성대신호 기반의 명령어인식기를 위한 특징벡터 연구

(Effective Feature Vector for Isolated-Word Recognizer using Vocal Cord Signal)

정 영 규 † 한 문 성 † 이 상 조 ††
(Young-Giu Jung) (Mun-Sung Han) (Sang-Jo Lee)

요 약 본 논문은 환경 노이즈를 원천적으로 차단하는 성대 마이크를 이용한 명령어 인식기를 개발한다. 성대마이크는 환경 노이즈의 효과를 최소화하는 장점이 있다. 그러나 고주파의 부재와 부분적인 포먼트 정보 손실 때문에, 성대마이크를 이용해서 개발된 ASR시스템은 표준마이크를 이용한 시스템에 비해 낮은 성능을 보인다. 이러한 문제 때문에 ASR시스템 개발에 성대마이크를 이용한 경우는 표준 마이크로부터 입력되는 정보 보안하는데 주로 사용된다. 본 논문은 한국어의 음운적 특징과 신호 분석을 통해 성대마이크만을 사용한 높은 성능의 ASR 시스템을 개발 할 수 있음을 보인다. 주파수 대역내 에너지 합을 이용하는 MFCC 알고리즘이 갖는 성대신호 분석의 문제점을 제시하고, 성대신호를 대상으로 보다 높은 성능을 갖는 특징추출 알고리즘의 조건을 제시한다. 이러한 조건은 (1) 민감한 band-pass filter와 (2) 유/무성음 분리를 위해 사용하는 특징벡터의 사용이다. 실험 결과 제안된 조건을 만족하는 ZCPA 알고리즘을 적용한 경우가 MFCC를 적용한 경우보다 약 16%정도의 높은 성능을 보인다. 그리고 CMS와 RASTA와 같은 channel normalization 알고리즘을 적용한 경우 약 2%의 성능 향상이 있다.

키워드 : 성대마이크, 성대명령어인식기, 한국어 음운자질, ZCPA, MFCC

Abstract In this paper, we develop a speech recognition system using a throat microphone. The use of this kind of microphone minimizes the impact of environmental noise. However, because of the absence of high frequencies and the partially loss of formant frequencies, previous systems developed with those devices have shown a lower recognition rate than systems which use standard microphone signals. This problem has led to researchers using throat microphone signals as supplementary data sources supporting standard microphone signals. In this paper, we present a high performance ASR system which we developed using only a throat microphone by taking advantage of Korean Phonological Feature Theory and a detailed throat signal analysis. Analyzing the spectrum and the result of FFT of the throat microphone signal, we find that the conventional MFCC feature vector that uses a critical pass filter does not characterize the throat microphone signals well. We also describe the conditions of the feature extraction algorithm which make it best suited for throat microphone signal analysis. The conditions involve (1) a sensitive band-pass filter and (2) use of feature vector which is suitable for voice/non-voice classification. We experimentally show that the ZCPA algorithm designed to meet these conditions improves the recognizer's performance by approximately 16 %. And we find that an additional noise-canceling algorithm such as RASTA results in 2% more performance improvement.

Key words : Throat microphone, Throat microphone signal analysis, Isolated-word recognition system, Korean Phonological Feature Theory, ZCPA, MFCC

† 비 회 원 : 한국전자통신연구원 스마트인터페이스연구팀 연구원
reraj@etri.re.kr
msh@etri.re.kr
†† 종신회원 : 경북대학교 컴퓨터공학과 교수
sjlee@knu.ac.kr
논문접수 : 2005년 9월 21일
심사완료 : 2006년 12월 15일

1. 서 론

오늘날의 컴퓨터 환경은 Ubiquitous compute, pervasive compute, wearable compute와 같은 이동 환경 내에서 사용자가 원할 때 적합한 서비스를 받을 수 있

는 방향으로 흘러가고 있다. 이러한 컴퓨터 환경에서 가장 중요한 기술중의 하나가 사용자 인터페이스이며, 여러 사용자 인터페이스 기술 중에서 음성 인식의 중요성은 누구도 부인할 수 없는 사실이다. 그러나 현재 음성인식 기술의 경우 위에서 언급한 환경과 같이 잡음이 많은 환경에서는 적정 수준의 성능을 내지 못한다. 이러한 문제를 해결하기 위해 주파수차감법에 의한 방법[1], soft-decision 필터링 방법[2], 최소평균제곱오차 추정방법[3], 그리고 인간청각기 특성을 이용한 방법등 다양한 잡음처리 방법이 제안되어왔으나 고소음의 모바일 환경에서 적절한 성능을 내기에는 매우 어려운 상황이다. 이에 환경 잡음의 영향을 받지 않는 골도마이크, 이어마이크, 넥마이크등 다양한 마이크를 통해 인식기를 개발하려는 시도가 있어 왔으나, 아직 제대로 된 인식 성능을 내는 인식기는 전무한 상태이다.

현재 Noise-free device를 사용한 인식 기술의 연구는 두 가지 형태로 이루어지고 있다. 하나는 Noise-free device만을 사용하여 인식기를 개발하는 연구[4]이고 다른 하나는 표준마이크와 결합하여 음성정보의 보조적 정보로 활용하는 연구[5-8]이다. 이 경우 음성 구간 검출이나 Speech Enhancement 등을 위해 Noise-free device가 사용된다. Noise-free deice만을 사용하는 음성인식 기술 연구는 인간의 발성과 비슷한 신호를 생성하는 디바이스의 개발에 중심을 두고 있다. 이러한 이유 때문에 디바이스의 신호 특성을 반영한 인식기 연구는 거의 이루어지지 않고 있다. 그러나 음성인식 기술에서 환경 노이즈 처리기술의 한계를 볼 때, Noise-free device를 이용한 음성인식 기술의 개발은 음성인식 시스템의 상용화에 큰 기여를 할 것으로 보인다.

