Review of Various Stages in Speaker Recognition System, Performance Measures and Recognition Toolkits

RUPALI PAWAR, Dr. R.M. JALNEKAR, Dr. J.S. CHITODE
Electronics &Telecommunication
Vishwakarma Institute of Technology Savitribai Phule University of Pune
Bibwewadi Pune
INDIA
rvspawar@rediffmail.com, director@vit.edu, j.chitode@gmail.com
http://www.vit.edu

Abstract: - Speaker Recognition is an important application of speech processing which enables to recognize the speaker with the help of unique features captured which characterise the voice of the speaker. The individual characteristics like pitch, fundamental frequency, formant frequency can be distinguishing components of the human speech signal. This paper makes an attempt to discuss various feature extraction techniques and Recognition techniques with their merits and demerits. It also discusses the pre-emphasis stage of the speaker recognition system. The standard databases available for speaker recognition along with the criterion for their selection are also reviewed. The paper gives an overview of various toolkits and performance parameters of Automatic Speaker Recognition System.

Key-Words: - Pre-processing, Framing, Feature Extraction, Generative and Discriminative Model, Toolkits, and Performance Measures: ROC, DET, EER

1 Introduction

Use of biometric system recognition has become more popular in recent days. Many applications require the authentication to confirm the identity of the individual who is requesting their service. The speech, face, iris, DNA, palm print recognition or their combination have replaced the recognition systems requiring PIN or passwords so as to secure a system. They are more secure compared to token based identification like identity card which might be lost or stolen. The biometric recognition system has overcome the necessity of remembering PIN and passwords for various banking and other important online transactions. [1]. The biometric identifications like speech or fingerprint or iris recognition characterize the unique feature of the individual and hence are more secure or reliable and preferred over other methods of identification where the possibility of infringement is comparatively higher.

Speaker Recognition is an important arena of Speech Processing and is an automatic process of recognizing who is speaking on the basis of information embedded in the speech wave. The speech wave comprises of features characterizing the speaker which are unique to a particular speaker. Speaker Recognition finds applications in Telephone banking, access control system and login to telephone aided shopping system and many more. Every Speaker Recognition System has two phases: training phase and testing phase. In the training phase the speech samples of individual speakers are accepted, pre-emphasis phase makes the speech signal noise free, the silence is removed and further processed for feature extraction. These features are stored for matching during the testing phase. In the last

few decades speech processing has found application in diverse fields. It can be used in education sector to learn and improvise the pronunciations of any language, in the domestic sector for issuing command to the various electronic gadgets like oven, washing machine etc., Robotics, Health care sector, Security and Military sector. There is increase in the applications along with the various unanswered challenges in this field like speaker variability, emotional states of the speaker, microphone characteristics, channel mismatch, room acoustics yet need to be addressed. An attempt to overcome these practical issues and challenges of speaker recognition system has attracted many

researchers to explore this domain over the past few decades.

In this paper section 2 discusses the various Speaker Recognition systems while Section 3 elaborates on the available database and its criterion for selection, Section 4, 5 & 6 emphasizes on the various stages of Speaker recognition like pre-emphasis, Feature extraction and Recognition. The performance parameters of Speaker Recognition is reviewed in section 7 while the various tool kits available for speaker recognition are discussed in section 8.

2 Speaker Recognition System

2.1 Speaker Dependent and Speaker Independent System

The system can be categorized as speaker dependent or speaker independent system. A Speaker dependent system is one which is tailored to recognize a particular speaker's speech. The speaker in such systems is initially enrolled and the features are stored in the database for further matching. The speaker independent system can recognize the speech of the speaker for whom the system is not trained for.

The speaker recognition system can also be categorized as text dependent and text independent system. In a text-dependent system, the spoken text is used to train and test the system, the system is constrained to use the same word or phrase [2].The recognition in this case is for a known set of vocabulary. In the text independent system there is no prior knowledge of the text to be spoken to the system. The recognition here can be for words the system is not trained for. The text independent recognition is more difficult but also more flexible. [3] The vocabulary used for the system can be of varied types: isolated words, connected words, continuous words or spontaneous words. The vocabulary can be small (tens of words), medium (hundreds of words) or large (ten thousands of words) very-large vocabulary (tens of thousands of words.) [4]. The Isolated word recognizers accept single word or utterances at a time unlike connected word recognizers which allow more than one utterance with some pause in between them. The continuous word recognizers allows the speaker to read a sentence or paragraph while spontaneous word recognizers accepts a natural speech of the speaker, which might be incorporated with natural pauses and utterances like 'aahs' and 'hmms'. The continuous and spontaneous word recognizers are challenging to implement as compared to the isolated and connected word recognizers as the silence and the word boundaries need to be taken care of. Algorithms for end point detection and silence removal can be used to detect the boundaries and remove the silence.

