MISSING FEATURE RECONSTRUCTION METHODS FOR ROBUST SPEAKER IDENTIFICATION

Xueliang Zhang *Hui Zhang* *Guanglai Gao*

Computer Science Department, Inner Mongolia University, Hohhot, China, 010021
cszxl@imu.edu.cn alzhu.san@163.com csggl@imu.edu.cn

ABSTRACT

In this study, we propose a reconstruction method to restore the degraded features for robust speaker identification. The proposed method is based on a hybrid generative model which consists of deep belief network (DBN) and restricted Boltzmann machine (RBM). Specifically, the noisy speech is firstly decomposed into time-frequency (T-F) representations. Then ideal binary mask (IBM) is computed to indicate each T-F point as reliable or unreliable. We reconstruct the unreliable ones by the proposed model iteratively. Finally, reconstructed feature is utilized to conventional speaker identification system. Experiments demonstrate that the proposed method achieves significant performance improvements over previous missing feature techniques under a wide range of signal-to-noise ratios.

Index Terms— Robust speaker identification, Missing feature techniques, Restricted Boltzmann machine, Deep belief network

1. INTRODUCTION

Speaker identification (SID) plays an important role in applications of security and access control. A typical speaker identification system includes feature extraction, speaker identity modeling and decision making using pattern classification methods. Commonly used speaker feature is short-time cepstral coefficients such as Mel-frequency cepstral coefficients (MFCCs) and recently proposed gammatone frequency cepstral coefficients (GFCCs) [1]. For speaker modeling, Gaussian mixture models (GMMs) are often used to describe the feature's distribution of speakers. According to the likelihoods of observing features given the speaker models, recognition decisions are made. However, such speaker recognition systems perform poorly when the input speech is corrupted by environmental noise, especially when the systems are trained on clean speech.

To deal with the robustness problem, filtering techniques such as spectral subtraction [2] are used which assume a priori knowledge of the noise spectrum. However, these methods do not perform well when the noise is non-stationary and this requirement limits the scope of the application. Other techniques rely on a statistical model of the noise, such as parallel model combination (PMC) [3]. Although PMC has been shown to be effective in stationary and non-stationary noise cases [3], it still needs noise knowledge to perform the adaptation.

Missing feature techniques are based on the observation that speech signals have a high degree of redundancy [4]. It means that, when knowledge of the noise is unavailable, one may ignore the severely corrupted speech data and base the recognition on the little corrupted data, or reconstruct the corrupted data by relatively clean ones [5]. These two approaches are called marginalization and reconstruction [4]. In reconstruction, the aim is to estimate the values for the unreliable (corrupted) components by conditioning on the reliable ones and the speech model. After producing a complete observation feature, recognition procedure performs in a conventional way. In marginalization, recognition is mainly based on reliable (uncorrupted) components with integrating over the unreliable ones.

Marginalization needs to modify the conventional recognition system and it can only work on the spectrographic features which are known to be less optimal than cepstral. At the same time, it is more time consuming for classification. By contrast, data reconstruction can use reconstructed spectrograms to generate cepstral vectors. Therefore, it can serve as front-end processing of conventional recognition system and perform more rapidly.

In this study, we follow the missing feature technique for robust speaker identification system by using reconstruction approach. For reconstruction, the main issue is to build speech prior model. Different with conventional method [4,6] using GMMs as the speech model, we proposed reconstruction methods based on a hybrid model which is composed of a RBM and a DBN. As a generative model, RBM and DBN are considered to be more powerful than GMM [7]. Generally speaking, it is helpful for reconstruction with a more accurate speech model.

The paper is organized as follows. In the next section, we present an overview of the proposed system. Section 3 gives the systemic evaluation. We conclude the paper in Section 4.

2. SYSTEM DESCRIPTION

2.1. Background of missing feature reconstruction

To reconstruct the missing feature, the feature vector x is firstly divided into reliable part x_r and unreliable part x_u : $x = (x_r, x_u)$. Reconstruction can be formulated as solving the parameter optimization problem:

$$\arg\max_{x_u} (p(x_r, x_u)) \tag{1}$$

where x_u is the parameter and x_r is the constraint. It means to find a configuration of x_u which leads to the maximal probability of joint configuration (x_u, x_r). Therefore, the first thing is to model the speech prior model $p(x)$. The second thing is to reconstruct unreliable features based on the speech prior model.

Previous researches, such as in [1,4,8,9], employed GMM to describe $p(x)$, which is trained using pooled training data. For reconstruction, x_u are estimated as the expectation of Gaussian component's mean conditioned on x_r. Increasing the number of mixture components of GMM is obviously helpful for reconstruction, which leads to more accurate description of $p(x)$.

