Unsupervised Speaker Adaptation of DNN-HMM by Selecting Similar Speakers for Lecture Transcription

Masato Mimura and Tatsuya Kawahara*
* Kyoto University, Academic Center for Computing and Media Studies,
Sakyo-ku, Kyoto 606-8501, Japan

***Abstract*—Unsupervised speaker adaptation of Deep Neural Network (DNN) is investigated for lecture transcription tasks, in which a single speaker gives a long speech and thus speaker adaptation is important. The proposed method selects similar speakers to the test data (test speaker) from the training database, which are used for retraining the baseline DNN. Several speaker characteristic features are defined for the speaker similarity measure. The feature based on Universal Background Model (UBM) and principal component analysis (PCA) achieves the best performance, resulting in a significant improvement from the baseline DNN and also from the adapted GMM-HMM system. The method is combined with a naive adaptation method using the initial ASR hypothesis of the test data, and an additional improvement is achieved.**

I. Introduction

In the past few years, Deep Neural Network (DNN) has demonstrated to outperform the conventional Gaussian Mixture Model (GMM)-based systems in a variety of automatic speech recognition (ASR) tasks [1]. We have confirmed that the DNN-HMM hybrid system achieves significantly better accuracy with faster decoding turnaround than GMM-HMM in our (offline) lecture transcription tasks which we address in this paper. DNN is flexible for handling a large-dimensional multi-frame feature vector and learning complex decision boundaries with a cascade of networks.

On the other hand, there is not an effective adaptation method of DNN to new speakers or environments, except for retraining the network with new data. This is contrastive to GMM-HMM that can be adapted with a solid statistical framework such as MLLR and MAP. Speaker adaptation is important in lecture transcription tasks because each speaker's speech is very long, typically from 15 to 90 minutes. Moreover, an unsupervised adaptation method which works using the test data is desirable because it is not practical to assume supervision adaptation data while it is allowed to conduct decoding the data several times. In the conventional GMM-HMM systems, speaker adaptation conducted via MLLR or MAP brings significant improvement of accuracy [2]. In fact, when speaker adaptation is applied, the performance of GMM-HMM is almost similar to that of DNN-HMM in our evaluation sets. However, the same adaptation scheme cannot be applied to DNN-HMM.

The objective of this study is to investigate effective methods of unsupervised speaker adaptation for lecture transcription tasks in which a single speaker gives a long speech. One naive or ideal method is to retrain DNN with the test data (test speaker) itself. However, since we cannot expect the correct supervision label for it, it is difficult to predict the adaptation scheme will work properly. Therefore, we present a scheme that retrieves training data similar to the current test data from the training database, which are then used for retraining DNN. In this paper, we investigate several speaker characteristic features which are computationally efficient and thus used for the similarity measure for this unsupervised adaptation scheme.

After a brief review on the baseline DNN-HMM system and previous work on speaker adaptation of DNN in Section 2 and 3, the detail of the proposed method is explained in Section 4. Experimental evaluations of the method are presented in two different lecture transcription tasks in Section 4. Here, we also report its combination with the naive retraining method using the test data.

II. Baseline DNN-HMM System and Performance

A. Baseline DNN-HMM Hybrid System

In this work, we adopt a standard DNN-HMM hybrid system as the baseline. It is made by replacing all GMMs of the GMM-HMM system with a single DNN to calculate emission probabilities of the triphone states. The DNN of our system is trained with a standard procedure [3][4] . The number of hidden layers is 7. Each hidden layer has 2048 sigmoidal nodes. The number of output nodes, which correspond to the tied states of the triphone HMMs, is 3015. The DNN was initialized with the stacked Restricted Boltzman Machines (RBMs) pre-trained using the contrastive divergence (CD1) method, and then fine-tuned by back-propagation (BP) with the minimum cross-entropy criterion. The unsupervised pre-training and supervised fine-tuning were performed using mini-batches of 256 samples with a stochastic gradient descent method. The training samples were shuffled prior to the training. The learning rate for BP was initialized with 0.08 and halved if the error rate for the development set was not improved at the end of each epoch. BP was iterated for up to 10 epochs. Other parameters of the DNN-HMM such as state transition probabilities were copied from the baseline GMM-HMM system. The raw acoustic feature at each analysis frame is 40-dimensional log Mel-filterbank outputs with their first

978-616-361-823-8

and second derivatives. The raw features were normalized to have a global mean of 0.0 and a variance of 1.0. Temporal contexts are incorporated by splicing 11 successive frames, so that the DNN has 1320 (=40*3*11) input nodes.

