DNN-based Feature Transformation for Speech Recognition Using Throat Microphone

Shengke Lin*, Takashi Tsunakawa*, Masafumi Nishida*, Masafumi Nishimura*
* Graduate School of Integrated Science and Technology, Shizuoka University, Shizuoka, Japan
E-mail: gs16050@s.inf.shizuoka.ac.jp
E-mail: {tuna, nishida, nisimura}@inf.shizuoka.ac.jp

***Abstract*—In this paper, we focus on utilizing a throat microphone as noise robust device because its signal is much less affected by surrounding noise than a conventional acoustic microphone signal. However, it can only record narrow frequency bands, and the microphone characteristics are also different from characteristics of acoustic microphone. Therefore, speech recognition performance is greatly degraded when a throat microphone is used as it is instead of a conventional acoustic microphone. To overcome this problem, we propose using a deep neural network (DNN)-based feature transformation method while also using model adaptation. We conducted a continuous digit recognition experiment. The result revealed that the proposed method improved the word error rate (WER) of using the throat microphone from 41.4% to 17.6%.**

I. Introduction

Speech recognition technology has become more and more widespread, but speech recognition performance needs to be further improved. One factor that degrades speech recognition performance is external noise, which causes problems such as failure of voice activity detection (VAD) and misrecognition. We focused on the throat microphone as a device that is barely affected by these external noises. In this application, a close-talk microphone and an array microphone are often used, but these have some problems. The close-talk microphone easily becomes worn out when worn for a long time, and the array microphone needs a wide installation place. Also, these microphones cannot prevent the mixing of noise completely. We therefore focus on a throat microphone, which can alleviate this problem. A throat microphone is a contact microphone that senses vibrations directly from the wearer's throat by way of sensors worn on the neck. It is robust against external noise because it barely picks up vibrations in the air.

There have been many studies [1]–[3] to improve the performance of speech recognition using throat microphones. One previous study [1] improved recognition accuracy by performing highly accurate VAD by using the signal of a throat microphone and performing speech recognition with an acoustic microphone signal. The reason the throat microphone is not directly used for speech recognition is that the acoustic mismatch degrades the accuracy of speech recognition in a general acoustic model assuming a usual microphone. A throat microphone's frequency characteristics are very different from those of the conventional acoustic microphones, so their acoustic mismatch leads to mis-recognition. Performance degradation due to mismatch must be prevented, as in the work of the [2]. However, it is not easy to prepare an enough training data for use in an acoustic model for large vocabulary speech recognition. Therefore, we try to reduce the mismatch with an existing acoustic model by mapping features of a throat microphone to features of an acoustic microphone in the feature space. Various methods [4]–[9] have been developed to expand bandwidth, and Gaussian mixture models (GMM) and artificial neutral networks (ANN) are often used. Kubota et al. [10] reported that a method combining k-means and a feedforward neural network (FFNN) [10] is very effective.

In contrast [10], we propose using long short-term memory (LSTM) that can handle time series data instead of an FFNN for feature transformation. We improved the recognition performance when using a throat microphone by training and transforming an independent deep neural network (DNN) in each feature dimension, because each dimension of mel-frequency cepstral coefficients (MFCCs) used as a feature has independent characteristics. Furthermore, we improve recognition performance without extra cost by adapting the acoustic model by using the transformed features as training data.

The remainder of the paper is as follows. Section 2 describes the basic layout of the system, the used corpus, feature transformation using DNN, feature transformation using k-means and DNN, and acoustic model adaptation. Section 3 describes the conditions and results of recognition performance experiments. Section 4 discusses conclusions and future prospects.

II. Method

A. Basic Layout of System

Fig. 1 shows the block diagram of our system. In the basic speech recognition process, VAD is performed on input speech, the features of an estimated speech section is extracted, and the speech is converted into text by inputting the extracted features to automatic speech recognition (ASR) software. In this paper, we improve recognition performance by first reducing the mismatch with the acoustic model by feature transformation using a DNN and then inputting the features extracted from the throat microphone to the ASR software. At this time, we try to improve recognition performance by classifying features of the input throat microphone by k-means and transforming with a DNN trained for each cluster. Also, we improve the recognition performance by adapting the acoustic model by using transformed features as adaptive data.

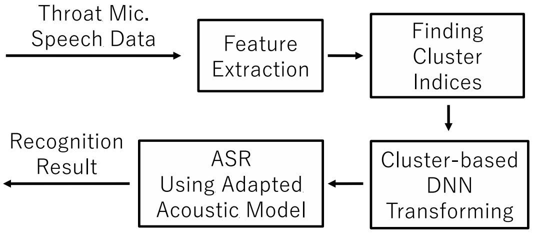

Fig. 1. DNN-based Feature Transformation

B. Corpus

We recorded parallel data of a throat microphone and an acoustic microphone for use in training of feature transformation. We use the same setup as Dupont et al. [11]. A recorder (ZOOM R24), throat microphone (Nanzu SH-12iK), and acoustic microphone (Sony ECM - CS3) were used. The audio sampling rate was 16,000 Hz.

