발음열 자동 변환을 이용한 한국어 음운 변화 규칙의 통계적 분석

이 경 님, 정 민 화
서강대학교 컴퓨터학과

Statistical Analysis of Korean Phonological Rules Using a Automatic Phonetic Transcription

Kyong-Nim Lee, Minhwa Chung
School of Computer Science, Sogang University
E-mail : {knlee, mchung}@sogang.ac.kr

Abstract

We present a statistical analysis of Korean phonological variations using automatic generation of phonetic transcription. We have constructed the automatic generation system of Korean pronunciation variants by applying rules modeling obligatory and optional phonemic changes and allophonic changes. These rules are derived from knowledge-based morphophonological analysis and government standard pronunciation rules. This system is optimized for continuous speech recognition by generating phonetic transcriptions for training and constructing a pronunciation dictionary for recognition. In this paper, we describe Korean phonological variations by analyzing the statistics of phonemic change rule applications for the 60,000 sentences in the Samsung PBS(Phonetic Balanced Sentence) Speech DB. Our results show that the most frequently happening obligatory phonemic variations are in the order of *liaison, tensification, aspirationalization,* and *nasalization of obstruent*, and that the most frequently happening optional phonemic variations are in the order of *initial consonant h-deletion, insertion of final consonant with the same place of articulation as the next consonants,* and *deletion of final consonant with the same place of articulation as the next consonants.* These statistics can be used for improving the performance of speech recognition systems.

I. 서론

음성 인식 시스템을 구성할 때 일반적으로 정확한 발음열을 반영하는 것이 인식률 향상에 도움이 되며, 음성 합성에서도 합성음의 명료성과 자연성을 높이기 위해서 발성 상황에 따라 여러 가지 형태의 발음열 생성이 필요하다. 이러한 이유로 음성학적 발음 특징에 대한 연구를 토대로 변환 규칙을 정의하고, 규칙에 기반해서 입력 문장을 보다 정확한 발음표기로 변환시키는 시스템들이 개발되었으며[1][2], 발음열 자동 생성 방법에 관한 다양한 연구가 이루어지고 있다[3].

본 논문에서는 [1]의 발음열 자동 생성 과정에서 적용된 음소 변동 규칙들의 통계적 자료를 기반으로 한국어 음운 변화 현상에 대한 분석을 수행하였다. [4]를 포함하여 기존 연구들은 한글 철자에 대한 통계적 분석이 대부분이며, [5]의 경우 발음사전에 기재된 약 66만개의 표제어에 대한 발음(음운)을 조사하여 음소와 음절들의 빈도수를 조사 분석한 통계 자료를 제시하였으나, 실제 문장에서 발생하는 형태소 및 어절 경계의 음운 변화 현상은 반영되지 않았으며 적용된 규칙에 대한 정보를 알 수 없다는 한계점이 있었다.

본 실험에 사용된 분석 대상은 트라이폰 기준으로 다양한 음운환경을 포함하며 음소열의 중복이 적고 고른 확률분포를 갖는 문장들의 집합이다. 실험 분석은 본 논문에서 정의한 음소 변동 규칙에 따른 발생 빈도수와 음소의 경계 위치에 따른 적용 양상에 대하여 초점을 맞추었다. 적용된 음소 변동 규칙들의 통계적 자

료를 기반으로 한국어 음운 변화 현상의 양상을 파악할 수 있었으며, 나아가 이러한 분석을 이용하여 음성인식기의 성능을 향상시키기 위한 분석자료로 활용할 수 있을 것이다.

II. 한국어 음소 변동 규칙 정의

발음열 생성 시스템의 중요 모듈은 문자열을 발음열로 변환하는 부분으로 문장, 끊어읽기 단위인 언절, 띄어쓰기 단위 어절, 그리고 단어 등 주어진 텍스트를 입력으로 받아 그에 대응하는 발음열을 생성하는 역할을 한다. 이 때 문자열에 대한 올바른 발음열을 생성하기 위해서는 해당 언어의 음운 현상에 대한 체계적이고, 정확한 분석이 필요하다. 본 논문에서는 음성학과 음운론 연구[6][7]을 기반으로 한국어에서 발생하는 음운 변화 현상을 정리하고, 문교부에서 제정한 표준어 규정[8]의 제 2부 표준 발음법을 참고하여 한국어의 대표적인 음소 변동 규칙 중 표 1과 같이 20개의 음소 변동 규칙을 채택하여 적용하였다.

