# 한국어 연속음성인식 시스템 구현을 위한 형태소 단위의 발음 변화 모델링*

정민화(서강대), 이경님(서강대)

<차 례>



<Abstract>

## Modeling Cross-morpheme Pronunciation Variations for Korean Large Vocabulary Continuous Speech Recognition

**Minhwa Chung, Kyong-Nim Lee**

In this paper, we describe a cross-morpheme pronunciation variation model which is especially useful for constructing morpheme-based pronunciation lexicon to improve the performance of a 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 phonological rules that can be applied to phonemes in within-morpheme and cross-morpheme.

The results of 33K-morpheme Korean CSR experiments show that an absolute reduction of 1.45% in WER from the baseline performance of 18.42% WER was achieved by modeling proposed pronunciation variations with a possible multiple context-dependent pronunciation lexicon.

* Keywords: Pronunciation Variation Modeling, Pronunciation Lexicon, Korean Phonological Rules, Large Vocabulary Continuous Speech Recognition.

* 이 논문은 서강대학교 2002년도 교내연구비 지원으로 수행되었습니다.

# 1. 서 론

음성인식 시스템을 구성하는데 있어서 한국어에서 발생하는 음운 변화 현상을 반영하는 방법으로는 일반적으로 크게 두 가지 접근 방법을 들 수 있다. 하나는 어휘모델 측면에서 발음사전 모델링을 통한 지식기반(knowledge-based)의 외재적(explicit) 접근 방법[5][6]과 음향모델 측면에서 공유(sharing)나 tying 기법을 이용한 모델기반(model-based)의 내재적(implicit) 접근 방법[8][9]을 들 수 있다. 본 논문에서는 언어학적 지식기반 접근 방법으로 대용량 연속음성인식에 필요한 형태소 단위의 발음사전 구축을 통하여 한국어 발음 변화 현상을 모델링 하는 방법을 제시한다.

일반적으로 발음사전에는 해당 표제어에 대해 하나의 대표 발음열 만을 포함하게 된다. 이는 고립단어 인식에서는 효율적일지 모르나 연속음성, 특히 대어휘를 인식 대상으로 하는 경우 여러 가지 고려사항이 발생한다. 동시조음(coarticulation) 현상으로 인한 음소문맥 제약 조건뿐만 아니라, 인접한 단어(더 작게는 형태소)간의 결합 환경에 따라 다양한 발음변화 현상이 발생하게 된다. 기존 발음 변화 모델링에 관한 다양한 연구[4][5] 결과, 발생 가능한 다양한 대체 발음열을 발음사전에 반영한 경우 시스템 성능이 향상된다는 사실은 이미 잘 알려져 있다.

한국어의 경우 언어학적 특성상 영어에서의 띄어쓰기 단위인 단어와는 다른 특성을 갖는다. 이때 사전을 구성하는데 있어 모든 형태소들의 가능한 조합을 표제어로 등록하는 것은 효율적이지 못하기 때문에, 대부분의 한국어 연속음성인식 시스템은 형태소를 사전의 표제어 및 디코딩 단위로 사용한다[10]. 형태소 기반의 발음사전을 구성하는데 있어서 형태소 내부에서 발생하는 발음 변화 현상과 형태소 경계에서 발생하는 발음 변화 현상에 대한 고려가 필요하다.

연속음성인식에 있어서 특히 형태소 경계에서의 발음 변화 현상은 매우 다양하게 일어나고 있다. 형태소 경계 발생은 주로 복합명사나 조사, 접미사 그리고 어미 등의 결합에 의해 생겨난다. 특히 경음화와 같은 일부 규칙은, 비록 같은 음소 문맥일지라도 형태소 내부와 형태소 경계를 포함하여 어절 경계에서의 발음이 다르게 실현되므로 경계 정보에 따라 발음 변화 현상이 서로 다르게 모델링 되어야 한다. 본 논문에서는 이를 더 구체화하여 형태소 내부, 복합어 경계, 형태소 경계, 어절 경계에서의 발음 변화 현상을 구분하여 모델링 하였다. 또한 발음열 생성 과정에서 적용된 음소 변동 규칙들의 통계적 자료를 기반으로 한국어 음운 변화 현상에 대한 분석을 수행하였다.

