CHAPTER 2

REVIEW OF THE PAST WORK
IN THE FIELD
Chapter 2

REVIEW OF THE PAST WORK IN THE FIELD

In this chapter, previous works carried out in the field of Speaker Recognition are briefly reviewed. It reviews various experiments and tries to highlight the important features used with their results in various approaches and recognition methods.

J.Pollack et al. [1] have examined the effect of several factors on voice identification. The factors considered for their study were the size of the class of possible voices, the duration of the speech signal, the frequency range of the speech signal, voicing and non-voicing speech characteristics and simultaneous presentation of several voices etc. Duration of the speech signal was found to be the most efficient factor of all these factors.

S.Pruzansky [2] used a pattern matching procedure for automatic recognition of talkers to study the effect of variation in pattern on recognition performance. A time-frequency-energy pattern was generated from common word utterance of ten speakers. Cross correlation of the reference patterns with the test pattern was performed and pattern with highest correlation
was chosen. Recognition score of this study was 89%. This study has also considered two dimensional patterns of time-energy and score was found to be very low. But spectral information alone gave the same result as the three dimensional pattern.

A real time speaker verification technique using computers have been presented by R.C.Lummis [3]. In this method utterances were represented by their pitch, amplitude, and formant frequency profiles and these functions were smoothed by low-pass filtering. Dissimilarity measures were then computed. Eight true speakers and 32 casual imposters were included in this experimental study. The acceptance-rejection criterion was adjusted a posteriori for equal rates of false acceptance and false rejection. An average error rate of 1% was obtained for the speakers.

A.E.Rosenberg [4] has presented a listener performance study for speaker verification task. In his subjective testing 10 listeners participated and each was presented with a paired comparison task between challenge and reference utterances. The reference utterance was from a true speaker while the challenge utterance was either from a true speaker or an imposter with equal likelihood. The overall average error was 4.2% for false acceptance and false rejection, while the best false acceptance rate was 1.6% and best false rejection rate was 0.5%.
S.K. Das and W.H. Mohn [5] have reported an experiment in automatic speaker verification using adaptive pattern recognition. One hundred and eighteen speakers and seven thousand phrases were used for the study. An average misclassification rate of one per cent with a "no decision" rate of ten per cent was obtained in this study.

Another reported work on implementation of an on-line speaker verification system by R.C. Lummis [6] converts the utterances to pitch and gain contours and comparison was made using automatic temporal registration. Updating capability was reported to be a novel aspect in this work along with graphic display.

A text-independent speaker recognition system was described by B.S. Atal [7]. The experiment included 10 female speakers and 12 predictor coefficients were extracted from the 50 msec. duration speech. The extracted features were represented as a vector and new sets of co-ordinates which minimized the intra speaker variances were determined by linear transformations of original vector space. Correlation measures were taken over several segments and average largest correlation was considered for recognition. An overall identification accuracy of 93% was reported in this study.
C.E. Williams and K.N. Stevens [8] tried to extract the parameters that reflect the emotional state of a speaker from a speech signal. They have also investigated the changes that happen in the speech signal of a speaker in different situations. A comparison was also made with real-life situations simulated by an actor. It was found that anger, fear and sad situations produce characteristic difference in contour of fundamental frequency, average speech spectrum, temporal characteristics, precision of articulation and waveform regularity of successive glottal pulses. It was also established that attributes for a given emotional situation are not consistent from one speaker to another.

Voice identification based on visual inspection of spectrograms was performed by O. Tosi et al. [9]. Experimental trials of this study correlated with forensic models have yielded an error of approximately 6% false identification and approximately 13% false elimination.

J.J. Wolf [10] has conducted a study to find out the efficient acoustic parameters for speaker recognition. He found that fundamental frequency, features of vowel and nasal consonant spectra, glottal source spectrum slope, word duration and word onset time are very useful parameters for speaker
recognition. A simple linear classification method was employed in this study. In the experiment using 17 such parameters, no speaker identification error was found for a set of 21 adult male speakers and an error rate of 2% was found for speaker verification.

Intensively trained professional mimics were tested with an automatic speaker verification method by R.C. Lummis and A.E. Rosenberg [11]. A computer was made use of to compare the mimics and the customer utterances. The features used in this system were pitch, level and first and third formant frequencies. They have reported that 27% of the best utterances of the best mimics were accepted corresponding to a false acceptance rate of 1.2% for non-nimicking impostors.

G.D. Hair and T.W. Rekieta [12] had experimentally tried speaker verification using phoneme spectra. In this study, phonemes were automatically edited and power density spectra was computed over 7.5 KHz bandwidth. Phoneme spectra for five repetitions were averaged to produce speaker standard pattern vector. In these types of pattern vectors, the error rates were found to be very low using a simple hypersphere decision rule.

D. Meltzer [13] considered formant values for men, women and children, combined with the fundamental frequency and
then presented to trained listeners. A respective score of 90.03% was obtained in this listening experiment.

