Text-independent Speaker Verification Using Convolutional Deep Belief Network and Gaussian Mixture Model

Ivan Rakhmanenko, Roman Meshcheryakov
Department of Information Systems Security
Tomsk State University of Control Systems and Radioelectronics
Tomsk, Russia
ria@keva.tusur.ru; mrv@tusur.ru

***Abstract*— There has been much interest in new deep learning approaches for representing and extracting high-level features for audio processing. In this paper convolutional deep belief network was used to generate new speech features for text-independent speaker verification. Structure and parameters of a convolutional deep belief network were described. New high-level speech features were extracted using proposed method. Relevance of speaker verification systems for mobile authentication was considered. Gaussian mixture model and universal background model speaker verification system used for experiments was described. Speaker verification accuracy using extracted features was evaluated on a 50 speaker set and a result is presented. Different layers and combinations of layers of convolutional deep belief network were used as a features for a text-independent speaker verification. High level features extracted by convolutional deep belief network were illustrated and analyzed. Reasons of insufficient verification accuracy were described. High-level features extracted by the third layer could be used for gender recognition.**

Keywords—speaker verification; speech features; gmm-ubm system; speech processing; cdbn; feature extraction; deep learning; neural networks

I. Introduction

There is an active conversion of practical methods that are used in user authentication systems from classical password-based methods to methods that are based on human biometrics. The voice, unlike the retina or fingerprints is considered to be less reliable for person identification or verification. However, in some cases speaker verification by voice is required. Particular attention should be devoted to speaker verification via mobile devices, as they have the microphone and computing capabilities that are necessary for speaker verification. Significantly, speaker verification on mobile devices could be combined with other verification methods, which allows taking advantage of multifactor authentication.

The application field of currently developed voice authentication systems includes multi-factor (biometric) authentication and access restriction systems, banking account management systems using voice biometrics in order to give speaker access to his banking account, national security and anti-terrorism issues. The use of speaker recognition systems that have even small possibility of mistake in such a sensitive application areas could be very dangerous.

Equal error rate value (EER) is one of the most common speaker verification accuracy measures used nowadays. EER is used both for text-dependent and text-independent automatic voice authentication systems. By now the best speaker recognition systems are characterized by 3-5% EER values. This accuracy is insufficient for modern speaker verification systems because even small probability of false acceptance is critical. If there are many speakers working with such systems, then mistakes will occur definitely, and such mistakes are unacceptable in systems granting access rights to confidential data or banking accounts.

Generally, low-level speech features are used for speaker verification, for example mel-frequency cepstral coefficients [1-4], linear prediction cepstral coefficients [5] and others. But attempts are made to use higher-level features, for example, extracting bottleneck features [6-8], constructing i-vectors based on a low-level representation [9-13], etc.

Based on how the brain processes incoming visual and audio signals, it can be assumed that the use of such features will improve the accuracy of speaker verification systems. In this paper a convolutional deep belief network (CDBN) was used to extract higher-level features and Gaussian mixture model with universal background model (GMM-UBM) system was used for speaker verification.

II. Speaker Verification Model

A. Gaussian Mixture Model

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities [1]. A GMM with M component Gaussian densities can be presented by the equation

$$p(x \mid \lambda) = \Sigma\{w_i, \mu_i, \Sigma_i\} \qquad (1)$$

where x is a D-dimensional continuous-valued data vector (i.e. measurement or features), w_i, $i = 1,...,M$, are the mixture weights, and $g(x|\mu_i, \Sigma_i)$, $i = 1,...,M$, are the component Gaussian

This work was supported by the Ministry of Education and Science of the Russian Federation within 1.3 federal program «Research and development in priority areas of scientific-technological complex of Russia for 2014-2020» (grant agreement № 14.577.21.0172 on October 27, 2015; identifier RFMEFI57715X0172).

densities with mean vector μ_i and covariance matrix Σ_i. The complete GMM is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. Each speaker is represented by his Gaussian mixture λ for speaker identification task. Gaussian mixture could be represented by the equation,

$$\lambda = \{w_i, \mu_i, \Sigma_i\} \tag{2}$$

There are two reasons for using Gaussian mixture densities as a representation of speaker identity [1]. The first reason is the intuitive notion that the individual component densities of the GMM may model some underlying set of acoustic classes, reflecting some general speaker-dependent vocal tract configurations. The second reason is the empirical observation that a linear combination of Gaussian basis functions is capable of representing a large class of sample distributions. A GMM can form smooth approximations to arbitrarily-shaped densities.