3만 3천 형태소급의 한국어 연속음성인식 시스템을 구성하고, 형태소 경계에서의 세부적인 발음 변화 모델링 및 문맥 종속적인 다중 발음열 사전의 구성을 통하여 베이스라인 시스템의 18.42% WER 보다 1.45% 향상된 결과를 얻을 수 있었다. 실험 결과를 통해 세분화된 형태소 경계에서의 음운 변화 모델링이 연속음성

인식의 성능 향상에 도움이 된다는 것을 확인할 수 있었다.

본 논문의 구성은 다음과 같다. 2장에서는 한국어 음운변화 현상의 적용 과정에 대하여 기술하고, 3장에서는 형태소 경계에서 발생하는 정교한 음운 변화 현상을 반영하기 위한 발음사전 모델링 방법을 제시한다. 4장에서는 제안한 기법의 성능을 비교 분석하여 그 결과를 제시하고, 마지막으로 5장에서 결론을 맺는다.

## 2. 한국어 음운 변화 현상

### 2.1. 한국어의 음운 변화 현상 모델링

말소리의 실현에 있어 단어의 철자는 음운 변화 과정을 통해 발음열로 바뀌게 된다. 음운 변화 과정은 음운 변화가 일어나는 위치에 따라 음소 변동 규칙과 변이음 규칙을 사용해서 설명할 수 있다. 해당 음소 문맥에 의해 하나의 음소가 다른 음소로 바뀌거나 탈락, 첨가되는 양상을 규칙화한 것을 음소 변동 규칙이라 하고, 하나의 음소가 발화 상에서 여러 변이음으로 실현되는 양상을 규칙화한 것을 변이음 규칙이라고 한다. 음소 변동 규칙은 규칙의 적용 양상에 따라 다시 필수 음소 변동 규칙과 수의적 음소 변동 규칙으로 나누어진다.

본 연구에서는 언어학적 지식[1][2]을 기반으로 한국어에서 발생하는 음운 변화 현상과 문교부에서 제정한 "표준어 규정 - 제 2부 표준 발음법"[3]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 주어진다. 이에 따라 실제 음소 문맥에 규칙이 적용되고, 음소 문맥에 따른 세부 규칙으로 총 816개의 음소 문맥을 얻을 수 있었다[7]. 자모음 분류에 의하면 자음 관련 규칙 775개, 모음 관련 규칙 41개이며, 필수와 수의적 규칙으로 분류하면, 필수 음소 변동 규칙은 757개, 수의적 음소 변동 규칙은 59개이다

### 2.2. 음소 변동 규칙의 예

본 절에서는 필수 음소 변동 규칙 중 대표적인 "경음화" 현상에 대해 구체적인 예를 제시하였다. 일반적으로 국어의 경음화 현상은 이완 장애음이 앞선 장애음과 만나서 무기 경음으로 바뀌는 현상이다. 분류 기준에 따라 다양한 방식으로 분석되어지는데 여기서는 대표적으로 5가지로 분류하여 정리하였다. 이 중 필수적으로 일반화하여 규칙화 할 수 있는 상위 3가지의 경우에 한하여 규칙으로 변환 처리하였다.

1) 장애음 뒤의 경음화:
   예) '국밥'(명사) →/국빱/, 꽃다발(명사) →/꼳따발/
2) 어간 종성 /ㄴ, ㅁ/ 뒤의 경음화:
   예) '신고' (신+고; 어간+어미) →/신꼬/, '감기' (감+기; 어간+어미) →/감끼/
3) 관형형 어미 '-(으)ㄹ' 뒤의 경음화:
   예) '갈 수도' →/갈쑤도/ (가+ㄹ+수+도; 어간+관형형 어미+의존명사+보조사)
4) 사이 시옷에 의한 경음화
   예) 냇가 →/낻까|내까/, 눈동자 →/눈똥자/, 문고리 →/문꼬리/
5) 한자음에 의한 경음화
   예) 시가(市價) →/시까/, 폐병(肺病) →/폐뼝/, 여권(旅券) →/여꿘/

