운율 정보를 이용한 한국어 위치 정보 데이타의 발음 모델링

(Pronunciation Variation Modeling for Korean Point-of-Interest Data Using Prosodic Information)

김 선 희 † 박 전 규 †† 나 민 수 ††† 전 재 훈 †††† 정 민 화 †††††
(Sunhe Kim) (Jeongue Park) (Minsoo Na) (Jehun Jeon) (Minwha Chung)

요 약 본 논문은 두 가지의 구조적 운율 정보, 즉 운율어와 음절 수를 이용하여 한국어 위치 정보 데이타의 발음모델링을 수행할 경우에 음성인식기의 성능을 평가하는 것을 목표로 하는 것이다. 먼저, 위치 정보 데이타가 운율어로 구성되어 있다는 전제 하에 운율어를 이용하여 위치 정보 데이타의 가능한 모든 발음을 생성하고, 다시 음절 수를 기준으로 발음변이 수를 조절하는 방법을 제시하였다. 제안한 방법에 의하여 9개의 테스트 세트와 9개의 학습 세트로 총 81개의 실험을 통하여 음성인식의 성능을 평가하였다. 실험 결과 운율어를 이용하여 발음 사전을 제작한 모든 경우에 베이스라인과 비교하여 성능이 향상되었다. 음절 수에 따라서 발음 변이의 수를 조절한 결과도 전체적으로는 3음절로 그 수를 제한한 경우에 가장 좋은 인식 성능을 얻을 수 있어서, 음절 수에 따른 발음 변이 수의 조절이 효과적임을 알 수 있었다. 제안한 방법과 같이 운율어와 음절수를 이용한 경우에 베이스라인의 WER 4.63%에서 최대 8.4%의 WER가 감소하였다.

키워드 : 발음모델링, 발음 변이, 운율어, 음절, 위치정보데이타

Abstract This paper examines how the performance of an automatic speech recognizer was improved for Korean Point-of-Interest (POI) data by modeling pronunciation variation using structural prosodic information such as prosodic words and syllable length. First, multiple pronunciation variants are generated using prosodic words given that each POI word can be broken down into prosodic words. And the cross-prosodic-word variations were modeled considering the syllable length of word. A total of 81 experiments were conducted using 9 test sets (3 baseline and 6 proposed) on 9 trained sets (3 baseline, 6 proposed). The results show: (i) the performance was improved when the pronunciation lexica were generated using prosodic words; (ii) the best performance was achieved when the maximum number of variants was constrained to 3 based on the syllable length; and (iii) compared to the baseline word error rate (WER) of 4.63%, a maximum of 8.4% in WER reduction was achieved when both prosodic words and syllable length were considered.

Key words : Pronunciation modeling, Pronunciation variation, Prosodic word, Syllable, Point-of-Interest data

· 본 연구는 KT와 정보통신연구진흥원을 통한 정보통신부 선도기반기술 개발사업의 연구비 지원으로 수행하였습니다. 본 논문은 제18회 한국 및 한국어 정보처리 학술대회에서 발표한 논문[15,16]을 토대로 작성되었습니다.

† 정 회 원 : 서울대학교 인문정보연구소 교수
sunhkim@snu.ac.kr
†† 비 회 원 : 한국전자통신연구원
jgp@etri.re.kr
††† 비 회 원 : 서울대학교 인지과학협동과정
dix39@snu.ac.kr
†††† 비 회 원 : 서강대학교 컴퓨터학과
jhjeon@snu.ac.kr
††††† 비 회 원 : 서울대학교 언어학과 교수
mchung@snu.ac.kr

논문접수 : 2006년 8월 13일
심사완료 : 2006년 8월 31일

1. 서 론

일반적으로 음성인식에서 발음 모델링은 어휘부(Lexicon)와 음향 모델(Acoustic Model), 그리고 언어 모델(Language Model)의 세 영역에서 가능하다. 이 세 영역에서의 발음 모델링은 서로 독립적이기도 하지만 어느 정도는 상호 의존적인 관계를 가지고 있어서 실제 음성인식 시스템에서는 세 영역 모두에서의 발음 모델링이 수행할 때 가장 효과적이고 실질적인 성능 향상을 가져올 수 있을 것으로 알려져 있다[1].

어휘부에서의 발음 모델링은 발음사전(lexicon) 기반

모델링이라고 하여 가능한 발음들을 생성하여 이를 탐색 과정과 효율적으로 통합하는 것이 관건이 된다. 가능한 발음들을 생성하는 발음 생성 방법으로는 지식 기반 방식[2,3]과 데이타 기반 방식[4,5]이 있다. 지식 기반이란 음성학 음운론과 같은 언어학적 지식을 이용하여 발음을 생성해 내는 것을 의미하고, 데이타 기반 방식이란 데이타에서 발음 변이 현상을 자동으로 추출하는 것을 의미하는 것으로 음소 인식기(phone recognizer)나 결정 트리 기반 규칙, 혹은 표준 비터비 알고리듬을 이용하는 방법 등이 있다.

