A REAL-TIME DEMONSTRATOR FOR THE 2D LOCALIZATION OF TWO SOUND SOURCES USING BLIND ADAPTIVE MIMO SYSTEM IDENTIFICATION

Anthony Lombard, Walter Kellermann

Multimedia Communications and Signal Processing
University of Erlangen-Nuremberg
Cauerstr. 7, 91058 Erlangen, Germany
Email: {lombard,wk}@LNT.de

Herbert Buchner

Deutsche Telekom Laboratories
Technical University Berlin
Ernst-Reuter-Platz 7, 10587 Berlin, Germany
Email: hb@buchner-net.com

ABSTRACT

A real-time demonstrator for the 2D localization of two sound sources using two microphone pairs is presented and evaluated. The scheme relies on Blind Source Separation (BSS) to adaptively identify the acoustical MIMO system, hence allowing the estimation of relative time delays for each source and each dimension. Extending our previously presented work [1], a mechanism to solve a pairing problem occuring in the multidimensional localization of several sources is described. It exploits the inherent signal extraction abilities of BSS. Experimental evaluations with large microphone apertures show that the demonstrator can accurately localize two speech sources in a 2D space, with a precision better than one degree.

Index Terms— Acoustic source localization, TDOA estimation, blind source separation, real-time demonstrator

1. INTRODUCTION

In this paper, we address the problem of localizing several simultaneously active broadband sound sources in a multidimensional space, under reverberant acoustical conditions. The localization procedure adopted here is based on the estimation of Time Differences Of Arrival (TDOA), where relative temporal signal delays (i.e., the TDOAs) have to be first estimated, before calculating the source position in a second step. By estimating one TDOA for each source and for each dimension, this two-step procedure can be applied for the localization of one source in several dimensions or for the localization of several sources in one dimension. However, the generalization to the simultaneous localization of multiple sources in several dimensions is not straightforward since it necessitates an intermediate step to solve a spatial ambiguity problem [1]. In the following, we describe a PC-based demonstrator capable of localizing two sources in a 2D space, thereby addressing both the multiple-source TDOA estimation and the spatial ambiguity issues in real-time. The different steps involved in the scheme are detailed in Sect. 2 and experimental results proving its effectiveness are provided in Sect. 3.

2. DESCRIPTION OF THE REAL-TIME LOCALIZATION DEMONSTRATOR

Figure 1 illustrates the three-step procedure involved in the demonstrator, where the microphone signals are delivered by the 4-sensor array depicted in Fig. 2. The real-time system is capable of localizing two simultaneously active sources. Therefore, if we assume

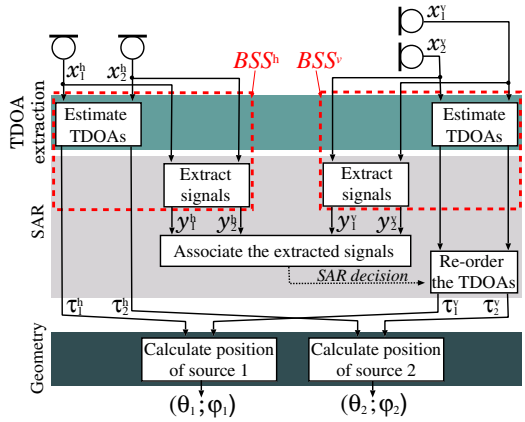

Fig. 1. Description of the implementation.

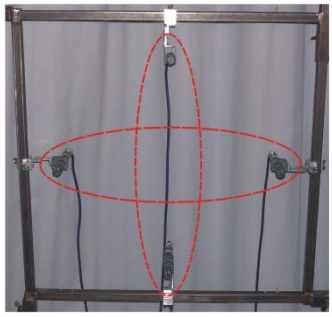

Fig. 2. 2D sensor array of four microphones.

sources in the far-field, a total of four Directions of Arrival (DOAs) need to be calculated using two TDOA estimators (one for each dimension, each computing a pair of TDOAs). With a first estimator measuring two TDOAs from the horizontal microphone pair, we can calculate the azimuth angles θ_1 and θ_2 of the first and the second source, respectively (the geometrical step in Fig. 1). Similarly we can obtain two elevation angles φ_1 and φ_2 from a second TDOA estimator measuring two TDOAs from the vertical microphone pair. However, without additional information on the relative source positions, we cannot determine if the source one with azimuth θ_1 has the elevation φ_1 or the elevation φ_2. The same problem occurs of course for the second source, hence giving rise to a *spatial ambiguity* phenomenon, regardless of the TDOA estimation method used [1]. Therefore, for the TDOA-based scheme to be successful, we need not only a good TDOA estimator but also a robust Spatial Ambiguity Resolver (SAR).

