한국인의 영어 인식을 위한 문맥 종속성 기반 음향모델/발음모델 적응*

오유리(GIST), 김홍국(GIST), 이연우(목포대), 이성로(목포대)

<차 례>

<Abstract>

Acoustic and Pronunciation Model Adaptation Based on Context dependency for Korean-English Speech Recognition

Yoo Rhee Oh, Hong Kook Kim, Yeonwoo Lee, Seong Ro Lee

In this paper, we propose a hybrid acoustic and pronunciation model adaptation method based on context dependency for Korean-English speech recognition. The proposed method is performed as follows. First, in order to derive pronunciation variant rules, an n-best phoneme sequence is obtained by phone recognition. Second, we decompose each rule into a context independent (CI) or a context dependent (CD) one. To this end, it is assumed that a different phoneme structure between Korean and English makes CI pronunciation variabilities while coarticulation effects are related to CD pronunciation variabilities. Finally, we perform an acoustic model adaptation and a pronunciation model adaptation for CI and CD pronunciation variabilities, respectively. It is shown from the Korean-English speech recognition experiments that the average word error rate (WER) is decreased by 36.0% when compared to the baseline that does not include any adaptation. In addition, the proposed method has a lower average WER than either the acoustic model adaptation or the pronunciation model adaptation.

* Keywords: Korean-English speech recognition, Acoustic model adaptation, Pronunciation model adaptation.

* 이 논문은 2007년 정부(교육과학부)의 재원으로 한국학술진흥재단의 지원을 받아 수행되었으며(KRF-2007-314-D00245) 또한 본 논문은 지식경제부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음(IITA-2008-C1090-0804-0007).

1. 서 론

국제화, 세계화 등의 흐름으로 모국어가 아닌 타국어를 사용하는 빈도가 증가하고 있다. 이에 따라, 타국어에 대한 음성인식(automatic speech recognition: ASR) 시스템의 개발이 요구되고 있다. 그러나 타국어 음성이 ASR 시스템의 입력으로 들어온 경우, 원어민 음성과 비교하였을 때 인식 성능이 크게 저하되는 경향이 있다[1]. 이는, ASR 시스템이 일반적으로 원어민 음성 데이터베이스(DB)를 이용하여 설계된 반면 원어민 음성과 타국어 음성 사이에 서로 다른 특성이 존재하기 때문이다. 다시 말해, ASR 시스템에 학습된 모델과 실제로 인식에 사용되는 음성 특징과의 차이가 발생하기 때문이다[2][3]. 이러한 문제점을 해결하기 위하여 타국어 음성 DB로 ASR 시스템을 학습시키는 방법이 있으나, 비용 및 시간 등의 측면에서 비효율적이다. 그러므로 타국어 음성에 대한 ASR 시스템의 성능을 향상시키기 위하여, 원어민 음성 DB로 학습된 ASR 시스템을 적응(adaptation)시키거나 타국어의 음성 특징이 원어민의 음성 특징을 반영하도록 보상시킬 필요가 있다.

타국어 음성에 대한 ASR 시스템의 성능을 향상시키기 위한 방법은 크게 발음모델 적응 방법, 음향모델 적응 방법, 언어모델 적응 방법, 그리고 각각의 방법들을 결합한 방법 등으로 분류할 수 있다. 첫째, 발음모델 적응 방법은 타국어 음성에 대한 발음 변이 규칙을 발음모델에 적용하는 것이다[4][5]. 이러한 방법으로, 음소인식과 결정트리(decision tree)를 이용한 데이터 기반(data-driven) 발음모델 적응 방법 등이 제안되어 왔다[6][7][8][9]. 둘째, 음향모델 적응 방법으로는 타국어 음성을 이용하여 음향모델을 변환, 적응시키는 방법들이 제안되어 왔다[4][10][11]. 셋째, 언어모델 적응 방법은 타국어 화자의 발성 방식 등 문법적 변이 특성을 언어모델에 적용하는 방법을 일컫는다[12]. 마지막으로 보다 나은 인식성능 향상을 위하여 여러 방법을 결합하는 방법이 있다[13]. 예를 들어 [13]에서는 발음모델에 따라 음향모델의 사용을 다르게 함으로써 음향모델 적응과 발음모델 적응 방식을 결합하는 방식을 제안하였다. 뿐만 아니라, maximum likelihood linear regression (MLLR), maximum a priori (MAP) 등 화자 적응 기법을 결합하기도 한다.

본 논문에서는 한국인의 영어 음성에 대한 ASR 시스템의 성능을 향상시키기 위하여 음향모델 적응 방법과 발음모델 적응 방법을 결합하는 방식을 제안한다. 특히, 한국인의 영어에 대한 발음변이 특성을 분석한 후, 발음변이 특성이 문맥에 따라 달리 나타나는 현상을 이용한다. 즉, 발음변이의 문맥 특성에 따라 음향모델 적응 방법이나 혹은 발음모델 적응 방법을 선택하여 적용하는 방식을 사용한다. 이를 위해 먼저, 결정트리 기반 간접적 데이터 기반 방식을 사용하여 한국인의 영어에 대한 발음변이 특성을 분석한다[14]. 구체적으로, 한국인의 영어에 대한 개발용 집합(development set)으로 음소인식을 수행하여 n-best 음소열을 획득한 후, C4.5와 같은 결정트리를 이용하여 발음변이 규칙을 추출한다[15]. 다음으로, 문맥

종속성에 따라 발음변이 규칙들을 분류하여, 문맥독립적(context-independent) 발음변이 규칙과 문맥종속적(context-dependent) 발음변이 규칙으로 나눈다. 여기서, 문맥종속 발음변이 규칙은 주변 음소열의 종류에 따라 발음변이가 발생하는 것으로 문맥에 따른 조음 현상에 해당한다. 또한, 문맥독립 발음변이 규칙은 주변 음소열에 상관없이 발음변이가 발생하는 것으로 모국어인 한국어와 타국어인 영어의 서로 다른 발음 공간(pronunciation space)에 의한 현상이다. 마지막으로, 문맥독립 발음변이 규칙과 문맥종속 발음변이 규칙에 대해 각각 음향모델 적응[16]과 발음모델 적응[14] 방법을 적용한다.

