On the Application of Hidden Markov Models for Enhancing Noisy Speech

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

Speech Research Department
AT&T Bell Laboratories
Murray Hill, NJ 07974

Department of Electrical Engineering[1]
Technion, Israel Institute of Technology
Haifa 32000, Israel

ABSTRACT

We propose a new algorithm for enhancing noisy speech which have been degraded by statistically independent additive noise. The algorithm is based upon modeling the clean speech as a hidden Markov process with mixtures of Gaussian autoregressive (AR) output processes, and the noise process as a sequence of stationary, statistically independent, Gaussian AR vectors. The parameter sets of the models are estimated using training sequences from the clean speech and the noise process. The parameter set of the hidden Markov model is estimated by the segmental k-means algorithm. Given the estimated models, the enhancement of the noisy speech is done by alternate maximization of the likelihood function of the noisy speech, once over all sequences of states and mixture components assuming that the clean speech signal is given, and then over all vectors of the original speech using the resulting most probable sequence of states and mixture components. This alternating maximization is equivalent to first estimating the most probable sequence of AR models for the speech signal using the Viterbi algorithm, and then applying these AR models for constructing a sequence of Wiener filters which are used to enhance the noisy speech.

1. Introduction

The problem of enhancing noisy speech is basically an estimation problem which requires knowledge of the probability distributions (PD's) of the speech signal and the noise process. In practice, however, these PD's are not known and the best which can be done is to use training sequences from the speech and the noise processes through which the unavailable statistics are learned. Direct application of the training sequences for approximating the conditional expected value results is a practically unacceptable solution [1]. An alternative approach, which has been proved useful in speech coding and recognition applications, is first to use training sequences for estimating parametric models for the PD's of the source and the noise, and then to implement the desired estimator of the original speech based upon the estimated PD's.

In this paper we apply the above modeling approach for enhancing speech signals which have been degraded by stationary, statistically independent, additive noise. The approach, however, can be extended to noise processes which are neither additive nor strictly stationary without any major difficulties. We use hidden Markov models (HMM's) with mixtures of Gaussian autoregressive (AR) output PD's for the speech signal, and a Gaussian AR model for the noise process. The estimation of the parameter sets of the models, and the enhancement process itself, are both optimal in the maximum likelihood (ML) sense. The estimation of the parameter set of the HMM is done using the segmental k-means algorithm, which jointly estimates the parameter set of the model and the sequence of states and mixture components which maximize the likelihood function of the clean speech [2]. The model for the noise is simply the centroid of the training sequence from that process.

Given the parameter sets of the speech and noise models, the enhancement of the noisy speech is done by alternate maximization of the likelihood function of the noisy speech, once over all sequences of states and mixture components assuming that the clean speech vectors are given, and then over all speech vectors using the most likely sequence of states and mixture components. The estimation of the most likely sequence of states and mixture components is done using the Viterbi algorithm, and it results in a sequence of AR models which are associated with the current estimate of the vectors of the clean speech signal. The estimation of the most likely sequence of clean speech vectors is done by Wiener filtering of the individual noisy speech vectors. The Wiener filter for each speech vector uses the AR model corresponding to this vector, from the most likely sequence of AR models, and the AR models for the noise process. The iterative procedure proceeds until some convergence criterion is satisfied.

The paper is organized as follows. In Section 2 we formulate the problem and specify the statistical models we use here. In Section 3 we describe the training procedure. In Section 4 we describe the enhancement algorithm. Finally, in Section 5 we describe the experiments used to evaluate the algorithm.

