CCIT Report #457
December 2003

Time Difference of Arrival Estimation of Speech Source in a Noisy and Reverberant Environment

Tsvi G. Dvorkind [a,*] and 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*

Abstract

Determining the spatial position of a speaker finds a growing interest in video conference scenario where automated camera steering and tracking are required. Speaker localization can be achieved with a dual step approach. In the preliminary stage 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, 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.

Key words: Source localization, non-stationarity, decorrelation
PACS:

1 Introduction

Determining the spatial position of a speaker finds a growing interest in video conference scenario where automated camera steering and tracking are required. Microphone array, which is usually used for speech enhancement in a noisy environment [1], can be used for the task of speaker localization as well. The related algorithms can be divided into two groups: single and dual

* Corresponding author
Email addresses: dvorkind@tx.technion.ac.il (Tsvi G. Dvorkind), gannot@eng.biu.ac.il (Sharon Gannot).
URL: http://www-sipl.technion.ac.il/ gannot (Sharon Gannot).

Preprint submitted to Signal processing *22 October 2003*

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 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 is a member of this group. In the second group of single step approaches we find the *maximum likelihood* (ML) algorithms, which estimate the source locus by applying maximum likelihood criterion. Usually, the ML formulation leads to algorithms involving maximization of the output power of a beamformer steered to potential source locations (i.e. [2], [3], [4], [5]). In the dual-step approaches group, the first algorithmic stage involves TDOA estimation from spatially separated microphone pairs. The *generalized cross correlation* (GCC) method presented by Knapp and Carter [6] 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 [8], [9]. Unfortunately, the GCC method suffers from another model inaccuracy. It is assumed that the noise field is uncorrelated, assumption that 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 is a practical problem in video conference applications. It usually occurs when a point source, e.g. computer fan, projector or a ceiling fan, exists. The authors in [10] suggested discriminating speaker from directional noise with a Gaussian mixture model. A different approach was presented in [11] and [12], 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) ATF-s for TDOA extraction [13]. Extension of that work, for spatially correlated noise was presented by Doclo and Moonen [14] [15]. Assuming that the noise correlation matrix is known (using a *voice activity detector* (VAD)), the authors present a time domain algorithm for TDOA estimation using *generalized eigenvalue decomposition* (GEVD) approach and a pre-whitening approach.

In this work, we tackle the TDOA estimation problem. Our model assumptions consider reverberations and spatially correlated noise scenarios [16],[17].

In [1] 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. Similarly to [1] and the preceding work by Shalvi and Weinstein [18] 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 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 [19], we have additionally exploited the fact that there is no correlation between the speaker and the directional noise. The authors in [19] showed that in an application of signals separation, imposing a decorrelation criterion on the estimated signals results ATF-s ratio estimation. The authors further suggested exploiting speech non-stationarity, resulting a set of decorrelation equations. However, the obtained equation set is nonlinear, and due to this nonlinearity an inherent *frequency permutation ambiguity* results. The authors in [19] 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 FIR model for the mixing channels and conducting the solution 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 [20] where joint diagonalization of correlation matrices is carried out in the time domain. Considering a convolutive mixture (due to room reverberation), researches 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 *finite impulse response* (FIR) constraint on the separating ATF-s [21],[22] or (equivalently) imposing smoothness in the frequency domain [23]. 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 simulative study shows that the decorrelation constraint presents improved TDOA estimation for the batch methods at low *signal to noise ratio* (SNR) conditions.

Special attention 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 approximated recursive solution of a nonlinear equation set. The method, notated by *Recursive Gauus* (RG), is applied to the nonlinear decorrelation equations, resulting a solution applicable for tracking scenarios. Opposed to the GCC based methods, our solutions deal with reverberant environment and directional additive noise scenario. Opposed to the subspace methods [13][14], the suggested algorithms are conducted in the frequency domain, resulting computationally more efficient implementations which do not rely on a VAD for prior knowledge of the noise characteristics. Furthermore, simulative 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. Finally, extensive experimental study is presented in Section 4.

2 Problem Formulation and Motivation

In this section the problem is formulated and the basic assumptions are presented. By analytical expression and by simulative 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 m-th microphone, $z_m(t)$, is:

$$z_m(t) = a_m(t) * s(t) + n_m(t) \; ; \; m = 1, \ldots, M \tag{1}$$

where $*$ stands for convolution, $s(t)$ is the source signal and $n_m(t)$ is the interference signal at the m-th 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 a time-varying ATF from the desired speech source to the m-th microphone. When $n_m(t)$ is a directional interference, we can state:

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

where $b_m(t)$ is the ATF between the noise $n(t)$ and the m-th 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 $s(t)$).

2.2 Usage of ATF-s Ratio for TDOA Extraction

Let $A_m(\omega)$ be the frequency response of the m-th ATF $a_m(t)$. Similarly, let $B_m(\omega)$ be the frequency response of $b_m(t)$. Define

$$\mathcal{H}_m(\omega) \triangleq \frac{A_m(\omega)}{A_1(\omega)}$$

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:

$$\begin{aligned} A_m(\omega) &= \alpha_{n_0} e^{-j\omega n_0} + \sum_{i=1}^{S_1} \alpha_{n_i} e^{-j\omega n_i};\ m = 2 \dots M \\ A_1(\omega) &= \beta_{p_0} e^{-j\omega p_0} + \sum_{i=1}^{S_2} \beta_{p_i} e^{-j\omega p_i} \end{aligned}$$

with $\alpha_{n_0}, \beta_{n_0}$ and n_0, p_0 being the amplitudes and the delays of the main peaks of $a_m(t)$ and $a_1(t)$ respectively. Then the ratio can be stated as:

$$\mathcal{H}_m(\omega) = \frac{\alpha_{n_0} e^{-j\omega n_0}}{\beta_{p_0} e^{-j\omega p_0}} e(\omega);\ e(\omega) = \frac{1+\sum_{i=1}^{S_1} \frac{\alpha_{n_i} e^{-j\omega n_i}}{\alpha_{n_0} e^{-j\omega n_0}}}{1+\sum_{i=1}^{S_2} \frac{\beta_{p_i} e^{-j\omega p_i}}{\beta_{p_0} e^{-j\omega p_0}}}.$$

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 $e(\omega)$ tends to be close to 1, and the peak of the corresponding $h_m(t)$ can be used to determine the TDOA [1]. 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 [24][25], the ATF-s relating each possible source position to each microphone were simulated. Six reverberation values (denoted by T_r) were considered. Ranging from low reverberation conditions ($T_r = 0.1$[sec]) to intense conditions ($T_r = 0.6$[sec]). Two approaches were examined. TDOA estimation using ATF-s ratio, and TDOA estimation using the GCC method [6]. Assume

[1] 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.

