발성 평가를 위한 영어 음성인식기의 개발

박전규, 이준조, 김영창, 허용수, 이석재*, 이종현**
동아시테크(주) 기술연구소, *연세대학교 영어영문학과, **LTI/CMU

Development of English Speech Recognizer for Pronunciation Evaluation

Jeon Gue Park, June-Jo Lee, Young-Chang Kim, Yongsoo Hur, Seok-Chae Rhee*, Jong-Hyun Lee**
R&D Center/Dong-A Seetech Co., Ltd, *Dept. of English Lang. and Lit./Yonsei University,
**Language Technology Institute/Carnegie Mellon University

jeongue@donga.co.kr, scrhee@yonsei.ac.kr

Abstract

This paper presents the preliminary result of the automatic pronunciation scoring for non-native English speakers, and shows the developmental process for an English speech recognizer for the educational and evaluational purposes. The proposed speech recognizer, featuring two refined acoustic model sets, implements the noise-robust data compensation, phonetic alignment, highly reliable rejection, key-word and phrase detection, easy-to-use language modeling toolkit, etc., The developed speech recognizer achieves 0.725 as the average correlation between the human raters and the machine scores, based on the speech database YOUTH for training and K-SEC for test.

I. 서론

본 논문은 비원어민의 영어 발성 능력을 평가하고 오류 발성에 대한 적절한 피드백을 지원하면서 궁극적으로 영어의 유창성(fluency) 평가를 목표로 하는 발성 평가용 음성인식엔진의 개발과 그 성능 평가에 관한 보고를 목적으로 한다.

현재까지 제시된 영어 발성 평가를 위한 음성인식 엔진들의 일반적인 접근법은 그림 1과 같이 요약할 수 있다[4][5]. 우선 고성능 잡음 처리 기술, 핵심어 추출 기술, 고신뢰도의 거절(rejection) 기능, 단어, 구절, 문장 등 다양한 평가 목표를 유연하게 모델링 하는 g2p(grapheme-to-phoneme) 모듈 및 언어 모델, 이를 수용하는 사전 및 언어 모델 도구, 자동 음소 분절화(phonetic segmentation) 모듈 등 일반적으로 현재 기술 수준의 음성인식기가 가져야 할 특성을 모두 포함해야할 뿐만 아니라 음향모델 학습 단계에서 사용하는 일부 기능 들도 통합되어져야 한다.

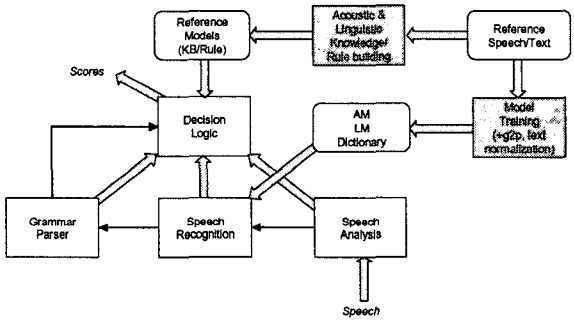

그림 1 발성인식을 위한 음성인식기의 기본 구조

한편 발성 인식이 아닌 해석의 관점에서 원어민뿐만 아니라 비원어민의 전형적인 발성 특성과 그 음향 음성학적 차이를 충실히 규명하기 위해서는 가능한한 다양하고도 충분한 지식을 시스템에 수용하는 노력이 필요하다[5]. 이를 위해 목표언어의 분절음 및 비분절음에 대한 특성을 음향음성학적인 지식을 총동원하여 지식베이스화하고, 발성평가 전문가의 지식이나 공인 영어발성 평가 기관에서 사용하는 수준의 평가 지침 및 지식을 지식베이스화하는 것이 필요하다. 마지막으로 위와 같은 점을 고려한 원어민과 비원어민의 발성 데이터베이스의 정교한 설계와 구축이 필수적이다[2].

이러한 지식베이스에 기반하는 음성인식 엔진 및 발성평가 시스템의 일반화 및 실세계 적용을 위해서는 잘 훈련된 전문성이 있는 인간 평가자와 기계 평가자 간의 평가 오류를 최소화하도록 시스템 튜닝과 개선 과정을 반복해야 하는 것이 필수적이다[4].

