HMM-Based Strategies for Enhancement of Speech Signals Embedded in Nonstationary Noise

Hossein Sameti, Hamid Sheikhzadeh, Li Deng, *Senior Member, IEEE,* and Robert L. Brennan

***Abstract*—An improved hidden Markov model-based (HMM-based) speech enhancement system designed using the minimum mean square error principle is implemented and compared with a conventional spectral subtraction system. The improvements to the system are: 1) incorporation of mixture components in the HMM for noise in order to handle noise nonstationarity in a more flexible manner, 2) two efficient methods in the speech enhancement system design that make the system real-time implementable, and 3) an adaptation method to the noise type in order to accommodate a wide variety of noises expected under the enhancement system's operating environment. The results of the experiments designed to evaluate the performance of the HMM-based speech enhancement systems in comparison with spectral subtraction are reported. Three types of noise—white noise, simulated helicopter noise, and multitalker (cocktail party) noise—were used to corrupt the test speech signals. Both objective (global SNR) and subjective mean opinion score (MOS) evaluations demonstrate consistent superiority of the HMM-based enhancement systems that incorporate the innovations described in this paper over the conventional spectral subtraction method.**

I. Introduction

SPEECH communication under noisy conditions is difficult and fatiguing. Speech sounds such as consonants, fricatives, and stops are often masked by noise, resulting in reduction of speech discrimination. The hearing impaired are at a considerable further disadvantage requiring an increase of between 2.5 and 12 dB SNR to achieve similar speech discrimination scores to those of normal hearing [1]. One characteristic of the design for future-generation hearing aids is to provide an effective front-end speech enhancement device. A major challenge in hearing aid design is to devise an effective speech enhancement strategy with the ability to cope with low SNR's (0–15 dB) and with the types of noise frequently encountered by hearing aid users, including speech weighted noise, low-frequency noise, and multitalker babble.

The main objective of speech enhancement is to improve one or more perceptual aspects of speech, such as overall quality, intelligibility for human or machine recognizers, or degree of listener fatigue. In the presence of background noise, the human auditory system is capable of employing effective mechanisms to reduce the effect of noise on speech perception. Although such mechanisms are not well understood at the present state of knowledge to allow the design of speech enhancement systems based on auditory principles, several practical methods for speech enhancement have already been developed. Digital signal processing (DSP) techniques for speech enhancement include spectral subtraction [2]–[4], adaptive filtering [5], [6], and suppression of nonharmonic frequencies [6]–[8]. Most of these techniques either require a second microphone to provide the noise reference [5], [9], [10], or require that the characteristics of noise be relatively stationary. Nevertheless, none of these requirements can be met in most of practical applications. Spectral subtraction, with no need for a second microphone and with the capability of handling noise nonstationarity to some extent, has been one of the relatively successful DSP methods. However, one major problem with this method is the annoying nonstationary "musical" [11] background noise associated with the enhanced speech. It also is incapable of coping with rapid variations in noise characteristics (e.g., simple noise amplitude variations). The basic advantage of this method is the implementation simplicity and relatively light computation requirements. We have developed a real-time spectral subtraction enhancement system using digital signal processors, which will be briefly described in this paper.

Enhancement methods that are based on stochastic models—hidden Markov models (HMM's)—have overcome the shortcomings of the DSP techniques by modeling both clean speech and noise and by accommodating the nonstationarity of speech and noise with multiple states connected with transition probabilities of a Markov chain. Using multiple states and mixtures in the HMM for noise enables the speech enhancement system to relax the assumption of noise stationarity. Another key aspect of our work described in this paper is real-time implementation of the speech enhancement system. We have successfully devised methods to reduce the system complexity and memory requirements. The HMM-based enhancement system we have implemented has the computational complexity similar to that of the spectral subtraction system. The HMM-based system is real-time implementable with its speech enhancement performance being significantly superior to the spectral subtraction method.

The organization of this paper is as follows. In Section II, a survey of the spectral subtraction and the maximum *a posteriori* (MAP), Approximate MAP (AMAP), and improved minimum mean square error (MMSE) approaches associated with the HMM-based enhancement system is carried out. In particular, a complete MMSE system with multiple states and

Manuscript received September 14, 1994; revised June 17, 1997. This work was supported by Unitron Industries Ltd., Ontario URIF fund, and by NSERC, Canada. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. John H. L. Hansen.

H. Sameti, H. Sheikhzadeh, and L. Deng are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ont., Canada N2L 3G1 (e-mail: deng@crg5.uwaterloo.ca).

R. L. Brennan is with the Unitron Industries Ltd., Kitchener, Ont., Canada N2G 4J3.

Publisher Item Identifier S 1063-6676(98)05940-9.

mixtures in the HMM's for both noise and clean speech is described in detail in Section II-B4. In Section III, a novel noise-type adaptation method for HMM-based enhancement systems is described. Two methods we developed in this study aiming at improving the implementation efficiency of the HMM-based enhancement system are presented in Section IV. The results of the experiments comparing various types of the speech enhancement systems are reported in Section V. Finally, Section VI contains the conclusions drawn from this study.

II. Speech Enhancement Methods

A. Spectral Subtraction

In this conventional method, a frequency-domain Wiener filter is constructed from the speech and noise spectral estimates at each time frame, which is then used to obtain the clean speech estimate. The noisy speech signal is first segmented into time-limited consecutive frames. Within each short-time frame, the clean speech $y(t)$, additive noise $v(t)$, and noisy speech $z(t)$ are all assumed to be stationary. Then the spectrum of the noisy speech

$$z(t) = y(t) + v(t) \tag{1}$$

is obtained as

$$Z(\omega) = Y(\omega) + V(\omega) \tag{2}$$

where $Z(\omega)$, $Y(\omega)$, and $V(\omega)$ are the power density spectra of $z(t)$, $y(t)$, and $v(t)$, respectively. A block diagram of the enhancement system we have implemented based on spectral subtraction is shown in Fig. 1 (similar to the system of [12]). In the operation of the system, a fast Fourier transform (FFT) is performed on each frame of the noisy signal to estimate the spectrum of the noisy speech. The estimate of the noise spectrum is updated during periods of nonspeech activity. An autocorrelation-based voicing and pitch detector is used for speech detection. When no speech is detected, the signal is assumed to be noise and the magnitude of noise spectral estimate, $|\hat{V}(\omega)|$, is updated as

$$|\hat{V}(\omega)|^2 = \Gamma_n |\hat{V}_{old}(\omega)|^2 + (1 - \Gamma_n)|Z(\omega)|^2 \tag{3}$$

where $|Z(\omega)|$ is the spectral magnitude of the current frame, Γ_n is a decay factor, and $|\hat{V}_{old}(\omega)|^2$ is the magnitude of noise spectral estimate before the update. The estimated noise spectral magnitude squared is then subtracted from short-time spectrum magnitude squared of the degraded speech estimated in frequency domain. Enhanced speech is obtained by reconstructing speech using the modified magnitude and the original (noisy) phase

$$Y(\omega) = [|Z(\omega)|^2 - |\hat{V}(\omega)|^2]^{1/2} e^{j\Theta_z(\omega)} \tag{4}$$

where $\Theta_z(\omega)$ is the phase of the degraded speech. In practice a Wiener filter of the form

$$H(\omega) = \left(\frac{|Z(\omega)|^2 - |\hat{V}(\omega)|^2}{|Z(\omega)|^2} \right)^{1/2} \tag{5}$$

