EURASIP Journal on Applied Signal Processing 2005:18, 2965–2978

A Block-Based Linear MMSE Noise Reduction with a High Temporal Resolution Modeling of the Speech Excitation

Chunjian Li
Department of Communication Technology, Aalborg University, 9220 Aalborg Ø, Denmark
Email: cl@kom.aau.dk

Søren Vang Andersen
Department of Communication Technology, Aalborg University, 9220 Aalborg Ø, Denmark
Email: sva@kom.aau.dk

Received 14 May 2004; Revised 11 March 2005

A comprehensive linear minimum mean squared error (LMMSE) approach for parametric speech enhancement is developed. The proposed algorithms aim at joint LMMSE estimation of signal power spectra and phase spectra, as well as exploitation of correlation between spectral components. The major cause of this interfrequency correlation is shown to be the prominent temporal power localization in the excitation of voiced speech. LMMSE estimators in time domain and frequency domain are first formulated. To obtain the joint estimator, we model the spectral signal covariance matrix as a full covariance matrix instead of a diagonal covariance matrix as is the case in the Wiener filter derived under the quasi-stationarity assumption. To accomplish this, we decompose the signal covariance matrix into a synthesis filter matrix and an excitation matrix. The synthesis filter matrix is built from estimates of the all-pole model coefficients, and the excitation matrix is built from estimates of the instantaneous power of the excitation sequence. A decision-directed power spectral subtraction method and a modified multipulse linear predictive coding (MPLPC) method are used in these estimations, respectively. The spectral domain formulation of the LMMSE estimator reveals important insight in interfrequency correlations. This is exploited to significantly reduce computational complexity of the estimator. For resource-limited applications such as hearing aids, the performance-to-complexity trade-off can be conveniently adjusted by tuning the number of spectral components to be included in the estimate of each component. Experiments show that the proposed algorithm is able to reduce more noise than a number of other approaches selected from the state of the art. The proposed algorithm improves the segmental SNR of the noisy signal by 13 dB for the white noise case with an input SNR of 0 dB.

Keywords and phrases: noise reduction, speech enhancement, LMMSE estimation, Wiener filtering.

1. INTRODUCTION

Noise reduction is becoming an important function in hearing aids in recent years thanks to the application of powerful DSP hardware and the progress of noise reduction algorithm design. Noise reduction algorithms with high performance-to-complexity ratio have been the subject of extensive research study for many years. Among many different approaches, two classes of single-channel speech enhancement methods have attracted significant attention in recent years because of their better performance compared to the classic spectral subtraction methods (a comprehensive study of spectral subtraction methods can be found in [1]). These two classes are the frequency domain block-based minimum mean squared error (MMSE) approach and the signal subspace approach. The frequency domain MMSE approach includes the noncausal IIR Wiener filter [2], the MMSE short-time spectral amplitude (MMSE-STSA) estimator [3], the MMSE log-spectral amplitude (MMSE-LSA) estimator [4], the constrained iterative Wiener filtering (CIWF) [5], and the MMSE estimator using non-Gaussian priors [6]. These MMSE algorithms all rely on an assumption of quasi-stationarity and an assumption of uncorrelated spectral components in the signal. The quasi-stationarity assumption requires short-time processing. At the same time, the assumption of uncorrelated spectral components can be warranted by assuming the signal to be infinitely long and wide-sense stationary [7, 8]. This infinite data length

assumption is in principle violated when using the short-time processing, although the effect of this violation may be minor (and is not the major issue this paper addresses). More importantly, the wide-sense stationarity assumption within a short frame does not well model the prominent temporal power localization in the excitation source of voiced speech due to the impulse train structure. This temporal power localization within a short frame can be modeled as a non-stationarity of the signal that is not resolved by the short-time processing. In [9], we show how voiced speech is advantageously modeled as nonstationary even within a short frame and that this model implies significant inter-frequency correlations. As a consequence of the stationarity and long frame assumptions, the MMSE approaches model the frequency domain signal covariance matrix as a diagonal matrix.

Another class of speech enhancement methods, the signal subspace approach, implicitly exploits part of the inter-frequency correlation by allowing the frequency domain signal covariance matrix to be nondiagonal. This class includes the time domain constraint (TDC) linear estimator and spectral domain constraint (SDC) linear estimator [10], and the truncated singular value decomposition (TSVD) estimator [11]. In [10], the TDC estimator is shown to be an LMMSE estimator with adjustable input noise level. When the TDC filtering matrix is transformed to the frequency domain, it is in general non-diagonal. Nevertheless, the known signal-subspace-based methods still assume stationarity within a short frame. This can be seen as follows. In TDC and SDC the noisy signal covariance matrices are estimated by time averaging of the outer product of the signal vector, which requires stationarity within the interval of averaging. The TSVD method applies singular value decomposition to the signal matrix instead. This can be shown to be equivalent to the eigen decomposition of the time-averaged outer product of signal vectors. Compared to the mentioned frequency domain MMSE approaches, the known signal subspace methods implicitly avoid the infinite data length assumption, so that the inter-frequency correlation caused by the finite-length effect is accommodated. However, the more important cause of inter-frequency correlation, that is, the non stationarity within a frame, is not modeled.

In terms of exploiting the masking property of the human auditory system, the above-mentioned frequency domain MMSE algorithms and signal-subspace-based algorithms can be seen as spectral masking methods without explicit modeling of masking thresholds. To see this, observe that the MMSE approaches shape the residual noise (the remaining background noise) power spectrum to one more similar to the speech power spectrum, thereby facilitating a certain degree of masking of the noise. In general, the MMSE approaches attenuate more in the spectral valleys than the spectral subtraction methods do. Perceptually, this is beneficial for high-pitch voiced speech, which has sparsely located spectral peaks that are not able to mask the spectral valley sufficiently. The signal subspace methods in [10] are designed to shape the residual noise power spectrum for a better spectral masking, where the masking threshold is found experimentally. Auditory masking techniques have received increasing attention in recent research of speech enhancement [12, 13, 14]. While the majority of these works focus on spectral domain masking, the work in [15] shows the importance of the temporal masking property in connection with the excitation source of voiced speech. It is shown that noise between the excitation impulses is more perceivable than noise close to the impulses, and this is especially so for the low-pitch speech for which the excitation impulses locate temporally sparsely. This temporal masking property is not employed by current frequency-domain MMSE estimators and the signal subspace approaches.

In this paper, we develop an LMMSE estimator with a high temporal resolution modeling of the excitation of voiced speech, aiming for modeling a certain non-stationarity of the speech within a short frame, which is not modeled by quasi-stationarity-based algorithms. The excitation of voiced speech exhibits prominent temporal power localization, which appears as an impulse train superimposed with a low-level noise floor. We model this temporal power localization as a non-stationarity. This non-stationarity causes significant inter-frequency correlation. Our LMMSE estimator therefore avoids the assumption of uncorrelated spectral components and is able to exploit the inter-frequency correlation. Both the frequency domain signal covariance matrix and filtering matrix are estimated as complex-valued full matrices, which means that the information about inter-frequency correlation are not lost and the amplitude and phase spectra are estimated jointly. Specifically, we make use of the linear-prediction-based source-filter model to estimate the signal covariance matrix, upon which a time domain or frequency domain LMMSE estimator is built. In the estimation of the signal covariance matrix, this matrix is decomposed into a synthesis filter matrix and an excitation matrix. The synthesis filter matrix is estimated by a smoothed power spectral subtraction method followed by an autocorrelation linear predictive coding (LPC) method. The excitation matrix is a diagonal matrix with the instantaneous power of the LPC residual as its diagonal elements. The instantaneous power of the LPC residual is estimated by a modified multipulse linear predictive coding (MPLPC) method. Having estimated the signal covariance matrix, we use it in a vector LMMSE estimator. We show that by doing the LMMSE estimation in the frequency domain instead of time domain, the computational complexity can be reduced significantly due to the fact that the signal is less correlated in the frequency domain than in the time domain. Compared to several quasi-stationarity-based estimators, the proposed LMMSE estimator results in a lower spectral distortion to the enhanced speech signal while having higher noise reduction capability. The algorithm applies more attenuation in the valleys between pitch impulses in time domain, while small attenuation is applied around the pitch impulses. This arrangement exploits the temporal masking effect and results in a better preservation of abrupt rise of the waveform amplitude while maintaining a large amount of noise reduction.