2. Problem Formulation

2.1 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 here HMM's with M states and mixtures of L Gaussian AR output processes at each state. Let $y \triangleq \{y_t,\ t=0,\ldots,T\}$, $y_t \in R^K$, be a sequence of K-dimensional vectors which represent the output from the model. Let $x \triangleq \{x_t,\ t=0,\ldots,T\}$, $x_t \in \{1,\ldots,M\}$, be a sequence of states. Let $h \triangleq \{h_t,\ t=0,\ldots,T\}$, $h_t \in \{1,\ldots,L\}$, be a sequence of mixture components. The pdf p_{λ_s} is given by

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

$p_{\lambda_s}(x)$ in (1) is the probability of the sequence of states x, and it is given by

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

where, $a_{x_{t-1}x_t}$ is the probability of being in state x_{t-1} (at time $t-1$) and in state x_t (at time t). $a_{x_{-1}x_0} \triangleq \pi_{x_0}$ denotes the probability of the initial state x_0. For $p_{\lambda_s}(h \mid x)$, the probability of choosing the sequence of mixture components, h, given the sequence of states x, and $p_{\lambda_s}(y \mid h,x)$, the pdf of the output sequence y given $\{x,h\}$, in (1), we make the following standard assumptions:

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

and

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

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 N_s-th order Gaussian AR output processes, we have

CH2561-9/88/0000-0533 $1.00 © 1988 IEEE

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

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

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

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

and this estimate is efficiently achieved using the Baum algorithm. In this work, however, we shall use the segmental k-means algorithm, which is significantly simpler to implement, yet it produces comparable results to the Baum algorithm. The estimation of the parameter set λ_x by the segmental k-means algorithm results from [2]

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

This estimate can be thought of as an approximation to the Baum estimator since it is obtained by replacing the double summation in (6) by the maximal term taken over all possible pairs of sequences (x,h).

2.2 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. Assume that the output from this model is a sequence of stationary, statistically independent, Gaussian AR K-dimensional vectors. Let N_v be the order of the AR process. Let $v \triangleq \{v_t, t=0, \ldots, T\}$ be a sequence of $T+1$ output vectors from the model. We have that

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

where, $V=\sigma_v^2(A_v^{\#}A_v)^{-1}$, and 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, $g_v \triangleq (g_v(0), g_v(1), \ldots, g_v(N_v))$, $g_v(0)=1$.

The noise modeling problem is that of finding the parameter set $\lambda_v \triangleq (\sigma_v^2, g_v(m), m=1, \ldots, N_v)$ by

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

for a given training sequence v from the noise process.

2.3 Speech Enhancement Problem

Given the parameter set λ_x 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 $z \triangleq \{z_t, t=0, \ldots, T\}$, $z_t=y_t+v_t$, the enhancement problem is that of estimating the sequence y of clean speech vectors by

$$\max_{y} \ln p_{\lambda_x\lambda_v}(y,z), \tag{10}$$

where,

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

due to the fact that the noise is additive and statistically independent of the signal. Note that the enhancement problem stated in (10) is consistent with the training procedure described in (6). It therefore has complexity similar to that associated with the maximization in (6). Here we perform the enhancement in a consistent manner with our training procedure which is described in (7). Specifically, the estimation of the speech signal results from

$$\max_{x,h,y} \ln p_{\lambda_x\lambda_v}(x,h,y,z), \tag{12}$$

where,

$$\begin{aligned} p_{\lambda_x\lambda_v}(x,h,y,z) &= p_{\lambda_v}(z \mid x,h,y)p_{\lambda_x}(x,h,y) \\ &= p_{\lambda_v}(z \mid y)p_{\lambda_x}(x,h,y) \\ &= p_{\lambda_v}(z-y)p_{\lambda_x}(x,h,y) \end{aligned} \tag{13}$$

due to the fact that given y, z and (x,h) are statistically independent.

3. Training of Speech and Noise Models

3.1 HMM Estimation

The estimation of the parameter set λ_x of the HMM for the clean speech is done by alternate maximization of the log likelihood $\ln p_{\lambda_x}(x,h,y)$, once over (x,h) assuming that λ_x is given, and then over λ_x assuming that (x,h) is known. Thus if each iteration comprises the estimation of (x,h) for a given λ_x and the estimation of a new λ_x based on (x,h), then the training algorithm generates a sequence of models with increasing likelihood. The procedure is stopped when a convergence criterion is satisfied, e.g., when the difference of the values of the log likelihood function (7) in two consecutive iterations is smaller than a given threshold. We now show how each of the two phases of each iteration is performed. We shall not discuss the convergence of the algorithm which can be shown by a standard argument from optimization theory.

