Source Localization in Reverberant Environments: Part II - Statistical Analysis

Tony Gustafsson*, Bhaskar D. Rao, and Mohan Trivedi

Abstract

The main difficulty in building robust practical systems for acoustical source localization using microphone arrays, is the effects of room-reverberation. In this paper, a statistical analysis is presented of the influence of room reverberation on source localization techniques. Using a statistical reverberation model presented in a companion paper, the Cramér-Rao lower bound for time-deley estimation and maximum likelihood estimators are derived. The probability of a large error is investigated, and also the performance of common time-delay estimation techniques is analyzed. An interesting outcome of the analysis is that the so-called PHAT time-delay estimator is shown to be optimal among a class of cross-correlation based time-delay estimators.

SP-EDICS: 2-ROOM

I. Introduction

Several approaches for acoustical source localization and speech acquisition, using microphone arrays, have appeared in the literature. Microphone arrays have been a promising solution for localization, while at the same time suppressing interference and noise. In a reverberant environment, the measured signals from a pair of microphones can be modeled as

$$\mathbf{x}(t) = \int_{-\infty}^{\infty} \mathbf{h}(t-\lambda)s(\lambda)d\lambda + \mathbf{n}(t), \tag{1}$$

where $\mathbf{x}(t) = [x_1(t)\ x_2(t)]^T$, $\mathbf{h}(t) = [h_1(t)\ h_2(t)]^T$, and $\mathbf{n}(t) = [n_1(t)\ n_2(t)]^T$. Here, $h_i(t)$ represents the impulse response of the acoustical transfer function from the source to the i^{th} microphone, $x_i(t)$ is the output of the i^{th} receiver, $s(t)$ is the unknown source signal, $n_i(t)$ is an additive noise term. See for example [13, Chapter 5] for a thorough treatment of the reverberation phenomenon.

This work was performed while T. Gustafsson was visiting University of California San Diego, Department of Electrical and Computer Engineering, 9500 Gilman Drive Mail Code 0407, La Jolla, CA 92093-0407 USA. email: tgustaf@ece.ucsd.edu. Support by the Swedish Foundation for International Cooperation in Research and Higher Education, and Telefonaktiebolaget LM Ericsson is gratefully acknowledged.

B. Rao is with University of California San Diego. email: brao@ece.ucsd.edu. This work was supported by UC DiMi Program # D97-17.

M. Trivedi is with University of California San Diego. email: trivedi@ece.ucsd.edu.

Most existing techniques for source localization, cf. [4], [2], [15], [17], are however based on the simpler model

$$\begin{aligned} x_1(t) &= s(t) + n_1(t) \\ x_2(t) &= s(t - \tau_0) + n_2(t), \end{aligned} \tag{2}$$

where τ_0 denotes the time-delay. This simple propagation model is not very realistic in practical environments. Among the few analyses available on the influence of reverberation on the estimate of τ_0, we mention [5], [9]. In [9], Ianniello studied the case with one source and two or three resolvable propagation paths. The effects of multi-path were then analyzed by computing a lower bound on the probability of large error estimates. However, since room reverberation consists of the superposition of a large number of echoes from different directions with different magnitudes, the results in [9] are limited in analyzing the effects of reverberation. In [5], Champagne et al. presented results using the image method for simulating the reverberant channel $\mathbf{h}(t)$. An interesting result of [5], was that the authors proposed a Cramér-Rao lower Bound (CRB) for the variance of the estimated time-delay, when reverberation is present. The introduced CRB showed good agreement with Monte-Carlo simulations. There was however no analytical motivation of why the introduced CRB should be relevant. Nevertheless, the numerical study in [5] clearly demonstrated the adverse effects of reverberation.

The purpose of the present paper is analyze of the problem of acoustic source localization in a reverberant environment. The statistical analysis is based on a model proposed in a companion paper [7]. Using the model of [7], several interesting results are derived. Among others, we will derive the CRB for estimation of τ_0.

II. Review of Single-path Propagation Time-delay Estimation (TDE)

The Generalized Cross Correlation (GCC) [12] method is probably the most popular method for estimating time-delays. Its popularity is due to high accuracy, and low computational complexity which is achieved by Fast Fourier Transform (FFT) implementations. In the GCC method, the estimated time-delay is obtained as

$$\hat{\tau} = \arg\max_{\tau} \hat{R}_{GCC}(\tau), \tag{3}$$

where

$$\hat{R}_{GCC}(\tau) \triangleq \int_{-\infty}^{\infty} |G(\omega)|^2 \hat{P}_{12}(\omega) e^{j\omega\tau}, \tag{4}$$

and $\hat{P}_{12}(\omega) \triangleq X_2(\omega) X_1^*(\omega)$ denotes the estimated cross-power spectrum. Here, $X_i(\omega)$ denotes the Fourier transform of $x_i(t)$ over a finite interval $0 \leq t \leq T$, the superscript $(\cdot)^*$ denotes complex conjugate, and $|G(\omega)|^2$ is a weighting function. Under certain conditions, the GCC method is the Maximum Likelihood (ML) estimator of τ_0 [12]. The CRB (assuming $s(t)$, $n_1(t)$ and $n_2(t)$ to be zero-mean, mutually uncorrelated wide-sense stationary Gaussian random processes with power spectra $P_{ss}(\omega)$, $P_{n_1n_1}(\omega)$ and $P_{n_2n_2}(\omega)$ respectively)

is further known to equal

$$\mathrm{CRB}_{sp}(\tau_0) = \left(2T \int_0^\infty \frac{\mathrm{SNR}(\omega)^2}{1+2\mathrm{SNR}(\omega)} \omega^2 d\omega \right)^{-1} \tag{5}$$

where we for notational simplicity assumed that $P_{n_1 n_1}(\omega) = P_{n_2 n_2}(\omega)$. The subscript $(\cdot)_{sp}$ indicates that the CRB is valid for the single-path propagation model (2). In the CRB expression (5), we also introduced the signal to noise ratio (SNR)

$$\mathrm{SNR}(\omega) = \frac{P_{ss}(\omega)}{P_{n_i n_i}(\omega)}. \tag{6}$$

¿From the above brief review we may conclude that TDE under single-path propagation is a well-understood problem. Much less is known about TDE when reverberation is present. Several authors have for example observed that the PHAse Transform (PHAT) method is more robust than other GCC methods in reverberant environments, cf. [3], [10], [16]. PHAT is also a member of the GCC class of algorithms, and is obtained with the following choice of weighting:

$$|G_{PHAT}(\omega)|^2 = \frac{1}{\left|\hat{P}_{12}(\omega)\right|}. \tag{7}$$

An analytical motivation of *why* PHAT is more robust than ML is however not available in the literature.

III. Brief Review of the Reverberation Model

We next briefly review the properties of the model introduced in [7]. Throughout the paper it will be assumed that only a sampled version of $\mathbf{x}(t)$ is available, i.e. $\{\mathbf{x}(nT_s)\}_{n=0}^{N-1}$ where T_s is the sampling interval. The source signal $s(t)$ is assumed band-limited, i.e. its power is zero outside the interval $[f_l, f_u]$ Hz. We further assume that frequency domain data are obtained from the DFT of the sequence $\mathbf{x}(nT_s)$, which implies a frequency domain sampling $\omega_k = 2\pi F_s/N$, $k = 0, 1, \cdots, N-1$, where $F_s = 1/T_s$. We will further assume that the considered array of microphones consists of a uniform linear array with M microphones. It is then assumed that the DFT of $\mathbf{x}(nT_s)$ obeys the model

$$\mathbf{X}(\omega_k) = \left(\mathbf{A}(\omega_k;\tau_0) + \mathbf{R}(\omega_k;\boldsymbol{\theta})\right) S(\omega_k),\ k = k_l, \cdots, k_u, \tag{8}$$

where

$$k_l = \mathrm{round}\left\{\frac{N f_l}{F_s}\right\} \tag{9}$$

$$k_u = \mathrm{round}\left\{\frac{N f_u}{F_s}\right\} \tag{10}$$

$$\mathbf{A}(\omega_k;\tau_0) = \begin{bmatrix} 1 & e^{-j\omega_k\tau_0} & \cdots e^{-j\omega_k\tau_0(M-1)} \end{bmatrix}^T. \tag{11}$$

The M-dimensional vector $\mathbf{R}(\omega_k;\boldsymbol{\theta})$ is assumed to be zero-mean, circularily symmetric Gaussian, and to have a covariance matrix $E_{\boldsymbol{\theta}}\left\{\mathbf{R}(\omega;\boldsymbol{\theta})\mathbf{R}(\omega;\boldsymbol{\theta})^H\right\} = \sigma^2\mathbf{I}_M$. Here $(\cdot)^H$ denotes complex conjugate transpose, $E_{\boldsymbol{\theta}}\{\cdot\}$

denotes the expectation operator with respect to all source/microphone positions, and

$$\sigma^2 = \frac{16\pi r^2 \beta^2}{\mathcal{A}(1-\beta^2)}. \tag{12}$$

In the above relationship, r denotes the distance from the source to the array, β denotes the reflection coefficient of the reflecting surfaces of the room in question, $\mathcal{A}$ denotes the total wall area, and $S(\cdot)$ denotes the DFT of the source signal.

As discussed in [7], $\mathbf{R}(\omega_k;\boldsymbol{\theta})$ and $\mathbf{R}(\omega_l;\boldsymbol{\theta})$ are in general correlated, unless the frequency separation $|\omega_k - \omega_l|$ is large. We will discuss this issue in more detal later on.

IV. The Cramér-Rao Lower Bound for TDE

We begin our statistical investigation by deriving the Cramér-Rao lower Bound (CRB) for estimation of τ_0, based on the model (8). It is not obvious what kind of assumptions we should impose on $S(\omega)$. Two options are immediate:

S1: Suppose that $\{S(\omega_k)\}_{k=k_l}^{k_u}$ are unknown but *deterministic* parameters.

S2: Suppose that $\{S(\omega_k)\}_{k=k_l}^{k_u}$ is a sequence of independent zero-mean Gaussian *random variables* with variances $P_{ss}(\omega_k)$.

Remark 1: Note, if $s(t)$ is a wide-sense stationary random processes, the Gaussianity, mutual independence, and variance of $S(\omega_k)$ assumed in *S2*, hold asymptotically (i.e. as $N \to \infty$) under mild assumptions. See for example [11, Chapter 15].

The outcome of the statistical analysis certainly depends on how $S(\omega_k)$ is modeled. In the narrowband sensor array processing case, the corresponding CRB's are commonly referred to as *deterministic/conditional* and *stochastic/unconditional* CRB's (see for example the discussion in [19]). In the sensor array processing literature, it is well known that the CRB corresponding to the deterministic case usually is too optimistic and hence unreachable. In the present scenario, the deterministic CRB would in general depend on the correlation length $\rho(T_{60})$ of $\mathbf{R}(\omega;\boldsymbol{\theta})$ (cf. the definition in [7]). Based on these observations, we have chosen to omit the deterministic case in the present investigation.

Consider next the stochastic case *S2*. Also in this case we face some technical problems. The reason for this is the assumed Gaussianity of $\mathbf{R}(\omega;\boldsymbol{\theta})$ *and* $S(\omega)$, leaving the distribution of their product undetermined. However, in the following we ignore this fact, and *heuristically* assume that also $\mathbf{X}(\omega_k)$ is Gaussian.

