

Efficient Speech Enhancement Based on Left-Right HMM with State Sequence Detection using LRT

¹J.J. Lee, ²J.H. Lee, ³K.Y. Lee

^{1,3}School of Electronic Engineering, SoongSil University

¹jjlee@ctsp.ssu.ac.kr, ³kylee@ssu.ac.kr

²Dept. of Internet Broadcasting, Dong-Ah Broad., Coll., Korea

²jes@sci.voice.edu

Abstract

Since the conventional HMM (Hidden Markov Model)-based speech enhancement methods try to improve speech quality by considering all states for the state transition, hence introduce huge computational loads inappropriate to real-time implementation. In the Left-Right HMM (LR-HMM), only the current and the next states are considered for a possible state transition so to reduce the computation complexity. We propose a new speech enhancement algorithm based on LR-HMM with state sequence detection using LRT (Likelihood Ratio Test). Experimental results show that the proposed method improves the speed up with little degradation of speech quality compared to the conventional method.

1. 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 noises. Speech enhancement based on HMM proposed by Ephraim [1,2] and Lee [3,4] has been a good method for this purpose. This method first estimates the HMM parameters from the training signals and, second, filters the corrupted signal by fixed numbers of the Kalman filters or the Wiener filters to get the estimated signal which can be expressed as a weighted sum of their output.

Since it considers that all of the transitions to all states from the current state are possible, however, the conventional HMM needs enormous computation complexity. Therefore, HMM with L states and M mixtures require $L \times M$ Kalman filters or Wiener filters to enhance the noisy speech. This vast computation loads obstruct the real-time implementation of HMM-based speech enhancement in spite of its considerably good performance.

In this paper, we propose a new speech enhancement algorithm based on LR-HMM with state sequence detection using LRT[5] to reduce much the computation complexity. Following the definition of LR-HMM, a new algorithm considers only the current state and next right-side state as a possible state for the state transition of the next frame. In the proposed method, we represent LRT by the probability ratio of two states current state and next right-side state, and then detect the state of current noisy speech with LRT. If we detect the state of current noisy speech, HMM with L states and M mixtures requires $2 \times M$ Kalman filters or Wiener filters to enhance the noisy speech. The proposed method reduces the computational loads.

The proposed method may cause a little but negligible SNR degradation of 0.12-0.5dB but greatly save the computation

time compared with the conventional HMM-based speech enhancement.

This paper is organized as follows. We describe the HMM-based speech enhancement in section II. This is followed by a description of the speech enhancement based on LR-HMM and the related state sequence detection algorithm in section III. The experimental results are shown in section IV and we then provide our conclusions in section V.

2. Speech enhancement based on HMM

2.1. Speech enhancement based on HMM and basic layout features

To represent the statistics of clean speech signal y , we consider an HMM with L states and M mixtures. Let $y = \{y(t), t = 1, 2, \dots, T\}$, $y(t) = \{y((t-1)N+1), \dots, y(tN)\}$ be the clean speech sequence. For HMM, let $s = \{s_t, t = 1, 2, \dots, T\}$, $s_t \in \{1, 2, \dots, L\}$ be a sequence of states corresponding to y , and $h = \{h_t, t = 1, 2, \dots, T\}$, $h_t \in \{1, 2, \dots, M\}$ be a sequence of mixture components corresponding to (s, y) , where t is the index for frame of time, T is total number of frames, and N is the frame length. Thus, at frame t , speech conditioned in h_t mixture on state s_t is expressed by a linear combination of its past values plus an excitation source, as

$$y(n) = B_{h_t|s_t}^T Y(n-1) + e_{h_t|s_t}(n), \quad (t-1)N+1 < n < tN \quad (1)$$

where, $B_{h_t|s_t}^T = [b_{h_t|s_t}(1), b_{h_t|s_t}(2), \dots, b_{h_t|s_t}(p)]^T$ is AR parameter vector under the state s_t . $Y(n-1) = [y(n-1), \dots, y(n-p)]^T$ is the sequence of past p observations, and excitation source $e_{h_t|s_t}(n)$ is Gaussian i.i.d. process with zero mean and variance $\sigma_{e_{h_t|s_t}}^2$.

When, the background noise $v(n)$ is assumed as white Gaussian process with zero mean and variance σ_v^2 , noisy speech sequence $z(t)$ can be represented by

$$z(t) = y(t) + v(t), \quad t = 1, 2, \dots, T \quad (2)$$

where $z(t) = \{z((t-1)N+1), \dots, z(tN)\}$ and $v(t) = \{v((t-1)N+1), \dots, v(tN)\}$.

