Vol.2018-MUS-119 No.26
Vol.2018-SLP-122 No.26
2018/6/16

Evaluation of DNN-based Speech Recognition for English Spoken by Japanese Learners

Jiang Fu† Yuya Chiba† Takashi Nose† Akinori Ito†

†Graduate School of Engineering, Tohoku University 6-6-5, Aramaki Aza Aoba, Aoba-ku, Sendai, Miyagi 980-8579 Japan

E-mail: †{fujiang@ecei, yuya@spcom.ecei, tnose@m, aito@spcom.ecei}.tohoku.ac.jp

Abstract: Regarding the assistance of computer-assisted language learning (CALL) systems to make foreign language learning easier, it is necessary to recognize the utterances of the learner with high accuracy. The quality of CALL systems mainly depends on the accuracy of automatic speech recognition (ASR). However, since pronunciation of non-native speakers is greatly different from that of native speakers, existing ASR system cannot well recognize speech accurately. To solve this problem, this research projects an acoustic model based on deep neural networks (DNN), which is trained by using ERJ (English Read by Japanese) database collected from 202 Japanese learners. Compared with traditional ASR systems, this new system significantly promotes speech recognition rate.

Keywords: Speech recognition, Deep neural networks, ERJ database, CALL

1. Introduction

With current globalization, English has become the most important second language (L2) for the people whose mother tongue language (L1) is not English. Using computer-assisted language learning (CALL) systems or computer-based learning tools to promote the level of English learning has become a hot topic in order to make foreign language learning easier, as the most effective way for non-native learners is peer-to-peer targeted teaching and the non-native learners can make full use of a CALL system with its flexible and convenient features. The conventional CALL system was often used for practicing "reading" and "writing" skills, and only a few CALL systems were equipped with a dialogue system as an efficient study method for "speaking" and "listening" [1]. To some extent, the quality of the dialogue system mainly depends on the accuracy of automatic speech recognition (ASR). Traditional speech recognition systems usually exploit GMM-HMM model as the acoustic model of LVCSR (Large Vocabulary Continuous Speech Recognition), which combines GMM (Gaussian Mixture Model) with HMM (Hidden Markov Model). As the great impact of deep learning on speech recognition, replacing GMM with DNN (Deep Neural Network) in the acoustic model helps to promote the recognition accuracy, compared to GMM-HMM results [2]. Therefore, we use DNN-HMM model as the acoustic model of LVCSR in our CALL system with ERJ (English Read by Japanese) database to evaluate the accuracy of ASR system compared with other conventional ASR systems.

The objective of this paper is to compare the performance of an acoustic model based on DNN-HMM to that of GMM-HMM by the non-native speech recognition task. The target language (L2) is English and the supposed mother tongue of learners (L1) is Japanese. We used the ERJ database as the material of training and testing the speech recognition, and Kaldi was used as the speech recognition system.

2. CALL Systems and Recognition of Non-native Speech

Speech-based CALL systems need two contradictory requirements for speech recognizer: the recognizer needs to distinguish the low-proficient speech from the native speech, and at the same time, the recognizer needs to recognize the learner's speech regardless of its proficiency. The former requirement comes from the task of pronunciation assessment, where the system automatically scores the learner's speech. The latter requirement is needed when the system makes conversation with the learner using the L2 language [1,3].

There have been many research studies on the first requirement. The most famous one is the Goodness of Pronunciation (GOP), based on ASR techniques for phone level pronunciation scoring of non-native speech [4], which is later widely used in utterance verification.

The second requirement is known as the non-native speech recognition. ASR systems with acoustic models trained on native speakers' data do not perform well when used to recognize speech of L2 learners. Therefore, a large number of research studies have been conducted on the methods for improving non-native speech recognition. With regard to the adaptation of acoustic model, a regression tree to the speakers' pronunciation proficiency had been adapted in acoustic models to enhance non-native speech recognition [5]. Furthermore, four acoustic model adaption methods have been compared with a small amount of non-native speech [6]. A proposed acoustic model adaptation method based on analyzing the second language speech pronunciation variants has a better speech recognition result than the conventional approach for non-native speech recognition [7]. Another research on non-native speaker adaptation proposed three interpolation techniques to improve non-native speech recognition [8]. In our former work, we developed three acoustic models of different levels on pronunciation proficiency for Japanese learners in a

† Tohoku University.

