Relative Transfer Function Identification Using Convolutive Transfer Function Approximation

Ronen Talmon, Israel Cohen, *Senior Member, IEEE*, and Sharon Gannot, *Senior Member, IEEE*

Abstract

In this paper we present a relative transfer function (RTF) identification method for speech sources in reverberant environments. The proposed method is based on the convolutive transfer function (CTF) approximation, which enables to represent a linear convolution in the time domain as a linear convolution in the short-time Fourier transform (STFT) domain. Unlike the restrictive and commonly used multiplicative transfer function (MTF) approximation, which becomes more accurate when the length of a time frame increases relative to the length of the impulse response, the CTF approximation enables representation of long impulse responses using short time frames. We develop an unbiased RTF estimator that exploits the non-stationarity and presence probability of the speech signal and derive an analytic expression for the estimator variance. Experimental results show that the proposed method is advantageous compared to common RTF identification methods in various acoustic environments, especially when identifying long RTFs, typical to real rooms.

Index Terms

Acoustic noise measurement, adaptive signal processing, array signal processing, speech enhancement, system identification.

I. Introduction

Identification of a relative transfer function (RTF) between two sensors is an important component of multichannel hands-free communication systems, particulary in reverberant and noisy environments [1], [2], [3], [4], [5]. Shalvi and Weinstein [6] proposed to identify the coupling between speech components received at two microphones by using the non-stationarity of the desired speech signal received at the sensors, assuming stationary additive noise and static RTF. By dividing the observation interval into a sequence of subintervals, the speech signal can be regarded as stationary in each subinterval, and non-stationary between subintervals. Thus, computing the cross power spectral density (PSD) of the sensor signals in each subinterval yields an overdetermined set of equations for two unknown variables: the RTF and the cross PSD of the sensors' noise signals. Estimates of these two variables

Ronen Talmon and Israel Cohen are with the Department of Electrical Engineering, Technion - Israel Institute of Technology, Haifa 32000, Israel (e-mail: ronenta2@techunix.technion.ac.il; icohen@ee.technion.ac.il).

Sharon Gannot is with the School of Engineering, Bar-Ilan University, Ramat-Gan, 52900, Israel (e-mail: gannot@eng.biu.ac.il).

This research was supported by the Israel Science Foundation (grant no. 1085/05).

are derived using the weighted least squares (WLS) approach. One limitation of the non-stationarity based method is that both the RTF and the noise PSD are estimated simultaneously through the same WLS optimization criterion. This restricts the RTF identification performance since it requires large weights in high signal-to-noise ratio (SNR) subintervals and low weights in low SNR subintervals, whereas the noise cross PSD estimate requires that the weights be inversely proportional to the SNR.

Cohen [7] proposed an RTF identification method, which solves the above conflict, by adding a-priori knowledge regarding speech presence during each observation interval. By using a voice activity detector (VAD), it is possible to separate the subintervals into two sets, one containing noise-only subintervals, while the other including subintervals where speech is present. The first set enables to find a reliable estimate for the noise cross PSD, while the second set of subintervals is employed for identifying the RTF using the already estimated cross PSD of the noise. Unfortunately, the above methods rely on the multiplicative transfer function (MTF) approximation [8]. The MTF approximation enables to replace a linear convolution in the time domain with a scalar multiplication in the short-time Fourier transform (STFT) domain. This approximation becomes more accurate when the length of a time frame increases, relative to the length of the impulse response. However, long time frames may increase the estimation variance, increase the computational complexity and restrict the ability to track changes in the RTF [8].

In this paper we present an RTF identification method based on the *convolutive* transfer function (CTF) approximation. This approximation enables representation of long impulse responses in the STFT domain using short time frames. We develop an unbiased RTF estimator that exploits the non-stationarity and presence probability of the speech signal. We derive an analytic expression for the estimator variance, and present experimental results that demonstrate the advantages of the proposed method over existing methods. Relying on the analysis of the system identification in the STFT domain with cross-band filtering [9], we show that the CTF approximation becomes more accurate than the MTF approximation, as the SNR increases. In addition, unlike existing RTF identification methods which are based on the MTF approximation, the proposed method enables flexibility in adjusting the lengths of time frames and the estimated RTF. Experimental results demonstrate that the proposed estimator outperforms the competing method when identifying long RTFs. We investigate the influence of important acoustic parameters on the identification accuracy. In particular, we find that the proposed method is advantageous in reverberant environments, when the distance between the sensors and the SNR are larger than certain thresholds.

This paper is organized as follows. In Section II, we formulate the RTF identification problem in the STFT domain. In Section III, we introduce the CTF approximation and propose an RTF identification approach suitable for speech sources in reverberant environments. Finally, in Section IV we present experimental results that demonstrate the advantage of the proposed method.

II. Problem Formulation

Let $s(n)$ denote a non-stationary speech source signal, and let $u(n)$ and $w(n)$ denote additive stationary noise signals, that are uncorrelated with the speech source. The signals are received by primary and reference microphones,

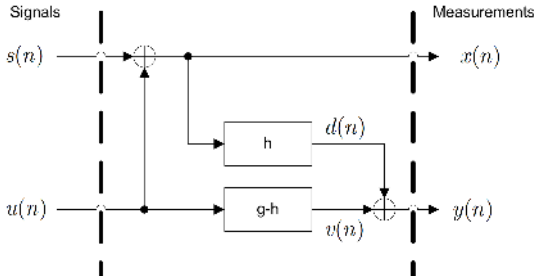


Fig. 1. RTF Model scheme with directional noise.

respectively:

$$x(n) = s(n) + u(n) \tag{1}$$

$$y(n) = h(n) * s(n) + w(n) \tag{2}$$

where * represents convolution. In this work, our goal is to identify the response $h(n)$. Usually $s(n)$ is not a clean speech source signal but a reverberated version, $s(n) = h_1(n) * \bar{s}(n)$, where $\bar{s}(n)$ is the clean speech signal and $h_1(n)$ is the room impulse response of the primary sensor to the speech source. Accordingly, $h_2(n) = h(n) * h_1(n)$ is the room impulse response of the reference sensor to the speech source, and $h(n)$ represents the *relative* impulse response between the microphones with respect to the speech source.

An equivalent representation of (1) and (2) is

$$y(n) = d(n) + v(n) = h(n) * x(n) + v(n) \tag{3}$$

$$v(n) = w(n) - h(n) * u(n) \tag{4}$$

where in (3) we have an LTI system with an input $x(n)$, output $y(n)$, and additive noise $v(n)$. The formulation in (3) cannot be considered as an ordinary system identification problem, since (4) indicates that $v(n)$ depends on both $x(n)$ and $h(n)$.

Here, we assume that the microphones' noise signals are generated by a single noise source in the room. Accordingly, the additive noise at the reference microphone $w(n)$ can be written as

$$w(n) = g(n) * u(n) \tag{5}$$

where $g(n)$ is the relative impulse response between the microphones with respect to the noise source signal. Such an RTF model scheme is represented in Fig. 1.

As in many speech enhancement applications, the signals can be divided into overlapping time frames and analyzed using the short-time Fourier transform (STFT). Common RTF identification methods [6] [7] assume that the support of $h(n)$ is finite and small compared to the length of the time frame. Then, (3) can be approximated in the STFT domain as

$$y_{p,k} = h_k x_{p,k} + v_{p,k} \tag{6}$$

where p is the time frame index, k is the frequency sub-band index and h_k is the RTF. This approximation is known as the multiplicative transfer function (MTF) approximation for modeling an LTI system in the STFT domain [8]. Using (6), the cross PSD between y and x can be written as

$$\phi_{yx}(p,k) = h_k \phi_{xx}(p,k) + \phi_{vx}(p,k). \tag{7}$$

Notice that in (7) we implicitly used the assumption that the speech is stationary in each time frame, which restricts the time frames to be relatively short ($<$ 40ms).

