On the Application of Hidden Markov Models for Enhancing Noisy Speech

YARIV EPHRAIM, MEMBER, IEEE, DAVID MALAH, FELLOW, IEEE, AND BIING-HWANG JUANG, SENIOR MEMBER, IEEE

Abstract—**A maximum *a posteriori* approach for enhancing speech signals which have been degraded by statistically independent additive noise is proposed. The approach is based upon statistical modeling of the clean speech signal and the noise process using long training sequences from the two processes. Hidden Markov models (HMM's) with mixtures of Gaussian autoregressive (AR) output probability distributions are used to model the clean speech signal. The model for the noise process depends on its nature. For Gaussian noise with a theoretically flat power spectral density considered here, a low-order Gaussian AR model is used. The parameter set of the HMM is estimated using the Baum or the EM (estimation-maximization) algorithm. The enhancement of the noisy speech is done by means of reestimation of the clean speech waveform using the EM algorithm. Efficient approximations of the training and enhancement procedures which involve explicit estimation of the state sequence associated with the clean speech are examined. This results in the segmental *k*-means approach for hidden Markov modeling, in which the state sequence and the parameter set of the model are alternately estimated. Similarly, the enhancement is done by alternate estimation of the state and observation sequences, using Viterbi decoding and time-varying Wiener filtering, respectively. An approximate improvement of 4.0–6.0 dB in signal-to-noise ratio (SNR) is achieved at 10 dB input SNR.**

I. INTRODUCTION

THE problem of enhancing speech signals which have been degraded by noise is basically an estimation problem in which a given function of the clean speech has to be estimated from a given sample function of the noisy speech so as to minimize the expected value of a given distortion measure between the clean and the estimated speech signals. Some examples are minimum mean-square error (MMSE) estimation of the clean speech waveform or of its sample spectrum function, given the noisy speech. The latter example is of particular interest since optimal autoregressive (AR) modeling of the clean speech, under the Itakura–Saito distortion measure, is achieved by AR modeling of its MMSE sample spectrum estimator [1]. Optimal solutions to the speech enhancement estimation problem, in the above-defined sense or in a sense of maximizing an appropriately chosen likelihood function, assume explicit knowledge of the joint probability distribution (PD) of the clean speech signal and the noise process. Such statistical knowledge is required in the above examples for evaluating the conditional expected value of the clean speech or of its sample spectrum, given the noisy speech, which are the optimal MMSE estimators of the speech waveform and its sample spectrum, respectively [2]. For the case considered in this paper where the speech is degraded by statistically independent additive noise, the PD's of the clean speech signal and the noise process must be known.

In practice, the PD of the speech signal is not known, and the PD of the noise process is rarely available. Instead, the statistics of the sources are implicitly given in terms of training sequences generated by the speech signal and the noise process. While such training sequences can be directly applied to optimally solve the specific estimation problem of interest simply by replacing statistical expectations with sample averages, this requires a substantial amount of memory and computation resources. This is due to the necessity of storing the two training sequences and reapplying them in estimating *each* vector of the clean speech [1]. An alternative approach is to first estimate the unknown PD's of the speech signal and the noise process from the given training sequences, and then to use the estimated PD's for constructing the desired signal estimators. This two-step approach is usually suboptimal, but has proven useful in the speech enhancement application considered here. In fact, this approach is not new in the area of speech processing, as it has been extensively used in speech recognition applications where explicit knowledge of the PD's of the acoustic signals is required in order to apply the likelihood ratio test.

For the two-step enhancement approach to be tractable, compact, but reliable, estimates of the PD's of the speech signal and the noise process must be obtained. This can be accomplished by modeling the sample PD of each process by a parametric PD which depends on a much smaller number of parameters than the number of samples in the training sequence itself. A useful class of models for speech signals is that of Markov sources or hidden Markov models (HMM's) [3]–[5]. These models assume that the PD of any vector of speech samples, at a given time instant, is parametric and is determined by the state at which the process is assumed to be in at that time. Furthermore, the transition from one state to another is Markovian of a given order, usually one. Such models have proven very successful in speech recognition applications

Manuscript received January 19, 1988; revised February 18, 1989.

Y. Ephraim and B.-H. Juang are with the Speech Research Department, AT&T Bell Laboratories, Murray Hill, NJ 07974.

D. Malah is with the Signal Processing Department, AT&T Bell Laboratories, Murray Hill, NJ 07974, on leave from the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel.

IEEE Log Number 8931331.

(see, e.g., [6]–[8]). Here, we shall use HMM's with mixtures of Gaussian AR output PD's to model the PD of the clean speech signal. The appropriate model for the noise process depends on the nature of the noise. Since in this paper we shall be concerned only with Gaussian noise with a theoretically flat power spectral density, a low-order AR model will be used.

The parameter set of the HMM for the speech signal is estimated by the maximum likelihood (ML) approach using the Baum [8]–[12] or the EM (estimation–maximization) reestimation algorithm [13]–[14]. This algorithm locally maximizes the probability density function (pdf) of the model observation sequence for the given training sequence of clean speech. The algorithm starts from an initial HMM, and iteratively generates a sequence of HMM's with nondecreasing likelihood values by maximizing in each iteration the so-called auxiliary function. An efficient approximation to the Baum algorithm, which will also be examined here, is given by the segmental k-means algorithm [15]. This algorithm locally maximizes the joint pdf of the state and observation sequences of the model for the given training sequence. This is done by alternate maximization of the joint pdf once over all sequences of states assuming a parameter set of the model is given, and then over all parameter sets assuming that the most likely sequence of state is available. The estimation of the most likely sequence of states is done by the Viterbi algorithm [16], and the estimation of the parameter set of the model is done by the Baum reestimation formulas [8]–[12]. The two algorithms perform similarly when there is a unique sequence of states which dominates the likelihood function of the HMM. The parameter set of the model for the noise process is obtained by AR modeling of the centroid covariance matrix of the training sequence from the noise process.

Given the parameter sets of the speech and noise models, a maximum *a posteriori* (MAP) approach for estimating the clean speech waveform is developed. The MAP enhancement algorithm is a straightforward application of the EM algorithm for reestimation of the clean speech waveform given the noisy speech [17]. The algorithm locally maximizes the conditional pdf of the clean speech given the noisy speech. The algorithm starts from the given noisy speech and generates a sequence of speech sample functions with nondecreasing likelihood values by maximizing in each iteration an appropriately defined auxiliary function. Note that this enhancement algorithm is consistent with the ML training procedure obtained by using the Baum algorithm.

An approximate MAP approach, which is consistent with the segmental k-means training procedure, is also developed and examined here. This algorithm locally maximizes the joint conditional pdf of the state and observation sequences associated with the clean speech, given the noisy speech. This is done by alternate maximization of the conditional pdf once over all state sequences assuming that the clean speech vectors are given, and then over the clean speech vectors assuming that the most likely state sequence is available. The estimation of the most likely sequence of states is done by applying the Viterbi algorithm to the current estimate of the clean speech signal. The estimation of the clean speech vectors is done by time-varying Wiener filtering of the noisy speech using the AR covariance matrices of the HMM associated with the most likely sequence of states and the AR covariance matrix of the noise model.

The approximate MAP speech enhancement approach is similar to the approach proposed by Lim and Oppenheim [18]. Both algorithms alternately estimate the original speech signal, and its power spectral density as represented by an AR model, for any vector of the speech signal. The two algorithms differ, however, in two major aspects. First, the enhancement in [18] is done on a frame-by-frame basis where estimation in one frame does not affect the estimation in adjacent frames. Here, the enhancement of the speech signal in a given time interval (which consists of several frames) is done simultaneously and estimates in adjacent frames are dependent due to the Markovian property of the model. The second difference between the two approaches is that in [18], any AR model for the current estimate of the speech signal can be chosen, while here, the estimate of the AR model is constrained to the finite set of predesigned AR models.

