PAPER

Acoustic Model Training Using Pseudo-Speaker Features Generated by MLLR Transformations for Robust Speaker-Independent Speech Recognition

Arata ITOH†*, *Nonmember*, Sunao HARA†**, Norihide KITAOKA†a), *Members*, *and* Kazuya TAKEDA†, *Fellow*

SUMMARY A novel speech feature generation-based acoustic model training method for robust speaker-independent speech recognition is proposed. For decades, speaker adaptation methods have been widely used. All of these adaptation methods need adaptation data. However, our proposed method aims to create speaker-independent acoustic models that cover not only known but also unknown speakers. We achieve this by adopting inverse maximum likelihood linear regression (MLLR) transformation-based feature generation, and then we train our models using these features. First we obtain MLLR transformation matrices from a limited number of existing speakers. Then we extract the bases of the MLLR transformation matrices using PCA. The distribution of the weight parameters to express the transformation matrices for the existing speakers are estimated. Next, we construct pseudo-speaker transformations by sampling the weight parameters from the distribution, and apply the transformation to the normalized features of the existing speaker to generate the features of the pseudo-speakers. Finally, using these features, we train the acoustic models. Evaluation results show that the acoustic models trained using our proposed method are robust for unknown speakers.

key words: *speech recognition, acoustic model training, pseudo speakers, feature generation, MLLR*

1. Introduction

In this paper, an acoustic model training method is proposed for robust, speaker-independent speech recognition using limited speech resources. Degradation of speech recognition performance is often due to a mismatch between model training and test conditions. There are many reasons for such mismatches: differences between individual speakers, recording equipment issues, background noise, etc. To compensate for such mismatches, adaptation techniques are often used [1]. Model-based adaptation methods, such as maximum a posteriori (MAP) adaptation [2] and maximum likelihood linear regression (MLLR) [3], transform acoustic models (usually hidden Markov models (HMMs)) to fit the target speaker or environment. These techniques, however, need a certain amount of adaptation data to estimate the parameters of the models.

Speaker adaptive training (SAT) [4] has also been proposed. In SAT, training data are normalized to a virtual *canonical* speaker, for whom the acoustic models are trained. In the recognition stage, adaptation parameters are estimated for the input speech and the models are adapted. This method achieves efficient adaptation.

Adaptation techniques which only need a small amount of target speech data, such as those using inter-speaker variation modeling like Eigenvoice [5], have also been proposed. In this framework, the super vectors of the mean parameters of the speaker-dependent acoustic models are used as bases, and the super vector of the new speaker-specific acoustic models is expressed as a linear combination of these bases. Eigenvoice only needs a small amount of target speech, because variation in speech and environment are expressed in a low-dimensional sub-space. Eigen-MLLR, which is a combination of MLLR and eigenvoice, was proposed in [6]. Principal component analysis (PCA) is applied to the MLLR transformation matrices to obtain bases, and then a new speaker's MLLR matrix is expressed as a linear combination of the matrices.

All of these adaptation methods, however, need adaptation data. We can only use a limited amount of speech data from the environment where the system is to be used, however, because the cost of collecting data in realistic environments is very high. We believe this assumption is realistic during the early use of such a speech application.

In this paper, we propose a novel speech feature generation-based speaker-independent model training method to compensate for the variation which occurs when using limited speech data resources. We do this by reversing the concept of adaptation. In the proposed method, we do not *remove* speaker variations but *add* them to the averaged speech features [7]. We assume that individual speech variation can be generated by adding individual differences to an "average" person. Speaker recognition using the MLLR transformation matrix [8] suggests that the linear transformation matrix can express individuality. We first obtain the MLLR transformation matrices from the speech data of a limited number of environmentally matched speakers and apply PCA to it to extract bases. We then construct pseudo-speaker transformation matrices from the statistical linear combination of the bases. Finally, speech features are generated by applying the constructed transformation matrices to the normalized speech features obtained from real speech

Manuscript received December 26, 2011.
Manuscript revised April 24, 2012.

†The authors are with Nagoya University, Nagoya-shi, 464–8603 Japan.
*Presently, with DENSO CORPORATION, Japan.
**Presently, with the Nara Institute of Science and Technology, Japan.
a) E-mail: kitaoka@nagoya-u.jp

DOI: 10.1587/transinf.E95.D.2479

and then used to train the speaker-independent (but environment adapted) acoustic models. Using this technique, we can easily obtain a huge amount of speech variations from a limited number of speakers in the target environments and make the acoustic models effective, despite inter-speaker variations [9].

The rest of this paper is organized as follows. Section 2 first outlines the MLLR adaptation method and then explains speech feature generation based on MLLR transformations. Section 3 discusses the experimental results of the proposed method and compares it with adaptation-based methods and speaker adaptive training. Section 4 concludes the paper and describes future work.

