Automatic Pronunciation Assessment of Korean Spoken by L2 Learners Using Best Feature Set Selection

Hyuksu Ryu[1], Hyejin Hong[2], Sunhee Kim[3], and Minhwa Chung[1]
[1] Department of Linguistics, Seoul National University, Seoul, REPUBLIC OF KOREA
E-mail: {oster01, mchung}@snu.ac.kr
[2] National Institute of Korean Language, Seoul, REPUBLIC OF KOREA
E-mail: honghyejin@korea.kr
[3] NAVER Corporation, Seongnam, REPUBLIC OF KOREA
E-mail: kim.sunhee@navercorp.com

Abstract—**This paper proposes a method for automatic pronunciation assessment of Korean spoken by L2 learners by selecting the best feature set from a collection of the most well-known features in the literature. The L2 Korean Speech Corpus is used for assessment modeling, where the native languages of the L2 learners are English, Chinese, Japanese, Russian, and Mongolian. In our system, learners' speech is forced-aligned and recognized using a native Korean acoustic model. Based on these results, various features for pronunciation assessment are computed, and divided into four categories such as RATE, SEGMENT, SILENCE, and GOP. Pronunciation scores produced by combining categories of features by multiple linear regression are used as a baseline. In order to enhance the baseline performance, relevant features are selected by using Principal Component Regression (PCR) and Best Subset Selection (BSS), respectively. The results show that the BSS model outperforms the baseline and the PCR model, and that features corresponding to speech segment and rate are selected as the relevant ones for automatic pronunciation assessment. The observed tendency of salient features will be useful for further improvement of automatic pronunciation assessment model for Korean language learners.**

I. Introduction

Computer-Assisted Language Learning (CALL) and Computer-Assisted Pronunciation Training (CAPT) systems have been developed in line with advances in speech technology [1]. In order to conduct effective CALL/CAPT, automatic pronunciation assessment of learners is required, where pronunciation scores can be rated by calculating the global scores over sentences or words of individual learners [1]. The global scores are based on various acoustic characteristics, such as rate of speech, or duration of speech segments and silences [2][3][4].

For automatic pronunciation assessment of English learners, the SpeechRater™ [5] by Education Testing Service (ETS) or the Versant system by Pearson [6] have been developed. The SpeechRater™ is the automated scoring system for speaking included in the TOEFL® examination. The system is used for pronunciation assessment in TOEFL® Practice Online (TPO). The Versant™ system provides speaking assessment solutions for English, Arabic, Dutch, French, and Spanish.

Although demand for Korean language education is steadily growing, studies regarding pronunciation assessment of Korean as a foreign language are limited to contrastive analysis between Korean and the native language (L1) of learners [7][8]. Furthermore, in the viewpoint of CALL/CAPT for the Korean language, except for some quantitative analyses [9][10][11][12], there has been a lack of studies regarding automatic pronunciation assessment in a Korean language context. As an early stage of developing a CAPT system for Korean language learners, this paper proposes a method for automatic pronunciation assessment of Korean spoken by L2 learners by selecting the best feature set from a collection of most well-known features in the literature.

The remaining part of this paper is organized as follows. Section II describes the related works about automatic pronunciation assessment. Methods for the experiments such as a pronunciation assessment model framework, and the corpora and statistical techniques used are explained in Section III. Section IV shows the results and discussion of the experiment, which is followed by the conclusion in Section V.

II. Related Works

Various features are used for automatic pronunciation assessment in many studies [2][3][4]. These features are derived from the segmentation of speech into phones. The rate of speech (ROS) is defined as the ratio between the number of speech phones and total duration [2]:

$$ROS = \frac{N_{phones}}{T_{total}} \tag{1}$$

, where T_{total} and N_{phones} mean total duration and number of phones, respectively. [2] reported that ROS had a correlation of 0.81 with the manual pronunciation scoring, and it is better than total duration.

Articulation rate (AR) is defined as the ratio between the number of phones and duration of speech without internal pauses [3]:

$$AR = \frac{N_{phones}}{T_{NoPause}} \tag{2}$$

, where $T_{NoPause}$ denotes duration of speech without internal pauses. [3][4] demonstrated that AR had a correlation of 0.83 with manual ratings for read speech, while it had weak correlation for spontaneous one.

Phonation time ratio (PTR) is defined as the ratio between the duration of speech without internal pauses and total duration [3]:

$$PTR = \frac{T_{NoPause}}{T_{total}} \tag{3}$$

This feature also had a strong correlation with manual ratings for read speech, according to [3].