The rest of this paper is organized as follows. In Section 2 the notations and assumptions used in the derivation of

LMMSE estimators are outlined. In Section 3, the nonstationary modeling of the signal covariance matrices is described. The algorithm is summarized in Section 4. In Section 5, the computational complexity of the algorithm is reduced by identifying an interval of significant correlation and by simplifying the modified MPLPC procedure. Experimental settings, objective, and subjective results are given in Section 6. Finally, Section 7 discusses the obtained results.

2. BACKGROUND

In this section, notations and statistic assumptions for the derivation of LMMSE estimators in time and frequency domains are outlined.

2.1. Time domain LMMSE estimator

Let $y(n,k)$, $s(n,k)$, and $v(n,k)$ denote the nth sample of noisy observation, speech, and additive noise (uncorrelated with the speech signal) of the kth frame, respectively. Then

$$y(n,k) = s(n,k) + v(n,k). \tag{1}$$

Alternatively, in vector form we have

$$\mathbf{y} = \mathbf{s} + \mathbf{v}, \tag{2}$$

where boldface letters represent vectors and the frame indices are omitted to allow a compact notation. For example $\mathbf{y} = [y(1,k), y(2,k), \ldots, y(N,k)]^T$ is the noisy signal vector of the kth frame, where N is the number of samples per frame.

To obtain linear MMSE estimators, we assume zero-mean Gaussian PDFs for the noise and the speech processes. Under this statistic model the LMMSE estimate of the signal is the conditional mean [16]

$$\hat{\mathbf{s}} = E[\mathbf{s}|\mathbf{y}] = \mathbf{C}_\mathbf{s}\left(\mathbf{C}_\mathbf{s} + \mathbf{C}_\mathbf{v}\right)^{-1}\mathbf{y}, \tag{3}$$

where $\mathbf{C}_\mathbf{s}$ and $\mathbf{C}_\mathbf{v}$ are the covariance matrices of the signal and the noise, respectively. The covariance matrix is defined as $\mathbf{C}_\mathbf{s} = E[\mathbf{s}\mathbf{s}^H]$, where $(\cdot)^H$ denotes Hermitian transposition and $E[\cdot]$ denotes the ensemble average operator.

2.2. Frequency domain LMMSE estimator and Wiener filter

In the frequency domain the goal is to estimate the complex DFT coefficients given a set of DFT coefficients of the noisy observation. Let $Y(m,k)$, $\theta(m,k)$, and $V(m,k)$ denote the mth DFT coefficient of the kth frame of the noisy observation, the signal, and the noise, respectively. Due to the linearity of the DFT operator, we have

$$Y(m,k) = \theta(m,k) + V(m,k). \tag{4}$$

In vector form we have

$$\mathbf{Y} = \boldsymbol{\theta} + \mathbf{V}, \tag{5}$$

where again boldface letters represent vectors and the frame indices are omitted. As an example, the noisy spectrum vector of the kth frame is arranged as $\mathbf{Y} = [Y(1,k), Y(2,k), \ldots, Y(N,k)]^T$ where the number of frequency bins is equal to the number of samples per frame N.

We again use the linear model. $\mathbf{Y}$, $\boldsymbol{\theta}$, and $\mathbf{V}$ are assumed to be zero-mean complex Gaussian random variables and $\boldsymbol{\theta}$ and $\mathbf{V}$ are assumed to be uncorrelated to each other. The LMMSE estimate is the conditional mean

$$\hat{\boldsymbol{\theta}} = E[\boldsymbol{\theta}|\mathbf{Y}] = \mathbf{C}_{\boldsymbol{\theta}}\left(\mathbf{C}_{\boldsymbol{\theta}} + \mathbf{C}_\mathbf{V}\right)^{-1}\mathbf{Y}, \tag{6}$$

where $\mathbf{C}_{\boldsymbol{\theta}}$ and $\mathbf{C}_\mathbf{V}$ are the covariance matrices of the DFT coefficients of the signal and the noise, respectively. By applying inverse DFT to each side, (6) can be easily shown to be identical to (3).

The relation between the two signal covariance matrices in time and frequency domains is

$$\mathbf{C}_{\boldsymbol{\theta}} = \mathbf{F}\mathbf{C}_\mathbf{s}\mathbf{F}^{-1}, \tag{7}$$

where $\mathbf{F}$ is the Fourier matrix. If the frame was infinitely long and the signal was stationary, $\mathbf{C}_\mathbf{s}$ would be an infinitely large Toeplitz matrix. The infinite Fourier matrix is known to be the eigenvector matrix of any infinite Toeplitz matrix [8]. Thus, $\mathbf{C}_{\boldsymbol{\theta}}$ becomes diagonal and the LMMSE estimator (6) reduces to the noncausal IIR Wiener filter with the transfer function

$$H_{WF}(\omega) = \frac{P_{ss}(\omega)}{P_{ss}(\omega) + P_{vv}(\omega)}, \tag{8}$$

where $P_{ss}(\omega)$ and $P_{vv}(\omega)$ denote the power spectral density (PSD) of the signal and the noise, respectively. In the sequel we refer to (8) as the Wiener filter or WF.

3. HIGH TEMPORAL RESOLUTION MODELING FOR THE SIGNAL COVARIANCE MATRIX ESTIMATION

For both time and frequency domains LMMSE estimators described in Section 2, the estimation of the signal covariance matrix $\mathbf{C}_\mathbf{s}$ is crucial. In this work, we assume the noise to be stationary. For the signal, however, we propose the use of a high temporal resolution model to capture the non-stationarity caused by the excitation power variation. This can be explained by examining the voice production mechanism. In the well-known source-filter model for voiced speech, the excitation source models the glottal pulse train, and the filter models the resonance property of the vocal tract. The vocal tract can be viewed as a slowly varying part of the system. Typically in a duration of 20 ms to 30 ms it changes very little. The vocal folds vibrate at a faster rate producing periodic glottal flow pulses. Typically there can be 2 to 8 glottal pulses in 20 ms. In speech coding, it is common practice to model this pulse train by a long-term correlation pattern parameterized by a long-term predictor [17, 18, 19]. However, this model fails to describe the linear relationship between the phases of the harmonics. That is, the long-term predictor alone does not model the temporal localization of power in the excitation source. Instead, we

apply a time envelope that captures the localization and concentration of pitch pulse energy in the time domain. This, in turn, introduces an element of non-stationarity to our signal model because the excitation sequence is now modeled as a random sequence with time-varying variance, that is, the glottal pulses are modeled with higher variance and the rest of the excitation sequence is modeled with lower variance. This modeling of non-stationarity within a short frame implies a temporal resolution much finer than that of the quasi-stationarity-based-algorithms. The latter has a temporal resolution equal to the frame length. Thus we term the former the high temporal resolution model. It is worth noting that some unvoiced phonemes, such as plosives, have very fast changing waveform envelopes, which also could be modeled as non-stationarity within the analysis frame. In this paper, however, we focus on the non-stationary modeling of voiced speech.

3.1. Modeling signal covariance matrix

The signal covariance matrix is usually estimated by averaging the outer product of the signal vector over time. As an example this is done in the signal subspace approach [10]. This method assumes ergodicity of the autocorrelation function within the averaging interval.

Here we propose the following method of estimating $\mathbf{C_s}$ with the ability to model a certain element of non-stationarity within a short frame. The following discussion is only appropriate for voiced speech. Let $\mathbf{r}$ denote the excitation source vector and $\mathbf{H}$ denote the synthesis filtering matrix corresponding to the vocal tract filter such that

$$\mathbf{H} = \begin{bmatrix} h(0) & 0 & 0 & \cdots & 0 \\ h(1) & h(0) & 0 & & \vdots \\ h(2) & h(1) & h(0) & & \\ \vdots & \vdots & & \ddots & 0 \\ h(N-1) & h(N-2) & \cdots & & h(0) \end{bmatrix}, \tag{9}$$

where $h(n)$ is the impulse response of the LPC synthesis filter. We then have

$$\mathbf{s} = \mathbf{Hr}, \tag{10}$$

and therefore

$$\mathbf{C_s} = E[\mathbf{ss}^H] = \mathbf{HC_rH}^H, \tag{11}$$

where $\mathbf{C_r}$ is the covariance matrix of the model residual vector $\mathbf{r}$. In (11) we treat $\mathbf{H}$ as a deterministic quantity. This simplification is common practice also when the LPC filter model is used to parameterize the power spectral density in classic Wiener filtering [5, 20]. Section 3.2 addresses the estimation of $\mathbf{H}$. Note that (10) does not take into account the zero-input response of the filter in the previous frame. Either the zero-input response can be subtracted prior to the estimation of each frame or a windowed overlap-add procedure can be applied to eliminate this effect.