2. Acoustic Model Training Based on Feature Generation Using MLLR Transformations for Pseudo-Speakers

Our proposed method consists of the following steps:

- estimation of the MLLR transformation matrices of speaker utterances recorded in the target environments;
- extraction of the bases of the MLLR transformation matrices;
- estimation of the basis weight distributions;
- construction of the pseudo-speaker's transformation matrix, and speech-feature generation by applying the transformation matrix to the speaker-normalized speech data;
- acoustic model training with the generated features.

The flow of the proposed method is summarized in Fig. 1. Here, we assume that we can use a certain amount of training data in the target environments, but the data do not include the test speakers. This assumption is reasonable because we are only able to collect a small amount of data in the environment where the application will actually be used when a new application has just been developed.

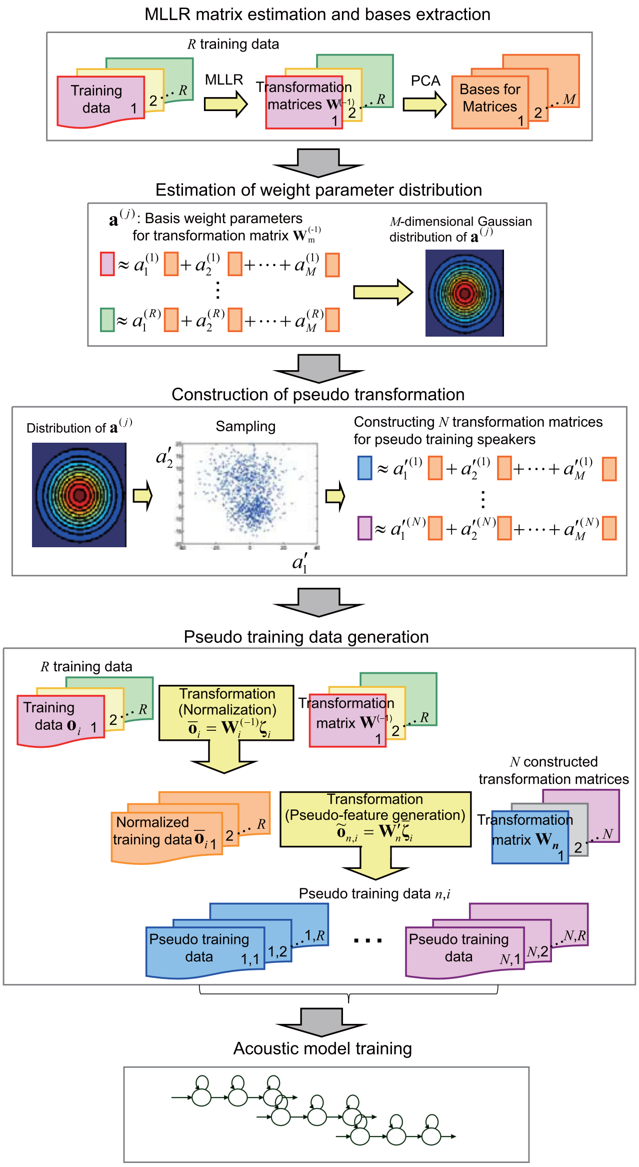

MLLR matrix estimation and bases extraction
R training data
Training data 1
2
R
MLLR
Transformation matrices $\mathbf{W}^{(-1)}$ 1
2
R
PCA
Bases for Matrices 1
2
M
Estimation of weight parameter distribution
$\mathbf{a}^{(j)}$: Basis weight parameters for transformation matrix $\mathbf{W}_m^{(-1)}$
$\approx a_1^{(1)} + a_2^{(1)} + \cdots + a_M^{(1)}$
$\approx a_1^{(R)} + a_2^{(R)} + \cdots + a_M^{(R)}$
M-dimensional Gaussian distribution of $\mathbf{a}^{(j)}$
Construction of pseudo transformation
Distribution of $\mathbf{a}^{(j)}$
Sampling
a'_2
a'_1
Constructing N transformation matrices for pseudo training speakers
$\approx a_1'^{(1)} + a_2'^{(1)} + \cdots + a_M'^{(1)}$
$\approx a_1'^{(N)} + a_2'^{(N)} + \cdots + a_M'^{(N)}$
Pseudo training data generation
R training data
Training data $\mathbf{o}_i$ 1
2
R
Transformation (Normalization) $\bar{\mathbf{o}}_i = \mathbf{W}_i^{(-1)}\zeta_i$
Transformation matrix $\mathbf{W}^{(-1)}$ 1
2
R
N constructed transformation matrices
Normalized training data $\bar{\mathbf{o}}_i$ 1
2
R
Transformation (Pseudo-feature generation) $\tilde{\mathbf{o}}_{n,i} = \mathbf{W}'_n\zeta_i$
Transformation matrix $\mathbf{W}_n$ 1
2
N
Pseudo training data n,i
Pseudo training data 1,1
1,2
1,R
Pseudo training data N,1
N,2
N,R
Acoustic model training

Fig. 1 Flow of the proposed feature generation-based acoustic model training.

