Exploring different attributes of source information for speaker verification with limited test data

Rohan Kumar Das, and S. R. Mahadeva Prasanna

Citation: The Journal of the Acoustical Society of America **140**, 184 (2016); doi: 10.1121/1.4954653
View online: http://dx.doi.org/10.1121/1.4954653
View Table of Contents: http://asa.scitation.org/toc/jas/140/1
Published by the Acoustical Society of America

Exploring different attributes of source information for speaker verification with limited test data

Rohan Kumar Das[a] and S. R. Mahadeva Prasanna
Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati-781039, Assam, India

(Received 14 December 2015; revised 8 June 2016; accepted 10 June 2016; published online 12 July 2016)

This work explores mel power difference of spectrum in subband, residual mel frequency cepstral coefficient, and discrete cosine transform of the integrated linear prediction residual for speaker verification under limited test data conditions. These three source features are found to capture different attributes of source information, namely, periodicity, smoothed spectrum information, and shape of the glottal signal, respectively. On the NIST SRE 2003 database, the proposed combination of the three source features performs better [equal error rate (EER): 20.19%, decision cost function (DCF): 0.3759] than the mel frequency cepstral coefficient feature (EER: 22.31%, DCF: 0.4128) for 2 s duration of test segments. © *2016 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4954653]

[DDS] Pages: 184–190

I. INTRODUCTION

The recent trend in the field of speaker verification (SV) is towards addressing issues from a practical deployment point of view.[1,2] From a generic framework and also providing robustness against spoofing, the text independent (TI) SV system is preferred. The major constraint of the TI SV system from the practical deployment point of view is the amount of speech data necessary for enrollment and testing. Limited speech data are favoured for providing user comfort. Although the i-vector based speaker modeling is the state of art approach for SV, the performance is found to deteriorate with limited speech data.[3,4] From the perspective of deployable systems, the desirable condition is sufficient train with limited test data. This work focuses on such a condition where the typical duration of enrollment data is about 3 min and that of test data is preferably $\leq$10 s.

Voice source features represent speaker characteristics in terms of modeling the glottal excitation signal generated from the voiced sound units that originate from source/filter model of speech. This information may be in terms of glottal pulse shape, fundamental frequency, and many such aspects of excitation source. There have been many efforts in the past to explore the source information that give a better SV performance when combined with the mel frequency cepstral coefficient (MFCC) feature of speech representing vocal tract aspect of speaker information.[5–8] This is due to the complementary nature of information captured by the source features in comparison to the system features. Also, as the excitation source features are less dependent on the amount of phonetic content, the duration of enrollment/testing can be less.[9,10] In earlier work,[11,12] a source feature called discrete cosine transform of the integrated linear prediction residual (DCTILPR) is computed. The score level fusion of this feature with MFCC feature was an improvement over the baseline system using only MFCC under limited test data conditions. This reinforced the earlier findings of the literature.[5–8] Further, the improvement in performance is found to be more evident with a decrease in the duration of test data. The work on source features is extended to introduce mel power difference of spectrum in the subband (MPDSS) feature to capture the source information in a different manner.[13,14] Fusion of this feature with MFCC at the score level is also found to be an improvement. But at the same time it is noticed that the information captured by the two features, DCTILPR and MPDSS, had different aspects of source characteristics. This motivated us to explore the different attributes of source information carried by different source features.

The linear prediction (LP) residual of speech that mostly contains excitation source information, does not contain second order relations as they are already extracted by LP analysis.[9,15] Therefore, when compact representation of the source feature by some signal processing method is obtained, it may not capture all the aspects of the source. Also, the noise like structure of the LP residual itself creates difficulty in compact representation of source information. This signifies the need for an alternative approach for source modeling. One such direction is capturing different attributes of source information and using them in combination for better speaker modeling.