We now model $\mathbf{r}$ as a sequence of independent zero-mean random variables. The covariance matrix $\mathbf{C_r}$ is therefore diagonal with the variance of each element of $\mathbf{r}$ as its diagonal elements. For voiced speech, except for the pitch impulses, the rest of the residual is of very low amplitude and can be modeled as constant variance random variables. Therefore, the diagonal of $\mathbf{C_r}$ takes the shape of a constant floor with a few periodically located impulses. We term this the temporal envelope of the instantaneous residual power. This temporal envelope is an important part of the new MMSE estimator because it provides the information of uneven temporal power distribution. In the following two subsections, we will describe the estimation of the spectral envelope and the temporal envelope, respectively.

3.2. Estimating the spectral envelope

In the context of LPC analysis, the synthesis filter has a spectrum that is the envelope of the signal spectrum. Thus, our goal in this subsection is to estimate the spectral envelope of the signal. We first use the decision-directed method [3] to estimate the signal power spectrum and then use the autocorrelation method to find the spectral envelope.

The noisy signal power spectrum of the kth frame $|\mathbf{Y}(k)|^2$ is obtained by applying the DFT to the kth observation vector $\mathbf{y}(k)$ and squaring the amplitudes. The decision-directed estimate of the signal power spectrum of the kth frame, $|\hat{\hat{\boldsymbol{\theta}}}(k)|^2$, is a weighted sum of two parts, the power spectrum of the estimated signal of the previous frame, $|\hat{\boldsymbol{\theta}}(k-1)|^2$, and the power-spectrum-subtraction estimate of the current frame's power spectrum:

$$\begin{aligned} |\hat{\hat{\boldsymbol{\theta}}}(k)|^2 &= \alpha |\hat{\boldsymbol{\theta}}(k-1)|^2 \\ &\quad + (1-\alpha)\max\left(|\mathbf{Y}(k)|^2 - E\left[|\hat{\mathbf{V}}(k)|^2\right], 0\right), \end{aligned} \tag{12}$$

where α is a smoothing factor $\alpha \in [0,1]$ and $E[|\hat{\mathbf{V}}(k)|^2]$ is the estimated noise power spectral density. The purpose of such a recursive scheme is to improve the estimate of the power-spectrum-subtraction method by smoothing out the random fluctuation in the noise power spectrum, thus reducing the "musical noise" artifact [21]. Other iterative schemes with similar time or spectral constraints are applicable in this context. For a comprehensive study of constraint iterative filtering techniques, readers are referred to [5]. We now take the square root of the estimated power spectrum and combine it with the noisy phase to reconstruct the so called intermediate estimate, which has the noise-reduced amplitude spectrum and a noisy phase. An autocorrelation method LPC analysis is then applied to this intermediate estimate to obtain the synthesis filter coefficients.

3.3. Estimating the temporal envelope

We propose to use a modified MPLPC method to robustly estimate the temporal envelope of the residual power. MPLPC is first introduced by Atal and Remde [17] to optimally determine the impulse position and amplitude of the excitation

in the context of analysis-by-synthesis linear predictive coding. The principle is to represent the LPC residual with a few impulses in which the locations and amplitudes (gains) of the impulses are chosen such that the difference between the target signal and the synthesized signal is minimized. In the noise reduction scenario, the target signal will be the noisy signal and the synthesis filter must be estimated from the noisy signal. Here, the synthesis filter is treated as known. For the residual of voiced speech, there is usually one dominating impulse in each pitch period. We first determine one impulse per pitch period then model the rest of the residual as a noise floor with constant variance. In MPLPC the impulses are found sequentially [22]. The first impulse location and amplitude are found by minimizing the distance between the synthesized signal and the target signal. The effect of this impulse is subtracted from the target signal and the same procedure is applied to find the next impulse. Because this way of finding impulses does not take into account the interaction between the impulses, reoptimization of the impulse amplitudes is necessary every time a new impulse is found. The number of pitch impulses p in a frame is determined in the following way. p is first assigned an initial value equal to the largest number of pitch periods possible in a frame. Then p impulses are determined using the abovementioned method. Only the impulses with an amplitude larger than a threshold are selected as pitch impulses. In our experiment, the threshold is set to 0.5 times the largest impulse amplitude in this frame. Having determined the impulses, a white noise sequence representing the noise floor of the excitation sequence is added into the gain optimization procedure together with all the impulses. We use a codebook of 1024 white Gaussian noise sequences in the optimization. The white noise sequence that yields the smallest synthesis error to the target signal is chosen to be the estimate of the noise floor. This procedure is in fact a multistage coder with p impulse stages and one Gaussian codebook stage, with a joint reoptimization of gains. Detailed treatment of this optimization problem can be found in [23]. After the optimization, we use a flat envelope equal to the square of the gain of the selected noise sequence to model the variance of the noise floor. Finally, the temporal envelope of the instantaneous residual power is composed of the noise floor variance and the squared impulses. When applied to noisy signals, the MPLPC procedure can be interpreted as a nonlinear least square fitting to the noisy signal, with the impulse positions and amplitudes as the model parameters.

4. THE ALGORITHM

Having obtained the estimate of the temporal envelope of the instantaneous residual power and the estimate of the synthesis filter matrix, we are able to build the signal covariance matrix in (11). The covariance matrix is used in the time LMMSE estimator (3) or in the spectral LMMSE estimator (6) after being transformed by (7).

The noise covariance matrix can be estimated using speech-absent frames. Here, we assume the noise to be stationary. For the time domain LMMSE estimator (3), if the noise is white, the covariance matrix $\mathbf{C_v}$ is diagonal with the noise variance as its diagonal elements. In the case of colored noise, the noise covariance matrix is no longer diagonal and it can be estimated using the time-averaged outer product of the noise vector. For the spectral domain LMMSE estimator (6), $\mathbf{C_V}$ is a diagonal matrix with the power spectral density of the noise as its diagonal elements. This is due to the assumed stationarity of the noise.[1] In the special case where the noise is white, the diagonal elements all equal the variance of the noise.

We model the instantaneous power of the residual of unvoiced speech with a flat envelope. Here, voiced speech is referred to as phonemes that require excitation from the vocal folds vibration, and unvoiced speech consists of the rest of the phonemes. We use a simple voiced/unvoiced detector

(1) Take the kth frame.
(2) Estimate the noise PSD from the latest speech-absent frame.
(3) Calculate the power spectrum of the noisy signal.
(4) Do power-spectrum-subtraction estimation of the signal PSD, and refine the estimate using decision-directed smoothing (equation (12)).
(5) Reconstruct the signal by combining the amplitude spectrum estimated by step (4) and the noisy phase.
(6) Do LPC analysis to the reconstructed signal. Obtain the synthesis filter coefficients and form the synthesis matrix $\mathbf{H}$.
(7) *IF* the frame is voiced
    Estimate the envelope of the instantaneous residual power using the modified MPLPC method.
(8) *IF* the frame is unvoiced
    Use a constant envelope for the instantaneous residual power.
(9) *ENDIF*
(10) Calculate the residual covariance matrix $\mathbf{C_r}$.
(11) Form the signal covariance matrix $\mathbf{C_s} = \mathbf{H}\mathbf{C_r}\mathbf{H}^H$ (equation (11)).
(12) *IF* time domain LMMSE:
    $\hat{\mathbf{s}} = \mathbf{C_s}(\mathbf{C_s} + \mathbf{C_v})^{-1}\mathbf{y}$ (equation (3)).
(13) *IF* frequency domain LMMSE: transform $\mathbf{C_s}$ to frequency domain $\mathbf{C}_{\boldsymbol{\theta}} = \mathbf{F}\mathbf{C_s}\mathbf{F}^{-1}$, filter the noisy spectrum $\hat{\boldsymbol{\theta}} = \mathbf{C}_{\boldsymbol{\theta}}(\mathbf{C}_{\boldsymbol{\theta}} + \mathbf{C_V})^{-1}\mathbf{Y}$ (equation (6)), and obtain the signal estimate by inverse DFT.
(14) *ENDIF*
(15) Calculate the power spectrum of the filtered signal, $|\hat{\boldsymbol{\theta}}(k-1)|^2$, for use in the PSD estimation of next frame.
(16) $k = k + 1$ and go to step (1).

Algorithm 1: TFE-MMSE estimator.

[1] In modeling the spectral covariance matrix of the noise we have ignored the inter-frequency correlations caused by the finite-length window effect. With typical window length, for example, 15 ms to 30 ms, the inter-frequency correlations caused by the window effect are less significant than those caused by the non-stationarity of the signal. This can be easily seen by examining a plot of the spectral covariance matrix.

