Deep Learning-based Automatic Pronunciation Assessment for Second Language Learners

Kohichi Takai[1,2], Panikos Heracleous[2], Keiji Yasuda[1,2], Akio Yoneyama[2]
[1]Nara Institute of Science and Technology, Japan
[2]KDDI Research, Inc., Japan

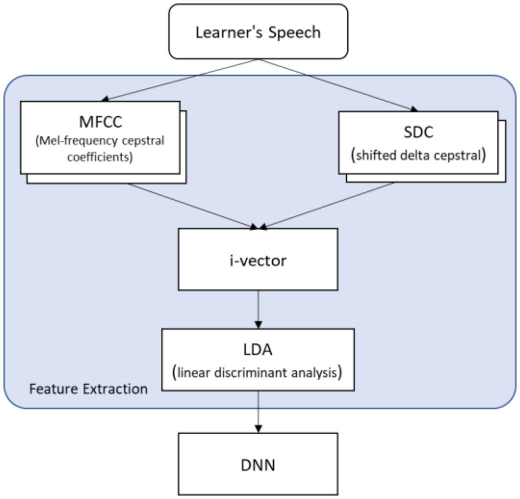
Introduction

- **Computer-aided language learning (CALL)** is of high importance **for English learning as a second language (ESL).**
- CALL is also useful for **shadowing-based pronunciation** and automatically providing pronunciation assessment.
- The proposed **text-independent method for pronunciation assessment** is based on deep neural networks (DNNs).
- The proposed method aims at providing CALL **without shadowing reference speech or acoustic models of native speakers.**

Method

The current study is based on DNNs and seeks to improve acoustic feature extraction. The following outlines the proposed method.

- Extract mel-frequency cepstral coefficients (MFCCs) and shifted delta cepstral (SDC) coefficients from speech samples every 10ms with a time window size of 20ms.
- Construct i-vectors from the whole utterance of MFCC and SDC features.
- Following i-vector extraction, apply linear discriminant analysis (LDA) to **reduce dimension size and improve evaluation performation**.
- The DNN has four hidden layers with 64 units and ReLu activation function.
- On the last layer, a fully-connected Softmax layer is added.



Learner's Speech

MFCC (Mel-frequency cepstral coefficients)

SDC (shifted delta cepstral)

i-vector

LDA (linear discriminant analysis)

Feature Extraction

DNN

Data Collections

924 speakers produced speech samples from a section of the shadowing materials. This resulted in 96,993 total speech samples.

Rank in overall criterion	Rank 1 (Beginner)	Rank 2	Rank 3 (Intermediate)	Rank 4	Rank 5 (Near native)
# of speech samples	3,433	6,698	11,165	11,737	63,960

Rank2: 1~3, Rank4: 3~5

Experiments

3-level re-scale

Below average (rank1,rank2), average (rank3), and above average (rank4, rank5)

Features (i-vector extraction)	dimension	Below average	Average	Above average	UAR	Pearson CC
MFCC	400	56.24	23.98	25.18	35.13	0.0236
MFCC+SDC	400	42.45	31.53	35.01	36.33	0.0568
MFCC+LDA	2	50.96	**62.95**	60.67	58.19	0.3928
MFCC+SDC+LDA	2	**63.14**	57.3	**72.75**	**64.4**	**0.4803**

Conclusion

- Unweighted average recall (UAR) was 64.4%, and the correlation was 0.48 **when using MFCC and SDC for i-vector extraction and LDA**,
- The improvement of audio feature extraction was useful for CALL.
- As future work, the current study will be compared with previous studies, and its effectiveness will be investigated.

Contact

KDDI Research, Inc., Japan
Kohichi Takai(ko-takai@kddi-research.jp)
Panikos Hracleous(pa-heracleous@kddi-research.jp)