Codebook-based Bayesian speech enhancement for nonstationary environments

Codebook-Based Bayesian Speech Enhancement for Nonstationary Environments

Sriram Srinivasan, *Member, IEEE*, Jonas Samuelsson, and W. Bastiaan Kleijn, *Fellow, IEEE*

***Abstract*—In this paper, we propose a Bayesian minimum mean squared error approach for the joint estimation of the short-term predictor parameters of speech and noise, from the noisy observation. We use trained codebooks of speech and noise linear predictive coefficients to model the *a priori* information required by the Bayesian scheme. In contrast to current Bayesian estimation approaches that consider the excitation variances as part of the *a priori* information, in the proposed method they are computed online for each short-time segment, based on the observation at hand. Consequently, the method performs well in nonstationary noise conditions. The resulting estimates of the speech and noise spectra can be used in a Wiener filter or any state-of-the-art speech enhancement system. We develop both memoryless (using information from the current frame alone) and memory-based (using information from the current and previous frames) estimators. Estimation of functions of the short-term predictor parameters is also addressed, in particular one that leads to the minimum mean squared error estimate of the clean speech signal. Experiments indicate that the scheme proposed in this paper performs significantly better than competing methods.**

***Index Terms*—Bayesian, codebooks, linear predictive coding, noise estimation, speech enhancement, speech processing, Wiener filtering.**

I. Introduction

ADVANCES in telecommunications over the last few decades have made *communications anywhere* a reality. Technological progress has made communication systems reliable and affordable, and mobile communication has now become ubiquitous. The freedom and flexibility provided by mobile communications introduces new challenges, one of the most prominent being the suppression of background acoustic noise. Mobile users communicate in different environments with varying amounts and types of background noise. Suppression of the background noise is important not only to improve the quality and intelligibility of speech but also to obtain a good performance of speech coding algorithms. Noise suppression systems also form a crucial front-end for the operation of speech recognition and speaker verification systems in noisy environments.

Manuscript received January 27, 2005; revised February 20, 2006. This work was supported in part by the European Commission under the ANITA project (IST-2001-34327). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Rainer Martin.

S. Srinivasan was with the Department of Signals, Sensors and Systems, Royal Institute of Technology (KTH), Stockholm SE-100 44, Sweden. He is now with Philips Research Laboratories, 5656AE Eindhoven, The Netherlands (e-mail: sriram.srinivasan@philips.com).

J. Samuelsson and W. B. Kleijn are with the Department of Signals, Sensors, and Systems, Royal Institute of Technology (KTH), Stockholm S-100 44, Sweden (e-mail: jonas.samuelsson@s3.kth.se; bastiaan.kleijn@s3.kth.se).

Digital Object Identifier 10.1109/TASL.2006.881696

Noise reduction remains a challenging problem largely due to the wide variety of background noise types and the difficulty in estimating their statistics. Examples of noise types include traffic noise in cities, multitalker babble noise in cafeterias, noise in subways, etc. Many noise suppression techniques fall into the category of single-channel algorithms that have only a single microphone to obtain the input signal, and are thus attractive in mobile applications due to cost and size factors. Examples of such methods include [1]–[5]. A problem of single-channel methods is that noise estimates need to be obtained from the noisy observation. This has proved to be a particularly difficult task, especially in nonstationary noise conditions.

Conventional approaches to noise estimation have been based on voice activity detectors (VADs). Traditional energy based VADs detect regions in the signal where speech is absent to update the noise statistics. With decreasing signal-to-noise ratio (SNR), reliable detection of pauses becomes increasingly difficult. Soft-decision VADs facilitate adaptation of the noise statistics even during speech activity. Examples of such methods can be found in [6]–[8]. However, the estimates are based on long-term averaging. Other noise estimation methods that do not rely on a VAD and adapt even during speech activity include [9], [10]. They typically employ a buffer of past noisy spectra from which the estimates are obtained. For example, the method described in [9] is based on the observation that the power of the noisy signal frequently decays to that of the noise signal, and this can be tracked by following the minima in the buffer. While on the one hand, the buffer needs to be large enough to ensure that it contains the minima, on the other hand large buffers make it difficult to deal with time-varying noise, which is the case in the practical scenarios mentioned earlier. In the remainder of this paper, to indicate the dependence on the buffer, we refer to the noise estimates produced by [9] as long-term estimates. Based on this buffer, the method produces an estimate for each frame.

In this paper, we present a Bayesian approach to estimate speech and noise spectra in nonstationary noise conditions. We obtain minimum mean squared error (MMSE) estimates of the speech and noise auto-regressive (AR) spectra, which are parameterized by the respective AR coefficients and the excitation variance (gain). The AR coefficients and the gain are commonly referred to as the short-term predictor (STP) parameters. *A priori* information about the speech and noise AR coefficients is modeled using trained codebooks. We perform joint estimation of the speech and noise STP parameters. This is in contrast to methods that first obtain a noise estimate, e.g., using [9], and then obtain the speech parameters in a second step. The noise estimate is typically obtained using a buffer of past frames, and this affects the accuracy of the resulting speech estimates in nonstationary noise environments. The proposed joint esti-

mation is performed online, on a frame-by-frame basis, based on the current observation frame unlike conventional noise estimation techniques that rely on a buffer of past frames. This ensures good performance in nonstationary environments, thus addressing a fundamental limitation of current noise estimation techniques. A potential problem of frame-by-frame gain computation is that the estimates may posses a high variance. To solve this problem, we also develop memory-based MMSE estimators. This paper is an extension of the work presented in [11] and includes memory-based estimation and detailed experimental evaluations in both the STP parameter domain and the speech signal domain.

The maximum-likelihood (ML) estimation first proposed in [12] and extended in [13] also uses *a priori* information about speech and noise and performs instantaneous gain computation. It was shown in [13] that the method provides superior performance compared to other methods using prior information such as [14]–[16]. While the AR coefficients were considered to be deterministic parameters in the ML scheme, in this paper, we treat them as random variables and obtain minimum mean squared error (MMSE) estimates. In terms of speech and noise codebooks, while in [12] and [13], one pair of speech and noise LP vectors was selected as the ML estimate, the MMSE estimate of the speech (noise) LP vector is a weighted sum of the speech (noise) codebook vectors. Similarly, the MMSE estimate of the speech and noise excitation variances is the weighted sum of the excitation variances corresponding to each pair of speech and noise codebook vectors and the noisy observation. Thus, the MMSE estimation can be seen as a soft-decision procedure that allows for a proportionate contribution from vectors according to their probability given the observation. The MMSE estimator takes into account the *a priori* probabilities of each of the speech and noise codebook vectors.

Bayesian MMSE estimation using *a priori* information has been addressed before, e.g., the methods based on hidden Markov models (HMMs) [4], [5], [16], [17]. In [4], the clean signal is modeled using Gaussian AR HMMs. The MMSE estimate of clean speech given the noisy speech is obtained as a weighted sum of MMSE estimators corresponding to each state of the HMM for the clean signal. However, the HMM-based systems treat the excitation variance as part of the *a priori* information. The MMSE estimate in [18] also treats the excitation variance as part of the *a priori* information. To account for the resulting mismatch in the level of the gain of the clean speech model during training and testing, the HMM methods usually include gain adaptation. Similarly, there is gain adaptation in the noise model too. For the speech model and models corresponding to stationary noise, an overall gain adjustment in time is sufficient. However to effectively deal with nonstationary noise, the gain adjustment needs to be performed either on a frame-by-frame basis or at a rate not slower than the rate at which the noise statistics change. Both forms of gain adaptation depend upon an estimate of the noise statistics, obtained from the observation. Consequently, the performance of these methods is limited by the performance of the underlying noise estimation algorithms in nonstationary environments.

In the method proposed in this paper, we avoid this problem by modeling prior information about the spectral shape alone and jointly computing the speech and noise gain on a frame-by-frame basis.

The remainder of this paper is organized as follows. In Section II, we give an overview of the codebook based maximum-likelihood estimation, including the joint gain estimation, which will be used in the proposed method. The Bayesian approach is introduced in Section III, where we first obtain the memoryless MMSE estimate of the speech and noise LP coefficients and their excitation variances in Section III-A, followed in Section III-B by estimates that incorporate memory. MMSE estimation of functions of the LP coefficients and excitation variances is discussed in Section III-C. The relation between the proposed approach and HMM-based methods is discussed in III-D. Experiments and results are discussed in Section IV and finally the conclusion is presented in Section V.

