한국어 연속음성인식을 위한 형태소 경계에서의 발음 변화 현상 모델링

이 경 님, 정 민 화
서강대학교 컴퓨터학과

Modeling Cross-morpheme Pronunciation Variation for Korean LVCSR

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

Abstract

In this paper, we describe a cross-morpheme pronunciation variation model which is especially useful for constructing morpheme-based pronunciation lexicon for Korean LVCSR. There are a lot of pronunciation variations occurring at morpheme boundaries in continuous speech. Since phonemic context together with morphological category and morpheme boundary information affect Korean pronunciation variations, we have distinguished pronunciation variation rules according to the locations such as within a morpheme, across a morpheme boundary in a compound noun, across a morpheme boundary in an eojeol, and across an eojeol boundary. In 33K-morpheme Korean CSR experiment, an absolute improvement of 1.16% in WER from the baseline performance of 23.17% WER is achieved by modeling cross-morpheme pronunciation variations with a context-dependent multiple pronunciation lexicon.

I. 서론

일반적으로 음성인식을 수행하는데 사용되는 발음사전에는 해당 표제어에 대해 하나의 발음열만을 포함하고 있다. 이는 고립단어 인식에서는 효율적일지 모르나 연속음성을 대상으로 하는 경우, 특히 대용량 연속음성인식의 경우에는 부적당하다. 왜냐하면, 연속음성의 경우 조음 현상으로 인하여 인접한 단어(더 작게는 형태소)간에 다양한 발음 변화 현상이 발생하기 때문이다. 한국어의 어절은 영어에서의 띄어쓰기 단위인 단어의 개념과는 다른 성질을 지닌다. 그러나, 발음사전을 구성하는데 있어서 모든 형태소들의 가능한 조합을 사전 표제어로 등록하는 것은 효과적이지 못하기 때문에, 대부분의 한국어 연속음성인식 시스템은 형태소 기반으로 사전을 구성하고 디코딩 단위로 삼는다.

형태소 기반의 다중 발음사전을 구성하는데 있어서 발생하는 대표적인 문제로서 일부 형태소의 경우 둘 이상의 이형태를 가지는 경우와 "할 수 있다"와 같이 어절 경계에서 발생하는 발음 변화를 들 수 있다. 영어의 경우 단어간에 발생하는 발음 변화 현상을 반영하기 위해 해당 단어열을 하나의 인식 단위로 묶어 모델링하는 Multi-word[4] 개념이 사용되고 있다. 이와 유사하게 한국어에서도 음가 기반의 의사 형태소나 여러 형태소를 연결하여 만든 결합 형태소가 사용되고 있다[5].

본 논문에서는 한국어 대용량 연속음성인식에 필요한 형태소 기반의 발음사전을 효율적으로 구축하기 위한 연구로서 형태소 경계에서 발생하는 발음 변화 현상을 모델링 하는 방법을 제시한다. 일반적으로 형태소 기반의 발음사전을 구성하는데 있어서 표제어 내부에서 발생하는 발음 변화 현상과 사전 경계에서 발생하는 발음 변화 현상에 대한 고려가 필요하다. 형태소 경계는 주로 복합명사나 조사, 접미사 그리고 어미 등의 결합에 의해 생겨난다. 특히 경음화와 같은 일부 규칙은, 비록 같은 음소 문맥일지라도 형태소 경계와 형태소 내부에서의 발음이 다르게 실현되므로 형태소 경계와 형태소 내부의 발음 변화 현상과는 다르게 모델링 되어야 한다. 본 논문에서는 이를 더 구체화하여 형태소 내부, 복합어 경계, 형태소 경계, 어절 경계에서의 발음 변화 현상을 구분하여 모델링 하였다.

실제로 세부적으로 고려된 발음 변화 현상을 반영하여 생성된 다중 발음사전을 사용한 음성인식 실험 결과, WER 측면에서 인식 성능이 1.16%가 향상된 결과를 얻을 수 있었으며, 이를 통해 세분화된 형태소 경계에서의 발음 변화 모델링이 연속음성인식의 성능 향상에 도움이 된다는 것을 확인할 수 있었다.

