음소 분류를 이용한 한국어 연속 숫자음 인식에 관한 연구

○ 임 재열, 안 수길
서울대학교 대학원 전자공학과

(A Study on the Recognition of Continuous Korean Digits using Phoneme Classification)

Jaeyeol RHEEM and Souguil ANN
Dept. of Electronics Eng. Seoul National Univ.

I. 서 론

오늘날 음성인식(speech recognition)은 급변하는 정보화 시대속에서 산업적으로 또는 공공적으로 광범위한 주목을 받고 있다. 외국의 경우에는 문장단위의 연속음 인식을 위해서 범국가적 연구 체제이 추진되어져와, 음성의 음향학적 측면외에도 언어학, 음성학, 문장학 등등의 언어 전반적 사항이 연구되어왔으며 인공지능(artificial intelligence) 과의 결합도 시도되고 있다[1,2]. 그러나 국내의 연구의 경우 제한된 어휘수의 고립단어, 또는 연결단어 인식에 관한 연구는 활발하나, 연속음 인식을 위한 한국어전반에 관한 연구는 아직 초보적인 단계에 있다[3,4]. 따라서 한국어의 연속음 인식을 위한 기본적인 연구가 수행되어야 할 필요성이 있다.

본 논문은 한국어 연속음 인식의 필요성을 인식하고 제한된 대상인 한국어 연속 숫자음에 대한 인식 시스템의 구현을 시도 했다. 이는 한국어 전반에 관한 언어학, 음운학, 문장학 등등의 광범위한 지식없이 적용할 수 있는 소 대상이기 때문이다. 그 특징으로 간략한 에너지 곡선을 이용한 새로운 비음구간 검출 알고리즘과 종성 내파음 구간 검출 알고리즘을 제시 했으며, optimal preemphasis factor의 이용과 기본주파수 검출시의 doubling과 halving 방지용 알고리즘을 사용했고, 포오만트 검출시 pole enhancement 방법을 적용했다. 그리고 화자 내부에 서의 발성 방법의 변동에서 오는 변동을 줄이고자 LPC 스펙트럼에서의 대역 에너지를 파라미터로 사용했다.

II. 연속음 인식 및 숫자음 구조

1. 연속음 인식

고립단어 인식을 위해서는 패턴 매칭(pattern matching) 방법이 주로 사용되고 있으나, 많은 단어를 인식할 때 계산량이 방대해지며, 조사및 어미변화에 따른 데이터량 또한 증가하고, 패턴의 표준화에 따른 문제점이 발생하게 된다. 따라서 음성 인식의 궁극적 목표인 자연스럽게 발성된 연속음성을 인식 또는 이해하기 위해서는 인식의 단위를 단어이하의 음소와 같은 미소 단위로 해야한다. 그리고 이러한 음소단위의 인식을 위해서는 개인차의 문제와 조음 결합이 가장 큰 어려움으로 지적되고 있다. 한편으로 무제한의 연속음을 인식하기 위해서는 단어 내부의 음소의 구별 외에도 문장으로서의 구조적 역할을 이해하기위하여 음성학외에도 언어학, 문장학, 어휘론과 문법, 강세, 억양변화 등의 운율현상등 언어 전반 사항을 연구해야된다. 그림 2-1에 일반적인 연속음인식 시스템의 구조를 보였다. 신호 처리 부분은 음성의 음향학적 특징 파라미터를 추출하는 부분으로, 이때 사용되는 파라미터는 화자나, 발음시간, 발음속도 등의 영향을 적게 받는 안정된 것을 사용하여야 한다. Bottom-Up Approach 부분은 추출된 파라미터를 이용하여 음성학적으로 분류하는 분류 인식에 해당되며 음소단위인 경우 음소분류(phoneme classification)를 하게된다. 여기서는 앞단에서 추출된 파라미터를 이용하여 특징 벡터를 추출한 후 분류 규칙(classification rule)에 의하여 유사성질을 지니는 음소단위의 유사군으로 분류된다. Top-Down Verification 부분은 앞단에서 얻은 음소 정보와 언어적인 문법구조를 이용하여 확인하는 부분이다. 이부분에서는 앞단의 분류된 음소 정보에 대하여 문법, 어휘, 발음 습관 등등에 관한 규칙을 이용하여, 수정및 편집(editing)을 수행함으로써 최종인식을 하게된다.