We recorded 3600 (220 min.) phoneme-balancing sentences of Japanese by 14 male speakers as training data and 1000 sentences of 11 consecutive digits by 10 male speakers as test data in soundproof room.

C. Feature Transformation using DNN

Fig.2 shows the network configuration of the DNN used for the transformation. For comparison, we also show the composition of FFNN with reference to a previous study [4].

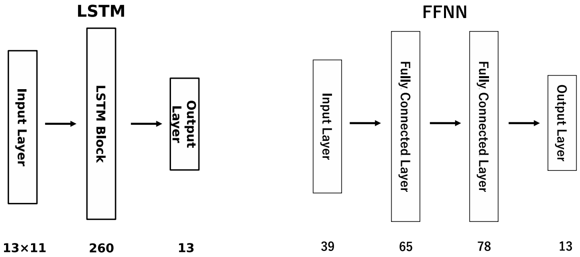

Fig. 2. Network structures of LSTM and FFNN

A 13-dimensional MFCC is used for input data and training data of LSTM. The input data are time series data of 11 frames combined with the preceding 10 frames. In total, 39 dimensions of a feature, which combines MFCC and Δ and $\Delta\Delta$ parameters, are used as input data in FFNN to consider the temporal change of speech at the transformation. The supervised data are a 13-dimensional MFCC.

We examine three training and transformation methods, considering that the low frequency of a low order MFCC signal and high frequency of a high order signal affect training and transformation of a DNN. Method 1 is a general method of training and transforming the network shown in Fig.3 by preparing n-dimensional input data and m-dimensional reference data as shown in Fig.2. Method 2 is to train and transform by preparing an internal network for each dimension with respect to the m-dimensional feature to be obtained as an output as shown in Fig. 4. The entire network is shown in Fig.2, in which the size of the output layer is changed to 1. We merge the internal networks with the merge layer and train with n-dimensional input data and m-dimensional reference data. Method 3 has the same network structure as method 2 but a different training method. It does not train the entire network but trains the internal network individually and restores trained parameters into the network in Fig.4. Training is performed by inputting the same n-dimensional input features and 1-dimensional reference data to each internal network.

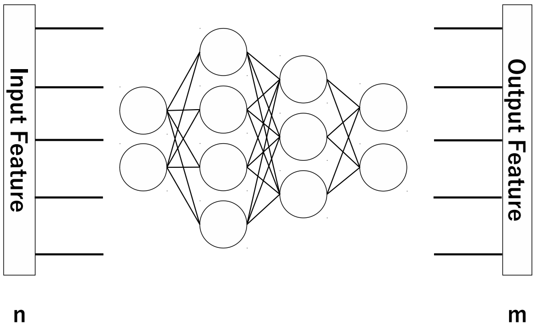

Fig. 3. Network structure of Method 1

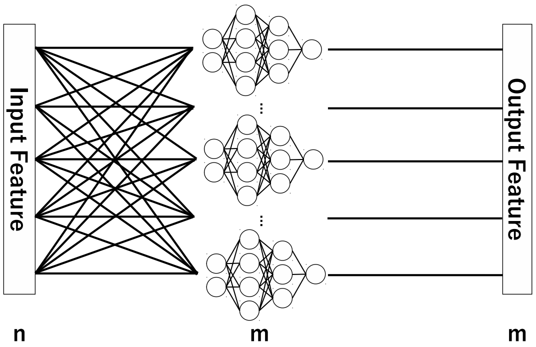

Fig. 4. Network structure of Methods 2 and 3
Training methods differ between 2 and 3.

D. Feature transformation using k-means and DNN

A method was previously developed to transform the features of the throat microphone by combining k-means and FFNN [5]. However, in this paper, we evaluate and compare the case of combining k-means and LSTM proposed in II-C. The bandwidth expansion consists of training and transformation, and the process is shown in Fig. 5 .

During training, all training data are used at the beginning to train one NN. Then, training data clusters are trained and classified by k-means. Next, as many NNs as k-means clusters are prepared and fine-tuned by using the training data classifying NN trained at the beginning. During transformation, input features are classified by k-means trained with training data, and each corresponding NN is transformed.

