한국어 음소 인식을 위한 기준 프레임 추출

김 범 국* 정 현 열*
*영남대학교 전자공학과

Typical Frame Extraction for Korean Phoneme Recognition

°Bum Koog Kim*, Hyun Yeol Chung*
*Dept. of Electronic Eng. Yeungnam University

요 약

음소를 인식의 기본으로하는 한국어 음성인식 시스템을 구현하기 위한 기초 연구의 일환으로서 각 음소의 특징을 가장 잘 표현하는 기준프레임 추출을 위한 연구를 수행하였다.

이를 위하여 먼저 선행 실험과 분산비 분석을 통해서 인식에 필요로한 시간 패턴의 길이를 추출한 후 이를 바탕으로 통계적 인식방법인 베이즈 결정법칙을 이용하여 시단 프레임으로부터 3프레임(1프레임 10ms)씩 시점을 1프레임씩 옮기면서 인식 실험을 행하여, 각 음소별 특징이 가장 풍부한 기준 프레임(3프레임)을 추출하였다. 그리고 이 기준 프레임을 중심으로 각 음소군별 인식 실험을 수행하여 그 결과를 시단을 기준으로한 경우와 비교 검토하고 한국어 전 음소별로 확장하여 인식 실험을 실시하였다.

이 실험 결과 모음의 경우 시단에서부터 5프레임, 파열음은 시단에서부터 5프레임사이, 마찰음은 3프레임에서부터 10프레임까지, 파찰음은 5프레임에서 9프레임까지, 비음과 유음의 경우 초성은 시단 프레임에서 6프레임, 종성은 종단으로부터 전 4프레임 구간이 인식률이 높게 나타나 이 부분의 특징이 인식에 가장 유효함을 알 수 있었다.

1. 서 론

음소를 인식의 기본단위로 하는 음성인식 시스템을 구현하기 위해서는 각 음소가 지속되는 시간 방향의 정보중 그 음소를 특징 지우는 변별적 특징이 가장 많이 포함된 위치(이하 기준프레임이라 함)가 어디인가를 파악할 수 있다면 음성의 합성, 인식시 매우 유용하게 이용될 수 있을 것이다.

본 연구실에서는 그 동안 한국어 음성인식 시스템을 구현하기 위한 그 기초적인 연구로서 한국어 전음소를 대상으로 한국어 모음의 분석과 인식[1], 파열자음의 분석, 인식[2,3], 마찰음및 파찰음의 분석과 인식[4], 비음과 유음의 인식을 위한 특징 추출에 관한 연구[4], 한국어 단음절에 포함된 음소 인식에 관한 연구[6], 다차원 척도법(*Multi-Dimension Scaling*)을 이용한 한국어 음소 분석[7]등의 연구를 통하여 음소를 단위로한 분석과 인식에 관한 연구를 수행 해 왔다.

이들 연구 결과를 참고로 본 논문에서는 각 음소의 특징을 가장 잘 표현하는 기준 프레임의 위치를 찾아 내고자 한다. 이를 위하여 먼저 선행 연구[1~7]를 참고로 인식에 필요로한 시간 패턴의 길이를 추출하고, 이를 바탕으로 시단 프레임으로부터 3프레임씩, 시점을 1프레임씩 옮겨 가면서 각 음소군별 실험을 통하여 전 음소의 기준 프레임(3프레임)을 추출한다. 이때 특징 파라메터로 21차원 캡스트럼계수를 사용한다.

이 결과를 이용하여 각 음소별 기준 프레임을 중심으로 특징 파라메터를 추출하여 인식 실험을 행한 결과와 시단에서부터 프레임을 증가 시키면서 인식 실험을 행한 결과를 비교 검토한다.

2. 음소의 분석

2.1 음성 자료

단음절 자료는 한국어 억순 사전중의 단어를 출현빈도순으로 나열하여 누적빈도 90% 이내에 들어가는 단음절 501개와 음소의 수가 적은 경음에 대해서는 누적빈도가 99.9% 까지에 들어가는 단음절로부터 뽑은 48개를 추가한 총 549개 이다.

이 단음절 자료를 방음실에서 한국인 성인남성 3인이 랜덤하게 각각 3회씩 자연스럽게 발성한 4941개로 한다.

