Int J Speech Technol (2011) 14: 49–64
DOI 10.1007/s10772-010-9087-8

Subsegmental, segmental and suprasegmental processing of linear prediction residual for speaker information

Debadatta Pati · S.R. Mahadeva Prasanna

Received: 27 August 2010 / Accepted: 30 November 2010 / Published online: 21 December 2010

Abstract This work processes linear prediction (LP) residual in the time domain at three different levels, extracts speaker information, and demonstrates their significance and also different nature for text-independent speaker recognition. The subsegmental analysis considers LP residual in blocks of 5 msec with shift of 2.5 msec to extract speaker information. The segmental analysis extracts speaker information by processing in blocks of 20 msec with shift of 2.5 msec. The suprasegmental speaker information is extracted by viewing in blocks of 250 msec with shift of 6.25 msec. The speaker identification and verification studies performed using NIST-99 and NIST-03 databases demonstrate that the segmental analysis provides best performance followed by subsegmental analysis. The suprasegmental analysis gives the least performance. However, the evidences from all the three levels of processing seem to be different and combine well to provide improved performance, demonstrating different speaker information captured at each level of processing. Finally, the combined evidence from all the three levels of processing together with vocal tract information further improves the speaker recognition performance.

Keywords Subsegmental · Segmental · Suprasegmental · LP residual · Source information · Speaker recognition

D. Pati · S.R.M. Prasanna (✉)
Department of Electronics and Communication Engineering,
Indian Institute of Technology Guwahati, Guwahati 781039, India
e-mail: prasanna@iitg.ernet.in

D. Pati
e-mail: debadatta@iitg.ernet.in

1 Introduction

The speaker information in the speech signal is attributed to the physiological and behavioral aspects of a person (Atal 1972). The physiological aspects are due to the vocal tract and excitation source that involved in the production (Wolf 1972). The behavioral aspect involves factors like speaking rate, accent etc. (Wolf 1972). The shape, size and the dynamics associated with the vocal tract contribute to the speaker characteristics. On the similar lines, the shape, size and the dynamics associated with the vocal folds contribute to the speaker characteristics. State of the art speaker recognition systems mostly use vocal tract related speaker information represented by the spectral or cepstral features like linear prediction cepstral coefficients (LPCC) or mel frequency cepstral coefficients (MFCC) (Furui 1981; Davis and Mermelstein 1980; Reynolds and Rose 1995). These features provide good recognition performance. The reason may be that, they nearly represent complete vocal tract information. However, under degraded conditions, the spectral or cepstral features give poor performance (Reynolds 1994). Hence their is a need for deriving robust features for speaker recognition. The speech production and perception theory indicate that source contains speaker information and also may be relatively more robust due to its impulsive nature (Mary and Yegnanarayana 2008). Motivated by this, attempts have been made in exploring methods for modeling the speaker information from the source (Atal 1972; Thevenaz and Hugli 1995; Hayakawa et al. 1997; Yegnanarayana et al. 2001; Farrus and Hernando 2009; Sonmez et al. 1998; Plumpe et al. 1999; Prasanna et al. 2006; Pati and Prasanna 2010; Zheng et al. 2007). These attempts demonstrate that source indeed contains significant speaker information. However, the recognition performance is not at par with the vocal tract information. The reason may be that the methods employed in representing the source information may not model all aspects of speaker information. By that we mean, LPCC or MFCC captures the formants and their bandwidth information characterizing the vocal tract completely, but pitch is only one aspect of speaker information due to source. Thus to further improve the performance of source features, methods need to be developed that tries to capture the complete source information. For this, the source signal needs to be derived from the speech. Earlier studies have shown that for proper linear prediction (LP) order (for example 8–20 in case of speech sampled at 8 kHz), the LP residual can be used as the best approximation of the source signal (Prasanna et al. 2006; Plumpe et al. 1999). The LP residual can be processed in time, frequency, cepstral or time-frequency domains to extract and model speaker information (Yegnanarayana et al. 2001; Prasanna et al. 2006; Hayakawa et al. 1997; Thevenaz and Hugli 1995). Processing the LP residual in time domain has the advantage that the artifacts of digital signal processing like block processing or windowing effect that creep in other domains of processing like frequency will be negligible. Thus processing LP residual in time domain is expected to model the speaker information in the best possible manner.

The existing attempts for processing LP residual in the time domain may be broadly grouped into subsegmental, segmental and suprasegmental levels. In Atal (1972), the temporal variation of pitch termed as pitch contour is used as the speaker information. The pitch contour spans over several segments and hence may be viewed as suprasegmental processing. Attempts have been made to use pitch as an additional parameter along with vocal tract features like MFCC at frame levels, which seem to improve the performance (Huang et al. 2008; Ezzaidi and Rouat 2004; Yegnanarayana et al. 2005). In these studies, pitch information is extracted for each segmental frame and appended to MFCC and hence may be treated under segmental processing. In Yegnanarayana et al. (2001), Prasanna et al. (2006), information from the LP residual is processed in blocks of 5 msec with one sample shift. In Murty et al. (2004), Murty and Yegnanarayana (2006) also, the LP residual phase computed from the analytic signal representation of the LP residual is processed in blocks of 5 msec with one sample shift. In these studies, the speaker information is implicitly captured using the auto associative neural network (AANN) models and demonstrated presence of speaker information. Since the block length is less than 20 msec, these studies may be viewed under subsegmental processing. All these studies are independent and use different approaches for extracting and modeling speaker information. An unified framework may be evolved where a given LP residual is processed at subsegmental, segmental and suprasegmental levels using a single signal processing approach and use the same to study the level of speaker information present at each level and also their differences. The present work proposes one such approach and hence it is termed as subsegmental, segmental and suprasegmental processing of LP residual for speaker information.

The present work processes the LP residual in blocks of 5 msec with 2.5 msec shift for subsegmental, 20 msec with 2.5 msec shift for segmental and 250 msec with 6.25 msec shift for suprasegmental, levels of processing. The 5 msec blocks of LP residual sample sequences in the time domain are used as feature vectors for modeling speaker information by Gaussian mixture modeling (GMM) technique to generate subsegmental speaker models. The 20 msec blocks of LP residual samples are first decimated by a factor of 4 to reduce its dimensionality and also to eliminate the information that has been modeled at the subsegmental level. The decimated LP residual sample sequences are modeled by GMM to generate segmental speaker models. The 250 msec blocks of LP residual samples are first decimated by a factor of 50 to reduce its dimensionality and also to eliminate the information that have been modeled both at the subsegmental and segmental levels. The decimated LP residual sample sequences are modeled by GMM to generate suprasegmental speaker models. All these models are independently tested using respective blocks of LP residual extracted from the test signals to evaluate the amount of speaker information present at each level. Finally the combination of evidences from all the three levels is made to observe their different nature of speaker information. The potential of combined source information is demonstrated by comparing and also combining its performance with a speaker recognition system using vocal tract feature. The earlier attempts of modeling speaker information from the LP residual in time domain use the AANN models for exploiting sequence information (Yegnanarayana et al. 2001; Prasanna et al. 2006). In the present work an alternative view is taken for the LP residual samples. The LP residual signal is like a random noise sequence, except for the pitch information. If we treat the residual signal as random noise, then the distribution of the samples will be Gaussian. Since the LP residual deviates from random noise due to pitch information, to that extent the distribution of the residual samples may be non-Gaussian in nature. However, this can be handled with the help of the GMM. Hence the motivation for using GMM for speaker modeling from LP residual.

The rest of the paper is organized as follows: Sect. 2 describes the proposed subsegmental, segmental and suprasegmental analysis of LP residual approach for modeling speaker information from the LP residual. This section will also describe the speaker recognition studies that have been performed using the proposed approach. Section 3 describes an alternative approach for subsegmental, segmental and suprasegmental analysis using the analytic signal concept and demonstrates its significance in modeling speaker information. Section 4 describes an alternative approach only for modeling suprasegmental information using the concept of instantaneous pitch. The last section summarizes the present work with a mention on the scope for future work.

