Codebook driven short-term predictor parameter estimation for speech enhancement

Codebook Driven Short-Term Predictor Parameter Estimation for Speech Enhancement

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

***Abstract*—In this paper, we present a new technique for the estimation of short-term linear predictive parameters of speech and noise from noisy data and their subsequent use in waveform enhancement schemes. The method exploits a priori information about speech and noise spectral shapes stored in trained codebooks, parameterized as linear predictive coefficients. The method also uses information about noise statistics estimated from the noisy observation. Maximum-likelihood estimates of the speech and noise short-term predictor parameters are obtained by searching for the combination of codebook entries that optimizes the likelihood. The estimation involves the computation of the excitation variances of the speech and noise auto-regressive models on a frame-by-frame basis, using the a priori information and the noisy observation. The high computational complexity resulting from a full search of the joint speech and noise codebooks is avoided through an iterative optimization procedure. We introduce a classified noise codebook scheme that uses different noise codebooks for different noise types. Experimental results show that the use of a priori information and the calculation of the instantaneous speech and noise excitation variances on a frame-by-frame basis result in good performance in both stationary and nonstationary noise conditions.**

***Index Terms*—Autoregressive models, codebooks, maximum-likelihood, nonstationary noise, short-term predictor, speech enhancement.**

I. Introduction

ENHANCEMENT of speech corrupted by additive background noise is a topic of long standing interest as it has applications in a wide range of areas. Examples of applications include mobile communications in hostile environments, hands-free telephony and speech recognition. Speech enhancement is also often considered as a useful preprocessing step to improve the performance of speech coders. We focus on single-channel speech enhancement systems where the only available input is the noisy speech. This is in contrast to multiple-channel systems that have more than one microphone and thus may provide additional information about the noise statistics. Single-channel systems are relevant especially in mobile communications where it is often not feasible to have multiple microphones due to cost and size factors.

Manuscript received March 11, 2004; revised October 4, 2004. This work was supported in part by the European Commission under ANITA Project IST-2001-34327. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Maurizio Omologo.

The authors are with the Sound and Image Processing Laboratory, Department of Signals, Sensors, and Systems, KTH Royal Institute of Technology, Stockholm 100 44, Sweden (e-mail: sriram.srinivasan@s3.kth.se; jonas.samuelsson@s3.kth.se; bastiaan.kleijn@s3.kth.se).

Digital Object Identifier 10.1109/TSA.2005.854113

Numerous single-channel noise suppression techniques such as Wiener filtering [1], subtractive type methods [2], [3], subspace based methods [4]–[6] and Kalman filtering methods [7] have been developed. A common feature of these systems is that they require some form of noise estimation to obtain information about noise statistics from the noisy observation. These estimation techniques include voice activity detection, estimation from initial silence segments, and, more recently, methods based on quantiles [8] and minimum statistics [9]. While the recent noise estimation techniques are designed to perform well even in nonstationary noise environments, performance still degrades with increasing nonstationarity. Methods such as those of [9] employ a buffer of past samples and the buffer is typically of the order of a few hundred milliseconds. A long-term estimate of the noise power spectrum is produced based on this buffer, and thus, performance in quickly varying noise conditions is limited by the buffer length. The dependence on the buffer arises due to the fact that there is no *a priori* information about the noise. In this paper, we provide a framework to address this fundamental limitation by using instantaneous estimates of speech and noise power spectra, estimated for each segment (typically 20–30 ms long), using *a priori* information about both speech and noise. We focus on the estimation of the short-term predictor (STP) parameters of speech and noise.

The STP parameters consist of the linear predictive (LP) coefficients and the excitation variance, which is the variance of the prediction error. These parameters can be used in different speech enhancement schemes. For example, the Kalman filtering approach [7] uses the linear predictive coefficients of speech and noise to form the state-space model. Other enhancement schemes that use AR spectra include [10]–[12]. The physiology of speech production constrains the speech LP coefficients to lie within a subset of all possible values, which can be modeled based on long sequences of training data. Such *a priori* information about the LP coefficients of speech has been exploited successfully in speech coding using trained codebooks [13]. The usefulness of *a priori* information about noise LP coefficients in speech enhancement has been demonstrated in hidden Markov model (HMM) based applications [14] and in codebook-based estimation [15], [16]. Such *a priori* information can be collected from a wide range of commonly occurring noise sources. Here we use trained codebooks of speech and noise LP coefficients for speech enhancement. In addition to the *a priori* information, in this paper we also use the long-term noise power spectral estimates obtained from the noisy observation to provide a safety-net for noise types that are not represented by the model.

Speech enhancement techniques that rely on *a priori* information about the power spectrum include recursive EM [17], codebook constrained Wiener filtering [10] and HMM based systems [11], [14]. The method of [17] does not use *a priori* information about noise and is unable to handle nonstationary colored noise. The methods in [10], [11], [14] depend on long-term noise estimates (or noise estimates from speech pauses) to obtain the speech variance or to perform model gain adaptation and hence perform poorly in nonstationary noise. These methods are considered in greater detail in Section III.

Estimation of the excitation variances of the speech and noise AR models on a frame-by-frame basis can play an important role in improved speech enhancement in nonstationary noise conditions. This dependence can be avoided if we have *a priori* information about speech and noise. A solution based on this principle was presented in [15]. The method uses two codebooks, one each for speech and noise AR spectral envelopes. For a given noisy frame of speech, for each pair of speech and noise entries from the respective codebooks, the excitation variances and a likelihood score are computed. The score captures the likelihood that the observed noisy frame is generated by a given pair of speech and noise spectral shapes, together with their variances. The pair of codebook entries and the associated excitation variances that globally maximize the likelihood score can then be used in an enhancement technique such as Kalman or Wiener filtering. Since the speech and noise excitation variances are estimated every frame, the method can deal with quickly varying noise types. A schematic diagram of this method is shown in Fig. 1. A similar approach was proposed in [18] in a voice decomposition context using Euclidean spectral distance to match the observed spectrum to the model spectrum (provided by the codebooks) and is discussed in more detail in Section III-B.

In this paper, we build upon the maximum-likelihood (ML) approach of [15] and provide a closed-form expression for the ML estimate of the excitation variances under certain assumptions. We present an iterative scheme that eliminates the need to search through a joint codebook. The method uses both estimated noise information and *a priori* information to reduce complexity and to achieve better performance. An interpolative search technique further improves performance by reducing the errors due to the limited precision of the codebooks. We also propose a classified noise codebook scheme, where we have multiple small noise codebooks, each trained for a particular noise type. For each segment of noisy speech, a classification is made and a particular noise codebook is selected, which is then used in the joint search with the speech codebook. We show that ML estimation of the speech and noise STP parameters is equivalent to minimizing the Itakura–Saito distortion between the observed and modeled noisy spectra. We also show experimentally that the ML estimation provides superior performance compared to the spectral matching using Euclidean spectral distance [18]. We provide experimental results to show that the proposed method using instantaneous excitation variance estimation performs better than the methods of [10], [14] that rely on long-term noise estimates (or estimates obtained from speech pauses) and the method of [17] that does not use *a priori* information about noise.

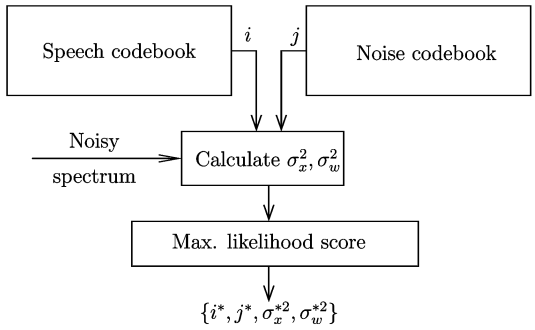

Fig. 1. Estimation of excitation variances and spectral shapes: i^*, j^* are the indexes of the selected entries from the speech and noise codebooks and σ_x^{*2}, σ_w^{*2} are the corresponding excitation variances.

The remainder of this paper is organized as follows. In Section II, we describe the ML estimation of the speech and noise short-term predictor parameters. In Section III, we compare the proposed estimation to related methods to highlight the differences. Implementation aspects are considered in Section IV where an iterative technique to reduce computational complexity, a classified noise codebook scheme and an interpolative search technique are described. Experiments and results are discussed in Section V and finally, conclusions are presented in Section VI.

II. Maximum Likelihood Estimation of Short-Term Predictor Parameters

In this section, we consider the problem of estimating both the speech and noise AR coefficients and the corresponding excitation variances, using the observed noisy spectrum and the *a priori* information contained in the speech and noise codebooks. A unified ML framework for obtaining the optimal speech and noise codebook combinations and computing the corresponding excitation variances is developed. 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 noisy speech, clean speech and noise respectively and are considered to be random processes. In the absence of background noise, under Gaussianity assumptions, the probability density of the speech samples given the LP parameters can be written as

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

where $\mathbf{x} = [x(0)x(1)\ldots x(K-1)]^T$, $\mathbf{a}_x = [1\ a_{x_1}\ldots a_{x_p}]^T$ is the vector of the AR coefficients of speech, K is the number of samples in a frame and $\mathbf{R}_x = \sigma_x^2(\mathbf{A}_x^T\mathbf{A}_x)^{-1}$, where $\mathbf{A}_x$ is the $K \times K$ lower triangular Toeplitz matrix with $[1\ a_{x_1}a_{x_2}\ldots a_{x_p}0\ldots 0]^T$ as the first column and σ_x^2 is the excitation variance of the speech AR model. The codebook index corresponding to the LP vector that maximizes the likelihood can be written as $i^* = \arg\max_i \max_{\sigma_x^2} p_{\mathbf{x}}(\mathbf{x}|\mathbf{a}_x^i; \sigma_x^2)$, where $\mathbf{a}_x^i$ is the i^{th} vector from the speech codebook.