B. Universal Background Model

Universal Background Model (UBM) is a GMM trained on large set of speech samples that was taken from big population of speakers expected during recognition. Given the data to train a UBM, there are many approaches that can be used to obtain the final model. The simplest is to pool all the data to train the UBM via the EM algorithm. One should be careful that the pooled data are balanced over the subpopulations within the data. For example, in using gender-independent data, one should be sure there is a balance of male and female speech. Otherwise, the final model will be biased toward the dominant subpopulation. The same argument can be made for other subpopulations such as speech from different microphones. Another approach is to train individual UBMs over the subpopulations in the data, such as one for male and one for female speech, and then pool the subpopulation models together [1].

In this paper, parameters for the UBM are trained using the EM algorithm, and a form of Bayesian adaptation is used for training speaker models. Number of mixtures used is 256, as EER is not decreasing for small speaker set when larger mixture numbers are used. Speaker models are derived by MAP adaptation, where only means are adapted with relevance factor r = 10. GMM-UBM system described in this section is based on MSR Identity Toolbox.

III. Convolutional Deep Belief Network

A. Convolutional Deep Belief Network

The main difference between the convolutional deep belief network [14, 15] and the usual deep belief network [16] is the use of a convolutional restricted Boltzmann machine (CRBM) [14] as a hidden network layer. The CRBM is similar to the RBM (restricted Boltzmann machine), but the weights between the hidden and visible layers are shared among all locations in the hidden layer. CRBM (Fig. 1) is a feature detector consisting of three layers - the visible layer V, the detection layer H and the pooling layer P. The visible units in case of audio processing are real-valued, and the hidden units are binary-valued.

B. CDBN Structure

The input layer is consisting of an $N_V \times Ch$ dimensional array of real-valued units, where N_V is the number of windows to which the audio signal is divided, and Ch is the number of channels of the spectrum. To construct the hidden layer, consider K $N_W \times Ch$ dimensional filter weights W^K (also referred to as “bases”). The hidden layer consists of K groups of $N_H \times Ch$ dimensional arrays (where $N_H = N_V\text{-}N_W\text{+}1$) with units in group k sharing the weights W^k. There is also a shared bias b_k for each group and a shared bias c for the visible units. The energy function of the CRBM (3) can then be defined as [15]:

$$\mathrm{E(v,h)} = \frac{1}{2}\sum_{i=1}^{N_V} v_i^2 - \sum_{k=1}^{K}\sum_{j=1}^{N_H}\sum_{r=1}^{N_W} h_j^k\, W_r^k\, v_{j+r-1} - \\ -\sum_{k=1}^{K} b_k \sum_{j=1}^{N_H} h_j^k - c\sum_{i=1}^{N_V} v_i \tag{3}$$

The detection and pooling layers both have K groups of units, and each group of the pooling layer has N_P x N_P binary units. For each $k \in \{1,\ldots,K\}$, the pooling layer P^k shrinks the representation of the detection layer H^k by a factor of C along each dimension, where C is a small integer such as 2 or 3.

The joint and conditional probability distributions are defined as follows (4-6):

$$\mathrm{P(v,h)} = \frac{1}{Z}\exp(-E(v,h)) \tag{4}$$

$$\mathrm{P}(h_j^k = 1|v) = sigmoid(\,(\widetilde{W}_j^k *_v v)_j + b_k) \tag{5}$$

$$\mathrm{P}(v_i|h) = Normal(\textstyle\sum_k(\widetilde{W^k} *_f h^k)_i + c, 1), \tag{6}$$

where *v is a “valid” convolution (5), *f is a “full” convolution (6) [15]. For m-dimensional feature vector and n-dimensional vector “valid” convolution should result in an (m-n+1)-dimensional vector and a “full” convolution should result in an (m+n-1)-dimensional vector.

Since all units in one layer are conditionally independent given the other layer, inference in the network can be efficiently performed using block Gibbs sampling [15].