In this work, one more source feature called residual mel frequency cepstral coefficient (RMFCC) is included along with the previously proposed DCTILPR and MPDSS features.[16] The different attributes are explored from the origin of these three source features and their usefulness for representing the source information when used in combination is investigated. The RMFCC is obtained by frame based processing of LP residual and provides compact representation of source information using cepstral analysis. Alternatively, DCTILPR is obtained by pitch synchronous analysis which provides compact representation of source

[a]Electronic mail: rohankd@iitg.ernet.in

information using discrete cosine transform (DCT). Even though the LP residual is common, the signal processing approach employed in both the cases are different. Accordingly, the source information representation may also be different. The source features explored in this work have been previously used for sufficient data conditions and found to be carrying speaker specific information. However, the novelty of this work lies in exploring the importance of the source features MPDSS, RMFCC, and DCTILPR for limited test data conditions and their different attributes of source information, which in combination give a better performance than standalone MFCC based vocal tract feature.

The rest of the work is organized in the following order: Section II describes the characteristics of the three source features, MPDSS, RMFCC, and DCTILPR used for speaker modeling. In Sec. III, the baseline SV system developed using the i-vector framework is described. Section IV experimentally demonstrates different attributes of source information. This work ends with summary and conclusion in Sec. V.

II. DIFFERENT ATTRIBUTES OF SOURCE FEATURES FOR SPEAKER MODELING

Source features give information about the glottal excitation in the form of pitch period, strength of excitation, glottal signal shape, etc. Since the glottis and associated muscle structure are unique for each individual, the information represented by the source features is expected to be unique for each speaker and can be utilized for SV.

A. MPDSS feature

The mel power difference of spectrum in subbands (MPDSS) feature exploits the spectral flatness of the LP residual spectrum.[13,14] Depending on the type of voice (say, hard or soft), the spectral flatness of the LP spectrum also changes. If the voice is hard, then it results in rapid and complete closure of vocal folds. Accordingly, the flow is discontinuous and excitation is more impulse-like. This results in more spectral flatness, equivalently less dynamic range, or less periodic nature. The MPDSS feature is an attempt to capture the periodicity aspect of excitation source information by using mel bank filters over subbands of power difference in spectrum. MPDSS [$M(m)$] can be defined as

$$M(m) = 1 - \frac{\left[\prod_{k=l_m}^{h_m} P(k)\right]^{1/N_m}}{\frac{1}{N_m}\sum_{k=l_m}^{h_m} P(k)}, \tag{1}$$

where $N_m = h_m - l_m + 1$ is the total number of samples, l_m and h_m denoting the first and last sample of the subband in the mth filter. The $P(k)$ corresponds to the power of the kth sample of the subband.

MPDSS is a subband version interpretation of spectral flatness measure. The ratio of the geometric mean (GM) to the arithmetic mean (AM) of the spectral samples gives the spectral flatness of a spectrum. If the spectral flatness is larger, then the dynamic range of the power spectrum is smaller and MPDSS values become closer to 0, denoting lesser periodicity. On the other hand, if the spectrum contains more difference between peaks and dips showing more dynamic range, the MPDSS values get closer to 1, indicating stronger periodicity. The subband based interpretation of this feature is better in the sense that the feature values equal to number of filters can be obtained. Also the use of mel filter bank helps in deciding subband spacing in a non-linear way which is motivated from the non-linear nature of human

