International Journal of Computer and Communication Technology

Volume 4 | Issue 1 Article 5

January 2013

Speaker Recognition using Supra-segmental Level Excitation Information

Debadatta Pati
debadatta@iitg.ernet.in

S. R. Mahadeva Prasanna
prasanna@iitg.ernet.in

Speaker Recognition using Supra-segmental Level Excitation Information

Debadatta Pati, S. R. Mahadeva Prasanna
debadatta@iitg.ernet.in, prasanna@iitg.ernet.in

Abstract**—*****Speaker specific information present in the excitation signal is mostly viewed from sub-segmental, segmental and supra-segmental levels. In this work, the supra-segmental level information is explored for recognizing speakers. Earlier study has shown that, combined use of pitch and epoch strength vectors provides useful supra-segmental information. However, the speaker recognition accuracy achieved by supra-segmental level feature is relatively poor than other levels source information. May be the modulation information present at the supra-segmental level of the excitation signal is not manifested properly in pith and epoch strength vectors. We propose a method to model the supra-segmental level modulation information from residual mel frequency cepstral coefficient (R-MFCC) trajectories. The evidences from R-MFCC trajectories combined with pitch and epoch strength vectors are proposed to represent supra-segmental information. Experimental results show that compared to pitch and epoch strength vectors, the proposed approach provides relatively improved performance. Further, the proposed supra-segmental level information is relatively more complimentary to other levels information.***

Keywords—Sub-segmental, Segmental, Supra-segmental, R-MFCC , Pitch and Epoch.

I. Introduction

Speaker recognition is the task of recognizing speakers based on the information available in their speech signal [1]. The task is either to identify or verify the identity of an unknown speaker. In case of identification, the most likely speaker of the test speech is identified by comparing with the stored reference models. Validating the identity claim by comparing the test speech with the claimed speaker model is the verification task. Depending on the text, text-dependent mode will use speech for the same text and no such constraint in case of text-independent mode. This study considers text-independent speaker identification and verification tasks.

Speaker characteristics in the speech signal is reflected mostly due to the differences in, the dimensions of the vocal tract, characteristics of vocal excitation and learning habits of the speakers [2], [3]. The vocal tract characteristic reflects the physiological structure of the speech production system and relatively more robust and less prone to the mimicry by imposters [4]. Therefore, state-of -the-art ASR system mostly use vocal tract information related features like mel frequency cepstral coefficient (*MFCC*) [5]–[7]. These features mostly characterize the formant structure that depends upon the shape and size of the vocal tract and hence provide good recognition performance. However, the performance of the *MFCC* severely degrades under noisy environment [8]. Thus, where available speech data is of poor quality, like telephonic speech, *MFCC* may not be a good choice. Hence, there is a need for deriving robust features for speaker recognition task. For this, the other component of the speech production system, the excitation source has been explored. The characteristics of the excitation source show both physiological and behavioral aspect of the speaker like pith and intonation, respectively. Thus, information present in the excitation signal relatively contributes more speaker specific information [2], [3]. Further, it was shown that features derived from the excitation signal are relatively more robust and require fewer amounts of data for speaker recognition [9]. Motivated by this, attempts have been made for exploring methods in extracting the speaker-specific information from the excitation signal, [9]–[15]. These attempts mostly try to capture the information attributed due to the vibration of the vocal folds and its strength. Vocal folds vibration depends upon the size of the vocal folds [6]. Since the physiological structure of the vocal folds is quite unique for a speaker, speaker specific characteristics are reflected in the nature of vocal folds vibration. These include, rate of vibration, nature of the periodicity of vibration, strength of the excitation at the instants of opening

Debadatta Pati and S. R. Mahadeva prasanna are with the Indian Institute of Technology, Guwahati - 781039, India.
Corresponding author phone: 0361-2582811, e.mail: debadatta@iitg.ernet.in

and closing and its variation from one instants to other. Unlike vocal tract features, it is difficult to represent all these information together in a single feature. The difficulty may be due to the non-availability of suitable signal processing tools/techniques and also due to the dynamic nature of the excitation.

Existing attempts on exploring the excitation signal mostly view the speaker-specific information from three different levels, called as *sub-segmental*, *segmental* and *supra-segmental* levels. *Sub-segmental* level mostly represents the excitation information present within one pitch period. This includes variation in the amplitude of the vibration within a glottal cycle and its timings like opening and closing instants. *Segmental* level mostly represents the source information present around two to three pitch periods. This includes rate of vocal folds vibration and its strength. *Supra-segmental* level represents the source information present around several pitch periods like pitch, harmonics and excitation strength contours that reflects the learning habits of the speaker. In [13], it was shown that segmental level provides best performance followed by sub-segmental level information. The supra-segmental level information provides the least performance. It may happen that the modulation information present in the supra-segmental level of the excitation signal is not manifested properly in pitch and epoch strength vectors. Due to the variation in the tension and mass lesions in vocal folds, local variations in the energy envelop called as modulation of the excitation signal at the supra-segmental level is also speaker dependant [16], [17]. Since, this information is different from pitch and epoch strength vectors; we may combine them to extract maximum speaker information from the supra-segmental level. Further, we may also benefited by combined use of pitch and epoch strength vectors with modulation together with sub-segmental and segmental levels information for complete representation of the source information. Thus, method needs to be developed to model the supra-segmental level modulation information.