2.1 Normalization of Training Speech

In MLLR, the mean vectors of the Gaussian distributions in the HMMs, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$, are adapted to a specific speaker by transformation:

$$\hat{\boldsymbol{\mu}} = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \tag{1}$$

where n is the dimension of a feature vector, $\mathbf{A}$ is a $n \times n$ matrix, and $\mathbf{b}$ is an n-dimensional vector. This equation can be rewritten as follows:

$$\hat{\boldsymbol{\mu}} = \mathbf{W}\boldsymbol{\xi}, \tag{2}$$

where $\boldsymbol{\xi} = (1, \mu_1, \ldots, \mu_n)^{\mathrm{T}}$. $\mathbf{W} = [\mathbf{b}\ \mathbf{A}]$ is an $n \times (n+1)$ matrix and can be estimated by the ML criterion using the EM algorithm with the auxiliary function [10]:

$$Q(\mathbf{W}, \bar{\mathbf{W}}) = K - \frac{1}{2}\sum_{m=1}^{M}\sum_{t=1}^{\mathrm{T}} \gamma_m(t)[K_m + \log|\boldsymbol{\Sigma}_m| + (\mathbf{x_t} - \mathbf{W}\boldsymbol{\xi}_m)'\Sigma_m^{-1}(\mathbf{x_t} - \mathbf{W}\boldsymbol{\xi}_m)], \tag{3}$$

where $\gamma_m(t)$ is the posterior probability of mixture component m at time t, K is a constant dependent only on the transition probabilities, and K_m is the normalization constant associated with Gaussian mixture component m, when given the adaptation data for a certain speaker (and/or environment) $\chi = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$.

We can consider the transformation matrix $\mathbf{W}$ as the expression of a speaker in an environment. We can then use the transformation inversely to transform a speaker-specific feature vector to the "average" speaker's feature vector most "suitable" for the speaker-independent models:

$$\bar{\mathbf{o}} = \mathbf{A}^{-1}\mathbf{o} - \mathbf{A}^{-1}\mathbf{b} = \mathbf{W}^{(-1)}\zeta, \tag{4}$$

where $\mathbf{o}$ and $\bar{\mathbf{o}}$ express an n-dimensional input feature vector and a normalized one, $\mathbf{W}^{(-1)} = \left[\, -(\mathbf{A}^{-1}\mathbf{b})\ \mathbf{A}^{-1} \right] \in \mathbf{R}^{n\times(n+1)}$ is a transformation matrix, and $\boldsymbol{\zeta} = \begin{bmatrix} \mathbf{1} & \mathbf{o}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} \in \mathbf{R}^{(n+1)\times 1}$ is an extended feature vector including a bias.

We obtain transformation matrix $\mathbf{W}_i^{(-1)}, (i = 1, \cdots, R)$ for speaker i of R speakers in the training data.

2.2 Basis Extraction Using PCA

We assume that transformation matrix $\mathbf{W}^{(-1)}$ consists of a linear combination of bases. One could use all the $\mathbf{W}_i^{(-1)}, (i = 1, \cdots, R)$ as bases. However, speech production is constrained by physical limitations such as vocal tract length. Such constraints should be reflected in the range of individual differences in the transformation matrix.

Thus, we apply PCA to the $n \times (n+1)$-dimensional R super vectors $\mathbf{V}_i (i = 1, \cdots, R)$, which are the concatenations of the columns in $\mathbf{W}_i^{(-1)}$s, and obtain M eigen vectors $\mathbf{V}_E^{(m)} (m = 1, \cdots, M)$ with the largest M eigenvalues as bases. This means that the transformation expressing individual differences is constrained as a linear combination of the basis super vectors, and that basis extraction is a blind estimation of the axes expressing speaker variation.

2.3 Estimation of Distribution of Weight Parameters

Using the bases extracted in the previous section, we express the individuality of a certain speaker $\mathbf{V}_j = \mathbf{a}^{(j)\mathrm{T}}(\mathbf{V}_E^{(1)}, \cdots, \mathbf{V}_E^{(M)})$, where $\mathbf{a}^{(j)} = (a_1^{(j)}, \cdots, a_M^{(j)})^{\mathrm{T}}$. We estimate the distribution of $\mathbf{a}^{(j)}$.