II. Codebook-Based ML Parameter Estimation

In this section, we provide a brief overview of the codebook-based ML estimation procedure, to establish the necessary background for the Bayesian estimation. We consider an additive noise model where speech and noise are independent

$$y(n) = x(n) + w(n) \tag{1}$$

where $y(n), x(n)$, and $w(n)$ represent the sampled noisy speech, clean speech, and noise, respectively. We use trained codebooks of speech and noise power spectral shapes parameterized as LP coefficients. The codebooks model only the envelope of the spectrum and not its fine structure. LP coefficients have been successfully used to encode the spectral envelope in low bit rate speech coding [19]. In the ML approach, the speech and noise codebook indices and the excitation variances corresponding to the vectors that the indices represent are obtained according to

$$\{i^*, j^*, {\sigma_x^2}^*, {\sigma_w^2}^*\} = \arg \max_{i,j,\sigma_x^2,\sigma_w^2} p\left(\mathbf{y}|\theta_x^i, \theta_w^j, \sigma_x^2, \sigma_w^2\right) \tag{2}$$

where σ_x^2 and σ_w^2 are the excitation variances of clean speech and noise, respectively, and $\theta_x^i = (a_{x_0}^i, \ldots, a_{x_p}^i)$ and $\theta_w^j = (a_{w_0}^j, \ldots, a_{w_q}^j)$ are the LP coefficients of clean speech and noise with p and q being the respective LP-model orders. $\mathbf{y} = [y(0)y(1)\ldots y(K-1)]^T$, where K is the number of samples in a frame. Let $A_x^i(\omega)$ and $A_w^j(\omega)$ denote the spectra of the ith speech codebook and jth noise codebook vectors given by

$$A_x^i(\omega) = \sum_{k=0}^{p} a_{x_k}^i e^{-j\omega k}$$
$$A_w^j(\omega) = \sum_{k=0}^{q} a_{w_k}^j e^{-j\omega k}. \tag{3}$$

We define the modeled noisy spectrum as $\hat{P}_y^{ij}(\omega) = (\sigma_x^2)/(|A_x^i(\omega)|^2) + (\sigma_w^2)/(|A_w^j(\omega)|^2)$. Under Gaussianity assumptions, it is well known that maximizing the log-likelihood is

equivalent to minimizing the Itakura–Saito distortion measure [20]. The Itakura–Saito measure between two spectra P_y and $\hat{P}_y$ is defined as [21]

$$d_{\mathrm{IS}}(P_y,\hat{P}_y)=\frac{1}{2\pi}\int_0^{2\pi}\left(\frac{P_y(\omega)}{\hat{P}_y(\omega)}-\ln\left(\frac{P_y(\omega)}{\hat{P}_y(\omega)}\right)-1\right)d\omega. \tag{4}$$

Using this fact, for the noisy case, the parameter estimation problem (2) is solved in [13] by finding the best spectral fit between the observed noisy power spectrum $P_y(\omega)$ and the modeled noisy power spectrum $\hat{P}_y^{ij}(\omega)$, with respect to the Itakura–Saito distortion measure. Codebook combinations that result in negative values for the variances are excluded from the search for the best fit. More formally, the codebook entries that are selected can be written as

$$\{i^*,j^*\}=\arg\min_{i,j}\left\{\min_{\sigma_x^2,\sigma_w^2} d_{\mathrm{IS}}(P_y(\omega),\frac{\sigma_x^2}{\left|A_x^i(\omega)\right|^2}+\frac{\sigma_w^2}{\left|A_w^j(\omega)\right|^2})\right\}. \tag{5}$$

For given $A_x(\omega)$ and $A_w(\omega)$, the excitation variances that minimize the Itakura–Saito distortion between P_y and $\hat{P}_y$ can be obtained under the assumption of small modeling errors by using a series expansion for $\ln(x)$ up to second-order terms. This assumption can be made valid by using a sufficiently large codebook and by using the envelope of the noisy signal instead of the periodogram for P_y. The resulting variances are given by the solution to the following system of equations [13]:

$$\mathbf{C}\begin{bmatrix}\sigma_x^2\\ \sigma_w^2\end{bmatrix}=\mathbf{D} \tag{6}$$

where $\mathbf{C}$ and $\mathbf{D}$ are given by

$$\mathbf{C}=\begin{bmatrix}\left\|\frac{1}{P_y^2(\omega)|A_x(\omega)|^4}\right\| & \left\|\frac{1}{P_y^2(\omega)|A_x(\omega)|^2|A_w(\omega)|^2}\right\|\\ \left\|\frac{1}{P_y^2(\omega)|A_x(\omega)|^2|A_w(\omega)|^2}\right\| & \left\|\frac{1}{P_y^2(\omega)|A_w(\omega)|^4}\right\|\end{bmatrix}$$
$$\mathbf{D}=\begin{bmatrix}\left\|\frac{1}{P_y(\omega)|A_x(\omega)|^2}\right\|\\ \left\|\frac{1}{P_y(\omega)|A_w(\omega)|^2}\right\|\end{bmatrix} \tag{7}$$

where $\|f(\omega)\|=\int|f(\omega)|d\omega$.

III. Bayesian MMSE Estimation

In this section, we describe various aspects of the Bayesian approach. We first derive the memoryless Bayesian MMSE estimates of the speech and noise short-term predictor (STP) parameters in Section III-A. In Section III-B, we derive the Bayesian estimates using the noisy observation for the current frame and the MMSE estimates of the STP parameters for the previous frame. The resulting framework is then used to obtain the MMSE estimates of a function of the STP parameters in Section III-C, which is shown to result in the MMSE estimate of the clean speech signal, given the noisy speech. Finally, we discuss the relation of the proposed approach to existing model-based Bayesian approaches in Section III-D.

A. Memoryless MMSE Estimation of STP Parameters

Let θ_x and θ_w denote the random variables corresponding to the speech and noise LP coefficients, respectively. Let σ_x^2 and σ_w^2 denote the random variables corresponding to the speech and noise excitation variances, respectively. We wish to jointly estimate the speech and noise LP coefficients and the excitation variances such that the mean squared error is minimized. Let $\theta=[\theta_x,\,\theta_w,\,\sigma_x^2,\,\sigma_w^2]$. The desired MMSE estimate can be written as [22, p. 113]

$$\hat{\theta}=\mathrm{E}\{\theta|\mathbf{y}\}. \tag{8}$$

We rewrite (8) as

$$\hat{\theta}=\int_\Theta\theta p(\theta|\mathbf{y})d\theta=\int_\Theta\theta\frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})}d\theta \tag{9}$$

where $\mathbf{y}=[y(0)\,y(2)\ldots y(N-1)]^T$ is the observed vector of noisy samples for the current frame, N is the frame length, $p(\theta|\mathbf{y})$ is the conditional probablity density function (pdf) of θ given $\mathbf{y}$ and $\hat{\theta}=[\hat{\theta}_x,\,\hat{\theta}_w,\,\hat{\sigma}_x^2,\,\hat{\sigma}_w^2]$. We model $p(\mathbf{y}|\theta)$ as a zero-mean Gaussian with variance R_x+R_w. We have $R_x=\sigma_x^2(\mathbf{A}_x^T\mathbf{A}_x)^{-1}$, where $\mathbf{A}_x$ is the $N\times N$ lower triangular Toeplitz matrix with $[1\,a_{x_1}a_{x_2},\ldots,a_{x_p}0,\ldots,0]^T$ as the first column, where N is the frame length. R_w is defined analogously. The integral is over the space $\Theta=\Theta_x\times\Theta_w\times\Sigma_x\times\Sigma_w$, where Θ_x,Θ_w represent the support-space of the vectors of speech and noise LP coefficients, and Σ_x,Σ_w represent the support-space for the speech and noise excitation variances. From the independence assumption in (1), we have

$$p(\theta)=p\left(\theta_x,\sigma_x^2\right)p\left(\theta_w,\sigma_w^2\right). \tag{10}$$

For simplicity, we assume that the spectral shapes and gains are independent so that $p(\theta_x,\sigma_x^2)=p(\theta_x)p(\sigma_x^2)$ and likewise for the noise. This is a simplifying approximation made for tractability.

Given θ_x, θ_w and the noisy speech $\mathbf{y}$, it is shown in the Appendix that the likelihood $p(\mathbf{y}|\theta)$ decays rapidly from its maximum value as a function of the deviation from the true excitation variances, which we approximate by the ML estimates $\sigma_x^{2,\mathrm{ML}}$ and $\sigma_w^{2,\mathrm{ML}}$ obtained using (6) and (7). This behavior can be expressed mathematically by approximating $p(\mathbf{y}|\theta)$ with $p(\mathbf{y}|\theta)\delta(\sigma_x^2-\sigma_x^{2,\mathrm{ML}})\delta(\sigma_w^2-\sigma_w^{2,\mathrm{ML}})$. Thus, we can approximate (9), as shown by (11) at the bottom of the next page, where $\delta(\cdot)$ is the Dirac-delta function, $\theta'=[\theta_x,\,\theta_w,\,\sigma_x^{2,\mathrm{ML}},\,\sigma_w^{2,\mathrm{ML}}]$. Note that we now have an integral only over the support-space of two sets of LP coefficients. The Dirac assumption on the conditional pdf and the ML estimation of the variances is an assumption made for tractability and computational efficiency. The analysis in the Appendix and the experimental results justify the validity

of this assumption. $p(\mathbf{y})$ serves as a normalization term and can be obtained as

$$p(\mathbf{y}) = \int_{\Theta_x}\int_{\Theta_w} p(\mathbf{y}|\theta_x, \theta_w, \sigma_x^{2,\mathrm{ML}}, \sigma_w^{2,\mathrm{ML}}) \times p(\theta_x)p(\theta_w)p(\sigma_x^{2,\mathrm{ML}})p(\sigma_w^{2,\mathrm{ML}})d\theta_x d\theta_w. \quad (12)$$