본 논문의 구성은 다음과 같다. 서론에 이어서 2장에서는 데이터 기반 방식으로 발음변이 규칙을 획득하고 분류하는 방식을 설명하고, 3장에서 분류된 발음변이 규칙을 이용하여 음향모델 적응과 발음모델 적응을 결합하는 방식을 제안한다. 다음으로 4장에서는 본 논문에서 제안한 방법을 이용한 음성인식 실험과 그 결과를 보인다. 마지막으로 5장에서 본 논문의 결론을 맺도록 한다.

2. 한국인의 영어에 대한 발음변이 분석 및 분류

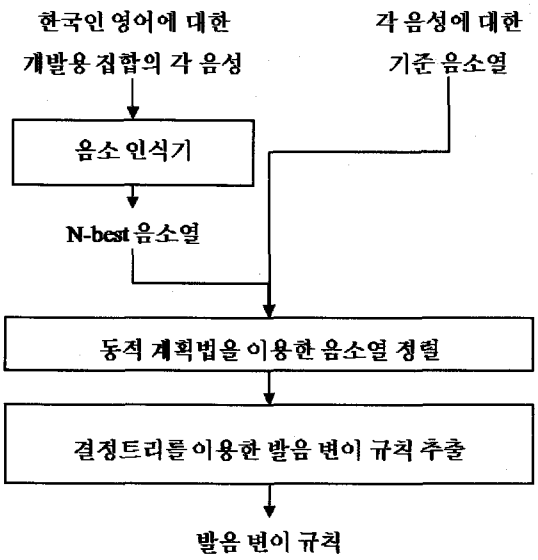
2.1. 데이터 기반 발음변이 분석

한국인의 영어에 대한 발음변이를 분석하기 위하여, <그림 1>과 같이 결정트리를 이용한 간접적 데이터 기반 방식을 사용한다. 먼저, 한국인의 영어에 대한 개발용 집합의 각 음성에 대하여 음소인식을 수행함으로써 n-best 음소열을 획득한다. 그 후, 동적 계획법(dynamic programming)을 이용하여 n-best 음소열을 정렬시킨다. 여기서 기준이 되는 음소열은 CMU 발음 사전을 바탕으로 생성된다[17]. 정렬된 음소열과 기준 음소열을 바탕으로, 음소 규칙 패턴을 식 (1)과 같이 획득한다.

$$L_1 - L_2 - X + R_1 + R_2 \rightarrow Y \tag{1}$$

여기서 음소 X는 왼쪽 두 음소(L_1과 L_2)와 오른쪽 두 음소(R_1과 R_2)가 있을 때 음소 Y로의 변이가 발생함을 의미한다.

다음으로, 획득된 음소 규칙 패턴을 입력으로 하는 결정트리로부터 발음변이 규칙을 추출한다. 본 논문에서는 결정트리로 Quinlan이 ID3 알고리즘을 확장하여 설계한 소프트웨어인 C4.5를 사용하였다[15]. 또한, 영어의 각 음소에 대하여 결정트리가 한 개씩 생성되고, 결정트리의 속성으로는 발성하는 음소 X와 X의 왼쪽 두 음소 L_1과 L_2, 오른쪽 두 음소 R_1과 R_2, X에 대한 변이 음소 Y 등이 사용된다. 다시 말해, 정렬된 음소열과 기준 음소열 사이의 음소간 사상(mapping)을 통하



<그림 1> 간접적 데이터 기반 발음변이 모델링 기법 적용 과정

여 음소 규칙 패턴을 획득한 후, 음소 규칙 패턴을 이용하여 C4.5 기반 결정트리를 생성한다. 그리고 각 음소 별 결정트리를 바탕으로 음소 별 발음변이 규칙들을 획득한다. 식 (2)는 C4.5 기반 결정트리를 통하여 획득된 음소 $phoneme_{target}$에 대한 발음 변이 규칙의 한 예이다.

$$\begin{aligned} &\text{Rule } rule_{id}: \\ &\qquad mPrevPrev = P_{L1}, mPrev = P_{L2} \\ &\qquad mNext = P_{R1}, mNextNext = P_{R2} \\ &\qquad \rightarrow \text{class } phoneme_{variant} \\ &\text{Default}: \\ &\qquad \text{class } phoneme_{default} \end{aligned} \tag{2}$$

여기서 $rule_{id}$는 발음 변이 규칙의 식별자를 나타내고, '$mPrevPrev = P_{L1}$, $mPrev = P_{L2}$, $mNext = P_{R1}$, $mNextNext = P_{R2}$'는 $rule_{id}$가 적용되는 문맥 정보이다. 즉, 식 (2)는 음소 $phoneme_{target}$이 '$P_{L1} - P_{L2} - phoneme_{target} + P_{R1} + P_{R2}$'의 문맥에 있을 때 음소 $phoneme_{variant}$로 변이되고, 그 외의 경우에는 default class인 $phoneme_{default}$로 발음됨을 의미한다. 데이터 기반 발음변이 분석에 대한 자세한 설명은 [14]에 기술되어 있다.

