한국음향학회지 제21권 제3호 pp. 323~329 (2002)

프레임 신뢰도 가중에 의한 강인한 음성인식

Frame Reliability Weighting for Robust Speech Recognition

조 훈 영*, 김 락 용*, 오 영 환*
(Hoon-Young Cho*, Lag-Young Kim*, Yung-Hwan Oh*)

*한국과학기술원 전산학과

(접수일자: 2002년 1월 31일; 채택일자: 2002년 2월 27일)

본 논문에서는 임의의 시점에서 발생하여 음성 신호의 일부분을 심하게 손상시키는 시간선택 잡음 (time-selective noise)을 보상하기 위한 프레임 신뢰도 가중 방법을 제안한다. 음성 프레임들은 서로 다른 정도의 신뢰도를 갖으며, 신뢰도는 프레임의 신호대잡음비 (signal-to-noise ratio)에 비례한다. 잡음이 일정한 경우에는 무음 구간에서 획득한 잡음 정보를 이용하여 프레임의 신호대잡음비 추정이 용이하나, 시간선택 잡음은 잡음 추정이 어렵다. 따라서, 본 연구에서는 프레임 신뢰도를 추정하기 위해 깨끗한 음성의 통계적 모델을 사용하였다. 제안한 MFR (model-based frame reliability) 방법은 참조 모델의 평균 벡터열과 입력 MFCC (mel-frequency cepstral coefficient) 특징 벡터열의 역변환에 의해 얻은 필터뱅크 에너지를 이용하여 프레임 신호대잡음비를 근사한다. 다양한 버스트 (burst) 잡음에 대한 인식 실험 결과, 제안한 방법은 프레임의 신뢰도를 효과적으로 나타낼 수 있었으며, 이 신뢰도를 우도 계산에서 가중치로 적용하여 인식 성능을 향상시킬 수 있었다.

핵심용어: 프레임 신뢰도, 신뢰도 가중, 프레임 SNR, 버스트 잡음, 잡음에 강한 음성인식
투고분야: 음성처리 분야 (2.5)

This paper proposes a frame reliability weighting method to compensate for a time-selective noise that occurs at random positions of speech signal contaminating certain parts of the speech signal. Speech frames have different degrees of reliability and the reliability is proportional to SNR (signal-to noise ratio). While it is feasible to estimate frame SNR by using the noise information from non-speech interval under a stationary noise situation, it is difficult to obtain noise spectrum for a time-selective noise. Therefore, we used statistical models of clean speech for the estimation of the frame reliability. The proposed MFR (model-based frame reliability) approximates frame SNR values using filterbank energy vectors that are obtained by the inverse transformation of input MFCC (mel-frequency cepstral coefficient) vectors and mean vectors of a reference model. Experiments on various burst noises revealed that the proposed method could represent the frame reliability effectively. We could improve the recognition performance by using MFR values as weighting factors at the likelihood calculation step.

Keywords: *Frame reliability, Reliability weighting, Frame SNR, Burst noise, Robust speech recognition*
ASK subject classification: *Speech signal processing (2.5)*

I. 서 론

음성인식 기술은 조용한 실험실 환경에서 높은 인식 수준에 도달하여 이 기술의 실용화를 위한 연구노력이 활발히 진행되어 왔다. 그럼에도 불구하고 실제 응용 환경에서 인식기를 사용할 경우, 응용 환경의 예측할 수 없는 잡음으로 인해 인식 성능이 급격히 저하된다. 응용 환경에 존재하는 잡음의 종류로는 크게 사용자 주변의 배경 잡음, 마이크로폰 또는 전화망 등 전송선 상에서

책임저자: 조훈영 (hycho@bulsai.kaist.ac.kr)
305-701 대전광역시 유성구 구성동 373-1
한국과학기술원 전산학과
(전화: 042-869-8720; 팩스: 042-869-3510)

발생하는 채널 잡음, 잡음의 영향으로 사용자의 발성 방식이 바뀌는 롬바드 효과 등을 들 수 있다. 이외에도 최근에 인터넷 및 이동 통신을 이용한 음성인식 응용 서비스가 증가함에 따라 단말기의 음성 부호화기에서 발생하는 정보 손실 및 유무선 전송경로 상의 패킷 손실도 성능 저하의 새로운 요인이 되고 있다.

