Combining Source and System Information for Limited Data Speaker Verification

Rohan Kumar Das[1], *Abhiram B*[2], *S R M Prasanna*[1], *A G Ramakrishnan*[2]

[1]Department of Electronics and Electrical Engineering ,
Indian Institute of Technology Guwahati, Guwahati-781039, India
[2]Department of Electrical Engineering ,
Indian Institute of Science Bangalore, Bangalore-560012, India
rohankd@iitg.ernet.in, abhiram1989@gmail.com, prasanna@iitg.ernet.in, ramkiag@ee.iisc.ernet.in

Abstract

Speaker verification using limited data is always a challenge for practical implementation as an application. An analysis on speaker verification studies for an i-vector based method using Mel-Frequency Cepstral Coefficient (MFCC) feature shows that the performance drops drastically as the duration of test data is reduced. This decrease in performance is due to insufficient phonetic coverage when we capture only the vocal tract feature. However the same can be improved if some source characteristics are taken into consideration. This paper attempts to improve the speaker verification performance using source characteristics. A recently proposed characterization of the voice source signal called the discrete cosine transform of the integrated linear prediction residual (DCTILPR) has been found to be useful as a speaker-specific feature. Speaker verification is performed over short test utterances in the NIST 2003 database using both the DCTILPR and MFCC features, and their score-level combination is found to give a significant performance improvement over the system using only the MFCC features.

Index Terms: speaker verification, short utterances, source features, DCTILPR, MFCC

1. Introduction

Research on speaker verification (SV) has expanded significantly over the years since its inception. However, while it comes to deployment as an application, the amount of speech data plays a significant role. Existing SV systems require a minimum amount of speech data so that sufficient phonetic content is covered for robust modeling. In some applications, we may not get this required amount of data, leading to poor system performance.

The i-vector [1] system has demonstrated the state-of-the-art approach for NIST speaker recognition evaluation (SRE). Its compact representation, computational efficiency and easy channel/session compensation makes it a benchmark for the SV task. The significant improvement in performance, achieved through the i-vector based system over other conventional SV systems [1] shows the potential for using it for SV under limited data conditions. In [2], an analysis of i-vector based SV system for short utterances is shown for different durations of train as well as test segments. From a practical system point of view, we consider sufficient training data and limited test data conditions. The analysis given in [2] for very less amount of test data ($<$10 s) shows that the performance drops significantly even though sufficient speech data is used during training. This trend of downfall in performance for limited test data motivates us to consider a speaker-specific feature which captures the speaker information with limited data. The literature shows that, though the voice source features are not as discriminative as vocal tract (or system) features, the fusion of the two can improve the accuracy [3, 4]. Also, studies of [5, 6] suggests that the amount of train/test data for the voice source features can be less compared to the amount of amount of data required in case vocal tract features.This is due to the fact that the voice source features do not depend much on phonetic content, where as the the robustness of vocal tract feature depends on the amount of phonetic content that it captures for a particular utterance. This shows the significance of using voice source features for limited data SV along with the conventional vocal tract features.

This paper concentrates on considering a source feature along with a vocal tract feature for improving the SV system performance for limited data test conditions. The studies in [7] have shown that the source feature discrete cosine transform of the integrated linear prediction residual (DCTILPR) captures relevant speaker information and gives significant performance in the case of speaker identification over standard NIST dataset. When it comes to limited data SV, using this source feature can certainly help to improve the system performance as it does not require sufficient phonetic coverage for robust modeling, which motivated us to consider the same. The performance evaluation is reported over NIST 2003 SRE database [8] for the state-of-the-art i-vector based SV system. Linear discriminant analysis (LDA) and within class covariance normalization (WCCN) [1] are applied as channel/session compensation techniques. Two parallel systems are developed using both DCTILPR and Mel-Frequency Cepstral Coefficient (MFCC) features for the stated i-vector based SV system. The SV system using only the MFCC features is considered as the baseline. The fusion of the above mentioned features is done at the score level, and it is found to give a significant improvement over the baseline results obtained for short utterance cases.

