論文 89-26-4-19

프레임간 에너지 차를 이용한 음성신호의 종성 폐쇄음 구간 검출에 관한 연구

(On the Interval Detection of Implosive Stop Sounds by Frame Energy Difference)

裵 明 振*, 崔 貞 雅*, 安 秀 桔*

(Myung Jin Bae, Jung Ah Choi and Souguil Ann)

要 約

음성 신호의 인식 시스템에서 분류 인식과정은 최종 인식과정에 필요한 복잡한 처리 과정을 줄여주기 때문에 유용하다. 본 논문에서는 종성 폐쇄음의 구간을 검출하는 새로운 분류인식 알고리듬을 제안하였다. 한국어의 종성 폐쇄음은 항상 모음 뒤에 발음되며, 그 특징은 모음 구간 내에 포함된다. 종성 폐쇄음 발음시에 연구개가 급히 닫히므로 에너지의 급격한 감소가 일어나고, 폐쇄 구간은 50~150 msec간 지속된다. 이러한 성질을 잘 나타내는 파라미터로 프레임간 에너지 차를 제안하였다.

Abstract

Preprocessing in speech recognition system is useful, for it reduces some of the complicated procedures required for the final recognition. In this paper, we suggest a new preprocessing algorithm for detecting the intervals of implosive stop sounds. Implosive stop sounds follow vowels in Korean language, and its characteristic is included in the region of vowels. When an implosive stop is pronounced, the velum is quickly closed, thus its energy decays abruptly and the closure lasts for about 50 to 150 msec. The enegy difference between adjacent frames is chosen as a parameter which represents well the above features.

I. 서 론

인간이 기계를 사용하게 된 이후로 기계와 인간의 자연스런 커뮤니케이션 방법이 끊임없이 연구되어 왔다. 그 하나가 일상의 의사소통 수단인 음성을 이용하는 것이다. 근래 기계와 전자공학의 눈부신 발달로 음성 인식의 필요성이 날로 증대하고 있으며 그에 대한 활발한 연구가 진행중이다.

음성 인식에 관한 연구는 이미 1950년대부터 진행이 되어 인식기가 개발되어 왔으나, 실용화는 빨리 이루어지지 못하였다. 1970년대 이후, 컴퓨터를 이용한 인식 방법이 연구되어, 제한된 범위의 고립 단어에 대해서는 이미 실용화 단계에 와있다. 또, 제한

*正會員, 서울大學校 電子工學科
(Dept. of Elec. Eng., Seoul Nat'l Univ.)
接受日字 : 1988年 7月 22日

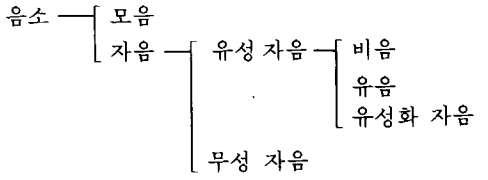
된 어휘, 정해진 문법의 범위 안에서는 문장 단위 인식도 가능하게 되어 Nippon의 DP-series나 Verbex 등의 인식기가 개발되어 있다. 이들은 주로 pattern matching 방법을 사용하고 있다.[11] 그러나, 단어수가 많아지면 그 계산량이 방대해질 뿐만 아니라, 조사 및 어미변화, 조음현상(coarticulation) 때문에 단어의 경계가 모호해져, 패턴의 표준화가 어려워진다.

따라서 궁극 목표인 연속음 인식을 위해서는 음소 단위의 인식이 바람직한 접근법이다. 인간이 음성을 인식할 때는 음운, 의미론, 운율, 문법, 상황등의 여러 가지를 종합해서 그 의미를 파악하게 된다. 인식기에도 이러한 언어학적 정보를 도입시켰을 때 높은 인식율을 얻을 수 있다. 이미 1973년에 Klatt와 Stevens가 스펙트로그램을 이용해 음소의 구별을 꾀하였고, 여기에 언어학적인 정보를 첨가시켰을 때 90 % 정도의 구별이 가능함을 밝혔다.[10] 음소 단위의 인식의 경우, 각 음소의 특징적인 파라미터를 추출하여 그 특징에 따라 몇가지 유사군(cluster)으로 분류하는 전처리 단계인 분류인식을 거쳐 최종 인식을 하는 것이 현재의 추세이다. 이렇게 하는 경우 최종 인식의 인식율을 높이고, 시간도 절약할 수 있다.[4]

