Training Robust Acoustic Models Using Features of Pseudo-Speakers Generated by Inverse CMLLR Transformations

Arata Itoh*, Sunao Hara*, Norihide Kitaoka*, and Kazuya Takeda*
*Nagoya University, Nagoya, Japan
E-mail: kitaoka@nagoya-u.jp Tel: +81-52-789-3626

***Abstract*—In this paper a novel speech feature generation-based acoustic model training method is proposed. For decades, speaker adaptation methods have been widely used. All existing adaptation methods need adaptation data. However, our proposed method creates speaker-independent acoustic models that cover not only known but also unknown speakers. We do this by adopting inverse maximum likelihood linear regression (MLLR) transformation-based feature generation, and then 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 MLLR transformation matrices for the existing speakers are estimated. Next we generate pseudo-speaker MLLR transformations by sampling the weight parameters from the distribution, and apply the inverse of the transformation to the normalized existing speaker features to generate the pseudo-speakers' features. Finally, using these features, we train the acoustic models. Evaluation results show that the acoustic models which are created are robust for unknown speakers.**

I. Introduction

A method for speaker-independent robust speech recognition with limited speech resources is proposed. The degradation of speech recognition performance is often due to the mismatch between the training and test conditions. There are many reasons for such mismatches: differences between individual speakers, recording equipment, surrounding noise, etc. To compensate for such mismatches, adaptation techniques are often used. Model-based adaptation, such as maximum a posteriori (MAP) adaptation [1] and maximum likelihood linear regression (MLLR) [2], 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)[3] has also been proposed. In SAT, training data are normalized to a "virtual" average speaker for whom the acoustic models are trained. In the recognition stage, input speech is also normalized and recognized using the acoustic models for the average speaker.

Adaptation techniques which only need a small amount of target speech data, such as those used by inter-speaker adaptation methods like Eigenvoice [4] have 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 needs a small amount of target speech because the variety in the speech and environments is expressed in a low dimensional sub-space. Eigen-MLLR, which is a combination of MLLR and eigenvoice, was proposed in [5]. 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 adaptation methods need adaptation data. We can only use limited speech data from the environment where the system is to be used, 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 in limited speech resources. We do this by reversing the concept of adaptation. In the proposed method, we do not *remove* the speaker variations; we *add* them to the averaged speech features. We assume that individual speech variation is generated by adding the individual differences to an "average" person. Speaker recognition using the MLLR transformation matrix [6] suggests that the linear transformation matrix expresses individuality. We first obtain the MLLR transformation matrices from a limited amount of speech data and apply PCA to it to extract a small number of bases. Then we generate pseudo-speaker transformation matrices from the statistical linear combination of the bases. Finally, the speech features are generated by applying the inverse transformation matrices to the normalized speech features 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 robust to the inter-speaker variations.

II. Acoustic model training based on feature generation using inverse MLLR transformation

Our proposed method consists of five steps: (1) estimation of the MLLR transformation matrices of speaker utterances recorded in the target environments; (2) extraction of the bases of the MLLR transformation matrices; (3) estimation of the basis weight distributions; (4) speech feature generation