Listener performance in children had been studied by S.Cort and T.Murry [14]. The study indicated that the repertoire rather than the actual duration of the sample provides cues to the children.

T.W.Rekieta and G.D.Hair [15] have tested mimic resistance of the speaker identification system using phoneme spectra. It was found in the study that mimic even after familiarization with the user voice, was unsuccessful in his attempt.

An evaluation study was conducted by A.E.Rosenberg [16] for the listener performance. Thirtytwo casual imposters and four professional mimics were employed with eight true speakers. The test pattern was from an impostor or a mimic against the reference pattern of a true speaker. (The study showed accepted results).

A statistical model based, decision-theoretic approach to speaker verification was reported by N.S.Jayant [17]. In this model, probability density function of a measurement was compared with the PDF measurement of the true pattern and accept, reject and refuse action decisions were taken.
An automatic speaker recognition method, using temporal variation of pitch in speech as speaker-identifying characteristic, is described by B.S. Atal [18]. The 20-dimensional vector representing the pitch contours were linearly transformed to maximize the inter speaker to intra speaker variance. A Euclidean distance measure was used for the decision. A 97% of identification accuracy was claimed with this method.

R.C. Lummis has reported a method for speaker verification by computers using speech intensity for temporal registration [19]. The features used in this method were voice pitch, low-frequency intensity and three lowest formant frequencies; all represented as functions of time. Before comparison of test and reference patterns, the time dimension of the test utterance is warped to optimally register it's intensity pattern onto the reference intensity pattern. This verification was conducted for moderate size of speakers and an error rate of less than 1% was obtained for verification solely based on voice pitch and intensity.

Linear Prediction characteristics were used for automatic speaker identification and verification by B.S. Atal [20]. The linear prediction coefficients, impulse response function, the auto-correlation function, the log area function and the cepstrum functions were used as inputs of an automatic
speaker recognition system. Six repetitions of 10 speakers were used. In all these parameters, cepstrum was found to be the most effective. An identification accuracy of 70% for 50 msec. speech and 98% for 0.5 sec. was reported. Verification accuracy of 83% for 50 msec. of speech and 98% for 1 sec. of speech was also claimed in this paper. For a text-independent speaker identification experiment, 93% of accuracy with a 2 sec. of speech was achieved.

An approach based on statistical properties of nasal spectra and study on nasal coarticulation indicated [21] that coarticulation between [m] and [v] have strong speaker dependence and this could be used for speaker identification.

Automatic speaker verification system which includes Linear Predictive parameters with the pitch and intensity analysis of sentence long utterances was presented by A.E.Rosenberg and M.R.Sambur [22]. A method for selecting optimum speaker dependent features have also been discussed in this paper. A verification error of 1% with casual imposters and 4% with well trained mimics were reported in this work.

A study to examine the most effective features for speaker identification by M.R.Sambur [23] revealed that, second resonance (around 1000 Hz) in /n/, third or fourth resonance
(1700-2000 Hz) in /m/, the values of the second, third and fourth formant frequencies in vowels and the average fundamental frequency of the speaker are the most important speaker dependent features. A probability of error criterion was used to determine the relative merits of the features. An identification experiment with 11 speakers using the best five features was also presented by him. One error was observed in 320 separate identification experiments.

J.E.Paul et al. [24] reported an analytical study for semi automatic speaker identification system. Study of discriminatory power of individual phonetic events along with the study of co articulation effects on some events was conducted in this work. Weighted Euclidean distance measure was used for measuring speaker similarity within a phonetic category. Desensitized Fisher discriminant was used to convert the individual distance measures into an overall measure of similarity between the speakers of two spoken utterances.

Speech amplitude, low pass (below 1 KHz) and high pass (above 1 KHz) zero crossing rates were used as weighted sum of orthonormal functions and these weighting coefficients were used to identify unknown speakers by D.A.Wasson and R.W.Donaldson [25]. A recognition rate of 96.6% was obtained with a speaker population size of 10 and using the same data for training and testing.
A long-term averaged speech spectrum of short sentences was used for feature parameter representation by S.Furui et al. [26]. Though variation was observed in pattern for individual talkers, by defining a suitable measure of the multidimensional distance, it was possible to recognize talker with unknown samples of 3 months or longer.

An overall review of the automatic speaker recognition was presented by B.S.Atal [27]. It discusses the speaker dependent properties of speech signal, methods of selecting efficient set of speech measurement and results of experimental studies of various methods. It also compares the performance of human listener and automatic method using computers. This paper deals with text-independent and text-dependent speaker recognition techniques.

Another review on speaker verification was presented by A.E.Rosenberg [28]. It discusses techniques, evaluation and implementation of various proposed speaker recognition system.

A.Furui and F.Itakura [29] have presented a method of speech wave analysis by Partial Auto Correlation (PARCOR) coefficients and fundamental frequency. Statistical measures like Averaged value, standard deviation and correlation coefficients between parameters in voiced portion etc., were
derived and used for the experiment. The experiment indicated that these statistical features are useful in discriminating speakers and also long term variation and combination of words were studied with these features.