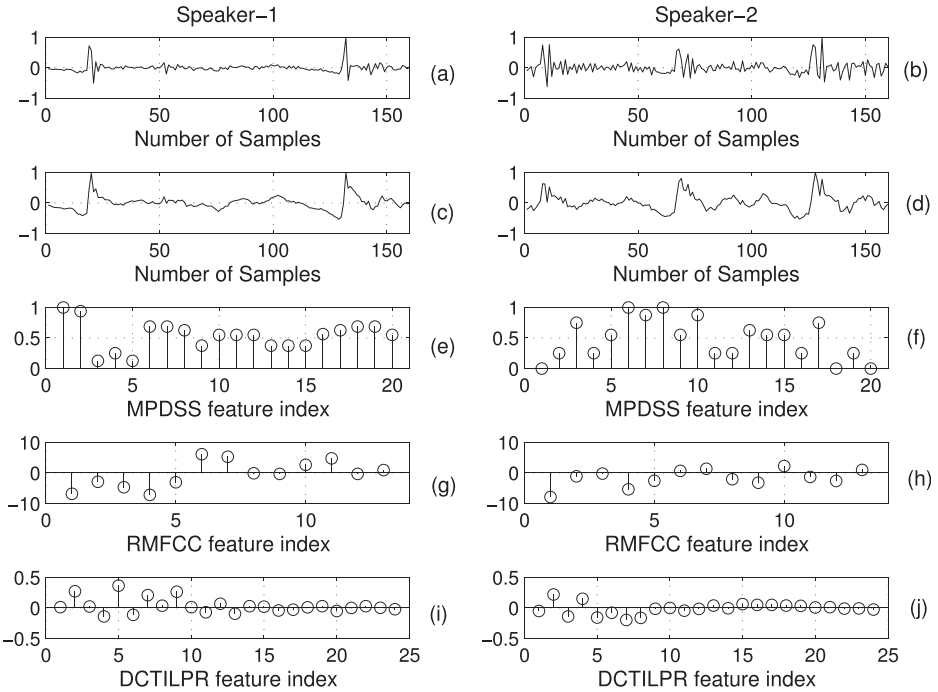

FIG. 1. Three source features for two different speakers for Speaker-1 and Speaker-2 from the vowel /a/ in the word "dark" for two different speakers. (a), (b) LP residual signals, (c), (d) ILPR, (e), (f) MPDSS feature, (g),(h) RMFCC feature, (i), (j) DCTILPR feature.

TABLE I. Performance of different features on NIST SRE 2003 dataset under limited duration test segments over i-vector framework.

	MFCC		MPDSS		RMFCC		DCTILPR	
Test Duration	EER (%)	DCF	EER (%)	DCF	EER (%)	DCF	EER (%)	DCF
10 s	5.81	0.1090	17.43	0.3269	12.96	0.2466	13.91	0.2497
5 s	10.52	0.1977	22.58	0.4250	18.88	0.3462	18.65	0.3460
3 s	16.94	0.3100	27.60	0.5202	23.62	0.4362	22.13	0.4077
2 s	22.31	0.4128	31.44	0.5958	27.55	0.5203	27.78	0.5198

auditory perception system.[13] In this case, 20 melbank filters are used, which form a 20-dimensional MPDSS feature for each speech frame.

B. RMFCC feature

The RMFCC involves processing the LP residual in the cepstral domain unlike the former feature MPDSS. The log magnitude spectrum of LP residual is passed through a non-uniform filter bank with triangular windows placed on the mel-frequency scale and then inverse discrete Fourier transform (IDFT) is computed to obtain RMFCC feature.[16] Let r(n) be the LP residual of a speech segment and R(w) its spectrum, the log magnitude of which is passed via melbank filters M_l for non-linear transformation. Then RMFCC feature $[R(k)]$ is computed in the following way,

$$R(k) = \mathrm{IDFT}[M_l(\log|R(w)|)]. \tag{2}$$

The first 13 dimensions along with 13-Δ and 13-ΔΔ are considered to form 39-dimensional RMFCC feature for short term processed speech signal with frames of 20 ms with a shift of 10 ms. Accordingly, RMFCC represents segmental level smoothed spectrum information due to mel filterbank. This, in turn, may correspond to the average glottal signal information.

C. DCTILPR feature

The DCTILPR source feature is obtained using the ILPR in temporal domain.[11,12] An epoch extraction algorithm is applied, and using these epochs, a voiced/unvoiced decision based on maximum normalized cross-correlation is applied as in Refs. 17 and 18. The epochs from the voiced regions are taken as glottal closure instants (GCIs) and for the interval between one GCI to the succeeding GCI, pitch synchronous DCT-II having compaction property is taken for compression. The first 24 coefficients are termed as DCTILPR feature. As the source feature DCTILPR captures the glottal signal shape information, pitch synchronous analysis is made for precisely capturing this aspect of source characteristics.

