발음 변이의 발음사전 포함 결정 조건을 통한 발음사전 최적화

전 재 훈[1], 정 민 화[2]
[1]서강대학교 컴퓨터학과, [2]서울대학교 언어학과

Pronunciation Lexicon Optimization with Applying Variant Selection Criteria

Je Hun Jeon[1] and Minhwa Chung[2]
[1]Department of Computer Science, Sogang University
[2]Department of Linguistics, Seoul National University
E-mail : {jhjeon, mchung}@snu.ac.kr

Abstract

This paper describes how a domain dependent pronunciation lexicon is generated and optimized for Korean large vocabulary continuous speech recognition(LVCSR). At the level of lexicon, pronunciation variations are usually modeled by adding pronunciation variants to the lexicon. We propose the criteria for selecting appropriate pronunciation variants in lexicon: (i) likelihood and (ii) frequency factors to select variants. Our experiment is conducted in three steps. First, the variants are generated with knowledge-based rules. Second, we generate a domain dependent lexicon which includes various numbers of pronunciation variants based on the proposed criteria. Finally, the WERs and RTFs are examined with each lexicon. In the experiment, 0.72% WER reduction is obtained by introducing the variants pruning criteria. Furthermore, RTF is not deteriorated although the average number of variants is higher than that of compared lexica.

I. 서론

하나의 단어가 발음될 때 사람에 따라 또는 주위 상황에 다르게 발음 된다. 이러한 발음 변이들은 자동음성 인식의 인식률 저하의 주요한 원인 중 하나로 자리 잡고 있다. 특히 자연스러운 발화에 있어서 발음 변이가 더 빈번 하게발생 하므로, 발음 변이를 모델링 하는 것은 인식률의 향상에 크게 기여 할 수 있을 것이다.

발음 변이의 모델링에는 주로 2 가지 방법론이 제시되어 왔다. 이러한 방법론은 정보의 종류에 따라 학습에 의한 방법과 지식 기반에 의한 방법으로[1] 분류 될 수 있다. 학습에 의한 방법은 음성 신호로부터 발음 변이를 추출 하는 상향식 방법이고, 지식 기반 방법은 현재 활용 가능한 언어학적 지식을 이용하는 하향식 방법이다. 이들 방법론들의 각각의 장단점으로 인해 상호간의 우위를 결정할 수 없지만, 지식 기반의 방법론은 범용적인 용도로서의 장점을 가지지만 자연스러운 발화의 변의 정보 표현에 단점을 가질 수 있다.

대부분의 음성인식 시스템은 발음 변이를 반영하기 위해 다중 발음 사전을 사용하게 된다. 발음 사전의 입장에서의 발음 변의 모델링은 발화 가능성이 있는 발음 변이들 중 적당한 변이를 사전에 포함하는 것이다. 가능한 모든 발음 변이를 발음 사전에 포함하는 것은 특정 단어의 인식률의 향상에 도움을 줄 수도 있지만, 인식기의 탐색 공간의 크기와 복잡도를 높여 인식률 하락의 요인이 될 수 도 있다. 따라서 가능한 모든 발음 변이중 적절한 발음변이를 선택해 발음 사전에 포함 하는 것이 인식률 향상에 필수적이다.

본 논문에서는 발음 변이들 중 발음 사전 포함 결정 조건을 제안하여 인식 성능 향상을 이루도록 하였다. 제안하는 결정 조건 첫 번째는, 발음 모델링에서 생성

된 발음 변이 적합성 값이고, 두 번째는 해당 코퍼스에서의 특정 발음 변이가 나타난 빈도 값이다. 본 논문에서는 결정조건의 타당성을 검증하기 위하여 한국어 대용량 인식 실험을 통해 여러 조건에서의 인식률의 변화에 대해에 대해 관찰하였고, 또한 발음 사전 변화에 따른 RTF(Real Time Factor)를 측정하여 제안된 방법에 의해 생성된 발음 사전과 기본 발음 사전의 성능 비교를 하였다.

