Source Localization in Reverberant Environments: Part I - Modeling

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

SP-EDICS: 2-ROOM

Abstract

Source localization using microphone arrays has for some time been an active and challenging research area. The main obstacle in designing robust practical systems is the effects of room reverberation. This paper, along with a companion, analyzes the influence of reverberation. In this paper, a new statistical model is developed that explains the effects of reverberation. The model is based on the theory of statistical room acoustics and is a generalization of the Ricean model used to model channels in digital communication. The properties of the statistical model is evaluated with both simulations using the image method, and with real data measurements. The agreements are quite good lending support to the model. The model is further validated in the companion paper where it is used to understand performance limits of source localization estimation accuracy in reverberant environments, and to understand the performance of source localization algorithms.

I. Introduction

Development of microphone array signal processing systems has been an active area of research. Microphone arrays have shown promise in acoustical source localization and speech acquisition. An example of this is

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.

interference suppression using beamforming. By steering the microphone array in certain look directions, interferers can be suppressed. However, the success of beamforming for interference suppression hinges on the knowledge of the location of the desired speaker. Knowledge about the location of the source is also crucial for automatic speaker tracking in video-conferencing [12], [13].

In the literature, several approaches have appeared to estimate the location of a speaker using microphone arrays. Among existing methods, those which are based on time-delay estimation (TDE) have gained a lot of attention, see for example [3], [2], [7], [10]. In this context, the received signals from a pair of microphones are modeled as

$$\begin{aligned} x_1(t) &= s(t) + n_1(t) \\ x_2(t) &= s(t - \tau_0) + n_2(t), \end{aligned} \tag{1}$$

where $x_i(t)$ $(i = 1, 2)$ is the output signal of the i^{th} receiver, $s(t)$ is the unknown source signal, $n_i(t)$ is an additive noise term assumed uncorrelated with $s(t)$, and τ_0 is the unknown time-delay. Assuming there are several such microphone pairs distributed over the spatial region, the location of the source can be estimated from triangulation of the estimated time-delays.

Considering the problem of acoustical source localization in, for example, an office environment, the single-path propagation model (1) is not realistic. In an office environment, the accuracy of the estimated time-delay is typically not limited by additive measurement noise as in (1). Instead, the accuracy is limited by room reverberation, and much less is known about TDE in reverberant environments. In a reverberant environment, the measured microphone signals could be modeled as

$$\begin{aligned} x_1(t) &= \int_{-\infty}^{\infty} h_1(t - \lambda) s(\lambda) d\lambda + n_1(t) \\ x_2(t) &= \int_{-\infty}^{\infty} h_2(t - \lambda) s(\lambda) d\lambda + n_2(t), \end{aligned} \tag{2}$$

where $h_i(t)$ represents the impulse response of the acoustical transfer function from the source to the i^{th} microphone. Now the time-delay τ_0 is "hidden" in the impulse responses $h_1(t)$ and $h_2(t)$. See for example [6, Chapter 5] for a thorough treatment of the reverberation phenomenon.

The empirical experience is that once the level of room reverberation rises above minimal levels, most

methods for TDE begin to exhibit dramatic performance degradations and become quite unreliable. The purpose of the present paper is to propose a *statistical* model of the impulse responses $h_i(t)$ which can serve as the basis for understanding accuracy limits of source localization in reverberant environments and for the development of robust algorithms. The model is derived using well-established theory of diffuse sound fields, and is inspired by similar concerns in digital mobile communication where so-called fading is a well-researched problem (see for example [8]). The introduced reverberation model is in fact quite similar to "Ricean fading". However, because of the broadband nature of the signal, the signal is divided into subbands and model is suitably adapted for each band. In a companion paper [4], the new model is applied to derive relevant Cramér-Rao lower bounds to understand localization accuracy limits. In addition, the accuracy of common time-delay estimators are analyzed, and the probability of an anomalous estimate is derived.

II. Elements of the New Model

In this section we will analyze the elements of the model that will be used to model the effects of reverberation. The model is built on some important results in room acoustics, and so the relevant results are reviewed and summarized in this context. We will high-light crucial assumptions, and make a couple of novel approximations to arrive at a simple but useful model.

A. General Assumptions

For simplicity, consider the case with two microphones. Using matrix notation, we can compactly write (2) as

$$\mathbf{x}(t) = \int_{-\infty}^{\infty} \mathbf{h}(t-\lambda)s(\lambda)d\lambda + \mathbf{n}(t), \tag{3}$$

where $\mathbf{x}(t) = [x_1(t)\ x_2(t)]^T$, $\mathbf{h}(t) = [h_1(t)\ h_2(t)]^T$ and $\mathbf{n}(t) = [n_1(t)\ n_2(t)]^T$. The impulse response $\mathbf{h}(t)$ is assumed to represent a linear and time-invariant (LTI) system, as indicated by the convolution representation (3). Considering common models of room transfer functions, such as wave-equation based modeling [6], the linearity assumption does not appear to be restrictive. Assuming time-invariance simply means that the positions of the microphones and the source are assumed stationary.

Since the main goal is to provide understanding of the reverberation phenomenon, we will neglect $\mathbf{n}(t)$ in the analysis. The dominating noise-sources in a typical office environment are:

1. Additive measurement noise due to the measuring equipment.

2. Interferers such as air-conditioning equipment and/or other speakers.

In a high-quality measurement equipment, the additive measurement noise is often negligible compared to the level of reverberation. Furthermore, most of the noise from an air-conditioning device in general has most of its power at frequencies below the frequencies where speech is dominant (i.e. 300-3500 Hz). Hence, neglecting $\mathbf{n}(t)$ is not too restrictive, assuming that other speakers are not present.

Assume next that the source is located at $\mathbf{r}_s$, and that the microphones are located at $\mathbf{r}_{m_1}$ and $\mathbf{r}_{m_2}$, respectively. The distance between the microphones is denoted as $d = \|\mathbf{r}_{m_1} - \mathbf{r}_{m_2}\|$. Our next assumption is the key assumption. We will in the rest of the paper assume that $\mathbf{h}(t)$ can be written as:

$$\mathbf{h}(t) = \mathbf{h}_d(t) + \mathbf{h}_r(t). \tag{4}$$

Here, subscript $(\cdot)_d$ denotes the direct-path propagation, and subscript $(\cdot)_r$ denotes the reverberant part. The direct-path impulse response $\mathbf{h}_d(t)$ contains the impulse response corresponding to free-space propagation of an omni-directional acoustical point source. We thus assume the following form of $\mathbf{h}_d(t)$:

$$\mathbf{h}_d(t) = \begin{bmatrix} 1 \\ \\ \delta(t-\tau_0) \end{bmatrix} \kappa(r) \tag{5}$$

where $\delta(\cdot)$ denotes Dirac's delta function and τ_0 denotes the difference in propagation times

$$\tau_0 = \frac{1}{c}\left(\|\mathbf{r}_s - \mathbf{r}_{m_1}\| - \|\mathbf{r}_s - \mathbf{r}_{m_2}\|\right). \tag{6}$$

