System Identification in the Short-Time Fourier Transform Domain with Cross-Band Filtering

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

Abstract

In this paper, we investigate the influence of cross-band filters on a system identifier implemented in the short-time Fourier transform (STFT) domain. We derive analytical relations between the number of cross-band filters, which are useful for system identification in the STFT domain, and the power and length of the input signal. We show that increasing the number of cross-band filters not necessarily implies a lower steady-state mean-square error (MSE) in subbands. The number of useful cross-band filters depends on the power ratio between the input signal and the additive noise signal. Furthermore, it depends on the effective length of input signal employed for system identification, which is restricted to enable tracking capability of the algorithm during time variations in the system. As the power of input signal increases or as the time variations in the system become slower, a larger number of cross-band filters may be utilized. The proposed subband approach is compared to the conventional fullband approach and to the commonly-used subband approach that relies on multiplicative transfer function (MTF) approximation. The comparison is carried out in terms of MSE performance and computational complexity. Experimental results verify the theoretical derivations and demonstrate the relations between the number of useful cross-band filters and the power and length of the input signal.

Index Terms

System identification, echo suppression, subband filtering, subband acoustic echo cancellers, short-time Fourier transform, time-frequency analysis.

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

The authors are with the Department of Electrical Engineering, Technion - Israel Institute of Technology, Technion City, Haifa 32000, Israel. E-mail addresses: kutiav@tx.technion.ac.il (Y. Avargel), icohen@ee.technion.ac.il (I. Cohen); tel.: +972-4-8294731; fax: +972-4-8295757.

I. INTRODUCTION

Identification of systems with long impulse responses is of major importance in many applications, including acoustic echo cancellation [1], [2], relative transfer function (RTF) identification [3], dereverberation [4], [5], blind source separation [6], [7] and beamforming in reverberant environments [8], [9]. In acoustic echo cancellation applications, a loudspeaker-enclosure-microphone (LEM) system needs to be identified in order to reduce the coupling between loudspeakers and microphones. A typical acoustic echo canceller (AEC) for an LEM system is depicted in Fig. 1. The far-end signal $x(n)$ propagates through the enclosure, which is characterized by a time-varying impulse response $h(n)$, and received in the microphone as an echo signal $d(n)$ together with the near-end speaker and a local noise. To cancel the echo signal, we commonly identify the echo path impulse response using an adaptive transversal filter $\hat{h}(n)$ and produce an echo estimate $\hat{d}(n)$. The cancellation is then accomplished by subtracting the echo estimate from the microphone signal. Adaptation algorithms used for the purpose of system identification are generally of a gradient type (*e.g.*, least-mean-square (LMS) algorithm) and are known to attain acceptable performances in several applications, especially when the length of the adaptive filter is relatively short. However, in applications like acoustic echo cancellation, the number of filter taps that need to be considered is several thousands, which leads to high computational complexity and slow convergence rate of the adaptive algorithm. Moreover, when the input signal to the adaptive filter is correlated, which is often the case in acoustic echo cancellation applications, the adaptive algorithm suffers from slow convergence rate [10].

To overcome these problems, block processing techniques have been introduced [10], [11]. These techniques partition the input data into blocks and perform the adaptation in the frequency domain to achieve computational efficiency. However, block processing introduces a delay in the signal paths and reduces the time-resolution required for control purposes. Alternatively, the loudspeaker and microphone signals are filtered into subbands, then decimated and processed in distinct subbands (*e.g.*, [12]–[18]). The computational complexity is reduced and the convergence rate is improved due to the shorter independent filters in subbands. However, as in block processing structures, subband techniques introduce a delay into the system by the analysis and synthesis filter banks. Moreover, they produce aliasing effects because of the decimation, which necessitates cross-band filters between the subbands [16], [19].

It has been found [16] that the convergence rate of subband adaptive filters that involve cross-band filters with critical sampling is worse than that of fullband adaptive filters. Several techniques to avoid cross-band filters have been proposed, such as inserting spectral gaps between the subbands [12], employing auxiliary

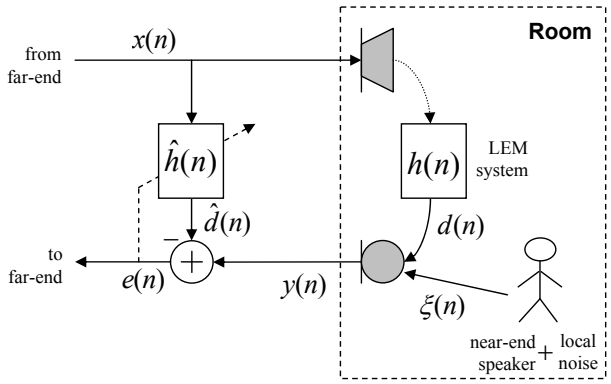


Fig. 1: A typical acoustic echo canceller (AEC) for a loudspeaker-enclosure-microphone (LEM) system.

subbands [15], using polyphase decomposition of the filter [17] and oversampling of the filter-bank outputs [13], [14]. Spectral gaps impair the subjective quality and are especially annoying when the number of subbands is large, while the other approaches are costly in terms of computational complexity. Some time-frequency representations, such as the short-time Fourier transform (STFT) have been introduced for the implementation of subband adaptive filtering [20]–[23]. A typical system identification scheme in the STFT domain is illustrated in Fig. 2. The block $\hat{\mathbf{H}}$ represents a matrix of adaptive filters which models the system $h(n)$ in the STFT domain. The off-diagonal terms of $\hat{\mathbf{H}}$ (if exist) correspond to the cross-band filters, while the diagonal terms represent the band-to-band filters. Recently, we analyzed the performance of an LMS-based direct adaptive algorithm used for the adaptation of cross-band filters in the STFT domain [24].

In this paper, we consider an offline system identification in the STFT domain using the least squares (LS) criterion, and investigate the influence of cross-band filters on its performance. We derive analytical relations between the input signal-to-noise ratio (SNR), the length of the input signal, and the number of cross-band filters which are useful for system identification in the STFT domain. We show that increasing the number of cross-band filters not necessarily implies a lower steady-state MSE in subbands. The number of cross-band filters, that are useful for system identification in the STFT domain, depends on the length and power of the input signal. More specifically, it depends on the SNR, *i.e.* the power ratio between the input signal and the additive noise signal, and on the effective length of input signal employed for system identification. The effective length of input signal employed for the system identification is restricted to enable tracking capability of the algorithm during time variations in the impulse response.

We show that as the SNR increases or as the time variations in the impulse response become slower (which enables to use longer segments of the input signal), the number of cross-band filters that should be

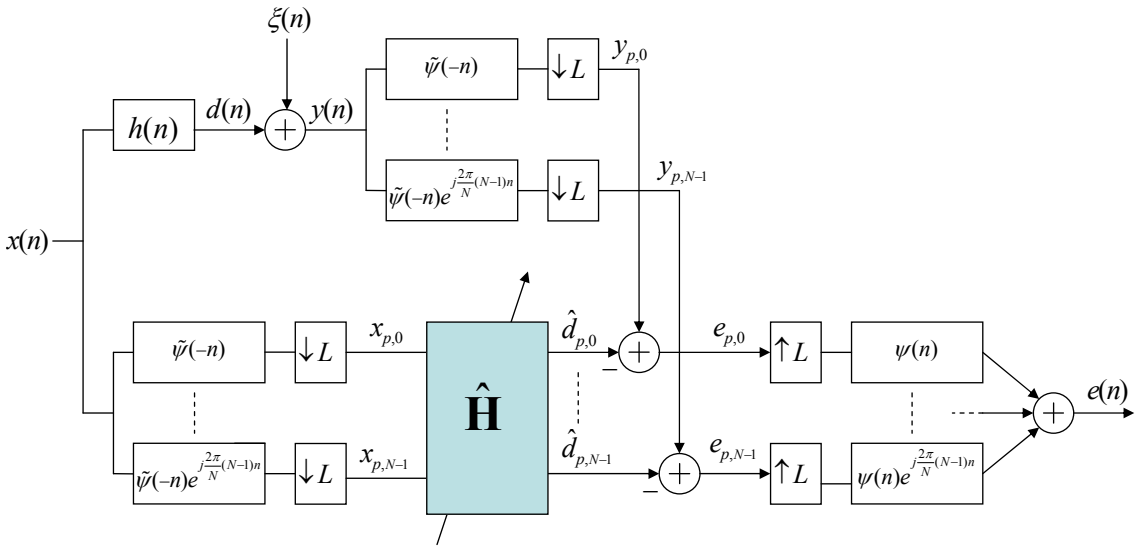


Fig. 2: System identification scheme in the STFT domain. The unknown system $h(n)$ is modeled by the block $\hat{\mathbf{H}}$ in the STFT domain.

estimated to achieve the minimal MSE increases. Moreover, as the SNR increases, the MSE that can be achieved by the proposed approach is lower than that obtainable by the commonly-used subband approach that relies on long STFT analysis window and multiplicative transfer function (MTF) approximation. Experimental results obtained using synthetic white Gaussian signals and real speech signals verify the theoretical derivations and demonstrate the relations between the number of useful cross-band filters and the power and length of the input signal.

The paper is organized as follows. In Section II, we briefly review the representation of digital signals and linear time-invariant (LTI) systems in the STFT domain and derive relations between the cross-band filters in the STFT domain and the impulse response in the time domain. In Section III, we consider the problem of system identification in the STFT domain and formulate an LS optimization criterion for estimating the cross-band filters. In Section IV, we derive an explicit expression for the attainable MMSE in subbands. In Section V, we explore the influence of both the input SNR and the observable data length on the MMSE performance. In Section VI, we address the computational complexity of the proposed approach and compare it to that of the conventional fullband and MTF approaches. Finally, in Section VII, we present simulation results which verify the theoretical derivations.

II. Representation of LTI Systems in the STFT Domain

In this section, we briefly review the representation of digital signals and LTI systems in the STFT domain. For further details, see *e.g.*, [25], [26]. We also derive relations between the cross-band filters in the STFT domain and the impulse response in the time domain, and show that the number of cross-band filters required for the representation of an impulse response is mainly determined by the analysis and synthesis windows employed for the STFT. Throughout this work, unless explicitly noted, the summation indexes range from $-\infty$ to ∞.

The STFT representation of a signal $x(n)$ is given by

$$x_{p,k} = \sum_m x(m)\tilde{\psi}^*_{p,k}(m)\,, \tag{1}$$

where

$$\tilde{\psi}_{p,k}(n) \triangleq \tilde{\psi}(n-pL)e^{j\frac{2\pi}{N}k(n-pL)}\,, \tag{2}$$

$\tilde{\psi}(n)$ denotes an analysis window (or analysis filter) of length N, p is the frame index, k represents the frequency-band index, L is the discrete-time shift (in filter bank interpretation L denotes the decimation factor as illustrated in Fig. 2) and * denotes complex conjugation. The inverse STFT, *i.e.*, reconstruction of $x(n)$ from its STFT representation $x_{p,k}$, is given by

$$x(n) = \sum_p \sum_{k=0}^{N-1} x_{p,k}\psi_{p,k}(n)\,, \tag{3}$$

where

$$\psi_{p,k}(n) \triangleq \psi(n-pL)e^{j\frac{2\pi}{N}k(n-pL)} \tag{4}$$

and $\psi(n)$ denotes a synthesis window (or synthesis filter) of length N. Throughout this paper, we assume that $\tilde{\psi}(n)$ and $\psi(n)$ are real functions. Substituting (1) into (3), we obtain the so-called completeness condition:

$$\sum_p \psi(n-pL)\tilde{\psi}(n-pL) = \frac{1}{N} \qquad \text{for all } n\,. \tag{5}$$

Given analysis and synthesis windows that satisfy (5), a signal $x(n) \in \ell_2(\mathbb{Z})$ is guaranteed to be perfectly reconstructed from its STFT coefficients $x_{p,k}$. However, for $L \leq N$ and for a given synthesis window $\psi(n)$, there might be an infinite number of solutions to (5); therefore, the choice of the analysis window is generally not unique [27], [28].