Table 1. SAR mechanism based on scores.

$sc_1(m) = \xi_{11}(m) \cdot \xi_{22}(m)$
$sc_2(m) = \xi_{12}(m) \cdot \xi_{21}(m)$
if $sc_1(m) \geq sc_2(m)$
y_1^1 is correlated with y_1^2 and y_2^1 is correlated with y_2^2.
$\Rightarrow$ The TDOAs are already in the right order.
$\begin{cases} position_1 = geometry(\hat{\tau}_1^1, \hat{\tau}_1^2) \\ position_2 = geometry(\hat{\tau}_2^1, \hat{\tau}_2^2) \end{cases}$
else
y_1^1 is correlated with y_2^2 and y_2^1 is correlated with y_1^2.
$\Rightarrow$ The TDOAs are permuted.
$\begin{cases} position_1 = geometry(\hat{\tau}_1^1, \hat{\tau}_2^2) \\ position_2 = geometry(\hat{\tau}_2^1, \hat{\tau}_1^2) \end{cases}$
end

2.1. The TDOA extraction

As can be seen from Fig. 1, the TDOA extraction is realized according to [2] by two BSS instances BSS^{h} and BSS^{v} running in parallel. BSS^{h} and BSS^{v} measure two TDOAs each (one per source), using the horizontal and the vertical microphone pairs of the 4-sensor array depicted in Fig. 2, respectively. Originally developed for the blind source separation of convolutive mixtures [3, 4], the TDOA estimation technique [2] performs a blind adaptive Multiple-Input-Multiple-Output (MIMO) system identification of the acoustical environment in the time domain.

For each BSS instance, two TDOA estimates $\hat{\tau}_1$ and $\hat{\tau}_2$ can be obtained after each BSS update as follows:

$$\hat{\tau}_1 = \arg\max_{\kappa} |w_{12}(\kappa)| - \arg\max_{\kappa} |w_{22}(\kappa)|, \quad (1)$$

$$\hat{\tau}_2 = \arg\max_{\kappa} |w_{11}(\kappa)| - \arg\max_{\kappa} |w_{21}(\kappa)|, \quad (2)$$

where w_{11}, w_{12}, w_{21} and w_{22} are the BSS adaptive filters after [2], identifying the acoustical propagation paths between each source and each microphone. Contrary to other widely used approaches like the Generalized Cross-Correlation (GCC) [5] or the Adaptive Eigenvalue Decomposition (AED) [6] algorithms, the BSS-based method accounts for both the room reverberation and the presence of multiple simultaneously active sound sources in its propagation model and is therefore well suited to the task considered here.

2.2. The spatial ambiguity resolver

The TDOA extraction described in Sect. 2.1 has been combined in [7] with a particle filter. There, the authors addressed the pairing problem encountered in the SAR task by considering more microphone pairs than the number of dimensions, hence introducing some redundancy. In this paper, to use a minimum number of microphone pairs (i.e., of BSS instances) and limit the computational complexity of the scheme, we rely on the ability of the BSS-based scheme to simultaneously estimate TDOAs for several sources and unravel the acoustical mixing system, thereby providing estimates of the original source signals at the BSS outputs. We actually tackle the SAR problem by measuring the correlation between the output signals of each BSS instance, which allows us to localize two sources in two dimensions using only two microphone pairs.

