JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 18, No. 6, pp. 1101-1107, Oct. 2017

자동음성인식 기술을 이용한 모바일 기반 발음 교수법과 영어 학습자의 발음 향상에 관한 연구

박 아 영[1*]
[1]창원중앙여자고등학교

The Study on Automatic Speech Recognizer Utilizing Mobile Platform on Korean EFL Learners' Pronunciation Development

A Young Park[1*]
[1]Changwon Jungang Girls' High School, 121, Danjeong-ro, Seongsan-gu, Changwon-si, Gyeongsangnam-do, Korea

[요 약]

본 논문은 스마트폰의 플랫폼에 내장되어 있는 자동음성인식 기술을 활용하여 영어 학습자의 발음에 대한 즉각적인 문자 피드백을 제공하는 모바일 기반 발음 교수법이 영어 학습자의 자음 발음 (V-B, R-L, G-Z) 인식과 출력에 미치는 영향에 대해 연구했다. 특히, 자동음성인식 기술을 이용한 모바일 기반 발음 교수법을 사용한 그룹, 전통적인 교사 중심의 발음 교수법 그룹, 그리고 이 둘을 합친 하이브리드 교수법 그룹으로 나누어 영어 학습자의 발음 평가 결과를 (인지, 출력) 비교, 분석했다. ANCOVA를 이용한 분석 결과, 영어 학습자의 발음 출력에 있어 하이브리드 교수법 그룹이 (M=82.71, SD =3.3) 전통적인 교수법 그룹 (M=62.6, SD =4.05) 보다 유의미하게 높은 결과를 나타냈다 (p<.05).

[Abstract]

This study explored the effect of ASR-based pronunciation instruction, using a mobile platform, on EFL learners' pronunciation development. Particularly, this quasi-experimental study focused on whether using mobile ASR, which provides voice-to-text feedback, can enhance the perception and production of target English consonants minimal pairs (V-B, R-L, and G-Z) of Korean EFL learners. Three intact classes of 117 Korean university students were assigned to three groups: a) ASR Group: ASR-based pronunciation instruction providing textual feedback by the mobile ASR; b) Conventional Group: conventional face-to-face pronunciation instruction providing individual oral feedback by the instructor; and the c) Hybrid Group: ASR-based pronunciation instruction plus conventional pronunciation instruction. The ANCOVA results showed that the adjusted mean score for pronunciation production post-test on the Hybrid instruction group (M=82.71, SD =3.3) was significantly higher than the Conventional group (M=62.6, SD =4.05) (p<.05).

색인어 : 발음, 자동음성인식, 피드백, 모바일 기반 언어 학습, 영어 학습자
Key word : Pronunciation, Automatic speech recognition, Feedback, Mobile learning, EFL learners

http://dx.doi.org/10.9728/dcs.2017.18.6.1101

Received 30 September 2017; **Revised** 22 October 2017
Accepted 25October 2017

***Corresponding Author; A Young Park**

Tel: +82-55-854-4058
E-mail: sonnik@hanmail.net

Ⅰ. Introduction

With increasing interest in digital devices in the field of second language (L2) acquisition, a variety of mobile technologies have been practiced in L2 pronunciation lessons. Among them, automatic speech recognition (ASR) has been highlighted in L2 pronunciation pedagogy [1]. ASR technology allows a computer and a mobile device to identify words that are read aloud or spoken into any sound-recording device. In the context of mobile devices (e.g. smartphone, tablet, personal digital assistant (PDA)), ASR applications or platforms that recognize the words and sentences which a person speaks into a microphone, and automatically convert them into written text [2]. In particular, recent developments of Multiple Spoken Language Technologies (MSLT) on the smartphone platform, such as the ASR technology voice-to-text (VTT), specifically in relation to text messaging, have greatly increased the potential application of smartphones in L2 pronunciation pedagogy. The VTT feature of the ASR in smartphone platforms provide feedback to encourage L2 learners to become more autonomous understand their pronunciation problems. Accordingly, VTT feature in ARS can be especially beneficial for L2 learners since they tend to be easily discouraged in their attempts to learn target pronunciation autonomously due to their limited abilities to monitor their pronunciation errors [3]. Moreover, conventional pronunciation lessons are unlikely to foster L2 learners' autonomy since conventional instruction has often focused on the teacher role such as monitoring errors and give feedback using recast and repetitions [4].

