INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Tandem Features for Text-dependent Speaker Verification on the RedDots Corpus

Md Jahangir Alam, Patrick Kenny, Vishwa Gupta

Computer Research Institute of Montreal (CRIM), Montreal, Canada
{jahangir.alam, patrick.kenny, vishwa.gupta}@crim.ca

Abstract

We use tandem features and a fusion of four systems for text-dependent speaker verification on the RedDots corpus. In the tandem system, a senone-discriminant neural network provides a low-dimensional bottleneck feature at each frame which are concatenated with a standard Mel-frequency cepstral coefficients (MFCC) feature representation. The concatenated features are propagated to a conventional GMM/UBM speaker recognition framework. In order to capture complementary information to the MFCC, we also use linear frequency cepstral coefficients and wavelet-based cepstral coefficients features for score level fusion. We report results on the part 1 and part 4 (text-dependent) tasks of RedDots corpus. Both the tandem feature-based system and fused system provided significant improvements over the baseline GMM/UBM system in terms of equal error rates (EER) and detection cost functions (DCFs) as defined in the 2008 and 2010 NIST speaker recognition evaluations. On the part 1 task (impostor correct condition) the fused system reduced the EER from 2.63% to 2.28% for male trials and from 7.01% to 3.48% for female trials. On the part4 task (impostor correct condition) the fused system helped to reduce the EER from 2.49% to 1.96% and from 5.9% to 3.22% for male and female trials respectively.

Index Terms: speaker verification, neural networks, GMM-UBM, tandem feature, RedDots

1. Introduction

In a speaker verification system the goal is to verify the claimed identity of an unknown speaker. Depending on the constraints imposed on the enrolment and test recordings it can be categorized as text-dependent or text-independent. Text-dependent speaker verification has gained a lot of research interest over the last few years in the speaker recognition community. In text-dependent speaker verification the classes to be recognized are speaker-lexical contents combinations rather than speakers as in text-independent speaker verification [4]. The lexical contents can be common for all speakers or unique. Because of the prior knowledge of the constrained phonetic contents text-dependent systems can provide better recognition accuracy for relatively short duration (1-3 seconds) enrolment and test recordings.

Early work on text-dependent speaker verification usually employed the dynamic time warping (DTW) method, which performs the speaker verification task by aligning enrolment and test feature sequences of different lengths and then applying temporal template matching [5]. Some recent work on text-dependent speaker verification [2-4, 6-10] uses GMM/UBM [2-4, 6-11], Hierarchical multi-Layer Acoustic Model (HiLAM) [11], Hidden Markov Model [7], Joint Factor Analysis (JFA) [4, 6-9] models on various text-dependent corpora (such as RSR2015 [11], CSLU, and a proprietary dataset collected at Concordia University [7]). As for the UBM/i-vector/PLDA system, direct application to the text-dependent condition is not as satisfactory as the text-independent case unless a large amount of background data is available [2, 12-14].

The RedDots challenge is an initiative of the RedDots project which was rolled out on 29 January 2016. The purpose of this challenge is to stimulate research efforts for text-dependent speaker verification over four parts (part 1- part 4) of RedDots corpus [1].

In this paper, we employ tandem features in the GMM/UBM framework [20] to improve the performance of text-dependent speaker verification on RedDots challenge corpus. Tandem features are obtained by concatenating deep bottleneck features, supplied by a senone-discriminant deep neural network, with Mel-frequency cepstral coefficients (MFCC) feature and then reducing the feature dimension using principal component analysis. In order to capture complementary information for score level fusion we also develop three systems based on MFCC (baseline system), linear frequency cepstral coefficients (LFCC) and wavelet-based cepstral coefficients (WCC) features. Although GMM/UBM system does not take into account lexical information in the same way as a HMM system such as HiLAM [11], it still shows a promising performance in text-dependent speaker verification.

