Proceedings of 20th International Congress on Acoustics, ICA 2010

23–27 August 2010, Sydney, Australia

Rapid Acoustic Model Adaptation Using Inverse MLLR-based Feature Generation

Arata ITO (1), Sunao HARA (1), Norihide KITAOKA (1) and Kazuya TAKEDA (1)

(1) Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya-shi, 464-8903 Japan

ABSTRACT

We propose a technique for generating a large amount of target speaker-like speech features by converting a large amount of prepared speech features of many speakers into features similar to those of the target speaker using a transformation matrix. To generate a large amount of target speaker-like features, the system only needs a very small amount of the target speaker's utterances. This technique enables the system to adapt the acoustic model efficiently from a small amount of the target speaker's utterances. To evaluate the proposed method, we prepared 100 reference speakers and 12 target (test) speakers. We conducted the experiments in an isolated word recognition task using a speech database collected by real PC-based distributed environments and compared our proposed method with MLLR, MAP and the method theoretically equivalent to the SAT. Experimental results proved that the proposed method needed a significantly smaller amount of the target speaker's utterances than conventional MLLR, MAP and SAT.

INTRODUCTION

Speech recognition performance degrades because of many factors such as noisy environments, speaking styles, and individual difference. In particular, speaker-independent speech recognition under various environments, as in the case of PC-based distributed speech recognition systems, becomes very difficult. To solve this problem, acoustic model adaptation to the specific speaker and the environment by Maximum Likelihood Linear Regression (MLLR) [1], or normalization-based training and recognition by Speaker Adaptive Training (SAT) [2], are often used and are very effective. However, it is necessary for MLLR to prepare some quantity of utterances matched to the target speaker and the environment. This problem is the same for SAT and any other adaptation methods. Hereafter, we discuss only speaker adaptation, but the discussion can include environmental adaptation.

In this paper, we propose a technique to generate large enough amounts of target speaker-like speech features by converting a large amount of prepared speech features of many speakers (reference speakers) into features similar to those of the target speaker using a transformation matrix obtained by Constrained MLLR (CMLLR) [3, 4] technique. To generate a large amount of target speaker-like features, the system needs a very small amount of the target speaker's utterances. Using the target speaker-like features, we can adapt the acoustic model efficiently. When applying this method to all the reference speakers, the method is almost equivalent to SAT theoretically. However, we combine similar reference speaker selection (SRSS) with the method, which cannot be used in the conventional SAT framework. Using SRSS, only the features of speakers originally similar to the target speaker are used for the adaptation, which makes adaptation more efficient than using all reference speakers. Moreover, we combine a Maximum A Posteriori (MAP) [5] criterion with the method, which is not involved in the SAT framework. Using MAP, the proposed method provides a large improvement of recognition performance. To evaluate the proposed method, we prepared 100 reference speakers and 12 target (test) speakers. We compared our proposed method with MLLR, MAP and the method theoretically equivalent to the SAT in an isolated word recognition task using a speech database collected by real PC-based distributed environments.

FEATURE GENERATION BY CMLLR TECHNIQUE

Maximum Likelihood Linear Regression (MLLR) is a technique for linearly transforming acoustic model parameters using regression matrices estimated by a small amount of adaptation utterances. Constrained MLLR (CMLLR) is a special type of MLLR that transforms mean and variance parameters of HMMs using a common regression matrix. The CMLLR method is used as a transformation of the characteristic domain. The feature transformation formula is

$$\hat{\mathbf{o}}(t) = \mathbf{A}\mathbf{o}(t) + \mathbf{b} = \mathbf{W}\boldsymbol{\zeta}(t) \tag{1}$$

where $\mathbf{o}(t)$ is an original input vector at time t, $\hat{\mathbf{o}}(t)$ is the transformed vector at time t, $\mathbf{W} = \left[\, \mathbf{b}^{\mathrm{T}}\ \mathbf{A}^{\mathrm{T}} \right]^{\mathrm{T}}$ is the $n \times (n+1)$ transformation matrix (where n is the dimensionality of the data) which is estimated a priori for a specific speaker using ML criterion, and $\boldsymbol{\zeta}(t) = \left[\, 1\ \ \mathbf{o}(t)^{\mathrm{T}} \right]^{\mathrm{T}}$ is the extended input vector. It is thought that the input voice becomes a voice of the speaker who has the "average" voice of the voice used for the acoustic model training.

