IOSR Journal of Electronics and Communication Engineering (IOSRJECE)
ISSN : 2278-2834 Volume 2, Issue 3 (July-Aug 2012), PP 39-45
www.iosrjournals.org

Self Determining Speaker Recognition by Three Level Segmental Processing Of Linear Prediction Residual

Gunda.Srikanth[1], T.V.Hyma Lakshmi[2]
[1,2]*(Electronics and Communication Enginering, SRKR Engineering College/ Andhra Univercity,India)*

Abstract :*This paper proposes a speaker specific source information at different levels.speaker recognition system exploits the source information (LP residual) present at different levels namely subsegmental, segmental &suprasegmental. The subsegmental analysis considers LP residual in blocks of 5 msec with shift of 2.5 msec to extract speaker information. The segmental analysis extracts speaker information by processing in blocks of 20 msec with shift of 2.5 msec. The suprasegmental speaker information is extracted by viewing in blocks of 250 msec with shift of 6.25 msec. The speaker recognizer studies performed using TIMIT (Texas Instruments and Massachusetts Institute of Technology) databases demonstrate that the segmental analysis provides best performance followed by subsegmental analysis. The suprasegmental analysis gives the least performance. However, the evidences from all the three levels of processing seem to be different and combine well to provide improved performance, demonstrating different speaker information captured at each level of processing. Finally, the combined evidence from all the three levels of processing together with vocal tract information further improves the speaker recognition performance.*

Keywords: *LP residual,Segmental,Source information ,Subsegmental,Suprasegmental.*

I. INTRODUCTION

1.Objective Of The Work

Speech is one of the most basic forms of communication among humans. Speech is a signal that carries information about the message to be conveyed, the characteristics of the speaker and the language of communication. The unique characteristics of the speech of a speaker are due to his/her anatomical and physiological factors. Anatomical factors relate to the physiological aspects of speech production mechanism, namely, the vocal tract system and the vocal folds. Physiological factors reflect the speaking habits of a person, such as speaking rate, accent and mannerisms. These speaker specific features are embedded in the speech signal, and hence, are useful in recognizing the speaker.

Every person has a unique voice using which that person can be recognized [1]. Automatic Speaker Recognition (ASR) [2] [3] is the task of recognizing a person from his/her voice by a machine. The objective of the present work is to explore the possibility of using excitation source characteristics of the speech production in the given speech signal as potential features for speaker recognition. The main idea in this approach is to capture the source characteristics related to glottal vibrations.

In this thesis, we address the issues involved in developing a text-independent speaker recognition system using source feature at different levels. The issue of capturing the speaker-specific source information is addressed using Gaussian mixture models (GMMs). The robustness of complete source feature and the amount of speech data required for automatic speaker recognition task are also discussed. TIMIT database is used in this study [4].

1.1 Background To Speaker Recognition

1.1.1 What is Speaker Recognition?

Form every day experience; it is clear that the speech signal carries information about the speaker. Very often we are able to recognize a speaker from his/her voice. Given this fact alone, scientific curiosity prompts us to investigate how the speech signal encodes information about its producer, and how reliably that information can be extracted. Speaker Recognition may be defined as an activity whereby a speech utterance is attributed to a person on the basis of its acoustic-phonetic or perceptual properties [5] [6].

Recognition often falls sharply when speakers attempt to disguise their voices [7]. This is reflected in machines, where accuracy decreases when mimics act as impostors. Humans appear to handle mimics better than machines do, easily perceiving when a voice is being mimicked [8]. If the target (intended) voice is familiar to the listener, he often associates the mimic voice with it. Certain voices are more easily mimicked than others, which lends evidences to the theory that different acoustic cues are used to distinguish different voices.

Speaker recognition is one area of artificial intelligence where machine performance can exceed human performance. Using short test utterances and a large number of speakers, machine accuracy often exceeds that of humans [7]. This is especially true for unfamiliar speakers, where the training time for humans to learn a new

voice is normally very long conditions was also reviewed in [2], where it was reported that human listeners are adept at using various cues to verify speakers in the presence of acoustic mismatch.

In naïve speaker recognition, the recognition is performed by untrained observers, for instance when answering the telephone call or hearing a voice in the next room. The decision is based on what is heard, and no special techniques are used. The term Automatic Speaker Recognition (ASR) first brings to mind the use of machines. Number of issues have to be addressed to get the speaker recognition work done by a machine. ASR makes it possible to verify the identity of a person trying to access the system by voice. Before addressing the various issues involved in ASR, the significance of this problem and the importance of such a system in our day-to-day life are discussed in the next section.

