Deep Neural Network-Based Speech Recognition with Combination of Speaker-Class Models

Tetsuo Kosaka*, Kazuki Konno* and Masaharu Kato*
*Graduate School of Science and Engineering, Yamagata University, Yonezawa, Japan
E-mail: {tkosaka@yz, tyh19923@st, katoh@yz}.yamagata-u.ac.jp

***Abstract*—This paper proposes a new speech recognition method based on speaker-class (SC) models. In previous studies based on this approach, Gaussian-mixture-model-based hidden Markov models (GMM-HMMs) have mainly been used as acoustic models. In this work, SC models that have deep neural network (DNN)-based HMM (DNN-HMM) structures are investigated and used for speaker-independent (SI) speech recognition. To realize SI speech recognition based on SC models, technological challenges must be solved so that unsupervised adaptation can be performed with only one utterance. To address this problem, we propose a new method of combining DNN outputs. In our experiments, five of 963 SC models were selected automatically, and DNN-HMM-based SC models were combined for each utterance. The results showed that the proposed method outperformed a baseline DNN-HMM system.**

I. Introduction

Recently, deep neural network (DNN)-based speech recognition has received high attention for its performance in large-vocabulary continuous speech recognition. However, the variety of speaker characteristics remains an open issue. To solve this problem, some speaker adaptation techniques have been proposed [1]–[3]. Techniques of speaker adaptation are helpful in terms of recognition performance; however, they require the extra effort that adaptation data be provided in advance. The goal of this work is to improve DNN-based recognition performance using speaker adaptation techniques without requiring adaptation data. Figure 1 shows the basic idea of unsupervised adaptation. Adaptation data are recognized once to derive recognized strings. A speaker-independent (SI) model is then adapted using adaptation data and recognized strings. Finally, an input utterance is recognized with the adapted model. The basic idea of the proposed method is shown in Figure 2. The two block diagrams are similar; however, an input utterance is recognized twice in the latter method and no adaptation data are required. The difference is that in the second block diagram, adaptation is accomplished by using the input utterance itself.

To realize SI speech recognition based on speaker adapted models, technological challenges must be solved so that unsupervised adaptation can be performed with only one utterance. In general, adaptation methods of DNN require some quantity of adaptation data because the number of adapted parameters is high. For example, 10 min of adaptation data were required for unsupervised adaptation in [1]. Conversely, only a few seconds of data can be used in the proposed scheme. For example, utterances with an average duration of 2.27 s are used as test data in our experiments.

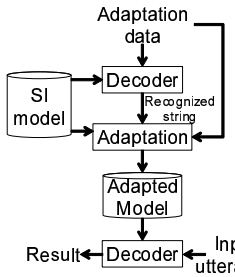

Fig. 1. *Block diagram of unsupervised speaker adaptation.*

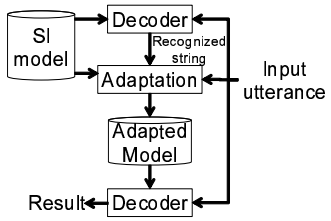

Fig. 2. *Block diagram of speaker independent speech recognition based on adaptation technique.*

To solve this problem, we utilize a recognition technique that uses a speaker-class (SC) model. The basic idea of this technique is that speakers near test speakers are selected from training speakers and those data are used to create an SC model. The techniques of SC-based speech recognition can be divided into two categories. One of the typical methods is to select cohort speakers for each evaluation speaker using adaptation data before recognition processing (e.g. [4], [5]). On the other hand, techniques of speaker-independent speech recognition using SC models have been proposed (e.g. [6], [7]). In the latter techniques, all speakers in the training data are clustered into speaker classes independent of the test speaker in the training step. In the recognition step, the most appropriate SC model is selected utterance by utterance and used for recognition. These works have mainly used Gaussian-mixture-model-based hidden Markov models (GMM-HMMs). Recently, Mimura et al. proposed a DNN-HMM-based adaptation method that was categorized as the former approach [8].

