Speech Enhancement Using Multiple Kalman Filter

(다중칼만필터를 이용한 음성향상)

Ki Yong Lee

● School of Electronic Engineering, Soongsil University, Seoul, Korea

ABSTRACT

In this paper, a Kalman filter approach for enhancing speech signals degraded by statistically independent additive nonstationary noise is developed. The autoregressive hidden markov model (ARHMM) is used for modeling the statistical characteristics of both the clean speech signal and the nonstationary noise process. In this case, the speech enhancement comprises a weighted sum of conditional mean estimators for the composite states of the models for the speech and noise, where the weights equal to the posterior probabilities of the composite states, given the noisy speech. The conditional mean estimators use a smoothing approach based on two Kalman filters with Markovian switching coefficients, where one of the filters propagates in the forward-time direction and the other one propagates in the backward-time direction with one frame. The proposed method is tested against the noisy speech signals degraded by Gaussian colored noise or nonstationary noise at various input signal-to-noise ratios. An approximate improvement of 4.7-5.2 dB in SNR is achieved at input SNR 10 and 15 dB. Also, in a comparison of conventional [7] and the proposed methods, an improvement of the about 0.3 dB in SNR is obtained with our proposed method.

I. INTRODUCTION

Speech enhancement attempts to minimize the effects of noise and to improve the performance in voice communication systems when their input signals are corrupted by background noise. Furthermore, when the noisy signal is assumed only to be available, speech enhancement requires explicit knowledge of the joint statistics of the clean speech signal and the noise process. There have been numerous studies on speech enhancement adopting the Wiener filter [1,2] and Kalman filter [3-5]. In [1], a time-varying autoregressive (AR) model is attributed to the speech signal and both the model and the speech signal are estimated from the given noisy speech signal using the maximum a posteriori probability (MAP). In [4], the Kalman filter with AR parameters based on minimum mean-square error (MMSE) estimation approach was proposed in time domain for filtering speech contaminated by additive white noise or colored noise. In this approach, the estimation is iteratively performed, once over the AR model assuming that the clean signal is available and then over the clean signal using the estimated model. In a comparison of Wiener and Kalman filter methods, the best performance was obtained with Kalman filtering [5].

Recently, the speech enhancement using hidden Markov model (HMM) is developed by Ephraim [6,7]. The speech and noise was modeled by the AR covariance matrices of the mixture HMM associated with the most likely sequence of the states and the AR covariance matrix of the white noise model, respectively. Then, the estimation of the clean speech is performed by applying time-varying Wiener filter to speech contaminated by the white noise contaminated speech. It is known that the speech quality enhanced by the HMM method is better than that obtained from the iterative Wiener filter system [1].

However, these conventional approaches are implemented for enhancing speech on the assumption that noise is additive stationary process such as white Gaussian or colored noise. If the noise is nonstationary with slowly varying statistics, we can not expect good performance in speech enhancement from those approaches. Since the Kalman filter can take advantage of the nonstationary process model and the recursive optimal estimation for real-time processing , we consider the speech enhancement using the Kalman filter with AR HMM represented by time-varying AR filters with its parameters switched by a Markov chain.

The problem addressed in this paper is a recursive method in time domain based on MMSE to enhance speech when only the speech contaminated by nonstationary noise is available for processing. To estimate the statistics of speech and noise, we use the mixture AR HMM and AR HMM with single mixture to model the speech and noise [8], respectively. Like the conventional HMM, the parameter set of the AR HMM is estimated by the maximum likelihood approach using the Baum reestimation and expectation-maximization algorithm from the given training speech and noise data [9,10]. Given the AR HMM parameter set of the speech and noise model, speech enhancement becomes a state estimation problem with the Kalman filter in a system with Markovian switching coefficient in control theory [11-13]. The state estimation is processed by the fixed interval smoothing using two Kalman filters with forward-backward direction or by Kalman filter with forward direction. The switching between the Kalman filters is governed by a finite-state Markov chain with the transition probabilities. This proposed enhancement method consists of the multiple Kalman filters and the outputs of which are weighted by a time-varying *a posteriori* probability. The coefficients of each Kalman filter consist of the ARHMM parameters of speech and noise estimated by training algorithm. Performance comparison between the proposed and conventional method accomplished in terms of signal to noise ratio (SNR) and sound spectrograms. An improvement of the approximate 4.9 dB in output SNR is achieved at input SNR with 10 dB and 15 dB.

The rest of the paper is organized as follows. In Section II, we formulate the problem and specify the speech and noise model with the HMM. In Section III, we describe the waveform based enhancement algorithm using a smoothing approach. In Section IV, we provide experimental results and conclusions

are given in Section V.

