성대 신호를 이용한 인식 시스템

RECOGNITION SYSTEM USING VOCAL-CORD SIGNAL

조 관현*, 한 문성**, 박 준석***, 정 영규****
Kwanhyun Cho, Munsung Han, Junseok Park, Younggyu Jeong

Abstract - This paper present a new approach to a noise robust recognizer for WPS interface. In noisy environments, performance of speech recognition is decreased rapidly. To solve this problem, We propose the recognition system using vocal-cord signal instead of speech. Vocal-cord signal has low quality but it is more robust to environment noise than speech signal. As a result, we obtained 75.21% accuracy using MFCC with CMS and 83.72% accuracy using ZCPA with RASTA.

Key Words : Speech recognition, Vocal-cord signal, Feature extraction, Wearable computer

1. 장 서론

음성인식 기술은 사용자에게 좀 더 편하고 자연스러운 인터페이스를 제공하며 최근 여러 가지 서비스에 사용되고 있다. 하지만 이런 장점에도 불구하고 실생활에 폭넓게 활용하기에는 아직 해결해야 할 몇 가지 문제점을 가지고 있다. 그 문제점 중 하나는 다양한 잡음에 의한 인식 성능의 저하이다. 이러한 잡음에 의한 성능저하를 해결하기 위해 음질 향상과 신호원 분리법, 잡음에 강인한 특징벡터 추출, 채널잡음 제거법 등의 여러 가지 연구가 시도되었다[1]-[4]. 본 연구에서는 이러한 문제를 해결하기 위해 음성 신호 대신 성대 신호를 이용하여 인식하는 방법을 제안하고 그 가능성을 실험을 하였다. 성대신호는 음성신호에 비해 성도를 거치면서 발생하는 공진의 영향을 별로 받지 못하기 때문에 신호의 명료성이 떨어지나 주변 잡음에 의해 신호의 왜곡을 거의 받지 않기 때문에 안정적인 인식성능을 보장할 수 있다. 본 연구에서는 MFCC(Mel-Frequency Cepstral Coefficients)와 ZCPA(Zero-Crossing with Peak Amplitudes)의 특징 벡터와 잡음처리를 위해 각각 CMS(Cepstral Mean Subtraction)와 RASTA(RelAtive SpecTrAl)를 사용하여 성대신호를 이용한 인식기에 대한 성능을 실험하였다.

본 논문의 나머지 부분은 다음과 같이 구성되었다. 먼저 2장에서는 VCSR(Vocal-Cord Signal Recognizer)의 시스템 구조와 동작 모드에 대해 알아보고 실험에 사용된 두 가지 종류의 특징벡터를 추출하는 과정은 3장에서, 인식률 실험과 그 결과를 4장에서 살펴본다. 마지막으로 본 연구의 결론과 차후 연구 계획에 대해 알아본다.

저자 소개
* 한국전자통신연구원 차세대 PC연구그룹 연구원
** 한국전자통신연구원 차세대 PC연구그룹 책임연구원
*** 한국전자통신연구원 차세대 PC연구그룹 팀장
****한국전자통신연구원 차세대 PC연구그룹 연구원

2. 장 SYSTEM ARCHITECTURE

WPS(Wearable Personal Station)[5]를 위한 입력방법으로 사용되는 VCSR은 시스템 설정이나 응용프로그램 제어와 같은 서비스를 제공한다. Fig. 1은 VCSR의 시스템 구조로 standalone mode와 server-client mode로 구성된다. Standalone mode는 WPS의 시스템 설정을 위한 것으로 한번에 40~50개의 단어를 인식 대상으로 한다. Server-client mode는 응용 프로그램 제어를 위해 사용되며 수백 개의 단어를 인식대상으로 한다. 이 경우 WPS는 client가 되어 특징 추출 및 잡음처리, 끝점검출의 작업을 수행하며 인식과정은 remote server에서 계산되어 결과가 client로 전송된다.

