Deep Neural Network based Acoustic Model Using Speaker-Class Information for Short Time Utterance

Hiroshi Seki* and Kazumasa Yamamoto† and Seiichi Nakagawa‡
Toyohashi University of Technology, Aichi, Japan
* E-mail: seki@slp.cs.tut.ac.jp † E-mail: kyama@slp.cs.tut.ac.jp ‡ E-mail: nakagawa@slp.cs.tut.ac.jp

***Abstract*—In speech recognition, it is preferable not to hypothesize the details, e.g., specific age and gender, of a target user. However, speaker independence is one of the things that degrades ASR performance. In this work, we propose a speaker adaptation method to recognize a short time utterance. There have been several studies on speaker-independent DNN-HMM in which i-vector is computed, and the additional information is combined with acoustic features. However, it is difficult to calculate i-vector accurately or apply speaker adaptation (e.g. fMLLR) when the utterance time is short (0.5sec∼). In our approach, we calculate the similarity score between the speaker class and the target utterance and utilize speaker class information configured in advance. As a precondition, we restrict the available time period to the first 50 frames per utterance for the recognition of short utterances. In experimental tests, we obtained a 4.0% relative WER gain compared to conventional DNN-HMM.**

I. Introduction

Recently, automatic speech recognition has been used in many applications as a human-machine interface. However, since a system can not identify the speaker and speech environment in advance, there is a problematic reduction in speech recognition performance owing to a mismatch between the input speech and the acoustic model's training data. To attain the performance required for a recognition system, an acoustic model that can consider various speakers and speech environments is essential.

Recently, deep neural networks (DNNs) have been applied to speech recognition and have outperformed the conventional Gaussian mixture model-based methods [1]. Several studies have focused on speaker-independent speech recognition and adaptation technique using DNN-HMM[2], [3], [4]. In these methods, the i-vector and speaker code are extracted from an utterance and these additional information are used to suppress the variation of acoustic features. However, it is difficult to calculate i-vector accurately when the utterance time is short[2]. In [5], more than 90% speaker identification rate is obtained when the duration of training and test utterances are more than 2.0 seconds. However, the performance of speaker identification decreased to 60% if the duration of training and test utterance are 30 and 0.5 seconds, respectively. The proposed method is related to the techniques such as training data clustering [6] and cepstral mean and variance normalization (CMN/CVN)[7], which have been widely used in GMM-HMM. In our method, we apply feature normalization by using speaker class configured in advance. The main difference between our method and i-vector-based method is the applicability to the short time utterance.

The rest of this paper is organized as follows: In Section 2, we describe how to utilize training data to obtain the speaker-class information. In Section 3, we describe the use of multiple acoustic models constructed by soft-clustering the training data. The experimental set up is presented in Section 4, and experimental results are presented in Section 5. We conclude the paper with a brief summary.

II. Speaker Class Normalization

A. Feature Normalization

Differences related to operation environment have made it not uncommon for acoustic feature mismatch to arise between a training set and a test set. This mismatch is part of what degrades performance under a real environment. Several approaches have been proposed to address this problem, including utterance-based cepstral mean normalization (CMN) and cepstral variance normalization(CVN). CMN and CVN for the i-th cepstral feature at frame t are

$$CMN \quad : \quad \hat{c}_i(t) = c_i(t) - \mu_i, \tag{1}$$

$$CVN \quad : \quad \hat{c}_i(t) = \frac{c_i(t)}{\sqrt{\sigma_i^2}} \tag{2}$$

where T is the total frames in the utterance, and μ_i and σ_i^2 are mean and variance of the i-th cepstral feature, respectively. Combining Eqs. (1) and (2), we obtain

$$CMVN : \hat{c}_i(t) = \frac{c_i(t) - \mu_i}{\sqrt{\sigma_i^2}}, \tag{3}$$

which suppresses the mismatch of acoustic features and enables us to build a robust system. We assert that these normalization methods can also be applied to mismatch among speakers. However, the completion of an utterance is required to process the CVN, and this constraint leads to a delay in the recognition process. In contrast, restriction of usable frames leads to the inaccurate estimation of statistics. To avoid this problem, the training data is divided into several classes and the distribution of cepstral features is modeled by the GMM with respect to each class[8]. Then, an utterance is classified to the nearest class using the first 50 frames, and the CMVN is applied using the mean and variance of the selected class. In this study, we first investigate how CMVN suppresses the variation of acoustic features caused by the diversity of speakers and how to improve the recognition accuracy.