TARGET SPEAKER-LIKE SPEECH FEATURE GENERATION BASED ON INVERSE TRANSFORMATION

We explain how to transform speech features of a certain speaker i (a reference speaker) close to those of a target speaker X by using transformation matrix $\mathbf{W}$ of the CMLLR method. The transformation procedure of speech features using the CMLLR transformation matrix is shown below:

1. Estimate transformation matrix $\mathbf{W}_i$ from utterances of reference speaker i, who has enough utterances, and the transcriptions of the utterances are known.
2. Multiply transformation matrix $\mathbf{W}_i$ to speech features of speaker i. Then, speech features of speaker i become those of the "average" speaker. In this paper, we call this

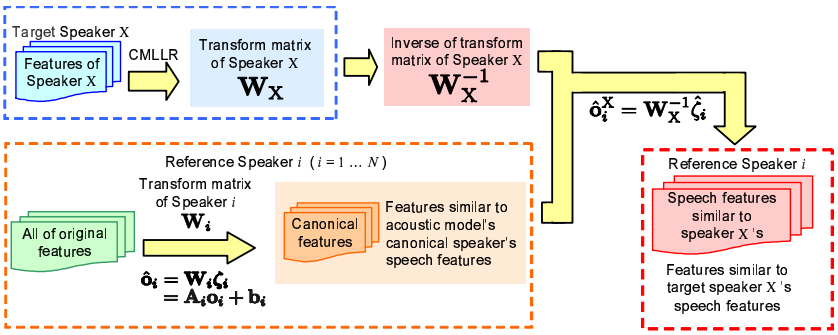

Figure 1: Target speaker-like feature generation method

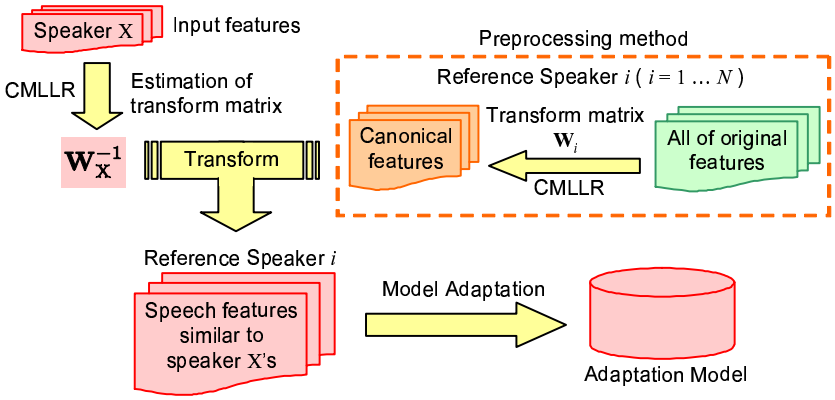

Figure 2: Adaptation method based on target speaker-like feature generation

procedure “canonicalization” and the features after the procedure a “canonical feature”.

3. Multiply the inverse matrix $\mathbf{W}_X^{-1}$ to the speech features of speaker i. $\mathbf{W}_X$ is the transformation matrix for the target speaker X ($X \neq i$) obtained as procedure 2. Then, speaker i's speech features are transformed into speech features similar to the target speaker X's. We call these transformed speech features “Speaker X-like speech features” of speaker i.
4. A large amount of “Speaker X-like speech features” can be obtained by doing these procedures for reference speaker $i = 1, \cdots, N$.