Each training speaker's super vector, $\mathbf{V}_i$, derived from the transformation matrix $\mathbf{W}_i^{(-1)}$ is approximated by a linear combination of $\tilde{\mathbf{V}}_i = \mathbf{a}^{(i)\mathrm{T}}(\mathbf{V}_E^{(1)}, \cdots, \mathbf{V}_E^{(M)})$. The weight vector $\mathbf{a}^{(i)}$ is obtained by the square error minimization criterion. With $\mathbf{a}^{(i)}$s for some training speakers, and an assumption of a type of distribution of $\mathbf{a}^{(j)}$, we can estimate the distribution parameters. We assume that $\mathbf{a}^{(j)}$ is distributed as an M-dimensional Gaussian.

2.4 Speech Feature Generation by MLLR Transformations for Pseudo-Speakers

Once we obtain the distribution of $\mathbf{a}^{(j)}$, we randomly pick N samples, $\mathbf{a}'^{(n)}$, $(n = 1, \cdots N)$, from the distribution. Using $\mathbf{a}'^{(n)}$, we construct N inverse transformations, $\mathbf{W}'^{(-1)}_n = \begin{bmatrix} -\left(\mathbf{A}'^{-1}_n \mathbf{b}'_n\right) & \mathbf{A}'^{-1}_n \end{bmatrix}$, by linear combination of the bases weighted by $\mathbf{a}'^{(n)}$. Then we obtain the transformation $\mathbf{W}'_n = [\mathbf{b}'_n \; \mathbf{A}'_n]$

Each constructed transformation $\mathbf{W}'_n$ corresponds to speaker characteristics of a pseudo-speaker n. We reverse the SAT technique [4] by applying the transformation to the normalized speech features to obtain a variety of speakers. We first apply the normalization matrix for training speaker i, $\mathbf{W}_i^{(-1)}$, to the speech features of speaker i and then apply the constructed transformation, $\mathbf{W}'_n$, to them to generate the speech features of pseudo-speaker n:

$$\tilde{\mathbf{o}}_{n,i} = \mathbf{A}'_n \bar{\mathbf{o}}_i + \mathbf{b}'_n \tag{5}$$

$$= \mathbf{W}'_n \bar{\boldsymbol{\zeta}}_i, \tag{6}$$

$$\bar{\mathbf{o}}_i = \mathbf{A}_i^{-1} \mathbf{o}_i - \mathbf{A}_i^{-1} \mathbf{b}_i \tag{7}$$

$$= \mathbf{W}_i^{(-1)} \boldsymbol{\zeta}_i, \tag{8}$$

$$(i = 1, \cdots, R)$$

where $\tilde{\mathbf{o}}_{n,i}$ is a generated feature of speaker n from the utterance by training speaker i, and $\boldsymbol{\zeta}_i = \begin{bmatrix} \mathbf{1} & \mathbf{o}_i^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$ and $\bar{\boldsymbol{\zeta}}_i = \begin{bmatrix} \mathbf{1} & \bar{\mathbf{o}}_i^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$ are extended feature vectors of training speech uttered by speaker i before and after normalization, respectively. Note that speaker n, who is not included in the training data, is a constructed pseudo-speaker. Applying this procedure using the training speech of speakers $i = 1, \cdots, R$ and pseudo-speakers $n = 1, \cdots, N$, we can obtain a large amount of training data for the acoustic models. This pseudo-speaker data is obtained from the distribution of the original training speakers. If there are enough training speakers to estimate the "correct" acoustic models, the results should be better than results using our method. However, we assume that we cannot realistically generate enough data to train acoustic models using real speakers, and thus we try to "interpolate" or "extra-polate" the parameters of training speakers.

2.5 Training Acoustic Models Using Generated Speech

Finally, we use the feature vectors generated by the technique described in the previous section to train the acoustic models. The pseudo-speakers' utterances have the speech features of constructed utterances, and are not the actual utterances of many human speakers' in the target environments. As a result, the acoustic models are expected to be robust at recognizing the utternaces of unknown speakers.

3. Experiments

3.1 Experimental Conditions

We collected field speech data using the *MusicNavi2* [11] spoken dialog-based music retrieval system. This system obtains user utterances from the Internet using loss-less speech compaction. Many anonymous users use this system. Figure 2 shows an example of the dialog obtained through the system. The underlined utterances are those of the user. The average signal-to-noise ratio (SNR) was 20.4 dB. Most of the users used the system in their homes. The utterances were spontaneous, but they spoke them in a relatively careful manner.

For recognition, we used a word-loop grammar with a vocabulary including all the words in the test utterances. There were no unknown words. We randomly selected 50 males and 50 females as training speakers. Utterances spoken by each training speaker were used as the training data. Training set 1 was 10 utterances from each subject (100 × 10 = 1000 utterances), and training set 2 was 30 utterances from each subject (100 × 30 = 3,000 utterances).