The rest of the paper is organized as follows. Section 2 describes the development of the i-vector based SV followed by channel/session compensation techniques for robust speaker modeling in case of sufficient data conditions. Section 3 provides the details of recently proposed DCTILPR feature used for SV and its significance for short utterances. In section 4, the SV experiments performed using MFCC and DCTILPR features and the combination of the two at the score level for short utterances are explained and their results are reported. Finally, a brief conclusion is presented in section 5.

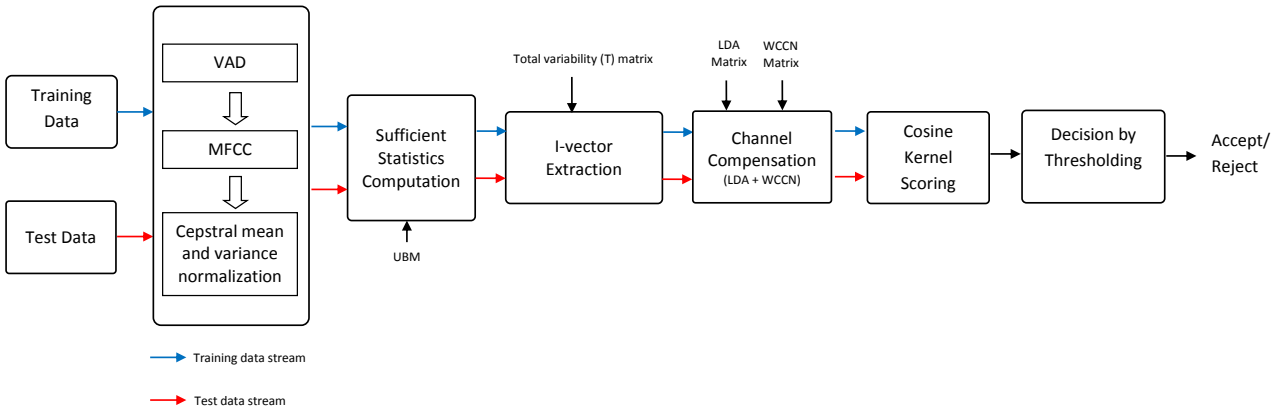

Figure 1: Block diagram of the i-vector based text-independent baseline SV system

2. Development of i-vector based baseline SV system: sufficient data conditions

The i-vector based speaker modeling has evolved from joint factor analysis(JFA) [9] which showed significant improvement over the traditional SV techniques. In contrast to JFA, i-vector based speaker modeling [1] considers both speaker and session space into a common space called total variability space which covers all the variabilities. In this kind of modeling, the Gaussian mixture model (GMM) mean supervectors [10] for a particular utterance are projected onto a low dimensional space called the total variability space, which gives a robust compact representation. These low dimensional vectors are called identity vectors or i-vectors. The matrix used for this transformation, which accounts for the dominant speaker as well as session/channel variabilities, is termed as Total variability matrix (T-matrix).

The i-vector based SV system as described in [1] is developed using the NIST SRE 2003 database. The NIST 2003 dataset contains data of 356 speakers (144 males and 212 females) for training their speaker models and 2559 test utterances for evaluating the performance of the SV system. Figure 1 shows the block diagram representation for the i-vector based text-independent baseline SV system. Both the train as well as test utterances undergo similar processing in this kind of system building. Preprocessing of the speech signals is done considering them as blocks of 20 ms with a shift of 10 ms. Energy based voice-activity detection (VAD) is performed for the speech utterances and the speech frames having energy greater than 0.07 times the average energy of the utterance are selected as frames of interest. 13 dimensional MFCC features including their first and second order derivatives are extracted for each of the frames, thus making up a 39 dimensional feature vector. The extracted features are then normalized to fit zero mean unit variance, i.e., cepstral mean subtraction (CMS) and variance normalization are performed for further processing.