II. 한국어 음소 변동 규칙

언어학적 지식[2]을 기반으로 한국어에서 발생하는 음운 변화 현상을 정리하고, 문교부에서 제정한 "표준어 규정 - 제 2부 표준 발음법"을 참고하여 한국어의 대표적인 20개의 음소 변동 규칙을 채택하여 적용하였다. 전체적으로 총 13개의 필수 음소 변동 규칙과 7개의 수의적 음소 변동 규칙으로 구성되며, 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 816개의 음소 문맥을 얻을 수 있었다[1].

다음은 한국어에서 발생하는 대표적인 음운 변화 현상인 "경음화" 현상에 대해 구체적인 예를 제시하였다. "경음화"는 이완 장애음이 앞선 장애음과 만나서 무기 경음으로 바뀌는 현상으로 기준에 따라 다양한 종류가 있는데, 여기서는 대표적인 3가지에 대해서 언급하고 이를 규칙으로 변환 처리하였다.

- 장애음 뒤에서의 경음화:
  예) '국밥'(명사)→/국빱/, 꽃다발(명사)→/꼳따발/
- 어간 종성 /ㄴ, ㅁ/ 뒤에서의 경음화:
  예) '신고' (신+고; 어간+어미) → /신꼬/
  '감기' (감+기; 어간+어미) → /감끼/
- 관형형 어미 '-(으)ㄹ' 뒤에서의 경음화:
  예) '갈 수도' → /갈쑤도/
  (가+ㄹ+수+도; 어간+관형형어미+의존명사+보조사)

III. 형태소 경계에서의 발음 변화 현상 모델링

3.1 어휘 모델에서의 발음 변화 모델링

발음 변화 모델을 어휘 모델 단계에서 통합시키기 위해 일반적으로 각각의 사전 엔트리에 가능한 발음열을 포함하는 다중 발음사전을 사용한다. 음향 관측모델 O가 주어졌을 때, 음성 인식의 목표는 확률 $P(W|O)$를 최대화하는 단어열 W를 찾는 것이다. 이는 Bayesian 규칙에 따라 다음과 같이 표현되며, 여기에서 P(W)는 언어 모델을, $P(O|W)$는 음향 모델과 발음사전으로부터 계산된 확률을 의미한다.

$$W^* = \arg\max_W P(W)P(O|W) \quad (1)$$

위의 식에 어휘 모델을 반영한 식으로 확장하면, 식 (1)은 (2)와 같이 수정된다. 여기서 $L_{w,k}$는 단어 W가 가질 수 있는 발음열 중, k번째 발음열을 의미한다. 수정된 식 (2)는 확률값을 최대로 만들기 위한 특정 발음열을 찾는다. 여기서 $P(O|L_{w,k})$는 발음 $L_{w,k}$에 대한 음향학적 확률값을 의미하며, $P(L_{w,k}|W)$는 단어 W가 $L_{w,k}$로 발음될 확률을 의미한다.

$$W^* = \arg\max_{w,k} P(W)P(O|L_{w,k})P(L_{w,k}|W) \quad (2)$$

이러한 발음 변화 현상을 모델링 하는데 있어서 가장 먼저 고려해야 할 것은 형태소 내부와 형태소 경계에서의 발음 변화 현상이다. 형태소 내부에서 발생하는 변화는 음소 환경에 따라 쉽게 모델링 될 수 있으며, 이것은 발음사전의 기본 발음열이 된다. 반면 형태소 경계에서의 발음 변화 현상은 음소 환경 뿐만 아니라 이웃하는 형태소의 결합 구성에 따라서도 영향을 받는다. 다음 표 1의 예제 "교육"은 기본적으로 /K JO JU KQ/으로 발음되며, 음소 문맥과 형태소 결합 정보에 따라 다양하게 발음되어지는 것을 볼 수 있다.

