Speaker Adaptation Based on System Combination Using Speaker-Class Models

Tetsuo Kosaka, Takashi Ito, Masaharu Kato and Masaki Kohda

Graduate School of Science and Engineering, Yamagata University, Yonezawa-city, Yamagata, Japan

Abstract

In this paper, we propose a new system combination approach for an LVCSR system using speaker-class (SC) models and a speaker adaptation technique based on these SC models. The basic concept of the SC-based system is to select speakers who are acoustically similar to a target speaker to train acoustic models. One of the major problems regarding the use of the SC model is determining the selection range of the speakers. In other words, it is difficult to determine the number of speakers that should be selected. In order to solve this problem, several SC models, which are trained by a variety of number of speakers, are prepared in advance. In the recognition step, acoustically similar models are selected from the above SC models, and the scores obtained from these models are merged using a word graph combination technique. The proposed method was evaluated using the Corpus of Spontaneous Japanese (CSJ), and showed significant improvement in a lecture speech recognition task.

Index Terms: speech recognition, speaker adaptation, speaker-class model, LVCSR

1. Introduction

Although high recognition performance has been achieved for read speech, it is well known that for spontaneous speech recognition, the performance has been rather poor. This problem can be addressed by gathering a large amount of speech data. In Japan, a spontaneous speech database "Corpus of Spontaneous Japanese" (CSJ) is available. This corpus consists of about 7M words with a total speech length of 650 h [1]. However, the technique of gathering a large amount of data may not always succeed because the corpus may include speech data with mismatched speaker characteristics. Although there are many limitations with respect to spontaneous speech recognition, in this study, we focus only on those pertaining to speaker characteristics. In particular, we focus on improving unsupervised batch adaptation. It is applicable, for example, for automatic transcription of speech in meetings.

On this issue, the use of a speaker-class (SC) model has been proposed. There are two major problems to be addressed: (1) the selection of speakers who are acoustically close to the test speaker and (2) determination of the selection range of speakers. To overcome the former problem, a speaker clustering method has been proposed [2]. In this technique, SC models are created in advance, and for a particular evaluation speaker, the most appropriate SC model is selected. An alternative is to select "cohort speakers" for each evaluation speaker and create an SC model by using the selected speakers [3] [4]. A basic solution to the problem of determining the selection range of speakers is to determine the number of selected speakers on an experimental basis [3] [4]. In this technique, the number of training speakers is fixed before the recognition step. However, the selection range of speakers depends on the evaluation speaker and the accuracy of the acoustic model. For example, if there is a training speaker whose characteristics are very close to those of an evaluation speaker, then it may have been redundant to select many cohort speakers.

In order to solve this problem, we propose a new SC-based technique that varies the selection range according to different recognition conditions. The basic concept of this technique is that several models, each having different number of speakers, are prepared in advance for each evaluation speaker, and then, the models that are closest to the evaluation speaker in terms of speaker characteristics are selected in the recognition step on the basis of likelihood. In order to implement the proposed method in the recognition step, we employ a system combination technique that combines several word graphs from different SC models to form a single word graph. After generating the word graph, the scores derived from the SC models closest to the evaluation speaker are merged.

In addition, we propose an unsupervised speaker adaptation based on SC models to further improve the recognition performance. In order to evaluate the performance of the proposed method, we performed recognition experiments on the CSJ using a large vocabulary continuous speech recognition (LVCSR) system.

2. System combination using speaker-class models

2.1. Speaker-class model

In the proposed method, several models that are trained by a variety of numbers of speakers are prepared in advance for each evaluation speaker. In order to create these SC models, training speakers whose characteristics are similar to those of the evaluation speaker must be selected. Generally, Gaussian mixture models (GMMs) are used to measure the similarity between training and evaluation speakers. Referring to the results of speaker recognition in [5], phoneme hidden Markov models (HMMs) are used instead of GMMs to measure the similarity. The procedure for creating SC models is as follows: First, speaker-dependent (SD) monophone HMMs for each training speaker are prepared to measure the similarity between training and evaluation speakers. Next, a likelihood calculation is carried out using a simple frame-synchronous beam search decoder with the SD HMMs and a phoneme-pair grammar. The first 20 utterances of each evaluation speaker are used for the likelihood calculation. This is because it was found, in a preliminary experiment, that a few utterances were insufficient for similarity measurement. Finally, all training speakers are arranged in the
order of likelihood for each evaluation speaker. In order to obtain various sizes of SC models, the number of selected speakers is varied. Thus, several SC models are trained for each evaluation speaker. Although the SC models are trained separately in this work, computational cost for model training can be reduced by using the techniques based on sufficient statistics [3].

### 2.2. System combination

In this section, we describe a combination approach to SC systems. The combination of multiple SC systems that have different speaker characteristics is expected to improve the recognition performance by making use of complementary information. In order to obtain acoustic models with different speaker characteristics, the number of training speakers used to train them is varied.