Here, the speed of sound propagation is denoted as c (generally specified as $c = 344$ m/s at 21° C). Furthermore, it is assumed that

$$\kappa(r) \simeq \sqrt{\frac{1}{4\pi r^2}}, \tag{7}$$

where r is the distance from the source to the point in-between the two microphones, i.e.

$$r \triangleq \left\| \mathbf{r}_s - \frac{1}{2}\left(\mathbf{r}_{m_1} + \mathbf{r}_{m_2}\right) \right\|. \tag{8}$$

Here we used the well-known fact that the power from an omni-directional point source decays as $1/4\pi r^2$. The above model of $\mathbf{h}_d(t)$ is a simplification of the actual scenario since we assume that both microphones receive equal amount of power from the direct-path propagation. Hence, we can only claim that the assumed form of $\mathbf{h}_d(t)$ is accurate in cases where $d \ll \|\mathbf{r}_s - \mathbf{r}_{m_1}\|$, $\|\mathbf{r}_s - \mathbf{r}_{m_2}\|$.

Next we proceed to establish the properties of $\mathbf{h}_r(t)$. We will assume that $\mathbf{h}_r(t)$ is due to *diffuse* sound, i.e. the sound energy has no direction associated with it and is received uniformly from all directions. In the following section we will review the properties of diffuse sound, and at the same time establish the *statistical* properties of $\mathbf{h}_r(t)$.

Remark 1: The assumption that $\mathbf{h}(t)$ can be split into a direct-path propagation $\mathbf{h}_d(t)$, and into a reverberant part $\mathbf{h}_r(t)$ is not entirely new. In a recent paper by Radlović et al. [9], a similar model was successfully used to investigate the robustness of equalization techniques in a reverberant environment. In [9], however, only point-to-point equalization was considered, and correlation between $h_1(t)$ and $h_2(t)$ was not taken into account.

B. Properties of $\mathbf{h}_r(t)$

The sound pressure at the microphone can be considered as being built up of a direct-path, plus several plane waves due to multiple reflections of the original sound from the walls. These reflections travel in different directions and encounter the walls at different angles of incidence. In the time domain, these reflections are perceived as delayed echoes with more or less random amplitudes.

The large number of echoes implies that the measured sound pressure can be quite different for different microphone locations. Studying an empty rectangular room, $\mathbf{h}(t)$ can be computed by solving the wave equation, i.e. the theory of *modal analysis*, see for example [6, Chapter 3]. This kind of "deterministic" modeling was recently applied for equalization of room transfer functions, see [5].

At higher frequencies, the complexity of modal analysis however increases to a point where exact mathematical analysis is no longer feasible. This "breakdown" in modal analysis is due to the large number of modes being excited by the source. A suitable tool for analyzing the behavior of room transfer functions for high

frequencies is to apply the theory of random (or diffuse) sound fields [6, Chapter 5]. The theory of statistical room acoustics closely describes the actual behavior if the following conditions are fulfilled [11]:

A1: The dimensions of the room are large relative to the wavelength of $s(t)$. For the frequencies of interest (for speech we are mainly interested in the band 300-3500 Hz), this condition is usually satisfied.

A2: The average spacing of the resonance frequencies of the room must be smaller than one third of their bandwidth. In a room with volume V (in m^3), and reverberation time[1] T_{60} (in seconds), this condition is fulfilled for all frequencies that exceed the "Schroeder large room frequency":

$$f_S = 2000\sqrt{\frac{T_{60}}{V}}. \tag{9}$$

For instance, in a "normal" office with reverberation time of $1s$ and volume $100m^3$, the statistical theory would be relevant for all frequencies above about 200 Hz, i.e. for all frequencies where speech energy is present!

A3: Both the source and the microphones are located in the interior of the room, at least a half-wavelength away from the walls.

Under the above conditions, the transfer function between the source and the microphone (excluding the direct-path) can accurately be modeled as a *random function*.

Define now the frequency response of $\mathbf{h}_r(t)$ as

$$\mathbf{H}_r(\omega) = \int_0^\infty \mathbf{h}_r(t)e^{-j\omega t}dt. \tag{10}$$

It is important to note that the randomness of the transfer function $\mathbf{H}_r(\omega)$ is not related to absolute time. That is, given a fixed source location and a fixed microphone location, $\mathbf{H}_r(\omega)$ will remain constant unless the room configuration or the positions of the microphones or the source, somehow are altered. To emphasize this fact, we define the vector $\boldsymbol{\theta} = [\mathbf{r}_s^T\ \mathbf{r}_{m_1}^T\ \mathbf{r}_{m_2}^T]^T$, and denote the *random* impulse response and its frequency response accordingly: $\mathbf{h}_r(t;\boldsymbol{\theta}) \leftrightarrow \mathbf{H}_r(\omega;\boldsymbol{\theta})$. Conditioned on a fixed $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$, $\mathbf{h}_r(t;\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}})$ is hence assumed to be a deterministic function.

[1]The reverberation time T_{60} is defined as the length of time for the sound intensity level in a room to decrease by 60 dB after the sound source is shut off.

Consider the k^{th} component of $\mathbf{H}_r(\omega;\boldsymbol{\theta})$, and write it as

$$H_{r_k}(\omega;\boldsymbol{\theta}) = H^r_{r_k}(\omega;\boldsymbol{\theta}) + jH^i_{r_k}(\omega;\boldsymbol{\theta}). \tag{11}$$

Here $H^r_{r_k}(\omega;\boldsymbol{\theta})$ and $H^i_{r_k}(\omega;\boldsymbol{\theta})$ are the real and imaginary parts of $H_{r_k}(\omega;\boldsymbol{\theta})$, respectively. We next cite a couple of interesting result from [11] (assuming *A1-A3* to be fulfilled):