In practice, the integrals in (11) and (12) are evaluated using numerical integration, as shown by (13) at the bottom of the page, where $\theta'_{ij} = [\theta_x^i, \theta_w^j, \sigma_{x,ij}^{2,\mathrm{ML}}, \sigma_{w,ij}^{2,\mathrm{ML}}]$, θ_x^i and θ_w^j are the ith speech codebook and jth noise codebook entries, respectively, $\sigma_{x,ij}^{2,\mathrm{ML}}$, $\sigma_{w,ij}^{2,\mathrm{ML}}$ are the maximum- likelihood estimates of the speech and noise excitation variances that depend on $\mathbf{y}, \theta_x^i$ and θ_w^j, and N_x, N_w are the speech and noise codebook sizes. To obtain (13) from (11), we discretized only the shapes θ_x and θ_w (represented by the codebooks) and not the excitation variances. Here, we assume that the codebooks model the probability density of the AR data. This is a valid assumption for codebooks with high dimensionality trained using the squared error distortion measure [23, ch. 5]. Since the excitation variances are completely determined given $\mathbf{y}$, θ_x and θ_w, we assume a noninformative prior for the excitation variances, i.e., we assume that they are uniformly distributed in the interval $[0, \sigma_{\max}^2]$. The exact value of $\sigma_{\max}^2$ is irrelevant since, for a uniform distribution, the terms cancel out in the numerator and denominator of (13). As in [13], codebook combinations that result in negative values for the excitation variances are excluded. Using the equivalence of the log-likelihood and the Itakura–Saito distortion, we can compute

$$p(\mathbf{y}|\theta_x^i, \theta_w^j, \sigma_{x,ij}^{2,\mathrm{ML}}, \sigma_{w,ij}^{2,\mathrm{ML}}) = C\exp(-d_{\mathrm{IS}}(P_y, \hat{P}_y^{ij,\mathrm{ML}})) \quad (14)$$

which allows an efficient computation in the frequency domain.[1] The term C, which is a constant with respect to the speech and noise STP parameters, also appears in the expression for $p(\mathbf{y})$, and thus cancels out in the numerator and denominator of (13). The estimate $\hat{\theta}$ can be used to construct a Wiener filter to obtain the enhanced speech

$$H_1(\omega) = \frac{\hat{\sigma}_x^2}{|\hat{A}_x(\omega)|^2} \Big/ \left(\frac{\hat{\sigma}_x^2}{|\hat{A}_x(\omega)|^2} + \frac{\hat{\sigma}_w^2}{|\hat{A}_w(\omega)|^2}\right) \quad (15)$$

where $\hat{A}_x(\omega), \hat{A}_w(\omega)$ are the spectra corresponding to $\hat{\theta}_x, \hat{\theta}_w$, respectively.

Since interpolation of LP coefficients can result in unstable filters, alternate representations are often used [19]. Representations that are guaranteed to result in stable synthesis filters include line spectral frequencies (LSFs), autocorrelation coefficients, reflection coefficients, and log-area ratios. Among these, it has been shown that LSFs result in the best performance and interpolation is often performed in this domain [19]. Thus, we perform the MMSE estimation in the LSF domain.

B. Memory-Based MMSE Estimation of STP Parameters

In this section, we exploit information from both the current and previous frames to derive the MMSE estimates of the STP parameters for the current frame. The motivation for doing so is that, in reality, parameters such as the speech and noise excitation variances are highly correlated across adjacent frames. Exploiting such correlation can result in estimates that have a reduced variance compared to the memoryless case. Since the memory is restricted to a small number of frames (in practice one 30-ms frame), the method retains its advantages of superior performance in nonstationary noise environments.

To incorporate memory, we would ideally like to derive a recursive estimator of the form $\hat{\theta}^n = \mathrm{E}\{\theta|\mathbf{y}_n, \mathbf{y}_{n-1}, \ldots\}$ where $\mathbf{y}_n$ is the vector of samples in frame n. However we did not find a mathematically tractable estimator that retains the instantaneous gain computation. Instead, we incorporate memory in the form of previous parameter estimates

$$\hat{\theta}^n = \mathrm{E}\{\theta|\mathbf{y}_n, \hat{\theta}^{n-1}\} \quad (16)$$

[1]To avoid problems with numerical precision, prior to taking the exponential, the maximum of the log-likelihood over all codebook entries can be subtracted from the log-likelihood corresponding to each codebook combination (i, j). The resulting probabilities are then normalized so that they add up to one.

$$\begin{aligned}\hat{\theta} &\approx \int_{\Theta} \theta \frac{p(\mathbf{y}\,|\,\theta)\delta\left(\sigma_x^2 - \sigma_x^{2,\mathrm{ML}}\right)\delta\left(\sigma_w^2 - \sigma_w^{2,\mathrm{ML}}\right)p(\theta_x)p(\theta_w)p\left(\sigma_x^2\right)p\left(\sigma_w^2\right)}{p(\mathbf{y})}d\theta \\ &= \int_{\Theta_x}\int_{\Theta_w} \theta' \frac{p\left(\mathbf{y}|\theta_x, \theta_w, \sigma_x^{2,\mathrm{ML}}, \sigma_w^{2,\mathrm{ML}}\right)p(\theta_x)p(\theta_w)p\left(\sigma_x^{2,\mathrm{ML}}\right)p\left(\sigma_w^{2,\mathrm{ML}}\right)}{p(\mathbf{y})}d\theta_x d\theta_w \end{aligned} \quad (11)$$

$$\begin{aligned}\hat{\theta} &= \frac{1}{N_x N_w}\sum_{i,j=1}^{N_x,N_w} \theta'_{ij} \frac{p\left(\mathbf{y}|\theta_x^i, \theta_w^j, \sigma_{x,ij}^{2,\mathrm{ML}}, \sigma_{w,ij}^{2,\mathrm{ML}}\right)p\left(\sigma_{x,ij}^{2,\mathrm{ML}}\right)p\left(\sigma_{w,ij}^{2,\mathrm{ML}}\right)}{p(\mathbf{y})} \\ p(\mathbf{y}) &= \frac{1}{N_x N_w}\sum_{i,j=1}^{N_x,N_w} p\left(\mathbf{y}|\theta_x^i, \theta_w^j, \sigma_{x,ij}^{2,\mathrm{ML}}, \sigma_{w,ij}^{2,\mathrm{ML}}\right)p\left(\sigma_{x,ij}^{2,\mathrm{ML}}\right)p\left(\sigma_{w,ij}^{2,\mathrm{ML}}\right) \end{aligned} \quad (13)$$

where $\hat{\theta}^n$ and $\hat{\theta}^{n-1}$ are the estimates of the STP parameter for frames n and $n-1$, respectively. $\hat{\theta}^n$ is the MMSE estimate given the observables $\mathbf{y}_n$ and $\hat{\theta}^{n-1}$ [22, p. 114]. In (16) and in the rest of the discussion, we drop the subscript in $\mathbf{y}_n$, and $\mathbf{y}$ refers to the current frame. Based on the theory developed in the previous section, we can rewrite (16) as

$$\hat{\theta}^n \approx \int_{\Theta_x}\int_{\Theta_w} \times\theta' \frac{p\left(\mathbf{y}|\theta_x,\theta_w,\hat{\theta}^{n-1},\sigma_x^{2,\mathrm{ML}},\sigma_w^{2,\mathrm{ML}}\right)p(\theta',\hat{\theta}^{n-1})}{p(\mathbf{y},\hat{\theta}^{n-1})}d\theta_x d\theta_w. \quad (17)$$

Given the noisy observation and the parameters for the current frame, we have $p(\mathbf{y}|\theta_x,\theta_w,\hat{\theta}^{n-1},\sigma_x^{2,\mathrm{ML}},\sigma_w^{2,\mathrm{ML}}) = p(\mathbf{y}|\theta_x,\theta_w,\sigma_x^{2,\mathrm{ML}},\sigma_w^{2,\mathrm{ML}})$. This follows from the fact that given the STP parameters for the current frame, which completely characterize the Gaussian pdf, the parameters from the previous frame do not affect the pdf. The probability that $\theta_x,\theta_w,\sigma_x^{2,\mathrm{ML}},\sigma_w^{2,\mathrm{ML}}$ are the correct parameters is embodied in the term $p(\theta',\hat{\theta}^{n-1})$. Thus, the memory in the system is modeled by the term $p(\theta',\hat{\theta}^{n-1})$ in (17). We have

$$\begin{aligned} p(\theta',\hat{\theta}^{n-1}) &= p\left(\theta_x,\theta_w,\sigma_x^{2,\mathrm{ML}},\sigma_w^{2,\mathrm{ML}},\hat{\theta}_x^{n-1},\hat{\theta}_w^{n-1},\hat{\sigma}_x^{2,n-1},\hat{\sigma}_w^{2,n-1}\right) \\ &\approx p\left(\theta_x,\sigma_x^{2,\mathrm{ML}},\hat{\theta}_x^{n-1},\hat{\sigma}_x^{2,n-1}\right) \\ &\quad\times p\left(\theta_w,\sigma_w^{2,\mathrm{ML}},\hat{\theta}_w^{n-1},\hat{\sigma}_w^{2,n-1}\right) \end{aligned} \quad (18)$$

where we used the assumption that the speech and noise parameters are independent. We note that while the independence assumption may not be strictly satisfied for the estimated parameters from the previous frame, we impose this restriction for simplicity and tractability. As before, we assume that the spectral shapes and the gains are independent so that $p(\theta_x,\sigma_x^{2,\mathrm{ML}},\hat{\theta}_x^{n-1},\hat{\sigma}_x^{2,n-1}) = p(\theta_x,\hat{\theta}_x^{n-1})p(\sigma_x^{2,\mathrm{ML}},\hat{\sigma}_x^{2,n-1})$ and likewise for the noise. We can now rewrite (17) as

$$\begin{aligned} \hat{\theta}^n &\approx \frac{1}{p\left(\mathbf{y},\hat{\theta}^{n-1}\right)}\int_{\Theta_x}\int_{\Theta_w}\theta' \\ &\quad\times p\left(\mathbf{y}|\theta_x,\theta_w,\sigma_x^{2,\mathrm{ML}},\sigma_w^{2,\mathrm{ML}}\right)p\left(\theta_x,\hat{\theta}_x^{n-1}\right) \\ &\quad\times p\left(\sigma_x^{2,\mathrm{ML}},\hat{\sigma}_x^{2,n-1}\right)p\left(\theta_w,\hat{\theta}_w^{n-1}\right) \\ &\quad\times p\left(\sigma_w^{2,\mathrm{ML}},\hat{\sigma}_w^{2,n-1}\right)d\theta_x d\theta_w \\ &= \frac{1}{p\left(\mathbf{y},\hat{\theta}^{n-1}\right)}\int_{\Theta_x}\int_{\Theta_w}\theta' \\ &\quad\times p\left(\mathbf{y}|\theta_x,\theta_w,\sigma_x^{2,\mathrm{ML}},\sigma_w^{2,\mathrm{ML}}\right)p\left(\hat{\theta}_x^{n-1}|\theta_x\right)p(\theta_x) \\ &\quad\times p\left(\sigma_x^{2,\mathrm{ML}},\hat{\sigma}_x^{2,n-1}\right)p\left(\hat{\theta}_w^{n-1}|\theta_w\right)p(\theta_w) \\ &\quad\times p\left(\sigma_w^{2,\mathrm{ML}},\hat{\sigma}_w^{2,n-1}\right)d\theta_x d\theta_w. \end{aligned} \quad (19)$$