USR: サイモン・アンド・ガーファンクル.
(“SIMON AND GARFUNKEL.”)
SYS: サイモン・アンド・ガーファンクルを検索しますか?
(Do you want to retrieve songs by “Simon and Garfunkel?)
USR: はい
(Yes)
SYS: 検索しています...60 曲見つかりました.
(Now retrieving songs ... 60 songs were found.)
アイ・アム・ア・ロック，明日に架ける橋，...
(“I am a rock,” “Bridge over troubled water,” ...)
USR: その曲.
(That one, please.)
SYS: 明日に架ける橋を再生します.
(Now playing “Bridge over troubled water.”)

Fig. 2 An exapmle of a dialog recorded by *MusicNavi2* system. Underlined are the user utterances used in the experiments.

Table 1 Experimental setup.

# Training speakers	100 (50 males and 50 females)
# Training utterances	Set 1: 1000 (10 per person, total 1597.6 [s]) Set 2: 3000 (30 per person, total 4876.6 [s])
# Test speakers	250 (160 males and 90 females) Exclusive with training sets
Amount of test data	12500 (50 uttr. per person)
Analysis	Pre-emphasis (coefficient = 0.97) 24 dimensional mel-filterbank
Features	12 MFCC (with CMN) + 12 ΔMFCC + 12 ΔΔMFCC +Δpower + ΔΔpower (Δ and ΔΔ derived from 5 frames)
Speech recognizer	Julius-4.1 [12]
Acoustic model structure	Gender-independent triphone HMM 3000 states, 16 mixtures per state
Language model	Word loop grammar
Dictionary	Words for MusicNavi2 (approx. 8000 words)

We first made MLLR matrices for real training speakers by applying global MLLR adaptation to 10 (Set 1) and 30 (Set 2) utterances by each speaker (that is to say, we used all the training data in set 1 and set 2). The seed models for speaker adaptation were trained using the Corpus of Spontaneous Japanese (CSJ) [13]. We used 967 lectures (a total of 228 hours) from the CSJ. The MLLR matrices obtained were used for both normalization of the real speakers' utterances and for basis extraction.

We used test utterances from 250 speakers (160 males and 90 females). Fifty utterances from each speaker were used as test data. The feature vector consisted of a 12-dimensional MFCC, their first and second derivatives, and the first and second derivatives of the power. The MFCCs were extracted using pre-emphasis with a coefficient of 0.97 followed by 24 dimensional mel-filterbank analysis. CMN was applied. Experimental setup conditions, including these, are summarized in Table 1. The CSJ models and proposed models received ML-training from flat-start models using an EM algorithm. The experimental setup shown in Table 1 was used for the original CSJ models, for the models trained using the proposed method, and for the MAP-adapted models described below.

Table 2 Relationship between cumulative proportions and number of bases.

Cumulative contribution ratio [%]		80	90	95
# of Bases	Training set 1	59	75	86
	Training set 2	56	73	84

For comparison, we performed MAP adaptation, which involves adaptation for the environment, and SAT, using all the training utterances.

3.2 Evaluation Results

3.2.1 Basis Extraction

We set the cumulative proportions to 80%, 90%, and 95% to extract the bases. The relationship between the cumulative contribution ratios and the number of bases is shown in Table 2. With a cumulative proportion of 80%, we need approximately half of the bases that are extracted from training speakers' utteraces.

3.2.2 Recognition Results

Using our proposed method, we generated 1,000 pseudo-speakers from the bases described in Table 2, using 600 real training speaker utterances randomly selected for each pseudo-speaker from the training data. These utterances were converted for each pseudo-speaker using the transformation method described above. The 600 randomly selected utterances were different for each pseudo-speaker. The weight distribution was estimated from 1,000 (set 1) or 3,000 (set 2) utterances, and the randomly selected utterances for a pseudo-speaker were transformed by a matrix constructed using the weight sampled from the distribution. Thus, we were able to obtain 600,000 training utterances, and we then used them to train the acoustic models[†]. For comparison, we also adapted the acoustic models trained using CSJ with the real training data described in Table 1. The average recognition rates and the standard deviations are shown in Table 3. The test speakers were all different from the training speakers, so the recognition rates in Table 3 can be seen as the recognition rates by environmentally adapted, speaker-independent acoustic models without any speaker adaptation. We also show the results using models trained using CSJ (CSJ) and MAP-adapted models (MAP).

The table shows that the larger the number of bases,

[†] A transformation for a pseudo-speaker obtained using our proposed method is not certain to express the same characteristics as a real human voice. Thus, too many pseudo-speakers may have an adverse effect on the statistics in the acoustic modeling. We decided the number of generated utterances from preliminary experiments. For a similar reason, the bases of the transformation should be extracted from some minimum number of real human voices. We preliminarily tested the bases extracted from 20 or 50 real training speakers, but the results were significantly inferior to the MAP models. We have to tackle the problem of creating more pseudo-speakers from fewer real speakers in the future.