The modulation in the envelope can be better modeled by sub-band level processing. However, due to non-stationary nature, it is difficult to perform direct sub-band processing across several segments of the excitation signal. In this work, alternative approach like residual mel frequency cepstral coefficients ($R - MFCC$) trajectories are used to model the modulation information. The computation of the $R - MFCC$ is similar to the conventional $MFCC$ computation except the use of the linear prediction (LP) residual signal [12], [14]. These cepstral coefficients essentially represent the variation in the strength of excitation at the segmental levels. The variation of the individual cepstral coefficient across several segments may be useful for modeling the supra-segmental level information. In this work we demonstrate the speaker specific nature of the R − MFCC trajectories and then describe a method to model the supra-segmental level modulation information. The significance of the proposed method is experimentally demonstrated from different speaker recognition studies.

The rest of the paper is organized as follows: Section II describes $R - MFCC$ cepstral trajectories and demonstrates its speaker specific nature. Section III describes the proposed cepstral trajectory vectors to model the supra-segmental level modulation information. In this section a combined feature is proposed for best possible way of representing the supra-segmental level information and demonstrates its usefulness for recognizing speakers. In Section IV, we evaluate the performance of the sub-segmental and segmental levels excitation information and made a comparison with the proposed supra-segmental information and finally a combined feature is proposed for complete representation of the excitation information. The performance of the proposed excitation feature is also compared with the conventional vocal tract information. The last section summarizes the present work with a mention on the scope for future work.

II. Speaker Specific Nature of Cepstral Trajectories

The cepstral coefficients derived from the segments essentially represent the oscillation in the sub-band energies. Hence, an individual cepstral trajectory nearly represents the variation in the sub-band energies across several segments. Thus, cepstral trajectories from the excitation signal can be used to model the supra-segmental level modulation information. Earlier studies have shown that cepstral coefficients derived from the mel bank spectrum of the LP residual are more effective in capturing the speaker information [12], [14]. Thus, individual $R - MFCC$ trajectories may be a good choice to model the modulation property of the excitation signal. It should be noted here that $R - MFCC$ feature represent the modulation in the excitation energy over a single segment. On the other hand, the cepstral trajectories represent the oscillation of the sub-band energies across several segments. Thus, speaker information from the cepstral trajectories may be viewed from the supra-segmental level. Usually, in speaker recognition studies, the first 13 coefficients excluding c_0 are used to rerepresent cepstral features. We use lesser (selective) cepstral trajectories to represent the modulation information. The reason for using the selective coefficients is to reduce the computational complexity and also, they all together may not be useful for speaker recognition. To select the cepstral coefficients, statistical *F-ratio* measure that evaluates the effectiveness of the feature coefficients may be used [18]. The

F-ratio of a cepstral coefficient is defined as the ratio of its variance of means and average intra-variance. Variance of means represents how the mean of a cepstral coefficient varies from speaker to speaker. Average intra-variance represents the variation of a cepstral coefficient within a speaker. An ideal cepstral coefficient should have large variance of means and small average intra-variance for discriminating speakers. *F-ratio* has been extensively used for measuring the discriminating ability and also selecting optimized feature for speaker recognition [2]. However, it should be noted here that cepstral coefficients with smaller *F-ratio* value may not be less effective in capturing the speaker information but may be redundant. Thus, when we purposefully want to select some few coefficients from a given set, *F-ratio* measure may be a good measure for selection.

Two separate data sets, called as *Set-1* and *Set-2* are used to select the cepstral coefficients. *Set-1* and S*et-2* consist of 90 speakers collected from NIST-99 and NIST-03 databases, respectively [19], [20]. NIST-99 is used as the representation of clean data collected over land-line and NIST-03 as relatively noisy data, since it is collected over mobile phones. Each speaker has training data of around 2 minutes and the testing data of at least 30 sec. Two sets are considered for robust conclusion. The *R−MFCC* coefficients are computed from 20 msec with a shift of 10 msec segments of the LP residual, using 24 mel filters as described below [12], [14].