본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 발음 모델링에 사용된 발음 변환 규칙의 구조에 대해 간단히 설명하고, 3장에서는 발음 사전 생성 방법과 제안하는 결정조건에 대해 자세히 설명하고, 4장에서는 다양한 발음 사전을 이용한 한국어 대용량 음성인식 실험 및 결과를 검토한 후, 5장에서 결론을 맺는다.

II. 발음 변화 규칙 정의

발음 변이 생성을 위해 사용된 규칙은 지식 기반의 규칙이다[2]. 한국어 발음 변화에 대한 언어학적 지식을 기반으로 발음 변화 규칙을 작성하였고, 규칙의 구조는 아래와 같다.

$$r : L\underline{G}R \rightarrow T with\ P_r \quad (1)$$

이 규칙에서 변화 될 자소 그룹 G는 좌측 음소 문맥 L과 우측 음소 문맥 R에서 의해 발음 T로 변화 하고, 그 변화의 적합성 값은 P_r로 표현 된다는 것이다. 자소 그룹 G는 자음 변화일 경우 연속하는 모든 자음을 하나의 그룹으로 나타내며, 모음일 경우 하나의 모음이 G로 구성 된다. 발음 T는 자소 그룹 G에 따라 여러 개의 음소로 변화 되어 질 수 있다.

발음 변화 규칙은 크게 13개의 규칙으로 구성되어 있다; (1) 음절말 중화, (2) 자음군 단순화, (3) 격음화, (4) 연음 규칙, (5) 유음화, (6) 장애음의 비음화, (7) 유음의 비음화, (8) 구개음화, (9) 경음화, (10) ㅎ-탈락, (11) ㄴ-첨가, (12) 전설 모음화, (13) 종성의 음가 변화. 세부 규칙은 해당 자소에 따라 변화 음소로 정의 되었다. 발음 변화 규칙의 적합성 값은 필수 음소 변동 규칙인 경우 1.0으로 정의 하고, 수의적 음소 변동 규칙인 경우 0..7~0.9 사이의 값을 여러 실험을 통해 정의 하였다.

III. 발음 사전의 생성

발음 사전의 표제어는 유사 형태소이고, 발음 변이는 해당 표제어의 인접한 좌우 형태소의 음소 문맥에 의한 발음 변화를 고려한 음소 문맥 의존적 발음 변이이다. 이 장은 해당 코퍼스에서 자동적으로 발음 사전을 생성하는 방법에 대해 설명하였다.

3.1 발음 변이의 생성

특정 단어의 음소 문맥 의존적 발음 변이는 코퍼스에 따라 다르게 나타 날 수 있다. 본 연구에서는 영역 최적화된 발음 사전을 생성하기 위하여 기존 발음 사전에 새로운 발음 변이를 추가하는 방법을사용하지 않고, 해당 코퍼스에 나타나는 발음 변이를 이용하여 새로운 발음 사전을 자동 생성하는 방법을 사용하였다. 발음 변이는 해당 코퍼스에서 정의된 발음 변화 규칙을 적용하여 생성하고, 발음 변이들의 적합성 값은 적용된 각 규칙의 적합성 값의 곱으로 정의하였다[3].

[그림 1]은 표제어 "어떻"의 발음 변이 생성 과정을 보여 주고 있다. 그림에서 "$"는 형태소의 경계를 의미하고, 이 형태소는 좌측 형태소의 종성 "ㄴ"과 우측 형태소의 초성 "ㄴ"을 음소 문맥으로 가진다. 주어진 예에서 각 노드는 정의된 규칙에서 자소 그룹 G의 경계를 나타내고, 아크는 발음 변화 T를 의미한다. 각 자소는 해당 되는 규칙을 적용하여 발음이 생성된다. 예를 들어 "ㄸ"은 좌측 음소(L) 문맥 "ㅓ"와 우측 음소 문맥(R) "ㅕ"를 가지는 경우 적합성 값 1.0을 거지고 "TQ+TT"의 음소로 변화 되고, 적합성 값 0.96을 가지고 "TT"로, 적합성 값 0.71을 가지고 "PQ+TT"로 변화 된다.

