Adaptive System Identification in the Short-Time Fourier Transform Domain Using Cross-Multiplicative Transfer Function Approximation

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

***Abstract*—In this paper, we introduce cross-multiplicative transfer function (CMTF) approximation for modeling linear systems in the short-time Fourier transform (STFT) domain. We assume that the transfer function can be represented by cross-multiplicative terms between distinct subbands. We investigate the influence of cross-terms on a system identifier implemented in the STFT domain and derive analytical relations between the noise level, data length, and number of cross-multiplicative terms, which are useful for system identification. As more data becomes available or as the noise level decreases, additional cross-terms should be considered and estimated to attain the minimal mean-square error (mse). A substantial improvement in performance is then achieved over the conventional multiplicative transfer function (MTF) approximation. Furthermore, we derive explicit expressions for the transient and steady-state mse performances obtained by adaptively estimating the cross-terms. As more cross-terms are estimated, a lower steady-state mse is achieved, but the algorithm then suffers from slower convergence. Experimental results validate the theoretical derivations and demonstrate the effectiveness of the proposed approach as applied to acoustic echo cancellation.**

***Index Terms*—Multiplicative transfer function (MTF), short-time Fourier transform (STFT), subband adaptive filtering, system identification.**

I. Introduction

IDENTIFYING linear time-invariant (LTI) systems in the short-time Fourier transform (STFT) domain has been studied extensively and is of major importance in many applications [1]–[7]. LTI system representation in the STFT domain generally requires crossband filters between subbands [1], [8]. To avoid the crossband filters, a multiplicative transfer function (MTF) approximation is often employed (e.g., [2], [5]). This approximation relies on the assumption that the support of the STFT analysis window is sufficiently large compared to the duration of the system impulse response, and that the transfer function of the system can be modeled as multiplicative. As the length of the analysis window increases, the MTF approximation becomes more accurate. However, the length of the input signal that can be employed for the system identification is usually finite to enable tracking during time variations of the system. Hence, as the length of the analysis window increases, fewer observations in each frequency bin become available.

Recently, we have investigated the influence of the analysis window length on the performance of a system identifier that relies on the MTF approximation [9]. We showed that the minimum mean-square error (mse) attainable under this approximation can be decomposed into two error terms. The first term, attributable to using a finite-support analysis window, is monotonically decreasing as a function of the window length, while the second term is a consequence of restricting the length of the input signal and is monotonically *increasing* as a function of the window length. Therefore, system identification performance does not necessarily improve by increasing the length of the analysis window. The signal-to-noise ratio (SNR) and the input signal length determine the optimal length of the window. We showed that as the SNR or input signal length decreases, a shorter analysis window should be used.

In this paper, we introduce cross-multiplicative transfer function (CMTF) approximation in the STFT domain. The transfer function of the system is represented by cross-multiplicative terms between distinct subbands, and data from adjacent frequency bins is used for the system identification. Two identification schemes are introduced: One is an offline scheme in the STFT domain based on the least-squares (LS) criterion for estimating the CMTF coefficients. In the second scheme, the cross-terms are estimated adaptively using the least-mean-square (LMS) algorithm [10]. We analyze the performances of both schemes and derive explicit expressions for the obtainable minimum mse (mmse). The analysis reveals important relations between the noise level, data length, and number of cross-multiplicative terms, which are useful for system identification. As more data becomes available or as the noise level decreases, additional cross-terms should be considered and estimated to attain the mmse. In this case, a substantial improvement in performance is achieved over the conventional MTF approximation. For every data length and noise level there exists an optimal number of useful cross-multiplicative terms, so increasing the number of estimated cross-terms does not necessarily imply a lower mse. Note that similar results have been obtained in the context of system identification with crossband filters [1].

Manuscript received May 18, 2007; revised September 18, 2007. This work was supported by the Israel Science Foundation under Grant. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sen M. Kuo.

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

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2007.910789

1558-7916/$25.00 © 2007 IEEE

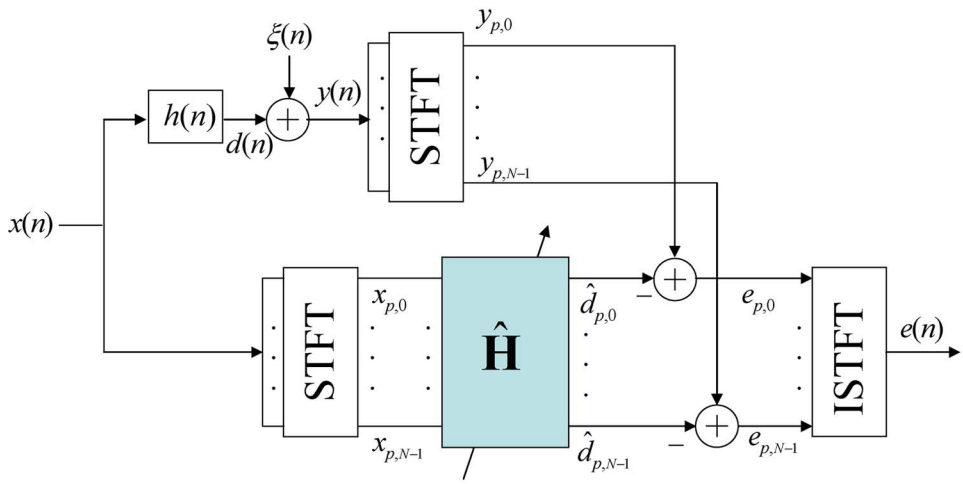

Fig. 1. System identification in the STFT domain. The unknown system $h(n)$ is modeled by the block $\hat{\mathbf{H}}$ in the STFT domain.

The main contribution of this work is a derivation of an explicit convergence analysis of the CMTF approximation, which includes the MTF approach as a special case. We derive explicit expressions for the transient and steady-state mse in frequency bins for white Gaussian processes. At the beginning of the adaptation process, the length of the data is short, and only a few cross-terms should be estimated, whereas as more data become available more cross-terms can be used to achieve the mmse. Consequently, the MTF approach is associated with faster convergence, but suffers from higher steady-state mse. Estimation of additional cross-terms results in a lower convergence rate, but improves the steady-state mse with a small increase in computational cost. Experimental results with white Gaussian signals and real speech signals validate the theoretical results derived in this work, and demonstrate the relations between the number of useful cross-terms and transient and steady-state mse.

This paper is organized as follows. In Section II, we introduce the CMTF approximation for system identification in the STFT domain. In Section III, we consider offline estimation of the cross-terms, and derive an explicit expression for the attainable mmse. In Section IV, we present an adaptive implementation of the CMTF estimation and analyze the transient and steady-state mse in subbands. Finally, in Section V, we present experimental results which verify the theoretical derivations.

II. CMTF Approximation

In this section, we introduce an CMTF approximation for system identification in the STFT domain. Throughout this work, unless explicitly noted, the summation indexes range from $-\infty$ to ∞.

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 [11]

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

where

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

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 ($0 \le k \le N-1$), L is the translation factor, and * denotes complex conjugation. A system identifier operating in the STFT domain is illustrated in Fig. 1, where the unknown system $h(n)$ is modeled in the STFT domain by a block $\hat{\mathbf{H}}$. Applying the STFT to $y(n)$, we have in the time–frequency domain [1]

$$y_{p,k} = d_{p,k} + \xi_{p,k}. \tag{4}$$

The signal component in the system output is related to its input in the STFT domain through crossband filters

$$d_{p,k} = \sum_{k'=0}^{N-1}\sum_{p'=0}^{M-1} x_{p-p',k'}h_{p',k,k'} \tag{5}$$

where $h_{p,k,k'}$ denotes a crossband filter of length M from frequency bin k' to frequency bin k. The crossband filters depend on both the system impulse response $h(n)$ and the STFT parameters. The widely used MTF approximation [9] avoids crossband filters by assuming 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 $h(n)$. In this case, $\tilde{\psi}(n-m)h(m) \approx \tilde{\psi}(n)h(m)$, and consequently (5) reduces to [12]

$$d_{p,k} \approx h_k x_{p,k} \tag{6}$$

where $h_k \triangleq \sum_{m=0}^{N_h-1} h(m)\exp(-j2\pi mk/N)$ and N_h is the length of $h(n)$. Note that the MTF approximation (6) approximates the time-domain linear convolution in (1) by a circular convolution of the input signal's pth frame and the system impulse response, using a frequency-bin product of the corresponding discrete Fourier transforms (DFTs). In the limit, for an infinitly long analysis window, the linear convolution would be exactly multiplicative in the frequency domain. This approximation is employed in some block frequency-domain methods, which attempt to estimate the unknown system in the frequency domain using block updating techniques (e.g., [13]–[16]).