In practice, we evaluate the integral in (19) using numerical integration

$$\begin{aligned} \hat{\theta}^n &= \frac{1}{N_x N_w p\left(\mathbf{y},\hat{\theta}^{n-1}\right)} \\ &\quad\times\sum_{i,j=1}^{N_x,N_w}\theta'_{ij}p\left(\mathbf{y}|\theta_x^i,\theta_w^j,\sigma_{x,ij}^{2,\mathrm{ML}},\sigma_{w,ij}^{2,\mathrm{ML}}\right) \\ &\quad\times p\left(\hat{\theta}_x^{n-1}|\theta_x^i\right)p\left(\hat{\theta}_w^{n-1}|\theta_w^j\right)p\left(\sigma_{x,ij}^{2,\mathrm{ML}},\hat{\sigma}_x^{2,n-1}\right) \\ &\quad\times p\left(\sigma_{w,ij}^{2,\mathrm{ML}},\hat{\sigma}_w^{2,n-1}\right) \end{aligned} \quad (20)$$

where

$$\begin{aligned} &p(\mathbf{y},\hat{\theta}^{n-1}) \\ &\quad= \frac{1}{N_x N_w}\sum_{i,j=1}^{N_x,N_w}p\left(\mathbf{y}|\theta_x^i,\theta_w^j,\sigma_{x,ij}^{2,\mathrm{ML}},\sigma_{w,ij}^{2,\mathrm{ML}}\right) \\ &\qquad\times p\left(\hat{\theta}_x^{n-1}|\theta_x^i\right)p\left(\hat{\theta}_w^{n-1}|\theta_w^j\right)p\left(\sigma_{x,ij}^{2,\mathrm{ML}},\hat{\sigma}_x^{2,n-1}\right) \\ &\qquad\times p\left(\sigma_{w,ij}^{2,\mathrm{ML}},\hat{\sigma}_w^{2,n-1}\right). \end{aligned} \quad (21)$$

As in the memoryless case, we assume that the codebooks model the probability density of the AR data and that the marginal pdf of the speech and noise excitation variances is uniform.

We approximate the joint distributions of the excitation variances $p(\sigma_x^2,\hat{\sigma}_x^{2,n-1})$ and $p(\sigma_w^2,\hat{\sigma}_w^{2,n-1})$ as bivariate Gaussians whose mean and covariance can be estimated from training data. The training data is in the form of pairs of excitation variances (obtained from clean speech or noise), corresponding to adjacent frames. The mean and the covariance depend on the level of the signal, which can differ during training and testing. This difference can be offset by scaling the mean and the covariance by a factor based on the long-term estimate of the excitation variance.

For the AR coefficients, we impose the Gaussian random walk (GRW) model [24, ch. 10] for the conditional prior pdfs. In the LSF domain, we have $p(\hat{\theta}_w^{n-1}|\theta_w^j) \sim \mathcal{N}(\theta_w^j,\Lambda_{\theta_w})$, i.e., we model the conditional pdf as a multivariate Gaussian with mean θ_w^j and variance Λ_{θ_w}, which is a $q\times q$ diagonal matrix. The kth diagonal entry $\lambda_{\theta_w}^k$ of Λ_{θ_w} determines how much the kth noise LSF component of the current frame can differ from the kth noise LSF component of the previous frame, i.e., the degree of smoothness is controlled by Λ_{θ_w}. A small value for Λ_{θ_w} corresponds to a smooth evolution of the parameters over time. The conditional pdfs corresponding to the speech parameters are defined analogously. The parameters Λ_{θ_w} and Λ_{θ_x} are obtained from training data (clean speech and noise, respectively) through a maximum-likelihood estimation.

C. MMSE Estimation of Functions of θ

The estimation framework represented by (11) and (17) can be used to obtain MMSE estimates of different parametric representations based on the LP coefficients. For simplicity, we consider the memoryless case here. Generalization to the memory-

based case is straightforward. For notational convenience, we define the function

$$f(\theta_x, \theta_w, \mathbf{y}) = \frac{p\left(\mathbf{y}|\theta_x, \theta_w, \sigma_x^{2,\mathrm{ML}}, \sigma_w^{2,\mathrm{ML}}\right) p(\theta')}{p(\mathbf{y})}. \quad (22)$$

The MMSE estimate of any function $g(\theta)$ can be obtained as

$$\begin{aligned}\hat{g}(\theta) &= \int_{\Theta} g(\theta) p(\theta|\mathbf{y}) d\theta \\ &= \int_{\Theta_x}\int_{\Theta_w} g(\theta') f(\theta_x, \theta_w, \mathbf{y}) d\theta_x d\theta_w. \end{aligned} \quad (23)$$

For example, let $g(\theta_x, \theta_w, \mathbf{y})$ be the Wiener filter defined as $H(\omega;\theta) = (\sigma_x^2)/(|A_x(\omega)|^2)/((\sigma_x^2)/(|A_x(\omega)|^2) + (\sigma_w^2)/(|A_w(\omega)|^2))$, where $A_x(\omega), A_w(\omega)$ are the spectra of the speech and noise LP coefficients θ_x, θ_w. The MMSE estimate $H_2(\omega)$ of the Wiener filter is obtained as

$$H_2(\omega) = \int_{\Theta_x}\int_{\Theta_w} H(\omega;\theta') f(\theta_x, \theta_w, \mathbf{y}) d\theta_x d\theta_w. \quad (24)$$

We note that the enhanced speech obtained by filtering $\mathbf{y}$ with the filter $H_2(\omega)$ is the MMSE estimate of the clean signal, $\mathrm{E}\{\mathbf{X}|\mathbf{y}\}$, where $\mathbf{X}$ is the random variable corresponding to clean speech. This can be seen if we write

$$\begin{aligned}\mathrm{E}\{\mathbf{X}|\mathbf{y}\} &= \int_{\Theta} p(\theta|\mathbf{y}) \mathrm{E}\{\mathbf{X}|\mathbf{y}, \theta\} d\theta \\ &= \int_{\Theta_x}\int_{\Theta_w} f(\theta_x, \theta_w, \mathbf{y}) \\ &\quad \times \mathrm{E}\left\{\mathbf{X}\,|\,\mathbf{y}, \theta_x, \theta_w, \sigma_x^{2,\mathrm{ML}}, \sigma_w^{2,\mathrm{ML}}\right\} d\theta_x d\theta_w. \end{aligned} \quad (25)$$

For Gaussian AR models, $\mathrm{E}\{\mathbf{X}|\mathbf{y}, \theta_x, \theta_w, \sigma_x^{2,\mathrm{ML}}, \sigma_w^{2,\mathrm{ML}}\}$ can be equivalently evaluated in the frequency domain as $H(\omega;\theta)\mathcal{Y}(\omega)$, where $\mathcal{Y}(\omega)$ is the Fourier transform of $\mathbf{y}$.

D. Relation to Existing Bayesian Approaches

In this section, we discuss similarities and differences to existing Bayesian speech enhancement approaches, specifically, the HMM-based approach discussed in [5]. Both the HMM used in [5] and codebook used here model the distribution of the AR parameters of the speech signal. The theoretical analysis in the estimation and use of such a model requires that the signal is stationary. In practice, both methods address the nonstationarity of the speech signal by performing the processing on a frame-by-frame basis, as speech can be described as a stationary process within a short frame of 20–30 ms.

The first difference between the HMM and codebook approaches lies in the manner in which they handle the nonstationarity of the noise signal, which in turn is related to the modelling and computation of the excitation variances. Since the HMM method models both the LP coefficients and the excitation variance as prior information, a gain adaptation is required to compensate for differences in the level of the excitation variance between training and operation. The gain adaptation factor is computed using the observed noisy gain and an estimate of the noise statistics obtained using, e.g., the minimum statistics approach [9]. Conventional noise estimation techniques are buffer-based techniques, where an estimate is obtained based on a buffer of several past frames, of the order of a few hundred milliseconds. Thus, such a scheme cannot react quickly to nonstationary noise. In the proposed approach, the codebook models only the LP coefficients, and the speech and noise excitation variances are optimally computed in a joint fashion on a frame-by-frame basis, using the current noisy observation. This enables the method to react quickly to nonstationary noise.