The paper is organized as follows. In Section II, we formulate the problem and specify the statistical models used here. In Section III, we describe the training procedures. In Section IV, we describe the enhancement algorithms. In Section V, we provide experimental results. Comments are given in Section VI.

II. Problem Formulation

A. HMM's for Clean Speech

Let p_{λ_s} be the pdf of an HMM for the clean speech signal where λ_s denotes the parameter set of the model. We consider HMM's with M states and mixtures of L Gaussian AR output processes at each state. Let $y \triangleq \{ y_t, t = 0, \cdots, T \}$, $y_t \in R^K$, be a sequence of K-dimensional vectors which represent the output from the model. Let $x \triangleq \{x_t, t = 0, \cdots, T\}$, $x_t \in \{1, \cdots, M\}$, be a sequence of states corresponding to y. Let $h \triangleq \{h_t, t = 0, \cdots, T\}$, $h_t \in \{1, \cdots, L\}$, be a sequence of mixture components corresponding to (x, y). The pdf p_{λ_s} is given by

$$p_{\lambda_s}(y) = \sum_x \sum_h p_{\lambda_s}(x, h, y)$$
$$= \sum_x \sum_h p_{\lambda_s}(x)\, p_{\lambda_s}(h|x)\, p_{\lambda_s}(y|h, x) \quad (1)$$

where $p_{\lambda_s}(x)$ is the probability of the sequence of states x, $p_{\lambda_s}(h|x)$ is the probability of the sequence of mixture components h given the sequence of states x, and $p_{\lambda_s}(y|h, x)$ is the pdf of the output sequence y given $\{x, h\}$. The probability $p_{\lambda_s}(x)$ is given by

$$p_{\lambda_s}(x) = \prod_{t=0}^{T} a_{x_{t-1}x_t} \quad (2)$$

where $a_{x_{t-1}x_t}$ denotes the transition probability from state x_{t-1} at time $t-1$ to state x_t at time t, and $a_{x_{-1}x_0} \triangleq \pi_{x_0}$ denotes the probability of the initial state x_0. For $p_{\lambda_s}(h|x)$ and $p_{\lambda_s}(y|h, x)$, we make the following standard assumptions:

$$p_{\lambda_s}(h|x) = \prod_{t=0}^{T} p_{\lambda_s}(h_t|x_t) \triangleq \prod_{t=0}^{T} c_{h_t|x_t} \tag{3}$$

and

$$p_{\lambda_s}(y|h, x) = \prod_{t=0}^{T} p_{\lambda_s}(y_t|h_t, x_t) \triangleq \prod_{t=0}^{T} b(y_t|h_t, x_t) \tag{4}$$

where $c_{h_t|x_t}$ is the probability of choosing the mixture h_t given that the process is in states x_t, and $b(y_t|h_t, x_t)$ is the pdf of the output vector y_t given (h_t, x_t). For zero-mean N_sth-order Gaussian AR output processes, we have

$$b(y_t|h_t = \gamma, x_t = \beta) = \frac{\exp\left\{-\frac{1}{2}y_t^{\#}S_{\gamma|\beta}^{-1}y_t\right\}}{(2\pi)^{K/2}\det^{1/2}(S_{\gamma|\beta})} \tag{5}$$

where $\#$ denotes vector transpose, $S_{\gamma|\beta} = \sigma_{\gamma|\beta}^2(A_{\gamma|\beta}^{\#}A_{\gamma|\beta})^{-1}$, $\sigma_{\gamma|\beta}^2$ 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 $N_s + 1$ elements of the first column constitute the coefficients of the AR process, $g_{\gamma|\beta} \triangleq (g_{\gamma|\beta}(0), g_{\gamma|\beta}(1), \cdots, g_{\gamma|\beta}(N_s))$, $g_{\gamma|\beta}(0) = 1$.

The modeling problem is that of estimating the parameter set $\lambda_s = (\pi, a, c, S)$ where $\pi \triangleq \{\pi_\beta\}$, $a \triangleq \{a_{\alpha\beta}\}$, $c \triangleq \{c_{\gamma|\beta}\}$, and $S \triangleq \{S_{\gamma|\beta}\}$ for $\alpha, \beta = 1, \cdots, M$ and $\gamma = 1, \cdots, L$, given a training sequence y from the speech signal. An ML estimate of the parameter set λ_s is obtained from

$$\max_{\lambda_s} \ln p_{\lambda_s}(y) = \max_{\lambda_s} \ln \sum_x \sum_h p_{\lambda_s}(x, h, y), \tag{6}$$

where the maximization is locally performed by the Baum reestimation algorithm [8]–[12]. The segmental k-means algorithm for estimating the parameter set of the model assumes that the double sum in (6) is dominated by a unique sequence of states and mixture components. Hence, the parameter set of the model is estimated along with the most likely sequence of states and mixture components by

$$\max_{x,h,\lambda_s} \ln p_{\lambda_s}(x, h, y). \tag{7}$$

The two estimation procedures are described in Subsections A and B of Section III.

B. AR Model for the Noise Process

Let p_{λ_v} be the pdf of the model for the noise process where λ_v is the parameter set of the model. For the Gaussian noise considered in this paper, we assume that

$$p_{\lambda_v}(v) = \prod_{t=0}^{T} p_{\lambda_v}(v_t) = \prod_{t=0}^{T} \frac{\exp\left\{-\frac{1}{2}v_t^{\#}V^{-1}v_t\right\}}{(2\pi)^{K/2}\det^{1/2}(V)} \tag{8}$$

where $v \triangleq \{v_t, t = 0, \cdots, T\}$, $v_t \in R^K$, is a sequence of $T+1$ K-dimensional output vectors and V is an N_vth-order AR covariance matrix. $V = \sigma_v^2(A_v^{\#}A_v)^{-1}$ where σ_v^2 and A_v are defined similarly to $\sigma_{\gamma|\beta}^2$ and $A_{\gamma|\beta}$, respectively. A_v is a $K \times K$ lower triangular Toeplitz matrix in which the first $N_v + 1$ elements of the first column constitute the coefficients of the AR process, $g_v \triangleq (g_v(0), g_v(1), \cdots, g_v(N_v))$, $g_v(0) = 1$.

The noise modeling problem is that of finding the parameter set $\lambda_v \triangleq (\sigma_v^2, g_v(m), m = 1, \cdots, N_v)$ given a training sequence v from the noise process. An ML estimate of λ_v is obtained from

$$\max_{\lambda_v} \ln p_{\lambda_v}(v), \tag{9}$$

and this maximization is equivalent to AR modeling of the centroid covariance matrix of the noise training sequence. The estimation of the noise model is discussed in Section III-C.

