잡음환경에 강인한 음성인식을 위해 SNR과 마스킹 효과를 이용한 적응 스펙트럼 차감법

김태준*°, 김종훈*, 이경모*, 이정현**
인하대학교 컴퓨터 정보공학과*, 인하대학교 컴퓨터 공학부**
{hope673*°, jhkim*, kmo21*}@nlsun.inha.ac.kr, jhlee@inha.ac.kr**

Adaptive Spectral Subtraction Method Using SNR and Masking Effect for Robust Speech Recognition in Noisy Environments

Tae-Jun Kim*°, Jong-Hun Kim*, Kyoung-Mo Lee*, Jung-Hyun Lee**
Dept. of Computer Science & Information Engineering, Inha University*
School of Computer Science & Engineering, Inha University**

요 약

스펙트럼 차감과정에서 발생하는 잔류 잡음을 제거하는 방법으로 파라메터를 이용하는 적응 스펙트럼 차감법이 있다. 이는 파라메터를 증가시켜 잔류 잡음을 감소시키는 방법이지만 파라메터를 과도하게 증가시킬 경우 음성 왜곡이 발생한다. 따라서, 적절한 파라메터를 추출하기 위하여 SNR이나, 마스킹 효과 등을 이용한 방법들이 제안되었으나 과도한 잡음의 제거로 인한 음성 왜곡 문제와 낮은 SNR에서 부정확한 파라메터의 추출 문제는 여전히 해결해야 할 과제로 남아있다. 본 논문은 기존의 SNR을 이용한 방법에 마스킹 효과를 적용한 수정된 적응 스펙트럼 차감법을 제안한다. 제안된 방법에서는 마스킹 임계치를 이용하여 잡음 추정값을 재 계산 함으로써 SNR을 향상시켰고, 이를 이용하여 파라메터를 추출함으로써 성능을 개선했다. 성능평가 결과, 제안한 차감법을 적용한 음성신호를 고립단어 음성인식 시스템에 적용했을 때 기존의 방법 보다 인식률이 향상된 것을 확인할 수 있었다.

1. 서론

스펙트럼 차감법은 잡음 환경에서 발생하는 배경 잡음을 제거하는 가장 간단하고 효과적인 방법이다. 하지만, 스펙트럼 차감 과정에서 발생하는 잔류 잡음은 음성인식의 성능을 크게 저하시키고 사람의 귀에 거슬리는 소리를 발생시킨다. 따라서, 과거부터 잔류 잡음을 제거하기 위해 파라메터를 이용하는 적응 스펙트럼 차감법이 제안 되어 왔다[1][2]. 이는 파라메터를 이용하여 배경잡음의 차감량을 조절하는 방법으로써 파라메터를 크게 하면 많은 배경잡음이 제거되어 잔류 잡음은 감소하지만 음성이 왜곡되는 문제점이 있고 반대로, 파라메터를 작게 하면 음성왜곡은 발생하지 않지만 적은 배경잡음이 제거되어 잔류 잡음을 증가시키는 문제점이 있어서, 적응 스펙트럼 차감법에서는 적절한 파라메터를 추출하는 기술이 필요하다. 기존의 기술로는 잡음 추정값으로 계산된 SNR을 이용하거나[1], 마스킹 효과를 이용하여 마스킹 임계치를 계산한 후 이 임계치를 가지고 파라메터를 추출하는 방법이 있다[2]. 하지만, 전자의 경우 잡음이 파라메터에 의해 필요이상 제거되어 음성 왜곡을 발생시키며, 후자의 경우 마스킹 임계치를 계산할 때 음성의 추정값을 사용하기 때문에 음성의 레벨이 낮은 구간에서 임계치가 낮게 측정되어 잡음이 덜 차감된다는 문제점을 가지고 있다. 본 논문에서는 기존의 적응 스펙트럼 차감 시스템의 파라메터 계산 방법의 문제점을 보완하기 위해 앞서 말한 SNR을 이용한 시스템을 기반으로 마스킹 효과를 적용하여 파라메터를 추출하는 수정된 적응 스펙트럼 차감법을 제안한다. 제안된 방법의 성능평가를 위해 기존의 적응 스펙트럼 차감법과 비교 평가하였다.

