한국어 음성 분할을 위한 특징 검출에 관한 연구

이 극*, 황 희융
서울대학교 공과대학 전자계산기공학과

A Study on the Feature Extraction for the Segmentation of Korean Speech

Lee Geuk, Hwang Hee Yeung
SEOUL NATIONAL UNIVERSITY

< ABSTRACT >

The speech recognition system usually consists of two modules, segmentation module and identification module. So, the performance of the system heavily depends on the segmentation accuracy and the segmentation unit. This paper is concerned with the agreeable features for segmentation in syllables. Total energy and two band width energy (LE:4000-5000Hz and HE:900-3100Hz) are suitable cues for segmentation. And we testify it through the experiment using connected digit.

1. 서론

디지틀 컴퓨터가 발전됨에 따라 인간과 기계 사이의 가장 자연스러운 대화의 매체인 음성을 이용하고자 하는 요구가 점점 커지고 있으며 디지틀 신호 처리 기술과 반도체 기술의 급속한 성장에 따라 음성인식에 대한 연구와 그 응용이 활발히 추진되고 있다. 음성인식 시스템은 1950년대에 나타난 최초의 모음 인식 시스템[1]에서 시작해서 지난 30여년간 많은 사람에 의해 꾸준히 진전되어 왔으며 일부 성공에도 불구하고 아직도 해결해야 될 산적한 문제에 직면해 있는 실정이다[2].

음성 인식은 인식하고자 하는 음성의 종류에 따라 고립단어 인식(isolated word recognition), 연결단어 인식(connected word recognition), 연속단어 인식(continuous word recognition)의 3가지로 분류되며 화자의 수에 따라 화자종속 인식(speaker dependent recognition)과 화자독립 인식(speaker independent recognition)으로 나눌 수 있다[3]. 고립단어 인식이란 단어들의 간격이 완전히 분리되어 발음된 단어를 인식하는 것을 말한다. 연결단어 인식은 비교적 제한된 어휘에 속한 단어들이 연결되어 발음된 것을 각 단어별로 구분해 인식하는 것이다. 이에 반하여 연속음성 이해 시스템은 자연스럽게 발음된 문장을 인식하여 그 의미까지 파악하는 과정을 포함하는 작업이다. 따라서 연속음성 이해를 위해서는 발음된 음성을 음소나 음절등의 기본 단위로 나누어서 이들을 식별한 다음, 최종적으로 언어학적 분석을 통해 의미를 찾아내는 복잡한 절차를 거쳐야한다.

2. 분할의 문제점

연결 혹은 연속 단어 인식 시스템에 있어서 인식 과정은 크게 분할(segmentation)과 식별(identification)의 두 단계로 구분할 수 있다. 분할이란 인식하기 위한 기본 분절로 음성신호를 자르는 과정을 말하며, 식별은 인식단위로 분할된 분절이 어떤것인가를 찾아내는 것을 말한다. 분할시 문제가 되는것은 어떤 단위로 분할하는 것이 분할을 쉽게할 수 있으며 또 어떤 단위로 분할된 분절이 더 쉽고 정확하게 인식 될 수 있느냐 하는 것이다. 대표적인 인식의 단위로서는 음소, 이음, 음절, 단어들이 있다.

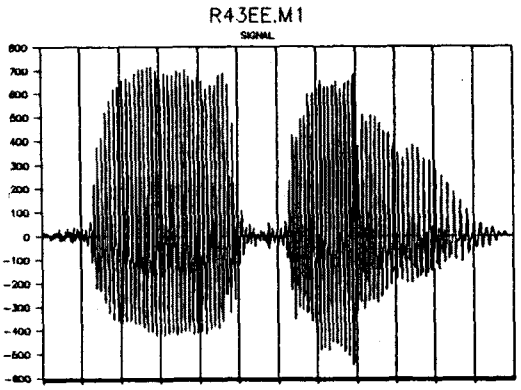
음소를 인식단위로 선정할 경우 유리한점은 한 언어에서 구별 될 수 있는 음소의 총 수는 많아야 50을 넘지 못한다는 점

