Recursive Estimation using the Hidden Filter Model for Enhancing Noisy Speech

*Yeong Tae Kang and *Ki Yong Lee

Abstract

A recursive estimation for the enhancement of white noise contaminated speech is proposed. This method is based on the Kalman filter with time-varying parametric model for the clean speech signal. Then, hidden filter model are used to model the clean speech signal. An approximation improvement of 4-5 dB in SNR is achieved at 5 and 10 dB input SNR, respectively.

Ⅰ. Introduction

Hidden markov models (HMM) have been used to estimate the model of speech signals for speech enhancement [1-2]. In HMM speech signals are first blocked into fixed-length frames. However, since the vocal tract and excitation can change substantially in a relatively short time, the framing often results in a poor temporal resolution for fast varying speech sounds such as plosives or diptongs. To represent the nonstationary nature of speech waveform, we assume that speech is the output of a time-varying system whose parameters depend on the states of the Markov chain. Hidden filter model (HFM) is an useful model with parametric form for the nonstationarity of speech waveform [3]. HFM has been proven successful in speech recognition application [4].

In this paper, we used the maximum likelihood approach using the Baum re-estimation algorithm to estimate parameter of HFM from the clean speech signal. Given the parameter of the speech, we propose a recursive estimation based on the Kalman filter for speech enhancement of white noise contaminated speech. This method does not need to frame the speech in training and enhancement procedure.

Ⅱ. The Hidden Filter Model for Speech Signal

The HFM is an autoregressive (AR) model with its parameters associated with markov chain states. Consider a first-order markov chain with L-states and a states transition matrix $\mathbf{A}=[a_{ij}]$ $i, j=1,\ldots,L$. Thus, at time t, the speech data conditioned on state i is described by

$$y(t)=\mathbf{B}_i^T\mathbf{Y}(t-1)+e_i(t), \tag{1}$$

where $\mathbf{B}_i^T=[b_i(1),\ldots,b_i(p)]$ is the vector of AR coefficients on state i, $\mathbf{Y}(t-1)=[y(t-1)\ldots y(t-p)]^T$ is the sequence of the past p observations, and the driving sequence $e_i(t)$ is zero mean Gaussian process with a variance σ_i^2.

Starting from an initial model λ_0, the objective function with M multiple training sequences is given by the Baum re-estimation algorithm [4], as

$$Q(\lambda,\lambda_0)=\sum_{m=1}^{M}\sum_{i,j=1}^{L}\sum_{t=1}^{T^m}\gamma_{ij}^m(t)\left[\ln a_{ij}+\ln\frac{1}{\sqrt{2\pi}\,\sigma_i}-\frac{(y^m(t)-\mathbf{B}_i^T\mathbf{Y}^m(t-1))^2}{2\sigma_i^2}\right] \tag{2}$$

where $\gamma_{ij}(t)$ is the a posterior probability of the transition from state i to state j given the observation sequence and the model λ_0.

We can obtain the model parameter $\lambda=\{\mathbf{A}=\{a_{ij}\},\ \mathbf{B}=\{\mathbf{B}_j\},\ \sigma=\{\sigma_i^2\},\ i, j=1,\ldots,L\}$ by maximizing the objective function (2) as described in [5]:

$$a_{ij}=\frac{\sum_{m=1}^{M}\sum_{t=1}^{T^m}\gamma_{ij}^m(t)}{\sum_{i=1}^{L}\sum_{m=1}^{M}\sum_{t=1}^{T^m}\gamma_{ij}^m(t)},$$

$$\mathbf{B}_j=\Big[\sum_{i=1}^{L}\sum_{m=1}^{M}\sum_{t=1}^{T^m}\gamma_{ij}^m(t)\mathbf{Y}^m(t-1)\mathbf{Y}^{m^T}(t-1)\Big]^{-1}\times\Big[\sum_{i=1}^{L}\sum_{m=1}^{M}\sum_{t=1}^{T^m}\gamma_{ij}^m(t)\,y^m(t)\mathbf{Y}^m(t-1)\Big],$$

$$\sigma_j^2=\frac{\sum_{i=1}^{L}\sum_{m=1}^{M}\sum_{t=1}^{T^m}\gamma_{ij}^m(t)(y(t)-\mathbf{B}_i^T\mathbf{Y}^m(t-1))^2}{\sum_{i=1}^{N}\sum_{t=1}^{T'}\gamma_{ij}^m(t)}$$

