Deep Learning-based Automatic Pronunciation Assessment for Second Language Learners

Kohichi Takai[1,2], Panikos Heracleous[2], Keiji Yasuda[1,2], and Akio Yoneyama[2]

[1] Nara Institute of Science and Technology, Japan
takai.koichi.tc1@is.naist.jp,ke-yasuda@dsc.naist.jp
[2] KDDI Research, Inc., Japan
{pa-heracleous,yoneyama}@kddi-research.jp

Abstract. For second language learners, computer-aided language learning (CALL) is of high importance. In recent years, the use of smart phones, tablets, and laptops has become increasingly popular; with this change, more people can use CALL to learn a second language. In CALL, automatic pronunciation assessment can be applied to provide feedback to teachers regarding the efficiency of teaching approaches. Furthermore, with automatic pronunciation assessment, students can monitor their language skills and improvements over time while using the system. In the current study, a text-independent method for pronunciation assessment based on deep neural networks (DNNs) is proposed and evaluated. In the proposed method, only acoustic features are applied, and native acoustic models and teachers' reference speech are not required. The method was evaluated using speech from a large number of Japanese students who studied English as a second language.

Keywords: Automatic pronunciation assessment · Acoustic features · Deep neural networks.

1 Introduction

English is one of the most widely spoken languages in the world and many people learn English as a second language (ESL). In addition to conventional in-class English learning, the importance of computer-aided language learning (CALL) increases. ESL includes four components namely, listening, reading, speaking, and writing. The current research focuses on English pronunciation assessment, which plays an important role in the speaking component of second language learning.

Previously, several studies addressed the problem of automatic pronunciation assessment using different features and grading approaches [4, 7, 11]. However, the majority of the studies reported require accurate native acoustic models with a large amount of training data, reference teachers' speech, and are usually text-dependent (i.e., known text of the uttered speech). These automatic assessments are only useful for shadowing-based pronunciation learning. In contrast, text-independent automatic pronunciation assessment without teacher's reference speech can expand possibility of CALL applications.

2 Methods

2.1 Overview of data collection

In the proposed method, automatic pronunciation assessment without reference native speech and known text of the uttered speech is being considered. Because of the simplicity in collecting the speech data, in the current study a speech shadowing framework and materials were used.

To evaluate the effectiveness of the proposed method for automatic pronunciation assessment, speech data on various materials and a large number of speakers were collected. Following the data collection, human raters were employed to annotate the collected speech samples. In this section, the data collection procedure and the data annotation are described.

Speaking materials and collected speech data For speaking material, 3,388 sentences of shadowing samples extracted from daily conversations were used. The materials were classified into five subsets reflecting the English proficiency level. In the current study, the TOEIC listening and reading test score was used to define difficulties of materials.

The materials also included the native reference speech samples, and, therefore, the speakers could use the native speech as a reference before producing the desired speech sample. This made it easy for speakers to produce difficult sentences and reduced the need for dictionaries. The speakers who participated in the data collection included Japanese students (45.53%), native Japanese English teachers (11.24%), and native English teachers (43.23%).

In total, 924 speakers produced speech samples from a part of the shadowing materials. Details of the collected speech data are shown in Table 1.

Table 1. Details of collected speech samples.

TOEIC score	# of sentences in shadowing materials	# of collected speech samples
∼400	729	57,601
400∼ 500	777	67,090
500∼ 600	938	65,060
700∼ 800	521	34,391
800∼	423	23,214
Total	3,388	247,356

Annotation by manual pronunciation evaluation A part of the collected speech data which consists of 96,993 speech samples were evaluated by human raters using four criteria as shown in Table 2.

Each speech sample was evaluated by two different English native raters using a 5-rank scale for each criteria. The two scores were averaged to provide the final score. Tables 3 shows the annotation results of overall annotation criterion.

Table 2. Criteria of subjective evaluation.

Criteria	Check Point	Rank 5	Rank 4	Rank 3	Rank 2	Rank 1
Overall	Intelligibility	Near native	3∼5	Intermediate	1∼3	Beginner
Pronunciation	Stress	>90%	>70%	>50%	>50%<40%	<40%
Intonation	Intonation/Stress	Appropriate tone place and pitch	3∼5	Often Appropriate	1∼3	Not appropriate
Fluency	Rhythm/Linking/Speed	Natural	3∼5	Fairly natural	1∼3	Not natural

Table 3. Subjective evaluation results in overall criterion.

Rank in overall criterion	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Total
# of speech samples	3,433	6,698	11,165	11,737	63,960	96,993

2.2 Results

The data set for preliminary DNN experiments were created by using a subset of data shown in Table 3. For pronunciation assessment, a 3-level scale was used, namely, below average (rank1 and rank2), average (rank3), and above average (rank4 and rank5) by merging the corresponding ranks. In the experiments reported in the current study, 935 speech samples for each class were used for training the DNN [6]. Other 924 speech samples for each class were used for the DNN evaluation.