Let, $i_r(n)$ be the integrated linear prediction residual (ILPR) corresponding to the LP residual $r(n)$ extracted between epoch locations j and $(j+1)$ of a speech segment. The respective DCTILPR feature $[D(k)]$ taken in pitch synchronous manner is given by

$$D(k) = \sum_{n=0}^{N-1} i_r(n)\cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right], \tag{3}$$

where N is the number of samples between the epoch locations j and $(j+1)$ and $k = 0, 1, 2, \ldots, N-1$.

TABLE III. Canonical correlation analysis (CCA) measure to highlight the nature of complementary characteristics from one feature to another.

Feature pairs	Correlation
MPDSS vs MFCC	0.89
RMFCC vs MFCC	0.91
DCTILPR vs MFCC	0.79
MPDSS vs RMFCC	0.91
MPDSS vs DCTILPR	0.81
RMFCC vs DCTILPR	0.82

D. Different attributes of source

Figure 1 shows the nature of LP residual, ILPR and the three source features for two speakers from TIMIT database for the vowel /a/ from word "dark."[19] The features MPDSS, RMFCC, and DCTILPR involve spectral, cepstral, and temporal domain processing of LP residual, respectively. The MPDSS and RMFCC features involve segmental processing on the speech signal. Thus, they model the excitation source information averaged over two to three pitch periods. On the contrary, DCTILPR feature is extracted by pitch synchronous analysis. Hence it models the source information within a pitch period representing the shape of glottal signal. Due to different domains of processing, different equations for extraction, and segmental vs pitch-synchronous ways of extracting information, each of these features is hypothesized to be capturing different attributes of an excitation source.

TABLE II. Performance under fusion of different source features with MFCC on NIST SRE 2003 dataset for limited duration of test segments and comparison to baseline MFCC feature based performance showing improvements highlighted by boldface numbers in each of the combination case.

	MFCC		MPDSS+MFCC			RMFCC+MFCC			DCTILPR+MFCC		
Test Duration	EER (%)	DCF	β_{opt}	EER (%)	DCF	β_{opt}	EER (%)	DCF	β_{opt}	EER (%)	DCF
10 s	5.81	0.1090	0.1	**5.56**	**0.1048**	0.15	**5.78**	**0.1087**	0.15	**5.33**	**0.0971**
5 s	10.52	0.1977	0.15	**10.12**	**0.1850**	0.25	**9.67**	**0.1829**	0.3	**8.45**	**0.1567**
3 s	16.94	0.3100	0.35	**15.04**	**0.2811**	0.3	**14.96**	**0.2790**	0.4	**12.46**	**0.2325**
2 s	22.31	0.4128	0.4	**19.96**	**0.3720**	0.4	**20.19**	**0.3767**	0.4	**17.71**	**0.3351**

TABLE IV. Performance under fusion of two source feature pairs and combined fusion of three source features showing different attributes on NIST SRE 2003 dataset.

Test Duration	DCTILPR+MPDSS			DCTILPR+RMFCC			MPDSS+RMFCC			Source fusion	
	β_{opt}	EER (%)	DCF	β_{opt}	EER (%)	DCF	β_{opt}	EER (%)	DCF	EER (%)	DCF
10 s	0.45	10.93	0.2052	0.65	12.33	0.2285	0.65	14.96	0.2815	10.57	0.1964
5 s	0.45	14.99	0.2785	0.4	13.37	0.2492	0.45	15.58	0.2955	11.97	0.2252
3 s	0.5	18.29	0.3413	0.5	**16.67**	**0.3109**	0.4	20.28	0.3802	**15.85**	**0.2854**
2 s	0.45	23.85	0.4456	0.45	**21.59**	**0.4065**	0.45	24.16	0.4578	**20.19**	**0.3759**

DCTILPR captures the glottal shape information of a speaker in a pitch synchronous manner.[11,12] But this does not capture the periodicity information of the signal denoting how much periodic it is. MPDSS feature is a variant of spectral flatness measure, and captures the periodicity information as the peak to dip ratio of the spectrum of a signal measures the periodicity.[13,14] Thus the periodicity attribute captured by the MPDSS is an additive information for representing the source characteristics. Further the RMFCC feature is extracted from the LP residual by its cepstral analysis. Due to the noise like structure of LP residual the information captured without any processing is less distinctive across speakers. The RMFCC feature computed from the subband energies, provides a segmental level smoothed spectrum information by capturing the strength of excitation.[16] Therefore the different attributes of these source features in combination is expected to enhance SV performance.

