On Multiplicative Transfer Function Approximation in the Short-Time Fourier Transform Domain

Yekutiel Avargel, *Senior Member, IEEE*, and Israel Cohen, *Senior Member, IEEE*

***Abstract*—The multiplicative transfer function (MTF) approximation is widely used for modeling a linear time invariant system in the short-time Fourier transform (STFT) domain. It relies on the assumption of a long analysis window compared with the length of the system impulse response. In this paper, we investigate the influence of the analysis window length on the performance of a system identifier that utilizes the MTF approximation. We derive analytic expressions for the minimum mean-square error (MMSE) in the STFT domain and show that the system identification performance does not necessarily improve by increasing the length of the analysis window. The optimal window length, that achieves the MMSE, depends on the signal-to-noise ratio and the length of the input signal. The theoretical analysis is supported by simulation results.**

***Index Terms*—Multiplicative transfer function, short-time Fourier transform, system identification.**

I. Introduction

IDENTIFICATION of linear time-invariant (LTI) systems in the short-time Fourier transform (STFT) domain is a fundamental problem in many practical applications [1]–[6]. To perfectly represent an LTI system in the STFT domain, cross-band filters between subbands are generally required [1], [7]. A widely-used approach to avoid the cross-band filters is to approximate the transfer function as multiplicative in the STFT domain. This approximation relies on the assumption that the support of the STFT analysis window is sufficiently large compared with the duration of the system impulse response, and it is useful in many applications, including frequency-domain blind source separation (BSS) [5], acoustic echo cancellation [2], relative transfer function (RTF) identification [3] and adaptive beamforming [6].

As the length of the analysis window increases, the multiplicative transfer function (MTF) approximation becomes more accurate. On the other hand, the length of the input signal that can be employed for the system identification must be finite to enable tracking during time variations in the system. Therefore, increasing the analysis window length while retaining the relative overlap between consecutive windows (the overlap between consecutive analysis windows determines the redundancy of the STFT representation), a fewer number of observations in each frequency-band become available, which increases the variance of the system estimate. Consequently, the mean-square error (MSE) in each subband may not necessarily decrease as we increase the length of the analysis window.

In this paper, we investigate the influence of the analysis window length on the performance of a system identifier that utilizes the MTF approximation. The MTF in each frequency-band is estimated offline using a least squares (LS) criterion. We derive an explicit expression for the MMSE in the STFT domain and show that it can be decomposed into two error terms. The first term is attributable to using a finite-support analysis window. As we increase the support of the analysis window, this term reduces to zero, since the MTF approximation becomes more accurate. However, the second term is a consequence of restricting the length of the input signal. As the support of the analysis window increases, this term increases, since less observations in each frequency-band can be used for the system identification. Therefore, the system identification performance does not necessarily improve by increasing the length of the analysis window. We show that the optimal window length depends on both the signal-to-noise ratio (SNR) and the input signal length. As the SNR or the input signal length increases, a longer analysis window should be used to make the MTF approximation valid and the variance of the MTF estimate reasonably low. The theoretical analysis is supported by simulation results.

The paper is organized as follows. In Section II, we present the MTF approximation and address the relation between the analysis window length and system identification performance. In Section III, we derive an explicit expression for the MMSE obtainable by using the MTF approximation. In Section IV, we investigate the influence of the window length on the MMSE. Finally, in Section V, we present simulation results that verify the theoretical derivations.

II. The MTF Approximation

Let an input $x(n)$ and output $y(n)$ of an unknown LTI system be related by

$$y(n) = h(n) * x(n) + \xi(n) \triangleq d(n) + \xi(n) \tag{1}$$

where $h(n)$ represents the impulse response of the system, $\xi(n)$ is an additive noise signal, $d(n)$ is the signal component in the system output, and $*$ denotes convolution. The STFT of $x(n)$ is given by [8]

$$x_{pk} = \sum_{m} x(m)\, \tilde{\psi}^{*}_{pk}(m) \tag{2}$$

Manuscript received August 24, 2006; revised September 27, 2006. This work supported by the Israel Science Foundation under Grant 1085/05. The associate editor coordinating the review of this paper and approving it for publication was Prof. Alfred Hanssen.

