음소 인식을 위한 특징 추출의 위치와 지속 시간 길이에 관한 연구

A Study on Duration Length and Place of Feature Extraction for Phoneme Recognition

김 범 국*, 정 현 열*
(Bum-Koog Kim*, Hyun-Yeol Chung*)

이 논문은 1994학년도 영남대학교 학술연구조성비에 의한것임

요 약

한국어 음성인식 시스템을 구현하기 위한 기초 연구로서 한국어 전음소를 대상으로 1) 각 음소의 특성을 가장 잘 나타내는 최적의 위치, 2) 최고의 인식률을 얻기 위한 적당한 지속시간길이를 찾기위해서 음소인식을 수행하였다.

인식실험을 위해 특징파라메터로 21차원 켑스트럼계수를 이용하여 베이즈 결정법칙으로서 세화자에 대한 종속인식실험을 행하였다.

인식실험결과 최고의 인식률을 보이는 최적의 특징추출의 위치는 모음에서는 10~50ms, 마찰음및 파찰음은 40~100ms, 비음, 유음은 10~50ms, 그리고 파열음은 10~50ms임을 알 수 있었다.

또, 35 전음소를 대상으로한 인식에 있어서는 최고의 인식률을 얻기위한 지속시간 정보의 길이는 60~70ms정도가 충분함을 알 수 있었다.

Abstract

As a basic research to realize Korean speech recognition system, phoneme recognition was carried out to find out : 1) the best place which represents each phoneme's characteristics, and 2) the reasonable length of duration for obtaining the best recognition rates.

For the recognition experiments, multi-speaker dependent recognition with Bayesian decision rule using 21 order of cepstral coefficient as a feature parameter was adopted.

It turned out that the best place of feature extraction for the highest recognition rates were 10~50ms in vowels, 40~100ms in fricatives and affricates, 10~50ms in nasals and liquids, and 10~50ms in plosives.

And about 70ms of duration was good enough for the recognition of all 35 phonemes.

I. 서 론

디지탈 신호처리기술과 통신기술및 컴퓨터의 발달로 인간과 기계간의 Man-Mechine Interface가 현실적 문제로 부각되고 있으며, 또한 최근에는 멀티미디어나 이동통신과 같은 다양한 정보매체를 통한 통신분야에 있어서도 음성 신호처리의 중요성이 대두되고 있다.

특히 음성인식분야에서는 1970년대부터 단모음, 숫자음 그리고 도시명등 주로 특정 단어를 대상으로

*영남대학교 전자공학과
접수일자 : 1994년 3월 31일

한 연구를 시작으로 최근에는 연속 숫자음, 단어 음성, 모음및 자음 음소, 대용량 단어 인식, 연속 음성 등에 대한 연구가 활발하게 진행되고 있으며, 자동통역 전화 시스템을 개발하기 위한 연구도 진행되고 있다[1-3].

숫자음이나 특정단어의 분석과 인식에 있어서와 같이 인식의 단위를 단어로 하는 경우 인식하려는 어휘수가 많아짐에 따라 대량의 기억 용량이 요구되며 비교적 많은 처리시간을 필요로 한다. 그러므로 음소를 인식의 기본단위로 하는 것이 유리하다. 음소를 인식의 기본단위로 한 음성 분석과 인식에 관한 연구로는 김 등의 음소를 이용한 한국어 음성 신호의 분석과 인식[4], 정 등의 파열자음의 분석[5], 파열자음의 인식[6]등의 연구가 있다.

이경우, 인식에 유용한 각음소편의 최소길이, 각음소내에서도 그음소의 성질을 가장 잘 나타내고 있는 부분이 어디인가에 대한 연구가 필요함에도 불구하고 아직 이에 대한 연구는 미진하다.

한편, Okada[7]등은 일본어의 어두 음소의 식별에서 특징추출 위치와 길이를 조사하였는데, 유성음에서는 파열시점등의 기준이 되는 frame보다도 1frame 전의 spectrum으로부터 4~5frame 구간이, 무성음에서는 음성 시단에 대응하는 frame으로부터 이후 4~5frame의 구간이 인식에 유효하다는 것을 보고하고 있으며, 또한 peaking 현상이 일어나지 않는 최적의 차원수는 패턴길이 1~5frame에 대해서 각각 15, 20, 20, 30, 30이라는 것을 확인하였다.