1.2 Limiting Factors In Speaker Recognition

The primary limiting factors in automatic speaker recognition are:

Insufficient Data: large amount of speech data is required to capture the speaker-specific information effectively. Too small a data may not be sufficient to give a complete representation of the speaker in the higher dimensional feature space.

Quality of Data: The performance of the system also depends upon the quality of the speech data being fed as input to the system. If the quality of recording is (distorted by background noise, distortion imposed by telephone transmission or tape recording), the performance might be poor.

Voice Disguise: There is always the possibility that a person is deliberately using voice disguise.

Present work is motivated by the first two limiting factors. Attempt to develop an approach which can work with small amount of speech data and also which is robust to noisy conditions is proposed in this work.

II. Work Done On The Speech Signal.

2.Representation Of Speech Production Mechanism

Speech is produced by exciting the vocal tract by the glottal excitation. From signal processing point of view, as shown in the Fig.2.1. Here, the vocal tract is replaced with filter, and the filter coefficients depend on the physical dimensions of the vocal tract. Glottal excitation is replaced with two types of signal generators, impulse train generator for voiced sounds and random number generator for unvoiced and fricative sounds.

If we denote the Fourier Transform of the source as **U(*f*)** and if we consider the vocal tract as a time-invariant linear system, represented as **H(*f*)**, however, is usually characterized by several peaks corresponding to resonances of the acoustic cavities that form the vocal tract. The spectral envelopes of the source and system for a sound unit /my name issrikanth/ are as shown in Fig.2.2 and Fig.2.3.

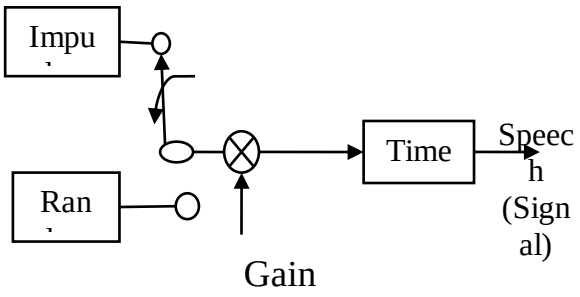


Fig.2.1

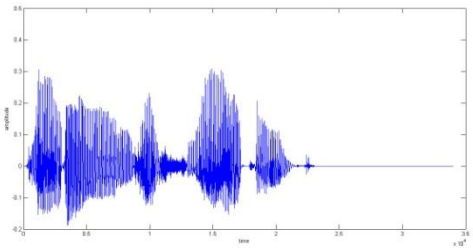


Fig.2.2.The time representation of the sound unit /my name is srikanth/

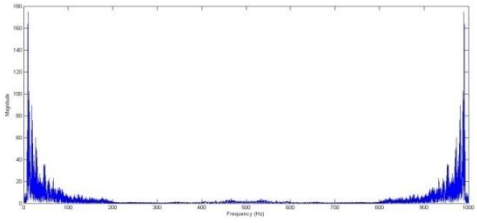


Fig.2.3. The frequency representation of the sound unit /my name is srikanth/

III. Work Done On Linear Prediction Residual

3.Linear Prediction Residual

One of the most powerful speech analysis techniques is the method of linear predictive analysis. 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. Since speech can be modeled as the output of a 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.

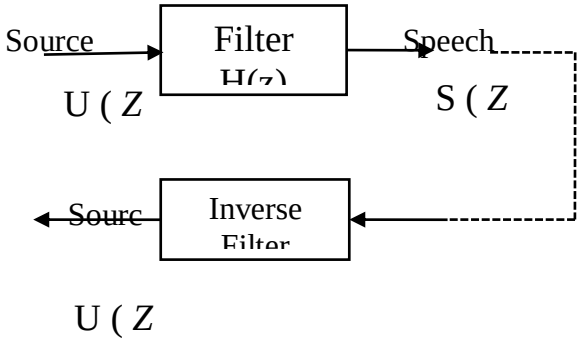


Fig.4 Filter and inverse filter representation of the speech production mechanism**.**

The discrete speech production representation of the same is as shown in the Fig.3.8. If the input signal is represented by ***u*(n)** and the output signal by ***s*(n),** then the transfer function of the system can be expressed as,

$$\boldsymbol{H}(\mathbf{z})=\frac{\boldsymbol{S}(\mathbf{z})}{\boldsymbol{U}(\mathbf{z})} \qquad (3.1)$$

Where $\boldsymbol{S}(\mathbf{z})$ and $\boldsymbol{U}(\mathbf{z})$ are z-transforms of $\boldsymbol{S}(\boldsymbol{n})$ and $\boldsymbol{U}(\boldsymbol{n})$ respectively.

