IDENTIFICATION OF THE RELATIVE TRANSFER FUNCTION BETWEEN MICROPHONES IN REVERBERANT ENVIRONMENTS

Ronen Talmon[1]*, Israel Cohen*[1] *and Sharon Gannot*[2]

[1]Department of Electrical Engineering, Technion - IIT, Haifa, 32000, Israel
[2]School of Engineering, Bar-Ilan University, Ramat-Gan, 52900, Israel

ABSTRACT

Recently, a relative transfer function (RTF) identification method based on the convolutive transfer function (CTF) approximation was developed. This method is adapted to speech sources in reverberant environments and exploits the non-stationarity and presence probability of the speech signal. In this paper, we present experimental results that demonstrate the advantages and robustness of the proposed method. Specifically, we show the robustness of this method to the environment and to a variety of recorded noise signals.

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

1. INTRODUCTION

One of the main challenges in identifying the relative transfer function (RTF) is its length, as the duration of the RTF in reverberant environments may reach several thousand taps. Identification of such long filters is computationally demanding and requires a large amount of observations. Therefore, a common approach is to use the multiplicative transfer function (MTF) approximation which enables to replace a linear convolution in the time domain with a scalar multiplication in the short time Fourier transform (STFT) domain [1] [2]. Unfortunately, this approximation becomes more accurate when the length of a time frame increases, relative to the length of the impulse response [3]. However, long time frames may increase the estimation variance, increase the computational complexity and restrict the ability to track changes in the RTF.

Recently, an RTF identification method based on the *convolutive* transfer function (CTF) approximation was presented [4]. This approximation enables representation of long impulse responses in the STFT domain using short time frames. Based on the analysis of the system identification in the STFT domain with cross-band filtering [5], it was shown that the CTF approximation becomes more accurate than the MTF approximation, as the signal to noise ratio (SNR) increases. In addition, this method exploits the non-stationarity and presence probability of the speech signal. In this paper, we show experimental results demonstrating the advantages of the RTF estimation method based on the CTF model. In particular, we show the robustness of this method to the environment and to a variety of recorded noise signals.

This paper is organized as follows. In Section 2, we formulate the RTF identification problem in the STFT domain. In Section 3, we review the RTF identification approach suitable for speech sources in reverberant environments. Finally, in Section 4 we present experimental results that demonstrate the advantage of the RTF identification method under the CTF approximation.

2. 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:

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

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

where $h(n)$ represents the coupling of the speech signal to the reference microphone, and * 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 problem is to have a linear time invariant (LTI) system, with an input $x(n)$, output $y(n)$ and additive noise $v(n)$, written as

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

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

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

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 $h(n)$. Thus, applying system identification method to an RTF identification problem leads to a biased estimation.

Dividing the observation interval into N_x overlapping time frames of length N with framing step L, we obtain according to [5] that a filter convolution in the time domain is represented as a sum of N cross-band convolutions in the STFT domain. Accordingly, (3) and (4) can be written in the STFT domain as

$$y_{p,k} = \sum_{k'=0}^{N-1} \sum_{p'} x_{p-p',k'} h_{p',k',k} + v_{p,k} \tag{5}$$

$$v_{p,k} = w_{p,k} - \sum_{k'=0}^{N-1} \sum_{p'} u_{p-p',k'} h_{p',k',k} \tag{6}$$

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

In order to simplify the analysis, we consider in (5) and (6) only band-to-band filters (i.e. $k = k'$). Then, (5) and (6) reduce to

$$y_{p,k} = \sum_{p'} x_{p-p',k} h_{p',k,k} + v_{p,k} \tag{7}$$

$$v_{p,k} = w_{p,k} - \sum_{p'} u_{p-p',k} h_{p',k,k}. \tag{8}$$

In (7) and (8) 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. Let $\mathbf{h}_{k',k}$ denote the cross-band filter from frequency band k' to frequency band k:

$$\mathbf{h}_{k',k} = [h_{0,k',k} \;\; h_{1,k',k} \; \cdots \; h_{N_h-1,k',k}]^T. \tag{9}$$