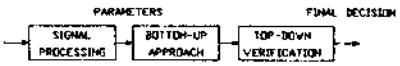


그림 2-1 연속음 인식 시스템의 구성

2. 한국어 숫자음의 음성학적 구조

본 논문에서 인식대상으로 한 숫자음은 '공', '일', ..., '구'의 10개의 단음절어를 연속해서 발음한 숫자이다. 그리고 숫자음 10개외에도 별도로 /에/발음도 포함시켰는데 이는 자연스러운 연속 숫자음을 얻고자 전화번호의 '-'를 /에/로 발음을 했기때문이다. 10개의 숫자음을 발음기호와 함께 표 2-1에 나타 냈다. 한국어 숫자음의 특징으로는 순수 모음으로 이루어진 숫자음 (이,오) 보다는 초성자음이 있는 숫자음(삼,사,칠,팔,구,공)이 많아 상대적으로 초성 자음의 유성음화 현상이 적게 일어나며, 단음절(monosyllabic) 성격을 띄고 있고, 초성자음과 종성자음을 알 경우에는 모음을 구체적으로 알지 못해도 구별할 수 있고, 어느 한 음소를 강화하 집중하지 않아도 추정할 수 있어 음소결합의 독립성이 좋다는 점을 들 수 있다.

표 2-1 한국어 숫자음

숫 자	분 류	발음 기호
영	전설모음 + 비음	/i/
이	전설모음	/i/
삼	무성음 + 전설모음 + 비음	/sam/
사	무성음 + 전설모음	/sa/
오	후설모음	/o/
육	전설모음 + 후설모음 + 무성음	/juk/
칠	무성음 + 전설모음 + 폐음	/t͡ɕʰil/
팔	무성음 + 전설모음 + 폐음	/pʰal/
구	무성음 + 후설모음	/ku/
공	무성음 + 후설모음 + 비음	/koŋ/

표 2-2 음소별 종류

음 소 명	종 류
모음(vowel)	/a/,/ɨ/,/o/,/u/,/jo/ (ㅏ, ㅡ, ㅗ, ㅜ,ㅛ)
활음(glide)	/j/ (ㅣ)
비음(nasal)	/m/,/n/ (ㅁ, ㄴ)
파열음(plosive)	/k/,/p/ (ㄱ, ㅂ)
파찰음(affricate)	/t͡ɕ/ (ㅈ)
마찰음(fricative)	/s/ (ㅅ)
폐쇄음(stop)	/k/ (6회 ㄱ)

III. 음성 인식 시스템의 구성

1. 시스템의 구성

한국어 연속 숫자음을 인식하기 위한 전체 구성은 아래의 그림 3-1과 같다.

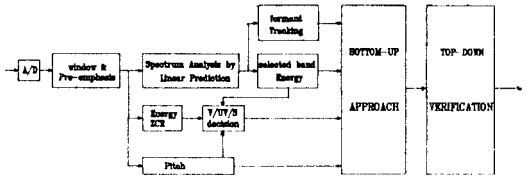


그림 3-1 시스템 구성

Microphone을 통해 입력된 음성 신호는 통과대를 통과하여 증폭된 후, 4KHz의 차단 주파수를 갖는 저역 여파기를 통과하여 12-bit A/D 변환기에 의해서 10KHz의 표본화 주파수로 표본화된다. A/D 변환된 신호는 기본 신호 처리로 25.6ms의 Hamming window를 통과하여 optimal pre-emphasis 된다. 그리고 14차의 선형예측 분석법이 매 5msec 마다 수행되며, 선형예측 계수는 Markel[8]의 autocorrelation 방법을 사용했다. 각 선형 예측계수들은 256 point/40Hz resolution의 스펙트럼 분석되어 5개 대역의 에너지를 구하게 된다. 이때 고려된 에너지 대역은 다음과 같다.

ferg0 :	0	5000Hz
ferg1 :	300	5000Hz
ferg2 :	100	900Hz
ferg3 :	640	2800Hz
ferg4 :	3700	5000Hz

기본 주파수의 추출 방법은 SIFT[7] 방법을 수정하여 사용했으며, 포오만트 추출은 McCandless[8,9]의 알고리즘을 수정하여 사용했다.