As stated before, a major problem in identifying acoustic impulse responses (AIRs) is their length. AIRs length is significantly influenced by the room reverberation time[1] [10], since the longer the reverberation time is, the longer it takes for the the AIR to convey most of its energy. For typical reverberant rooms with T_{60} of several hundred milliseconds, the MTF approximation restricts the time frames to be much larger than T_{60}, but then the speech signal cannot be assumed stationary during such long time frames. In this work we address the problem of RTF identification in the STFT domain using short time frames, without resorting to the MTF approximation.

Let N_x denote the number of time frames, let N denote the length of a time frame in the STFT domain and let L denote the framing step. According to [9], [12], [13] a filter convolution in the time domain can be represented as a sum of N cross-band convolutions in the STFT domain. The cross-band filters are used for canceling the aliasing caused by sampling in each frequency sub-band [14]. Accordingly, (1) and (2) can be written in the STFT domain as

$$x_{p,k} = s_{p,k} + u_{p,k} \tag{8}$$

$$y_{p,k} = \sum_{k'=0}^{N-1} \sum_{p'} s_{p',k'} h_{p-p',k',k} + w_{p,k} \tag{9}$$

where p is the time frame index, k and k' are the frequency sub-band indices and $h_{p,k',k}$ are the cross-band filter coefficients between frequency bands k' and k of length N_h. The length of $y_{p,k}$ is given by $N_y = N_x + N_h - 1$. Similarly, an STFT representation of (3) and (4) is given by

$$\begin{aligned} y_{p,k} &= d_{p,k} + v_{p,k} \\ &= \sum_{k'=0}^{N-1} \sum_{p'} x_{p-p',k'} h_{p',k',k} + v_{p,k} \end{aligned} \tag{10}$$

$$v_{p,k} = w_{p,k} - \sum_{k'=0}^{N-1} \sum_{p'} u_{p-p',k'} h_{p',k',k}. \tag{11}$$

Let $\mathbf{h}_{k',k}$ denote the cross-band filter from frequency band k' to frequency band k:

$$\mathbf{h}_{k',k} = [h_{0,k',k} \quad h_{1,k',k} \quad \cdots \quad h_{N_h-1,k',k}]^T \tag{12}$$

[1] The room reverberation time is the time for the acoustic energy to attenuate by 60dB, after a sound source is stopped [11]. This value is usually denoted by T_{60}.

and let $\mathbf{h}_k$ denote a column stack concatenation of the cross-band filters $\{\mathbf{h}_{k',k}\}_{k'=0}^{N-1}$

$$\mathbf{h}_k = [\mathbf{h}_{0,k}^T \ \ \mathbf{h}_{1,k}^T \ \cdots \ \mathbf{h}_{N-1,k}]^T . \tag{13}$$

Note that due to the non causality of the cross-band filter $h_{p,k',k}$, the time index p should have ranged differently according to the number of non causal coefficients of $h_{p,k',k}$. However, we assume that an artificial delay has been introduced into the system output signal $y(n)$ in order to compensate for those non causal coefficients. Let

$$\mathbf{X}_k = \begin{bmatrix} x_{0,k} & 0 & \cdots & \cdots & 0 \\ x_{1,k} & x_{0,k} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & x_{0,k} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ x_{N_x-1,k} & \vdots & \vdots & \vdots & x_{N_x-N_h,k} \\ 0 & x_{N_x-1,k} & \vdots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & x_{N_x-1,k} \end{bmatrix} \tag{14}$$

be an $N_y \times N_h$ Toeplitz matrix constructed from the STFT coefficients of the input signal x in the kth sub-band, and let Δ_k^x be a concatenation of $\{\mathbf{X}_k\}_{k=0}^{N-1}$

$$\Delta_k^x = [\mathbf{X}_0 \ \cdots \ \mathbf{X}_{N-1}]^T . \tag{15}$$

Similarly, define Δ_k^u as a concatenation of $\{\mathbf{U}_k\}_{k=0}^{N-1}$, where $\mathbf{U}_k$ is an $N_y \times N_h$ Toeplitz matrix constructed from the STFT coefficients of the noise signal u. Then, we can represent (10) and (11) in a matrix form as

$$\mathbf{y}_k = \mathbf{d}_k + \mathbf{v}_k = \Delta_k^x \mathbf{h}_k + \mathbf{v}_k \tag{16}$$

$$\mathbf{v}_k = \mathbf{w}_k - \Delta_k^u \mathbf{h}_k \tag{17}$$

where

$$\mathbf{y}_k = \begin{bmatrix} y_{0,k} & y_{1,k} & \cdots & y_{N_y-1} \end{bmatrix}^T \tag{18}$$

being $\mathbf{v}_k$ and $\mathbf{w}_k$ defined similarly.

Identification of the system $\mathbf{h}_k$ from (16) based on several cross-band filters is presented and analyzed extensively in [9]. However, the signal x is assumed to be uncorrelated with the additive noise v, which is clearly not the case in RTF identification as seen in (17). Thus, applying this method to an RTF identification problem leads to a biased estimation.

III. RTF Identification using CTF Approximation

A. The CTF Approximation

In order to simplify the analysis, we consider in (10) and (11) only band-to-band filters (i.e. $k = k'$). Then, (10) and (11) reduce to

$$\begin{aligned} y_{p,k} &= \sum_{p'} x_{p-p',k} h_{p',k,k} + v_{p,k} \\ &= x_{p,k} * h_{p,k,k} + v_{p,k} \end{aligned} \tag{19}$$

$$\begin{aligned} v_{p,k} &= w_{p,k} - \sum_{p'} u_{p-p',k} h_{p',k,k} = \\ &= w_{p,k} - u_{p,k} * h_{p,k,k}. \end{aligned} \tag{20}$$

For more details see [9], where an extensive discussion is given on the STFT domain representation with only a few cross-band filters. In (19) and (20) we have approximated the convolution in the time domain as a convolution between the STFT samples of the input signal and the corresponding band to band filter. Using our previous notation, we can also write (19) and (20) in a matrix form as

$$\mathbf{y}_k = \mathbf{X}_k \mathbf{h}_{k,k} + \mathbf{v}_k \tag{21}$$

$$\mathbf{v}_k = \mathbf{w}_k - \mathbf{U}_k \mathbf{h}_{k,k} \tag{22}$$

B. The Proposed Method

By taking the expectation of the frame by frame multiplication of the two observed signals $y_{p,k}$ and $x_{p,k}$, we obtain from (21)

$$\Phi_{yx}(k) = \Psi_{xx}(k)\mathbf{h}_{k,k} + \Phi_{vx}(k) \tag{23}$$

where $\Psi_{xx}(k)$ is an $N_y \times N_h$ matrix and its (p,l)th term is

$$\left[\Psi_{xx}(k)\right]_{p,l} = E\left\{x_{p-l,k} x^*_{p,k}\right\} \triangleq \psi_{xx}\left(p,l,k\right) \tag{24}$$

and $\Phi_{yx}(k)$ and $\Phi_{vx}(k)$ are $N_y \times 1$ vectors, given as

$$\Phi_{yx}(k) = \left[\begin{array}{ccc} \phi_{yx}(0,k) & \cdots & \phi_{yx}(N_y-1,k) \end{array}\right]^T \tag{25}$$