그러나 Noise-free device를 음성인식에 사용하기에는 많은 문제점이 존재한다. 주요한 문제점으로 제안된 주파수 정보와 낮은 포먼트 정보로 들 수 있다. 이러한 문제를 해결하기 위해, 가장 중요한 연구중의 하나는 디바이스의 특성에 적합한 특징벡터를 찾는 것이다. 본 논문은 성대마이크의 신호를 분석하여 성대마이크 신호 분석에 적합한 특징벡터를 제시한다. 그리고 우리는 기존의 알고리즘에 이러한 특징을 반영하여 높은 노이즈 환경에서도 높은 성능을 보이는 명령어 인식기를 개발한다.

2. 관련연구

Noise-free device를 사용한 인식 기술의 연구는 일본과 미국 등에서 부분적으로 이루어지고 있다. Nakajima et al.[4]는 non-audible murmur(NAM) 인식을 위해 피부에 붙이는 형태의 새로운 입력 인터페이스를 제안하였다. 이러한 Stethoscopic microphone은 흡입 디스크와 폴리에스테르 판으로 구성된다. NAM sampling을 사용하여 학습 시킨 결과 자음은 거의 인식되지 않았으며, 모음은 잘 인식되었다. 이러한 문제를 해결하기 위해 자음, 모음에 대해 균형있는 power ratio를 제공하는 최적의 센싱 위치를 제시하였다.

Jou et al.[5]은 성대 마이크를 사용하여 녹음된 soft whisper를 인식하기 위한 다양한 adaptation 기술들을 비교 설명한다. 본 논문에서 사용된 Adaptation들은 Maximum likelihood linear regression, feature-space adaptation 그리고 downsampling, sigmoidal low-pass filter나 linear multivariate regression을 통한 재학습 방법들이 있다. 또한 Zheng et al.[6]은 bone-conductive microphone과 regular air-conductive microphone, In-ear microphone, Throat microphone들을 결합하여 높은 노이즈 환경에서 음성 검출, speech enhancement와 인식률 향상에 대한 연구를 진행하였다.

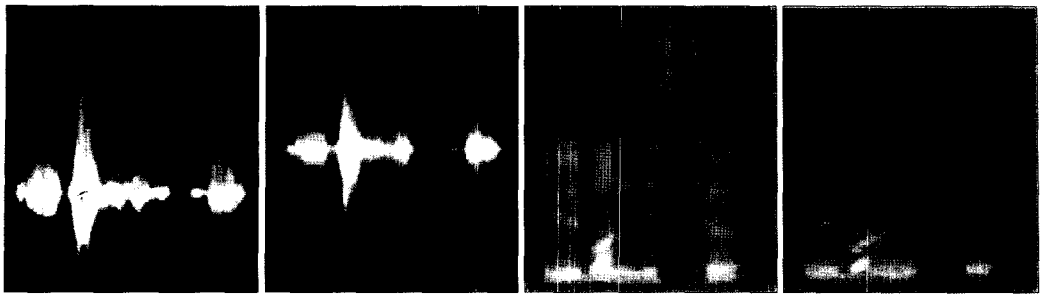
그리고 Dupont, et al. [7] 은 성대 마이크와 일반 마이크의 입력 신호를 동시에 받은 다음 각각의 acoustic models에 의해 제공되는 확률벡터를 결합하여 음성인식의 성능을 향상 시키는 방법을 제안한다. Graciarena et al. [8]은 성대마이크를 일반 마이크의 보완 센서로 사용하여 두 마이크로부터 들어온 입력신호에서 노이즈 Mel-cepstral feature를 추출하여 clean standard microphone mel-cepstral features로 변환하는 알고리즘을 제안하였다. 이와 같이 대부분의 성대마이크를 이용한 연구는 기존의 표준마이크의 노이즈 특징을 보안하는 용도로 사용되어 왔다. 그러나 우리는 성대마이크만을 사용하여 환경 노이즈에 강인한 화자독립 명령어 인식기를 개발한다.

3장에서 성대신호의 특징을 분석하고 한국어의 언어자질을 이용하여 성대신호 분석에 적합한 특징벡터의 조건을 제시한다. 그리고 우리는 이러한 특징벡터의 조건을 만족하는 ZCPA와 음성인식에 널리 사용되는 MFCC와의 성능비교를 통해 제안된 조건의 타당성을 검증한다. 그리고 4장에서 성대신호 분석에 적합한 특징추출 알고리즘을 제안하고 5장에서는 다양한 실험을 통해 MFCC알고리즘이 갖는 성대신호 모델링의 한계점을 보이고, 제안된 조건을 만족하는 ZCPA알고리즘이 성대신호 인식기에 높은 성능을 가짐을 보인다.

