음소변동규칙의 적합도 조정을 통한 연속음성인식 성능향상

나민수[1], 정민화[2]
[1]서울대학교 인지과학 협동과정, [2]서울대학교 언어학과

Improving the Performance of the Continuous Speech Recognition by Estimating Likelihoods of the Phonetic Rules

Minsoo Na[1], Minhwa Chung[2]
[1]Interdisciplinary Program in Cognitive Science, Seoul National University
[2]Department of Linguistics, Seoul National University
E-mail : {dix39, mchung}@snu.ac.kr

Abstract

The purpose of this paper is to build a pronunciation lexicon with estimated likelihoods of the phonetic rules based on the phonetic realizations and therefore to improve the performance of CSR using the dictionary. In the baseline system, the phonetic rules and their application probabilities are defined with the knowledge of Korean phonology and experimental tuning. The advantage of this approach is to implement the phonetic rules easily and to get stable results on general domains. However, a possible drawback of this method is that it is hard to reflect characteristics of the phonetic realizations on a specific domain. In order to make the system reflect phonetic realizations, the likelihood of phonetic rules is reestimated based on the statistics of the realized phonemes using a forced-alignment method. In our experiment, we generates new lexica which include pronunciation variants created by reestimated phonetic rules and its performance is tested with 12 Gaussian mixture HMMs and back-off bigrams. The proposed method reduced the WER by 0.42%.

I. 서론

HMM(Hidden Markov Model) 기반의 음성인식에서 발음사전은 개별 인식어휘에 대한 PLU(Phone-Like Unit) 단위의 발음열 정보를 제공하고 이 정보는 인식 단계에서 탐색공간을 구성하기 위한 필수요소로 사용된다.

음성인식을 위한 발음열을 생성하는 기본적인 방식은 자소문맥, 형태소 정보 등의 조건을 발음정보인 음소로 사상하는 음소변동규칙을 정의하는 것이다. 음소변동규칙을 도출하는 방식에 따라 발음열 자동생성은 지식기반 접근방식[6]과 학습기반 접근방식[4, 5]으로 분류된다.

학습기반 발음열 생성방식은 발음열 생성의 불규칙성을 반영하기 위하여 음운현상이 발생한 결과인 실제 음성 데이터로부터 음소변동규칙을 유도한다. 음성학, 음운론, 형태론 등의 이론에 근거하여 음소변동규칙을 정의하는 지식기반 발음열 생성방식은 음운현상과 관련하여 연구되어 온 명시적인 지식을 발음열 모델링에 도입할 수 있다는 장점을 가진다[1].

발음열 생성의 문제점은 사전의 구성 시 음운현상을 반영하기 위해서 자소와 형태소 정보를 사용하는데 각 음소변동규칙의 조건과 음소의 사상관계가 항상 일대일 대응이 되지는 않는다는 것이다.[2, 3] 일대일 대응이 되는 조건에 대해서는 지식 또는 학습에 의해 만들어진 규칙에 의해 발음열을 생성할 수 있지만 그렇지 않은 조건에 대해서는 별도의 사전을 유지하여 예외처리하거나 하나의 인식어휘에 대한 다중의 발음열을 확률적으로 정의하는 등의 노력이 필요하다.

본 논문에서는 실제 음성 데이터의 음소관찰을 통해 음소변동규칙의 중요도를 판단하고 적합도를 조정하여 가능성이 높은 다중 발음열을 생성하였다. 생성된 발음사전이 음성인식의 성능을 향상할 수 있음을 실험을 통해서 보인다.

본 논문의 구성은 다음과 같다. 2장에서는 음성데이터의 자소 및 형태소 조건의 음소 실현을 관찰하기 위한 강제인식과정을 기술하고 3장에서는 각 조건의 상대적 빈도를 고려하여 음소변동규칙의 적합도를 조정하는 과정을 살펴본다. 4장에서는 HMM 기반의 연속 음성인식 실험을 실시하여 지식기반 발음열 생성 시스템과 성능을 비교하고 결론을 내린다.

