Volume: 17 Issue: 01 Pages: 6719-6727 (2025) ISSN: 0975-0290

Automatic Pronunciation Evaluation and Feedback Generation System based on Resource-efficient factorized TDNN and Phoneme Error Pattern Model

Il Han
Institute of Information Technology, High-Tech Research and Development Center, **Kim Il Sung** University, Pyongyang, Democratic People's Republic of Korea
Email: il.han0104@ryongnamsan.edu.kp
Hyok Kwak
Institute of Information Technology, High-Tech Research and Development Center, **Kim Il Sung** University, Pyongyang, Democratic People's Republic of Korea
Email: abcdef@gmail.com
Chung-Hyok O
Institute of Information Technology, High-Tech Research and Development Center, **Kim Il Sung** University, Pyongyang, Democratic People's Republic of Korea
Email: abcdef@gmail.com
Kang-Song Choe
Institute of Information Technology, High-Tech Research and Development Center, **Kim Il Sung** University, Pyongyang, Democratic People's Republic of Korea
Email: abcdef@gmail.com
Chol-Nam Om
Institute of Information Technology, High-Tech Research and Development Center, **Kim Il Sung** University, Pyongyang, Democratic People's Republic of Korea
Email: cn.om1011@ryongnamsan.edu.kp

ABSTRACT

In this paper we describe an automatic pronunciation error detection and feedback generation system for non-native second language learners by using deep acoustic model based on factorized TDNN and language model with phoneme error model. Our system builds language model considering phoneme error patterns and gives a useful feedback for learners. Deep acoustic model consists of TDNN-F with grouped fully-connected layers and shuffle operation. This network architecture maintains recognition accuracy like traditional TDNN and costs less then it. Also, our system evaluates pronunciation proficiency of utterance in word level and phoneme level based on confidence from Minimum Bayesian Risk decoder, feedback is generated on phone error model of L2 learners. This system based on resource-efficient deep acoustic architecture can be deployed in resource-limited mobile devices.

Keywords - Computer-Aided Pronunciation Training, mispronunciation error detection, deep acoustic model, factorized time delay neural network, feedback generation

Date of Submission: May 29.2025 Date of Acceptance: June 19.2025

I. INTRODUCTION

Recently with the development of speech recognition, the CAPT (Computer-aided practice training) [1] system has been extensively studied to automatically evaluate the pronunciation by introducing this technique into the pronunciation education.

Due to development of modern information technology in past decades, CAPT has helped L2 learners much since it is an efficient tool for improving language skills in classroom or indoors instead of skillful teacher. [2][3] These CAPT systems help L2 learners to improve speaking skills. CALL evaluates the pronunciation of L2 learners and give feedback on mispronunciation in phone-level, word-level [4] or sentence-level.

GOP [8] (Goodness of Pronunciation) is a widely used index for mispronunciation detection. Extended neural networks including normal and mispronunciations have been explored to provide diagnostic information. [5] To further improve detection performance, classifier-based evaluation methods have been proposed. [6][7][9]

Pronunciation Errors are classified into two types, i.e., phonemic error and prosodic error. [10] We only focus on phonemic error.

Many works have been carried out on the evaluation of pronunciation with reference text.

In 2000, Witt and Young first introduced a Goodness of Pronunciation (GOP) method, a variation of posterior probability for phone level pronunciation scoring. [8] This method is widely used in mispronunciation detection task.