Due to the finite length of the input signal, the MTF approximation results in insufficient accuracy of the system estimate, even for a long analysis window. This inaccuracy is attributable to the fact that fewer observations become available in each frequency band [9]. Furthermore, the exact STFT representation of the system in (5) implies that the drawback of the MTF approximation may be related to ignoring cross-terms between subbands. Using data from adjacent frequency bins and including cross-multiplicative terms between distinct subbands, we may improve the system estimate accuracy without significantly increasing the computational cost.

Specifically, let $h_{k,k'}$ be a cross-term from frequency bin k' to frequency bin k and let $d_{p,k}$ be approximated by $2K+1$ cross-terms around frequency bin k, i.e.,

$$d_{p,k} \approx \sum_{k'=k-K}^{k+K} h_{k,k' \bmod N} x_{p,k' \bmod N}. \tag{7}$$

Note that for $K=0$, (7) reduces to the MTF approximation (6). Equation (7) represents the CMTF approximation for modeling an LTI system in the STFT domain.

III. Offline System Identification

In this section, we consider an offline scheme for estimating the CMTF coefficients using an LS optimization criterion for each frequency bin and derive an explicit expression for the obtainable mmse.

Let

$$\mathbf{x}_k = [x_{0,k}\, x_{1,k}\, \cdots\, x_{P-1,k}]^T \tag{8}$$

denote a finite-length time-trajectory of x_{pk} for frequency bin k, and let the vectors $\mathbf{y}_k$, $\mathbf{d}_k$, and $\boldsymbol{\xi}_k$ be defined similarly. Then, (4) can be written in vector form as

$$\mathbf{y}_k = \mathbf{d}_k + \boldsymbol{\xi}_k. \tag{9}$$

Let $\mathbf{X}_k = [\mathbf{x}_{(k-K) \bmod N} \cdots \mathbf{x}_{(k+K) \bmod N}]$ and let

$$\tilde{\mathbf{h}}_k = \left[h_{k,(k-K) \bmod N}\ \cdots\ h_{k,(k+K) \bmod N}\right]^T \tag{10}$$

denote $2K+1$ cross-terms for frequency bin k. Then, the CMTF approximation (7) can be written in vector form as

$$\mathbf{d}_k = \mathbf{X}_k \tilde{\mathbf{h}}_k. \tag{11}$$

The LS estimate of $\tilde{\mathbf{h}}_k$ is therefore given by

$$\begin{aligned}\hat{\tilde{\mathbf{h}}}_k &= \arg\min_{\tilde{\mathbf{h}}_k} \|\mathbf{y}_k - \mathbf{X}_k \tilde{\mathbf{h}}_k\|^2 \\ &= \left(\mathbf{X}_k^H \mathbf{X}_k\right)^{-1} \mathbf{X}_k^H \mathbf{y}_k\end{aligned} \tag{12}$$

where we assume that $\mathbf{X}_k^H \mathbf{X}_k$ is not singular. Substituting (12) into (11), we obtain an estimate of the desired signal in the STFT domain, using $2K+1$ cross-terms.

A. MSE Analysis

We now derive an explicit expression for the mmse in the STFT domain. To make the analysis mathematically tractable we assume that $x_{p,k}$ and $\xi_{p,k}$ are zero-mean white Gaussian signals with variances σ_x^2 and σ_ξ^2, respectively, and that they are statistically independent. The Gaussian assumption of the corresponding STFT signals underlies the design of many speech-enhancement systems [17] and can be justified by a version of the central limit theorem [18,Th. 4.4.2]. The following mse analysis is closely related to that derived in [1], and the reader is referred to there for further details.

The (normalized) mse is defined as

$$\epsilon(K) = \frac{1}{E_d} \sum_{k=0}^{N-1} E\{\|\mathbf{d}_k - \hat{\mathbf{d}}_k\|^2\} \tag{13}$$

where $E_d \triangleq \sum_{k=0}^{N-1} E\{\|\mathbf{d}_k\|^2\}$, and $\hat{\mathbf{d}}_k = \mathbf{X}_k \hat{\tilde{\mathbf{h}}}_k$. Substituting (12) into (13), the mse can be expressed as

$$\begin{aligned}\epsilon(K) =& \frac{1}{E_d} \sum_{k=0}^{N-1} E\left\{\left\|\mathbf{X}_k \left(\mathbf{X}_k^H \mathbf{X}_k\right)^{-1} \mathbf{X}_k^H \boldsymbol{\xi}_k\right\|^2\right\} \\ &+ \frac{1}{E_d} \sum_{k=0}^{N-1} E\left\{\left\|\left[\mathbf{I}_P - \mathbf{X}_k \left(\mathbf{X}_k^H \mathbf{X}_k\right)^{-1} \mathbf{X}_k^H\right] \mathbf{d}_k\right\|^2\right\}\end{aligned} \tag{14}$$

where $\mathbf{I}_P$ is the identity matrix of size $P \times P$. Equation (14) can be rewritten as

$$\epsilon(K) = \epsilon_1 + 1 - \epsilon_2 \tag{15}$$

where

$$\epsilon_1 = \frac{1}{E_d} \sum_{k=0}^{N-1} E\left\{\boldsymbol{\xi}_k^H \mathbf{X}_k \left(\mathbf{X}_k^H \mathbf{X}_k\right)^{-1} \mathbf{X}_k^H \boldsymbol{\xi}_k\right\} \tag{16}$$

$$\epsilon_2 = \frac{1}{E_d} \sum_{k=0}^{N-1} E\left\{\mathbf{d}_k^H \mathbf{X}_k \left(\mathbf{X}_k^H \mathbf{X}_k\right)^{-1} \mathbf{X}_k^H \mathbf{d}_k\right\}. \tag{17}$$

Let $\mathbf{h}_{k,k'}$ denote the crossband filter from frequency bin k' to frequency bin k

$$\mathbf{h}_{k,k'} = [h_{0,k,k'} \quad h_{1,k,k'} \quad \cdots \quad h_{M-1,k,k'}]^T \tag{18}$$

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

$$\mathbf{c}_k = \begin{bmatrix}\mathbf{h}_{k,0}^T & \mathbf{h}_{k,1}^T & \cdots & \mathbf{h}_{k,N-1}^T\end{bmatrix}^T. \tag{19}$$

In addition, let us assume that $x_{p,k}$ is variance-ergodic and that the length P of the time-trajectories is sufficiently large, so that $(1/P)\sum_{p=0}^{P-1} x_{p,k}x_{p+s,k'}^* \approx E\{x_{p,k}x_{p+s,k'}^*\}$. Accordingly, using the fourth-order moment factoring theorem for zero-mean complex Gaussian samples [19] and following a similar analysis to that given in [1], we obtain an explicit expression for $\epsilon(K)$ as follows:

$$\epsilon(K) = \frac{a(K)}{\eta} + b(K) \tag{20}$$

where

$$a(K) \triangleq \frac{(2K+1)N}{P\sum_{k=0}^{N-1}\|\mathbf{c}_k\|^2} \tag{21}$$

$$b(K) \triangleq 1 - \frac{(2K+1)}{P}, - \frac{\sum_{k=0}^{N-1}\sum_{m=0}^{2K}|h_{0,k,(k-K+m) \bmod N}|^2}{\sum_{k=0}^{N-1}\|\mathbf{c}_k\|^2} \tag{22}$$

and $\eta = \sigma_x^2/\sigma_\xi^2$ denotes the SNR. Equations (20)–(22) represent the mmse obtained by using LS estimates of $2K+1$ cross-terms. The mmse $\epsilon(K)$ is a monotonically decreasing function of η. Furthermore, it is easy to verify from (21) and (22) that $\epsilon(K+1) > \epsilon(K)$ for low SNR, and $\epsilon(K+1) \leq \epsilon(K)$ for high SNR. Hence, $\epsilon(K)$ and $\epsilon(K+1)$ must intersect at a certain SNR value, denoted by $\eta_{K+1\to K}$. That is, for SNR values higher than $\eta_{K+1\to K}$, a lower mse can be achieved by estimating $2(K+1)+1$ cross-terms rather than only $2K+1$ cross-terms. Employing the conventional MTF approximation (i.e., ignoring all the cross-terms), yields the minimal mse only when the SNR is lower than $\eta_{1\to 0}$. The SNR intersection point $\eta_{K+1\to K}$, obtained by requiring that $\epsilon(K+1) = \epsilon(K)$, is given by

$$\eta_{K+1\to K} = \frac{2N}{2\sum_{k=0}^{N-1}\|\mathbf{c}_k\|^2 + P\sum_{k=0}^{N-1} f_k(K)} \tag{23}$$

where

$$f_k(K) = |h_{0,k,(k-K-1) \bmod N}|^2 + |h_{0,k,(k+K+1) \bmod N}|^2. \tag{24}$$

Since $\eta_{K+1\to K}$ is inversely proportional to P, the number of cross-terms that should be estimated in order to achieve the mmse increases as we increase P. Note that we implicitly assume that during P frames the system impulse response does not change, and the estimated cross-terms are updated every P frames. Therefore, in case time variations in the system are slow, we can increase P and correspondingly increase the number of estimated cross-terms to achieve a lower mse. These relations indicate that for a given power and length of the input signal, there exists an optimal number of estimated cross-terms that achieves the minimal mse. Note that similar mse behavior was demonstrated in the context of system identification with cross-band filters [1].