Some system combination approaches (e.g., ROVER [6] and confusion network combination (CNC) [7]) are proposed to improve the performance of LVCSR systems. In this study, we use the word graph combination technique that has been proposed by Chen et al. [8]. In this combination technique, word graphs are directly integrated. Unlike in conventional combination approaches such as ROVER or CNC, the timing information for all word hypotheses is well preserved.

Fig. 1 shows the block diagram of the proposed system. A one-pass frame-synchronous search algorithm using beam searching has been adopted in the first pass. The search algorithm calculates the acoustic and language likelihood to obtain word graphs. In the first pass, several SC models are used as acoustic models, and a bigram is used as a language model. After several word graphs are obtained, they are combined to form a single graph. In the second pass, a trigram is used to rescore the combined graph, and thus, recognition results are obtained.

Here, we describe the algorithm of the system combination. Suppose that there are $N$ word graphs, $W_1, W_2, \ldots, W_N$, to be combined. If two arcs $q_1$ in $W_1$ and $q_2$ in $W_2$ are equal, then the two word graphs $W_1$ and $W_2$ can be combined as

$$W_1 + W_2 = \{ q = q_1 + q_2 | q_1, q_2 \in W_1 \} \cup \{ q_1 | q_1 \not\in W_2 \}.$$

(1)

Two equal arcs have the same word ID, start time, and end time. The word graph combination for all systems can be obtained as

$$W = W_1 + W_2 + \cdots + W_N = \sum_{i=1}^{N} W_i.$$

(2)

We compare some scoring methods when several arcs are combined. In the proposed method, several SC models that have different speaker characteristics are used. Each arc has several scores obtained from the SC models. In this case, models that have higher likelihood values are expected to be closer to the target speaker. These models are considered to be suitable for recognizing utterances of the target speaker. On the basis of the above consideration, we propose the following scoring methods.

**Maximum score (MAX)** The merged score is simply obtained by selecting the maximum value from among $N$ systems.

$$score_{MAX}(q) = \max \{ score(q_1), score(q_2), \ldots, score(q_N) \}.$$

(3)

where $score(q_k)$ is the log-likelihood value of system $k$ at arc $q$.

**Average score (AVE)** The merged score is obtained by averaging all scores.

$$score_{AVE}(q) = \frac{1}{N} \sum_{k=1}^{N} score(q_k).$$

(4)

**N-best selection (NB)** The $score(q_k)(k = 1, \ldots, N)$ are sorted by their log-likelihood values and the top $N_s$ scores are merged as

$$score_{NB}(q) = \frac{1}{N_s} \sum_{k=1}^{N_s} score(q_k).$$

(5)

This method selects speaker models that are closer to the target speaker. Since the selection is conducted at each arc separately, the selected model sets may differ among arcs. However, the number of selected models ($N_s$) is common to all arcs.

**Likelihood-based selection (LS)** The scores are sorted by their log-likelihood values and the top $n_q$ scores are merged as

$$score_{LS}(q) = \frac{1}{n_q} \sum_{k=1}^{n_q} score(q_k).$$

(6)

Here, $n_q$ is automatically determined by the difference in the log-likelihood value from the best score. Since the selection is conducted at each arc separately, the number of selected models may differ among arcs.

### 3. Unsupervised speaker adaptation

In order to demonstrate the effectiveness of the proposed unsupervised speaker adaptation based on the SC models, we conduct recognition experiments. In these experiments, an initial model is adapted by using labels obtained from the recognition results. We compare the following methods.

**SI-SI (baseline)** A speaker-independent (SI) model is used as the initial model. Speaker adaptation is carried out by using the recognition results of the SI model as labels for adaptation at the first iteration. At subsequent iterations, the results of the adapted model are used as labels.

**SI-SC** An SI model is used as the initial model. Labels for adaptation are generated by recognizing evaluation data using system combination of several SC models at the first iteration. At subsequent iterations, the results of the adapted model are used as labels.

**SC-SC** A “1/4” SC model is used as the initial model (see Table 2). The method of generating adaptation labels is the same as that in the SI-SC method.

The above-mentioned methods are summarized in Table 1. For adaptation, we use the conventional maximum likelihood linear regression (MLLR) technique [9]. In the MLLR adaptation, the Gaussian mean parameters are updated. The mixture weights are also updated by the maximum likelihood estimation. The number of regression classes is automatically determined by the
amount of adaptation data. In these experiments, 9 to 16 regression classes were used.