The state equations for Kalman filter are as

$$Y(n) = F(s_t, h_t)Y(n-1) + G e_{s_t|h_t}(n) \quad (3)$$

$$z(n) = H^T Y(n) + v(n) \quad (4)$$

where

$$Y(n) = \begin{bmatrix} y(n) \\ y(n-1) \\ \vdots \\ y(n-(p-1)) \end{bmatrix}$$

$$F(s_t, h_t) = \begin{bmatrix} b_{h_t|s_t}(1) & b_{h_t|s_t}(2) & \cdots & b_{h_t|s_t}(p-1) & b_{h_t|s_t}(p) \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}$$

$$H = G = [1 \ 0 \ \cdots \ 0]^T$$

Given noisy speech, the estimation of $Y(n)$ is considered as a conditional mean as

$$\hat{Y}(n) = \{E\{Y(n) | z(t)\}\} \quad (5)$$

$$= \int_{-\infty}^{\infty} Y(n) p(Y(n) | z(t)) dY(n).$$

In (5), conditional distribution function is given as

$$p(Y(n) | Z(t))$$

$$= \sum_{j=1}^L \sum_{m=1}^M p(Y(n) | s_t = j, h_t = m, z(t)) p(s_t = j, h_t = m | z(t)) \quad (6)$$

By substituting (6) into (5), we can derive the estimation of $\hat{Y}(n)$ as

$$\hat{Y}(n) = \sum_{j=1}^L \sum_{m=1}^M \hat{Y}_{h_t|s_t}^{j,m}(n) p(s_t = j, h_t = m | z(t)) \quad (7)$$

In the above equation, $\hat{Y}_{h_t|s_t}^{j,m}(n)$ is conditional mean estimation given $s_t = j$ and $h_t = m$. $\hat{Y}_{h_t|s_t}^{j,m}(n)$ is obtained by Kalman filter and $p(s_t = j, h_t = m | z(t))$ is obtained by Bayesian theorem and property of probability. Since $L \times M$ numbers of Kalman filters are considered to get $\hat{Y}(n)$, the computation loads becomes so huge not to implement the real-time system.

3. Speech enhancement based on Left-Right HMM and state sequence detection algorithm

3.1. Speech enhancement based on Left-Right HMM

If we have the known current state s_t^* , (7) is rewritten by LR-HMM as

$$\hat{Y}(n) = \sum_{s_t^* = j}^{j+1} \sum_{m=1}^M \hat{Y}_{h_t|s_t^*}^{j,m} p(s_t^* = j, h_t = m | z(t)) \quad (8)$$

Then, since $2M$ Kalman filters are considered to get $\hat{Y}(n)$, the computation loads can be decreased. To get $\hat{Y}(n)$, $\hat{Y}_{h_t|s_t^*}^{j,m}$ and the weight $p(s_t^* = j, h_t = m | z(t))$ are necessary to be calculated. Given the known s_t^* is known, to calculate, the Kalman algorithm for $\hat{Y}_{h_t|s_t^*}^{j,m}$ can be expressed as

$$\hat{Y}_{m|s_t^*}^{j,m}(n) = F(s_t^* = j, h_t = m) \hat{Y}_{m|s_t^*}^{j,m}(n-1) + K_{m|s_t^*}^{j,m}(n) \cdot \{z(n) - H^T F(s_t^* = j, h_t = m) \hat{Y}_{m|s_t^*}^{j,m}(n-1)\} \quad (9)$$

$$M_{m|s_t^*}^{j,m}(n) = F(s_t^* = j, h_t = m) P_{m|s_t^*}^{j,m}(n-1) F^T(s_t^* = j, h_t = m) + G Q G^T \quad (10)$$

$$K_{m|s_t^*}^{j,m}(n) = M_{m|s_t^*}^{j,m}(n) H^T [V + H M_{m|s_t^*}^{j,m}(n) H^T]^{-1} \quad (11)$$

$$P_{m|s_t^*}^{j,m}(n) = M_{m|s_t^*}^{j,m}(n) - K_{m|s_t^*}^{j,m}(n) H M_{m|s_t^*}^{j,m}(n) \quad (12)$$

The weight $p(s_t = j, h_t = m | z(t))$ must be obtained beforehand by using Bayesian theorem as

$$p(s_t = j, h_t = m | z(t))$$

$$= \frac{p(z(t) | s_t = j, h_t = m, z(t-1)) p(s_t = j, h_t = m | z(t-1))}{p(z(t) | z(t-1))} \quad (13)$$