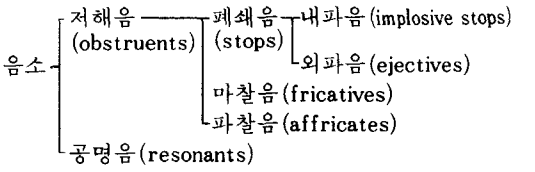
특히 한국어를 보면 초성, 중성, 종성의 세 부분이 어울려 한 음절을 형성한다. 음소 단위 인식시 음소의 특징에 따라 시간영역에서 분류한 후, 주파수 영역에서 이에 대한 최종 인식의 법칙을 적용하는 것이 효과적이다.[5] 종성의 폐쇄음을 특히 내파음이라하는데, 내파음은 전이 파형을 가지므로 넓은 프레임 단위로 분석하면 평균 효과가 일어나 검출이 어렵다. 이 논문에서는 시간영역에서의 에너지 차이를 이용해 내파음 구간을 검출하는 새로운 알고리듬을 제시하였다.

II. 음성 신호의 특징과 분류

한국어의 음소를 언어학적으로 분류하면 다음과 같다.



이것을 음성 인식을 위해 조음 방법에 따라 분류해 보면 다음과 같다.[9]



한국어에서 종성으로는 형태상으로 14개의 기본 자음과 그 외에 겹받침이 가능하다. 그 조음방법에 따라 공명음과 저해음으로 나뉘어 볼 수 있는데, 이중 저해음에 대한 분류를 나타내면 표 1과 같다.

표 1. 한국어의 저해음 분류
Table 1. Classification of obstruents of korean language.

조음 방법	힘	조음점		
		입술	이 뒷부분	연구개 근처
파열음	기본	ㅂ	ㄷ	ㄱ
	경음	ㅃ	ㄸ	ㄲ
	격음	ㅍ	ㅌ	ㅋ
파찰음	기본		ㅈ	
	경음		ㅉ	
	격음		ㅊ	
마찰음	기본		ㅅ	
	경음		ㅆ	

종성에 저해음이 오는 경우 음절이 단독 발음되거나, 뒤에 비음 이외의 자음으로 시작되는 음절이 나타날 때는 그 계열의 가장 윗 계열 소리인 /ㄱ, ㄷ, ㅂ/으로 중화된다. 그 예를 살펴보면,

'ㄱ': 호박벌, 남녘땅, 깎다
'ㄷ': 옷, 꽃집, 밑바닥
'ㅂ': 입고서, 앞뜰, 잎

이들의 음성학적 특징은 공기 흐름의 측면에서 볼때, 초성 발성시와는 달리 공기가 밖으로 유출되어 나는 소리가 아니라 일단 폐에서 유출된 공기를 발성중 어떤 부분을 급격히 닫음으로써 생성된다는 점이다.[8,9]

III. 폐쇄음 검출의 기존방법

초성 폐쇄음을 발성한 경우에 다른 음소 부류와 구분짓는 데는 닫혀진 공기를 방출하는 순간에 에너지의 파열에 뒤따르는 무음 구간을 이용하는 방법이 제안되었다. 또한 유성 폐쇄음과 무성 폐쇄음을 구분하는 방법은 VOT(voicing onset time)을 사용한 것

이 있다. VOT는 닫혀진 공기를 방출할 때부터 후두가 진동을 시작할 때까지 걸리는 시간을 말한다. 어떤 유성 파열음인 경우에는 후두가 닫혀진 주기 동안에도 진동을 계속하게 된다. 이것은 스펙트럼에서 "voice bar"로 나타난다.[10]

VOT외에도 다른 성질이 폐쇄음에서 나타날 수 있는데 닫혀진 구간이 75msec보다 작으면 유성 폐쇄음으로, 130msec보다 길면 무성 폐쇄음으로 인식한다.[10] 최근의 방법은 기본 주파수가 천천히 증가하면 유성 파열음으로 인식하고, 기본 주파수가 급히 낮아지면 무성 파열음으로 인식하게 된다.