Several studies investigated the correlation between the log posterior HMM-likelihood score and the manual score. A Goodness of Pronunciation (GOP) score to detect individual pronunciation error was proposed in [13], however, GOP was also used for pronunciation assessment by averaging it over an entire sentence [14]. The GOP is calculated as follows:

$$\begin{aligned} GOP &\equiv \frac{|\log p(q_i|O_i)|}{N_i} \\ &\approx \left| \frac{\log(\frac{p(O_i|q_i)}{\max_{j=1}^{J} p(O_i|q_i)})}{N_i} \right| \\ &= \left| \frac{\log(p(O_i|q_i))}{N_i} - \frac{\log(\max_{j=1}^{J} p(O_i|q_i))}{N_i} \right| \\ &= \left| p_{q_i(forced)} - p_{q_i(recognition)} \right| \end{aligned} \tag{4}$$

, where N_i and $p(O_i|q_i)$ mean the number of frames composing acoustic segment O_i and the probability of observing O_i given the phone q_i, respectively. By (4), GOP is defined as the log posterior likelihood $p(q_i|O_i)$ and calculated by the difference between the log likelihood from the forced-alignment and recognition. [14] presented that GOP over an entire sentence had high correlation with human ratings.

ETS presented various features for automatic pronunciation scoring of non-native spontaneous speech in tests of spoken English, such as TOEFL® [5]. They reported 29 candidate features for automatic pronunciation scoring. Most of them are related to fluency, such as the number or duration of words or silences. They also considered disfluency, repetition and language model score, since these were focused on scoring of spontaneous speech. By applying multiple linear regression and classification and regression tree (CART), they achieved 0.57 of correlation between the machine and human scores.

Lastly, Pearson [15] developed features for pronunciation assessment of English learners, for example, content, duration, and spectral features. The content feature is based on the latent semantic analysis (LSA) considering linguistic content spoken by learners. The durational feature is calculated by comparison of duration of phones produced by learners with native speech statistics, and spectral features - by the difference in log likelihood between native and non-native ASR. [15] combined the features using neural network regression and reported 0.826 of correlation with the human score.

As described above, the existing studies dealt with several features for automatic pronunciation assessment of learners. The features presented by previous studies are categorized in terms of their own characteristics. For example, ROS and AR are features regarding the 'rate' of speakers' speech. Furthermore, many features proposed by ETS [5] are related to the frequency or duration of 'silences' and 'speech segments' including mean, standard deviation, and mean absolute deviation. Although the previous studies have revealed that the suggested features are helpful for the performance of pronunciation assessment, the features have different influence on the performance of automatic pronunciation assessment in terms of their own category. Therefore, it is important to investigate relevance of feature categories on the pronunciation assessment. However, considering relevance of feature categories on the performance of automatic pronunciation scoring has not been studied significantly. In this study, we perform comprehensive analysis using a collection of the most well-known features according to the corresponding categories and select the most salient features for improvement of assessment performance.

III. Method

A. Assessment Modeling Framework

In order to perform automatic pronunciation assessment for Korean language learners, we organize a pronunciation assessment modeling framework as shown in Fig. 1.

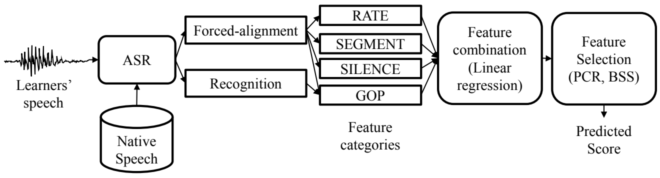

Fig. 1. Assessment modeling framework

As illustrated in Fig. 1, learners' speech is forced-aligned and recognized through an automatic speech recognizer (ASR) trained on Korean native speech. Using the results of forced-alignment and recognition, we calculate features for pronunciation assessment.

Among various features as described in Section II, we select and categorize features into GOP, RATE, SEGMENT, and SILENCE, as shown in Table I. RATE deals with the features related to the pace that learners speak. In the category of RATE, ROS and AR use the number of syllables, while Wpsec and Wpsecutt use the number of words. The category of SEGMENT includes the features about duration or frequency of segments including mean and standard/absolute deviation, while SILENCE treats features regarding silences. Since durational features and spectral features presented by [15] require nonnative ASR and statistics of native speech, these features will be implemented in further work. Furthermore, among the features proposed by [5], we eliminate ten features regarding spontaneous speech such as LM score, disfluency, repetition, and the number of unique words, since this study focuses on the read speech. Details of features used in this study are shown in Table I.