3. 성대신호의 특징

3.1 성대신호의 신호적인 특징

성대 마이크를 통해 입력된 신호를 녹음한 후 청취해 보면 어느 정도 잡음이 포함되어 있지만 사람이 분별하지 못할 정도는 아니다. 그림 1은 동일한 명령어에 대해서 일반마이크를 이용한 경우(a)와 성대마이크를 이용한



(a) 일반 MIC (b) 성대 MIC (a) 일반 MIC (b) 성대 MIC

그림 1 일반 및 성대 마이크에 의한 출력

경우(b)를 나타낸 것으로 진폭의 차이는 있으나 두 신호가 비슷한 형태를 보여줌을 알 수 있다. 그러나 스펙트럼 상에서는 매우 큰 차이를 보인다.

그림 1과 같은 스펙트럼상의 정보량의 차이는 ASR시스템을 개발할 때 성능 차이로 나타난다. ASR시스템에서의 성능차이 분석을 위해 음성인식에 널리 사용되는 특징추출 알고리즘인 MFCC를 이용하여 실험한다. 실험환경은 두 개의 마이크를 통해 동시에 발성한 50개의 명령어 셋에 대해 100명분의 학습데이타와 25명분의 테스트데이타를 이용한다. 실험결과 성대마이크를 이용한 경우가 일반마이크를 이용한 경우에 비해 약 30%의 성능 저하가 발생한다. 표 1은 일반마이크와 성대마이크에 대한 MFCC 알고리즘의 성능차이를 비교한 것이다.

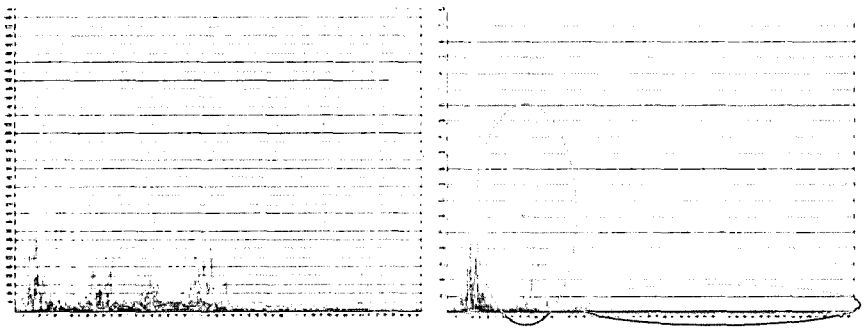
표 1 마이크에 따른 성능 비교

	일반 마이크	성대마이크
인식률 (%)	97.14	67.8

표 1을 통해 MFCC 알고리즘의 한계를 어느 정도 예측할 수 있다. 그림 1에서 알 수 있듯이 성대신호는 고주파 정보가 없고, 저주파 대역에서도 부분적인 포먼트 손실이 발생한다. 만약 MFCC가 인간의 청각이 갖는 민감도를 갖지 못한다면 이러한 성능차이는 부분적으로 설명될 수 있다. 인간의 귀는 140dB SPL의 넓은 입력 신호에 반응하며, 2Hz의 미세한 신호 차이도 감지한다. 그러나 MFCC가 갖는 특성인 band 별 에너지 합으로는 인간 귀가 가지는 민감성을 따라 갈 수 없다.

MFCC의 성대신호 분석에 대한 한계를 보다 정확하게 측정하기 위해 음성신호와 성대신호를 MFCC[9] 알고리즘을 통과시킨 후의 정보량을 분석한다. 그림 2는 16kHz, 16bit의 wave 데이타를 대상으로 Pre-emphasis, Hamming-Windowing을 한 후 FFT 수행 결과를 모든 프레임에 대해서 주파수 영역에서 보여준다.

그림 2에서 가로축은 주파수영역을 256개로 나눈 인덱스이고 세로축은 주파수영역에 포함된 에너지 값이다. 그리고 다양한 색은 개별 프레임을 나타낸다. 그림 2에서 보듯이 2kHz이하에서는 두 신호의 에너지 분포가 비슷한 분포를 보인다. 그러나 2kHz에서부터 성대신호의 정보량이 감소하기 시작되고, 4kHhz이상에서는 거의 정보가 없음을 알 수 있다. 또한 저주파 영역에서도 많은 정보의 손실이 발생됨을 알 수 있다. 따라서 MFCC와 같이 고정된 시간-주파수 분해만으로는 정보의 양이 현저히 떨어지는 성대 특성을 모델링하기에 많은 문제점이 있음을 간접적으로 알 수 있다.



(a) 음성신호의 FFT결과 (b) 성대신호의 FFT결과

그림 2 프레임별 주파수영역에서의 에너지 변화

이러한 문제점을 해결하는 방안으로 Ghitza에 의해 제안된 Ensemble Interval Histogram(EIH)[10] 알고리즘을 제안한다. EIH 알고리즘은 band pass cochlear 필터들로 이루진 청각 모델로써 보다 미세한 신호에 대해서도 반응하는 알고리즘이다.

3.2 음운 자질(Phonological Feature)과 성대신호와의 관계

우리는 앞 절을 통해 성대신호 분석에서의 문제점을 제시하였다. 그리고 이러한 문제점을 해결하는 방법으로 인간의 청각 구조를 모델링 한 알고리즘의 필요성을 강조하였다. 본 절에서는 언어의 특성을 분석하여 성대 신호 분석에 적합한 특징벡터를 제안한다.

발성기관은 폐, 기관, 후두, 인두, 코, 입, 입술등이 있는데, 이들이 일체가 되어 폐에서 입술로 이어지는 복잡한 관강을 형성한다. 인간이 이러한 발성기관을 통해 의미를 갖는 말을 생성하기 위해서는 그 나라의 문자가 갖는 음운적 특징을 오랜 시간 훈련을 해야 가능하다. 이렇듯 문자의 음운적 특징은 특징 분석의 가이드 라인을 제시해줄 수 있다.