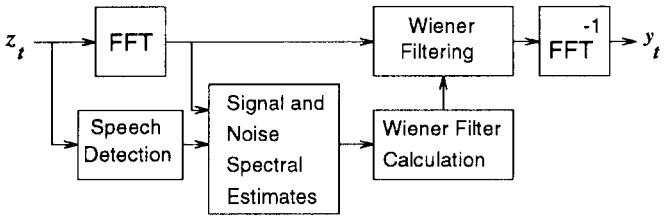

Fig. 1. Spectral subtraction enhancement system block diagram.

is constructed for every frame and the formula in (4) is calculated by the linear system operation

$$Y(\omega) = H(\omega)Z(\omega). \tag{6}$$

This method relies on the critical assumption that noise is stationary so long as the noise spectrum is not updated.

B. HMM-Based Enhancement Methods

Among stochastic enhancement methods, HMM-based methods have been most successful. HMM has long been used for speech modeling with applications to speech recognition and enhancement. There are significant differences in applying HMM's for recognition and enhancement purposes, however.

In speech recognition, a separate model for every speech unit (feature, phoneme, or word) is trained. This model is to contain the ordered sequence of stochastic properties for the utterance corresponding to that speech unit. Therefore it has to be left-to-right, i.e., transitions from a higher-indexed state to a lower-indexed state are generally prohibited. For a left-to-right model, if similar states (corresponding to similar signal properties) are to happen in different time frames, they have to be assigned as different states while they contain the same statistical information. The objective in speech recognition is to find models with maximal separation so that they give as different likelihoods for a single testing token as possible. This requires that the model best preserve the distinctive statistical properties of the training data.

The modeling problem in speech enhancement is rather different. The objective is to average out the noise signal and extract the general spectral characteristics of speech regardless of the phoneme, word, or sentence pronounced. This is done to distinguish speech from noise and not to distinguish different units of speech. Thus the structure of the speech model for enhancement should be different from that for speech recognition. First, we wish to accommodate all the speech characteristics in a single, compact model. Second, the model is not supposed to capture distinctive properties of speech within different utterances; rather, it is to capture the global characteristics of speech. Third, the temporal order of the states in the model need not be constrained since there is a single, global model for speech and different state sequences for the same state ensemble can represent distinct utterances. As a result, the speech model for enhancement is structured to be ergodic; i.e., there are no constraints on the transition probabilities of the HMM. This also makes the model less redundant since each distinct spectral shape of speech or noise needs to be represented only once in the model.

1) Training HMM's for Clean Speech and for Noise: The HMM's employed throughout this work for clean speech and for noise are ergodic mixture autoregressive hidden Markov models (AR-HMM's) [13]. These HMM's enable us to parametrically model the speech and noise spectral shapes. The output probability density function (pdf) of each mixture of the HMM's is assumed to be Gaussian AR. The likelihood of a training data sequence given the HMM (for clean speech or for noise) parameters is written in terms of the transition probabilities, mixture weights, and conditional output pdf [14]. For implementation efficiency, the output pdf is approximated by the sum of the products of the data and model autocorrelation coefficients [13].

The model parameter set for an HMM with M states and L mixtures is defined as $\lambda_y = (\pi, a, c, h)$ where $\pi = \{\pi_\beta\}$ is the set of initial state probabilities, $a = \{a_{\alpha\beta}\}$ is the set of state transition probabilities, $c = \{c_{\gamma|\beta}\}$ is the set of mixture weights, and $h = \{h_{\gamma|\beta}\}$ with $h_{\gamma|\beta}$ being the AR parameter set of a zero-mean N_yth order Gaussian AR output process corresponding to state and mixture pair (β, γ), $h_{\gamma|\beta} = \{h_{\gamma|\beta}(0), h_{\gamma|\beta}(1), \cdots, h_{\gamma|\beta}(N_y), \sigma^2_{\gamma|\beta})\}$, $h_{\gamma|\beta}(0) = 1$, $\sigma^2_{\gamma|\beta}$ being the variance (AR gain) for $\alpha, \beta = 1, \cdots, M$ and $\gamma = 1, \cdots, L$. Given a K-dimensional training data sequence $y = \{y_t\}$, $y_t \in R^K$, a maximum likelihood (ML) estimate of the parameter set λ_y is obtained and maximized through the Baum reestimation algorithm [13]. Alternately, the segmental K-means algorithm can be used to maximize the parameter set along the dominant state and mixture sequence.

Since the Baum and the segmental K-means algorithms optimize their objective functions locally, it is important to devise a good initial model. Vector quantization using Itakura–Saito distortion measure [15] (LPC-VQ [16]) is used in our system to estimate the initial model parameters. The generalized Lloyd algorithm (GLA) [17] is used to design the VQ codebook. To obtain the initial estimate for (π, a, c) parameters, the training data sequence is encoded according to the designed codebook. π, a, and c are then obtained by estimating the frequency functions associated with them.

Merhav and Ephraim have shown [18] that if the vector size K is sufficiently large, the Baum algorithm, the segmental k-means algorithm, and the LPC-VQ will generate similar model estimates. Further, as $K \to \infty$, the asymptotic performances of the three methods will be the same.

2) MAP Enhancement Method: For MAP estimation of the clean speech signal [14], the estimation-maximization (EM) algorithm [19] is employed in constructing our speech enhancement system. Let k denote the iteration index (initially set to zero). First, the weight sequence

$$q(\beta, \gamma|y(k)) \triangleq \{q_t(\beta, \gamma|y(k)), t = 0, \cdots, T\} \tag{7}$$

is evaluated for all possible states β, mixtures γ, and time frames t using the forward-backward algorithm [20]. $q_t(\beta, \gamma|y(k))$ is the conditional probability of being in state β and choosing mixture γ at time frame t given an estimate of the clean speech $y(k)$

$$q_t(\beta, \gamma|y(k)) = P(s_t = \beta, m_t = \gamma|y(k)). \tag{8}$$

Associated with each state and mixture pair (β, γ), there is a set of AR (LPC) coefficients that can be used in combination with the noise AR process to form a Wiener filter $H_{\gamma|\beta}(\theta)$. A new estimate of the clean speech is calculated by filtering the noisy speech through a weighted sum of the Wiener filters, the weights being $q_t(\beta, \gamma|y(k))$ for each time frame t. This estimate of clean speech is then used to find a probability sequence $q(\beta, \gamma|y(k))$, thus supplying a new sequence of Wiener filters and another estimate of the clean speech. This iterative process continues until a preset convergence criterion is reached. In the first iteration, noisy speech is used as an estimate of the clean speech. For each time frame, t, enhancement is done efficiently in frequency domain using [14]:

$$y_{t,\theta}(k+1) = \left[\sum_{\beta=1}^{M}\sum_{\gamma=1}^{L} q_t(\beta, \gamma|y(k)) H^{-1}_{\gamma|\beta}(\theta)\right]^{-1} z_{t,\theta} \tag{9}$$

where y and z denote the clean signal and the noisy signal, respectively, and subscript θ indicates the frequency domain components. A speech enhancement system which we developed based on the above MAP algorithm is shown in Fig. 2.