2.2 Speaker Identification and Speaker Verification

Speaker Recognition system performs two basic tasks: Speaker Identification and Speaker Verification. In a system which is trained for a set of speakers, the task of determining which amongst the set of speakers is speaking is Speaker Identification. This is also called as closed set speaker identification as population of speakers is known and one amongst the given set of speaker's voice is tested. The speaker to be tested belongs to the existing (enrolled) database of speakers. If the speaker to be tested does not belong to the trained database then such identification is called open set identification. Thus the closed set identification is where the group members are known, their speaker profile can be acquired and stored in the database. In open-set identification the unknown individual can come from general population [5]. Identification is always carried out against a finite, known pool of individuals, it is not possible to identify arbitrary people.

Speaker verification is the task of determining whether the speaker is the one who he claims to be. During speaker verification the features of the speaker to be verified are tested or compared with the claimed identity present in the trained database. If the two match and cross a certain threshold then the claimed speaker is accepted if it doesn't then the claim is rejected. The various phases of speaker recognition system are Pre-emphasis, Analysis, Feature extraction and Recognition as depicted in Figure 1.

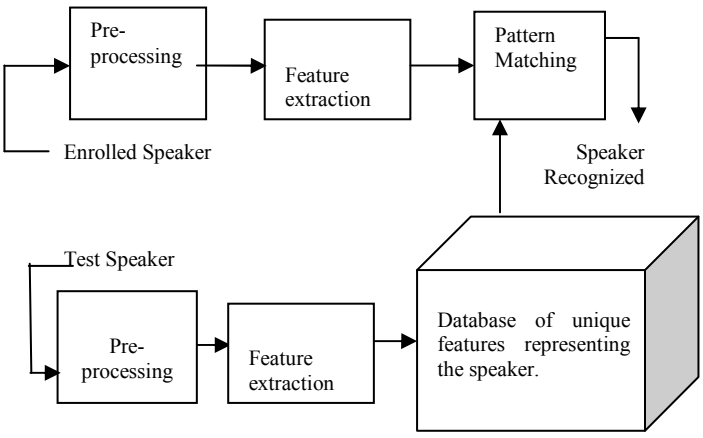


Fig. 1: Block Diagram of Speaker Recognition System

3. Database For Speaker Recognition

In a Speaker Recognition System the voice to be used for training phase can be recorded or a standard database can be used. In the last few decades with increase in speech related applications-recognition, forensic and many more the necessity of varied database has increased. The standard database available for varied applications allows the researchers to evaluate and compare the performance of a system with the existing system and thereby verify and validate the results. The different audio formats that can be used to record a speech are Uncompressed Audio Format, Lossless Compressed Audio Format and Lossy Compressed Audio Format. The audio format most commonly used by researchers is the wav format or the waveform audio file format. It is a standard format used mainly in Windows PCs. A wav file has 3 chunks: the RIFF chunk, the FORMAT chunk and the DATA chunk. The RIFF chunk is 12 bytes, the FORMAT chunk is 24 bytes and the DATA chunk has variable number of bytes depending on the length of data. The format chunk specifies the data format, length of the FORMAT chunk, information about the sample rate, channel number and more,while the data chunk contains the actual sample data. [6]

The necessity of proper data selection in any Speaker Recognition System is to reduce the time required for further pre-processing. Depending on the necessity of the application and research area Standard database and native language database is used by many researchers. The data selection for speaker recognition should satisfy certain criterion. [7]

3.1 Less Intra Speaker Variability

The intra-speaker variation can originate from a variable speaking rate, changing emotions and environmental noise. The variance brought by different speakers is denoted as inter-speaker variance and is caused by the individual variability in vocal systems involving source excitation, vocal tract articulation, lips and/or nostril radiation. If the inter-speaker variability dominates the intra-speaker variability, speaker recognition is feasible. Multiple sessions of recording helps estimating intra speaker variability [8].The recorded wave files during the time of training and testing should have less intraspeaker variability for a robust speaker recognition system.

3.2 Good Performance

The performance of a system should be good even if it is trained for larger corpus. The recorded voice should be noise free. The database selected should have utterances that should address the variation in large population. Usually it is observed that the performance of a system degrades as the vocabulary changes from small to large or from isolated words to continuous or spontaneous words. [9] A robust system should perform well under any condition the database should have files recorded under noise free environmental like sound booths.

