Volume 3, Issue 11, November 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Performance Evaluation of Different Modelling Techniques for Text-Independent Speaker Recognition System

R. Rajeswara Rao*
Department of CSE,
JNTU Kakinada, India

Abstract— In this paper, through different experimental studies we demonstrate that the excitation component of speech can be exploited for text independent 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 Ergodic Hidden Markov Models (EHMMs). 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 (EHMMs). It is demonstrated that the proposed speaker recognition system using excitation information and EHMMs requires significantly less amount of data both during training as well as testing, compared to other speaker recognition systems, which uses other information. We demonstrate the speaker recognition studies on TMIT database.

Keywords— Ergodic Hidden Markov Models (EHMMs), LPC, MFCC.

I. INTRODUCTION

Speech is produced from a time varying vocal tract system excited by a time varying excitation source [15][19][6]. 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, anti-resonances and spectral roll-off characteristics. These filter characteristics are usually 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 [3][21][15][9][10][4][19]. 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 [8] Incorporation of pitch and duration (prosody) information into speaker recognition system has also been studied [25]. 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 [26][11]. 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 [12]. 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 [2]. Wakita has reported an experiment using the LP residual energy for vowel recognition and also for speaker recognition [24]. 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 [8]. The use of cepstrum computed over the LP residual was also proposed for Speaker recognition [23]. 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. 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 [22]. 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 [22]. 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 [20][1][26] or by random noise [14]. It appears that significant speaker-specific excitation information may be present in the segmental and suprasegmental features of the residual. The present work focuses 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 [14]. 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.

II. SPEECH FEATURE EXTRACTION

The selection of the best parametric representation for acoustic data is an important task in the design of any text-independent speaker recognition system. The acoustic features should fulfill the following requirements. Be of low dimensionality to allow a reliable estimation of parameters of the Automatic speaker recognition systems. Be independent of the speech and recording environment.

A. Pre processing

The task begins with the pre-processing of the speech signal collected from each speaker. The speech signal is sampled at 16000 samples/sec. In the pre-processing stage, the given speech utterance is pre-emphasized, blocked into a number of frames and windowed. The frame size chosen is 25 ms which corresponds 400 samples and a frame shift between frames is 10 ms which corresponds to 160 samples has been taken. Before doing the pre-emphasis, a check is done to see whether the amplitude of the utterance is sufficient. Depending on the amplitude in each frame, the frame is weighted. The pre-processing task is carried out in a sequence of steps as explained below.

a. Amplitude Check

The given speech utterance should be checked to see that the amplitude of the samples in the speech frames are sufficient so that meaningful features can be extracted from it. The check on the amplitude of the given speech utterance is done as follows.

- The maximum amplitude of samples in each frame is obtained.
- The mean (μ) and standard deviation (σ) of these amplitudes is computed.
- If the $(\mu+3*\sigma)$ of this is less than a pre-determined threshold or the standard deviation is less than a certain minimum value, then the speaker is requested to speak again.

A statistical analysis of the frames has shown that the threshold required for the amplitude is around 500. The minimum variation level for the amplitude is around 150. These two factors are necessary to identify the speech frames in the presence of noise.

b. Frame Weighting

The speech frames should be classified into frames containing speech information and non-speech frames, so that only speech frames are used to extract the features. Weighting of the frames is done as follows.

- Using the mean and standard deviation values obtained in the previous step, the maximum limit for amplitudes is computed as $a_{\max} = \mu + \sigma$ and minimum limit as $a_{\min} = a_{\max} * 0.25$.
- Frames with the maximum amplitudes less than $a_{\min}$ are declared as silence frames for a given weight of 0.0. Those with amplitude greater than or equal to $a_{\max}$ are given a value of 1.0.
- Those with amplitude in-between are linearly weighted based on the maximum amplitude in the frame.
- If the number of silence frames obtained is below a certain fraction of the total number of frames then the minimum limit set is changed to $1.1 * a_{\min}$ and the frame weights and the number of silence frames are recomputed. This step is repeated till the required number of silence frames is obtained. Studies show that atleast 10 % of the frames in the entire utterance which is being used presently can be classified as silence frames.

c. Pre-Emphasis

The given speech samples in each frame are passed through a first order filter to spectrally flatten the signal and make it less susceptible to finite precision effects later in the signal processing task. The pre-emphasis filter used has the

form $H(z)=1-az^{-1}$, $0.9 \le a \le 1.0$. In fact, it is sometimes better to difference the entire speech utterance before frame blocking and windowing.

d. *Windowing*

After pre-emphasis, each frame is windowed using a window function. The windowing ensures that the signal discontinuities at the beginning and end of each frame is minimized. The window function used is the Hamming window given below,

$$W(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1 \quad (1)$$

Where N is the number of samples in the frame.