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 $\mathbf{X}_k$ be an $N_y \times N_h$ Toeplitz matrix constructed from the STFT coefficients of the input signal x in the kth sub-band. Similarly, let $\mathbf{U}_k$ be an $N_y \times N_h$ Toeplitz matrix constructed from the STFT coefficients of the noise signal u. Then, we can write (7) and (8) in a matrix form as

$$\mathbf{y}_k = \mathbf{X}_k \mathbf{h}_{k,k} + \mathbf{v}_k \tag{10}$$

$$\mathbf{v}_k = \mathbf{w}_k - \mathbf{U}_k \mathbf{h}_{k,k} \tag{11}$$

where

$$\mathbf{y}_k = \left[y_{0,k} \;\; y_{1,k} \; \cdots \; y_{N_y-1,k} \right]^T \tag{12}$$

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

3. RTF IDENTIFICATION METHOD

By taking expectation of the cross multiplication of the two observed signals y and x in the STFT domain, we obtain from (10)

$$\Phi_{yx}(k) = \Psi_{xx}(k)\mathbf{h}_{k,k} + \Phi_{vx}(k) \tag{13}$$

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

$$[\Psi_{xx}(k)]_{p,l} = E\left\{ x_{p-l,k} x^*_{p,k} \right\} \triangleq \psi_{xx}(p,l,k) \tag{14}$$

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

$$\Phi_{yx}(k) = \left[\begin{array}{ccc} \phi_{yx}(0,k) & \cdots & \phi_{yx}(N_y-1,k) \end{array} \right]^T \tag{15}$$

$$\Phi_{vx}(k) = \left[\begin{array}{ccc} \phi_{vx}(0,k) & \cdots & \phi_{vx}(N_y-1,k) \end{array} \right]^T \tag{16}$$

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 (11):

$$\Phi_{vx}(k) = \Phi_{wu}(k) - \Psi_{uu}(k)\mathbf{h}_{k,k} \tag{17}$$

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

$$\Phi_{wu}(k) = \left[\begin{array}{ccc} \phi_{wu}(k) & \cdots & \phi_{wu}(k) \end{array} \right]^T \tag{18}$$

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

$$[\Psi_{uu}(k)]_{p,l} = E\left\{ u_{p-l,k} u^*_{p,k} \right\} \triangleq \psi_{uu}(l,k) \tag{19}$$

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, 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 $\Psi_{xx}(k) = \Psi_{ss}(k) + \Psi_{uu}(k)$, where $\Psi_{ss}(k)$ is defined similarly to (14). Thus, from (13) and (17), we have

$$\Phi_{yx}(k) = \Psi_{ss}(k)\mathbf{h}_{k,k} + \Phi_{wu}(k). \tag{20}$$

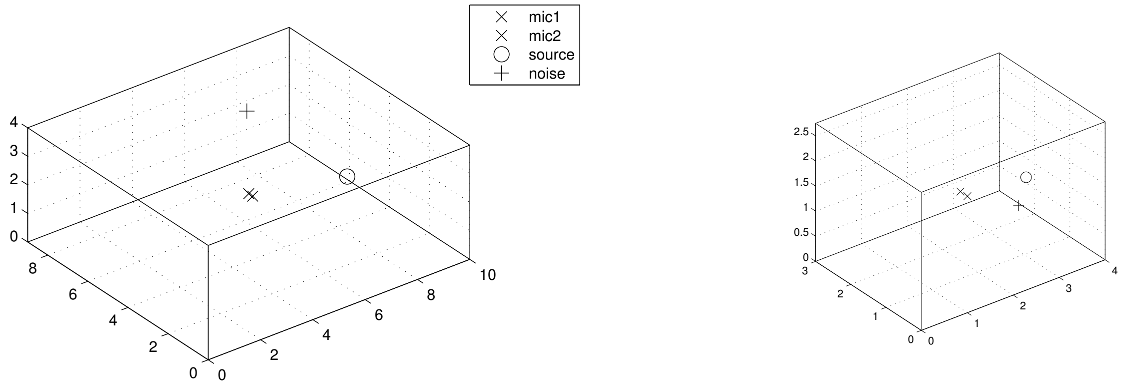
Now, writing (20) in terms of the PSD estimates, we obtain