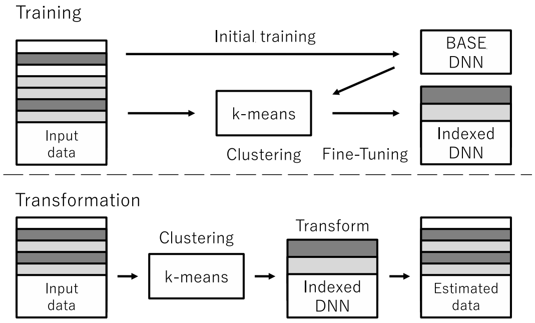

Fig. 5. Training and transformation method when k-means and DNN are combined

In the case of a combination of k-means and FFNN, the features use $MFCC + \Delta + \Delta\Delta$. However, in the case of a combination of k-means and LSTM, k-means uses $MFCC + \Delta + \Delta\Delta$ as features, but LSTM uses time series data of MFCC without the Δ parameter. The flow of training is shown in Fig. 6. The 13-dimensional MFCC feature sequence is x, and the n-th frame of the feature sequence is x_n. k-means trains $x + \Delta x + \Delta\Delta x$ as features, and LSTM trains $(x_n, x_{n-1}, x_{n-2}, \cdots, x_{n-10})$ as one piece of data. $x + \Delta x + \Delta\Delta x$ is calculated from x_n and classified with k-means. After classification, each LSTM is fine-tuned with data of $(x_n, x_{n-1}, x_{n-2}, \cdots, x_{n-10})$.

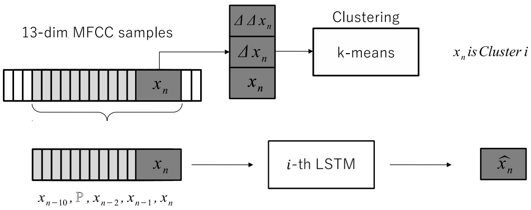

Fig. 6. Feature flow when k-means and LSTM are combined

E. Acoustic Model Adaptation

Fig. 7 shows the 1st and 9th dimensions of the MFCC features transformed by LSTM. As shown in the upper part of Fig. 7 , the 1st to 5th order features of the transformed MFCC are considerably close to the features of the acoustic microphone, and the transformation worked well. On the other hand, as shown in the lower part of Fig. 7 , the output features tended to smooth along the time scale at the 6th or higher orders, and features slightly different from those of the acoustic microphone were observed.

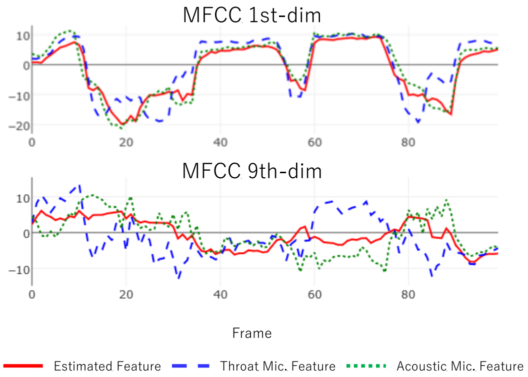

Fig. 7. Feature transformed by LSTM

For this reason, acoustic mismatch with a general acoustic model was less than before transformation, but it still existed. Therefore, the acoustic model was adapted using the transformed features as adaptation data. The mean vector of acoustic model was adapted only using maximum likelihood linear regression (MLLR). First, after global MLLR adaptation with clustering number 1, the clustering number was changed to 32 and further adaptation was done with MLLR.

III. Experiments

In this section, we describe the experimental conditions and the results of the recognition performance comparative experiment conducted using the continuous digit recognition task.

A. Experimental condition

Consecutive digit recognition was performed on Julius, and a large vocabulary acoustic model was created from the ASJ-JNAS corpus. The language model is a digit recognition grammar with an indefinite number of digits. However, to eliminate the effect of speaker adaptation, the test data and training data contained some identical speakers, and cross validation was performed excluding the same speaker. Therefore, on average, the training data is 3000 sentences (200 minutes). The MFCC features were extracted with a frame length of 25 ms and a shift length of 10 ms, and cepstral mean normalization (CMN) was applied.

The optimization algorithm of the DNN is AdaDelta, and the activation function uses identity mapping. The mini batch size at training was 4096, the initial learning rate was 0.01, and the number of iterations was 100. The number of clusters of k-means is set to 10 as in the previous study [5]. Also, although the feature transformed by the DNN is 13 dimensions, the features of the acoustic model is 25 dimensions of 12th order MFCC and its Δ parameter and power. Therefore, the Δ parameter was newly extracted for the features after the transformation, and the adaptation of the acoustic model and speech recognition were performed. Only the mean vector of acoustic model was adapted with supervised MLLR.

TensorFlow is used for a training of DNN, scikit-learn is used for a training of k-means, and features are extracted and the acoustic model is adapted by using the Hidden Markov Model Toolkit (HTK).