잡음 환경에서 인식기의 성능이 떨어지는 이유는 잡음의 영향으로 인식기의 학습 환경과 사용 환경간에 불일치가 발생하기 때문이며, 이러한 불일치를 보완하기 위한 기존의 연구는 잡음에 근본적으로 강한 특징을 추출하는 방식 (robust feature extraction), 잡음 음성에서 잡음을 추정하고 제거한 뒤 특징 벡터를 추출하는 음질 개선 (speech enhancement) 및 인식 모델의 파라미터를 잡음 환경에 적응하는 모델 기반의 잡음 보상 (model-based noise compensation) 등으로 구분할 수 있다[1]. 이와 같은 기존의 연구 결과들에 의해 잡음 환경에서 일부 성능 개선이 가능하였으나, 대부분이 시간에 따라 일정한 정상적 (stationary) 잡음 혹은 시간에 따라 천천히 변하는 잡음에 대해 효과적인 방법들이다. 반면에 실생활에 존재하는 잡음은 음악소리처럼 시간에 따라 급격히 변하거나, 자동차 경적소리, 문 닫히는 소리, 키보드나 마우스 누르는 소리처럼 시간 및 주파수 영역에서 음성의 일부분을 심하게 손상시키는 잡음이 많은 비중을 차지하고 있어 이를 위한 잡음처리 연구가 필요하다.

이와 관련하여 시간-주파수 영역에서 잡음에 의해 지배되는 부분을 손실 영역 또는 비신뢰 영역이라 하고, 이 영역을 통계적 방법에 의해 채워넣거나, 우도 (likelihood) 계산에서 제외하는 손실 데이터 이론에 관한 연구가 최근에 다수 이루어지고 있다[2,3]. 또한 이와 유사한 개념으로 시간 영역에서 프레임들의 신호대잡음비를 정규화하여 프레임 신뢰도로 사용하고 이를 DTW (dynamic time warping) 기반의 인식 과정에서 두 벡터간 거리의 가중치로 적용함으로써 잡음에 의한 정보손실이 적은 프레임이 인식 결과에 더 크게 기여하게 하는 분절 (segmental) 신호대잡음비 가중 방식이 연구되기도 하였으며[4], 스펙트럼 차감법 (spectral subtraction)으로 잡음을 제거하여 얻은 음성 추정치의 분산을 프레임별로 계산하고 이를 패턴 비교 단계에서 가중치로 적용하는 방식[5] 등이 연구되었다. 잡음이 일정한 경우는 무음 구간의 잡음 정보를 획득할 수 있어 손실 데이터 이론에서 시간-주파수 영역의 각 부분에 대해 손실 여부를 결정하거나, 프레임 신뢰도를 위한 분절 신호대잡음비의 계산이 용이하다. 그러나 잡음이 시간에 따라 급격히 변하는 경우는 무음 구간의 잡음 정보를 활용할 수 없으므로 시간-주파수 영역의 각 부분에 대한 신호대잡음비 추정이 여전히 어려운 문제로 남아 있다.

본 연구에서는 시간 영역의 특정 위치에서 발생하여 음성구간의 일부분을 오염시키는 시간선택 잡음에 대해 인식 성능을 향상시키고자 한다. 이 경우 잡음의 추정이 어려우므로 본 연구에서는 잡음을 직접 추정하는 대신에 입력 음성과 근사한 HMM (Hidden Markov Model) 참조 모델을 이용하여 잡음 음성의 프레임 신호대잡음비를 근사하고, 이를 인식점수 계산 단계에서 신뢰도로 적용하는 MFR (model-based frame reliability)을 제안한다. 또한, 참조 모델의 평균 벡터와 입력 벡터간의 거리 정보를 신뢰도로 활용하는 DFR (distance-based frame reliability) 방식을 제안하여 MFR과 비교한다. 본 논문의 2장에서는 프레임 신뢰도 가중의 개요를 설명하고, 3장에서는 제안한 모델 기반의 프레임 신뢰도 및 이를 이용한 인식 방법에 대해 기술한다. 4장에서 실험 및 결과를 기술하고 마지막으로 5장에서 결론을 맺는다.

II. 프레임 신뢰도 가중

HMM (hidden Markov model)을 이용한 기존의 음성인식 시스템에서 음성 신호는 일련의 특징 벡터들로 표현되며, 이 벡터들은 해독단계에서 우도에 동일한 정도로 기여한다. 예를 들어 HMM λ에 대한 특징 벡터열 $Y=\{y_1, y_2, \cdots, y_T\}$의 상태열을 $S=\{s_1, s_2, \cdots, s_T\}$라고 하면, Y의 우도는 다음의 식 (1)과 같이 계산된다.

