Improvement of distant-talking speaker identification using bottleneck features of DNN

Takanori Yamada[1], *Longbiao Wang*[2], *Atsuhiko Kai*[1]

[1]Graduate School of Engineering, Shizuoka University, Hamamatsu 432-8561, Japan
[2]Nagaoka University of Technology, Nagaoka 940-2188, Japan

yamada@spa.sys.eng.shizuoka.ac.jp, wang@vos.nagaokaut.ac.jp, kai@sys.eng.shizuoka.ac.jp

Abstract

In this paper we propose bottleneck features of deep neural network for distant-talking speaker identification. The accuracy of distant-talking speaker recognition is significantly degraded under reverberant environment. Feature mapping or feature transformation has been shown efficacy in channel-mismatch speaker recognition. Bottleneck feature derived from multi-layer network, which is a nonlinear feature transformation method, has been shown efficacy in automatic speech recognition (ASR) system. In this study, bottleneck features extracted from deep neural networks (DNNs) which employ an unsupervised pre-training method are used as nonlinear feature transformation for distant-talking speech. The speaker identification experiment was performed on large-scale distant-talking speech set, with reverberant environments different to the training environments. The proposed bottleneck features achieved a relative error reduction of 46.3% compared with conventional MFCC. Moreover, a combination of likelihoods of bottleneck features and MFCC achieved a furthermore improvement.

Index Terms: speaker recognition, bottleneck features, pre-training, deep neural network, reverberant speech

1. Introduction

Due to the effects of reverberation, the accuracy of distant-talking speaker identification is significantly reduced. Many techniques such as dereverberation method, feature wraping and feature transformation have been proposed for robust distant-talking speaker recognition. The most general approach may be cepstral mean normalization (CMN) [1]. However, length of impulse response in a distant-talking environment is usually much longer than analysis window size of short-term spectral analysis. Therefore, the CMN cannot compensate for late reverberation. Some spectral subtraction-based dereverberation methods have been proposed for distant-talking speaker recognition [2-4]. They treated late reverberation as additive noise, and a noise reduction technique based on spectral subtraction was proposed to suppress the late reverberation. To construct a more robust representation of each cepstral feature distribution, feature warping method was proposed [2][5]. This method warps the distribution of a cepstral feature stream to a standardized distribution over a specified time interval. In addition, feature transformation approach was proposed for robust distant-talking speaker recognition [6]. The transformation is applied to the distorted feature before mapping them to a normal distribution, and aims to decorrelate the feature vectors, making them more amendable to diagonal covariance Gaussian mixture model (GMM).

Bottleneck features extracted by a multi-layer perceptron (MLP) can be used a non-linear feature transformation and dimentionality reduction [7]. An MLP was trained by backpropagation (BP) algorithm from random initial parameters. And then the bottleneck features were extracted by dimensionality reduction of several frames of cepstral coefficients. The combination of bottleneck features and cepstral coefficient is better than the conventional MFCC. However, deep networks of MLP with many hidden layers have a high computational cost, and can't learn much further away from the top layer. In addition, it does not lead to improved accuracy even after a long time training by the BP algorithm using the initial random value. Recently, deep belief networks (DBNs) which employ an unsupervised pretraining method using restricted Boltzmann machine (RBM) have been proposed to train better initial values of deep networks [8]. A RBM is obtained by adding constraints that there is no connection of the same layer and connected to the symmetry to the Boltzmann machine. Its parameter is determined by the Greedy layer-wise training. This training method minimize the error between the input feature and the inverse transform feature of the input. After pretraining, DNNs are obtained by discriminative trained using BP algorithm from DBNs which are multi-layer structure stacked RBMs. DNNs with pretraining have been shown better performance than the conventional MLP without pretraining on automatic speech recognition [9] and large vocabulary business search task [10] etc.