Computation of R −MFCC *coefficients:*

The discrete Fourier transform (DFT) of the LP residual e (n) is given by

$$E(k)=\sum_{n=0}^{N\text{-}1} e(n)e^{j\frac{2\pi}{N}}nk,\ \ 0\le k\le N-1$$

(1)

Where, N is the number of points used to compute the DFT. The mel warped spectrum of E(k) is computed as

$$E(m)=\sum_{k=0}^{N\text{-}1}|\,E(k)\,|^2\ H_m(k),\ \ 0\le m\le M-1$$

(2)

Where, Hm (k) is the m^th filter weights and M is the number of filters in the mel filter bank. Then, the cepstral coefficients c(n) are computed from the mel warped spectrum E(m) as

$$c(n)=\sum_{m=0}^{M\text{-}1}\log_{10}(E(m))\cos[n(m-\frac{1}{2})\frac{\pi}{M},\ \ 0\le n\le C-1$$

(3)

Where, C (usually C < M) is the number of cepstral coefficients. The zeroth coefficient, c0 is excluded. since it represents the average log-energy of the residual signal that carries little speaker information.

The *F-ratio* value of 13 individual *R − MFCCs* for both sets is given in the Table I. It can be observed from third and sixth rows of this table that, the first five higher *F-ratio* value coefficients for both sets are from their first seven coefficients.

For example, cepstral trajectories c_{tjr1}, c_{tjr2}, c_{tjr3}, c_{tjr4}, c_{tjr7} in case of *Set-1* and c_{tjr1}, c_{tjr2}, c_{tjr4}, c_{tjr6}, c_{tjr7} for *Set-2*. The common higher *F-ratio* value cepstral coefficients in both cases are c_{tjr1}, c_{tjr2}, c_{tjr4}, c_{tjr6}. Therefore, we consider these four coefficient trajectories to represent the *supra-segmental* level modulation information.

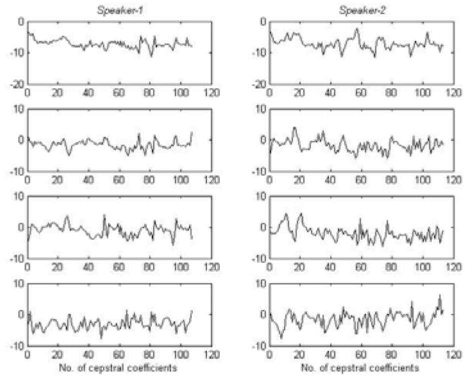

Fig.1. Examples of four R −MFCC (c_{tjr1}, c_{tjr2}, c_{tjr4}, c_{tjr6}) trajectories from two male speakers' common utterance.

Figure 1 shows the example of c_{tjr1}, c_{tjr2}, c_{tjr4}, c_{tjr6} trajectories for *Speaker-1* and *Speaker-2*. In both cases, the text of the speech signal remains same. So that, any variations in the cepstral trajectories may be due to their speaker dependant characteristics. It can be observed that in each case, apart from their duration differences, the variation in the sequence of cepstral trajectories are also significantly different across speakers. This shows that cepstral trajectories are speaker dependant. This is indeed we observe from the speaker identification and verification studies made in the next section.

III. SPEAKER RECOGNITION STUDIES USING CEPSTRAL TRAJECTORY FEATURES

In the previous section we observe that the temporal variations in the sequence of cepstral trajectory samples are different from speaker to speaker. In this section we demonstrate the significance of the information present in cepstral trajectories from different speaker recognition studies. For identification experiment, GMM approach is used to build the speaker models and decision is taken based on the log likelihood ratio (LLR) [7]. The identification experiment is conducted on *Set-1* and *Set-2*. The speaker of the model having highest LLR is identified as the speaker. The identification accuracy is expressed in terms of percentage. In case of verification task, state-of-the-art GMM-universal background model (GMM-UBM) approach is used. The UBM is built from approximately forty hours of speech data collected from 200 speakers (100 males and 100 females from switchboard database) and serves as the imposter model. The Gaussian mixture speaker models are built by adaption of UBM. Only the means are adapted and the weights and variances of the speaker models and the UBM remain same. For a given test utterance, the LLR is given by

$$LLR = \log P(s\lambda_c) - \log P(s\lambda_u)$$

(4)

Where, $P(s\lambda_c)$ and $P(s\lambda_u)$ are the likelihoods given by the claimed speaker model and the UBM, respectively.

The verification experiment is conducted on whole NIST-03 database [20]. The database consists of 356 targets speakers. There are totally 2559 test utterances with duration of 15-45 sec. Each test utterance is tested against 11 hypothesized speakers that include the genuine speaker and 10 imposters. The performance is given by detection error trade-off (DET) based on genuine and imposter LLRs [21]. From DET, equal error rate (EER) is found such that false acceptance rate (FAR) is equal to false rejection rate (FRR). EER is expressed in percentage.

The speaker specific features from cepstral trajectories are represented by sequence of 10 cepstral values with a shift of one value. The sequence of 10 cepstral coefficients that span across 10 segments is considered to capture supra-segmental level information. Every sample shift is considered to get the maximum number of feature vectors.