TABLE I
DESCRIPTION OF FEATURES USED IN THIS STUDY

Category	Feature	Description
GOP	GOP	Goodness of pronunciation
RATE	ROS	Rate of speech
	AR	Articulation rate
	PTR	Phone-time ratio
	Wpsec	Speech articulation rate
	Wpsecutt	Speaking rate
SEGMENT	Globsegdur	Duration of entire transcribed segment, including all pause
	Segdur	Total duration of segment w/o pauses
	Wdpchk	Average length of speech chunks
	Secpchk	Average duration of speech chunks
	Secpchkmeandev	Mean absolute deviation of speech chunks in seconds
	Wdpchkmeandev	Mean absolute deviation of speech chunks in words
SILENCE	Numsil	Number of silence events
	Silpwd	Duration of silences normalized by response lenght in words
	Silpsec	Duration of silences normalized by total words duration
	Silmean	Average duration of silences
	Silmeandev	Mean deviation of silences
	Longpfreq	Frequency of long pauses (0.2s)
	Longpmn	Mean duration of long pauses
	Longpwd	Frequency of long pauses normalized by response length in words
	Longpmeandev	Mean deviation of long pauses
	Silstddev	Standard deviation of silence duration
	Longpstdev	Standard deviation of long pauses

Lastly, in the feature combination step of Fig. 1, we compute pronunciation score from the features using multiple linear regression. However, features could have redundancy for each other, since they are likely to be inter-correlated. Thus, in order to extract relevant features and minimize redundancy, we apply several statistical techniques in the step of feature extraction. In this study, we use the Principal Component Regression (PCR) and Best Subset Selection (BSS), which are two frequently used and reliable statistical methods [16].

B. Corpora and ASR Settings

For pronunciation assessment modeling, we use L2 Korean Speech Corpus (L2KSC), which is built for studying Korean as a foreign language [17]. The corpus is composed of 229 learners with various native languages. In addition, the speech of 46 Korean native speakers is also included. The learners' proficiency in Korean is distributed from beginner to advanced levels. The speech data selected for the assessment modeling consists of 990 sentences in total uttered by 35 Korean native speakers and 130 learners whose L1 is English, Chinese, Japanese, Russian, and Mongolian. Each speaker produces six sentences from the traditional story "The North Wind and the Sun". The training set consists of 840 sentences spoken by 140 speakers whose L1 is English, Chinese, Japanese, and Korean, with 35 speakers in each group. The test set is composed of 150 sentences produced by 25 speakers whose L1 is English, Chinese, Japanese, Mongolian, and Russian, with 5 speakers in each group.

Furthermore, a Korean native read speech data which is composed of a phonetically balanced set of 51,500 sentences produced by 526 Korean native speakers is also used for ASR training. Context-independent monophone acoustic models are trained on the speech data by using HTK version 3.4 [18]. The number of Gaussian mixtures is increased up to 16. In addition, a pronunciation dictionary containing Korean canonical pronunciation and a phone-level language model are used for the forced-alignment and recognition. The utterances for assessment modeling are recognized and forced-aligned using the acoustic model. The phone error rate for the recognition is 26.44%.

C. Manual Rating

For the sentences uttered by Korean language learners, we perform manual ratings. Two Korean language education specialists whose L1 is Korean and two Korean naive native listeners participated in the rating. They were asked to rate the pronunciation of learners in a 9-point Likert scale. They rated 780 sentences of 130 Korean language learners. The distribution of manual ratings is illustrated in Fig. 2. As can be seen, the histogram of manual ratings shows a Gaussian distribution.

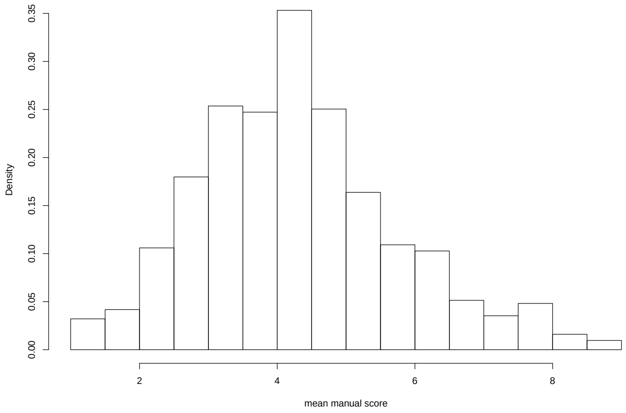

Fig. 2. The distribution of manual ratings

Intra-class correlation coefficient (ICC) [19] was calculated to assess reliability among raters. The ICC is calculated within the framework of analysis of variance (ANOVA) as follows.