B. Experimental results

Table I shows the results of executing the continuous digit recognition task for each DNN training method mentioned in II-C. As shown in the Table I , Method 2 has a superior performance to Method 1. It is speculated that preparing an internal network and separating the network parameters for each dimension may enable the feature of the dimension to be captured successfully. Moreover, Method 3 performs superiorly to Method 2. Since these two network structures are identical, it is presumed that propagation of training works well by training networks in feature dimensions individually. In this paper, we adopt Method 3, which performs the best, and the results for Method 3 are shown in the feature transformation by DNNs unless otherwise stated.

TABLE I
Results of continuous digit recognition task for each training method.

Training Method	LSTM WER	FFNN WER
Method 1	30.3%	43.9%
Method 2	25.1%	41.4%
Method 3	24.0%	41.3%

Table II shows the results of adapting the acoustic model by using MLLR. The recognition performance was better when LSTM and MLLR were combined than when only MLLR is applied to the throat microphone. Although the performance is further improved by combining the k-means, the training method based on the LSTM is better for suppressing the calculation cost because k-means requires a high calculation cost.

TABLE II
Result of continuous digit recognition task for each bandwidth expansion method

Microphone. (transformation method)	WER (not adapted)	WER (MLLR adapted)
AM[a] (BASELINE)	4.4%	None
TM[b] (BASELINE)	41.4%	23.9%
TM[b] (FFNN)	41.3%	33.6%
TM[b] (k-means + FFNN)	29.0%	24.5%
TM[b] (LSTM)	24.0%	20.5%
TM[b] (k-means + LSTM)	21.9%	17.6%

[a] Acoustic Microphone
[b] Throat Microphone

IV. Conclusions

We presented a method to transform features of a throat microphone into features of an acoustic microphone with a deep neural network (DNN) using long short-term memory (LSTM). We found that recognition accuracy was improved by using an internal network for each feature dimension and individually these networks rather than using a network trained by using all the features collectively. Experimental results showed that the proposed method can achieve a higher recognition accuracy than a simpler method of adapting an acoustic model. For future work, we will conduct experiments in noisy environments and attempt to improve the feature transformation performance.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Numbers (16K01543, 16H01817, 16K13028) and Telecommunications Advancement Foundation.

References

[1] Tomas Dekens, Werner Verhelst, François Capman, Frédéric Beaugendre, “Improved speech Recognition in Noisy Environments by Using a Throat Microphone for Accurate Voicing Detection,” *Signal Processing Conference,* pp. 1978-1982, 2010.

[2] Martin Graciarena, Horacio Franco, Greg Myers, Cregg Cowan, Federico Cesari et al., “Combination of Standard and Throat Microphones for Robust Speech Recognition in Highly Noisy Environments,” *Interspeech,* pp. 1-4, 2004.

[3] MA Tuğtekin Turan, Engin Erzin, “Source and Filter Estimation for Throat-Microphone Speech Enhancement,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 24, no. 2, pp. 265-275, 2016.

[4] A. Shahina and B. Yegnanarayana, “Mapping Speech Spectra from Throat Microphone to Close-Speaking Microphone: A Neural Network Approach,” *EURASIP Journal on Advances in Signal Processin,* vol. 2007, no. 2, pp. 1-10, 2007.

[5] Karthika 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] Md Sahidullah, Rosa Gonzalez Hautamäki, Dennis Alexander Lehmann Thomsen, Tomi Kinnunen, Zheng-Hua Tan, Ville Hautamäki, Robert Parts, Martti Pitkänen, “Robust Speaker Recognition with Combined Use of Acoustic and Throat Microphone Speech,” *Interspeech,* pp. 1720-1724, 2016.

[7] Anuradha S Nigade, JS Chitode, “Throat Microphone Signals for Isolated Word Recognition Using LPC,” *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 2, no. 8, pp. 401-407, 2012.

[8] Ho Seon Shin, Hong-Goo Kang, Tim Fingscheidt, “Survey of Speech Enhancement Supported by a Bone Conduction Microphone,” *Proceedings of 10. ITG Symposium Speech Communication,* pp. 1-4, 2012.

[9] Srinivas Desai, E. Veera Raghavendra, B. Yegnanarayana, Alan W Black, Kishore Prahallad, “Voice Conversion Using Artificial Neural Networks,” *IEEE International Conference on Acoustics, Speech and Signal Processing,* pp. 3893-3896, 2009.

[10] Ryosuke Kubota, Eiji Uchino, Noriaki Suetake, “Tone Quality Improvement of Bone Conduction Voice by Using Neural Gas Network and Local Conversion Models,” *Fuzzy System Symposium,* pp. 112-113, 2009.

[11] Stéphane Dupont, Christophe Ris, Damien Bachelart, “Combined Use of Close-Talk and Throat Microphones for Improved Speech Recognition Under Non-Stationary Background Noise,” *COST278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction,* pp. 1-4, 2004.