The feature vectors are derived from each chosen cepstral trajectories and modeled independently. The evidence from individual trajectories is combined at the score level. For combination, linear and logical *OR* combination schemes are used [22], [23]. In case of linear combination, the respective scores are weighted by their performances and combined. For example, the LLR of the combined system, LLRs, is given by the following relation:

$$LLR_s = \sum_{i=1}^{S} \frac{R_i}{\sum_{i=1}^{S} R_i} \times LLR_i$$

(5)

Where, S is the number of systems combined, LLR_i and R_i are LLR and identification performance of the ith system, respectively. In case of verification task, the R_i in equation 5 is replaced by the reciprocal of respective EER and then

TABLE 1
F-ratio VALUE OF R-MFCCS FOR Set-I and Set-II

Cepstral Coefficients	*Set-I*												
	c_{tjr1}	c_{tjr2}	c_{tjr3}	c_{tjr4}	c_{tjr5}	c_{tjr6}	c_{tjr7}	c_{tjr8}	c_{tjr9}	c_{tjr10}	c_{tjr11}	c_{tjr12}	c_{tjr13}
F-ratio	10.23	9.11	9.18	12.57	8.68	8.71	5.73	4.64	4.55	3.07	2.70	2.30	2.96
Order (Descend)	c_{tjr4}	c_{tjr1}	c_{tjr3}	c_{tjr2}	c_{tjr6}	c_{tjr5}	c_{tjr7}	c_{tjr8}	c_{tjr9}	c_{tjr10}	c_{tjr13}	c_{tjr11}	c_{tjr12}
Cepstral Coefficients	*Set-II*												
	c_{tjr1}	c_{tjr2}	c_{tjr3}	c_{tjr4}	c_{tjr5}	c_{tjr6}	c_{tjr7}	c_{tjr8}	c_{tjr9}	c_{tjr10}	c_{tjr11}	c_{tjr12}	c_{tjr13}
F-ratio	7.58	6.65	5.97	6.79	5.54	7.17	6.98	5.37	4.81	5.93	5.75	4.91	3.13
Order (Descend)	c_{tjr1}	c_{tjr6}	c_{tjr7}	c_{tjr4}	c_{tjr2}	c_{tjr3}	c_{tjr10}	c_{tjr11}	c_{tjr5}	c_{tjr8}	c_{tjr12}	c_{tjr9}	c_{tjr13}

the scores of the combined system is computed accordingly.

The simple linear combination of scores with predefined weights may give wrong decision [24]. The potential of the combined system is further verified from the logical *OR* combination. In this scheme we use the ground truth information for decision. In case of identification, if any one system is giving the correct decision, we consider it as a correct decision. In case of verification, the true scores around the mean of the good system are modified based on the information provided by the poor system [13], [14]. The Comb_2 scheme ensures the performance of the good system unaffected and at the same time exploits the evidences from the poor system. The linear and logical *OR* combinations are abbreviated as Comb_1 and Comb_2, respectively.

TABLE II

SPEAKER IDENTIFICATION AND VERIFICATION RESULTS CEPSTRAL TRAJECTORIES, PITCH AND EPOCH STRENGTH VECTORS. *Supra*=$t_0+a_0+C_{tjr}$, REPRESENTS COMPLETE SUPRA-SEGMENTAL LEVEL SOURCE INFORMATION.

Feature		Performance (%)		
		Identification		Verification
		Set-I	Set-II	
c_{tjr1}		40	27	31.39
c_{tjr2}		34	22	31.21
c_{tjr4}		21	12	32.02
c_{tjr6}		26	17	32.83
C_{tjr}	Comb_1	56	37	26.73
	Comb_2	70	41	21.72
t_0+a_0	Comb_1	56	37	26.73
	Comb_2	70	41	21.72
Supra	Comb_1	56	37	26.73
	Comb_2	70	41	21.72

The results of the speaker identification and verification studies using cepstral trajectory feature vectors and their different combinations are given in the Table II. The results show that each cepstral trajectory feature vector contains speaker information. In case of more noisy speech their performance is relatively less. This may be due to the fact that cepstral processing is affected by noise. Further, the evidences provided by cepstral trajectory vectors are different. This can be observed from the confusion patterns of detailed identification results of S*et-1* shown in Fig. 2. In the confusion pattern, principal diagonal represent correct identification and the rest represent miss classification. In each case, the confusion pattern is entirely different. The decisions for both true and false cases are different. This indicates that they reflect different aspect of source information and can be combined to further improve the recognition accuracy.

In this work the combined c_{tjr1}, c_{tjr2}, c_{tjr4}, c_{tjr6} vectors is abbreviated as C_{tjr}. The performance of C_{tjr} vectors for both sets from $Comb_1$ and $Comb_2$ schemes are given in fifth row of the Table II. In case of *Set-1*, the best individual performance, 40% from c_{tjr1} is improved to 56% and 70% for $Comb_1$ and$Comb_2$ schemes, respectively. In case of *Set-2*, the best individual performance, 27% from ctjr1 is improved to 37% and 41% for $Comb_1$ and $Comb_2$ schemes, respectively. Similarly, in case of verification task, the best individual performance, 42.41% from c_{tjr2} is improved to 41.59% and 26.87% for $Comb_1$ and $Comb_2$ schemes, respectively.

