2000년도 한국음향학회 학술발표대회 논문집 제19권 제1(s)호

전처리 기법에 따른 잡음음성의 인식성능 비교

손 종 목, 이 용 주, 배 건 성
경북대학교 전자·전기공학부

Comparison of Recognition Performance of Noisy Speech Depending on Preprocessing Methods

Jong Mok Son, Yong Ju Lee, and Keun Sung Bae
School of Electronic and Electrical Engineering, Kyungpook National University
E-mail : sjm@palgong.knu.ac.kr

요 약

본 연구에서는 부가잡음에 의한 음성신호의 왜곡에 대해 다양한 음성개선 기법을 전처리기로 도입하여 HMM(Hidden Markov Model)에 기반한 음성인식 시스템의 인식성능을 평가하였다. 음성개선 기법으로는 MMSE(Minimun Mean Square Error) STSA(Short-Time Spectral Amplitude Estimator) 기법과 웨이브렛 영역에서의 UWD(Undecimated Wavelet Denoising), CWD(Conventional Wavelet Denoising) 기법을 적용하였다. 잡음이 없는 데이터로 훈련한 음성인식시스템에 잡음음성을 입력할때 각 음성개선기법을 전처리기로 사용하여 신호대잡음비(Signal to Noise Ratio)에 따른 인식성능을 비교하였다.

Ⅰ. 서 론

최근 음성인식이 인간과 기계 사이의 자연스러운 통신을 위한 가장 중요한 수단으로 인식되어 이와 관련된 연구가 꾸준히 이루어져 왔으며, 일부 응용분야에서 성공적으로 적용되어 다양한 제품들이 출시되고 있다. 하지만, 좀 더 다양한 응용분야에 음성인식 기술을 적용하기 위해서는 실제 환경에 존재하는 여러가지 주변 잡음에 강인한 특성을 가지는 인식 시스템이 요구된다. 음성인식시스템의 성능은 훈련환경과 인식시의 주변환경이 다를 때 크게 저하되는 것으로 알려져 있다. 만약 인식기가 사용되는 환경을 미리 알고있다면, 주변환경을 고려한 훈련을 통하여 음성신호의 왜곡에 의한 인식성능 저하를 상당히 줄일 수 있다. 그러나, 실제 환경에서는 그 주변환경이 다양하게 변화하기 때문에 훈련시 이를 충분히 고려할 수 없다. 때문에, 왜곡된 음성신호에 대해 인식성능을 향상시키기 위해서는 인식시에 바람직하지 못한 왜곡을 어느정도 제거할 필요가 있다.

음성인식시스템의 인식성능을 저하시키는 음성신호의 왜곡은 크게 부가잡음에 의한 왜곡과 채널특성에 의한 왜곡으로 나눌 수 있다. 부가잡음은 음성신호에 합의 형태로 나타나고, 채널특성에 의한 왜곡은 음성신호에 대해 컨벌루션의 형태로 나타난다. 때문에, 부가 잡음에 왜곡된 음성신호의 개선은 주파수 영역이나 웨이브렛 영역에서, 채널특성에 의한 왜곡은 CMN(Cepstral Mean Normalization), CDCN(Codeword Dependent Cepstral Normalization) 방법과 같이 켑스트럼 영역에서 연구가 많이 이루어졌다[1,2].

본 연구에서는 부가잡음에 의한 음성신호의 왜곡에 대해 다양한 음성개선기법을 전처리기로 도입하여 HMM에 기반한 음성인식시스템의 성능 저하를 줄이고자 하였으며, 각 음성개선기법에 대해 신호대잡음비에 따른 인식성능을 알아보았다. 음성개선 기법으로는 주파수 영역에서 프레임별 음성부재 확률을 고려한 MMSE STSA 기법과 웨이브렛 영역에서의 UWD, CWD 기법을 사용하였다[3,4,5].

본 논문의 구성은 다음과 같다. Ⅰ장 서론에 이어 Ⅱ장에서는 MMSE STSA, UWD, CWD 음성개선기법에 대해

설명하고, Ⅲ장에서는 실험환경을 기술하며 부가잡음에 의해 왜곡된 신호에 각각의 음성개선기법을 전처리기로 사용하였을 경우의 인식결과를 제시하고, 결과를 검토한다. 마지막으로 Ⅳ장에서 결론을 맺고 향후 연구방향을 제시한다.