한국어는 음소 문자로써 자음과 모음으로 이루어져 있으며 이를 음절단위로 조합해서 글자를 나타낸다. 모음은 총 21개로 모두 유성성의 특징을 갖는다. 자음의 경우 총 19자인데 형태와 위치에 따라 유성음이 되기도 하고 무성음이 되기도 한다. 표2는 한국어 자음의 분류이다.

한국어가 음절을 이루는 원리는 자음+모음+자음 또는 자음+모음, 모음+자음, 모음들 중에 한가지 경우이다. 그리고 이러한 음절은 그 차체로 음운자질을 갖거나 발성할 때 음운자질을 갖게 된다. 음운자질이란 어떤 음운이 갖고 있는 고유의 특성으로 크게 유성성(voiced), 모음성(vocalic)/자음성(consonantal), 성절성(sylabic), 공명성(sonorant)/장애성(obstruent)으로 나뉜다[11-13]. 다음은 음운자질에 대한 개략적 설명이다.

- 유성성 : 유성음과 무성음의 구분을 뜻하는 것으로 성대의 떨림 유무에 대한 자질.
- 모음성/자음성 : 모음과 유성자음을 구분하기 위한 자질로써 모든 모음은 모음성은 가지나 자음성은 없으며 유성자음은 모음성과 자음성을 동시에 가진다. 그리고 그 밖의 자음은 모음성은 갖고 있지 않으나 자음성을 갖는 것으로 구분할 수 있다.
- 성절성 : 음절의 정점을 이루는 분절음의 특징으로 모음이 가지는 대표적인 자질
- 공명성/장애성 : 똑 같은 입의 크기에 따라 소리가 멀리가는 정도를 나타내는 자질.

이러한 음운자질은 발성기관과 밀접한 관계를 가진다. 본 논문은 성대의 떨림과 관계된 자질인 유성성 및 모음성/자음성의 음운 자질을 이용하여 성대신호의 특징을 모델링한다. 표 1의 자음 중에서 유성성을 갖은 자음은 비음과 유음이다. 그리고 나머지는 무성음에 속한다. 그러나 나머지 무성 자음 중에서 성대의 긴장을 통해 발성되는 자음이 있다. 경음(ㅃ,ㄸ,ㅉ,ㅆ,ㄲ)이 이것에 속한다.

다음으로 중요한 특징은 음운 변화이다. [+자음성, -유성성] 을 갖는 (ㅂ,ㄷ,ㄱ,ㅍ,ㅌ,ㅋ,ㄲ,ㅅ,ㅆ,ㅈ,ㅊ,ㅎ) 이 비음인 (ㅁ,ㄴ,ㅇ) 을 만나 [+자음성, +유성성]으로 바뀌는 현상이다. 이것을 비음화라고 한다. 또 하나는 유성음화 현상이다. 이는 유성성의 자질을 갖는 음운들 사이에 무성장애음이 유성음으로 변화는 현상이다. 한국어 자음에서는 (ㄱ,ㄷ,ㅂ,ㅊ,ㅎ)이 유성음들 사이에서 유성음화 (intersonorant obstruent voicing) 된다. 이렇듯 19개의 자음 중에서 9개의 자음이 유성성을 가지며 나머지 10개의 자음도 비음화와 유성음화 현상으로 인해 유성음화 되는 현상이 단어 내에서 빈번히 나타난다. 이러한 한국어의 음운적 특징 때문에 성대마이크만을 이용한 명령어인식기 개발이 가능하다. 또한 이러한 특징을 볼 때 성대신호로의 특징벡터로써 유/무성음 분리에 사용되는 특징벡터- zero-crossing, peak, period pitch, -가 보다 유용 하리라는 사실을 알 수 있다.

유/무성음 분리를 위한 방법으로는 음향학적 특징을 이용한 신호처리기법 기반의 알고리즘이 가장 많이 사용되고 있다[14]. 이러한 시도들은 유성음의 피치주기뿐만 아니라 여러가지 특징 파라미터들을 이용하여 주기성 판단에 의한 유/무성음검출을 향상시키려고 하였다.

본 논문은 민감한 청각 구조를 반영하는 EIH 알고리즘의 특성과 유/무성음 검출에 적합한 특성인 피치와 영교차의 특성을 모두 갖는 ZCPA(Zero-Crossings with Peak Amplitudes)[15]와 band 별 에너지 합을 특징으

표 2 한국어 자음의 분류(Classification of the Korean consonants)

구별요인	양순성	전설성			후설성	후두성
		정지성	파찰성	마찰성		
평 음	ㅂ	ㄷ	ㅈ	ㅅ	ㄱ	
경 음	ㅃ	ㄸ	ㅉ	ㅆ	ㄲ	
격 음	ㅍ	ㅌ	ㅊ		ㅋ	ㅎ
비 음	ㅁ	ㄴ			ㅇ	
유 음		ㄹ				

로 사용하는 MFCC와 성능을 비교하여 앞절에서 분석한 성대신호의 특성이 타당함을 실험을 통해 보인다.

