DOI: http://dx.doi.org/10.16933/sfle.2021.35.1.91

Evaluating Score Reliability of Automatic English Pronunciation Assessment System for Education*

Hong, Yeonjung
(Korea University)

Nam, Hosung**
(Korea University)

Hong, Yeonjung & Nam, Hosung. (2021). Evaluating score reliability of automatic English pronunciation assessment system for education. *Studies in Foreign Language Education, 35*(1), 91-104.

The purpose of this study was to evaluate the pronunciation score reliability of an automatic pronunciation assessment system for education, SpeechPro, a commercially released and patented system but without a score reliability test. So it is necessary to ensure the commercial system's reliability. The method is to measure score agreement between SpeechPro and human raters. The database used is a paid English speech corpus of the native speakers and non-native speakers with score annotations of the three English raters. First, the inter-rater agreement was measured, and then the agreement between SpeechPro's scores and the raters' average scores were measured. The following 5 metrics were used: Pearson correlation coefficient, standardized mean difference, quadratic weighted kappa, exact percentage agreement, and 1-point adjacent percentage agreement. The results are that human-machine agreement is significantly identical to human-human agreement according to all the metrics used, proving the score reliability of SpeechPro. This provides a logical justification required before the comparison with other automatic pronunciation assessment systems.

Ⅰ. Introduction

The purpose of this study is to evaluate score reliability of an automatic pronunciation assessment system developed using automatic speech recognition (ASR) technology. The system is called SpeechPro (Patent No. 40-2020-0157130, 2020) and it was developed by NAMZ Labs, MediaZen. It has already been released commercially and is being used by many users, but its reliability has not been verified by comparison with well-trained human raters. Therefore, the focus of this study is to validate the reliability of the already released commercial product.

* This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2019S1A5A8033772).
** First author: Hong, Yeonjung, Corresponding author: Nam, Hosung

www.kci.go.kr

SpeechPro is a text-dependent pronunciation evaluator which outputs scores for the four language units: phoneme, syllable, word, and sentence. The acoustic model of SpeechPro is a deep neural network hidden Markov model (DNN-HMM), which is widely used for training ASR system because of its stability and high recognition accuracy (Dahl, Yu, Deng, & Acero, 2011). In this study, SpeechPro trained with native speech data of North American English was used, so a user's speech was measured based on American English accent. The method of evaluation was to measure agreement between human-rated word scores and machine-rated word scores on non-native English speech. The metrics used for measuring human-machine score agreement were Pearson correlation coefficient, standardized mean difference, quadratic weighted kappa, exact percentage agreement, and 1-point adjacent percentage agreement.

To test the reliability of SpeechPro's performance is important because it is impossible to move on to the next step to expand the system's features and functions without the score reliability ensured. The demand for computer-assisted pronunciation training (CAPT) system and computer-assisted language learning (CALL) in general will keep increasing due to the accelerated social trends towards distance learning using online platforms (Hanna, Barr, Hou, & McGill, 2020).

The outline of this paper is as follows: Literature Review section reviews previous literature about automatic pronunciation assessment and inter-rater reliability. Methodology section explains the overall architecture of SpeechPro, describes the non-native English speech database with human raters' scores, and delineates analysis methods. Results section reports human-human agreement results and human-machine agreement results. Discussion and Conclusion sections summarizes the results, limitations of this study, and suggests future research topics.

Ⅱ. Literature Review

The development of ASR technology naturally led to its application to pronunciation evaluation (Witt & Young, 1998; Herron, Menzel, Atwell, Bisiani, Danelozzi, Morton, & Schmidt, 1999; Menzel, Herron, Morton, Bomaventura, & Howarth, 2001; Beatty, 2003; Kim, 2006; Tafazoli, Huertas Abril, & Gómez Parra, 2019). It is because ASR technology enables learners to receive quantitative and immediate feedback on pronunciation accuracy for granular language units, which is more resource-consuming if human instructors are to take the role of assessment and feedback provision.