표 1. 예제 "교육"에 대한 다양한 발음열

해당 발음열	형태소 결합 예제	$P(L_{w,k}\|w)$
K JO JU G	교육+이	0.46222
K JO JU KQ	교육+과	0.40889
G JO JU KQ	사회+교육	0.04444
G JO JU G	참+교육+이	0.03333
K JO JU KH	교육+해	0.02889
K JO JU NX	교육+만	0.00667
KK JO JU G	대학+교육+이	0.00667
KK JO JU KQ	대학+교육	0.00667
G JO JU NX	의무+교육+만	0.00222

3.2 형태소 경계에서의 발음 변화 현상

입력 텍스트 문장의 각 문자열을 음소열로 변환하는데 있어서 형태소 경계에서 발생하는 문맥들은 종종 중의성을 내포한다. 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. 한국어의 형태소 경계에서의 발음 변화 현상을 설명하는 데 있어서 몇가지 주목할 사항은 같은 음소 문맥 정보를 갖더라도 형태소 경계 정보와 품사 정보에 따라 적용되는 규칙이 다르다는 것과 언절 내

어절의 경계에서 나타나는 발음 변화 현상은 '경음화', 'ㄴ-첨가', '연음규칙', '격음화', '장애음의 비음화', '변자음화'로 한정된다는 것이다. 여기서 언절은 끊어 읽기 단위로 음운론적인 단위로 볼 수 있으며, 하나 이상의 어절이 모여서 언절로 이루어지므로 형태소 경계에서의 발음 변화 현상을 설명하는데 있어서 규칙이 적용되는 위치가 어절 경계인지 내부인지에 따라 달라져야 한다. 위와 같은 내용을 바탕으로 형태소 경계에서의 발음 변화 현상을 적용 위치에 따라 분류해 보면 다음과 같다. 여기서 '+'는 형태소 경계를 '#'은 언절 내부에서의 어절 경계를 나타낸다.

- 어절 내 형태소 경계
  예) 어간+어미: '신+다'→/신따/
  명사+주격조사: '교육+이'→/교유기/, '솜+이'→/소미/
- 복합명사 내의 형태소 경계:
  예) 명사+명사(복합명사): '솜+이불'→/솜니불/
- 언절 내 어절 경계:
  예) 명사#명사: '대학#교육'→/대학꾜육/
  관형사형전성어미#비단위성의존명사: 'ㄹ#수'→/ㄹ쑤/

3.3 형태소 경계에서의 발음 변화 현상 모델링

한국어의 특징이 잘 반영된 발음열을 생성하려면 주어진 문장에 대해 올바른 형태소열로 태깅하여 그 정보를 사용해야 한다. 형태음운론적 분석을 통해 예제 문장 "신발을 신고"에 적용된 음소 변동 규칙의 세부 규칙 표현은 아래 표 2와 같다. 음소 문맥 항의 L3은 음소 변동이 일어나는 음절 경계의 앞 음절의 종성을 나타내고, R1은 뒷음절 초성을 나타낸다. 변환 코드는 해당 음소 문맥에 대한 음소의 변동 결과를 나타낸다. 경계 정보를 반영하기 위하여 '어/형/복/내/수/다'와 같이 규칙이 적용되는 범위를 나타내는 flag를 사용하여 음소 변동 규칙 오토마타를 구성하였다. 반면, 수의적 음소 변동 규칙의 경우에는 경계 정보와 관계없이 어디서나 적용될 수 있다. 예제 문장의 경우 규칙 4.8과 9.113번이 각각 적용되어 /신바를신꼬/로 변환되며, 규칙 14.1과 14.9에 의해 /심바를싱꼬/라는 추가 음소열도 얻어낼 수 있다.

표 2. 세부 필수 음소 변동 규칙 예제

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
ㄴ	ㄱ	→	ㄴ	ㄲ	9	113	0 1 1 0 0 0
ㄴ	ㅂ	→	ㅁ	ㅂ	14	1	-
ㄴ	ㄲ	→	ㅇ	ㄲ		9	-

Ⅳ. 실험결과 및 분석

4.1 실험 데이터베이스 및 베이스라인 시스템

연속 HMM을 기반으로 한 화자 독립 시스템을 기반으로 음성인식 실험을 수행하였다. 본 실험에는 39차 MFCC 특징 벡터를 사용하였으며, 6개의 Gausian 분포를 갖는 HMM 모델을 사용하였다. 실험 대상으로는 삼성 PBS(Phone Balanced Sentence; 음소균형문장) 음성 데이터베이스를 사용하였다. 학습 데이터로는 43,000문장을, 테스트에는 학습에 참여하지 않은 화자 2000문장의 발화 가운데 10회 이상 발생 기준으로 OOV가 없는 686문장을 선택하였다. 인식 대상 어휘수는 33K 형태소이며, 인식에 사용된 언어 모델은 370K 크기의 back-off bigram으로 테스트 문장을 기준으로 perplexity는 120.87이며, 엔트로피는 6.92bits, bigram hit ratio는 87.88%이다. 문장 분석은 형태소 분석 결과에 품사 태그가 부착된 형태를 기준으로 하였다. 분석 결과 각각 한 문장 당 9.2어절, 한 어절 당 2.1 형태소, 한 형태소 당 1.9음절로 구성되었으며, 형태소 경계는 약 608,777회 발생하였다.

