한국인 영어학습자 발화의 자동음성인식을 위한 영어 음성사전 구성*

장태엽
(한국외국어대학교 영어학과)

1. 서론

본 연구는 영어 자동음성인식 시스템에 한국어 학습자의 발화오류 및 발음변이를 음성사전에 등재시켜 전반적인 음성인식 성능을 향상시키는 것을 목적으로 한다. 즉, 이러한 음성인식 시스템은 영어 모국어 화자의 영어음성 뿐 아니라 한국어의 액센트가 섞인 한국인의 영어발화에 대한 인식률을 높일 수 있도록 음성사전에 하나의 어휘에 대하여 다양한 변이형태를 등재한다는 것이다.

인간의 음성언어는 여러 가지 요인에 의하여 각종 변이(variation)를 산출하게 되며 이러한 변이는 음향신호(acoustic signal)의 일관성을 저해하여 청자의 지각 또는 음성인식기의 인식을 떨어뜨리는 직접적인 요인이 된다. 이러한 변이를 청자의 인지기능이 처리하려면 정형에서 벗어난 발음열까지도 해당 어휘와 연계시킬 수 있는 신호해석의 범위를 넓혀야 한다. 음소, 음성 또는 이와 유사한 인식단위(phone-like-unit)로 구현된 음성인식기로서는 입력 신호의 인식을 위한 알고리즘이 수행될 때 다양한 형태로 나타난 일련의 발음열 음향신호를 이에 상응하는 언어학적 단위인 개별 어휘로 연계시키는 작업이 효율적으로 진행되어야 한다. 좀 더 구체적으로 표현하면, 음성사전에 나열된 각 어휘에 대해 정형화된 하나의 발음 이외에도 자주 나타나는 발음변이들을 포함시켜 줌으로써 인식수행에 도움을 줄 수 있게 된다.

적절한 발음변이사전을 구성하면 음성인식의 성능이 향상됨을 확인한 연구는 이미 많이 진행되었다(Heine 외 1998, Ward 외 2003, Nock & Young 1998, Yang & Martens 2000, Kessens 외 2003, Wester 2003, Kam & Lee 2002, Jang 2006 등). 하지만 본 연구에서처럼 외국어 학습자의 오류발음을 체계적으로 추출하여 발음변이사전을 구성함으로써 이미 구축되어 있는 음성인식기를 특별한 변형이나 재훈련 없이 바로 사용하여 인식률 향상을 시도한 연구는 유례가 드물다. Goronzy 외 (2004)는 독일어 인식기의 발음사전에 영어액센트를 가진 변이들을 추가시키는 방법으로 독일어인식성능의 향상을 보고하고 있지만 시험적인 환경에서의 실험에 그쳤고 변이의 선별을 통한 사전 성능의 최적화 등의 방법을 사용하지 않았다는 면에서 본 연구와 다르다. 좀 더 이르게, Bonaventura 외 (1998)는 영어, 스페인어, 이탈리아어 등의 언어에서 모국어 액센트가 외국어의 음성인식에 미치는 영향에 관해 논의하고 있지만 외국어 액센트로 인해 인식률이 저하된다는 확인에 그치고 있고 구체적인 개선을 위한 본격적인 실험을 수행하지 않았다. 음성인식 기법을 이용해 한국인의 발화오류를 분석한 연구로는 Kim 외 (2004)가 있다. 그 연구의 주요 목표는 영어 학습자의 오류를 진단하고 오류발화자에게 피드백을 주어 교육적인 효과를 추구하는 것이므로 인식률 향상을 궁극적 목표로 하는 본 연구와는 다르다.

본 연구에서는 먼저 한국인 영어학습자의 부정확한 영어발음이 실제 영어음성인식률을 저하시킴을 밝히고, 이에 대한 부분적인 해결책으로 학습자의 가능한 오류발화변이를 고려하

* 이 논문은 2005학년도 한국외국어대학교 교내 학술연구비 지원에 의하여 작성되었음.

여 최적의 사전크기의 조정단계를 거쳐 음성인식률을 향상시키는 방법을 제시하고자 한다.

2. 음성 데이터

본 연구에서 사용된 음성 데이터는 크게 두 종류이다. 첫째, 음성인식기를 구축하는 데 사용될 원어민 발화 영어데이터와, 둘째, 이 음성인식기가 한국인 화자의 발음을 어떻게 인식하는 지를 테스트하게 될 한국인 영어발화 음성데이터이다.

먼저, 원어민 발화 데이터는 음성정보기술산업지원센터(SiTEC)에서 구입한 English02 코퍼스 중 문장발화 데이터를 사용하였다. 이 데이터는 원래 400명의 20-30대 남녀 원어민 화자의 음성으로 구성되어 있는데 그 중 300명분을 구입하여 사용하였으며, 233개의 단어를 조합한 124개의 문장세트에서 각 화자가 일정 문장을 조용한 사무실 환경에서 읽어 총 3,678개의 문장토큰으로 이루어져 있다. 본 논문에서는 이 코퍼스를 세트-1 이라고 이름 붙이기로 한다.