2.2 분석 방법

단음절 음성자료는 그림 1에 보인 바와 같이 4.5KHz LPF을 통과한 후 10KHz 12bit A/D 변환기를 통해 음성 데이타로 변환되고, 29CH BPF(Q=6, 1/6 octave, 250Hz ~ 6300Hz)를 통과 시켜 분석 된다.(프레임 길이:10ms, 분석창 길이:20ms) 각 음소에는 시작에 의해 시단, 중심, 종단 프레임이라는 시간적 레이블을 부여한다.

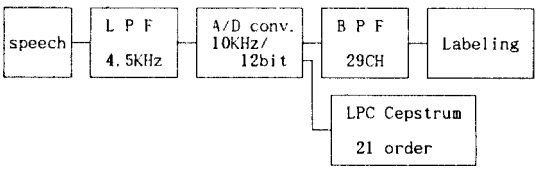

그림 1. 특징 파라메터 추출의 흐름도

이 레이블을 참고로하여 각 프레임으로부터 *power*성분을 포함하는 *21*차원 *LPC* 켑스트럼계수를 구하여 음소별 데이타 베이스를 구성한다. 각 음소별 자료수는 표 1과 같다.

표 1. 각 음소의 자료수.

(P*, T*, K*, l*, m*, n* : 종성)

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

2.2.1 복칭 분석

각 음소의 기준 프레임을 추출하기 위해 선행 연구들에서 얻은 결과와 분산 분석[4-5]을 통해 인식 실험에 필요로한 각 음소군별 지속시간 정보의 길이를 조사하고, 이 결과를 이용하여 인식 실험을 실시한다. 표 2에 단음절에 포함된 각 음소에 대한 지속시간의 평균과 표준편차를 보인다.

표 2. 단음절에 포함된 각 음소에 대한 지속시간의 평균과 표준편차. (지속시간 : ms)

음소환경		V	V+C	C+V	C+V+C	음소환경		V+C	C+V	C+V+C
음소	자료수	평균편차	평균편차	평균편차	평균편차	음소	자료수	평균편차	평균편차	평균편차
a	990	415 46	161 41	328 52	143 42	kh	135		121 15	111 21
ə	963	402 43	172 47	317 48	143 84	m	261		97 41	85 26
o	633	425 34	180 48	310 47	155 45	m*	414	401 69		178 59
u	666	427 47	173 62	312 48	107 53	n	198		97 25	84 26
ɯ	573	394 47	93 24	330 48	92 41	n*	513	290 61		259 64
i	648	408 38	167 36	325 54	115 48	ŋ	648	291 87		277 78
e	180	444 37		314 53	125 34	l*	675	304 59		284 53
ɛ	306	427 39		340 56	129 37	r	279		82 41	79 43
j	270	(b)125 30	(c)329 72			c	342		82 17	83 18
w	198	(b)113 31	(c)257 30			c'	162		62 15	58 12
p	369			47 18	45 17	ch	216		146 24	127 21
p'	135			31 5	31 5	h	216		122 34	109 30
ph	198			110 17	92 18	s	342		180 36	170 31
t	333			46 16	46 16	s'	162		174 42	157 53
t'	207			32 5	31 5	P	162	45 10		39 12
th	153			99 19	91 14	T	414	39 8		37 9
k	467			71 16	63 19	K	549	36 8		35 11
k'	225			40 9	38 8	total	13202			

단, (m*, n* : 어미), (P, T, K : 어미),
(반모음 j, w경우, (b)는 반모음이 어두, (c)는 어두 자음에 후속하는 경우)

2.2.2 분산 분석

단일 특징량인 경우에 있어서 *class*의 식별에 이용하는 특징량으로서는 class간의 분산이 크고 동시에 *class*내의 분산이 작은 것이 요구된다.

따라서, *class*간의 분산비를 구하므로서 식별에 유효한 켑스트럼계수와 지속시간 정보의 길이와 위치를 찾아 낼 수 있다.

이와 같은 특징의 평가량으로서 양분산의 비가 이용되며, 그 분산비의 정의식은

$$Y=\frac{\frac{1}{(g-1)}\sum_{k=1}^{g}(\mu_k-\mu)^2}{\frac{1}{(g(n-1))}\sum_{k=1}^{g}\sum_{i=1}^{n}(X_{ik}-\mu_k)^2} \qquad (1)$$

여기서, X_{ik} : *class* k 특징량의 i번째의 관측치

$(k=1,2,\ldots,g,\quad i=1,2,\ldots,n)$

μ_K : *class* k 특징량의 평균치

μ : 전체 특징량의 평균치

와 같이 나타난다.