B. Speaker Class Incorporation

DNN has the ability to model complex distribution. For this reason, we assume that the CMVN procedure may be modeled in the network automatically. Thus, we also trained the network with additional inputs which represents speaker information. Figure 1 shows an overview of how the DNN incorporates speaker-class information.

In the training, first we cluster the training data on the basis of acoustic feature similarity (see Section Ⅲ.B). After that, we compute a set of GMM likelihood between speaker-class GMMs and the first 50 frames of the target utterance. Those likelihoods are fed into the DNN as a speaker information. Therefore, the number of units added to the network is corresponding to the number of clusters.

III. Soft Clustering

A. Utterance to Speaker Clustering

In order to identify the speaker class, we use GMM likelihood in the logarithm domain as follows [9]:

$$L(X|\lambda_i) = \log p(X|\lambda_i) = \sum_{t=1}^{T} \log p(x_t|\lambda_i), \qquad (4)$$

where T is the available frames in the utterance (in the experiment, $T = 50$) and λ_i is the GMM for class i. Each GMM was trained by the training data of each class.

B. Division of Training Data

We assume that the generation of a more detailed cluster would further suppress the variation of acoustic features. Therefore, we increase the number of classes using soft-clustering technique shown in [9]. When doing this, we use overlap-allowed clustering in order to prevent the reduction of training data in each class.

IV. Experimental Setup

A. Database

To ensure an age- and gender-independent speech recognition system, we used three types of corpus, summarized in Table I. The database used for the adult class is the ASJ+JNAS[10], [11] database consisting of 133 male and 164 female speakers aged 18 to 59. This corpus consists of 20,337 (≈33 h) and 25,056 (≈44 h) sentences uttered by males and females, respectively. The database for the elder class is the S-JNAS [12] database consisting of 151 male and 150 female speakers aged 60 to 90. This corpus consists of 24,081 (≈53 h) and 24,061 (≈53 h) sentences uttered by males and females, respectively. The database for the child class is the CIAIR-VCV [13] database consisting of 140 male and 138 female speakers aged 6 to 12. This corpus consists of 7,538 sentences and 3,993 words (≈11 h) and 7,744 sentences and 3,910 words (≈11 h) uttered by males and females, respectively. In the CIAIR-VCV corpus, the child class was mainly composed of speech obtained from the reading of fairy tales. However, the language model we used in the experiment was trained by newspapers. As a result, the child class's out-of-vocabulary rate was 13.8 or 13.6%, while the rates for the elder and adult class were 0.5% and 2.1%, respectively.

TABLE I
Training data.

AS+JNAS		
Gender	Male	Female
Age	18-59	18-59
# speakers	133	164
# utterances	20,337(≈ 33h)	25,056(≈ 44h)
S-JNAS		
Gender	Male	Female
Age	60-90	60-90
# speakers	151	150
# utterances	24,081(≈ 53h)	24,061(≈ 53h)
CIAIR-VCV		
Gender	Male	Female
Age	6-12	6-12
# speakers	140	138
# utterances	7,538(+3993,≈ 11h)	7,744(+3910, ≈ 11h)

Each corpus contains male and female speech data, so we divide the training data into six basic classes: adult-male (AM), adult-female (AF), elder-male (EM), elder-female (EF), child-male (CM), and child-female (CF). Test data for each class were 100 sentences. The average number of frames per utterance is 540 frames. Although our aim is to recognize a short utterance, there was no appropriate test set to evaluate speaker adaptation/recognition on short utterance. Therefore, we restrict the available frames to the first 50 frames per utterance and calculate speaker class or GMM likelihoods from those data. The beginning of the speech in utterance was detected manually. In the experiment, we refer to these initial speaker class as *6-class-init*.

