ISSN (Print) : 2319-8613
ISSN (Online) : 0975-4024

A Review on Speaker Recognition

S.Sujiya [#1], Dr.E.Chandra [*2]
Department of Computer Science, Bharathiar University, Coimbatore, India
suji.sreedharan@gmail.com
crcspeech@gmail.com

***Abstract*—Automatic speaker recognition system plays a vital role in verifying identity in many e-commerce applications as well as in general business interactions, forensics, and law enforcement. Today many employees have access to their company's information system by logging in from home. Also Internet services and telephone banking are widely used by private and corporate sectors. Therefore to protect one's resources or information confidentially with simple password is not consistent and secure in the technological world of today. There are two major applications of speaker recognition technologies. If the speaker claims to be of an assured identity and the voice is used to verify this claim, it is called as verification or authentication. On the other hand, identification is the work of determining an unknown speaker's identity. This paper depicts the fundamentals of automatic speaker recognition, concerning feature extraction and speaker modeling.**

Keyword- Speaker Recognition, Classification, Feature Extraction, Speaker Modelling.

I. INTRODUCTION

Automatic Speaker Recognition is the field of study deals with digital signal processing correlated to the recognition of the people based on their voice. Every individual are unique in their vocal tract shapes, larynx sizes, and other parts of their voice production organs. In addition to these physical differences, each individual speaker has his/her characteristic manner of speaking, related to accent, rhythm, intonation style, pronunciation pattern, selection of vocabulary [1]. Voice of a person is the most popular biometric identification technique used for authenticating and monitoring human using their speech signal.

The objective of speaker recognition is to decide which speaker is present based on the individual's utterance. Speaker recognition is the procedure of automatically recognizing who is speaking by using the speaker related information included in speech waves to verify identities being claimed by people accessing systems. Speaker identity is linked with physiological and behavioural characteristics of the speech production system of an individual speaker [2].

- The physiological module of speaker recognition is the physical shape of the subject's voice tract.
- The behavioural module is the physical movement of jaws, tongue and larynx

II. SPEAKER RECOGNITION

Speaker recognition comprises activities which attempt to link a speech sample to its speaker through its acoustic properties [4]. Speech signal is a multidimensional acoustic wave which deals with information regarding characteristics of a speaker, spoken phrase, speaker emotions, additional noise, channel transformations etc. Every individual voice is unique personal trait. For indistinguishable voice, the two individuals should have alike vocal mechanism and identical coordination of their articulators, which is least feasible. However, some amount of difference also occurs in the speech exemplars obtained from the same speaker. This occurrence is due to the fact that a speaker cannot accurately imitate the same utterance again and again. Even, the signature of an individual also shows complete variation from trails to trials [2][3].

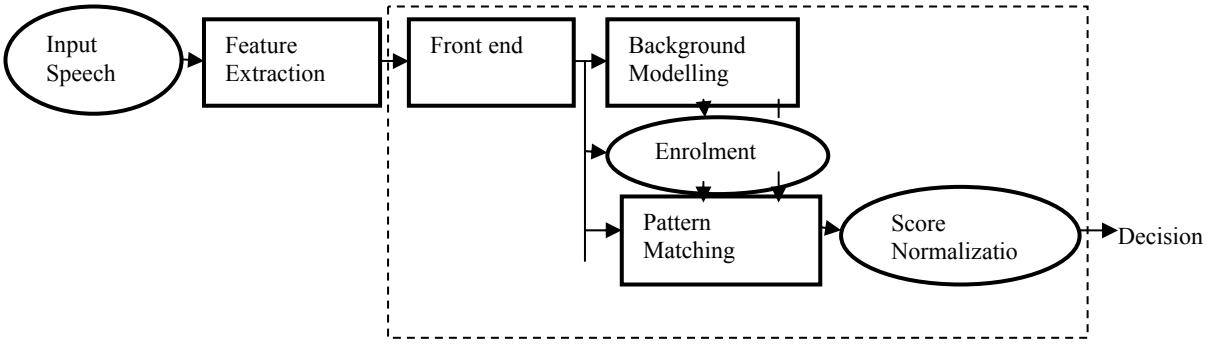

Fig.1.Automatic Speaker Recognition System

Figure 1 shows the components of an automatic speaker recognition system. In speaker recognition the initial process is feature extraction. In feature extraction module the raw signals are transformed into feature vectors in

which speaker specific properties are emphasized and statistical redundancies is concealed. In the enrollment mode, a speaker model is trained using feature vectors of the target speaker. In recognition mode, the feature vectors extracted from the unknown person's utterance of the individual are with the system database and a similarity score is generated. Final decision is made by the decision module based on the similarity score. Virtually all state-of-the-art speaker recognition systems use a set of background speakers or cohort speakers. This is done to enhance the robustness and computational efficiency of the recognizer. In the enrolment phase, background speakers are used as the negative examples in the training of a discriminative model [4], or in training phase a universal background model from which the target speaker models are adapted. In the recognition phase, background speakers are used in the normalization of the speaker match score [5][8]

