한국어 음성인식을 위한 음성학 기반의 유사음소단위 집합 설계*

홍혜진(서울대), 김선희(서울대), 정민화(서울대)

<차 례>

1. 서론
2. 기존 유사음소단위 집합 분석
   2.1. 국외 기관의 유사음소단위 집합
   2.2. 국내 기관의 유사음소단위 집합
3. 유사음소단위 집합 설계
4. 실험 및 결과
   4.1. 폰단위 인식
   4.2. 고립단어인식
   4.3. 연속음성인식
5. 결론

<Abstract>

A Phonetics Based Design of PLU Sets for Korean Speech Recognition

Hyejin Hong, Sunhee Kim, Minhwa Chung

This paper presents the effects of different phone-like-unit (PLU) sets in order to propose an optimal PLU set for the performance improvement of Korean automatic speech recognition (ASR) systems. The examination of 9 currently used PLU sets indicates that most of them include a selection of allophones without any sufficient phonetic base. In this paper, a total of 34 PLU sets are designed based on Korean phonetic characteristics and the effects of each PLU set are evaluated through experiments. The results show that the accuracy rate of each phone is influenced by different phonetic constraint(s) which determine(s) the PLU sets, and that an optimal PLU set can be anticipated through the phonetic analysis of the given speech data.

* Keywords: Phone-like-unit, Subword unit, Korean speech recognition.

* 이 논문은 KT와 정보통신부 및 정보통신연구진흥원의 IT 성장동력기술개발사업의 일환으로 수행된 연구입니다(2006-S-036-01, 신성장동력산업용 대용량/대화형 분산/내장처리 음성인터페이스 기술 개발).

1. 서 론

일반적으로 대용량 연속음성인식 시스템에서는 음성인식을 위한 기본 단위로 하위단어단위(subword unit)를 정의하여 사용한다. 이는 연속음성의 경우, 단어 경계를 찾아내기가 쉽지 않고, 단어 사이의 음운 현상을 고려하여야 하며, 또한 인식 대상이 되는 단어의 수가 증가함에 따라 시스템의 계산량이 급격하게 증가하기 때문이다. 대부분의 대용량 연속음성인식 시스템에서는 음성인식의 기본 단위로 음소를 기반으로 확장하여 정의된 유사음소단위(phone-like-unit: PLU)를 주로 사용한다. 유사음소단위는 음향 모델과 어휘 모델의 기본 단위로서, 그것이 어떻게 정의되느냐에 따라 음성인식 시스템의 성능이 결정적인 영향을 받게 된다. [1][2]에서 음절(syllable), 반음절(demi-syllable) 등을 사용하는 방법이 제안되기도 하였으나, 음절, 반음절 등의 하위단어단위는 유사음소단위에 비해 학습이 어렵다는 단점이 있다.

유사음소단위는 언어에 따라 다른 특성을 보이므로 일반적으로 음성학이나 언어학을 전공한 사람들에 의하여 정의되어 사용되어 왔으며, 그 설정 방법론에 관한 연구는 국내외에서 많은 연구가 진행되지 않은 분야 가운데 하나이다. 유사음소단위 집합은 국내외 모두 연구 기관에 따라 자체적으로 정의하여 다르게 사용하고 있으며, 이렇게 정의된 유사음소단위 집합은 거의 수정되지 않고 모든 영역에 동일하게 사용되었다. 그러나 기존 정의된 유사음소단위 집합이 다양한 음향적 환경과 새로운 영역에 적절하지 않을 수 있다는 점을 고려하여 [3]-[5]에서 유사음소단위를 학습 데이터로부터 자동으로 추출하는 방법을 제안하였으나, 실험 결과 수동으로 정의한 유사음소단위 집합을 이용한 경우의 성능에 미치지 못하거나 거의 유사한 결과를 보였다.

국내의 경우에 유사음소단위 집합에 관한 관련 연구로는 [6]-[11] 등을 들 수 있다. [6]은 HM-Net 문맥의존 음향모델링에 적합한 유사음소단위에 관한 연구로 음소를 기반으로 39개의 유사음소단위를 정의하여 그 유효성을 확인하였다. [8]에서는 음성학적 자질을 고려한 음소 및 변이음 모델을 인식에 이용하는 방법을 제안하였는데, 이는 기본 음소를 바탕으로 그 변이음을 트라이폰(triphone)으로 모델링하여 인식 성능이 향상됨을 보였다. [9]는 한국어의 변이음 규칙 가운데 일부를 이용하여 발음 사전을 생성하였을 때 인식 성능이 향상됨을 보고하였다. [10]은 음성 코퍼스 구축을 위한 분절음 연구로 음성학 기반의 분절음 레이블링을 제안하였고, [11]은 한국어 분절음의 음성·음운론적 특성에 관한 연구로 음성인식 및 합성에 언어학적 지식을 활용하는 방안을 제시하였는데, 이 두 연구를 토대로 음성합성이나 음성인식의 기본 단위를 설정한 경우는 보고되지 않았다. 즉, 기존의 연구들은 음소를 기반으로 하여 문맥 의존적 변이음을 트라이폰을 이용하여 모델링하거나[6][8], 일부 변이음 규칙을 이용하여 어휘부를 보완하거나[9], 혹은 그 유효

성 검증 없이 음성학 및 음운론적 지식을 변이음 생성에 이용할 것에 대한 제안으로, 이들 모두를 엄밀한 의미에서 음성학적 지식을 체계적으로 이용하여 유사음소 단위를 정의하고자 한 시도로 평가하기에는 미흡하다고 할 수 있다.

본 논문은 한국어 음성인식 시스템의 성능 향상을 위한 기초 연구로서, 우선 기존에 국내외 각 기관에서 정의하여 사용하고 있는 유사음소단위 집합을 검토한 후, 이를 기반으로 한국어의 음성 현상 제약[1]에 따라 가능한 유사음소단위 집합을 설계하고 인식 실험을 통하여 이를 평가하는 것을 그 목적으로 한다. 이러한 연구는 음성학을 바탕으로 한국어의 음성적 특성을 유사음소단위 집합에 체계적으로 반영하여 각각의 집합들에 대하여 그 유용성을 실험을 통하여 밝힘으로써, 궁극적으로 대용량 음성인식 시스템의 음향 모델 기본 단위로 적합한 유사음소단위 집합을 제안하여 음성인식의 성능 향상에 기여할 것으로 기대한다.

2. 기존 유사음소단위 집합 분석

2.1. 국외 기관의 유사음소단위 집합

국외 기관의 기존 유사음소단위 집합 가운데 대표적인 것으로는 영어의 경우 TIMIT 61, TIMIT 39, CMU 39, WSJ 46 등이 있다. <표 1>에서 보여 준 TIMIT 61은 광역 음성 전사(transcription at a broad phonetic level)를 기반으로 한 것으로 파열음의 경우에 묵음 구간과 파열 구간을 구분하였으며, 또한 [ɦ], [ü], [ɾ], [ʔ] 등의 변이음이 포함된다[12]. <표 2>에 보이는 바와 같이 TIMIT 39과 CMU 39는 모두 음소 기반 집합이다[13]. WSJ 46의 경우는 음소 기반 집합을 확장한 형태로서, 21개 모음(3개의 이중모음 및 [ə] 포함), 24개 자음(파열음 6개, 마찰음 8개, 파찰음 2개, 비음 3개, 반모음 5개), 그리고 묵음(silence)으로 구성된다. 프랑스어의 경우는 LIMSI에서 사용하는 유사음소단위 집합이 있으며, 이는 14개의 모음(3개의 비모음 포함)과 20개의 자음(파열음 6개, 마찰음 6개, 비음 3개, 반모음 5개), 그리고 묵음(silence)으로 구성된다[14].

2.2. 국내 기관의 유사음소단위 집합

본 논문에서는 산업계, 학계 등의 국내의 각 기관에서 자체적으로 정의하여 사용하고 있는 유사음소단위 집합 총 9개를 검토하였다. [15]에서 제시된 4개 기관의

---

1) 본 논문에서는 음소간의 교체가 아닌 하나의 음소가 환경에 따라 다른 변이음으로 실현되는 현상을 가리키기 위해 '음성 현상'이란 용어를 사용하였다. '음성 현상 제약'은 '변이음 규칙' 등의 용어로도 사용되고 있다[9].

<표 1> TIMIT 61

IPA	유사음소	예	IPA	유사음소	예
[ɑ]	aa	bob	[ɬ]	ix	debit
[æ]	ae	bat	[i]	iy	beet
[ʌ]	ah	but	[ǰ]	jh	joke
[ɔ]	ao	bought	[k]	k	key
[ɑʷ]	aw	bout	[k$^{\square}$]	kcl	k closure
[ə]	ax	about	[l]	l	lay
[əʰ]	ax-h	potato	[m]	m	mom
[ɚ]	axr	butter	[n]	n	noon
[ɑʸ]	ay	bite	[ŋ]	ng	sing
[b]	b	bee	[ɾ̃]	nx	winner
[b$^{\square}$]	bcl	b closure	[o]	ow	boat
[č]	ch	choke	[ɔʸ]	oy	boy
[d]	d	day	[p]	p	pea
[d$^{\square}$]	dcl	d closure		pau	pause
[ð]	dh	then	[p$^{\square}$]	pcl	p closure
[ɾ]	dx	muddy	[ʔ]	q	glottal storp
[ɛ]	eh	bet	[r]	r	ray
[l̩]	el	bottle	[s]	s	sea
[m̩]	em	bottom	[š]	sh	she
[n̩]	en	button	[t]	t	tea
[ŋ̍]	eng	Washington	[t$^{\square}$]	tcl	t closure
	epi	epenthetic silence	[Θ]	th	thin
[ɝ]	er	bird	[ʊ]	uh	book
[e]	ey	bait	[u]	uw	boot
[f]	f	fin	[ü]	ux	toot
[g]	g	gay	[v]	v	van
[g$^{\square}$]	gcl	g closure	[w]	w	way
[h]	hh	hay	[y]	y	yacht
[ɦ]	hv	ahead	[z]	z	zone
[I]	ih	bit	[ž]	zh	azure
-	h#	utterance initial and final silence			

집합과 [6]에서 제시된 1개 집합, 저자들의 소속기관에서 자체적으로 사용하고 있는 3개의 집합을 그 분석 대상으로 하였다. 분석 대상이 된 기관과 각 기관에서 사용된 집합 수는 <표 3>과 같다.