2 Processing of LP residual in time domain

In LP model of speech production, each sample of speech is predicted as a linear combination of the past p samples, where p represents the order of prediction (Makhoul 1975). If $s(n)$ is the present sample, then it is predicted by the past p samples as

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \tag{1}$$

where, $a_k s$ are the LP coefficients (LPCs) computed by minimizing the mean square prediction error. The error between the actual and the predicted sample value is called as the prediction error or LP residual and is given by

$$r(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k) \tag{2}$$

The LP residual $r(n)$ is obtained by passing the speech signal through an inverse filter $A(z)$ given by

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k} \tag{3}$$

The predicted samples $\hat{s}(n)$ model the vocal tract information in terms of (LPCs) (Atal 1974). The suppression of this information from the speech signal $s(n)$ that results in the LP residual $r(n)$ is therefore mostly contains information about the source. So the source signal can be approximated by the LP residual. The representation of source information in the LP residual depends upon the order of prediction. In Prasanna et al. (2006), it was shown that for a speech signal sampled at 8 kHz, the LP residual extracted using LP order in the range 8–20 best represents the speaker-specific source information. In this study, LP residual computed using 10th order LP analysis followed by inverse filtering the speech signal sampled at 8 kHz is used as the source signal. Example of the speech and LP residual signal is shown in Fig. 1(a) and (b), respectively. The instants around the peaks in the LP residual are termed as epochs (Ananthapadmanabha and Yegnanarayana 1979; Murthy and Yegnanarayana 2008). Significant speaker-specific source information is present around the region of the epochs (Murty and Yegnanarayana 2006). This includes the strength and rate of occurrence of the epochs and their temporal variations across several glottal cycles.

In this section we describe the methods employed in processing the LP residual to extract speaker information. In extracting such information we consider subsegmental, segmental and suprasegmental level processing of LP residual. In subsegmental processing, features are derived to represent the speaker information present mostly within one glottal cycle. In segmental processing, features are derived to represent the speaker information mostly related to pitch and energy of the excitation present across 2–3 glottal cycles. In suprasegmental level processing, features are derived to represent the prosodic aspects of the speaker present across about 25–50 glottal cycles.

GMM approach is used to build the speaker models (Reynolds and Rose 1995; Reynolds 1995). Decision is taken based on the log-likelihood ratio (LLR). Recognition experiments are conducted for both identification and verification tasks. In case of identification, the speaker of the model having highest LLR is identified as the speaker.

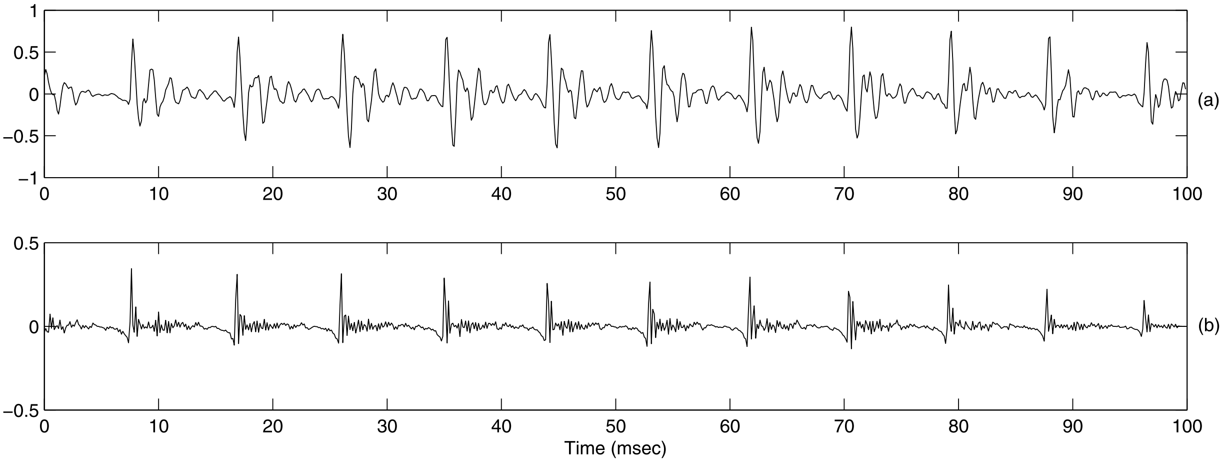

Fig. 1 Speech and LP residual. (**a**) Voiced segment of speech. (**b**) Corresponding 10th order LP residual

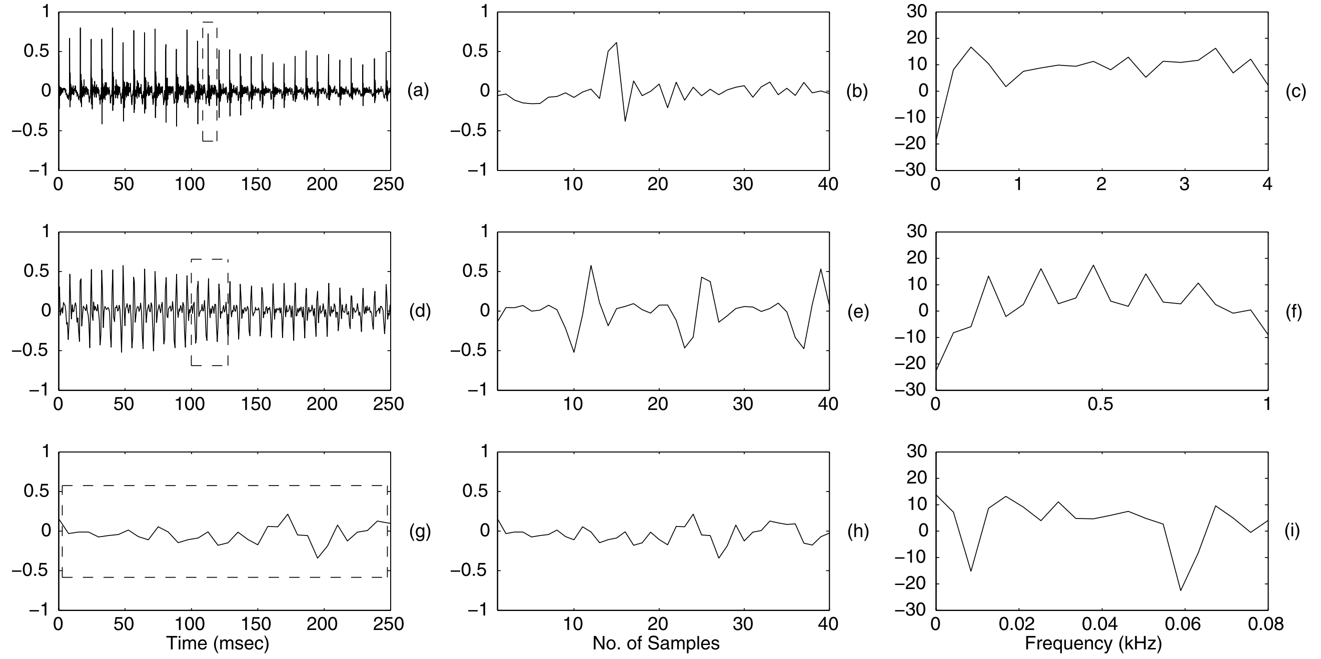

Fig. 2 Temporal sequences and their spectra from subsegmental, segmental and suprasegmental processing of LP residual. (**a**) LP residual. (**b**)–(**c**) Subsegmental sequence and its spectrum, respectively. (**d**) LP residual decimated by a factor 4. (**e**)–(**f**) Segmental sequence and its spectrum, respectively. (**g**) LP residual decimated by a factor 50. (**i**)–(**j**) Suprasegmental sequence and its spectrum, respectively. The *dotted box* in (**a**), (**d**) and (**g**) represents the nature of the LP residual that will be processed at subsegmental, segmental and suprasegmental levels, respectively

The experiment is conducted on two subsets of NIST-99 and NIST-03 database (Przybocky and Martin 2000; Nist speaker recognition evaluation plan 2003). NIST-99 is used as representation of clean data collected over land line and NIST-03 as relatively noisy data, since it is collected over mobile phones. Each subset consists of 90 speakers (48 males and 42 females) having matched condition and testing data of at least 30 sec. The performance is expressed in terms of identification accuracy expressed in percentage. The speaker verification study is conducted on the whole NIST-03 database. The performance is given by detection error tradeoff (DET) curve based on genuine and imposter LLRs (Martin et al. 1997). From the DET curve, equal error rate (EER) is found by choosing a threshold such that false acceptance rate (FAR) is equal to false rejection rate (FRR). EER is expressed in percentage. All the speaker recognition studies are performed for text-independent case, where there is no restriction on the type of text used for recording the speech during training and testing.

2.1 Speaker information from subsegmental processing of LP residual

At the subsegmental level, speaker information present mostly within one glottal cycle is modeled. This information may be attributed to the activity like opening and closing glottal characteristics. To model this information, the LP residual is blocked into frames of 5 msec with a shift of 2.5 msec. For 5 msec at 8 kHz, the frames have 40 samples. One such frame is shown in the Fig. 2(b) and its spectrum is shown in Fig. 2(c). The largest amplitude of the samples of the vector indicate the strength of excitation. The samples in the vector represent the sequence information of glottal cycle. Since these frames are obtained from the LP residual sampled at 8 kHz, they will have excitation source information present as the fine variations represented by frequency components up to 4 kHz. These frames of LP residual samples in the time domain are used as the feature vectors to represent the speaker information at the subsegmental level and used for speaker recognition experiments. The nature of the LP residual signal that will be processed at the subsegmental level is the one shown in Fig. 2(a). This is nothing but the original LP residual.