발성 평가의 또 다른 목표는 비원어민의 발성 능력

개선을 위한 교수시스템의 개발에 있다. 이러한 시스템은 비원어민의 잘못된 발성이나 발음 습관을 개선하도록 우선 잘못된 발성을 포착한 다음 다양한 발성관련 지식을 토대로 음성, 문자, 동영상 등이 복합된 멀티미디어 피드백 메커니즘을 설계하고 구현하는 것이 중요시된다[3][9].

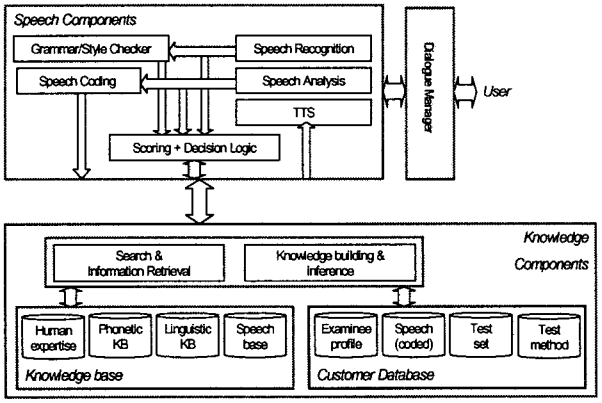

그림 2 발성 평가 시스템의 제안 구조

그림 2.는 음성인식기, 음성합성기, 문법해석기, 음성코딩, 지식베이스와 그 추론 엔진 등 음성처리 및 자연어처리 기술이 결합된 발성 평가 또는 교수 시스템의 예를 제시하고 있다. 본 논문에서는 이와 같은 시스템에 핵심적으로 적용이 가능한 음성인식 엔진의 개발 개념을 소개하고 그 성능 평가에 대해 기술한다.

II. 음성인식 엔진의 개발

2.1 시스템의 개요

본 논문에서 사용된 음성인식 엔진은 동아시테크(주)와 카네기멜른대학교의 연구진이 공동 개발한 연속음성인식 엔진인 SeeVoice를 근간으로 하며 음소 분절화, 고성능 잡음처리, 고신뢰도 거절기능, 핵심 단어 및 구절 인식, class n-gram 등을 지원하는 실시간 디코더를 특징으로 한다.

HMM 기반의 학습엔진을 사용해서 두 가지의 반연속(semi-continuous) HMM 음향모델(AM)을 생성해서 사용하고 있다. 하나는 DARPA의 RM 데이터베이스와 같이 현재 미국영어를 8개 권역으로 나누어 균형있게 선정해서 기본 음향모델(default AM)을 만들었다. 다른 하나는 기본 모델에 어린이 화자의 특성이 강조되도록 MLLR 재학습 기법을 통해 생성된 어린이용 음향 모델(kid AM)이다. 각각의 간단한 특성은 표 1.에 제시되어 있다. 두 가지 모두 다양한 잡음소스를 고려해서 별도로 수집된 잡음 데이터베이스를 활용해서 잡음 HMM 모델을 생성해서 사용하고 있다. 각 AM은 4개의 코드북(12차 MFCC, 12차 delta MFCC, 12차 delta-delta MFCC, 3차 power-delta power-delta delta power)에 기초한 senonic HMM으로 정의된다.

표 1 기본 및 어린이용 음향모델의 특성

AM	CIP	CDP	잡음모델수
default	44	62,932	7
kid	53	35,624	10

발성 사전은 수작업으로 튜닝된 CMU 사전에 기초하며 여기에 존재하지 않는 새로운 단어의 음소열은 별도로 작성된 g2p 엔진을 적용하여 생성한다. 발성사전, g2p, 언어모델 생성 모듈은 하나의 LM(Language Model) 유틸리티에 포함되는데, 이 유틸리티는 엔진에서 사용하는 언어 자원을 통합 관리하여 목표 시스템 저작을 지원한다.

