Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

Signal Processing 85 (2005) 177–204

SIGNAL PROCESSING

www.elsevier.com/locate/sigpro

Time difference of arrival estimation of speech source in a noisy and reverberant environment

Tsvi G. Dvorkind[a,*], Sharon Gannot[b]

[a]*Faculty of Electrical Engineering, Technion, Technion City, 32000 Haifa, Israel*
[b]*School of Electrical Engineering, Bar-Ilan University, 52900 Ramat-Gan, Israel*

Received 30 October 2003; received in revised form 23 September 2004

Abstract

Determining the spatial position of a speaker finds a growing interest in video conference scenarios where automated camera steering and tracking are required. Speaker localization can be achieved with a dual-step approach. In the preliminary stage a microphone array is used to extract the *time difference of arrival* (TDOA) of the speech signal. These readings are then used by the second stage for the actual localization. In this work we present novel, frequency domain, approaches for TDOA calculation in a reverberant and noisy environment. Our methods are based on the speech quasi-stationarity property, noise stationarity and on the fact that the speech and the noise are uncorrelated. The mathematical derivations in this work are followed by an extensive experimental study which involves static and tracking scenarios.
© 2004 Elsevier B.V. All rights reserved.

Keywords: Source localization; Non-stationarity; Decorrelation; TDOA

1. Introduction

Determining the spatial position of a speaker finds a growing interest in video conference scenarios where automated camera steering and tracking are required. Microphone arrays, which are usually used for speech enhancement in a noisy environment [22], can be used for the task of speaker localization as well [3,6,8,9,11,20,27]. The related algorithms can be divided into two groups: single and dual-step approaches. In single step approaches the source location is determined directly from the measured data (i.e. the received signals at the microphone array). In the dual-step approaches, the location estimate is obtained by applying two algorithmic stages. First, *time difference* (or *time delay*) *of arrival* (TDOA) estimates are obtained from different microphone

*Corresponding author. Tel.: +972 4 8294751; fax: +972 4 8292795.
E-mail addresses: dvorkind@tx.technion.ac.il (T.G. Dvorkind), gannot@eng.biu.ac.il (S. Gannot).
URL: http://www.eng.biu.ac.il/~gannot.

0165-1684/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2004.09.014

pairs. Then, these TDOA readings are used for determining the spatial position of the source.

Single step approaches can be further divided into two groups. The first group is the high-resolution spectral estimation methods. The well-known *multiple signal classification* (MUSIC) algorithm [35] is a member of this group. So is the work in [21] which considers direction of arrival (DOA) estimation with a uniform circular arrays that outperforms MUSIC-like algorithms at low *signal-to-noise ratio* (SNR) for similar computational loads. Though the mentioned algorithms can perform DOA estimation of multiple sources they are mainly suited for narrow-band signals. We note, however, that extension of those algorithms for a wide-band signals do exist. See for example [12,40]. In the second group of single step approaches we find the *maximum-likelihood* (ML) algorithms, which estimate the source locus by applying the ML criterion. Usually, the ML formulation leads to algorithms involving maximization of the output power of a beamformer steered to potential source locations (i.e. [3,6,8,9,11]).

In the dual-step approaches group, the first algorithmic stage involves TDOA estimation from spatially separated microphone pairs. The *maximum-likelihood generalized cross correlation* (ML-GCC)[1] method presented by Knapp and Carter [27] is considered to be the classical solution for this algorithmic stage. However, the GCC method assumes a reverberant-free model such that the *acoustical transfer function* (ATF), which relates the source and each of the microphones, is a pure delay. Champagne et al. showed this approximation to be inaccurate in reverberant conditions, which frequently occur in enclosed environments [7]. Consequently, algorithms for improving the GCC method in presence of room reverberation were suggested [5,38]. Unfortunately, the GCC method suffers from another model inaccuracy. It is assumed by the GCC model that the noise field is uncorrelated, an assumption which usually does not hold. Thus, the GCC method cannot distinguish between the speaker and a directional interference, as it tends to estimate the TDOA of the stronger signal. Directional interference usually occurs when a point source, e.g. computer fan, projector or a ceiling fan, exists. The authors in [31] suggested discriminating speaker from directional noise with a Gaussian mixture model. A different approach was presented in [10,30], where *higher order statistics* (HOS) was employed for TDOA estimation of a non-Gaussian source and correlated Gaussian noise.

Recently, subspace methods were suggested for TDOA estimation. Assuming spatially uncorrelated noise, Benesty suggested a time domain algorithm for estimating the (truncated to shorter length) impulse responses for TDOA extraction [2]. Extension of that work, for spatially correlated noise was presented by Doclo and Moonen [13,15]. Assuming that the noise correlation matrix is known (using a *voice activity detector* (VAD)), the authors presented a time domain algorithm for TDOA estimation using a *generalized eigenvalue decomposition* (GEVD) approach and a pre-whitening approach.

In this work, we tackle the TDOA estimation problem. Hence, the proposed solutions are members of the dual-step approaches. We address the TDOA extraction based on a single microphone pair. The second algorithmic stage, i.e. the actual localization based on multiple TDOA readings which are extracted from additional microphone pairs [4,18,25], is not addressed in this work. Our model assumptions consider reverberation and spatially correlated noise scenarios [17,19]. Specifically, we consider a single speaker in a stationary noise environment. In [22] the speaker's ATF-s ratio was used as part of a beamformer in a speech enhancement application. Here, we exploit this quantity for the source localization application. Particularly, we show that the TDOA reading can be extracted from the location of the maximal peak in the corresponding impulse response. Similar to [22] and the preceding work by Shalvi and Weinstein [37] we also assume that the interfering noise is relatively stationary, and present a framework where the ATF-s ratio and a noise related term are estimated simultaneously without any VAD employment. Quasi-stationarity of the speech and stationarity of the noise are exploited to derive batch and recursive

[1] For brevity we will simply notate this by GCC.

solutions. The importance of the recursive solution manifests itself in tracking scenarios, where the estimated ATF-s ratio and noise statistics might slowly vary with time. Following the work in [41], we have additionally exploited the fact that there is no correlation between the speaker and the directional noise. The authors in [41] showed that in an application of signal separation, imposing a decorrelation criterion on the estimated signals results in ATF-s ratio estimation. The authors further suggested exploiting speech non-stationarity, resulting in a set of decorrelation equations. However, the obtained equation set is nonlinear, and due to this nonlinearity an inherent *frequency permutation ambiguity* results [32]. The authors in [41] did not give a closed form solution for the resulting, frequency domain, nonlinear equation set. Instead, it was suggested to solve the problem iteratively, by assuming a simplified finite impulse response (FIR) model for the mixing channels and solving in the time domain. To maintain simplicity of the solution, we are solving the problem in the frequency domain. Furthermore, we do not assume the simplified mixing channel.

The obtained decorrelation equations are closely related to *blind source separation* (BSS) problems. Gannot and Yeredor considered the case of instantaneous mixture of a non-stationary signal with a stationary noise [23], where joint diagonalization of correlation matrices is carried out in the time domain. Considering a convolutive mixture (due to room reverberation), researchers suggested solving the nonlinear frequency domain decorrelation equations by applying joint diagonalization of the PSD matrices obtained from different time epoches. Special attention is given to the inherent frequency permutation problem, which is usually solved by imposing an FIR constraint on the separating ATF-s [32,36] or (equivalently) imposing smoothness in the frequency domain [34]. In our contribution we exploit the stationarity of one of the sources (the directional noise) to resolve frequency permutations. No FIR constraint is employed, and the estimated ATF-s ratio is exploited for TDOA extraction. Our simulation study shows that the decorrelation constraint presents improved TDOA estimation for the batch methods at low SNR conditions.

Special emphasis is given for deriving a recursive solution applicable for tracking scenarios. Since the involved decorrelation equation set is nonlinear, we present a general framework for an approximate recursive solution of a nonlinear equation set. The method notated by *recursive Gauss* (RG) is applied to the nonlinear decorrelation equations, resulting in a solution applicable for tracking scenarios. Opposed to the GCC-based methods, our solutions deal with reverberant environment and correlated noise field. Opposed to the subspace methods [2,15], the proposed algorithms are conducted in the frequency domain, resulting in computationally more efficient implementations which do not rely on a VAD for prior knowledge of the noise characteristics. Furthermore, simulation study shows that the suggested algorithms are suitable for tracking scenarios, while the subspace method fails to lock on the TDOA readings, which constantly change due to source movement.

The outline of this work is as follows. In Section 2 we present the model assumptions and suggest the use of ATF-s ratio quantity for TDOA estimation. Section 3 presents the TDOA estimation algorithms, exploiting speech quasi-stationarity, noise stationarity and the fact that there is no correlation between the speech and the noise. Extensive experimental study is presented in Section 4. Finally, several practical considerations are presented in Section 5.

2. Problem formulation and motivation

In this section the problem is formulated and the basic assumptions are presented. By analytical expression and by simulation study, we justify the use of ATF-s ratio for TDOA extraction.

2.1. Basic model assumptions

Define a set of M microphones for which the measured signal at the mth microphone, $z_m(t)$, is

$$z_m(t) = a_m(t) * s(t) + n_m(t); \quad m = 1, \ldots, M, \qquad (1)$$

where * stands for convolution, $s(t)$ is the source signal and $n_m(t)$ is the interference signal at the mth

microphone. t stands for the discrete time index. Naturally, we assume that the interference signal is uncorrelated with the source signal. $a_m(t)$ is an impulse response from the desired speech source to the mth microphone. When $n_m(t)$ is a directional interference, we can state

$$n_m(t) = b_m(t) * n(t); \quad m = 1, \ldots, M, \tag{2}$$

where $b_m(t)$ is the impulse response between the noise $n(t)$ and the mth microphone. $s(t)$ is assumed to be quasi-stationary, while the interference signals are assumed to be stationary (or at least more stationary than the speech signal $s(t)$). This will be defined more precisely in the sequel.

2.2. Usage of ATF-s ratio for TDOA extraction

Let $A_m(\omega)$ be the frequency response of the mth impulse response $a_m(t)$. Define

$$\mathscr{H}_m(\omega) \triangleq \frac{A_m(\omega)}{A_1(\omega)} \tag{3}$$

the ATF-s ratio and its corresponding impulse response $h_m(t)$. Usually, the desired TDOA value can be extracted from $h_m(t)$ by estimating its peak value location. Assume that

$$A_m(\omega) = \alpha_{n_0} e^{-j\omega n_0} + \sum_{i=1}^{L_m} \alpha_{n_i} e^{-j\omega n_i}; \quad m = 2 \ldots M,$$

$$A_1(\omega) = \beta_{p_0} e^{-j\omega p_0} + \sum_{i=1}^{L_1} \beta_{p_i} e^{-j\omega p_i}$$

with $\alpha_{n_0}, \beta_{p_0}$ and n_0, p_0 being the amplitudes and the delays of the largest peaks (not necessarily relating to the first arrival peaks) of $a_m(t)$ and $a_1(t)$, respectively. Note that we have omitted the microphone indices dependence of the gains and the delays for brevity. L_m and L_1 are the lengths of the above impulse responses. The respective ATF-s ratio can be stated as

$$\mathscr{H}_m(\omega) = \frac{\alpha_{n_0} e^{-j\omega n_0}}{\beta_{p_0} e^{-j\omega p_0}} e_m(\omega);$$

$$e_m(\omega) = \frac{1 + \sum_{i=1}^{L_m} (\alpha_{n_i} e^{-j\omega n_i} / \alpha_{n_0} e^{-j\omega n_0})}{1 + \sum_{i=1}^{L_1} (\beta_{p_i} e^{-j\omega p_i} / \beta_{p_0} e^{-j\omega p_0})}.$$

At low reverberation, where $|\alpha_{n_0}| \gg |\alpha_{n_i}|$ and $|\beta_{p_0}| \gg |\beta_{p_i}|$; $(i \neq 0)$ the error multiplicative term $e_m(\omega)$ tends to be close to 1, and the peak of the corresponding $h_m(t)$ can be used to determine the TDOA.[2] Experimental study supports this approach.

2.2.1. Preliminary simulation

To justify the use of the ATF-s ratio for TDOA extraction the following simulation was carried out. In a rectangular room with dimensions [4,7,2.75], 125 possible source locations were considered, by uniformly distributing 5 positions along each axis. A pair of microphones was placed near the center of the room at coordinates [2,3.5,1.375] and [1.7,3.5,1.375]. Using the image method [1,33], the ATF-s relating each possible source position to each microphone were simulated. Six reverberation values, denoted by T_r,[3] were considered. Ranging from low reverberation conditions ($T_r = 0.1$ s) to moderate conditions ($T_r = 0.6$ s). Two approaches were examined. TDOA estimation using ATF-s ratio, and TDOA estimation using the GCC method [27]. Assume a noise-free case, such that $z_m(t) = a_m(t) * s(t)$; $m = 1, \ldots, M$. The ATF-s ratio can be estimated from the cross-PSD divided by the auto-PSD:

$$\frac{\Phi_{z_m z_1}(\omega)}{\Phi_{z_1 z_1}(\omega)} = \frac{A_m(\omega) A_1^*(\omega) \Phi_{ss}(\omega)}{A_1(\omega) A_1^*(\omega) \Phi_{ss}(\omega)} = \mathscr{H}_m(\omega), \tag{4}$$

where $\Phi_{ss}(\omega)$ is the speech PSD at the estimation frame and * stands for conjugation. In practice, however, PSD-s will be estimated using a finite support observation window. Suppose that Welch method [42] is applied for the PSD estimation, using window $w(t)$ of length P. Denote by $\hat{\Phi}_{z_i z_j}(\omega)$ the cross-PSD estimate of z_i with z_j. Note that $\lim_{P\to\infty}(\hat{\Phi}_{z_m z_1}(\omega)/\hat{\Phi}_{z_1 z_1}(\omega)) = \mathscr{H}_m(\omega)$. However, for a finite length analysis window $w(t)$ (i.e. P is finite) $(\hat{\Phi}_{z_m z_1}(\omega)/\hat{\Phi}_{z_1 z_1}(\omega)) \neq \mathscr{H}_m(\omega)$ as the PSD

[2] We note that $h_m(t)$ is a non-causal impulse response, since ATF-s are usually non-minimum phase. Thus, evaluation of the ATF-s ratio in the Z domain, contains poles both inside and outside the unit circle.

[3] The reverberation time is the time for the acoustic energy to drop by 60 dB from its original value. This time is set using Eyring formula [28]: $T_r = (-13.82/c(L_x^{-1} + L_y^{-1} + L_z^{-1}) \ln \beta)$ with β the reflection coefficient ($\beta \in (0, 1)$), L_x, L_y, L_z are the rectangular room dimensions and c is the sound propagation speed (approximately 340 m/s in air).

