Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 48 (2006) 1243–1261
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Extraction of speaker-specific excitation information from linear prediction residual of speech

S.R. Mahadeva Prasanna [a,*], Cheedella S. Gupta [b], B. Yegnanarayana [b]

[a] Department of Electronics and Communication Engineering, Indian Institute of Technology Guwahati, Guwahati 781 039, Assam, India
[b] Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai 600 036, Tamil Nadu, India

Received 15 October 2005; received in revised form 13 March 2006; accepted 19 June 2006

Abstract

In this paper, through different experimental studies we demonstrate that the excitation component of speech can be exploited for speaker recognition studies. Linear prediction (LP) residual is used as a representation of excitation information in speech. The speaker-specific information in the excitation of voiced speech is captured using the AutoAssociative Neural Network (AANN) models. The decrease in the error during training and recognizing correct speakers during testing demonstrates that the excitation component of speech contains speaker-specific information and is indeed being captured by the AANN models. The study on the effect of different LP orders demonstrates 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 excitation information. It is also demonstrated that the proposed speaker recognition system using excitation information and AANN models requires significantly less amount of data both during training as well as testing, compared to the speaker recognition system using vocal tract information. Finally the speaker recognition studies on NIST 2002 database demonstrates that even though, the recognition performance from the excitation information alone is poor, when combined with evidence from vocal tract information, there is significant improvement in the performance. This result demonstrates the complementary nature of the excitation component of speech.

Keywords: Speaker recognition; Excitation information; LP residual; AANN model; Vocal tract information

1. Introduction

Speech is produced from a time varying vocal tract system excited by a time varying excitation source (O'Shaughnessy, 1987; Rabiner and Juang, 1993; Deller et al., 2000). The resulting speech signal contains information about message, speaker, language and emotional status. For analysis and processing of speech signals, the vocal tract system is modeled as a time varying filter, and the excitation as voiced or unvoiced or plosive or combination of these types. The time varying filter characteristics capture variations in the shape of the vocal tract system in the form of resonances, antiresonances and spectral roll-off characteristics. These filter characteristics are usually

* Corresponding author. Tel.: +91 361 2582513; fax: +91 361 2690762.
E-mail addresses: prasanna@iitg.ernet.in (S.R. Mahadeva Prasanna), yegna@cs.iitm.ernet.in (B. Yegnanarayana).

0167-6393/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2006.06.002

represented by spectral features for each short (10–30 ms) segment of speech, and we call these features as *system features*. This representation of speech has been extensively exploited for developing speaker recognition systems (Atal, 1976; Rosenberg, 1976; O'Shaughnessy, 1986; Furui, 1996; Furui, 1997; Campbell, 1997; Reynolds et al., 2000).

Speaker-specific information is also present in the suprasegmental characteristics of a speech signal. These characteristics include word usage (idiolect), variation in pitch, duration of words, speaking rate, speaking style, loudness, phonetics and idiosyncrasies. Doddington has developed a speaker recognition system based on the word usage or idiolect alone (Doddington, 2001). Incorporation of pitch and duration (prosody) information into speaker recognition system has also been studied (Weber et al., 2002). With sufficient amount of training and test data, it may be possible to capture speaker-specific information from the suprasegmental characteristics and hence may help in significantly enhancing the performance of speaker recognition systems, especially, under degraded conditions. But some of these suprasegmental characteristics are higher level production features, and are difficult to characterize (Yegnanarayana et al., 1992; Madhukumar, 1993). Moreover, these features vary significantly for the same speaker depending on the manner in which the speech is uttered. Further, as mentioned above, a large amount of data is needed to extract the speaker-specific information from the suprasegmental characteristics of a speech signal. Therefore, it is difficult to reliably extract and represent speaker-specific information present at the suprasegmental level for developing speaker recognition systems.

There is yet another component in speech, which is largely ignored in most speech analysis techniques. It is the residual of the speech signal obtained after suppressing the vocal tract characteristics from the signal. The Linear Prediction (LP) analysis may be used for suppressing the vocal tract characteristics (Makhoul, 1975). This is achieved by first predicting the vocal tract information from the signal and then suppressing it by inverse filter formulation. The resulting signal is termed as the LP residual and contains mostly information about the excitation source. In this work the features extracted from the LP residual are referred to as *source features*. Atal has used pitch information extracted from the residual signal for speaker recognition studies (Atal, 1972). Wakita has reported an experiment using the LP residual energy for vowel recognition and also for speaker recognition (Wakita, 1976). It has also been shown that a combination of Linear Prediction Cepstral Coefficients (LPCCs) and energy of the LP residual gives better speaker recognition performance compared to using only LPCCs (Faundez and Rodriguez, 1998). The use of cepstrum computed over the LP residual was also proposed for speaker recognition (Thevenaz and Hugli, 1995). Combination of LPCC and LP residual cepstrum was shown to reduce the error rate in speaker recognition (Liu and Palm, 1997). In all these studies, no specific attempts are made to explore the speaker-specific excitation information present only in the residual of speech. Further, the LP residual may contain more speaker-specific information than those represented by pitch, residual energy and residual cepstrum parameters. Hence a detailed exploration to know the speaker-specific excitation information present in the residual of speech is needed and hence the motivation for the present work.

It has been shown that humans can recognize people by listening to the LP residual signal (Feustel et al., 1989). This may be attributed to the speaker-specific excitation information present at the segmental (10–30 ms) and suprasegmental levels (1–3 s). The presence of speaker-specific information at the segmental and suprasegmental levels can be established by generating signals that retain specific features at these levels. For instance, speaker-specific suprasegmental information (intonation and duration) can be perceived in the signal which has impulses of appropriate strength at each pitch epoch in the voiced region, and at random instances in the unvoiced regions. Such a signal can be generated by first finding the instants of significant excitation of speech and then weighting them with appropriate strengths as discussed in (Smits and Yegnanarayana, 1995). Instants of significant excitation correspond to pitch epochs in case of voiced speech and some random excitation instants like onset of burst events in case of unvoiced speech (Smits and Yegnanarayana, 1995). The LP residual has the additional information of the glottal pulse characteristics in the samples between two pitch epochs. Perceptually the signals will be different if these samples (related to the glottal pulse characteristics) are replaced by synthetic model signals (Rosenberg, 1971; Ananthapadmanabha and Yegnanarayana, 1979) or by random noise (Murthy et al., 2004). It appears that significant speaker-specific excitation information may be present in the segmental and suprasegmental features of the residual. The present work focusses on extracting

speaker-specific excitation information present at the segmental level of the residual.

At the segmental level, each short segment of the LP residual can be considered to belong to one of the five broad categories, namely, voiced, unvoiced, plosive, silence and mixed excitation. The voiced excitation is the dominant mode of excitation during speech production. Further, if voiced excitation is replaced by random noise excitation, it is difficult to perceive the speaker's identity (Murthy et al., 2004). In this paper we demonstrate that the speaker characteristics are indeed present at the segmental level of the LP residual, and they can be reliably extracted using neural network models.

The rest of the paper is organized as follows: In Section 2 we examine the characteristics of the LP residual, and discuss issues involved in extracting the speaker-specific information from the residual. In Section 3 we discuss Autoassociative Neural Network (AANN) models to capture the speaker-specific information from the residual. Section 4 describes the database used in the study. Speaker recognition studies are described in Section 5. In Section 6 the performance of speaker recognition systems based on the features from the LP residual and the features representing the vocal tract system are examined for different orders of LP analysis. The proposed speaker recognition system, based on the LP residual, may not require large amounts of data. This aspect is examined in Section 7. In Section 8 we discuss the speaker recognition studies using the database of NIST 2002 speaker recognition evaluation. Summary and conclusions of this study and the scope for future work are given in Section 9.

2. Speaker characteristics in the LP residual

In the linear prediction analysis of speech each sample is predicted as a linear weighted sum 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

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

The difference between the actual and predicted sample value is termed as the prediction error or residual, which is given by

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

where $\{a_k\}$ are the linear prediction coefficients. The linear prediction coefficients are typically determined by minimizing the mean squared error over an analysis frame. The coefficients can be obtained by solving the set of p normal equations,

$$\sum_{k=1}^{p} a_k R(n-k) = -R(n), \quad n = 1, 2, \ldots, p \tag{3}$$

where

$$R(k) = \sum_{n=0}^{N-(p-1)} s(n)s(n-k), \quad k = 0, 1, \ldots, p \tag{4}$$

and $\{s(n)\}$ are the speech samples.