The authors are with the Department of Electrical Engineering, Technion–Israel Institute of Technology, Technion City, Haifa 32000, Israel (e-mail: kutiav@tx.technion.ac.il; icohen@ee.technion.ac.il).

Color versions of Figs. 1 and 2 are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2006.888292

where

$$\tilde{\psi}_{pk}(m) = \tilde{\psi}(m - pL)\, e^{j(2\pi/N)k(m-pL)} \tag{3}$$

denotes a translated and modulated window function, $\tilde{\psi}(n)$ is a real-valued analysis window of length N, p is the frame index, k represents the frequency-bin index, L is a discrete-time shift and * denotes complex conjugation. Applying the STFT to $d(n)$ yields

$$\begin{aligned} d_{pk} &= \sum_m \sum_\ell h(\ell)\, x(m-\ell)\, \tilde{\psi}^*_{pk}(m) \\ &= \sum_m x(m) \sum_\ell h(\ell)\, \tilde{\psi}^*_{pk}(m+\ell)\,. \end{aligned} \tag{4}$$

Let us assume that the analysis window $\tilde{\psi}(n)$ is long and smooth relative to the impulse response $h(n)$ so that $\tilde{\psi}(n)$ is approximately constant over the duration of $h(n)$. Then $\tilde{\psi}(n-m)\, h(m) \approx \tilde{\psi}(n)\, h(m)$, and by substituting (3) into (4), we obtain [9]

$$d_{pk} \approx h_k\, x_{pk} \tag{5}$$

where $h_k \triangleq \sum_m h(m) \exp(-j2\pi mk/N)$. The approximation in (5) is the well-known MTF approximation for modeling an LTI system in the STFT domain. In the limit, for an infinitly long smooth analysis window, the transfer function would be exactly multiplicative in the STFT domain. However, since practical implementations employ finite length analysis windows, the MTF approximation is never accurate.

Let P denote the number of samples in a time-trajectory of x_{pk}, let $\mathbf{x}_k = [\, x_{0,k} \quad x_{1,k} \quad \cdots \quad x_{P-1,k}\,]^T$ denote a time-trajectory of x_{pk} at frequency-bin k, and let the vectors $\mathbf{y}_k$, $\mathbf{d}_k$ and $\boldsymbol{\xi}_k$ be defined similarly. Then,

$$\mathbf{y}_k = \mathbf{d}_k + \boldsymbol{\xi}_k \tag{6}$$

and the MTF approximation can be written in a vector form as

$$\mathbf{d}_k = \mathbf{x}_k\, h_k\,. \tag{7}$$

The LS estimate of h_k is therefore given by

$$\begin{aligned} \hat{h}_k &= \arg\min_{h_k} \|\mathbf{y}_k - \mathbf{x}_k\, h_k\|^2 \\ &= \frac{\mathbf{x}_k^H \mathbf{y}_k}{\mathbf{x}_k^H \mathbf{x}_k}\,. \end{aligned} \tag{8}$$

Clearly, as N, the length of the analysis window, increases, the MTF approximation becomes more accurate. However, the length of the input signal is generally finite[1] and the overlap between consecutive analysis windows is chosen to be fixed (the ratio N/L determines the redundancy of the STFT representation). Hence, increasing N yields shorter time-trajectories (smaller P) and less observations in each frequency-band can be used for the system identification, which increases the variance of $\hat{h}_k$. Therefore, we need to find an appropriate window length, which is sufficiently large to make the MTF approximation valid, and sufficiently small to make the system identification performance most satisfactory. In the following sections, we investigate the relation between the analysis window length and the system identification performance, and show that the optimal window length depends on both the SNR and the input signal length.

[1]Note that the length of the input signal is related to the update rate of $\hat{h}_k$ as we assume that during that period the system remains constant. Therefore, a finite length input signal is practically employed for system identification, to enable tracking the time variations in $h(n)$.