Consider the case where we have the output signal and the system and have to compute the input signal. The above equation can be expressed as

$$\boldsymbol{S}(\mathbf{z}) = \mathbf{H}(\mathbf{z})\boldsymbol{U}(\mathbf{z}) \qquad (3.2)$$

$$\boldsymbol{U}(\mathbf{z}) = \frac{\boldsymbol{S}(\mathbf{z})}{\boldsymbol{H}(\mathbf{z})} \qquad (3.3)$$

$$\mathbf{U}(\mathbf{z}) = \frac{\mathbf{1}}{\boldsymbol{H}(\mathbf{z})}\boldsymbol{S}(\mathbf{z}) \qquad (3.4)$$

$$\boldsymbol{U}(\mathbf{z}) = \boldsymbol{A}(\mathbf{z})\boldsymbol{S}(\mathbf{z}) \qquad (3.5)$$

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

Linear Prediction models the output $\boldsymbol{s}(\boldsymbol{n})$ as the linear function of past outputs and present and past inputs. Assuming an all-pole model for the vocal tract, the signal $\boldsymbol{s}(\boldsymbol{n})$ can be expressed as a linear combination of past values and some input $\boldsymbol{u}(\boldsymbol{n})$as shown below.

$$\boldsymbol{s}(\boldsymbol{n}) = -\sum_{\boldsymbol{k}=1}^{\boldsymbol{p}} \boldsymbol{a}_{\boldsymbol{k}}\boldsymbol{s}(\boldsymbol{n}-\boldsymbol{k}) + \boldsymbol{G}\boldsymbol{u}(\boldsymbol{n}) \qquad (3.6)$$

Where **G** is a gain factor.

Now assuming that the input $\boldsymbol{u}(\boldsymbol{n})$ is unknown, the signal $\boldsymbol{s}(\boldsymbol{n})$ can be predicted only approximately from a linear weighted sum of past samples. Let this approximation of $\boldsymbol{s}(\boldsymbol{n})$be $\boldsymbol{s}\hat{}(\boldsymbol{n})$**,** where

$$\boldsymbol{s}\hat{}(\boldsymbol{n}) = -\sum_{\boldsymbol{k}=1}^{\boldsymbol{p}} \boldsymbol{a}_{\boldsymbol{k}}\boldsymbol{s}(\boldsymbol{n}-\boldsymbol{k}) \qquad (3.7)$$

Then the error between the actual value $\boldsymbol{s}(\boldsymbol{n})$and the predicted value $\boldsymbol{s}\hat{}(\boldsymbol{n})$ is given by

$$\boldsymbol{e}(\boldsymbol{n}) = \boldsymbol{s}(\boldsymbol{n}) - \boldsymbol{s}\hat{}(\boldsymbol{n}) = \boldsymbol{G}\boldsymbol{u}(\boldsymbol{n}) \qquad (3.8)$$

3.1 Representation of lp residual signal

This LP residual, which is generated by LP analysis, is usually ignored in all the major applications of speech analysis like speaker recognition. Only LPC coefficients are used to compute the feature vectors. But the residual signal is rich with source characteristics, which are also speaker-specific. Hence, the information present in the residual signal can be used for speaker recognition task. In the next section we shall review the work done in the direction of using the LP residual signal for speaker recognition task.

Though it is a known fact that residual contains information regarding the source, which is speaker-specific, not much work has been done in utilizing this information for speaker recognition task [2].

A few attempts have been made to utilize the speaker information present in the residual signal. In the residual signal is converted into an one-sided autocorrelation sequence and the reflection coefficients are computed by performing FFT based cepstrum is used to extract the information present in the residual signal. Some authors often summarize the whole residual in just one number represented by F_0, the fundamental frequency. But the residual on the whole carries richer information than the fundamental frequency.

