한국어 대어휘 연속음성 인식용 발음사전 자동 생성 및 최적화*

이경님(성신여대), 정민화(서울대)

<차 례>

<Abstract>

Building a Morpheme-Based Pronunciation Lexicon for Korean Large Vocabulary Continuous Speech Recognition

Kyong-Nim Lee, Minhwa Chung

In this paper, we describe a morpheme-based pronunciation lexicon useful for Korean LVCSR. The phonemic-context-dependent multiple pronunciation lexicon improves the recognition accuracy when cross-morpheme pronunciation variations are distinguished from within-morpheme pronunciation variations. Since adding all possible pronunciation variants to the lexicon increases the lexicon size and confusability between lexical entries, we have developed a lexicon pruning scheme for optimal selection of pronunciation variants to improve the performance of Korean LVCSR. By building a proposed pronunciation lexicon, an absolute reduction of 0.56% in WER from the baseline performance of 27.39% WER is achieved by cross-morpheme pronunciation variations model with a phonemic-context-dependent multiple pronunciation lexicon. On the best performance, an additional reduction of the lexicon size by 5.36% is achieved from the same lexical entries.

* Keywords: Pronunciation variation modeling, Continuous speech recognition.

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

1. 서 론

음성인식 기술은 초창기의 경우 소규모 고립 단어 인식이나 제한된 연결어 인식을 위주로 개발되어 왔지만, 점차적으로 언어학적 특성을 포함하여 다양한 종류의 변화 현상을 다루어야 하는 대규모 어휘의 연속어 인식까지 그 기술의 범위가 확대되어 왔다. 특히, 최근에는 받아쓰기(dictation)를 비롯하여 방송뉴스(broadcast news)나 강연(lecture speech)의 오디오에 있는 음성 및 음향을 인식하여 응용하는 다양한 연구가 진행되고 있다.

이러한 기능을 수행하기 위해서는 음성인식 시스템이 수용할 수 있는 어휘가 수만 단어 이상이 되어야 하며, 다양한 문법구조 및 발음현상 등이 고려되어야 한다. 특히, 음성은 매 순간 변화하는 특성이 있기 때문에 사람의 말소리도 발성할 때마다 다르게 실현된다. 발음이 변화하는 원인으로는 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 스타일에 의한 변화를 들 수 있다[1]. 이러한 현상들로 인하여 실제 발성된 음성과 인식 문장 사이에 불일치성이 발생하게 되며, 자연스러운 문장을 인식하는데 있어서 어려움을 주는 요인이 된다. 본 논문에서는 여러 현상들 중에서 한국어가 가지는 음성학 및 음운론적 특성에 기반을 둔 발음 변화 모델링 기법을 적용함으로써 음성 인식 성능을 향상시키고자 한다.

음성 인식 시스템을 구성하는데 있어서 발음 변화 현상에 대한 모델링 기법은 크게 두 가지 방식으로 접근한다. 구체적으로 외재적(explicit) 방식으로는 어휘모델 측면에서 지식기반(knowledge-based)[2] 또는 데이터기반(data-driven)의 발음사전 모델링을 통한 접근 방법[3][4]을 들 수 있다. 내재적(implicit) 방식으로는 음향모델 측면에서 데이터를 이용하여 forced alignment를 거친 음향모델 재학습(retraining) 방법과 모델기반(model-based)의 공유(sharing)나 tying 기법을 이용한 접근 방법[5]을 들 수 있다. 대부분의 시스템 구성에서는 인식하고자 하는 어휘 목록과 이에 해당하는 발음열을 수록한 발음사전(Pronunciation Lexicon)을 사용하여 기본적인 발음 변화 현상을 해결하는 것이 일반적이다.

한국어는 표음문자이지만 음운 변화 현상이 발생하게 됨으로써 해당 문자 표기열과는 다르게 소리값이 실현된다. 동시조음(coarticulation) 현상뿐만 아니라, 언어학적 특성상 형태소 경계에서의 발음이 다르게 실현되므로 인접한 문맥정보와 경계 여부에 따라 서로 다르게 모델링 되어야 한다. 기존 연구로 텍스트 기반에 음성인식 후처리 측면에서 음운 접속정보와 형태소 접속정보 등을 활용하여 원래의 문자열로 복원하는 역추적 규칙이 제안되었다[6]. 하지만 이 방식은 어휘수가 증가하게 되면 탐색 과정이 복잡해질 뿐만 아니라 가능한 후보 결과가 늘어남에

따라 시스템 자체에 부하를 가져올 수도 있다. 따라서 일반적으로 발음 변화에 따른 표면형과 기저형의 불일치는 발음사전을 사용하여 해결한다. 보통 해당 표제어에 대해 하나의 대표 발음열(canonical pronunciation)만을 포함하기 때문에 인접한 단어 간의 결합 환경에 따른 다양한 변화 현상을 반영하기 어려운 문제점은 발생 가능한 다양한 대체 발음열(alternative pronunciation)을 다중 발음사전에 추가 반영함으로써 음성 인식 시스템 내부에서 해결하고자 하였다[2]. 그러나 사전에 수록된 정보만으로는 일반적인 발음 변화를 모두 반영하지 못할 뿐만 아니라, 관찰되는 모든 현상들을 사전에 추가하는 경우 오히려 단어간 혼동을 유발함으로써 인식 시스템의 성능을 하락시킬 수 있다. 따라서 관찰된 발음 변화에 대한 요인을 분석하여 발생되는 현상들에 대한 체계적인 모델링이 필요하게 된다.