그런데, 음성 인식 시스템에서 가능한 발음들을 생성하여 발음 사전에 추가하는 경우에는 음향적 복잡도(acoustic confusability)를 증가시켜 새로운 오류를 야기할 수 있다. 따라서 적절하게 발음 변이 수를 조절하여 이러한 오류를 감소하는 방법들도 제안되었는데, 일반적으로 발음 변이 수를 조절하는 기준으로는 (i) 최고 유사도, (ii) 신뢰도, (iii) 변이음 사이의 복잡도 등이 이용된다[1].

본 논문에서 위치 정보(Point-of-Interest: POI) 데이타란 행정구역 및 지명, 인명, 상호명과 같은 위치 관련 어휘로 구성된 데이타로서, 이는 텔레메틱스를 비롯한 다른 지명 정보 관련 분야의 응용 프로그램의 개발에 필수적이다. POI의 발음 모델링의 문제는 먼저 POI데이타의 가능한 모든 발음을 생성하고, 이 생성된 모든 발음으로부터 다시 음향적 복잡도를 줄이기 위하여 발음 변이 수를 조절하여 발음 사전을 생성하여 수행될 수 있다. 일반적으로 운율 정보를 음성인식에 이용한 연구들에 있어서는 대부분 운율의 음향적 정보를 이용하는데 반하여[7,8], 본 연구에서는 운율어나 음절 수와 같은 운율의 구조적 정보를 이용하여 발음 모델링을 할 경우에 음성인식의 성능이 향상될 수 있음을 보이고자 한다.

본 논문은 먼저, 2장에서 위치 정보 데이타를 위하여 운율어 기반의 다중 발음 생성 시스템을 제안하고, 3장에서는 음절 수에 따라 발음 변이 수를 조절하여 생성한 발음사전을 제시한다. 4장에서는 실험 및 그 결과를 제시한 다음, 5장의 결론으로 마무리한다.

2. 위치 정보 데이타의 다중 발음 생성

POI 데이타는 모두 특정 장소를 나타내는 고유명사에 해당하고 새로운 상호명의 출현에 따라 많은 신조어를 포함하는 등, 일반적인 텍스트 데이타와는 다른 특성을 보인다. 구조적으로는 '전망좋은집'과 같이 수식어를 포함한 명사구의 형태를 보이는 경우도 있고, '홍도야울지마라'와 같이 하나 이상의 문장으로 이루어진 경우도 있으나 대부분의 경우는 '한국교육문화사'와 같이 명사들이 결합된 복합 명사구로 구성되어 있다.

POI의 다중 발음 생성은 두 가지의 언어학적 특징과 관련이 있다[6]. 첫째, POI는 모두 특정 장소를 나타내는 고유명사에 해당하는데, 고유 명사가 음성합성이나 음성인식을 위한 발음 생성에 있어서 문제가 된다는 것은 여러 다른 언어에서도 지적된 바 있다[9]. 인도 유럽어의 고유 명사의 문제는 언어들 사이의 공유하는 알파벳에 기인하는데 반하여, 한국어의 경우는 불규칙 발음과 관련이 있다. [10]에 의하면 53,750 문장으로 이루어진 한국어의 일반 텍스트 코퍼스에서 불규칙 발음은 6.67%가 포함되어 있는 것으로 보고되었다. 대부분의 한국어 불규칙 발음이 명사에서 나타나므로, 많은 고유 명사와 복합명사구로 구성되어 있는 위치 정보 데이타는 텍스트 코퍼스와 비교할 때 그보다 많은 불규칙 발음이 포함되어 있을 것으로 예상된다.

둘째로, POI는 대부분은 두 개 이상의 명사가 결합된 복합 명사구로 구성되어 있어서, 하나의 복합명사구를 구성하고 있는 명사가 다른 많은 복합 명사구에도 나타난다. 예를 들면 '한국교육문화사'의 경우, 그것을 구성하고 있는 각각의 명사 '한국', '교육', '문화사'는 웹상에서 수집한25만 개의 POI에서 각각 3,538회, 523회, 307회 출현하다. 따라서, 불규칙 발음을 포함하는 명사가 다른 여러 개의 복합명사구를 형성한다면, 이러한 불규칙 발음 복합 명사구를 찾기 위해서는 그 하위 구성요소인 불규칙 발음을 포함하는 단어들을 찾는 것이 선행되어야 한다. 또 한편으로는, 2개 이상의 명사가 결합되는 경우에는 그 구성 요소인 단어들이 결합할 때 여러 다른 발음 변이가 관찰되므로 이러한 구성요소 경계에서 발음 변이 현상을 생성해 내기 위해서는 마찬가지로 복합명사구의 하위 구성요소를 찾아내야 한다.