II. 자소열과 음소열의 정렬

음성 데이터로부터 자소에 대한 음소실현 빈도를 관찰하기 위해서 데이터의 음성에 대한 음소 전사가 필요하다. 음소 전사 방법은 크게 수작업에 의한 방법과 인식기를 이용해 자동 생성시키는 방법으로 나누어 볼 수 있다. 수작업 전사는 전문 지식이 필요할 뿐만 아니라 작업에 많은 시간과 노력이 필요하다. 또한 최적의 발음 전사를 이용해 규칙을 유도하더라도 해당 인식기에서 사용되는 음향 모델의 특성으로 인하여 WER의 감소를 보장할 수 없다. 이에 반해 인식기의 특성을 반영 할 수 있는 자동 전사 방법은 수작업 전사보다 빠른 시간에 이루어질 수 있고 저비용으로도 가능하다.

본 연구에서는 음소열은 탐색알고리즘에 의한 자동 전사 방법인 강제인식을 사용하여 얻었다. 강제인식은 43,000 발화의 PBS (Phonetically Balanced Sentence) 음성 데이터에 대해 HTK(Hidden Markov Toolkit)의 HVite 함수로 수행했다. 자소열은 음성데이터의 형태소 태깅이 된 텍스트 코퍼스에서 참고하였다. 텍스트 코퍼스의 각 음절은 이전음절과 현재음절의 형태소 종류에 따라 음절경계(단어경계, 형태소경계, 형태소내부 등)의 종류가 결정되고 이후 자소-음소 정렬과정에서 사용된다.

표 1.. 강제인식 예

자소열: (보험) 혜택을 (받게)
강제인식:
1) (M) JE TH EH K WW L (P)
2) (ㅁF) 예태글 (ㅂI)

강제인식로 얻어진 음소열은 음절내의 삼성 위치와 비교해서 탈락, 종성의 복합자음 및 연음 등에 의한 음소 이동, 인식단위에 따르는 자소 이동 등 정렬 과정을 거치고 그 결과로 각 자소에 해당하는 음소의 배열을 찾는다.

현재 인식대상은 '혜택을' 이고 자소 문맥은 '보험'과 '받게'이다. 강제인식 결과는 1)의 PLU 형식으로 출력되고 출력결과를 자소형식으로 해석하면 2)와 같다. 각 PLU는 삼성의 위치에 따라 동일한 자소를 다른 기호로 표시한다. 1)의 'K'는 초성 'ㄱ'을 나타낸다.

표 2.. 자소-음소 정렬 예

자소열: (ㅁ) ㅎ ㅖ n ㅌ ㅐ n ㄱ ㅇ ㅡ ㄹ (ㅂ)
음소열: (M) n JE n TH EH n K N WW L (P)
자소-음소 정렬:
ㅁ+ㅎ -> ㅁ+ㅇ (M+n)
n+ㅌ -> n+ㅌ (n+TH)
ㄱ+ㅇ -> n+ㄱ (n+K)
ㄹ+ㅂ -> ㄹ+ㅂ(L+P)

강제인식 결과열을 표현형식, 삼성 위치에 따라 정렬하여 표 2와 같이 음소열을 정렬하고 이와 상응하는 음소변동규칙을 찾기 위해서 이전 음절의 종성과 다음 음절의 초성을 단위로 자소열과 음소열을 그룹화한다.

음소변동규칙은 이전음절의 종성자소와 현재음절의 초성자소 등의 자소문맥을 입력으로 음운현상을 반영하여 출력 음소를 결정하고 동일한 조건(자소)에서도 음절의 경계와 형태소의 종류에 따라 다른 규칙이 적용될 수 있다. 모음에 대한 음운현상은 자음과 독립적으로 처리한다.

베이스라인 시스템으로 사용한 기존의 발음열 생성 시스템[6]에서 음운현상의 종류를 총 13종류로 분류하고 각 현상의 자소문맥에 따라 세분하여 각 음소변동규칙으로 음운현상을 근사화 하였으며 필수 음소변동규칙에 대해 1, 수의적 음소변동규칙에 대해서 0.7~0.9 사이의 적합도를 부여했다. 제안한 시스템에서는 베이스라인 시스템의 음운현상 분류와 음소변동규칙의 종류를 그대로 사용했다.