본 논문에서는 형태소 단위의 한국어 연속 음성 인식용 발음사전을 자동 생성하고 최적화 기법을 제시한다. 형태소 및 어절 경계에서 발생하는 다양한 음운 변화 현상을 반영하기 위해 학습용 코퍼스 문장을 이용하여 음소 문맥과 형태소 결합 정보에 따라 서로 다른 규칙을 적용한 형태소 단위의 어휘 목록과 해당 발음열을 자동 생성하였다. 음향학적 측면에서는 forced alignment 과정을 통해 주어진 문장에 대해 발음사전에 기재된 다중 발음열 중 적합한 음소열을 찾음으로써 음성 데이터와 음향 모델간의 최적화된 정렬을 수행하여 음향 모델을 재학습 하였다. 마지막으로 인식 단계에서 사용되는 발음사전의 최적화를 위해 사전 크기 및 혼잡도를 줄이기 위한 사전 pruning 과정을 적용하였다. 해당 어휘에 대해 발생 가능한 발음열의 가짓수와 발생 빈도 등을 고려한 가중치 적용을 통해 발음사전에 기재될 발음들을 선택적으로 선별하는 사전 최적화 기법을 제시하였다.

실험에는 방송뉴스 영역을 대상으로 기존의 3만 3천 형태소급 낭독체 연속음성인식용 음향모델에 방송뉴스 데이터를 적응하여 재구성된 음향모델을 사용하였다. 베이스라인으로는 형태소 경계에서의 발음 변화 모델을 적용한 문맥 종속적인 다중 발음사전을 사용하였으며, 단일 발음사전과 비교하여 현저한 성능 향상을 확인할 수 있었다. 인식용 사전으로는 기존의 확률기반 pruning 기법보다 논문에서 제안한 단어빈도율과 대표 발음열의 수를 고려한 사전 pruning 기법을 적용한 경우에 있어서 안정적인 성능 향상을 보였다. 베이스라인을 기준으로 WER의 절대적 에러 감소량이 최대 0.56%일 때, 사전 크기는 약 5.36% 정도 감소하였다.

본 논문의 구성은 다음과 같다. 2장에서는 연속음성인식 시스템을 구성하기 위한 한국어 음운변화 현상의 적용 과정에 대하여 기술하고, 3장에서는 음성인식용 발음사전을 최적화하기 위한 pruning 기법을 이용한 발음사전 모델링 방법을 제시한다. 4장에서는 제안한 기법의 성능을 비교 분석하여 그 결과를 제시하고, 마지막으로 5장에서 결론을 맺는다.

2. 한국어 연속음성인식 시스템 구성에서의 발음 변화 모델링

2.1 연속 음성 인식 시스템 구성

음성 인식 시스템을 구성하는 대표적인 요소로는 크게 음향모델, 어휘모델, 언어모델을 들 수 있다. 음성 인식의 최종 목표는 음향 관측모델 O가 주어졌을 때, 확률 $P(W|O)$를 최대화하는 단어열 W를 찾는 것이므로, Bayesian 규칙에 따라 다음과 같이 확률의 조합으로 표현된다.

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

발음 변화 현상을 반영하는 일반적인 방법으로는 어휘 모델 측면에서 해당 엔트리에 대응하는 가능한 대체 발음열을 포함하는 다중 발음사전을 사용하는 것이다. 위의 식에 어휘모델을 반영하여 확장하면, 식 (1)은 (2)와 같이 수정된다. 여기서 $L_{W,k}$는 단어 W가 가질 수 있는 발음열 중, k번째 발음열을 의미하며, 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.

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

잘 설계된 어휘 모델은 발화자의 음성과 음향모델, 그리고 외재적인 발음 모델 사이에서의 불일치성을 가져오는 부정적인 요소들을 줄일 수 있다. 불일치 현상을 보상하기 위한 대표적 기법으로는 음향모델 적응 방법과 발음사전 적응 방법을 들 수 있으며, 두 가지를 결합한 방법이 더 좋은 결과를 가져올 것이다. 구체적으로 변이음 적용이나 재학습 등을 통한 음향모델 적응 방법과 다양하고 정교한 음운 변화 규칙을 적용하는 사전 적응 방법을 들 수 있다.

2.2 언어학적 지식에 기반한 발음열 자동 생성

대부분 대용량 연속 음성인식 시스템에서는 기본 인식단위로 음소(Phoneme)나 음소와 유사한 단위(PLU; Phoneme-like Unit) 등을 주로 사용하며, 발성된 단어나 문장을 인식하기 위해서는 기본 subword 단위들로 구성된 어휘 발음사전을 필요로 하게 된다. 발음사전은 보통 주어진 어휘 리스트와 이에 해당하는 표준 발음 표기(baseform)를 인식 단위의 열로 나열하여 구성한다. 보통 이러한 발음 표기열은 표준 발음사전을 참조하거나 음운학적 지식을 가진 전문가에 의해서 작성된다. 그러나 한국어에 관한 표준화된 전자 발음사전이 존재하지 않을 뿐만 아니라, 개발 분

야마다 전문가에 의해 발음열을 구축하기에는 응용 영역이 방대하고, 어휘량이 증가함에 따라 수작업에 의한 비용 소모와 일관성 유지 문제가 발생된다. 따라서 영역에 의존적이지 않으면서 한국어의 음운 변화 현상이 반영된 발음열 자동 변환 생성기의 사용이 필수적이다.

입력 텍스트 문장의 문자열을 음소열로 변환하는데 있어서 형태소 경계에서 발생하는 문맥들은 종종 중의성을 내포한다. 특히 한국어 문장은 하나 이상의 형태소들이 결합된 어절들로 구성되므로, 형태소를 디코딩 단위로 삼는 경우 경계에서 발생하는 음운 변화 현상이 고려되어야 한다. 몇 가지 주목할 사항은 같은 음소 문맥 정보를 갖더라도 형태소 경계 정보와 품사 정보에 따라 적용되는 규칙이 다르다는 것과 어절 경계에서 나타나는 현상이 일부 규칙으로 제한된다는 것이다.