Assume that an initial estimate of the parameter set of the model is given. Then, the estimation of (x,h) which maximizes $\ln p_{\lambda_x}(x,h,y)$ can be done by applying the Viterbi algorithm using the following path metric

$$\ln \pi_\beta + \ln c_{\gamma|\beta} + \ln b(y_0 \mid x_0=\beta, h_0=\gamma), \tag{14}$$

for $t=0$, and

$$\ln a_{\alpha,\beta} + \ln c_{\gamma|\beta} + \ln b(y_t \mid x_t=\beta, h_t=\gamma), \tag{15}$$

for $1\le t\le T$, where $\alpha,\beta=1, \ldots, M$ and $\gamma=1, \ldots, L$. Let the resulting sequences of states and mixture components be denoted by (x^*,h^*).

Assume that (x^*,h^*) is given. Then, the estimation of a new parameter set, say λ_x', is done by

$$\begin{aligned} \max_{\lambda_x'} \ln p_{\lambda_x'}(x^*,h^*,y) = \max_{\lambda_x'}\{ &\ln \pi'_{x_0^*} + \sum_{t=1}^{T} \ln a'_{x_{t-1}^* x_t^*} \\ &+ \sum_{t=0}^{T} \ln c'_{h_t^* \mid x_t^*} + \sum_{t=0}^{T} \ln b'(y_t \mid x_t^*, h_t^*)\}, \end{aligned} \tag{16}$$

subject to the following constraints.

$$\pi'_\beta \ge 0, \ \sum_{\beta=1}^{M} \pi'_\beta = 1, \quad a'_{\alpha,\beta} \ge 0, \ \sum_{\beta=1}^{M} a'_{\alpha,\beta} = 1, \quad c'_{\gamma|\beta} \ge 0, \ \sum_{\gamma=1}^{L} c'_{\gamma|\beta} = 1, \tag{17}$$

and $S'_{\gamma|\beta}$ is positive definite, for all $\alpha,\beta=1, \ldots, M$ and $\gamma=1, \ldots, L$.

The maximization of (16) over π'_β is trivial since x_0^* is given. Hence,

$$\pi'_\beta = 1 \text{ for } \beta=x_0^*, \text{ and } \pi'_\beta = 0 \text{ for } \beta\ne x_0^*. \tag{18}$$

The maximization of (16) over $a'_{\alpha,\beta}$, subject to the constraints (17), is equivalent to the following problem.

$$\begin{aligned} &\max_{\{a'_{\alpha,\beta}\}_{\beta=1}^{M}} \{\sum_{\beta=1}^{M} \sum_{t\in\tau_{\alpha,\beta}} \ln a'_{x_{t-1}^* x_t^*}\} \\ &\text{subject to: } a'_{\alpha,\beta} \ge 0, \ \sum_{\beta=1}^{M} a'_{\alpha,\beta} = 1, \end{aligned} \tag{19}$$

where, $\tau_{\alpha,\beta} \triangleq \{1\le t\le T; \ x_{t-1}^*=\alpha, \ x_t^*=\beta\}$. From the definition of $\tau_{\alpha,\beta}$ we have that (19) is equivalent to

$$\begin{aligned} &\max_{\{a'_{\alpha,\beta}\}_{\beta=1}^{M}} \{\sum_{\beta=1}^{M} \mid \tau_{\alpha,\beta} \mid \ln a'_{\alpha,\beta}\} \\ &\text{subject to: } a'_{\alpha,\beta} \ge 0, \ \sum_{\beta=1}^{M} a'_{\alpha,\beta} = 1, \end{aligned} \tag{20}$$