B. Computational Complexity

The computational complexity of the proposed approach requires the solution of LS normal equations $(\mathbf{X}_k^H\mathbf{X}_k)\hat{\hat{\mathbf{h}}}_k = \mathbf{X}_k^H\mathbf{y}_k$ [see (12)] for each frequency bin. This results in $P(2K+1)^2 + (2K+1)^3/3$ arithmetic operations when using the Cholesky decomposition [20]. Computing the desired signal estimate (11) results in an additional $2P(2K+1)$ arithmetic operations. Assuming P is sufficiently large and neglecting the computations required for the forward and inverse STFTs, the complexity associated with the CMTF approach is given by

$$O_{\mathrm{CMTF}}(K) = O[NP(2K+1)^2]. \tag{25}$$

We observe that the computational complexity obtained by using the CMTF approximation is $(2K+1)^2$ times larger than that obtained by using the MTF approximation. However, incorporating cross-terms into the system model may yield lower mse for stronger and longer input signals.

IV. Adaptive System Identification

In this section, we adaptively update the cross-terms in frequency bins by the LMS algorithm [10] and derive explicit expressions for the transient and steady-state mse in subbands.

Let $\hat{d}_{p,k}$ be an estimate of $d_{p,k}$ using $2K+1$ adaptive cross-terms around the frequency bin k, i.e.,

$$\hat{d}_{p,k} = \sum_{k'=k-K}^{k+K} x_{p,k'}\hat{h}_{k,k'}(p) \tag{26}$$

where $\hat{h}_{k,k'}(p)$ is an adaptive cross-term that represents an estimate of the CMTF $h_{k,k'}$ at frame index p (recall that due to periodicity of the frequency bins, the summation index k' is related to frequency bin $k' \bmod N$). Let $\hat{\mathbf{h}}_k(p) = [\hat{h}_{k,k-K}(p)\ \hat{h}_{k,k-K+1}(p)\ \cdots\ \hat{h}_{k,k+K}(p)]^T$ denote $2K+1$ adaptive cross-terms at the kth frequency bin, and let $\mathbf{x}_k(p) = [x_{p,k-K}\ x_{p,k-K+1}\ \cdots\ x_{p,k+K}]^T$ be the input data vector corresponding to $\hat{\mathbf{h}}_k(p)$. Then, the estimated desired signal $\hat{d}_{p,k}$ from (26) can be rewritten as

$$\hat{d}_{p,k} = \mathbf{x}_k^T(p)\hat{\mathbf{h}}_k(p). \tag{27}$$

The $2K+1$ adaptive cross-terms are updated using the LMS algorithm as

$$\hat{\mathbf{h}}_k(p+1) = \hat{\mathbf{h}}_k(p) + \mu e_{p,k}\mathbf{x}_k^*(p) \tag{28}$$

where

$$e_{p,k} = y_{p,k} - \hat{d}_{p,k} \tag{29}$$

is the error signal in the kth frequency bin, $y_{p,k}$ is defined in (4), and μ is a step-size. Let $\mathbf{h}_k$ be a vector containing the first

element in each of the $2K+1$ crossband filters around the kth frequency bin, i.e.,

$$\mathbf{h}_k = [h_{0,k,k-K} \quad h_{0,k,k-K+1} \quad \cdots \quad h_{0,k,k+K}]^T. \tag{30}$$

In addition, let $\bar{\mathbf{h}}_{k,k'} = [h_{1,k,k'} \ \cdots \ h_{M-1,k,k'}]^T$ be the last $M-1$ elements of the crossband filter $\mathbf{h}_{k,k'}$ [as defined in (18)], let $\boldsymbol{\chi}_k(p) = [x_{p,k} \ x_{p-1,k} \ \cdots \ x_{p-M+1,k}]^T$, and let $= [x_{p-1,k} \ \cdots \ x_{p-M+1,k}]^T$. Then, defining

$$\mathbf{g}_k(p) = \hat{\mathbf{h}}_k(p) - \mathbf{h}_k \tag{31}$$

as the misalignment vector and substituting (4), (5), and (27) into (29), the error signal can be written as

$$e_{p,k} = \tilde{\mathbf{x}}_k^T(p)\tilde{\mathbf{c}}_k + \bar{\mathbf{x}}_k^T(p)\bar{\mathbf{c}}_k - \mathbf{x}_k^T(p)\mathbf{g}_k(p) + \xi_{p,k} \tag{32}$$

where $\tilde{\mathbf{c}}_k, \bar{\mathbf{c}}_k, \tilde{\mathbf{x}}_k(p)$, and $\bar{\mathbf{x}}_k(p)$ are the column-stack concatenations of $\{\mathbf{h}_{k,k'}\}_{k'\in\mathcal{L}}, \{\bar{\mathbf{h}}_{k,k'}\}_{k'=k-K}^{k+K}, \{\boldsymbol{\chi}_{k'}(p)\}_{k'\in\mathcal{L}}$, and $\{\bar{\boldsymbol{\chi}}_{k'}(p)\}_{k'=k-K}^{k+K}$, respectively, and $\mathcal{L} = \{k' | k' \in [0, N-1]$ and $k' \notin [k-K, k+K]\}$. Substituting (32) into (28), the LMS update equation can be expressed as

$$\begin{aligned}\mathbf{g}_k(p+1) = &\left[\mathbf{I} - \mu\mathbf{x}_k^*(p)\mathbf{x}_k^T(p)\right]\mathbf{g}_k(p) \\ &+ \mu\left[\tilde{\mathbf{x}}_k^T(p)\tilde{\mathbf{c}}_k\right]\mathbf{x}_k^*(p) \\ &+ \mu\left[\bar{\mathbf{x}}_k^T(p)\bar{\mathbf{c}}_k\right]\mathbf{x}_k^*(p) + \mu\xi_{p,k}\mathbf{x}_k^*(p)\end{aligned} \tag{33}$$

where $\mathbf{I}$ is the identity matrix.

A. MSE Analysis

We proceed with the mean-square analysis of the adaptation algorithm under the assumptions made in Section III-A. The analysis relies on the common assumption that $\mathbf{x}_k(p)$ is independent of $\hat{\mathbf{h}}_k(p)$ (e.g., [21], [22]).