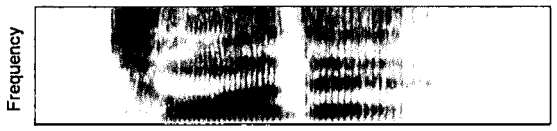
$$\Pr(Y, S \mid \lambda) = \Pr(y_1 \mid s_1)\Pr(y_2 \mid s_2) \cdots \Pr(y_T \mid s_T) \cdot \Pr(S \mid \lambda) \quad (1)$$

일반적인 인식기에서는 시간 영역에서 일정한 간격으로 특징을 추출하므로, 식 (1)에 의해 입력 신호에서 상대적으로 길고 정상적인 부분이 인식 결과에 지배적인 영향을 미치게 된다[6]. 그러나 신호에서 빠르게 변하는 부분도 인식에 중요한 정보를 포함하고 있어 이 부분의 프레임 해상도를 높이는 가변 프레임율 (variable frame rate; VFR)[6]이 제안되었고, 이 외에도 가변 정보율 (variable information rate; VIR)[7] 등에 관한 연구로 잡음이 없는 음성에서도 음성인식에 효과적인 음성 정보가 시간 영역에서 비균일하게 분포함을 알 수 있었다.

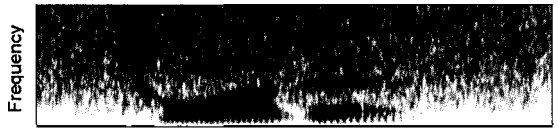
잡음이 존재하는 환경에서 음성 정보의 분포는 잡음의

크기가 시간에 따라 일정한 경우와 잡음의 크기가 시간에 따라 변하는 경우로 구분하여 생각해 볼 수 있다. 먼저 잡음의 크기가 시간에 따라 일정한 경우, 음성의 크기는 시간에 따라 동적으로 변하므로 매 프레임의 신호대잡음비가 가변적이다. 둘째로 잡음의 크기가 시간에 따라 변하는 경우는 실제 응용 환경에 존재하는 대부분의 잡음 특성에 해당하며, 음성과 잡음의 크기가 동시에 변하므로 프레임 신호대잡음비는 잡음이 일정한 경우와 마찬가지로 가변적이다. 음성 신호에서 신호대잡음비가 낮은 프레임은 잡음에 의해 상대적으로 심하게 손상되어 신호대잡음비가 높은 프레임에 비해 정보 손실이 크게 발생한다. 그림 1에서 (a), (b) 및 (c)는 각각 깨끗한 음성의 스펙트로그램과 정상적 잡음에 의해 오염된 음성의 스펙트로그램 및 시간 영역에서 불규칙하게 발생하여 음성의 일부분을 손상시키는 시간선택 잡음이 섞인 음성의 스펙트로그램을 나타낸다.

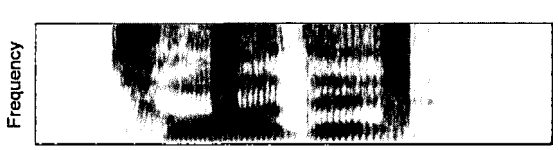
이와 같이 잡음이 존재하는 대부분의 응용 환경에서는 잡음의 특성에 따라 시간 영역에서 정보가 비균일하게 분포하며 잡음에 의해 심하게 손상되어 인식에 유효한 정보량이 적어진 프레임일수록 신뢰도가 낮다고 볼 수 있다. 기존의 음성인식 시스템들에서는 각 프레임의 신뢰도를 고려하지 않고 모든 프레임이 동일한 비중으로 패턴 비교에 참여하였으나, 신뢰도가 높은 프레임일수록 인식 결과에 큰 영향을 미치도록 가중함으로써 인식 성능을 더욱 향상시킬 수 있다. 식 (1)에 프레임 신뢰도 $\gamma(t)$를 반영한 우도는 식 (2)와 같이 나타낼 수 있다. 이 때, 신뢰도 $\gamma(t)$는 0에서 1사이의 정규화된 값으로서 이 값이 1에 가까울수록 해당 프레임이 높은 신뢰도를 갖으며, 후보 단어의 결정에 더 큰 영향을 주게 된다.



(a) 깨끗한 음성(/청와대/)의 스펙트로그램
(a) Spectrogram of a clean speech



(b) 백색 잡음에 의해 시간 영역이 균일하게 오염된 음성의 스펙트로그램
(b) Spectrogram of a speech corrupted by white noise. The signal is contaminated uniformly in time domain.