국어의 경음화 현상은 문맥에 따라, 개인 또는 지방에 따라 실현에 차이를 보이기도 하여 지금까지 끊임없는 연구의 대상이 되어왔다. 특히 한자어나 복합어 내부의 경음화는 동일한 음운론적 환경, 형태론적 환경에서 선택적으로 실현되기 때문에 기준을 알기 어렵다. 규칙으로 생성 가능한 단어 이외에는 예외 발음사전을 사용하여 규칙이 적용되기 전에 생성되도록 하였다.

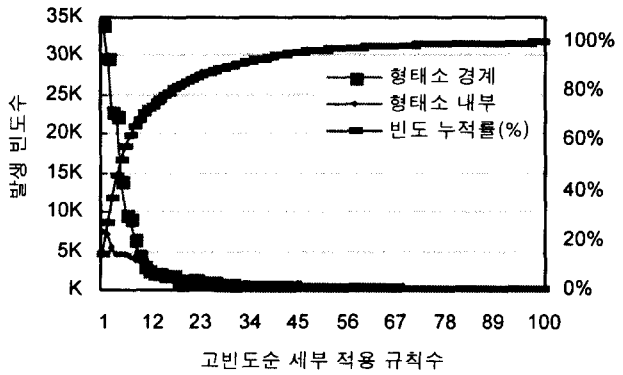
## 2.3. 필수 음소 변동 규칙의 통계적 분석

13개 범주의 필수 음소 변동 규칙이 적용된 자료를 분석한 결과[7], 가장 빈도수가 높은 규칙은 연음법칙이며, 경음화, 격음화, 장애음의 비음화 순으로 발생 빈도수를 기록하였다. 필수 음소 변동 규칙의 세부 규칙 발생 횟수 및 분포 다음 표 1과 같다. 발생 빈도수를 기준으로 가장 많이 적용된 필수 음소 변동 세부 규칙은 형태소 내부의 경우 규칙 번호 4.4(연음규칙; 'ㄴ+ㅇ→∅+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 '있/paa+다/ef→/읻+따/'의 경우를 들 수 있다.

<표 1> 적용 빈도수 기준 상위 5개의 필수 음소 변동 규칙
(규칙 번호는 [7] 참조, 예제의 tag 표기는 형태소 품사의 대분류 기호)

| 순위 | 형태소 내부 | | | 형태소 경계 | | |
|---|---|---|---|---|---|---|
| | 규칙 번호 | 적용규칙 | 예제 | 규칙 번호 | 적용규칙 | 예제 |
| 1 | 4.4 | ㄴ+ㅇ→Ø+ㄴ | 운영→/우녕/ | 9.72 | ㅆ+ㄷ→ㄷ+ㄸ | 있+다→/읻+따/ |
| 2 | 4.8 | ㄹ+ㅇ→Ø+ㄹ | 졸음→/조름/ | 4.4 | ㄴ+ㅇ→Ø+ㄴ | 기분+을→/기부+늘/ |
| 3 | 9.4 | ㄱ+ㅅ→ㄱ+ㅆ | 박수→/박쑤/ | 4.1 | ㄱ+ㅇ→Ø+ㄱ | 국+이→/구+기/ |
| 4 | 9.1 | ㄱ+ㄱ→ㄱ+ㄲ | 학교→/학꾜/ | 4.8 | ㄹ+ㅇ→Ø+ㄹ | 수술+이→/수수+리/ |
| 5 | 5.1 | ㄴ+ㄹ→ㄹ+ㄹ | 천리→/철리/ | 4.17 | ㅁ+ㅇ→Ø+ㅁ | 중심+에→/중시+메/ |

다음 <그림 1>은 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.




