Wiener Filtering을 이용한 잡음환경에서의 음성인식

Speech Recognition in Noisy Environments using Wiener Filtering

김 진 영* · 엄 기 완* · 최 홍 섭**

(Jin-Young Kim* · Ki-Wan Eom* · Hong-Sub Choi**)

ABSTRACT

In this paper, we present a robust recognition algorithm based on the Wiener filtering method as a research tool to develop the Korean Speech recognition system. We especially used Wiener filtering method in cepstrum-domain, because the method in frequency-domain is computationally expensive and complex.

Evaluation of the effectiveness of this method has been conducted in speaker -independent isolated Korean digit recognition tasks using discrete HMM speech recognition systems. In these tasks, we used 12th order weighted cepstral as a feature vector and added computer simulated white gaussian noise of different levels to clean speech signals for recognition experiments under noisy conditions.

Experimental results show that the presented algorithm can provide an improvement in recognition of as much as from 5% to 20% in comparison to spectral subtraction method.

1. 서 론

음성인식 기술은 man-machine interface 방법 중 가장 활용도가 높은 기술로써 오랫동안 세계 각국에서 실용화를 위한 연구를 계속해 오고 있다. 이러한 음성인식 기술의 최종목표는 임의의 화자가 발성한 연속적인 음성을 실시간 처리로 높은 인식률을 갖는 인식 시스템의 개발이라 할 수 있다.

그러나 현재의 음성인식 시스템 개발에서 가장 어렵고 필요한 부분으로 실제 시스템이 운영되는 환경과 연구실의 개발 환경과의 차이에 의한 배경잡음의 변화에 대해 적절히 대처할 수 있는 잡음처리 기술을 들 수 있다.

즉, 지금까지 음성인식 시스템은 실험실 환경에서는 우수한 성능을 나타내지만, 학습환경과 인식환경이 일치하지 않을 때, 그 성능은 급격하게 저하된다. 이는 주로 학습환경에 없던 잡음

* 전남대학교 전자공학과
** 대진대학교 전자공학과

의 영향 때문이며, 이런 잡음의 영향에 견인한 인식 알고리듬을 개발하기 위한 노력 또한 다양한 방법으로 진행되어 왔다.[1]-[3]

이 논문의 구성은 다음과 같다. 2장에서는 잡음처리기술로서 Wiener filtering방법과 음성인식 시스템에서의 알고리듬 구현방법에 대해 스펙트럼 차감법과 함께 설명한다. 그리고 3장에서는 본 연구에서 수행한 음성인식 실험과 결과를 검토한 후, 4장에서 향후 연구방향과 함께 결론을 맺는다.

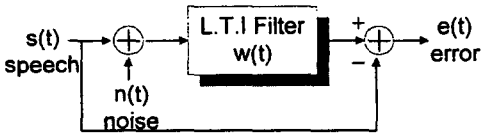
2. 잡음처리 기술

2. 1 Wiener filtering

지금까지 고전적인 적응 필터링 기술의 하나인 Wiener filter는 잡음환경에서의 음성인식 및 음질개선(speech enhancement)분야에 적용되어 우수한 성능을 보여 왔다.[4],[5]

그림 1에서 Wiener filter는 최소평균 제곱오차(Minimum Mean Square Error) 의미에서 원래의 음성신호 s(t)와 필터링된 신호와의 차 e(t)가 최소가 되도록 한다. 그러므로 이 필터는 원래 파형에 충실하면서 신호 대 잡음비(SNR)를 최대가 되도록 한다.