2. Extraction of tandem features

Fig. 1 presents a schematic diagram for extracting tandem features from the deep neural network (DNN) bottleneck and MFCC features. A DNN can be used as a means of extracting features, known as bottleneck features, for used by a classifier. Bottleneck features are extracted by placing a hidden layer, which has relatively small number of nodes compared to the size of other layers, in between the input and output layers [21-22]. In speech related applications these features are widely being employed for improving recognition accuracy [21-25]. In order to extract bottleneck features from the RedDots [1] and LibriSpeech [18] corpora we trained a senone discriminant DNN using 1141 hours of audio recordings from the Fisher and Switchboard corpora. As there is no background data in the RedDots challenge

Copyright © 2016 ISCA

http://dx.doi.org/10.21437/Interspeech.2016-1465

corpus we use LibriSpeech (11126 recordings from 146 speakers) as our background data to train the UBM. The inputs to the DNN are TRAP (TempoRAl Pattern) [26] feature parameters computed from 31 frames (15 frames from each side of the current frame) of the log filterbank features. The DNN has 4 hidden layers (with sigmoid activation function) of size 1000 nodes each and a linear bottleneck layer of size 64 (2nd to last hidden layer). The output layer, which is the classification layer, is a softmax of dimension 3925 i.e., the output layer computes posteriors for 3925 senones. The models were trained by cross-entropy criterion followed by 2 iterations of MMI (or sequence) training.

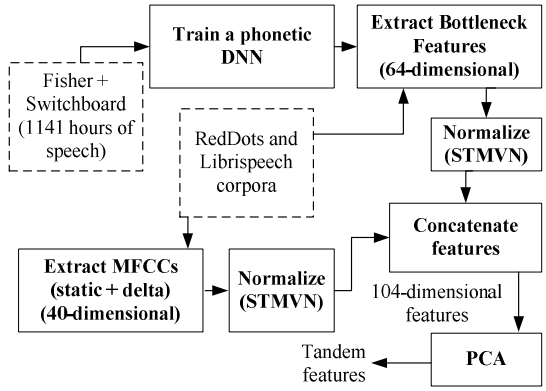

Figure 1: Extraction of tandem features.

Once the senone discriminant DNN is trained the bottleneck features for all RedDots and background data (LibriSpeech) are extracted. 40-dimensional MFCCs (20 static including the log energy + 20 deltas) are also extracted for the above mentioned corpora. Features are normalized using short-time mean and variance normalization with a window of 151 frames (75 frames from each side of the current frame). Tandem features are then obtained by concatenating this high level (i.e, bottleneck) and low level spectral features (i.e., MFCC) and reducing the dimension by applying principal component analysis (PCA). The tandem features are propagated to a standard GMM/UBM speaker verification framework.

3. Baseline system

Our baseline system for text-dependent speaker verification task is based on MFCC (Mel-frequency cepstral coefficients) features and uses the conventional GMM/UBM speaker recognition framework. MFCC is the most widely used form of feature extraction for speech and speaker recognition. MFCC processing begins with pre-emphasis, typically using a 1st order high-pass filter. Short-time Fourier Transform analysis is performed using a hamming window, and triangular shape Mel-frequency integration is performed for auditory spectral analysis. The logarithmic nonlinearity stage follows, and the final features are obtained through the use of a Discrete Cosine Transform (DCT). Static feature dimension is 20. The log energy computed from the raw speech signal is included instead of 0-th cepstrum. Final 60-dimensional MFCCs are obtained by appending delta and double delta coefficients and normalizing the features with a STMVN method.

4. Complementary systems for fusion

Score level fusion of multiple systems that carry complementary information to each other is widely used in speaker recognition systems. In order to capture complementary information to the MFCC, we also develop linear frequency cepstral coefficients (LFCC) and wavelet-based cepstral coefficients (WCC) features-based systems to perform score level fusion.

The LFCC features extraction follows the same processing steps as the MFCC. The only difference is that in LFCC linear frequency filterbank are used in place of Mel-frequency filterbank. It has been found in [17] that the LFCC features help to reduce error rates when fused with MFCC in the score level.