CALL system [9]. It is also discussed that utterance selection and verification in non-native speech data can significantly reduce the error rate [10]. Besides the improvement on acoustic modeling in non-native speech recognition, using different phoneme sets and a special pronunciation lexicon can be useful in the predetermination of already known mother tongue speakers for better performance on non-native speech recognition [11,12].

All of the above studies are based on traditional GMM-HMM ASR systems. Thanks for the advances in both machine learning method and rapidly developing computer hardware, DNNs promote a lot on the native speech recognition [2]. However, only a few works have been done for DNN-based non-native speech recognition. For example, Chen and Cheng used DNN-based acoustic models for recognition of non-native Mandarin [13]. Cheng et al. used the DNN-based acoustic models for assessing the pronunciation of a learner [14].

In this paper, we will adapt DNNs to speech recognition of English spoken by Japanese learners compared with the traditional GMM-HMM systems.

3. DNN-based Acoustic Modeling of Non-native Speech

Over the last few years, as already shown in Sect. 2, researchers proposed various approaches to promote the speech recognition accuracy for non-native speech. Especially in CALL systems, the system performance in the process of interaction with users depends on the accuracy of ASR. In this paper, we want to investigate the accuracy on non-native speech recognition using DNN-based acoustic modeling method.

The acoustic model has been based on the GMM-HMM for a long time. Neural network technology has not been widely used on speech recognition in early days. One of the reasons was that the number of layers in neural networks at that time was small, and increasing the number of layers and the number of nodes would make the neural network training some difficult or even impossible to achieve.

The deep learning techniques developed in recent years have shown that neural networks are more suitable for acoustic modeling [2]. There are usually two steps to training DNNs. First, use pre-training to initialize the network weights; Then use the backpropagation algorithm to tune the network.

In the process of DNN-based acoustic modeling in Kaldi toolkit [15], the training is not immediately started from utterance-level transcriptions, but from the phoneme-to-audio alignments which were generated by a GMM-HMM system. 13 Mel-frequency cepstral coefficients (MFCCs) along with the normalized energy and their first and second order derivatives are extracted in the baseline GMM-HMM system. Feature vectors are typically counted every 10ms with an overlapping analysis window of 25ms. The recipe begins with the monophone training, then first triphone pass which comprises 2000 regression tree leaves and 11000 Gaussians and second triphone pass which comprises 2500 regression tree leaves and 20000 Gaussians will be conducted. Other additional steps, such as linear discriminant analysis (LDA), maximum likelihood linear transform (MLLT) and speaker adaptive training (SAT), will be performed later.

As a tool for classification, basic neural networks have been implemented into the original system, which have input nodes corresponding to the dimensions of the split FBANK or MFCCs features (13 MFCCs, the energy and their first and second order derivatives) and output nodes corresponding to context dependent triphones of GMM-HMM system. The part of DNN is trained using 90% of the training data for training and the remaining 10% for evaluation. In the step of NNET1 in Kaldi toolkit, the neural network sets a simpler sigmoid function in hidden layers and softmax function in an output layer. Meanwhile, NNET2 sets basic hyperbolic tangent function or p-norm nonlinearity in hidden layers.

4. Experiments with ERJ Database

4.1 Data preparation in ERJ database

The English Read by Japanese (ERJ) database is a corpus of non-native English utterances spoken by Japanese students. [16,17]. The ERJ database is widely used for many research studies on the development of the CALL systems [18-20], non-native speech recognition [21] and synthesis [22].

English speech samples spoken by 100 male and 102 female Japanese students are included in the database. All the sentences are divided into eight groups (S1 to S8), and all the words are divided into five groups (W1 to W5). The required amount of the recording in a sentence group is about 120 sentences and in a word group is about 220 words. Therefore, each sentence and each word were read by nearly 25 speakers and 20 speakers, respectively. Table 1 and Table 2 show the details of sentence subsets and word subsets, respectively.

TABLE 1(A)
SENTENCE SUBSETS IN ERJ DATABASE

	S1	S2	S3	S4
# of sentences	120	123	122	123
# of words	742	838	823	823
Male	13	12	11	11
Female	13	13	11	12

TABLE 1(B)
SENTENCE SUBSETS IN ERJ DATABASE

	S5	S6	S7	S8
# of sentences	124	123	123	122
# of words	814	899	882	879
Male	13	14	12	14
Female	13	13	13	14

In order to build the ASR system, firstly, sentence parts of the ERJ database were chosen and analyzed with considering the phones. The setting of phonemic symbols in data preparation used the default phonemic symbols in the CMU pronouncing dictionary [23], which are listed in Table 3.