As the training database, we used 967 academic oral presentations in the Corpus of Spontaneous Japanese (CSJ) [5], which consists of 799 male and 168 female speakers. The total amount of the training data is 257 hours. The baseline GMM-HMM was trained using the same training database. The acoustic feature for the GMM-HMM consists of 12-dimensional MFCC, ΔMFCC, ΔΔMFCC, Δpower and ΔΔpower. Cepstrum mean and variance normalization (CMN/CVN) and vocal tract length normalization (VTLN) were applied for each lecture (speaker). The GMM-HMM were trained discriminatively using the minimum phone error (MPE) criterion. The number of the tied states is 3015 and each state has a Gaussian mixture of 16 components.

B. Baseline Performance in Lecture Transcription Tasks

We have set up two lecture transcription tasks in this work. All lectures are distributed by Kyoto University OpenCourseWare (OCW) website[1]. They were not given as a normal course in our university, but arranged as symposia open to public. We took two series of symposia which are most popular in the OCW website. One is symposia on disaster control, which attracts special attention in Japan after the large earthquake and tsunami in 2011. We selected six lectures as a test set, referred to as OCW-SHINSAI hereafter. The other series of symposia is by the Center for iPS Cell Research and Application (CiRA), whose director was awarded the Nobel Prize in 2012. We selected six lectures as a test set, referred to as OCW-CiRA. The baseline language model was trained using the transcriptions of the CSJ, and adapted to each task using related document texts such as newspapers and web pages. For the ASR decoder, Julius rev.4.2 [2] was used with some modifications for DNN-HMM.

Performance of the baseline DNN-HMM and GMM-HMM systems is shown in Table I. Here, the results of unsupervised speaker adaptation of the baseline GMM-HMM with MAP and MLLR using the ASR results by the baseline system are also given. The DNN-HMM exhibited much better accuracy than the baseline GMM-HMM. With MLLR-based speaker adaption, however, the performance of the GMM-HMM is almost similar to that of the DNN-HMM. Here, MLLR works better than MAP, because MAP is more sensitive to ASR errors in the initial recognition hypothesis used in the unsupervised adaptation scheme.

The primary goal of these tasks is to transcribe lectures offline as accurately as possible, so a multi-pass recognition strategy is allowed. Speaker adaptation is important since each test data is long and given by a single speaker, and it has actually shown to be effective for the GMM-HMM. Therefore, we explore speaker adaptation methods for the DNN-HMM system.

[1] http://ocw.kyoto-u.ac.jp

[2] http://julius.sourceforge.jp/

TABLE I
BASELINE SYSTEM PERFORMANCE (WORD ACCURACY)

	OCW-SHINSAI	OCW-CiRA
DNN-HMM	63.69	79.40
GMM-HMM	55.01	75.41
+MAP adaptation	57.73	76.31
+MLLR adaptation	62.84	78.46

III. RELATED WORK ON SPEAKER ADAPTATION OF DNN-HMM

Speaker adaptation of DNN-HMM is not straightforward because standard techniques of MLLR and MAP for GMM-HMM is not applicable. A simple method is to design a linear transformation to weights of links to input and/or output nodes, which is analogous to fMLLR and global MLLR [6]. Some work introduced a special network layer for speaker adaptive training [7]. Another naive method is to retrain DNN with new data [8]. It is shown to work much better than the linear transformation method for task adaptation [9]. Recently, the retraining approach has also been applied to speaker adaptation [10][11].

Since DNN has a huge number of free parameters, it will easily fall into overfitting with a small amount of data. Yu et al. [10] proposed a method to control the adaptation procedure by monitoring the KL-divergence from the baseline model. Liao et al. [11] introduced L2-regularization to effectively control speaker adaptation. These studies showed that the overfitting problem is mitigated by some regularization methods, but the problem becomes not so serious when the amount of adaptation data gets large (longer than 10 minutes), which is true in the lecture transcription tasks.