Due to the assumed whiteness of $S(\omega_k)$ (i.e. *S2*), $\mathbf{X}(\omega_{k_1})$ and $\mathbf{X}(\omega_{k_2})$ are independent. This observation holds irrespective of the value of the coherence bandwidth $\rho(T_{60})$. To see this, study the expected value of the product of two arbitrary elements $\mathbf{X}(\omega_k)$ and $\mathbf{X}(\omega_l)$ $(k \neq l)$:

$$\begin{aligned} &E_{\boldsymbol{\theta},S}\left\{ \left(\mathbf{A}(\omega_k)S(\omega_k) + \mathbf{R}(\omega_k;\boldsymbol{\theta})S(\omega_k)\right)\left(\mathbf{A}(\omega_l)S(\omega_l) + \mathbf{R}(\omega_l;\boldsymbol{\theta})S(\omega_l)\right)^H \right\} \\ &\quad = E_S \underbrace{\left\{S(\omega_k)S(\omega_l)^*\right\}}_{=0} E_{\boldsymbol{\theta}}\left\{ \left(\mathbf{A}(\omega_k) + \mathbf{R}(\omega_k;\boldsymbol{\theta})\right)\left(\mathbf{A}(\omega_l) + \mathbf{R}(\omega_l;\boldsymbol{\theta})\right)^H \right\} = 0. \end{aligned} \tag{13}$$

The result follows from the whiteness of $S(\omega_k)$ and the mutual independence of $\mathbf{R}(\omega;\boldsymbol{\theta})$ and $S(\omega)$. For the above result to hold true, the expectation operator must be defined as the ensemble average over all signal realizations, *and* over all source/microphone positions, denoted as $E_{\boldsymbol{\theta},S}\{\cdot\}$. Note that if $S(\omega_{k_1})$ and $S(\omega_{k_2})$ were correlated, or if we used the deterministic setting *S1*, the relationship (13) would not be true, and the CRB computation would become more complicated.

Proposition 1: Suppose that $\{\mathbf{X}(\omega_k)\}_{k=k_l}^{k_u}$ is a sequence of independent zero-mean Gaussian random variables with

$$E_{\boldsymbol{\theta},S}\left\{\mathbf{X}(\omega_k)\mathbf{X}(\omega_k)^H\right\} = P_{ss}(\omega_k)\left(\sigma^2\mathbf{I}_M + \mathbf{A}(\omega_k;\tau_0)\mathbf{A}(\omega_k;\tau_0)^H\right) \quad (14)$$

$$E_{\boldsymbol{\theta},S}\left\{\mathbf{X}(\omega_k)\mathbf{X}(\omega_k)^T\right\} = 0, \quad (15)$$

where $P_{ss}(\omega_k) > 0$. Then, any unbiased estimator $\hat{\tau}$ of the true time-delay τ_0 satisfies

$$E_{\boldsymbol{\theta},S}\left\{(\hat{\tau}-\tau_0)^2\right\} \geq \mathrm{CRB}_{rev}(\tau_0;M) = \frac{\sigma^4}{\sum_{k=k_l}^{k_u}\mathrm{Tr}\left\{\mathbf{P}_\mathrm{k}\mathbf{D}_\mathrm{k}\mathbf{P}_\mathrm{k}\mathbf{D}_\mathrm{k}\right\}} \quad (16)$$

where subscript "*rev*" indicates CRB for the reverberation model,

$$\mathbf{D}_k \triangleq \mathbf{A}(\omega_k;\tau_0)\mathbf{D}^H(\omega_k;\tau_0) + \mathbf{D}(\omega_k;\tau_0)\mathbf{A}^H(\omega_k;\tau_0) \quad (17)$$

$$\mathbf{P}_k \triangleq \mathbf{I}_M - \frac{\mathbf{A}(\omega_k;\tau_0)\mathbf{A}^H(\omega_k;\tau_0)}{\sigma^2+M} \quad (18)$$

$$\mathbf{D}(\omega_k;\tau_0) \triangleq \frac{\partial\mathbf{A}(\omega_k;\tau)}{\partial\tau}, \quad (19)$$

where the right hand side of (19) is evaluated at $\tau=\tau_0$.

Proof: See Appendix A.

To gain some insight into the above results, we next evaluate the CRB for the interesting case when $M=2$:

Corollary 1: Under the same assumptions as in Proposition 1, but for the special case $M=2$, expression (16) simplifies to

$$\mathrm{CRB}_{rev}(\tau_0;M=2) = \left(2\frac{\mathrm{SRR}^2}{1+2\mathrm{SRR}}\sum_{k=k_l}^{k_u}\omega_k^2\right)^{-1} \quad (20)$$

where the Signal to Reverberation Ratio (SRR) is defined as

$$\mathrm{SRR} \triangleq \frac{1}{\sigma^2} = \frac{\mathcal{A}(1-\beta^2)}{16\pi r^2\beta^2}. \quad (21)$$

Proof: Straightforward calculations which are omitted.

It is interesting to note the similarity between (20) and the CRB expression for single-path time-delay estimation in additive noise (5). The expression (20) does however not contain the power spectrum of the source signal, which is an important distinction.

Remark 2: Next we would like to comment on how to include the effects of additive measurement noise in the above CRB expressions. Assuming that the additive noise $n_i(t)$ is zero-mean, white, Gaussian, and with variance η^2, the CRB expression for the case with additive measurement noise present reads as

$$\widetilde{\mathrm{CRB}}_{rev}(\tau_0; M=2) = \left(2 \sum_{k=k_l}^{k_u} \frac{\mathrm{SNRR}(\omega_k)^2}{1+2\mathrm{SNRR}(\omega_k)} \omega_k^2 \right)^{-1} \tag{22}$$

where the signal to noise and reverberation ratio (SNRR) is defined as

$$\mathrm{SNRR}(\omega) = \frac{P_{ss}(\omega)\frac{1}{4\pi r^2}}{P_{ss}(\omega)\frac{4\beta^2}{\mathcal{A}(1-\beta^2)} + \eta^2}. \tag{23}$$

Note that $P_{ss}(\omega)$ now appears in the CRB, in contrast to (20).

A promising approach to analyze the accuracy of $\hat{\tau}$ (as defined in equation (3)) was recently proposed in [5]. It was suggested that the single-path CRB (5) is valid also in reverberant environments, with the important modification that $\mathrm{SNR}(\omega)$ is modified to account for reverberation. The "equivalent SNR" suggested in [5] reads as

$$\left(\mathrm{SNR}_{eq}(\omega)\right)_i = \frac{|H_i(\omega;0)|^2 P_{ss}(\omega)}{P_{n_i n_i}(\omega) + |H_i(\omega;\beta) - H_i(\omega;0)|^2 P_{ss}(\omega)}, \tag{24}$$

where $H_i(\omega;0)$ denotes the transfer functio from the source to the i^{th} microphone in case of no reverberation, and $H_i(\omega;\beta)$ denotes the same transfer function with reverberation included. We note that the SNRR in (23) corresponds to the average value of SNR_{eq} (24), assuming that the quantity $(H_i(\omega;\beta) - H_i(\omega;0))$ corresponds to diffuse sound. We have however, in contrast to [5], derived $\mathrm{CRB}_{rev}(\tau_0)$ under precise modeling assumptions.

V. Analysis of GCC Techniques

A. The Probability of an Anomalous Estimate

Knowledge of $CRB_{rev}(\tau_0)$ for a particular room configuration is certainly an important factor to consider. However, the derived CRB is rather local in the sense that it is reachable only for large SRR. In practical applications, the limiting factor of the performance is rather the fact that GCC-based localization methods suffer from outliers, simply because the "wrong peak" of the GCC function $\hat{R}_{GCC}(\tau)$ is selected. It is therefore of great interest to analyze the probability of an anomalous estimate, or in other words, the probability of selecting the wrong peak of $\hat{R}_{GCC}(\tau)$. For the case with single-path propagation and additive measurement noise, Ianniello analyzed this probability in a classical paper [8]. The purpose of the following section is then to extend Ianniello's analysis to include reverberation.

In addition to the general assumptions introduced in Section III, we introduce the following assumptions:
E1: Assume that the sampled source signal $s(nT_s)$ is a zero mean white ergodic random process:

$$E_S\left\{s(nT_s)\right\} = 0 \tag{25}$$

$$E_S\left\{s(nT_s)s((n+l)T_s)\right\} = R_{ss}\delta_l, \tag{26}$$

where δ_l denotes Kronecker's delta function.

E2: Assume that $m \ll \tilde{N} \ll N$, where $\tilde{N} \triangleq F_s/\rho(T_{60})$.

The basic idea of the analysis technique is as follows. For a given realization $\mathbf{h}(t;\boldsymbol{\theta})$, the conditional cross-correlation between $x_1(t)$ and $x_2(t)$ equals

$$E_S\{x_1(t)x_2(t-\tau)|\boldsymbol{\theta}\} = h_1(\tau;\boldsymbol{\theta}) * R_{ss}(\tau) * h_2(-\tau;\boldsymbol{\theta}), \tag{27}$$

where $*$ denotes convolution, and $R_{ss}(\tau)$ denotes the auto-correlation function of $s(t)$. Hence, even if $s(t)$ is ergodic and if $N \to \infty$, it cannot be expected that the maximum of $E_S\{x_1(t)x_2(t-\tau)|\boldsymbol{\theta}\}$ appears in the vicinity of τ_0. For N large (as dictated by Assumption E2), the analysis should focus on the behavior of the random variable $E_S\{x_1(t)x_2(t-\tau)|\boldsymbol{\theta}\}$. The measurement time N will hence not appear in our analysis.

Consider now the GCC function, which is computed using $\mathbf{x}(nT_s)$ to produce the sequence $\{\hat{R}_{GCC}(nT_s)\}$ where $n = -\tau_m, -\tau_m+1, \cdots, \tau_m$ and $\tau_m = d/(cT_s)$. Here, τ_m denotes the maximum relative time-delay. *If* the sequence $\{\hat{R}_{GCC}(nT_s)\}$ consists of independent random variables, a reasonable definition of an anomalous event is the following:

$$\mathcal{E} \triangleq \left[\hat{R}_{GCC}(nT_s) > \hat{R}_{GCC}(n_0T_s) \text{ for at least one } [-\tau_m, \tau_m] \ni n \neq n_0\right] \tag{28}$$

where n_0 corresponds to the true time-delay (to simplify the presentation we assume that $n_0 = \tau_0/T_s$ and τ_m are integers). The above definition of the anomalous event $\mathcal{E}$ is essentially taken from [8]. In [8], the anomalous event $\mathcal{E}$ was defined using the correlation length of $s(t)$. More precisely, if T_c denotes the correlation length of $s(t)$, $\mathcal{E}$ was defined by sampling $\hat{R}_{GCC}(\tau)$ with sampling interval T_c.

If we next denote the number of available samples of the GCC function as $m = 2\tau_m + 1$, the probability of an outlier can be computed from (see [8])

$$\text{Prob[Outlier]} \simeq \text{Prob}[\mathcal{E}] = 1 - \int_{-\infty}^{\infty} p(z_0) \left\{\int_{-\infty}^{z_0} p(z_n)dz_n\right\}^{m-1} dz_0. \tag{29}$$

Here we introduced the following notations:

- $z_0 \triangleq \hat{R}_{GCC}(n_0T_s)$, i.e. the value of the GCC function for the true time-delay.
- $z_n \triangleq \hat{R}_{GCC}(nT_s)$, for $n \neq n_0$.
- $p(z_0)$: the probability density function (PDF) of z_0.
- $p(z_n)$: the PDF of z_n for $n \neq n_0$.