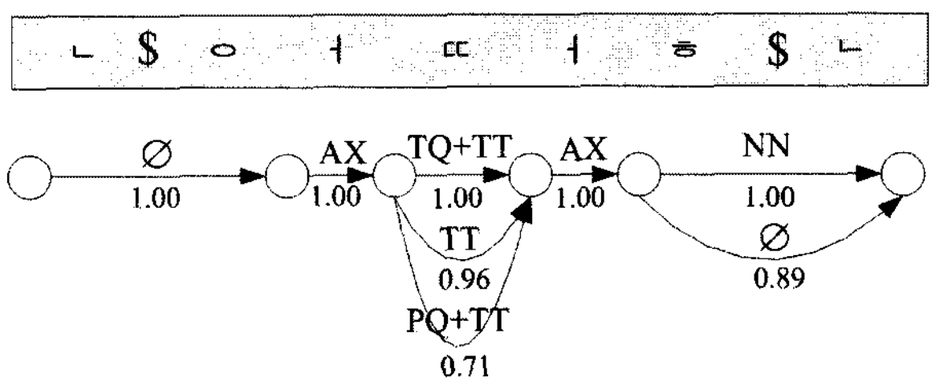

그림 1. 형태소 "어떻"의 발음 변화 예제

예에서의 형태소 "어떻"의 발음 변이들은 각 아크의 가능한 모든 경로를 탐색하는 것으로 생성되어 진다. 최종적인 발음 변이와 적합성 값은 아래 [표 1]과 같이 생성된다.

표 1. 형태소 "어떻"의 발음 변이와 적합성 값

적합성 값	발음 변이
1.0000	AX TQ TT AX NN
0.9545	AX TT AX NN

0.8821	AX TQ TT AX
0.8421	AX TT AX
0.7051	AX PQ TT AX NN
0.6221	AX PQ TT AX

3.2 발음 사전의 생성과 최적화

발음 사전은 생성된 모든 발음 변이 중 적절한 발음 변이를 선정하는 과정으로 이루어진다. 본 연구에서는 발음 변이 선택 기준으로 2가지 결정 조건을 제시 하였다. 첫 번째는 각 발음 변이들의 적합성 값(*F1*)이다. 발음 사전의 표제어들은 여러 개의 발음 변이를 가질 수 있고, 각 발음 변이들은 생성시에 참조된 음소 문맥에 따라 고유의 적합성 값을 가지게 된다. 같은 발음 변이를 가지지만 참조된 음소 문맥이 다를 경우 두 가지 이상의 적합성을 가질 수 있다. 하지만 발음 사전에서는 하나의 특정 발음 변이에 대해 하나의 적합성 값을 하나로 표현해야 하므로 아래 식 (2)와 같이 적합성 값(F1)을 정의하였다.

$$P(v) \approx \sqrt[n]{P(v_1)+P(v_2)+\ldots+P(v_n)} : F1 \quad (2)$$

이 식에서 $P(v)$는 발음 변이 v의 적합성 값이고, v_i는 i번째 발음 변이를 의미 한다.

두 번째 선택 기준은 해당 코퍼스에서의 각 발음 변이들의 생성 빈도 수(*F2*)로 아래 식 (3)과 같이 정의하였다.

$$P(v) = \frac{C(v)}{C(w)} : F2 \quad (3)$$

이 식에서 $C(w)$는 형태소 w가 해당 코퍼스에서 나타난 빈도 이고, $C(v)$는 형태소 w에 대한 발음 변이 v가 생성된 빈도수 이다.