Some available standard databases are listed in Table 1. [8] [10] [11]

Name	No. of Speakers	No. of Utterances	Speaking Style	Sampling Frequency
TIMIT	530	6300 sent.	Reading	16khz
NTIMIT	530	6300 sent.	Reading	8 khz
RM1	144	15024 sent.	Reading	20khz
RM2	4	10608 stent	Reading	20khz
Switch Board	69	35 dialogues	Continuous /Spontaneous	8khz
ATIS2	351	12000 utterances	Spontaneous	16khz
ELSDSR [10]	22 speakers	11 sentences	Reading	16 kHz
Yoho[11]	138 speakers	1380 verification sessions	Reading	8kHz

Table 1: Some available standard databases

4. Pre-Processing

In speaker recognition applications where the input audio file can be from a standard database or recorded, the first step would be pre-processing of the signal which involves endpoint detection, silence and noise removal, framing, pre-emphasis etc. Removal of unvoiced or silence from the signal and endpoint detection is crucial for speaker recognition. [12].

4.1 Pre-Emphasis

Noise can significantly decrease the performance of a speaker recognition system. To solve this problem, many noise reduction algorithms have been developed. Most of them assume that the noise level is constant, or is to be evaluated in the course of the algorithm. In order to the flatten speech spectrum, a pre-emphasis filter is used before spectral analysis. Its aim is to compensate the high-frequency part of the speech signal that was

suppressed during the human sound production mechanism. The most used filter is a high-pass FIR filter whose transfer function is

$$H(z)=1-az-1$$

4.2 Categorization as silence, voiced and unvoiced

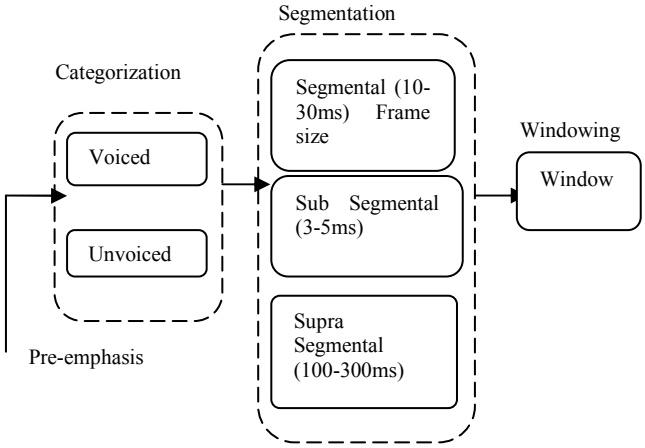
Speech sounds are composed of sequence of sounds called phonemes produced as a result of acoustical excitation of the vocal tract. The vocal tract shape is determined from the position of the vocal organs, and speech is produced by controlling the speech production model using the vocal tract area. [12]

The cross section area of the vocal tract varies from 0 to 20 square cm and is dependent on the position of the articulators. Nasal tract begins at the soft palate called velum and ends at the nostrils. Depending on whether the vocal cord vibrates sound produced can be broadly categorised as voiced and unvoiced sound. The vocal cords are tensed for sounds like a/e/i and vibrate to produce voiced sound. The vocal cords vibrate periodically and when air flows from the lungs resulting in a speech waveform which is quasi-periodic in nature. Air flows through vocal cords into vocal tract in discrete puffs. The vocal cords do not vibrate for sounds like s/f resulting in aperiodic or random speech waveform called unvoiced sound. The separation of the speech signal into voiced, unvoiced, and silence provides a preliminary acoustic segmentation of speech, which is important for speech analysis. [13]

Short time energy (STE), short time zero crossing (STZC) and short time autocorrelation are some of the techniques which can be used to separate voiced, unvoiced and silence in a speech signal. Combination of STE, STZC or auto correlation can enhance the accuracy of such categorization.

4.3 Segmentation

Speech data contains features that convey speaker's identity. This includes speaker –specific information due to the excitation source, vocal tract and behavioural traits .To obtain good representation of these speaker characteristics, speech data can be analysed using segments of different duration : sub segmental (3-5mS), segmental(10-30mS) and supra segmental analysis (100-300mS) respectively. The sub segmental analysis helps in capturing the information due to the excitation source, while the segmental analysis includes the information due to movement or constriction of the vocal tract. The frames of larger duration in Supra segmental analysis extract information due to behavioural characteristics of a speaker. Features like duration, intonation and energy fall in this category. The formants, cepstral coefficients can be the useful features extracting information due to movement of the vocal tract, categorizing it under segmental analysis. The excitation source features like pitch and glottal source parameters fall under sub segmental analysis. [14].



4.4 Windowing

The short term time domain analysis helps in computing the time domain features like energy, zero crossing and auto correlation. The spectral features of speech signal are not apparent in time domain and hence the necessity of frequency domain analysis.

