한국음향학회지 제29권 제7호 pp. 468~474 (2010)

잡음환경에서 음성인식 성능향상을 위한 바이너리 마스크를 이용한 스펙트럼 향상 방법

Method for Spectral Enhancement by Binary Mask for Speech Recognition Enhancement Under Noise Environment

최 갑 근*, 김 순 협*
(Gab-Keun Choi, Soon-Hyob Kim)

*광운대학교 대학원 컴퓨터공학과
(접수일자: 2010년 8월 10일; 수정일자: 2010년 8월 30일; 채택일자: 2010년 9월 7일)

음성인식의 실용화에 가장 저해되는 요소는 배경잡음과 채널잡음에 의한 왜곡이다. 일반적으로 배경잡음은 음성인식 시스템의 성능을 저하시키고 이로 인해 사용 장소의 제약을 받게 한다. DSR (Distributed Speech Recognition) 기반의 음성인식 역시 이와 같은 문제로 성능 향상에 어려움을 겪고 있다. 이러한 문제를 해결하기 위해 다양한 잡음제거 알고리듬이 사용되고 있으나 낮은 SNR환경에서 부정확한 잡음추정으로 발생하는 스펙트럼 손상과 잔존 잡음은 음성인식기의 인식환경과 학습환경의 불일치를 만들게 되어 인식률을 저하시키는 원인이 된다. 본 논문에서는 이와 같은 문제를 해결하기 위해 잡음제거 알고리듬으로 MMSE-STSA 방법을 사용하였고 손상된 스펙트럼을 보상하기 위해 Ideal Binary Mask를 이용하였다. 잡음환경 (SNR 15 ~ 0 dB)에 따른 실험결과 제안된 방법을 사용했을 때 향상된 스펙트럼을 얻을 수 있었고 향상된 인식성능을 확인했다.

핵심용어: 스펙트럼향상, 음성인식
투고분야: 음성처리 분야 (2)

The major factor that disturbs practical use of speech recognition is distortion by the ambient and channel noises. Generally, the ambient noise drops the performance and restricts places to use. DSR (Distributed Speech Recognition) based speech recognition also has this problem. Various noise cancelling algorithms are applied to solve this problem, but loss of spectrum and remaining noise by incorrect noise estimation at low SNR environments cause drop of recognition rate. This paper proposes methods for speech enhancement. This method uses MMSE-STSA for noise cancelling and ideal binary mask to compensate damaged spectrum. According to experiments at noisy environment (SNR 15 dB ~ 0 dB), the proposed methods showed better spectral results and recognition performance.

Keywords: Spectrum Enhancement, Noisy Speech Recognition
ASK subject classification: Speech Signal Processing (2)

I. 서 론

정보통신산업의 비약적인 발전은 컴퓨팅 환경의 극단적 변화를 가져왔다. 현재의 컴퓨팅환경은 불과 몇 년 전까지만 해도 음성통신이 위주였던 모바일 단말기에 대해 네트웍 기능과 다양한 응용소프트웨어를 사용할 수 있도록 성능이 강화된 스마트폰이 대중적으로 보급되기 시작하였다. 하지만 휴대가 간편하고 다양한 응용소프트웨어를 사용할 수 있는 스마트폰 역시 입력장치로 터치패드와 키패드를 사용하는 과거의 방식을 채택하고 있다. 이와 같은 입력장치의 불편함을 극복하고자 근래 들어 음성을 이용하여 웹을 검색하는 응용소프트웨어가 개발되고 상당한 성과를 거두면서 스마트폰 등에 대한 음성인식 기술의 관심이 나날이 확대되고 있다. 특히 스마트폰 등에 적용할 수 있는 음성인식 기술로 음성의 특징 추출은 단말기가 담당하고 연산량과 메모리용량을 많이 요구하는 인식부분은 원격서버가 담당하도록 하는 분산음성인식

책임저자: 최 갑 근 (cocomm@kw.ac.kr)
139-701 서울시 노원구 월계동 447-1 광운대학교 컴퓨터공학과
(전화: 02-940-5123; 팩스: 02-941-8919)

