음소경계 정보를 이용한 한국어 숫자음 인식에 관한 연구

최 관 묵 , 임 동 철 , 이 행 세
아주대학교 전자공학과

A Study on Korean Digit Recognition by Using Phoneme Boundary Information

Goan Mook Choi, Dong Chul Lim and Haing Sei Lee
School of Electronics Engineering, Ajou University
E-mail : cgmook@hanmail.net

Abstract

Recognition rate of Korean digit is lower than that of other words because it is composed of similar phonemes. In this paper, a new method is proposed for the improvement of recognition rate by using the phoneme boundary information. In addition, the proposed method rarely increase cost because phoneme boundary is found by using simple method.

We experimented with speech data of one man and then obtained results of enhanced speech recognition rate.

1. 서론

숫자음 인식은 실제 상용 분야에서 많이 사용될 수 있는 유용한 분야이지만, 숫자음이 유사한 음소들로 구성되어 있어 일반적으로 인식률이 낮은 편이다.

본 연구는 구현이 간단하면서도 고립단어 인식에 있어서 좋은 성능을 보여주고 있는 동적 시간정합 알고리즘에 음소경계 정보를 사용하여 한국어 숫자음 0부터 9까지의 인식률을 향상시키는 방법을 제안한다.

본 연구에서 제안하는 방법은 두 가지로 음소경계 검출을 통해 같은 수의 음소로 이루어진 숫자들을 그룹으로 묶어 같은 그룹 안에 속한 숫자들만을 인식 비교 대상으로 하는 방법과, 같은 그룹내의 숫자들 중에서 유사한 음소로 이루어진 숫자음을 대상으로 음소별 비교를 통하여 선택된 음소가 포함된 숫자음을 선택하는 방법으로 음소 경계 검출시의 작은 연산량 증가만으로 인식률을 향상시킬 수 있다.

2. 동적 시간정합의 개요

동일인이 같은 단어를 발성하는 경우에도 발성할 때마다 단어의 시간적 길이 뿐만 아니라 구성 음소별 지속시간이 변화한다. 이 문제를 효과적으로 처리할 수 있는 방법이 Vinsyuk, Sakoe 와 Chiba 에 의해 제안된 동적 프로그래밍(dynamic programing, DP)에 기반한 시간축의 비선형 신축에 의한 정합법이다[1][2]. 동적 시간정합은 서로 다른 두 개의 자료에서 비선형의 최적의 경로를 찾아 서로 다른 길이의 특징 벡터를 비교하는 방법이다. 그림 1은 동적 시간정합에 의해 찾아진 최적의 경로를 나타내는 그림이다. 그림에서 i(k)와 j(k)

가 정합되었을 때 두 자료간의 차이가 최소가 됨을 의미한다. 즉, i(k)와 j(k)는 같은 음소에 포함되어있다는 것을 알 수 있다.

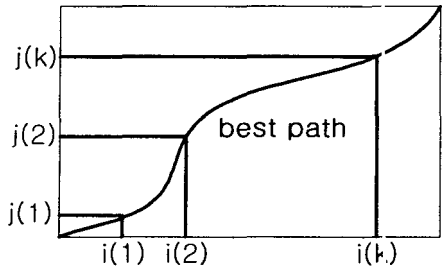


그림 1. DTW search

최적 경로는 표준패턴 열과 입력특징 벡터열의 최소거리를 찾아내는 식(1)에 의해 찾아진다[3].

$$D_{\min}(i_k, j_k) = \min_{(i_{k-1}, j_{k-1})} \{ D_{\min}[(i_{k-1}, j_{k-1}) + d[(i_k, j_k)|(i_{k-1}, j_{k-1})]] \quad (1)$$

이를 정리하면 아래와 같다.