Speech is a non-stationary signal hence computing the frequency spectrum of the whole signal will not help us fetch speaker specific information. Frame of 10-30 mS of a speech signal will give the significant time-varying spectral information. The formant frequencies plotted for voiced and unvoiced speech are different. For voiced speech the magnitude of lower frequencies is successively larger than the magnitude of the higher formant frequencies (enhances low frequency & suppresses high frequency) and vice-versa for unvoiced speech.

The next step in processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. Windowing is a point wise multiplication of the frame and the window function. The Hamming window is popularly used, the hamming window function is given as:

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right)$$

$$0 \le n \le N \quad .$$

Where N = size of frame.

In [15] two approaches of window function: fixed and adjustable are discussed. It justifies the use of Hanning and Hamming window over rectangular and Blackman window due to its optimum main lobe and good side lobe attenuation. The adjustable Kaiser window has the disadvantage of higher computational complexity due to the use of Bessel functions in the calculation of the window coefficient.

5. Feature Extraction

Feature extraction is the process of retaining useful information of the signal which contribute to unique identity of a speaker. It discards redundant and unwanted information, however, in practice, while removing the unwanted information, one may also lose some useful information. [16]. Feature extraction may also involve transforming the signal into a form appropriate for the models used for recognition. Extracted features should accomplish some criteria while dealing with the speech signal such as:

Have less speaker variability
Be robust against noise and distortion
Occur frequently and naturally in speech
Be difficult to impersonate/mimic
Not be affected by the speaker's health or long-term variations in voice. [17]

Some of the feature extraction techniques are Linear Predictive Coding (LPC), Linear Prediction Cepstral Coefficient (LPCC), Mel Frequency Cepstral Coefficient (MFCC), AMFCC (Autocorrelation MFCC) perpetual linear prediction (PLP), Wavelet.

5.1 Linear Predictive Coding

LPC is a model based on a mathematical approximation of the vocal tract represented by a tube of varying diameter [l8] LPC is the most common technique of spectral analysis. Linear prediction models the human vocal tract as an infinite impulse response (IIR) system. The vocal tract forms the tube, which is characterized by its resonances, which gives rise to formants. LPC analyses the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The process of removing the formants is called inverse filtering, and the remaining signal after the subtraction of the filtered modelled signal is called the residue.

5.2 Mel Frequency Cepstral Coefficients

MFCC is popularly used as it represents the human auditory system well. The MFCC is a representation defined as the real cepstrum of a windowed short-time signal derived from the fast Fourier transform of the speech signal. [17]. It transforms the signal in frequency domain and is less prone to noise. The Mel scale is represented by following formula

$$Mel_f = 2595 \ln (1+f/700)$$

Mel_f is the Mel frequency in Mel and f is the linear frequency in Hz. Figure 3 shows the steps involved in computing MFCC of the signal.

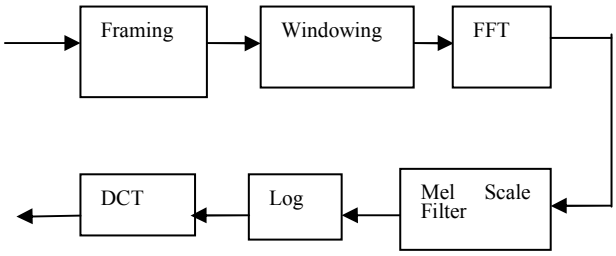


Fig 3: Feature Extraction Using MFCC

5.3 Some other approaches of Feature extraction

BFCC (Bark Frequency Cepstral Coefficient): It is similar to MFCC implementation with where Mel bank filters are replaced by Bark scale filters. Mathematically bark scale filters are represented by the formula

fbark=[6ln(f/600+(f/600)2+1)0.5] [19]

Perceptual Linear Prediction (PLP) is based on the human auditory system and it discards the irrelevant speech information to extract the features. PLP and MFCC are based on human auditory system and logarithmically spaced filters these techniques have good response than that of the LPC. [14]

LPCC is easily obtained by the least square method using a set of recursive formula. LPCC needs much less computational time to be extracted from speech signal compared to MFCC. Noise robust spectral estimation is possible using AMFCC (Autocorrelation MFCC) and BFCC (Bark Frequency Cepstral Coefficient). [17]

In [19] comparative analysis of MFCC, LPCC and MFCC is done for Hindi vocabulary. The MFCC proves to be better than conventional LPCC and BFCC feature extraction techniques. Table 2 puts forth literature survey of comparison of features extracted using different techniques with regards to their performance.