G.R.Doddington [30] has developed an entry control system using voice verification. The system was a text dependent one and chooses word from a set of 16 monosyllabic words. The impostor rejection rate of the system was 99.3% and user acceptance rate is 99.9% with a verification time of 5.6 sec.

An evaluation study of an automatic speaker verification system over telephone lines has been performed by R.A.Rosenberg [31]. The experiment included 100 male and female speakers and acoustic analysis of a fixed sentence long utterance was conducted. Computation in the system was done off line and updating of analyzed data with accepted utterance was also tried. Error rates of approximately 10% for new customers and 5% for adapted customers were reported.

An approach to speaker recognition using orthogonal linear prediction parameters was proposed by M.R.Sambur [32]. The orthogonal parameters obtained by linear transformation of Linear Prediction parameters were found to be important speaker characterizing features. The experiment conducted with this approach
was using the same sentence spoken by 21 male speakers. For identification and verification, the recognition accuracy exceeded 99% with high quality input speech. Telephone speech produced an accuracy above 96% and in text independent speaker identification a 94% of accuracy was obtained with high quality speech.

S. Furui [33] extracted two types of speaker dependent feature parameters. One was a long time average spectrum of a short sentence and the other the statistical parameters derived from PARCOR coefficients and fundamental frequency. He observed that by eliminating vocal chord characteristics stable personal information could be obtained.

Speaker-Speech interaction was investigated by R.L.Kashyap [34]. It considered problem of clustering speaker to aid speaker verification and optimal choice of phonemes for speaker recognition. An experiment for both speech and speaker recognition using the same speech data base was also conducted.

A survey of automatic speaker recognition system was presented by V.V.S.Sarma and B.Yegnanarayana [35]. Along with the review of the state of art, this paper also discusses computer algorithms developed for extraction and evaluation of several features like glottal waves, spectrographic information, linear prediction coefficients, LPC contours etc. The problem of data reduction and computational needs are also discussed in this paper.
Temporal variation of talker dependent features was analyzed by S. Furui [36]. Temporal variations within and between talkers of feature parameters related to exciting source and vocal tract characteristics of speech sound were analyzed. The variation of exciting source was found to be large with longer time interval and small in shorter time intervals. Talker recognition experiments indicated that the removal of exciting source characteristics from original speech improves talker recognition rate.

An automatic speaker recognition system over communication channel was designed by M. J. Hunt et al. [37]. Statistics on fundamental frequency and spectral shape information by real-time cepstrum were used in this study. Fundamental frequency was found to be not degrading over the communication channel. A reasonably good performance was reported in this study.

The structure of modular system for automatic speaker identification and verification for security system and forensic voice identification was described by E. Bunge [38]. The system response time of less than 1 sec. and the error rate of less than 1% were reported.

An investigation on long time feature averaging for speaker recognition was conducted by J. D. Markel et al. [39].
The parameters used for the study were pitch, gain and reflection coefficients. A thorough observation of the between-to-within speaker variance ratio indicated that performance was improved by long time averaging of the parameter set. The second and sixth reflection coefficients averages showed higher variance ratio.

A report on the AUROS (Automatic Recognition of Speakers by computers) project was presented by E.Bunge et al. [40]. It elaborately described the feature extraction, classification methods and finally the experiment and its results. This paper claimed a false acceptance rate of 0.94% and false rejection rate of 0.87%.

A comparative study of two speaker recognition methods, using Linear Prediction model had been conducted by L.Fasolo and G.A.Mian [41]. The experiment was conducted with 500 phrases of 10 speakers recorded over a period of 3 months.

Techniques for extraction of speaker specific features had been explained by P.Jesorsky et al. [42]. In this paper a new, minimum-maximum-locating method for time normalization is presented. It also discussed the feature extraction techniques and some experimental results.
Another method of text-independent speaker recognition which almost resembles the human perception technique had been developed by L.L. Pfeifer [43]. The vowel sounds were chosen as the basis and vowel pooling was done without vowel recognition. The sequential analysis in a dynamic fashion tested the samples until a required confidence level has been achieved.

J.D. Markel and S.B. Davis [44] have conducted experiments on Text-independent speaker identification with long term averaged features of large data base. It was found that if the averaging interval is increased, the probability of correct identification increases monotonically. For an average interval of 39 seconds a text independent speaker recognition rate of 98.05% was obtained.

A canonical discriminant analysis was applied to some sub-spaces of observation space and personal factor spaces were constructed which were used for text-independent speaker identification by H. Matsumoto et al. [45]. Decision was made by likelihood measure derived from a posteriori probabilities in the factor spaces. An identification accuracy compatible to human listeners has been claimed using 21 dimensional observation vectors obtained from voiced segments of every 40 msec.
H.M.Dante and V.V.S.Sarma [46] developed a sequential identification technique for large number of speakers. Some predetermined stages were included to reduce by classifying the number of speakers and a tree classifier is used for final decision.