1) Transient Performance: The transient mse is defined by

$$\epsilon_k(p) = E\{|e_{p,k}|^2\}. \tag{34}$$

Using the whiteness property of the input signal, and substituting (32) into (34), the mse can be expressed as

$$\epsilon_k(p) = \sigma_\xi^2 + \sigma_x^2(\|\tilde{\mathbf{c}}_k\|^2 + \|\bar{\mathbf{c}}_k\|^2) + \sigma_x^2 E\{\|\mathbf{g}_k(p)\|^2\}. \tag{35}$$

In order to find an explicit expression for the transient mse, a recursive formula for $E\{\|\mathbf{g}_k(p)\|^2\}$ is required. From (33), we obtain

$$\begin{aligned}&E\{\|\mathbf{g}_k(p+1)\|^2\} \\ &\quad= E\left\{\left\|\left[\mathbf{I} - \mu\mathbf{x}_k^*(p)\mathbf{x}_k^T(p)\right]\mathbf{g}_k(p)\right\|^2\right\} \\ &\qquad+ \mu^2 E\left\{\left\|\left[\tilde{\mathbf{x}}_k^T(p)\tilde{\mathbf{c}}_k\right]\mathbf{x}_k^*(p)\right\|^2\right\} \\ &\qquad+ \mu^2 E\left\{\left\|\left[\bar{\mathbf{x}}_k^T(p)\bar{\mathbf{c}}_k\right]\mathbf{x}_k^*(p)\right\|^2\right\} \\ &\qquad+ \mu^2 E\left\{\left\|\xi_{p,k}\mathbf{x}_k^*(p)\right\|^2\right\}.\end{aligned} \tag{36}$$

Using the independence assumption, and the fourth-order moment factoring theorem for zero-mean complex Gaussian samples, the first term on the right side of (36) can be expressed as (see Appendix A)

$$\begin{aligned}&E\left\{\left\|\left[\mathbf{I} - \mu\mathbf{x}_k^*(p)\mathbf{x}_k^T(p)\right]\mathbf{g}_k(p)\right\|^2\right\} \\ &\qquad= \left[1 - 2\mu\sigma_x^2 + 2\mu^2\sigma_x^4(K+1)\right]E\{\|\mathbf{g}_k(p)\|^2\}.\end{aligned} \tag{37}$$

The evaluation of the last three terms in (36) is straightforward, and they can be expressed as

$$\begin{aligned}&\mu^2 E\left\{\left\|\left[\tilde{\mathbf{x}}_k^T(p)\tilde{\mathbf{c}}_k\right]\mathbf{x}_k^*(p)\right\|^2\right\} \\ &\qquad= \mu^2\sigma_x^4\|\tilde{\mathbf{c}}_k\|^2(2K+1)\end{aligned} \tag{38a}$$

$$\begin{aligned}&\mu^2 E\left\{\left\|\left[\tilde{\mathbf{x}}_k^T(p)\bar{\mathbf{c}}_k\right]\mathbf{x}_k^*(p)\right\|^2\right\} \\ &\qquad= \mu^2\sigma_x^4\|\bar{\mathbf{c}}_k\|^2(2K+1)\end{aligned} \tag{38b}$$

$$\begin{aligned}&\mu^2 E\left\{\left\|\xi_{p,k}\mathbf{x}_k^*(p)\right\|^2\right\} \\ &\qquad= \mu^2\sigma_\xi^2\sigma_x^2(2K+1).\end{aligned} \tag{38c}$$

Substituting (37) and (38) into (36), we have an explicit recursive expression for $E\{\|\mathbf{g}_k(p)\|^2\}$

$$E\{\|\mathbf{g}_k(p)\|^2\} = \alpha(K)\,E\{\|\mathbf{g}_k(p-1)\|^2\} + \beta_k(K) \tag{39}$$

where

$$\alpha(K) \triangleq 1 - 2\mu\sigma_x^2 + 2\mu^2\sigma_x^4(K+1) \tag{40}$$

$$\beta_k(K) \triangleq \mu^2\sigma_x^2(2K+1)\left[\sigma_\xi^2 + \sigma_x^2(\|\tilde{\mathbf{c}}_k\|^2 + \|\bar{\mathbf{c}}_k\|^2)\right]. \tag{41}$$

Equations (35) and (39)–(41) represent the mse behavior in the kth frequency bin using $2K+1$ adaptive cross-terms.

2) Stability: It is easy to verify from (35) and (39) that a sufficient condition for mse convergence is that $|\alpha(K)| < 1$, which results in the following condition on the step-size μ:

$$0 < \mu < \frac{1}{\sigma_x^2(K+1)}. \tag{42}$$

The upper bound of μ is inversely proportional to K, and as the number of cross-terms increases, a lower step-size value should be utilized, which may result in slower convergence. An optimal step-size that results in the fastest convergence for each K is obtained by differentiating $\alpha(K)$ with respect to μ, which yields

$$\mu_{\text{opt}} = \frac{1}{2\sigma_x^2(K+1)}. \tag{43}$$

By substituting (43) into (40), we obtain

$$\alpha_{\text{opt}}(K) = 1 - \frac{1}{2(K+1)}. \tag{44}$$

Expectedly, we have $\alpha_{\text{opt}}(K) < \alpha_{\text{opt}}(K+1)$, which indicates that faster convergence is achieved by decreasing K.

3) Steady-State Performance: We proceed with analyzing the steady-state performance of the adaptive algorithm. Let us first consider the mean convergence of the misalignment vector

$\mathbf{g}_k(p)$. From (33), and by using the whiteness property of $x_{p,k}$, it is easy to verify that $E\{\mathbf{g}_k(\infty)\} = 0$; hence

$$E\{\hat{\mathbf{h}}_k(\infty)\} = \mathbf{h}_k \tag{45}$$

where $\mathbf{h}_k$ is defined in (30). This indicates that the adaptive cross-terms converge in the mean to the first element in the corresponding crossband filters. Substituting (45) for $\hat{\mathbf{h}}_k(p)$ in (35), we find the minimum mse obtainable in the kth frequency bin:

$$\epsilon_k^{\min} = \sigma_\xi^2 + \sigma_x^2(\|\tilde{\mathbf{c}}_k\|^2 + \|\bar{\mathbf{c}}_k\|^2). \tag{46}$$

Now, substituting (46) into (35), the steady-state mse can be expressed as

$$\epsilon_k(\infty) = \epsilon_k^{\min} + \sigma_x^2 E\{\|\mathbf{g}_k(\infty)\|^2\}. \tag{47}$$

Provided that μ satisfies (42), such that the mean-square convergence of the algorithm is guaranteed, the steady-state solution of (39) is given by

$$E\{\|\mathbf{g}_k(\infty)\|^2\} = \frac{\beta_k(K)}{1-\alpha(K)}. \tag{48}$$

Substituting (40) and (41) into (48), we obtain an explicit expression for $E\{\|\mathbf{g}_k(\infty)\|^2\}$. Accordingly, (47) can be written, after some manipulations, as

$$\epsilon_k(\infty) = \frac{2-\mu\sigma_x^2}{2-2\mu\sigma_x^2(K+1)}\epsilon_k^{\min}. \tag{49}$$

Equations (46) and (49) provide an explicit expression for the steady-state mse in frequency-bins. Note that $\epsilon_k^{\min}$ implicitly depends on K (it is actually a decreasing function of K), and therefore the influence of the number of estimated cross-terms on the steady-state mse $\epsilon_k(\infty)$ is not clear from (49). However, since a smaller step-size is used for larger K [see (42)], a lower steady-state mse is expected as we increase the number of estimated cross-terms.

B. Computational Complexity

The adaptation formula given in (28) requires $2K+2$ complex multiplications, $2K+1$ complex additions, and one complex subtraction to compute the error signal. Note that each arithmetic operation is not carried out every input sample, but once for every L input samples, where L denotes the decimation factor of the STFT representation. Thus, the adaptation process requires $4(K+1)$ arithmetic operations for every L input samples. Moreover, computing the desired signal estimate in (26) results in an additional $4K+1$ arithmetic operations. Hence, the proposed adaptive approach requires $8K+5$ arithmetic operations for every L input samples and each frequency bin. When compared to the MTF approach ($K = 0$), the proposed approach involves an increase of only $8K$ arithmetic operations for every L input samples and every frequency bin.

C. Discussion

The expressions derived for the analysis of offline and adaptive schemes (Sections III and IV) are related to the problem of model-order selection, where in our case the model order is determined by the number of estimated cross-multiplicative terms. Selecting the optimal model complexity for a given data set is a fundamental problem in many system identification applications [23]–[29], and many criteria have been proposed for this purpose. The Akaike information criterion (AIC) [28] and the minimum description length (MDL) [29] are among the most popular choices. Generally, the estimation error can be decomposed into two terms: a bias term, which is monotonically decreasing as a function of the model order, and a variance term, which is respectively monotonically *increasing*. The optimal model order is affected by the level of noise in the data and the length of the observable data. As the SNR increases or as more data is employable, the optimal model complexity increases, and correspondingly additional cross-terms can be estimated to achieve lower mse. At the beginning of the adaptation process, the length of the data is short, and only a few cross-terms are estimated. As the adaptation process proceeds, more data can be used, additional cross-terms can be estimated, and lower mse can be achieved. These points will be demonstrated in the next section.

