Speech Recognition with Large-Scale Speaker-Class-Based Acoustic Modeling

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

***Abstract*—This paper investigates speaker-independent speech recognition with speaker-class models. In previous studies based on this method, the number of speaker classes was relatively small and it was difficult to improve the performance significantly over the baseline. In this work, as many as 500 speaker-class models are used to enable more precise modeling of speaker characteristics. In order to avoid a lack of training data for each speaker-class model, a soft clustering technique is used in which a training speaker is allowed to belong to several classes. In the recognition experiments, a slight improvement in performance was obtained using a conventional method with several tens of speaker-class models. In contrast, a significant improvement was obtained using an unsupervised soft clustering method with several hundred speaker-class models. In addition, the results indicated a possibility of reducing the error rate drastically if the speaker-class model selection was conducted more effectively.**

I. Introduction

The variety in speaker characteristics is one of the major problems for speaker-independent speech recognition. Many techniques have been proposed to solve this problem. For instance, the use of a speaker-class (SC) model has been proposed. The techniques of SC-based speech recognition can be divided in to two categories. One of the typical methods is to select cohort speakers for each evaluation speaker using adaptation data before recognition processing [1][2][3][4]. The data on the selected speakers are used to create an SC model. Since this technique requires adaptation data, it is considered a certain type of speaker adaptation. On the other hand, techniques of speaker-independent speech recognition using SC models have been proposed [5][6][7]. In these 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. Such techniques are considered speaker-independent (SI) speech recognition because adaptation data are not required. These will be referred to as speaker clustering techniques in this paper. Comparing the two methods, the speaker clustering method is the more useful, because adaptation data are not required. We focus on this technique in the present work.

For recognition tasks involving speakers with a broad range of ages, the speaker clustering technique has proven useful. Enami et al. [7] showed that using a system based on speaker clustering improved the recognition performance for speech corpora including three generations of speakers (child, adult, and elder) aged between 6 and 90 years. In contrast, it was difficult to improve the performance significantly over that of a SI system for adult-only speech data, because acoustic characteristics are to some extent similar among speakers of the same age.

TABLE I
Comparison of speaker-class-based methods

Previous work	Type	Hard or Soft	Number of classes
Proposed	Speaker clustering	hard / soft	500
[7]	Speaker clustering	soft	30
[6]	Speaker clustering	hard	16
[5]	Speaker clustering	hard	170
[1][2][3][4]	Cohort speakers	-	-

Increasing the number of SC models is one of the solutions to the problems of this method. In previous studies of speaker clustering methods, the number of speaker classes was relatively small. Table I lists the SC-based methods. Some large-scale speech corpora representing a total of more than 1000 speakers have been developed. In Japan, a spontaneous speech database known as the Corpus of Spontaneous Japanese (CSJ) is available. Thus, we can conduct large-scale SC-based acoustic modeling using such large amounts of training data. In this work, we attempt to use several hundred SC models created using an unsupervised speaker clustering technique. Increasing the number of SC models is expected to enable more precise modeling of speaker characteristics.

Although the number of SC models can be increased using a large-scale speech corpus, excessively increasing the number of models causes a lack of training data for each SC model. Hence, a soft clustering technique was proposed [7] in which a training set was allowed to become associated with several classes. A similar approach was proposed by Jouvet et al. [8]. In this paper, a soft clustering approach is used to avoid the lack of training data when increasing the number of classes. Although the number of SC models was relatively large in [5], the performance gain was not as significant because the technique was based on hard clustering in which each training speaker belongs to a single class.

In order to improve the recognition performance using large-scale SC modeling, the soft clustering approach is employed in this work. In addition, since no comparison in performance between soft and hard clustering was made in [7], the effectiveness of soft clustering has not been clear. We investigate the effectiveness of soft clustering approach by comparing the

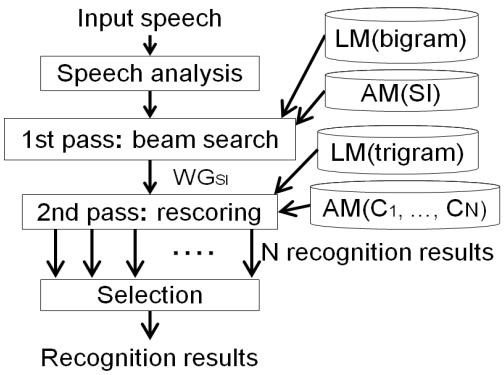

Fig. 1. *Block diagram of the proposed recognition system.*

two approaches.