In this paper, we propose a robust distant-talking speaker identification method using the bottleneck features extracted from DNNs. We consider that the DNNs can transform the reverberant speech feature to a new feature space close to clean speech feature which has more discriminative classification ability for distant-talking speaker recognition. In addition, by using multiple contexts (frames) for input data, the bottleneck features can reduce the influence of reverberation over frames. Bottleneck features extracted from the DNNs are used for training the conventional speaker-dependent GMM. Furthermore, the likelihood of MFCC and that of the bottleneck features are combined linearly.

2. Deep Neural Networks

Although DNNs can be seen as improved MLPs, discriminative training method is the same. The advantages of DNN is pretraining which can determine the good initial value of the multi-layer networks. However, pretraining of multi-layer networks is computationally expensive. Therefore, we used RBM for efficient computation.

2.1. Restricted Boltzmann Machine

RBM is a bipartite graph shown in Figure 1. It has visible and

Copyright © 2013 ISCA

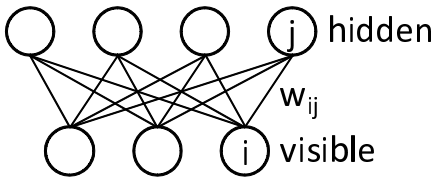

Figure 1: *Graphical representation of the RBM.*

hidden layer in which visible units that represent observations are connected to hidden units that learn to represent features using weighted connection. An RBM is restricted that there are novisible-visible or hidden-hidden connections. Different types of RBM is used in the case of binary or real-valued input. Bernoulli-Bernoulli RBMs used to convert binary stochastic variables to binary stochastic variables. Gaussian-Bernoulli RBMs is used to convert real-valued stochastic variables to binary stochastic variables.

In a Bernoulli-Bernoulli RBMs, the weights on the connections and the biases of the individual units define a probabillity distribution over the joint states of the visible and hidden units via an energy function. The energy of a joint configuration is:

$$E(\mathbf{v},\mathbf{h}|\theta) = -\sum_{i=1}^{\mathcal{V}}\sum_{j=1}^{\mathcal{H}} w_{ij}v_i h_j - \sum_{i=1}^{\mathcal{V}} a_i v_i - \sum_{j=1}^{\mathcal{H}} b_j h_j \quad (1)$$

where $\theta = (\mathbf{w}, \mathbf{a}, \mathbf{b})$ and w_{ij} represents the symmetric interaction term between visible unit i and hidden unit j while a_i and b_j are their bias term. $\mathcal{V}$ and $\mathcal{H}$ are the numbers of visible and hidden units.

The probability that an RBM assigns to a visible vector $\mathbf{v}$ is:

$$p(\mathbf{v}|\theta) = \frac{\sum_{\mathbf{h}} \exp(-E(\mathbf{v},\mathbf{h}))}{\sum_{\mathbf{v}}\sum_{\mathbf{h}} \exp(-E(\mathbf{v},\mathbf{h}))} \quad (2)$$

Since there are no hidden-hidden connections, the conditional distribution $p(\mathbf{h}|\mathbf{v},\theta)$ is factorial and is given by:

$$p(h_j = 1|\mathbf{v},\theta) = \sigma(b_j + \sum_{i=1}^{\mathcal{V}} w_{ij}v_i) \quad (3)$$

where $\sigma(x) = (1+\exp(-x))^{-1}$. Similarly, since there are no visible-visible connections, the conditional distribution $p(\mathbf{v}|\mathbf{h},\theta)$ is factorial and is given by:

$$p(v_i = 1|\mathbf{h},\theta) = \sigma(a_i + \sum_{j=1}^{\mathcal{H}} w_{ij}h_j) \quad (4)$$

In a Gaussian-Bernoulli RBMs, the energy of a joint configuration is:

$$E(\mathbf{v},\mathbf{h}|\theta) = \sum_{i=1}^{\mathcal{V}} \frac{(v_i - a_i)^2}{2} - \sum_{i=1}^{\mathcal{V}}\sum_{j=1}^{\mathcal{H}} w_{ij}v_i h_j - \sum_{j=1}^{\mathcal{H}} b_j h_j \quad (5)$$

The conditional distribution $p(\mathbf{h}|\mathbf{v},\theta)$ is factorial and is given by:

$$p(v_i = 1|\mathbf{h},\theta) = \mathcal{N}(v_i; a_i + \sum_{j=1}^{\mathcal{H}} w_{ij}h_j, 1) \quad (6)$$

where $\mathcal{N}(\mu, V)$ is a Gaussian with mean μ and variance V.