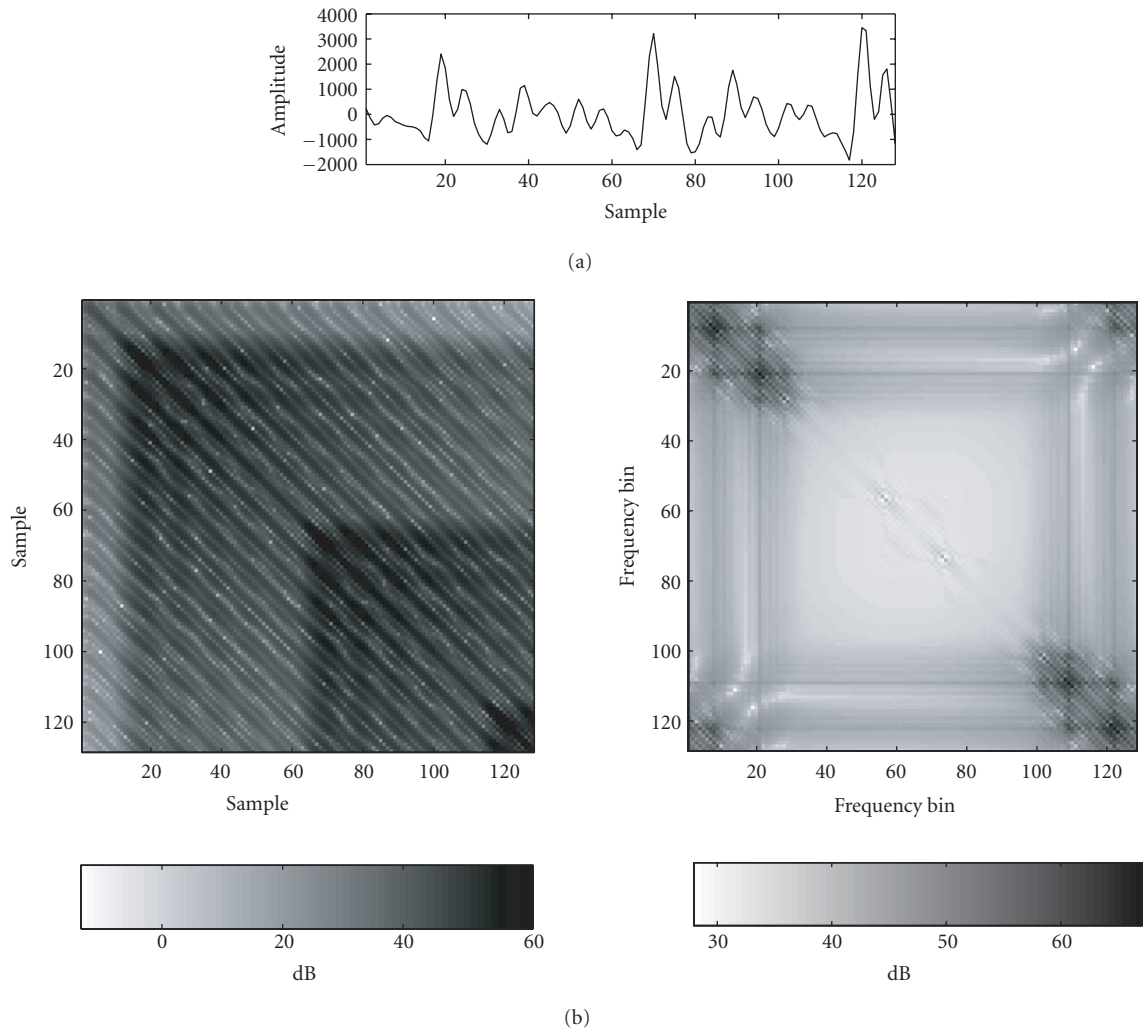

FIGURE 1: (a) The voiced speech waveform and (b) its time domain (left) and frequency domain (right) (amplitude) covariance matrices estimated with the nonstationary model. Frame length is 128 samples.

that utilize the fact that voiced speech usually has most of its power concentrated in the low frequency band, while unvoiced speech has a relatively flat spectrum within 0 to 4 kHz. Every frame is lowpass filtered and then the filtered signal power is compared with the original signal power. If the power loss is more than a threshold, the frame is marked as an unvoiced frame, and vice versa. Note however that even for the unvoiced frames, the spectral covariance matrix is non-diagonal because the signal covariance matrix $\mathbf{C_s}$, built in this way, is not Toeplitz. Hereafter, we refer to the proposed approach as the time-frequency-envelope MMSE estimator (TFE-MMSE), due to its utilization of envelopes in both time and frequency domains. The algorithm is summarized in Algorithm 1.

5. REDUCING COMPUTATIONAL COMPLEXITY

The TFE-MMSE estimators require inversion of a full covariance matrix $\mathbf{C_s}$ or $\mathbf{C}_{\boldsymbol{\theta}}$. This high computational load prohibits the algorithm from real-time application in hearing aids. Noticing that both covariance matrices are symmetric and positive definite, Cholesky factorization can be applied to the covariance matrices, and the inversion can be done by inverting the Cholesky triangle. A careful implementation requires $N^3/3$ operations for the Cholesky factorization [24] and the algorithm complexity is $\mathcal{O}(N^3)$. Another computation intensive part of the algorithm is the modified MPLPC method. In this section we propose simplifications to these two parts.

Further reduction of complexity for the filtering requires understanding the inter-frequency correlation. In the time domain the signal samples are clearly correlated with each other in a very long span. However, in the frequency domain, the correlation span is much smaller. This can be seen from the magnitude plots of the two covariance matrices (see Figure 1).

For the spectral covariance matrix, the significant values concentrate around the diagonal. This fact indicates that a small number of diagonals capture most of the inter-frequency correlation. The simplified procedure is as follows.

Half of the spectrum vector $\boldsymbol{\theta}$ is divided into small segments of l frequency bins each. The subvector starting at the jth frequency is denoted as $\boldsymbol{\theta}_{\text{sub},j}$, where $j \in [1, l, 2l, \ldots, N/2]$ and $l \ll N$. The noisy signal spectrum and the noise spectrum can be segmented in the same way giving $\mathbf{Y}_{\text{sub},j}$ and $\mathbf{V}_{\text{sub},j}$. The LMMSE estimate of $\boldsymbol{\theta}_{\text{sub},j}$ needs only a block of the covariance matrix, which means that the estimate of a frequency component benefits from its correlations with l neighboring frequency components instead of all components. This can be written as

$$\hat{\boldsymbol{\theta}}_{\text{sub},j} = \mathbf{C}_{\boldsymbol{\theta}_{\text{sub},j}} \left(\mathbf{C}_{\boldsymbol{\theta}_{\text{sub},j}} + \mathbf{C}_{V_{\text{sub},j}}\right)^{-1} \mathbf{Y}_{\text{sub},j}. \tag{13}$$

The first half of the signal spectrum can be estimated segment by segment. The second half of the spectrum is simply a flipped and conjugated version of the first half. The segment length is chosen to be $l = 8$, which, in our experience, does not degrade performance noticeably when compared with the use of the full matrix. Other segmentation schemes are applicable, such as overlapping segments. It is also possible to use a number of surrounding frequency components to estimate a single component at a time. We use the nonoverlapping segmentation because it is computationally less expensive while maintaining good performance for small l. When the signal frame length is 128 samples and the block length is $l = 8$, using this simplified method requires only $8 \times 8^3/128^3 = 1/512$ times of the original complexity for the filtering part of the algorithm with an extra expense of FFT operations to the covariance matrix. When l is set to values larger than 24, very little improvement in performance is observed. When l is set to values smaller than 8, the quality of enhanced speech degrades noticeably. By tuning the parameter l, an effective trade-off between the enhanced speech quality and the computational complexity is adjusted conveniently.

In the MPLPC part of the algorithm, the optimization of the impulse amplitude and the gain of the noise floor brings in heavy computational load. It can be simplified by fixing the impulse shape and the noise floor level. In the simplified version, the MPLPC method is only used for searching the locations of the p dominating impulses. Once the locations are found, a predetermined pulse shape is put at each location. An envelope of the noise floor is also predetermined. The pulse shape is chosen to be wider than an impulse in order to gain robustness against estimation error of the impulse locations. This is helpful as long as noise is present. The pulse shape used in our experiment is a raised cosine waveform with a period of 18 samples and the ratio between the pulse peak and the noise floor amplitude is experimentally determined to be 6.6. Finally, the estimated residual power must be normalized. Although the pulse shape and the relative level of the noise floor are fixed for all frames, experiments show that the TFE-MMSE estimator is not sensitive to this change. The performance of both the simplified procedure and the optimum procedure is evaluated in Section 6. Figure 2 shows the estimated envelopes of residual in the two ways.

