한국어 연속음인식에 관한 연구 (유성음 분류및 단모음 인식)

On the Classification of Voiced Sound and the Recognition of Vowels for Korean Continuous Speech

* 하 판 봉 (Ha, P. B.)
** 이 철 희 (Lee, C. H.)
*** 방 승 찬 (Bang, S. C.)
**** 안 수 길 (Ann, S. G.)

요 약

본 논문에서는 우리나라 음성의 유성음을 모음, 비음 및 유성화 자음으로 분류하는 알고리즘을 기술하였다. 먼저 기존의 PITCH검출 알고리즘에 의하여 음성을 유성음과 무성음으로 나눈 뒤, 단지 정규화된 1차 상관계수, 영교차율, LOG에너지 및 LOG 에너지의 끝짜기 검출만을 이용하여, 유성음은 모음, 비음 및 유성화자음으로 분류하고 무성음은 실제의 무성음과 묵음으로 분류하였다.

그리고 이렇게 분류된 모음에 대하여 단모음 인식을 행하였다. 단지 한 FRAME으로 모음을 대표하였기 때문에 메모리 크기와 인식 시간을 줄였다.

여기서 UP & DOWN 및 수정된 영교차율을 새로이 정의하여 적용한 결과 만족한 결과를 얻을 수 있었다. LPC 배개변수 및 전력 스펙트럼도 단모음 인식의 FEATURE로 사용하였다. 그리고 각 FEATURE의 성능을 비교하였다. 이들 FEATURE을 잘 조합하여 2단계 인식을 행한 결과 92%의 높은 인식율을 얻을 수 있었다.

ABSTRACT

In this paper, the classification of Korean voiced sound into vowel, nasal and voiced consonant is studied. First using available pitch extraction algorithm, speech is classified as two classes; voiced and

*, **, ***, **** 서울대학교 전자공학과 통신이론실

unvoiced sounds. And then voiced sound is classified as one of vowel, nasal and voiced consonant, and unvoiced sound as one of actual unvoiced sound and silence by using only normalized first-order auto-correlation coefficient, zero-crossing rate, LOG energy and valley of LOG energy contour as features.

For classified vowel, recognition of Korean vowels in continuous speech is considered. Only one frame(128 samples) is taken as a reference, so that memory size is reduced and recognition time is saved.

New features, UP & DOWN and modified zero-crossing rate are defined and successfully applied. LPC parameters and power spectrum are also used as features. And performance of each feature is calculated. With best combination of these features and two step recognition, high recognition rate(92%) is achieved.

1. 서 론

음성 인식은 신호처리의 가장 중요한 분야의 하나로 오래전부터 많은 사람이 관심을 갖고 연구하였다[2]~[8]. 한국어에 대해서는 주로 숫자음과 고립단어인식에 대한 연구가 발표되었다[9]~[10]. 음성 인식은 크게 고립 단어 단위에 의한 방법과 음소 단위에 의한 방법을 생각할 수 있다.

고립 단어 단위의 음성인식은 먼저 FILTER BANK, LPC 분석 등에 의하여 FEATURE를추출하고, 기준 PATTERN과 DISTANCE를 계산하여 인식을 행한다. 또한 시간의 차이를 해결하기 위하여 DYNAMIC PROGRAMMING기법이 도입되었다[6]~[8]. 이 방법은 단어 수가 많지 않을 때는 높은 인식율을 얻을 수 있지만 단어 수가 증가함에 따라 인식 시간의 증대와 MEMORY 용량의 증대, 인식율의 저하하는 문제가 있다.

이에 반해 음소 단위에 의한 인식은 고립 단어단위의 인식에 비해 다소 인식율은 떨어지지만 단어 수의 증가에 따른 MEMORY 용량의 증대, 인식 시간의 증가, 인식율의 저하 등의 문제에 있어서는고립 단어 단위의 인식보다 유리하다.

2. 한국어 음성 인식의 문제점과 특수성

한국어는 타 언어와 비교하여 다음과 같은 문제점이 있다. 특히 이러한 문제점은 연속음 인식을 행할 때 두드러지게 나타나게 된다.