$$\begin{aligned}
E_{\boldsymbol{\theta}}\left\{\mathbf{H}_r(\omega;\boldsymbol{\theta})\right\} &= \mathbf{0} && (12)\\
\varphi_{rr}(\Delta\omega) &\triangleq \frac{E_{\boldsymbol{\theta}}\left\{H^r_{r_k}(\omega;\boldsymbol{\theta})H^r_{r_k}(\omega+\Delta\omega;\boldsymbol{\theta})\right\}}{\sqrt{E_{\boldsymbol{\theta}}\left\{(H^r_{r_k}(\omega;\boldsymbol{\theta}))^2\right\}E_{\boldsymbol{\theta}}\left\{(H^r_{r_k}(\omega+\Delta\omega;\boldsymbol{\theta}))^2\right\}}}\\
&= \frac{1}{1+\left(\Delta\omega\frac{T_{60}}{13.8}\right)^2}, && (13)\\
\varphi_{ii}(\Delta\omega) &\triangleq \frac{E_{\boldsymbol{\theta}}\left\{H^i_{r_k}(\omega;\boldsymbol{\theta})H^i_{r_k}(\omega+\Delta\omega;\boldsymbol{\theta})\right\}}{\sqrt{E_{\boldsymbol{\theta}}\left\{(H^i_{r_k}(\omega;\boldsymbol{\theta}))^2\right\}E_{\boldsymbol{\theta}}\left\{(H^i_{r_k}(\omega+\Delta\omega;\boldsymbol{\theta}))^2\right\}}}\\
&= \frac{1}{1+\left(\Delta\omega\frac{T_{60}}{13.8}\right)^2}, && (14)\\
\varphi_{ri}(\Delta\omega) &\triangleq \frac{E_{\boldsymbol{\theta}}\left\{H^r_{r_k}(\omega;\boldsymbol{\theta})H^i_{r_k}(\omega+\Delta\omega;\boldsymbol{\theta})\right\}}{\sqrt{E_{\boldsymbol{\theta}}\left\{(H^r_{r_k}(\omega;\boldsymbol{\theta}))^2\right\}E_{\boldsymbol{\theta}}\left\{(H^i_{r_k}(\omega+\Delta\omega;\boldsymbol{\theta}))^2\right\}}}\\
&= \frac{\Delta\omega\frac{T_{60}}{13.8}}{1+\left(\Delta\omega\frac{T_{60}}{13.8}\right)^2}. && (15)
\end{aligned}$$

It is important to note that the expectation operator should be interpreted as the ensemble average over all possible values of $\boldsymbol{\theta}$ satisfying *A3*, which we indicated with the notation $E_{\boldsymbol{\theta}}\{\cdot\}$.

At least two important conclusions can be drawn from equations (12)-(15):

1. For $\Delta\omega = 0$, the real and imaginary parts of $H_{r_k}(\omega;\boldsymbol{\theta})$ are uncorrelated.

2. For $\Delta\omega$ "sufficiently large", $H_{r_k}(\omega;\boldsymbol{\theta})$ and $H_{r_k}(\omega+\Delta\omega;\boldsymbol{\theta})$ are uncorrelated.

To give a more precise meaning of "$\Delta\omega$ sufficiently large", we define a "coherence bandwidth" $\rho(T_{60})$ as

$$\rho(T_{60}) \triangleq \frac{1}{2\pi}\int_{-\infty}^{\infty}\varphi_{rr}(\Delta\omega)d(\Delta\omega) \simeq \frac{7}{T_{60}}. \tag{16}$$

The chosen definition of $\rho(T_{60})$ is rather arbitrary. An alternative definition is

$$\frac{1}{1+\left(2\pi\rho(T_{60})\frac{T_{60}}{13.8}\right)^2} = \frac{1}{2} \Rightarrow \rho(T_{60}) \simeq \frac{2}{T_{60}}. \tag{17}$$

Consequently, $H_{r_k}(\omega_1;\boldsymbol{\theta})$ and $H_{r_k}(\omega_2;\boldsymbol{\theta})$ can be considered approximately uncorrelated if $|\omega_1-\omega_2| \geq \rho(T_{60})$.

Before we conclude our review of the statistical properties of $\mathbf{H}_r(\omega;\boldsymbol{\theta})$, there are two more important questions that should be addressed. The first unresolved issue is the amount of mutual *spatial* correlation between $H_{r_1}(\omega;\boldsymbol{\theta})$ and $H_{r_2}(\omega+\Delta\omega;\boldsymbol{\theta})$, and the second issue is how to find the variance of $\mathbf{H}_r(\omega;\boldsymbol{\theta})$.

The most difficult issue is to analyze the spatial correlation between $H_{r_1}(\omega;\boldsymbol{\theta})$ and $H_{r_2}(\omega+\Delta\omega;\boldsymbol{\theta})$. To the best of the authors knowledge, no general results are available in the literature. However, if it is assumed that the sound source consists of a single sinusoid with angular frequency ω, it can been shown that [6, Chapter 8]

$$\frac{E_{\boldsymbol{\theta}}\{x_1(t)x_2(t)\}}{\sqrt{E_{\boldsymbol{\theta}}\{x_1^2(t)\}E_{\boldsymbol{\theta}}\{x_2^2(t)\}}} = \frac{\sin\left(\frac{\omega d}{c}\right)}{\frac{\omega d}{c}}, \tag{18}$$

which holds true if $\mathbf{h}_r(t;\boldsymbol{\theta})$ corresponds to diffuse sound, and if $\mathbf{h}_d(t)$ is negligible. If the source energy is not concentrated to a single frequency, ω in (18) can be replaced with the mean of the highest and lowest frequencies of the source signal. Expression (18) is only approximately true in this case. Similarly to the definition of the *coherence bandwidth* $\rho(T_{60})$ we can define a *spatial correlation distance* as

$$\int_{-\infty}^{\infty} \frac{\sin\left(\frac{\omega d}{c}\right)}{\frac{\omega d}{c}} d(d) = \frac{c}{\omega/2\pi}. \tag{19}$$

Hence, we consider $H_{r_1}(\omega;\boldsymbol{\theta})$ and $H_{r_2}(\omega;\boldsymbol{\theta})$ approximately uncorrelated if the spatial separation d is larger than $2\pi c/\omega$. For speech signals, $\omega \simeq \pi(3500-300)$, which implies that the spatial correlation is negligible if $d > 0.2m$.

The final issue to resolve, is how to compute the variance of the random variable $H_{r_k}(\omega;\boldsymbol{\theta})$. This problem has however previously been studied in the literature, see for example [6, Chapter 5], where we find that

$$q^2 \triangleq E_{\boldsymbol{\theta}}\left\{H_{r_k}(\omega;\boldsymbol{\theta})(H_{r_k}(\omega;\boldsymbol{\theta}))^*\right\} = \frac{4\beta^2}{\mathcal{A}(1-\beta^2)}, \tag{20}$$

where β $(0 \leq \beta \leq 1)$ is the reflection coefficient, $\mathcal{A}$ denotes the total wall area of the room, and $(\cdot)^*$ denotes complex conjugate.

Before we conclude the section, we would like to discuss the statistical distribution of $\mathbf{H}_r(\omega;\boldsymbol{\theta})$. It is natural to assume that $\mathbf{H}_r(\omega;\boldsymbol{\theta})$ has a complex-valued circularly symmetric (i.e. $\varphi_{ri}(0)=0$) Gaussian distribution, denoted as

$$\mathbf{H}_r(\omega;\boldsymbol{\theta}) \in \mathcal{N}\left(\mathbf{0}, q^2\mathbf{I}_2\right), \tag{21}$$

where $\mathbf{I}_m$ denotes the $m \times m$ identity matrix. Here, the spatial correlation has been neglected in accordance with the above discussion. At least for large values of T_{60}, the assumption on Gaussianity can be motivated from the central limit theorem, simply due to the superposition of a large number of echoes with random phases and amplitudes. See also the discussion in [6, Chapter 3], where it is shown that the magnitude of the sound pressure, assuming diffuse sound propagation and a single frequency excitation, has a Rayleigh distribution, which gives support to the assumed Gaussianity of $\mathbf{H}_r(\omega;\boldsymbol{\theta})$.