$$\hat{\Phi}_k = \hat{\Psi}_k \mathbf{h}_{k,k} + \mathbf{e}_k \tag{21}$$

where $\mathbf{e}_k$ denotes the PSD estimation error, and

$$\hat{\Phi}_k \triangleq \hat{\Phi}_{yx}(k) - \hat{\Phi}_{wu}(k) \tag{22}$$

$$\hat{\Psi}_k \triangleq \hat{\Psi}_{ss}(k) = \hat{\Psi}_{xx}(k) - \hat{\Psi}_{uu}(k). \tag{23}$$



(a)

(b)

Fig. 1. Simulated setups illustration. (a) Hall setup. (b) Office setup.

A weighted least square (WLS) solution to (21) is of the form[1]:

$$\hat{\mathbf{h}}_{k,k} = \left(\hat{\Psi}_k^H W_k \hat{\Psi}_k\right)^{-1} \hat{\Psi}_k^H W_k \hat{\Phi}_k \tag{24}$$

where W_k is the weight matrix. This yields an RTF identification estimator carried out in the STFT domain using the CTF approximation. This 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, while, the stationary noise signals PSDs $\hat{\psi}_{uu}(l,k)$ and $\hat{\phi}_{wu}(k)$ can be obtained from silent periods (where the speech signal is absent).

4. EXPERIMENTAL RESULTS

In this section, the RTF identification method using the CTF approximation is tested in simulated environment, and compared with Cohen's competing method [2] using the MTF approximation. A speech source signal drawn from TIMIT database [6] is sampled at 8 kHz and used in the experiments. In addition, three noise signals recorded from NOISEX-92 database [7] are used (Airconditioner noise, destroyer room noise and high frequency noise) with variance that varies to control the SNR level. The STFT is implemented using Hamming windows of length $N = 512$ with 75% overlap. The acoustic room impulse responses are generated using a simulator [8] of Allan and Berkley's image method [9]. The responses are measured in two rectangular rooms. The first is a large hall, 10 m wide by 9 m long and 4 m high with reverberation time set to 400ms. The second is a typical office or a small living room, 3 m wide by 4 m long and 2.75 m high with reverberation time set to 200ms. The primary microphone is located at the center of each room, and the reference microphone with several spacings from it. In addition, a speech source and a noise source are located in both rooms. Figure 1 illustrates the described setups.

For evaluating the identification performance, we use a measure of the signal blocking factor (SBF) [2] [4] defined by

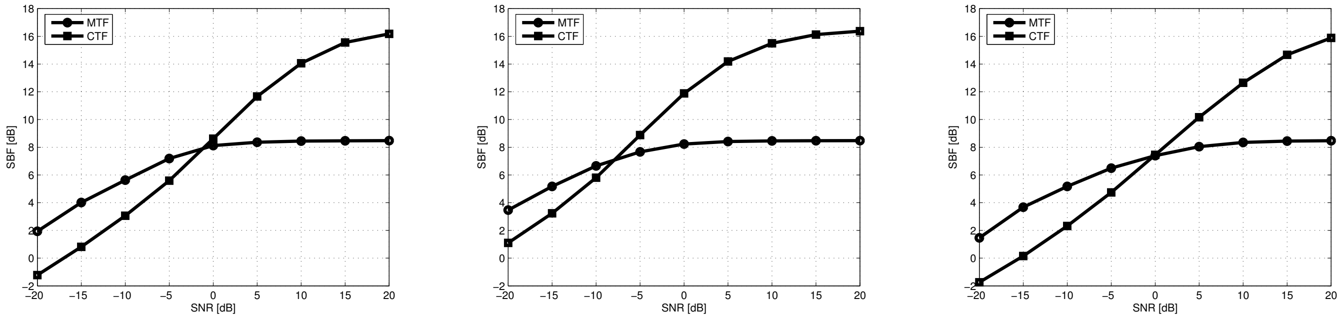
$$\text{SBF} = 10\log_{10}\frac{E\left\{s^2(n)\right\}}{E\left\{r^2(n)\right\}} \tag{25}$$