1) **조사의 발달** : 한국어에서는 조사가 발달하여 하나의 명사가 수 많은 고립단어군을 이룬다. 즉 연음, 경음화 법칙 등에 의하여 두개의 고립단어라기보다는 하나의 고립 단어라고 볼 수 있다.

예) ~을(를), ~가, ~은(는), ~에게(에), ~와(과), ~으로, ~의, …

2) **어미의 발달** : 한국어의 형용사, 동사는 많은 어미 변화를 하여 하나의 동사, 형용사가 수 많은 고립 단어군을 이룬다.

예) 생각하고, 생각하여, 생각하니, 생각하지만, 생각했고, 생각할, 생각했지만, 생각하기 때문에.

이러한 문제점으로 인하여 고립 단어 단위에 의한 한국어 연속음 인식을 행하려면 비록 적은 수의 단어를 인식하려고 하여도 많은 어려움이 따르게 된다.

이에 반해 한국어는 타 언어에 비하여 다음과 같은 특수성이 있어 새로운 인식 ALGORITHM을 생

각할 수 있다.

1) 타 언어에 비하여 음절의 분리가 정확하다. 그리고 음절 단위로 또박 또박 발음하여도 의미 전달에 아무런 지장이 없다.

2) 억양(ACCENT)이 의미상 중요하지 않다.

3) 글자(음절)의 사용 빈도수가 적은 수에 국한되어 있다. 즉 한글에서는 현재 사용하는 약 1,500여 글자 가운데 270번째 글자에서 90%의 누적백분율을 보이고 있다[04].

이상의 한국어의 문제점과 특수성을 고려할 때 한국어 연속음 인식은 고립 단어 단위에 의한 방법은 거의 불가능하고 음소 단위에 의한 방법 또는 음절 단어에 의한 방법을 고려해야만 한다. 그러기 위해서는 먼저 한국어 음소의 통계치에 대한 연구가 요청되고, 연속음이 입력되었을 때 음절 단위로 분류하는 ALGORITHM의 개발이 요청된다. 특히 자음소에 대한 통계치는 많은 사람에 대한 DATA 분석, 각FEATURE별 통계적 특성, 새로운 PARAMETER의 개발이 요청된다.

본 논문에서는 음절 분리의 첫번째 단계로서 유성음을 모음, 비음, 유성화자음으로 분류하는 ALGORITHM과 음소 인식의 한단계로서 단모음 인식에 관하여 연구하였다.

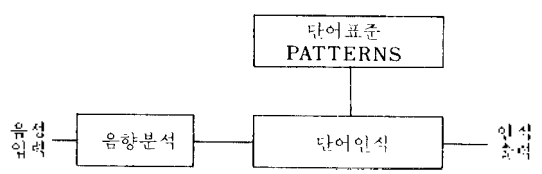

그림 1 고립단어 단위에 의한 단어 인식.

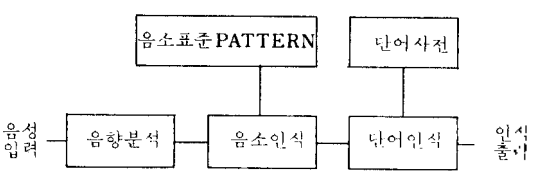

그림 2 음소 단위에 의한 단어 인식.

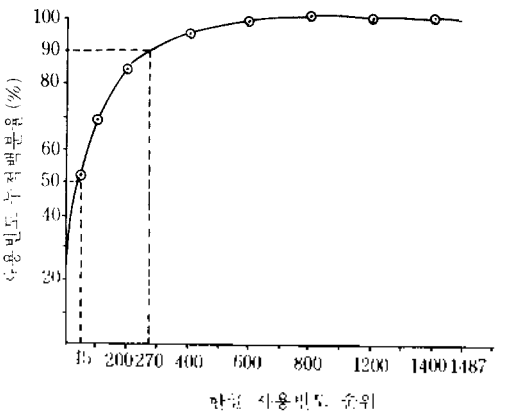

그림 3 한글 사용도수 누적백분율과 빈도 순위[05].