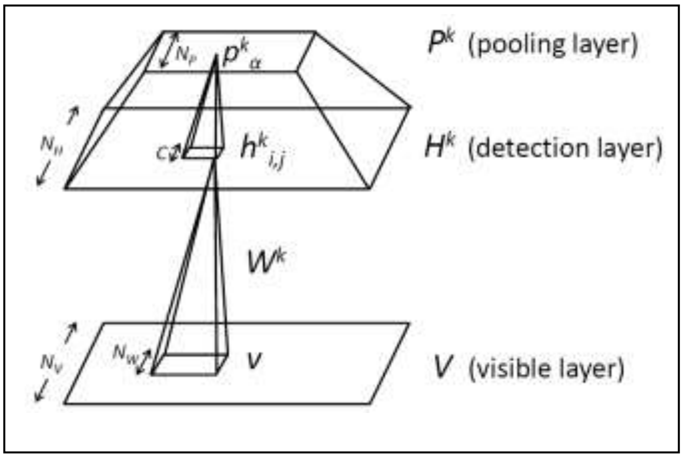

Fig. 1. Convolutional Restricted Boltzmann Machine with probabilistic max-pooling

C. *Layer-wise Training*

The use of an additional pooling layer makes it possible to reduce the amount and detail of the data supplied to the next hidden layer, which makes it possible to extract higher-level features in the data. This also reduces the computational load on following layers and filters out random noise.

A convolutional deep belief network is a composition of simple convolutional restricted Boltzmann machines. This fact allows the hidden layer of each CRBM to serve as a visible layer for the next CRBM. Thereby, a quick layer-wise training technique could be performed to train a CDBN. In order to estimate the gradient, the contrastive divergence approximation is applied to each sub-network, beginning with the first pair of layers. The data from the training set is fed to the visible layer of the CRBM, the following hidden layers take their input from the output of the next CRBM's hidden layers.

IV. Experimental Evaluation

A. *Speech Corpus*

The experiments were conducted using speech database containing collection of speech from 25 male and 25 female speakers. This speech database includes speech samples of sentences from science fiction stories. The total length of speech for each speaker is at least 6 minutes consisting of 50 speech segments of various lengths. Each speaker was recorded using medium-quality microphone, 8000 Hz sampling rate, 16 Bit sample size.

All 50 speaker set was divided equally for male and female speakers on the UBM training set consisting of 30 speakers and speakers' training set consisting of 20 speakers. For MAP adaptation of speakers' models 40 speech segments was taken. Remaining 10 utterances of each speaker was used for testing verification system. Overall, 4000 tests were done for each feature set, having 10 positive (true speaker) and 190 negative (imposter) tests for each speaker.

B. *Experimental Evaluation*

After training phase, that consists of UBM training and speakers' models adapting, starts test phase. For each test speech segment verification scores (log-likelihood ratios) are calculated using speaker GMM and UBM models (7). Using different decision thresholds hypothesized speaker model was accepted or rejected.

$$\Lambda(X) = \log p(X|\lambda_{hyp}) - \log p(X|\lambda_{ubm}) \tag{7}$$

Two different verification metrics was used for evaluating speaker verification system: EER and minimum detection cost function with SRE 2008 parameters (minDCF).

In order to test speaker verification accuracy using CDBN, network structure and parameters should be specified. In this paper CDBN consisting of three connected layers was used for experimental evaluation. The first and second layers consist of 300 bases, the third of 60. The input layer consists of 80 neurons (Ch = 80). Spectrogram of the speech is extracted from the audio, PCA whitening is applied. Lower dimension spectrogram is fed to the input layer. The data given to the visible layer is selected by 20 ms windows with a 10 ms offset. For each base in the hidden layers, the filter dimension N_W = 6 and the convolution factor C = 3 was used. Parameters for the first and second layers of the network were taken from [15]. Parameters for the third layer were selected by the authors independently. Using presented parameters CDBN for audio processing was trained.

As a result of training CDBN, three trained layers of the network were obtained, the outputs of each of which could be used as a features for GMM-UBM speaker verification system. To assess the verification system accuracy using obtained features, the CDBN layers outputs were fed to the Gaussian mixture model, which was used as a classifier. Also, for features of each layer separate UBM was trained.