For the purpose of building the universal background model (UBM) [10] and the T-matrix, Switchboard Corpus II cellular data of 1872 utterances is used as development data. The development data undergoes the same kind of preprocessing as mentioned in the case of train/test data. A gender-independent UBM of 1024 mixtures is trained using a subset of the development data of approximately 10 hours with equal amount of male and female speech. The entire development data is used to train a T-matrix of 400 columns which captures all the variability present in the speech data. Since the low dimensional i-vector representation is derived from the T-matrix, the i-vector based speaker modeling has both the speaker and channel information, and it requires some channel/session compensation methods for modeling only the speaker information for robust SV. For this purpose, 150 dimensional LDA and full dimensional WCCN are applied by learning the respective matrices using the development data.

The zeroth and the first order statistics (GMM mean supervectors) are computed from the train and test feature vectors which are then used along with the T-matrix to estimate the i-vectors as mentioned in [1]. LDA and WCCN are then applied on the i-vectors for channel/session compensation. Finally, cosine kernel scoring is done between the channel compensated train and test i-vectors to get the similarity scores. Table 1 shows the i-vector baseline system performance under sufficient data conditions. We can observe that the system performance improves significantly after the channel/session compensation is done using LDA and WCCN.

Table 1: *Performance of the baseline i-vector system on the NIST SRE 2003 dataset for sufficient data conditions*

Without Compensation		**With Compensation**	
EER(%)	**DCF**	**EER(%)**	**DCF**
4.74	0.0858	2.4	0.0474

3. Source features: DCTILPR

The voice source-based feature we extract from the speech signal is called the DCTILPR. The integrated linear prediction residual (ILPR) [11] is used as a voice source estimate, and its pitch synchronous discrete cosine transform (DCT) coefficients are taken as the feature vector. The DCTILPR has been shown to perform on par with existing voice source-based speaker-specific features in a speaker identification task [7]. Here, we use the DCTILPR features in a SV task on the NIST SRE 2003 database.

Figure 2 [7] shows the block diagram to extract the DCTILPR. The energy-based VAD described in section 2 is applied on the speech signal to get the frames with significant voice activity. On these frames, an epoch extraction algorithm [11] is

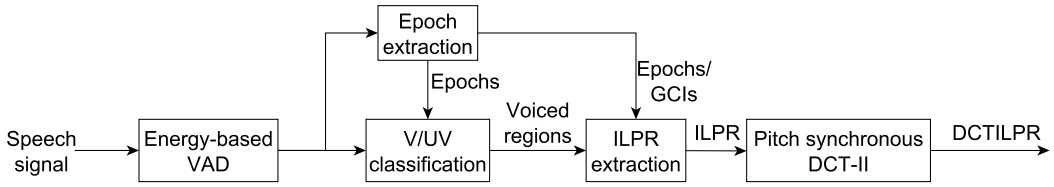

Figure 2: Block diagram of the method to extract DCTILPR

applied, and using these epochs, a voiced/unvoiced (V/UV) decision based on maximum normalized cross-correlation is applied as in [12]. Only the voiced regions are retained for further processing, and the ILPR is extracted on the voiced regions as in [11]. Using the epochs in the voiced regions as glottal closure instants (GCIs) and considering the interval between two successive GCIs as a pitch period, pitch synchronous DCT-II is obtained to get the DCTILPR. As shown in [7], the first 24 DCT coefficients capture the speaker information contained in the voice source, and are taken as the feature vector.

It has also been shown in [7] that the DCTILPR captures speaker-specific information which is not captured by the MFCCs. Thus, we combine the classifiers trained using the MFCCs and the DCTILPR as follows:

$$\boldsymbol{S}_{combi} = \boldsymbol{\alpha}\boldsymbol{S}_{dctilpr} + (\mathbf{1} - \boldsymbol{\alpha})\boldsymbol{S}_{mfcc} \quad (1)$$

where $\boldsymbol{S}_{dctilpr}$ and $\boldsymbol{S}_{mfcc}$ represent the scores obtained using DCTILPR and MFCC features respectively, with the i-vector based SV system. α is a scalar between 0 and 1, the optimal value of which is chosen for fusion of the two scores to give $\boldsymbol{S}_{combi}$.