The results of identification and verification experiments are given in the second column of the Tables 1 and 2, respectively. In these tables the performance of the vocal tract based features namely, MFCC is also given. It is to be cautioned at this stage that the speaker verification system using MFCC and GMM is a baseline system without any normalization techniques. Hence the performance itself is poor compared to the state-of-the-art on NIST-03 (Nist speaker recognition evaluation plan 2003). Since our objective is only relative comparison among source and vocal tract features, we have settled to the baseline system. The results

Table 1 Speaker identification performance (in %) of subsegmental (*Sub*), segmental (*Seg*), suprasegmental (*Supra*) and spectral (*MFCC*) information for two subsets of 90 speakers. $Src-1$ represents combination of *Sub* and *Seg* source information. $Src-2$ represents combination of *Sub*, *Seg* and *Supra* source information

Database	*Sub*	*Seg*	*Supra*	$Src-1$		$Src-2$		*MFCC*	$Src-2+MFCC$	
				$Comb-1$	$Comb-2$	$Comb-1$	$Comb-2$		$Comb-1$	$Comb-2$
NIST-99	64	60	31	64	71	68	76	87	84	96
NIST-03	57	58	13	60	67	60	67	66	70	79
Relative Degradation	11	3	58	6	6	12	12	24	17	18

Table 2 Speaker verification performance of *Sub*, *Seg*, *Supra*, $Src-1$, $Src-2$ and *MFCC* information for whole NIST-03 database

Database	*Sub*	*Seg*	*Supra*	$Src-1$		$Src-2$		*MFCC*	$Src-2+MFCC$	
				$Comb-1$	$Comb-2$	$Comb-1$	$Comb-2$		$Comb-1$	$Comb-2$
NIST-03	41.01	26.96	44.49	32.02	23.21	32.25	21.22	22.94	27.78	17.43

show that subsegmental features provide good performance and hence contain speaker information. However the performance is comparatively poorer than the vocal tract features. The reason may be that the subsegmental features contain only one aspect of source information. The performance can be improved by using additional information from segmental and suprasegmental levels.

It is interesting to observe that the performance of both subsegmental source information and vocal tract features degrade in case of NIST-03, as expected. However, the amount of degradation in the performance is relatively less in case of subsegmental source information, about 11%, as against to 24% in case of vocal tract features. This demonstrates the relative robustness of source information present at the subsegmental level.

2.2 Speaker information from segmental processing of LP residual

At the segmental level, speaker information present in two to three glottal cycles is modeled. This information may be attributed mostly to pitch and energy. Speaker information represented by variations within a glottal cycle have already been modeled by subsegmental analysis. In segmental level processing of LP residual, other information that can be observed at the segmental level needs to be emphasized. For this we propose to decimate the LP residual by a factor 4 so that the sampling rate becomes 2 kHz and we may have source information up to 1 kHz. The decimated LP residual is shown in Fig. 2(d). Even after decimation, the dominant speaker information at the segmental level, that is, pitch and energy information, still can be preserved. Moreover, in segmental level processing, LP residual frames of 20 msec duration are used as the feature vectors. For 20 msec at 8 kHz, the feature vectors with 160 samples is of very large dimension for building the models. By decimating the LP residual by a factor 4, the dimension of the feature vectors is reduced to 40 samples per vector which is equal to the subsegmental feature vectors length. Since the LP residual is decimated by a factor 4, we prefer to compute the feature vectors for every 2.5 msec frame shift so that the number of feature vectors will remain same as the subsegmental features. One such feature vector derived from the decimated LP residual is shown in Fig. 2(e). It contains mainly the pitch and energy information. The fine variations within the glottal cycle are suppressed by smoothing. Similar observation can also be made from the spectrum of the feature vector shown Fig. 2(f). The periodicity and the amplitude of the spectrum clearly represent the pitch and energy information. This observation indicate that segmental feature vectors reflect different aspect of source information compared to subsegmental feature vectors. This will also be confirmed from the comparison study in Sect. 2.4.

The effectiveness of these features are evaluated from the identification and verification experiments. The results are given in the third column of the Table 1 and Table 2, respectively. The high performance show that segmental features contain good speaker information, even better than those contained at the subsegmental level. This shows that the pitch and energy may be dominating speaker-specific source information. Further, the recognition performance is comparatively poor than vocal tract features. The same reason of incomplete representation of speaker information may be attributed. The segmental source features are relatively more robust compared to both vocal tract as well as subsegmental features, since it shows only about 3% relative degradation in the performance from NIST-99 to NIST-03 database.

2.3 Speaker information from suprasegmental processing of LP residual

Subsegmental processing models speaker information up to 4 kHz. Segmental processing models speaker information up to 1 kHz. Beyond that LP residual also contains some speaker information at very low frequency range, that is, may be less than 100 Hz. For example the variation in pitch and energy across several glottal cycles (Atal 1972; Farrus and Hernando 2009). In capturing such information, we need to process the LP residual at the suprasegmental level, for example, with frames of 100–300 msec range. For the LP residual sampled at 8 kHz, the feature vectors from such frames will be of very large dimension for building models. We prefer to decimate the LP residual by a factor 50 so that the sampling rate becomes 160 Hz and we may have the source information up to 80 Hz. The dimension of the feature vector is also reduced by 50 factor. Further, the high frequency information that is already modeled by subsegmental and segmental level processing will be smoothed out. Therefore in suprasegmental level processing of LP residual, we decimate the LP residual by a factor of 50 and process in frames of 250 msec with shift of 6.25 msec. The frame size is decided so that the dimension of the feature vectors will remain same as in subsegmental and segmental processing. However, the minimum possible frame shift in this case is 6.25 msec which corresponds to one same shift. Figure 2(h) shows a suprasegmental feature derived from the decimated residual shown in Fig. 2(g). The fast varying components of the original LP residual are eliminated and it mostly represent the long term variations. This can also be observed from the spectrum of the shown feature vector from Fig. 2(i). Information present in the smoothed spectrum is up to 80 Hz. The periodicity and other high frequency related information are absent.

The speaker information present in these features is verified from the recognition experiments as performed earlier. The results of the identification and verification experiments are given in the fourth column of the Table 1 and Table 2, respectively. Results show that suprasegmental level features contain some speaker information. Further, the recognition performance is significantly poor compared to subsegmental, segmental and vocal tract information. The poor result indicates that the suprasegmental features may have large intra-speaker variability. The other major factor is text-independent mode of operation. However, it may contain different aspect of speaker information and hence may combine well with other features.

2.4 Combining evidences from subsegmental, segmental and suprasegmental levels of LP residual

By the way of deriving each feature, the information present at subsegmental, segmental and suprasegmental levels are different and hence may reflect different aspect of speaker-specific source information. By comparing their recognition performance it can be observed that the segmental features provide best performance. Thus the segmental features may have more speaker-specific evidence compared to other level features. The different performances in the recognition experiments indicate the different nature of speaker information present. In this section we use confusion patterns and scatter diagrams to further explain the different nature of the speaker information present in the proposed features and their usefulness for combined use in speaker recognition.

In case of identification, the confusion pattern of features is considered as an indication of the different nature of information present. In the confusion pattern, principal diagonal represents correct identification and the rest represents miss classification. Figures 3 and 4 show the confusion patterns of the identification results conducted for all the proposed features using NIST-99 and NIST-03 databases, respectively. In each case, the confusion pattern is entirely different. The decisions for both true and false identification are different. This indicates that they reflect different aspect of source information. This may help in combining the evidences to fur-

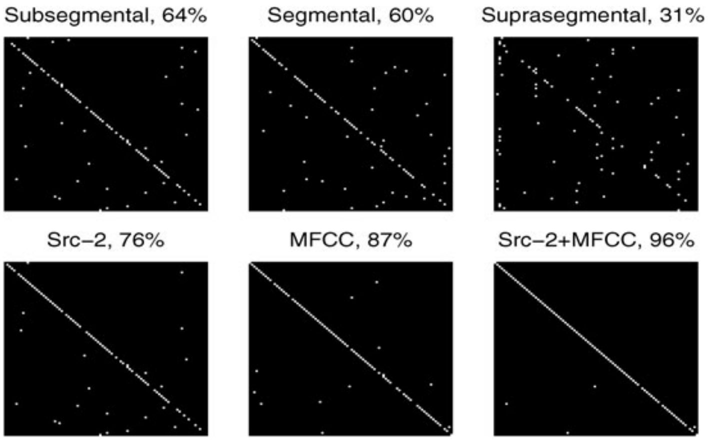

Fig. 3 Confusion patterns of *Sub*, *Seg*, *Supra*, *Src* − 2 and *MFCC* information for identification of 90 speakers from NIST-99 database