B. Acoustic Models

The speech was analyzed using a 25-ms Hamming window with a pre-emphasis coefficient of 0.97 and shifted with a 10-ms frame advance.

1) Syllable-Based Acoustic Model: The basic unit of Japanese is the syllable and there are 116 context-independent syllables in total. In this study, we used left context (vowels and pause: a, i, u, e, o, N, qs, sil), which leads to 928 left context-dependent syllables in total [14]. Each HMM consists of four states, so the number of output units increases to 928 $\times$ 4. To reduce the number of output units, we used tied 3 state syllables (TC3), which tied the latter three states of the syllable. If the latter three states are tied, only the first state is a left context-dependent syllable (or states) and the others consist of context-independent syllables (or state) (#output units: $1 \times 928 + 3 \times 116 = 1276$).

2) GMM-HMM: In the context-independent 116 syllable-based HMM training, we used the EM algorithm, after which 928 context-dependent syllable-based HMMs were trained using the MAP estimation algorithm. Each HMM has four states and each distribution is represented with 32 mixture diagonal Gaussians. When we train the model using all classes, the distribution of each HMM is represented with 128 mixture diagonal Gaussians. The feature vector consists of 12 MFCCs along with their first and second derivatives and the first and

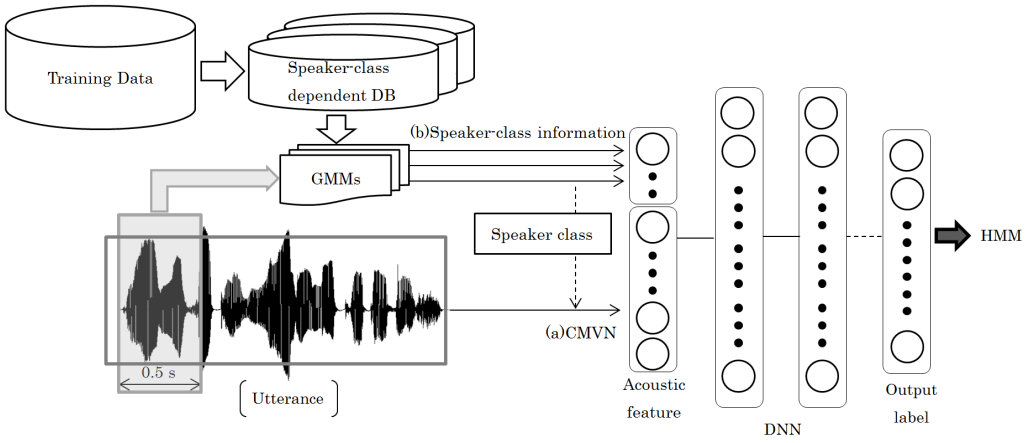

Fig. 1. Overview of speaker class incorporation for short time utterance speech recognition.

second derivatives of the logarithm power. These were trained with HTK[15].

3) DNN-HMM: For the DNN-HMM, we used 12 MFCCs along with their first and second derivatives and the first and second derivatives of the logarithm power across 11 frames. The features will be normalized to zero mean and unit variance using all training data except for the speaker-class based CMVN. The training targets were obtained from the forced alignment using the corresponding tied 3 state context-dependent syllable GMM-HMM. To reduce the computational time, the network was fine-tuned by using a rectifier function as an activation function. The network had the following architecture in all experiments: 428 input units, 3 hidden layers with 2,048 hidden units, and 1,276 output units.

C. Language Model and Decoder

A tri-gram based language model was trained on the Mainichi newspaper corpus (75 months, 11,533,739 words in total, vocabulary size of 20,000 words).

As the decoder, we used SPOJUS++(SPOken Japanese Understanding System) WFST version[16].