When there is background acoustic noise, then in the absence of any speech enhancement system, the index is simply chosen as $\arg\max_i \max_{\sigma_x^2} p_{\mathbf{x}}(\mathbf{y}|\mathbf{a}_x^i; \sigma_x^2)$. If an estimate of the noise LP

vector $\mathbf{a}_w = [1\, a_{w_1} \ldots a_{w_q}]^T$ is available, then the ML estimate of the clean speech LP vector given the noisy observation and the estimated noise LP vector can be written as

$$i^* = \arg\max_i \max_{\sigma_x^2,\sigma_w^2} p_{\mathbf{y}}(\mathbf{y}|\mathbf{a}_x^i, \mathbf{a}_w; \sigma_x^2, \sigma_w^2) \tag{3}$$

where the dependence on the excitation variances is explicitly shown. $p_{\mathbf{y}}(\mathbf{y}|\mathbf{a}_x, \mathbf{a}_w; \sigma_x^2, \sigma_w^2)$ is zero-mean Gaussian with its covariance matrix given by $R_x + R_w$, where the noise covariance matrix R_w is defined analogous to R_x. Generalizing to the case where we have a noise codebook instead of a single noise estimate, the ML estimate of the speech and noise codebook entries can be written as

$$\{i^*, j^*\} = \arg\max_{i,j} \max_{\sigma_x^2,\sigma_w^2} p_{\mathbf{y}}(\mathbf{y}|\mathbf{a}_x^i, \mathbf{a}_w^j; \sigma_x^2, \sigma_w^2). \tag{4}$$

If we let the frame length approach infinity, the covariance matrices can be described as circulant and are diagonalized by the Fourier transform. The logarithm of the likelihood in (4) can then be written as [19]

$$\begin{aligned} L &= \ln\left(p_{\mathbf{y}}(\mathbf{y}|\mathbf{a}_x^i, \mathbf{a}_w^j; \sigma_x^2, \sigma_w^2)\right) \\ &= \int_0^{2\pi} -\frac{P_y(\omega)}{\frac{\sigma_x^2}{|A_x^i(\omega)|^2} + \frac{\sigma_w^2}{|A_w^j(\omega)|^2}} + \ln\left(\frac{1}{\frac{\sigma_x^2}{|A_x^i(\omega)|^2} + \frac{\sigma_w^2}{|A_w^j(\omega)|^2}}\right) d\omega \end{aligned} \tag{5}$$

where $A_x^i(\omega)$ is the spectrum of the i^{th} vector from the speech codebook and $A_w^j(\omega)$ is the spectrum of the j^{th} vector from the noise codebook and are given by

$$A_x^i(\omega) = \sum_{k=0}^{p} a_{x_k}^i e^{-j\omega k}, \quad A_w^j(\omega) = \sum_{k=0}^{q} a_{w_k}^j e^{-j\omega k}. \tag{6}$$

Combining (4) and (5), we have (7), as shown at the bottom of the page, where $d_{\mathrm{IS}}(P_y(\omega), \hat{P}_y(\omega))$ is the Itakura–Saito distortion between $P_y(\omega)$ and $\hat{P}_y(\omega)$ given by [20]

$$\begin{aligned} &d_{\mathrm{IS}}(P_y(\omega), \hat{P}_y(\omega)) \\ &\quad = \frac{1}{2\pi}\int_0^{2\pi} \left(\frac{P_y(\omega)}{\hat{P}_y(\omega)} - \ln\left(\frac{P_y(\omega)}{\hat{P}_y(\omega)}\right) - 1\right) d\omega. \end{aligned} \tag{8}$$

Let $\hat{P}_y^{ij}(\omega) = \sigma_x^2/|A_x^i(\omega)|^2 + \sigma_w^2/|A_w^j(\omega)|^2$. To complete the estimation, the excitation variances that minimize $d_{\mathrm{IS}}(P_y(\omega), \hat{P}_y^{ij}(\omega))$ need to be determined. Assuming that the modeling error between $P_y(\omega)$ and $\hat{P}_y^{ij}(\omega)$ is small, using a series expansion for $\ln(x)$ up to second order terms, it can be shown that [20]

$$d_{\mathrm{IS}}(P_y(\omega), \hat{P}_y^{ij}(\omega)) \approx \frac{1}{2} d_{\mathrm{LS}}(P_y(\omega), \hat{P}_y^{ij}(\omega)) \tag{9}$$

where $d_{\mathrm{LS}}(P_y(\omega), \hat{P}_y^{ij}(\omega))$ is the log-spectral distortion between the observed noisy spectrum and the noisy spectrum obtained from the model

$$\begin{aligned} &d_{\mathrm{LS}}(P_y(\omega), \hat{P}_y^{ij}(\omega)) \\ &= \frac{1}{2\pi}\int_0^{2\pi} \left|\ln\left(\frac{\sigma_x^2}{|A_x^i(\omega)|^2} + \frac{\sigma_w^2}{|A_w^j(\omega)|^2}\right) - \ln(P_y(\omega))\right|^2 d\omega. \end{aligned} \tag{10}$$

Given $A_x^i(\omega)$ and $A_w^j(\omega)$, the corresponding optimal excitation variances can be determined by differentiating (10) with respect to σ_x^2 and σ_w^2, setting the result to zero and solving the resulting set of simultaneous equations. First we simplify (10) to ensure that the resulting equations are linear

$$\begin{aligned} &d_{\mathrm{LS}}(P_y(\omega), \hat{P}_y^{ij}(\omega)) \\ &\quad = \frac{1}{2\pi}\int_0^{2\pi} \left|\ln\left(\frac{\frac{\sigma_x^2}{|A_x^i(\omega)|^2} + \frac{\sigma_w^2}{|A_w^j(\omega)|^2}}{P_y(\omega)}\right)\right|^2 d\omega \\ &\quad \approx \frac{1}{2\pi}\int_0^{2\pi} \left|\frac{\frac{\sigma_x^2}{|A_x^i(\omega)|^2} + \frac{\sigma_w^2}{|A_w^j(\omega)|^2} - P_y(\omega)}{P_y(\omega)}\right|^2 d\omega \end{aligned} \tag{11}$$

where we used the approximation $\ln(1+x) \approx x$, for small x, i.e., small modeling errors. This approximation can be made valid by using the spectral envelope of the observed noisy speech instead of the periodogram. Partial differentiation with respect to σ_x^2 and σ_w^2 yields

$$\begin{aligned} &\int_0^{2\pi} \frac{\sigma_x^2|A_w^j|^2 + \sigma_w^2|A_x^i|^2 - P_y|A_x^i|^2|A_w^j|^2}{P_y|A_x^i|^2|A_w^j|^2}\left(\frac{1}{P_y|A_x^i|^2}\right) d\omega = 0, \\ &\int_0^{2\pi} \frac{\sigma_x^2|A_w^j|^2 + \sigma_w^2|A_x^i|^2 - P_y|A_x^i|^2|A_w^j|^2}{P_y|A_x^i|^2|A_w^j|^2}\left(\frac{1}{P_y|A_w^j|^2}\right) d\omega = 0 \end{aligned}$$

where the dependency on ω has not been shown, to facilitate notation. The resulting solution can be written as

$$\mathbf{C}\begin{bmatrix}\sigma_x^2 \\ \sigma_w^2\end{bmatrix} = \mathbf{D} \tag{12}$$

$$\begin{aligned} \{i^*, j^*\} &= \arg\max_{i,j}\left\{\max_{\sigma_x^2,\sigma_w^2} \int_0^{2\pi} -\frac{P_y(\omega)}{\frac{\sigma_x^2}{|A_x^i(\omega)|^2} + \frac{\sigma_w^2}{|A_w^j(\omega)|^2}} + \ln\left(\frac{1}{\frac{\sigma_x^2}{|A_x^i(\omega)|^2} + \frac{\sigma_w^2}{|A_w^j(\omega)|^2}}\right) d\omega\right\} \\ &= \arg\max_{i,j}\left\{\max_{\sigma_x^2,\sigma_w^2} \int_0^{2\pi} -\frac{P_y(\omega)}{\frac{\sigma_x^2}{|A_x^i(\omega)|^2} + \frac{\sigma_w^2}{|A_w^j(\omega)|^2}} + \ln\left(\frac{P_y(\omega)}{\frac{\sigma_x^2}{|A_x^i(\omega)|^2} + \frac{\sigma_w^2}{|A_w^j(\omega)|^2}}\right) - \ln(P_y(\omega))\, d\omega\right\} \\ &= \arg\min_{i,j}\left\{\min_{\sigma_x^2,\sigma_w^2} d_{\mathrm{IS}}\left(P_y(\omega), \frac{\sigma_x^2}{|A_x^i(\omega)|^2} + \frac{\sigma_w^2}{|A_w^j(\omega)|^2}\right)\right\} \end{aligned} \tag{7}$$

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

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

$$\mathbf{D} = \begin{bmatrix} \left\| \frac{1}{P_y(\omega)|A_x^i(\omega)|^2} \right\| \\ \left\| \frac{1}{P_y(\omega)|A_w^j(\omega)|^2} \right\| \end{bmatrix}$$

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

The estimation procedure can be summarized as follows. For each pair of speech and noise spectral shapes from the respective codebooks, the excitation variances are calculated according to (12) and the distortion $d_{\text{IS}}(P_y(\omega), \hat{P}_y^{ij}(\omega))$ is evaluated. Codebook combinations that result in a negative value for either the speech or noise excitation variance are discarded since they are infeasible due to the nonnegativity constraints on the variances. The speech and noise power spectra globally minimizing the distortion measure are determined. These power spectra together with the corresponding excitation variances represent the ML estimate of the speech and noise short-term predictor parameters. We note that for each frame of noisy speech, the noise codebook is augmented with the long-term estimate of the noise LP vector obtained from the observation, using [9] for example. One of the trained noise codebook entries is chosen if it provides a better representation of the underlying noise than the long-term estimate, which is chosen otherwise. The ML estimation algorithm is described in Table I.