The first term of the numerator can be approximated as follows;

$$p(z(t) | s_t = j, h_t = m, z(t-1))$$

$$= \prod_{n=1}^N p(z(n) | s_t = j, h_t = m) \quad (14)$$

Also, the right side of (14) can be expressed by normal distribution $N[\cdot]$ as

$$p(z(n) | s_t = j, h_t = m) = N[\hat{Y}_{m|j}(n), H P_{m|j} H^T] \quad (15)$$

In (13), $p(s_t = j, h_t = m | z(t-1))$ can be represented by the given Markov process

$$p(s_t = j, h_t = m | z(t-1))$$

$$= \sum_{i=1}^M p(s_t = j, h_t = m | s_{t-1} = i, h_{t-1} = l, z(t-1)) \times p(s_{t-1} = i, h_{t-1} = l, z(t-1)) \quad (16)$$

The first term of (16) can be rewritten as

$$p(s_t = j, h_t = m | s_{t-1} = i, h_{t-1} = l, z(t-1))$$

$$= p(h_t = m | s_t = j, s_{t-1} = i, h_{t-1} = l, z(t-1)) \times p(s_{t-1} = i | s_{t-1}, h_{t-1} = l, z(t-1)) \quad (17)$$

Since h_t and s_t are statistically independent from each other, the second term can be rewritten as

$$p(h_t = m | s_t = j, s_{t-1} = i, h_{t-1} = l, z(t-1)) = c_{m|j} \quad (18)$$

$$p(s_{t-1} = i | s_{t-1}, h_{t-1} = l, z(t-1)) = p(s_{t-1} = j | s_{t-1} = i) = a_{ij} \quad (19)$$

Equation (16) can be rewritten by substituting (18) and (19) into it as

$$p(s_t = j, h_t = m | z(t-1))$$

$$= \sum_{i=1}^M a_{ij} c_{m|j} p(s_{t-1} = i, h_{t-1} = l | z(t-1)) \quad (20)$$

In (13), since the denominator is independent from j , it becomes a scale factor. Therefore, $p(s_t = j, h_t = m | z(t))$ can be calculated by using the previous probabilities as

$$p(s_t = j, h_t = m | z(t))$$

$$= K_t N_{m|j} \sum_{i=1}^M a_{ij} c_{m|j} p(s_{t-1} = i, h_{t-1} = l | z(t-1)) \quad (21)$$

In (21), K_t is a scale factor the summation of which is 1 as

$$\sum_{j=i}^{i+1} \sum_{m=1}^M p(s_t = j, h_t = m | z(t)) = 1$$

3.2. Detection of state sequence s_t based on the LRT

In the speech enhancement using LR-HMM, we assume that the current state s_t^* is known. However, we cannot practically know the state. Therefore, we consider the state sequence s_t detection algorithm based on the LRT.

In the LR-HMM with $s_{t-1} = j$, s_t is j or $j+1$. To detect s_t , the probabilities of those two states j and $j+1$ is compared. By

taking the logarithm for the probability ratio of two states j and $j+1$, likelihood ratio $\Lambda(t)$ can be defined by

$$\Lambda(t) \equiv \log \left(\frac{p(s_t = j | z(t))}{p(s_t = j+1 | z(t))} \right) \begin{matrix} \geq \\ < \end{matrix} 0 \quad (22)$$

$$s_t = j$$

$$s_t = j+1$$

In the above equation, $\Lambda(t) \geq 0$ means that the state of the current frame is more likely to be j than $j+1$ and, therefore, $s_t = j$. On the contrary, when $\Lambda(t) < 0$, $s_t = j+1$ is statistically true. In (22), $p(s_t | z(t))$ is the probability and is obtained from following equation.

$$p(s_t = j | z(t)) = \sum_{m=1}^M p(s_t = j, h_t = m | z(t)) \quad (23)$$

Equation (23) can be easily obtained from (21).

4. Experimental results

The proposed method is examined in enhancing speech signals degraded by statistically independent additive stationary white Gaussian noise at input signal-to-noise ratios (SNR) with 0, 5, 10 and 20dB. The input SNR is defined as the ratio of the average power of the signal to the average power of the noise. Training for the HMM of clean speech was performed using 50 sentences from 2-male and 3-female. Speech was sampled at 11, 025Hz. The order of each AR model is 15. In the enhancement test, neither the speaker nor the speech material used for testing was in the training set. Test data consisted of "An-nyeong-ha-se-yo (Good morning)" by a male and female. The speech sequence for enhancement was recorded in a manner similar to that for training.