본 논문에서는 '음소 변동 규칙'과 '변이음 규칙'을 단계별로 적용한 발음열 자동 생성 시스템[7]을 사용하였다. 언어학적 지식을 기반으로 한국어에서 발생하는 음운 변화 현상과 문교부에서 제정한 표준어 규정 표준 발음법을 참고하여 대표적인 20개의 음소 변동 규칙으로 정리하고, 해당 음소 문맥 별로 총 816개의 세부 규칙으로 나누어 적용하였다. 이 중 형태소 경계에서 적용되는 규칙 중 형태소 내부에서 적용되는 규칙과 달리 47개의 음소 문맥에서 서로 다른 규칙이 변별적으로 적용되었다.

2.3 데이터에 기반한 형태소 경계에서의 음운 변화 현상 반영

영어와 같은 서양언어에서는 띄어쓰기 단위인 단어(word)를 기본 단위로 사용하지만, 한국어는 언어학적 특성상 단어 단위의 정의가 모호하다. 한국어 문장은 어절 단위로 띄어쓰기를 하며, 어절 안에는 띄어쓰기 및 구분 기호가 없이 결합된 형태소들로 구성된다. 따라서 어휘 사전을 구성하는데 있어 모든 형태소들의 가능한 조합을 표제어로 등록하는 것은 효율적이지 못하기 때문에, 대부분의 한국어 연속음성인식 시스템은 형태소를 사전의 표제어 및 디코딩 단위로 사용한다[8].

고립단어 인식에서는 주어진 어휘 목록에 해당하는 발음 표기열을 발음사전에 기재하면 되지만, 연속음성인식의 경우 동시조음 현상을 포함하여 음운 변화 현상이 형태소 경계에서도 발생하게 된다. 이때 어휘 목록을 바탕으로 발음열을 생성하게 되면 시작과 끝 부분에서 발생하는 변화 현상을 반영하기 힘들다. 연속음성의 경우 특히 형태소 경계에서의 발음 변화 현상이 매우 다양하게 일어난다. 형태소 경계는 주로 복합명사나 조사, 접미사 그리고 어미 등의 결합에 의해 생겨나며, 특히 경음화와 같은 일부 규칙은 비록 같은 음소 문맥일지라도 형태소 내부, 형태소 경계, 그리고 어절 경계에서의 발음이 다르게 실현되므로 위치 정보에 따라 발음 변화 현상이 서로 다르게 모델링 되어야 한다.

형태소 경계에서의 현상을 반영하기 위한 방법으로는 음소 문맥 정보와 형태

소 접속정보 등을 활용하여 탐색 공간을 확장하는 방법과 경계에서 발생 가능한 음소문맥들을 예측하여 발음사전에 추가하는 방법을 들 수 있다. 이와 같이 인식용 언어모델 네트워크를 확장하거나 사전에 예측 가능한 발음열을 추가하는 경우, 그 후보 수가 급격하게 증가하게 됨에 따라 오히려 시스템 부하가 발생할 수도 있다. 예측 가능한 현상들을 분석하여 반영하는 것이 좋은 성능을 가져올 수도 있겠지만 모든 정보를 사용한다고 반드시 좋은 결과만을 얻을 수 있는 것은 아니다.

본 논문에서는 학습용 텍스트 코퍼스를 이용하여 해당 어휘가 주어진 문장 내 환경에 따라 실제 발생 가능한 발음열로 구성된 데이터 기반의 발음사전을 생성하였다. 한국어의 특징을 잘 반영하기 위해 주어진 입력 문장에 대해 형태소 분석을 수행하여 품사 정보와 경계 정보를 구하였다. 최종적으로 문장 단위의 형태소 분석 결과를 입력으로 받아 발음열 생성기를 이용하여 형태소 단위의 어휘 목록과 발음사전을 자동 생성하였다. 기존 연구[7]에서도 형태소 경계에 문맥 종속적인 다중 발음사전을 사용한 경우 단일 발음사전과 비교하여 1.45%(상대적 7.9%) 정도의 WER를 감소시켰다. 전반적으로 형태소 경계 정보와 품사 범주를 고려한 음운 변화 모델링을 통해 향상된 인식 결과를 얻을 수 있었다.

2.4 음향 모델 측면에서의 발음 변화 현상 모델링

일반적인 음향 모델 학습 과정은 다음과 같다. 먼저 임의로 초기 HMM 모델을 정의한 후, 학습용 음성 데이터와 해당 음소 전사열을 이용하여 음소 문맥을 반영하지 않은(context-independent) 모노폰을 학습한다.

보다 정확한 학습을 수행하기 위해, 학습용 음성 데이터를 모노폰 단위로 학습된 1차 음향 모델을 이용하여 음소 단위 인식을 수행한다. 이때 인식 목록은 전체 어휘를 대상으로 하지 않고, 주어진 문장의 정답을 주고 해당 단어에 해당하는 대체 발음열들 중에서 likelihood가 가장 큰 음소열을 결과값으로 출력한다. 이러한 과정을 forced alignment라고 하며, 발음사전의 다중 발음열 중 적합한 음소열을 찾아 음성 데이터와 음향 모델간의 최적화된 정렬을 수행하고 음향 모델을 재학습한다. 이때 사용되는 학습용 발음사전에는 다양한 발음 변화를 포함시킴으로써 간접적으로 변화 현상을 반영한다.

다음은 재학습 과정으로 모노폰 단위로는 조음 현상을 반영하기 어렵기 때문에 트라이폰으로 확장하여 학습을 수행한다. 전사 파일로부터 트라이폰 목록을 만들고 먼저 학습된 모노폰으로부터 초기 트라이폰 모델을 만들게 된다. 이때 트라이폰은 단어 내에서 발생된 것들만 학습되기 때문에 학습에서 나오지 않는 트라이폰에 대한 보완은 임의의 두 트라이폰 사이에 출력 확률 분포가 유사한 상태의 경우 공유하도록 하는 tied-state 트라이폰[9]을 사용하였다. 여기에 공유되는 음소들은 음향학적 결정 트리(phonetic decision tree)를 구성한 후 임의의 음소 모델에

대한 음향학적인 특성을 판단하게 하는 하향식 방법을 채택하였다. 본 논문에서는 한국어가 가지는 음성학 및 음운론적 특성을 반영하여 조음 방식과 조음 위치에 따라 분류된 212 종류의 질문 집합을 사용하였다. 발음사전의 모든 표제어들로부터 발생 가능한 트라이폰 목록(seen & unseen triphone)을 만들어 낸 후, 음향학적 결정 트리를 이용하여 실제 생성된 트라이폰과 가장 가까운 값을 공유하는 트라이폰을 재구성하여 음향 모델을 학습한다.