4. Experimental results and analysis

The significant performance of the i-vector based system for the sufficient data conditions as discussed in section 2 motivates us to use this system for the case of short utterances too. The test segments for NIST 2003 SRE range between 15-45 s of duration. The i-vector based system is then evaluated by varying the duration of test data from 2 s to 10 s to analyze the performance in the case of short test utterances for MFCC features. From Table 2, we can see that, as the duration of the test utterance is decreased, the SV performance degrades significantly. Also, we can see that the results improve significantly after channel/session compensation is done for short utterances.

Table 2: *Results of i-vector system for different duration test segments on NIST 2003 dataset using MFCC features*

Test Utterance Duration	System Performance			
	Without Compensation		With Compensation	
	EER(%)	DCF	EER(%)	DCF
Full	4.74	0.0858	2.4	0.0474
20 s	5.51	0.1021	3.38	0.0606
15 s	6.41	0.1188	4.33	0.0813
10 s	8.85	0.1620	5.8	0.1090
5 s	13.91	0.2631	10.52	0.1977
3 s	19.82	0.3662	16.94	0.3100
2 s	25.38	0.4784	22.31	0.4128

Table 3: *Results of i-vector system for different duration test segments on the NIST 2003 dataset using DCTILPR features*

Test Utterance Duration	System Performance			
	Without Compensation		With Compensation	
	EER(%)	DCF	EER(%)	DCF
10 s	24.93	0.45	13.91	0.25
5 s	27.59	0.52	18.65	0.35
3 s	31.84	0.58	22.13	0.41
2 s	34.73	0.65	27.78	0.52

An i-vector based SV system is developed using DCTILPR features in the same way as described in section 2. As we are concentrating on SV for short utterances, we evaluate the performance of this system only for cases of 10 s or less. The results obtained for short utterances using DCTILPR features are shown in Table 3. It is clearly visible from Table 3 that the system using only DCTILPR features performs poorly in comparison to the system using MFCC features as shown in Table 2. This has been shown to be mainly due to the ILPR having more handset variability than the MFCCs [7].

The fusion of the MFCC and DCTILPR features are done at the score level using Equation 1 for the optimal value of $\boldsymbol{\alpha}$ in the range 0 to 1. The Table 4 shows the performance of the system for the combination of the stated source and system features for short utterances. It can be inferred from Table 4 that the performance improves significantly over that of the baseline for short utterances after fusion of the DCTILPR due to the additional speaker information present in it, which is not captured by the vocal tract features. $\boldsymbol{\alpha}_{opt}$ varies between 0.15 and 0.4 for different cases, which shows that the DCTILPR features must be given a weightage in that range for optimal performance. We can observe that the improvement in EER over the baseline is more and more pronounced as the duration of the test data decreases (5.81%-5.33%= 0.48% in the 10 s case to 22.31%-17.71%= 4.6% in the 2 s case, with compensation). Also, $\boldsymbol{\alpha}_{opt}$ increases as the duration of the test data decreases (0.15 in the 10 s case to 0.4 in the 2 s case, with compensation). Thus, the importance of the source feature increases as the test data duration decreases. As ILPR has handset variability issues, the performance improvement is more significant after the channel/session compensation. Figure 3 shows the evolution of equal error rate (EER) vs alpha($\boldsymbol{\alpha}$) for the 2 s case with channel/session compensation by applying LDA and WCCN. We can see that, if an optimal weightage of 0.4 is given to the DCTILPR features, they improve the EER by 4.6%. Thus, combining information from a source feature improves the system performance in case of limited data SV.

Table 4: *Performance comparison of i-vector system for short test segments on NIST 2003 dataset for the baseline system vs. proposed system of fusing both DCTILPR and MFCC features*

Test Utterance Duration	Performance- Baseline System				Performance- Proposed System					
	Without Compensation		With Compensation		Without Compensation			With Compensation		
	EER(%)	DCF	EER(%)	DCF	α_{opt}	EER(%)	DCF	α_{opt}	EER(%)	DCF
10 s	8.85	0.1620	**5.81**	**0.1090**	0.25	7.90	0.1491	0.15	**5.33**	**0.0971**
5 s	13.91	0.2631	**10.52**	**0.1977**	0.3	12.20	0.2290	0.3	**8.45**	**0.1567**
3 s	19.82	0.3662	**16.94**	**0.3100**	0.3	16.98	0.3213	0.4	**12.46**	**0.2325**
2 s	25.38	0.4784	**22.31**	**0.4128**	0.3	23.08	0.4313	0.4	**17.71**	**0.3351**