본 연구에서는 음소를 인식의 기본 단위로 하는 한국어 음성 인식 시스템 개발을 목표로 하여 그 기초연구로서 한국어 전 음소를 인식에 적당한 특징추출의 위치와 지속시간 정보의 길이의 인식에 대한 영향을 조사하기로 한다. 이를 위하여 특징 파라미터를 typical frame을 포함한 전, 후 1frame, 시단으로부터 3frame, 제2frame~4frame과 같이 시단으로부터 11frame까지 시점을 1frame씩 옮겨가면서 3frame씩의 시간방향 특징과 typical 1frame, typical frame을 포함한 전 1frame(2frame), typical frame을 포함한 후 1frame(2frame), typical frame을 포함한 전, 후 1frame, 시단으로부터 3frame, 4frame, 5frame, 6frame, 7frame, 8frame, 10가지 경우로 분류해서 인식 실험을 실시하여 인식에 유효한 특징추출의 위치와 인식을 위한 시간방향의 최적정보의 길이를 찾아내고자 한다.

이를 위해 먼저 인식시 사용할 인식알고리듬, 인식에 필요한 차원압축방법, 인식실험에 대해 기술한후 인식결과를 분석, 검토하여 인식을 위한 최적 특징추출의 위치, 최적 시간방향정보의 길이를 찾아내고자 한다.

Ⅱ. 베이즈 결정법칙

Spectral 거리척도는 표준적인 spectrum과 입력 spectrum의 유사성을 나타내는 척도이다. 그러나, 일반적으로 표준적인 spectrum을 설정하는 것이 어렵다. 특히, 음성은 context나 화자에 의해서 spectrum이 변동하기 때문에 입력 spectrum은 어떤 모집단으로부터 발생된 것이라고 보는 통계적 결정이론을 사용하는 것이 유리하리라 생각된다. 본 연구에서는 이의 대표적인 방법중의 하나인 bayes 결정법칙[8]을 이용하여 인식실험을 실시한다. 이하 이에 대해 설명한다.

인식해야 할 category(c_1, c_2,...,c_m)에 대한 특징 파라미터의 모집단을 각각 w_1, w_2, w_3,...,w_m로 하고, 관측된 pattern의 특징 파라미터를 vector $X=(x_1, x_2,...x_n)^t$으로 나타내면 X가 w_i로부터 생기는 확률은 $P(w_i|X)$이다.

따라서, 이 값이 최대가 되는 w_i가 결정되면 X의 category를 c_i로 판정하면 되는데 이것을 bayes의 결정법칙이라 한다.

이 확률을 다음과 같이 표현할 수 있다.

$$P(w_i|X)=\frac{P(w_i)\,P(X|w_i)}{\sum_{j=1}^{m} P(X|w_j)}=\frac{P(w_i)\,P(X|w_i)}{P(X)} \quad (1)$$

$P(w_i|X)$는 직접 구하기가 어렵지만 (1)식을 이용하면 $P(X)$는 w_i에는 관계가 없기 때문에 분자의 값이 최대가 되는 w_i를 구하면 된다.

$P(w_i)$는 category w_i의 선행확률이고 다수의 sample을 관측하므로서 근사적인 추정을 할 수 있다. 또한, $P(X|w_i)$는 w_i의 모집단으로부터 다수의 sample을 관측하므로서 추정할 수 있다.

$P(X|w_i)$를 다수의 sample로부터 구하는 경우 $P(X|w_i)$을 파라메터릭(parameteric) 표현으로 나타낼 수 없을 때는 히스토그램을 작성하므로서 확률을 추정할 수 있으나 X가 유한개의 값만을 가질 경우에

만 유용하고 일반적으로는 그렇지 못하다.

그러므로, $P(X|w_i)$을 어떤 종류의 함수형으로 가정해서 그 함수의 파라미터를 다수의 sample로부터 추정한 파라미터의 표현을 사용하는데 일반적으로 정규분포로 가정하는 경우가 많다.

이때

$$P(X|wi) = \frac{1}{(2\pi)^{n/2}|\sum_i|^{1/2}} \{-1/2\,(X-\mu_i)^t \sum i^{-1}\,(X-\mu_i)\} \quad (2)$$

여기서, μ_i는 X의 평균치, $\sum_i$는 공분산행렬,

$$\mu_i = E(X),\ \sum_i = E[(X-\mu_i)(X-\mu_i)^t]$$

로 된다.

결국 (1)식에 대수를 취하고 w_i에 관계없는 요소를 제거하면

$$g_i(x) = \log P(w_i) - 1/2\log|\sum i| - 1/2[X-\mu i)^t \sum i-1(X-\mu i)] \quad (3)$$

로 된다.

식(3)에서 $g_i(x)$가 최대가 되는 category c_i를 결정하면 된다.

다시 말하면 class내의 patern의 분포가 (2)식의 다차원정규분포로 가정할 수 있는 경우 (3)식을 기초로한 결정법은 평균 오인식률을 최소로 한다고 하는 bayes결정법칙으로 된다.

Ⅲ. K-L변환법에 의한 특징 파라미터의 차원압축

Cepstrum 계수에 의한 시계열을 하나의 vector로서 취급하여 인식 실험을 행할 경우에는 계산량의 방대하게 된다. 여기서는 이러한 계산량을 줄이기 위한 방법으로 K-L변환법(Karhuen-Loeve expension)[9]을 사용한다.