III. DEVELOPMENT OF BASELINE SV SYSTEM

The i-vector based modeling approach is considered for the SV studies.[3] This approach defines a single subspace assumed to capture all the variabilities such as speaker, channel/session, etc., denoted by the total variability space. This subspace transforms the Gaussian mixture model (GMM) mean supervector of an utterance to a low dimensional vector, called i-vector, that possess the dominant speaker specific information. The projection matrix used for the transformation is termed as total variability matrix (T-matrix). Different channel/session compensation techniques are applied over it to nullify the channel and session effects.

In this work, a baseline system considering 39-dimensional MFCC feature (13-base + 13-velocity + 13-acceleration coefficients) is developed over the NIST SRE 2003 database in the i-vector framework. The database consists of 356 speaker's data containing a population of 212 female and 144 male speakers. There are 2559 test utterances from these speakers in the database that are used for verification against the 356 speaker model set. The typical duration for enrollment session of the speakers is about 2–3 min and the testing session files are of 15–45 s duration. For studies under limited data, the duration of the testing sessions are made $\leq$10 s. Segmental processing is done over the speech segments with 20 ms frame size with a frame shift of 10 ms to select the speech regions by using energy based voice activity detection (VAD) method and cepstral mean variance normalized features of those regions are retained. Switchboard Corpus-II dataset is used as a development data, from which 251 male and 251 female speaker's data are used for building the gender independent universal background model (UBM) of 1024 Gaussian mixtures. The sufficient statistics (zeroth and first order statistics) of the train, test, and development data are extracted using the trained UBM parameters. The sufficient statistics of the development data are then used to train a T-matrix of 400 columns. The compact low dimensional representation, i-vectors of train, test, and development data are estimated using the T-matrix on respective sufficient statistics. Further, linear discriminant analysis (LDA) and within class covariance normalization (WCCN) matrices are learned using the development data i-vectors to use them for channel/session compensation. Finally, the SV task is performed as per the NIST SRE 2003 evaluation plan, that has 2559 genuine claims and 25 935 impostor claims from the test set, which is kept standardized for evaluating system performance.[20] The validation of test claims are made by taking cosine kernel between the train and test i-vectors obtained by i-vector based speaker modeling approach.

In a similar way, three parallel systems are built using the source features MPDSS, RMFCC, and DCTILPR over i-vector framework. Table I shows the performance of the baseline system using MFCC features along with the three systems based on the source features for limited test data conditions ($\leq$10 s) in terms of equal error rate (EER) and decision cost function

TABLE V. Performance under fusion of two source features with MFCC and its comparison to (Source fusion+MFCC) indicating better results for fusion of three source attributes when combined to MFCC.

Test Duration	(DCTILPR+MPDSS)+MFCC		(DCTILPR+RMFCC)+MFCC		(MPDSS+RMFCC)+MFCC		Source Fusion+MFCC	
	EER (%)	DCF	EER (%)	DCF	EER (%)	DCF	EER (%)	DCF
10 s	**5.24**	**0.0975**	**5.19**	**0.0979**	**5.42**	**0.1008**	**5.10**	**0.0965**
5 s	**8.40**	**0.1531**	**8.36**	**0.1553**	**9.58**	**0.1776**	**8.18**	**0.1524**
3 s	**12.15**	**0.2242**	**11.97**	**0.2215**	**13.96**	**0.2578**	**11.47**	**0.2148**
2 s	**17.12**	**0.3216**	**16.98**	**0.3262**	**18.34**	**0.3466**	**16.08**	**0.3025**

(DCF). As reported in the literature, the performance of the individual source features is significantly lower as compared to MFCC. However, as the amount of test data decreases, the performance degradation in the case of source features is relatively less as compared to MFCC. This is due to the fact that the voice source features are less dependent on the phonetic content unlike the system features that models the speaker characteristics based on more coverage of acoustic space for a speaker, capturing the phonetic variation. Moreover, the MFCC features model the vocal tract characteristics to a large extent, whereas each of the source features model one aspect of the source information. Thus, under limited test data conditions, even though a single source feature does not perform better than MFCC, the fusion of multiple source features may help in better characterization of source information, which in turn can help in improving SV performance.