An investigation to find out the possibility of application of zero crossing analysis data in speaker identification was carried out by C.Basztura and J.Jurkeiwicz [47]. An experiment with 20 male speakers proved the possibility of tracking the individual characteristics from their voices using the zero crossing analysis of speech.

An entry control system with voice identification with less than 6 seconds using four randomly chosen monosyllabic words amongst sixteen words was designed by G.R.Doddington [48]. A sequential decision strategy was used in the method. The data storage requirement is 9408 bits for the whole 16 words.

The long term zero crossing analysis of speech signal was used for speaker identification by C.Basztura and W.Majewski [49]. This work presented a method for defining minimum length of window, in which the statistical distribution is stationary in successive zero crossing intervals.
Speaker verification by human listeners over several communication systems has been studied by C.A. Mc Gomical et al. [50]. The study reveals that speaker verification by human listeners cannot be performed as accurately over mixed speech transmission system as over the same transmission system.

Dynamic programming was successfully applied to selection of features in text independent speaker identification by R.S. Cheung and B.A. Eisenstein [51].

T. Tran Cao and F. Spronck [52] have presented a work on the use of measurements of fundamental period and rate of zero crossings in speaker identification. They measured the rate of zero passage of audio signal and the first two derivatives of French non-nasal vowels taking two time windows: the fundamental period and a standard interval of 10 ms.

Dynamic programming procedures to extract the personal characteristics using PARCOR parameters and the fundamental frequency were described by S. Saito and S. Furui [53]. It was found that the rate of similarity from the matched path is close to the diagonal within the speaker whereas it is spread around the diagonal in the case of match between the speakers. The error rate in this method was found to be reduced to about 2%, when tested with two words.
Y. Grenier [54] presented another experiment for speaker identification, where Linear Prediction parameters were found to use covariance matrix. For decision, two methods were used viz., Mahalanobis distance measure and the likelihood ratio.

An elaborate study on speaker verification through telephone suggests that normalization is required for the telephone voice. A. Ichikawa et al. [55], first conducted an experiment on identification through microphone using autocorrelation coefficient, linear prediction coefficients, PARCOR parameters and Log area ratio and found that PARCOR parameters give an identification score of 99.8%. When PARCOR parameters were adopted to telephone speech, it was found that the score was as low as 65%. The authors suggested that a technique involving average-self-inverse filtering which normalizes the linear distortion of actual telephone voice and selected band analysis with the use of pitch frequency for normalization of non-linear distortion improves the verification score to 94%.

S. Nakakagawa and T. Sakai [56] reported the results of a statistical analysis on static features of voiced sound spectra and three dimensional representation of consonants on the basis of dynamic features of the spectra. The paper also presented different conclusions on the use of these spectra.
A mini-computer based speaker recognition scheme was presented by V.V.S.Sarma et al. [57]. A simple scheme for recognizing 8 speakers using long term averaged features extracted from short code words was found to give a recognition accuracy of 94%.

An evaluation of the speaker recognition method using LPC parameters in different environment viz., quiet room, dialed up telephone line via direct hook up and suction cup tap were performed by G.A.Mian [58]. Sentences were manually segmented and evaluation was conducted on phoneme, on breath groups and the whole sentence using minimum distance classifier.

M.Shridhar and M.Baraniecki [59] have discussed the problem of noise in the speaker verification system and investigated the possibility of a speaker verification algorithm derived from an orthogonal parameter representation of speech.

A system for on line speaker verification was described by U.Hofker and P.Jesorsky [60]. This paper discusses various considerations for pre-processing, feature extraction and classification to get a better performance.

W.D.Voiers [61] developed a listener method of speaker recognition in which he employed factor analysis to identify the elementary perceptual parameters of individual differences in speech.
H.M.Dante, V.V.S.Sarma and G.R.Dattatreya [62] presented a multistage decision scheme for speaker recognition. In this method one feature was used for decision at each stage and final decision was made when the number of classes becomes less than a predetermined value. The procedure was based on an optimal stochastic control problem and a population of 60 speakers was tested in this method.

A speaker recognition system suitable for forensic application was designed by E.Bunge [63]. He also investigated the influence of telephone transmission, which is common in forensic applications.

Another speaker verification system based on a feature set of selected short-time spectra, long-time spectra, intensity contour and stationary contour was developed by U.Hoefker et al. [64].

A speaker verification with two stage classifier employing only 391 bits of reference storage on a standard magnetic identity card was presented by R.Geppert et al. [65].

R.Dubes et al. [66] used a method based on choral speech for speaker recognition. The authors claimed that when the voice channel is translated into choral speech, many of the
factors affecting the system will be averaging out and the system will become more or less independent of text, language, time and media of recording.

A review on Principles of Automatic Speaker Recognition was given by P. Jesorsky [67]. In this he describes speaker specific features, pre-processing and parameter representation of speech signal, feature extraction methods by statistical analysis, segment analysis and contour analysis and also different classification schemes. Finally he investigates various recognition schemes available.