We now proceed with an STFT representation of LTI systems. Let $h(n)$ denote a length Q impulse response of an LTI system, whose input $x(n)$ and output $d(n)$ are related by

$$d(n) = \sum_{i=0}^{Q-1} h(i)x(n-i)\,. \tag{6}$$

In the STFT domain, we obtain after some manipulations (see Appendix I)

$$d_{p,k} = \sum_{k'=0}^{N-1} \sum_{p'} x_{p',k'} h_{p-p',k,k'} = \sum_{k'=0}^{N-1} \sum_{p'} x_{p-p',k'} h_{p',k,k'}\,, \tag{7}$$

where $h_{p-p',k,k'}$ may be interpreted as a response to an impulse $\delta_{p-p',k-k'}$ in the time-frequency domain (the impulse response is translation-invariant in the time axis and is translation varying in the frequency axis). The impulse response $h_{p,k,k'}$ in the time-frequency domain is related to the impulse response $h(n)$ in the time domain by

$$h_{p,k,k'} = \{h(n) * \phi_{k,k'}(n)\}\big|_{n=pL} \triangleq \bar{h}_{n,k,k'}\big|_{n=pL}\,, \tag{8}$$

where $*$ denotes convolution with respect to the time index n and

$$\begin{aligned} \phi_{k,k'}(n) &\triangleq e^{j\frac{2\pi}{N}k'n} \sum_{m} \tilde{\psi}(m)\psi(n+m)e^{-j\frac{2\pi}{N}m(k-k')} \\ &= e^{j\frac{2\pi}{N}k'n}\psi_{n,k-k'}\,, \end{aligned} \tag{9}$$

where $\psi_{n,k}$ is the STFT representation of the synthesis window $\psi(n)$ calculated with a decimation factor $L = 1$. Equation (7) indicates that for a given frequency-band index k, the temporal signal $d_{p,k}$ can be obtained by convolving the signal $x_{p,k'}$ in each frequency-band k' ($k' = 0, 1, \ldots, N-1$) with the corresponding filter $h_{p,k,k'}$ and then summing over all the outputs. We refer to $h_{p,k,k'}$ for $k = k'$ as a band-to-band filter and for $k \neq k'$ as a cross-band filter. Cross-band filters are used for canceling the aliasing effects caused by the subsampling. Note that equation (8) implies that for fixed k and k', the filter $h_{p,k,k'}$ is noncasual in general, with $\left\lceil \frac{N}{L} \right\rceil - 1$ noncasual coefficients. In echo cancellation applications, in order to consider those coefficients, an extra delay of $\left(\left\lceil \frac{N}{L} \right\rceil - 1\right) L$ samples is generally introduced into the microphone signal ($y(n)$ in Fig. 1) [13]. It can also be seen from (8) that the length of each cross-band filter is given by

$$N_h = \left\lceil \frac{Q+N-1}{L} \right\rceil + \left\lceil \frac{N}{L} \right\rceil - 1\,. \tag{10}$$

To illustrate the significance of the cross-band filters, we apply the discrete-time Fourier transform (DTFT) to the undecimated cross-band filter $\bar{h}_{n,k,k'}$ (defined in (8)) with respect to the time index n and obtain

$$\bar{H}_{k,k'}(\theta) = \sum_{n} \bar{h}_{n,k,k'} e^{-jn\theta} = H(\theta)\tilde{\Psi}(\theta - \frac{2\pi}{N}k)\Psi(\theta - \frac{2\pi}{N}k')\,, \tag{11}$$

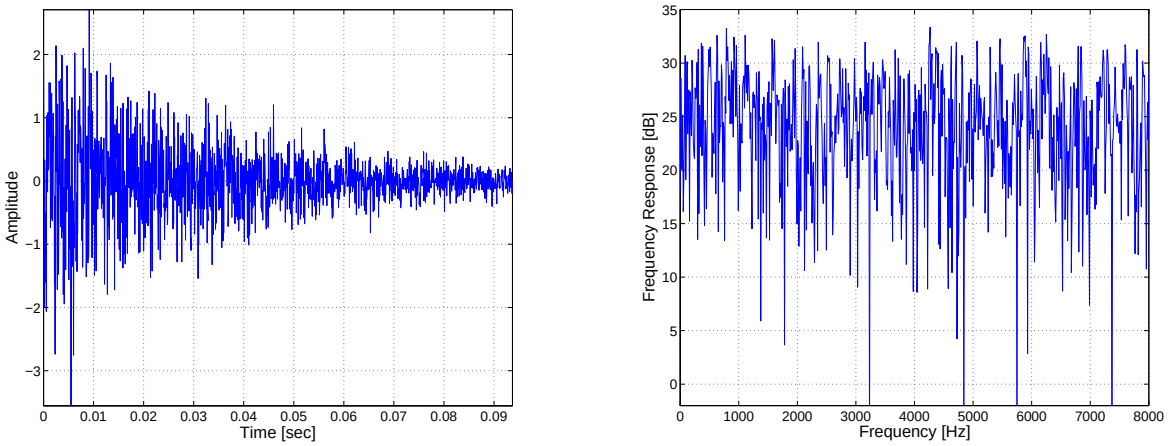
where $H(\theta)$, $\tilde{\Psi}(\theta)$ and $\Psi(\theta)$ are the DTFT of $h(n)$, $\tilde{\psi}(n)$ and $\psi(n)$, respectively. Had both $\tilde{\Psi}(\theta)$ and $\Psi(\theta)$ been ideal low-pass filters with bandwidth $f_s/2N$ (where f_s is the sampling frequency), a perfect STFT representation of the system $h(n)$ could be achieved by using just the band-to-band filter $h_{n,k,k}$, since in this case the product of $\tilde{\Psi}(\theta - \frac{2\pi}{N}k)$ and $\Psi(\theta - \frac{2\pi}{N}k')$ is identically zero for $k \neq k'$. However, the bandwidths of $\tilde{\Psi}(\theta)$ and $\Psi(\theta)$ are generally greater than $f_s/2N$ and therefore, $\bar{H}_{k,k'}(\theta)$ and $\bar{h}_{n,k,k'}$ are not zero for $k \neq k'$. One can observe from (11) that the energy of a cross-band filter from frequency-band k' to frequency-band k decreases as $|k - k'|$ increases, since the overlap between $\tilde{\Psi}(\theta - \frac{2\pi}{N}k)$ and $\Psi(\theta - \frac{2\pi}{N}k')$ becomes smaller. As a result, relatively few cross-band filters need to be considered in order to capture most of the energy of the STFT representation of $h(n)$.

Figure 3 illustrates a synthetic LEM impulse response based on a statistical reverberation model, which assumes that a room impulse response can be described as a realization of a nonstationary stochastic process $h(n) = u(n)\beta(n)e^{-\alpha n}$, where $u(n)$ is a step function (*i.e.*, $u(n) = 1$ for $n \geq 0$, and $u(n) = 0$ otherwise), $\beta(n)$ 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 our example, α corresponds to $T_{60} = 300$ ms (where $f_s = 16$ kHz) and $\beta(n)$ has a unit variance.

To compare the cross-band filters obtained for this synthetic impulse response with those obtained in anechoic chamber (*i.e.*, impulse response $h(n) = \delta(n)$), we employed a Hamming synthesis window of length $N = 256$, and computed a minimum energy analysis window $\tilde{\psi}(n)$ that satisfies (5) for $L = 128$ (50% overlap) [27]. Then we computed the undecimated cross-band filters $\bar{h}_{n,k,k'}$ using (8). Figures 4(a) and (b) show mesh plots of the $\left|\bar{h}_{n,1,k'}\right|$ and contours at -40 dB (values outside this contour are lower than -40 dB) for $h(n) = \delta(n)$ and for the synthetic impulse response depicted in Fig. 3. Figure 4(c) shows an ensemble averaging of $\left|\bar{h}_{n,1,k'}\right|^2$ over realizations of the stochastic process $h(n) = u(n)\beta(n)e^{-\alpha n}$ which is given by

$$E\left\{\left|\bar{h}_{n,1,k'}\right|^2\right\} = u(n)e^{-2\alpha n} * \left|\phi_{1,k'}(n)\right|^2 . \tag{12}$$

Recall that the cross-band filter $h_{p,k,k'}$ is obtained from $\bar{h}_{n,k,k'}$ by decimating the time index n by a factor of L (see (8)). We observe from Fig. 4 that most of the energy of $\bar{h}_{n,k,k'}$ (for both anechoic chamber and the LEM reverberation model) is concentrated in the eight cross-band filters, *i.e.*, $k' \in \{(k+i) \bmod N |\ \ i = -4, \ldots, 4\}$; therefore, both impulse responses may be represented in the time-frequency domain by using only eight cross-band filters around each frequency-band. As expected from (11), the number of cross-band filters required for the representation of an impulse response is mainly determined by the analysis and synthesis windows, while the length of the cross-band filters (with respect



(a) (b)

Fig. 3: (a) A synthetic LEM impulse response: $h(n) = \beta(n)e^{-\alpha n}$ and (b) its frequency response. $\beta(n)$ is unit-variance white Gaussian noise and α corresponds to $T_{60} = 300$ ms (sampling rate is 16 kHz).

to the time index n) is related to the length of the impulse response.

III. System Identification in the STFT Domain

In this section, we consider system identification in the STFT domain and address the problem of estimating the cross-band filters of the system using an LS optimization criterion for each frequency-band. Throughout this section, scalar variables are written with lowercase letters and vectors are indicated with lowercase boldface letters. Capital boldface letters are used for matrices and norms are always ℓ_2 norms.