III. SPEAKER CHARACTERISTICS IN THE LP RESIDUAL

Speech signals, as any other real world signals, are produced by exciting a system with source. A simple block diagram representation of the speech production mechanism is shown in the Fig.1. Vibrations of the vocal folds, powered by air coming from the lungs during exhalation, are the sound source for speech. Hence, as can be from Fig. 1, the glottal excitation forms the source, and the vocal tract forms the system. One of the most powerful speech analysis technique is the method of linear predictive analysis [10]. The philosophy of linear prediction is intimately related to the basic speech production model. The Linear Predictive Coding (LPC) analysis approach performs spectral analysis on short segments of speech with an all-pole modeling constraint [10].

Since speech can be modeled as the output of linear, time-varying system excited by a source, LPC analysis captures the vocal tract system information in terms of coefficients of the filter representing the vocal tract mechanism. Hence, analysis of speech signal by LP results in two components, namely the synthesis filter on one hand and the residual on the other hand. In brief, the LP residual signal is generated as a by product of the LPC analysis, and the computation of the residual signal is given below.

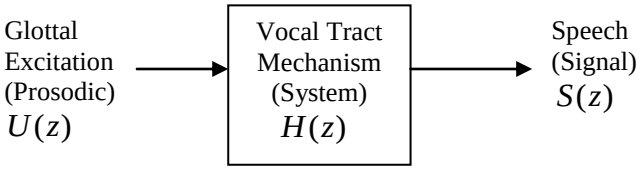


Fig. 1 Source and System representation of speech production mechanism.

If the input signal is represented by U_n and the output signal by S_n, then the transfer of the system can be expressed as,

$$H(z) = \frac{S(z)}{U(z)} \quad (2)$$

Where S(z) and U(z) are z-transforms of S_n and U_n respectively.

Consider the case where we have output signal and the system and have to compute the input signal. The above equation can be expressed as $S(z) = H(z)U(z)$

$$U(z) = \frac{S(z)}{H(z)} \quad (3)$$

$$U(z) = \frac{1}{H(z)} S(z) \quad (4)$$

$$U(z) = A(z)S(z) \quad (5)$$

Where $A(z) = 1/H(z)$ is the inverse filter representation of the vocal tract system.

Linear prediction models the output S_n as the linear function of past outputs and present and past inputs. Since prediction is done by a linear function, the name linear prediction. Assuming an all-pole for the vocal tract, the signal S_n can be expressed as linear combination of past values and some input U_n as shown below.

$$Sn = -\sum_{k=1}^{p} a_k S_{n-k} + GU_n \quad (6)$$

Where G is a gain factor.

Now assuming that the input U_n is unknown, the signal S_n can be predicted only approximately from a linear weighted sum of past samples. Let this approximation of S_n be $\tilde{S}_n$, where

$$\tilde{S}_n = \sum_{k=1}^{p} a_k S_{n-k} \quad (7)$$

Then the error between the actual value S_n and predicted value $\tilde{S}$ is given by $e_n = S_n - \tilde{S}_n$. This error e_n is nothing but LP residual of signal is shown in Fig 2.