III. 음소변동규칙의 적합도 조정

기존의 지식기반 시스템에서 다중 발음열을 생성하기 위해 각 음소변동규칙에 대한 적합도를 정의하였다. 음소변동규칙의 적합도를 결정하는 방식은 음소변동규칙의 분류에 따라 일괄적으로 부여하는 것으로 자소문맥에 따라 구분되는 음소변동규칙 간의 상대적 비중의 차이는 고려되지 않았다. 음성데이터의 음운현상을 관찰한 예를 보면 이전 음절의 종성과 현재 음절의 초성이 각각 /ㄱ/, /ㅅ/일 때 경음화 된 경우는 문맥조

건의 81%이고 /ㄱ/, /ㅈ/일 때 경음화 된 경우는 조건의 63%로 같은 종류의 음운현상(경음화)라도 자소 문맥에 따라 실현되는 확률에 차이가 있을 수 있음을 알 수 있다. 음성 데이터에서 상대적으로 더 높은 빈도로 발견된 음소변동규칙이 발음열 생성 시에도 높은 적합도를 가지도록 조정한다.

3.1 음소변동규칙의 적합도

음성 데이터의 음운현상을 음소변동규칙의 적합도에 반영하기 위해서 2장과 같이 자소문맥에 따라 실현된 음소열을 음소변동규칙의 입출력 형식으로 정렬했다. 자소입력 조건과 텍스트 코퍼스의 자소열의 음절경계 정보를 사용하여 해당하는 음소변동규칙을 탐색하고 매 발견시마다 규칙의 상대적 발견빈도(r_n)를 높였다.

$$p_n = 0.8 + 0.2 * r_n,\ where\ 0 \le r_n \le 1\ and\ 1 \le n \le 0.$$

$$r_n = \frac{c_n}{c_1 + c_2 + c_3}, 1 \le n \le 3$$

c_n은 각 자소문맥에서 적용될 수 있는 음소변동규칙의 발견빈도를 나타낸다. r_n은 같은 자소문맥에서 발생할 수 있는 다른 음운현상의 전체발견빈도($c_1+c_2+c_3$)에 대한 해당 음운현상의 발견빈도(c_n)로 표현된다. p_n은 하나의 자소문맥에 대한 규칙의 적합도를 나타내고 각 자소문맥은 1~3 개의 음소변동규칙을 가질 수 있다. 데이터에서 발견되지 않았으나 실험 시에는 발생할 수 있는 음운현상이 있으므로 음소변동규칙의 적합도를 발견빈도에 비례하도록 하고 규칙이 음성데이터에서 발견되지 않더라도 최소한 0.8의 적합도를 가지도록 했다.

3.2 조정 적합도 분석

발음열 생성 시 적용 가능한 음소변동규칙이 다수인 경우는 대부분 음절경계가 단어경계일 때이다. 단어경계에서 이전음절의 종성이 /ㄱ/인 경우 각 자소문맥에서 두 종류의 음소변동규칙의 적용이 가능했다. 첫 번째 규칙(표 3. 음운현상1)의 적합도가 1로, 그리고 두 번째 규칙의 적합도가(표 3. 음운현상2)로 0.9로 부여되었다. 제안한 시스템은 이를 음소통계에 기반하여 다음과 같이 조정하였다.

/ㄱ, ㄹ/, /ㄱ, ㅁ/, /ㄱ, ㅇ/, /ㄱ, ㅎ/ 등의 문맥에서 기존의 음소변동규칙의 적합도와 비교할 때 적합도의 상대적 크기가 역전되어 나타났다. 따라서 데이터의 통계에 따라 발음열 생성 시 각 조건에 대해서 음운현상을 반영하지 않고 자소를 그대로 음소로 변환하는 음소변동규칙의 적합도를 더 높였다.

표 3. 자소-음소 정렬로 조정된 적합도

자소문맥	음운현상 1	적합도	음운현상 2	적합도
/ㄱ, ㄱ/	/ㄱ,ㄲ/	0.947	/ㄱ, ㄱ/	0.853
/ㄱ, ㄴ/	/ㅇ, ㄴ/	0.902	/ㄱ, ㄴ/	0.898
/ㄱ, ㄷ/	/ㄱ, ㄸ/	0.958	/ㄱ, ㄷ/	0.842
/ㄱ, ㄹ/	/ㅇ, ㄴ/	0.816	/ㄱ, ㄹ/	0.984
/ㄱ, ㅁ/	/ㅇ, ㅁ/	0.886	/ㄱ, ㅁ/	0.914
/ㄱ, ㅂ/	/ㄱ, ㅃ/	0.940	/ㄱ, ㅂ/	0.859
/ㄱ, ㅅ/	/ㄱ, ㅆ/	0.962	/ㄱ, ㅅ/	0.838
/ㄱ, ㅇ/	/n, ㄱ/	0.800	/ㄱ, ㅇ/	1.000
/ㄱ, ㅈ/	/ㄱ, ㅉ/	0.926	/ㄱ, ㅈ/	0.874
/ㄱ, ㅎ/	/n, ㅋ/	0.800	/ㄱ, ㅎ/	1.000

