L-R HMM 갖는 문장 종속 음성 향상 방법

이종주, 이기용
숭실대학교 정보통신전자공학부

Text Dependent Speech Enhancement based L-R HMM

J.J. Lee, K.Y. Lee
School of Electronic Engineering, Soong Sil University
jjlee@ctsp.ssu.ac.kr , kylee@ssu.ac.kr

요 약

본 논문에서는 Left-Right HMM 모델에 기초를 둔 음질 향상 방법을 제안하였다. 기존 HMM에 기초를 둔 음질 향상 방법은 ergodic HMM에 기초를 두고 음질을 향상시켰다. 본 논문에서는 Left-Right HMM이 현재 상태에서 다음상태로만 변하는 성질을 이용하여 현재의 상태를 결정하여 다음 프레임에서 현재와 다음 상태에서만 계산을 하는 방법을 사용하였다. 그 결과 기존의 방법에 비해 많은 시간을 줄일 수 있었다.

I. 서론

음성 향상이란, 입력 신호가 배경잡음에 의해 오염되었을 때 음성 통신 시스템에서 성능을 향상시키고, 잡음의 영향을 최소화 하는 것이다. 오직 잡음 섞인 음성만을 알고 있을 때, 음성 향상을 위해서는 깨끗한 음성과 잡음 신호의 joint 통계성을 알고 있어야 한다. 음성 향상 법에 있어서, 위너필터와 칼만 필터에 관한 수많은 연구들이 있었다. 최근, Ephraim은 Hidden Markov Model(HMM)을 사용하는 음성 향상 방법을 제안하였다 [1, 2]. 그러나, Ergodic HMM에 기초를 둔 음질 향상 방법은 모든 상태에 대해 계산을 하므로, 계산량이 많아 실시간 처리에 부적절하다. 예를 들어 상태가 L이고, mixture가 M인 경우 LxM개의 칼만 필터나 위너 필터가 필요하다. 그러나, Left-Right HMM에서는 현재 상태에서 현재 상태나 바로 다음 상태로만 천이될 수 있으므로, 현재 상태를 아는 경우 모든 상태에 걸쳐서 계산할 필요가 없이, 현재 상태와 다음 상태에서만 계산을 수행한다.

본 논문에서는 Left-Right HMM에서 현재의 상태를 결정하여, 다음 프레임에서 현재 상태와 다음 상태에서만 계산하여 계산량을 줄이는 음질 향상 알고리즘을 제안하였다. 이 방법을 사용하면 기존의 방법에 비해 SNR은 약 0.2~0.4dB 정도 떨어지나, 시간을 66%정도 절약할 수 있었다. 본 논문은 II장은 HMM에 기초한 음성 향상, III장은 Left-Right HMM에 기초한 음성 향상과 상태 결정 알고리즘, IV장은 실험 결과, V장은 결론으로 구성되어있다.

II. HMM에 기초한 음성 향상

깨끗한 음성 신호 $y(n)$를 각각의 상태에서 L개의 상태와 M mixture의 가우시안 AR모델로 표현하기 위해, y에 대응하는 상태열을 $s=\{s_t, t=1,2,.....,T\}, s_t \in \{1,2...,L\}$라 두고, (s,y)에 대응하는 mixture 열을 $h=\{h_t, t=1,2,.....,T\}, h_t \in \{1,2...,M\}$라고 하면 깨끗한 음성 모델을 아래식으로 표현된다.

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

이다. 이 때, $\mathbf{B}_{h_t|s_t}^T=[b_{h_t|s_t}(1)...,b_{h_t|s_t}(p)]^T$는 s_t 상태의 AR 계수들이다. $Y(n-1)=[y(n-1)..y(n-p)]^T$는 과거 p개의 관측열이고, $e_{h_t|s_t}(n)$는 평균이 0이고, 분산이 $\sigma_{h_t|s_t}^2$인 가우시안 모델이다.

배경 잡음 $v(n)$은 평균이 0이고, 분산이 σ_v^2인 백색 가우시안 모델이고, $e_{h_t|s_t}(n)$과 $v(n)$은 상관관계가 없다고 가정한다.