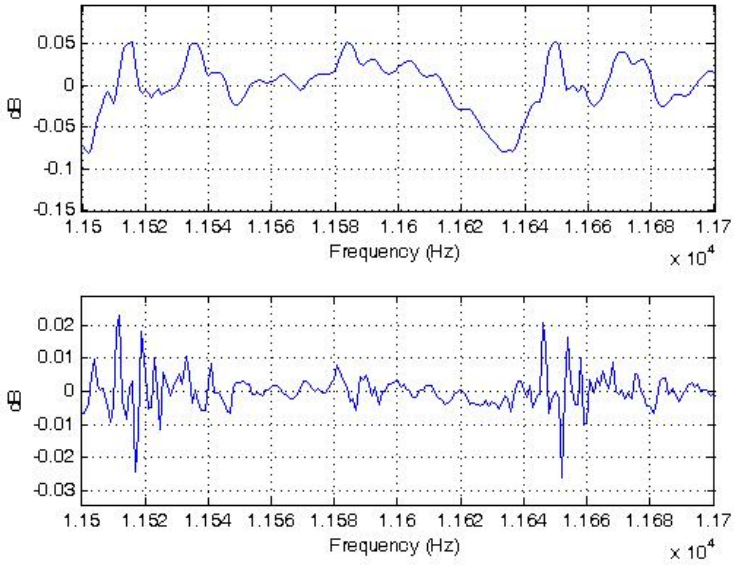
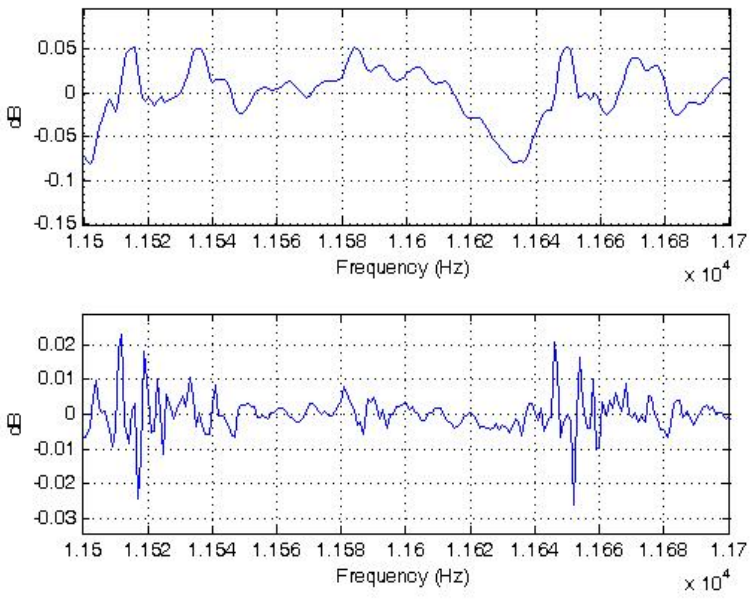


Fig. 2 Actual signal and its LP residual

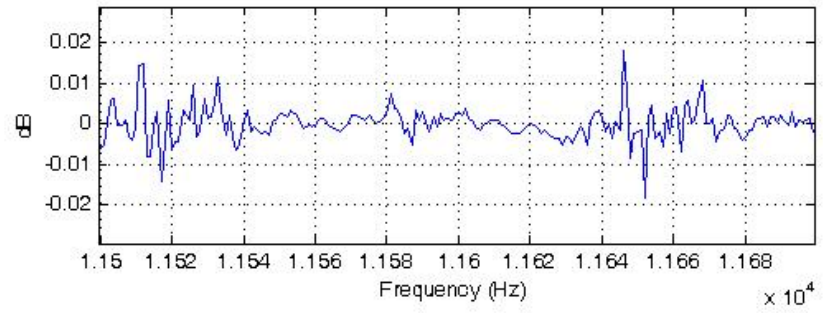
As the envelope of the short-time spectrum corresponds to the frequency response of the vocal tract shape, one can observe the short time spectrum of the LP residual for different LP orders and the corresponding signal LP spectra to determine the extent of the vocal tract information present in the LP residual. As the order of the LP analysis is increased, the LP spectrum approximates the short-time spectral envelope better. The envelope of the short-time spectrum corresponds to the frequency response of the vocal tract shape, thus reflecting the vocal tract system characteristics. For a low order, say 2, as shown in Fig.3, the LP spectrum may pick up only prominent resonance, and hence the residual will still have a large amount of information about the vocal tract system. Thus the spectrum of the residual contains most of the information of the spectral envelope. As LP order increases the information about the vocal tract system decreases as shown in Fig.2.