Ⅱ. 음성개선기법

2.1 MMSE STSA

부가잡음에 의해 오염된 신호는 다음과 같이 나타낼 수 있다.

$$y(t)=x(t)+d(t) \tag{1}$$

여기서, $x(t)$와 $d(t)$는 각각 잡음에 의해 오염되지 않은 원신호와 잡음 신호를 나타낸다. 식 (1)을 주파수 영역으로 옮겼을때 각 주파수 성분은 식 (2)와 같이 나타낼 수 있다.

$$Y_k=X_k+D_k \tag{2}$$

여기서, 첨자 k는 신호의 k번째 주파수 성분임을 나타낸다. 신호의 각 주파수 성분은 아래와 같이 크기 및 위상 성분으로 나타낼 수 있다.

$$\begin{aligned} X_k &= A_k e^{j\alpha_k} \\ Y_k &= R_k e^{j\beta_k} \end{aligned} \tag{3}$$

인간의 청각 특성은 신호의 위상보다 진폭에 더욱 민감하므로 원 신호의 추정은 진폭 A_k를 추정하는 문제로 간략화 된다. 부가적으로, 신호의 각 주파수 성분들이 독립적이라고 가정하면, A_k의 최소 자승 에러 추정은 식(4), (5)로 구해진다[3].

$$\upsilon_k=\frac{\xi_k}{1+\xi_k}\gamma_{ki} \tag{4}$$

$$\text{where, } \xi_k=\frac{\lambda_x(k)}{\lambda_d(k)},\quad \gamma_k=\frac{R_k^2}{\lambda_d(k)}$$

$$\begin{aligned} \widehat{A_k} &= E\{A_k|y(t),0\le t\le T\} \\ &= E\{A_k|Y_0,Y_1,\ldots\} \\ &= E\{A_k|Y_k\} \\ &= \Gamma(1.5)\frac{\sqrt{\upsilon_k}}{\gamma_k}M(\upsilon_k)R_k \\ &= G_{MMSE}(\xi_k,\gamma_k)R_k \end{aligned} \tag{5}$$

, where

$$M(\theta)=e^{-\frac{\theta}{2}}\left[(1+\theta)I_0\left(\frac{\theta}{2}\right)+\theta I_1\left(\frac{\theta}{2}\right)\right]$$

여기서, $\lambda_x(k)$와 $\lambda_d(k)$는 각각 k번째 주파수 성분에서의 원신호와 잡음 파워의 추정치이다. $\Gamma(\cdot)$는 gamma 함수를 나타내며($\Gamma(1.5)=\sqrt{\pi}/2$), ξ_k는 priori SNR, γ_k는 posterior SNR을 나타낸다. I_0, I_1은 각각 영차와 일차의 변형 베셀 함수를 나타낸다. 식 (5)에 음성부재확률을 도입하면 식 (6)과 같이 나타낼 수 있다.

$$\widehat{A_k}=\frac{\Lambda(\xi_k,\gamma_k,q_k)}{1+\Lambda(\xi_k,\gamma_k,q_k)}G_{MMSE}(\xi_k,\gamma_k)R_k \tag{6}$$

여기서, $\Lambda(\cdot)$는 다음과 같이 정의되는 generalized likelihood ratio이다.

$$\Lambda(Y_k,q_k)=\mu_k\frac{P(Y_k|H_k^1)}{P(Y_k|H_k^0)} \tag{7}$$

$$\text{, where } \mu_k=\frac{(1-q_k)}{q_k}$$

식 (7)에서 q_k는 k번째 주파수 성분의 존재하지 않을 확률이고, H_k^1, H_k^0는 각각 음성 존재와 부존재의 상태를 나타낸다.

2.2 UWD 및 CWD

웨이브렛 변환은 신호처리의 관점에서 대역통과 필터의 출력으로 볼 수 있다. 때문에, 그림 1과 같은 형태의 필터뱅크를 구성하여 웨이브렛 분해를 할 수 있다. 입력신호가 저역통과 필터와 고역통과 필터를 거치게 되면 한 레벨의 웨이브렛 변환이 수행되며, 이를 반복하여 원하는 레벨까지 얻는다. 일반적인 웨이브렛 변환에서는 필터를 통과한뒤 간축과정이 들어가 다음 스케일에서 같은 필터를 사용할 수 있지만, 비간축 웨이

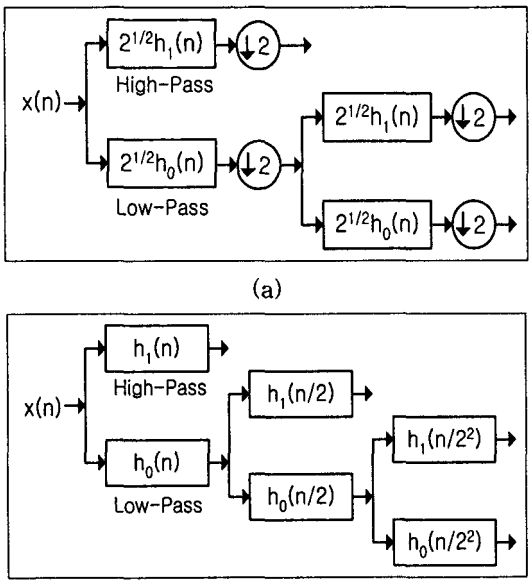

그림 1. 이산 웨이브렛에 대한 dyadic 필터뱅크 구조
(a) Conventional discrete wavelet transform(CWD)
(b) Undecimated wavelet transform(UWD)