The objective distortion measure adopted is the output_SNR defined by

$$\text{output_SNR} = 10 \log_{10} \frac{\frac{1}{T} \sum_{t=1}^T y^2(t)}{\frac{1}{T} \sum_{t=1}^T [y(t) - \hat{y}(t)]^2}$$

where T is the total number of frame in the utterance. We compared the performance of the proposed method and conventional method. Table 1 shows the comparison results with various SNR between the conventional method and the proposed one. Generally, for the enhanced performance, the proposed method reduces the computational loads with little degradation of enhanced performance of 0.12-0.5dB.

Figure 1 shows the speech enhancement performance under 10dB of SNR. (a) and (b) are a clean speech and its corrupted speech, respectively. (c) and (d) are the enhanced results by the conventional method and the proposed method, respectively. Listening test did not tell noticeable difference.

5. Conclusion

In this paper, we proposed a speech enhancement algorithm based on LR-HMM with state sequence detection using LRT. This method can reduce the computational complexity with

little degradation of speech quality compared with the previous method based on the ergodic HMM model.

The speech enhancement based on LR-HMM with state detection using LRT can be widely used for speaker verification and speech recognition application when it is modified to enhance the speech quality of each speaker in the text-dependent speaker verification.

Table 1: the competition between conventional method and proposed method(state: 7, mixture: 5)

Input SNR[dB]	Conventional method		Left-Right model with detection	
	Output SNR[dB]	No. of Kalman Filter	Output SNR[dB]	No. of Kalman Filter
0	8.5	$L \times M$	7.99	$2 \times M$
5	11.35		10.86	
10	15		14.81	
20	24.2		24.08	

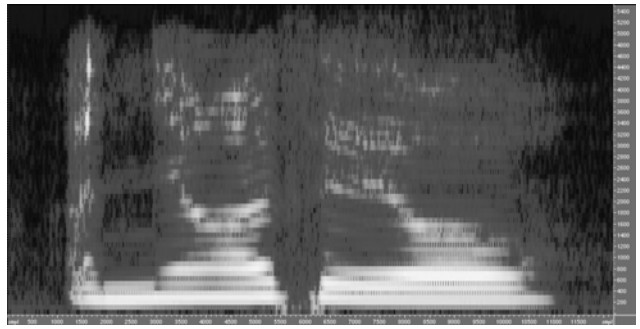
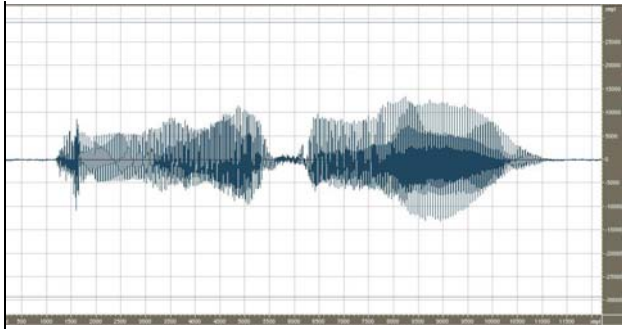
ACKNOWLEDGEMENT

This research was supported in part by University IT Research Center Project

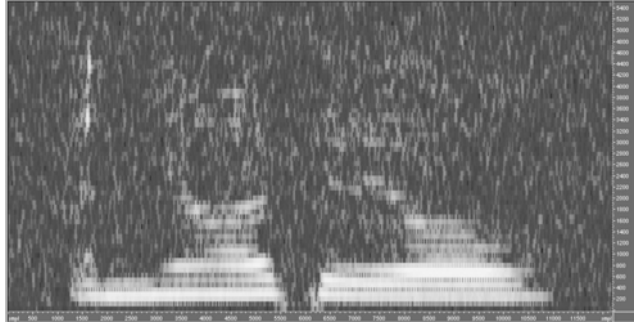
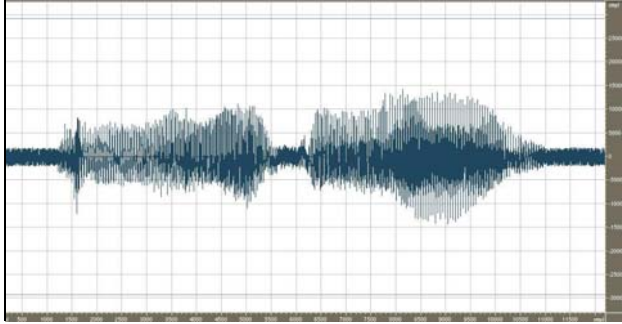
This work was supported (in part) by Biometrics Engineering Research Center, (KOSEF)