The second difference is that the HMM-based method obtains MMSE estimates of the clean speech signal as opposed to the codebook approach that obtains MMSE estimates of the speech and noise STP parameters. Let $\mathbf{X}$ denote the random variable corresponding to the clean speech signal. Given the noisy observations, the HMM method obtains the expected value of $\mathbf{X}$ and its functions such as the spectral magnitude and the log-spectral magnitude. The proposed codebook method obtains the expected value of θ given the noisy observations for the current and previous frames. The framework developed here also allows the MMSE estimation of arbitrary functions of the STP parameters as discussed in Section III-C, where the MMSE estimate of one such function is shown to result in the expected value of $\mathbf{X}$ given the noisy observations. We also note that the proposed technique of instantaneous frame-by-frame gain computation can be incorporated into the HMM-based scheme. This is, however, beyond the scope of this paper.

IV. Experiments

In this section, we describe the experiments performed to evaluate the performance of the MMSE estimation scheme. We first describe the experimental setup and the objective quality measures used in the evaluation. This is followed by an analysis of the memoryless and memory-based estimators. Next, we evaluate the performance of the proposed estimation scheme in the short-term predictor parameter domain. This includes a comparison to the estimates obtained using the long-term noise estimates [9]. Then, we compare the performance of the proposed MMSE method to the HMM-based estimation scheme [16] and the Ephraim–Malah system [25] in the speech signal domain. This is followed by a discussion on computational complexity. The section concludes with a description of the listening tests performed to evaluate perceptual quality.

A. Experimental Setup

The test set consisted of ten speech utterances, five male and five female, from the TIMIT database, resampled at 8 kHz. A ten-bit speech codebook of dimension ten was trained with 10 min of speech from the TIMIT database using the generalized Lloyd algorithm (GLA)[26]. The training data did not include the test utterances. A frame length of 240 samples was used with 50% overlap between adjacent frames. The frames were windowed using a Hanning window. The noise types considered were highway noise (obtained by recording noise on a freeway as perceived by a pedestrian standing at a fixed point), siren noise (a two-tone siren recorded inside a stationary emergency vehicle), speech babble noise (from Noisex-92), and

white Gaussian noise. An artificial nonstationary white noise (White-NS) was also used and was generated by alternating the variance of white Gaussian noise every 500 ms between σ^2 and $5\sigma^2$, where the actual value of σ^2 depends on the desired SNR. The noise codebooks were trained using the GLA with two minutes of training data. The noise samples used in the training and testing were different. For highway and white noise, the noise LP order was 6. For babble noise, the LP order was 10. For siren noise, which typically exhibits strong harmonics, the LP order was 16. The codebook for White-NS was the same as that for white noise. The number of vectors in the noise codebooks were empirically chosen to be 4, 8, 16, and 2 for highway, white, babble, and the two-tone siren noise, respectively [13]. For each frame, the classified noise codebook scheme discussed in [13] was used to select a noise codebook using an ML criterion based on the noisy observation. As in [13], to provide robustness towards unknown noise types, in addition to the trained entries, the noise codebook had one additional entry that was replaced each frame with the long-term estimate provided by [9].

B. Objective Quality Measures

The objective measures of quality used in this section are SNR, segmental SNR (SSNR), log-spectral distortion (SD), and perceptual evaluation of speech quality (PESQ). The SNR (in decibels) for an utterance was computed as

$$\mathrm{SNR} = 10\log_{10}\left(\frac{\sum_{t=1}^{T} x^2(t)}{\sum_{t=1}^{T}(x(t)-\hat{x}(t))^2}\right) \tag{26}$$

where $\hat{x}(t)$ is the modified (noisy or enhanced) speech, and T is the number of samples in the utterance. The SSNR was computed as the average of the SNR for each frame in the utterance. For the nth Hanning windowed frame, the instantaneous SD between the clean speech AR envelope $A_n(\omega)$ and the AR envelope of the processed signal $\hat{A}_n(\omega)$ was computed as

$$\mathrm{SD}_n = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left(10\log_{10}\frac{|A_n(\omega)|^2}{|\hat{A}_n(\omega)|^2}\right)^2 d\omega}.$$

The SD for an utterance was computed as the average of the instantaneous SD for the individual frames. While computing SSNR and SD, frames corresponding to silent segments were excluded [27]. PESQ scores were computed according to [28].

C. Memoryless Versus Memory-Based MMSE Estimation

From the experiments, it was observed that memory corresponding to the speech spectral shape and the speech excitation variance had little or no influence on the results. Using memory corresponding to the noise parameters was seen to result in a significant reduction of outliers in the noise excitation variances, as seen in Fig. 1. The figure plots the excitation variances for two noise types, highway and white, with and without memory. The true excitation variances are also plotted for reference. It can be seen that incorporating memory results in smoother estimates.

Table I quantifies the reduction in the variance of the estimates of the noise excitation variances. The table shows the mean and the variance of the normalized squared error between the true and the estimated noise excitation variances. The normalized squared error for frame i is defined as

$$\mathrm{MSE}_{\mathrm{norm}} = \left(\frac{\sigma_{w,i}^2 - \hat{\sigma}_{w,i}^2}{\bar{\sigma}_w^2}\right)^2 \tag{27}$$

where $\sigma_{w,i}^2$ and $\hat{\sigma}_{w,i}^2$ are the true and estimated noise excitation variances for the ith frame, and the normalizing factor $\bar{\sigma}_w^2$ is computed as the mean of the true excitation variances over all the frames.

We note that, in general, it is not meaningful to consider the excitation variances independently of the AR spectra. Accurate estimates of the speech excitation variance result in poor performance when combined with poor estimates of the gain normalized AR coefficients. For the noise estimates, the mean squared error values of the LSF coefficients obtained with and without memory, were not very different (less than 0.2-dB difference). Thus, in this case, it is meaningful to look at the excitation variances independently. Estimates of the excitation variances that track the nonstationarities well and yet exhibit low variance provide good perceptual performance. As seen in Table I, incorporating memory achieves a significant reduction in the variance of the error at the same or a lower mean.

To analyze the effect of memory in the speech signal domain, we compare the mean and the variance of the squared error between the clean speech and the enhanced speech obtained with and without memory in Table II. Enhanced speech was obtained

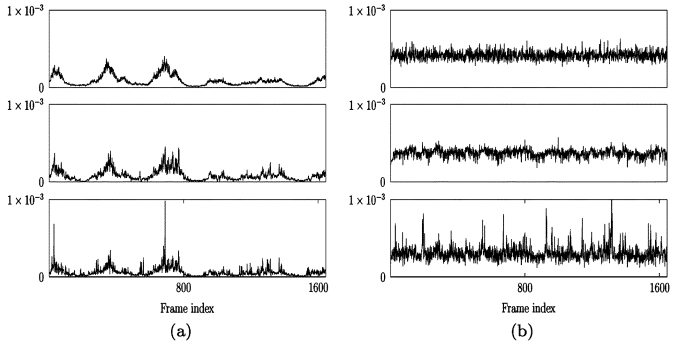

Fig. 1. Plot of the true and estimated noise excitation variances with and without memory. (a) Highway noise. (b) White noise. In each figure, the top plot corresponds to the true values of the excitation variances, the middle plot to memory-based estimates and the bottom plot to memoryless estimates.

TABLE I
MEAN AND VARIANCE OF THE NORMALIZED SQUARED ERROR BETWEEN THE TRUE AND ESTIMATED NOISE EXCITATION VARIANCES, WITH AND WITHOUT MEMORY. RESULTS ARE AVERAGED OVER TEN UTTERANCES AT 10-dB INPUT SNR

	Mean		Variance	
Noise	Memoryless	Memory	Memoryless	Memory
Highway	0.43	0.35	3.25	0.63
White	0.13	0.11	0.42	0.01
Babble	0.71	0.59	4.38	2.00
Siren	1.69	1.36	134.3	52.7
White-NS	0.26	0.27	0.63	0.18

TABLE II
MEAN AND VARIANCE OF THE SQUARED ERROR BETWEEN THE CLEAN AND ENHANCED SPEECH WAVEFORMS WITH AND WITHOUT MEMORY. RESULTS ARE AVERAGED OVER TEN UTTERANCES AT 10-dB INPUT SNR

	Mean $\times 10^{-4}$		Variance $\times 10^{-6}$	
Noise	Memoryless	Memory	Memoryless	Memory
Highway	3.74	3.41	1.96	1.01
White	3.57	3.28	1.08	0.61
Babble	7.07	5.51	5.18	1.82
Siren	1.97	1.90	0.44	0.32
White-NS	3.06	2.97	1.78	1.11

TABLE III
MEAN SQUARED ERROR IN LSF DOMAIN AVERAGED OVER TEN UTTERANCES AT 10-dB INPUT SNR FOR LSF COEFFICIENTS CORRESPONDING TO NOISY SPEECH, THE PROPOSED BAYESIAN ESTIMATE, AND THOSE OBTAINED USING LONG-TERM NOISE ESTIMATES (LT)

Noise	Noisy	Proposed	LT
Highway	0.0151	0.0096	0.0158
White	0.0295	0.0168	0.0278
Babble	0.0142	0.0118	0.0160
Siren	0.0260	0.0088	0.0291
White-NS	0.0231	0.0155	0.0246

using the memoryless and the memory-based version of the Wiener filter defined in (24). Again, it can be seen that the memory-based estimator achieves a significant reduction in the variance of the error at the same or a lower mean. In the remainder of this section, we consider only the memory based estimator.

D. Evaluation in the STP Parameter Domain

In this section, we compare the performance of the codebook-based Bayesian estimator (with memory) in the short-term predictor parameter domain. We first look at the mean squared error (mse) per dimension between the true and estimated speech LSF coefficients, averaged over ten utterances. For comparison, we present the mse values between the clean and the noisy LSF coefficients, and those corresponding to the LSF coefficients estimated from speech obtained in a subtractive manner from the long-term noise estimate of [9].[2] While computing the mse, frames corresponding to silence were excluded [27]. These results are shown in Table III. It can be seen that the proposed MMSE estimator results in significantly lower mse values compared to those obtained with the noisy speech, and with the long-term noise estimates. In some cases, LT results in worse values than the noisy case. This is explained by the fact that while the subtractive approach improves the SNR, it is not necessarily optimal for the mse for the LSF coefficients. In Table IV, we show the corresponding log-spectral distortion values, without the inclusion of the excitation variances. Values with the excitation variance included are presented in Table V.