$$\Phi_{vx}(k) = \left[\begin{array}{ccc} \phi_{vx}(0,k) & \cdots & \phi_{vx}(N_y-1,k) \end{array}\right]^T \tag{26}$$

where $E\{\cdot\}$ denotes mathematical expectation, $\phi_{yx}(p,k)$ denotes the cross PSD between the signals $y(n)$ and $x(n)$, $\phi_{vx}(p,k)$ denotes the cross PSD between the signals $v(n)$ and $x(n)$ and $\psi_{xx}(p,l,k)$ denotes the cross PSD between the signal $x(n)$ and its delayed version $x'(n) \triangleq x(n-lL)$, all at time frame p and frequency k. Since the speech signal $s(n)$ is uncorrelated with the noise signal $u(n)$, by taking mathematical expectation of the cross multiplication of v and x in the STFT domain, we get from (22):

$$\Phi_{vx}(k) = \Phi_{wu}(k) - \Psi_{uu}(k)\mathbf{h}_{k,k} \tag{27}$$

where $\Phi_{wu}(k)$ is an $N_y \times 1$ vector, given as

$$\Phi_{wu}(k) = \begin{bmatrix} \phi_{wu}(k) & \cdots & \phi_{wu}(k) \end{bmatrix}^T \tag{28}$$

and $\Psi_{uu}(k)$ is an $N_y \times N_h$ matrix and its (p, l)th term is given by

$$\begin{aligned} \left[\Psi_{uu}(k)\right]_{p,l} &= E\left\{u_{p-l,k}u^*_{p,k}\right\} \\ &\triangleq \psi_{uu}\left(p, l, k\right) = \psi_{uu}\left(l, k\right) \end{aligned} \tag{29}$$

where $\phi_{wu}(k)$ denotes the cross PSD between the signals $w(n)$ and $u(n)$, and $\psi_{uu}(l, k)$ denotes the cross PSD between the signal $u(n)$ and its delayed version $u'(n) \triangleq u(n - lL)$, both at frequency bin k. It is worth noting that since the noise signals are stationary during our observation interval (it is sufficient to assume that the noise statistics are changing slowly compared to the speech statistics [7]), the noise spectrum terms are independent of the time frame index.

Once again, by exploiting the fact that the speech signal $s(n)$ and the noise signal $u(n)$ are uncorrelated, we obtain from (1)

$$\Psi_{xx}(k) = \Psi_{ss}(k) + \Psi_{uu}(k) \tag{30}$$

where $\Psi_{ss}(k)$ is defined similarly to (24). Substituting (27) into (23) and using (30), we have

$$\Phi_{yx}(k) = \Psi_{ss}(k)\mathbf{h}_{k,k} + \Phi_{wu}(k). \tag{31}$$

Now, writing (31) in terms of the PSD estimates, we have

$$\hat{\Phi}_k = \hat{\Psi}_k\mathbf{h}_{k,k} + \mathbf{e}_k \tag{32}$$

where $\mathbf{e}_k$ denotes the PSD estimation error (See Appendix A), and

$$\hat{\Phi}_k \triangleq \hat{\Phi}_{yx}(k) - \hat{\Phi}_{wu}(k) \tag{33}$$

$$\hat{\Psi}_k \triangleq \hat{\Psi}_{ss}(k) = \hat{\Psi}_{xx}(k) - \hat{\Psi}_{uu}(k). \tag{34}$$

A weighted least square (WLS) solution to (32) is of the form[2]:

$$\hat{\mathbf{h}}_{k,k} = \left(\hat{\Psi}^H_k W_k \hat{\Psi}_k\right)^{-1} \hat{\Psi}^H_k W_k \hat{\Phi}_k \tag{35}$$

where W_k is the weight matrix. This yields the proposed RTF identification method carried out in the STFT domain using the CTF approximation. The suggested estimator requires estimates of the PSD terms $\phi_{yx}(p, k)$, $\phi_{wu}(k)$, $\psi_{xx}(p, l, k)$ and $\psi_{uu}(l, k)$. We can estimate $\hat{\phi}_{yx}(p, k)$ and $\hat{\psi}_{xx}(p, l, k)$ directly from the measurements, whereas the stationary noise signals PSDs $\hat{\psi}_{uu}(l, k)$ and $\hat{\phi}_{wu}(k)$ can be obtained from measurements in passages where the speech signal is absent. In practice we can determine the speech presence probability and use MCRA [15] or IMCRA [16] methods for the PSD estimation.

[2] Assuming $\left(\Psi^H_k W_k \Psi_k\right)$ is not singular. Otherwise, a regularization in needed.

The covariance matrix of $\hat{\mathbf{h}}_{k,k}$ is given by [17]

$$\mathbf{cov}\left\{\hat{\mathbf{h}}_{k,k}\right\} = \left(\Psi_k^H W_k \Psi_k\right)^{-1} \Psi_k^H W_k \mathbf{cov}\left(\mathbf{e}_k\right) \times W_k \Psi_k \left(\Psi_k^H W_k \Psi_k\right)^{-1} \tag{36}$$

where $\mathbf{cov}\left(\mathbf{e}_k\right)$ is the covariance matrix of $\mathbf{e}_k$ and its (p, l)th element is given by (See Appendix A)

$$\left[\mathbf{cov}(\mathbf{e}_k)\right]_{p,l} = \psi_{vv}\left(l - p, k\right)\psi_{ss}\left(p, p - l, k\right) \tag{37}$$

and $\psi_{ss}(p, l, k)$ and $\psi_{vv}(l, k)$ are defined similarly to (24) and (29) respectively. According to [18], the weight matrix W_k that minimizes the estimator variance is

$$W_k = \left(\mathbf{cov}\left(\mathbf{e}_k\right)\right)^{-1}. \tag{38}$$

Substituting (38) into (35) yields

$$\hat{\mathbf{h}}_{k,k} = \left(\hat{\Psi}_k^H \left(\mathbf{cov}\left(\mathbf{e}_k\right)\right)^{-1} \hat{\Psi}_k\right)^{-1} \hat{\Psi}_k^H \left(\mathbf{cov}\left(\mathbf{e}_k\right)\right)^{-1} \hat{\Phi}_k. \tag{39}$$

The proposed estimator in (39) is often referred to as the best linear unbiased estimator (BLUE) [17]. By substituting (38) into (36), we obtain the variance of the proposed estimator

$$\mathbf{cov}\left\{\hat{\mathbf{h}}_{k,k}\right\} = \left(\Psi_k^H \left(\mathbf{cov}\left(\mathbf{e}_k\right)\right)^{-1} \Psi_k\right)^{-1}. \tag{40}$$

C. Particular Case

When the STFT samples of the signals are uncorrelated, i.e.

$$\psi_{xx}(p, l, k) = \phi_{xx}(p, k)\delta(l) \tag{41}$$

$$\psi_{uu}(p, l, k) = \phi_{uu}(k)\delta(l) \tag{42}$$

We can substitute (41) and (42) into (23) and (27) yielding

$$\Phi_{yx}(k) = \Phi_{xx}(k)h_{0,k,k} + \Phi_{vu}(k) \tag{43}$$

$$\Phi_{vu}(k) = \Phi_{wu}(k) - \Phi_{uu}(k)h_{0,k,k}. \tag{44}$$

In this case $h_k \triangleq h_{0,k,k}$ can be regarded as the multiplicative transfer function for each frequency bin k. The proposed estimator and the estimation variance are (See Appendix B) respectively

$$\hat{h}_{0,k,k} = \frac{\langle \hat{\phi}_{yx}(p, k) - \hat{\phi}_{wu}(k)\rangle_p}{\langle \hat{\phi}_{ss}(p, k)\rangle_p} \tag{45}$$

$$\mathbf{var}\left\{\hat{h}_{0,k,k}\right\} = \frac{\phi_{vv}(k)}{N_y \langle \phi_{ss}(p, k)\rangle_p} \tag{46}$$

where $\langle\cdot\rangle_p$ is an average operator over the time frame p.

