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-560012, India
{rohankd, prasanna}@iitg.ernet.in, abhiram1989@gmail.com, 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], i-vector based SV system for short utterances is analyzed 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 source feature which captures complementary 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] suggest that the amount of train/test data can be less for the voice source features than that for vocal tract features. This is due to the fact that the voice source features do not depend much on phonetic content, whereas the robustness of vocal tract feature depends on the amount of phonetic content that it captures for a particular utterance. This motivates us to use voice source features along with the conventional vocal tract features for limited data SV.

This paper focuses 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 a good speaker identification on 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 two systems are fused 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 i-vector based SV followed by channel/session compensation techniques for robust speaker modeling for 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 combination of two at the score level for short utterances are explained and their results are reported. Finally, a brief conclusion is presented in section 5.

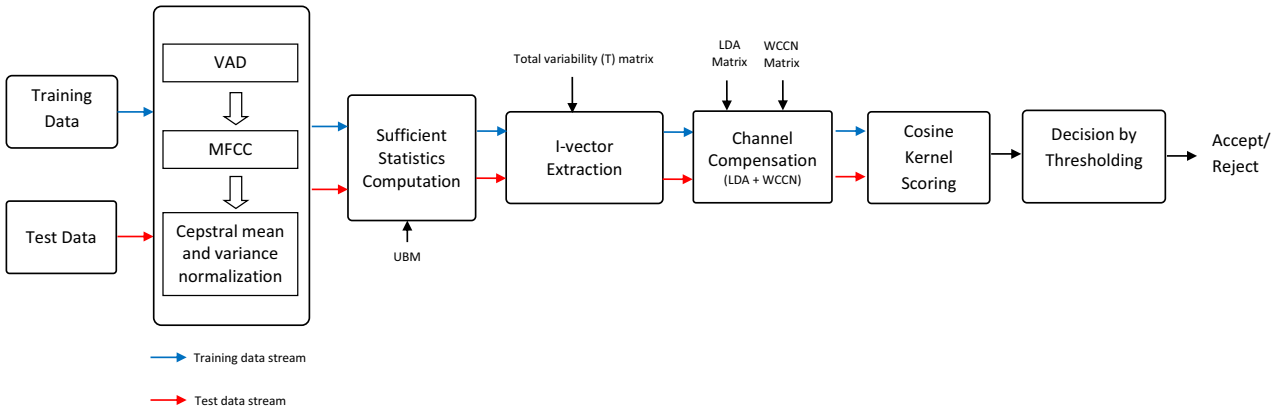

Figure 1: Block diagram of the i-vector based text-independent baseline SV system [9]

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) [10] 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 [11] for a particular utterance are projected on to 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 male and 212 female speakers) for training their speaker models and 2559 test utterances for evaluating the performance of SV system. Figure 1 [9] shows the block diagram of 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. The speech signals are processed 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 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 cepstral variance normalization (CVN) are performed for further processing.

For the purpose of building the universal background model (UBM) [11] 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 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 variabilities 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 based baseline SV system performance in terms of equal error rate (EER) and decision cost funciton (DCF) 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 NIST SRE 2003 dataset using MFCC features 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) [12] 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 NIST SRE 2003 database.

Figure 2 [7] shows the block diagram to extract the DCTILPR. The energy-based VAD described in Section 2 is ap-

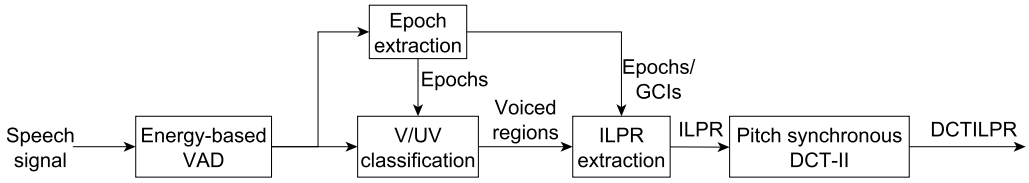

Figure 2: Block diagram of the method to extract DCTILPR [7]

plied on the speech signal to get frames with significant voice activity. On these frames, an epoch extraction algorithm [12] is applied, and using these epochs, a voiced/unvoiced (V/UV) decision based on maximum normalized cross-correlation is applied as in [13]. Only the voiced regions are retained for further processing, and the ILPR is extracted on the voiced regions as in [12]. 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.