*Department of Electronics Engineering, Changwon National University

Manuscript Received June 17, 1996.

Ⅲ. Recursive Estimation for Speech Enhancement

If the speech is degraded by statistically independent additive zero mean white Gaussian noise, we can construct a state-space form with markov states $s(t) \in \{1, \ldots, L\}$ at time t, as;

$$\mathbf{Y}(t) = \Phi(s(t))\mathbf{Y}(t-1) + Ge(s(t)), \tag{3}$$

$$z(t) = H^T \mathbf{Y}(t) + w(t) \tag{4}$$

where $\Phi(s(t)) = \begin{bmatrix} \mathbf{B}^T_{s(t)} \\ \mathbf{I} \quad \mathbf{0} \end{bmatrix}$, $G = [1 0 \cdots 0]^T$, $H = [1 0 \cdots 0]^T$, and $w(t)$ is white Gaussian noise with variance σ_w^2, and assume that $\mathbf{Y}(0)$, $e(s(t))$, and $w(t)$ are mutually independent.

Given the noisy speech $\mathbf{Z}(t) = \{z(1) \cdots z(t)\}$, the estimate $\hat{\mathbf{Y}}(t)$ of clean speech $y(t)$ is given by the conditional mean

$$\hat{\mathbf{Y}}(t) = E\{\mathbf{Y}(t)|\mathbf{Z}(t)\} = \int_{-\infty}^{\infty} \mathbf{Y}(t)\, p(\mathbf{Y}(t)|\mathbf{Z}(t))\, d\mathbf{Y}(t). \tag{5}$$

By definition [6], the conditional density function of (5) can be written as

$$p(\mathbf{Y}(t)|\mathbf{Z}(t)) = \sum_{j=1}^{L} p(\mathbf{Y}(t)|s(t)=j, \mathbf{Z}(t))\, p(s(t)=j|\mathbf{Z}(t)). \tag{6}$$

Substituting (6) into (5), and interchanging integration and summation, the estimate $\hat{\mathbf{Y}}(t)$ is obtained by

$$\hat{\mathbf{Y}}(t) = \sum_{j=1}^{L} \hat{\mathbf{Y}}_j(t)\, p(s(t)=j|\mathbf{Z}(t)) \tag{7}$$

where $\hat{\mathbf{Y}}_j(t) = \int_{-\infty}^{\infty} \mathbf{Y}(t)\, p(\mathbf{Y}(t)|s(t)=j, \mathbf{Z}(t))\, d\mathbf{Y}(t)$ is the conditional mean estimate of $\mathbf{Y}(t)$ given $s(t)=j$.

The estimate $\hat{\mathbf{Y}}(t)$ of (7) is a weighted sum of the L individual estimates $\hat{\mathbf{Y}}_j(t)$. The weighting factor $p(s(t)=j|\mathbf{Z}(t))$ is the probability that the individual estimators are correct ones for the given noisy speech $\mathbf{z}(t)$.