2.2. Modeling the $p(x)$ by RBM

RBM is a type of undirected graphical model constructed from a layer of hidden units and a layer of visible units with no visible-visible or hidden-hidden connections. RBM defines the joint probability of the visible and hidden units as

$$p(v, h) = \frac{1}{Z} e^{-E(v,h)} \tag{2}$$

where v and h denote a visible and hidden layer respectively. Z is the partition function to ensure $p(v, h)$ is a probability distribution.

If we assume visible units are Gaussian random variables with unit variance, we can define the energy function E as formula (3). While if the visible units are Bernoulli random variables, E is formulated as (4). For real-valued features, we use Gaussian-Bernoulli RBM.

$$E(v, h) = \sum_{i \in v} (v_i - a_i)^2 - \sum_{j \in h} b_j h_j - \sum_{i,j} v_i h_j w_{ij} \tag{3}$$

$$E(v, h) = -\sum_{i \in v} a_i v_i - \sum_{j \in h} b_j h_j - \sum_{i,j} v_i h_j w_{ij} \tag{4}$$

where v_i and h_j denote the ith and jth units of v and h, a_i and b_j denote the bias of visible layer and hidden layer respectively, and w_{ij} denotes the weight between v_i and h_j.

Because there are no direct connections between hidden units, hidden unit h_j is independent conditional on v.

$$p(h_j = 1|v) = \sigma(b_j + \Sigma_i w_{ij} v_i) \tag{5}$$

where $\sigma(x) = (1 + e^{-x})^{-1}$ is sigmoid function.

Similarly, visible unit v_i is independent conditional on h. If v is Gaussian random variable, its conditional distribution is defined as (6). And if v is Bernoulli random variable, its conditional distribution is defined as (7).

$$p(v_i|h) = \mathcal{N}(v_i; a_i + \Sigma_j w_{ij} h_j, 1) \tag{6}$$

$$p(v_i = 1|h) = \sigma(a_i + \Sigma_j w_{ij} h_j) \tag{7}$$

where $\mathcal{N}$ denotes the normal distribution.

We use Gaussian-Bernoulli RBM to describe the speech prior model. The acoustic features are normalized to unit variance. During the training phase, the parameters a_i, b_j, w_{ij} of RBM are initialized by small random values. The RBM is trained by the contrastive divergence method. The details could be found in [10].

2.3. Modeling $p(x)$ by Hybrid Model

The structure of the proposed model illustrated in Fig. 1, which is comprised of a DBN and a RBM. The training phase has two parts. The one is to build the DBN. The other is to create the RBM. We train the DBN layer by layer as RBMs. The first layer of DBN is Gaussian-Bernoulli RBM and the higher layers are Bernoulli-Bernoulli RBMs. Similarly, the input acoustic features are normalized to unit variance. We get the top layer representation of DBN by propagating the acoustic features through three layers using the sequence of posterior distributions (5). Then we combine the top layer representation of DBN with acoustic feature together to train a new RBM. It should be mentioned that part of the visible units of this RBM are Gaussian random variables (corresponding to acoustic feature) and part of ones are Bernoulli random variables (corresponding to top representation of DBN). It can be proven that the conditional distribution and the posterior distribution can still be represented by formula (5) and (6), (7) respectively, because of the conditional independence between visible and hidden units.

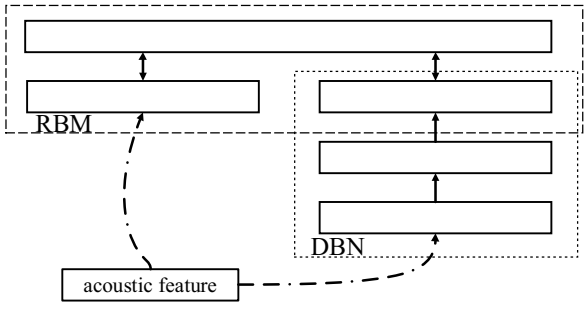

Fig. 1. the structure of the proposed method for speech prior model

3. EXPERIMENT AND RESULTS

3.1. Database

The systems are evaluated using the 2002 NIST Speaker Recognition Evaluation corpus [11]. The dataset contains 330 speakers. Each speaker has an around 2-minute-long telephone recording with 8kHz sampling rate. As in [6], each recording is divided into 5-second-long pieces, and 2 of them are employed to test and remaining ones are used for training. To study the performance of the proposed system, the test utterances are mixed with 5 different types of noise, which are speech shaped noise (SSN), white noise, babble noise, factory noise and cocktail party noise. Each noise is mixed with test utterances at 5 signal-to-noise ratio (SNR) levels from -6 dB to 18 dB at 6-dB intervals.