2.2. 문맥종속/문맥독립 발음변이 규칙

본 절에서는, 발음 변이 규칙을 문맥종속 발음변이 규칙과 문맥독립 발음변이 규칙으로 분류하는 방법에 대해서 설명한다. 먼저, 문맥종속 발음변이는 음소에 대한 이음(allophone)과 같이 제한된 특정 음소열 환경에서 발생하는 것이다. 예를 들어 'this spring'을 말한다고 가정할 때, 'this */DH IH S/*'의 마지막 음소와 'spring */S P R IH NG/*'의 첫 음소는 */S/*로 인접해 있으므로 */DH IH S S P R IH NG/* 대신에 */DH IH S P R IH NG/*로 발음되곤 한다.[1] 이러한 현상은 조음 현상에 의하여 발생된 것으로, 음소 */S/*에 대한 변이 발음이 발생한 것이라 볼 수 있다. 따라서 음소 */S/*에 대한 발음 변이 규칙은 식 (3)과 같이 정의될 수 있다.

$$
\begin{aligned}
&\text{Rule } S\_rule_1: \\
&\qquad mPrevPrev = P_{1,L1},\ mPrev = /S/ \\
&\qquad mNext = P_{1,R1},\ mNextNext = P_{1,R2} \\
&\qquad \rightarrow /sil/ \\
&\text{Rule } S\_rule_2: \\
&\qquad mPrevPrev = P_{2,L1},\ mPrev = P_{2,L2} \\
&\qquad mNext = /S/,\ mNextNext = P_{2,R2} \\
&\qquad \rightarrow /sil/ \\
&\text{Default}: \\
&\qquad \text{class}/S/
\end{aligned}
\tag{3}
$$

여기서 */sil/*은 음소의 삭제를 의미한다. 또한, 식 (3)에서 음소 */S/*는 규칙 S_rule_1 또는 규칙 S_rule_2가 만족하는 상황에서 삭제되고, 그 외의 상황에서는 default class인 */S/*로 발음됨을 의미한다.

문맥종속 발음변이 규칙과는 달리, 문맥독립 발음변이 규칙은 화자의 모국어에 존재하지 않는 발음에 의한 것이다. 예를 들어, 한국인 화자가 영어 단어인 'five'를 말한다고 가정할 때 */F AY V/* 대신에 */P AY B/*로 발음할 수 있다. 이는 한국어 음소 구성에는 */F/*나 */V/* 등이 없기 때문에 한국인 화자는 이와 유사한 발음인 */P/*나 */B/* 등으로 발음하기도 한다. 따라서 음소 */F/*에 대한 발음변이 규칙은 아래의 식 (4)와 같이 정의될 수 있다. 식 (4)에서 음소 */F/*는 일반적으로 default class인 */P/*로 발음되고, 예외 상황인 규칙 F_rule_1와 규칙 F_rule_2 등에서는 $\text{class}\,phoneme_{variant1}$, $\text{class}\,phoneme_{variant2}$ 등으로 변경되어 발음됨을 의미한다.

1) 본 논문에서의 발음에 대한 표기는 uppercase ARPAbet 형식을 따르며, 이는 baseline ASR 시스템에서의 음소 표기와 동일하다.

$$
\begin{aligned}
&\text{Rule } F\_rule_1: \\
&\qquad mPrevPrev = P_{1,L1},\ mPrev = P_{1,L2} \\
&\qquad mNext = P_{1,R1},\ mNextNext = P_{1,R2} \\
&\qquad \rightarrow \text{class } phoneme_{variant1} \\
&\text{Rule } F\_rule_2: \\
&\qquad mPrevPrev = P_{2,L1},\ mPrev = P_{2,L2} \\
&\qquad mNext = P_{2,R1},\ mNextNext = P_{2,R2} \\
&\qquad \rightarrow \text{class } phoneme_{variant2} \\
&\text{Default}: \\
&\qquad \text{class /P/}
\end{aligned}
\tag{4}
$$

위의 두 가지 예를 통하여, 문맥종속 발음변이 규칙에서는 $phoneme_{default}$가 $phoneme_{target}$과 같으나, 문맥독립 발음변이 규칙에서는 $phoneme_{default}$가 $phoneme_{target}$과 다름을 보인다. <표 1>은 문맥종속 발음변이 규칙과 문맥독립 발음 변이 규칙에 대하여 $phoneme_{default}$와 $phoneme_{target}$의 관계, 발생 원인 등을 정리한 것이다.

<표 1> 문맥종속 발음변이 규칙과 문맥독립 발음변이 규칙의 비교

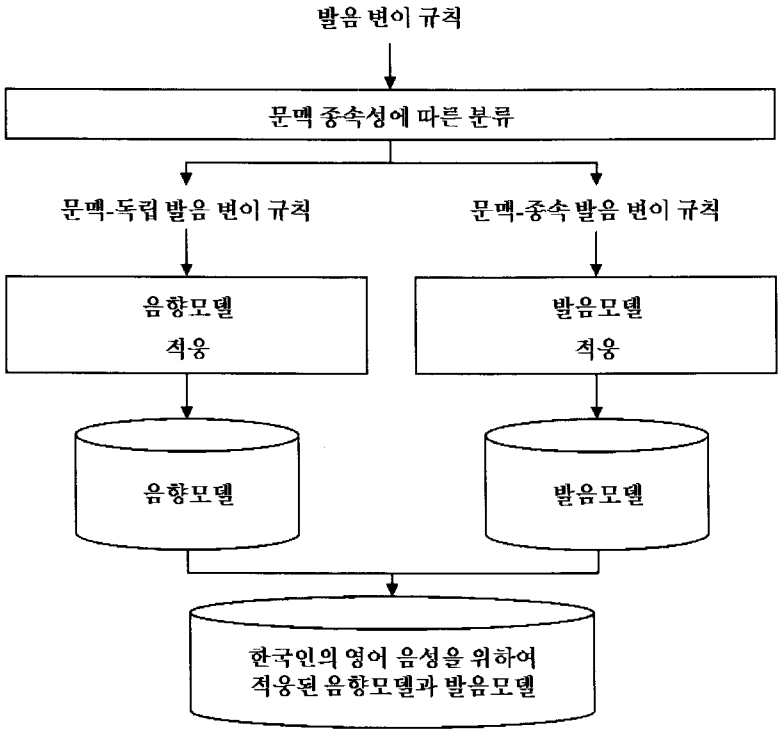
	발음변이 규칙	
	문맥종속	문맥종속
$phoneme_{default}$ vs. $phoneme_{target}$	동일	상이
발생 원인	조음 현상	상이한 음소 체계