These results coincide with the estimator and estimation variance under the MTF assumption introduced in [7]. It is worthwhile noting that when using the MTF approximation and setting the time frame to be larger than the support of the acoustic impulse response, the assumption that the STFT samples of the signals are uncorrelated

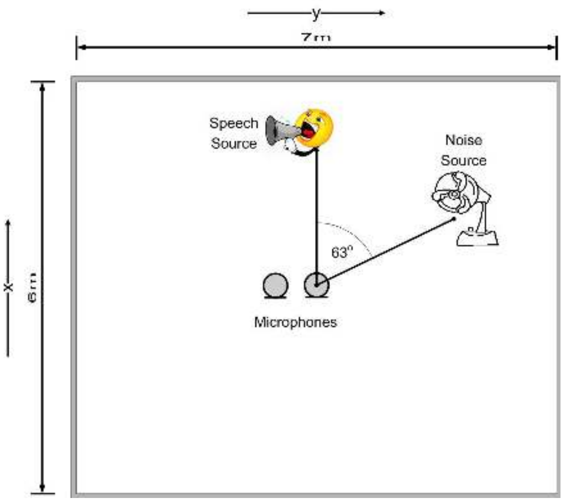


Fig. 2. Experimental setup.

becomes more accurate. In addition, we obtain the same results also in case the cross-band filters contain a single tap (i.e. $N_h = 1$).

IV. EXPERIMENTAL RESULTS

In this section we evaluate the performance of the proposed method using the CTF approximation, and compare it with Cohen's RTF identification method [7] using the MTF approximation in various environments.

In the following experiments we use Habets' simulator [19] for simulating acoustic impulse responses, based on Allen and Berkley's image method [20]. The responses are measured in a rectangular room, 6 m wide by 7 m long and 2.75 m high. We locate the primary microphone at the center of the room, at $(3\text{m}, 3.5\text{m}, 1.375\text{m})$, and the reference microphone at $(3\text{m}, 3.5\text{m} - d, 1.375\text{m})$ with several spacings d. A speech source at $(5\text{m}, 3.5\text{m}, 1.375\text{m})$ is 2 m distant from the primary microphone[3], and a noise source is placed at $(4\text{m}, 5.5\text{m}, 1.375\text{m})$. Figure 2 shows an illustration of the room setup. In each experiment this setup (the speech and noise sources and the two microphones) is rotated 16 times around the center of the room in azimuth steps of 22.5° (with respect to the room floor) and the results are obtained by averaging over these rotated setups.

The signals are sampled at 8 kHz. The speech source signal is a recorded speech from TIMIT database [21] and the noise source signal is a computer generated white zero mean Gaussian noise with variance that varies to control the SNR level. It is worthwhile noting that we obtained similar results using recorded (colored) stationary noise signals. The microphone measurements are generated by convolving the source signals with the corresponding simulated impulse responses. The STFT is implemented using Hamming windows of length $N = 512$ with 75% overlap. The relative impulse response is infinite but both methods approximate it as a finite response filter. Under

[3] Creating a far-end field configuration.

the MTF approximation, the RTF length is determined by the length of the time frame, whereas under the CTF approximation the RTF length can be set as desired. In the following experiments we set the estimated RTF length to be $1/8$ of the room reverberation time T_{60}. This particular ratio was set since empirical tests produced satisfactory results. In addition, we used a short period of noise-only signal at the beginning of each experiment for estimating the noise signals PSD. In practice, it can be performed adaptively using a VAD based on MCRA [15] or IMCRA [16] methods.

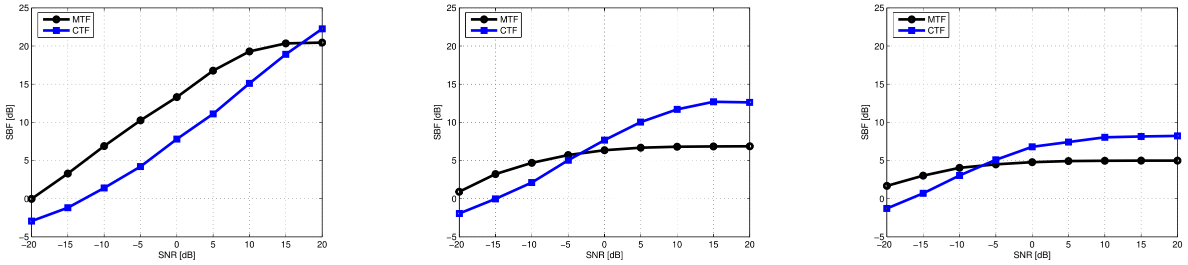
For evaluating the identification performance, we use a measure of the signal blocking factor (SBF) defined by

$$\text{SBF} = 10 \log_{10} \frac{E\left\{s^2(n)\right\}}{E\left\{r^2(n)\right\}} \tag{47}$$

where $E\{s^2(n)\}$ is the energy contained in the speech received at the primary sensor, and $E\{r^2(n)\}$ is the energy contained in the leakage signal $r(n) = h(n) * s(n) - \hat{h}(n) * s(n)$. The leakage signal represents the difference between the reverberated speech at the reference sensor and its estimate given the speech at the primary sensor. This parameter indicates the ability to block the desired signal in generalized sidelobe canceler (GSC) techniques and produce reference noise signals [1] [2]. It has a major effect on the amount of signal distortion at an adaptive beamformer output. It is worthwhile noting that the time domain impulse response $\hat{h}(n)$ is not directly reconstructed from the filter estimate in the STFT domain $\hat{\mathbf{h}}_{k,k}$, obtained from (39). First, the output of the convolution $\hat{h}(n) * s(n)$ is calculated in the STFT domain using $\hat{\mathbf{h}}_{k,k}$. Second, the time domain leakage signal $r(n)$ is calculated using inverse STFT.

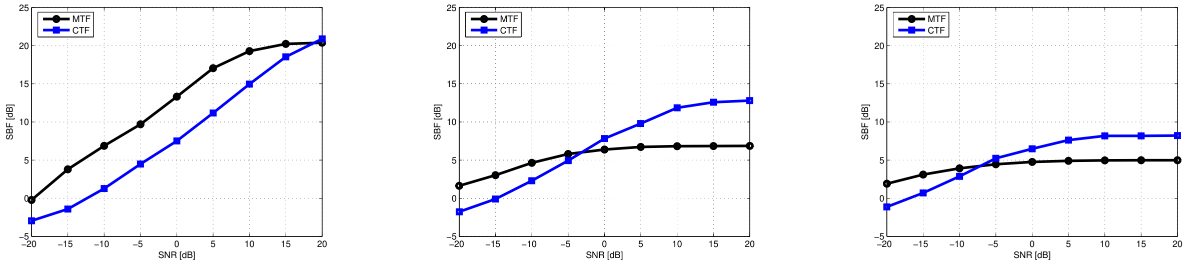
Deriving an explicit expression for the MSE obtained by the proposed estimator, taking the CTF approximation into account, is mathematically untraceable due to the correlation between the additive noise and the reference signal. In case of high SNR level at the primary microphone, the MSE analysis of the system identification in the STFT domain with cross-band filters [9] guarantees better performance for identification based on the CTF approximation rather than identification that relies on the MTF model. While the model complexity increases under the CTF approximation, as the SNR level increases and the data becomes more reliable, a larger number of parameters can be accurately estimated, thus enabling better identification.