2. 관련 연구

2.1 파라메터를 이용하는 적응 스펙트럼 차감법

기존의 스펙트럼 차감법의 문제점은 음성신호를 처리하는 과정에서 잔류 잡음이 발생된다는 것이다. 적응 스펙트럼 차감법은 잔류 잡음을 제거하기 위하여 식(1)과 같이 파라메터 α와 β를 추가하여 잔류 잡음이 가청 되는 것을 최소화 시킨다.

$$|\hat{S}(w)|^2 = \begin{cases} |Y(w)|^2 - \alpha|\hat{D}(w)|^2, & if\ |Y(w)|^2 > \alpha|\hat{D}(w)|^2 \\ \beta|\hat{D}(w)|^2, & otherwise \end{cases} \quad (1)$$

♦ Overestimation factor $\alpha(\alpha > 1)$: short-time 스펙트럼에서 추정된 잡음은 α에 의해 필요이상으로 제거되어진다. 이는 잔류 잡음을 줄일 수는 있지만 음성의 왜곡을 발생시킨다.

♦ Spectral flooring $\beta(0 \le \beta \ll 1)$: β는 배경잡음을 추가 시켜 잔류 잡음을 감소 시키는 파라메터이다. 잔류 잡음을 줄일 수 있는 반면 배경잡음을 증가시킨다[3].

2.2 마스킹 효과와 잡음 마스킹 임계치 계산

큰 음에 의해 작은 음이 들리지 않게 되는 것을 마스킹 효과라 하며 작은 음은 큰 음에 의해 계산된 마스킹 임계치 이상의 음만 가청 되어 진다. 이때, 마스킹을 하는 큰 음을 마스커, 마스킹 되는 작은 음을 마스키 라고 한다. 마스킹의 종류에는 주파수 마스킹과 시간적 마스킹이 있고 주파수 마스킹에서 마스킹 효과가 일어나는 주파수의 폭을 크리티컬 밴드라고 부르며 마스킹 임계치를 계산하는데 사용되어진다[4]. 식(2)는 마스킹 임계치를 계산하는 식이다.

$$T(f_m, f) = \begin{cases} T_{\max}(f_m)\left(\frac{f}{f_m}\right)^{28}, & if\ f \le f_m \\ T_{\max}(f_m)\left(\frac{f}{f_m}\right)^{-10}, & if\ f > f_m \end{cases} \quad (2)$$

f과 f_m는 각각 마스키와 마스커의 주파수이고 $T_{max}(f_m)$은 주파수 f_m에 따른 상대적인 마스킹 임계치로써 마스커와 마스키의 tone-like 또는 noise-like특성에 따라서 결정이 된다. 즉, tone-like특성 이라면 -(14.5+i)dB의 값을, noise-like특성

이라면 -5.5dB의 값을 갖는다. 하지만, 본 논문에서는 계산에 소요되는 시간을 줄이기 위하여 음성신호는 낮은 크리티컬 밴드에서는 tone-like특성을 갖고 높은 크리티컬 밴드에서는 noise-like특성을 갖는다는 사실을 기반으로 계산하였다[2].

3. SNR과 마스킹 효과를 이용한 적응 스펙트럼 차감법

본 논문에서 제안하는 적응 스펙트럼 차감 시스템의 전체 구성은 그림 1과 같다.

첫번째 단계로 입력된 신호를 음성과 비음성 구간으로 검출하고 검출된 비음성 구간에 대하여 잡음 추정값($|\hat{D}(w)|$)을 계산한다. 이 잡음 추정값은 이후에 마스킹 효과를 적용한 새로운 잡음 추정값을 계산할 때 사용되어 진다. 음성구간에 대해서는 윈도우함수(Hanning Window)를 곱하여 프레임들로 나누고 나눠진 프레임들에 대하여 FFT(Fast Fourier Transform)을 사용하여 주파수 축으로 변환한다. 두번째 단계에서는 각 프레임마다 마스킹 임계치를 계산하게 된다. 임계치 계산을 위해 마스커에 해당하는 잡음이 추가되지 않는 음성신호($|S(w)|$)가 필요하지만 계산이 불가능 하므로 음성신호의 추정값을 사용하게 된다. 음성신호의 추정값은 앞서 계산한 잡음 추정값을 해당 프레임의 잡음이 섞인 음성신호($|Y(w)|$)에 차감을 해줌으로써 얻을 수 있다. 추정된 음성신호는 크리티컬 밴드 분석을 통하여 마스킹 임계치를 계산하는데 사용 되어진다[2]. 세번째 단계에서는 앞서 계산된 마스킹 임계치를 고려하여 해당 프레임의 임계치 보다 높아 차단되지 않고 가청 되는 잡음을 계산 한다. 이렇게 계산된 잡음 추정값은 첫번째 단계에서 계산된 잡음 추정값과 구별하기 위하여 앞으로 마스킹 효과를 고려한 잡음 추정값 $|\hat{D}m(w)|$라고 표기 하겠다. 네번째 단계에서는 마스킹 효과를 고려한 SNR을 가청 되는 음성신호와 마스킹 효과를 고려한 잡음 추정값의 비율로 생각할 수 있다. 마찬가지로 계산된 SNR은 기존의 SNR과 구별 하기 위하여 SNR_m이라고 표기 하겠다. 마지막 단계에서는 계산된 SNR_m을 사용하여 파라메터를 추출하고 추출된 파라메터를 적응 스펙트럼 차감법에 적용시킨다.

