論文
6-2-6

웨이브렛 변환을 이용한 음성의 적응 잡음 제거
(Adaptive Noise Reduction of Speech using Wavelet Transform)

임형규(Hyung-kyu Im)[1], 김철수(Cheol-su Kim)[2]

요 약

본 논문은 잡음 환경의 음성 인식을 위하여 음성에 부가된 잡음을 제거하는 방법으로 프레임 단위로 웨이브렛 변환을 하여 웨이브렛 계수의 표준편차를 이용하여 시간 적응 임계값을 정하는 새로운 방법을 제안한다. 음성의 특성을 고려하기 위하여 고주파 성분을 많이 가지는 무성음의 경우는 첫 번째 스케일의 detail 신호에서, 저주파 성분을 많이 가지는 유성음의 경우는 세 번째 스케일의 approximation 신호의 표준편차를 이용하여 시간 적응 임계값을 설정하였다. 또한 제안한 방법으로 잡음을 제거한 후에도 묵음구간에 잔여 잡음이 존재하게 되므로 묵음구간을 검출하여 묵음구간의 잔여 잡음을 제거하였다. 실험을 통해 제안한 방법이 일반적인 웨이브렛 변환과 웨이브렛 패킷 변환을 이용한 방법보다 SNR과 MSE측면에서 향상됨을 확인 할 수 있었다.

Abstract

This paper proposed a new time adapted threshold using the standard deviations of Wavelet coefficients after Wavelet transform by frame scale. The time adapted threshold is set up using the sum of standard deviations of Wavelet coefficient in level 3 approximation and weighted level 1 detail. Level 3 approximation coefficients represent the voiced sound with low frequency and level 1 detail coefficients represent the unvoiced sound with high frequency. After reducing noise by soft thresholding with the proposed time adapted threshold, there are still residual noises in silent interval. To reduce residual noises in silent interval, a detection algorithm of silent interval is proposed.

From simulation results, it is demonstrated that the proposed algorithm improves SNR and MSE performance more than Wavelet transform and Wavelet packet transform does.

논문접수 : 2005. 4. 11.
심사완료 : 2005. 5. 11.

1) 정회원 : 서남대학교

2) 정회원 : 서남대학교

I. 서 론

음성인식 시스템의 실용화가 늘어남에 따라 최근에는 주변 잡음에 대한 인식 시스템의 성능저하가 문제시되고 있다. 이러한 이유는 잡음이 없거나 비교적 조용한 실험실에서는 우수한 성능을 나타내는 음성인식 시스템의 성능이 잡음환경의 영향을 고려하지 않은 음성인식 시스템의 실제 환경에서는 성능이 급격하게 감소하게 된다. 따라서 잡음에 의해 오염된 음성신호에 잡음을 제거하는 기술인 음성 개선은 음성신호처리 분야에서 매우 중요하다고 볼 수 있다.

최근 들어 활발히 연구 되고 있는 웨이브렛 변환(wavelet transform)은 해석하고자 하는 주파수 성분에 따라 가변할 수 있는 창함수의 크기를 제공한다. 웨이브렛 변환에서 창함수는 기저함수(basis function)라고 불리어진다. 즉 가변하는 기저함수에 의한 웨이브렛 변환의 다중해상도 특성과 시간-주파수 국부성은 통계적 특성을 모르거나 시간적으로 예측하기 힘든 비정상적(non-stationary)인 신호해석에 매우 유용한 것으로 밝혀졌다[1][2].

본 논문에서는 웨이브렛 변환을 이용한 잡음제거 방법에서 가장 중요한 부분인 웨이브렛 계수들을 두 그룹, 즉, 잡음 성분의 영향을 많이 받는 계수들과 영향을 적게 받는 계수들로 나누는 기준인 임계값을 정하는 방법을 제안한다. 또한 제안한 방법으로 잡음을 제거하고 난 후에도 묵음구간에 잔여 잡음이 존재하게 되는데 묵음구간을 검출하기 위한 방법을 제안하며 이를 이용하여 묵음구간에 존재하는 잔여 잡음을 제거한다. 성능 평가를 위해서 SNR(Signal to Noise Ratio)과 MSE(Mean Squared Error)을 계산하여 기존의 웨이브렛 변환과 웨이브렛 패킷 변환과 비교하였다.