C. Speech Enhancement Problem

Given the parameter set λ_s of an HMM for the clean speech signal, the parameter set λ_v for the AR model for the noise process, and a sequence of K-dimensional noisy vectors $z \triangleq \{z_t, t = 0, \cdots, T\}$, $z_t = y_t + v_t$, the enhancement problem considered here is that of estimating the sequence y of clean speech vectors by the MAP estimation approach as follows:

$$\max_y \ln p_{\lambda_s\lambda_v}(y, z) = \max_y \ln \sum_x \sum_h p_{\lambda_s\lambda_v}(x, h, y, z) \tag{10}$$

where

$$p_{\lambda_s\lambda_v}(y, z) = p_{\lambda_s}(y)\, p_{\lambda_v}(z|y) = p_{\lambda_s}(y)\, p_{\lambda_v}(z - y) \tag{11}$$

due to the fact that the noise is additive and statistically independent of the signal, and

$$\begin{aligned} p_{\lambda_s\lambda_v}(x, h, y, z) &= p_{\lambda_s\lambda_v}(z|x, h, y)\, p_{\lambda_s}(x, h, y) \\ &= p_{\lambda_v}(z|y)\, p_{\lambda_s}(x, h, y) \\ &= p_{\lambda_v}(z - y)\, p_{\lambda_s}(x, h, y) \end{aligned} \tag{12}$$

due to the fact that given y, z and (x, h) are statistically independent. Note that since $p_{\lambda_s\lambda_v}(z) = \int p_{\lambda_s\lambda_v}(y, z)\, dy$ is independent of y, the problem (10) is equivalent to

$$\max_y \ln p_{\lambda_s\lambda_v}(y|z), \tag{13}$$

where $p_{\lambda_s\lambda_v}(y|z) = p_{\lambda_s\lambda_v}(y, z)/p_{\lambda_s\lambda_v}(z)$. Furthermore, the enhancement problem stated in (13) is consistent with the ML training procedure described in (6). In the approximate MAP enhancement procedure which is consistent with the segmental k-means training algorithm (7), it is assumed that the double sum in (10) is dominated by a unique sequence of states and mixture components. Hence, the sequence of clean speech vectors is estimated along with the most likely sequence of states and mixture components by

$$\max_{x,h,y} \ln p_{\lambda_s\lambda_v}(x, h, y, z). \tag{14}$$

Similarly to (13), the problem in (14) is equivalent to

$$\max_{x,h,y} \ln p_{\lambda_s\lambda_v}(x, h, y|z). \tag{15}$$

The two MAP enhancement procedures (13) and (15) are described in Section IV.

III. Training of Speech and Noise Models

The formulation of the speech modeling problem as given in Section II considers the estimation of the parameter set of the model from a single training sequence of speech. In this paper, however, multiple training sequences which are assumed to be statistically independent have been used. Hence, we provide the algorithms for the more general case of modeling using N training sequences of speech. Let $y_{T_n} \triangleq \{y_{t,n}, t = 0, \cdots, T_n\}$ be a sequence of $T_n + 1$ K-dimensional vectors, and let $y \triangleq \{y_{T_n}, n = 1, \cdots, N\}$ be the set of N such sequences. Let $x_{T_n} \triangleq \{x_{t,n}, t = 0, \cdots, T_n\}$ and $h_{T_n} \triangleq \{h_{t,n}, t = 0, \cdots, T_n\}$ be, respectively, the sequences of states and mixture components corresponding to the nth utterance of the training sequence. Finally, let $x \triangleq \{x_{T_n}, n = 1, \cdots, N\}$ and $h \triangleq \{h_{T_n}, n = 1, \cdots, N\}$.

A. Baum Reestimation Algorithm

The likelihood function to be maximized is given by

$$\ln p_{\lambda_s}(y) = \sum_{n=1}^{N} \ln p_{\lambda_s}(y_{T_n}). \tag{16}$$

Local maximization of (16) can be achieved by the Baum reestimation algorithm [8]–[12]. This algorithm generates a sequence of HMM's with nondecreasing likelihood values (16). Each iteration of the Baum algorithm starts with an old set of parameters, say λ_s, and estimates a new set of parameters, say λ_s', by maximizing the following auxiliary function:

$$\phi(\lambda_s') = \sum_{n=1}^{N} \sum_{x_{T_n}} \sum_{h_{T_n}} p_{\lambda_s}(x_{T_n}, h_{T_n}|y_{T_n}) \ln p_{\lambda_s'}(x_{T_n}, h_{T_n}, y_{T_n}) \tag{17}$$

over λ_s', subject to the constraints $\pi_\beta' \geq 0$, $\sum_{\beta=1}^{M} \pi_\beta' = 1$, $a_{\alpha\beta}' \geq 0$, $\sum_{\beta=1}^{M} a_{\alpha\beta}' = 1$, $c_{\gamma|\beta}' \geq 0$, $\sum_{\gamma=1}^{L} c_{\gamma|\beta}' = 1$, and $S_{\gamma|\beta}' = \sigma_{\gamma|\beta}^2 (A_{\gamma|\beta}^{\#} A_{\gamma|\beta})^{-1}$ for $\alpha, \beta = 1, \cdots, M$ and $\gamma = 1, \cdots, L$. The algorithm is stopped when a convergence criterion is satisfied, e.g., when the difference of the values of the likelihood function (16) in two consecutive iterations is smaller than or equal to a given threshold. Convergence of the model sequence generated by the Baum algorithm was discussed in [14].

The above constrained maximization of the auxiliary function results in the following reestimation formulas:

$$\pi_\beta' = \frac{1}{N} \sum_{n=1}^{N} \sum_{\gamma=1}^{L} q_{0,n}(\beta, \gamma) \tag{18}$$

$$a_{\alpha\beta}' = \frac{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \sum_{\gamma=1}^{L} q_{t,n}(\alpha, \beta, \gamma)}{\sum_{\beta=1}^{M} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \sum_{\gamma=1}^{L} q_{t,n}(\alpha, \beta, \gamma)} \tag{19}$$

$$c_{\gamma|\beta}' = \frac{\sum_{n=1}^{N} \sum_{t=0}^{T_n} q_{t,n}(\beta, \gamma)}{\sum_{\gamma=1}^{L} \sum_{n=1}^{N} \sum_{t=0}^{T_n} q_{t,n}(\beta, \gamma)} \tag{20}$$

and the parameters of the AR output PD's are obtained from

$$\min_{S_{\gamma|\beta}} \{\operatorname{tr} R_{\gamma|\beta}' S_{\gamma|\beta}^{-1} - \ln \det S_{\gamma|\beta}^{-1}\}$$

$$R_{\gamma|\beta}' \triangleq \frac{\sum_{n=1}^{N} \sum_{t=0}^{T_n} q_{t,n}(\beta, \gamma) y_{t,n} y_{t,n}^{\#}}{\sum_{n=1}^{N} \sum_{t=0}^{T_n} q_{t,n}(\beta, \gamma)} \tag{21}$$

where

$$q_{t,n}(\alpha, \beta, \gamma) \triangleq \frac{\sum_{\{x_{T_n}: x_{t-1,n}=\alpha,\ x_{t,n}=\beta\}} \sum_{\{h_{T_n}: h_{t,n}=\gamma\}} p_{\lambda_s}(x_{T_n}, h_{T_n}, y_{T_n})}{\sum_{x_{T_n}} \sum_{h_{T_n}} p_{\lambda_s}(x_{T_n}, h_{T_n}, y_{T_n})}, \quad 0 < t \leq T_n \tag{22}$$

is the conditional probability, under p_{λ_s}, of being in state α at time $t-1$, in state β at time t, and choosing mixture component γ while in state β, given the nth utterance of the speech training sequence, and

$$q_{t,n}(\beta, \gamma) \triangleq \frac{\sum_{\{x_{T_n}: x_{t,n}=\beta\}} \sum_{\{h_{T_n}: h_{t,n}=\gamma\}} p_{\lambda_s}(x_{T_n}, h_{T_n}, y_{T_n})}{\sum_{x_{T_n}} \sum_{h_{T_n}} p_{\lambda_s}(x_{T_n}, h_{T_n}, y_{T_n})}, \quad 0 \leq t \leq T_n \tag{23}$$

is the conditional probability, under p_{λ_s}, of being in state β at time t and choosing mixture component γ while in state β, given the nth utterance of the speech training sequence. The reestimation formulas (19)–(21) are valid provided that the terms in the denominators of these expressions are greater than zero. If any of these conditions is not satisfied, then the affected reestimated parameter can be arbitrarily chosen up to the constraints associated with the problem (17), without affecting the

likelihood value. For example, if the denominator of (19) equals zero for a particular α, then any $a'_{\alpha\beta}$, $\beta = 1, \cdots, M$, which satisfies $\sum_{\beta=1}^{M} a'_{\alpha\beta} = 1$ can be chosen.