We want to improve the recognition performance of DNN-HMM-based systems by using the latter approach. The simplest technique based on the latter approach involves the acoustically closest SC model to an input utterance being selected on the basis of a likelihood criterion. However, one

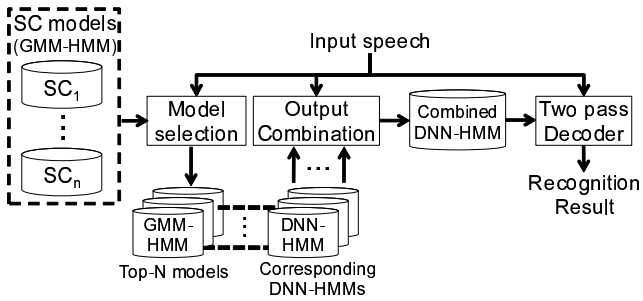

Fig. 3. *Block diagram of the proposed recognition system.*

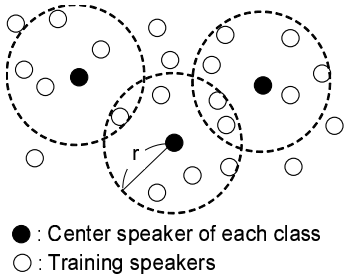

Fig. 4. *Conceptual diagram of soft clustering.*

problem is that a suitable model is not always selected, especially when the utterance is very short. In our previous work, a 20.48% word error rate (WER) was obtained with the single model selection scheme using the GMM-HMM-based SC model, whereas a 12.40% WER was obtained in the ideal condition in which the best SC model was selected manually [9]. With careful attention to the likelihood values of SC models, an SC model with the best performance did not always demonstrate the highest likelihood; however, it was usually ranked relatively high in the likelihood list. In that context, we propose a method for output combination of SC-based DNNs. In this method, the observation probabilities of DNNs are merged by using weight factors. The proposed method is evaluated with the Corpus of Spontaneous Japanese (CSJ) task.

The remainder of this paper is organized as follows: Sec. II introduces the proposed speech recognition technique using SC models. Section III describes the conditions of the speech recognition experiments and the conditions of the SC modeling. Section IV describes the results of the speech recognition experiments. Section V provides our conclusions.

II. Speech recognition using speaker-class models

A. Overview

Figure 3 shows a block diagram of the proposed recognition system. The basic idea is that the N best SC models are selected based on a likelihood criterion and are used for model combination. However, it is difficult to prepare many SC models that have a DNN-HMM structure in advance because this requires a massive amount of calculation time. To save on calculation time, SC models with a GMM-HMM structure are used for model selection instead of DNN-HMMs. From the preliminary experiment, we confirmed that likelihood values obtained from GMM-HMM and DNN-HMM achieve high correlation. For this reason, GMM-HMMs can be used for model selection. After selecting the top N GMM-HMMs, the corresponding DNN-HMMs are selected. In this case, the corresponding two models are trained using the same training speakers. The selected DNN-HMMs are combined by the proposed combination algorithm. Finally, input speech is recognized by using the combined DNN-HMM.

B. Speaker-class model

In the proposed system, speaker-class models based on both GMM-HMM and DNN-HMM structures are used. The former is used for model selection, and the latter for recognition. The same clustering algorithm is used for both. The algorithm is based on soft clustering in which data elements can belong to more than one cluster, as proposed in [9]. This algorithm is a modified version of the hard clustering proposed in [10]. The merit of the algorithm in [10] is that no initial parameter except for the number of clusters is needed. We apply this algorithm to the soft-clustering method. In soft clustering, only a cluster radius and the number of clusters are required as initial parameters.