한국인 발화 영어데이터는 역시 같은 기관에서 구입한 K-SEC 코퍼스(이석재 외 2003)를 사용하였으며 그 중 위 English02의 데이터들과 비슷한 발화 형식으로 구성된 set5를 사용하였다. 이 데이터는 총 322명의 한국어화자와 15명의 원어민 화자의 음성을 포함하고 있으며 출신지역, 남녀 성비 등에 대한 균형이 잡혀있으며 한국인 화자 중 영어권 거주 경험이 있는 화자는 전체의 1%를 넘지 않는다. 총 180단어로 구성된 36개의 문장을 각 화자가 발화하여 총 10220개의 발화토큰을 구성하고 있다.[1] 본 연구에서 세트-2로 부르게 될 이 코퍼스를 세 가지의 하위분류 세트로 나누어서 첫 번째 하위세트(세트-2a)는 사전확장용, 두 번째 하위세트는(세트-2b)는 한국인 음성인식테스트용, 세 번째 하위세트(세트-2c)는 원어민 음성인식테스트용으로 각각 사용되었다. 세트-2b와 세트-2c는 음성인식기 구축이나 사전확장 등의 목적으로 전혀 사용하지 않고 성능테스트를 위해서만 사용하여 공정한 평가를 도모하였다. 특히 세트-2b의 화자의 수를 세트-2c와 동일하게 15명으로 설정하고 세트-2b에 포함된 화자는 세트-2a와도 겹치지 않게 하여 토큰과 화자가 완전히 독립되게 구성하여 비교실험의 신뢰성을 확보하는데 노력하였다.

표 1 은 위에서 설명한 코퍼스 및 그 하위세트의 구성과 본 실험에서의 용도를 요약해서 보여주고 있다.

---

1) 잘못된 발화나 기계적 처리의 오류로 인한 토큰을 제외한 숫자이다.

표 1. 연구에 사용된 코퍼스의 분류와 명세

		세트-1	세트-2a	세트-2b	세트-2c
원래의 코퍼스		English02	K-SEC		
발화유형		낭독체 영어문장	낭독체 영어문장		
단어수		233	180		
문장수		124	36		
총 토큰수		3,678	9,240	468	512
화자정보	유형	원어민	한국인	한국인	원어민
	수(남,녀)	300(150, 150)	307	15(8, 7)	15(8, 7)
	국적	미국	한국	한국	미국, 캐나다, 뉴질랜드
녹음정보	환경	조용한 사무실	조용한 사무실		
	AD	16kHz, 16bit	16kHz, 16bit		
본 연구에의 용도		음성인식기구축	한국인 발화오류를 포함한 사전 구성	한국인 발화의 인식성능	원어민 발화의 인식성능

3. 한국인 영어학습자의 음성과 원어민의 음성에 대한 음성인식 테스트

영어 음성인식기는 정상적으로 영어 원어민의 음성을 훈련데이터로 사용하여 구축한다.[2] 그러나 이렇게 구축된 음성인식기가 한국인 학습자의 영어음성에 대해서도 같은 인식률을 보이는 지에 대한 기존의 연구는 없었다. 따라서 본 연구에서는 이에 대한 문제를 제기하고 영어 원어민의 음성을 모델링하여 음성인식기를 구축한 뒤에, 이것이 한국인의 영어음성을 인식하는 데 있어 원어민과의 얼마나 차이를 보이는지를 조사하고자 한다.

3.1. 영어 음성인식기

영어의 음성인식기는 Jang (2006)에서 구축된 형태를 본 연구의 데이터 입출력에 적합하도록 수정하여 사용하였으며 이 인식기의 구축과정은 다음과 같이 요약될 수 있다. 가장 보편적이면서도 현재까지 가장 뛰어난 성능을 보이는 HMM (Hidden Markov Model) 방식이 사용되었으며 40개의 유사음성단위(phone-like unit)를 설정하고 상태(state)수 3의 연속음성인식을 수행하도록 설계되었고, 사용된 음향파라미터는 12개의 MFCC (Mel Frequency Cepstral Coefficients), 1개의 Energy, 이들의 일차 도함수(13개), 이차 도함수(13개) 등 모두 39개이다. 기본적인 음성모델링이 진행된 후, 문장 발화에서 빈번히 발생 할 수 있는 짧은 휴지(short pause)를 기본휴지(silence) 모델로부터 추출하여 추가로 설정한 후, 적절한 Gaussian 분포의 mixture 증가 등의 방법으로 인식기의 성능을 향상시켰다. 일련의 작업들은 음성인식기 구축을 위한 공개소프트웨어인 HTK (HMM Tool Kit, 버전 3.1, Young 외 1996)을 이용하여 진행되었다.

이 때 사용된 데이터는 표 1에서 나타난 바와 같이 세트-1의 발화 토큰들이다. 300명에

2) 물론 한국인 학습자의 발화오류를 포함하도록 음성인식기를 구성하면 성능이 국지적으로 최적화 되겠지만 원래 목적인 원어민의 음성에 대한 정확도가 상대적으로 하강할 것이며, 학습자의 영어구사 수준 등의 다양성으로 인해 실질적으로 모든 학습자의 발화오류를 모델링하는 인식기를 구현하는 것 자체가 비현실적이다.

달하는 남녀 화자의 발화가 훈련데이터로 사용되었으므로 화자독립성을 구현한 것으로 판단할 수 있지만 훈련에 사용된 토큰의 개수가 3,600여개에 그쳐 여러 가지 변수에 강인한(robust) 최고성능의 인식기라고 판정할 수는 없다. 하지만 본 연구의 목적이 음성인식기의 성능제고가 아니라 음성인식기에 사용되는 사전의 구성과 역할에 초점이 맞추어져 있음을 고려하여 더 많은 데이터를 공급하는 작업을 진행하지 않았다.