The common method to validate the reliability of machine-based pronunciation evaluation system is to prove its performance is similar to that of human evaluators (Kim, 2006; Chung, Jang, Yun, Yun, & Sa, 2008; Cincarek, Gruhn, Hacker, Nöth, & Nakamura, 2009; Loukina, Zechner, Yoon, Zhang, Tao, Wang, Lee, & Mulholland, 2017; Wang, Zechner, & Sun, 2018; Prafianto, Nose, Chiba, & Ito, 2019). Given non-native speakers' speech data and its transcript, human expert raters' scores on the speech are collected and their reliability as reference scores is evaluated by measuring inter-rater agreement. The next step is to measure human-machine agreement between the validated reference scores and automatically generated scores by the machine of the interest.

www.kci.go.kr

According to Yun (2009), human raters' performance can be random and arbitrary due to individual differences on perception and recognition so that inter-rater agreement and intra-rater agreement are not as high as expected due to according to Yun (2009). Therefore, given that the raters' expertise as instructors in ESL domain is guaranteed and that a set of objective scoring rubric is provided to the raters, randomness or inconsistency found in their scoring results can be ignored as noise.

This study replicates the methods used in the previous studies which measured score accuracy of an ASR-based pronunciation assessment system (Kim, 2006; Chung et al., 2008; Cincarek et al., 2009; Loukina et al., 2017; Wang et al., 2018; Prafianto et al., 2019). The five metrics used in this study were a comprehensive list of those used in the previous studies. The common methodology to evaluate a machine score's reliability is to measure its agreement with human raters' scores, and there are various metrics to be used. Chung et al. (2008) employs Pearson correlation coefficient and tests its p-value to measure correlation among the machine scores, Korean raters' scores and native English speakers' scores on Korean learners of English speech. Cincarek et al. (2009) measures inter-rater agreement on native English speech from six different countries speakers' speech from six countries using Pearson correlation coefficient and open correlation coefficient. Loukina et al. (2017) examines the validity of SpeechRater^{SM} developed by Educational Testing Service with the following five metrics: Pearson correlation coefficient, standardized mean difference, quadratic weighted kappa, exact percentage agreement, and 1-point adjacent percentage agreement. Wang et al. (2018) uses Pearson correlation coefficient, standardized mean difference, and quadratic weighted kappa to evaluate the performance of SpeechRater^{SM}.

Ⅲ. Methodology

1. SpeechPro

This study uses SpeechPro, an ASR-based pronunciation evaluation system. Given speech and text data, SpeechPro measures the degree of pronunciation accuracy of the speech for the text, and provides percentage scores for the following 4 units: phoneme, syllable, word, and sentence.

1) System Overview

Figure 1 describes the five parts of SpeechPro layout: Choosing a chapter of a book, choosing a sentence from the selected chapter, recording, receiving pronunciation accuracy scores represented as quality scores, and some features related with fluency level.

As it is shown in Figure 2, in the quality score section, each word of a sentence is presented as a clickable button which plays back an audio chunk for the word selected and shows a table of a word score, syllable scores, and phoneme scores. Language units with scores are colored green if it ranges from 80 to 90, orange if it ranges from 60 to 80, and red under 60. The color coding is intended to help users

www.kci.go.kr

receive more intuitive feedback with detailed score information.

Figure 2 shows the score results of the case where the user pronounced 'leave[liv]' when it is supposed to be read 'leaf[lif]'. As mentioned already, the syllable score 67 is the average of the phoneme scores 92, 69 and 40, and as the word 'leaf' is a mono-syllabic word, the word score is also 67. The sentence score is the average of the four word scores.