2.5 한국어의 언어적 특성에 따른 고려 사항

이상적으로는 다양한 종류의 변화 스타일을 시스템 내부에서 모두 다루어야 하지만, 현재 state-of-the-art 인식기로는 한 시스템 내에 모델링 하는 것이 쉽지 않을 뿐만 아니라 세분화된 현상들을 한꺼번에 해결하기 힘들다. 따라서 별도의 모델링을 통하여 시스템 구성에 따라 단계별로 적용하는 방식으로 접근한다.

한국어는 불규칙 활용 및 축약 현상을 포함하여 여러 음운 현상에 의하여 하나의 형태소가 여러 가지 음소열로 발음될 수 있고, 하나의 음소열이 여러 형태소의 발음을 대표 할 수도 있다. 따라서 언어학적 특성상 사전의 표제어와 실제 발음 정보 사이에는 다음과 같은 고려 사항이 필요하다.

첫 번째로는 같은 표제어(spelling)이지만 서로 다른 의미를 갖는 동형이의어(homographs)의 경우를 들 수 있다. 의미에 따라 발음이 다른 경우로 시제에 따라 다르게 실현되는 'read'나 한국어의 경우 '사적'(historic-史的, personal-私的), '성적'(score-成績, sexual-性的), '대가'(authority-大家, price-代價) 등을 예로 들 수 있다. 이런 경우에는 일반적으로 사전에는 하나의 표제어만을 포함하고 간단하게 발음사전에 발음열을 추가하여 처리한다.

두 번째로는 서로 다른 의미를 갖는 단어들에 대해 하나의 엔트리를 갖지만, 같은 철자와 발음을 공유하는 동음이의어(homonyms)을 들 수 있다. 예를 들면 'right'는 방향과 긍정의 표현을 갖으며, 한국어의 경우 여러 의미를 내포하는 '가다', '묻다', '배' 등을 들 수 있다. 이러한 동음이의어는 단어 의미정보가 언어모델이나 의미적 파싱 정보에 영향을 미치기 때문에 문제가 될 수 있다. 다중 표제어를 갖는 인식용 사전으로 구성하는 경우, 언어모델에서 중의성을 해결 할 수 있지만 사전 엔트리 수의 증감과 더불어 데이터 부족 문제를 가져올 수 있다.

세 번째로는 서로 다른 의미를 갖는 단어들이지만 같은 발음열을 공유하는 경우로서 일반적으로 서로 다른 철자를 갖는 동음이자(homophones)를 들 수 있다. 예를 들면 두 단어 'two' 와 'too'는 서로 다른 표제어를 갖지만 발음은 같은 경우지만, 서로 다른 엔트리로 기재하여 언어모델에서 각각 다르게 적용함으로써 인식과정에서 해결할 수 있도록 하였다. 이는 인간의 음성인식과 같은 이치로 주어진 문맥에 따라 해당 단어에 맞게 인식하게 된다. 한국어에서는 주로 발음은 같으나

글자가 서로 다른 표기를 갖는 용언구에서 많이 볼 수 있다. 예를 들면, '같', '갖', '갔', '갗' 등의 경우 표준 발음열 /K AA TQ/를 갖는다. 반면 뒤에 오는 어미에 따라 서로 다른 발음열을 갖기도 하는데, '같어', '갖어', '갔어' 등과 같이 활용에 따라 다르게 실현되는 경우에는 해당 변형 발음열 /K AA TH/, /K AA ZH/, /K AA SS/을 사전에 추가 기재함으로써 해결한다.

때로는 자동 음성인식 시스템에서 이러한 지식들을 태스크와 대화에 의존적인 언어모델로 통합하여 사용하기도 한다. 다중을 허용하는 대신 변별적이여야 하며, 다중 발음열을 사용하는 경우 발음 확률값을 이용하여 혼잡도를 감소시키는 데 사용하기도 한다. 발생 확률이 낮은 발음의 경우 발음에 대한 확률값을 발음사전 대신에 때로는 언어모델의 일부로 정의되기도 한다.

3. 음성인식용 발음사전 생성 및 최적화

3.1 경계 문맥에 종속적인 다중 발음사전 생성

학습 과정에서는 학습용 음성 데이터에서 발생 가능한 모든 대체 발음열을 포함하는 발음사전을 사용하지만, 인식 과정에서는 추가 어휘가 발생할 뿐만 아니라 응용 영역에 따라 인식 어휘가 바뀌게 되므로 인식 성능과 속도를 고려하여 인식용 발음사전을 준비하여야 한다. 경계 정보에서 발생하는 발음 변화를 반영하기 위해서는 실제 문장에 나타나는 문맥 정보를 이용하여 구축하는 것이 가능성 높은 정보를 제공하며, 텍스트 양이 많을수록 보다 정교한 정보를 구할 수 있을 것이다. 본 논문에서는 방송뉴스 전사 문장과 신문 기사 및 교과서를 포함한 텍스트 문서 등에서 수집한 약 275,686 문장을 사용하였다. 앞 장에서 기술한 바와 같이 자동으로 어휘 리스트와 발음열 정보를 구하였다.