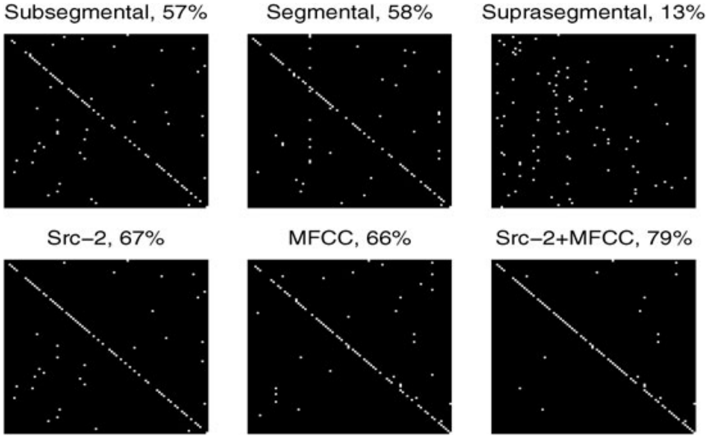

Fig. 4 Confusion patterns of *Sub*, *Seg*, *Supra*, *Src* − 2 and *MFCC* information for identification of 90 speakers from NIST-03 database

ther improve the recognition performance from the source perspective.

For combination we use score level fusion and logical *OR* combination scheme (Mashao and Skosan 2006). In this work the score level and logical *OR* combinations are abbreviated as *Comb* − 1 and *Comb* − 2, respectively. In the score level fusion, the respective scores are weighted by their performances and linearly combined. For example, the log-likelihood ratio (LLR) of the combined system, LLR_c, is given by the following relation:

$$LLR_c = \sum_{i=1}^{C} \frac{P_i}{\sum_{i=1}^{C} P_i} \times LLR_i \tag{4}$$

where, C is the number of systems combined, LLR_i and P_i are the LLR and identification performance of the i^{th} system, respectively. In case of verification mode, the P_i in the above equation is replaced by the reciprocal of the respective equal error rate (EER) and then the scores of the combined system is computed accordingly. The performance of the linearly combined systems are given in Table 1 under the columns with heading *Comb* − 1 for different cases. In all the cases, the performance is improved compared to their respective individual performance. In case of NIST-99 database, the performance is improved from 64% to 68% and in case of NIST-03 database from 57% to 60%. It should be noted here that the small improvement in the performance should not be confused with the worth of combined use of all the features as a best representation of the source information. It is because the performance of the combination system also depends on the combination scheme employed.

It is well known that, simple linear combination with predefined weights may not necessarily provide the best result (Zheng et al. 2007). This is because, fusion of scores may result in a wrong decision. To get a feel of the potential of the combined use of the features in representing the source information, we use logical *OR* combination. In this combination, if any one system is giving correct decision, we consider it as a correct decision. The performance of the *OR* combined systems are also given in the Table 1. The results show that the maximum benefit we can achieve from the proposed features for the NIST-99 and NIST-03 databases are 76% and 67%, respectively. This result shows that if we have a suitable combination scheme, we will benefit by the proposed features. Further in comparison with the vocal tract information, the confusion patterns of the combined system is different from the vocal tract system. By combining evidences from both the features, the respective performances given in the last column of the Table 1 are improved. This indicates that the proposed feature is well combined with the vocal tract information.

In case of verification, as suggested in Zheng et al. (2007), the different aspect of speaker information in the three features are verified from their distribution of scores for imposter and genuine trails. Distribution of two dimensional (2-D) LLR scores for genuine and imposter trials among subsegmental, segmental and suprasegmental features are shown in Figs. 5(a)–(c), respectively. In these figures 'o' represent genuine and 'x' represent imposter speaker. In the regions marked as marked as *I* and *II*, the respective features give different decision. For example, in region *I*, feature represented by x-axis rejects, but the other one accepts. Similarly in region *II*, feature represented by x-axis accepts but the other one rejects. Further, in these regions, some genuine rejected and imposters accepted by one feature are corrected by other. These observations indicate the different nature of speaker information present in these features. In combining the evidences, we use two combination techniques such as linear and logical *OR* combination scheme. In linear combination, weighted scores are combined linearly. In logical *OR* combination, the true scores around the mean provided by the good system are modified based on the information provided by the poor system. In case of linear combination, the performance

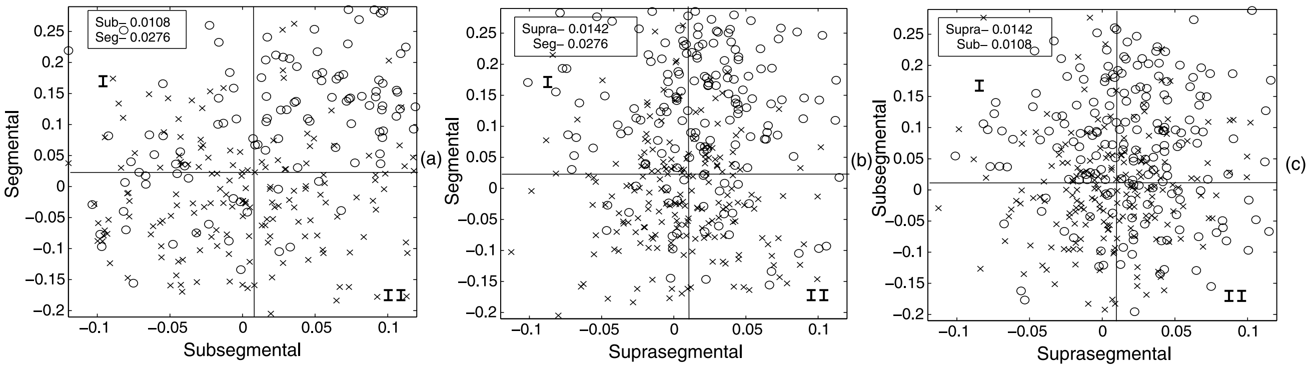

Fig. 5 Distribution of 2-D LLR scores of, (**a**) Subsegmental and segmental information, (**b**) Suprasegmental and segmental information, (**c**) Suprasegmental and subsegmental information

is decreased. The reason may be as mentioned earlier. In case of logical *OR* combination, the performance achieved for the combined system as shown in Table 2 is 21.22% which is even better than the MFCC features. This shows that it is indeed possible to get better performance from the source than vocal tract information, provided we have suitable combination technique. In case of combining the evidences from the proposed feature with MFCC, performance is further improved by the logical *OR* combination scheme.

From this section we observe that the combined use of subsegmental, segmental and suprasegmental features provide useful speaker-specific source information. This information is also well combined with the vocal tract information to improve the recognition accuracy. Further, individually the subsegmental, segmental and suprasegmental features are not providing recognition performance at par with vocal tract information. The reason may be that each of them represent one aspect of speaker information due to source. Further, the results given in Tables 1 and 2 show that, the combination of the subsegmental, segmental and suprasegmental level information performs slightly better compared to vocal tract information. These results are interesting because they demonstrate that it is indeed possible to achieve speaker recognition performance using only excitation source information, which is either comparable or even better compared to the vocal tract information.

The speaker information in the LP residual may be attributed to both the amplitude values and the sequence knowledge. In the next section we describe a method in extracting the subsegmental, segmental and suprasegmental speaker information by separating the amplitude and sequence information of the LP residual. The method involves analytic signal representation of the LP residual. Since the amplitude and sequence information are two different aspects of speaker information, their combined effect may provide improved performance.

3 Speaker information using analytic signal representation of LP residual

In the previous section, speaker information from the LP residual was derived by direct processing of the LP residual at the subsegmental, segmental and suprasegmental levels. The dominant speaker information present in these three levels of processing mostly represents the amplitude and sequence information of the source. When the LP residual is processed directly, the effect of amplitude values dominate over the sequence information, especially, around the instants of glottal closure (Murty and Yegnanarayana 2006). It may therefore be better to separate the amplitude and sequence information and then process them independently. One approach to achieve this is with the use of analytic signal representation of the LP residual (Cohen 1995). In this representation, the magnitude of the analytic signal of LP residual represents the amplitude values of the LP residual and the cosine of the phase of the analytic signal represents the sequence information. Thus the analytic signal representation of the LP residual may help in exploiting the amplitude and sequence information separately. We propose to derive the subsegmental, segmental and suprasegmental features from the analytic signal representation of the LP residual.