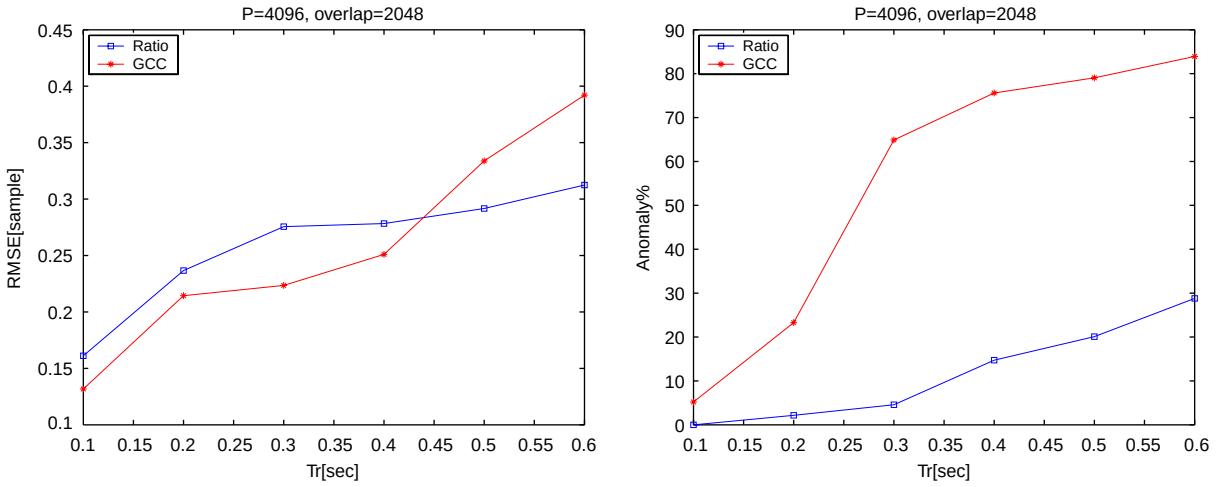

Fig. 1. Simulative model test. Long observation frame.

estimates are smoothed by a circular convolution over the $[0, 2\pi)$ interval and exact elimination of common terms in the nominator and denominator does not occur. However, for implementing a tracking system, where fast changes in $\mathscr{H}_m(\omega)$ might occur, only short observation intervals can be used. Furthermore, fast update rate and low-complexity calculations can be obtained with short observation frames. Thus, in the simulation to follow, we will present two approaches. First, long observation intervals are considered. For this purpose P was set to 4096 samples.[4] While this allows us to evaluate the ATF-s ratio for TDOA extraction, this is less practical for tracking applications. We then proceed by evaluating the PSD-s with shorter frames, i.e. $P = 256$ samples. As will be seen shortly, reasonable performance (with respect to TDOA estimation) can still be obtained. For the simulation purposes it is assumed that $s(t)$ is white, such that $\Phi_{ss}(\omega)$ is constant $\forall\omega \in [0, 2\pi)$. In practice, speech signals are non-white, and might require longer observation intervals for obtaining meaningful data in each frequency bin. Using the Welch method with Hanning window of length P, 50% overlap and 10 (weighted) periodograms in total, the PSD-s are estimated. For each source position 10 realizations of $s(t)$ are conducted, resulting in a Monte-Carlo simulation of 1250 evaluations in total. From the evaluated ATF-s ratio, the corresponding (two-sided) impulse response is extracted. To obtain sub-sample precision, the calculated impulse response is sinc-interpolated on a finer 0.1 sample resolution grid (i.e. using Shannon interpolation scheme). Finally, the TDOA is evaluated by extracting the position of the maximal peak of the impulse response. Divergence of more than one sample from the true TDOA (which is known from the geometry of the problem) is considered to be anomaly.[5] Non-anomalous estimations are considered for calculating the *root mean square error* (RMSE). Fig. 1 presents the result for $P = 4096$ samples and Fig. 2 for $P = 256$ samples. As can be seen from Fig. 1, non-anomalous estimations achieve low RMSE. However, the anomaly percentage, presents a basic difference between the methods. Note that by using the long support window the GCC method is rendered useless at $T_r = 0.3\,\mathrm{s}$ due to divergence from the ideal, reverberant-free, model. This result is compatible with the one presented by Champagne et al. [7]. On the other hand, the ATF-s ratio model maintains low anomaly percentage even at high reverberations.

[4]Throughout this work the sampling frequency is 8000 Hz.

[5]This is a reasonable value considering the microphone separation distance and the stated sampling rate. This value was chosen after visual inspection of the TDOA estimation errors, for both methods, and for all tested reverberation times.

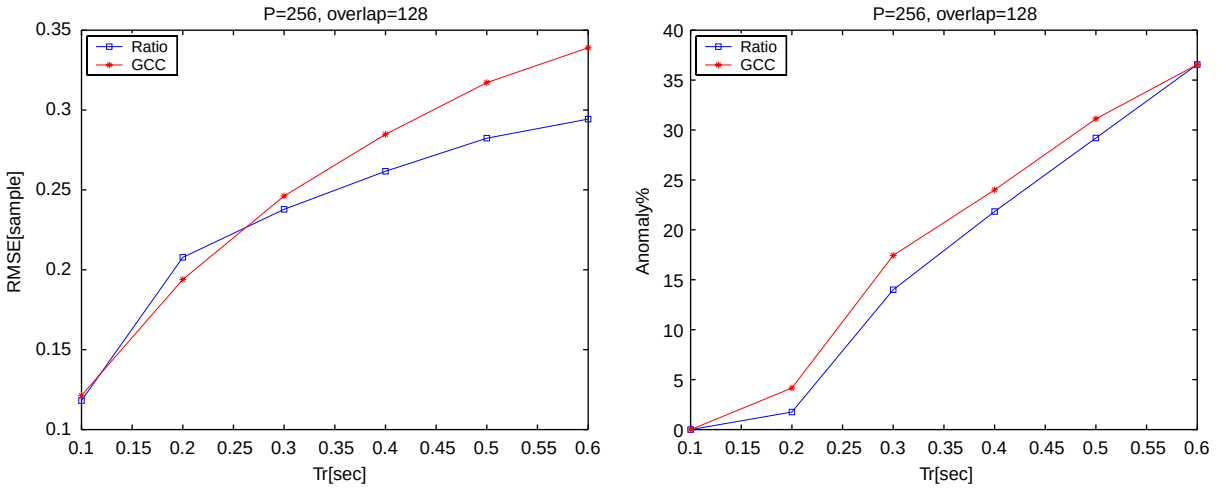

Fig. 2. Simulative model test. Short observation frame.

Fig. 2 presents the TDOA estimation results based on short observation interval. As can be seen, the methods still have small RMSE. From the presented anomaly percentage we can see that by evaluating the PSD-s with small P we can actually improve the GCC robustness to reverberation (note that the analysis carried out in [7] exploited long observation frames). Examining the anomaly percentage for the ATF-s ratio method, we notice an increase for large T_{r} values (up-till 37% at $T_{\mathrm{r}} = 0.6\,\mathrm{s}$, instead of 29% at the long frame case). However, we aim to mid range reverberation of $T_{\mathrm{r}} = 0.2-0.3\,\mathrm{s}$. Furthermore, the use of additional, spatially separated microphone pairs, which will produce additional TDOA readings, is expected to improve the actual localization. It is worth mentioning that the GCC method still suffers from another modeling assumption: it assumes uncorrelated measurement noise. In the sequel, we will demonstrate that the GCC method is rendered useless in the presence of correlated noise and low SNR conditions, while the algorithms derived in Section 3 present robust behavior.

3. Algorithm derivation—TDOA

In this section we address the problem of ATF-s ratio estimation. Quasi-stationarity of the speech signal, stationarity of the noise signal and the fact that speech and noise signals are uncorrelated are exploited for deriving several algorithms.

3.1. Speech quasi-stationarity

An unbiased method for estimating $\mathscr{H}_m(\omega)$, exploiting the speech signal quasi-stationarity, was first presented in [22], based on a method derived in [37]. Noting that the speaker and the noise source are uncorrelated, we can state the following equation:

$$\Phi_{z_i z_j}(\omega) = A_i(\omega)A_j^*(\omega)\Phi_{ss}(\omega) + B_i(\omega)B_j^*(\omega)\Phi_{nn}(\omega) \tag{5}$$

with $\Phi_{z_i z_j}$ being the cross-PSD of z_i and z_j, $\Phi_{ss}(\omega)$ is the speech auto-PSD and $\Phi_{nn}(\omega)$ is the noise auto-PSD. $B_m(\omega)$ is the frequency response of $b_m(t)$.[6] Examining (5), we note that

$$\Phi_{z_m z_1}(\omega) - \mathscr{H}_m(\omega)\Phi_{z_1 z_1}(\omega) = \Phi_{b_m^1}(\omega), \tag{6}$$

where

$$\Phi_{b_m^1}(\omega) = (\mathscr{G}_m(\omega) - \mathscr{H}_m(\omega))|B_1(\omega)|^2\Phi_{nn}(\omega) \tag{7}$$

is a noise-only term, and we define $\mathscr{G}_m(\omega) \triangleq (B_m(\omega)/B_1(\omega))$ to be the noise ATF-s

[6] Though our expressions consider a single directional interference, all the derivations can be extended for a multiple (stationary) interferers in a straight forward manner.

ratio. In practice, however, stationarity of the speech signal can be assured only over short time intervals. Consider an observation interval of length NP for which the noise signal can be regarded stationary and the ATF-s time invariant, while the speech signal statistics is changing. However, by dividing the observation interval into N consecutive frames (of length P each), the speech signal is regarded stationary for each frame. Hence, notating the frame index by $n = 1, \ldots, N$ the speech signal auto-PSD at the nth frame can be written as $\Phi_{ss}(n, \omega)$ (this is the quasi-stationarity assumption for speech signals).

By evaluating (6) for each frame, an overdetermined set of equations for $\mathscr{H}_m(\omega)$ is obtained. This set can be solved by virtue of the *least squares* (LS) method [26]. The resultant frequency domain algorithm is now presented.

Exploiting the quasi-stationarity property of the speech and defining

$$\hat{\Phi}_{b_m^1}(n,\omega) \triangleq \hat{\Phi}_{z_m z_1}(n,\omega) - \mathscr{H}_m(\omega)\hat{\Phi}_{z_1 z_1}(n,\omega); \quad n = 1, \ldots, N,$$

where, $\hat{\Phi}_{z_i z_j}(n, \omega)$ is an estimate of the PSD of z_i and z_j at the nth frame, Eq. (6) becomes a set of equations for $\mathscr{H}_m(\omega)$. This overdetermined set for $\mathscr{H}_m(\omega)$ can also be stated as

$$\hat{\Phi}_{z_m z_1}(n,\omega) = \mathscr{H}_m(\omega)\hat{\Phi}_{z_1 z_1}(n,\omega) + \Phi_{b_m^1}(\omega) + \xi(n,\omega); \quad n = 1, \ldots, N, \tag{8}$$

where, $\xi(n,\omega) \triangleq \hat{\Phi}_{b_m^1}(n,\omega) - \Phi_{b_m^1}(\omega)$ is an error term, which is minimized in the LS sense, using the overdetermined set (8). The noise-only term $\Phi_{b_m^1}(\omega)$ which is regarded stationary, and the ATF-ratio $\mathscr{H}_m(\omega)$, which is assumed to be slow time varying, are independent of the frame index (n). We denote this set of equations (or the equivalent relation in (6)) as the *first form of stationarity* (S1). The *weighted LS* (WLS) solution [26] to (8) is

$$\begin{bmatrix} \hat{\mathscr{H}}_m(\omega) \\ \hat{\Phi}_{b_m^1}(\omega) \end{bmatrix} = (\mathbf{A}^\dagger \mathbf{W} \mathbf{A})^{-1} \mathbf{A}^\dagger \mathbf{W} \underline{\hat{\Phi}}_{z_m z_1}(\omega) \tag{9}$$

with

$$\mathbf{A} \triangleq \begin{bmatrix} \hat{\Phi}_{z_1 z_1}(1,\omega), 1 \\ \vdots \\ \hat{\Phi}_{z_1 z_1}(N,\omega), 1 \end{bmatrix}; \quad \underline{\hat{\Phi}}_{z_m z_1}(\omega) \triangleq \begin{bmatrix} \hat{\Phi}_{z_m z_1}(1,\omega) \\ \vdots \\ \hat{\Phi}_{z_m z_1}(N,\omega) \end{bmatrix}.$$

$\mathbf{W}$ is an optional weight matrix and † stands for Hermitian transpose. In practice, for a non-moving source, $\mathbf{W}$ is set to the identity matrix.

Alternatively, using the same assumptions as before and by evaluating the connection between $\Phi_{z_m z_m}(\omega)$ and $\Phi_{z_1 z_m}(\omega)$ a *second form of stationarity* (S2) can be stated. Examine

$$\hat{\Phi}_{z_m z_m}(n,\omega) = \mathscr{H}_m(\omega)\hat{\Phi}_{z_1 z_m}(n,\omega) + \Phi_{b_m^2}(\omega) + \xi_2(n,\omega); \quad n = 1, \ldots, N, \tag{10}$$

where $\Phi_{b_m^2}$ is also a stationary noise-only term

$$\Phi_{b_m^2}(\omega) = (\mathscr{G}_m(\omega) - \mathscr{H}_m(\omega))B_1(\omega)B_m^*(\omega)\Phi_{nn}(\omega) \tag{11}$$

and similar to the definition of $\xi(n,\omega)$, we have the error term $\xi_2(n,\omega) \triangleq \hat{\Phi}_{b_m^2}(n,\omega) - \Phi_{b_m^2}(\omega)$, with $\hat{\Phi}_{b_m^2}(n,\omega) \triangleq \hat{\Phi}_{z_m z_m}(n,\omega) - \mathscr{H}_m(\omega)\hat{\Phi}_{z_1 z_m}(n,\omega)$.

This second form of stationarity has LS solution similar to (9)

$$\begin{bmatrix} \hat{\mathscr{H}}_m(\omega) \\ \hat{\Phi}_{b_m^2}(\omega) \end{bmatrix} = (\mathbf{B}^\dagger \mathbf{W} \mathbf{B})^{-1} \mathbf{B}^\dagger \mathbf{W} \underline{\hat{\Phi}}_{z_m z_m}(\omega) \tag{12}$$

with

$$\mathbf{B} \triangleq \begin{bmatrix} \hat{\Phi}_{z_1 z_m}(1,\omega), 1 \\ \vdots \\ \hat{\Phi}_{z_1 z_m}(N,\omega), 1 \end{bmatrix}; \quad \underline{\hat{\Phi}}_{z_m z_m}(\omega) \triangleq \begin{bmatrix} \hat{\Phi}_{z_m z_m}(1,\omega) \\ \vdots \\ \hat{\Phi}_{z_m z_m}(N,\omega) \end{bmatrix}.$$

The importance of this second form of stationarity will be clarified in the next subsection, where we relate it (and the first form of stationarity) to the decorrelation criterion.

3.2. Decorrelation criterion

Until this point we estimated $\mathscr{H}_m(\omega)$ based on noise stationarity and the speech quasi-stationarity characteristics. Though the lack of correlation between the speech and the noise term was already

used in deriving (5) it is interesting to incorporate this property directly as a part of the criterion. Namely, imposing the fact that the speaker and the interference noise must be uncorrelated.