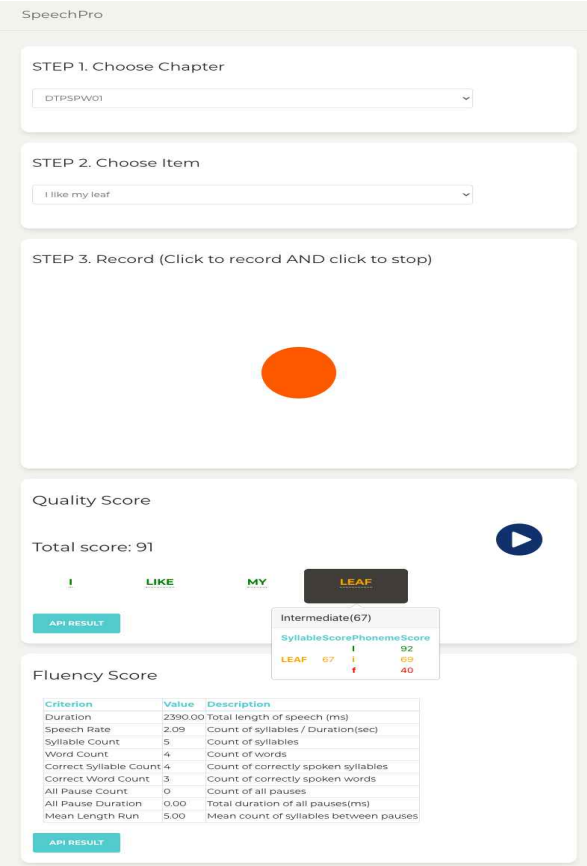

Figure 1. The layout of the SpeechPro system

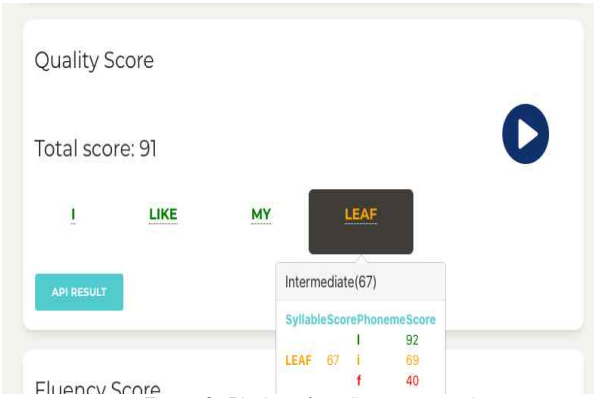

Figure 2. Display of quality score results

www.kci.go.kr

2) Model Components

SpeechPro is composed of an acoustic model, a pronunciation model, and a scoring model. The role of the acoustic model is to provide log-likelihood for each time frame of speech signal to be classified as one of the phonemes in the language. The DNN-HMM based acoustic model was trained using Kaldi, the most popular ASR toolkit (Povey, Ghoshal, Boulianne, Burget, Glembek, Goel, Hanneman, Motlicek, Qian, Schwarz, Silovsky, Stemmer, & Vesely, 2011). For training the model, Librispeech (Panayotov, Chen, Povey, & Khudanpur, 2015) and Wall Street Journal were used, both of which are American English read-speech databases and 1,573 hours in total. Pronunciation model maps each word of the inputted sentence to its corresponding phoneme sequence. For modelling American English pronunciation, Carnegie Mellon University pronouncing dictionary (CMUdict) is used, which is an open-source pronunciation dictionary for North American English that contains over 134,000 words and their pronunciations (Carnegie Mellon University, 2015). Since CMUdict is machine-readable, it is widely used for speech engineering domain where pronunciation modelling is required. The scoring model of SpeechPro outputs a percentage score which is phone log-likelihood normalized by native speakers' phone log-likelihood statistics.

3) Scoring Procedure

The scoring procedure has 6 steps in total.

(1) The user selects a transcript and reads it aloud.
(2) The transcript is mapped to the phoneme sequence using the pronunciation model.
(3) The phone sequence is forced-aligned to the recorded speech by the acoustic model so that the time each phone occurs in the recording is determined. In other words, each 20ms frame of the recording is labeled with a phone symbol in the order of the sequence from the transcript.
(4) Log-likelihood of the labeled phone for each time frame is measured by the acoustic model.
(5) The native speakers' phone log-likelihood statistics obtained from Librispeech are used for normalizing the user's log-likelihood value for each time frame. The normalized score ranges from 0 to 100.
(6) Phone score is the average of the normalized scores for the time frames within a phone. Word score is the average of the phone scores, and sentence score is the average of the word scores.

2. Database Description

For testing score reliability of SpeechPro, accented English pronunciation evaluation corpus (Speechocean, 2020) was used. For convenience, the database is referred to as APE in the rest of this paper. It is composed of native English speech, non-native English speech, and scores for each word and sentence rated by 3 raters who were recruited by APE creators. The total duration of the database is 11.38 hours spoken

www.kci.go.kr

by 22 speakers.

