Speaker Recognition Techniques: A Survey

[1]Umaisa Hassan, [2]Sukhvinder Kaur, [3]Dr. Rahul Malhotra

[1]Student, [2]Assistant Professor, [3]Principal
[1]Electronics and Communication Engineering
Haryana, India

Abstract: Speech is the most usual form of human communication and is an individual's unique characteristics. Speech processing has emerged as one of the important application area of digital signal processing. The objective of automatic speaker recognition is to extract, characterize and recognize the information about speaker identity. In this, the features of speech signal are first extracted using Discrete wavelets transform (DWT), Linear Predictive Coefficients (LPC), Irregular Discrete Wavelet Packet Transform (DPWT), Pyknogram, Mel Frequency cepstral coefficient (MFCC)). For verifications these features are matched using feature matching algorithms T-Test distance matrix, Kullback–Leibler Distance Metric (KL2), and Bayesian Information Criterion (BIC). Its performance can be evaluated by using Detection Error Trade-off (DET) & Receiver Operating Characteristics (ROC) curve. It can be implemented in biometric, Access Control, transaction authentication, Law Enforcement, speech data management.

Index Terms **– Discrete wavelet transform(DWT), Pyknogram, Mel Frequency cepstral coefficient (MFCC), Non-linear energy operator T-Test, Kullback–Leibler Distance Metric (KL2), Bayesian Information Criterion (BIC), Detection Error Trade-off (DET) and Receiver Operating Characteristics (ROC).**

I. INTRODUCTION

Speech is the most natural way for human communication. It conveys several types of information from the speech production and perception point of view. The speech signal conveys message and language information also called linguistic information. Also, information about speaker's emotional and physiological characteristics can be obtained from speech signal[1]. It also gives information about the environment in which the speech was produced and the medium through which it was transmitted. Hence, speech signal carries lots of information which is encoded in a complex form. Humans can effortlessly decode most of this information. This has inspired researchers to develop systems that automatically extract and process the huge amount of information in speech.

Some of the broad areas where speaker recognition technology is being currently used are discussed below:

- Access Control: Speaker recognition systems are used for physical facilities. Recently, they are used for providing access to computer networks or websites .It is also used for automatic resetting of password.
- Transaction Authentication: For telephone banking it can be used for account access control. It can also provide more security while verification for more sensitive transactions. It can also provide verification of user for remote electronic and mobile purchases (e- and m-commerce).
- Law Enforcement: Speaker recognition systems are used for home-parole and prison call monitoring. A check can be kept on parolees by calling them at random times to confirm that they are at home. Similarly, inmates can be monitored before any outgoing call. These speaker recognition systems can be used to testify aural/spectral inspections of voice samples for forensic analysis.
- Speech Data Management: It can be used in voice mail browsing or intelligent answering machines. Speaker recognition can be used to tag incoming voice mail with speaker name. It can also be used for speech skimming or audio mining applications. Recorded meetings or video can be commented with speaker tags for indexing and filing.
- Personalization: In voice-web or device customization, speaker recognition systems can be used to get personal settings based on user verification for multi-user site or device.

The remaining of this paper is organized as follows: section 2 introduces speaker recognition, section 3 contains feature extraction, section 4 contains feature matching, and section 5 explains performance and evaluation and conclusion is described in the last section.

II. SPEAKER RECOGNITION

Speaker recognition is the identification of a person from characteristics of voices. Recognizing the speaker can simplify the task of translating speech in systems that have been trained on specific voices or it can be used to authenticate or verify the identity of a speaker as a part of security process. Each speaker recognition system has two phases enrollment phase & verification phase. During enrollment, the speaker's voice is recorded and typically a number of features are extracted to form a voice print. For identification systems, the utterance is compared against multiple voice prints in order to determine the best match while verification systems compare an utterance against a single voice print. Because of the process involved, verification is faster than identification. Speaker recognition falls into two categories text dependent & text independent. Fig.1 shows the block diagram of speaker recognition system.