잡음이 음성에 통계적으로 독립적으로 부가된 음성은 아래의 식으로 표현된다.

$$z(t)=y(t)+v(t) \quad , \quad t=1,2,.....,T \tag{2}$$

$y(t)=\{y(n),(t-1)N+1\le n\le tN\}$는 깨끗한 음성 벡터이고 $v(t)=\{v(n),(t-1)N+1\le n\le tN\}$는 부가된 배경 잡음 벡터이다.

칼만 필터는 아래 식과 같이 음성과 백색 잡음의 상태 공간 모델을 확장시키는데 기초를 둔다.

$$Y(n)=F(s_t,h_t)Y(n-1)+Ge_{s_t|h_t}(n) \tag{3}$$

$$z(n) = H^T Y(n) + v(n) \tag{4}$$

위 식에서, 각 요소들은 아래와 같다.

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

$$F(s_t, h_t) = \left[B^T_{h_t|s_t}\right] \quad , \quad H = G = [1\ 0\ \ldots\ 0]^T$$

$e_{h_t|s_t}(n)$ 과 $w(n)$ 는 상관 관계가 없다(uncorrelated)고 가정 했으므로, 분산 행렬을 아래와 같이 표현할 수 있다.

$$E\left[e_{s_t|h_t}(n) e^T_{s_t|h_t}(n)\right] = Q(s_t = j, h_t = m) = \left[\sigma^2_{s_t|h_t}\right]$$

$$E\left[v(n) v^T(n)\right] = V = \sigma_v$$

$$E\left[e_{s_t|h_t}(n) v^T(n)\right] = 0$$

잡음 음성이 주어져있을 때, $\hat{Y}(n)$ 을 추정하는 것은 조건 평균으로 주어진다.

$$\begin{aligned} \hat{Y}(n) &= \{E\{Y(n) \mid z(t)\} \\ &= \int_{-\infty}^{\infty} Y(n)\, p(Y(n) \mid z(t))\, dY(n) \end{aligned} \tag{5}$$

위 (5)식의 조건 분포 함수를 아래식과 같이 쓸 수 있다.

$$\begin{aligned} & p(Y(n) \mid Z(t)) \\ &= \sum_{j=1}^{L} \sum_{m=1}^{M} p(Y(n) \mid s_t = j, h_t = m, z(t))\, p(s_t = j, h_t = m \mid z(t)) \end{aligned} \tag{6}$$

(6)을 (5)식에 대입하고 적분과 합계를 바꾸면 , $\hat{Y}(n)$ 추정은 아래식으로 구할 수 있다.

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

위 식에서 $\hat{Y}_{h_t|s_t}(n) = \int_{-\infty}^{\infty} Y(n) p(Y(n) \mid s_t = j, h_t = m, z(t))\, dY(n)$ 는 $s_t = j$, $h_t = m$,일 때 $Y(n)$ 의 조건 평균 추정식이며, (7)식에서 알 수 있듯이 $\hat{Y}(n)$ 추정은 각각 L 상태 추정 $\hat{Y}_{h_t|s_t}(n)$의 합으로 구한다.

각각의 $\hat{Y}_{h_t|s_t}(n)$은 multiple kalman filter로 구할 수 있는데, 이 필터가 L 상태와 M mixture을 가지고 있는 경우, L×M개의 칼만 필터가 필요하므로, 많은 계산량이 요구된다.

III. Left-Right HMM에 기초한 음성 향상과 상태(s_t^*)결정 알고리즘

a. Left-Right HMM에 기초한 음성 향상 방법

그러나, 상태열 $s_1, s_2, \ldots\ldots, s_T$ 을 알고 있다고 가정한다면, (7)식은 아래와 같이 다시 쓸 수 있다.