Consider the STFT-based system identification scheme as illustrated in Fig. 2. The input signal $x(n)$ passes through an unknown system characterized by its impulse response $h(n)$, obtaining the desired signal $d(n)$. Together with the corrupting noise signal $\xi(n)$, the system output signal is given by

$$y(n) = d(n) + \xi(n) = h(n) * x(n) + \xi(n)\,. \tag{13}$$

Note that the noise signal $\xi(n)$ may often include a useful signal, as in acoustic echo cancellation where it consists of the near-end speaker signal as well as a local noise. From (13) and (7), the STFT of $y(n)$ may be written as

$$y_{p,k} = d_{p,k} + \xi_{p,k} = \sum_{k'=0}^{N-1} \sum_{p'=0}^{N_h-1} x_{p-p',k'} h_{p',k,k'} + \xi_{p,k}\,, \tag{14}$$

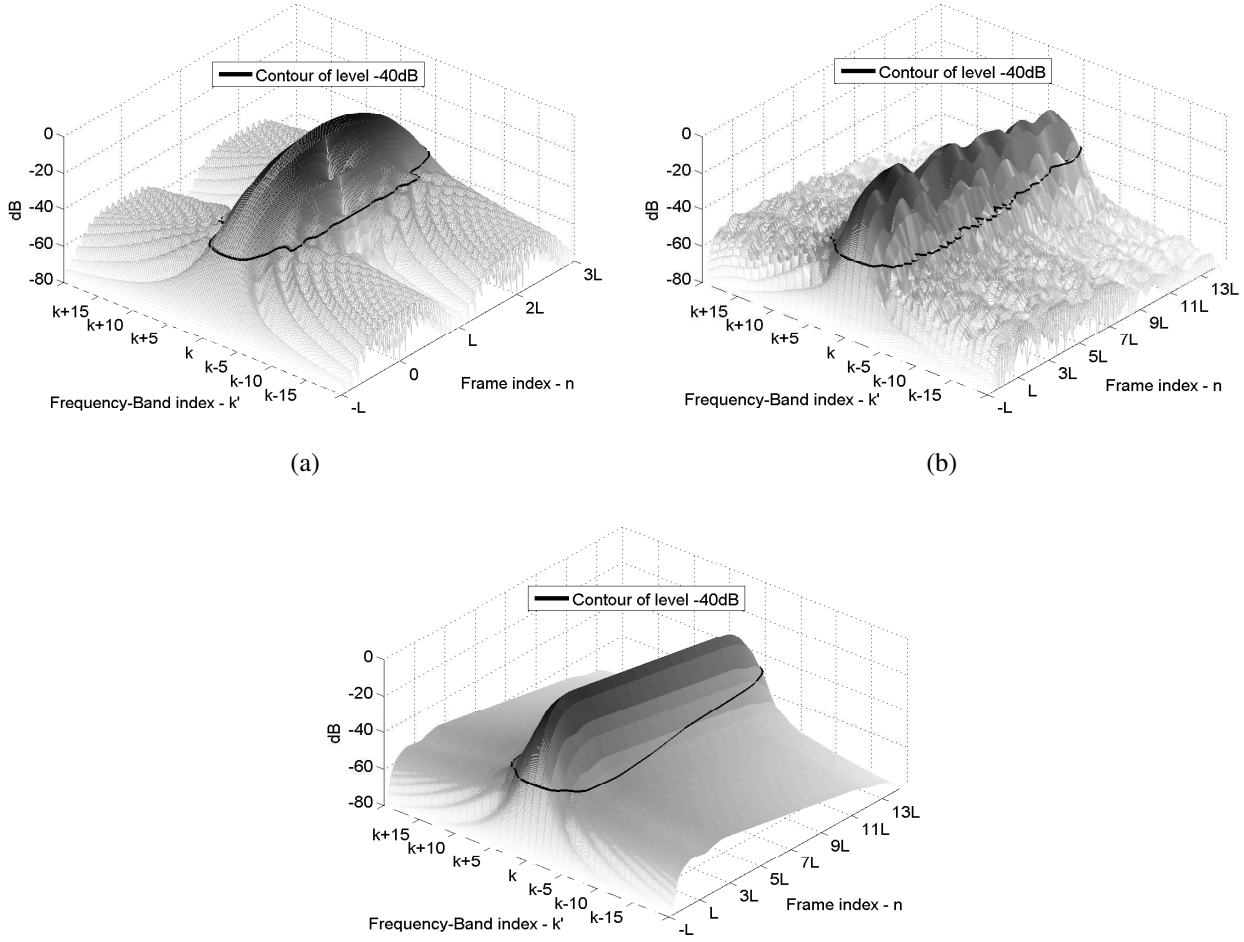


Fig. 4: A mesh plot of the cross-band filters $|\bar{h}_{n,1,k'}|$ for different impulse responses. (a) An anechoic chamber impulse response: $h(n) = \delta(n)$. (b) An LEM synthetic impulse response: $h(n) = u(n)\beta(n)e^{-\alpha n}$, where $u(n)$ is a step function, $\beta(n)$ is zero-mean unit-variance white Gaussian noise and α corresponds to $T_{60} = 300$ ms (sampling rate is 16 kHz). (c) An ensemble averaging $E|\bar{h}_{n,1,k'}|^2$ of the impulse response given in (b).

where N_h is the length of the cross-band filters. Here, we do not consider the case where the cross-band filters in the k-th frequency-band are shorter than the band-to-band filter, as in [16]. We assume that all the filters have the same length N_h. Defining N_x as the length of $x_{p,k}$ in frequency band k, we can write the length of $y_{p,k}$ for a fixed k as $N_y = N_x + N_h - 1$. It is worth noting that due to the noncasuality of the filter $h_{p,k,k'}$ (see Section II), the index p' in (14) should have ranged from $-\left\lceil \frac{N}{L} \right\rceil + 1$ to $N_h - \left\lceil \frac{N}{L} \right\rceil$, where $\left\lceil \frac{N}{L} \right\rceil - 1$ is the number of noncasual coefficients of $h_{p,k,k'}$. However, we assume that an artificial delay

of $\left(\lceil \frac{N}{L} \rceil - 1\right) L$ samples has been introduced into the system output signal $y(n)$ in order to compensate for those noncasual coefficients, so the signal $y_{p,k}$ in (14) corresponds to the STFT of a delayed signal $y\left(n - \left(\lceil \frac{N}{L} \rceil - 1\right) L\right)$. Therefore, both p and p' take on values starting with 0 rather than with $-\lceil \frac{N}{L} \rceil + 1$.

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

$$\mathbf{h}_{k,k'} = \left[\begin{array}{cccc} h_{0,k,k'} & h_{1,k,k'} & \cdots & h_{N_h-1,k,k'} \end{array} \right]^T \tag{15}$$

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

$$\mathbf{h}_k = \left[\begin{array}{ccccc} \mathbf{h}_{k,0}^T & \mathbf{h}_{k,1}^T & \cdots & \cdots & \mathbf{h}_{k,N-1}^T \end{array} \right]^T . \tag{16}$$

Let

$$\mathbf{X}_k = \left[\begin{array}{ccccc} x_{0,k} & 0 & \cdots & \cdots & 0 \\ x_{1,k} & x_{0,k} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{N_y-1,k} & \cdots & \cdots & \cdots & x_{N_y+N_h-2,k} \end{array} \right] \tag{17}$$

represent an $N_y \times N_h$ Toeplitz matrix constructed from the input signal STFT coefficients of the k-th frequency-band, and let $\mathbf{\Delta}_k$ be a concatenation of $\{\mathbf{X}_k\}_{k=0}^{N-1}$ along the column dimension

$$\mathbf{\Delta}_k = \left[\begin{array}{ccccc} \mathbf{X}_0 & \mathbf{X}_1 & \cdots & \cdots & \mathbf{X}_{N-1} \end{array} \right] . \tag{18}$$

Then, (14) can be written in a vector form as

$$\mathbf{y}_k = \mathbf{d}_k + \boldsymbol{\xi}_k = \mathbf{\Delta}_k \mathbf{h}_k + \boldsymbol{\xi}_k \, , \tag{19}$$

where

$$\mathbf{y}_k = \left[\begin{array}{cccc} y_{0,k} & y_{1,k} & y_{2,k} & \cdots & y_{N_y-1,k} \end{array} \right]^T \tag{20}$$

represents the output signal STFT coefficients of the k-th frequency-band, and the vectors $\mathbf{d}_k$ and $\boldsymbol{\xi}_k$ are defined similarly.

Let $\hat{h}_{p',k,k'}$ be an estimate of the cross-band filter $h_{p',k,k'}$, and let $\hat{d}_{p,k}$ be the resulting estimate of $d_{p,k}$ using only $2K$ cross-band filters around the frequency-band k, *i.e.*,

$$\hat{d}_{p,k} = \sum_{k'=k-K}^{k+K} \sum_{p'=0}^{N_h-1} \hat{h}_{p',k,k' \bmod N} x_{p-p',k' \bmod N} \, , \tag{21}$$

where we exploited the periodicity of the frequency-bands (see an example illustrated in Fig. 5). Let $\hat{\tilde{\mathbf{h}}}_k$ be the $2K+1$ estimated filters at frequency band k

$$\hat{\tilde{\mathbf{h}}}_k = \left[\begin{array}{cccc} \hat{\mathbf{h}}_{k,(k-K) \bmod N}^T & \hat{\mathbf{h}}_{k,(k-K+1) \bmod N}^T & \cdots & \cdots & \hat{\mathbf{h}}_{k,(k+K) \bmod N}^T \end{array} \right]^T , \tag{22}$$

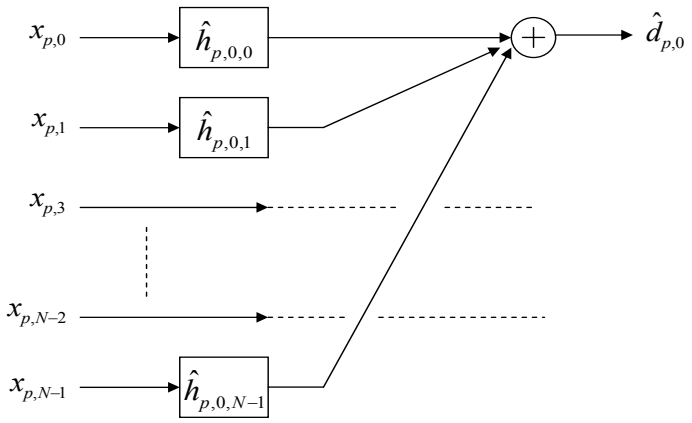


Fig. 5: Cross-band filters illustration for frequency-band $k = 0$ and $K = 1$.

where $\hat{\mathbf{h}}_{k,k'}$ is the estimated cross-band filter from frequency-band k' to frequency-band k, and let $\tilde{\mathbf{\Delta}}_k$ be a concatenation of $\{\mathbf{X}_{k'}\}_{k'=(k-K) \bmod N}^{(k+K) \bmod N}$ along the column dimension

$$\tilde{\mathbf{\Delta}}_k = \left[\begin{array}{cccccc} \mathbf{X}_{(k-K) \bmod N} & \mathbf{X}_{(k-K+1) \bmod N} & \cdots & \cdots & \mathbf{X}_{(k+K) \bmod N} \end{array} \right] . \tag{23}$$

Then, the estimated desired signal can be written in a vector form as

$$\hat{\mathbf{d}}_k = \tilde{\mathbf{\Delta}}_k \hat{\tilde{\mathbf{h}}}_k \, , \tag{24}$$

Note that both $\hat{\tilde{\mathbf{h}}}_k$ and $\hat{\mathbf{d}}_k$ depend on the parameter K, but for notational simplicity K has been omitted. Using the above notations, the LS optimization problem can be expressed as

$$\hat{\tilde{\mathbf{h}}}_k = \arg \min_{\tilde{\mathbf{h}}_k} \left\| \mathbf{y}_k - \tilde{\mathbf{\Delta}}_k \tilde{\mathbf{h}}_k \right\|^2 . \tag{25}$$

The solution to (25) is given by

$$\hat{\tilde{\mathbf{h}}}_k = \left(\tilde{\mathbf{\Delta}}_k^H \tilde{\mathbf{\Delta}}_k \right)^{-1} \tilde{\mathbf{\Delta}}_k^H \mathbf{y}_k \, , \tag{26}$$

where we assumed that $\tilde{\mathbf{\Delta}}_k^H \tilde{\mathbf{\Delta}}_k$ is not singular[1]. Substituting (26) into (24), we obtain an estimate of the desired signal in the STFT domain at the k-th frequency-band, using $2K$ cross-band filters. Our objective is to analyze the MSE in each frequency-band, and investigate the influence of the number of estimated cross-band filters on the MSE performance.

[1]In the ill-conditioned case, when $\tilde{\mathbf{\Delta}}_k^H \tilde{\mathbf{\Delta}}_k$ is singular, matrix regularization is required [29].

IV. MSE ANALYSIS

In this section, we derive an explicit expression for the MMSE obtainable in the k-th frequency-band[2]. To make the following 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. We also assume that $x_{p,k}$ is statistically independent of $\xi_{p,k}$. The Gaussian assumption of the corresponding STFT signals is often justified by a version of the central limit theorem for correlated signals [30, Theorem 4.4.2], and it underlies the design of many speech-enhancement systems [31], [32].