6. References

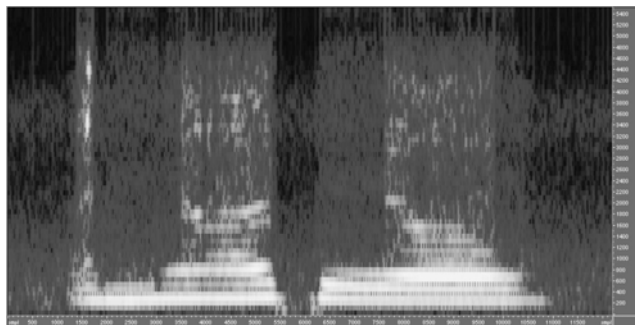
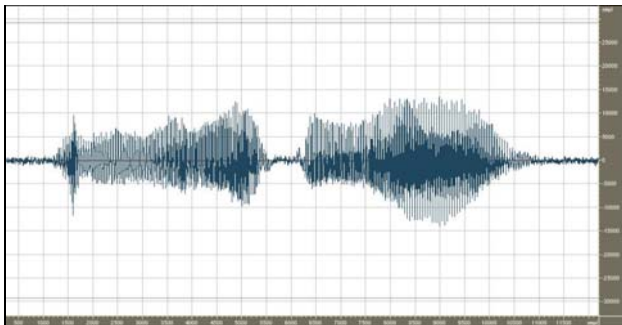
- [1] Y. Ephraim, D. Malah, B.-H. Juang, " On the application of hidden Markov models for enhancing noisy speech," *IEEE Trans. Acoustic. Speech Process.* Vol.37, no.12, pp.1846-1856, Dec., 1989
- [2] Y. Ephraim, " A Bayesian approach for speech enhancement using hidden Markov models," *IEEE Trans. Signal Processing*, vol. 41, pp.725-735, Apr., 1992
- [3] Ki Yong Lee, JaeYeal Rheem, " Smoothing approach using forward-backward Kalman filter with Markov Switching Parameters for Speech Enhancement," *Signal Processing*, vol. 80, pp.2579-2588, Apr., 2000
- [4] Souhwan Jung and Ki Yong Lee, "Time-Domain Approach Using Multiple Kalman Filters and EM Algorithm to Speech Enhancement with Nonstationary Noise," *IEEE Trans. Speech and Audio Processing*, vol. 8, no.3 pp.282-291, May, 2000
- [5] Porat, B. Friedlander, B., "On the generalize likelihood ratio test for a class of nonlinear detection problems," *IEEE Trans. Signal Processing*, vol. 41, no. 11 pp.569-578, Nov., 1993
- [6] Qi Li, "A detection Approach to Search-Space Reduction for HMM State Alignment in speaker Verification," *IEEE Trans. Speech and Audio Processing*, vol. 9, no.5 pp.569-578, July, 2001
- [7] Todd K. Moon, Wynn C. Stirling, *Mathematical Methods and Algorithms for signal processing*, Prentice hall, 2000



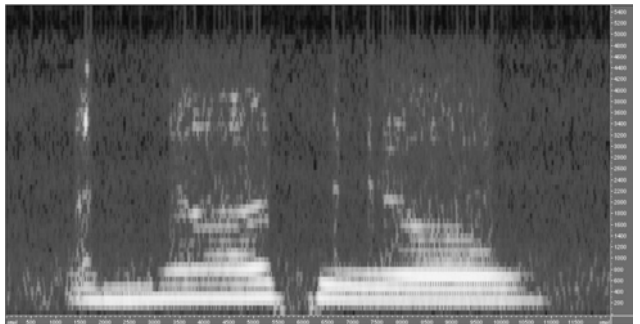
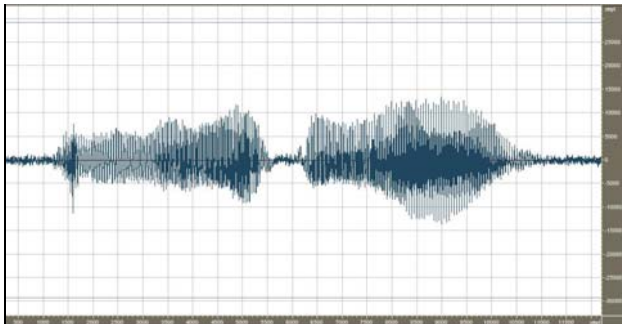
(a) Clean Speech



(b) 10dB Noisy Speech



(c) Enhanced Speech(Conventional Method)



(d) Enhanced Speech(Proposed Method)

Figure 1: Speech Wave and Spectrum