한국음향학회지 제20권 제3호 pp. 56~62 (2001)

손실 데이터 이론을 이용한 강인한 음성 인식

Robust Speech Recognition Using Missing Data Theory

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

*한국과학기술원 전자전산학과

(접수일자: 2000년 12월 20일; 채택일자: 2001년 3월 5일)

본 논문에서는 손실이 발생하는 상황에서 높은 인식률을 유지하기 위해서 손실 데이터 이론을 음성 인식기에 적용하였다. 손실 데이터 이론은 일반적으로 이용되는 통계적 정합 방법인 은닉 마코프 모델 (HMM: hidden Markov model) 중 연속 Gaussian확률 밀도 함수를 이용하여 음성 특징들의 출력 확률을 나타내는 경우에 쉽게 적용할 수 있다는 장점을 갖고 있다. 손실 데이터 이론의 방법 중 계산량이 적고 인식기에 적용이 쉬운 주변화 (marginalization)방법을 사용하였으며 특징 벡터의 특정 차수나 시간열의 손실 검출 방법은 음성 신호의 에너지와 주위 배경 잡음의 에너지의 차이가 임계치보다 작게 되는 부분을 찾는 주파수 차감 방법을 이용하였다. 본 논문에서 제안한 손실 영역의 신뢰도 평가는 분석 구간이 모음일 확률을 계산해서 비교적 잉여 정보가 많이 포함된 모음화된 구간의 손실만을 처리하도록 하였다. 제안한 방법을 사용하여 여러 잡음 환경에 대해서 기존의 손실 데이터 처리 방법만을 사용한 경우보다 452 단어의 화자독립 단어 인식 실험을 수행한 결과 오류율 측면에서 평균적으로 약 12%의 성능 향상을 얻을 수 있었다.

핵심용어: 손실 데이터 이론, 잡음에 강인성, 주파수 차감, 모음화 확률, 음성 인식
투고분야: 음성처리 분야 (2,5)

In this paper, we adopt a missing data theory to speech recognition. It can be used in order to maintain high performance of speech recognizer when the missing data occurs. In general, hidden Markov model (HMM) is used as a stochastic classifier for speech recognition task. Acoustic events are represented by continuous probability density function in continuous density HMM (CDHMM). The missing data theory has an advantage that can be easily applicable to this CDHMM. A marginalization method is used for processing missing data because it has small complexity and is easy to apply to automatic speech recognition (ASR). Also, a spectral subtraction is used for detecting missing data. If the difference between the energy of speech and that of background noise is below given threshold value, we determine that missing has occurred. We propose a new method that examines the reliability of detected missing data using voicing probability. The voicing probability is used to find voiced frames. It is used to process the missing data in voiced region that has more redundant information than consonants. The experimental results showed that our method improves performance than baseline system that uses spectral subtraction method only. In 452 words isolated word recognition experiment, the proposed method using the voicing probability reduced the average word error rate by 12% in a typical noise situation.

Keywords: *Missing data theory, Noise robustness, Spectral subtraction, Voicing probability, Speech recognition,*
Ask subject classification: *Speech signal processing (2,5)*

책임저자: 김락용 (kly@bulsai.kaist.ac.kr)
305-701 대전시 유성구 구성동 한국과학기술원 전산학과
(전화: 042-869-8720; 팩스: 042-869-3510)

I. 서 론

지난 수년간 잡음에 강인한 음성 인식 방법에 관한 방대한 연구가 국내외적으로 이루어져 왔다[1,2]. 이 방법들 중의 다수는 음성 인식의 전처리 단계에서 잡음이 섞인 신호에 대해 스펙트럼 차감 방법이나 필터링 등을 통해 잡음을 제거한다[3,4]. 또한 최근에 연구되기 시작한 음성 정보의 부분적 손실이나 불완전한 음성에 대한 사람의 인지에 대한 여러 연구 결과를 통해 잡음에 손상된 음성 신호의 명료도 (intelligibility)를 유지시키는 특징들에 대해 많은 논문들이 발표되고 있다[1,5,6]. 실제 사용하는 음성 통신 채널에 따라 주파수상의 손실인 필터링 및 시간 영역에서 주위의 높은 소리에 의해 마스킹 (masking)되어 들을 수 없는 부분이 발생한다. 따라서 조용한 실험실 환경에서 광대역의 손실 없는 모든 정보를 이용해 학습된 인식 모델과 실제 사용 환경과의 차이로 인해 인식기의 성능이 급격히 저하된다.