Thus to improve the performance of source features, methods need to developed that tries to capture the complete source information. Source information contained in the LP residual of the speech signal [7]. The LP residual can be processed in time, frequency, cepstral or time –frequency domains to extract and model information [8]. Processing of LP residual in time-domain has the advantage that the artifacts of digital signal processing like block processing or windowing effect that creep in other domains of processing like will be negligible. Thus processing LP residual in time-domain is expected to model the speaker information in the best possible manner. Much work is not done on processing of LP residual at different levels. A unified frame work may be evolved where a given LP residual is processed at sub-segmental, segmental and supra-segmental levels. Extract features at each level and model the speaker information using Gaussian Mixture Models.

It is quoted in the literature that information extracted from the residual signal using the above mentioned techniques yielded improvement in the performance when combined with the existing information . From signal processing concepts, it is a known fact that the spectrum based features (like MFCCs), capture the gross level information. But the residual signal has much flatter spectrum (like white noise) representing the source characteristics rather than those of the vocal tract. The spectrum of the signal and the residual for a speech segment are as shown in the Fig. 3.1. Spectral representation of the residual signal might not yield complete information present in the residual, the challenge is to extract complete source information effectively.

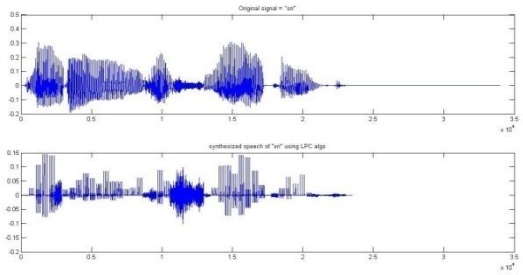


Fig. 3.1 showing the original speech signal and lpc for that speech signal/my name is srikanth/

3.2 Speaker information from sub segmental processing of LP residual.

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

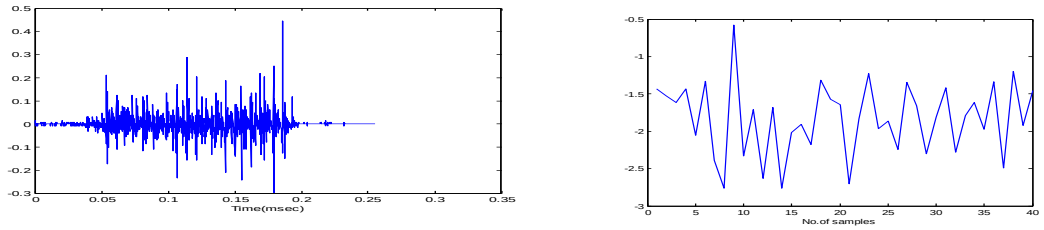


Fig. 3.2: Subsegmental of a) LP residual & (b) No. of Samples

The results show that sub segmental features provide good performance and hence contain speaker information. However the performance is comparatively poorer than the vocal tract features. The reason may be that the sub segmental features contain only one aspect of source information.

3.3Speaker information from segmental processing of LP residual

At the segmental level, speaker information present in two to three glottal cycles is modeled. This information may be attributed mostly to pitch and energy. Speaker information represented by variations within a glottal cycle has already been modeled by sub segmental analysis. In segmental level processing of LP residual, other information that can be observed at the segmental level needs to be emphasized. For this we propose to decimate the LP residual by a factor 4 so that the sampling rate becomes 2 kHz and we may have source information up to 1 kHz.. Even after decimation, the dominant speaker information at the segmental level, that is, pitch and energy information still can be preserved. Moreover, in segmental level processing, LP residual frames of 20 msec duration are used as the feature vectors. For 20 msec at 8 kHz, these feature vectors with 160 samples are of very large dimension for building the models. By decimating the LP residual by a factor 4, the dimension of the feature vectors is reduced to 40 samples per vector which is equal to the subsegmental feature vectors length.

Since the LP residual is decimated by a factor 4, we prefer to compute the feature vectors for every 2.5 msec frame shift so that the number of feature vectors will remain same as the subsegmental features. It contains mainly the pitch and energy information. The fine variations within the glottal cycle are suppressed by smoothing.The periodicity and the amplitude of the spectrum clearly represent the pitch and energy information. This observation indicates that segmental feature vectors reflect different aspect of source information compared to subsegmental feature vectors.

The high performance shows that segmental features contain good speaker information, even better than those contained at the sub segmental level. This shows that the pitch and energy may be dominating speaker-specific source information. Further, the recognition performance is comparatively poor than vocal tract features. The same reason of incomplete representation of speaker information may be attributed. The segmental source features are relatively more robust compared to both vocal tract as well as sub segmental features and also requires fewer databases. The experiments conducted on TIMIT database.