Figure 3(a)-(c) shows the SBF curves obtained by both methods as a function of the SNR at the primary microphone. We observe that the RTF identification based on CTF approximation achieves higher SBF than the RTF identification based on MTF approximation in higher SNR conditions, whereas, the RTF identification that relies on MTF model achieves higher SBF in lower SNR conditions. Since the RTF identification using CTF model is associated with greater model complexity, it requires more reliable data, meaning, higher SNR values. In addition, as the environment becomes more reverberant, the intersection point value between the SBF curves decreases, implying that the RTF identification using CTF model outperforms the RTF identification based on MTF model starting from lower SNR conditions. In Fig. 3(a), the reverberation time is $T_{60} = 0.125$s and the intersection point between the SBF curves is at SNR of 17dB. As the reverberation time increases in Fig. 3(b) and (c) ($T_{60} = 0.25$s and $T_{60} = 0.5$s, respectively), the intersection point values decrease to lower SNR values (-3dB and -6dB, respectively). We also observe that the gain for 20 dB SNR is much higher in the case of $T_{60} = 0.25$s



(a) (b) (c)

Fig. 3. SBF curves obtained by using the MTF and CTF approximations under various SNR conditions. The time frame length is $N = 512$ with 75% overlap, and the distance between the primary and reference microphones is $d = 0.3$m. (a) Reverberation time $T_{60} = 0.125$s. (b) Reverberation time $T_{60} = 0.25$s. (c) Reverberation time $T_{60} = 0.5$s.

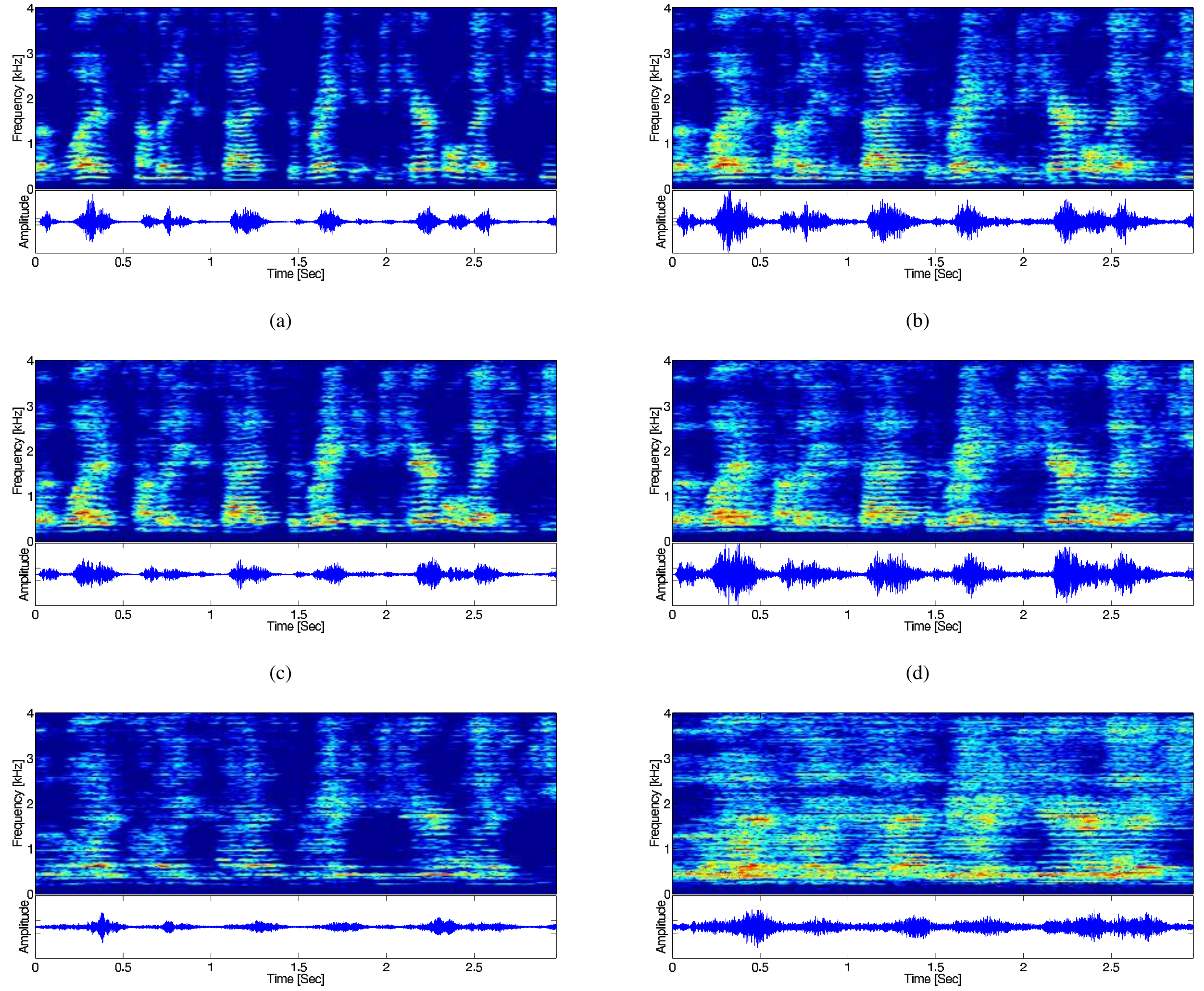


(a) (b) (c)

Fig. 4. SBF curves under the same setup as of Fig. 3 with additional uncorrelated Gaussian noise. (a) Reverberation time $T_{60} = 0.125$s. (b) Reverberation time $T_{60} = 0.25$s. (c) Reverberation time $T_{60} = 0.5$s.

than in the case of $T_{60} = 0.5$s. In the case of $T_{60} = 0.5$s (i.e longer impulse response) the model mismatch using only a single band-to-band filter is larger than the model mismatch in the case of $T_{60} = 0.25$s. Thus, in order to obtain larger gain in the later case, more cross-band filters should be employed to represent the system. More details and analytic analysis is presented in [9]. Generally the microphones introduce additional static noise into the measurements. We demonstrate the robustness of the proposed method in Fig. 4(a)-(c), where we repeat the last experiment with additional uncorrelated additive Gaussian noise. We observe that the improvement of the RTF identification based on the CTF method is slightly degraded (e.g. in Fig. 4(a) the intersection point is moved to the right, compared with Fig. 3(a)). The additional additive uncorrelated noise reduces the effective SNR of the RTF identification and thus, as previously claimed, the RTF identification that relies on the CTF approximation becomes less advantageous.