In addition, although many studies have been conducted to examine the effect of applying mobile ASR in the L2 pronunciation classrooms, the majority of them have focused on mainly mobile ASR-based commercial applications, (e.g. [2], [6], [7]), not mobile platforms. However, it is important to explore whether mobile ASR-based platforms are efficient tools for L2 learners' pronunciation development. This is because mobile platforms, unlike commercial applications, offer free, easy access features, which can be used without Internet connection and do not require downloads. Accordingly, this study explored whether ASR using a smartphone platform can be an effective pedagogical tool to enhance the pronunciation teaching and learning of target English pronunciation by comparing it to conventional instruction on EFL learners' pronunciation development.

Ⅱ. Background

2-1 ASR Technology in L2 Pronunciation Teaching

For the past two decades, the development of ASR in mobile devices such as smartphones, tablets, PDAs and media players has expended Mobile-Assisted Language Learning (MALL) [9]. Among diverse algorithms for ASR in mobile devices, the Hidden Markov Model (HMM) is one of the most predominant algorithms and has proven to be an effective method of dealing with speech at the sentence, word, or text level [7]. HMM computes the probable match between the input it receives and phonemes contained in a database of hundreds of native speaker recordings [10]. In short, ASR based on HMM algorithms computes how close the phonemes of a spoken input are to a corresponding modeland it provides highly reliable feedback of pronunciation information to the L2 learners [2], [11]. A number of scholars have advocated that ASR based on HMM has several benefits for L2 learners' pronunciation development (e.g. [11]). First, ASR can offer written feedback which helps L2 learners to identify individual pronunciation errors.Secondly, ASR's feedback possibly prevents learners from developing incorrect pronunciation habits [12]. While in a conventional pronunciation lesson teachers have limited time to observe individual pronunciation performance and provide customized feedback, ASR can provide opportunities for L2 learners to practice these tasks independently [13]. Thirdly, ASR can reduce L2 learners' learning anxiety because it can foster a safer space to practice pronunciation without face-to-face interaction with teachers [1].

Despite of all the benefits mentioned above, ASR has been criticized for low rates of accurate recognition for non-native speakers of the language such as L2 learners [6]. However, recently ASR seems to facilitate pronunciation improvement for diverse populations of learners [2], [10]. For example, [13] conducted ASR based instruction using software for L2 leaners. The results of [13]'s study suggest that ASR was helpful for learners in teaching pronunciation, especially for L2 learners who have limited L2 exposure and strong foreign accents. Therefore, the current study hypothesized that ASR based instruction can be beneficial for L2 pronunciation development of EFL (English as Foreign Language) learners.

2-2 Studies on Using ASR for Pronunciation Teaching

A number of ASR studies examined the effectiveness of ASR from software or applications and reported that it is useful in the development of L2 learners' pronunciation [6]. For example, [14] investigated the effect of the Microsoft Speech Application Software Development Kit in developing an oral skills training website for EFL learns. The ASR-based instruction allowed EFL

learners to practice their oral skills, such as speaking and pronunciation, and to obtain immediate feedback on their performance. The results suggest that most teachers and EFL learners enjoyed using this software because it could help them improve their English oral skills. [14] also pointed out that the ASR-based learning environment encouraged learners to produce more output in a low-L2 anxiety environment. In addition, [15] study explored the mobile application *TipTopTalk!* for L2 pronunciation training based on the minimal-pairs technique. SLT including speech recognition and text-to-speech conversion were integrated in this software. The results showed that L2 learners with low proficiency levels made relatively more progress than the rest. However, [15] suggested that it is desirable to design specific and individualized feedback in future versions of the application to avoid the performance drop detected after the protracted use of the tool. Finally, [7]'s study demonstrated that the ASR-based mobile application, *iFlytek Voice Input*, has a positive impact on improving English pronunciation accuracy for EFL learners. In line with [15], [7] also found that this *iFlytek Voice Input* can only give a written feedback to tell EFL learners which words they mispronounced, but fail in giving them the correct pronunciation needed to help improve their pronunciation.

