

잡음에 강인한 음성인식을 위한 환경 파라미터 변환에 관한 연구

강철호, 흥미정
광운대학교 전자 통신 공학과

A Study on Environment Parameter Compensation Method for Robust Speech Recognition

Mee Jeong Hong, Chul Ho Kang
Dept. of Electronics Communication Engineering, Kwangwoon Univ.

요약

최근 음성 인식 기술의 발전으로 음성 인식 시스템의 실용화가 점차 증가함에 따른 가장 큰 문제점은 음성 인식기의 인식환경과 학습환경과의 차이로 인해 음성 인식기의 성능이 급격히 떨어지는데 있다. 이를 해결하기 위해 본 논문에서는 기존의 잡음처리 방법 중 CMS(Cepstral Mean Subtraction)와 환경 잡음 (부가 잡음, 채널 왜곡)을 동시에 추정하는 최신 모델 보상 기법인 VTS(Vector Taylor Series)를 소개하고 그 성능을 비교하였다.

1. 서론

최근에 음성인식 시스템의 실용화가 활발히 진행되고 있는 가운데 잡음에 강인한 음성인식 시스템의 필요성이 매우 중요하게 요구되고 있는 실정이다.

이는 잡음이 없거나 비교적 조용한 실형실 환경에서 우수한 성능을 나타내는 음성인식 시스템이 환경 잡음 (주위 잡음)이 존재하는 환경에서는 성능이 급격히 떨어지기 때문이다. 따라서 음성인식 시스템의 실용화를 위해서는 잡음에 대한 대책이 반드시 필요하다.

음성인식 시스템의 성능을 저하시키는 음향적, 환경적 변화의 여러 가지 요인 중 특히 인식이 수행되는 환경에 따라 음성인식 시스템의 훈련환경과 인

식 환경이 서로 다르기 때문에 나타난다. 훈련환경과 인식환경과의 차이를 야기시키는 요인으로는 배경잡음, 마이크로 폰, 통신 채널, 화자의 발성형태 등을 들 수 있다.

본 논문에서는 음성인식 시스템의 성능을 저하시키는 요인 중 부가 잡음과 채널 왜곡을 동시에 감소시키는 방법으로 일반적으로 알려진 기존의 방법들을 간단히 살펴 보고, 그 중 CMS를 이용한 환경잡음의 전처리 방법과 최신 기법인 VTS를 비교하며, 깨끗한 음성으로 보상 가능하도록 하고자 한다.

2. 기존의 환경 잡음 전처리 방법

환경에 강인한 음성인식 시스템을 구현하는 방법은 음성 강화(Speech Enhancement), 잡음에 강한 특징 추출(Robust Feature Extraction), 잡음에 강한 거리측정(Robust Distance Measure), 모델에 기반을 둔 보상방법(Model-based Compensation)을 들 수 있다.

2.1 SS (Spectral subtraction)

비음성구간에 존재하는 부가적인 잡음의 평균치를 전체음성 스펙트럼 값에 대해서 모두 차감하는 간단한 방법이다. 이는 음성신호의 초기목음구간에 부가적인 잡음이 존재한다는 가정에 따른다.

2.2 CMS (Cepstral Mean Subtraction)

아래와 같이 cepstral 영역에서의 전체 평균값을 차감함으로써 필터링과 유사한 효과를 가지게 하는 방법이다.

$$c_{\hat{x}}[n] = c_x[n] - \frac{1}{N} \sum_{n=1}^N c_x[n]$$

2.3 RASTA

음성 스펙트럼의 각 성분 내에서 음성에 비해 느리게 변화하는 부분을 아래의 전달함수를 고역통과 필터를 거쳐 억압되도록 하는 방법이다.

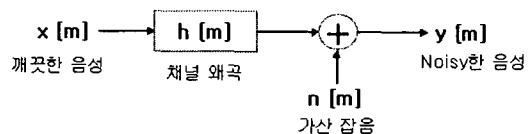
$$H(z) = \frac{z^4(0.2 + 0.1z^{-1} - 0.1z^{-3} - 0.2z^{-4})}{(1 - 0.98z^{-1})}$$

3. VTS(Vector Taylor Series) 이론

3-1. 잡음 환경의 모델링

훈련환경과 인식환경이 차이가 날 때 음성인식기의 성능은 저하된다. 두 환경 사이의 차이를 야기시

키는 요인은 그림 1과 같이 크게 두 가지로 분류할 수 있는데 기계소리나 다른 화자의 음성등과 같은 부가적인 잡음(additive noise)과 방안에서의 반향, 마이크로폰, 통신 채널 등의 선형 필터링(linear filtering)을 들 수 있다.



