, "Time delay estimation via multichannel cross-correlation," in *Proc. of ICASSP*, 2005, vol. III, pp. 49–52.

[8] N. Aoshima, "Computer-generated pulse signal applied for sound measurement," *J. Acoust. Soc. Am.*, vol. 69(5), pp. 1484–1488, 1981.

[9] A. Brutti, M. Omologo, and P. Svaizer, "Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays," in *Interspeech*, 2005, pp. 2337–2340.