(c) 2회의 백색 버스트 잡음에 의해 시간 영역의 일부분이 오염된 음성의 스펙트로그램
(c) Spectrogram of a speech corrupted by white burst noise. The signal is contaminated partly in time domain.

그림 1. 가산 잡음에 의해 오염된 음성의 스펙트로그램 비교
Fig. 1. Comparison of speech spectrograms corrupted by additive noises.

$$\Pr(Y, S \mid \lambda) = \Pr(y_1 \mid s_1)^{\gamma(1)} \Pr(y_2 \mid s_2)^{\gamma(2)} \cdots \Pr(y_T \mid s_T)^{\gamma(T)} \cdot \Pr(S \mid \lambda) \quad (2)$$

III. 모델기반 프레임 신뢰도 가중

시간선택 잡음은 자동차 경적소리, 문 닫히는 소리, 키보드 소리, 총소리와 같이 특정 시간 영역에서 발생하여 음성의 일부분을 오염시키며 그 외의 부분에는 영향을 미치지 않는다. 잡음이 일정한 경우에는 무음 구간에서 획득한 잡음 정보를 이용하여 프레임의 신뢰도 추정이 용이하나, 시간선택 잡음은 잡음 추정이 어렵다. 따라서 본 논문에서 제안한 방법은 프레임 신뢰도를 추정하기 위해 깨끗한 음성의 통계적 모델을 사용한다. 즉 잡음 음성에 대응하는 연속 HMM의 상태들에 포함된 스펙트럼 정보를 이용하여 매 프레임에서 신호대잡음비를 추정하고, 이를 정규화하여 프레임 신뢰도로 사용한다.

3.1. 모델기반 프레임 신뢰도 (MFR)

제안한 방법은 연속 HMM의 평균 벡터열과 입력 MFCC 특징 벡터열의 코사인 역변환에 의해 얻은 필터뱅크 에너지를 이용하여 프레임 신호대잡음비를 근사한다. 입력 MFCC 벡터열을 $\{y_t\}$라 하고, 이에 대한 최대 우도 HMM의 최적 상태열을 $\{s_t\}$, 각 상태에 포함된 평균 벡터로 구성된 벡터열을 $\{x_t\}$라고 하자. MFCC 추출 과정에서 코사인 변환 행렬을 C라 하면, 코사인 역변환과 지수함수에 의해 t번째 프레임에서 잡음 음성과 깨끗한 음성의 멜 필터뱅크 에너지의 근사치를 각각 식 (3)과 (4)처럼 구할 수 있다.

$$y_t^e = \exp(C^{-1} y_t^c) \quad (3)$$

$$x_t^e = \exp(C^{-1} x_t^c) \quad (4)$$

식 (4)의 필터뱅크 에너지 벡터 x_l^e에서 i번째 필터의 저역, 고역 차단주파수 및 중심주파수에 해당하는 FFT 인덱스를 각각 f_L, f_H, f_C라 하고 음성의 파워스펙트럼을 $|X(f)|^2$라 하면, 이 필터를 통과한 음성의 에너지 $x_l^e(i)$는 식 (5)와 같이 표현할 수 있다. 식에서 $\omega_i(f)$는 저역 및 고역 차단주파수에서 0이고 중심주파수에서 1이며, 그 사이에서는 선형적으로 변하는 i번째 멜 필터의 주파수 응답으로서 f_C와 f_H 사이에서 $\omega_i(f)=1-\omega_{i+1}(f)$인 특성을 갖는다.

$$x_l^e(i)=\sum_{f=f_L}^{f_C}\omega_i(f)\cdot|X(f)|^2+\sum_{f=f_c}^{f_H}\omega_i(f)\cdot|X(f)|^2 \quad (5)$$

식 (5)와 같은 멜 필터뱅크의 특성에 의하여 임의의 주파수 f는 인접한 두 필터의 에너지 계산에 참여하게 되며, 이 때 각 필터에 해당하는 주파수 응답을 각각 $\omega'(f)$ 및 $\omega''(f)$라고 하면 음성의 에너지는 식 (6)과 같이 필터뱅크 에너지의 총합으로 구할 수 있다.

$$\begin{aligned}P_{speech}(t)&=\sum_i x_l^e(i)\\&=\sum_f\omega'(f)\cdot|X(f)|^2+\sum_f\omega''(f)\cdot|X(f)|^2\\&=\sum_f\omega'(f)\cdot|X(f)|^2+\sum_f(1-\omega'(f))\cdot|X(f)|^2\\&=\sum_f|X(f)|^2\end{aligned} \quad (6)$$