마지막으로, 인식기의 구축 과정에서 사용된 음성사전에 대한 설명이 필요하다. 한 단어에 대해 기본적인 발음형태만을 가진 사전의 사용 대신 그 기본 사전에 발음변이를 생성한 후 인식에 효과적으로 이용되는 변이를 자동화된 방법으로 선별하여 구성한 발음변이사전을 사용하였다는 점이다. 이처럼 원어민의 발음변이를 생성, 선별하는 과정을 이용하여 한국인 화자의 주요 오류발음을 변이로 취급하고 사전을 구성함으로써 영어음성인식기가 한국인 영어발화를 효과적으로 인식토록 하는 것이 본 연구의 핵심적인 과정이다.

3.2. 인식테스트 방법

위에서 설명한 대로 음성인식기를 구축한 후 한국인 영어학습자의 음성과 원어민의 음성에 대한 인식실험을 수행하였다. 이때 사용한 데이터 세트는 각각 표 1에 나타난 세트-2b와 세트-2c이며 테스트 데이터의 디코딩에 그 효과가 입증된 Viterbi 알고리즘(Forney 1973)에 의한 디코딩 방식을 사용하였다.

인식테스트의 수행에 필요한 추가적 조치의 하나는 음성사전의 재구성이다. 인식기의 구축에 이용된 데이터 세트-1의 단어는 테스트에 사용된 세트-2b,c의 단어구성과 다르므로 사전을 다시 구성해야만 인식테스트를 수행할 수 있다. 즉 세트-1의 233단어에 대한 발음변이사전은 이미 구성되어 있으므로, 세트-2b,c의 180단어에 대한 발음사전을 구성한 다음 두 사전을 결합하는 과정이다. 이때 세트-2b,c를 위한 사전은 발음변이를 포함하지 않고 기본형(baseform) 발음만으로 구성하게 된다. 그 이유는 발음 변이를 구성하는 작업 자체가 한국인의 영어발음 오류 유형을 포함하게 한다는 취지이며 그 작업이 본 연구의 핵심적인 과정이 되기 때문이다.

마지막으로, 음성인식에 필수적인 부분의 하나는 언어모델(language model)이다. 본 연구에서는 이러한 언어모델의 역할이 결정적이지 않았음을 밝힌다. 왜냐하면, 단순히 인식률의 향상을 위해서라면, 한국인 영어발화의 텍스트 중 직접적으로 테스트에 사용되지 않는 세트-2a의 문장 전사를 모체로 *N*-gram 모델을 만드는 것이 필요하겠지만 문장의 개수가 36개에 불과한 코퍼스인 경우 인식과정 자체가 음향모델(acoustic model)과는 무관하게 언어모델에 전적으로 의존하는 현상이 발생하게 되고 결과적으로 100%에 가까운 인식률을 나타낼 것이 확실하다. 이렇게 되면, 본연구의 주된 목표인 발음오류의 모델링을 통한 음성인식 기능 향상의 정도를 판단할 수 없는 상태가 될 가능성이 많아진다. 이를 방지하기 위해 본 연구에서는 세트-2a의 텍스트와 더불어 테스트 데이터와는 무관한 세트-1의 텍스트를 혼합하여 bigram 언어모델을 구성한 후 언어모델의 가중치를 일정치 이상 넘지 않게 하여 인식을 수행하였다.

본 연구에서는 음성인식의 성능을 (1a)와 같이 계산된 단어인식 정확도(word accuracy)를 사용하여 나타내었고 음성인식 실험에서 빈번히 언급되는 단어 오인식률은 (1b)에 나타난 바와 같이 단순히 '100-단어인식률'로 나타낼 수 있다.

(1) a. 단어인식률(word accuracy, %)

$$100 \times \frac{N - (\text{대치}(substitution) + \text{삽입}(insertion) + \text{탈락}(deletion))}{N(\text{전체인식토큰의 수})}$$

b. 단어 오인식률(word error rate, WER, %)

100 - 단어인식률

3.3. 인식률 비교

이러한 방식으로 비교해본 원어민과 한국인 학습자의 발화 인식률의 결과는 표 2와 그림 1에 나타내었다.

표 2. 한국인발화와 원어민발화의 인식률(%) 비교

언어모델가중치	0	5	10	15	20
원어민발화인식률	27.83	70.60	84.04	87.78	88.19
한국인발화인식률	8.45	46.60	63.75	71.23	74.07

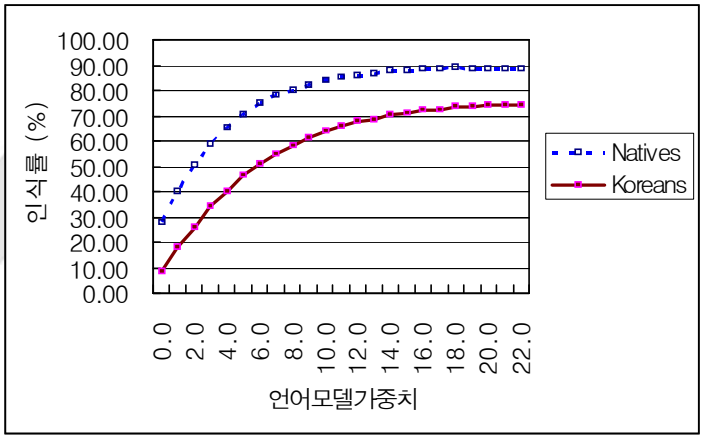

그림 1. 한국인발화와 원어민발화의 인식률 비교

위에 나타난 결과는, 언어모델 가중치의 크기와는 상관없이 한국인의 발화에 대한 인식률이 현저히 떨어지고 있음을 보여준다. 특히 언어모델의 가중치가 낮을 때 더 큰 차이가 나타나는 것을 볼 때, 원어민의 음향모델이 한국인의 발화를 인식하는 데 상대적으로 낮은 성능을 발휘함을 보임을 확인할 수 있다. 같은 문장의 내용을 같은 환경에서 발화한 두 테스터 데이터 세트가 이처럼 현격한 차이를 나타내는 결과는 한국인 영어발화의 오류에 기인한다고 결론짓는데 무리가 없음을 증명하고 있는 것이다.