[2]An estimate of the power spectrum of clean speech was obtained in a subtractive fashion using the long-term noise estimate according to $\hat{P}_x = \max(P_y - \hat{P}_w^{\mathrm{LT}}, 0)$, where $\hat{P}_w^{\mathrm{LT}}$ is the long-term noise estimate. The autocorrelation was obtained through an inverse Fourier operation, from which the LSFs were computed.

TABLE IV
SD (IN DECIBELS) OF SPEECH SPECTRAL SHAPES, WITHOUT INCLUDING THE EXCITATION VARIANCE, AVERAGED OVER TEN UTTERANCES AT 10-dB INPUT SNR FOR NOISY SPEECH, THE PROPOSED BAYESIAN ESTIMATE, AND USING LONG-TERM NOISE ESTIMATES (LT)

Noise	Noisy	Proposed	LT
Highway	4.7	4.0	4.5
White	6.6	5.3	6.3
Babble	4.3	4.2	4.4
Siren	6.5	3.8	6.6
White-NS	5.5	5.0	5.4

TABLE V
SD (IN DECIBELS) OF SPEECH SPECTRA INCLUDING THE EXCITATION VARIANCE, AVERAGED OVER TEN UTTERANCES AT 10-dB INPUT SNR FOR NOISY SPEECH, THE PROPOSED BAYESIAN ESTIMATE, AND USING LONG-TERM NOISE ESTIMATES (LT)

Noise	Noisy	Proposed	LT
Highway	8.4	5.6	6.4
White	13.9	7.3	10
Babble	7.7	6.1	6.4
Siren	11.8	6.3	9.9
White-NS	11.3	7.6	10.2

E. Comparison With Related Enhancement Systems

Thus far, we have evaluated the performance of the proposed system in the short-term predictor parameter domain. In this section, we evaluate[3] the enhanced speech signal in terms of SNR, SSNR, SD, and PESQ. SSNR is reported to have a better correlation to subjective quality than SNR. Nevertheless, SNR, which evaluates the squared error, is interesting in the study of an MMSE estimator.

Based on the method presented in this paper, the enhanced signal can be obtained in two different ways. The first corresponds to filtering the noisy speech with $H_1(\omega)$ defined in (15). This filter is constructed using the MMSE estimates of the short-term predictor parameters. The second approach to obtain the enhanced signal is to use the filter $H_2(\omega)$ defined by (24). As discussed in Section III-C, using $H_2(\omega)$ results in the optimal MMSE estimate of the clean speech signal given the noisy speech. In our experiments too, $H_2(\omega)$ resulted in slightly better results in terms of the objective measures. Hence, we present results for the enhanced speech obtained using $H_2(\omega)$, with memory.

We also provide comparisons with a Wiener filter (WF) scheme using long-term noise estimates [9], the Ephraim–Malah (EM) short-time spectral amplitude estimator [25] using long-term noise estimates, and the HMM-based MMSE approach as described in [16]. For the EM method, computaion of the *a priori* SNR was performed using the decision directed approach with a smoothing factor of $\beta = 0.98$ [25]. For the HMM-based system, as suggested in [16], the speech model had five states with five mixture components in

[3]To be consistent with the evaluation in Section IV-D, SD was computed using LP coefficients extracted from segments that were Hanning windowed. In [11], a rectangular window was used.

TABLE VI
SNR VALUES (IN DECIBELS) AVERAGED OVER TEN UTTERANCES AT 10-dB INPUT SNR FOR ENHANCED SPEECH OBTAINED USING THE PROPOSED SCHEME, THE HMM METHOD, THE EPHRAIM–MALAH METHOD (EM), AND THE WIENER FILTER USING LONG-TERM NOISE ESTIMATES (WF)

Noise	H_2	HMM	EM	WF
Highway	15.3	13.4	14.1	13.0
White	15.5	15.0	14.4	15.1
Babble	13.2	12.1	13.0	12.5
Siren	17.9	17.8	11.5	11.5
White-NS	15.9	10.8	10.9	10.6

TABLE VII
SSNR VALUES (IN DECIBELS) AVERAGED OVER TEN UTTERANCES AT 10-dB INPUT SNR, FOR THE NOISY SPEECH, AND FOR ENHANCED SPEECH OBTAINED USING THE PROPOSED SCHEME, THE HMM METHOD, THE EPHRAIM–MALAH METHOD (EM), AND THE WIENER FILTER USING LONG-TERM NOISE ESTIMATES (WF)

Noise	Noisy	H_2	HMM	EM	WF
Highway	1.9	8.7	6.9	6.8	5.9
White	0.7	7.8	7.6	6.3	6.3
Babble	1.3	6.5	4.6	4.7	4.3
Siren	0.7	12.2	10.8	2.2	2.0
White-NS	5.2	10.1	6.3	5.4	6.2

each state. For each of the noise types considered here, separate noise HMMs were trained. The noise HMMs had three states with three mixture components in each state as in [16]. The LP orders in the noise HMMs were the same as the LP orders in the noise codebooks. For the two-tone siren noise, a special HMM was trained, with two states and one mixture component in each state. The training data used to train the codebooks was used to train the HMMs as well. Model gain adaptation and noise HMM selection was performed in [16] using data from segments detected as noise-only regions. In our implementation, this was modified to use the more accurate noise estimates provided by [9] on a frame-by-frame basis. The HMM method with this modification provided better results (in terms of SNR and SSNR) than the original HMM approach (results with the original approach for this data set are reported in [13]).

It can be seen from Tables VI–IX that, in general, the proposed scheme performs better than the HMM-based method, the Ephraim–Malah method (EM) and Wiener filtering using long-term noise estimates, especially for the nonstationary noise types. The performance gain is significant in terms of SSNR, SD, and PESQ. For the stationary noise types, e.g., white noise, the methods exhibit similar performance to the reference methods as expected, since long-term noise estimates are accurate in this case.

The performance of the HMM method in siren and highway noise conditions provides a useful insight into its operation. The two-tone siren noise considered here was generated by a nonmoving source and recorded by a stationary listener. Thus, once the nonstationarity of the siren is captured by the two-state HMM during training, it can accurately model the noise. On the other hand, for changing noise types such as highway noise, as discussed in Section I, the HMM method is unable to perform well since its gain adaptation is based on long-term noise estimates. To verify this behavior, the experiment was repeated (using the same siren codebook and HMM) with siren noise modulated by a 0.1-Hz sine wave, to simulate a siren (for e.g., in a vehicle) approaching and leaving the listener. The results are shown in Table X. It can be seen that the proposed method is able to handle the nonstationarity, and performs significantly better than the HMM scheme.

Also interesting is the poor performance of the HMM method for White-NS. The reason for this is that there was no noise HMM trained on White-NS, just as there was no noise codebook trained on White-NS. The white noise codebook was expected to handle this case as well. This was done to show the advantage of treating the spectral shape and the gain independently. With the proposed scheme, it is sufficient to model only the spectral shape of the noise.

TABLE VIII
SD VALUES (IN DECIBELS) AVERAGED OVER TEN UTTERANCES AT 10-dB INPUT SNR FOR THE NOISY SPEECH, AND FOR ENHANCED SPEECH OBTAINED USING THE PROPOSED SCHEME, THE HMM METHOD, THE EPHRAIM–MALAH METHOD (EM), AND THE WIENER FILTER USING LONG-TERM NOISE ESTIMATES (WF)

Noise	Noisy	H_2	HMM	EM	WF
Highway	4.7	3.8	4.9	4.0	4.5
White	6.6	5.4	5.8	5.3	5.9
Babble	4.3	4.2	5.9	4.3	4.6
Siren	6.5	3.7	3.9	6.9	7.0
White-NS	5.5	5.2	5.7	5.2	5.6

TABLE IX
PESQ VALUES AVERAGED OVER TEN UTTERANCES AT 10-dB INPUT SNR FOR THE NOISY SPEECH, AND FOR ENHANCED SPEECH OBTAINED USING THE PROPOSED SCHEME, THE HMM METHOD, THE EPHRAIM–MALAH METHOD (EM), AND THE WIENER FILTER USING LONG-TERM NOISE ESTIMATES (WF)

Noise	Noisy	H_2	HMM	EM	WF
Highway	2.4	3.0	2.6	2.9	2.6
White	2.1	2.9	2.6	2.7	2.5
Babble	2.4	2.7	2.5	2.7	2.5
Siren	2.3	3.3	3.0	2.4	2.3
White-NS	2.3	2.9	2.3	2.4	2.5

TABLE X
SNR, SSNR, SD (ALL IN DECIBELS), AND PESQ SCORES CORRESPONDING TO THE MODULATED SIREN NOISE AT 10-dB INPUT SNR

	Noisy	H_2	HMM
SNR	10	18.2	15.5
SSNR	3.2	13.5	9.5
SD	6.0	3.4	4.6
PESQ	2.4	3.3	2.9

F. Computational Complexity

In comparison to methods such as the Ephraim–Malah scheme and the Wiener filter based on long-term noise estimates, model-based schemes such as the proposed approach and the HMM-based methods suffer from an increase in computational complexity. This is the price to be paid for the improved performance in nonstationary noise environments. The complexity is directly related to the model size, e.g., the number of codebook vectors, or the number of states and mixture components in the HMM. In [13], an iterative scheme to reduce computational complexity resulting from an exhaustive search of the speech and noise codebooks is proposed and can be adopted in the method proposed in this paper as well. It is also relevant to mention that the HMM and codebook approaches lend themselves in a straightforward fashion to parallel processing, which can result in a significant speedup. For example, in principle, one processor can be assigned to compute the likelihood $p(\mathbf{y}|\theta_x, \theta_w, \sigma_x^{2,\mathrm{ML}}, \sigma_w^{2,\mathrm{ML}})$ corresponding to each combination of speech and noise codebook vectors. The amount of time required for the resulting computations is then independent of the model size. A final step of weighted summation then produces the MMSE estimate. While this is an extreme case, in general, a speedup can be obtained with the use of more than one processor, and the resulting computational complexity is determined by the model size and the number of processors.