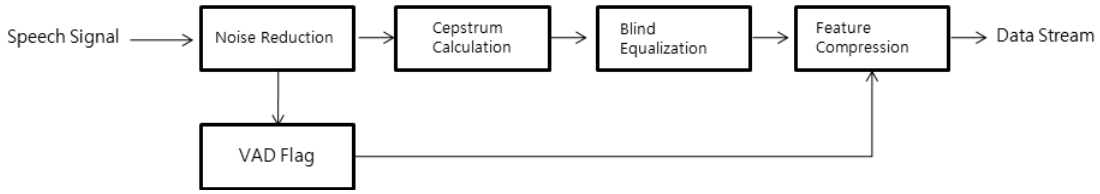

그림 1. 분산 음성인식의 AFE (Advanced Front End)
Fig. 1. Advanced Front End of Distributed Speech Recognition.

(Distributed Speech Recognition)이 연구되고 있다. 분산음성인식에서는 배경잡음을 효과적으로 제거하고 정확한 음성 특징 추출을 위한 다양한 연구가 시도 되고 있으며 이를 위해 ETSI (European Telecommunication Standard Institute)에서는 표준을 제정하고 있고 특징 추출과 관련하여 FE (Front End)가 제정된 이후 잡음제거와 모델 보상 기법 등을 보강한 AFE (Advanced Front End)가 제정되었고 기존의 방법들과 비교하여 우수한 성능을 보이며 현재까지 가장 높은 인식률을 보이고 있으나 불안정하게 변화하는 배경잡음으로 인한 인식성능 저하는 크게 개선되지 못하고 있다. 그림 1은 ETSI에서 제정한 AFE (Advanced Front End) 표준안이다 [1, 2].

음성인식 성능의 가장 큰 저해요소는 배경잡음으로 인한 인식환경과 훈련환경의 불일치가 가장 크다. 일반적으로 음성인식을 위한 훈련용 음성은 조용한 환경에서 수집된 음성을 사용한다. 하지만 실생활에서 발생하는 다양한 배경잡음은 인식환경이 훈련환경과 차이를 보이게 하는 가장 큰 요소이다. 분산음성인식 역시 사용 환경에 따라 발생하는 배경잡음의 처리는 인식률 개선에 중요한 요소가 된다. 이와 같이 배경잡음에 의한 인식률 저하를 막기 위해 잡음을 제거하는 다양한 방법들이 소개되고 있는데 음성향상기술이 그 대표적인 기술이다 [3].

음성향상 알고리듬 중에서 일반적으로 널리 사용되고 있는 알고리듬은 Wiener 필터 방식이며, 음성신호의 스펙트럼에 대한 MMSE (Minimum Mean Square Error) 추정 기반의 필터가 사용된다. 이보다 개선된 것은 음성신호와 잡음신호의 스펙트럼에 대한 사전 확률분포를 가정하고 통계모델에 근거하여 음성신호의 스펙트럼의 크기를 추정하는 MMSE-STSA (Short Time Spectral Amplitude) 방법이 좋은 성능을 보이나 음성통신에서 음성품질향상을 목표로 연구된 알고리듬으로 인식상황에 적합하지 않을 수 있으며 특히 잡음제거 후 음악잡음과 잔존잡음이 남아 인식률을 저하시키게 되어 음성인식을 위해서는 적절한 알고리듬의 선택이 필요하다 [3, 4].

본 논문에서는 잡음제거 후 발생하는 잔존잡음과 음악잡음, 스펙트럼 왜곡 등을 보상하기 위해 Hu, Wang이 소개한 CASA (Computational Auditory Scene Analysis) 기반 IBM (Ideal Bit Mask)를 이용하였다. IBM은 시간 주파수 공간에서 약한 신호는 강한신호에 마스킹 된다고 보고 마스킹 임계값을 정한 후 에 목표한 신호는 1을 주고 그 외의 간섭신호에 대해 0을 주어 마스킹 공간을 만든다. 본 논문에서는 이와 같은 마스킹공간을 잡음이 제거된 음성향상 신호에 곱해주어 불필요한 잔존잡음과 스펙트럼 왜곡을 최소화 시킬 수 있도록 스펙트럼을 보상하였다 [5-7].

II. 음성향상

분산음성인식을 위한 AFE의 설계에서 가장 중요한 것은 인식환경과 학습 환경의 불일치를 줄일 수 있도록 하는 것이다. 이와 관련하여 최근 들어 음성향상을 이용한 잡음제거방식이 많이 연구되고 있다 [3].