II. 웨이브렛을 이용한 잡음제거 방법

웨이브렛을 이용한 thresholding 방법의 기본 원리는 백색 가우시안 잡음(white gaussian noise)에 의해 오염된 신호를 웨이브렛 변환했을 때 각 스케일에 포함된 잡음 성분은 신호 성분의 크기보다 상대적으로 작은 값을 가지므로 적절한 임계값(λ) 이하의 값을 제거한 후 다시 합성함으로써 효과적으로 잡음을 제거할 수 있다는 것이다[3][4].

임계값 λ는 웨이브렛 변환(WT)일 경우 식(2)를 사용하며, 웨이브렛 패킷 변환(WPT)일 경우 식(3)을 사용한다.

$$T_{soft}(X) = \begin{cases} sgn(X)(|X|-\lambda) & ,|X|\geq\lambda \\ 0 & ,|X|<\lambda \end{cases} \quad (1)$$

$$\lambda = \sigma\sqrt{2\log(N)}$$

$$\lambda = \sigma\sqrt{2\log(N\log_2 N)}$$

N 은 신호의 샘플수이고, σ 는 선택되어진 웨이브렛 계수의 표준편차이다. 표준편차는 잡음이 포함된 웨이브렛 계수의 중간값을 이용하여 식(4)와 같이 계산한다.

$$\sigma = median/0.6745$$

III. 시간 적응 임계값

웨이브렛 변환으로 얻어지는 웨이브렛 계수가 음성 신호 구간과 잡음 신호 구간에서 다른 특성을 가진다는 것을 이용하였다[5].

3.1 시간 적응 임계값을 이용한 잡음 제거

본 논문에서는 주어진 음성 신호를 프레임 단위로 나누어 각각의 프레임에 대하여 웨이브렛 변환한 후 cD1과 cA3를 이용하면 잡음 환

경 하에서도 음성 구간을 검출할 수 있음을 이용하여 시간 적응 임계값을 생성한다. 이는 음성신호에 존재하는 파열음, 마찰음 및 파찰음의 경우 신호의 에너지는 유성음구간에 비해 상대적으로 작지만 주파수 영역에서 고주파 부분에 많은 에너지를 가지게 되며, 유성음 구간의 경우 저주파 부분에 많은 에너지를 가지게 됨을 이용한 것이다. cD1의 프레임 단위 표준편차는 S_{D1}^k로 표시하고 cA3의 표준편차는 S_{A3}^k로 표시한다.

본 논문에서는 그림 1과 같이 S_{A3}^k의 출력 파형이 음성 파형과 비슷함을 이용하여 시간 적응 임계값으로 사용할 수 있음을 알게 되었다.

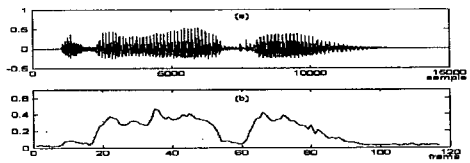

그림 1. (a)음성 파형 (b) S_{A3}^k의 파형

Fig. 1 (a) Speech signal (b) S_{A3}^k

그러나 음성파형에 존재하는 고주파 성분의 파찰음을 저주파 성분을 표시하는 S_{A3}^k에서는 제대로 나타내지 못한다. 따라서 고주파 성분인 파찰음의 정보를 가지고 있는 S_{D1}^k를 S_{A3}^k에 적절하게 더함으로써 음성파형 전부를 나타낼 수 있다.

이러한 성질을 이용하여 웨이브렛 영역에서 음성검출을 위한 파라미터를 식(5)와 같이 정의하여 사용할 수 있다[5].

$$T^k = S_{A3}^k + \alpha S_{D1}^k \qquad k = 1, \cdots, N.$$

S_{A3}^k은 세 번째 approximation 스케일의 표준편차이며, α는 weighting factor이고, S_{D1}^k은 첫 번째 detail 스케일의 표준편차이다. k는 프레임의 수이며, 가중치 α 값은 [5]에서와 같이 '6'의 값을 사용하여 음성 구간을 검출하였다.