<표 2> TIMIT/CMU 39

IPA	유사음소	예	IPA	유사음소	예
[ɑ]	aa	bob	[l]	l	lay
[æ]	ae	bat	[m]	m	mom
[ʌ]	ah	but	[n]	n	noon
[$ɑ^y$]	ay	bite	[ŋ]	ng	sing
[$ɑ^w$]	aw	bout	[o]	ow	boat
[b]	b	bee	[p]	p	pea
[č]	ch	choke	[r]	r	ray
[d]	d	day	[s]	s	sea
[ð]	dh	then	[š]	sh	she
[ɾ]	dx	muddy	[t]	t	tea
[ɛ]	eh	bet	[Θ]	th	thin
[ɚ]	er	butter	[ʊ]	uh	book
[e]	ey	bait	[u]	uw	boot
[f]	f	fin	[v]	v	van
[g]	g	gay	[w]	w	way
[h]	hh	hay	[y]	y	yacht
[I]	ih	bit	[z]	z	zone
[i]	iy	beet	[$ɔ^y$]	oy	boy
[j]	jh	joke		bcl	
[k]	k	key			

<표 3> 분석 대상이 된 국내 기관의 유사음소단위 집합

기관	집합수
ETRI[15]	2
삼성[15]	1
KT[2)]	1
SiTEC[15]	1
영남대[6]	1
서울대	3
계	9

2) [15]에는 ETRI, 삼성, KT, SiTEC의 4개 기관의 집합 5개가 제시되었는데, KT의 경우에는 [15]에 제시된 집합 대신 해당 기관의 협조로 그 기관에서 현재 실제로 사용하고 있는 집합을 분석하였다.

<표 4> 자음 유사음소단위 분석 결과

			음소	유사음소단위	집합수
파열음	양순음	평음	/ㅂ/	{p}, {b}, {p˜}	2
				{p}, {p˜}	4
				{p}	3
		경음	/ㅃ/	{p=}	모두 포함
		격음	/ㅍ/	{pʰ}	모두 포함
	치조음	평음	/ㄷ/	{t}, {d}, {t˜}	2
				{t}, {t˜}	3
				{t}	4
		경음	/ㄸ/	{t=}	모두 포함
		격음	/ㅌ/	{tʰ}	모두 포함
	연구개음	평음	/ㄱ/	{k}, {g}, {k˜}	2
				{k}, {k˜}	4
				{k}	3
		경음	/ㄲ/	{k=}	모두 포함
		격음	/ㅋ/	{kʰ}	모두 포함
파찰음	구개치조음	평음	/ㅈ/	{ʧ}, {ʤ}	1
				{ʧ}	8
		경음	/ㅉ/	{ʧ=}	모두 포함
		격음	/ㅊ/	{ʧʰ}	모두 포함
마찰음	치조음	평음	/ㅅ/	{s}, {ɕ}	1
				{s}	8
		경음	/ㅆ/	{s=}, {ɕ=}	1
				{s=}	7
	성문음	평음	/ㅎ/	{h}, {ɦ}, {ç}, {x}	1
				{h}, {ɦ}	1
				{h}	7
유음			/ㄹ/	{l}, {ɾ}	5
				{l}, {ɻ}, {ɾ}	1
				{l}	3
비음	양순음		/ㅁ/	{m}, {m˜}	2
				{m}	7
	치조음		/ㄴ/	{n}, {n˜}, {ɲ}	1
				{n}, {n˜}	2
				{n}	6
	연구개음		/ㅇ/	{ŋ}	모두 포함

<표 4>는 한국어의 자음 음소와 이를 바탕으로 기존 유사음소단위 집합을 분석한 결과를 나타낸 것이다.[3] 9개의 유사음소단위 집합은 크게 음소를 기반으로 정의된 집합을 사용하는 경우(ETRI(2), 영남대)와 여러 가지 음성 현상을 고려하여 변이음을 추가한 경우로 분류될 수 있다. 변이음이 추가된 경우는 설측음화, 장애

3) 음소는 / /로 표기하였으며, 음소와의 구별을 위해 유사음소단위는 { }로 표기하였다.

음의 불파음화, 비음의 불파음화, 유성음화, 구개음화 등과 같은 음성 현상을 고려하여 이를 반영하는 변이음을 별개의 유사음소단위로 설정한 것을 알 수 있다.

고려된 음성 현상에 따라 자음의 유사음소단위 집합을 분류하면 <표 5>와 같다. <표 5>를 보면 이미 언급한 바와 같이 2개 기관에서 음소를 기반으로 유사음소단위 집합을 설정하였고, 한 기관의 경우는 음소와 이 5가지 음성 현상을 모두 고려하여 유사음소단위 집합을 정의하였으며, 나머지의 경우는 몇 가지의 음성 현상이 특별한 근거 없이 선택적으로 고려되었음을 볼 수 있다. 또한, 개별적인 음성 현상이 고려되는 경우에 있어서도 해당 현상이 체계적으로 고려되지 않음을 발견하였다. 예를 들면, ETRI(1)의 경우에 불파음을 별개의 유사음소로 설정하는데, [t˺]를 제외한 [p˺], [k˺]만을 별개의 유사음소단위로 정의하였다. 그러나 /ㅂ/, /ㄷ/, /ㄱ/ 등의 장애음은 음절말에서 각각 [p˺], [t˺], [k˺]와 같이 무성 장애음으로 실현되는 것이 일반적인 데 반하여 이와 같이 [t˺]를 제외한 것에 대해서는 그 이유를 짐작하기가 쉽지 않다.

모음의 경우에는 총 9개 집합을 대상으로 모음 유사음소단위를 분석한 결과, 두 개의 집합을 제외하고는 이중모음을 반모음 유사음소단위와 단모음 유사음소단위의 결합으로 정의하지 않고, 독립된 하나의 유사음소단위로 정의하여 사용하고 있었다. <표 6>에서 볼 수 있듯이 대부분 이중 모음은 하나의 단위로 정의하였고, 모음에 있어서는 표기는 다르나 실제로 동일하게 발음되는 것으로 여겨지는 모음들은 통합하여 사용하는 데 중점을 둔 것을 알 수 있다.

이와 같이 기존 유사음소단위 집합을 자음과 모음으로 각각 나누어 분석한 결과, 상당수의 유사음소단위 집합이 음성학적인 관점에서 볼 때 체계적 분류 없이 자의적으로 정의되어 사용되고 있음을 알 수 있었다. 이와 같이 자의적으로 유사

<표 5> 자음의 유사음소단위 집합 분석

	음소	설측음화	장애음의 불파음화	비음의 불파음화	구개음화	유성음화	비고
ETRI(1)							장애음의 불파음화:/ㄱ, ㅂ/
ETRI(2)							
삼성							
KT							{ç}, {x} 포함
SiTEC							
영남대							
서울대(1)							구개음화:/ㅅ, ㄹ/
서울대(2)							
서울대(3)							

<표 6> 모음의 유사음소단위 집합 분석

	단모음	이중모음	이중모음 (반모음+단모음)	비고
ETRI(1)				단모음 /ㅚ/, 이중모음 /ㅞ/ 통합
ETRI(2)				단모음 /ㅚ/, 이중모음 /ㅞ/ 통합
삼성				이중모음 /ㅙ/, /ㅞ/ 통합
KT				초성 자음에 따라 모음 구분4)
SiTEC				
영남대				이중모음 /ㅙ/, /ㅞ/ 통합
서울대(1)				
서울대(2)				
서울대(3)				

음소단위를 정의할 경우에는 개별적으로 고려한 변이음들이 인식 성능에 미치는 영향을 파악하기가 어렵게 된다. 따라서 본 논문에서는 개별 변이음이 인식 성능에 미치는 영향을 분석하기 위해 음성학적 지식을 기반으로 하여 한국어의 음성 현상을 고려하여 유사음소단위를 정의하고, 기존 유사음소단위 집합에서 정의된 유사음소단위가 음성인식에서 유용한지에 대해 체계적으로 검토하고자 한다.

3. 유사음소단위 집합 설계

한국어 음소와 그 변이음에 관한 기존의 연구들에 의하여 대표적인 변이음 규칙으로는 모음의 무성화, 순음화, 경구개음화, 경구개음화 및 원순음화, 연구개음화, 구개수음화, 평음의 유성음화, 마찰음화 및 유성음화, 유음변이, 불파음화(비파화), 무성화, 장음화 등이 있다[11](cf. [10][16]-[19]). 본 논문에서는 기존의 기관들이 유사음소단위 집합을 설계할 때 고려한 음성 현상 제약을 중심으로 유사음소단위 집합을 체계적으로 설계하여 각각의 제약이 음성인식의 성능에 미치는 영향을 고찰하고자 한다.

기존의 기관들에서 고려한 음성제약 현상들과 그 내용은 다음과 같다.

(1) 설측음화: 음절말 /ㄹ/은 설측음으로 조음된다.
(2) 장애음의 불파음화: 음절말 장애음은 불파된다.

---

4) KT의 내부지침에 의하면 초성 자음에 따라 단모음이 2가지의 변이음으로 실현된다고 한다. 초성으로 /ㅋ, ㅌ, ㅍ, ㅊ, ㅅ, ㅎ/이 오는 경우와 이외의 자음이 오는 경우에 대해 각각 별개의 유사음소단위를 설정하였다.