The proposed estimation scheme can handle both quickly changing noise envelopes and quickly changing energy. The different entries of the noise codebook deal with changing noise envelopes while the instantaneous estimation of the excitation variances handles changing noise energy. If the nonstationarity is only due to changing energy, then a single-entry noise 'codebook' consisting only of the long-term noise estimate, together with the instantaneous energy estimation can provide good performance. In this case, the proposed estimation scheme uses only the gain-normalized spectral envelope of noise from the long-term estimate and computes the gain through an independent optimization. In general, since the spectral envelope of the noise can also vary, the use of a noise codebook can be expected to further improve performance. This is indeed observed in the experiments. We also note that the method handles inaccuracies in the speech model during speech pauses by assigning a zero excitation variance to speech if necessary and concentrating the energy in the noise model.

The estimated STP parameters of speech and noise can be used in several applications. In this paper, we focus on noise reduction for waveform enhancement. If $\{i^*, j^*, \sigma_x^{2*}, \sigma_w^{2*}\}$ represent the optimal codebook entries and excitation variances, then the corresponding STP parameters can be used to construct a Wiener filter for speech waveform enhancement

$$H(\omega) = \frac{\frac{\sigma_x^{2*}}{|A_x^{i*}(\omega)|^2}}{\frac{\sigma_x^{2*}}{|A_x^{i*}(\omega)|^2} + \frac{\sigma_w^{2*}}{|A_w^{j*}(\omega)|^2}}. \tag{13}$$

TABLE I
ALGORITHM FOR ML ESTIMATION OF SPEECH AND NOISE STP PARAMETERS USING *A-PRIORI* INFORMATION. N_x, N_w ARE THE SPEECH AND NOISE CODEBOOK SIZES, RESPECTIVELY

```
For each frame of noisy speech
D = ∞
for i = 1 : N_x
  for j = 1 : N_w
    {σ_x^{i 2}, σ_w^{j 2}} = arg min_{σ_x^2, σ_w^2} d_IS(P_y(ω), P̂_y^{ij}(ω))
    if σ_x^{i 2}, σ_w^{j 2} ≥ 0
      if d_IS(P_y(ω), P̂_y^{ij}(ω)) < D
        {i*, j*, σ_x^{2*}, σ_w^{2*}} = {i, j, σ_x^{i 2}, σ_w^{j 2}}
        D = d_IS(P_y(ω), P̂_y^{ij}(ω))
      endif
    endif
  end
end
```

The STP parameter estimates can also be used in a Kalman filter. Other applications where the estimated speech parameters can be used include LP based speech coding and speech recognition.

III. DISCUSSION

In this section, we compare the proposed ML estimation scheme to other enhancement schemes that employ *a priori* information. The codebook constrained Wiener filter (CCWF) approach of [10] is discussed in Section III-A. The spectral matching perspective and the method of [18] are discussed in Section III-B. This is followed by a discussion on the gain adaptation employed by HMM based methods to highlight the differences from the proposed instantaneous gain estimation. We conclude this section with a discussion on the recursive EM approach described in [17]. The performance of CCWF, the HMM based method and the recursive EM approach is compared to that of the proposed method through experiments in Section V-J. Results using the method of [18] are provided in Section V-E.

A. Codebook Constrained Wiener Filtering

In codebook constrained Wiener filtering [10], a trained codebook of speech LP coefficients is used to provide intra-frame constraints in an iterative Wiener filter framework. At each iteration an estimate of the vector of clean speech LP coefficients is obtained from the current estimate of clean speech and is replaced by the closest (with respect to a selected distortion measure) entry from the codebook. The LP vector from the codebook is used to construct a Wiener filter to obtain the next estimate of clean speech. The iterations continue until convergence, which is said to occur when the same codebook entry is selected in consecutive iterations. In [10], the Itakura–Saito distortion measure was used and it was shown that the method converges typically within three iterations. However the method uses long-term noise spectral estimates to obtain the initial estimate of clean speech. Also, the excitation variance of clean speech is obtained in a subtractive manner using the noisy variance and the estimated noise variance. Consequently, while the method performs well in stationary noise, performance degrades with increasing nonstationarity. This is confirmed by the experimental results in Section V-J.

TABLE II
COMPARISON OF THE DIFFERENT ESTIMATION APPROACHES

Method	Distortion measure used to select codebook entries	Distortion measure used for variance calculation	Small error assumption
ML	Itakura-Saito	Itakura-Saito (equiv. Log-spectral distortion)	Yes
LS	Log-spectral distortion	Log-spectral distortion (equiv. Itakura-Saito)	Yes
LL [15]	Itakura-Saito	Modified spectral distance	No
ES [18]	Euclidean spectral distance	Euclidean spectral distance	No

B. Spectral Matching Perspective

In Section II, we developed the ML estimator of the speech and noise short-term predictor parameters. It can be seen from (7) that obtaining the ML estimate involves minimizing the Itakura–Saito distortion between the observed noisy spectrum $P_y(\omega)$ and the modeled noisy spectrum $\hat{P}_y^{ij}(\omega)$ for all codebook combinations. This leads to a spectral matching perspective, i.e., the short-term predictor parameters can be obtained by minimizing a spectrally based distortion measure between the observed and modeled noisy spectra. As an alternative to the Itakura–Saito distortion, we can use log-spectral distortion. The variance estimation remains identical if we employ the assumption of small errors while the distortion measure is given by (10). Note that we use the small error assumption only for evaluating the excitation variances and not while computing the distortion measure to select the best codebook combination. The approach of Sugiyama in [18] minimizes Euclidean spectral distance for voice decomposition using AR VQs as voice models. The minimization is actually performed in the autocorrelation domain. Both the correlation distance and LPC cepstral distance were considered and the former was found to result in better estimates. Since the Fourier transform is unitary and hence preserves distance, minimizing the correlation distance is equivalent to minimizing Euclidean spectral distance. It is shown later in the experiments in Section V-E that the ML approach (Itakura–Saito distortion) and the log-spectral distortion based estimators exhibit noticeably superior performance compared to the method of [18]. The method described in [15] also optimizes the log-likelihood in the codebook search. The variance estimation however minimizes a modified Euclidean spectral distance $||(P_y(\omega)|A_x(\omega)|^2|A_w(\omega)|^2 - \sigma_x^2|A_w(\omega)|^2 - \sigma_w^2|A_x(\omega)|^2)^2||$, resulting in a mismatch between the criterion used in the variance estimation and the criterion (log-likelihood) used in the global search to obtain the speech and noise codebook entries. By using the small error assumption, the unified ML approach presented in this paper avoids this mismatch. Table II compares the different alternatives from the spectral matching perspective.

C. Comparison to Gain-Adapted HMM Schemes

HMM based speech enhancement methods obtain an estimate of the clean speech signal using Gaussian AR HMMs for both the clean signal and the noise. As mentioned earlier, both the LP coefficients and the excitation variance (gain) are considered as *a priori* information [21]. This naturally leads to a mismatch in the gain term during training and testing. Thus, some form of gain adaptation is essential. For the MAP estimation described in [21], gain-normalized HMMs are trained for the clean speech signal. Let $\lambda = (\lambda_x, \lambda_w)$, where λ_x denotes the parameter set for the gain-normalized HMM for the clean signal and λ_w denotes the parameter set for the noise HMM. At time instant t, gain-adapted MAP signal estimation is then performed according to

$$\hat{\mathbf{x}}_t = \max_{\mathbf{x}_t} \max_{g_t>0} p_\lambda(\mathbf{x}_t, \mathbf{y}_0^t | g_0^t) \tag{14}$$

where $\mathbf{x}_t$ is the vector of clean speech samples at time t (corresponding to a single frame), $\mathbf{y}_0^t$ is the sequence of vectors of noisy samples up to time t, g_0^t is the gain contour of the speech model and $p_\lambda(\mathbf{x}_t, \mathbf{y}_0^t | g_0^t)$ is the joint pdf of $\mathbf{x}_t$ and $\mathbf{y}_0^t$, given the gain contour g_0^t and the complete parameter set λ. It is important to note that g_t is optimized based on the noisy observation and the parameter set of the noise model. Since the noise model is obtained either during speech pauses or from long-term estimates [21], [14], this form of gain adaptation still suffers from poor performance in nonstationary noise. A similar gain adaptation is performed for the MMSE estimator, whose performance is compared to the proposed instantaneous estimation scheme in Section V-J.