각 음소의 시단에서부터 *150ms*에서 *200ms* 구간 (*15*프레임에서 *20*프레임)의 시간 변화에 대하여 각 음소군별로 나누어 *(1)*식에 의해 구해진 분산 분산 분석을 통해 파열음의 경우는 시단에서 *10*프레임 정도가, 마찰음과 파찰음은 시단에서 *14*프레임 정도가 식별에 유효함을 알 수 있었다. 그리고, 비음과 유음은 시단에서 *11*프레임 정도가 식별에 유효함을 알 수 있었다.

그림 2에 분산비 분석의 일예를 보인다.

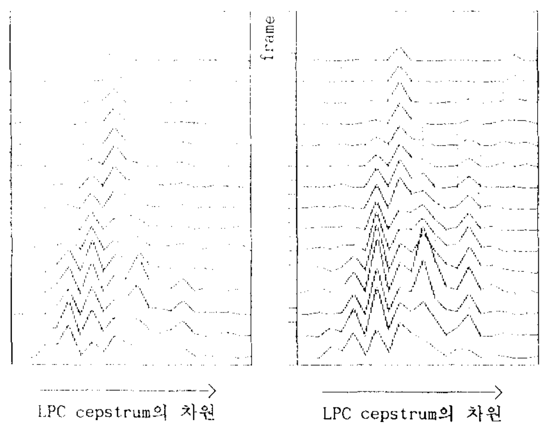

그림 2A) /ㅁ,ㄴ,ㄹ/의 분산비. 그림 2B) /ㅁ,ㄹ/의 분산비.

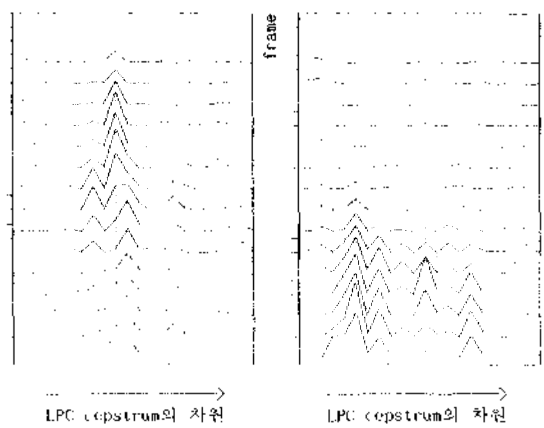

그림 2C) /ㄴ,ㄹ/의 분산비. 그림 2D) /ㅁ,ㄴ/의 분산비.

그림 2. 분산비 분석의 일예(비음과 유음의 경우).

3. 실험및 고찰

3.1 기준 프레임 추출

인식 실험에 있어서는 *3*인 화자와 *2*회의 발성을 표준 패턴으로하고 나머지 *1*회의 발성을 입력으로 통계적 인식 방법인 베이즈 결정 법칙(Bayesian decision rule)[8]을 이용하여 인식 실험을 행하고 *3*회 반복한 평균을 인식률로 한다.

선행 실험과 분산 분석의 결과를 참고로하여 인식에 유효한 특징추출의 위치를 조사하기 위해서는 음소군별로 각 음소의 시단으로부터 *3*프레임, 제 *2*프레임으로부터 *4*프레임과 같은 순서로 시단으로부터 *15*프레임까지 시점을 *1*프레임씩 옮겨 가면서 *3*프레임씩에 대한 시간방향 특징에 대해 인식 실험을 실시하였다.

그림 3에 *30ms*에 대한 특징 추출 위치별 인식률의 변화를 나타내었으며, 표 3에 각 음소별 기준 프레임(*3*프레임)을 추출한 결과를 나타낸다.

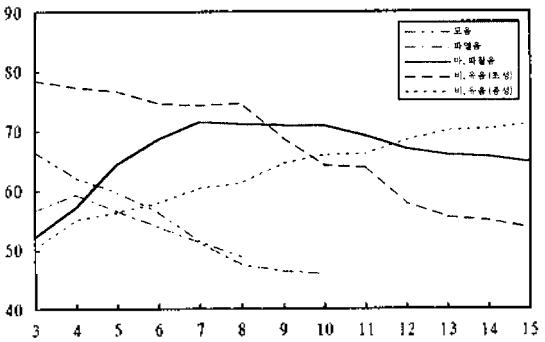

그림 3. *30ms*대한 특징 추출 위치별 인식률의 변화.