To test and compare speaker verification system accuracy, GMM-UBM speaker verification system with parameters and features from [17] was used. This features includes 14 mel-frequency cepstral coefficients (MFCC) and features obtained using greedy Add-del algorithm including 13 MFCC, 10 delta MFCC, 2 double delta MFCC, voicing probability, 1 linear prediction coefficient (LPC) and 1 line spectral pair (LSP).

Results of the experimental evaluation are given in Table I. Based on the results, it can be concluded that none of the feature sets extracted by the CDBN gives an accuracy of speaker verification system more than a standard feature set consisting of 14 MFCC. A feature set obtained by the greedy add-del algorithm shows the best verification accuracy.

In order to use information of different levels, combinations of GMM classifiers using separate CDBN layers were used. Combinations of different feature level classifiers did not increase the verification accuracy, compared to the classifiers using separate feature levels.

TABLE I. Test Accuracy For Speaker Verification Using Different Features

№	**Feature Set Evaluation**		
	Feature Set	***% EER***	***minDCF* 100***
1	CDBN L1	2,00	0,997
2	CDBN L2	3,50	1,740
3	CDBN L3	10,00	5,765
4	CDBN L1 + CDBN L2	2,00	1,197
5	CDBN L1 + CDBN L3	2,00	1,121
6	CDBN L2 + CDBN L3	3,29	1,926
7	CDBN L1 + CDBN L2 + CDBN L3	2,00	1,327
8	MFCC	1,00	0,925
9	Greedy Add-del	0,58	0,623

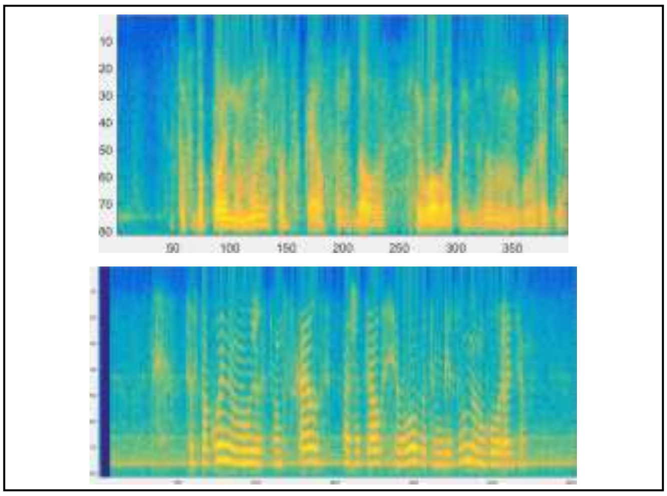

Fig. 2. Spectrogram of the same phrase for male (top) and female (bottom) speaker

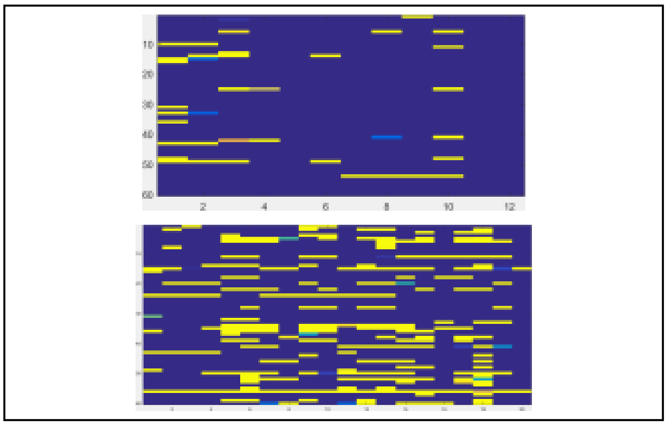

Fig. 3. Third layer neurons activations of the same phrase for male (top) and female (bottom) speaker

C. Discussion

There could be different reasons for low accuracy of speaker verification system using CDBN features. Deep learning methods work better on a large dataset used for training, so a small amount of training speech samples is one of the possible reasons for the low accuracy of the verification system. Another reason could be the GMM as a classifier, as it could not provide an opportunity to show better results.

Nevertheless, attention should be given to visual representation of a speech signal and CDBN neurons activations. Fig. 2 shows spectrogram of a same phrase for a male and female speaker. There could be seen a significant difference between male and female speaker saying the same phrase on a third layer of CDBN (Fig. 3). This fact could be used for gender recognition, using CDBN outputs as a features.