The (normalized) MSE is defined by

$$\epsilon_k(K) = \frac{E\left\{\left\|\mathbf{d}_k - \hat{\mathbf{d}}_k\right\|^2\right\}}{E\left\{\left\|\mathbf{d}_k\right\|^2\right\}}, \tag{27}$$

Substituting (24) and (26) into (27), the MSE can be expressed as

$$\begin{aligned}\epsilon_k(K) &= \frac{1}{E\left\{\left\|\mathbf{d}_k\right\|^2\right\}} E\left\{\left\|\left[1 - \tilde{\mathbf{\Delta}}_k\left(\tilde{\mathbf{\Delta}}_k^H\tilde{\mathbf{\Delta}}_k\right)^{-1}\tilde{\mathbf{\Delta}}_k^H\right]\mathbf{d}_k\right\|^2\right\} \\ &\quad + \frac{1}{E\left\{\left\|\mathbf{d}_k\right\|^2\right\}} E\left\{\left\|\tilde{\mathbf{\Delta}}_k\left(\tilde{\mathbf{\Delta}}_k^H\tilde{\mathbf{\Delta}}_k\right)^{-1}\tilde{\mathbf{\Delta}}_k^H\boldsymbol{\xi}_k\right\|^2\right\}. \end{aligned} \tag{28}$$

Equation (28) can be rewritten as

$$\epsilon_k(K) = 1 + \epsilon_1 - \epsilon_2, \tag{29}$$

where

$$\epsilon_1 = \frac{1}{E\left\{\left\|\mathbf{d}_k\right\|^2\right\}} E\left\{\boldsymbol{\xi}_k^H\tilde{\mathbf{\Delta}}_k\left(\tilde{\mathbf{\Delta}}_k^H\tilde{\mathbf{\Delta}}_k\right)^{-1}\tilde{\mathbf{\Delta}}_k^H\boldsymbol{\xi}_k\right\} \tag{30}$$

and

$$\epsilon_2 = \frac{1}{E\left\{\left\|\mathbf{d}_k\right\|^2\right\}} E\left\{\mathbf{d}_k^H\tilde{\mathbf{\Delta}}_k\left(\tilde{\mathbf{\Delta}}_k^H\tilde{\mathbf{\Delta}}_k\right)^{-1}\tilde{\mathbf{\Delta}}_k^H\mathbf{d}_k\right\}. \tag{31}$$

To proceed with the mean-square analysis, we derive simplified expressions for ϵ_1 and ϵ_2. Recall that for any two vectors $\mathbf{a}$ and $\mathbf{b}$ we have $\mathbf{a}^H\mathbf{b} = tr(\mathbf{a}\mathbf{b}^H)^*$, where the operator $tr(\cdot)$ denotes the trace of a matrix. Then ϵ_1 can be expressed as

$$\epsilon_1 = \frac{1}{E\left\{\left\|\mathbf{d}_k\right\|^2\right\}} tr\left(E\left\{\boldsymbol{\xi}_k\boldsymbol{\xi}_k^H\right\} E\left\{\tilde{\mathbf{\Delta}}_k\left(\tilde{\mathbf{\Delta}}_k^H\tilde{\mathbf{\Delta}}_k\right)^{-1}\tilde{\mathbf{\Delta}}_k^H\right\}\right)^*. \tag{32}$$

[2]We are often interested in the time-domain MMSE, *i.e.*, in the MMSE of $\hat{d}(n)$. However, the time-domain MMSE is related to the sum of MMSEs in all the frequency-bands.

The whiteness assumption for $\xi_{p,k}$ yields $E\left\{\boldsymbol{\xi}_k\boldsymbol{\xi}_k^H\right\} = \sigma_\xi^2\mathbf{I}_{N_y\times N_y}$, where $\mathbf{I}_{N_y\times N_y}$ is an identity matrix of size $N_y \times N_y$. Using the property that $tr(AB) = tr(BA)$ for any two matrices A and B, we have

$$\begin{aligned}\epsilon_1 &= \frac{1}{E\left\{\|\mathbf{d}_k\|^2\right\}}\sigma_\xi^2 E\left\{tr\left(\tilde{\boldsymbol{\Delta}}_k^H\tilde{\boldsymbol{\Delta}}_k\left(\tilde{\boldsymbol{\Delta}}_k^H\tilde{\boldsymbol{\Delta}}_k\right)^{-1}\right)^*\right\}\\ &= \frac{1}{E\left\{\|\mathbf{d}_k\|^2\right\}}\sigma_\xi^2 E\left\{tr\left(\mathbf{I}_{(2K+1)N_h\times(2K+1)N_h}\right)^*\right\}\\ &= \frac{\sigma_\xi^2 N_h\left(2K+1\right)}{E\left\{\|\mathbf{d}_k\|^2\right\}}\,. \end{aligned} \tag{33}$$

Using (19), $E\left\{\|\mathbf{d}_k\|^2\right\}$ can be expressed as

$$E\left\{\|\mathbf{d}_k\|^2\right\} = \mathbf{h}_k^H E\left\{\boldsymbol{\Delta}_k^H\boldsymbol{\Delta}_k\right\}\mathbf{h}_k\,, \tag{34}$$

and by using the whiteness property of $x_{p,k}$, the (m,l)-th term of $E\left\{\boldsymbol{\Delta}_k^H\boldsymbol{\Delta}_k\right\}$ is given by

$$\begin{aligned}\left(E\left\{\boldsymbol{\Delta}_k^H\boldsymbol{\Delta}_k\right\}\right)_{m,l} &= \sum_n E\left\{x_{n-l\bmod N_h,\left\lfloor\frac{l}{N_h}\right\rfloor}x^*_{n-m\bmod N_h,\left\lfloor\frac{m}{N_h}\right\rfloor}\right\}\\ &= \sum_n \sigma_x^2\delta(l\bmod N_h - m\bmod N_h)\delta\left(\left\lfloor\frac{l}{N_h}\right\rfloor - \left\lfloor\frac{m}{N_h}\right\rfloor\right)\\ &= N_x\sigma_x^2\delta(l-m)\,. \end{aligned} \tag{35}$$

Accordingly, $E\left\{\boldsymbol{\Delta}_k^H\boldsymbol{\Delta}_k\right\}$ is a diagonal matrix, and (34) reduces to

$$E\left\{\|\mathbf{d}_k\|^2\right\} = \sigma_x^2 N_x\left\|\mathbf{h}_k\right\|^2\,. \tag{36}$$

Substituting (36) into (33), we obtain

$$\epsilon_1 = \frac{\sigma_\xi^2 N_h\left(2K+1\right)}{\sigma_x^2 N_x\left\|\mathbf{h}_k\right\|^2}\,. \tag{37}$$

We now evaluate ϵ_2 defined in (31), assuming that $x_{p,k}$ is variance-ergodic [33] and that N_x is sufficiently large. More specifically, we assume that $\frac{1}{N_x}\sum_{p=0}^{N_x-1}x_{p,k}x^*_{p+s,k'} \approx E\left\{x_{p,k}x^*_{p+s,k'}\right\}$. Hence, the (m,l)-th term of $\tilde{\boldsymbol{\Delta}}_k^H\tilde{\boldsymbol{\Delta}}_k$ can be approximated by

$$\begin{aligned}\left(\tilde{\boldsymbol{\Delta}}_k^H\tilde{\boldsymbol{\Delta}}_k\right)_{m,l} &= \sum_n x_{n-l\bmod N_h,\ k-K+\left\lfloor\frac{l}{N_h}\right\rfloor\ \bmod N}x^*_{n-m\bmod N_h,\ k-K+\left\lfloor\frac{m}{N_h}\right\rfloor\ \bmod N}\\ &\approx N_x E\left\{x_{n-l\bmod N_h,\ k-K+\left\lfloor\frac{l}{N_h}\right\rfloor\ \bmod N}x^*_{n-m\bmod N_h,\ k-K+\left\lfloor\frac{m}{N_h}\right\rfloor\ \bmod N}\right\} \end{aligned} \tag{38}$$

which reduces to (see Appendix II)

$$\left(\tilde{\boldsymbol{\Delta}}_k^H\tilde{\boldsymbol{\Delta}}_k\right)_{m,l} \approx N_x\sigma_x^2\delta(l-m)\,. \tag{39}$$

Substituting (39), (36) and the definition of $\mathbf{d}_k$ from (19) into (31), we obtain

$$\epsilon_2 = \frac{1}{\sigma_x^4 N_x^2 \left\|\mathbf{h}_k\right\|^2} \mathbf{h}_k^H \mathbf{\Omega}_k \mathbf{h}_k \tag{40}$$

where $\mathbf{\Omega}_k \triangleq E\left\{\mathbf{\Delta}_k^H \tilde{\mathbf{\Delta}}_k \tilde{\mathbf{\Delta}}_k^H \mathbf{\Delta}_k\right\}$. Using the fourth-order moment factoring theorem for zero-mean complex Gaussian samples [34], $\mathbf{\Omega}_k$ can be expressed as (see Appendix III)

$$\mathbf{\Omega}_k = \sigma_x^4 N_x \left[N_h \left(2K+1\right) \mathbf{I}_{N\cdot N_h \times N\cdot N_h} + N_x \tilde{\mathbf{I}}_{N\cdot N_h \times N\cdot N_h}\right] , \tag{41}$$

where $\tilde{\mathbf{I}}_{N\cdot N_h \times N\cdot N_h}$ is a diagonal matrix whose (m,m)-th term satisfies

$$\left(\tilde{\mathbf{I}}_{N\cdot N_h \times N\cdot N_h}\right)_{m,m} = \begin{cases} 1, & m \in \mathcal{L}_k(K) \\ 0, & otherwise \end{cases} \tag{42}$$

where $\mathcal{L}_k(K) = \left\{\left[(k-K+n_1) \bmod N\right] N_h + n_2 \mid \; n_1 \in \{0,\ldots,2K\}, n_2 \in \{0,\ldots,N_h-1\}\right\}$. Substituting (41) into (40), we obtain

$$\epsilon_2 = \frac{N_h\left(2K+1\right)}{N_x} + \frac{\sum_{m=0}^{2K} \left\|\mathbf{h}_{k,(k-K+m) \bmod N}\right\|^2}{\left\|\mathbf{h}_k\right\|^2} . \tag{43}$$

Finally, substituting (37) and (43) into (29), we have an explicit expression for $\epsilon_k(K)$:

$$\epsilon_k(K) = 1 + \frac{N_h\left(2K+1\right)}{N_x}\left[\frac{\sigma_\xi^2}{\sigma_x^2\left\|\mathbf{h}_k\right\|^2} - 1\right] - \frac{\sum_{m=0}^{2K} \left\|\mathbf{h}_{k,(k-K+m) \bmod N}\right\|^2}{\left\|\mathbf{h}_k\right\|^2} . \tag{44}$$

Expression (44) represents the MMSE obtained in the k-th band using LS estimates of $2K$ cross-band filters. It is worth noting that $\epsilon_k(K)$ depends, through $\mathbf{h}_k$, on the time impulse response $h(n)$ and on the analysis and synthesis parameters, *e.g.*, N, L and window type (see (8)). However, in this paper, we address only with the influence of K on the value of $\epsilon_k(K)$.

V. Relations between MMSE and SNR

In this section, we explore the relations between the input SNR and the MMSE performance. The MMSE performance is also dependent on the length of the input signal, but we first consider a fixed N_x, and subsequently discuss the influence of N_x on the MMSE performance.