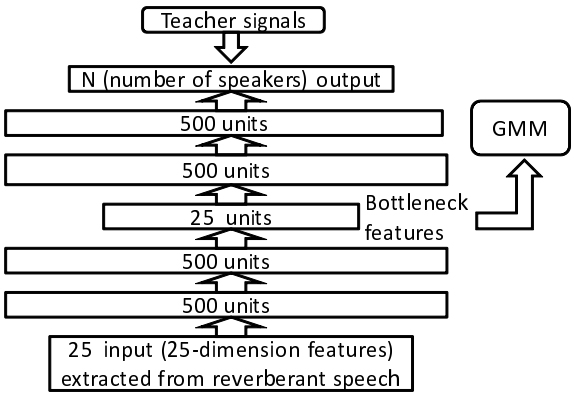

Figure 2: *Flowchart of the bottleneck feature extraction.*

Maximum likelihood estimation of RBM is to maximize the log likelihood log$p(\mathbf{v}|\theta)$ for the parameters θ. Therefore, the weight update equation is given by:

$$\Delta w_{ij} = \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model} \quad (7)$$

where $\langle\cdot\rangle_{data}$ is the expectation that v_i and h_j are on together in the training set and $\langle\cdot\rangle_{model}$ is the same expectation calculated from the model. Because compute $\langle v_i h_j \rangle$ is expensive, using contrastive divergence (CD) approximation for the compute gradient. It is possible to compute $\langle v_i h_j \rangle$ by once the Gibbs sampling.

2.2. DNNs structure and training

Deep Belief Networks (DBNs) is configured hierarchically by connecting the pretrained RBM. In order to obtain a pretrained RBM, we trains Gaussian-Gaussian RBM first, and later trains the Bernoulli-Bernoulli RBM. The top layer of DNNs uses a softmax layer. The softmax operation is given by:

$$p(l|\mathbf{h}) = \frac{\exp(b_l + \sum_i h_i w_{il})}{\sum_m \exp(b_m + \sum_i h_i w_{im})} \quad (8)$$

where b_l is the bias of the label and w_{il} is the weight from hidden unit i in top layer to label l.

After configure the DBNs using RBM, DBNs is discriminative trained by using BP algorithm to maximize the log probability of the class labels. In general, the DBNs after discriminative training are called DNNs.

3. Bottleneck Features

Bottleneck features are extracted from one of the internal layers of the multi-layer network. Multi-layer network to obtain the bottleneck features is shown in Figure 2. The number of hidden units of the internal layers is smaller than the other layers. This structure bottleneck layer can be treated as a low-dimensional nonlinear function of input features. In addition, it is possible to enhance the identification ability of bottleneck features by discriminative training, and it is expected to mitigate the influence of reverberation for speaker identification.

Both of MLP without pretraining and DNNs with pretraining are used as multi-layer networks. The initial value of MLP is generated randomly from -0.5 to 0.5 and the initial value of DBNs is determined by the unsupervised pretraining. After initialization, a supervised discriminative training is performed for

Table 1: *Details of recording conditions for impulse response measurement. "RT60 (s)": reverberation time in room.*

impulse response no	room	RT60 (s)
(a) CENSREC-4 database for training		
1	Japanese style room	0.40
2	Japanese style bath	0.60
3	elevator hall	0.75
(b) RWCP database for test		
4	tatami-floored room	0.47
5	echo room (panel)	1.30

both MLP without pretraining and DBNs with pretraining. Finally, the bottleneck features extracted from the bottleneck layer of DNNs are used to train speaker-dependent model.