III. MSE Analysis

In this section, we derive an explicit expression for the MMSE in the STFT domain under the assumptions of the MTF approximation and a finite-length input signal. To make the analysis mathematically tractable we assume that the input signal $x(n)$ and the noise signal $\xi(n)$ are uncorrelated zero-mean white Gaussian signals with variances σ_x^2 and σ_ξ^2, respectively. The system identification performance is evaluated using the (normalized) MSE of the output signal in the STFT domain, defined by

$$\epsilon = \frac{\sum_{k=0}^{N-1} E\left\{\left\|\mathbf{d}_k - \hat{\mathbf{d}}_k\right\|^2\right\}}{\sum_{k=0}^{N-1} E\left\{\|\mathbf{d}_k\|^2\right\}}\,. \tag{9}$$

where $\hat{\mathbf{d}}_k = \mathbf{x}_k \hat{h}_k$. Substituting (8) into (9), the MSE can be expressed as

$$\epsilon = 1 + \epsilon_1 - \epsilon_2 \tag{10}$$

where

$$\epsilon_1 = \frac{\sum_{k=0}^{N-1} E\left\{\left(\mathbf{x}_k^H \mathbf{x}_k\right)^{-1} \boldsymbol{\xi}_k^H \mathbf{x}_k \mathbf{x}_k^H \boldsymbol{\xi}_k\right\}}{\sum_{k=0}^{N-1} E\left\{\|\mathbf{d}_k\|^2\right\}} \tag{11}$$

and

$$\epsilon_2 = \frac{\sum_{k=0}^{N-1} E\left\{\left(\mathbf{x}_k^H \mathbf{x}_k\right)^{-1} \mathbf{d}_k^H \mathbf{x}_k \mathbf{x}_k^H \mathbf{d}_k\right\}}{\sum_{k=0}^{N-1} E\left\{\|\mathbf{d}_k\|^2\right\}}\,. \tag{12}$$

Using (4) and the assumption that $x(n)$ is white, we obtain

$$E\left\{\|\mathbf{d}_k\|^2\right\} = P\sigma_x^2 \sum_m r_{\tilde{\psi}}(m)\, r_h(m)\, e^{-j(2\pi/N)km} \tag{13}$$

where $r_f(n) = \sum_m f(n+m) f^*(m)$ denotes the cross-correlation sequence of $f(n)$. Assuming that x_{pk} is variance-ergodic and that P is sufficiently large, so that $(1/P)\sum_{p=0}^{P-1} |x_{pk}|^2 \approx E\left\{|x_{pk}|^2\right\}$, we have

$$\mathbf{x}_k^H \mathbf{x}_k = P\, \sigma_x^2\, r_{\tilde{\psi}}(0)\,. \tag{14}$$

Using the STFT representations of $x(n)$ and $\xi(n)$ (as defined in (2)), it can be verified that

$$\begin{aligned} E\left\{\boldsymbol{\xi}_k^H \mathbf{x}_k \mathbf{x}_k^H \boldsymbol{\xi}_k\right\} &= \sum_{p,p'=0}^{P-1} E\left\{\xi^*_{pk}\xi_{p'k}\right\} E\left\{x_{pk} x^*_{p'k}\right\} \\ &= P\sigma_x^2\sigma_\xi^2 \sum_p r^2_{\tilde{\psi}}(pL)\,. \end{aligned} \tag{15}$$

Substituting (13)–(15) into (11), we obtain

$$\epsilon_1 = \frac{\sigma_\xi^2}{\sigma_x^2} \frac{N \sum_p r_{\tilde{\psi}}^2(pL)}{r_{\tilde{\psi}}(0) \sum_{k=0}^{N-1} \sum_m r_{\tilde{\psi}}(m) r_h(m) e^{-j(2\pi/N)km}}. \quad (16)$$

To simplify the expression for ϵ_2, we substitute the STFT representations of $x(n)$ and $d(n)$ into $E\{\mathbf{d}_k^H \mathbf{x}_k \mathbf{x}_k^H \mathbf{d}_k\} = \sum_{p=0}^{P-1} \sum_{p'=0}^{P-1} E\{d_{pk}^* x_{pk} d_{p'k} x_{p'k}^*\}$, and obtain

$$E\{\mathbf{d}_k^H \mathbf{x}_k \mathbf{x}_k^H \mathbf{d}_k\} = \sum_{p=0}^{P-1} \sum_{m,n} \tilde{\psi}_{pk}(m) \tilde{\psi}_{pk}^*(n) \sum_{p'=0}^{P-1} \sum_{m',n'} \tilde{\psi}_{p'k}(m') \tilde{\psi}_{p'k}^*(n') \times \sum_{i,j} h(m-i) h(n'-j) E\{x(i)x(n)x(j)x(m')\}. \quad (17)$$