3. 한국인의 영어인식을 위한 음향모델/발음모델 적응의 결합

한국인의 영어에 대한 ASR 시스템의 성능을 향상시키기 위하여, <그림 2>와 같이 발음변이 규칙의 문맥 종속성 여부에 따라 음향모델 적응과 발음모델 적응을 구별하여 적용한다. 즉, 2장에서 기술한 바와 같이 한국인의 영어에 대한 발음변이 규칙을 획득하고 각각의 규칙을 문맥독립 혹은 문맥종속 발음변이 규칙으로 분류한다. 그 후, 문맥독립 발음변이 규칙에 대해서는 음향모델 적응을, 문맥종속 발음변이 규칙에 대해서는 발음모델 적응을 적용한다.

3.1. 음향모델 적응

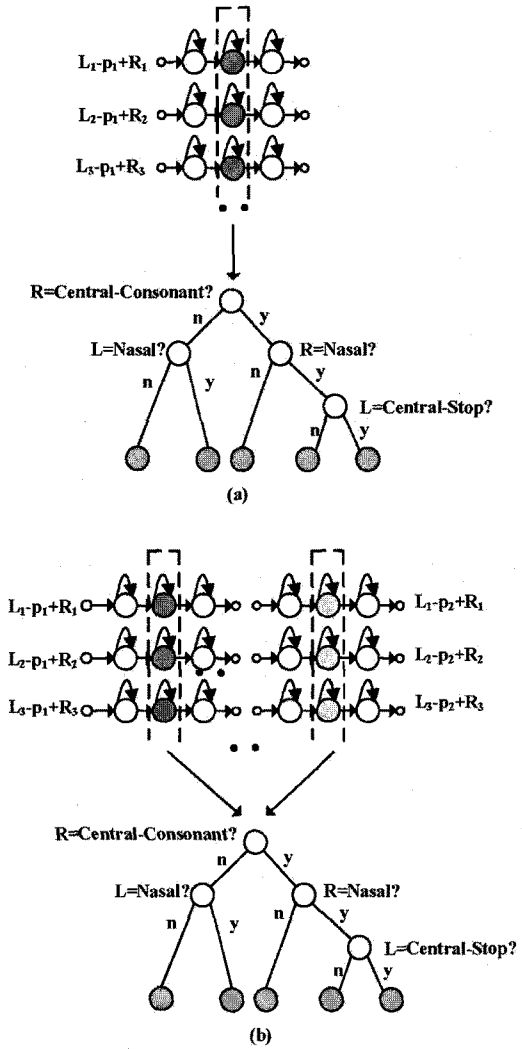
본 장에서는, 2장에서 획득된 문맥독립 발음변이 규칙을 이용하여 음향모델을 적응시키는 방법에 대하여 기술한다. 음향모델 적응은 [16]에서 제안한 방식을 근



<그림 2> 한국인의 영어 음성에 대한 ASR 시스템을 위한 발음변이 규칙의 문맥 종속성에 따른 음향모델/발음모델 적응의 결합

간으로 한다. 하지만, [16]에서와의 차이점은 [16]에서는 monophone 기반 confusion matrix를 이용하여 발음변이를 획득한 반면, 본 논문에서는 2장에서 언급한 바와 같이 간접적 데이터 기반 방식을 이용한다. 문맥독립 발음변이 규칙이 발견되면, 발음변이를 보상하기 위하여 음향모델을 상태 공유단계에서 적응시킨다.

<그림 3>은 본 논문에서 사용하는 음향모델 적응의 주요 과정을 보여 준다. 여기서 음소 /p_1/은 발음변이가 없는 것이고, 음소 /p_2/는 문맥독립 발음변이 /p_1/가 있는 경우이며 두 음소에 대한 결정트리는 각각 <그림 3>의 (a)와 (b)와 같다. 이를 식 (1)과 같이 표현하면 각각 *-*-/p_1/+*+*→/p_1/, *-*-/p_2/+*+*→/p_1/이 되며, 문맥독립 발음변이가 있는 *-*-/p_2/+*+*→/p_1/에 대해 /p_2/→/p_1/로 간단히 나타내기로 한다. 그림에서 보는 바와 같이 발음변이가 없는 음소 /p_1/의 경우, 결정트리의 부모 노드(root node)에는 중심 음소가 /p_1/인 모든 triphone 모델이 놓인다. 반면, 문맥독립 발음변이가 있는 음소 /p_2/의 경우, 결정트리의 부모 노드에는 중심 음소가 /p_2/인 모든 triphone 모델뿐 아니라 변이 음소인 /p_1/이 중심 음소인 모든 triphone 모델도 함께 놓인다. 그 후, 결정트리의 부모 노드에 놓인 triphone 모델들은 미리 준비된



<그림 3> 음향모델 적응을 위한 상태 공유 도식[16]; (a) 발음변이가 없는 음소 /p_1/에 대한 결정 트리 (b) 문맥독립 발음변이가 있는 음소 /p_2/에 대한 결정트리

결정 질문들에 따라 결정트리의 말단 노드로 군집화된다. 또한 결정트리의 각 말단 노드에 대하여, 군집화된 triphone 모델들 중 가우시안 분포(Gaussian distribution)의 분산(variance)이 가장 큰 모델을 대표 모델로 결정한다.