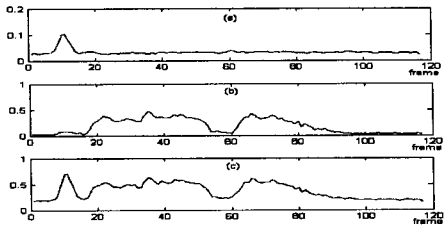

그림 2. (a) S_{D1}^k (b) S_{A3}^k (c) T^k

Fig. 2. (a) S_{D1}^k (b) S_{A3}^k (c) T^k

α 값을 '6'으로 설정한 T^k의 그래프를 살펴보면 그림 2와 같다. 그림 2에서 T^k의 값을 살펴보면 S_{D1}^k의 값을 '6'을 곱하여 S_{A3}^k에 더함으로써 기저선이 원래 0.03 정도에서 0.28 정도로 상승했음을 볼 수 있다. 즉, S_{D1}^k의 기저선의 6배가 더해져서 T^k의 기저선이 상승하였다. [5]의 논문에서와 같이 끝점 검출만을 한다면 문제가 되지 않는다. 그러나 본 논문에서는 T^k 값을 시간 적응 임계값으로 활용할 수 있도록 다음과 같은 과정을 거친다.

1) S_{D1}^k의 기저선을 제거하기 위해서 S_{D1}^k의 평균값을 구한 후, S_{D1}^k에서 이 평균값만큼 빼준다(그림 3(a)).

$$S_{D1}^{k'} = S_{D1}^k - mean(S_{D1}^k)$$

2) $S_{D1}^{k'}$의 최대값과 S_{A3}^k의 최대값과의 비율을 구한다.

$T^{k'}$의 값을 다음 식과 같이 구한다(그림 3(b)).

$$T^{k'} = S^{k}_{A3} + \beta \frac{\max(S^{k}_{A3})}{\max(S^{k'}_{D1})} S^{k'}_{D1}$$

여기에서 β 는 조정 파라미터이다.

3) 파라미터 값의 최대값이 '1' 이 되도록 정규화 과정(최대값 조정 과정)을 거친다 (그림 3(c)).

$$T^{k''} = T^{k'} \times \frac{1}{\max(T^{k'})}$$

4) 재샘플링을 통하여 시간 적응 임계값을 구한다(그림 3(d)). 즉, $T^{k''}$ 는 원 신호의 샘플수가 아닌 프레임의 수이므로 원 신호의 샘플수로 재샘플링을 통하여 시간 적응 임계값을 결정할 수 있다.

5) 마지막으로 파라미터 $T^{k''}$ 를 이용하여 시간 적응 임계값을 식 (2)에서 구한 λ 값을 곱하여 사용한다(그림 3(e)).

$$\lambda^{k'} = \gamma\lambda(1 - T^{k''})$$

여기에서 γ 는 조정 파라미터이다.

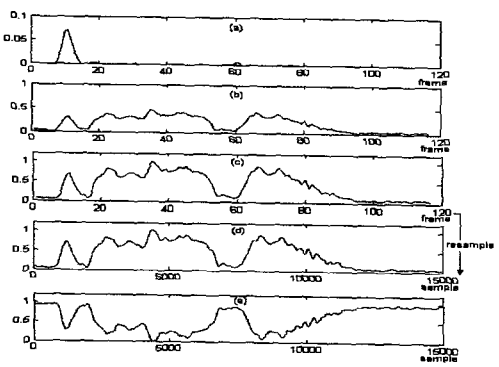

그림 3. (a)기저선 제거 (b) $T^{k'}$ (c) $T^{k''}$ (d)재샘플링 (e) $\lambda^{k'}$

Fig. 3 (a)baseline elimination (b) $T^{k'}$ (c) $T^{k''}$ (d)re-sampling (e) $\lambda^{k'}$

잡음 신호와 시간 적응 임계값의 비교과정을 거치면서 soft thresholding을 이용하여 잡음을 제거한다. 그러나 이러한 방법으로 잡음을 제거하여도 묵음구간에서의 잔여잡음이 존재하게 된다.

3.2 묵음구간 잔여잡음 제거

실험에 사용된 음성신호의 초기부분과 끝부분은 잡음 신호만이 존재한다고 가정하고 원 신호의 앞부분과 뒷부분의 일부를 이용하여 묵음구간을 검출하기 위한 파라미터로 사용한다. 본 연구에서는 음성신호의 앞부분과 뒷부분의 1500샘플을 이용하여 묵음구간 검출 임계값으로 사용하였다. 256샘플의 프레임일 경우 앞부분과 뒷부분 각각 5개의 프레임을 이용하였다. 만약 프레임 크기가 바뀔 경우 이용되는 프레임 수도 바뀌게 된다.

