주파수 영역에서의 고립단어에 대한 음성 특징 추출

조영훈, 박은명, 강홍석, 박원배
경북대학교 정보통신학과
전화 : (053) 940-8897 / (053) 950-5508

Speech Feature Extraction for Isolated Word in Frequency Domain

Young Hoon Cho, Eun Myung Park, Hong Suk Kang, Won Bae Park
Department of Information & Communication Kyungpook National University
E-mail : yhcho@inc.knu.ac.kr

Abstract

In this paper, a new technology for extracting the feature of the speech signal of an isolated word by the analysis on the frequency domain is proposed. This technology can be applied efficiently for the limited speech domain.

In order to extract the feature of speech signal, the number of peaks is calculated and the value of the frequency for a peak is used. Then the difference between the maximum peak and the second peak is also considered to identify the meanings among the words in the limited domain. By implementing this process hierarchically, the feature of speech signal can be extracted more quickly.

I. 서론

인간이 정보전달을 위해 가장 효과적이고 보편적이며 편리한 수단은 음성이다. 이러한 음성은 인간과 기계간의 정보전달 매개체로서 역할과 이용이 증대되고 있다. 또한 최근 여러 응용분야에서 음성인식기술이 다양하게 사용되어지고 있다.

기존의 음성인식 시스템에 대한 연구에는 음성의 시간적 변화를 모델링하는 천이확률과 스펙트럼 변화를 모델링하는 출력확률로 구성하여 주어진 모델과의 확률적인 추정값을 사용하여 유사도를 계산하는 HMM(Hidden Markov Model)[1]을 이용하거나, 기준이 되는 음성신호의 패턴과 입력된 음성신호를 비교하여 유사도를 찾는 DTW(Dynamic Time Warping)을 이용하는 방법[2], 사람의 정보처리 과정을 모델링 하여 간단하고 많은 처리요소들을 병렬로 연결하여 학습을 통해 입력패턴에 존재하는 정보를 찾아내어 처리하는 ANN(Artificial Neural Network)[2] 등이 있고, 이러한 방법들을 통해 음성인식률이나 인식할 수 있는 음성의 수를 증가시키는 것이 주된 목적이 되어왔다.

그러나 최근 많은 응용분야에서 사용되는 음성인식 시스템은 몇 개의 한정된 단어만을 사용하여 입력된 음성에 대한 빠른 응답을 요구한다.

이러한 응용분야에서는 기존의 컴퓨터 시스템에서 사용되어왔던 음성인식기술과는 다른 제한된 domain에서 실시간 음성인식기술이 필요하게 되었다.

제한된 domain에서의 격리단어는 단어구간을 파악하기가 쉽고 음성의 끝점 추출이 용이하다. 또한 인식단어가 20개정도로 적기 때문에 유사한 단어가 적고 데이터 베이스도 작아지는 이점이 있다.

화자의 음성 입력에 대해 좀더 빠른 응답과 주변 잡음에 영향을 적게 받기 위해서는 정확한 음성 인식 능력이 떨어지는 단점이 있다. 그러나 제한된 domain에서의 음성인식에 있어 정확한 인식능력은 큰 의미를 가지지 않으므로 거부율을 낮춤으로써 인식률을 높일 수 있다.

본 논문에서는 한정된 domain에서 고립단어에 대해 주파수영역에서 음성의 특징을 찾아내는 방법을 제시하고자 한다. 2장에서는 주파수 영역에서 고립단어를 인식하기 위한 특징추출 방법에 대해 논하고 마지막 3장에서는 실제적으로 음성을 인식하는 과정과 2장에서 논한 특징추출 방법의 장점과 문제점을 논한다.

II. 주파수 영역에서 고립단어 인식

1. 고립단어 인식 시스템의 개요

일반적으로 음성인식 시스템은 크게 음성을 입력받는 부분, 입력받은 음성에서 고립단어를 추출하기 위한 끝점 검출 부분, 추출된 고립단어에 대한 특징을 추출하는 부분, 마지막으로 추출된 특징으로 음성을 인식하는 부분으로 나뉘어 진다.

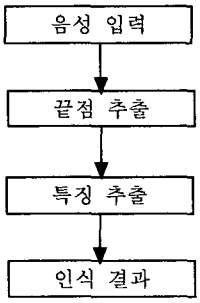


그림 1. 일반적인 음성인식 시스템의 구조

음성인식 시스템에서 가장 중요한 부분이 바로 음성의 특징을 추출하는 부분이다. 이러한 음성의 특징을 추출하기 위해 몇 가지 계층적인 단계를 거친다.