Volume: 17 Issue: 01 Pages: 6719-6727 (2025) ISSN: 0975-0290

In 2008, Zhang introduced the Scale Logarithmic Posterior Probability Method (SLPP) in Chinese mispronunciation detection to achieve considerable performance improvement. [11]
Wang [12] showed that GOP method combined with error pattern detector in phonetic mispronunciation detection with serial and parallel structure reduces the average error rate and improves the diagnosis feedback.
Yan and Gong [13] introduced an acoustic model refined by Minimum Phone Error (MPE) to pronunciation proficiency estimation. Qian [14] minimized the pronunciation detection errors for English learners with a HMM model trained Minimum Word Error (MWE).
The mispronunciation detection task can be formulated as a binary classification problem.
Many methods based on classifiers have been introduced for mispronunciation detection to improve performance. Using the decision tree-based approach, a threshold for different phones was set to achieve significant improvement compared to the conventional threshold. [15] Truong shows that Linear Discriminant Analysis (LDA) based approach outperforms GOP and decision tree-based approaches. [16] Strike [17] used Supper Vector Machine (SVM) combining various features to improve the vowel substitution error and performance. In [18], Doremalen et al built SVM to detect the substitution error of the Dutch vowel and improve performance to combine different features. In 2009, Wei et al. [19] applied SVM to detect the correct and incorrect pronunciation of Chinese syllables and to improve the discriminative ability of pronunciation variation with the pronunciation space model. In addition, SVM was also used as the baseline classifier for mispronunciation detection. [20] [21].
As the DNN-HMM based approach is superior to the traditional GMM-HMM based approach, the GOP score estimated from DNN output is closer to the human score than GMM-based system. [22]
Recently, deep neural networks (DNNs) has significantly improved the discrimination of acoustic models in speech recognition. [23] Qian et al [24] proposed a method for mispronunciation detection of English speakers with Deep Belief Networks (DBNs) and improved pronunciation estimation relative error rates of word pronunciation. For English pronunciation evaluation, a GOP score estimated from DNN correlates better with human score than that from GMM.
Hu et al [25] investigated different pitch embedding methods in a DNN acoustic model and detected mispronunciation caused by lexicon stress or lexicon tone to improve the discrimination of acoustic models. In a previous work [26], Hu et al proposed a method to improve the mispronunciation detection in English and Chinese using deep acoustic model and transfer learning based on logistic regression classifier. Li et al. [27] suggested a method for detecting phonetic accents and pitch accents by MD-DNN.
A pronunciation model [28] that uses explicit knowledge of pronunciation error patterns increases discrimination of whether a phoneme is pronounced correctly or not. It is shown that the modern method based on this pronunciation model, the proposed method, and shows similar classification performance to the expert.
In [29], Mispronunciation detection problem was regarded as an anomaly diagnosis problem.
A pre-trained approach using speech data from two languages (students' native and target languages) for the verification of non-native speakers, and a language-generated representation learning was proposed. [30] [32]
Moustroufas and Diagalakis compared the sentence-level likelihood scores from acoustic models of random speakers using acoustic models of native and non-native. [33]
The final goal of pronunciation evaluation is to give a reasonable feedback so that help learners to improve pronunciation as fast as possible. Also, an acoustic model in CAPT must be robust in environment noises and has low real time factor in mobile devices.
A pronunciation method by DNN in previous works only detects whether of phoneme correct or incorrect, thus has no way to give feedback for correction method of pronunciation. However, DNN is limited by its inability to capture wide temporal context compared to RNN. In real environments, there exists a certain amount of noise and echoes, and DNN cannot solve this problem.
Recently, Time-Delay Neural Network (TDNN) [40] [43] [44], which is widely used in speech recognition, is better than DNN in the performance and computational costs.
In this paper, we modified TDNN architecture to decrease computational costs for CAPT. To develop a dialogue-based CAPT system, we proposed pronunciation evaluation system and feedback generation system based on resource-efficient TDNN architecture.
The main contributions of this paper are as follows.
First, we propose a deep acoustic model consisting of factorized TDNN with grouped fully-connected (FC) layers and shuffle operations.
Second, we constructed language model that takes into account the pronunciation errors and online decoder system which detects mispronunciation and gives feedback generation.
The rest of paper is organized as follows. In Section 2, we introduce our novel method and the system configurations. Section 3 presents the discussion of the experimental results. Finally, we present the conclusions.

2. PROPOSED METHOD

In this section, we describe the proposed deep acoustic model and the system configuration.

2.1 TDNN-F with grouped layers and shuffle operations

In order to model the temporal dynamics in speech signals well, we must have model which can effectively deal with the long-term temporal contexts
In [25] [26] [29], DNN has limitation which does not capture long term dependency. Recurrent neural networks (RNNs) which use a dynamically changing context over all the sequence rather than a fixed context has shown to achieve state-of-art performance. [31] However due to the recurrent connections in RNN, parallelization during

Volume: 17 Issue: 01 Pages: 6719-6727 (2025) ISSN: 0975-0290

training and decoding cannot be implemented as in feed-forward neural networks.

We improved TDNN [43] (known as 1-D CNN) architecture which is robust to noise and reverberation. This neural network can increase the input context to 300ms. This capability enables the neural network to learn the representation of reverberation and noise in a wide context. Rapid adaptive technique is used using iVector, which captures speaker and environmental information, is very useful in neural network adaptation.

The common method to reduce the number of parameters of the pre-trained neural network is to use singular value decomposition (SVD). [41] Using SVD, we restricted one to semi-orthogonal at C = AB. TDNN-F [44] is a factored form of TDNN. As shown in Fig. 1, we decompose more than two FC layers, one being semi-orthogonal.

The first layer is a splicing layer which concatenates all the L1 of M-dimensional features. (1×M→L1×M) In the first FC, it acts as a bottleneck and reduces the input dimension from M to K. (L1×M →1×K, K<<M) After that, the L2 frames are concatenated again(1×K→L2×K). In the second FC layer projection is applied (L2×K→1×N) After the projection layer, we apply activation as ReLU and Batch Normalization (BN). A dropout (0.5) is also used. A statistically scaled (bypass=0.66) residual connection is applied between the input and output modules. The structure of the neural network is as follows. L1 = L2 = 2.

Our TDNN-F is a variant of TDNN-F which modifies FC layers. The goal is to reduce computational costs and maintain performance. As Fig 1 b) shows, the FC layer is replaced by grouped fully-connected (GFC) layer. In this case computational costs are decreased. For example, if two groups were used, the required FLOPs would be halved.

A shortcoming of GFC layers is the lack of feature exchange across groups. To overcome this limitation, we insert a shuffle layer [45], enabling information exchange across groups, and maintained performance.

