Robust seed model training for speaker adaptation using pseudo-speaker features generated by inverse CMLLR transformation

Arata Itoh [#1], Sunao Hara [#2], Norihide Kitaoka [#3], Kazuya Takeda [#3]

[#] *Department of Information Science, Nagoya University*
Furo-cho, Chikusa-ku Nagoya Aichi, 464-8603 Japan
[3] kitaoka@nagoya-u.jp

***Abstract*—In this paper, we propose a novel acoustic model training method which is suitable for speaker adaptation in speech recognition. Our method is based on feature generation from a small amount of speakers' data. For decades, speaker adaptation methods have been widely used. Such adaptation methods need some amount of adaptation data and if the data is not sufficient, speech recognition performance degrade significantly. If the seed models to be adapted to a specific speaker can widely cover more speakers, speaker adaptation can perform robustly. To make such robust seed models, we adopt inverse maximum likelihood linear regression (MLLR) transformation-based feature generation, and then train our seed 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 is 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 seed models. Using this seed models, we obtained better speaker adaptation results than using simply environmentally adapted models.**

I. Introduction

A method to train acoustic models for speech recognition suitable for speaker adaptation based on feature generation 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 recording equipment, surrounding noise, individual speakers, 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.

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.

Here, we assume that if we can obtain enough data in the environment where the system is to be used, we can adapt or train acoustic models to fit test environments. Using such environment-adapted acoustic models as the seed models for speaker adaptation, the adaptation is expected to be done more successful because there is no environmental mismatch and only the speaker differences should be compensated.

In reality, however, we can only use limited speech data from such an environment, 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-adaptation seed model training method which trains acoustic models robustly from the 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 speech features averaged in the target environment. We assume that individual speech variation is generated by adding the individual differences to an "average" person. If we can obtain the sufficient amount of training data, it is possible to make robust acoustic models covering the variation of the speech and thus the models are suitable for adaptation seed models. Here, we simulate the variation of the speech artificially. Speaker recognition using the MLLR transformation matrix [6] suggests that the linear transformation matrix expresses individuality. We first obtain the MLLR transformation matrices from limited number of speakers' 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 which cover thousands of speakers in the target environments. Such acoustic models have wide variances and thus target speaker to adapt is probably covered. Such condition makes the adaptation easier.

The remainder of this paper is organized as follows. Section 2 describes our proposed feature generation-based acoustic model training. Section 3 explains the benefit of the use of the acoustic models trained by the feature generation as the seed models of speaker adaptation. Experimental results are shown in Section 4. We conclude this paper in Section 5.

II. Acoustic Model Training Based on Feature Generation using Inverse MLLR Transformation

Our proposed seed model training 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 applying the inverse transformation matrix to the speaker-normalized speech data; (5) acoustic model training with generated features. The flow of the 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 assumption is practical 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 $\mathbf{W}_i, (i = 1, \cdots, R)$ 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.