이다. 그러나 같이 표기되는 음소들도 이웃 음소들에 의해 강하게 영향을 받아 각기 다르게 발음되므로 실제 발음된 음성의 음소 분절들은 음소와 일대 일 대응 관계가 성립되지 않으며 더우기 발음된 음성을 음소별로 정확히 분할하는 것은 거의 매우 어려운 일이다[4]. 음절을 인식단위로 사용하는 경우 음소 인식에서 나타나는 화자의 불규칙성, 감소(reduction), 동화(assimilality), 상호조음(coarticulation) 현상 등에는 덜 민감하나 음절 자체도 이웃 음절에 영향을 받으므로 유성음이나 무성음이 연속적으로 발음될때 분할의 어려움이 역시 존재한다[5]. 우리나라의 경우 실제 사용되는 음절의 수는 약 1096개에 이르고 있다[6]. 단어인식의 경우 이웃간의 영향은 거의 미미하나 가능한 단어의 수는 거의 무한대로 많아지므로 사용 어휘가 제한되어야 하며 표준 패턴이 커져서 기억 장소를 많이 차지하고 계산량도 증가하게 된다. 본 논문은 음절의 연결상태와는 독립적으로 작용할수 있는 음절분할 단서(cue)를 찾는데 그 목적이 있다.

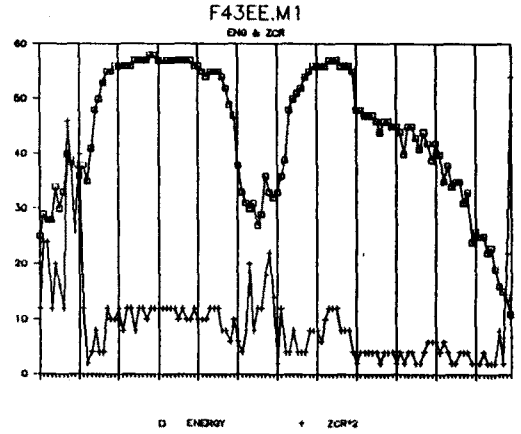
3. 음성 시료 및 특징 분석

음성시료는 가장 보편적으로 널리 쓰이는 음성인 숫자음(공 - 구)을 택하였으며 1인의 남성화자가 발음한 2자리 숫자음 100여개와 3자리 숫자 80여개를 사용하였다.

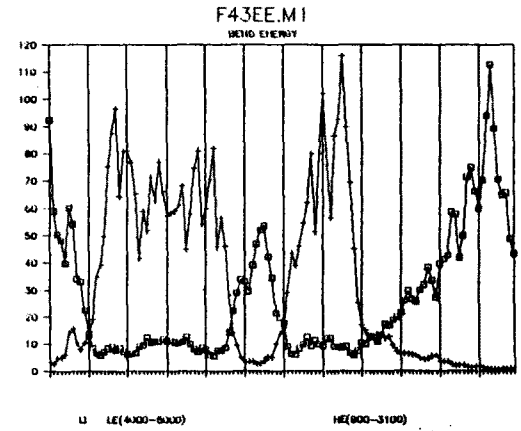
음향분석을 위해 대수 에너지와 FFT(Fast Furier Transform) 결과로 얻어진 중대역(HE: 900 - 3100Hz 사이)와 고대역(LE: 4000 - 5000Hz 사이)의 에너지를 사용했다. 특히 고대역 에너지의 peak값과 중대역 에너지의 dip이 대수 에너지와 함께 음절분할을 위한 주요한 특징으로 나타남을 그림1과 2에서 볼 수 있다. 입력된 음성은 각 프레임의 대수 에너지가 30을 넘으면 일단 음성이 있는 부분으로 간주하며 몇프레임 이상 30 이하로 떨어지면 음성이 종료된 것으로 판단한다. 음성 지속 부분이 시작점에서 부터 40 프레임 이상 계속되면 음절의 종료가 가까워진다고 생각하고 40에서 70 프레임 사이에서 대수 에너지의 dip들, 고대역 에너지의 peak들, 중대역 에너지의 dip들이 존재하는지 조사해서 일단 분할의 후보점들로 등록해 둔다. 등록된 후보점들은 각 dip이나 peak의 날카로운 정도(sharpness)와 근처 다른 dip이나 peak과의 상대적인 차이를 보고 가장 타당한 한 점을 음절의 경계로 선택한다. 그림1의 경우는 대수 에너지 만으로 음절의 시점과 종점을 결정할 수 있는 경우이며 그림2는 고대역의 peak나 중대역의 dip에 의해 분할점이 나타나는 경우이다.