3.2. Speaker Identification System

The missing feature techniques can be used as a front-end processing. Therefore, we construct a conventional SID system as in [6].

64-dimensional Gammatone feature (GF) of input speech is first extracted. As in [6], we convert channels 11-64 of GF to GFCC using discrete cosine transform (DCT) and keep the lower 23-order of GFCC coefficients.

The speaker identification system employs Gaussian mixture model (GMM) and universal background (UBM). Specifically, UBM is a GMM with 1024 Gaussian components which is trained using all the pooled training data. Each speaker model is adapted from UBM using the utterances of individual speaker [12]. Compared with individually trained GMMs, GMM-UBM scores much faster and is more discriminative. In Fig. 2, we show the recognition rates of GMM-UBM on unprocessed noisy utterances.

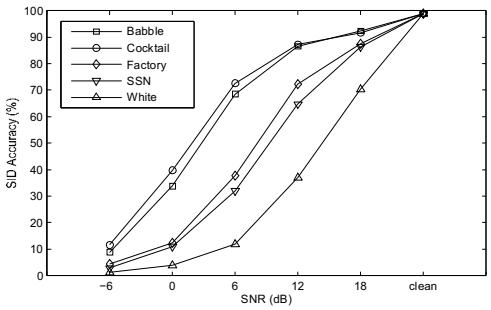

Fig. 2. SID performance of baseline system for different noises on various SNR.

3.3. Mask Computation

Before applying reconstruction method, we should firstly mark the reliable and the unreliable part of acoustic feature. As noted above, SID system is built on GFCC which is converted from GF. Therefore, we employ the ideal binary mask (IBM) to indicate the reliable and the unreliable part. IBM is defined as follows [6]:

$$IBM(t,f) = \begin{cases} 1 & \text{if } SNR(t,f) > 0 \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

T-F units are indexed by time t and frequency f, where (t, f) denote the T-F unit in time frame t and frequency channel f. $IBM(t, f)$ and $SNR(t, f)$ are the IBM and the local signal-to-noise ratio of the T-F unit, respectively. when given premixed target and interference signals, the IBM can be easily computed. Here, we marked a feature as the reliable part if $IBM(t, f) = 1$ or as the unreliable part if $IBM(t, f) = 0$.

3.4. Comparison of Reconstruction Approaches

We should mention that reconstruction approaches try to impute corrupted GF not GFCC. The whole procedure of robust speaker identification is as follows: 1) training the speech model using 54-dimentional GF (channels of 11-64); 2) using the IBM to indicate the reliable and unreliable units on GF map of noisy speech; 3) reconstructing corrupted GF by speech models; 4) converting the reconstructed GF into GFCC and recognizing the speaker by SID system.

To evaluate the performance of the proposed reconstruction method, other two methods are presented here. One is GMM-based reconstruction method [5]. And the other is based on RBM, as described in section 2.2.

If there is no reliable unit in a frame, the optimization problem has no constraint. Therefore, we have to perform frame selection before reconstruction. Our method is quite simple. We choose a frame if it has at least one reliable unit.

The GMM-based reconstruction approach is implemented according to [4]. We train a 2048-component GMM with diagonal covariance using all the pooled training data. The posteriori probability of the Gaussian component is calculated given the reliable GF units.We estimate the unreliable GF units as the expectation of Gaussian components' mean. More details could be found [4, 6].

The RBM-based approach employs a Gaussian-Bernoulli restricted Boltzmann machine with 200 hidden units to build speech model. One iteration of contrastive divergence [10] is utilized to train the RBM. The learning rate is 0.001 and batch size is 100. During the reconstructing phase, we firstly initialize the input feature by keeping the reliable part x_r and setting the unreliable part x_u to zero. We compute the hidden layer representation conditioned on the input feature by formula (5). Then we reconstruct the feature by formula (6). After that, the x_u is replaced by the corresponding part of reconstructed feature. The entire procedure runs several times. We also analyze the influence of RBM with different hidden units to SID accuracy, which could be seen in Fig. 4.