1) Speaker Description

The 22 speakers are grouped as native English speakers and non-native English speakers. The native English speakers are from USA and the non-native speakers are from India, Spain, Portugal, China, and Japan. The detailed information of speaker distribution is described in Table 1.

Table 1. Distribution of speakers and sentences in APE

Region	# Female	# Male	# Speakers (%)	# Sentences
US	2	2	18.20%	1,279
China	2	2	18.20%	1,513
Japan	2	2	18.20%	1,100
India	2	2	18.20%	1,158
Portugal	2	2	18.20%	1,598
Spain	1	1	9.00%	800
Total	**11**	**11**	**100%**	**7,448**

2) Rater Description

There are 3 human raters who are native English speakers from California, USA. They are English teachers experienced in teaching English as a second language. The raters transcribed each recording and rated pronunciation scores based on their own transcriptions. So, the prompt sentences that the speakers were told to read and the transcript sentences that raters transcribed can be different. The scoring units are word and sentence and the score is given in 1 to 5 scale, 1 being poor, 5 being excellent.

Word scoring standard is if its phoneme, syllable and word pronunciation is articulated in a correct manner, if its delivery is intelligible for native speakers, and if the volume and speed are appropriate.

3) Data Preprocessing

Out of 7,448 recordings in APE, the recordings which do not meet the following conditions were deleted:

(1) Transcripts of the 3 human raters show an exact match.
(2) Every word and sentence is labeled with scores from 1 to 5.
(3) Phonemes mapped to the transcript are found in the native English phone statistics which is used in the analysis.

The total of 139 recordings do not match those conditions. For example, some words are not rated by the raters by mistake not meeting the second condition, and a word 'sh-shooting' from one of the transcripts is mapped to 'SH_S SH_B UW1_I T_I IY0_I NG_E' but 'SH_S' is not found in the reference English phone statistics data since the phone 'SH' is usually not pronounced independently within a word. The number of

www.kci.go.kr

the selected sentences for analysis is 7,309 after the deletion.

Next, a list of the words with multiple pronunciations such as 'read' is sorted out and the correct pronunciation is chosen considering the context the word is used. The pronunciation model is updated with the determined pronunciation for every word.

3. Analysis Method

1) Metrics

The following 4 types of metrics are used to test human-human agreement and human-machine agreement.

(1) Pearson Correlation Coefficient (PCC) and open Pearson Correlation Coefficient (open-PCC)

Pearson correlation coefficient is defined as the covariance of the two variables divided by the product of their standard deviations. It is commonly employed for measuring the association between variables. Variables of interest can be both continuous and ordinal values. Pearson correlation coefficient ranges from -1 to 1, where -1 means perfect negative relationship, 0 means no relationship, and 1 means perfect positive relationship.

Since we have 3 human ratings, open Pearson Correlation coefficient is additionally measured, which is Pearson correlation coefficient between one rater's score and the average of the other 2 raters' scores.

(2) Standardized Mean Difference (SMD)

Standardized mean difference is defined as difference in mean outcome between groups divided by standard deviations of outcome among participants. This metric is employed when there are differences in measurement scales (Hedges, Pustejovsky, & Shadish, 2012). The absolute value of standardized mean difference is the size of the difference between the groups. The value 0 of standardized mean difference means that there is no difference between the groups in comparison. The larger the absolute value of standardized mean difference, the more different the two groups of values are.

(3) Quadratic Weighted Kappa (QKW)

Quadratic weighted kappa measures the agreement between 2 ratings and it is a commonly used metric for human-machine agreement (Chen & He, 2013). It ranges from -1 to 1, where 1 means the two raters show complete agreement, 0 means they show random agreement, and -1 means complete disagreement.

(4) Exact Percentage Agreement (EPA) and 1-point Adjacent Percentage Agreement (APA)

Exact percentage agreement measures the degree of exact agreement between 2 ratings. Additionally, 1-point adjacent percentage agreement is measured where the two scores that have 1 point difference are

www.kci.go.kr

also considered as showing agreement.

2) Target of Analysis

The units of pronunciation scores provided in APE are a word and a sentence but they are rated with different standards; word score standards are segmental level while sentence score standards are suprasegmental level. However, every type of score rated by SpeechPro is segmental level since its sentence scores are simply the average values of word scores within a sentence. Therefore, only the word scores of the 3 human raters and SpeechPro are the target of analysis in this study. The sample size of word scores from 7,309 sentences are 75,945.