본 논문에서는 손실 데이터 존재시 정확한 통계치를 추정하는 방법[7,8]을 음성 인식에 적용하였다. 채널 제한된 전화 음성이나 주위 잡음에 의해서 음성 신호의 특정 대역이 차폐되는 현상이 생기며 이때 발생한 손실된 부분에 의한 왜곡으로 학습된 인식기와의 정합 과정에서 오류가 발생한다. 이와 같은 손실을 효과적으로 처리하여 인식 과정에서 손실된 부분에 의해서 나머지 부분이 영향을 받는 누수 현상을 극복하고자 하였다. 음성 인식에서 일반적으로 이용되는 통계적 정합 방법인 은닉 마코프 모델 (hidden Markov model)중에 연속 Gaussian 확률 밀도 함수를 이용하여 음성 특징들의 출력 확률을 나타내는 경우에 쉽게 적용할 수 있다는 장점을 갖고 있다.

II. 손실 데이터 이론 및 음성 인식에의 응용

실제 환경에서는 여러 형태의 정보들이 왜곡되어 사람의 오감을 통해 인지된다. 그러나 원래의 정보들이 옳게 인지되는 것이 목표지만 여러 요인에 의해 차폐 (occlusion)가 발생한다. 생략과 간섭 현상은 특히, 사람의 정보 입력의 대부분을 차지하는 시각과 청각에서 손실 없는 지각을 위해서 중요한 역할을 한다[9]. 즉, 일상생활에서 사람은 인지되는 손실된 정보로부터 손실 전의 시각이나 청각 정보를 연상이나 남아 있는 잉여 정보로부터 복구한다. 청각에서 주위 사람이 떠드는 잡음 신호인 혼성잡음 (babble)에 노출되거나 듣고자 하는 소리보다 주위의 잡음이 큰 경우에 발생하는 마스킹효과 (masking effect)에 의해 시각에서의 차폐와 유사한 현상이 나타난다. 시각이나 청각 정보의 차폐에 의해 그림이나 소리의 일부분이 손상되며 손상된 부분을 손실 데이터라 한다. 또한, 손실 데이터를 처리하는 통계적 처리 기법을 손실 데이터 이론이라고 부른다.

본 논문에서는 사람의 손실 데이터에 대한 강인하고 유연한 대응을 정량화하여 음성 인식에 적용함으로써 성능 개선을 이루고자 한다. 손실 데이터 이론을 음성 인식에 적용하기 위해서는 해결해야 할 두 가지 문제가 있다. 첫 번째 문제는 입력된 데이터 중에서 손실된 부분을 찾는 문제이고, 두 번째 문제는 찾아진 손실 데이터를 재구성하는 문제이다. 기존의 연구에 의하면 첫 번째 문제는 주파수 차감 방법을 통해 차감된 주파수 성분이 임계치 값 이하로 내려간 부분을 손실된 부분으로 간주하였으며[10] 두 번째 문제를 위해 데이터 대치 방법 (data imputation)과 주변화 (marginalization)방법이 발표되었다[11].

2.1. 음성 인식에의 응용

조용한 환경에서 발음된 음성 신호를 이용하는 음성 인식기는 선형 채널 왜곡인 저대역 통과 필터나, 고대역 통과 필터, 대역 통과 필터 등을 거치거나 잡음이 가산되어지는 경우, 그리고 인터럽트에 의해 입력 음성이 중단되어 손실이 발생한 경우에 매우 급격한 성능 저하 현상을 보인다. 저하 현상들을 보상하기 위해 멜 (mel)주파수 단위의 필터 뱅크를 통과한 주파수 대역별 에너지와 손실 데이터 이론을 이용하여 좋은 개선 효과를 얻을 수 있었다. 그림 1은 이를 블럭도로 나타낸 것이다[10].

그림 1에서 사용하는 인식기는 HMM (hidden Markov model)의 상태에서 출력 확률을 표현함에 있어 켑스트럼으로 변환하지 않고 직접 멜 (mel) 단위의 필터뱅크 출력 에너지 (MFB: mel-scale filter bank)를 특징 벡터로 사

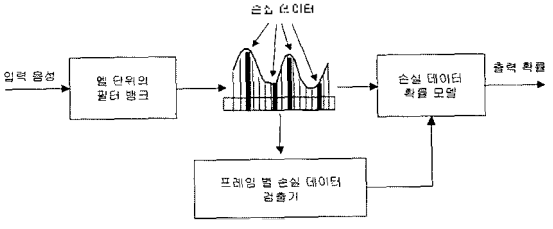


그림 1. 손실 데이터 이론을 사용하는 음성 인식기의 블록도
Fig. 1. Block diagram of robust speech recognition using missing data theory.

용한다. 이와 같이 스펙트럼의 크기정보를 직접 사용하면 입력 특징변수의 양은 증가하지만, 손상되고 손실된 특징 벡터의 직접적인 처리가 가능하다는 장점이 있다. 스펙트럼의 크기를 사용하는 특징 벡터에서 손실 데이터 검출기를 이용해서 음성인지 잡음인지를 검출한다. 사용하는 손실 데이터 검출기는 간단한 임계치를 사용하는데 임계치 이하가 되는 입력 특징 벡터의 차수들을 손실 (missing) 되었다고 결정한다. 각각의 스펙트럼의 크기의 특징에 손실인지 여부를 결정하고 이를 이용하여 주어진 모델의 출력 확률을 변화시킨다. 간단한 적용으로는 손실 데이터에 대한 확률 계산을 하지 않음으로써 손실 데이터의 특징 정보가 포함되지 않은 확률 밀도 함수를 구해 인식기에 적용하는 것이다.