음성향상 방법은 다양하게 소개 되었으나 연산 및 잡음제거 성능이 비교적 우수한 것으로 알려진 통계적 모델기반의 MMSE (Minimum Mean Square Error)는 음성신호 크기 스펙트럼을 추정하는 방식이 매우 효과적인 것으로 알려져 있다. MMSE-STSA (Minimum Mean Square Error Short-Time Spectral Amplitude) 예측기는 음성과 잡음에 대해 통계적으로 독립가우시안 확률변수 요소로 모델링한 것에 기반 한다. 향상된 음성은 잡음 신호의 원 위상을 혼합한 MMSE-STSA 예측기를 사용하여 구성된다. 주파수 영역과 FFT를 이용한 신호스펙트럼 예측으로 분석된다 [4].

잡음신호 $x(t)$에서 k번째 스펙트럴 요소의 MMSE (Minimum Mean Square Error) 진폭 예측기는 다음에 의해 주어진다.

$$\widehat{A}_k = G_k R_k \tag{1}$$

여기서 R_k는 $x(t)$에서 k번째 스펙트럴 요소의 진폭이다. 그리고 다음과 같이 주어진 G_k는

$$G_k = \frac{\sqrt{\pi}}{2} \cdot \frac{\sqrt{V_k}}{SNR_{post_k}} \cdot M[V_k] \quad (2)$$

수식 (2)에서 V_k는 다음과 같이 계산되어진다.

$$V_k = \left(\frac{SNR_{prio_k}}{1+SNR_{prio_k}} \right) \cdot SNR_{post_k} \quad (3)$$

여기서 SNR_{prio_k}와 SNR_{post_k}는 각각 사전과 사후 SNR이 된다. 수식 (2)에서 함수 M[ ]은 다음과 같이 평가된다.

$$M[\theta] = \exp(\frac{-\theta}{2})[(1+\theta)I_0(\frac{\theta}{2}) + \theta I_1(\frac{\theta}{2})] \quad (4)$$

식 (4)에서 I_0와 I_1은 각각 0과 1차의 수정된 Bessel 함수를 표현한다.

n번째 분석 프레임에서 k번째 스펙트럴 요소에 대한 사전 SNR은 다음과 같이 정의된다.

$$SNR_{prio_k}(n) = \alpha \left(\frac{\widehat{A_k^2}(n-1)}{\widehat{\lambda_k^2}(n-1)} \right) + (1-\alpha) P[SNR_{post_k}(n) - 1] \quad (5)$$

여기서 $0 \le \alpha \le 1, \lambda_k$는 잡음의 k번째 스펙트럼 요소의 분산이고 P[ ]는 반파 정류 연산자이며 다음과 같이 정의된다.

$$P[x] = \begin{cases} x & \text{if } x \ge 0 \\ 0 & otherwise \end{cases} \quad (6)$$

사후 SNR, SNR_{post_k}는 R_k^2와 $\lambda_d(k)$를 사용하여 계산되어지는데 R_k^2는 k번째 스펙트럴 요소의 제곱진폭이며, $\lambda_d(k)$는 k번째 스펙트럴 요소의 분산이다 [4].

$$SNR_{post_k} = \frac{R_k^2}{\lambda_d(k)} \quad (7)$$

III. IBM (Ideal Binary Mask)을 이용한 스펙트럼 향상방법

CASA (Computational Auditory Scene Analysis)의 연산 목표는 IBM (Ideal Binary Mask)를 얻기 위한 것으로 시간주파수 영역에서 강한에너지가 약한 에너지를 마스킹 한다고 보고 그것을 구분하기 위해 국부기준 (LC : *Local Criteria*)을 평가해 기준보다 약한 에너지를 간섭신호 및 잡음신호로 분리한다. 간섭신호 및 잡음신호에 대해서는 0을 주어 구분하고 국부기준보다 큰 에너지에 대해서는 원하는 신호로 보며 1을 준다. 시간 주파수 표현은 인간의 와우각에 기초한 데이터 표현방법으로 Cocktail Party 효과와 같이 사람의 청각시스템이 우수한 잡음제거 효과를 거두는데 초점을 두고 연구된 분야이다 [7]. 식 (8)은 IBM에 대해 정의하고 있다.