K-L변환법이란 여러개의 양적 변수들 사이의 관계를 분석하여 이 변수들의 선형결합으로 표시되는 주성분을 찾고 이중에서 중요한 몇개의 주성분으로 전체의 변동을 설명하고자 하는 다변량분석법이 K-L변환법으로 자료의 요약이나 선형관계식을 통하여 차수를 감소시켜 해석을 용이하게 하는데 목적이 있다. 즉, 다차원 공간에 대해서 관측 vector분포의 불균일성을 이용하고 통계적으로 최적인 차원으로 감소시키는 방법이다. 간단히 K-L 변환법에 대해서 정리하면 다음과 같다.

n차원의 관측 vector X의 공분산 행렬을 S라 하면

$$S = \frac{1}{N-1}\sum_{i=1}^{N}(Xi-\mathbf{X})\,(Xi-\mathbf{X})^t \quad (4)$$

여기서 Xi : i번째의 관측 vector

$\mathbf{X}$: 전체의 관측 vector의 평균 vector

N : 전체의 sample수

로 된다.

K-L 변환법에서 선형변환행렬 A는 다음의 특징평가함수 J(a)의 최대화 조건을 만족하는 vector a_j (j=1, 2, 3....m)로서 구성된다.

$$J(a) = \frac{a^t \cdot S \cdot a}{a^t \cdot a} \quad (5)$$

a는 다음의 고유치 문제를 풀어서 얻어진 고유 vector로 된다.

$$S \cdot a - \lambda \cdot a = 0 \quad (6)$$

고유치는 $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq \ldots \geq \lambda_n \geq 0$

에 대응하는 고유 vector $a_1, a_2, \ldots, a_m$의 vector에 의해 A를 구성한다.

$$A = [a_1, a_2, \ldots, a_m]$$

공분산 행렬 S는 대칭행렬이기 때문에

$$a_1{}^t \cdot a_j = 0,\ a_1{}^t \cdot S \cdot a_j = 0\ (i \neq j)$$

가 성립한다.

즉, 고유 vector간에는 직교하고 얻어진 특징 vector의 각 요소간에는 무상관으로 된다. 그러나 이 K-L변환법에 의해 얻어진 특징에 대해서는 본질적으로 저축에 고분산성의 음운성의 정보가 집약된다고 확증할 수는 없다.

Ⅳ. 인식실험

4.1 한국어 음소

한국어는 음소로서 단모음 10개, 반모음 2개, 중모음 12개, 그리고 자음 19개로 구성되며, 이들의 조합에 의해 실제 약 1000개의 단음절이 사용되고 있다[10]. 또, 한국어의 어휘는 명사의 약 90%, 용언 어간의 약 86%가 1음절및 2음절로 구성된다[11].

본 연구에서는 모음으로서는 기본 8모음(/ㅏ(a)/, /ㅓ(ə)/, /ㅗ(o)/, /ㅜ(u)/, /ㅡ(m)/, /ㅣ(i)/, /ㅐ(ɛ)/, /ㅔ(e)/)과 반모음 2개(/j/, /w/) 자음으로서는 19개(/ㅂ/(p)/, /ㅃ(p′)/, /ㅍ(ph)/, /ㄷ(t)/, /ㄸ(t′)/, /ㅌ(th)/, /ㄱ(k)/, ㄲ(k′), /ㅋ(kh)/, /ㅁ(m)/, /ㄴ(n)/, /ㅇ(ŋ)/, /ㄹ(r)/, /ㅈ(c)/, /ㅉ(c′)/, /ㅊ(ch)/, /ㅎ(h)/, /ㅅ(s)/, /ㅆ(s′)/)와 종성자음 6개(/ㅁ(m*)/, /ㄴ(n*)/, /ㄹ(l*)/, /(P*)/, /ㄷ(T*)/, /ㄱ(K*)/)을 포함하는 총 35개 음소를 대상으로 한다. 이때, 종성자음중(/ㅂ(p)/, /ㅃ(p′)/, /ㅍ(ph)/)은 (/ㅂ(p)/)로, (/ㅂ(p)/)로, (/ㄷ(t)/, /ㅌ(th)/)은 (/ㄷ(t)/로 (/ㄱ(k)/, /ㄲ(k′)/, /ㅋ(kh)/)은 (/ㄱ(k)/)로 나타낸다.

4.2 음성데이터

단음절의 자료는 한국어 역순 사전 중의 단어를 음절별의 출현 빈도순으로 나열하여 누적 빈도 90%이내에 들어 있는 단음절 501개와 음소의 수가 적은 경음 /ㅃ(p′)/, /ㄸ(t′)/, /ㄲ(k′)/, /ㅆ(c′)/, /ㅉ(s′)/에 대해서는 누적 빈도 99.9%까지에 들어가는 다음절로부터 뽑은 48개를 추가한 총 549개로 한다.