<그림 1> 적용된 필수 음소 변동 규칙의 분포도

## 2.4. 형태소 범주에 따른 규칙 적용 분석

한국어는 형태소의 범주에 따라 서로 다른 음소열로 발음열이 실현된다. 입력 문자열 '신고'의 경우와 같이 어간과 어미의 결합인지 하나의 명사인지에 따라 다

르게 발화된다. 여기서는 필수 음소 변동 규칙이 형태소의 범주에 따라 어간, 어미, 조사, 명사 · 부사 · 관형사(default), 복합어로 분리하여 수행된 결과를 분석하였다. 표 2는 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 2> 형태소 범주 별로 적용된 음소 변동 규칙 분석[7]

| 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% |

## 3. 형태소 경계에서의 발음 변화 현상 모델링

### 3.1 음성인식 시스템 구현에서의 발음 변화 현상 모델링

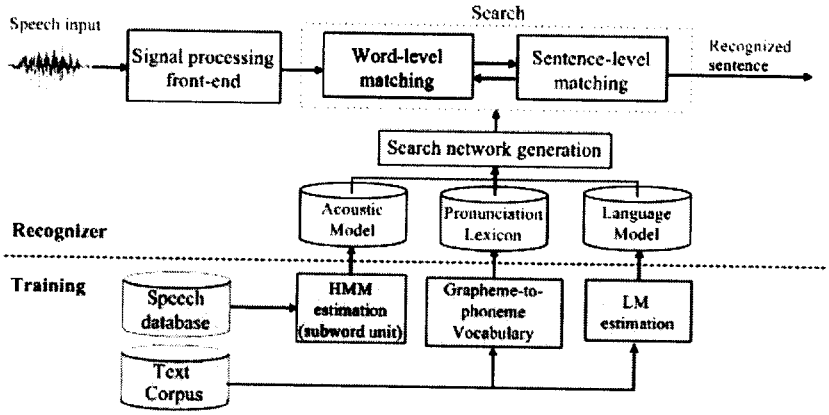
음성 인식 시스템을 구성하는 대표적인 요소로는 크게 음향모델, 어휘모델, 언어모델을 들 수 있다. 음성 인식의 최종 목표는 음향 관측모델 $O$가 주어졌을 때, 확률 $P(W|O)$를 최대화하는 단어열 $W$를 찾는 것이므로, Bayesian 규칙에 따라 다음과 같이 표현되며, 여기에서 P(W)는 언어 모델을, $P(O|W)$는 음향 모델로부터 계산된 확률을 의미한다.

$$\hat{\mathbf{W}} = \arg\max_{\mathrm{W}} P(\mathbf{W} \mid \mathrm{O}) = \arg\max_{\mathrm{W}} P(\mathrm{O} \mid \mathbf{W})\mathrm{P}(\mathbf{W}) \qquad (1)$$

어휘 모델 측면에서 발음 변화 현상을 반영하는 일반적인 방법으로는 해당 엔트리에 대응하는 가능한 대체 발음열을 포함하는 다중 발음사전을 사용하는 것이다. 위의 식에 어휘 모델을 반영한 식으로 확장하면, 식 (1)은 (2)와 같이 수정된다. 여기서 $L_{w,k}$는 단어 $W$가 가질 수 있는 발음열 중, $k$번째 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 $W$가 $L_{w,k}$로 발음될 확률을 의미한다.

$$\hat{\mathbf{W}} = \arg\max_{\mathrm{W},k} P(\mathbf{W})P(\mathrm{O}\,|\,\mathrm{L}_{W,k})P(\mathrm{L}_{W,k}\,|\,\mathbf{W}) \tag{2}$$




<그림 2> 연속음성인식 시스템 구성도

음성은 매 순간 변화하는 특성이 있기 때문에 사람이 발성하는 말소리도 매 시간에 따라 다르게 실현된다. 이러한 발음 변화의 원인들은 일반적으로 1) 잡음과 같은 환경에 의한 변화, 2) 발화자의 연령, 성별, 물리적 구조와 조건에 따른 음향 신호로의 변이, 3) 화자들 사이의 방언이나 사투리의 차이에 의한 단어의 발음 변이, 4) 단어 사이에 발생하는 음운론적 문맥에서 발생하는 조음현상에 의한 변이, 5) 문법적 사용 또는 스타일로의 변화를 들 수 있다.[4][5]