4. 한국인의 발화오류를 포함하는 음성사전의 구성

위의 음성인식 실험에서 살펴본 것처럼 한국인 학습자의 영어발화는 원어민 발화에 비해 인식기능의 저하를 유발하였고 이를 개선시키기 위해 음성사전의 구성을 인위적으로 변화시키는 작업을 수행하였다.

음성사전의 재구성은 크게 두 가지 과정으로 나뉜다. 첫째, 각 어휘에 대해 가능한 한 많은 한국인의 영어발화 오류를 모두 포함시켜서 최대한의 발음변이 사전을 구성하는 것이다. 둘째, 이렇게 구성된 최대변이사전에 포함된 변이들 중 현실적으로 나타나지 않는 변이들을 제거시켜서 사전의 크기를 적당하게 조절하는 작업이다.

4.1. 기본 사전의 구성

변이를 생성하기 위해서는 각 단어 당 하나씩의 정형화된 발음열을 가진 기본형(baseform) 사전을 구성하는 일이 선행되어야 한다. 본 연구의 음성인식에 사용되는 단어들은 크게 두 세트로써 (1)음성인식기를 구현할 때 사용했던 코퍼스 세트-1의 233개 어휘와 (2)사전의 성능 테스트를 위해 새로 마련된 코퍼스 세트-2의 180개 어휘이다. 세트-1의 어휘에 대해서는 앞에서도 설명했듯이 발음변이를 포함한 사전이 이미 마련되어 있으므로 다시 구성할 필요가 없고, 세트-2의 어휘들에 대해서만 기본 사전을 구성하게 된다. 각 어휘에 대한 기본 발음열은 CMU 발음사전(Carnegie Mellon University pronouncing dictionary, 버전 0.6)을 이용하여 구성하였다.[3] 즉, 각 단어에 해당하는 발음열을 CMU 발음사전에서 찾아와서 음소세트의 필요한 변형을 거치는 간단한 작업이다. 그 후, 이미 준비되어 있는 세트-1의 발음사전과 합쳐서 정렬하는 과정을 겪으면 기본사전이 만들어진다.

표 3. 기본사전에 나타난 어휘와 발음의 예. (본 논문의 각 음성/음운 표기는 CMU 음소세트를 사용하였다.)

어휘	기본 발음	
	CMU phoneset	IPA 표기
alive	ax l aa ih v	ə l aɪ v
Henry	hh eh n r iy	h ɛ n r i
lake	l ey k	l e k
surprise	s axr p r ay z	s ə r p r aɪ z

본 연구에서 사용하는 기본 사전은 세트-1의 어휘 및 원어민 변이(233)와 세트-2의 어휘 및 기본 변이(180)를 합친 수에서 서로 중복되는 단어를 뺀 수인 총 408개의 변이개수를 가진 것으로 구성하게 되었다.

4.2. 최대변이의 생성

위에서 설명한 대로 기본사전이 구성되었으므로 각 어휘의 발음에 대한 변이형태를 최대한 생성해내는 과정이 필요하다. 여기서의 변이란 영어원어민이 생성해내는 발음의 변이가 아니라 한국인 화자가 정확하지 않은 발음을 구사할 때 나타나는 오류 형태로 국한한다. 이를 추출하기 위해 두 가지 방법을 사용하였다. 첫째, 한국인의 영어발음 오류를 다룬 기존의 문헌을 이용하는 방법이며, 본 연구에서는 정국(2005), 김기섭(2002), 구희산(2000), 장태엽(2005) 등에 기술되어 있는 발음오류나 발음오류를 만들어내는 규칙들을 선별하여 도입하였다. 두 번째 방법은, 실제 한국인의 발화나 발화데이터를 청취하거나 음향분석을 통해 분석해 내는 방법이다. 본 연구를 위해서는 코퍼스 세트-I의 데이터를 일부 청취하면서 발생하는 오류를 직관적으로 찾아낸 경우가 있었다. 하지만 주로 첫 번째 방법에 의해

3) 본 연구가 진행되는 현재, http://www.speech.cs.cmu/cgi-bin/cmudict에 공개되어 있다.

오류변이를 생성하였다.

이렇게 생성된 변이는 표 4에 정리하였다.

표 4. 발음오류변이 생성에 사용된 음운현상과 예

구분	음운현상	오류 발화의 예
a	마찰음을 폐쇄음으로 대체	file [f ay l] -> [p ay l] very [v eh r iy] -> [b eh r iy] bath [b ae th] -> [b ae s]
b	유음의 대체	lace [l ey s] -> [r ey s] right [r ay t] -> [l ay t]
c	비음의 유음화	only [ow n l iy] -> [ow l l iy]
d	유음의 비음화	Henry [h eh n r i] -> [h eh n n i]
e	반전음의 탈락	car [k a r] -> [k a]
f	ae 모음의 상승	bad [b ae d] -> [b eh d]
g	활음의 모음화	pie [p ay] -> [p aa ih]
h	활음 탈락	year [y iy r] -> [iy r]
i	구개음화	seat [s iy t] -> [sh iy t]
j	마찰음의 폐찰음화	measure [m eh zh ax r] -> [m eh jh ax r]
k	모음이완	food [f uw d] -> [f uh d]
l	모음첨가	lake [l ey k] -> [l ey k ax]
m	모음의 비원순음화	wall [w ao l] -> [w ax l]

표 4에서 나타난 13가지 현상들에 대해 환경을 설정하고 규칙화하여 모든 기본 발음에 적용시켜 자동으로 오류변이들을 추출할 수 있도록 스크립트를 작성하여 진행하였다. 단 이 규칙들은 전통적인 형태의 음운규칙 즉, 기저형에서 표면형으로 도출하는 생성음운론적 규칙으로 이해될 필요는 없으며 때로는 국지적으로 적용될 수도 있는 음성규칙(phonetic rule)이라고 볼 수 있다.