한편 주어진 부대역 내부에서 스펙트럼이 거의 변하지 않는다고 가정할 때, 잡음 음성의 필터뱅크 에너지 $y_l^e(i)$에서 음성의 필터뱅크 에너지 $x_l^e(i)$를 차감하고, 이들의 총합을 구하여 현재 프레임에서 잡음의 에너지를 식 (7)과 같이 근사할 수 있다.

$$P_{noise}(t)\approx\sum_i\left(\sqrt{y_l^e(i)}-\sqrt{x_l^e(i)}\right)^2 \quad (7)$$

제안한 MFR은 이와 같이 구한 프레임 신호대잡음비를 0에서 1사이의 값으로 정규화한 것으로 식 (8)과 같다. 식 (8)에서 $g(\cdot)$는 입력 음성의 최대 및 최소 프레임 신호대잡음비 값을 기준으로 현재 프레임의 신호대잡음비를 정규화하는 함수이다.

$$\gamma_{MFR}(t)=g\left(10\cdot\log_{10}\left(\frac{P_{speech}(t)}{P_{noise}(t)}\right)\right) \quad (8)$$

이와 다른 방식으로서 제안한 DFR (distance-based frame reliability)은 입력 특징벡터의 일부분에 왜곡이 심하게 발생할수록 참조 벡터와의 거리가 상대적으로 멀어지는 사실을 이용한 방법으로서 식 (9)와 같이 프레임 신뢰도를 정의한다. 이 방법은 MFR에 비해 계산량이 적으며 특징추출 방법과 무관하다.

$$\gamma_{DFR}(t)=g\left(\frac{1}{\sum_i(y_l^e(i)-x_l^e(i))^2}\right) \quad (9)$$

3.2. 프레임 신뢰도의 가중

앞 절에서 구한 프레임 신뢰도를 기존의 비터비 알고리즘에 가중치로 적용하여 입력 신호의 프레임별 신뢰도에 따라 각 프레임이 인식 점수에 다른 정도의 기여를 하도록 수정된 기존의 가중 비터비 알고리즘은 다음과 같다[5].

Step 1: Initialization. For each state i of HMM,

$$\delta_1(i)=\pi_i\times[b_i(x_1)]^{\gamma(1)}$$
$$\Psi_1(i)=0$$

Step 2: Iteration. For $2\le t\le T$ and $\forall j$,

$$\delta_t(j)=\max_i[\delta_{t-1}(i)\times a_{ij}]\times[b_j(x_t)]^{\gamma(t)}$$
$$\Psi_t(j)=\arg\max_i[\delta_{t-1}(i)\times a_{ij}]$$

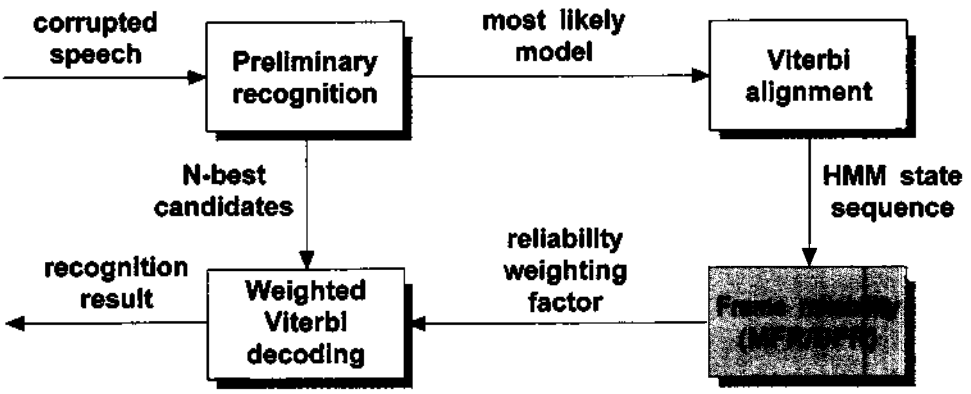


그림 2. 프레임 신뢰도 가중을 적용한 음성인식 절차
Fig. 2. Overall recognition procedure for an ASR system using frame reliability weighting.

Step 3: Termination.

$$P^* = \max_{s \in s_f} [\delta_T(s)]$$

본 논문에서는 $\gamma(t)$로서 MFR 또는 DFR을 사용하며 $\gamma(t)=1$이면 t 번째 프레임이 인식단어의 결정에 미치는 영향이 최대치가 되고, $\gamma(t)=0$이면 이 프레임에서 관측 확률이 무조건 1이 되므로 인식단어의 결정에 아무런 영향을 미치지 않게 된다. 그림 2는 지금까지 기술한 프레임 신뢰도 가중을 적용한 인식절차를 보인다.

