PAPER *Special Section on Processing Natural Speech Variability for Improved Verbal Human-Computer Interaction*

Unsupervised Speaker Adaptation Using Speaker-Class Models for Lecture Speech Recognition

Tetsuo KOSAKA[†a)], ***Member***, **Yuui TAKEDA**[††], **Takashi ITO**[†], ***Nonmembers***, **Masaharu KATO**[†], ***and* Masaki KOHDA**[†], ***Members***

SUMMARY In this paper, we propose a new speaker-class modeling and its adaptation method for the LVCSR system and evaluate the method on the Corpus of Spontaneous Japanese (CSJ). In this method, closer speakers are selected from training speakers and the acoustic models are trained by using their utterances for each evaluation speaker. One of the major issues of the speaker-class model is determining the selection range of speakers. In order to solve the problem, several models which have a variety of speaker range are prepared for each evaluation speaker in advance, and the most proper model is selected on a likelihood basis in the recognition step. In addition, we improved the recognition performance using unsupervised speaker adaptation with the speaker-class models. In the recognition experiments, a significant improvement could be obtained by using the proposed speaker adaptation based on speaker-class models compared with the conventional adaptation method.

key words: *speech recognition, speaker adaptation, speaker-class model, LVCSR, corpus of spontaneous Japanese*

1. Introduction

High recognition accuracy has been achieved for read speech with a large vocabulary continuous speech recognition (LVCSR) system. However, it is well known that rather poor performance is reported for spontaneous speech recognition. Although there are full of problems to be solved in spontaneous speech recognition, we focus on the problem of speaker characteristics in this paper. Gathering a large amount of speech data has been a general solution of the problem. A large scale speech corpus containing a total speech length of 2100–2300 hours is available for research organizations participating in DARPA EARS program now [1], [2]. 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 hours [3]. However, gathering a large amount of data cannot always succeed because mismatched speech data may be included in a corpus. There are some reasons for the acoustic mismatch. For example, a variation of speaker characteristics, a speaking style and a recording environment often have adverse effects upon the performance of speech recognition.

The aim of this work is to improve the recognition performance of spontaneous speech by using unsupervised batch-type adaptation for acoustic models. The batch-type adaptation is mainly applied to the automatic transcription of meetings and lectures. It is useful in those applications, since there is no need for real-time speech recognition.

Adaptation techniques can be classified into two methods, i.e. supervised and unsupervised modes. The unsupervised method is more difficult to improve the recognition performance because recognition results are used for adaptation. In particular, it is difficult to improve the performance of spontaneous speech because it has relatively higher error rate. In order to improve the unsupervised adaptation performance, we propose a novel unsupervised speaker adaptation method by using speaker-class models.

On the issue of speaker characteristics, the use of speaker-class model has been proposed. There are two major problems to be solved: 1) how to select speakers who are acoustically close to the test speaker, and 2) how to determine the selection range of speakers. For the former problem, it has been proposed that a speaker clustering method in which speaker-class models are created in advance and the most proper speaker-class model is selected [4]. It will be referred to 'speaker clustering' method in this paper. An alternative is to select 'cohort speakers' for each evaluation speaker and to create speaker-class model by using the selected speakers [5], [6]. Comparing between the two methods, the 'cohort speakers' method is thought to be more effective because it takes the evaluation speaker into account for creating speaker-class model. A basic solution for the problem of determining the selection range of speakers is that the number of selected speakers is determined on an experimental basis [5], [6]. However, the selection range of speakers depends on the evaluation speaker and the accuracy of acoustic models. For example, if there is a training speaker whose characteristics are very close to the evaluation speaker, many cohort speakers may not be necessary. It is difficult to estimate how many speakers should be selected for creating the speaker-class model.

In order to solve the problem, we employ the selection method of speaker-class model based on likelihood basis. Several models which have variety of speaker range are prepared for each evaluation speaker in advance, and the most proper model is selected in recognition step. Tani et al. defined the distance between speaker models, and proposed the method of determining the number of speakers based on the distance [7]. However, if the accuracy of the speaker

Manuscript received November 30, 2009.

Manuscript revised February 25, 2010.

[†] The authors are with the Graduate School of Science and Engineering, Yamagata University, Yonezawa-shi, 992–8510 Japan.