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 $r(n) = h(n) * s(n) - \hat{h}(n) * s(n)$. This parameter has a major effect on the amount of signal distortion and noise reduction at an adaptive beamformer output.

Figure 2 shows the SBF curves obtained by both methods as a function of the SNR at the primary microphone in the small room. 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. In addition, Fig. 2 shows the RTF identification method robustness to various noise signals, as both performances and intersection points values of the curves are nearly the same in the presence of the three noise signals.

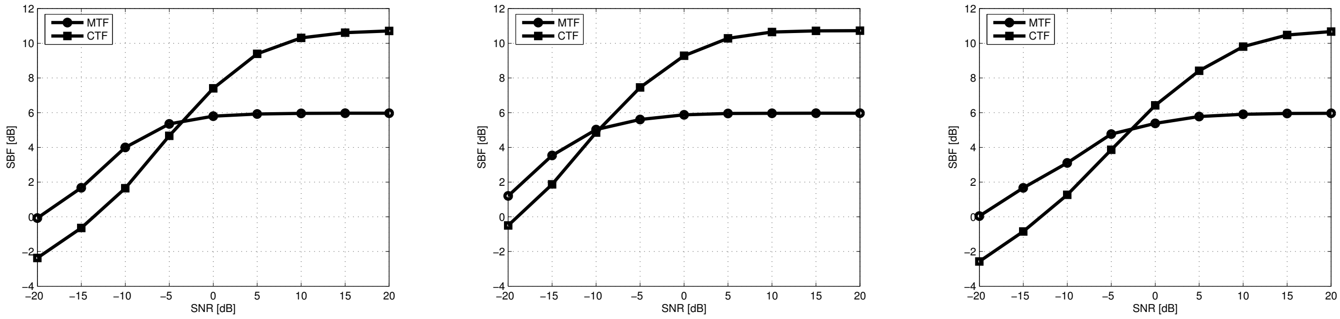
Similar results are obtained in Fig. 3, where the identification is carried out in the large room setup, which demonstrates the RTF identification method robustness to various of setups and room sizes. In the large room the curves shapes and intersection points values have the same characteristics, however both methods achieve lower SBF values as the room dimensions increase and the environment is more reverberant.

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



(a) (b) (c)

Fig. 2. SBF curves obtained under various SNR conditions in the small room setup. The distance between the primary and reference microphones is $d = 0.2$m. (a) Airconditioner noise. (b) Destroyer room noise. (c) High frequency noise.



(a) (b) (c)

Fig. 3. SBF curves obtained under various SNR conditions in the large room setup. The distance between the primary and reference microphones is $d = 0.2$m. (a) Airconditioner noise. (b) Destroyer room noise. (c) High frequency noise.

Figure 4 shows the SBF curves obtained as a function of the distance between the primary and reference microphones denoted by d. The RTF identification method under the CTF approximation becomes more advantageous as the coupling between the microphones becomes more complicated as a result of either larger room dimensions or increased distance between the microphones. In addition, the robustness of the RTF identification to the presence of a variety of noise signals is obtained once again.