The residual in Eq. (2) is obtained by passing the speech signal through the inverse filter $A(z)$, given by

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k} \tag{5}$$

The LP spectrum $|H(w)|^2$ is given by

$$|H(w)|^2 = \left| \frac{G}{1 + \sum_{k=1}^{p} a_k \mathrm{e}^{-\mathrm{j}wk}} \right|^2 \tag{6}$$

where G is the gain parameter given by the minimum mean squared error

$$G^2 = \min_{\{a_k\}} \left\{ \sum_n e^2(n) = \sum_{k=0}^{p} a_k R(k) \right\} \tag{7}$$

Fig. 1 shows a segment of voiced speech, its LP residual, short-time spectrum and the 8th order LP spectrum. As the order of the LP analysis is increased, the LP spectrum approximates the envelope of the short-time spectrum better. Through out the paper we use the term *spectrum* to refer to the *power spectrum*, which is related to the autocorrelation function and in turn represents the second order statistics of the signal. The envelope of the short-time spectrum approximates the frequency response of the vocal tract shape, thus reflecting the characteristics of the vocal tract system. Typically the vocal tract system is characterized by maximum of five resonances in the 0–4 kHz range. Therefore an LP order in the range 8–14 seems to be appropriate for a speech signal sampled at 8 kHz. For low orders, the LP spectrum may pick up only the prominent resonance peak as shown in Fig. 2(a) for $p = 2$. In this case the residual will still have significant information about the vocal tract system. Thus the spectrum of the residual

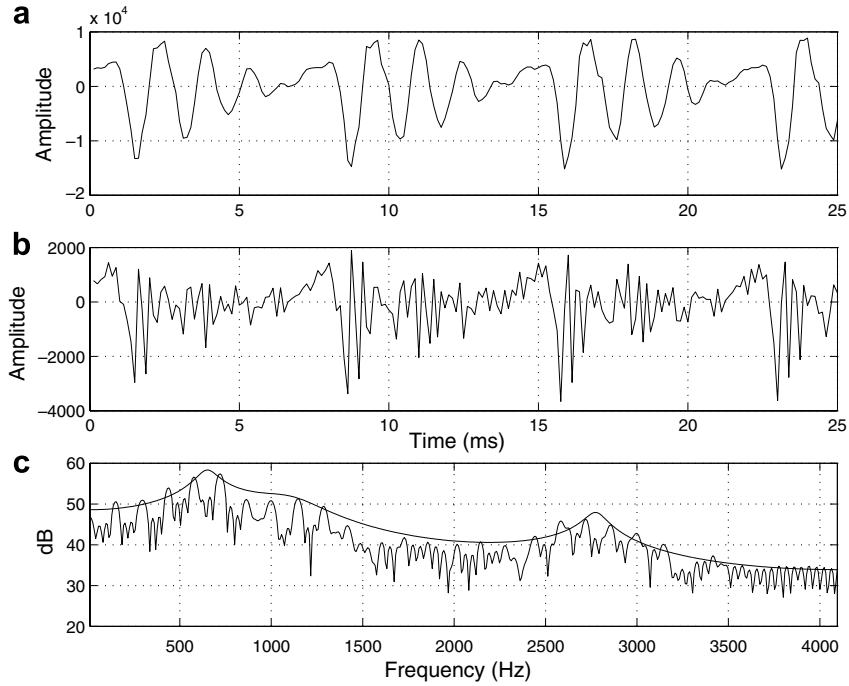

Fig. 1. (a) Segment of voiced speech, and its (b) LP residual and (c) short-time spectrum superimposed with an 8th order LP spectrum.

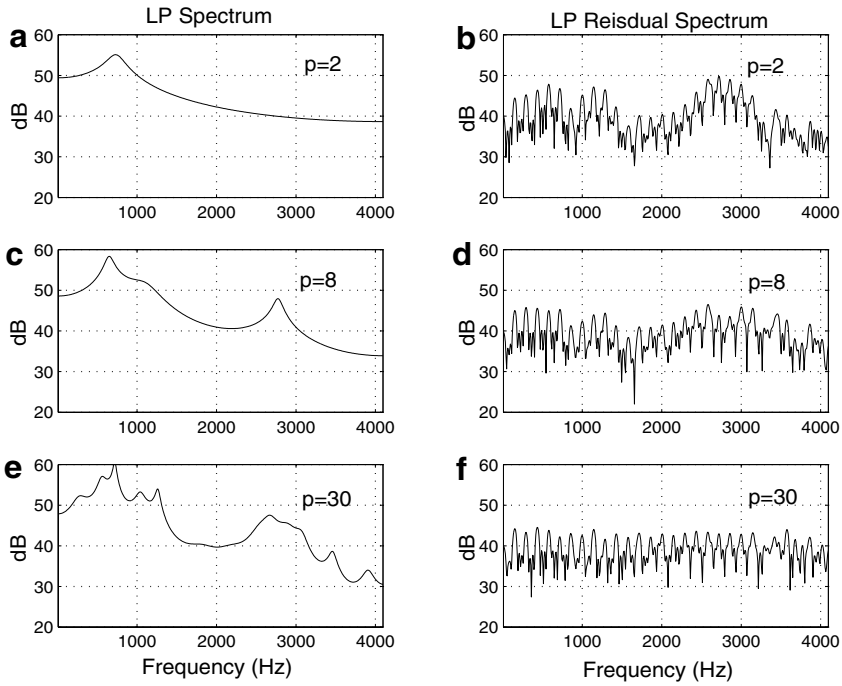

Fig. 2. (a) LP spectrum and (b) residual spectrum for LP order 2. (c) LP spectrum and (d) residual spectrum for LP order 8. (e) LP spectrum and (f) residual spectrum for LP order 30.

(Fig. 2(b)) shows significant information of the spectral envelope. On the other hand, if a large LP order is used, then the LP spectrum contains several spurious peaks as shown in Fig. 2(e) for $p = 30$.

These spurious peaks affect the residual when the speech signal is passed through the corresponding inverse filter.

When proper LP order is used, the LP residual mostly contains the excitation source information. Among the different categories of excitation, it is conjectured that voiced excitation contains significant speaker-specific information, as the corresponding glottal vibrations may be distinct for a given speaker (Plumpe et al., 1999). The differences in the rate of glottal vibration, shape of the glottal pulse and strength of excitation may be attributed to the speaker characteristics. The strength of excitation depends on the rate at which the glottal closure takes place. The strength is indicated approximately by the large residual error around the instant of significant excitation in each pitch period (Smits and Yegnanarayana, 1995). In the next section we discuss extraction of the speaker-specific excitation information present in the LP residual using AANN models.

3. AANN models for capturing speaker-specific information

Since LP analysis extracts the second order statistical features through the autocorrelation coefficients, the LP residual does not contain any significant second order relations corresponding to the shape of the vocal tract. That is why the autocorrelation function of the LP residual has low correlation values for nonzero time lags, like that for a white noise process (Makhoul, 1975). We conjecture that the speaker-specific information may be present in some higher order relations among the samples of the residual signal. It is not clear how this information can be extracted from the residual signal. Statistical features like higher order moments of the distribution of samples of the residual do not seem to capture the desired speaker-specific information. It is conjectured that extraction of such an information may involve nonlinear processing. Since neural network models can be trained to capture the nonlinear information present in the signal, we explore these models in this study. In particular, we explore AANN models to extract the desired information from the residual samples (Yegnanarayana et al., 2001; Yegnanarayana et al., 2005). The extraction of speaker-specific excitation information from the LP residual using AANN models was first demonstrated in (Yegnanarayana et al., 2001) for text-independent speaker recognition. Also, in (Yegnanarayana et al., 2005) the same study was demonstrated for text-dependent speaker recognition. In both these studies a fixed LP order and small database of about 20–30 speakers were used. The present work differs from these earlier studies in the following ways: (i) the effect of LP order on the manifestation of speaker-specific excitation information in the LP residual is studied, (ii) requirement of significantly less amount of data for speaker recognition using speaker-specific excitation information from the LP residual and AANN models is demonstrated, and (iii) complementary nature of speaker-specific excitation information is demonstrated on a large standard database.

AANN models are basically feed forward neural network (FFNN) models which try to map an input vector onto itself, and hence the name autoassociation or identity mapping (Yegnanarayana, 1999; Haykin, 1999). It consists of an input layer, an output layer and one or more hidden layers. The number of units in the input and output layers are equal to the size of the input vectors. The number of nodes in the middle hidden layer is less than the number of units in the input or output layers. The middle layer is also the dimension compression hidden layer. The activation function of the units in the input and output layers are linear, whereas the activation function of the units in hidden layer can be either linear or nonlinear. The performance of AANN models can be interpreted in different ways, depending on the problem and the input data. If the data is a set of feature vectors in the feature space, then the performance of AANN models can be interpreted either as linear and nonlinear principal component analysis (PCA) or distribution capturing of the input data (Diamantaras and Kung, 1996; Ikbal et al., 1999; Kishore and Yegnanarayana, 2001). On the other hand, if the AANN is presented directly with signal samples, such as speech signal, the network captures the implicit linear/nonlinear relations among the samples (Anjani et al., 2000; Reddy, 2001; Gupta, 2003). This can be used for speech enhancement if the input is noisy (Anjani et al., 2000), and for capturing the higher order relations among the samples in case the input is LP residual (Reddy, 2001; Gupta, 2003). This is because significant part of the second order relations among the samples in the speech signal are removed in the LP residual through LP analysis. Notice that if the input is only samples of noise, then the training error of the network does not reduce during training, indicating that there are no

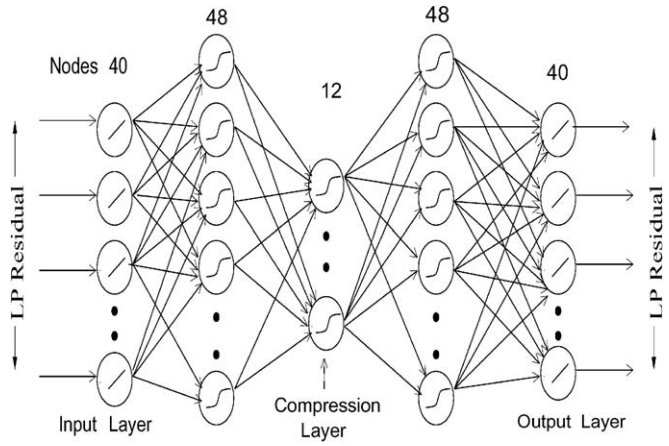

Fig. 3. Structure of AANN model used for capturing speaker-specific excitation features.

relations among the samples to capture. It is the interpretation of capturing the nonlinear relations among the samples, that we are exploiting in our studies to develop speaker models from LP residual (Reddy, 2001; Yegnanarayana et al., 2001; Gupta, 2003; Yegnanarayana et al., 2005).