where, $\mid \tau_{\alpha,\beta} \mid$ is the cardinality of $\tau_{\alpha,\beta}$. The solution of this problem is

$$a'_{\alpha,\beta} = \mid \tau_{\alpha,\beta} \mid \ / \sum_{\beta=1}^{M} \mid \tau_{\alpha,\beta} \mid, \tag{21}$$

provided that $|\tau_{\alpha,\beta}|>0$. Otherwise, $a'_{\alpha,\beta}$ can be arbitrarily chosen, subject to the constraints in (19), since its value does not affect the likelihood function. Note that the estimate in (21) has the intuitive interpretation of being the number of transitions along the most probable path from state α to state β, normalized by the number of instances in which state α has been visited.

The estimation of $c'_{\gamma|\beta}$ by maximizing (16) subject to the constraints (17) is done similarly to the estimation of $a'_{\alpha,\beta}$. It results in

$$c'_{\gamma|\beta} = |\eta_{\gamma|\beta}| \Big/ \sum_{\gamma=1}^{L} |\eta_{\gamma|\beta}|, \tag{22}$$

where, $\eta_{\gamma|\beta} \triangleq \{0\le t\le T;\ h_t^*=\gamma,\ x_t^*=\beta\}$, provided that $|\eta_{\gamma|\beta}|>0$. Otherwise, $c'_{\gamma|\beta}$ can be arbitrary chosen from the feasible set, since its value does not affect the likelihood function (16). This estimate of $c'_{\gamma|\beta}$ has the intuitive interpretation of being the ratio of the number of time instances for which the γ-th mixture component associated with state β has been chosen for the most probable path, and the number of time instances any other mixture component associated with the β-th state has been chosen for the same path.

The parameters of the AR process associated with state β and mixture component γ, results from

$$\min_{S'_{\gamma|\beta}} \{\mathrm{tr}\, R_{\gamma|\beta} S'^{-1}_{\gamma|\beta} - \ln\det R_{\gamma|\beta} S'^{-1}_{\gamma|\beta}\} \tag{23}$$

where,

$$R_{\gamma|\beta} \triangleq \sum_{t\in\eta_{\gamma|\beta}} y_t y_t^{\#}, \tag{24}$$

provided that the set $\eta_{\gamma|\beta}$ is not empty. Otherwise, any positive definite covariance matrix of the given AR structure, $S'_{\gamma|\beta}$, can be chosen without affecting the value of the log likelihood (16). The minimization problem in (23) has a unique solution provided that $R_{\gamma|\beta}$ is positive definite [3]. This unique solution results from a set of linear equations similar to those associated with the "covariance method" of linear prediction analysis.

A simpler solution to the estimation of the parameters of each AR process is possible if the pdf (5) is approximated by

$$b(y_t \mid h_t=\gamma, x_t=\beta) = \frac{\exp\{-\frac{1}{2}\sum_{m=-N_s}^{N_s} r_t(m) r_{\gamma|\beta}(m)/\sigma^2_{\gamma|\beta}\}}{(2\pi\sigma^2_{\gamma|\beta})^{K/2}} \tag{25}$$

where

$$r_t(m) \triangleq \sum_{n=0}^{K-|m|-1} y_t(n) y_t(n+|m|)$$

$$r_{\gamma|\beta}(m) \triangleq \sum_{n=0}^{N_s-|m|-1} g_{\gamma|\beta}(n) g_{\gamma|\beta}(n+|m|). \tag{26}$$

This approximation results from replacing the vector $A_{\gamma|\beta}y_t$ in (5) by $g_{\gamma|\beta}\otimes y_t$, which is the convolution of the two sequences $g_{\gamma|\beta}(n)$ and $y_t(n)$. Such an approximation is reasonable if $K\gg N_s$. On substituting (25) into (16) the problem becomes

$$\min_{\sigma^2_{\gamma|\beta},\, g_{\gamma|\beta}} \{\sum_{m=-N_s}^{N_s} \bar{r}(m) r_{\gamma|\beta}(m)/\sigma^2_{\gamma|\beta} + K\ln\sigma^2_{\gamma|\beta}\}, \tag{27}$$

where,

$$\bar{r}(m) \triangleq \frac{1}{|\eta_{\gamma|\beta}|} \sum_{t\in\eta_{\gamma,\beta}} r_t(m). \tag{28}$$

The minimization in (27) is a standard problem in linear prediction analysis and it is achieved by applying the so called "autocorrelation method" of linear prediction to the autocorrelation sequence $\{\bar{r}(m)\}$, provided that this sequence is positive definite.