4. Experiments

4.1. Experimental Setup

In this study, the bottleneck features for distant-talking speaker identification were extracted from MFCC features. 1-frame of 25-dimensional MFCC features was used as the input of the DNNs. There are 25 hidden units in the bottleneck layer and 500 hidden units in non-bottleneck hidden layer. The number of hidden-layer of DNNs is 5. MFCC features were normalized with mean of the entire training data. The DNN training was carried out using stochastic mini-batch gradient descent with a minibatch size of 100 samples. 50 epochs with a learning rate of 0.1 were used for all layers during pretraining and 1000 epochs with a learning rate of 0.1 were used for all layers during finetuning.

Distant-talking speaker identification was evaluated in artifical reverberant speech for the sake of convenience. Five kinds of multi-channel impulse responses were selected from Real World Computing Partnership (RWCP) sound scene database for test [11] and the CENSREC-4 database for training [12], which were convoluted with clean speech to create artifical reverberant speech. A large scale database, Japanese Newspaper Article Sentence (JNAS) [13] corpus, was used as clean speech. The utterances of training set is composed of 50 male speakers with 10 utterances taken from each speaker. 20 utterances of each speaker were used as test data. The mean length of utterance of the training set was 3.90 sec, and that of the test set was 5.63 sec.

Table 1 lists the impulse responses for train set and test set. Impulse responses were measured at position 2.0 m for the RWCP database and 0.5 m for the CENSREC-4 database from the microphone, respectively.

Table 2 gives the conditions for speaker recognition. 25-dimension MFCCs and GMMs with 128 mixtures were used. GMMs were trained using one kind of reverberant speech corresponding to one kind of impulse response.

Three kinds of methods were compared in this study. Method 1 is to train the GMM using MFCC and used as a baseline (denoted as 'MFCC'). Method 2 uses the bottleneck features extracted from the MLP without pretraining (denoted as 'BF-MLP'). Method 3 uses the bottleneck features extracted from the DNNs with pretraining (denoted as 'BF-DNNs').

Table 2: *Conditions for speaker recognition.*

sampling frequency	16 kHz
frame length	25 ms
frame shift	10 ms
feature space	25 dimensions with CMN (12 MFCCs + Δ + Δ power)
acoustic model	GMMs with 128 diagonal covariance matrices

Table 3: *Distant-talking speaker identification rates (%).*

Features	RT60 for training (s)			Ave.
	0.40	0.60	0.75	
(a) RT 60 for test = 0.47 s				
MFCC	90.4	76.5	87.3	84.7
BF-MLP	83.2	91.6	92.6	89.1
BF-DNNs	92.9	91.1	92.8	92.3
(b) RT 60 for test = 1.30 s				
MFCC	89.3	81.6	85.1	85.3
BF-MLP	78.2	91.3	94.4	88.0
BF-DNNs	90.9	90.6	93.3	91.6

4.2. Experimental Results

Table 3 show the result of distant-talking speaker identification. In Table 3, BF-DNNs outperformed the MFCC for all conditions, and the average relative error reduction rate was 46.3%. Although there are large variations in the performance of the MFCC in terms of the training set, BF-DNNs based method has relative small variance and it is robust in all reverberant environments. The result of MFCC using training data with 0.60 s reverberation time is not good. The reason may be that the Japanese style bath is a special reverbrant environment. On the other hand, the results of BF-DNNs trained by Japanese style bath environment are close to that of other conditions. This indicates that the bottleneck features can potentially transform the feature space of reverberant speech to the feature space of the relative clean speech. Comparing with the BF-DNNs and BF-MLP, BF-DNNs based method is better than BF-MLP based method in average. However, BF-MLP is better than BF-DNNs when reverberant speech with RT 60 of 0.60 s or 0.75 s is using as training data. Therefore, it is necessary to deal with the reverberant speech with long reverberation time for effective pretraining. It is considered that applying dereverberation before bottleneck feature extraction-based transformation may improve the overall performance.

A linear combination of the likelihood of MFCC and BF-DNNs was also evaluated. The weights of MFCC likelihood and BF-DNNs are 0.2 and 0.8, respectively, which are determined empirically. The results are shown in Table 4. In Table 4, the combination method was better than all of the individual methods.