2.2. 손실데이터 검출

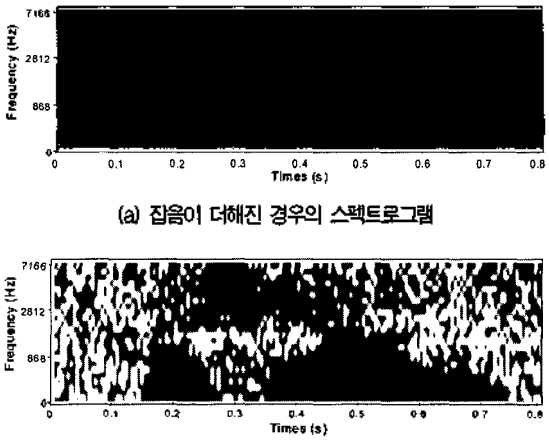
일반적으로 손실 데이터 검출에 주로 쓰이는 방법은 주파수 차감으로서 이를 이용하여 주변 잡음에 의해 음성 신호가 차폐되는 부분을 찾는다. 즉 주위 잡음에 대한 잡음의 통계적인 특성을 추정하고 이를 음성 신호와 비교하여, 주어진 임계치 이하의 값을 갖는 부분을 손실되었다고 검출하는 방법이다. 많이 사용되는 임계치 결정 방법은 두 가지이다. 첫 번째 방법은 주파수 차감 후 얻어지는 개선된 음성 신호의 에너지가 음수가 되는 값을 기준 임계치로 사용하는 방법이고 다른 방법은 신호 대 잡음비 (SNR)를 이용하는 방법으로 SNR이 음수가 될 때 음성이 배경 잡음에 의해 차폐된다고 판정하는 방법이다. 먼저 음수 에너지를 기준으로 사용하는 방법은 다음 식과 같이 표시할 수 있다.

$$|X(\omega)|^2 = |X(\omega)+N(\omega)|^2 - |N(\omega)|^2 < 0 \qquad (1)$$

여기서 $|X(\omega)|^2$는 전력 주파수 차감에 의해서 잡음이 제거된 음성신호의 에너지이다. 또한 각 ω에 대해 $|N(\omega)|^2$는 추정된 배경 잡음의 평균 에너지이다. 따라서 추정된 음성 신호의 에너지가 음수가 되는 부분을 손실이라고 검출한다. 또 다른 방법은 신호 대 잡음비 (SNR)를 이용하는 방법으로 만일 추정된 SNR이 0 dB보다 작아지면 손실이 발생했다고 판정하게 된다. 이를 식으로 나타내면 다음과 같다.

$$\log\left(\frac{|X(\omega)|^2}{|N(\omega)|^2}\right) < 0 \quad or \quad |X(\omega)|^2 < |N(\omega)|^2 \qquad (2)$$

그림 2는 특징 벡터로 24차 mel 단위의 필터뱅크 에너지를 사용하고 10 dB 레벨로 백색 잡음을 더한 경우에 대해서 식 1을 사용하여 손실 데이터를 검출한 예이다. 그림에서 알 수 있듯이 잡음에 의해서 차폐가 되는 비교적 작은 에너지를 갖는 부분이 손실된 것으로 희게 나타났다.



(a) 잡음이 더해진 경우의 스펙트로그램

(b) 잡음 음성에 대한 손실 데이터 검출 결과

그림 2. 손실 데이터의 검출 결과
Fig. 2. Result of detecting missing data block.

2.3. 손실 데이터 처리

손실 데이터로 검출된 부분은 크게 두 가지 방법에 의해서 다루어진다[11]. 즉, 손실 전의 값으로 적절한 예측을 통해서 대치하는 데이터 대치 방법 (data imputation)과 손실된 부분을 버리고 안정적으로 존재하는 값을 이용하는 주변화 (marginalization)방법이다. 데이터 대치 방법은 실제 추정을 통해서 손실된 부분을 재구성함으로써 음성 신호의 개선 (speech enhancement)에 사용될 수 있다. 본 논문에서는 인식기의 성능 개선에 중점을 두는 간단한 방법인 주변화를 사용하였다.