Figure 1 shows the outline of this procedure. The number of regression classes in CMLLR is one. We call the features applying the procedures “canonicalized” features. It is assumed that the transcription of every reference speaker's utterance is not changed after the transformation.

MODEL ADAPTATION BY USING SIMILAR SPEAKER-DEPENDENT SPEECH FEATURE

Figure 2 is an outline of the model adaptation technique by the generated target speaker's speech features. When the target speaker's input voice is obtained, the reference speakers' voices are converted by the transformation method explained above, and adaptation models are generated by performing model adaptation by generated “Speaker X-like speech features”. This adaptation can always be supervised because transcriptions of the reference speakers' utterances are already known.

EXPERIMENTAL DATA AND CONDITIONS

We use a real environmental speech database collected by using a voice interactive music search engine *MusicNavi2* [6]. Utterances in the MusicNavi2 database are collected by way of a WEB interface on users' PCs. The speakers in the MusicNavi2 database are numerous and consist of various kinds of persons using a variety of PCs and microphones under various environments. Most of the utterances are isolated word utterances and all of the utterances have been transcribed manually. The word dictionary used by the recognition experiment was made from the text of the transcription of the utterances. That is, the unknown word did not exist in the recognition experiment. The vocabulary size was about 8,000. The number of reference speakers was 100 (50 males and 50 females). The average number of utterances of each reference speaker was 175, and the average utterance length of each utterance was 1.627 s. The number of target (test) speakers was 12 (6 males and 6 females), and each of them had an average of 150 utterances. The baseline acoustic models were speaker-independent triphone HMMs (“CSJ models”) that had been trained from the Corpus of Spontaneous Japanese (CSJ) [7]. HMMs have 3,000 states, each of them with 16 mixtures. The feature vectors used had 38 dimensions (12 MFCC, 12 ΔMFCC, 12 ΔΔMFCC, Δpower, and ΔΔpower) in total. The recognition decoder used was Julius-4.1 [8].

Speech features of 100 reference speakers were adapted to each target speaker as a preprocessing by the proposed method.

EXPERIMENT

CONDITION 1: EVALUATION OF ADAPTATION MODEL BY TRANSFORMED SPEECH FEATURES

The proposed adaptation models were made by using speech features generated by the proposed method, and the word recognition experiment was performed. This experiment was supervised; that is to say, the transformation matrix for transformation from “average” features to “target” features was estimated by using the correct transcriptions of the target speaker's utterances made manually. The model adaptation method was

