A Nonstationary Autoregressive HMM and Its Application to Speech Enhancement

Ki Yong Lee* and Jae Yeol Rheem**
*Dept. of Electronics Engr., Changwon National University Changwon, Kyungnam-Do 641-773, Korea
Tel. +82-551-79-7527, Fax:+82-551-81-5070, E-mail: kylee@sarim.changwon.ac.kr
** Dept. of Electronics Engr., Korea Institute of Technology and Education, Chonan 330-600, Korea

ABSTRACT

Since speech sounds, such as fricative, glides, liquids, diphthongs, and transition regions between phones, reveal the most notable nonstationary nature, we propose the nonstationary autoregressive (AR) HMM with state-dependent polynomial function for modeling the nature of speech. Then, the nonstationary AR model has parameters depend on the states of the Markov chain. It is designed to handle the speech signal at the frame level, where it is represented by the signal, rather than dealing with feature vectors directly. Also, we proposed a new speech enhancement based on the nonstationary AR HMM and the IMM algorithm under white noise condition. The proposed enhancement is the weighted sum of the parallel Kalman filters with interacting rule by IMM algorithm. The simulation results shows that the proposed method offers performance gains relative to the previous results [7] with slightly increased complexity.

I. Introduction

Speech enhancement is the problem of enhancing a given sample function of noisy speech signal to improve the performance of voice communications whose input signal is noisy. When the noisy signal is only assumed available for processing, speech enhancement becomes a set of particular in estimation and information theory. The solutions could be found if the joint statistics of the signal and noise were explicitly available [1]. The hidden Markov model (HMM) and hidden filter model (HFM) are useful methods to estimate probability distributions of speech and noise signal in speech enhancement. In the standard HMM [2,3] and HFM [4,5], individual states are assumed to be stationary stochastic sequence. This stationary-state assumption appears to be reasonable when a state is intended to represent piece-wise stationary segment of speech. Since speech sounds, such as fricative, glides, liquids, diphthongs, and transition regions between phones, reveal the most notable nonstationary nature, we can not expect to obtain the better performance by the conventional speech enhancement methods based on such assumption.

Recently, to overcome these problems, an approach based on the trend HMM [6] is suggested for speech enhancement [7]. In this approach speech signal is blocked by samples into fixed-length frames and each frames are modeled by time-varying autoregressive model controlled by Markov switching sequence. Given the trended HMM trained from clean speech, a recursive estimation for speech enhancement comprises a weighted sum of Kalman filters operating separately in parallel. Thus the interactions between the parallel filters are ignored.

In this study, we propose the nonstationary autoregressive (AR) HMM with state-dependent polynomial function for modeling the nature of speech. Then, the nonstationary AR model has parameters depend on the states of the Markov chain. Our model is formally very similar to the trend HMM [3], but it is designed to handle the speech signal at the frame level, where it is represented by the signal, rather than dealing with feature vectors directly. We also propose a new speech enhancement based on the nonstationary AR HMM and the interactive multi model (IMM) algorithm under white noise condition. In this approach the estimator of speech is the weighted sum of the parallel Kalman filters. As the IMM algorithm handles the interactions between the parallel filters in an efficient way, enhancement performance is improved about 0.8 dB without much increase in complexity.

II. The Nonstationary AR HMM for Speech Model

In nonstationary AR HMM, speech signal is blocked by samples into fixed-length frames and modeled by nonstationary AR model with frame-varying polynomial function controlled by Markov switching sequences at each frame. It is designed to handle the speech signal at the frame level, where it is represented by the signal, rather than dealing with feature vectors directly. Then, at n-th frame speech signal conditioned on state i is

---

This work was supported by KOSEF(971-0917-105-1)

expressed as a linear combination of its past valued plus excitation source as

$$y\left(N\left(n-1\right)+t\right)=\sum_{k=1}^{p}\sum_{m=0}^{M}B_i^k\left(m\right)n^m y\left(N\left(n-1\right)+t-k\right)$$
$$e_i\left(N\left(n-1\right)+t\right).\quad t=1,\ldots,N \tag{1}$$

where $\sum_{m=0}^{M}B_i^k\left(m\right)n^m$ in the first term is the state-dependent polynomial function of order M, the second term $e_i\left(\cdot\right)$ is the excitation source with state-dependent variance σ_i^2. N and n is the frame length and number, respectively.