[††] The author is with the Faculty of Engineering, Yamagata University, Yonezawa-shi, 992–8510 Japan.

a) E-mail: tkosaka@yz.yamagata-u.ac.jp

DOI: 10.1587/transinf.E93.D.2363

Table 1 Comparison of speaker-class based methods.

previous works	speaker selection	determination of selection range
[4]	speaker clustering	likelihood
[5], [6]	cohort speakers	-
[7]	cohort speakers	distance
proposed	cohort speakers	likelihood

class model changes, a re-setup of a threshold value is required. Table 1 shows the classification of speaker-class methods. To briefly summarize the merits of this work: 1) Speaker-class models depending on evaluation speaker are used, and 2) the range of speaker selection is determined automatically based on likelihood basis.

In order to improve the unsupervised adaptation performance further, we propose a speaker adaptation method based on speaker-class models. The error rate is crucial for the unsupervised adaptation because recognition results are used for adaptation in the unsupervised mode. In the proposed method, recognition results by speaker-class models are used as transcriptions of unsupervised adaptation. From the recognition experiments, we show that accurate transcriptions are important for unsupervised adaptation, while initial model itself is not significant.

The rest of this introduction reviews some related works. The speaker-clustering approach utilizes the correlation among speakers. Two similar approaches were proposed, i.e. reference speaker weighting (RSW) and eigenvoice (EV) [8], [9]. In the former method, the parameters of a set of models are adapted to match the recognized speaker based on the recognized speaker's similarity to a set of reference speakers. The basic idea of the latter method is not essentially different from the RSW method except the reference vectors are computed. Mak et al. proposed the combination of reference speaker selection and the RSW method [10]. The method consists of two processes. First, the top M reference speakers are selected based on likelihood basis. After the speaker selection, the RSW procedure is conducted. Since the number of M is fixed in the speaker selection step, the issue of the selection range of speakers cannot be solved. Then the method differs from our approach. The results of that work showed that the combination of the speaker selection and the RSW was effective. Then, it is expected that the combination method of our speaker selection approach and the RSW is effective, however, it is out of our investigation in this paper.

2. Speech Recognition with Speaker-Class Models

2.1 Overview

The basic idea of speaker-class modeling is that closer speakers are selected from training speakers and acoustic models are trained by using their utterances for each evaluation speaker. It is expected that recognition performance will be improved because an acoustic mismatch between input speech and models becomes small. Since the speaker-class models are created for each evaluation speaker separately, the models are more suitable for input speech than speaker-class models which are created independently of evaluation speaker. One of the major issues of speaker-class model is determining the selection range of speakers. So, we employ the following method to determine the range. Several models which have a variety of speaker range are prepared for each evaluation speaker in advance. In the recognition step, the most proper model is selected for each utterance based on likelihood basis. By way of comparison, model selection not for each utterance but for each speaker is also conducted. Since a feature variation of utterances is large in spontaneous speech condition, the former method is expected to show better recognition performance. The details of speaker-class modeling are described in the following subsections.

2.2 Speaker-Class Modeling

In order to create speaker-class models, closer training speakers must be selected. Generally, Gaussian Mixture Models (GMMs) are used to measure similarity between training and evaluation speakers. In the example of [5], the likelihood values of the input speech uttered by the evaluation speaker are calculated by speaker-dependent GMMs. Using the values, the training speakers are ordered according to the similarity to the evaluation speaker. Referring to the results of speaker recognition in [11], phoneme HMMs are used to measure it instead of GMMs. In that paper, a speaker vector-based speaker identification was conducted and HMM-based system showed better results than GMM-based system. Based on the results, HMMs are used to measure similarity in this paper.

The procedure for creating the speaker-class model is as follows. First, speaker-dependent (SD) monophone HMMs for each training speaker are prepared to measure similarity. Next, likelihood calculation is carried out by using a simple frame-synchronous beam search decoder with a phone-pair grammar. The first 20 utterances of each evaluation speaker are used for likelihood calculation. This is because that it was found that a few utterances were not enough for similarity measurement in a preliminary experiment. Finally, all training speakers are arranged in the likelihood order for each evaluation speaker by using the SD HMMs. In order to obtain various sizes of speaker-class models, the number of speakers is varied and several speaker-class models are trained for each evaluation speaker. In addition, we perform unsupervised speaker adaptation with recognition results derived from speaker-class models. Performance improvement can be expected by using more accurate labels from speaker-class models. We use MLLR which is widely used for speaker adaptation. In this paper, we compare the following adaptation methods: 1) Comparison between speaker-independent (SI) model and speaker-class model is conducted to clarify which model is suitable for initial model to be adapted, and, 2) comparison between label sets derived from SI model and speaker-class model to clarify which label set is better to use for adaptation.