4. 성대신호에 적합한 특징추출 알고리즘

성대 마이크로 입력된 성대신호는 일반 마이크로 입력된 음성신호화는 많은 차이가 있음을 3장을 통해서 설명하였다. 본 장에서는 3장의 분석을 기반으로 음성신호 분석에 널리 사용되는 MFCC 와 유/무성은 검출에 적합한 특성을 갖는 피크와 영교차에 기반하는 ZCPA를 비교하여 성대신호 모델링에 적합한 특징추출 알고리즘을 제안한다. 그리고 좀더 높은 인식성능을 위해 채널 노이즈 제거 알고리즘인 CMS와 RASTA (RelAtive SpecTrA)[16]를 추가하여 보다 높은 인식성능을 갖는 성대신호 명령어 인식기를 개발한다. 먼저 두 알고리즘에 대해 개략적인 설명을 한 후 실험을 통해 성대 신호에 적합한 특징벡터를 규명한다.

4.1 Mel-Frequency Cepstral Coefficient(MFCC)

인간의 청각 시스템의 특성을 반영한 MFCC 알고리즘은 다른 특징들보다 보다 좋은 성능을 제공한다고 보고되며, 일반적으로 Mel-cepstrum은 critical band filters을 사용하여 얻을 수 있다. 인간의 귀가 낮은 주파수 영역에서는 분해능력이 높고 높은 주파수대에서 분해능이 떨어지므로 1KHz이하에서는 선형적으로 filter를 적용하고 그 이상에서는 log 스케일로 필터를 적용한다. 그림 3은 Mel-cepsturm을 얻기 위한 MFCC의 개략적인 구조도이다.

4.2 Zero-Crossings with Peak Amplitudes(ZCPA)

Nonlinear stages와 band pass cochlear filter-bank로 구성된 ZCPA는 upward zero-crossing 시간 사이의 Peak amplitude 값을 특징으로 사용하는 특징 추출

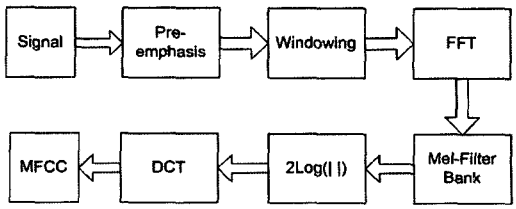


그림 3 MFCC의 처리과정

알고리즘으로 청각 특성을 모델링한 EIH을 기반으로 하고 있다. ZCPA는 zero-crossing을 이용하여 시간 정보를 반영하고, peak 정보를 이용하여 intensity 정보를 반영한다. 그림 4는 ZCPA의 블록 다이어그램이다. 현재 시스템에서 사용하는 ZCPA 알고리즘은 Cochlear filter 필터로 FIR 필터를 사용한다. 그리고 filter bank의 주파수 응답은 16개의 Hamming band-pass window들로 구성되어 있다.

식 (1)은 시간 t에 ZCPA의 출력이다.

$$y(t:i) = \sum_{channel} \sum_{k=1} \delta_{ij_k} f(A_k), \qquad 1 \le i \le N \tag{1}$$

k는 각 channel에서 upward zero-crossing의 수이고, N은 frequency bin의 수이다. 그리고 j_k는 k번째와 (k+1)번째 zero crossing을 이용하여 계산된 frequency bin의 인덱스이고 A_k는 k번째와 k+1번째 사이의 peak amplitude이다. 그리고 δ_{ij}는 Kronecker delta이고 f()는 monotonic function 이다. 여기서는 log 함수가 사용된다.

5. 실험 및 평가

화자독립 성대신호 명령어 인식기 개발을 위한 실험

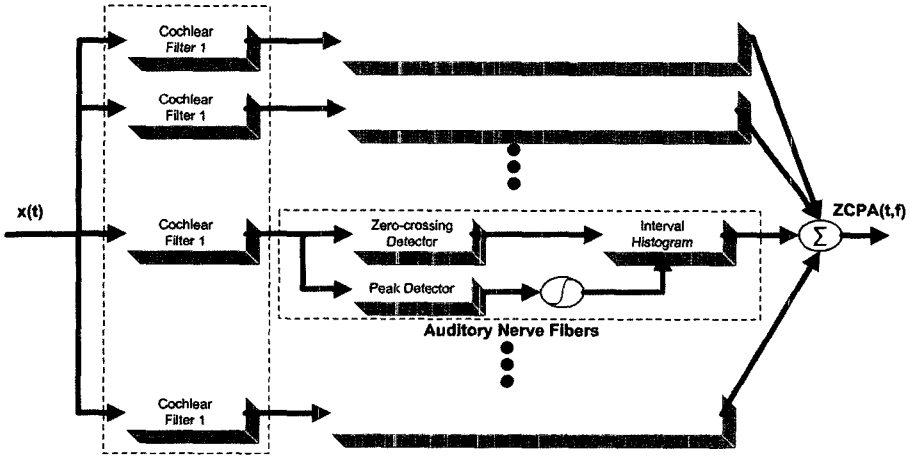


그림 4 ZCPA 블록 다이어그램

환경은 다음과 같다. 동일한 화자에 대해 일반 마이크와 성대 마이크를 통해서 녹음한 100명분 50단어를 학습데이타로 사용하고 25명분 50단어를 테스트데이타로 사용한다. 학습 모델은 소규모 고립어 인식에 높은 성능을 보이는 TDNN을 사용한다. 발성데이타는 16 kHz에 16bit으로 샘플링 된 PCM 데이타를 사용한다.