The parameter set

$$\lambda=\left\{a_{ij},\left[\sum_{m=0}^{M}B_i^1\left(m\right)n^m,\sum_{m=0}^{M}B_i^2\left(m\right)n^m,\ldots,\sum_{m=0}^{M}B_i^p\left(m\right)n^m\right]^T,\sigma_i^2,i,j=1,\ldots,L\right\}$$

of the nonstationary AR-HMM for clean speech is estimated from training of speech signals, where a_{ij} is transition probability. We define the observation sequence $O_1^T=\left\{o_1,o_2,\ldots o_T\right\}$, where $o_n=\left\{y\left(N\left(n-1\right)+1\right),y\left(N\left(n-1\right)+2\right),\ldots,y\left(Nn\right)\right\}$ and T is number of total frame. As with the standard HMM, we used the expectation-maximization (EM) or Baum-Welch algorithm for parameter estimation. Each iteration of the EM algorithm starts with an old set of parameters, say λ_0, and estimation a new set of parameters, say λ, by maximizing the following objective function with training sequences of speech:

$$\begin{aligned}Q\left(\lambda,\lambda_0\right)&=\sum_S p\left(S\middle|O_1^T,\lambda_0\right)\log p\left(O_1^T,S\middle|\lambda\right)\\&=\sum_{i,j=1}^{L}\sum_{n=1}^{T}\gamma_{ij}\left(n\right)\left[\log a_{ij}\right.\\&\left.+\log\prod_{t=1}^{N}\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{\left(e_i\left(N\left(n-1\right)+t\right)\right)^2}{2\sigma_i^2}\right)\right]\end{aligned} \tag{2}$$

where $\boldsymbol{S}=\left\{s(1),s(2),\ldots,s(T)\right\}$ is state sequence and $\gamma_{ij}\left(n\right)$ is a posteriori probability of the transition from state i to state j given the observation sequence and the model λ_0.

For $M=0$, the nonstationary AR HMM becomes the standard HMM. Given the clean speech, we can easily estimate a nonstationary AR HMM parameter by EM based training algorithm [9].

III. Speech Enhancement Algorithm

We assume the noisy speech is only available for processing as

$$z(t)=y(t)+v(t) \tag{3}$$

where $z(t)$ is noisy speech and $v(t)$ white noise with zero-mean and variance σ_v^2.

Let $\mathbf{Z}^n=\left\{z_1,\ldots,z_n\right\}, z_n=\left[z\left(\left(n-1\right)N\right)\cdots z\left(n\cdot N-1\right)\right]$ be measurement sequence vector. Using the law of total probability, the minimum mean square error (MMSE) estimator is given by

$$\hat{Y}_n=E\left\{Y_n\middle|\mathbf{Z}^n\right\}=\sum_{i=1}^{L}E\left\{Y_n\middle|\mathbf{Z}^n,s_n=i\right\}P\left(s_n=i\middle|\mathbf{Z}^n\right) \tag{4}$$

where $Y_n=\left[y\left(\left(n-1\right)N\right)\ldots y\left(n\cdot N-1\right)\right]$ is speech vector and s_n is state sequence.

This estimator comprises a weighted sum of conditional mean estimator for the composite states of the signal and noise, where the weights are the probabilities of theses states given the noisy signal. Since given s_n the signal Y_n is Gaussian and the noise process is also assumed Gaussian, the conditional mean estimator $Y_{n,i}=E\left\{Y_n\middle|\mathbf{Z}^n,s_n=i\right\}$ can be independently evaluated for each element. Let $\hat{y}_{n,i}(t)$ be the t-th element of $\hat{Y}_{n,i}$. Then (4) can be written as

$$\hat{y}_n(t)=\sum_{i=1}^{L}\hat{y}_{n,i}(t)P\left(s_n=i\middle|\mathbf{Z}^n\right).$$
$$\text{for } t=(n-1)N\ldots nN-1 \tag{5}$$

In this case, the conditional mean estimator $\hat{y}_{n,i}(t)$ is obtained from $\mathbf{Z}^n$ using a Kalman filter.