2. 유성음/무성음/무음구간 결정

무음(silence) 구간과 음성 구간의 구별은 대역 에너지 ferg1을 이용한다. 이 에너지가 300에서 5000Hz의 대역을 이용하는 것은 voiced bar, 또는 voiced stop은 무음 구간에 포함시키기 위해서이다. ferg1의 함수에서 음성 신호 시작 전후의 각 100msec 동안의 배경 잡음에 대한 평균 값을 얻은 후, 이 값에대해 3dB 이상의 frame에 대해서는 음성구간으로 간주하며 그 미만에 대해서는 무음 구간으로 간주한다. 이때 무음 구간 중에도 pitch가 0이 아닌 경우를 특히 voiced silence라고 명명한다.

threshold = background noiselevel + 3dB (3-1)

음성 구간에 대해서는 pitch의 존재에 따라서 간단히 유성음 구간과 무성음 구간을 구분한다.

IV. 사전 분할 및 음소 분류 (Bottom-Up Approach)

Bottom-Up approach는 그림 4-1과 같이 사전 분할(primary sementation) 부분과 음소 분류(phoneme classification) 부분으로 구성되어있다.

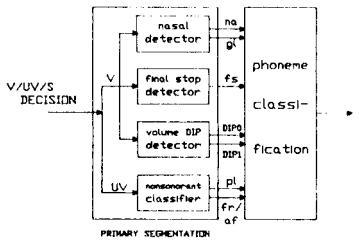


그림 4-1 Bottom-Up Approach

1. 비음 구간의 검출 및 분류

일반적으로 연속음 인식에서는 3절에서 정의될 volume DIP을 이용하여 비음및 기타 유성 자음 또는 낮은 에너지의 모음 구간을 검출한다[11]. 그러나, 한국어 숫자음의 경우 조성자음의 영향으로 연속된 다음 단어와의 조음현상이 잘 일어 나지않음으로써 DIP으로 검출하지 못하는 구간이 자주 발생하게된다. 이것을 방지하고자 시간영역에서의 비음구간 검출기를 선행 시켰다.

그림 4-2 에서와같이 각 frame 마다의 에너지 차를 간략화 시켜서 일정 패턴으로 비음 구간을 검출할 수 있다. Smoothing은 3 point Banning window(1/4, 1/2, 1/4)를 사용했으며, 에너지 차는 식(4-1)로 정의되고, median smoothing[12]을 이용했다. 그림 4-3에는 비음 구간의 패턴과 확인 알고리즘을 제시했으며 그림 4-4에는 그 결과를 보였다.

ergd(i) = out(i-1) - out(i) (4-1)

비음 패턴으로 검출된 부분은 비음과 유음 그리고 기타 유성음 구간으로 확인 알고리즘을 통해 1차 확인된후 음소분류를 하게된다. 1차 확인시의 파라미터는 다음과 같다.

rf24	ferg4와 ferg2의 평균 비율
rf01	평균 영교차율
r40	ferg0와 ferg4의 평균 비율
ae	평균 에너지
emax	입력 신호의 최대 에너지
af4	ferg4의 평균 에너지

그림 4-2 비음구간 검출도

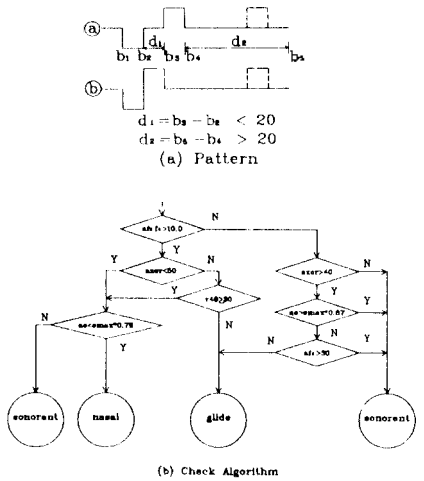


그림 4-3 비음구간 패턴및 확인 알고리즘

2. 종성 내파음구간의 검출 및 분류

종성 내파음(final impulsive stop) 구간의 검출도 비음 검출과 같이 간략화된 에너지차 패턴을 이용하여 구할 수 있다. 특히 종성 내파음 뒤에는 반드시 무음 구간이 오며, 에너지 차가 현격하다[13]는 사실을 이용하여 1차 확인후 음소 분류를 거치게 된다.