2.3 Algorithm

The training procedure of the speaker class models is described below.

1. Create SD HMMs by using training data consist of S speakers. Then, the number of SD HMMs is S. The SD HMMs for speaker q_s are referred to as Λ_{q_s}.
2. Let $\boldsymbol{X} = \{x_1, x_2, \cdots, x_M\}$ be the set of the input features of an evaluation speaker where M is the number of utterances. In the experiments, M is set to 20. Calculate $P(\boldsymbol{X}|\Lambda_{q_s})$ by using a standard speech recognizer.
3. Sort the training speakers in the order of likelihood basis, i.e. $q_1, q_2, \cdots, q_S$ when $P(\boldsymbol{X}|\Lambda_{q_1}) > P(\boldsymbol{X}|\Lambda_{q_2}) > \cdots > P(\boldsymbol{X}|\Lambda_{q_S})$
4. Make a classification of the training speakers on the basis of likelihood in order to generate N speaker-classes. Then, $S_1 < S_2 < \cdots < S_N = S$ where S_n is the number of speakers in class n.
5. Create N speaker-class models $\{\Theta_n\}$ based on the Maximum Likelihood (ML) criterion by using each speaker-class data set. For example, the speaker-class model for the class n is trained by $\{\boldsymbol{X}^{(q_1)}, \boldsymbol{X}^{(q_2)}, \cdots, \boldsymbol{X}^{(q_{S_n})}\}$ where $\boldsymbol{X}^{(q_{S_n})}$ is the whole training data for speaker q_{S_n}.

In the recognition step, the optimal number of n for speaker-class model Θ_n is determined automatically based on likelihood basis. There are two methods for determination:

Utterance The optimal model is selected on a utterance-by-utterance basis by using

$$\hat{n} = \operatorname*{argmax}_n P(x_i|\Theta_n), \tag{1}$$

where x_i is i-th utterance of the evaluation speaker. The utterance is recognized by the selected model $\Theta_{\hat{n}}$. Therefore, the model may be replaced at every utterance.

Speaker The model is selected by using whole utterances of the evaluation speaker. Let L is the number of utterances of the evaluation speaker. Then, $\hat{n}$ is set as

$$\hat{n} = \operatorname*{argmax}_n \prod_{i=1}^{L} P(x_i|\Theta_n). \tag{2}$$

Whole utterances of the evaluation speaker are recognized by $\Theta_{\hat{n}}$.

Here we will discuss the issue of calculation cost in the recognition step. Since it mainly depends on decoding process, it increases with the number of speaker-class (SC) models considered. It is expected that it will be reduced if the SC models are used only at the second decoding pass in multi-pass decoder. This will be the subject of the future investigation.

2.4 Acoustic Models

Since the speaker class model is selected based on likelihood basis in our approach, there is an advantage that the re-setup of a threshold value in distance based methods is not required if the accuracy of the speaker class model changes. In order to demonstrate the merit in the experiments, both block-diagonal HMMs and diagonal HMMs are used.

In our implementation, the output probability distribution is represented by multiple Gaussian mixture densities:

$$b(\boldsymbol{x}) = \sum_{w=1}^{W} \lambda_w \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_w, \Sigma_w)$$

$$= \sum_{w=1}^{W} \lambda_w \frac{1}{(2\pi)^{D/2}|\Sigma_w|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_w)^T \Sigma_w^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_w)\right\}, \tag{3}$$

where λ_w is the mixture weight of w-th component, $\boldsymbol{\mu}_w$ is mean vector and Σ_w is covariance matrix. In the diagonal case, the covariance matrix of Gaussian mixture is given by the diagonal elements:

$$\Sigma_w = diag(\sigma_1^2, \sigma_2^2, \cdots, \sigma_D^2) \tag{4}$$

The block-diagonal matrix, in which correlations between static, delta or delta-delta coefficients are assumed to be zero, can be given as:

$$\Sigma_w = \begin{bmatrix} M_1 & O & O \\ O & M_2 & O \\ O & O & M_3 \end{bmatrix}, \tag{5}$$

where M_i is a d_i-dimensional symmetric matrix and d_i satisfies:

$$\sum_{i=1}^{3} d_i = D. \tag{6}$$

The matrix M_1 represents correlations among static coefficients. Also, M_2 for delta coefficients and M_3 for delta-delta coefficients are used.