First, a speaker-dependent (SD) model is prepared for each training speaker to measure the similarity between training speakers. All SD models are clustered, and the clustering result is used to create SC models. In the algorithm, the cluster with the maximum sum of distances is divided step by step. Distances between pairs of SD models are calculated in advance to prepare a distance table that can reduce the calculation cost. Based on the results of the aforementioned clustering, a center speaker is calculated for each cluster. The center speaker is determined by measuring the sum of distances from each speaker belonging to the cluster and taking the minimum. Speakers within a predetermined radius of the center speaker are regarded as members of the cluster. The concept of clustering is shown in Fig. 4. Using the above algorithm, some speakers will be assigned to more than one cluster.

In the algorithm, the distance between SD models must be calculated. GMM-HMM is used for the structure of the SD model. The distance between two HMMs M_1 and M_2 with the same structure is defined as follows:

$$D(M_1, M_2) \triangleq \frac{1}{NM} \sum_{i=1}^{N} \sum_{m=1}^{M} d(b^1_{im}, b^2_{ig(m)}), \qquad (1)$$

where N is the number of states, M is the number of mixture components, and b_{im} is the observation probability at state i and mixture component m. Note that $g(m)$ is the mixture permutation function that minimizes the value of the distance. Transition probability parameters are omitted from the distance calculation. The SI model is used as the initial model of each SD model. Therefore, two mixture components that belong to different SD models but have the same mixture

component number and state will possess similar acoustic features. Because of this, we assume that

$$g(m) = m. \tag{2}$$

The Bhattacharyya distance measure is employed to calculate the distance d. This measure is symmetric and is guaranteed to be nonnegative.

C. Model selection

The top N SC models that are acoustically close to input speech are selected from a large number of SC models on the basis of a likelihood criterion. The model selection algorithm is as follows:

1) Decoding processes are conducted using multiple SC GMM-HMMs for each utterance.
2) From the results of the above step, the N best SC GMM-HMMs are selected on the basis of a likelihood criterion.
3) The corresponding DNN-HMMs are selected and used for output combination.

D. Output combination

We propose a combination algorithm of outputs from multiple DNNs. The observation probability of DNN-HMM is calculated as

$$p(\boldsymbol{x}|s_i) = \frac{p(s_i|\boldsymbol{x})p(\boldsymbol{x})}{p(s_i)}, \tag{3}$$

where $p(s_i|\boldsymbol{x})$ is the state posterior probability estimated from the DNN, $p(s_i)$ is the state prior probability, and $p(\boldsymbol{x})$ is the prior probability of input features and can be ignored. In the proposed technique, multiple observation probabilities are combined by weighting factors. Assume that $p^m(\boldsymbol{x}|s_i)$ is the observation probability of the m-th DNN. The probability after combination is then calculated as

$$b_i(\boldsymbol{x}) = \sum_m w_{im} p^m(\boldsymbol{x}|s_i). \tag{4}$$

In addition, weights can be tied across states to reduce the number of parameters.

In our experiments, two weighting methods, maximum-likelihood (ML) estimation-based weighting and equal weighting, are tested. In the former method, the weight factor estimation is carried out for each input utterance. For ML estimation, a phoneme symbol string must be prepared for each utterance. To obtain the phoneme string, a decoding process is accomplished by using an SI model. Note that the phoneme string contains some errors.

E. Decoding process

A two-pass decoder is used for recognition. A one-pass algorithm that involves a frame-synchronous beam search is adopted in the first pass. The search algorithm calculates the acoustic and language likelihoods to obtain a word graph. The abovementioned DNN-HMMs in which the observation probabilities are combined are used as the acoustic models in the first pass. Moreover, a bigram is used as the language model. Once the word graph is obtained, rescoring processes are conducted during the second pass. A trigram is used as the language model in this step.