In [1], the correlation between the output signals was measured in the time domain using the Cross-Correlation Function (CCF). Alternatively, the correlation can be measured in the frequency domain, based on the Magnitude Squared Coherence (MSC) function. Following the notations introduced in Fig. 1, the MSC between the i^{th} output of BSS^{h} and the j^{th} output of BSS^{v} is defined for each DFT bin ν and for each processing block m as:

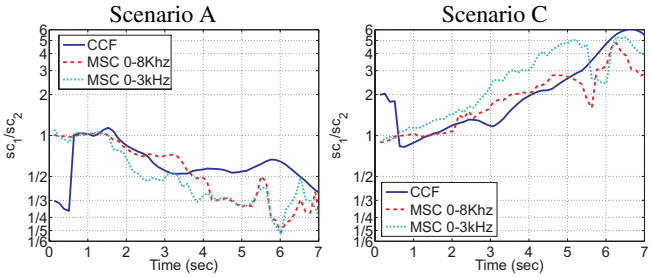

Fig. 3. Example runs for two source constellations (scenarios A and C in Sect. 3) of the SAR scores, when using the CCF and when averaging the MSC (here computed based on the Welch's method) over the entire bandwidth or over the first three kilohertz.

$$\hat{\Gamma}_{ij}^{(\nu)}(m) \triangleq \hat{\Gamma}_{y_i^{\mathrm{h}} y_j^{\mathrm{v}}}^{(\nu)}(m) = \frac{|\hat{S}_{y_i^{\mathrm{h}} y_j^{\mathrm{v}}}^{(\nu)}(m)|^2}{\hat{S}_{y_i^{\mathrm{h}} y_i^{\mathrm{h}}}^{(\nu)}(m) \cdot \hat{S}_{y_j^{\mathrm{v}} y_j^{\mathrm{v}}}^{(\nu)}(m)}, \quad (3)$$

where $\hat{S}_{y_i^{\mathrm{h}} y_j^{\mathrm{v}}}^{(\nu)}(m)$, $\hat{S}_{y_i^{\mathrm{h}} y_i^{\mathrm{h}}}^{(\nu)}(m)$ and $\hat{S}_{y_j^{\mathrm{v}} y_j^{\mathrm{v}}}^{(\nu)}(m)$ denote power spectral densities estimated using either the Welch's averaged, modified periodogram method [8], or using a first-order recursive smoothing. To obtain a single correlation measure for each signal pair and for each processing block, the MSC can be averaged over the DFT bins:

$$\xi_{ij}^{\mathrm{MSC}}(m) \triangleq \xi_{y_i^I y_j^J}^{\mathrm{MSC}}(m) = \frac{1}{N/2+1} \sum_{\nu=0}^{N/2} \hat{\Gamma}_{ij}^{(\nu)}(m), \quad (4)$$

where N is the DFT length. Like the MSC estimate (3), $\xi_{ij}^{\mathrm{MSC}}(m)$ is a measure with values between 0 and 1, which results from an inherent signal power normalization. To obtain slightly more robust and quicker SAR decisions, it is also possible to perform the average (4) only over frequency ranges approximating the spectral support of the source signals (e.g., for speech, between 0 and 3kHz). This is because BSS offers better signal separation at frequencies where the excitation signals are strong.

At each block instant, the SAR can therefore rely on a set of four correlation measures $\xi_{11}(m)$, $\xi_{12}(m)$, $\xi_{21}(m)$ and $\xi_{22}(m)$, where $\xi_{ij}(m)$ can be CCF-based like in [1], or MSC-based like in (4). An SAR procedure is presented in Table 1. In our case, we have to decide between two cases (the "vertical" TDOAs are either in the right order or permuted, see Fig. 1). The procedure consists in attributing a score to both possibilities (sc_1 and sc_2 in Table 1). By choosing the possibility with highest score, a fast and unambiguous SAR decision can actually be made, as can be seen in Fig. 3 for two example trials. The algorithm's parameters used here are the same as those used for the evaluations in Sect. 3.

2.3. The geometry step

As a final step in the localization scheme of Fig. 1, the estimated TDOAs have to be translated into a spatial location for each source. For simplicity, we resort here to the far-field assumption. The geometry step therefore reduces to a simple mapping of two TDOAs τ^{h} and τ^{v} obtained from the horizontal and the vertical microphone pairs respectively, into an azimuth angle $\theta = \arcsin(c\tau^{\mathrm{h}}/f_{\mathrm{s}}d^{\mathrm{h}})$ and an elevation angle $\varphi = \arcsin(c\tau^{\mathrm{v}}/f_{\mathrm{s}}d^{\mathrm{v}})$. d^{h} and d^{v} are respectively the horizontal and vertical microphone spacings, c is the sound velocity and f_{s} is the sampling rate.