3. Experimental Set-Up

3.1 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 and at a quantization size of 16 bits. The short-time analysis is performed by means of a Hamming window, having length 25 ms at a rate of 8 ms. The 13-dimensional feature vector (12-dimensional MFCC and log power) is derived from the digitized samples for each frame. Also the delta and the delta-delta feature vectors are calculated from MFCC feature vector and 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 word 2-gram and 3-gram is used for recognition. In the first pass, word graph is generated with acoustic models and 2-gram language model. In the second pass, 3-gram language model is applied for re-scoring the word graph and recognition result is obtained. Decoding is performed by a one-pass algorithm in which a frame-synchronous beam search and tree-structured lexicon are applied in the first pass. The 2-gram and 3-gram are trained from text data containing 2,668 lectures in the CSJ and the total number of words is 6.68M. Those lectures consist of both academic presentations and extemporaneous presentations. Trained language models have 47,099 word-pronunciation entries. A set of shared state triphones is used as acoustic model.

3.2 Speaker-Class Model

In this section, the condition of speaker-class modeling is described. Speech data of both academic presentations and extemporaneous presentations are used as training set for acoustic modeling. The total number of lectures we used for modeling is 2,667 and the total speech length is about 447 hours. One lecture is uttered by one speaker, then the total number of speakers is 2,667 (1594 males, 1073 females). Note that some speakers give plural lectures. The steps of a creation procedure are as follows:

1. Speaker-dependent (SD) monophone HMMs are trained for each training speaker, and are used for a speaker selection. A model topology is a left-to-right HMM with 3 states. The number of mixture components is 12. Since the number of training speakers is 2,667, 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.
3. Speaker-class models are trained for each evaluation speaker based on the result of the above speaker order. In order to obtain various sizes of speaker-class models, the number of speakers is varied and seven speaker-class models are trained for each evaluation speaker (see Table 2). In the experiments, both block-diagonal HMMs and diagonal HMMs are used. The block-diagonal system has 3000 tied states with 32 mixture components per state, and the diagonal system has 3000 states with 16 components per state.

Table 2 Number of speakers for each speaker-class model.

Proportion of all	#speakers
1/64	42
1/32	84
1/16	167
1/8	334
1/4	667
1/2	1334
ALL	2667

3.3 Evaluation Set

The evaluation set we use is ‘testset1’ which consists of academic presentations uttered by 10 male speakers. This is one of the standard test sets in CSJ corpus. Experimental results of each research group can be compared each other by using this test set. The total speech length is 1.7 hours.

4. Result and Discussion

4.1 Selection of Speaker Class Model

In order to investigate the effects of model precision on recognition performance, we conducted recognition experiments by using both the block-diagonal models and the diagonal models. Figure 1 shows the recognition results. In the figure, 1/64—ALL are the results of speaker-class models without likelihood selection. The best WER result of the diagonal covariance models was 20.26% by using 1/32 model. In contrast, the best WER of the block-diagonal models was 19.11% by using 1/4 model. The reason why the optimal number of training speakers of the block-diagonal model increased was lack of training data when the number of training speakers was limited. Since the optimal number of training speakers depends on the accuracy of a model, the framework which chooses the optimal model automatically is required. Then, we proposed the automatic selection of the optimal number of speakers by a likelihood basis. In Fig. 1, ‘Speaker’ and ‘Utterance’ are the results of the automatic selection. The ‘Speaker’ means one speaker-class model is selected for each evaluation speaker. The ‘Utterance’ means that a selected speaker-class model is varied by the likelihood of each utterance. From the results, the automatic selection by the ‘Utterance’ method shows equivalent or better performance than every speaker-class model. Table 3 shows the recognition results for each speaker. The best performance for each speaker is indicated in boldface. In this table, the optimal speaker-class model differs among evaluation speakers. This is the reason why the performance