Define

$$\theta_k(n) \triangleq \sum_m h(n-m) \tilde{\psi}_{0,k}^*(m) \quad (18)$$

$$\phi_k(n) \triangleq \sum_m \theta_k(n+m) \tilde{\psi}_{0,k}^*(m). \quad (19)$$

Then, using the fourth-order moment factoring theorem for zero-mean real Gaussian samples [10], we can express (17) as

$$E\{\mathbf{d}_k^H \mathbf{x}_k \mathbf{x}_k^H \mathbf{d}_k\} = \sigma_x^4 P^2 \left| \sum_m \theta_k(m) \tilde{\psi}_{0,k}(m) \right|^2 + \sigma_x^4 P \sum_p \phi_k(pL) \phi_k^*(-pL) + \sigma_x^4 P \sum_p r_{\tilde{\psi}}(m) r_h(m) e^{j(2\pi/N)kpL} \quad (20)$$

where we assumed that $\tilde{\psi}(n)$ is a symmetric function (*i.e.*, $\tilde{\psi}(n) = \tilde{\psi}(-n)$). Using (13), (14), and (20) we obtain an explicit expression for ε_2 that, together with ε_1 in (16), can be substituted into (10), which yields

$$\epsilon = 1 - a + \frac{1}{P}\left(\frac{b}{\eta} - c\right) \quad (21)$$

where $\eta \triangleq \sigma_x^2/\sigma_\xi^2$ denotes the SNR and

$$a \triangleq \frac{1}{R} \sum_{k=0}^{N-1} \left| \sum_m \theta_k(m) \tilde{\psi}_{0,k}(m) \right|^2 \quad (22a)$$

$$b \triangleq \frac{N}{R} \sum_p r_{\tilde{\psi}}^2(pL) \quad (22b)$$

$$c \triangleq \frac{1}{R} \cdot \sum_{k=0}^{N-1} \left\{ \sum_p \phi_k(pL) \phi_k^*(-pL) + \sum_p r_{\tilde{\psi}}(pL) r_h(pL) e^{j(2\pi/N)kpL} \right\} \quad (22c)$$

where $R \triangleq r_{\tilde{\psi}}(0) \sum_{k=0}^{N-1} \sum_m r_{\tilde{\psi}}(m) r_h(m) e^{-j(2\pi/N)km}$. Expectedly, we observe from (21) that as the SNR increases, a lower MSE can be achieved.

IV. Optimal Window Length

In this section, we investigate the relation between the length of the analysis window and the MMSE obtainable by using the MTF approximation. Rewrite (21) as

$$\epsilon = \epsilon_N + \epsilon_P \quad (23)$$

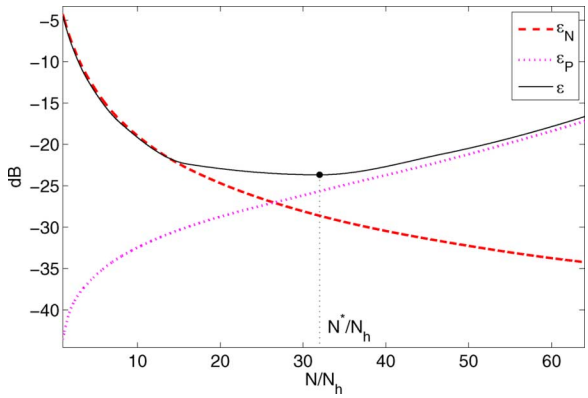

Fig. 1. Theoretical MSE curves as a function of the ratio between the analysis window length (N) and the impulse response length (N_h), obtained for a 0 dB SNR.

where $\epsilon_N = 1 - a$ and $\epsilon_P = (1/P)(b/\eta - c)$. Then, the error ϵ_N is attributable to using a finite-support analysis window. For sufficiently large N, we can apply the approximation $\tilde{\psi}(n-m)h(m) \approx \tilde{\psi}(n)h(m)$ to (22a) and verify that $a = 1$ and $\epsilon_N(N \to \infty) = 0$. On the other hand, the error ϵ_P is a consequence of restricting the length of the input signal. It decreases as we increase P, and reduces to zero when $P \to \infty$.