복합명사구의 하위 구성요소를 추출해 내기 위해서 1차적으로 형태소 분석기를 사용하는 것을 생각해 볼 수 있는데, 이 경우에는 최소한 두 가지의 문제가 예상된다. 먼저, 대부분의 복합명사구는 두 개 이상의 명사가 결합되어 있어서 형태소 분석 과정은 결국 명사를 찾아내어 분석해 내는 과정이 되는데, 위에서 언급한 대로 위치 정보 데이타는 많은 고유 명사와 신조어를 포함하고 있어서 기존의 형태소 분석기로는 제대로 데이타를 처리하기가 어려울 것이다. 뿐만 아니라, 한국어의 경우에 실제로 복합명사나 합성어에서 불규칙 발음이 관찰되는데, 이러한 복합명사가 형태소 분석기에 의하여 각각의 명사로 분할되게 되면 불규칙 발음을 포함하는 단어를 제대로 추출해 낼 수 없게 된다.

따라서, 복합명사구로 이루어진 POI의 가능한 모든 발음을 생성해 내기 위해서는 위에서 언급한 두 가지 문제와 관련하여 복합명사구를 구성하고 있는 요소로 명사가 아닌 다른 단위에 의한 분할이 필요하게 되는데,

[6]은 [11]에서 제안된 운율어(Prosodic word)의 개념을 이용하여 불규칙 발음을 추출하는 방법을 제안하였다. [11]에 의하면 운율어란 "강세구로 실현될 수 있는 분절음의 최소 연쇄(the minimal sequence of segments which can be produced as one Accentual Phrase (AP)"라고 정의하였는데, 이는 바꾸어 발하면 운율어란 따로 끊어서 발화할 수 있는 분절음의 최소 연쇄가 된다.

그림 1은 위와 같이 정의된 운율어를 기반으로 위치 정보 데이타의 다중 발음을 생성하기 위한 방법을 나타낸 것이다. 이러한 방법은 새로운 데이타가 추가될 때마다 반복되어 실행될 수 있도록 설계된 것으로, 전체 과정은 데이타나 시스템을 이용하는 과정인 (A), (B), (C) 세 부분과 수동 처리 부분(Manual Review)으로 나뉘어져 있다.

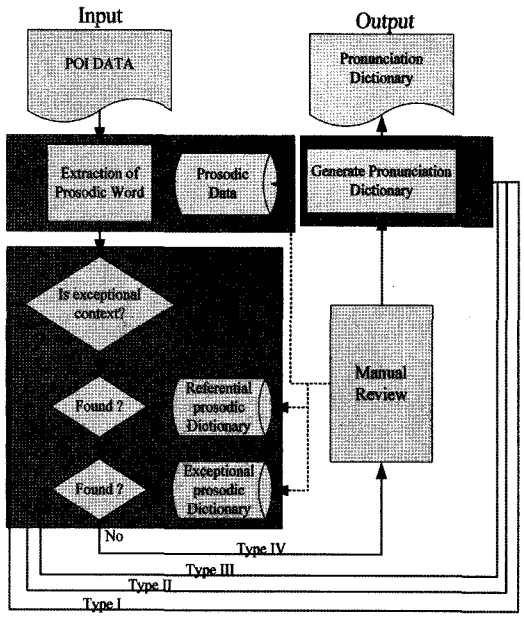

그림 1 위치 정보 데이타의 다중 발음열 생성 과정: (A) 입력된 위치 정보 데이타를 운율어로 분할; (B) 정의된 사전을 이용하여 불규칙 발음 운율어 검출; (C) 문자음성변환기를 이용하여 다중 발음 생성

(A) 과정은 입력된 위치 정보 데이타를 운율어로 분할하는 과정이다. 초기 운율어 사전은 수작업으로 만들어서 사용하게 되는데, 이때 각각의 운율어는 원 데이타에서의 위치 및 빈도 정보와 함께 수록되고, 수록된 운율어는 다시 어두 운율어 사전(Dic-f), 어중 운율어 사전(Dic_m), 어말 운율어 사전(Dic-r)으로 분리된다. 입력된 데이타를 운율어로 분할하는 알고리즘은 다음과 같다.