Denoting the SNR by $\eta = \sigma_x^2/\sigma_\xi^2$, (44) can be rewritten as

$$\epsilon_k(K) = \frac{\alpha_k(K)}{\eta} + \beta_k(K) , \tag{45}$$

where

$$\alpha_k(K) \;\triangleq\; \frac{N_h}{N_x\left\|\mathbf{h}_k\right\|^2}\left(2K+1\right) , \tag{46}$$

$$\beta_k(K) \;\triangleq\; 1 - \frac{N_h\left(2K+1\right)}{N_x} - \frac{1}{\left\|\mathbf{h}_k\right\|^2}\sum_{m=0}^{2K}\left\|\mathbf{h}_{k,(k-K+m) \bmod N}\right\|^2 . \tag{47}$$

From (45), the MMSE $\epsilon_k(K)$ for fixed k and K values, is a monotonically decreasing function of η, which expectedly indicates that higher SNR values enable a better estimation of the relevant cross-band filters. Moreover, it is easy to verify from (46) and (47) that $\alpha_k(K+1) > \alpha_k(K)$ and $\beta_k(K+1) \leq \beta_k(K)$. Consequently $\epsilon_k(K)$ and $\epsilon_k(K+1)$ are two monotonically decreasing functions of η that satisfy

$$\begin{aligned} &\epsilon_k(K+1) > \epsilon_k(K), \quad \text{for} \quad \eta \to 0 \ \text{(low SNR)}, \\ &\epsilon_k(K+1) \leq \epsilon_k(K), \quad \text{for} \quad \eta \to \infty \ \text{(high SNR)}. \end{aligned} \tag{48}$$

Accordingly, these functions must intersect at a certain SNR value $\eta_k\left(K+1 \to K\right)$, that is, $\epsilon_k(K+1) \leq \epsilon_k(K)$ for $\eta \geq \eta_k\left(K+1 \to K\right)$, and $\epsilon_k(K+1) > \epsilon_k(K)$ otherwise (see typical MSE curves in Fig. 6). For SNR values higher than $\eta_k\left(K+1 \to K\right)$, a lower MSE value can be achieved by estimating $2(K+1)$ cross-band filters rather than only $2K$ filters. Increasing the number of cross-band filters is related to increasing the complexity of the system model [35], as will be explained in more details at the end of this section.

The SNR-intersection point $\eta_k\left(K+1 \to K\right)$ is obtained from (45) by requiring that $\epsilon_k(K+1) = \epsilon_k(K)$

$$\eta_k\left(K+1 \to K\right) = \frac{\alpha_k(K+1) - \alpha_k(K)}{\beta_k(K) - \beta_k(K+1)}\,. \tag{49}$$

Substituting (46) and (47) into (49), we have

$$\eta_k\left(K+1 \to K\right) = \frac{2N_h}{2N_h\left\|\mathbf{h}_k\right\|^2 + N_x\left(\left\|\mathbf{h}_{k,(k-K-1)\bmod N}\right\|^2 + \left\|\mathbf{h}_{k,(k+K+1)\bmod N}\right\|^2\right)}\,. \tag{50}$$

Since the cross-band filter's energy $\left\|\mathbf{h}_{k,k'}\right\|^2$ decreases as $|k-k'|$ increases (see Section II), we have

$$\eta_k\left(K \to K-1\right) \leq \eta_k\left(K+1 \to K\right)\,. \tag{51}$$

Specifically, the number of cross-band filters, which should be used for the system identifier, is a monotonically increasing function of the SNR. Estimating just the band-to-band filter and ignoring all the cross-band filters yields the minimal MSE only when the SNR is lower than $\eta_k\left(1 \to 0\right)$.

Another interesting point that can be concluded from (50) is that $\eta_k\left(K+1 \to K\right)$ is inversely proportional to N_x, the length of $x_{p,k}$ in frequency-band k. Therefore, for a fixed SNR value, the number of cross-band filters, which should be estimated in order to achieve the minimal MSE, increases as we increase N_x. For instance, suppose that N_x is chosen such that the input SNR satisfies $\eta_k\left(K \to K-1\right) \leq \eta \leq \eta_k\left(K+1 \to K\right)$, so that $2K$ cross-band filters should be estimated. Now, suppose that we increase the value of N_x, so that the same SNR now satisfies $\eta_k\left(K+1 \to K\right) \leq \eta \leq \eta_k\left(K+2 \to K+1\right)$. In this case, although the SNR remains the same, we would now prefer to estimate $2(K+1)$ cross-band

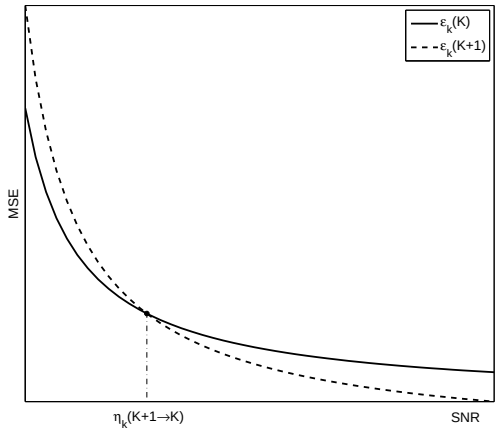


Fig. 6: Illustration of typical MSE curves as a function of the input SNR showing the relation between $\epsilon_k(K)$ (solid) and $\epsilon_k(K+1)$ (dashed).

filters rather than $2K$. It is worth noting that N_x is related to the update rate of $\hat{h}_{p,k,k'}$. We assume that during N_x frames the system impulse response does not change, and its estimate is updated every N_x frames. Therefore, a small N_x should be chosen whenever the system impulse response is time varying and fast tracking is desirable. However, in case the time variations in the system are slow, we can increase N_x, and correspondingly increase the number of cross-band filters.

It is worthwhile noting that the number of cross-band filters determines the complexity of system model. As the model complexity increases, the empirical fit to the data improves (*i.e.*, $\left\|\mathbf{d}_k - \hat{\mathbf{d}}_k\right\|^2$ can be smaller), but the variance of parametric estimates increases too (*i.e.*, variance of $\hat{\mathbf{d}}$), thus possibly worsening the accuracy of the model on new measurements [35]–[37], and increasing the MSE, $\epsilon_k(K)$. Hence, the appropriate model complexity is affected by the level of noise in the data and the length of observable data that can be employed for the system identification. As the SNR increases or as more data is employable, additional cross-band filters can be estimated and lower MMSE can be achieved.

VI. Computational Complexity

In this section, we address the computational complexity of the proposed approach and compare it to the conventional fullband approach and to the commonly-used subband approach that relies on the multiplicative transfer function (MTF) approximation. The computational complexity is computed by

counting the number of arithmetic operations[3] needed for the estimation process in each method.

A. Proposed subband approach

The computation of the proposed subband approach requires the solution of the LS normal equations (see (26))

$$\left(\tilde{\mathbf{\Delta}}_k^H \tilde{\mathbf{\Delta}}_k\right) \hat{\tilde{\mathbf{h}}}_k = \tilde{\mathbf{\Delta}}_k^H \mathbf{y}_k \tag{52}$$

for each frequency-band. Assuming that $\tilde{\mathbf{\Delta}}_k^H \tilde{\mathbf{\Delta}}_k$ is nonsingular, we may solve the normal equations in (52) using the Cholesky decomposition [38]. The number of arithmetic operations involved in forming the normal equations and solving them using the Cholesky decomposition is $N_y \left[(2K+1) N_h\right]^2 + \left[(2K+1) N_h\right]^3 /3$ [38]. As the system is identified, the desired signal estimate is computed by using (24), which requires $2N_y N_h (2K+1)$ arithmetic operations. In addition to the above computations, we need to consider the complexity of implementing the STFT. Each frame index in the STFT domain is computed by applying the discrete Fourier transform (DFT) on a short-time section of the input signal multiplied by a length N analysis window. This can be efficiently done by using fast Fourier transform (FFT) algorithms [39], which involve $5N \log_2 N$ arithmetic operations. Consequently, each STFT frame index requires $N + 5N \log_2 N$ arithmetic operations (the complexity of the ISTFT is approximately the same). Since the subband approach consists of two STFT (analysis filter bank) and one ISTFT (synthesis filter bank), the overall complexity of the STFT-ISTFT operations is $3N_y (N + 5N \log_2 N)$. Note that we also need to calculate the minimum energy analysis window by solving (5); however, since we compute it only once, we do not consider the computations required for its calculation. Therefore, the total number of computations required in the proposed approach is

$$\begin{aligned} & N \left\{ N_y \left[(2K+1) N_h\right]^2 + \left[(2K+1) N_h\right]^3 /3 + 2N_y (2K+1) N_h \right\} \\ & +3N_y (N + 5N \log_2 N) \quad \text{arithmetic operations .} \end{aligned} \tag{53}$$

Assuming that N_y is sufficiently large (more specifically, $N_y > (2K+1) N_h/3$) and that the computations required for the STFT-ISTFT calculation can be neglected, the computational complexity of the subband approach with $2K$ cross-band filters in each frequency-band can be expressed as

[3]An arithmetic operation is considered to be any complex multiplication, complex addition, complex subtraction, or complex division.

$$O_{SB}^{K}(N_h, N_y) = O\left(NN_y\left[(2K+1)\,N_h\right]^2\right). \tag{54}$$

B. Fullband approach

In the fullband approach, we consider the following LS optimization problem:

$$\hat{\mathbf{h}} = \arg\min_{\mathbf{h}} \|\mathbf{y} - \mathbf{X}\mathbf{h}\|^2, \tag{55}$$

where $\mathbf{X}$ is the $M \times Q$ Toeplitz matrix constructed from the input data $x(n)$, M is the observable data length, $\mathbf{y}$ is the $M \times 1$ system output vector constructed from $y(n)$ and $\hat{\mathbf{h}}$ is the $Q \times 1$ system estimate vector. In this case, the LS normal equations take the form of

$$\left(\mathbf{X}^H\mathbf{X}\right)\hat{\mathbf{h}} = \mathbf{X}^H\mathbf{y}. \tag{56}$$

As in the subband approach, forming the normal equations, solving them using the Cholesky decomposition and calculating the desired signal estimate, require $MQ^2 + Q^3/3 + 2MQ$ arithmetic operations. For sufficiently large M (*i.e.*, $M > Q/3$), the computational complexity of the fullband approach can be expressed as

$$O_{FB}(Q, M) = O\left(MQ^2\right). \tag{57}$$

A comparison of the fullband and subband complexities is given in subsection VI-D, by rewriting the subband complexity in terms of the fullband parameters (Q and M).