표 2. 형태소 "어떻"의 F1, F2 값

#	발음 변이	F1	F2
1	AX TQ TT AX NN	1.000	0.077
2	AX TT AX NN	0.955	0.077
3	AX TQ TT AX	0.948	1.000
4	AX TT AX	0.906	1.000
5	AX TQ TT AX TQ	0.810	0.846
6	AX TT AX TQ	0.774	0.846
7	AX PQ TT AX NN	0.709	0.077
8	AX PQ TT AX	0.673	1.000
9	AX PQ TT AX TQ	0.574	0.846

[표 2]는 형태소 "어떻"의 F1과 F2 값의 예이다. 이 형태소는 전체 코퍼스에서 13번 나타났고, 발음 변이의 수는 총 9개이다. 위 예에서 첫 번째, 두 번째 발음 변이는 *F1*값은 높지만 *F2*값이 낮다. 이것은 이 발음 변이들이 생성될 때 높은 적합성 값을 가지고 생성 되지만 전체 코퍼스에서 나타나는 빈도는 낮다는 것을 의미한다. 세 번째, 네 번째 변이음은 생성되는 빈도 많으며, 생성시 적합성 값도 높음을 알 수 있다. 여덟 번째 발음 변이는 발음 생성 적합성 값은 낮지만, 생성되는 빈도는 많음을 알 수 있다.

음성 인식 실험에서 사용되는 발음 사전은 이러한 두 가지 선택 기준을 사용하여 다양하게 생성하였고 각각의 기준에 대한 타당성을 비교 검토 하였다.

Ⅳ. 인식실험 및 결과

본 실험에서는 제안된 방법론의 타당성을 검토하기 위하여 한국어 대용량 음성 인식을 수행하였다. 실험은 45K 문장 882K 형태소로 이루어진 낭독체 코퍼스를 사용하였다. 이 코퍼스 중 43K문장은 음향 모델의 학습과 언어 모델 생성에 사용하였고, 학습에 사용되지 않은 2K 문장 중 OOV를 포함하지 않는 600 문장을 인식 실험에 사용하였다. 음향 모델은 HTK를 이용하여 학습하였고, 12개의 믹스쳐를 가지는 트라이폰(tri-phone)으로 구성하였다. 사용된 언어 모델은 백오프 트라이그램(back-off tri-gram)이며, 음향 모델 학습에 사용된 문장을 기본으로 학습하였으며, 37M 형태소를 가지는 신문 코퍼스를 이용하여 추가 학습하였다. 언어 모델에 대한 테스트 문장의 복잡도(Perplexity)는 83.2이다. 인식에 사용된 디코더는 본 연구실의 원 패스 세미-다이나믹 디코더[4]를 사용하였다. 발음 사전의 표제어는 언어 모델 학습에 사용된 코퍼스들에서 빈도수가 높은 25K 유사형태소로 구성하였다.

4.1 발음 사전의 구성 환경

본 실험에는 4가지 종류의 발음 사전을 사용하였다. 기본 발음 사전(*K0*)는 발음 변이 생성시 최고의 적합성 값은 가지는 발음 변이만을 포함하는 것으로 정의하였고, 이 사전의 평균 발음 변이 수는 1.45개 이다. 비교 사전은 발음 변이 선택 결정 조건 *F1, F2*에 따라 3가지로 생성 하였다. 첫 번째는 *F1*만 적용한 발음 사전(*K1*)이고, 두 번째는 *F2*만 적용한 발음 사전(*K2*)이고, 세 번째는 *F1, F2*를 동시에 적용한 발음 사전(*K3*)이다. 또한 비교 사전 *K1, K2, K3*는 1.5개~2.1개의 다양한 평균 발음 변이 수를 가지도록 구성하였다.