Our observations are a mixture of the filtered speech $s_m(t) \triangleq a_m(t) * s(t)$ and the noise $n_m(t)$. As for directional noise $n_m(t) \triangleq b_m(t) * n(t)$, the cross-PSD matrix of the first and the mth microphone can be written as

$$\mathbf{P}(\omega) \triangleq \begin{bmatrix} \Phi_{z_1 z_1}(\omega) & \Phi_{z_1 z_m}(\omega) \\ \Phi_{z_m z_1}(\omega) & \Phi_{z_m z_m}(\omega) \end{bmatrix}, \tag{13}$$

where $\Phi_{z_i z_j} = A_i(\omega) A_j^*(\omega) \Phi_{ss}(\omega) + B_i(\omega) B_j^*(\omega) \Phi_{nn}(\omega)$. Applying an unmixing transformation $\mathbf{U}(\omega)$ to $[Z_1(\omega)\ Z_m(\omega)]^{\mathrm{T}}$ such that the output PSD matrix $\mathbf{R}(\omega) = \mathbf{U}(\omega)\mathbf{P}(\omega)\mathbf{U}^{\dagger}(\omega)$ is diagonal yields decorrelated outputs. We show now that a by-product of the diagonalization process will lead us to an estimate of $\mathscr{H}_m(\omega)$. In particular, and without loss of generality, by setting

$$\mathbf{U}(\omega) = \begin{pmatrix} u_1(\omega) & -1 \\ -u_2(\omega) & 1 \end{pmatrix}$$

and constraining the off-diagonal elements of $\mathbf{R}(\omega)$ to zero we obtain the (nonlinear) decorrelation criterion

$$u_2^*(\omega)(\Phi_{z_m z_1}(\omega) - u_1(\omega)\Phi_{z_1 z_1}(\omega)) = \Phi_{z_m z_m}(\omega) - u_1(\omega)\Phi_{z_1 z_m}(\omega). \tag{14}$$

Note that (14) is a single (nonlinear) equation in two unknowns. Eq. (14) was derived in [41], and it was iteratively solved in the time domain for a simplified version of the mixing channel, where the problem was constrained to FIR decoupling filters. The authors in [41] suggested to exploit speech quasi-stationarity to obtain a set of equations for $u_1(\omega)$ and $u_2(\omega)$. Indeed, by exploiting the quasi-stationarity property of the speech, Eq. (14) becomes a set of equations, obtained by evaluating the PSD-s at different frame indices

$$u_2^*(\omega)(\underline{\hat{\Phi}}_{z_m z_1}(\omega) - u_1(\omega)\underline{\hat{\Phi}}_{z_1 z_1}(\omega)) \approx \underline{\hat{\Phi}}_{z_m z_m}(\omega) - u_1(\omega)\underline{\hat{\Phi}}_{z_1 z_m}(\omega) \tag{15}$$

with

$$\underline{\hat{\Phi}}_{z_m z_1}(\omega) \triangleq \begin{bmatrix} \hat{\Phi}_{z_m z_1}(1,\omega) \\ \vdots \\ \hat{\Phi}_{z_m z_1}(N,\omega) \end{bmatrix};$$

$$\underline{\hat{\Phi}}_{z_1 z_1}(\omega) \triangleq \begin{bmatrix} \hat{\Phi}_{z_1 z_1}(1,\omega) \\ \vdots \\ \hat{\Phi}_{z_1 z_1}(N,\omega) \end{bmatrix};$$

$$\underline{\hat{\Phi}}_{z_m z_m}(\omega) \triangleq \begin{bmatrix} \hat{\Phi}_{z_m z_m}(1,\omega) \\ \vdots \\ \hat{\Phi}_{z_m z_m}(N,\omega) \end{bmatrix},$$

where N is the number of evaluated frames. For $N \geqslant 2$ we have enough equations to solve the problem, though the expressions are still nonlinear in $u_1(\omega)$ and $u_2(\omega)$. Simple assignment shows that the pair $\{u_2(\omega) = \mathscr{G}_m(\omega), u_1(\omega) = \mathscr{H}_m(\omega)\}$ as well as the pair $\{u_1(\omega) = \mathscr{G}_m(\omega), u_2(\omega) = \mathscr{H}_m(\omega)\}$ solves the equations at hand. This is referred to as the *frequency permutation ambiguity* problem[7] [32]. The authors in [41] did not present a solution to (15). In particular, they avoided the permutation problem inherent in (15), by solving the problem (iteratively) in the time domain. In this contribution we solve (15) directly to obtain an estimate for $\mathscr{H}_m(\omega)$. Furthermore, we tackle the permutation problem by exploiting noise stationarity.

3.3. Decorrelation algorithms

To maintain simplicity of the solution, we wish to solve the problem in the frequency domain. The main attraction of the frequency domain approach is its ability to translate the problem from convolutive mixture to an instantaneous mixture. Noting that the equation set (15) is nonlinear in $u_2(\omega)$ and $u_1(\omega)$, the Gauss method is employed (Appendix A). Though other search algorithms can be applied, this method was chosen due to its simplicity and since a simple way for deriving a recursive solution for it exists. This recursive

[7] Indeed this is a difficulty, since permutations in each frequency prevents consistent construction of $\mathscr{H}_m(\omega)$.

solution, which we will address in the sequel, enables tracking of a moving source.

3.3.1. Linear solution

We start by presenting a simple and non-iterative way for obtaining an estimate of $u_1(\omega) = \mathscr{H}_m(\omega)$ from set (15). Special attention will be given to avoid the permutation problem, i.e. the solution $u_1(\omega) = \mathscr{G}_m(\omega)$.

Experimental results revealed that the first (and second) form of stationarity perform well at reasonable SNR, but at negative SNR values their estimate of $\mathscr{H}_m(\omega)$ deteriorates. On the other hand, it is assumed that for negative SNR values, the estimated noise bias terms ($\hat{\Phi}_{b_m^1}(\omega)$ in (9) and $\hat{\Phi}_{b_m^2}(\omega)$ in (12)) can be reliably obtained.[8] Using (7) and (11) it is evident that

$$\frac{\Phi_{b_m^2}(\omega)}{\Phi_{b_m^1}(\omega)} = \mathscr{G}_m^*(\omega). \tag{16}$$

Thus, a possible initialization for $u_2^*(\omega)$ is

$$u_2^*(\omega) = \frac{\hat{\Phi}_{b_m^2}(\omega)}{\hat{\Phi}_{b_m^1}(\omega)}. \tag{17}$$

This assignment has a twofold advantage. First, using this initialization, the set (15) becomes a *linear* set in $u_1(\omega)$. Thus, LS solution can be obtained

$$\hat{\mathscr{H}}_m(\omega) = (\underline{V}^\dagger\, \underline{V})^{-1}\underline{V}^\dagger[\underline{\hat{\Phi}}_{z_m z_m}(\omega) - u_2^*(\omega)\underline{\hat{\Phi}}_{z_m z_1}(\omega)], \tag{18}$$

where

$$\underline{V} \triangleq \underline{\hat{\Phi}}_{z_1 z_m}(\omega) - u_2^*(\omega)\underline{\hat{\Phi}}_{z_1 z_1}(\omega)$$

and $u_2^*(\omega)$ is set according to (17). Second, by setting $u_2^*(\omega) = \mathscr{G}_m^*(\omega)$, $u_1(\omega)$ must tend to become $\mathscr{H}_m(\omega)$, thus overcoming the frequency permutation problem. The resultant algorithm is notated by *linear decorrelation* (LD) and is summarized in Fig. 3.

[8] In general, there is an inherent tradeoff in the algorithm. While estimating noise bias terms and the speaker's ATF-ratio in a single LS formulation, an accurate solution for both cannot be obtained for very high and very low SNR conditions simultaneously.

(1) Estimate $\Phi_{b_m^1}(\omega)$ using (9) and $\Phi_{b_m^2}(\omega)$ using (12).
(2) Estimate $u_2^*(\omega) = \mathscr{G}_m^*(\omega)$ using (17).
(3) Estimate $\mathcal{H}_m(\omega)$ using (18).

Fig. 3. Linear decorrelation (LD) algorithm. Batch solution.

The stated solution is a batch solution, i.e. all the available data are used at once. A recursive solution, directly applicable to the tracking problem, will be presented in the sequel.

3.3.2. Decorrelation and first form of stationarity

We now present an iterative solution to (15) based on the Gauss method. In the previous section the LD algorithm resolved the permutation problem by simply relying on a proper initialization for $u_2(\omega)$. An alternative approach (which also exploits noise stationarity), is to solve sets (15) and (8) *simultaneously* as one large LS problem. Concatenating these equations we get

$$\begin{bmatrix} \underline{\hat{\Phi}}_{z_1 z_m}(\omega) & \underline{\hat{\Phi}}_{z_m z_1}(\omega) & -\underline{\hat{\Phi}}_{z_1 z_1}(\omega) & \underline{0} \\ \underline{\hat{\Phi}}_{z_1 z_1}(\omega) & \underline{0} & \underline{0} & \underline{1} \end{bmatrix} Times \begin{bmatrix} \mathscr{H}_m(\omega) \\ \mathscr{G}_m^*(\omega) \\ \mathscr{H}_m(\omega)\mathscr{G}_m^*(\omega) \\ \Phi_{b_m^1}(\omega) \end{bmatrix} \approx \begin{bmatrix} \underline{\hat{\Phi}}_{z_m z_m}(\omega) \\ \underline{\hat{\Phi}}_{z_m z_1}(\omega) \end{bmatrix}, \tag{19}$$

where $\underline{0}$ and $\underline{1}$ stand for column vectors (of proper dimensions) of zeros and ones, respectively. Denote the parameter set by

$$\underline{\theta} \triangleq [\mathscr{H}_m(\omega), \mathscr{G}_m^*(\omega), \Phi_{b_m^1}(\omega)]^{\mathrm{T}}.$$

Denote the left-hand side of (19) by

$$\begin{aligned} \underline{h}(\underline{\theta}) \triangleq\ & \mathscr{H}_m(\omega)\begin{bmatrix} \underline{\hat{\Phi}}_{z_1 z_m}(\omega) \\ \underline{\hat{\Phi}}_{z_1 z_1}(\omega) \end{bmatrix} + \mathscr{G}_m^*(\omega)\begin{bmatrix} \underline{\hat{\Phi}}_{z_m z_1}(\omega) \\ \underline{0} \end{bmatrix} \\ & - \mathscr{H}_m(\omega)\mathscr{G}_m^*(\omega)\begin{bmatrix} \underline{\hat{\Phi}}_{z_1 z_1}(\omega) \\ \underline{0} \end{bmatrix} \\ & + \underline{\Phi}_{b_m^1}(\omega)\begin{bmatrix} \underline{0} \\ \underline{1} \end{bmatrix}. \end{aligned} \tag{20}$$

Then, Gauss iterations (see Appendix A) take the form

$$\underline{\theta}^{(l+1)} = \underline{\theta}^{(l)} + (\mathbf{H}(\underline{\theta}^{(l)})^{\dagger}\mathbf{H}(\underline{\theta}^{(l)}))^{-1}\mathbf{H}(\underline{\theta}^{(l)})^{\dagger} \times(\underline{d} - \underline{h}(\underline{\theta}^{(l)})), \tag{21}$$

where the superscript denotes the iteration index, $\mathbf{H}(\underline{\theta}^{(l)})$ is the gradient matrix at the lth iteration:

$$\mathbf{H}(\theta^{(l)}) \triangleq \nabla_{\theta}\,\underline{h}(\underline{\theta})|_{\underline{\theta}=\underline{\theta}^{(l)}} = \begin{bmatrix} \underline{\Phi}_{z_1z_m}(\omega) - \mathscr{G}_m^{(l)^*}(\omega)\underline{\hat{\Phi}}_{z_1z_1}(\omega), & \underline{\Phi}_{z_mz_1}(\omega) - \mathscr{H}_m^{(l)}(\omega)\underline{\hat{\Phi}}_{z_1z_1}(\omega), & \underline{0} \\ \underline{\hat{\Phi}}_{z_1z_1}(\omega), & \underline{0}, & \underline{1} \end{bmatrix} \tag{22}$$

and

$$\underline{d} \triangleq \begin{bmatrix} \underline{\hat{\Phi}}_{z_mz_m}(\omega) \\ \underline{\hat{\Phi}}_{z_mz_1}(\omega) \end{bmatrix}. \tag{23}$$

Two stopping criterions can be considered. First, the residual norm $\|\underline{\theta}^{(l+1)} - \underline{\theta}^{(l)}\|$ can be limited to a predefined threshold. Second, the number of the iterations can be limited a priori. We note that as in all gradient-based algorithms convergence to local minima might occur.

The resultant algorithm is denoted by *Gauss and first form of stationarity* (GS1) and is summarized in Fig. 4.

3.4. Recursive estimation

In real life scenarios we have to cope with slow changes of the noise statistics and the ATF-s (due to speaker movement). A sequential solution will allow us to perform low-complexity, low-latency algorithms which can be applied directly to the tracking problem.

3.4.1. Recursive linear LS

By applying the RLS equations (B.3) to the S1 algorithm, using forgetting factor $\alpha < 1$, slow variations of $\mathscr{H}_m(\omega)$ are trackable. This recursive solution for S1, notated by RS1, is summarized in Fig. 5.

Similarly, recursive solution can be derived for the LD algorithm. The recursive version of the LD algorithm, notated by RLD, is summarized in Fig. 6.

3.4.2. Recursion for GS1

Algorithms which employ the nonlinear decorrelation equation (14) can be solved recursively using the RG method, presented in Appendix C. For the GS1 algorithm, the parameter set is $\underline{\theta} = [\mathscr{H}_m(\omega), \mathscr{G}_m^*(\omega), \Phi_{b_m^1}(\omega)]^{\mathrm{T}}$ and the update stage includes the evaluation of *two* equations. Consider the nth time instance for which we receive the measurements $\underline{h}_n(\underline{\theta}) \approx \underline{d}_n$ with:

$$\begin{aligned} \underline{h}_n(\underline{\theta}) \triangleq\ & \mathscr{H}_m(\omega)\begin{bmatrix} \underline{\hat{\Phi}}_{z_1z_m}(n,\omega) \\ \underline{\hat{\Phi}}_{z_1z_1}(n,\omega) \end{bmatrix} + \mathscr{G}_m^*(\omega)\begin{bmatrix} \underline{\hat{\Phi}}_{z_mz_1}(n,\omega) \\ 0 \end{bmatrix} \\ & - \mathscr{H}_m(\omega)\mathscr{G}_m^*(\omega)\begin{bmatrix} \hat{\Phi}_{z_1z_1}(n,\omega) \\ 0 \end{bmatrix} + \Phi_{b_m^1}(\omega)\begin{bmatrix} 0 \\ 1 \end{bmatrix}; \\ d_n \triangleq\ & \begin{bmatrix} \hat{\Phi}_{z_mz_m}(n,\omega) \\ \hat{\Phi}_{z_mz_1}(n,\omega) \end{bmatrix}. \end{aligned} \tag{24}$$

The gradient matrix of $\underline{h}_n(\underline{\theta})$ is

$$\mathbf{H}_n(\underline{\theta}) = \begin{bmatrix} \hat{\Phi}_{z_1z_m}(n,\omega) - \mathscr{G}_m^*(\omega)\hat{\Phi}_{z_1z_1}(n,\omega), & \hat{\Phi}_{z_mz_1}(n,\omega) - \mathscr{H}_m(\omega)\hat{\Phi}_{z_1z_1}(n,\omega), & 0 \\ \hat{\Phi}_{z_1z_1}(n,\omega), & 0, & 1 \end{bmatrix}. \tag{25}$$

(1) Denote $\underline{\theta} = [\mathcal{H}_m(\omega), \mathcal{G}_m^*(\omega), \Phi_{b_m^1}(\omega)]^T$.
(2) Initialize $\mathcal{G}_m^{(0)*}(\omega)$ as in the LD algorithm, $\mathcal{H}_m^{(0)}(\omega)$ as the output of the LD algorithm and $\Phi_{b_m^1}^{(0)}(\omega)$ from the LS solution of (9).
(3) Calculate $\underline{h}(\underline{\theta})$ using (20).
(4) Calculate $\mathbf{H}(\underline{\theta})$ using (22).
(5) Set $\underline{d}$ as in (23).
(6) Iterate (21) till a pre-defined convergence criterion is reached.