A five layer AANN model with the structure shown in Fig. 3 is used. The structure of the network used in our study is $xL48N12N48NxL$, where x refers to the number of LP residual samples per frame, L refers to linear units and N to nonlinear units. A tanh(.) is used as the nonlinear activation function. The structure of the network was determined experimentally. The performance of the network does not depend critically on the structure of the network (Reddy, 2001; Gupta, 2003). When the input to an AANN consists of samples of random noise, then the network weights will not converge. On the other hand, if blocks of speech samples or LP residual samples are given as input, the error between the input (also the desired output) and the actual output is reduced during training, indicating that there is some relation among the samples. As the number (x) of LP residual samples per block is increased, then the relations over longer length of the block are captured. But, if the length of the block exceeds a pitch period, then the effect of pitch period also influences the training of the network. Therefore the number of samples per block are mostly limited to less than a pitch period. If the number of units in the dimension compression layer is large, then too many details in the input data may be captured, and these details may not be consistent across several blocks. If the number of units in the compression layer is very small (say 4 or 5), then important speaker-specific information may be missing. The training error is an indication of the minimum number of units required in the compression layer. Typically the training error reaches a low value when the number of units in the compression layer are increased to about 12, and thereafter the error does not significantly reduce even if the number of units are increased. Note that a lower number is preferable as it reduces the size (in terms of the weights) of the network.

4. Database used for the study

In general, speaker recognition refers to both speaker identification and speaker verification. Speaker identification is the task of identifying a given speaker from a set of speakers. In the closed-set speaker identification no speaker outside the given set is used for testing. Speaker verification is the task of verifying the identity claim of a given speaker. The result of speaker verification is either to accept or reject the claim of the speaker. In this paper we consider closed-set identification task for small data sets of 20 speakers each, in order to study the effects of various parameters in extracting the speaker-specific information. We consider speaker verification task on large standard database to study the complementary information of features of excitation source component and vocal tract system.

For speaker identification studies we use speech data collected over three different channels, namely, microphone, telephone and cellular phone. The microphone data was collected in the laboratory environment from 20 speakers. Speech data was collected over the same channel in two sessions for

each speaker. One minute of data was collected in each session. The data from one session was identified for training, and the other session data for testing. This data is referred as MIC (microphone) data throughout this study.

The telephone channel data was selected from the NIST 99 evaluation development database (Martin, 1999). The database contains 230 male and 309 female speakers. Among these, 80 male speakers were chosen at random, and four sets TEL1, TEL2, TEL3 and TEL4, each of 20 speakers, were formed. Each speaker's speech was collected over the same channel in different sessions. One minute of speech data from one of the sessions is used for training, and 1 min of data from other session for testing. The four sets of telephone data provide representative variations in the telephone channel.

Cellular phone data was chosen from the NIST 2001 evaluation development database (Martin, 2001). Out of the total 45 male speakers, 20 speakers were chosen at random to form CEL set. One minute of speech data is used for training and 1 min of data for testing. In all the cases the speech signal was sampled at 8 kHz sampling frequency.

Through out this study, small closed set identification experiments are done to demonstrate the feasibility of capturing the speaker-specific information separately from the system features and from the source features. The closed set identification studies are further used to examine the effect of different LP orders on the manifestation of speaker-specific excitation information in the LP residual. Requirement of significantly less amount data for the speaker recognition using speaker-specific excitation information and AANN models is also demonstrated using closed set identification studies. The speaker verification studies are used to demonstrate the complementary nature of the speaker-specific excitation information from the LP residual on large standard databases. For speaker verification studies, the complete set of NIST 2002 database is used (Martin, 2002).

5. Speaker identification studies

In this section speaker identification studies using LP residual and AANN models are described (Yegnanarayana et al., 2001; Gupta, 2003). In the present study the block of LP residual samples are normalized before using it for training and testing. The normalization is needed to avoid the large fluctuations of the signal amplitudes in different regions, such as in the weak and strong voiced sounds. The normalization involves dividing each sample in the block by the square root of the total energy of the signal in the block. The normalization does not destroy the implicit relations among the samples in the block, and it is these relations we hope to capture by training the AANN. In the training phase, one AANN model is trained separately for each speaker. During the testing phase, the models are used to decide the identity of the speaker for each test datum.

The block diagram for the training phase of the proposed speaker identification studies using LP residual is shown in Fig. 4. For both training and testing all the 1 min data for each speaker is used. The effective duration is only about 40 s, as only the high voiced speech data in each of the 1 min data is used for the study. A voiced and unvoiced detection is applied to each frame based on a pitch extraction algorithm (Prasanna and Yegnanarayana, 2004).

The LP residual of the speech signal is computed using an 8th order LP analysis. Blocks of 40 samples of the LP residual, corresponding to 5 ms of data,

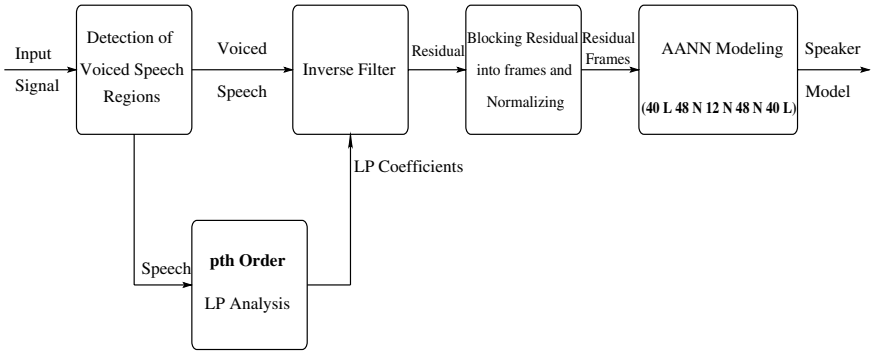

Fig. 4. Block diagram of training phase of speaker-identification system using LP residual.

are used as input to the AANN. Successive blocks are formed with a shift of one sample. There is a possibility that in case of some speakers samples from more than one pitch period may be included in the block of 40 samples. In such cases, the effect of pitch period is also included in the speaker-specific excitation information extracted from the LP residual. Each block of 40 samples is normalized before giving as input to the network. The weights of the network are initialized to random values in the range -1 to 1. The network is trained for 60 epochs using the back propagation learning algorithm (Haykin, 1999). The choice of the number of epochs for training is mostly dictated by the training error and the time taken for computation of the weights of the AANN. One AANN model is developed for each speaker.

The block diagram of the testing phase of the proposed speaker identification using LP residual is shown in Fig. 5. For testing, the LP residual is derived from the high voiced segments of the test speech data. Blocks of 40 samples of normalized LP residual are given as input. The output of each model is compared with its input to compute the squared error for each block. The error (E_i) for the ith block is transformed into a confidence value using $C_i = \exp(-\lambda E_i)$, where the constant value $\lambda = 1$ is used throughout this study. The confidence value will be larger for smaller values of the error, that is, for blocks matching with the corresponding models. The value of C_i will be low for large error value, thus giving less confidence to blocks not matching with their respective models. A given test utterance is compared with each of the speaker models to obtain the average confidence value $C = (1/N)\sum_{i=1}^{N} C_i$ for each model, where N is number of blocks in the test utterance. The average confidence value is used to evaluate the performance of the test utterances with respect to a given model.

The test data of each of the 20 speakers belonging to a particular set is tested against the models of all the 20 speakers in the set. The average confidence value for each of the 20 models for the given test data is computed, and this confidence value is used to rank the speaker models. Ideally, a genuine speaker should have the highest confidence value, and thus have rank one.

For comparison, a speaker identification system using the LP spectral features representing the vocal tract system is developed (Yegnanarayana et al., 2001; Kishore, 2001). The block diagrams for the training and testing phases are shown in Figs. 6 and 7, respectively. In this case the distribution of the feature vectors is used as speaker-specific informa-

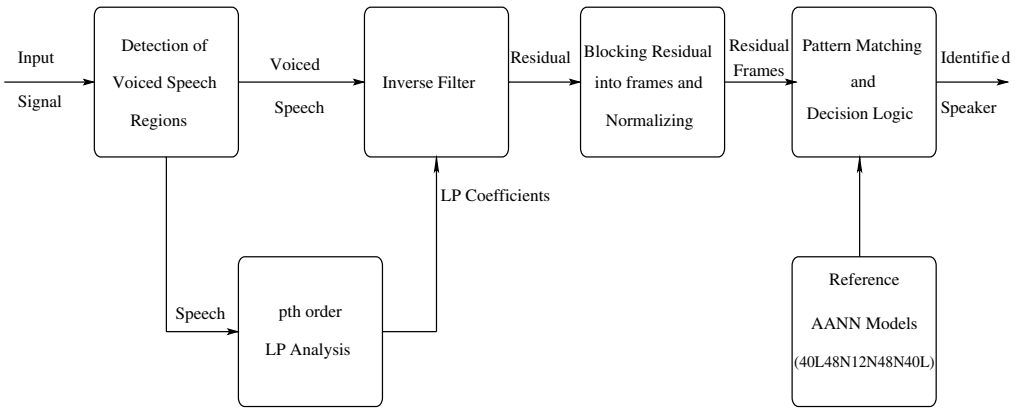

Fig. 5. Block diagram of testing phase of speaker-identification system using LP residual.