G. Evaluation of Perceptual Quality

To evaluate the perceptual quality, we compare the proposed scheme to the noise suppression system of the selectable mode vocoder (SMV) [29]. The SMV includes a noise suppression module that operates on the input signal prior to the encoding/decoding process. The SMV noise suppression system (SMV-NS) requires estimates of the background noise and contains mechanisms to update the background noise estimates based on the observed noisy input. It is a frequency domain technique and frequency bins in the noisy spectrum are grouped together to obtain 16 channels. An attenuation factor is determined for each of the 16 channels, which is applied to all the frequency bins in that channel. Details regarding the exact implementation are described in [29].

The SMV-NS system is a perceptually well tuned standardized system, which in informal listening tests clearly outperformed the reference systems considered in the previous section. To make a fair comparison, a well-tuned reference system, not tuned by the authors is best suited. Hence, the choice of SMV-NS for the subjective evaluation. Moreover, since the SMV-NS is perceptually optimized and not optimized for objective measures such as SNR or SD, it gives poor objective results and objective comparisons with the SMV is not fair. Thus, we use the SMV-NS only for subjective tests.

Noisy speech at 10-dB input SNR was processed by the standard SMV and the signal at the output of the decoder was used as the first signal in the evaluation. To generate the second signal, the output of the proposed enhancement system H_2 was processed by the SMV, with its noise suppression module disabled. Thus, the encoding/decoding operation is identical in both systems; they differ only in the noise suppression module.

TABLE XI
SCALE USED TO RATE THE QUALITY OF THE SECOND UTTERANCE RELATIVE TO THAT OF THE FIRST

3	Much better
2	Better
1	Slightly better
0	About the same
-1	Slightly worse
-2	Worse
-3	Much worse

TABLE XII
RESULTS FROM THE LISTENING TEST WITH 95% CONFIDENCE INTERVALS. TEN LISTENERS PARTICIPATED IN THE TEST. POSITIVE VALUES INDICATE A PREFERENCE FOR THE PROPOSED METHOD (SEE TABLE XI)

Noise	Score
Highway	0.3 ± 0.19
White	1.0 ± 0.17
Babble	0 ± 0.17
Siren	2.1 ± 0.16
White-NS	2.1 ± 0.15

To perform a more precise evaluation than an AB preference test, a test similar to the comparison category rating (CCR) [30] was conducted. Listeners were presented with a pair of utterances (one processed by the reference system and the other processed by the proposed system) in each trial. The order of presentation was random. To eliminate any biasing due to the order of the algorithms within a pair, each pair of enhanced utterances was presented twice, with the order switched. Listeners were asked to rate the quality of the second utterance relative to that of the first according to the scale in Table XI.

Ten listeners participated in the test. For each noise type, ten utterances were used. The results from the listening test, together with the 95% confidence intervals are shown in Table XII. It can be seen that for the strongly nonstationary noise types such as siren noise and White-NS, there is a clear preference for the proposed approach. There is also a preference for the white noise case. For highway and babble noise, both systems perform about the same. We note here that the SMV noise suppression system is a perceptually well-tuned system. The proposed MMSE scheme could also benefit from similar perceptual tuning in which case it could be expected to outperform the SMV system for all the noise types.

V. CONCLUSION

In this paper, Bayesian MMSE estimators of the speech and noise short-term predictor parameters were developed using codebooks of linear predictive coefficients to model the prior information. It was shown that the proposed scheme provides superior performance compared to methods that rely on long-term noise estimates, in both stationary and nonstationary environments. Memory-based estimation was seen to significantly reduce both the mean and the variance of the squared error. Memory was found to be useful only for the

noise parameters. Estimation of functions of the short-term predictor parameters was also addressed. From the experiments, it was seen that the proposed MMSE scheme performed significantly better than the HMM-based MMSE scheme, the Ephraim–Malah scheme, and the Wiener filter using long-term noise estimates, in terms of SNR, SSNR, SD, and PESQ. In terms of subjective quality, the proposed scheme was seen to perform better than the standard SMV noise suppression scheme for white noise, siren noise, and nonstationary white noise, while the two systems performed about the same for the other noise types. The use of codebooks results in an increase in computational complexity compared to the Ephraim–Malah scheme or the Wiener filter, which is the price to be paid for the improved performance.

The framework developed in this paper is general and is neither limited to linear predictive coefficients, nor to the codebook structure. Alternate parametric models may be employed, while retaining the proposed estimation framework with instantaneous gain computation. Future work could focus on incorporating the instantaneous gain estimation into methods based on Gaussian mixture models, HMMs, and particle-filter schemes.

APPENDIX

For given θ_x, θ_w and the noisy speech $\mathbf{y}$, we investigate the behavior of $p(\mathbf{y}|\theta)$ as a function of the excitation variances σ_x^2 and σ_w^2. In particular, we are interested in the behavior of the likelihood as a function of the deviation of the excitation variances from their true values, which we approximate by their maximum-likelihood estimates σ_x^2 and σ_w^2 obtained using (6) and (7). We first consider the case where noise is not present. In the absence of background noise, under Gaussianity assumptions, the probability density of the speech samples given the LP parameters can be written as

$$p_{\mathbf{x}}\left(\mathbf{x}|\mathbf{a}_x,\sigma_x^2\right)=\frac{1}{(2\pi)^{N/2}|\mathbf{R}_x|^{1/2}}\exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{R}_x^{-}1\mathbf{x}\right) \tag{28}$$

where $\mathbf{x}=[x(0)x(1)\dots x(N-1)]^T$, $\mathbf{a}_x=[1\,a_{x_1}\,a_{x_2}\dots a_{x_p}]^T$ and $\mathbf{R}_x=\sigma_x^2(\mathbf{A}_x^T\mathbf{A}_x)^{-1}$, where $\mathbf{A}_x$ is the N × N lower triangular Toeplitz matrix with $[1a_{x_1}a_{x_2}\dots a_{x_p}0\dots0]^T$ as the first column. Since the frame length ($N=240$ samples) is large compared to the LP order ($p=10$), the covariance matrix $\mathbf{R}_x$ can be described as circulant and is hence diagonalized by the discrete Fourier transform [31]. We have $\mathbf{R}_x=F^H\sigma_x^2\Lambda_xF$, where $F=[f_{pq}]$ denotes the discrete Fourier transform matrix whose (p,q)th entry is given by $f_{pq}=(1)/(\sqrt{N})\exp((-j2\pi pq)/(N))$, the superscript ^{H} denotes complex conjugate transpose and Λ_x is a diagonal matrix containing the eigenvalues of $(\mathbf{A}_x^T\mathbf{A}_x)^{-1}$. The diagonal entries of $\sigma_x^2\Lambda_x$, the eigenvalue matrix of $\mathbf{R}_x$, correspond to the spectral components of $\mathbf{x}$. The kth diagonal entry of Λ_x is given by $\lambda_x(k)=(1)/(|A_x(k)|^2)$, $A_x(k)=(1)/(\sqrt{N})\sum_{i=0}^{N-1}a_{x_i}\exp((-j2\pi ik)/(N))$, where $a_{x_0}=1$ and $a_{x_i}=0$ for $p+1\le i\le N-1$.

We wish to study the effect of a deviation δ_x in the excitation variance σ_x^2 on $p_{\mathbf{x}}(\mathbf{x}|\mathbf{a}_x,\sigma_x^2)$ as $\mathbf{x}$ and $\mathbf{a}_x$ (and thus Λ_x) remain unchanged. Let $\mathbf{R}_x'=F^H(\sigma_x^2+\delta_x)\Lambda_xF$. We have

$$\begin{aligned}
&p_{\mathbf{x}}\left(\mathbf{x}|\mathbf{a}_x,\sigma_x^2+\delta_x\right)\\
&\quad=\frac{1}{(2\pi)^{N/2}|\mathbf{R}_x'|^{1/2}}\exp\left(-\frac{1}{2}\mathbf{x}^T{\mathbf{R}_x'}^{-1}\mathbf{x}\right)\\
&\quad=\underbrace{\frac{1}{(2\pi)^{N/2}(\sigma_x^2+\delta_x)^{\frac{N}{2}}\prod_{i=1}^{N}\sqrt{\lambda_{x_i}}}}_{A}\\
&\qquad\times\underbrace{\exp\left(-\frac{1}{2}\sum_{i=1}^{N}\frac{|\mathcal{X}_i|^2}{(\sigma_x^2+\delta_x)\,\lambda_{x_i}}\right)}_{B}
\end{aligned} \tag{29}$$

where $\mathcal{X}_i$, $1\le i\le N$ are the discrete Fourier transform coefficients of $\mathbf{x}$ and λ_{x_i} are the diagonal entries of Λ_x. We note that δ_x can take values in the range $[-\sigma_x^2,\infty)$. For positive values of δ_x, as δ_x increases, the denominator grows and the exponential in term B converges to one. Thus, the behavior of the likelihood is dominated by $(\sigma_x^2+\delta_x)^{(-N)/(2)}$. Since N is typically large, this indicates a rapid decay as the deviation δ_x grows. For negative values of δ_x, the exponential term B dominates and an exponential decay of the likelihood occurs.