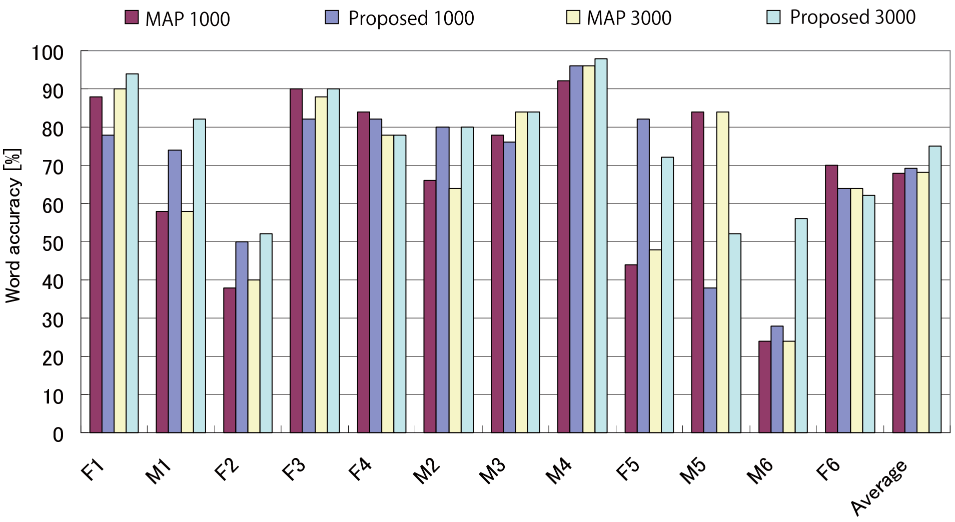


Fig. 3 Recognition rates for each speaker when using cumulative contribution ratio of 95%. M*n* and F*n* express male speaker number *n* and female speaker number *n*, respectively. MAP 1000 and Proposed 1000 are results using training set 1, and MAP 3000 and Proposed 3000 are results using training set 2.

Table 3 Recognition rates [%] and standard deviations of acoustic models trained using pseudo-speaker utterances, those adapted by MAP, and those trained using only CSJ. Note that the CSJ models did not use any training data and thus the recognition results for training sets 1 and 2 were identical.

Methods[†]		Proposed			MAP	CSJ
Cumu. contrib. ratio		80%	90%	95%		
Training set 1	Recog. rate	64.1	65.5	66.5	67.9	60.8
	Std. dev.	17.3	17.2	16.9	19.6	20.7
Training set 2	Recog. rate	70.5	70.7	70.8	69.2	60.8
	Std. dev.	16.4	16.4	16.7	19.5	20.7

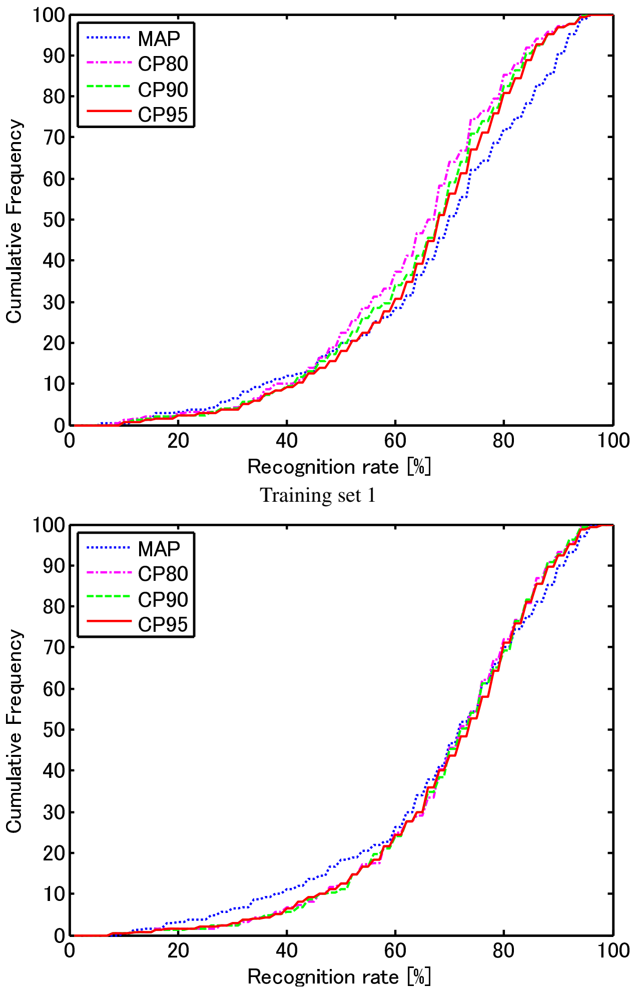
the better the recognition rates, especially regarding training set 1, which consisted of 10 utterances by 100 speakers. The recognition rate for training set 1 using the proposed method is comparable to the rate using MAP adaptation. Using training set 2, the proposed method outperforms MAP adaptation.