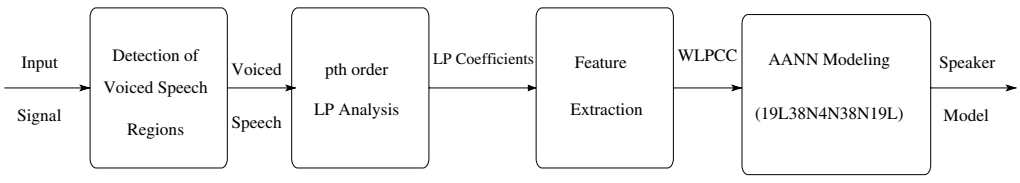

Fig. 6. Block diagram of training phase of speaker-identification system using system (LPCC) features.

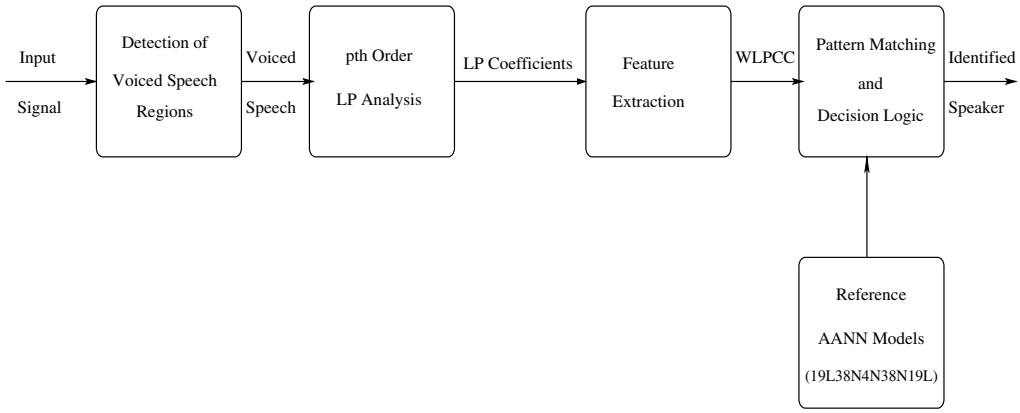

Fig. 7. Block diagram of testing phase of speaker-identification system using system (LPCC) features.

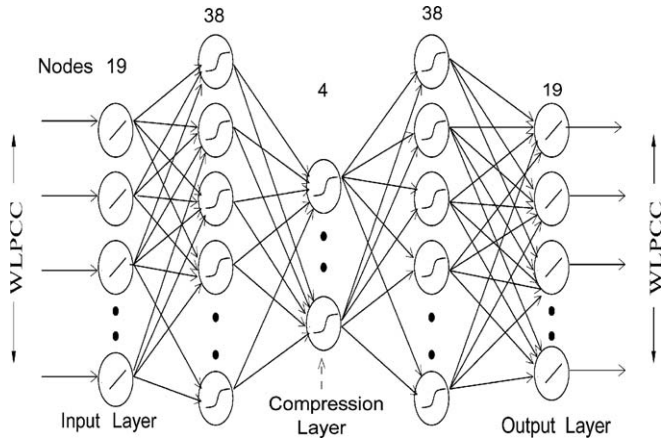

Fig. 8. Structure of AANN model used for capturing speaker-specific system features.

tion (Yegnanarayana et al., 2001; Kishore, 2001). The distribution is captured using an AANN model shown in Fig. 8. The model uses a feature vector consisting of 19-dimensional Weighted LP Cepstral Coefficients (WLPCC), which are derived from the 8th order LP analysis. The 19-dimensional LPCCs (c_ns) are derived from the 8 LP coefficients (a_ns) using the recursive relation between them (Deller et al., 2000). The inverse Fourier transform of LPCCs give the log LP spectrum (Deller et al., 2000). The larger the number of LPCCs, the better will be the representation of the LP spectrum. Also, since the values of the LPCCs (c_ns) are low for large values of the index n, the LPCCs are linearly weighted to derive the feature vector. Thus, for each frame, nc_n, $n = 1, 2, \ldots, 19$, is used as feature vector.

The distribution of the feature vectors is usually different for different speakers. Each AANN model is trained with feature vectors derived from the training data of the speaker. The feature vectors are computed for every frame of 20 ms, shifted by 10 ms. The model is trained using back-propagation learning algorithm for 60 epochs. Note that the AANN model (Fig. 8) used for capturing the distribution of spectral feature vectors in the 19-dimensional feature space is different both in structure and in training from the AANN model (Fig. 3) used earlier to capture the relations among the samples in the LP residual. Also the studies made here are for closed-set speaker identification for each of the 4 sets of 20 speakers.

The ranking of the speakers is done during testing. The speaker with highest confidence score is assigned rank 1, the second highest rank 2 and so on. The ranks of different speakers in both (source and system based) speaker identification systems for the data set TEL1 are given in Table 1. The dif-

Table 1
Performance of speaker recognition using source and system features

	Speaker no.																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Rank of Model 1	2	1	1	1	1	1	4	1	1	1	1	2	1	1	1	13	1	1	1	1
Rank of Model 2	1	1	1	1	1	2	1	1	1	8	1	1	1	1	1	1	1	1	1	1

The table shows the rank of the speaker obtained by matching with 20 speakers of TEL1 set. Model 1 refers to AANN models using source features and Model 2 refers to AANN models using system features.

Table 2
Performance of the identification systems based on system and source features for a set of 80 speakers

Type of speaker recognition system	# Models with	
	Rank = 1	Rank = 1&2
Model 1 (source features)	64/80	71/80
Model 2 (system features)	70/80	73/80

The system features are based on the weighted LPCC derived from 8th order LP analysis. The source features are based on the LP residual signal.

ferent ranks assigned by the two speaker identification systems for a given speaker in some cases may indicate the complementary nature of the speaker-specific information captured in the two systems. The performance of both speaker identification systems for the data sets TEL1, TEL2, TEL3 and TEL4 is summarized in Table 2. From the table it is evident that both excitation source features and vocal tract system features seem to give good performance. It is important to note that the source features are derived from the LP residual signal, which is obtained after removing the significant part of spectral envelope information.

6. Effect of LP order on speaker identification

In the speaker identification studies discussed in the previous section, an 8th order LP analysis was used. But the extent of speaker-specific excitation information that is present in the LP residual may depend on the order of the predictor in the LP analysis and this issue is studied in this section. The LP residual is extracted from the speech for a given LP order and one AANN model is trained for 60 epochs for each speaker. The extent of speaker-specific information in the LP residual can be understood from the trend of the training error. The training error curves of the AANN models for LP residuals obtained from different LP orders are shown in Fig. 9 for one speaker for the TEL1 data set. For reference, the training error curve for random noise sequence is also shown in the figure. For low LP orders (<8), the training error values are low. This is because, in these cases there is significant information about the vocal tract shape in the LP residual. Thus there is significant second order correlation information which the network tries to capture. For LP orders in the range 8–20, the LP residual contains mostly the information about the excitation source. The network thus tries to capture the speaker-specific information present in the excitation component. For high LP orders (>30), the training error is high. This may be because the spurious spectral nulls in the inverse filter may be affecting the speaker-specific information present in the excitation source component when the information is extracted from the LP residual. When the AANN model is trained with random noise sequence, the training error is high and also flat, indicating that no information is present in the data for the network to learn. One can attribute the low training errors for LP orders in the range 8–20 mainly to the speaker-specific information present in the excitation source component.

One can verify that the speaker-specific information in the LP residual is indeed captured by the proposed AANN models by testing them for each set of data as described earlier. The identification performance is shown as percentage in Figs. 10 and 11. In Fig. 10 the performance for the four telephone data sets is given. It shows the general trend, and at the same time it also shows the differences in the different telephone channels. For comparison of results of telephone data with microphone and cellphone data, only one of the telephone channels, namely TEL1, is used in all the following studies. Fig. 11 shows the performance for MIC, TEL1 and CEL data sets, each of 20 speakers. For lower LP orders (<8) the performance of the recognition system is poor. This is due to the fact that the LP residual has significant system features as explained earlier. For LP orders in the range 8–20, the recognition system gives good performance, as most of the information corresponding to the shape of the vocal tract system is removed.

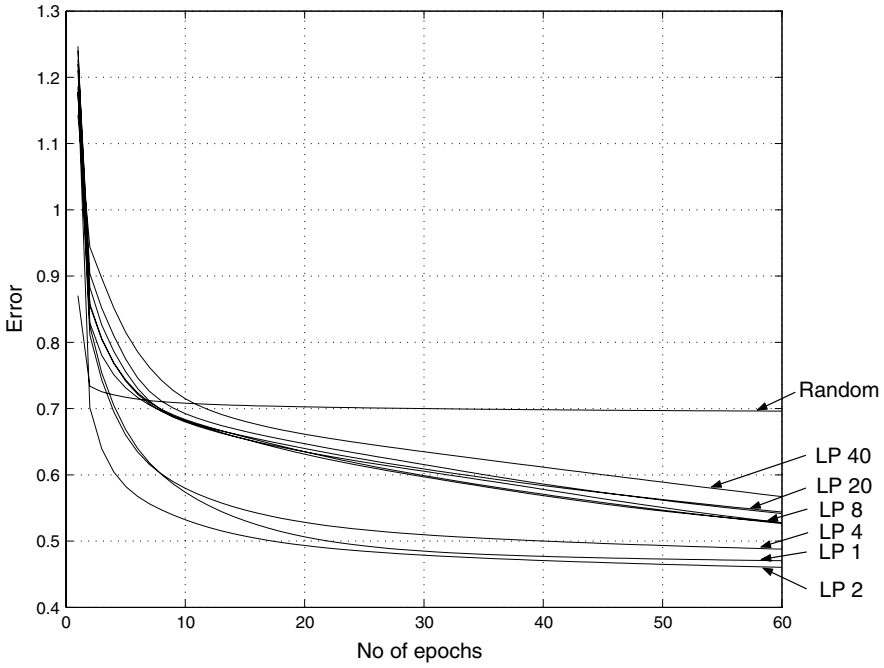

Fig. 9. Training error curves of AANN models for LP residuals extracted for different LP orders and random noise.