A scheme for automatic speaker identification for a large population based on multistage, decision tree classifier was proposed by H.M. Dante and V.V.S. Sarma [68]. A large number of classes, after comparing with the given pattern, were rejected in the first stage based on a subset of the features. The final decision was taken using the remaining features after a predetermined number of classifications. In this scheme the authors used a population size of 30.

The importance of mood of the speaker in speaker recognition problem was investigated by A.I. Menkiti [69]. He examined the speaker identification in two methods viz., aural and by analysis for different speaker moods.
The problem of variability of speaker voice with organic change as well as change of time was found to be tackled by applying cluster analysis. M.H.Kuhn [70] has shown that with this technique of excluding extremely untypical samples from the training session, the performance was improved and the error rate of 20% was reduced to 4%.

A system for access control using speaker verification implemented on a mini—computer was developed by M.H.Kuhn [71]. Here he proposed a method of reduction of storage size, so that the pattern can be stored in a magnetic identity card and the verification system will have no restriction on the population size of speakers.

M.Baraniecki and M.Shridhar [72] introduced a new scheme to verify speaker from a speech corrupted by noise with unknown statistics. When preprocessed with an adaptive noise cancelling algorithm, the verification performance was improved from 45% to 95%.

R.E.Wohlford et al. [73] conducted a comparative study of four speaker verification methods. The four methods were: short and long term averages, cepstral measurements of long term spectral averages, orthogonal linear prediction of speech waveforms and long term averages of LPC reflection coefficients combined with pitch and overall power.
S.Furui and A.E.Rosenberg [74] have suggested a new technique where the fixed text utterance were represented by time functions of cepstral coefficients expanded by an orthogonal polynomial. A dynamic programming technique was also used to bring sample utterance representation into time registration with reference pattern. The decision, based on the overall distance, which was again a sum of local distances along the optimal path associated with the time registration.

A real time automatic speaker recognition system was implemented on a mini-computer by C.E.Chafei [75]. The four characteristic features are extracted from the pitch of the speaker. The system worked satisfactorily for 15 speakers with utterances taken over a period of 6 months.

A low cost microprocessor based speaker verification system was presented by M.H.Khun [76]. The verification procedure involves Bayes classifier with histogram approximation of the probability density. The system requires, 256 bytes of RAM memory and this was tested for 127 speakers. The reference was updated after every successful verification and the verification time was about 8 sec. with 1% false rejection and 2% false acceptance.

H.M.Dante and V.V.S.Sarma [77] have studied speaker verification based on a theoretical model useful for forensic
application. Three methods using single feature, multiple independent measurements of single feature and multiple independent features were analysed in this paper.

M.H.Kuhn et al. [78] in another paper have discussed the differences of customer acceptance in different application areas like banking and military environment. He also indicated the possibility of online user evaluation.

A study by H.Ney and M.H.Kuhn [79] revealed that the time frequency matrix obtained from a 13 channel filter bank, normalized with respect to long time averaged spectrum and normalized via dynamic programming could be a very strong feature set for telephone line speaker recognition.

Orthogonal parameters formed from the eigen values and eigen vectors of the covariance matrix of the speech of a speaker were found to be very useful in speaker verification. R.E.Bogner [80] studied this technique and also the factors that effect the distortion over telephone line. The method was based on covariance of logarithmic spectral estimates and the results gave an accuracy of about 95%.

A.E.Rosenberg and K.L.Shilpey [81] proposed a speaker identification and verification system combined with a word recognition system in which a template distance measure was made use of.
The auto correlation function was subjected for study in speaker recognition on telephone lines. H.Ney [82] came to the conclusion that the clipped auto correlation function was superior to simple auto correlation function both in the simplicity of computation and reduced dynamic variability. Dynamic programming was used for time registration and an error rate of around 2% was obtained.

A text-independent real time speaker recognition system was developed by E.H.Wrench Jr. [83]. A test conducted with 30 speakers with 10 seconds of speech gave a recognition accuracy of 93-100%.

A dynamic programming based feature selection was attempted by M.Shridhar et al. [84]. He has also investigated the possibility of the use of orthogonal linear prediction parameters in the text-independent speaker recognition. A verification accuracy of 96.5% was claimed with 8 optimally chosen parameters generated from 100 seconds of time spaced voiced speech for reference pattern and 5 seconds of speech for test pattern.

A composite model of speech for speaker and word recognition was presented by R.J.Fontana and M.S.Fox [85]. The method comprised of estimating the underlying sub-source using
data compression technique and the switch sequences were derived from these estimates. These switch sequences were then compared in time domain and decision was made via variation distance.

A fixed text approach for telephone line speaker recognition, based on the cepstral coefficients obtained from LPC analysis of time functions and removing the frequency response distortions introduced from the transmission lines was adopted by S. Furui [86]. Time registration and distance calculation were performed by dynamic programming technique and the decision was made on the overall distance measure. A recognition accuracy less than 1% was obtained in spite of different transmission conditions.