3. 유성음 분류

음성인식 분야에 있어서 일정한 한 분석 구간의 음성을 세분화할 수 있으면 그것을 이용해서 연속음성을 음소나 음절 단위로 처리하는 것이 가능하다. 보통 음성은 유성음, 무성음 및 묵음으로 크게 분류된다. 유성음으로는 모음과 유성자음이 있다. 유성자음은 다시 비음, 유음 및 유성화 자음으로분류된다.

이 절에서는 유성음을 이와같이 분류하는 알고리즘을 기술한다. 알고리즘을 간단하게 하기 위하여 정규화된 1차 상관계수(C1), 영교차율(ZCR), LOG에너지(LE) 및 LOG 에너지의 골짜기 검출(VA)만을 사용했다. 음성을 우선 기존의 PITCH 검출알고리즘에 의해서[02]유성음과 무성음으로 나누어 놓고, 유성음은 모음, 비음 및 유성화 자음으로 분류한다. 무성음은 영교차율과 LOG에너지를 사용하여 실제의 무성음과 묵음으로 나눈다. 남성의 경우 PITCH가 25ms까지 커지지 때문에 유·무성음을 구분할 때는 1FRAME을 256 SAMPLE (32 ms)로 하고, 유성음을 다시 분류할 때는 한 FRAME을 128 SAMPLE (16 ms)로 했다. 비음은 파형

에 있어서 굴곡이 모음이나 유성화 자음보다 심하지 않으므로 정규화된 1차 상관계수

$$C1 = \sum_n S(n)\,S(n-1) / \sum_n S^2(n) \qquad (1)$$

가 거의 1에 가깝다는 성질과 영교차율

$$ZCR = 1/2 \cdot \sum_n ABS(SIGN(S(n+1)) - SIGN(S(n))) \qquad (2)$$

의 통계에 의하여 분류된다. 경계치는 각각 THC1 =0.975, THZCR=11로 선정한다. 유성화 자음은 영교차율이 큰 경우와 작은 경우가 있는데, 작은 경우는 유음사이에 있는 자음의 구간이 짧은 경우에서이다. 우선 영교차율로부터 영교차율이 클 경우의 유성화 자음을 모음으로부터 분류한 뒤 분류되지 않은 유성화 자음은 유성음사이에서 천이구간으로만 존재하더라도 자음은 입안을 닫아서 발음하는 것이기 때문에 에너지에 있어서 골짜기가 생긴다. 이와같은 골짜기는 정상속도로 발음했을 때 5개의 128 SAMPLE의 FRAME (80ms) 내에서 일어나는데 5개 FRAME에서 16 SAMPLE씩 WINDOW을 취하면서 LOG에너지

$$LE = 10 \cdot LOG(1/16 \cdot \sum_n S^2(n)) \qquad (3)$$

를 계산하여 그것들이 골짜기 형태인지를 점검한 뒤 골짜기 FEATURE

$$VA = LE(1) + LE(40) - 2 \cdot MIN(LE(1)) - ABS(LE(1) - LE(40)) \qquad (4)$$

에 의해 유성화 자음인지를 판정한다. 경계치는 각각 THWZ=30, THVA=6으로 정했다.

이전까지 분류된 것에 대해 문장 구문에 맞게 SMOOTHING을 한다. 이때 SMOOTHING은 3개의 FRAME을 가지고 하는 데 그 알고리즘은 그림 4와 같다. 표 1은 이미 알고 있는 FRAME에 대한 결과를 분석한 것이다. 이것은 SMOOTHING하기 이전의 결과인 데, SMOOTHING을 했을 때도 유성화 자음을 제외한 나머지에 대해서는 비슷한 양상을 보였다. 표 2는 SMOOTHING한 후 유성화자음에 대한 결과 분석이다.