물론 본 연구에서 사용된 이 현상들이 한국인이 발화하는 발음오류의 전부 또는 대부분을 나타내지는 않으며 많은 다른 오류의 유형이 새로이 발견되어 발음오류변이로 추가될 필요가 있다. 특히 영어의 음절구조나 운율적 특징이 한국어와 다름으로 인해 발생하는 여러 가지 오류발화는 후속연구로 미룬다.

또한, 위에 나타난 현상들은 변이음으로의 변화를 수용하여 발화하는 데 실패해서 생기는 오류 등은 포함하지 않았다(예: 연구개음화 된 설측음으로 발화 하지 않는 오류, /s/뒤에서 성문음화되는 무성폐쇄음 등). 이런 상대적으로 미세한 발음현상들은 사전의 발음변이로 처리하지 않고 음성단위의 모델링에서 통계적인 변이로 수용하여 처리할 수 있다고 가정한다. 하지만 궁극적으로 비모국어의 발화를 발음변이 모델링으로 처리할 때에 이러한 변이의 어디까지를 수용하는 것이 적절한 것인가는 관련된 실험디자인을 면밀하게 구성하여 진행해본 후에 내릴 결론이며 본 연구의 범위를 벗어나는 것으로 판단한다.

어쨌든, 위의 13가지 현상을 규칙화해서 생성해낸 오류발음을 포함하고 있는 최대 오류발음변이사전은 표 5와 같은 모양을 지닌다.

표 5. 오류발음변이를 생성하여 기본사전을 확장한 최대 오류발음변이사전의 예. 각 단어의 변이 중 가장 위쪽의 변이가 기본 발음

어휘	기본 발음 및 오류발음변이
alive	ax l ay v ax l aa ih b ax l aa ih b ax ax l aa ih v ax l ay b ax l ay b ax ax l ay v ax
Henry	hh eh n r iy hh eh l l ih hh eh l l iy hh eh l r ih hh eh l r iy hh eh n n ih hh eh n n iy hh eh n r ih
lake	l ey k l eh ih k l eh ih k ax l ey k l ey k ax r eh ih k r ey k r ey k ax

표 5에서 볼 수 있듯이, 하나의 기본발음에 여러 가지 규칙이 동시에 적용될 수 있으며 이처럼 복수의 규칙이 적용될 경우 기하급수적으로 변이의 총 개수가 늘어나게 됨을 알 수 있다. 기본 사전의 총 변이인 408개에서 표 4에 나타난 13개의 규칙을 적용시킨 결과 총 1392개의 변이를 가진 최대 오류발음변이사전으로 확대되었다.

4.3. 발음변이의 선별

생성된 모든 오류발음을 포함하고 있는 표 5와 같은 최대 발음변이사전은 자동음성인식에 사용될 경우 시스템의 혼잡도(confusability)을 과도하게 증가시켜 오히려 성능을 떨어뜨릴 가능성이 있다(Kessens 외 2003). 직관적인 판단에 근거하더라도 기계적으로 생성된 모든 발음변이들이 실제로 발화되리라고 판단하기는 힘들다. 그러므로 이 발음변이들 중 빈번하게 발화되는 발음변이의 형태를 파악하여 그 순서를 정하고 일정기준에 못 미치는 변이들을 삭제하는 선별작업이 필요하다.

본 연구의 구체적인 목표중 하나는 음성인식기의 성능향상에 있음을 감안하여, 선별작업은 연구자 또는 영어교육 담당자의 직관적인 판단에 의존하기 보다는 음성인식기가 자동으로 판단하도록 하는 과정을 거치는 것이 실용적이라 할 수 있다. 그래서 본 연구에서는 사전에 등재된 각 단어의 개별 발음변이형이 음성인식기에 입력되는 실제 그 단어의 발화토큰과 얼마나 유사하다고 판정할 수 있는 지를 결정하는 실험을 수행하였으며 그 구체적인 방법은 다음과 같다.

먼저 이 판단을 내릴 인식기는 본 논문에 이미 기술한 실험(3장 참조)에서 사용된 것을 재사용하기로 한다. 그리고 인식용 데이터로는 아직 사용한 적이 없는 세트-2a를 활용한다.

그런데 현재의 실험은 단순한 인식테스트가 아니라 각 주어진 단어에 대해서 어떤 음성열을 매치시키는 가를 판단하는 작업이므로 입력으로 작용하는 각 발화 토큰이 어떤 내용의 발화인지에 대한 정보를 미리 제공하고 각 음성과 단어의 경계를 결정하는 역할만 요구하는 것이다. 즉 자동 음성 세그멘테이션(segmentation)에 사용되는 강제정렬(forced alignment) 방식을 이용하여 각 토큰의 음성열을 추출하고 이 음성열이 최대변이사전의 변이 중 어떤 것과 가장 유사한 지를 음향정보에 의존하여 확률적으로 선택하게 하는 과정이다. 이 과정을 겪은 후 한 단어의 각 변이가 음성인식기에 의해 몇 번 선택되었는지에 대한 빈도수(frequency)가 자동으로 계산되고 주어진 코퍼스에 포함되어 있는 해당 단어의 총수를 기준으로 각 변이의 출현확률이 쉽게 얻어진다. 이 과정은 (1) 같이 표현 할 수 있다.