One real time speaker verification system implemented on a mini-computer was introduced by M. De George [87]. The techniques for optimization like word selection, microphone selection, training technique and reference template updating were discussed in this paper. The particular work claimed a recognition accuracy of 98.8% with a speaker base of 27 males and 15 females.

A software developed for interactive text-independent speaker verification was presented by A. Kohen and I. Froind [88]. The different aspects of the problem and the results of a verification system were also discussed.
A comparative study of statistical features and dynamic features were conducted by S. Furui [89]. The speech wave transformed into a set of log area ratios and fundamental frequency were used for the extraction of statistical and dynamic features. In the case of statistical features, the mean value and standard deviation for each time function and correlation matrix between these functions were calculated for voiced regions of speech. Using dynamic features, time registration of time functions was performed. The long term effect of these features were also studied in this work.

H. Ney [90] reported a method for speaker recognition of telephone speech using the intensity contours and fundamental period. The fundamental frequency was measured by estimation procedure and a real time operation was assured by implementing the system on a mini-computer or a microprocessor.

A study on the effects of acoustic features on speaker identification by K. Itoh and S. Saito [91] concluded that frequency spectrum envelope (FSE) is contributing too much to the speech and only by removing FSE the fundamental frequency and tempo become significant. The study was conducted by using synthesized speech signal processed by PARCOR speech analysis-synthesis technique.
S. Furui [92] developed a scheme which employed the statistical feature of cepstrum viz., mean value, standard deviation, Fourier coefficients of each parameter and cross correlation between parameters. More than 99.9% of accuracy was obtained using a microphone. False acceptance rate of 2% and false rejection rate of 3% was achieved for an online verification experiment using telephone speech.

Another comparative study on statistical and dynamic features using for text-dependent speaker recognition method presented by S. Furui [93] suggested that the statistical features have advantages in calculation, memory size for feature and recognition. It also attempted to combine these two features and the effectiveness of the spectral equalization technique.

H. Ney and R. Gierioff [94] introduced a feature weighting technique in speaker recognition. In which, weights were predetermined for individual feature components according to the ability to distinguish between classes of speakers. These weights are depending on both time and frequency.

R. Schwartz et al. [95] have investigated the application of probability density estimation to text-independent speaker identification. The study was performed for two parametric and one non-parametric PDF estimation and the performance was found to be better in non-parametric estimation.
A method for text-independent speaker recognition was devised by N.Mohanakrishnan et al. [96]. In this method, they included LPC, reflection cepstrum, log area ratio coefficients, speech power spectrum parameters and inverse filter spectral coefficients, as features and selected any two features for the first stage of recognition. If there was any disagreement in the result, another feature was made use of and the result was found to improve.

A unified scheme for speaker verification was proposed by M.Shridhar et al. [97]. They used an orthogonal linear prediction model for verification and claimed a good result for the test speech recorded in a noisy environment. The test speech was preprocessed using a modified adaptive noise cancellation filter.

H.Ney [98] suggested a speaker recognition method which employs the spectrogram of a sentence. The time synchronization of the test and reference template was done using dynamic programming and the decision was made by the dissimilarity measure after comparison of the two spectrograms.

A text-independent speaker recognition procedure described by M.Shridhar and N.Mohanakrishnan [99] claimed an accuracy greater than 99%. The procedure was implemented at two stages using two different parameter sets and a confirmation was done with a third set of parameter if controversy arose in the first stages.
A comparative study by J. Wolf et al. [100] showed that the probabilistic classifier is much more superior to that of a minimum distance classifier. A data base created from a radio channel speech was tested and compared with the performance of lab quality speech in this study.

Another method based on a statistical model of the speaker's vector quantized speech was introduced by K. P. Li and E. H. Wrench Jr. [101]. The frequency-occurring vectors or characters formed a model of multiple points with n-dimensional speech space instead of the usual single point models. The distance was also measured depending on the statistical distribution of the models. An accuracy as high as 96% was claimed by this particular study.

A comparative study of the distance measures used in the speaker recognition methods was conducted by M. Shridhar et al. [102]. Among the four commonly used methods viz., Mahalanobis distance, maximum a posteriori probability, nearest neighbour criterion and the correlation distance measures, the nearest neighbour criterion with a modification was found to be superior.

From a previous experimental analysis on the telephone speech, based on the F0 statistics and low order cepstrum coefficient, the results are found to be poor. In a study by M. J. Hunt [103] feature set based on the frequencies of peaks in
the short term smoothed spectrum was found to perform better for
the telephone speech, mainly because of the greater resistance to
the noise and non-linear distortion.

D.Helling and H.W.Strube [104] have described a speaker
identification and verification system in real time using signal
processors. In this method, a Euclidean distance approach was
used for decision and two feature reduction techniques were
discussed.