손실 데이타 이론은 높은 인식 성능을 보이는 Gaussian 혼합 밀도 함수를 이용하는 HMM 인식기에 쉽게 적용이 가능하다. HMM 모델 학습시 forward-backward 알고리듬에서 각 상태에 대한 Gaussian 혼합 출력 확률 밀도 함수를 추정하게 된다. $\boldsymbol{x}$는 입력 MFB 특징 벡터를 의미하고, C_i는 HMM에 대한 상태를 나타낸다. 특징 벡터 $\boldsymbol{x}$의 일부인 $\boldsymbol{x}_p$는 현재 안정적으로 계산된 특징 차수로 이루어진 특징 벡터의 일부를 의미하고 나머지는 손실되었다고 할 경우에 출력 확률 밀도 함수는 $f(\boldsymbol{x}_p \mid C_i)$로 바뀐다. 대각 행렬만으로 단순화시킨 covariance Gaussian 혼합 밀도 모델을 사용할 때 원래의 출력 확률 밀도 함수는 식 3과 같이 단일 분산 Gaussian 밀도 함수들의 가중 합으로 표시할 수 있다.

$$f(\boldsymbol{x} \mid C_i) = \sum_{j=1}^{M} w_j \prod_{i=1}^{D} N(m_{ij}, \sigma_{ij}^2) \quad (3)$$

여기서 M은 혼합 밀도수를 w_j 는 혼합 밀도 함수들의 가중치를 나타낸다. D는 입력 특징 벡터의 차수를, 그리고 $N(m_{ij}, \sigma_{ij}^2)$은 입력 특징 벡터 $\boldsymbol{x}_i$와 j번째 혼합 밀도 함수에 대해서 평균 m_{ij}와 분산 σ_{ij}^2를 갖는 단일 분산 Gaussian 밀도 함수를 나타낸다. 혼합 밀도 함수의 일부 차수가 손실되었을 경우 위의 식은 다음과 같이 표시된다.

$$f(\boldsymbol{x} \mid C_i) = \sum_{j=1}^{M} w_j \underbrace{\prod_{i} N(m_{ij}, \sigma_{ij}^2)}_{\text{(present)}} \underbrace{\prod_{i} N(m_{ij}, \sigma_{ij}^2)}_{\text{(missing)}} \quad (4)$$

위의 식 4에서 각각의 mixture 성분은 존재하는 특징 벡터의 차수에 대한 일차의 Gaussian성분과 손실된 특징 차수에 대한 일차의 Gaussian 성분의 곱으로 표시된다. 수정된 음성의 출력 확률 밀도 함수 $f(\boldsymbol{x}_p \mid C_i)$는 원래의 출력 밀도 함수를 손실 특징 벡터의 차수 별로 적분해 곱함으로써 얻을 수 있다. 즉 식 5와 같다.

$$f(\boldsymbol{x}_p \mid C_i) = \int f(\boldsymbol{x} \mid C_i) d\boldsymbol{x}_m \quad (5)$$

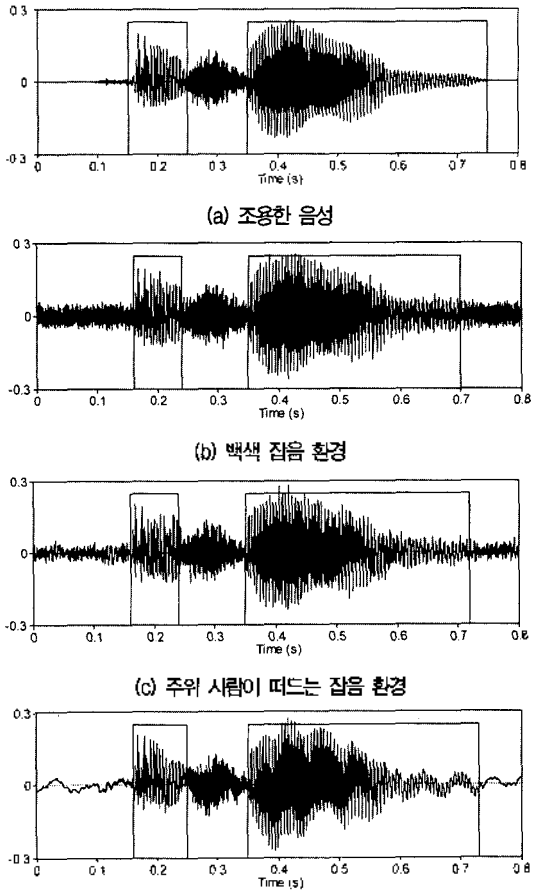
식 5의 적분에 의해 식 4의 오른쪽 항은 1이 되며 수정된 음성의 출력 밀도 함수는 다음과 같다.

$$f(\boldsymbol{x}_p \mid C_i) = \sum_{j=1}^{M} w_j \underbrace{\prod_{i} N(m_{ij}, \sigma_{ij}^2)}_{\text{(present)}} \quad (6)$$

식 6은 원래의 출력 밀도 함수 계산에서 손실된 특징 벡터의 차수에 대응되는 항을 넣지 않고 계산함으로써 쉽게 얻을 수 있다. 손실 특징 벡터의 발생에 따라서 변화하는 출력 확률 밀도 함수의 계산은 적은 수의 추가 연산에 의해 수행되므로 전체적인 복잡도를 증가시키지 않고 현재 입력되는 특징 벡터의 손실 여부를 알 수 있다면 시간에 따라 변화하는 필터링이나 잡음에 쉽게 대처할 수 있다는 장점이 있다.