V. Experimental Results

A. GMM-HMM and DNN-HMM

(a) GMM-HMM

In the experiment, each class GMM-HMM was trained using the corresponding speaker class. The experimental results are listed in Tables II and III for the cases of known and unknown speaker class, respectively. When we trained one model using all training data, the average WER was 15.4%. The average WER of the known class model (6 GMMs) was 13.0%. These results show that the diversity of speakers decreases the performance. The first row in Table III shows how many frames were used to classify the utterance into speaker class. When we further increase the number of clusters (12 class soft), the average WER was 12.9% (all frames in an utterance) and 14.4% (50 frames), respectively. Compared to the 1class (1 GMM, baseline), the average WER of the 12-class-soft (12 GMMs) using 50 frames got better performance(14.4%). These results show the effectiveness of multiple acoustic models based on soft-clustering technique.

(b) DNN-HMM

Table II shows the experimental results for the case of known class using DNN-HMM. When we trained one model using all training data, the average WER was 11.2%. The average WER of known class models (6 DNNs) was also 11.2%. Compared to the GMM-HMM, the DNN-HMM was more robust against speaker variation.

B. Incorporation of Speaker Class Information

Lastly, we investigated the performance of the DNN-HMM that incorporated speaker-class information. In this experiment, as shown in Table III, we also used all or 50 frames of the utterance to identify the speaker class. When we focus on the 6-class-init with 50 frames, the average WER of the speaker-class-dependent CMVN was 10.9%, and with the addition of the likelihood it was 10.8%. These results show the incorporation of speaker-class information provides better results than the 1-class DNN-HMM (11.2%) even if the available time period was only 50 frames. The 6-class-init (DNN) with likelihood got the best performance, obtained a 4.0% relative gain compared with 1-class DNN (1 DNN) and 6-class-init (6 DNNs), that is, from 11.2% to 10.8%. Additionally, we conducted a significance test between 1 class (baseline, 1 DNN) and 6-class-init (likelihood, 1 DNN). As a result, these methods are statistically significant at the 10% level ($p = 0.084$). When we combined the two methods ($CMVN$ and $Likelihood$), the average WER of the combination method was not improved. These results show that the combination of two approaches does not provide complementary function. We also conducted experiments using 12-class-soft to investigate whether further increase of cluster could lead the improvement

TABLE II
Word error rate for the GMM-HMM and DNN-HMM (Class: Known)[%].

Model	Training data	AM	AF	EM	EF	CM	CF	Ave.
GMM	1class (1 GMM, baseline)	9.2	8.3	10.4	8.3	32.3	24.2	15.4
	6 class init (6 GMMs)	6.5	6.4	10.4	6.6	26.8	21.1	13.0
DNN	1 class (1 DNN, baseline)	5.5	4.5	7.1	6.2	23.5	20.0	**11.2**
	6 class init (6 DNNs)	6.1	4.9	7.1	5.3	22.8	21.1	**11.2**

TABLE III
Word error rate for the GMM-HMM and DNN-HMM (Class: Unknown)[%].

Model	Training data	All frames							50 frames						
		AM	AF	EM	EF	CM	CF	Ave.	AM	AF	EM	EF	CM	CF	Ave.
GMM	6 class init (6 GMMs)	6.8	7.5	11.2	9.1	28.0	22.3	14.1	8.3	10.2	17.2	8.2	28.9	23.9	16.1
	12 class soft (12 GMMs)	7.1	5.9	8.8	7.0	28.4	20.4	12.9	8.7	6.0	11.3	8.1	28.9	23.7	14.4

Model	Training data	Speaker-class Incorporation	All frames	50 frames						
			Ave.	AM	AF	EM	EF	CM	CF	Ave.
DNN	6 class init (1 DNN)	CMVN(class)	10.8	5.5	4.3	6.5	5.3	23.3	20.2	**10.9**
		Likelihood	10.8	5.3	4.3	6.7	5.2	23.1	19.9	**10.8**
		CMVN(class)+ Likelihood	11.1	5.5	4.3	7.0	6.5	23.7	20.5	11.2
	12 class soft (1 DNN)	CMVN(class)	18.0	5.5	4.7	6.8	4.9	24.4	20.9	11.2
		Likelihood	10.9	5.7	4.7	7.1	6.5	23.3	18.7	**11.0**
		CMVN(class)+Likelihood	11.3	5.6	4.2	7.0	5.2	25.0	21.2	11.4

as same as GMM-HMM. The performance of 12-class-soft (1 DNN) shows almost the same as the 6-class-init (1 DNN) unlike 12 class soft clustering GMMs. These results show the increase of clusters could represent more detailed speaker information. On the other hand, it suffers the lack of training data to achieve better generalization. The average WER of CMVN using 12-class-soft also decreased to 18.0%. This is considered that the increase of speaker class makes it difficult to estimate adequate mean and variance, because the number of training data for specific classes may decrease.