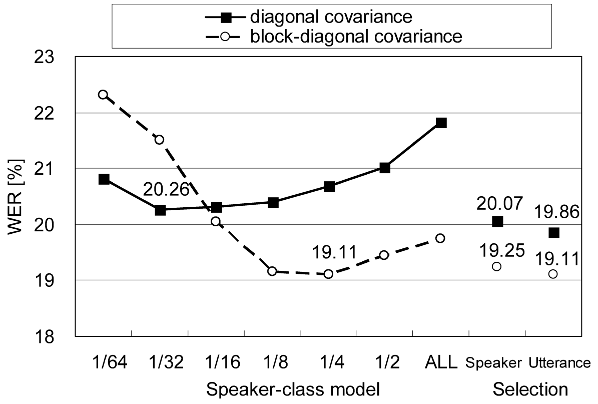

Fig. 1 The recognition results of speaker-class models and the automatic selection of speaker model by likelihood basis.

Table 3 The recognition results of the speaker-class models by using the block-diagonal model. [Word Error Rate(%)]

speaker ID	1/16	1/8	1/4	1/2	ALL
0097	7.72	7.50	7.38	**7.23**	**7.23**
0110	9.56	**9.40**	9.56	9.56	9.79
0137	18.46	17.34	17.63	**17.00**	17.54
0106	29.72	28.72	**28.36**	28.62	28.62
0112	12.09	**10.69**	11.00	11.13	11.66
0156	32.94	**31.30**	31.68	31.84	32.62
0051	14.96	14.02	**13.94**	14.73	15.24
0121	24.68	23.05	**22.31**	23.22	23.22
0123	17.98	**17.81**	18.31	19.36	18.98
0011	23.01	23.12	**22.47**	23.01	23.72
average	20.05	19.16	**19.11**	19.44	19.75

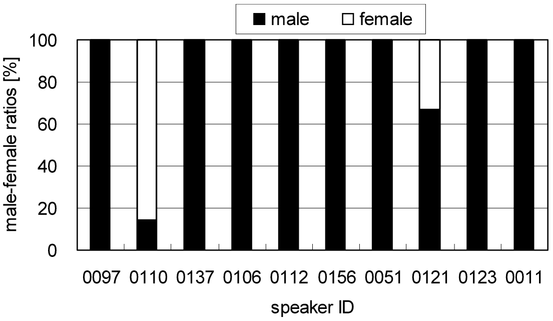

Fig. 2 Male-female ratios of training speakers.

improvement could be obtained by automatic selection. If the number of the selected speakers is fixed at the optimal point on average, it is not necessarily the case that the optimal number is selected for each speaker. Comparing between the two selection methods, the ‘Utterance’ method shows better performance. This means that the variation of characteristics of each utterance is large.

We investigated what kind of training speakers were selected for speaker-class models in the case of 1/64 (42 speakers). Figure 2 shows male-female ratios of the selected speakers. Note that the all evaluation speakers are male. For almost all speakers, the proportions of male speakers are higher, however, the result of speaker ‘0110’ shows a different trend. In this case, the proportion of male speakers is small, even though ‘0110’ is a male speaker. We have listened to his presentation, and found out that his voice quality was high-pitched. It is well known that the recognition performance of this speaker drops by using male-dependent model. Then, it is considered that his voice has female-like characteristics. In Table 3, the results of ‘ALL’ was obtained by using gender-independent (GI) models. The evaluation set ‘testset 1’ consists of male speakers only. In this case, gender-dependent (GD) models are usually effective for the recognition. In the experiments by using GD models, we obtained a WER of 20.09%. This is because female-like speakers such as speaker ‘0110’ have adverse effects on the recognition performance.

We tried to visualize the result of speaker selection by using the COSMOS method [12]. In the method, the distribution of the acoustic models is plotted in a two-dimensional diagram by use of multi-dimensional linear measurement. We employed the Bhattacharyya distance measures to calculate the similarity of two probability distributions. Figure 3 shows the result of the distribution of training speakers. The cohort speakers for speaker ‘0110’ and speaker ‘0011’ are plotted on all other training speakers. It is found that the cohort speakers for ‘0110’ are mainly selected from female speakers and the speaker class models of ‘0110’ and ‘0011’ are definitely different in speaker space.