Fig. 1 shows the MSE curves ϵ, ϵ_N and ϵ_P as a function of the ratio between the analysis window length, N, and the impulse response length, N_h, for a 0-dB SNR (for other simulation parameters see Section V). As expected, we observe that ϵ_N is a monotonically decreasing function of N, while ϵ_P is a monotonically increasing function (since P decreases as N increases). Consequently, the total MSE, ϵ, may reach its minimum value for a certain optimal window length N^*, *i.e.*,

$$N^* = \arg\min_N \epsilon. \quad (24)$$

In the example of Fig. 1, we obtained that N^* is approximately $32\,N_h$.

The optimal window length represents the trade-off between the number of observations in time-trajectories of the STFT representation and accuracy of the MTF approximation. Equation (23) implies that the optimal window length depends on the relative weight of each error, ϵ_N or ϵ_P, in the overall MSE ϵ. Since ϵ_P decreases as we increase either the SNR, η, or the length of the time-trajectories, P, we expect that the optimal window length N^* would increase as η or P increases. Denote by N_x the length of the input signal. Then, the number of samples in a time-trajectory of the STFT representation is $P \approx N_x/L$. For given analysis window and overlap between consecutive windows (given N and N/L), P is proportional to the length of the input signal. Hence, the optimal window length generally increases as N_x increases. Recall that the impulse response is assumed time invariant during N_x samples, in case the time variations in the system are slow, we can increase N_x, and correspondingly increase the analysis window length in order to achieve lower MMSE. These points will be further demonstrated in the next section.

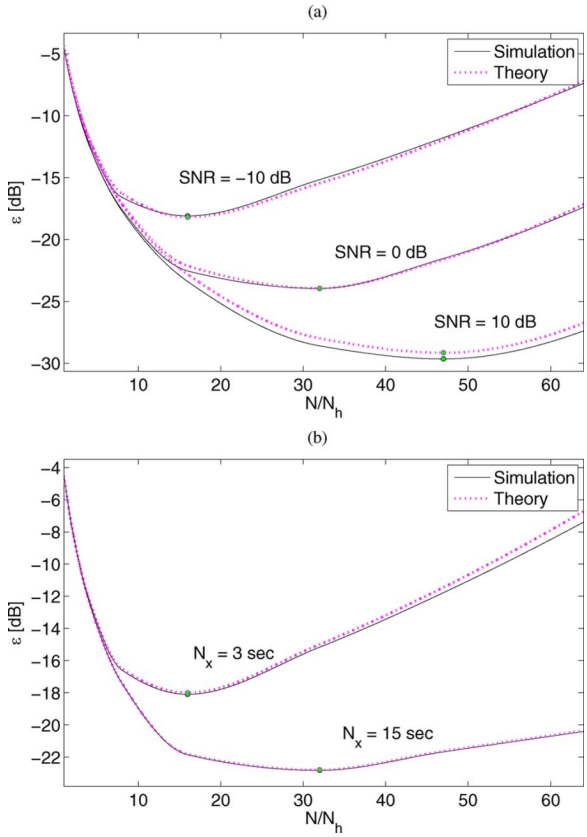

Fig. 2. Comparison of simulation (solid) and theoretical (dashed) MSE curves as a function of the ratio between the analysis window length (N) and the impulse response length (N_h). (a) Comparison for several SNR values (input signal length is 3 s). (b) Comparison for several signal lengths (SNR is -10 dB).