Note that the standard deviations of the recognition rates using the proposed method are smaller than those with MAP, suggesting that the acoustic models trained using our proposed method were robust for handling speaker variations. To see the tendency of the difference in performance between MAP and the proposed method, we examine the results of sample test speakers. Figure 3 shows the recognition rates for each test speaker randomly sampled from all the test speakers. We found that even recognition rates for speakers with good results using MAP degraded slightly, while recognition rates for most speakers with low recognition rates with MAP were improved. To confirm this observed tendency, we describe the result of all the test speakers in another way. Cumulative test speaker frequencies and recognition rates are shown in Fig. 4. The number of speakers with low recognition rates is significantly smaller using the proposed method than with MAP. This suggests that speaker construction using our method produces a wider range of speaker variations. For this reason, the proposed training feature generation method works robustly for unknown speakers, especially those with originally low recognition rates. Inversely, our method did not perform well with speakers who originally had high recognition rates. This may be the effect of pseudo-speakers which were very different from real human voices, with the result that the distributions in HMM states were broader than they needed to be. We can think of some possible reasons. A linear transformation for a pseudo human should be applied to a subset of real training data. That is to say, a constraint between n and i in Eqs. (5) and (6) should be applied. Furthermore, a linear transformation by itself may not be sufficient to express the characteristics of a human voice. Investigation of these issues are the subject of future work.

3.3 Comparison with SAT

The method proposed in this paper was inspired by SAT, in which all training speech is normalized by transformations such as inverse MLLR before being used for training. Models are then adapted to a specific speaker. Our method, however, generates speaker variation to train the acoustic models.

[†] It should be noted that we should have compared these results with an experiment in which we trained acoustic models using the original training data (that is, training using only set 1 or set 2), but we could not do this under the same conditions because of a lack of training data. We conducted preliminary experiments with a small size of acoustic models and discovered that even using training set 2 and HMMs with 500 states, our proposed method (75.5% recognition rate) outperformed HMMs trained using original data (71.1%), because of over-training. When using HMMs with 300 states, result using original data (72.2%) were comparable to results using our method (74.0%). (The test data in this footnote is a subset of that in Table 1)



Training set 1

Training set 2

Fig. 4 Cumulative speaker frequency and recognition rates. CP*n* expresses the cumulative contribution ratio of *n* in the proposed method. The nearer to the X-axis the line is, the better the recognition performance.

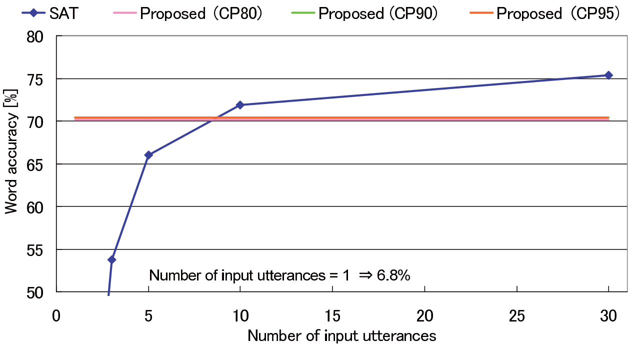


Fig. 5 Comparison of proposed method and SAT. Proposed method was performed using training data set 2. CP*n* expresses the cumulative proportion of *n* in the proposed method. The difference in *n* in CP*n* did not have a big effect on recognition performance, thus all the lines for the proposed methods with different CP*n* values overlap in the figure.

The crucial difference between SAT and our method is that SAT needs adaptation data for a specific speaker, but our method does not.

We compared the performance of SAT and our method. In the SAT framework, to make normalized acoustic models we adapted CSJ models described in Sect. 3.1 to the 3000 training utterances (training set 2) normalied by CMLLR, and we assume that the normalization parameters of the transformation matrix are estimated from 1, 3, 5, 10, and 30 input utterances in the test phase. Here, we transform the input features for normalization, not the models, which is theoretically identical to the model transformation using CMLLR [10]. The recognition results are shown in Fig. 5. In this figure, we used training set 2 for the proposed method and the results are identical to those using training set 2 in Table 3. SAT performs better with more than ten adaptation utterances, but our method performs well without adaptation data.

4. Conclusion

In this paper, we proposed a feature generation-based acoustic model training method. Linear transformations of features corresponding to pseudo-speakers were constructed by the linear combination of principal components of inverse MLLR transformation matrices for a limited number of training speakers. These transformations were applied to the normalized speech features of training speakers. Pseudo-speakers were expected to represent the speakers which were not included in the training data. Our method outperforms adaptation-based methods when the amount of training data in the test environments is limited, especially for speakers with low speech recognition rates.