D. Comparison to Recursive EM

A recursive procedure for estimating channel distortion and additive noise statistics for speech recognition is described in [17]. The method is based on the batch EM algorithm described in [22] and is modified to obtain parameter estimates on a frame-by-frame basis to operate in nonstationary environments. The pdf of the clean speech spectral vector is modeled by a Gaussian mixture model (GMM) with M components having zero mean and diagonal covariance matrices. The clean speech GMM parameters are obtained using a training database. The noise parameters consist of the correlation coefficients up to lag q, from which the noise spectrum is obtained. No *a priori* information about noise is used. A frequency domain objective function is formulated and the noise parameters are identified as those that optimize the objective function. If $\lambda^{(n)}$ denotes the vector of noise correlation coefficients for the n^{th} frame and $\Lambda^{(n)} = (\lambda^{(1)}, \ldots, \lambda^{(n)})$, the recursive update can be written as

$$\lambda^{(n+1)} = \lambda^{(n)} + \epsilon \hat{I}(\lambda^{(n)}, \Lambda^{(n)})^{-1} \hat{S}(\lambda^{(n)}, \Lambda^{(n)}) \tag{15}$$

where ϵ is the step size, $\hat{S}(\lambda^{(n)}, \Lambda^{(n)})$ and $\hat{I}(\lambda^{(n)}, \Lambda^{(n)})$ are estimates of the first and second derivatives of the objective function $Q_n(\lambda^{(n)}, \Lambda^{(n)})$, and are obtained recursively using estimates from the previous and the current frame.

For white noise, λ is a scalar (the variance) and only one unknown parameter needs to be estimated. In this case, the recursive update tracks nonstationarities in the noise level well. However as the number of unknown parameters increases (for colored noise), the recursion is unable to track changes on a frame-by-frame basis. The batch EM algorithm [22] uses the entire data to obtain noise estimates and was shown to track noise generated by an AR(2) process. In the recursive version where instantaneous estimates are produced, performance is observed to be poor for colored noise (cf. Section V-J). The method proposed in this paper overcomes this drawback by using *a priori* information about noise. In particular, for nonstationary colored noise such as the two-tone siren noise and the highway noise considered in the experiments (Section V-J), the recursive EM method is unable to track the rapid changes in the spectral shape of noise. This is a clear example where using *a priori* information about noise provides significant gains in performance.

As with other methods that incorporate the variances into the trained model, a drawback with the recursive EM technique is the need for gain adaptation to compensate for the mismatch in the gain during training and testing. We emphasize again that gain adaptation is required regardless of whether the noise is stationary or not. Even for stationary noise, gain adaption is required to match the gain in the trained speech model to the test case. In the HMM based methods, it was possible to adapt the speech model gain based on the noise estimate and the noisy observation. In the recursive estimation approach of [17], such a scheme is not possible since no independent noise estimates are available. By separating the gain term from the codebooks, we avoid the need for such adaptation in our method.

IV. Implementation Aspects

In this section, we discuss modifications to the codebook-based ML estimation scheme that aim to reduce computational complexity and improve performance. We first present an iterative parameter estimation scheme in Section IV-A that addresses the high computational complexity of the joint speech-noise codebook search. This is followed by a description of a classified noise codebook scheme in Section IV-B. Finally, an interpolation scheme that addresses the limited precision of the speech codebook is presented in Section IV-C.

A. Iterative Parameter Estimation

In the first step, the long-term estimate of the spectral shape of noise is obtained for example using the minimum statistics approach [9]. This estimate is used to search through the speech codebook to obtain the ML estimate of the speech spectral shape. The search involves calculating the excitation variances according to (12) and evaluating (8) for each speech codebook entry. The optimal speech spectral shape that results from this search is now used to find the best noise spectral shape from the noise codebook.

The iterative procedure of alternately finding the optimal speech and noise codebook entries and the associated variances continues until convergence. Convergence is said to occur when there is no improvement in the distortion measure used in the search. It may happen at the first step of the iteration that the long-term estimate obtained from the observation is better than all entries in the noise codebook. In this case, the iterative procedure is terminated immediately. Each iteration does not increase the value of the chosen distortion measure. This, together with the fact that the codebooks are of finite size, guarantees convergence. The complete algorithm is presented in Table III.

TABLE III
Iterative Search Algorithm

For each frame of noisy speech
$i = 0$
$a_w^{(i)}$=Estimated noise spectral shape
repeat
  $i := i + 1$
  Search speech CB with $a_w^{(i-1)}$ to obtain $a_x^{(i)}, \sigma_x^{2(i)}, \sigma_w^{2(i)}$
  Search noise CB with $a_x^{(i)}$ to obtain $a_w^{(i)}, \sigma_x^{2(i)}, \sigma_w^{2(i)}$
until convergence

A large reduction in complexity results from not having a joint search through both codebooks. Consequently, it is possible to increase the size of the codebooks to provide better signal representation. We note that the iterative scheme may converge to a local optimum. This can be overcome to some extent by selecting at each stage the $N^{(\mathrm{s})}$ best entries from the speech (noise) codebooks instead of just one. We refer to these $N^{(\mathrm{s})}$ best entries as a subset.

It is possible that the iterative procedure continues until the search complexity (number of codebook combinations considered) is equivalent to that of a full search of the joint codebook. However, in practice, convergence was found to occur quite early in the iterative procedure. Often, convergence occurred within a single iteration. This is discussed in more detail in Section V-F. To see the reduction in complexity due to the iterative method, let N_x, N_w denote the number of entries in the speech and noise codebooks respectively, let $N_x^{(\mathrm{s})}$, $N_w^{(\mathrm{s})}$ be the cardinality of the speech and noise subsets, and let P denote the number of iterations. The noniterative technique requires searching through $N_x N_w$ combinations of codebook entries. The number of searches for the iterative method is $P(N_x^{(\mathrm{s})} N_w + N_x N_w^{(\mathrm{s})}) \approx P N_x N_w^{(\mathrm{s})}$ since the speech codebook is generally larger than the noise codebook. Thus, the new method provides a reduction in complexity by a factor $N_w / P N_w^{(\mathrm{s})}$.

B. Classified Noise Codebooks

The use of a noise codebook and instantaneous estimation of speech and noise excitation variances provides good performance in highly nonstationary noise conditions [16]. Choosing an appropriate noise codebook size is critical. If the noise codebook is too small, it may not result in an accurate description of the observed noise. On the other hand, with increasing noise codebook size and multiple noise sources, we obtain a very general description of the noise parameter space, which nullifies the advantage of a restricted parameter space provided by the *a priori* information. A similar behavior has also been observed in the HMM based method described in [14], where the authors observe that training a single large noise HMM with various noise types not only increases computational complexity but also degrades performance by introducing more sources of error.

In our context, for a sufficiently large noise codebook trained on various noise sources, it is possible that several pairs of vectors from the speech and noise codebooks provide a good fit to the observed noisy spectrum resulting in ambiguity. In such a situation, the speech and noise codebook entries that maximize the likelihood score may no longer be the speech and noise codebook entries that represent the underlying speech and noise data. This is related to the uniqueness of the solution [15].

To address these issues, we propose a classified noise codebook scheme, where we have multiple small noise codebooks, each trained for a particular noise type. We first obtain a long-term estimate of the noise spectrum using the minimum statistics approach [9], which corresponds to an estimate obtained from multiple past frames. We denote this long-term estimate as $\bar{A}_w(\omega)$. For each segment of noisy speech, a classification is made using this long-term estimate and a particular noise codebook is chosen. The selected noise codebook is then used in the subsequent ML search. Thus, the parameter estimation can be viewed as a two-step process. In the first step, a single noise codebook is selected from a set of noise codebooks. The long-term estimate $\bar{A}_w(\omega)$ is used to select a particular codebook. The speech codebook does not figure in this step. The second step corresponds to the regular codebook search outlined in Section II using the speech codebook and the selected noise codebook. We note that the selected noise codebook is augmented with the long-term estimate $\bar{A}_w(\omega)$ obtained from the noisy observation using [9] to provide robustness to noise sources not adequately represented in the pre-trained codebooks.

To perform the classification, we consider each noise codebook as a Gaussian mixture model, with equal weights for all the mixture components. The mixture (codebook) that results in the highest likelihood for the current observation frame is chosen as the codebook for the current segment. The resulting ML classifier can be written as

$$n^* = \arg\max_n \frac{1}{M_n} \sum_{m=1}^{M_n} p(\mathbf{w}|\mathbf{a}_w^{n,\,m}), \quad 1 \le n \le M \tag{16}$$

where $\mathbf{w}$ is the vector of noise samples, $\mathbf{a}_w^{n,\,m}$ is the m^{th} vector in the n^{th} noise codebook, M_n is the size of the n^{th} codebook and M is the number of noise codebooks. To obtain n^*, we use the equivalence of the log-likelihood and the Itakura–Saito measure so that (16) can be equivalently written as

$$n^* = \arg\min_n \frac{1}{M_n} \sum_{m=1}^{M_n} \exp\left(d_{\mathrm{IS}}(\bar{A}_w(\omega), A_w^{n,\,m}(\omega))\right) \tag{17}$$

where d_{IS} is the Itakura–Saito measure and $A_w^{n,\,m}(\omega)$ is the spectrum corresponding to $\mathbf{a}_w^{n,\,m}$.