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

위 식에서 s_t^* 는 알고 있는 상수 값으로 결정하는 방법은 뒤에 설명하였다. $\hat{Y}(n)$을 구하기 위해서는 $\hat{Y}_{h_t|s_t}(n)$과 가중치 $p(s_t^*, h_t = m \mid z(t))$를 계산해야 한다. s_t^*를 알고 있을 때, $\hat{Y}_{h_t|s_t}(n)$을 계산하기 위한 칼만 필터 알고리즘은 다음과 같다.

$$\begin{aligned} \hat{Y}_{m|s_t^*}(n) = F(s_t^*, h_t = m) & \hat{Y}_{m|s_t^*}(n-1) \\ & + K_{m|s_t^*}(n) \left\{ z(n) - H^T F(s_t^*, h_t = m) \hat{Y}_{m|s_t^*}(n-1) \right\} \end{aligned} \tag{8}$$

$$\begin{aligned} M_{m|s_t^*}(n) = F(s_t^*, h_t = m) & P_{m|s_t^*}(n-1) F^T(s_t^*, h_t = m) \\ & + G Q(s_t^*, h_t = m) G^T \end{aligned} \tag{9}$$

$$K_{m|s_t^*}(n) = M_{m|s_t^*}(n) H^T \left[V + H M_{m|s_t^*}(n) H^T \right]^{-1} \tag{10}$$

$$P_{m|s_t^*}(n) = M_{m|s_t^*}(n) - K_{m|s_t^*}(n) H M_{m|s_t^*}(n) \tag{11}$$

s_t^* 를 알고 있을 때 , 가중치 $p(s_t^*, h_t = m \mid z(t))$의 계산은 아래와 같다.

$$\begin{aligned} & p(s_t^*, h_t = m \mid z(t)) \\ & = p(h_t = m \mid s_t^*, z(t))\, p(s_t^* \mid z(t)) = c_{m|s_t^*} \end{aligned} \tag{12}$$

위 식을 (7)에 대입하면, 최종적으로 추정하는 신호 $\hat{Y}(n)$를 구할 수 있다.

$$\hat{Y}(n) = \sum_{m=1}^{M} \hat{Y}_{h_t|s_t^*} c_{m|s_t^*} \tag{13}$$

b. 상태(s_t^*) 결정

Left-Right Hmm을 사용해서 칼만 필터를 수행하기 위해서는 각 프레임의 상태열 $s_1, s_2, \ldots\ldots, s_T$ 알고 있어야 한다. 이러한 상태를 찾는 일은 확률값 $p(s_t = j \mid z(t))$를 사용한다. $p(s_t = j \mid z(t))$는 아래식으로 구할 수 있다.

$$p(s_t = j \mid z(t)) = \sum_{m=1}^{M} p(s_t = j, h_t = m \mid z(t)) \tag{14}$$

Left-Right HMM 모델에서 전의 프레임 상태가 $s_{t-1} = i$ 라고 주어져 있을 때, 현재 프레임의 상태는 i 또는 i+1이 된다. t프레임에서 상태 s_t가 i 또는 i+1이라는 가정하에 확률값을 비교해서 구할 수 있다. i, i+1 일 때의 확률의 비를 D(t)라고 두고 아래식과 같이 정의한다.

$$D(t)=\frac{p(s_t=i|z(t))}{p(s_t=i+1|z(t))}\begin{matrix}s_t=i\\ \geq\\ <\\ s_t=i+1\end{matrix}1 \quad (15)$$

(16)식에서 D(t)≥1이면 현재 프레임의 상태는 $s_t=i$ 이고, D(t)<1 이면 현재 프레임의 상태는 $s_t=i+1$이 된다. s_t가 정해지면 그 값을 s_t^*라고 둔다. $p(s_t=j,h_t=m|z(t))$의 계산은 아래식과 같이 베이시안 법칙으로 나타낼 수 있다.

$$p(s_t=j,h_t=m|z(t)) = \frac{p(z(t)|s_t=j,h_t=m,z(t-1))p(s_t=j,h_t=m|z(t-1))}{p(z(t)|z(t-1))} \quad (16)$$

(16)식의 첫번째 요소를 아래식으로 근사화시킬 수 있다.

$$p(z(t)|s_t=j,h_t=m,z(t-1))=\prod_{n=1}^{N}p(z(n)|s_t=j,h_t=m) \quad (17)$$

(17) 오른쪽 식의 계산은 아래와 같이 정규분포($N[\cdot]$) 식으로 나타내어진다.