다음 표 3의 예제는 Two-level 모델에 입각하여 입력 문자열 '한국'에 대해 기저형과 표면형의 불일치 문제를 보여준다. 이러한 불 일치성으로 인하여 음향모델 학습 과정에서 실제 입력된 음성과는 다른 음소로 학습됨에 따라 음향모델이 오염되는 현상이 발생한다. 또한 인식 과정에서 보면, 실제 인식된 음소열 시퀀스와 단어가 일치하지 않는 현상이 발생하게 된다. 이러한 문제들은 음성인식 시스템 구성할 때 일반적으로 발음사전(Pronunciation Lexicon) 사용하여 해결한다.

<표 3> Two-level 모델 예제; 문자열과 발음열 대응 관계

| 분류 1 | 분류 2 | | 예 제 | |
|---|---|---|---|---|
| 기저형 (Baseform) | 문자열 (Canonical transcription) | | 한국 | 싶은 |
| 표면형 (Surface form) | 발음열 | 음소 전사열(phonemic transcription) | /항국/ | /시픈/ |
| | | 음성 전사열(phonetic transcription) | [H AA NX G UW KQ] | [S IY PH WW N] |

이 모든 현상들을 분석하여 발음 변화를 모델링 하는 것이 더 좋은 성능을 보일 수 있겠지만 주어진 모든 현상들을 한 시스템 내에 모델링 한다는 것이 그리 쉬운 것이 아니며, 또한 모든 정보를 사용하여 반드시 좋은 결과를 얻을 수 있는 것은 아니다. 본 논문에서는 한국어가 가지는 음성학 및 음운론적 특성 위주의 음운 변화 현상을 반영한 음성 인식 시스템 설계를 목표로 접근 방법을 수행하였다.

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 환경뿐만 아니라 이웃하는 형태소의 결합 구성에 따라서도 영향을 받는다. 예제 "교육"의 경우 기본적으로 /K JO JU KQ/으로 발음되며, 음소 문맥과 형태소 결합 정보에 따라 9가지의 대체 발음열 리스트로 다양하게 발음되는 것을 볼 수 있다[8]. 이러한 변화 현상을 반영한 발음사전을 사용하여 음성인식기의 성능을 향상시키는 것이 목표이다.

### 3.2 형태소 경계에서의 음운 변화 현상 모델링

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 하나이상의 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.

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

한국어의 특징이 잘 반영된 발음열을 생성하려면 주어진 문장에 대해 올바른 형태소 열로 태깅하여 그 정보를 사용해야 한다. 형태 음운론적 분석을 통해 예제 문장 "신발을 신고"에 적용된 음소 변동 규칙의 세부규칙 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 4> 세부 필수 음소 변동 규칙 예제

| 음소문맥 | | | 변환코드 | | 규칙번호 | 세부규칙번호 | 적용범위 |
|---|---|---|---|---|---|---|---|
| L3 | R1 | | L3 | R1 | | | 어/형/복/내/수/다 |
| ㄹ | ㅇ | → | ∅ | 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 | - |