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:

$$\frac{\Phi_{z_m z_1}(\omega)}{\Phi_{z_1 z_1}(\omega)} = \frac{A_m(\omega) {A_1}^*(\omega) \Phi_{ss}(\omega)}{A_1(\omega) {A_1}^*(\omega) \Phi_{ss}(\omega)} = \mathcal{H}_m(\omega)$$

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 [26] is applied for the PSD estimation, using window $w(t)$ of length P. Denote $\hat{\Phi}_{z_i z_j}(\omega)$ the PSD estimate of z_i with z_j. Only for $P \to \infty$ the statement $\frac{\hat{\Phi}_{z_m z_1}(\omega)}{\hat{\Phi}_{z_1 z_1}(\omega)} \to \mathcal{H}_m(\omega)$ holds. However, for a finite length analysis window $w(t)$ this does not hold as the PSD estimates are smoothed by a circular convolution over the $[0, 2\pi)$ interval and exact elimination of common terms in the nominator and denominator of the stated ratio is not possible. However, for implementing a tracking system, where fast changes in $\mathcal{H}_m(\omega)$ 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[2]. 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]. However, as will be seen shortly, reasonable performance (with respect to TDOA estimation) can 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 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 subsample precision, the calculated impulse response, is Shannon interpolated with 0.1[sample] resolution. Finally, the TDOA is evaluated by extracting the position of the maximal peak of the impulse response. Divergence of more than 2 samples from the true TDOA (which is known from the geometry of the problem) is considered to be anomaly. Non-anomalous estimations are considered for calculating the *root mean square error* (RMSE). Figure 1 presents the result for $P = 4096$[samples] and Figure 2 for $P = 256$[samples]. As can be seen from Figure 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 render useless at $T_r = 0.3$[sec] due to divergence from the ideal, reverberant-free, model. This

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

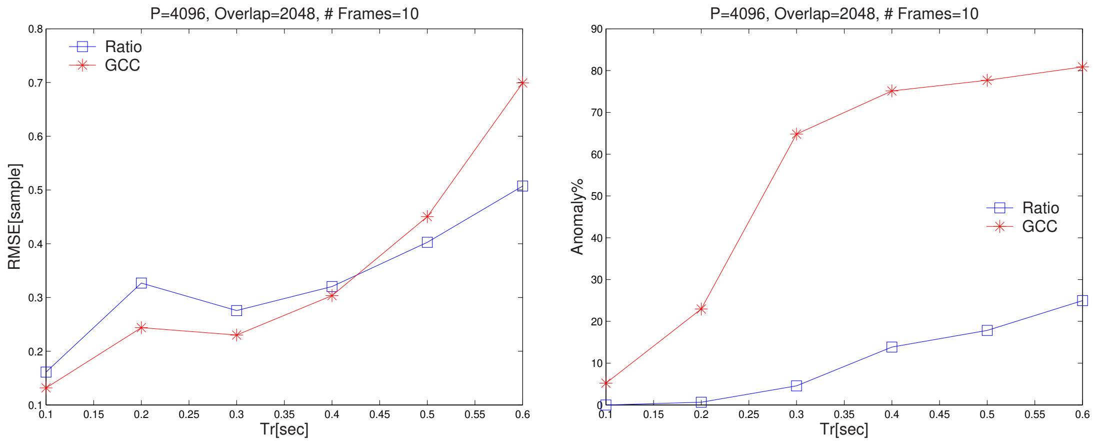

Fig. 1. Simulative model test. Long observation frame.

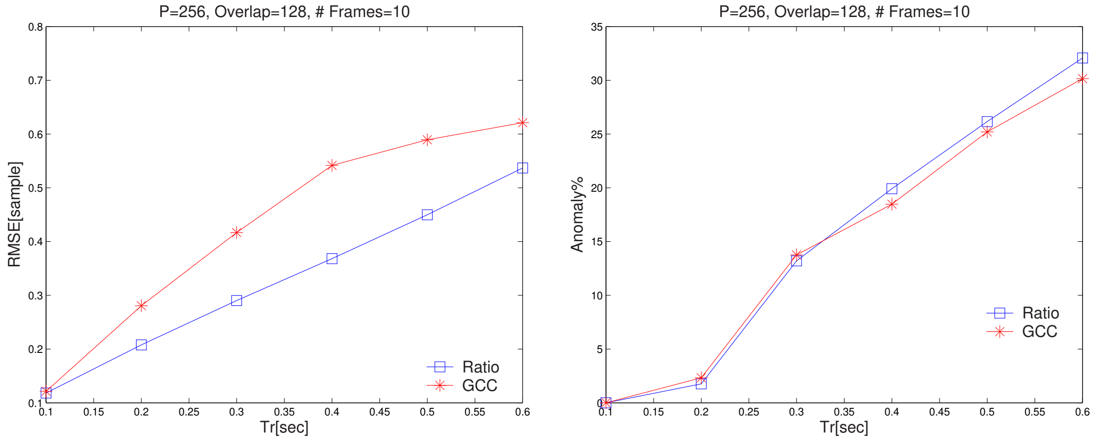

Fig. 2. Simulative model test. Short observation frame.

result is compatible with the one presented by Champagne et al. in [7]. On the other hand, the ATF-s ratio model maintains low anomaly percentage even at high reverberations. Figure 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 estimation (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 32% for $T_r = 0.6[sec]$, instead of 25% for the long frame case). However, aiming to mid range reverberation of $T_r = 0.2, 0.3$[sec] and exploiting additional, spatially separated, microphone pairs we hope to achieve reasonable performance despite the use of small P. It is worth mentioning that the GCC method still suffers from another modelling assumption; it assumes uncorrelated measurement noise. In the sequel, we will demonstrate that the GCC method render 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 $\mathcal{H}_m(\omega)$, exploiting the speech signal quasi-stationarity, was first presented in [1], based on a method derived in [18]. Noting that the speaker and the noise source are uncorrelated, we can state the following equation:

$$\Phi_{z_i z_j}(\omega) = A_i(\omega) A_j{}^*(\omega) \Phi_{ss}(\omega) + B_i(\omega) B_j{}^*(\omega) \Phi_{nn}(\omega) \tag{3}$$

with $\Phi_{z_i z_j}(\omega)$ being the PSD of z_i and z_j, $\Phi_{ss}(\omega)$ is the speech auto-PSD and $\Phi_{nn}(\omega)$ is the noise auto-PSD. Examining (3), we note that

$$\Phi_{z_m z_1}(\omega) - \mathcal{H}_m(\omega) \Phi_{z_1 z_1}(\omega) = \Phi_{b_1}(\omega) \tag{4}$$

where

$$\Phi_{b_1}(\omega) = (\mathcal{G}_m(\omega) - \mathcal{H}_m(\omega)) \left| B_1(\omega) \right|^2 \Phi_{nn}(\omega) \tag{5}$$

is a noise-only term. In practice however, stationarity of the speech signal can be assured only over small time intervals. Consider an analysis interval for which the noise signal can be regarded stationary and the ATF-s time invariant, while the speech signal statistics is changing (quasi-stationarity assumption for speech signals). Dividing the observation interval into N consecutive frames, an overdetermined set of equations for $\mathcal{H}_m(\omega)$ is obtained. This set can be solved by virtue of the *least squares* (LS) method. The resultant frequency domain algorithm is now presented.