표3. 각 음소별 기준 프레임.

음 소	기준프레임 (3프레임)	음 소	기준프레임 (3프레임)
ㅏ (a)	1~3	ㅂ (P)	3~5
ㅓ (ə)	1~3	ㄷ (T)	1~3
ㅗ (o)	2~4	ㄱ (K)	3~5
ㅜ (u)	3~5	ㅁ (m)	4~6
ㅡ (ɯ)	2~4	ㄴ (n)	3~5
ㅣ (i)	1~3	ㄹ (r)	1~3
ㅔ (e)	1~3	ㅁ (m*)	13~15
ㅐ (ɛ)	2~4	ㄴ (n*)	13~15
ㅂ (p)	2~4	ㄹ (l*)	13~15
ㅃ (p')	1~3	ㅇ (ŋ)	12~14
ㅍ (ph)	2~4	ㅈ (c)	5~7
ㄷ (t)	2~4	ㅉ (c')	7~9
ㄸ (t')	1~3	ㅊ (ch)	7~9
ㅌ (th)	2~4	ㅅ (s)	6~8
ㄱ (k)	3~5	ㅆ (s')	8~10
ㄲ (k')	2~4	ㅎ (h)	3~5
ㅋ (kh)	4~6		

그림 3과 표 3으로부터 모음의 경우는 시단 프레임에서 *5* 프레임사이, 파열음은 시단에서 *5*프레임사이, 마찰음과 파찰음은 *5*프레임으로부터 *10*프레임사이, 그리고 비음과 유음의 경우 초성은 시단에서 *6*프레임, 종성은 종단으로부터 전 *4*프레임 구간이 인식에 유효한 정보가 가장 많이 포함되어 있음을 알 수 있었다.

또한, 위의 결과로부터 인식에 적당한 시간방향의 길이를 알아보기 위해서 비교적 인식에 공헌도가 높은 구간, 즉, 모음의 경우는 시단에서 *8*프레임, 파열음의 경우는 시단에서 *10*프레임, 마찰음과 파찰음의 경우는 시단에서 *15*프레임, 비음과 유음의 경우는 시단에서 *15*프레임 사이의 구간에 대해서 인식 실험을 실시하였다.

이때 기준 프레임을 포함한 구간 추출은 다음과 같이 한다.

1)모음은 기준 프레임을 포함한 시단에서부터 추출하여 인식 실험을 하였다.

2)파열음은 기준 프레임 전 *1*프레임으로부터 프레임을 증가 시키면서 추출한다.

3)마찰음,파찰음의 경우는 기준 프레임 전 *1*프레임으로부터 프레임 증가 시키면서 추출한다.

4)비음과 유음의 경우는 초성은 기준 프레임을 포함한 시단부터 프레임수를 증가시키면서 추출하고, 종성은 종단에서부터 전 방향으로 프레임을 증가 시키면서 추출하였다

이상의 결과와 시단에서 3프레임을 기준으로 2프레임씩 증가 시키면서 인식 실험을 행한 결과를 비교 검토하였다.

이 결과를 각각 그림 4, 5, 6, 7에 나타내었다.

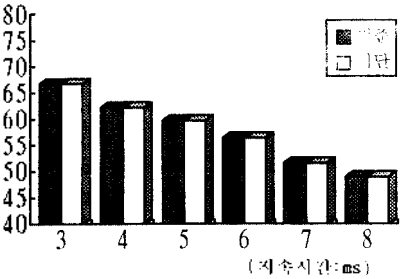

그림 4. 모음의 평균 인식률.

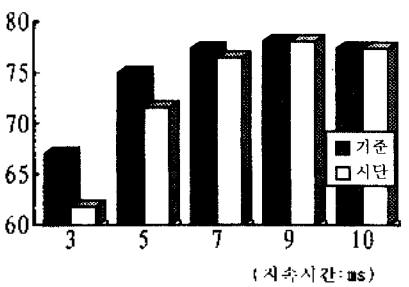

그림 5. 파열음의 평균 인식률.

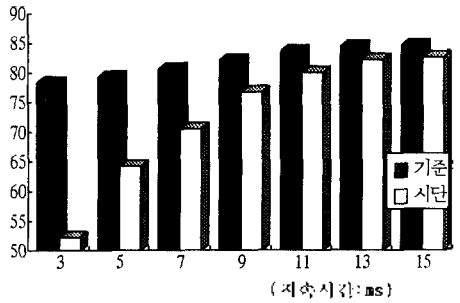

그림 6. 마찰음과 파찰음의 평균 인식률.