The probability measures $q_{t,n}(\alpha, \beta, \gamma)$ and $q_{t,n}(\beta, \gamma)$ can be efficiently calculated using the forward–backward formulas as follows:

$$q_{t,n}(\alpha, \beta, \gamma) = \frac{\sum_{\xi=1}^{L} F_{t-1,n}(\alpha, \xi)\, B_{t,n}(\beta, \gamma)\, a_{\alpha\beta} c_{\gamma|\beta}\, b(y_{t,n}|\gamma, \beta)}{\sum_{\alpha,\beta=1}^{M} \sum_{\xi,\gamma=1}^{L} F_{t-1,n}(\alpha, \xi)\, B_{t,n}(\beta, \gamma)\, a_{\alpha\beta} c_{\gamma|\beta}\, b(y_{t,n}|\gamma, \beta)}, \quad 0 < t \le T_n \tag{24}$$

$$q_{t,n}(\beta, \gamma) = \frac{F_{t,n}(\beta, \gamma)\, B_{t,n}(\beta, \gamma)}{\sum_{\beta=1}^{M} \sum_{\gamma=1}^{L} F_{t,n}(\beta, \gamma)\, B_{t,n}(\beta, \gamma)}, \quad 0 \le t \le T_n \tag{25}$$

where

$$F_{0,n}(\alpha, \gamma) = \pi_\alpha c_{\gamma|\alpha} b(y_{0,n}|\gamma, \alpha)$$

$$F_{t,n}(\alpha, \gamma) = \sum_{\nu=1}^{M} \sum_{\mu=1}^{L} F_{t-1,n}(\nu, \mu)\, a_{\nu\alpha} c_{\gamma|\alpha} b(y_{t,n}|\gamma, \alpha), \quad 0 < t \le T_n \tag{26}$$

$$B_{T_n}(\beta, \gamma) = 1$$

$$B_{t,n}(\beta, \gamma) = \sum_{\nu=1}^{M} \sum_{\mu=1}^{L} B_{t+1,n}(\nu, \mu) \cdot a_{\beta\nu} c_{\mu|\nu} b(y_{t+1,n}|\mu, \nu), \quad 0 \le t < T_n. \tag{27}$$

The minimization problem in (21) has a unique solution provided that $R'_{\gamma|\beta}$ is positive definite [19, Theorem 2], [20]. The minimizing AR parameter set can be found by AR modeling of $R'_{\gamma|\beta}$ using a variant of the covariance method of linear prediction [21, Corollary 2], [22, p. 14]. An approximate solution can be obtained by the autocorrelation method of linear prediction if end-block effects are neglected, more specifically, if the Kth-order vector $A_{\gamma|\beta} y_{t,n}$ is considered as the convolution result of $y_{t,n}$ with $g_{\gamma|\beta}$. The two vectors are identical in their first K elements, but the vector which results from convolution of $y_{t,n}$ with $g_{\gamma|\beta}$ has $N_s - 1$ additional elements. This approximation was found reasonable for the values of $K = 128$ and $N_s = 10$ used here. In this case, it can be shown by substituting $R'_{\gamma|\beta}$ and $S_{\gamma|\beta}$ into (21) that the AR parameter set is obtained from AR modeling of the autocorrelation function given by

$$r'_{\gamma|\beta}(m) \triangleq \frac{\sum_{n=1}^{N} \sum_{t=0}^{T_n} q_{t,n}(\beta, \gamma)\, r_{t,n}(m)}{\sum_{n=1}^{N} \sum_{t=0}^{T_n} q_{t,n}(\beta, \gamma)} \tag{28}$$

where

$$r_{t,n}(m) = \frac{1}{K} \sum_{k=0}^{K-|m|-1} y_{t,n}(k)\, y_{t,n}(k + |m|), \quad m = -N_s, \cdots, N_s.$$

The likelihood function (16), which has to be evaluated in checking convergence of the Baum algorithm, can be efficiently calculated similarly to the calculation of the denominator of (23) using the forward–backward formulas, as is shown in (25).

B. Segmental k-Means Algorithm

The likelihood function to be maximized in this case is given by

$$\ln p_{\lambda_s}(x, h, y) = \sum_{n=1}^{N} \ln p_{\lambda_s}(x_{T_n}, h_{T_n}, y_{T_n}) \tag{29}$$

and the maximization is performed over (x, h) and λ_s for a given y. This is done by alternate maximization of the likelihood function once over (x, h) assuming λ_s is given, and then over λ_s assuming that the most likely sequence of states and mixture components, say (x^*, h^*), is available. Thus, if each iteration comprises the estimation of (x, h) for a given λ_s and the estimation of a new λ_s based upon (x^*, h^*), then this training algorithm generates a sequence of models with nondecreasing likelihood. The procedure is stopped when a convergence criterion is satisfied, e.g., when the difference of the values of the likelihood function (29) in two consecutive iterations is smaller than or equal to a given threshold. Convergence of the model sequence generated by this algorithm was considered in [23] using standard arguments from optimization theory [14], [24, p. 187], [25].

The maximization of (29) over (x, h) is achieved by applying the Viterbi algorithm for each utterance of the training sequence independently, using the following path metric:

$$\ln \pi_\beta + \ln c_{\gamma|\beta} + \ln b(y_{0,n}|h_{0,n} = \gamma, x_{0,n} = \beta)$$

for $t = 0$ and

$$\ln a_{\alpha\beta} + \ln c_{\gamma|\beta} + \ln b(y_{t,n}|h_{t,n} = \gamma, x_{t,n} = \beta) \tag{30}$$

for $1 \le t \le T_n$ where $\alpha, \beta = 1, \cdots, M$ and $\gamma = 1, \cdots, L$. Given the most likely sequence (x^*, h^*), a new parameter set is obtained from maximization of the auxiliary function

$$\sum_{n=1}^{N} \sum_{x_{T_n}} \sum_{h_{T_n}} \delta(x_{T_n} - x^*_{T_n}, h_{T_n} - h^*_{T_n}) \ln p_{\lambda'_s}(x_{T_n}, h_{T_n}, y_{T_n}) \tag{31}$$

over λ'_s, subject to the constraints associated with (17), where $\delta(\cdot)$ denotes a Dirac function. Comparing (17) and (31) shows that this maximization can be performed using

exactly the same reestimation formulas (18)–(21) with $q_{t,n}(\alpha, \beta, \gamma)$ and $q_{t,n}(\beta, \gamma)$ being replaced by

$$q_{t,n}(\alpha, \beta, \gamma) \triangleq \sum_{\left\{x_{T_n}: \substack{x_{t-1,n}=\alpha \\ x_{t,n}=\beta}\right\}} \sum_{\left\{h_{T_n}: h_{t,n}=\gamma\right\}} \cdot \delta(x_{T_n} - x^*_{T_n}, h_{T_n} - h^*_{T_n}) = \begin{cases} 1 & \alpha = x^*_{t-1,n}, \beta = x^*_{t,n}, \gamma = h^*_{t,n} \\ 0 & \text{otherwise,} \end{cases} \quad 0 < t \le T_n. \quad (32)$$

$$q_{t,n}(\beta, \gamma) \triangleq \sum_{\{x_{T_n}: x_{t,n}=\beta\}} \sum_{\{h_{T_n}: h_{t,n}=\gamma\}} \cdot \delta(x_{T_n} - x^*_{T_n}, h_{T_n} - h^*_{T_n}) = \begin{cases} 1 & \beta = x^*_{t,n}, \gamma = h^*_{t,n} \\ 0 & \text{otherwise,} \end{cases} \quad 0 \le t \le T_n. \quad (33)$$