Exploiting the quasi-stationarity property of the speech and defining:

$$\hat{\Phi}_{b_1}(n, \omega) \triangleq \hat{\Phi}_{z_m z_1}(n, \omega) - \mathcal{H}_m(\omega) \hat{\Phi}_{z_1 z_1}(n, \omega); \qquad n = 1, \ldots, N$$

where, $\hat{\Phi}_{z_i z_j}(n, \omega)$ is an estimate of the PSD of z_i and z_j at the n-th frame, Equation (4) become a set of equations for $\mathcal{H}_m(\omega)$. This overdetermined set for $\mathcal{H}_m(\omega)$ can also be stated as:

$$\hat{\Phi}_{z_m z_1}(n, \omega) = \mathcal{H}_m(\omega) \hat{\Phi}_{z_1 z_1}(n, \omega) + \Phi_{b_1}(\omega) + \xi(n, \omega); \qquad n = 1, \ldots, N \tag{6}$$

where, $\xi(n, \omega) \triangleq \hat{\Phi}_{b_1}(n, \omega) - \Phi_{b_1}(\omega)$ is an error term, which is minimized in the LS sense, using the overdetermined set (6). The noise-only term $\Phi_{b_1}(\omega)$

which is regarded stationary, and the ATF ratio $\mathcal{H}_m(\omega)$, 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 depicted in (4)) as the *first form of stationarity* (S1). The *weighted LS* (WLS) solution to (6) is:

$$\begin{bmatrix} \hat{\mathcal{H}}_m(\omega) \\ \hat{\Phi}_{b_1}(\omega) \end{bmatrix} = \left(\mathbf{A}^\dagger \mathbf{W} \mathbf{A}\right)^{-1} \mathbf{A}^\dagger \mathbf{W} \underline{\hat{\Phi}}_{z_m z_1}(\omega) \tag{7}$$

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, a *second form of stationarity* (S2) can be stated. Examine

$$\hat{\Phi}_{z_m z_m}(n,\omega) = \mathcal{H}_m(\omega)\hat{\Phi}_{z_1 z_m}(n,\omega) + \Phi_{b_m}(\omega) + \xi_2(n,\omega); \qquad n = 1,\ldots,N \tag{8}$$

where $\Phi_{b_m}(\omega)$ is also a stationary noise-only term:

$$\Phi_{b_m}(\omega) = (\mathcal{G}_m(\omega) - \mathcal{H}_m(\omega))B_1(\omega){B_m}^*(\omega)\Phi_{nn}(\omega). \tag{9}$$

This second form of stationarity has LS solution similar to (7):

$$\begin{bmatrix} \hat{\mathcal{H}}_m(\omega) \\ \hat{\Phi}_{b_m}(\omega) \end{bmatrix} = \left(\mathbf{B}^\dagger \mathbf{W} \mathbf{B}\right)^{-1} \mathbf{B}^\dagger \mathbf{W} \underline{\hat{\Phi}}_{z_m z_m}(\omega) \tag{10}$$

with

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

3.2 Decorrelation Criterion

Until this point we estimated $\mathcal{H}_m(\omega)$ based on noise stationarity and the speech quasi-stationarity characteristics. Though this led us to an attractive closed-form solutions, it is interesting to evaluate the influence of an additional constraint. Namely, the fact that the speaker and the interference noise are 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 m-th microphone can be written as:

$$\mathbf{P}(\omega) \triangleq \begin{bmatrix} \Phi_{z_1 z_1}(\omega) & \Phi_{z_1 z_m}(\omega) \\ \Phi_{z_m z_1}(\omega) & \Phi_{z_m z_m}(\omega) \end{bmatrix} \tag{11}$$

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 $\begin{bmatrix} Z_1(\omega) & Z_m(\omega) \end{bmatrix}^T$ such that the output PSD matrix $\mathbf{R}(\omega) \triangleq \mathbf{U}(\omega)\mathbf{P}(\omega)\mathbf{U}^\dagger(\omega)$ is diagonal yields decorrelated outputs. We show now that a by-product of the diagonalization process will lead us to an estimate of $\mathcal{H}_m(\omega)$. In particular, setting:

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

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

$$u_2^*(\omega)\left(\Phi_{z_m z_1}(\omega) - u_1(\omega)\Phi_{z_1 z_1}(\omega)\right) = \Phi_{z_m z_m}(\omega) - u_1(\omega)\Phi_{z_1 z_m}(\omega). \tag{12}$$

Note that (12) is a single (nonlinear) equation in two unknowns. Equation (12) was derived in [19], 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 [19] 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, equation (12) becomes a set of equations, obtained by evaluating the PSD-s at different frame indices:

$$u_2^*(\omega)\left(\underline{\hat{\Phi}}_{z_m z_1}(\omega) - u_1(\omega)\underline{\hat{\Phi}}_{z_1 z_1}(\omega)\right) \approx \underline{\hat{\Phi}}_{z_m z_m}(\omega) - u_1(\omega)\underline{\hat{\Phi}}_{z_1 z_m}(\omega) \tag{13}$$

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 $N \geq 2$ we have enough equations to solve the problem, though the expressions are still non-linear in $u_1(\omega)$ and $u_2(\omega)$. Simple assignment shows that the pair $\{u_2(\omega) = \mathcal{G}_m(\omega), u_1(\omega) = \mathcal{H}_m(\omega)\}$ as well as the pair $\{u_1(\omega) = \mathcal{G}_m(\omega), u_2(\omega) = \mathcal{H}_m(\omega)\}$ solves the equations at hand. This is referred to as the *frequency permutation ambiguity* problem[3]. The authors in [19] did not present a solution to (13). In particular, they avoided the permutation problem inherent in (13), by solving the problem (iteratively) in the time domain. In this contribution we solve (13) directly to obtain an estimate for $\mathcal{H}_m(\omega)$. 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 (13) is nonlinear in $u_2(\omega)$ and $u_1(\omega)$, the Gauss method is employed. 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 solution, which we will address in the sequel, allows 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 $u_1(\omega) = \mathcal{H}_m(\omega)$ from the set (13). Special attention will be given to avoid the permutation problem, i.e. the solution $u_1(\omega) = \mathcal{G}_m(\omega)$.