The proposed method is evaluated with the CSJ task. In the experiments, various numbers of SC models and various class sizes are tested. In addition, a comparison between the hard and soft clustering techniques is made.

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

II. Speech recognition using speaker-class models

A. Overview

Fig. 1 shows a block diagram of the proposed recognition system. In the proposed system, the calculation cost is large because a decoding process must be conducted many times. In order to mitigate this problem, a two-pass decoder is used. 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. This calculation cost is much larger than that of the second pass. Therefore, only a speaker-independent (SI) model is used as the acoustic model in the first pass. Moreover, a bigram is used as the language model. Once the single word graph is obtained, rescoring processes are conducted using multiple SC models ($C_1, \ldots, C_N$) in the second pass. A trigram is used as the language model in this step. Thus, multiple recognition results are obtained for each utterance. A suitable recognition result is selected for each utterance based on a likelihood criterion.

B. Speaker clustering method

In this work, we utilize hard and soft clustering methods for creating SC models. The proposed soft clustering technique is based on the Bhattacharyya distance measure and is a modified version of the hard clustering proposed in [5]. We now describe the algorithms of the hard and soft clustering methods. First, a speaker-dependent (SD) model set is prepared for each training speaker to measure similarity between training speakers. All SD model sets are clustered and the clustering result is used for creating SC models. The hard clustering algorithm used was originally proposed in [9]. The merit of this algorithm is that no initial parameter except the number of clusters is needed. This algorithm has been successfully applied to tree-structured speaker clustering [5]. We apply this algorithm in the soft clustering method. In the proposed soft clustering, only a cluster radius and the number of clusters are required as initial parameters.

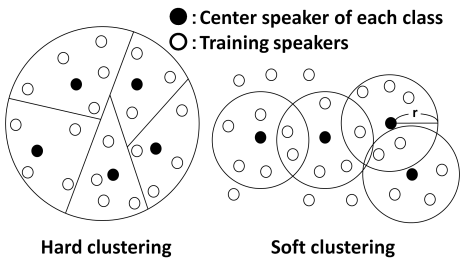

Fig. 2. *Conceptual diagram of hard and soft clustering.*

In the hard clustering 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. The details of this algorithm are given in [5]. The procedure of the soft clustering is as follows. Based on the results of the aforementioned hard 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 the clustering is shown in Fig. 2. Using the above algorithm, some speakers will be assigned to more than one cluster. Note that some speakers may not be assigned to any cluster. The problem of this point is described in Sec. IV.

C. Distance between speaker models

As described in the previous section, a distance between SD models must be calculated in the clustering algorithm. The distance between two hidden Markov models (HMMs) M_1 and M_2 with the same structure is defined as follows [5]:

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

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 number 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

$$g(m) = m. \qquad (2)$$

The Bhattacharyya distance measure is employed to calculate the distance d. This measure is symmetric and is guaranteed not to be negative.

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 and 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 13-dimensional feature, which consists of the 12 Mel-frequency cepstral coefficients (MFCCs) and the log power, is derived from the digitized samples for each frame. Moreover, the delta and delta-delta features are calculated from the MFCCs and the log power, so the total number of dimensions is 39. The 39-dimensional parameters are normalized by the cepstral mean normalization (CMN) method. A two-pass search decoder with a bigram and trigram is used for recognition (see Fig. 1). In the first pass, a word graph is generated using an SI model set and the bigram language model. Decoding is performed using a one-pass algorithm that involves a frame-synchronous beam search and a tree-structured lexicon. In the second pass, SC model sets and the trigram language model are applied to rescore the word graph, and thus, multiple recognition results are obtained. A suitable result is selected for each utterance based on the likelihood criterion. 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) that consists of academic presentations given by 10 male speakers. This is one of the standard test sets in the CSJ. The total speech length is 1.7 h.