Fig. 4. Gauss and first form of stationarity (GS1) algorithm. Iterative, batch solution.

(1) Use $\underline{\theta} = [\mathcal{H}_m(\omega), \Phi_{b_m^1}(\omega)]^T$.
(2) Apply (B.3) with: $\underline{a}_n^T = [\hat{\Phi}_{z_1 z_1}(n,\omega), 1]$ and $y_n = \hat{\Phi}_{z_m z_1}(n,\omega)$.

Fig. 5. Recursive solution for S1 (RS1).

(1) Use the current estimate of $\Phi_{b_m^1}(\omega)$ available from RS1 algorithm.
(2) Apply (B.3) with $\underline{\theta} = [\mathcal{H}_m(\omega), \Phi_{b_m^2}(\omega)]^T$, $\underline{a}_n^T = [\hat{\Phi}_{z_1 z_m}(n,\omega), 1]$ and $y_n = \hat{\Phi}_{z_m z_m}(n,\omega)$ for recursive estimation of $\Phi_{b_m^2}(\omega)$.
(3) Evaluate $u_2^*(\omega) = \mathcal{G}_m^*(\omega)$ using (17).
(4) Apply (B.3) with $\underline{\theta} = \mathcal{H}_m(\omega)$, $\underline{a}_n = \hat{\Phi}_{z_1 z_m}(n,\omega) - u_2^*(\omega)\hat{\Phi}_{z_1 z_1}(n,\omega)$ and $y_n = \hat{\Phi}_{z_m z_m}(n,\omega) - u_2^*(\omega)\hat{\Phi}_{z_m z_1}(n,\omega)$ for recursive estimation of $\mathcal{H}_m(\omega)$.

Fig. 6. Recursive solution for LD (RLD).

Using notations as in (C.2), the measurements for the LS problem, at the nth time instance take the simple form

$$\underline{y}_n = \underline{d}_n - \underline{h}_n(\hat{\underline{\theta}}(n-1)) + \mathbf{H}_n(\hat{\underline{\theta}}(n-1))\,\hat{\underline{\theta}}(n-1) = \begin{bmatrix} \hat{\Phi}_{z_m z_m}(n,\omega) - \hat{\mathcal{H}}_m(n-1,\omega)\hat{\mathcal{G}}_m^*(n-1,\omega)\hat{\Phi}_{z_1 z_1}(n,\omega) \\ \hat{\Phi}_{z_m z_1}(n,\omega) \end{bmatrix}, \tag{26}$$

where $\hat{\mathcal{H}}_m(n-1,\omega)$, $\hat{\mathcal{G}}_m^*(n-1,\omega)$ is the estimation of $\mathcal{H}_m(\omega)$, $\mathcal{G}_m^*(\omega)$ available after $n-1$ measurements. Since for each time instant we have two equations, the form of RLS depicted in Appendix D should be used, namely, for each time instant we perform two RLS iterations, one for each equation. The resultant recursive algorithm is denoted by RGS1 and summarized in Fig. 7.

4. Experimental study

In this section we assess the proposed algorithms, namely, S1, LD, and GS1, and compare them with the classical GCC algorithm [27] and the recently proposed subspace method (GEVD algorithm) presented by Doclo and Moonen in [15]. The latter is notated by DM.

4.1. TDOA estimation—simulation setup

We start by describing the simulation setup for TDOA estimation. Throughout this study, the sampling frequency is $F_s = 8000$ Hz. Speech signals are drawn from the TIMIT database [29] and the noise source is the speech-like noise drawn from the NOISEX-92 [39] database. Throughout the simulations speech sentences and the directional interference are filtered by the respective

Notate the current time instance by n and the sequential number of the evaluated equation by $2n+m$; $m \in \{1,2\}$. Evaluate (B.3) with:

(1) $\underline{a}^{\dagger}_{2n+m}$ is the m-th row of the 2×3 matrix $\mathbf{H}_n(\hat{\underline{\theta}}(n-1))$. $\mathbf{H}_n$ is evaluated according to (25).
(2) The current measurement y_{2n+m} is the m-th row of $\underline{y}_n$. $\underline{y}_n$ is evaluated according to (26).
(3) According to Appendix D, the forgetting factor α should be switched to 1, whenever $m \neq 1$.

Fig. 7. Recursive solution for GS1 (RGS1).

ATF-s, and summed at different SNR values to create the received microphone signals. Most of the simulations consider ATF-s created with the image method [1,33]. We also consider a static scenario simulation for which the ATF-s were obtained beforehand using real room recordings.

4.1.1. Evaluated algorithms

For the static scenarios, we evaluate the proposed batch algorithms (S1, LD, GS1). For the tracking scenario, we evaluate the recursive forms of the algorithms (RS1, RLD, RGS1). In both cases, we compare the TDOA estimation results with the classical GCC method and the subspace DM method.

Unless stated differently, the setup for the DM method is as follows:

(1) The ATF-s length is underestimated to 170 samples. This value was found to be sufficient for TDOA estimation at $T_{\mathrm{r}} = 0.25\,\mathrm{s}$, while higher values were not considered as they increased the computational demands.
(2) LMS sub-sampling is set to 10 samples.
(3) LMS step-size of 10^{-8} is used.
(4) First 20,000 samples of the noise signal are used for noise covariance matrix estimation.

For the GCC method, the entire available data of each experiment are used to produce the PSD estimates.

For all evaluated methods sub-sample TDOA calculation is performed using sinc-interpolation, on a $(T_{\mathrm{s}}/10)\,\mathrm{s}$ resolution grid, where T_{s} is the sample interval.

4.1.2. Figures of merit

The quality of the TDOA estimation algorithms is assessed by the following figures of merit:

(1) *Anomaly percentage*. Within each experiment we have defined a certain anomaly threshold. TDOA estimates which resulted in an error above the threshold were regarded as anomalies. The anomaly threshold was defined in accordance with the used sampling rate (8 kHz), the microphone separation (30 cm in most of the experiments to follow) and the TDOA difference between the noise source and the speaker. Histogram plots of the TDOA estimation errors were also used to determine the threshold.
(2) *Root mean square error* (*RMSE*) *in sample units*. The RMSE value is obtained only from non-anomalous estimates.
(3) For tracking scenario, the perceptual impression of the estimated TDOA values with respect to their true trajectory is an important figure of merit. Nevertheless, we calculated RMSE and anomaly percentage for these scenarios as well.

4.1.3. PSD estimation

Throughout the simulation we have conducted the PSD estimation using the Welch method [42]. For tracking purposes it is important to evaluate short observation intervals as the ATF-s themselves vary with time. For this purpose, and throughout the simulations, PSD estimates were obtained with Hanning analysis windows of length 256 samples and 50% overlap. Ten (weighted) periodograms were used for each PSD estimate.

For static scenarios, we allowed for 10 non-overlapping frames for each LS formulation. For statistical significance we repeated the experiments in a Monte-Carlo simulation (180 trials). For tracking scenarios it is important to achieve fast update rate in the TDOA readings. For this purpose, and opposed to the static scenarios, overlapping frames are used. In particular, in each new frame the recent periodogram is considered while the oldest periodogram is discarded. This results in strong overlapping between frames. During the tracking scenarios, the RLS algorithm is employed, where we have used a forgetting factor of $\alpha = 0.8222$.[9]

4.2. TDOA estimation—static scenarios

We start by evaluating static scenarios. Namely, scenarios for which the speaker is not moving and time invariant ATF-s relate its position with each microphone. Though for static scenarios there is no inherent constraint on the data length that can be used, we used short analysis window (as in the tracking scenario to follow). We note that the usage of small window support should reduce the reverberation effects on the GCC method, as was previously presented in Section 2.

4.2.1. Simulated ATF-s

For the first static scenario we used room dimensions of [4,7,2.75] (all dimensions are in meters). Microphone pair is placed at [2,3.5,1.375], [1.7,3.5,1.375]. Noise source positioned at [1.5,4,2.08] and speech source is placed at [2.53,4.03,2.67]. As a result, the true TDOA for the speech source is 3 samples, and for the noise source is -2.6 samples. Various reverberation times and SNR values are tested and the ATF-s are simulated using the image method [1,33]. Figs. 8 and 9 present the histogram plots of the TDOA estimation results for the various methods. If the absolute value of the TDOA estimation error was larger than 2 samples, it was considered to be an anomaly. This threshold was determined after visual inspection of the TDOA estimation histograms.

As can be seen from Fig. 8, at low reverberation conditions ($T_r = 0.1$ s) and high SNR (5 dB) all methods perform well (this might exclude the GCC method that even at these mild conditions has 16% anomaly). When we test the reverberation of $T_r = 0.5$ s, even in the high SNR level, the performance of the subspace method DM and the GCC result rapidly deteriorates. It seems that despite the use of short support analysis window P, the GCC still suffers from the lack of reverberant model. The subspace method becomes inadequate probably due to the underestimated impulse-response length. Possibly, this can be solved at the expense of increased complexity, by modeling a longer impulse responses and considering larger amounts of data. On the other hand, the simulation shows that the proposed frequency domain methods present low anomaly results. As can be seen from Fig. 9, this is also the case at mid-range reverberation $T_r = 0.25$ s and at lower SNR conditions. Note that at low SNR the decorrelation-based algorithms LD, GS1 outperform the stationarity-based algorithm S1. Furthermore, at low SNR conditions the GCC is rendered useless, since it locks on the directional interference TDOA instead of the speaker TDOA. We note that the DM method, which exploits a priori knowledge of noise covariance matrix, does not deteriorate at the low SNR conditions. However, it is still outperformed by GS1. Evaluation of the RMSE (for the non-anomalous experiments) demonstrates that the TDOA estimates of the proposed methods are extracted with high accuracy. The DM method presents a higher deviation from the true TDOA.

Next, we consider an experiment which tests the estimation accuracy of the proposed methods at various SNR conditions. Specifically, we demonstrate the advantage of the decorrelation-based methods at low SNR conditions. Using the same geometrical settings as described in this subsection, we have simulated the ATF-s at $T_r = 0.1$ s. Instead of a speech signal we have filtered a white Gaussian noise and changed its variance across frames. PSD estimates were obtained using large

[9] This value for α was set after applying some calculations involving the maximal movement speed to be considered and the asymptotic amount of data used for the TDOA estimation. More details can be found in [16].

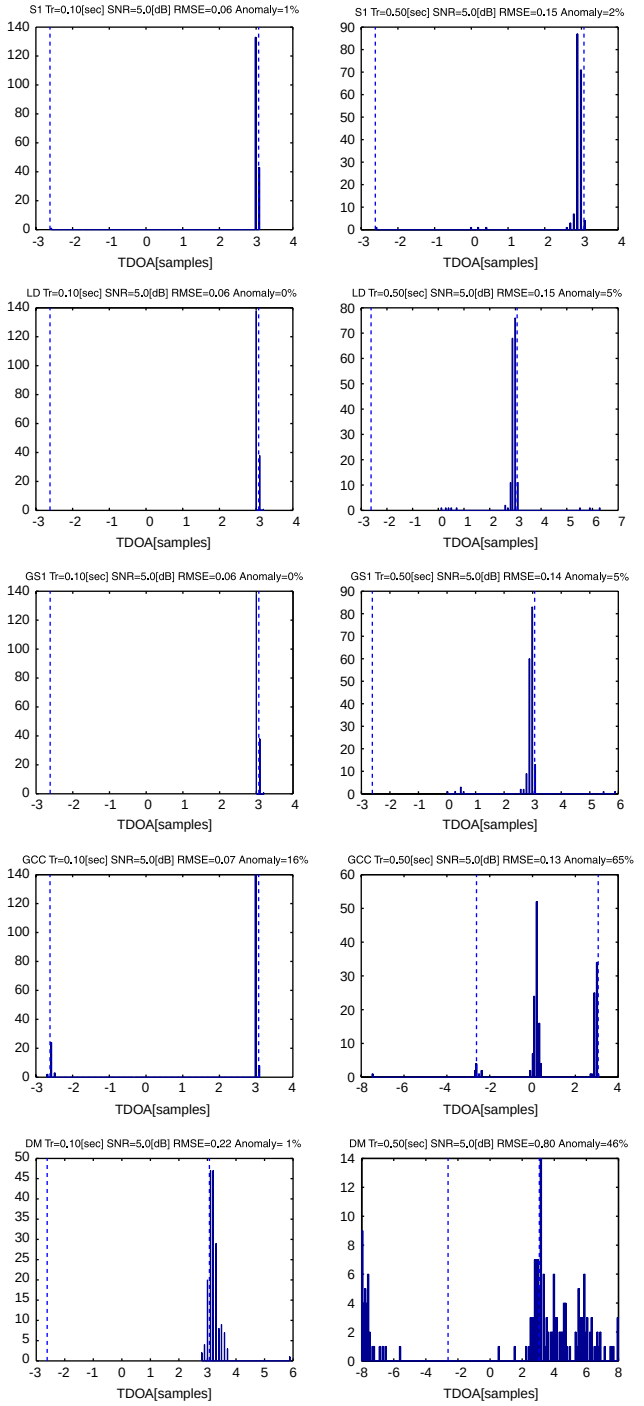

Fig. 8. Simulated ATF-s. TDOA estimation histogram plots at $T_r = 0.1$ s and $T_r = 0.5$ s. SNR = 5 dB. Dashed line at TDOA = −2.6 sample is the interference TDOA. Dashed line at TDOA = 3 sample is the original speaker TDOA. The methods name, T_r value, SNR level, RMSE of non-anomalous estimates (in samples) and the anomaly percentage for each method is stated in the title of each plot.