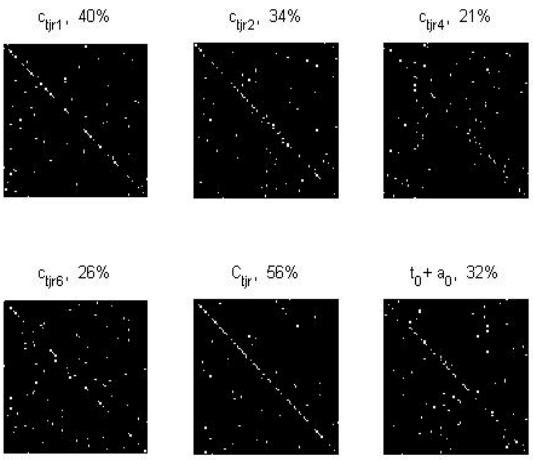

Fig.2. Confusion patterns of cepstral trajectory, combined pitch and epoch strength vectors from identification results of *Set-1*.

The improvement in the recognition accuracy of from C_{tjr} feature indicates that the *supra-segmental* level information present in cepstral trajectories can be effectively represented by combined representation of c_{tjr1}, c_{tjr2}, c_{tjr4}, c_{tjr6} vectors.

A. Complimentary Nature of Cepstral Trajectory with Pitch and Epoch Strength Vectors

To demonstrate the complementary nature of the C_{tjr} feature with pitch and epoch strength vectors, we evaluate the speaker recognition performance of pitch and epoch vectors as suggested in [13], [14]. Pitch and epoch strength values are computed by using event based fundamental frequency estimation method [13], [25], [26]. The detail computational procedure of this approach is given in [13]. Pitch and epoch strength vectors called as, t_0 and a_0 vectors are represented by every ten pitch and epoch strength values with a shift of one value, respectively [13]. The combined use of pitch and epoch strength vectors is abbreviated as $t_0 + a_0$ vectors.

The recognition performance of $t_0 + a_0$ vectors is given in the seventh column of the Table II. It can be observed that the performance of the $t_0 + a_0$ vectors is relatively poor than C_{tjr}. This may due to large intra-speaker variability of $t_0 + a_0$ and also due to text-independent mode of operation. However, from the confusion patterns of $t_0 + a_0$ vectors shown in Fig. 2, it can be observed that the evidence provided by $t_0 + a_0$ and C_{tjr} is different and hence may be combined for effective representation of the *supra-segmental* level information. In this work, the combined evidences from $t_0 + a_0$ and C_{tjr} are represented by *Supra*. The results of the *Supra* feature are given in the eighth row of the Table II. For both identification and verification tasks the best performance provided by C_{tjr} vector is further improved when combined with $t_0 + a_0$ vectors. Further, for more noisy speech the performance of $t_0 + a_0$ and C_{tjr} feature vectors is affected. For example, in case of identification task, the performance of $t_0 + a_0$ and C_{tjr} feature vectors degrades by 59% and 34%, respectively. However, the corresponding degradation in case of *Supra* feature is relatively less, around 32%, as against 59% in case of $t_0 + a_0$ vectors. It shows that $t_0 + a_{0\,+}\, C_{tjr}$ representation is relatively more robust against noise. Thus, we conclude that combined representation of cepstral trajectory, pitch and epoch strength vectors may be the best possible way of representing the *supra-segmental* level information.

IV. SPEAKER SPECIFIC EXCITATION INFORMATION

The speaker information from the excitation signal is modeled from *sub-segmental*, *segmental* and *supra-segmental* levels. In this section, we evaluate the speaker recognition performance of sub-*segmental* and *segmental* levels excitation information then made a comparison with *supra-segmental* level feature. The evidences from all three levels are combined to represent the complete excitation information. Finally, a comparison is made between the vocal tract and excitation information for speaker recognition task.

A. Speaker Recognition using Sub-segmental Information

In [13], the LP residual and its analytic representation are processed in blocks of 5 msec with a shift of 2.5 msec to model the sub-segmental level information. It was shown that the LP residual processed in *sub-segmental* blocks provide useful information which is relatively more complimentary to other levels source information. Therefore, in this work the LP residual is processed in blocks of 5 msec with a shift of 2,5 msec to model the *sub-segmental* level information. The LP residual *sub-segmental* blocks are called as *Sub* features. It should be noted here that the LP residual is directly processed to obtain the *Sub* feature and provides lossless information. The speaker recognition results of the *Sub* feature is given in the first row of the Table III. The identification accuracy achieved by *Sub* feature for *Set−1* and *Set−2* is 64% and 57%, respectively. The relative degradation in the performance from *Set − 1* to *Set − 2* is around 10%. In case of the verification task, the EER achieved is 23.75%. Due to lossless representation of the information, the *Sub* feature provides good recognition accuracy.