B. Speaker-class model

The CSJ is used to train the SI and SC models. The total number of lectures used for training is 2667. Each lecture is given by one speaker. Therefore, the total number of speakers is also 2667. Note that some speakers gave several lectures. The total speech length is approximately 447 h. The SI model is a set of shared-state triphones (an HMnet) that has 3000 tied states with 32 mixtures of diagonal covariance Gaussians per state. For speaker clustering, SD monophonic HMMs are trained for each training speaker at the beginning. The model structure is a left-to-right HMM with three states, and the number of mixture components is 12. The 2667 SD models are clustered by the algorithm described in Sec. II. The number of classes is set between 10 and 500. The speaker radius for soft clustering is set between 200 and 250. After the speaker clustering step is conducted, 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

Fig. 3 shows the recognition results for each clustering condition. In the figure, r indicates the value of the cluster radius for the soft clustering. The SI model was used for a baseline in the experiments. For the hard clustering, an increase in the number of classes produces a sharp decrease in recognition performance. The reason that the performance drops is clearly a lack of training data for each class due to the greater number of classes. For the hard clustering, we stopped before creating more than 100 classes because there was no possibility of improving the performance. On the other hand, the recognition performance of the soft clustering is almost always better than the baseline of the SI model. As the number of classes increases, the performance of the soft clustering tends to improve. The best WER of 20.48% was obtained on the condition that the number of classes was 300. The reason that better results were obtained is apparently the property of the soft clustering by which a training speaker was allowed to belong to several classes. This property allows the clustering to prevent deterioration from an insufficient amount of training data.

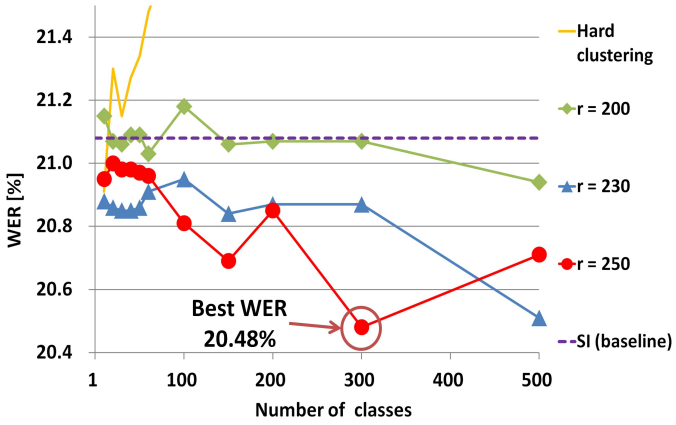

Fig. 3. *WERs of various numbers of classes. Results of both soft and hard clustering are plotted.*

Fig. 4 shows the relationship between the number of classes and the ideal WER, which is the WER if the system selects the highest-performing model for every utterance. It is clear that there are considerable differences between the ideal WERs and the WERs obtained in the recognition experiments. The results indicate the possibility of reducing the error rate by up to 41.2 % if the SC model selection were successful. This means that the SC model itself is very effective for speaker-independent speech recognition. This suggests that there is room for further research into the model selection method. Furthermore, there is another behavior to note in the results. Fig. 4 shows that the ideal WERs increase with r. This tendency is reasonable, because reducing the cluster radius of the soft clustering leads to more precise modeling of speaker characteristics. However, this tendency is totally opposite to the tendency in Fig. 3. This implies that the likelihood-based selection does not work well when the SC models do not include many training speakers.

We also examined the relationship between likelihood and the WER of each SC model. Fig. 5 shows the relationship for a certain utterance. There are 501 points plotted in the figure, and each point represents one SC model created by the soft clustering with the radius set to 230. One of those