In brief, an overview of the empirical ASR studies proposes a valuable research design for the present study. First, most of the ASR studies focused on commercial software or applications, neglecting others. However, this can be problematic because there are diverse useful ASR-based tools other than software and applications such as mobile platforms [16]. Therefore, the current study investigated the effect of mobile platforms that had free, easy access and could be used without Internet connection and downloading in comparison to commercial applications. Secondly, several studies found ASR's limitation in terms of L2 pronunciation development. Although ASR provides written feedback on L2 learners' pronunciation errors, it is not enough to correct or improve L2 learners' pronunciation. Therefore, to increase the effect of the ASR, the current study adopted the hybrid instruction model to supplement the limitation of the mobile ASR. Keeping these findings in mind, the research questions of the current study were set as follows:

RQ1: To what extent do ASR-based, Conventional, and Hybrid (Conventional + ASR) pronunciation instruction impact Korean EFL learners' perception and production of pronunciation development?

RQ2: What is the most effective instruction in Korean EFL learners' perception and production of pronunciation development, among ASR-based, Conventional, and Hybrid (Conventional + ASR) pronunciation instructions?

Ⅲ. Methodology

3-1 Participants

This experimental study consists of three intact classes of 117 Korean university students which were assigned to three groups: a class of 35 students who received ASR-based instruction, a class of 31 students who received conventional instruction, and a class of 48 students who received Hybrid instruction. All participants were freshmen with an average age of 19 years and 8 months. The three intact classes chosen for the current study were all low-intermediate level classes. Their average score of TOEIC was 487. Most of the participants began learning English at age 10 as a formal education from elementary and secondary schools. Despite that they had learned English for about 9 years, they did not have many opportunities to practice speaking English, and, therefore, they did not have the chance to practice and learn English pronunciation.

3-2 Treatment

As a treatment, the English pronunciation sessions were taken once a week for 20 minutes for one academic semester lasting 15 weeks and were taught by the same instructor. Target pronunciation for this pronunciation sessions were selected before this intervention based on participants' needs analysis, which identified their learning needs and difficulties. The selected Englishconsonants minimal pairs by the needs analysis for the current study were V-B, R-L, and G-Z.

As shown in Fig. 1, ASR Group received ASR-based pronunciation instruction and were given immediate textual feedback via text message on the smartphone platform. During 20-minute pronunciation sessions, pronunciation training activities consisted of reading aloud the target minimum pairs and phrases in English using the smartphone platform independently. After each reading attempt, students were provided with immediate written text feedback by the smartphone.

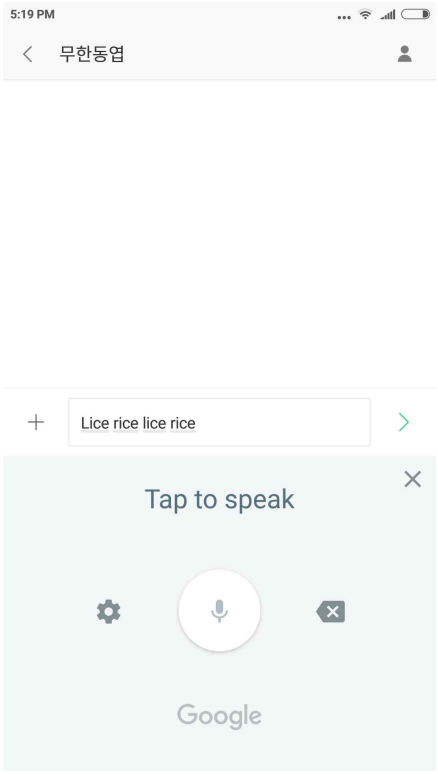

그림 1. 스마트폰의 플랫폼에 내장되어 있는 자동음성인식 기술을 활용한 영어학습자의 발음에 대한 문자 피드백
Fig. 1. Written visual feedback from ASR-based instruction by the smartphone platform

The Conventional Group received conventional face-to-to pronunciation instruction in the form of individual oral feedback by the instructor. Although the Conventional Group did not have access to mobile ASR, they also had to complete the same activities that the ASR Group did. They read aloud the same target English consonants minimal pairs and had weekly 20-minute sessions with an English instructor who provided immediate oral feedback such as recast and repetitions of their pronunciation.

