Bottleneck feature-mediated DNN-based feature mapping for throat microphone speech recognition

Takahito Suzuki*, Jun Ogata†, Takashi Tsunakawa*, Masafumi Nishida* and Masafumi Nishimura*
*Shizuoka University, Shizuoka, Japan
E-mail: suzuki.takahito.14@shizuoka.ac.jp
†National Institute of Advanced Industrial Science and Technology, Japan

Abstract—Throat microphones are more robust to environmental noises than usual acoustic microphones such as close-talk microphones because they detect speech signals through skin vibrations rather than by air transmission. Throat microphones, however, cannot be used in conventional speech recognition systems because their acoustic characteristics are much different from those of the acoustic microphones. In this study, we propose a deep neural network (DNN)-based feature mapping method for throat microphone speech recognition. To utilize a large amount of training data recorded by acoustic microphones and effectively reduce the acoustic mismatch between the throat and acoustic microphones, we tried to use the bottleneck features to mediate between them. Evaluation results for a large-vocabulary speech recognition task of Japanese free conversation revealed that the proposed system had a 45.8% lower character error rate (75.5% → 40.9%) than the typical MFCC system trained from the acoustic microphone data.

I. INTRODUCTION

Deep neural networks (DNN) with high discrimination performance have recently been applied to speech recognition, and the recognition accuracy was reported to be almost the same level as that of a human's transcription of an English telephone conversation (Switchboard) [1]. However, recognition performance is still much degraded in highly non-stationary noise environments and needs to be further improved. As a method to suppress external noise, a throat microphone, which closely adheres to the throat and receives the vibrations directly from the skin by the piezo element, can be used [2][3]. It is more robust to external noise than a standard acoustic microphone, which receives vibrations of air. However, there is a big difference in acoustic characteristics between throat and acoustic microphones. Therefore, the recognition accuracy of a throat microphone deteriorates when using an acoustic model trained by speech collected with an acoustic microphone because of acoustic mismatch. Furthermore, it is difficult to train acoustic models from scratch with the speech of a throat microphone because it can use only a limited amount of data.

To solve this acoustic mismatch of the throat microphone in speech recognition, various methods have been proposed [4]-[6]. Lin et al. [6] proposed a DNN-based feature mapping from the throat microphone to the acoustic microphone by using acoustic models of a conventional acoustic microphone. Its feature transformation is from the mel-frequency cepstral coefficient (MFCC) of the throat microphone to the MFCC of the acoustic microphone.

In distant-talking speech-recognition task, Hiwaman et al. [7] proposed a method of DNN-based feature mapping from the MFCC feature space of a single distant microphone (SDM) to the bottleneck feature space of an individual head microphone (IHM) to suppress acoustic mismatch between IHM and SDM. The bottleneck feature (BNF) is extracted from the bottleneck layer of DNN that is trained to discriminate phonemes. The BNF has intrinsic information of phonemes and is more effective for phoneme discrimination than conventional features such as MFCC. In the work of Hiwaman et al. [7], the weights of DNN for feature mapping were initialized to the weights of DNN for extracting the BNF of an IHM and then fine-tuned with the MFCC of an SDM as the input signal and the BNF of the IHM as the supervised signal. Das et al. [8] reported that a DNN that was trained with English alignments and then retrained with a limited number of Turkish alignments had higher Turkish phoneme discrimination accuracy than a DNN that was randomly initialized and then trained with only Turkish alignments. Hence, feature mapping accuracy was assumed to be improved by devising the initial weights of DNN as done by Hiwaman et al. [7].

Inspired by the Hiwaman's work, we introduce the BNF-based feature mapping approach into throat-microphone speech-recognition task. In this study, we propose a DNN-based feature mapping from the MFCC of throat-microphone speech to the BNF of acoustic-microphone speech. By using this feature mapping as a pre-processing, a normal acoustic model trained from large-sized acoustic-microphone data can be applied on the following recognition process. Furthermore, we study an initialization technique of DNN to improve the feature mapping.