III. The New Model

A. Modeling

We are now in position to propose the new reverberation model. Assuming that we have recorded the signal $\mathbf{x}(t)$ for $0 \leq t \leq T$, the frequency domain representation of $\mathbf{x}(t)$ reads as

$$\begin{aligned} \mathbf{X}(\omega) &= \int_0^T \mathbf{x}(t) e^{-j\omega t} dt \simeq \mathbf{H}(\omega) S(\omega) \\ &= \left(\mathbf{H}_d(\omega;\tau_0) + \mathbf{H}_r(\omega;\boldsymbol{\theta})\right) S(\omega), \end{aligned} \tag{22}$$

where

$$\mathbf{H}_d(\omega;\tau_0) = \begin{bmatrix} 1 \\ \\ e^{-j\omega\tau_0} \end{bmatrix} \kappa(r). \tag{23}$$

Here it was assumed that the integration time T is large so that windowing distortions of the Fourier transform are negligible. To simplify the notation, we next scale the measured output by a factor $1/\kappa(r)$, and define the following quantities and their respective Fourier transforms:

$$\mathbf{a}(t) \stackrel{\triangle}{=} \frac{1}{\kappa(r)} \mathbf{h}_d(t) \leftrightarrow \mathbf{A}(\omega;\tau_0) \tag{24}$$

$$\mathbf{r}(t;\boldsymbol{\theta}) \stackrel{\triangle}{=} \frac{1}{\kappa(r)} \mathbf{h}_r(t;\boldsymbol{\theta}) \leftrightarrow \mathbf{R}(\omega;\boldsymbol{\theta}) \in \mathcal{N}\left(0, \sigma^2 \mathbf{I}_2\right), \tag{25}$$

where

$$\sigma^2 \stackrel{\triangle}{=} q^2/\kappa(r)^2. \tag{26}$$

Assume next that the source signal $s(t)$ is band-limited, i.e. its power is zero outside the interval $[f_l, f_u]$ Hz. The signal bandwidth is then $B \stackrel{\triangle}{=} f_u - f_l$ Hz. We also assume that f_l is larger than the Schroeder large

room frequency f_S, defined in (9). Suppose next that the microphone output $\mathbf{x}(t)$ is sampled with sampling frequency F_sHz (assuming that $f_u \leq F_s/2$), to produce the sequence $\mathbf{x}(nT_s),\ n = 0, 1, \cdots, N-1$, where $T_s = 1/F_s$ and $N = \text{round}\{TF_s\}$. For sampled data, the Fourier transform (22) is usually computed using the Discrete Fourier Transform (DFT). Hence, up to within a scaling,

$$\mathbf{X}(\omega_k) \simeq \sum_{n=0}^{N-1} \mathbf{x}(nT_s) e^{-j\omega_k nT_s}. \tag{27}$$

The DFT operation produces the following sampling of the frequency axis:

$$\omega_k = \frac{2\pi F_s k}{N},\ k = 0, 1, \cdots, N-1. \tag{28}$$

To find a suitable model, we consider the following two cases:

M1: Suppose that $\rho(T_{60}) \leq F_s/N$. In this case, we find the following model (for simplicity, only frequencies in the interval $[0, F_s/2]$ are included):

$$\mathbf{X}(\omega_k) = \left(\mathbf{A}(\omega_k; \tau_0) + \mathbf{R}(\omega_k; \boldsymbol{\theta})\right) S(\omega_k),\ k = k_l, \cdots, k_u, \tag{29}$$

where

$$k_l = \text{round}\left\{\frac{Nf_l}{F_s}\right\} \tag{30}$$

$$k_u = \text{round}\left\{\frac{Nf_u}{F_s}\right\}, \tag{31}$$

and round$\{\cdot\}$rounds towards the nearest integer value. Hence, if the coherence bandwidth is smaller than the frequency sampling induced by the DFT, the sequence $\{\mathbf{R}(\omega_k; \boldsymbol{\theta})\}_{k=k_l}^{k_u}$ consists of *uncorrelated* Gaussian random vectors.

M2: Suppose next that $\rho(T_{60}) > F_s/N$. This case is somewhat more difficult to handle. Due to the large (relative the DFT sampling of the frequency axis) coherence bandwidth, the correlation between adjacent values of $\mathbf{R}(\omega_k; \boldsymbol{\theta})$ is not negligible. In theory it is of course possible to include this correlation in a parametric manner (see for example Remark 2 below). For now, we however avoid such precise modeling, and conclude that $\{\mathbf{R}(\omega_k; \boldsymbol{\theta})\}_{k=k_l}^{k_u}$ is a sequence of *correlated* Gaussian random vectors.

Remark 2: An interesting possibility to incorporate the frequency correlation of $\mathbf{R}(\omega_k;\boldsymbol{\theta})$ could be as follows. Suppose that $\mathbf{R}(\omega_k;\boldsymbol{\theta})$ is generated from an Auto-Regressive (AR) random process:

$$\mathbf{R}(\omega_k;\boldsymbol{\theta}) = \alpha\mathbf{R}(\omega_{k-1};\boldsymbol{\theta}) + \mathbf{w}(k), \tag{32}$$

where $\mathbf{w}(k)$ is a white Gaussian noise sequence. The real-valued scalar α should then be chosen such that the frequency correlation of $\mathbf{R}(\omega_k;\boldsymbol{\theta})$ approximately satisfies (12)-(15). Since this approach complicates the investigation, analysis of the model (32) is deferred to future research.

B. Discussion

Although analysis of reverberation at first seems difficult, we have demonstrated that it is possible to derive a fairly simple model for explaining the effects of reverberation. The crucial assumptions are

1. The reverberant part of the impulse response $\mathbf{r}(t;\boldsymbol{\theta})$ can be considered equivalent with diffuse sound propagation.

2. By considering the coherence bandwidth $\rho(T_{60})$, a sequence $\{\mathbf{R}(\omega_k;\boldsymbol{\theta})\}_{k=k_l}^{k_u}$ of random vectors can be constructed. The correlation function of the random vectors $\{\mathbf{R}(\omega_k;\boldsymbol{\theta})\}_{k=k_l}^{k_u}$ then depends on the coherence bandwidth $\rho(T_{60})$.

One conclusion of the above discussion is that the effects of reverberation is similar to that of additive noise. The most important distinction is that the variance of the "noise term" $\mathbf{R}(\omega_k;\boldsymbol{\theta})S(\omega_k)$ is proportional to the power of the input signal. From an engineering standpoint, this means that unlike in the additive noise case, simply increasing the signal level will not be able to alleviate the problem.