잡음이 섞인 음성 신호의 음성구간과 묵음구간의 특성을 살펴보면 음성구간의 경우 해당 프레임의 에너지와 표준편차는 크며, 묵음 구간의 경우 해당 프레임의 에너지와 표준편차는 작은 경향이 있다. 이를 이용하여 각 프레임에 해당하는 샘플 값들을 제곱한 후 더해서 앞에서 구한 T^{k} 값으로 곱한 값을 SF(silence frame)으로 정의하며 식(10)과 같이 쓸 수 있으며, 묵음 구간을 검출하기 위한 임계값은 λ_{SF_k} 으로 정한다. 식(10)및 식(11)은 프레임 크기가 256 샘플일 경우이다.

$$SF_k = T^{k} \times \sum_{i=1}^{256} (F^{k}_{i})^2$$

$$\lambda_{SF_k} = \left(\sum_{k=1}^{5} SF_k + \sum_{k=N-4}^{N} SF_k \right) / 10$$

여기에서 F^{k}_{i} 는 k번째 프레임의 각각의 샘플 값들을 의미한다.

SF_k 의 처음 5개의 값과 마지막 5개의 값의 평균을 묵음 구간을 검출하기 위한 임계값으로 사용한다. 만약 SF_k 의 값이 λ_{SF_k} 보

다 작으면 묵음 구간으로 간주하고 잡음이 제거된 신호에서 이 프레임에 해당하는 샘플 값들은 '0' 으로 만들어 묵음 구간의 잔여 잡음을 제거하며, SF_k 의 값이 λ_{SF_k} 보다 크면 음성구간으로 간주하며 원래의 샘플 값을 변경시키지 않는다. 즉, 식(10)및 식(11)과 같은 간단한 식을 이용하여 음성의 시작점과 끝점을 검출하는 끝점 검출을 수행할 수 있다.

Ⅳ. 실험 방법 및 결과

4.1 실험 방법

실험에 사용된 음성은 16kHz 샘플링되고 16bit로 양자화된 "청와대" 와 "1월 15일 까지 할 수 있는 데요 "를 발음한 음성신호를 사용하였다. 잡음음성을 위해서는 원음성에 white gaussian noise를 첨가하여 0dB에서 20dB 까지의 잡음신호를 만들어 실험을 수행하였다. 또한 잡음 제거의 객관적인 성능 평가를 위하여 SNR과 MSE를 계산한다[6]. 이를 수식으로 나타내면 다음과 같다.

$$SNR(\delta) = 10 \log \frac{\sum_i \hat{s}_i^2}{\sum_i n_i^2} \quad (12)$$

$$MSE(\delta) = \frac{\sum_{i=0}^{N} (s_i - \hat{s}_i)^2}{N} \quad (13)$$

여기에서 s, n, $\hat{s}$ 는 각각 원 신호, 잡음, 원 신호의 추정값을 나타내며, N은 샘플수이다.

4.2 실험 결과

음성 특징 벡터의 추출에 필요한 계산량에 영향을 줄 뿐만 아니라 음성 특징 벡터의 성질에도 영향을 주는 프레임 크기와 주파수 대역 분할을 수행하는 웨이브렛 모함수를 결정한다.

프레임 크기에 따른 성능을 평가하기 위해서 "청와대" 음성샘플에 white gaussian noise 10dB를 첨가하여 잡음신호를 만들어 실험을 수행하였다. 웨이브렛 함수는 다우베치(Daubechies)함수 8tap을 사용하였으며 분석 프레임의 1/2을 overlap 하였다.

표 1의 결과를 보면 256 샘플의 경우가 다른 프레임 크기의 경우보다 SNR에서는 2%~10%, MSE에서는 6%~30%의 성능 차이가 있어서 256 샘플의 경우가 제안한 방법에 적합함을 알 수 있다.

표 1. 프레임 크기에 따른 SNR 과 MSE의 평가
Talbe 1. SNR and MSE with different frame size.

프레임 크기	SNR	MSE
100	11.018	0.627
128	11.923	0.509
150	11.144	0.609
200	11.738	0.531
256	**12.179**	**0.480**
300	11.558	0.554

최적의 웨이브렛 모함수를 찾기 위하여 기존 논문에서 음성의 잡음 제거를 위하여 사용한 웨이브렛 모함수들을 통하여 제안한 알고리즘에 맞는 웨이브렛 모함수를 찾도록 한다[6]. 실험은 "청와대" 음성샘플에 white gaussian noise 10dB를 첨가하여 잡음신호를 만들어 실험을 수행하였으며 분석 프레임은 앞의 실험에서 우수한 성능을 보인 256샘플로서 128샘플을 overlap 하였다.