4.1. Wavelet-based cepstral coefficients

The wavelet transform is a transformation that provides time-frequency representation of a signal. In this work Daubechies' 3 tap filters are used to compute 6 level wavelet transform [15]. For a given wavelet tree the wavelet packet transform (WPT) is computed. This yields a sequence of subband signals (or WPT coefficients) at the leaves of the tree. Let $x(c,n)$ represent the c-th subband signal, where c is subband index and n is the sample or WPT coefficient index. Standard NEO of $x(c,n)$ can be expressed as a special case of the following k-th order ($k=0,1,2,...$) and l-th lag ($l=1,2,3,...$) generalized discrete energy operator:

$$\Psi_{k,l}\left(x(c,n)\right) = x(c,n)x(c,n+k)... - x(c,n-l)x(c,n+k+l). \quad (1)$$

For k=0 and l=1, eqn. (1) reduces to the standard NEO. Since NEO is an energy operator and energy is a positive quantity, in order to avoid any negative values in eqn. (1) (if $x(c,n)x(c,n+k) < x(c,n-l)x(c,n+k+l)$ for k=0, l=1) we have taken the absolute values of eqn. (1)

The nonlinear energy operator (NEO) is applied to the WPT coefficients and the output of NEO is averaged to get the subband energy. The NEO approach has many attractive features such as simplicity, efficiency, and adaptability to instantaneous signal variations [16].

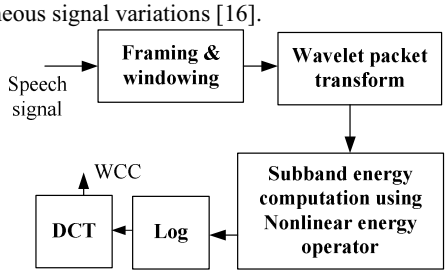

Figure 2: Wavelet-based cepstral coefficients (WCC) extraction.

The logarithmic nonlinearity stage follows, and the final 13-dimensional static WCC (including the log energy) features are obtained through the use of a Discrete Cosine Transform (DCT). Delta and double delta coefficients are appended making 39-dimensional WCC features. Short time mean and variance normalization is applied to normalize the features. Fig. 2 shows overall processing steps for WCC features.

5. Experiments

A Gaussian Mixture Model - Universal Background Model (GMM/UBM) framework [20] was used as the backend for the text-dependent speaker verification tasks. Experimental results are reported on the part 1 and part 4 (text-dependent) of the RedDots corpus [1] under three types of nontarget trials as depicted in table 1. Based on tandem, MFCC, LFCC and WCC features four sub-systems denoted as Tandem system, MFCC, LFCC, and WCC, respectively, were developed. A fused system was also developed by fusing the scores of these sub-systems. Performance evaluation metrics used in this work are: the Equal Error Rate (EER), the old detection cost function (DCF08) and the new detection cost function (DCF10) as defined in NIST speaker recognition evaluation of 2008 and 2010, respectively.

Table 1. *Definition of target trials and three types of nontarget trials: target correct (TC), target wrong (TW), impostor correct (IC), impostor wrong (IW)*

	Target trials
TC	Target speaker pronouncing the correct lexical contents
	Three types of Nontarget trials
TW	Target speaker pronouncing the wrong lexical contents
IC	Impostor pronouncing the correct lexical contents
IW	Impostor pronouncing the wrong lexical contents

5.1. RedDots challenge corpus

The RedDots corpus [1, 19] contains audio recordings from 62 speakers including 49 male speakers and 13 female speakers from 21 countries. In the current release of this corpus there are 473 male and 99 female sessions. The number of recordings per session is 24 and the average duration of the recordings is 3 seconds. This corpus is comprised of four parts: part 1 consists of 10 sentences common to all speakers, part 2 contains 10 sentences unique to each speaker, part 3 contains 2 free choice recordings and part 4 contains free text sentences that are unique across sessions. For part 4 enrollment can be text dependent or text prompted. In this work we conducted experiments on part 1 and part 4 (text dependent) tasks only. Since no development (or background) data was released with the challenge corpus we used LibriSpeech as background data by taking 11126 recordings from 146 speakers (female: 73, male: 73).

5.2. Experimental setup

A 512-component gender-independent UBM with diagonal covariances was trained on all the background data. During enrolment a target speaker model is trained by MAP (maximum a posteriori) adaptation of the parameters of the UBM using the target speaker's enrolment utterances. For UBM adaptation we used single iteration with a relevance factor of 2. Each trained speaker model is scored against all the test utterances. No score normalization was used.