C. Noise Model Estimation

The estimation problem of the parameter set of the AR model for the noise process results from substituting (8) into (9). This problem is equivalent to

$$\min_V \{\text{tr } R_v V^{-1} - \ln \det V^{-1}\} \quad (34)$$

where

$$R_v \triangleq \frac{1}{T+1} \sum_{t=0}^{T} v_t v_t^{\#}.$$

This problem is similar to that associated with the estimation of the parameter set of each AR output process of the HMM. An approximation solution is obtained from AR modeling of the autocorrelation function given by

$$r_v(m) \triangleq \frac{1}{T+1} \sum_{t=0}^{T} \frac{1}{K} \sum_{k=0}^{K-|m|-1} v_t(k)\, v_t(k + |m|), \quad m = -N_v, \cdots, N_v. \quad (35)$$

D. Implementation Issues

The segmental k-means algorithm is initialized by a parameter set λ_s obtained from AR model vector quantization of the given training sequences. The model designed by the segmental k-means algorithm is used as the initial model for the Baum algorithm. The vector quantizer (VQ) is designed for the Itakura–Saito distortion measure using the standard generalized Lloyd algorithm [27]. Initially, an M-entry code book is designed by successive splitting of codewords corresponding to lower order code books, starting from the centroid of the training sequences. The splitting of only one codeword, that with the largest residual energy, is applied in each increase of a code book order. Thus, the tree design approach of [27] is used, but with a nonbinary or a pruned tree. This approach has the advantage that the number of states need not be a power of two. Once the code book representing the states' codewords has been designed, the training sequences are clustered using this code book. The vectors in each cluster are then used for designing the codewords representing the mixture components of that state using the same splitting strategy. The resulting $M \times L$ code book is used as the initial parameter set for the output AR processes of the HMM. The initial estimate of (π, a, c) is obtained from decoding of the training sequences using the designed VQ and estimating the frequency at which each initial state is used, each state transition occurs, and each mixture is chosen for each state.

In implementing the Baum algorithm, a proper scaling of the forward and backward probabilities is applied. In particular, the recursive scaling proposed in [28], in which $F_{t,n}(\alpha, \gamma)$ is normalized by $\Sigma_{\alpha,\gamma} F_{t,n}(\alpha, \gamma)$ and $B_{t,n}(\beta, \gamma)$ is normalized by $\Sigma_{\beta,\gamma} B_{t,n}(\beta, \gamma)$ for each (t, n), is used. Furthermore, the calculation of each summand of the forward-backward formulas and of the probabilities $q_{t,n}(\alpha, \beta, \gamma)$ in (24) is done in the log domain due to the relatively small numbers involved. In this case, the values of $\ln b(y_t | h_t = \gamma, x_t = \beta)$ for each (t, n) are shifted into the dynamic range of the computer prior to their summation, simply by subtracting $\max_{\gamma,\beta} \ln b(y_t | h_t = \gamma, x_t = \beta) - \ln D$ where D is the largest number which can be represented on the computer, from each term. This shift is compensated for automatically by the above scaling and it has no effect on the reestimation formulas. The evaluation of the likelihood (16), however, is affected by both the scaling and the shifting, and hence, the likelihood calculated as the denominator of (25) has to be modified appropriately.

IV. Speech Enhancement Algorithm

In this section, we first apply the EM algorithm for MAP estimation of the clean speech signal given the noisy speech. Then we present an approximate MAP approach in which the enhancement is performed based upon the most likely sequence of states and mixture components.

A. EM Algorithm

Let z be a given sequence of $T + 1$ K-dimensional vectors of noisy speech. Let $\lambda \triangleq (\lambda_s, \lambda_v)$. Let $y(k) \triangleq \{y_t(k), t = 0, \cdots, T\}$, $y_t(k) \in R^K$, be a current estimate of the speech signal. Similarly, let $y(k+1)$ be a new estimate of the speech signal. Using Jensen's inequality and the fact that given y, (x, h) and z are statistically independent, we have that

$$\begin{aligned} &\ln p_\lambda(y(k+1)|z) - \ln p_\lambda(y(k)|z) \\ &= \ln \sum_{x,h} \frac{p_\lambda(x, h, y(k)|z)}{p_\lambda(y(k)|z)} \frac{p_\lambda(x, h, y(k+1)|z)}{p_\lambda(x, h, y(k)|z)} \\ &= \ln \sum_{x,h} p_\lambda(x, h | y(k)) \frac{p_\lambda(x, h, y(k+1)|z)}{p_\lambda(x, h, y(k)|z)} \\ &\ge \sum_{x,h} p_\lambda(x, h | y(k)) \ln \frac{p_\lambda(x, h, y(k+1)|z)}{p_\lambda(x, h, y(k)|z)} \\ &\triangleq \phi(y(k+1)) - \phi(y(k)) \end{aligned} \quad (36)$$

where

$$\phi(y(k+1)) \triangleq \sum_{x,h} p_\lambda(x, h \mid y(k)) \cdot \ln p_\lambda(x, h, y(k+1) \mid z). \quad (37)$$

Hence, maximization of $\phi(y(k+1))$ over $y(k+1)$ results in $\ln p_\lambda(y(k+1)\mid z) \geq \ln p_\lambda(y(k)\mid z)$ where equality holds if and only if $y(k+1) = y(k)$ almost everywhere $p_\lambda(x, h \mid y(k))$. This standard argument of the EM algorithm implies that a MAP estimate of the speech waveform can be achieved by reestimation of the speech waveform through the maximization of the auxiliary function $\phi(y(k+1))$.

On substituting (1)–(5), (8), and (12) into (37), and setting the gradient of $\phi(y(k+1))$ with respect to $y(k+1)$ to zero, we obtain the following reestimation formula for the clean speech signal:

$$y_t(k+1) = \left[\sum_{\beta,\gamma} q_t(\beta, \gamma \mid y(k)) H_{\gamma\mid\beta}^{-1}\right]^{-1} z_t, \quad 0 \leq t \leq T \quad (38)$$

where $q_t(\beta, \gamma \mid y(k))$ is defined similarly to (23) with y_{T_m} being replaced by $y(k)$, and $H_{\gamma\mid\beta}$ is a Wiener filter for the output Gaussian process from state β and mixture γ and the Gaussian noise process (8):

$$H_{\gamma\mid\beta} \triangleq S_{\gamma\mid\beta}(S_{\gamma\mid\beta} + V)^{-1}. \quad (39)$$

The probability measure $q_t(\beta, \gamma \mid y(k))$ is calculated by (25) using the forward–backward formulas (26)–(27).