Figure 5(a)-(f) shows waveforms and spectrograms of the speech and leakage signals obtained by the proposed and competing methods. In Fig. 5(c) and (e) we observe that the leakage signal obtained by the RTF identification that relies on CTF approximation is much lower than the leakage signal obtained by the RTF identification based on MTF approximation. Similar results are obtained in Fig. 5(d) and (f) where the reverberation time is longer, and



(a)

(b)

(c)

(d)

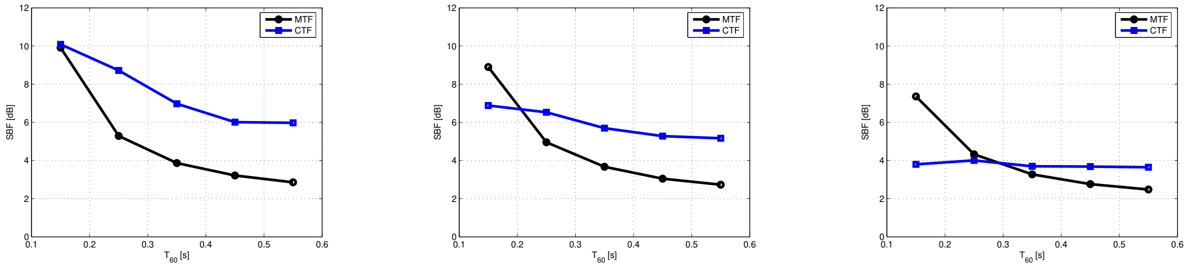
(e)

(f)

Fig. 5. Waveforms and spectrograms obtained under SNR = 15dB. The time frame length is $N = 512$ with 75% overlap, and the distance between the primary and reference microphones is $d = 0.3$m. (a) Speech signal $s(n)$ with reverberation time $T_{60} = 0.25$s. (b) Speech signal $s(n)$ with reverberation time $T_{60} = 0.5$s. (c) Leakage signal $r(n)$ based on the MTF model with reverberation time $T_{60} = 0.25$s. (d) Leakage signal $r(n)$ based on the MTF model with reverberation time $T_{60} = 0.5$s. (e) Leakage signal $r(n)$ based on the CTF model with reverberation time $T_{60} = 0.25$s. (f) Leakage signal $r(n)$ based on the CTF model with reverberation time $T_{60} = 0.5$s.

hence, the leakage signals have greater amplitudes in comparison with Fig. 5(c) and (e).

Figure 6(a)-(c) shows the SBF curves obtained as a function of the reverberation time. Increasing the reverberation time results in longer acoustic impulse responses, and consequently the RTF identification using CTF approximation yields higher SBF than that obtained by the RTF identification based on MTF approximation. On the other hand, the RTF identification using MTF model performs better than the RTF identification using CTF model in less reverberant environments. In addition, the higher the SNR conditions are, the more advantangeous the RTF identification based on CTF model is. In Fig 4(a), where the SNR value is 5dB, the RTF identification using CTF approximation

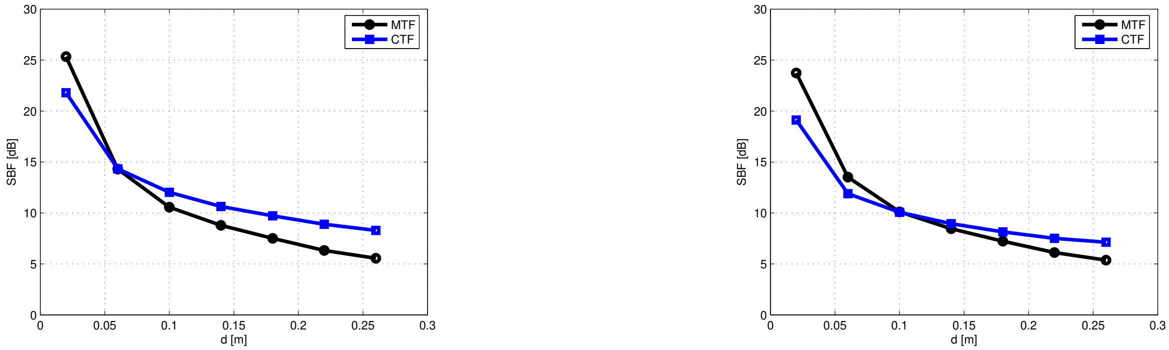


(a)

(b)

(c)

Fig. 6. SBF curves for the compared methods in various T_{60} conditions. The time frame length is $N = 512$ with 75% overlap, and the distance between the primary and reference microphones is $d = 0.3$m. (a) SNR = 5dB. (b) SNR = 0dB. (c) SNR = -5dB.



(a)

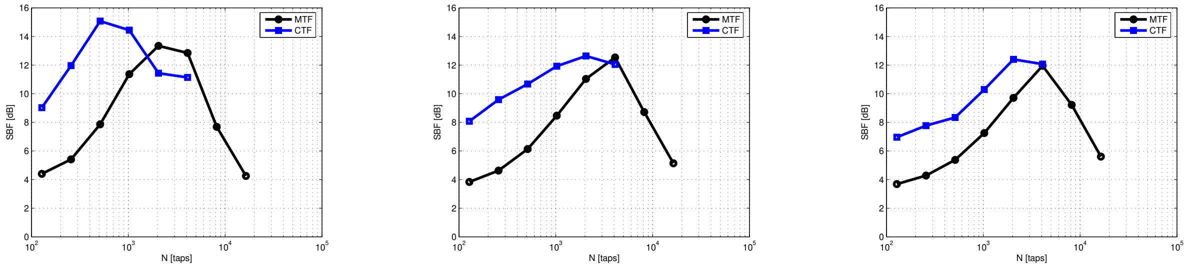
(b)

Fig. 7. SBF curves for the compared methods in various distances between the primary and reference microphones d. The time frame length is $N = 512$ with 75% overlap, and the reverberation time is $T_{60} = 0.5$s. (a) SNR = 5dB. (b) SNR = 0dB.

outperforms the RTF identification that relies on MTF approximation. However, in Fig. 6(b) and (c), where the SNR values are lower (0dB and -5dB, respectively), the RTF identification based on CTF model yields better results when the reverberation times are long enough (the intersection points values between the SBF curves are at 0.2s and 0.3s, respectively).

Figure 7(a)-(b) shows the SBF curves obtained as a function of the distance between the primary and reference microphones d. The coupling between the microphones becomes more complicated as the distance between the microphones increases. Hence, the RTF is more difficult to identify and requires longer FIR representation. The RTF identification that relies on CTF model performs better than the RTF identification using MTF approximation when the distance between the microphones is large. A comparison of Fig. 7(a) and (b) indicates that the intersection point between the curves decreases as the SNR increases.

In the following experiment we compare the competing methods for various time frame lengths. Under the MTF approximation, longer time frames enable identification of a longer RTF at the expense of fewer observations in each frequency bin. Thus, under the MTF model, controlling the time frame length controls both the representation



(a)

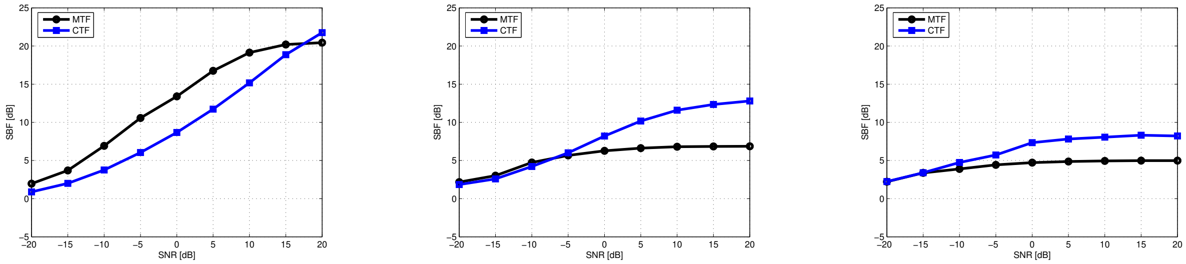
(b)

(c)

Fig. 8. SBF curves for the compared methods using various time frame lengths N. The SNR level is 15dB, and the distance between the primary and reference microphones is $d = 0.3$m. (a) Reverberation time $T_{60} = 0.2$s. (b) Reverberation time $T_{60} = 0.3$s. (c) Reverberation time $T_{60} = 0.4$s.