(A) 음성 신호

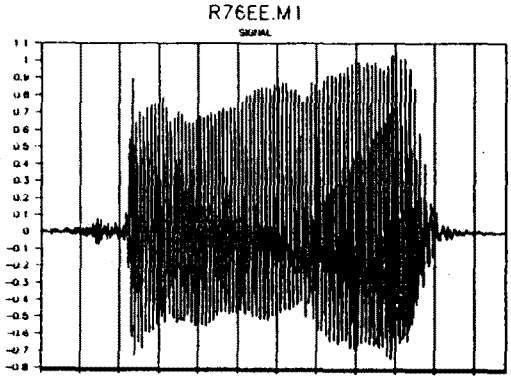


(B) 대수 에너지와 영교차율

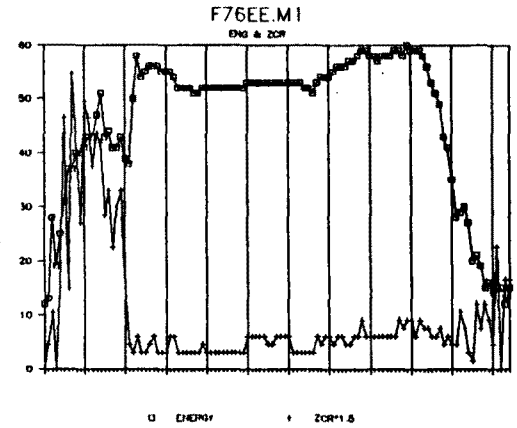


(C) 중대역과 고대역 에너지

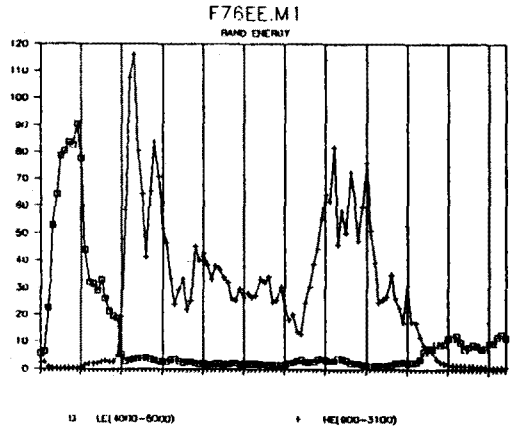
그림1. 음성 '사삼'의 그래프들



(A) 음성 신호



(B) 대수 에너지와 영교차율



(C) 중대역과 고대역 에너지

그림2. 음성 '철욱'의 그래프들

4. 결론 및 앞으로의 연구 방향

본 연구에 의하면 FFT에서 유도되는 중대역 에너지의 dip과 고대역 에너지의 peak가 음절분할에 중요한 단서가 됨을 보여준다. 또한 중대역 에너지는 유성음 부분에 집중적으로 나타나며 고대역 에너지는 무성음이나 묵음 부분에 집중되고 있음을 나타내고 있다. 이들의 특징은 앞으로 연구가 되어야 할 음소 단위 분할의 가능성을 시사하고 있다. 그러나 음소 분할을 위해서는 정교한 판단결정 정책(decision making strategy)가 필요할 것으로 예측되므로 인공지능언어나 기법의 도입이 필요할 것이다.

5. 참고문헌

[1] D.Raj Reddy, "Speech Recognition by Machine : A Review," Proc. of IEEE, Vol.64, No.4, pp501 - 531, Apr. 1976.

[2] 이 극, 김 원준, 황 희웅, "편향 영교차율과 에너지를 이용한 한국어 단음절어 인식에 관한 연구," 대한전기학회 전산기연구회 추계 학술 발표 논문집, pp.16 - 19, Nov. 1986.

[3] Wayne A. Lea, Trends in Speech Recognition, Prentice - Hall, 1980.

[4] G. Mercier and et al., "Automatic Segmentation, Recognition of Phonetic Units and Traning in the Keel Speech Recognition System," IEEE Int. Conf. ASSP, pp 2000 - 2003, 1982.

[5] 이 길행, 한국어 인식을 위한 효과적인 음절 분할에 관한연구," 한국과학 기술원 석사 학위 논문, 1986.

[6] 허 웅, 국어 국문학, 샘문화사, 1985.

[7] 정 영조, 한국어 연결 단위를 위한 인식 단위의 연구, 한국과학 기술원 석사 학위 논문, 1987.

[8] G.Bristow, Electronic Speech Recognition, Collins, 1986.

[9] P.Regel, "A Module for Acustic - Ponetic Transcription of Fluent Spoken German Speech," IEEE Tran. on ASSP, Vol.30, No.3, pp440 - 550, Jun.1982.

[10] 오 영환, "단어 음성의 시점, 종점, 결정 및 유성, 무성, 무음 분류 알고리즘 , 정보과학회 논문지, Vol.12, No.1, pp 8 - 15, Feb. 1985.