Experimental results revealed that the first (and second) form of stationarity perform well at reasonable *signal to noise ratio* (SNR), but at negative SNR values their estimate of $\mathcal{H}_m(\omega)$ deteriorates. On the other hand, it is speculated that for negative SNR values, the estimated noise bias terms ($\hat{\Phi}_{b_1}(\omega)$ in (7) and $\hat{\Phi}_{b_m}(\omega)$ in (10)) can be reliably obtained[4]. Using (5) and (9) it is evident that

$$\frac{\Phi_{b_m}(\omega)}{\Phi_{b_1}(\omega)} = \mathcal{G}_m^*(\omega) \tag{14}$$

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

$$u_2^*(\omega) = \frac{\hat{\Phi}_{b_m}(\omega)}{\hat{\Phi}_{b_1}(\omega)} \tag{15}$$

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

[4] 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.

This assignment has a twofold advantage. First, using this initialization, the set (13) becomes a **linear** set in $u_1(\omega)$. Thus, LS solution can be obtained:

$$\hat{\mathcal{H}}_m(\omega) = \left(\underline{A}^{\dagger}\underline{A}\right)^{-1} \underline{A}^{\dagger}[\underline{\hat{\Phi}}_{z_m z_m}(\omega) - u_2^*(\omega)\underline{\hat{\Phi}}_{z_m z_1}(\omega)] \tag{16}$$

where

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

and $u_2^*(\omega)$ is set according to (15). Second, by setting $u_2^*(\omega) = \mathcal{G}_m^*(\omega)$, $u_1(\omega)$ must tend to become $\mathcal{H}_m(\omega)$, thus overcoming the frequency permutation problem. The resultant algorithm is notated by *Linear Decorrelation* (LD) and is summarized in Figure (3).

(1) Estimate $\Phi_{b_1}(\omega)$ using (7) and $\Phi_{b_m}(\omega)$ using (10).
(2) Estimate $u_2^*(\omega) = \mathcal{G}_m^*(\omega)$ using (15).
(3) Estimate $\mathcal{H}_m(\omega)$ using (16).

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

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

3.3.2 Gauss and First Form of Stationarity

We now present an iterative solution to (13) 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 the sets (13) and (6) **simultaneously** as one large LS problem. Concatenating these equations we get:

$$\begin{bmatrix} \underline{\hat{\Phi}}_{z_1 z_m}(\omega) & \underline{\hat{\Phi}}_{z_m z_1}(\omega) & -\underline{\hat{\Phi}}_{z_1 z_1}(\omega) & \underline{0} \\ \underline{\hat{\Phi}}_{z_1 z_1}(\omega) & \underline{0} & \underline{0} & \underline{1} \end{bmatrix} \begin{bmatrix} \mathcal{H}_m(\omega) \\ \mathcal{G}_m^*(\omega) \\ \mathcal{H}_m(\omega)\mathcal{G}_m^*(\omega) \\ \Phi_{b_1}(\omega) \end{bmatrix} \approx \begin{bmatrix} \underline{\hat{\Phi}}_{z_m z_m}(\omega) \\ \underline{\hat{\Phi}}_{z_m z_1}(\omega) \end{bmatrix} \tag{17}$$

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

$$\underline{\theta} \triangleq [\mathcal{H}_m(\omega), \mathcal{G}_m^*(\omega), \Phi_{b_1}(\omega)]^T.$$

Denote also:

$$\begin{aligned}\underline{h}(\underline{\theta}) \triangleq {} & \mathcal{H}_m(\omega)\begin{bmatrix}\underline{\hat{\Phi}}_{z_1 z_m}(\omega)\\ \underline{\hat{\Phi}}_{z_1 z_1}(\omega)\end{bmatrix} + \mathcal{G}_m^*(\omega)\begin{bmatrix}\underline{\hat{\Phi}}_{z_m z_1}(\omega)\\ \underline{0}\end{bmatrix} + \\ & + \mathcal{H}_m(\omega)\mathcal{G}_m^*(\omega)\begin{bmatrix}-\underline{\hat{\Phi}}_{z_1 z_1}(\omega)\\ \underline{0}\end{bmatrix} + \Phi_{b_1}(\omega)\begin{bmatrix}\underline{0}\\ \underline{1}\end{bmatrix}\end{aligned} \tag{18}$$

Then Gauss iterations take the form:

$$\underline{\theta}^{(l+1)} = \underline{\theta}^{(l)} + \left(\mathbf{H}(\underline{\theta}^{(l)})^{\dagger}\mathbf{H}(\underline{\theta}^{(l)})\right)^{-1}\mathbf{H}(\underline{\theta}^{(l)})^{\dagger}\left(\underline{d} - \underline{h}(\underline{\theta}^{(l)})\right) \tag{19}$$

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

$$\begin{aligned}&\mathbf{H}(\theta^{(l)}) \triangleq \nabla_{\theta}\underline{h}(\underline{\theta})\big|_{\underline{\theta}=\underline{\theta}^{(l)}} = \\ &= \begin{bmatrix}\underline{\Phi}_{z_1 z_m}(\omega) - \mathcal{G}_m^{(l)^*}(\omega)\underline{\hat{\Phi}}_{z_1 z_1}(\omega), & \underline{\hat{\Phi}}_{z_m z_1}(\omega) - \mathcal{H}_m^{(l)}(\omega)\underline{\hat{\Phi}}_{z_1 z_1}(\omega), & \underline{0}\\ \underline{\hat{\Phi}}_{z_1 z_1}(\omega), & \underline{0}, & \underline{1}\end{bmatrix}\end{aligned} \tag{20}$$

and

$$\underline{d} \triangleq \begin{bmatrix}\underline{\hat{\Phi}}_{z_m z_m}(\omega)\\ \underline{\hat{\Phi}}_{z_m z_1}(\omega)\end{bmatrix}. \tag{21}$$

Two stopping criterions can be considered. First, the residual norm $\left\|\underline{\theta}^{(l+1)} - \underline{\theta}^{(l)}\right\|$ can be limited to a predefined threshold. Second, the number of the iterations can be limited apriori.

The resultant algorithm is denoted by *Gauss and First Form of Stationarity* (GS1) and is summarized in Figure 4.