IV. COMBINATION OF DIFFERENT ATTRIBUTES

As demonstrated in earlier studies, the source features perform well in combination with MFCC features, especially for limited test data conditions.[8,12,14] The study is extended for the RMFCC feature to view the trend. Each of the source features MPDSS, RMFCC, and DCTILPR are fused at the score level as given by Eq. (4).

$$S_c = \beta S_s + (1-\beta) S_m, \tag{4}$$

where S_s, S_m, and S_c denotes the scores obtained using particular source feature, MFCC feature and the combination of the two, respectively. The scores obtained from two features are fused with different weights varying between 0 to 1 in steps of 0.05 and the optimal weight value β is considered for which the performance in terms of EER is minimal that is tuned on the development set.

Table II shows the performance of the fusion of each of the three source features with the MFCC feature. The improvement in performance for each case is more apparent as the duration of the test speech segment is reduced. To study the different attributes in terms of correlation measure for each of the source features, canonical correlation analysis (CCA) is performed among the source features and MFCCs. Table III shows that there is some complementary nature of information carried by each feature in combination with another (correlation value being less than 1), DCTILPR showing more complementary information to the other two source features as well as MFCC. The three source features are combined at the score level similar to the case as given by Eq. (4) considering scores obtained from two features at a time. Table IV shows the results for fusion of different source features that give improvement on fusion to one another depicting different nature of source information carried by each of the source features. The trend shown by CCA is also reflected while performing fusion of different features as can be observed from Tables II and IV. This is due to the fact that the greater the amount of complementary information for each of the feature pairs, the more it helps to achieve better performance on fusion. As each of the source features showed improvement on fusion as a pair of two features, depicting different aspects of source, the fusion of all the three source features is carried out by averaging the scores and the performances are reported in last column of Table IV. Due to the combination of the three different source features carrying different attributes of source, the performance obtained for 2 and 3 s cases outperform the MFCC feature for respective cases by a larger margin than that observed from fusion of only DCTILPR and RMFCC features. This result infers that, even though any one source feature does not provide good performance, when combined they outperform MFCC. Thus source features may be capturing different attributes of excitation source and, hence, significant improvement in performance is obtained when they are combined.

Finally, the combination of two source features at a time with MFCC is carried out and then the fusion of all the three