3) Design of Analysis

(1) Human-human score agreement

Inter-rater agreement among the 3 raters is measured to test the reliability of the reference scores before measuring the performance of SpeechPro. The reference scores for the word unit are discrete numeric variables ranging from 1 to 5. The following 5 metrics are suitable for the type of scores of interest: PCC and open-PCC, SMD, QWP, EPA, and APA.

(2) Human-machine score agreement

The human-machine score agreement is measured between the SpeechPro's word scores and the 3 human raters' average word scores. SpeechPro's scores and the averaged reference scores are both continuous numeric variables but do not share the same range of values, PCC and SMD are used to measure agreement between those two score types.

The other metrics, QWK, EPA and APA, are also used to measure the agreement but since they are applicable only to discrete values, both the SpeechPro scores and the averaged reference scores are converted to 5 discrete numerical values for the analysis.

Ⅳ. Results

1. Human-Human Agreement

Table 2 describes the overall distribution of the reference scores of each human rater for the word scores and those for non-natives speakers only.

www.kci.go.kr

Table 2. Mean and Standard Deviations of Each Human Rater's Word Scores

		Score distribution					
		H1		H2		H3	
Score scale	Sample size	Mean	Std.	Mean	Std.	Mean	Std.
1-5	75,945	4.40	0.83	4.27	1.08	3.95	1.02

Figure 3 shows the frequency of each score label for the 3 human raters.

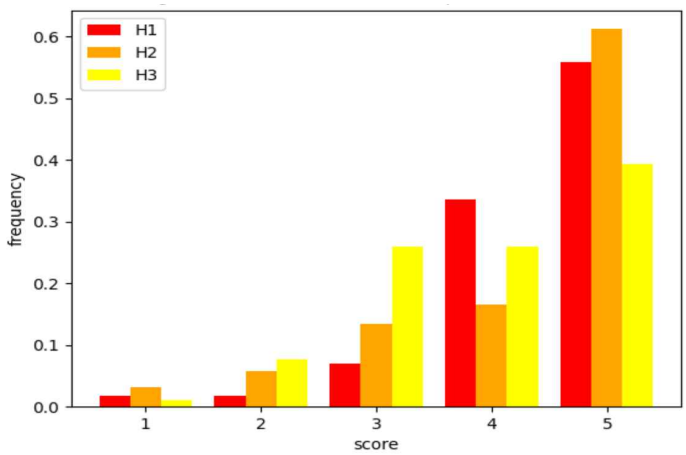

Figure 3. Histogram of reference word score frequencies for each rater

Table 3 shows human-human agreement measured by PCC and open-PCC. The effect size r of every correlation is provided in Table 3 and its p-value is lower than 0.001, proving that it is statistically significant.

Table 3. Human-Human Agreement Measured by PCC and open-PCC

PCC				open-PCC			
H1-H2	H2-H3	H3-H1	Average	H1-H2,3	H2-H1,3	H3-H1,2	Average
0.55***	0.27***	0.47***	0.43	0.64***	0.46***	0.40***	0.50

***: $p < 0.001$, ** : $p < 0.01$, * : $p < 0.1$

Table 4 shows human-human agreement measured by QWK, EPA and APA, SMD.

Table 4. Human-Human Agreement Measured by QWK, EPA, APA, and SMD

	H1-H2	H2-H3	H3-H1	Average
QWK	0.53	0.26	0.41	0.40
EPA (%)	58.02%	39.62%	52.00%	49.88%
APA (%)	91.34%	80.35%	98.31%	90.00%
|SMD|	0.13	0.30	0.48	0.30

www.kci.go.kr

2. Human-Machine Agreement

The distribution of SpeechPro's word scores for the 75,945 words are plotted in the upper left of Figure 2 using histogram. The mean and the standard deviation of the word scores are 82.47 and 18.16, respectively. The histogram of the averaged reference scores which are continuous numeric values from 1 to 5 is in the upper right of Figure 4. The lower left is the discretized version of SpeechPro scores, and the lower right is that of the averaged reference scores.