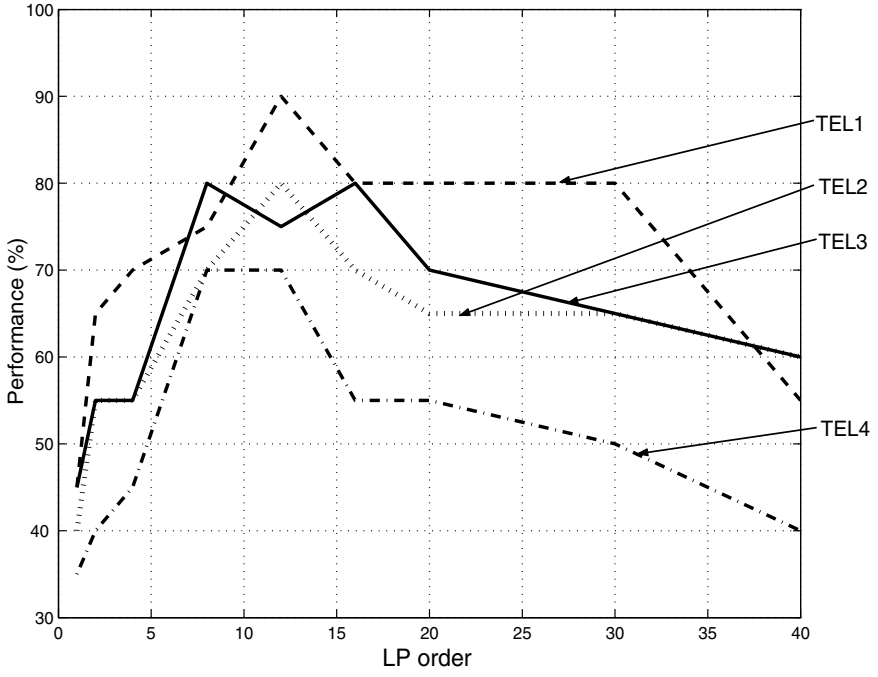

Fig. 10. Performance of the speaker recognition system based on speaker-specific source information for different LP orders. TEL1, TEL2, TEL3 and TEL4 are different telephone channels data. Each set has 20 speakers. One minute of data is used for training and 1 min of data for testing.

For LP orders greater than 30, the speaker-specific information in the LP residual is masked due to the effects of the spurious nulls in the spectrum of the inverse filter, and also due to the emphasis of high

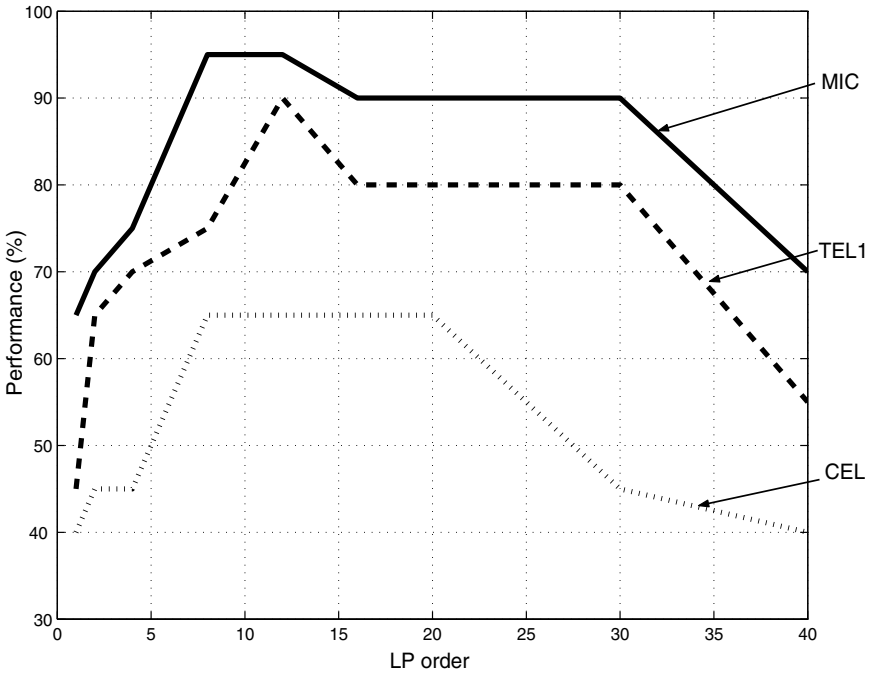

Fig. 11. Performance of the speaker recognition system based on speaker-specific source information for different LP orders. MIC, TEL1 and CEL refer to microphone, telephone and cellular data, respectively. Each set has 20 speakers. One minute of data is used for training and 1 min of data for testing.

frequency noise. Hence the performance is poor for large (>30) order of LP analysis.

From the above studies, we may conclude that the optimal range of the LP order for speaker recognition is in the range of 8–20 for speech signals sampled at 8 kHz. The variation of the peak performance for each set in Figs. 10 and 11 is due to the quality of the data in the set. Microphone (MIC) data set gives the best performance and cellular (CEL) data gives the worst performance.

A similar study was conducted to understand the presence of speaker-specific information in the vocal tract system features for different LP orders. The network structure used for the study is 19*L* 38*N*4*N*38*N*19*L*, which is same as described in the previous section. As shown in the network structure, only 19 weighted LPCCs are used as feature vectors irrespective of the order of the LP analysis. Note that the structure of this AANN model is different from the AANN structure used earlier to capture the information in the excitation component.

The speaker models built using the weighted LPCC features for different LP orders are tested as described in the previous section. The performance of the system for different LP orders is shown in Fig. 12. For low LP orders (<8), the performance of the system is low, as it cannot capture the speaker-specific information present in all the resonances of the vocal tract system. For LP orders in the range 8–20, the speaker-specific information is best represented in the weighted LPCC features, and hence the performance is high. For high LP orders (>30), the performance is again low due to the presence of spurious peaks in the LP spectrum (equivalently spurious nulls in the spectrum of the inverse filter).

From these studies we may conclude that the optimal range of LP order for speaker recognition using speech signals sampled at 8 kHz is 8–20. It is interesting to note that intuitively we feel that for low LP orders (<8), the missing speaker-specific information due to the vocal tract system information is present in the LP residual, and hence can be captured by the model. But in fact the presence of the vocal tract information in the LP residual degrades the performance. This is because of the differences in the way the speaker-specific information is captured by the two types of AANN models. The speaker-specific information is best represented either in the vocal tract system features or in the excitation source features, when the LP order is in the optimal range of 8–20. It is also important to note that the degradation in speech data used for

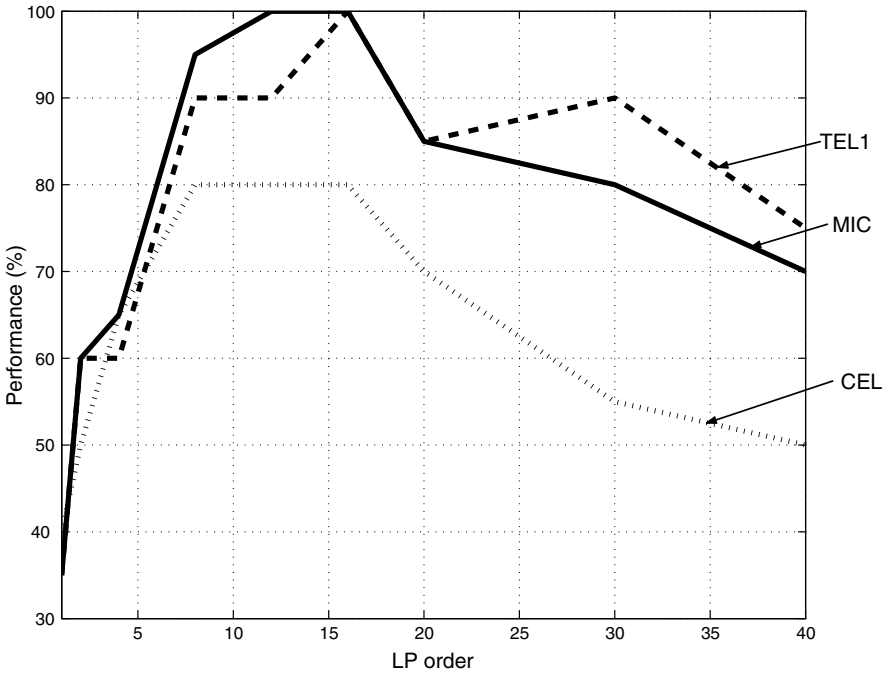

Fig. 12. Performance of the speaker recognition system based on vocal tract system information for different LP orders. MIC, TEL1 and CEL refer to microphone, telephone and cellular data, respectively. Each set has 20 speakers. One minute of data is used for training and 1 min of data for testing.

training and testing can affect the performance of the speaker identification system. Hence the performance is poor for noisy CEL data set, compared to MIC data set. The low performance for CEL data set may also be due to the compression code used to represent the speech data for transmission over cellular phones. In all the following studies unless specified, we use an LP order of 8.