The analytic signal of the LP residual $r_a(n)$ corresponding to the LP residual $r(n)$ is given by (Cohen 1995)

$$r_a(n) = r(n) + jr_h(n) \tag{5}$$

where $r_h(n)$ is the Hilbert transform of $r(n)$ and is given by

$$r_h(n) = IFT[R_h(\omega)] \tag{6}$$

where

$$R_h(w) = \begin{cases} -jR(w), & 0 \leq w < \pi \\ jR(w), & 0 > w \geq -\pi \end{cases} \tag{7}$$

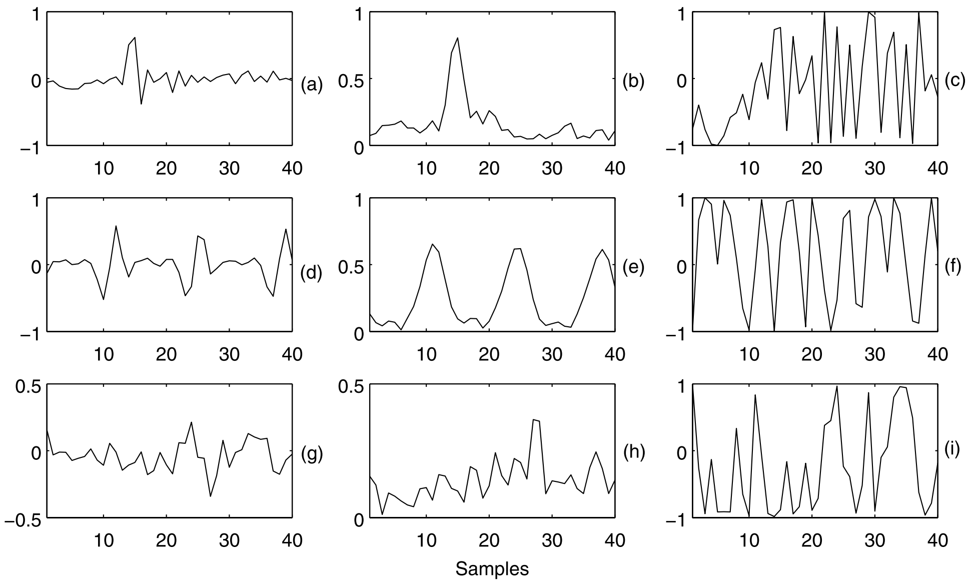

Fig. 6 Decomposition of subsegmental, segmental and suprasegmental feature vectors using analytic signal representation. (**a**) Subsegmental feature vector. (**b**)–(**c**) HE and RP of subsegmental feature vectors, respectively. (**d**) Segmental feature vector. (**e**)–(**f**) HE and RP of segmental feature vectors, respectively. (**g**) Suprasegmental feature vector. (**h**)–(**i**) HE and RP of suprasegmental feature vectors, respectively

$R(\omega)$ is the Fourier transform of $r(n)$ and IFT denotes the inverse Fourier transform. The magnitude of the analytic signal, called as the Hilbert envelope (HE) of the LP residual is given by (Murty and Yegnanarayana 2006)

$$|r_a(n)| = \sqrt{r^2(n) + r_h^2(n)} \tag{8}$$

and the cosine of the phase, called as the residual phase (RP) is given by Murty and Yegnanarayana (2006)

$$\cos(\theta(n)) = \frac{Re(r_a(n))}{|r_a(n)|} = \frac{r(n)}{|r_a(n)|} \tag{9}$$

The procedure to compute the subsegmental, segmental and the suprasegmental feature vectors from HE and RP of the LP residual is same as described earlier except the input sequence. In one case the input will be HE and the other case it will be RP. Example of subsegmental information derived from the LP residual and HE of the LP residual are shown in Figs. 6(a) and (b), respectively. The unipolar nature of the HE helps in suppressing the bipolar variations representing sequence information and emphasizing only the amplitude values. As a result, the amplitude information in the subsegmental sequence of the LP residual is further emphasized by its HE counterpart. Similar observation can also be made in case of segmental and suprasegmental levels processing as shown in Figs. 6(d) and (e), and Fig. 6(g) and (h), respectively. On the other hand, the residual phase represents the sequence information of the residual samples. Figures 6(c), (f) and (i) show the residual phase of the subsegmental, segmental and suprasegmental processing, respectively. In all these cases, the amplitude information is absent. Hence analytic signal representation provides amplitude and sequence information of the LP residual samples independently. In (Murty and Yegnanarayana 2006), it was shown that information present in the residual phase significantly contributes to the speaker recognition. We propose that, the information present in the HE may also contribute well to speaker recognition. Further, as they reflect different aspect of the source information, the combined representation of both the evidences may be more effective for speaker recognition. We conduct different experiments to verify this proposal. The observation from all these experiments are described next.

Subsegmental, segmental and suprasegmental sequences are derived from the HE and RP of the LP residual. In this study subsegmental, segmental and suprasegmental sequences derived from the LP residual, HE of the LP residual and phase of the LP residual are called as the residual features, HE features and RP features, respectively. The potential of the HE and RP features are verified from different recognition experiments. For fair comparison with the residual features, the experimental conditions remain same as mentioned earlier, except for the use of the HE and RP features.

The speaker identification performances of these features for both the databases are given in Tables 3 and 4 and the verification performances for whole NIST-03 database is given in Table 5. In these tables the performance of the residual features are also given for comparison purpose. For both the tasks, the performance of individual HE and RP features is comparatively poorer than their corresponding residual features. Because, as mentioned earlier, HE and RP features independently represent two different aspects of the information that is present in the residual features. The different nature of the information present in the HE and RP features can also be observed from their confusion patterns obtained from the identification tasks. Figure 7 shows the confusion patterns of the identification results conducted for HE and

Table 3 Speaker identification performance of residual, HE, RP and HE + RP features for 90 speakers from NIST-99 database

Feature		*Sub*	*Seg*	*Supra*	$Src - 2$		*MFCC*	$Src - 2 + MFCC$	
					$Comb - 1$	$Comb - 2$		$Comb - 1$	$Comb - 2$
Residual		64	60	31	68	76	87	84	96
HE		44	56	8	66	71		88	94
RP		49	69	17	69	73		86	93
HE + RP	$Comb - 1$	57	69	13	74	88		87	98
	$Comb - 2$	64	78	22					

Table 4 Speaker identification performance of residual, HE, RP and HE + RP features for 90 speakers from NIST-03 database

Feature		*Sub*	*Seg*	*Supra*	$Src - 2$		*MFCC*	$Src - 2 + MFCC$	
					$Comb - 1$	$Comb - 2$		$Comb - 1$	$Comb - 2$
Residual		57	58	13	60	67	66	70	79
HE		32	39	7	47	54		70	76
RP		23	51	14	48	56		69	77
HE + RP	$Comb - 1$	40	54	12	58	72		70	83
	$Comb - 2$	48	59	17					

Table 5 Speaker verification performance of residual, HE, RP and HE+RP features for whole NIST-03 database

Feature		*Sub*	*Seg*	*Supra*	$Src - 2$		*MFCC*	$Src - 2 + MFCC$	
					$Comb - 1$	$Comb - 2$		$Comb - 1$	$Comb - 2$
Residual		41.01	26.96	44.49	32.25	21.22	22.94	27.78	17.43
HE		45.52	32.92	45.66	36.27	22.31		26.92	21.01
RP		41.73	26.83	45.84	31.39	22.13		20.01	20.46
HE + RP	$Comb - 1$	43.90	27.19	44.94	33.28	20.41		22.99	16.67
	$Comb - 2$	30.12	21.36	32.83					

RP features using NIST-99 database. At each level, the confusion patterns of the HE and RP features are different. Their decisions for both true and false identification are different. This indicates that the information present in HE features is different from that of RP features. By combining individual evidences, the respective performances may be further improved.

There are two approaches that can be used for combining evidences from HE and RP. In one approach, at each level, HE and RP can be combined independently (vertically) and this evidence at each level can be further combined to obtain overall source information. Alternatively, the HE and RP from all the three levels can be combined first (horizontally) and then these combined HE and RP evidences are further combined to obtain complete source information. From the experimental results we observed that the later approach seem to give better performance. The reason may be that HE and RP information from all the three levels together may combine well to become more speaker-specific, because their origin is same.