V. Experimental Results

In this section, we present two experiments to demonstrate the theoretical results. The first examines the proposed approach under white Gaussian signals, whereas the second experiment is carried out in an acoustic echo cancellation scenario using real speech signals. The performance of both offline and adaptive schemes are evaluated, and a comparison is made with the conventional fullband approach. The evaluation includes objective quality measures, a subjective study of temporal waveforms, and informal listening tests. For the adaptive system identification, we use the normalized LMS (NLMS) algorithm [10] for updating the cross-terms,[1] instead of the LMS algorithm that was used for the analysis. That is, the update formula (28) is now modified to

$$\hat{\mathbf{h}}_k(p+1) = \hat{\mathbf{h}}_k(p) + \frac{\mu}{\|\mathbf{x}_k(p)\|^2} e_{p,k}\mathbf{x}_k^*(p) \tag{50}$$

where $0 < \mu < 2$. In the following experiments, we use a Hamming synthesis window of length N with 50% overlap (i.e., $L = 0.5N$), and a corresponding minimum-energy analysis window that satisfies the completeness condition [30]. The sample rate is 16 kHz.

A. Performance Evaluation for White Gaussian Input Signals

In the first experiment, we examine the system identifier performance in the STFT domain for white Gaussian signals. The input signal $x(n)$ and the additive noise signal $\xi(n)$ are uncorrelated zero-mean white Gaussian processes with variances σ_x^2 and σ_ξ^2, respectively. The lengths of the signals are 3 s. We model the impulse response as a nonstationary stochastic process with an exponential decay envelope, i.e., $h(n) = u(n)\beta(n)e^{-\alpha n}$, where $u(n)$ is the unit step function, $\beta(n)$ is a unit-variance zero-mean white Gaussian noise, and α

[1]The LMS algorithm is used in Section IV in order to make the mean-square analysis mathematically tractable. Most adaptive filtering applications, however, employ the NLMS algorithm, and it is used here for performance demonstration.

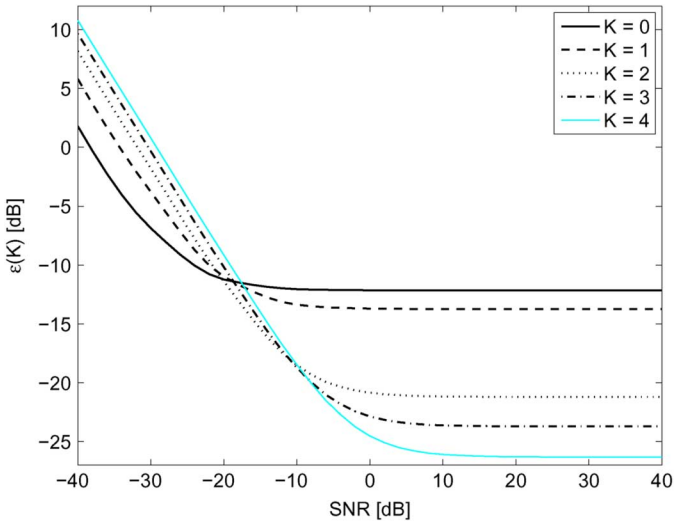

Fig. 2. MSE curves as a function of the SNR using LS estimates of the cross-terms (offline scheme), for white Gaussian signals of length 3 s.

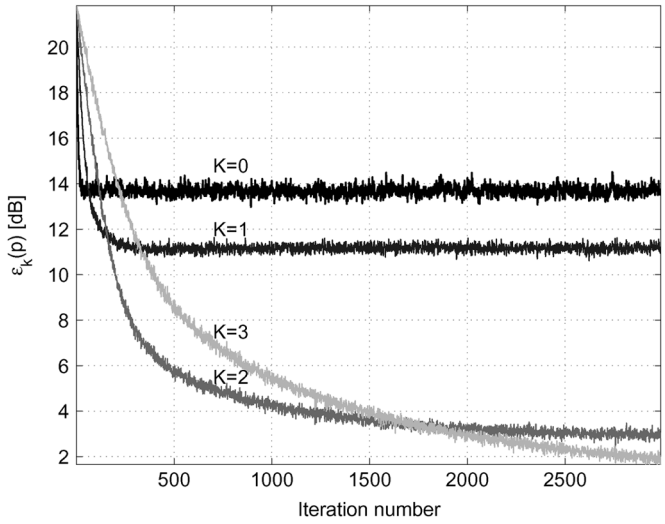

Fig. 3. Transient mse curves, obtained by adaptively updating the cross-terms via (50), for white Gaussian signals of length 12 s and SNR $= 30$ dB.

TABLE I
AVERAGE RUNNING TIME IN TERMS OF CPU FOR SEVERAL K VALUES, OBTAINED USING LS ESTIMATES OF THE CROSS-TERMS. THE LENGTH OF THE INPUT SIGNAL IS 3 S

K	Running Time [sec]
0 (MTF)	0.1168
1	0.3388
2	0.4073
3	0.5014
4	0.7442

TABLE II
AVERAGE RUNNING TIME IN TERMS OF CPU FOR SEVERAL K VALUES AS OBTAINED BY ADAPTIVELY UPDATING THE CROSS-TERMS. THE LENGTH OF THE INPUT SIGNAL IS 12 S

K	Running Time [sec]
0 (MTF)	0.1845
1	0.1936
2	0.2042
3	0.2156

is the decay exponent. In the following, we use $\alpha = 0.02$. To maintain the large analysis-window support assumption, which the CMTF approximation relies on, the length of the impulse response is chosen to be eight times shorter than the length of the analysis window ($N = 128$ and $N_h = 16$). Fig. 2 shows the mse curves $\epsilon(K)$, obtained by the offline scheme using (13), as a function of the SNR. The cross-terms are estimated using the LS criterion [see (12)]. The results confirm that as the SNR increases, the number of cross-terms that should be estimated to achieve the minimal mse increases. We observe that when the SNR is lower than -20 dB, the conventional MTF approximation ($K = 0$) yields the minimal mse. For higher SNR values, the estimation of five cross-terms per frequency-bin ($K = 2$) enables a substantial improvement of 10 dB in the mse. Similar results are obtained for longer signals, with the only difference being that the intersection points of the mse curves move toward lower SNR values [as expected from (23)]. The complexity of the proposed approach is evaluated by computing the central processing unit (CPU) running time[2] of the LS estimation process for each K. The average running time in terms of CPU seconds is summarized in Table I. We observe, as expected from (25), that the running time of the proposed approach increases as more cross-terms are estimated. For instance, the process of estimating five cross-terms ($K = 2$) is approximately four times slower than that of the MTF approach.

Fig. 3 shows the transient mse curves $\epsilon_k(p)$ for frequency bin $k = 1$ and SNR of 30 dB, as obtained by adaptively updating the cross-terms using (50). The length of the signals is 12 s, and the results are averaged over 1000 independent runs. Since the step-size μ should be inversely proportional to K to ensure convergence [see (42) and (43)], we choose $\mu = 0.1/(K+1)$. The results confirm that as more data is employed in the adaptation process, a lower mse is obtained by estimating additional cross-terms. Clearly, as K increases, a lower steady-state mse $\epsilon_k(\infty)$ is achieved; however, the algorithm then suffers from slower convergence. The conventional MTF approach yields faster convergence, but higher steady-state mse. Table II shows the average running times in terms of CPU seconds, as obtained by the adaptive scheme. Expectedly, higher running time is obtained by increasing K (see Section IV-B). However, in contrast to the offline scheme (Table I), the additional computational cost of estimating more cross-terms is small in the adaptive scheme. Including five cross-terms ($K = 2$), for instance, decreases the steady-state mse by approximately 11 dB, with only a small increase of 10% in computational complexity, when compared to the MTF approach ($K = 0$).

B. Acoustic Echo Cancellation Application

In the second experiment, we demonstrate the proposed approach in an acoustic echo cancellation application [31]–[33] using real speech signals. The experimental setup is depicted in Fig. 4. We use an ordinary office with a reverberation time T_{60}

[2]The simulations were all performed under MATLAB; v.7.2, on a Pentium IV 2.2-GHz PC with 1 GB of RAM, running Microsoft Windows XP v.2002.