7. Effect of size of data for speaker identification

Traditionally speaker identification systems based on the vocal tract system features follow statistical approach (Atal, 1976; Rosenberg, 1976; O'Shaughnessy, 1986; Furui, 1996; Furui, 1997; Campbell, 1997; Reynolds et al., 2000). The statistical methods capture the speaker variability in terms of the Probability Density Function (PDF) of the feature vectors of the speaker in the feature space. The performance of these systems depends on the amount of data available both for training and testing. If the data available is small, the distribution of the feature vectors in the feature space is sparse, and hence the recognition performance is poor during testing. In the proposed speaker identification system based on *source* features, the speaker-specific information is captured in terms of the higher order relations present among the samples of the residual signal, and not in terms of the PDF of the feature vectors of the speaker.

In the speaker recognition studies discussed so for, 1 min (effectively about 40 s) of speech data was used for generating the speaker models. Using different amounts of training data for generating the speaker models, the effect of size of the training data on the performance of the system can be studied. All the systems are evaluated independently using 1 min (effectively about 40 s) test data as explained earlier. The results are shown in Fig. 13. It is evident from the figure that about 6 s of data (that is, 6 s of voiced speech data extracted from the beginning of 1 min of training data) is enough for capturing the speaker-specific information. This is because the speaker-specific information in the LP residual depends less critically on the spectrum of the sound unit, as the spectral envelope information corresponding to the sound unit is removed in the residual. In the rest of the paper, unless otherwise specified, only 6 s of high voiced speech data is used for training the models using LP residual.

To examine the effect of size of the test data on the performance of the system, we consider different

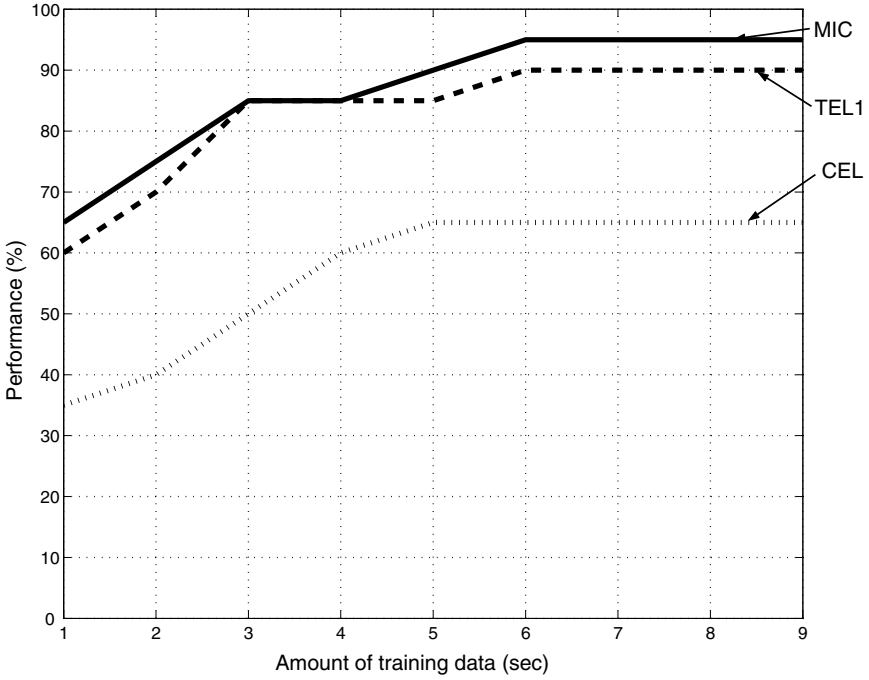

Fig. 13. Performance of the proposed speaker recognition system based on source features for different sizes of training data using 8th order LP analysis. Different amounts of training data (1–9 s) is used and testing was done with 1 min data.

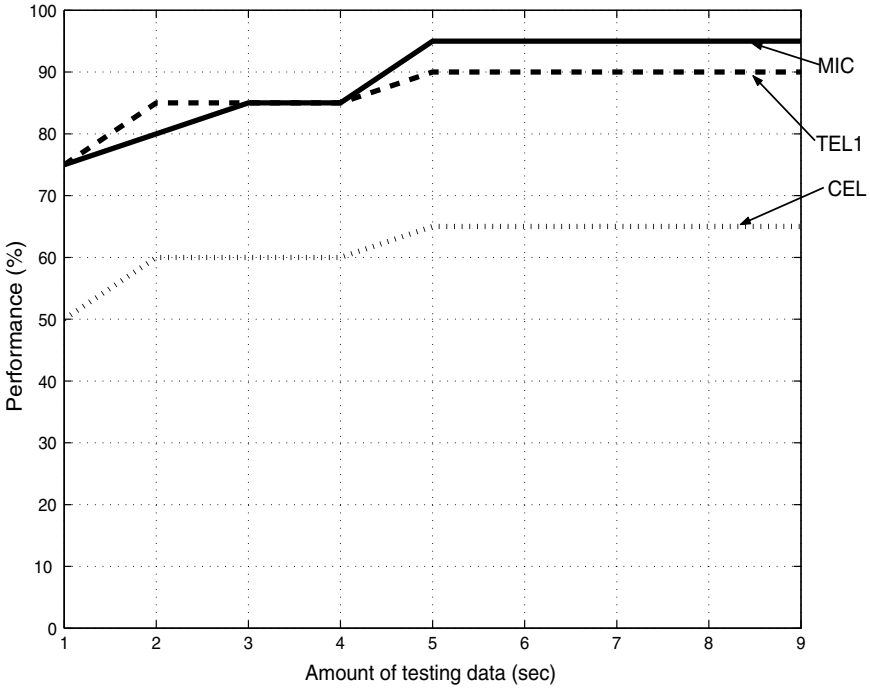

Fig. 14. Performance of the proposed speaker recognition system based on source features for different sizes of testing data using 8th order LP analysis. Models are trained with 6 s of high voiced speech and tested with different amounts of testing data (1–9 s).

cases, each using different amount of test data. All models are trained with 6 s of speaker data for this study. The performance variations with respect to the amount of the test data is shown in Fig. 14.

From the figure it can be seen that about 5 s of voiced speech data extracted from the beginning of 1 min of test data is sufficient for testing the models in this case. Because of this from now onwards

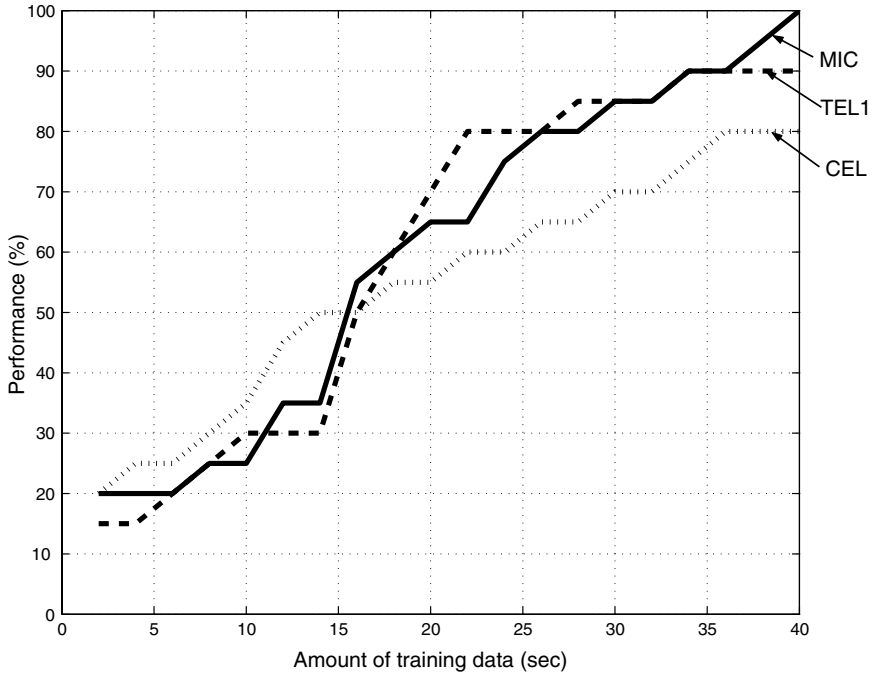

Fig. 15. Performance of the speaker recognition system based on system features for different sizes of training data 8th order LP analysis. Different amounts of training data (1–40 s) is used and testing was done with 1 min data.