4. Experimental set-up

4.1. LVCSR system

In this section, we describe our LVCSR system, which is used for recognition experiments. 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 13-dimensional feature (12-dimensional MFCC and log power) is derived from the digitized samples for each frame. Also, the delta and delta-delta features are calculated from the MFCC feature and the log power. Then, 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 using a bigram and trigram is used for recognition. In the first pass, a word graph is generated using acoustic models and the bigram language model. In the second pass, the trigram language model is applied to re-score the word graph, and thus, the recognition results are obtained. Decoding is performed using a one-pass algorithm in which a frame-synchronous beam search and a tree-structured lexicon are applied in the first pass. The bigram and trigram models are trained from text data containing 2,668 lectures in the CSJ, and the total number of words is 6.68 M. Trained language models have 47,099 word- pronunciation entries. A set of shared-state triphones is used as an acoustic model.

4.2. Speaker-class model

In this section, the conditions for speaker-class modeling are described. The total number of lectures used for model training is 2,667, and the total speech length is about 447 h. One lecture is given by one speaker; therefore, the total number of speakers is 2,667. Note that some speakers give several lectures. The steps for the creation of a SC model are as follows:

1. Speaker-dependent (SD) monophone HMMs are trained for each training speaker and are used for speaker selection. The model topology is a left-to-right HMM with three states. The number of mixture components is 12. The number of training speakers is 2,667. Then, the number of model sets of SD HMMs is 2,667.

2. All training speakers are arranged in the order of likelihood corresponding to a given evaluation speaker by using the above SD HMMs. 20 utterances from the evaluation data are used for likelihood calculation.

3. Speaker-class models are trained for each evaluation speaker on the basis of the speaker order described above. In order to obtain various sizes of SC models, the number of speakers is varied, and seven SC models are trained for each evaluation speaker (see Table 2). The experiments use block-diagonal HMMs in which the correlations between static, delta or delta-delta coefficients are assumed to be zero. The models have 3000 tied states with 32 mixture components per state.

4.3. Evaluation set

We use the “testset1” evaluation set, which consists of academic presentations given by 10 male speakers. This is one of the standard test sets in the CSJ corpus. Experimental results of each research group can be compared by using this test set. The total speech length is 1.7 h.

5. Experimental results

5.1. Performance of speaker-class model

Table 3 shows the recognition results of each SC model without considering the system combination. In the figure, 1/64 - ALL indicate the type of SC model (see Table 2). The best word error rate (WER) was 19.11% using the 1/4 model. Thus, the performance of the SC model showed better results than SI the model (19.75%). When the number of training speakers was limited, the recognition performance dropped. It would appear that the reason why the performance dropped was lack of training data.

5.2. System combination

Fig. 2 shows the performance in terms of WERs for the N-best based combination method. The x-axis indicates the number of selected models. The WER of the SI model was 19.75%, and that of SC model was 19.11% (see Table 3). All results of the system combination indicated a better performance than the SI and SC systems. The best WER of 18.65% could be obtained with the NB method by using the top five models. The results suggest that the proposed system combination by using model selection is more effective than the AVE method, where the scores of all models are used. This further implies that the use of closer models is more effective.

Table 4 shows the recognition performance of the LS method. In the experiments, the number of selected models was determined on the basis of likelihood. Models in which the difference calculated by subtracting its score from the best score was less than the threshold value were selected for model combination. Thus, when the threshold value was set to 0, it was equivalent to MAX scoring method, whereas when it was set to ∞, it was equivalent to AVE scoring method. In the experiments, the best WER of 18.50% could be obtained by setting the threshold to 100; the result showed better performance than the NB method. In the NB method, the number of combined models for each arc is fixed, whereas in the LS method, this number differs among arcs. This implies that it is better to vary the number of combined models associated with each arc.

The experimental results of the various recognition techniques are summarized in Fig. 3. All the system combination
5.3. Unsupervised speaker adaptation

We conducted experiments using an unsupervised speaker adaptation to investigate the effectiveness of SC models. The experimental results are shown in Fig. 4. Each adaptation procedure was conducted until the recognition performance saturated. The SI model was used as an initial model for the SI–SI and SI–SC methods. The performance of the SI model was 19.75%. In the SC–SC method, the 1/4 SC model was used as the initial model, and the WER of the SC model was 19.11%. The recognition results of the system combination using the LS (WER: 18.50%) were used as labels for adaptation in the SI–SC method; however, the difference between the SI–SC and SC–SC methods was small. From the results, we observe that accurate transcriptions are more important than the performance of the initial model.

6. Conclusions

In this paper, we proposed a new approach using speaker-class (SC) models for an LVCSR system. One of the major problems regarding the use of the SC model is determining the selection range of the speakers. In order to solve this problem, we proposed a system combination technique where several word graphs from different SC models were combined. The proposed method was evaluated using the Corpus of Spontaneous Japanese (CSJ). The results of the experiments indicated that the performance of the proposed system combination approach was superior to that of the conventional SC-based system. In addition, the SC-based unsupervised adaptation method achieved a significant improvement with regard to recognition performance as compared with the conventional MLLR adaptation method.

Figure 2: Results of N-best-based combination method.

Figure 3: Comparison of several recognition techniques.

Figure 4: Results of unsupervised speaker adaptation.

7. References