4.2 실험 결과 및 분석

인식 실험에서 각 사전에 따른 WER(Word Error Rate)의 변화는 아래 [그림 2]와 같다. *K0* 사전의 경우 15.26%의 WER을 보였고, *K1*, *K2*, *K3* 사전의 경우 평균 발음이 2.1개를 가질 때 최적의 인식률 값을 보였다. *K1*의 경우 최적의 인률 값을 보일 때도, *K0*와 비교할 경우 인식률 향상을 보이지 않았다. 이는 적합성 값만으로 발음을 선택할 경우, 발음 생성 빈도*(F2)*가 낮지만 적합성 값*(F1)*이 높은 발음 변이 들이 다수 포함됨으로 테스트 문장의 발음 변이를 실지로 반영하지 못한 것으로 보인다. *K2*의 경우는 *K0* 사전 보다는 낮은 WER를 가졌지만 큰 차이가 나지 않았다. 하지만 F1과 F2를 동시에 적용한 *K3* 사전의 경우는 *K0*와 비교할 경우 0.72%의 WER를 줄일 수 있었다.

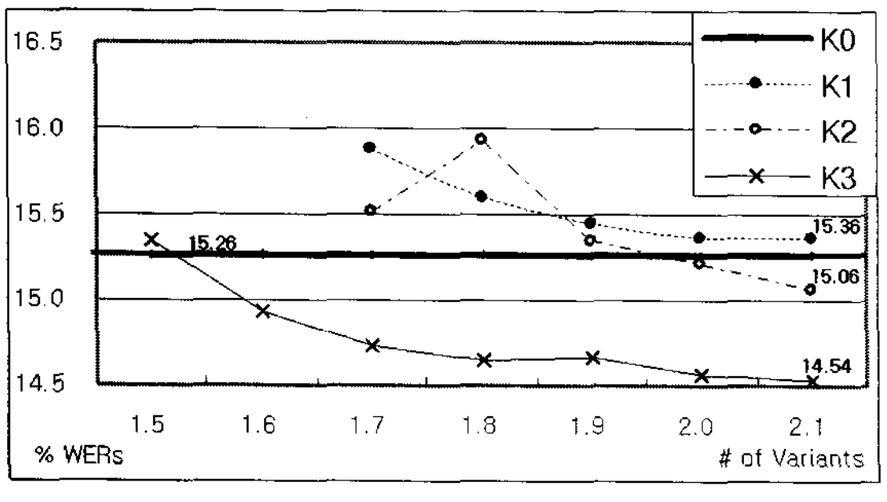

그림 2. 각 사전에 따른 WERs 변화

위 실험에서 나타난 바와 같이 기본 발음 사전 *K0*와 비교할 때, 최고 인식률을 보이는 *K3* 사전의 경우 균 발음 수는 0.65개가 많다. 이 수를 기본 표제어 수로 환산할 경우 16K이상의 발음 변이들을 더 포함하게 된다. 이 경우 탐색 공간의 크기는 발음 변이 수의 증가 보다 더 가파른 증가를 보일 수 있으며, 이는 탐색공간의 복잡도를 높일 수 있다. 본 실험에서는 RTF(Real Time Factor)의 변화를 통해 실제 복잡도 변화를 추측 하여 보았다. [그림 3] 각 사전의 최적 인식률을 보이는 경우의 평균 발음 수, WER, RTF를 나타낸 것이다. 그림에서 나탄 바와 같이 RTF의 변화는 평균 발음 수의 변화에 따라 증가하는 것이 아니라 각 발음 사전의 특성, 즉 해당 음향 모델의 특성을 얼마나 잘 모델링 하는가에 따라 다른 것을 볼 수 있다.