As in Fig. 1.c), the grouped shuffle TDNN-F operation starts with a layer that partitions the feature vectors by the number of groups. Then, feature vectors in each group. (L1 frames of M/G dimension) are spliced. Next, we apply the bottleneck layer to each group and get k/G dimension. After the bottleneck layer, L2 frames of k/G-dimension features are spliced. The second FC layer (projection layer) is applied each group. After the projection layer, ReLU activation function and BN are employed. A dropout (0.5) and residual connection with a bypass factor of 0.66.

2.2 Minimum Bayesian Risk Decoder with confidence

Fig. 2 illustrates the Minimum Bayesian Risk [35] (MBR) decoding process, where R is a one-best path, L is a lattice. In MBR options, we set decode_mbr as true for MBR decoding or set false to perform MAP estimation. The WER is the lowest when MAP estimation is performed.

The confidence of each word can be conducted from MBR decoding. During decoding, confidence based on posterior probability are estimated. By means of MBR decoding [35], we can find the lattices that outputs the minimum word error rate, and obtain the confidence level of the maximum of each word in the lattices being 1, minimum is 0, and start time and end time.

R=MBR-decode(L)
1: R<-One-best path of L
2: loop
3: R<-normalization of $eps(R)$
4: $(\hat{L}, \gamma(\cdot,\cdot))$←Acc-Stats(L,R)
5: $\Delta Q \leftarrow 0$
6: for q=1 to $|R|$
7: $\hat{r} \leftarrow max_s\, \gamma\,(q,s)$
8: $\Delta Q \leftarrow \Delta Q + \gamma(q,r_q) - \gamma(q,\hat{r})$
9: $r_q \leftarrow \hat{r}$
10: if r_q!=0
11: $confidence \leftarrow 0$
12: for j=0 to $|\gamma_q|$
13: if $\gamma_{q,j}$=r_q
14: $confidence \leftarrow \gamma_{q,j}(s)$
15: end for
16: end for
17: if $\Delta Q = 0$ then
18: break
19: end if
20: end loop
21: $R \leftarrow remove - eps(R)$

Fig.2. Confidence from MBR decoding

The decoder gives a confidence level from 0 to 1 for each word. A confidence close to 1 means that the posterior probability is maximum, and a confidence close to 0 means that the likelihood is very low.

2.3 Phoneme error patterns

The pronunciation errors in the non-native corpus are classified as follows.

- Deletion: When pronouncing, the phoneme in the word was deleted.
- Insertion: The phoneme was inserted before or after another phoneme.
- Substitution: The phoneme was replaced by another phoneme similar to that.
- Distortion: The phoneme was replaced by acoustic like phone.

These errors can be classified to consonant error and vowel error. Since the learners are accustomed to mother language pronunciation, there are many instances of the incorrect pronunciation of short vowels in English, with long or short vowels in English, without paying attention to the quantitative aspects of the English vowels at the beginning of English studying. The length of the English vowel not only makes the difference distinct, but also makes the meaning of the word different.