MFCC-LPCC	Used in less intra speaker variability and availability of rich spectral analysis tool. Performance better than MFCC or LPCC used independently [20]
MFCC	Frequency domain Represents human auditory system Robust and reliable to noise and estimation errors [18] Better than LPC[14] or LPCC[21]
LPC	Common technique of spectral analysis

	Time domain Suffers variation in amplitude of the speech signal due to noise[22]
LPCC	Robust and reliable to noise and estimation errors[17]
PLP	Based on Human auditory system[17] Better response than LPC

Table 2: Comparison of some Feature Extraction Techniques

6. Speaker Recognition

The speaker recognition can be categorized under Acoustic Phonetic, Pattern Recognition, Artificial Intelligence Approach and the Connectionist approach. The Pattern recognition can be either template based or stochastic approach. The template based approach has advantage over Acoustic phonetic approach as errors due to classification or segmentation of smaller units like phonemes can be avoided [21]. The template based approach becomes expensive and impractical with increase in vocabulary. The stochastic approach being a probabilistic approach can be dealt with data having uncertain or missing information, it also handles speaker variability, confusing sounds and homophones. The ANN is a hybrid of the two approaches mentioned above.

Some of the modelling techniques are Vector Quantization, Dynamic Time warping, Gaussian Mixture Model, Hidden Markov Model, Support Vector Machine and Neural Networks. According to the training paradigm, models can also be classified into generative and discriminative models.

6.1 Generative and Discriminative Model

A generative model is a model for randomly generating observable data, given some hidden parameters. It specifies a joint probability distribution over observation and label sequences. Generative models are used in machine learning for either modeling data directly (i.e., modeling observations drawn from a probability density function), or as an intermediate step to forming a conditional probability density function. A conditional distribution can be formed from a generative model through Bayes' rule. Generative models include: Gaussian mixture model and other types of mixture model.

Discriminative models, also called conditional models, are a class of models used in machine learning for modeling the dependence of an unobserved variable 'y' on an observed variable 'x' Within a probabilistic framework, this is done by modeling the conditional probability distribution P(y|x).Discriminative models used in machine learning include: Support Vector Machine, Neural Network and Linear Regression.

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) in contrast, model the boundary between speakers. [17]. Combination of the two generative and discriminative which are also called as hybrid models can be used to enhance the recognition rate.

6.2 Vector Quantization (VQ)

Quantization is an important aspect of data compression or coding and is the process of approximating continuous amplitude signals by digital (discrete amplitude) signals. The independent quantization of each signal value or parameter is termed scalar quantization, while the joint quantization of a block of parameters is termed block or vector quantization. The utility of VQ in ASR lies in the efficiency of using compact codebooks. The approach of combining advantages of VQ with DTW helps overcome quantization errors and improves rate of recognition. [22]

6.3 Dynamic Time Warping (DTW)

Dynamic Time Warping is a dynamic programming technique in which the entire problem is divided into a small number of steps each requiring a decision to be made based on the local distance measures. The overall decision is made depending on these smaller decisions. Thus it uses a divide and conquer approach. DTW aims to overcome one of the prime causes of variability in speech; the global and local variation in speech rate. This is achieved using non-linear time alignment. [23]

In [24] the system is speaker dependent and is tested for 10 English digits .The technique called as crosswords reference templates (CWRT) is used to generate the reliable templates to improve the recognition rate from 85.3% to 99%.

6.4 Gaussian Mixture Model (GMM)

Gaussian Mixture Model is a stochastic model which has become the de facto reference method in speaker recognition. GMM is the most frequently and successfully employed density estimators in speaker recognition [25]. The GMM can be considered as an extension of the VQ model, in which the clusters are overlapping. [16]. In GMM each speaker has the independent GMM model. Text independent (TI) recognition can be done using Gaussian mixture models.

According to [26], the Gaussian models are somewhat crude in which they model the gross characteristics of the speaker's distribution.

6.5 Hidden Markov Model (HMM)

Hidden Markov Models (HMMs) are learnable finite stochastic automates. A Hidden Markov Model consists of two stochastic processes. The first stochastic process is a Markov chain, externally not visible that is characterized by states and transition probabilities. The second stochastic process produces emissions observable at each moment, depending on a state-dependent probability distribution. [27]

[28] Proposes the Genetic Algorithm for training HMM algorithm. Six Chinese vowels are accepted as data for experimentation. It proves to be efficient method with improved speed and accuracy especially for speech signal which is degraded by noise.

7. Performance of ASR

The performance of the Automatic Speaker Recognition system can be analysed based on accuracy and the speed of the system. The accuracy of a system can be a measure from False Acceptance Ratio and False Rejection Ratio. False acceptance Ratio is the percent of invalid inputs which are accepted as valid while False Rejection Ratio is the percent of valid inputs which are rejected assuming they are invalid.

Some other performance measures for a speaker recognition system, Decision Error Rate (DET), Equal Error Rate (EER), Half Total Error Rate (HTER) are discussed in this section.