[그림 1] 환경 모델 블록도

Clean한 음성을 나타내는 vector x 가 주위환경으로부터 영향을 받아 새로운 vector y 를 만들었다고 가정하자. 이때 vector y 는 noisy 음성을 나타내고 다음과 같은 식으로 나타낼 수 있다.

$$y = x + g(x, a_1, a_2, \dots) \quad (1)$$

여기에서 함수 $g(\cdot)$ 는 환경함수(Environment function)이고, a_1, a_2 는 환경을 나타내는 파라미터들(vector, 상수, 행렬, ...)이다. 함수 $g(\cdot)$ 은 환경파라미터들에 대한 지식은 필요 없을지라도 완벽하게 알려진 것으로 가정한다.

이 경우에 있어 clean 음성과 noisy 음성 사이의 관계는 log-spectral영역에서 표현할 경우

$$y[k] = x[k] + g(x[k], h[k], n[k]) \quad (2)$$

이고, 벡터 표현 식으로 나타내면,

$$y = x + g(x, h, n) \quad (3)$$

이다

여기에서 환경함수 $g(x, h, n)$ 는 다음과 같이 나타낼 수 있다.

$$g(x, h, n) = h + 10 \log_{10} \left(i + 10^{\frac{n-x-h}{10}} \right) \quad (4)$$

이 경우 i 는 단위 벡터이며, 모든 vector의 차수는 L 이다. VTS에 접근하기 위한 첫번째 가정은 환경 파라미터는 다음과 같은 vector 들이다.

$$h = \begin{bmatrix} h[0] \\ \vdots \\ h[L-1] \end{bmatrix}, n = \begin{bmatrix} n[0] \\ \vdots \\ n[L-1] \end{bmatrix} \quad (5)$$

여기에서, 성분 $h[k]$ 는 채널 $|H(\omega_k)|^2$ 의 Power Spectrum 의 k 번째 log spectral mel 성분이다. $n[k]$ 도 유사하다. 두 번째 가정은 clean 음성의 log-spectrum random variable 는 다음과 같은 Gaussian 분포의 mixture에 의해 나타낼 수 있다.

$$P(x_i) = \sum_{k=0}^{K-1} P_k N_{x_i}(\mu_{x,k}, \sum_{x,k}) \quad (6)$$

여기에서 K 는 mixture 개수이다.

3-2. VTS 근사화 방법

y 의 확률 밀도 함수에 대한 해를 얻기 위해서는 확률 분포가 Gaussian 분포가 되도록 단순화 시킨다. 이를 위해서 환경 벡터함수 $g(x, a_1, a_2, \dots)$ 을 VTS 근사화에 의해 대치한다. 이를 기반으로 x 와 y 사이의 관계는

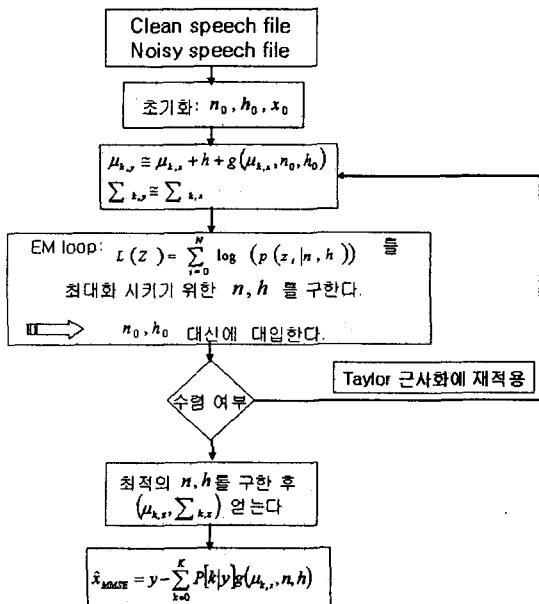
$$\begin{aligned} y &= x + g(x_0, a_1, a_2, \dots) + g'(x_0, a_1, a_2, \dots)(x - x_0) \\ &\quad + \frac{1}{2} g''(x_0, a_1, a_2, \dots)(x - x_0)(x - x_0) + \dots \end{aligned} \quad (7)$$

환경 모델에 의거한 Taylor series는 아래와 같이 나타낼 수 있다.

$$\begin{aligned} y &= x + g(x_0, h, n) + g'(x_0, h, n)(x - x_0) \\ &\quad + \frac{1}{2} g''(x_0, h, n)(x - x_0)(x - x_0) \end{aligned} \quad (8)$$

(8)식을 기반으로 하여 y 의 평균벡터와 공 분산 행렬을 구하고, EM기법을 사용하여, 부가잡음(n)과 채널왜곡(h)를 추정한 후 MMSE방법을 이용하여 noisy 한 음성이 주어질 때 clean 음성을 계산하게 된다.