In combining the evidences we employ both $Comb - 1$ and $Comb - 2$ combination schemes described earlier. The identification performance of the various combinations for NIST-99 and NIST-03 databases are given in Tables 3 and 4, respectively and the verification performance for the whole NIST-03 database is given in Table 5. The results show that for less noisy data (i.e. NIST-99), the performance achieved from combined HE and RP features is better than the residual feature. For noisy data (i.e. NIST-03), for both the tasks, the performance is slightly poor than the residual feature. The reason may be the quality of the data and the combination technique employed. For example in case of combination scheme $Comb - 2$, the recognition performance is improved. Further in noisy condition, with MFCC features, the combined representation of the HE and RP features is providing better performance than the residual feature. This shows the robustness of the combined HE and RP representation of the source in providing the additional information to the MFCC feature. From this observation we conclude that combined representation of HE and RP features may be better than the residual feature alone.

The above observations indicate that complete information present in the source can be represented by the combined representation of the HE and RP features. Further, to achieve maximum benefit, it may be better to

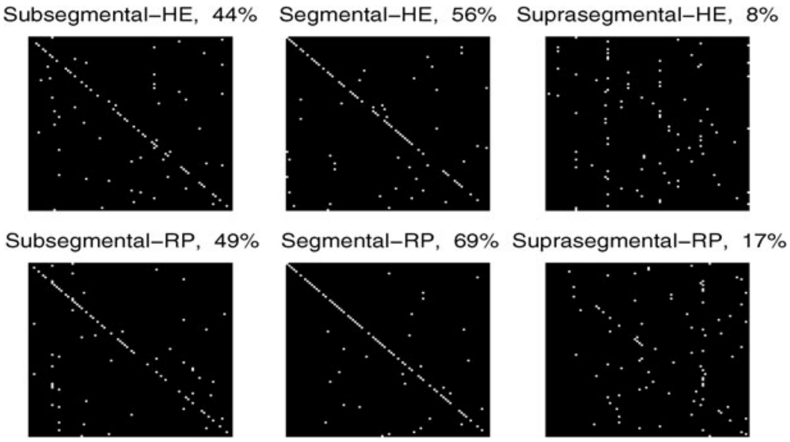

Fig. 7 Confusion patterns of HE and RP features for identification of 90 speakers from NIST-99 database

first combine the HE and RP at subsegmental, segmental and suprasegmental levels separately and then combine them. The speaker recognition performance of the information present in the segmental level is comparatively better than the other two levels. The segmental level features namely, pitch and energy seem to be more speaker-specific. The recognition performance of the information present in the suprasegmental level is very poor compared to the other levels. The suprasegmental level information may have large intra-speaker variability and also due to the text-independence. In the next section, we propose an alternative approach for extracting only suprasegmental level speaker information using the instantaneous pitch concept proposed in Yegnenarayana and Murthy (2009). This study enables us to understand whether poor performance is due to the level of processing or the method employed.

4 Suprasegmental speaker information using instantaneous pitch and epoch strength

In this section an alternative approach is employed for extracting the suprasegmental level information. The objective is to verify the effectiveness of the proposed method employed in extracting the suprasegmental level information using LP residual described in Sect. 2.3. The excitation source at the suprasegmental level mostly contains the pitch contour and epoch strength contour information. Epoch strength represents the strength at the instant of glottal closure in case of voiced speech (Murty and Yegnanarayana 2006). In an alternative approach, we directly compute the pitch and epoch strength contours and then use them as features to represent the suprasegmental level information. To compute the pitch and epoch strength values, we use the recently proposed instantaneous pitch estimation method (Yegnenarayana and Murthy 2009; Murthy and Yegnanarayana 2008, 2009). The advantage of using this method is that it computes the instantaneous pitch values and hence gives accurate values for pitch and epoch strength contours. A brief description of this method is given below.

Instantaneous pitch estimation method locates the glottal closure instants (GCIs) by passing the speech signal through a zero-frequency resonator twice. The zero-frequency resonator is a second order infinite impulse response (IIR) filter located at 0 Hz (Murthy and Yegnanarayana 2009). The purpose of passing the speech signal twice is to reduce the effects of all (high frequency) resonances (Murthy and Yegnanarayana 2008). Passing the speech signal twice through a zero frequency resonator is equivalent to four times successive integration. This will result a filtered output that grows/decays as a polynomial function of time. The trend in the filtered signal is removed by subtracting the local mean computed over an interval corresponding to the average pitch period. The resulting mean subtracted signal is called as *zero-frequency filtered signal*. Following steps are involved in processing the speech signal to derive the *zero-frequency filtered signal*.

(1) Difference the speech signal $s(n)$

$$x(n) = s(n) - s(n-1) \tag{10}$$

(2) Pass the difference speech signal $x(n)$ twice through zero-frequency resonator

$$y_1(n) = -\sum_{k=1}^{2} a_k y_1(n-k) + x(n) \tag{11}$$

and

$$y_2(n) = -\sum_{k=1}^{2} a_k y_2(n-k) + y_1(n) \tag{12}$$

where, $a_1 = -4$, $a_2 = 6$, $a_3 = -4$ and $a_4 = 1$

(3) compute the average pitch period using the autocorrelation over a 20 msec speech segment

(4) Remove the trend in $y_2(n)$ by subtracting the mean computed over average pitch period. The resulting signal

$$y(n) = y_2(n) - \frac{1}{2N+1} \sum_{m=-N}^{N} y_2(n+m) \qquad (13)$$

is the *zero-frequency filtered signal*. Here, $2N + 1$ corresponds to the number of samples in the window used for mean substraction.

The positive zero crossings in the *zero-frequency filtered signal* correspond to the locations of the GCIs (Murthy and Yegnanarayana 2008). The interval between successive positive zero-crossings gives the instantaneous pitch period t_0. The reciprocal, $f_0 = \frac{1}{t_0}$ is the instantaneous pitch frequency (Yegnenarayana and Murthy 2009). The slope of the *zero-frequency filtered signal* around the zero crossings corresponding to the location of the epochs gives a measure of epoch strength a_0 (Murthy and Yegnanarayana 2009).

The zero-frequency resonator filter out a mono-component centered around the 0 frequency from the speech signal. However, in case of the telephonic speech, the frequency components below 300 Hz are heavily damped. The output of the zero-frequency resonator obtained from processing the telephonic speech may not give correct estimation of the pitch and epoch strength. To avoid this difficulty, we purpose to use the positive zero-crossings in the *zero-frequency filtered signal* derived from the HE of the LP residual for computation of pitch and epoch strength contours. Due to impulse-like nature of the LP residual, the information about the fundamental frequency will spread across all the frequencies including the zero frequency. The purpose of using the HE is to emphasize the peaks around the GCIs in each glottal cycle (Ananthapadmanabha and Yegnanarayana 1979; Yegnanarayana and Prasanna 2010).

To verify the effectiveness of the proposed approach, we compute the epochs from a telephonic speech and compare them with the estimated epochs from the *zero-frequency filtered signal* of the corresponding clean speech. For this, we collect the speech data of a speaker from TIMIT and NTIMIT databases (Zue et al. 1990; Jankowski et al. 1990). For both the cases the text of the speech remains same. The speech data collected from TIMIT database represents the clean speech and from the NTIMIT database represents the corresponding telephonic speech. Figures 8(a) and (b) show a segment of clean speech and the corresponding *zero-frequency filtered signal* derived from the clean speech, respectively. The arrows in the *zero-frequency filtered signal* indicate the location of the positive zero-crossings. It can be observed that the instants of the positive zero-crossings in the *zero-frequency filtered signal* clearly indicate the location of the epochs. Further, Figs. 8(c), (d) and (e) show the segment of telephonic speech of the same text as in case of