식 (4)의 문맥독립 발음변이 /F/→/P/를 예로 들면, <그림 3>의 (b)의 상태공유 기법을 이용한 음향모델 적응을 적용한다. 즉, /F/에 대한 결정트리의 부모 노드에 중심음소가 /P/이거나 /F/인 모든 triphone 모델을 놓은 후, 결정 질문을 이용하여 군집화한다. 그 후, 결정트리의 말단 노드에 놓인 음향모델들 중 가우시안 분포의 분산이 가장 큰 음향모델을 대표 모델로 선택한다.

3.2. 발음모델 적응

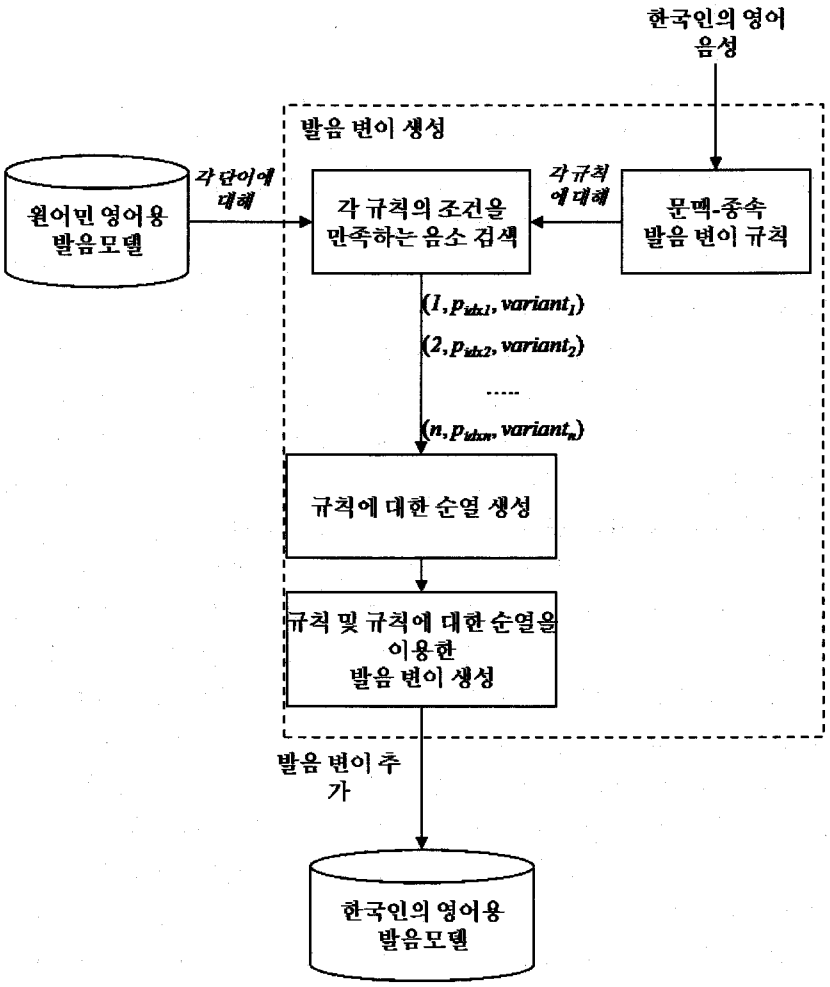
문맥종속 발음변이로 판단된 경우, [14]에서 제안된 발음모델 적응을 수행한다. 다시 말해, 2장에서 설명된 간접적 데이터 기반 방식을 이용하여 한국인의 영어 음성에 대한 문맥종속 발음변이 규칙들을 획득한다. 그리고 문맥종속 발음변이 규칙들을 만족하는 각 단어의 변이 발음들을 포함시킴으로써 다중 발음모델을 생성한다. [14]에서 제안한 방식과 다른 점은, 본 논문에서는 간접적 데이터 기반 방식으로 획득된 발음변이 규칙들 중 문맥종속 발음변이 규칙만을 사용한다는 점이다. 이와 같이 발음변이 규칙 중 문맥종속 발음변이 규칙만을 선별하여 사용함으로써, 다중 발음모델의 복잡도(confusability)를 감소시키고 인식 속도를 빠르게 할 수 있다. 뿐만 아니라, 한 단어에 대하여 여러 개의 발음을 가지는 다중 발음모델의 경우, 한 개 이상의 단어가 유사한 발음을 가지게 되는 등의 이유로 혼잡도가 증가하여 인식 성능이 저하되는 문제점을 줄일 수 있다 [18].

