운율 정보를 이용한
한국어 위치 정보 데이터의 발음 모델링

김선희*, 박전규**, 전재훈*, 나민수***, 정민화****
*서울대학교 인문정보연구소, **한국전자통신연구원 음성/언어정보연구센터,
***서울대학교 인지과학협동과정, ****서울대학교 언어학과
sunhkim@snu.ac.kr, jgp@etri.re.kr, dix39@snu.ac.kr, jhjeon@snu.ac.kr, mchung@snu.ac.kr

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

Sunhee Kim*, Jeon-Gue Park**, Je Hun Jeon*, Minsoo Na***, Minhwa Chung****
* Center for Humanities and Information, SNU
** Speech/Language Information Research Center, ETRI
*** Interdisciplinary Program in Cognitive Science, SNU
**** Department of Linguistics, SNU

요 약

일반적으로 운율 정보를 음성인식에 이용한 연구들에 있어서는 대부분 운율의 음향적 정보를 이용하는데 반하여, 본 연구에서는 운율어나 음절 수와 같은 운율의 구조적 정보가 인식률 향상에 기여함을 보인다. 본 논문은 두 가지 운율 정보, 즉 운율어와 음절 수를 이용하여 발음모델링을 할 경우에 음성인식기의 성능을 평가하는 것을 목표로 하는 것으로, 먼저, 운율어를 이용하여 위치 정보 데이터의 가능한 모든 발음을 생성하고, 다시 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, 베이스라인의 WER 4.63% 에서 최대 8.4%의 WER 가 감소하였다. 위치 정보 데이터의 음절 수에 따라서 발음 변이의 수를 조절한 결과도 전체적으로는 3 음절로 그 수를 제한한 경우, 6 음절 이상 단어에서는 4 음절로 제한한 경우에 가장 좋은 인식 성능을 얻을 수 있어서, 음절 수에 따른 발음 변이 수의 조절이 효과적임을 알 수 있었다.

1. 서론

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 데이터 기반 방식[4, 5] 이 있다. 지식 기반이란 음성학 음운론과 같은 언어학적 지식을 이용하여 발음을 생성해 내는 것을 의미하고, 데이터 기반 방식이란 데이터에서 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, 가능한 발음들을 생성하는 발음 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 데이터(Point-of-Interest: POI)란 행정구역 및 지명, 인명, 상호명과 같은 위치 관련 어휘로 구성된 데이터로서, 이는 텔레메틱스를 비롯한 다른 지명 정보 관련 분야의 응용 프로그램의 개발에 필수적이다. POI 의 발음 모델링의 문제는 운율어를 이용하여 POI 데이터의 가능한 모든 발음을 생성하고[6], 다시 음향적 복잡도를 줄이기 위하여 발음변이 수를 조절하여 발음 사전을 생성하는 것이다. 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 성능 향상의 문제를 다룬다.

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

2. 위치 정보 데이터의 다중 발음 생성

본 장에서 [6]에서 제안된 POI 의 다중 발음 생성 방법을 간략하게 소개한다. POI 의 다중 발음 생성은 그 두 가지 언어학적 특징과 관련이 있다. 첫째, POI 는 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.

둘째로, POI 는 대부분은 두 개 이상의 명사가 결합된 복합 명사구로 구성되어 있어서, 하나의 복합명사구를 구성하고 있는 명사가 다른 많은 복합 명사구에도 나타난다. 불규칙 발음을 포함하는 명사가 다른 여러 개의 복합명사구를 형성한다면, 이 불규칙 발음을 포함하는 단어들을 찾아 내야 할 필요가 있다. 또 한편으로는, 2 개 이상의 명사가 결합되는 경우에는 그 구성 요소인 단어들이 결합할 때 여러 다른 발음 변이가 관찰되므로 이러한 구성요소 경계에서 발음 변이 현상을 생성해 낼 필요가 있게 된다.

이렇게 복합명사구로 이루어진 POI 의 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 [10]에서 제안된 운율어(Prosodic word)의 개념을 이용하여 가능한 발음 변이 현상을 생성해 낸다. [10]에 의하면 운율어란 "강세구로 실현될 수 있는 분절음의 최소 연쇄(the minimal sequence of segments which can be prod uced as one Accentual Phrase (AP)" 라고 정의하였는데, 이는 바꾸어 발하면 운율어란 따로 끊어서 발화할 수 있는 분절음의 최소 연쇄가 된다.

[6]에서 POI 데이터의 다중 발음 생성 과정은 3 가지의 세부 과정으로 나뉜다: (A) 입력된 POI 를 운율어로 분할, (B) 정의된 사전을 이용하여 불규칙 발음 운율어 검출, (C) 문자음성변환기를 이용하여 다중 발음 생성. 제안한 방법에 따르면 '한국교육문화사' 란 단어는 다음과 같이 3 개의 운율어로 분석할 수 있고, 이와 같이 분석될 때 이 단어는 다음의 4 가지의 발음으로 실현될 수 있다.