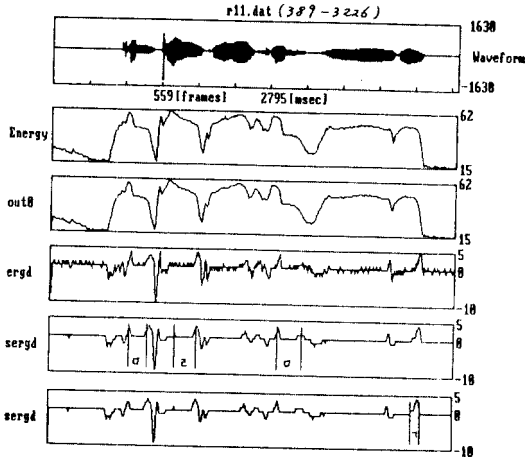


그림 4-4 비음구간및 종성내파음구간 검출결과

3. DIP에의한 모음 구간의 검출 및 분류

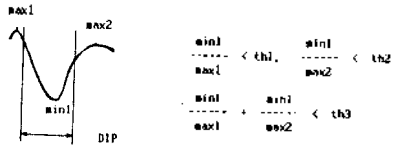


그림 4-5 DIP의 정의

DIP의 조건은 위와 같으며, 조건을 만족하는 골짜기에서의 최대 기울기를 가지는 두점 사이를 DIP으로 정의한다[11]. 실험에 사용된 에너지대역은 640-2800Hz(DIP0)와 0-5000Hz(DIP1)이다. 검출된 DIP은 각각 독립적으로 취급되어 음소 분류를 거치게된다.

검출된 DIP과 나머지 유성음 구간에 대하여 독립적으로 포오만트 및 파라미터 확인을 통하여 음소 분류를 행한다. DIP0과 DIP1은 편집 규칙(editing rule)에 의해서 결합되며, 편집규칙은 음소의 선호도를 반영한다.

4. 무성음 구간의 분류

무성음구간은 사전 분류에의하여 파열음군 /p, k/와 마찰음군 /s, ts/으로 분류되며, 확인 되지않은 무성음 구간에 대해서는 무성음 구간(ns: nonsonorant)으로 둔다. 무성음 구간의 분류 기준은 통계적 특성에의하며 무성음 구간의 분류시는 별도로 포오만트 주파수의 확인을 하지 않았다. 사용된 파라미터는 평균 영교차율, pre-emphasis factor, af2f4 등이다.

5.포오만트 주파수에 의한 음소 분류

사전 분할된 각 부분중 유성음인 경우에는 포오만트 주파수를 확인한다. 포오만트 주파수는 첫째, 둘째, 세째 포오만트 F1, F2, F3가 고려되었으며 유성음 구간의 전이성도를 판별하기위하여, 평균 F2 값에대한 F2의 변화량 을 고려했다. 본 논문에서는 이 변동량이 300Hz를 넘을 때는 급격히 변화하고 있다고 판별했다. 그리고 판별되지 못한유성음 구간에 대 해서는 /?/로 명명했다.

표 4-1 모음의 포오만트 주파수

모음	포오만트 영역
/ㅓ/	F1 : 550 - 800 F2 : 1100 - 1400 F3 : 2500 - 2800
/ㅚ/	F1 : 200 - 390 F2 : 1600 - 2250 F3 : 2500 - 3280
/ㅜ/	F1 : 200 - 400 F2 : 600 - 800 , 1450 - 1700 F3 : 1800 - 2750
/ㅐ/	F1 : 410 - 580 F2 : 1900 - 1990 F3 : 2410 - 2750
/ㅗ/	F1 : 300 - 430 F2 : 520 - 790 , 1500 - 2250 F3 : 2050 - 2980

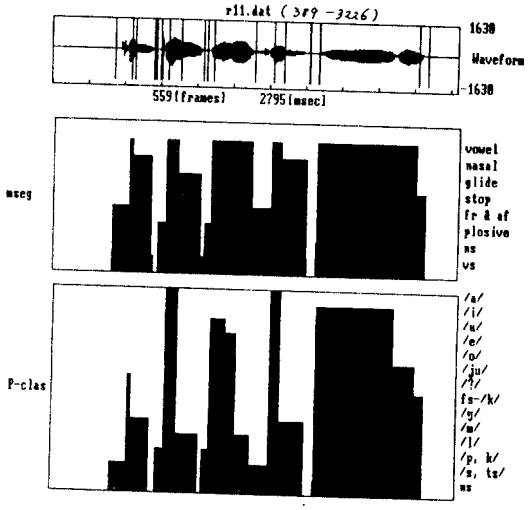
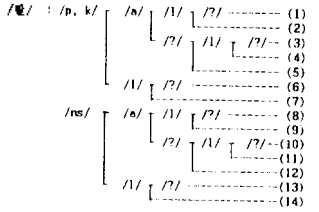