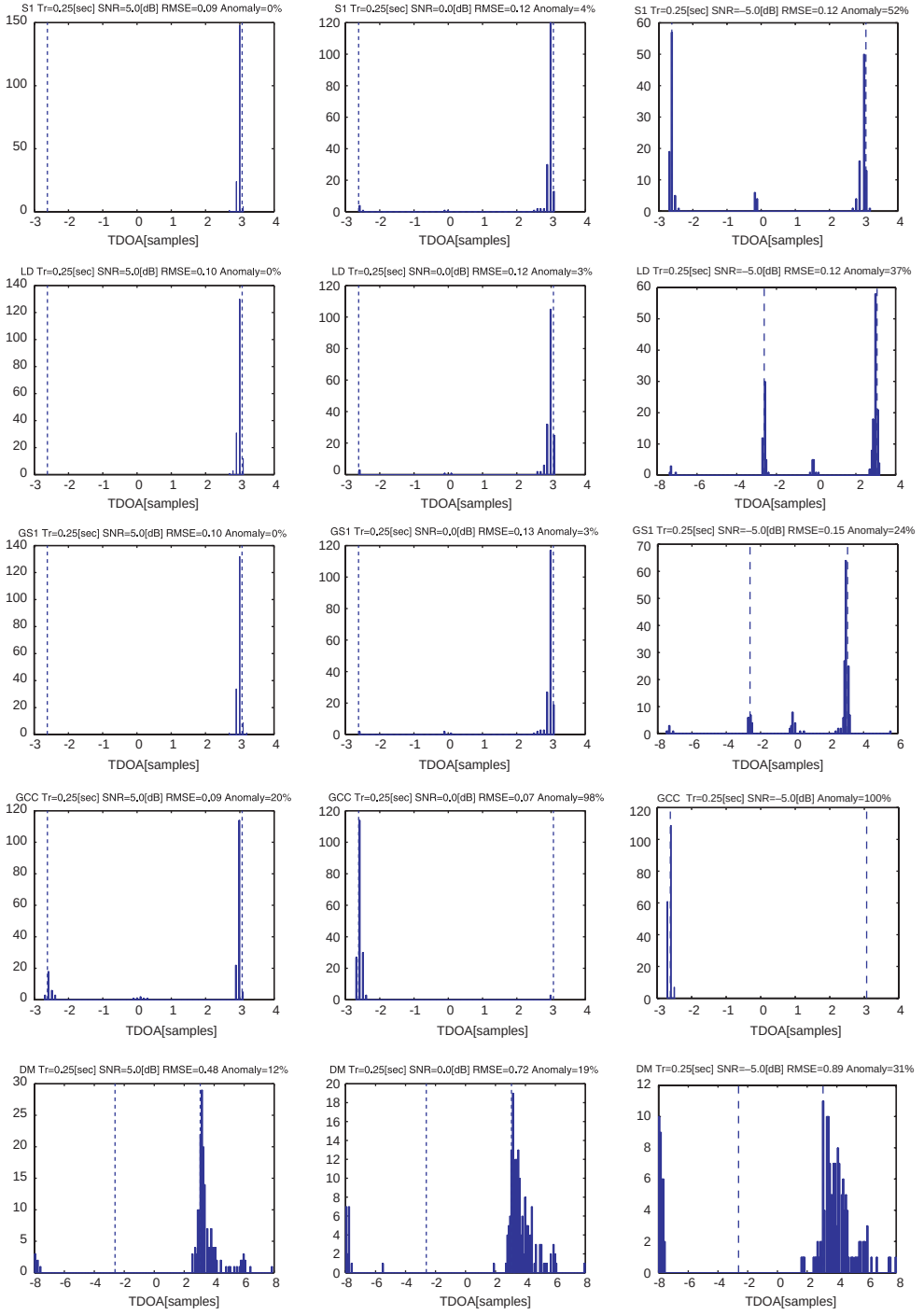

Fig. 9. Simulated ATF-s. TDOA estimation histogram plots at various SNR conditions and $T_r = 0.25$ s. Dashed line at TDOA $= -2.6$ sample is the interference TDOA. Dashed line at TDOA $= 3$ sample is the original speaker TDOA. The methods name, T_r value, SNR level, RMSE of non-anomalous estimates (in samples) and the anomaly percentage for each method is stated in the title of each plot.

frames of length 1024. $\mathscr{H}_m(\omega)$ was estimated using 3 methods; S1, LD and GS1. Notating by $\hat{\mathscr{H}}_m(\omega)$ the estimated quantity, and by $\hat{h}_m(t)$ the respective time domain impulse response, the error term

$$e \triangleq \frac{\sqrt{\sum_{t=-8}^{8}|h_m(t)-\hat{h}_m(t)|^2}}{\sqrt{\sum_{t=-8}^{8}|h_m(t)|^2}} \tag{27}$$

was evaluated. This normalized error describes how well the true impulse response was estimated around the zone of interest.[10] The results, which are presented in Fig. 10, indeed support the conjecture that the (smoothed) ATF-s ratio can be estimated more accurately by the GS1 method at low SNR conditions.

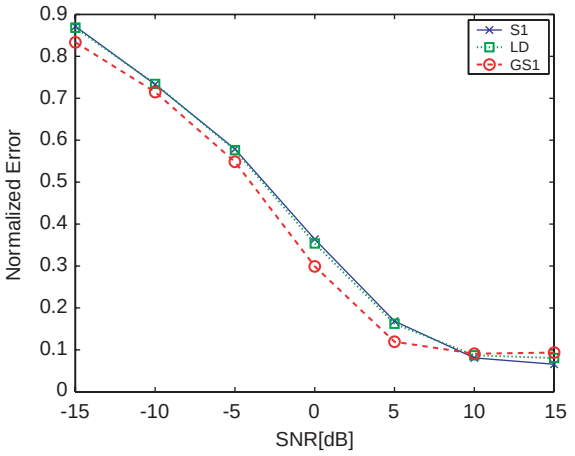

Fig. 10. Estimation error of the truncated impulse response using the proposed batch algorithms. The error is calculated according to (27).

4.2.2. Real room ATF-s

An actual room configuration is depicted in Fig. 11. Using real room recordings, the ATF-s were calculated beforehand and then used in the simulations. Fig. 11 also presents the a_1 impulse response which relates the source with one of the microphones. The respective reverberation time is less than 0.1 s, indicating low reverberation conditions. Using the geometry of the problem and the obtained ATF-s, the speaker's TDOA value is 1.5 sample and the directional noise TDOA is -1.1 sample. Figs. 12 and 13 present the TDOA estimation histograms for the evaluated algorithms, at various SNR conditions. Within this experiment, while taking into account the distribution of the obtained TDOA estimates, the anomaly threshold was set to mid way between the speech TDOA and the noise TDOA, i.e. deviations of more than 1.3 samples from the speaker's TDOA were considered to be anomalies.

Several phenomena are manifested by Figs. 12 and 13. Note that anomaly results demonstrate how well the proposed decorrelation-based methods outperform the S1 method at low SNR conditions. As in the previous experiment, here as well we see the tendency of the GCC method to lock on the stronger signal. Fig. 12 also shows that the subspace DM method performs reasonably well at negative SNR conditions, revealing small anomaly results. However, we remind the reader that the DM method uses a priori knowledge of the noise statistics, which is unavailable for the proposed methods. We note that within the positive SNR values, the proposed methods outperform the subspace DM method by presenting lower anomaly and RMSE results.

Next, we demonstrate by simulation that the proposed methods are sensitive to the noise stationarity assumption. For that purpose we consider the same setting as in Fig. 11, where the noise source is now set to be another, non-correlated speaker (taken from the TIMIT database [29] as well). The SNR is set to 0 dB and 88 experiments, using different speech sections are evaluated. The TDOA estimation results are summarized as histograms in Fig. 14. As can be seen, the proposed methods cannot distinguish between the speaker and the interference as both signals are non-stationary and of the same energy level. Due to the same energy level, the GCC method cannot select the desired TDOA as well. The figure also shows that though the subspace method does not lock on the noise TDOA reading, it fails to cope with this problem as well.

[10] Taking into account the microphone separation and the sampling rate, we note that a search zone of $[-8, 8]$ samples is sufficient for evaluating the estimation accuracy.

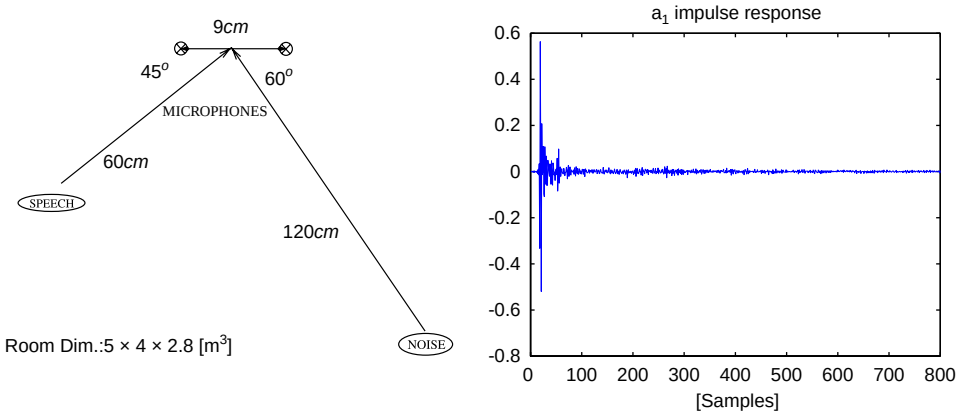

Fig. 11. Real room ATF-s. Left: Geometric configuration. Right: a_1 impulse response.

4.3. TDOA estimation—tracking scenario

We proceed by discussing the tracking scenario in which a moving speaker is considered. Room dimensions and the noise source position are as in the first static scenario, depicted in Section 4.2.1. The speaker trajectory is set to an helix with radius $R = 1.5$ m around the reference microphone, at movement speed of 0.5 m/s and for a total movement time of $T = 30$ s. The speaker Cartesian position as a function of time $t \in [0, T]$ is,

$$x(t) = 2 + R\cos(2\pi ft), \quad y(t) = 3.5 + R\sin(2\pi ft),$$
$$z(t) = 1 + \tfrac{t}{T}$$

with $f = 0.0529$ Hz. This trajectory is depicted in Fig. 15. TDOA estimation results are presented with respect to the microphone pair placed at [2,3.5,1.375], [2.3,3.5,1.375]. Sampling every 3.75 cm along the speaker trajectory, the ATF-s between the speaker and the microphones are simulated using the image method and used to filter the speech. Reverberation time is set to 0.25 s. The mean SNR for the 30 s long signal is set to a relatively high value of 10 dB for producing reasonable results. The TDOA extraction procedures are the same as in the static scenario. However, for the proposed methods, we now solve the LS problem recursively with a forgetting factor smaller than 1 and use overlapping frames.

4.3.1. Tracking scenario—evaluation

We proceed by presenting estimation results for 6 methods. Recursive forms of S1 (RS1), LD (RLD) and GS1 (RGS1) are evaluated and compared with the GCC and DM methods for the tracking scenario. Here, we further consider the adaptive eigenvalue decomposition method, proposed by Benesty[11] [2] and denoted here by EVD. For the latter, a step-size of 10^{-7} is used. In order of the subspace methods (i.e. EVD and DM) to work in the tracking scenario, Doclo [14] proposed to slightly modify the algorithms by introducing intermediate initializations, reducing the LMS sub-sampling to 1 sample and using underestimated ATF-s 20 taps long. Using an anomaly threshold of 2 samples, Fig. 16 presents the TDOA estimation plots for the different methods.

As can be seen from Fig. 16 the subspace methods have difficulties in locking on the relatively fast changing ATF-s, thus introducing large anomaly percentage. We note that despite the relatively high mean SNR, the instantaneous SNR might be low. This causes the EVD method, and especially the GCC method, to lock on the noise TDOA reading (which is approximately at 4.2 samples) during low-SNR time epoches. As the DM method takes into account the noise field, it

[11] Since the mean SNR is relatively high, applying this subspace method is at place here.

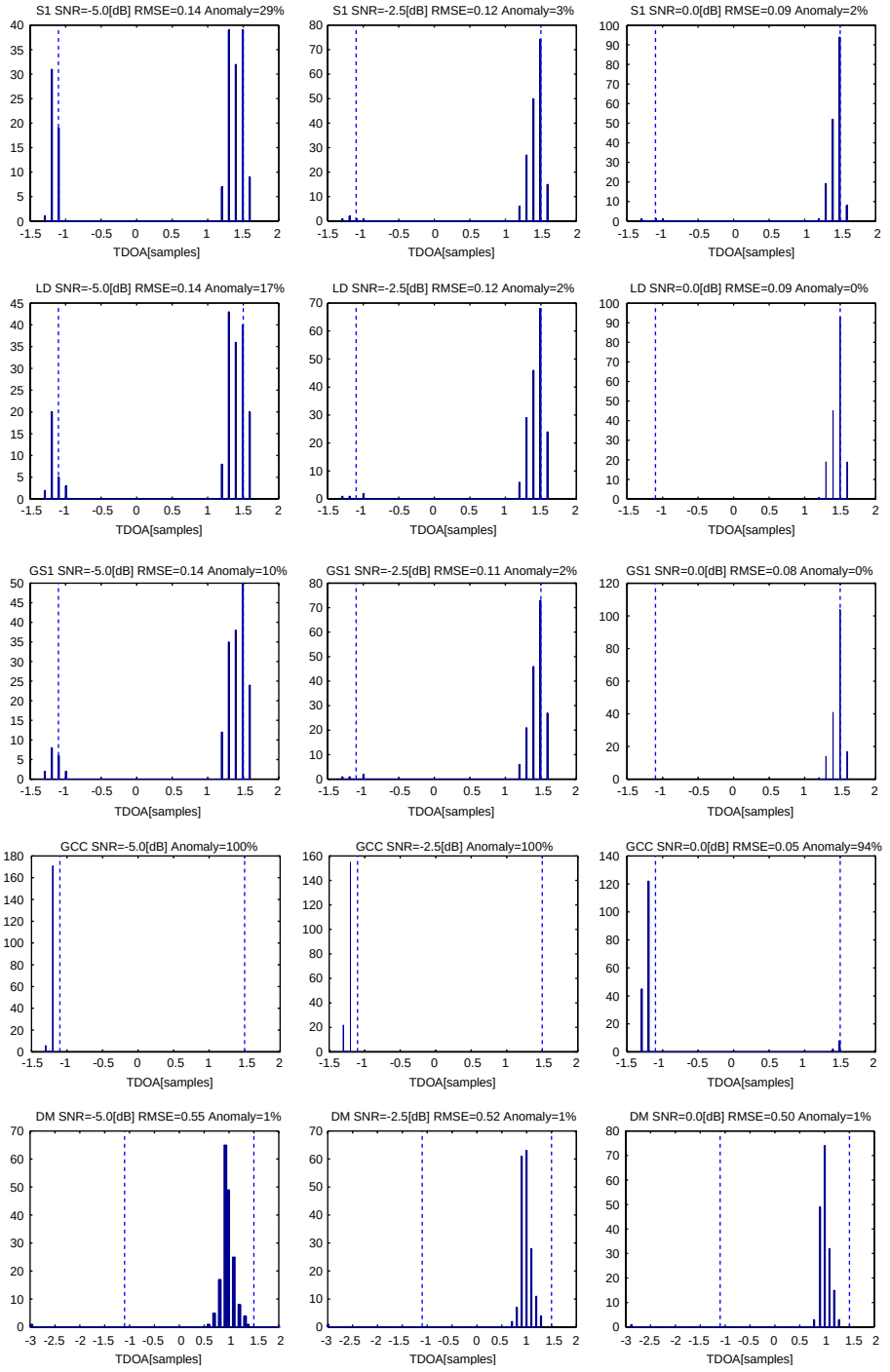

Fig. 12. Room measured ATF-s. TDOA estimation histogram plots at low SNR conditions. Dashed line at TDOA $= -1.1$ sample is the interference TDOA. Dashed line at TDOA $= 1.5$ sample is the original speaker TDOA. The methods name, SNR level, RMSE of non-anomalous estimates (in samples) and anomaly percentage for each method is stated in the title of the plot.