### 3.3 Forced Alignment를 이용한 음향 모델 학습

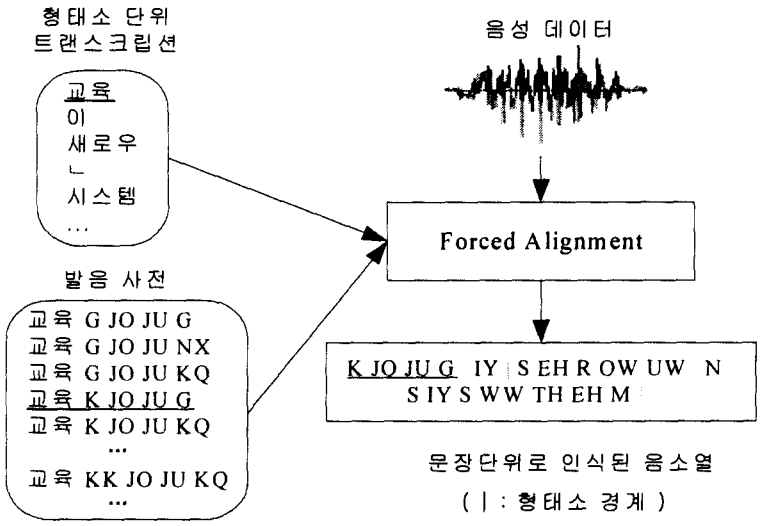
일반적인 음향 모델 학습 과정에는 Baum-Welch 알고리즘이 사용된다. 먼저 임의로 초기 HMM 모델을 정의한 후, 학습 집합(set)의 음성 데이터와 해당 전사 파일(transcription file)을 이용하여 음소 문맥을 반영하지 않은(context-independent) 모노폰을 학습한다. 모노폰 단위로는 조음 현상(coarticulation)을 반영하기 어렵기 때문에 트라이폰 으로 확장하여 재학습을 수행한다. 전사 파일로부터 트라이폰 목록을 만들고 먼저 학습된 모노폰 으로 부터 초기 트라이폰 모델을 만들게 된다. 이때 트라이폰은 학습시 단어 내에서 생성된 것들만이 학습된다. 학습에서 나오지 않는 트라이폰과의 학습 문제를 보완하기 위해 이전에는 음향학적으로 비슷한 모델끼리 공유하는 일반화된 트라이폰(Generalized triphone)을 사용했으나, 요즘에는 임의의 두 트라이폰 사이에 그 출력 확률 분포가 매우 유사한 상태가 있을 때 상태를 공유하도록 하는 tied-state 트라이폰[9]이 사용된다. 이때 공유되는 음소들은 주어진 데이터로부터 군집화(Clustering) 하는 상향식 방법과 언어적인 지식을 이용하는 음향학적 결정 트리(phonetic decision tree)를 구성한 후 임의의 음소 모델에 대한 음향학적인 특성을 판단하게 하는 하향식 방법에 의해 만들어지는데, 본 시스템에서는 하향식 방법을 이용하였다. 결정 트리는 조음 방식과 조음 위치에 따라 분류되도록 구성하였다. 이러한 식으로 트라이폰의 학습이 완료되면 발음사전의 모든 표제어들로부터 발생 가능한 트라이폰 목록(seen & unseen triphone)을 만들어 낸 후, 음향학적 결정 트리를 이용하여 실제 생성된 트라이폰과 비슷한 트라이폰들을 생성해 낸다.

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 텍

스트 전사물을 일관 변환하거나 자동 생성을 통해 학습 모델을 구하지만, 정확한 학습과 객관적 평가를 위해서는 실제 발음열에 가까운 음소열 들이 필요하다. 본 실험에서는 직접 발성된 문장을 귀로 듣고 사람이 받아 적기를 수행한 청각 발음열 43,000문장을 이용하여 모노폰 단위로 학습된 1차 음향 모델을 생성하였다.

보다 정확한 학습을 수행하기 위해, 학습하고자 하는 음성 문장을 넣고 모노폰 단위로 학습된 음향 모델을 이용하여 폰 인식을 수행하였다. 이때 인식 범위는 전체 어휘가 아니라, 주어진 문장의 정답을 주고 해당 단어에 해당하는 대체 발음열들 중에서 likelihood가 가장 큰 음소열을 결과 값으로 출력한다. 이 과정을 forced alignment라고 하며, 이 과정을 거쳐 학습에 사용될 새로운 발음열을 생성한다.

각 목적에 부합하는 다양한 발음 변화를 포함하는 발음사전을 자동 생성하고, 그림 3과 같은 forced alignment 과정을 통하여 음향 모델을 개선하였다. 이 과정을 통해 주어진 문장에 대해 발음사전의 다중 발음열 중 적합한 음소열을 찾아 음성 데이터와 음향 모델간의 최적화된 정렬을 하는데 사용하였다. 또한 평가를 위해 인식 과정에서도 같은 발음사전을 사용하였다.