그림 1. Wiener Filter



주파수 영역에서 Wiener filter의 전달함수는 다음 식으로서 정의된다.

$$W(w)=\frac{gP_S(w)}{gP_S(w)+P_N(w)} \qquad (1)$$

$$where \quad g=\frac{R_Y(0)-R_N(0)}{R_S(0)}$$

여기서,

PS(w) : 기준 음성신호의 전력 밀도스펙트럼
PN(w) : 잡음신호의 전력 밀도스펙트럼
PY(w) : 입력 음성신호의 전력 밀도스펙트럼
R(0) : 각 신호의 0번째 자기상관 계수

PN(w)를 구하기 위해서는 음성에 첨가된 잡음의 형태를 미리 알고 있거나 또는, 음성이 없는 묵음 구간에서 잡음의 통계적 특성을 측정할 수 있다는 가정이 필요하다.

그러면 Wiener filtering된 음성신호와의 로그스펙트럼의 오차는 다음과 같이 쓸 수 있다.

$$\begin{aligned} Error &= \int \left| \ln(\,W(w)^2 \cdot P_Y(w)) - \ln(gP_S(w))\,\right|^2 dw \\ &= \int \left| \ln\left(\frac{P_Y(w)\cdot gP_s(w)}{(gP_S(w)+P_N(w))^2}\right)\right|^2 dw \end{aligned} \tag{2}$$

(가) 셉스트럼 영역에서의 해석

식 (2)에서 전력 스펙트럼을 구하는데 많은 계산량이 필요 하게 된다. 그러므로 식(2)의 Error함수를 셉스트럼 영역(cepstrum domain)에서 다시 정의하기 위해, 각 음성신호를 AR(Autoregressive) 모델링하여 로그 전력스펙트럼을 셉스트럼으로 바꾼다.

그러면 식(2)의 Error함수는 셉스트럼 영역에서 다음과 같이 나타낼 수있다.

$$Error = \sum((C_Y(n)+C_S(n) - \ln(g) - 2C_W(n))^2 \tag{3}$$

여기서,

$C_Y(n) = cepst(\ln P_Y(w))$: cepstral of input signal

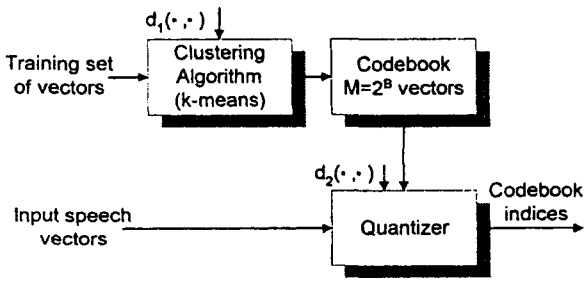
$C_S(n) = cepst(\ln P_S(w))$: cepstral of reference signal

$$C_W(n) = cepst\left(\ln\left(\frac{\alpha_S^2}{|A_S(e^{-jw})|^2} + \frac{\alpha_N^2}{g\cdot|A_N(e^{-jw})|^2}\right)\right)$$

(나) 음성인식 시스템에 적용

다음 그림 2는 벡터양자화 과정을 나타낸다. 위에서 유도한 식(3)의 Error 함수를 코드북벡터와의 거리측정 함수 d2(· , ·)에 사용였으며, 기준신호로는 코드북 벡터의 12차 셉스트럼 계수를 이용하였다.

그림 2. Vector Qantization



그러므로, 식(3)에서와 같이 $C_w(n)$를 구하기 위해서는 먼저 셉스트럼 계수로 구성된 codebook의 codeword를 13개의 자기상관 계수로 바꾸고, 여기에 잡음신호로 부터 구한 자기상관 계수를 이득 g로 나눈 값을 합하여 12차 가중 셉스트럼 계수를 구한다.

