SPEECH ENHANCEMENT USING A-PRIORI INFORMATION WITH CLASSIFIED NOISE CODEBOOKS

Sriram Srinivasan, Jonas Samuelsson and W. Bastiaan Kleijn

Dept. of Signals, Sensors and Systems
KTH (Royal Institute of Technology), Stockholm, Sweden
{sriram.srinivasan, jonas.samuelsson, bastiaan.kleijn}@s3.kth.se

ABSTRACT

This paper focuses on the estimation of short-term linear predictive parameters from noisy speech and their subsequent use in waveform enhancement schemes. We use a-priori information in the form of trained codebooks of speech and noise linear predictive coefficients. The excitation variances of speech and noise are determined through the optimization of a criterion that finds the best fit between the noisy observation and the model represented by the two codebooks. Improved estimation accuracy and reduced computational complexity result from classifying the noise and using small noise codebooks, one for each noise class. For each segment of noisy speech, the classification scheme selects a particular noise codebook. Experimental results show good performance, especially under non-stationary noise conditions. Listening tests confirm that the new method outperforms conventional speech enhancement systems.

1. INTRODUCTION

With the ubiquitous use of mobile communications, enhancing speech subjected to background acoustic noise is a problem that has received much interest. Among the solutions proposed are the classic subtractive type method [1], Kalman filter techniques [2] and subspace based methods [3] to name a few. In this work we focus on methods that use a-priori information about speech and noise [4] [5][6]. The a-priori information consists of trained codebooks of speech and noise auto-regressive spectral shapes parameterized as linear predictive (LP) coefficients. For each frame of noisy speech, the speech and noise LP parameters and the respective excitation variances that are most likely to have resulted in the observed noisy spectrum are computed. In [5], the speech and noise spectra and excitation variances were used to construct a Wiener filter to enhance the noisy speech.

As shown in [5], the codebook-based method can handle highly non-stationary noise types since the instantaneous speech and noise excitation variances are computed for each segment (typically 20-30 ms) of noisy speech. This is in contrast to most other enhancement schemes that rely on estimating the noise statistics based only on the noisy observation. These noise estimation techniques include [7] and [8], which provide reasonably accurate estimates for stationary noise types. However, they typically employ a buffer of past samples whose length is of the order of several hundred milliseconds and thus do not react well to quickly changing noise conditions. The use of a noise codebook in addition to using noise information estimated from the observation, together with instantaneous estimation of the variances as in [5], overcomes this problem.

In this paper, we propose a classified noise codebook scheme. In the classified scheme, instead of a single noise codebook, we have a number of noise codebooks, each trained on a particular noise type. For each input segment of noisy speech, one noise codebook is selected. The classification is based on a conventional estimate of the noise spectrum, obtained using minimum statistics [7] for example. Thus the classification uses an average estimate of the noise obtained from multiple frames. We propose a maximum-likelihood classification scheme to select a single codebook. The system is easily extendable to different noise types with the addition of the appropriate noise codebook. Since the individual noise codebooks are typically smaller than a single noise codebook trained for all noise types, computational complexity is reduced. Smaller codebooks also mean that the advantage due to a-priori information is retained. Large codebooks, trained on different noise types lose this advantage to some extent, since with increasing size they provide a less restrictive representation of the noise parameter space. In this case, the speech and noise codebook entries that maximize the likelihood score in the joint codebook search may no longer be the speech and noise codebook entries that represent the underlying speech and noise data.

A similar classified scheme is used in [9] in the context of hidden Markov model (HMM) based enhancement using multiple noise HMMs, where a single noise HMM is selected during periods of non-speech activity. The selected noise HMM is used until the next occurrence of non-speech 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 an average estimate of the noise obtained from the observation. A more important difference is that in the method proposed here, the excitation variances are computed for each frame to better deal with non-stationary noise.

2. CODEBOOK BASED PARAMETER ESTIMATION

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. We have trained codebooks of speech and noise spectral shapes parameterized as LP coefficients. We consider 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 [10]. For each frame, the noisy spectrum can be modelled by a combination of speech and noise LP spectral shapes from the respective codebooks, together with their excitation variances. Given the spectral shapes and excitation variances, the modelled noisy spectrum can be written as

$$\hat{P}_y(\omega) = \frac{\sigma_x^2}{|A_x(\omega)|^2} + \frac{\sigma_w^2}{|A_w(\omega)|^2}, \tag{2}$$

where σ_x^2 and σ_w^2 are the excitation variances of clean speech and noise respectively, and

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

where $\theta_{\mathbf{x}} = (a_{x_0}, \ldots, a_{x_p}), \theta_{\mathbf{w}} = (a_{w_0}, \ldots, a_{w_q})$ are the LP coefficients of clean speech and noise with p,q being the respective LP-model orders. The parameters to be estimated are $\{\sigma_x^2, \sigma_w^2, \theta_{\mathbf{x}}, \theta_{\mathbf{w}}\}$.