In the future, we will use more real speech data to generate a huge amount of feature vectors in order to produce an accurate and robust acoustic model. Currently, we only use PCA to constrain freedom of combination, but we need to investigate an appropriate constraint for speech generation. In a sub-space with an appropriate constraint, we can generate a huge number of more accurate unknown speaker utterances, which can then be used to train a universal model.

Our model covers a broad variety of speakers, and thus we expect it to be a good seed model for further speaker adaptation research. We are now investigating the appropriateness of our models for speaker adaptation [9].

References

[1] K. Shinoda, “Acoustic model adaptation for speech recognition,” IEICE Trans. Inf. & Syst., vol.E93-D, no.9, pp.2348–2362, Sept. 2010.

[2] J.-L. Gauvain and C.-H. Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains,” IEEE Trans. Speech Audio Process., vol.2, no.2, pp.291–298, 1994.

[3] C.J. Leggetter and P.C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density HMMs,” Comput. Speech Lang., vol.9, no.2, pp.171–185, 1995.

[4] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, “A compact model for speaker-adaptive training,” Proc. ICSLP, pp.1137–1140, 1996.

[5] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, “Rapid speaker adaptation in eigenvoice space,” IEEE Trans. Speech Audio Process., vol.8, no.6, pp.695–707, 2000.

[6] K.-T. Chen, W.-W. Liau, H.-M. Wang, and L.-S. Lee, "Fast speaker adaptation using eigenspace-based maximum likelihood linear regression," Proc. ICSLP, pp.742–745, 2000.

[7] A. Itoh, S. Hara, N. Kitaoka, and K. Takeda, "Training robust acoustic models using features of pseudo-speakers generated by inverse CMLLR transformation," Proc. APSIPA ASC 2011, Oct. 2011.

[8] J. Silovsky, P. Cerva, and J. Zdansky, "MLLR transforms based speaker recognition in broadcast streams," Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions, pp.423–431, 2009.

[9] A. Itoh, S. Hara, N. Kitaoka, and K. Takeda, "Robust seed model training for speaker adaptation using pseudo-speaker features generated by inverse CMLLR transformation," Proc. ASRU2011, pp.169–172, 2011.

[10] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," Comput. Speech Lang., vol.12, no.2, pp.75–98, 1998.

[11] S. Hara, C. Miyajima, K. Itou, and K. Takeda, "Data collection system for the speech utterances to an automatic speech recognition system under real environments," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.J90-D, no.10, pp.2807–2816, Oct. 2007.

[12] T. Kawahara and A. Lee, "Open-source speech recognition software Julius," JSAI, vol.20, no.1, pp.41–49, 2005.

[13] S. Furui, K. Maekawa, and H. Isahara, "A Japanese national project on spontaneous speech corpus and processing technology," Proc. ASR2000, pp.244–248, 2000.

Arata Itoh receive B.E. and M.E. degrees from Nagoya University, Nagoya, Japan, in 2009 and 2011, respectively. He joined DENSO CORPORATION, Kariya, Japan, in 2011. He was engaged in the research on the speech recognition when he was a student of Nagoya University.

Sunao Hara received B.E., M.E. and Ph.D. degrees from Nagoya University, Nagoya, Japan, in 1997, 1999, and 2011, respectively. He is now assistant professor of Nara Institute of Science and Technology, Nara, Japan. His research interests include speech recognition and spoken dialog systems.

Norihide Kitaoka received the B.E. and M.E. degrees from Kyoto University, Kyoto, Japan, in 1992 and 1994, respectively, and the Dr. Eng. degree from Toyohashi University of Technology, Toyohashi, Japan, in 2000. He joined DENSO CORPORATION, Kariya, Japan, in 1994. He joined the Department of Information and Computer Sciences, Toyohashi University of Technology, as a Research Associate in 2001 and was a Lecturer from 2003 to 2006. Since 2006, he has been an Associate Professor with the Department of Media Science, Graduate School of Information Science, Nagoya University, Nagoya, Japan. His research interests include speech processing, speech recognition, and spoken dialog.

Kazuya Takeda received the B.E., M.E., and Dr. Eng. degrees from Nagoya University, Nagoya, Japan, in 1983, 1985, and 1994, respectively. In 1986, he joined the Advanced Telecommunication Research Laboratories, Kyoto, Japan, where he was involved in two major projects, namely, speech database construction and speech synthesis system development. In 1989, he moved to KDD R&D Laboratories, Saitama, Japan, where he participated in the construction of voice-activated telephone extension systems. Since 1995, he has been with Nagoya University. His research interest covers a wide range of digital signal processing applications such as acoustic, speech, music, and human behavior signals. Dr. Takeda is a Board Member of the Acoustical Society of Japan.