The iterative algorithm described above for estimating the parameter set of the HMM for the speech signal is started from some initial estimate of λ_y. This estimate is obtained here by initial clustering of the training sequence into $M\times L$ AR models, using the standard generalized Lloyd algorithm for AR model vector quantization. First a code book of size M is designed and the training sequence is decoded into the M state codewords. For each state we design a code book of size L, using the subtraining sequence assigned to that state, by repeatedly splitting the codeword representing that state. The resulting $M\times L$ codewords are used as the initial parameters for the AR processes of the HMM. An initial estimate for $c_{\gamma|\beta}$ is obtained by first decoding the subtraining sequence corresponding to the β-th state codeword using the L mixture codewords, and then counting the relative frequency of appearance of each of the L "mixture" codewords. The probability of the initial state is chosen to be equal for each of the M states. The state transition probability was chosen to be $a_{\alpha,\alpha}=0.8$, $\alpha=1,\ldots,M$, and $a_{\alpha,\beta}=0.2/(M-1)$, $\alpha,\beta=1,\ldots,M$, $\alpha\neq\beta$.

3.2 Noise Model Estimation

The estimation problem of the parameter set of the AR model for the noise process results from substituting (8) into (9). It is given by,

$$\min_{V}\{\mathrm{tr}\, R_v V^{-1} - \ln\det R_v V^{-1}\}, \tag{29}$$

where,

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

This is the same problem as that associated with the estimation of the parameters of each AR output process of the HMM. As done in that case, we approximate p_{λ_v} by

$$p_{\lambda_v} = \frac{\exp\{-\frac{1}{2}\sum_{t=0}^{T}\sum_{m=-N_v}^{N_v} \rho_t(m)\rho_v(m)/\sigma_v^2\}}{(2\pi\sigma_v^2)^{(T+1)K/2}} \tag{30}$$

where

$$\rho_t(m) \triangleq \sum_{n=0}^{K-|m|-1} v_t(n) v_t(n+|m|)$$

$$\rho_v(m) \triangleq \sum_{n=0}^{N_v-|m|-1} g_v(n) g_v(n+|m|). \tag{31}$$

The estimation of λ_v is done by applying the "autocorrelation method" of linear prediction analysis to the autocorrelation sequence given by

$$\bar{\rho}(m) = \frac{1}{T+1}\sum_{t=0}^{T}\rho_t(m), \tag{32}$$

provided that this sequence is positive definite.

4. Speech Enhancement Algorithm

The enhancement of the noisy speech is performed iteratively by alternate maximization of the log likelihood $\ln p_{\lambda_y\lambda_v}(x,h,y,z)$, defined in (13), once over (x,h), assuming that y is given, and then over y, assuming that (x,h) is known. Given an initial estimate of the clean speech to be enhanced, the estimation of the best sequence of states and mixture components is done by applying the Viterbi algorithm using the following path metric

$$\ln\pi_\beta + \ln c_{\gamma|\beta} + \ln b(y_0 \mid x_0=\beta, h_0=\gamma) + \ln p_{\lambda_v}(z_0-y_0), \tag{33}$$

for $t=0$, and

$$\ln a_{\alpha,\beta} + \ln c_{\gamma|\beta} + \ln b(y_t \mid x_t=\beta, h_t=\gamma) + \ln p_{\lambda_v}(z_t-y_t), \tag{34}$$

for $1\le t\le T$, where $\alpha,\beta=1,\ldots,M$ and $\gamma=1,\ldots,L$. Let the resulting sequence of states and mixture components be denoted by (x^*,h^*).