1. 입력된 POI에 대해서 Dict_f를 이용해 어두에 사용 가능한 운율어 목록(fw_i)을 검출하고, Dict_r을 이용해 어미에 사용 가능한 운율어 목록(rw_i)을 검출한 다음, 각 리스트에서 $i=0$을 기준으로 운율어로 나눌지 결정하는 검색을 시작한다. (단 음절 길이가 가장 긴 단어가 리스트에서 0번째 기록되고, 음절 길이가 긴 순서대로 위치시킨다.)
2. fw_i와 rw_i의 단어 경계가 일치하면 출력 List에 기록하고 단계 8로 간다.
3. 각 단어에서 i가 기록된 리스트 이상이면 탐색을 중단하고 단계 7로 간다.
4. fw_i와 rw_i의 단어 경계가 서로 Cross 되어 있으면 $D(fw_i, fw_{i+1})$, $D(rw_i, rw_{i+1})$을 이용해 그 값이 작은 단어의 첨자를 $i=i+1$한 다음 2단계로 돌아 간다.
5. fw_i와 rw_i의 단어 경계가 서로 만나지 않으면 $D(fw_i, fw_{i+1})$, $D(rw_i, rw_{i+1})$을 이용해 그 값이 큰 단어를 출력 List에 기록한다.
6. 전체 POI의 나머지 부분에 대해서 Dict_m을 이용해 가능한 운율어 목록을 다시 구성한다. 단 앞 단계에서 fw_i가 선택 된 경우 신규 리스트 이름을 fw_i로, rw_i가 선택된 경우 이름을 rw_i로 하고 $i=0$로 한 다음 2단계로 돌아 간다.
7. 입력 POI가 기존에 정의 되지 않은 운율어를 포함한 것으로 간주하고, 운율어로 더 이상 분할하지 않고 출력 종료 한다.
8. 출력 목록을 운율어 목록으로 출력 하고 종료 한다.

* $F(w_i)=\sqrt{freq(w_i)}\times len(w_i)$ (검색된 단어 w_i가 하나의 운율어로 나누어질 Measure 함수. $freq(w_i)$는 w_i가 사전 작성시 사용된 빈도 수, $len(w_i)$는 w_i의 음절 길이)

* $D(w_i, w_{i-1})=\dfrac{F(w_i)-F(w_{i+1})}{len(w_i)-len(w_{i+1})}$ (긴 단어 w_i를 포기하고 다음 단어 w_{i+1}를 선택할 함수)

(B)는 [10,12]가 제안한 불규칙 발음 검출 방법을 위치 정보 데이타에 적용하여 불규칙 발음 운율어를 검출하는 과정이다. [10]은 일반적으로 예외로 처리되어 사전에 무작위로 추가되는 불규칙 발음이 일정한 환경에서 관찰되는 3가지의 음운 현상이라고 규명하고, 이러한 연구 결과에 따라 주어진 데이타로부터 불규칙 발음을 검출하는 방법을 제안했다. [10,12]에 의한 불규칙 발음과 관련된 음운 현상과 그 환경은 다음과 같다.

표 1 불규칙 발음 음운 현상과 그 환경

(C: a consonant of a consonant cluster; V: a vowel or diphthong)
(I) Lexical Tensification
(II) Nasalization of the lateral
(III) /n/-Epenthesis, Neutralization/ Simplification + Liaison

	p	t	s	c	k	l	V
m							
n						(II)	
N			(I)				
l							
V							
C							(III)

(A) 과정에서 분할된 운율어는 [6]의 제안과 같이 (B)과정에서 먼저 이 운율어가 불규칙 발음이 나타나는 음운 환경에 해당하는지 아닌지 여부에 따라 먼저 주어진 운율어가 불규칙 발음 환경에 속하는 않는 경우는 Type I로 분류된다. 불규칙 음운 환경에 속하는 운율어인 경우는 다시 정의된 사전에 포함되어 있는지의 여부에 따라 Type II와 Type III로 나뉜다. Type II는 운율어 가운데 불규칙 음운 환경에 속하지만 규칙적인 음운 현상을 보이는 운율어들(여기에서는 '참조 운율어'라고 지칭함)이고, Type III는 불규칙 발음 운율어이다. 마지막으로 불규칙 음운 환경에 속하지만 참조 운율어도 불규칙 운율어도 아닌 운율어들은 Type IV로 분류되어 전문가에 의한 수작업으로 Type III나 Type IV로 분류되게 된다(Manual Review). (A)와 (B) 과정에서 새로 검출된 모든 종류의 운율어들은 다음 번 작업 이전에 각각 업테이트하게 된다.

이와 같이 (A)와 (B) 및 수동 분류를 통하여 운율어로 분할된 POI는 최종적으로 (C)과정에서 문자음성변환기를 이용하여 가능한 모든 발음을 생성하게 된다. 여기에서 채용하는 문자음성변환기는 지식기반 시스템으로서, 발음을 생성하기 위한 음소변동 규칙으로는 다음과 같은 10개의 규칙으로 이루어져 있다[12-14]: (1) 종성중화, (2) 자음군 단음화, (3) 장애음 뒤의 경음화, (4) 격음화, (5) 장애음의 비음 동화, (6) 유음화, (7) 이중 비음 동화(장애음의 비음화+/ㄹ/의 비음화), (8) 연음(재음절화), (9) 구개음화, (10) 단모음화.