In (16), $(1/M_n)\sum_{m=1}^{M_n} p(\mathbf{w}|\mathbf{a}_w^{n,\,m})$ can be interpreted as the mean of the likelihoods corresponding to each codevector in the n^{th} noise codebook. If a noise codebook contains codevectors that are very different from each other, as is the case with a two-tone siren noise for instance, the likelihood of the individual codebook vectors may vary strongly. Thus, the mean codebook likelihood may be low even when a single codevector results in a high likelihood. As a result, the classifier given by (17) may fail. Motivated by the above theory, an alternate classification technique is to consider the maximum of the likelihood of the codevectors instead of the mean. The corresponding classifier is given by

$$n^* = \arg\max_n \left\{ \max_{1 \le m \le M_n} p(\mathbf{w}|\mathbf{a}_w^{n,\,m}) \right\}, \quad 1 \le n \le M. \tag{18}$$

We use the classifier given by (18) in the experiments. Fig. 2 provides a schematic diagram of the classified scheme. We note that though the classification is performed using long-term noise estimates, the different entries in the codebook permit variations in the spectral shape. What the classified scheme cannot handle instantaneously is when the distribution (pdf) of the noise (considering only the spectral shape) varies rapidly since the long-term estimate cannot adapt immediately. However, in most practical situations, the noise distribution does not change rapidly.

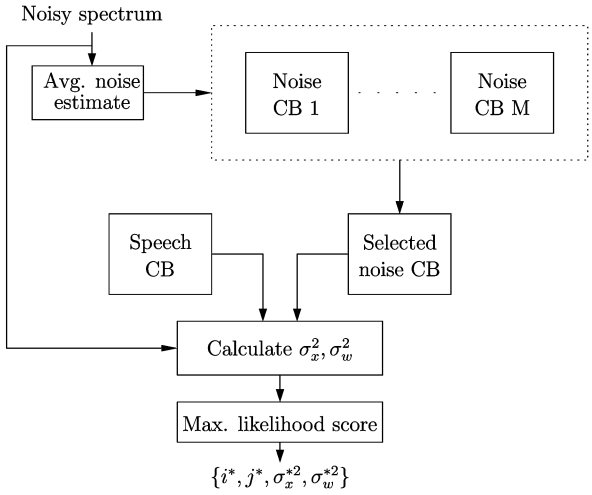

Fig. 2. Classified noise codebook scheme. Using noise information estimated from the noisy observation, a single noise codebook is chosen, which is used in the subsequent ML search. i^*, j^* are the indexes of the selected entries from the speech and noise codebooks and σ_x^{*2}, σ_w^{*2} are the corresponding excitation variances.

Since we have one codebook for each noise type, good descriptions of the noise sources can be obtained while the ambiguity mentioned earlier is avoided. A similar classified scheme is used in [14] in the context of HMM based enhancement using multiple noise HMMs. In [14], a single noise HMM is selected during periods of nonspeech activity. The selected noise HMM is used until the next occurrence of nonspeech activity when a new selection is made. In the classified scheme proposed in this paper, we perform a classification for each frame of noisy speech using the long-term estimate of the noise obtained from the observation. As mentioned earlier, another important difference is that in the method proposed here, the excitation variances are computed for each frame. In [14], the noise excitation variance of the different components of the selected noise HMM, which is part of the *a priori* information, is scaled by a gain factor only when a new selection is made during nonspeech activity.

C. Interpolation Scheme

An important feature of the proposed method using *a priori* information is that the estimated clean speech LP vector is pro-

TABLE IV
INTERPOLATIVE SEARCH

1. Obtain index using codebook-based ML search.
2. Form interpolative codebook containing NK vectors (using N interpolation steps and K nearest neighbors of the index chosen in step 1).
3. Repeat ML search with interpolative codebook to obtain final estimate.

duced from a codebook trained with speech data and is hence guaranteed to posses speech-like properties. For a given codebook size, performance can be improved if, while retaining the advantage due to *a priori* information, we can reduce the errors resulting from the limited precision of the codebooks. Performing an interpolative search is a natural way to achieve improved performance. Given two centroids from the codebook, we generate a set of points between the centroids and search for the point in the set that maximizes the likelihood. The starting point is the centroid that was selected as the ML estimate. Performing an interpolation along the lines between the ML centroid and each of its neighbors ensures that we reach a point that results in a likelihood not smaller than the codebook-based ML value. We define the nearest neighbor of a centroid as the codebook entry with the smallest Itakura–Saito distortion to the centroid. The set of interpolation points is referred to as the interpolation codebook.

In higher dimensions, the number of neighbors becomes very large and this is the case with a codebook of speech LP coefficients. As an approximation, the search can be performed using K nearest neighbors of the codebook-based ML estimate where K is related to the intrinsic dimensionality of the speech data and can be determined empirically. For a given speech codebook, the K nearest neighbors of each centroid are pre-computed and stored in a table. The search is described in Table IV.

Since the interpolation is always between two vectors from the speech codebook, the interpolation codebook consists of LP vectors that are speech-like. The interpolation scheme presented here is different from the one discussed in [23], which is an MMSE estimation of the STP parameters from a fixed set of trained (including the excitation variance) STP parameters. Here, we dynamically generate the interpolation codebook based on the vector selected from the fixed codebook and perform instantaneous variance estimation.

V. EXPERIMENTS AND RESULTS

We begin this section with a description of the objective measures of speech quality that we use and the experimental setup. This is followed by a description of a number of experiments to evaluate the performance of various aspects of the codebook-based enhancement system. Experiments were conducted to compare the performance of the ML (Itakura–Saito based) parameter estimator, log-spectral distortion and Euclidean spectral distance based estimators. Other experiments include determination of the noise codebook size and evaluation of the iterative search, the classified noise codebook scheme and the interpolation method. We also compare the performance of the proposed system to the HMM based system described in [14], the codebook constrained Wiener filtering described in [10] and the recursive EM method of [17]. Finally, the parameters obtained from the codebook-based parameter estimation scheme presented here are used in the noise suppression system of the enhanced variable rate codec (EVRC-NS) [24] and the resulting enhanced speech is compared to the output of the regular EVRC-NS through listening tests.

A. Objective Quality Measures

The objective measures of speech quality used were signal-to-noise ratio (SNR), segmental signal-to-noise ratio (SSNR) and mean log-spectral distortion (SD). The SNR (in decibels) for an utterance was computed as

$$\mathrm{SNR} = 10\log_{10}\left(\frac{\sum_{t=1}^{T} x^2(t)}{\sum_{t=1}^{T}(x(t)-\hat{x}(t))^2}\right) \tag{19}$$

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 (20 ms) in the utterance. For the n^{th} 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

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

The SD for an utterance was computed as the average of the instantaneous SD for the individual frames. A tenth order LP analysis was performed to obtain the AR envelopes. Frames whose average energy was 40 dB below the long-term average energy of the utterance were excluded in the computation of SSNR and SD [25]. We also used the Perceptual Evaluation of Speech Quality (PESQ) [26], an ITU recommendation that has been reported to have a high correlation to subjective quality.

B. Experimental Setup

A 10-bit speech codebook of linear predictive coefficients of dimension 10 was trained using the generalized Lloyd algorithm (GLA) [27] with 10 minutes of speech from the TIMIT database [28] using the Itakura–Saito measure. The sampling frequency was 8000 Hz. A frame length of 240 samples with 50% overlap was used. The frames were windowed using a Hanning window. The test set consisted of ten speech utterances, five male and five female, not included in the training. Experiments were conducted for noisy speech at 10 dB input SNR for 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 an emergency vehicle), speech babble noise (from Noisex-92) and white Gaussian noise. The highway noise captures the sound of vehicles approaching the listener and moving away and is thus nonstationary. The siren noise is nonstationary since it switches periodically between

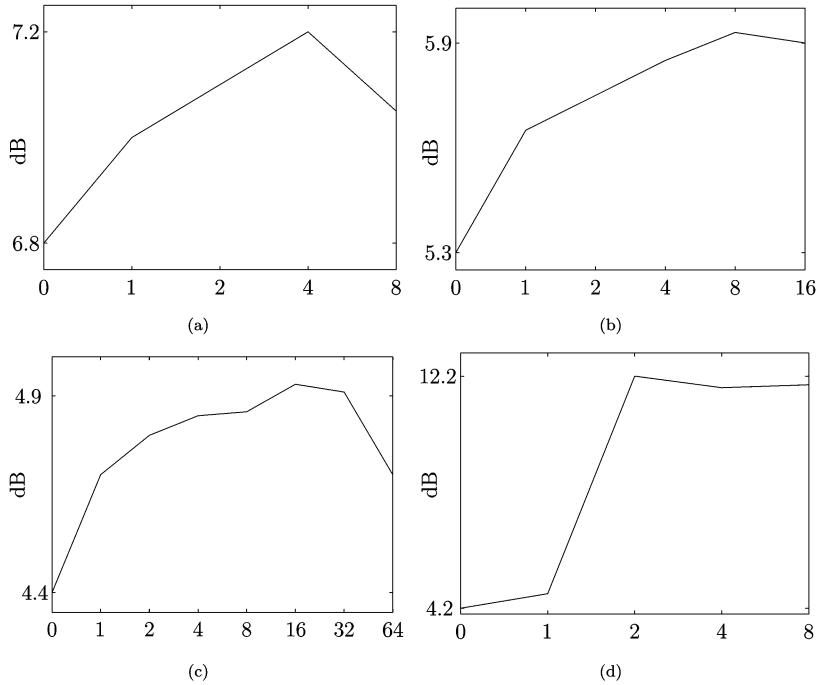

Fig. 3. Segmental SNR values for varying number of noise codebook entries. The zero-entry codebook corresponds to using noise information estimated from the observation only (no *a priori* information). (a) Highway. (b) White. (c) Babble. (d) Siren.