Note that the TDOA estimator described in Sect. 2.1 and based on time-domain system identification allows to use large microphone

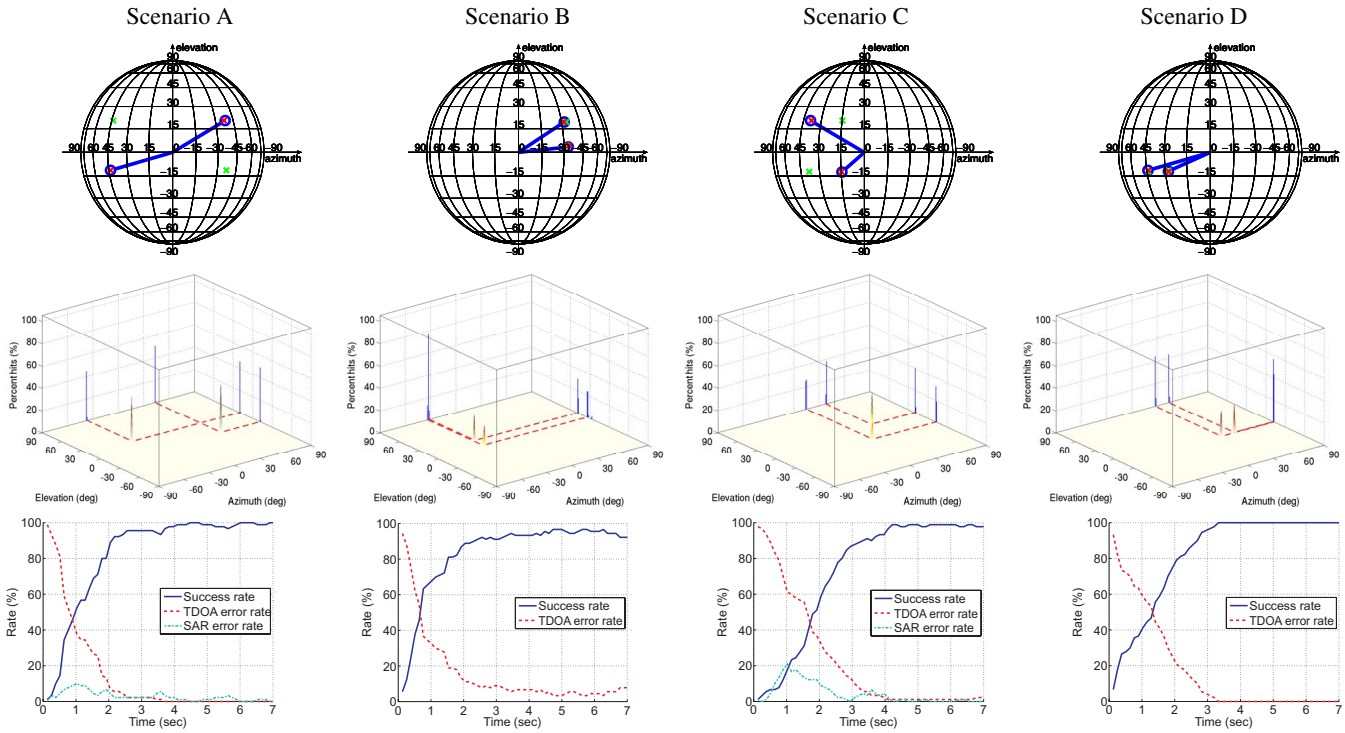

Fig. 4. Results for two static sources in four scenarios:
- *The first row* shows the estimated location after convergence (blue lines), the red crosses showing the reference source locations and the green crosses showing what happens for scenarios A and C when the reference elevations are permuted (i.e., after a wrong SAR decision).
- *The second row* shows the 2D histograms of the estimated source locations with their marginal histograms obtained for each scenario from the set of 90 7sec-long trials. Reference source locations are shown by red dashed lines.
- *The third row* shows the evolution of the success rates (margin 10^o) and error rates (either due to a DOA estimation error, or due to a wrong SAR decision when the DOAs were correct). For each scenario, the results have been averaged over all sources, dimensions and trials, the rates adding up to 100%.

spacings since it is not affected by spatial aliasing effects. The full potential of large microphone arrays can then be exploited to obtain a good spatial resolution. To further improve the spatial resolution of the localizer at a low computational cost, the filters of the BSS unmixing system can be interpolated before performing the effective TDOA estimations (1) and (2), hence delivering fractional delays but without increasing the sampling rate for the BSS operations.