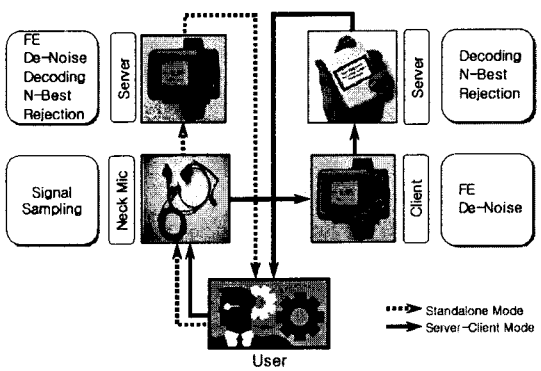

Fig. 1 System architecture overview

3. 장 특징추출

Neck microphone으로부터 획득한 성대신호는 특징 추출 과정과 잡음처리 과정을 거쳐 특징 벡터를 구성한다. 본 실험에서는 8KHz, 16bit Linear PCM 으로 sampling된 성대신호에 대해 다음과 같이 두 가지 종류의 특징벡터를 사용하여 성능을 비교해 보았다.

3.1 절 MFCC

MFCC는 일반적으로 음성인식에서 많이 사용되는 특징벡터로 비교적 좋은 성능을 나타낸다. Fig. 2는 MFCC를 추출하는 과정을 나타내는 블록 다이어그램이다.

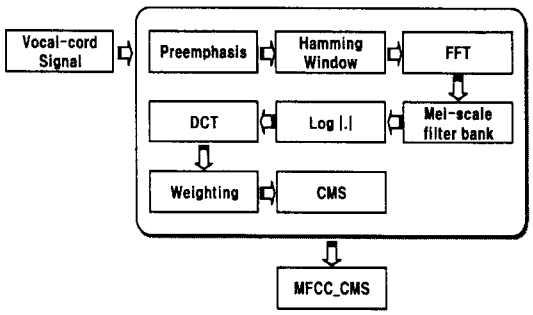

Fig. 2 The overall diagram of MFCC extraction

MFCC 특징벡터를 추출하는 과정은 다음과 같다. 먼저 preemphasis를 수행하고 Hamming window가 적용된 20ms 프레임의 성대 신호에 대해 단구간 분석을 한다. 매 10ms마다 26개의 mel-scale 필터뱅크 에너지를 계산하여 log를 취하고 DCT를 수행한다. 이와 같은 과정을 통해 추출된 13차 MFCC를 기반으로 delta-MFCC를 계산하여 26차 MFCC 특징벡터를 구성한다. 또한 잡음처리에 대한 실험을 위해 CMS을 적용하여 각각의 발성에 대해 두 가지의 특징벡터를 추출한다.

3.2 절 ZCPA (Zero-Crossing with Peak Amplitudes)

인간의 청각 지각을 기반으로 하는 ZCPA 모델[6]은 다른 청각 모델에 비해 간단하지만 잡음에 강인한 특성을 보인다. ZCPA 모델은 Fig. 3에서처럼 대역통과 와우각 필터뱅크와 비선형 단계(nonlinear state)로 구성되어 있다. 와우각 필터뱅크는 와우각 안에 있는 기저막에 따라 다양하게 위치한 주파수 민감도를 나타내며 해밍 대역통과 필터들로 구현된다. 청각 신경섬유는 자극을 받음과 동시에 반응하며 이러한 반응은 각 대역통과 필터의 출력신호의 상향하는(up-going) 영교차로 나타나고 연속적인 교차점간의 간격은 주파수 히스토그램에 적용된다. 인접한 신경발화 간격의 역(inverse)은 모아져서 주파수 히스토그램으로 표현된다. 또한 연속적인 영교차 간의 각각의 극대치가 탐지되면, 이 극대치는 흥분발화율(firing rate)을 나타내기 위한 주파수(frequency bin)에 대한 비선형 가중 요소로 이용된다. 모든 필터 채널의 히스토그램은 결합되어 이 청각 모델의 결과값을 나타내게 된다.