표 2. 웨이브렛 모함수에 따른 SNR 과 MSE의 평가
Table 2. SNR and MSE with different Wavelet.

모함수	SNR	MSE
db32	12.232	0.474
db8	12.179	0.480
db4	12.083	0.490
sym8	11.872	0.515
sym4	12.215	0.476
sym3	**12.240**	**0.473**
coif5	11.841	0.519
coif3	11.995	0.500

표 2의 SNR과 MSE의 비교에서 웨이브렛 변환에서는 symlets 8tap 함수가, 웨이브렛 패킷 변환 에서는 다우베치 32tap 함수가 적합하지만, 제안한 방법에서는 symlets 3tap 함수가 다른 웨이브렛 모함수보다 더 좋은 성능을 보여주었다.

위의 두 실험에서 우수하게 나타난 256 샘플의 분석 프레임과 symlets 3tap 모함수를 이용하여 white gaussian noise을 5dB 단위로 0dB에서 20dB까지 섞인 음성 샘플을 생성하여 실험하였다. 다음은 각각의 음성에 대한 실험 결과이다.

4.2.1 "청와대 " 음성의 실험 결과

그림 4와 같이 0dB white gaussian noise을 첨가하였을 경우 잡음 제거 결과는 WT(Wavelet Transfrom)과 WPT(Wavelet Packet Trasnform)을 비교하였을 때 SNR과 MSE 모두 약간의 성능 향상이 있었다.

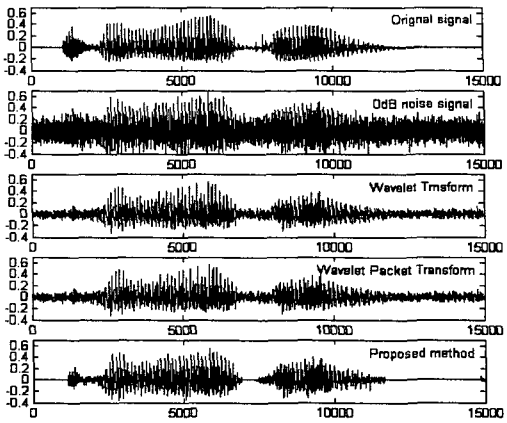

그림 4. 0dB white gaussian noise를 첨가한 "청와대" 음성의 잡음 제거
Fig. 4. Noise reduction of "청와대" signal with 0dB white gaussian noise.

표 3과 표 4는 제안한 알고리즘이 잡음 dB가 증가할수록 기존의 WT와 WPT보다 성능 개선 정도가 우수함을 보여주고 있다.

표 3 "청와대" 음성의 SNR 비교
Table 3. SNR comparison of "청와대" signal.

잡음 크기	WT	WPT	Proposed
0dB	5.885	5.819	6.912
5dB	8.177	7.640	9.671
10dB	9.987	8.358	12.240
15dB	11.643	8.592	14.517
20dB	13.079	8.682	16.308

표 4 "청와대" 음성의 MSE 비교
(단위 : 1.0e-003)
Table. 4 MSE comparison of "청와대" signal.

잡음 크기	WT	WPT	Proposed
0dB	2.043	2.074	1.613
5dB	1.206	1.364	0.854
10dB	0.794	1.156	0.473
15dB	0.543	1.095	0.280
20dB	0.390	1.073	0.185

4.2.2 "1월 15일 까지 할 수 있는 데요 " 음성의 실험 결과

그림 5과 같이 10dB white gaussian noise을 첨가하였을 때, 제안한 알고리즘의 경우 WT와 WPT의 경우와 비교하여 다소 양호하게 복원함을 알 수 있다.

표 5와 표 6은 음성 "1월 15일 까지 할 수 있는 데요 " 에 대한 잡음 크기에 따른 SNR과 MSE 값을 보여준다. 0dB의 경우 WT 및 WPT와 약간의 차이가 있었으나 잡음 dB가 증가할수록 제안한 알고리즘의 성능이 우수함을 보여준다.