3) Approximate MAP Method: Approximate MAP (AMAP) enhancement system [14], a block diagram of which is shown in Fig. 3, is a simplified approximation of the MAP algorithm. For AMAP, a single state and mixture pair is assumed to dominate the sequence at each time frame, thus constraining the filter weights to be one for only one state and mixture pair and zero for the others. Given an estimate of the clean speech signal, the estimation of the most likely sequence of states and mixtures is carried out by applying the Viterbi algorithm using the path metric

$$\begin{aligned} &\ln \pi_\beta + \ln c_{\gamma|\beta} + \ln b(y_0(k)|m_0 = \gamma, s_0 = \beta) \\ &\qquad \text{for } t = 0, \\ &\ln a_{\alpha\beta} + \ln c_{\gamma|\beta} + \ln b(y_t(k)|m_t = \gamma, s_t = \beta) \\ &\qquad \text{for } 1 \le t \le T \end{aligned} \tag{10}$$

where $\alpha, \beta = 1, \cdots, M$, $\gamma = 1, \cdots, L$, and s_t and m_t are the state and mixture at time frame t, respectively. At each t, a frame of noisy speech, z_t, is enhanced using the Wiener filter corresponding to the most probable state and mixture pair.

Both MAP and AMAP enhancement algorithms are iterative since they use the enhanced speech as an estimate of the clean speech theoretically required by the formulations. Neither of these methods is capable of handling nonstationary noise due to the calculation of the filter weights being based on the clean signal information only and ignores noise variations.

4) Improved MMSE Enhancement Method: Finally, an improved minimum mean square error enhancement system, based on the algorithm first described in [21], has been developed in this study. In this MMSE system, a multiple state and mixture noise model was employed to accommodate nonstationarity in noise. Fig. 4 shows a simplified block diagram of the system. The MMSE enhancement system is designed to optimize the function

$$\hat{g}(y_t) = E\{g(y_t)|z_0^t\} \tag{11}$$

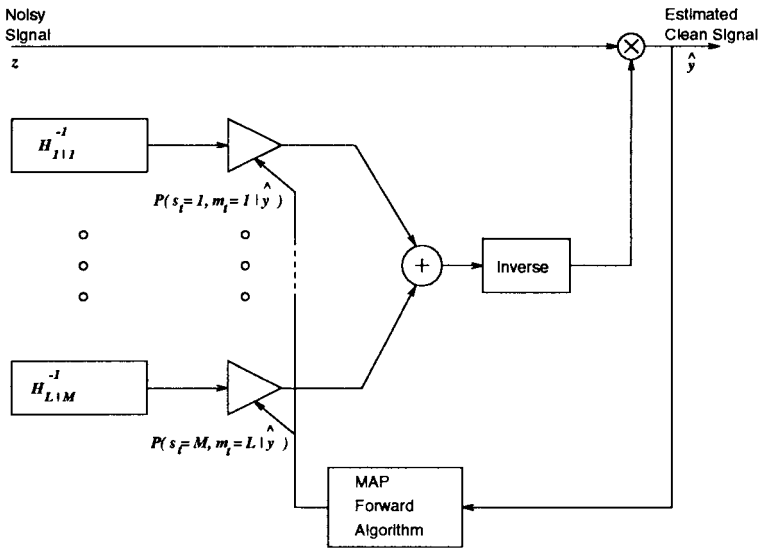

Fig. 2. MAP enhancement block diagram.

where $g(\cdot)$ is a function on R^K, $E\{\cdot|\cdot\}$ denotes conditional expectation, and $z_0^t \triangleq \{z_0, \cdots, z_t\}$ is the noisy speech data from time zero up to time t. The forward algorithm for a single state and mixture noise model [20] is extended in this study to multiple state and mixture noise model. Let n_t and p_t denote the state and mixture of noise at time frame t. Let N and P denote the number of states and mixtures of the noise HMM. $\hat{g}(y_t)$ can be calculated from

$$\hat{g}(y_t) = \sum_{\beta=1}^{M}\sum_{\gamma=1}^{L}\sum_{\xi=1}^{N}\sum_{\delta=1}^{P} q_t(\beta, \gamma, \xi, \delta|z_0^t)E \cdot \{g(y_t)|z_t, s_t = \beta, m_t = \gamma, n_t = \xi, p_t = \delta\} \quad (12)$$

where

$$q_t(\beta, \gamma, \xi, \delta|z_0^t) \triangleq P(s_t = \beta, m_t = \gamma, n_t = \xi, p_t = \delta|z_0^t) = \frac{F_t(\beta, \gamma, \xi, \delta, z_0^t)}{\sum_{\beta=1}^{M}\sum_{\gamma=1}^{L}\sum_{\xi=1}^{N}\sum_{\delta=1}^{P} F_t(\beta, \gamma, \xi, \delta, z_0^t)} \quad (13)$$

is the posterior probability of speech state β and mixture γ, and noise state ξ and mixture δ at time t given the noisy signal z_0^t. In (13)

$$F_0(\beta, \gamma, \xi, \delta, z_0) \triangleq \pi_\beta \cdot c_{\gamma|\beta} \cdot \pi_\xi \cdot c_{\delta|\xi} b(z_0|\beta, \gamma, \xi, \delta) \quad (14)$$

$$F_t(\beta, \gamma, \xi, \delta, z_0^t) \triangleq \sum_{\{s_0^{t-1}:\, s_t=\beta\}} \sum_{\{m_0^{t-1}:\, m_t=\gamma\}} \sum_{\{n_0^{t-1}:\, n_t=\xi\}} \sum_{\{p_0^{t-1}:\, p_t=\delta\}} \cdot \prod_{\tau=0}^{t} a_{s_{\tau-1}s_\tau} \cdot c_{m_\tau|s_\tau} a_{n_{\tau-1}n_\tau} \cdot c_{p_\tau|n_\tau} \cdot b(z_\tau|s_\tau, m_\tau, n_\tau, p_\tau) \quad (15)$$

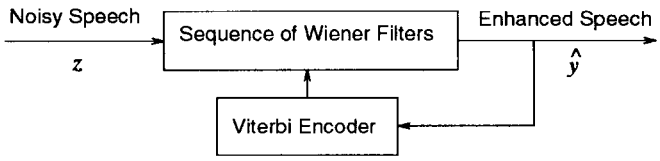

Fig. 3. AMAP enhancement block diagram.

for $t > 0$, where

$$b(z_t|s_t, m_t, n_t, p_t) = \frac{\exp\left\{-\frac{1}{2} z_t^{Tr}(\Sigma_{m_t|s_t} + \Sigma_{p_t|n_t})^{-1} z_t\right\}}{(2\pi)^{K/2}[\det(\Sigma_{m_t|s_t} + \Sigma_{p_t|n_t})]^{1/2}} \quad (16)$$

is the conditional pdf of the noisy signal z_t given that the clean signal is in state s_t with mixture component m_t and the noise frame corresponds to state n_t and mixture p_t [superscript Tr in (16) denotes matrix transposition]. In (16), $\Sigma_{\gamma|\beta} = \sigma^2_{\gamma|\beta}(A^T_{\gamma|\beta}A_{\gamma|\beta})^{-1}$ is the covariance matrix of the Gaussian output process associated with state β and mixture γ of speech AR-HMM, $\sigma^2_{\gamma|\beta}$ is the variance of the innovation process of the AR source, and $A_{\gamma|\beta}$ is a $K \times K$ lower triangular Toeplitz matrix in which the first $N_y + 1$ elements of the first column constitute the coefficients of the AR process, $h_{\gamma|\beta}(\cdot)$. Similarly, $\Sigma_{\delta|\xi}$ is the covariance matrix of the Gaussian output process associated with state ξ and mixture δ of noise AR-HMM. Note that for Gaussian HMM's representing speech and noise, the noisy process z_0^t is also a Gaussian process.