Assume that (x^*,h^*) is given. Then, a new estimate of the speech signal, say $\{\hat{y}_t\}$, is obtained by

$$\max_{y_t}\{\ln b(y_t \mid x_t^*, h_t^*) + \ln p_{\lambda_v}(z_t-y_t)\}. \tag{35}$$

On substituting (5) and (8) into (35) the problem becomes

$$\min_{y_t}\{y_t^{\#} S^{-1}_{h_t^*|x_t^*} y_t + (z_t-y_t)^{\#} V^{-1}(z_t-y_t)\}. \tag{36}$$

The solution of (36) is easily shown to be

$$\hat{y}_t = S_{h_t^*|x_t^*}(S_{h_t^*|x_t^*} + V)^{-1} z_t, \tag{37}$$

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.

The estimate $\hat{y}_t$ can be efficiently implemented in the frequency domain if $S_{k_t^* | x_t^*}$ and V are approximated by their asymptotically equivalent circulant covariance matrices. Since both covariance matrices correspond to some AR processes, such approximations are always possible. Let

$$f_{k_t^* | x_t^*}(\theta) \triangleq \sigma^2_{k_t^* | x_t^*} / |A_{k_t^* | x_t^*}(\theta)|^2$$

and

$$f_v(\theta) \triangleq \sigma_v^2 / |A_v(\theta)|^2, \tag{38}$$

where, $A_{k_t^* | x_t^*}(\theta)$ and $A_v(\theta)$ are the Fourier transforms of $g_{k_t^* | x_t^*}$ and g_v, respectively, be the asymptotic power spectral densities associated with the two AR processes. Let

$$S_{k_t^* | x_t^*} \sim C(f_{k_t^* | x_t^*}(\theta))$$
$$V \sim C(f_v(\theta)), \tag{39}$$

where, $C(f_{k_t^* | x_t^*}(\theta))$ and $C(f_v(\theta))$ are the asymptotically equivalent circulant covariance matrices of $S_{k_t^* | x_t^*}$ and V, respectively. Using some basic properties of circulant matrices and their inverses, we have that

$$S_{k_t^* | x_t^*}(S_{k_t^* | x_t^*} + V)^{-1} \sim C\left(\frac{f_{k_t^* | x_t^*}(\theta)}{f_{k_t^* | x_t^*}(\theta) + f_v(\theta)}\right) \tag{40}$$

and, hence,

$$\hat{y}_{t,\theta} = \frac{f_{k_t^* | x_t^*}(\theta)}{f_{k_t^* | x_t^*}(\theta) + f_v(\theta)} z_{t,\theta}, \tag{41}$$

where, $\hat{y}_{t,\theta}$ and $z_{t,\theta}$ are the Fourier transforms of $\hat{y}_t$ and z_t, respectively.

The iterative enhancement algorithm described above is started from $\hat{y}_t = z_t$, i.e., we use the noisy source as our initial estimate of the clean speech.

5. Experimental Results

The algorithm for speech enhancement described above was used to enhance speech signals which were degraded by statistically independent additive white noise at signal to noise ratio (SNR) values of 0, 5, 10, 15, and 20 dB. The values of the parameters of the algorithm, namely the number of states M, the number of mixture components for each states L, the order of each autoregressive output process N_s, and the order of the AR model for the noise N_v, were experimentally determined. Since the noise examined here is white with theoretically flat power spectral density, the order of its AR model was chosen to be N_v=4. The order of each AR output process of the HMM for the clean speech was chosen to be N_s=10, which is a commonly used value in linear predictive analysis of speech signals. The product $M \times L$ determines the total number AR codewords used in modeling the clean speech signal. To determine this number we performed the following experiment.