(3) 비음의 불파음화: 음절말 비음은 불파된다.
(4) 구개음화: [i], [j] 앞의 /ㄴ, ㅅ, ㄹ/은 구개음으로 조음된다.
(5) 유성음화: 유성음 사이의 장애음은 유성음으로 실현된다.
(6) 이중모음: 이중모음은 [j], [w]의 반모음과 단모음의 결합이다.

(1)~(5)는 자음에 관련된 제약이며, (6)은 이중모음을 반모음과 단모음의 결합으

<표 7> 설계한 유사음소단위 집합

	유사음소단위 개수	음소	설측음화	장애음의 불파음화	비음의 불파음화	구개음화	유성음화	이중모음 (반모음+단모음)
1	41							
2	42							
3	44							
4	43							
5	45							
6	44							
7	46							
8	47							
9	45							
10	46							
11	48							
12	47							
13	49							
14	48							
15	50							
16	51							
17	46							
18	47							
19	49							
20	48							
21	50							
22	49							
23	51							
24	52							
25	50							
26	51							
27	53							
28	52							
29	54							
30	53							
31	55							
32	56							
D	52							
SV	43							

로 정의하느냐에 관한 모음에 관련된 제약이다.

이와 같은 6가지 제약들을 결합하여 가능한 유사음소단위 집합을 <표 7>과 같이 설계하였다. 이때, 이중모음에 관련된 제약을 고려하여 모음 집합을 설계하는 경우 자음에 대해서는 구개음화를 제외한 모든 제약을 고려하였다[5].

4. 실험 및 결과

3장에서 설계한 총 34개의 유사음소단위 집합을 대상으로 폰단위 인식, 고립단어인식 및 연속음성인식 실험을 수행하였다[6].

4.1. 폰단위 인식

폰단위 인식 실험에는 phonetically balanced word (PWB) 고립단어 코퍼스가 사용되었다. 이 가운데 371,520 모노폰(monophone)을 학습에 사용하였으며, 43,344 모노폰을 인식 실험에 사용하였다. 이때, 이중모음을 반모음과 단모음의 결합으로 정의한 SV 집합의 경우에는 학습에는 416,520 모노폰이 사용되었으며, 48,594 모노폰을 대상으로 인식 실험을 수행하였다. 실험은 HTK toolkit 3.3에서 수행하였으며, 폰 바이그램(phone 2-gram) 가중치는 5.0이었다.

다음 <표 8>은 인식 실험에서 사용된 음성 데이터에서 각각의 음성 현상 제약에 의하여 영향을 받는 모노폰의 개수를 나타낸다.

<표 8> 음성 현상 제약에 따라 영향을 받는 모노폰의 개수

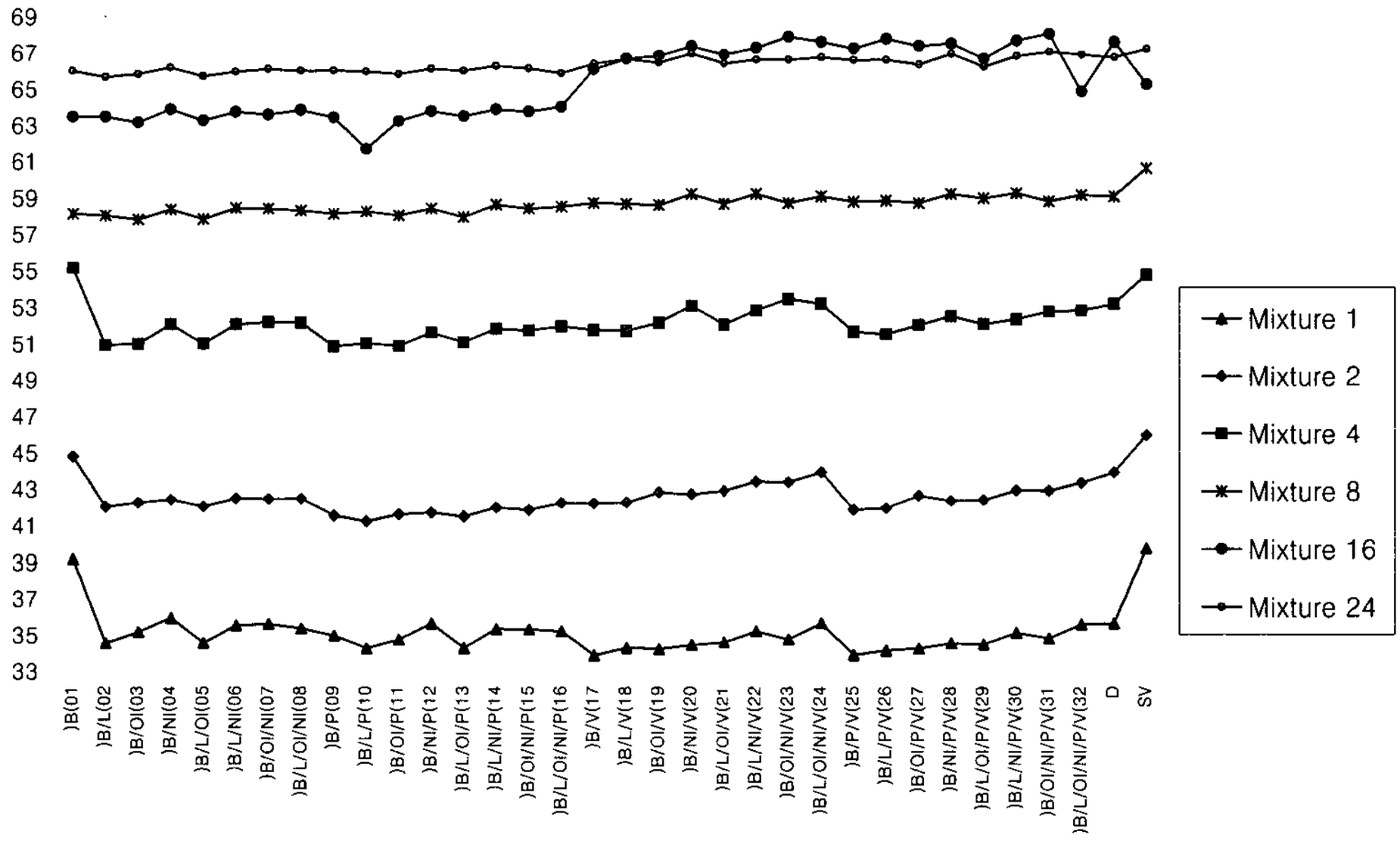
음성 현상 제약	모노폰의 개수
설측음화	1,078
장애음의 불파음화	1,652
비음의 불파음화	2,240
구개음화	798
유성음화	4,074
이중모음(반모음+단모음)	5,250

<그림 1>은 폰단위 인식 실험의 결과를 각각 그래프로 나타낸 것이다. 34개의

---

5) <표 7>에서 SV는 이중모음 제약을 고려하여 이중모음을 반모음과 단모음의 결합으로 설계한 집합이며, D는 SV 집합과 자음 조건은 동일하나 이중모음 제약을 고려하지 않은 집합이다. 4장에 제시될 그래프에도 해당 집합에 대해 동일한 표기를 하였다.

6) 4장에 제시된 모든 그래프에서는 각 유사음소단위 집합에서 고려한 음성 현상 제약을 다음과 같이 간략 표기하였다. (B: 음소, L: 설측음화, OI: 장애음의 불파음화, NI: 비음의 불파음화, P: 구개음화, V: 유성음화)



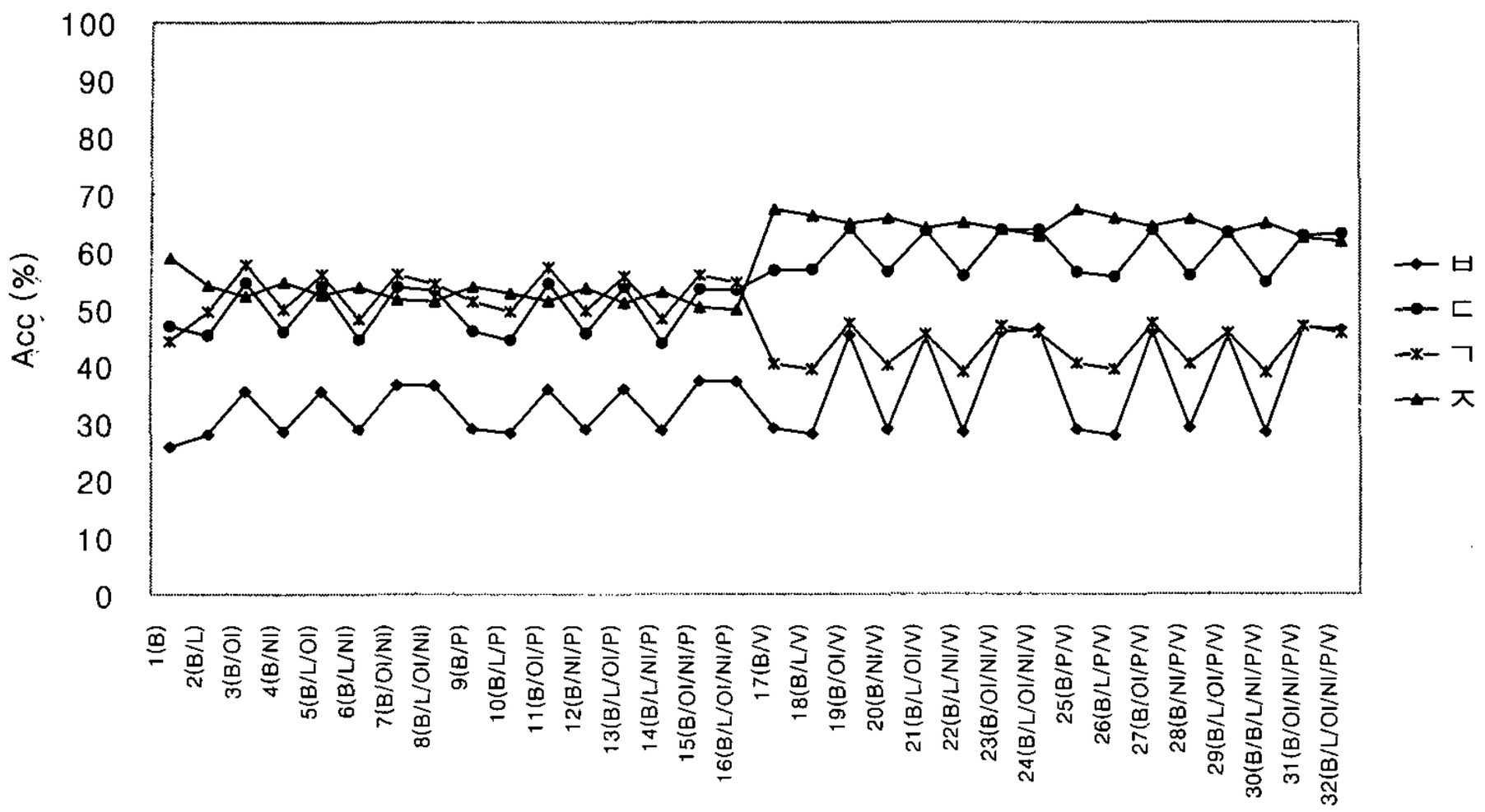
<그림 1> 믹스처 개수에 따른 폰단위 인식률

유사음소단위 집합에 따른 인식률의 변화를 믹스처(mixture) 개수를 1개에서 24개까지 증가시키면서 관찰하였다.