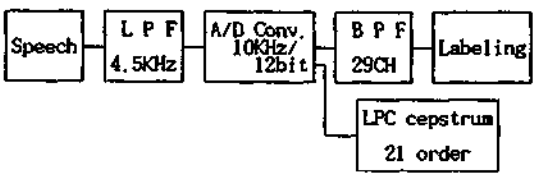


그림. 1. 분석의 흐름도
Fig 1. Block diagram of analysis.

음성 자료는 이들 단음절을 한국인 성인 남성 화자 3인이 방음실에서 랜덤하게 각 3회씩 자연스럽게 발성한 총 4941개로 한다.

음성 자료는 그림 1에 나타낸 것과 같이 4.5kHz Low-Pass Filter를 통과한 후 샘플링 주파수 10kHz, 양자화정도 12bit의 A/D converter를 통해 이산 데이타로 변환되고 29Channel Band-Pass Filter (Q=6 단공진, 1/6octave간격, 중심 주파수 250Hz-63Hz)를 통과시켜 분석된다(frame길이 : 10ms, 분석 window길이 : 20ms).

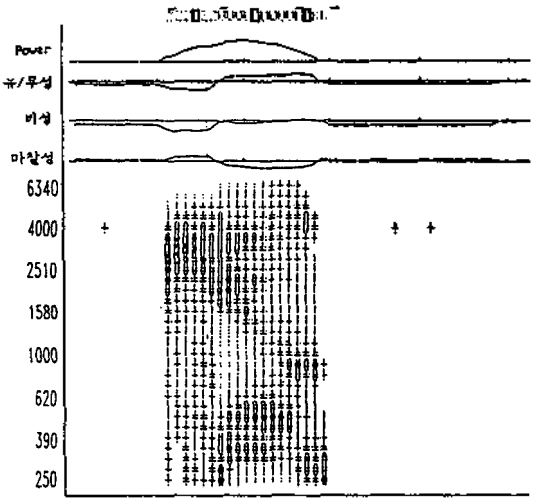
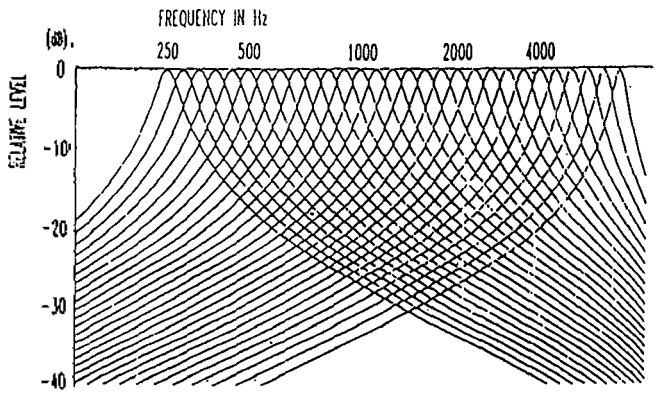


그림 3. Labeling의 일례(단음절 "죽").
Fig 3. An example of labeling (in the case of mono-syllable "zuk")



No. of channels : 29 channels
Center freq. : 250~6300 Hz
Interval : 1/6 oct.
Characteristics : single tuned (Q=6)
Analysis period : 10 ms/frame

CH	Hz	CH	Hz
1	250	16	1410
2	281	17	1580
3	316	18	1780
4	354	19	2000
5	400	20	2240
6	445	21	2500
7	500	22	2820
8	561	23	3160
9	630	24	3550
10	710	25	4000
11	793	26	4470
12	890	27	5000
13	1000	28	5600
14	1120	29	6300
15	1260		

그림 2. 29CH BPF의 주파수 특성
Fig 2. Frequency characteristics of 29 channel BPF.

표 1. 음소별 자료의 수
Table. 1 Number of Phoneme data

phoneme	number	phoneme	number
ㅏ (a)	990	ㅋ (kh)	135
ㅓ (ə)	963	ㅁ (m)	261
ㅗ (o)	633	ㅁ (m*)	414
ㅜ (u)	666	ㄴ (n)	198
ㅡ (ɯ)	573	ㄴ (n*)	513
ㅣ (i)	648	ㅇ (ŋ)	648
ㅔ (e)	180	ㄹ (l*)	675
ㅐ (ɛ)	306	ㄹ (r)	279
ㅑ (j)	270	ㅈ (c)	342
ㅕ (ʌ)	198	ㅉ (c')	162
ㅂ (p)	369	ㅊ (ch)	216
ㅃ (p')	135	ㅎ (h)	216
ㅍ (ph)	198	ㅅ (s)	342
ㄷ (t)	333	ㅆ (s')	162
ㄸ (t')	207	ㅂ (P*)	162
ㅌ (th)	153	ㄷ (T*)	414
ㄱ (k)	467	ㄱ (K*)	549
ㄲ (k')	225	Total	13202

(P, T*, K*, l*, m*, n* : 중성을 의미함)

그림2에 29CH BPF의 주파수 특성을 나타낸다.