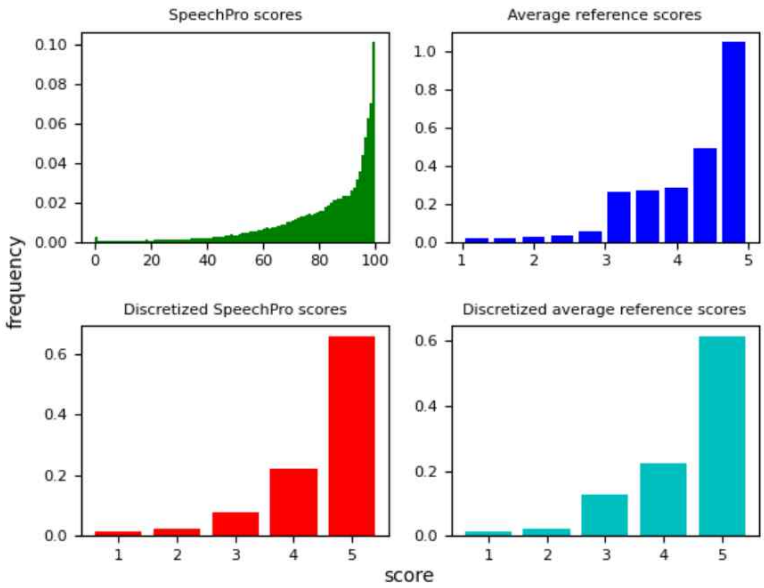

Figure 4. Histogram of SpeechPro scores, average reference scores, the discretized SpeechPro scores and the discretized average reference scores

Table 5 shows human-machine agreement measured by PCC, SMD, QWK, EPA and APA. For measuring PCC and SMD, the SpeechPro's original scores and the averaged reference scores are used. The p-value of the correlation coefficient 0.50 is lower than 0.001, showing statistical significance. Considering the average PCC and the average open-PCC of inter-rater agreement is 0.43 and 0.50, respectively, the value 0.50 is within the range of that of reference scores. The |SMD| of human-machine agreement, 0.35, is higher than the averaged |SMD| of human-human agreement. But the maximum value of human-human |SMD| is 0.48, which can be interpreted that this metric also tells the human-machine scores are showing agreement.

The other 3 metrics are applied to the scores converted to 5 discrete scores. The QWK of human-machine agreement is even higher than the averaged QWK of human-human agreement. Similarly, EPA and APA of human-machine agreement are greater than those of human-human agreement.

www.kci.go.kr

Table 5. Human-Machine Agreement Versus Human-Human Agreement

	PCC	|SMD|	QWK	EPA	APA
Human-Machine	0.50***	0.35	0.46	59.37%	94.05%
Human-Human (averaged)	0.43(0.50)	0.30	0.40	49.88%	90.00%

V. Discussion and Conclusion

This study aimed to validate the automatic pronunciation score reliability of the patented commercial product, SpeechPro. The score annotated English speech database of the native and non-native speakers was used. The human-human score agreement and human-machine score agreement were measured in turn. The purpose of measuring human-human score agreement was to verify the quality of the reference scores, and that of measuring human-machine score agreement was to test the machine scores' reliability. The five metrics conventionally used to measure score agreement were used. The results showed that SpeechPro's scores were significantly correlated and showing agreements with the averaged reference scores. This indicates SpeechPro's performance is verified.

The results of the statistical analysis show that human-machine agreement is higher than the human-human agreement, which could provoke doubts on the raters' qualification. However, the APE corpus authors clearly mention in its document that they recruited well-trained native English speaking ESL teachers from the same region in the USA. The rubric for pronunciation scoring was also provided to the raters. Table 3 also shows inter-rater agreement is statistically significant. But the human-machine agreement being higher than the human-human agreement still needs explanation. According to Yun (2009), due to human raters' individual differences in perceptive skills, inconsistency in inter-rater agreement and intra-rater agreement is inevitable. So, the human-human agreement from the result of this study can be understood as reflecting the inevitable inconsistency. Meanwhile, the human-machine agreement was measured between the averaged human score and the machine score, where intra-machine agreement is guaranteed to be consistent and inter-/intra-rater agreement inconsistency is diluted by averaging. This could explain the human-machine agreement being higher than the human-human agreement in this study.