그림 2에서 나타낸 인식절차는 먼저 1차 인식에 의해 N-best 후보들을 선택한 후, 그 중 최적 후보를 참조 패턴으로 사용하여 MFR 또는 DFR 값을 계산하고, N-best 후보들에 대해 가중 비터비 알고리즘을 적용하여 최종적인 인식결과를 얻는다.

IV. 실험 및 결과

4.1. 실험 환경

제안한 프레임 신뢰도 가중 방법을 검증하기 위해 100단어 규모의 고립단어 인식실험을 수행하였다. 실험에 사용한 음성 데이터베이스는 국어공학센터의 PBW (phoneme balanced word)-452이며[8], 이 중에서 100개의 단어를 임의로 선정하였다. 학습 자료로는 각 단어별로 남녀 화자 50명에 대한 2회의 발성을 사용하였으며, 평가 자료로는 남녀 화자 30명이 발성한 3000개의 발성을 사용하였다. 평가 자료에는 SINR (signal-to-impulsive noise ratio) 10, 5, 0, -5, -10 dB로 백색 버스트 잡음 및 총소리 잡음 (machine gun noise)을 가산하였다. 이때 사용한 SINR은 다음 식과 같이 정의된다[9].

$$SINR = \frac{P_{signal}}{\alpha \cdot P_{noise}} \tag{10}$$

식 (10)에서 α는 신호에서 잡음에 의해 오염된 부분의 비율을 나타내며, P_{signal}과 P_{noise}는 각각 음성과 잡음의 에너지를 의미한다. 본 연구에서는 1회의 버스트 잡음이 각 입력 길이의 5%에 해당하는 지속 시간을 갖도록 하여, 각 입력의 10% 및 20%를 오염하였다. 10% 오염의 경우 각 입력 음성의 1/3 및 2/3 지점에 잡음을 가산하였고, 20% 오염의 경우 입력 음성의 1/5, 2/5, 3/5, 4/5 지점에 잡음을 가산하였다. 특징 벡터로는 0차 계수를 포함한 12차의 MFCC를 사용하였다.

4.2. 실험 결과

먼저 입력 음성에서 오염된 부분의 비율과 오염의 정도에 따른 인식기의 성능을 살펴보았다. 표 1은 깨끗한 음성 및 SINR 10, 5, 0, -5, -10 dB의 백색 버스트 잡음 음성에 대한 인식결과를 나타내며, 그림 3은 표 1에서 음성이 10%

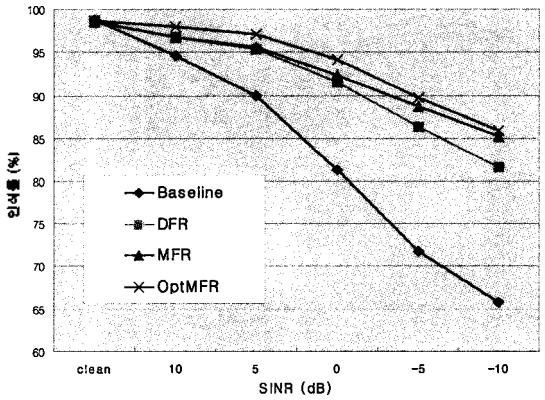


그림 3. 백색 버스트 잡음에 의해 각 입력 음성의 10%가 오염된 경우의 단어 인식률(%) 비교

Fig. 3. Comparison of word accuracy (%) for white burst noises corrupting 10% of each utterance.

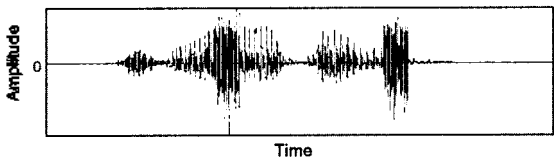
표 1. 백색 버스트 잡음에 의해 10% 및 20%가 오염된 음성에 대한 단어 인식률 (%)

Table 1. Word accuracy (%) for white burst noise corrupting 10% and 20% of each utterance.