Note, the expression (29) for the outlier probability only makes sense if the sequence $\hat{R}_{GCC}(nT_s)$ consists of mutually independent random variables. The strategy for analyzing Prob[Outlier] should then be clear:

1. Determine under which conditions the sequence $\{\hat{R}_{GCC}(nT_s)\}_{n=-\tau_m}^{\tau_m}$ consists of mutually independent random variables.

2. Find the PDF's $p(z_0)$ and $p(z_n)$.

3. Evaluate the integral (29).

We then have the following interesting result:

Proposition 2: Under the general assumptions introduced in Section III and under assumptions *E1-E2*, the sequence $\{\hat{R}_{GCC}(nT_s)\}_{n=-\tau_m}^{\tau_m}$ consists of mutually independent random variables with the following distributions (assuming that $|G(\omega)|^2 = 1$):

$$\hat{R}_{GCC}(n_0 T_s) \in \mathcal{N}\left(R_{ss}, \frac{R_{ss}^2}{\tilde{N}}\left(\frac{2}{\mathrm{SRR}} + \frac{1}{\mathrm{SRR}^2}\right)\right) \tag{30}$$

$$\hat{R}_{GCC}(nT_s) \in \mathcal{N}\left(0, \frac{R_{ss}^2}{\tilde{N}}\left(\frac{2}{\mathrm{SRR}} + \frac{1}{\mathrm{SRR}^2}\right)\right). \tag{31}$$

Proof: See Appendix B

To compute the probability of an outlier, it only remains to compute the integral (29) using the results in Proposition 2. This integral has to be implemented using numerical integration. A couple of remarks are in place:

Remark 3: Since $\tilde{N}$ is proportional to T_{60}, one may erroneously think that the variance of $\hat{R}_{GCC}(nT_s)$ decreases as the reverberation time increases. This is however not correct since 1/SRR increases at a faster rate than $\tilde{N}$, as T_{60} increases.

Remark 4: Previously we assumed that $s(t)$ is white. In a practical setup with $s(t)$ representing speech, the whiteness assumption is typically violated. To avoid these difficulties, Ianniello used the correlation length of $s(t)$ in his definition of the event $\mathcal{E}$. However, we conjecture that Proposition 2 is relevant also for colored signals, assuming that we apply a GCC method (such as PHAT) that employs pre-whitening.

Remark 5: In the numerical experiments, we will study source signals which are spectrally flat within an interval $[f_l, f_u]$. A simple approach to modify the variance expressions in Proposition (2), to accommodate for the band-pass characteristics of $s(t)$, is to replace $\tilde{N}$ with $2\tilde{N}(f_u - f_l)/F_s$. This modification is applied in the numerical examples.

With the above results on the probability of selecting the wrong peak of the GCC function, we can also predict how accurately it is possible to estimate the time-delay in a reverberant environment. For that purpose, assume that the variance of the GCC estimate $\hat{\tau}$ equals $\mathrm{CRB}_{rev}(\tau_0)$ in cases where the correct peak is selected. Furthermore, in cases where the wrong peak is selected, we assume that $\hat{\tau}$ is uniformly distributed in the interval $[-d/c, d/c]$. Then the variance of the estimated time-delay $\hat{\tau}$ approximately can be found from the following expression:

$$E_{\boldsymbol{\theta},S}\left\{(\hat{\tau} - \tau_0)^2\right\} \simeq (1 - \mathrm{Prob}[\mathcal{E}])\,\mathrm{CRB}_{rev_s}(\tau_0) + \mathrm{Prob}[\mathcal{E}]\frac{d^2}{3c^2}. \tag{32}$$

The variance expression (32) is then a more realistic (and more pessimistic!) bound than $\mathrm{CRB}_{rev}(\tau_0)$.

B. Variance of $\hat{\tau}$

As previously mentioned, several authors have noticed that PHAT performs better than other GCC methods in reverberant conditions. The purpose of the following section is to provide an analytical motivation of this empirical observation. As in the previous section, we will focus on the behaviour of the random variable $E_S\{x_1(t)x_2(t-\tau)|\boldsymbol{\theta}\}$. Hence, it is assumed that N is large. As in the proof of Proposition 2, it is assumed that the frequency axis is sampled according to the coherence bandwidth of $\mathbf{R}(\omega;\boldsymbol{\theta})$ (cf. equations (104) and (105)).

For any choice of the GCC weighting $|G(\omega)|^2$, the estimated time-delay can approximately be written as[1]

$$\hat{\tau} = \arg\min_{\tau} V(\tau), \tag{33}$$

where

$$V(\tau) \triangleq -\left\{ \sum_{k=k_l}^{k_u} |G(\phi_k)|^2 \left(P_{12}(\phi_k;\boldsymbol{\theta}) e^{j\phi_k\tau} + P_{12}(\phi_k;\boldsymbol{\theta})^* e^{-j\phi_k\tau} \right) \right\} \tag{34}$$

$$\hat{P}_{12}(\phi_k;\boldsymbol{\theta}) = P_{ss}(\phi_k) \left(e^{-j\phi_k\tau_0} + \underbrace{R_2(\phi_k;\boldsymbol{\theta}) + R_1^*(\phi_k;\boldsymbol{\theta}) e^{-j\phi_k\tau_0}}_{\epsilon(\phi_k)} + \underbrace{R_2(\phi_k;\boldsymbol{\theta}) R_1^*(\phi_k;\boldsymbol{\theta})}_{\tilde{\epsilon}(\phi_k)} \right). \tag{35}$$

Furthermore,

$$\phi_k = \frac{2\pi F_s}{\tilde{N}},\ k = -\frac{\tilde{N}}{2}, \cdots, \frac{\tilde{N}}{2} - 1 \tag{36}$$

$$\tilde{N} = \frac{F_s}{\rho(T_{60})}. \tag{37}$$

In the above, $\hat{P}_{12}(\phi_k;\boldsymbol{\theta})$ denotes the Fourier-transform of $E_S\{x_1(t)x_2(t-\tau)|\boldsymbol{\theta}\}$. The basic idea in the following is to analyze the accuracy of $\hat{\tau}$ (defined as in (33)) for small levels of the reverberation power σ^2. To perform the analysis, we assume that $\hat{\tau}$ is consistent in the sense that $\hat{\tau} \to \tau_0$ as $\sigma \to 0$. As is well-known from the literature on sensor array signal processing, a sufficient condition for guaranteeing consistency, is that

$$d \leq \frac{c}{2f_l}, \tag{38}$$

where c denotes the propagation speed. Since $\hat{\tau}$ minimizes $V(\tau)$, we have $V'(\hat{\tau}) = 0$, where $V'(\hat{\tau})$ denotes the gradient of $V(\tau)$ evaluated at $\hat{\tau}$. For high SRR, a first order Taylor expansion yields

$$0 = V'(\tau_0) + V''(\tau_0)\left(\hat{\tau} - \tau_0\right) + o_p\left(|V'(\tau_0)|\right), \tag{39}$$

where $V''(\tau_0)$ denotes the Hessian, and $o_p(\cdot)$ is order in probability. It now follows that

$$\hat{\tau} - \tau_0 = -\frac{1}{Z} V'(\tau_0) + o_p\left(|V'(\tau_0)|\right). \tag{40}$$

[1]This approximation assumes that $P_{ss}(\omega)$ is a "smooth" function of ω, in relation to the coherence-bandwidth $\rho(T_{60})$.

where

$$Z \stackrel{\triangle}{=} \lim_{\sigma \to 0} V''(\tau_0). \tag{41}$$

For high SRR, and N large, we find that the mean square error of the estimation error is given by

$$E_{\boldsymbol{\theta}} \left\{ (\hat{\tau} - \tau_0)^2 \right\} = \sigma^2 \frac{K}{Z^2} + o(\sigma^2), \tag{42}$$

where

$$K \stackrel{\triangle}{=} \lim_{\sigma \to 0} \frac{1}{\sigma^2} E_{\boldsymbol{\theta}} \left\{ \left(V'(\tau_0) \right)^2 \right\}. \tag{43}$$

Note then that

$$E_{\boldsymbol{\theta}} \left\{ \epsilon(\phi_k) \epsilon(\phi_k)^* \right\} = 2\sigma^2 \tag{44}$$

$$E_{\boldsymbol{\theta}} \left\{ \tilde{\epsilon}(\phi_k) \tilde{\epsilon}(\phi_k)^* \right\} = \sigma^4 \tag{45}$$

$$E_{\boldsymbol{\theta}} \left\{ \epsilon(\phi_k) \tilde{\epsilon}(\phi_k)^* \right\} = 0. \tag{46}$$

Since we in equation (42) neglect all terms that are of order $o(\sigma^2)$, $P_{12}(\phi_k; \boldsymbol{\theta})$ is approximated as

$$P_{12}(\phi_k; \boldsymbol{\theta}) \simeq P_{ss}(\phi_k) \left(e^{-j\phi_k \tau_0} + \epsilon(\phi_k) \right) \tag{47}$$

Next, we compute the gradient and Hessian matrices as

$$V'(\tau) = \frac{\partial}{\partial \tau} V(\tau) = -\frac{1}{2} \sum_{k=k_l}^{k_u} |G(\phi_k)|^2 \left(P_{12}(\phi_k; \boldsymbol{\theta}) (j\phi_k) e^{j\phi_k \tau} + P_{12}(\phi_k; \boldsymbol{\theta})^* (-j\phi_k) e^{-j\phi_k \tau} \right) \tag{48}$$

$$V''(\tau) = \frac{\partial}{\partial \tau} V'(\tau) = \frac{1}{2} \sum_{k=k_l}^{k_u} |G(\phi_k)|^2 \phi_k^2 \left(P_{12}(\phi_k; \boldsymbol{\theta}) e^{j\phi_k \tau} + P_{12}(\phi_k; \boldsymbol{\theta})^* e^{-j\phi_k \tau} \right). \tag{49}$$

Since $P_{12}(\phi_k; \boldsymbol{\theta}) \to P_{ss}(\phi_k) e^{-j\phi_k \tau_0}$ as $\sigma \to 0$, we find that the Hessian, evaluated at $\tau = \tau_0$, can be written as

$$Z \stackrel{\triangle}{=} \lim_{\sigma \to 0} V''(\tau_0) = \sum_{k=k_l}^{k_u} |G(\phi_k)|^2 \phi_k^2 P_{ss}(\phi_k). \tag{50}$$