첫 번째 단계에서는 주파수 영역에서 peak 값을 가지는 영역의 수를 찾아낸다. 두 번째 단계에서는 그 영역들이 가지는 주파수 값을 계산한다. 그리고 마지막 단계에서는 가장 높은 peak 값과 그 다음의 peak 값의 차이를 이용한다.

격리단어 인식에 사용한 음성자료는 10명의 이름을 각각 10번 발음하여 사용한다. 각 음성은 8KHz로 sampling하여 FFT를 사용하여 주파수 영역으로 변환시켜 특징을 추출한다.

음성자료(괄호안은 표기방식)				
조영훈(조)	박은명(박)	강홍석(강)	이종열(이)	홍준혁(홍)
최진실(최)	채시라(채)	윤도현(윤)	정시은(정)	서우석(서)

표 1. 격리단어 인식에 사용한 음성자료

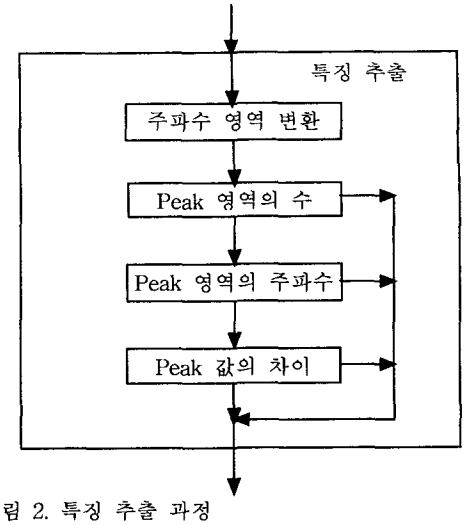


그림 2. 특징 추출 과정

2. Peak 영역 검색

첫 단계에서 수행하는 작업은 peak 영역의 수를 찾는 것이다. 입력된 음성을 주파수 영역으로 변환하면 그림 3 과 같은 결과를 얻을 수 있다.

일반적으로 주파수 영역에서의 음성신호는 몇 개의 peak값을 가지는 영역으로 이루어진다. 그림 4 에서 보듯이 peak영역의 수는 입력된 음성에 따라 달라진다. 이러한 peak영역의 수로 입력된 음성신호를 몇 가지 그룹으로 구분할 수 있다.

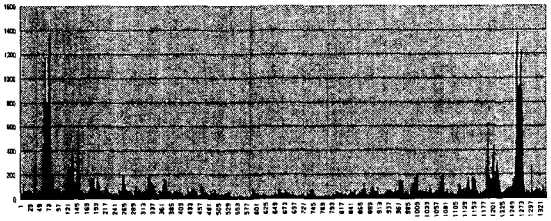
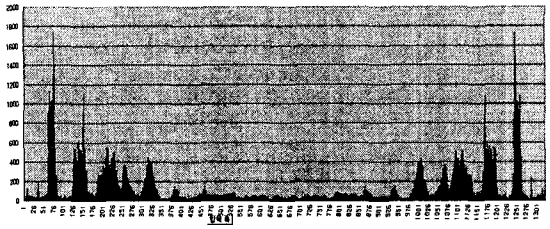


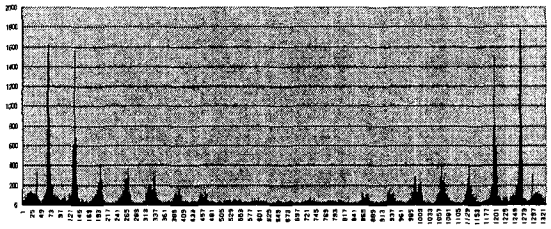
그림 3. 주파수 영역에서의 음성신호

Magnitude값을 어떻게 두는지에 따라 peak값이 달라질 수 있다. 표 2 는 음성데이터 [강홍석]에 대한 주파수 영역에서 임계값에 따른 peak영역의 수를 보여준다. 표 2 에서 보는 바와 같이 Magnitude 임계값을 500으로 두었을 경우 발음한 음성데이터에 대한 peak 영역의 수가 2개로 나타나고 그 값의 변화도 거의 일정함을 알 수 있다. 이 임계값은 나머지 9개의 음성에 대해서도 유사하게 나타난다. 또한 표 2 에서처럼 임계값이 300이하일 경우 peak 영역 수의 변화가 심하고 표 3에서 보는 바와 같이 임계값이 800이상일 경우 거

의 대부분의 음성데이터에서 peak영역의 수가 1이 됨을 알 수 있다. 임계값을 500으로 두는 것이 각 음성 데이터들을 구분하기에 가장 적합함을 알 수 있다.