Once the repeated sentences were selected out in sentence part of ERJ database, totally 39 phones were counted in every subset of ERJ sentence part. After counting the phonemic symbols in revised sentence subsets in ERJ database, finally the training set was determined with S1, S2, S4, S5, S6, and S7, the development set was S3 and the test set was S8.

TABLE 2
WORD SUBSETS IN ERJ DATABASE

# of words	S1 226	S2 224	S3 225	S4 222	S5 229
Male	13	12	11	11	20
Female	13	13	11	12	20

TABLE 3
PHONEMIC SYMBOLS IN CMU PRONOUNCING DICTIONARY

B, D, G, P, T, K, JH, CH, S, SH, Z, ZH, F, TH, V, DH, M, N, NG, L, R, W, Y, HH, IY, IH, EH, EY, AE, AA, AW, AY, AH, AO, Y, OW, UH, UW, ER

4.2 Language modeling

We modified CMU pronouncing dictionary as the baseline pronunciation lexicon by changing all the stressed phonemes to non-stressed phonemes, because Japanese speakers' English pronunciation generally is different from native speakers, and if we have more data, it could improve the modeling of the specific phone in the acoustic model.

As for the language model, we selected the sentence parts of ERJ database and removed all repeated sentences. We developed bigram and trigram language models from 980 different sentences in ERJ database.

TABLE 4
BEST EXPERIMENT RESULTS IN DEV SET (GMM-HMM)

Gram Phone	Bi Mono	Bi Tri	Tri Mono	Tri Tri
WER[%]	7.54	7.95	7.09	7.83
SER[%]	24.43	26.61	22.56	25.93

TABLE 5
BEST EXPERIMENT RESULTS IN TEST SET (GMM-HMM)

Gram Phone	Bi Mono	Bi Tri	Tri Mono	Tri Tri
WER[%]	7.70	5.32	6.90	5.11
SER[%]	23.48	20.47	22.26	19.75

4.3 Results of monophone and triphone training

In this part, the monophone training and triphone training were run in Kaldi Toolkit. The language model scale has been set from 10 to 50 with a beam width of 13 to find the lowest WER (word error rate) and the lowest SER (sentence error rate).

The results are presented in Tables 4 and 5. From the results above, using the trigram language model contributes a bit to revising performance of this ASR system. Furthermore, monophone training is better than triphone training with the development set, but not with the test set.

After the GMM-HMM model was established, we can do more about speech recognition with ERJ database using deep learning method.

4.4 Results of DNN training

The setting details of DNN training are listed in Table 6. We used four conditions (NN11, NN21, NN22 and NN23), in which NN11 stands for NNET1 setup in Kaldi and the other three NNs stand for NNET2 setup in the same toolkit. As the initial step, we regarded NN11 as an experimental neural network training with its basic sigmoid activation function and NNET2 setup as the main focus in our ASR system for the future application. The feature vectors of NN11 were extracted with FBANK features, as previous research studies have shown FBANK features are more

TABLE 6
SETTINGS IN DNN TRAINING

	NN11	NN21	NN22	NN23
Feature	40 FBANK	40 MFCCs	40 MFCCs	40 MFCCs
Hidden Type	sigmoid	tanh	p-norm	p-norm E.L.
Input Nodes	440	360	360	360
HidLayer HidDim	4 1024	4 1024	4 p-norm	4 p-norm
Learning Rate	0.008	0.02	0.02	0.02
Output Nodes	3168	1551	1551	1551

TABLE 7
BEST EXPERIMENT RESULTS IN DEV SET

	NN11	NN21	NN22	NN23
WER[%]	2.99	4.99	3.13	2.81
SER[%]	12.75	18.29	13.22	12.28

TABLE 8
BEST EXPERIMENT RESULTS IN TEST SET

	NN11	NN21	NN22	NN23
WER[%]	2.05	3.42	2.31	1.97
SER[%]	9.24	15.42	10.62	9.19

suitable in neural network training [24,25]. Concerning the hidden layer, the four-layer neural network is enough in most of the databases [26]. We used MFCCs feature without dimension reduction applied to NNET2 setup, which was slightly different with FBANK. Three different activation functions were chosen in NNET2 setup for the investigation. Regarding the language model, only the trigram model was modified during the training. Results of DNN training are showed in Tables 7 and 8.