Consider next the computation of K. Applying the approximation (47), we obtain

$$\begin{aligned} V'(\tau_0) &\simeq -\frac{1}{2} \sum_{k=k_l}^{k_u} |G(\phi_k)|^2 (j\phi_k) P_{ss}(\phi_k) \left\{ (e^{-j\phi_k \tau_0} + \epsilon(\phi_k)) e^{j\phi_k \tau_0} - (e^{j\phi_k \tau_0} + \epsilon(\phi_k)^*) e^{-j\phi_k \tau_0} \right\} \\ &= -\frac{1}{2} \sum_{k=k_l}^{k_u} |G(\phi_k)|^2 (j\phi_k) P_{ss}(\phi_k) \left\{ \epsilon(\phi_k) e^{j\phi_k \tau_0} - \epsilon(\phi_k)^* e^{-j\phi_k \tau_0} \right\}, \end{aligned} \tag{51}$$

where $\{\epsilon(\phi_k)\}$ is a sequence of circularly symmetric zero-mean independent Gaussian random variables. Hence,

using the assumption that $\epsilon(\phi_k)$ and $\epsilon(\phi_l)$ are independent for $\phi_k \neq \phi_l$, it follows that

$$K = \lim_{\sigma\to 0} \frac{1}{\sigma^2} E_{\boldsymbol{\theta}} \left\{ \sum_{k=k_l}^{k_u} \frac{1}{4} |G(\phi_k)|^4 \phi_k^2 P_{ss}(\phi_k)^2 \left\{ \epsilon(\phi_k) e^{j\phi_k \tau_0} - \epsilon(\phi_k)^* e^{-j\phi_k \tau_0} \right\} \right.$$
$$\left. \left\{ \epsilon(\phi_k) e^{j\phi_k \tau_0} - \epsilon(\phi_k)^* e^{-j\phi_k \tau_0} \right\}^* \right\} \tag{52}$$
$$= \lim_{\sigma\to 0} \frac{1}{\sigma^2} \left(\sum_{k_l=1}^{k_u} \frac{1}{4} |G(\phi_k)|^4 \phi_k^2 P_{ss}(\phi_k)^2 2\sigma^2 \right) = \frac{1}{2} \sum_{k=k_l}^{k_l} |G(\phi_k)|^4 \phi_k^2 P_{ss}(\phi_k)^2,$$

and the derivation of $E_{\boldsymbol{\theta}} \left\{ (\hat{\tau} - \tau_0)^2 \right\}$ is complete.

The mean square error (42) clearly depends on how the weighting sequence $|G(\phi_k)|^2$ is chosen. An interesting question is how $|G(\phi_k)|^2$ should be chosen for lowest possible error variance. The optimal choice of $|G(\phi_k)|^2$ is provided in the following result:

Proposition 3: The lowest possible error variance $E_{\boldsymbol{\theta}} \left\{ (\hat{\tau} - \tau_0)^2 \right\} \simeq \sigma^2 \frac{K}{Z^2}$ is obtained if

$$|G_{opt}(\phi_k)|^2 = \frac{1}{P_{ss}(\phi_k)}. \tag{53}$$

Proof: Define the following matrices

$$\boldsymbol{\Delta}^T \triangleq \left[\phi_{k_l} \sqrt{P_{ss}(\phi_{k_l})} \quad \cdots \quad \phi_{k_u} \sqrt{P_{ss}(\phi_{k_u})} \right] \tag{54}$$
$$\mathbf{G} \triangleq \operatorname{diag} \left\{ |G(\phi_{k_l})|^2, \cdots, |G(\phi_{k_u})|^2 \right\} \tag{55}$$
$$\boldsymbol{\Sigma} \triangleq \operatorname{diag} \left\{ P_{ss}(\phi_{k_l}), \cdots, P_{ss}(\phi_{k_l}) \right\}. \tag{56}$$

Then it is easy to see that

$$\frac{K}{Z^2} = \frac{1}{2} \left(\boldsymbol{\Delta}^T \mathbf{G} \boldsymbol{\Delta} \right)^{-1} \boldsymbol{\Delta}^T \mathbf{G} \boldsymbol{\Sigma} \mathbf{G} \boldsymbol{\Delta} \left(\boldsymbol{\Delta}^T \mathbf{G} \boldsymbol{\Delta} \right)^{-1}. \tag{57}$$

Assuming that $P_{ss}(\phi_k) > 0$ for $k = k_l, \cdots, k_u$, it follows from well-known matrix optimization results (see for example [14, Appendix II.2]) that the best possible weightings are given by $\mathbf{G}_{opt} = \boldsymbol{\Sigma}^{-1}$, and the proof is complete.

In general the quantity $P_{ss}(\phi_k)$ is unknown. However, note that

$$\left| E_{\boldsymbol{\theta},S} \left\{ \hat{P}_{12}(\phi_k) \right\} \right| = \left| P_{ss}(\phi_k) e^{-j\phi_k \tau_0} \right| = P_{ss}(\phi_k). \tag{58}$$

Hence, a natural estimate of $P_{ss}(\phi_k)$ is $\hat{P}_{ss}(\phi_k) = \left| \hat{P}_{12}(\phi_k) \right|$. It is then interesting to see that the resulting estimator corresponds to the PHAT time-delay estimator, compare with equation (7). The above calculations have shown that PHAT in reverberant environments should be considered as the prime choice among the GCC estimators, which agrees well with previous empirical observations cf. [3], [10], [16].

VI. ML Estimation of τ_0

The previous section dealt with the problem of analyzing the accuracy of TDE in reverberant environments. Although we showed that PHAT may be considered as the best GCC-estimator, is of interest to study the

actual ML estimator of τ_0. It should also be noted that PHAT is applicable only to the case with $M = 2$ microphones, whereas the ML estimator can include measurements from $M \geq 2$ microphones. In the following, we study two different approaches to finding the ML estimate of τ_0. These two approaches are obtained by using slightly different models for $\mathbf{X}(\omega_k)$.

A. "Stochastic" ML

In the first case, we assume that the sequence $\{\mathbf{X}(\omega_k\}_{k=k_l}^{k_u}$ consists of independent and Gaussian random variables:

$$\mathbf{X}(\omega_k) \in \mathcal{N}\left(0, P_{ss}(\omega_k)\left(\sigma^2\mathbf{I}_M + \mathbf{A}(\omega_k;\tau_0)\mathbf{A}(\omega_k;\tau_0)^H\right)\right). \tag{59}$$

Concentrating the likelihood function with respect to $\{P_{ss}(\omega_k)\}_{k=k_l}^{k_u}$, straighforward caluclations show that the ML estimator of τ_0 and σ^2 is obtained by minimizing the following criterion function:

$$\begin{aligned} V_{SML}(\tau,\sigma^2) = &\sum_{k=k_l}^{k_u} M\log\left\{\mathbf{X}(\omega_k)^H\left(\mathbf{I}_M - \frac{\mathbf{A}(\omega_k;\tau)\mathbf{A}(\omega_k;\tau)^H}{\sigma^2+M}\right)\mathbf{X}(\omega_k)\right\} \\ &+ (k_u - k_l + 1)\left(1 + \frac{M}{\sigma^2}\right). \end{aligned} \tag{60}$$

The criterion-function $V_{SML}(\cdot)$ clearly depends on both σ^2 and τ. Unfortunately $V_{SML}(\cdot)$ cannot be concentrated with respect to σ^2, which leads to a prohibitively high computational complexity. Therefore, we consider the resulting estimator to be of theoretical interest only.

B. Approximate ML

To find an estimator that is more attractive from a computational point of view, we next propose an Approximative ML (AML) estimator. To arrive at the desired solution, it is assumed that the sequence $\{\mathbf{X}(\omega_k)\}_{k=k_l}^{k_u}$ consists of independent and Gaussian random variables with the following distribution:

$$\mathbf{X}(\omega_k) \in \mathcal{N}\left(\mathbf{A}(\omega_k)S(\omega_k), \sigma_k^2\mathbf{I}_M\right). \tag{61}$$

Hence, it is now assumed that $\{S(\omega_k)\}_{k=k_l}^{k_u}$ consists of unknown but *deterministic* parameters. We further ignore the fact that the variance of $\mathbf{X}(\omega_k)$ is proportional to $|S(\omega_k)|^2$. Instead, we allow the parameters σ_k^2 to be frequency dependent. Note, the introduced model clearly violates the assumed frequency correlation of $\mathbf{R}(\omega;\boldsymbol{\theta})$, since $\mathbf{X}(\omega_{k_1})$ and $\mathbf{X}(\omega_{k_2})$ are assumed independent, while assuming that $S(\omega_k)$ is deterministic. We however ignore this fact in the derivation of the AML-estimator.

Concentrating the resulting likelihood function with respect to all nuisance parameters, the following criterion function is obtained:

$$V_{AML}(\tau) = \sum_{k=k_l}^{k_u}\log\left\{\left\|\mathbf{\Pi}_k^{\perp}(\tau)\mathbf{X}(\omega_k)\right\|^2\right\}, \tag{62}$$

where $\mathbf{\Pi}_k^{\perp}(\tau)$ denotes the projection matrix onto the orthogonal complement of the space spanned by $\mathbf{A}(\omega_k;\tau)$:

$$\mathbf{\Pi}_k^{\perp}(\tau) \stackrel{\triangle}{=} \mathbf{I}_M - \frac{\mathbf{A}(\omega_k;\tau)\mathbf{A}(\omega_k;\tau)^H}{M}. \tag{63}$$

The time-delay τ_0 is estimated by minimizing $V_{AML}(\tau)$ with respect to τ:

$$\hat{\tau} = \arg\min_{\tau} V_{AML}(\tau). \tag{64}$$

The main advantage with the AML approach, is that $V_{AML}(\tau)$ is a function of τ only, which is an important observation considering real-time implementations. The criterion function $V_{AML}(\tau)$ is unfortunately non-linear in τ. To speed up the calculations, we only compute $V_{AML}(\tau)$ on the grid $\tau = nT_s$ for lag-values $|n| = 0, 1, \cdots, \tau_m$. Finally, quadratic interpolation is used to refine the estimate so obtained [8].

Remark 6: In the literature on direction finding for wide-band signals, estimators quite similar to the AML estimator have appeared. For example, in [6] the deterministic ML estimator of τ_0 was derived assuming σ_k^2 to be constant over the frequency band of interest. The resulting cost-function is identical to (62), with the distinction that $\log\{\cdot\}$ is missing.

VII. A Robust Procedure for Source Localization

The purpose of this section is to propose a new source localization method. The key observation is that the reverberation model strongly (via the factor $\kappa(r)^2$) depends on the distance between the speaker and the microphones. Although not included in the model, there is in practice also a strong dependence on the orientation of the speaker. Given a number of microphones distributed over the spatial region of interest, it is clear that time-delays estimated by microphones close to the speaker should be favored in the localization procedure. We should also favor the microphones (if any!) which the speaker are facing. However, if only audio information is available, it is far from obvious how to decide which of the estimated time-delays that should be trusted the most. In the following, a novel procedure for *weighting* of the estimated time-delays will be proposed.

Suppose that P different microphone pairs are distributed over the spatial region. The estimated time-delays are collected in the vector

$$\hat{\boldsymbol{\tau}} \stackrel{\triangle}{=} \left[\hat{\tau}_1, \quad \cdots, \quad \hat{\tau}_P\right]^T. \tag{65}$$

Next we wish to estimate the unknown location of the source, denoted $\mathbf{r}_s$. Given $\hat{\boldsymbol{\tau}}$, several different approaches to solve this problem have been proposed, see for example [2], [17], [18]. Here we focus on a weighted least squares approach: define the estimated source location as

$$\hat{\mathbf{r}}_s = \arg\min_{\mathbf{r}_s} \left(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}(\mathbf{r}_s)\right)^T \mathbf{W} \left(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}(\mathbf{r}_s)\right), \tag{66}$$

where the vector $\boldsymbol{\tau}(\mathbf{r}_s)$ contains the theoretical time-delays as a function of $\mathbf{r}_s$, and $\mathbf{W}$ is a positive definite weighting matrix. In the following we will consider the case when $\mathbf{W}$ is diagonal, i.e. $\mathbf{W} = \text{diag}\{w_1, \cdots, w_P\}$. Hence, we implicitly assume that the estimated time-delays are statistically independent. Considering (66), it is intuitively clear that the elements of $\mathbf{W}$ should be inversely proportional to the variance of $\hat{\boldsymbol{\tau}}$. Such a weighting strategy would be optimal if $\hat{\boldsymbol{\tau}}$ is Gaussian.