한국어의 경우 주로 자음 변화가 심하기 때문에 이에 대한 연구는 많지만 모음에 대한 연구는 체계적이지 못하다. 주로 사투리나 방언에 대한 연구가 많으며, 특히 발화 속도나 습관에 따라 변화가 다양하여 텍스트 형태와 분석 자료만을 가지고 규칙화하기 어렵다. 사투리나 방언의 경우는 텍스트에 이미 반영된 철자만을 대상으로 하고, 표준 발음법에서 제시된 변화 현상만을 그 대상으로 삼아 모음 관련 규칙을 반영하였다.

표 1에서 규칙 이름에 *가 표시된 것은 수의적 음소 변동 규칙을 나타내며, 총 13개의 필수 음소 변동 규칙과 7개의 수의적 음소 변동 규칙을 갖는다. 각 음소 변동 규칙들은 적용되는 음소 문맥 별로 다시 세부 규칙 번호가 주어지고, 이에 따라 실제 음소 문맥에 규칙이 적용된다. 음소 문맥에 따른 세부 규칙의 수는 총 816개이며, 자음과 모음을 기준으로 분류하면 자음 관련 규칙 775개와 모음 관련 규칙 41개이며, 필수 적용 규칙과 수의적 규칙으로 분류하면 필수 음소 변동 규칙 757개와 수의적 음소 변동 규칙 59개이다.

III. 형태음운론적 분석에 기반한 발음열 자동 생성

본 논문에서 사용된 발음열 자동 생성기 알고리즘은 한국어의 음운 변화 규칙을 다음과 같이 3단계로 나누어 진행된다. 해당 음소 문맥에 의해 하나의 음소가 다른 음소로 바뀌거나, 탈락, 첨가되는 양상을 규칙화

표 1. 음소 변동 규칙과 표준 발음법의 대응 관계

음소 변동 규칙					표준 발음법	
분류		규칙 이름	규칙 번호	세부 규칙수	장	항
자음 관련 규칙	종성 규칙	음절말 중화	1	117	4	9
		자음군 단순화	2	256		8, 10, 11
		격음화	3	21		12
		연음법칙	4	42		13~ 15
	음의 동화	유음화	5	10	5	20
		장애음의 비음화	6	34		18, 7장 30항
		유음의 비음화	7	19		19
		변자음화*	14	17		21
	구개음화		8	3	5	17
	경음화		9	136	6	23~28
	첨가	ㄴ-첨가*	11	30	7	29
		중복 자음화*	13	6		30
	탈락	종성 ㅎ-탈락	10	1	4	12-4
		초성 ㅎ-탈락*	15	5	-	-
		동일 조음위치 자음탈락*	12	7	-	-
모음 관련 규칙	자음 첫소리 '의' 단모음화		16	18	2	5항 다만3
	용언의 활용형 '져, 쪄, 쳐' 단모음화		17	3		5항 다만1
	'례' 단모음화*		18	17		5항 다만2
	첫 음절 외 '의' 단모음화*		19	2		5항 다만4
	용언 어미 '어' 이중모음화*		20	1	5	22

한 것을 '음소 변동 규칙'이라 정의하고, 표준 발음 생성을 위한 필수 음소 변동 규칙과 비표준 발음을 포함하여 화자의 습관 및 환경에 따라 발생 가능한 수의적 음소 변동 규칙을 단계별로 적용하였다. 마지막으로 하나의 음소가 음성 환경, 말의 속도와 스타일에 따라서 여러 가지 음가를 가지는 변이음 생성 규칙을 적용하였다.

3.1 연속음성인식을 위한 발음열 자동생성

한국어의 음운 변화는 음소의 배열과 형태소의 종류에 따라서 영향을 받는다. 같은 음소의 배열이라 하더라도 그 음소열이 '하나의 형태소 내부에 있는가', '형태소 경계에 위치하는가', 또는 '어절 경계에 위치하는가'에 따라 각기 다른 음운 변화 현상을 보여준다. 특히 한국어 문장은 하나 이상의 형태소들이 결합된 어절들로 구성되므로 형태소를 디코딩 단위로 삼는 경우 형태소 및 어절 경계에서 발생하는 음운 변화 현상이 반영되어야 한다. 또한 같은 음소 문맥 정보를 갖더라