<표 1> 음성인식용 발음사전 엔트리 예제

표제어	발음열
한국	H AA NG G UW KQ
한국 (2)	H AA NG G UW G
십	S IY PQ
싶	S IY PQ
싶 (2)	S IY PH
고	K OW
고 (2)	KK OW
고 (3)	G OW
고 (4)	KH OW
은	WW N
케이비에스	K EY IY B IY EY SS WW

<표 1>은 다중 발음열을 갖는 발음사전의 예이다. 단어 뒤에 숫자가 붙은 경우는 추가된 대체 발음열을 나타낸다. 강세(stress)가 의미 분별에 크게 영향을 미치는 영어와는 달리 한국어에서는 문맥에 따른 변이음 현상이 오히려 많은 영향을 미치게 된다. 주로 표제어의 초성에 변이음이 존재하면 그 수만큼 엔트리 수가 증가하게 된다. 또한 조사나 어미와 같은 연결 어휘의 문맥 정보에 따라 음운 변화가 발생하게 되면 엔트리 수가 추가된다. 다른 언어와는 달리 한국어에서는 일반적인 현상이기 때문에, 이를 고려하여 사전 크기를 조절해야 한다.

일반적으로 표제어 대 발음열 비율이 1.2 배 정도인 다른 언어들과는 달리 본 논문에서 사용된 코퍼스에서 발생한 전체 어휘를 대상으로는 약 1.32배 정도, 10회 이상 발생한 어휘 목록을 대상으로는 약 1.8배 정도의 비율이 되는 것을 확인할 수 있었다. 분석결과 "교육"의 경우 음소 문맥과 형태소 결합 정보에 따라 9가지의 다양한 대체 발음으로 표현되는 것을 확인할 수 있었다. 영어 switchboard 영역에 대한 실험을 예로 들면, 실제 음성 데이터로부터 "the"에 대하여 36개의 서로 다른 발음열이 발견되었고, 실험에는 38 종류의 발음열이 학습용으로 사용하였으며, 이 중 학습 집합에 있던 절반 정도가 테스트 집합에서 관찰되어졌다. 그러나 관찰된 변화 현상을 모두 사전에 포함하면 서로 다른 35개의 단어에서도 발음이 중복되는 문제가 발생하게 된다[10].

3.2 확률 기반의 발음사전 pruning 기법

발음 변화 모델링의 도전적인 과제 중 하나는 어떤 발음을 선택하여 자동 음성 인식 시스템의 성능을 최대로 갖게 하는 발음사전을 구성하느냐이다. 대부분의 발음사전은 언어학적 지식을 기반으로 작성되지만 인식 성능의 관점에서는 고려되지 않았다. 본 논문에서는 주어진 학습 데이터를 기반으로 음성 인식의 성능 향상에 도움이 되는 최적의 발음열을 찾아내는 방식에 중점을 두었다.

거의 사용되지 않는 발음열은 정확성 보다 오히려 오류를 야기하게 되며, 엔트리 사이에 음향학적 혼잡도가 증가되어 인식 성능의 저하를 예측하게 한다. 이와 관련하여 정적 표준 발음사전에 어떻게 발음 변화열을 추가해야 인식률을 증가시킬 수 있을 것인가에 대해 다양한 연구들이 진행되어 왔다[11]. 음소문맥 종속 다중 발음사전은 주어진 해당 어휘가 갖는 다양한 변화 현상을 반영할 수 있는 반면 발음열의 종류가 늘어나게 되면서 사전 크기가 점차 증가한다. 이를 해결하는 방법으로 사전에 기재될 대체 발음열을 선택하기 위해 pruning 기법을 활용한다. 충분히 큰 코퍼스에서 가장 많이 관찰되는 변화를 선택하는 것이 일반적이며, 이러한 pruning을 위한 기준으로는 카운트 기반(count-based), 확률 기반(probability-based), 또는 엔트로피 기반(entropy-based)으로 발음열의 후보열을 결정하는 다양한 접근방법 등이 기존 연구들에서 사용되었다[3][12].

기존의 전통적인 pruning scheme에서는 대부분 확률기반의 접근방법이 좋은 성능을 보여 왔다. 적용 규칙에 대한 확률이나 발음에 대한 확률을 기준으로 threshold 값에 따라 추가할 발음열의 수를 조절할 수 있다. 본 논문에서는 해당 어휘에 대해 예측 가능한 발음열의 확률값을 기준으로 튜닝된 threshold 보다 큰 값을 갖는 대체 발음열을 대상으로 사전에 추가 기재한다. 발음 빈도수에 의한 기준은 다음과 같이 PF(pronunciation frequency)로 정의하고, 주어진 단어 w_i 일 때 발음열 v_j로 실현되는 정규화된 확률값은 식 (3)과 같이 계산된다.

$$pf_{ij} = P(v_j | w_i) \qquad (3)$$

해당 어휘를 기준으로 발생한 발음 빈도수에 기반한 순위 전략(pronunciation frequency-based ranking strategy)으로 발음열 후보열들은 위의 식 (3)에 의해 계산된 값을 기준으로 pruning이 적용된다.

3.3 단어빈도를 가중치로 적용한 확률기반 사전 pruning 기법

화자의 다양성(speaker variation) 문제는 화자별 독립 발음사전을 두어 해결할 수 있지만, 경계에서 발생하는 음운 변화의 다양성은 학습 코퍼스에서 발생된 현상들을 통계적으로 처리하여 하나의 사전에 반영하는 것이 신뢰할만한 결과를 얻을 것이다. 중국어의 경우 같은 발음을 공유하지만 서로 다른 의미와 문자열을 갖는 동음이의어가 주를 이루기 때문에 단어간 혼잡도를 줄이기 위해 IWF(inverse word frequency)를 고려하는 것이 일반적이다[13]. 2.5절에서 본 바와 같이 한국어는 대조적으로 동일한 어휘 목록에 대해 서로 다른 발음들을 갖는 경우가 주를 이루게 됨으로 언어적 특성에 맞게 적용하여야 한다.