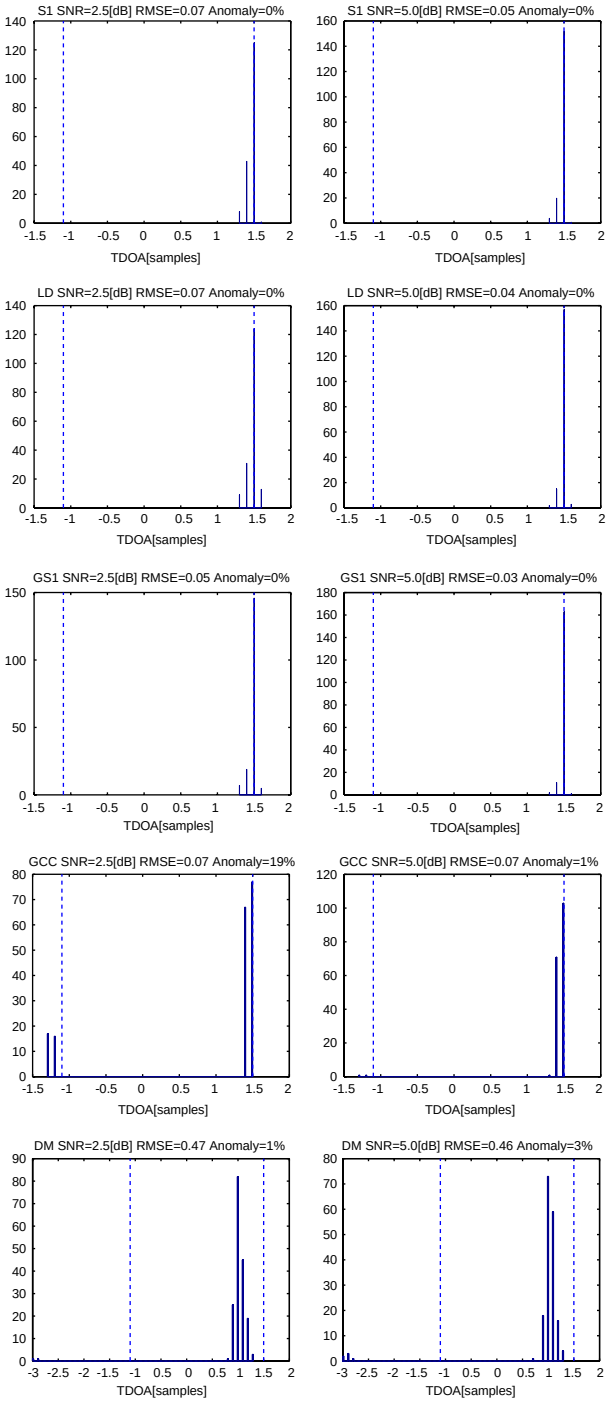

Fig. 13. Room measured ATF-s. TDOA estimation histogram plots at high SNR conditions. Dashed line at TDOA = −1.1 sample is the interference TDOA. Dashed line at TDOA = 1.5 sample is the original speaker TDOA. The methods name, SNR level, RMSE of non-anomalous estimates (in samples) and anomaly percentage for each method is stated in the title of the plot.

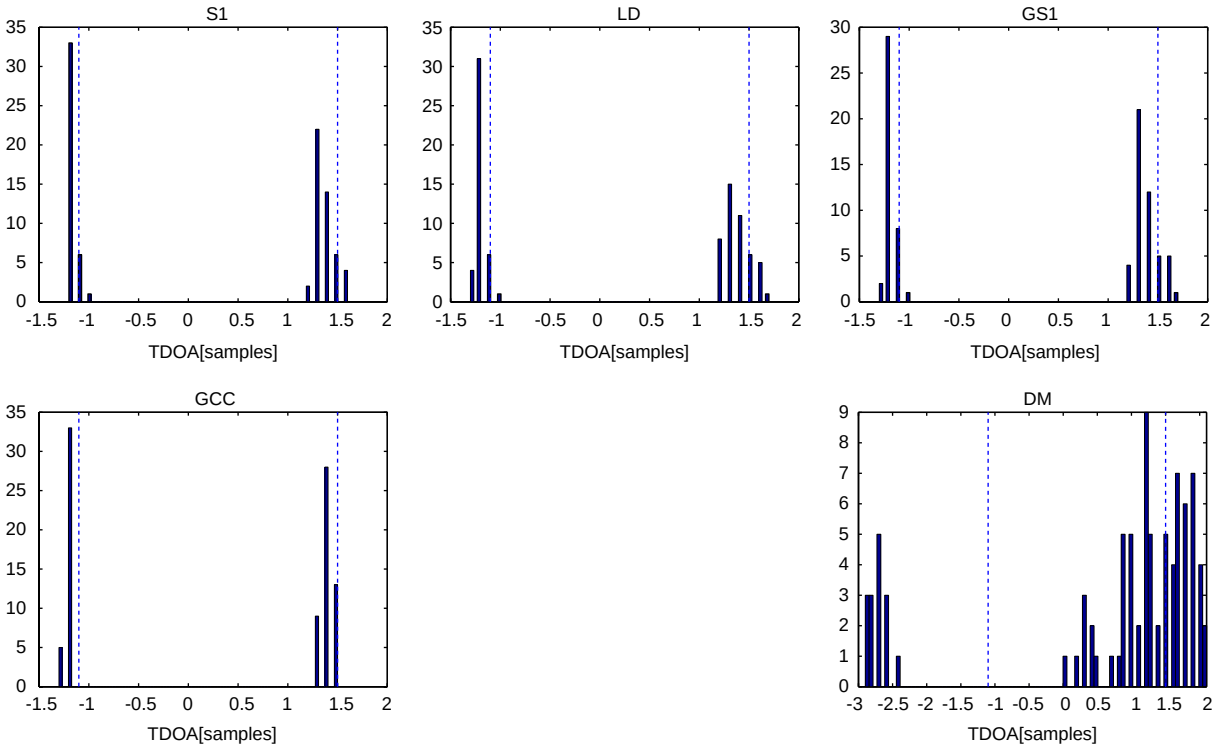

Fig. 14. TDOA estimation histogram plots. The interference is a second speaker and the SNR level is set to 0 dB. Dashed line at TDOA = −1.1 sample is the interference TDOA. Dashed line at TDOA = 1.5 sample is the original speaker TDOA.

does not have the EVD tendency to lock on the noise, but still many of its readings are erroneous, especially when the speech TDOA is close to the noise TDOA. In contrast, the proposed methods (RS1, RLD, RGS1) usually manage to track the changes in the speaker TDOA. We note, however, that due to the memory introduced by the RLS-based algorithm, time instances where wrong TDOA is estimated, cause the estimated trajectory to slightly distract from the real trajectory. This puts a limitation on the ability of the proposed algorithms to track fast moving source. In fact, we note that by varying the value of the forgetting factor α one can counter balance between tracking abilities and the (stationary) noise suppression abilities of the proposed algorithms. We further note that the decorrelation methods (RLD, RGS1) did not yield an improvement in the tracking scenario compared with the simpler RS1 method. This is probably due to the additional approximation which is present in the RG method

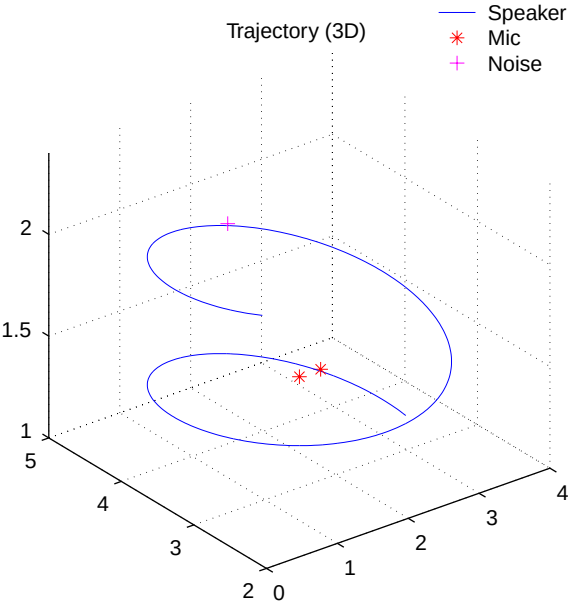

Fig. 15. Speaker trajectory.

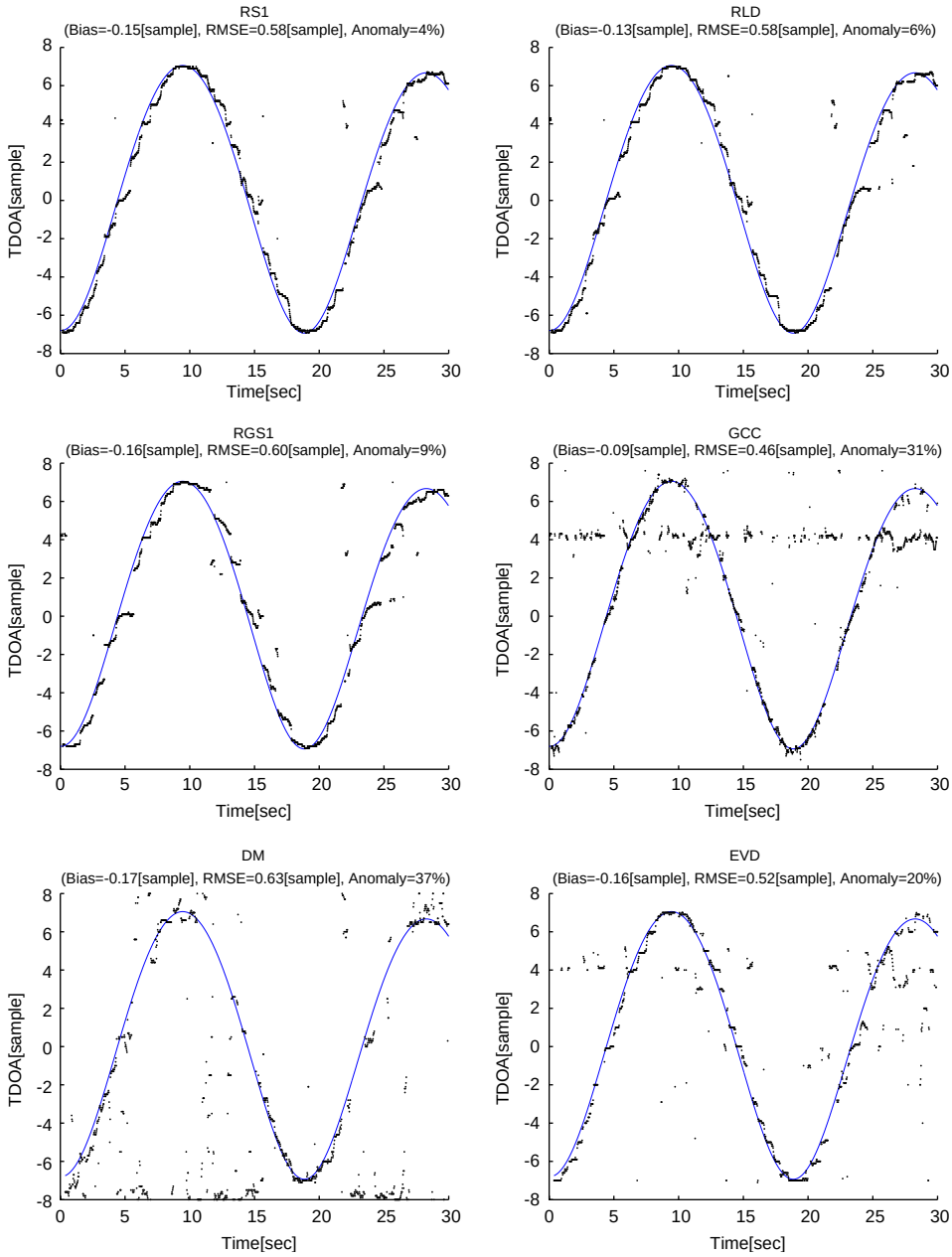

Fig. 16. Moving source scenario, TDOA estimation results. Solid line: True TDOA. Dots: Estimation results. The method's name, its bias, RMSE and anomaly results are presented in the title of each plot.

and not suitable for this relatively fast changing scenario.