임의의 시간 t에서의 ZCPA 모델의 결과값은 다음과 같이 표시된다.

$$y(t,i) = \sum_{channel} \sum_{k=1}^{K-1} \delta_{ij} f(A_k),\ 1 \le i \le N \qquad (1)$$

여기서 K는 각 필터 채널에서의 상향하는 영교차점 수를 나타내고 N은 주파수 영역(frequency bin)의 수를 나타낸다. j_k는 k번째와 $(k+1)$번째의 영교차점을 이용하여 계산된 주파수 영역이고 A_k는 k번째와 $(k+1)$번째 사이의 극대치이다. δ_{ij}는 크로네커 델타(Kronecker delta) 함수이고 $f()$는 자극강도와 청각 신경세포의 위상고정(phase-locking) 정도 간의 관계를 나타내기 위한 monotonic 함수이다.

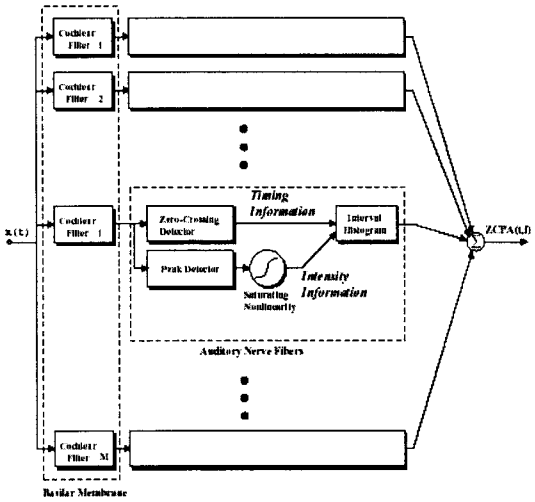

Fig. 3 The overall diagram of ZCPA extraction

Fig. 3은 ZCPA의 추출과정을 나타내는 블록 다이어그램이다. Fig. 3의 과정에 의해 계산된 ZCPA는 16차의 ZCPA 와 RASTA filtering을 거친 16차 ZCPA_RASTA를 구성한다.

4. 장 실험 및 결과

4.1 절 Experimental setup

본 실험에 사용된 Database는 300명의 남성 화자에 의해 발성된 80,000개의 단어로 구성된 학습 데이터와 60명의 남성 화자에 의해 발성된 1,6000 단어의 테스트 데이터로 구성되었다. 성대신호는 넥마이크를 사용하여 8Khz, 16bit로 샘플링된다. 한 프레임은 샘플링된 160개의 데이터로 구성되고 80개 샘플을 중첩하여 매 프레임을 구성한다. 단구간 분석을 위해 사용하는 윈도우의 크기는 256이고 26개의 mel-scale 필터뱅크를 사용하였다.

MFCC_0_D, ZCPA의 두 종류 특징벡터와 각 특징벡터에 잡음처리를 적용한 MFCC_0_D_Z, ZCPA_RASTA에 대해 성능 비교 실험을 수행하였다. 인식실험은 HTK ver3.1을 사용

하여 수행했고 Tri-phone 기반 HMM 모델을 사용하였으며 각 모델은 5개의 state로 구성되었고 한 state당 3개의 Gaussian Mixture 분포를 사용하였다.

4.2 절 Experimental results

Fig. 4는 잡음처리를 하지 않은 26차 MFCC와 16차 ZCPA를 사용하여 인식 실험을 한 결과이다. 테스트 데이터에 대한 인식률은 MFCC 26차에 대해서는 55.16%, ZCPA 16차에 대해서는 64.0%의 인식률로 모두 70%이하로 일반적인 조용한 환경에서의 음성인식이 90% 이상의 인식률을 보이는 것에 비하면 상대적으로 저조한 인식률을 보인다. 이것은 성대신호 자체의 품질이 일반 음성신호에 비해 많이 떨어져 모델의 변별력이 감소하기 때문이다.