5.3. Results and discussion

In this work we reported text-dependent speaker verification results both on male and female trials of the part 01 and part 04 (text-dependent (td)) of the RedDots challenge corpus. Nontarget trials are of three types: impostor correct (IC), impostor wrong (IW) and target wrong (TW) as defined in table 1. These acronyms are used in reporting the results in tables 2-5.

Tables 2 & 3 present verification results on part 01 task under these three types of nontarget trials for the male and female trials, respectively. On the other hand tables 4 & 5 present verification results on part 04 (td) tasks for the male and female trials, respectively under IC, TW and IW nontarget types. It is observed from tables 2-5 that on both tasks the fused system and tandem system yielded significant reduction in error rates over the baseline in all three evaluation metrics under IC, TW, and IW nontarget trials conditions. Because of the mismatched lexical contents tandem system (and hence fused system) is more discriminative in TW (i.e., target speaker pronouncing the wrong pass-phrase) and IW (i.e., nontarget speaker pronouncing the wrong pass-phrase) conditions than in IC (i.e., nontarget speaker uttering the correct pass-phrase) condition. In the TW and IW conditions the tandem and fused systems achieved EER below 1%. In the IC nontarget type the lowest EER (2.28%) was obtained the fused system for part 1 task - male trials, and for female trials EER obtained by this system is 3.48%.

Although the performance of WCC system is worse than the other systems (including the baseline) it seems to carry complementary information to other systems. When fused with other three systems WCC feature helped to reduce the error rates, except in the case of IW where EER of fused system (without WCC) is slightly lower than the fused system. In order to find out how much the WCC system is contributing in the fused system we developed another fused system by excluding WCC system's scores. By comparing the results from the last rows of table 5 it is evident that wavelet cepstral coefficients feature carries information which is complementary to other systems and therefore, helped to reduce error rates when fused with other systems.

6. Conclusions

In this work we presented tandem features and a fusion of four systems for reducing text-dependent speaker verification error rates on the tasks of RedDots corpus. The tandem features were formed by the concatenation of a bottleneck feature, provided by a senone-discriminant deep neural network, with a standard Mel-frequency cepstral coefficients (MFCC) feature and reducing the feature dimension with the help of PCA algorithm. The tandem features were propagated to a conventional GMM/UBM framework for text-dependent speaker verification. To get benefited from score level fusion we also developed MFCC, linear frequency cepstral coefficients and wavelet cepstral coefficients features-based systems. Reported results on the part 1 and part 4 (text-dependent) tasks of RedDots corpus showed that both the tandem feature-based system and fused system provided significant reduction in error rates over traditional baseline system in terms of EER (%) and DCF08 and DCF10 metrics in impostor correct, impostor wrong and target wrong nontarget trials. On the part 1 task (impostor correct condition) the fused system reduced the EER from 2.63% to 2.28% for male trials and from 7.01% to 3.48% for female trials. On the part4 task (impostor correct condition) the fused system helped to reduce the EER from 2.49% to 1.96% and from 5.9% to 3.22% for male and female trials respectively.

In target wrong and impostor wrong conditions the tandem and fused systems achieved EER below 1%. Based on above observations it can be concluded that tandem feature is very promising for speaker recognition with short duration utterances.

Table 2. *Speaker verification results on male trials of part 01 task of the RedDots corpus in terms of EER (%), DCF08, and DCF10 in three non target trials types.*

EER (%)/DCF08/DCF10			
Task: Part 01 / Trials: Male			
System	Non-target type		
	IC	TW	IW
MFCC	2.63/0.12/0.42	4.84/0.26/0.57	0.48/0.02/0.10
LFCC	3.32/0.15/0.48	4.96/0.28/0.65	0.77/0.03/0.15
WCC	11.7/0.5/0.99	12.0/0.7/0.99	4.28/0.2/0.9
Tandem system	3.81/0.17/0.55	0.83/0.03/0.07	0.42/0.006/0.026
Fused system	**2.28/0.1/0.34**	**0.56/0.02/0.03**	**0.15/0.003/0.01**

Table 3. *Speaker verification results on female trials of part 01 task of the RedDots corpus in terms of EER (%), DCF08, and DCF10 in three non* target trials types.