도 발음열 생성시 현 위치와 품사 정보에 따라 변화하는 현상이 달라지기도 한다. 이러한 한국어의 특징을 잘 반영하여 발음열을 생성하려면 주어진 문장을 형태소 분석하고, 올바른 형태소열로 태깅하여 그 정보를 이용해야 한다. 그림 1은 형태음운론적인 분석을 통해 '신발을 신고, 신고하러 갔다'라는 문장이 음소 문맥과 경계 위치에 따라 해당 음소 변동 규칙이 적용되는 과정을 나타낸다. 주어진 문장은 끊어읽기 단위인 2개의 언절로서 총 4개의 어절과 9개의 형태소로 구성된다. 그림에서 어절 경계는 '//'으로 표현하였으며, 형태소 경계는 '+'로, 품사 태그 정보는 형태소 뒤에 '/'를 붙여 기재하였다.

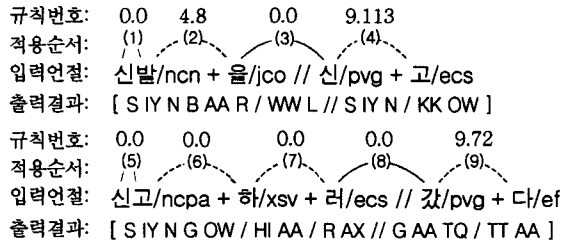

그림 1. 음소 문맥과 경계 정보에 따른 규칙 적용

위 예제에 적용된 필수 음소 변동 규칙의 세부규칙 표현은 아래 표 2와 같다. 음소 문맥 항의 L3는 음소 변동이 일어나는 음절 경계의 앞 음절의 종성을 나타내고, R1은 뒷음절 초성을 나타낸다. 변환 코드는 해당 음소 문맥에 대한 음소의 변동 결과를 나타낸다. 적용범위는 세부규칙의 적용범위와 적용 양상을 나타낸다[1]. 그림 1의 필수 음소 변동 규칙 적용 과정을 보면, (1)(5)에는 형태소 내부 규칙, (2)(4)(6)(7)(9)에는 형태소 경계 규칙, (3)(8)에는 어절간 규칙이 경계 위치에 따라 서로 다르게 적용하였다. 입력 언절 1)과 2)의 '신고'라는 단어는 서로 같은 철자이지만, '신고'가 어간과 어미의 결합인 경우 경음화 규칙 중 세부규칙 9.113에 의해 /신꼬/로 변환되고, 명사인 경우 변화없이 /신고/로 발화된다. 일반적으로 어절간에 일어나는 음운 변화 현상은 수의적 변이음 규칙으로 (8)에서 유성음화 규칙이 적용된 것을 볼 수 있다.

표 2. 그림 1에 적용된 세부 필수 음소 변동 규칙

음소 문맥		→	변환 코드		규칙 번호	세부 규칙 번호	적용범위
L3	R1		L3	R1			어/형/복/내/수/다
ㄹ	ㅇ	→	Ø	ㄹ	4	8	1 1 0 0 0 0
ㅆ	ㄷ	→	ㄷ	ㄸ	9	72	1 1 1 1 0 0
ㄴ	ㄱ	→	ㄴ	ㄲ		113	0 1 1 0 0 0

Ⅳ. 실험결과 및 분석

4.1 실험 데이터베이스

실험 결과의 합당성을 뒷받침하기 위해서는 본 실험에 사용한 DB의 검증 및 분석이 필요하다. 본 논문에서는 발생 가능한 모든 음운 현상을 포함하며, 가능한 다양한 트라이폰 모델을 포함하도록 설계된 삼성 PBS(Phone Balanced Sentence) 음성 DB의 문장을 실험에 사용하였다. 문장 분석은 형태소 분석 결과에 품사 태그가 부착된 형태를 기준으로 하였다. 분석 결과 한 문장 당 9.2어절, 한 어절 당 2.1 형태소, 한 형태소 당 1.9음절로 구성되었으며, 형태소 경계는 약 608,777회 발생하였다.