III. 모음화 확률 (voicing probability)을 이용한 손실 데이터 검출

모음화된 (voicing)구간은 pitch가 존재하는 영역을 말하며 여러 함수를 통해 모음화된 영역을 찾을 수 있다. 일반적으로 많이 사용되는 방법으로는 영교차율 (zero-crossing rate), 1차 자기상관 계수 등이 이용된다. 또한 전체 주파수



(a) 조용한 음성

(b) 백색 잡음 환경

(c) 주위 사람이 떠드는 잡음 환경

(d) 자동차 운행 잡음 환경

그림 3. 여러 잡음 환경에서의 모음화 구간 검출의 예 (모든 잡음은 SNR이 10 dB임)
Fig. 3. Examples of finding voiced region in several noise conditions (when SNR is 10dB).

대역의 크기의 자승에 대한 pitch가 존재하는 저주파 대역의 자승의 비를 구해서도 찾을 수 있다. 실제 응용의 경우에는 하나의 방법이 아닌 여러 방법을 같이 적용하거나 통계적 패턴 인식에서 사용하는 방법 등을 결합해서 사용한다. 본 논문에서는 정규화된 교차 상관함수 (cross correlation function)와 동적 프로그램 (dynamic programming)을 이용하는 RAPT (Robust Algorithm for Pitch Tracking)방법을 사용하였다[12]. RAPT 방법에서는 원래의 표본화율 (sampling rate)과 이보다 낮은 표본화율로 얻어진 두 가지 음성신호에 대한 정규화된 교차 상관함수를 이용하며 동적 프로그램을 후처리로 이용해서 일련의 정규화된 교차 상관함수의 peak로 이루어진 pitch가 존재하는 구간과 무성음 혹은 묵음구간의 후보를 찾게 된다. 따라서 pitch가 존재하는 구간과 그렇지 않은 부분으로 나누는 과정에서 얻어지는 모음화 확률 (voicing probability)은 연속적인 확률 값이 아닌 0 또는 1 의 이진

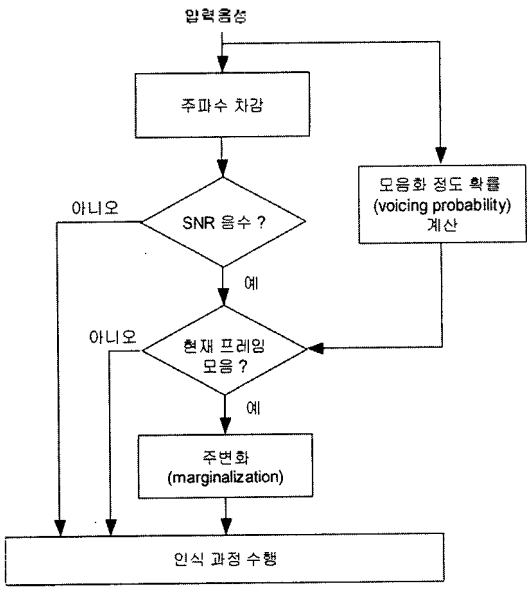


그림 4. 모음화 확률을 활용한 제안된 손실 데이터 추출방법의 블록도

Fig. 4. Block diagram of the proposed method incorporating voicing probability for detecting missing data.

확률값을 갖는다.그림 3은 본 논문에서 사용한 RAPT방법을 여러 종류의 잡음 (SNR 10dB)의 경우에 대해 모음화 구간의 검출을 수행한 결과이다. 조용한 음성의 경우를 기준으로 할 때 1 프레임 정도의 오차의 정확도로 검출됨을 알 수 있다.

그림 4는 모음화 확률이 결합된 손실 데이터 이론의 처리 블럭도이다. 손실 데이터 여부는 주파수 차감을 통하여 음수의 SNR이 되는 부분으로 결정하며 상대적으로 낮은 SNR을 갖는 자음 부분은 주파수 차감에 의한 왜곡 현상이 심하므로 그대로 사용하고 SNR이 높은 모음 부분에 대해서 마스킹 레벨 혹은 잡음의 에너지보다 음성 신호의 에너지가 낮은 대역 (필터뱅크의 차수)을 손실된 것으로 판정한다. 손실 데이터는 주변화 (marginalization)를 통해 인식기에서 인식 과정에서 주어진 단어의 HMM모델의 상태에 대한 출력 확률을 계산시에 포함시키지 않도록 함으로써 성능 향상을 얻고자 하였다.