<그림 4>는 문맥종속 발음변이 규칙을 이용하여 발음모델을 적응시키는 과정을 보인다. 먼저, 원어민 영어를 위하여 제작된 발음모델의 각 단어에 대하여, 문맥종속 발음변이 규칙의 조건을 만족하는 음소들이 있는지를 검색한다. 그리고 각 단어에 대하여 문맥종속 발음변이 규칙의 조건을 만족하는 문맥종속 발음변이 규칙들을 $(Rule_{id}, P_{idx}, phoneme_{variant})$의 형태로 열거한다. 여기서, $Rule_{id}$는 문맥종속 발음변이 규칙에 대한 식별자, P_{idx}는 단어의 발음음소열에서 해당 음소의 위치, $phoneme_{variant}$는 문맥종속 발음변이 규칙에 해당하는 변이 발음을 나타낸다. 다음으로, 한 단어 내에서 여러 개의 문맥종속 발음변이 규칙이 적용되는 경우, 문맥종속 발음변이 규칙의 모든 가능한 조합을 찾아낸다. 어떠한 단어에 적용되는 문맥종속 발음변이 규칙이 $Rule_1$, $Rule_2$인 경우를 예로 들면, 모든 가능한 문맥종속 발음변이 규칙의 조합은 $(Rule_1)$, $(Rule_2)$, $(Rule_1, Rule_2)$가 된다. 마지막으로, 발음모델의 각 단어에 대하여, 발견된 문맥종속 발음변이 규칙 및 이에 대한 조합을 이용하여 변이 발음을 추가함으로써 한국인의 영어를 위한 발음모델을 생성한다.

4. 음성인식 실험

4.1. Baseline ASR 시스템

원어민 영어 음성으로 영어 ASR 시스템을 학습시키기 위하여 Wall Street Journal database (WSJ0)의 일부를 사용하였다[19]. WSJ0은 대어휘급 연속 ASR 시스템의 성능을 평가하기 위하여 구축된 5,000 단어급 음성 데이터베이스이다. 학습용 집합은, Sennheiser 근거리 마이크로폰과 몇몇 원거리 마이크로폰으로 녹음된



<그림 4> 문맥종속 발음변이 규칙을 이용한 발음모델 적응 과정.

7,138개의 발화음성으로 구성된다. 또한, 각 발화음성은 16 kHz로 샘플링되어 샘플당 16bit로 저장되어져 있다. 그리고 음성인식용 특징으로, 12차 mel-frequency cepstral coefficient (MFCC)와 로그 에너지를 추출하였으며 이에 대한 1차, 2차 미분계수를 계산하여 총 39차 특징 벡터를 사용하였다. 또한, 켑스트럼 평균 정규화 (cepstral mean normalization) 기법과 에너지 정규화(energy normalization) 기법을 적용하였다.

Baseline ASR 시스템의 음향모델은, 3개의 상태(state)를 가지며 left-to-right 형태의 hidden Markov model (HMM)로서, 문맥종속적이고 4 개의 혼합밀도(mixture)를 가진다. 또한 음향모델의 학습을 위하여 HTK v.3.2 toolkit 을 사용하였다[20]. 뿐만 아니라, 모든 triphone 모델은 41 개의 monophone 모델에서 확장되었으며, triphone

모델의 각 상태(state)는 결정트리에 의하여 군집화되었다[21]. 결과적으로, baseline ASR 시스템의 음향모델은 8,360개의 triphone 모델 및 5,356개의 상태로 구성되며, 본 논문에서는 이를 AM0로 칭하기로 한다.

본 논문에서 사용한 한국인의 영어 음성으로는 Speech Information Technology & Industry Promotion Center (SiTEC)에서 제공하는 한국인의 영어발화 음성 DB (K-SEC)의 일부이다[22]. K-SEC은 원어민 화자와 한국인 화자에 의한 영어 발화음성을 포함하는 데이터베이스로서, 본 논문에서는 개발용 집합(development set), 평가용 집합(test set) 등으로 사용하였다. 개발용 집합은 1명의 한국인 화자가 1,103개의 고립 단어를 발화한 음성으로 구성된다. 또한, 평가용 집합은 49명의 한국인 화자와 7명의 원어민 화자가 각각 14개의 문장을 발화한 음성으로 구성되며, 여기서 한 문장 당 평균 10.4개의 단어를 포함한다.

모국어와 타국어의 특성 차이에 의한 음향모델/발음모델의 영향을 판단하기 위하여, 평가용 집합에 사용된 문장들만으로 언어모델을 생성하였으며 back-off bigram 언어모델을 사용하였다. 또한, 각 단어에 대한 발음은 CMU에서 제공하는 발음사전을 이용하여 전사하였으며, CMU 발음사전에 존재하지 않는 단어에 대해서는 직접 전사하였다[17]. 본 논문에서는 baseline ASR 시스템의 발음모델을 PM0로 칭하기로 한다.

4.2. 문맥 종속성에 기반한 음향모델/발음모델 적응

본 논문에서 제안하는 문맥 종속성에 기반한 음향모델/발음모델 적응 방법에 대한 성능을 평가하기 위하여, 먼저 2장에서 기술한 방식으로 발음변이 규칙을 획득 및 분류하였다. 즉, 개발용 집합의 발화 음성으로 음소인식을 수행하여 20-best 음소열을 얻었다. 그 후, 기준 음소열과 인식된 음소열을 정렬하고 발음 변이 규칙을 생성하였다. 다음으로, 발음변이 규칙들을 문맥종속 발음변이 규칙과 문맥독립 발음변이 규칙으로 분류하였다. 이 과정을 통하여 획득된 문맥독립 발음변이 규칙은 */G/→/sil/*, */L/→/R/*, */TH/→/DH/*, */ZH/→/Z/*, 그리고 */G/→/T/*이다. 여기서, */R/*, */TH/*, */DH/*, 그리고 */ZH/*는 한국어의 음소 체계에 없는 음소로서, 획득된 문맥독립 발음변이 규칙이 서로 다른 언어 간의 음소 체계에 의한 영향을 반영함으로 해석할 수 있다.