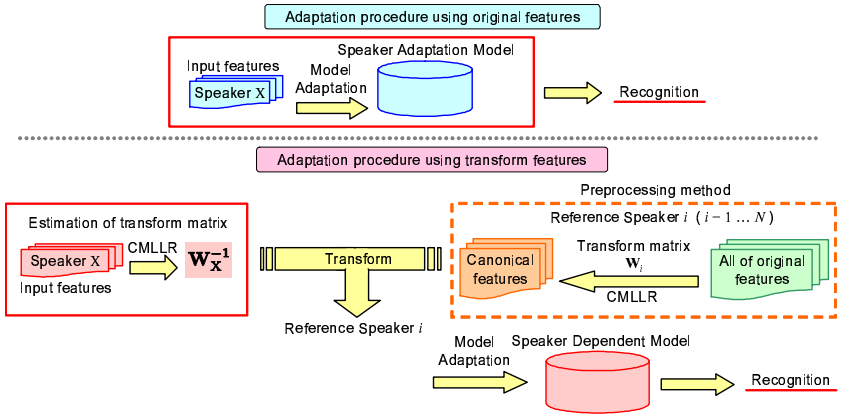

Figure 3: Conventional (top) and proposed (bottom) adaptation procedures

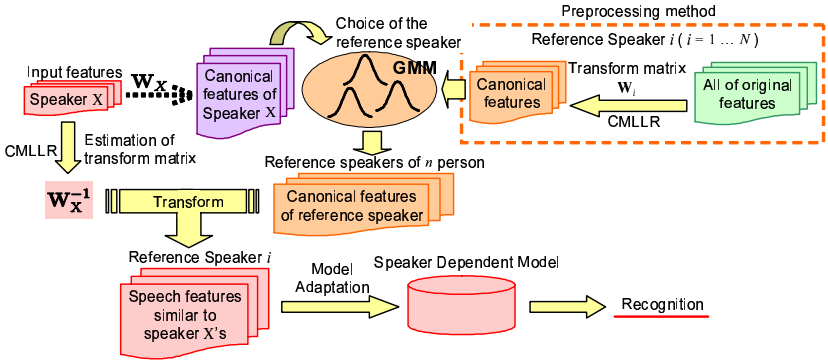

Figure 4: Adaptation procedure using reference speakers selected by GMM

supervised MLLR (the number of regression classes was 32). The model adaptation with speech features generated by the proposed method is always supervised using the transcriptions of utterances *a priori*. For example, when we make adaptation models for target speaker 1, the speech features of the reference speaker converted by the transformation matrix estimated from all the utterances of target speaker 1 ("Target speaker 1-like features") are used for adaptation by MLLR. For comparison, we also made environmentally adapted models as a baseline by MLLR using original speech features of reference speakers. Adaptation data were all speech features of reference speakers not converted as target speaker-like; that is, it is thought of as task and environment adaptation for PC-based music retrieval. We performed recognition experiments with 1: CSJ models, 2: environmentally adapted models (ENV models), and 3: speaker-adapted models by the proposed method. In the proposed method, all the utterances of each target speaker were used for estimating the transformation matrix.

CONDITION 2: EVALUATION OF MODEL ADAPTATION BY SMALL AMOUNT OF TARGET SPEAKER UTTERANCES

Supervised adaptation

Here, the supervised adaptation means that the manual transcriptions of the input utterances are used. We compare two types of adaptation models. The first is an adaptation model adapted by speech features converted by transformation matrix. The second is a speaker adaptation model adapted by input utterances, which is the conventional method. We investigated the performance change according to the number of input utterances and compared the word accuracies of these two models. Figure 3 shows the outlines of those experiments. We

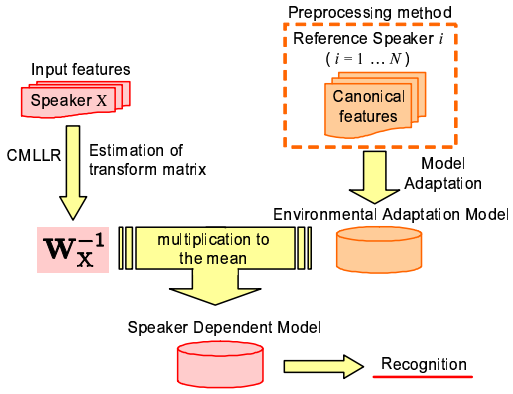

Figure 5: Conversion of HMM parameter by transformation matrix

used MLLR (number of regression classes was 32) or MAP as model adaptation methods. The baseline model is the CSJ model and environmental adaptation model (ENV model). As adaptation data, we used 1, 3, 5, 10, 30, 50, and 80 utterances from the beginning of the time series of the utterances recorded when each input speaker actually used the system. As for testing, we used 50 utterances from the end of each input speaker's time series.

Unsupervised adaptation

The unsupervised adaptation means that the transcriptions of the input utterances are unknown. Thus, recognition results are used as the transcription of the input utterances. Only the parts indicated by the red boxes in Fig.3 are unsupervised. Other ex-