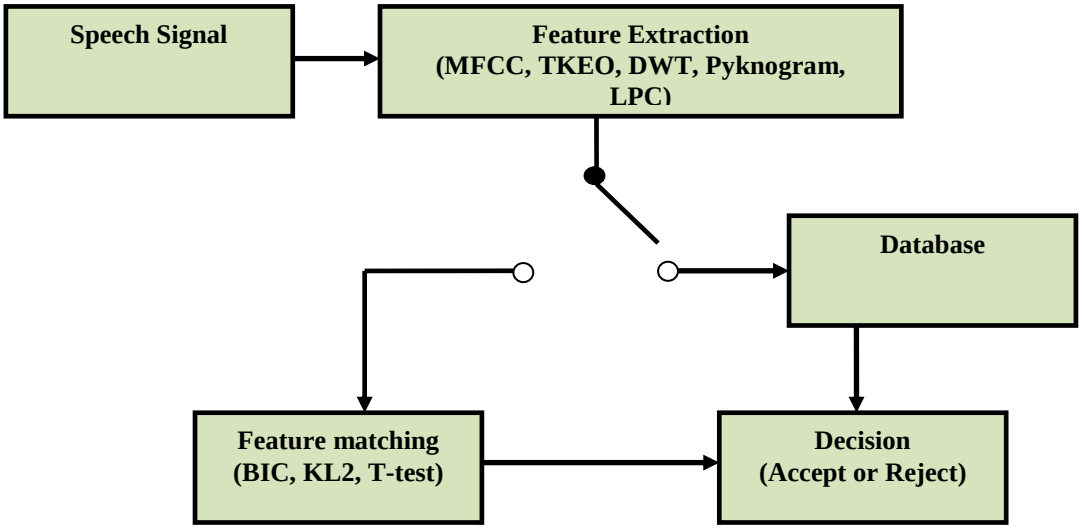

Fig1 : Block diagram of speaker recognition system.

2.1 Front Processing

It is the process of conversion of high quality speech signal into a comparatively low quality speech signal subspace by keeping the speaker discriminative content. **.** The speaker's voice signal contains both the features of speech and the speaker personality characteristics. And also the speaker's model is not derived from the speech signal, but received by extracting features from the speech signal. In other words, the model is a speech feature model of speaker. The test tone is compared and matched to the speaker's model only after features parameter extracted, and the training speech get the model after feature extracted so feature extraction is important part in speaker recognition system. It also computes the performance of the recognition system Feature extraction process includes a sampling, a quantization, a pre – emphasis, a windowing, and a feature extraction.

2.2 Pre-emphasis

Pre-processing of a signal includes analog to digital conversion, filtering of undesired signal, segmentation into different frame size because of negative spectral in the natural speech signal. The slope of this negative speech signal is nearly 20dB/decade because of the physiological feature of speech production. The original physiological signal has a small amplitude at high frequency in comparison to low-frequency formants[2]. Energy in human voice signal decreases as the frequency increases. Energy also increases in different parts of the level of the signal by the amount inversely proportional its frequency.

2.3 Windowing

The next step in the processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. The concept here is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame. If we define the window as $w(n)$, $0 \leq n \leq N-1$, where N is the number of samples in each frame, then the result of windowing of signal is $y(n) = x(n) * w(n)$ where $x(n)$ is the speech signal being processed. Windowing is followed by framing, where the speech signal is made stationary by dividing it into overlapping fixed duration segments called frames which can capture the speaker specific characteristics.

III. FEATURE EXTRACTION

Feature extraction requires much attention because recognition performance relies heavily on the feature extraction phase. It is very difficult to obtain the data that embedded in the speech signal, so features are extracted from speech signal. Different techniques for feature extraction are LPC, MFCC, Pyknogram, DPWT, and Teager Kaiser Operator.

3.1 Linear Predictive Coefficients (LPC)

LPC (Linear Predictive coding) analyzes 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 is called the residue. In LPC system, each sample of the signal is expressed as a linear combination of the previous samples[3]. Human speech is produced in the vocal tract which can be approximated as a variable diameter tube. The linear predictive coding (LPC) model is based on a mathematical approximation of the vocal tract represented by this tube of a varying diameter. At any particular time, t, the speech sample s (t) is represented as a linear sum of the p previous samples. The most important aspect of LPC is the linear predictive filter which allows the value of the next sample to be determined by a linear combination of previous samples. Linear predictive coding may reduce bit rate significantly and at this reduced rate the speech has a distinctive synthetic sound and there is a noticeable loss of quality. However, the speech is still audible and it can still be easily understood. Since there is information loss in linear predictive coding, it is a lossy form of compression. LPC does not represent the vocal tract characteristics from the glottal dynamics and also is takes more time and computational cost to create the model of each speaker. Linear Predictive Coding is an analysis/synthesis technique to lossy speech compression that attempts to model the human production of sound instead of transmitting an estimate of the sound wave.