초성 폐쇄음을 인식하기 위한 많은 연구가 진행되어 왔지만 종성 폐쇄음인 내파음(implosive stops) 검출에 대한 연구는 거의 이루어지지 않았다. 다만 내파음의 성분은 중성을 이루는 모음 부분에 포함된다는 사실이 알려져 있다.[10] 이것을 프레임 단위로 처리한다면 평균 효과가 일어나 내파음의 전이 구간을 잘 관측할 수 없게된다. 또한 주파수 영역에서 분석하려면 내파음이 발생하는 구간을 시간 영역에서 정확히 찾은 다음에 분석을 수행해야만 성질을 파악할 수 있게 된다. 따라서 음성 신호를 인식하고자 할 때 구간 검출이 선행되어야 하는데 본격적인 분석법을 적용하기 전에 분류 인식 과정으로 내파음이 발생하는 구간을 찾는 방법을 제안하고자 한다.

Ⅳ. 종성 폐쇄음 구간 검출

내파음이 속해 있는 파형의 성질을 살펴보기 위하여 /오/와/육/의 파형과 그에 대한 에너지 변화(energy contour)를 그림 1과 그림 2에 각각 나타내었다. 여기에서는 에너지를 계산하는 대신에 진폭의 절대값의 합을 사용하였다.

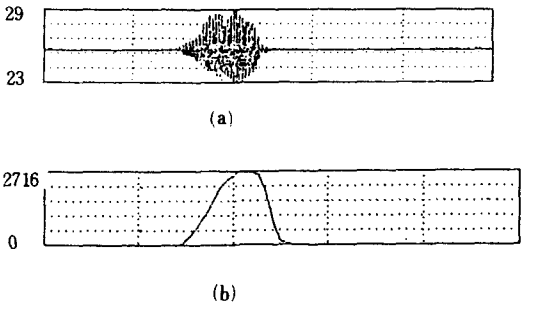


그림 1. 음성 신호 /오/에 대한 파형과 에너지
Fig. 1. Speech waveform and energy contour for /ou/.

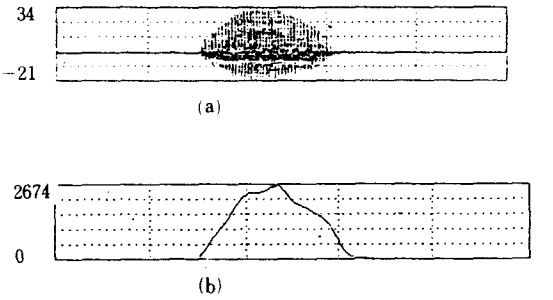


그림 2. 음성 신호 /육/에 대한 파형과 에너지
Fig. 2. Speech waveform and energy contour for /yuk/.

$$E(i)=\sum_{n=i}^{i+N-1}|S(n)| \qquad (1)$$

여기서는 8KHz의 표본 주파수로 표본화된 음성 신호의 표본 256개(32msec)를 한 프레임의 길이로 사용하였으며 반 프레임씩 겹치게 하여 절대값 진폭의 합을 계산하였다.

그림 2(b)를 살펴보면 내파음 /ㄱ/받침은 별도의 파형으로 나타나지 않고 중성 /ㅜ/에 포함되어 나타나고 있음을 알 수 있다. 내파음이 종성에 나타나면 그림 1(b)의 에너지 변화에 비해 빠른 시간에 연구개를 막아버리기 때문에 에너지의 감소 현상이 아주 급히 일어나고 있음을 알 수 있다. 또한 일정 기간 동안 공기를 막은 상태가 지속되므로 일정 기간의 무음 구간이 있은 후 다음 음이 발음된다는 점이다.