of the data in the STFT domain and the estimated RTF. On the other hand, under the CTF model, the length of the estimated RTF can be set independently from the time frame length. Thus, under the CTF approximation, controlling the time frame length controls only the representation of the data in the STFT domain. Figure 8(a)-(c) shows the SBF curves obtained by the proposed and competing methods as a function of the time frame length N with a fixed 75% overlap. It is worthwhile noting that this experiment is most favorable to the competing method since the number of variables under the MTF model increases as the time frame increases, while the number of estimated variables under the CTF model is fixed (since the RTF length is fixed, longer time frame yields shorter band-to-band filters). We observe that using the RTF identification method based on MTF model requires longer time frames for longer T_{60} in order to achieve optimal performance. In addition, we observe a trade-off as the time frame increases between increasing the length of the estimated RTF and decreasing the estimation variance. Similar trade-off can be observed for the RTF identification that relies on CTF approximation. As the time frame length increases, the band-to-band filters become shorter and easier to identify, whereas less frames of observations are available. This trade-off between the length of the band-to-band filters and the number of data frames is studied for the general system identification case in [9]. We can also observe that the optimal performance of the RTF identification method under the CTF approximation is achieved using shorter time frames compared with the optimal performance achieved by the RTF identification method that relies on the MTF model. The RTF identification method based on CTF approximation performs better using short time frames, which enable greater flexibility and reduced computational complexity. In addition, the RTF identification method under the MTF approximation doesn't reach the optimal performance of the RTF identification method under the CTF model. Since the model mismatch using the MTF approximation is too large, it cannot be compensated by taking longer time frames and estimating more variables. On the other hand, the CTF approximation enables better representation of the input data by appropriately adjusting the length of time frames, while the estimated RTF length is set independently according to the reverberation time.

Now, we demonstrate the performance of the proposed method in the presence of diffused noise, which is used to model many practical noise fields, e.g. a moving car interior. The diffused noise is simulated as a spherical



(a)

(b)

(c)

Fig. 9. SBF curves for the compared methods under various SNR conditions with diffused noise. The time frame length is $N = 512$ with 75% overlap, and the distance between the primary and reference microphones is $d = 0.3$m. (a) Reverberation time $T_{60} = 0.125$s. (b) Reverberation time $T_{60} = 0.25$s. (c) Reverberation time $T_{60} = 0.5$s.

noise field according to [22], [23]. Figure 9(a)-(c) shows the SBF curves obtained as a function of the SNR at the primary microphone in the presence of diffused noise. The performance of both proposed and competing methods in the presence of diffused noise is similar to the performance achieved in the presence of directional noise in Fig. 3(a)-(c). We observe that both methods show increased SBF in low SNR values and that the RTF identification using CTF model becomes advantageous starting from lower SNR levels (the intersection points between the curves is shifted to the left, compared with Fig. 3).

V. CONCLUSION

We have proposed a relative transfer function identification method for speech sources in reverberant environments. The identification is carried out in the STFT domain, without using the common and restrictive MTF approximation. Instead, we have used the convolutive transfer function approximation, which supports the representation of long transfer functions with short time frames. An unbiased estimator for the RTF was developed and analytic expressions for its variance were presented. We have investigated the performance of the proposed method in various acoustic environments, and demonstrated improved RTF identification when the SNR is high or when the time variations of the transfer functions are relatively slow. The input signal used for the RTF identification is of finite length to enable tracking of time variations. Hence, RTF identification that relies on the MTF approximation is significantly influenced by the time frame length. Long time frames enable identification of a long RTF, but then fewer observations are available in each frequency bin, which may increase the estimation variance. The proposed algorithm, on the other hand, enables better representation of the input data by appropriately adjusting the length of time frames, and better RTF identification by appropriately adjusting the length of the RTF in each subband. Following the attractive results, we intend to develop an adaptive solution, in order to support dynamic environments, and to incorporate the proposed identification method into a beamforming application.

APPENDIX A
DERIVATION OF (37)

From (31) and (32) we get

$$\mathbf{e}_k = \left(\hat{\Phi}_{yx}(k) - \Phi_{yx}(k)\right) - \left(\hat{\Phi}_{wu}(k) - \Phi_{wu}(k)\right) \\ - \left(\hat{\Psi}_{ss}(k) - \Psi_{ss}(k)\right) \mathbf{h}_{k,k}. \tag{48}$$

Using (4), (23) and (30), we have

$$\mathbf{e}_k = \left(\hat{\Phi}_{vx}(k) - \Phi_{vx}(k)\right) - \left(\hat{\Phi}_{vu}(k) - \Phi_{vu}(k)\right) \\ = \left(\hat{\Phi}_{vs}(k) - \Phi_{vs}(k)\right) \tag{49}$$

where $\Phi_{vs}(k)$ and $\Phi_{vu}(k)$ are defined similarly to (26) and (28) respectively. Now, assuming the STFT samples have zero mean and using the fact that $v(n)$ is a noise only signal uncorrelated with $s(n)$, we get

$$\phi_{vs}(p,k) = E\left\{v_{p,k} s^*_{p,k}\right\} = 0. \tag{50}$$

Thus, the cross PSD estimation using cross periodograms yields

$$\mathbf{cov}\left(\hat{\phi}_{vs}(p,k)\hat{\phi}^*_{vs}(p',k)\right) = E\left\{v_{p,k} s^*_{p,k} v^*_{p',k} s_{p',k}\right\} \\ = E\left\{v_{p,k} v^*_{p',k}\right\} E\left\{s^*_{p,k} s_{p',k}\right\} \\ = \psi_{vv}\left(p'-p,k\right) \psi_{ss}\left(p,p-p',k\right) \tag{51}$$

where * represents complex conjugation and $\psi_{ss}(p,l,k)$ and $\psi_{vv}(l,k)$ are defined similarly to (24) and (29) respectively. Finally, by combining (49) and (51), we obtain (37).

APPENDIX B
DERIVATION OF (45)-(46)

Similarly to (41) and (42) we get

$$\psi_{ss}(p,l,k) = \phi_{ss}(p,k)\delta(l) \tag{52}$$

$$\psi_{vv}(p,l,k) = \phi_{vv}(k)\delta(l). \tag{53}$$

By substituting (52) and (53) into (37), we have that $\mathbf{cov}\left(\mathbf{e}_k\right)$ is a diagonal matrix and its pth diagonal term is

$$\left[\mathbf{cov}\left(\mathbf{e}_k\right)\right]_{p,p} = \phi_{vv}(k)\phi_{ss}(p,k) \tag{54}$$

which is the cross PSD estimation variance of $\hat{\phi}_{vs}(k)$ using cross periodograms [18]. Thus, from (39) and (40) using (54) we get

$$\hat{h}_{0,k,k} = \left(\sum_p \left[\hat{\Phi}^H_{ss}(k)\right]_p \left[\mathbf{cov}\left(\mathbf{e}_k\right)\right]^{-1}_{p,p} \left[\hat{\Phi}_{ss}(k)\right]_p\right)^{-1} \\ \times \sum_p \left[\hat{\Phi}^H_{ss}(k)\right]_p \left[\mathbf{cov}\left(\mathbf{e}_k\right)\right]^{-1}_{p,p} \left[\hat{\Phi}_{yx}(k) - \hat{\Phi}_{wu}(k)\right]_p \tag{55}$$

$$\mathbf{var}\left\{\hat{h}_{0,k,k}\right\} = \left(\sum_{p}\left[\Phi_{ss}^{H}(k)\right]_{p}\left[\mathbf{cov}\left(\mathbf{e}_{k}\right)\right]_{p,p}^{-1}\left[\Phi_{ss}(k)\right]_{p}\right)^{-1}. \tag{56}$$

Now, by substituting the elements of $\mathbf{cov}\left(\mathbf{e}(k)\right)$ and $\Phi_{ss}(k)$ into (55) and (56), we obtain

$$\hat{h}_{0,k,k} = \frac{\sum_{p}\left(\hat{\phi}_{yx}(p,k) - \hat{\phi}_{wu}(k)\right)}{\sum_{p}\hat{\phi}_{ss}(p,k)} \tag{57}$$

$$\mathbf{var}\left\{\hat{h}_{0,k,k}\right\} = \left(\frac{1}{\phi_{vv}(k)}\sum_{p}\phi_{ss}(p,k)\right)^{-1}. \tag{58}$$

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their constructive comments and helpful suggestions. The authors would also like to thank Dr. Emanuel Habets for his advise on room impulse response simulations.