Volume: 17 Issue: 01 Pages: 6719-6727 (2025) ISSN: 0975-0290

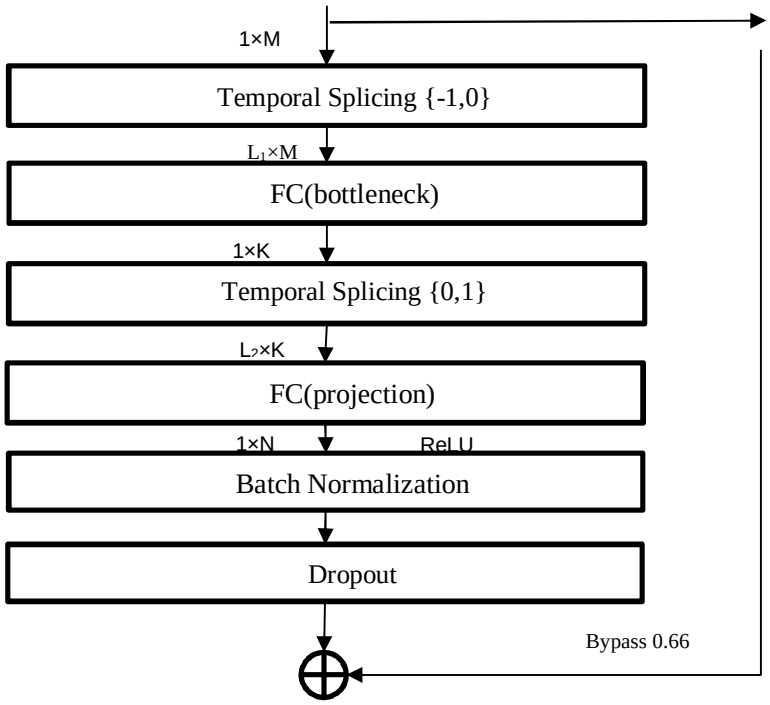

Fig. 1. a) TDNN-F modules

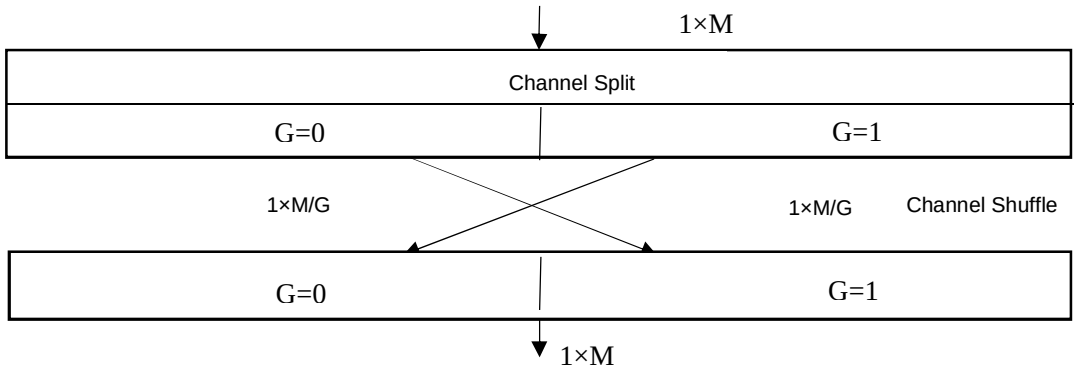

Fig.1.b) Detailed operation of the TDNN-F with shuffle operation and two groups

Volume: 17 Issue: 01 Pages: 6719-6727 (2025) ISSN: 0975-0290

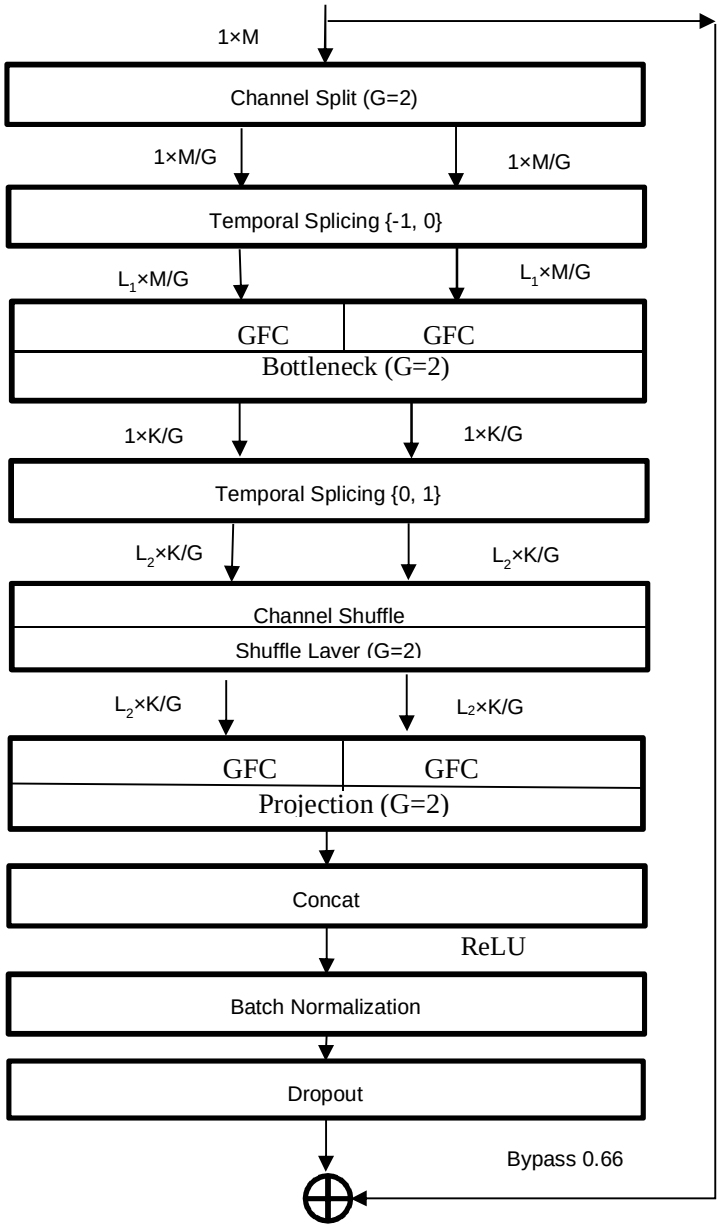

Fig.1.c) TDNN-F-2 modules

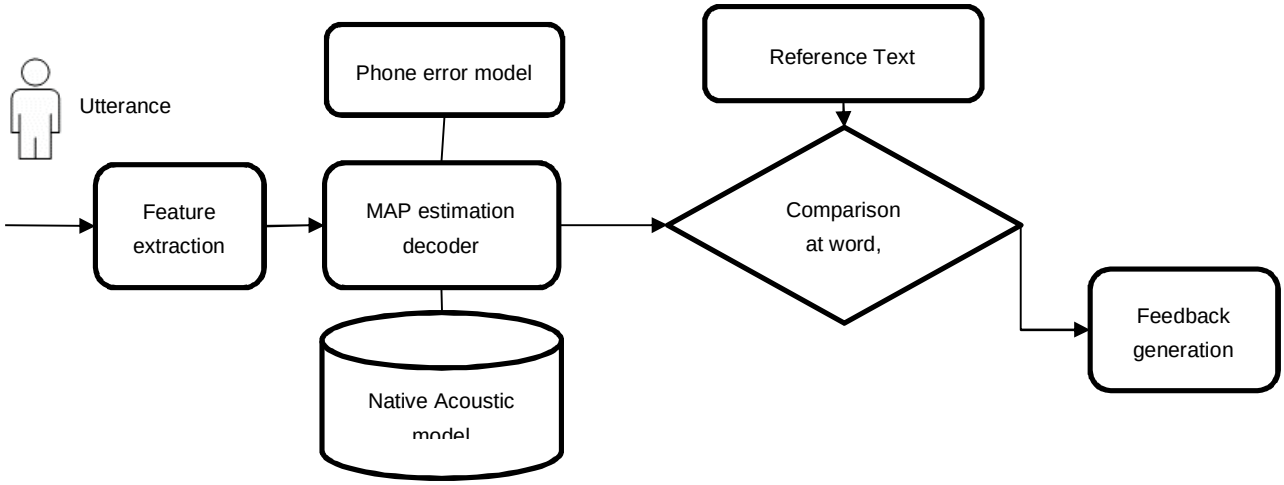

Fig 3. The pipeline of proposed system

Volume: 17 Issue: 01 Pages: 6719-6727 (2025) ISSN: 0975-0290

For example, word “live” (pronunciation is /lɪv/) has the meaning of ‘live’ and (leave)/liːv/ has the meaning of ‘leave’. Both words are composed of three phonemes, the first and third speech being the same, and the length of the central vowel is different, but this makes the meaning completely different. For examples, there are live/leave, ship/sheep, cup/carp. There are also errors in diphthong and triphthong.

There is some commonality in English and Korean consonants in the form of cracked consonants such as /p/, /b/, /t/, /d/, /k/ and /g/. However, English consonant pronunciation has some characteristics different from that of Korean consonants.

Poor phonological discrimination of voiced and voiced consonants. Many learners make English pronunciation and pronounce unvoiced consonants as voiced consonants.

In particular, there are many phenomena that occur behind the silent consonants that come to the end of the word, such as voiced consonants, by unnecessarily adding the Korean vowel "to".

There is also a tendency for many learners to use silent consonants as a consonant in Korean

Also, many learners have difficulty to distinguish between tongue-end and tongue-floor sounds. English learners have the most difficulty in pronunciation tongue-tips.

We assume that English learners using the CAPT are primarily aware of the pronunciation of each individual word in the sentence, that English level is not low, and that they didn’t intentionally mispronounce.

These learners often confuse acoustically similar consonants and vowels.

To detect such confusing pronunciation errors, we define a set of confusions in the vowel and consonant.

The phoneme pairs belonging to the vowel confusion set are composed mostly of pairs of long vowels/short vowels and long vowels/doubles. We defined the phoneme error patterns discussed above as confusing phones and extended the language model to this regard.

Table 1. Confusing phoneme patterns

confusion set of vowels		Confusion set of consonants	
e	eɪ	s	θ