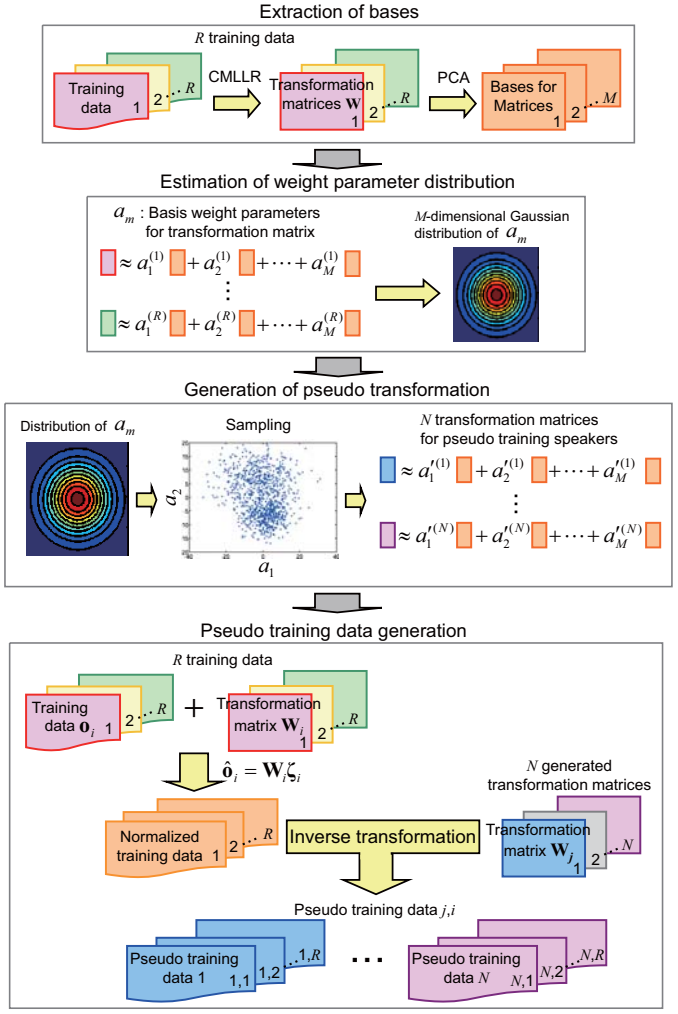


Fig. 1. Flow of the proposed feature generation-based acoustic model training

applying the inverse transformation matrix to the speaker-normalized speech data; (5) acoustic model training with generated features. The flow of the proposed method is summarized in Fig. 1. Here, we assume that we can use a certain amount of the training data in the target environments, but the data do not include the test speakers. This was necessary because we just developed this new application and were only able to collect a small amount of data in the field where the application was used.

A. Normalization of training speech

We adopted Constrained MLLR (CMLLR)[7], [8] to normalize the training speech:

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

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

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

B. Basis extraction using PCA

We assume that transformation matrix $\mathbf{W}$ consists of a linear combination of bases. One could use all the $\mathbf{W}_i, (i = 1, \cdots)$ as bases, but the number of components in $\mathbf{W}_i$ is large $(n \times (n+1))$. 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 $\mathbf{W}_i$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. We also consider basis extraction as the blind estimation of speaker variances.

C. 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 $\mathbf{a}^{(j)} = (a_1^{(j)}, \cdots, a_M^{(j)})^{\mathrm{T}} (i = 1, \cdots, R)$. We estimate the distribution of $\mathbf{a}^j$.

Each training speaker's super vector derived from the transformation matrix is approximated by a linear combination of $\tilde{\mathbf{V}}_i = \mathbf{a}^{(i)\mathrm{T}}(\mathbf{V}_E^1, \cdots, \mathbf{V}_E^M)$. Weight $\mathbf{a}^{(i)}$ is obtained by the square error minimization criterion. With $\mathbf{a}^{(j)}$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.

D. Speech feature generation by inverse MLLR transformation

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 generate N MLLR transformations, $\mathbf{W}'_n = [\mathbf{b}'_n \ \mathbf{A}'_n]$, by linear combination of the bases weighted by $\mathbf{a}'^{(n)}$.

Each generated transformation corresponds to a pseudo-speaker. We reverse the SAT technique [3] to obtain a variety of speakers by applying the transformation to the normalized speech features. We first apply the normalization matrix for training speaker i, $\mathbf{W}_i$, to the speech features of speaker i and then apply the inverse of the generated transformation, $\mathbf{W}'^{(-1)}_n$, to them to generate the speech feature of pseudo-speaker n:

$$\begin{aligned}\tilde{\mathbf{o}}_n &= \mathbf{A}'^{-1}_n \hat{\mathbf{o}}_i - \mathbf{A}'^{-1}_n \hat{\mathbf{b}}'_n & (2)\\ &= \mathbf{W}'^{(-1)}_n \hat{\boldsymbol{\zeta}}_i, & (3)\\ \hat{\mathbf{o}}_n &= \mathbf{W}_i \boldsymbol{\zeta}_i, & (4)\\ &\quad (i = 1, \cdots, R)\end{aligned}$$

where $\tilde{\mathbf{o}}_n$ is a generated feature of speaker n and $\boldsymbol{\zeta}_i = \left[\begin{array}{cc}\mathbf{1} & \mathbf{o}_i^{\mathrm{T}}\end{array}\right]^{\mathrm{T}}$ and $\hat{\boldsymbol{\zeta}}_i = \left[\begin{array}{cc}\mathbf{1} & \hat{\mathbf{o}}_i^{\mathrm{T}}\end{array}\right]^{\mathrm{T}}$ 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 generated 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 much more training data for the acoustic models. These pseudo-speakers are obtained from the distribution of original training speakers. If there are training speakers enough to estimate the "correct" acoustic models, the results should be better than our method. Here, we assume the situation that we cannot use enough data to train acoustic models and thus we try to "interpolate" or "extra-polate" the parameters of speakers.

E. 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 training data consist not only of existing speaker utterances but also the utterances of other generated speakers. As a result the acoustic models are expected to be robust at recognizing unknown speaker utterances.