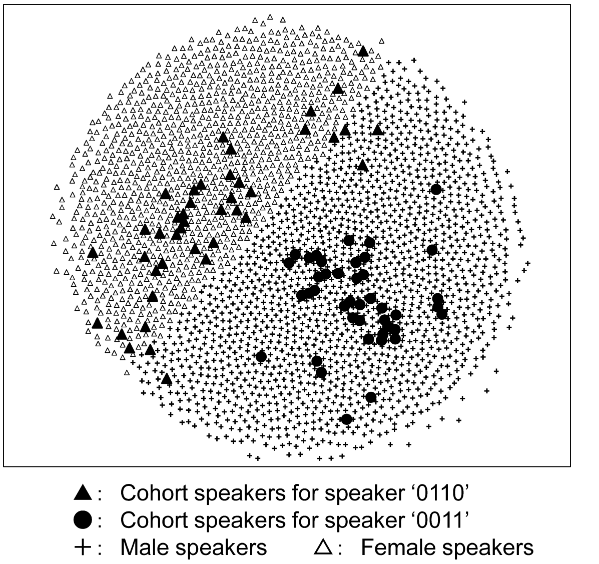

Fig. 3 Scatter plot of the distribution of the training speakers by the COSMOS method.

4.2 Unsupervised Speaker Adaptation

In order to investigate the effectiveness of the speaker-class model in unsupervised speaker adaptation, we conducted some unsupervised speaker adaptation experiments. Figure 4 shows the block-diagram of the method of unsupervised speaker adaptation. An initial model is adapted by using labels obtained from the recognition results to create an adapted model. We compare the following methods:

Baseline: The SI model is used as an 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 the following iterations, the results of the adapted model are used as labels.

Method 1: The ‘1/4’ speaker-class model is used as an initial model. Speaker adaptation is carried out by using the recognition results of the speaker-class model without likelihood selection as labels for adaptation.

Method 2: The SI model is used as an initial model. Speaker adaptation is carried out by using likelihood selection of the recognition results of the speaker-class models as labels for adaptation.

Method 3: The ‘1/4’ speaker-class model is used as an initial model. Speaker adaptation is carried out by using likelihood selection of the speaker-class models as la-

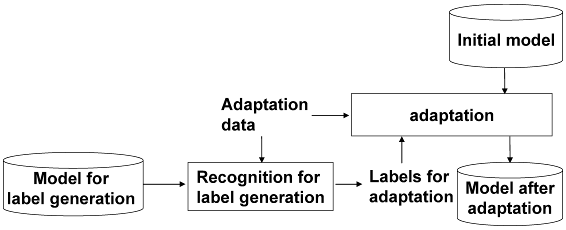

Fig. 4 The block-diagram of the method of unsupervised speaker adaptation.

Table 4 Summary of unsupervised speaker adaptation.

		labels for adaptation	
method	initial model	model for label generation	likelihood selection
baseline	SI	SI	-
method 1	speaker-class	speaker-class	no
method 2	SI	speaker-class	yes
method 3	speaker-class	speaker-class	yes
method 4	speaker-class	SI	-

bels for adaptation.

Method 4: The '1/4' speaker-class model is used as an 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 the following iterations, the results of the adapted model are used as labels.

Those methods are summarized in Table 4. The block-diagonal models were used for adaptation, and the likelihood selection was performed with 1/2, 1/4 and 1/8 speaker-class models. Each adaptation procedure was conducted until performance was saturated. In the MLLR adaptation, the Gaussian mean parameters were updated. The mixture weights were also updated by the maximum likelihood estimation. The number of regression classes is automatically determined by the amount of adaptation data. 9 to 16 regression classes were used in the adaptation procedure.