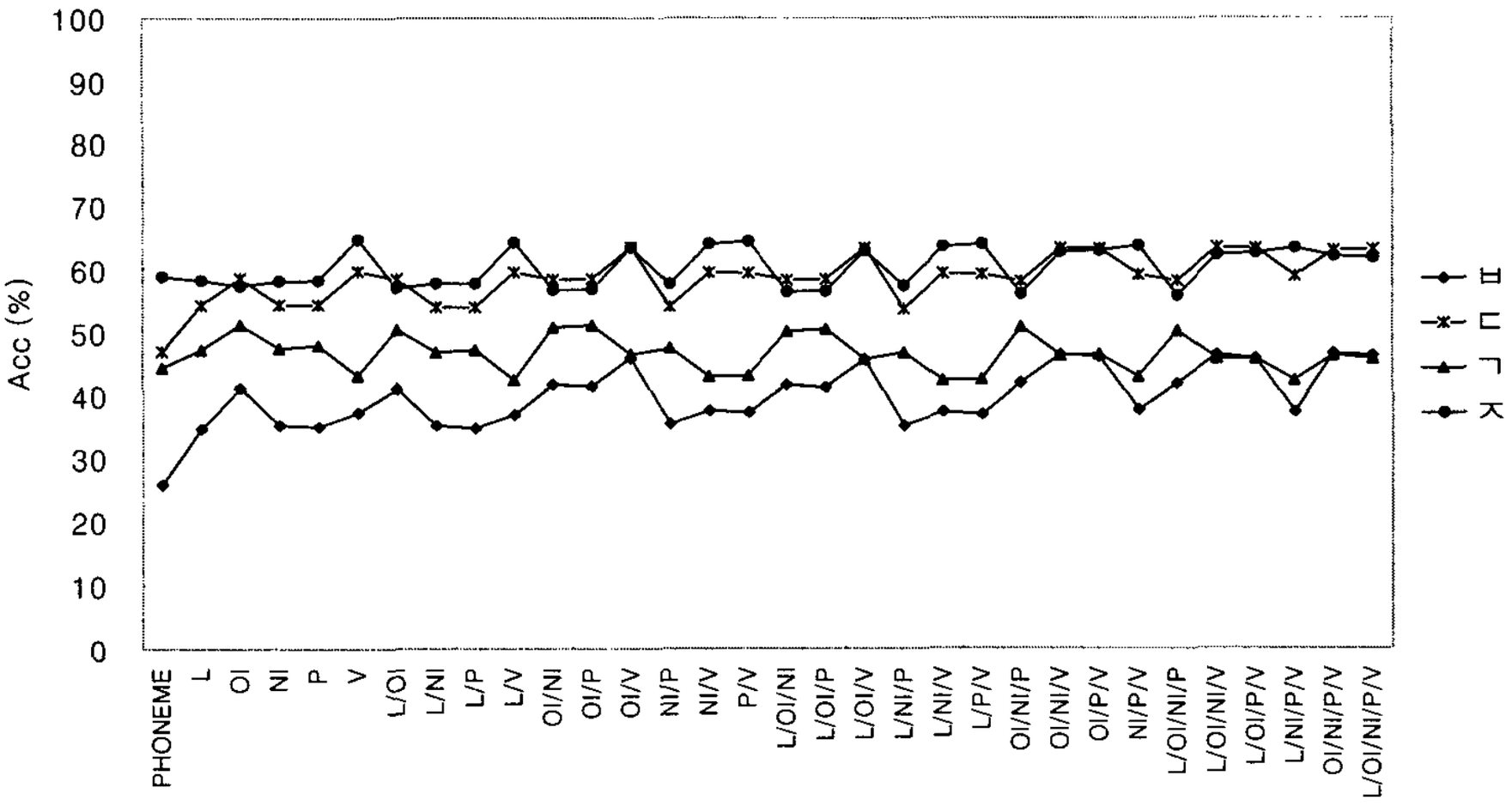
인식 실험 결과, 믹스처가 1~7개까지는 음소 기반 집합이 가장 좋은 인식률을 보였으나, 믹스처가 8개 이상인 경우에는 대체로 유사음소단위 수가 많을수록 인식률이 좋은 경향을 보였다. 이것은 믹스처의 개수가 증가하면서 믹스처의 개수가 적은 경우와는 달리 각 유사음소단위의 모델링이 제대로 이루어지면서 그 변별력이 강화되었다고 볼 수 있다. 이와 같이 믹스처가 폰단위 인식 결과에 상당한 영향을 주었음을 확인하였다.

다음으로, 각 제약이 음소 인식률에 미치는 영향을 확인하기 위하여 설계한 유사음소단위 집합을 대상으로 각 집합별 음소 인식률과 제약 고려에 따른 음소별 평균 인식률을 구하였다. 여기에서 음소 인식률이란 각 음소의 변이음에 대한 인식률의 평균으로, 예를 들어, /ㅂ/에 대한 음소 인식률은 {p}, {b}, {p˺}의 인식률의 평균을 말한다. 제약 고려에 따른 음소 인식률이란 제약에 따라 다른 음소별 인식률로서 예를 들면, 설측음화 제약을 고려한 집합의 인식률의 경우에는 설측음화 이외의 제약 고려 여부에 관계없이 설측음화 제약을 고려한 집합 모두에 대한 평균 인식률이다.

유사음소단위 집합별 음소 인식률과 음성 현상 제약 고려에 따른 평균 인식률을 분석한 결과, 음성 현상 제약 고려 여부가 각 모노폰의 인식률에 많은 영향을