The estimate $y_t(k+1)$ can be efficiently implemented in the frequency domain if $S_{\gamma\mid\beta}$ and V are approximated by their asymptotically equivalent circulant covariance matrices. Since the two matrices are covariance matrices of AR processes, such an approximation is always possible provided that $|A_{\gamma\mid\beta}(\theta)|^2 \geq m > 0$ and $|A_v(\theta)|^2 \geq m > 0$ for some m where $A_{\gamma\mid\beta}(\theta)$ and $A_v(\theta)$ are the Fourier transforms of $g_{\gamma\mid\beta}$ and g_v, respectively [26]. Let

$$f_{\gamma\mid\beta}(\theta) \triangleq \sigma_{\gamma\mid\beta}^2 / \left|A_{\gamma\mid\beta}(\theta)\right|^2 \quad (40)$$

and

$$f_v(\theta) \triangleq \sigma_v^2 / \left|A_v(\theta)\right|^2 \quad (41)$$

be the asymptotic power spectral densities associated with the two AR processes. Then

$$S_{\gamma\mid\beta} \sim C(f_{\gamma\mid\beta}(\theta)), \quad V \sim C(f_v(\theta)) \quad (42)$$

where $C(f_{\gamma\mid\beta}(\theta))$ and $C(f_v(\theta))$ are the asymptotically equivalent circulant covariance matrices of $S_{\gamma\mid\beta}$ and V, respectively. Using some basic properties of circulant matrices and their inverses [26], we have that

$$S_{\gamma\mid\beta}(S_{\gamma\mid\beta} + V)^{-1} \sim C(H_{\gamma\mid\beta}(\theta)) \quad (43)$$

where

$$H_{\gamma\mid\beta}(\theta) \triangleq \frac{f_{\gamma\mid\beta}(\theta)}{f_{\gamma\mid\beta}(\theta) + f_v(\theta)}. \quad (44)$$

Hence,

$$y_{t,\theta}(k+1) \approx \left[\sum_{\beta,\gamma} q_t(\beta, \gamma \mid y(k)) H_{\gamma\mid\beta}^{-1}(\theta)\right]^{-1} z_{t,\theta} \quad (45)$$

where $y_{t,\theta}(k+1)$ and $z_{t,\theta}$ are the Fourier transforms of $y_t(k+1)$ and z_t, respectively.

B. Approximate MAP Algorithm

The approximate MAP enhancement algorithm results from alternate maximization of $\ln p_{\lambda_s\lambda_v}(x, h, y, z)$, defined in (12), once over (x, h) assuming that y is given, and then over y assuming that (x, h) is available. Given an estimate of the clean speech signal, say $y(k)$, the estimation of the most likely sequence of states and mixture components is done by applying the Viterbi algorithm using the path metric

$$\ln \pi_\beta + \ln c_{\gamma\mid\beta} + \ln b(y_0(k) \mid h_0 = \gamma, x_0 = \beta) + \ln p_{\lambda_v}(z_0 - y_0(k)) \quad (46a)$$

for $t = 0$ and

$$\ln a_{\alpha\beta} + \ln c_{\gamma\mid\beta} + \ln b(y_t(k) \mid h_t = \gamma, x_t = \beta) + \ln p_{\lambda_v}(z_t - y_t(k)) \quad (46b)$$

for $1 \leq t \leq T$ where $\alpha, \beta = 1, \cdots, M$ and $\gamma = 1, \cdots, L$. Note that since the last term in each of (46a) and (46b) is independent of the states and the mixture components, these terms can be ignored in performing the Viterbi decoding. Let the resulting sequence of states and mixture components be denoted by (x^*, h^*). Given (x^*, h^*), a new estimate of the speech signal, say $\{y_t(k+1)\}$, is obtained from

$$\max_{y_t} \left\{\ln b(y_t \mid x_t^*, h_t^*) + \ln p_{\lambda_v}(z_t - y_t)\right\} \quad (47)$$

for all $0 \leq t \leq T$. On substituting (5) and (8) into (47) and setting the gradient of the resulting function with respect to y_t to zero, we get

$$y_t(k+1) = S_{h_t^*\mid x_t^*}(S_{h_t^*\mid x_t^*} + V)^{-1} z_t \quad (48)$$

which is equivalent to Wiener filtering of the noisy speech using the covariance matrix of the AR process corresponding to the most probable state and mixture component at time t and the stationary covariance matrix of the noise. Equation (48) can be efficiently implemented in the frequency domain similarly to (38). The entire algorithm which comprises alternate application of (46) and (48) is described in Fig. 1.

The reestimation algorithms (45) and (48) are initialized from the noisy speech, i.e., $y(0) = z$, and the algorithms are stopped when a convergence criterion similar to that used in Section II is satisfied. The convergence of these algorithms will not be discussed here as convergence can be shown using standard arguments from optimization theory, in particular, the Global Convergence Theorem [24, p. 187].

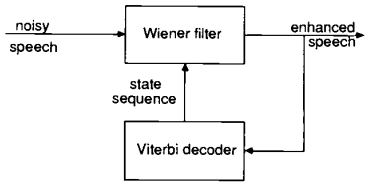

Fig. 1. A block diagram for the approximate MAP enhancement approach.

V. Experimental Results

The speech enhancement approach described in this paper was examined in enhancing speech signals which have been degraded by statistically independent additive Gaussian white noise at signal-to-noise ratio (SNR) values of 5, 10, 15, and 20 dB. The two training procedures for designing the clean speech model, namely, the Baum and the segmental k-means, were examined and compared. Similarly, the two enhancement procedures, (45) and (48), were applied and compared. Training was performed using 100 sentences of clean conversational speech spoken by ten speakers through a telephone handset. Enhancement tests were performed on eight sentences spoken by four speakers and recorded in a manner similar to that of the training set. The speech material and the speakers used for training were different from those used for testing. The model for the noise process was estimated directly from the noisy speech, using an initial interval whose length was about 10 percent of the length of the utterance to be enhanced, and in which speech was not present.

In all of our experiments, the dimension of the speech vectors was $K = 128$ at a sampling rate of 8 kHz. Training was done using nonoverlapped frames, while enhancement was performed using frames of speech which overlapped each other by 64 samples. A Hanning analysis window was applied to the speech frames during training and enhancement. The synthesis of the enhanced signal from the individually processed frames was done using the standard short time Fourier transform overlap and add technique [29]. The order of each AR output process of the HMM was set to $N_s = 10$, which is a commonly used value in linear predictive analysis of speech signals. The order of the AR model for the noise process was set to $N_v = 4$ since the noise examined here has a theoretically flat power spectral density. The iterative algorithms for designing the models and for performing the enhancement were terminated whenever the difference in likelihood values at two consecutive iterations, normalized by the older likelihood value, was less than or equal to 10^{-5}.

The number of states M and mixture components for each state L were experimentally determined by examining the enhancement results obtained using different values of (M, L) at input SNR of 10 dB. Table I shows the minimum and maximum SNR values achieved in this experiment. The case VQ-AMAP represents enhancement results obtained using the initial HMM designed by the vector quantization approach described in Section III-D and the approximate MAP enhancement approach described in (48). The case SEG-AMAP represents enhancement results obtained using segmental k-means training and approximate MAP enhancement. The case ML-MAP represents enhancement results obtained using ML Baum training and MAP enhancement (45) approaches. Finally, the cases VQ-CLN and SEG-CLN represent some theoretical performance bounds within the proposed framework for speech enhancement. Here, the clean speech was used for estimating the most likely sequence of states and mixture components, and the noisy speech was filtered through a time-varying Wiener filter, which at each time instant was constructed from the spectrum of the AR process associated with the estimated state and mixture component and the spectrum of the AR model of the noise process. In the VQ-CLN case, decoding is performed by applying AR model vector quantization to the clean speech on a frame-by-frame basis using the $M \times L$ VQ designed for initializing the segmental k-means algorithm. In the SEG-CLN case, Viterbi decoding was applied to the clean speech using the model designed by the segmental k-means algorithm. The major difference between the two cases is that the decoding in the VQ-CLN case is memoryless, while the SEG-CLN version incorporates the Markovian memory. The setup of this specific experiment is described in Fig. 2.