(1) Denote $\underline{\theta} = [\mathcal{H}_m(\omega), \mathcal{G}_m^*(\omega), \Phi_{b_1}(\omega)]^T$.
(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 $\Phi_{b_1}^{(0)}(\omega)$ from the LS solution of (7).
(3) Calculate $\underline{h}(\underline{\theta})$ using (18).
(4) Calculate $\mathbf{H}(\underline{\theta})$ using (20).
(5) Set $\underline{d}$ as in (21).
(6) Iterate (19) till a pre-defined convergence criterion is reached.

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

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 (A.3) to the S1 algorithm, using forgetting factor $\alpha < 1$, slow variations of $\mathcal{H}_m(\omega)$ are trackable. This recursive solution for S1 (Notated by RS1) is summarized in Figure 5.

(1) Use $\underline{\theta} = [\mathcal{H}_m(\omega), \Phi_{b_1}(\omega)]^T$.
(2) Apply (A.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)

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

(1) Use the current estimate of $\Phi_{b_1}(\omega)$ available from RS1 algorithm.
(2) Apply (A.3) with $\underline{\theta} = [\mathcal{H}_m(\omega), \Phi_{b_m}(\omega)]^T$, $\underline{a}_n^T = [\hat{\Phi}_{z_1 z_m}(n, \omega), 1]$ and $y_n = \hat{\Phi}_{z_m z_m}(n, \omega)$ for recursive estimation of $\Phi_{b_m}(\omega)$.
(3) Evaluate $u_2^*(\omega) = \mathcal{G}_m^*(\omega)$ using (15).
(4) Apply (A.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)

3.4.2 Recursion for GS1

Algorithms which employ the nonlinear decorrelation equation (12) can be solved recursively using the *Recursive Gauss* (RG) method, presented in Appendix B. For the GS1 algorithm, the parameter set is $\underline{\theta} = [\mathcal{H}_m(\omega), \mathcal{G}_m^*(\omega), \Phi_{b_1}(\omega)]^T$ and the update stage includes the evaluation of **two** equations. Consider the

n-th time instance for which we receive the measurements $\underline{h}_n(\underline{\theta}) \approx \underline{d}_n$ with:

$$\begin{aligned}\underline{h}_n(\underline{\theta}) \triangleq\ & \mathcal{H}_m(\omega)\begin{bmatrix}\hat{\Phi}_{z_1 z_m}(n,\omega)\\ \hat{\Phi}_{z_1 z_1}(n,\omega)\end{bmatrix} + \mathcal{G}^*_m(\omega)\begin{bmatrix}\hat{\Phi}_{z_m z_1}(n,\omega)\\ 0\end{bmatrix} - \\ & -\mathcal{H}_m(\omega)\mathcal{G}^*_m(\omega)\begin{bmatrix}\hat{\Phi}_{z_1 z_1}(n,\omega)\\ 0\end{bmatrix} + \Phi_{b_1}(\omega)\begin{bmatrix}0\\ 1\end{bmatrix}; \qquad (22)\end{aligned}$$

$$\underline{d}_n \triangleq \begin{bmatrix}\hat{\Phi}_{z_m z_m}(n,\omega)\\ \hat{\Phi}_{z_m z_1}(n,\omega)\end{bmatrix}.$$

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

$$\mathbf{H}_n(\underline{\theta}) = \begin{bmatrix}\hat{\Phi}_{z_1 z_m}(n,\omega) - \mathcal{G}^*_m(\omega)\hat{\Phi}_{z_1 z_1}(n,\omega), & \hat{\Phi}_{z_m z_1}(n,\omega) - \mathcal{H}_m(\omega)\hat{\Phi}_{z_1 z_1}(n,\omega), & 0\\ \hat{\Phi}_{z_1 z_1}(n,\omega), & 0, & 1\end{bmatrix}. \qquad (23)$$

Using the previous notations, the measurements for the LS problem, at the n-th time instance take the simple form:

$$\begin{aligned}\underline{y}_n =& \underline{d}_n - \underline{h}_n(\hat{\underline{\theta}}(n-1)) + \mathbf{H}_n(\hat{\underline{\theta}}(n-1))\hat{\underline{\theta}}(n-1) = \qquad (24)\\ =& \begin{bmatrix}\hat{\Phi}_{z_m z_m}(n,\omega) - \hat{\mathcal{H}}_m(n-1,\omega)\hat{\mathcal{G}}^*_m(n-1,\omega)\hat{\Phi}_{z_1 z_1}(n,\omega)\\ \hat{\Phi}_{z_m z_1}(n,\omega)\end{bmatrix}\end{aligned}$$

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 instance we have two equations, the form of RLS depicted in Appendix C should be used. Namely, for each time instance we perform two RLS iterations, one for each equation. The resultant recursive algorithm is denoted by RGS1 and summarized in Figure 7.

4 Experimental Study

In this chapter we assess the proposed algorithms, namely S1, LD, and GS1, and compare them with the classical GCC algorithm [6] and the recently proposed subspace method (GEVD algorithm) presented by Doclo and Moonen in [14]. The latter is notated by DM.

Notate the current time instance by n and the sequential number of the evaluated equation by $2n+m;\ m \in \{1,2\}$. Evaluate (A.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 (23).
(2) The current measurement y_{2n+m} is the m-th row of $\underline{y}_n$. $\underline{y}_n$ is evaluated according to (24).
(3) According to Appendix C, the forgetting factor α should be switched to 1, whenever $m \neq 1$.

Fig. 7. Recursive solution for GS1 (RGS1)

4.1 TDOA Estimation - Simulation Setup

We start by describing the simulation setup for TDOA estimation. Throughout this work, the sampling frequency is $F_s = 8000$[Hz]. Speech signals are drawn from the TIMIT database [27] and the noise source is the speech-like noise drawn from the NOISEX-92 [28] database. Throughout the simulations speech sentences and the directional interference are filtered by the respective 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 [24][25]. 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 $T_r = 0.25$[sec].
(2) LMS sub-sampling is set to 10 samples.
(3) LMS step-size of 10^{-8} is used.
(4) First 20000 samples are used for noise covariance matrix estimation.

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

For all evaluated methods sub-sample TDOA calculation is performed using

Shannon interpolation, on a $\frac{T_s}{10}$[sec] 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. Anomaly is defined as divergence of more than 2 samples from the actual TDOA (known in advance from the geometry of the problem).
(2) *Root Mean Square Error* (RMSE) in sample units. When the anomaly percentage figure of merit is used, 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 the most important figure of merit.

4.1.3 PSD estimation

Throughout the simulation we have conducted the PSD estimation using the Welch method [26]. 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. 10 (weighted) periodograms were used for each PSD estimate. For static scenarios, we allowed for 10 non-overlapping frames for each LS formulation. For statistical significants 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 $\alpha = 0.8222$.