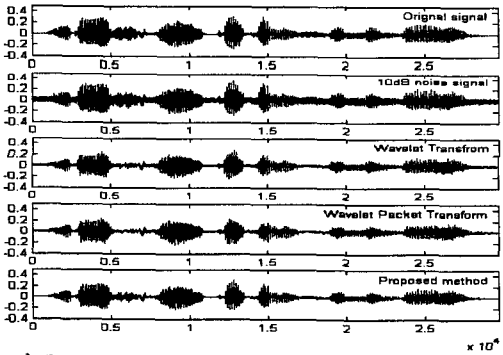

그림 5. 10dB white gaussian noise를 첨가한 "1월 15일 까지 할 수 있는 데요" 음성의 잡음 제거

Fig 5. Noise reduction of "1월 15일 까지 할 수 있는 데요" signal with 10dB white gaussian noise.

표 5 "1월 15일 까지 할 수 있는 데요" 음성의 SNR
Table 5. SNR of "1월 15일 까지 할 수 있는 데요"

잡음 크기	WT	WPT	Proposed
0dB	5.197	5.071	6.493
5dB	7.053	6.486	8.807
10dB	8.565	7.044	10.985
15dB	10.091	7.220	13.099
20dB	11.517	7.281	14.935

표 6 "1월 15일 까지 할 수 있는 데요" 음성의 MSE (단위 : 1.0e-003)
Table 6. MSE of "1월 15일 까지 할 수 있는 데요"

잡음 크기	WT	WPT	Proposed
0dB	0.715	0.736	0.531
5dB	0.466	0.531	0.311
10dB	0.329	0.467	0.188
15dB	0.231	0.449	0.116
20dB	0.167	0.442	0.076

V. 결 론

본 논문은 잡음 환경의 음성 인식을 위하여 음성에 부가된 잡음을 제거하는 방법으로 프레임 단위로 웨이브렛 변환을 수행한 후 웨이브렛 계수의 표준편차를 이용하여 시간 적응 임계값을 정하는 새로운 방법을 제안하였다. 음성의 특성을 고려하여 고주파 성분을 많이 가지는 무성음의 경우는 첫 번째 스케일의 detail 신호에서, 저주파성분을 많이 가지는 유성음의 경우는 세 번째 스케일의 approximation 신호의 표준 편차를 구하여 시간 적응 임계값을 설정하였다. 또한 제안한 방법으로 잡음을 제거하고 난 후에도 묵음구간에 잔여 잡음이 존재하게 되므로 묵음구간을 검출하여 묵음구간의 잔여 잡음을 제거하였다.

원음성에 white gaussian noise를 5dB단위로 0dB에서 20dB 까지 첨가하여 잡음을 제거한 실험에서 프레임 크기와 웨이브렛 모함수의 변화에 따른 성능 평가를 통해 256 샘플과 sym 3tap 함수가 제안한 방법에 적합함을 증명하였다. 단어와 문장의 음성 샘플 잡음제거 실험을 통하여 일반적인 웨이브렛 변환과 웨이브렛 패킷 변환을 이용한 방법보다 SNR과 MSE가 향상됨으로써 시간 적응 임계값 방법이 효과적으로 잡음을 제거할 수 있음을 증명하였다.

참고문헌

[1] Daubechies, I., "The Wavelet transform, time- frequency localization and signal analysis," *IEEE Trans. on information theory*, vol. 36, no. 5, pp. 961-1005, 1990.

[2] Michel Misiti, Yves Misiti, Georges Oppenheim, Jean-Michel Poggi, *Wavelet Toolbox for Use with MATLAB*, The Math Work Inc, 2001.

[3] Bahoura, M., Fouat, J., "Wavelet speech enhance- ment based on the teager energy operator," *IEEE Signal Process.* Lett. 8(1), pp. 10-12, 2001.

[4] Chang, S., Kwon, Y., Yang S., Kim I., "Speech Enhancement for Non-stationary Noise Environment by adaptive Wavelet Packet," *IEEE Trans. ASSP,* vol. 1, pp. I-561-I-564, 2002.

[5] 석종원, "Wavelet Transform-Based Speech Signal Processing: Speech Enhancement and Endpoint Detection", 박사학위논문, 경북대학교, 1999.

[6] 김현기, 이상운, 홍재근, "이산 웨이브렛 변환영역에서의 스펙트럼 차감법을 이용한 잡음제거" , 멀티미디어학회 논문지, 제4권, 제4호, pp. 306-315, 2001.