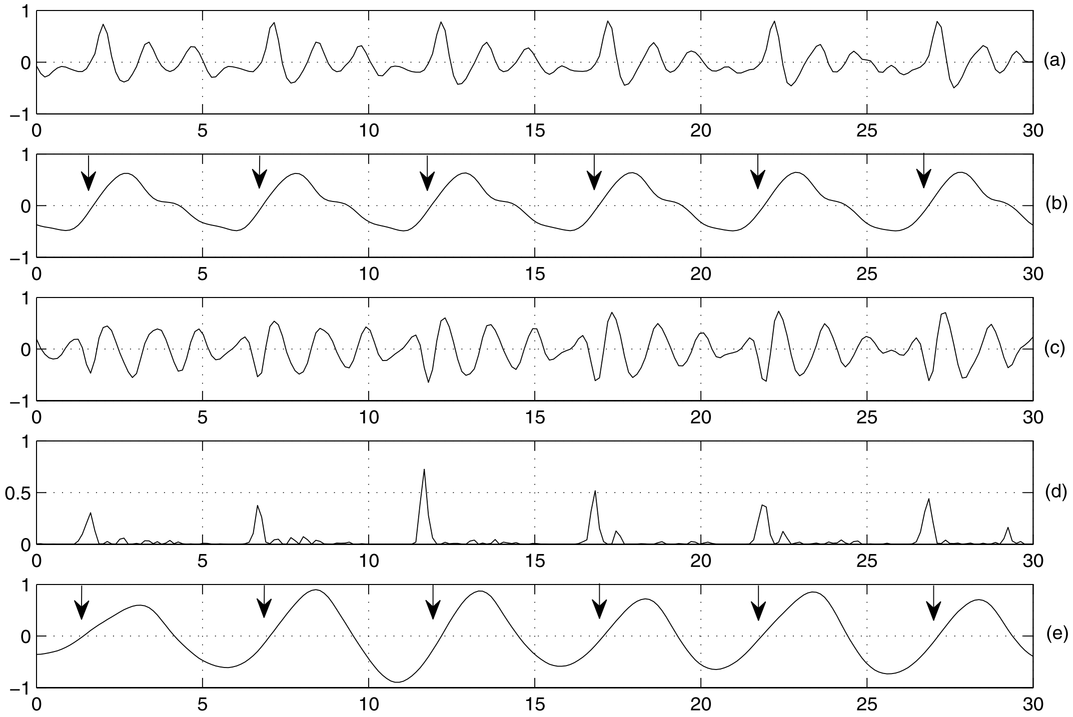

Fig. 8 Estimation of pitch period from clean and telephonic speech signal. (**a**) Clean speech. (**b**) *zero-frequency filtered signal* derived from the speech signal in (**a**). (**c**) Speech signal of the same text as in (**a**) collected over telephone channel. (**d**) The HE of the LP residual of the speech signal in (**c**). (**e**) *Zero-frequency filtered signal* derived from the signal in (**d**). The location of the positive zero-crossings in the filtered signal (**b**) and (**e**) are shown by *arrows*

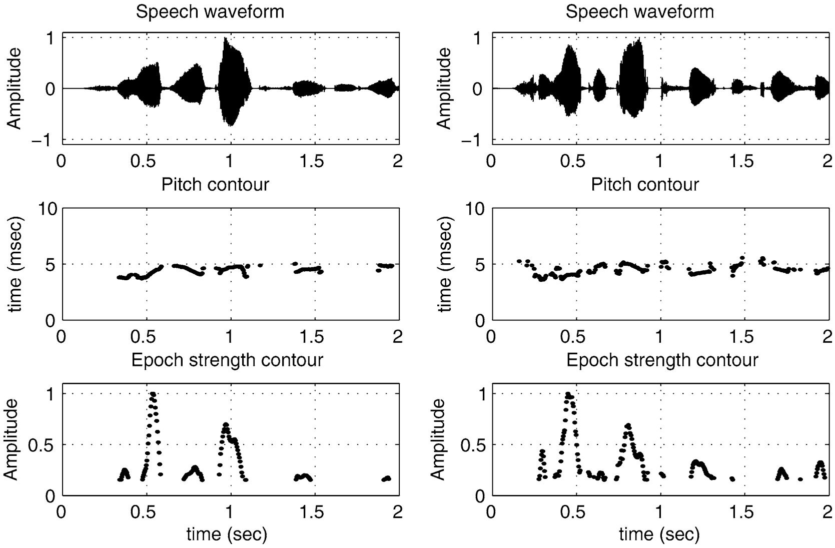

Fig. 9 Examples of speech waveforms, pitch and epoch strength contours computed using zero-frequency filtering approach of two female speakers collected from TIMIT database. (*Left*) *Speaker* − 1. (*Right*) *Speaker* − 2

clean speech, HE of the LP residual of the telephonic speech and *zero-frequency filtered signal* derived from the HE of the LP residual of the telephonic speech, respectively. It can be observed that the time instants of the zero-crossings indicated by arrows in the *zero-frequency filtered signal* corresponds to the original epochs shown in Fig. 8(b). From this observation we conclude that in case of telephonic speech, the *zero-frequency filtered signal* derived from the HE of the LP residual can be used to compute the pitch and epoch strength.

In this work, zero-frequency filtering approach as described above is used for computation of pitch and epoch strength contours. Figure 9 shows the examples of pitch and epoch strength contours of two female speakers collected from TIMIT database. It can be observed that contours are significantly different across the speakers. This shows that pitch and epoch strength contours contain speaker-specific information. Since the contours are computed across a longer segment of the voiced speech, the speaker-specific information present in them is attributed to the suprasegmental level. The suprasegmental level information is usually extracted from 100–300 msec segments and hence we need on an average around 25–50 pitch values to represent a feature vector. Since the nature of pitch and epoch strength contours have large intra-speaker variability, the dimension of the feature vectors consisting of 25–50 values may not seem to be effective for the recognition task. Further, pitch and epoch strength are computed from voiced speech only. The number of feature vectors obtained with 25–50 dimension may be comparatively less.

With large dimension and less number of feature vectors, speaker information may not be modeled well. Thus due to intra-speaker variability and poor modeling, matching may be difficult. For this reason we prefer to use lower dimension feature vectors. To select suitable dimension, we conduct a speaker identification study for different dimensions of pitch values for 30 speakers set collected from NIST-99 database. In this experiment the feature vectors are made by sequence of pitch values with a shift of one pitch value. The reason for considering every sample shift of the pitch values is to get the maximum number of feature vectors. The result of this experiment is shown in Fig. 10. From this figure we observe that with increase in the dimension from 1 to 10 the performance is increased. With further increase in dimension, the performance is decreased. The reason may be that with increase in dimension, the intra-speaker variability may also be increased. So we use 10 pitch values with shift of one value to represent pitch feature vectors. We call them as t_0 *vectors*. Similarly we use 10 epoch strength values with shift of one value to represent epoch strength features. We call them as a_0 *vectors*. The combined evidences from t_0 and a_0 *vectors* called as $(t_0 + a_0)$ *vectors* is used as the complete

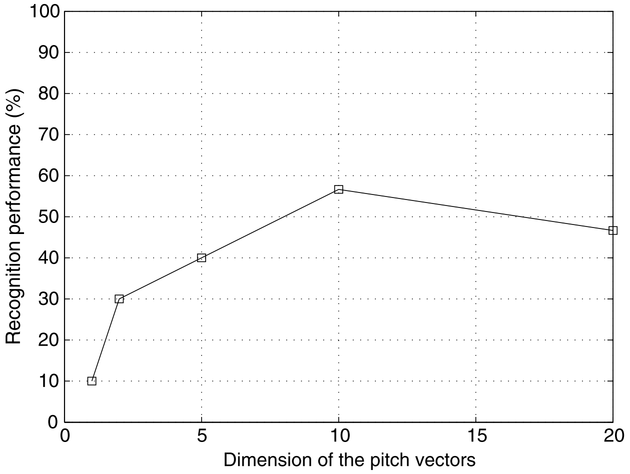

Fig. 10 Speaker identification performance of pitch vectors for different dimension

Table 6 Comparison of speaker identification and verification performances of *Supra*, $t0$ and $a0$ feature vectors

Feature		*Performance*		
		Identification		*Verification*
		$NIST-99$	$NIST-03$	
Supra	$Perf-1$	31	13	44.49
	$Perf-2$	31	17	32.83
$t0vectors$		29	18	45.39
$a0vectors$		9	7	49.27
$(t0+a0)vectors$	$Perf-1$	32	13	45.32
	$Perf-2$	33	21	31.21
$Sub+Seg+Supra$	$Perf-1$	74	60	33.28
	$Perf-2$	88	72	20.41
$Sub+Seg$+$(t0+a0)vectors$	$Perf-1$	74	59	31.16
	$Perf-2$	90	74	19.78
$Sub+Seg+Supra+mfcc$	$Perf-1$	87	70	22.99
	$Perf-2$	98	83	16.67
$Sub+Seg$+$(t0+a0)vectors+mfcc$	$Perf-1$	88	69	25.34
	$Perf-2$	96	86	16.53
mfcc		87	66	22.94

suprasegmental information of the source. In combining the evidences we employ $Comb-1$ and $Comb-2$ techniques as described earlier. Using these feature vectors the recognition experiments are conducted.