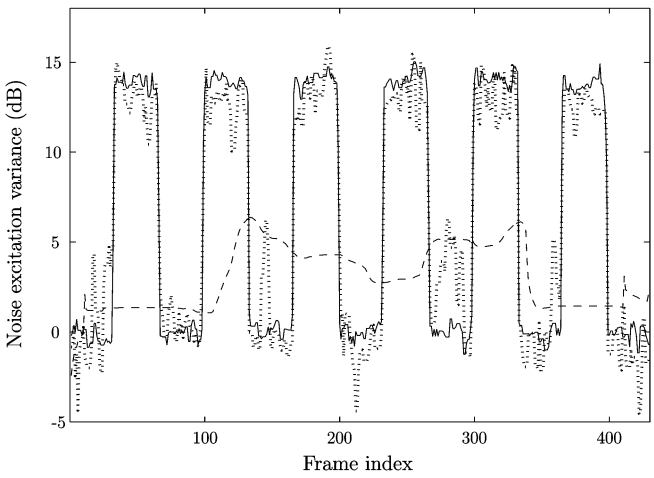

Fig. 4. Estimation of noise excitation variances for White-NS at 10 dB input SNR: true values (solid), ML estimate (dotted) and long-term estimate (dashed).

two tones that have distinct spectral shapes. 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 $5\sigma^2$ (see Fig. 4). The noise codebooks were trained using the GLA algorithm with two minutes of training data. The noise samples used in the training and testing were different. The objective quality measures for the noisy input at 10 dB SNR are provided in Table V.

Enhanced speech was obtained by applying the Wiener filter (13) to the noisy speech, without any perceptual tuning. Evaluation of the objective quality measures on the resulting enhanced speech provides a framework to study the performance of each feature of the proposed parameter estimation scheme. We note that the output of the Wiener filter is not free from residual noise,

TABLE V
SSNR, SD (BOTH IN dB) AND PESQ (MOS SCORE) VALUES AVERAGED OVER TEN UTTERANCES AT 10 dB INPUT SNR FOR THE NOISY INPUT

Noise	SSNR	SD	PESQ
Highway	1.9	3.3	2.4
White	0.7	4.6	2.1
Babble	1.3	3.2	2.4
Siren	0.7	4.7	2.3
White-NS	5.2	3.7	2.3

which can be reduced through careful tuning. In a later section we provide subjective results when using our parameter estimates in the noise suppression system of the enhanced variable rate codec [24], which is a well-tuned system.

C. Envelope Versus Periodogram

To evaluate the excitation variances of speech and noise using (12), we require the observed noisy power spectrum. The model spectrum based on the speech and noise codebooks provides the spectral envelope. Trying to fit the DFT-based periodogram to the model-based smooth envelope violates the assumption of small errors, resulting in incorrect excitation variances. So we use the envelope of the observed noisy speech instead of the DFT-based periodogram. One way to obtain the envelope is to use the AR-spectrum given by

$$P_y(\omega) = \frac{\sigma_y^2}{|A_y(\omega)|^2}, \quad A_y(\omega) = \sum_{k=0}^{r} a_{y_k} e^{-j\omega k} \tag{20}$$

where a_{y_k} are the order-r AR-coefficients of the noisy signal and σ_y^2 is the corresponding excitation variance.

TABLE VI
SSNR, SD (BOTH IN dB) AND PESQ (MOS SCORE) VALUES AVERAGED OVER TEN UTTERANCES AT 10 dB INPUT SNR FOR THE LS AND ML BASED ESTIMATORS, THE LOG-LIKELIHOOD (LL) APPROACH OF [15], THE SQUARED EUCLIDEAN SPECTRAL DISTANCE (ES) BASED ESTIMATOR [18] AND WIENER FILTERING BASED ON LONG-TERM NOISE ESTIMATES (LT)

	SSNR					SD					PESQ				
Noise	LS	ML	LL	ES	LT	LS	ML	LL	ES	LT	LS	ML	LL	ES	LT
Highway	7.2	7.2	6.7	5.2	5.9	3.0	3.0	3.1	3.3	3.0	2.7	2.7	2.6	2.5	2.6
White	5.8	5.9	5.8	5.1	5.8	4.0	4.0	4.2	4.7	4.1	2.6	2.6	2.6	2.4	2.5
Babble	4.8	4.9	4.6	2.9	4.3	3.3	3.3	3.6	3.6	3.1	2.5	2.6	2.4	2.4	2.5
Siren	12.1	12.2	11.1	9.2	2.0	2.6	2.6	2.8	3.0	4.9	3.2	3.2	3.1	3.0	2.3
White-NS	9.0	9.6	9.0	6.5	6.2	3.7	3.8	3.7	4.4	3.8	2.6	2.5	2.6	2.5	2.5

D. Noise Codebook Size

For the highway, white, babble and siren noise considered here, experiments were conducted to choose the best noise codebook size. For each noise type, the codebook-based parameter estimation was performed using noise codebooks of varying sizes. We present the results for the ML estimator. To focus on the effect of the noise codebook size alone, a full search of the speech and noise codebook was performed, instead of the iterative search. Also, for each noise type, the appropriate noise codebook was used, i.e., we assumed ideal classification. For highway and white noise, the noise LP order was 6. For babble noise, which is speech like, the LP order was 10. For siren noise, which typically exhibits strong harmonics, the LP order was 16. It was observed that objective measures such as the segmental SNR values of the enhanced speech increased up to a certain noise codebook size, after which they began to decrease. The initial increase in segmental SNR with codebook size is intuitive since small codebooks do not adequately describe the noise spectral shapes. The decrease can be attributed to the fact that with increasing size the ambiguity discussed in Section IV-B begins to play a role.

Fig. 3 shows the segmental SNR values for the different noise types, as a function of the number of noise codebook entries. The speech codebook size was fixed at 10 bits. For each frame, the noise codebooks were augmented with the long-term noise information estimated from the noisy observation using [9]. Also shown in the figure is the result for the case where the noise codebook consists of only the long-term noise information. This is denoted in the figure by a codebook with 0 entries. It can be seen that for all noise types, using *a priori* information is better than just using the long-term noise information estimated from the observation. This is consistent with the observations made in Section II. As expected, there is a large gain due to the *a priori* information for siren noise, which is nonstationary. A similar trend was observed with the other objective quality measures. Based on these results, codebook sizes of 4, 8, 16, and 2 entries were found to be optimal for highway, white, babble, and siren noise respectively. The real-world siren noise considered here consists of two tones, and thus two codebook entries were sufficient. These codebook sizes are used in the rest of the experiments. For these codebook sizes, the percentage of frames for which a noise codebook entry was preferred over the long-term noise estimate was 71%, 95%, 94%, and 98% for highway, white, babble and siren noise, respectively.

E. Performance of the Different Estimators

In this section, we compare the performance of the four different estimators listed in Table II. The noise codebook sizes were those of Section V-D. For each noise type, we used the appropriate noise codebook, i.e., we assumed an ideal classifier. Results with the classified scheme are presented in Section V-G. We used a full search of the speech and noise codebooks instead of the iterative search. This allows us to focus on the performance of the estimators. The estimated parameters were used in the Wiener filter (13). Table VI shows the SSNR, the SD and the PESQ scores for the different estimators. Also shown are the values obtained using a Wiener filter (implemented according to [1]) with long-term (LT) estimates of the noise spectrum obtained from [9] only. It can be seen that the ML estimator (using Itakura–Saito distortion) results in better performance than all the other estimators. The LS, ML and LL estimators all perform better than the ES estimator suggesting that spectral measures in the log domain are better. They also perform better than Wiener filtering with long-term noise estimates. For stationary white noise, using long-term estimates performs as well as the ML estimator, which is expected since there is no added advantage due to the frame-by-frame variance calculation in stationary noise environments. For the nonstationary noise types such as highway noise, siren noise and white-NS, the estimators proposed in this paper have a significant advantage. Fig. 4 provides a plot of the ML estimate and the long-term estimate of the noise excitation variance for white-NS. It can be seen that the ML estimator tracks the nonstationarity instantaneously. We use the ML based estimator in the experiments that follow.

F. Evaluation of the Iterative Scheme

To see that the iterative scheme does not result in loss of performance compared to a full search of the speech and noise codebooks, experiments were conducted with and without the iterative scheme. For each noise type, we used the appropriate noise codebook, i.e., we assumed an ideal classifier. Siren noise and highway noise were excluded from this test as their noise codebook contain only two and four entries respectively. The speech and noise subset sizes were fixed at 2 entries. It was observed that the iterative scheme converged within two iterations in most cases as shown in Table VII. Table VIII compares the performance of the iterative scheme to the full search. It can be seen that there is no significant loss in performance due to the iterative scheme. For nonstationary noise sources such as babble

TABLE VII
PERCENTAGE OF FRAMES WITH CONVERGENCE OCCURRING AFTER 1, 2 AND 3 ITERATIONS FOR A SUBSET SIZE OF 2 FOR BOTH SPEECH AND NOISE

Noise	1	2	3
White	78.4	21.3	0.3
Babble	73.5	25.6	0.9

TABLE VIII
SSNR, SD (BOTH IN dB) AND PESQ (MOS SCORE) VALUES AVERAGED OVER TEN UTTERANCES AT 10 dB INPUT SNR FOR THE ITERATIVE (ITER) AND FULL SEARCH (FS) SETUPS

	SSNR		SD		PESQ	
Noise	Iter	FS	Iter	FS	Iter	FS
White	5.9	5.9	4.1	4.0	2.6	2.6
Babble	4.8	4.9	3.3	3.3	2.5	2.6

noise which require a codebook with several entries, the iterative scheme results in a significant reduction in computational complexity.