A study to find the best features for consonant and
speaker recognition by S.Nakagawa and M.Sakamoto [105] showed
that FFT cepstrum and LPC cepstrum are the most suitable para-
eters. In this work, it was specifically found that 18th FFT
cepstrum at an analysis frame length of 25.6 ms. with 5 ms. of
shift between frames are most suitable for the speaker recognition.

M.K.Krasner et al. [106] conducted a study on the
speaker recognition using a radio channel speech data, which is
of poor quality and full of noise. This work suggested ways to
extract robust feature sets and a method for modelling and
classification.

A performance verification study of the communication
system for speaker identification was carried out by
P.E.Papamichialis and G.R.Doddington [107]. The listeners were
made to identify the utterance comparing with a reference utterance and the experiment was repeated for processed and unprocessed speech, from same and different data base.

H.Hollien [108] suggested that speaker identification can be effectively done using multiple vectors. In his paper he considered four speech vectors viz., Speaker fundamental frequency, long term speech spectra, vowel formant tracking and temporal analysis vector.

In another approach for speaker identification by C.B.A.Shaw Cross et al. [109] two dimensional half-plane lattice parameters of spectrogram were employed. An FFT and Wigner transform were used to process the raw speech data into a two dimensional form suitable for lattice modelling.

A speaker verification system constructed from off-the shelf components consisting of both hardware and software was explained by D.E.Crabbs and D.P.Conard [110]. They described the system, starting from the analog speech acceptance, feature extraction and template creation to the decision classification and also performance verification.

H.Tananka et al. [111] proposed a novel approach for speaker identification based on a new parameter named Time Sequence Matrix. In this method the zero crossing interval was
extracted and plotted into a two dimensional 10x10 matrix. The experimental results gave an accuracy of 94% with this method.

An operational evaluation of the speaker verification system was presented by D.E.Crabbs and J.R.Clymer [112]. The optimization of the system performance based on the results of the experiment was also explained by them.

A speaker identification based on known voices familiarity and narrow-band coding was presented by A.S.Nielson and K.R.Stern [113]. A listening test was conducted over an unprocessed and LPC processed channels from the voices of speakers who were familiar with the listeners and it was found that the processed speech was better for identification.

The channel variability problem was analysed for speaker identification by H.Gish et al. [114]. They have shown that the channel invariant features can discard more speaker dependent information and found that it was more effective when channel variability was incorporated during training.

Vocoded speech was put into test for speaker recognition by S.S.Everett [115]. Six different voice processors were used in the input side of the speaker recognition system. An accuracy as high as 95% was obtained for different input filter bandwidths.
The use of vector quantization in speaker verification was tried by F.K. Soong et al. [116]. A vector quantization code book was used as an efficient means of characterizing the short term spectral features of a speaker. A data base consisting of isolated digits was used for experiment and 98% accuracy for speaker identification was obtained. The system was also tested for different code book size, number of digits, different recording sessions etc.

Another work on vector quantization was reported by J.T. Buck et al. [117]. For an experiment on 16 speaker population, they obtained a false rejection rate of 0.8% and false acceptance rate of 0.0%.

A speaker verification system for access control discusses the issues of interfacing different microphones and also the environmental variation. This was studied by W. Flix and M. De George [118].

H. Kashiwagi et al. [119] used spectral envelope of the linear prediction residual of speech for speaker identification. The spectral envelope was obtained from the low time portion of the cepstrum of the residual and a Euclidean distance between the envelopes were used for identification. An identification rate above 80% was obtained by this method.
H.Noda [120] developed a speaker verification system for forensic application where concatenation of inventory of recorded syllabic units was tried for fixing the threshold. Here a text independent speaker verification via telephone lines was conducted. It was found that this method is superior to the normal fixed threshold.

A.E.Rosenberg and K.L.Shipley [121] have introduced a talker recognition system in tandem with a talker independent isolated word recognition. A template based approach was adopted for both these recognitions. An experiment with relatively large population gave an identification error of 3.6 to 14% and an error rate of 8% when tested in a speaker verification mode.

Performance of the human speaker recognition of LPC voice processor was studied by Z.Uzdy [122] This study pointed out that for a very good speaker recognition a high-frequency data bandwidth is necessary.

M.Ganesan et al. [123] have demonstrated this speaker identification and verification based on an acoustic data of Hindi vowels using a sound spectrograph. The features extracted from the spectrograph was fed to the computer for creation of templates. A classification was done based on the Euclidean distance between the test and reference templates. The experiment was done by changing the parameters and utterances and results were compared.
A general review of the speaker recognition methods was given by G.R. Doddington [124]. He has discussed the different methods of speaker recognition like listening, visual and computer techniques and the limitations of each method. The paper also deals with text-dependent and text-independent methods and also covers the state of the art.

G. Audiso and A. Caramella [125] have developed a speaker verification algorithm which evaluate the probability density function using a Parzen estimator with a hyperconic kernel. A method for evaluating the radius of the kernel was also given in this paper.

A system using template matching for both text-dependent and text-independent speaker recognition was presented by A.L. Higgins [126]. A 10 sec. speech conversation could be identified without any error using the above template matching method.