TABLE III

SPEAKER RECOGNITION RESULTS OF EXCITATION AND VOCAL TRACT FEATURES. *Src=Sub+Seg+Supra*, REPRESENTS THE COMPLETE EXCITATION INFORMATION.

Feature		Performance (%)		
		Identification		Verification
		Set-I	Set-II	
Sub		64	57	23.75
Seg	Comb$_1$	82	51	16.39
	Comb$_2$	88	61	12.69
Supra	Comb$_1$	64	43	25.15
	Comb$_2$	77	53	20.09
Src	Comb$_1$	83	61	14.13
	Comb$_2$	97	72	9.62
MFCC		87	66	7.27
Src+MFCC	Comb$_1$	91	66	6.56
	Comb$_2$	98	82	7.27

B. Speaker recognition using Segmental Information

The *segmental* level information is extracted by processing the vocal excitation signal in blocks of two to three pitch periods. Since the speech signal is assumed to be stationary at the *segmental* level, the vocal excitation signal is processed both in time and frequency domains to model the *segmental* level information. In [14], a comparison is made on processing the LP residual in time and frequency domains for modeling the *segmental* level information. It was shown that with a small compromise in recognition performance, frequency domain processing provides compact way of representing the *segmental* level information. In frequency domain, the *segmental* level information is captured from the parameterizations of the LP residual sub-band magnitude spectra. The purpose of using the sub-band spectrum is that, obtaining a global value from the spectrum may not likely to show good speaker-dependant characteristics. In [14], cepstral analysis and spectral flatness measure were made on residual sub-band spectra to capture the energy and periodicity information, respectively. It was shown that *R−MFCC* and mel power difference of spectrum in sub-band (*M − PDSS*) feature vectors derived from mel warped spectrum well represent the energy and periodicity information of the excitation signal, respectively. The combined evidences from *R−MFCC* and *M−PDSS* features well represent the *segmental* level excitation information. Thus, in this work the combination of *R−MFCC* and *M −PDSS*, called as *Seg* is used to represent the segmental level excitation information.

The procedure to compute *R−MFCC* feature is described in Section II. The first 13 coefficients excluding c_0 are used as *R − MFCC* feature. The cepstral mean subtraction is performed to eliminate the channel effect [27]. The procedure to compute *M − PDSS* is given below [14].

Computation of M − PDSS *feature:*

The *M −PDSS* feature is computed from spectral flatness measure of the power differences in mel sub-band spectrum. The spectral flatness essentially represents the periodicity nature of the spectrum. For example, more flat spectrum is less periodic. The spectral flatness is measured as the ratio of the geometric mean to the arithmetic mean of the spectral samples. In [14], spectral flatness measured from 20 mel sub-band spectra is used as the components of *M − PDSS* feature vector. The mathematical expression for computation of M –PDSS feature components v (m) is given below [14], [28].

$$v(m) = 1 - \frac{\left[\prod_{k=l_m}^{h_m} P_m(k)\right]^{\frac{1}{N_m}}}{\frac{1}{N_m}\sum_{k=l_m}^{h_m} P_m(k)} \qquad (6)$$

Where, $P_m(k) = \left[E(k)H_m(k)\right]^2$, is the residual mel sub-band power spectrum, l_m, h_m are the lower and upper limits of the sample frequency points and $N_m = h_m - l_m + 1$ is the sample number of frequency points of the m^th filter. Each component of the v (m) is used to represent *M − PDSS*.

The speaker recognition results of the *Seg* feature is given in third row of the Table III. The maximum benefit we can achieve for *Set−1* and *Set−2* is 88% and 61%, respectively. The relative degradation in the performance from *Set* – 1 to *Set* − *2* is around 30%. In case of verification task, the minimum EER achieved is 12.69%. The performance achieved by *Seg* indicates the presence of the speaker information in the *segmental* level of the excitation signal. The performance of the *Supra* feature is also given in third row of the Table III for comparison. By comparing the results from *Sub*, *Seg* and *Supra* features it can be observed that, the *segmental* level information provides best performance followed by *sub-segmental* level information. The *supra-segmental* level information still provides least performance. Further, the relative degradation in the performance due to noise is more in case of *supra-segmental* level information. It may happen that *supra-segmental* level excitation information has large intra-speaker variability.. However, one should not be confused with the usefulness of the *supra-segmental* level information. Because, this information is different from *sub-segmental* and *segmenta*l levels [13]. By combining evidences from *sub-segmental*, *segmental* and *supra-segmental* levels, we may achieve improved recognition accuracy. This is indeed we observe from the speaker recognition results given in fourth column of the Table III. In all cases the performance of individual levels excitation information is improved. Hence, it is suggested that the combined use of evidences from *Sub*, *Seg* and *Supra* features may be the best possible way of representing the complete source information.