사람의 지각에 있어서 모음과 같이 잉여 정보가 많이 있는 부분의 확실한 손실만을 인식 과정에서 손실된 것으로 처리하기 위해 모음화 확률을 도입하여 손실 영역에 대한 신뢰도를 평가하는 방법을 제안하였다. 제안한 방법은 여러 잡음에 관계없이 모음화된 영역을 일관되게 찾고 이를 기반으로 이미 찾아진 손실 영역에 대한 신뢰도 평가를 수행한 후 최종적으로 손실된 데이터만을 인식과정에서 제외하고 인식을 수행하는 실험을 수행하였다. 자세한 실험 결과는 4장에서 기술하고자 한다.

IV. 실험 및 결과

4.1. 실험 환경

본 논문에서 사용한 음성 데이터베이스는 국어공학센터에서 만든 PBW-452로 음소별로 균등하게 분포되도록 만든 고립 단어 DB이다. 총 화자 수는 총 70명으로 이들 중 50명은 HMM모델의 학습용으로 나머지 20명은 인식 테스트를 위해 사용하였다. 화자별 452단어를 2번씩 반복해서 발음하였다.

사용한 단어 모델로는 전체 단어 모델로 연속 HMM모델을 사용하였다. 모델별로 가변 상태 수가 아닌 상태 수를 14개로 고정시켜 사용하였으며 각각의 상태에서 출력 확률 밀도는 3개의 mixture를 가지는 Gaussian 분포로 나타내었다. 또한 본 논문에서는 두 종류의 특징 벡터를 추출해 비교 실험하였다. 멜 (MEL)단위의 필터 뱅크 에너지 (MFBE)와 잡음에 강인하도록 RelAtive SpecTrAl (RASTA) 처리된 임계 대역단위의 필터 뱅크 에너지 (CBE-RASTA: critical band energy with rasta filtering)를 사용하였다. 또한 잡음 환경 실험에 사용한 잡음을 모의하기 위해 잡음에 대한 데이터 베이스인 NOISEX-92를 활용하였다[13]. 사용한 잡음 원은 백색잡음 (WGN), 주위 사람들이 떠드는 혼성 잡음 (BAB), 그리고 유색 잡음인 자동차 잡음 (CAR)의 세 종류이다.

4.2. 특정 대역의 손실에 대한 손실데이터 이론의 성능 평가

먼저 손실 데이터 처리 여부에 대한 성능 평가를 위해 특징 벡터의 특정 차수를 임의로 손실시켰다. 멜 단위의 필터 뱅크의 출력 에너지를 특징 벡터로 사용한 기본 시스템의 경우에 채널의 왜곡내지는 손실이 발생할 때의 결과 (no-MDT) 와 손실 데이터 이론을 적용하여 손실이 있는 대역의 값을 빼고 신뢰된 부분만을 갖고 인식과정에 사용하는 주변화 (marginalization) 과정을 적용시킨 경우 (MDT) 의 결과를 비교해 보았다. 여기서 MDT는 missing data theory를 의미한다.

그림 5는 주어진 24 차의 멜 단위의 필터뱅크 에너지 (MFBE)를 특징 벡터로 사용한 경우에 해당 주파수 대역을 나타내는 특징 차수의 값을 임의로 손실시킬 경우 오류율의 변화이다. 결과에서 알 수 있듯이 5차 이상의 차수

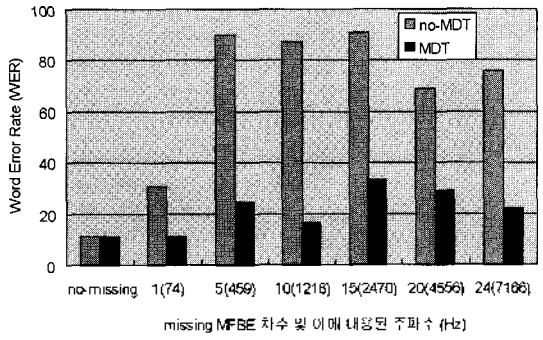


그림 5. 특정 대역이 손실될 경우의 손실 데이터 이론을 적용한 인식기의 성능

Fig. 5. Performance of recognition system incorporating missing data theory in case of a specific feature element missing.

(450 Hz 이상)가 손실되었을 경우는 아무런 처리를 하지 않으면 오류율은 90%까지 증가되지만 손실된 차수의 값을 주변화 시키는 손실 데이터 처리를 거치면 오류율은 30% 정도로 감소하게 된다. 또한 손실된 차수에 대응되는 주파수 성분의 상대적인 중요도를 실험 결과에서 알 수 있다. 중심 주파수가 2470 Hz 인 15차 멜 단위 주파수 성분이 손상되었을 경우 가장 큰 오류율을 보였다.