The Hybrid Group received a combination of ASR-based pronunciation instruction and conventional pronunciation instruction. Pronunciation instruction consisted of two phases. First, students practiced their target consonants minimal pairs independently using the ASR-based smartphone platform to identify their pronunciation problems. Second, students received face-to-face oral feedback from the instructor focusing on problematic pronunciation identified by the smartphone platform feedback.

3-3 Pronunciation Test

To assess the participants' perception and production of target pronunciation before and after the intervention, the modified version of [1]'s pronunciation test was adopted. [1]'s test has been widely used in SLA research since it effectively measures both perception and production of pronunciation. To make test items suitable to measure target pronunciation, English consonant minimal pairs (V-B, R-L, and G-Z), this study revised and removed items from [1]'s test and included items that aim to test minimal pairs of target pronunciation selected for the current study. The first section of [1]'s test focused on perception of target pronunciation including two questions: 1) Circle the word you hear 2) Circle the word you hear from the minimal pairs. And the second section focused on production of pronunciation including two questions: 1) Say the words 2) Say the minimal pairs. The total number of test items for each section was 10. To save all the spoken pronunciation performance of participants during the test, mobile ASR on an iPod Touch using a commercial ASR application, *Nuance Dragon Dictation*, a speaker independent dictation system designed for continuous speech recognition was used. The reason behind is that mobile platform was not an efficient tool to save the large amount of data systematically. Fig. 2 shows a example of pronunciation test saved in *Nuance Dragon Dictation*. The maximum score for the pronunciation performance test was 100 and the minimum was 0. To maintain reliability of scoring and to minimize rater bias, the participants' tests were scored by two English instructors who were native speakers of English with considerable experience in grading pronunciation in speaking tests. The grading inter-rater reliability was 0.73.

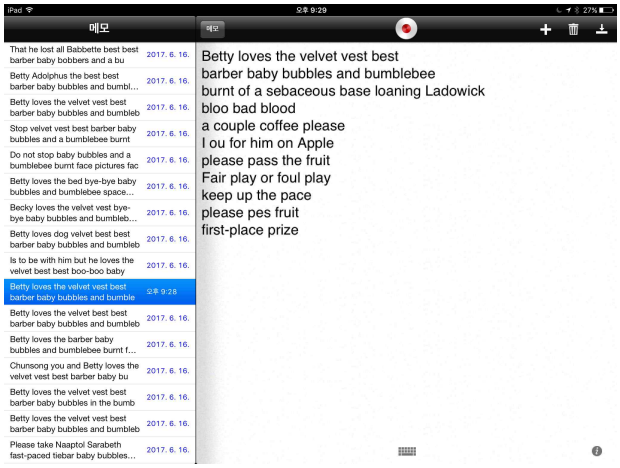

그림 2. 목표 발음 평가를 위한 발음 시험 (발음 출력) 예시
Fig. 2. Pronunciation Test (Production) for target minimum paris of target pronunciation

4. Data Analysis

To address research question 1 (RQ1), a paired t-test was conducted to examine whether there was a significant difference between pronunciation test scores (perception and production of pronunciation tests respectively) at Time 1 (pre-test) and Time 2 (post-test) for each group (the ASR group, the Conventional group, and the Hybrid group).

To address research question 2 (RQ2), firstly an independent t-test was carried out to examine whether there was an initial difference among the three groups' pronunciation test scores at Time 1 (pre-test). The independent variable was the type of treatment (the ASR, Conventional, Hybrid instruction) and the dependent variable was the pronunciation test scores on the pre-test. Secondly, for perception test scores pronunciation, a one-way between-groups analysis of variance (ANOVA) was applied to compare the changes among the three groups between Time 1 and Time 2. However, for productive pronunciation, due to initial difference of among the three groups in the t-test at Time 1, a one-way between-groups analysis of covariance (ANCOVA) was applied to control initial differences between the three groups and for a more statistically accurate comparison among the three groups at Time 2.