Remark 3: Consider next an extension to a microphone array. Assume that the considered array of microphones consists of a uniform linear array with M microphones. Assuming that the spacing between the

microphones equals d, and that the source is sufficiently far away, it follows that

$$\mathbf{a}(t) = \begin{bmatrix} 1 \\ \delta(t-\tau_0) \\ \vdots \\ \delta(t-(M-1))\tau_0 \end{bmatrix}. \tag{33}$$

Definition (33) relies on the assumption that the sound wave transmitted from the source is plane. For $M > 2$ the validity of that assumption is questionable in a room environment, simply due to geometrical considerations.

Remark 4: Attempting to localize human speakers, the above discussion should not be taken too literally. This since it cannot be expected that the position of a human speaker is completely stationary. As he/she is speaking, the head moves, which affects $\mathbf{H}(\omega)$. Our empirical experience is that these movements actually makes it easier to localize the speaker, compared to a stationary loudspeaker. A potential explanation is that the head-movements ensures that the recorded microphone signal consists of several realizations of the reverberant transfer function $\mathbf{R}(\omega;\boldsymbol{\theta})$. Hence, there is a chance of decreasing the effects of reverberation by segmenting the available data and performing averaging over different realizations of the involved estimates of the auto-and cross-spectra (assuming a frequency-domain based approach for estimation of the time-delay). Furthermore, we must keep in mind that sound originating from a human speaker tends to be more directional than sound from the assumed omni-directional point source.

C. Similarity with Ricean Fading in Digital Mobile Communication

It is interesting to note that the introduced model for the transfer function $\mathbf{H}(\omega)$ essentially corresponds to a so-called "frequency selective fading" model, commonly used for modeling multi-path propagation in mobile communication. If we assume that $\mathbf{R}(\omega;\boldsymbol{\theta})$ is Gaussian, the reverberation model is equivalent with "Ricean fading". Note also that the reverberation model corresponds to Rayleigh fading when the direct-path propagation $\mathbf{A}(\omega;\tau_0)$ is absent. See for example [8] for a discussion on fading in the context of digital mobile communication. Here we just point out that Ricean fading usually occurs when the direct line of

sight communication is distorted with reflections from a large number of scatterers. Note that the factor $\kappa(r)$ then corresponds to the "propagation path-loss". Furthermore, expressions (12)-(15) also appears in statistical models of fading. Then the quantity $T_{60}/13.8$ usually goes under the name "delay-spread". Note, finally, that the analogy with fading in mobile digital communication was the reason to introduce the term "coherence bandwidth" for $\rho(T_{60})$. This since the coherence bandwidth is a well-established quantity in the digital communication community.

IV. Experimental Results

Synthetic Data

The purpose of the first example is to investigate to which extent the assumption of diffuse sound is valid. In particular, we will compare the theoretical expressions (12)-(15) with the outcome of a Monte-Carlo simulation. For the simulations, we consider a rectangular room with plane reflective surfaces. Each boundary is characterized by its reflection coefficient β, $0 \leq \beta \leq 1$, which is assumed to be identical for all walls. The dimensions of the room along the x, y, and z-axes are denoted L_x, L_y and L_z, respectively. In our simulations we have studied a scenario where $L_x = 10$, $L_y = 6.6$, and $L_z = 3$ (all measures in meters). Furthermore, we assume that there is an omni-directional acoustical point source present in the interior of the room. The radiated signal is measured by omni-directional microphones, which are located in the interior of the room (compare with Assumption *A3*). Except for the source and the microphones, the room is assumed to be empty.

In the paper, we have relied on the assumption that the diffuse part of the room transfer function is a random function, where the randomness is with respect to the positions of the source and the microphone. To generate independent realizations of the room transfer function, we apply the following procedure:

1. Define

$$\tilde{\mathbf{r}}_{m_1} = \begin{bmatrix} d/2 & 0 & 0 \end{bmatrix}^T \tag{34}$$

$$\tilde{\mathbf{r}}_{m_2} = \begin{bmatrix} -d/2 & 0 & 0 \end{bmatrix}^T. \tag{35}$$

The source is assumed to be located at $\tilde{\mathbf{r}}_s$, such that the distance from the source to the point in between the microphones is $r = \|\tilde{\mathbf{r}}_s\|$. The true time-delay is then obtained as

$$\tau_0 = \frac{1}{c}\left(\|\tilde{\mathbf{r}}_s - \tilde{\mathbf{r}}_{m_1}\| - \|\tilde{\mathbf{r}}_s - \tilde{\mathbf{r}}_{m_2}\|\right). \tag{36}$$

This configuration of $\tilde{\mathbf{r}}_s$, $\tilde{\mathbf{r}}_{m_1}$, and $\tilde{\mathbf{r}}_{m_2}$, is fixed for each Monte-Carlo run.

2. For each Monte-Carlo run, generate a random translation $\mathbf{y}$ and a 3×3 random rotation matrix $\mathbf{G}$ (i.e. $\mathbf{G}^T\mathbf{G} = \mathbf{I}_3$), and let

$$\mathbf{r}_s = \mathbf{y} + \mathbf{G}\tilde{\mathbf{r}}_s \tag{37}$$

$$\mathbf{r}_{m_1} = \mathbf{y} + \mathbf{G}\tilde{\mathbf{r}}_{m_1} \tag{38}$$

$$\mathbf{r}_{m_2} = \mathbf{y} + \mathbf{G}\tilde{\mathbf{r}}_{m_2}. \tag{39}$$

Since r and τ_0 are invariant to the above translation and rotation of the coordinate system, the direct-path transfer function $\mathbf{H}_d(\omega)$ is also invariant. Here the random translation $\mathbf{y}$ is uniformly distributed in the room volume, and the rotation angles used to define $\mathbf{G}$ are uniformly distributed in the interval $[-\pi, \pi]$.

3. If either $\mathbf{r}_s$, $\mathbf{r}_{m_1}$, or $\mathbf{r}_{m_2}$ is located outside the room, return to 2). Else, compute $\mathbf{H}(\omega)$ using $\mathbf{r}_s$, $\mathbf{r}_{m_1}$, and $\mathbf{r}_{m_2}$, as input data.