(1) $P(V_i|W_p) = C(V_i) / C(V_{i...N})$

$P(V_i|W_p)$: 단어 W_p에서 i번째 변이가 나타날 확률

$C(V_i)$: V_i가 인식된 총 개수

$C(V_{i...N})$: 단어 W_p의 변이가 나타나는 총 개수

예를 들어, 단어 'bat'의 변이가 [b ae t], [b ae t ax], [b eh t], [b eh t ax] 4개로 생성되어 있고 각각의 빈도수가 4, 3, 8, 2 라면 각각의 확률은 0.24(4/17), 0.18(3/17), 0.47(8/17), 0.12(2/17)가 된다.

이 과정에 의해 계산된 확률을 각 변이에 기술한 확률오류변이사전은 표 6의 형태로 나타낼 수 있다.

표 6. 확률오류변이사전의 예: 최대 오류발음변이사전에 각 변이의 확률이 주어진 형태. 각 단어의 변이 중 가장 위쪽의 변이가 기본형 발음

어휘	기본 발음 및 오류발음변이	확률
alive	ax l ay v	0.72385
	ax l aa ih b	0.02092
	ax l aa ih b ax	0.00418
	ax l aa ih v	0.10879
	ax l ay b	0.05021
	ax l ay b ax	0.02929
	ax l ay v ax	0.06276
Henry	hh eh n r iy	0.13109
	hh eh l l ih	0.02622
	hh eh l l iy	0.12734
	hh eh l r ih	0.03371
	hh eh l r iy	0.06367
	hh eh n n ih	0.12734
	hh eh n n iy	0.45693
	hh eh n r ih	0.03371
lake	l ey k	0.48133
	l eh ih k	0.09129
	l eh ih k ax	0.02905
	l ey k ax	0.00415
	r eh ih k	0.07054
	r ey k	0.31120
	r ey k ax	0.01245

표에 나타난 단어 ‘alive'와 'lake'의 경우는 기본형(baseform)의 확률이 가장 높으므로 한국인에 의해서도 비교적 정확도가 높이 발화되었음을 알 수 있고 ‘Henry'의 경우 [hh eh n n ih] ([hɛnni])와 같은 오류발음이 가장 빈번하게 나타남을 확인할 수 있다. 기본형의 확률은 결국 한국인 학습자들이 그 어휘를 얼마나 오류 없이 발화할 수 있는지에 대한 잣대로서 활용할 수도 있을 것이다.

이렇게 확률오류변이사전이 완성되면 어떤 변이를 수용하고 어떤 변이를 삭제할 것인가는 적정한 임계값(threshold, T)을 어떻게 설정하는가에 따라 결정되며, 본 연구에서의 결정과정은 뒤 5.2에서 기술하였다.

이 확률오류변이사전은 음성인식기의 성능과 별도로 개별 한국인 영어학습자의 수준이나 오류유형을 판정하는데도 유용하다. 개별학습자의 음성발화를 테스트 토큰으로 입력하고 확률을 계산해낸 후 높은 확률값을 보인 오류를 따로 모아 성향을 분석하면 빈번한 오류의 유형을 파악할 수 있고 전체적으로 기본 발음이나 적법한 발음변이를 발화한 확률을 모아보면 전체적인 발음의 정확도를 판정할 자료로 이용이 가능할 것이다. 나아가서 발음의 정확도 측정을 컴퓨터와 음성인식기법을 이용해 보다 객관적으로 할 수 있는 도구를 제작하는데도 도움이 될 것이다.

5. 오류발음변이사전을 이용한 음성인식

5.1 오류발음사전의 효용성

위에서 언급한 것처럼 오류발음변이사전은 그 자체로 많은 효용성을 지닌다고 판단할 수 있지만 이 사전이 직접적으로 음성인식기의 인식률 향상에 도움이 되는 지를 검증해볼 필요가 있다. 이 작업은 새로 구성한 오류변이사전을 이미 구축해 놓은 음성인식기와 연동하고 데이터 세트-2b, 세트-2c를 테스트 데이터로 사용하여 실험하였다. 즉, 위 3장에서 진행하였던 실험에 새로운 사전을 투입하여 사용하는 것이며 이렇게 나온 결과 또한 위 3장에서의 한국인의 인식결과와 대비해서 분석하였다.

우선 표 7에서는 앞의 표 2에 나타난 오류발음변이를 사용하지 않은 상태와 사용한 상태에 대한 비교를 제시하였다. 오류변이를 사용한 아래의 경우는 발음변이의 수를 줄이는 선별과정을 거치지 않은 상태, 즉 모든 발화변이를 다 포함하고 있는 경우이다.