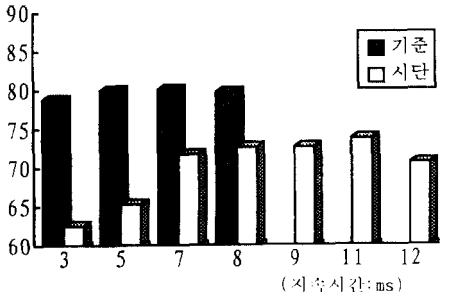

그림 7. 비음과 유음의 평균 인식률.

위의 그림으로부터 다음과 같은 사실을 알 수 있었다.

1)모음의 경우 시단에서 *5*프레임, 2)파열음은 기준 프레임 전 *1*프레임으로부터 *9*프레임, 3)마찰음과 파찰음의 경우, 기음/ㅎ/은 시단 프레임에서 *14* 프레임, 그외 음소는 기준 프레임 전 *1*프레임에서부터 *14*프레임, 4)비음과 유음의 경우는 초성은 시단 프레임에서 *7*프레임, 종성은 종단 프레임을 기준으로 전방향으로 *7*프레임에서 인식률이 높게나타나 시단에서 프레임수를 증가 시키면서 인식 실험을 행하는 경우보다 효과적임을 알 수 있었다.

또한, 한국어 전 음소를 대상으로 기준 프레임을 중심으로 프레임수를 증가 시키면서 인식 실험을 행한 결과 인식률의 향상을 기대할 수 있었다.

4. 결 론

단음절에 포함된 전 음소를 대상으로 분석과 인식 실험을 통하여 각 음소별 특징을 가장 잘 포함하는 기준 프레임(*3*프레임)을 추출하였고, 음소군별로 보면 모음의 경우는 시단 프레임에서 *5*프레임사이, 파열음은 시단에서 *5*프레임사이, 마찰음과 파찰음은 *5*프레임으로부터 *10*프레임사이, 그리고 비음과 유음의 경우 초성은 시단에서 *6*프레임, 종성은 종단으로부터 전 *4*프레임 구간으로 나타났다.

이를 이용하여 음소군별로 기준 프레임을 포함한 인식 실험을 실시하여 다음과 같은 사실을 확인하였다.

1)모음의 경우 시단에서 *5*프레임, 2)파열음은 기준 프레임 전 *1*프레임으로부터 *9*프레임, 3)마찰음과 파찰음의 경우, 기음/ㅎ/은 시단 프레임에서 *14* 프레임, 그외 음소는 기준 프레임 전 *1*프레임에서부터 *14*프레임, 4)비음과 유음의 경우는 초성은 시단 프레임에서 *7*프레임, 종성은 종단 프레임을 기준으로 전방향으로 *7*프레임에서 인식률이 높게나타나 이 부분의 정보가 인식에 중요함을 알 수 있었다.

참고문헌

(1) H.Y.Chung, S.Makino and K.Kido, "Analysis and recognition of Korea isolated vowels using foramnt frequency", J.Acoust.Jpn,9,5,pp.225-232(1988).

(2) 鄭鉉烈外, "韓國語破裂子音の分析", 日本音響學會, 1-3-3 (Oct.1988)

(3) 鄭鉉烈外, "韓國語語頭破裂子音の認識", 日本音響學會 1-2-21 (Mar.1989)

(4) 정석재, "마찰음및 파찰음의 분석과 인식", 영남대학 석사학위논문,(1991)

(5) 김범국, 정현일, "비음과 유음의 인식을 위한 특징추출에 관한 연구", 한국음향학회 영남지부 학술발표회 논문집,(1994)

(6) 김범국, "한국어 단음절에 포함된 음소 인식에 관한 연구", 영남대학 석사학위논문(1992)

(7) 권영욱, "MDS법을 이용한 한국어 음소 분석", 영남대학 석사학위논문(1991)

(8) 岡田美智男, 牧野正三, 城戸健一, "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)

(9) 中川聖一, "確率モデルによる音聲認識", 電子情報通信學會 (1988)

(11)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)

(12)Toshiro Haga, Shigeji Hashimoto, "回歸分析と主成分分析"(1986)