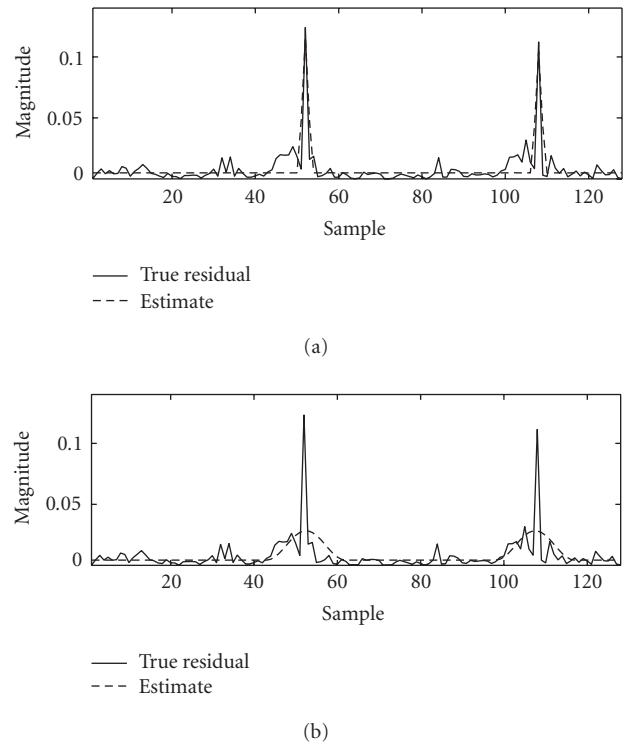

Figure 2: Estimated magnitude envelopes of the residual by (a) the complete MPLPC method and (b) the simplified MPLPC method.

6. RESULTS

Objective performance of the TFE-MMSE estimator is first evaluated and compared with the Wiener filter [2], the MMSE-LSA estimator [4], and the signal subspace method TDC estimator [10]. For the TFE-MMSE estimator, both the complete algorithm and the simplified algorithms are evaluated. For all estimators the sampling frequency is 8 kHz, and the frame length is 128 samples with 50% overlap. In the Wiener filter we use the same decision-directed method as in the MMSE-LSA and the TFE-MMSE estimator to estimate the PSD of the signal. An important parameter for the decision-directed method is the smoothing factor α. The larger the α, the more noise is removed and more distortion imposed to the signal, because of more smoothing made to the spectrum. In the MMSE-LSA estimator with the aforesaid parameter setting, we found experimentally $\alpha = 0.98$ to be the best trade-off between noise reduction and signal distortion. We use the same α for the WF and the TFE-MMSE estimator as for the MMSE-LSA estimator. For the TDC, the parameter μ ($\mu \geqq 1$) controls the degree of oversuppression of the noise power [10]. The larger the μ, the more attenuation to the noise but larger distortion to the speech. We choose $\mu = 3$ in the experiments by balancing the noise reduction and signal distortion.

All estimators run with 32 sentences from different speakers (16 male and 16 female) from the TIMIT database [25] added with white Gaussian noise, pink noise, and car

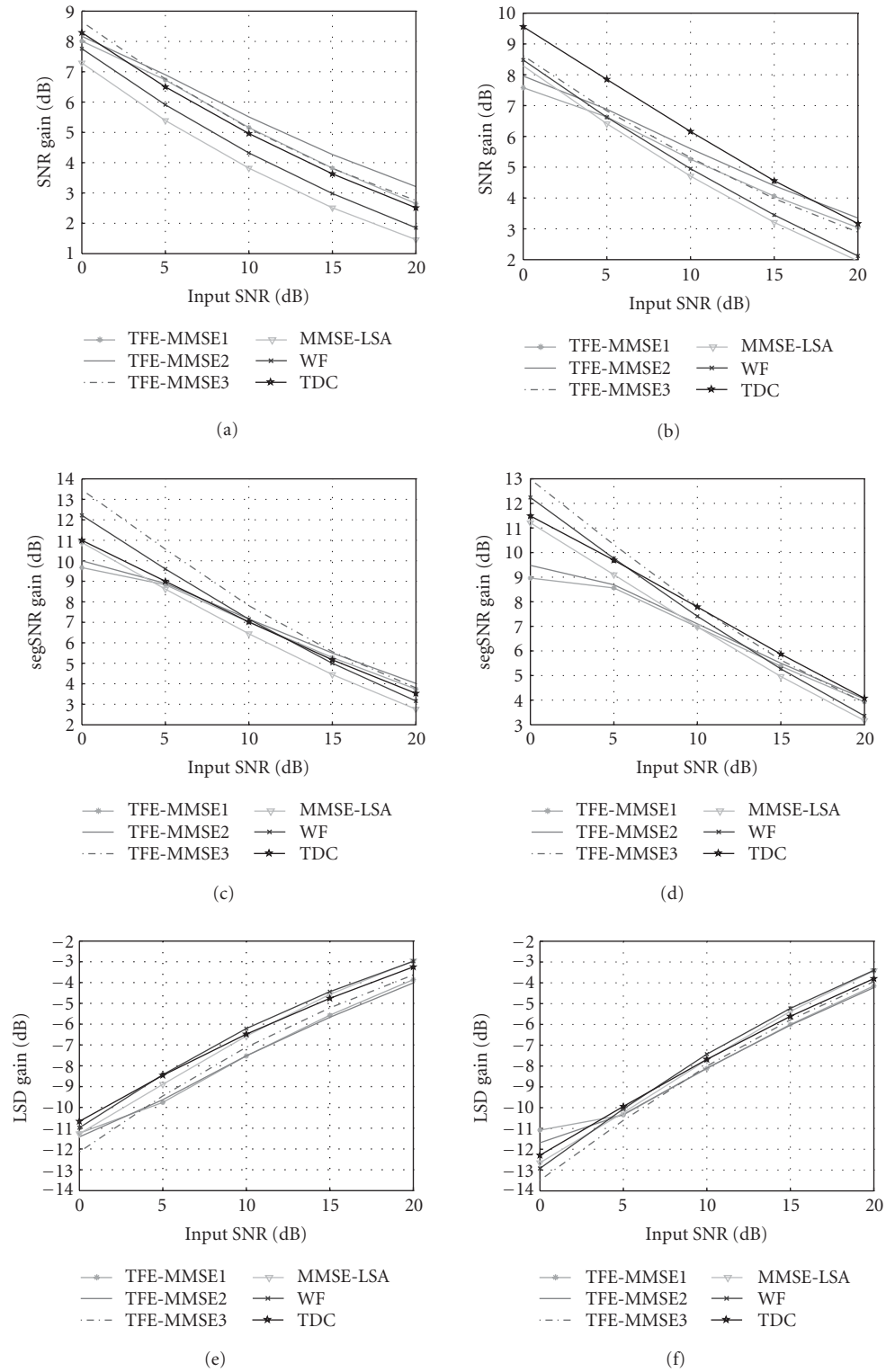

Figure 3: (a), (b) SNR gain, (c), (d) segSNR gain, and (e), (f) log-spectral distortion gain for the white Gaussian noise case. (a), (c), and (e) are for male speech and (b), (d), and (f) are for female speech.

noise in SNR ranging from 0 dB to 20 dB. The white Gaussian noise is computer generated, and the pink noise is generated by filtering white noise with a filter having a 3 dB per octave spectral power descend. The car noise is recorded inside a car with a constant speed. Its spectrum is more low-pass than the pink noise. The quality measures used include

Table 1: Preference test between WF and TFE-MMSE3 with additive white Gaussian noise.

Gender	Approach	15 dB	10 dB	5 dB
Male	WF	8%	7%	37%
speaker	TFE	92%	83%	63%
Female	WF	37%	33%	58%
speaker	TFE	63%	67%	42%

Table 2: Preference test between MMSE-LSA and TFE-MMSE3 with additive white Gaussian noise.