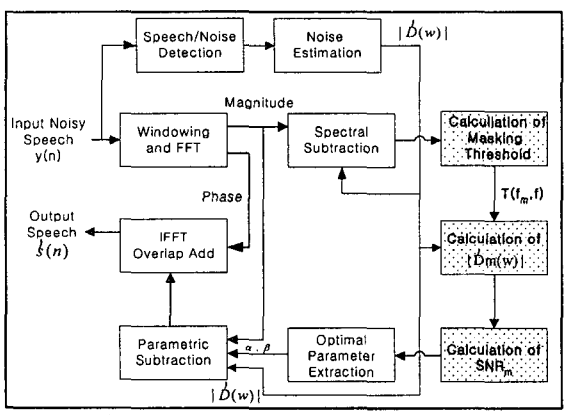

그림1. 제안된 적응 스펙트럼 차감 시스템의 구성도

3.1 $|\hat{D}m(w)|$ 계산

SNR이 높은 구간에서는 음성신호가 잡음신호 보다 크기 때문에 상대적으로 마스킹 임계치도 높아져서 대부분의 잡음이 차단되지만, 낮은 SNR 구간에서는 잡음신호의 에너지가 높아 마스킹 임계치에 의해 차단되지 않고 가청 되는 잡음 $|\hat{D}m(w)|$이 존재하게 된다. $|\hat{D}m(w)|$는 식(3)와 식(4)에 의해 계산되어 진다.

For each frame i

$$if(P_{T_i(w)} \geq P_{\hat{D}_i(w)}) \ then \quad P_{\hat{D}m_i(w)} = 0 \tag{3}$$

$$if(P_{T_i(w)} < P_{\hat{D}_i(w)}) \ then \quad P_{\hat{D}m_i(w)} = P_{\hat{D}_i(w)} - P_{T_i(w)} \tag{4}$$

Next i

i는 프레임 인덱스에 해당하며, $P_{T_i(w)}$는 마스킹 임계치의 전력 스펙트럼, $P_{\hat{D}_i(w)}$는 잡음 추정값의 전력 스펙트럼, $P_{\hat{D}m_i(w)}$는 마스킹 되지 않은 잡음 추정값의 전력 스펙트럼이다. 계산된 마스킹 임계치의 전력값과 잡음 전력값을 비교하여 마스킹 임계치가 크거나 같다면 해당 주파수에서의 잡음은 음성에 의해 마스킹 되어 가청 되지 않는다. 그러므로, 마스킹 되지 않은 잡음 전력 스펙트럼의 계산은 식(3)과 같으며 반대로, 마스킹 임계치가 추정된 잡음 보다 작아 그 주파수에서 잡음이 가청 될 경우 식(4)와 같이 잡음 전력 스펙트럼에서 마스킹 임계치의 전력값을 차감 하게 되고, 그 값이 해당 주파수에서 마스킹 효과를 고려한 잡음 추정값의 전력 스펙트럼이 된다.

3.2 SNR_m 계산

인간이 청감으로 느끼는 음의 크기와 물리적으로 측정한 음의 크기는 다르다. 기존의 SNR이 물리적으로 측정한 신호와 잡음의 비율이라면, 본 논문에서 제안한 SNR_m은 식(5)과 같이 인간의 청각 심리적 특성인 마스킹 효과를 고려한 가청 되는 신호와 가청 되는 잡음의 비율로 생각할 수 있다. 식(5)의 분자 항은 식(6)과 같이 가청 되는 음성과 가청 되는 잡음으로 세분 할 수 있고, 가청 되는 음성은 잡음이 없는 깨끗한 음성신호와 잡음에 의해 마스킹 되는 음성신호의 차이로 생각 할 수 있다. 그러나, 음성구간에서 잡음에 의해 마스킹 되는 음성은 무시할 수 있으므로 식(6)의 분자 항은 식(7)과 같이 쓸 수 있다. "clean speech power" 항목은 잡음이 추가된 음성신호의 전력 스펙트럼에서 잡음 추정값의 전력 스펙트럼을 차감하여 구할 수 있고, "audible noise power" 항목은 앞서 계산한 마스킹 효과를 고려한 잡음 추정값의 전력 스펙트럼에 해당하므로 식(7)은 식(8)과 같이 된다. 마지막으로, 식(8)의 분자 항을 정리하면 마스킹 효과를 고려한 SNR_m은 식(9)을 통해 얻을 수 있다.