A. *Classification of Speaker Recognition*

Speaker recognition can be classified into a number of categories. Figure 2 below provides the various classifications of speaker recognition.

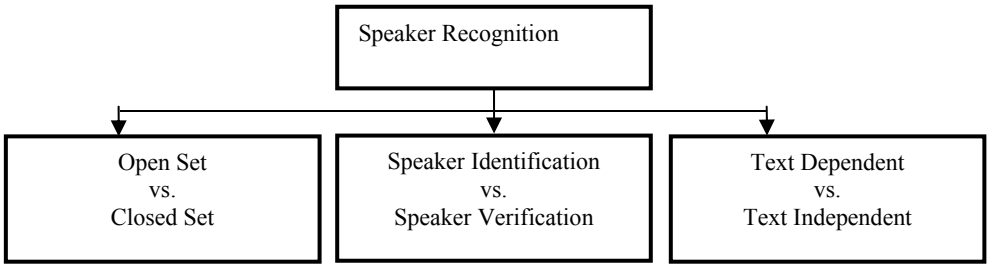

Fig. 2. Classification of Speaker Recognition

1. Open Set Vs Closed Set

Speaker recognition can be categorized into open set and closed set speaker recognition. This group of classification is based on the set of trained speakers available in a system. [5].

Open Set: An open set system can have any number of trained speakers. The speakers and the number of speakers can be anything greater than one in open set.

Closed Set: A closed set system has only a fixed number of users registered to the system.

2**.** *Identification vs. Verification*

Automatic speaker composed of identification and verification is often considered to be the most natural and economical methods for preventing unauthorized access to physical locations or computer systems [4].

Speaker identification: Speaker identification is the study of identifying a speaker of a given utterance amongst a set of known speakers. The unknown speaker is identified as the speaker whose model best matches the input utterance.[6]

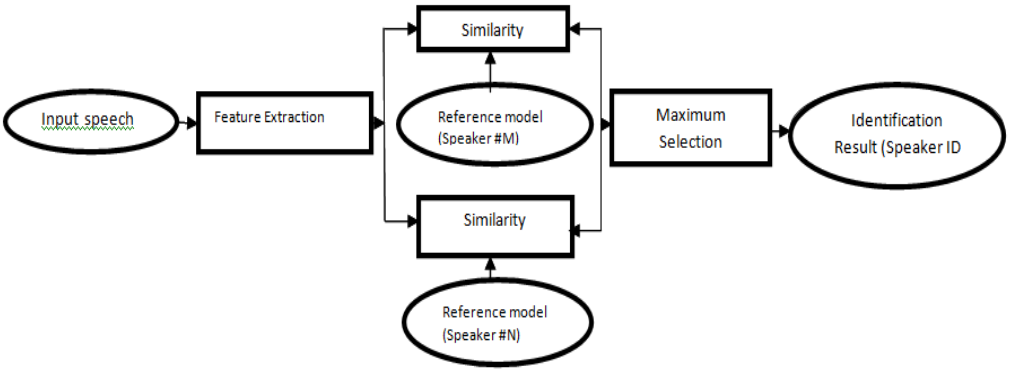

Fig. 3**.** Speaker Identification

Speaker verification: Speaker verification is the method of accepting or rejecting the identity claim of a speaker. Speaker verification is a more direct and converged effort leading to either acceptance or rejection of the claimed identity of a speaker. To be precise, this analysis concludes whether a speaker is the one who he/she claims to be [4]. It can be considered as a true-or-false binary decision problem. Basically referred to as the open-set problem, because this task requires distinguishing a claimed speaker's voice known to the system from a

potentially large group of voices unknown to the system. In today's technological world verification is the basis for most speaker recognition applications and the most commercially feasible task [1][5].