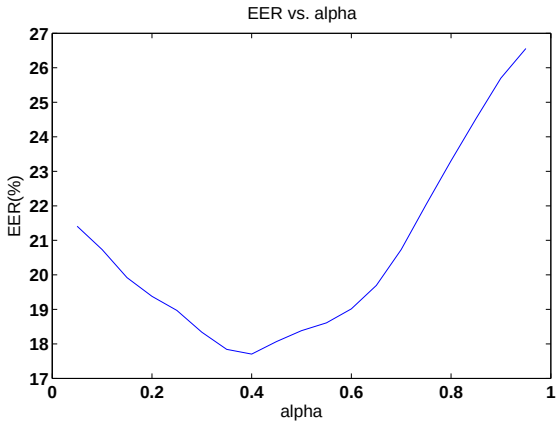

Figure 3: EER vs. alpha (α). An EER absolute improvement of 4.6% at α=0.4 ($\alpha_{opt} = 0.4$) can be observed for the case of 2 s test data, indicating that the DCTILPR features provide speaker information not captured by the MFCCs and improve performance in short test utterance cases

5. Conclusions

Limited data SV is a challenge to the speech community for implementation of a practical system. The paper presents the significance of source information in SV system for short test utterances. The initial studies for the baseline system using vocal tract features for short test utterances shown in the paper, is improved on addition of the source feature DCTILPR at the score level. The significant improvement in performance is due to the different/additional speaker information present in the source feature. The importance of the source feature becomes more significant as the duration of the test data is reduced. An absolute improvement of 4.6% EER is reported for test data of 2 sonds after channel/session compensation.

6. Acknowledgement

This work is part of the ongoing project on the development of “Speech based multi-level person authentication system” funded by the Department of Electronics & Information Technology (DeitY), Govt. of India.

7. References

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *Audio, Speech, and Language Processing, IEEE Trans. on*, vol. 19, no. 4, pp. 788 –798, may 2011.

[2] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, “i-vector based speaker recognition on short utterances,” in *Interspeech 2011*, 2011.

[3] K. Murthy and B. Yegananarayana, “Combining evidence from residual phase and mfcc features for speaker recognition,” *IEEE Signal Processing Letters*, vol. 13(1), pp. 52–55, 2006.

[4] J. Gudnason and M. Brookes, “Voice source cepstrum coefficients for speaker identification,” in *Proc. ICASSP*, 2008, pp. 4821–4824.

[5] S. R. M. Prasanna, C. Gupta, and B. Yegananarayana, “Extraction of speaker specific information from linear prediction residual of speech,” *Speech Communication*, vol. 48, pp. 1243–1261, 2006.

[6] W. Chan, N. Zheng, and T. Lee, “Discrimination power of vocal source and vocal tract related features for speaker segmentation,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15 (6), pp. 1884–1892, 2007.

[7] A. G. Ramakrishnan, B. Abhiram, and S. R. M. Prasanna, “Characterization of the voice source using DCT for speaker information,” *IEEE Signal Processing Letters*, Submitted on 18 February, 2014.

[8] *“The NIST Year 2003 Speaker Recognition Evaluation Plan”, NIST , Feb 2003*.

[9] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Joint factor analysis verses eigenchannels in speaker verifiaction,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15,no.4, pp. 1435–1447, May 2007.

[10] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[11] A. P. Prathosh, T. V. Ananthapadmanabha, and A. G. Ramakrishnan, “Epoch extraction based on integrated linear prediction residual using plosion index,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, issue 12, pp. 2471 – 2480, 2013.

[12] T. V. Ananthapadmanabha, A. P. Prathosh, and A. G. Ramakrishnan, “Detection of closure-burst transitions of stops and affricates in continuous speech using plosion index,” *Journal of the Acoustical Society of America*, vol. 135, no. 1, pp. 460–471, 2014.