3. EXPERIMENTAL RESULTS

The localization performance of the real-time demonstrator was assessed for two sources using the 4-sensor array depicted in Fig. 2. All microphone signals were recorded at the sampling frequency 48kHz before being downsampled to the sampling frequency $f_s =$ 16kHz for real-time processing. To guarantee a good spatial resolution, relatively large horizontal and vertical microphone spacings d^{h}=d^{v}=47cm in each dimension were used and fractional TDOAs were obtained (Sect. 2.3) using an interpolation factor of three.

For each BSS instance, a filter coefficient update occured every 128ms. To limit the computational complexity of the scheme, the length of the BSS adaptive filters used in the TDOA extraction (Sect. 2.1) was restricted to 128 samples only. Although this does not allow to fully identify the acoustical MIMO mixing system and its late reflections, it suffices to correctly estimate the direct propagation paths and the early reflections, hence obtaining good TDOA estimates and providing enough signal separation for the SAR step. The SAR procedure of Table 1 was applied using the MSC-based correlation measure proposed in Sect. 2.2, averaging over the first three kilohertz only. The Welch's averaged, modified periodograms [8] in (3) were computed by sectioning observation intervals of one second into 50%-overlapping blocks of 1024 samples each.

3.1. Results for static sources

Various 7sec-long speech segments were played by two loudspeakers in a living-room-like environment ($T_{60} \approx 300$ms). The four source constellations depicted in Fig. 4 were considered, conducting 90 trials with different speech mixtures in each run. References for the source locations were obtained during the recordings by playing five seconds of white noise by a single loudspeaker placed at each considered source location. The recorded signals served to compute the GCC-PHAT function at the sampling rate 48kHz and extract the "reference" TDOAs accordingly [5], the TDOAs being mapped into DOAs like in Sect. 2.3. Note that scenarios B and D are particular cases since the sources have a common (or very similar) DOA in one dimension (a common azimuth for scenario B, a common elevation for scenario D). Actually, as can be seen from the first row of Fig. 4, scenarios B and D are not subject to localization errors due to wrong SAR decisions, contrary to scenarios A and C where a wrong SAR decision can cause a considerable mismatch for both sources.

Figure 4 provides an overview of the localization results achieved by the demonstrator in the four considered scenarios. We

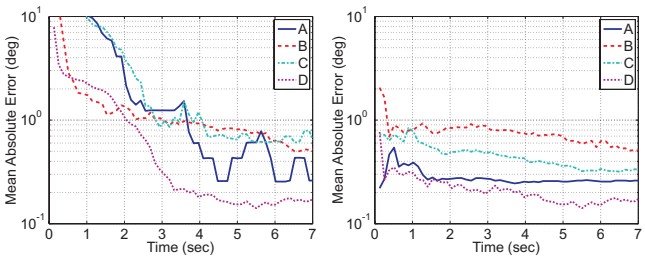

Fig. 5. Mean absolute error of the DOA estimates averaged over azimuth and elevation angles for two static sources. Left: taking all trials into account. Right: taking only the successful trials into account, the proportion of successful trials with 10^o margin being depicted as "Success rate" in Fig. 4.

see from the histograms that correct source locations were obtained most of the time in all four scenarios. This demonstrates that wrong DOA estimations occured relatively rarely in all scenarios, from the start till the end of each trial. However, analyzing the transient behavior of the demonstrator depicted in the third row of Fig. 4, we see that a certain convergence time was necessary to obtain the four DOAs reliably and approach the line of 100% success. In fact, although a first source could usually be found almost immediately, the TDOA estimation algorithm needed up to a few seconds to localize the second source. As a consequence, it sometimes happened during the first 2-3 seconds of each trial that the TDOA estimators delivered only one TDOA (actually twice the same) in each dimension, which is seen as a failure in the plots of Fig. 4. Finally, we can observe that as long as the TDOA estimates were correct, errors due to a wrong SAR decision were very limited. This confirms the good results shown in Fig. 3 and proves the effectiveness of the MSC-based SAR procedure. This is a very important result since it allows us to avoid introducing some redundancy in the estimation process (Sect. 2.2) and makes it possible to perform the source localization based on two microphone pairs only.