The key problem is how the variance of $\hat{\boldsymbol{\tau}}$ can be derived from the available data. The variance of $\hat{\boldsymbol{\tau}}$ seems quite difficult to estimate, and we resort to approximative procedures. For that purpose, recall Proposition 2, which stated that $\hat{R}_{GCC}(nT_s)$, for $n \neq n_0$, is Gaussian with zero-mean and variance proportional to $2\sigma^2 + \sigma^4$. Note next that $\text{CRB}_{rev}(\tau_0; M = 2)$ (20) is proportional to $(2\sigma^2 + \sigma^4)$. Hence,

$$\text{Var}(\hat{R}(nT_s)) \sim \text{CRB}(\tau_0; M = 2), \tag{67}$$

where $\sim$ denotes "proportional to". The basic idea is hence to estimate the quantity $\text{Var}(\hat{R}(nT_s))$, and use this estimate instead of the unknown variance of $\hat{\boldsymbol{\tau}}$.

The proposed algorithm for robust source localization reads as follows:

1. Compute the GCC function for each microphone pair, i.e. evaluate $\{\hat{R}^p_{GCC}(nT_s)\}_{n=-\tau_m}^{\tau_m}$ for $p = 1, \cdots, P$
2. Compute the p^{th} time-delay as

$$\hat{n}_p = \arg\max_n \hat{R}^p_{GCC}(nT_s). \tag{68}$$

3. Compute the p^{th} weight as

$$\hat{w}_p = \frac{1}{\sum_{n \neq \hat{n}_p} \left(\hat{R}^p_{GCC}(nT_s) \right)^2}, \tag{69}$$

where the summation is restricted to lag-values in the interval $[-\tau_m, \tau_m]$.

4. Minimize (66) with respect to $\mathbf{r}_s$, using $\mathbf{W} = \text{diag}\{\hat{w}_1, \cdots, \hat{w}_P\}$.

In practice we have noted that a slight modification of $\hat{w}_p$ is beneficial. The observation we have made, is that realizations of $\hat{R}^p_{GCC}(nT_s)$ where the difference between the largest and the second largest value of $\hat{R}^p_{GCC}(nT_s)$ is small, should have a smaller weight. Hence, in our experiments, the following value of $\hat{w}_p$ has been applied:

$$\hat{w}_p = \frac{\Delta^p}{\sum_{n \neq \hat{n}_p} \left(\hat{R}^p_{GCC}(nT_s) \right)^2}, \tag{70}$$

where Δ^p denotes the difference between the largest and the second largest value of $\hat{R}^p_{GCC}(nT_s)$.

Before we proceed to the numerical examples, it should be noted that Proposition 2 holds under rather restrictive assumptions (such as $s(nT_s)$ being white). Hence, we believe that the proposed procedure for robust source localization should be used together with PHAT. This since the PHAT weighting achieves a kind of pre-filtering, which should make it more likely that the conditions of Proposition 2 are fulfilled.

VIII. Numerical Examples

A. TDE Accuracy

In the following, the sampled source signal $s(nT_s)$ is generated by filtering of a zero-mean white Gaussian noise sequence with unit variance through a 6^{th} order band-pass Butterworth filter. The cut-off frequencies of the Butterworth filter were chosen as $f_l = 300$ Hz and $\bar{f}_u = 3500$ Hz, respectively. The signal bandwidth then roughly corresponds to the frequency band of human speech. The sampled source signal is subsequently convolved with the simulated room response $\mathbf{h}(nT_s)$, which results in the sampled microphone signal $\mathbf{x}(nT_s)$. To generate independent realizations of $\mathbf{R}(\omega;\boldsymbol{\theta})$, we apply the image method [1], and the procedure described in [7].

In the following, three different time-delay estimators are studied:

- CC: Ordinary cross-correlation. That is, choose the GCC weighting as $|G(\omega)|^2 = 1$.
- PHAT.
- AML.

All algorithms are evaluated on a grid $|n|T_s = 0, 1, \cdots, \tau_m$. To refine the estimated time-delay, quadratic interpolation is applied as in [8]. In all simulations, we further use $N = 2048$ samples to estimate the time-delay, and for each value of the reverberation time T_{60}, 100 Monte-Carlo simulations are performed. In the derivation of the outlier probability, we used the coherence band-width $\rho(T_{60})$ to construct uncorrelated snapshots of $\mathbf{R}(\omega;\boldsymbol{\theta})$. Furthermore, in [7], two reasonable definitions were introduced: $\rho(T_{60}) = 2/T_{60}$ and $\rho(T_{60}) = 7/T_{60}$. The outcome of the theoretical outlier probability certainly depends on which definition we chose. Our experience is that a value somewhere in-between these extremes (say $\simeq 3/T_{60}$) gives the best agreement with the empirical results. In the simulations, we for completeness include both definitions $\rho(T_{60}) = 2/T_{60}$ and $\rho(T_{60}) = 7/T_{60}$. The same remark applies to the derived expression for the lowest possible error variance of the GCC-method, and also in this case we include both definitions of $\rho(T_{60})$. In the numerical experiments, we say that an outlier has occurred if the estimation error is larger than T_s.

Consider next Fig. 1. The predicted expressions for the CRB and the probability of an outlier show a good agreement with the empirical accuracy. The empirical variance of the GCC methods follows closely their theoretical counterparts. It is interesting to note that for T_{60} small, the definition $\rho(T_{60}) = 2/T_{60}$ seems accurate, whereas the definition $\rho(T_{60}) = 7/T_{60}$ seems more accurate for larger reverberation times. However, when $T_{60} > 0.1s$, the probability of outliers becomes dominant, and the estimation error is no longer tolerable. The AML estimator reaches the CRB only for small values of T_{60}, and suffers also from outliers. Hence, in this case there seems to be no benefits from using AML.

Next we consider a possible remedy for decreasing the probability of an outlier. The scenario is identical to the previous one, i.e. $d = 1m$. However, we now include two additional microphones in-between the two previous ones, i.e. along the x-axis there are microphones at $x = -1/2, -1/3, 1/3, 1/2\ m$. The GCC methods

use only the ones at $x = -1/2, 1/2$ (as in the previous simulation), whereas the AML method uses data from all four microphones. The outcome of 100 Monte-Carlo simulations is illustrated in Fig. 2. For high SRR, the AML method offers only a marginal improvement. However, from Fig. 2 it should be noted that the AML method does not suffer from outliers until $T_{60} \simeq 0.5s$, whereas the GCC methods produce outliers already for $T_{60} \simeq 0.25s$. Hence, simultaneous processing of multiple microphones seems to be an effective way of reducing the probability of outliers.

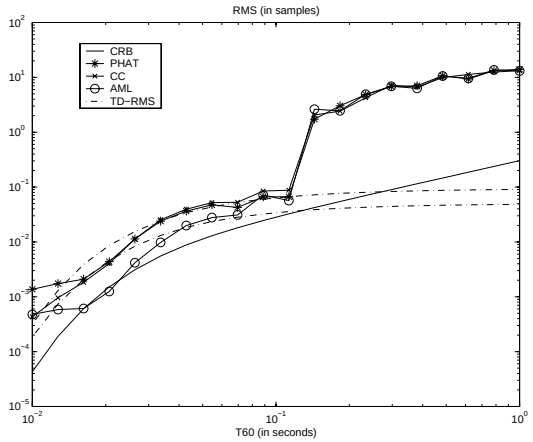
Finally, we once again consider the case $M = 2$, but we modify the input signal slightly. The source signal is now generated by superimposing two sinusoids on the previous bandpass random signal. The sinusoids have unit amplitude, frequencies 2500Hz and 1250Hz, and the phases of the sinusoids are chosen randomly for each realization. The outcome of 100 Monte-Carlo runs is shown in Fig. 3. From Fig. 3, we note that PHAT performs much better than CC, which agrees well with the theoretical results. Note especially that CC suffers from outliers already when $T_{60} > 0.09s$, whereas PHAT performs satisfactorily until $T_{60} > 0.25s$. Note also that the theoretical outlier-probability agrees well with the empirical one, even if the applied source signal is colored.

B. Robust Source Localization Using Real Data

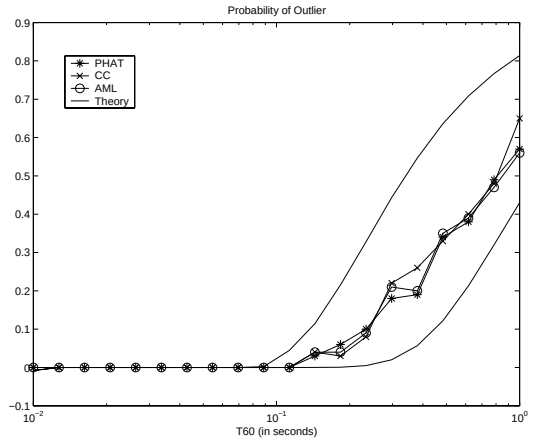
In the final example, we will present experimental results applying the robust method described in Section VII. For this purpose, consider Figure 5, which illustrates the microphone arrangement on the left wall of Figure 4. An identical array with four microphones is placed in the ceiling.

The microphone outputs are sampled with $F_s = 16$kHz, and the source signal is a male speaker reading from the US tax-law. For each source location/orientation, we collect 10 seconds of data. The source location is estimated as outlined in Section VII. Furthermore, the cross-power spectrum is obtained by averaging 10 consecutive realizations of $\hat{P}_{12}(\omega)$, where each $\hat{P}_{12}(\omega)$ is computed using $N = 1024$ samples of $\mathbf{x}(t)$.

The outcome of these experiments is shown in Fig. 6. The difference between the two cases in Fig. 6 is that the speaker orientation is varied. Several interesting observations can be made. First we note that the accuracy of the system is highly dependent on the orientation of the speaker. We also note that the proposed strategy for weighting increases the accuracy significantly. Note, for example, the results in Fig. 6b. This case is obviously difficult, and the un-weighted method produces estimates which are useless. The outcome of the weighted method, is that we at least estimate the bearing from one array accurately, whereas information from the other array is completely ignored. Hence, if there are sufficiently many microphones available, so that the speaker always is relatively close to a couple of them, we believe that the proposed weighting-scheme offers a robust solution for source localization.



(a) Root mean square error, in samples, for various estimators.



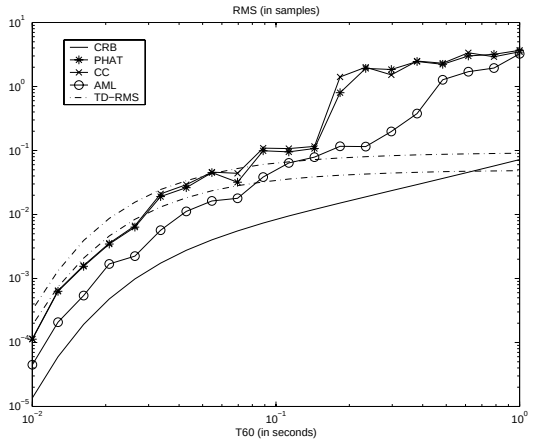
(b) Empirical probability of an anomalous estimate.