Each estimate $\hat{\mathbf{Y}}_j(t)$ is found from a modified Kalman filter given by

$$\hat{\mathbf{Y}}_j(t) = \Phi(s(t)=j)\hat{\mathbf{Y}}_j(t-1) + K_j(t)\{z(t) - H^T \Phi(s(t)=j)\hat{\mathbf{Y}}_j(t-1)\}, \tag{8}$$

$$M_j(t) = \Phi(s(t)=j)\, P_j(t-1)\Phi^T(s(t)=j) + GQ(s(t)=j)G^T, \tag{9}$$

$$K_j(t) = M_j(t)H^T [HM_j(t)H^T + R]^{-1}, \tag{10}$$

$$P_j(t) = M_j(t) - K_j(t)HM_j(t). \tag{11}$$

where $Q(s(t)=j) = \sigma_j^2 \mathbf{I}$ and $\mathbf{R} = \sigma_w^2 \mathbf{I}$ are the covariance matrix of excitation $e(s(t))$and noise $w(t)$, respectively.

The weighting factor $p(s(t)=j|\mathbf{Z}(t))$ becomes as, using $\mathbf{Z}(t) = \{z(t), \mathbf{Z}(t-1)\}$ and Bayes rule;

$$p(s(t)=j|\mathbf{Z}(t)) = \frac{p(z(t)|s(t)=j, \mathbf{Z}(t-1))\, p(s(t)=j|\mathbf{Z}(t-1))}{p(z(t)|\mathbf{Z}(t-1))}. \tag{12}$$

This first term of the numerator of (12) can be approximated by

$$p(z(t)|s(t)=j, \mathbf{Z}(t-1)) \approx N[H^T \Phi(s(t)=j)\hat{\mathbf{Y}}_j(t-1),\ HM_j(t)H^T + R] \tag{13}$$

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

The second term of the numerator of (12) is the predicted probability given by,

$$p(s(t)=j|\mathbf{Z}(t-1)) = \sum_{i=1}^{L} a_{ij}\, p(s(t-1)=i|\mathbf{Z}(t-1)). \tag{14}$$

Since the denominator term of (12) is independent of j, it becomes a scale factor. Therefore, $p(s(t)=j|\mathbf{Z}(t))$ is can be efficiently calculated using the previous weighting factor as

$$p(s(t)=j|\mathbf{Z}(t)) = c_t\, p(z(t)|s(t)=j, \mathbf{Z}(t-1)) \sum_{i=1}^{L} a_{ij}\, p(s(t-1)=i|\mathbf{Z}(t-1)) \tag{15}$$

where c_t is a scale factor determined at time t and guaranteeing that the sum of all the weighting factor is equal to one;

$$\sum_{i=1}^{L} p(s(t)=j|\mathbf{Z}(t)) = 1.$$

With initial conditions $p(s(0)=j|\mathbf{Z}(0)) = 1/L$, $\hat{\mathbf{Y}}(0) = 0$, $P_j(0) = 0$, for $j = 1, \ldots, L$, (9) is processed first, followed by (10-11), (8), (15), and then (7). The enhanced speech signal $\hat{y}(t)$ is equal to the first component $\hat{y}(t)$ of the estimated $\hat{\mathbf{Y}}(t)$. However, the last component of the estimated $\hat{\mathbf{Y}}(t+p-1)$ at time $t+p-1$ will give a better estimate of the speech signal at the t-th instant. Thus we delay the computation of $\hat{\mathbf{Y}}(t)$ until the $(t+p-1)$ th instant. The enhanced speech signal at the t-th instant is finally obtained as

$$\hat{y}(t) = \frac{[1 0 \cdots 0]}{[0 \cdots 0 1]} \hat{\mathbf{Y}}(t)\,(t+p-1). \tag{16}$$

Ⅳ. Experimental Results

The proposed enhancement approach was examined in

Table 1. SNR(dB) Performance of the proposed method

Input SNR	0	5	10	15	20
without delay	5.94	9.09	13	17.18	21.14
general(HMM)	7.25	10.27	13.68	17.58	21.69
with delay	7.31	10.78	14.27	18.17	22.19