7.1 Receiver Operating Characteristic (ROC)

The measure of accuracy was calculated by plotting the ROC curves (introduced in 1950). Plotting error rates (FA and FR) against each other in order to find appropriate confidence thresholds for the classification of the recognition results. After completion of recognition a decision has to be made about the correctness compared with the input speech. When the speech content is unknown, the decision about the correctness of the ASR result is made on the basis of a confidence score and a confidence threshold. If the confidence score is below the threshold the recognized result is incorrect else it is correct. There are four possibilities correct rejection (CR), false rejection (FR), false acceptance (FA) and correct acceptance (CA). In order to find a proper trade-off between (FR) and (FA), ROC curves are used for finding the best solution.

Two types of ROC curves are commonly used.

1. Plot of a detection rate against an error rate.
2. Plot of one error rate against another error rate. This is also called Decision error trade off or DET [29]

7.2 Equal Error Rate EER

The receiver operating characteristic (ROC) curve adopted from psychophysics is used for evaluating speaker recognition systems. Two conditions are considered for the input utterances: s, the condition that the utterance belongs to the speaker, and n, the condition that utterance doesn't belong to the speaker. Two decision conditions also exist: S, the condition that the utterance is accepted as that of the speaker, and N, the condition that the utterance is rejected. These conditions combine to make up the four conditional probabilities.

P(S|s) is the probability of correct acceptance.

P(S|n) the probability of false acceptance (FA)

P(N|s) the probability of false rejection (FR)

P(N|n) the probability of correct rejection.

Since the relationships

$P(S|s)+P(N|s)=1$ and $P(S|n)+P(N|n)=1$

Speaker Recognition Systems can be evaluated using the two probabilities P(S|s) and P(S|n). If these two values are assigned to the vertical and horizontal axes respectively, and if the decision criterion (threshold) of accepting the speech as being that of the speaker is varied.

Figure 4 shows the performance of B is always superior to that of A and C corresponds to a chance performance. The curve of error rate Vs the decision criterion 'a' depicts the tight decision criterion, where it makes the imposter difficult to falsely accept the system. Thereby increasing the possibility of rejecting speaker's .The loose criterion depicted by 'b' allows the speaker to be consistently accepted also falsely accepting imposters. 'a' and 'b' represent strict and lax decision criterion respectively.

If the FR rate is specified, the corresponding FA rate is obtained as the intersection of the ROC curve with the vertical line indicating the FR rate. EER corresponds to a threshold at which the FR rate is equal to the FA rate as indicated by c in Figure 4b. The criterion is usually set a posteriori for each

individual speaker or for a set of test speakers. The EER point corresponds to the intersection of the ROC curve with the straight line of 45 degrees, indicated in Figure 4a. Although the EER performance measure rarely corresponds to a realistic operating point, it is quite a popular measure of the ability of a system to separate impostors from speaker. [30]

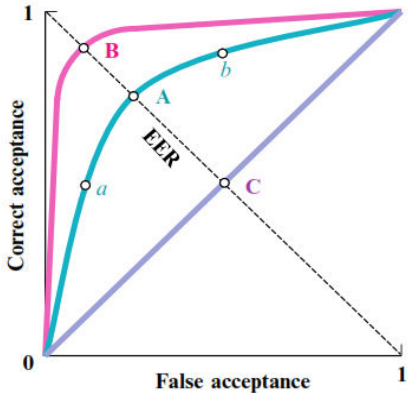


Fig 4 (a)[30]
Performance charachteristics of Speaker A,B and C.

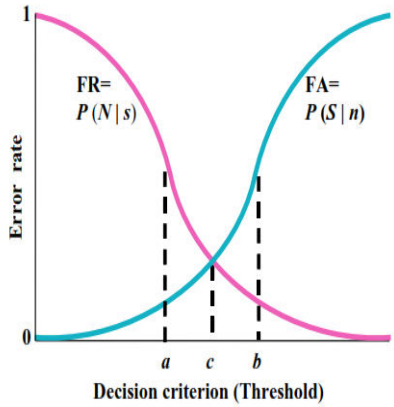


Fig 4(b)[30]
Relationship between error rate and decision criterion

7.2.1 Half Total Error Rate (HTER)

Another popular measure is the half total error rate (HTER), which is the average of the two error rates FR and FA. It can also be seen as the normalized cost function assuming equal costs for both errors, where cost function in NIST speaker recognition evaluations is defined as weighted sum of the two types of errors.[30]

7.3 Decision Error Trade-off (DET)

Recognition behaviour is well captured by DET curve which shows the relationship between the probability of miss P(m)and probability of false acceptance P(f).DET is the essence of ROC curve plotted on the axis scaled by normal derivatives rather than linearly[31]