Fig. 1. Simulation results for the case with two microphones, as a function of the reverberation time. Here the input signal consists of filtered white noise. $d = 1m$, $r = 3m$, and $\tau_0 = 2T_s$. TD-RMS shows the theoretical value of the standard deviation of the GCC estimate, assuming optimal weighting.

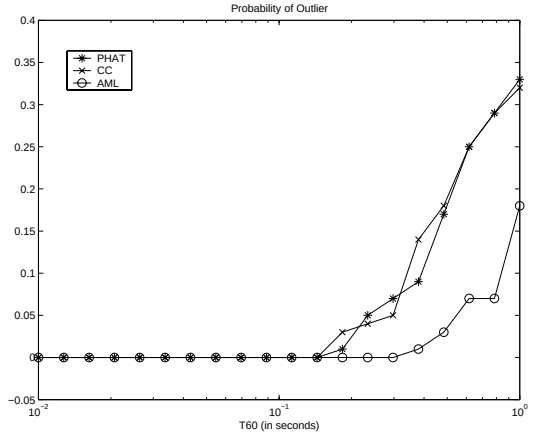
IX. Conclusions

Several practical studies in the literature have indicated that the problem of localizing acoustical sources in reverberant environments is difficult. The main difficulty is to design systems that are robust against room reverberation. There has hence been an interest in understanding and analyzing the performance when room reverberation is present.

Based on the new model introduced in a companion paper, the "fading-like" reverberation model was in the present paper applied to perform a statistical analysis. Among others, the Cramér-Rao lower Bound for



(a) Root mean square error, in samples, for various estimators.



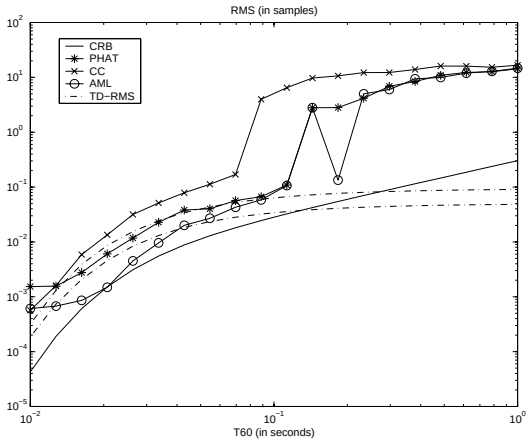
(b) Empirical probability of an anomalous estimate.

Fig. 2. Simulation results for the case with four microphones and a bandpass source signal, as a function of the reverberation time. Here $d = 1/3m$, $r = 3m$, and $\tau_0 = 0T_s$. TD-RMS shows the theoretical value of the standard deviation of the GCC estimate, assuming optimal weighting.

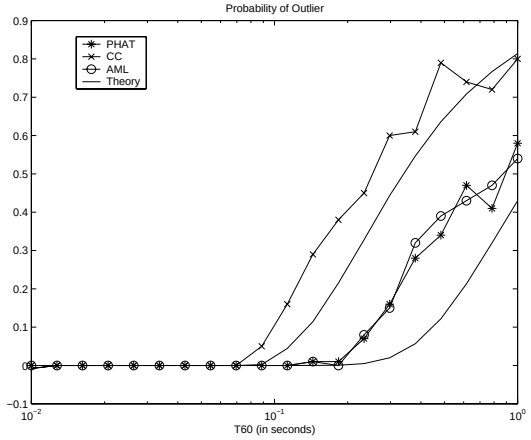
the variance of the estimated time-delay was derived. Also, an analytical expression for the probability of an anomalous estimate was computed.

The derived CRB, together with the probability of an anomalous estimate, then provide tools for analyzing systems for source localization. Unfortunately, the obtained results are rather pessimistic, in the sense that they clearly show the severe effects of reverberation. The variance of the estimated time-delay is typically tolerable only if the speaker is close to the microphones, or if the level of reverberation is small.

The derived results however indicate a possible remedy. The derived CRB depends strongly on the number



(a) Root mean square error, in samples, for various estimators.



(b) Empirical probability of an anomalous estimate.

Fig. 3. Simulation results for the case with two microphones, as a function of the reverberation time. Here the input signal consists of filtered white noise with two superimposed sinusoids. $d = 1m$, $r = 3m$, and $\tau_0 = 2T_s$. TD-RMS shows the theoretical value of the standard deviation of the GCC estimate, assuming optimal weighting.

of microphones used to estimate the bearing. Hence, a more complete study of the attainable accuracy when several microphones simultaneously are processed is highly desired. Another possibility for designing more robust estimators, is to combine audio and video information. The main idea then is to use video information to provide initial estimates of the positions of potential speakers. Then it should be possible to decrease the probability of anomalous estimates, by properly incorporating the video-information. For example, the video-information can tell us in which range the true time-delay should be found, which potentially could decrease the probability of an anomalous estimate.

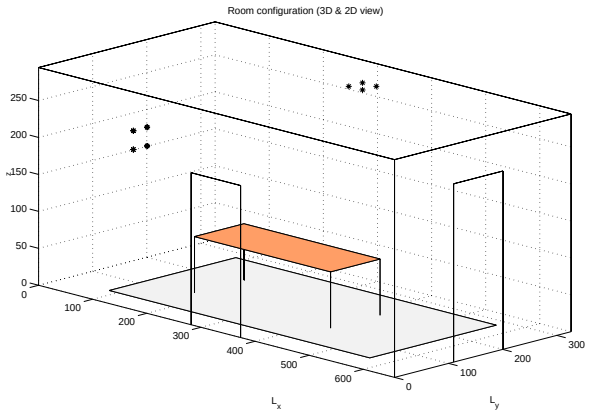


Fig. 4. Room configuration (dimensions in cm). + - microphones)

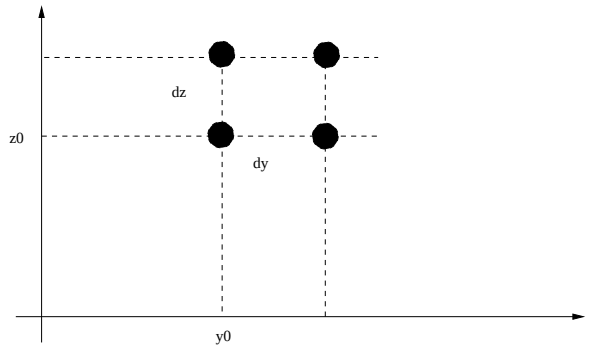


Fig. 5. Microphone arrangement. Here, $y_0 = 1.65$, $z_0 = 1.55$, and $\Delta y = \Delta z = 0.25$ (all measures are in meters).

Appendix

I. Proof of Proposition 1

Collect all unknown parameters in the vector

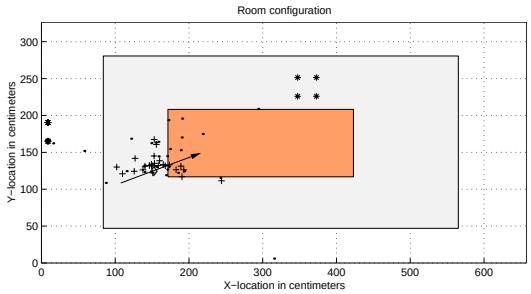
$$\boldsymbol{\xi}_0 = \begin{bmatrix} P_{ss}(\omega_{k_l}) & \cdots & P_{ss}(\omega_{k_u}) & \sigma^2 & \tau_0 \end{bmatrix}^T, \tag{71}$$

and collect all measurements in the vector $\mathbf{X} = \begin{bmatrix} \mathbf{X}(\omega_{k_l})^T & \cdots & \mathbf{X}(\omega_{k_u})^T \end{bmatrix}^T$. The likelihood function can then be written as

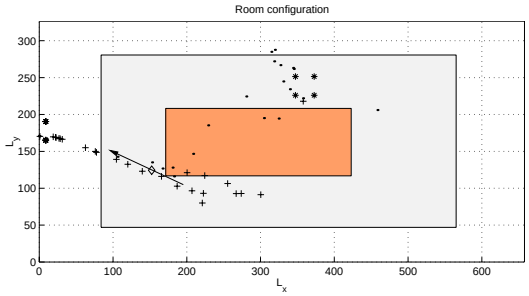
$$l\left(\mathbf{X};\boldsymbol{\xi}_0\right) = \frac{1}{\pi^{mN}\,\det\{\mathbf{R}_{XX}(\boldsymbol{\xi}_0)\}} \exp\left\{-\mathbf{X}^H \mathbf{R}_{XX}(\boldsymbol{\xi}_0)^{-1}\mathbf{X}\right\} \tag{72}$$

where $\det\{\cdot\}$ denotes the determinant. Furthermore, the $MN \times MN$ matrix $\mathbf{R}_{XX}(\boldsymbol{\xi}_0)$ is block-diagonal due to the assumed whiteness of the sequence $\{S(\omega_k)\}_{k=k_l}^{k_u}$: $\mathbf{R}_{XX}(\boldsymbol{\xi}_0) = \mathrm{diag}\{\mathbf{R}_{X_{k_l}X_{k_l}}(\boldsymbol{\xi}_0), \cdots, \mathbf{R}_{X_{k_u}X_{k_u}}(\boldsymbol{\xi}_0)\}$. The covariance matrix of any unbiased estimator $\hat{\boldsymbol{\xi}}$ fulfills

$$E_{\boldsymbol{\theta},S}\left\{\left(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}_0\right)\left(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}_0\right)^T\right\} \geq \mathbf{F}(\boldsymbol{\xi}_0)^{-1} \tag{73}$$



(a) Estimated source locations.



(b) Estimated source locations.