실험에 앞서 이연철[17]의 넥마이크를 이용한 인식기의 성능을 살펴 본다. 이연철은 넥마이크로 입력된 음성신호에 대한 인식 연구에서 화자종속으로 다양한 특징추출 알고리즘을 대상으로 성능을 테스트하였다. 대상으로 한 특징추출 알고리즘은 LPC와 MFCC, MFCC+CMS 이다. 표 3은 이연철의 넥마이크를 이용한 화자종속 명령어 인식기의 성능 측정 결과이다.

표 3 성대마이크를 이용한 화자종속 명령어 인식기의 특징벡터별 인식률[17]

특징벡터	LPC Cepstrum	MFCC	MFCC+CMS
성대마이크	52.3	50.0	75.0

표 3의 실험 결과를 통해 두 가지 사실을 알 수 있다. 첫째 본 논문의 인식기 모듈이 매우 최적화 되어있음을 알 수 있다. 표 4의 실험 결과와 비교해보면 화자독립인 우리의 시스템이 화자 종속인 표 3의 결과와 거의 비슷한 성능을 갖는다. 둘째 성대신호 분석에 MFCC와 같은 주파수 대역의 에너지 합을 사용한 경우 높은 성능을 내기가 어려움을 알 수 있다. 이연철의 연구 결과는 화자종속임에도 불구하고 인식률이 75%가 넘지 않는다.

본 논문은 두 번째 사실을 보다 정확하게 측정하기 위해 몇 가지 실험을 수행한다. 실험에 앞서 본 시스템 성능의 객관성을 확보하기 위해 음성데이타를 사용하여 각각의 특징추출 알고리즘에 대해 성능을 측정한다. 특징추출 알고리즘은 MFCC와 ZCPA를 상용한다. 음성데이타와 성대 데이타는 동일화자가 동시에 발성한 데이타를 녹음한 것이다. 표 4는 음성신호에 대한 화자독립 명령어 인식기의 인식률 측정 결과이다.

실험 결과 음성데이타에 대해 두 알고리즘은 거의 동일한 성능을 보임을 알 수 있다. 그러나 두 특징추출 알고리즘을 성대신호에 적용 한 경우에는 높은 성능차이가 발생한다.

표 4 음성신호에 대한 화자 독립 명령어 인식기의 인식률

특징추출 알고리즘	MFCC(13차)	ZCPA
인식률 (%)	97.14	98.86

본 논문은 MFCC의 성대신호 모델링 한계를 보다 정확하게 분석하기 위해 다양한 실험을 수행한다. MFCC를 이용한 실험은 두 가지 가능성을 확인하기 위한 실험이다. 첫째 성대신호의 주파수 대역내에 얼마나 유용한 정보가 포함되어있는가? 둘째 MFCC 알고리즘이 갖는 성대신호 분석의 한계점은 무엇인가? 실험은 아래의 방법으로 수행한다.

- MFCC(12차) + Energy(1차)
- MFCC(12차) + Energy(1차) + CMS
- MFCC(12차) + Energy(1차) + Delta(13차) + CMS
- 8 kHz로 down sampli1ng한 후 MFCC(12차) + Energy(1차) + CMS
- 5 kHz이하 주파수 영역에 대해 MFCC(12차) + Energy(1차) + CMS
- Rectangle band-pass filter-bank가 적용된 MFCC 사용
- 선형 band-pass filter 영역 변화 후 특징 추출 (MFCC+CMS)

각 실험 결과는 표 5와 같다.

성대 신호의 주파수 대역내에 있는 정보의 유용성을 실험하기 위해 3가지 실험을 수행한다. 성대 신호가 5 kHz이상의 영역에는 신호의 특성이 나타나지 않으므로 Mel-filer를 5 kHz 이하 영역에 대해서만 24개의 Mel-filer bank를 적용한 경우와 filter의 모양을 Rectangle로 변화시킨 경우, 그리고 1KHz 이하에 적용된 선형적 필터를 1KHz씩 오른쪽으로 shift하면서 어느정도 인식률 변화가 발생하는지 특정한다. 결과는 기존의 MFCC+CMS 보다 성능이 못한 것으로 나타났다. 이러한 실험결과로 성대신호의 주파수 정보는 음성 신호의 주파수 정보의 서브 집합임을 알 수 있다.

두 번째 실험은 MFCC 알고리즘이 갖는 성대신호분

표 5 MFCC를 이용한 성대신호 특징 분석

특징추출 알고리즘	인식률
MFCC(12차) + Energy(1차)	67.8
MFCC(12차) + Energy(1차) + CMS	75.0
MFCC(12차) + Energy(1차) + Delta(13차) + CMS	77.6
8 kHz down sampling 후 MFCC(12차) + Energy(1차)	60.5
5 kHz 이하 주파수 영역에 대해 MFCC(12차) + Energy(1차) + CMS	68.5
Rectangle band-pass filter-bank가 적용된 MFCC(12차)+Energy(1차)+CMS	71.5
선형 band-pass filter-bank의 영역 변화를 통한 성능평가(MFCC+CMS)	70~73

석의 한계점을 관찰하기 위한 것이다. 첫째는 동일한 인식기(TDNN)를 이용하여 음성과 성대 데이타에 대하여 인식률 테스트를 수행하였다. 결과는 약 30%정도 성대 인식기가 낮은 성능을 보였다. 두 번째 실험은 저주파내 정보 손실이 인식기의 성능에 어느정도 영향을 미치는가를 측정한다. 이를 위해 인식 및 학습 데이타를 8kHz로 down-sampling 한 후 성능을 측정한다. 이 실험의 결과 16 kHz 16bit으로 sampling된 데이타를 대상으로 한 경우보다 15% 더 낮은 성능을 보인다. 실험 결과를 통해, 성대신호 인식기의 성능저하의 원인이 고주파 정보의 부재뿐만 아니라, 저주파 대역내 정보 손실 또한 많은 영향을 미침을 알 수 있다.