Ⅳ. Results

For RQ1, a paired t-test results demonstrate that the mean test score of the Hybrid group significantly increased between Time 1 and Time 2 in terms of both perception and production of target pronunciation as summarized in Table 1 below. However, the ASR group and the Conventional group showed a significant increase only in production not in perception.

표 1. t-test를 통한 Time1 과 Time2 시험 점수 평균 비교
Table 1. Comparison of Mean Score by a Paired t-test Between Time1 and Time2

Group	Mean	Std. Deviation	T	df	Sig. (2-tailed)
Hybrid (Perception)	8.80	14.06	4.338	47	.00
Hybrid (Production)	65.02	21.07	21.377	47	.00
ARS (Perception)	5.42	15.89	2.022	34	.051
ARS (Production)	51.43	21.91	13.884	34	.00
Conventional (Perception)	6.21	19.16	1.805	30	.081
Conventional (Production)	34.68	27.35	7.058	30	.00

These finding reveal that all three instructions (Hybrid, ASR and Conventional) had a significantly positive effect on Korean EFL learners' production of pronunciations. Meanwhile, in term of perception of pronunciation, the ASR and Conventional groups failed to have significantly positive impacts on L2 learners' development. That is, among the three instructions, only the Hybrid instruction, which provided a combination of written textual feedback by mobile ASR and an instructors' oral feedback, was beneficial in developing both perception and production of pronunciations. ASR-based instruction alone was insufficient to enhance both perception and production of L2 learners' pronunciation. This finding of the current study is in line with existing ASR studies such as [1] and [2]'s studies. But it is worth noting that there is one difference between [2]'s study and the current studies' result. That is, the current study included the Hybrid group, which was not usually included in previous ASR studies, and which emphasize the efficiency of the mobile ASR technology.

For RQ2, a one-way ANOVA result showed that the mean test score of the Hybrid group increased at Time 2 the most, the ASR group increased second to most, and the Conventional group increased the least in terms of perception of pronunciation as summarized Fig. 3.

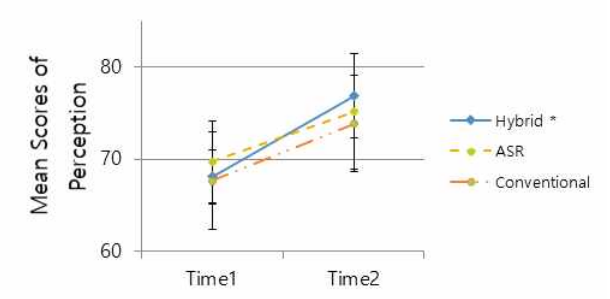

그림 3. ANOVA를 통한 Time1 과 Time2 시험 점수 (발음 인지) 평균 비교
Fig. 3. Comparison of Mean Score of Perception by an One-Way ANOVA between Time1 and Time2

This pattern was observed similarly in production of pronunciation as presented in Fig. 4. In terms of production of pronunciation, a one-way ANCOVA result shows that the Hybrid group increased at Time 2 the most, the ASR group increased second to most, and the Conventional group increased the least.

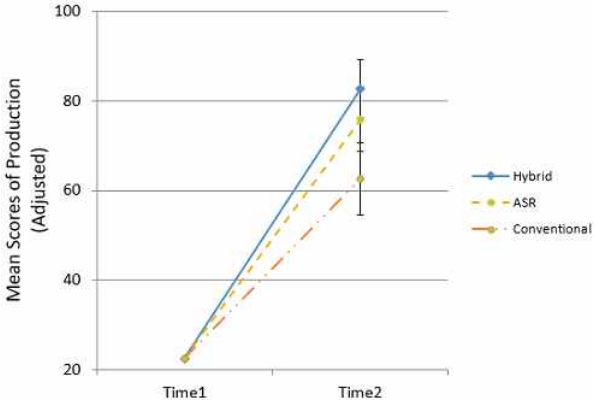

그림 4. ANCOVA를 통한 Time1 과 Time2 시험 점수 (발음 출력) 평균 비교

Fig. 4. Comparison of Mean Score of Production by an One-Way ANCOVA between Time1 and Time2