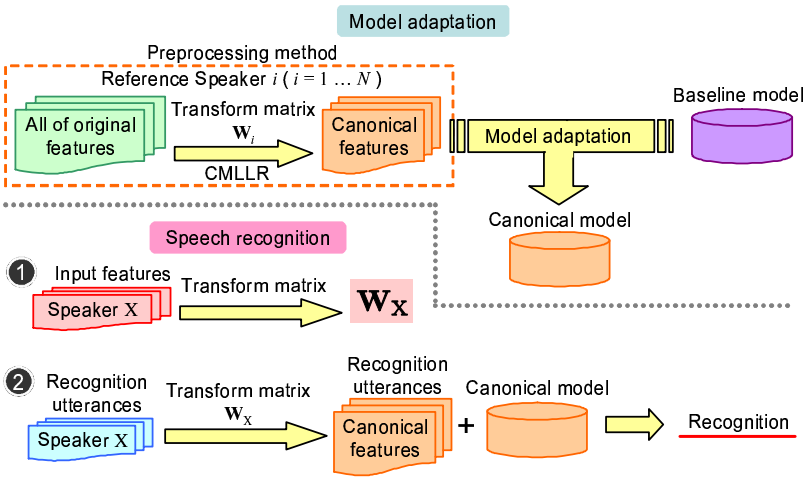

Figure 6: Experiment procedure using SAT-like model

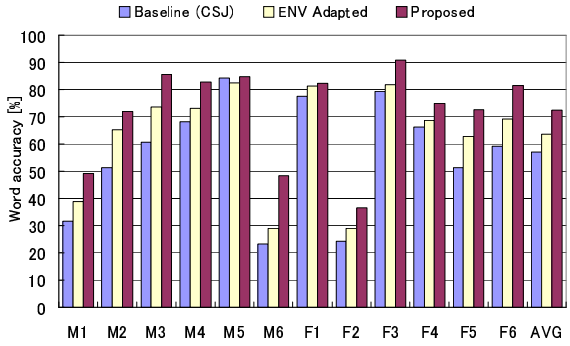

Figure 7: Comparison among original, environmentally adapted, and speaker-model adapted by proposed method

perimental conditions are all the same as the previous section. This means that the model adaptation of our proposed method is performed in a supervised manner.

CONDITION 3: SIMILAR REFERENCE SPEAKER SELECTION (SRSS) BY GMM

We adopt a similar speaker selection method to our proposed method. We modeled reference speakers by Gaussian Mixture Models (GMMs). Then, N reference speakers similar to the input (target) speaker are selected from 100 reference speakers by GMM likelihoods to the input utterances and only these N speakers are used for adaptation. Each GMM was made using the canonical speech features of each reference speaker, and the number of mixtures was 128. This similar speaker selection was performed based on the likelihood, and the top $N(=10)$ reference speakers with the N largest likelihoods were selected for generating the input speaker-like features. Figure 4 shows the outline of this method. The test speaker-like speech features were generated from 10 reference speakers to make the adaptation models. We used only the CSJ model as a baseline and MLLR as the model adaptation method.

CONDITION 4: CONVERSION OF HMM PARAMETERS BY TRANSFORMATION MATRIX

We test the inverse transformation of the acoustic model's mean parameters by inverse CMLLR transformation matrix. Figure 5 shows the outline of this experiment. The adaptation models were made from baseline models by using all canonicalized speech features of the reference speaker by MLLR. The ENV model was used as a baseline model.

CONDITION 5: EVALUATION OF SAT-LIKE MODEL

We test the speech recognition experiment using acoustic models that were adapted to canonicalized speech features of the reference speaker. Figure 6 shows the outline of this experiment. The adaptation models were made from baseline models by using all canonicalized speech features of the reference speaker by MLLR. When speech recognition was done, the recognition utterances canonicalized by using the transformation matrix estimated from the input (target) speaker's utterances were used. This experiment can be considered simply as Speaker Adaptive Training (SAT) because of using the acoustic models that were adapted to canonicalized speech features of the reference speaker beforehand. The CSJ model and ENV model were used as baseline models.