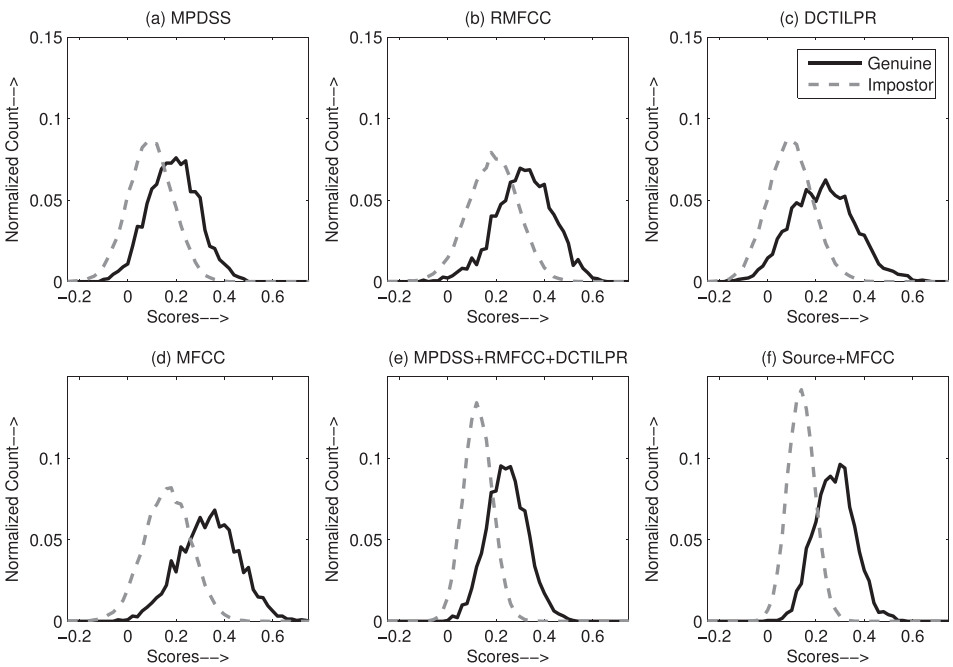

FIG. 2. Histograms of scores using different features and their combination for 2 s duration test data case.

TABLE VI. Area of overlap of genuine and impostor score histograms indicating better separability for three source features fusion and their fusion with MFCC features.

Feature	Overlap (%)
MPDSS	62.17
RMFCC	55.20
DCTILPR	52.94
MFCC	42.90
Source fusion	**39.09**
Source+MFCC	**30.85**

source features along with MFCC is made, the results of which are reported in Table V. In all the cases, significant improvements are seen, which are more prominent when all three source features carrying different attributes of each source are fused with MFCC, indicating the importance of each source feature for excitation source characterization that helps in improving SV performance. Thus, all the source features are found to be necessary for better results and their necessity increases with reduction of test data. Figure 2 shows the histogram of the scores obtained from the 2559 genuine and 25 935 impostor claims of NIST SRE 2003 database, for 2 s test data case for different features and their combination. It indicates more separability of genuine and impostor scores for the fusion of the three source features than baseline MFCC-based system. To quantify the same, the area of overlap (in %) is computed for the genuine and impostor score histograms and is shown in Table VI. Also, separability enhances on fusion of MFCC with three source features. Figure 3 illustrates the detection error tradeoff (DET) curve trend for different features and combination of features for the case of 2 s test data. The combination of three source features gives better performance than standalone vocal tract features. Also the fusion of three source features with the vocal tract information enhances the baseline performance based on MFCC features by a large margin.

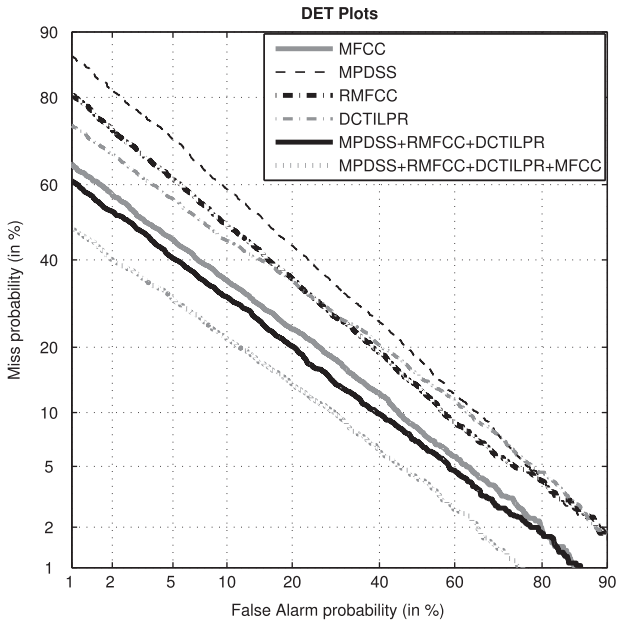

FIG. 3. DET plots obtained using different features and their combination for 2 s duration test data case.