Equation (12) shows that the MMSE estimator of $g(y_t)$ given z_0^t is a weighted sum of the individual MMSE estimators of the output processes generated by the clean speech's HMM, where the weights are the probabilities that the individual estimators are the correct ones for the given noisy signal. The conditional expectation in (12) is given by

$$E\{g(y_t)|z_t, s_t, m_t, n_t, p_t\} = \int g(y_t) p_{\lambda_y\lambda_v}(y_t|z_t, s_t, m_t, n_t, p_t)\, dy_t \quad (17)$$

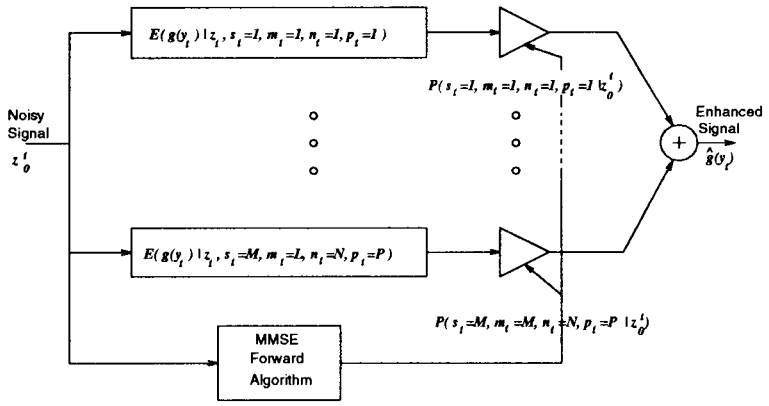

Fig. 4. MMSE enhancement block diagram.

where $p_{\lambda_y \lambda_v}(y_t|z_t, s_t, m_t, n_t, p_t)$ denotes the conditional pdf of the clean signal y_t given state s_t with mixture component m_t at time t, and the noisy signal z_t. The exact evaluation of (17) for nonlinear function $g(\cdot)$ is not trivial. For $g_1(y_t) \triangleq \{Y_t(k), k = 0, \cdots, K-1\}$ where $Y_t(k)$ is the kth component of the discrete Fourier transform (DFT) of y_t, (17) is derived to be the conventional Wiener filter. For some other functions such as

$$g_2(y_t) \triangleq \{|Y_t(k)|, k = 0, \cdots, K-1\} \quad (18)$$

$$g_3(y_t) \triangleq \{|Y_t(k)|^2, k = 0, \cdots, K-1\} \quad (19)$$

$$g_4(y_t) \triangleq \{\log|Y_t(k)|, k = 0, \cdots, K-1\} \quad (20)$$

where (17) has also been evaluated [21].

Using the system shown in Fig. 4, no iterations are necessary for the MMSE enhancement. This shows the superiority of MMSE system and its higher capability for noise cancelling in comparison with the MAP enhancement system, which needs to iterate many times to achieve an acceptable result. The more significant superiority of the MMSE system over the MAP system, however, is its ability to deal with nonstationary noise due to its inherent capability to calculate filter weights given the noisy signal instead of an estimate of the clean signal. In this work we chose the function $g(y_t)$ to be the DFT of y_t for implementation simplicity.

Equations (12)–(16) indicate that computation of $\hat{g}(y_t)$ is very costly in terms of computational complexity. For each frame t, a large number of filter weights has to be calculated using expensive calculation of (15) and (16). This makes the enhancement procedure very time consuming and far from real-time implementable. To solve this problem, we devised two methods by which the computational load dropped considerably. These methods will be described in Section IV.

III. Noise Adaptation Algorithm

In general, there are a large number of diversified types of noise, with very time-varying spectral characteristics, in the environment in which speech enhancement systems are intended to be deployed (e.g., the system as a front-end of advanced digital hearing aid). It is always an advantage for the enhancement system to have *a priori* knowledge about the noise nature. Enhancement methods which make assumptions about the noise type are deficient in terms of functionality under various corrupting noise types. The HMM-based enhancement systems are inherently relying on the type of training data for noise. Expectedly, such a system can handle only the type of noise that has been used for training noise HMM. Therefore, data from various noise types should be used for training the noise HMM. This creates the problem of a large model size for the noise HMM, making the search space expand linearly with the number of noise types with computation cost growing drastically. Furthermore, the unwanted large search space deteriorates the system performance by introducing more sources of error in the MMSE forward algorithm.

A novel noise adaptation algorithm is devised in this work 1) to enable the system to handle arbitrary types of corrupting noise and 2) to avoid up-growth in computation complexity and preserve the real-time implementation capability of the system. This algorithm, with the block diagram shown in Fig. 5, carries out noise-model selection and adaptation of the variances (LPC gains) of the Gaussian AR processes associated with the noise HMM's. During intervals of nonspeech activity, a Viterbi algorithm is performed on noise data using different noise models. By scaling the gain term in every HMM mixture by a single factor and performing the Viterbi scoring, the model gain is coarsely optimized. The noise HMM generating the best score is selected and a fine scaling adjustment is carried out to adapt to the noise level using the Viterbi algorithm again. This procedure has been motivated by our earlier work [22] and is based on the assumption that noise training sequences with similar characteristics but varying levels result in AR-HMM's differing only in the AR gains (not in spectral shapes). In order to avoid confusing unvoiced speech (mainly fricatives) with nonspeech segments contaminated with noise, only segments more than 100 ms long are used for noise model updating. Since generally no fricative or other unvoiced phoneme lasts longer than 100 ms, the system will not mistake speech with pure noise intervals.