IV. FEATURE EXTRACTION OF LP RESIDUAL SIGNAL

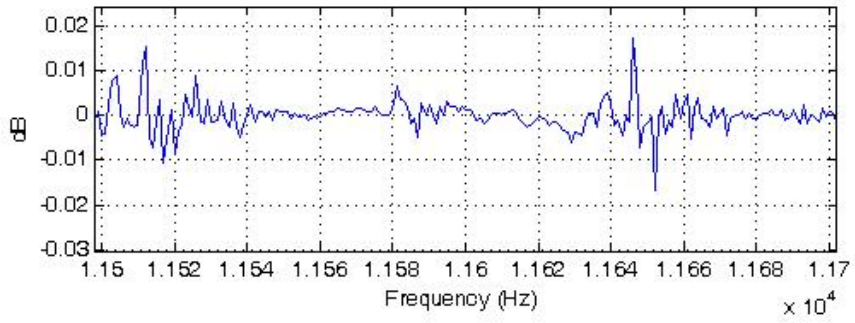
MFCC is the best known and most popular, and this feature has been used. MFCC's are based on the known variation of the human ear's critical bandwidths with frequency. The MFCC technique makes use of two types of filter, namely, linearly spaced filters and logarithmically spaced filters. To capture the phonetically important characteristics of speech, signal is expressed in the Mel frequency scale. This scale has a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. Normal speech waveform may vary from time to time depending on the physical condition of speakers' vocal cord. Rather than the speech waveforms themselves, MFFCs are less susceptible to the said variations [18].



(a)



(b)



(c)

Fig. 3. (a) Source information in LP residual of order 2
(b) Source information in LP residual of order 10
(c) Source information in LP residual of order 30.

A. Motivation to use Mel Frequency Ceptral Coeffien (MFCC)

Since our interest is in capturing global features which correspond to source excitation, the low frequency or pitch components are to be emphasized. To fulfil this requirement it is felt that MFCC are most suitable as they emphasize low frequency and de-emphasize high frequencies.

B. MFCC

In this phase the digital speech signal is partitioning into segments (frames) with fixed length 10-30 ms from which the features are extracted due to their spectral qualities. Spectrum is achieved with fast Fourier transformation [12]. Then an arrangement of frequency range to mel scale follows according to relation

$$f_{mel} = 2595 \log\left(1 + \frac{f_{Hz}}{700}\right) \qquad (7)$$

By logarithm of amplitude of mel spectrum and applying reverse Fourier transformation we achieve frame cepstrum:

$$mel - cepstrum(frame) = FFT^{-1}\left[mel(\log|FFT(frame)|)\right] \qquad (8)$$

The FFT-base cepstral coefficients are computed by taking IFFT of the log magnitude spectrum of the Speech signal. The mel-warped cepstrum is obtained by inserting a intermediate step of transforming the frequency scale to place less emphasis on higher frequencies before taking the IFFT [12][13][17].

V. PARAMETRIC APPROACHES

A. Ergodic Hidden Markov Model

The Hidden Markov Models (HMM) is a doubly embedded stochastic process where the underlying stochastic process is not directly observable. HMMs can be used as probabilistic speaker models for both text-dependent and independent speaker recognition. An HMM not only models the underlying speech sounds but also the temporal sequencing of the sounds. This temporal modelling is advantageous for text-dependent tasks. For text-dependent speaker recognition task, HMM-based methods have achieved significantly better recognition [15][16][17].

Since the stressful cues contained in an utterance cannot be assumed as specific sequential events in the signal, an ergodic or fully connected HMM structure becomes more appropriate than LeftToRight (LTR) structure because every state in the ergodic structure can be reached in a single step from every other state. An ergodic or fully connected that is derived from ergodic or fully connected HMM has been used in this work. The transition matrix, A, of this structure can be written in terms of the b_{ij} coefficients (positive coefficients) as,

$$A = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix}$$

In training phase, an HMM for each speaker is obtained by estimating the parameters of model using feature vectors from the training data. The parameters of HMM are:

State transition probability distribution: It is represented by A=$[a_{ij}]$ where

$$aij = P(q_{t+1} = j \mid q_t = i) \quad 1 \le i, j \le N$$

defines the probability of transition from state i to j at time t.