우선 에너지가 급속히 감소하는 현상을 검출하기 위해서는 프레임간 에너지의 차이를 계산하고 그 값이 최대 에너지에 비해 어느 정도인가를 계산할 필요가 있다. 인접한 프레임간 에너지 차이는 다음과 같이 계산할 수 있다.

$$\begin{aligned}D(i)&=E(i)-E(i+1)\\&=\sum_{n=i}^{i+N/2-1}|S(n)|+\sum_{n=i+N/2}^{i+N-1}|S(n)|\\&\quad-\sum_{n=i+N/2}^{i+N-1}|S(n)|-\sum_{n=i+N}^{i+3N/2-1}|S(n)|\\&=\sum_{n=i}^{i+N/2-1}|S(n)|-\sum_{n=i+N}^{i+3N/2-1}|S(n)|\\&=Eh(i)-Eh(i+2)\end{aligned} \qquad (2)$$

(E(.) : 프레임 평균에너지
D(.) : 프레임간 에너지 차
Eh(.) : 반 프레임 평균에너지)

즉, 반 프레임 구간마다 평균에너지를 구한 다음에 하나 걸러서 계산된 반 프레임 에너지를 빼면 인근

한 프레임간의 에너지 차이를 구할 수 있다. 반면 한 프레임의 에너지는 인근한 반 프레임간의 에너지값을 더하기만 하면 식(1)의 결과가 얻어지기 때문에 인근한 프레임간의 에너지 차이를 계산하기 위한 별도의 계산 과정이 필요하지 않게 된다.

이렇게 계산된 프레임간 에너지 차의 형태를 그림 3에 나타내었다. 내파음이 포함되지 않은 /오/를 나타내는 그림 3(a)와 포함된 /육/의 그림 3(b)의 형태를 보면 연구개를 빨리 닫기 때문에 종성부분에서 +쪽으로 에너지 변화가 상당히 큼을 알 수 있다.

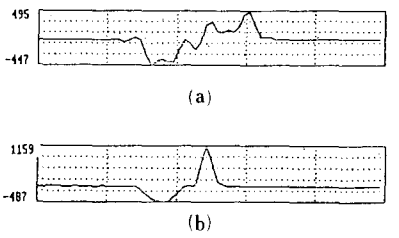


그림 3. 숫자음 /오/와 /육/의 프레임간 에너지 차
Fig. 3. Frame energy difference for speech signals /ou/ and /yuk/.

실험적으로 구해 본 결과에 의하면 종성 폐쇄음의 특징은, 내파음이 발생할 때는 중성의 유성음이 갖는 최대 에너지값에 비해 프레임간 에너지 차이가 30% 이상인 구간이 항상 발생하였다. 다음에는 연구개가 닫히는 시간이 내파음의 존재 유무를 결정짓는 요인이 된다. 종성에서 내파음이 발생할 때는 에너지가 감소하는 시간이 48msec에서 160msec정도를 소요한다는 사실을 실험을 통해 확인하였다. 에너지가 영으로 떨어지는 구간을 측정할 때에는 식(2)로 계산된 그림 3의 프레임간 에너지 차에서 구하는 것이 좋다. 내파음은 무음 구간을 가져야 하기 때문에 무음이 검출되는 부분에서부터 뒤로 에너지의 차를 계산하면 +쪽으로 봉우리가 나타나게 되는데 이 봉우리가 차지하는 프레임의 수를 측정하면 된다. +인 피크가 끝나는 프레임의 에너지값이 중성 최대 에너지와 비슷한데, 이것은 중성 최대 에너지의 95%이상이다. 또, 비음이나 유음 구간과의 차이는 프레임간 최대 에너지 차가 0으로 떨어지는 데는 10~50msec가 소요된다는 것이다.

V. 실험 및 결과

종성 폐쇄음이 존재하는 구간을 검출할 때 사용한 시뮬레이션 장비는 A/D와 D/A 변환기가 부착된 IBM-PC/AT를 사용하였다. 우선 숫자음 (0-9)을 남성 화자 8명과 여성 화자 2명이 각 5번씩 마이크로폰을 통해 직접 발음할 때에 표본화된 데이타를 하드 디스크에 저장시켰다.