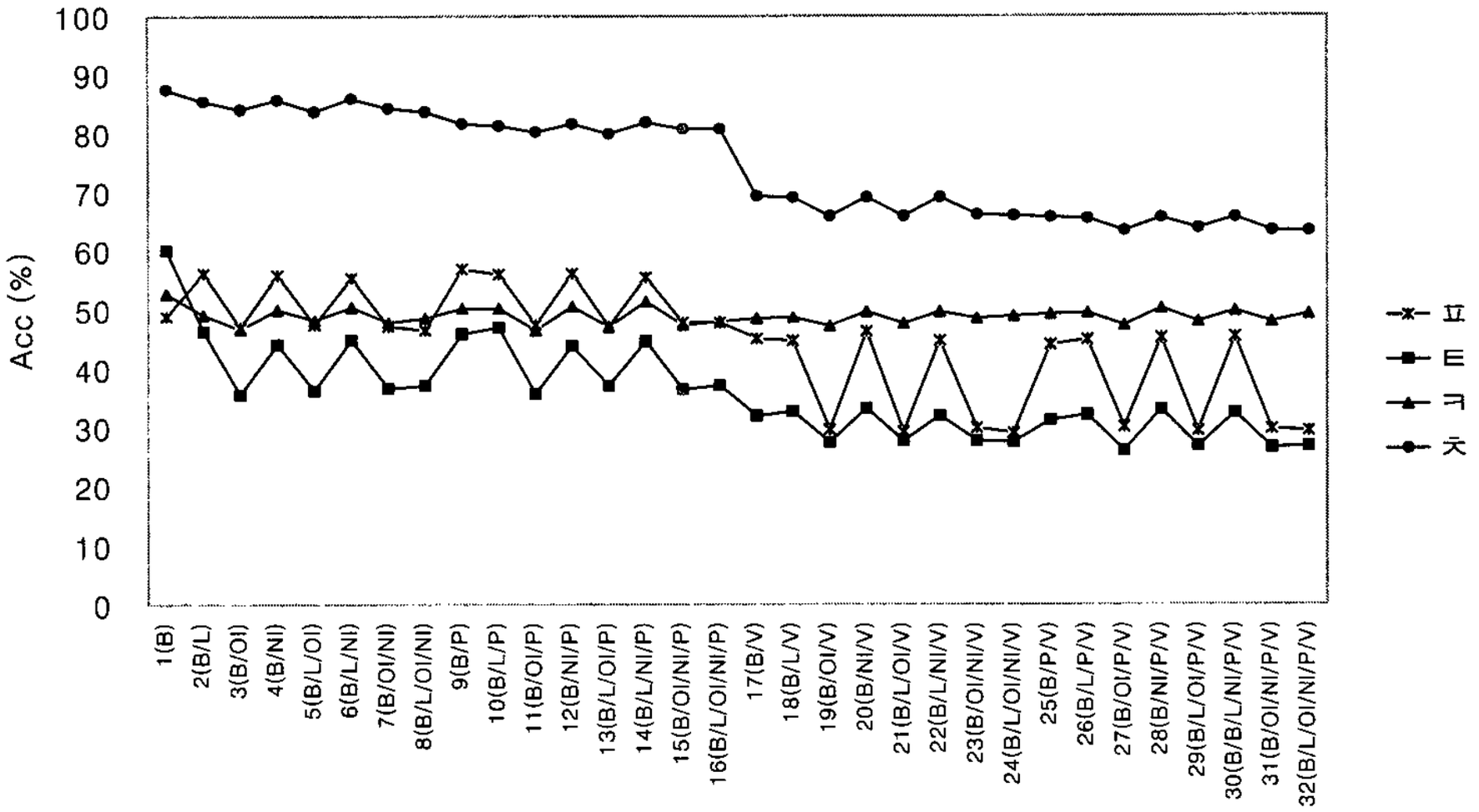


<그림 2> 유사음소단위 집합별 파열·파찰 평음 음소 인식률

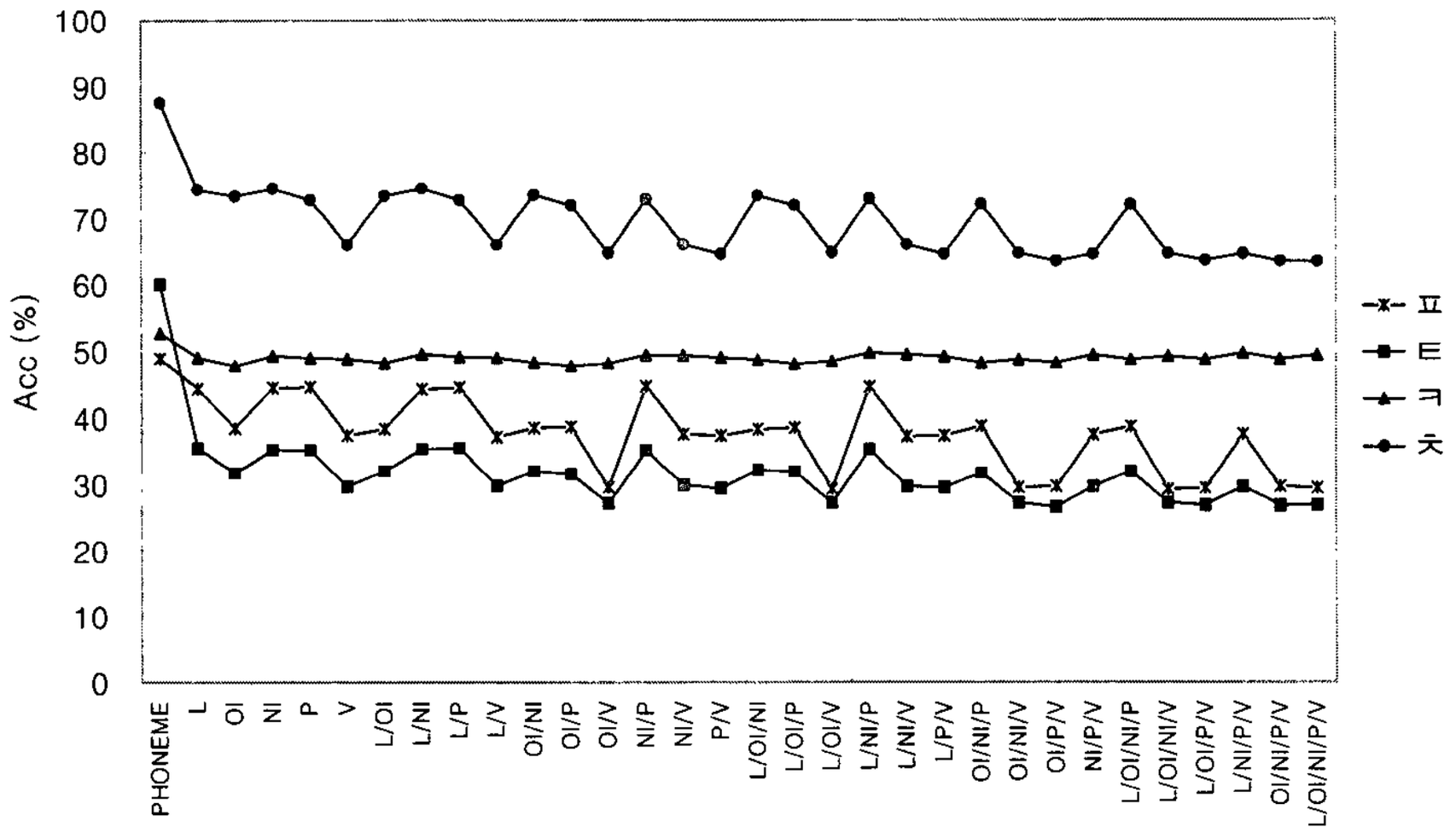


<그림 3> 제약 고려에 따른 파열·파찰 평음 평균 인식률

미치는 것으로 보인다. 예를 들어, 파열 평음 /ㅂ, ㄷ, ㄱ/는 <그림 2>, <그림 3>과 같이 장애음의 불파음화 현상을 고려했을 경우 뚜렷한 인식률 향상을 보였다. 반면, 파열 평음인 /ㅍ, ㅌ, ㅋ/는 <그림 4>, <그림 5>에서 보는 것과 같이 장애음의 불파음화 현상을 고려했을 경우에 반대로 인식률이 저하되는 경향을 보였다. 이