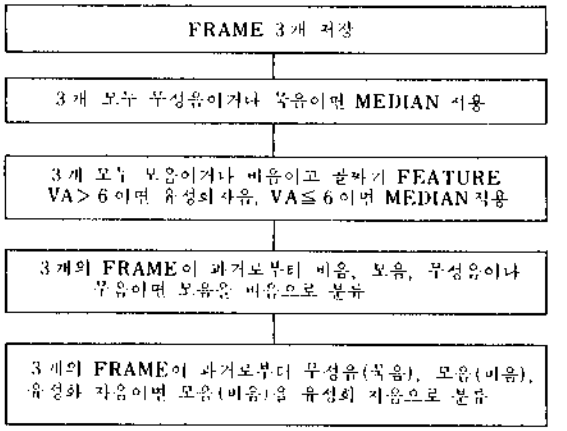

그림 4 SMOOTHING 알고리즘.

표 1 SMOOTHING 전의 판정 결과

	모 음	비 음	유성화 자 음	무성음	묵 음
모음(148)	140	3	–	1	4
비음(88)	34	51	–	–	3
유 성 화 자음(37)	29	1	7	–	–
무성음(39)	1	1	7	29	1
묵음(200)	–	–	–	–	200

표 2 SMOOTHING 후의 유성화 자음의 판정 결과.

	모 음	비 음	유성화 자 음	무성음	묵 음
유 성 화 자음(41)	3	1	37	–	–

4. 단모음 인식

4.1 FEATURE의 추출

1) 영교차율

기본적인 FEATURE의 하나로 주파수 특성을 나타내 주고 (2)식과 같이 정의된다.

2) UP & DOWN

새로 도입된 FEATURE의 하나로 증감이 바뀌는 횟수로 정의한다. 이는 다음의 두가지 이유로도 입되었다.

a) 높은 주파수 성분이 강한 모음이라도(예:"이") 작게 발음되는 경우 고주파성분이 약하게 나타난다.

b) 영교차율의 경우 낮은 주파수가 섞여 있는 경우 영교차율이 낮게 나타난다.

$$\text{UP \& DOWN} = 1/2 \cdot \sum_{n} (\text{SIGN}(S(n+1) - S(n)) - \text{SIGN}(S(n) - S(n-1))) \quad (5)$$

3) 수정된 영교차율

차수 M인 수정된 영교차율의 경우 한번 영교차가 일어나면 그 다음 M SAMPLE동안에는 영교차로 간주하지 않는다. 도입 이유는 강한 낮은 주파수 성분에 높은 주파수가 섞여있는 경우 낮은 주파수를 찾기 위함이다.

4) LPC ANALYSIS

LPC PARAMETER에서는 여러가지 FEATURE를 생각할 수 있다. 가장 대표적인 것의 하나는 LPC 계수로부터 FORMANT를 구하고 그 특성을 비교하였다.

5) 전력 SPECTRUM

주어진 FRAME을 DFT하여 주파수 영역에서 특성을 비교하였다. 또 주파수 대역을 여러 대역으로 나누어 각 모음별 특성을 비교하였다.

4.2 DISTANCE MEASURE

본 논문에서는 REFERENCE PATTERN의 크기를 가능한 줄이기 위해 모음의 대표값을 한 FRAME(128 SAMPLES)만을 사용하였다. 이 경우 메모리 용량이 적으므로 후에 여러가지 경우에 쉽게 확장할 수 있다. 그리고 몇 가지 FEATURE에 대해서는 통계치를 구하여 성능비교를 하였다.

음성은 경우에 따라서 빠르게 또는 천천히 발음될 수 있다. 이러한 시간적인 변동을 보완해 주기 위해 LINEAR TIME ALIGNMENT 또는 DYNAMIC TIME WARPING이 도입되었다. 본 논문에서는 대표값을 한 FRAME으로 하였기 때문에 TIME ALIGNMENT문제는 발생하지 않는다. 그 결과 인식 시간에 많은 감축을 볼 수 있다.

여러가지 FEATURE에 대하여 인식 실험을 하였으므로 각 FEATURE에 대하여 DISTANCE MEASURE가 달라진다.

1) CROSS-CORRELATION

REFERENCE PATTERN과 TEST PATTERN을 서로 CROSS-CORRELATION하여 유사도를 측정한다. 이 값은 크기가 큰 부분에 큰 영향을 받게 된다.