$$ICC = \frac{\sigma_{\alpha}^2}{\sigma_{\alpha}^2 + \sigma_{\epsilon}^2} \tag{5}$$

, where σ_{α}^2 and σ_{ϵ}^2 denote variance of raters and residuals, respectively. According to (5), the ICC measures the proportion of total variance due to differences among raters. It is a widely used measure of inter-rater reliability for quantitative ratings [20]. Instead of the Pearson correlation, we use the ICC as inter-rater reliability, since we have four raters. The Pearson correlation is valid in a case of two raters. The calculation of the ICC values is performed using R package 'ICC' [21]. The ICC for the manual rating is 0.694, which belongs to the range of high reliability (0.61-0.80), according to [20].

D. Statistical Methods

We applied several statistical methods for prediction of pronunciation scores of learners as described in Fig. 1. First of all, multiple linear regression is carried out for score combination among the features. In our model, the features described in Table I and an average manual score of four raters are used as predictor and response variables, respectively. Multiple linear regression was carried out using R 3.2.4 [22].

Principal Component Regression (PCR) and Best subset selection (BSS) were applied to improve the performance of the assessment by reducing the dimensions, and to extract the best features set. The PCR model extracts linearly uncorrelated (orthogonal) predictors called 'Principal Components (PC)' from highly inter-correlated features. There are two criteria to select the optimal number of components. One is to determine the number based on the proportion of variance of dependent variable explained by the components and the other one is the root mean square error (RMSE). The optimal number of the components is usually determined at the point of a considerable change of slope, which is called an elbow, in the scree plots [23]. The PCR model is implemented using R packages 'pls' [24].

The Best Subset Selection (BSS) approach fits a separate least squares regression for each possible combination of the features and performs an exhaustive search for the best subsets of the features which minimize Bayesian information criterion (BIC). The BSS can be performed by R package 'leaps' [25]. In order to find the optimal parameters for PCR and BSS, we use 10-fold cross-validation on the training set.

E. Assessment Modeling

As mentioned in Section III-A, we use four categories of features, such as GOP, RATE, SEGMENT, and SILENCE, for assessment modeling. Multiple linear regression, PCR, and BSS are applied to the assessment models for score prediction. The details of assessment modeling are as follows.

In the setting of assessment modeling in Table II, multiple linear regression is applied to the model (1) to (8-1). Among the models, we compare the performance of automatic pronunciation assessment in term of combination of categories to investigate relevance of the feature categories. The PCR and BSS are applied to the model with the highest performance to extract the best feature set as the model (9) and (10), respectively, as shown in Table IV. We use the Pearson correlation between predicted score and the average manual rating score among raters as a measure of the models' performance.

TABLE II
DETAILS OF ASSESSMENT MODELING

No.	Model	Details
1	GOP	GOP features
2	SIL	SILENCE features
2-1	SIL + GOP	(2) + GOP
3	SEG	SEGMENT features
3-1	SEG + GOP	(3) + GOP
4	R	RATE features
4-1	R + GOP	(4) + GOP
5	SIL + SEG	SILENCE + SEGMENT
5-1	SIL + SEG + GOP	(5) + GOP
6	SIL + R	SILENCE + RATE
6-1	SIL + R + GOP	(6) + GOP
7	SEG + R	SEGMENT + RATE
7-1	SEG + R + GOP	(7) + GOP
8	SIL + SEG + R	SILENCE + SEGMENT + RATE
8-1	FULL	(8) + GOP

IV. Results and Discussion

A. Comparing Feature Categories

As mentioned in Section III-E, we compared the performance of assessment models in terms of combination of feature categories, as shown in Table III.