각 음소에는 시찰에 의해 시단, 중심, 종단 frame이라는 시간적 label을 부여한다. Fig.3에 labeling의 예를 보인다. 이 label을 참고하여 각 frame별 21차원 LPC cepstrum 계수를 추출하여 음소별 데이터 베이스를 구성하며 이렇게하여 작성한 음소별 자료수는 13,202개이다.

4.3 인식결과 및 검토

인식 실험에 있어서는 3인 화자의 2회의 발성을 표준 pattern, 나머지 1회의 발성을 입력으로 3절에서 기술한 식별 방법으로 인식하여 3회 반복한 평균을 인식률로 하는 open 인식 실험을 실시한다. 특징추출의 위치를 조사하기 위해서는 각 음소군 별로 그림4와 같이 typical frame을 포함한 전, 후 1frame, 시단으로부터 3frame, 제 2frame으로부터 4frame과 같은 순서로 시단으로부터 11frame까지 시점을 1frame씩 옮겨 가면서 3frame씩의 시간방향 특징을 추출하여 인식 실험을 통하여 특징 추출의 위치를 확인한다. 이때 특징 벡터의 차원수는 63차원을 25차원으로 압축하였다.

그림5에 모음, 파열음, 마찰음및 파찰음, 비음및 유음의 특징추출 위치별 인식결과를 나타낸다.

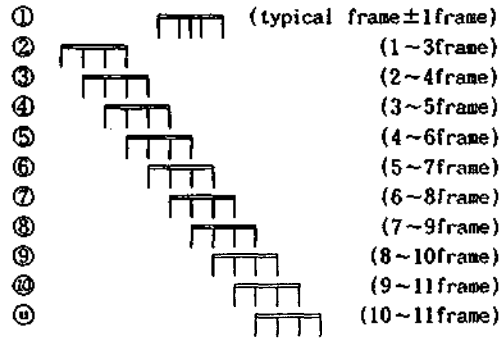


그림 4. 특징추출의 위치.
Fig 4. Place of feature extraction.

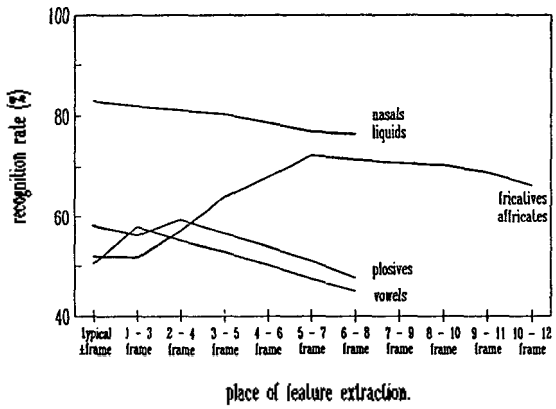


그림 5. 특징추출의 위치에 따른 인식률의 변화.
Fig 5. Variation of recognition rates according to the place of feature.

그림5로부터 특징추출의 위치는 모음의 경우 시단에서부터 3frame이 인식률이 가장 높게 나타나 시단에서부터 4~5frame이 중요함을 알 수 있다. 그리고, 파열음의 경우 2frame으로부터 4frame까지가 인식률이 높게 나타나 시단에서 5frame정도가 중요함을 알 수 있다.

마찰음과 파찰음의 경우는 5frame으로부터 7frame에서 최고의 인식률을 보이며, 특징추출의 위치는 4frame에서 10frame까지가 중요함을 알 수 있었다. 또한, 비음과 유음의 경우는 typical frame을 포함한 전, 후 1frame에서 최고의 인식률을 보이며, 특징추

출의 위치는 typical frame을 포함하며 비교적 인식률이 높게 나타나는 시단에서부터 5frame정도가 유효함을 알 수 있었다.

인식에 사용되는 지속시간 길이에 대해서는 위의 결과를 참고로 하여 그림6과 같이 지속시간 길이를 typical 1frame, typical을 포함한 전 1frame, typical을 포함한 후 1frame, typical frame을 포함한 전, 후 1frame(3frame), 시단에서부터 3frame, 시단에서부터 4frame, 5frame, 6frame, 7frame, 8frame의 10가지 경우로 frame수를 증가시키면서 검토하고, 계산량을 줄이기 위해 특징 벡터의 차원수를 K-L 변환을 이용하여 1frame(21차원)은 압축하지 않은 21차원 그대로, 2frame 경우는 42차원은 20차원으로, 3frame인 경우는 즉 63차원을 25차원으로, 4frame(84차원), 5frame(105차원)도 25차원으로, 6frame(126차원), 7frame(147차원), 8frame(168차원)은 30차원으로 각각 압축해서 인식 실험에 이용한다.