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 have considered the selection of 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 $[2.53, 4.03, 2.67]$. Various reverberation times and SNR values are tested and the ATF-s are simulated using the image method [24][25]. For this experiment, the percentage of anomalies is calculated and only non-anomalous estimates are involved in calculating the RMSE. The anomaly and RMSE results are summarized in Tables 1 and 2 respectively. As can be seen from Table 1, at low reverberation conditions ($T_r = 0.1$[sec]) 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 severe reverberation of $T_r = 0.5$[sec], even in the high

T_r[sec]	SNR[dB]	S1	LD	GS1	GCC	DM
0.10	5.0	1	0	0	16	1
0.50	5.0	2	5	5	65	46
0.25	5.0	0	0	0	20	12
0.25	0.0	4	3	3	98	19
0.25	-5.0	52	37	24	100	31

Table 1
Static scenario with simulated ATF-s. Anomaly results.

T_r[sec]	SNR[dB]	S1	LD	GS1	GCC	DM
0.10	5.0	0.06	0.06	0.06	0.07	0.22
0.50	5.0	0.15	0.15	0.14	0.13	0.80
0.25	5.0	0.09	0.10	0.10	0.09	0.48
0.25	0.0	0.12	0.12	0.13	0.07	0.72
0.25	-5.0	0.12	0.12	0.15	-	0.89

Table 2
Static scenario with simulated ATF-s. RMSE results.

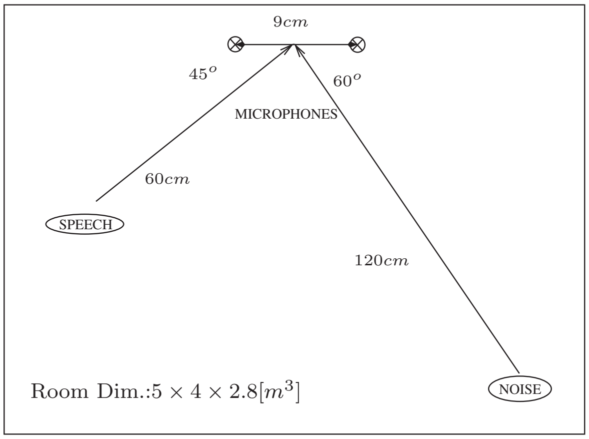

Fig. 8. Real room ATF-s. Geometric configuration.

SNR level, the performance of the subspace method DM and the GCC rapidly deteriorate. It seems that despite the use of short support analysis window P, the GCC still lacks the reverberant model. The subspace methods becomes inadequate probably due to the underestimated ATF length. Possibly, this can be solved on the expense of increased complexity. On the other hand, the simulation shows that the proposed frequency domain methods present low anomaly results. This is also the case at mid-range reverberation $T_r = 0.25$[sec] 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 render 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 fast 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 is extracted with high accuracy. The DM method presents slightly higher deviation from the true TDOA.

4.2.2 Real Room ATF-s

The actual room configuration is depicted in Figure 8. Using real room recordings, the ATF-s were calculated beforehand and then used in the simulations. From the geometry of the problem, and from the obtained ATF-s it was calculated that the speaker's TDOA is 1.5[sample] and the directional noise TDOA is -1.1[sample]. Table 3 presents the RMSE for the evaluated algorithms, at various SNR values. Within this experiment anomaly is not considered. Several phenomena are manifested by Table 3. Note that the proposed methods have lower RMSE than the GCC algorithm. This also holds for the DM algorithm, except for the very low SNR$= -5$[dB] (however, we note that the DM algorithm uses a priori knowledge of the noise covariance matrix, while in the proposed methods the noise bias term is directly estimated from the

SNR[dB]	S1	LD	GS1	GCC	DM
-5.0	1.44	1.11	0.86	2.70	0.77
-2.5	0.46	0.41	0.41	2.71	0.62
0.0	0.36	0.09	0.08	2.65	0.60
2.5	0.07	0.07	0.05	1.19	0.67
5.0	0.05	0.04	0.03	0.30	0.93

Table 3
Static scenario with real room ATF-s. RMSE results.

noisy observation). We also note that the RMSE value of 2.7[samples] presented by the GCC method at negative SNR is mainly caused by the bias of the method, due to its tendency to lock on the stronger (correlated) signal (bias level of $2.6-2.7$ samples is the value that diverts the GCC readings from the speaker's TDOA to the noise TDOA). We also note, that as before, slight improvement is obtained by the decorrelation methods (LD, GS1) compared with the stationarity method (S1) at low SNR conditions.

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 Subsection 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 $T = 30$[sec]. The speaker Cartesian position as a function of time $t \in [0, T]$ is,

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

with $f = 0.0529[Hz]$. This trajectory is depicted in Figure 9. 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[sec]. The mean SNR for the 30[sec] long signal is set to a relatively high value of 10[dB] to produce 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 one and use overlapping frames.

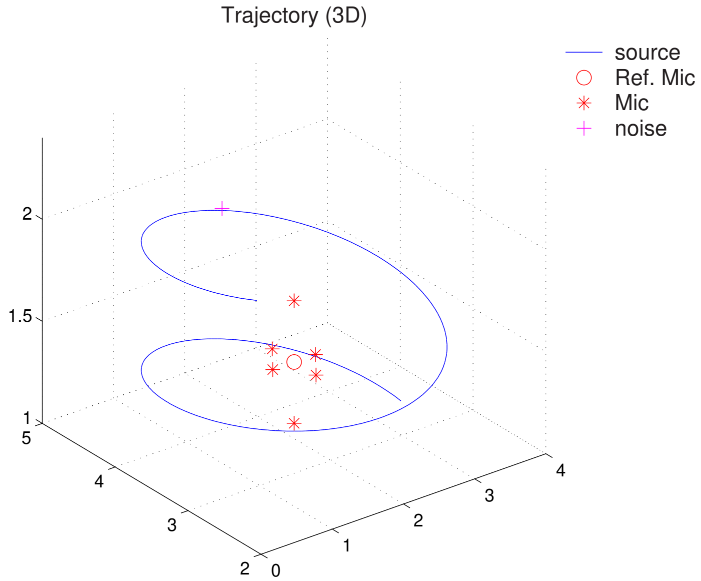

Fig. 9. Speaker trajectory

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[5] [13] and denoted here by EVD. For the latter, a step size of 10^{-7} is used. In order for the subspace methods (i.e. EVD and DM) to work in the tracking scenario, Doclo [29] proposed to slightly modify the algorithms by introducing intermediate initializations, reducing the LMS sub-sampling to 1 sample and using underestimated ATF-s of 20 samples. Figure 10 presents the TDOA estimation plots for the different methods. As can be seen from Fig. 10 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 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. Nevertheless, the obtained performance is significantly su-