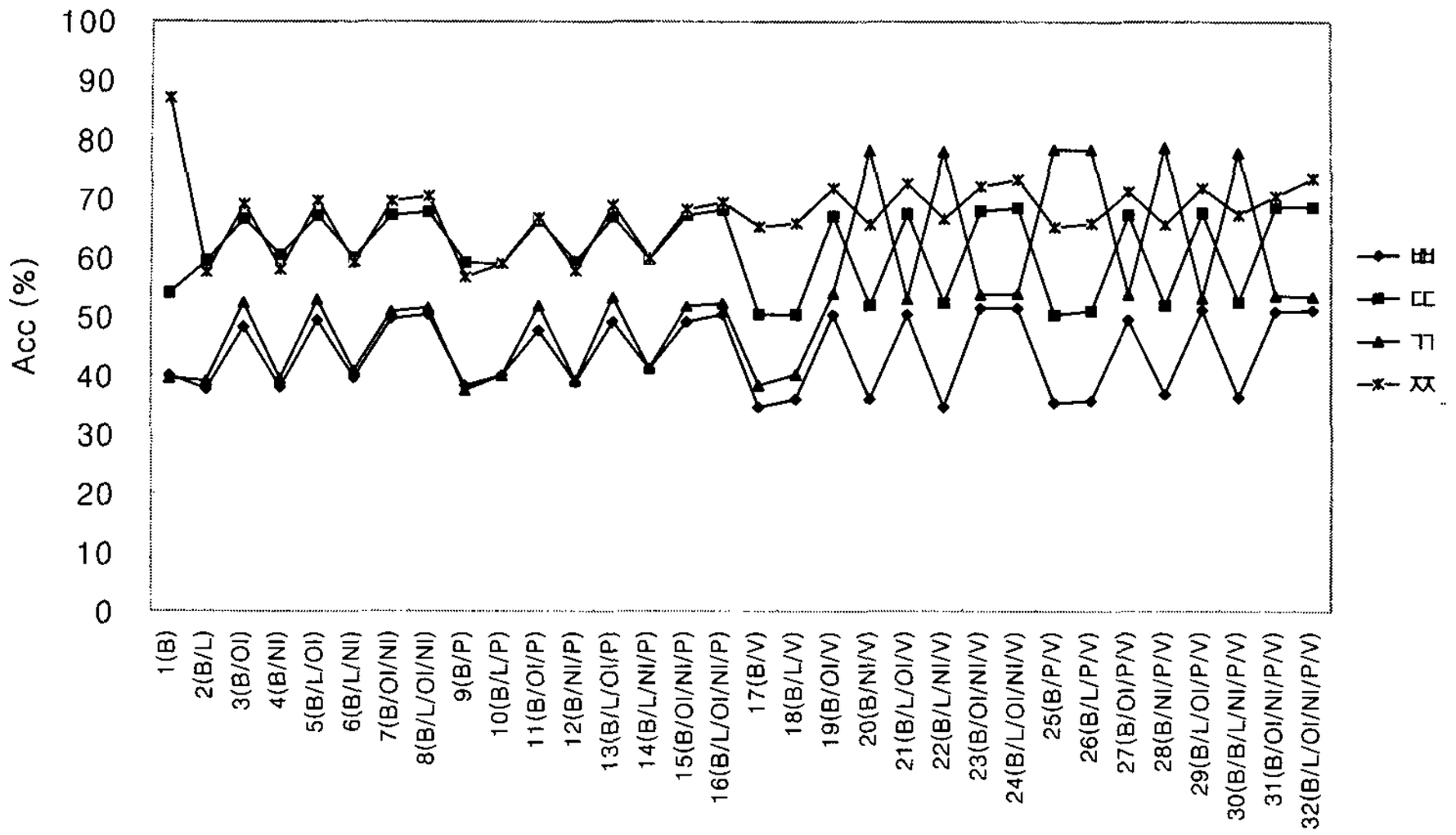


<그림 4> 유사음소단위 집합별 파열·파찰 격음 음소 인식률

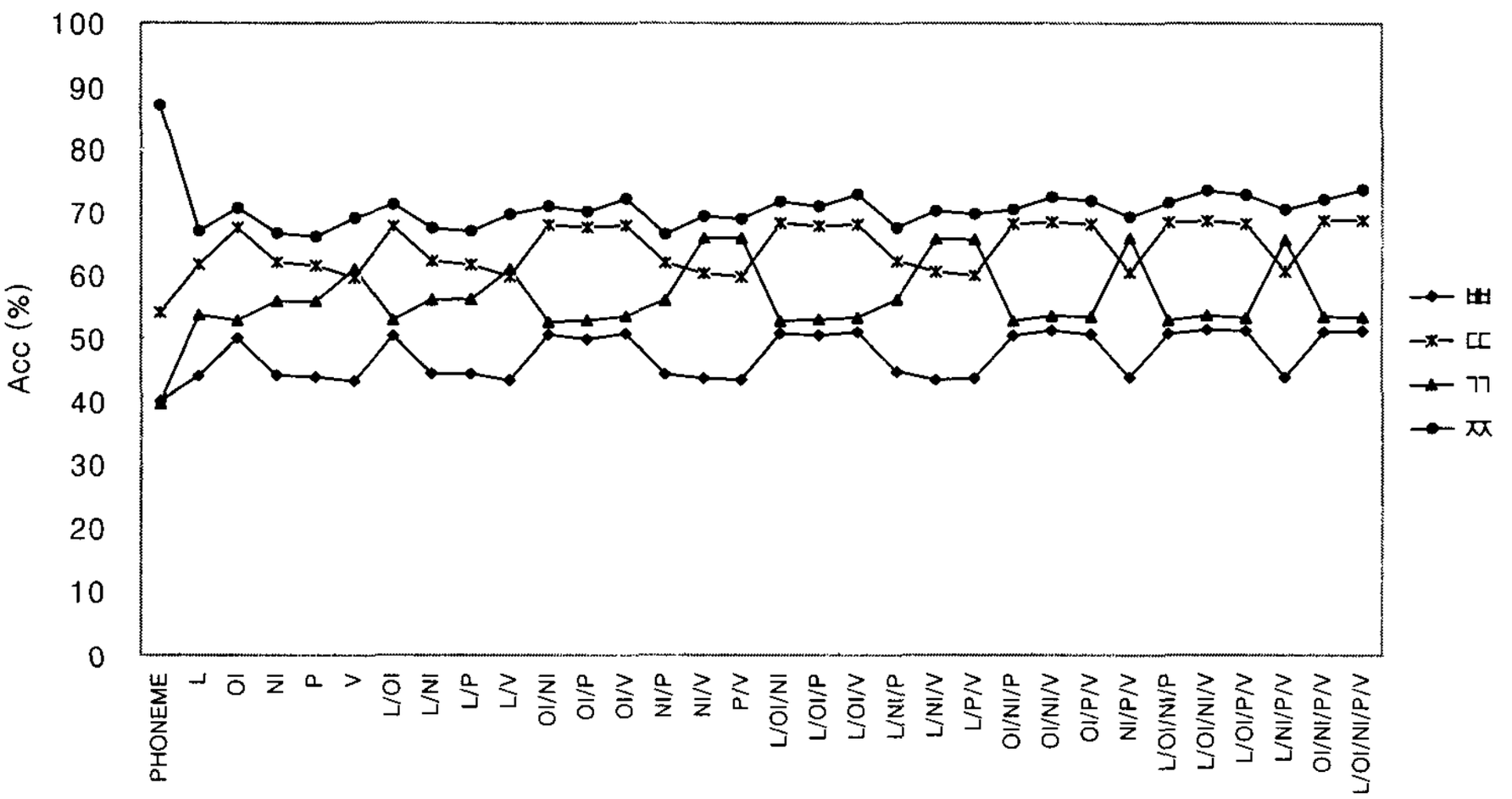


<그림 5> 제약 고려에 따른 파열·파찰 격음 평균 인식률

때, 파열 경음인 /ㅃ, ㄸ, ㄲ/는 <그림 6>, <그림 7>에서 보는 바와 같이 파열 평음의 경우와 마찬가지로 인식률이 향상되었다. 또 다른 예로, 유성음화를 고려하는 경우에 /ㅎ/의 인식률은 향상되지만, 반대로 /ㅍ, ㅌ, ㅊ/ 등의 격음은 인식률이 저하되었음을 확인하였다.



<그림 6> 유사음소단위 집합별 파열·파찰 경음 음소 인식률



<그림 7> 제약 고려에 따른 파열·파찰 경음 평균 인식률

이와 같이 음성 현상 제약을 고려함에 따라 어떤 모노폰의 경우에는 인식률이 향상되지만 반대로 인식률이 저하되는 모노폰도 존재하는데, 본 논문에서 고려하고 있는 음성 현상 제약과 각 모노폰의 인식률과의 관계를 요약하면 <표 9>와 같다.

<표 9> 음성 현상 제약에 따라 영향을 받는 음소

인식률이 향상된 음소	음성 현상 제약	인식률이 저하된 음소
/ㄹ/, /ㅢ/, /ㅓ/	설측음화	/ㅇ/, /ㅋ/
/ㅂ/, /ㄷ/, /ㄱ/, /ㅃ/, /ㄸ/, /ㄲ/, /ㅉ/, /ㅇ/, /ㅘ/, /ㅙ/	장애음의 불파음화	/ㅍ/, /ㅌ/, /ㅋ/, /ㅜ/, /ㅖ/
/ㅁ/, /ㄴ/, /ㅛ/, /ㅖ/	비음의 불파음화	/ㅇ/, /ㅢ/, /ㅓ/
/ㄴ/, /ㅇ/, /ㅣ/	구개음화	/ㅆ/, /ㅠ/, /ㅢ/
/ㅂ/, /ㄷ/, /ㄱ/, /ㅈ/, /ㅎ/, /ㅇ/, /ㄴ/	유성음화	/ㅍ/, /ㅌ/, /ㅊ/, /ㄸ/, /ㅑ/, /ㅒ/

이와 같은 폰단위 인식 결과에 따르면 음성 현상 제약을 고려할 때 인식률이 저하되는 모노폰에 비해 인식률이 향상되는 모노폰이 더 많이 분포할 때 전반적인 인식 결과가 좋을 것으로 예상할 수 있다. 즉, 최적의 유사음소단위 집합을 구성하기 위해서는 코퍼스 분석을 통해 각 제약의 영향을 받는 모노폰의 분포를 확인하고 이를 바탕으로 유사음소단위 집합을 매번 재구성해야 할 것으로 보인다.

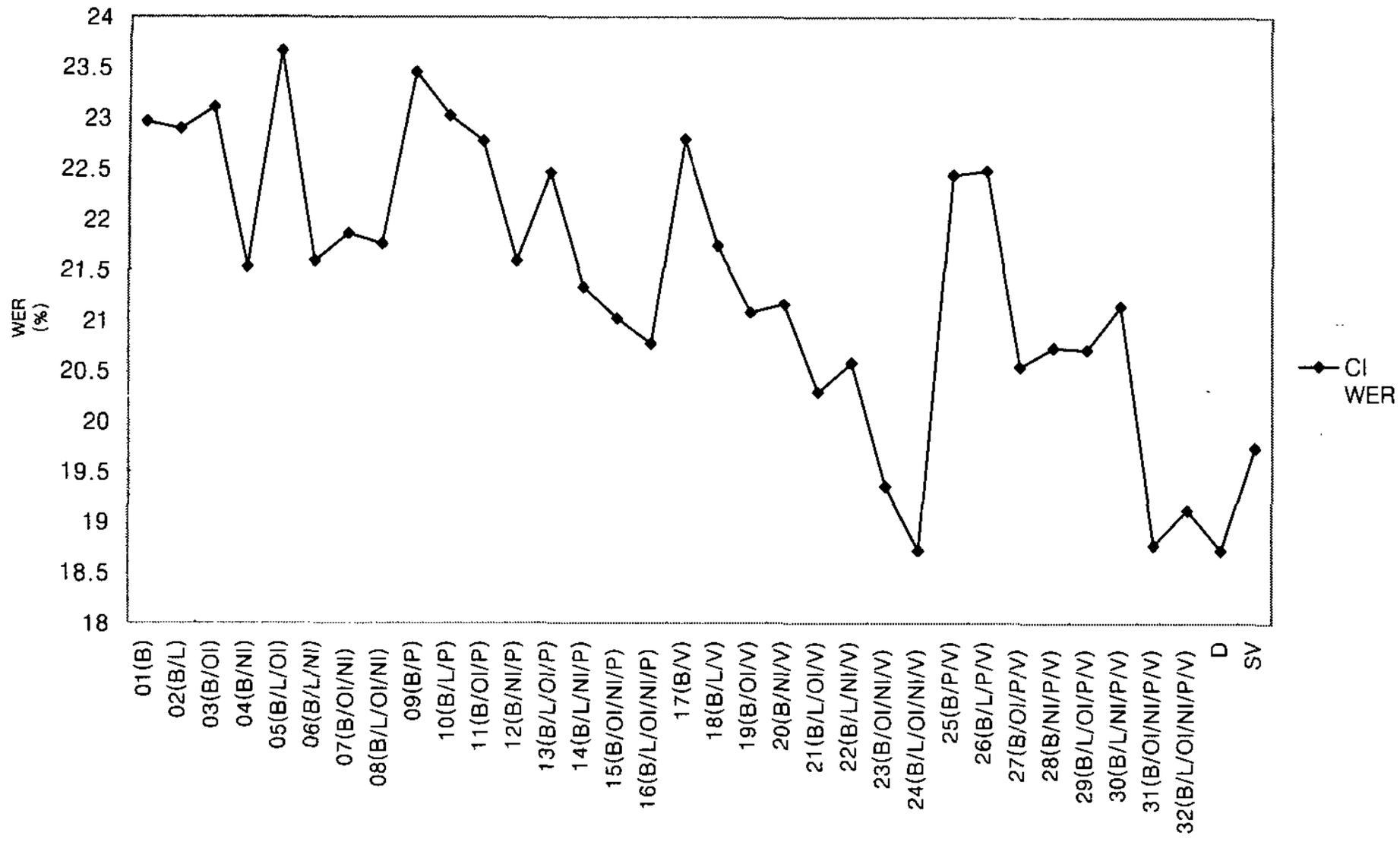
음성현상 제약 외에 폰단위 인식 결과에 영향을 주는 요소로 믹스처를 확인하였는데, 폰 바이그램 역시 폰단위 인식 결과에 상당한 영향을 주는 것으로 보인다. 예를 들어, /ㅣ/의 인식률은 구개음화를 고려했을 때 향상되었는데, 이는 {ɲ}, {ɕ} 등의 구개음화가 고려된 유사음소단위의 영향에 의한 것으로 추측할 수 있다.