$$IBM(t,f) = \begin{cases} 1, \text{if } T(t,f) - N(t,f) > LC \\ 0, otherwise \end{cases} \quad (8)$$

여기서 $T(t,f)$는 DFT (Descrete Fourior Transform)를 실시한 후 잡음이 없는 이상적인 음성 신호 주파수 스펙트럼 크기이며, $N(t,f)$는 잡음 시간 주파수 스펙트럼 크기이다. LC (*Local Criteria*)는 국부 기준 값이 된다. LC는 일반적으로 SNR 0 dB를 사용하는 경우가 많다. 통계 기반 잡음 추정을 이용한 음성향상 알고리듬들은 일반적으로 각각의 프레임에서 음성의 존재여부를 계산하고 잡음을 제거한다. 따라서 프레임 내에서 변화하는 잡음의 크기를 정확하게 추정하지 못하면 잔존잡음과 음악잡음에 의해 스펙트럼을 손상시키게 되며 이 같은 문제는 비안정적 잡음신호가 부가된 경우 더욱 심각해진다. 본 논문에서는 변화하는 잡음을 최대한 제거하기 위해 긴 구간에서 식 (9)와 같이 각각의 주파수 빈에 대한 평균값을 계산한다.

$$\mu(k) = \frac{1}{L} \sum_{l=1}^{L-1} N(l,k) \quad (9)$$

식 (9)에서 L은 프레임의 개수이고, k는 주파수 빈 인덱스, l은 프레임 인덱스이다. 결국 LC는 $LC \ge \mu(k)$가 되도록 모든 주파수 빈 인덱스에 대해 값이 설정된다. 이와 같은 과정을 거쳐도 잡음의 변화가 매우 심하고 SNR이 낮은 경우에는 LC보다 큰 잡음성분이 남게 된다. 따라서 의도하지 않은 잡음 성분이 남은 상태에서 마스크를 사용하면 오히려 성능이 나빠질 수 도 있다. 이러한 문제를 해결하기 위해서 본 논문에서는 기본주파수 (F_0)와 하모닉스가 존재하는 영역에 바이너리 인덱스를 조사하여 잔존잡음과 스펙트럼 왜곡을 보정한다. 식 (10)은 기본주파수 (F_0)와 하모닉스 존재여부를 조사한다.

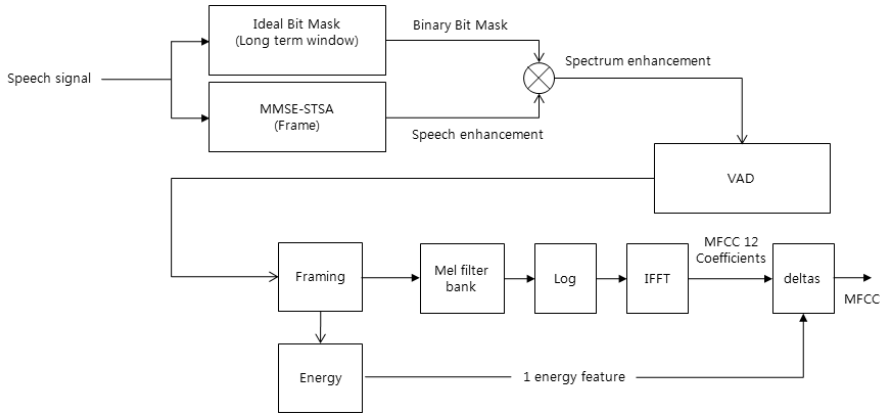

그림 2. MFCC에서 MMSE-STSA와 IBM을 이용한 스펙트럼 향상과정
Fig. 2. Spectrum enhancement process Using MMSE-STSA and IBM in MFCC.