This work was partially supported by the European Commission under the ANITA project (IST-2001-34327).

The above parameter estimation problem can be solved by finding the best spectral fit between the observed and the modelled noisy spectrum, with respect to a particular distortion measure. In general this is a difficult problem, but it can be solved by restricting the search space using a-priori information in the form of trained codebooks of speech and noise spectral shapes. For each combination of $\theta_{\mathbf{x}}, \theta_{\mathbf{w}}$ from the speech and noise codebooks, the excitation variances can be obtained by minimizing $d(P_y(\omega), \hat{P}_y(\omega))$, where d is a distortion measure and $P_y(\omega)$ is the observed noisy spectrum. The parameter set resulting in a global minimum of $d(P_y(\omega), \hat{P}_y(\omega))$, for all codebook combinations is selected as the optimal solution to the estimation problem. More formally, the codebook entries that are selected can be written as

$$\{i^*, j^*\} = \arg\min_{i,\,j} \Big\{ \min_{\sigma_x^2, \sigma_w^2} d(P_y(\omega), \frac{\sigma_x^2}{|A_x^i(\omega)|^2} + \frac{\sigma_w^2}{|A_w^j(\omega)|^2}) \Big\}, \quad (4)$$

where $A_x^i(\omega)$ and $A_w^j(\omega)$ correspond to the spectra of the i^{th} and j^{th} speech and noise codebook entries respectively. The conditions for this technique to result in a unique solution are described in [4]. When the distortion measure is chosen as the Itakura-Saito measure [11], we obtain maximum-likelihood estimates since maximizing the log-likelihood is equivalent to minimizing the Itakura-Saito distortion [6]. For given $A_x(\omega)$ and $A_w(\omega)$, the excitation variances that minimize the Itakura-Saito distortion between $P_y(\omega)$ and $\hat{P}_y(\omega)$ are obtained from the following system of equations [5]:

$$\mathbf{C} \begin{bmatrix} \sigma_x^2 \\ \sigma_w^2 \end{bmatrix} = \mathbf{D}, \quad (5)$$

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

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

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

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

3. CLASSIFIED NOISE CODEBOOKS

The use of a noise codebook and instantaneous estimation of speech and noise excitation variances provides good performance in highly non-stationary noise conditions [5]. Choosing an appropriate noise codebook size is critical as it affects performance in different ways. From the point of view of computational complexity due to the joint search between the speech and noise codebooks, it is helpful to have small noise codebooks. 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, we obtain a description of the noise parameter space that becomes more and more complete. This is especially the case if a single noise codebook is trained with different noise sources. 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 [4].

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 conventional estimate of the noise spectrum using the minimum statistics approach [7]. We note that this corresponds to an average estimate of the noise spectrum obtained from multiple past frames. For each segment of noisy speech, a classification is made using this average estimate and a particular noise codebook is chosen. The selected noise codebook is then used in the subsequent maximum-likelihood search. Thus, the parameter estimation can be viewed as a two-step process. In the first step, a single noise codebook is chosen from a set of noise codebooks. An estimate of the noise LP vector obtained from multiple frames of the noisy observation 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 2 using the speech codebook and the selected noise codebook. We note that the selected noise codebook is augmented with the vector of noise LP parameters estimated from the noisy observation using [7] 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 a given observation is chosen as the codebook for the current segment. The resulting maximum-likelihood classifier can be written as

$$n^* = \arg\max_{n} \frac{1}{M_n} \sum_{m=1}^{M_n} p(\mathbf{w}|\mathbf{a}_w^{n,m}), \quad 1 \le n \le N, \quad (6)$$

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 N is the number of noise codebooks. From the equivalence of the log-likelihood and the Itakura-Saito measure, (6) can be equivalently written as

$$n^* = \arg\min_{n} \frac{1}{M_n} \sum_{m=1}^{M_n} \exp\big(d_{\mathrm{IS}}(\bar{A}_w(\omega), A_w^{n,m}(\omega))\big), \quad (7)$$

where d_{IS} is the Itakura-Saito measure, $A_w^{n,m}(\omega)$ is the spectrum corresponding to $\mathbf{a}_w^{n,m}$ and $\bar{A}_w(\omega)$ is the spectrum corresponding to the vector of noise LP parameters estimated from the noisy observation using the minimum statistics approach [7] for example. Thus $\bar{A}_w(\omega)$ is an average noise estimate, obtained from multiple frames.