The main difference between the production results and the perception results was that in the perception results the mean score of the Hybrid group increased significantly more than two other groups. Meanwhile, the hybrid group and the ASR group increased significantly more than Conventional group in production. In short, in production of target pronunciation development L2 learners significantly benefit more from the Hybrid and ASR-based instructions compared to the Conventional instruction. This finding lends support to a number of studies such as [2], [17] that show how ASR-based instruction can give reliable and helpful feedback on pronunciation improvement over the intervention.

Among the three instructions in the current study, Hybrid instruction was the most effective in both perception and production of target pronunciation development. One possible explanation for the Hybrid group's improving the most compared to the other groups was that through Hybrid instruction, ASR-based instruction and Conventional instruction compensate each other and create a synergy effect. That is, written feedback from the ASR-based instruction might only identify learners' pronunciation errors. However, to correct and improve learners' pronunciation errors, it is required that they have more specific customized teacher's feedback on how to practice and perform the accurate target pronunciation from Conventional instruction.

Ⅴ. Conclusion

The result of the current study provided evidence in line with existing ASR literature that ASR-based instruction via the smartphone platform successfully promotes L2 learners' pronunciation [6]. Moreover, the current study demonstrates that integrating ASR-based instruction into the conventional instruction significantly improves not only L2 learners' production but also perception of target pronunciation compared to the ASR-based instruction only. At least one possible explanation for this result is that ASR only instruction is useful in detecting learners' pronunciation errors, but not in correcting those errors.

The finding of the current study suggests several pedagogical implications. First, the overall success of the significant pronunciation improvement suggests that the Mobile ASR-based learning environment is propitious for the L2 learners' pronunciation development since it can provide VTT feedback to L2 leaners to identify their own pronunciation errors. This feedback allows L2 learners to be aware of their problems in pronunciation, which is critical procedure in resolving these problems [4]. Secondly, different from to previous ASR studies, the results of the present study do not suggest that, due to its significantly more positive impact, the ASR-based instruction can replace conventional instruction in L2 pronunciation classrooms. This is because the results revealed that applying ASR alone is not enough to correct and improve these pronunciation errors. Rather, ASR is more like a useful teaching tool which can complement Conventional instruction. This finding of the current study reveals that incorporation of ASR-based instruction has great potential offering new ways of constructing the learning experience of L2 learners' pronunciation lessons while fundamentally changing the balance between classroom and individual learning. Finally, the L2 teachers should be trained to be able to successfully integrate ASR-based instruction by the mobile platform into their pronunciation lessons to reduce their teaching burden and increase the autonomy of their students. Furthermore, it may seem that rapid growth of digital technologies are replacing the teachers' role in the L2 classroom [18]. However, the findings of the current study suggest that the advent of digital technologies continues to transform the L2 teachers' roles to fill the gap between technology and human capability when they willing to embrace the new technological development in their L2 classroom in this digital era.

The results of this study should be viewed in light of its limitations, the most obvious of which is the fact that intact classes were used rather than setting up truly experimental condition. From an experimental standpoint, it would be ideal to have a control group in the current study. However, from a pedagogical point, it was not ethical to apply no treatment at all for one class. In addition, because of a short intervention duration and a limited research scope that covered only three selected minimal pairs for the target pronunciation, findings of the current study need to be considered as tentative rather than conclusive.

Future research should incorporate a long-term intervention in order to more thoroughly explore the dynamic nature of the impact of the mobile ASR on the wide range of target pronunciation development.

References

[1] A. K. Elimat and A. F. AbuSeileek, "Automatic Speech Recognition Technology as an Effective Means for Teaching Pronunciation," *JALT CALL Journal*, Vol. 10, No. 1, pp. 21-47, 2014.

