S7.2

SPEECH ENHANCEMENT BASED UPON HIDDEN MARKOV MODELING

Yariv Ephraim, David Malah[1]*, and Biing-Hwang Juang*

AT&T Bell Laboratories
Murray Hill, NJ 07974

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. A low order Gaussian AR model is used for the wide-band Gaussian noise considered here. 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. An approximate improvement of 4.0-6.0 dB in signal to noise ratio (SNR) is achieved at 10 dB input SNR.

I. Introduction

In [1], a model based approach for enhancing speech signals degraded by statistically independent additive noise was proposed. In this approach the unknown probability distributions (PD's) of the speech signal and the noise process are first estimated from long training sequences from the two processes, and then estimation of the clean speech signal is applied using the estimated statistics. Hidden Markov models (HMM's) were used for the clean speech, and a low order Gaussian autoregressive (AR) model was used for the noise process. The noise process considered in [1] was a wide-band Gaussian noise. The parameter set of the HMM for the speech signal was estimated using the segmental k-means approach, which is an approximate maximum likelihood (ML) modeling approach. The estimation of the clean speech signal was done using an approximate maximum a-posteriori (MAP) estimation approach, in which the joint probability density function (pdf) of the state sequence, the clean signal, and the noisy speech is locally maximized over all state sequences and clean signals.

In this paper we examine ML hidden Markov modeling of the clean speech and *exact* MAP estimation of the clean speech given the noisy speech. The modeling is done using the well known Baum algorithm, and enhancement is performed using the *EM* (estimation-maximization) algorithm [2], [3]. The MAP 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 non-decreasing likelihood values by maximizing in each iteration an appropriately defined auxiliary function. The enhancement algorithm developed here is examined and compared with the approximate MAP approach of [1].

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

$$\begin{aligned} p_{\lambda_s}(y) &= \sum_x \sum_h p_{\lambda_s}(x,h,y) \\ &= \sum_x \sum_h p_{\lambda_s}(x) p_{\lambda_s}(h \mid x) p_{\lambda_s}(y \mid h,x), \end{aligned} \tag{1}$$

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

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

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 $p_{\lambda_s}(h \mid x)$, and $p_{\lambda_s}(y \mid h,x)$, we make the following standard assumptions:

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

and

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

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

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

where # denotes vector transpose, $S_{\gamma\mid\beta} = \sigma_{\gamma\mid\beta}^2 (A_{\gamma\mid\beta}^{\#} A_{\gamma\mid\beta})^{-1}$, $\sigma_{\gamma\mid\beta}^2$ is the variance of the innovation process of the AR source, and $A_{\gamma\mid\beta}$ 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, $g_{\gamma\mid\beta} \triangleq (g_{\gamma\mid\beta}(0), g_{\gamma\mid\beta}(1),\cdots, g_{\gamma\mid\beta}(N_s))$, $g_{\gamma\mid\beta}(0)=1$.

The modeling problem is that of estimating the parameter set $\lambda_s=(\pi,a,c,S)$, where $\pi \triangleq \{\pi_\beta\}$, $a \triangleq \{a_{\alpha,\beta}\}$, $c \triangleq \{c_{\gamma\mid\beta}\}$, and $S \triangleq \{S_{\gamma\mid\beta}\}$, 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

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

and this maximization is locally performed by the Baum reestimation algorithm. The segmental $k-$ means algorithm for estimating the parameter set of the model used in [1], 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 ML estimation procedure is described in Section III-A.

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 with a

[1] D. Malah is on sabbatical leave from the Department of Electrical Engineering, Technion-Israel Institute of Technology, Haifa 32000, Israel.

theoretically flat power spectral density considered in this paper, we assume that

$$p_{\lambda_v}(v) = \prod_{t=0}^{T} p_{\lambda_v}(v_t)$$

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

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_v-th order AR covariance matrix. $V=\sigma_v^2(A_v^{\#}A_v)^{-1}$, where σ_v^2 and A_v are defined similarly to $\sigma^2_{\gamma|\beta}$ 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-B.

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

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

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

$$p_{\lambda_s\lambda_v}(x,h,y,z) = p_{\lambda_v}(z \mid x,h,y) p_{\lambda_s}(x,h,y)$$

$$= p_{\lambda_v}(z \mid y) p_{\lambda_s}(x,h,y)$$

$$= p_{\lambda_v}(z-y) p_{\lambda_s}(x,h,y) \tag{12}$$

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

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

where $p_{\lambda_s\lambda_v}(y \mid z) = p_{\lambda_s\lambda_v}(y,z)/p_{\lambda_s\lambda_v}(z)$. The approximate MAP enhancement procedure developed in [1] assumes 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

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

The MAP enhancement procedure (13) is 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 algorithm 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 n-th 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. This algorithm generates a sequence of HMM's with non-decreasing 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,

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

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 AR covariance matrices $S_{\gamma|\beta}'$, 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.