iː	ɪ	p	f
æ	e	f	v
ʌ	ɑː	t	d
ɜː	ɔː	k	g
uː	ʊ	l	r
ə	ɜː	v	b
ɔː	əʊ	b	p
ɒ	ɔː	ð	d
ɔ	əʊ	ʤ	ʧ
əʊ	ɒ		

For example, word “this” in the case of this, the pronunciation is /ðɪs/, but considering the confusing phoneme, we generate a mispronounced word with the pronunciation /dɪs/. This is denoted as this_1. Thus, we generate these meaningless words in the corpus and build a trigram language model.

Also, the pronunciation of the asset s / æset /some people mispronounce as / ɑː set /. In this case we denote it as asset_1 (the pronunciation symbol consists of / ɑː set /).

Generally, there are 16,000 words in the English little dictionary. Generally, for speech recognition, only those words are present in the language corpus. However, considering the error patterns, we extended the language corpus by about 9 times.

2.4 System configuration

The system engine is based on acoustic model trained native corpus and extended language model. The proposed system consists of a preprocessing module, a recognition module, an evaluation module and a feedback module.

The speech features of the speaker are then MBR decoded. After speaking, the recognition results are compared with the reference text and the evaluation is done by word and phoneme. The phone sequence from text are compared in Levinshtein distance. The comparison gives feedback on deletion, insertion, and replacement of phoneme.

3. EXPERIMENT AND RESULT

3.1 Database

Two databases are utilized in our research.

a) Non-native corpus

We use the ERK (English Read by Korean) database as non-native corpus.

This corpus consists of 2 000 English sentences uttered by 100 graduate students. All the speakers are non-native, and their mother tongue is Korean. A third of the speakers are female. Four English expert made the scores of each sentence, word, and phoneme. They scored every sentence on accuracy and fluency. The scoring metric of word and sentence is listed in Table 2 and Table 3.

Table 2. The scoring metric of sentence

9-10	The overall pronunciation of the sentences is excellent, which accurate phonology and no obvious pronunciation mistakes.
7-8	The overall pronunciation of the sentences is good, with a few pronunciation mistakes
5-6	The overall pronunciation of the sentence is understandable, with many pronunciation mistakes and accent, but it does not affect the understanding of basic meanings
3-4	Poor, clumsy and rigid pronunciation of the sentence as a whole, with serious pronunciation mistakes
0-2	Extremely poor pronunciation and only one or two words are recognizable.