(1) /한국교육문화사/: 한국-교육-문화사('-'는 운율어 경계를, 띄어쓰기는 끊어 읽기, 즉 강세구의 경계를 표시함)

a. 한국 교육 문화사 [hankuk kyoyuk munhwasa]
b. 한국교육 문화사 [hankuk' yoyuk munhwasa]
c. 한국 교육문화사 [hankuk kyoyu ŋ munhwasa]
d. 한국교육문화사 [hankuk' yoyu ŋ munhwasa]

아래는 POI 가운데 불규칙 발음 운율어를 포함하는 경우에 가능한 모든 발음을 도출한 예이다.

(2) /즉석김밥나라/: 즉석-김밥-나라

a. 즉썩 김빱 나라 [cIs' ek kimp' ap nala]
b. 즉썩낌빱 나라 [cIs' ek' imp' ap nala]
c. 즉썩 김빰나라 [cIs' ek kimp' amnala]
d. 즉썩낌빰나라 [cIs' ek' imp' amnala]

3. 운율 정보를 이용한 발음 사전 생성

가능한 모든 발음 변이를 생성하기 위하여 [6]은 POI 를 운율어로 분할한 다음, 운율어의 내부에서는 음운 규칙을 필수적으로 적용하고, 운율어의 경계에 음운규칙을 수의적으로 적용하여 가능한 모든 발음 변이를 생성해 내었다. 따라서, POI 의 길이가 길수록(혹은 음절 수가 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 생성하면, 음향적 복잡도에 의하여 오류가 증가할 수 있는데, 여기에서 이러한 오류를 줄이기 위하여 음절수를 기준으로 발음 변이의 수를 조절하는 방법을 제시한다.

베이스라인 시스템은 운율어를 이용하지 않은 발음사전으로서, 이는 다시 표준 발음 사전(bDic_1best), 가능한 모든 발음 사전(bDic-all), N-Best 발음(bDic_Nbest) 사전의 3 개로 구성된다.

운율 정보를 기반으로 제안한 방법에 의한 성능을 평가하기 위하여 음절 수와 운율어 수를 기준으로 분류한 250k POI 를 근거로 하여 6 개의 발음 사전을 설계하였다. <표 1>에 의하면 POI 를 구성하는 운율어 수와 생성된 발음 변이의 수가 음절 수에 비례하여 많아지는데, 특히 발음 변이는 음절 수가 6, 8, 10 일 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, 다음 3 개는 <표 1>의 결과에 따라서 5 음절 이하인 경우는 발음 변이의 수를 1 개로, 6 음절 이상인 경우에는 발음 변이의 수를 각각 2 개(Dic_2), 3 개(Dic-3), 4 개(Dic_4)로 제한한 3 개의 사전을 설계하였다. <표 2>는 제안하는 6 개의 발음 사전을 음절 수에 따른 각각 발음 변이 수가 어떻게 조절되었는지를 보여 주는 표이다.

<표 1> 음절 수에 따라 분류한 POI 의 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

<표 2> 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

4. 실험 및 결과

4.1. 실험 환경

학습을 위해서는 SiTEC 과 ETRI 에서 각각 독자적으로 제작된 103,082 발화로 구성된 2 개의 음성 코퍼스를 사용하였다. SITEC 에서 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 SM-10A)을 이용하여 190 명으로부터 녹음한 8,516 발화로 구성되었다. ETRI 에서 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. 테스트 용 음성 코퍼스로는 학습용 코퍼스 녹음에 참여하지 않은 38 명으로부터 4,149 개의 녹음된 발화를 이용하였다.

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

각각의 발음 사전을 위하여 음향 모델 학습은 16 개 믹스쳐 트라이폰, 30 개 특징 벡터의 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

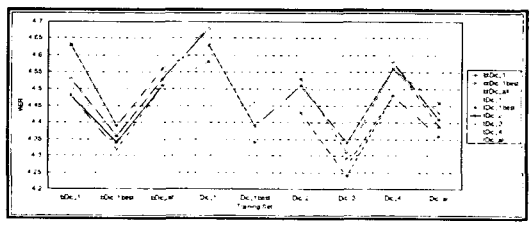
3 개의 베이스라인 사전과 6 개의 제안한 사전을 각각 학습과 테스트에 모두 이용하여 총 81 개 실험을 수행하였고, 그 결과는 다음 <표 3>과 같다. 가장 좋은 결과는 음절 수를 2 개와 3 개로 제한하여 제작한 발음 사전 Dic_2, Dic_3 이었고(4.24 of WER), 가장 나쁜 결과는 베이스라인에서 표준 발음 사전을 이용하였을 때이다(4.63 of WER). 베이스라인에 비하여 제안한 방법으로 최대 8.4%로 WER 가 감소하였다.