[그림 2]는 VTS근사화의 순서도를 나타낸 것이다.



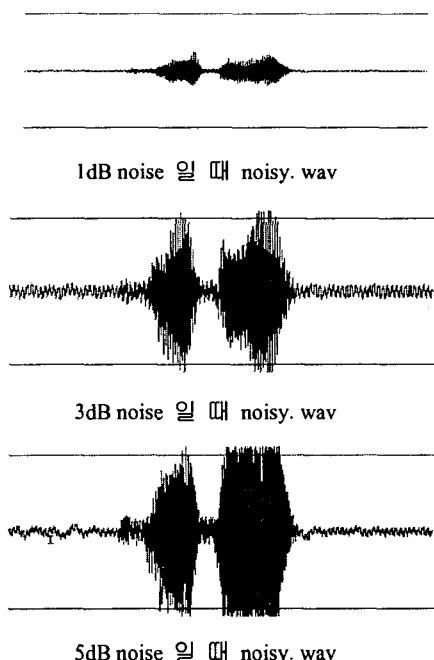
[그림 2] 0차 VTS 근사화 블록도

4. 실험 및 고찰

첫 번째 실험은 “안녕하세요” 라고 발음한 깨끗한 음성과 1dB, 3dB 5dB의 noise(white Gaussian)를 Convolution 하여 얻은 noisy 음성으로, 두 번째 실험은 실제 잡음환경(거리 잡음, 전철 잡음)에서 얻은 noisy음성으로 CMS기법과 VTS방법을 비교하였다.

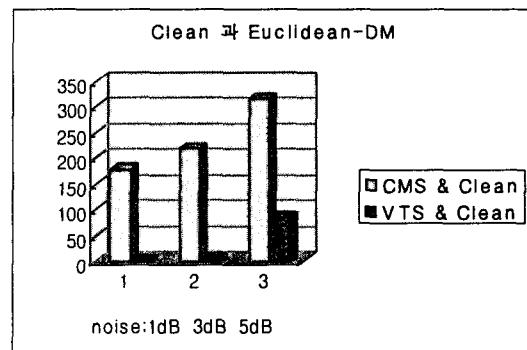
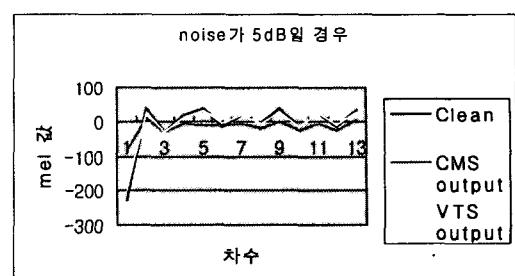


[그림 3] “안녕하세요” (clean.wav)



[그림 4] 1dB, 3dB, 5dB 경우의 noisy speech

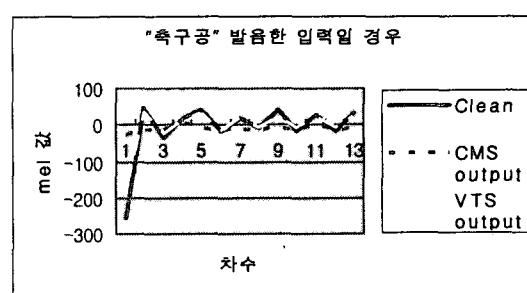
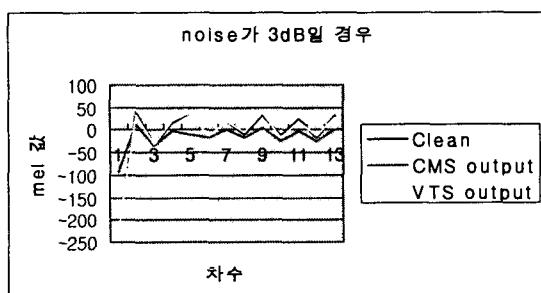
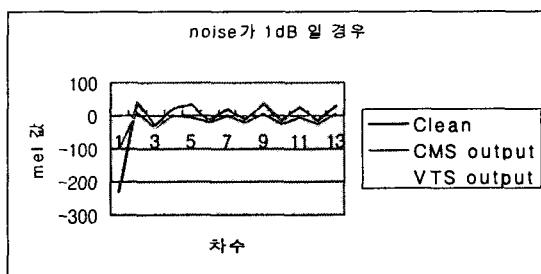
그림 4와 같이 잡음의 3경우 데시벨 별로 얻은 noisy한 음성을 입력으로 하여 CMS 와 VTS 를 수행한 결과를 그림 5에 나타내었다.