C. Multiplicative transfer function (MTF) approach

The MTF approximation is widely-used for the estimation of linear system in the STFT domain. Examples of such applications include frequency-domain blind source separation (BSS) [40], STFT-domain acoustic echo cancellation [23], relative transfer function (RTF) identification [3] and multichannel processing [8], [41]. Therefore, it is of great interest to compare the performance of the proposed approach to that of the MTF approach. In the above-mentioned applications, it is commonly assumed that the support of the STFT analysis window is sufficiently large compared with the duration of the system impulse response, so the system is approximated in the STFT domain with a single multiplication per frequency-band and no cross-band filters are utilized. Following this assumption, the STFT of the system output signal $y(n)$ is approximated by [42]

$$y_{p,k} \approx H_k x_{p,k} + \xi_{p,k}\,, \tag{58}$$

where $H_k \triangleq \sum_m h(m) \exp\left(-j2\pi mk/N\right)$. The single coefficient H_k is estimated using the following LS optimization problem:

$$\hat{H}_k = \arg\min_{H_k} \left\|\mathbf{y}_k - H_k \mathbf{x}_k\right\|^2\,, \tag{59}$$

where $\mathbf{y}_k$ was defined in (19) and $\mathbf{x}_k$ is the first column of $\mathbf{X}_k$ (defined in (17)). The solution of (59) is given by

$$\hat{H}_k = \frac{\mathbf{x}_k^H \mathbf{y}_k}{\left\|\mathbf{x}_k\right\|^2}\,. \tag{60}$$

In contrast with the fullband and the proposed approaches, the estimation of the desired signal in the MTF approach does not necessitate the inverse of a matrix. In fact, it requires only $N\left(5N_y+1\right) + 3N_y\left(N + 5N\log_2 N\right)$ arithmetic operations. Neglecting the STFT-ISTFT calculation (the second term), the computational complexity of the MTF approach can be expressed as

$$O_{MTF}(N_y) = O\left(NN_y\right)\,. \tag{61}$$

D. Comparison and Discussion

To make the comparison of the above three approaches tractable, we rewrite the complexities of the subband approaches in terms of the fullband parameters by using the relations $N_y \approx M/L$ and $N_h \approx Q/L$. Consequently, (54) and (61) can be rewritten as

$$O_{SB}^K(Q,M) = O\left(MQ^2 \frac{N\left(2K+1\right)^2}{L^3}\right) \tag{62}$$

and

$$O_{MTF}(M) = O\left(N\frac{M}{L}\right)\,. \tag{63}$$

A comparison of (57), (62) and (63) indicates that the complexity of the proposed subband approach is lower than that of the fullband approach by a factor of $L^3/\left[N\left(2K+1\right)^2\right]$ but higher than that of the MTF approach by a factor of $\left[Q\left(2K+1\right)/L\right]^2$. For instance, for $N=256$, $L=0.5N$, $Q=1500$ and $K=4$ the proposed approach complexity is reduced by a factor 100, when compared to the fullband

approach complexity and increased by a factor 10^4, when compared to the MTF approach complexity. However, the relatively high computational complexity of the fullband approach is compensated with a better MSE performance of the system identifier (see Section VII). On the other hand, the substantial low complexity of the MTF approach results in an insufficient accuracy of the system estimate, especially when the large window support assumption is not valid (*e.g.*, when long impulse response duration is considered). This point will be demonstrated in Section VII.

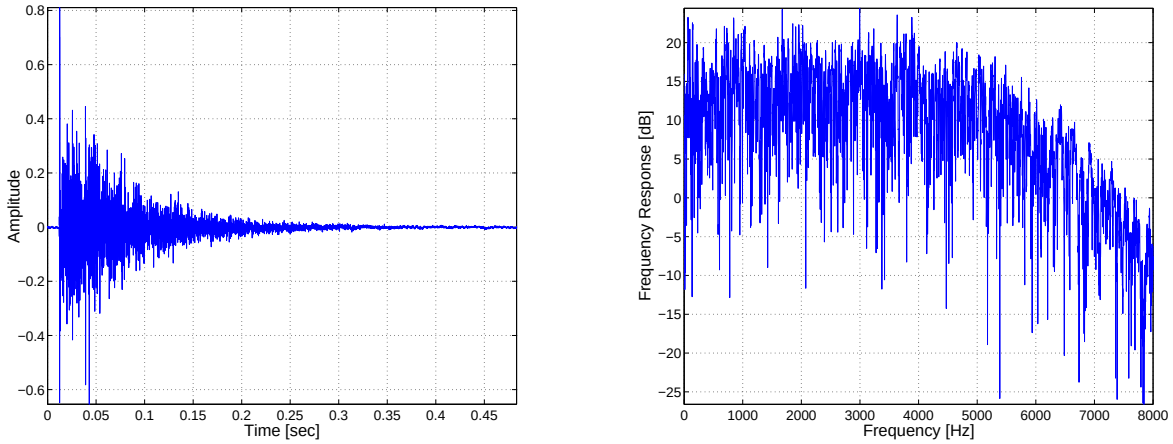
It can be seen from (62) that the computational complexity of the proposed approach increases as we increase the number of cross-band filters. However, as was shown in the previous section, this does not necessarily imply a lower steady-state MSE in subbands. Consequently, under appropriate conditions (*i.e.*, low SNR or fast time variations in the system), a lower MSE can be attained in each frequency-band with relatively few cross-band filters, resulting in low computational complexity. It is worth noting that the complexities of both the fullband and the proposed approaches may be reduced by exploiting the Toeplitz and block-Toeplitz structures of the corresponding matrices in the LS normal equations ($\mathbf{X}^H\mathbf{X}$ and $\tilde{\mathbf{\Delta}}_k^H\tilde{\mathbf{\Delta}}_k$, respectively) [38].

VII. Experimental Results

In this section, we present experimental results that verify the theoretical derivations obtained in sections IV and V. The signals employed for testing include synthetic white Gaussian signals as well as real speech signals. The performance of the proposed approach is evaluated for several SNR and N_x values and compared to that of the fullband approach and the MTF approach. Results are obtained by averaging over 200 independent runs.

We use the following parameters for all simulations presented in this section: Sampling rate of 16 kHz; A Hamming synthesis window of length $N = 256$ (16 ms) with 50% overlap ($L = 128$), and a corresponding minimum energy analysis window which satisfies the completeness condition (5) [27]. The impulse response $h(n)$ used in the experiments was measured in an office which exhibits a reverberation time of about 300 ms. Figure 7 shows the impulse and frequency responses of the measured system. The length of the impulse response was truncated to $Q = 1500$.

In the first experiment, we examine the system identifier performance in the STFT domain under the assumptions made in Section IV. That is, the STFT of the input signal $x_{p,k}$ is a zero-mean white Gaussian process with variance σ_x^2. Note that, $x_{p,k}$ is not necessarily a valid STFT signal, as not always a sequence whose STFT is given by $x_{p,k}$ may exist [43]. Similarly, the STFT of the noise signal $\xi_{p,k}$ is also a zero-mean white Gaussian process with variance σ_ξ^2, which is uncorrelated with $x_{p,k}$. Figure 8



(a) (b)

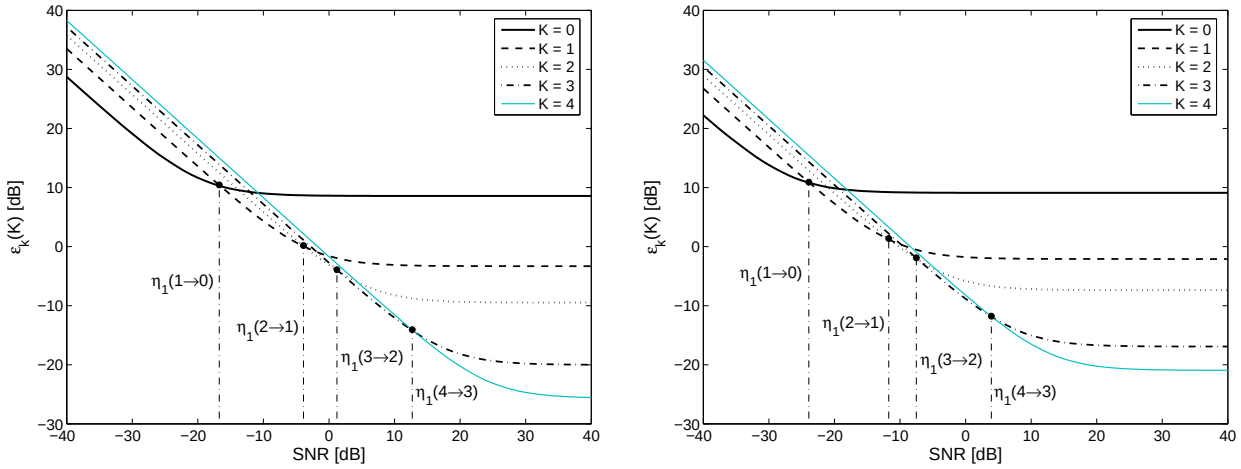
Fig. 7: (a) Measured impulse response and (b) its frequency response (sampling frequency=16kHz).

shows the MSE curves for the frequency-band $k = 1$ as a function of the input SNR for $N_x = 200$ and $N_x = 1000$ (similar results are obtained for the other frequency-bands). The results confirm that as the SNR increases, the number of cross-band filters that should be estimated to achieve a minimal MSE increases. We observe, as expected from (51), that the intersection-points of the MSE curves are a monotonically increasing series. Furthermore, a comparison of Figs. 8(a) and (b) indicates that the intersection-points values decrease as we increase N_x, as expected from (50). This verifies that when the signal length increases (while the SNR remains constant), more cross-band filters need to be used in order to attain the MMSE.

In the second experiment, we demonstrate the proposed theory on subband acoustic echo cancellation application (see Fig. 1). The far-end signal $x(n)$ is a speech signal and the local disturbance $\xi(n)$ consists of a zero-mean white Gaussian local noise with variance σ_ξ^2. The echo canceller performance is evaluated in the absence of near-end speech, since in such case a double-talk detector (DTD) is often applied in order to freeze the system adaptation process. Commonly used measure for evaluating the performance of conventional AECs is the echo-return loss enhancement (ERLE), defined in dB by

$$\mathrm{ERLE}(K) = 10\log\frac{E\left\{d^2(n)\right\}}{E\left\{\left(d(n)-\hat{d}_K(n)\right)^2\right\}}, \tag{64}$$

where $\hat{d}_K(n)$ is the inverse STFT of the estimated echo signal using $2K$ cross-band filters around each frequency-band. The ERLE performance of a conventional fullband AEC, where the echo signal

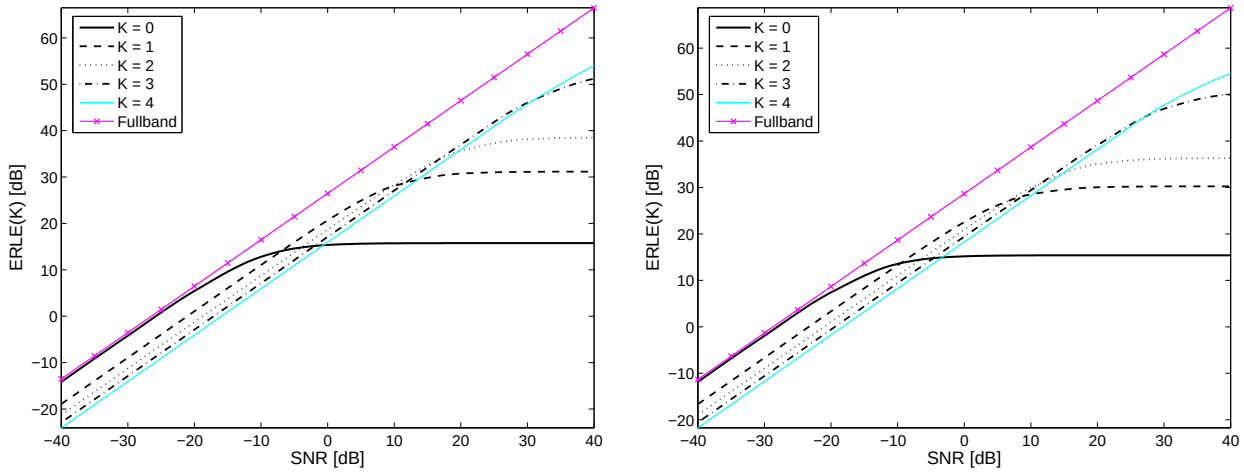


(a) (b)

Fig. 8: MSE curves as a function of the input SNR for white Gaussian signals. (a) $N_x = 200$. (b) $N_x = 1000$.

is estimated by (55), is also evaluated. Figure 9 shows the ERLE curves of both the fullband and the proposed approaches as a function of the input SNR obtained for a far-end signal of length 1.5 sec (Fig. 9(a)) and for a longer signal of length 2.56 sec (Fig. 9(b)). Clearly, as the SNR increases, the performance of the proposed algorithm can be generally improved (higher ERLE value can be obtained) by using a larger number of cross-band filters. Figure 9(a) shows that when the SNR is lower than -7 dB, estimating just the band-to-band filter ($K = 0$) and ignoring all the cross-band filters yields the maximal ERLE. Incorporating into the proposed AEC two cross-band filters ($K = 1$) decreases the ERLE by approximately 5 dB. However, when considering SNR values higher than -7 dB, the inclusion of two cross-band filters ($K = 1$) is preferable. It enables an increase of $10-20$ dB in the ERLE relative to that achieved by using only the band-to-band filter. Similar results are obtained for a longer signal (Fig. 9(b)), with the only difference that the intersection-points of the subband ERLE curves move towards lower SNR values. A comparison of the proposed subband approach with the fullband approach indicates that higher ERLE values can be obtained by using the latter, but at the expense of substantial increase in computational complexity. The advantage of the fullband approach in terms of ERLE performance stems from the fact that ERLE criterion is defined in the time domain and fullband estimation is also performed in the time domain.