EER (%)/DCF08/DCF10			
Task: Part 01 / Trials: Female			
System	Non-target type		
	IC	TW	IW
MFCC	7.01/0.28/0.39	7.10/0.29/0.52	2.32/0.09/0.23
LFCC	6.12/0.23/0.5	6.16/0.31/0.63	2.42/0.08/0.22
WCC	18.0/0.69/0.87	13.0/0.66/0.97	6.5/0.36/0.89
Tandem system	5.32/0.25/0.46	0.47/0.019/0.019	**0.23**/0.012/0.06
Fused system	**3.48/0.19/0.34**	**0.38/0.01/0.01**	0.41/**0.009/0.02**

Table 4. *Speaker verification results on male trials of part 04 task (td) of the RedDots corpus in terms of EER (%), DCF08, and DCF10 in three non* target trials types.

EER (%)/DCF08/DCF10			
Task: Part 04 (td) / Trials: Male			
System	Non-target type		
	IC	TW	IW
MFCC	2.49/0.11/0.39	5.6/0.3/0.7	0.49/0.02/0.13
LFCC	3.2/0.14/0.45	5.97/0.37/0.79	0.72/0.03/0.20
WCC	12.5/0.51/0.98	14.7/0.96/0.99	4.56/0.22/0.98
Tandem system	2.95/0.14/0.49	0.81/0.03/0.15	0.25/0.004/0.02
Fused system	**1.96/0.09/0.28**	**0.54/0.02/0.10**	**0.15/0.003/0.01**

Table 5. *Speaker verification results on female trials of part 04 (td) tasks of the RedDots corpus in terms of EER (%), DCF08, and DCF10 in three non* target trials types.

EER (%)/DCF08/DCF10				
Task: Part 04 (td) / Trials: Female				
System		Non-target type		
		IC	TW	IW
1	MFCC	5.9/0.23/0.34	6.13/0.25/0.56	1.71/0.07/0.22
2	LFCC	5.4/0.19/0.45	5.5/0.27/0.66	2.11/0.07/0.25
3	WCC	17.12/0.67/0.89	13.9/0.69/0.97	6.2/0.32/0.93
4	Tandem system	4.38/0.2/0.39	**0.36**/0.02/0.07	**0.12**/0.006/0.046
5	Fused system	**3.22/0.15/0.3**	**0.36/0.016/0.04**	**0.20/0.006/0.03**
6	Fused system (without WCC)	3.62/0.16/0.32	0.39/0.017/0.043	0.12/0.006/0.042

7. References

[1] Kong Aik Lee, Anthony Larcher, Guangsen Wang, Patrick Kenny, Niko Brummer, David van Leeuwen, Hagai Aronowitz, Marcel Kockmann, Carlos Vaquero, Bin Ma, Haizhou Li, Themos Stafylakis, Jahangir Alam, Albert Swart and Javier Perez, The RedDots Data Collection for Speaker Recognition, *Proc. Interspeech, Dresden Germany, Sept. 2015.*

[2] A. Larcher, K.-A. Lee, B. Ma, and H. Li, "Phonetically constrained PLDA modeling for text-dependent speaker verification with multiple short utterances," in Proc. ICASSP , Vancouver, Canada, May 2013.

[3] ——, "The RSR2015 database for text-dependent speaker verification using multiple pass-phrases," in Proc. Interspeech, Portland OR, Sept. 2012.

[4] Kenny, P., Stafylakis, T., Alam, J., Ouellet, P., and Kockmann, M., "Joint Factor Analysis For Text-Dependent Speaker Verification," *Proc. Odyssey Speaker and Language Recognition Workshop, Joensuu, Finland, June 2014.*

[5] S. Furui, "Cepstral analysis technique for automatic speaker verification," IEEE Trans. Acoust. Speech Signal Process., vol. 29 (2), pp. 254–272, 1981.