From (1)-(3), we can construct state space model of the form

$$\mathbf{y}_{n,i}(t)=\Phi(n=i)\mathbf{y}_{n,i}(t-1)+\mathbf{e}_i(t) \tag{6}$$
$$z_n(t)=H\mathbf{y}_{n,i}(t)+v(t) \tag{7}$$

where $H=\left[1\quad 0\quad\cdots\quad 0\right]$

$$\mathbf{y}_{n,i}(t)=\left[y\left(N\left(n-1\right)+t\right)\ldots y\left(N\left(n-1\right)+t-p+1\right)\right]^T,$$

$$\mathbf{e}_i(t)=\left[e_i\left(N\left(n-1\right)+t\right)\quad 0\quad\cdots\quad 0\right]^T,$$

$$\Phi(n=i)=\begin{bmatrix}\sum_{m=0}^{M}B_i^1(m)n^m & \sum_{m=0}^{M}B_i^2(m)n^m & \cdots & \sum_{m=0}^{M}B_i^p(m)n^m\\ 1 & 0 & \cdots & 0\\ \vdots & 1 & \ddots & \vdots\\ 0 & \cdots & 0 & 1\end{bmatrix}$$

, and $b_{m,i}^k$ is k-th element of $B_i(m)$.

From standard Kalman filtering theory for white source and measurement noise, the state vector estimator is

$$\begin{aligned}\hat{\mathbf{y}}_{n,i}(t)&=\Phi(n=i)\bar{\mathbf{y}}_{n,i}(t-1)\\&+\mathbf{K}_i(t)\left[z(t)-H\Phi(n=i)\bar{\mathbf{y}}_{n,i}(t-1)\right]\end{aligned} \tag{8}$$

The gain and error covariance equations are

$$\mathbf{K}_i(t) = \mathbf{P}_i(t|t-1)H\left[\sigma_v^2 + H^T\mathbf{P}_i(t|t-1)H\right]^{-1} \quad (9)$$

$$\mathbf{M}_i(t) = \Phi(n=i)\overline{\mathbf{P}}_i(t-1)\Phi^T(n=i) + G^T(n=i)G(n=i) \quad (10)$$

$$\mathbf{P}_i(t) = \left[I - \mathbf{K}_i(t)H\right]\mathbf{P}_i(t|t-1) \quad (11)$$

where

$$\overline{\mathbf{y}}_{n,i}(t-1) = E\left\{\mathbf{y}_{n,i}(t-1)\middle|s_n = i, \mathbf{Z}^{n-1}\right\}, \quad (12)$$

$$\overline{\mathbf{P}}_i(t-1) = E\left\{\left[\mathbf{y}_n(t-1) - \overline{\mathbf{y}}_{n,i}(t-1)\right][\cdots]^T\middle|s_n = i, \mathbf{Z}^{n-1}\right\}, \quad (13)$$

and $\mathbf{P}_i(t|t-1)$ is an a priori error covariance matrix of i-th state Kalman filter.

To derive the equations for $\overline{\mathbf{y}}_{n,i}(t-1)$ and $\overline{\mathbf{P}}_i(t-1)$, we first introduce the following equation on the basis of the total probability law:

$$p\left(\mathbf{y}_n(t-1)\middle|s_n = i, \mathbf{Z}^{n-1}\right) = \sum_j \left\{p\left(\mathbf{y}_n(t-1)\middle|s_{n-1} = j, s_n = i, \mathbf{Z}^{n-1}\right)\right\} \cdot p\left(s_{n-1} = j\middle|s_n = i, \mathbf{Z}^{n-1}\right) \quad (14)$$

As s_n is independent of $y_n(t-1)$ if s_n is known, we easily obtain

$$p\left(\mathbf{y}_n(t-1)\middle|s_{n-1} = j, s_n = i, \mathbf{Z}^{n-1}\right) = p\left(\mathbf{y}_n(t-1)\middle|s_{n-1} = j, \mathbf{Z}^{n-1}\right) \quad (15)$$

Substituting of this and of the following:

$$p\left(s_{n-1} = j\middle|s_n = i, \mathbf{Z}^{n-1}\right) = a_{ij}\frac{p\left(s_{n-1} = j\middle|\mathbf{Z}^{n-1}\right)}{p\left(s_n = i\middle|\mathbf{Z}^{n-1}\right)} \quad (16)$$

in (14) yields

$$p\left(\mathbf{y}_n(t-1)\middle|s_n = i, \mathbf{Z}^{n-1}\right) = \sum_j a_{ij}p\left(s_{n-1} = j\middle|\mathbf{Z}^{n-1}\right) \cdot \frac{p\left(\mathbf{y}_n(t-1)\middle|s_{n-1} = j, \mathbf{Z}^{n-1}\right)}{p\left(s_n = i\middle|\mathbf{Z}^{n-1}\right)} \quad (17)$$