Table 3. The scoring metric of word

10	The pronunciation of the word is perfect
7-9	Most phones in this word are pronunced correctly but ha ve accents
4-6	Less than 30% of phones in this word are wrongly pronu

Volume: 17 Issue: 01 Pages: 6719-6727 (2025) ISSN: 0975-0290

	ced
2-3	More than 30% of phones in this word are wrongly pron unced. In another case, the word is mispronunced as som e toher word
0-1	The pronunciation is hard to distinguish or no voice

b) Native corpus

As a native corpus, we used the Tedium database [42]. The detailed information of the database is as follows.

- It consists of 2,351 spoken dialogues in NIST sphere format (SPH).
- The recording time is 452 hours in total.
- There were 2,351 aligned sentences in STM format.
- Development set and test set
- Lexicon dictionary

3.2 Training

We use speech recognition framework Kaldi5.5.[34]

After the monophone training, we train the triphone context-dependent GMM-HMM. After LDA+fMMR+SAT transformation (tri4b), it is ended with TDNN-F-HMM. We take 40-dimensional MFCC and 30-dimensional ivector as input feature.

During the training process, data argumentation technique is used. After applying speed perturbation and volume perturbation, NoiseX-92 noises and echo and room impulse responses were mixed. These data include room impulse responses (RIRs) and noises.

These noises used are first, a point source noise (843 noise recordings, background or foreground noise).[36] Second, the RWCP sound scene database and various room impulse responses [37]. There is a total of 325 real room impulse responses. Third, the simulated room impulse response signals [38].

The acoustic model consists of twelve proposed TDNN-F layers. The activation function in each TDNN-F layer is ReLU. The results in the test set are shown in Table 3.

The phoneme accuracy in the test set is 97%, which is like in TDNN-Fs

Table 4. PER of recipes

	Triphone	TDNN	TDNN-F	TDNN-F-2
DEV	27.83	12.41	4.19	4.27
TEST	25.11	11.75	3.24	3.47

3.3 Feedback generation

After analyzing pronunciation error, it provides feedback on substitution, deletion, and insertion errors.

For example, when he uttered a sentence “He lives in Brussels now-he’s a teacher at the international school there” we would say that if he did not utter the word “a”, we would say that “a” was deleted, i.e., not uttered.

For example, when he uttered the sentence “Ben, that’s nice,” we mean that the /b/ phoneme in Ben was replaced by /D/, if the recognition result was “then, that’s nice”.

When he utters the sentence “three, why” we mean that if result is “three, oh, why”, “oh” is inserted.

Of course, detecting mispronunciation is not the ultimate goal. An intuitive and effective feedback allows foreign language learners to not only know the pronunciation error but also learn how to correct it. Only displaying score message is not enough. The message must instruct how to correct.

For example, when uttering the sentence “What do people really want for their birthday,” the recognition result is perceived as “do->the”, meaning that /d/ in /du:/ is distorted /ð/ in the pronunciation of the word do is uttered as /u://i:/. Then, the following feedback message is displayed on the screen:

“You pronounced the /d/ in ‘do’ as / ð / in that. You should pronounce as /d/ in door. Try again,” thus show learners how to utter phonemes correctly. Comparing canonical speech with its own speech, in result, learner can perceive the difference between correct and incorrect phonemes. Also, clicking the phoneme image, allows learner to play the corresponding native audio.

3.4 Results analysis in ERK

The ERK dataset was tested. We compare the performance by Pearson correlation coefficient (r).

In the test set, 200 speech files were tested, each of which was spoken by 20 random utterances.

The word error rate and phone error rate are evaluated for every utterance. We also found and compared the correlation between the score from our system and human score.

The correlation coefficients of scores scored by the expert in the ERK sentence by sentence are given in Table 5.

Table 5. The inter-inter correlations in ERK

	R0	R1	R2	R3
R0		0.554	0.655	0.67
R1			0.594	0.575
R2				0.652

The average coefficient of four experts is 0.62±0.05. The correlation between R0 and R3 is the highest. The correlation between our method and human scores is r = 0.507

The results for each sentence in test set is as follows.

The correlation between character accuracy and human scores is r = 0.477.The correlation between word accuracy and human scores is r = 0.472.The correlation between phoneme accuracy and human scores is r = 0.507.

Table 6. Correlations of different methods in ERK

	Our met hod	Character accuracy-human	Word accuracy-huma n**Error! Reference source not found.**	Human-h uman
Correl ation	0.50 7	0.477	0.472	0.617

The Pearson correlation coefficient of the proposed method is very similar to the human score, with 30% improvement over the previous method.