$$p(z(n)|s_t=j,h_h=m,z(t-1)) = N[\hat{y}_{mj}(n),HP_{mj}H^T] \quad (18)$$

(16)식의 두번째 요소 $p(s_t=j,h_t=m|z(t-1))$는 주어진 Markov 과정으로 나타낼 수 있다.

$$p(s_t=j,h_t=m|z(t-1)) = \sum_{l=1}^{M}p(s_t=j,h_t=m|s_{t-1}=i,h_{t-1}=l,z(t-1))\times p(s_{t-1}=i,h_{t-1}=l,z(t-1)) \quad (19)$$

위 식에서 첫번째 요소는 아래식으로 다시 쓸 수 있다.

$$p(s_t=j,h_t=m|s_{t-1}=i,h_{t-1}=m,z(t-1)) = p(h_t=m|s_t=j,s_{t-1}=i,h_{t-1}=l,z(t-1))\times p(s_{t-1}=i,s_{t-1},h_{t-1}=l,z(t-1)) \quad (20)$$

h_t 와 s_t 은 서로 독립이므로 위 식의 첫번째 두번째 요소를 아래와 같이 다시 쓸 수 있다.

$$p(h_t=m|s_t=j,s_{t-1}=i,h_{t-1}=l,z(t-1))=c_{mj} \quad (21)$$

$$p(s_t=j|s_{t-1}=i,h_{t-1}=l,z(t-1))=p(s_t=j|s_{t-1}=i)=a_{ij} \quad (22)$$

(21), (22)식을 (19)식에 대입하면 아래식과 같이 나타내어진다.

$$p(s_t=j,h_t=m|z(t-1)) = \sum_{l=1}^{M}a_{ij}c_{mj}p(s_{t-1}=i,h_{t-1}=l|z(t-1)) \quad (23)$$

(16)식의 분모는 상태 j에 독립적이므로, 이 요소는 스케일 인수이다. 그러므로, $p(s_t=j,h_t=m|z(t))$는 전의 확률을 사용해서 계산할 수 있다.

$$p(s_t=j,h_t=m|z(t)) = D_t N_{mj}\sum_{l=1}^{L}a_{ij}c_{mj}p(s_{t-1}=i,h_{t-1}=l|z(t-1)) \quad (24)$$

(24)식에서 D_t는 가중치 요소들이 모든 합이 1이 되게 하는 스케일 인수이다.

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

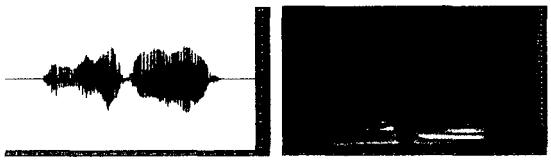
전체적인 음성 알고리즘의 순서는 [그림2]에 나타내었다.

IV. 실험 및 결과

제안된 방법은 SNR이 각각 0dB, 5dB, 15dB 에서 백색 잡음이 부가되었을 때 음성향상 결과를 보여주고 있다.

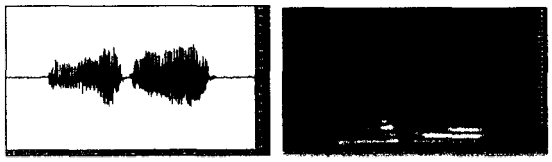
학습은 한 명의 여성이 발성한 깨끗한 음성으로, 6개의 "안녕하세요" 문장을 사용하였다. 테스트는 동일 문장으로 학습에 사용하지 않은 문장을 사용하였다. 이 실험에서, 샘플링 주파수는 11,025Hz이고, 깨끗한 음성의 AR 모델은 15차이다. HMM에서 5 상태, 4 mixture를 사용하였고, Pentium IV 2.0Gz PC에서 실험 하였다.

[그림1]은 SNR 5dB 환경에서 음질 향상 결과를 나타내고 있다. (a),(b)는 각각 깨끗한 음성과 잡음이 부가된 음성이고, (c),(d)는 기존의 방법과 제안된 방법으로 음질을 향상 시킨 결과 이다. 들어본 결과는 많은 차이를 느낄 수 없었다. [표1]은 각 dB 별로 음질이 향상된 결과를 비교하고 있다. 전체적으로 기존의 방법에 비해 제안되 방법의 성능이 약 0.4~0.8dB 정도 떨어지나, 시간은 약 66%정도 감소하는 것을 볼 수 있다.