The MMSE enhancement system does not require noise model updating as often as the spectral subtraction method, since it can handle noise nonstationarity within a specific noise type due to use of the multiple state and mixture noise model. Noise model update here is only to switch to the model

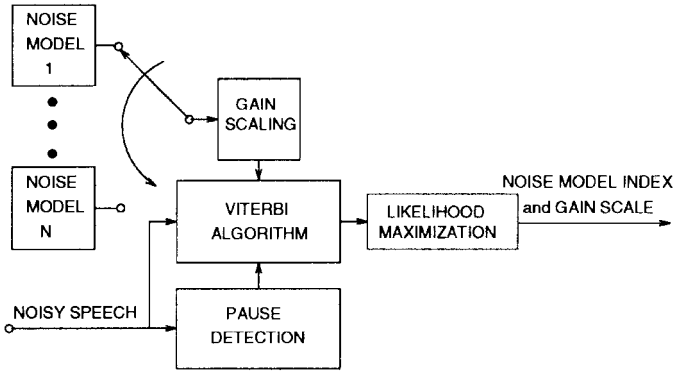

Fig. 5. Block diagram of the noise adaptation method.

representing a new noise type if required. Selection of different spectra and gains within a specific type of noise is carried out by the forward algorithm for each speech frame. A corrupting noise with continuously variable power can easily be handled by the MMSE method without the requirement to update the noise model type; in contrast, spectral subtraction method fails to follow the continuous noise power variations. This method of noise model selection can successfully cope with noise level variations as well as different noise types as long as the corrupting noise has been modeled during the training process. Further, the method keeps the noise model sufficiently compact so that excessive computation cost in enhancement is avoided. Assume that a three-state and three-mixture HMM is required to model each noise type, and assume that five noise types are to be dealt with. Without the noise adaptation algorithm, 45 possible output distributions have to be searched to select a noise pdf. Using the noise adaptation algorithm, this search space is reduced to only nine output distributions at a time. The only extra computation is due to the selection of the appropriate noise model once every few seconds during the nonspeech activity.

IV. Efficient Implementation of the MMSE Enhancement Algorithm

With a real-time implementable system as an objective, the MMSE enhancement algorithm is to be efficiently implemented for reducing the computation requirement of the system to that comparable to the conventional DSP method. For this, the following methods have been devised in implementing our speech enhancement system.

A. Double Pruning the MMSE Forward Calculation

Calculation of MMSE forward probability and filter weights which constitutes a major computational load, is carried out according to (13). For speech and noise HMM's of sizes $M \times L$ and $N \times P$, these equations call for calculation of $M \times L \times N \times P$ filter weights and the same number of pdf values for each time frame t. Since the majority of these weights are negligible due to their extremely small values (orders as little as 10^{-200}), an efficient pruning method was devised and implemented to reduce the computation cost, as well as the memory requirement, of the system.

Equation (15) can be rewritten in the following recursive form:

$$F_t(\beta, \gamma, \xi, \delta, z_0^t) = \left[\sum_{\beta'=1}^{M}\sum_{\gamma'=1}^{L}\sum_{\xi'=1}^{N}\sum_{\delta'=1}^{P} F_{t-1}(\beta', \gamma', \xi', \delta', z_0^{t-1}) \cdot a_{\beta'\beta} \cdot a_{\xi'\xi}\right] \cdot c_{\gamma|\beta} \cdot c_{\delta|\xi} \cdot b(z_t|\beta, \gamma, \xi, \delta) \tag{21}$$

where $b(\cdot)$ is calculated from (16). For pruning, $b(\cdot)$ values are first normalized by their maximum value. [This does not affect the filter weights since the forward probabilities appear both in the numerator and denominator in (13).] Then all the $b(\cdot)$ values less than an empirically determined certain threshold are deleted and (21) is calculated only for the remaining $b(\cdot)$'s. A second pruning is performed for F_{t-1} and only the significant values of F_{t-1} are used to calculate (21). This double pruning method allows the computation cost of the enhancement process to be independent of the size of the speech and noise HMM's, since the number of saved filter weights does not directly depend on the model size. Without this pruning, however, the computation cost would increase proportionally with the speech and noise model sizes.

B. Approximating the pdf of Noisy Speech

Calculation of (16) is very costly because of the $K \times K$ matrix inversion ($K = 256$ in our system) of the covariance matrix and multiplication of matrices with dimensions as large as K (for $K = 256$, computation cost is of order $K^3 = 1.6 \times 10^7$). Since the summation of two AR processes is not necessarily an AR process, the assumption of structured covariance matrix for noisy speech (in order for $\Sigma_{z_t} = \Sigma_{\gamma|\beta} + \Sigma_{\delta|\xi}$ to be decomposable into Toeplitz matrices comprised of AR coefficients of process z_t [23]) is generally invalid. To avoid the expensive calculation, an approximation method was devised for the inversion of the noisy covariance matrix. For any process z_t, the covariance matrix can be written in the form [24]

$$\Sigma_{z_t}^{-1} = CP^{-1}C^T \tag{22}$$

where C and P are upper triangular and diagonal $(K+1) \times (K+1)$ matrices, respectively, of the forms

$$C = \begin{bmatrix} 1 & a_1(1) & a_2(2) & \cdots & a_K(K) \\ 0 & 1 & a_2(1) & \cdots & a_K(K-1) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \tag{23}$$

$$P = \text{diag}(r_{zz}(0), E_1, E_2, \cdots, E_K) \tag{24}$$

where $a_i(j)$ is the ith coefficient of the jth order linear predictor for the process z_t, $r_{zz}(i)$ is the ith autocorrelation coefficient of the process z_t, and E_i is the squared prediction error for the ith order linear predictor. The exponent term in (16) needs

$$D = z_t^{Tr}\Sigma_{z_t}^{-1}z_t \tag{25}$$

to be calculated. From (22) we can write

$$D = z_t^{Tr} C P^{-1} C^{Tr} z_t = [z_t^{Tr} C (P^{-1})^{1/2}][(P^{-1})^{1/2} C^{Tr} z_t] \tag{26}$$

so

$$D = [z_t^{Tr} U^{Tr}][U z_t] = [U z_t]^{Tr}[U z_t] \tag{27}$$

where $U = (P^{-1})^{1/2} C^{Tr}$. Thus, the inversion of the $K \times K$ matrix is avoided but the problem of multiplying large matrices still remains to be resolved. Note that $\det(\Sigma_{z_t}) = \det(P)$ with P being a diagonal matrix, so $\det(\Sigma_{z_t})$ is found by calculating the product of diagonal elements of the matrix P. To resolve the second computation problem, instead of calculating U, approximated U is calculated by considering the process z_t as an AR process of a higher order than the orders of either of the two processes y_t and v_t (the clean signal and noise). For an AR process of order p, for $j \geq p$ we have

$$a_j(i) = \begin{cases} a_p(i) & i = 1, 2, \cdots, p \\ 0 & i > p \end{cases} \tag{28}$$

$$E_j = E_p. \tag{29}$$

Therefore, U will be a Toeplitz matrix after pth row in (30), shown at the bottom of the page. U can be separated into two parts; the first part comprising of the first p rows and the second part of the other $K-p$ rows. Multiplication of the first p rows is done easily due to the small value of p compared to K ($p = 14$ and $K = 256$ in our system). The second part of the matrix has a circular structure, and for implementation efficiency the output pdf can be approximated by the sum of the products of the autocorrelation coefficients of the data and of the model AR parameters [14] as follows.