4.2 발음 변화 모델링 성능 평가

해당 실험용 데이터베이스를 학습하기 위해서는 학습용 발음열이 필요하다. 그러나, 한국어에 관한 표준화된 전자 발음사전이 존재하지 않을 뿐만 아니라 전문가에 의한 정확한 발음열을 구축하기에는 방대한 양이다. 일반적으로는 주어진 텍스트를 기반으로 자동 생성을 통해 학습 모델을 구하지만, 생성기를 통해 구축된 발음사전을 객관적으로 평가하기 위해서는 실제 발음열에 가까운 음소열들이 필요하다. 본 실험에서는 직접 음성 문장을 귀로 듣고 사람이 받아적기를 수행한 청각 전사열 43,000문장을 이용하여 모노폰 학습을 수행하여 1차적으로 음향 모델을 만들었다. 보다 정확한 학습을 하기 위해, 발음열 자동 생성기를 사용하여 각 목적에 부합하는 다양한 발음 변화를 포함하는 발음사전을 가지고, 그림 1과 같은 과정을 통하여 음향 모델을 개선하였다. 이 과정을 통해 주어진 문장에 대해 발음사전의 다중 발음열 중 적합한 열을 찾아 음성 데이터와 음향 모델간의 최적화된 정렬을 하는데 사용하였다. 또한 평가를 위해 같은 발음사전을 인식에 사용하였다. 이 실험에서는 성능 평가를 위해 두 가지 척도를 사용하였다. 그림 1과 같이 학습된 음향모델을 사용하여 forced alignment 과정을 통해 인식해서 살아 남은 문장을 카운트 한 FAR(Forced Alignment Rate)이고, 다른 하나는 형태소 단위의 에러 발생 수치를 말해주는 WER(Word Error Rate)이다.

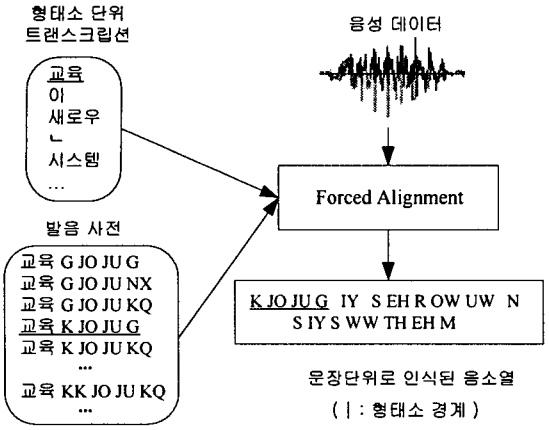

그림 1. Forced Alignment 과정

베이스라인으로는 좌우 음소 문맥만을 고려한 형태소 내부의 발음 변화 현상만을 반영한 발음사전을 사용한 경우이다. 표 3의 2번째 항목과 5번째 이후 항목들은 한국어의 경우 해당 품사 정보에 따라 다르게 발음되는 경우를 반영하기 위해 POS, 즉 품사정보를 고려하여 생성한 발음사전을 사용한 경우이다. FAR의 변화는 없지만 WER의 경우 인식률 측면에서는 약간의 성능이 향상된 것을 볼 수 있었다. 본 논문에서 제안한 형태소 경계에서의 발음 변화 현상을 반영한 결과, 형태소 경계에서의 문맥 종속적인 다중 발음사전을 사용한 경우 7.77%의 FAR 향상을 가져왔다. WER 측면에서도 경계에서의 발음 변화 현상을 반영한 경우 0.75% 정도의 WER를 감소할 수 있었으며, 3.2절에서 제시한 바와 같이 형태소 경계에 대해 내부에서 적용되는 규칙과 달리 변별적으로 고려하여 47개의 음소 문맥에서 다른 규칙이 적용되었다. 실험 결과 추가로 약 0.4%의 WER가 감소하였다. 결과적으로 형태소 경계 정보와 품사 범주를 모두 고려한 발음 변화 모델링에서 가장 좋은 인식률을 얻을 수 있었다.