4. For each Monte-Carlo run, generate a new realization of the source signal $s(t)$.

In this manner, it is possible to compute independent realizations of $\mathbf{R}(\omega;\boldsymbol{\theta})$, without affecting the direct-path transmission[2]. To generate sampled versions of the acoustical impulse response $\mathbf{h}(t)$, we apply the image-method [1]. In all simulations the sampling frequency is chosen as $F_s = 10$ kHz. The impulse response $\mathbf{h}(t)$ theoretically extends to infinity. The simulated $\mathbf{h}(t)$ is however truncated to 6000 (0.6 s) samples. For each realization of $\mathbf{h}(nT_s)$ we compute the frequency response function $\mathbf{H}(\omega)$ at $\simeq$2500 equidistant frequencies in the interval $[300, 5000]$Hz. We next compute the reverberant part of the frequency response by subtracting

[2]The same procedure will be applied in [4], where it is more crucial to keep $\mathbf{A}(\omega;\tau_0)$ fixed while generating independent realizations of $\mathbf{R}(\omega;\boldsymbol{\theta})$. This since the accuracy of TDE is studied in [4].

the direct-path transfer function

$$\mathbf{H}_r(\omega_k;\boldsymbol{\theta}_m) = \mathbf{H}(\omega_k) - \mathbf{H}_d(\omega_k;\tau_0). \tag{40}$$

Here we introduced the notation $(\boldsymbol{\theta}_m)$ to indicate that $\mathbf{H}_r(\omega_k;\boldsymbol{\theta}_m)$ is conditioned on the m^{th} realization of the random vector $\boldsymbol{\theta}_m$. For a particular outcome of $\boldsymbol{\theta}_m$, we can now, for example, estimate the auto-correlation function (denoted as upper-case $C(\cdot\,;\boldsymbol{\theta}_m)$) of the real part of $H_{r_1}(\omega_k;\boldsymbol{\theta}_m)$ as

$$\hat{C}_{r_1r_1}(l;\boldsymbol{\theta}_m) = \frac{1}{L}\sum_{k=1}^{K-l} H^r_{r_1}(\omega_k;\boldsymbol{\theta}_m)H^r_{r_1}(\omega_{k+l};\boldsymbol{\theta}_m), \tag{41}$$

and similarly for the normalized cross-correlation (denoted as lower-case $c(\cdot\,;\boldsymbol{\theta}_m)$) between the real and imaginary parts of $H_{r_1}(\omega_k;\boldsymbol{\theta}_m)$:

$$\begin{aligned}&\hat{c}_{r_1i_1}(l;\boldsymbol{\theta}_m)\\ &\quad= \frac{\frac{1}{L}\sum_{k=1}^{K-l} H^r_{r_1}(\omega_k;\boldsymbol{\theta}_m)H^i_{r_1}(\omega_{k+l};\boldsymbol{\theta}_m)}{\sqrt{\frac{1}{L}\sum_{k=1}^{K}(H^r_{r_1}(\omega_k;\boldsymbol{\theta}_m))^2\frac{1}{L}\sum_{k=1}^{K}(H^i_{r_1}(\omega_k;\boldsymbol{\theta}_m))^2}},\end{aligned} \tag{42}$$

where K denotes the number of frequency points in the interval $[300, 5000]$Hz. The final estimate of the involved correlation function is then defined as

$$\hat{C}_{r_1r_1}(l) \triangleq \frac{1}{W}\sum_{m=1}^{W}\hat{C}_{r_1r_1}(l;\boldsymbol{\theta}_m) \tag{43}$$

where W denotes the number of independent realizations of $\mathbf{H}(\omega)$. The corresponding normalized correlation function $\hat{c}_{r_1r_1}(l)$ is similarly defined.

In order to apply the image method, we first need to compute the value of the reflection coefficient β which results in the given reverberation time. For a given value of T_{60}, the reflection coefficient β is here computed from *Eyring's* formula[6]:

$$\beta = \exp\left\{-\frac{13.82}{(\frac{1}{L_x}+\frac{1}{L_y}+\frac{1}{L_z})cT_{60}}\right\}. \tag{44}$$

In Fig.1 the estimated values of $\hat{C}_{r_1r_1}(l)$ and $\hat{c}_{r_1i_1}(l)$ are illustrated for the case $T_{60} = 0.5s$, which according to (44) corresponds to $\beta = 0.87$. In Fig. 1, also the theoretical expressions for the involved correlation functions

are included, i.e.

$$C_{r_1 r_1}(\Delta\omega) = \frac{q^2/2}{1+\left(\Delta\omega\frac{T_{60}}{13.8}\right)^2} \quad (45)$$

$$c_{r_1 i_1}(\Delta\omega) = \frac{\Delta\omega\frac{T_{60}}{13.8}}{1+\left(\Delta\omega\frac{T_{60}}{13.8}\right)^2}. \quad (46)$$

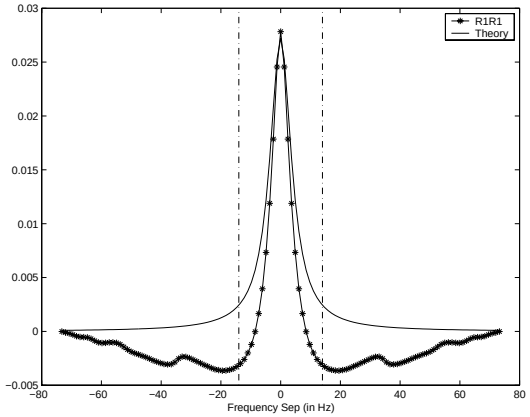
The results in Fig.1 indicate a surprisingly good match between the theoretical and estimated correlation functions. It is especially important to observe the correlation values at $\Delta\omega = \rho(T_{60}) = 7/T_{60}$. Both $\hat{C}_{r_1 i_1}$ and $\hat{c}_{r_1 i_1}$ are close to zero for $\Delta\omega > \rho(T_{60})$, which illustrates the idea behind the definition of the coherence bandwidth. Note also the almost perfect agreement between $q^2/2$ and the empirical variance of $H_r^r(\omega)$ and $H_r^i(\omega)$.

Next we turn our attention to the correlation between $H_{r_1}(\omega)$ and $H_{r_2}(\omega+\Delta\omega)$. In the analysis in Section II-B, we noted that few results are available on the theoretical spatial correlation function. It is hence of interest to investigate $\hat{c}_{r_1 r_2}$ and $\hat{c}_{r_1 i_2}$ by means of Monte-Carlo simulations. In Fig. 2, the outcome of $\hat{c}_{r_1 r_2}$ and $\hat{c}_{r_1 i_2}$ is illustrated for two different microphone separations ($d = 1m$ and $d = 0.1m$, respectively). We note that the spatial correlation increases as d decreases, as expected. However, even if the microphones are as close as $d = 0.1m$, the normalized spatial correlation does not exceed 0.2 for $\Delta\omega > \rho(T_{60})$. The main modeling error seems to be that $\hat{c}_{r_1 r_2}(0) \simeq 0.3$, when $d = 0.1m$. However, if the distance is increased to $d = 0.25m$, the value of $\hat{c}_{r_1 r_2}(0)$ drops to approximately 0.15. Hence, also the assumption that the spatial correlation is negligible seems quite applicable, at least for microphone separations larger than $\simeq 0.25m$. This observation agrees well with the theoretical investigation in Section II-B.