Ⅳ. 인식실험 및 결과

4.1 실험환경

음소변동규칙의 적합도 조정에 따르는 음성인식의 성능의 변화를 알아보기 위해서 HTK(Hidden Markov Toolkit)를 사용하여 인식실험을 실시했다. PBS 43,000 문장의 음성데이터를 학습 데이터로 사용하고 학습문장에 포함되지 않은 600문장을 인식용 데이터로 사용하였다. 음성신호는 25ms 단위의 윈도우를 10ms씩 이동하여 추출하였다. 음향모델은 Triphone 기반의 HMM을 12개의 Gaussian Mixture로 확장하여 학습했다. 언어모델은 backoff-bigram을 사용하였다. 제안한 시스템을 통해 생성한 발음사전의 평가 시 새로 만들어진 발음사전을 사용하여 음향모델을 재학습하였다.

4.2 발음사전 비교

조정된 음소변동규칙의 적합도가 음성인식성능에 주는 영향을 비교하기 위하여 기존의 지식기반 발음열 생성 시스템(베이스라인 시스템)의 발음사전과 조정된 발음사전을 사용하여 4.1의 환경에서 음성인식실험을 수행했다.

실험에서 사용된 발음사전의 종류는 두 가지이고 각 발음사전은 평균변이음의 수에 따라 다시 6종류로 분류된다.

첫 번째 발음사전은 베이스라인 시스템으로 자동생성한 발음사전이고 KBD라고 한다. 두 번째 발음사전은 베이스라인 시스템의 음소변동규칙의 적합도를 수정하여 생성한 발음사전으로 DRD라고 한다. 각 발음사전은 최대 15개의 발음열을 가질 수 있다. 각 단어에 대한 발음열 후보의 적합도가 해당 단어의 후보 발음열 중 가장 높은 적합도에 비해 정해진 컷오프 비율보다 높다면 해당 후보의 발음열을 발음사전에 기록한다. 이렇게 컷오프 비율보다 높은 상대적 적합도를 가지는 발음열 수의 총합을 단어 수의 총합으로 나누어 발음사전의 평균 변이음의 수를 구한다.

표 4. 베이스라인 시스템의 발음사전 평가

	Var.	Cor.	Subs.	Del.	Ins.	WER	SER
KBD	1.3	80.72	14.66	4.62	2.54	21.82	94.50
	1.5	82.38	13.84	3.78	2.86	20.48	93.50
	1.7	83.44	12.85	3.71	2.66	19.22	91.67
	1.9	83.59	12.74	3.67	2.58	18.99	91.33
	2.1	83.47	12.83	3.70	2.64	19.17	91.50
	2.3	83.13	13.11	3.71	2.74	19.57	92.00

표 5. 제안한 시스템의 발음사전 평가

	Var.	Cor.	Subs.	Del.	Ins.	WER	SER
DBD	1.3	81.95	14.09	3.96	2.77	20.83	93.67
	1.5	82.68	13.46	3.86	2.83	20.15	93.67
	1.7	83.74	12.56	3.70	2.62	19.07	92.00
	1.9	84.00	12.48	3.52	2.57	18.57	91.16
	2.1	84.01	12.51	3.47	2.60	18.58	91.67
	2.3	83.47	13.09	3.43	2.75	19.28	92.00

두 발음사전 구성방법의 음성인식 성능을 비교한 결과 제안한 시스템의 발음사전으로 음성인식 시스템을 구성하였을 때 오류(WER, Word Error Rate)가 18.99%에서 18.57%로 최대 0.42% 감소하였다. 음소변동규칙 적용의 적합도 조정으로 실제 음성데이터에서 통계적으로 빈번하게 발생한 규칙이 발음열 생성에 더 비중이 크게 작용하였고, WER의 향상은 적합도가 큰 음소변동규칙에 의해 생성된 발음열이 발음사전에 포함되어 탐색 네트워크의 구성이 개선된 결과로 해석된다.