II. Speech and Noise Models

a. mixture AR hidden markov model for clean speech

To represent the statistics of the clean speech signal, we consider ARHMM with L-states and mixtures of M Gaussian AR output processes at each state. Let $y=\{y(n),\ n=1,\ldots,T\}$, $y(n)=\{y((n-1)N+1),\ \ldots,y(nN)\}$ be the observation sequence, $s=\{s(n),\ n=1,\ldots,T\}$, $s(n)\in\{1,\ldots,L\}$, be a sequence of states corresponding to y, and $h=\{h(n),\ n=1,\ldots,T\}$, $h(n)\in\{1,\ldots,M\}$, be a sequence of mixture components corresponding to (s, y). Thus, at frame n speech conditioned in $h(n)$ mixture on state $s(n)$ is expressed by a linear combination of its past values plus an excitation source, as

$$y(t)=\boldsymbol{B}_{h(n)|s(n)}^{T}\boldsymbol{Y}(t-1)+e_{h(n)|s(n)}(t),\quad (n-1)N+1\le t\le N\cdot n \tag{1}$$

where $\boldsymbol{B}_{h(n)|s(n)}^{T}=\left[b_{h(n)|s(n)}(1),\ldots,b_{h(n)|s(n)}(p)\right]$ is the vector of AR coefficients, $\boldsymbol{Y}(t-1)=\left[y(t-1)\ldots y(t-p)\right]^{T}$ is the sequence of the past p observations, and the excitation source $e_{h(n)|s(n)}(t)$ is Gaussian i.i.d. process with zero mean and variance $\sigma^{2}_{h(n)|s(n)}$.

The likelihood $p(y)$ of the observation sequence is obtained as

$$p(y)=\sum_{s}\sum_{h}p(s,h,y)=\sum_{s}\sum_{h}\prod_{n=1}^{T}a_{s(n-1)s(n)}\,c_{h(n)|s(n)}\,p\left(y(n)\middle|h(n),s(n)\right) \tag{2}$$

where $a_{s(n-1)s(n)}$ denotes the transition probability from state $s(n-1)$ at fram $n-1$ to state $s(n)$ at time n, and $c_{h(n)|s(n)}$ is the probability of choosing the mixture $h(n)$ provided that the process is in state $s(n)$. According to (1), the transformation from the excitation sequences $\left\{e_{h(n)|s(n)}(t),\ (n-1)N+1\le t\le nN\right\}$ to y has Jacobian 1. Given $(h(n),\ s(n))$, the conditional pdf $p\left(y(n)\middle|h(n),s(n)\right)$ is given by

$$p\left(y(n)\middle|h(n),s(n)\right)=\prod_{t=(n-1)N+1}^{nN}\frac{1}{\sqrt{2\pi}\sigma_{h(n)|s(n)}}\cdot\exp\left\{-\frac{\left(y(t)-\boldsymbol{B}_{h(n)|s(n)}^{T}\boldsymbol{Y}(t-1)\right)^{2}}{2\sigma^{2}_{h(n)|s(n)}}\right\} \tag{3}$$

The parameter set $\lambda_{y}=\left\{a_{ij},\ c_{m|j},\ B_{m|j},\ \sigma_{m|j},\ i,j=1,\ldots L\right.$ and $\left.m=1,\ldots,M\right\}$ of the ARHMM for the clean speech is estimated from training sequences of clean speech signals. As with the standard mixture ARHMM [8], we used the Baum-Welch algorithm [9] for parameter estimation.

b. HMM for noise signal

Assume that the noise is additive and statistically independent of the speech signal. Previous works assumed that noise was stationary white noise or stationary colored noise. However, the real noise, such as computer fan noise and car noise etc., generally has the characteristics of exhibits nonstationary with time-varying statistics. To model the nonstationary noise, we consider HMM with K states for the noise process. Let $v=\{v(n),\ n=1,\ldots,T\}$, $v(n)=\{v((n-1)N+1),\ldots,v(nN)\}$ be the observation sequence, $x=\{x(n),\ n=1,\ldots,T\}$, $x(n)\in\{1,\ldots,K\}$, be a sequence of states corresponding to v. Then the noise $v(t)$ is modeled by an AR process with order q conditioned on state k as

$$v(t)=\boldsymbol{C}_{x(n)}^{T}\boldsymbol{V}(t-1)+w_{x(n)}(t),\ (n-1)N+1\le t\le nN \tag{4}$$

where $\boldsymbol{C}_{x(n)}^{T}=\left[c_{x(n)}(1),\ldots,c_{x(n)}(q)\right]$ is the vector of AR coefficients, $\boldsymbol{V}(t-1)=\left[v(t-1)\ldots v(t-q)\right]^{T}$ is the sequence of past q observations, and $\sigma^{2}_{x(n)}$ is the variance of the innovations process of an AR source. The pdf $p_{\lambda_v}(v)$ of noise is given by