1) 초기화

For j=1, 2, ..., J

$D_{\min}(1, j) = d[(1, j)|(0, 0)]$

Next j

2) 반복

For i=2, 3, ..., I

For j=1, 2, ... J

Compute $D_{\min}(i, j)$

Record path

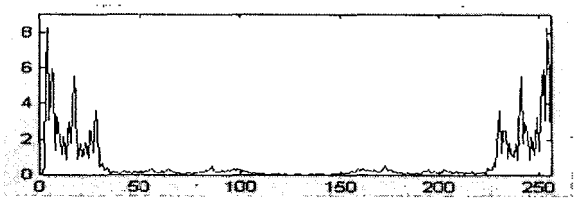
Next j

Next i

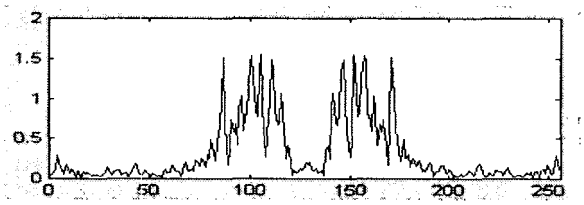
3. 음소경계 정보를 이용한 동적 시간정합

3.1 스펙트럼 정보를 이용한 음소경계 검출

음소 경계검출은 음성인식에 있어서 가장 기본적인 문제이면서도 어려운 문제중의 하나이다. 특히 유성음간의 음소구분은 그 경계도 애매할 뿐만 아니라 수작업으로 한다해도 구분하기가 쉽지 않은 일이다. 현재 음소구분은 단구간 에너지나 영교차율, 스펙트럼 등의 여러 가지 특징 파라미터를 사용하여 행해지고 있으며 신경망을 이용하여 검출하는 방법 등이 발표되고 있으나 본 연구는 한국어 숫자음 0부터 9까지의 단음절을 대상으로 하기 때문에 구성 음소의 수는 1개에서 3개까지로 비교적 간단하다. 그래서 본 연구에서는 음소 경계 검출시의 연산량 증가를 줄이기 위해서 음성의 스펙트럼 정보만을 이용하려 한다. 음성을 단구간 Discrete Fourier Transform(DFT)하여 magnitude 값을 취하면 그림 2와 같이 음성구간의 주파수 특성에 따라 무성음과 유성음을 구분할 수 있다[4].



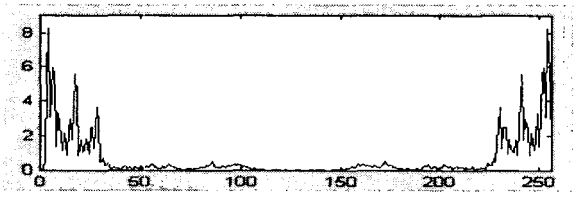
(a)



(b)

그림 2. 단구간 음성을 DFT한 경우
(a) 7의 중성 모음 'ㅣ'구간
(b) 7의 초성 자음 'ㅊ'구간

또한 그림 3에서 보듯이 유성음에서 formant의 변화로 인해 저주파 에너지가 밀집되어 있는 영역이 변하는 것을 감지하여 유성음간의 경계도 검출할 수 있다.



(a)

(b)

그림 3. 단구간 음성을 DFT한 경우
(a) 7의 중성 모음 'ㅣ'구간
(b) 7의 종성 자음 'ㄹ'구간

위의 특징을 이용한 제안된 음소경계 검출 방법은 아래와 같다.

1) 유·무성음 경계 구함

각 음성을 256 frame 크기의 단위로 DFT 하여 128 point 이하의 값은 0으로 대체 시킨다.

전체 frame의 average magnitude 값을 구한 후 frame내에서 유성음의 formant를 포함시킬 수 있게 충분한 영역 128×(2/3)지점까지의 저주파 영역의 average magnitude값을 구한다. frame을 이동시켜 가면서 위의 과정을 반복하여 저주파 영역의 값이 전체 frame값의 0.9배 이상인 구간이 연속적으로 100ms이상이면 그러한 frame의 시작점을 유성음 구간의 시작으로 간주한다. 유성음으로만 이루어진 경우도 음성의 시작 부근에서 노이즈로 인해 유·무성음의 경계가 검출될 수 있으므로 유·무성음 경계점이 음성 시작점의 50ms 이내에 있을 경우 무시한다.