The rest of this paper is organized as follows. Section 2 describes the method of DNN-based feature mapping for throat microphone speech recognition, Section 3 describes conditions and results of recognition experiments in a Japanese large vocabulary continuous speech recognition (LVCSR) task, and Section 4 discusses conclusions and future works.

II. PROPOSED METHOD

A. Overview of our throat mic. speech recognition system

Fig. 1 shows the block diagram of the throat-microphone speech-recognition system. First, this system extracts a 13-dimensional MFCC of throat microphone input by applying

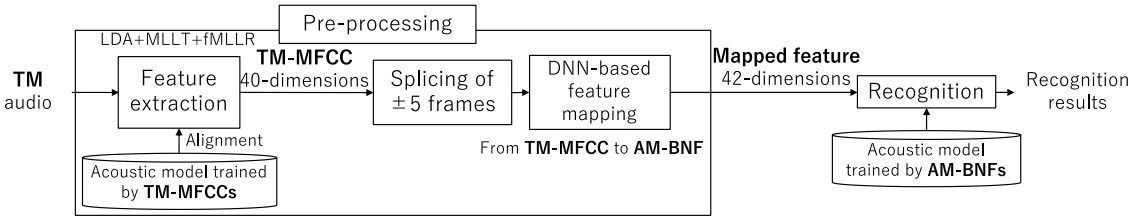

Fig. 1 Block diagram of our throat-microphone speech-recognition system (TM: throat microphone, AM: acoustic microphone)

cepstrum mean normalization (CMN) and then splices 4 frames on each side of the current frame. After that, the current frame is compressed to 40 dimensions with linear discriminant analysis (LDA), and maximum linear likelihood transformation (MLLT) is applied. Then, feature-space maximum linear likelihood regression (fMLLR) is also applied. In this study, we refer to the 40-dimensional feature vector as throat microphone (TM)-MFCC. To estimate the fMLLR transform, the alignment information of the throat microphone is obtained by using a GMM-HMM trained only with throat microphone speeches. Next, the system carries out the feature mapping from the 440-dimentional feature vector obtained by splicing ±5 frames of the TM-MFCC to the BNF of the acoustic microphone (AM-BNF). The feature vector estimated by the DNN-based mapping is input to the recognition system that has the GMM-HMM trained by AM-BNFs taken from a typical speech corpus.

B. *Training of DNN for extracting BNF*

The architecture of the DNN for extracting the BNF is shown in Fig. 2. The acoustic microphone (AM)-MFCC is extracted from speech data of the acoustic microphone as well as the TM-MFCC in pre-processing. Then the 440-dimensional features obtained by splicing ±5 frames of the AM-MFCC are used for training of the DNN as an input signal. The state alignments of the acoustic-microphone speeches are estimated by using a GMM-HMM trained only with the AM-MFCCs and used for training of the DNN as a supervised signal. The DNN is pre-trained by stacked denoising auto-encoders (SdA) and fine-tuned. Then a 42-dimentional BNF is estimated from the middle layer of the DNN.

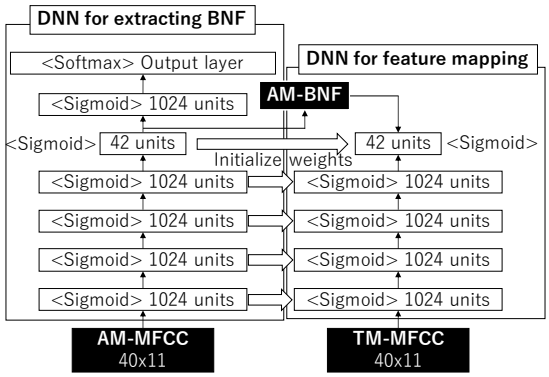

Fig. 2 Architectures of DNN for extracting BNF and for feature mapping

C. *Training of DNN for feature mapping*

Fig. 3 shows the training process of DNN for the feature mapping (FM-DNN). The AM-BNF and TM-MFCC are extracted from parallel data simultaneously recorded by the acoustic and throat microphones. The input signal for training of the FM-DNN is the 440-dimensional features obtained by splicing ±5 frames of the TM-MFCC, and the supervised signal is the AM-BNF. The FM-DNN has the same architecture as the DNN for extracting the BNF cut up to the bottleneck layer as shown in Fig. 2. The weights of the FM-DNN are initialized to the weights of the DNN for extracting BNF and fine-tuned.