Fig. 6. Experimental results using real data. Here "+" denotes the robust method, and "·" denotes the original unweighted method. The included arrow illustrates the orientation of the speaker.

where $\mathbf{F}(\boldsymbol{\xi}_0)$ denotes the Fisher information matrix. For notational simplicity, argument $(\boldsymbol{\xi}_0)$ is omitted in the following. The $(i,j)^{th}$ element of $\mathbf{F}$ can be evaluated as, see for example [11, Chapter 15],

$$(\mathbf{F})_{i,j} = \mathrm{Tr}\left\{ \mathbf{R}_{\mathrm{XX}}^{-1} \frac{\partial \mathbf{R}_{\mathrm{XX}}}{\partial \boldsymbol{\xi}_{0_{\mathrm{i}}}} \mathbf{R}_{\mathrm{XX}}^{-1} \frac{\partial \mathbf{R}_{\mathrm{XX}}}{\partial \boldsymbol{\xi}_{0_{\mathrm{j}}}} \right\}. \tag{74}$$

To simplify the notation, partition $\mathbf{F}$ in agreement with the following partitioning of $\boldsymbol{\xi}_0 = [\boldsymbol{\eta}^T \ \sigma^2 \ \tau_0]^T$ (where $\boldsymbol{\eta}^T = [P_{ss}(\omega_{k_l}) \ \cdots \ P_{ss}(\omega_{k_u})]$):

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}^{11} & \mathbf{F}^{12} & \mathbf{F}^{13} \\ (\mathbf{F}^{12})^T & \mathbf{F}^{22} & \mathbf{F}^{23} \\ (\mathbf{F}^{13})^T & (\mathbf{F}^{23})^T & \mathbf{F}^{33} \end{bmatrix}. \tag{75}$$

The different elements of $\mathbf{F}$ are evaluated in the following:

- $\mathbf{F}^{11}$: Computing the derivative of $\mathbf{R}_{XX}$ with respect to $\boldsymbol{\eta}_i$ and $\boldsymbol{\eta}_j$, we find that

$$(\mathbf{F}^{11})_{ij} = \mathrm{Tr}\left\{ \mathbf{R}_{\mathrm{X_iX_i}}^{-1} \left(\sigma^2 \mathbf{I}_{\mathrm{M}} + \mathbf{A}(\omega_{\mathrm{i}};\tau_0)\mathbf{A}(\omega_{\mathrm{i}};\tau_0)^{\mathrm{H}}\right) \mathbf{R}_{\mathrm{X_jX_j}}^{-1} \left(\sigma^2 \mathbf{I}_{\mathrm{M}} + \mathbf{A}(\omega_{\mathrm{j}};\tau_0)\mathbf{A}(\omega_{\mathrm{j}};\tau_0)^{\mathrm{H}}\right) \right\} \delta_{\mathrm{i,j}} \tag{76}$$

where $\delta_{i,j}$ denotes Kronecker's delta function. Using the matrix inversion lemma, it follows that

$$\mathbf{R}_{X_iX_i}^{-1} = \frac{1}{\sigma^2 P_{ss}(\omega_i)} \left(\mathbf{I}_M - \frac{\mathbf{A}(\omega_i;\tau_0)\mathbf{A}(\omega_i;\tau_0)^H}{\sigma^2 + M} \right), \tag{77}$$

where we used the fact that $\mathbf{A}(\omega_i;\tau_0)^H \mathbf{A}(\omega_i;\tau_0) = M$. The expression for $(\mathbf{F}^{11})_{ij}$ then simplifies to

$$(\mathbf{F}^{11})_{ij} = \frac{M}{P_{ss}^2(\omega_i)} \delta_{i,j}, \tag{78}$$

and finally

$$\mathbf{F}^{11} = \mathrm{diag}\left\{ \frac{M}{P_{ss}^2(\omega_{k_l})}, \cdots, \frac{M}{P_{ss}^2(\omega_{k_u})} \right\}. \tag{79}$$

- $\mathbf{F}^{22}$: Since the derivative of $\mathbf{R}_{X_iX_i}$ with respect to σ^2 equals $P_{ss}(\omega_i)\mathbf{I}_M$,

$$\mathbf{F}^{22} = \sum_{k=k_l}^{k_u} P_{ss}^2(\omega_i) \mathrm{Tr}\left\{ \mathbf{R}_{\mathrm{X_iX_i}}^{-2} \right\}. \tag{80}$$

If we once again apply expression (77), straightforward calculations result in the following expression:

$$\mathbf{F}^{22} = \frac{MN}{\sigma^4} \left(1 - \frac{2}{\sigma^2 + M} + \frac{M}{(\sigma^2 + M)^2} \right). \tag{81}$$

- $\mathbf{F}^{33}$: Next we compute the derivative of $\mathbf{R}_{X_iX_i}$ with respect to the time-delay τ_0. We then find that

$$\frac{\partial}{\partial \tau_0} \mathbf{R}_{X_iX_i} = P_{ss}(\omega_i) \mathbf{D}_i, \tag{82}$$

where

$$\mathbf{D}_i \stackrel{\triangle}{=} \frac{\partial}{\partial \tau_0} \mathbf{A}(\omega_i;\tau_0)\mathbf{A}(\omega_i;\tau_0)^H = \mathbf{A}(\omega_i;\tau_0)\mathbf{D}^H(\omega_i;\tau_0) + \mathbf{D}(\omega_i;\tau_0)\mathbf{A}^H(\omega_i;\tau_0) \tag{83}$$

$$\mathbf{D}(\omega_i;\tau_0) = \frac{\partial \mathbf{A}(\omega_i;\tau_0)}{\partial \tau_0}. \tag{84}$$

It is then easy to see that

$$\mathbf{F}^{33} = \frac{1}{\sigma^4} \sum_{k=k_l}^{k_u} \mathrm{Tr}\left\{ \mathbf{P}_{\mathrm{k}} \mathbf{D}_{\mathrm{k}} \mathbf{P}_{\mathrm{k}} \mathbf{D}_{\mathrm{k}} \right\}, \tag{85}$$

where

$$\mathbf{P}_k \stackrel{\triangle}{=} \mathbf{I}_M - \frac{\mathbf{A}(\omega_k;\tau_0)\mathbf{A}^H(\omega_k;\tau_0)}{\sigma^2 + M}. \tag{86}$$

- Along the same lines, it can be shown that

$$\mathbf{F}^{12} = \frac{M}{\sigma^2} \left(1 - \frac{1}{\sigma^2 + M} \right) \left[\frac{1}{P_{ss}(\omega_{k_l})}, \cdots, \frac{1}{P_{ss}(\omega_{k_u})} \right]^T \tag{87}$$

$$\mathbf{F}^{13} = \frac{1}{\sigma^2} \left[\mathrm{Tr}\{\mathbf{P}_{\mathrm{k_l}} \mathbf{D}_{\mathrm{k_l}}\}, \cdots, \mathrm{Tr}\{\mathbf{P}_{\mathrm{k_u}} \mathbf{D}_{\mathrm{k_u}}\} \right]^T \tag{88}$$

$$\mathbf{F}^{23} = \frac{1}{\sigma^4} \sum_{k=k_l}^{k_u} \mathrm{Tr}\left\{ \mathbf{P}_{\mathrm{k}} \mathbf{D}_{\mathrm{k}} \mathbf{P}_{\mathrm{k}} \right\}. \tag{89}$$

To simplify the CRB expression, we next note that $\mathbf{F}^{13} = \mathbf{0}$. This follows since

$$\begin{aligned} \mathrm{Tr}\{\mathbf{P}_{\mathrm{i}}\mathbf{D}_{\mathrm{i}}\} &= \frac{\sigma^2}{\sigma^2+M}\left(\mathrm{Tr}\{\mathbf{A}(\omega_{\mathrm{i}};\tau_0)\mathbf{D}(\omega_{\mathrm{i}};\tau_0)^{\mathrm{H}}\} + \mathrm{Tr}\{\mathbf{D}(\omega_{\mathrm{i}};\tau_0)\mathbf{A}(\omega_{\mathrm{i}};\tau_0)^{\mathrm{H}}\}\right) \quad (90)\\ &= \frac{\sigma^2}{\sigma^2+M}\left(\left(j\omega_i\sum_{l=1}^{M-1} l\right) + \left(j\omega_i\sum_{l=1}^{M-1} l\right)^*\right) = 0. \quad (91)\end{aligned}$$

Similarly it can be shown that $\mathbf{F}^{23} = \mathbf{0}$. Hence, the expression for $\mathbf{F}$ reads as

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}^{11} & \mathbf{F}^{12} & \mathbf{0} \\ (\mathbf{F}^{12})^T & \mathbf{F}^{22} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{F}^{33} \end{bmatrix}, \quad (92)$$

and we finally find that

$$E_{\boldsymbol{\theta},S}\left\{(\hat{\tau}-\tau_0)^2\right\} \geq (\mathbf{F}^{33})^{-1}. \quad (93)$$

II. Proof of Proposition 2

For large N, the estimated cross-correlation can be written as

$$\hat{R}(nT_s) \simeq E\{x_1(t)x_2(t-\tau);\boldsymbol{\theta}\} = \frac{1}{N}\sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} P_{12}(\omega_k;\boldsymbol{\theta})e^{j\omega_k nT_s}, \quad (94)$$

where

$$P_{12}(\omega_k;\boldsymbol{\theta}) = R_{ss}\left(e^{-j\omega_k n_0 T_s} + R_2(\omega_k;\boldsymbol{\theta}) + R_1^*(\omega_k;\boldsymbol{\theta})e^{-j\omega_k n_0 T_s} + R_2(\omega_k;\boldsymbol{\theta})R_1^*(\omega_k;\boldsymbol{\theta})\right), \quad (95)$$

denotes the cross-power spectrum, conditioned on $\boldsymbol{\theta}$. Here, $\mathbf{R}(\omega_k;\boldsymbol{\theta}) = [R_1(\omega_k;\boldsymbol{\theta})\ R_2(\omega_k;\boldsymbol{\theta})]^T$. Note, when computing $\hat{R}(nT_s)$, also frequencies that does not satisfy the Schroeder large room frequency are included in the summation (94). We will however neglect this fact in the following.

We next proceed to determine the statistical properties of $\hat{R}(nT_s)$. For that purpose, it is useful to first determine the properties of $P_{12}(\omega_k;\boldsymbol{\theta})$. Evaluating the expected value of $P_{12}(\omega_k;\boldsymbol{\theta})$, we find that

$$E_{\boldsymbol{\theta}}\{P_{12}(\omega_k;\boldsymbol{\theta})\} = R_{ss}e^{-j\omega_k n_0 T_s}. \quad (96)$$

We can now show that the expected value of $\hat{R}(nT_s)$ equals

$$E_{\boldsymbol{\theta}}\left\{\hat{R}(nT_s)\right\} = \begin{cases} R_{ss} & \text{for } n = n_0 \\ 0 & \text{for } n \neq n_0 \end{cases}. \quad (97)$$

To see this, note that

$$E_{\boldsymbol{\theta}}\left\{\hat{R}(nT_s)\right\} = \frac{R_{ss}}{N}\sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} e^{j(n-n_0)\omega_k T_s} = \frac{R_{ss}}{N}\sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} e^{j(n-n_0)\frac{2\pi k}{N}}. \quad (98)$$

Result (97) then follows from the orthogonality properties of complex sinusoids.