2) 유·유성음 경계 구함

DFT한 frame의 중간지점부터 frame 시작 부분까지 frame의 내에서 크기가 작은 window를 사용하여 window의 average magnitude값을 구하면서 역으로 탐색하여 나간다. 이렇게 구한 값이 일정 threshold 값을 넘는 점을 찾아 frame의 저주파 영역에 밀집된 에너지 경계로 한다. 저주파 에너지 경계를 계속 찾아내면서 음성의 끝점까지 위의 과정을 반복한다. 이렇게 구한 저주파 에너지 경계가 비슷한 값을 유지하다가 그 값이 심하게 변하는 frame을 찾아 그 frame을 formant가 변하는 점으로 하여 유성음간의 경계로 한다. 음소경계 부근에는 스펙트럼의 변화가 심하므로 경계 부근에서 50ms이내의 formant의 변화로 인해 검출된 유·유성음 경계는 무시한다. window 크기는 10 point 정도로 하고 threshold 값은 너무 크면 formant의 변화를 감지하지 못하고 너무 작으면 formant의 순간적인 변화에도 반응하므로 1정도로 한다.

멜 주파수 켑스트럼 계수(Mel frequency cepstrum coefficient, MFCC)를 특징 벡터로 사용할 경우 따로 DFT를 할 필요가 없어 연산량의 증가는 거의 없지만 위의 방법을 사용하여 음소경계 검출을 하면 2나 5와 같은 하나의 음소로 이루어진 숫자도 음성의 끝부분에서 formant들이 사라짐으로 인해 잘못된 유성음 경계가 검출된다. 또, 4와 9같은 경우도 2개의 음소로 이루어져 있지만 위와 같은 이유로 유성음경계가 하나 더 검출되어 3개의 음소로 구분된다. 그래서 2개의 음소로 이루어진 0, 1, 2, 5, 6 과 3개의 음소로 이루어진 3, 4, 7, 8, 9 의 2개의 그룹으로 나누어진다.

3.2 음소경계 정보를 이용한 동적 시간정합

본 절에서는 음소경계를 사용하여 숫자음의 인식률을 향상시키기 위해 2가지 방법을 제안한다. 제안 1은 3.1절에서 제안된 음소경계 검출 방법을 사용하여 음소경계를 찾아내고, 구성 음소의 수별로 그룹을 형성(0, 1, 2, 5, 6 과 3, 4, 7, 8, 9) 같은 그룹 안에 속한 음성만을 비교 대상으로 하여 오인식이 일어날 수 있는 숫자음들을 비교 대상에서 격리시키는 방법이다. 제안 1의 방법은 DP의 연산량을 줄일 수 있을 뿐 아니라 인식률을 향상시킬 수 있다. 제안 2는 제안 1에 의해 같은 그룹 안에 속하게 된 숫자들 중에서 유사한 음소로 이루어져 오인식이 자주 되는 1과 2, 3과 4의 인식률을 높이기 위한 방법으로 전체 minimum cost에서 음소별 부분값을 각각 비교하여 작은 값을 가지는 음소를 선택하는 방법이다. 일반적인 동적 시간정합을 이용한 방법은 숫자음 같이 유사한 발음이 많은 경우에 있어서 오인식이 일어날 가능성이 커지게 된다. 예를 들어 숫자 1이나 2같은 경우는 1의 종성인 'ㄹ' 구간만이 차이가 나기 때문에 전체적으로 볼 때 그 차이가 무시되고 오인식되는 경우가 많이 생긴다. 실제로 그림 4와 같이 음소경계를 이용하여 각 음소영역별(p_1, p_2, p_3) 부분 값을 관찰한 결과 3과 4는 3의 종성인 'ㅁ'영역에서의 분명한 차이에도 불구하고 초성과 중성의 음소가 같은 4로 오인식되는 경우가 많이 발생하였다. 제안 2는 이러한 점을 해결하기 위해 3과 4는 초성과 중성의 음소가 같으므로 종성의 음소만을 비교하여 작은 값을 가지는 음소가 속한 숫자음을 선택하는 방법이다. 1과 2의 경우도 마찬가지이다.

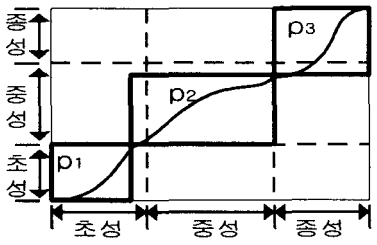


그림 4. 음소경계 파라미터

이를 정리하면 아래와 같다.