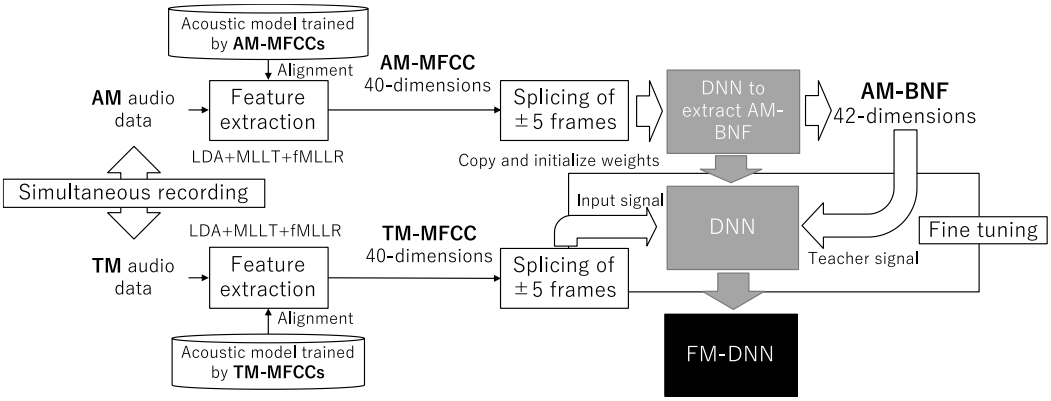

Fig. 3 Pipeline for training the DNN for feature mapping (TM: Throat microphone, AM: Acoustic microphone, FM-DNN: DNN for feature mapping)

III. EXPERIMENTS

A. Datasets

Parallel data used for training feature mapping was recorded simultaneously with a throat microphone (NANZU SH – 12jK) and an acoustic microphone (Sony EXM - CS 3) using a multi-track recorder (ZOOM R24). The sampling rate of the audio data is 16,000 Hz. We recorded 8 male speakers reading 504 Japanese phoneme balanced sentences for about 3 hours. The throat microphone speech data of this parallel data was also used for training the GMM-HMM of the throat microphone.

As test data, free conversations of 10 male speakers for about 19 minutes were recorded. The contents of these free conversations were group discussions. Each group consisted of three people. The speech data has less external noise because the conversations were recorded in a quiet environment. The speakers of the test data are not included in the training data.

Approximately 240 hours from the Corpus of Spontaneous Japanese (CSJ) were used for training data of the GMM-HMM of the acoustic microphone, and about 114 hours from the CSJ were used for training the DNN to extract the BNF.

B. Experimental conditions

Kaldi [9] was used for feature extraction, acoustic model training, and recognition experiments. Kaldi+PDNN [10] was also used to train the DNN to extract the BNF. In this training, the mini batch size was set to 256, the initial learning rate to 0.01, and number of iterations to 8. Keras was used to train the FM-DNN. In this, the mini batch size was set to 4096, and the initial learning rate to 0.001, and number of iterations to 100. The 3-gram language model was used and was generated from transcripts of the CSJ.

C. Experimental results

Three types of experiments were conducted to evaluate the effectiveness of the proposed throat-microphone speech-recognition system.

C-1. COMPARISON WITH CONVENTIONAL SYSTEMS

To evaluate the performances of four conventional systems and the proposed system, we conducted five recognition experiments.