Considering the case where noise is present, assuming large frames, we can write the covariance matrix of the noisy speech as

$$\mathbf{R}_y=\mathbf{R}_x+\mathbf{R}_w=F^H\left(\sigma_x^2\Lambda_x+\sigma_w^2\Lambda_w\right)F \tag{30}$$

where Λ_w is a diagonal matrix containing the eigenvalues of $(\mathbf{A}_w^T\mathbf{A}_w)^{-1}$. Let $\mathbf{R}_y'=F^H[(\sigma_x^2+\delta_x)\Lambda_x+(\sigma_w^2+\delta_w)\Lambda_w]F$. We have

$$\begin{aligned}
&p_{\mathbf{y}}\left(\mathbf{y}|\mathbf{a}_x,\mathbf{a}_w,\sigma_x^2+\delta_x,\sigma_w^2+\delta_w\right)\\
&\quad=\frac{1}{(2\pi)^{N/2}|\mathbf{R}_y'|^{1/2}}\exp\left(-\frac{1}{2}\mathbf{y}^T{\mathbf{R}_y'}^{-1}\mathbf{y}\right)\\
&\quad=\frac{\exp\left(-\frac{1}{2}\mathbf{y}^TF^H[(\sigma_x^2+\delta_x)\Lambda_x+(\sigma_w^2+\delta_w)\Lambda_w]^{-1}F\mathbf{y}\right)}{(2\pi)^{N/2}\left[\prod_{i=1}^{N}\left(\sigma_x^2+\delta_x\right)\lambda_{x_i}+\left(\sigma_w^2+\delta_w\right)\lambda_{w_i}\right]^{1/2}}\\
&\quad=\frac{\exp\left(-\frac{1}{2}\sum_{i=1}^{N}|\mathcal{Y}_i|^2\left(\sigma_x^2+\delta_x\right)\lambda_{x_i}+\left(\sigma_w^2+\delta_w\right)\lambda_{w_i}\right)}{(2\pi)^{N/2}\left[\prod_{i=1}^{N}\left(\sigma_x^2+\delta_x\right)\lambda_{x_i}+\left(\sigma_w^2+\delta_w\right)\lambda_{w_i}\right]^{1/2}}\\
&\quad=\frac{1}{(2\pi)^{N/2}\prod_{i=1}^{N}\left[\left(\sigma_x^2+\delta_x\right)\lambda_{x_i}+\left(\sigma_w^2+\delta_w\right)\lambda_{w_i}\right]^{1/2}}\\
&\qquad\times\exp\left(-\frac{1}{2}\sum_{i=1}^{N}\frac{|\mathcal{Y}_i|^2}{\left(\sigma_x^2+\delta_x\right)\lambda_{x_i}+\left(\sigma_w^2+\delta_w\right)\lambda_{w_i}}\right)
\end{aligned} \tag{31}$$

where $\mathcal{Y}_i$, $1\le i\le N$ are the discrete Fourier transform coefficients of $\mathbf{y}$ and δ_w,λ_{w_i} are defined analogously to δ_x,λ_{x_i}, respectively. In the case when both δ_x and δ_w are positive or both δ_x and δ_w are negative, the behavior of the likelihood is similar to the speech-only case. For positive values of δ_x and negative values of δ_w (or *vice versa*), we rely on the assumption that the

speech and noise spectral shapes are sufficiently different, i.e., the vectors $[\lambda_{x_1} \dots \lambda_{x_N}]^T$ and $[\lambda_{w_1} \dots \lambda_{w_N}]^T$ are linearly independent so that a positive δ_x cannot compensate a negative δ_w at all frequency indices simultaneously. Thus, the errors add up resulting in a decay of the likelihood.

REFERENCES

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

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

[3] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. Signal Process.*, vol. 39, no. 8, pp. 1732–1742, Aug. 1991.

[4] Y. Ephraim, "A minimum mean square error approach for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1990, vol. 2, pp. 829–832.

[5] ——, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80, no. 10, pp. 1526–1555, Oct. 1992.

[6] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[7] J. Ramirez, J. C. Segura, C. Benitez, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Commun.*, vol. 42, pp. 271–287, Apr. 2004.

[8] D. Malah, R. V. Cox, and A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 1999, vol. 2, pp. 789–792.

[9] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[10] V. Stahl, A. Fischer, and R. Bippus, "Quantile based noise estimation for spectral subtraction and Wiener filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Jun. 2000, vol. 3, pp. 1875–1878.

[11] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based Bayesian speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2005, vol. 1, pp. 1077–1080.

[12] M. Kuropatwinski and W. B. Kleijn, "Estimation of the excitation variances of speech and noise AR-models for enhanced speech coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2001, vol. 1, pp. 669–672.

[13] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 163–176, Jan. 2006.

[14] M. Sugiyama, "Model based voice decomposition method," in *Proc. ICSLP*, Oct. 2000, vol. 4, pp. 684–687.

[15] Y. Zhao, S. Wang, and K. C. Yen, "Recursive estimation of time-varying environments for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2001, vol. 1, pp. 225–228.

[16] H. Sameti, H. Sheikhzadeh, and L. Deng, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 5, pp. 445–455, Sep. 1998.

[17] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Trans. Signal Process.*, vol. 40, no. 4, pp. 725–735, Apr. 1992.

[18] M. Kuropatwinski and W. B. Kleijn, "Minimum mean square error estimation of speech short-term predictor parameters under noisy conditions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2003, vol. 1, pp. 96–99.

[19] K. K. Paliwal and W. B. Kleijn, , W. B. Kleijn and K. K. Paliwal, Eds., "Quantization of LPC parameters," in *Speech Coding and Synthesis*. Amsterdam, The Netherlands: Elsevier Science B.V., 1995, ch. 12, pp. 433–468.

[20] F. Itakura and S. Saito, "A statistical estimation method for speech spectral density and formant frequencies," *Electron. Commun. Jpn.*, vol. 53-A, pp. 36–43, 1970.

[21] R. M. Gray, A. Buzo, A. H. G Jr., and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 367–376, Aug. 1980.

[22] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*. Upper Saddle River, NJ: Prentice-Hall, 2000.

[23] R. M. Gray, *Source Coding Theory*. Boston, MA: Kluwer, 1990.

[24] A. Papoulis and S. U. Pillai, *Probability, Random Variable and Stochastic Processes*, 4th ed. New York: McGraw-Hill, 2002.

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

[26] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, no. 1, pp. 84–95, Jan. 1980.

[27] N. S. Jayant and P. Noll, *Digital coding of waveforms*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[28] *Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for end-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs*, ITU-T Rec. P.862, 2001.

[29] *Selectable Mode Vocoder (SMV) Service Option for Wideband Spread Spectrum Communication Systems, Version 3.0*, 3GPP2 Document C.S0030-0, Jan. 2004.

[30] *Methods for subjective determination of transmission quality Annex E*, ITU-T Rec. P.800, 1996.

[31] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. Englewood Cliffs: Prentice-Hall, 1993.

Sriram Srinivasan (S'04–M'06) received the Ph.D. degree in telecommunications from the Department of Signals, Sensors, and Systems, Royal Institute of Technology (KTH), Stockholm, Sweden, in 2005.

From April to June 2005, he was a Visiting Researcher at the Telecommunications Laboratory, University of Erlangen-Nuremberg, Germany. He is currently working as a Senior Scientist at Philips Research Laboratories, Eindhoven, The Netherlands. His research interests include single and multichannel speech enhancement.

Jonas Samuelsson was born in Vallentuna, Sweden, in 1971. He received the M.Sc. degree in electrical engineering and the Ph.D. degree in information theory, both from Chalmers University of Technology, Gothenburg, Sweden, in 1996 and 2001, respectively.

He held a Senior Researcher position at the Department of Speech, Music, and Hearing, Royal Institute of Technology (KTH), Sweden, from 2002 to 2003. In 2004, he became a Research Associate at the Department of Signals, Sensors, and Systems, KTH. His research interests include signal compression, quantization theory, and speech and audio processing. He is currently working on speech enhancement and source and channel coding for future wireless networks.

W. Bastiaan Kleijn (F'99) received the M.S. degree in electrical engineering from Stanford University, Stanford, CA, the M.S. degree in physics and the Ph.D. degree in soil science, both from the University of California, Riverside, and the Ph.D. degree in electrical engineering from Delft University of Technology, Delft, The Netherlands.

He worked on speech processing at AT&T Bell Laboratories from 1984 to 1996, first in development and later in research. Between 1996 and 1998, he held guest professorships at Delft University of Technology, Vienna University of Technology, Vienna, Austria, and the Royal Institute of Technology (KTH), Stockholm, Sweden. He is now a Professor at KTH and heads the Sound and Image Processing Laboratory in the Department of Signals, Sensors, and Systems. He is also a founder and former Chairman of Global IP Sound AB where he remains Chief Scientist.

Prof. Kleijn is an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS, is on the Editorial Boards of the IEEE *Signal Processing Magazine*, and the *EURASIP Journal of Applied Signal Processing*, and was an Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He has been a member of several IEEE technical committees, and a Technical Chair of ICASSP'99, the 1997 and 1999 IEEE Speech Coding Workshops, and a General Chair of the 1999 IEEE Signal Processing for Multimedia Workshop.