4.3. 특징 벡터에 따른 손실 데이터이론의 성능 비교

기존 논문에서 사용된 멜 주파수 단위의 필터 뱅크 출력 에너지 (MFBE)를 특징 벡터로 사용하는 경우를 기본 시스템으로 하고, 기본 시스템에서 손실 데이터 이론을 적용한 경우와 전처리로 주파수 차감 (Spectral Subtraction)을 적용한 경우 (SS: spectral subtraction) 에 대해 비교 실험을 수행하였다. 표 1의 결과를 보면 특징 벡터로 MFBE 와 손실 데이터 처리로 주변화 방법 (MG: marginalization)을 사용한 경우가 SNR이 15 dB 이상의 경우에 손실 데이터 처리를 하지 않은 기본 시스템에 비해서 성능 향상이 있었다. 전처리로 음질 개선을 위해 사용한 주파수 차감 방법을 적용한 결과는 잡음의 크기에 관계없이 기본 시스템에 비해 성능의 향상됨을 알 수 있다. RASTA처리를 결합한 임계대역 특징 벡터 (CBE-RASTA)를 사용하는 시스템의 인식 결과는 전처리로 주파수 차감을 적용한 경우보다 잡음에 강인하였다. 또한 손실 데이터 이론을 적용하지 않은 기본 시스템에 비해서는 오류율 측면에서 평균적으로 64.4%의 성능 개선을 얻을 수 있었다.

4.4. 모음화 확률을 이용하는 제안된 손실 데이터 검출 방법의 성능

손실 데이터 검출시 간단한 주파수 차감 방법에 의해

표 1. 멜 단위의 필터뱅크 에너지를 사용하는 기본 시스템과 RASTA처리한 임계 대역 필터뱅크 에너지를 사용한 제안된 시스템의 성능 (오류율) 비교

Table 1. Comparison of recognition performance by word error rate (WER (%) in both baseline system using mel-scale filter bank energy and proposed system using RASTA filtered critical filter bank energy.

SNR (dB)	특징 벡터 및 손실 처리 여부			
	MFBE	MFBE + MG	SS + MFBE	CBE-RASTA + MG
Inf.	10.9	10.9	10.9	10.5
25	67.4	50.7	35.3	15.5
20	84.7	76.6	46.4	16.1
15	93.3	91.0	59.6	17.8
10	96.6	96.2	74.4	23.7
5	97.8	97.7	88.2	43.2
0	98.8	98.9	96.6	69.1

표 2. 손실 데이터 검출에 주파수 차감 방법만 사용하는 기본시스템과 모음화 확률을 이용하는 제안된 시스템의 성능 (오류율) 비교

Table 2. Comparison of recognition performance by word error rate (WER (%) in both baseline system using spectral subtraction and proposed system incorporating voicing probability for detecting missing data.

SNR (dB)	CBE-RASTA + MG			CBE-RASTA + MG + PRV		
	WGN	BAB	CAR	WGN	BAB	CAR
Inf.	10.5	10.5	10.5	10.5	10.5	10.5
25	15.5	15.6	15.4	12.1	12.2	12.1
20	16.1	16.9	15.5	12.5	13.6	12.2
15	17.8	23.3	15.9	13.7	20.3	12.3
10	23.7	42.7	16.8	19.5	42.3	12.9
5	43.2	73.8	17.9	39.7	75.9	14.1
0	69.1	93.1	20.3	67.5	93.9	15.9

해당 특징 차수의 값이 주어진 임계치 이하의 값이 될 경우에 손실 데이터로 검출하는 기존의 방법은 추가적인 계산의 증가가 적다는 장점이 있지만 음성에서 자음의 경우는 모음에 비해 상대적으로 낮은 에너지를 갖음으로 인해서 주위의 잡음에 쉽게 차폐되어 거의 모두 손실 처리가 된다. 따라서 인식 오류가 증가하게 된다. 이와 같은 오류를 모음화 확률 (voicing probability)을 계산해서 손실 영역의 신뢰도 평가를 수행하는 후처리을 사용하여 개선시킬 수 있었다. 여러 잡음 환경에 대한 제안된 방법의 성능은 표 2에 나타나 있다. 표 2에서 CBE-RASTA + MG는 특징 벡터로 CBE-RASTA를 사용하고 손실 데이터 처리 방법으로 주변화 방법 (MG)을 사용한 경우이고, CBE-RASTA + MG + PRV는 모음화 확률 (PRV)을 사용하는 제안된 방법을 의미한다. 음성과 유사한 광대역 잡음인 babble잡음의 경우는 15 dB까지는 성능 향상이 있으나 그 이상의 잡음

레벨에서는 성능 향상이 없었다. 하지만 제안된 방법을 통해서 백색 잡음이나 협대역 유색 잡음의 일종인 자동차 잡음의 경우에는 오류율 측면에서 평균적으로 각각 11%, 21.9% 감소시키는 성능 개선을 얻을 수 있었다.