Gender	Approach	15 dB	10 dB	5 dB
Male	LSA	4%	25%	46%
speaker	TFE	96%	75%	54%
Female	LSA	25%	42%	50%
speaker	TFE	75%	58%	40%

the SNR, the segmental SNR, and the log-spectral distortion (LSD). The SNR is defined as the ratio of the total signal power to the total noise power in the sentence. The segmental SNR (segSNR) is defined as the average ratio of signal power to noise power per frame. To prevent the segSNR measure from being dominated by a few extreme low values, since the segSNR is measured in dB, it is common practice to apply a lower power threshold ϵ to the signals. Any frame that has an average power lower than ϵ is not used in the calculation. We set ϵ to 40 dB lower than the average power of the utterance. The segSNR is commonly considered to be more correlated to perceived quality than the SNR measure. The LSD is defined as [26]

$$\mathrm{LSD} = \frac{1}{K}\sum_{k=1}^{K}\left[\frac{1}{M}\sum_{m=1}^{M}\left(20\log_{10}\frac{|X(m,k)|+\epsilon}{|\hat{X}(m,k)|+\epsilon}\right)^2\right]^{1/2}, \quad (14)$$

where ϵ is to prevent extreme low values. We again set ϵ to 40 dB lower than the average power of the utterance. Results of the white Gaussian noise case are given in Figure 3. TFE-MMSE1 is the complete algorithm, and TFE-MMSE2 is the one with simplified MPLPC and reduced covariance matrix ($l = 8$). It is observed that the TFE-MMSE2, though a result of simplification of TFE-MMSE1, has better performance than the TFE-MMSE1. This can be explained as follows. (1) Its wider pulse shape is more robust to the estimation error of impulse positions. (2) The wider pulse shape can model to some extent the power concentration around the impulse peaks, which is overlooked by the spiky impulses. For this reason, in the following evaluations we investigate only the simplified algorithm.

Informal listening tests reveal that, although the speech enhanced by the TFE-MMSE algorithm has a significantly clearer sound (less muffled than the reference algorithms), the remaining background noise has musical tones. A solution to the musical noise problem is to set a higher value to the smoothing factor α. Using a larger α sacrifices the SNR and LSD slightly at high input SNRs, but improves the SNR and LSD at low input SNRs, and generally improves the segSNR significantly. The musical tones are also well suppressed. By setting $\alpha = 0.999$, the residual noise is greatly reduced, while the speech still sounds less muffled than for the reference methods. The reference methods cannot use a smoothing factor as high as the TFE-MMSE: experiments show that at $\alpha = 0.999$ the MMSE-LSA and the WF result in extremely muffled sounds. The TDC also suffers from a musical residual noise. To suppress its residual noise level to as low as that of the TFE-MMSE with $\alpha = 0.999$, the TDC requires a μ lager than 8. This causes a sharp degradation of the SNR and LSD and results in very muffled sounds. The TFE-MMSE2 estimator with a large smoothing factor ($\alpha = 0.999$) is hereafter termed TFE-MMSE3 and its objective measures are also shown in the figures. To verify the perceived quality of the TFE-MMSE3 subjectively, preference tests between the TFE-MMSE3 and the WF, and between the TFE-MMSE3 and the MMSE-LSA are conducted. The WF and the MMSE-LSA use their best value of smoothing factor ($\alpha = 0.98$). The test is confined to white Gaussian noise and a limited range of SNRs. Three sentences by male speakers and three by female speakers at each SNR level are used in the test. Eight unexperienced listeners are required to vote for their preferred method based on the amount of noise reduction and speech distortion. The utterances are presented to the listeners by a high-quality headphone. The clean utterance is first played as a reference, and the enhanced utterances are played once, or more if the listener finds this necessary. The results in Tables 1 and 2 show that (1) at 10 dB and 15 dB the listeners clearly prefer the TFE-MMSE over the two reference methods, while at 5 dB the preference on the TFE-MMSE is unclear; (2) the TFE-MMSE method has a more significant impact on the processing of male speech than on the processing of female speech. At 10 dB and above, the speech enhanced by TFE-MMSE3 has barely audible background noise, and the speech sounds less muffled than the reference methods. There is one artifact heard in rare occasions that we believe is caused by remaining musical tones. It is of very low power and occurs some times at speech presence. The two reference methods have higher residual background noise and suffer from muffling and reverberance effects. When SNR is lower than 10 dB, a certain speech-dependent noise occurs at speech presence in the TFE-MMSE3 processed speech. The lower the SNR is, the more audible this artifact is. Comparing the male and female speech processed by the TFE-MMSE3, the female speech sounds a bit rough.

The algorithms are also evaluated for pink noise and car noise cases. The objective results are shown in Figures 4 and 5. In these results the TDC algorithm is not included because the algorithm is proposed based on the white Gaussian noise assumption. An informal listening test shows that the perceptual quality in the pink noise case for all the three algorithms is very similar to that in the white noise case, and that in the car noise case all tested methods have very similar perceptual quality due to the very lowpass spectrum of the noise.

A comparison of spectrograms of a processed sentence (male "only lawyers love millionaires") is shown in Figure 6.

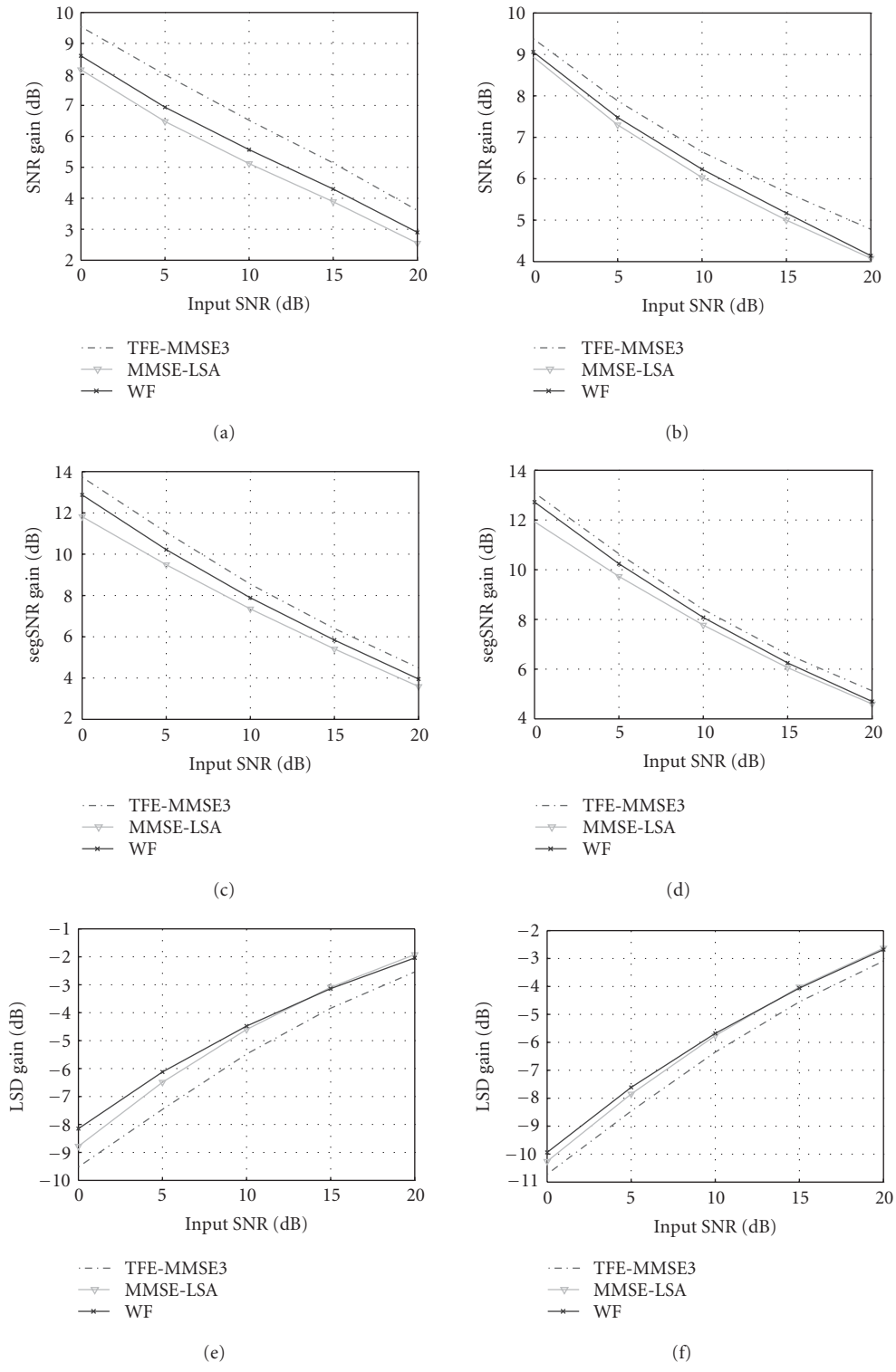

Figure 4: (a), (b) SNR gain, (c), (d) segSNR gain, and (e), (f) log-spectral distortion gain for the pink noise case. (a), (c), and (e) are for male speech and (b), (d), and (f) are for female speech.