브렛 변환에서는 필터를 통과한 신호를 그대로 유지하는 반면, 다음 스케일의 신호를 얻기 위해 필터 계수사이에 '0'을 삽입한다. 위의 웨이브렛 변환을 이용한 음성개선 과정은 다음과 같다[5]. 우선, CWD와 UWD를 위해서 각각 그림 1의 (a), (b)를 이용해 웨이브렛 변환을 수행하고, 무성음에서 신호손실이 많이 발생하는 것을 막기 위해 무성음 구간을 검출하여 유성음 구간과 처리를 달리한다. 그리고, 변환된 영역에서 아래와 같은 Soft-thresholding 기법을 적용하여 음성을 개선한다[6,7].

$$T_{soft}(X)=\begin{cases} sgn(X)(|X|-\lambda), & |X|\geq\lambda \\ 0 & , |X|<\lambda \end{cases} \qquad (8)$$

이때, Threshold 설정을 위한 잡음레벨의 추정은 MAD(Median Absolute Deviation)을 사용하였다. Soft-thersholding을 적용한 웨이브렛 계수를 이용하여 역 웨이브렛 변환을 수행함으로써 개선된 음성신호를 구하였다.

Ⅲ. 실험 및 검토

실험환경은 다음과 같다. 다양한 음성개선기법을 음성인식시스템의 전처리로 사용하였을 경우, 각 개선기법에 따른 인식률을 보기위해 Continuous HMM을 사용하여 고립단어 인식기를 구현하였다. 음성데이터로는 16 kHz, 16 bits로 샘플링된 ETRI (Electronics and Telecommunications Research Institute)의 445 DB 중 임의로 89개의 단어를 선정하여 사용하였으며, 훈련용 화자의 데이터를 사용하여 CHMM을 훈련하고, 테스트용 10명의 화자(남자 6명, 여자 4명)의 데이터를 인식실험에 사용하였다. 테스트용 데이터에 백색 가우시안 잡음을 첨가하여 25, 20, 15, 10[dB]의 신호대 잡음비를 가지는 실험데이터를 만들었다. 각 신호대잡음비에 대한 실험데이터 수는 1777개이다.

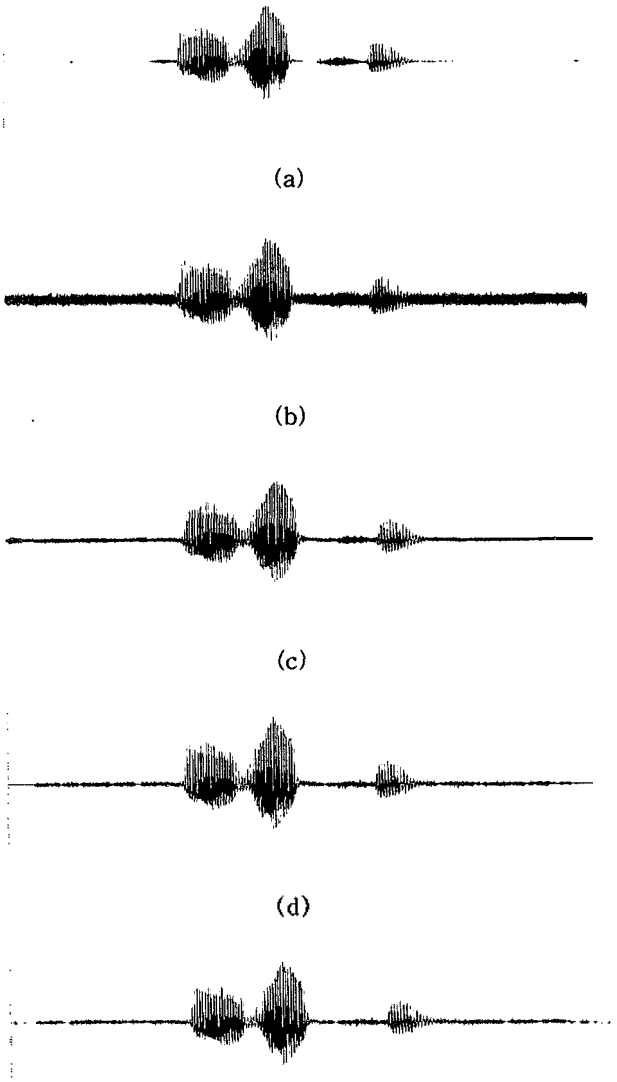

그림 2. 음성개선 예
(a) Clean data (b) Noisy data(10 [dB])
(c) Enhanced data with MMSE STSA (d) Enhanced data with UWD (e) Enhanced data with CWD