$$F(l)=\sum_{k=1}^{N-1} IBM(l,k) \tag{10}$$

$$H(l)=\begin{cases}1\,,\ \text{if } F(l)>\delta\\ 0\,,\ otherwise\end{cases} \tag{11}$$

$$BM(l,k)=H(l)\circ IBM(l,k) \tag{12}$$

식 (10)에서 $F(l)$은 샘플링레이트가 8 kHz, FFT 크기가 256일 때 각각의 주파수 빈 인덱스의 크기는 31.25 Hz가 된다. 이에 따라 $F(l)$ 값은 DC ~ 625 Hz인 20개 정도의 주파수 빈이 존재 하는 영역이 적당한 것으로 본 연구에서 조사되었다. N의 값이 20에서 음성의 영역을 잘 구분하는 것은 음성의 주요한 성분이 분포하고 있는 주파수가 600 Hz 이하에서 발견되기 때문이다. 즉 기본주파수와 이에 따른 하모닉 스펙트럼 크기 등 모든 것이 가장 뚜렷하게 나타나는 영역으로 볼 수 있다. 여기서 δ의 값은 프레임 내에서 1을 가진 주파수 빈의 합이며 15정도가 적절한 것으로 조사되었다. 프레임에 따라서 조금씩 다른 값은 가지고 있으나 낮은 SNR에서도 음성에너지가 600 Hz이하의 스펙트럼 크기는 최대값을 가진다. 이러한 결과는 식 (11)에서 음성의 존재 여부를 평가하게 되는 기준이 되고 음성이 존재하지 않는 프레임에 0을 주게 되어 보다 명확한 구분을 할 수 있게 한다. 따라서 낮은 SNR에서도 끝점 검출성능이 비교적 우수하고 손상된 스펙트럼을 향상하기 위한 마스크를 계산하기가 용이하다. 식 (11)을 통해 얻어진 음성 존재 여부 마스크 $H(l)$는 식 (12)에서 $IBM(k,l)$과 곱해지게 되고 그 결과는 식 (13)에서 음성향상의 출력$S_{enhance}(l,k)$과 곱해지게 되는데 마스크의 값이 바이너리이기 때문에 1로 설정된 주파수 빈의 값만이 출력되게 되어 향상된 스펙트럼 $\acute{S}(l,k)$를 얻게 된다. 그림 2는 제안된 알고리듬을 수행하기 위한 절차를 보여준다.

$$\acute{S}(l,k)=S_{enhance}(l,k)\circ BM(l,k) \tag{13}$$

IV. 실험 및 성능분석

본 논문에서는 음성향상 후 제안된 IBM (Ideal Binary Mask)을 이용한 스펙트럼 향상 성능을 알아보기 위해 MMSE-STSA 알고리듬에 의해 음성 향상된 출력과 제안된 MMSE-STSA와 IBM을 함께 사용한 출력을 스펙트로그램과 바이너리 비트 마스크등으로 비교해 보았다. 사용된 잡음의 종류는 Buccaneer, destroyerengine, f16, factory, hfchannel, pink, volvo이고, 잡음조건은 SNR 15 dB ~ 0 dB를 사용하여 성능을 측정하였다. 또한 스펙트럼이 향상된 상태에서 음성인식을 실시하여 음성향상만 이용하여 인식한 결과와 비교 실험을 실시하였다. 음성인식 실험을 위해서는 HMM (Hidden Markov Model)모델을 이용하였고, 훈련을 위해서는 Baum-Welch reestimation 알고리듬을 이용하였으며 본 연구를 위해 직접 제작된 인식기를 사용하였다. 잡음 데이타 베이스는 전술한 잡음의 종류를 갖고 있는 NOISEX-92를 사용하였으며 잡음 음성은 clean 음성에 잡음을 인위적으로 부가하여 사용하였으며 특별히 라우드 스피커를 이용하여 잡음을 재생하며 인식실험을 하는 온라인 방식의 실험을 실시하였다. 온라인 실험을 위해서는 표 1에서 보여주는 주파수 응답 평탄도가 높은 모니터 스피커와 파워 앰프를 이용하였으

표 1. On-line 음성인식 실험을 위한 장비
Table 1. Equipment for online speech recognition experiment.

아이템	제조사	사양
라우드 스피커	Whafedale	70 Hz ~ 40 kHz/8o hm
파워 앰프	Rotel RB1050	70 watt/ch.

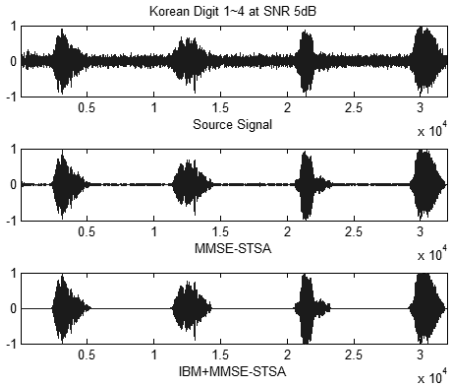

그림 3. SNR 5 dB에서 한국어 숫자음 1~4에 대한 잡음제거 결과
Fig. 3. Result of Noise reduction for Korean digit 1~4 at SNR 5 dB.