[이종열]에 대한 주파수 영역에서의 신호



[박은명]에 대한 주파수 영역에서의 신호

그림 4. 서로 다른 음성신호에 대한 주파수 영역에서의 값

임계값	Peak 영역의 수									
	1	2	3	4	5	6	7	8	9	10
100	5	5	8	18	6	20	20	8	6	7
200	3	4	5	4	5	8	5	3	3	5
300	2	2	3	3	3	4	4	2	2	3
400	2	2	2	2	3	3	3	2	2	3
500	2	2	2	1	2	2	3	2	2	2
600	1	2	2	1	1	1	3	1	2	1
700	1	2	2	1	1	1	2	1	1	1
800	1	1	1	1	1	1	1	1	1	1
900	1	1	1	1	1	1	1	1	1	1
1000	1	1	1	1	1	1	1	1	1	1

표 2. 음성 데이터 [강홍석]에 대한 임계값에 따른 peak영역의 수

임계값	Peak 영역의 수(괄호안은 빈도)									
	조	박	강	이	홍	최	채	윤	정	서
400	4(7)	3(7)	2(6)	4(7)	2(8)	3(6)	2(8)	3(7)	3(6)	5(7)
500	4(8)	3(7)	2(8)	2(9)	2(8)	3(8)	2(7)	3(9)	2(7)	5(8)
600	3(5)	2(7)	1(5)	2(6)	2(8)	2(6)	2(6)	2(7)	2(7)	4(7)
700	2(6)	2(8)	1(7)	2(5)	2(9)	2(7)	1(6)	2(6)	2(5)	3(6)

표 3. 각 임계값에 대한 음성데이터들의 peak 영역의 수

3. 특정 임계값에서의 주파수 값

임계값 500에서의 각 발성음에 대한 주파수 값은 표 2와 같다. 임계값 이상의 Magnitude를 가지는 영역의 주파수 값은 표 3 에서 볼 수 있듯 대략적으로 peak영역의 수에 비례적인 값을 가지지만 peak영역의 수가 같더라도 주파수 값이 차이가 난다는 것을 알 수 있다.

	조	박	강
Peak영역의 수	4	3	2
주파수 값(Hz)	1550~1620	1180~1270	840~910
	이	홍	최
Peak영역의 수	2	2	3
주파수 값(Hz)	740~830	810~890	1050~1180
	채	윤	정
Peak영역의 수	2	3	2
주파수 값(Hz)	760~820	1140~1210	810~880
	서		
Peak영역의 수	5		
주파수 값(Hz)	1910~1990		

표 3. 임계값 500에서의 주파수

4. Peak 영역의 값의 차이

일반적으로 음성신호는 저주파부분에서 Magnitude 값이 크고 고주파로 갈수록 작아진다. 특정 임계값을 초과하는 영역들의 값의 차를 계산하기 보다 두 번째 영역의 Magnitude값이 얼마인가를 계산하는 것이 수행시간을 훨씬 단축시킬 수 있다. 이러한 peak영역의 차를 이용하기 위해 첫 번째 단계에서 peak영역의 수를 계산할 때 peak영역의 수가 처음으로 1이 되는 임계값을 찾아 그 임계값을 사용한다. 표 4 는 표 1 에서 peak영역의 수가 처음으로 1이 되는 임계값이다.

음성	Peak 영역의 수가 최초로 1이 되는 값									
데이터	1	2	3	4	5	6	7	8	9	10
조	700	800	800	700	800	800	800	800	800	600
박	X	X	X	X	X	X	X	X	X	X
강	600	800	800	500	600	600	800	600	700	600
이	800	800	800	600	800	700	800	800	800	800
홍	X	X	1000	X	X	X	X	X	X	1000
최	900	900	1000	900	X	1000	900	X	900	900
채	800	700	700	700	800	900	700	700	800	700
윤	1000	1000	X	X	900	1000	1000	1000	900	X
정	X	X	X	X	X	X	X	X	X	X
서	700	700	700	600	900	700	600	800	700	800

표 4. Peak 영역의 수가 처음으로 1이 되는 임계값

표 5 는 각 음성데이터에서 peak영역의 수가 처음 1이 되는 임계값들 중 가장 빈도가 높은 임계값과 그 빈도를 보여준다. 표 5에서 보는바와 같이 peak 영역의 수가 처음 1이 되는 임계값도 각 음성데이터들을 특징지을 수 있다.