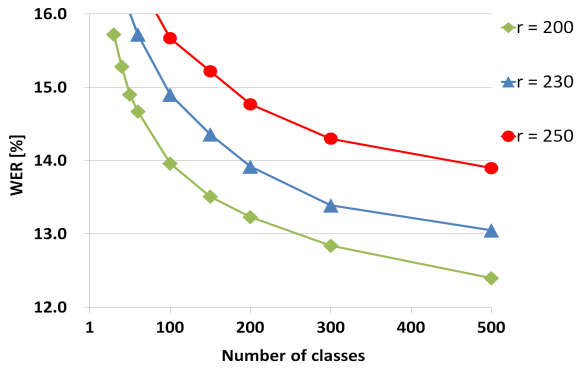

Fig. 4. *Relationship between the number of classes and the ideal WER.*

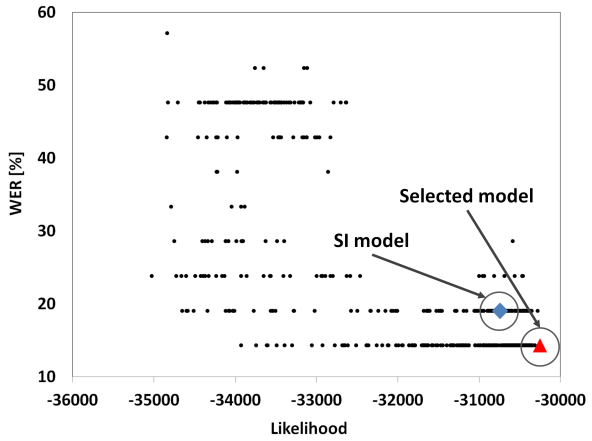

Fig. 5. *Relationship between likelihood and WER of SC models for a certain utterance (ID = 136). The selection method worked well for this utterance.*

plots represents the SI model and is marked by a diamond. A triangle represents the model selected by the likelihood-based criterion. Based on the results in this figure, the model selected by the criterion yields the highest performance. Thus, the model selection worked well for this utterance. Fig. 6 shows the relationship for another utterance by the same speaker. In this case, the model selection procedure did not select the best-performing model. There are many models that yield better performance than the selected model. We must ascertain why the most appropriate model cannot be selected by the likelihood criterion.

V. Conclusions

In this paper, we investigated speaker-independent speech recognition using SC models. The number of speaker classes was relatively small in previous studies. In this work, as many as 500 SC models were used to enable more precise modeling of speaker characteristics.

From the results of the experiments, a limited performance improvement was obtained with the conventional method in which several tens of SC models were used. In contrast, significant improvement was achieved using large-scale SC modeling. In addition, the soft clustering technique became very effective when the number of SC models increased. The performance of the hard clustering decreased sharply as the

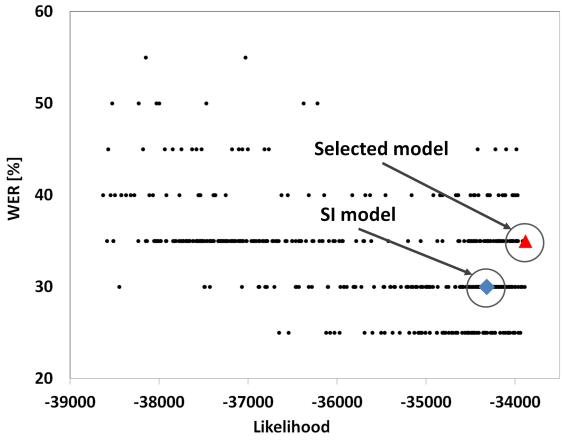

Fig. 6. *Relationship between likelihood and WER of SC models for a certain utterance (ID = 001). The selection method did not work well for this utterance.*

number of models increased, because the amount of training data for each class model became insufficient.

The results indicate the possibility of reducing the error rate by up to 41.2 % if the SC model selection were successful. This means that the SC model itself is very effective for speaker-independent speech recognition. However, the selection criterion is not adequate. We applied the likelihood criterion in this work. We will conduct a review and analyze the selection criterion in the future.

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] S.Yoshizawa, A.Baba, K.Matsunami, Y.Mera, M.Yamada, and K.Shikano, "Unsupervised speaker adaptation based on sufficient HMM statistics of selcted speakers," in Proc. of ICASSP2001, 2001, pp. 341-344.

[2] 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.

[3] 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.

[4] M.Tani, T.Emori, Y.Ohnishi, T.Koshinaka, and K.Shinoda, "Speaker selection for unsupervised speaker adaptation based on HMM sufficient statistics," in IPSJ SIG Technical Reports, 2007- SLP-69-15, 2007, pp. 85-89.

[5] T. Kosaka, S. Matsunaga, and S. Sagayama, "Speker-independent speech recognition based on tree-structured speaker clustering," Computer Speech and Language, vol. 10, 1996.

[6] 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.

[7] D.Enami, F.Zhu, K.Yamamoto, and S.Nakagawa, "Soft-clustering technique for training data in age- and gender-independent speech recognition," in Proc. of APSIPA2012, 2012, pp. 1-4.

[8] D.Jouvet and N.Vinuesa, "Classification margin for improved class-based speech recognition performance," in Proc. of ICASSP2012, 2012, pp. 4285-4288.

[9] N.Sugamura, K.Shikano, and S.Furui, "Isolated word recognition using phoneme-like templates," in Proc. of ICASSP83, 1983, pp. 723-726.