For a zero-mean pth-order Gaussian AR output process with the AR parameter set of $a_p \triangleq \{a_p(0), a_p(1), \cdots, a_p(p)\}$, $a_p(0) = 1$ and gain σ^2 and observation z_t with vector size K, if $K \gg p$ then the output pdf can be approximated by

$$b(z_t) = \frac{\exp\{-\alpha/(2\sigma^2)\}}{(2\pi\sigma^2)^{K/2}} \tag{31}$$

where α is defined as

$$\alpha \triangleq r_t(0) R_p(0) + 2 \sum_{m=1}^{p} r_t(m) R_p(m). \tag{32}$$

The terms $r_t(m)$ and $R_p(m)$ are simply autocorrelation sequences defined as

$$r_t(m) = \sum_{n=0}^{K-m-1} z_t(n) z_t(n+m) \tag{33}$$

$$R_p(m) = \sum_{n=0}^{p-m-1} a_p(n) a_p(n+m). \tag{34}$$

Using the above method, generation of covariance matrices of clean speech, Σ_{y_t}, and of noise, Σ_{v_t}, separately for calculating Σ_{z_t} is avoided. Instead, the autocorrelation coefficients of the clean speech and noise processes are calculated from their AR coefficients. Assuming additivity and independence of the noise and original speech signal, their autocorrelation coefficients are added for the autocorrelation coefficients of the noisy speech to be obtained. Levinson–Durbin [24] recursion is performed on the calculated autocorrelation coefficients to find the AR coefficients of the noisy process, a_p, and the error prediction terms, E_i. The matrix U can then be calculated. However, the dominant part in calculating D [from (27)] is the part due to the lower (circulant) segment of U since $K \gg p$. Moreover, this part of calculation is further simplified by approximating D with α as shown in (32). Hereby, the computation cost for calculating the noisy process pdf (16) is drastically reduced from the order of K^3 to the order of $K \times p$.

V. Speech Enhancement Experiments

A. Speech Enhancement System Overview

The speech data used in the speech enhancement experiments reported in this section were selected from the sentences in the TIMIT data base. One hundred sentences spoken by 13 different speakers with a sampling rate of 16 kHz were used for training the clean speech model. One frame of speech covers 256 speech samples (equivalent to 16 ms). No interframe overlap was used in training the speech model. In all the experiments, the speech model consisted of five states and five mixtures. The sentences used for enhancement tests were selected such that there were no common sentences or speakers between the enhancement and training sets. A 50% overlap between adjacent frames was used in the enhancement procedure.

A block diagram of the implemented MMSE enhancement system is shown in Fig. 6. Each frame of noisy speech

$$U = \begin{bmatrix} r_{zz}(0) & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ a_1(1) & E_1^{-(1/2)} & 0 & 0 & 0 & 0 & \cdots & 0 \\ a_2(2) & a_2(1) & E_2^{-(1/2)} & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_p(p) & \cdots & a_p(1) & E_p^{-(1/2)} & 0 & 0 & \cdots & 0 \\ 0 & a_p(p) & \cdots & a_p(1) & E_p^{-(1/2)} & 0 & \cdots & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & a_p(p) & \cdots & a_p(1) & E_p^{-(1/2)} & 0 \\ 0 & 0 & 0 & 0 & a_p(p) & \cdots & a_p(1) & E_p^{-(1/2)} \end{bmatrix} \tag{30}$$

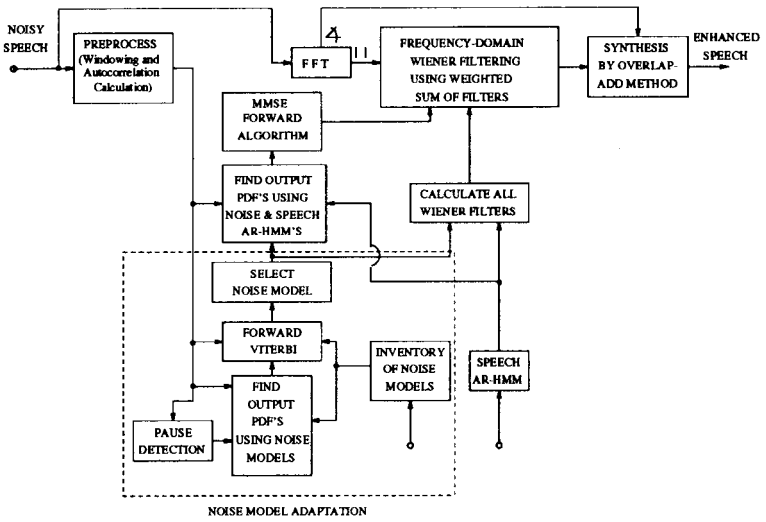

Fig. 6. Block diagram of the MMSE enhancement system.

is first preprocessed, having its AR coefficients extracted. The components inside the dashed lines in Fig. 6 implement the noise-model adaptation method described in detail in Section III. Briefly, the noisy signal during long periods of nonspeech activity are first fed into a Viterbi-like forward algorithm. Then, the likelihood for each pretrained noise HMM is calculated and compared with likelihoods for the other noise HMM's and the model associated with the highest likelihood determines the selected noise model. Using the selected noise HMM parameters and the clean speech model, the preprocessed noisy speech is input to the MMSE forward algorithm [specified in (13)], which generates the weights for the Wiener filters.

In the meantime, all Wiener filters for each combination of the state and mixture pairs in the speech and noise models are calculated. A single weighted filter is constructed for each frame of noisy speech using the calculated filter weights and the pretrained Wiener filters. The filtering of the noisy signal is carried out using the weighted filter. This generates the spectral magnitude of the enhanced speech signal. Using this magnitude and the noisy speech's phase information, an inverse FFT is performed to obtain the time-domain enhanced speech via the standard overlap-and-add method [25].

In speech enhancement experiments, three different types of noise were used: white noise, simulated helicopter noise (obtained by modulating the white noise with a 5-Hz sinusoidal waveform), and multitalker lowpass noise that was recorded in a lively cocktail party environment. Since the MAP and AMAP algorithms cannot cope with nonstationary noise, they were used only for the sentences corrupted with white noise. Spectral subtraction and the MMSE methods can handle noise nonstationarity, hence their performances on all three types of noise were compared. For the MMSE enhancement, the noise HMM's we have implemented contained three states and three mixtures. The noise models containing five states and five mixtures have also been used in a few tests and we found that they did not result in notable improvements over the three states and three mixtures noise models.

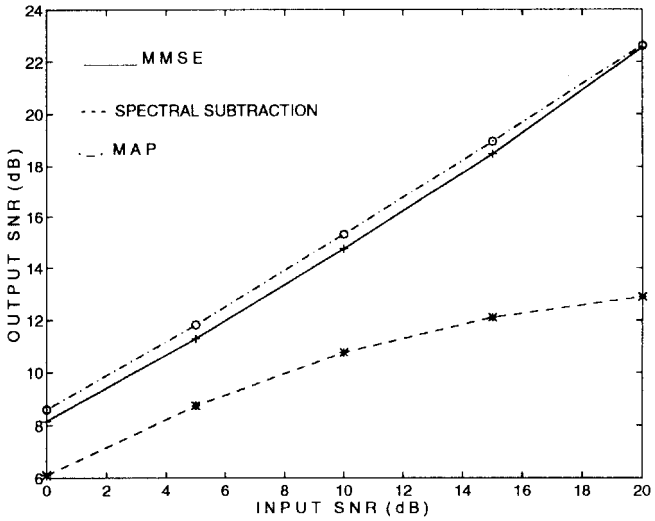

Fig. 7. Comparison of MAP, MMSE, and spectral subtraction systems for white noise corrupted speech signals.