RECOGNITION RESULT

Figure 7 shows the result of CONDITION 1. The vertical axis is average word accuracy, and the horizontal axis is target speakers. The left bar, the center bar, and the right bar show the results by the CSJ models, the models adapted to original speech features by MLLR, and the models adapted to transformed speech features by MLLR (Proposed), respectively.
From the results, we know that word accuracies of all target speakers were improved by the proposed adaptation method. The models by the proposed method improved the word accuracy by about 10 points in the average word accuracy of 12 target speakers compared with the CSJ model. This resulted because the proposed method generated enough speech features of target speaker-like features from the reference speaker's features.
Figures 8 and 9 show the results of CONDITIONS 2 and 3 in supervised and unsupervised fashions, respectively. Figures 10,11,12, and 13 show the results of CONDITIONS 2 and 4. We used the CSJ model and ENV model in supervised and unsupervised adaptation as explained in the captions. The vertical axis is average word accuracy, and the horizontal axis is number of input utterances.
A common tendency by both MLLR and MAP in the proposed method was that the word accuracy decreased when the input utterance was 1, whereas when 3, 5, and 10 utterances were inputted, the word accuracy was improved by the proposed method more than the conventional speaker adaptation models and the baseline models. These were also common in both supervised and unsupervised adaptation. However, the perfor-

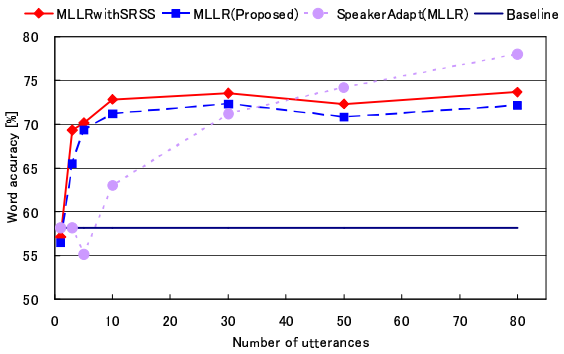

Figure 8: Result of SRSS with supervised adaptation according to number of input utterances (baseline (blue line): CSJ model)

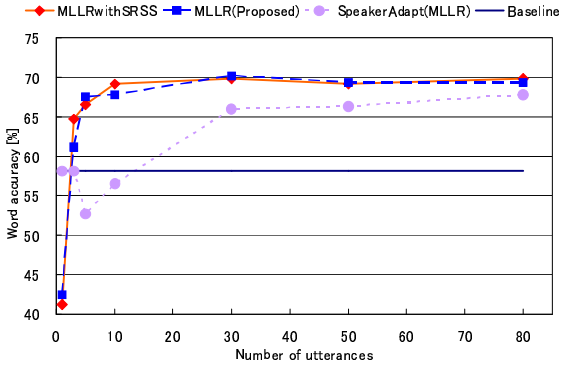

Figure 9: Result of SRSS with unsupervised adaptation according to number of input utterances (baseline (blue line): CSJ model)

mance of the conventional speaker adaptation models was better than proposed, as the number of utterances increased to 30, 50, and 80 because the performance of the adaptation models by the proposed method was saturated. It is thought that the transformation matrix did not express the target speaker's characteristics well with one utterance, and the estimation of the transformation matrix was accurate as the number of input utterances increased. In addition, the word accuracy was improved by the proposed method faster than the conventional speaker adaptation of MLLR and MAP. The performance of unsupervised adaptation by the proposed method was a little lower than that of totally supervised adaptation, but the performance of the proposed method was higher than that of the conventional speaker adaptation. Influence of false recognition became small by acoustic models adapted to the reference speakers' utterances because the adaptation could be supervised at any time even if the recognition of the input utterance failed and estimation of the transformation matrix became unstable. That is to say, the proposed method was effective in the case of a small amount of input utterances.

The recognition performance was improved by selecting a similar reference speaker by GMM. Target speaker's characteristic was expressed better by using the speakers who were selected by SRSS.