1) 식(1)에 의하여 구하여진 최종 minimum cost 값의 차이가 작은 두 음성이 3과4, 1과2인 경우만을 대상으로 하여 각 숫자의 종성 음소에 해당하는 부분값 p_a 와 p_b를 구함
2) p_a 와 p_b를 비교하여 작은 값을 가지는 음소를 선택한다.
3) 선택된 음소가 포함된 숫자를 선택한다.

4. 실험 및 결과 분석

제안된 방법과 기존의 방법의 성능을 비교하기 위해서 실험실에서 11㎑의 8bit로 녹음된 잡음이 없는 남성 1인의 음성을 사용하였고, 각 숫자의 발음 횟수는 25회로 이중 3개는 reference 음성으로 나머지 22개는 test 음성으로 사용하여 총 220개의 음성 데이터를 가지고 화자종속 시험을 하였다. 특징 벡터는 12차 MFCC를 사용하였다.

성능 평가의 결과는 표 1과 같다. 제안 1은 구성 음

표 1. 각 숫자음의 인식률 단위(%)

숫자음	0	1	2	3	4	5	6	7	8	9	Total
기존의 방법	100	100	95	72.7	100	100	100	100	100	90.9	95.9
제안 1	100	100	95	77.3	100	100	100	100	100	100	97.3
제안 2	100	100	95	90.9	100	100	100	100	100	100	98.6

소의 수에 따라 2개의 그룹으로 나누어 같은 그룹 안의 대상만을 인식 비교대상으로 한 방법의 결과이다. 제안 2는 같은 그룹 내에서 음소별 부분 값으로 비교를 한 후 선택된 음소를 포함한 숫자로 인식한 결과로서 동적 시간정합 결과에 미치는 영향을 줄이기 위해 두 음성의 minimum cost의 차가 20% 이내인 경우에 한해서만 실행을 하여 값의 차이가 많이 나는 음성은 비교대상에서 제외하였다.

표 1에서 보는 바와 같이 일반적인 동적 시간정합 알고리즘에 의한 인식률은 95.9%이다. 제안 1의 방법에 의해 9와 3의 오인식률이 개선되어 97.3%로 전체적으로 1.4%의 인식률을 향상시켰다. 제안 2의 방법에 의해 인식률이 낮았던 3의 인식률을 18.2% 개선시켜 인식률은 98.6%로 일반적인 방법에 비해 전체적으로 2.7%의 인식률을 향상시킬 수 있었다.

5. 결론

본 연구는 한국어 숫자음 인식에 많이 발생하던 특정 유사음성의 오인식을 제안된 음소경계 정보를 이용한 방법으로 개선시켜 일반적인 방법에 비해 1.4%와 2.7% 높은 인식률을 나타냈다. 또한 멜 주파수 캡스트럼 계수 추출시 DFT를 해야하므로 음소경계 검출을 위해 따로 DFT 할 필요가 없어 연산량의 증가도 크지 않다.

구성 음소의 개수에 따른 그룹화가 정확하게 되었을 경우 더욱 인식률을 향상시킬 수 있으므로 보다 정확한 음소 경계 검출방법의 연구가 필요하고 숫자음 이외의 유사음소로 구성된 단어로의 확장과 더 많은 음성 데이터를 이용한 시험이 필요하다.

참고문헌

[1] H. Sakoe and S. Chiba, "Dynamic programing optimization for spoken word recognition," *IEEE Trans. Acoustics, Speech, Signal Proc.*, **ASSP-26(1)**, 43-49, Feb. 1978.

[2] Lawrence Rabiner and Biing-Hwang Jung, *Fundamentals of speech recognition*, (Prentice Hall,1993) chap. 4, pp 200-232.

[3] John R. Deller, Jr, John G. Proakis and John H. L. Hansen, *Discrete-Time Processing of Speech Signals*, (Macmillan Publishing Company, 1993), chap. 11, pp 623-633.

[4] B.Yegnanarayana, C. d'Alessandro, and V. Darsinos. "An iterative algorithm for decomposition of speech signals into periodic and aperiodic componests," *IEEE Transactions on Speech and Audio Processing*, **6(1)**, 1-11, Jan 1998.