며 인식 실험을 위한 마이크는 Sony vaio note book VGN-CR35L에 내장된 콘덴서 마이크를 이용하였다 [8]. SNR의 측정을 위해 잡음과 음성을 따로 녹음하며 입력 이득 값을 설정하였고 스피커와 마이크간의 거리는 1 m로 설정하였다. 인식실험을 위한 어휘는 한국어 숫자음을 사용하였다.

그림 3에서 보여주는 바와 같이 잡음이 섞인 음성에 대해 MMSE-STSA 알고리듬을 이용하여 음성향상을 하게 되면 비교적 좋은 성능을 보이기는 하지만 잔존잡음이 많이 남아 있는 것을 확인 할 수 있다. 잔존잡음은 음성의 특징을 추출할 때 왜곡된 특징이 추출되는 원인이 된다. 따라서 음성향상을 하고 난 후에 계산된 IBM을 이용하여 바이너리 마스크 연산을 하게 되면 이와 같은 잔존잡음을 많이 줄일 수 있게 된다. 여기서 IBM은 충분히 긴 구간의 스펙트럼 구간 정보를 이용하여 수행하는데 프레임단위가 아니라 각각의 k번째 스펙트럴 요소들에 대한 전체 스펙트럴 에너지 기댓값을 사용하여 마스크의 LC값을 설정하고 식 (12)에서 음성의 가장 큰 에너지가 존재하는 제 1 포만트 지역의 바이너리 값을 조사하여 보조적으로 마스크를 보완하여 잔존잡음을 제거한다.

그림 3에서 IBM+MMSE-STSA를 보면 MMSE-STSA가 제거 하지 못한 잔존잡음을 최대한 제거하는 것을 확인 할 수 있다. 낮은 SNR상황에서 잡음의 종류가 무색잡음이고 안정적 잡음이면 잡음제거 효과는 탁월하다. 하지만 잡음의 종류가 SSN (Speech Shape Noise)의 특징을

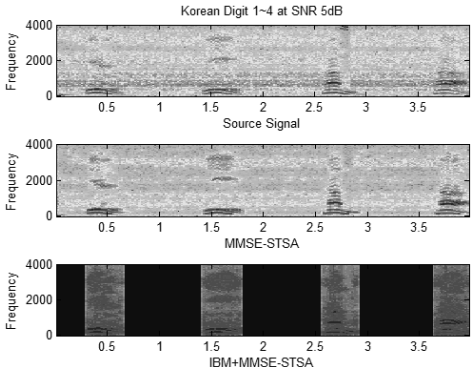

그림 4. SNR 5 dB에서 한국어 숫자음 1~4에 대한 스펙트로그램
Fig. 4. Spectrogram for Korean digit 1~4 at SNR 5 dB.

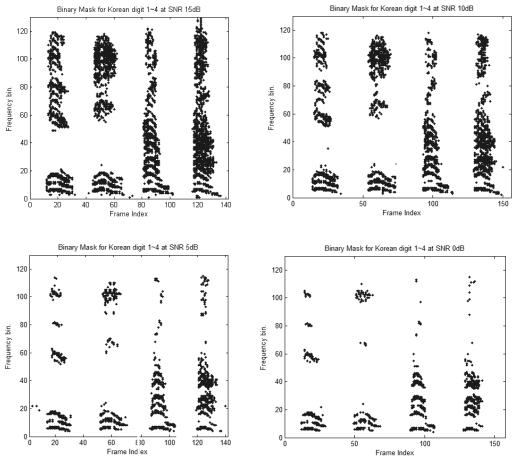

그림 5. SNR 변화에 대한 바이너리 마스크
Fig. 5. Binary Mask for SNR change.

가지게 되고 비안정적이며 낮은 SNR이면 여전히 잡음으로 인하여 마스크의 LC값 설정이 곤란하지만 단순히 음성향상만을 수행한 결과보다는 잔존잡음과 중요 스텍트럼 요소를 향상시키는 것을 확인 할 수 있다. 그림 3에서 보여지는 잡음의 종류는 f16잡음이 사용되었다.