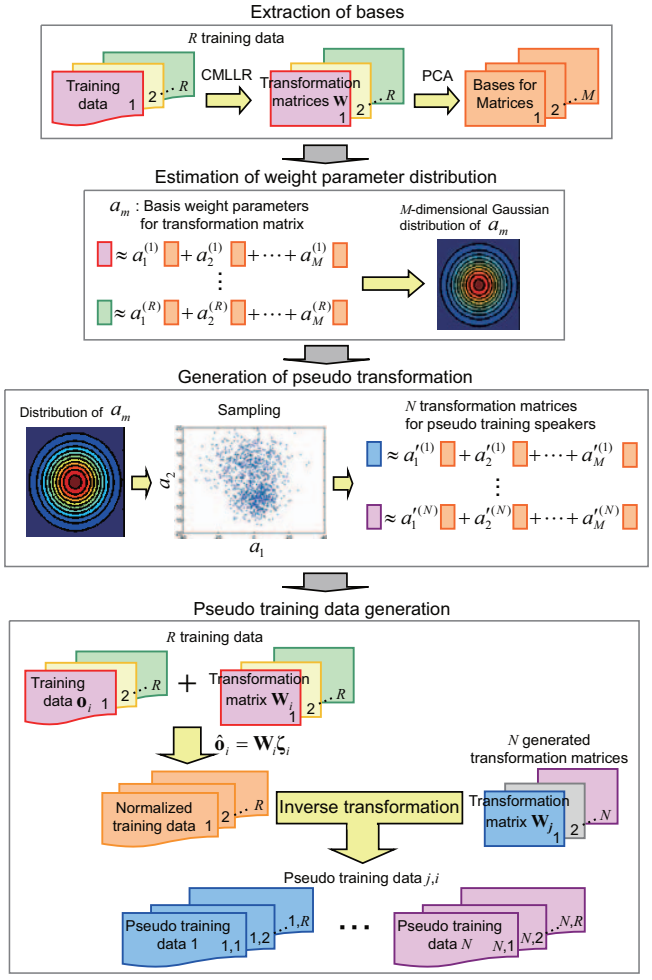

Fig. 1. Flow of the proposed feature generation-based acoustic seed model training

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 $\mathbf{a}^{(j)} = (a_1^{(j)}, \cdots, a_M^{(j)})^{\mathrm{T}} (j = 1, \cdots, R)$. 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, $\mathbf{W'}_n^{(-1)}$, to them to generate the speech feature of pseudo-speaker n:

$$\tilde{\mathbf{o}}_n = \mathbf{A'}_n^{-1}\hat{\mathbf{o}}_i - \mathbf{A'}_n^{-1}\hat{\mathbf{b}}'_n \quad (2)$$

$$= \mathbf{W'}_n^{(-1)}\hat{\boldsymbol{\zeta}}_i, \quad (3)$$

$$\hat{\mathbf{o}}_n = \mathbf{W}_i\boldsymbol{\zeta}_i, \quad (4)$$

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

where $\tilde{\mathbf{o}}_n$ is a generated feature of speaker n and $\boldsymbol{\zeta}_i = \left[\ \mathbf{1} \ \ \mathbf{o}_i^{\mathrm{T}}\right]^{\mathrm{T}}$ and $\hat{\boldsymbol{\zeta}}_i = \left[\ \mathbf{1} \ \ \hat{\mathbf{o}}_i^{\mathrm{T}}\right]^{\mathrm{T}}$ 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.

E. Training acoustic seed 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.

III. Effect of generation-based seed model training on the speaker adaptation

As a result of the training expressed in Section II, the acoustic seed models are expected to cover a broad range of individual differences of speakers. Such models may be robust for unknown speaker utterances, but they have possibility to be too broad to discriminate the appropriate phonetic categories.

These models, however, have some levels of likelihood for outlier speakers' utterances. Because of this characteristic, these models are expected to be suitable for the speaker adaptation seed models.

IV. 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.

TABLE I
Experimental setup

# Training speakers	100 (50 males and 50 females)
# Training utterances	3000 (30 uttr. per person)
# Test speakers	12 (6 males and 6 females) Exclusive with training sets
Amount of test data	600 (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 300 states, 16 mixtures per state
Language model	Word loop grammar
Dictionary	Words for MusicNavi2 (approx. 8000 words)

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 was 30 utterances from each subject ($100 \times 30 = 3000$ utterances).

We used test utterances from 12 speakers (6 males and 6 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.

B. Basis extraction

We set the cumulative proportion to 95% to extract the bases, resulting that we extracted 84 bases.

C. Recognition results using feature generation-based acoustic model training

First, we perform a recognition experiment using acoustic models trained by proposed feature-generation based training without speaker adaptation.