제안한 방법에 따르면 '한국교육문화사'란 단어는 (1)과 같이 3개의 운율어로 분석할 수 있고, 이와 같이 분석될 때 이 단어는 다음의 4가지의 발음으로 실현될 수 있다. (2)는 POI 가운데 불규칙 발음 운율어를 포함하는 경우인 '즉석김밥나라'의 가능한 모든 발음을 도출한 예이다. ('{}'는 운율어 경계를 표시함)

(1) /한국교육문화사/

Sub process	Derived forms and rules
(A)	{hankuk} {kyoyuk} {munhwasa}
(B)	해당 사항 없음
(C)	{hankuk} {kyoyuk} {munhwasa}
	{hankuk' yoyuk} {munhwasa}
	융합
	{hankuk} {kyoyuŋmunhwasa}
	비음화
	{hankuk'yoyuŋmunhwasa}
	융합 + 비음화

(2) /즉석김밥나라/

Sub process	Derived forms and rules
(A)	{cIksʌk} {kimpap} {nala}
(B)	{cIksʌk} {kimp'ap} {nala}
	어휘적경음화
(C)	{cIks'ʌk} {kimp'ap} {nala}
	경음화
	{cIs' ʌk' imp'ap} {nala}
	융합
	{cIs' ʌk} {kimp'amnala}
	비음화
	{cIs' ʌk'imp'amnala}
	융합+비음화

3. 운율 정보를 이용한 발음 사전 생성

이와 같이 가능한 모든 발음 변이를 생성하기 위하여 POI를 운율어로 분할한 다음, 운율어의 내부에서는 음운 규칙을 필수적으로 적용하고, 운율어의 경계에 음운 규칙을 수의적으로 적용하여 가능한 모든 발음 변이를 생성해 내었다. 따라서, POI의 길이가 길수록(혹은 음절수가 많을수록) 포함되어 있는 운율어의 수는 많아지는데, 이렇게 구성 운율어의 수가 많아지면 운율어 경계에서 수의적으로 적용되는 음운규칙이 많아지고, 결과적으로 많은 발음 변이가 생성되게 된다. 위에서 이미 지적한 대로 이렇게 가능한 모든 발음을 생성하게 되면, 음성 인식 시스템에서 음향적 복잡도가 높아져 오류가 증가할 수 있는데, 여기에서는 이러한 오류를 줄이기 위하여 음절수를 기준으로 발음 변이의 수를 조절하는 방법을 제시한다.

베이스라인 시스템은 운율어를 이용하지 않은 발음사

표 2 음절 수에 따라 분류한 POI의 운율어의 수와 발음 변이의 수(pw: prosodic word, pro: pronunciations)

syllable length	# of words	average pw	average pro
1	12	1	1
2	1326	1	1.1
3	4661	1	1.2
4	2229	1.8	1.3
5	2019	2	1.4
6	1032	2.5	2
7	1131	2.8	1.9
8	434	3.1	3
9	236	3.6	3
10	137	4	4.7
11	57	4.5	4.2
12	33	5.1	5
13	16	5.6	11.5
14	7	6.3	6.3
15	1	7	16
16	1	6	2
18	1	6	33

전으로서, 이는 다시 표준 발음 사전(bDic_1best), 가능한 모든 발음 사전(bDic-all), N-Best 발음(bDic_Nbest) 사전의 3개로 구성된다.

운율 정보를 기반으로 제안한 방법에 의한 성능을 평가하기 위하여 음절 수와 운율어 수를 기준으로 분류한 250k POI를 근거로 하여 6개의 발음 사전을 설계하였다. 표 2에 의하면 POI를 구성하는 운율어 수와 생성된 발음 변이의 수가 음절 수에 비례하여 많아지는데, 특히 발음 변이는 음절 수가 6, 8, 10일 때 각각 증가하는 폭이 큰 것을 볼 수 있다. 제안하는 사전은 먼저 베이스라인과 마찬가지로 표준 발음 사전(Dic_1best), 가능한 모든 발음 사전(Dic-all), N-Best 발음(Dic_Nbest) 사전을 운율어를 이용하여 생성하고, 다음으로는 표 2의 결과에 따라서 5음절 이하인 경우는 발음 변이의 수를 1개로, 6음절 이상인 경우에는 발음 변이의 수를 각각 2개(Dic_2), 3개(Dic-3), 4개(Dic_4)로 제한한 3개의 사전을 설계하였다. 표 3은 제안하는 6개의 발음 사전을 음절 수에 따른 각각 발음 변이 수가 어떻게 조절되었는지를 보여 주는 표이다.