한국어에서의 기능어는 어미, 조사, 단위성 의존명사 등과 같은 품사들로서 영어의 기능어 분류와는 달리 주로 접미사(suffix)에 해당한다. 별도의 기능어 모델링으로 해결할 수도 있지만 코퍼스에서 많이 발생하는 만큼 관찰되는 발음의 다양성을 보장할 수 있다. 연속어에서는 일반적으로 고빈도 단어일수록 더 많은 수의 발음 변화열이 발생하게 된다. 기존의 확률 기반 pruning 기법은 표제어간의 가중치는 고려하지 않고 주어진 단어 내부에서의 발음 빈도만을 고려했기 때문에 상대적으로 낮은 확률을 갖는 발음열이 선택되지 않는 경우가 발생된다. 이와 같이 어휘 규모가 증가함에 따라 다양한 현상들이 관찰되는 변화량을 고려하여 전반적으로 사전 엔트리에 대한 발음 종류에 대한 최적화가 필요하게 된다. 기존 연구[14]에서는 주어진 단어 w_i가 가질 수 있는 최대 발음열의 개수를 제한하는데 log-count pruning을 사용하여 발음사전을 구성하는 연구가 시도되었다.

본 논문에서는 해당 학습 데이터에서 관찰된 발음열의 종류와 어휘 발생빈도에 대하여 log-count를 적용하여 제안하는 해당 단어 w_i에 대한 단어 빈도수(GWF:

Generalized Word Frequency)는 다음 수식 (4)와 같이 정의한다.

$$gwf_i = \log_{10}(baseform(w_i) \cdot count(w_i)) \quad (4)$$

여기서 baseform(w_i)는 해당 표제어가 코퍼스에서 관측되는 발음열 종류의 가짓수를 표현하며, 변화에 대한 비율을 고려하여 단어별 발생 빈도수와 더불어 가중치로 적용한다. 본 논문에서는 발음사전 최적화에 사용되는 confidence score값은 앞서 구한 PF와 GWF를 결합하여 최종적으로 $pf_{ij} \cdot gwf_i$ 로 정의한다. 인식용 발음사전의 최적화를 위해 [14]의 연구에서는 해당 표제어가 가질 수 있는 추가 발음열의 수를 정해진 개수로 한정하였지만, 본 연구에서는 엔트리 단위로 계산된 confidence score를 기준으로 pruning을 수행한다. 주어진 단어 w_i에 대해 해당 발음열 v_j의 score는 앞서 제안한 PF · GWF 값을 기준으로 계산되며, 주어진 threshold 값을 넘지 못하는 발음열 목록들은 베이스라인 사전에서 제거함으로써 발음사전을 최적화한다.

4. 실험 결과 및 분석

4.1 실험 환경 및 베이스라인 시스템

연속 HMM을 기반으로 한 화자 독립 시스템인 HTK(Hidden Markov Toolkit)[9]를 사용하여 음성인식 실험을 수행하였다. 본 실험에는 16kHz로 샘플링 된 음성신호에 대해 25ms의 해밍 윈도우를 사용하여 프레임 분석을 한 후, 프레임 단위로 13차의 MFCC와 그 delta 및 delta-delta 계수를 포함한 39차의 벡터를 사용하였다. HMM 모델은 기본적으로 6개의 gaussian 분포를 사용하였으며, 별도로 12개의 gaussian을 갖는 잡음(garbage) 모델을 사용하였다.

기본 음향 모델은 기존의 3만 3천 형태소급의 한국어 연속음성인식 시스템 구성[7]을 위한 음향모델을 기반으로 새로운 인식 영역에 해당하는 방송 뉴스 데이터를 사용하여 적응을 수행하였다. 실험 대상으로는 1997년 2월부터 1998년 12월 사이 50일 분의 KBS 9시 뉴스 중에서 앵커와 기자에 해당하는 음성 부분만 선별 수집한 25시간 분량의 한국어 방송 뉴스 데이터를 사용하였다. 실험에는 두 집합 중에서도 잡음 정도에 따라 clean 음성 부분만을 대상으로 삼았다. 총 발화수는 14.4k, 형태소 단위로 어휘 크기는 21.7k이며, 인식 실험에는 5,573 발화 문장의 clean 레벨 데이터만을 음향 모델 적응과 테스트에 사용하였다. 이 중 성능 평가에 사용한 테스트 발화는 총 311 문장이다.

4.2 방송 뉴스 특성에 따른 실험 환경

입력 철자 전사 문장을 이용하여 통계적 방법으로 학습 코퍼스에서 적절한 어휘 목록을 추출하고, 발생 가능한 모든 음소 전사열(phonetic transcriptions)을 구한다. 이를 기반으로 어휘 목록과 발음사전을 자동으로 생성하고, 언어모델 생성을 위한 학습 코퍼스로도 사용한다. 이때 사전의 구성에 따라 성능에 영향을 미치게 되는데, 인식 가능한 사전의 크기가 작을수록 리스트 선택이 명료해지기 때문에 인식률이 향상되는 요인이 되지만, 해당 리스트 중에 인식 단어가 없는 경우에는 인식이 불가능하기 때문에 오류를 초래하게 된다. 따라서 적절한 coverage를 갖는 어휘 목록을 설계하는 것이 중요하다.

본 논문에서 사용한 학습용 텍스트 코퍼스는 총 264,721 발화 문장으로 고유 형태소 수는 64,273이고, 한 발화 문장 당 평균 11.88 어절과 25.25 형태소를 수록한 데이터베이스이다. 학습 코퍼스를 기반으로 적정 수준의 coverage 구성을 위해 분석한 결과, 20k 어휘 목록 구성시 98.4%의 coverage를 나타냈고, 25k 어휘 목록에서는 99%의 coverage를 보이고 있다. [5][8]에서도 사전의 크기를 정할 때 WSJ, LeMond 코퍼스의 사전과 유사한 coverage를 갖도록 하였다. 참고로 WSJ 코퍼스에 대한 20k 사전의 coverage는 97.5%, LeMond 코퍼스에 대한 40k 사전은 coverage가 97.6%이다. 본 실험에서는 학습 코퍼스에서의 어휘 발생 빈도가 10회 이상을 기준으로 한 20.4k 어휘 목록(coverage 98.32%)을 베이스라인으로 사용하였다. 다음 장에서는 다양한 사전 pruning 기법을 적용한 인식용 발음사전의 성능을 비교 평가한다.