It has recently become standard to plot the error curve on a normal deviate scale (Martin et al., 1997), in which case the curve is known as the detection error trade-offs (DETs) curve. With the normal deviate scale, a speaker recognition system whose speaker and impostor scores are normally distributed, regardless of variance, will result in a linear scale with a slope equal to -1.The better the system is, the closer to the origin the curve will be. In practice, the score distributions are not exactly Gaussian but are quite close to it. The DET curve representation is therefore more easily readable and allows for a comparison of the system's performances over a large range of operating conditions.[30]

8. Speaker Recognition Toolkits

Exploring the domain has become more convenient today as researchers who are not able develop a complete speech recognition system can focus on a specific problem by using speech recognition toolkits.

Some of the Speaker Recognition toolkits available are listed below in Table 3

Sr. No.	ToolKit	Platform	Features/Objectives
1	ALIZE	C++	Open Source[32] To encourage the laboratories to evaluate new proposals using the toolkit and both standard databases and protocols like NIST SRE ones;
2.	SPEAR	Combination of Python & C++	Open Source Includes all the processing stages from the front-end feature extractor to the final steps of decision and evaluation. Gaussian mixture models, inter-session variability, joint factor analysis and total variability (i-vectors). The tool chains can be easily evaluated on well-known databases such as NIST SRE and MOBIO[33]

3.	MSR identity toolbox	Matlab	provides tools for speaker recognition using both the GMM-UBM and i-vector paradigms and performance measures EER[34]

Table 3: Some of the Speaker Recognition toolkits available

9. Applications & Challenges of Speaker Recognition System

There is usually a trade-off between Recognition rate and the number of enrolment sessions and duration of the enrolment, test session duration of speech. [35] Apart from choosing better algorithms one needs to take into account various other factors that contribute to the accuracy of the recognition system. The acoustics of the room where the speech is recorded, the quality of microphone, the distance between the speaker and the microphone amounting to echo in the recorded sound or delay in speech during the recording session, channel mismatch are some of the challenges which need to be taken care of. Certain other challenges which need to be addressed from speaker's point of view are the accent and the speed of reading in the train and the test phase. Apart from this the emotional state of the speaker (anxiety, stress etc.), variation in voice due to cold or cough, age of the speaker.

The speaker recognition system finds application in wide area, it can be used in access control system to public safety. A robust Speaker recognition system can be used for authentication of the user in finance and other domains, can play a key role in forensic , biometrics, computer network security, corpus, databases, identification of persons, site security monitoring.

10. Conclusion

This paper gives a survey about various phases used in speech recognition. The availability of standard databases and various steps involved in pre-processing of a signal are discussed. The paper reviews the feature extraction, recognition techniques and performance measures of Speaker Recognition System. This paper makes an attempt to give an overview of various toolkits that can be used for speech Recognition. Although there is a lot of work done in the field of speaker recognition in the past few decades yet there is scope of improving the performance with regards to recognition rate and robustness of a system. Challenges with regards to intra and inter speaker variability, noisy environment, choice of microphone, channel disturbances need to be addressed allowing researchers to explore this domain and propose a robust system in all respects.

References:

[1] Jain, A. Ross, and S. Prabhakar, An introduction to biometric recognition, *IEEE Trans.Circuits Systems Video Technology*, Vol.14, No.1, 2004, pp. 4–20.

[2] Douglas A. Reynolds, Automatic speaker recognition using Gaussian mixture speaker models," *The Lincoln Laboratory Journal* 1995

[3] Douglas A. Reynolds,An Overview of Automatic Speaker Recognition Technology, *Proceedings of ICASSP"*, 2002,pp 300–304

[4] http://www.speec.cmu.edu/comp.speech/ Section6/Q6.

[5] Roberto Togneri and Daniel Pullella, An overview of Speaker Identification: Accuracy and Robustness Issues", *IEEE Circuits and Systems Magazine*, 2011 pp 23-61

[6] Dr. Shaila Apte, "Speech And Audio Processing, "ISBN 13:9788126534081,Wiley Publication

[7] Arkadiusz Nagórski, Lou Boves Herman Steeneken, In Search of Optimal Data Selection For Training of Automatic Speech recognition, *Automatic Speech Recognition and Understanding.* 2003, IEEE Workshop.