$$p_{\lambda_v}(v)=\sum_{x}p_{\lambda_v}(v,x)=\sum_{x}\prod_{n=1}^{T}\tilde{a}_{x(n-1)x(n)}\,p_{\lambda_v}\left(v(n)\middle|x(n),\boldsymbol{V}_n(0)\right) \tag{5}$$

where $\tilde{a}_{x(n-1)x(n)}$ denotes the transition probability from state $x(n-1)$ at time instant $n-1$ to state $x(n)$ at n, and $p\left(v(n)\middle|x(n),\boldsymbol{V}_n(0)\right)$ is the conditional pdf of the output $v(n)$ given the sequence $x(n)$ of noise states and initial values $\boldsymbol{V}_n(0)=\{v(1-q),\ldots,v(0)\}$:

$$p\left(v(n)\middle|x(n)=i\right)=\prod_{t=(n-1)N+1}^{nN}\frac{1}{\sqrt{2\pi}\sigma_{x(n)}}\cdot\exp\left\{-\frac{\left(v(t)-\boldsymbol{C}_{x(n)}^{T}\boldsymbol{V}(t-1)\right)^{2}}{2\sigma^{2}_{x(n)}}\right\} \tag{6}$$

The parameter set $\lambda_{v}=\left\{\tilde{a}_{ij},\boldsymbol{C}_{j},\sigma_{j},i,j=1,\ldots,K\right\}$ of the HMM for the noise is also estimated using the Baum algorithm for speech model.

The noise model (4) becomes stationary white noise model of [6,7] or colored noise model of [4] for $K=1$ and the standard stationary white Gaussaian noise model [4] for $K=1$ and $q=0$, respectively.

III. Speech Enhancement using smoothing approach

In this section, we derive the speech enhancement using the Kalman filter with a priori knowledge of both speech and noise statistics from section II. Both speech and noise are represented by AR models. We assume that only the noisy speech sequence $z(n)=\{z(t),(n-1)N+1\le t\le nN\}$ is available for speech enhancement, represented by

$$z(n)=y(n)+v(n),\qquad n=1,2,\ldots,T \tag{7}$$

where $y(n)=\{y(t),(n-1)N\le t\le nN\}$ and $v(n)=\{v(t),(n-1)N\le t\le nN\}$. Note that the indexing on nN

is n-th frame with blocklength N.

Then, the MMSE signal estimator $\hat{y}(n)$ of clean speech $y(n)$, given $\boldsymbol{Z}(n)=\{z(n)\cdots z(1)\}$, can be written using the Bayes theorem as follows [14]:

$$\begin{aligned}\hat{y}(n) &= E\left[y(n)\middle|z(n), z(n-1),\ldots,z(1)\right] \\ &= \sum_{y(n)} y(n)p\left(y(n)\middle|\boldsymbol{Z}(n)\right)\end{aligned} \tag{8}$$

The $p\left(y(n)\middle|\boldsymbol{Z}(n)\right)$ in (8) is the conditional pdf of the clean speech signal $y(n)$ given the noisy signal $\boldsymbol{Z}(n)$ and can be derived similarly to [7]. Then, $p\left(y(n)\middle|\boldsymbol{Z}(n)\right)$ can be expressed

$$p\left(y(n)\middle|\boldsymbol{Z}(n)\right)=\sum_{\bar{x}(n)} p\left(y(n)\middle|\boldsymbol{Z}(n),\bar{x}(n)\right)p\left(\bar{x}(n)\middle|\boldsymbol{Z}(n)\right) \tag{9}$$

where $p\left(\bar{x}(n)\middle|\boldsymbol{Z}(n)\right)$ is the conditional probability of the composite state $\bar{x}(n)$ of the noisy signal at time n given the noisy signals, and $p\left(y(n)\middle|\boldsymbol{Z}(n),\bar{x}(n)\right)$ is the conditional pdf of the clean signal at time n given the noisy signal and its composite state at time n. Combining (8) and (9) yields upon interchanging the order of summation

$$\hat{y}(n)=\sum_{\bar{x}(n)} \hat{y}_{\bar{x}(n)}(n)p\left(\bar{x}(n)\middle|\boldsymbol{Z}(n)\right) \tag{10}$$

where

$$\begin{aligned}\hat{y}_{\bar{x}(n)}(n) &= \sum_{y(n)} y(n)p\left(y(n)\middle|\boldsymbol{Z}(n),\bar{x}(n)\right) \\ &= E\left[y(n)\middle|\boldsymbol{Z}(n),\bar{x}(n)\right]\end{aligned} \tag{11}$$

which can be computed by using a vector Kalman filtering algorithm. Since an estimate of the vector is produced at each time instant n, we call a direct implementation of (11) the vector Kalman filter.