4. 실험 및 결과

4.1 실험 환경

학습을 위해서는 SiTEC과 ETRI에서 각각 독자적으로 제작된 103,082 발화로 구성된 2개의 음성 코퍼스를 사용하였다. SITEC에서 제작된 음성 코퍼스는 저속 주행환경(30~60 Km/h)과 고속 주행환경(70~90 Km/h)에서 장착된 마이크(AKG C400-BL)와 헤드셋(Shure

표 3 음절 수를 기준으로 설계한 발음 사전과 그 발음 변이 수(Dic_1: N-best; Dic_1best: 1-best; Dic_all: all possible pronunciations; Dic_2: maximum number limited to 2; Dic_3: maximum number limited to 3; Dic_4: maximum number limited to 4)

syllable length	Dic_1	Dic_ 1 best	Dic_ all	Dic_2	Dic_3	Dic_4
1	1	1		1	1	1
2	1	1		1	1	1
3	1	1		1	1	1
4	1	1		1	1	1
5	1	1		1	1	1
6	2	1		2	2	2
7	2	1		2	2	2
8	3	1		2	2	3
9	3	1		2	2	3
10	4	1		2	3	3
11	4	1		2	3	3
12	5	1		2	3	4
13	5	1		2	3	4
14	5	1		2	3	4
15	5	1		2	3	4
16	2	1		2	3	4
18	5	1		2	3	4

SM-10A)을 이용하여190명으로부터 녹음한 8,516발화로 구성되었다. ETRI에서 제작된 음성 코퍼스는 94,566 발화의 텔레메틱스 코퍼스로서 여러 다른 주행 환경에서 장착된 마이크(AKG C400-BL)와 헤드셋(Altec Lansing AHS302)을 이용하여 녹음한 것이다. 테스트용 음성 코퍼스로는 학습용 코퍼스 녹음에 참여하지 않은 38명으로부터 4,149개의 녹음된 발화를 이용하였다.

음성 신호는 25mm 해밍 윈도우를 이용하여 2단계 위너 필터를 통하여 처리하였다. 학습과 테스트에는 모두 30차 계수(13차 MFCC, 13차 델타 MFCC, 13차 델타 델타 MFCC 가운데 처음의 4차 계수)를 특징 벡터로 추출하였다.

각각의 발음 사전을 위하여 음향 모델 학습은 16개 믹스쳐 트라이폰, 30개 특징 백터의 특징 스트림으로, 다음의 11 단계에 따라 수행되었다.

- STEP 1: Uniform segmentation for monophone (Context Independent (CI) phone) seed model
- STEP 2: Monophone model training of mixture 1 with Viterbi Baum-Welch algorithm
- STEP 3: Forced alignment for multiple pronunciation
- STEP 4: Monophone model training of mixture 3 with Viterbi Baum-Welch algorithm, mixture = 3

- STEP 5: Triphone (Context Dependent (CD) phone) cloning from 1 mixture monophone
- STEP 8 : Triphone seed model training with given alignment file
- STEP 9 : Triphone model training with Viterbi Baum-Welch algorithm
- STEP 10 : Triphone tying
- STEP 11 : Tied-triphone model training of mixture 2 to 16 with Viterbi Baum-Welch training

4.2 실험 결과

3개의 베이스라인 사전과 6개의 제안한 사전을 각각 학습과 테스트에 모두 이용하여 총 81개 실험을 수행하였고, 그 결과는 다음 표 3과 같다. 가장 좋은 결과는 음절 수를 2개와 3개로 제한하여 제작한 발음 사전 Dic_2, Dic_3이었고(4.24 of WER), 가장 나쁜 결과는 베이스라인에서 표준 발음 사전을 이용하였을 때이다 (4.63 of WER). 베이스라인에 비하여 제안한 방법으로 최대 8.4%로 WER가 감소하였다.

위에서 언급한 대로 베이스라인과 제안한 사전의 차이는 기본적으로 운율어의 사용 여부에 있는데, 표 4에서 보는 바와 같이 운율어를 이용하여 발음 사전을 생성한 경우가 그렇지 않은 경우에 비하여 인식 성능이 향상 된 것으로 나타났다. 또한, 이러한 인식 성능의 향상은 6음절 이상인 경우에 더 명확하게 나타난 것을 볼 수 있었다. 다음 그림 2는 각 음절 수에 따라 발음 변이 수를 조절한 결과를 특별히 잘 보여주고 있는데, 최대