Peak영역의 수가 1이 되는 임계값(괄호안은 빈도)				
조	박	강	이	홍
800(7)	X	600(5)	800(8)	X
최	채	윤	정	서
900(6)	700(6)	1000(6)	X	700(5)

표 5. Peak영역의 수가 1이 되는 임계값과 빈도

III. 결론

입력된 음성신호를 주파수 영역으로 변환하고, 첫 번째 단계를 거쳐 peak영역의 수를 계산한다.

	Peak 영역의 수			
	2	3	4	5
음성데이터	강, 이, 홍, 채, 정	박, 최, 윤	조	서

표 6. 각 음성데이터들의 Peak 영역의 수

그중 동일한 peak영역의 수를 가지는 음성데이터들을 다시 두 번째 단계를 거쳐 각각의 음성데이터들이 가지는 주파수 영역을 살펴본다.

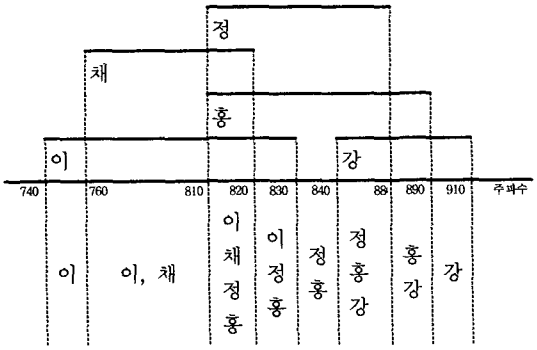


그림 5. Peak영역의 수가 2인 음성데이터들이 차지하는 주파수 영역

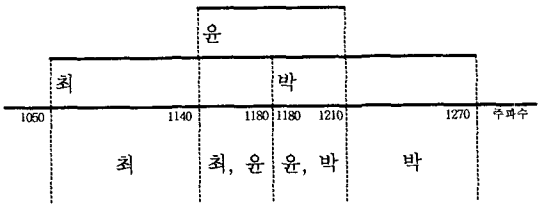


그림 6. Peak영역의 수가 3인 음성데이터들이 차지하는 주파수 영역

각 주파수 영역에서 겹치는 부분을 가지는 음성데이터는 마지막 단계를 거쳐 구분해 낼 수 있다.

이러한 방법은 기존의 음성인식에서 음성 특징 추출 방법에 비해 입력된 음성을 주파수 영역으로 보낸 뒤 주파수 영역에서의 신호에 대해 특별한 가공 없이 바로 특징을 추출해 낼 수 있고, 계층적인 단계를 거치므로 특징 추출에 있어 processing time을 줄일 수 있으며, 시간영역의 음성신호를 주파수 영역으로 보내고 간단한 작업만을 수행하므로 시스템에서 차지하는 cost를 줄일 수 있는 장점이 있다.

위 결과에서 [홍준혁] 과 [정시은] 의 경우 거의 구분해 낼 수 없는 것을 알 수 있다. 주파수 영역에서 각 음성을 구분해 낼 수 있는 특징을 더 찾는다면 좀 더 확장된 domain에서 적용해 볼 수 있을 것이고 신뢰성있는 음성인식 시스템을 구현할 수 있을 것이다.

참고문헌

[1] 정훈, "HMM을 이용한 실시간 화자독립 고립단어 인식에 관한 연구 ", 강원대학교, 1996

[2] 최지봉, "음성 파라미터 추출 방식에 따른 신경회로망을 이용한 음성인식", 강원대학교, 1997

[3] 도삼주, "전화 음성의 격리 단어 인식에 관한 연구", 한국 과학 기술원, 1990

[4] 이행세, "음성 인식 기법", 청문사, 1999

[5] 이영호, 정홍, "음절을 기반으로 한 한국어 음성인식", 전자공학회논문지, Vol. 31-B, No.1, Jan, 1994

[6] Gulmezoglu MB, Dzhafarov V, Keskin M, and Barkana A, "A Novel Approach to Isolated Word Recognition", IEEE Transactions on Speech & Audio Processing, Vol.7 No.6, Nov. 1999, pp. 620-628

[7] Montri Karnjanadecha and Stephen A. Zahorian, "Signal Modeling for Isolated Word Recognition", IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol.1, Mar. 1999 pp.293-296