(1) A system using the TM-MFCC as the input feature and a GMM-HMM trained with the AM-MFCC (AM GMM-HMM) as an acoustic model
(2) A system using the BNF of the throat microphone as the input feature and a GMM-HMM trained with the AM-BNF (AM Tandem) as an acoustic model
(3) A system using the TM-MFCC as the input feature and a GMM-HMM trained with the TM-MFCC (TM GMM-HMM) as an acoustic model
(4) A system using the BNF of the throat microphone as the input feature and a GMM-HMM trained with BNF of the throat microphone (TM Tandem) as an acoustic model

TABLE I
CHARACTER ERROR RATE (CER) OF CONVENTIONAL METHODS AND PROPOSED METHOD

Acoustic model	Type of input feature	CER [%]
(1) AM GMM-HMM	MFCC	74.6
(2) AM Tandem	BNF	96.6
(3) TM GMM-HMM	MFCC	52.4
(4) TM Tandem	BNF	48.9
(5) AM Tandem (Proposed)	Mapped feature	42.2

(5) A system using the feature mapped by the DNN as the input feature and a GMM-HMM trained with the AM-BNF (AM Tandem) as an acoustic model (proposed system)

Each system recognized some speech data of the throat microphone. The BNFs input to systems (2) and (4) are extracted from the DNNs trained with the acoustic and throat microphones, respectively. In these experiments, the FM-DNN is randomly initialized. Table I shows experimental results. System (1) had a character error rate (CER) of 74.6% because of a big acoustic mismatch between the acoustic and throat microphones. Furthermore, system (2) had a 96.6% CER because the DNN trained with the AM-MFCC cannot transform the TM-MFCC accurately into a feature space of the AM-BNF. Meanwhile, system (5) shows feature mapping could extract features that suppress the mismatch with the AM-BNF. The GMM-HMM trained with a large amount of the acoustic microphone's speech data has higher phoneme discrimination performance than the GMM-HMM trained with only about 3 hours of throat microphone data. As a result, the proposed system (5) had higher recognition accuracy than systems (3) and (4).

C-2. FEATURE MAPPING METHODS

To verify the effectiveness of the proposed feature mapping from the TM-MFCC to AM-BNF, we experimented with four mapping methods as shown in Table II. TM-BNF is extracted from DNN trained only with throat microphone.

In all methods, the architecture of hidden layers of the FM-DNN is the same as that shown in Fig. 2, and the weights of FM-DNN were randomly initialized. In speech recognition, the input feature is obtained from the mapped feature by applying CMN, LDA, and fMLLR as well as the TM-MFCC. When the feature input to the FM-DNN is the MFCC, the number of units of the input layer is 440 in order to input the feature obtained by splicing ±5 frames of the TM-MFCC. For the same reason, when the feature input to the FM-DNN is the BNF, the number of units of the input layer is 462. When the output feature is the MFCC, the supervised signal is a 13-dimentional MFCC of the acoustic microphone by applying CMN and hence the number of units of the output layer is 13. The acoustic model is the GMM-HMM trained with the AM-MFCC. Meanwhile, when the output feature is the BNF, the supervised signal is a 42-dimentional BNF of the acoustic microphone by applying CMN and hence the number of units

TABLE II
CHARACTER ERROR RATE (CER) FOR EACH METHOD OF FEATURE MAPPING FROM MFCC OR BNF OF THROAT MICROPHONE TO MFCC OR BNF OF ACOUSTIC MICROPHONE

Feature mapping method	CER [%]
(1) TM-MFCC to AM-MFCC	48.0
(2) TM-BNF to AM-MFCC	59.1
(3) TM-BNF to AM-BNF	51.1
(4) TM-MFCC to AM-BNF (Proposed)	42.2

TABLE III
CHARACTER ERROR RATE (CER) OF USING FEATURE ESTIMATED BY FM-DNN INITIALIZED FOR EACH METHOD

Initialization method	CER [%]
(1) Random	42.2
(2) TM-DNN	41.8
(3) AM-DNN (Proposed)	40.9

of output layer is 42. The acoustic model is the GMM-HMM trained with the AM-BNF. Table II shows experimental results. In the results, the proposed method (4) achieved a lower CER than the conventional feature mapping from the TM-MFCC to the MFCC of the acoustic microphone [6]. Methods (2) and (3) using the TM-BNF as the input signal had higher CERs than methods (1) and (4) using the TM-MFCC.