2.2 잡음 처리

본 논문에서 사용한 잡음 처리 기술은 데이터에 의한 방법과 보상기법에 의한 두 가지이다[6]. 우선 전체 데이터에 대한 학습을 수행하여 이를 베이스라인 시스템(baseline)으로 놓고, 전체 데이터중에서 15dB 이상의 음성 데이터를 선정해서 학습을 수행한 결과(15dBup), CDCN을 사용해서 학습을 수행한 결과, 마지막으로 VTS를 사용해서 학습한 결과를 비교 분석하고 있다. 그림 3.은 각 방법에 의한 인식 실험 결과를 비교하고 있다. 이때 기본 인식조건은 데이터나 보상기법에 의한 전처리 외에 표 1.에서 제시하는 바와 같이 잡음유형별로 HMM 모델을 사용한다는 것이다.

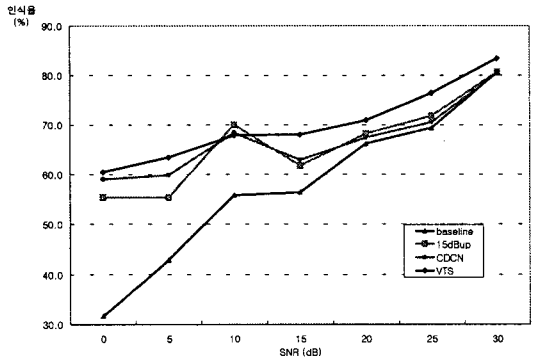

그림 3 잡음처리 방법에 의한 성능 향상

2.3 거절 기능

음성인식 엔진의 거절 기능은 기본적으로 음소 및 단어 신뢰도점수(confidence score) 계산[8]에 근거하며 최종 거절 기능의 수행을 위해 50개의 은닉층 노드로 구성된 단층 BP 신경망을 구현했다. [8]에서는 9개의

인식기에 기반한 특징값을 동원해서 거절 기능을 수행하고 있지만 본 논문에서는 부가적으로 다음과 같은 7가지의 특징값을 추가해서 사용하고 있다. 이러한 특징값에는 1. frame duration(단어별 프레임 길이), 2. normalized likelihood(프레임 길이에 의해 정규화된 로그 우도), 3. likelihood score(단어별 우도), 4. likelihood score path(현재 단어까지의 경로를 고려한 우도 점수), 5. top senone score(현재 단어에서 가장 큰 senone 점수), 6. lattice density(해당 단어로의 in-bound, out-bound 링크의 합), 7. phone perplexity (단어내 음소들의 평균 perplexity) 등이 포함된다. 생성된 신경망의 가중치 인자 셑을 인식 및 거절에 그대로 적용한다.

특징값의 추가에 따른 실험결과는 표 2.와 같다. 이때 성능평가를 위해 식 1.과 같은 AER(Annotation Error Rate)[8]을 정의하고 있다. 이러한 거절 기능의 구현을 위해 음향모델 학습에 사용되지 않은 RM 데이터중 1680문장(16,198단어)셑을 선정해서 실험에 사용했다. 기본 엔진의 성능은 이 테스트 데이터에 대해 95.728%의 단어 인식율을 나타낸다.

$$AER = \frac{vmber\ of \in correctly\ assigned\ tags}{total\ vmber\ of\ tags}$$

(식 1)

표 2 거절 성능의 향상 (단위: %)

특징수	AER	false alarm	missing
9	3.21	1.12	2.09
16 (9+7)	3.09	1.11	1.98

2.4 발성 평가를 위한 실험용 데이터베이스

음향 모델의 학습과 별도로 비원어민의 발성 정확도를 평가하기 위해 본 논문에서 사용하고 있는 발성 데이터베이스는 두 종류로서 하나는 카네기스피치사의 어린이 발성 데이터베이스인 YOUTH, 두 번째는 K-SEC의 한국인 어린이 발성 데이터인 K-SEC이다.

YOUTH는 56명의 남자 어린이와 79명의 여자 어린이로 구성된 총 135명의 미국 본토 어린이가 61개의 단어와 118개의 문장으로부터 구성된 평균 188개의 문장씩 발성한 총 25,122개의 문장으로 구성되어 있다. 녹음 조건은 Andrea NC-72 헤드셑을 사용해서 16kHz, 16bit linear PCM, 모노로 녹음되어 있다[2].

K-SEC은 한국학술진흥재단의 협동연구지원과제로 추진중인 "한국인의 영어발음 음성코퍼스(K-SEC) 설계, 구축 및 그 활용방안에 관한 연구"의 연구결과로 구축된 영어발성 데이터베이스이다. 본 연구에서는 이 가운데 10명의 한국인 어린이와 2명의 원어민 어린이를 선정하고 각각 36개의 문장을 문장단위로 수작업으로 레이블링하여 실험에 사용하고 있다. 원래 DAT로 녹음된 데이터를 역시 16kHz, 16bit linear PCM, 모노로 샘플링해서 사용했다.

Ⅲ. 실험 결과 및 토의

3.1 실험

우선 학습 데이터(YOUTH)에 자동 음소 분절화를 통해 음소 로그사후확률 점수(phone log-posterior probability score), 음소기반 지속시간 점수(phone -based duration score), 음절시간점수(syllabic timing score) 등의 세 가지[4] 특징값을 추출했다. 다음 원어민의 음향 음성학적 특징을 바탕으로 하는 평가 시스템의 성능 변화를 측정하기 위해 다섯 가지의 태스크를 설정하였다. 이때 태스크는 문맥종속음소(triphone)과 문맥독립음소(uniphone)를 우선적으로 고려하고, 음소단위별로 벡터양자화(VQ)와 스칼라양자화(SQ)를 적용하는 것으로 세분하였다. 마지막으로 목표 음소 및 코드워드별 통계량으로서 평균과 표준편차에 근거하는 사전확률(prior probability)를 추정했다. YOUTH 데이터로부터는 총 50개의 uniphone과 4,431개의 triphone이 생성되었다. 태스크는 다음으로 정의했다.

1. baseline: uniphone과 triphone에 대해 각 음소별 특성의 통계량을 추정
2. uniphone: uniphone을 대상으로 음소별 특성의 통계량을 추정. 각 음소에 대해 VQ를 적용하여 16개의 코드워드를 생성
3. triphone: triphone별로 단일한 음소별 특성의 통계량을 추정
4. VQ: baseline의 uniphone과 triphone 데이터를 대상으로 해당 음소에 속한 데이터에 대해 최대 4개의 코드워드로 구성되는 clustering을 수행한 다음 각 cluster별로 음소별 특성의 통계량을 추정
5. SQ: 4. VQ와 달리 Phone Duration에 대해서만 양자화를 수행. 각 triphone에 대해 최대 4개의 코드워드로 구성되는 VQ를 수행한 다음 해당 cluster에 대해 음소별 특성의 통계량을 추정

테스트를 위해서는 K-SEC의 12명 어린이 발성에 대해 문장별로 5명의 한국인 음성학자 및 3명의 원어민 평가자 총 8명이 평가한 뒤 이를 산술평균하여 인간 평가자의 평가 결과로 사용했다. 전체 문장에 대해 분석된 uniphone의 수는 42개, triphone의 수는 1,135개이다. 학습 데이터와 동일한 조건으로 문장별로 세 가지 특성값을 계산하였다. 다음 YOUTH로부터 추정

한 통계량과 사전확률에 대해 테스트 데이터의 사후확률(posterior probability)을 추정한다. 이때 원어민의 사후확률의 최대치와 최소치를 기준으로 정규화를 통해 100점 만점으로 환산된 기계평가 점수를 산출했다.

인간 평가자에게는 문장별로 음소, 단어, 문장의 유창성을 종합적(holistic)으로 판단하여 100점 만점을 기준으로 점수를 부여하도록 하였다. 전체 문장에 대해 한국인 음성학자 평균과 원어민 평가자 평균간의 상관지수는 r=0.906, 한국인 음성학자간의 채점결과는 평균 r=0.834, 원어민 평가자간의 채점결과는 평균 r=0.637을 나타냈다. 한편 인간 평가자와 기계평가 결과간 편차의 평균은 7.79점이며 표준편차는 7.04를 나타냈다.

표 3.은 지금까지의 평가 조건에 따른 평가 결과를 요약하고 있다. KOR은 한국인 음성학자 평균과 기계평가, NAT는 원어민 평가자 평균과 기계 평가, TOT는 전체 8명의 인간 평가자 평균과 기계 평가 간의 상관계수를 각각 의미한다. 표 3.에서 보는 바와 같이 본 실험에서는 VQ 추정에 의한 방법이 가장 높은 상관지수를 나타내고 있다.

표 3. baseline 시스템의 성능 상관계수(r=)

추정 대상	KOR	NAT	TOT
baseline	0.602	0.635	0.633
uniphone	0.557	0.619	0.601
triphone	0.627	0.698	0.677
VQ	0.685	0.733	0.725
SQ	0.657	0.702	0.695

기존의 방법에서는 실험에 사용된 사용자의 유창성 수준을 1-5 또는 1-6단계로 구분하고 있지만[1][4][7] 본 논문에서는 편의상 100점 만점으로 설정하고 있어서 단계별 평가를 적용할 경우 보다 높은 수준의 상관도가 기대된다.

Ⅳ. 결론

본 논문에서는 비원어민의 발성 평가를 위한 영어음성인식기의 구현과 그 기본적인 성능 평가에 대해 기술하고 있다. 정확하고 신뢰성있는 평가 시스템의 구현을 위해서는 음향 음성학적인 지식, 평가와 관련한 전문 지식 등을 지식베이스화하고 이를 유연하게 통합 또는 추론함으로써 인간 평가자의 수준에 이르도록 지속적으로 시스템을 튜닝하는 과정이 필수적이다.

본 논문에서는 그 초기단계로서 발성해석을 위한 기본 파라미터로서의 특징들을 음소 사후 확률, 음소 지속시간 및 음절 타이밍에 중점을 맞추어서 평가에 사용하고 있다. 이러한 특징을 적용한 결과 인간 평가자와 기계 평가간의 상관지수는 최대 0.725를 얻었다.

앞으로의 연구 개발은 현재 음성인식 엔진 위주의 분절음 평가 방식에 추가적으로 발성해석의 관점에서 생성되는 스펙트럼 특성, 강세나 억양을 비롯한 운율 특성 등을 통해 비분절적 요소에 대한 특성 및 파라미터를 추가함으로써 보다 향상된 인간 평가자와의 상관지수를 얻는 방향으로 이루어질 것이다.

참고문헌

[1] H. de Jong *et. al.*, "Relating PhonePass Overall Scores to the Council of Europe Framework Level Descriptors," *Technology in Language Education*, pp. 51-58, 2001.

[2] M. Eskenazi, "Pinpointing Pronunciation Errors in Children's Speech: Examining the Role of the Speech Recognizer," *2002 PMLA Workshop*, pp. 48-52, 2002.

[3] M. Eskenazi, "*Fluency - a Testbed for Foreign Language Tutoring,*" Technical Report, Carnegie Mellon University, 2002.

[4] H. Franco *et. al.*, "Combination of Machine Scores for Automatic Grading of pronunciation Quality," *Speech Communication 30*, pp. 121-130, 2000.

[5] J. Komissarchik *et. al.*, "BetterAccent Tutor - Analysis and Visualization of Speech Prosody," *InSTILL*, pp. 86-89, August, 2000.

[6] B. Raj, *Reconstruction of Incomplete Spectrograms for Robust Speech Recognition*, Ph.D Thesis, ECE Dept., CMU, April, 2000.

[7] C. Teixeira *et al.*, "Prosodic Features for Automatic Text-independent Evaluation of Degree of Nativeness for Language Learner," *2000 ICSLP*, pp. 187-190, October 2000.

[8] R. Zhang *et al.*, "Word Level Confidence Annotation using Combination of Features," *Eurospeech 2001*, pp. 2105-2108, 2001.

[9] 김효숙, "한국인을 위한 영어발음교정 시스템 'Dr. Speaking'소개," *대한음성학회 창립 25주년 기념 학술대회*, pp.47-50, 2002.

**본 논문은 정보통신부의 산업기술개발사업 (AA-2002-A3-0275-0001) 지원으로 수행되었습니다.