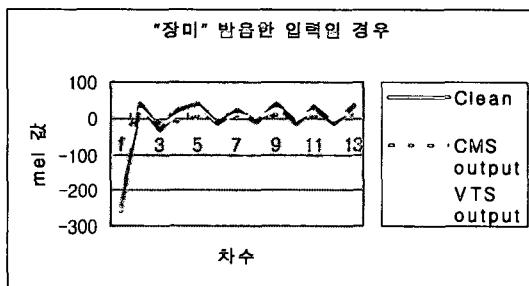


[그림 5] 각각의 noise 데시벨 경우에 CMS와 VTS 의 결과값들과 Clean과 Euclidean Distance Measure이용하여 비교한 결과.

그림 5로부터 CMS 에 비해 VTS 결과 값이 Clean 과 훨씬 더 비슷함을 보였으며, 잡음이 증가할수록 마찬가지로 VTS결과는 CMS 결과보다 Clean 과 유사 함을 보였다. 이것은 Euclidean Distance Measure를 이용하여 비교해 보면 더욱 정확히 확인할 수 있다.

앞서 말한 바와 같이 두 번째 실현은 거리잡음과 전철잡음인 실제 환경에서 녹음한 실제 데이터를 입력으로 했을 경우 비교한 결과를 다음 그림에서 볼 수 있다.





[그림 6] 실제 데이터 입력인 경우 비교한 결과

그림 6에서 보는 바와 같이 실제 환경에서 녹음한 음성 데이터를 사용해 본 결과 VTS 가 Clean 과 많이 유사하기 때문에 Clean한 음성으로 보상할 확률이 크다는 사실을 다시 한 번 확인 할 수 있었다.

5. 결론

본 연구에서는 환경잡음에 강인한 voice portal system의 인식률 향상을 목표로 음성 인식 시스템의 성능을 저하시키는 요인 중 부가 잡음과 채널 왜곡을 동시에 감소시키는 방법으로 기존의 방법 중 CMN 처리 방법과 최신 기법인 VTS를 도입하여 그 유효성을 비교 검토 하였다.

실험 결과, 잡음의 데시벨을 증가 시켰을 때 얻은 음성 데이터를 입력으로 한 경우 VTS결과가 CMS보다 훨씬 더 Clean한 음성과 유사함을 볼 수 있었고, 실제 환경잡음(거리잡음, 전철잡음)에서 녹음한 데이터를 입력으로 한 경우 동일한 결과를 얻을 수 있었다. 이것으로 잡음과 채널 왜곡을 동시에 추정하는 기법인 VTS 을 적용할 경우 Clean한 음성으로 보상 할 수 있는 가능성이 훨씬 크다는 사실을 알 수 있었다.

향후 현재까지 검토한 결과를 바탕으로 여러 가지 실제 환경잡음에서 얻은 음성데이터를 이용하여 인터넷용 Voice Portal System에 적용시켜 인식기의 성능을 향상시키고자 한다.

[참고 문헌]

- [1] P. Moreno, "Speech Recognition in Environments", PhD thesis, Carnegie

Mellon Univ, April 1996

- [2] P. J. Moreno, B. Raj, and R.M. "A Vector Taylor Series Approach for Environment-Independent Speech Recognition," Proc. ICASSP-96
- [3] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "Hmm adaptation using vector taylor series for noisy speech recognition", Processing of ICSLP,2000
- [4] M. F. Gales, "Model-Based Techniques for Noise Robust Speech Recognition", Ph.D. Thesis, Engineering Department, Cambridge University, Sept 1995
- [5] N.S. Kim, D.Y. Kim, and C.K. Un, "Environment compensation based on VTS with noise statistics", IEEE Signal Processing Letter, submitted for publication.
- [6] 김명수, 정현열, "Histogram 처리와 Noise Threshold를 이용한 음성 인식기의 환경잡음 처리 성능향상", 영남대학교 정보통신학과, 1998