표 1. 음성개선기법에 따른 인식률 비교(%)

SNR[dB] / Method	25	20	15	10
Noisy	78.22	43.11	15.31	4.61
MMSE STSA	95.10	90.43	77.88	50.82
CWD	91.23	83.68	54.25	19.92
UWD	88.24	81.26	58.38	23.52

* Clean speech의 인식률 : 98.76 %

실험과정은 다음과 같다. 부가잡음에 오염된 신호를 신호대잡음비별로 각 음성개선기법을 적용하여 개선된 음성 데이터를 생성하였다. 생성된 데이터를 preemphasis 계수 0.95로 전처리한 후, 20 ms의 해밍 윈도우를 10 ms 간격으로 오버랩하여 구간단위 분석하였으며, 각 구간에서 1차의 에너지와 13차의 멜 켑스트럼을 구하고, 현재 구간을 포함한 전후 각 3구간(전체 7구간)의 정보를 이용하여 1차의 차분 에너지와 13차의 차분 멜 켑스트럼을 구하여 인식실험에 사용하였다.

그림 2에 각 음성개선기법을 적용하였을때 개선된 음성신호를 나타내었다. 10 [dB]의 신호대잡음비를 가지는 음성데이터에 대하여 각 개선기법을 적용하였을 때, MMSE STSA 방식은 파형이 매끄럽게 보이는 반면 뮤지컬 잡음이 들렸고, 웨이브렛을 이용한 개선방식의 경우 클릭 잡음이 들렸었다.

각각의 음성개선기법을 음성인식시스템의 전처리기로 사용하였을때, 개선기법에 따른 각 신호대잡음비에서의 인식결과를 표 1에 나타내었다. 표 1은 인식시스템에 음성개선기법을 전처리기로 사용하는 것 외에 부가적인 기법을 인식시스템에서 사용하지 않았을 경우에 대한 실험 결과이다. 잡음에 왜곡되지 않은 음성 데이터에 대해서는 98.76 %의 인식률을 나타내었다. 반면, 음성개선기법을 전처리기로 사용하지 않았을 경우, 부가 잡음에 왜곡된 데이터에 대해서는 신호대잡음비에 따라 인식률이 급격히 나빠지는 것을 볼 수 있다. 전체적으로 음성부재 확률을 고려한 MMSE STSA 방식의 성능이 웨이브렛 방식에 비해 우수한 결과를 보였으며, 웨이브렛 영역에서의 UWD 및 CWD 기법은 서로 비슷한 성능을 보였다.

Ⅳ. 결 론

본 연구에서는 부가잡음에 의한 음성신호의 왜곡에 대해 주파수 영역의 MMSE STSA, 웨이브렛 영역의 CWD, UWD 방식의 음성개선기법을 음성인식시스템의 전처리기로 도입하여, 잡음음성에 대한 인식성능을 평가하였다.

음성개선기법별로 각각 신호대잡음비에 따른 인식률을 구해본 결과 MMSE STSA방식이 다른 방식에 비해서 높은 성능을 나타내었다. 향후 다양한 잡음에 대한 실험과 함께, 채널 특성에 기인한 음성신호의 왜곡을 개선해 주는 방법에 관한 연구가 실제 환경에서 음성인식 기술을 적용하기 위해서 이루어져야한다.

본 연구는 한국과학재단의 특정기초연구과제 (과제번호 : 1999-2-303-001-3) 연구비 지원으로 수행되었습니다.

참 고 문 헌

[1] Alejandro Acero, "Acoustical and Environmental Robustness in Automatic Speech Recognition," *Ph.D. thesis, Carnegie Mellon University*, 1990.

[2] Fu-Hua Liu, Richard M. Stern, Xuedong Huang, Alejandro Acero, "Efficient Cepstral Normalization for Robust Speech Recognition," *Proc. of the sixth ARPA Workshop on Human Language Technology*, 1993.

[3] Yariv Ephraim and David Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109-1121, 1984.

[4] Liu Zhibin, Xu Naiping, "Speech Enhancement Based on Minimum Mean-Square Error Short-Time Spectral Estimation and Its Realization," *IEEE International Conference on Intelligent Processing Systems*, pp. 1794-1797, 1997.

[5] 한미경, 배건성, "웨이브렛 변환을 이용한 음성개선에 관한 연구," *한국음성과학회 제7회 학술발표회 논문집*, pp. 165-172. 1999.

[6] D.L. Donoho, "De-Noising by Soft-Tresholding," *IEEE Trans. on Information Theory*, pp. 961-1005, 1995.

[7] Jongwon Seok and Keunsung Bae, "Speech Enhancement with Reduction of Noise Components in the Wavelet Domain," *International Conference on Acoustics, Speech and Signal Processing*, pp. 425-455, 1997.