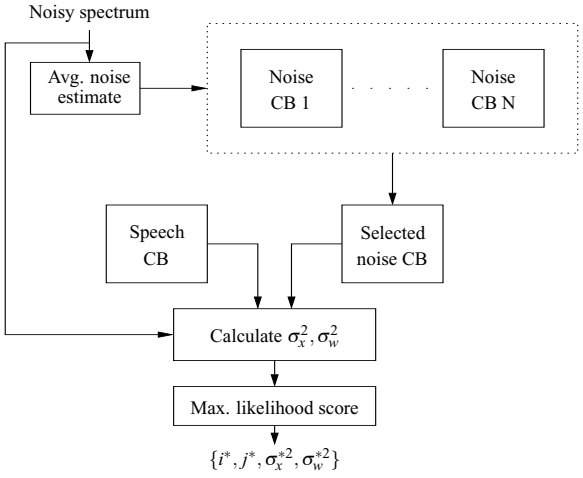

Figure 1: The classified noise codebook scheme: Using noise information estimated from the noisy observation, a single noise codebook is chosen which is used in the subsequent maximum-likelihood search. i^*, j^* are the indices of the selected entries from the speech and noise codebooks and $\sigma_x^{*2}, \sigma_w^{*2}$ are the corresponding excitation variances.

In (6), $\frac{1}{M_n} \sum_{m=1}^{M_n} p(\mathbf{w}|\mathbf{a}_w^{n,m})$ 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 siren noise for instance, 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

$$n^* = \arg\max_n \{ \max_{1 \le m \le M_n} p(\mathbf{w}|\mathbf{a}_w^{n,m}) \}, \quad 1 \le n \le N. \tag{8}$$

We use the classifier given by (8) in the experiments. Figure 1 provides a schematic diagram of the classified scheme.

The advantages of the classified scheme include the small size of the individual codebooks, which addresses the complexity issue. Also, as we have one codebook for each noise type, a good description of the noise source can be obtained and the ambiguity discussed earlier is avoided. Another important advantage of the classified scheme is that it is possible to have different LP model orders for different noise types. This was found to be particularly useful in enhancing speech corrupted by siren noise for example. Noise types such as siren noise that exhibit strong harmonics need a high order in the LP analysis compared to other noise types such as car noise. The order of the LP analysis of the observation is then suitably modified, depending on the noise codebook that is chosen. It is also possible to have different codebook sizes for different noise types.

4. EXPERIMENTS

In this section, we describe the experiments performed to evaluate the performance of the proposed classified noise codebook scheme. A 10-bit speech codebook of dimension 10 was trained using the generalized Lloyd algorithm [12] with 10 minutes of speech from the TIMIT database [13] 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 Hann 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 signal-to-noise ratio (SNR) for highway noise (obtained by recording noise on a freeway as perceived by a pedestrian standing at a fixed point), speech babble noise, siren noise and white Gaussian noise. The noise samples used in the training and testing were different. The objective measures of speech quality used were signal-to-noise ratio (SNR), segmental signal-to-noise ratio (SSNR) and log-spectral distortion (SD). The SNR for an utterance was computed as

$$\text{SNR} = 10\log_{10}\left(\frac{\sum_{t=1}^{T} x^2(t)}{\sum_{t=1}^{T}(x(t)-\hat{x}(t))^2}\right), \tag{9}$$

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. The SD was computed according to [11].

4.1 Choosing the 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 described in section 2 was performed using noise codebooks of varying sizes. To focus on the effect of the noise codebook size alone, the appropriate noise codebook was used, without performing the classification (i.e., we assume an ideal classifier). Performance of the classified scheme is discussed in section 4.2. Using the estimated STP parameters, a Wiener filter was constructed as

$$H(\omega) = \frac{\sigma_x^2/|A_x(\omega)|^2}{\sigma_x^2/|A_x(\omega)|^2 + \sigma_w^2/|A_w(\omega)|^2}. \tag{10}$$

Enhanced speech was obtained by applying the Wiener filter to the noisy speech. 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 noise codebook begins to represent a more complete description of the noise parameter space rather than a restrictive description that captures only noise-specific characteristics. As observed earlier, in this case, the speech and noise codebook entries that maximize the likelihood score in the joint codebook search may no longer be the speech and noise codebook entries that represent the underlying speech and noise data.

Figure 2 shows the segmental SNR values for the different noise types, as a function of the number of noise codebook entries. For each frame, the noise codebooks were augmented with the noise information estimated from the noisy observation using [7]. Also shown in the figure is the result for the case where the noise codebook consists of only the estimated 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 noise information estimated from the observation. As expected, there is a large gain due to the a-priori information for siren noise, which is non-stationary. Based on the segmental SNR values from the experiments, codebook sizes of 4,8,16,2 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.