V. SUMMARY AND CONCLUSION

The work explores different attributes of excitation source information for source-based speaker modeling. The three source features, namely, MPDSS, RMFCC, and DCTILPR are found to capture different aspects of source information, which are the periodicity, smoothed spectrum information, and shape of glottal signal, respectively. The nature of different information is visible on the fusion of multiple features which becomes more evident with reduction of test segment lengths. The combination of these three features outperforms the MFCC features for very low amounts of test data cases. Also, the three source features, in combination with MFCC features, provide better performance than their individual combination due to the presence of different attributes of source characteristics. The future work should focus on moving this good performance of source-based modeling to sufficient data cases.

[1]K. A. Lee, A. Larcher, H. Thai, B. Ma, and H. Li, "Joint application of speech and speaker recognition for automation and security in smart home," in *INTERSPEECH* (2011), pp. 3317–3318.
[2]S. Dey, S. Barman, R. K. Bhukya, R. K. Das, B. C., Haris, S. R. Mahadeva Prasanna, and R. Sinha, "Speech biometric based attendance system," in *National Conference on Communications 2014*, IIT Kanpur, February 2014.
[3]N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," IEEE Trans. Audio, Speech, Lang. Process. **19**(4), 788–798 (2011).
[4]A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, "i-vector based speaker recognition on short utterances," in *Interspeech* (2011).
[5]J. Gudnason and M. Brookes, "Voice source cepstrum coefficients for speaker identification," in *Proc. ICASSP* (2008), pp. 4821–4824.
[6]M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," IEEE Trans. Speech Audio Process. **7**(5), 569–586 (1999).
[7]B. Yegnanarayana, K. Sharat Reddy, and S. P. Kishore, "Source and system features for speaker recognition using AANN models," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'01)* (2001), Vol. 1, pp. 409–412.
[8]K. Sri Rama Murty and B. Yegananarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," IEEE Signal Process. Lett. **13**(1), 52–55 (2006).
[9]S. R. Mahadeva Prasanna, C. Gupta, and B. Yegananarayana, "Extraction of speaker specific information from linear prediction residual of speech," Speech Commun. **48**, 1243–1261 (2006).
[10]W. Chan, N. Zheng, and T. Lee, "Discrimination power of vocal source and vocal tract related features for speaker segmentation," IEEE Trans. Audio, Speech, and Language Process. **15**(6), 1884–1892 (2007).
[11]A. G. Ramakrishnan, B. Abhiram, and S. R. Mahadeva Prasanna, "Voice source characterization using pitch synchronous discrete cosine transform for speaker identification," J. Acoust. Soc. Am. **137**, EL469–EL475 (2015).
[12]R. K. Das, B. Abhiram, S. R. Mahadeva Prasanna, and A. G. Ramakrishnan, "Combining source and system information for limited data speaker verification," in *Interspeech* (2014), Singapore.
[13]D. Pati and S. R. Mahadeva Prasanna, "A comparative study of explicit and implicit modelling of subsegmental speaker-specific excitation source information," Sadhana **38**(4), 591–620 (2013).
[14]R. K. Das, D. Pati, and S. R. Mahadeva Prasanna, "Different aspects of source information for limited data speaker verification," in *National Conference on Communications (NCC)* (2015).
[15]K. Sri Rama Murty, V. Boominathan, and K. Vijayan, "Allpass modeling of lp residual for speaker recognition," in *International Conference on Signal Processing and Communications (SPCOM)*, July 2012, pp. 1–5.

[16]D. Pati and S. R. Mahadeva Prasanna, "Speaker information from subband energies of linear prediction residual," in *National Conference on Communications (NCC)*, Jan. 2010, pp. 1–4.
[17]A. P. Prathosh, T. V. Ananthapadmanabha, and A. G. Ramakrishnan, "Epoch extraction based on integrated linear prediction residual using plosion index," IEEE Trans. Audio, Speech, Lang. Process. **21**(12), 2471–2480 (2013).
[18]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," J. Acoust. Soc. Am. **135**(1), 460–471 (2014).
[19]W. Fisher, G. Doddington, and K. Goudie-Marshall, "The DARPA speech recognition research database: Specifications and status," in *Proc. of DARPA Workshop on Speech Recognition* (1986), pp. 93–99.
[20]"The NIST Year 2003 Speaker Recognition Evaluation Plan," NIST, Feb. 2003.