The mean absolute error of the DOA estimates averaged over all four angles is depicted in Fig. 5 for all scenarios. In the left figure, where every 90 trials were taken into account in the ensemble average, the relatively large error found during the first seconds of the simulations is due to the localization of one source only instead of two, as mentioned above already. However, only accounting for the trials where both sources were found, the DOA estimates are very accurate, the expected DOA estimation error lying well under one degree in all scenarios.

3.2. Results for moving sources

The localization scheme was also assessed in the presence of a fixed source and a moving source under the same conditions as in Sect. 3.1. Additionally, to avoid outliers due to moving sources, a median filter of length 19 was applied on the extracted TDOAs, as well as on the SAR decisions. Fig. 6 depicts the localization results obtained for this scenario. Approximate references for the source locations were computed during the recordings, based on geometrical measurements. As already observed for fixed sources in Sect. 3.1, we see that a convergence delay of 2-3 seconds was necessary to follow each source movement and find the new positions accurately. Although the approximate references do not allow us to draw some conclusions on the exact accuracy of the localization procedure in this case (see Sect. 3.1 for a more precise evaluation), it shows that the demonstrator was able to correctly find and track the sources over the entire duration of the experiment.

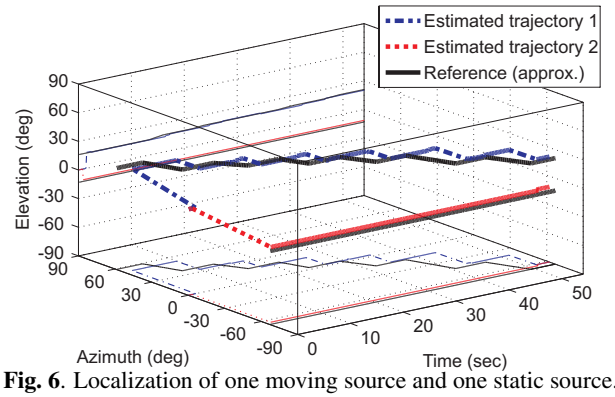

Fig. 6. Localization of one moving source and one static source.

4. CONCLUSIONS

A BSS-based real-time demonstrator for the localization of two sound sources has been presented and assessed. The experimental evaluations conducted for static and moving sources and using large microphone apertures show that the BSS-based algorithm can accurately localize two speech sources in a 2D space within a few seconds and with a precision better than one degree. The procedure proposed to solve the SAR issue proved to be reliable, which allows us to operate with two microphone pairs only, without relying on any prior knowledge on the source positions.

5. REFERENCES

[1] A. Lombard, H. Buchner, and W. Kellermann, "Multidimensional localization of multiple sound sources using blind adaptive MIMO system identification," in *Proc. IEEE Int. Conf. on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, Heidelberg, Germany, Sep. 2006.

[2] H. Buchner, R. Aichner, J. Stenglein, H. Teutsch, and W. Kellermann, "Simultaneous localization of multiple sound sources using blind adaptive MIMO filtering," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Philadelphia, PA, USA, Mar. 2005.

[3] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 1, pp. 120–134, Jan. 2005.

[4] R. Aichner, H. Buchner, F. Yan, and W. Kellermann, "A real-time blind source separation scheme and its application to reverberant and noisy environments," *Signal Processing*, vol. 86, no. 6, pp. 1260–1277, June 2006.

[5] C.H. Knapp and G.C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24, pp. 320–327, August 1976.

[6] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Am.*, vol. 107, pp. 384–391, Jan 2000.

[7] F. Antonacci, D. Riva, D. Saiu, A. Sarti, M. Tagliasacchi, and S. Tubaro, "Tracking multiple acoustic sources using particle filtering," in *Eur. Signal Processing Conf. (EUSIPCO)*, Florence, Italy, Sep. 2006.

[8] P. D. Welch, "The use of the Fast Fourier Transform for the estimation of power spectra," *IEEE Transactions on Audio Electroacoustics*, vol. 15, no. 2, pp. 70–73, June 1967.