Fig. 4. Experimental setup. A speakerphone (Phoenix Audio DUET Executive Conference Speakerphone) is connected to a laptop using its USB interface. Another speakerphone without its cover shows the placement of the built-in microphone and loudspeaker.

of about 100 ms. The measured acoustic signals are recorded by a DUET conference speakerphone, Phoenix Audio Technologies, which includes an omnidirectional microphone near the loudspeaker (more features of the DUET product are available at: http://phnxaudio.com.mytempweb.com/?tabid=62). The far-end signal is played through the speakerphone's built-in loudspeaker, and received together with the near-end signal by the speakerphone's built-in microphone. The small distance between the loudspeaker and the microphone yields relatively high SNR values, which may justify the estimation of more cross-terms. Employing the MTF approximation in this case, and ignoring all the cross-terms, may result in insufficient echo reduction. It is worth noting that estimation of crossband filters [1], rather than CMTF, may be even more advantageous, but estimation of crossband filters results in a significant increase in computational complexity. In this experiment, the signals are sampled at 16 kHz. A far-end speech signal $x(n)$ is generated by the loudspeaker and received by the microphone as an echo signal $d(n)$ together with a near-end speech signal and local noise [collectively denoted by $\xi(n)$]. The distance between the near-end source and the microphone is 1 m. According to the room reverberation time, the effective length of the echo path is 100 ms, i.e., $N_h = 1600$. We use a synthesis window of length 200 ms (corresponding to $N = 3200$), which is twice the length of the echo path. The influence of the window length on the performance is investigated in the sequel (see Section V-C). A commonly used quality measure for evaluating the performance of acoustic echo cancellers (AECs) is the echo-return loss enhancement (ERLE), defined in dB by

$$\text{ERLE}(K) = 10\log_{10}\frac{E\{y^2(n)\}}{E\{e_K^2(n)\}} \tag{51}$$

where

$$e_K(n) = y(n) - \hat{d}_K(n) \tag{52}$$

is the error signal, and $\hat{d}_K(n)$ is the inverse STFT of the estimated echo signal using $2K+1$ cross-terms in each frequency bin.

TABLE III
ECHO-RETURN LOSS ENHANCEMENT (ERLE) FOR SEVERAL K VALUES AND VARIOUS ANALYSIS WINDOW LENGTHS (N). THE EFFECTIVE LENGTH OF THE ECHO PATH IS $N_h = 1600$

K	ERLE [dB]			
	$N = 4N_h$	$N = 2N_h$	$N = N_h$	$N = 0.75N_h$
0 (MTF)	14.21	9.74	9.72	8.59
1	17.32	14.29	11.9	10.58
2	16.89	16.19	14.03	12.72
4	7.37	12.29	14.47	12.79
Fullband	18.5	18.5	18.5	18.5

Fig. 5(a)–(c) show the far-end signal, near-end signal, and microphone signal, respectively. Note that a double-talk situation (simultaneously active far-end and near-end speakers) occurs between 4.65 and 6.1 s (indicated by two vertical dotted lines). Since such a situation may cause divergence of the adaptive algorithm, a double-talk detector (DTD) is usually employed to detect near-end signal and freeze the adaptation [34], [35]. Since the design of a DTD is beyond the scope of this paper, we manually choose the periods where double-talk occurs and freeze the adaptation in these intervals. Fig. 5(d)–(g) show the error signal $e_K(n)$ obtained by using $K = 0, 1, 2,$ and 4, respectively, where the cross-terms are adaptively updated by the NLMS algorithm using a step-size $\mu = 1/(K+1)$. The performance of a conventional fullband AEC, where the echo signal is estimated in the time domain [33], is also evaluated [see Fig. 5(h)]. The NLMS algorithm is used for the fullband approach with a step-size value of 0.01 to insure stability.

Table III shows the ERLE values computed after convergence of the adaptive algorithms for various window lengths: $N = 4N_h, 2N_h, N_h,$ and $0.75N_h$ (the influence of the analysis window length N on the performance will be addressed in Section V-C). Clearly, the proposed CMTF approach is considerably more advantageous, in terms of ERLE, than the conventional MTF approach. For example when $N = 2N_h$, a substantial increase of 4.5 dB in the ERLE is obtained by estimating only 3 cross-terms ($K = 1$), whereas an additional 1.9-dB increase is achieved by including five cross-terms ($K = 2$). We observe from Fig. 5 that at the beginning of the adaptation, the convergence rate is slower for larger K, which initially results in higher error. The slower convergence is attributable to the relatively small step-size forced by estimating more cross-terms [see (42)]. However, as the adaptation proceeds, a smaller error is attained as more cross-terms are estimated. The results indicate that the optimal number of cross-terms that should be estimated in order to achieve the maximal ERLE is 5 ($K = 2$). It is worth noting, however, that a higher ERLE could be achieved for $K = 4$, if the adaptation process was longer. Subjective listening tests confirm that the proposed CMTF approach achieves a perceptual improvement in speech quality over the conventional MTF approach (audio files are available online [36]).

A comparison of the proposed approach with the fullband approach indicates that the latter achieves the maximal ERLE value (see Table III), and its convergence rate is inferior only to the MTF approach. However, the high ERLE value is achieved

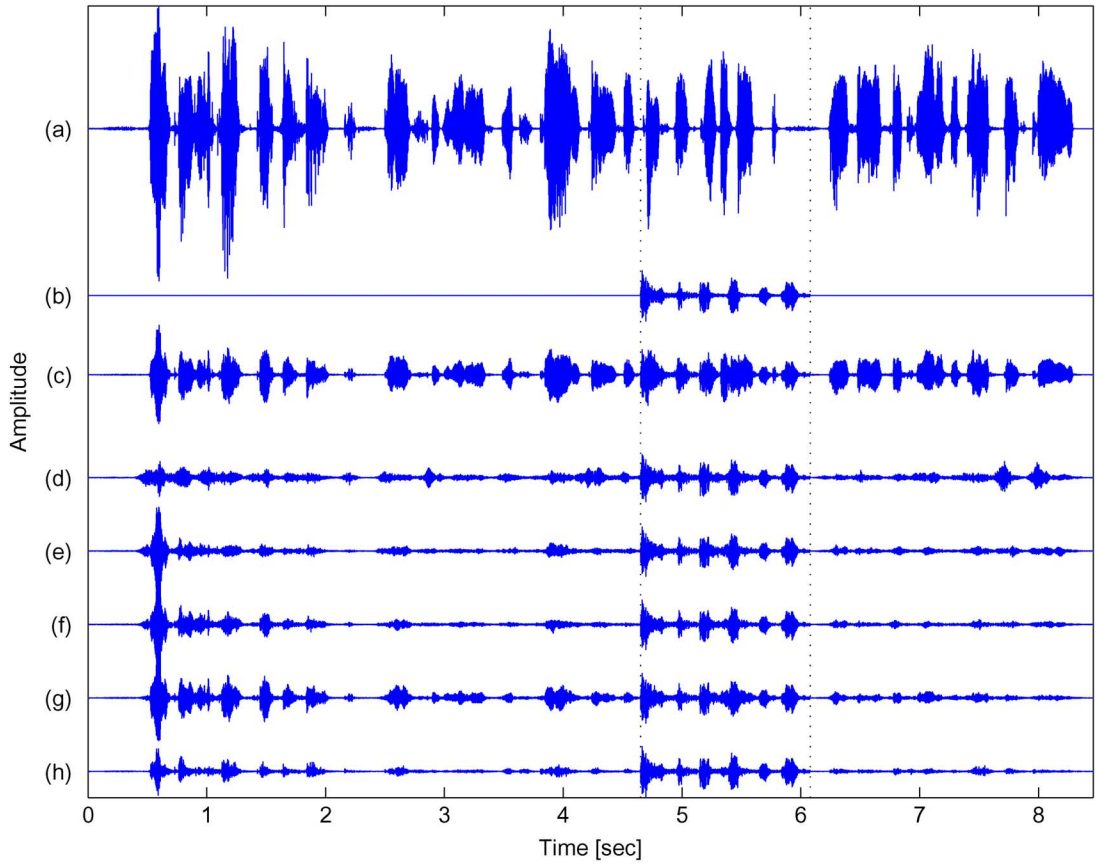

Fig. 5. Speech waveforms and error signals $e_K(n)$, obtained by adaptively updating the cross-terms via (50). A double-talk situation is indicated by vertical dotted lines. (a) Far-end signal. (b) Near-end signal. (c) Microphone signal. (d)—(h) Error signals for $K = 0$ (MTF approximation), $K = 1, K = 2, K = 4$, and the conventional fullband approach, respectively. The length of the analysis window is twice the length of the echo path $(N = 2N_h)$.

at the expense of a substantial increase in computational complexity. Specifically for $N = 2N_h$, running time measurements indicate that the fullband approach is approximately 33 times slower (233 s) than the proposed approach (7 s). Moreover, note that the performance improvement achieved by the fullband approach is not very significant (2.3 dB for $N = 2N_h$, when compared to $K = 2$), so that one can alternatively employ the CMTF approach with five cross-terms $(K = 2)$ to achieve computational efficiency. It should be noted that the relatively slow convergence of the proposed CMTF approach is a consequence of using a very long analysis window, which reduces the update rate of the adaptive cross-terms (assuming that the relative overlap between consecutive windows is retained). Due to the long echo path impulse response, a relatively long window is necessary to maintain the large support assumption. In fact, the CMTF approach (for any K) would suffer from slow convergence and bad tracking capabilities whenever the unknown system impulse response is long. As a result, applications like relative transfer function (RTF) identification [2], in which the unknown impulse response is much shorter, might be more suitable for using the CMTF approximation.