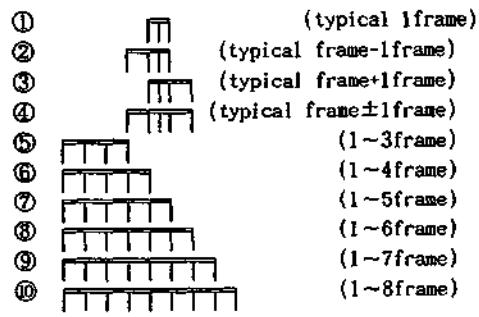


그림 6. 지속시간정보의 같이 추출.
Fig 6. Extraction of duration length.

표 2.1, 2.2에 각 음소별 인식 결과를 나타낸다. 또, 그림7에는 인식 조건별 평균을 나타낸다.

표 2.1, 2.2와 그림7로부터 특징 파라미터의 길이를 typical 1frame경우와 typical frame을 포함한 전, 후 1frame경우 즉, 2frame경우 각 음소의 특징을 충분하게 표현하지 못함을 알 수 있다. 그리고, typical frame을 포함한 전, 후 1frame보다 시단으로부터

표 2.1 지속시간길이의 변화에 대한 인식률.
Table 2.1 Recognition rates for feature length variation. (open test)

	①1frame (21 ord)	②typ-1fr (20 ord)	③typ+1fr (20 ord)	④typ±1fr (25 ord)	⑤1~3fr (25 ord)
ㅏ(a)	23.6	18.2	18.3	18.8	58.8
ㅓ(ə)	21.5	14.3	19.0	14.6	30.5
ㅗ(o)	39.0	38.1	28.4	36.7	57.9
ㅜ(u)	24.8	29.7	23.9	23.0	42.6
ㅡ(ɯ)	24.9	21.2	21.9	24.3	37.4
ㅣ(i)	38.0	31.0	29.9	36.6	55.4
ㅔ(e)	56.7	58.3	50.6	53.3	77.8
ㅐ(ɛ)	34.3	27.5	29.1	22.5	26.8
ㅚ(j)	63.3	77.8	67.4	76.7	66.7
ㅟ(ɯ)	24.2	39.4	50.5	53.0	33.8
ㅂ (p)	28.5	48.0	43.4	68.3	52.8
ㅃ (p')	37.8	55.6	55.6	64.4	65.2
ㅍ (ph)	33.3	45.5	33.8	48.5	61.1
ㄷ (t)	25.2	45.0	51.1	54.1	61.9
ㄸ (t')	47.8	65.2	65.2	79.7	77.3
ㅌ (th)	23.5	51.0	47.7	56.9	57.5
ㄱ (k)	20.5	25.0	42.3	46.2	53.6
ㄲ (k')	25.3	54.7	50.2	61.3	56.9
ㅋ (kh)	33.3	33.3	48.9	33.3	35.6
ㅂ (P*)	1.9	1.9	3.1	1.0	1.2
ㄷ (T*)	2.2	0.7	3.6	4.3	2.4
ㄱ (K*)	4.9	1.6	5.5	4.4	6.6
ㅁ (m)	71.3	78.2	70.9	80.5	76.7
ㅁ (m*)	73.2	84.1	75.1	84.1	77.3
ㄴ (n)	65.2	83.3	71.7	75.8	72.2
ㄴ (n*)	54.4	64.3	75.0	64.9	72.9
ㅇ (ŋ)	55.1	61.1	63.6	60.2	62.5
ㄹ (l*)	86.2	93.5	90.8	94.2	92.6
ㄹ (r)	59.1	66.7	58.8	67.7	66.6
ㅈ (c)	28.9	22.8	48.0	35.1	36.8
ㅉ (c')	38.9	53.7	57.4	72.2	30.9
ㅊ (ch)	50.5	56.9	54.2	63.9	36.1
ㅎ (h)	52.8	76.4	57.9	79.2	76.9
ㅅ (s)	34.2	43.0	49.4	50.9	63.5
ㅆ (s')	79.6	79.6	75.3	83.3	45.7
평 균	39.5%	47.0%	46.8%	51.2%	52.3%

표 2.2 지속시간길이의 변화에 대한 인식률.
Table 2.2 Recognition rates for feature length variation. (open test)