III. Experimental set up

A. Recognition system

In this section, we describe our recognition system. In the speech analysis module, a speech signal is digitized at a sampling frequency of 16 kHz with a quantization size of 16 bits. The length of the analysis frame is 25 ms, and the frame period is set to 8 ms. A 25-dimensional feature, which consists of the log mel-filter bank (FBANK) features and the log power, is derived from the digitized samples for each frame. Moreover, the delta and delta-delta features are calculated from the 25-dimensional feature, so the total number of dimensions is 75 per frame. The input layer of the DNN uses 75 coefficients with a temporal context of 11 frames, summing to a total of 825 input features. The DNN has seven hidden layers with 2048 hidden units in each layer. The final output layer has 3003 units, corresponding to the total number of HMM states. The bigram and trigram models are trained on textual data containing 2668 lectures from the CSJ, and the total number of words is 6.68M. We used an evaluation set (testset1) consisting of academic presentations given by ten male speakers. This is one of the standard test sets in the CSJ.

B. Speaker-class model

The CSJ is used to train the SI and SC models. The total number of lectures used for training is 963. Each lecture is given by one speaker. Therefore, the total number of speakers is also 963. Note that some speakers gave several lectures. The total speech length is approximately 203 h. The SI model is a set of shared-state triphones with 3003 tied states. For speaker clustering, SD monophonic GMM-HMMs are used for measuring the distance between training speakers. The SD model is then trained for each training speaker in advance. The model structure is a left-to-right HMM with three states, and the number of mixture components is 12. The 963 SD models are clustered by the algorithm described in Sec. II. From the results of the GMM-HMM-based SC models, a growing number of SC models (more than 300) led to an improvement in recognition performance [9]. For this reason, we set the number of speaker classes to 963 (i.e. equivalent to the number of training speakers), and the speaker radius for soft clustering is set to 180. After the speaker-clustering step is completed, SC models are trained using an SI model as the initial model. The structure of each SC model is the same as that of the SI model.

IV. Experimental results

To clarify the effects of the SC model itself, we conducted preliminary experiments in which the output combination process was omitted. In these experiments, only one SC model was selected from the 963 SC models and was used for recognition. Table I shows WERs of these experiments. In this table, *baseline* represents the results of the baseline SI model. *Utterance* means that SC model selection was done for every input utterance. For comparison, the results of model selection for every evaluation speaker are indicated (*speaker*). The average number of utterances per speaker was 122.

TABLE I
RECOGNITION RESULTS IN WER(%) OF SINGLE SC MODEL SELECTION.

baseline	utterance	speaker
15.12	15.01	14.86

The results of model selection show better performance than the results of the baseline. The same SC model was not always selected for each utterance even if the evaluation speaker was the same. We considered that there is a variation of acoustic features for each utterance. For the case of *speaker*, the best results could be obtained. This means that it is slightly difficult to select a suitable SC model by using one short utterance.

Table II shows the recognition results of the proposed methods. In these experiments, two types of model selection were applied. In *5SC*, that the top five models were selected for each utterance by a likelihood criterion, and those models were combined by the proposed output combination algorithm. In *5SC+SI*, five SC models and the SI model were used (six models in total). For the method of *weight tying*, weights were tied across states. In addition, we tested the effects of transition probability estimation by an ML criterion, in which all transition probabilities of the combined DNN-HMM were estimated for each utterance.

In *5SC* experiments, the best performance (14.91%) could be obtained without weight tying and transition estimation. This result is better than that of the single-SC-model selection shown in Table I. The method of single-model selection strongly depends on the performance of model selection. From the results of the comparison between *utterance* and *speaker* shown in Table I, it is considered that the selection performance by the likelihood criterion is insufficient. Conversely, the proposed method utilizes the outputs of multiple SC models. This helps relieve the problem of model selection. Compared between *5SC* and *5SC+SI*, the latter showed the better performance. Because the SI model covers a wide range of speaker characteristics, it is also expected to relieve the problem of model selection. Finally, 14.87% as the best performance could be obtained without weight tying and transition estimation. To determine if the weight factor estimation by ML is effective, we tested the equal weighting method where all weights are set equally in this condition. The WER of this experiment was 14.89%. The difference between the ML estimation and the equal weighting method is small.