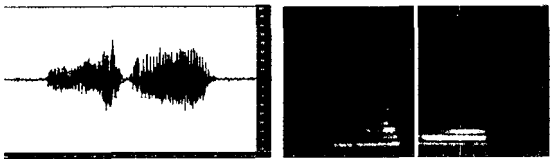


(a)Clean Speech

(b) 5 dB Noisy Speech



(c) Enhanced Speech(conventional method)

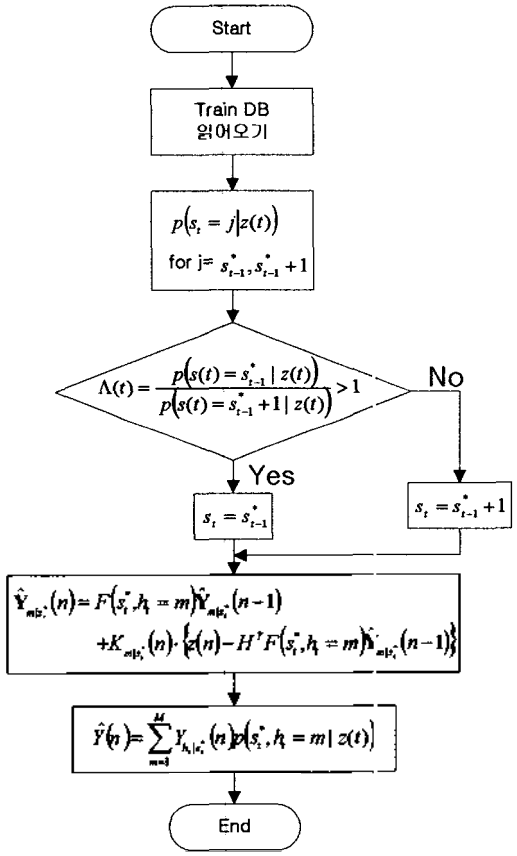


(d) Enhanced Speech(proposed method)

[그림1] 음질 향상 결과 파형

	Ergodic model		Left-Right model	
Input SNR[dB]	Output SNR[dB]	시간 (초)	Output SNR[dB]	시간 (초)
0	8.5	6	8.3	2
5	11.35		11.01	
10	15		14.6	
20	24.2		23.95	

[표1] 기존 방법과 제안된 방법의 음질 향상 결과 비교



[그림2] multiple kalman filter 수행 알고리즘

V.결론

본 논문에서는 Left-Right HMM 모델에 기초를 둔 음질 향상 방법을 제안하였다. 기존의 ergodic HMM 모델을 이용한 방법에 비해 SNR은 약간 떨어지나, 시간을 66%정도 절약할 수 있었다. 하지만, 상태 결정 방법이 아직은 불안정하여 이 부분에 대한 연구가 좀 더 필요하다. 이 방법은 발성하는 문장이 고정되어 있는 화자 인식이나 음성 인식 분야에서 개인별 음성 향상 방법으로 발전시킨다면, 화자 인식이나 음성 인식 분야에서 좀 더 좋은 성능을 얻을 수 있게 될 것이라 생각한다.

참고 문헌

[1] Y. Ehpraim, " A Bayesain approach for speech enhancement using hidden Markov models," *IEEE Trans. Signal Processing*, vol. 41, pp.725-735, Apr, 1992

[2] Y. Ehpraim, D, Malah, B.-H. Juang, " On the application of hidden Markov models for enhancing noisy speech," *IEEE Trans. Acoustic. Speech Process.* ASSP-37 1846-1856

[3] Ki Yong Lee, Jae Yeal Rheem, " Smoothing approach using forward-backward Kalman filter with Markov Switching Parameters for Speech Enhancement," *IEEE Trans. Signal Processing*, vol. 80, pp.2579-2588, Apr, 2000

[4] Ki Yong Lee, Katsuhiko Shirai, " Efficient Recursive Estimation for Speech Enhancement in Colored Noise," *IEEE Signal Processing Letters*, vol. 3, no.7 pp.196-199, July, 1996