2) MEAN SQUARE ROOT

REFERENCE PATTERN과 TEST PATTERN과의 차를 제곱하여 합산 또는 평균한다. 이 값은 주로 차이가 큰 부분에 의하여 결정된다.

3) VARINNCE DISTANCE

새로 정의한 값으로 다음과 같이 정의된다. 도입 이유는 음성의 강약의 변화에 영향을 덜 받게 하기 위함이다.

R(n): REFERENCE FRAME

T(n): TEST FRAME

OF: OFFSET TO PREVENT BEING DEVIDED BY ZERO

$$AVERAGE = \sum_{n} ((R(n)+OF)/(T(n)+OF)) \quad (6)$$

$$VARIANCE\ DIFFERENCE = \sum_{n} ((R(n)+OF)-(T(n)+OF)-AVG)^2 \quad (7)$$

4) ITAKURA DISTANCE

LPC 계수에 대하여 사용하였다[5].

4.3 통계 특성

새로 정의한 FEATURE의 경계치에 이용한 FEATURE의 각 모음별 특성은 표 3～5와 같다. 수정된 영교차율의 경우 "아"와 "에"는 상당히 떨어진 특성을 보여주고 있다. 실제 인식에서는 이 둘이 서로 잘못 인식되는 경우가 상당히 많으므로 수정된 영교차율의 통계 특성을 이용하여 ERROR을 크게 낮출 수 있다.

표 3 영교차율의 통계.

	평 균	표 준 편 차
아	27.9	4.8
어	19.1	5.5
오	13.0	3.6
우	12.2	1.8
이	12.1	3.5
에	15.7	3.2

표 4 UP & DOWN의 통계.

	평 균	표준편차	최 대 값	최 소 값
아	38.4	4.9	53.4	34.8
어	32.2	7.4	40.3	16.4
오	20.3	10.3	35.5	12.9
우	21.2	8.9	37.0	13.8
이	49.9	10.8	70.0	31.6
에	32.9	9.7	41.4	22.6

표 5 수정된 영교차율의 통계(M=3)

	평 균	표준편차	최 대 값	최 소 값
아	26.8	5.0	34.0	17.0
어	18.0	5.0	24.2	12.1
오	11.8	2.0	13.2	5.5
우	11.2	1.5	13.2	5.5
이	9.6	1.9	12.4	7.2
에	13.8	1.7	15.9	12.5

4.4 인식 실험

새로 정의한 FEATURE의 경계치 그리고 기존의 FEATURE등을 사용하여 인식 실험을 행하였다. 먼저 중요한 몇 개의 FEATURE의 PERFORMANCE을 살펴 보면 표 6～8과 같다.

이상의 인식 결과에서 어떠한 FEATURE도 단독으로는 만족할 만한 인식율을 보여주지 못했다. 그리고 "오"에서는 가장 많은 ERROR가 발생하였다. "오"는 주로 "우" 또는 "어"로 잘못 인식하였다"오"의 인식율을 높히기 위해 "오"에 대하여 PERFOR-

표 6 LPC PARAMETER(ITAKURA DISTANCE).

	정	오	
아	20	0	
어	11	1	
오	2	17	
우	9	3	
아	2	15	
에	1	7	
	45	43	인식율 : 51.1%

표 7 전력 SPECTRUM(MEAN SQUARE ROOT)

	정	오	
아	20	0	
어	7	5	
오	4	15	
우	11	1	
이	11	6	
에	5	3	
	58	30	인식율 : 65.9%

표 8 16 BAND 전력 SPECTRUM, OFFSET = =1000(VARIANCE DISTANCE).

	정	오	
아	5	15	
어	10	2	
오	8	11	
우	7	5	
이	17	0	
에	4	4	
	51	37	인식율 : 60.0%

표 9 모든 FEATURE 및 경계치를 사용한 경우.