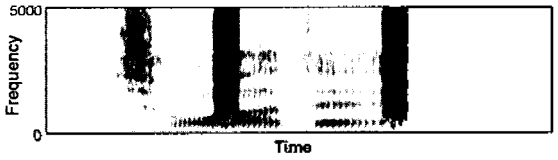
	오염률	SINR					
		clean	10 dB	5 dB	0 dB	-5 dB	-10 dB
Baseline	10%	98.8	94.6	90.0	81.3	71.7	65.7
	20%	98.8	92.4	82.6	67.6	54.5	44.0
DFR	10%	98.6	96.7	95.4	91.6	86.4	81.6
	20%	98.6	94.8	89.4	81.2	68.9	59.0
MFR	10%	98.5	96.7	95.6	92.3	88.7	85.2
	20%	98.5	95.6	92.6	85.2	74.4	65.0
OptMFR	10%	98.6	98.0	97.1	94.1	89.8	85.9
	20%	98.6	97.5	94.7	87.1	75.7	66.0

오염된 경우의 인식 결과를 그래프로 나타낸 것이다.

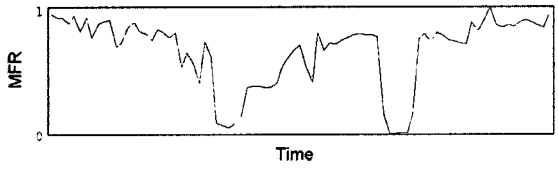
표 1에서 오염 비율이 고정된 상태에서 오염의 정도가 심할수록 인식률이 떨어짐을 볼 수 있다. 또 동일한 SINR에서 오염 비율이 큰 경우 인식률이 저하되었다. 본 연구에서 제안한 MFR 및 DFR 가중에 의해 인식률이 향상되었으며 잡음이 클수록 더 큰 성능 향상을 나타내었다. 또한 MFR이 DFR 방식에 비해 더 높은 성능을 보여 프레임 신뢰도를 더 효과적으로 표현함을 알 수 있었다. 표에서 OptMFR은 입력 신호의 정답 모델이 주어진 경우의 MFR



(a) 음성 파형 (/청와대/)
(a) Waveform



(b) 스펙트로그램
(b) Spectrogram



(c) MFR 값
(c) MFR sequence

그림 4. 2회의 백색 버스트 잡음에 의해 오염된 경우
Fig. 4. Example of speech contaminated by white burst noise.

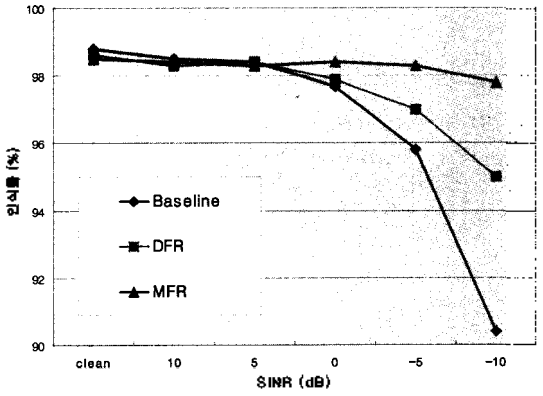


그림 5. 총소리 잡음에 의해 20%가 오염된 음성에 대한 단어 인식률(%)의 비교
Fig. 5. Comparison of word accuracy (%) for machine gun noises corrupting 20% of each utterance.

성능으로서, 1차 인식에 의해 입력 신호에 가장 근사한 모델을 찾는 MFR 성능의 상한선으로 볼 수 있다. 그림 3에 나타난 바와 같이 MFR은 OptMFR에 비해 성능면에서 크게 뒤떨어지지 않아 1차 인식의 성능이 MFR 결과에 큰 영향을 미치지 않았음을 알 수 있다. 그림 4는 SINR 0 dB의 백색 버스트 잡음에 의해 10%가 오염된 경우의 음성 파형 및 스펙트로그램, 그리고 제안한 MFR에 의해 구한 프레임 신뢰도를 나타낸다. 이 그림에서 잡음에 의해 손상된 부분에서 프레임 신뢰도가 급격히 저하됨을 볼 수 있다.

그림 5에서는 보다 실제적인 잡음에 해당하는 총소리 잡음에 대한 실험 결과를 나타낸다. 이 경우에도 잡음이 증가함에 따라 백색 버스트 잡음과 비슷한 경향을 나타내었다. 총소리 잡음은 신호의 에너지가 짧은 시간 영역에 집중되어 있어서 동일 SINR의 백색 버스트 잡음에 비해 성능을 크게 저하시키지 않았다.