처리 과정은 그림 4의 순서도에 따라 처리하였다. 유성/무성/무음 구간 검출 동안에 반 프레임마다 평균에너지와 최대 에너지값을 함께 계산하였다. 일정 길이의 모음 구간이 유성/무성/무음 구간 검출 시에 찾아지고 나서 무음 구간이 찾아지면 진행순서를 거꾸로 하여 내파음 구간을 검출하게 된다.

Ⅳ에 열거된 특징에 따라 판정 기준은, 최대 에너지값을 Emx, 프레임간의 에너지 차가 +일 동안의 봉우리의 최대값을 Dmx, 봉우리가 끝나는 부분의 에너지값을 Emx1, 봉우리가 차지하는 프레임 수를 I, 최대 차 에너지로부터 차 에너지가 0이 되는 때 까지의 프레임 수를 J라 할 때,

$$Dmx > (Emx1 * 0.3)$$
$$3 \le I \le 10$$
$$Emx1 > (Emx * 0.95)$$
$$0 < J < 4 \qquad (3)$$

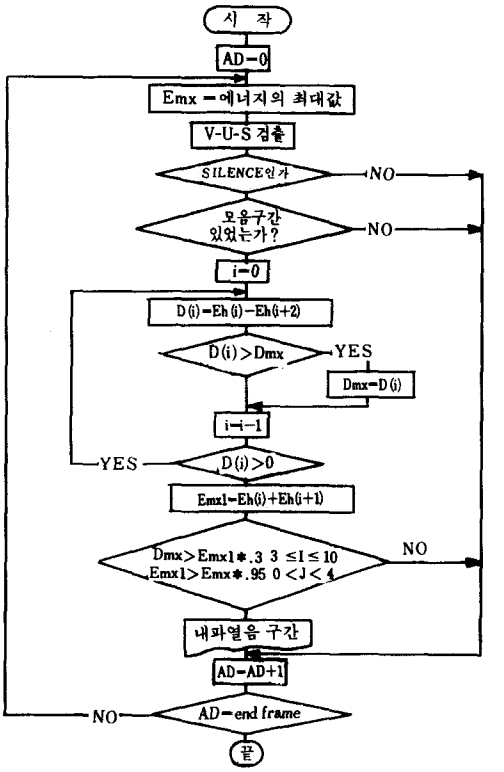


그림 4. 내파열음 구간을 찾기 위한 순서도
Fig. 4. Flowchart for the detection of final stop interval.

표 2. 화자 ㄱ의 숫자음 발음 결과
Table 2. Result for speaker ㄱ.

File Name	J	I	Dmx/Emx1	Emx1/Emx	SCORE
kkh1-00.dat	12	14	.1889151	0	0
kkh1-11.dat	0	0	0	3.264953E-02	0
kkh1-22.dat	0	0	0	3.942584E-02	0
kkh1-33.dat	0	2	1.265823E-02	3.164557E-02	0
kkh1-44.dat	1	2	.0183953	3.287671E-02	0
kkh1-55.dat	1	2	3.520242E-03	8.297712E-03	0
kkh1-66.dat	2	7	.3158329	.9998349	4
kkh1-77.dat	0	2	4.154166E-03	1.015463E-02	0
kkh1-88.dat	7	14	.1576635	1.137245E-02	0
kkh1-99.dat	11	13	.2654288	.8656057	0

표 3. 화자 ㄴ의 숫자음 발음 결과
Table 3. Result for speaker ㄴ.

File Name	J	I	Dmx/Emx1	Emx1/Emx	SCORE
kds1-00.dat	9	14	.1545933	0	0
kds1-11.dat	11	13	.1773519	.9212543	0
kds1-22.dat	3	14	.1091405	1.803547	0
kds1-33.dat	12	14	.2427459	.935598	0
kds1-44.dat	7	14	.2019517	.7167254	0
kds1-55.dat	4	13	.1850467	.9996262	1
kds1-66.dat	3	7	.4265734	.9996319	4
kds1-77.dat	0	2	1.274993E-02	2.260214E-02	0
kds1-88.dat	9	14	.136376	2.118989E-02	0
kds1-99.dat	5	10	.2002618	.9613875	1

의 각 조건을 만족할 때 score를 1씩 부여해 score가 3이상이 되면 종성 폐쇄음 구간으로 판정하였다. 이 과정의 순서도는 그림 4에 제시하였다. 화자 ㄱ, ㄴ에 대한 결과는 표 2, 3에 보였다.