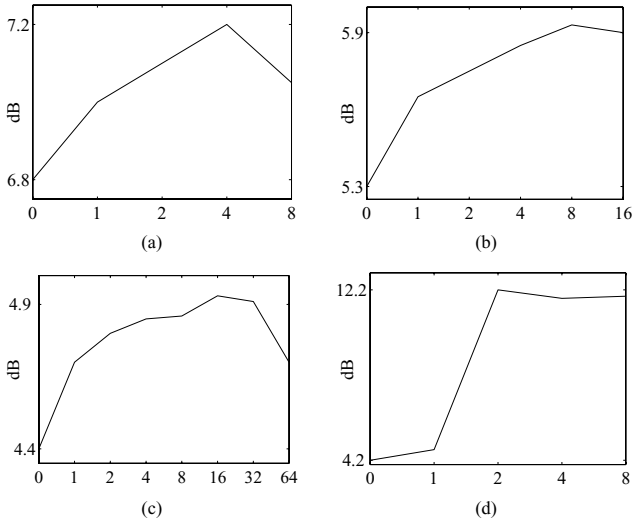

Figure 2: Segmental SNR values for varying number of noise codebook entries. The zero-entry codebook corresponds to using noise information estimated from the observation only (no a-priori information). (a) Highway. (b) White. (c) Babble. (d) Siren.

We note that the codebook-based parameter estimation method discussed in this paper can deal with two types of non-stationarity, namely, varying spectral shape and varying noise energy. The noise codebook handles varying noise spectral shape. The estimation of excitation variances of speech and noise for each observation frame handles quickly changing noise energy.

4.2 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. 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 (8). 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. A common LP order of 6 was used for all noise types.

Enhanced speech was obtained by applying the Wiener filter to the noisy speech. The classifier given by (8) performed better than the classifier in (6)and was thus used in the experiments.

It can be seen from table 1 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. There is large improvement for siren noise. This is also due to the fact that it is possible to have a higher LP order for siren noise in the classified scheme. Different LP model orders for different noise types are not possible in the unclassified scheme. 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.

	SNR		SSNR		SD	
Noise	C	NC	C	NC	C	NC
Highway	14.3	12.9	7.2	5.7	3	3.1
White	14.7	13.6	5.9	5.2	4	4.2
Babble	11.8	11.4	5	4.4	3.3	3.5
Siren	16.5	11.6	11	4	2.8	4.8

Table 1: SNR, segmental SNR (SSNR) and spectral distortion (SD) values in dB averaged over ten utterances at 10 dB input SNR for the classified (C) and non-classified (NC) setups.

4.3 Enhancement system

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 noise suppression system of the enhanced variable rate codec (EVRC-NS) [14]. 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 spectrum, we use the modelled noisy spectrum obtained from the classified noise codebook scheme. The modelled noisy spectrum is given by (2). For the estimate of the background noise power spectrum for each frame, we use $\hat{P}_w(\omega) = \frac{\sigma_w^2}{|A_w(\omega)|^2}$, 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 [14]. 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 at its native frame rate as in [14] with no changes. 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 2 that there is a strong preference for the proposed method for the highway, babble and siren noise. As expected, there is only a slight advantage for white noise, which is stationary and thus easy to estimate using conventional noise estimation techniques.

	Highway	White	Babble	Siren
Score (%)	87	63	80	81

Table 2: Preference for proposed method averaged over all listeners.

5. CONCLUSIONS

A classified noise codebook scheme with a maximum-likelihood classifier has been proposed for the codebook-based short-term predictor parameter estimation method. Experiments show that using classified noise codebooks results in improved performance compared to using a single noise codebook. The small size of the individual noise codebooks reduces computational complexity arising due to the joint search between the speech and noise codebooks. The estimates of the speech and noise spectra obtained from the classified scheme were used in an enhancement algorithm based on the EVRC noise suppression system. Listening tests show that the resulting system performs better than the regular EVRC noise suppression system.

REFERENCES

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

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

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

[4] 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, 2001, pp. 669–672.

[5] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Speech enhancement using a-priori information," in *Proc. Eurospeech*, Sept. 2003, pp. 1405–1408.

[6] ——, "Estimation of short-term predictor parameters for coding and enhancement of noisy speech," *to appear in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 2004.

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

[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, 2000, pp. 1875–1878.

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

[10] 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: Elsevier Science B.V., 1995, ch. 12, pp. 433–468.

[11] R. M. Gray, A. Buzo, A. H. Gray Jr., and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoustics, Speech Signal Processing*, vol. 28, no. 4, pp. 367–376, Aug. 1980.

[12] 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.

[13] "DARPA-TIMIT," *Acoustic-phonetic continuous speech corpus, NIST Speech Disc 1-1.1*, 1990.

[14] TIA/EIA/IS-127, *Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems*, July 1996.