C. Speaker Recognition using Vocal Tract Information

We also verify the potential of the proposed source feature (*Src*) with the conventional vocal tract information (*MFCC*). For this, we evaluate the performance of the *MFCC* features.

The *MFCC* feature is computed from 20 msec with a shift of 10 msec segment of speech using 24 overlapping mel filters [5]–[7]. The set of first 13 *MFCC*s excluding c_0 are used to represent *MFCC* feature. The experimental conditions remain same for fair comparison.

The performance of the *MFCC* feature is given in fifth row of the Table III. For both identification and verification tasks, the individual performance of the *MFCC* feature is significantly better than the proposed *Src* feature. However, it is interesting to note that, if suitable combination technique is available, then one can also able to achieve better identification accuracy from the source feature itself. For example, in case of Comb_2 scheme, the identification accuracy achieved by *Src* for *Set-1* and *Set-2* is 97% and 72%, as against 87%

and 66% in case of *MFCC* feature, respectively. Further, since *MFCC* and *Src* represent two different aspect of the speaker information present in the speech signal, they may be combined together to further improve the recognition accuracy. The results of combined *MFCC* and *Src* are given in the last row of the Table III. The maximum benefit we achieve in case of combining the vocal tract and excitation information is better than individual *MFCC* feature. This shows that the source provides complimentary evidence to vocal tract information to further improve the recognition accuracy.

V. CONCLUSION

The objective of this work was to experimentally evaluate the potential of the *supra-segmental* level excitation information for recognizing speakers. We explore the excitation signal at the *supra-segmental* level and propose $R - MFCC$ trajectory vectors to model the modulation information. From different speaker recognition studies we observed that, the proposed cepstral trajectory vectors well model the modulation information and provides complimentary information to pitch and epoch strength vectors. The combined evidence from cepstral trajectory together with pitch and epoch strength vectors (*Supra*) provides improved recognition accuracy and hence may be the best possible way of representing the *supra-segmental* level excitation information. We also evaluate the effectiveness of the *sub-segmental* and *segmental* levels information. We found that *segmental* level provides best performance followed by *sub-segmental* level information. The *supra-segmental* level information provides least performance. However, combining the evidences from all the levels (*Src*), the performance of the *segmental* level information is further improved. Hence, it is suggested that the proposed *Src* feature may be the best possible way of representing the complete excitation information for speaker recognition. Further, the performance of *Src* is relatively poor than the conventional vocal tract information (*MFCC*). However, the performance of the *MFCC* feature is further improved by using complimentary information from *Src*.

The recognition accuracy achieved by *Supra* is still poor than *sub-segmental* and *segmental* levels information. It is also observed that performance of *Src* is still inferior compared to *MFCC* in real time application. This may be due to the method employed for extraction of the excitation information. For example, there is no parameterizations is involved in modeling the *sub-segmental* level information. Any parameterizations like modeling the glottal flow may provide relatively more effective information [29]. The evidence from the parameterizations of the *sub-segmental* level information together with other levels information may further improve the recognition accuracy from excitation prospective. Further, the performance of the combined system is also depends upon the combination scheme employed. New combination technique needs to be developed to exploit the same. For this, amount of representative and discriminating information captured by each feature measurements may be useful [30].

ACKNOWLEDGMENT

This work is supported by the UK-INDIA Education and Research Initiative (UKIERI) project titled ”study of source features for speech synthesis and speaker recognition” between IIT Guwahati, IIIT Hyderabad and CSTR, University of Edinburgh, UK.

REFERENCES

[1] J. P. Campbell, Jr., “Speaker recognition: A tutorial,” Proc. IEEE, vol. 85, no. 9, pp. 1437–1462, Sept. 1997.

[2] J. J. Wolf, “Efficient acoustic parameters for speaker recognition,” J. Acoust. Soc. .Amer. vol. 51, no. 2, pp. 2044–2055, 1972.

[3] B. S. Atal, “Automatic speaker recognition based on pitch contours,” J. Acoust. Soc. Amer., vol. 52, no. 6, pp. 1687–1697, 1972.

[4] D. O. Shaughnessy, “Speaker recognition,” vol. 3, Oct. 1986, pp. 4–17.

[5] S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Trans. Acoust. Speech and Signal Process., vol. 28, no. 28, pp. 357–366, Aug. 1980.

[6] J. R. Deller jr., John H. L. Hansen and J G Proakis, Discrete-Time Processing of Speech Signal, 2nd ed. New York: IEEE Press, 2000.

[7] D. A. Reynolds and R. C. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models,” IEEE Trans. Speech and Audio Process., vol. 3, no. 1, pp. 4–17, Jan. 1995.

[8] D. A. Reynolds, “Experimental evaluation of features for robust speaker identification,” IEEE Trans. Speech Audio Process., vol. 2, no. 4, pp. 639–643, Oct. 1994.

[9] S. R. M. Prasanna, C. S. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," Speech Commun., vol. 48, pp. 1243–1261, Jun. 2006.