It is worthwhile noting that the relatively small ERLE values obtained by both fullband and subband approaches, may be attributable to the nonlinearity introduced by the loudspeaker and its amplifier. Estimating the overall nonlinear system by the LTI model in (1) yields a model mismatch that degrades the system estimate accuracy. Several techniques for nonlinear acoustic echo cancellation have been proposed (e.g., [37], [38]). However, combining such techniques with the CMTF approximation is beyond the scope of this paper.

C. Influence of the Analysis Window Length

Next, we investigate the influence of the STFT analysis window length (N) on the CMTF performance. We repeated the last experiment with various window lengths and computed the ERLE for each K (see Table III). As expected, the performance of the CMTF approach can be generally improved by using a longer analysis window. This is because CMTF heavily relies on the assumption of a long analysis window compared to the length of the system impulse response. Note that the fullband approach outperforms the proposed approach in terms of steady-state ERLE, even for a long analysis window $(N = 4N_h)$. We observe that as the window length increases, fewer cross-terms should be estimated to achieve the maximal ERLE. For instance, when the length of the window is equal to that of the impulse response $(N = N_h)$, nine cross-terms should be estimated $(K = 4)$, whereas when the window length is increased by a factor of 4 $(N = 4N_h)$, the maximal ERLE is achieved with the estimation of only three cross-terms $(K = 1)$. Further increasing the window length would ultimately make the MTF approach a preferable choice, with no cross-terms. This phenomenon is due to the fact that by increasing the analysis window length while retaining the relative overlap between consecutive windows (i.e., the ratio N/L is fixed),

TABLE IV
ECHO-RETURN LOSS ENHANCEMENT (ERLE) FOR SEVERAL K VALUES, IN THE PRESENCE OF NARROWBAND NOISE UNDER VARIOUS SNR CONDITIONS

K	ERLE [dB]			
	SNR= -5 dB	SNR= 0 dB	SNR= 5 dB	SNR= 10 dB
0 (MTF)	8.14	9.17	9.56	9.68
1	13.73	14.12	14.25	14.28
2	15.68	16.05	16.16	16.19
4	12.15	12.25	12.28	12.29
Fullband	12.46	15.39	17.09	17.89

fewer observations in each frequency bin are available, which increases the variance of the system estimate. Thus, the optimal model order decreases, and correspondingly fewer cross-terms need to be estimated to achieve higher ERLE.

D. Performance Evaluation Under Presence of Narrowband Noise Signal

In the third experiment, we demonstrate the effectiveness of the proposed approach over the fullband approach in the presence of a narrowband noise signal. The noise signal is generated using a white Gaussian signal to excite a bandpass filter with bandwidth of 150 Hz and a center frequency of 7.8 kHz. The resulting narrowband noise signal is then added to the microphone signal $y(n)$, and the experiment described in Section V-B is repeated under various SNR conditions. Table IV shows the ERLE obtained for SNR values of -5, 0, 5, and 10 dB, and for analysis window of length $N = 2N_h$. Clearly, as the SNR increases, the performance of the proposed approach, as well as that of the fullband approach, is generally improved. We observe that the performance degradation of the proposed CMTF approach, when compared to the noiseless scenario (see Table III), is less substantial than that of the fullband approach. Moreover, when considering low SNR values, the CMTF approach outperforms the fullband approach. For instance, for -5 dB SNR, incorporating five cross-terms $(K = 2)$ enables an increase of 3.2 dB in the ERLE relative to that achieved by the fullband approach. This is attributable to the fact that the noise is present in only a few frequency bins. By using the proposed approach, the system estimate is degraded only in these particular frequency bins, and the overall estimate is less affected by the noise. In the fullband approach, however, the estimation is carried out in the time domain, so the influence of the noise is much more devastating. This experiment shows that for narrowband noise, the ERLE and computational efficiency can be improved by using the proposed CMTF approach, compared to using the fullband approach.

VI. CONCLUSION

We have introduced an CMTF approximation for identifying an LTI system in the STFT domain. The cross-terms in each frequency bin are estimated either offline by using the LS criterion, or adaptively by using the LMS (or NLMS) algorithm. We have derived explicit relations between the attainable mmse and the power and length of the input signal. We showed that the number of cross-terms that should be utilized in the system identifier is larger for stronger and longer input signals. Consequently, for high SNR values and longer input signals, the proposed CMTF approach outperforms the conventional MTF approximation. This improvement is due to the fact that data from adjacent frequency-bins becomes more reliable and may be beneficially utilized for the system identification.

In addition, we have analyzed the transient and steady-state mse performances obtained by adaptively estimating the cross-terms. We showed that the MTF approximation yields faster convergence, but also results in higher steady-state mse. As the adaptation process proceeds, more data is employable, and lower mse is achieved by estimating additional cross-terms. Accordingly, during rapid time variations of the system, fewer cross-terms are useful. However, when the system time variations become slower, additional cross-terms can be incorporated into the system identifier and lower mse is attainable.

Experimental results corresponding to an acoustic echo cancellation scenario have demonstrated the advantage of the proposed approach. It is shown that a substantial improvement is achieved over the MTF approximation without significantly increasing the computational cost. Moreover, compared to the conventional fullband approach, the proposed approach yields a substantial decrease in computational complexity with only a slight degradation in performance. Furthermore, for additive narrowband noise, the CMTF approach outperforms the fullband approach. It should be noted that for reasons of convergence rate, applications that involve short impulse responses (e.g., identification of speech source coupling between sensors [39]) are more suitable for using the CMTF approximation due to the requirement of a large STFT analysis-window support.

Adaptive control of cross-terms is related to filter-length control [40]–[44]. Filter-length control algorithms dynamically adjust the number of filter taps and provide a balance between complexity, convergence rate and steady-state performance. By employing filter-length control techniques, an algorithm for adaptively controlling the number of cross-terms may be developed for both faster convergence rate and smaller steady-state mse. This may further improve the performance in many applications that employ the MTF approximation.

APPENDIX A
DERIVATION OF (37)

Using the independence assumption of $\mathbf{x}_k(p)$ and $\hat{\mathbf{h}}_k(p)$, the first term on the right of (36) can be expressed as

$$\begin{aligned} &E\left\{\left\|\left[\mathbf{I}-\mu\mathbf{x}_k^*(p)\mathbf{x}_k^T(p)\right]\mathbf{g}_k(p)\right\|^2\right\} \\ &\quad = E\{\|\mathbf{g}_k(p)\|^2\} - 2\mu E\left\{\mathbf{g}_k^H(p)\mathbf{A}_k(p)\mathbf{g}_k(p)\right\} \\ &\quad\quad +\mu^2 E\left\{\mathbf{g}_k^H(p)\mathbf{B}_k(p)\mathbf{g}_k(p)\right\} \end{aligned} \tag{53}$$

where

$$\mathbf{A}_k(p) = E\left\{\mathbf{x}_k^*(p)\mathbf{x}_k^T(p)\right\} \tag{54}$$

and

$$\mathbf{B}_k(p) = E\left\{\mathbf{x}_k^*(p)\mathbf{x}_k^T(p)\mathbf{x}_k^*(p)\mathbf{x}_k^T(p)\right\}. \tag{55}$$

Using the whiteness property of $x_{p,k}$, $\mathbf{A}_k(p)$ reduces to

$$\mathbf{A}_k(p) = \sigma_x^2 \mathbf{I}_{2K+1} \tag{56}$$

where $\mathbf{I}_{2K+1}$ is the identity matrix of size $2K+1 \times 2K+1$. The (m,ℓ)th term of $\mathbf{B}_k(p)$ in (55) can be written as

$$[\mathbf{B}_k(p)]_{m,\ell} = \sum_r E\{x_{p,k-K+r} x^*_{p,k-K+r} x_{p,k-K+\ell} x^*_{p,k-K+m}\} \tag{57}$$

where the index r sums over integer values for which the subscripts of x are defined. By using the fourth-order moment factoring theorem for zero-mean complex Gaussian samples [[19], p. 90], (57) can be rewritten as

$$\begin{aligned}[\mathbf{B}_k(p)]_{m,\ell} = &\sum_r E\{x_{p,k-K+r} x^*_{p,k-K+r}\} \\ &\times E\{x_{p,k-K+\ell} x^*_{p,k-K+m}\} \\ &+ \sum_r E\{x_{p,k-K+r} x^*_{p,k-K+m}\} \\ &\times E\{x_{p,k-K+\ell} x^*_{p,k-K+r}\}\end{aligned} \tag{58}$$

where by using the whiteness property of $x_{p,k}$, we obtain

$$[\mathbf{B}_k(p)]_{m,\ell} = \sigma_x^4 \sum_r \delta(\ell - m) + \sigma_x^4 \sum_r \delta(r-m)\delta(r-\ell). \tag{59}$$

Since r ranges from 0 to $2K+1$, $\mathbf{B}_k(p)$ in (57) reduces to

$$\mathbf{B}_k(p) = 2\sigma_x^4 (K+1) \mathbf{I}_{2K+1}. \tag{60}$$

Substituting (56) and (60) into (53) yields (37).