B. Results Using Objective Evaluation

A global measure of signal-to-noise ratio (SNR) was used as the objective evaluation criterion throughout this work, which is calculated by

$$\mathrm{SNR} = 10 \log \frac{\sum_{n=1}^{K} y^2(n)}{\sum_{n=1}^{K} [y(n) - \hat{y}(n)]^2} \tag{35}$$

where K is the frame-length, $y(n)$ is the clean speech signal, and $\hat{y}(n)$ the enhanced speech signal. In our tests, input SNR's varied from 0 to 20 dB. Spectral subtraction and several types of HMM-based systems were implemented for enhancement. Figs. 7–9 show the output SNR's of these enhancement systems averaged over ten different test sentences corrupted

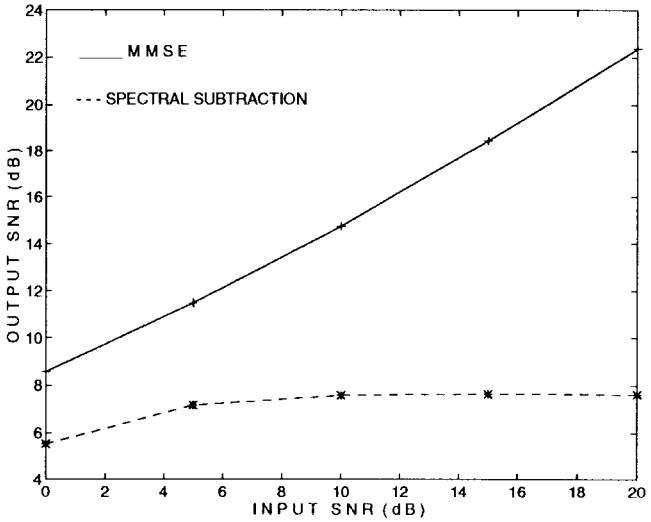

Fig. 8. Comparison of MMSE and spectral subtraction systems for helicopter noise corrupted speech signals.

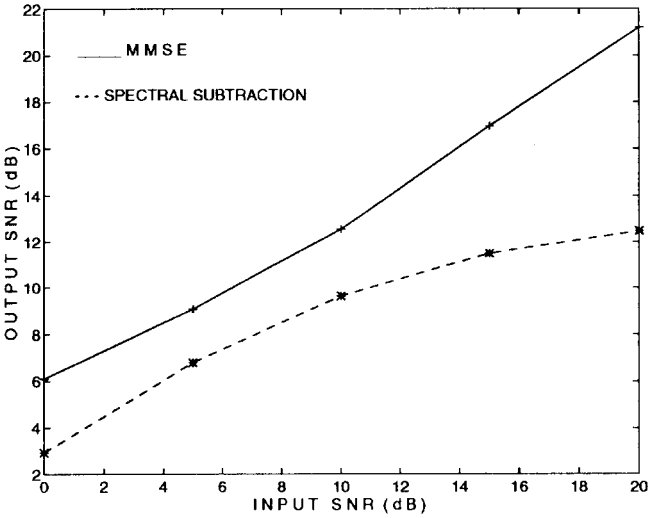

Fig. 9. Comparison of MMSE and spectral subtraction systems for multitalker noise.

by the white noise, helicopter noise, and multitalker noise, respectively.

As evident from Figs. 7–9, the HMM-based systems always outperform the spectral subtraction system. For the white noise case, the HMM-based systems have an advantage of at least 2.5 dB SNR over the spectral subtraction system, and since the noise is stationary, the performances of the MMSE and MAP systems are similar to each other. For the two nonstationary-noise cases (Figs. 8 and 9), while the MMSE system results in almost linear input–output relation with respect to the SNR values, the spectral subtraction system tends to saturate in output SNR at high input SNR's and falls behind the MMSE system by at least 2.5 dB even at low input SNR's. The spectral subtraction system fails to handle noise nonstationarity that is as simple as the simulated, highly regular helicopter noise. In fact, for input SNR's of greater that about 10 dB, the spectral subtraction method deteriorates the signal such that the output

TABLE I
FIVE-POINT ADJECTIVAL SCALES FOR QUALITY AND IMPAIRMENT, AND ASSOCIATED SCORES (AFTER JAYANT AND NOLL)

Score	Impairment
5 (Excellent)	Imperceptible
4 (Good)	(Just) Perceptible but not Annoying
3 (Fair)	(Perceptible and) Slightly Annoying
2 (Poor)	Annoying (but not Objectionable)
1 (Bad)	Very Annoying (Objectionable)

SNR is lower than the input SNR. These results are consistent with the results of subjective evaluations presented in the next section. In these cases, listeners prefer the unprocessed noisy sentence over the enhanced one using the spectral subtraction method.

C. Results Using Subjective Evaluation

For the spectral subtraction system, we found that the process of dynamic reduction of spectral energy always introduces an audible artifact, a "musical"-like signal-dependent interference. Since the spectral subtraction algorithm raises the SNR without knowledge about speech characteristics, low-amplitude speech signals such as stops tend to be lost at input SNR's below 5 dB. This reduces effectiveness of the algorithm in enhancing speech intelligibility. Under low input SNR conditions, the problem of musical noise bothered the listeners extensively. Although the SNR's were improved in these cases, some listeners could not tolerate the musical noise. For the higher input SNR tests (10 dB and more), the noise reduction was not carried out efficiently and the musical noise was also generated although not as strong as the low input SNR cases. In all cases, some listeners preferred the nonprocessed signal over the enhanced one.

On the other hand, since the HMM-based systems use speech information already embedded in the trained model, their output intelligibility should be always better than the spectral subtraction method at a cost of higher system implementation complexity. This ought to be particularly true for the MMSE enhancement strategy, since it is capable of coping with noise nonstationarities. The SNR results presented in Section V-B have indirectly reflected this fact.

To test the above inferences, mean opinion score (MOS) comparative evaluations were conducted for the MMSE system and the spectral subtraction system. Both of the systems were scored by five native English speakers using the scoring criterion established in Table I.

Fig. 10 shows the MOS results averaged over ten test sentences contaminated by the three types of noise (denoted by W for white noise, H for simulated helicopter noise, and M for multitalker noise), each at 0, 5, and 10 dB input SNR levels. The results show that the MMSE system consistently outperforms the spectral subtraction system by one score on average. In general, the MOS results are consistent strongly with the SNR objective evaluations reported in Section V-B.