Next, we evaluate the computational costs and model size of networks

Volume: 17 Issue: 01 Pages: 6719-6727 (2025) ISSN: 0975-0290

In TDNN, FC layers costs 2LMN+N FLOPs and has parameters of LMN+N. In TDNN-F, first FC costs 2L1MK+K FLOPs, size is L1MK+K, second FC costs 2L2KN+N FLOPs and size is L2KN+N. In proposed TDNN-F, grouped FC costs 2L1(M/G)K+K FLOPs, has parameters of L1(M/G)K+K. Next FC layer costs L2(K/G)N+N FLOPs and has parameters of L2(K/G)+N. To compare the computational cost of TDNN and TDNN-F effectively, we assume that the input and output dimensions are equal (M=N). The bottleneck dimension is 1/10 of the input dimension. (K=M/10, L=3, L1=L2=[L/2]) Under these assumptions, the computational cost of the TDNN neural network is 6M2+M FLOPs, model size is 3M2+M. The computational cost of the TDNN-F neural network is 2/5M2+10/11×M FLOPs, the parameters of model is M2/5+11×M/10. The computational costs of TDNN-F with grouped FC is M2/5+11M/10 FLOPs, the parameters of model are M2/10+11×M/10.

Thus, TDNN-F requires 90% less computational costs than TDNN. TDNN-F with grouped FC layers requires 50% less costs and model size then TDNN-F. Consequently, compared to the neural network composed by TDNN and the neural network composed by TDNN-F, the number of parameters decreased by 50%, and the accuracy maintained.

The real-time factor (RTF) decreased by a factor of 0.8.

3.5 Feedback generation analysis

The system evaluated the effectiveness of the instruction and playing audio/voice feedback. The tester tested the case of only instruction and listening to speech and corrected pronunciation. The examiner selected a dialog from textbook. After evaluation, he received a pronunciation feedback. Table 7 shows the effectiveness between message feedback and self-audio feedback on the phoneme's correction.

Table7. Statistics showing how feedback help learners correct their pronunciation errors.

	Corrected after viewing instruction	Corrected after playing audio
Deletion	18	19
Insertion	5	5
Sustition	164	231

The results show that the feedback is not so good as the type of mispronunciation errors. Learners can correct about half of the errors from feedback and modify the shape of the mouth, tongue and lip. Playing audio/voice is not very helpful. However, the substitution error can be replaced during the speaking, although it is correct in the first attempt. This is because the learner's habits and the background from his/her mother tongue, and it is difficult to correct at once.

4. CONCLUSION

In this paper, we propose a method to detect mispronunciation detection and feedback generation using a deep acoustic model based on a factorized TDNN with grouped FC and shuffle operation. Language model is generated from phone error patterns. Unlike the previous method, the proposed method is based on the native acoustic model and the trigram language model, pronunciation is evaluated by phone error rate.

The proposed TDNN-F neural network costs 50% less than the TDNN-F and the model size is reduced by halves. It can be deployed in mobile devices. Feedback with details of error and multimedia can help learner to correct mispronunciation.

REFERENCES

[1] Neri, A., Cucchiarini, C., Strik, H., 2002a. Feedback in computer assisted pronunciation training: technology push or demand pull. In: Proceedings of the International Conference on Spoken Language Processing, pp. 1209–1212.

[2] M. C. Pennington, "Computer-Aided Pronunciation Pedagogy: Promise, Limitations, Directions," Computer Assisted Language Learning, vol. 12, no. 5, pp. 427–440, 1999

[3] M. A. Salteh and K. Sadeghi, "Teachers and Students Attitudes Toward Error Correction in L2 Writing," The Journal of Asia TEFL, vol. 12, no. 3, pp. 1–31, 2015.

[4] S. Robertson, C. Munteanu, and G. Penn, "Pronunciation Error Detection for New Language Learners." in INTERSPEECH, 2016, pp. 2691–2695

[5] R. Srikanth and J. Salsman, "Automatic Pronunciation Scoring And Mispronunciation Detection Using CMUSphinx," in Proceedings of the Workshop on Speech and Language Processing Tools in Education, 2012, pp. 61–68.

[6] Franco, H., Neumeyer, L., Digalakis, V., Ronen, O., 2000. Combination of machine scores for automatic grading of pronunciation quality. Speech Communication 30, 121–130.

[7] Molina, C., Yoma, N.B., Wuth, J., Vivanco, H., Asr based pronunciation evaluation with automatically generated competing vocabulary and classifier fusion. Speech Communication 51, 485–498.

[8] Witt, S.M., Young, S.J., 2000. Phone-level pronunciation scoring and assessment for interactive language learning. Speech Communication 30, 95–108.

[9] Witt, S., Young, S., 1997a. Computer-assisted pronunciation teaching based on automatic speech recognition. Language Teaching and Language Technology Groningen, The Netherlands, 25–35.

[10] Witt, S., Young, S.J., 1997b. Language learning based on non-native speech recognition, in: Proc. Eurospeech, pp. 633–636.

[11] Zhang, F., Huang, C., Soong, F.K., Chu, M., Wang, R.H., 2008.Automatic mispronunciation detection for Mandarin. Proc. ICASSP- 2008. IEEE, pp. 5077–5080.

[12] Wang et al, 2012. Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training. Proc. ICASSP-2012. IEEE, pp. 5049–5052.

[13] Yan, K., Gong, S., 2011. Pronunciation proficiency evaluation based on discriminatively refined acoustic models. IJITCS 3, 17–23.

Volume: 17 Issue: 01 Pages: 6719-6727 (2025) ISSN: 0975-0290

[14] Qian, X., Soong, F.K., Meng, H.M., 2010. Discriminative acoustic model for improving mispronunciation detection and diagnosis in computeraided pronunciation training (CAPT). Proc. InterSpeech-2010. ISCA, pp. 757–760.