마지막으로 시간 변화율을 적용했을 때의 인식기 성능을 측정하였다. 이 경우 Delta만을 사용하였는데 약 2%의 성능 향상이 있을 뿐이었다. 앞의 세 실험을 종합해보면 성대신호와 같이 주파수 대역의 정보량이 현저히 낮고, 제안된 주파수 대역 정보를 갖는 신호를 모델링하기에 MFCC 알고리즘과 같이 대역내 에너지 합만을 특징으로 사용하는 알고리즘은 적합하지 않음을 알 수 있다. 본 논문은 이러한 한계점을 극복하고 성대마이크만을 사용한 ASR시스템을 개발하기 위한 특징추출 알고리즘의 조건을 제시한다.

- 인간의 청각과 비슷한 민감한 band-pass 필터가 필요하다. 그림 2의 주파수 분석을 통해 알 수 있듯이 FFT와 같은 band 별 에너지만을 가지고는 부족한 정보를 갖는 성대신호로부터 유용한 정보를 추출할 수 없다.
- 유성음 추출에 적합한 특징벡터를 사용 할 필요가 있다. 한국어의 언어 자질 분석에서 성대신호는 유성음의 분석이 매우 중요하다. 따라서 특징추출시 유성음 분석을 강화함으로써 성능을 보다 높일 수 있다.

본 논문은 제안된 조건을 갖춘 ZCPA 알고리즘으로 성대신호 명령어 인식기를 개발한다. 표 6은 ZCPA 알고리즘을 성대신호에 적용한 결과이다. 표 4와 비교해보면, MFCC에 비해 약 16%정도의 높은 성능을 보임을 알 수 있다. 그리고 채널 노이즈 제거를 위해 RASTA를 적용한 경우 약 2%의 성능 향상을 보인다.

또한 성대신호 분석에 대한 MFCC와 ZCPA 특징 추출 알고리즘의 성능차이는 명령어 집합 분석에서도 나타난다. 본 실험에 사용된 명령어 집합 중에서 "일번리스트"와 "이번리스트", "삼번리스트"와 "사번리스트", "일번항목"과 "이번항목", "삼번항목"과 "사번항목" 등의 8개 명령어 집합의 경우 ZCPA가 MFCC에 비해 높은 성능을 보인다. 이러한 결과의 원인은 본 논문 3장에서 제시한 한국어 음운자질에서 설명된다. "일번리스트"의 "일" 에서의 "ㄹ"은 유성 자음에 속한다. 그러나 "ㄹ"은 발음 특성상 주파수 영역에서 매우 약하게 나타난다. 따라서 "일번리스트"와 "이번리스트"를 구분하기 위해서는 매우 약하게 나타나는 유성자음을 인식할 수 있는 특징벡터를 추출해야한다. 그러나 MFCC 알고리즘은 예시한 단어의 형태로 포함된 "ㄹ"이나 "ㅁ"을 분석하기에 ZCPA보다 적합하지 않음을 실험을 통해 알 수 있다. 위의 결과를 통해 본 논문에서 제시한 성대신호 분석을 위한 조건이 타당함을 알 수 있다.

표 6 ZCPA를 이용한 화자독립 성대신호 명령어 인식기의 인식률

특징추출 알고리즘	ZCPA	ZCPA+RASTA
인식률	83.63	85.16

마지막으로 그림 5는 특징추출 알고리즘에 따른 화자독립 성대신호 명령어 인식기의 인식률 측정결과이다. ZCPA에 RASTA를 적용한 경우에 MFCC에 CMS를 적용한 경우보다 약 10%의 높은 인식률을 보인다.

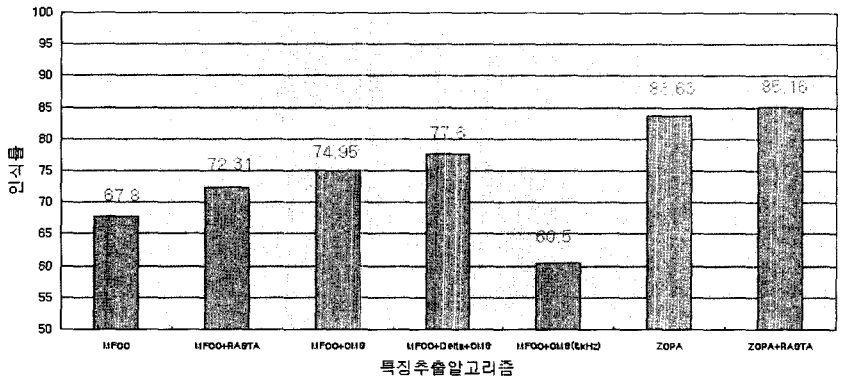


그림 5 특징추출 알고리즘에 따른 성대신호 명령어 인식기의 성능

5. 결 론

본 논문은 노이즈를 원천적으로 차단하는 성대 마이크를 이용하여 화자독립 명령어 인식기를 개발한다. 그러나 성대 마이크의 출력신호를 스펙트럼상에서 분석해 보면 일반마이크의 출력에 비해 포먼트 정보가 현저히 떨어지고 고주파 정보가 거의 나타나지 않은 특징을 보인다.