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their constructive comments and helpful suggestions.

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.*, vol. 15, no. 4, pp. 1305–1319, May 2007.

[2] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 451–459, Sep. 2004.

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

[4] C. Avendano, "Acoustic echo suppression in the STFT domain," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, New Paltz, NY, Oct. 2001, pp. 175–178.

[5] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1–3, pp. 21–34, Nov. 1998.

[6] Y. Lu and J. M. Morris, "Gabor expansion for adaptive echo cancellation," *IEEE Signal Process. Mag.*, vol. 16, no. 2, pp. 68–80, Mar. 1999.

[7] X.-G. Xia, "System identification using chirp signals and time-variant filters in the joint time-frequency domain," *IEEE Trans. Signal Process.*, vol. 45, no. 8, pp. 2072–2084, Aug. 1997.

[8] 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.

[9] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time Fourier transform domain," *IEEE Signal Process. Lett.*, vol. 14, no. 5, pp. 337–340, May 2007.

[10] S. Haykin, *Adaptive Filter Theory*. Upper Saddle River, New Jersey: Prentice-Hall, 2002.

[11] M. R. Portnoff, "Time-frequency representation of digital signals and systems based on short-time Fourier analysis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 1, pp. 55–69, Feb. 1980.

[12] C. Avendano, "Temporal processing of speech in a time-feature space," Ph.D. dissertation, Oregon Graduate Inst. Sci. Technol., Beaverton, Apr. 1997.

[13] M. Dentino, J. M. McCool, and B. Widrow, "Adaptive filtering in the frequency domain," *Proc. IEEE*, vol. 66, no. 12, pp. 1658–1659, Dec. 1978.

[14] D. Mansour and J. A. H. Gray, "Unconstrained frequency-domain adaptive filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-30, no. 5, pp. 726–734, Oct. 1982.

[15] P. C. W. Sommen, "Partitioned frequency domain adaptive filters," in *Proc. 23rd Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, Nov. 1989, pp. 677–681.

[16] J. J. Shynk, "Frequncy-domain and multirate adaptive filtering," *IEEE Signal Process. Mag.*, vol. 9, no. 1, pp. 14–37, Jan. 1992.

[17] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[18] D. R. Brillinger, *Time Series: Data Analysis and Theory*. Philadelphia, PA: SIAM, 2001.

[19] 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.

[20] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, MD: Johns Hopkins Univ. Press, 1996.

[21] L. L. Horowitz and K. D. Senne, "Performance advantage of complex LMS for controlling narrow-band adaptive arrays," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, no. 3, pp. 722–736, Jun. 1981.

[22] K. Mayyas, "Performance analysis of the deficient length LMS adaptive algorithm," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2727–2734, Aug. 2005.

[23] L. Ljung and T. Söderström, *Theory and Practice of Recursive Identification*. Cambridge, MA: MIT Press, 1983.

[24] F. D. Ridder, R. Pintelon, J. Schoukens, and D. P. Gillikin, "Modified AIC and MDL model selection criteria for short data records," *IEEE Trans. Instrum. Meas.*, vol. 54, no. 1, pp. 144–150, Feb. 2005.

[25] G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, 1978.

[26] P. Stoica and Y. Selen, "Model order selection: A review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, Jul. 2004.

[27] G. C. Goodwin, M. Gevers, and B. Ninness, "Quantifying the error in estimated transfer functions with application to model order selection," *IEEE Trans. Autom. Control*, vol. 37, no. 7, pp. 913–928, Jul. 1992.

[28] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. AC-19, no. 6, pp. 716–723, Dec. 1974.

[29] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

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

[31] J. Benesty, T. Gänsler, D. R. Morgan, T. Gdnsler, M. M. Sondhi, and S. L. Gay, *Advances in Network and Acoustic Echo Cancellation*. New York: Springer, 2001.

[32] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*. New York: Wiley, 2004.

[33] C. Breining, P. Dreiseitel, E. Hänsler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, and J. Tlip, "Acoustic echo control," *IEEE Signal Process. Mag.*, vol. 16, no. 4, pp. 42–69, Jul. 1999.

[34] J. Benesty, D. R. Morgan, and J. H. Cho, "A new class of doubletalk detectors based on cross-correlation," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 2, pp. 168–172, Mar. 2000.
[35] J. H. Cho, D. R. Morgan, and J. Benesty, "An objective technique for evaluating doubletalk detectors in acoustic echo cancelers," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 6, pp. 718–724, Nov. 1999.
[36] Y. Avargel, homepage. [Online]. Available: http://siglab.technion.ac.il/~yekutiel
[37] A. Guerin, G. Faucon, and R. L. Bouquin-Jeannes, "Nonlinear acoustic echo cancellation based on Volterra filters," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 672–683, Nov. 2003.
[38] A. Stenger and W. Kellermann, "Adaptation of a memoryless preprocessor for nonlinear acoustic echo cancelling," *Signal Process.*, vol. 80, pp. 1747–1760, Sep. 2000.
[39] I. Cohen, "Identification of speech source coupling between sensors in reverberant noisy environments," *IEEE Signal Process. Lett.*, vol. 11, no. 7, pp. 613–616, Jul. 2004.
[40] F. Riera-Palou, J. M. Noras, and D. G. M. Cruickshank, "Linear equalisers with dynamic and automatic length selection," *Electron. Lett.*, vol. 37, no. 25, pp. 1553–1554, Dec. 2001.
[41] R. C. Bilcu, P. Kuosmanen, and K. Egiazarian, "A new variable length LMS algorithm: Theoretical analysis and implementations," in *Proc. 9th IEEE Int. Conf. Electron., Circuits, Syst.*, Sep. 2002, vol. 3, pp. 1031–1034.
[42] Y. Gu, K. Tang, H. Cui, and W. Du, "Convergence analysis of a deficient-length LMS filter and optimal-length sequence to model exponential decay impulse response," *IEEE Signal Process. Lett.*, vol. 10, no. 1, pp. 4–7, Jan. 2003.
[43] Y. Gu, K. Tang, and H. Cui, "LMS algorithm with gradient descent filter length," *IEEE Signal Process. Lett.*, vol. 11, no. 3, pp. 305–307, Mar. 2004.
[44] Y. Gong and C. F. N. Cowan, "An LMS style variable tap-length algorithm for structure adaptation," *IEEE Trans. Signal Process.*, vol. 53, no. 7, pp. 2400–2407, Jul. 2005.

Yekutiel Avargel (S'06) received the B.Sc. degree in electrical engineering from the Technion—Israel Institute of Technology, Haifa, in 2004, where he is currently pursuing the Ph.D. degree in electrical engineering.

From 2003 to 2004, he was a Research Engineer at RAFAEL Research Laboratories, Haifa, Israel Ministry of Defense. Since 2004, he has been a Research Assistant and a Project Supervisor with the Signal and Image Processing Lab (SIPL), Electrical Engineering Department, Technion. His research interests are statistical signal processing, system identification, adaptive filtering, and digital speech 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, 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. He is a Coeditor of the Multichannel Speech Processing section of the *Springer Handbook of Speech Processing* and serves as a 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.

Dr. Cohen received the Technion Excellent Lecturer awards in 2005 and 2006. He serves as Associate Editor of the IEEE SIGNAL PROCESSING LETTERS. He served as Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING.