The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 16, No. 2, pp.131-136, Apr. 30, 2016. pISSN 2289-0238, eISSN 2289-0246

http://dx.doi.org/10.7236/JIIBC.2016.16.2.131
JIIBC 2016-2-16

비원어민 교수자 음성모델을 이용한 자동발음평가 시스템

An automatic pronunciation evaluation system using non-native teacher's speech model

박혜빈*, 김동헌**, 정진우***

Hye-bin Park*, Dong Heon Kim, Jinoo Joung*****

요 약 외국어 학습에서 발음학습은 가장 중요한 부분 중 하나이다. 발음학습 과정은 학습자의 발음에 대해 정확한 평가와 잘못된 발음이 있을 경우 적절한 피드백을 주어 이를 개선시키는 작업을 포함한다. 숙련된 평가자의 평가는 비용에서, 비숙련 원어민들의 평가는 일관성에서 문제가 있기 때문에 이를 보완할 수 있는 자동발음평가 시스템에 대한 연구가 진행되고 있으며 자동음성인식 기술의 활용이 각광받고 있다. 본 연구에서는 자동음성인식 기술과 비원어민 교수자의 음성 모델을 기반으로 단어 수준에서 학습자의 발음 정확성과 유창성을 평가하는 시스템을 구축하였고, 이를 통해 학습자들이 자신의 발음을 정확히 평가받고 평가결과에 따라 적절한 피드백을 받을 수 있도록 하였다. 또한 시스템의 성능평가를 통해 발음 정확성과 유창성에 대한 자동평가결과가 전반적으로 학습자의 실제 영어실력을 정확히 구분한다는 것을 확인하였다.

Abstract An appropriate evaluation on learner's pronunciation has been an important part of foreign language education. The learners should be evaluated and receive proper feedback for pronunciation improvement. Due to the cost and consistency problem of human evaluation, automatic pronunciation evaluation system has been studied. The most of the current automatic evaluation systems utilizes underlying Automatic Speech Recognition (ASR) technology. We suggest in this work to evaluate learner's pronunciation accuracy and fluency in word-level using the ASR and non-native teacher's speech model. Through the performance evaluation on our system, we confirm the overall evaluation result of pronunciation accuracy and fluency actually represents the learner's English skill level quite accurately.

Key Words : Automatic Pronunciation Evaluation, Automatic Speech Recognition (ASR)

Ⅰ. 서 론

외국어 학습에 있어 발음 학습은 가장 중요한 부분 중 하나이다. 학습자는 자신의 발음에 대해 정확한 평가를 받고 발음에 문제가 있을 경우 적절한 피드백을 받아 해당 부분을 교정함으로써 발음을 향상시킬 수 있다. 따라서 발음평가의 정확성과 오류 발음에 대한 피드백은 발음 학습에 있어서 매우 중요한 요소가 된다. 하지만, 사람이 발음을 평가할 경우 평가자에 따라 동일한 발음에 대해 상이한 결과를 낼 수 있으며, 같은 평가자라 하더라도 일관된 평가를 내리기 힘들다[1]. 이러한 점을 보완하기 위하여 컴퓨터로 학습자의 발음을 평가하는 자동발음평

*준회원, 상명대학교 컴퓨터과학과
**정회원, (주)지앤넷
***정회원, 상명대학교 컴퓨터과학과 (교신저자)
접수일자 : 2016년 1월 27일, 수정완료 : 2016년 3월 2일
게재확정일자 : 2016년 4월 8일

Received: 27 January, 2016 / Revised: 2 March, 2016 / Accepted: 8 April, 2016
*Corresponding Author: jjoung@smu.ac.kr
Dept. of Computer Science, Sangmyung University, korea

가 시스템이 연구되고 있다[5-9].

근래 연구되는 자동발음평가 시스템은 대부분 자동음성인식 기술(Automatic Speech Recognition: ASR)을 활용하여 구축되고 있다. 자동음성인식은 은닉 마르코프모델(Hidden Markov Model: HMM)을 기반으로 하는데[16] 입력이 들어오면 모든 기준 패턴과 비교한 후 가장 유사한 패턴을 찾아 출력한다[3]. 자동음성인식의 특성상 기준 패턴을 준비하기 위한 훈련용 음성코퍼스가 필요하며, 자동발음평가 시스템에서 이것은 원어민의 발음이 된다[3]. 따라서 자동발음평가 시스템은 학습자의 발음이 원어민 발음과 얼마나 유사한지를 계산하고 그것을 바탕으로 평가점수를 내리게 된다. 그런데 ASR 시스템은 보통 원어민 음성으로 훈련시키기 때문에 학습자의 음성을 인식한 결과는 원어민을 대상으로 인식한 결과보다 정확성과 신뢰도가 현격히 떨어진다[2]. 따라서 ASR이 학습자의 음성을 잘 처리하기 위해서는 acoustic model adaptation을 통해 인식공간을 확장해 줄 필요가 있다.

본 연구에서는 비원어민 교수자의 음성모델을 구축하고, 이를 바탕으로 단어 수준의 발음 정확성과 유창성을 평가하는 시스템을 구축하였다. 먼저 음성인식의 정확성을 높이기 위해서 acoustic model adaptation을 통해 음향모델을 확장하였다. 그리고 Sphinx4 음성 인식 엔진을 사용하여 발음 정확성을 평가하였고, 음성에서 단어 단위로 강도(intensity)를 추출하여 발음 유창성을 계산하였다. 다음 장에서는 자동발음평가 시스템에 관련된 연구를 소개하고 3장에서는 새로운 시스템을 제안하고 4장에서는 제안한 시스템에 대한 성능평가 결과를 소개하며 5장에서는 결론과 향후 연구방향을 제시한다.

II. 관련연구

CAPT (Computer-Aided Pronunciation Training) 시스템은 학습자의 발음을 평가하고 학습자들이 발음을 연습할 수 있는 환경을 만들어 주기 위해 설계되었다. CAPT 시스템은 크게 평가모듈과 피드백모듈로 나뉘는데 그중 평가모듈에서 발음평가는 일반적으로 전체적(holistic) 오류 검출과 국소적(pinpoint) 오류 검출로 나뉜다[5]. 전체적 오류 검출은 많은 음성 샘플을 검사하여 화자의 전반적인 유창성을 평가하고 국소적 오류 검출은 단어나 그 하위 수준에서 특정 발음 오류를 식별한다[5].

국소적 오류 검출을 기반으로 하는 대부분의 발음평가시스템은 ASR 모듈이 포함되어 있으며 ASR은 오류발음 검출과 피드백 생성을 지원한다. ASR은 음성인식 결과와 그 결과에 대한 확률 혹은 신뢰도 점수를 시스템에 제공하고 시스템은 결과물을 사용하여 음성을 평가하게 된다. 다음은 ASR을 이용한 연구들이다.

[5]는 CALL(Computer Aided Language Learning) 시스템을 이용하여 외국인들의 오류발음을 식별하는 문제에 집중했고, [6]은 발화에 대한 사전지식 없이 외국인 학습자의 발음을 평가하는 다양한 기술을 제안했다. [7]은 외국인들에게 아랍어를 가르치는 CAPL 시스템을 개발했는데, 발음 오류를 검출하기 위해 음성인식기와 음소 지속시간 분류 알고리즘을 사용했다. [8]은 HTK(Hidden Markov Model Toolkit)와 CMU(Carnegie Mellon University) Sphinx 음성인식기를 사용하여 발음점수화 시스템을 설계하였고, Timit Database로 훈련된 모델을 사용하여 인도영어음성을 평가했다. [9]도 Sphinx 음성인식기로 오류발음을 식별하고 forced-alignment와 edit distance를 이용하여 1과 10사이의 점수로 피드백을 주는 시스템을 설계하였다.

CAPT 시스템은 ASR에 크게 의존하기 때문에 ASR 성능이 전체 시스템의 성능을 결정하게 된다[2]. 그런데 ASR은 학습자와 달리 발음과 문법에서 흠이 없는 원어민의 음성 데이터로 모델을 훈련시키기 때문에 학습자 음성에 대해서는 결과의 정확성과 신뢰도가 떨어진다[2]. 따라서 인식공간을 확장하여 인식결과의 정확성과 신뢰도를 올릴 필요가 있다.

[13]은 각 단어에 대한 발음과 목표 학습자의 발음 변이를 포함한 확장된 발음사전을 생성하여 음성인식을 했다. 확장된 발음사전에서 발음 변이는 언어전이론(language transfer theory)을 기반으로 했다.

[2]도 확장된 발음사전을 생성하여 음성인식을 수행했는데, [13]과는 달리 한국인 영어 학습자가 영어 문장을 발화한 데이터를 모아서 한국인의 발음 변이에 관한 지식을 추출했고, 이러한 지식을 바탕으로 음소 수준의 발음 전사로부터 음소 발음 변이 규칙을 학습하기 위해 오류주도학습에 의한 접근방법을 설계하였다.

[14]는 음소 수준에서 발음을 평가하기 위해 신뢰도 점수 GOP (Goodness Of Pronunciation)를 계산하였다. GOP 점수는 사용되는 모델에 종속적이기 때문에 인식결과를 증진시키기 위해 화자 적응(speaker adaptation)을

사용하여 화자의 특정 스펙트럼 특성을 적응시켰다. 이때, 특정 음소 오류 패턴을 제외한 화자정규화를 제공하기 위해 MLLR (Maximum Likelihood Regression) 알고리즘을 적용하였다.

Ⅲ. 자동발음평가 시스템

본 연구에서는 단어 수준의 발음 정확성과 유창성을 평가하는 시스템을 구축하여 학습자들이 자신의 발음을 평가받고 평가결과에 따라 적절한 피드백을 받도록 하였다. 전체 시스템의 개요는 아래의 그림과 같으며, 시스템은 front-end, ASR, pronunciation evaluator로 나뉜다.

1. front-end

학습자가 시스템 상에 띄어진 문장을 읽으면, 이를 녹음하여 ASR에 해당 문장 정보와 함께 학습자의 음성을 넘기게 된다. 만약 ASR에서 유효하지 않은 데이터가 전송된다면 사용자의 발음 정확성이 임계치 이하라 판단하고 해당 문장을 다시 읽도록 유도한다.

ASR이 유효한 데이터를 전송한다면 데이터를 적절히 파싱하여 학습자의 발음 정확도와 유창성에 대한 점수와 함께 잘못 발음한 단어에 대한 정보를 화면에 출력해준다. 발음 정확도는 학습자가 얼마나 정확하게 발음을 했는가를 평가하고 발음 유창성은 학습자가 얼마나 목표와 유사하게 발음 했는지를 평가한다.

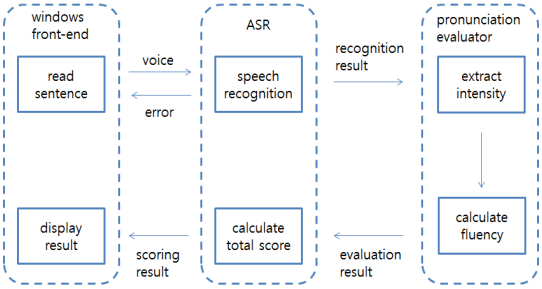

그림 1. 시스템 전체 개요
Fig. 1. System Diagram

2. ASR

front-end에서 음성이 전송되면 ASR에서는 해당 음성을 인식하여 텍스트와 시간정보를 얻는다. 발음 정확도를 계산하기 위하여 먼저 front-end에서 전달된 값을 바탕으로 DB에서 원 문장에 대한 텍스트를 얻는다. 그리고 Needleman-Wunsch[10] 알고리즘을 이용하여 원 문장과 음성인식 결과 텍스트를 정렬시킨다. 정렬시킨 결과는 다음의 표와 같이 나타난다.

표 1. 문자열 정렬 결과
Table 1. The result of string alignment

원 문장	people do some pretty crazy things on drugs and this crime would *** fit the bill
인식 결과	people do some pretty crazy things on drugs and this crime would the GOOD the AND

정렬시킨 결과에서 삽입(I), 삭제(D), 대치(S)가 이루어진 단어의 수를 센 후 아래의 공식에 따라 발음 정확도를 구한다. 이때 totalword는 원 문장에서 전체 단어수가 된다.

$$Accuracy = (1-(I+D+S)/totalword)*100 \quad (1)$$

구한 발음 정확도가 임계치 이하이면, ASR은 front-end에 오류 메시지를 보내고 front-end로부터 음성을 기다리게 된다. 반면, 발음 정확도가 임계치 보다 높을 경우 pronunciation evaluator로 음성인식 결과로 나온 텍스트와 시간정보, 음성을 전달한다.

pronunciation evaluator에서 평가결과가 전달되면, 발음 정확도와 발음 유창성을 아래 식에 대입하여 발음 정확도에 대한 발음 유창성 점수를 계산하게 된다.

$$total\ \ score = 100*(1-fluency/accuracy) \quad (2)$$

계산이 완료되면 ASR은 front-end로 평가와 관련된 모든 데이터를 전송한다.

3. pronunciation evaluator

ASR에서 음성과 인식결과(텍스트, 시간정보)가 들어오면, pronunciation evaluator는 입력 음성에서 단어 단위로 강도를 추출한다. 추출된 강도를 바탕으로 단어별 최솟값, 최댓값, 평균값, 평균증가량, 평균감소량, 단어지속시간을 계산한다. 이 값들을 각각 평균을 내어 전체평균 최솟값(TAvgMin), 전체평균 최댓값ITAvgMax), 전체평균 평균값(TAvgAvg), 전체평균 평균증가량(TAvgPos), 전체평균 평균감소량(TAvgNeg), 전체평균 단어지속시간(TAvgLen)을 구한다. DB에서 해당 문장에 대한 목표의 통계적 데이터를 가져와서 아래 식에 따라 유창성 점수를 구한다. 이때 아래 첨자로 t가 표시된 것이 목표의 데이터가 된다.

$$fluency = 100 * AVG(|(TAvgMin - TAvgMin_t)/TAvgMin_t|, |(TAvgMax - TAvgMax_t)/TAvgMax_t|, |(TAvgAvg - TAvgAvg_t)/TAvgAvg_t|, |(TAvgPos - TAvgPos_t)/TAvgPos_t|, |(TAvgNeg - TAvgNeg_t)/TAvgNeg_t|, |(TAvgLen - TAvgLen_t)/TAvgLen_t|) \quad (3)$$

Ⅳ. 실험 및 결과

자동발음평가 시스템의 성능을 평가하기 위하여 국내 영어학원에서 발음학습 관련 음원을 제공받았다. 학원에서 자체적으로 분류한 교수자, 우수학습자(high), 중간학습자(mid), 부진학습자(low) 네 종류의 음원을 제공받아 본 시스템이 세 그룹의 학습자들을 정확히 구분할 수 있는 지를 실험하였다.

ASR로 sphinx4[11] 음성인식 엔진을 사용하였고, 언어모델은 제공받은 음원에 대한 스크립트로 생성하였다. 발음사전은 카네기멜론 대학에서 제공하는 cmudict을 사용하였고, 음향모델은 voxforge_en_sphinx[12]를 기반으로 하여 목표인 교수자의 음성을 적응시켰다. 이때, 영어학원 으로부터 두 가지 버전(fast, slow)의 음성을 제공받았기 때문에 두 버전을 구분하기 위하여 fast 음성만을 적응시킨 모델과 slow만을 적응시킨 모델 2개를 생성했다. fast 버전의 학습자 음성에서는 fast 버전 음향모델을 사용했고, slow 버전의 학습자 음성에서는 slow 버전 음향모델을 사용하여 테스트했다.

1. 발음 정확도 평가

우수, 중간, 부진학습자들에게 5개 문장을 한번은 빠르게 한번은 느리게 말하게 한 후 그 음성을 녹음하여 테스트 파일로 사용하였다. 10개 파일에 대한 발음 정확도 평가결과는 아래 표2와 같다.

우수학습자는 발음 정확도가 평균 100%, 중간학습자는 평균 92.8%, 부진학습자는 73.2%로 정확도에 따라 영어실력이 구분됨을 볼 수 있다. 모든 음성파일의 평가결과가 발음 평가 임계치(50%)를 넘겼기 때문에, 발음 유창성 평가에서도 사용하였다.

표 2. 발음 정확도 평가
Table 2. pronunciation accuracy evaluation result

version	person	accuracy					
		1	2	3	4	5	avg
fast	high	100	100	100	100	100	100
	mid	79	87	100	93	100	91.8
	low	86	80	59	93	82	80
slow	high	100	100	100	100	100	100
	mid	100	94	100	93	82	93.8
	low	58	67	92	58	57	66.4

2. 발음 유창성 평가

10개 파일에 각각 강도를 뽑아낸 후 식(3)에 따라 계산한 발음 유창성은 다음 표 3과 같으며 소수 둘째자리에서 반올림하였다.

표 3. 발음 유창성 평가
Table 3. pronunciation fluency evaluation result

version	person	accuracy					
		1	2	3	4	5	avg
fast	high	10.0	4.1	6.3	7.6	6.6	6.9
	mid	13.5	13.1	16.6	13.6	14.3	14.2
	low	9.0	16.0	8.3	14.2	12.0	11.9
slow	high	5.5	15.2	17.1	27.8	14.7	16.1
	mid	6.2	24.6	13.0	15.5	28.2	17.5
	low	22.5	23.1	15.6	12.8	18.7	18.6

표 2와 표 3의 결과를 식(2)에 대입하여 구한 발음 정확도에 대한 발음 유창성은 다음 표4와 같다.

표 4. 발음 정확도에 대한 발음 유창성 평가
Table 4. pronunciation fluency evaluation based on pronunciation accuracy

version	person	accuracy					
		1	2	3	4	5	avg
fast	high	90.0	95.9	93.6567	92.3	93.4	93.1
	mid	82.9	84.9	83.4	85.4	85.7	84.5
	low	89.6	80.0	85.9	84.7	85.3	85.1
slow	high	94.5	84.8	82.9	72.2	85.3	83.9
	mid	93.8	73.9	87.0	83.3	65.6	81.3
	low	61.2	65.5	83.0	77.9	67.1	72.1

fast 버전에서 mid와 low가 구분이 잘 안되기는 하지만 전반적으로 우수학습자의 평균 발음 유창성이 88.5%, 중간학습자가 82.9%, 부진학습자가 78.6%로 발음 실력이 구분이 되는 것을 알 수 있다.

V. 결 론

본 연구를 통해 ASR를 기반으로 하는 자동발음평가 시스템에 대한 연구와 원어민 음성으로 훈련된 ASR이 가지는 문제점을 해결하기 위한 연구에 대해 살펴보았다. 기존의 연구들과 달리 단어 수준에서 학습자의 발음 정확성과 유창성을 평가하는 시스템을 구축하였고, ASR 인식 결과의 정확성과 신뢰도를 높이기 위해 원어민 음향모델에 학습자가 목표로 하는 집단의 음성으로 acoustic model adaptation을 하였다. 이를 통해 학습자들이 자신의 발음을 정확히 평가받고 평가결과에 따라 적절한 피드백을 받을 수 있도록 하였다. 시스템은 영어학원에서 제공받은 음원으로 평가하였고, 평가결과 발음 정확성과 유창성에 대한 점수가 전반적으로는 영어실력(상, 중, 하)를 구분하였다.

하지만, 발음 유창성 평가를 할 때, 음원에서 강도만을 뽑아 했기 때문에 학습자들의 성별에 따른 발음 차이를 고려하지 못했다. 이는 실험결과에서 오류로 나타났다. 따라서 학습자들의 성별에 영향을 받지 않는 특징을 추가로 뽑아서 발음 유창성을 평가하는 연구를 차후에 진행 할 것이다.

References

[1] Weonhee Yun. 2009. Discrepancy between Korean and Native English Raters Evaluating the English Pronunciation Spoken by Korean Learners of English. The Journal of Linguistic Science 48, 201-217.

[2] Jonghoon Lee. 2012. Error Simulation-based Pronunciation Feedback for Korean English Learners. PhD thesis, Division of Electrical and Computer Engineering Pohang University of Science and Technology.

[3] Weonhee Yun. 2012. The Objectives of English Pronunciation Evaluations and the Usability of Machine Scoring. The Journal of Linguistic Science 61, 167-184.

[4] Hyunsong Chung, Tae-yeoub Jang, Weonhee Yun, Ilsung Yun, Jaejin Sa. 2008. A Study on Automatic Measurement of Pronunciation Accuracy of English Speech Produced by Korean Learners of English. Language and Linguistic 42, 165-196

[5] Peabody, M. A. 2011. Methods for Pronunciation Assessment in Computer Aided Lanugage Learning. PhD thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.

[6] Moustroufas, N. and Digalakis, V. 2007. Automatic pronunciation evaluation of foreign speakers using unknown text. In Comput. Speech Language, 219-230.

[7] Sherif Mahdy Abdou, Salah Eldeen Hamid, M. R. A. S. O. A.-H. M. S. and Nazih, W. 2006. Computer aided pronunciation learning system using speech recognition techniques, in Interspeech.

[8] Chitralekha Bhat, K.L. Srinivas, P. R. 2010. Pronunciation scoring for indian english learners using a phone recognition system. In Proceedings of the First International Conference on Intelligent Interactive Technologies and Multimedia, 135-139.

[9] Srikanth, R. and Salsman, L. B. J. 2012. Automatic Pronunciation Evaluation And Mispronunciation Detection Using CMUSphinx. In 24th International Conference on Computational Linguistics 61-68.

[10] Needleman, Saul B., and Christian D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology 48.3, 443-453.

[11] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel. 2004. Sphinx-4: A flexible open source framework for speech recognition. Sun Microsystems Inc. Technical Report SML1 TR2004-0811.

[12] Hauswald, Johann, et al.. 2015. Sirius: An open end-to-end voice and vision personal assistant and its implications for future warehouse scale computers. Proceedings of the Twentieth International Conference on Architectural Support for Programming Languages and Operating Systems. ACM.

[13] Harrison, A. M., Lau, W. Y., Meng, H. M., and Wang, L. 2009. Improving mispronunciation detection and diagnosis of learners' speech with context-sensitive phonological rules based on language transfer. In INTERSPEECH 2787-2790.

[14] Witt, S. M., and Young, S. J. 1997. Language learning based on non-native speech recognition. In Eurospeech.

[15] Kim, S. D., Kim, W. S., & Woo, I. S. 2011. A Study on the Multilingual Speech Recognition using International Phonetic Language. Journal of the Korea Academia-Industrial cooperation Society, 12(7), 3267-3274.

[16] Jong-Young Ahn, Sang-Bum Kim, Su-Hoon Kim, Kang-In Hur, 2011. A study on Voice Recognition using Model Adaptation HMM for Mobile Environment Journal of Institute of Internet, Broadcasting and Communication (IIBC).

저자 소개

박 혜 빈(준회원)

• 2012년~현재 : 상명대학교 컴퓨터과학과 재학 중

<주관심분야 : 인공지능, 음성인식, 유무선 네트워크>

김 동 헌(정회원)

• 1983년 : 서울대학교 졸업 (학사)
• 1996년 : 서강대학교 대학원 졸업 (석사)
• 2013년 : 전남대학교 대학원 졸업 (전자상거래학박사)
• 1984~2000년 : IBM 워싱턴 D.C Voice 연구소 3년 근무, 한국 IBM AS/400 사업본부장, 한국 IBM Netfinity 사업본부장
• 2000년~현재 : ㈜지앤넷 대표이사

<주관심분야 : 음성인식, 인공지능>

정 진 우(정회원)

• 1997년 : NYU Polytechnic Institute 졸업(Ph.D in EE)
• 1997~2005년 : 삼성종합기술원 재직
• 2005현재 : 상명대학교 컴퓨터과학과 교수

<주관심분야: 인공지능, 음성인식, 유무선 네트워크, SoC design, Embedded system