<그림 3> Forced Alignment 과정

# 4. 실험 결과 및 분석

## 4.1 실험 환경 및 베이스라인 시스템

연속 HMM을 기반으로 한 화자 독립 시스템[9]을 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 데이터베이스[1]를 사용하였다. 학습 데이터로는 43,000문장을, 테스트에는 학습에 참여하지 않은 화자 2000문장의 발화 가운데 10회 이상 발생 기준으로 OOV가 없는 680문장을 선택하였다. 인식 대상 어휘 수는 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[8].

## 4.2 성능 평가

본 실험에서는 성능 평가를 위해 두 가지 척도를 사용하였다. 그림 3과 같이 학습된 음향모델을 사용하여 forced alignment 과정을 통해 인식해서 나온 음소 인식 결과와 청각 발음열을 폰 단위로 비교한 PAR (Phone Accuracy Rate)로 실제 발성된 발음에 얼마나 근접한 결과를 얻었는지 알아보는 평가 수치이다. 다른 하나는 형태소 단위의 에러 발생 수치를 말해주는 단어 인식률(WER: Word Error Rate)이다.

전체 실험 결과는 표 5와 같다. 베이스라인으로는 좌우 음소 문맥만을 고려한 형태소 내부의 음운 변화 현상만을 반영한 발음사전을 사용한 경우이다. 표 5의 2번째 항목과 5번째 이후 항목들은 한국어의 경우 해당 품사 정보에 따라 다르게 발음되는 경우를 반영하기 위해 POS, 즉 형태소 태그를 고려하여 생성한 발음사전을 사용한 경우이다. PAR의 큰 변화는 없지만 WER의 경우 인식률 측면에서는 약간의 성능이 향상된 것을 볼 수 있었다. 본 논문에서 제안한 형태소 경계에서의 음운 변화 현상을 반영한 결과, 형태소 경계에서의 문맥 종속적인 다중 발음사전을 사용한 경우 8.45%의 PAR 향상을 가져왔다. WER 측면에서도 경계에서의 음운 변화 현상을 반영한 경우 0.78% 정도의 WER를 감소할 수 있었으며, 3.2절에서 제시한 바와 같이 형태소 경계에 대해 내부에서 적용되는 규칙과 달리 변별적으로

1) Lab of Samsung Advanced Institute of Technology has funded our construction of the PBS DB.

고려하여 47개의 음소 문맥에서 다른 규칙이 적용되었다. 실험 결과 약 0.13% 가량 추가로 WER가 감소하였다. 결과적으로 형태소 경계 정보와 품사 범주를 모두 고려한 음운 변화 모델링에서 가장 좋은 인식률을 얻을 수 있었다.

추가 적용 범위를 확장하여 어절 경계에서 빈번히 발생하는 '관형형 어미 '-(으)ㄹ' 뒤에서의 경음화'를 반영한 결과, 비록 PAR에는 거의 변화가 없었지만, WER가 0.23% 정도 추가 감소를 보였다. 대부분 경음화가 적용된 경우 비단위성 단위명사의 단음절 단어로서 추가 발음열이 한정되어 있지만, 한 어절 내에 여러 개의 명사가 결합된 경우, 경계에 발생하는 음운 변화 규칙을 모두 반영하여 발음사전에 추가하면 오히려 혼잡도를 증가시키는 요인으로 작용하여 인식률 감소를 야기할 수도 있다.

<표 5> 음운 변화 현상을 반영한 모델에 따른 성능 비교

| No | 발음 사전 | 엔트리 수 | PAR (%) | WER (%) |
|---|---|---|---|---|
| 1 | 베이스라인 (형태소 내부) | 33,398 | 84.91 | 18.42 |
| 2 | 형태소 내부 + 품사 정보 | 33,400 | 84.97 | 18.21 |
| 3 | 형태소 내부 + 형태소 경계 + 품사정보(형태소 내부와 경계 같은 규칙 적용) | 39,733 | 93.42 | 17.33 |
| 4 | 형태소 내부 + 형태소 경계(내부와 경계에 서로 다른 세부 규칙 적용) | 39,722 | 93.46 | 17.28 |
| 5 | 형태소 내부 + 형태소 경계 + 품사정보(세부 규칙 적용) | 39,725 | 93.49 | 17.20 |
| 6 | 형태소 내부 + 형태소 경계 + 품사정보 + 어절 경계의 (관형형 어미) 경음화 현상 (세부 규칙 적용) | 39,735 | 93.50 | 16.97 |