V. Conclusion

In this paper convolutional deep belief network was used to generate new speech features for text-independent speaker verification. New high-level speech features were extracted using proposed method. Speaker verification system based on GMM-UBM speaker verification system was used to assess speaker verification accuracy using extracted CDBN features. None of the feature sets extracted by the CDBN gives an accuracy of speaker verification system more than a standard feature set consisting of 14 MFCC. Nevertheless, speaker verification methods using presented features could be combined with methods using different speech features to obtain better verification accuracy.

References

[1] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, “Speaker verification using adapted Gaussian mixture models,” Digit. Signal Process., vol. 10, no. 1-3, pp. 19-41.

[2] K.S. Ahmad, A.S. Thosar, J.H. Nirmal, and V.S. Pande, “A unique approach in text independent speaker recognition using MFCC feature sets and probabilistic neural network,” Adv. in Pattern Recog. (ICAPR), pp. 1-6.

[3] A. Jain, and O.P. Sharma, “Evaluation of MFCC for speaker verification on various windows,” Recent Adv. and Innov. in Engr. (ICRAIE), 2014, pp. 1-6.

[4] M.J. Alam, T. Kinnunen, P. Kenny, P. Ouellet, and D. O’Shaughnessy, “Multitaper MFCC and PLP features for speaker verification using i-vectors,” Speech Comm., 2013, vol. 55, no. 2, pp. 237-251

[5] Y. Kawakami, L. Wang, A. Kai, and S. Nakagawa, “Speaker Identification by Combining Various Vocal Tract and Vocal Source Features,” Int. Conf. on Text, Speech, and Dialogue, 2014, pp. 382-389

[6] M. McLaren, L. Ferrer, and A. Lawson, “Exploring the role of phonetic bottleneck features for speaker and language recognition,” Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5575-5579.

[7] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, “A novel scheme for speaker recognition using a phonetically-aware deep neural network,” Acoustics, Speech and Sig. Process. (ICASSP), 2014, pp. 1695-1699.

[8] F. Richardson, D. Reynolds, and N. Dehak, “Deep neural network approaches to speaker and language recognition,” IEEE Sig. Process. Let., 2015, vol. 22, no. 10, pp. 1671-1675.

[9] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet and J. Alam, “Deep neural networks for extracting baum-welch statistics for speaker recognition,” Proc. Odyssey, 2014, pp. 293-298.

[10] T. Stafylakis, P. Kenny, V. Gupta, J. Alam, and M. Kockmann, “Compensation for phonetic nuisance variability in speaker recognition using DNNs,” Odyssey: The Speaker and Lang. Recognition Workshop, 2016, pp. 340-345.

[11] O. Kudashev, S. Novoselov, T. Pekhovsky, K. Simonchik, and G. Lavrentyeva, “Usage of DNN in speaker recognition: advantages and problems,” Int. Symp. on Neural Networks, 2016, pp. 82-91.

[12] D. Snyder, D. Garcia-Romero, and D. Povey, “Time delay deep neural network-based universal background models for speaker recognition,” Aut. Speech Recognition and Understanding (ASRU), 2015, pp. 92-97.

[13] O. Ghahabi, and J. Hernando, “Deep belief networks for i-vector based speaker recognition,” Acoustics, Speech and Sig. Process. (ICASSP), 2014, pp. 1700-1704.

[14] H. Lee, R. Grosse, R. Ranganath, and A.Y. Ng, “Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations,” Proc. of the 26th Annual Int. Conf. on Machine Learning, 2009, pp. 609-616.

[15] H. Lee, P. Pham, Y. Largman, and A.Y. Ng, “Unsupervised feature learning for audio classification using convolutional deep belief networks,” Adv, in Neural Info. Process. Systems, 2009, pp. 1096-1104.

[16] G.E. Hinton, S. Osindero, and Y.W. The, “A fast learning algorithm for deep belief nets,” Neural Computation, 2006, vol. 18, no. 7, pp. 1527-1554.

[17] I.A. Rakhmanenko, and R.V. Meshcheryakov, “Identification features analysis in speech data using GMM-UBM speaker verification system,” SPIIRAS Proc., 2017, vol. 52, pp. 32-50.