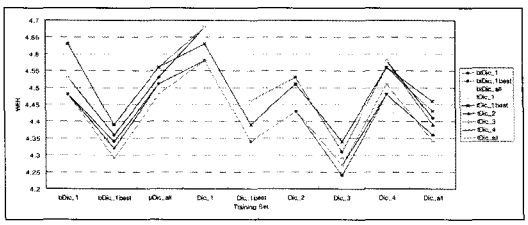

그림 2 전체 POI에 대한 WER

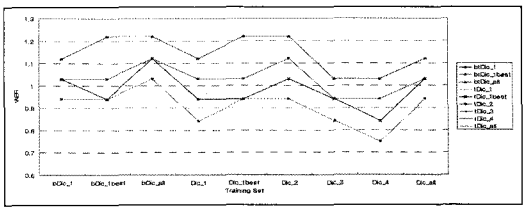

그림 3 6음절 이상 POI에 대한 WER

발음 변이 수를 3개로 제한하여 실험한 경우(DIc_3)에 가장 좋은 성능을 보이는 것을 볼 수 있다. 그림 3은 6음절 이상 단어들만 따로 실험한 결과를 나타내는데, 이 경우에는 음절 수를 4로 제한한 경우가 가장 그 성능이 좋은 것으로 나타났다.

5. 결 론

본 논문은 두 가지 운율의 구조적 특성, 즉 운율어와 음절 수를 이용하여 발음모델링할 경우에 음성인식기의 성능을 평가하는 것을 목표로 하는 것으로, 먼저, 운율어를 이용하여 POI데이타의 가능한 모든 발음을 생성하고, 음절 수를 기준으로 발음변이 수를 조절하는 방법을 제시한 다음, 제안한 방법에 의하여 생성한 발음사전을 이용하여 음성인식의 성능을 평가하였다. 실험 결과 운율어를 이용하여 발음 사전을 제작한 모든 경우에 베이스라인과 비교하여 성능이 향상됨을 보였는데, 베이스라인 WER 4.63 에서 최대 8.4% WER 가 감소하였다. POI의 음절 수에 따라서 발음 변이를 조절한 결과도 전체적으로 3음절로 제한한 경우, 6음절 이상 단어에서는 4음절로 제한한 경우에 가장 좋은 인식 성능을 얻을 수 있어서, 음절 수에 따른 발음 변이 수의 조절이 효과적임을 알 수 있었다.

이러한 연구는 우선 POI 데이타를 이용한 음성인식 시스템의 성능 향상에 기여하였음을 보였을 뿐만 아니라, 일반적으로 운율 정보를 음성인식에 이용한 연구들에 있어서는 대부분 운율의 음향적 정보를 이용하는데 반하여, 본 연구에서는 운율어나 음절 수와 같은 운율의

표 4 WER(Word Error Rate)로 나타난 81개의 실험 결과

	btDic_1	btDic_1best	btDic_all	tDic_1	tDic_1best	tDic_2	tDic_3	tDic_4	tDic_all
bDic_1	4.53	4.63	4.53	4.48	4.63	4.48	4.48	4.48	4.48
bDic_1best	4.36	4.39	4.34	4.32	4.39	4.34	4.32	4.32	4.29
bDic_all	4.53	4.56	4.52	4.51	4.56	4.51	4.51	4.51	4.48
Dic_1	4.68	4.68	4.68	4.58	4.63	4.58	4.58	4.58	4.58
Dic_1best	4.46	4.46	4.46	4.34	4.39	4.34	4.34	4.34	4.34
Dic_2	4.52	4.53	4.51	4.43	4.51	4.43	4.43	4.43	4.43
Dic_3	4.29	4.31	4.29	4.27	4.34	4.24	4.24	4.27	4.27
Dic_4	4.58	4.56	4.58	4.48	4.56	4.48	4.48	4.48	4.52
Dic_all	4.41	4.43	4.39	4.36	4.46	4.36	4.36	4.36	4.34

구조적 정보를 이용하여 인식률 향상을 보인 것으로, 언어학적 이론을 음성 인식과 접목시킨 학제적 연구로서도 그 의미가 있다고 하겠다. 향후 본 연구에서 제시한 방법론을 다른 언어에도 적용할 수 있을 것으로 기대된다.

참 고 문 헌

[1] H. Strik and C. Cucchiarini, "Modeling Pronunciation Variation for ASR: A Survey of the Literature," *Speech Communication* 29, pp. 225-246, 1999.

[2] J. M. Kessens, M. Wester, H. Strik, "Improving the performance of Dutch CSR by modeling within word and corss-word pronunciation variation," *Speech Communication* 29, pp. 193-207, 1999.

[3] J. H. Jeon, S. Wee, M, Chung, "Generating Pronunciation Dictionary by Analyzing Phonological Variations Frequently Found in Spoken Korean," *Proc. of International Conference on Speech Processing*, pp. 519-523, 1997.

[4] E. Fosler-Lussier, "Multi-level decision trees for static and dynamic pronunciation models," *Proc. Eurospeech 1999*, 1999.