G. Evaluation of the Classified Scheme

To evaluate the advantage due to the classified scheme, noisy speech at 10 dB input SNR was processed by the codebook based enhancement system with and without classified noise codebooks. We used a full search of the speech and noise codebooks instead of the iterative search. This allows us to focus on the performance of the classified scheme. Four different noise types were considered: highway, white, babble and siren noise. In the classified scheme, four separate noise codebooks, one for each noise type, were used together with the classifier (18). The same codebook was used for both the stationary and nonstationary white noise types. The noise LP order was 6 for highway and white noise, 10 for babble noise, and 16 for siren noise. In the unclassified setup, a single noise codebook was formed by concatenating the individual noise codebooks. Enhanced speech was obtained by applying the Wiener filter to the noisy speech. The classifier given by (18) performed better than the classifier in (16) and was used in the experiments.

It can be seen from Table IX that the classified scheme results in improved performance compared to a single noise codebook. In the unclassified scheme, it was found that sometimes entries from the concatenated noise codebook that did not correspond to the actual noise type were selected. We note that along with the improvement in performance, there is also a reduction in computational complexity due to the small size of the individual noise codebooks. The experiments to evaluate the ML-based estimator in Section V-E and the experiments with the classified codebooks in this section use the same noise codebook sizes. The slight difference in performance is explained by the fact that in Section V-E we assumed an ideal classifier whereas we perform the actual classification here. It is also possible that for some noise types (such as babble noise here), the classified scheme results in slightly improved performance compared to using an ideal classifier since there is a greater choice of codebooks and thus it is possible in certain frames that a different codebook contains a better representation of the noise.

TABLE IX
SSNR, SD (BOTH IN dB) AND PESQ (MOS SCORE) VALUES AVERAGED OVER TEN UTTERANCES AT 10 dB INPUT SNR FOR THE CLASSIFIED (C) AND NON-CLASSIFIED (NC) SETUPS

	SSNR		SD		PESQ	
Noise	C	NC	C	NC	C	NC
Highway	7.2	5.9	3.0	3.5	2.7	2.6
White	5.9	5.3	4.0	4.2	2.5	2.4
Babble	5.0	4.8	3.3	3.5	2.6	2.5
Siren	11.0	10.2	2.8	3.2	3.1	2.8
White-NS	8.9	7.7	3.6	3.7	2.7	2.5

TABLE X
SSNR, SD (BOTH IN dB) AND PESQ (MOS SCORE) VALUES AVERAGED OVER TEN UTTERANCES AT 10 dB INPUT SNR FOR THE CASES WITH ALL NOISE CODEBOOKS (CB), WITHOUT THE CODEBOOK FOR THE NOISE TYPE IN QUESTION (NCB) AND A SINGLE CODEBOOK OF SIZE 0 (CB-0)

	SSNR			SD			PESQ		
Noise	CB	NCB	CB-0	CB	NCB	CB-0	CB	NCB	CB-0
Highway	7.2	6.5	6.8	3.0	3.3	3.1	2.7	2.7	2.7
White	5.9	5.0	5.3	4.0	4.1	4.1	2.5	2.5	2.5
Babble	5.0	4.2	4.4	3.3	3.1	3.2	2.6	2.6	2.6
Siren	11.0	2.4	3.2	2.8	4.8	4.5	3.1	2.4	2.5
White-NS	8.9	7.6	7.8	3.6	3.6	3.7	2.7	2.6	2.6

TABLE XI
SSNR, SD (BOTH IN dB) AND PESQ (MOS SCORE) VALUES AVERAGED OVER TEN UTTERANCES AT 10 dB INPUT SNR WITH (I) AND WITHOUT (NI) INTERPOLATION AND USING A 11-BIT SPEECH CODEBOOK (CB-11)

	SSNR			SD			PESQ		
Noise	I	CB-11	NI	I	CB-11	NI	I	CB-11	NI
Highway	7.2	7.2	7.2	2.8	2.9	3.0	2.7	2.7	2.7
White	5.9	5.8	5.9	3.8	3.9	4.0	2.6	2.6	2.6
Babble	5.2	4.9	4.9	3.1	3.3	3.3	2.6	2.6	2.6
Siren	12.2	12.2	12.2	2.5	2.5	2.6	3.2	3.3	3.2
White-NS	10.5	9.2	9.6	3.4	3.8	3.8	2.7	2.5	2.5

H. Robustness to Noise Types

Experiments were performed to evaluate the robustness of the estimation scheme to noise types not represented in the codebooks. For each of the noise types considered here, the noise codebook trained on that type was excluded in the experiment, for e.g., for highway noise, the classified scheme was run using codebooks for babble, white and siren noise. Results are presented in Table X in the column NCB. We also provide the results for the case (column CB-0) where for each frame the noise codebook consists of only the long-term noise estimate for that frame (codebook of size 0 in Section V-D). This represents the case where the method does not get 'confused' with the wrong noise types.

Several conclusions can be drawn from Table Section V-D. First, the importance of using noise codebooks (column CB in the table) is reiterated. Second, for the nonstationary noise types, NCB performs better than LT in Table VI suggesting that the proposed ML estimation of the STP parameters followed by Wiener filtering is better than simple Wiener filtering

TABLE XII
SSNR, SD (BOTH IN dB) AND PESQ (MOS SCORE) VALUES AVERAGED OVER TEN UTTERANCES AT 10 dB INPUT SNR FOR THE PROPOSED SYSTEM WITH CLASSIFICATION AND INTERPOLATION (CI), THE HMM BASED SYSTEM, THE CODEBOOK-CONSTRAINED WIENER FILTER APPROACH (CCWF) AND THE RECURSIVE EM METHOD (REM)

	SSNR				SD				PESQ			
Noise	CI	HMM	CCWF	REM	CI	HMM	CCWF	REM	CI	HMM	CCWF	REM
Highway	7.2	5.9	6.5	4.8	2.8	3.0	3.2	3.6	2.7	2.6	2.4	2.6
White	6.1	6.1	7.2	7.1*	3.8	3.6	4.0	3.9*	2.6	2.5	2.2	2.6*
Babble	5.2	4.0	5.3	5.0	3.1	3.1	3.2	3.3	2.6	2.5	2.4	2.5
Siren	11.1	6.8	3.4	2.0	2.2	3.4	4.8	5.4	3.2	2.7	2.5	2.3
White-NS	9.6	6.9	6.5	7.8*	3.4	4.0	3.7	3.6*	2.7	2.3	2.2	2.5*

* These results were obtained by using the update equations tailored for the white noise case, i.e., where the only unknown parameter to be estimated was the variance. Using the general form with higher AR model orders resulted in worse values.

using the long-term noise estimate even when the noise is not adequately represented in the codebooks. Third, there is some loss in performance due to ambiguity between the long-term noise estimate and entries in the noise codebooks (comparing NCB to CB-0). The above mentioned ambiguity and the dependence on long-term estimates for robustness to new noise types is the price we pay for the improved performance in nonstationary noise environments, attained by using *a priori* noise information.

I. Evaluation of the Interpolation Scheme

Experiments were performed with and without interpolation using a 10-bit speech codebook to evaluate the performance gain due to the interpolation scheme. The size of the interpolation codebook was 100 entries. We perform the interpolation in the line spectral frequency (LSF) domain. To focus on the gain due to interpolation alone, we use a full search of the speech and noise codebooks and assume an ideal classifier, i.e., we use the appropriate noise codebook for each noise type. Experiments were also conducted using a 11-bit speech codebook without interpolation to study the advantage due to interpolation in a 10-bit codebook. As expected, it can be seen from Table XI that interpolation results in an improvement in SD values, due to the increased precision in the representation of the speech LP coefficients compared to not using interpolation. An interesting observation is that interpolation results in SD values that are similar or better than those obtained using a 11-bit speech codebook. Thus, interpolation provides good performance while reducing computational complexity.

J. Comparison to Related Systems

The codebook-constrained Wiener filter (CCWF) method [10], the HMM based enhancement system presented in [14] and the recursive EM method (REM) [17] were implemented for comparison. The minimum statistics approach [9] was used for obtaining the long-term noise spectral estimates for use in CCWF. We use the 10-bit speech codebook from the previous experiments. For the HMM based system, as suggested in [14], the speech model had five states with five mixture components in 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 [14]. The training data used to train the speech codebook was used to train the HMM as well. During periods of speech inactivity, the Viterbi algorithm was performed on the noise data, and the noise HMM resulting in the highest likelihood was selected. The model gain adaptation was performed as described in [14]. We compare the CCWF method, the HMM system and the REM technique to the proposed system with classified noise codebooks and interpolation.

It can be seen from Table XII that the proposed method performs better than the HMM based system for the nonstationary noise types. As mentioned earlier, the instantaneous variance calculation on a frame-by-frame basis plays an important role in improving performance. The HMM based methods cannot be modified in a straight-forward manner to include instantaneous variance calculation. For stationary white Gaussian noise, instantaneous variance calculation does not result in any added advantage and the HMM based method performs well.

The proposed method also outperforms CCWF for the nonstationary noise types. CCWF has a higher SSNR for babble noise and white noise. The higher SSNR can be attributed to an overall stronger attenuation by CCWF. However this was observed to have an adverse effect on low energy speech segments. SD and PESQ values are worse for CCWF for all noise types.