Next we would like to compute the variance of $\hat{R}(nT_s)$. Keeping in mind that $\mathrm{Var}_{\boldsymbol{\theta}}\{\hat{R}(nT_s)\} = E_{\boldsymbol{\theta}}\{\hat{R}^2(nT_s)\} - E^2_{\boldsymbol{\theta}}\{\hat{R}(nT_s)\}$, we study

$$\begin{aligned} E_{\boldsymbol{\theta}}\left\{\hat{R}^2(nT_s)\right\} &= \frac{1}{N^2}\sum_{k,l} E_{\boldsymbol{\theta}}\left\{P_{12}(\omega_k;\boldsymbol{\theta})P^*_{12}(\omega_l;\boldsymbol{\theta})\right\} e^{j(\omega_k-\omega_l)nT_s} \\ &= \frac{1}{N^2}\sum_{k\neq l} E_{\boldsymbol{\theta}}\left\{P_{12}(\omega_k;\boldsymbol{\theta})P^*_{12}(\omega_l;\boldsymbol{\theta})\right\} e^{j(\omega_k-\omega_l)nT_s} + \frac{1}{N^2}\sum_{k=l} E_{\boldsymbol{\theta}}\left\{P_{12}(\omega_k;\boldsymbol{\theta})P^*_{12}(\omega_k;\boldsymbol{\theta})\right\}. \end{aligned} \tag{99}$$

For $\omega_k \neq \omega_l$ we find that

$$\begin{aligned} &E_{\boldsymbol{\theta}}\left\{P_{12}(\omega_k;\boldsymbol{\theta})P^*_{12}(\omega_l;\boldsymbol{\theta})\right\} - E_{\boldsymbol{\theta}}\left\{\hat{P}_{12}(\omega_k)\right\} E_{\boldsymbol{\theta},S}\left\{\hat{P}_{12}(\omega_l)^*\right\} \\ &= R^2_{ss} E_{\boldsymbol{\theta}}\left\{\left(e^{-j\omega_k n_0 T_s} + R_2(\omega_k;\boldsymbol{\theta}) + R^*_1(\omega_k;\boldsymbol{\theta})e^{-j\omega_k n_0 T_s} + R_2(\omega_k;\boldsymbol{\theta})R^*_1(\omega_k;\boldsymbol{\theta})\right)\right. \\ &\quad \left.\left(e^{-j\omega_l n_0 T_s} + R_2(\omega_l;\boldsymbol{\theta}) + R^*_1(\omega_l;\boldsymbol{\theta})e^{-j\omega_l n_0 T_s} + R_2(\omega_l;\boldsymbol{\theta})R^*_1(\omega_l;\boldsymbol{\theta})\right)^*\right\} - R^2_{ss}e^{jn_0T_s(\omega_l-\omega_k)}. \end{aligned} \tag{100}$$

Defining the quantity (for explicit expressions, see [7])

$$\Psi\left(\omega_k-\omega_l\right) \stackrel{\triangle}{=} E_{\boldsymbol{\theta}}\left\{R_1(\omega_k;\boldsymbol{\theta})R_1(\omega_l;\boldsymbol{\theta})^*\right\} = E_{\boldsymbol{\theta}}\left\{R_2(\omega_k;\boldsymbol{\theta})R_2(\omega_l;\boldsymbol{\theta})^*\right\}, \tag{101}$$

equation (100) can be written as

$$\begin{aligned} &E_{\boldsymbol{\theta},S}\left\{\hat{P}_{12}(\omega_k)\hat{P}^*_{12}(\omega_l)\right\} - E_{\boldsymbol{\theta},S}\left\{\hat{P}_{12}(\omega_k)\right\} E_{\boldsymbol{\theta},S}\left\{\hat{P}_{12}(\omega_l)^*\right\} \\ &\quad = R^2_{ss}\left(\Psi\left(\omega_k-\omega_l\right) + \Psi\left(\omega_k-\omega_l\right)e^{jn_0T_s(\omega_l-\omega_k)} + \Psi\left(\omega_k-\omega_l\right)^2\right). \end{aligned} \tag{102}$$

The technical complication in the analysis is that terms like $\Psi(\omega_k-\omega_l) \neq 0$ due to the frequency correlation of $\mathbf{R}(\omega;\boldsymbol{\theta})$. The analysis can in principle be carried out using the function $\Psi\left(\omega_k-\omega_l\right)$. We choose however a simpler, but approximative, approach. The main observation is that $\hat{R}(nT_s)$ for large N approximately can be written as

$$\hat{R}(nT_s) \simeq \frac{1}{\tilde{N}}\sum_{k=-\frac{\tilde{N}}{2}}^{\frac{\tilde{N}}{2}-1} P_{12}(\phi_k;\boldsymbol{\theta})e^{j\phi_k nT_s}, \tag{103}$$

where we have re-sampled the frequency axis according to the coherence bandwidth $\rho(T_{60})$ of $\mathbf{R}(\omega;\boldsymbol{\theta})$:

$$\phi_k = \frac{2\pi F_s}{\tilde{N}},\ k = -\frac{\tilde{N}}{2},\cdots,\frac{\tilde{N}}{2}-1 \tag{104}$$

$$\tilde{N} = \frac{F_s}{\rho(T_{60})}. \tag{105}$$

The approximation (103) is based on the fact that $P_{12}(\omega_k;\boldsymbol{\theta})$ and $P_{12}(\omega_l;\boldsymbol{\theta})$ are highly correlated for a small frequency separation $|\omega_k-\omega_l|$. In the following it will be assumed that $\Psi(\cdot)$ is negligible for $|\omega_k-\omega_l| \geq 2\pi\rho(T_{60})$. We then find that

$$E_{\boldsymbol{\theta}}\left\{P_{12}(\phi_k;\boldsymbol{\theta})P^*_{12}(\phi_l;\boldsymbol{\theta})\right\} = \begin{cases} R^2_{ss}\left(1+\frac{1}{SRR}\right)^2 & \text{for } \phi_k = \phi_l \\ R^2_{ss}e^{jn_0T_s(\phi_l-\phi_k)} & \text{for } \phi_k \neq \phi_l \end{cases}, \tag{106}$$

and

$$E_{\boldsymbol{\theta}}\left\{P_{12}(\phi_k;\boldsymbol{\theta})P_{12}^*(\phi_l;\boldsymbol{\theta})\right\} - E_{\boldsymbol{\theta}}\left\{P_{12}(\phi_k;\boldsymbol{\theta})\right\}E_{\boldsymbol{\theta}}\left\{P_{12}(\phi_l;\boldsymbol{\theta})^*\right\} = 0 \text{ for } \phi_k \neq \phi_l. \tag{107}$$

Returning to equation (99), we find that

$$\begin{aligned} E_{\boldsymbol{\theta}}\left\{\hat{R}^2(nT_s)\right\} &\simeq \frac{R_{ss}^2}{\tilde{N}^2}\sum_{k\neq l} e^{j(\phi_k-\phi_l)(n-n_0)T_s} + \frac{1}{\tilde{N}}R_{ss}^2\left(1+\frac{1}{SRR}\right)^2 \\ &= R_{ss}^2\frac{\tilde{N}-1}{\tilde{N}}\delta_{n-n_0} - \frac{R_{ss}^2}{\tilde{N}}(1-\delta_{n-n_0}) + \frac{1}{\tilde{N}}R_{ss}^2\left(1+\frac{1}{SRR}\right)^2 \\ &= R_{ss}^2\delta_{n-n_0} + \frac{1}{\tilde{N}}R_{ss}^2\left(1+\frac{1}{SRR}\right)^2 - \frac{R_{ss}^2}{\tilde{N}}. \end{aligned} \tag{108}$$

Since $E_{\boldsymbol{\theta}}\{\hat{R}(nT_s)\} = R_{ss}\delta_{n-n_0}$, it consequently follows that

$$\mathrm{Var}_{\boldsymbol{\theta}}\left\{\hat{R}(nT_s)\right\} \simeq \frac{1}{\tilde{N}}R_{ss}^2\left(1+\frac{1}{SRR}\right)^2 - \frac{R_{ss}^2}{\tilde{N}} = \frac{R_{ss}^2}{\tilde{N}}\left(\frac{2}{SRR}+\frac{1}{SRR^2}\right). \tag{109}$$

The proof is complete if we can establish that $\hat{R}(n_1T_s)$ and $\hat{R}(n_2T_s)$ are independent for $n_1 \neq n_2$, and that $\hat{R}(nT_s)$ is Gaussian for all n. The Gaussianity follows from the central limit theorem, since we assume $\tilde{N} \gg m$. Finally, we would like to establish that $\hat{R}(n_1T_s)$ and $\hat{R}(n_2T_s)$ are uncorrelated. Details are omitted, but calculations completely analogous to the previous ones (and invoking the approximation (103), show that for $n_1 \neq n_2$

$$E_{\boldsymbol{\theta}}\left\{\left(\hat{R}(n_1T_s) - E_{\boldsymbol{\theta}}\left\{\hat{R}(n_1T_s)\right\}\right)\left(\hat{R}(n_2T_s) - E_{\boldsymbol{\theta}}\left\{\hat{R}(n_2T_s)\right\}\right)\right\} \simeq 0, \tag{110}$$

as required.

References

[1] J.B. Allen and A. Berkeley. "Image Method for Efficiently Simulating Small-room Acoustics". *Journal of the Acoustical Society of America*, 65(4):943–950, April 1979.

[2] M. S. Brandstein, J. E. Adcock, and H. F. Silverman. "A Closed-Form Estimator for use with Room Environment Microphone Arrays". *IEEE Trans. on Speech and Audio Processing*, 5(1):45–50, January 1997.

[3] M.S. Brandstein. "A Pitch-based Approach to Time-delay Estimation of Reverberant Speech". In *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 1997.

[4] M.S Brandstein and H.F. Silverman. "A Practical Methodology for Speech Source Localization with Microphone Arrays". *Computer, Speech, and Language*, 11(2):91–126, April 1997.

[5] B. Champagne, S. Bédard, and A. Stéphenne. "Performance of Time-delay Estimation in the Presence of Room Reverberation". *IEEE Trans. on Speech and Audio Processing*, 4(2):148–152, March 1996.

[6] M.A. Doron, A.J. Weiss, and H. Messer. "Maximum-Likelihood Direction Finding of Wide-Band Sources". *IEEE Trans. on Signal Processing*, 1(2):411–417, January 1993.

[7] T. Gustafsson and B.D. Rao. "Source Localization in Reverberant Environments: Modeling". Submitted to IEEE Trans. on Speech and Audio Processing for possible publication, 2000.

[8] J.P. Ianniello. "Time Delay Estimation Via Cross-Correlation in the Presence of Large Estimation Errors". *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 30(6):998–1003, December 1982.

[9] J.P. Ianniello. "High-resolution Multipath Time Delay Estimation for Broadband Random Signals". *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 36(3):320–327, March 1988.

[10] E.E. Jan and J. Flanagan. "Sound Source Localization in Reverberant Environments using an Outlier Elimination Algorithm". In *Proc. of ICSLP*, pages 1321–4, Philadelphia, USA, 1996.

[11] S. M. Kay. *Fundamentals of Statistical Signal Processing.* Prentice-Hall, Englewood Cliffs, NJ, 1993.

[12] C.H. Knapp and G.C. Carter. "The Generalized Correlation Method for Estimation of Time Delay". *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 24(4):320–326, August 1976.

[13] H. Kuttruff. *Room Acoustics.* John Wiley, 1973.

[14] L. Ljung. *System Identification: Theory for the User.* Prentice-Hall, Englewood Cliffs, NJ, 1987.

[15] M. Omologo and P. Svaizer. "Acoustic Source Location in Noisy and Reverberant Environment Using CSP Analysis". In *Proc. ICASSP*, pages 921–924, Atlanta, USA, 1996.

[16] M. Omologo and P. Svaizer. "Use of the Crosspower-Spectrum Phase in Acoustic Event Localization". *IEEE Trans. on Speech and Audio Processing*, 5(3):288–292, May 1997.

[17] H. C. Schau and A. Z. Robinson. "Passive Source Localization Employing Intersecting Spherical Surfaces from Time-of-Arrival Differences". *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 35(8):1223–1225, August 1987.

[18] J.O. Smith and J.S. Abel. "Closed-Form Least-Squares Source Location Estimation from Range-Difference Measurements". *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 35(12):1661–1669, December 1987.

[19] P. Stoica and A. Nehorai. "Performance Study of Conditional and Unconditional Direction-of-arrival Estimation". *IEEE Trans. on Acoustics, Speech and Signal Processing*, 38(10):1783–95, Oct. 1990.