[5] M. Riley,W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, M. Saraclar, C. Wooters, G. Zavaliagkos, "Stochastic pronunciation modeling from hand-labeled phonetic corpora," *Speech Communication* 29, pp. 209-224, 1999.

[6] S. Kim, J. H., Jeon, M. Na, M. Chung, "Irregular Pronunciation Detection for Korean Point-of-Interest Data Using Prosodic Word," *말소리*, 제57권, pp. 123-137, 2006.

[7] K. Hirose and K. Iwano, "Detection of prosodic word boundaries by statistical modeling of mora transitions of fundamental frequency contours and its use for continuous speech recognition," *Proc. IEEE International Conference on Acoustics Speech & Signal Processing*, Vol.3 pp. 1763-1766, 2000.

[8] [Prosody-2001] Prosody in Speech Recognition and Understanding, ISCA Tutorial and Research Workshop (ITRW), Molly Pitcher Inn, Red Bank, NJ, USA, October 22-24, 2001, ISCA Archive, http://www.isca-speech.org/archive/prosody_2001.

[9] A. Sethy, S. Narayanan, S. Parthasarthy, "A Syllable Based Approach for Improved Recognition of Spoken Names," *Proc. ISCA Tutorial and Research Workshop, PMLA*, pp. 33-35, 2002.

[10] S. Kim, "Phonology of Exceptions for Korean Grapheme-to-Phoneme Conversion," *Proc. Interspeech 2004-ICSLP*, pp. 1285-1288, 2004.

[11] S.-A. Jun, *The Phonetics and Phonology of Korean Prosody: Intonational Phonology and Prosodic Structure*, Garland Publishing Inc., New York : NY., 1996.

[12] S. Kim, J. Ahn, S.-H. Kim, Y.-H. Lee, "A Korean Grapheme-to-Phoneme Conversion System Using Selection Procedure for Exceptions," *Proc. Interspeech 2004-ICSLP*, pp. 1905-1908, 2004.

[13] J. H. Jeon, S. Cha, M. Chung,, J. Park, "Automatic Generation of Korean Pronunciation Variants by Multistage Applications of Phonological Rules," *Proc. of the International Conference on Spoken Language Processing*, pp. 1943-1946, 1998.

[14] J. H. Jeon and M. Chung, "Automatic Generation of Domain-Dependent Pronunciation Lexicon with Data-Driven Rules and Rule Adaptation," *Proc. Interspeech-2005*, pp. 1337-1340, 2005.

[15] 김선희, 박전규, 나민수, 전재훈, 정민화, "운율 정보를 이용한 한국어 위치 정보 데이타의 발음 모델링", *제18회 한글 및 한국어 정보처리 학술대회 논문집*, pp. 51-56, 2006.

[16] 김선희, 전재훈, 나민수, 정민화, "운율어를 이용한 한국어 위치 정보 데이타의 다중 발음 사전 생성", *제18회 한글 및 한국어 정보처리 학술대회 논문집*, pp. 183-188, 2006.

김 선 희

1985년 연세대학교 불어불문학과(학사) 1986년 파리7대학 언어학과(석사). 1990년 프랑스 고등사회과학대학원 언어학과(박사). 1991년~2001년 연세대학교 시간강사. 2000년~2001년 L&H Korea 책임연구원. 2002년 3월~2004년 8월 광운대학교 연구교수. 2004년 9월~2005년 8월 한국과학기술원 연구교수. 2005년 9월~현재 서울대학교 연구교수. 관심분야는 언어학(음성학, 음운론), 음성언어처리

박 전 규

1987년 한국외국어대학교(학사). 1989년 한국외국어대학교 전산과(석사). 1991년~1999년 한국전자통신연구원. 2000년~2001년 L&H Korea. 2001년~2004년 동아시테크㈜. 2004년~현재 한국전자통신연구원. 관심분야는 음성언어처리, 자연어처리

나 민 수

2004년 한동대학교 전산전자공학부(학사). 2006년 서울대학교 인지과학 협동과정(석사). 2006년~현재 서울대학교 인지과학 협동과정 박사과정. 관심분야는 음성언어처리, 자연어처리

전 재 훈

1996년 서강대 컴퓨터학과(학사). 1998년 서강대 컴퓨터학과(석사). 1998년 3월~2003년 5월 삼성전자 선임연구원. 2004년 2월~현재 서강대 컴퓨터학과 박사과정. 관심분야는 음성언어처리, 자연어처리

정 민 화

1984년 서울대학교 제어계측공학과(학사). 1988년 University of Southern California 전기공학과(석사). 1993년 University of Southern California 전기공학과(박사). 1993년 12월~1994년 7월 한국통신 연구개발원 선임연구원. 1994년 9월~2004년 8월 서강대학교 컴퓨터학과 부교수. 2004년 9월~현재 서울대학교 언어학과 부교수. 관심분야는 음성언어처리, 자연어처리