추가 실험으로 하나의 어절 안에 여러 개의 명사가 결합된 경우, 명사의 경계에서 발생하는 음운 변화 규칙을 적용하지 않은 실험에서는 오히려 FAR와 인식률이 감소하는 현상을 보였다. 이로서 합성명사의 내부의 경계에서 일어나는 발음 변화 현상에 대한 고려가 사전 크기의 증가를 가져오지만 인식 성능의 향상을 위해서 필요하다는 것을 예상할 수 있다. 또한 적용 범위를 확장하여 어절 경계에서 발생하는 '관형형 어미 '-(으)ㄹ' 뒤에서의 경음화'를 반영한 결과, 비록 인식률이 감소하였지만 가장 좋은 FAR을 얻을 수 있었다. 실제 발성된 발음에 근접한 결과를 얻었지만, 대부분 경음화가 적용된 비단위성 단위명사의 경우 단음절 단어로서, 발음사전에 추가된 발음열이 오히려 혼잡도를 증가시켜 인식률이 감소했음을 추측할 수 있다.

표 3. 발음 변화 모델링에 따른 성능 비교

	발음사전	엔트리 수	FAR (%)	WER (%)
1	베이스라인(형태소 내부)	33,398	75.80	23.17
2	1 + POS	33,400	75.80	23.16
3	1 + 형태소 경계	39,733	83.57	22.42
4	1 + 형태소 경계(세부규칙)	39,722	83.57	22.04
5	4 + POS	39,725	83.57	**22.01**
6	5 + 합성명사 분리	37,398	82.71	22.08
7	5 + 관형형 어미 경음화	39,735	**83.62**	22.14

V. 결론

이 논문에서는 한국어 대용량 연속음성인식에 필요한 형태소 기반의 발음사전을 효과적으로 구축하기 위한 방법 중 하나로서 형태소 경계에서의 발음 변화 모델을 제시하였다. 문맥 종속적인 다중 발음사전을 구성하고, 형태소 결합 정보에 따라 형태소 내부와 형태소 경계에서 발생하는 음운 변화 규칙을 구별하여 모델링한 경우 가장 좋은 인식률을 얻을 수 있었다. 반면, 모든 가능한 발음 변이를 사전에 추가하는 경우 사전크기가 증가함에 따라 혼잡도가 증가하여 에러를 유발하게 된다. 이러한 문제를 해결하기 위해 최적의 발음 변이를 선택적으로 적용하는 것이 중요하다. 선택 기준으로는 일반적으로 발생 빈도가 많이 사용되며, 이 외에도 엔트로피나 likelihood와 같은 다양한 선택 기준 방법이 사용되고 있다[3]. 향후 연구과제로서 보다 정확한 발음열을 반영하는 방법과 시스템 성능 향상을 위하여 최적의 발음사전을 구성하기 위한 다양한 접근 방법에 대한 연구를 수행하고 있다.

감사의 글

본 연구는 과기부 국책연구개발사업의 뇌신경정보학과제 (M1-0107-01-0003) 지원으로 수행되었으며, 본 실험에 사용한 삼성종합기술원의 PBS 음성 DB 사용 허가에 감사 드립니다.

참고문헌

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

[2] 이호영, *국어음성학*, 태학사, 1996.

[3] H. Strik, C. Cucchiarini, "Modeling Pronunciation Variation for ASR: A Survey of literature," Speech Communication, 29(2-4): pp.225-246, 1999.

[4] M. Finke and A. Waibel, "Speaking Mode Dependent Pronunciation Modeling in Large Vocabulary Conversational Speech Recognition," Proc. of Eurospeech, pp.2379-2382, 1997.

[5] Y.-H. Park and M. Chung, "Automatic Generation of Concatenate Morpheme Based Language Models for Korean LVCSR," Proc. of ICSP, pp.633-637, 2001.