TABLE IV
Word error rate for the DNN-HMM (Class: Known).

Model	Training data	Ave.
DNN	CMVN(all frames in utterance)	10.2 %
	CMVN(50 frames in utterance)	20.6 %

We conducted several additional comparisons as a reference. Table IV shows the average WER obtained under various CMVN conditions for the case of known speaker-class. The procedure to train the DNN is equal to other speaker-class based CMVN except for normalization unit. When we applied $CMVN$ per utterance with available number of frames $T = allframes$ (about 640 frames), the average WER was 10.2% (oracle case). However, the CMVN using only 50 frames degraded the accuracy (20.6%), making it unsuitable for use with short time utterance. It is obvious that our proposed methods provide sufficient speaker information even if the uttered time is short.

VI. Conclusion

In this work, we investigated the use of speaker-class information to train the DNN. In experiments, both class-dependent CMVN and the additional input of speaker class information to the network outperformed the conventional DNN-HMM. These results demonstrate that speaker class, which was estimated from only the first 50 frames in an utterance, provides important information to suppress the diversity of speakers. However these experiments are conducted under clean and read speech, and the gains are relatively small. In the future, this approach is also applicable to actual environment such as noisy speech.

References

[1] G. Hinton, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," Signal Processing Magazine, IEEE, pp. 82-97, 2012.

[2] Y. Liu, P. Karanasou, T. Hain, "An investigation into speaker informed DNN front-end for LVCSR," ICASSP, pp. 4300-4304, 2015.

[3] G. Saon, H. Soltau, D. Nahamoo, M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," ASRU, pp. 55-59, 2013.

[4] O. A. Hamid, H. Jiang, "Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code," ICASSP, pp. 7942-7946, 2013.

[5] M. Tsujikawa, T. Nishikawa, T. Matsui, "Study on i-vector based speaker identification for short utterances," IEICE Technical Report, pp. 65-70, 2015 (in Japanese).

[6] M. Padmanabhan, L.R. Bahl, D. Nahamoo, M.A. Picheny, "Speaker clustering and transformation for speaker adaptation in large-vocabulary speech recognition systems," ICASSP, pp. 701-704, 1996.

[7] O. Viikki, D. Bye, K. Laurila, "A recursive feature vector normalization approach for robust speech recognition in noise," ICASSP, pp. 733-736, 1998.

[8] A.Y. Nakano, S. Nakagawa, K. Yamamoto, "Distant speech recognition using a microphone array network," IEICE Trans. on information and systems, pp. 2451-2462, 2010.

[9] D. Enami, F. Zhu, K. Yamamoto, S. Nakagawa, "Soft-clustering technique for training data in age-and gender-independent speech recognition," APSIPA, pp. 1-4, 2012.

[10] ASJ, http://research.nii.ac.jp/src/ASJ-JIPDEC.html

[11] K. Itou, et al., "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," The journal of the acoustical society of Japan(E), pp. 199-206, 1999.

[12] S-JNAS, http://research.nii.ac.jp/src/S-JNAS.html

[13] CIAIR-VCV, http://research.nii.ac.jp/src/CIAIR-VCV.html

[14] S. Nakagawa, K. Hanai, K. Yamamoto, N. Minematsu, "Comparison of syllable-based HMMs and triphone-based HMMs in Japanese speech recognition," ASRU, 1999.

[15] HMM Toolkit, http://htk.eng.cam.ac.uk/

[16] Y. Fujii, K. Yamamoto, S. Nakagawa, "Large vocabulary speech recognition system: SPOJUS++," MUSP, pp. 110-118, 2011.