Table I shows that the three proposed speech enhancement schemes, VQ-AMAP, SEG-AMAP, and ML-MAP, provide very similar SNR improvement for all of the examined values of (M, L). Furthermore, this SNR improvement is about 0.5 dB lower than that obtained in the VQ-CLN and SEG-CLN cases which use the clean speech for performing the decoding. The SNR improvement obtained in the latter two cases is essentially identical. Careful informal listening tests indicate that for a given (M, L), the three enhancement schemes, VQ-AMAP, SEG-AMAP, and ML-MAP, provide very similar enhanced speech quality. In some cases, however, the ML-MAP approach provided slightly better results than the other two procedures. The best enhancement results were obtained using the five-state five-mixture model. For this case, the enhanced speech has almost no residual noise, it is reasonably intelligible, and it contains fewer gross estimation errors than the enhanced speech obtained using $M = 8$, $L = 4$ or $M = 16$, $L = 8$. Those gross estimation errors are due to decoding errors which result in an incorrect filter selection. The enhanced speech corresponding to VQ-CLN and SEG-CLN sounds identical, a fact which implies the unimportance of the Markovian memory in decoding the speech signal, given the clean speech, in this application. The differences between the best enhanced speech signals and the speech signals obtained in the VQ-CLN or SEG-CLN cases are generally small. In both cases, the input noise was effectively removed. The speech obtained using VQ-CLN or SEG-CLN is somewhat crisper, but somewhat noiser than the enhanced signals obtained using either VQ-AMAP, SEG-AMAP, or ML-MAP.

TABLE I
ENHANCEMENT RESULTS FOR DIFFERENT NUMBER OF STATES AND MIXTURE COMPONENTS AT 10 dB INPUT SNR

State/Mixture	VQ-AMAP	SEG-AMAP	ML-MAP	VQ-CLN	SEG-CLN
5/5	14.23-15.93	14.25-15.95	14.10-15.84	14.73-16.45	14.72-16.44
8/4	14.24-15.76	14.26-15.75	14.26-15.70	14.75-16.51	14.75-16.50
16/8	14.15-15.85	14.16-15.82	14.04-15.72	15.04-16.72	15.04-16.70

In a complementary experiment to the SEG-CLN case, we performed the enhancement in a manner similar to that shown in Fig. 2, but with decoding being applied to the noisy speech rather than to the clean speech. A block diagram of this system is shown in Fig. 3. The difference between this enhancement scheme and the algorithm described in Fig. 1 is that here the decoding is based upon the noisy speech, while in Fig. 1, the decoding is based upon the current estimate of the clean speech. In fact, if the algorithm described in Fig. 1 is initialized by the given noisy speech, i.e., $y(0) = z$, then the enhanced speech obtained at the first iteration of that algorithm and the enhanced speech obtained in this experiment are identical. The purpose here was to examine the importance of correct decoding and to prove superiority of the approximate MAP algorithm over the intuitive approach of Fig. 3. For this experiment, the SNR values of the enhanced signals were in the range of 12.38-13.26 dB at input SNR of 10 dB. The quality of the enhanced signal obtained here was significantly worse than that of the speech signal obtained in either CLN-SEG or ML-MAP. In particular, the enhanced signal contained a significant level of nonwhite residual noise.

In order to examine further the theoretical bounds of performance of the proposed speech enhancement approach, we repeated the experiments referred to as VQ-CLN for much larger code book VQ's. Specifically, we designed AR model VQ's for the clean speech, with 64, 128, and 256 codewords, using the binary tree design approach of [27]. The SNR of the noisy speech was again 10 dB. The resulting SNR of the enhanced signals was in the range of 15.04-16.79 dB for the 256 codeword code book, 15.01-16.71 dB for the 128 codeword code book, and 14.93-16.58 dB for the 64 codeword code book. The quality and intelligibility of the enhanced speech were fairly good in all three cases, with minor differences among them.

The good quality of the enhanced speech obtained when decoding is done using the clean speech, on the one hand, and the very bad quality of the enhanced speech obtained when decoding is performed using only the noisy speech, on the other hand, indicate that correct decoding of the speech signal is crucial for successful speech enhancement. Given the correct decoding of the noisy speech, a time-varying Wiener filter, which at each time instant is constructed from the power spectral density of a finite-order AR process representing the true power spectral density of the speech vector at that time, performs satisfactorily, even when a relatively small number of quantized AR spectra, e.g., 64, are used.

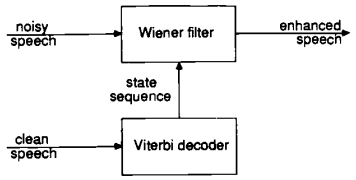

Fig. 2. A block diagram of a theoretical speech enhancement system in which signal decoding is done using the clean speech.

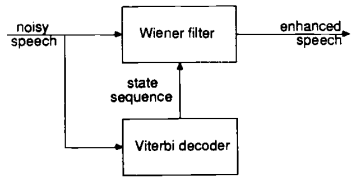

Fig. 3. A block diagram of an open loop version of the system of Fig. 1.

TABLE II
MINIMUM AND MAXIMUM SNR VALUES OBTAINED BY USING THE ML-MAP ENHANCEMENT APPROACH WITH THE FIVE-STATE FIVE-MIXTURE MODEL AT DIFFERENT INPUT SNR (SNR-IN) VALUES

SNR-IN	ML-MAP	VQ-CLN	Iterations
5.00	10.50-11.96	11.12-12.87	10-19
10.00	14.10-15.84	14.73-16.45	10-17
15.00	18.24-19.61	18.63-20.14	10-13
20.00	22.53-23.63	22.76-23.92	11-21

Table II focuses on the ML-MAP approach with five-state five-mixture model, and shows minimum and maximum values of SNR of the enhanced speech obtained at values of different input SNR. The minimum and maximum number of iterations used in each case is also shown. The table also provides a comparison to the theoretical bounds obtained in the VQ-CLN case.

Informal listening to the enhanced speech signals indicates that at 5 dB input SNR, the enhancement was effective only for some of the sentences, while for the other sentences, it introduced some noticeable distortions. At the higher input SNR values of 15 and 20 dB, very good enhanced speech quality was obtained. The noise was completely removed and the speech was minimally distorted. The crispness and naturalness of the original speech were well preserved.

VI. COMMENTS

We proposed a new approach for enhancing speech signals which have been degraded by statistically independent additive noise. The approach capitalizes on statisti-

cal modeling of the clean speech and the noise process using long training sequences from the two processes. Given the estimated statistics of the speech and the noise processes, a MAP estimation approach was developed and implemented using the EM algorithm. An efficient approximation of the MAP enhancement approach, in which time-varying Wiener filtering of the noisy speech and Viterbi decoding of the enhanced speech are alternately applied, was developed. The approach was tested for enhancing speech degraded by white noise. It proved especially useful for enhancing noisy speech with SNR greater than or equal to 10 dB.

We opt for HMM's due to their general acceptability as reliable models for speech signals in the speech recognition community. The most natural way to use these models in speech enhancement applications is for simultaneous enhancement of the entire utterance of noisy speech. This, however, is not the only way the proposed approach can be implemented, and a frame-by-frame enhancement implementation is possible, for example, by considering $\max_{y_t} p_\lambda(y_t|z_t)$. This frame-by-frame version of the MAP enhancement algorithm can also be efficiently implemented using a slightly different forward formula from that used here. The estimate obtained in simultaneous enhancement of all the frames in the noisy input utterance is usually more accurate than that obtained on a frame-by-frame basis since the number of noisy speech samples upon which the estimation is based is larger. Since this is the first paper on the subject, our goal was to establish a benchmark on the performance of the proposed approach. Hence, we focused on the simultaneous enhancement of all the frames in each given noisy input utterance. While the simultaneous estimation of the entire noisy utterance is not practical in real-time speech communication applications because of the long time delay it requires, it may be useful in other applications, such as recognition of noisy speech. In the latter case, the delay introduced does not play any role since recognition is naturally performed on the basis of the entire input utterance. For real-time communication applications, either the frame-by-frame implementation of the MAP enhancement approach mentioned above or the approximate MAP approach with tolerated delay achieved by sequential application of the Viterbi algorithm [30] can be used.