V. 결 론

본 연구에서는 기존 손실 데이터 이론을 적용한 음성 인식시스템의 문제점인 손실 데이터 검출의 신뢰성 평가를 위해 모음화 확률을 사용하는 방법을 제안하였다. 제안된 방법은 주파수 차감에 의해 찾아진 손실 부분에 대한 신뢰도 평가를 위해 모음화 확률 (voicing probability)을 사용하는 방법이다. 여러 잡음 환경에 대해서 voicing probability를 사용하지 않는 방법에 비해 오류율 측면에서 평균적으로 약 12%의 성능 향상을 얻을 수 있었다.

청각 기관에서 지각하는 비선형 주파수 대역인 임계 대역 필터군의 출력 에너지를 RASTA 처리하여 특징 벡터로 사용함으로써 잡음에 강인한 결과를 얻을 수 있었다. 실험 결과에 의하면 CBE-RASTA와 손실 데이터 이론을 결합할 경우에 특징 벡터로 MFBE를 사용한 기본 시스템에 비해서 오류율 측면에서 평균적으로 약 64% 라는 오류율 감소를 얻을 수 있었다.

앞으로는 손실 데이터 검출시 발생하는 오류가 인식기에 끼치는 영향을 정량적으로 평가하고 각 임계 대역별로 마스킹 단계 (masking level)를 다르게 적용하는 손실 데이터 검출 방법을 적용하면 보다 높은 성능 개선을 얻으리라 기대된다.

참 고 문 헌

1. Y. Gong, "Speech Recognition in Noise Environments: A Survey," *Speech Communication*, vol. 16, pp. 261-291, 1995.
2. S. Furi, "Recent Advances in Robust Speech Recognition," *Proc. ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, pp. 11-20, 1997.
3. S. F. Boll, "Suppression of Acoustic Noise in Speech using Spectral Subtraction," *IEEE Trans. on Acoust. Speech Signal Processing*, ASSP-27, pp. 113-120, 1979.
4. J. D. Koo and Gibson and S.D. Gray, "Filtering of Colored Noise for Speech Enhancement and Coding" , Proc. *IEEE Internat. Conf. Acoustic Speech Signal Processing*, pp. 349-352, 1989.
5. R. P. Lippmann, "Accurate Consonant Perception without Mid-Frequency Speech Energy," *IEEE Trans on Speech and Audio Processing*, vol. 4, No. 1, pp. 66-69, 1996.
6. J. B. Allen, "How do Humans Process and reconize Speech ?," *IEEE Trans on Speech and Audio Processing*, vol. 2, No. 4, pp. 567-577, 1994.
7. J. A. Rodrick and B.R. Donald, "Analysis of Incomplete Multivariate Data," John Wiley & Sons, 1987.
8. J. L. Schafer, "Analysis of Incomplete Multivariate Data," Chapman & Hall, 1997.
9. A. S. Bregman, "Auditory Scene Analysis:The Perceptual Organization of Sound," The MIT Press, 1990.
10. R. P. Lippmann, B. A. Carlson, "Using missing feature theory to additive select features for robust speech recognition with interruptions filtering and noise," *in Proc. Eurospeech*, vol. 1, pp. 37-40, 1997.
11. A. Vizinho, "Missing Data Theory, Spectral Subtraction and Signal-to-Noise Estimation for Robust ARS:an Integrated Study," *Proc. Euro. Conf. Speech Commnun. Technology*, pp. 2407-241, 1999.
12. D. Talkin, "A Robust Algorithm for Pitch Tracking," *Speech Coding and Synthesis, Chapter 14*, W. B. Kleijn, and K. K. Paliwal Eds., Elsevier, 1995.
13. A. Varga and H. Steeneken, "Assessment for Autometic 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, 1993.

저자 약력

● 김 락 용 (Lag-Young Kim)

1989년 2월: 연세대학교 전자공학과 졸업 (공학사)
1991년 2월: 연세대학교 본대학원 전자공학과 졸업 (공학석사)
1991년 1월~현재: LG 전자기술원 정보기술 (연) MI 그룹 선임연구원
1998년 3월~현재: 한국과학기술원 전자전산학과 전산학 전공 박사과정 재학중
※ 주관심분야: 음성인식, 음성신호처리

● 조 훈 영 (Hoon-Young Cho)

1995년 8월: 한국과학기술원 전산학과 (학사)
1998년 2월: 한국과학기술원 전산학과 (석사)
1998년 3월~현재: 한국과학기술원 전자전산학과 전산학 전공 박사과정 재학중
※ 주관심분야: 음질개선, 잡음에 강한 음성인식

● 오 영 환 (Yung-Hwan Oh)

1972년: 서울대학교 공과대학 (학사)
1974년: 서울대학교 교육대학원 (석사)
1980년: Tokyo Institute of Technology 정보공학전공 (박사)
1981년~1985년: 충북대학교 컴퓨터 공학과 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, 전문가 시스템