온라인 실험을 통해 얻어진 결과는 그림 4와 같다. 실험을 위해 라우드 스피커를 이용하여 f16 잡음을 재생하고 화자가 직접 마이크에 녹음한 것을 처리한 결과이다. 그림 4에서 볼 수 있듯이 MMSE-STSA는 잔존 잡음이 남아 있다. 또한 잔존잡음은 음악잡음의 형태이고 음성인식의 끝점 검출에서 임계값 설정에 영향을 주게 된다. 한편 MMSE-STSA의 출력 결과에 바이너리 마스크를 곱해준 결과를 본다면 마스크 공간을 만들 때 음성의 중요성분이 존재 하는 1 포만트 지역을 조사하여 식 (11)의 조건에 의해 프레임의 마스크가 0으로 처리된 것을 볼 수 있다. 또한 음성이 존재하는 영역에서도 식 (9)에 의해 계산되어진 LC값의 영향으로 잔존 잡음은 억제되어있는 것을 볼 수 있다.

그림 5는 잡음조건에 따라 Binary Mask의 출력결과에

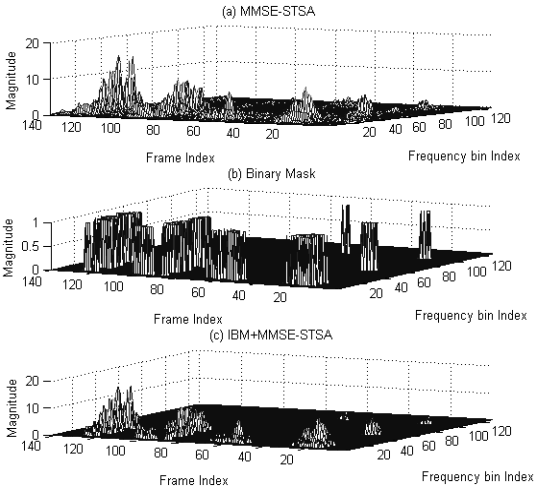

그림 6. SNR 5 dB에서 IBM+MMSE-STSA의 성능
Fig. 6. Performance of IBM+MMSE-STSA at SNR 5 dB.

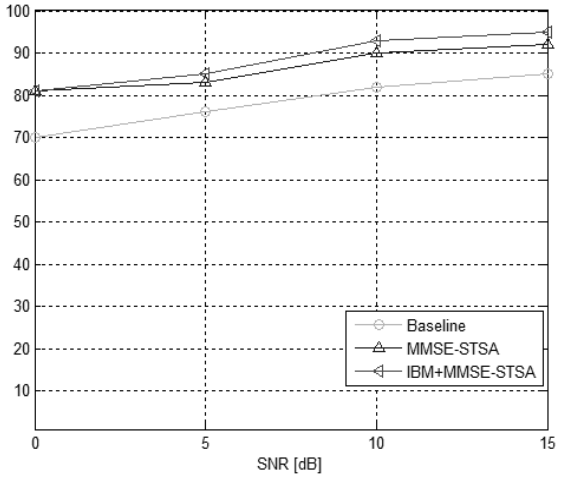

그림 7. SNR 변화에 따른 인식 성능 (Word Accuracy (%))
Fig. 7. Recognition Performance for SNR change.

보조적으로 기본주파수 F_0와 하모닉스를 카운트 하여 마스크를 보상한 결과를 보여준다. 그림 5의 위쪽 좌, 우는 SNR 15 dB와 10 dB에 대한 바이너리 마스크를 보여주며 비교적 성능이 안정적이다. 하지만 아래쪽 좌, 우는 SNR 5 dB와 SNR 0 dB에 대한 바이너리 마스크로 잡음의 크기가 커져 LC값이 커지는 결과에 따라 중요 스펙트럼 정보가 많이 사라진 것을 볼 수 있고 마스크의 성능이 떨어지는 것이 확인된다. 하지만 그림 6에서 보여지는 바와 같이 MMSE-STSA 알고리듬을 수행한 음성향상 결과와 함께 사용하게 되면 그림 6의 (a)에서 보여지는 잔존 잡음들이 (c)에서 마스크의 도움으로 사라지게 되어 스펙트럼 향상에는 도움이 되는 것을 확인 할 수 있다.