And the denominator of (17) can be written as

$$p\left(s_n = i\middle|\mathbf{Z}^{n-1}\right) = \sum_j a_{ij}p\left(s_{n-1} = j\middle|\mathbf{Z}^{n-1}\right) \quad (18)$$

From (12),(13) and(17), we computes the mixed initial conditions $\overline{\mathbf{y}}_{n,i}(t-1)$, $\overline{\mathbf{P}}_i(t-1)$ for the Kalman filter matched to $s_n = i$, according to the following equations:

$$\overline{\mathbf{y}}_{n,i}(t-1) = \sum_j a_{ij}\frac{p\left(s_{n-1} = j\middle|\mathbf{Z}^{n-1}\right)\hat{\mathbf{y}}_{n-1,j}(t-1)}{p\left(s_n = i\middle|\mathbf{Z}^{n-1}\right)}, \quad (19)$$

$$\overline{\mathbf{P}}_i(t-1) = \sum_j a_{ij}\frac{p\left(s_{n-1} = j\middle|\mathbf{Z}^{n-1}\right)}{p\left(s_n = i\middle|\mathbf{Z}^{n-1}\right)} \cdot \left[P_j(t-1) + \left[\hat{\mathbf{y}}_{nj}(t-1) - \overline{\mathbf{y}}_{nj}(t-1)\right][\cdots]^T\right] \quad (20)$$

In this step we introduce the approximation. In the following segments, all parameters are initialized using the values from the previous segment.

By Bayes rule, probability $P\left(s_n = i\middle|\mathbf{Z}^n\right)$ of (5) is rewritten as

$$P\left(s_n = i\middle|\mathbf{Z}^n\right) = \frac{P\left(z_n\middle|s_n = i, \mathbf{Z}^{n-1}\right)P\left(s_n = i\middle|\mathbf{Z}^{n-1}\right)}{P\left(z_n\middle|\mathbf{Z}^{n-1}\right)} \quad (21)$$

Since the first term of the numerator $P\left(z_n\middle|s_n = i, \mathbf{Z}^{n-1}\right)$ can be approximated by a Gaussian density function, the element of $P\left(z_n\middle|s_n = i, \mathbf{Z}^{n-1}\right)$ is established from the Kalman filter (8)-(10) as

$$P\left(z_n(t)\middle|s_n = i, \mathbf{Z}^{n-1}\right) \approx N\left[H\Phi(n=i)\hat{\mathbf{y}}_{n,i}(t-1), \sigma_v^2 + H^T\mathbf{P}_i(t|t-1)H\right]$$

where $N[\cdot,\cdot]$ is normal distribution function.

Then, $P\left(z_n\middle|s_n = i, \mathbf{Z}^{n-1}\right)$ is rewritten as

$$P\left(z_n\middle|s_n = i, \mathbf{Z}^{n-1}\right) \approx \prod_{t=1}^{N} P\left(z_n(t)\middle|s_n = i, \mathbf{Z}^{n-1}\right) \quad (22)$$

The second term of the numerator $P\left(s_n\middle|\mathbf{Z}^{n-1}\right)$ can be rewritten as

$$P\left(s_n\middle|\mathbf{Z}^{n-1}\right) = \sum_{j=1}^{L} P\left(s_n = i\middle|s_{n-1} = j\right)P\left(s_{n-1} = j\middle|\mathbf{Z}^{n-1}\right) = \sum_{j=1}^{L} a_{ij}P\left(s_{n-1} = j\middle|\mathbf{Z}^{n-1}\right) \quad (23)$$

Since $P\left(s_{n-1} = j\middle|\mathbf{Z}^{n-1}\right)$ is known from the previous recursive calculation and $P\left(z_n\middle|\mathbf{Z}^{n-1}\right)$, being common to all terms, act as a normalization factor, $P\left(s_n = i\middle|\mathbf{Z}^n\right)$ can now be computed by

$$P\left(s_n = i\middle|\mathbf{Z}^n\right) = (const)\prod_{t=1}^{N} P\left(z_n(t)\middle|s_n = i, \mathbf{Z}^{n-1}\right) \cdot \sum_{j=1}^{L} a_{ij}P\left(s_{n-1} = j\middle|\mathbf{Z}^{n-1}\right) \quad (24)$$

With initial condition $P\left(s_0 = i\middle|\mathbf{Z}^0\right) = \frac{1}{L}$, $\hat{\mathbf{y}}_{0,i}(0) = \mathbf{0}$ and $\mathbf{P}_i(0) = 10 \cdot \mathbf{I}$ at the first speech segment, for $i=1,\ldots,L$, filtering is performed in the following order. The first is the mixing step denoted by (18)-(20) and the next Kalman filtering step is processed by (8)-(11). Then the probability calculation follows from (24) and finally the output is generated in (5). Note that the mixing which is represented by (18), (20) and is the key of the IMM algorithm can not be found in the previous algorithm [7].