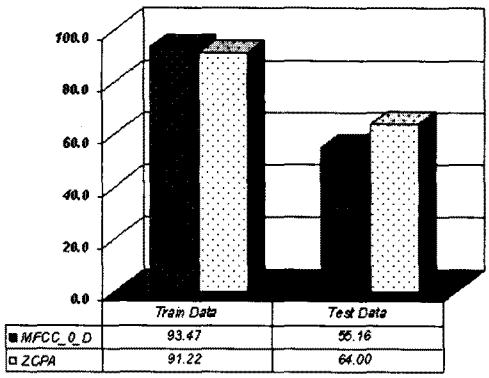

Fig. 4 Recognition result without noise filtering

Fig. 5.는 MFCC와 ZCPA를 추출하는 과정에서 잡음제거를 위해 각각 CMS와 RASTA를 적용하여 추출한 특징벡터에 대한 인식 실험 결과이다. CMS를 적용한 MFCC의 경우는 75.21%, RASTA를 적용한 ZCPA는 83.72%의 성능을 보였다. 잡음처리를 적용하지 않은 결과에 비해 약 20%정도의 인식률이 향상되었다. 이러한 결과는 성대신호가 주변잡음에 영향을 받지 않지만 성대신호 자체에 상당한 채널 잡음이 존재함을 알 수 있다. 또한 이러한 채널 잡음은 일반적으로 음성인식에서 사용되는 잡음제거 알고리즘을 사용해도 충분히 효과가 있음을 알 수 있다.

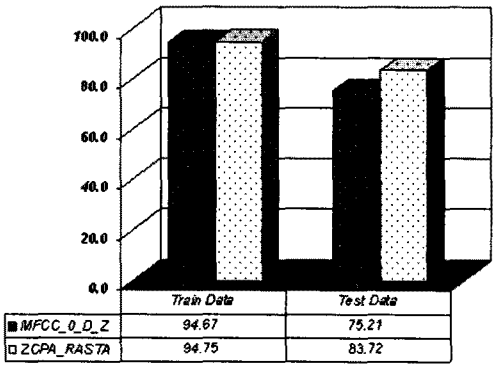

Fig. 5 Recognition result with noise filtering

Fig.4와 fig.5의 결과에서 알 수 있듯이 성대신호를 이용한 인식기는 음성인식에 비해 다소 낮은 인식률을 보이지만 주변 소음이 심한 환경에서는 음성인식의 성능저하에 대한 문제를 해결해 줄 수 있는 대안이 될 수 있을 것이다.

5. 장 결론

본 논문에서는 주변잡음에 강인하고 안정적인 인식 성능을 보장하기 위해 성대신호를 이용한 인식모델을 제안하고 인식 실험을 통해 가능성을 제시하였다. 실험 결과, 음성인식에서 주로 사용되는 MFCC와 CMS를 사용하여 75.21%의 인식률을 얻었고 잡음에 좀 더 강인하다고 알려져 있는 ZCPA와 RASTA를 사용하여 83.72%의 인식률을 얻었다.

이러한 인식률은 조용한 환경에서 음성을 이용한 인식률에 비하면 다소 떨어지지만 주변잡음이 심한 거리나 자동차 내부, 공장 등과 같은 환경에서는 음성인식의 성능저하를 대체할 수 있는 방법이 될 수 있다.

앞으로 MFCC나 ZCPA보다 좀 더 성대신호를 잘 표현할 수 있는 특징추출 방법에 때한 연구와 WPS와 같은 소형 단말에 적합한 인식기의 개발이 병행될 것이다.

참 고 문 헌

[1] S.F.Boll: "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", IEEE Trans. ASSP, Vol.27, No.2, pp.113-120, 1979.

[2] J.C.Segura, A.de la Torre, M.C.Benitez, and A.M. Peinado: "Model Based Compensation of the Additive Noise for Continuous Speech Recognition. Experiments using AURORA II Database and Tasks," EuroSpeech' 01, Vol.I, pp.221-224, 2001.

[3] ETSI ES 202 050 V1.1.1, "Distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," 2003.

[4] P.J.Moreno, B.Raj, and R.M.Stern: A Vector Tylor Series Approach for Environmental-Independent Speech Recognition," ICASSP'96, pp.733-736, 1996.

[5] 정보통신연구진흥원, "IT839 전략 기획보고서-차세대 PC," 2004.6.

[6] D.kim. S.Lee, and R.M.Kil. Auditory processing of speech signals for robust speech recognition in real-world noisy environments. IEEE Transactions on Speech and Audio Processing, 7(1):55-69, 1999.

[7] 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.

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

[9] T.Sreenivas and R.Niederjohn. Zero-crossing based spectral analysis and svd sepctral analysis for formant frequency estimation in noise. IEEE Transactions on Signal Processing, 40(2): 282-293, 1992.