We summarize the experimental results in Table III. According to the sign test, the difference between *baseline* and *single SC selection* is not statistically significant; however, both proposed methods are significant at the level of 5%.

The proposed method takes 1.24 times the calculation cost of the single model selection method except for the cost of top-N model selection. This is because forward calculation for DNNs is performed on graphics processing units (GPUs) and the percentage of forward calculation against total calculation cost is only 3.4% for each DNN. In contrast, a huge amount of calculation time is needed for top-N model selection in the current implementation because a decoding process is carried out for likelihood calculation. There are some methods to save model selection time as follows: 1) using GMMs instead of GMM-HMMs, 2) reducing the number of speaker classes, and 3) use of tree-structured speaker clustering [6]. We will try to reduce the calculation cost in the future.

TABLE II
RECOGNITION RESULTS IN WER(%) OF COMBINATION METHODS.

Type	WER (%)			
5SC	15.04	14.99	15.15	14.91
5SC+SI	15.05	15.00	15.13	**14.87**
Transition	yes	no	yes	no
Weight tying	yes	yes	no	no

TABLE III
SUMMARY OF RECOGNITION EXPERIMENTS.

Type	baseline	single SC selection	proposed (equal weights)	proposed (ML estimation)
WER (%)	15.12	15.01	14.89	14.87

V. CONCLUSIONS

In this paper, we investigated DNN-HMM-based speech recognition using SC models. In the proposed method, the top N of 963 SC models were selected for each utterance by a likelihood criterion, and the N outputs of DNNs were merged to be used for the observation probability of DNN-HMM. In the experiments, five SC models or five SC models and one SI model were combined. The proposed method showed significant improvement over the baseline. In contrast, the single SC selection scheme could not achieve significant improvement.

ACKNOWLEDGMENT

This work was supported in part by a Grant-in-Aid for Scientific Research (KAKENHI 25330183) from the Japan Society for the Promotion of Science.

REFERENCES

[1] H. Liao, "Speaker adaptation of context dependent deep neural networks," in Proc. of ICASSP2013, 2013, pp. 7947–7950.
[2] D.Yu, K.Yao, H.Su, G.Li and F.Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition, " in Proc. of ICASSP2013, 2013, pp. 7893–7897.
[3] S. Xue, O. A.-Hamid, H. Jiang, and L. Dai, "Direct adaptation of hybrid DNN/HMM model for fast speaker adaptation in LVCSR based on speaker code, " in Proc. of ICASSP2014, 2014, pp. 6339–6343.
[4] M.Padmanabhan, L.R. Bahl, D.Nahamoo, and M.Picheny, "Speaker clustering and transformation for speaker adaptation in speech recognition systems, " Trans. on Speech and Audio Proc., vol. 6, no. 1, pp. 71–77, 1998.
[5] T. Kosaka, T. Ito, M. Kato, and M. Kohda, "Speaker adaptation based on system combination using speaker-class models, " in Proc. of Interspeech2010, 2010, pp. 546–549.
[6] T. Kosaka, S. Matsunaga, and S. Sagayama, "Speker-independent speech recognition based on tree-structured speaker clustering, " Computer Speech and Language, vol. 10, no. 1, pp. 55–74, 1996.
[7] Y.Zhang, J.Xu, Z.-J. Yan, and Q. Huo, "An i-vector approach to training data clustering for improved speech recognition, " in Proc. of Interspeech2011, 2011, pp. 789-792.
[8] M. Mimura and T. Kawahara, "Unsupervised speaker adaptation of DNN-HMM by selecting similar speakers for lecture transcription, " in Proc. of APSIPA2014, 2014, pp. 1–4.
[9] K. Konno, M. Kato, and T. Kosaka, "Speech recognition with large-scale speaker-class-based acoustic modeling," in Proc. of APSIPA2013, 2013, pp. 1–4.
[10] N. Sugamura, K. Shikano, and S. Furui, "Isolated word recognition using phoneme-like templates, " in Proc. of ICASSP83, 1983, pp. 723-726.