C-3. INITIALIZATION OF FM-DNN

To assess whether initializing the weights of the FM-DNN reduces CER, we experimented with three initialization methods.

(1) Random initialization (Random)
(2) Weights of DNN for extracting the BNF of the throat microphone (TM - DNN)
(3) Weights of DNN for extracting the BNF of the acoustic microphone (AM - DNN) (Proposed)

We conducted recognition experiments using features mapped by each DNN. The acoustic model is the GMM-HMM trained by the AM-BNF. Table III shows the experimental results. The accuracy of feature mapping was improved by initializing the weights of the FM-DNN using the weight of the DNN for extracting the BNF instead of initializing randomly. Finally, by using the proposed method (3), the CER was reduced to 40.9% by initializing with the weight of the DNN for extracting the AM-BNF.

From the above, the proposed method had a lower CER (40.9%) than the conventional methods. On the other hand, in the results of recognition of acoustic microphone speech data recorded simultaneously with the test data using the GMM-HMM trained with the AM-MFCC, the CER was 29.0%. When the test data with less external noise was used, the throat microphone still had inferior recognition accuracy to the acoustic microphone.

IV. CONCLUSION

In this study, we proposed DNN-based feature mapping from the MFCC of a throat microphone to the BNF of an acoustic microphone by using an acoustic model trained by a large amount of the BNF of an acoustic microphone. The proposed method performed much better than the acoustic models trained only with speech data of a throat microphone or an acoustic microphone. Moreover, our feature mapping method achieved higher recognition accuracy than the conventional feature mapping method that transforms the MFCC of the throat microphone into the MFCC of the acoustic microphone. We also found that feature mapping accuracy was improved by initializing the weights of the FM-DNN using the weight of the DNN for extracting the AM-BNF instead of initializing randomly. For future work, we will try to improve feature mapping accuracy and conduct experiments in noisy environments.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Numbers (16H01817) and (16K01543).

REFERENCES

[1] G. Saon et al., "English Conversational Telephone Speech Recognition by Humans and Machines," *Proc. Interspeech 2017*, pp. 132–136, 2017.
[2] T. Dekens, W. Verhelst, F. Capman, F. Beaugendre, "Improved Speech Recognition in Noisy Environments by Using a Throat Microphone for Accurate Voicing Detection," *Signal Processing Conference*, pp. 1978–1982, 2010.
[3] W. Amano, K. Noguchi, R. Takeda, K. Honma, "Automatic Speech Recognition Using Throat Microphone Under Highly-Noisy Environments," *journal of EICA*, pp. 182–186, 2014, in Japanese
[4] A. Shahina, B. Yegnanarayana, "Mapping Speech Spectra from Throat Microphone to Close-Speaking Microphone: A Neural Network Approach," *EURASIP Journal on Advances in Signal Processing,* vol. 2007, no. 2, pp. 1–10, 2007.
[5] K. Vijayan K. Sri Rama Murty, "Comparative Study of Spectral Mapping Techniques for Enhancement of Throat Microphone Speech," *Twentieth National Conference on Communications,* pp. 1–5, 2014.
[6] S. Lin, T. Tsunakawa, M. Nishida, M. Nishimura, "DNN-based Feature Transformation for Speech Recognition Using Throat Microphone," *APSIPA ASC 2017*, pp. 596–599, 2017.
[7] I. Himawan et al., "Learning Feature Mapping Using Deep Neural Network Bottleneck Features for Distant Large Vocabulary Speech Recognition," *ICASSP*, pp. 4540–4544, 2015.
[8] A. Das, M. Hasegawa-Johnson, "Cross-lingual Transfer Learning during Supervised Training in Low Resource Scenarios," *INTERSPEECH*, pp. 3531–3535, 2015.
[9] D. Povey et al., "The Kaldi Speech Recognition Toolkit," IEEE *2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society*, 2011
[10] Y. Miao, "Kaldi+PDNN: Building DNN-based ASR Systems with Kaldi and PDNN," arXiv:1401.6984, 2014.