실험에 사용한 500개의 단음절어 중에서 종성에 폐쇄음 구간을 가진 단어는 숫자음 /육/이다. 이 단어들에 대해 위에 제시한 종성 폐쇄음 구간 검출 과정을 수행하였을 때 분류된 단어는 /육/ 발음 모두와 하나의 /이/ 발음이었다. 이것은 발음 연습을 하지않은 화자가 /이/를 발성할 때 짧고 강하게 발음하였기 때문이다.

이 알고리듬은 연결 단어의 중간 음절에 종성 폐쇄음이 존재하는 경우의 구간 검출에 더 잘 적용될 수 있다. 단음절어는 종성에 폐쇄음이 존재하지 않을 때에도 에너지의 급격한 감소와 무음 구간이 뒤따르나 중간 음절에서는 종성 폐쇄음이 존재할 때에만 이러한 현상이 일어나기 때문이다. 연결 숫자음에 대한 실험 결과를 표 4에 제시하였다. 음절 중간의 종성 폐쇄음도 검출할 수 있음을 알 수 있다.

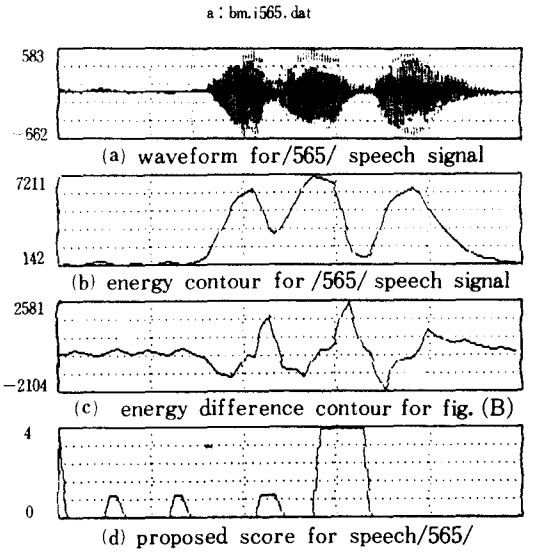


그림 5. 연결숫자음 /565/에 대한 결과파형
Fig. 5. Result waveform of connected digits/ /565/.

표 4. 연결 숫자음/555/와 /565/에 대한 파라미터와 처리결과
Table 4. Parameters and processing results of connected digits/555/ and /565/.

File Name	J	I	Dmx/Emx1	Emx1/Emx	SCORE
bmj555.dat	0	0	.1561822	.5771149	0
bmj555.dat	2	3	.2603037	.6920824	1
bmj555.dat	0	0	9.174312E-02	.4261468	0
bmj555.dat	0	1	.1176471	.6735294	0
bmj555.dat	0	0	3.069502E-03	6.688219E-02	0
bmj555.dat	0	0	9.427757E-03	6.688219E-02	0
bmj555.dat	2	8	.1935979	.7042425	2
bmj565.dat	0	1	.3227848	.663924	1
bmj565.dat	0	1	.2091255	.4340304	0
bmj565.dat	0	1	.3778802	.6513825	1
bmj565.dat	1	2	.309403	.6313614	1
bmj565.dat	2	6	.3578758	.9480865	4

Ⅵ. 결　　론

음성을 인식하는 방법은 상당히 복잡할 뿐 아니라 처리 과정이 종속적으로 이루어지고 있기 때문에 처

리 방법을 분산시켜 각각을 병행해서 처리할 수 있도록 하는 방법이 필요하다. 이에 대한 해결법의 하나는 본격적인 인식을 수행하기전에 간단하고 통계적인 방법을 적용하여 기능별 분류를 먼저 수행하는 분류 인식 방법을 사용하는 것이다. 본 논문에서는 분류 인식의 하나로 종성 폐쇄음의 구간을 검출하는 새로운 방법을 제안하였다.