In the third experiment, we compare the proposed approach to the MTF approach and investigate the influence of the STFT analysis window length (N) on their performances. We use a 1.5 sec length input speech signal and a white additive noise, as described in the previous experiment. A truncated impulse



(a) (b)

Fig. 9: ERLE curves for the proposed subband approach and the conventional fullband approach as a function of the input SNR for a real speech input signal. (a) Signal length is 1.5 sec ($N_x = 190$); (b) Signal length is 2.56 sec ($N_x = 322$).

response with 256 taps (16 ms) is used. Figure 10 shows the ERLE curves of both the MTF and the proposed approaches as a function of the input SNR obtained for an analysis window of length $N = 256$ (16 ms, Fig. 10(a)) and for a longer window of length $N = 2048$ (128 ms, Fig. 10(b)). In both cases we have $L = 0.5N$. As expected, the performance of the MTF approach can be generally improved by using a longer analysis window. This is because the MTF approach heavily relies on the assumption that the support of the analysis window is sufficiently large compared with the duration of the system impulse response. As the SNR increases, using the proposed approach yields the maximal ERLE, even for long analysis window. For instance, Fig. 10(b) shows that for 20 dB SNR the MTF algorithm achieves an ERLE value of 20 dB, whereas the inclusion of two cross-band filters ($K = 1$) in the proposed approach increases the ERLE by approximately 10 dB. Furthermore, it seems to be preferable to reduce the window length, as seen from Fig. 10(a), as it enables an increase of approximately 7 dB in the ERLE (for a 20 dB SNR) by using the proposed method. A short window is also essential for the analysis of nonstationary input signal, which is the case in acoustic echo cancellation application. However, a short window support necessitate the estimation of more cross-band filters for performance improvement, and correspondingly increases the computational complexity.

Another interesting point that can be concluded from Fig. 10 is that for low SNR values, a higher ERLE can be achieved by using the MTF approach, even when the large support assumption is not valid

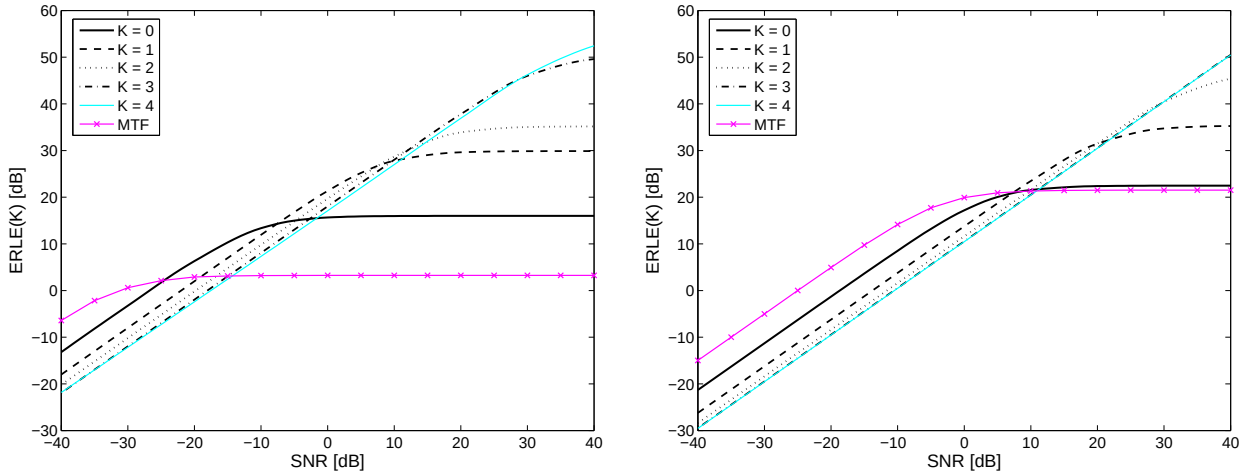


Fig. 10: ERLE curves for the proposed subband approach and the commonly-used multiplicative transfer function (MTF) approach as a function of the input SNR for a real speech input signal and an impulse response 16 ms length. (a) Length of analysis window is 16 ms ($N = 256$); (b) Length of analysis window is 128 ms ($N = 2048$).

(Fig. 10(a)).

VIII. Conclusions

We have derived explicit relations between the attainable MMSE in subbands and the power and length of the input signal for a system identifier implemented in the STFT domain. We showed that the MMSE is achieved by using a variable number of cross-band filters, determined by the power ratio between the input signal and the additive noise signal, and by the effective length of input signal that can be used for the system identification. Generally the number of cross-band filters that should be utilized in the system identifier is larger for stronger and longer input signals. Accordingly, during fast time variations in the system, shorter segments of the input signal can be employed, and consequently less cross-band filters are useful. However, when the time variations in the system become slower, additional cross-band filters can be incorporated into the system identifier and lower MSE is attainable. Furthermore, each subband may be characterized by a different power ratio between the input signal and the additive noise signal. Hence, a different number of cross-band filters may be employed in each subband.

The strategy of controlling the number of cross-band filters is related to and can be combined with step-size control implemented in adaptive echo cancellation algorithms, *e.g.*, [44], [45]. Step-size control is designed for faster tracking during abrupt variations in the system, while not compromising for higher

MSE when the system is time invariant. Therefore, joint control of step-size and the number of cross-band filters may further enhance the performance of adaptive echo cancellation algorithms.

APPENDIX I
DERIVATION OF (7)

Using (1) and (6), the STFT of $d(n)$ can be written as

$$d_{p,k} = \sum_{m,l} h(l)x(m-l)\tilde{\psi}^*_{p,k}(m) \tag{65}$$

Substituting (3) into (65), we obtain

$$\begin{aligned} d_{p,k} &= \sum_{m,l} h(l) \sum_{k'=0}^{N-1} \sum_{p'} x_{p',k'}\psi_{p',k'}(m-l)\tilde{\psi}^*_{p,k}(m) \\ &= \sum_{k'=0}^{N-1} \sum_{p'} x_{p',k'} h_{p,k,p',k'} \end{aligned} \tag{66}$$

where

$$h_{p,k,p',k'} = \sum_{m,l} h(l)\psi_{p',k'}(m-l)\tilde{\psi}^*_{p,k}(m) \tag{67}$$

may be interpreted as the STFT of $h(n)$ using a composite analysis window $\sum_m \psi_{p',k'}(m-l)\tilde{\psi}^*_{p,k}(m)$. Substituting (2) and (4) into (67), we obtain

$$\begin{aligned} h_{p,k,p',k'} &= \sum_{m,l} h(l)\psi(m-l-p'L)e^{j\frac{2\pi}{N}k'(m-l-p'L)}\tilde{\psi}(m-pL)e^{-j\frac{2\pi}{N}k(m-pL)} \\ &= \sum_{l} h(l) \sum_{m} \tilde{\psi}(m)e^{-j\frac{2\pi}{N}km}\psi\left((p-p')L-l+m\right)e^{j\frac{2\pi}{N}k'((p-p')L-l+m)} \\ &= \{h(n) * \phi_{k,k'}(n)\}\,|_{n=(p-p')L} \triangleq h_{p-p',k,k'}\,, \end{aligned} \tag{68}$$

where $*$ denotes convolution with respect to the time index n, and

$$\phi_{k,k'}(n) \triangleq e^{j\frac{2\pi}{N}k'n} \sum_{m} \tilde{\psi}(m)\psi(n+m)e^{-j\frac{2\pi}{N}m(k-k')}\,. \tag{69}$$

From (68), $h_{p,k,p',k'}$ depends on $(p-p')$ rather than on p and p' separately. Substituting (68) into (66), we obtain (7)-(9).

APPENDIX II
DERIVATION OF (39)

Using the whiteness property of $x_{p,k}$, the (m,l)-th term of $\tilde{\mathbf{\Delta}}_k^H \tilde{\mathbf{\Delta}}_k$ given in (38) can be derived as

$$\begin{aligned}\left(\tilde{\mathbf{\Delta}}_k^H \tilde{\mathbf{\Delta}}_k\right)_{m,l} &\approx N_x E\left\{x_{n-l \bmod N_h,\ k-K+\left\lfloor \frac{l}{N_h}\right\rfloor \bmod N} x^*_{n-m \bmod N_h,\ k-K+\left\lfloor \frac{m}{N_h}\right\rfloor \bmod N}\right\} \\ &= N_x \sigma_x^2 \delta\left(l \bmod N_h - m \bmod N_h\right) \\ &\quad \times \delta\left(\left(k-K+\left\lfloor \frac{l}{N_h}\right\rfloor\right) \bmod N - \left(k-K+\left\lfloor \frac{m}{N_h}\right\rfloor\right) \bmod N\right). \end{aligned} \tag{70}$$

Therefore, $\left(\tilde{\mathbf{\Delta}}_k^H \tilde{\mathbf{\Delta}}_k\right)_{m,l}$ is nonzero only if $l \bmod N_h = m \bmod N_h$ and $\left(k-K+\left\lfloor \frac{l}{N_h}\right\rfloor\right) \bmod N = \left(k-K+\left\lfloor \frac{m}{N_h}\right\rfloor\right) \bmod N$. Those conditions can be rewritten as

$$l = m + rN_h \qquad \text{for } r = 0, \pm 1, \pm 2, \ldots \tag{71}$$

and

$$k-K+\left\lfloor \frac{l}{N_h}\right\rfloor = k-K+\left\lfloor \frac{m}{N_h}\right\rfloor + qN \quad \text{for } q = 0, \pm 1, \pm 2, \ldots . \tag{72}$$

Substituting (71) into (72), we obtain

$$r = qN \ \ ; \quad q = 0, \pm 1, \pm 2, \ldots . \tag{73}$$

However, recall that $0 \le l, m \le (2K+1)N_h - 1 \le NN_h - 1$, then it is easy to verify from (71) that

$$\max\{|r|\} = N - 1 . \tag{74}$$

From (73) and (74) we conclude that $r = 0$, so (71) reduces to $m = l$ and we obtain (39).