이 논문에서 사용된 삼성 PBS 60,000문장과 연속음성 인식기의 언어모델 생성을 위해 수집 가공한 신문 및 방송 뉴스 대상의 7M 형태소 텍스트 코퍼스로부터 발음열 생성기를 사용해서 얻은 트라이폰 수를 비교 분석하였다. 6배 이상의 텍스트 크기를 갖는 7M 형태소를 기준으로 실험에 사용된 DB는 1회 발생 기준으로 약 79%, 5회 이상 발생 기준으로 약 74%의 트라이폰을 포함하고 있으며, 가능한 트라이폰을 균형적으로 포함하도록 설계되었기 때문에 일반 텍스트에서 발생하는 현상보다 신뢰성 있는 결과를 보여준다.

4.2 형태소 범주에 따른 규칙 적용

한국어는 형태소의 범주에 따라 서로 다른 음소열로 발음열이 실현된다. 그림 1의 '신고'의 경우와 같이 어간과 어미의 결합인지 하나의 명사인지에 따라 다르게 발화된다. 여기서는 필수 음소 변동 규칙이 형태소의 범주에 따라 어간, 어미, 조사, 명사·부사·관형사(default), 복합어로 분리하여 수행된 결과를 분석하였다. 표 3은 규칙 적용 범위에 따라 분류된 음소 변동 규칙 오토마타를 참조하여 얻은 결과로, 명사 프로세스의 경우 입력 형태소 중 34.4%가 변동 규칙이 적용되어 다른 음소열로 변화하였다.

표 3. 형태소 범주 별로 적용된 음소 변동 규칙 분석

형태소 범주	입력 형태소 수	적용된 음소 변동 규칙 수		
		필수	수의	발생비율
명사	593,666	79,940	124,120	34.4%
어간	119,501	14,289	26,222	33.9%
어미	210,741	32,348	1,871	16.2%
조사	236,649	14,513	1,692	6.5%
복합어	40	39	5	110%
합계	1,160,597	141,129	153,910	25.4%

4.3 적용된 음소 변동 규칙의 통계적 분석

그림 2는 13개 범주의 필수 음소 변동 규칙이 적용된 결과를 분석한 것으로, 가로축은 적용된 음소 변동 규칙이며, 세로축은 형태소 내부와 형태소 경계에서 적용된 규칙의 발생 횟수이다. 분석 결과 가장 많이 적용된 규칙은 연음법칙이며 경음화, 격음화, 장애음의 비음화 순이다. 발생 빈도수를 기준으로 가장 많이 적용된 필수 음소 변동 세부 규칙은 형태소 내부의 경우 규칙 번호 4.4(연음규칙; 'ㄴ+ㅇ→Ø+ㄴ')이며, 형태소 경계의 경우 9.72(경음화; 'ㅆ+ㄷ→ㄷ+ㄸ')이다. 예를 들면 규칙 4.4는 '운영/ncn→/우녕/'과 같이 받침 'ㄴ'이 연음이 되어 다음 음절의 초성으로 이동하는 현상이며, 9.72는 형태소 경계에서의 적용된 경음화 규칙으로 '있/paa+다/ef→/읻/+/따/'의 경우를 들 수 있다.

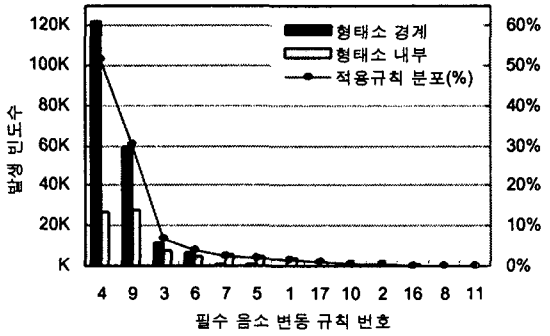

그림 2. 필수 음소 변동 규칙 발생 횟수 및 분포

그림 3은 7개 범주의 수의적 음소 변동 규칙이 적용된 결과를 분석한 것으로 필수 음소 변동 규칙보다 발생 빈도수가 비교적 적다. 수의적 음소 변동은 형태소 경계 정보에 따라 발화 현상이 달라지지는 않으나, 경계에 따라 발음사전에 기재되는 음소열이 변화하므로 분류하여 분석하였다. 다만 모음화 규칙 18, 19, 20은 음절의 중성 변화 규칙으로 형태소 경계에서는 발생하지 않는다.