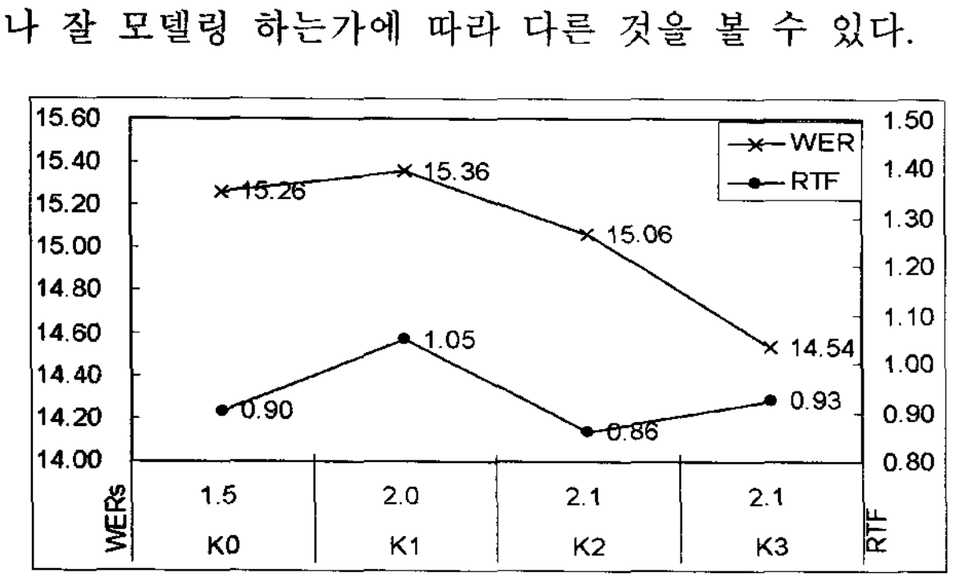

그림 3. 각 사전에 따른 RTF의 변화

요약하면 발음 변이들의 사전 포함 결정 조건을 적용하여 기본 발음 사전에 비해 0.72%의 WER을 줄일 수 있었으며, 평균 발음 변이 수는 크게 증가하지만 RTF의 크게 변화 하지 않음을 볼 수 있었다.

V. 결론

본 논문에서는 발음 사전의 최적화를 위해 두 가지의 발음 변이들의 발음 사전 포함 결정 조건을 제안하였다. 결정 조건 첫 번째는 발음 변이 생성시 발음 적합성 값이고, 두 번째는 발음 변이들의 발생 빈도 값이다. 제안한 결정 조건을 적용함으로써 기본 사전에 비해 0.72%의 WER을 줄일 수 있었다. 또한 평균 발음 수가 늘어나는 것에 비하여 RTF는 크게 변화 하지 않는 것을 볼 수 있었다. 향후 과제로 제안된 결정 조건뿐만 아니라 발음 모델링 기법의 다양화, 즉 학습 기반의 발음 모델과, 지식 기반 발음 모델의 비교와 두 방법의 장점을 결합하는 방법론에 대한 연구를 통해 발음 사전 최적화에 대한 추가 연구가 필요하다.

VI. 감사의 글

이 연구(논문)는 과학기술부 지원으로 수행하는 21세기 프론티어 연구개발사업(인간기능 생활지원 지능로봇 기술개발사업)의 일환으로 수행되었습니다.

참고문헌

[1] H. Strik, "Pronunciation adaptation at the lexical level", Proc. of the ISCA Tutorial & Research Workshop (ITRW) on Adaptation Methods For Speech Recognition, Sophia-Antipolis, France, pp. 123-131, 2001.

[2] K. N. Lee, J. H. Jeon, and M. Chung, "Automatic Generation of Pronunciation Variants for Korean Continuous Speech Recognition," The Journal of the Acoustical Society of Korea, volume 20, pp.35-43, 2001.

[3] J. Jeon and M. Chung, "Automatic Generation of Domain-Dependent Pronunciation Lexicon with Data-Driven Rules and Rule Adaptation," Proceedings of Interspeech 2005, Lisbon Portugal, pp. 1337-1340, 2005.

[4] D.-H. Ahn, M. Chung, "A one pass semi-dynamic network decoder based on language model network," in Proc. of 7th EUROSPEECH '01, Aalborg, Denmark, September 2001.