$$SNR_m = 10 \cdot \log\left(\frac{audible\ noisy\ speech\ power}{audible\ noise\ power}\right) \tag{5}$$

$$= 10 \cdot \log\left(\frac{audible\ speech\ power + audible\ noise\ power}{audible\ noise\ power}\right) \tag{6}$$

$$= 10 \cdot \log\left(\frac{clean\ speech\ power + audible\ noise\ power}{audible\ noise\ power}\right) \tag{7}$$

$$= 10 \cdot \log\left(\frac{|Y(w)|^2 - |\hat{D}(w)|^2 + |\hat{D}m(w)|^2}{|\hat{D}m(w)|^2}\right) \tag{8}$$

$$= 10 \cdot \log\left(\frac{|\hat{S}(w)|^2}{|\hat{D}m(w)|^2} + 1\right) \tag{9}$$

마스킹 임계치가 충분히 커서 잡음이 모두 마스킹 될 경우를 고려하면 SNR_m은 아래 식(10)을 통해 계산 되어진다. 이때, SNR_{max}값은 20dB이다.

$$if\ |\hat{D}m(w)|^2 = 0\ then \quad SNR_m = SNR_{max}$$
$$if\ |\hat{D}m(w)|^2 > 0\ then \quad SNR_m = 10 \cdot \log\left(\frac{|\hat{S}(w)|^2}{|\hat{D}m(w)|^2} + 1\right) \tag{10}$$

3.3 파라메터 추출과 적응 스펙트럼 차감법에 적용

식(10)을 통해 계산된 SNR_m를 사용하여 파라메터 α를 계산하는 방법은 식(11)과 같다

$$\alpha = \begin{cases} 5 & , \quad if\ SNR_m < -5dB \\ \alpha_0 - \frac{3}{20} SNR_m & , \quad if\ -5dB \le SNR_m \le 20dB \\ 1 & , \quad if\ SNR_m > 20dB \end{cases} \qquad (11)$$

α_0는 SNR_m = 0dB에서의 α값(α_0=4)이다. β 값은 기존의 SNR 방법을 사용하여 낮은 SNR_m에서는 0.02와 0.06사이의 값을, 높은 SNR_m에서는 0.005와 0.02의 사이의 값으로 계산하였다[5]. 마지막으로 추출된 파라메터를 식(1)에 대입하면 제안한 적응 스펙트럼 차감을 수행할 수 있다.

4. 실험 및 평가

본 장에서는 먼저, 제안된 적응 스펙트럼 차감법을 사용하여 계산한 SNR_m과 파라메터 α값을 기존의 방법과 비교해 본다. 다음으로 다양한 잡음 환경에서 녹음한 컴퓨터 활용단어 50개를 대상으로 본 논문에서 제안한 방법과 기존의 적응 스펙트럼 차감법을 수행한 결과를 고립단어 음성인식 시스템에 적용 하여 인식률 값을 비교 하였다. 사용한 실험 데이터는 실험실과 사무실, 공장지대, 움직이는 차 안에서 각각 녹음한 컴퓨터 활용단어 50개이며, 이 데이터는16kHz로 샘플링 되고 16bit로 A/D변환 하였으며, 창 함수로는 Hanning Window를 사용하였다. 그림 2은 기존의 SNR과 제안한 SNR_m과의 비교 그래프이다. Y축은 향상된 SNR을 알아보기 쉽게 수치화 한 SNR 이득값(G_{SNR})이고 식(12)을 통해 계산되어 진다.

$$G_{SNR}(\%) = \frac{SNR_m - SNR}{SNR} \times 100 \qquad (12)$$

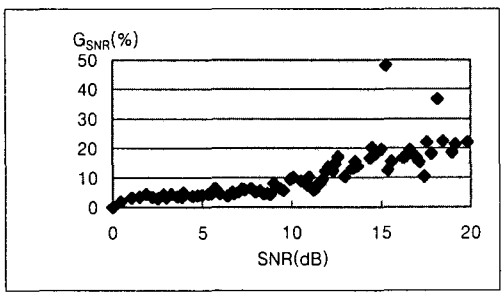

그림 2. 기존의 SNR에 대해 제안한 SNR_m의 증가율

그림 2에서 보는 바와 같이 0~20dB의 SNR영역에서 G_{SNR}은 5%이상의 값을 보이고 있다. 그림 3은 기존의 SNR을 이용하여 계산된 α값과 제안한 SNR_m을 이용하여 계산된 α값의 비교 그래프이다.