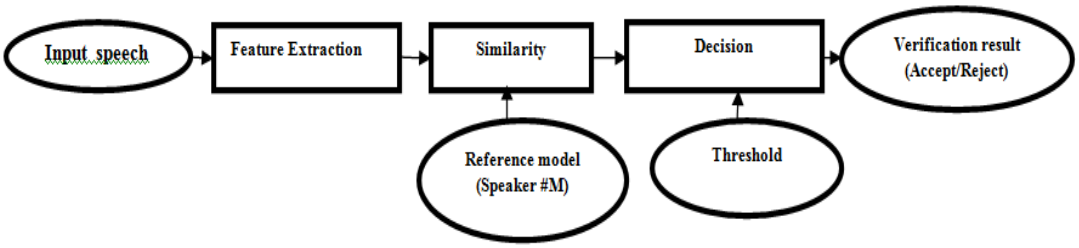

Fig. 4. Speaker Verification

3. *Text-Dependent vs. Text Independent*

Text-Dependent: In text-dependent recognition the test utterance is the same to the text used in the training phase. The test speaker has prior knowledge of the system.

Text-Independent: In, text-independent recognition the test speaker doesn't have any knowledge about the contents of the training phase and can speak anything [8].

III. FEATURE EXTRACTION

According to speaker recognition, feature extraction is the process of retaining useful relevant information of the speech signal by rejecting redundant and irrelevant information .It is a process of analysis of speech signal. Various techniques for extracting features for speaker recognition are Mel-Frequency Cepstrum Coefficients (MFCC), Linear Prediction Coding (LPC), Linear Predictive Cepstral Coefficients (LPCC) and Perceptual Linear Predictive Cepstral Coefficients (PLPC).[9]

A. Mel Frequency Cepstral Coefficients (MFCC)

MFCC is one of the most popular feature extraction technique used in speaker identification and verification process. It is based on the human peripheral auditory method. According to human perception of the frequency contents of sounds for speech signals, it does not track a linear scale. Because of human perception behavior which does not follow linear scale that is above 1000 Hz, a log scale above 1000Hz is taken which is called as Mel Scale. This Mel scale indicates linearity up to 1000Hz and logarithmic above 1000Hz. Hence for each tone of actual frequency, a subjective pitch is measured on different scale called as Mel Scale. Formula to calculate the estimated mels for a given frequency f in Hz:

$$mel\ (f) = 2595 * log10(1 + f/700)$$

The log (mel) spectrum is converted back to time. The end result is called the Mel frequency cepstrum coefficients (MFCC). The human ear is responsive to both the static and dynamic characteristic of a signal and the MFCC mainly focus on the static characteristics [8][9].

B. Linear Predictive Coding (LPC)

In Linear Predictive Coding the analysis of the speech signal is achieved by estimation of the formants. Effects of formants from the speech signal are removed by LPC, and estimate the intensity and frequency of the remaining buzz. This method of removing the formants is known as inverse filtering, and the remaining signal is called the Residue. In LPC technique, each sample of the speech signal is conveyed as a linear combination of the previous samples [1][11]. This is called a linear predictor and hence it is called as linear predictive coding.

C. Linear Predictive Cepstral Coefficients (LPCC)

LPCC is a popular technique and widely used to extract the features from speech signal. LPCC parameters can effectively describe energy and frequency spectrum for sound frames. The base of explaining acoustic signals spectrum, modelling and pattern recognition is set by the result of increasing logarithm which restrains the fast change of frequency spectrum, more centralized and superior for short-time character and it is because of cepstrum resulting from original spectrum. One of the common short term spectral measurements presently used are LPC derived cepstral coefficients (LPCC) and their regression coefficients LPCC shows the differences of the biological arrangement of human vocal tract and is computed through iteration from the LPC Parameters to the LPC Cepstrum [11].

D. Perceptual Linear Predictive Cepstral Coefficients (PLPCC)

This technique is based on magnitude spectrum of the speech analysis window. MFCC and LPC are cepstral techniques and PLPCC is a temporal technique in feature extraction phase. The steps followed to calculate the coefficients of the PLPCC are:

First, compute the power spectrum of a windowed speech.

Second, for sampling frequency of 8 kHz perform grouping of the results to 23 critical bands using bark scaling.

Third, to simulate the power law of hearing, carry out loudness equalization and cube root compression.

Fourthly, perform inverse Fast Fourier Transform (IFFT).

Fifth, one is execute LP analysis by Levinson- Durbin algorithm.