4.2. 고립단어인식

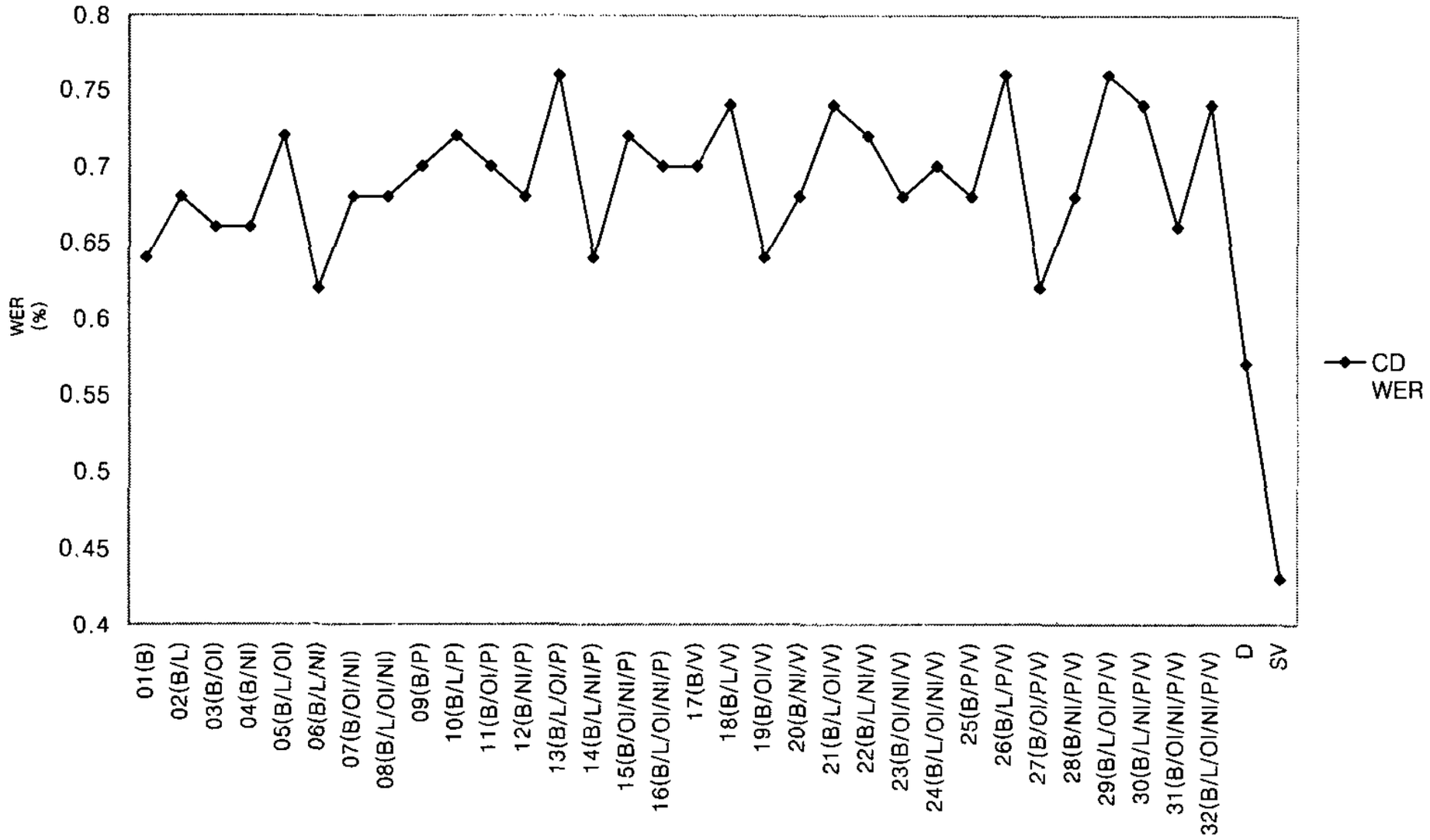
고립단어인식 실험에는 폰단위 인식 실험에서 사용한 것과 동일한 452단어 PBW 코퍼스를 사용하였다. 학습에는 120명의 발화가 사용되었으며, 14명의 발화를 인식 데이터로 사용하였다. 실험은 HTK toolkit 3.3에서 수행하였다.

문맥 독립(context independent) 단위로 인식 실험을 수행한 결과(<그림 8>), 유사음소단위 집합의 변화가 인식률에 영향을 준다고 볼 수 있으나, 그 구체적인 영향은 해석이 어려웠다. 설측음화, 장애음의 불파음화, 비음의 불파음화, 유성음화 제약을 고려한 24번 집합이 단어 오인식률(word error rate: WER) 18.84%로 인식 결과가 가장 좋았는데, 이는 폰단위 인식에서도 믹스처 16개를 기준으로 하여 폰단위 오인식률(phone error rate: PER) 32.4%를 보여 모든 집합의 평균 34.6%와 비교하였을 때 비교적 좋은 인식 결과를 보인 집합이다.

반면, 문맥 종속(context dependent) 단위로 인식 실험을 수행한 결과(<그림 9>), 유사음소단위 집합의 변화가 인식 결과에 큰 영향을 준다고 보기는 어려웠다. 즉, 트라이폰(triphone)를 사용하는 경우에는 유사음소단위 집합에 거의 영향을 받지 않고 거의 비슷한 인식률을 보였다. 이것은 문맥 종속 단위가 문맥 정보를 포함하기 때문에 세분화된 유사음소단위 집합을 사용하지 않아도 문맥 독립 단위로 인식을 수행하는 것에 비해 변별력을 획득할 수 있기 때문이라고 본다.



<그림 8> 고립단어인식 결과(CI)



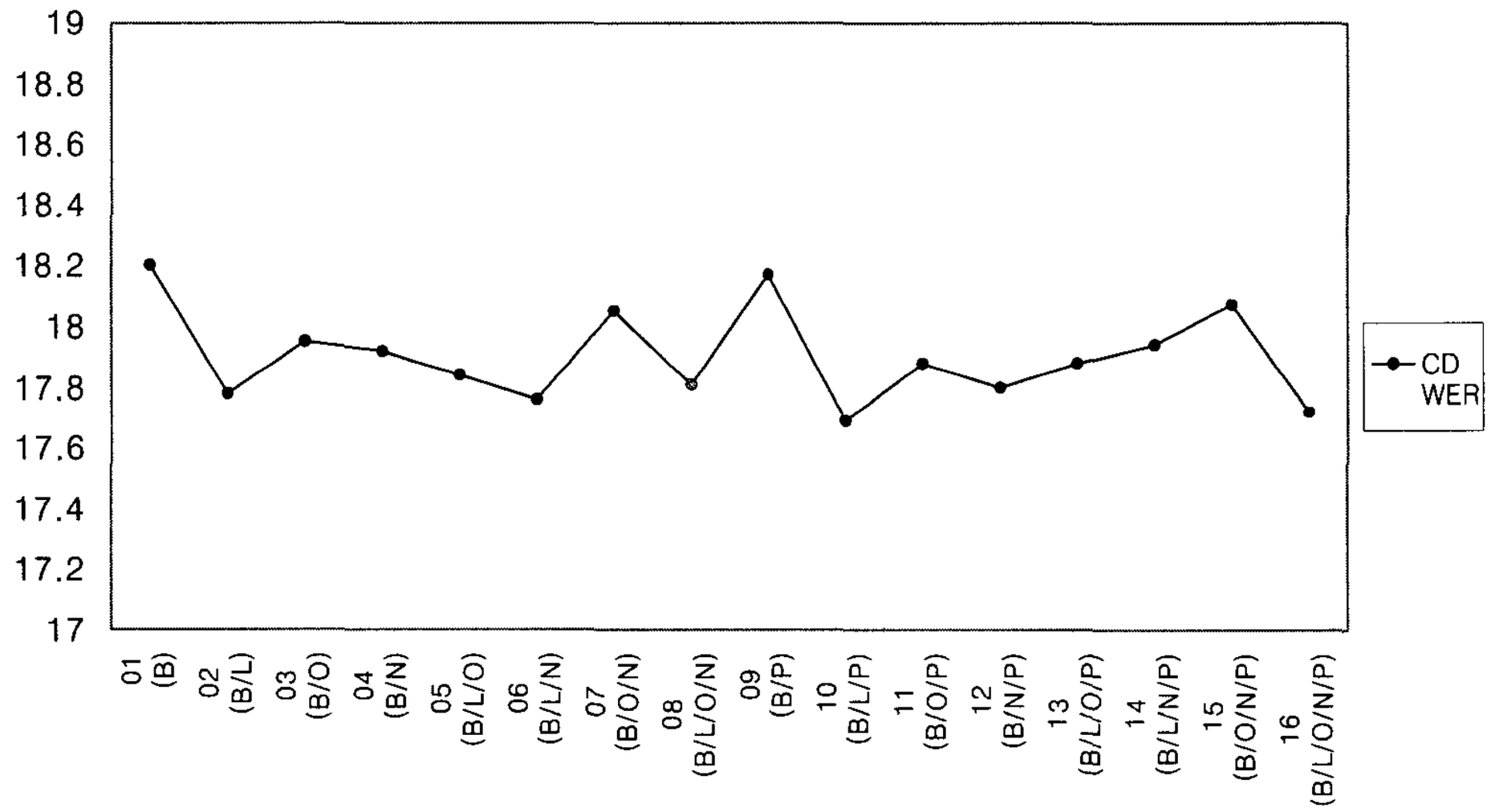
<그림 9> 고립단어인식 결과(CD)

4.3. 연속음성인식

연속음성인식 실험은 3장에서 설계한 34개의 집합 가운데 1~16번 집합을 대상으로 수행하였다. 연속음성인식 실험에는 연속음성 낭독체 코퍼스가 사용되었다. 이 가운데 43,000문장을 학습에 사용하였으며, 1,616문장을 인식 실험에 사용하였다. 단어인식 과정에서 사용한 발음사전의 표제어는 39,693개였고, 서브워드 모델은 각 유사음소단위에 대해 학습된 HMM이다. 이 HMM은 모노폰에서 확장된 음소 문맥 정보를 포함하는 트라이폰으로 학습되었다. 언어모델은 트라이그램 모델을 사용하여 실험을 수행하였다. 이때, 음성 현상 제약에 따라 영향을 받는 모노폰의 개수는 <표 10>과 같다.