[10] P. Thevenaz and H. Hugli, "Usefulness of the LPC-residue in text independent speaker verification," Speech Commun., vol. 17, pp. 145–157, Aug. 1995.

[11] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," IEEE Signal Process. Lett., vol. 13, no. 1, pp. 52–55, Jan. 2006.

[12] D. Pati and S. R. M. Prasanna, "Speaker information from sub-band energies of linear prediction residual," in Proc. NCC 2010, pp. 1–4.

[13] ——, "Subsegmental, segmental and suprasegmental processing of linear prediction residual for speaker information," Accepted for publication, Int. J. of Speech Technology, Springer.

[14] ——, "Processing of linear prediction residual in spectral and cepstral domains for speaker information," Communicated to Int. J. of Speech Processing, Springer.

[15] ——, "Speaker recognition from excitation source prospective," IETE Technical Review, vol. 27, no. 2, pp. 138–157, Mar 2010.

[16] M. Farrus and J. Hernando, "Using jitter and shimmer in speaker verification," IET signal proc., vol. 3, no. 4, pp. 247–257, Nov 2009.

[17] Kreiman J. and Gerrat B. R., "Perception of aperiodicity in pathological voice," J. Acoust. Soc. Amer., vol. 117, pp. 2201–2211, 2005.

[18] S. Pruzansky and M. V. Mathews, "Talker-Recognition procedure based on Analysis of Variance," J. Acoust. Soc. Amer., vol. 36, no. 11, pp. 2041–2047, 1964.

[19] M. Przybocky and A. Martin, "The NIST-1999 speaker recognition evaluation- An overview," Digital signal processing, vol. 10, pp. 1–18, 2000.

[20] "NIST speaker recognition evaluation plan," in Proc. NIST Speaker Recognition Workshop, College Park, MD, 2003.

[21] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in Proc. Eur. Conf. on Speech Communication Technology, Rhodes, Greece, vol. 4, 1997, pp. 1895–1898.

[22] D. J. Mashao and M. Skosan, "Combining classifier decisions for robust speaker identification," Pattern Recognition, vol. 39, pp. 147–155, Jan. 2006.

[23] J. J. Hall and S. N. Srihari, "Decision combination in multiple classifier systems," IEEE Trans. Patt. Anal. and Mach. intell., vol. 16, pp. 66–75, Jan. 1994.

[24] N. Zheng, T. Lee, and P. C. Ching, "Integration of complimentary acoustic features for speaker recognition," IEEE signal process. Lett., vol. 14, no. 3, pp. 181–184, March 2007.

[25] K. S. R. Murthy and B. Yegnanarayana, "Epoch extraction from speech signal," IEEE Trans. Audio Speech and Language Process., vol. 16, no. 8, pp. 1602–1613, November 2008.

[26] B. Yegnenarayana and K. S. R. Murthy, "Event based instantaneous fundamental frequency estimation from speech signals," IEEE Trans. Audio, Speech and Language Process., vol. 17, no. 4, pp. 614–624, May 2009.

[27] S. Furui, "Cepstral analysis technique for automatic speaker verification," IEEE Trans. Acoust. Speech, and Signal Process., vol. 29, no. 2, pp. 254–272, Apr. 1981.

[28] S. Hayakawa, K. Takeda and F. Itakura, "Speaker identification using harmonic structure of lp-residual spectrum," Biometric personal Aunthentification, Lecture notes, Springer, Berlin, vol. 1206, pp. 253–260, 1997.

[29] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modelling of glottal flow derivative waveform with application to speaker identification," IEEE Trans. Speech and Audio Process., vol. 7, no. 5, pp. 569–586, Sep. 1999.

[30] R.Padmanabhan and H. A. Murthy, "Acoustic feature diversity and speaker verification," in INTERSPEECH 2010, Sept., Makuhari, Chiba, Japan, 2010, pp. 2010–2013.

Debadatta Pati was born in India in 1971. He received the B.E. degree in industrial electronics from college of engineering, Osmanabad, Aurangabad University, India, in 1996 and the M.E. degree in Electronics and Telecommunication from University College of Engineering, Burla, India, in 2003. He is currently pursuing the Ph.D. degree in Electronics and Electrical Engineering at Indian Institute of Technology Guwahati, India.

His research interests are in speech and speaker recognition.

S. R. Mahadeva Prasanna was born in India in 1971. He received the B.E. degree in electronics engineering from Sri Siddartha Institute of Technology, Bangalore University, Bangalore, India, in 1994, the M.Tech. degree in industrial electronics from the National Institute of Technology, Surathkal, India, in 1997, and the Ph.D. degree in computer science and engineering from the Indian Institute of Technology Madras, Chennai, India, in 2004. He is currently an Associate Professor in the Department of Electronics and Electrical Engineering, Indian Institute of Technology, Guwahati.

His research interests are in speech and signal processing, application of AI tools for pattern recognition tasks in speech, and signal processing.