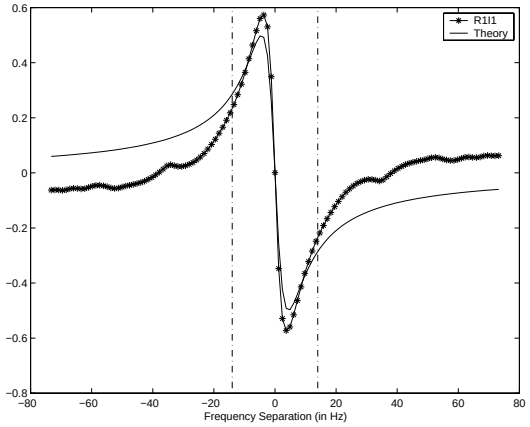
In Fig. 3, we study the spatial correlation in a different manner. In Fig. 3, we plot $\hat{c}_{r_1 r_2}(0)$ as a function of d. As expected, the spatial correlation decreases as a function of d.

Real Data

Our final example deals with real data. At our disposal, we had an office with dimensions $L_x = 6.6m$, $L_y = 3.3m$, and $L_z = 2.9m$. The room was empty, except for a table in the middle of the room. We placed a microphone at location $\mathbf{r}_m = [0.1\ 1.9\ 1.5]^T$. The position of the microphone then actually violates assumption



(a) $\hat{C}_{r_1 r_1}(l)$.



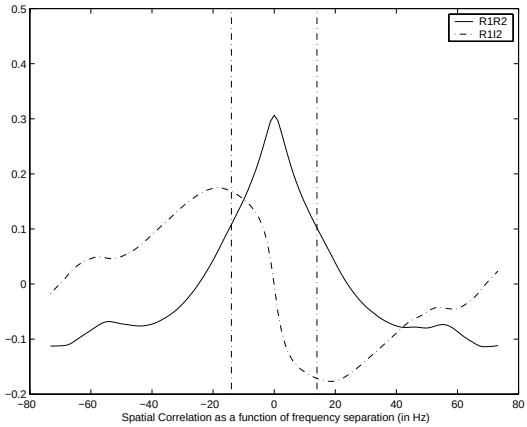
(b) $\hat{c}_{r_1 i_1}(l)$.

Fig. 1. Synthetic Data: Estimated correlation functions of $H_{r_1}(\omega;\boldsymbol{\theta})$. Here $r = 3m$, the number of Monte-Carlo runs equals $W = 250$, and $T_{60} = 0.5$s. Vertical dash-dotted line indicates the coherence bandwidth $\rho(T_{60}) = 7/T_{60}$, and the solid line illustrates the theoretical correlation functions.

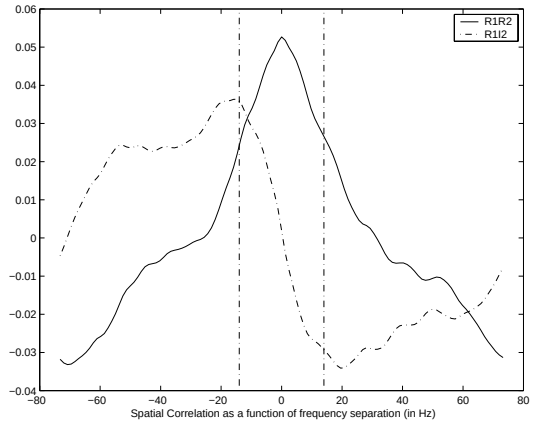
A3, since the microphone is not located in the interior of the room.

We next measured the transfer function from various source positions to the fixed microphone[3]. See Fig. 4 for an illustration of the various source positions. The locations missing in Fig. 4 were due to the table present in the room. For all source positions, the loudspeaker was located at a fixed height of $1.1m$. The transfer functions were estimated using a chirp-signal as input $(100-500\text{Hz})$. Note, since we study low frequencies, we study the frequency band where the assumption of diffuse sound is the least appropriate due to the Schroeder

[3]In these recordings we used a "Radio Shack omni-directional loud-speaker", Cat.No. 40.1352.



(a) $d = 0.1$m.



(b) $d = 1$m.

Fig. 2. Synthetic Data: Spatial correlation for two different values of the microphone separation. Here $r = 3m$, the number of Monte-Carlo runs equals $W = 250$, $T_{60} = 0.5$s, and $\tau_0 = 0$. Vertical dash-dotted line indicates the coherence bandwidth $\rho(T_{60}) = 7/T_{60}$.

large room frequency. The microphone outputs were sampled with sampling frequency $F_s = 2000$ Hz, and the estimated transfer functions were computed at 1024 frequency points in the interval $[0, Fs/2]$ Hz using the Matlab-function `tfe.m`.

Given the set of 128 estimated transfer functions, we compute $\hat{c}_{r_1 r_1}$ and $\hat{c}_{r_1 i_1}$ as in the previous example. However, before we investigate these results, we first study the ensemble average of the envelope of the

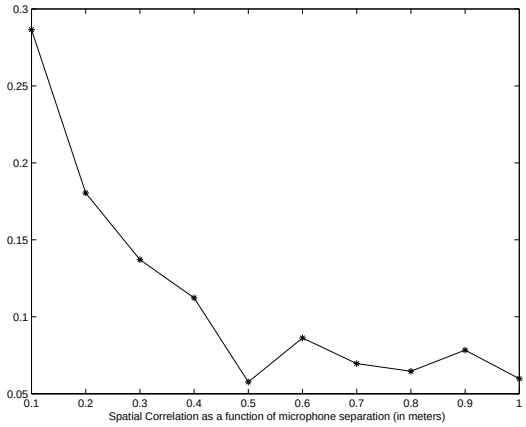


Fig. 3. Synthetic Data: Spatial correlation, as a function of microphone separation. Here $r = 3m$, the number of Monte-Carlo runs equals $W = 250$, $T_{60} = 0.5$s, and $\tau_0 = 0$.

estimated impulse response:

$$\bar{\hat{h}}(nT_s) \triangleq \frac{1}{128} \sum_{k=1}^{128} \hat{h}_k^2(nT_s), \tag{47}$$

where $\hat{h}_k(t)$ denotes the estimated impulse response from the k^{th} loudspeaker position. If $h_k(t)$ is due to diffuse sound propagation only, $\bar{\hat{h}}(nT_s)$ should decay as $e^{-13.8nT_s/T_{60}}$, cf. [11]. Hence, $\log\{\bar{\hat{h}}(nT_s)\}$ should decay linearly as a function of n. In Fig. 5, the outcome of $\log\{\bar{\hat{h}}(nT_s)\}$ is illustrated, where a straight-line approximation of the reverberant part is illustrated as well. From the straight-line approximation, we estimate the reverberation time as approximately $\hat{T}_{60} \simeq 0.35s$, resulting in a Schroeder large room frequency $f_S \simeq 150$Hz. Using $\hat{T}_{60}$, we can further compute the theoretical values of $c_{r_1r_1}$ and $c_{r_1i_1}$. Hence, in Fig. 6, we illustrate $c_{r_1r_1}$ and $c_{r_1i_1}$ together with $\hat{c}_{r_1r_1}$ and $\hat{c}_{r_1i_1}$.