The effectiveness of the methods employed in extracting the suprasegmental information is verified by comparing the recognition performances of the two approaches. The results of the identification and verification experiments are given in Table 6. The performance of the suprasegmental feature derived from the decimation of the LP residual (*Supra*) is given for comparison. In this table $Perf-1$ represent the maximum performance of a feature vector that can be achieved either from LP residual or by combination of its analytic signal decomposition. $Perf-2$ represents the maximum performance that can be achieved by using even the logical OR combination, that is, $Comb-2$. For example, in case of identification task for NIST-03 database, $Perf-1$ is 13% and $Perf-2$ is 17%. The results show that for both identification and verification tasks, the combined representation of the pitch and epoch strengths is providing nearly same performance as compared to the suprasegmental features derived from the decimation of the LP residual. Similar observation is also made when they are combined with other two levels information of the source. The combined performance in both the cases are nearly same. It is also observed that both are providing almost same additional information to MFCC. When they are combined with MFCC separately, the combined performance in both the case is almost same. This observation indicates that the effectiveness of the information present in pitch and epoch strength contours and in suprasegmental features derived from the LP residual are almost same. Further in case of pitch and epoch strengths features there is a slight improvement in the performance in some cases, but the computation involved in this approach is all together different. For unified processing, we therefore recommend that the features derived from the LP residual decimated by a factor 50 can be used to represent the source information at the suprasegmental level. These studies indicate that the information present at the suprasegmental level may be less effective due to large intra-speaker variability and also due to text-independent mode of operation.

5 Summary and conclusion

In this work an unified framework is proposed for the extraction of complete source information by the time domain analysis of the LP residual. Speaker specific information in the LP residual include those within one glottal cycle, pitch and energy across two to three glottal cycles, and variation of the pitch and energy across several glottal cycles. In the proposed method, speaker information within one glottal cycle is extracted by the subsegmental processing of the LP residual. The pitch and energy information is extracted by the segmental processing of the LP residual. Pitch and energy contour information is extracted from the suprasegmental processing of the LP residual. To model the speaker information effectively using GMM, the segmental and suprasegmental level information is decimated by a factor of 4 and 50, respectively. Experimental results show that subsegmental, segmental and suprasegmental levels contain speaker information. Further combining the evidences from each level, the performance improvement indicates the different nature speaker information at each level. In direct processing of the LP residual the effect of the amplitude dominate the sequence information. To minimize this, the amplitude and sequence information is captured independently using the analytic signal representation of the LP residual. The combination of amplitude and sequence information seem to be a better choice. At the individual level, information provided by segmental level of the LP residual is most effective compared to the other two levels. The information provided at the suprasegmental level processing of the LP residual is poor due to intra-speaker variability and text-independence. This is also confirmed by an alternative approach using pitch and epoch strength contours to capture the suprasegmental information.

In this work the excitation source information is extracted by processing the LP residual in the time domain. The time domain processing of the LP residual is computationally intensive. Because the waveform itself is directly modeled. To explore the possibility of compact parametric representation of the excitation information, LP residual can be processed from the other domains like frequency or cepstrum. This has to be done by keeping in view of the blocking effect that is present in these domains. Further, in this work we use the combination scheme based on logical *OR* to demonstrate the potential of source evidence. New combination techniques need to be explored to exploit the same.

References

Ananthapadmanabha, T. V., & Yegnanarayana, B. (1979). Epoch extraction from linear prediction residual for identification of closed glottis interval. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *ASSP-27*, 309–319.

Atal, B. S. (1972). Automatic speaker recognition based on pitch contours. *The Journal of the Acoustical Society of America*, *52*(6), 1687–1697.

Atal, B. S. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *The Journal of the Acoustical Society of America*, *55*(6), 1304–1312.

Cohen, L. (1995). *Time-frequency analysis: theory and application. Signal processing series*. Englewood Cliffs: Prentice Hall.

Davis, S. B., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(28), 357–366.

Ezzaidi, H., & Rouat, J. (2004). Pitch and MFCC dependent GMM models for speaker identification systems. In *IEEE int. conf. on electrical and computer eng.*: *Vol. 1*.

Farrus, M., & Hernando, J. (2009). Using jitter and shimmer in speaker verification. *Signal Processing*, *3*(4), 247–257.

Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *29*(2), 254–272.

Hayakawa, S., Takeda, K., & Itakura, F. (1997). Speaker identification using harmonic structure of lp-residual spectrum. In *Lecture notes in computer science: Vol. 1206. Audio- and video-based biometric personal authentification* (pp. 253–260). Berlin: Springer.

Huang, W., Chao, J., & Zhang, Y. (2008). Combination of pitch and MFCC GMM supervectors for speaker verification. In *IEEE int. conf. on audio, language and image process (ICALIP)* (pp. 1335–1339).

Jankowski, C., Kalyanswamy, A., Basson, S., & Spitz, J. (1990). NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database. In *Int. conf. on acoust. speech and signal process. (ICASSP)*, Albuquerque, NM (pp. 109–112).

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, *63*(4), 561–580.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., & Przybocki, M. (1997). The DET curve in assessment of detection task performance. In *Proc. Eur. conf. on speech communication technology*, Rhodes, Greece, Vol. *4* (pp. 1895–1898).

Mary, L., & Yegnanarayana, B. (2008). Extraction and representation of prosodic features for language and speaker recognition. *Speech Communication*, *50*, 782–796.

Mashao, D. J., & Skosan, M. (2006). Combining classifier decisions for robust speaker identification. *Pattern Recognition*, *39*, 147–155.

Murthy, K. S. R., & Yegnanarayana, B. (2008). Epoch extraction from speech signal. *IEEE Transactions on Speech and Audio Processing*, *16*(8), 1602–1613.

Murthy, K. S. R., & Yegnanarayana, B. (2009). Characterization of glottal activity from speech signal. *IEEE Signal Processing Letters*, *16*(6), 469–472.

Murty, K. S. R., & Yegnanarayana, B. (2006). Combining evidence from residual phase and MFCC features for speaker recognition. *IEEE Signal Processing Letters*, *13*(1), 52–55.

Murty, K. S. R., Prasanna, S. R. M., & Yegnanarayana, B. (2004). Speaker specific information from residual phase. In *Int. conf. on signal proces. and comm. (SPCOM)*.

Nist speaker recognition evaluation plan (2003). In: Proc. NIST speaker recognition workshop, College Park, MD.

Pati, D., & Prasanna, S. R. M. (2010). Speaker information from subband energies of linear prediction residual. In *Proc. NCC* (pp. 1–4).

Plumpe, M. D., Quatieri, T. F., & Reynolds, D. A. (1999). Modelling of glottal flow derivative waveform with application to speaker identification. *IEEE Transactions on Speech and Audio Processing*, *7*(5), 569–586.

Prasanna, S. R. M., Gupta, C. S., & Yegnanarayana, B. (2006). Extraction of speaker-specific excitation information from linear prediction residual of speech. *Speech Communication*, *48*, 1243–1261.

Przybocky, M., & Martin, A. (2000). The NIST-1999 speaker recognition evaluation- an overview. *Digital Signal Processing*, *10*, 1–18.

Reynolds, D. A. (1994). Experimental evaluation of features for robust speaker identification. *IEEE Transactions on Speech and Audio Processing*, *2*(4), 639–643.

Reynolds, D. A. (1995). Speaker identification and verification using gaussian mixture speaker models. *Speech Communication*, *17*, 91–108.

Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, *3*(1), 4–17.

Sonmez, K., Shriberg, E., Heck, L., & Weintraub, M. (1998). Modeling dynamic prosodic variation for speaker verification. In *Proc. ICSLP' 98*: *Vol. 7* (pp. 3189–3192).

Thevenaz, P., & Hugli, H. (1995). Usefulness of the LPC-residue in text-independent speaker verification. *Speech Communication*, *17*, 145–157.

Wolf, J. J. (1972). Efficient acoustic parameters for speaker recognition. *The Journal of the Acoustical Society of America*, *51*(2), 2044–2055.

Yegnanarayana, B., & Prasanna, S. R. M. (2010). Analysis of instantaneous F0 contours from two speakers mixed signal using zero frequency filtering. In *Int. conf. on acoust. speech and signal process. (ICASSP)*, Dallas, Texas, USA (pp. 5074–5077).

Yegnanarayana, B., Reddy, K. S., & Kishore, S. P. (2001). Source and system feature for speaker recognition using AANN models. In *Proc. IEEE int. con. acoust. speech and signal processing*, Salt Lake City, UT, USA, May 2001 (pp. 409–412).

Yegnanarayana, B., Prasanna, S. R. M., Zachariah, J. M., & Gupta, C. S. (2005). Combining evidences from source, suprasegmental and spectral features for fixed-text speaker verification study. *IEEE Transactions on Speech and Audio Processing*, *13*(4), 575–582.

Yegnenarayana, B., & Murthy, K. S. R. (2009). Event based instantaneous fundamental frequency estimation from speech signals. *IEEE Transactions on Audio, Speech and Language Processing*, *17*(4), 614–624.

Zheng, N., Lee, T., & Ching, P. C. (2007). Integration of complimentary acoustic features for speaker recognition. *IEEE Signal Processing Letters*, *14*(3), 181–184.

Zue, V., Seneff, S., & Glassa, J. (1990). Speech database development at MIT: Timit and beyond. *Speech Communication*, *9*(4), 351–356.