Table 2: *Performance of the i-vector system on NIST SRE 2003 dataset using DCTILPR features for sufficient data conditions*

Without Compensation		**With Compensation**	
EER (%)	**DCF**	**EER (%)**	**DCF**
21.55	0.3594	12.01	0.2112

An i-vector based SV system similar to the one described in Section 2 is developed with the DCTILPR as the speaker-specific features. In this case, since the features are not in the cepstral domain, CMS is not done. The features are normalized with respect to the positive peak amplitude of the ILPR, as in [7]. The other blocks in the i-vector based SV system are implemented in the same way as described in Section 2. It has also been shown in [7] that the DCTILPR captures speaker-specific information which is not captured by the MFCCs. However the DCTILPR alone does not give significant performance as mentioned in [7], which can be seen by comparing Table 2 with the baseline system performance in Table 1. This is because the DCTILPR features suffer from more handset variability than the MFCCs, as shown in [7]. It is also found in [7] that the combination of the two features significantly improves the performance. Thus, for better speaker modeling the combined score $\boldsymbol{S_c}$ of the classifiers trained using the MFCCs and the DCTILPR is obtained as follows:

$$\boldsymbol{S_c} = \boldsymbol{\alpha S_d} + (\mathbf{1} - \boldsymbol{\alpha}) \boldsymbol{S_m} \tag{1}$$

where $\boldsymbol{S_d}$ and $\boldsymbol{S_m}$ 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 $\boldsymbol{S_c}$.

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 3, 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 for short utterances with channel/session compensation.

Table 3: *Results of baseline i-vector system on NIST 2003 dataset using MFCC features for limited duration test segments*

Test Utterance Duration	**System Performance for MFCC Features**			
	Without Compensation		**With Compensation**	
	EER (%)	**DCF**	**EER (%)**	**DCF**
10 s	8.85	0.1620	5.81	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 4: *Results of i-vector system on NIST 2003 dataset using DCTILPR features for limited duration test segments*

Test Utterance Duration	**System Performance for DCTILPR Features**			
	Without Compensation		**With Compensation**	
	EER (%)	**DCF**	**EER (%)**	**DCF**
10 s	24.93	0.4471	13.91	0.2497
5 s	27.59	0.5182	18.65	0.3460
3 s	31.84	0.5797	22.13	0.4077
2 s	34.73	0.6537	27.78	0.5198

The performance of the i-vector based SV system developed using DCTILPR features is then evaluated for different durations of test data to compare with the baseline system developed using MFCC features. Since the study is 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 4. Clearly the system using only DCTILPR features performs poorly in comparison to the system using MFCC features. 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 based systems is performed at the score level using Equation 1 for the optimal value of $\boldsymbol{\alpha}$ in the range 0 to 1. Table 5 shows the performance of the proposed system for the combination of stated source and system features for short utterances. It can be inferred from Table 5 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

Table 5: *Performance of the proposed i-vector system for short test segments on NIST 2003 dataset fusing DCTILPR and MFCC features and absolute performance improvement over the baseline system*

Test Utterance Duration	Performance- Proposed System						Absolute Performance Improvement			
	Without Compensation			With Compensation			Without Compensation		With Compensation	
	α_{opt}	EER (%)	DCF	α_{opt}	EER (%)	DCF	EER (%)	DCF	EER (%)	DCF
10 s	0.25	7.90	0.1491	0.15	**5.33**	**0.0971**	0.95	0.0129	**0.48**	**0.0119**
5 s	0.3	12.20	0.2290	0.3	**8.45**	**0.1567**	1.71	0.0341	**2.07**	**0.0380**
3 s	0.3	16.98	0.3213	0.4	**12.46**	**0.2325**	2.84	0.0449	**4.48**	**0.0775**
2 s	0.3	23.08	0.4313	0.4	**17.71**	**0.3351**	3.07	0.0471	**4.60**	**0.0777**