We believe that the main contribution of this paper is in establishing a fairly flexible statistical framework for studying speech enhancement, and in demonstrating encouraging preliminary results using the proposed approach. Rather than attributing arbitrary PD's to the speech signal and the noise process, as was often done in the past, we estimate these statistics from training sequences from the clean speech and the noise process. This statistical knowledge was used here in deriving MAP estimators for the clean speech signal; however, it can easily be applied to derive other estimators such as MMSE estimators for the speech signal or for its sample spectrum [32]. Since, given the models for the speech and noise, we apply optimal (in a given sense) estimators to the noisy speech, better enhancement results are expected as the modeling, especially that of the clean speech signal, is better understood. Some insight into this problem can be obtained from the experience accumulated in the speech recognition area where hidden Markov modeling has been used for a long time. The modeling problems in the two cases, however, are not identical since in speech recognition, the training sequences usually represent the same spoken word and the variability is smaller than in training for speech enhancement. The models used here may be refined, for example, by supplementing the AR estimate of the power spectral density of each vector of the clean speech by a pitch model similar to that used in [31]. In addition, better modeling may be achieved if the order of the model, i.e., the number of states, mixture components, and AR coefficients, is determined in a more optimal way than the experimental approach used here.

An important issue not studied in this paper is the initialization of the iterative enhancement procedures. Since the iterative algorithms always converge to local maxima, the initialization may be important, and initial estimates other than that of the noisy speech used here may prove useful.

Acknowledgment

The authors wish to thank S. Shitz and A. Dembo for fruitful discussions during this work. They also thank L. R. Rabiner and the anonymous reviewers for critical reading of the manuscript and their comments which improved the presentation.

References

[1] Y. Ephraim and R. M. Gray, "A unified approach for encoding clean and noisy sources by means of waveform and autoregressive model vector quantization," *IEEE Trans. Inform. Theory*, vol. IT-34, pp. 826-834, July 1988.

[2] H. L. Van Trees, *Detection, Estimation and Modulation Theory*. New York: Wiley, 1968.

[3] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379-423, 623-656, 1948.

[4] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[5] A. B. Poritz, "Hidden Markov models: A guided tour," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1988, pp. 7-13.

[6] J. D. Ferguson, Ed., *Proc. Symp. Appl. Hidden Markov Models to Text and Speech*, IDA-CRD, Princeton, NJ, 1980.

[7] L. R. Bahl, F. Jelinek, and R. L. Mercer, "A maximum likelihood approach to continuous speech recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-5, pp. 179-190, Mar. 1983.

[8] 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.

[9] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Statist.*, vol. 41, pp. 164-171, 1970.

[10] L. E. Baum, "An inequality and associated maximization technique in statistical estimation of probabilistic functions of Markov processes," *Inequalities*, vol. 3, no. 1, pp. 1-8, 1972.

[11] L. A. Liporace, "Maximum likelihood estimation for multivariate observations of Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 729-734, Sept. 1982.

[12] A. B. Poritz, "Linear predictive hidden Markov models and the speech signal," in *Proc. Symp. Appl. Hidden Markov Models to Text and Speech*, J. D. Ferguson, Ed., IDA-CRD, Princeton, NJ, 1980, pp. 88-142, and summarized in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1982, pp. 1291-1294.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likeli-

hood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. B*, vol. 39, pp. 1-38, 1977.
[14] C. F. J. Wu, "On the convergence properties of the EM algorithm," *Ann Statist.*, vol. 11, no. 1, pp. 95-103, 1983.
[15] L. R. Rabiner, J. G. Wilpon, and B.-H. Juang, "A segmental k-means training procedure for connected word recognition," *AT&T Tech. J.*, pp. 21-40, May-June 1986.
[16] G. D. Forney, Jr., "The Viterbi algorithm," *Proc. IEEE*, vol. 61, pp. 268-278, Mar. 1973.
[17] B. R. Musicus, "An iterative technique for maximum likelihood parameter estimation on noisy data," S.M. thesis, M.I.T., Cambridge, 1979.
[18] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 197-209, June 1978.
[19] A. Dembo, "The relation between maximum likelihood estimation of structured covariance matrices and periodograms," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 1661-1662, Dec. 1986.
[20] A. Q. Nguyen, "On the uniqueness of the maximum likelihood estimate of structured covariance matrices," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 1249-1251, Dec. 1984.
[21] R. M. Gray, A. H. Gray, Jr., G. Rebolledo, and J. E. Shore, "Rate-distortion speech coding with a minimum discrimination information distortion measure," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 708-721, Nov. 1981.
[22] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*. New York: Springer-Verlag, 1976.
[23] B.-H. Juang and L. R. Rabiner, "The segmental k-means algorithm for estimating parameters of hidden Markov models," submitted for publication.
[24] D. G. Luenberger, *Linear and Nonlinear Programming*. Reading, MA: Addison-Wesley, 1984.
[25] M. J. Sabin and R. M. Gray, "Global convergence and empirical consistency of the generalized Lloyd algorithm," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 148-155, Mar. 1986.
[26] R. M. Gray, "Toeplitz and circulant matrices: II," Stanford Electron. Lab., Stanford, CA, Tech. Rep. 6504-1, Apr. 1977.
[27] A. Buzo, A. H. Gray, Jr., R. M. Gray, and J. D. Markel, "Speech coding based upon vector quantization," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 562-574, Oct. 1980.
[28] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," *Bell Syst. Tech. J.*, vol. 62, pt. 1, pp. 1035-1074, Apr. 1983.
[29] R. E. Crochiere, "A weighted overlap-add method of short-time Fourier analysis/synthesis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 99-102, Feb. 1980.
[30] L. L. Scharf, D. D. Cox, and J. Masreliez, "Modulo-2π phase sequence estimation," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 615-620, Sept. 1980.
[31] J. M. Tribolet and R. E. Crochiere, "Frequency domain coding of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 512-530, Oct. 1979.
[32] Y. Ephraim, "Minimum mean square error speech enhancement based upon hidden Markov modeling," submitted for publication.

Yariv Ephraim (S'82-M'84) was born in Cairo, Egypt, on September 9, 1951. He received the B.Sc., M.Sc., and D.Sc. degrees from the Technion—Israel Institute of Technology, Haifa, in 1977, 1979, and 1984, respectively, all in electrical engineering.

During 1984-1985 he was a Rothschild Postdoctoral Fellow at the Information Systems Laboratory, Stanford University, Stanford, CA. Since 1985 he has been a Member of Technical Staff in the Speech Research Department, AT&T Bell Laboratories, Murray Hill, NJ. His research interests are statistical signal processing, and speech modeling, coding, enhancement, and recognition.

David Malah (S'67-M'71-SM'84-F'87), for a photograph and biography, see p. 1679 of the November 1989 issue of this TRANSACTIONS.

Biing-Hwang Juang (S'79-M'81-SM'87), for a photograph and biography, see p. 804 of the June 1989 issue of this TRANSACTIONS.