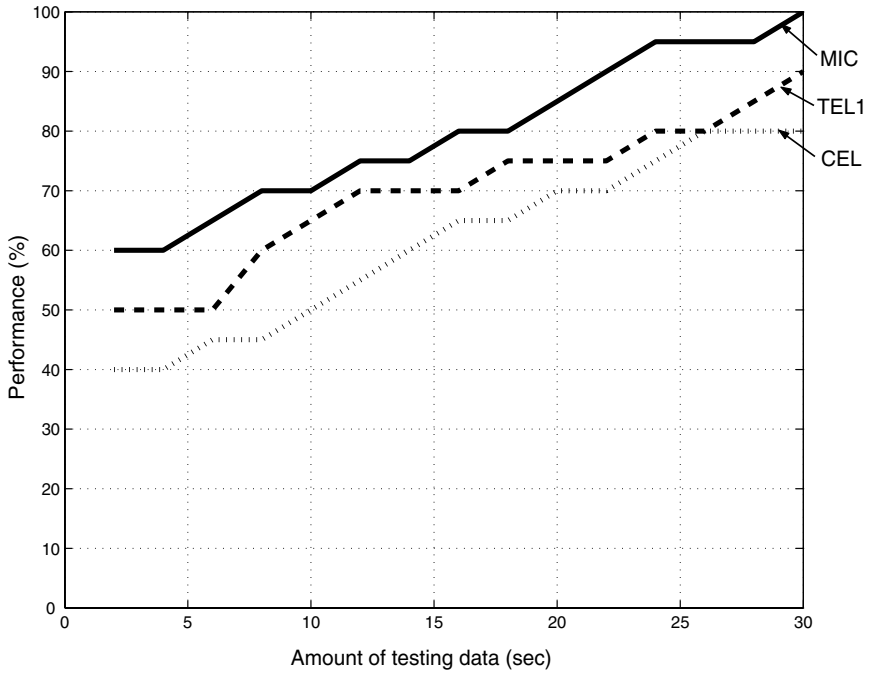

Fig. 16. Performance of the speaker recognition system based on system features for different sizes of testing data using 8th order LP analysis. Models are trained with 6 s of high voiced speech and tested with different amounts of testing data (1–30 s).

for all the studies involving LP residual about 6 s of voiced speech data is used both for training and testing the models.

For comparison, results obtained for different amounts of training and test data for speaker identification based on system features are shown in Figs. 15 and 16. The results show significant reduction in the performance when the quantity of the data is reduced. This is because the distribution of the system features of a speaker can be captured well only when there is sufficient amount of data representing all the different types of sound units both during training and testing.

From this study we can conclude that the proposed speaker identification system based on the source features requires less amount of data, as compared to the identification system based on system features. Hence the performance of the identification system can be improved by generating more than one model for each speaker for a given amount of data, and likewise more tests can be made using different test segments from a given test data. It may be possible to combine the evidence from various combinations of these models and tests for taking a decision.

8. Speaker verification studies on a large database

The proposed method for extracting speaker-specific information from the LP residual is evaluated on NIST 2002 speaker recognition evaluation database (Martin, 2002). As mentioned earlier, the studies on large database discussed in this section are for speaker verification task. The training data for each target speaker is about 2 min (110–130 s) of speech from a single conversation. The duration of the test segment varied from 15 to 45 s. The database has 3570 test utterances, where each test utterance is to be evaluated against 11 speaker models, among which one may be genuine speaker and the other ten are impostor speakers. Thus there are 39,270 test trials in the speaker verification task.

The objective of this study is mainly to demonstrate that some complementary information is available in the excitation source component of speech, which can be exploited for improving the performance of speaker verification systems developed using system (spectral) features. The speaker verification studies using the speaker-specific excitation information from the LP residual is conducted as follows: To reduce the computation time, a 6th order LP analysis is used to derive the LP residual from down sampled (4 kHz) speech signal. Only the first 6 s of the voiced speech data derived from the available (approximately 2 min) training data is used to train the neural network model for each speaker. A block size of 20 samples and a block shift of one sample are used both for training and testing. The network structure is $20L16N5N16N20L$. The optimal order (6) for LP analysis and the structure of the AANN model for the LP residual were obtained by conducting a separate set of experiments on the down sampled data. For testing, the first 6 s of voiced speech data from the available (15–45 s) test data is used for each speaker. For each block of test data, the mean squared error between the input block values and the output of the speaker's model is computed. The error is transformed to the confidence value. For the same block of test data the confidence values are obtained from 20 background models also (Yegnanarayana et al., 2002). The background models are the speaker models derived from the NIST 2001 speaker recognition evaluation database (Martin, 2001). The confidence value of the speaker model is normalized using the mean and standard deviation of the confidence values obtained for the background models. The average of the normalized confidence values of all the blocks in the test utterance was given as the score of the speaker model.

For the 3570 test utterances there are totally 39,270 (3570×11) test trials. Testing was done for all the trails and the scores obtained from these trials were submitted for NIST 2002 speaker recognition evaluation as IITM2 system (Yegnanarayana et al., 2002). The performance of the system evaluated using the key released by the NIST 2002 evaluation for the primary evaluation condition (Martin, 2002) and different recording conditions like INside building (IN), OUTside building (OUT), and inside VEHicle (VEH) are shown in Table 3. One of the metric used for the evaluation was the equal error rate (EER), which is the point at which the false acceptance and false rejection rates are equal. The results show that the performance of the system based on the LP residual alone is poor compared to the state of the art system submitted for the evaluation (Martin, 2002).

A separate speaker recognition system based on the system features was also developed, and the results of the same were submitted as IITM1 system for NIST 2002 evaluation (Yegnanarayana et al., 2002). Weighted LPCCs (WLPCCs) were used as feature vectors. The WLPCCs of the voiced frames

Table 3
Performance of the speaker recognition systems evaluated on NIST 2002 database

Type of system	Primary	IN	OUT	VEH
IITM2	23.8	25.8	19.9	21.6
IITM1	17.2	18.6	14.8	17.6
IITM3	15.2	15.8	11.4	14.2
OGI1	8.6	9.1	8.5	10.8
OGI1 + IITM1	8.0	8.6	7.7	10.8
OGI1 + IITM2	7.8	8.3	8.1	10.8
OGI1 + IITM1 + IITM2	7.1	8.1	6.9	9.0

IITM2 refers to the speaker recognition system based on speaker-specific information derived from the LP residual. IITM1 is the speaker recognition system based on the spectral features and AANN models. IITM3 is a combination of IITM1 and IITM2 systems. OGI1 is the speaker recognition system based on the spectral features and UBM–GMM models (Kajarekar et al., 2002; Kajarekar et al., 2003). The evaluation condition is called primary, and the different recording conditions are INside building (IN), OUTside building (OUT) and inside VEHicle (VEH). The performance is given in terms of % EER.

from the training data of about 2 min duration were used as feature vectors for building speaker models. AANN models ($19L38N4N38N19L$) were used to capture the distribution of the feature vectors for each speaker. During testing the voiced frames from the entire 1 min data was used. The mean squared error between each test frame and the output of the speaker model was computed. The average confidence value for the test utterance was normalized both with respect to the model and the test utterance (Yegnanarayana et al., 2002). The scores of the complete 39,270 test trials were submitted for evaluation. The performance of the IITM1 system computed using the key from NIST 2002 evaluation is tabulated in Table 3. The performance of the IITM1 system is better compared to the IITM2 system. But the complementary nature of the information in the IITM2 system is evident in the IITM3 system, where the IITM1 and IITM2 are combined by adding the scores.

For comparison, the results of the Oregon Graduate Institute system (OGI1) which used melfrequency cepstral coefficients as feature vectors and based on Universal Background Model–Gaussian Mixture Model (UBM–GMM) framework for NIST 2002 evaluation (Kajarekar et al., 2002) is considered in the present study. The block diagrams for the training and testing phases of the OGI system are shown in Figs. 17 and 18, respectively. The performance of this system is also shown in Table 3. The details of the OGI1 system are given

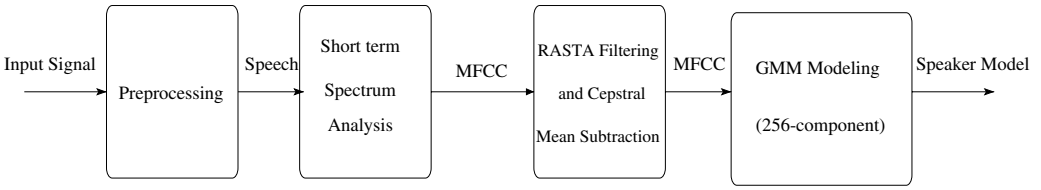

Fig. 17. Block diagram of training phase of speaker-verification system using system (MFCC) features developed at OGI.

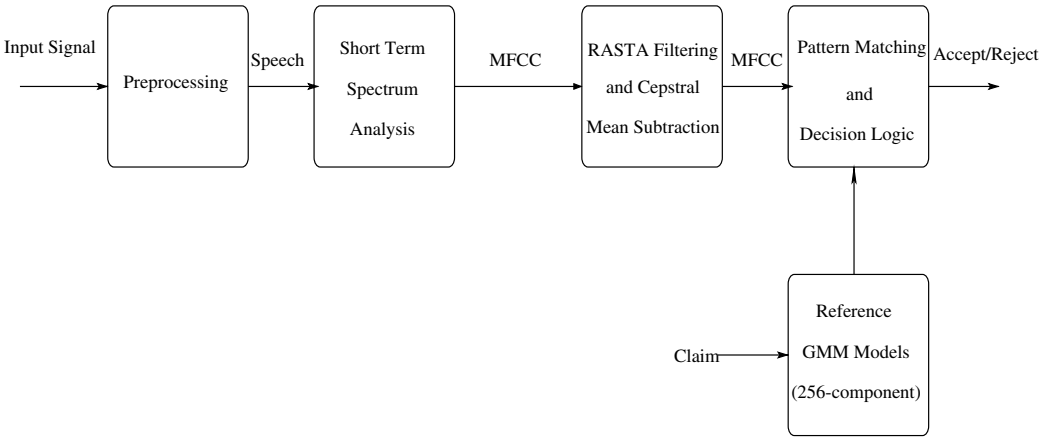

Fig. 18. Block diagram of testing phase of speaker-verification system using system (MFCC) features developed at OGI.

in (Kajarekar et al., 2002; Kajarekar et al., 2003). The superior performance of the OGI system is due to the use of 39 dimension feature vector, selection of specific sound units and significantly improved normalization techniques. The feature vector consists of 13 melcepstral coefficients, 13 delta melcepstral coefficients and 13 delta–delta melcepstral coefficients.