그림 4-6 사전 분할및 음소 분류 결과

V. Top-Down Verification

이부분에서는 각 숫자음에 대한 음소가능 조합을 규정하여 음소 분류결과를 편집하게 된다. 음소 가능 조합에는 연속음 발음시에 발생하는 조음 현상이 포함되며, 또한 음소 분류시의 가능한 오분류 조합도 포함됨으로써 음소 분류시의 오분류를 수정할 수 있게된다. 그에로써 숫자음 /팔/에 대한 음소 조합은 다음과 같다.



VI. 실험 및 결과

1. 실 험

실험에 사용된 데이타는 일반 컴퓨터실의 잡음환경에서 20대 남성 화자에의해 발음된 11개의 전화 번호이며, 발음 속도는 일상의 빠른 대화속도에 맞추어 2.8초 동안에 7자리 전화 번호를 발음함으로써 평균 3.7 음절/초이다. 음소 분류시에는 각 데이타의 통계자료를 근거로 분류되었으며 Top-Down 확인단에서는 숫자음에 대한 음소 조합 정보와 /에/모음에 대한 정보만을 줌으로써, 임의의 연속 숫자음의 인식이 가능케 했다.

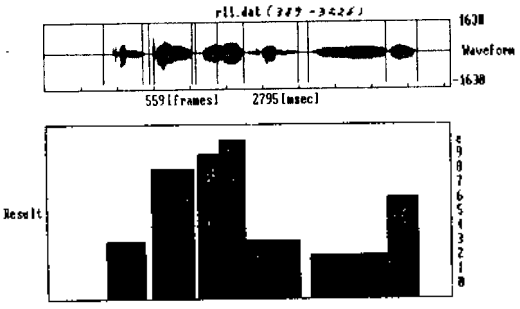


그림 5-1 최종 인식 결과

표 5-1 음소 분류 결과

음소	분류된 음소												
	a	i	u	o	ju	/?/	fs	k	ŋ	m	l	/p,k/s,ts/	e
a	25					3					3		
i		9				3					7		
u			2	1		9							
o				10		3					1		
e	1	1				1					2		9
ju					6								
fs-k							6						
ŋ								4	1		1		
m									8				
l											16		
p						3						13	
k												14	
s			1										18
ts													7

표 5-2 최종 인식 결과

	데 이 터	인식 결과	비 교
1	389-3226	389-3226	
2	446-8989	446-8989	
3	878-2670	878-2670	
4	434-3599	434-3599	
5	879-5290	879-5290	
6	482-6556	4826556	- → 2
7	385-2624	38 2824	5 →
8	878-2037	878-2037	
9	647-4173	64 4173	7 → -
10	355-0109	355-0109	
11	446-8989	446-8989	

2. 실험 결과 및 결론

음소 분류 결과 무성음구간은 모두 올체 분류되었으며, 비음 검출기는 모든 비음을 검출하는 외에 묽음 및 유성음 부분을 검출했으나 뒤에 오는 음소 분류시 모두 올바로 판정되었다. 중성 내파음의 경우에도 모두 바른게 판정되었다.그러나 모음의 판정에는 음소 분류시 많은 오판정을 했다. 특히 한 모음구간이 여러 DP으로 나뉘어 질 경우에는 한 모음 내에서도 여러 모음으로 판명되는 경우가 다소있었으며 중성 자음으로 비음이나 묽음이 오는 경우에는 중성 모음을 구별하지 못하는 경우가 있었다. 그에로 /아/ 모음의 경우 삽과 팔의 비음및 묽음에 의하여 비확인 모음(/?/)으로 분류된 경우가 10% 발생했으며, 특히 팔의 묽음 /마/의 영향으로 묽음(/l/)으로 분류된 경우가 10% 발생했다. 각 음소의 분류결과는 표 5-1에 있다.