The proposed reconstruction method is comprised of a DBN and a RBM. The DBN has three layers which are one visible and two hidden layers. The size of the visible layer corresponds to GF vector size which is 54. The sizes of the second and the third layers are both 50. To train the RBM of hybrid model, we combine the GF and its top layer representation of DBN. Hence, the size of the visible layer of RBM is 104 (54-dimensional GF plus 50-dimensional top layer representation of DBN). The size of the hidden layer of RBM is set to 100. During the reconstructing phase, the acoustic feature is initialized by keeping the reliable part x_r and setting the unreliable part x_u to zero. We propagate the acoustic feature through DBN. Then we combine the top layer representation of DBN with the acoustic feature as the input of RBM. The unreliable part x_u is filled with the reconstructed feature of RBM. The entire procedure runs several times until the reconstructed feature becomes stable.

Fig. 3. illustrates the reconstructed features using the three methods. The features reconstructed by GMM are too smooth to lose the details. Results of RBM without over-smoothing are better than GMM's. But the RBM only reconstruct the local structure and lost the big picture. DBN's results are more close to the expected ones. And we can see the DBN have reconstructed the big picture and not to be over-smoothing.

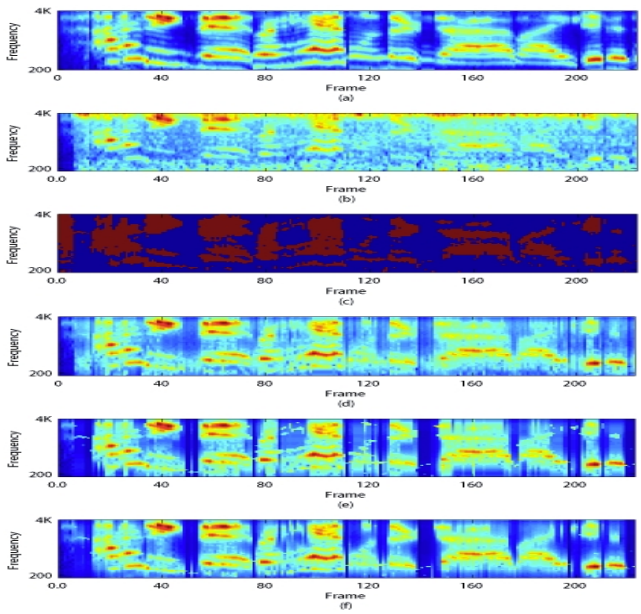

Fig. 3. Comparison of reconstruction methods for a utterance mixed with factory noise at 0dB. (a) cochleagram of the utterance. (b) cochleagram of the mixture. (c) Ideal Binary Mask of the mixture. (d) Reconstructed cochleagram by GMM. (e) Reconstructed cochleagram by RBM. (f) Reconstructed cochleagram by the proposed method.

3.5. Evaluation Results

We employ SID accuracy rate as the metric to evaluate the three reconstruction methods. Table I shows the SID performances of the three reconstruction methods: the GMM-based, RBM-based and the proposed method. As shown in the table, the proposed method obtains the best results for all types of noisy utterances. Compared with GMM-based and RBM-based methods, the average SID accuracy rates of the proposed method are improved 8.2% and 3.4% respectively. Especially in low SNR cases, the improvements are more obvious. We also compare the numbers of parameters of three models. In Table II, we can see that the GMM has the largest number of parameters among three models, but its performance is the worst. It shows that RBM-based and DBN-based method are more powerful to build speech module. Meanwhile, the number of parameters of the proposed model is a bit more than that of RBM.

To compare the RBM and the proposed model further, we investigate the performance of RBMs with different numbers of hidden units. In Fig. 4, it shows the average SID rate on all kinds of noisy conditions and all SNR levels. As shown in Fig. 4, enlarging the hidden units improves the SID performance when the hidden layer size is small (from 10 to 50). However, the performance didn't change significantly as the size keeping increasing. The reason is that in DBN, the local characteristics are taken care of using the lower layers while higher-order and highly non-linear statistical structure in the input is modeled by the higher layers [7]. We exploit the top layer representations of DBN as additional constraints in the parameter optimization problem.

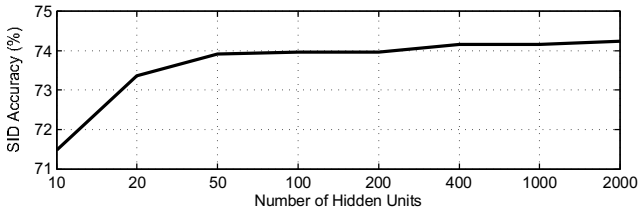

Fig. 4. performances of RBM reconstruction method with different number of hidden units.