We generated 1000 pseudo-speakers from the bases described in Section IV-B, 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 used the acoustic models trained by the Corpus of Spontaneous Japanese (CSJ)[11], and those adapted by MAP from the CSJ models with the real training data described in Table I (MAP). The average recognition rates and the standard deviations are shown in Table II. The test speakers were all different from the training speakers, so the recognition rates in Table II can be seen as the recognition rate without any speaker adaptation (but with environmental adaptation in the cases of "MAP" and "Proposed.").

TABLE II
RECOGNITION RATES [%] USING ACOUSTIC MODELS TRAINED USING PSEUDO-SPEAKER UTTERANCES, THOSE TRAINED USING CSJ DATABASE, AND THOSE ADAPTED BY MAP

	CSJ	MAP	Proposed
Recog. rate	58.3	68.5	73.3

The table shows that the models adapted by the proposed method outperforms not only the CSJ models but also MAP-adapted models.

D. Recognition results by speaker-adapted models from those by generation-based training

We applied speaker adaptation methods to the CSJ models, environment-adapted models using MAP, and models trained using features generated by our proposed method. The results are shown in Figs. 2 and 3. Figures 2 and 3 are the results by MLLR adaptation and MAP adaptation, respectively, and the lines with marks show the tendencies by changing the number of utterances used for adaptation.

We can find that the performance differences between models without adaptation are kept after the adaptation, even after the adaptation using 80 utterances for each speaker. After the adaptation using a small number of utterances as five, the performance improvement by the adaptation using the models trained by our method as the seed models is larger than the others. This result implies that the our proposed models covers broad variations of speakers and adapted robustly using small nuber of utterances.

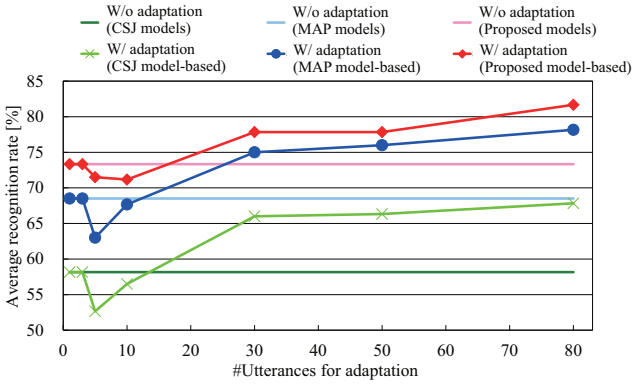

Fig. 2. Recognition results using MLLR speaker adaptation from CSJ models, MAP adapted models, and models adapted by proposed method.

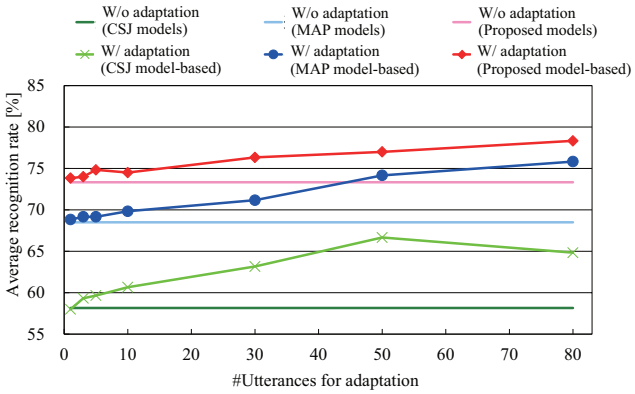

Fig. 3. Recognition results using MAP speaker adaptation from CSJ models, MAP adapted models, and models adapted by proposed method.

V. CONCLUSION

In this paper, we proposed a generative acoustic model training based on the generation of pseudo speech features using a linear combination of principal components of MLLR transformation matrices. Our models outperform conventional MAP environment-adapted models. Using our models as seed models, speaker adaptation was effectively performed.

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. We believe that robust acoustic models are more suitable for adaptation seed models.

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. Make, and H. Isahara, "A Japanese national project on spontaneous speech corpus and processing technology," Proc. ASR2000, pp. 244-248, 2000.