종성 폐쇄음이 갖는 성질중에 중요한 것은 에너지의 감소가 급속히 일어나 감소되는 시간이 보통 50~150msec 정도를 차지하고, 그리고 얼마간의 무음 구간이 발생한다는 사실이다. 본 논문에서는 이러한 사실을 잘 나타내는 파라미터로 프레임간의 에너지 차를 제안하였다.

프레임간의 에너지 차를 이용하면 분류 인식용 파라미터로 많이 사용하고 있는 기존의 에너지 파라미터를 그대로 사용하기 때문에 별도의 계산 과정이 요구되지 않게된다. 프레임간 에너지 차를 파라미터로 사용하여 10명의 화자가 각 5번씩 발음한 숫자음에 대해 종성의 폐쇄음이 포함된 단어들을 찾았을 때 숫자 발음 /이/ 하나만이 잘못 검출되었고, /육/의 종성 폐쇄음 구간은 모두 검출되어 99.8%의 검출율을 얻었다.

參 考 文 獻

[1] L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Inc., 1978.

[2] L.R. Rabiner and Bernard Gold, *Theory and Application of Digital Signal Processing*, Prencice-Hall, Inc., 1975.

[3] Ian H. Witten, *Principles of Computer Speech*, Academic Press, 1982.

[4] Byeungwoo Jeon, "A study on the recognition of Korean isolated digits using clustering techniques," Seoul National University, MA Pap, Jan. 1987.

[5] P. Demichelis, R. DeMori, P. Lafaro, and M.O'Kane, "Computer recognition of plosive sounds using contextual information," *IEEE Trans.* ASSP-31, no. 2, pp. 359-377, Apr. 1983.

[6] G. Fant *Acoustic Theory of Speech Production*, Mouton,s' Gravenhage, 1960.

[7] Myung Jin Bae, Ik Joo Chung, and Souguil ANN, "The extraction of nasal sound using G-peak in continued speech," KIEE vol. 24, Mar. 1987.

[8] 허웅, 국어 음운학 -우리말 소리의 오늘, 어제-, 샘 문화사, 1985.

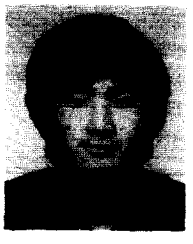
[9] 정명우, 백용학, 송석만, 이정수, 이원국, 현대 음운론 개설, 서린 문화사, 1982.

[10] International Series in Natural Philosophy, *Mechanisms of Speech Recognition*, vol. 85, Pergamon Press, 1978.

[11] Jean-Paul Haton, *Automatic Speech Analysis and Recognition*, D. Reidel Publishing Company, 1982.

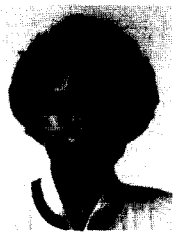
[12] Jungah CHOI, Myungjin BAE, Souguil ANN, "On the interval detection of an implosive stop sound in speech signals," 음성통신 및 처리기술 WORKSHOP 논문집, 한국 음향 학회, MAY, 1987. *

著 者 紹 介



裵 明 振(正會員)

1957年 5月 20日生. 1981年 2月 숭실대학교 전자공학과 졸업 공학사학위 취득. 1983年 2月 서울대학교 대학원 전자공학과 졸업 공학석사학위 취득. 1987年 8月 서울대학교 대학원 전자공학과 박사과정 수료. 1989年 4月~현재 호서대학교 전자공학과 조교수. 주관심분야는 음성 신호 처리이며, 신호 처리및 통신공학등임.



崔 眞 雅(準會員)

1964年 10月 17日生. 1987年 2月 서울대학교 전자공학과 졸업 공학사학위 취득. 1989年 2月 서울대학교 대학원 전자공학과 졸업 공학석사학위 취득. 1989年 4月~ 현재 서울대학교 대학원 전자공학과 박사 과정.

●

安 秀 桔 (正會員) 第26卷 第3號 參照

현재 서울대학교 전자공학과 교수