A speaker recognition system for the field use (outside the lab environment), has been explained by M.C.K. Yang et al. [127]. The system makes use of a regression technique for the combination of several features.

Assuming that the recognition of speech is a prerequisite for a good quality speaker recognition, M. De George and
W.Fix [128] have presented a speaker verification system integrated to an access control environment which utilizes short term spectral and temporal features. A field experiment was also conducted by them using this system.

Cepstral features and energy are extracted in real time and used in the speaker verification system by M.Burnbaum et al. [129]. A dynamic time warping method was used for distance measurement and the reference template was updated after each successful verification. Since the system was used in a dialed telephone line, channel normalization and noise floor were used to reduce the telephone line variation. An equal error rate of 19% was claimed by the authors.

S.Furui [130] has given an overview of the speaker recognition technology and discusses various issues in this particular field. The paper focuses on the effect of long term spectral variability on the recognition accuracy and the ways to reduce this effect.

An optimal decision threshold was determined by N.Fakotakis et al. [131]. This was calculated from the distribution of the intra and inter speaker distances by minimizing the false acceptance and false rejection errors.

An overview of speaker recognition was presented by D.O'Shaughnessy [132]. The paper reviews various recognition
techniques, and different distance measures, timing considerations and dynamic time warping. It also discusses template matching technique and use of dynamic and statistical features. The author looks into various techniques like vector quantization, cepstral analysis and use of orthogonal LPC parameters. In general the paper touches upon almost all the aspects of speaker recognition.

A text-dependent speaker verification using vector quantization source coding was presented by D.K.Burton [133]. A source code book was designed to represent a speaker based on his utterance and deviation is measured in this code book. The experiment resulted into a false acceptance rate as low as 0.7% and 0.6% false rejection rate with a speaker population of 16 true speakers and 11 imposters.

A PC based speaker verification, using statistically averaged parameters of speech was proposed by Michailov and D.Milev [134]. This method requires smaller memory capacity and computation power. In this method, the parametric vector representation was transformed into a new one, which reduces the influence of phonetic content and the condition of recording and also the information redundancy of the LPC.

The technique of frequency warping was found to be very effective for speaker verification in noise by H.Noda [135]. In this method the higher energy portion like formant regions were
expanded and low energy portion which is affected by the noise was shortened in the spectrum. Then an LPC analysis on this frequency warped spectrum was performed and the auto correlation function derived from this to form an all pole model. The spectral distance measures from this method was found to be very effective in speaker verification especially in noisy environments.

A new method of Circular Hidden Markov Model (CHMM) was applied in speaker identification by Y.C.Zheng and B.Z.Yuan [136]. A distinct reference CHMM was produced for each person using Baum's forward and backward algorithm. Classification was effected depending up on the highest probability and results show 94% speaker recognition accuracy.

A template based approach for speaker verification, using different distance measures was presented by G.Velius [137]. He has also studied the variation by changing the length of window and the parameters chosen and the order of LPC-cepstrum analysis. Euclidean, inverse variance weighting, differential mean weighting, Khan's simplified weighting, Mahalanobis distance, and the Fisher linear discriminant were the distance measures tested. It was found that the performance varies depending upon the vocabulary and an average performance of 5% Equal Error Rate was obtained.
J.Wilbur and F.J.Taylor [138] used a derivative of Wigner Distribution function called smoothed discrete Wigner distribution. This was proved to be significantly more consistent estimate over differing samples of a given word. With this the formant frequencies are well defined and consistent over the samples, which helped in speaker identification.

A Texas Instruments TMS32020 digital signal processor based speaker verification system was designed by J.B.Attili et al. [139]. They have used a 37 dimensional feature vector consisting of 12 PARCOR coefficients, 12 log-area coefficients, 12 LPC cepstrum coefficients and one normalized gain coefficient. An overall error rate of 1.9% in text independent verification and 0.94% in text dependent verification was observed using this hardware based system which carried out the verification in almost real time.

2.1 PRESENT WORK

Text-dependent speaker recognition has shown a better performance compared to the text-independent recognition. The a priori knowledge of the text gives an opportunity of modelling the phonemes. In this present work a model with knowledge base and feed back network is proposed for speech production. Attempts have been made to relate the measurable parameters with this model, since, the actual parameters in the model are not measurable.
Two approaches have been attempted for speaker identification. Both the methods are employing the popular template matching approach. In the first approach, a Fixed Text Phoneme Model has been used. Using entropy from the probability of occurrence of the text and the entropy obtained from the timings of the spoken text, a template is generated for comparison.

In the second approach, a Similarity Measure Method is made use of. The feature measurements from the spoken fixed-text are taken for the creation of the template. A weightage for each measure is assigned when a similarity is observed during comparison. The sum total of the weightages are taken as the similarity measure for each speaker template. A knock-out method is also used for the selection of the feature measurements depending on their performance. A good performance is obtained using a set of final feature measurements.