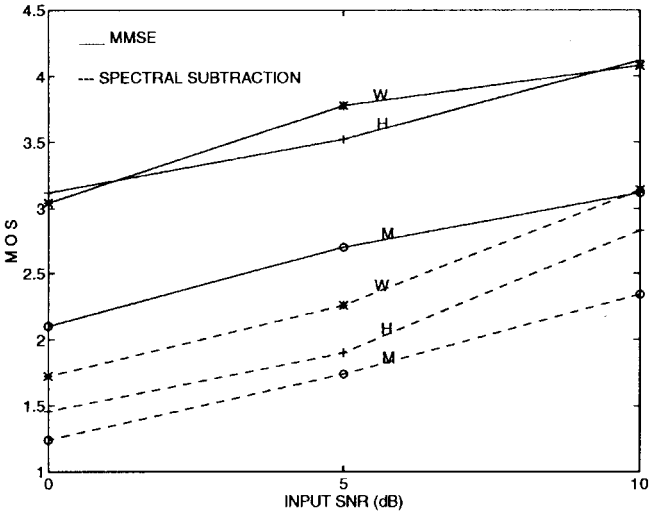

Fig. 10. MOS results for MMSE and spectral subtraction systems averaged over ten sentences evaluated by five listeners. W: white noise. H: helicopter noise. M: multitalker noise.

VI. Conclusion

The principal contribution of this study is its demonstration that the use of general statistical characteristics of speech, as partially captured by the HMM trained from a large corpus of clean speech data, is beneficial in improving the performance of speech enhancement systems. The HMM-based MMSE speech enhancement system is shown to be consistently superior in performance to the spectral subtraction based system in handling noise nonstationarity. This superiority is demonstrated by both subjective and objective evaluations for three different types of noise and for the SNR values ranging from 0 to 20 dB.

The second contribution of this study is its development of the novel noise-model adaptation method that is highly efficient in reducing the noise-model size and in reducing the noise-model training time. This makes the HMM-based MMSE speech enhancement system capable of handling a wide variety of noise types, as well as handling a wide variation in the noise power. The noise-model adaptation method also results in a considerable reduction of computational cost associated with processing noisy speech data.

On the Sun Sparc2 workstation in which all our speech enhancement algorithms were developed, the several optimization methods employed in our system implementation that have been described in this paper currently yield an execution speed of about 0.1 times the real-time speed of speech utterances for the most successful HMM-based MMSE algorithm. Therefore, the algorithm is fully capable of being implemented in real-time using DSP processors.

Acknowledgment

The authors thank Dr. H. Arndt for his encouragement, discussions, and support of this work.

References

[1] B. Hagerman, "Clinical measurements of speech reception threshold in noise," *Tech. Audiol.*, vol. ISSN 0280-6819, Report TZ, p. 13, June 1983.

[2] J. S. Lim, "Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 471–472, Oct. 1978.

[3] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 113–120, Apr. 1979.

[4] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, pp. 1586–1604, Dec. 1979.

[5] B. Widrow *et al.*, "Adaptive noise cancelling: Principles and applications," *Proc. IEEE*, vol. 63, pp. 1692–1716, Dec. 1975.

[6] R. H. Frazier, S. Samsam, L. D. Braida, and A. V. Oppenheim, "Enhancement of speech by adaptive filtering," in *Proc. ICASSP*, 1976, pp. 251–253.

[7] J. S. Lim, A. V. Oppenheim, and L. D. Braida, "Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 354–358, Aug. 1978.

[8] T. W. Parsons, "Separation of speech from interfering of speech by means of harmonic selection," *J. Acoust. Soc. Amer.*, vol. 60, pp. 911–918, Oct. 1976.

[9] O. Mitchell, C. Ross, and G. Yates, "Signal processing for a cocktail party effect," *J. Acoust. Soc. Amer.*, vol. 50, pp. 656–660, Aug. 1971.

[10] S. F. Boll and D. C. Pulsipher, "Suppression of acoustic noise in speech using two microphone adaptive noise cancellation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 752–753, Dec. 1980.

[11] S. F. Boll, "Speech enhancement in the 1980s: Noise suppression with pattern matching," in *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Eds. New York: Marcel Decker, 1991, ch. 10, pp. 309–325.

[12] D. G. Jamieson and R. Brennan, "Evaluation of speech enhancement strategies for normal and hearing-impaired listeners," in *Proc. Workshop on Speech Processing in Adverse Conditions*, Cannes-Mandelieu, France, 1992, pp. 154–157.

[13] B.-H. Juang and L. R. Rabiner, "Mixture autoregressive hidden Markov models for speech signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 1404–1413, Dec. 1985.

[14] Y. Ephraim, D. Malah, and B.-H. Juang, "On the application of hidden Markov models for enhancing noisy speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-37, pp. 1846–1856, Dec. 1989.

[15] R. M. Gray, A. Buzo, A. H. Gray, and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 367–376, Aug. 1980.

[16] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression.* Boston, MA: Kluwer, 1992.

[17] A. Buzo, A. H. Gray, R. M. Gray, and J. D. Markel, "Speech coding based on vector quantization," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. ASSP-28, pp. 562–574, Oct. 1980.

[18] N. Merhav and Y. Ephraim, "Hidden Markov modeling using a dominant state sequence with application to speech recognition," *Comput. Speech Lang.*, vol. 5, pp. 327–339, 1991.

[19] C. F. J. Wu, "On the convergence properties of the EM algorithm," *Ann. Stat.*, vol. 11, pp. 95–103, 1983.

[20] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–285, Feb. 1989.

[21] Y. Ephraim, "A minimum mean square error approach for speech enhancement," *Proc. ICASSP*, pp. 829–832, 1990.

[22] H. Sheikhzadeh and L. Deng, "Waveform-based speech recognition using hidden filter models: Parameter selection and sensitivity to power normalization," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 80–91, Jan. 1994.

[23] R. M. Gray, "Toeplitz and circulant matrices: A review," Tech. Rep., Stanford Univ., Stanford, CA, Apr. 1993.

[24] S. M. Kay, *Modern Spectral Estimation.* Englewood Cliffs, NJ: Prentice-Hall, 1988.

[25] J. S. Lim and A. V. Oppenheim, *Advanced Topics in Signal Processing.* Englewood Cliffs, NJ: Prentice-Hall, 1988.

Hossein Sameti, photograph and biography not available at the time of publication.

Hamid Sheikhzadeh received the B.S. and M.S. degrees in electrical engineering from Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran, in 1986 and 1989, respectively. He is currently pursuing the Ph.D. degree in electrical engineering.

Since 1990, he has been a Research and Teaching Assistant in the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research interests include signal processing and speech processing, with particular emphasis on speech recognition, speech enhancement, and auditory modeling.

Li Deng (S'83–M'86–SM'91) received the B.S. degree in biophysics from the University of Science and Technology of China in 1982, and the M.S. and Ph.D. degrees in electrical engineering from the University of Wisconsin-Madison in 1984 and 1986, respectively.

He was with INRS-Telecommunications, Montreal, P.Q., Canada, working on large vocabulary speech recognition from 1986 to 1989. Since 1989, he has been with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ont., Canada, where he is currently Full Professor. From 1992 to 1993, he conducted sabbatical research at Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, working on statistical models of speech production and the related speech recognition algorithms. His research interests include acoustic-phonetic modeling of speech, speech recognition, synthesis, and enhancement, speech production and perception, statistical methods for signal analysis and modeling, nonlinear signal processing, neural network algorithms, computational phonetics and phonology for the world's languages, and auditory speech processing.

Robert L. Brennan, photograph and biography not available at the time of publication.