V. 결 론

본 논문에서는 시간 영역에서 부분적으로 오염된 음성을 대상으로 인식성능을 높이고자 하였다. 이를 위하여 음성의 매 프레임마다 신뢰도를 측정하고, 신뢰도가 높은 프레임이 최종적인 인식 결과에 더 큰 영향을 주도록 하였다. 무음 구간에서 잡음 정보의 수집이 비교적 용이한 정상적 잡음과는 달리 음성의 일부분이 오염된 경우에 잡음의 추정이 어려우므로 본 연구에서는 깨끗한 음성의 통계적 모델과 입력 신호를 정렬하고, 특징추출의 역과정에 의해 구한 참조 모델의 필터뱅크 에너지와 입력 신호의 필터뱅크 에너지를 이용하여 프레임 신호대잡음비와 비례하는 프레임 신뢰도를 계산하였다. 실험 결과 제안한 MFR 방법은 오염된 지역에서 신뢰도가 급격히 저하되었으며, 신뢰도를 비터비 알고리즘에 가중치로 적용하여 인식 성능을 높일 수 있었다. 제안한 방법은 잡음의 종류 및 시간적인 발생 위치에 대해 무관하다는 장점이 있으며, 향후에는 다양한 비정상적 잡음에 대한 프레임 신뢰도 추정방법에 대한 연구가 필요하다.

참 고 문 헌

1. Y. Gong, "Speech Recognition in Noise environments: A Survey," *Speech Communication*, vol. 16, 261-291, 1995.
2. M. Cooke, P. Green, L. Josifovski and A. Vizinho, "Robust automatic speech recognition with missing and unreliable

acoustic data," *Speech Communication*, vol. 34, 267–285, 2001.
3. 김락용, 조훈영, 오영환, "손실 데이터 이론을 이용한 강인한 음성 인식," 한국음향학회지, 20(3), 56–62, 2001.
4. H. Kobatake and Y. Matsunoo, "Degraded Word Recognition based on Segmental Signal-to-Noise Ratio Weighting," *Proc. IEEE Int. Conf. Acoustic Speech Signal Processing*, vol. I, 425– 428, 1994.
5. N. B. Yoma, F. McInnes and M. Jack, "Weighted Matching Algorithms and Reliability in Noise Cancelling by Spectral Subtraction," *Proc. IEEE Int. Conf. Acoustic Speech Signal Processing*, vol. 2, 1171–1174, 1997.
6. K. M. Ponting and S. M. Peeling, "The use of variable frame rate analysis in speech recognition," *Computational Speech and Language*, vol. 5, 169–179, 1991.
7. I. J. Choi, C. K. Un and N. S. Kim, "Speech recognition based on variable information rate model," *Electronics Letters*, vol. 33, 749–750, 1997.
8. 이용주, "음성데이터베이스의 현황 및 과제," 제13회 음성통신 및 신호처리 워크샵, 13 (1), 279–287, 1996.
9. S. V. Vaseghi and B. P. Milner, "Speech Recognition in Impulsive Noise," *Proc. IEEE Int. Conf. Acoustic Speech Signal Processing*, vol. 1, 437–440, 1995.

저자 약력

● 조 훈 영 (Hoon-Young Cho)

1995년 8월: 한국과학기술원 전산학과 (학사)
1998년 2월: 한국과학기술원 전산학과 (석사)
1998년 3월~현재: 한국과학기술원 전자전산학과 전산학전공 박사과정 재학중
※ 주관심분야: 잡음에 강한 음성인식, 음질 개선, 신호 분리

● 김 락 용 (Lag-Young Kim)

1989년 2월: 연세대학교 전자공학과 (학사)
1991년 2월: 연세대학교 본대학원 전자공학과 (석사)
2002년 2월: 한국과학기술원 전자전산학과 전산학전공 (박사)
1991년 1월~현재: LG 전자기술원 정보기술 (연) MI 그룹 선임연구원
※ 주관심분야: 음성인식, 음성신호처리

● 오 영 환 (Yung-Hwan Oh)

1972년: 서울대학교 공과대학 (학사)
1974년: 서울대학교 교육대학원 (석사)
1980년: Tokyo Institute of Technology 정보공학전공 (박사)
1981년~1985년: 충북대학교 컴퓨터공학과 조교수
1983년~1984년: University of California (Davis) 연구교수
1995년~1996년: Carnegie-Mellon University 연구교수
1985년~현재: 한국과학기술원 전자전산학과 전산학전공 교수
※ 주관심분야: 음성인식, 음성합성, 음성코딩, 화자인식, 대화관리, 신경회로망, 전문가시스템