From these results, DNN training significantly improved the results of WER and SER. Especially the third method of NNET2 has a comparable result with that of the NNET1 method. Above all, DNN training enhances the performance of the ASR system.

4.5 Relationship between the proficiency and WER

In the ERJ database, we already have evaluation scores about every Japanese learner's English speaking level judged by professional teachers. The distribution of the segmental scores is shown in Fig. 1. Here, we divided all speakers into three classes: LOW, MID and HIGH, where the speaker having a lower score than the 1st quantile were classified into LOW, those having higher score than the 3rd quantile were HIGH, and the others were MID. To investigate the relationship between the proficiency and WER among all the students in ERJ database, cross-validation has been applied in ERJ database to redefine the training set and test set for decoding. P-norm ensemble activation function in NNET2 was selected in DNN training step which has the same parameter with NN23 showed in Sect. 4.4.

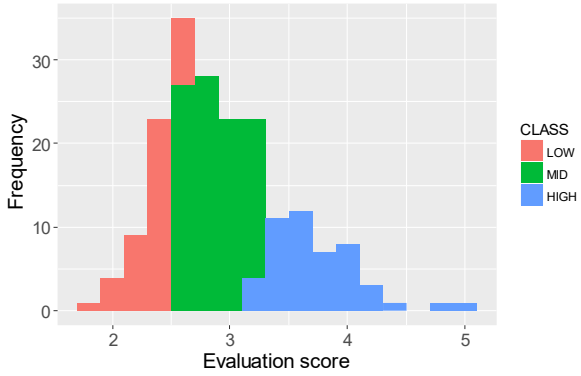

Fig. 1 Histogram of proficiency scores of utterances in ERJ.

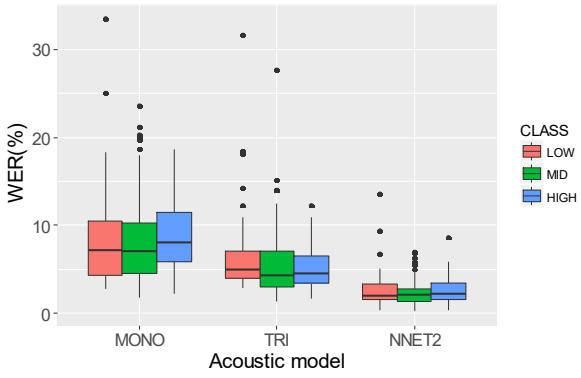

Fig. 2 Boxplots of WER for three acoustic models, dependent to the segmental score of the speakers.

Fig. 2 shows the relationship between the WER from our established ASR system and segmental score from teachers for both GMM-HMM and DNN-HMM. From these results, DNN training significantly improved the results of WER almost in every Japanese learner's decoding. This ASR system cannot well automatically predict the learner's English fluency, but still to the other hand, using this system can bring enough high speech recognition rates for application in future English CALL system.

5. Conclusions

We developed an ASR system with the sentence part of ERJ database. To clarify the advantage of DNN-HMM trained acoustic model embedded in this ASR system, we also established the traditional GMM-HMM acoustic model to be compared. The results indicated DNN-based method significantly improves the accuracy of non-native speech recognition. In addition, this DNN-based speech recognizer has a farsighted meaning to our further CALL system.

Reference

[1] Lee, S., Noh, H., Lee, J., Lee, K., & Lee, G. G., POSTECH approaches for dialog-based English conversation tutoring. In Proc. 2010 APSIPA annual summit and conference, pp. 794-803, 2010.

[2] Hinton, G., Deng L., Yu, D., E. Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., N. Sainath, T., et al., Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, Vol. 29, No. 6, 82-97, 2012.

[3] Raux, A., & Eskenazi, M., Using task-oriented spoken dialogue systems for language learning: potential, practical applications and challenges. In Proc. InSTIL/ICALL Symposium, pp. 147-150, 2004.

[4] Witt, S., & Young, S. J., Language learning based on non-native speech recognition. In Proc. Fifth European Conference on Speech Communication and Technology, pp. 633-636, 1997.

[5] Minematsu, N., Kurata, G., & Hirose, K., Integration of MLLR adaptation with pronunciation proficiency adaptation for non-native speech recognition. In Proc. Seventh International Conference on Spoken Language Processing, pp. 529-531, 2002.