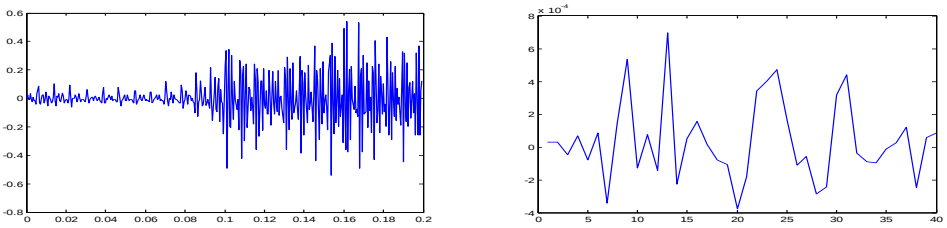


Fig. 3.3: Segmental of (a) LP residual & (b) No. of Sample

3.4 .Speaker information from suprasegmental processing of LP residual

Sub segmental processing models speaker information up to 4 kHz. Segmental processing models speaker information up to 1 kHz. Beyond that LP residual also contains some speaker information at very low frequency range, that is, may be less than 100 Hz. For example the variation in pitch and energy across several glottal cycles . In capturing such information, we need to process the LP residual at the suprasegmental level, for example, with frames of 100–300 msec range. For the LP residual sampled at 8 kHz, the feature vectors from such frames will be of very large dimension for building models.

We prefer to decimate the LP residual by a factor 50 so that the sampling rate becomes 160 Hz and we may have the source information up to 80 Hz. The dimension of the feature vector is also reduced by 50 factors. Further, the high frequency information that is already modeled by subsegmental and segmental level processing

will be smoothed out. Therefore in suprasegmental level processing of LP residual, we decimate the LP residual by a factor of 50 and process in frames of 250 msec with shift of 6.25 msec. The frame size is decided so that the dimension of the feature vectors will remain same as in subsegmental and segmental processing. However, the minimum possible frame shift in this case is 6.25 msec which corresponds to one same shift. Figure 3.4(a) shows a suprasegmental feature derived from the decimated residual shown in Fig. 3.4(b). The fast varying components of the original LP residual are eliminated and it mostly represents the long term variations. Information present in the smoothed spectrum is up to 80 Hz. The periodicity and other high frequency related information are absent. .

Results show that supra segmental level features contain some speaker information. Further, the recognition performance is significantly poor compared to sub segmental, segmental and vocal tract information. The poor result indicates that the supra segmental features may have large intra-speaker variability. The other major factor is text independent mode of operation. However, it may contain different aspect of speaker information and hence may combine well with other features.

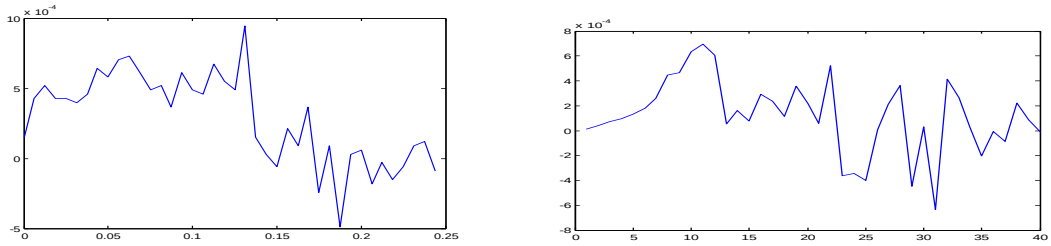
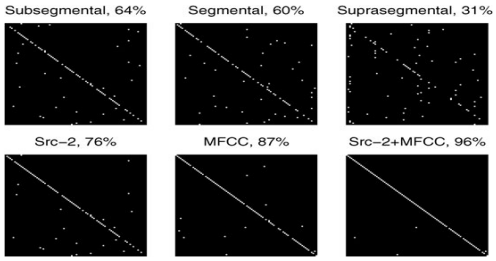