For the evaluation, the recalls of each class, the unweighted average recall (UAR) (i.e., mean of the class recalls), and the Pearson correlation coefficient metrics were used. Mel-frequency cepstral coefficients (MFCCs) [8] concatenated with shifted delta cepstral (SDC) coefficients [1, 9] were extracted from the speech signal every 10 ms with a time window of 20 ms. The MFCC and SDC features were used to construct the i-vectors [3] used for training and evaluation.

Gaussian mixture models (GMMs) supervectors are widely used in speaker recognition. The GMM supervectors are obtained by concatenating the means of an adapted GMM. The main disadvantage of supervectors is the high dimensionality, which imposes high computational and memory costs. To overcome these problems, the i-vectors were introduced, which represent the whole utterance by a small number of factors, explaining also the variability of speaker, language, emotion, and channel. In the current method, the i-vectors are used as features for pronunciation scoring. Following i-vector extraction, linear discriminant analysis (LDA) [5] was also applied to further improve the class discrimination ability.

The classification experiments were based on DNNs. A DNN is a feed-forward neural network with many (i.e., more than one) hidden layers. The main advantage of DNNs compared to shallow networks is the better feature expression and the ability to perform complex mapping. In the current study, four hidden layers with 64 units and ReLu activation function were used. On top, a fully-connected Softmax layer was added. The number of batches was set to 512, and 500 epochs were used.

Table 4 shows the results achieved. As shown, when using LDA, significant improvements were obtained. When using MFCC and SDC features with LDA, a 64.4% UAR and a 0.48 correlation were achieved. These results are comparable or even superior to other similar state-of-the-art approaches [2, 10].

Table 4. Individual recalls for the three classes.

Features (ivector extraction)	Below average	Average	Above average	UAR	Pearson CC
MFCC	56.24	23.98	25.18	35.13	0.0236
MFCC + SDC	42.45	31.53	35.01	36.33	0.0568
MFCC + LDA	50.96	62.95	60.67	58.19	0.3928
MFCC + SDC + LDA	63.14	57.30	72.75	64.40	0.4803

3 Conclusions

In the current paper, a method for automatic pronunciation assessment for second language learners was presented. The method is based on DNNs, and the results obtained were very promising. Data collection of a large number of speakers was also introduced. Evaluating the method using a larger amount of non-native speech data is currently in progress.

References

1. Bielefeld, B.: Language identification using shifted delta cepstrum. In Fourteenth Annual Speech Research Symposium (1994)
2. Chen, L.Y., Jang, J.S.R.: Automatic Pronunciation Scoring using Learning to Rank and DP-based Score Segmentation. in Proc. of Interspeech p. 761764 (2010)
3. Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P.: Front-End Factor Analysis for Speaker Verification. IEEE Transactions on Audio, Speech, and Language Processing **19(4)**, 788–798 (2011)
4. Franco, H., Neumeyer, L., Ramos, M., Bratt, H.: Exploring Deep Learning Architecures for Automatically Grading Non-native Spontaneous Speech. in Proc. of ICASSP pp. 6140–6144 (2016)
5. Fukunaga, K.: Introduction to Statistical Pattern Recognition, 2nd ed. New York: Academic Press, ch. 10 (1990)
6. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., Kingsbury, B.: Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups ," IEEE Signal Processing Magazine **29, Issue:6**, 82–97 (2012)

7. Nicolao, M., Beeston, A.V., Hain, T.: Automatic Assesement of English Learner Pronunciation Using Discriminative Classifiers. in Proc. of ICASSP pp. 5351–5355 (2015)
8. Sahidullah, M., Saha, G.: Design, Analysis and Experimental Evaluation of Block Based Transformation in MFCC Computation for Speaker Recognition. Speech Communication **54 (4)**, 543565 (2012). https://doi.org/doi:10.1016/j.specom.2011.11.004
9. T.-Carrasquillo, P., Singer, E., Kohler, M.A., Greene, R.J., Reynolds, D.A., Jr., J.D.: Approaches to language identification using gaussian mixture models and shifted delta cepstral features. in Proc. of ICSLP2002-INTERSPEECH2002 pp. 16–20 (2002)
10. Witt, S., Young, S.: Phone-level pronunciation scoring and assessment for interactive language learning S.M. Speech Communication **30**, 95108 (2000)
11. Yue, J., Shiozawa, F., Toyama, S., Yamauchi, Y., K. Ito, D.S., Minematsu, N.: Automatic Scoring of Shadowing Speech based on DNN Posteriors and their DTW. in Proc. of Interspeech pp. 1422–1426 (2017)