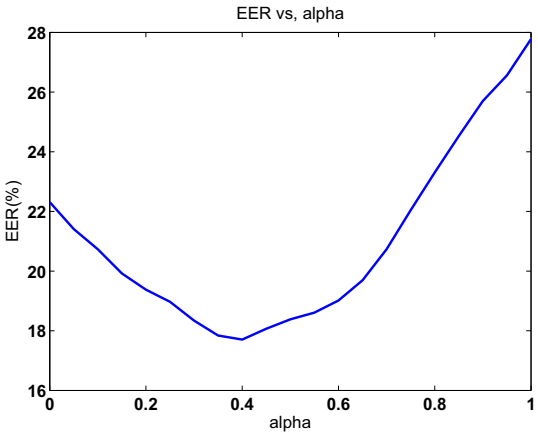

Figure 3: EER vs. alpha ($\boldsymbol{\alpha}$). An absolute improvement of 4.6% EER at $\boldsymbol{\alpha}$=0.4 ($\alpha_{opt} = 0.4$) can be observed after compensation for the case of 2 s test data, indicating that the DCTILPR features carry speaker information not captured by the MFCCs and improve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

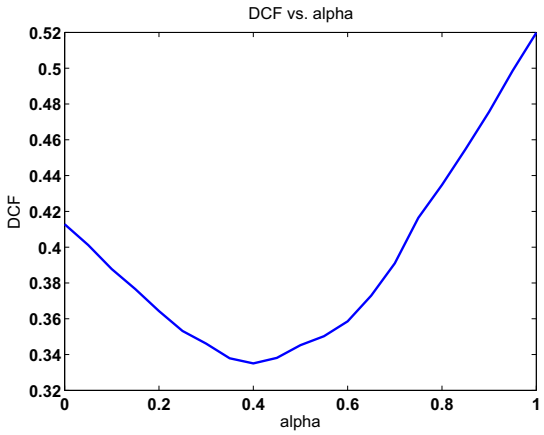

Figure 4: DCF vs. alpha ($\boldsymbol{\alpha}$). An absolute improvement of 0.0777 DCF at $\boldsymbol{\alpha}$=0.4 ($\alpha_{opt} = 0.4$) can be observed after compensation for the case of 2 s test data, indicating that the DCTILPR features carry speaker information not captured by the MFCCs and improve performance in short test utterance

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 variation of 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%.

The DCF also follows the same trend in improvement like the EER. It is more significant as the duration of the test data is reduced (0.1090-0.0971= 0.0119 in the 10 s case to 0.4128-0.3351= 0.0777 in the 2 s case, with compensation). Figure 4 shows the variation of DCF vs alpha ($\boldsymbol{\alpha}$) for the 2 s case with channel/session compensation. It is observed that for an optimal value of $\boldsymbol{\alpha}$ = 0.4, an absolute improvement of 0.0777 in the value of DCF. Thus, combining information from a source feature improves the system performance in the case of limited data SV.

5. Conclusion

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 performance of the baseline system using vocal tract features for short test utterances improves 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 and 0.0777 DCF are obtained for test data of 2 s after channel/session compensation. We intend to evaluate the performance of the combination of source and system features for short utterances on the larger NIST SRE 2012 database in future.

6. Acknowledgement

This work is part of the ongoing project on the development of “Speech based multi-level person authentication system” funded by the e-security division of 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," *IEEE Trans. on Audio, Speech, and Language Processing*, 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, "A characterization of the voice source using pitch synchronous discrete cosine transform for speaker information," *JASA Express Letters*, Submitted on 14 May, 2014.

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

[9] S. Dey, S. Barman, R. K. Bhukya, R. K. Das, Haris B C, S. R. M. Prasanna, and R. Sinha, "Speech biometric based attendance system," in *National Conference on Communications*, 2014.

[10] 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.

[11] 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.

[12] 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.

[13] 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.