However, using the fact that is a Gauss-Markov process we need only the conditioning of being in the composite state at time n instead of its entire part history [15,16].

Therefore, $\hat{y}_{\bar{x}(n)}(t)$, $(n-1)N+1\le t\le nN$, can be obtained recursively by a smoother using two Kalman filters with forward and backward direction or conventional Kalman filter conditioned on the composite state. To develop a speech enhancement algorithm based on Kalman filtering and the assumption of nonstationary noise, we begin with the mixture HMM with AR source model in (1) and the observation model in (7), and reformulate them into a canonical state space form with Markov switch sequences $(s(n), h(n), x(n))$ at n-th frame as:

$$\bar{\boldsymbol{Y}}(t)=\Phi\left(s(n),h(n),x(n)\right)\bar{\boldsymbol{Y}}(t-1)+G\bar{e}\left(s(n),h(n),x(n)\right) \tag{12}$$

$$z(t)=H^T\bar{\boldsymbol{Y}}(t) \tag{13}$$

where $\bar{\boldsymbol{Y}}(t)=\begin{bmatrix}\boldsymbol{Y}(t)\\ V(t)\end{bmatrix}$, with $\boldsymbol{Y}(t)=\left[y(nN-t)\ldots y(nN-t-p+1)\right]^T$ and $\boldsymbol{V}(t)=\left[v(nN-t)\ldots v(nN-t-q+1)\right]^T$,

$$\Phi\left(s(n),h(n),x(n)\right)=\begin{bmatrix}\Phi_y\left(s(n),h(n)\right) & \boldsymbol{0}\\ \boldsymbol{0} & \Phi_v\left(x(n)\right)\end{bmatrix},$$

$$\Phi_y\left(s(n),h(n)\right)=\begin{bmatrix}\boldsymbol{B}_{h(n)|s(n)} & \\ \boldsymbol{0} & \boldsymbol{I}\end{bmatrix},\ \Phi_v\left(x(n)\right)=\begin{bmatrix}C_{x(n)}^T & \\ \boldsymbol{0} & \boldsymbol{I}\end{bmatrix},$$

$$\bar{e}\left(s(n),h(n),x(n)\right)=\begin{bmatrix}e\left(s(n),h(n)\right)\\ w\left(x(n)\right)\end{bmatrix},\quad G=\begin{bmatrix}G_y & \boldsymbol{0}\\ \boldsymbol{0} & G_v\end{bmatrix},$$

$H^T=\left[H_y{}^T\ \ H_v{}^T\right]$, $G_y=H_y{}^T=[10\ldots0]$ $G_v=H_v{}^T=[10\ldots0]$.

We assume that $e(s(n),h(n))$ and $w(x(n))$ are uncorrelated so that

$$\begin{aligned}\boldsymbol{Q}\left(s(n),h(n),x(n)\right) &= E\left[\bar{e}\left(s(n),h(n),x(n)\right)\bar{e}^T(\cdot)\right] \\ &= \begin{bmatrix}\sigma_{y_{h(n)|x(n)}}^2 & 0\\ 0 & \sigma_{v_{x(n)}}^2\end{bmatrix}\end{aligned}$$

Eq. (12) and (13) bears the form of a linear system with a set of AR coefficients associated with Markov states driven by a white Gaussian noise process with variances following Markov states. Also, this state-space model is the model of so-called the noise free or perfect measurements problem in the estimation literature [18].

Since the composite state sequence $\bar{x}(n)$ is $\{s(n),h(n),x(n)\}$, the estimate $\hat{\bar{\boldsymbol{Y}}}(t)$ is obtained by

$$\begin{aligned}\hat{\bar{\boldsymbol{Y}}}(t) &= \sum_{k=1}^{K}\sum_{j=1}^{L}\sum_{m=1}^{M}\hat{\bar{\boldsymbol{Y}}}_{m|j,k}(t) \\ &\quad\cdot p\left(s(n)=j,h(n)=m,x(n)=k\middle|\boldsymbol{Z}(n)\right)\end{aligned} \tag{14}$$

This estimator comprises a weighted sum of conditional mean estimator using a Kalman filter for the composite states of the signal and noise, where the weights are the probabilities of these states given the noisy signal. If the state sequences of the speech and noise for a given noisy speech are known, the most appropriate filter from the predesigned set of filters can be applied to the noisy speech and optimal estimation of clean speech signal can be performed. There are $L\times M\times K$ possible state sequences.