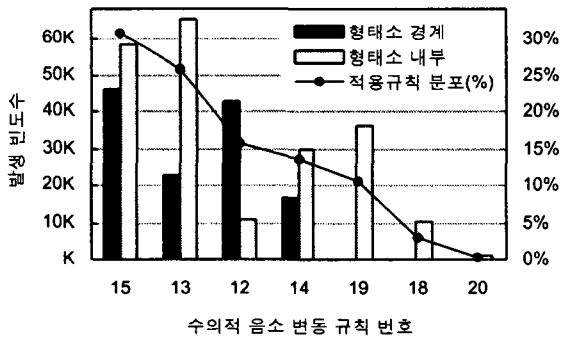

그림 3. 수의적 음소 변동 규칙 발생 횟수 및 분포

수의적 음소 변동 규칙인 '동일 조음위치 자음 탈락'과 '중복 자음화'는 발화 속도에 따른 발생 환경이 서로 상반되는 음운 변화 현상이다. '있/paa+다/ef'의 표준 발음은 /읻따/이지만 빠르게 말하는 경우 자음이 탈락하여 /이따/로 발화되며, '부터/jxc'의 경우 천천히 또박또박 발화하는 경우 중복 자음 'ㄷ'이 종성에 추가되어 /붇터/로 발화하거나 '아파트'가 /압파트/로 발화되기도 한다. 이와 같이 화자가 빠르게 발화하는 경우 '중복 자음화'가 잘 발생하지 않으며, 천천히 또박또박 발화하는 경우에는 '동일 조음위치 자음 탈락'에 의한 현상이 잘 나타나지 않는다[6]. 이러한 현상은 개발하는 음성 인식 시스템의 종류나 환경에 따라 조절해야 하는 요소로 필요에 따라 사용할 수 있도록 출력 선택 기능을 추가하였다.

분석 결과, 형태소 내부의 경우 관형격 조사 '의'가 /에/로 변화하는 규칙이 가장 많이 발생하였다. 실제 낭독체 문장을 대상으로 녹음할 때에는 임의로 '에'로 발성하도록 유도하였다. 그 다음은 규칙번호 15.1(초성 ㅎ-탈락; 'Ø+ㅎ→Ø+Ø')로 '화해/ncn→/화애/'나 '공개/ncpa+해/xsv→/공개애/'와 같이 약한 유성음 'ㅎ'이 탈락되는 규칙이 형태소 내부에서나 경계에서 빈번하게 적용되었다.

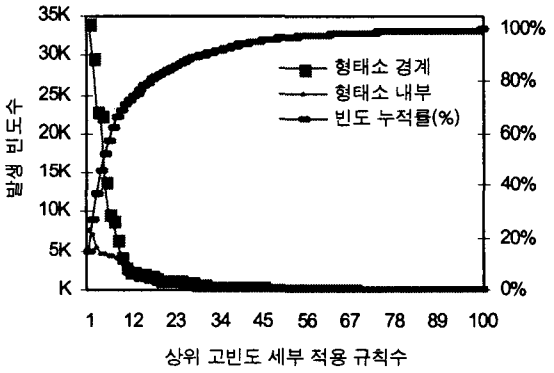

그림 4. 적용된 필수 음소 변동 규칙의 분포도

위의 그림 4는 적용된 필수 음소 변동 세부 규칙을 고빈도 순으로 정렬한 상위 100개의 규칙 분포도이다. 필수 음소 변동 세부 규칙 757개 중 192가지의 규칙이 삼성 PBS 60,000문장에서 적용되었으며, 총 289,169회의 변동 규칙이 발생하였다. 형태소 경계와 내부에서 모두 포함하여 1000번 이상 발생한 규칙은 상위 36번째 규칙까지이며, 100번 이상은 상위 82번째까지이다. 이 중 평균 상위 100개의 규칙으로 약 99.67%의 적중률을 보였다.

음성인식에서의 발음변이를 모델링하는 대표적인 방법으로는 발음사전을 사용하는 것이다. 표제어 내부에서 일어나는 음운 변화 현상은 발음사전에 등록하여 해결할 수 있으나 경계 부분에서 발생하는 변화 현상을 반영하기 위해 발음사전에 가능한 모든 발음을 등록하는 경우에는 표제어 수가 증가할 뿐만 아니라 인식 속도와 인식률에 나쁜 영향을 미치게 된다. 이를 해결하기 위한 방안으로 이 논문에서 소개된 분석 결과를 활용하여 빈번히 발생하는 음운 변화 현상만을 발음사전에 추가하여 활용할 수 있을 것이다.