And final step is to convert LP coefficients into cepstral coefficients

The relationship between frequency in Bark and frequency in Hz is specified as in[12]

$$f(\text{bark}) = 6 * \text{arcsin h}(f(\text{Hz})/600)$$

IV. SPEAKER MODELING

In Speaker modelling two types of models are extensively used in recognition systems:

- Stochastic models
- Template models

The stochastic model exploits the advantage of probability theory. This is achieved by speech production process as a parametric random process. It assumes that the parameters of the essential stochastic process can be estimated accurately, in a definite manner. In parametric methods assumption is made about generation of feature vectors but the non-parametric methods are free from several assumptions about data generation. The template model (non-parametric method) attempts to generate a model for speech production process for a particular user in a non-parametric manner. This is done using sequences of feature vectors extracted from multiple utterances of the same word by the same person. Template models used to dictate early work in speaker recognition because it facilitates without making any prediction about how the feature vectors are being created. Hence the template model is naturally more reasonable. However, recent work in stochastic models has exposed them to be more flexible, thus allowing for generation of better models for speaker recognition process. The state-of-the-art in feature matching techniques used in speaker recognition includes Gaussian Mixture Modelling (GMM), Dynamic Time Warping (DTW) and Vector Quantization (VQ) and ANN.

In a speaker recognition system, the process of representing each speaker in an efficient and unique approach is known as vector quantization. It is the process of mapping vectors from a large vector space to a finite number of regions in that gap. Each region is abbreviated a cluster and represented by its centre called a code word. A codebook is a collection of all code words. Hence for multiple users there should be multiple codebooks each representing the corresponding speaker. The data is thus significantly compressed and accurately represented [13]. Without quantization of the feature vectors, computational complexity of a system would be very large as there would be large number of feature vectors. In a speaker recognition system, feature vectors are usually contained in vector space, which are obtained from the feature extraction described above. When vector quantization process goes to achievement, only remnants are a few representative vectors, and these vectors are collectively known as the speaker's codebook. The codebook then hand out as template for the speaker, and is involved when testing a speaker in the system [14][11].

A. Vector Quantization

Vector quantization are template based models for text-independent and text-dependent speaker recognition A speaker recognition system must be capable to estimate probability distributions of the computed feature vectors. Storing every single vector that is generated from the training mode is impractical, since these distributions are distinct over a high-dimensional space. It is often easier to start by quantizing each feature vector to one of a moderately small number of template vectors, with a process called vector quantization. The technique of VQ deals with extracting a small number of representative feature vectors. It is the efficient means of characterizing the speaker specific features [12].

The training features are clustered to generate a codebook for each speaker [13]. In the recognition stage, the tested speaker is compared to the codebook of each speaker and the distance is measured to identify the speaker. The problem of speaker recognition belongs to a much broader topic in scientific engineering called pattern recognition. The main goal of pattern recognition is to organize objects of interest into one of a number of classes. The objects of interest are broadly called patterns and sequences of acoustic vectors are extracted from an input speech using the techniques of vector quantization. The classes refer to individual speakers. Since the classification procedure applied on extracted features, it is also designated as feature matching [13].

Codebook Formation

Figure 5 shows a conceptual diagram to demonstrate the recognition process. In the figure, only two speakers and two dimensions of the acoustic space are shown. The circles refer to the acoustic vectors from speaker 1 and triangles denotes speaker 2. In the training phase, a speaker-specific VQ codebook is created for each known speaker by clustering his/her training acoustic vectors. The centroids are shown in the figure by circles and

triangles for speaker 1 and 2, respectively. The distance from a vector to the nearby codeword of a codebook is called a Vector Quantization distortion. In the recognition phase, an input utterance of an unknown voice is vector-quantized using each trained codebook and the total VQ distortion is calculated. The speaker corresponding to Vector Quantization codebook with smallest total distortion is recognized [13][14].