APPENDIX III
DERIVATION OF (41)

The (m,l)-th term of $\mathbf{\Omega}_k$ from (40) can be written as

$$\begin{aligned}(\mathbf{\Omega}_k)_{m,l} &= \sum_{n,r,q} E\left\{x_{r-n \bmod N_h,\ k-K+\left\lfloor \frac{n}{N_h}\right\rfloor \bmod N} x^*_{r-m \bmod N_h, \left\lfloor \frac{m}{N_h}\right\rfloor} \right. \\ &\quad \left. \times\ x_{q-l \bmod N_h, \left\lfloor \frac{l}{N_h}\right\rfloor} x^*_{q-n \bmod N_h,\ k-K+\left\lfloor \frac{n}{N_h}\right\rfloor \bmod N}\right\} . \end{aligned} \tag{75}$$

By using the fourth-order moment factoring theorem for zero-mean complex Gaussian samples [34], (75) can be rewritten as

$$\begin{aligned}(\mathbf{\Omega}_k)_{m,l} &= \sum_{n,r,q} E\left\{x_{r-n \bmod N_h,\ k-K+\left\lfloor \frac{n}{N_h}\right\rfloor \ \bmod N} x^*_{q-n \bmod N_h,\ k-K+\left\lfloor \frac{n}{N_h}\right\rfloor \ \bmod N}\right\} \\ &\quad \times E\left\{x^*_{r-m \bmod N_h,\left\lfloor \frac{m}{N_h}\right\rfloor} x_{q-l \bmod N_h,\left\lfloor \frac{l}{N_h}\right\rfloor}\right\} \\ &\quad + \sum_{n,r,q} E\left\{x_{r-n \bmod N_h,\ k-K+\left\lfloor \frac{n}{N_h}\right\rfloor \ \bmod N} x^*_{r-m \bmod N_h,\left\lfloor \frac{m}{N_h}\right\rfloor}\right\} \\ &\quad \times E\left\{x_{q-l \bmod N_h,\left\lfloor \frac{l}{N_h}\right\rfloor} x^*_{q-n \bmod N_h,\ k-K+\left\lfloor \frac{n}{N_h}\right\rfloor \ \bmod N}\right\}. \end{aligned} \tag{76}$$

Using the whiteness property of $x_{p,k}$, we can write (76) as

$$(\mathbf{\Omega}_k)_{m,l} = \omega_1 + \omega_2\,, \tag{77}$$

where

$$\omega_1 = \sigma_x^4 \sum_{n,r,q} \delta(r-q)\,\delta(r-q+l \bmod N_h - m \bmod N_h)\,\delta\left(\left\lfloor \frac{m}{N_h}\right\rfloor - \left\lfloor \frac{l}{N_h}\right\rfloor\right) \tag{78}$$

and

$$\begin{aligned}\omega_2 &= \sigma_x^4 \sum_{n,r,q} \delta(n \bmod N_h - m \bmod N_h)\,\delta\left(\left(k-K+\left\lfloor \frac{n}{N_h}\right\rfloor\right) \bmod N - \left\lfloor \frac{m}{N_h}\right\rfloor\right) \\ &\quad \times \delta(n \bmod N_h - l \bmod N_h)\,\delta\left(\left(k-K+\left\lfloor \frac{n}{N_h}\right\rfloor\right) \bmod N - \left\lfloor \frac{l}{N_h}\right\rfloor\right). \end{aligned} \tag{79}$$

Recall that n ranges from 0 to $(2K+1)N_h - 1$, and that r and q range from 0 to $N_y - 1$ (although for fixed m, l and n values only N_x values of r and q contribute), (78) reduces to

$$\omega_1 = \sigma_x^4 N_x (2K+1) N_h \delta(m-l)\,. \tag{80}$$

We now proceed with expanding ω_2. It is easy to verify from (79) that m and l satisfy $m \bmod N_h = l \bmod N_h$ and $\left\lfloor \frac{m}{N_h}\right\rfloor = \left\lfloor \frac{l}{N_h}\right\rfloor$, therefore $m = l$. In addition, n satisfies both

$$n \bmod N_h = m \bmod N_h \tag{81}$$

and

$$\left(k-K+\left\lfloor \frac{n}{N_h}\right\rfloor\right) \bmod N = \left\lfloor \frac{m}{N_h}\right\rfloor\,, \tag{82}$$

where (82) can be rewritten as

$$k-K+\left\lfloor \frac{n}{N_h}\right\rfloor = \left\lfloor \frac{m}{N_h}\right\rfloor + hN, \quad \text{for } h = 0, \pm 1, \pm 2, \ldots\,. \tag{83}$$

Writing n as $n = \left\lfloor \frac{n}{N_h} \right\rfloor N_h + n \bmod N_h$, we obtain

$$n = m - (k - K - hN) N_h, \quad \text{for } h = 0, \pm 1, \pm 2, \dots . \tag{84}$$

From (84), one value of n, at the most, contributes to ω_2 for a fixed value of m. Therefore, we can bound the range of m, such that values outside this range will not contribute to ω_2. Since $n \in \{0, 1, ..., (2K+1)N_h - 1\}$, we can use (84) to obtain

$$\begin{aligned} m &\in \{ (k - K - hN) N_h + n |\ n \in \{0, 1, ..., (2K+1)N_h - 1\}, h = 0, \pm 1, \pm 2, \dots \} \\ &= \{ (k - K + n_1 - hN) N_h + n_2 |\ n_1 \in \{0, 1, ..., 2K\}, \\ &\qquad n_2 \in \{0, 1, ..., N_h - 1\}, h = 0, \pm 1, \pm 2, \dots \} . \end{aligned} \tag{85}$$

Now, since the size of $\mathbf{\Omega}_k$ is $N_h N \times N_h N$, m should also range from 0 to $N N_h - 1$ and therefore, (85) reduces to

$$m \in \{ [(k - K + n_1) \bmod N] N_h + n_2 |\ n_1 \in \{0, 1, ..., 2K\}, n_2 \in \{0, 1, ..., N_h - 1\} \} . \tag{86}$$

Finally, since ω_2 is independent of both r and q, it can be written as

$$\omega_2 = \sigma_x^4 N_x^2 \delta (m - l) \delta (m \in \mathcal{L}_k(K)) \tag{87}$$

where $\mathcal{L}_k(K) = \{ [(k - K + n_1) \bmod N] N_h + n_2 |\ n_1 \in \{0, 1, ..., 2K\}, n_2 \in \{0, 1, ..., N_h - 1\} \}$. Substituting (80) and (87) into (77), and writing the result in a vector form yields (41).

ACKNOWLEDGEMENT

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

REFERENCES

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

[2] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A Prctical Approach*. New Jersey: John Wiley & Sons, Inc., 2004.

[3] I. Cohen, "Relative transfer function identification using speech signals," *Special Issue of the IEEE Trans. Speech and Audio Processing on Multi-channel Signal Processing for Audio and Acoustics Applications*, vol. 12, no. 5, pp. 451–459, September 2004.

[4] Y. Huang, J. Benesty, and J. Chen, "A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 5, pp. 882–895, September 2005.

[5] M. Wu and D. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 774–784, May 2006.

[6] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 774–784, May 2006.

[7] F. Talantzis, D. B. Ward, and P. A. Naylor, "Performance analysis of dynamic acoustic source separation in reverberant rooms," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1378–1390, July 2006.

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

[9] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *IEEE Trans. Speech and Audio Processing*, vol. 12, no. 6, pp. 561–571, November 2004.

[10] S. Haykin, *Adaptive Filter Theory*, 4th ed. New Jersey: Prentice-Hall, 2002.

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

[12] H. Yasukawa, S. Shimada, and I. Furukawa, "Acoustic echo canceller with high speech quality," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. Dallas, Texas: IEEE, Apr. 1987, pp. 2125–2128.

[13] W. Kellermann, "Analysis and design of multirate systems for cancellation of acoustical echoes," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. New-York City, USA: IEEE, Apr. 1988, pp. 2570–2573.

[14] M. Harteneck, J. M. Páez-Borrallo, and R. W. Stewart, "An oversampled subband adaptive filter without cross adaptive filters," *Signal Processing*, vol. 64, no. 1, pp. 93–101, Mar. 1994.

[15] V. S. Somayazulu, S. K. Mitra, and J. J. Shynk, "Adaptive line enhancement using multirate techniques," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. Glasgow, Scotland: IEEE, May 1989, pp. 928–931.

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

[17] S. S. Pradhan and V. U. Reddy, "A new approach to subband adaptive filtering," *IEEE Transactions on Signal Processing*, vol. 47, no. 3, pp. 655–664, Mar. 1999.

[18] B. E. Usevitch and M. T. Orchard, "Adaptive filtering using filter banks," *IEEE Transactions on Circuits and Systems II*, vol. 43, no. 3, pp. 255–265, Mar. 1996.

[19] A. Gilloire and M. Vetterli, "Adaptive filtering in subbands," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. New-York City, USA: IEEE, Apr. 1988, pp. 1572–1575.

[20] 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.

[21] C. Avendano and G. Garcia, "STFT-based multi-channel acoustic interference suppressor," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. Salt-Lake City, Utah: IEEE, May 2001, pp. 625–628.

[22] Y. Lu and J. M. Morris, "Gabbor expansion for adaptive echo cancellation," *IEEE Signal Processing Mag.*, vol. 16, pp. 68–80, Mar. 1999.

[23] C. Faller and J. Chen, "Suppressing acoustic echo in a spectral envelope space," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 13, no. 5, pp. 1048–1062, Sep. 2005.

[24] Y. Avargel and I. Cohen, "Performance analysis of cross-band adaptation for subband acoustic echo cancellation," *submitted to Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC)*, Paris, France, Sep. 2006.

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

[26] S. Farkash and S. Raz, "Linear systems in Gabor time-frequency space," *IEEE Transactions on Signal Processing*, vol. 42, no. 3, pp. 611–617, Jan. 1998.

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

[28] S. Qian and D. Chen, "Discrete Gabor transform," *IEEE Transactions on Signal Processing*, vol. 41, no. 7, pp. 2429–2438, Jul. 1993.

[29] A. Neumaier, "Solving ill-conditioned and singular linear systems: A tutorial on regularization," *SIAM Rev.*, vol. 40, no. 3, pp. 636–666, Sep. 1998.

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

[31] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transaction on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[32] Y. Ephraim and I. Cohen, "Recent advancements in speech enhancement," in *The Electrical Engineering Handbook*, 3rd ed., R. C. Dorf, Ed. Boca Raton: CRC, 2006.

[33] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. Singapore: McGRAW-Hill, 1991.

[34] D. G. Manokis, 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.

[35] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[36] L. Ljung, *System Identification: Theory for the User*, 2nd ed. Upper Saddle River, New Jersey: Prentice-Hall, 1999.

[37] 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. and Measurement*, vol. 54, no. 1, pp. 144–150, February 2005.

[38] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 3rd ed. Baltimore, MD: The Johns Hopkins University Press, 1996.

[39] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

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

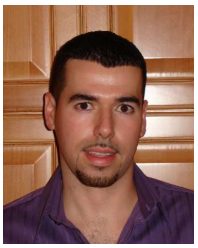
[41] I. Cohen, "Multichannel post-filtering in nonstationary noise environments," *IEEE Transactions on Signal Processing*, vol. 52, no. 5, pp. 1149–1160, May 2004.

[42] C. Avendano, "Temporal processing of speech in a time-feature space," Ph.D. dissertation, Oregon Graduate Institute of Science & Technology, April 1997.

[43] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. ASSP-32, no. 2, pp. 236–243, Apr. 1984.

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

[45] A. Mader, H. Puder, and G. U. Schmidt, "Step-size control for acoustic echo cancellation filters- an overview," *Signal Processing*, vol. 80, pp. 1697–1719, Sep. 2000.



Yekutiel Avargel received the B.Sc. degree in electrical engineering in 2004 from the Technion — Israel Institute of Technology, Haifa, Israel. He is currently pursuing the Ph.D. degree in electrical engineering at the Technion.

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) of the 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 in 1990, 1993 and 1998, respectively, all from the Technion – Israel Institute of Technology, Haifa, Israel.

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

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

LIST OF TABLES

LIST OF FIGURES