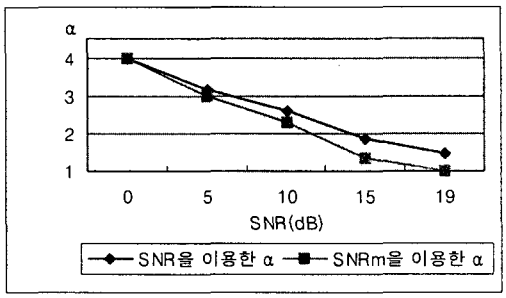

그림 3. SNR과 SNR_m을 이용하여 계산된 α값의 비교

그림 4는 사무실배경 잡음에서 "시스템종료"의 음성 파형 (a)와 이 음성 파형을 본 논문에서 제안한 적응 스펙트럼 차감법에 적용시켰을 때의 음성파형 (b)를 보여주고 있다.

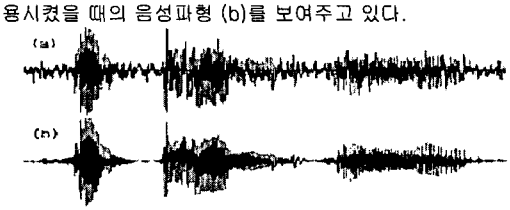

그림 4. 사무실 배경잡음에서 "시스템종료"를 제안한 적응 스펙트럼 차감법에 적용시켰을 때의 음성 파형

그림 5는 잡음이 섞인 음성신호를 제안한 방법과 기존의 방법에 각각 적용하여 잡음을 제거 하고 고립단어 음성인식 시스템에 적용했을 때의 인식률이다. 제안한 적응 스펙트럼 차감법이 잡음환경에서 기존의 방법 보다 효과적임을 알 수 있다.

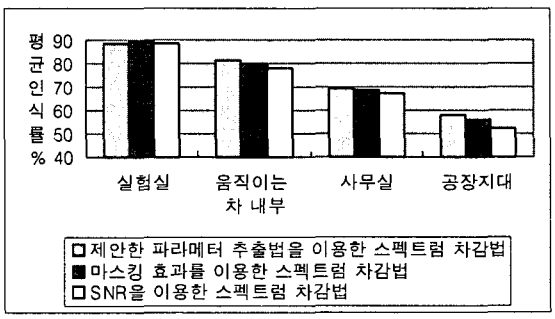

그림 5. 잡음환경에서 음성 인식기를 이용한 인식률 비교

5. 결론

본 논문에서는 기존의 적응 스펙트럼 차감법의 성능을 개선하고자 SNR과 마스킹 효과를 함께 고려한 적응 스펙트럼 차감법을 제안하였다. 제안한 방법의 성능평가를 위해 기존의 고립단어 음성인식 시스템에 적용한 결과 인식률이 향상된 것을 확인 할 수 있었다. 향후 연구과제로는 여전히 낮은 SNR에서 보다 적절한 파라메터 추출을 위한 연구가 필요하며, 고립단어만이 아닌 연속단어에서도 제안한 방법을 적용시키는 지속적인 연구가 필요하다.

Acknowledgement

본 연구는 정보통신부 대학 IT 연구센터 육성 지원사업의 연구결과로 수행되었습니다.

참고문헌

[1] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," IEEE Trans. Acoustic, Speech, Signal Processing, vol. ASSP-27, pp. 113-120, Apr. 1979.

[2] A. H. Tewfik, "Low-bit Trans parent Audio Compression using Adapted Wavelets," IEEE Transactions on Signal Processing, vol.41, No.12, pp. 3463-3479, Dec. 1993.

[3] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," IEEE Trans. on Speech and Audio Proc., vol. 7, pp. 126- 137, March 1999.

[4] T. Usagawa, "Speech Parameter Extraction in Noisy Environment Using a Masking Model," IEEE Trans. on Speech and Signal Pro., vol. ICASSP-94, pp:II/81 - II/84, April 1994.

[5] M. Berouti, "Enhancement of Speech Corrupted by Acoustic Noise," Proc. IEEE ICASSP, pp.208-211, April 1979.