Observation symbol probability distribution: It is given by $B = b_j(k)$, in which

$$b_j(k) = P(O_t = V_k \mid q_t = j) \quad 1 \le k \le M$$

defines the symbol distribution in state j, j=1, 2, … N

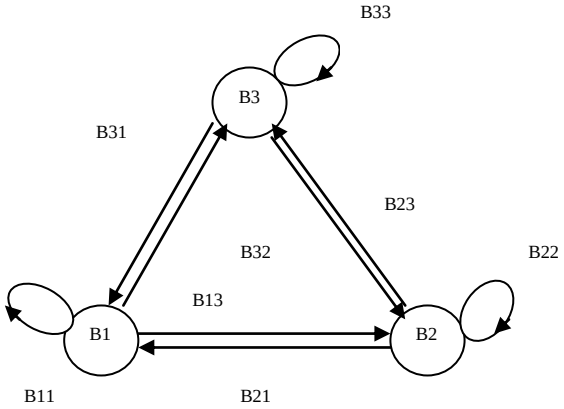


Fig.4 Three State Ergodic Hidden Markov Model

The initial state distribution: It is given by $\prod = [\pi]$, where

$$\pi_i = P(q_1 = i) \quad 1 \le i \le N$$

Here, N is the total number of states, and q_t is the state at time t. M is the number of distinct observation symbols per state, and O_t is the observation symbol at time t.

The model parameters can be collectively represented as $\lambda = \{A_i, B_i, \pi_i\}$ for i = 1… M. Each speaker in a speaker identification system can be represented by a HMM and is referred to by the speaker's respective models λ.

In the testing phase, $P(O \mid \lambda)$ for each model is calculated [17], where $O = (O_1 O_2 O_3 \ldots O_T)$ is the sequence of the test feature vectors. The goal is to find the probability, given the model, that the test utterance belongs to that particular model. The speaker model that gives the highest score is declared as the identified speaker.

B. Experimental Setup

The system has been implemented in Matlab7 on Windows XP platform. We have trained the HMMs using Gaussian Components as 2, 4, 8, and 16 by varying the HMM states from 2 to 4, and for training speech duration of 30 sec. Testing is performed using different test speech durations such as 1 sec., 2 sec., and 3 sec..

The steps involved in the proposed algorithm for text independent speaker recognition system are as follows:

Training Phase:

for each speaker P_j from speaker list N do
  for each speech signal S_i of speaker Pj
    Preprocess of speech S_i
    Compute $\hat{S}_i$ using LP approximation
    Compute LP residual

$$e_i = S_i - \hat{S}_i$$

  for each sample of e_i from K samples do
    Extract MFCC features f_k from e_i
  end
end
    Initialize HMM model parameters $\lambda_j = (A, B, \pi)$
    Train λ_j for optimal solution using EM algorithm
end

Testing Phase:

for each speaker P_j from speaker list N do
for each speech signal S_i of speaker Pj
Preprocess of speech S_i
Compute $\hat{S}_i$ using LP approximation
Compute LP residual

$$e_i = S_i - \hat{S}_i$$

for each sample of e_i from K samples do
Extract MFCC features f_k from e_i
end
end

For each model $\lambda_1 \lambda_2 \ldots \lambda_N$ do
Using the Viterbi decoding process calculate $P(O \mid \lambda_j)$, *where* $P(O \mid \lambda_j)$ is the probability of the observation sequence $O(o_1 o_2 \ldots o_T)$
end
Calculate 1-best result for a given testing speech signal using

$$arg \quad \max_{1 \le j \le N} P(O \mid \lambda_j)$$

End

VI. EXPERIMENTAL EVALUATION

A. Data base used for 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 identification task for TIMIT Speaker database.

The TIMIT corpus of read speech has been designed to provide speaker data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speaker recognition systems. TIMIT contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. We consider 200 speakers out of 630 speakers for speaker recognition. Maximum of 30 seconds of speech data is used for training and minimum of 1 sec of data for testing. In all the cases the speech signal was sampled at 16 kHz sampling frequency. Throughout this study, closed set identification experiments are done to demonstrate the feasibility of capturing the speaker-specific information from the system features and from the source features. Requirement of significantly less amount data for the speaker recognition using speaker-specific excitation information and Gaussian mixture models is also demonstrated.