Therefore, the estimate of the state vector $\hat{\bar{\boldsymbol{Y}}}(t)$ becomes a weighted sum of the individual Kalman estimators $\hat{\bar{\boldsymbol{Y}}}_{m|j,k}(t)$, where the weights are *a priori probabilities* of the $L\times M\times K$ composite states. Therefore, the problem of speech enhancement in (14) is divided into the estimation of $\hat{\bar{\boldsymbol{Y}}}_{m|j,k}(t)$ and computation of $p\left(s(n)=j,h(n)=m,x(n)=k\middle|\boldsymbol{Z}(n)\right)$

A. The estimation of $\hat{\bar{\boldsymbol{Y}}}_{m|j,k}(t)$

Each sample estimate of $\hat{\bar{\boldsymbol{Y}}}_{m|j,k}(t)$ is obtained by using forward-backward Kalman filtering to further improve state smoothing. Then, the smoothed estimate and covariance can be expressed as

$$\begin{aligned}\hat{\bar{\boldsymbol{Y}}}_{m|j,k}(t|N) &= P_{m|j,k}(t|N)\Big[\left(P_{m|j,k}^{f}(t|t)\right)^{-1}\hat{\bar{\boldsymbol{Y}}}_{m|j,k}^{f}(t|t) \\ &\quad+\left(P_{m|j,k}^{f}(t|t)\right)^{-1}\hat{\bar{\boldsymbol{Y}}}_{m|j,k}^{f}(t|t)\Big]^{-1}\end{aligned} \tag{15}$$

$$P_{m|j,k}(t|N)=\left[\left(P^{f}_{m|j,k}(t|t)\right)^{-1}+\left(P^{b}_{m|j,k}(t|t+1)\right)^{-1}\right]^{-1} \quad (16)$$

This form of the fixed-interval smoother is referred to as the two-filter form because the smoothed estimate $\hat{\bar{Y}}_{m|j,k}(t|N)$ is obtained as the combination of the forward time filtered estimate $\hat{\bar{Y}}^{f}_{m|j,k}(t|t)$ and the estimate $\hat{\bar{Y}}^{b}_{m|j,k}(t|t+1)$ generated using a backward-time filter.

-Forward Kalman filtering

$$\hat{\bar{Y}}^{f}_{m|j,k}(t)=\Phi(j,m,k)\hat{\bar{Y}}^{f}_{m|j,k}(t-1)+K^{f}_{m|j,k}(t)\left\{z(t)-H^{T}\Phi(j,m,k)\hat{\bar{Y}}^{f}_{m|j,k}(t-1)\right\}, \quad (17)$$

$$M^{f}_{m|j,k}(t)=\Phi(j,m,k)P^{f}_{m|j,k}(t-1)\Phi^{T}(j,m,k)+GQ(j,m,k)G^{T},$$

$$K^{f}_{m|j,k}(t)=M^{f}_{m|j,k}(t)H\left[H^{T}M^{f}_{m|j,k}(t)H\right]^{-1},$$

$$P^{f}_{m|j,k}(t)=M^{f}_{m|j,k}(t)-K^{f}_{m|j,k}(t)HM^{f}_{m|j,k}(t).$$

- Backward Kalman filtering

$$\hat{\bar{Y}}^{b}_{m|j,k}(t|t)=\Phi(j,m,k)\hat{\bar{Y}}^{b}_{m|j,k}(t+1|t+1)+K^{b}_{m|j,k}(t)\cdot\left\{z(t)-H^{T}\Phi(j,m,k)\hat{\bar{Y}}^{b}_{m|j,k}(t+1|t+1)\right\}, \quad (18)$$

$$M^{b}_{m|j,k}(t-1|t)=\Phi(j,m,k)P^{b}_{m|j,k}(t|t)\Phi^{T}(j,m,k)+GQ(j,m,k)G^{T},$$

$$K^{b}_{m|j,k}(t)=M^{b}_{m|j,k}(t|t+1)H\left[H^{T}M^{b}_{m|j,k}(t|t+1)H\right]^{-1},$$

$$P^{b}_{m|j,k}(t|t)=M^{b}_{m|j,k}(t|t+1)-K^{b}_{m|j,k}(t)HM^{b}_{m|j,k}(t|t+1)$$

where $Q(s(n)=j,h(n)=m)=\sigma^{2}_{m|j}I$ is the covariance matrix of $e(s(n)=j,\ h(n)=m)$. Since the $\left[H^{T}M^{f}_{m|j,k}(t)H\right]^{-1}$ and $\left[H^{T}M^{b}_{m|j,k}(t)H\right]^{-1}$ have the scalar value, it excludes the need of inverse matrix procedure.