	⑥ 4frame (25 ord)	⑦ 5frame (25 ord)	⑧ 6frame (30 ord)	⑨ 7frame (30 ord)	⑩ 8frame (30 ord)
ㅏ(a)	50.0	44.5	74.5	41.8	42.7
ㅓ(ə)	25.4	25.9	41.4	25.2	24.6
ㅗ(o)	56.0	53.3	64.8	55.2	48.1
ㅜ(u)	41.9	41.4	63.5	40.1	38.3
ㅡ(ɯ)	31.6	31.2	40.7	31.7	28.6
ㅣ(i)	51.1	47.7	69.4	40.7	42.1
ㅔ(e)	78.9	83.3	78.3	81.7	76.6
ㅐ(ɛ)	25.5	24.5	32.4	21.6	27.5
ㅚ(j)	71.1	71.1	77.8	74.4	73.3
ㅟ(ɯ)	33.8	36.4	59.1	39.4	42.4
ㅂ (p)	54.7	53.7	54.5	63.4	63.4
ㅃ (p')	63.7	66.7	73.3	75.6	73.3
ㅍ (ph)	68.2	69.7	68.2	86.4	72.7
ㄷ (t)	66.4	60.4	49.5	65.8	65.8
ㄸ (t')	80.7	85.5	75.4	81.2	84.1
ㅌ (th)	62.7	70.6	60.8	76.5	70.6
ㄱ (k)	60.7	60.3	59.6	65.4	64.7
ㄲ (k')	65.3	80.0	70.7	88.0	88.0
ㅋ (kh)	40.7	53.3	60.0	64.4	71.1
ㅂ (P*)	2.5	3.7	14.8	5.6	3.7
ㄷ (T*)	3.1	3.6	12.3	6.5	8.0
ㄱ (K*)	9.7	10.4	23.0	8.2	9.8
ㅁ (m)	76.7	82.8	73.6	81.6	80.5
ㅁ (m*)	76.8	79.7	78.3	75.4	76.8
ㄴ (n)	74.2	77.3	71.2	75.8	74.2
ㄴ (n*)	74.7	59.1	63.7	61.4	66.7
ㅇ (ŋ)	64.4	61.6	65.7	61.1	62.0
ㄹ (l*)	93.2	95.6	94.7	96.9	94.7
ㄹ (r)	71.7	76.3	79.6	83.9	83.9
ㅈ (c)	40.1	38.6	34.2	43.0	50.5
ㅉ (c')	47.5	75.9	72.2	83.3	87.0
ㅊ (ch)	47.7	52.8	55.6	75.0	80.6
ㅎ (h)	80.1	86.1	91.7	80.6	86.1
ㅅ (s)	63.2	54.4	46.5	57.9	62.3
ㅆ (s')	43.2	64.8	68.5	64.8	68.5
평 균	54.2%	56.8%	60.6%	59.4%	59.8%

#(P*, T*, K*, m*, n*, l* : 종성을 의미함)

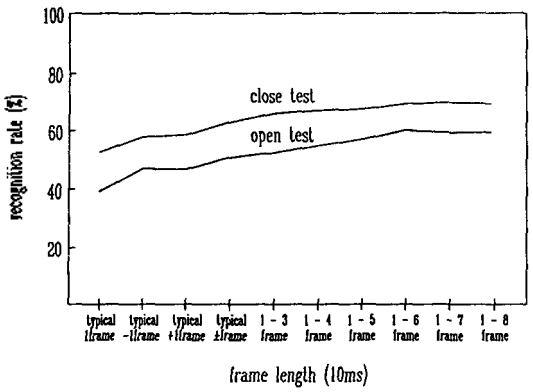


그림 7. 특징추출의 길이의 변화에 대한 인식률.
Fig 7. Recognition rates for variation of feature extraction length.

3frame인 경우가 인식율이 높게 나타나는 경향을 보여 시단에서 부터 frame수를 8frame까지 증가시키면서 인식 실험을 한 결과 open실험에서는 6frame일 때 close실험에서는 7frame일 경우가 최고의 인식률을 나타내고 8frame이상일 경우에는 더 이상 인식률의 상승을 볼 수 없었다. 따라서, 전 음소를 대상으로 볼때 지속시간의 정보가 시단에서부터 6frame~7frame 정도가 유효함을 알 수 있다. 이는 일본어의 경우보다 1~2frame정도 길게 나타났다.

V. 결 론

한국어 음성인식 시스템 개발을 위한 기초 연구로서 음소인식에 유효한 특징추출위치와 적당한 지속시간 정보의 길이를 확인하기 위해 음소군별 인식 실험과 한국어 전 음소를 대상으로 인식 실험을 행한 결과 다음과 같은 사실을 확인할 수 있었다.

1) 특징추출의 위치는 모음의 경우 시단에서부터 3frame이 인식률이 가장 높게 나타나 시단에서부터 4~5frame이 중요함을 알 수 있었고, 파열음의 경우는 2frame으로부터 4frame까지가 높게 나타나 시단에서 5frame정도가 유효함을 알 수 있었다. 또 마찰음과 파찰음의 경우는 5frame으로부터 10frame까지가 유효함을 알 수 있었다. 그리고, 비음과 유음의 경우는 typical frame을 포함한 전, 후 1frame에서 최고의 인식률을 보이며, 특징 추출의 위치는 typical frame을 포함하는 시단에서부터 5frame정도가 중요함을 알 수 있었다.