[8] M.A.Anusuya and S.K.Katti "Speech Recognition by Machine: A Review" *International journal of computer science and Information Security,* 2009

[9] Wu, Rong Zhang, Alexander Rudnick , Data Selection for speech Recognition , *Workshop on Automatic Speech Recognition and Understanding,* 2007IEEE

[10] http://www.imm.dtu.dk/

[11] yoho data Testing with The YOHO CD-ROM Voice Verification Corpus *Joseph P.Campbell*

[12] Honda M., (2003),"Human Speech Production Mechanisms" *NTT Technical Review*, Vol 2

[13] Qi Y.,Hunt R.B.,Voiced -Unvoiced-Silence Classifications of Speech using Hybrid Features and a Network Classifier*", IEEE Transactions on Speech And Audio Processing*, 1993, Vol.1,No. 2

[14] H.S. Jayanna , S.R.Mahadeva Prasanna, Analysis, Feature Extraction, Modeling and Testing Techniques for Speaker Recognition, *IETE Tech Rev*, 2009, 26:181-90

[15] Saurabh Singh Rajput , Dr.S.S. Bhadauria, Implementation of FIR Filter using efficient Window Function and its Application in Filtering a speech Signal, *International Journal of Electronics, Electrical and Mechanical Controls* , 2012

[16] Kesarkar M. Feature Extraction For Speech Recogniton"M.Tech. *Credit Seminar Report, Electronic Systems Group, EE. Dept,IIT Bombay*.2003

[17] Tze Fen Li , Shui-Ching Chang,Speech recognition of mandarin syllables using both linear predictive coding cepstra and Mel frequency cepstra ", ELSEVIER, *Pattern Recognition,* Volume 36, Issue 11, November 2003, Pages 271–272

[18] Performance Analysis of Lip synchronization using LPC, MFCC and PLP Speech Parameters, *International Conference on Computational Intelligence and Communication Networks*

[19] Taabish Gulzar , Anand Singh, Sandeep Sharma "Comparative Analysis of LPCC, MFCC and BFCC for the Recognition of Hindi Words using Artificial Neural Networks, *International Journal of Computer Applications (0975 – 8887)* Volume 101– No.12, September 2014

[20] S. B. Davis and P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoustic, Speech, Signal Processing,* 1980, 28(4), 357- 366.

[21] Wiqas Ghai, Navdeep Singh, Literature Review on Automatic Speech Recognition *International Journal of computer Application*, 2012 Vol 41-No-8 pp 42-50

[22] R.K.Moore, "Twenty things we still dont know about speech ", *Workshop on Progress and Prospects of speech Research and Technology* , 1994

[23] Bin Amin T. and Mahmood I., Speech Recognition using Dynamic Time Warping, *2nd International Conference on Advances in Space Technologies*, 2008,pp.74-79

[24] Abdulla W.H., Chow D. and Sin G.,Cross-words Reference Template for DTW-based Speech Recognition Systems", *Conference on Convergent Technologies for Asia-Pacific Region* 2003 vol 4.

[25] Sheeraz Memon, Margaret Lech, Namunu Maddage, Information Theoretic Expectation Maximization based Gaussian Mixture Modeling for Speaker Verification, *International Conference on Pattern Recognition , IEEE Computer Society* 2010

[26] H. Gish and M. Schmidt. , Text-independent speaker identification. *IEEE Signal Processing Magazine*,1994, pp 18-32,

[27] Hidden Markov Models, Theory and Applications Edited by Przemyslaw Dymarski, ISBN 978-953-307-208-1, 326 pages, Publisher: *InTech, Chapters published* April 19, 2011 under CC BY-NC-SA 3.0 license DOI: 10.5772/601

[28] Zhao Lishuang , Han Zhiyan , "Speech Recognition System Based on Integrating feature and HMM", *International Conference on Measuring Technology and Mechatronics Automation* 2010

[29] Tibor Fabian Confidence Measurement Technique in Automatic Speech Recognition and Dialog Management *Lehrstuhl fur Mensch-Maschine-Kommunikation Technische Universit¨at Munchen,* 2008

[30] Peter J. Barger, Sridha Sridharan, On the Performance and Use of Speaker Recognition Systems for Surveillance, *Proceedings of the IEEE International Conference on Video and Signal Based Surveillance (AVSS'06)*0-7695-2688-8/06, 2006

[31] Sadaoki Furui,, Speech and Speaker Recognition Evauation, *L. Dybkjær et al. (eds.), Evaluation of Text and Speech Systems, 1–27.* 2007 Springer.

[32] Jean-Francois Bonastre, Frederic Wils , Sylvain Meignier "ALIZE, A FREE TOOLKIT FOR SPEAKER RECOGNITION", *ICASSP* 2005

[33] Elie Khoury, Laurent El Shafey, Sebastien Marcel, "Spear: An open Source Tool Box for Speaker Recognition Based on Bob", *ICASSP*_2014

[34] Seyed Omid Sadjaidi , Malcolm Slaney and Larry Heck, "MSR Identity Tool box : A Matlab Toolbox for speaker Recognition Research"

[35] Joseph P. Campbell, Speaker Recognition