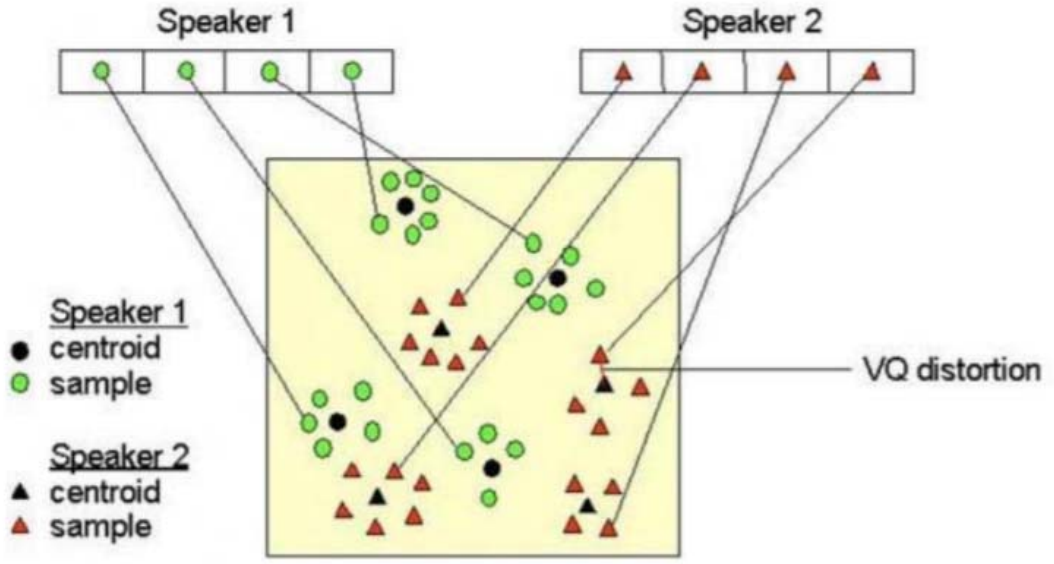

Figure 5: Codebook Formulation

B. Dynamic Time Warping

Dynamic time warping is basically template based models uses principle of dynamic programming [principle of optimality].This is used to compute overall distortion between the two speech templates. Comparing the template with incoming speech might be achieved via a pair-wise comparison of the feature vectors in each. The problem with this approach is that if constant windowpane spacing is used, the length of the input and stored sequences is unlikely to be the identical. Moreover, in case of word, there will be dissimilarity in the length of individual phonemes. The matching process needs to balance for length differences by considering the non-linear nature of the length differences within the words. This is achieved by dynamic time warping algorithm, this algorithm is used to find optimal alignment between two sequences of feature vectors, which allows for stretched and compressed sections of the sequence [7]. The two sequences of observations are positioned on the sides of a grid with the unknown sequence on down the grid and the stored template up on the left of the grid. Both sequences start position on the bottom left of the grid. Inside each cell we can establish a distance measure comparing the corresponding elements of the two sequences [14][15].

C. Gaussian Mixture Model (GMM)

GMM is a parametric method best used to model speaker identities due to the fact that Gaussian components have the capability of representing some general speaker dependent spectral shapes. The Gaussian mixture model (GMM) is the most popular models for text-independent and text dependent speaker recognition, respectively. According to the training paradigm, models can also be categorized into generative model and discriminative model. The generative models such as GMM and VQ estimate the feature distribution within each speaker. The discriminative models such as artificial neural networks (ANNs) and support vector machines (SVMs. A Gaussian Mixture Model (GMM) is a parametric probability density function signifies as a weighted sum of Gaussian component densities. GMMs are generally utilized as a parametric model of the probability distribution of continuous measurements in a biometric system, such as vocal-tract interrelated spectral features in a speaker recognition system. GMM parameters are projected from training data by iterative Expectation-Maximization (EM) algorithm or Maximum A Posteriori (MAP) estimation from a well-trained prior model[8][15][16].

A Gaussian mixture model is a weighted sum of M component Gaussian densities is represented by the equation

$$p(x|\lambda) = X\, M_i = 1\, w_i\, g(x|\mu_i, \Sigma_i) \quad (1)$$

where x denotes D-dimensional continuous-valued data vector (i.e. measurement or features), w_i , i = 1, . . . , M, are the mixture weights, and $g(x|\mu_i , \Sigma_i)$, i = 1, . . . , M, are the component Gaussian densities. Every component density is a D-variate Gaussian function of the form

$$g(x|\mu_i, \Sigma_i) = 1\ (2\pi)D/2|\Sigma_i|\ 1/2\ \exp -\ 1/2\ (x - \mu_i)'\ \Sigma - 1\ i\ (x - \mu_i) \quad (2)$$

with mean vector μi and covariance matrix Σ_i . The mixture weights convince the constraint that PM $_i$=1 w_i = 1. The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weights from all factor densities. These parameters are collectively represented by the notation,

$$\lambda = \{w_i , \mu_i , \Sigma_i\}\ i = 1, \ldots, M. \quad (3)$$

GMMs are often used in biometric systems, especially in speaker recognition systems, due to their capability of representing a large class of sample distributions. One of the powerful attributes of the GMM is its ability to form even approximations to arbitrarily shaped densities. The classical uni-modal Gaussian model denotes feature distributions by a position (mean vector) and a elliptic shape (covariance matrix) and a vector quantizer (VQ) or nearest neighbor model represents a distribution by a discrete set of characteristic templates [12][13][14].