TABLE III
PERFORMANCE BY COMPARISON OF FEATURE CATEGORIES

No.	Model	Corr. (w/o GOP)	Corr. (w/ GOP)
1	GOP		0.569
2	SIL	0.833	0.842
3	SEG	0.842	0.847
4	R	0.867	0.867
5	SIL + SEG	0.859	0.860
6	SIL + R	0.883	0.884
7	SEG + R	0.885	0.888
8	SIL + SEG + R	0.888	**0.889**

Table III has two columns of correlation, which are correlation without GOP on the left and including GOP on the right. The column including GOP denotes the models with the '-1' in Table II. In Table III, RATE features have better performance than SILENCE, SEGMENT, and GOP features, when comparing the models from (1) to (4). It denotes that features such as rate of speech and articulation rate have stronger effect on the performance than frequency or duration of speech segments or silences, and GOP. It is also observed that combinations of feature categories induce similar tendency of the performance, when comparing the model from (5) to (7). Comparing models (5) and (6), appending RATE features to SILENCE has higher correlation than appending SEGMENT features. It is also observed that features regarding duration or frequency of segments are more influential than those of silences. This relationship between SEGMENT and SILENCE is shown by comparing the models (6) and (7), since adding SEGMENT is better than adding SILENCE. Therefore, the categories of RATE and SEGMENT are more relevant for pronunciation scoring of Korean language learners, than SILENCE and GOP.

Stacking categories shows better correlation than single category. Appending GOP features also helps to improve the performance, although the model using GOP only shows the lowest correlation of 0.569. The effect of GOP, however, is slightly small, compared with other categories. In addition, the full model using all categories presents the highest correlation (0.889, written in boldface).

B. Selecting the Relevant Features

In order to extract the best feature subset, PCR and BSS are applied to the full model, which shows the highest correlation. The results are described in Table IV.

TABLE IV
PERFORMANCE OF MODELS WITH FEATURE SELECTION

No.	Model	Correlation
8-1	Full	0.889
9	(8-1) + PCR	0.890
10	(8-1) + BSS	0.895

In the PCR of the model (9), the optimal number of the component is determined as nine by finding the point of a considerable change of slope in the scree plots of RMSE and variance explained by the number of components. The model (9) applying PCR leads to a slight improvement in performance. Even if the absolute difference between the full model and PCR is quite slight, the result presents that using only nine principal components shows similar or better performance comparing with the model using all features.

The BSS shows that the model with eight features has the minimum BIC. The model (10) applying the BSS presents the highest performance among all models in Table III and Table IV. In addition, the selected model using BSS is significant ($F(8, 831) = 670.1$, $p < 2.2\text{e}{-16}$). By using BSS, approximately 5.4% of relative improvement is achieved, comparing to the full model.

The details of selected features and the corresponding weights are presented in Table V. The table is sorted by the absolute value of weights. As can be seen in Table V, all features except GOP and Silpsec correspond to the category of RATE and SEGMENT. Furthermore, features regarding silences are almost excluded in the selected features. The result observed from Table V is compatible with the relevance of categories discussed in Section IV-A. Thus, the categories of RATE and SEGMENT are considered as more relevant than SILENCE and GOP for prediction of learners' pronunciation score.

V. CONCLUSION

This paper proposes a method for automatic pronunciation assessment of Korean spoken by L2 learners by selecting the best feature set from a collection of the most well-known features in the literature. We categorized the existing representative features for automatic pronunciation assessment according to their descriptions and extracted relevant features by applying statistical techniques such as PCR and BSS to improve the performance. Firstly, the feature categories show better performance in the order of RATE, SEGMENT, SILENCE, and GOP. Secondly, the results show that the assessment model with BSS had the highest performance. By the result of selected features from the BSS model, most of the salient features correspond to speech rate and segments. The observed tendency of salient features will be useful for further improvement of automatic pronunciation assessment model for Korean language learners.

TABLE V
SELECTED FEATURES AND THE CORRESPONDING WEIGHTS

No.	Feature	Category	Weight
1	PTR	RATE	-5.550
2	Silpsec	SILENCE	-2.154
3	Secpchk	SEGMENT	1.578
4	ROS	RATE	1.342
5	Wpsec	RATE	1.192
6	Wdpchk	SEGMENT	-0.783
7	Secpchkmeandev	SEGMENT	0.411
8	GOP	GOP	-0.001

In this work, however, we were limited to using the existing features for pronunciation assessment. In spite of this limitation, it is noteworthy that it is one of the first studies to validate the existing features and the approach on Korean learners. The result of this work can further serve as a baseline for developing a CAPT system. In addition, the existing features presented in this study are mostly related to how fast or how long the learners speak. It is necessary to develop features regarding other aspect, such as pronunciation quality. Thus, in a future study, assessment modeling using novel features will be investigated based on the results of this paper.

ACKNOWLEDGMENT

This work was partly supported by the project, "Developing a pronunciation assessment system for foreign learners of Korean," funded by NAVER Corporation in Seongnam, Korea. This research was partly supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Republic of Korea (No. 2015R1D1A2A01061378).