III. EXPERIMENTS

A. Experimental conditions

We collected real-field speech data using the *MusicNavi2* [9] spoken dialog-based music retrieval system. This system obtains user utterances from the Internet using loss-less speech compaction. Many anonymous users can use this system.

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 ($100 \times 10 = 1000$ utterances), and training set 2 was 30 utterances from each subject ($100 \times 30 = 3000$ utterances).

We used test utterances from 250 speakers (160 males and 90 females). Fifty utterances from each speaker were used as test data. A feature vector consisted of a 12-dimensional MFCC, their first and second derivatives, and the first and second derivatives of the power. Experimental setup conditions, including these, are summarized in Table I.

For comparison, we performed MAP adaptation using all the training utterances, which is the adaptation for the environment, and SAT in the way of [2].

B. Evaluation results

1) Basis extraction: We set the cumulative proportions to 80%, 90%, and 95% to extract the bases. The relation between the cumulative contribution ratios and the number of bases is shown in Table II. With a cumulative proportion of 80%, we need approximately half of the bases that are extracted from the number of training speakers.

TABLE I
EXPERIMENTAL SETUP

# Training speakers	100 (50 males and 50 females)
# Training uttrerances	Set 1: 1000 (10 uttr. per person) Set 2: 3000 (30 uttr. per person)
# Test speakers	250 (160 males and 90 females) Exclusive with training sets
Amount of test data	12500 (50 uttr. per person)
Features	12 MFCC + 12 Δ + 12 $\Delta\Delta$ +Δpower + $\Delta\Delta$power
Speech recognizer	Julius-4.1[10]
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)

TABLE II
RELATION 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

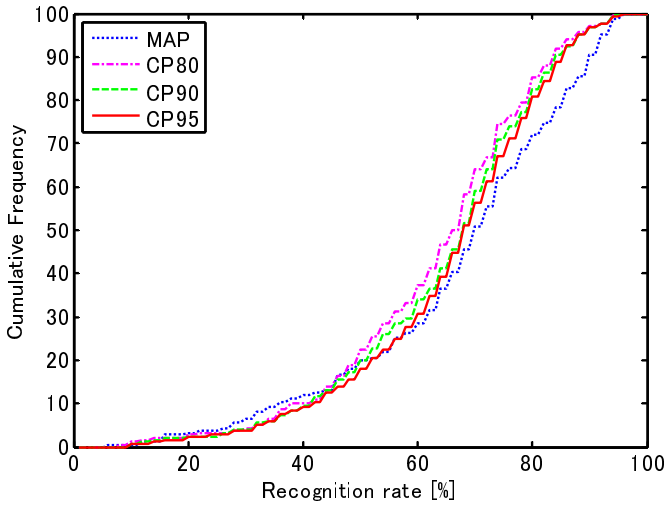
2) Recognition results: Using our proposed method, we generated 1000 pseudo-speakers from the bases described in Table II, randomly selected 600 real training speaker utterances from the training data, and converted these utterances for each pseudo-speaker. Thus we were able to obtain 60,000 training utterances. We trained the acoustic models using these utterances. For comparison, we also adapted the acoustic models trained using the Corpus of Spontaneous Japanese (CSJ)[11] with the real training data described in Table I. The average recognition rates and the standard deviations are shown in Table III. The test speakers were all different from the training speakers, so the recognition rates in Table III can be seen as the recognition rate without any speaker adaptation (but with environmental adaptation).

The table shows that the larger the number of the bases, 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. With Training set 2, the proposed method outperforms MAP adaptation. Note that the standard deviations of the recognition rates with the proposed method are smaller than those with MAP, suggesting that the acoustic models were robust for handling speaker variations. The cumulative test speaker frequencies and the recognition rates are shown in Fig. 2. The numbers of speakers with low recognition rates are significantly smaller using the

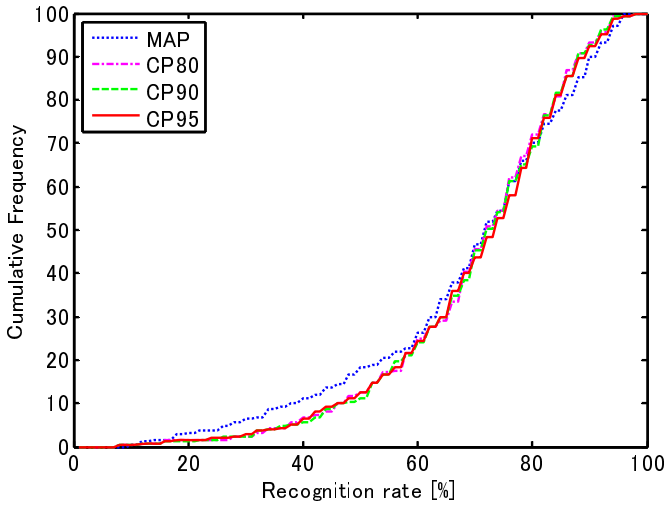
[1]We have to note that we should compare these results with the case we train acoustic models using original training data, but we could not do on the same condition because of the lack of training data. We conducted a preliminary experiments with small size of acoustic models and obtained that even using the training set 2 and HMMs with 500 states, the result of our method outperformed that by HMMs trained using original data because of the over-training. Using HMMs with 300 states, training original data were comparable with our method.