D. Artificial Neural Networks

The Artificial Intelligence approach is a fusion of the acoustic phonetic approach and pattern recognition approach. It exploits the ideas and concepts of Acoustic phonetic and pattern recognition methods. Knowledge based approach uses the information concerning linguistic, phonetic and spectrogram. The main benefits of ANN include their discriminant-training power, a flexible architecture that permits simple use of contextual information, and weaker hypothesis about the statistical distributions. The main drawback are that their optimal structure has to be selected by trial-and-error procedures, the need to partition the available train data in training and cross-validation sets, and the fact that the temporal structure of speech signals remains complicated to handle. It can be used as binary classifiers for speaker verification systems to separate the speaker and the non speaker classes. It is also used multi-category classifiers for speaker identification purposes [6], [8], [10].

TABLE I. Speech Corpus for speaker recognition [18][19].

Name of the speech corpus	**No of Speakers**
TIMIT speech corpus	630 speakers (438 male and 192 female)
SIVA speech corpus(Italian speech corpus)	840 speakers
POLYVAR Speech Corpus	143 speakers (85 male and 58 female)
POLYCOST Speech Corpus	133 speakers (74 male and 59 female)
KING Speech Corpus	51 male speakers
YOHO Speech Corpus	138 speakers (106 male and 32 female)
SWITCHBOARD Speech Corpus	Switchboard I consists of 543 speakers and Switchboard II consists of 657 speakers
NIST 2001 SRE Speech Corpus	174 speakers (74 male and 100 female)
NIST 2002 SRE Speech Corpus	330 speakers (139 male and 191 female)
NIST 2004 SRE Speech Corpus	616 speakers (248 male and 368 female)
TSID Speech Corpus	35 speakers (four female, 31 male),
Speaker Recognition Corpus (OGI)	100 speakers consisting of 47 male and 53 female speakers
ANDOSL	129 speakers 67 female and 62 male,
Digit-SPL	males and 19 females,

V. SUMMARY

This paper represents an overview of automatic speaker recognition. The recognition accuracy of speaker recognition systems under controlled conditions is high. In feature extraction, high level features highlights behavioral characteristics of speakers, such as prosody (pitch, duration, and energy) (phonetic information pronunciation, emotion, stress, idiolect word usage conversational patterns, or other acoustic events. These differences in the speaking habits result from the manner in which people have learned to use their speech mechanism, but at the same time the sociolinguistic background, the education and the socio-economic environment plays a vital role in these differences. The main problem, as reported in different studies of this kind of systems is it's essential for more information for both training and testing phases if compared to low-level feature systems and are also easily forged. However, in practical scenario many negative factors are encountered including mismatched handsets for training and testing, limited training data, unbalanced text, background noise and non-cooperative users. The well standard techniques of robust feature extraction, feature normalization, model-domain compensation and score normalization methods are required for speaker

recognition. The technology advancement as denoted by NIST evaluations in the recent years has addressed several technical challenges such as text/language dependency, channel effects, speech durations, and cross-talk However, many research problems remain to be addressed and should be improved in human-related error sources like emotion variability, misspoken phrases, poorly recorded/noisy samples, insufficient number of comparable words, Extreme emotional states (e.g. stress or duress) Change in physical state of the speaker, Channel mismatch or mismatch in recording, different pronunciation speed speaker's health aging etc., should be carefully analyzed before implementing speaker recognition system.

VI. CONCLUSION

In today's technological world security places a great setback for confidential information. Speaker recognition is a multi-disciplinary branch of biometrics which can be used for speaker Identification and Verification for protecting confidential information. Therefore, in order to prevent un-authorized access there is a need to develop a voice based recognition system which provides a solution for financial transaction and personal data privacy that would reduce the high-tech computer theft. In this review paper various feature extraction techniques and modelling techniques used in speaker recognition is discussed which can be extended in future for developing a real time application towards speaker identification and verification system for securing confidential data.