These studies also showed that the effect of unsupervised adaptation is far below that of supervised adaptation. We can reason that the retraining method is apparently not robust for ASR errors in the initial recognition hypothesis used for unsupervised adaptation, as in the MAP adaptation for GMM-HMM. The problem may be mitigated by filtering based on a confidence measure of the hypothesis, but will not be solved completely because the confidence measure is not perfect for filtering. Moreover, erroneous samples in the initial recognition result would be very useful for improving the speaker adaptation performance. More recently, Saon et al. proposed to incorporate i-vector, which will be explained in Section 4.1, to input features as a speaker-characterics [12]. Our scheme conducts direct speaker adaptation by retraining DNN.

IV. SPEAKER ADAPTATION USING TRAINING DATA OF SIMILAR SPEAKERS

In this paper, we present a speaker adaptation scheme that selects training data similar to the test data for enhancing DNN. In the lecture transcription tasks, this is equivalent to selection of similar speakers to the speaker of the test data. The adaptation data are selected from the training database with the supervision labels, so the network retraining is guaranteed to work appropriately.

This approach was explored in GMM-HMM [13][14], and the key issue is the definition of the similarity measure of

speakers, which is accurate and computationally efficient. Use of a likelihood ratio such as the KL-divergence would be accurate, but it involves likelihood computation for long speech data, which is impractical. In this paper, therefore, we investigate rather simple features to represent speaker characteristics, which are inspired by the progress in the speaker recognition research. The similarity measure is defined by the Euclidean distance or cosine distance between these features representing speaker characteristics. The N-most similar speakers are selected from the training database, and their data are used for retraining DNN for speaker adaptation to the test data.

Specifically, the following four features are investigated.

A. UBM-GMM-based Features

We adopt an approach of Universal Background Model (UBM), which has been widely used for speaker recognition. In this approach, a speaker-independent GMM called UBM is trained using all speakers' data of the training database. For each speaker, a speaker-dependent GMM is built by conducting MAP adaptation to the UBM using the speaker's utterances, and the mean vectors of the adapted GMM are stacked to make up a GMM super-vector (GMM-SV) m_s for speaker s.

Since the GMM super-vector is high-dimensional and apparently redundant, compact and informative feature extraction has been investigated, and recently i-vector and its variants are proposed [15][16]. In this study, we conduct principal component analysis (PCA) on the super-vector to reduce the dimensionality. Here, (d-dimensional) super-vector m_s for speaker s is approximated by

$$m_s \simeq m_0 + Tw_s \tag{1}$$

where m_0 is the speaker-independent super-vector defined from the UBM (UBM super-vector). The normalized super-vectors $(m_s - m_0)$ of all S speakers are concatenated to make a large ($S \times d$-dimensional) matrix, on which PCA is conducted. As a result, M principal components are extracted to make up a ($M \times d$-dimensional) matrix T, which corresponds to the total variability matrix in [16].

The GMM-PCA feature is defined by

$$w_s = T^t(m_s - m_0) \tag{2}$$

as the components in T are supposed to be orthonormal. Unlike i-vector [16], we do not formulate an EM algorithm [15] which estimates T and w_s simultaneously, because each speaker's session in our tasks is long enough to build a reliable speaker-dependent model with the MAP adaptation.

For our training database, a UBM with 256 mixture components was trained using the 38-dimensional MFCC-based features used for GMM-HMM. CMN/CVN and VTLN were not applied to preserve speaker information. Then, the UBM was adapted to each of the 967 speakers in the database. The dimension of the resulting GMM super-vectors (GMM-SV) is 9728 (=256*38). As the result of PCA, 200 eigen-vectors with the largest eigen-values were extracted. A 200-dimensional feature (GMM-PCA) is also computed for each speaker. Two similarity measures of speakers are defined by the (negative) Euclidean distance for the 9728-dimensional GMM-SV and the cosine distance for the 200-dimensional GMM-PCA, respectively.