TABLE III
RECOGNITION RATES [%] AND STANDARD DEVIATIONS USING ACOUSTIC MODELS TRAINED USING PSEUDO-SPEAKER UTTERANCES AND THOSE ADAPTED BY MAP

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



Training set 1



Training set 2

Fig. 2. Cumulative speaker frequency against recognition rates. CPn 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.

proposed method than with MAP. This suggests that speaker generation produces a wide range of speaker variations. For this reason, the proposed training feature generation method works robustly for unknown speakers, especially those with low recognition rates. Inversely, our method did not perform well for the speakers originally with high recognition rates.

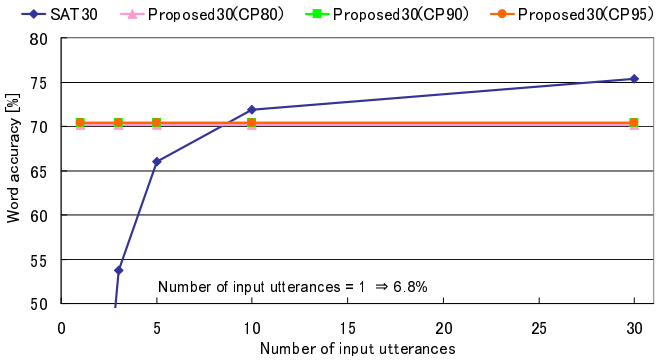


Fig. 3. Comparison of proposed method and SAT. CPn expresses cumulative proportion of n in the proposed method. In the figure, all the lines for CPns overlaps.

This may be the effect of pseudo-speakers which were far from real humans, resulting that the distributions in HMM states were too broad than they need to be.

C. Comparison with SAT

The method proposed in this paper is inspired by SAT, in which all the training speech is normalized and used for training, and in which the test utterances are normalized as well. However, our method generates speaker variation to train the acoustic models.

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

We also compared the performance of SAT and our method. In the SAT framework, we assume that the normalization parameters of the transformation matrix are estimated from 1, 3, 5, 10, and 30 input utterances with which to normalize the test data. The recognition results are shown in Fig. 3. SAT performs better with more than ten adaptation utterances, but our method performs well without adaptation data.

IV. CONCLUSION

In this paper, we proposed generative acoustic model training based on the generation of pseudo speech features using a linear combination of principal components of MLLR transformation matrices. 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 to produce an accurate and robust acoustic model. Currently we only use PCA to constrain the freedom of combination, but we need to investigate an appropriate constraint for speech generation. In an appropriate constraint subspace, we can generate a huge number of more accurate unknown speaker utterances to train a universal model.

REFERENCES

[1] 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, pp. 291–298, 1994.

[2] M.J.F. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition,” Computer Speech and Language, vol. 12, no. 2, pp. 75–98, 1998.

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

[4] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, “Rapid speaker adaptation in eigenvoice space,” IEEE Trans. Speech and Audio Processing, vol. 8, pp. 695–707, 2000,

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

[6] S. Jan, C. Petr, and Z. Jindrich, “MLLR transforms based speaker recognition in broadcast streams,” Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions, pp. 423–431, 2009.

[7] M. J. F. Gales and P.C. Woodland, “Mean and variance adaptation within the MLLR framework,” Computer Speech and Language, vol. 10, pp. 249–264, 1996.

[8] V. V. Digalakis, D. Rtischev, and L. G. Neumeyer, “Speaker adaptation using constrained estimation of Gaussian mixtures,” IEEE Trans. Speech and Audio Processing, vol. 3, No. 5, 1995.

[9] 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. on Inf. & Syst., vol J90-D, No. 10, pp. 2807–2816, 2007. (in Japanese)

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

[11] S. Furui, K. Maekawa, and H. Isahara, “A Japanese national project on spontaneous speech corpus and processing technology,” Proc. ASR2000, pp. 244-248, 2000.