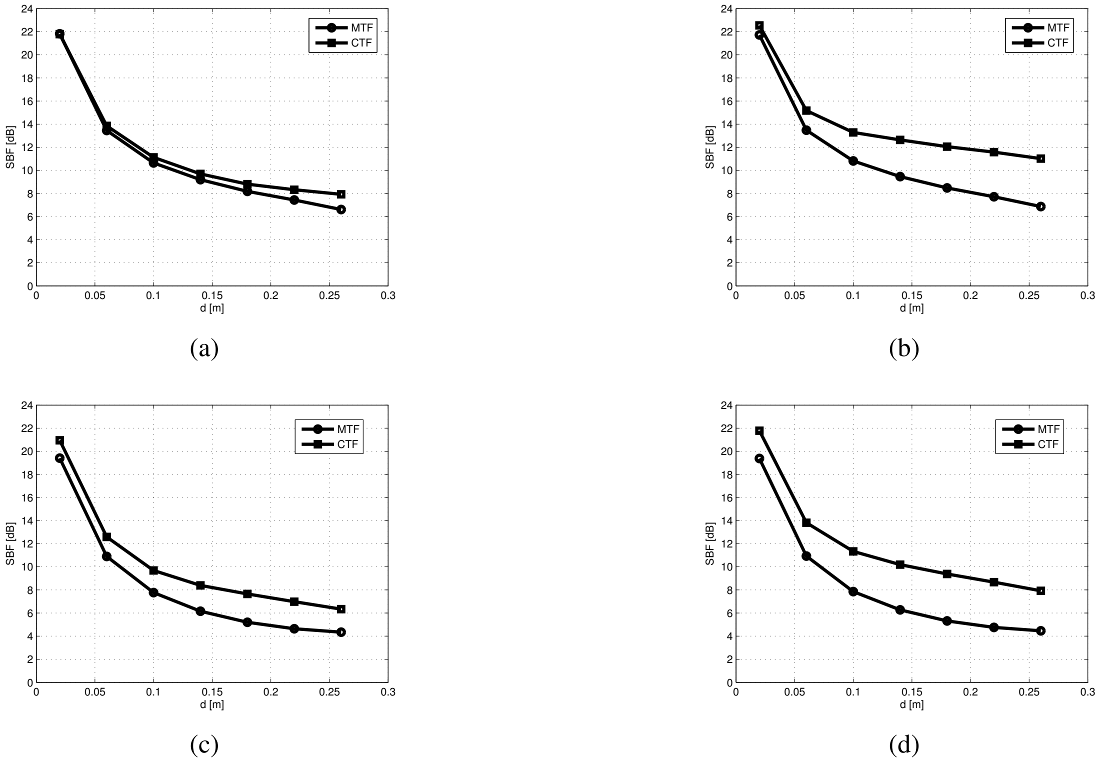
5. DISCUSSION

Investigating the performance of the RTF identification method using the CTF approximation in various acoustic environments showed an improved RTF identification when the SNR is high or when the coupling between the microphones becomes more complicated. Since the RTF identification using CTF model is associated with greater model complexity, it requires more reliable data, meaning, higher SNR values. Furthermore, it was shown that the RTF identification method is robust to a variety of stationary additive noise signals, to the reverberation time and to the room dimensions and setup. It is also worthwhile noting that the RTF identification method under the CTF approximation enables important advantages over competing methods that rely on the MTF approximation. 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 CTF approximation 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.

6. REFERENCES

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

[2] I. Cohen, "Relative transfer function identification using



(a)

(b)

(c)

(d)

Fig. 4. SBF curves for the compared methods in various distances between the primary and reference microphones d in SNR $=$ 0dB. (a) Airconditioner noise in the small room. (b) Destroyer room noise in the small room. (c) Airconditioner noise in the large room. (b) Destroyer room noise in the large room.

speech signals," *IEEE Transactions on Speech and Audio Processings*, vol. 12, no. 5, pp. 451–459, Sep 2004.

[3] 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.

[4] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *submitted to IEEE Transaction on audio, speech and Language*, 2008.

[5] Y. Avargel and I. Cohen, "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.

[6] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic-phonetic continous speech database," National Inst. of Standards and Technology (NIST), Gaithersburg, MD, Feb 1993.

[7] A. Varga and H. J. M. Steeneken, "Assesment of automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," Speech Communication, vol. 12, no. 3, pp. 247-251, Jul 1993.

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

[9] 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, pp. 943–950, 1979.