It is interesting to investigate the performance of the suggested algorithms in a stationary diffused noise field as well. Diffused noise field is typical to car environments. The respective coherence function between two sensors is given by $\gamma(\omega) = (\Phi_{z_i z_j}(\omega)/\sqrt{\Phi_{z_i z_i}(\omega)\Phi_{z_j z_j}(\omega)}) = (\sin(\omega(d/c))/\omega(d/c))$ (where d is the microphone separation distance and c is the sound

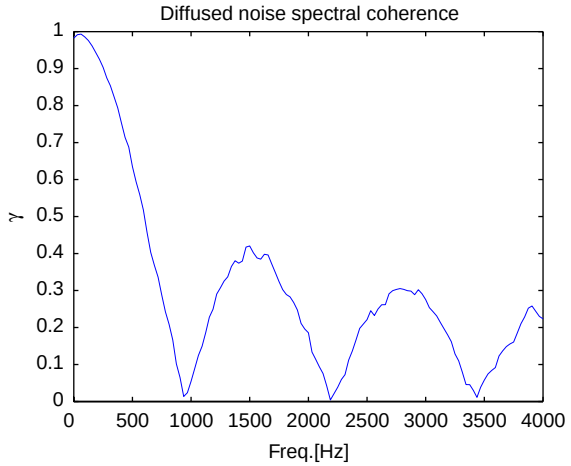

Fig. 17. Coherence function for the diffused noise.

propagation speed). This coherence function exhibits correlation between sensors at the low-frequency band and almost no correlation at the high-frequency band. In our study we used a simulated diffused noise, which coherence is depicted in Fig. 17. The obtained performance is demonstrated in Fig. 18. As expected, the proposed methods are robust to the noise field as long as it is stationary, and track the true TDOA of the speaker. As at most of the frequencies, the additive noise is present and spatially correlated, the GCC method has poor performance. Interestingly, most of its estimates are located near the boundary of the TDOA search zone (less than [−8, 8] samples for 30 cm microphone separation). We note that we have noticed this phenomenon for static

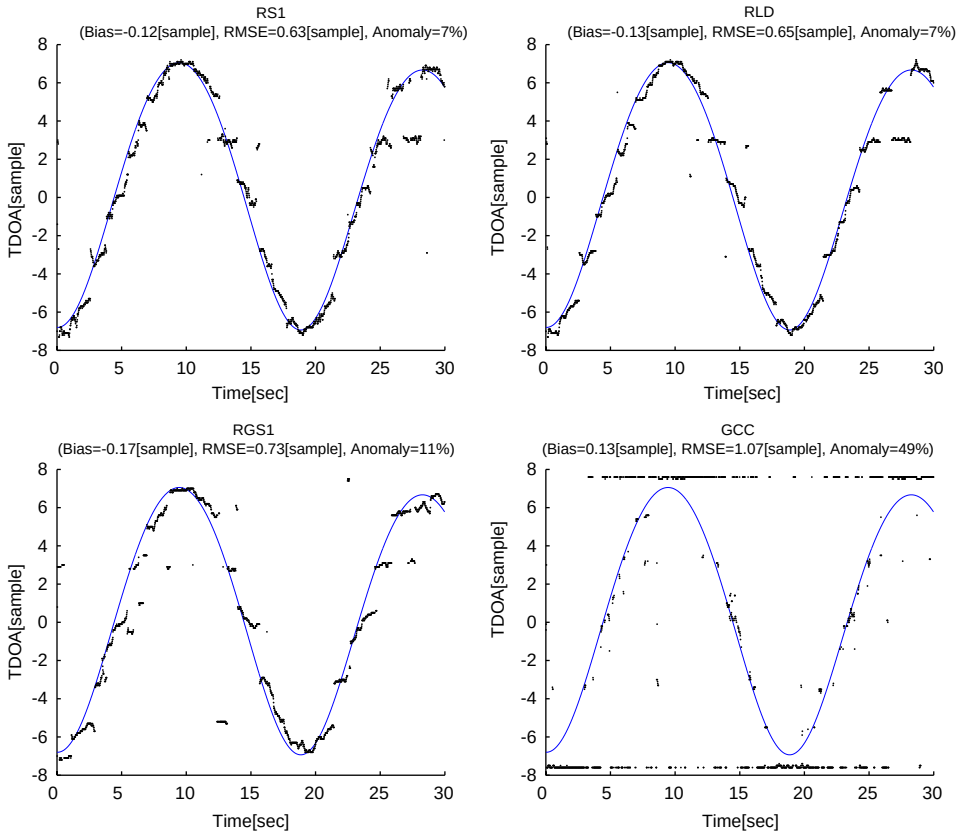

Fig. 18. TDOA estimation results in diffused noise. Solid line: True TDOA. Dots: Estimation results. The method's name, its bias, RMSE and anomaly results are presented in the title of each plot.

scenarios as well, and for different settings of the search zone.

As the simpler RS1 method performed better in the tracking scenarios only the latter is presented in the evaluation to follow.

4.3.2. Switching scenario—evaluation

Consider the following simulation which is typical for a video conference scenario. Two speakers, located at two different and fixed locations alternately speak. The camera should be able to maneuver from one person to the other. As in the previous experiment we have used room dimensions [4,7,2.75], $T_r = 0.25$ s, mean SNR of 10 dB and a horizontal microphone pair at [2,3.5,1.375], [2.3,3.5,1.375]. Here, we consider additional TDOA estimation results, obtained by a vertical microphone pair, placed at [2,3.5,1.375], [2,3.5,1.675]. One speaker was placed at the position [2.75,4.75,2.436] and the other speaker at [1.47,4.03,2.674]. A directional interference was placed at the position [2,4.207,2.082]. Fig. 19 presents the TDOA estimation results by the RS1 algorithm (which gave the best results), for the previously mentioned horizontal and vertical microphone pairs. For this experiment, and in face of the results, anomaly was defined as divergence of more than 0.5 sample from the true TDOA. As can be seen from the figure, for the stated scenario the algorithm demonstrates excellent tracking capabilities.

5. Practical considerations

5.1. VAD

As previously mentioned, and opposed to the DM algorithm, the suggested algorithms do not rely on a VAD for the estimation of noise statistics. This is achieved by simultaneously estimating noise related terms and speech related terms in a single LS framework. Unfortunately, as a result, both cannot be estimated reliably in very low and very high SNR conditions. While in high SNR conditions this is not a difficulty (as the noise related term is regarded as a nuisance parameter), during low SNR conditions we suggest the usage of the decorrelation-based methods. This is particularly useful for static scenarios, where the approximations made by the RG method may be avoided. We note that for high SNR conditions and tracking scenarios we recommend the simpler S1 method.

5.2. Computational efficiency

Next, we wish to consider the computational complexity for the suggested frequency-domain algorithms. Denote by P the periodogram length and by K the periodogram shift involved in the Welch PSD estimation. Applying one iteration of the RLS algorithm on a parameter set $\theta \in \mathscr{C}^p$

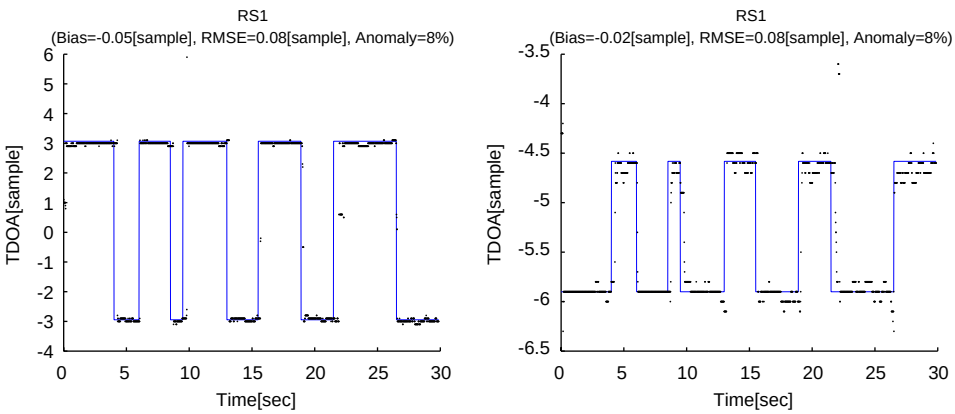

Fig. 19. TDOA estimation results for the RS1 method in the switching scenario. Solid line: True TDOA. Dots: Estimation results. Left plot: TDOA estimation results with respect to a horizontal microphone pair. Right plot: TDOA estimation results with respect to a vertical microphone pair. The method's bias, RMSE and anomaly results are presented in the title of each plot.

involves $10p^2 + 12p$ real multiplications and one complex division. Noting that RLS iteration is performed in each frequency bin and that there are $P/2$ frequencies to evaluate, the total number of real multiplications performed by the RLS is $P(10p^2 + 12p)/2$. The suggested frequency domain algorithms further involve one IFFT operation and interpolation. Assuming that the interpolation is conducted for S samples[12] with a 1/10 sample resolution, the last stage involves approximately $2P\log_2 P + 10S^2$ real multiplications. Consider for example the RS1 algorithm. Cross-PSD $\Phi_{z_m z_1}(\omega)$ and auto-PSD $\Phi_{z_1 z_1}(\omega)$ can be compactly evaluated for every new K samples using

$$2\left(P + 2P\log_2 P + \frac{3P}{2}\right) = 2P(2.5 + 2\log_2 P)$$

real multiplications. Considering the RLS iterations and the time domain post-processing, the computational burden *per sample* is

$$\frac{2P(2.5 + 2\log_2 P) + P(10p^2 + 12p)/2 + 2P\log_2 P + 10S^2}{K}$$

real multiplications. For RS1 algorithm $p = 2$. Assuming that $S = 17$, $P = 256$, $K = 128$ this yields approximately 193 multiplications per sample. We note that this burden is higher than the one imposed by the simpler GCC method, but it is usually much lower than the burden imposed by the subspace methods.

6. Summary

In this work novel TDOA estimation algorithms, based on the ATF-s ratio $\mathscr{H}_m(\omega)$ for TDOA extraction, were presented. Speech quasi-stationarity, noise stationarity and the fact that there is no correlation between the speech and the noise were used for $\mathscr{H}_m(\omega)$ estimation. Noise stationarity was employed for resolving frequency permutation ambiguity, inherent to the frequency domain decorrelation criterion. Simulation results revealed superiority over the classical *generalized cross correlation* (GCC) method and the recently proposed subspace method. Preliminary experiments showed that the use of short support analysis window can improve the robustness of the GCC algorithm to reverberation. Computational considerations, presented in Section 5, revealed that the suggested frequency domain methods result in relatively low computational costs. Special care was given to recursive implementation which is applicable for the tracking scenario. This resulted in a general formulation, notated by *recursive Gauss*, for recursive solution of a nonlinear equation set.

[12]The region of interest for conducting the interpolation is bounded by the microphone pair separation.

Acknowledgements

We would like to thank Dr. Simon Doclo from K.U. Leuven, Belgium for generously making his simulation code available for us and for his valuable remarks. We also would like to thank the anonymous reviewers for their thorough work and valuable comments.

Appendix A. The Gauss method

Let $\underline{\theta} \in \mathbb{C}^p$ be an unknown $p \times 1$ parameter vector, which is measured through K nonlinear equations $\underline{h}$ resulting a measurement vector $\underline{v}$

$$\underline{h}(\underline{\theta}) = \underline{v}.$$

Expansion of $\underline{h}(\underline{\theta})$ around $\underline{\theta}^{(0)}$, using first-order approximation becomes

$$\underline{h}(\underline{\theta}) \approx \underline{h}(\underline{\theta}^{(0)}) + \mathbf{H}(\underline{\theta}^{(0)})(\underline{\theta} - \underline{\theta}^{(0)}) \tag{A.1}$$

with $\mathbf{H}$ being a $K \times p$ gradient matrix such that $\mathbf{H}_{k,q} = \partial \underline{h}_k / \partial \underline{\theta}_q$. Thus

$$\mathbf{H}(\underline{\theta}^{(0)})\,\underline{\theta} \approx \underline{v} - \underline{h}(\underline{\theta}^{(0)}) + \mathbf{H}(\underline{\theta}^{(0)})\underline{\theta}^{(0)}.$$

When $K > p$, this is an overdetermined set which can be solved in the LS sense, resulting the iterative algorithm

$$\underline{\theta}^{(l+1)} = (\mathbf{H}(\underline{\theta}^{(l)})^{\dagger}\mathbf{H}(\underline{\theta}^{(l)}))^{-1}\mathbf{H}(\underline{\theta}^{(l)})^{\dagger}(\underline{v} - \underline{h}(\underline{\theta}^{(l)}) + \mathbf{H}(\underline{\theta}^{(l)})\underline{\theta}^{(l)}) \tag{A.2}$$

which is equivalent to

$$\underline{\theta}^{(l+1)} = \underline{\theta}^{(l)} + (\mathbf{H}(\underline{\theta}^{(l)})^{\dagger}\mathbf{H}(\underline{\theta}^{(l)}))^{-1}\mathbf{H}(\underline{\theta}^{(l)})^{\dagger} \times(\underline{v} - \underline{h}(\underline{\theta}^{(l)})). \tag{A.3}$$

Appendix B. Recursive least squares

Sequential solution to the linear LS problem $\mathbf{A}\,\underline{\theta} \approx \underline{y}$ can be obtained on a frame-by-frame basis, using the *recursive least squares* (RLS) algorithm. Consider a weighted LS (WLS) problem for estimating the parameter set $\underline{\theta} \in \mathbb{C}^p$ based on N equations:

$$\hat{\underline{\theta}}(N) = \arg\min_{\underline{\theta}}(\mathbf{A}_{1:N}\,\underline{\theta} - \underline{y}_{1:N})^{\dagger}\mathbf{W}_{1:N}(\mathbf{A}_{1:N}\,\underline{\theta} - \underline{y}_{1:N}) \tag{B.1}$$

with

$$\mathbf{W}_{1:N} = \begin{bmatrix} \alpha^{N-1} & 0 & \dots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \alpha & 0 \\ 0 & \dots & 0 & 1 \end{bmatrix} \tag{B.2}$$

a diagonal $N \times N$ weight matrix, with the nth element along the diagonal set to α^{N-n}. α is the forgetting factor, $0 < \alpha \leqslant 1$. $\mathbf{A}_{1:N}$ stands for an $N \times p$ matrix and $\underline{y}_{1:N}$ is an $N \times 1$ measurement vector

$$\mathbf{A}_{1:N} \triangleq \begin{bmatrix} \underline{a}_1^{\dagger} \\ \vdots \\ \underline{a}_N^{\dagger} \end{bmatrix}; \quad \underline{y}_{1:N} \triangleq \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

with $\underline{a}_n$; $n = 1, \dots, N$ a $p \times 1$ vector. Then, the recursive solution to (B.1) takes the known form (see for example [24,26]):

$$\underline{K}_n = \frac{\mathbf{P}_{n-1}\underline{a}_n}{\alpha + \underline{a}_n^{\dagger}\mathbf{P}_{n-1}\underline{a}_n},$$

$$\hat{\underline{\theta}}(n) = \hat{\underline{\theta}}(n-1) + \underline{K}_n(y_n - \underline{a}_n^{\dagger}\hat{\underline{\theta}}(n-1)),$$

$$\mathbf{P}_n = \left(\sum_{t=1}^{n} \alpha^{n-t}\underline{a}_t\underline{a}_t^{\dagger}\right)^{-1} = (\mathbf{P}_{n-1} - \underline{K}_n\underline{a}_n^{\dagger}(\mathbf{P}_{n-1})\frac{1}{\alpha}, \tag{B.3}$$

where $\mathbf{P}_n$ is the weighted inverse. To avoid direct calculation of the initial inverse $\mathbf{P}_0$, a common approach is to use the diagonal initialization $\mathbf{P}_0 = \beta\mathbf{I}$ with $\beta \gg 1$.

Appendix C. Recursive nonlinear least squares

In this appendix, a method is derived for recursive estimate of a nonlinear LS problem. The method first resolves the nonlinearities by first-order approximation, as in the Gauss method. Then, by proper approximation, a recursion is derived. We denote this recursive procedure by RG.