In the post-evaluation studies, we provided scores of IITM1 and IITM2 systems for OGI to explore methods for combining the scores. The results of combining the scores (simple weighted sum) are reproduced in Table 3 (Kajarekar et al., 2003). Combining the scores from the proposed IITM1 and IITM2 systems with OGI1 system improves the performance significantly. This shows that the speaker-specific information derived from the LP residual and from the weighted LPCC seems to have some information complementary to the speaker-specific information present in the OGI1 system.

9. Summary and conclusions

In this work we have demonstrated the importance of information in the excitation component of speech for speaker recognition task. Linear prediction residual was used to represent the excitation information. Performance of the recognition experiments show that AANN models can capture some speaker-specific excitation information from the LP residual. Performance of the system for different orders of LP analysis shows that the optimal range for the LP order is 8–20 for speech signals sampled at 8 kHz. The speaker-specific excitation information may be present in the higher order relations among the samples of the LP residual. The recognition performance depends on the number of samples selected for each block to capture the speaker-specific excitation information. Larger the number, the better is the performance, although smaller number reduces computational complexity due to smaller size of the AANN model.

Presently, we are taking the average of the confidences of all the blocks for evaluating the performance of the speaker recognition system. A better approach would be to determine suitable weighting for each block, depending on the nature of the speech signal in that block. The amount of training as well as testing data required in the case of the proposed speaker recognition system is significantly less compared to the recognition systems based on vocal tract system features. Hence for the same amount of data, we can have multiple models and multiple test segments, providing multiple evidences to take a decision about the speaker. This may improve the recognition performance for a given amount of data.

The objective in this paper was mainly to demonstrate the significance of the speaker-specific excitation information present in the linear prediction residual for speaker recognition. We have not made any attempt to optimize the parameters of the model used for feature extraction, and also the decision making stage. Therefore the performance of speaker recognition may be improved by optimizing the various design parameters.

References

Ananthapadmanabha, T.V., Yegnanarayana, B., 1979. Epoch extraction from linear prediction residual for identification of closed glottis interval. IEEE Trans. Acoust. Speech Signal Process. ASSP-27, 309–319.

Anjani, A.V.N.S., 2000. Autoassociate neural network models for processing degraded speech. MS thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai 600 036, India.

Atal, B.S., 1972. Automatic speaker recognition based on pitch contours. J. Acoust. Soc. Amer. 52 (6), 1687–1697.

Atal, B.S., 1976. Automatic recognition of speakers from their voices. Proc. IEEE 64 (4), 460–475.

Campbell, J.P., 1997. Speaker recognition: a tutorial. Proc. IEEE 85, 1436–1462.

Deller Jr., J.R., Hansen, J.H.L., Proakis, J.G., 2000. Discrete-Time Processing of Speech Signals. IEEE Press, New York.

Diamantaras, K.I., Kung, S.Y., 1996. Principal Component Neural Networks: Theory and Applications. John Wiley and Sons, New York.

Doddington, G., 2001. Speaker recognition based on idiolectal differences between speakers. In: Proc. European Conf. on Speech Processing, Technology (EUROSPEECH), Aalborg, Denmark, pp. 2521–2524.

M. Faundez, D. Rodriguez, 1998. Speaker recognition using residual signal of linear and nonlinear prediction models. In: Proc. Internat. Conf. on Spoken Language Processing.

Feustel, T.C., Velius, G.A., Logan, R.J., 1989. Human and machine performance on speaker identity verification. Speech Technol., 169–170.

Furui, S., 1996. An overview of speaker recognition technology. In: Lee, C.H., Soong, F.K., Paliwal, K.K. (Eds.), Automatic Speech and Speaker Recognition. Kluwer Academic, Boston, Chapter 2.

Furui, S., 1997. Recent advances in speaker recognition. Pattern Recognition Lett. 18, 859–872.

Gupta, C.S., 2003. Significance of source features for speaker recognition. MS thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai 600 036, India.

Haykin, S., 1999. Neural Networks: A Comprehensive Foundation. Prentice-Hall Inc., New Jersey.

Ikbal, M.S., Misra, H., Yegnanarayana, B., 1999. Analysis of autoassociative mapping neural networks. In: Proc. Internat. Joint Conf. on Neural Networks (IJCNN), USA, pp. 854–858.

Kajarekar, S.S., Adami, A.G., Hermansky, H., 2002. Ogi submission – NIST 2002 one-speaker detection task. In: Proc. NIST Speaker Recognition Workshop, Vienna, VA, USA.

Kajarekar, S.S., Adami, A.G., Hermansky, H., 2003. Novel approaches for one and two-speaker detection. In: Proc. European Conf. on Speech Processing, Technology (EUROSPEECH).

Kishore, S.P., 2001. Speaker verification using autoassociative neural network models. MS thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai 600 036, India.

Kishore, S.P., Yegnanarayana, B., 2001. Online text-independent speaker verification system using autoassociative neural network models. In: Proc. Internat. Joint Conf. on Neural Networks (IJCNN), Washington, DC, USA.

Liu, J.H.L., Palm, G., 1997. On the use of features from prediction residual signal in speaker recognition. In: Proc. European Conf. Speech Processing, Technology (EUROSPEECH), pp. 313–316.

Madhukumar, A.S., 1993. Intonation knowledge for speech systems for an Indian language. Ph.D. thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai 600 036, India.

Makhoul, J., 1975. Linear prediction: a tutorial review. Proc. IEEE 63, 561–580.

Martin, A., 1999. NIST 1999 speaker recognition evaluation plan. In: Proc. NIST Speaker Recognition Workshop, USA.

Martin, A., 2001. NIST 2001 speaker recognition evaluation plan. In: Proc. NIST Speaker Recognition Workshop, USA.

Martin, A., 2002. NIST 2002 speaker recognition evaluation plan. In: Proc. NIST Speaker Recognition Workshop, USA.

Murthy, K.S.R., Prasanna, S.R.M., Yegnanarayana, B., 2004. Speaker-specific information from residual phase. In: Internat. Conf. on Signal Processing and Communications, Bangalore, India, pp. 516–519.

O'Shaughnessy, D., 1986. Speaker recognition. IEEE ASSP Mag. 3, 4–17.

O'Shaughnessy, D., 1987. Speech Communication: Human and Machine. Addison-Wesley, New York.

Plumpe, M.D., Quatieri, T.F., Reynolds, D.A., 1999. Modeling of the glottal flow derivative waveform with application to speaker identification. IEEE Trans. Speech Audio Process. 1, 569–586.

Prasanna, S.R.M., Yegnanarayana, B., 2004. Extraction of pitch in adverse conditions. In: Proc. IEEE Internat. Conf. on Acoust., Speech, Signal Processing (ICASSP), Montreal, Canada.

Rabiner, L.R., Juang, B.H., 1993. Fundamentals of Speech Recognition. Prentice-Hall, Englewood Cliffs, NJ.

Reddy, K.S., 2004. Source and system features for speaker recognition. MS thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai 600 036, India.

Reynolds, D.A., Quateri, T.F., Dunn, R.B., 2000. Speaker recognition using adapted gaussian mixture models. Digital Signal Process. 10, 19–41.

Rosenberg, A.E., 1971. Effect of glottal pulse shape on the quality of natural vowels. J. Acoust. Soc. Amer. 49, 583–590.

Rosenberg, A.E., 1976. Automatic speaker verification: a review. Proc. IEEE 64 (4), 475–487.

Smits, R., Yegnanarayana, B., 1995. Determination of instants of significant excitation in speech using group delay function. IEEE Trans. Speech Audio Process. 3, 325–333.

Thevenaz, P., Hugli, H., 1995. Usefulness of lpc residue in text-independent speaker verification. Speech Commun. 17, 145–157.

Wakita, H., 1976. Residual energy of linear prediction to vowel and speaker recognition. IEEE Trans. Acoust. Speech Signal Process. 24, 270–271.

Weber, F., Manganaro, L., Peskin, B., Shriberg, E., 2002. Using prosodic and lexical information for speaker identification. In: Proc. IEEE Internat. Conf. on Acoust. Speech Signal Processing (ICASSP), Orlando, FL, USA, pp. 141–144.

Yegnanarayana, B., 1999. Artificial Neural Networks. Prentice-Hall, New Delhi, India.

Yegnanarayana, B., Madhukumar, A.S., Ramachandran, V.R., 1992. Robust features for applications in speech and speaker recognitions. In: Proc. ESCA Workshop on Speech in Adverse Conditions, Cannes Mandelieu, France.

Yegnanarayana, B., Reddy, K.S., Kishore, S.P., 2001. Source and system features for speaker recognition using AANN models. In: Proc. IEEE Internat. Conf. on Acoust. Speech Signal Processing (ICASSP), Salt Lake City, Utah, USA, pp. 409–412.

Yegnanarayana, B., Prasanna, S.R.M., Gangashetty, S.V., Gupta, C.S., Guruprasad, S., Dhananjay, N., 2002. IIT Madras speaker verification system. In: Proc. NIST Speaker Recognition Workshop, Vienna, VA, USA.

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. IEEE Trans. Speech Audio Process. 13 (4), 575–582.