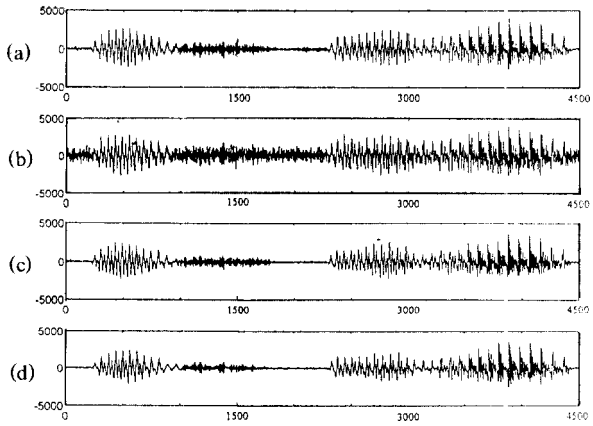

Figure 1. Speech signals : (a) clean speech, (b) noisy speech(5dB) (c) enhanced speech without delay, (d) enhanced speech with delay

enhancing speech signals which have been degraded by statistically independent additive white Gaussian noise at signal-to-noise ratio (SNR) values of 0, 5, 10, 15. 20 dB. Training was performed using 8 sentences of clean speech by two male speaker. Enhancement tests were performed on 2 sentences of test sentences spoken by a different one speaker. In experiment, speech is sampled at 12kHz, the order of each AR process is 12 and the number of states of HFM is 5, respectively. Table 1. shows performance of proposed method with a delay and without a delay. By the proposed method, an approximate improvement of 4 dB in SNR is achieved at various input SNR. Also, results for the method with delay show an additional 1.1 dB and 0.4 dB improvement over the method without delay and conventional HMM[1], respectively. In Fig. 1, we show a noisy speech signal processed by the proposed method under SNR value of 5 dB.

V. Conclusion

We proposed a new approach for enhancing the speech signal which have been degraded by statistically independent additive white Gaussian noise. Given the trained HFM from clean speech, the recursive estimation based on the Kalman filter is developed. This approach is comprised of a fixed set of estimators operating in parallel with each individual estimator weighted by the probabilities that its own are correct ones for the given noisy speech. The noise model and HFM used in here is simple model. Presently, we are studying about recursive estimation with the mixture HFM for speech enhancement under the colored noise.

References

1. 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.
2. 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.
3. H. Sheikhzadeh, H. and L. Deng, "Waveform-based speech recognition using hidden filter models : Parameter selection and sensitivity to power normalization," *IEEE Trans. Speech and Audio Processing*, vol. 2, pp. 80-89, Jan. 1994.
4. A. B. Poritz, "Hidden markov models : A guided tour," in *Proc. ICASSP*, pp. 7-13, 1988.
5. K. Y. Lee and K. Shirai, "Recursive estimation for speech enhancement using the hidden filter model," Proc. Acoust. Soc. Jpn Spring Meeting'95, pp. 63-64, 1995.
6. K. Y. Lee, *et al.*, "Robust recursive estimation for linear systems with non-Gaussian state and measurement noise," in *Proc. ICASSP*, pp. III-500-503, 1994.

▲Yeong Tae Kang

Yeong Tae Kang received the B. S., M.S. degrees in Electronics Engineering from Changwon National University, Kyeongnam, Korea, in 1990, 1996, respectively. His interests are statistical signal processing, speech enhancement and digital communication.

▲Ki Yong Lee

Ki Yong Lee received the B.S. degree in 1983 from SoongSil University, Seoul. Korea, and the M.S. and Ph.D. degrees in 1985 and 1991, respectively, from Seoul National University, all in Electronics Engineering. During 1994-1995 he was a KOSEF Postdoctoral Fellow at Waseda University, Tokyo, and at Signal Processing Group,

Edinburgh University, Edinburgh, Scotland. Also, he was a JSPS short-term Postdoctoral Fellow in 1996. Since 1991 Dr. Lee has been an Assistant Professor at the Department of Electronics Engineering at Changwon National University, Kyeongnam, Korea. His interests are in the application of estimation theory and statistics to problem solving in signal processing, speech enhancement and digital communication.