[6] Wang, Z., Schultz, T., & Waibel, A., Comparison of acoustic model adaptation techniques on non-native speech. In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 540-543, 2003.

[7] Oh, Y. R., Yoon, J. S., & Kim, H. K., Acoustic model adaptation based on pronunciation variability analysis for non-native speech recognition. Speech Communication, 49(1), 59-70, 2007.

[8] Tan, T. P., & Besacier, L., Acoustic model interpolation for non-native speech recognition. In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1009-1012, 2007.

[9] 安斎、咸、伊藤：「日本人英語発話からの文法誤り検出」情報処理学会研究会資料 Vol. 2011-SLP-85, No. 15, 2011

[10] Van Doremalen, J., Cucchiarini, C., & Strik, H., Optimizing automatic speech recognition for low-proficient non-native speakers. EURASIP Journal on Audio, Speech, and Music Processing, 2010(1), 973-954, 2010.

[11] Wang, X., & Yamamoto, S., Second language speech recognition using multiple-pass decoding with lexicon represented by multiple reduced phoneme sets. In Sixteenth Annual Conference of the

International Speech Communication Association, pp.1265-1269, 2015.
[12] Wang, X., & Yamamoto, S., Speech recognition of english by Japanese using lexicon represented by multiple reduced phoneme sets. IEICE Transactions on Information and Systems, 98(12), 2271-2279, 2015.
[13] Chen, X., & Cheng, J., Deep neural network acoustic modeling for native and non-native Mandarin speech recognition. In Proc. Ninth International Symposium on Chinese Spoken Language Processing (ISCSLP), pp. 6-9, 2014.
[14] Cheng, J., Chen, X., & Metallinou, A., Deep neural network acoustic models for spoken assessment applications. Speech Communication, 73, 14-27, 2015.
[15] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al., The Kaldi speech recognition toolkit. In IEEE Workshop on Automatic Speech Recognition and Understanding, 2011.
[16] English Speech Database Read by Japanese Students, http://research.nii.ac.jp/src/en/UME-ERJ.html
[17] Makino, T., & Aoki, R., English read by Japanese phonetic corpus: an interim report. Research in Language, 10(1), 79-95, 2012.
[18] Minematsu, N., Okabe, K., Ogaki, K., & Hirose, K., Measurement of objective intelligibility of Japanese accented English using ERJ (English Read by Japanese) database. In Proc. Twelfth Annual Conference of the International Speech Communication Association, pp.1481-1484, 2011.
[19] Luo, D., Qiao, Y., Minematsu, N., Yamauchi, Y., & Hirose, K., Regularized-MLLR speaker adaptation for computer-assisted language learning system. In Proc. Eleventh Annual Conference of the International Speech Communication Association, pp.594-597, 2010.
[20] Ito, A., Tsutsui, R., Makino, S., & Suzuki, M., Recognition of English utterances with grammatical and lexical mistakes for dialogue-based CALL system. In Proc. Ninth Annual Conference of the International Speech Communication Association, pp.2819-2822, 2008.
[21] Wang, X., Kato, T., & Yamamoto, S., Phoneme set design based on integrated acoustic and linguistic features for second language speech recognition. IEICE Transactions on Information and Systems, 100(4), 857-864, 2017.
[22] Oshima, Y., Takamichi, S., Toda, T., Neubig, G., Sakti, S., & Nakamura, S., Non-native text-to-speech preserving speaker individuality based on partial correction of prosodic and phonetic characteristics. IEICE Transactions on Information and Systems, 99(12), 3132-3139, 2016.
[23] The CMU Pronouncing Dictionary, http://www.speech.cs.cmu.edu/cgi-bin/cmudict
[24] Yoshioka, T., Chen, X., & J. F. Gales, M., Impact of singlemicrophone dereverberation on DNN-based meeting transcription systems. In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 5527-5531, 2014.
[25] Mohamed, A., Hinton, G., & Penn, G., Understanding how deep belief networks perform acoustic modelling. In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 4273-4276, 2012.
[26] Pan, J., Liu, C., Wang, Z., Hu, Y., & Jiang, H., Investigation of deep neural networks (DNN) for large vocabulary continuous speech recognition: Why DNN surpasses GMMs in acoustic modeling. In Proc. Eighth International Symposium on Chinese Spoken Language Processing (ISCSLP), pp. 301-305, 2012.