We designed AR model vector quantizers for the clean speech, using the generalized Lloyd algorithm, with 64, 128, and 256 codewords. Each of these quantizers can be considered as an HMM with one state and equiprobable mixture components, or alternatively, as an HMM with as many states as codewords, one mixture component per state, with all initial and state transition probabilities the same. In this experiment, the vector quantizers replace the HMM for the clean speech, and the selection of a specific codeword for a given input noisy speech was done by the "nearest neighbor" rule (in the Itakura-Saito sense) using the *clean* speech vector corresponding to the noisy input vector. Each vector of the noisy speech was filtered using a Wiener filter which was based on the chosen codeword and the AR model for the noise process. No iterations were needed here since the selection of an AR codeword from the code book for the clean speech was done using the clean signal itself and hence was the best possible codeword selection. The quality of the enhanced speech signal obtained in this manner was surprisingly good even when only 64 codewords were used. At 10 dB input SNR the enhanced speech obtained using 256 codewords sounded almost the same as the original clean speech.

The experiment described above obviously demonstrates the best performance which may never be achieved by a system of the type examined here, since the selection of an AR model for a given input noisy speech is based on the clean speech. This experiment does, however, cast light on several important aspects of the speech enhancement problem. First, it shows that the concept of representing the power spectral density of a given vector of speech by the power spectral density of a finite order AR process is adequate for speech enhancement purposes. Second, it shows that only coarse quantized versions of the power spectral densities of speech, e.g., those obtained using a 64 vector code book, are needed in speech enhancement applications. Third, the experiment proves that the proper selection of an AR codeword for a given noisy input vector is the key to the success of the algorithm. In our system, where only the noisy speech is given, the selection of an AR codeword for a given noisy input vector is performed using the Viterbi algorithm which chooses the most probable codeword based on the current noisy input vector, as well as other speech frames in the neighborhood of the analysed vector.

The above experiment provides some guidelines for choosing the number of states and the number of mixture components for each state. It shows that the product of M and L should be in the range of from 64 to 256. In our experiments we obtained the best results using 32 states and a maximum of 8 mixture components per state. The actual number of mixture component per state was automatically determined by the algorithm for initial clustering of the training sequence (see Section 3.1). We used here the rule of continuing splitting the codewords corresponding to a given state, until either the maximum number of mixture components per state is achieved or an empty cell is detected, whichever occurs first.

Table 1 shows typical SNR improvement obtained using the proposed algorithm. In these experiments we used 100 sentences of clean conversational speech, spoken by 10 different speakers, recorded using a telephone handset, for training an HMM for the clean speech. For testing we used 2 sentences from 2 speakers, where the speech material and the speakers were different from those used for training. The AR model for the noise was estimated from the actual noise sample which was added to the clean speech to produce the noisy speech. The enhancement of the entire tested speech sample was done simultaneously, i.e., in each iteration the most probable sequence of states and mixture components corresponding to the *entire* speech sample to be enhanced was first found and then the Wiener filters were applied. We used frames of 128 samples of speech, sampled at 8 kHz, which overlap each other by 64 samples. The synthesis of the enhanced signal was done using the standard overlap and add technique.

Informal listening tests showed that the noise level of the enhanced signal was significantly lower than that of the input noisy speech and this was achieved without noticeable degradation of the speech signal itself. The crispness of the original speech was preserved and no muffling of sounds, which is usually associated with enhanced speech signals, was detectable. The enhanced signal was, however, accompanied by a residual noise which sounded like a mixture of wide band noise and "musical noise." The level of the musical noise is significantly lower than that obtained using "spectral subtraction" based speech enhancement systems.

SNR IN	SNR OUT	SNR IN	SNR OUT
0.0	7.79	15.0	17.14
5.0	11.02	20.0	20.61
10.0	14.65		

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," to be published in *IEEE Trans. Inform. Theory*.

[2] L. R. Rabiner, J. G. Wilpon, and B.-H Juang, "A segmental k-means training procedure for connected word recognition," *AT&T Technical Journal*, pp. 21-40, May-June, 1986.

[3] 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.