4.3 성능 평가

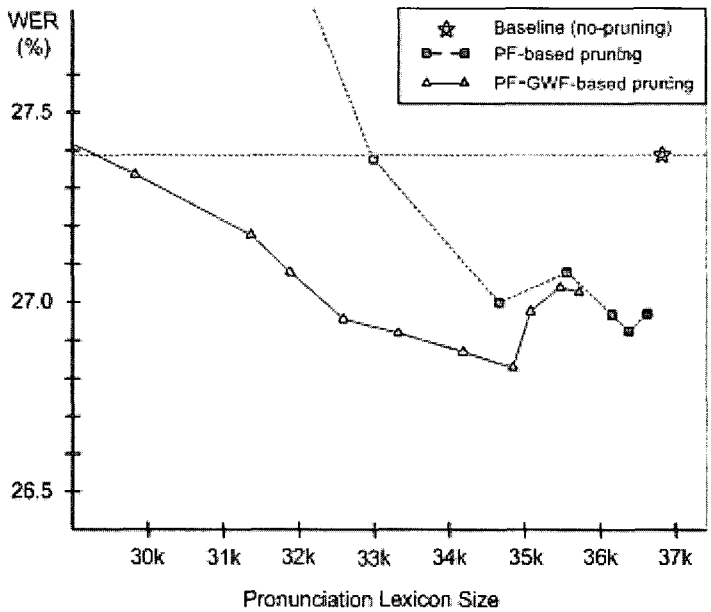
본 실험에서는 20.4k 크기를 갖는 어휘 목록을 바탕으로 서로 다른 인식용 발음사전을 구성하여 인식 실험을 수행하였다. 성능 평가를 위한 두 가지 척도로는 형태소 단위의 에러 발생 수치를 말해주는 WER(Word Error Rate)과 발음사전 크기(Lexicon Size)를 사용하였다. 발음사전 구성에 따른 최종 실험 결과는 <표 2>와 같다. 사전 pruning 기법을 적용한 경우에는 인식률을 평가 기준으로 최대 성능일 때의 수치와 turning된 threshold 값에 대한 정보를 기록하였다.

사전 pruning 기법을 적용한 경우에는 인식률을 평가 기준으로 최대 성능일 때의 수치와 turning된 threshold 값에 대한 정보를 기록하였다. <표 2>의 1번째 항목에 해당하는 단일 발음사전은 형태소 경계에서 발생하는 음운 변화는 제외하고, 표제어 내부의 좌우 음소 문맥만을 고려하여 생성한 발음사전으로 사전 크기는 표제어 수와 동일하지만 다중 발음사전과 비교하면 WER가 12.9%나 현저한 차이를 확인할 수 있다. 또한 해당 표제어에 대해 최적의 대표 발음열 하나만을 갖는

<표 2> 다양한 Pruning 기법을 이용한 발음사전 모델링 성능 비교

발음사전	엔트리 수 (발음사전 크기)	단일 발음사전 대비 비율	WER (%)
단일 발음사전 (형태소 경계 변화 미반영)	20,375	1	40.29
1-best 발음사전 (형태소 경계 변화 반영)	20,375	1	31.86
베이스라인 (다중 발음사전 - no pruning)	36,822	1.8	27.39
PF (*threshold* : 0.003)	36,181	1.78	26.93
PF · GWF (*threshold* : 0.015)	34,848	1.71	26.83

1-best 발음사전의 경우에도 다중 발음사전과 비교하여 4.47% 정도 인식률 차이가 나는 것을 확인할 수 있다. 결과적으로 형태소 경계에서의 음운 변화 모델링이 반영된 다중 발음사전에서 향상된 인식률을 얻을 수 있었다. 본 실험에서의 베이스라인은 앞서 설명한 형태소 경계 정보와 품사 범주를 모두 고려한 세부적인 발음 변화 모델링 및 문맥 종속적인 다중 발음사전을 기반으로 한다. 하나의 표제어 당 1.8개의 발음 변화열을 포함하는 베이스라인 인식용 발음사전의 WER는 27.39%이다.



<그림 1> 제안한 pruning 기법에 따른 발음사전 크기 및 인식 성능 비교

<그림 1>은 베이스라인 사전을 기반으로 앞서 제안한 순위 전략에 의한 confidence score값을 기준으로 적용된 사전 pruning에 따른 실험 결과이다. 각각 PF는 발음 빈도에 따라 순위화 하며, PF · GWF는 발음과 단어 빈도가 고려된 순위

전략을 의미한다. <그림 1>에서 보는 바와 같이 PF · GWF 방식이 기존의 PF 방식과 비교해 봤을 때, 사전의 크기 감소에 따른 인식 성능의 변화가 안정적임을 확인할 수 있다. PF 방식의 경우, 해당 단어에 갖는 발음 변화가 다양할수록 상대적으로 적게 발생되는 발음열의 score 값이 전체적으로 낮아질 수밖에 없다. 구체적으로 "정책"과 "가입"을 예로 들면, 코퍼스에서 발견된 어휘 빈도수(count(w_i))는 각각 3220회와 545회이며, 관찰된 발음 변화열의 종류(baseform(w_i))는 각각 12가지와 8가지이다. PF 방식으로 pruning을 수행하면 score를 기준으로 각각 상위 7개의 발음열을 수록하게 되지만, PF · GWF의 경우에는 각각 8개와 6개의 발음열이 사전에 수록된다.