TABLE II
COMPARISON OF SPEAKER SIMILARITY MEASURES (WORD ACCURACY)

	OCW-SHINSAI	OCW-CiRA
unadapted DNN-HMM	63.69	79.40
HMM-SV	64.73	80.10
HMM-xform	65.14	80.24
GMM-SV	64.95	80.41
GMM-PCA	65.31	80.44

TABLE III
RESULTS FOR TEST SET (WORD ACCURACY)

lecture ID	OCW-SHINSAI	OCW-CiRA
baseline	63.69	79.40
adapt with similar speakers	65.31	80.44
adapt with test data	67.23	80.87
combination of above two	67.98	81.17

B. GMM-HMM-based Features

We also investigate features based on triphone GMM-HMM. MLLR adaptation of the baseline GMM-HMM is conducted for each speaker using the phone labels to obtain a speaker-dependent GMM-HMM. In this process, an MLLR transformation matrix is estimated. We can use this MLLR matrix (HMM-xform) as a speaker characteristic feature [17] . In addition, the mean vectors of the adapted GMM-HMM (except the silence model) are stacked to make up an HMM super-vector (HMM-SV). It is very high-dimensional (1,827,649 in our setting). Two more similarity measures of speakers are defined by the (negative) Euclidean distance for MLLR-xform and GMM-SV.

V. EXPERIMENTAL EVALUATIONS

Experimental evaluations were conducted using the baseline DNN-HMM and the test sets described in Section 2. After N-best speakers are selected from the training database, the baseline DNN is retrained for additional 10 epochs of BP using the speech data of these speakers. A relatively small learning rate ($= 0.001$) was used for retraining to prevent overfitting.

A. Comparison of Speaker Similarity Measures

First, we compared the four features for the speaker similarity measure. In this experiment, 5-best speakers were selected from the training database. The results are listed in Table II.

All methods realized a significant improvement from the baseline performance. Note that this improvement gives a large margin over the adapted GMM-HMM system (Table I). Among the two test sets, a larger improvement was obtained for the test set of OCW-SHINSAI which had lower baseline performance. Although the differences among the four methods are not large, the GMM-PCA feature yielded the best performance consistently for both test sets. This is attributed to compact and informative feature representation attained through PCA.

B. Effect of Number of Selected Speakers

Next, we investigate the effect of the number of selected speakers. There is a trade-off between similarity and the

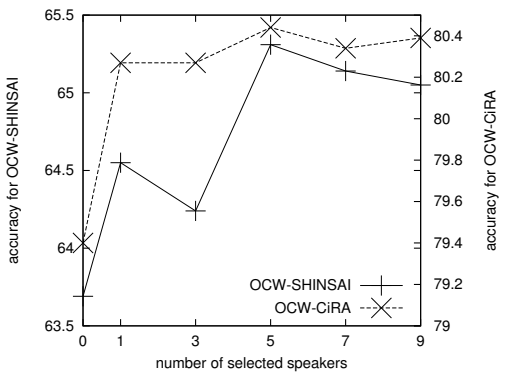

Fig. 1. Effect of number of selected speakers (word accuracy) (0 means unadapted case)

amount of adaptation data; the more speakers are selected, the less similar speakers would be incorporated. The result by changing the number of speakers (N) used for adaptation is presented in Fig. 1. In this experiment, the GMM-PCA feature was used for the speaker similarity measure.

From this result, we concluded that N=5 was best, although there is not a significant drop before N=10. In this case (N=5), the amount of the adaptation data for each test speaker was about 80 minutes on average.

C. Combination with Adaptation using Test Data

We also tried the unsupervised adaptation method using the initial ASR result of the test data. The baseline DNN is retrained using the test data with the initial recognition hypothesis as a supervision label, which is error prone. The method can be combined with the proposed method. In this case, the baseline DNN is retrained using the similar speakers' data with correct labels and the test data with a error-prone label.

The overall results are shown in Table III. The adaptation method using the test data also realizes a significant improvement in accuracy. It is generally observed that the method performs better than the adaptation method using similar speakers when the test data is longer because more speaker-dependent data are incorporated, although there is not a statistically significant difference between the two methods for a majority of the lectures. The combination of the two adaptation methods (last row in the two tables) achieved a further improvement in both test sets. It shows a synergetic effect of the two methods. It is noteworthy that the adaptation methods are effective especially for the lectures whose baseline performance is relatively low, and a drastic improvement is obtained for two female speakers among the test sets. This is because the training database is biased with male speakers, and demonstrates the effect of the speaker adaptation.