The experimental results are shown in Fig. 5. The best WERs of the baseline and each method are shown in the figure. Also, the recognition results of SI, baseline and method 2 for each speaker are shown in Table 5. The baseline method obtained 17.50% as the best WER of iteration sequence, and 17.14% for the method 1, 17.02% for the method 2, 17.13% for the method 3, and 17.33% for the method 4 were achieved, respectively. All the proposed methods could achieve the performance improvement compared with the baseline adaptation method (statistically significant at level of 1% by the sign test.) We obtained the best WER of 17.02% by using the method 2, however, the difference among the proposed methods excepting the method 4 is small. Although the difference in WERs among the methods 1, 2, and 3 was small, the method 2 consistently outperformed the methods 3 and 4. In the method 2, SI model is adapted by using the labels derived by several speaker-class models. Then, the initial model and the models for label generation differ. The unsupervised cross-validation adaptation method has been proposed by [13], where models for E-step and M-step are purportedly-separated. There is a possibility that the method 2 causes an effect similar to the cross-validation method. This will be the subject of the future investigation.

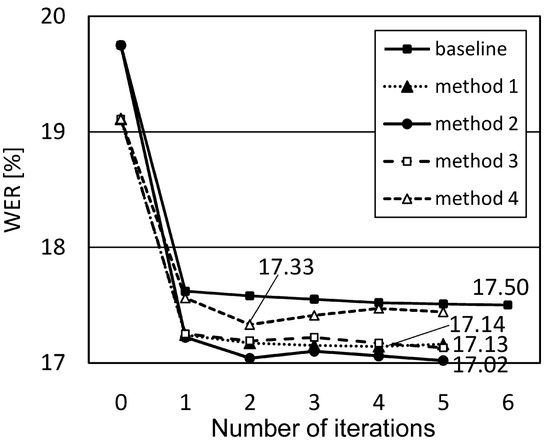

Fig. 5 Recognition results of unsupervised speaker adaptation by using '1/4' speaker-class model.

Table 5 WERs of unsupervised speaker adaptation for each speaker (%).

speaker ID	SI	baseline	method 2
0097	7.23	6.54	**6.46**
0110	9.79	**9.10**	**9.10**
0137	17.54	17.24	**16.41**
0106	28.62	**23.48**	23.64
0112	11.66	8.95	**8.30**
0156	32.62	28.92	**28.11**
0051	15.24	13.24	**13.12**
0121	23.22	**21.19**	21.36
0123	18.98	18.27	**16.93**
0011	23.72	21.47	**20.26**
average	19.75	17.50	**17.02**

From the results, all the proposed methods are able to improve the recognition performance. The reason why performance improvement of the proposed method 1, 2 and 3 could be achieved is that the accuracy of labels for adaptation were improved by likelihood selection. The method 4 is not so effective compared with other proposed methods. In this method, SI based transcriptions are used as the adaptation labels. Then, we can see that accurate transcriptions are more important rather than the performance of an initial model. We can conclude that adaptation labels obtained by speaker-class models are effective.

5. Conclusion

In this paper, we have proposed a new speaker-class modeling and its adaptation method. In order to determine the selection range of speakers, a suitable model for input speech was selected from speaker-class models which had different speaker-size by likelihood basis. The result showed that speaker-class modeling in which closer speakers were selected automatically was more effective rather than speaker-independent modeling. Furthermore, speaker adaptation was preformed to investigate whether the speaker-class model was effective. The results showed that the method in which the recognition results of speaker-class models were used as labels for adaptation was effective.

Mak et al. proposed the combination of reference speaker selection and the RSW method, and showed a significant improvement [10]. We plan to develop the combination of our speaker selection method and the RSW method. In addition, we are now studying the speaker class model selection based on the word graph combination [14]. By using the combination method, speaker class models can be selected for each word hypothesis. It is expected that the word level selection is more effective than the utterance level selection which has been proposed in this paper.

References

[1] G. Evermann, H.Y. Chan, M.J.F. Gales, B. Jia, D. Mrva, P.C. Woodland, and K. Yu, "Training LVCSR systems on thousands of hours of data," Proc. ICASSP2005, pp.209–212, 2005.

[2] S.F. Chen, B. Kingsburg, L. Mangu, D. Povey, G. Saon, H. Saltau, and G. Zweig, "Advances in speech transcription at IBM under the DARPA EARS program," IEEE Trans. Audio, Speech and Language Processing, vol.14, no.5, pp.1596–1608, 2006.

[3] S. Furui, M. Nakamura, T. Ichiba, and K. Iwano, "Analysis and recognition of spontaneous speech using corpus of spontaneous Japanese," Speech Commun., vol.47, pp.208–219, 2005.