V. Simulation Results

In this section, we present simulation results which verify the theoretical analysis. We use a synthetic room impulse response $h(n)$ based on a statistical reverberation model, which generates a room impulse response as a realization of a nonstationary stochastic process $h(n) = u(n)\beta(n)e^{-\alpha n}$, where $u(n)$ is a step function, $\beta(n)$ is a zero-mean white Gaussian noise and α is related to the reverberation time T_{60} (the time for the reverberant sound energy to drop by 60 dB from its original value). In the following simulations, the length of the impulse response is set to 16 ms, the sampling rate is 16 kHz, α corresponds to $T_{60} = 50$ ms and $\beta(n)$ is unit-variance zero-mean white Gaussian noise. We use a Hamming synthesis window with 50% overlap ($L = 0.5N$), and a corresponding minimum energy analysis window which satisfies the completeness condition [11]. The signals $x(n)$ and $\xi(n)$ are uncorrelated zero-mean white Gaussian. Fig. 2 shows the MSE curves, both in theory and in simulation, as a function of the ratio between the analysis window length and the impulse response length. Fig. 2(a) shows the MSE curves for SNR values of -10, 0 and 10 dB, obtained with a signal length of 3 s (corresponding to $N_x = 48{,}000$), and Fig. 2(b) shows the MSE curves for signal lengths of 3 and 15 s, obtained with a -10 dB SNR. The experimental results are obtained by averaging over 100 independent runs. Clearly, the theoretical analysis well describes the MSE performance achievable by using the MTF approximation. As the SNR or the signal length increases, a lower MSE can be achieved by using a longer analysis window. Accordingly, as the power of the input signal increases or as the time variations in the system become slower (which enables one to use of a longer input signal), a longer analysis window should be used to make the MTF approximation appropriate for system identification in the STFT domain.

VI. Conclusions

We have derived explicit relations between the MMSE and the analysis window length, for a system identifier implemented in the STFT domain and relying on the MTF approximation. We showed that the MMSE does not necessarily decrease with increasing the window length, due to the finite length of the input signal. The optimal window length that achieves the MMSE depends on the SNR and length of the input signal.

It is worthwhile noting, that the stationarity of the input signal should also be taken into account when determining the appropriate window length. For nonstationary input signals it may be necessary to use a shorter analysis window for more efficient representation in the STFT domain. Furthermore, the performance analysis is evaluated based on a normalized MSE in the STFT domain. One may also be interested to analyze the MSE in the time-domain, which is a topic for further research.

Acknowledgment

The authors thank the anonymous reviewers for their helpful comments. They also thank Phoenix Audio Technologies for providing audio equipment and for their helpful technical support.

References

[1] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, to be published.

[2] C. Faller and J. Chen, "Suppressing acoustic echo in a spectral envelope space," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 13, no. 5, pp. 1048–1062, Sep. 2005.

[3] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. Speech Audio Process. (Special Issue on Multichannel Signal Processing for Audio and Acoustics Applications)*, vol. 12, no. 5, pp. 451–459, Sep. 2004.

[4] C. Avendano, "Acoustic echo suppression in the STFT domain," in *Proc. IEEE Workshop on Application of Signal Processing to Audio and Acoustics*, New Paltz, NY, Oct. 2001, pp. 175–178.

[5] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[6] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[7] A. Gilloire and M. Vetterli, "Adaptive filtering in subbands with critical sampling: Analysis, experiments, and application to acoustic echo cancellation," *IEEE Trans. Signal Process.*, vol. 40, no. 8, pp. 1862–1875, Aug. 1992.

[8] M. R. Portnoff, "Time-frequency representation of digital signals and systems based on short-time Fourier analysis," *IEEE Trans. Signal Process.*, vol. ASSP-28, no. 2, pp. 55–69, Feb. 1980.

[9] C. Avendano, Temporal processing of speech in a time-feature space Oregon Grad. Inst. Sci. & Tech., 1997, Ph.D. dissertation.

[10] D. G. Manolakis, V. K. Ingle, and S. M. Kogon, *Statistical and Adaptive Signal Processing: Spectral Estimation, Signal Modeling, Adaptive Filtering, and Array Processing*. Boston, MA: McGraw-Hill, 2000.

[11] J. Wexler and S. Raz, "Discrete Gabor expansions," *Signal Process.*, vol. 21, pp. 207–220, Nov. 1990.