b. The computation of $p\left(s(n)=j,\ h(n)=m,x(n)=k|Z(n)\right)$

The weighting factor $p\left(s(n)=j,\ h(n)=m,x(n)=k|Z(n)\right)$ becomes, using $Z(n)=\{z(n),\ Z(n-1)\}$ and Bayes rule:

$$p\left(s(n)=j,\ h(n)=m,x(n)=k|Z(n)\right)=\frac{p\left(z(n)|s(n)=j,\ h(n)=m,\ x(n)=k,Z(n-1)\right)}{p\left(z(n)|Z(n-1)\right)}\cdot p\left(s(n)=j,\ h(n)=m,x(n)=k|Z(n-1)\right). \quad (19)$$

The first term $p\left(z(n)|s(n)=j,\ h(n)=m,x(n)=k,\ Z(n-1)\right)$ is the conditional probability density of the observation $z(n)$, if the past observations $Z(n-1)$ and the particular state sequence $\{s(n)=j,\ h(n)=m,\ x(n)=k\}$ are given. This can be approximated to a Gaussian density, where the mean and covariance can be calculated by using the Kalman filter matched to the sequence $s(n)$, $h(n)$, and $x(n)$, i.e.,

$$p\left(z(n)|s(n)=j,\ h(n)=m,x(n)=k,Z(n-1)\right)=\prod_{t=1}^{N}p\left(z(t)|s(n)=j,\ h(n)=m,x(n)=k\right)$$

where $p\left(z(t)|s(n)=j,\ h(n)=m,x(n)=k\right)=N\left[\hat{\bar{Y}}_{m|j,k}(t),\ HP_{m|j,k}(t)H^{T}\right]$ (20)

where $N[\ldots]$ denotes a normal distribution.

Since the $(s(n)=j,\ h(n)=m)$ and $x(n)=k$ are mutually independent, we can recast the second term in (19) as

$$p\left(s(n)=j,h(n)=m,x(n)=k|Z(n-1)\right)=p\left(s(n)=j,\ h(n)=m|Z(n-1)\right)\cdot p\left(x(n)=k|Z(n-1)\right) \quad (21)$$

The first term in (21) is the predicted probability that will be generated by the Markov process,

$$p\left(s(n)=j,h(n)=m|Z(n-1)\right)=\sum_{i=1}^{I}\sum_{n=1}^{M}p\left(s(n)=i,\ h(n)=m|s(n-1)=i,h(n-1)=n,Z(n-1)\right)\cdot p\left(s(n-1)=i,\ h(n-1)=n|Z(n-1)\right), \quad (22)$$

where we can rewrite the first term as

$$p\left(s(n)=j,h(n)=m|s(n-1)=i,h(n-1)=n,Z(n-1)\right)=p\left(h(n)=m|s(n)=j,s(n-1)=i,h(n-1)=n,Z(n-1)\right)\cdot p\left(s(n)=j|s(n-1)=i,h(n-1)=n,Z(n-1)\right). \quad (23)$$

Since the $h(n)$ and $s(n)$ are independent of $Z(n-1)$, and the probability law for the Markovian chain s(t) is completely specified by the transition probabilities, the first and second term in (23) is rewritten respectively as

$$p\left(h(n)=m|s(n)=j,s(n-1)=i,h(n-1)=n,Z(n-1)\right)=c_{j|m} \quad (24)$$

and

$$p\left(s(n)=j|s(n-1)=i,h(n-1)=n,Z(n-1)\right)=p\left(s(n)=j|s(n-1)=i\right)=a_{ij} \quad (25)$$

Substituting (24) and (25) into (22) yields

$$p\left(s(n)=j,h(n)=m|Z(n-1)\right)=\sum_{i=1}^{I}\sum_{n=1}^{M}c_{m|j}a_{ij}\cdot p\left(s(n-1)=i,\ h(n-1)=n|Z(n-1)\right). \quad (26)$$

Similarly as in (24)-(26), the second term in (21) is written as

$$p\left(x(n)=k|Z(n-1)\right)=\sum_{l=1}^{K}\tilde{a}_{kl}p\left(x(n-1)=l|Z(n-1)\right) \quad (27)$$

Therefore, substituting (25) and (26), (21) is rewritten as

$$p\left(s(n)=j,h(n)=m,x(n)=k|Z(n-1)\right)=\sum_{i,\ l=1}^{I}\sum_{n=1}^{M}\sum_{l=1}^{K}\tilde{a}_{kl}c_{m|j}\cdot a_{ij}p\left(s(n-1)=i,\ h(n-1)=n,x(n-1)=k|Z(n-1)\right) \quad (28)$$