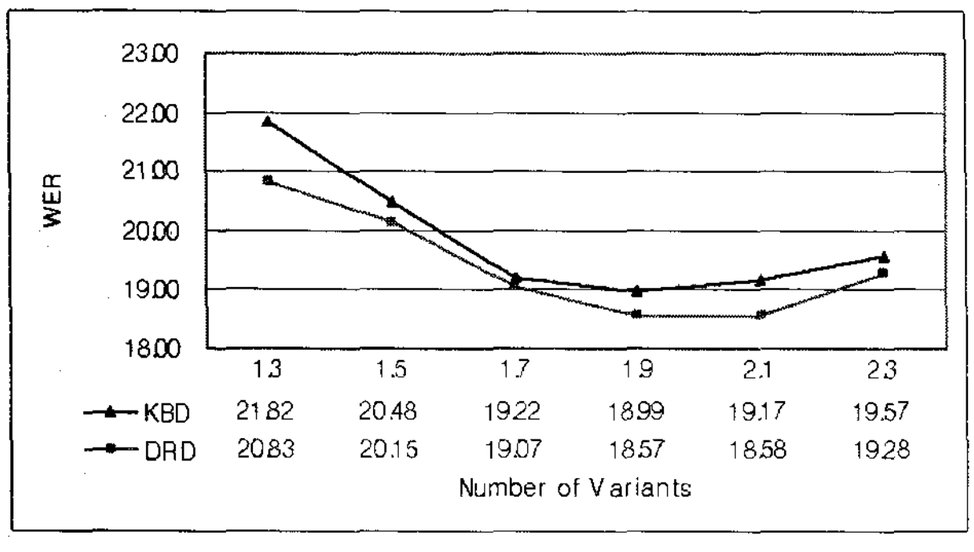

그림 1. WER 기준의 인식 성능 비교

V. 결론

본 논문에서는 음성 데이터에서 고빈도로 발견되는 음운현상이 발음열 생성 시에 상대적으로 높은 적합도를 가진 음소변동규칙으로 반영되도록 규칙의 적합도를 조정하였다. 데이터에서 음운현상의 관찰은 강제인식 기법을 사용하여 수행했고 자소-음소 정렬결과에 따라 음소변동규칙을 조정했다. 기존에 정의된 음소변동규칙의 테두리에서 적합도를 조정하여 음성인식 성능이 향상할 수 있었다.

음성데이터의 자소-음소 정렬결과를 통해서 인식기의 인식범주를 관찰한 결과 기정의 된 음소변동규칙을 통해 한정적으로 설명할 수 없는 거나 비직관적인 음소열로 인식하기도 하였다. 이 같은 데이터 또는 인식기의 특징을 반영하여 발음열 생성 시스템에 정의되지 않은 규칙을 활용하고 발음열 생성에 반영하는 학습방식도 고려하여야 한다.

VI. 감사의 글

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

참고문헌

[1] Strik, H. "Pronunciation Adaptation at the Lexical Level" Proc. ITRW Adaptation Methods for Speech Recognition, pp123-130, 2001.

[2] M. Adda-Decker, L. Lamel. "Pronunciation variants across system configuration, language and speaking style," Speech Communication, Nov. 1999, v.29 n.2-4, pp 83-98.

[3] E. Fosler-Lussier, N. Morgan. "Effects of speaking rate and word frequency on pronunciations in convertional speech," Speech Communication, Nov. 1999, v.29 n.2-4, pp 137-158.

[4] Houda Mokbel, D. Juvet, "Derivation of the optimal set of phonetic transcriptions for a word from its acoustic realizations," Speech Communication, 1999, v.29 n.2-4 pp 49-64.

[5]Jeon,, J. H., M. Chung, "Automatic Generation of Domain-Dependent Pronunciation Lexicon with Data-Driven Rules and Rule Adaptation," Proc. Interspeech-2005, pp1337-1340, 2005.

[6] 전재훈, "형태음운학적 분석에 기반한 한국어 발음열 자동 생성," 서강대학교 전자계산학과 석사학위논문, 1997.