7. DISCUSSION

The results show that for male speech, the TFE-MMSE3 estimator has the best performance in all the three objective measures (SNR, segSNR, and LSD). For female speech, the TFE-MMSE3 is the second in SNR, the best in LSD, and among the best in segSNR. The TFE-MMSE3 estimator allows a high degree of suppression to the noise while

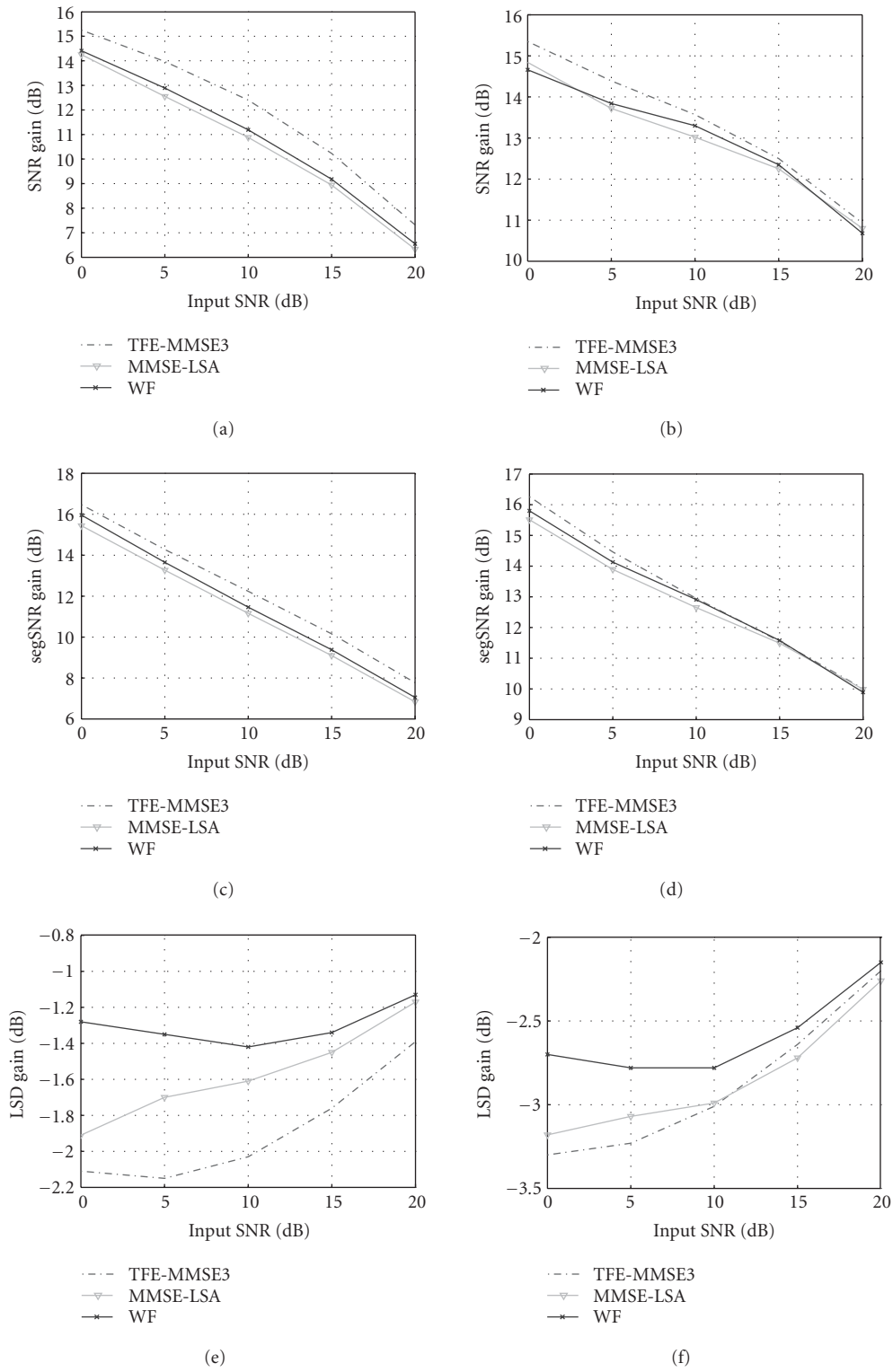

Figure 5: (a), (b) SNR gain, (c), (d) segSNR gain, and (e), (f) log-spectral distortion gain for the car noise case. (a), (c), and (e) are for male speech and (b), (d), and (f) are for female speech.

maintaining low distortion to the signal. The speech enhanced by the TFE-MMSE3 has a very clean background and a certain speech-dependent residual noise. When the SNR is high (10 dB and above), this speech-dependent noise is very well masked by the speech, and the resulting speech sounds clean and clear. As spectrograms in Figure 6 indicate, the

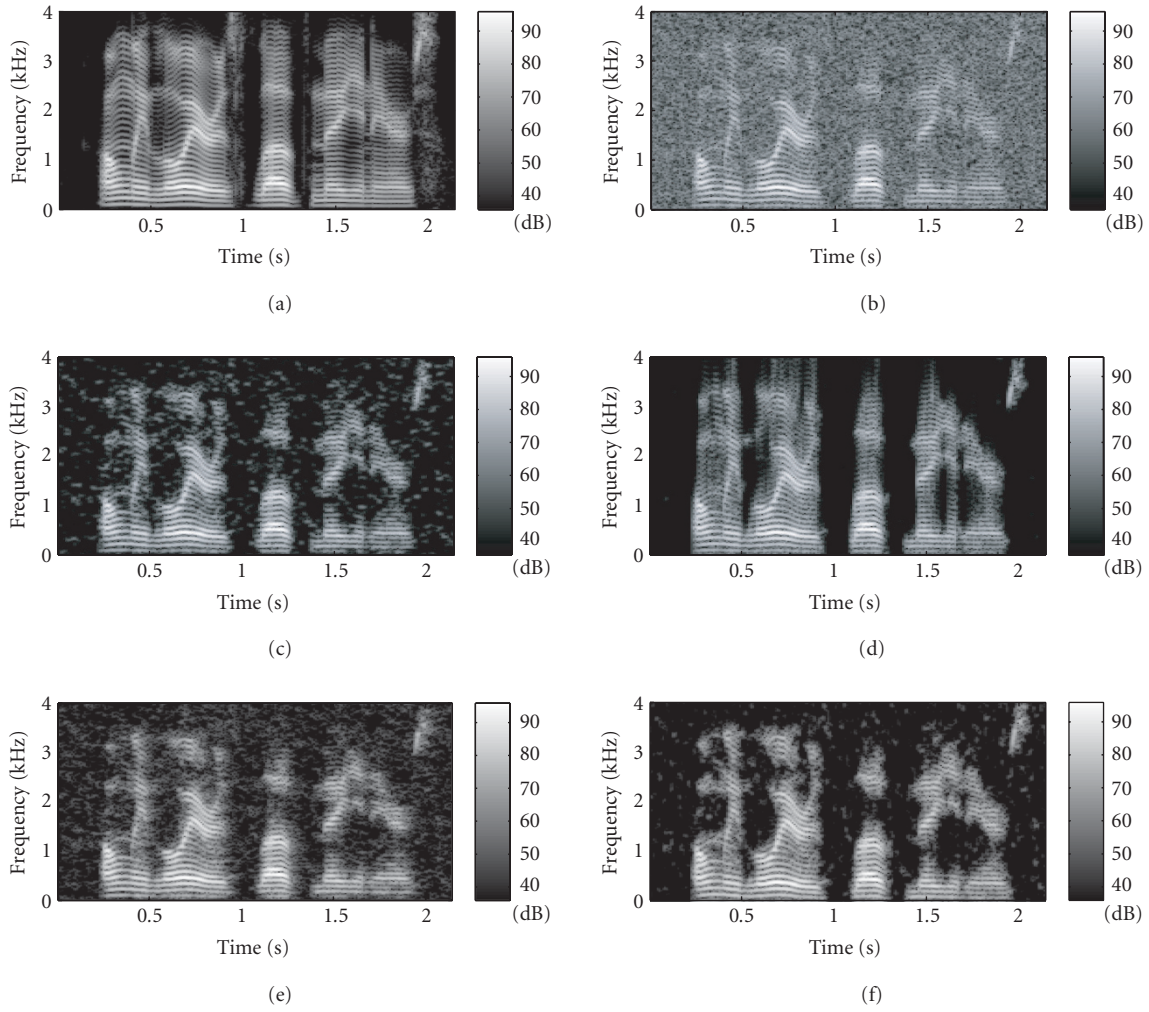

Figure 6: Spectrograms of enhanced speech. Input SNR is 10 dB. (a) Clean signal, (b) noisy signal, (c) TDC processed signal, (d) TFE-MMSE3 processed signal, (e) MMSE-LSA processed signal, and (f) WF processed signal.