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

$$\min_{S_{\gamma|\beta}} \{ \mathrm{tr}\, R_{\gamma|\beta}' S^{-}_{\gamma|\beta} - \ln \det S^{-}_{\gamma|\beta} \} \tag{21}$$

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

where

$$q_{t,n}(\alpha,\beta,\gamma) \triangleq \frac{\sum_{\left\{x_{T_n}:\, \substack{x_{t-1,n}=\alpha \\ x_{t,n}=\beta}\right\}} \sum_{\left\{h_{T_n}:\, h_{t,n}=\gamma\right\}} p_{\lambda_s}(x_{T_n}, h_{T_n}, y_{T_n})}{\sum_{x_{T_n}} \sum_{h_{T_n}} p_{\lambda_s}(x_{T_n}, h_{T_n}, y_{T_n})},$$

$$0 < t \leq T_n \tag{22}$$

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 n-th utterance of the speech training sequence, and

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

$$0 \leq t \leq T_n \tag{23}$$

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 n-th 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 satisfy $\sum_{\beta=1}^{M} a'_{\alpha,\beta}=1$ 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.

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

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

where

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

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

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

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

The minimization problem in (21) has a unique solution provided that $R'_{\gamma|\beta}$ is positive definite. 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. An approximate solution can be obtained by the autocorrelation method of linear prediction if end-block effects are neglected. More specifically, if the K-th 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

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

where

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

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. 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

$$\min_V \{\operatorname{tr} R_v V^{-1} - \ln\det V^{-1}\}, \qquad (29)$$

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. Approximate solution is obtained from AR modeling of the autocorrelation function given by

$$r_v(m) \triangleq \frac{1}{T+1}\sum_{t=0}^{T}\frac{1}{K}\sum_{k=0}^{K-|m|-1} v_t(k) v_t(k+|m|) \qquad m=-N_v, \cdots, N_v. \quad (30)$$

IV. EM Speech Enhancement Algorithm

In this section we apply the *EM* algorithm for MAP estimation of the clean speech signal given the noisy speech. 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

$$\begin{aligned}
\ln p_\lambda(y(k+1)\mid z) - \ln p_\lambda(y(k)\mid z) &= \ln\sum_{x,h}\frac{p_\lambda(x,h,y(k)\mid z)}{p_\lambda(y(k)\mid z)}\,\frac{p_\lambda(x,h,y(k+1)\mid z)}{p_\lambda(x,h,y(k)\mid z)}\\
&= \ln\sum_{x,h} p_\lambda(x,h\mid y(k))\frac{p_\lambda(x,h,y(k+1)\mid z)}{p_\lambda(x,h,y(k)\mid z)}\\
&\geq \sum_{x,h} p_\lambda(x,h\mid y(k))\ln\frac{p_\lambda(x,h,y(k+1)\mid z)}{p_\lambda(x,h,y(k)\mid z)}\\
&\triangleq \phi(y(k+1)) - \phi(y(k)), \qquad (31)
\end{aligned}$$

where

$$\phi(y(k+1)) \triangleq \sum_{x,h} p_\lambda(x,h\mid y(k)) \ln p_\lambda(x,h,y(k+1)\mid z). \qquad (32)$$

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 (32), and setting the gradient of $\phi(y(k+1))$ with respect to $y_t(k+1)$ to zero, we obtain the following reestimation formula for the clean speech signal.

$$y_t(k+1)=\left[\sum_{\beta,\gamma} q_t(\beta,\gamma\mid y(k)) H^{-}_{\gamma|\beta}\right]^{-1} z_t, \qquad 0\leq t\leq T, \qquad (33)$$

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

$$H_{\gamma|\beta} \triangleq S_{\gamma|\beta}(S_{\gamma|\beta}+V)^{-1}. \qquad (34)$$

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 similarly to the implementation of the approximate MAP approach in [1]. The reestimation algorithm (33) is initialized from the noisy speech, i.e., $y(0)=z$, and the algorithm is stopped when a convergence criterion similar to that used in Section II is satisfied.

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, the approximate MAP and the exact MAP (33) approaches, were applied and compared. Training was performed using 100 sentences of clean conversational speech spoken by 10 speakers through a telephone handset. Enhancement tests were performed on 8 sentences spoken by 4 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% 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 8kHz. Training was done using non-overlapped 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. 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 HMM's designed by the standard AR model vector quantization approach, and the approximate MAP enhancement approach described in [1]. The VQ model was used for initializing the segmental k–means algorithm, and the segmental k–means model was used to initialize the Baum algorithm. 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 (33) 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 by a time varying Wiener filter. At each time instant, the filter 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, AR model vector quantization was applied 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 *VQ–CLN* case is memoryless while the *SEG–CLN* version incorporates the Markovian memory.

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 0.5dB 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 on 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* is somewhat crisper and somewhat noisier than the enhanced signals obtained using either of *VQ–AMAP*, *SEG–AMAP*, or *ML–MAP*.

Table I: Enhancement results for different number of states and mixture components at 10 dB input SNR.

M/L	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

Table II focuses on the *ML–MAP* approach with the 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 with the theoretical bounds obtained in the *VQ–CLN* case.

Table II: Minimum and maximum SNR values obtained by using the *ML–MAP* approach with M=L=5.

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

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 statistical 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. This approach 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 $\max_{y_t} p_\lambda(y_t \mid z_t)$. This frame-by-frame version of the MAP enhancement algorithm can also be efficiently implemented using slightly different forward-backward formulas than those 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.

References

[1] Y. Ephraim, D. Malah, and B.-H. Juang, "On the application of hidden Markov models for enhancing noisy speech," *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 533-536, April 1988.

[2] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the *EM* algorithm," *J. Royal Stat. Soc. B*, vol. 39, pp. 1-38, 1977.

[3] B. R. Musicus, "An iterative technique for maximum likelihood estimation with noisy data," S. M. Thesis, MIT, 1979.