REFERENCES

[1] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832–844, 2009.

[2] C. Cucchiarini, H. Strik, and L. Boves, "Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms," *Speech Communication*, vol. 30, no. 2-3, pp. 109–119, 2000.

[3] C. Cucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learnersâĂŹ fluency by means of automatic speech recognition technology," *The Journal of the Acoustical Society of America*, vol. 107, no. 2, pp. 989–999, 2000.

[4] C. Cucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech," *Journal of the Acoustical Society of America*, vol. 111, no. 6, pp. 2862–2873, 2002.

[5] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken English," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.

[6] R. Downey, H. Farhady, R. Present-Thomas, M. Suzuki, and A. van Moere, “Evaluation of the usefulness of the versant for English test: A response,” *Language Assessment Quarterly*, vol. 5, no. 2, pp. 160–167, 2008.

[7] J. Cui, “The education of Korean pronunciation for students speaking Chinese,” *Bilingual Research*, vol. 20, pp. 309–343, 2002.

[8] J.-Y. Han, J.-S. Choi, H.-Y. Lee, J.-Y. Park, K.-M. Lee, H.-Y. Cho, and J. Cui, *Teaching Korean Pronunciation*. Seoul, Korea: Hollym, 2003.

[9] H. Hong, S. Kim, and M. Chung, “A corpus-based analysis of Korean segments produced by Japanese learners,” in *SLaTE (Workshop on Speech and Language Technology in Education) 2013*, pp. 189–192, 2013.

[10] S. H. Yang, M. Na, and M. Chung, “Modeling pronunciation variations for non-native speech recognition of Korean produced by Chinese learners,” in *SLaTE (Workshop on Speech and Language Technology in Education) 2015*, pp. 95–99, 2015.

[11] S. H. Yang and M. Chung, “Automatic classification of retroflex segmental variations for Korean produced by Chinese,” in *ICSS (International Conference on Speech Science) 2015*, pp. 105–106, 2015.

[12] S. H. Yang, H. Ryu, and M. Chung, “A corpus-based analysis of Korean segments produced by Chinese learners,” in *APSIPA ASC (Asia-Pacific Signal and Information Processing Association Annual Summit and Conference) 2015*, pp. 583–586, 2015.

[13] S. M. Witt and S. J. Young, “Phone-level pronunciation scoring and assessment for interactive language learning,” *Speech Communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[14] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, “Automatic scoring of pronunciation quality,” *Speech Communication*, vol. 30, no. 2-3, pp. 83–93, 2000.

[15] A. Metallinou and J. Cheng, “Using deep neural networks to improve proficiency assessment for children English language learners,” in *Interspeech 2014*, pp. 1468–1472, 2014.

[16] M. Febrero-Bande, P. Galeano, and W. González-Manteiga, “Functional principal component regression and functional partial least-squares regression: An overview and a comparative study,” *International Statistical Review*, vol. 83, no. 1, pp. 1–23, 2015.

[17] S.-c. Rhee, J.-a. Kim, and J.-w. Chang, “Design and construction of speech corpus for Korean as a foreign language (L2KSC),” *The Journal of Chinese Language and Literature*, vol. 33, pp. 35–53, 2005.

[18] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, and D. Povey, *The HTK book (for HTK version 3.4)*. 2006.

[19] P. E. Shrout and J. L. Fleiss, “Intraclass correlations: uses in assessing rater reliability,” *Psychological bulletin*, vol. 86, no. 2, pp. 420–428, 1979.

[20] J. L. Fleiss, B. Levin, and M. C. Paik, *Statistical methods for rates and proportions*. Wiley series in probability and statistics, Hoboken, N.J.: J. Wiley, 3rd ed., 2003.

[21] M. E. Wolak, D. J. Fairbairn, and Y. R. Paulsen, “Guidelines for estimating repeatability,” *Methods in Ecology and Evolution*, vol. 3, no. 1, pp. 129–137, 2012.

[22] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[23] R. Wehrens, *Chemometrics with R: Multivariate data analysis in the natural sciences and life sciences*. Use R!, Heidelberg; New York: Springer,, 2011.

[24] B. H. Mevik and R. Wehrens, “The pls package: Principal component and partial least squares regression in R,” *Journal of Statistical Software*, vol. 18, no. 2, pp. 1–23, 2007.

[25] T. Lumley and A. Miller, *leaps: regression subset selection*, 2009. R package version 2.9.