2) 지속 시간 정보의 길이는 typical 1frame으로 했을 경우와 typical frame을 포함한 전, 후 1frame, 즉 2frame인 경우 시간방향 정보의 부족으로 인하여 각 음소의 특징을 충분하게 표현하지 못함을 알 수 있었다.

3) typical frame을 포함한 전, 후 1frame보다 시단으로부터 3frame인 경우가 인식률이 높게 나타나는 경향을 보였다. 이를 참고로 하여 시단으로부터 지속시간 정보를 8frame까지 증가시키면서 인식 실험을 행한 결과 open실험에서는 시단으로부터 6frame까지 인식률의 증가를 보인다. 7frame이후는 인식률이 포화함을 보였으며, close실험에서는 7frame일 때 최고의 인식률을 보여 시간방향의 정보는 일본어의 경우보다 1frame~2frame 긴 7frame(70ms)정도가 적당함을 알 수 있었다.

4) 전 음소를 대상으로 볼 때 지속시간 정보의 길이를 6frame~7frame으로 했을때 최고의 인식률을 보여 이 정도의 지속시간 정보가 인식에 적당함을 알 수 있었다.

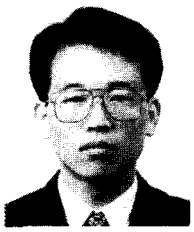
이상의 결과는 한국어 음성의 인식, 합성, 분석등의 시스템 설계에 있어 설계자가 기본적으로 고려해야할 사항인 특징 파라미터의 종류, 특징추출의 위치, 지속시간 정보의 길이등을 결정하는데 중요한 기초자료가 되리라 생각된다.

참 고 문 헌

1. 온종관 외, "음소모델링을 이용한 한국어 대용량단어 인식에 관한 연구," 음성통신및 신호처리 워크샵, 19-24(Aug. 1989)
2. 최영재 외, "한국통신의 자동통역 전화시스템 개발 구상," Korea-Japan Joint Workshop, 14-21 (July. 1991)
3. 이종락, "한국통신의 자동통역 연구현황" 음성통신및 신호처리 워크샵(Aug. 1992)
4. 김영일 외, "음소를 이용한 한국어 음성 신호의 분석과 인식," 한국 음향 학회, 6, 2(1987).
5. 鄭鉉烈, 牧野定三, 城戶健一, "韓國語 破裂子音の分析," 日本音響學會 1-3-3 (Oct. 1988)
6. 鄭鉉烈, 牧野定三, 城戶健一, "韓國語 語頭 破裂子音の認識," 日本音響學會 1-2-21 (Mar. 1989)
7. Okada et al, "スペクトルの時間變化 パターンによる語頭音素の識別," 日本電子 情報通信學會誌 J70A, 8, 1174-1185 (1987)
8. 中川聖一, "確率モデルによる音響認識," 電子情報通信學會(1988)

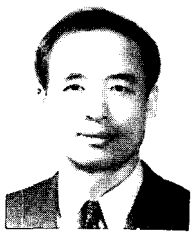
9. Toshiro Haga, Shigeji Hashimoto, 回歸分析と主成分析(1986)
10. 許熊, "國語 音韻學," 正音社, 207-226 (1985)
11. 梅田博之, "韓國語の音聲學的 研究," 螢雪出版社, 35-37 (1983)
12. 岡田美智男, "語頭子音の統計的 分析と認識に關する研究," 東北大學 博士學位論文, 15-17 (1984)
13. 鄭鉉烈, "韓國語音聲の分析と認識に關する基礎研究," 東北大學 博士學位論文 (Jan. 1989)
14. L.R.Rabiner and R.W.Schafer, Processing of Speech Signals, Prentice-Hall, (1978).

▲김 범 국 1964년 12월 26일생



1990년 2월 : 영남대학교 수학과 졸업(이학사)
1992년 2월 : 영남대학교 대학원 전자공학과 졸업(공학석사)
1992년 3월~현재 : 대구전문대학, 대구공업전문대학 강사
1994년 8월~현재 : 영남대학교 전자공학과 정보통신전공 박사과정 재학
※ 주관심분야 : 음성신호처리및 그응용등임

▲정 현 열 1951년 11월 26일



1975년 : 영남대학교 전자공학과 졸업(공학사)
1981년 : 영남대학교 대학원 전자공학과 졸업(공학석사)
1985년 5월~1986년 3월 : 일본 東北대학 응용정보학 연구센터 연구생
1989년 4월 : 일본 동북대학 대학원 정보공학과 졸업(공학박사)
1989년 3월~현재 : 영남대학교 전자공학과 부교수
1992년 7월~1993년 7월 : 미국 Carnegie Mellon Uni. Robotics연구소 Visiting Research Scholar
※ 주관심분야 : 음성신호처리및 그응용등임