The recognition performance of MAP improved more than that of MLLR. This was because the influence of speech features converted into the "incorrect" direction at the feature conversion by unstable estimation of the CMLLR transformation matrix was reduced by MAP. The speech feature conversion of the proposed method converted all of the reference speaker's utterances by one transformation matrix estimated from the target speaker's utterances. There were speakers greatly different from the target speaker and the features converted by one transformation matrix may be far from those of the target speaker. MAP uses the prior distribution of parameters; thus the model parameters are updated considering the priors. That is, the mean parameters of HMMs become close to the observation when the distance between observation and the prior is small, and the parameters remain close to the parameters of the base models when the distance is large. That is, in MAP, the mean parameters of HMMs become close to the parameters of the baseline models for the data that emerge by feature transformation, and the influence of the outliers can be reduced.

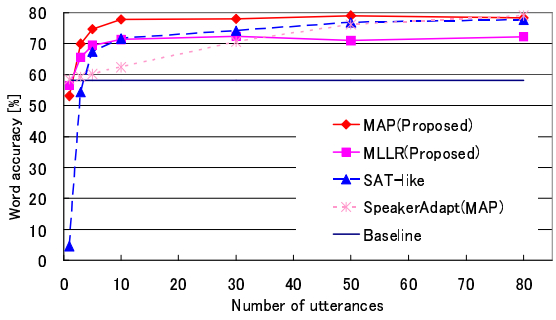

Figure 10: Result of supervised adaptation (baseline (blue line): CSJ model)

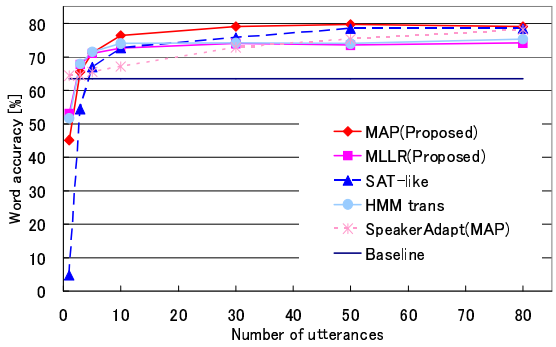

Figure 11: Result of supervised adaptation (baseline (blue line): ENV model)

From Figs.10,11,12, and 13, there was no difference in comparison among the models with transformed parameters of HMMs by the transformation matrix directly ("HMM trans") and MLLR adaptation models by using transformed features of the proposed method ("MLLR(Proposed)"); however, the performance of MAP adaptation models by using transformed features of the proposed method ("MAP(Proposed)") was better than "HMM trans". Unnecessary parameters were transformed because "HMM trans" linearly transformed all the model parameters by one transformation matrix. On the other hand, parameters that were not necessary to update were not updated because each model parameter was updated in MAP according to the adaptation data. The performance difference between "HMM trans" and "MAP(Proposed)" arises from this difference.

DIFFERENCE BETWEEN PROPOSED METHOD AND SAT

In the recognition experiment using Speaker Adaptive Training (SAT)-like models ("SAT-like"), the speech features for recognition were canonicalized by the transformation matrix estimated using input speech features. This processing can be expressed by the following formula.

$$\mathcal{N}(\mathbf{Ao}+\mathbf{b};\boldsymbol{\mu},\boldsymbol{\Sigma}) = |\mathbf{A}^{-1}|\mathcal{N}(\mathbf{o};\mathbf{A}^{-1}(\boldsymbol{\mu}-\mathbf{b}),\mathbf{A}^{-1}\boldsymbol{\Sigma}\mathbf{A}^{-1^{\mathrm{T}}})$$

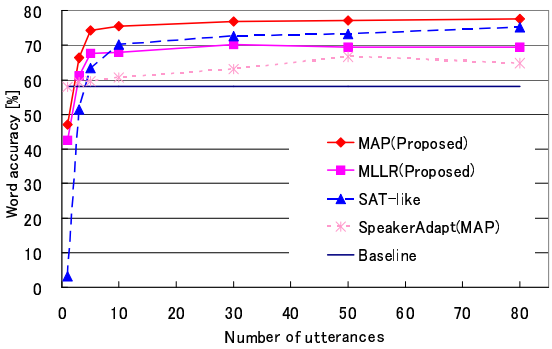

Figure 12: Result of unsupervised adaptation (baseline (blue line): CSJ model)