<표 3> WER (Word Error Rate) 로 나타난 81 개의 실험 결과

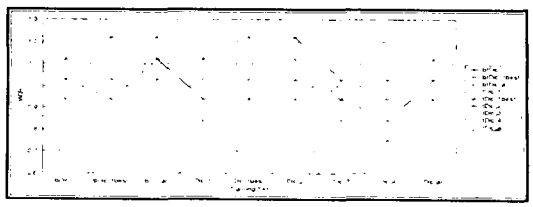
	btDic_1	btDic_1best	btDic_all	tDic_1	tDic_1best	tDic_2	tDic_3	tDic_4	tDic_all
bDic_1	4.53	4.63	4.53	4.48	4.63	4.48	4.48	4.48	4.48
bDic_1best	4.36	4.39	4.34	4.32	4.39	4.34	4.32	4.32	4.29
bDic_all	4.53	4.56	4.52	4.51	4.56	4.51	4.51	4.51	4.48
Dic_1	4.68	4.68	4.68	4.58	4.63	4.58	4.58	4.58	4.58
Dic_1best	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

위에서 언급한 대로 베이스라인과 제안한 사전의 차이는 기본적으로 운율어의 사용 여부에 있는데, <표 3>에서 보는 바와 같이 운율어를 이용하여 발음 사전을 생성한 경우가 그렇지 않은 경우에 비하여 인식 성능이 향상 된 것으로 나타났다. 또한, 이러한 인식 성능의 향상은 6 음절 이상인 경우에 더 명확하게 나타난 것을 볼 수 있었다. 다음 <그림 1>은 각 음절 수에 따라 발음 변이 수를 조절한 결과를 특별히 잘 보여주고 있는데, 최대 발음 변이 수를 3 개로 제한하여 실험한 경우(Dic_3)에 가장 좋은 성능을 보이는 것을 볼 수 있다. <그림 2>는 6 음절 이상 단어들만 따로 실험한 결과를 나타내는데, 이 경우에는 음절 수를 4 로 제한한 경우가 가장 그 성능이 좋은 것으로 나타났다.

<그림 1> 전체 POI 에 대한 WER



<그림 2> 6 음절 이상 POI 에 대한 WER



5. 결론

본 논문은 두 가지 특성 운율적 특성, 즉 운율어와 음절 수를 이용하여 발음모델링할 경우에 음성인식기의 성능을 평가하는 것을 목표로 하는 것으로, 먼저, 운율어를 이용하여 POI 데이터의 가능한 모든 발음을 생성하고, 다시 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. POI 의 음절 수에 따라서 발음 변이를 조절한 결과도 전체적으로 3 음절로 제한한 경우, 6 음절 이상 단어에서는 4 음절로 제한한 경우에 가장 좋은 인식 성능을 얻을 수 있어서, 음절 수에 따른 발음 변이 수의 조절이 효과적임을 알 수 있었다.

이러한 연구는 우선 POI 데이터를 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 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.

감사의 글

이 논문은 산업자원부 지원 뇌신경정보학사업의 "뇌정보처리의 인지신경 기전에 기반한 대화형 멀티모달 사용자 인터페이스 개발" 과제의 연구비 지원으로 수행되었습니다.

참고문헌

1. Strik, H. and Cucchiarini, C. "Modeling Pronunciation Variation for ASR: A Survey of the Literature". *Speech Communication* 29: pp225-246, 1999.
2. Kessens et al., "Improving the performance of Dutch CSR by modeling within word and corss-word pronunciation variation," *Speech Communication* 29, pp193-207, 1999.
3. Jeon, J. H., S. Wee, M, Chung, "Generating Pronunciation Dictionary by analyzing Phonological Variations Frequently Found in Spoken Korean", *Proc. of International Conference on Speech Processing*, pp519-523, 1997.
4. Fosler-Lussier, E., "Multi—level decision trees for static and dynamic pronunciation models," *Proc. Eurospeech 1999*, 1999.
5. Riley et al., "Stochastic pronunciation modeling from hand-labeled phonetic corpora," *Speech Communication* 29, pp209-224, 1999.
6. Kim, S., J. H., Jeon, M. Na, M. Chung, "Irregular Pronunciation Detection for Korean Point-of-Interest Data Using Prosodic Word", *말소리*, 제 57 권, pp123-137, 2006.
7. Hirose, K. 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. Sethy, A., 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. Jun, S.-A., *The Phonetics and Phonology of Korean Prosody: Intonational Phonology and Prosodic Structure*, Garland Publishing Inc., New York : NY., 1996.