However, one limitation of this study is that the speakers of APE did not read the same prompt sentences. It means the vocabulary level was not controlled. According to (Loukina et al., 2017), the entropy of a word can be a significant factor that affects pronunciation. With the same prompts read by every speaker, intra-rater and inter-rater agreement would be more consistent.

Also, it is important to further verify the system's performance by comparing SpeechPro with other automatic pronunciation assessment systems. Then to expand the system's features, it is requested to develop a module for scoring suprasegmentals such as stress, rhythm and intonation, since both the segmentals and suprasegmentals constitute the holistic quality of pronunciation (Prafianto et al., 2019; Kim, 2020). The score reliability of the suprasegmentals would be evaluated the same way as it is done in this

www.kci.go.kr

study with the sentence scores in APE.

SpeechPro will be useful for the purpose of English education for both teachers and learners. Automatized pronunciation assessment lessens instructors' burdens to listen to the audio and score them. Learners can receive immediate feedback on their speech with percentage score for each granular language unit, which is expected to enhance the learners' self-assessment and self-correction skills. To examine the effectiveness of SpeechPro on pronunciation enhancement, the longitudinal user study should be done with the specific pronunciations known to be hard for Korean college students using SpeechPro (Kim & Oh, 2019).

References

Beatty, K. (2003). *Teaching and researching computer-assisted language learning.* London: Person Education.

Carnegie Mellon University. (2015, Jul 15). The CMU pronouncing dictionary. Retrieved January 23, 2021, from the World Wide Web: http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Chen, H., & He, B. (2013, October). Automated essay scoring by maximizing human-machine agreement. *In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing,* 1741-1752.

Chung, H., Jang, T., Yun, W., Yun, I., & Sa, J. (2008). A study on automatic measurement of pronunciation accuracy of English speech produced by Korean learners of English. *Language and Linguistics, 42*, 165-196.

Cincarek, T., Gruhn, R., Hacker, C., Nöth, E., & Nakamura, S. (2009). Automatic pronunciation scoring of words and sentences independent from the non-native's first language. *Computer Speech & Language, 23*(1), 65-88.

Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2011). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing, 20*(1), 30-42.

Hanna, L., Barr, D., Hou, H., & McGill, S. (2020). An investigation of Modern Foreign Language (MFL) teachers and their cognitions of Computer Assisted Language Learning (CALL) amid the COVID-19 health pandemic. *arXiv preprint arXiv:2010.13901.*

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2012). A standardized mean difference effect size for single case designs. *Research Synthesis Methods, 3*(3), 224-239.

Herron, D., Menzel, W., Atwell, E., Bisiani, R., Danelozzi, F., Morton, R., & Schmidt, J. (1999). Automatic localization and diagnosis of pronunciation errors for second-language learners of English. *Paper presented at the 6th European Conference on Speech Communication and Technology*, September 5-9, 1999, Budapest, Hungary.

www.kci.go.kr

Kim, I. S. (2006). Automatic speech recognition: Reliability and pedagogical implications for teaching pronunciation. *Educational Technology & Society, 9*(1), 322-334.

Kim, M. (2020). A study of rhythm improvements and relevant linguistic factors in the pronunciation of English learners. *Studies in Foreign Language Education, 34*(1), 237-261.

Kim, T. S., & Oh, Y. (2019). A PBL approach to English pronunciation education: A case study of first-year students in English education department. *Studies in Foreign Language Education, 33*(4), 325-354.

Loukina, A., Zechner, K., Yoon, S. Y., Zhang, M., Tao, J., Wang, X., Lee, C. M., & Mulholland, M. (2017). Performance of automated speech scoring on different low- to medium-entropy item types for low-proficiency English learners. *ETS Research Report Series, 2017*(1), 1-17.

Menzel, M., Herron, D., Morton, R., Bomaventura, P., & Howarth, P. (2001). Interactive pronunciation training. *ReCALL, 13*(1), 67-78.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. Paper presented at *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hanneman, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The Kaldi speech recognition toolkit. Paper presented at *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

Prafianto, H., Nose, T., Chiba, Y., & Ito, A. (2019). Improving human scoring of prosody using parametric speech synthesis. *Speech Communication, 111*, 14-21.