¿From Fig. 6 we note a good agreement with the theoretical expressions, and in Fig. 5 we see that $\log\{\bar{\hat{h}}(nT_s)\}$ is close to a straight line (at least for $n > 40$ samples). Note that these results are obtained with the direct-path propagation included, in contrast to our examples using synthetic data. This is probably the reason why the theoretical correlation decays more rapidly than $\hat{c}_{r_1r_1}$ as the frequency separation increases. The direct-path propagation is included since it is difficult to subtract the direct-path transfer function when dealing with real data. In Figure 5, we further notice three different characteristics of $\log\{\bar{\hat{h}}(nT_s)\}$: direct-path propagation for $5 \leq n \leq 15$, dominant early reflections for $15 \leq n \leq 40$, and late reverberation for $n > 40$.

The main modeling error of the proposed reverberation model in then that the dominant early reflections are neglected and lumped together with the reverberant part.

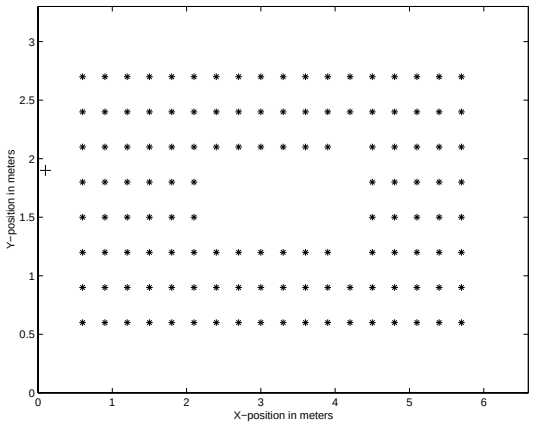


Fig. 4. Locations of the loudspeaker used to estimate room transfer functions. Here "+" indicates the location of the microphone, and "∗" indicates the loudspeaker positions.

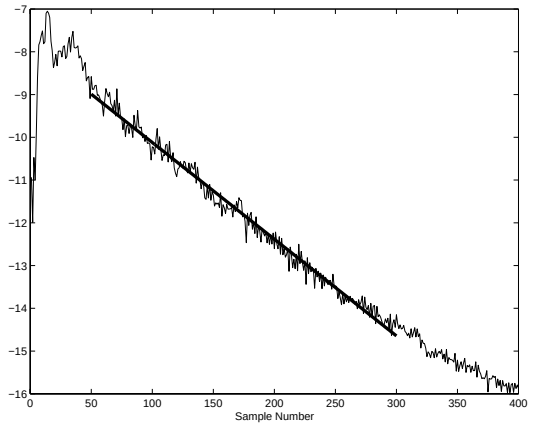
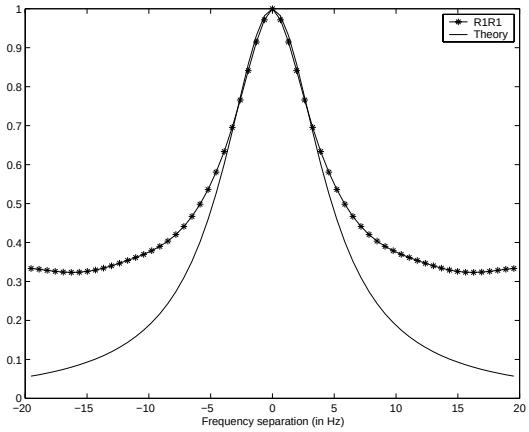


Fig. 5. Real Data: Logarithm of $\bar{\hat{h}}(nT_s)$.

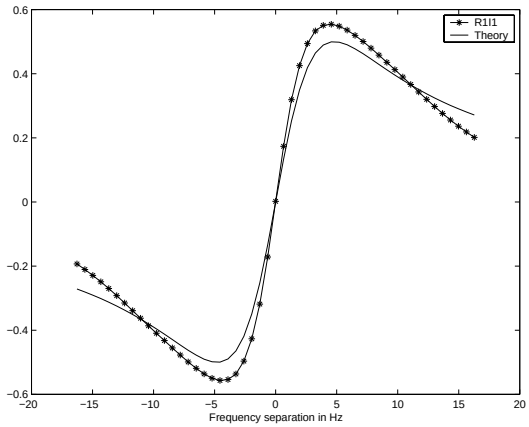
V. Conclusions

Robust microphone-based source localization is an important ingredient in several multimedia signal-processing systems. However, experience has shown that the problem of localizing acoustical sources in reverberant environments is difficult. Hence, for quite some time there has been an interest in understanding and analyzing the performance of localization techniques when room reverberation is present.

For this purpose, a new reverberation model was proposed. This model is developed exploiting known



(a) $\hat{c}_{r_1 r_1}(l)$.



(b) $\hat{c}_{r_1 i_1}(l)$.

Fig. 6. Real Data: Estimated normalized correlation functions of $\mathbf{H}(\omega)$, using real data.

results in room acoustics along with Most of the presented results were previously known in the literature. that the room transfer function consists of a direct-path propagation and a reverberation tail, where the reverberation tail of the room impulse response is assumed to describe diffuse sound propagation.

Simulated examples, and also real data recordings, indicated that the proposed model accurately can describe actual room transfer functions. In a companion paper, the statistical model will among others be utilized to analyze common time-delay estimators.

Acknowledgement

The authors would like to express their gratitude to Mr. Jonathon Vance who provided the real-data recordings.

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

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

[5] Yoichi Haneda, Shoji Makino, and Yutaka Kaneda. "Multiple-Point Equalization of Room Transfer Functions by Using Common Acoustical Poles". *IEEE Trans. on Speech and Audio Processing*, 5(4):325–333, July 1997.

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

[7] 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.

[8] J.G. Proakis. *Digital Communications, third edition*. McGraw-Hill, 1995.

[9] B.D. Radlović, R.C. Williamson, and R.A. Kennedy. "Equalization in an Acoustic Reverberant Environment: Robustness Results". *IEEE Trans. on Speech and Audio Processing*, 8(3):311–319, May 2000.

[10] 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.

[11] M.R. Schroeder. "Frequency Correlation Functions of Frequency Responses in Rooms". *Journal of the Acoustical Society of America*, 34(12):1819–1823, 1963.

[12] C. Wang and M. Brandstein. "Multi-source Face Tracking with Audio and Visual Data". In *IEEE Third Workshop on Multimedia Signal Processing*, pages 169–174, Copenhagen, Denmark, 1999.

[13] C. Wang and M. S. Brandstein. "A Hybrid Real-Time Face Tracking System". In *Proc. ICASSP*, pages 3737–3740, Seattle, USA, 1998.