B. Performance of speaker Recognition

The system has been implemented in Matlab7 on windows XP platform. The result of the study has been presented in Table1. We have used LP order of 12 for all experiments. We have trained the model using Gaussian mixture components as 16 for training speech length as . 30 sec. Testing is performed using test speech length as 3 sec., Here, recognition rate is defined as the ratio of the number of speakers identified to the total number of speakers tested.

VII. RESULTS

The speaker recognition rate is defined as the ratio of the number of speakers recognized to the total number of speakers tested. We have compared the results obtained by the proposed new approach with the some recent works which were discussed. In these works, features used, number of speakers, duration of training speech and duration of testing speech are different. Table 1, shows comparative analysis of these parameters for speaker recognition performance. The study made on the size of data for testing showed that 3 sec. of data is enough to test the speaker in terms of proposed source characteristics. The amount of training as well as testing data required in the case of text-independent speaker recognition system based on source feature is significantly less compared to the existing systems based on the vocal tract system features. Our proposed method which is tabulated in last row of the Table 1, reported that recognition performance is on par with the recent speaker models.

Table 1. Comparison to the other recent speaker models

Model	**Features**	**No. of Speakers**	**Training duration**	**Testing duration**	**Recognition Performance**
GMM [5]	Pitch + R-cep	35	20 sec.	20 sec.	98.5%
AANN [8]	Formants + LPC	50	84 sec.	48 sec.	100%
GMM [9]	MFCC + Phase	35	20 sec.	16 sec.	95.7%
GMM [6]	PLAR	168	24 sec.	6 sec.	98.81%
AANN [10]	Source information	20	6 sec.	6 sec.	95%

Proposed ergodic HMM based approach	Source information	200	30 sec.	3 sec.	99.5%

VIII. CONCLUSION

The effectiveness of source feature derived from the speech signal for text-independent speaker recognition task has been established. In this work, we proposed source feature-based text-independent speaker recognition using left-right HMM and ergodic HMMs. Here the speaker variability in terms of time varying source characteristics like glottal excitation, intonation and prosody are modeled. It is established that the source characteristics such as glottal vibrations and the prosody features such as intonation and duration can be effectively captured with HMMs than the Gaussian mixture models. The significance of ergodic HMM for automatic text-independent speaker recognition system has been established by carrying out exhaustive experiments. The effect of various parameters on the performance of speaker recognition system using GMM and HMMs was presented. The study has been made on the issues related to number of mixture components, number of states, size of data for training and testing to get good recognition performance. The amount of training as well as testing data required in the case of text-independent speaker recognition system based on source feature is significantly less compared to the existing systems based on the vocal tract system features.

REFERENCES

[1] G Anantha padmanabha, T.V., and 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.

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

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

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

[5] Reynolds D. A., and Rose R. C., " Robust Text-Independent Speaker Identification using Gaussian Mixture Models,'' IEEE-Transactions on Speech and Audio Processing, vol. 3, no. 1, pp. 72-83,1995.

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

[7] 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.

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

[9] 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.

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

[11] Gish H., Krasner M., Russell W., and Wolf J., "Methods and experiments for text-independent speaker recognition over telephone channels," Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 11, pp. 865-868, Apr. 1986.

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

[13] Molau S., Pitz M., Schluter R., and Ney H., "Computing Mel-frequency cepstral coefficients on the power spectrum," Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, pp. 73-76, May. 2001.

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

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

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

[17] Picone J. W., "Signal modeling techniques in speech recognition," Proceedings of IEEE, vol. 81, no. 9, pp. 1215-1247, Sep. 1993.

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

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

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

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

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

[23] Thevenaz P., Hugli H., 1995. *Usefulness of LPC residue in text-independent speaker verification*. Speech Commun. 17, 145–157.

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

[25] Weber F., Manganaro L., Peskin B., and 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.

[26] Yegnanarayana B., 1999. *Artificial Neural Networks*. Prentice-Hall, New Delhi, India.