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

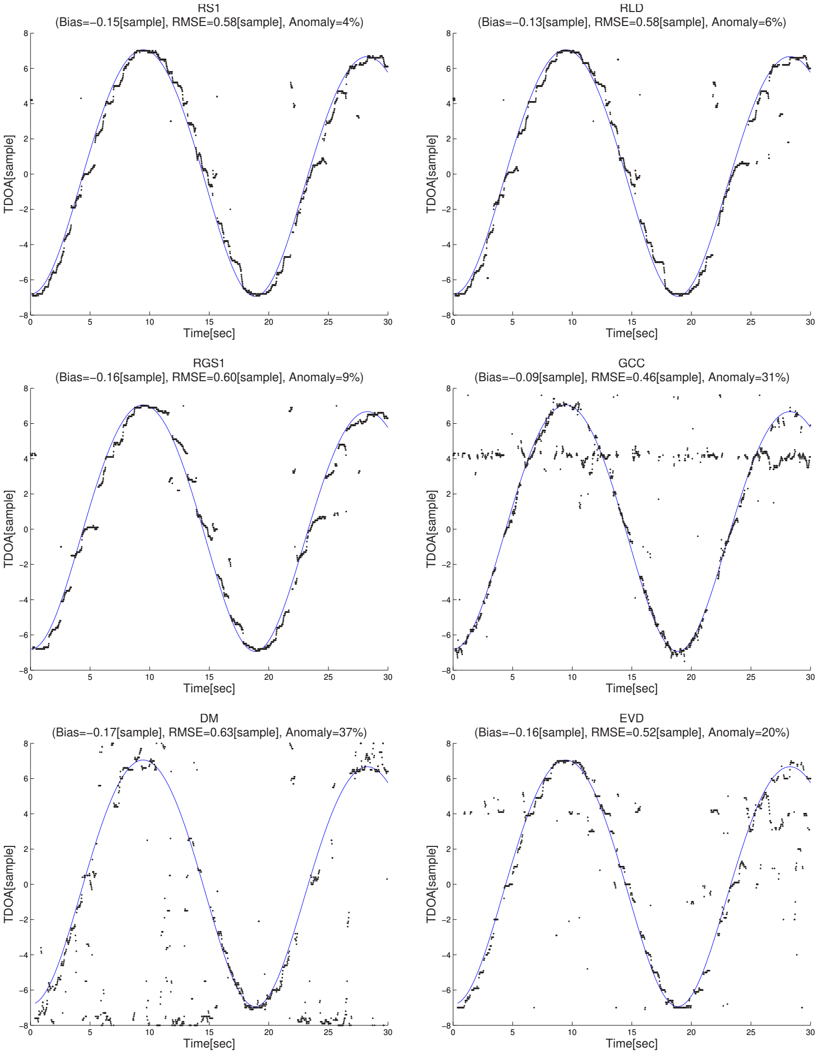

Fig. 10. TDOA estimation results. Solid line: True TDOA. Dots: Estimation results. The method's name, its bias, RMSE and anomaly results are presented in the title of each plot.

perior to the other methods. We further note that, as the the decorrelation methods (RLD, RGS1) did not yield an improvement in the tracking scenario compared with the simpler RS1 method, 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. For this scenario, using the same settings as in the previous experiment, simulation was conducted with one speaker located at the position $[2.75, 4.75, 2.436]$

and the other at $[1.47, 4.03, 2.674]$. A directional interference was placed at the position $[2, 4.207, 2.082]$. Figure 11 presents the TDOA estimation results by the RS1 algorithm (which gave the best results), for the previously mentioned microphone pair. For this experiment, anomaly was defined as divergence of more than 0.5 [sample] from the true TDOA. As can be seen from the plot, for the stated scenario and microphone pair, the algorithm demonstrates excellent tracking capabilities.

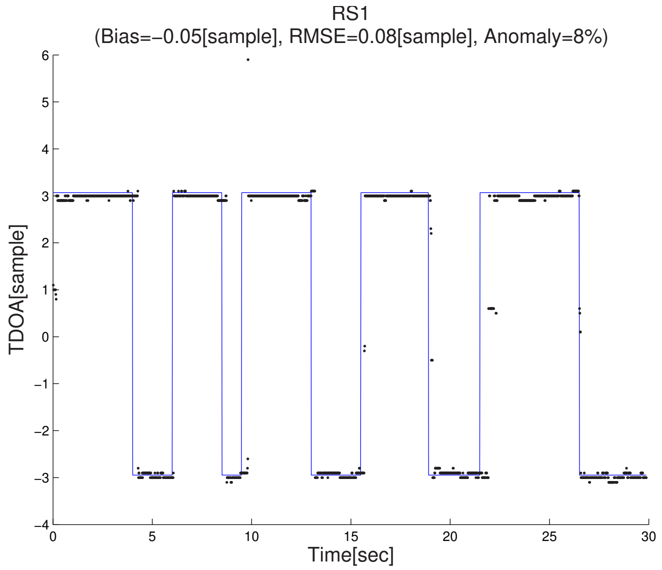

Fig. 11. TDOA estimation results. Solid line: True TDOA. Dots: Estimation results. The method's name, its bias, RMSE and anomaly results are presented in the title of the plot.

5 Summary

In this work novel TDOA estimation algorithms, based on the ATF-s ratio $\mathcal{H}_m(\omega)$ for TDOA extraction, were presented. Speech quasi-stationarity, noise stationarity and the the fact that there is no correlation between the speech and the noise were used for $\mathcal{H}_m(\omega)$ 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. Usage of short support analysis window was considered for improving GCC robustness to reverberation. Computational considerations, presented in Appendix D, revealed that the suggested frequency domain methods result in 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* (RG), for recursive solution of a nonlinear equation set.

6 Acknowledgement

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.

A 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:

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

with

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

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

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

with $\underline{a}_n;\; n = 1, \ldots, N$ a $p \times 1$ vector. Then, the recursive solution to (A.1) takes the known form (see for example [30],[31]):

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

$$\hat{\underline{\theta}}(n) = \hat{\underline{\theta}}(n-1) + \underline{K}_n\left(y_n - \underline{a}_n^{\dagger}\hat{\underline{\theta}}(n-1)\right) \tag{A.3}$$

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

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 $\beta \gg 1$.

B 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 *Recursive Gauss* (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:

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

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 (B.1) is:

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

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:

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

we will minimize:

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

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:

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

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 (A.2).