The proposed method clearly outperforms REM for the nonstationary colored noise types. As discussed before, estimation accuracy of REM drops with an increase in the number of unknown parameters. The method was found to perform well when tracking only the variance of white noise. The general colored noise version of REM (with higher AR order) resulted in poor estimates when applied to white noise. By using *a priori* information about the spectral shape of noise, the proposed method is able to provide good performance in both white and colored noise.

We note that the objective measures of the proposed method with classification and interpolation differ only slightly from those presented in Section V-I where the system included interpolation but assumed an ideal classifier.

K. Evaluation of Perceptual Quality

The parameter estimation described in this paper can be incorporated in several state of the art speech enhancement systems. In this work, we use the parameter estimates in the

TABLE XIII
PREFERENCE FOR PROPOSED METHOD AVERAGED OVER ALL LISTENERS. TEN UTTERANCES HAVING AN INPUT SNR OF 10 dB WERE USED FOR EACH NOISE TYPE

	Highway	White	Babble	Siren	White-NS
Score (%)	84	64	81	80	81

noise suppression system of the enhanced variable rate codec (EVRC-NS) [24]. We use the proposed system with classified noise codebooks and interpolation. The EVRC-NS requires estimates of the background noise and contains mechanisms to update the background noise estimates based on the observed noisy input. Here, we use the noise estimates obtained from the classified noise codebook scheme. The EVRC-NS is a frequency domain technique and frequency bins in the noisy spectrum are grouped together to obtain 16 channels. A frequency dependent gain factor is applied to each bin to obtain the enhanced spectrum. In our implementation, since we work with AR-spectra that do not contain the fine structure, this grouping is not necessary and we retain the individual frequency bins. For computing the frequency dependent gain factor, instead of the noisy power spectrum, we use the modeled noisy power spectrum obtained from the classified noise codebook scheme. For the estimate of the background noise spectrum for each frame, we use $\hat{P}_w(\omega) = \sigma_w^2/|A_w(\omega)|^2$, where $A_w(\omega)$ is the noise spectrum corresponding to the noise codebook entry selected for that frame and σ_w^2 is the corresponding excitation variance.

For consistency with our parameter estimation technique, we use a frame length of 240 samples with 50% overlap. The frames were windowed using a Hann window. The rest of the processing is the same as in [24]. The observed noisy spectrum is modified by the frequency dependent gain factor and is transformed back to the time domain to obtain the enhanced speech. The regular EVRC-NS used in the comparison was run without any modifications as described in [24]. We focus only on the enhancement system and do not perform the encoding/decoding operation.

AB listening tests were conducted to evaluate the performance of the proposed method. The number of listeners was 10. Enhanced speech obtained using the regular EVRC-NS was compared to the enhanced speech obtained using the EVRC-NS with the codebook-based parameter estimates. The noisy speech had a 10 dB input SNR. The methods were evaluated in pairwise comparisons on each of the noisy utterances. 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. It can be seen from Table XIII that there is a strong preference for the proposed method for the highway, babble, siren and nonstationary white noise. As expected, there is only a slight advantage for white noise which is stationary, and thus its parameters can be well estimated using conventional noise estimation techniques.

VI. CONCLUSIONS

We have presented a new technique to estimate the AR spectra of speech and noise for use in speech enhancement. We use *a priori* information about both speech and noise parameterized as LP coefficients. We derived ML estimates of the speech and noise codebook entries and their excitation variances. A key feature of the proposed method is that the excitation variances of the AR models are computed for each observation frame, thus enabling the method to work well in nonstationary noise. It was seen that the ML estimation can be viewed from a spectral matching perspective that allows us to compute the short-term predictor parameters by minimizing a distortion measure (Itakura–Saito for the ML case) between the observed noisy spectrum and the *a priori* information based model spectrum. It was shown that using the Itakura–Saito measure resulted in better performance compared to using other spectral measures, in particular the Euclidean spectral distance as in [18]. In addition to *a priori* noise information, the proposed method also uses long-term noise information estimated from the noisy observation, which serves as a safety-net for noise types not represented in the codebooks. The iterative search technique addresses the computational complexity arising due to the joint search of the speech and noise codebooks. The use of a classified noise codebook scheme results in a scalable system, reduces computational complexity and improves performance by reducing ambiguity. The scalability lies in the fact that in order to incorporate *a priori* information about a new noise source, we only need to add the appropriate codebook. Such *a priori* information can be collected from a wide range of commonly occurring noise sources. Experiments show that the proposed method performs well resulting in significant noise suppression. The estimates of the speech and noise spectra obtained from the method can be used in several state-of-the-art speech enhancement systems. In this work, we used these estimates in the noise suppression system of the enhanced variable rate codec [24]. Results from AB listening tests confirm the superior performance of the proposed method. Future work will focus on obtaining codebook-based maximum *a-posteriori* and minimum mean-squared error estimates with instantaneous excitation variance estimation.

ACKNOWLEDGMENT

The authors wish to thank the anonymous reviewers for their valuable comments that helped in significantly improving the presentation of the paper.

REFERENCES

[1] S. V. Vaseghi, *Advanced Signal Processing and Digital Noise Reduction*. New York: Wiley, 1998, ch. 6.
[2] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 2, pp. 126–137, Mar. 1999.
[3] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.
[4] Y. Ephraim and H. L. van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, pp. 251–266, Jul. 1995.
[5] H. Lev-Ari and Y. Ephraim, "Extension of the signal subspace speech enhancement approach to colored noise," *IEEE Signal Process. Lett.*, vol. 10, no. 4, pp. 104–106, Apr. 2003.
[6] U. Mittal and N. Phamdo, "Signal/noise KLT based approach for enhancing speech degraded by colored noise," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 2, pp. 159–167, Mar. 2000.

[7] 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. 9, pp. 1732–1742, Aug. 1991.
[8] V. Stahl, A. Fischer, and R. Bippus, "Quantile based noise estimation for spectral subtraction and Wiener filtering," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 3, Jun. 2000, pp. 1875–1878.
[9] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 4, pp. 504–512, Jul. 2001.
[10] T. Sreenivas and P. Kirnapure, "Codebook constrained Wiener filtering for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 4, pp. 383–389, Sep. 1996.
[11] Y. Ephraim, "A minimum mean square error approach for speech enhancement," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, Apr. 1990, pp. 829–832.
[12] B. Logan and T. Robinson, "Adaptive model-based speech enhancement," *Speech Commun.*, vol. 34, no. 4, pp. 351–368, Jul. 2001.
[13] K. K. Paliwal and W. B. Kleijn, "Quantization of LPC parameters," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Amsterdam, The Netherlands: Elsevier, 1995, ch. 12, pp. 433–468.
[14] 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.
[15] 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. Acoustics, Speech, Signal Processing*, vol. 1, May 2001, pp. 669–672.
[16] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Speech enhancement using a-priori information," in *Proc. Eurospeech*, Sep. 2003, pp. 1405–1408.
[17] Y. Zhao, S. Wang, and K.-C. Yen, "Recursive estimation of time-varying environments for robust speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, May 2001, pp. 225–228.
[18] M. Sugiyama, "Model based voice decomposition method," in *Proc. ICSLP*, vol. 4, Oct. 2000, pp. 684–687.
[19] U. Grenander and G. Szego, *Toeplitz Forms and their Applications*, 2nd ed. New York: Chelsea, 1984.
[20] R. M. Gray, A. Buzo, A. H. Gray 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.
[21] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Trans. Signal Processing*, vol. 40, no. 4, pp. 725–735, Apr. 1992.
[22] Y. Zhao, "Frequency-domain maximum likelihood estimation for automatic speech recognition in additive and convolutive noises," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 255–266, May 2000.
[23] 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. Acoustics, Speech, Signal Processing*, vol. 1, Apr. 2003, pp. 96–99.
[24] "Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems,", TIA/EIA/IS-127, 1996.
[25] N. S. Jayant and P. Noll, *Digital Coding of Waveforms*. Englewood Cliffs, NJ: Prentice-Hall, 1989.
[26] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ) – A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, May 2001, pp. 749–752.
[27] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Communications*, vol. COM-28, no. 1, pp. 84–95, Jan. 1980.
[28] "DARPA-TIMIT," *Acoustic-Phonetic Continuous Speech Corpus, NIST Speech Disc 1-1.1*, 1990.

Sriram Srinivasan (S'04) received the M.Sc. degree in mathematics in 1999 and the M.Tech degree in computer science in 2001 from the Sri Sathya Sai Institute of Higher Learning, India. He is currently pursuing the Ph.D. degree in the Department of Signals, Sensors, and Systems, KTH Royal Institute of Technology, Stockholm, Sweden.

He was a Software Engineer at Mascon Communication Tech, India, during 2001–2002. During April 2005 to June 2005, he was a Visiting Researcher at the Telecommunications Laboratory, University of Erlangen-Nuremberg, Germany. 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 in 1996, and the Ph.D. degree in information theory in 2001, both from Chalmers University of Technology, Gothenburg, Sweden.

He held a Senior Researcher position at the Department of Speech, Music, and Hearing, KTH Royal Institute of Technology, Stockholm, 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'04) received the Ph.D. degree in electrical engineering from the Delft University of Technology, Delft, The Netherlands, the Ph.D. degree in soil science, and the M.S. degree in physics, both from the University of California, and the M.S. degree in electrical engineering from Stanford University, Stanford, CA.

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, and KTH Royal Institute of Technology, Stockholm, Sweden. He is now 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.

Dr. Kleijn is an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS. He is on the Editorial Boards of the *IEEE Signal Processing Magazine* and the *EURASIP Journal of Applied Signal Processing*, and has been 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.