표 7. 오류발음변이의 사용여부에 따른 한국인발화와 원어민발화의 인식률(%) 비교

언어모델가중치		0	5	10	15	20
오류변이를 사용하지 않음	원어민발화	27.83	70.60	84.04	87.78	88.19
	한국인발화	8.45	46.60	63.75	71.23	74.07
오류변이를 최대로 사용함 (괄호는 증감)	원어민발화	25.87 (-1.96)	67.58 (-3.02)	84.45 (+0.45)	88.25 (+0.47)	89.51 (+1.32)
	한국인발화	11.80 (+3.35)	47.99 (+1.39)	65.98 (+2.23)	74.49 (+3.26)	76.51 (+2.44)

우선 한국인 발화의 인식률을 살펴보면 오류변이를 사용하였을 때 인식률이 언어모델의 가중치의 증감에 상관없이 증가되는 경향을 볼 수 있다. 한편 언어모델 가중치가 0일 때의 인식률 증가는 3.35%, 가중치가 20일 때의 인식률 증가는 2.44%로서 언어모델을 인식에 많이 활용할수록 인식률 증가는 다소 둔화되고 있다. 인식 프로세스에 언어모델보다 음향모델을 많이 사용할 때 음성사전의 효과가 상대적으로 더 커진다는 것은 사전의 확대가 곧 인식률 증가와 연관되어 있음을 의미한다.

원어민의 경우 언어모델 가중치가 적은 경우(0, 5의 경우)에 오류변이 사전을 최대로 사용하면 사용하지 않았을 때 보다 오히려 인식률이 떨어지는 현상(표 7의 괄호에서 음수 값)을 볼 수 있는데 이는 혼잡도(confusability)의 증가에 따른 영향으로 판단된다. 특히 표 7의 결과는 최대 오류발음변이사전을 사용하여 인식을 수행하였을 경우이므로 적절한 선별작업을 거쳐서 사전의 크기를 조절할 필요가 있음을 암시한다. 원어민 발화의 경우도 언어가중치가 조금 높아짐으로써, 사전의 확대와 인식률 향상이 연관되어가는 경향을 보이며 이 결과는 언어모델과 음향모델의 적정한 조정을 통해 인식기능을 최적화시키는 작업이 필요함을 보여주기도 한다.

5.2 사전의 크기 결정

표 8과 그림 2는 언어모델을 20으로 고정하고 사전의 크기를 조정하는 역할을 하는 임계치인 T값에 따라 변화하는 사전의 크기와 그 사전을 사용했을 때의 한국인 화자 발화의 인식률을 보여주는 결과이다.

표 8. T값의 변화에 따른 사전의 크기(발음오류변이의 수) 및 인식률(%)의 변화

T 값	사전의 크기 (총 변이의 수)	인식률 (%)
0.0	875	77.75
0.1	657	79.14
0.2	**593**	**80.08**
0.3	552	79.93
0.4	519	79.35
0.5	508	78.54
0.6	495	78.06
0.7	486	77.97
0.8	478	76.37
0.9	476	75.88
1.0	473	74.07

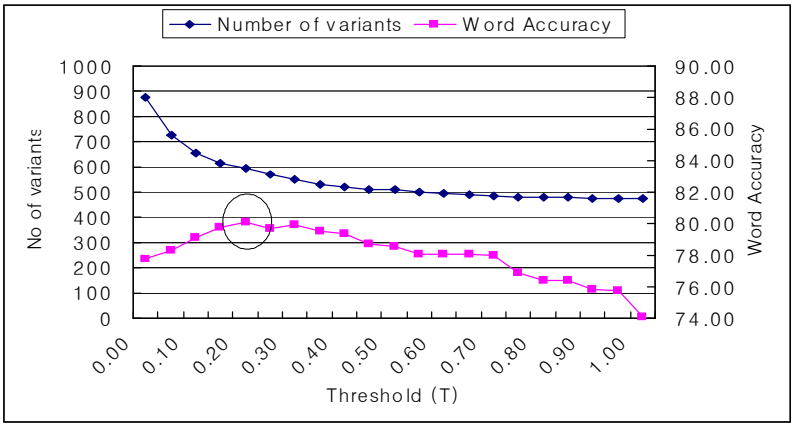

그림 3. T값의 변화에 따른 사전의 크기 및 인식률의 변화

표와 그림에 나타나 있는 것처럼, T값이 커질수록(즉 1에 가까울수록) 탈락되는 변이의 개수가 늘어나 최종 사전의 크기는 줄어들고, T값이 0에 가까울수록 많은 변이가 잔류하게 되어 사전은 크기가 늘어나게 된다. T값이 1인 경우, 즉 어떠한 변이도 허용하지 않고 각 단어별로 기본 발음만을 허용하는 사전으로 인식을 수행할 경우는 가장 인식률이 저하되는 것을 확인할 수 있다.

그러면 어떤 부분에서 최적의 T값을 고정시키고, 즉 사전의 크기를 결정하고 인식기에 투입하는가 하는 문제는 인식률의 변화를 기준으로 판단할 수 있다. 표 8, 또는 그림 2에서 인식률이 최고조에 달하는 점은 T값이 0.2(인식률은 80.8%)일 때이다. 즉 이 인식환경에서는 593개의 변이형을 포함하는 사전이 한국인 학습자들의 발음오류를 적절히 표현하여 음성인식에 실질적인 도움을 주는 것이다.

이 T값은 오류변이사전의 크기, 사용되는 데이터베이스의 특성, 화자 집단의 성향, 음성인식실험의 목적 등에 따라 달라질 수 있다.

표 9는 선별작업을 거쳐 최종사전에 수록된 변이형들의 예이다. 표에 나타난 위의 세 단어의 경우는 표 6의 모든 변이에서 확률값이 T(0.2)값 보다 높은 변이들로 구성된 것을 볼 수 있다.