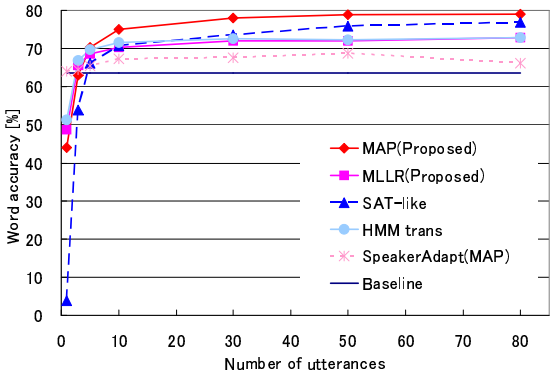

Figure 13: Result of unsupervised adaptation (baseline (blue line): ENV model)

$\mathcal{N}(\mathbf{o};\boldsymbol{\mu},\boldsymbol{\Sigma})$ represents a Gaussian density that has mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. That is, canonicalization of the speech features for recognition is equivalent to the transformation of the mean and variance of the acoustic model. When the number of input utterances was small, the performance of "SAT-like" had decreased because a large amount of model parameters was converted by a small amount of information, whereas as the number of input utterances increases, the performance of "SAT-like" had improved because the estimation accuracy of the transformation matrix improved. For this reason, the performance of "SAT-like" was better than that of "MLLR(Proposed)" and "HMM trans".
The proposed method constructs the specific speaker models by using existing acoustic models. Speakers similar to the target speaker can be prepared from an existing database in the proposed method, and the target speaker's characteristic can be expressed better by using a similar speaker's voice. Moreover, it is also possible to use not only MLLR but also MAP as the model adaptation technique. The acoustic model's performance can be improved by using the similar reference speaker selection (SRSS) and MAP.

CONCLUSION

In this paper, we focused on the transformation matrix of MLLR. We proposed a technique to generate target speaker-like speech features by inversely transforming canonicalized speech features of other speakers by the target speaker's transformation matrix, and proposed a model adaptation technique using generated speech features. To evaluate the proposed method, the models were adapted by MLLR and MAP, and we tested them on isolated word speech recognition using real environmental voices. As a result of the experiments, the word accuracy was improved by the transformed speech features. The proposed method provided robust model adaptation when a small amount of input utterances was obtained. Moreover, similar reference speaker selection (SRSS) and MAP, which are not usually adopted in the SAT framework, can be combined in the proposed method. A bigger performance improvement was achieved by combining SRSS and MAP.
We will work on evaluation of the proposed method with a different database in the future.

REFERENCES

[1] M. J. F. Gales. Maximum likelihood linear transformations for hmm-based speech recognition. *Computer Speech and Language*, 12:75–98, 1998.

[2] Tasos Anastasakos, John Mcdonough, Richard Schwartz, and John Makhoul. A compact model for speaker-adaptive training. *in Proc. ICSLP*, pages 1137–1140, 1996.

[3] M.J.F. Gales and P.C. Woodland. Mean and variance adaptation within the mllr framework. *Computer Speech and Language*, 10:249–264, 1996.

[4] S. Young, J. Odell, D. Ollason, and P. Woodland. The htk book, (for htk version 3.4), chapter9. pages 136–147, 2006.

[5] J.-L. Gauvain and C.-H. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Trans. Speech Audio Process.*, 2:291–298, 1994.

[6] Sunao HARA, Chiyomi MIYAJIMA, Katsunobu ITOU, and Kazuya TAKEDA. Data collection system for the speech utterances to an automatic speech recognition system under real environments. *IEICE trans.*, J90-D(10):2807–2816, 2007.

[7] The national institute for japanese language. http://www.kokken.go.jp/en/.

[8] Julius. http://julius.sourceforge.jp/.