[4] T. Kosaka, S. Matsunaga, and S. Sagayama, "Speaker-independent speech recognition based on tree-structured speaker clustering," Comput. Speech Lang., vol.10, pp.55–74, 1996.

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

[6] M. Padmanabhan, L.R. Bahl, D. Nahamoo, and M. Picheny, "Speaker clustering and transformation for speaker adaptation in speech recognition systems," IEEE Trans. Speech Audio Process., vol.6, no.1, pp.71–77, 1998.

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

[8] T.J. Hazen and J.R. Glass, "A comparison of novel techniques for instantaneous speaker adaptation," Proc. Eurospeech97, pp.2047–2050, 1997.

[9] R. Kuhn, P. Nguyen, J.-C. Junqua, L. Goldwasser, N. Niedzielski, S. Finche, K. Field, and M. Contolini, "Eigenvoices for speaker adaptation," Proc. ICSLP98, pp.1771–1774, 1998.

[10] B. Mak, T.-C. Lai, and R. Hsiao, "Improving reference speaker weighting adaptation by the use of maximum-likelihood reference speakers," Proc. ICASSP2006, pp.229–232, 2006.

[11] T. Kosaka, T. Akatsu, M. Katoh, and M. Kohda, "Speaker vector-based speaker identification with phonetic modeling," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.J90-D, no.12, pp.3201–3209, Dec. 2007.

[12] M. Shozakai and G. Nagino, "Two-dimensional visualization of acoustic space by multidimensional scaling," IPSJ SIG Technical Reports, 2004-SLP-52-23, pp.129–136, 2004.

[13] T. Shinozaki, Y. Kubota, and S. Furui, "Unsupervised cross-validation adaptation algorithms for improved adaptation performance," Proc. ICASSP2009, pp.4377–4380, 2009.

[14] I.F. Chen and L.-S. Lee, "A new framework for system combination based on integrated hypothesis space," Proc. Interspeech2006, pp.533–536, 2006.

Tetsuo Kosaka was born in Miyagi, Japan, in 1960. He received B.E., M.E. and Ph.D. degrees from Tohoku University, Miyagi, Japan, in 1984, 1986 and 1997 respectively. In 1986, he joined CANON Inc., Tokyo, Japan. From 1991 to 1995, he had been a researcher at ATR Interpreting Telephony Research Laboratories, Kyoto, Japan. Since 2002, he has been working at Yamagata University, Yamagata, Japan, and he is currently an Associate Professor at Graduate School of Science and Engineering, Yamagata University. His research interests are speech processing and its applications. He received the Paper Award of the IEICE, Japan, in 1996. He is a member of the IEEE, the ASJ and the IPSJ.

Yuui Takeda was born in Yamagata, Japan, in 1984. She received B.E. degree from Yamagata University, Yamagata, Japan, in 2008. She is currently an engineer at Catalog Inc. Her research interests are speech recognition and speaker characteristics.

Takashi Ito was born in Yamagata, Japan, in 1985. He received B.E. degree from Yamagata University, Yamagata, Japan, in 2008. Since 2008, he is currently a master's degree student at Graduate School of Science and Engineering, Yamagata University. His research interest is speech recognition. He is a member of the ASJ.

Masaharu Kato was born in Osaka, Japan, in 1969. He received B.E., and M.E. degrees from Yamagata University, Yamagata, Japan, in 1991 and 1993 respectively. Since 1993, he has been working at Yamagata University, and he is currently a Research Associate at Graduate School of Science and Engineering, Yamagata University. His research interest is speech recognition. He is a member of the ASJ.

Masaki Kohda was born in Aichi, Japan, in 1942. He received B.E., M.E. and Ph.D. degrees from Nagoya University, in 1965, 1967 and 1979 respectively. In 1967, he joined Electrical Communication Laboratories of Nippon Telegraph and Telephone Corporation, Tokyo, Japan. Since 1987, he had been working at Yamagata University, and he is currently a professor emeritus at Yamagata University. His research interests are speech recognition, speaker recognition, speech synthesis, speech coding and language modeling. He is a member of the Information Processing Society of Japan, the Acoustical Society of Japan, the Japanese Society for Artificial Intelligence and the Association for Natural Language Processing.