표 9. 최종 발음사전의 예

어휘	기본 발음 및 오류발음변이
alive	ax l ay v
Henry	hh eh n r iy hh eh n n iy
lake	l ey k r ey k
likes	l ay k s l ay k s ax r ay k s ax
would	w uh d uh d

6. 결론

실험을 통해 얻어낸 본 연구의 주요 발견 사항은 첫째, 원어민의 발화를 모델링하여 구축한 자동음성인식시스템으로 한국인 영어학습자의 발화를 인식할 경우 뚜렷한 성능의 저하가 온다는 점이다. 둘째, 한국인 영어학습자의 발음오류들을 규칙화하고 이를 발음변이로 간주하고 이를 포함하는 음성사전을 구성한 후 적절한 임계치를 발견하여 변이의 개수를 인위적으로 조작하는 방법으로 국지적이지만 최적의 변이사전을 구현할 수 있게 된다. 더불어 본 연구에서 사용된 방식은 한국인의 영어발음의 숙련도나 주요 발음오류의 유형 등을 정량적인 분석을 통해 밝힐 수 있는 가능성을 보여주었다.

이와 같이 긍정적인 결과에도 불구하고 영어학습자 발화의 전체적인 인식률 향상을 위해서는 갈 길이 먼 것이 사실이다. 오류변이사전을 사용하지 않았을 때보다 영어학습자 발화의 인식 성능이 어느 정도 상승하였지만 여전히 원어민의 발화의 인식률에 비하면 크게 떨어지고 있다. 이를 해결하기 위해서는 본 연구에서 사용하였던 오류발음변이의 분석 외에 다른 여러 가지 분석과 기법의 개발이 병행되어야 할 것으로 판단된다.

본 연구의 의의중 하나는 확장성이다. 영어와 한국어 뿐 만 아니라 어떠한 2개의 언어가 있고 이 중 하나가 모국어이고 다른 하나가 학습되고 있는 외국어일 경우 본 연구에서 진행된 것과 동일한 방법을 적용시켜 인식성능을 향상시키거나 외국어 교육에 응용할 수 있을 것이다.

참고문헌

구희산(2000). 『영어음성학』, 한국문화사

김기섭(2002). 『음향 분석과 영어 발음교육』, 한국문화사.

이석재, 이숙향, 강석근, 이용주(2003). 「한국인의 영어 음성 코퍼스 설계 및 구축」, 『말소리』 46, 160-173.

장태엽(2005). 「한국어 영어학습자의 영어음성 데이터베이스 구축에 관한 연구」, 『언어와 언어학』 35, 293-310.

정국(2005). 『영어음운론의 이해』, 한국문화사.

Bonaventura, Patrizia, Filippo Gallocchio, Jean-Francois Mari, and Giorgio Micca. (1998), “Speech recognition methods for non-native pronunciation variations”, in Strik etal. 1998, 14-22.

Forney, G. D.(1973), “The Viterbi algorithm”, *Proceedings of the IEEE* 61, 268-278.

Goronzy, Silke, Stefan Rapp, and Ralf Kompe. 2004. “Generating non-native pronunciation variants for lexicon adaptation”, *Speech Communication* 42(1), 109-123.

Heine, Henrik, Gunnar Evermann, and Uwe Jost.(1998), “An HMM-based probabilistic lexicon”, in Strik et al. 1998, 57-62.

Kim, Jong-Mi, Chao Wang, Mitch Peabody, and Stephanie Seneff.(2004), "An

interactive English pronunciation dictionary for Korean learners", *Proceedings of Interspeech (ICSLP) 2004*, Jeju, Korea, 1677-80.

Jang, Tae-Yeoub.(2006), "Generation and selection of English pronunciation variations using knowledge-based rules and speech recognition techniques," Studies in *Phonetics, Phonology, and Morphology* 12(2), 362-375. The Phonology-Morphology Circle of Korea.

Kam, Patgi and Tan Lee.(2002), "Modelling pronunciation variation for Cantonese speech recognition", *Proceedings of ISCA Workshop on Pronunciation Modeling and Lexicon Adaptation (PMLA 2002),* Estes Park, Colorado.

Kessens, Judith M., Catia Cucchiarini, and Helmer Strik.(2003), "A data-driven method for modeling pronunciation variation", *Speech Communication* 40(4), 517-534.

Nock, H. J., and S. J. Young. "Detecting and correcting poor pronunciations for multiword units", in Strik et al. 1998, 85-90.

Strik, Helmer, Judith M. Kessens, and Mirjam Wester (eds.).(1998), *Modeling pronunciation variation for automatic speech recognition*, Rolduc, The Nethelands. European Speech Communication Association, University of Nijemegen.

Ward, Wayne, Holly Krech, Xiuyang Yu, Keith Herold, George Figgs, Ayako Ikeno, Dan Jurafsky, and William Byrne.(2002), "Lexicon adaptation for LVCSR: speaker idiosyncracies, non-native speakers, and pronunciation choice", *Proceedings of ISCA Workshop on Pronunciation Modeling and Lexicon Adaptation (PMLA 2002),* Estes Park, Colorado.

Wester, Mirjam.(2003), "Pronunciation modeling for ASR-knowledge-based and data-derived methods", *Computer Speech and Language* 17(1), 69-85.

Yang, Qian, and Jean-Pierre Martens.(2000), "On the importance of exception and cross-word rules for the data-driven creation of Lexica for ASR", *Proceedings of 11th ProRisc Workshop*, Veldhoven, The Netherlands, 589-593.

Young, Steve, J. Jansen, Dave Ollason, and Phil Woodland.(1996), *HTK Book.* Entropic.