Since the denominator term of (19) is independent of j and m, it becomes a scale factor. Therefore, weighting factor $p\left(s(n)=j,h(n)=m,x(n)=k|Z(n)\right)$ can be calculated

recursively using the previous weighting factor as

$$p\left(s(n)=j,h(n)=m,x(n)=k\middle|Z(n)\right)$$
$$=D_n\cdot N_{m|j,k}\sum_{i=1}^{L}\sum_{l=1}^{M}\sum_{t=1}^{K}\tilde{a}_{kl}c_{m|j}a_{ij}$$
$$p\left(s(n-1)=i,h(n-1)=n,x(n-1)=k\middle|Z(n-1)\right) \quad (29)$$

where D_n is a scale factor determined at time t, and guarantees that the sum of all the weighting factors is equal to one;

$$\sum_{k=1}^{K}\sum_{j=1}^{L}\sum_{m=1}^{M}p\left(s(n)=j,h(n)=m,x(n)=k\middle|Z(n)\right)=1. \quad (30)$$

Finally, item needed in the computation of (29) is the initial probability of falling to each state of speech and noise model at time zero. However, in our experiments it was found that the recursive method is relatively insensitive to the choice of initial probabilities of state.

Then, the enhanced speech signal $\hat{y}(t)$ is equal to the first component of the estimated $\hat{\bar{Y}}(t)$ as

$$\hat{y}(t)=\left[\underbrace{1}_{p}\underbrace{0\ldots0}_{q}\right]\hat{\bar{Y}}(t|N) \quad (31)$$

or $$\hat{y}(t)=\left[\underbrace{0\ldots0}_{p-1}\underbrace{1}_{1}\underbrace{0\ldots0}_{q}\right]\hat{\bar{Y}}(t+p-1|N).$$

IV. EXPERIMENTAL RESULTS

The proposed enhancement approach was examined in enhancing speech signals degraded by statistically independent additive stationary Gaussian colored noise and nonstationary noise at the input signal-to-noise ratio (SNR) with 0, 5, 10, 15, and 20 dB. The input SNR is defined as the ratio of the average power of the signal to the average power of the noise.

Training for mixture AR HMM of clean speech was performed using 8 min of conventional speech from 8 speakers, e.g., 4 males and 4 females. The speech is sampled at 12kHz and observation vectors are formed by applying a Hamming window of 256 samples without overlap. The order of each AR model is 12, which is a commonly used value in linear predictive analysis of speech signals. In enhancement test, neither the speakers nor the speech material used for testing were in the training set. The test data consisted of three sentences originally spoken by a male and a female. Then, the speech sequence for enhancement recorded in a manner similar to that of the training.

First, we examined the performance of the proposed method under colored noise. For a colored noise, we used the car noise sequences. The model for the colored noise process was estimated directly from the noisy speech, using an initial interval in which speech was not present. Then, the model is assumed to be one-state AR HMM with 8-th order. Table 1 shows the performance of proposed method with various states number, mixture components for each state, and input SNRs under the colored noise with single state. The best enhancement results were obtained using the eight-state six-mixture model. Table 2 show the performance comparisons between the proposed method and the conventional method based HMM using the Weiner filter for the eight-state six mixture model at various input SNR values. An approximate improvement of 0.3 dB in output_SNR is achieved at SNR 10 and 15 dB, compared to the speech enhancement method based on the HMM with Weiner filter. Although the output_SNR is slightly improved by the proposed method, we can not distinguish the difference from two methods by informal listening test.

Second, we examined the performance of the proposed method under nonstationary noise. The nonstationary noise for testing and training was artifically generated by the randomly switching of two AR model with 8-th order. The two-state AR HMM with 8-th order trained to model the noise process from the generated noise signal. Table 3 shows the performance of the proposed method with the additive nonstationary noise at input signal-to-noise ratio (SNR) values with 0, 5, 10, 15, and 20 dB. The proposed method yields good results, too. However, as the conventional method [7], the proposed speech enhancement method was also found less effective at the low input SNR with 5 dB, although the output SNR is 12.2 dB.

In this experiment, we assume correct knowledge about the statistics of excitation source and noise variance. In general, however, since the recording conditions during training and testing may be different, the variances of excitation source in speech model and noise source are unknown or perfectly unknown. When noisy speech signal was degraded by noise signal with unknown variance, the proposed method had poor results and even diverge when the input has low SNR. As the input SNR increases, the output_SNR of the proposed method with no knowledge of the noise statistics was improved and the adverse phenomena mentioned above were significantly reduced. Therefore, when the statistics of noise model structure are known, the proposed method would produce the same good results on colored noise or nonstationary noise.