Speechocean, (2020, September, 14). King-ASR-704: Accented English Pronunciation Evaluation Corpus (Word level). Retrieved January 20, 2021, from the World Wide Web: http://en.speechocean.com/datacenter/details/1328.html

Witt, S., & Young, S. (1998). Computer-assisted pronunciation teaching based on automatic speech recognition. In S. Jager, J. Nerbonne, & A. van Essen (Eds.), *Language Teaching & Language Technology* (pp. 25-35). The Netherlands: Swets & Zeitlinger.

Wang, Z., Zechner, K., & Sun, Y. (2018). Monitoring the performance of human and automated scores for spoken responses. *Language Testing, 35*(1), 101-120.

Yun, W., (2009). An Analysis of the Korean Inter-rater Difference in Evaluating English Pronunciations of Korean Speakers. *Studies in Foreign Language Education, 23*(2), 85-103.

Tafazoli, D., Huertas Abril, C. A., & Gómez Parra, M. E. (2019). Technology-based review on Computer-Assisted Language Learning: A chronological perspective. *Pixel-Bit: Revista de Medios y Educación, 54*, 29-43.

미디어젠(주). (2020). 특허출원제40-2020-0157130. 서울:특허청.

www.kci.go.kr

<Korean Abstract>

홍연정, 남호성. (2021). 교육적 활용을 위한 자동 영어 발음 평가 시스템의 점수 신뢰도 평가. *외국어교육연구, 35*(1), 91-104.

이 연구는 교육적으로 활용되는 자동 영어 발음 평가 시스템 SpeechPro의 발음 점수 신뢰도를 평가하는 것을 목적으로 한다. SpeechPro는 특허 출원과 상업적 출시가 이루어졌으나 연구적으로 점수 신뢰도가 확인되지 않았기 때문에, 본 연구에서는 기존 출시 제품의 안정성을 검증하고자 한다. 평가 방법은 SpeechPro와 전문 채점자 간의 점수 상관성을 측정하는 것이다. 이를 위해 원어민과 비원어민의 영어 발화와 3명의 영어 교육 전문가가 채점한 단어별 발음 점수가 함께 태깅되어 있는 유료 데이터베이스를 사용하였다. 먼저 평가자간 점수 차이를 비교하여 상관성을 측정하였고, 다음으로 SpeechPro 점수와 평가자 3명의 평균 점수 간의 상관성을 측정하였다. 측정 방법으로는 피어슨 상관 계수 (Pearson Correlation Coefficient), 표준화된 평균차 (Standardized Mean Difference), 2차 가중 카파 (Quadratic Weighted Kappa), 완전 퍼센트 일치율 (Exact Percentage Agreement), 1점 인접 퍼센트 일치율 (1-point Adjacent Percentage Agreement)를 사용하였다. 평가자간 점수 일치도와 SpeechPro-평가자간 점수 일치도가 5개의 측정치 전부로부터 동일하다는 결과를 보임으로써, SpeechPro의 발음 평가 점수 신뢰성이 입증되었다. SpeechPro와 전문 채점자 간의 상관성이 확인됨으로써 타 자동 발음 평가 시스템과의 비교 연구를 진행할 수 있는 논리적 근거가 마련되었다.

Key words: Automatic pronunciation assessment, Inter-rater agreement, Machine score reliability, CAPT / 자동 발음 평가, 평가자간 일치, 기계 점수 신뢰도, CAPT

Examples in: English
Applicable Languages: English
Applicable Levels: University

Hong, Yeonjung
PhD. candidate
Department of English Language and Literature, Korea University
145 Anam-ro, Seongbuk-gu, Seoul, Republic of Korea 02841
TEL: (02) 3290-1991
E-MAIL: yvonne_yj_hong@korea.ac.kr

Nam, Hosung
Professor
Department of English Language and Literature, Korea University
145 Anam-ro, Seongbuk-gu, Seoul, Republic of Korea 02841
TEL: (02) 3290-1991
E-MAIL: hnam@korea.ac.kr

received in January 26, 2021
revised version received in February 07, 2021
revised version accepted in February 09, 2021

www.kci.go.kr