<표 10> 음성 현상 제약에 따라 영향을 받는 모노폰의 개수

음성 현상 제약	모노폰의 개수
설측음화	4,134
장애음의 불파음화	3,577
비음의 불파음화	8,736
구개음화	1,977



<그림 10> 연속음성인식 결과

인식 실험 결과, <그림 10>에서 보는 것처럼 음소 기반 집합보다는 모든 변이음 기반 집합의 경우에 더 낮은 WER를 보였다. 16개 집합 가운데에서 설측음화, 구개음화 제약을 고려한 10번 집합이 WER 17.69%로 인식 결과가 가장 좋았는데,

이것은 음소 기반 집합(WER 18.20%)에 비해 상대적으로 2.80%의 향상도를 보인 것이다. 그러나 폰단위 인식, 고립단어인식 실험에서는 10번 집합의 인식 결과는 각각 PER 38.33%(믹스처 16개 기준), WER 23.02%(CI), WER 0.72%(CD)로, 집합의 평균인 PER 34.6%, WER 21.35%(CI), WER 0.68%(CD)에 비교해 볼 때, 좋지 않은 결과를 보인다.

5. 결 론

본 논문은 한국어 음성인식 시스템의 성능 향상을 위하여 변별력이 강화된 새로운 유사음소단위 집합 설계를 위한 기초 연구로서, 음성학적 지식을 바탕으로 국내외 주요기관의 기존 유사음소단위 집합을 검토한 다음, 이를 기반으로 새로운 유사음소단위 집합을 설계하고 인식 실험을 통하여 그 성능을 평가하였다.

국내 기관(9개)에서 사용하는 유사음소단위 집합들을 검토한 결과, 음성학적 측면에서의 체계적인 분류가 부족하며 자의적으로 유사음소단위 집합을 정의하여 사용하고 있음을 알 수 있었다. 따라서 본 논문에서는 음소 및 변이음 기반으로 유사음소단위 집합 34개를 설계하여 폰단위 인식, 고립단어인식 및 연속음성인식 실험을 수행하여 한국어 음성인식에 영향을 미치는 음성 현상과 변이음을 분석하였다.

폰단위 인식 실험 결과, 변이음 기반 유사음소단위 집합을 구성하는 음성 현상 제약의 고려 여부가 각 모노폰의 인식률에 많은 영향을 미치는 것을 확인하였다. 최적의 유사음소단위 집합을 구성하기 위해서는 코퍼스 분석을 통해 각 제약의 영향을 받는 모노폰의 분포를 확인하고 이를 바탕으로 유사음소단위 집합을 매번 재구성해야 할 것으로 보인다. 그러나 인식 영역에 따라 유사음소단위 집합을 새로 구성하는 것은 음향모델의 재학습, 발음열의 재생성 등의 문제점이 있어 현실적으로 불가능하다. 따라서 새로운 영역의 코퍼스나 음향적 환경이 상이한 코퍼스에서도 대체적으로 높은 성능을 보이는 유사음소단위 집합을 설계하기 위해서는 phonetically balanced sentence (PBS)와 같은 음소 균형 코퍼스를 바탕으로 기본 유사음소단위 집합을 설계한 후, 이를 음성 분포가 다른 여러 다양한 코퍼스 환경에서 그 성능을 평가하는 방안을 생각해 볼 수 있다. 이때, 응용 영역의 특성에 따라 달라질 수 있는 음성 현상을 반영할 수 있도록 음향모델 적응과 발음사전의 발음열 생성 파라미터 조정, 디코더의 제약사항 반영 등을 통해 유사음소단위 집합과 음성 현상의 대응 관계에 관한 후속 연구가 필요할 것이다.

고립단어인식 실험에서는 전반적으로 음소 기반 유사음소단위 집합에 비하여 변이음 기반의 유사음소단위 집합을 사용하는 경우에 인식률이 향상되었으나, 문맥 종속 단위로 인식 실험을 수행한 경우에는 문맥 독립 단위로 인식을 수행하는

경우에 비해 각 유사음소단위 집합에 따른 인식 성능의 변화가 적었다. 이러한 결과는 문맥 독립 단위나 다이폰(diphone)을 사용하는 내장형 솔루션의 경우에 직접 이용될 수 있을 것이다.

연속음성인식에서는 음소 기반 집합보다는 모든 변이음 기반 집합의 경우에 더 낮은 WER를 보여 변이음 기반의 유사음소집합과 인식률 사이에 어떠한 상관관계가 있다는 것을 볼 수 있었으나, 유사음소단위 집합의 변화가 연속음성 인식의 성능에 미치는 구체적인 영향을 해석하는 연구가 후속적으로 수행되어야 할 것이다.

참 고 문 헌

[1] M. Hunt, M. Lennig, P. Mermelstein, "Experiments in syllable-based recognition of continuous speech", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 5, No. 1, pp. 880-883, 1980.

[2] W. Weigel, G. Ruske, "Continuous speech recognition using syllabic segmentation demisyllable hidden Markov models", *Proc. EUROSPEECH*, pp. 1017-1020, 1989.

[3] R. Singh, B. Raj, R. M. Stern, "Automatic generation of phone sets and lexical transcriptions", *Proc. ICASSP*, pp. 1691-1694, 2000.

[4] R. Singh, B. Raj, R. M. Stern, "Automatic generation of subword units for speech recognition systems", *IEEE Transactions on Speech and Audio Processing*, Vol. 10, No. 2, pp. 89-99, 2002.

[5] Y. Liu, P. Fung, "Automatic phone set extension with confidence measure for spontaneous speech", in *Proc. EUROSPEECH*, pp. 2741-2744, 2003.

[6] 임영춘, 오세진, 김광동, 노덕규, 송민규, 정현열, "음성인식에서 문맥의존 음향모델의 성능향상을 위한 유사음소단위에 관한 연구", *한국음향학회지*, 제22권, 제5호, pp. 388-402, 2003.

[7] 서영주, 성철재, 이정철, 한민수, 이영직, "음성학적 지식에 기반한 한국어 변이음 집단화 수형도의 구현", *제13회 음성통신 및 신호처리 학술대회 논문집(KSCSP)*, 제13권, 제1호, pp. 344-347, 1996.

[8] 김회린, 이항섭, "음성학적 지식 기반 변이음 모델을 이용한 가변 어휘 단어 인식기", *한국음향학회지*, 제16권, 제2호, pp. 31-35, 1997.

[9] 이경님, 정민화, "연속음성 인식에서의 발음사전 최적화를 위한 변이음 규칙 적용 방법", *제19회 음성통신 및 신호처리 학술대회 논문집(KSCSP)*, 제19권, 제1호, pp. 149-152, 2002.

[10] 이숙향, 신지영, 김봉완, 이용주, "음성 코퍼스 구축을 위한 SiTEC 분절음·운율 레이블링 기준의 검토 및 제안", *말소리*, 제46호, pp. 127-143, 2003.

[11] 정국, 구희산, 이찬도, 김종미, 한선희, "음성 인식/합성을 위한 국어의 음성-음운론적 특성 연구", *한국음향학회지*, 제13권, 제6호, pp. 31-44, 1994.

[12] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren,

"The DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM", NTIS order number PB91-100354, 1990.

[13] K. F. Lee, H. W. Hon, "Speaker-independent phone recognition using hidden Markov models", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 37, No. 11, pp. 1641-1648, 1989.

[14] J. L. Gauvain, L. F. Lamel, "Speaker-independent phone recognition using BREF", *Proc. DARPA Speech and Natural Language Workshop*, pp. 344-349, 1992.

[15] 김상훈, 이용주, "음성 DB 표준화", *제20회 음성통신 및 신호처리 학술대회 논문집 (KSCSP)*, pp. 181-184, 2003.

[16] 이호영, *국어 음성학*, 태학사, 1996.

[17] 이호영, "한국어의 변이음 규칙과 변이음의 결정 요인들", *말소리*, 제21호, pp. 144-175, 1993.

[18] 김종미, "자음의 단어내 음운환경별로 본 음가변화", *한국음향학회지*, 제13권, 제5호, pp. 69-76, 1994.

[19] 이호영, 지민제, 김영송, "동시조음에 의한 변이음들의 음향적 특성", *한글*, 제220호, pp. 5-28, 1993.

접수일자: 2008년 2월 20일
게재결정: 2008년 3월 17일

▶ 홍혜진(Hyejin Hong)
주소: 151-745 서울특별시 관악구 관악로 599 서울대학교 인문관 1동 426호
소속: 서울대학교 언어학과
전화: 02) 880-9039
E-mail: souble1@snu.ac.kr

▶ 김선희(Sunhee Kim)
주소: 151-745 서울특별시 관악구 관악로 599 서울대학교 인문관 5동 313호
소속: 서울대학교 인문학연구원
전화: 02) 880-7735
E-mail: sunhkim@snu.ac.kr

▶ 정민화(Minhwa Chung) : 교신저자
주소: 151-745 서울특별시 관악구 관악로 599 서울대학교 인문관 3동 406호
소속: 서울대학교 언어학과
전화: 02) 880-9195
E-mail: mchung@snu.ac.kr