For the better estimation of the speech, we delayed the computation of $\hat{y}_n(t)$ until the $(t+p-1)$ the instant. The speech sample estimate at time instant t is finally

obtained by

$$\hat{y}_n(t) = \sum_{j=1}^{L} H\hat{\mathbf{y}}_{n,i}(t) P\left(s_n = i \middle| \mathbf{Z}^n\right) \quad (25)$$

Hence, all Kalman filters are tried and each estimate is assigned a probability of being the best signal estimate. Since the MMSE signal estimate is constructed as the average of the individual estimate weighted by their probabilities, this estimator is soft decision estimation approach.

IV. Experimental Results

We discuss the performance of the proposed speech enhancement. This method was tested using Gaussian white noise at input signal-to-noise ratio (SNR) greater than or equal to 5 dB. The SNR is defined as the ratio between the average power of the signal and the average of the noise.

The nonstationary AR HMM with $M=1$ was estimated from a training data set which consisted of 7-min of conventional speech from four speakers, two male and two females. The raw speech data was in the form of digitally sampled signal at 12kHz. A Hamming window of duration 25.6 msec was applied every 20 msec within each window. 12-order AR coefficients were computed. We choose to evaluate the proposed method with the nonstationary AR HMM based separately multiple model. The method was tested on speech signals different from those used for training, and the speakers of the training and test speech material were not the same. The test data consisted of two sentences originally spoken by a male and a female.

Table 1 compares the performance of the proposed method with that of the conventional method.

V. Conclusions

In this study, we proposed, implemented and evaluated a speech enhancement based on nonstationary AR HMM and IMM. The principal motivation of this method is to parametrically describe continuously-varying transitional acoustic patterns of speech in a more natural and a more structural manner than the conventional method developed and widely used in the past. In this approach the estimator of speech is the weighted sum of the parallel Kalman filters. These filters are operating interactively instead of separately. The enhanced performance is improved by considering the interactions between the parallel filters.

Reference

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

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

[3] Y. Ephraim, "A bayesian estimation approach for speech enhancement using hidden markov models," *IEEE Trans. Signal Processing*, vol.40, pp. 725-735, Apr. 1992.

[4] K.Y. Lee and K. Shirai, "Recursive estimation using the mixture hidden filter model for enhanceing noisy speech," *Proc. IASTED Int'l Conf. Signal and Processing(SIP-95)*, pp. 43-46, 1995.

[5] K.Y. Lee and K. Shirai, "Efficient recursive estimation for speech enhancement in colored noise," *IEEE Signal Processing Letters*, vol. 3, no. 7, pp.196-199, 1996. [5] K.Y. Lee and K. Shirai, "Efficient recursive estimation for speech enhancement in colored noise," *IEEE Signal Processing Letters*, vol. 3, no. 7, pp.196-199, 1996.

[6] L. Deng, "A generalized hidden markov model with state-conditioned trend functions of time for the speech signal," *EURASIP Signal Processing*, 27, pp. 65-78, 1992

[7] K.Y.Lee, J.Y.Rheem, and K.Shirai, "Recursive estimation based on the trend HMM in speech enhancement," *Proc. IEEE-APCCAS*, Nov. 1996.

[8] H.A.P.Blom and Y.Bar-Shalom, "The interacting multiple model algorithm for systems with Markovian switching coefficients," *IEEE Automatic Control*, vol.33, pp.780-783, Aug. 1988

[9] A.P. DEmpster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," J. Roy. Statist. Soc., Vol. 39, 1977. pp. 1-38

Table 1. Output SNR Performance

input SNR	HMM	Proposed method	
		SMM	IMM
5	10.1	10.5	11
10	14.2	14.7	15
15	18.3	18.7	19
20	22.2	22.5	22.8