VI. Conclusion

We have presented an unsupervised speaker adaptation method for DNN-HMM systems for lecture transcription tasks, in which speaker adaptation is important because of a large amount of the adaptation data. In order to complement the naive adaptation method using the initial ASR result of the test data, we propose an adaptation scheme that selects training data similar to the test data. We investigate several speaker characteristic features used in speaker recognition to define the speaker similarity measure. These features are efficiently computed without likelihood computation on the speech data. In the experimental evaluation, they all achieved a significant improvement in accuracy from the baseline DNN-HMM, but the GMM-PCA feature showed the best performance. The method is combined with the adaptation method using the test data, resulting to an additional improvement. The adaptation method is particularly effective for the speakers whose baseline performance is low.

The ASR transcripts have been edited and proofed by human editors, and the final texts are now used as caption of these lectures broadcast from the OCW website.

Acknowledgments: This work was supported by JST CREST and JSPS Grant-in-Aid for Scientific Research.

References

[1] G.E.Hinton, L.Deng, D.Yu, G.Dahl, A.Mohamed, N.Jaitly, A.Senior, V.Vanhoucke, P.Nguyen, T.Sainath, and B.Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] H.Nanjo and T.Kawahara, "Language model and speaking rate adaptation for spontaneous presentation speech recognition," *IEEE Trans. Audio, Speech & Language Process.*, vol. 12, no. 4, pp. 391–400, 2004.

[3] A.Mohamed, G.Dahl, and G.Hinton, "Acoustic modelling using deep belief networks," *IEEE Trans. Audio, Speech, & Language Proc.*, vol. 20, no. 1, pp. 14–22, 2012.

[4] G.E.Dahl, D.Yu, L.Deng, and A.Acero, "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition," *IEEE Trans. Audio, Speech, & Language Proc.*, vol. 20, no. 1, pp. 30–42, 2012.

[5] K.Maekawa, "Corpus of Spontaneous Japanese: Its design and evaluation," in *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003, pp. 7–12.

[6] V.Abrash, "Mixture input transformations for adaptation of hybrid connectionist speech recognizers," in *Eurospeech*, 1997.

[7] S.M.Siniscalchi, J.Li, and C-H.Lee, "Hermitian polynomial for speaker adaptation of connectionist speech recognition systems," *IEEE Trans. Audio, Speech & Language Process.*, vol. 21, no. 10, pp. 2152–2161, 2013.

[8] J.P.Neto, C.Martins, and L.B.Almeida, "Speaker adaptation in a hybrid HMM-MLP recognizer," in *ICASSP*, 1996.

[9] Y.Xiao, Z.Zhang, S.Cai, J.Pan, and Y.Yan, "A initial attempt on task-specific adaptation for deep neural network besed large vocabulary continuous speech recognition," in *Proc. INTERSPEECH*, 2012.

[10] D.Yu, K.Yao, H.Su, G.Li, and F.Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *ICASSP*, 2013, pp. 7893–7897.

[11] H.Liao, "Speaker adaptation of context dependent deep neural networks," in *ICASSP*, 2013, pp. 7947–7951.

[12] G.Saon, H.Soltau, D.Nahamoo, and M.Picheny, "Speaker adaptation of neural netowork acoustic models using i-vectors," in *Proc. ASRU*, 2013, pp. 55–59.

[13] R.Gomez, T.Toda, H.Saruwatari, and K.Shikano, "Improving rapid unsupervised speaker adaptation based on HMM sufficient statistics," in *ICASSP*, vol. 1, 2006, pp. 1001–1004.

[14] M.Mimura and T.Kawahara, "Fast speaker normalization and adaptation based on BIC for meeting speech recognition," in *APSIPA*, 2011.

[15] P.Kenny, G.Boulianne, and P.Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Audio, Speech & Language Process.*, vol. 13, no. 3, pp. 345–354, 2005.

[16] N.Dehak, P.Kenny, R.Dehak, P.Dumouchek, and P.Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech & Language Process.*, vol. 19, no. 4, pp. 788–798, 2011.

[17] A.Stolcke, L.Ferrer, S.Kajarekar, E.Shriberg, and A.Venkataraman, "MLLR transforms as features in speaker recognition," in *EUROSPEECH*, 2005, pp. 2425–2428.