V. CONCLUSION

We proposed a new approach in time domain for enhancing speech signals degraded by statistically independent additive stationary or nonstationary colored Gaussian noise. A speech enhancement is developed by MMSE estimation based on the estimated statistics of the both speech and noise process from long training sequence. We used a HMM with mixtures of Gaussian AR output probability distributions and a HMM with single mixture to model the speech and noise, respectively. The mixture models are equivalent to a large HMM with simple states, together with additional constraints on the possible transitions between states. The parameter set of the ARHMM for the speech and noise is estimated by the maximum likelihood approach using the Baum reestimation algorithm from the given training speech and noise data. When the noisy speech signal is assumed only available, then the MMSE estimation for speech enhancement comprises a weighted sum of conditional mean estimators for the composite states of the models for the speech and noise, where the weights equal the posterior probabilities of the composite states given the noisy speech. The conditional mean estimators use a smoothing approach based on two Kalman filters with Markovian switching coefficients, where one of the filters propagates in the forward-time direction and the other one propagates in the backward-time direction in one frame. This enhancement algorithm using the modified Kalman filtering algorithm is easier to implement than the HMM based on Weiner filter since it is a noniterative estimator. This approach does not require the transformation of speech in the enhancement procedure like conventional HMM with Weiner filter. In our experimental test, we obtain the performance about 14.8-15.3 and 19.0-19.5 dB at input SNR 10 and 15 dB under nonstationary or stationary noise, respectively.

REFERENCES

[1] J.S. Lim and A.V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 197-210, June 1978.

[2] J.H.L. Hansen and M.A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Signal Processing*, vol.39, no.4, pp. 795-805, Apr. 1991.

[3] B.G. Lee, K.Y. Lee, and S. Ann, "An EM-based approach for parameter enhancement with an application to speech signals," *Signal Processing*, vol. 46, pp. 1-14, 1995.

[4] J.D. Gibson, B. Koo, and S.D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. Signal Processing*, vol. 39, pp. 1732-1742, Aug. 1991.

[5] K.K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in Proc. *IEEE Int. Conf., Speech, Signal Processing (Dallas, TX)*, Apr. 6-9. 1987, pp. 177-180.

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

[7] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov model," *IEEE Trans. Signal Processing*, vol. SP-41, pp. 725-735, Apr. 1992

[8] B-H. Juang and L.R. Rabiner, "Mixture autoregressive hidden Markov models for speech signals," *IEEE Trans. Acoust. Speech, Signals Proc*, vol. 33, pp. 1404-1414, Dec. 1985.

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

[10] 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, no.1, pp. 1-38, 1977

[11] D.T. Magil, "Optimal adaptive estimation of sampled stochastic processes," IEEE *Trans. Automatic Control*, vol. AC-10, pp. 434-439, Oct. 1965.

[12] G.A. Ackerson and K.S. Fu, "On state estimation in switching environments," IEEE Trans. Automat. Contr., vol.AC-15, pp. 10-17, Feb. 1970.

[13] K. Watanabe, *Adaptive Estimation and Control: Partitioning Approach*. Prentice-Hall International, 1992.

[14] A.P. Sage and J.L. Melsa, *Estimation Theory with Applications to Communications and Control*, New York: McGraw-Hill, 1971.

[15] R.L. Moose, "An adaptive state estimation solution to maneuvering target problem," *IEEE Trans. Automatic Control*, vol. AC-20, pp. 359-362, June 1975.

[16] T.K. Tugnait, "Detection and estimation for abruptly changing systems," *Automatica*, 18, pp. 607-615, Sept. 1982.

[17] B.D.O. Anderson and J.B. Moore, Optimal Filtering. Englewood Cliffs, NJ:Prentice-Hall, 1979.

[18]. K.Y. Lee and K. Shirai, "Efficient recursive estimation for speech enhancement in colored noise," *IEEE Signal Proc. Letters*, vol. 3, no.7, 1996.

Tabel 1. Output_SNRs for different number of state/mixture under various input SNRs.

State/mixture	input SNRs (dB)			
	0	5	10	15
4 / 4	7.9	11.3	14.8	18.9
8 / 4	8.3	11.8	15.1	19.1
8 / 6	8.8	12.2	15.5	19.3
12 / 4	8.5	12.2	15.4	19.2
12 / 6	8.8	12.1	15.5	19.4

Tabel 2. Comparisions of output_SNRs between the conventional HMM and proposed method with state 8/4 under various input SNRs.

Input SNRs (dB)	Methods	
	HMM with Weiner filter	proposed method
0	8.5	8.8
5	11.8	12.2
10	15.2	15.4
15	18.9	19.2

Table 3. Output_SNRs of the proposed method under nonstationary noise

Noise model	Input SNRs				
	0	5	10	15	20
two-state HMM	8.7	12.2	15.1	19.5	22.2