PF · GWF 방식으로 단어 빈도수와 발음열의 변화 종류를 보상 적용하여 pruning을 수행한 결과, 사전 비율이 1.5배(사전 크기: 29,846) 미만으로 줄어도 베이스라인 성능보다 낮은 에러율을 유지하는 것을 확인할 수 있었다. 반면 사전의 크기가 점차 감소함에 따라 어느 시점 이후로는 WER가 증가하게 되며, 1-best 단일 발음사전의 오류율 수준까지 점차 증가하게 된다. 이는 pruning에 의해 발음열이 삭제되면서 발음 변화의 다양성이 감소함에 따라 나타나는 현상으로 볼 수 있다. 실험 분석 결과, PF · GWF 방식의 경우 WER가 26.83%일 때 최고의 인식율을 보였으며, 최종적으로 베이스라인을 기준으로 약 0.56%의 에러율과 5.36%의 사전 크기를 추가로 감소하였다. 이 실험을 통해 최적의 발음열을 선택하는 방식이 시스템의 성능을 향상시키는 것을 확인할 수 있었다.

5. 결 론

본 논문에서는 한국어 대어휘 연속음성 인식에 필요한 형태소 기반의 발음사전을 구성하고 최적화 하는 과정을 제시하였다. 형태소 경계에서의 발음 변화 현상에 초점을 맞추었으며, 문맥 종속적인 다중 발음사전을 기반으로 형태소 결합 정보와 품사의 종류에 따라 형태소 내부와 형태소 경계에서 발생하는 음운 변화 규칙을 세분화하여 모델링 함으로써 성능 향상을 꾀하였다. 이와 함께 가능한 발음 변이를 사전에 추가함으로서 생기는 일부 어휘들에서의 혼잡도 증가와 이로 인해 유발되는 오류 발생 문제를 해결하기 위해 최적의 발음 변화열을 선택적으로 적용하는 사전 pruning 기법을 적용하였다. 본 논문에서는 학습용 코퍼스를 기반으로 해당 어휘당 발음열의 발생 빈도에 대한 확률과 단어 빈도와 발음 변화열의 다양성을 보상 적용한 PF · GWF 방식의 순위 전략에 따른 pruning 방법을 제안하였다. 제안된 pruning 기법을 적용하여 최적화된 인식용 발음사전에 대한 실험 결과, 사전 크기의 감소와 인식률 향상이라는 결과를 얻을 수 있었다. 향후 과제로는 규칙이나 학습 데이터에 기반한 변화 현상뿐만 아니라, 화자의 발화 스타일

을 포함하여 규칙에 의해 생성되지는 않지만 일반적으로 많이 발생되는 발음 정보를 반영하는 연구가 필요하다. 실제 음성 데이터로부터 해당 정보를 구하여 사전에 반영하거나 적응화 하는 방법과 음성인식 시스템에서 내부적으로 공유할 수 있는 다양한 발음 표현을 구조적으로 해결 할 수 있는 방법에 대한 연구를 수행할 예정이다.

참 고 문 헌

[1] M. Weintraub, K. Taussig, K. Hunicke-Smith and A. Snodgrass, "Effect of speaking style on LVCSR performance," *Proc. of the Spoken Language Processing(ICSLP '96)*, pp. 1036-1038, 1996.

[2] H. Strik, C. Cucchiarini, "Modeling pronunciation variation for ASR: a survey of the literature," *Speech Communication*, Vol. 29, pp.225-246, 1999.

[3] Q. Yang and J.-P. Martens, "Data-driven lexical modeling of pronunciation variations for ASR," *Proc. of the Spoken Language Processing(ICSLP '00)*, pp. 750-753, 2000.

[4] M. Wester, "Pronunciation modeling for ASR knowledge-based and data-derived methods," *Computer Speech and Language*, Vol. 17, pp. 69-85, 2003.

[5] P. C. Woodland, C. J. Leggetter *et al.*, "The 1994 HTK large vocabulary speech recognition system," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '95)*, pp. 73-76, 1995.

[6] 장영성, 정민화, "음운 변화 역추적 규칙을 적용한 한국어 음성 언어 형태소 분석 모델," 봄 *한국정보과학회*, vol. 24, no. 1, pp. 483-486, 1996.

[7] 정민화, 이경님, "한국어 연속음성인식 시스템 구현을 위한 형태소 단위의 발음 변화 모델링", *말소리지*, pp. 107-121, 3월, 2004년.

[8] O.-W. Kwon and J. Park, "Korean large vocabulary continuous speech recognition with morpheme-based recognition units," *Speech Communication*, Vol. 39, pp. 287-300, 2003.

[9] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, *The HTK Book* (for HTK Version 3.2), *EntropicCambridge Research Laboratory*, 2002.

[10] D. McAllaster, L. Gillick, F. Scattone and M. Newman, "Fabricating conversational speech data with acoustic models: A program to examine model-data mismatch," *Proc. of the Spoken Language Processing(ICSLP '98)*, pp. 1847-1850, 1998.

[11] S. Deligne, B. Maison and R. Gopinath, "Automatic generation and selection of multiple pronunciations for dynamic vocabularies," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP '01)*, pp. 565-568, 2001.

[12] E. Fosler-Lussier and N. Morgan, "Effects of speaking rate and word frequency on pronunciation in conversational speech," *Speech Communication*, Vol. 29, pp. 137-158, 1999.

[13] M. Tsai, F. Chou and L. Lee, "Improved pronunciation modeling by properly integrating

better approaches for baseform generation, ranking and pruning," *Proc. of Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology(PMLA)*, pp. 77-82, 2000.

[14] E. Fosler-Lussier and G. Williams, "Not just what, but also when: Guided automatic pronunciation modeling for broadcast news", *In: DARPA Broadcast News Workshop*, pp. 171-174, 1999.

접수일자: 2005년 8월 22일
게재결정: 2005년 9월 23일

▶ 이경님(Kyong-Nim Lee)
주소: 136-742 서울시 성북구 동선동 3가 249-1 성신여자대학교
소속: 자연과학대학 미디어정보학부
전화: 02) 920-7615
FAX: 02) 920-2250
E-mail: knlee@sungshin.ac.kr

▶ 정민화(Minhwa Chung)
주소: 서울시 관악구 신림9동 산 56-1 서울대학교
소속: 인문대학 언어학과
전화: 02) 880-9195
FAX: 02) 882-2451
E-mail: mchung@snu.ac.kr