<표 2>는 평가용 집합에 대한 다양한 ASR 시스템의 음성인식 성능을 보인다. 여기서 비교된 ASR 시스템은 다음과 같다.

1) baseline ASR 시스템(AM0 + PM0)
2) 문맥독립 발음변이 규칙을 이용하여 음향모델만을 적응시킨 ASR 시스템(적응된 AM + PM0)
3) 문맥종속 발음변이 규칙을 이용하여 발음모델만을 적응시킨 ASR 시스템

(AM0 + 적응된 PM1)

4) 발음변이 규칙 전체를 이용하여 발음모델만을 적응시킨 ASR 시스템(AM0 + 적응된 PM2)
5) 문맥독립/문맥종속 발음변이 규칙을 이용하여 음향모델/발음모델을 모두 적응시킨 ASR 시스템(적응된 AM + 적응된 PM1).

먼저, baseline ASR 시스템인 (AM0 + PM0)의 평균 단어오인식률(word error rate: WER)은 10.3%이다. 문맥독립 발음변이 규칙을 이용하여 음향모델만을 적응시킨 모델(적응된 AM + PM0)과 문맥종속 발음변이 규칙을 이용하여 발음모델만을 적응시킨 모델(AM0 + 적응된 PM1)의 경우, 평균 WER이 각각 9.5%와 9.46%로 감소됨을 볼 수 있었다. 뿐만 아니라, [14]에서 제안된 발음모델 적응과 같이 발음변이 규칙 전체를 이용하여 발음변이 규칙 전체를 이용하여 발음모델을 적응시킨 (AM0 + 적응된 PM2) 경우, 평균 WER이 8.98%로 감소됨을 확인할 수 있었다. 마지막으로, 음향모델과 발음모델을 모두 적응시킨 모델(적응된 AM + 적응된 PM)의 경우, 평균 WER은 8.65%임을 알 수 있었다. 실험 결과를 통하여, 음향모델만을 적응시킨 경우와 발음모델만을 적응시킨 경우보다, 본 논문에서 제안하는 방식으로 음향모델과 발음모델을 모두 적응시킨 경우 한국인의 영어에 대한 ASR 인식 성능을 보다 향상시킬 수 있음을 알 수 있었다.

<표 2> 평가용 집합에 대한 다양한 ASR 시스템의 WER (%) 성능 비교

	화자			
	한국인	원어민	평균	WER 상대적 감소율 (%)
Baseline (AM0 + PM0)	19.92	0.68	10.30	-
적응된 AM + PM0	18.12	0.88	9.50	7.8
AM0 + 적응된 PM1	18.23	0.68	9.46	8.2
AM0 + 적응된 PM2	17.28	0.68	8.98	12.8
적응된 AM + 적응된 PM1	16.51	0.78	8.65	16.0

추가적으로, 개발용 집합의 발화 음성으로 음소인식을 수행한 결과 중 20-best 음소열을 사용하는 대신, 5-best, 10-best, 15-best, 20-best, 30-best, 40-best 등으로 하여 발음변이 규칙을 획득하였다. N-best를 달리하여 생성된 발음변이 규칙으로 음향모델과 발음모델을 적응시킨 ASR 시스템의 성능은 <표 3>과 같다.

먼저, baseline ASR 시스템인 (AM0+PM0)의 평균 WER는 10.3%이다. 5-best, 10-best, 15-best, 20-best, 30-best, 40-best 등의 음소인식 결과를 사용하여 획득된 발음변이 규칙을 이용하여 음향모델과 발음모델을 적응시킨 ASR 시스템의 평균 WER은 각각 36.0%, 28.7%, 31.4%, 16.1%, 29.4%, 26.4%이다. 즉, 5-best 음소인식 결과로 획득된 발음변이 규칙을 이용하여 제안한 방식으로 음향모델과 발음모델

을 적응시킨 경우, 평균 WER는 상대적으로 36.0% 감소됨을 확인할 수 있었다.

<표 3> 문맥독립 발음변이 규칙 획득에 사용된 n-best 음소인식 결과에 따른 ASR 시스템의 WER (%) 성능 비교

	화자			
	한국인	원어민	평균	WER 상대적 감소율 (%)
Baseline (AM0+PM0)	19.92	0.68	10.30	-
5-best (AM0 + 적응된 PM1)	12.11	1.08	6.60	36.0
10-best (AM0 + 적응된 PM1)	14.10	0.59	7.35	28.7
15-best (AM0 + 적응된 PM1)	13.25	0.88	7.07	31.4
20-best (AM0 + 적응된 PM1)	16.51	0.78	8.65	16.1
30-best (AM0 + 적응된 PM1)	13.96	0.59	7.28	29.4
40-best (AM0 + 적응된 PM1)	14.09	1.08	7.59	26.4

5. 결 론

본 논문에서는 한국인의 영어 음성에 대한 ASR 시스템의 인식 성능을 향상시키기 위하여, 발음변이 규칙의 문맥 종속성에 따라 음향모델과 발음모델을 적응시키는 방법을 제안하였다. 먼저, 문맥 종속성에 따른 분류는, 문맥독립 발음변이 규칙은 모국어와 타국어의 다른 언어 체계에 의한 것이고 문맥종속 발음변이 규칙은 조음 현상에 의한 것이라는 가정에 근거하였다. 제안하는 문맥 종속성에 기반한 음향모델/발음모델 적응은 크게 한국인의 영어 분석, 문맥 종속성에 따른 발음변이 분류, 발음변이 규칙에 따른 음향모델/발음모델 적응 등 세 가지 과정으로 이루어졌다. 한국인의 영어에 대한 ASR 실험을 통하여, 원어민 음성으로 학습된 baseline ASR 시스템과 비교하여 제안한 적응 방식을 적용하였을 때 약 36.0%의 단어오인식률(WER) 감소를 보였다.