## 5. 결 론

이 논문에서는 한국어 대용량 연속음성 인식에 필요한 형태소 기반의 발음사전을 효과적으로 구축하기 위한 과정 중 하나로서 특히 형태소 경계에서의 발음 변화 현상 모델링에 초점을 맞추었다. 다양한 조건에 대한 인식 실험을 수행한 결과, 문맥 종속적인 다중 발음사전을 기반으로 형태소 결합 정보와 품사의 종류에 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 사전 크기가 증가함에 따라 혼잡도가 증가하여 에러를 유발하게 된다. 이러한 문제를 해결하기 위해 최적의 발음 변이를 선택적으로 적용

하는 것이 중요하다. 선택 기준으로는 일반적으로 발생 빈도가 많이 사용되며, 이외에도 엔트로피나 likelihood와 같은 다양한 선택 기준 방법이 사용되고 있다[4]. 향후 연구과제로서 보다 정확한 발음열을 반영하는 방법과 시스템 성능 향상을 위하여 최적의 발음사전을 구성하기 위한 다양한 접근 방법에 대한 연구를 수행하고 있다.

## 참 고 문 헌


[1] 이기문, 김진우, 이상억, 국어음운론, 학연사, 2000.

[2] 이호영, *국어음성학*, 태학사, 1996.

[3] *표준어 규정 - 제 2부 표준 발음법*, 문교부 고시 제 88-2호, 1988.

[4] H. Strik, C. Cucchiarini, "Modeling Pronunciation Variation for ASR: A Survey of literature", *Speech Communication*, Vol. 29, No. 2-4, pp.225-246, 1999.

[5] H. Strik, 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.

[6] Kyong-Nim, Minhwa Chung, "Modeling Cross-morpheme Pronunciation Variations for Korean Large Vocabulary Continuous Speech Recognition", *Proc. of the 8th European Conference on Speech Communication and Technology (EUROSPEECH '03)*, pp. 261-264, Geneva, Switzerland, Sept.. 1-4, 2003

[7] Kyong-Nim, Minhwa Chung, "Statistical Analysis of Korean Pronunciation Variations", *Proc. of the 15th International Congrees Phonetic Science (ICPhS '03)*, pp.2745-2748, Barcelona Spain, August 3-9, 2003.

[8] M. Finke, A. Waibel, "Speaking Mode Dependent Pronunciation Modeling in Large Vocabulary Conversational Speech Recognition", *Proc. of the European Conference on Speech Communication and Technology (EUROSPEECH '97)*, pp.2379-2382, 1997.

[9] P. C. Woodland, C. J. Leggetter et al., "The 1994 HTK Large Vocabulary Speech Recognition System", *Proc. of 1995 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '95)*, Detroit, MI, USA, May 1995.

[10] Y.-H. Park, D.-H. Ahn, M. Chung, "Morpheme-based Lexical Modeling for Korean Broadcast News Transcription", *Proc. of 8th Europoean Conference on Speech Communication and Technology (EUROSPEECH '03)*, pp.1129-1132, Geneva, Switzerland, Sep. 2003.

접수일자: 2004년 2월 16일
게재결정: 2004년 3월 15일

▶ 정민화(Minhwa Chung)
주소: 121-742 서울시 마포구 신수동 1 서강대학교
소속: 서강대학교 컴퓨터학과
전화: 02) 705-8496
FAX: 02) 704-8273
E-mail: mchung@sogang.ac.kr

▶ 이경님(Kyong-Nim Lee)
주소: 121-742 서울시 마포구 신수동 1 서강대학교
소속: 서강대학교 컴퓨터학과
전화: 016) 201-7492
E-mail: knlee@sogang.ac.kr