그림 2에서 보여지는 바와 같이 향상된 스펙트럼은 VAD (Voice Activity Detection) 알고리듬을 실행하게 된다. 스펙트럴 에너지를 기반으로 하는 VAD 알고리듬들은 특히 잔존잡음으로 인하여 안정된 문턱값을 획득하는데 어려움이 있다. 본 논문에서 제안된 방법은 음성을 향상한 후에 잔존잡음을 최대한 제거하고 음성검출을 시도하여 안정된 문턱 값을 설정하는데 도움을 준다.

그림 6은 SNR 변화에 따른 인식률을 보여준다. 실험결과 낮은 SNR과 비안정적 잡음에서는 여전히 음성향상의 결과만으로 인식실험을 했을 때와 별 차이가 없다. 하지만 SNR이 비교적 높은 구간에서는 IBM의 LC설정이 비교적 좋은 개선 성능을 보여준다.

V. 결 론

본 논문은 잡음환경에서 음성인식 시스템의 성능을 보완하기 위해 음성의 특징추출을 하는 전처리 부분에 잡음을 제거하기 위한 음성향상 알고리듬을 이용하여 인식성능을 개선하는 연구를 하였다. 음성향상을 위해 사용한 MMSE-STSA는 통계기반의 비교적 우수한 성능의 알고리듬으로 알려져 있으나 잡음을 제거한 후 발생하는 잔존잡음과 음악잡음 등은 잘못된 음성의 검출과 왜곡된 특징을 추출하게 되는 이유가 된다. 본 논문에서는 성능향상을 위해 음성향상을 한 후에 IBM (Ideal Binary Mask)를 이용하여 잔존잡음과 음악잡음을 최대한 제거하였다. 하지만 비교적 높은 SNR (15 dB ~ 10 dB)에서는 우수한 성능을 갖는 IBM이 비안정적 잡음을 갖는 낮은 SNR (5 dB ~ 0 dB)에서는 잔존잡음 제거에는 효과적이나 인식성능 향상을 위해 스펙트럼을 향상 시키고 보상하는 효과는 한계가 있었다. 본 연구에 의하여 음성향상 알고리듬을 이용하여 잡음을 제거한 후에 IBM을 함께 사용하여 음성의 스펙트럼을 향상시키면 개선된 성능을 확인할 수 있으나 잡음의 종류가 SSN (Speech Shape Noise)이면서 비안정적이고 낮은 SNR에서는 인식성능 개선에 어려움이 있었다. 이것은 손상된 스펙트럼이 많아 충분히 보상되지 않은 것이 원인으로 보인다. 보다 높은 음성인식률의 향상을 위해서는 모델적응, 특징보상과 같은 연계 알고리듬을 이용하면 개선의 여지가 많을 것으로 본다.

감사의 글

본 연구는 2010학년도 광운대학교 연구년에 의하여 연구 되었습니다.

참 고 문 헌

1. ETSI standard document, *Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms*, ETSI ES 201 108 v.1.1.1 (2000-02), Feb. 2002.
2. ETSI standard document, *Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms*, ETSI ES 202 050 v.1.1.3 (2003-11), Nov. 2003.
3. R. Flynn, E jones, "Robust Distributed Speech Recognition using Speech Enhancement", *IEEE Tansactions on Consumer Electronics*, vol. 54, no. 3, pp. 1267-1273, 2008. 8.
4. Ephraim, Y., Malah, D. "Speech enhancement Using a minimum mean square error short-time spectral amplitude estimator", *IEEE Trans. Acoust., Speech Signal Process.*, vol. 32, pp. 1109-1121, 1984.
5. A. S. Bregman, *Auditory Scene Analysis. Cambridge*, MA: MIT Press, 1990.
6. N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2236-2252, 2003.
7. R. Lyon, "A computational model of filtering, detection, and compression in the cochlea," *in Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82.*, vol. 7, pp. 1282-1285, 1982.
8. A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247-251, July 1993.

저자 약력

•최 갑 근 (Gab-Keun Choi)
한국음향학회지 제28권 제5호 참조

•김 순 협 (Soon-Hyob Kim)
한국음향학회지 제27권 제7호 참조