참 고 문 헌

[1] A. D. Lawson, D. M. Harris, J. J. Grieco, "Effect of foreign accent on speech recognition in the NATO N-4 corpus", *Proc. Eurospeech*, pp. 1505-1508, 2003.

[2] D. V. Compernolle, "Recognizing speech of goats, wolves, sheep and ... non-natives", *Speech Communication*, Vol. 35, Nos. 1-2, pp. 71-79, 2001.

[3] L. M. Arslan, J. H. L. Hansen, "A study of temporal features and frequency characteristics in American English foreign accent", *Journal of the Acoustical Society of America*, Vol. 102, No. 1, pp. 28-40, 1997.

[4] R. Gruhn, K. Markov, S. Nakamura, "A statistical lexicon for non-native speech recognition",

Proc. ICSLP, pp. 1497-1500, 2004.

[5] A. Raux, "Automated lexical adaptation and speaker clustering based on pronunciation habits for non-native speech recognition", *Proc. ICSLP*, pp. 613-616, 2004.

[6] H. Strik, C. Cucchiarini, "Modeling pronunciation variation for ASR: A survey of the literature", *Speech Communication*, Vol. 29, Nos. 2-4, pp. 225-246, 1999.

[7] E. Fosler-Lussier, "Multi-level decision trees for static and dynamic pronunciation models", *Proc. Eurospeech*, pp. 463-466, 1999.

[8] I. Amdal, F. Korkmazasky, A. C. Suredan, "Data-driven pronunciation modelling for non-native speakers using association strength between phones", *Proc. ASRU*, pp. 85-90, 2000.

[9] S. Goronzy, S. Rapp, R. Kompe, "Generating non-native pronunciation variants for lexicon adaptation", *Speech Communication*, Vol. 42, No. 1, pp. 109-123, 2004.

[10] G. Stemmer, S. Steidl, C. Hacker, E. Noth, "Adaptation in the pronunciation space for non-native speech recognition", *Proc. ICSLP*, pp. 2901-2904, 2004.

[11] J. J. Morgan, "Making a speech recognizer tolerate non-native speech through Gaussian mixture merging", *Proc. InSTIL/ICALL Symposium on Computer Assisted Learning*, pp. 213-216, 2004.

[12] J. R. Bellegarda, "An overview of statistical language model adaptation", *Proc. ITRW on Adaptation Methods for Speech Recognition*, pp. 165-174, 2001.

[13] G. Bouselmi, D. Fohr, I. Illina, "Combined acoustic and pronunciation modelling for non-native speech recognition", *Proc. Interspeech*, pp. 1449-1452, 2007.

[14] M. Kim, Y. R. Oh, H. K. Kim, "Non-native pronunciation variation modeling using an indirect data driven method", *Proc. ASRU*, pp. 231-236, 2007.

[15] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993.

[16] Y. R. Oh, J. S. Yoon, H. K. Kim, "Acoustic model adaptation based on pronunciation variability analysis for non-native speech recognition", *Speech Communication*, Vol. 49, No. 1, pp. 59-70, 2007.

[17] R. Weide, *The CMU Pronunciation Dictionary, release 0.6*, Carnegie Mellon University, 1998.

[18] M.-Y. Tsai, F.-C. Chou, L.-S. Lee, "Improved pronunciation modelling by properly integrating better approaches for baseform generation, ranking and pruning", *Proc. Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology (PMLA)*, pp. 77-82, 2002.

[19] D. B. Paul, J. M. Baker, "The design for the Wall Street Journal-based CSR corpus", *Proc. ICSLP*, pp. 899-902, 1992.

[20] S. J. Young, *et al.*, *The HTK Book (for HTK Version 3.2)*, Microsoft Corporation, Cambridge University Engineering Department, 2002.

[21] S. J. Young, P. C. Woodland, "Tree-based state tying for high accuracy acoustic modeling", *Proc. ARPA Human Language Technology Workshop*, pp. 307-312, 1994.

[22] S.-C. Rhee, S.-H. Lee, Y.-J. Lee, S.-K. Kang, "Design and construction of Korean-spoken English corpus", *Proc. ICSLP*, pp. 2769-2772, 2004.

접수일자: 2008년 11월 20일
게재결정: 2008년 12월 24일

▶ 오유리(Yoo Rhee Oh)
주소: 500-712 광주광역시 북구 오룡동 1번지 광주과학기술원
소속: 광주과학기술원(GIST) 정보통신공학과 휴먼컴퓨팅 연구실
전화: 062) 970-3121
E-mail: yroh@gist.ac.kr

▶ 김홍국(Hong Kook Kim) : 교신저자
주소: 500-712 광주광역시 북구 오룡동 1번지 광주과학기술원
소속: 광주과학기술원(GIST) 정보통신공학과 휴먼컴퓨팅 연구실
전화: 062) 970-2228
E-mail: hongkook@gist.ac.kr

▶ 이연우(Yeonwoo Lee)
주소: 전남 무안군 청계면 도림리 61번지 목포대학교
소속: 목포대학교 공과대학 정보공학부 정보통신공학전공
전화: 061) 450-2745
E-mail: ylee@mokpo.ac.kr

▶ 이성로(Seong Ro Lee)
주소: 전남 무안군 청계면 도림리 61번지 목포대학교
소속: 목포대학교 공과대학 정보공학부 정보전자공학전공
전화: 061) 450-2436
E-mail: srlee@mokpo.ac.kr