[6] Patrick Kenny, Themos Stafylakis, Jahangir Alam and Marcel Kockmann, "An I-Vector Backend for Speaker Verification," *Proc. Interspeech, Dresden Germany, Sept. 2015.*

[7] ——, "JFA Modeling with Left-to-Right Structure and a New Backend for Text-Dependent Speaker Recognition," *Proc. ICASSP, Brisbane, Australia, April 2015.*

[8] Patrick Kenny, Themos Stafylakis, Md. Jahangir Alam, Pierre Ouellet and Marcel Kockmann, "In-Domain versus Out-of-Domain Training for Text-Dependent JFA," Proc. INTERSPEECH, Singapore, September 2014.

[9] Themos Stafylakis, Patrick Kenny, Jahangir Alam, and Marcel Kockmann, "Speaker and Channel Factors in Text-Dependent Speaker Recognition," *IEEE Trans. Audio, Speech and Language Proc.,* vol. 24(1), pp, 65-78, January 2016.

[10] Yuan Liu,Yanmin Qian, Nanxin Chen, Tianfan Fu, Ya Zhang, Kai Yu, "Deep feature for text-dependent speaker verification," Speech Comm., vol. 73, pp. 1-13, October 2015.

[11] A. Larcher, K.-A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," in Speech Comm., vol. 60, pp. 56-77, May 2014.

[12] E. Variani, X. Lei, E. McDermott, I. Lopez-Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in IEEE International Conference on Acoustics, Speech, and Signal Processing, Florence, Italy, May 2014.

[13] H. Aronowitz and O. Barkan, "On leveraging conversational data for building a text dependent speaker verification system," in Proc. Interspeech, Lyon, France, pp. 243–247, Sept. 2013.

[14] T. Stafylakis, P. Kenny, P. Ouellet, J. Perez, M. Kockmann, and P. Dumouchel, "I-Vector/PLDA Variants for Text-Dependent Speaker Recognition," Aug. 2013. [Online]. Available:http://www.crim.ca/perso/patrick.kenny

[15] R. Sarikaya, B. L. Pellom, and J. H. L. Hansen, "Wavelet packet transfrom features with application to speaker identification," in *Nordic Signal Processing Symposium*, 1998.

[16] Maragos, J. F. Kaiser, and T. F. Quatieri, On amplitude and frequency demodulation using energy operators, IEEE Trans. Signal Processing , vol. 41, no. 4, pp. 1532–1550, April 1993.

[17] Jahangir Alam, Patrick Kenny and Themos Stafylakis, Combining Amplitude and Phase-Based Features for Speaker Verification with Short Duration Utterances, *Proc. Interspeech, Dresden Germany, Sept. 2015*

[18] Vassil Panayotov, Guoguo Chen, Daniel Povey and Sanjeev Khudanpur, "LibriSpeech: an ASR corpus based on public domain audio books," proc. of ICASSP, 2015.

[19] The RedDots challenge: https://sites.google.com/site/thereddotsproject/reddots-challenge.

[20] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," Digital Signal Process., vol. 10, pp. 19–41, 2000.

[21] Yao Tian, Meng Cai, Liang He, Jia Liu, "Investigation of Bottleneck Features and Multilingual Deep Neural Networks for Speaker Verification," Proc. Interspeech 2015.

[22] F. Richardson, D. Reynolds, and N. Dehak, "Deep Neural Network Approaches to Speaker and Language Recognition," *IEEE Signal Processing Letters*, Vol. 22, No. 10, pp. 1671-1675, October 2015.

[23] G. Hinton et al., "Deep neural networks for acoustic modeling in speech recognition," IEEE Signal Process. Mag. , pp. 82–97, Nov. 2012.

[24] Y. Song, B. Jiang, Y. Bao, S. Wei, and L.-R. Dai, "I-vector representation based on bottleneck features for language identification," Electron. Lett. , pp. 1569–1580, 2013.

[25] P. Matejka, L. Zhang, T. Ng, H. S. Mallidi, O. Glembek, J. Ma, and B. Zhang, "Neural network bottleneck features for language identification," in Proc. IEEE Odyssey , pp. 299–304, 2014.

[26] F Grezl, TRAP-Based Probabilistic Features for Automatic Speech Recognition, Dept. of Computer Graphics & Multimedia, Brno University of Technology, Brno, Czech Republic, Doctoral Thesis, 2007.