[2] D. Liakin, W. Cardoso, and N. Liakina, "Learning L2 pronunciation with a mobile speech recognizer: French /y/," *CALICO Journal*, Vol. 32, No. 1, pp. 12-25, 2015.

[3] J. A. Foote and K. McDonough, "Using shadowing with mobile technology to improve L2 pronunciation," *Journal of Second Language Pronunciation*, Vol. 3, No. 1, pp. 34-56, 2017.

[4] A. Baker, "Exploring teachers' knowledge of second language pronunciation techniques: teacher cognitions, observed classroom practices, and student perceptions," *TESOL Quarterly*, Vol. 48, No. 1, pp. 136-163, 2014.

[5] M. Celce-Murcia, D. Brinton, and J. Goodwin, *Teaching pronunciation,* 2nd ed. Cambridge: Cambridge University Press, 2010.

[6] S. M. Mccrocklin, "Pronunciation learner autonomy: The potential of Automatic Speech Recognition," *System*, Vol. 57, pp. 25-42, 2016.

[7] M. Li, M. Han, Z. Chen, Y. Mo, X. Chen, and X. Liu, "Improving English Pronunciation Via Automatic Speech Recognition Technology," in *Proceeding of Educational Technology (ISET), 2017 International Symposium*, pp. 224-228, 2017.

[8] I. S. Lee, "Applications of English education with remote wireless mobile devices," *Journal of Digital Contents Society*, Vol. 14, No. 2, pp. 255-262, 2013.

[9] E. J. Song, "A study on the system for on-line education by mobile," *Journal of Digital Contents Society*, Vol. 6, No. 3, pp. 149-155, 2005.

[10] R. Hincks, "Speech technologies for pronunciation feedback and evaluation," *ReCALL*, Vol. 15, pp. 3–20, 2003.

[11] J. van Doremalen, C. Cucchiarini, and H. Strik, "Automatic pronunciation error detection in non-native speech," *Journal of the Acoustical Society of America*, Vol. 134, No. 2, pp. 1336-1347, 2013.

[12] H. Franco, H. Bratt, R. Rossier, V. Rao Gadde, E. Shriberg, V. Abrash, and K. Precoda, "EduSpeak: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications," *Language Testing*, Vol. 27, No. 3, pp. 401–418, 2012.

[13] A. Neri, C. Cucchiarini, H. Strik, and L. Boves, "The pedagogy-technology interface in Computer Assisted Pronunciation Training," *Computer Assisted Language Learning,* Vol. 15, pp. 441-447, 2002.

[14] H. Chen, "Developing and evaluating an oral skills training website supported by automatic speech recognition technology," *Recall Journal*, Vol. 23, No. 1, pp. 59–78, 2011.

[15] C. Gonzalez-Ferreras and V. Carde ´noso-Payo, "Improving L2 Production with a Gamified Computer-Assisted Pronunciation Training Tool, TipTopTalk!," in *Proceeding of IberSPEECH 2016: IX Jornadas en Tecnolog´ıas del Habla and the V Iberian SLTech Workshop,* pp. 177–186, 2016.

[16] M. Gutiérrez-Colon Plana, P. Gallardo-Torrano, and M. Grova, "SMS as a learning tool: an experimental study," *The Eurocall Review*, Vol. 20, No. 2, pp. 33–47, 2012.

[17] S. Petersen, R. Sell, and J. Watts, "Let the students lead the way: An exploratory study of mobile language learning in a classroom," in *Proceedings 10th World Conference on Mobile and Contextual Learning (mLearn)*, Beijing, China: Beijing Normal University, pp. 55-61, 2011.

[18] J. M. Howard and A. Scott, "Any time, any place, flexible pace: Technology-enhanced language learning in a teacher education programme," *Australian Journal of Teacher Education*, Vol. 42, No. 6, pp. 51-68, 2017.

박 아영(A Young Park)

2009년 : 한국교원대학교 대학원 (영어교육석사)

2015년 : 영국 브리스톨 대학교 (교육박사-영어교육)

2002년~현 재: 교육공무원

※관심분야 : 디지털 교육, 디지털 교과서