C 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 modelling errors. N time instances can be augmented to a matrix form $\underline{z}_{1:N} \approx \mathbf{H}_{1:N}\underline{\theta}$ where

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

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

$$\hat{\underline{\theta}} = \left(\mathbf{H}_{1:N}{}^{\dagger}\mathbf{W}_{1:N}\mathbf{H}_{1:N}\right)^{-1} \mathbf{H}_{1:N}{}^{\dagger}\mathbf{W}_{1:N}\underline{z}_{1:N} \tag{C.1}$$

Our goal is to evaluate (C.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 $0 < \alpha \leq 1$ 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:

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

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 (C.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 $\alpha \leq 1$ 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

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

D Computational Complexity

In this section we address the computational complexity of the proposed 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 $\underline{\theta} \in \mathcal{C}^p$ involves $10p^2 + 12p$ real multiplications and one complex division. Noting that RLS iteration is performed in each frequency bin and that there are $\frac{P}{2}$ frequencies to evaluate, the total number of real multiplications performed by the RLS is $\frac{P(10p^2+12p)}{2}$. The suggested frequency domain algorithms further involve one IFFT operation and interpolation. Assuming that the interpolation is conducted for S samples[6] with a $\frac{1}{10}$ 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

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

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

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

real multiplications. For RS1 algorithm $p = 2$. Assuming that $S = 17, P = 256, K = 128$ this yields approximately 193 multiplications per sample. Recall that this algorithmic stage is only the first one in the localization task. However, we note that the computational complexity of the second algorithmic stage (namely, localization from the TDOA estimates) is negligible comparing with the first algorithmic stage.

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

References

[1] S. Gannot, D. Burshtein, E. Weinstein, Signal Enhancement Using Beamforming and Non-Stationarity with application to Speech, IEEE Trans. on Sig. Proc. 49 (8) (2001) 1614–1626.

[2] S. T. Birchfield, D. K. Gillmor, Fast Bayesian Acoustic Localization, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, Florida, May 2002 2 (2002) 1793–1796.

[3] C. Chen, R. Hudson, Maximum-Likelihood Acoustic Source Localization: Experimental Results, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, Florida, May 2002 3 (2002) 2949–2952.

[4] C. Chen, R. Hudson, K. Yao, Maximum-Likelihood Source Localization and Unknown Sensor Location Estimation for Wideband Signals in the Near-Field, IEEE transactions on Signal Processing 50 (8) (2002) 1843–1854.

[5] 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, Orlando, Florida, May 2002 3 (2002) 2965–2968.

[6] C. Knapp, G. Carter, The Generalized Correlation Method for Estimation of Time Delay, IEEE Trans. Acoust., Speech, Signal Processing 24 (4) (1976) 320–327.

[7] B. Champagne, S. Bédard, A. Stéphenne, Performance of Time-Delay Estimation in the presence of Room Reverberation, IEEE Trans. Acoust., Speech, Signal Processing 4 (2) (1996) 148–152.

[8] A. St$\acute{e}$phene, B. Champagne, A New Cepstral Prefiltering Technique for Estimating Time Delay Under Reverberant Conditions, Signal Processing 59 (1997) 253–266.

[9] M. Brandstein, H. Silverman, A Robust Method for Speech Signal Time-Delay Estimation in Reverberant Rooms, Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing (1997) 375–378.

[10] 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, Orlando, Florida, May 2002 1 (2002) 893–896.

[11] C. Nikias, R. Pan, Time delay Estimation in Unknown Gaussian Spatially Correlated Noise, IEEE Trans. Acoust., Speech, Signal Processing 36 (11) (1988) 1706–1714.

[12] H. H. Chiang, C. L. Nikias, A New Method for Adaptive Time-Delay Estimation for Non-Gaussian Signals, IEEE Trans. Acoust., Speech, Signal Processing 38 (2) (1990) 209–219.

[13] J. Benesty, Adaptive Eigenvalue Decomposition Algorithm for Passive Acoustic Source Localization, Acoustic Society of America 107 (1) (2000) 384–391.

[14] S. Doclo, M. Moonen, Robust Adaptive Time Delay Estimation for Speaker Localisation in Noisy and Reverberant Acoustic Environments, EURASIP Journal on Applied Signal Processing 2003 (11) (2003) 1110–1124.

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

[16] T. Dvorkind, S. Gannot, Speaker Localization in a Reverberant Environment, IEEE proceedings, The 22nd convention of Electrical and Electronics Engineers in Israel. (2002) 7–9.

[17] T. Dvorkind and 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.

[18] O. Shalvi, E. Weinstein, System Identification Using Nonstationary Signals, IEEE Trans. Signal Proc. 44 (8) (1996) 2055–2063.

[19] E. Weinstein, M. Feder, A. V. Oppenheim, Multi-Channel Signal Seperation by Decorrelation, IEEE transactions on Speech and Audio Processing 1 (4) (1993) 405–413.

[20] S. Gannot, A. Yeredor, Noise Cancellation with Static Mixtures of a Nonstationary Signal and Stationary Noise, EURASIP Journal on Applied Signal Processing 2002 (12) (2002) 1460–1472.

[21] L. Parra, C. Spence, Convolutive Blind Seperation of Non-Stationary Sources, IEEE transactions on Speech and Audio Processing 8 (3) (2000) 320–327.

[22] D. Schobben, P. Sommen, A Frequency Domain Blind Signal Separation Method Based on Decorrelation, IEEE transactions on signal processing 50 (8) (2002) 1855–1865.

[23] K. Rahbar and J. P. Reilly, Blind Source Seperation for MIMO Convolutive Mixtures, in: 3rd Int. Conf. on ICA and BSS, San Diego, California, USA, 2001, pp. 242–247.

[24] J.B. Allen and D.A. Berkley, Image Method for Efficiently Simulating Small-Room Acoustics, J. Acoust. Soc. Am 65 (4) (1979) 943–950.

[25] 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.

[26] 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 transactions on Audio and Electroacoustics AU-15 (2) (1967) 70–73.

[27] National Institute of Standards and Technology, The DARPA TIMIT acoustic-phonetic continuous speech corpus, CD-ROM NIST Speech Disc 1-1.1 (Oct. 1991).

[28] A. Varga and 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 Comm. 12 (1993) 247–251.

[29] S. Doclo, Modification of Robust Time-Delay Estimation in Highly Adverse Acoustic Environments for Tracking Scenarios, Private communication (Aug. 2003).

[30] T. Kailath, A. Sayed, B. Hassibi, Linear Estimation, Prentice Hall, 2000.

[31] S. Haykin, Adaptive Filter Theory, 3rd Edition, Prentice Hall, 1996.