[15] Ito, A., Lim, Y.-L., Suzuki, M., Makino, S., 2005. Pronunciation error detection method based on error rule clustering using a decision tree. Proc. Eurospeech-2005. ISCA, pp. 173–176.

[16] Truong, K., 2004. Automatic Pronunciation Error Detection in Dutch as a Second Language: An Acoustic–phonetic Approach. Master's Thesis, Utrecht University, The Netherlands.

[17] Strik, H., Truong, K., de Wet, F., Cucchiarini, C., 2009. Comparing different approaches for automatic pronunciation error detection. Speech Commun. 51 (10), 845–852

[18] van Doremalen, J., Cucchiarini, C., Strik, H., 2009. Automatic detection of vowel pronunciation errors using multiple information sources. Proc. ASRU-2009. IEEE, pp. 580–585.

[19] Wei, S., Hu, G., Hu, Y., Wang, R.H., 2009. A new method for mispronunciation detection using support vector machine based on pronunciation space models. Speech Commun. 51 (10), 896–905.

[20] Jie, J., Xu, B., 2009. Exploring the automatic mispronunciation detection of confusable phones for Mandarin. Proc. ICASSP-2009. IEEE, pp. 4833–4836.

[21] Yoon, S.-Y., Hasegawa-Johnson, M., Sproat, R., 2010. Landmark-based automated pronunciation error detection. Proc. InterSpeech-2010. ISCA, pp. 614–617.

[22] Hu, W., Qian, Y., Soong, F.K., 2013. A new DNN-based high quality pronunciation evaluation for computer-aided language learning (CALL). Proc. InterSpeech-2013. ISCA, pp. 1886–1890

[23] Hinton, G.E., Deng, L., Yu, D., Dahl, G.E., Rahman Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., Kingsbury, B., 2012a. Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Process. Mag. 29 (6), 82–97.

[24] Qian, X., Meng, H.M., Soong, F.K., 2012. The use of DBN-HMMs for mispronunciation detection and diagnosis in L2 English to support computer-aided pronunciation training. Proc. InterSpeech-2012. ISCA.

[25] Hu, W., Qian, Y., Soong, F.K., 2014a. A DNN-based acoustic modeling of tonal language and its application to Mandarin pronunciation training. Proc. ICASSP-2014. IEEE, pp. 3230–3234.

[26] Hu, W., Qian, Y., Soong, F.K., Wang, Y., 2015. Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers. Speech Communication 67, 154–166.

[27] Li, K., Mao, S., Li, X., Wu, Z., Meng, H., 2018. Automatic lexical stress and pitch accent detection for L2 English speech using multi-distribution deep neural networks. Speech Communication 96, 28–36.

[28] Shiran Dudy and Steven Bedrick and Meysam Asgari and Alexander Kain, 2018. Automatic analysis of pronunciations for children with speech sound disorders. Computer Speech & Language 50, 62-84.

[29] Mostafa Ali Shahin, Beena Ahmed, 2019. Anomaly detection based pronunciation verification approach using speech attribute features, Speech Communication 111, 29–43.

[30] Longfei Yang, Kaiqi Fu, Jinsong Zhang, Takahiro Shinozaki,., 2021. Non-native acoustic modeling for mispronunciation verification based on language adversarial representation learning, Neural Networks Oct;142:597-607

[31] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE, 2013, pp. 6645–6649.

[32] Tu, M., Grabek, A., Liss, J., Berisha, V., 2018. Investigating the role of L1 in automatic pronunciation evaluation of L2 speech, in: Proc. Interspeech, pp. 1636–1640.

[33] Moustroufas, N., Digalakis, V., 2007. Automatic pronunciation evaluation of foreign speakers using unknown text. Computer Speech and Language 21,219–230

[34] D. Povey et al., "The Kaldi speech recognition toolkit," in Proc. ASRU, 2011.

[35] Xu, H., Povey, D., Mangu, L., Zhu, J. (2011). Minimum bayes risk decoding and system combination based on a recursion for edit distance. Computer Speech & Language, 25(4), 802–828.

[36] https://www.freesound.org/

[37] https://github.com/tomkocse/sim-rir-preparation/

[38] http://www.openslr.org/resources/26/sim_rir.zip

[39] Jiang Fu, Yuya Chiba,Takashi Nose, Akinoir Ito., 2020. Automatic assessment of English proficiency for Japanese learners without reference sentences based on deep neural network acoustic models. Speech Communication 84, 154–166.

[40] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme recognition using time-delay neural networks," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 37, no. 3, pp. 328–339, Mar. 1989.

[41] J. Xue, J. Li, and Y. Gong, "Restructuring of deep neural network acoustic models with singular value decomposition." in Interspeech, 2013, pp. 2365–2369.

[42] https://www.openslr.org/51/

[43] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in INTERSPEECH, 2015, pp. 3214–3218.

[44] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in Proc. Interspeech, 2018, pp. 3743–3747.

[45] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: an extremely efficient convolutional neural network for mobile devices," in Proc. CVPR, 2018, pp. 6848–6856.