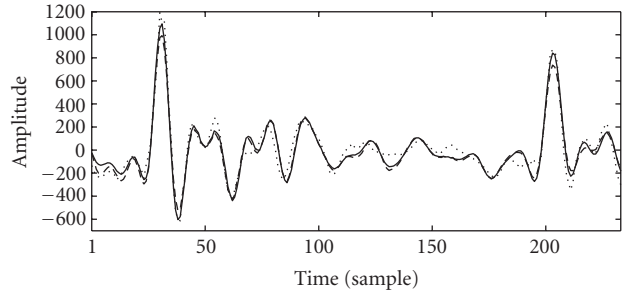

Figure 7: Comparison of waveforms of enhanced signals and the original signal. Dotted line: original, solid line: TFE-MMSE, dashed line: WF.

clearer sound is due to a better preserved signal spectrum and a more suppressed background noise. At SNR lower than 5 dB, although the background still sounds clean, the speech-dependent noise becomes audible and perceived as a distortion to the speech.The listeners preference starts shifting from the TFE-MMSE3 towards the MMSE-LSA that has a more uniform residual noise, although the noise level is high. The conclusion here is that at high SNR, it is preferable to remove background noise completely using the TFE-MMSE estimator without major distortion to the speech. This could be especially helpful at relieving listening fatigue for the hearing aid user, whereas, at low SNR, it is preferable to use a

noise reduction strategy that produces uniform background noise, such as the MMSE-LSA algorithm.

The fact that female speech enhanced by the TFE-MMSE estimator sounds a little rougher than the male speech is consistent with the observation in [15], where male voiced speech and female voiced speech are found to have different masking properties in the auditory system. For male speech, the auditory system is sensitive to high frequency noise in the valleys between the pitch pulse peaks in the time domain. For the female speech, the auditory system is sensitive to low frequency noise in the valleys between the harmonics in the spectral domain. While the time domain valley for the male speech is cleaned by the TFE-MMSE estimator, the spectral valleys for the female speech are not attenuated enough; a comb filter could help to remove the roughness in the female voiced speech.

In the TFE-MMSE estimator, we apply a high temporal resolution non-stationary model to explain the pitch impulses in the LPC residual of voiced speech. This enables the capture of abrupt changes in sample amplitude that are not captured by an AR linear stochastic model. In fact, the estimate of the residual power envelope contains information about the uneven distribution of signal power in time axis. In Figure 7 the original signal waveform, the WF enhanced signal waveform, and the TFE-MMSE enhanced signal waveform of a voiced segment are plotted. It can be observed in this figure that by a better model of temporal power distribution the TFE-MMSE estimator represents the sudden rises of amplitude better than the Wiener filter.

Noise in the phase spectrum is reduced by the TFE-MMSE estimator. Although human ears are less sensitive to phase than to power, it is found in recent work [27, 28, 29] that phase noise is audible when the source SNR is very low. In [27] a threshold of phase perception is found. This phase-noise tolerance threshold corresponds to an SNR threshold of about 6 dB, which means, for spectral components with local SNR smaller than 6 dB, that it is necessary to reduce phase noise. The TFE-MMSE estimator has the ability of enhancing phase spectra because of its ability to estimate the temporal localization of residual power. It is the linearity in the phase of harmonics in the residual that makes the power be concentrated at periodic time instances, thus producing pitch pulses. Estimating the residual power temporal envelope enhances the linearity of the phase spectrum of the residual and therefore reduces phase noise in the signal.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their many constructive suggestions, which have largely improved the presentation of our results. This work was supported by The Danish National Centre for IT Research Grant no. 329, and Microsound A/S.

REFERENCES

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[2] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

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

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[5] J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Signal Processing*, vol. 39, no. 4, pp. 795–805, 1991.

[6] R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02)*, vol. 1, pp. I-253–I-256, Orlando, Fla, USA, May 2002.

[7] W. B. Davenport Jr. and W. L. Root, *An Introduction to the Theory of Random Signals and Noise*, McGraw-Hill, New York, NY, USA, 1958.

[8] R. M. Gray, *Toeplitz and circulant matrices : a review*, [Online], available: http://ee.stanford.edu/~gray/toeplitz.pdf, 2002.

[9] C. Li and S. V. Andersen, "Inter-frequency dependency in MMSE speech enhancement," in *Proc. 6th Nordic Signal Processing Symposium (NORSIG '04)*, pp. 200–203, Espoo, Finland, June 2004.

[10] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 4, pp. 251–266, 1995.

[11] M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech enhancement from noise: a regenerative approach," *Speech Communication*, vol. 10, no. 1, pp. 45–57, 1991.

[12] D. E. Tsoukalas, J. N. Mourjoupoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Trans. Speech Audio Processing*, vol. 5, no. 6, pp. 497–514, 1997.

[13] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 2, pp. 126–137, 1999.

[14] K. H. Arehart, J. H. L. Hansen, S. Gallant, and L. Kalstein, "Evaluation of an auditory masked threshold noise suppression algorithm in normal-hearing and hearing-impaired listeners," *Speech Communication*, vol. 40, no. 4, pp. 575–592, 2003.

[15] J. Skoglund and W. B. Kleijn, "On time-frequency masking in voiced speech," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 4, pp. 361–369, 2000.

[16] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1993.

[17] B. Atal and J. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '82)*, vol. 7, pp. 614–617, Paris, France, May 1982.

[18] B. Atal, "Predictive coding of speech at low bit rates," *IEEE Trans. Commun.*, vol. 30, no. 4, pp. 600–614, 1982.

[19] B. Atal and M. R. Schroeder, "Adaptive predictive coding of speech signals," *Bell System Technical Journal*, vol. 49, no. 8, pp. 1973–1986, 1970.

[20] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 26, no. 3, pp. 197–210, 1978.

[21] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 2, no. 2, pp. 345–349, 1994.
[22] A. M. Kondoz, *Digital Speech, Coding for Low Bit Rate Communications Systems*, John Wiley & Sons, Chichester, UK, 1999.
[23] N. Moreau and P. Dymarski, "Selection of excitation vectors for the CELP coders," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 1, pp. 29–41, 1994.
[24] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, Md, USA, 1996.
[25] L. F. Lamel, J. Garafolo, J. Fiscus, W. Fisher, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continuous speech corpus," NTIS, Springfield, Va, USA, 1990, CDROM.
[26] J. M. Valin, J. Rouat, and F. Michaud, "Microphone array post-filter for separation of simultaneous non-stationary sources," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, vol. 1, pp. I-221–I-224, Montreal, Quebec, Canada, May 2004.
[27] P. Vary, "Noise suppression by spectral magnitude estimation—mechanism and theoretical limits," *Signal Processing*, vol. 8, no. 4, pp. 387–400, 1985.
[28] H. Pobloth and W. B. Kleijn, "On phase perception in speech," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '99)*, vol. 1, pp. 29–32, Phoenix, Ariz, USA, March 1999.
[29] J. Skoglund, W. B. Kleijn, and P. Hedelin, "Audibility of pitch–synchronously modulated noise," in *Proc. IEEE Workshop on Speech Coding For Telecommunications Proceeding*, vol. 7-10, pp. 51–52, Pocono Manor, Pa, USA, September 1997.

Chunjian Li received the B.S. degree in electrical engineering from Guangxi University, China, in 1997, and the M.S. degree in digital communication systems and technology from Chalmers University of Technology, Sweden, in 2003. He is currently with the Digital Communications Group (DICOM) at Aalborg University, Denmark. His research interests include digital signal processing and speech processing.

Søren Vang Andersen received his M.S. and Ph.D. degrees in electrical engineering from Aalborg University, Aalborg, Denmark, in 1995 and 1999, respectively. Between 1999 and 2002 he was with the Department of Speech, Music and Hearing at the Royal Institute of Technology, Stockholm, Sweden, and Global IP Sound AB, Stockholm, Sweden. Since 2002 he has been an Associate Professor with the Digital Communications (DICOM) Group at Aalborg University. His research interests are within multimedia signal processing: coding, transmission, and enhancement.