Consider a nonlinear equation set for a $p \times 1$ parameter vector $\underline{\theta} \in \mathbb{C}^p$

$$\underline{h}_{1:N}(\underline{\theta}) = \underline{d}_{1:N}$$

with

$$\underline{h}_{1:N}(\underline{\theta}) \triangleq \begin{bmatrix} h_1(\underline{\theta}) \\ \vdots \\ h_N(\underline{\theta}) \end{bmatrix}; \quad \underline{d}_{1:N} \triangleq \begin{bmatrix} d_1 \\ \vdots \\ d_N \end{bmatrix}.$$

Applying first-order approximation around an initial guess $\underline{\theta}^{(0)}$ (as with the Gauss method) we obtain

$$\underline{h}_{1:N}(\underline{\theta}^{(0)}) + \mathbf{H}_{1:N}(\underline{\theta}^{(0)})(\underline{\theta} - \underline{\theta}^{(0)}) \approx \underline{d}_{1:N}, \tag{C.1}$$

where $\mathbf{H}_{1:N}$ is the $N \times p$ gradient matrix

$$\mathbf{H}_{1:N}(\underline{\theta}) \triangleq \begin{bmatrix} H_1(\underline{\theta}) \\ \vdots \\ H_N(\underline{\theta}) \end{bmatrix}$$

with $H_n(\underline{\theta}) = \nabla_{\underline{\theta}} h_n(\underline{\theta})$ the gradient row vector of $h_n(\underline{\theta})$. According to the Gauss method, the iterative LS solution to the linearized set (C.1) is

$$\underline{\theta}^{(l+1)} = (\mathbf{H}_{1:N}(\underline{\theta}^{(l)})^{\dagger}\mathbf{H}_{1:N}(\underline{\theta}^{(l)}))^{-1}\mathbf{H}_{1:N}(\underline{\theta}^{(l)})^{\dagger} \times(\underline{d}_{1:N} - \underline{h}_{1:N}(\underline{\theta}^{(l)}) + \mathbf{H}_{1:N}(\underline{\theta}^{(l)})\underline{\theta}^{(l)}),$$

where the superscript denotes the iteration number. Consider the next measurement $h_{N+1}(\underline{\theta}) = d_{N+1}$ available at time instance $N+1$. In order to estimate $\underline{\theta}$ we will use all the available measurements simultaneously. Though we could approximate all $N+1$ equations at the current estimate

$\underline{\theta}^{(l+1)}$, we will do so *only* for the new equation. Namely, instead of minimizing in the LS sense the following residual norm

$$\min_{\underline{\theta}} \|\underline{d}_{1:N+1} - (\underline{h}_{1:N+1}(\underline{\theta}^{(l+1)}) + \mathbf{H}_{1:N+1}(\underline{\theta}^{(l+1)})(\underline{\theta} - \underline{\theta}^{(l+1)}))\|$$

we will minimize

$$\min_{\underline{\theta}} \left\| \begin{array}{c} \underline{d}_{1:N} - (\underline{h}_{1:N}(\underline{\theta}^{(l)}) + \mathbf{H}_{1:N}(\underline{\theta}^{(l)})(\underline{\theta} - \underline{\theta}^{(l)})) \\ d_{N+1} - (h_{N+1}(\underline{\theta}^{(l+1)}) + \mathbf{H}_{N+1}(\underline{\theta}^{(l+1)})(\underline{\theta} - \underline{\theta}^{(l+1)})) \end{array} \right\|.$$

The reason for this approximation is to keep past solutions intact, i.e. when new equation becomes available there is no need to update past solutions based on the new equation, thus, enabling a recursive solution to be derived. Now, using *stochastic approximation*, i.e. replacing the iteration index by the time index, a sequential algorithm is obtained. To summarize the procedure, an estimate for $\underline{\theta}$ at the current time instance n (denoted by $\hat{\underline{\theta}}(n)$) is obtained by solving the following LS problem sequentially using the RLS procedure

$$\hat{\underline{\theta}}(n) = \arg\min_{\underline{\theta}} \left\| \begin{bmatrix} H_1(\hat{\underline{\theta}}(0)) \\ \vdots \\ H_n(\hat{\underline{\theta}}(n-1)) \end{bmatrix} \underline{\theta} - \underline{y}_{1:n} \right\|, \quad \text{(C.2)}$$

where

$$\underline{y}_{1:n} \triangleq \begin{bmatrix} d_1 - h_1(\hat{\underline{\theta}}(0)) + H_1(\hat{\underline{\theta}}(0))\hat{\underline{\theta}}(0) \\ \vdots \\ d_n - h_n(\hat{\underline{\theta}}(n-1)) + H_n(\hat{\underline{\theta}}(n-1))\hat{\underline{\theta}}(n-1) \end{bmatrix}$$

with $\hat{\underline{\theta}}(0)$ the initial estimate for the parameter set. Recalling that in tracking problems the parameter set $\underline{\theta}$ might slowly vary with time, a common practice is to apply the RLS algorithm with a diagonal weight matrix, as depicted in (B.2).

We also note that the resulting algorithm can be viewed as a special case of the *extended Kalman filter*. The detailed derivation is out of the scope of this work and can be found in [18].

Appendix D. Recursive least squares for multiple readings

Assume a scenario in which for each time instance we have K scalar measurements $\underline{z}_t \in \mathbb{C}^K$ related to an unknown $p \times 1$ parameter vector $\underline{\theta} \in \mathbb{C}^p$ by a linear $K \times p$ transformation $\mathbf{H}_t$

$$\underline{z}_t \approx \mathbf{H}_t \underline{\theta}.$$

The approximation is due to the fact that the measurements are noisy, or due to slight modeling errors. N time instances can be augmented to a matrix form $\underline{z}_{1:N} \approx \mathbf{H}_{1:N} \underline{\theta}$ where

$$\underline{z}_{1:N} \triangleq \begin{bmatrix} \underline{z}_1 \\ \vdots \\ \underline{z}_N \end{bmatrix}; \quad \mathbf{H}_{1:N} \triangleq \begin{bmatrix} \mathbf{H}_1 \\ \vdots \\ \mathbf{H}_N \end{bmatrix}.$$

The *weighted LS* (WLS) solution for $\underline{\theta}$, using non-negative weight matrix $\mathbf{W}_{1:N}$ (of size $KN \times KN$) is

$$\hat{\underline{\theta}} = (\mathbf{H}_{1:N}^{\dagger} \mathbf{W}_{1:N} \mathbf{H}_{1:N})^{-1} \mathbf{H}_{1:N}^{\dagger} \mathbf{W}_{1:N} \underline{z}_{1:N}. \quad \text{(D.1)}$$

Our goal is to evaluate (D.1) recursively. If the parameters slowly change, a common approach is to apply a diagonal weight matrix $\mathbf{W}_{1:N}$ with powers of a forgetting factor $0 < \alpha \leqslant 1$ along its diagonal. Note, that for measurements associated with the same time instance, we wish to apply the same factor, since equations of the same time instance have equal importance. Such weight matrix can be represented recursively as

$$\mathbf{W}_{1:N} = \begin{bmatrix} \alpha \mathbf{W}_{1:N-1} & \mathbf{0} \\ \mathbf{0}^{\dagger} & \mathbf{I} \end{bmatrix}; \quad \mathbf{W}_{1:1} = \mathbf{I},$$

where $\mathbf{I}$ and $\mathbf{0}$ stand for the identity and zero matrices of sizes $K \times K$ and $(N-1)K \times K$, respectively. Though it might seem that in order to derive a recursive solution for (D.1) a $K \times K$ matrix inversion should be made in each RLS iteration, in practice the complexity can be further reduced. This is obtained by applying the well-known RLS algorithm with a minor twist. Consider a *single* equation which is updated into the recursion. We must check if this new equation belongs to the next time instance. If so, a memory factor $\alpha \leqslant 1$ is applied. If this is not the case and we are evaluating one of the K equations of the current time instance, a memory factor of 1 is

used. Thus, in order to derive a recursion, where the update stage considers a *single* equation, the forgetting factor should vary. Notating the time instance by n and the sequential number of the equation by $nK + k$ (where $k \in \{1, \ldots, K\}$) the forgetting factor becomes

$$\text{forgetting factor} = \begin{cases} \alpha; & k = 1, \\ 1 & \text{otherwise.} \end{cases}$$

References

[1] J.B. Allen, D.A. Berkley, Image method for efficiently simulating small-room acoustics, J. Acoust. Soc. Am. 65 (4) (1979) 943–950.

[2] J. Benesty, Adaptive eigenvalue decomposition algorithm for passive acoustic source localization, Acoust. Soc. Am. 107 (1) (2000) 384–391.

[3] S.T. Birchfield, D.K. Gillmor, Fast Bayesian acoustic localization, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, Orlando, Florida, May 2002, pp. 1793–1796.

[4] M.S. Brandstein, J.E. Adcock, H.F. Silverman, A closed-form location estimator for use with room environment microphone arrays, IEEE Trans. Speech Audio Process. 5 (1) (1997) 45–50.

[5] M. Brandstein, H. Silverman, A robust method for speech signal time-delay estimation in reverberant rooms, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1997, pp. 375–378.

[6] M. Cetin, D.M. Malioutov, A.S. Willsky, A variational technique for source localization based on a sparse signal reconstruction perspective, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 3, Orlando, Florida, May 2002, pp. 2965–2968.

[7] B. Champagne, S. Bédard, A. Stéphenne, Performance of time-delay estimation in the presence of room reverberation, IEEE Trans. Acoust. Speech Signal Process. 4 (2) (1996) 148–152.

[8] C. Chen, R. Hudson, Maximum-likelihood acoustic source localization: experimental results, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 3, Orlando, Florida, May 2002, pp. 2949–2952.

[9] C. Chen, R. Hudson, K. Yao, Maximum-likelihood source localization and unknown sensor location estimation for wideband signals in the near-field, IEEE Trans. Signal Process. 50 (8) (2002) 1843–1854.

[10] H.H. Chiang, C.L. Nikias, A new method for adaptive time-delay estimation for non-Gaussian signals, IEEE Trans. Acoust. Speech Signal Process. 38 (2) (1990) 209–219.

[11] J. DiBiase, H. Silverman, M. Brandstein, Robust Localization in Reverberant Rooms, Springer, Berlin, 2001.

[12] E.D. DiClaudio, R. Parisi, G. Orlandi, Multi-source localization in reverberant environments by root-music and clustering, in: ICASSP-2000, 2000, pp. 921–924.

[13] S. Doclo, Multi-microphone noise reduction and dereverberation techniques for speech applications, Ph.D. Thesis, Katholieke Universiteit Leuven, May 2003.

[14] S. Doclo, Modification of robust time-delay estimation in highly adverse acoustic environments for tracking scenarios, Private communication, August 2003.

[15] S. Doclo, M. Moonen, Robust adaptive time delay estimation for speaker localisation in noisy and reverberant acoustic environments, EURASIP J. Appl. Signal Process. 2003 (11) (2003) 1110–1124.

[16] T.G. Dvorkind, Speaker localization in a reverberant and noisy environment, Master's Thesis, Technion—Israel Institute of Technology, December 2003.

[17] T. Dvorkind, S. Gannot, Speaker localization in a reverberant environment, IEEE Proceedings, The 22nd Convention of Electrical and Electronics Engineers in Israel, 2002, pp. 7–9.

[18] T. Dvorkind, S. Gannot, Speaker localization exploiting spatial-temporal information, in: International Workshop on Acoustic Echo and Noise Control, Kyoto, Japan, 2003, pp. 295–298.

[19] T. Dvorkind, S. Gannot, Approaches for time difference of arrival estimation in a noisy and reverberant environment, in: International Workshop on Acoustic Echo and Noise Control, Kyoto, Japan, 2003, pp. 215–218.

[20] D.R. Fischell, C.H. Coker, A speech direction finder, in: Proceedings of the 1984 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-84), 1984, pp. 19.8.1–19.8.4.

[21] J. Fuchs, On the application of the global matched filter to DOA estimation with uniform circular arrays, IEEE Trans. Signal Process. 49 (4) (2001) 702–709.

[22] S. Gannot, D. Burshtein, E. Weinstein, Signal enhancement using beamforming and non-stationarity with application to speech, IEEE Trans. Signal Process. 49 (8) (2001) 1614–1626.

[23] S. Gannot, A. Yeredor, Noise cancellation with static mixtures of a nonstationary signal and stationary noise, EURASIP J. Appl. Signal Process. 2002 (12) (2002) 1460–1472.

[24] S. Haykin, Adaptive Filter Theory, third ed., Prentice-Hall, Englewood Cliffs, NJ, 1996.

[25] Y. Huang, J. Benesty, G.W. Elko, R.M. Mersereau, Real-time passive source localization: a practical linear-correction least-squares approach, IEEE Trans. Speech Audio Process. 9 (8) (2001) 943–956.

[26] T. Kailath, A. Sayed, B. Hassibi, Linear Estimation, Prentice-Hall, Englewood Cliffs, NJ, 2000.

[27] C. Knapp, G. Carter, The generalized correlation method for estimation of time delay, IEEE Trans. Acoust. Speech Signal Process. 24 (4) (1976) 320–327.

[28] H. Kurtuff, Room Acoustics, third ed., Elsevier, London, UK, 1991.

[29] National Institute of Standards and Technology, The DARPA TIMIT acoustic-phonetic continuous speech corpus, CD-ROM NIST Speech Disc 1-1.1, October 1991.
[30] C. Nikias, R. Pan, Time delay estimation in unknown Gaussian spatially correlated noise, IEEE Trans. Acoust. Speech Signal Process. 36 (11) (1988) 1706–1714.
[31] T. Nishiura, S. Nakamura, K. Shikano, Talker localization in a real acoustic environment based on DOA estimation and statistical sound source identification, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, Orlando, Florida, May 2002, pp. 893–896.
[32] L. Parra, C. Spence, Convolutive blind separation of non-stationary sources, IEEE Trans. Speech Audio Process. 8 (3) (2000) 320–327.
[33] P. Peterson, Simulating the response of multiple microphones to a single acoustic source in a reverberant room, J. Acoust. Soc. Am. 76 (5) (1986) 1527–1529.
[34] K. Rahbar, J.P. Reilly, Blind source separation for MIMO convolutive mixtures, in: Third International Conference on ICA and BSS, San Diego, California, USA, 2001, pp. 242–247.
[35] R. Schmidt, Multiple emitter location and signal parameter estimation, IEEE Trans. Antennas and Propagation AP-34 (3) (1986) 276–280.
[36] D. Schobben, P. Sommen, A frequency domain blind signal separation method based on decorrelation, IEEE Trans. Signal Process. 50 (8) (2002) 1855–1865.
[37] O. Shalvi, E. Weinstein, System identification using nonstationary signals, IEEE Trans. Signal Process. 44 (8) (1996) 2055–2063.
[38] A. Stéphene, B. Champagne, A new cepstral prefiltering technique for estimating time delay under reverberant conditions, Signal Process. 59 (1997) 253–266.
[39] A. Varga, H.J.M. Steeneken, Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems, Speech Commun. 12 (1993) 247–251.
[40] H. Wang, M. Kaveh, Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources, IEEE Trans. Acoust. Speech Signal Process. 33 (4) (1985) 823–831.
[41] E. Weinstein, M. Feder, A.V. Oppenheim, Multi-channel signal separation by decorrelation, IEEE Trans. Speech Audio Process. 1 (4) (1993) 405–413.
[42] P.D. Welch, The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms, IEEE Trans. Audio Electroacoust. AU-15 (2) (1967) 70–73.