또 다른 접근 방법으로는 인식 단위(표제어)의 경계 부분에서 일어날 수 있는 음운 변화 현상을 인식 네트워크에 적용시키는 방법이 있다. 인식 단위의 경계 부분에서 일어날 수 있는 모든 가능한 음소 문맥을 인식 전에 미리 인식 네트워크에 적용하는 방법으로 앞 표제어의 종성과 뒤의 초성의 쌍으로 나타낼 수 있는 모든 쌍에서 음운 변화 현상이 일어나는 것이 아니라 일정한 규칙에 따라 특정한 쌍에서만 일어나게 된다. 특히 형태소 내부와 형태소 경계에서 발생하는 현상이 다를 뿐만 아니라 음소 문맥에 따라 발생 가능한 네트워크만을 확장하는 것이 효율적이므로 이 논문에서 소개된 자료를 활용하여 인식기의 성능을 향상시킬 수 있다. [9]의 연구에서는 이러한 분석 자료를 이용하여 트리 구조의 인식 네트워크의 공유 효율을 높이고 이로 인해 네트워크의 크기를 줄일 수 있도록 인식 중에 음소 문맥을 이용해 인식 네트워크에 음운 변화 현상을 적용시키는 방법을 제안하였다.

V. 결론

정확한 발음열을 생성하기 위해 한국어가 가지는 언어학적 지식과 문교부 제정 표준어 규정을 기반으로 음운 변화 규칙을 분석하고, 이를 통해 정의된 음소 변동 규칙과 변이음 규칙을 다단계로 적용하여 가능한 모든 발음열을 생성하였다. 정의된 음소 변동 규칙들이 실제 적용되는 현상을 분석하기 위하여 트라이폰 기반의 PBS 60,000문장에 발음열 자동 생성기를 적용하여 나온 결과를 통계적으로 분석하였다. 분석 결과는 음소변동을 모델링 한 분류에 따른 빈도수와 음소의 경계 위치에 따른 적용양상에 대하여 초점을 맞추었다. 적용된 음소 변동 규칙들의 통계적 자료를 기반으로 한국어 음운 변화 현상 양상을 파악할 수 있었으며, 나아가 이러한 분석을 이용하여 음성 인식기의 성능을 향상시키기 위한 자료로 활용할 수 있을 것이다. 일반적으로는 가능한 모든 음운 변화 현상을 분석하여 모델링 하는 것이 정확한 음운변이를 반영할 수 있으나, 혼잡도 증가와 변별력 감소 문제 및 인식 네트워크 확장시 가능한 음소 문맥을 적용하는 경우 적용 규칙수가 필요 이상으로 많아지기 때문에 본 논문에서 통계적으로 분석된 음운 변화 현상을 사용함으로써 시스템 개발에 유용하게 사용할 수 있을 것이다.

감사의 글

본 실험에 사용한 삼성종합기술원의 PBS 음성 DB 사용허가에 감사 드립니다.

참고문헌

[1] 이경님, 전재훈, 정민화, 한국어 연속음성 인식을 위한 발음열 자동 생성, 한국음향학회지, 제20권, 제2호, pp. 35-43, 2001.

[2] B. Kim, W. Lee, G. Lee, J. Lee, Unlimited vocabulary grapheme-to-phoneme conversion for Korean TTS, Proc. of ACL-COLING 98, pp.675-679, 1998.

[3] H. Strik and C. Cucchiarini, Modeling Pronunciation Variation for ASR: Overview and Comparison of Methods, Proc. of the ESCA workshop 'Modeling pronunciation variation for automatic speech recognition', pp.137-144, 1998.

[4] 김홍규, 강범모, 한글 사용빈도의 분석, 고려대학교 민족문화연구소, 1997.

[5] 한국방송공사, 표준 한국어 발음 대사전, 1993.

[6] 이기문, 김진우, 이상억, 국어음운론, 학연사, 2000.

[7] J. Clark and C. Yallop, An Introduction to Phonetics and Phonology, Oxford, 1995.

[8] 표준어 규정, 문교부 고시 제88-2호, 1988.

[9] 김한준, 음소 문맥과 음운 변화 현상을 이용한 한국어 연속 음성 인식, 서강대학교 컴퓨터학과 석사 학위 논문, 2001.