5. Conclusions and Future Work

In this paper, we proposed a robust distant-talking speaker identification method by using bottleneck features. Bottleneck features extracted from MLP and DNNs were used to train the GMM for speaker identification. The proposed bottleneck features of DNNs achieved a relative error reduction of 46.3% compared with conventional MFCC. Comparing with

Table 4: *Distant-talking speaker identification rates using combined score of MFCC and BF-DNNs (%).*

RT60 for training (s)	MFCC	BF-DNNs	Combination
(a) RT 60 for test = 0.47 s			
0.40	90.4	92.9	93.7
0.60	76.5	91.1	91.3
0.75	87.3	92.8	93.5
Ave.	84.7	92.3	92.8
(b) RT 60 for test = 1.30 s			
0.40	89.3	90.6	92.9
0.60	81.6	90.6	91.4
0.75	85.1	93.3	94.7
Ave.	85.3	91.6	93.0

MLP without pretraining and DNNs with pretraining, the results showed that the pretraining was effective for distant-talking speaker identification. Moreover, the speaker recogniton performance was furthermore improved by combining the likelihoods of bottleneck features and MFCC.

In this study, we did not use the benefits of DNNs ability to deal with context of multiple frames which is expected to reduce the influence of the reverberation caused over multiple frames. In the future, we will use multiple frames of features as input data of multi-layer network. In addition, so far we only used a single reverberant environment in the training set to train speaker model. We will try to train a reverberant model with multiple reverberant environments in the training set.

6. References

[1] S. Furui, "Cepstral Analysis Technique for automatic speaker verification, " IEEE Trans. Acoust. Speech Signal Process., vol. 29, no. 2, pp. 254-272, 1981.

[2] Q. Jin, T. Schultz and A. Waibel, "Far-field speaker recognition, " IEEE Trans. ASLP, vol. 15, no. 7, pp. 2023-2032, 2007.

[3] Y. Pan, "Robust Speech Recognition on Distant Microphones, " Thesis in submission, Carnegie Mellon University, 2007.

[4] L. Wang, Z. Zhang and A. Kai, "Hands-free speaker identification based on spectral subtraction using a multi-channel least mean square approach, " Proc. of ICASSP 2013 (Accepted).

[5] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification, " Proc. of Speaker Odyssey 2001 conference, June, 2001.

[6] D. Zhu, B. Ma, H. Li and Q. Huo, "A generalized feature transformation approach for channel robust speaker verification, " Proc. of ICASSP 2007, April, 2007.

[7] Y. Konig, L. Heck, M. Weintraub and K. Sonmez, "Nonlinear discriminant feature extraction for robust text-independent speaker recognition, " Proc. of RLA2C, ESCA workshop on Speaker Recognition and its Commercial and Forensic Applications, pp. 72-75, 1998.

[8] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks, " Science, vol. 313, no. 5786, pp. 504-507, 2006.

[9] A. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic Modeling using Deep Belief Networks, " IEEE Trans. on Audio Speech and Language Process., vol. 20, pp. 12-22, 2012.

[10] D. Yu and M. L. Seltzer, "Improved Bottleneck Features using Pretrained Deep Neural Networks, " Proc. of Interspeech2011, pp. 237-240, 2011.

[11] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura and T. Yamada, "Acoustical Sound Database in Real Environments for Sound Scene Understanding and Hands-Free Speech Recognition, " Proc. of LREC2000, pp. 965-968, May, 2000.

[12] M. Nakayama et al., "CENSREC-4: Development of Evaluation Framework for Distant-talking Speech Recognition under Reverberant Environments, " Proc. of Interspeech2008, pp. 968-971, Sep. 2008.

[13] K. Itou, M. Yamamoto, K. Takeda, T. Kakezawa, T. Matsuoka, T. Kobayashi, K. Shikano, S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research, " J. Acoust Soc Jpn (E). 20(3), 199-206, 1999.