본 논문은 성대신호의 주파수 대역에서의 특성을 관찰하기 위해 음성인식에 널리 사용되는 MFCC 알고리즘을 이용하여 실험한다. 이러한 실험 결과 성대신호 내에 포함된 정보는 음성신호의 서브 집합임을 알 수 있고, 제안된 주파수와 포먼트의 부분적 손실을 갖는 신호를 분석하기에 MFCC와 같은 band별 에너지합을 사용하는 특징추출 알고리즘으로는 적합하지 않음을 다양한 실험으로 보였다. 그리고 이를 기반으로 성대신호분석에 적합한 알고리즘의 조건을 제시한다. 제안된 조건은 FIR필터와 같은 보다 민감한 주파수 분석이 가능한 필터가 필요하며, 유성음 특징을 추출할 수 있는 특징벡터를 사용하는 것이 좋다. 이러한 접근 방법의 검증을 위해 제안된 조건을 가장 많이 만족하는 ZCPA 알고리즘을 이용하여 명령어 인식기를 개발한다.

실험결과 MFCC보다 16% 높은 인식률을 보였다. 그리고 채널 노이즈 제거를 위해 RASTA 필터를 적용한 결과 85.16%의 높은 성능을 보였다. 이것은 MFCC에 CMS를 적용한 결과보다 약 10%높은 성능이다.

참 고 문 헌

[1] S. F. Boll, "Suppression of acoustic noise speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-27, 113-120, Apr., 1979.

[2] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Processing*, 28, 137-145, Apr. 1980.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, 33, 443-445, Apr. 1985.

[4] Nakajima. Y, Kashioka. H, Shikano. K and Campbel. N, "Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin," ICASSP'03, volume 5, pp 708-11, 2003.

[5] S. C. Jou, T. Schultz, and A. Waibel, "Adaptation for Soft Whisper Recognition Using a Throat Microphone," in Proc. ICSLP, Jeju Island, Korea, Oct 2004.

[6] Zhengyoun Zhang, Zicheng Liu, Sinclair. M, Acero. A, Li Deng, Droppo. J, Xuedong Huang. Yanli Zheng, "Multi-sensory microphones for robust speech detection, enhancement and recognition," ICASSP'04, page: iii-781-4 vol.3, May 2004.

[7] S. Dupont, C. Ris, 2004, "Combined use of close-talk and throat microphones for improved speech recognition under non-stationary background noise," proc. of Robust 2004 (Workshop (ITRW) on Robustness Issues in Conversational Interaction), Norwich, Aug. 2004.

[8] M. Graciarena. H. Franco, K. Sonmez, H Bratt, "Combining Standard and Throat Microphones for Robust Speech Recognition," in IEEE Signal Processing Letters, Vol. 10 No. 3, pp. 72-74, March 2003.

[9] Donghoon Hyun, Chulhee Lee, "Optimization of mel-ceptrum for speech recognition," IEEE SMC '99 Conference Proceedings Volume 1, pp. 500-503, Oct. 1999.

[10] O. Ghitza, "Auditory models and human performances in tasks related to speech coding and speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 1, part II, pp. 115-132, 1994.

[11] 구현옥, *국어 음운학의 이해*, 한국문화사, 1999.

[12] 정경일 외, *한국어의 탐구와 이해*, 박이정출판사, 2000.

[13] 신지영, 차재은, *우리말 소리의 체계: 국어 음운론 연구의 기초를 위하여*, 한국문화사, 2003.

[14] C. K. Un and S. C. Yang, "A Pitch extraction algorithm based on LPC inverse filtering and AMDF," *IEEE Trans. Acoust., Speech Signal Processing*, ASSP-25, 565-572, Dec. 1977.

[15] Doh-Suk Kim, Soo-Young Lee, Rhee M. Kil "Auditory Processing of Speech Signals for Robust Speech Recognition in Real-Word Noisy Environments," IEEE Tran. Speech and Audio Processing, vol., 7 No. 1, Jan., 1999.

[16] H. Hermansky and N. Morgan, "RASTA processing of speech," IEEE Trans. Speech Audio Processing, vol. 2, pp. 578-589, Oct. 1994.

[17] 이연철, 이상운, 홍훈섭, 한문성, 마평수, "넥마이크로 입력된 음성 신호에 대한 인식 연구", 제 18회 한국정보처리학회, 제9권 제2호, 2002.

정 영 규

2000년 2월 신라대학교 전자계산학과(이학사). 2006년 6월 경북대학교 컴퓨터공학과(박사수료). 2004년~현재 전자통신연구원 연구원. 관심분야는 자연어처리, 음성인식, 멀티모달 인식

한 문 성

1977년 2월 서울대학교 수학과(이학사) 1988년 8월 인디아나 대학교 전산학과(박사과정). 1988년~현재 전자통신연구원 책임연구원. 관심분야는 음성인식, 멀티모달 인식, 뇌공학

이 상 조

1974년 2월 경북대학교 수학교육과(이학사). 1976년 2월 한국 과학 기술원(이학석사). 1993년 2월 서울대학교 컴퓨터공학과(공학박사). 1976년~현재 경북대학교 컴퓨터공학과 교수. 관심분야는 언어처리, 지식처리, 정보검색, 기계학습, 시멘틱웹