Fig.3.4: Suprasegmental of (a) LP residual & (b) No. of Samples

3.5 Combining evidences from sub segmental, segmental and supra segmental Levels of LP residual

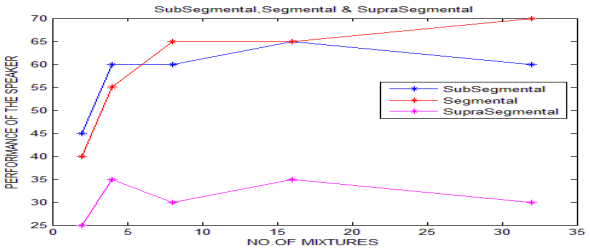
By the way of deriving each feature, the information present at sub segmental, segmental and supra segmental levels are different and hence may reflect different aspect of speaker specific source information. By comparing their recognition performance it can be observed that the segmental features provide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

In case of identification, the confusion pattern of features is considered as an indication of the different nature of information present. In the confusion pattern, principal diagonal represents correct identification and the rest represents miss classification. Fig3.5shows the confusion patterns of the identification results conducted for all the proposed features using TIMIT databases, respectively. In each case, the confusion pattern is entirely different. The decisions for both true and false identification are different. This indicates that they reflect different aspect of source information. This may help in combining the evidences to further improve the recognition performance from the source perspective.



Fig,3.5 Confusion patterns of *Sub, Seg, Supra, Src-2(source)* and *MFCC* information for identification of 80 speakers from TIMIT database

IV. Performance of Speaker Recognition at different levels



The system has been implemented in Matlab7 on Windows XP platform. We have used LP order of 12 for all experiments. We have trained the model using Gaussian mixture components as 4, 8, 16, 32 and 64 for different training speech durations as 10 sec., 20 sec., 30 sec., and 5 sec. Here, recognition rate is defined as the ratio of the number of speakers identified to the total number of speakers tested.

V. Conclusion

The effectiveness of complete source features derived from the speech signal at different levels for text-independent speaker recognition task has been established. In this work, we proposed different source features at subsegmental, segmental and suprasegmental levels such as pitch and energy and long-term variations were computed for text-independent speaker recognition using GMM. Here the speaker variability in terms of time varying source characteristics like pitch, energy and long term variations are modeled.

The effect of various parameters on the performance of speaker recognition system using GMM was presented. The study has been made on the issues related to number of mixture components, 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.

ACKNOWLEDGEMENT

I wish to express my profound gratitude and sincere thanksto to my supervisor T.Y.Hyma Lakshmi, Assistant professor,Dept of ECE, SRKREngg College, Bhimavaram for providingan initiative to this project and giving valuable timelysuggestion for the thesis work and helped me with his wide.

References

[1] B. S. Atal, "Automatic recognition of speakers from their voices," Proc. IEEE, vol.64, pp. 460-475, Apr. 1976.
[2] G. R. Doddington, "Speaker recognition-identifying people by their voices," Proc. IEEE, Vol. 73, pp. 1651-1664, Nov. 1985.
[3] D. O. 'Shaughnessy, "Speaker recognition," IEEE ASSP Magazine, pp. 4-17, 1986.
[4] "The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus," (CD-ROM), NIST Speech Disc 1-1.1, NTIS-1990.
[5] S. Furui, "An overview of speaker recognition technology", in Automatic Speech and Speaker Recognition (C.-H. Lee, F. K. Soong, and K. K. Paliwal, eds.), ch. 2, pp. 31-56, Boston: Kluwer Academic, 1996.
[6] A. E. Rosenberg, "Automatic speaker verification: A review," Proc. IEEE, vol. 64, pp. 475-487, Apr. 1976.
[7] Rama Chellapa, Charles L. Wilson and SaadSirohey, "Human and machine recognition of faces: A survey," Proc. IEEE vol. 83, no. 5, pp. 705-740, May 1995.
[8] Pati, D., &Prasanna, S. R. M. (2010). Speaker information from subband energies of linear prediction residual. In *Proc. NCC* (pp. 1–4).
[9] Atal, B. S. (1972). Automatic speaker recognition based on pitch contours. *The Journal of the Acoustical Society of America*, *52*(6), 1687–1697.

Authors

Gunda.Srikanthborn on JAN 1st 1989 received B.Tech degree in 2010 in Electronics and Communications Engineering from St.Anns College of Engineering and Technology chirala and curently persuing M.Tech inCommunication Systemsin S.R.K.R. engineering college bhimavaram.

T.V.Hyma Lakshmireceived M.Tech degree from JNTU Univercity Anantapur and currently working as an Assistant professor in SRKR Engineering College in Electronics and Communications Engineering department.

.