그리고 잡음신호의 자기상관 계수는 다음 식(4)와 같이 묵음구간에서 수 프레임 동안의 평균값으로 구한다.

$$R_N(n)=\frac{1}{m}\sum_{i=1}^{m}R_i(n)$$
$$\text{for} \quad n=0,1,\cdots p \tag{4}$$

p : 셉스트럼 차수

$R_i(n)$: i 번째프레임의 자기상관계수

2.2 스펙트럼 차감법

스펙트럼 차감법(spectral subtraction)은 고전적인 잡음 보상법으로 Boll 등에 의해 제안되었다. 이는 음성신호와 배경 잡음신호 사이의 상관계수가 0이고, 잡음신호는 정제적(stationary)이라는 가정하에서, 잡음이 섞인 음성신호의 스펙트럼에서 잡음신호의 스펙트럼 성분만을 제거함으로서, 원래 음성신호를 추정하는 방법이다.[6],[7]

이를 식으로 표현하면 다음과 같다.

$$\widehat{P_S}(\omega)=P_Y(\omega)-E[P_N(\omega)] \tag{5}$$

여기서,

$P_Y(w)$: 잡음이 섞인 음성신호의 short-time power spectrum

$P_N(w)$: 잡음신호의 short-time power spectrum

3. 실험 및 결과

3. 1 음성데이타

인식실험에서 사용하는 음성 데이타는 25명의 화자가 한국어 숫자음 '공' 부터 '구' 까지 10개를 2회씩 발음한 것을 사용하였다. (25명×10발음×2회=500개) 이렇게 녹음된 데이타는, 8kHz로 샘플링하였고 묵음구간과 음성구간을 구분하기 위한 끝점검출을 하였다.[8] 잡음환경에서의 인식실험을 위해, 식(6)으로 정의되는 SNR 값이 3, 6, 9, 12, 15 dB가 되도록 백색 가우시안 잡음을 끝점검출된 음성신호에 인위적으로 첨가하였다.

$$SNR = 10 \log_{10} \frac{\frac{1}{n}\sum_{i=1}^{n} P_S(i)}{P_N} \quad (6)$$

여기서 $P_S(i)$는 i번째 프레임의 음성신호 전력이고, P_N은 잡음신호의 평균전력이다.

3. 2 특징벡터 추출

음성인식 실험에 사용하는 특징벡터는 12차 선형예측 가중 셉스트럼 계수로서 다음과 같이 구한다.

- Windowing

A/D변환된 음성신호는 256샘플을 윈도우의 한 프레임으로 하고 128샘플 간격으로 윈도우를 씌운다.

사용된 Hamming 윈도우 함수는 식(7)과 같다.

$$w(n) = 0.54 - 0.46\cos\frac{\pi i}{(N-1)} \quad (7)$$

- 전처리(Pre-emphasis)

음성신호의 고주파 성분에 대한 감쇄를 보상하기 위해 전달함수가 $H(z) = 1 - 0.98z^{-1}$인 1차 고역통 과 필터를 사용하여 전처리를 하였다.

- 가중 셉스트럼 계수

각 프레임에서 12개의 자기상관 계수를 구하고 이 계수들로 부터 Levinson-Durbin 알고리즘을 이용, 12개의 선형 셉스트럼 계수를 구한다. 그리고 이들 셉스트럼 가중함수로는 식(8)의 Lifter를 사용하였다.

$$h(n) = 1 + 6.5\sin(\frac{n\pi}{16}) \quad \text{for } 0 \le n \le 11 \quad (8)$$

3. 3 음성인식 시스템

음성인식 시스템에서 Wiener filtering방법의 성능을 평가하기 위해 벡터양자화기와 HMM을 이용한 화자독립 음성인식 시스템을 구현하였다.[9] VQ를 위한 Codebook은 128개의 Codeword로 이루어져 있으며, 단어 단위의 3상태 HMM모델을 사용하였다. 학습은 2회씩 발음한 것 중 하나씩 뽑아 단어 당 총 25개를 사용하고 인식실험에서는 구축된 Database를 모두 사용, 인식실험을 수행하였다.

3. 4 인식실험 결과

이 논문에서는 SNR이 각각 3, 6, 9, 12, 15dB일 때 잡음제거 방법을 사용하지 않은 경우(U. E)와 스펙트럼 차감법을 사용한 경우(P. S), 그리고 Wiener filtering방법을 적용한 경우(W. F)에 대해 각각 인식실험을 했으며 이들의 실험결과를 표1에 비교하였다.