	정	오	
아	20	0	
어	12	0	
오	11	8	
우	12	0	
이	17	0	
에	7	1	
	80	8	인식율 : 90.9%

MANCE가 좋은 FEATURE로 1차 인식을 하고 1차 인식 결과 "어", "오", "우"로 인식한 부분중 사용한 FEATURE들의 6단모음에 대하여 DISTANCE를 구하여 작은 순위로 1에서 6까지의 가중치를 부여했을 때 그 가중치들의 평균들간의 차가 0.6이하인 부분에 대하여 2차 인식을 행하였다. 그 결과 전체 인식율은 92%가 되고 ERROR 발생도 "오"에 편중되지 않았다.

5. 결 론

본 논문에서는 1음절에서 3음절까지의 50종류의 단어 음성에 대해서 기존의 PITCH추출 알고리즘에 의해서 먼저 유성음과 무성음으로 나눈 후, 그것을 다시 유성음은 모음, 비음 및 유성화 자음으로 분류하고 무성음은 묵음과 실제의 무성음으로분류하는 알고리즘을 제시하였다. 그 결과로 모음, 무성음 및 묵음의 경우에는 대부분을 제대로 인식했고 비음의 경우에는 그것이 존재하는 핵심 부분은 거의 옳게 했으며 유성화 자음인 경우에는 모음사이에 있는 'ㅎ'을 제외한 나머지는 대부분 옳게 분류해냈다.

이렇게 분류된 모음에 대해서 모음을 한 FRAME으로 대표하여 THRESHOLD외 몇가지 FEATURE을 사용하여 다음절 단어중 단모음 인식을 행한 결과 비교적 좋은 결과를 얻었다. 그 결과 적은 양의 DATA에서 여러가지 특성을 구하여 여러 위치에서 발음되는 모음 인식이 가능함을 보여주었다. 앞으로 대표값을 하나가 아니라 여러가지 경우에서 추출하다면 좀 더 광범위한 음성인식에 적용할 수 있으리라 기대된다.

1. L.R. Rabiner and R.W. Schafer, "Digital Processing of Speech Signals", Englewood Cliffe, N.J., Prentice-Hall, 1978.
2. J.D. Markel, "The SIFT Algorithm for Fundamental Frequency Estimation", IEEE Trans. on Audio Electroacoust., Vol. AU-20, Dec. 1972.

3. B.S. Atal and L.R. Rabiner, "Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Application to Speech Recognition", IEEE Trans. on Acoutics, Speech and Signal Processing, Vol. ASSP-24, No. 8, pp.201-212, June 1976.
4. L.R. Rabiner and M.R. Samber, "Application of an LPC Distance Measure to the Voiced-Unvoiced-Silence Detection Problem", IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-25, No. 4, August 1977.
5. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Regnition", IEEE Trans. on Acoutics, Speech and Signal Processing, Vol. ASSP-28, No. 1, pp.67-72, Feb. 1975.
6. H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", IEEE. Trans. on Acoutics, Speech and Signal Processing, Vol. ASSP-26, No. 1, pp.48-49.
7. C.S. Myers and L.R. Rabiner, "A Comparative Study of Several Dynamic Time-Warping Algorithm for Connected Word Recognition", BSTJ, pp.11389-1409.
8. C.S. Myers, L.R. Rabiner and A.E. Rosenberg, "On the Use of Dynamic Time Warping for Word Spotting and Connected Word Recognition", BSTJ, Vol. 60, No. 3, pp.303-325, Mar. 1981.
9. 오 영환, "숫자 음성 자동 인식에 관한 실험", 대한전자공학회지, Vol. 15, No. 6, Dec. 1978.
10. 한희, 김 순협, 박 규태, "특성 대역 에너지를 이용한 한국어 기본 숫자음성의 자동 인식에 관한 연구", 대한전자공학회지, Vol. 19, No. 3, June 1982.
11. 우메다 히로유끼, "한국어의 음성학적 연구", 형설출판사, 1983.
12. 허 웅, "국어 음운학", 샘문화사, 1985.
14. 한글기계화연구소, "한글기계화연구 I", pp. 5~28, Aug. 1975.
15. 강 전희, "한글, 한자전산화 문제에 관한 고찰", pp. 19~24. 전자공학회잡지, Vol. 11, No. 1, Feb. 1984.