4. CONCLUSION

In this paper, we propose reconstruction method based on a hybrid model to improve the robustness of speaker recognition systems. The experiment results show that the proposed method outperforms the conventional GMM-based method. So far as we know, this is the first time to reconstruct the corrupted speech feature using RBM and DBN for robust speaker identification.

Table 1. SID accuracy (%) of the three reconstruction methods. GMM-based, RBM-based and proposed method.

Babble	**-6dB**	**0dB**	**6dB**	**12dB**	**18dB**	**Avg.**
GMM	35.61	66.82	84.85	90.90	95.00	74.64
RBM	53.79	75.91	87.73	93.03	95.76	81.24
Proposed	55.91	79.55	90.75	92.88	95.91	83.00

Cocktail	**-6dB**	**0dB**	**6dB**	**12dB**	**18dB**	**Avg.**
GMM	39.39	69.85	86.51	90.76	95.15	76.33
RBM	56.67	79.39	88.18	93.03	95.91	82.64
Proposed	57.58	80.15	89.70	93.33	95.61	83.27

Factory	**-6dB**	**0dB**	**6dB**	**12dB**	**18dB**	**Avg.**
GMM	22.73	53.79	76.67	87.88	94.24	67.06
RBM	39.24	64.09	80.91	88.64	92.12	73.00
Proposed	41.82	71.97	86.21	91.97	96.76	77.47

SSN	**-6dB**	**0dB**	**6dB**	**12dB**	**18dB**	**Avg.**
GMM	18.33	44.39	71.82	85.91	88.94	61.88
RBM	27.88	52.42	70.61	79.39	86.36	63.33
Proposed	31.36	63.64	83.33	91.06	95.00	72.88

White	**-6dB**	**0dB**	**6dB**	**12dB**	**18dB**	**Avg.**
GMM	23.33	53.48	76.06	85.76	93.18	66.36
RBM	26.82	57.27	79.09	89.09	94.09	69.27
Proposed	25.30	55.00	80.30	90.45	95.45	69.30

Table 2. the number of parameters of three reconstruction methods.

	GMM	**RBM**	**Proposed**
Number of parameters	54×2048×2	54×200	54×50+2×50×50 +(54+50) ×100
Total	221K	11K	18K

5. ACKNOWLEDGEMENTS

This research was supported in part by the China National Nature Science Foundation (No.61263037, No.61365006) and the University Science Research Project No.NJZY13007.

REFERENCES

[1] Yang Shao and DeLiang Wang, "Robust speaker identification using auditory features and computational auditory scene analysis," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 1589–1592.

[2] Javier Ortega-García and Joaquín González-Rodríguez, "Overview of speech enhancement techniques for automatic speaker recognition," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*. IEEE, 1996, vol. 2, pp. 929–932.

[3] Lit Ping Wong and Martin Russell, "Text-dependent speaker verification under noisy conditions using parallel model combination," in *Acoustics, Speech, and Signal Processing, 2001. ICASSP 2001. IEEE International Conference on*. IEEE, 2001, vol. 1, pp. 457–460.

[4] Martin Cooke, Phil Green, Ljubomir Josifovski, and Ascension Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech communication*, vol. 34, no. 3, pp. 267–285, 2001.

[5] Roberto Togneri and Daniel Pullella, "An overview of speaker identification: Accuracy and robustness issues," *Circuits and Systems Magazine, IEEE*, vol. 11, no. 2, pp. 23–61, 2011.

[6] Xiaojia Zhao, Yang Shao, and DeLiang Wang, "Casa-based robust speaker identification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 5, pp. 1608–1616, 2012.

[7] Abdel-rahman Mohamed, Geoffrey Hinton, and Gerald Penn, "Understanding how deep belief networks perform acoustic modelling," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4273–4276.

[8] Bhiksha Raj, Michael L Seltzer, and Richard M Stern, "Reconstruction of missing features for robust speech recognition," *Speech Communication*, vol. 43, no. 4, pp. 275–296, 2004.

[9] Tobias May, Steven van de Par, and Armin Kohlrausch, "Noise-robust speaker recognition combining missing data techniques and universal background modeling," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 108–121, 2012.

[10] Geoffrey Hinton, "A practical guide to training restricted boltzmann machines," *Momentum*, vol. 9, no. 1, 2010.

[11] Alvin Martin and M Przybocki, "The nist year 2002 speaker recognition evaluation plan," 2001, [ONLINE]. Available: http://www.itl.nist.gov/iad/mig/tests/sre/2002/2002-spkrec-evalplan-v60.pdf.

[12] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1, pp. 19–41, 2000.