표 1. 인식실험 결과 단위 : %

SNR	3	6	9	12	15	clean
U. E	18.5	19.0	23.5	32.3	44.8	97.92
P. S	21.9	30.4	41.3	60.0	72.1	
W. F	30.3	52.4	65.5	70.7	76.6	

4. 결 론

이 논문에서는 잡음환경 하에서 음성인식 시스템의 성능을 향상시키기 위한 방법으로 Wiener filtering방법을 셉스트럼영역에서 적용하였으며, 숫자 음을 대상으로 화자독립 음성인식 실험을 통하여 그 성능을 평가하였다. 실험결과에서 볼 수 있듯이 Wiener filtering방법을 적용한 음성인식 시스템의 성능은, 잡음제거 방법을 사용하지 않은 경우에 대해서 10% - 40%, 그리고 스펙트럼 차감법에 비해서는 5% - 20% 정도 향상됨을 알 수 있다.

그러나 본 논문에서 사용한 실험데이터가 숫자 음이라는 제한된 음성 data base이므로 통계적인 결론을 얻기에는 부족한 면이 많다. 따라서 앞으로 더 많은 화자 및 데이터에 대한 실험을 통해 제안된 방법의 성능에 대한 통계적인 검증이 필요하겠다. 그리고 벡터양자화 과정에서의 계산속도를 향상시키는 방법에 대한 연구도 앞으로의 과제라 생각된다.

참 고 문 헌

[1] Jean-Claude Junqua & Jean-Paul Haton. 1996. Robustness in Automatic Speech Recognition. Kruwer Academic Publishers.

[2] A. Acero & R. M. Stern. 1990. "Environmental robustness in the automatic speech recognition." Proc. ICASSP-90, 849-852.

[3] S. Kay. 1980. "Noise compensation for autoregressive spectral estimation." IEEE Trans. on ASSP, ASSP-28(3), 292-303.

[4] A. Berstein & I. Shallom. 1991. "An hypothesized Wiener filtering approach to noisy speech recognition." Proc. ICASSP-91, 913-916.
[5] Ephraim Y., Malah D., and Juang B. 1988. "On the application of hidden Markov models for enhancing noisy speech." Proc. ICASSP-88, .533-536.
[6] S. F. Boll. 1979. "Suppression of acoustic noise in speech using spectral subtracting." IEEE Trans. on ASSP, ASSP-27(2), 113-120.
[7] J. S. Lim. 1985. "Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise." IEEE Trans. on ASSP, ASSP-26(5), 471-472.
[8] L. R. Rabiner & M. R. Sambur. 1975. "An algorithm for determining the endponts of isolated utterances." BS서, .297-315.
[9] S. E. Levison & L. R. Rabiner. 1983. "An introduction to the application of the theory of probabilistic function of a Markov process to automatic speech recognition." BSTJ 62(4), 1035-1073.

접수일자 : '97. 1. 25.
게재결정 : '97. 2. 19.

▲ 김진영
광주광역시 북구 용봉동 300
전남대학교 전자공학과 (우편번호 : 500-757)
Tel : (062) 520-6398 FAX : (062) 514-6472
e-mail: kimjin@dsp.chonnam.ac.kr

▲ 엄기완
광주광역시 북구 용봉동 300
전남대학교 전자공학과 (우편번호 : 500-757)
Tel : (062) 514-6472(O) FAX : (062) 514-6472
e-mail: eom@dsp.chonnam.ac.kr

▲ 최홍섭
경기도 포천군 포천읍 선단리 산11-1
대진대학교 공과대학 전자공학과 487-800
Tel : (0357) 539-1903 FAX : (0357) 539-1900
e-mail: hschoi@road.daejin.ac.kr