REFERENCES

[1] Hossein Zeinali et.al., Deep Neural Networks and Hidden Markov Models in i-vector-based Text-Dependent Speaker Verification, Odyssey 2016, June 21-24, 2016, Bilbao, Spain.
[2] Local Spectral Variability Features for Speaker Verification. Digital Signal Processing Volume 50, March 2016, Pages 1–11
[3] [John H.L., Speaker Recognition by Machines and Humans, IEEE SIGNAL PROCESSING MAGAZINE [74] November 2015
[4] Varun Sharma et.al. A Review On Speaker Recognition Approaches And Challenges, International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 5, May - 2013 ISSN: 2278-0181.
[5] Proceedings of the World Congress on Engineering 2011 Vol II WCE 2011, July 6 - 8, 2011, London, U.K.
[6] Zia Saquib et.al. A Survey on Automatic Speaker Recognition Systems. Communication in Computer and Information Science January 2010.
[7] Joseph.P et.al. Speaker Recognition: A Tutorial, Proceedings of the IEEE, vol. 85, no. 9, September 1997.
[8] Citation: Mathur S, Choudhary SK, Vyas JM (2013) Speaker Recognition System and its Forensic Implications. 2: 723 doi: 10.4172/ scientificreports.723.
[9] Seiichi Nakagawa et.al. Speaker Identification and Verification by Combining MFCC and Phase Information., IEEE transactions on audio, speech, and language processing, vol. 20, no. 4, may 2012.
[10] Ms. Arundhati et.al., Speaker Identification, Signal & Image Processing : An International Journal (SIPIJ) Vol.2, No.2, June 2011
[11] Hema A. Murthy et.al., Robust Text-Independent Speaker Identification over Telephone Channels., IEEE transactions on speech and audio processing, vol. 7, no. 5, September 1999.
[12] Tomi Kinnunen et.al.,Real-Time Speaker Identification and Verification., IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 1, January 2006.
[13] Reynolds, D.A.: A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification. PhD thesis, Georgia Institute of Technology (1992).
[14] Reynolds, D.A., Rose, R.C.: Robust Text-Independent Speaker Identification using Gaussian Mixture Speaker Models. IEEE Transactions on Acoustics, Speech, and Signal Processing 3(1) (1995) 72–83.
[15] Tomi Kinnunen et.al. An Overview of Text-Independent Speaker Recognition: from Features to Supervectors, Preprint submitted to Speech Communication July 1, 2009.
[16] Dr E.Chandra et.al.,A Study on Speaker Recognition System and Pattern classification Techniques, International Journal Of Innovative Research In Electrical, Electronics, Instrumentation And Control Engineering Vol. 2, Issue 2, February 2014.
[17] Alfredo Maesa et.al.,Text Independent Automatic Speaker Recognition System Using Mel-Frequency Cepstrum Coefficient and Gaussian Mixture Models, Journal of Information Security, 2012, 3, 335-340 http://dx.doi.org/10.4236/jis.2012.34041 Published Online October 2012 (http://www.SciRP.org/journal/jis).
[18] Haris B C, G Pradhan, ET.AL.,Multi-Variability Speech Database for Robust Speaker Recognition, 978-1-61284-091-8/11/$26.00 ©2011 IEEE.
[19] Utpal Bhattacharjee et.al.,Development of a Speech Corpus for Speaker Verification Research in Multilingual Environment, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-6, January 2013.

AUTHOR PROFILE

S.Sujiya is a Ph.D. Research Scholar at Bharathiar University Coimbatore. She received her BCA Degree in Providence College for women, niligires in 2008, and MCA degree from Avinashilingam University, Coimbatore in 2011. She completed her MPhil Degree in SNS College of Arts and Science, Coimbatore in 2015. Her research interests include Speech and speaker recognition.

Dr.E.Chandra, Professor and Head in the Department of Computer Science, Bharathiar University, Coimbatore, Tamil Nadu, India. She has more than 20 years of teaching experience and 18 years of Research Experience. Her Area of Specialization includes Neural Networks, Speech Recognition System. She has authored more than 47 papers published in refereed International journals and presented 45 papers in national and International Conferences. She has obtained funding projects from UGC in the field of speech signal processing. She is the Board of Studies Member for various affiliated Institutions, Member of various Professional Society Inspection Commission and Reviewer for International Journals.