REFERENCES

[1] S. Gannot, D. Burshtein, and E. Weinstein, “Signal enhancement using beamforming and nonstationarity with applications to speech,” *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[2] S. Gannot and I. Cohen, “Speech enhancement based on the general transfer function GSC and postfiltering,” *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 6, pp. 561–571, Nov 2004.

[3] T. G. Dvorkind and S. Gannot, “Time difference of arrival estimation of speech source in a noisy and reverberant environment,” *Signal Processing*, vol. 85, pp. 177–204, 2005.

[4] J. Chen, J. Benesty, and Y. Huang, “A minimum distortion noise reduction algorithm with multiple microphones,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, pp. 481–493, Mar. 2008.

[5] E. Warsitz, A. Krueger, and R. Haeb-Umbach, “Speech enhancement with a new generalized eigenvector blocking matrix for application in a generalized sidelobe canceller,” in *in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Las Vegas*, 2008, pp. 73–76.

[6] O. Shalvi and E. Weinstein, “System identification using nonstationary signals,” *IEEE Transactions on Signal Processing*, vol. 40, no. 8, pp. 2055–2063, Aug 1996.

[7] I. Cohen, “Relative transfer function identification using speech signals,” *IEEE Transactions on Speech and Audio Processings*, vol. 12, no. 5, pp. 451–459, Sep 2004.

[8] Y. Avargel and I. Cohen, “On multiplicative transfer function approximation in the short time Fourier transform domain,” *IEEE Signal Processing Letters*, vol. 14, pp. 337–340, 2007.

[9] ——, “System identification in the short time Fourier transform domain with crossband filtering,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, May 2007.

[10] E. Habets, “Single- and multi-microphone speech dereverberation using spectral enhancement,” Ph.D. dissertation, Technische Universiteit Eindhoven, The Netherlands, Jun 2007.

[11] I. 3382:1997, “Acoustics - Measurement of the reverberation time of rooms with reference to other acoustical parameters,” 1997.

[12] M. Portnoff, “Time frequency representation of digital signals and systems based on short-time Fourier analysis,” *IEEE Transactions on Signal Processing*, vol. ASSP-28, no. 1, pp. 55–69, Feb 1980.

[13] S. Farkash and S. Raz, “Linear systems in Gabor time-frequency space,” *IEEE Transactions on Signal Processing*, vol. 42, no. 3, pp. 611–617, Jan 1994.

[14] A. Gilloire and M. Vetterli, “Adaptive filtering in subbands with critical sampling: Analysis, experiments and applications to acoustic echo cancellation,” *IEEE Transactions on Signal Processing*, vol. 40, no. 8, pp. 1862–1875, Aug 1992.

[15] I. Cohen, “Noise estimation by minima controlled recursive averaging for robust speech enhancement,” *IEEE Signal Processing Letters*, vol. 9, pp. 12–15, Jan 2002.

[16] ——, “Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, Sep 2003.

[17] S. M. Kay, *Fundamentals of Statistical Signal Processing*, A. V. Oppenheim, Ed. Prentice Hall, 1993, vol. I.

[18] D. Manolakis, V. Ingle, and S. Kogan, *Statistical and Adaptive Signal Processing: Spectral Estimation, Signal Modeling, Adaptive Filtering and Array Processing*. New York: McGraw-Hill, 2000.

[19] E. A. P. Habets, "Room impulse response (RIR) generator," http://home.tiscali.nl/ehabets/rir_generator.html, Jul. 2006.

[20] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, p. 943950, 1979.

[21] J. S. Garofolo, "DARPA TIMIT acoustic-phonetic continous speech corpus cd-rom," National Inst. of Standards and Technology, Gaithersburg, MD, Feb 1993.

[22] N. Dal-Degan and C. Prati, "Acoustic noise analysis and speech enhancement techniques for mobile radio applications," *Signal Processing*, vol. 15, pp. 43–56, 1988.

[23] E. A. P. Habets, I. Cohen, and S. Gannot, "Generating non-stationary multi-sensor signals under a spatial coherence constraint," *to appear in Journal of the Acoustical Society of America*, 2008.

Ronen Talmon received the B.A degree in mathematics and computer science in 2005 from the Open University, Israel. He is currently pursuing the M.Sc degree in electrical engineering at the Technion – Israel Institute of Technology, Haifa, Israel.

From 2000 to 2005, he was a software developer and researcher at a technological unit of the Israeli Defense Forces. His research interests are statistical signal processing, system identification, speech enhancement and array processing.

Israel Cohen (M'01-SM'03) received the B.Sc. (Summa Cum Laude), M.Sc. and Ph.D. degrees in electrical engineering from the Technion – Israel Institute of Technology, Haifa, Israel, in 1990, 1993 and 1998, respectively.

From 1990 to 1998, he was a Research Scientist with RAFAEL research laboratories, Haifa, Israel Ministry of Defense. From 1998 to 2001, he was a Postdoctoral Research Associate with the Computer Science Department, Yale University, New Haven, CT. In 2001 he joined the Electrical Engineering Department of the Technion, where he is currently an Associate Professor. His research interests are statistical signal processing, analysis and modeling of acoustic signals, speech enhancement, noise estimation, microphone arrays, source localization, blind source separation, system identification and adaptive filtering.

Dr. Cohen received in 2005 and 2006 the Technion Excellent Lecturer awards. He served as Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and IEEE SIGNAL PROCESSING LETTERS, and as guest editor of a special issue of the *EURASIP Journal on Advances in Signal Processing* on Advances in Multimicrophone Speech Processing and a special issue of the *EURASIP Speech Communication Journal* on Speech Enhancement. He is a coeditor of the Multichannel Speech Processing section of the *Springer Handbook of Speech Processing* (Springer, 2007).

Sharon Gannot (S'92-M'01-SM'06) received his B.Sc. degree (summa cum laude) from the Technion – Israel Institute of Technology, Haifa, Israel in 1986 and the M.Sc. (cum laude) and Ph.D. degrees from Tel-Aviv University, Israel in 1995 and 2000 respectively, all in electrical engineering.

In the year 2001 he held a post-doctoral position at the department of Electrical Engineering (SISTA) at K.U.Leuven, Belgium. From 2002 to 2003 he held a research and teaching position at the Faculty of Electrical Engineering, Technion-Israel Institute of Technology, Haifa, Israel. Currently, he is a Senior Lecturer at the School of Engineering, Bar-Ilan University, Israel.

Dr. Gannot is an Associate Editor of the EURASIP Journal of Applied signal Processing, an Editor of a special issue on Advances in Multi-microphone Speech Processing of the same journal, a guest editor of ELSEVIER Speech Communication journal and a reviewer of many IEEE journals and conferences. Dr. Gannot is a member of the Technical and Steering committee of the International Workshop on Acoustic Echo and Noise Control (IWAENC) since 2005. His research interests include parameter estimation, statistical signal processing and speech processing using either single- or multi-microphone arrays.

LIST OF TABLES

LIST OF FIGURES