최종인식 결과 11개의 전화 번호에있는 숫자음 77개에 대해서는 2개의 오인식이 발생해 97.4%의 인식률을 얻었으며 /에/ 모음을 고려한 전체 88개의 음절에 대해서는 4개의 오인식이 발생해 95.5%의 인식률을 얻었다. 표 5-2에있는 Top-Down approach후의 최종 인식결과를 보면 주된 오인식이 모음 /에/에 있음을 알 수 있다. 이것은 /에/ 모음의 앞 또는 뒤에 중성 모음이 올 경우에, /에/ 모음이 중성 모음과 유사해지거나 또는 중성 모음이 /에/ 모음과 유사하게 되어 /에/모음과 중성 모음의 중간적인 성질을 띄게 되기 때문이다. 한편 77개의 숫자음 중 2개의 오인식의 경우 '7 '가 /치레/로 발음된 경우와 '5-'가 /에/로 판정된 경우이므로 순수 숫자음만을 가지고 실험할 경우 더 높은 인식률이 예상된다.

VI. 결 론

본 논문은 한국어 연속음중 제한된 대상인 숫자음에 대한 음소단위의 인식을 시도 했다. 인식기의 구성은 일반적인 연속음인식 시스템의 구성을 따랐으며 optimal pre-emphasis factor를 파라미터의 하나로 사용했고, SIFT 알고리즘과 McCandless의 포오만트 추적 알고리즘을 수정해서 적용했다. 그리고 간략화된 에너지차의 패턴을 이용하는, 한국어 숫자음에 맞는 비음구간의 검출 방법과 중성 내파음의 검출 방법을 제안했다. 그리고 '-'를 /에/로 한 화자가 발음한 전화번호를 대상으로 인식 실험한 결과 숫자음만의 경우 97.4%, /에/를 포함한 경우 95.5%의 인식률을 얻었다. 본논문은 한국어 연속음 인식 시스템의 구성을 위한 앞으로의 시스템구현에 발판이 될 수 있겠으며, 특히 새로이 제안된 비음 및 중성 내파음의 검출 방법은 앞으로도 유용하게 사용될 수 있으리라고 본다. 그리고 앞으로의 과제는 한국어의 모든 음소를 분류하기 위한 분류 파라미터의 개발과 그 시스템의 구현, 그리고 무제한의 화자 인식을 위한 정규화의 방법 연구등이될 것이다.

참고 문헌

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

[2] Jean-Paul Haton, Automatic Speech Analysis and Recognition D,Reidel Publishing Company, 1981.

[3] 이 철희, "한국어 연속음 중 단모음 인식에 관한 연구," 서울대학교 대학원 공학석사학위 논문, 1986.

[4] 이 용주, 김 경태, 차 균현, "한국어 단모음의 성별, 연령별 특징 변화 및 인식," 음성 통신 및 처리 기술 Workshop 논문집, 1987.

[5] D.Ter Haar, Mechanisms of Speech Recognition, International series in natural Philosophy, Vol,85, Pergamond Press.

[6] J.D,Markel and A.H,Gray, Jr., Linear Prediction of Speech, Springer Verlag Berlin Heidelberg, New York, 1976.

[7] Markel, J.D., "The SIFT Algorithm fr Fundamental Frequency Estimation," IEEE Trans, AU-20, 367-377, 1972.

[8] Stephanie S.McCandless, "An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra," IEEE Trans, ASSP-22, No.2,135 141 ,April, 1974.

[9] Stephanie Seneff, "Modifications to Formant Tracking Algorithm of April 1974," IEEE Trans, ASSP-22, pp,192-193, April, 1976.

[10]배명진,임재열,안수길,"음성 발생 모델로 부터 G-peak를 이용한 음성 에너지 추출에 관한 연구," KIEE, Vol, 24, pp, 12-17,1987

[11]Clifford J, Weinstein, Stephanie S,McCandless, LEE F,Mondshein, and Victor W,Zue, "A System for Acoustic Phonetic Analysis of Continuous Speech," IEEE Trans, ASSP-23, No,1, pp,54-67, Feb., 1975.

[12]L,R,Rabiner, M,R,Sambur, and C,E,Schmidt, "Applications of a Nonlinear Smoothing Algorithm to Speech Processing," IEEE Trans, ASSP-23, No,6, pp,552-557, Dec., 1975.

[13]최 정아, 배 명진, 안 수길, "음성 신호에서의 중성내파음 구간 검출에 관한 연구," 음성통신 및 처리기술 Workshop 논문집, 1987.