3.2 Mel Frequency Cepstral Coefficients (MFCC)

Mel frequency cepstral coefficients (MFCC) is probably the best known and most widely used for both speech and speaker recognition. A Mel is a unit of measure based on human ear's perceived frequency[4]. MFCC"s are based on the known variation of the human ear's critical bandwidths with frequency; filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. It is a technique based on hearing behavior that cannot recognize frequencies over 1KHz. The signal is expresses in the MEL scale, which is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. Psychophysical studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale[2]. MFCC features are based on the known variation of the human ear's critical bandwidths with frequency.

3.3 Teaser-Kaiser Energy Operator (TKEO)

The Teaser-Kaiser Energy Operator (TKEO) is a powerful nonlinear operator proposed by Kaiser, capable of extracting the signal energy based on mechanical and physical considerations. It has been successfully used in various speech applications. This operator can be used to detect frequency and/or amplitude variations in a signal. The output of TKEO can represent their spectral content of the signals having frequency less than sampling frequency. Since the frequency variation in the compressed signal is less than the one in the original signal, the problem of cross-terms is reduced by using TKEO[5]. The TKEO measure is more effective than the traditional energy measure in detecting important parts of signal in a very noisy environment.
For a band limited digital signal, this operator can be approximated by:

$$[x\,(n)\,] = x\,2\,n - x\,n - 1\,(n + 1) \quad (1)$$

3.4 Discrete Wavelet Transform (DWT)

DWT is a powerful mathematical tool in many areas of science and engineering especially in the field of speech and image compression which uses multi resolution filters banks for the signal analysis[6]. A wavelet is a basic idea of the wavelet transform is to represent an arbitrary signal „S" as a super position of a set of such wavelets or basis functions. These basis functions are obtained from a signal prototype wavelet called the mother wavelet by dilation and translation.

3.5 Irregular Discrete Wavelet Packet Transform (DWPT)

DWT decomposes the data in a dyadic form, and the recursive decomposition only act on the low frequency content that is "approximation". The high frequency content is preserved as "detail". Approximation is abbreviated as A and Detail is abbreviated as D.

3.6 Enhanced Spectrogram: Pyknogram

The enhanced spectrogram, called Pyknogram, were first introduced to facilitate formant tracking and are calculated by applying multiband demodulation in the framework of the AM-FM modulation model[5]. Overlaps in speech data can be detected by using Pyknogram. In Pyknogram, the resonances (formants) and harmonic structure of speech are enhanced by decomposing the spectral sub-bands into amplitude and frequency components[7]. The frequency and amplitude components of a given subband, x(n), are calculated using equation of Teaser Kaiser Energy Operator (TKEO) as follows:

$$f = \frac{1}{2\pi}\arccos\left(1 - \left(\frac{\psi|\,x(n) - x(n-1)|}{2\psi|\,x(n)|}\right)\right) \quad (2)$$

$$|\,a\,| = \sqrt{\frac{\psi|\,x(n)|}{sin^2(2\pi f)}} \quad (3)$$

The weighted average of the instantaneous frequency components are used to derive a short-time estimate value for the dominant frequency in each subband over a fixed period of time, in this case the duration of a time-frame (typically 12 msec) and is revealed in eq. (4).

$$F_w(t) = \frac{\sum_t^{n+T} f(n)a^2(n)}{\sum_t^{n+T} a^2(n)} \quad (4)$$

where f(n) and a(n) are the instantaneous frequency and amplitude functions calculated for each sample in the t^{th} frame over the frame length (T samples per frame).

IV. Feature Matching

It means finding corresponding features from two similar datasets based on a search distance. Many distance measure algorithm were proposed in past for speaker identification. Mostly used are BIC, the generalized likelihood ratio, cross likelihood ratio, and KL distance.

4.1 Bayesian Information Criterion (BIC)

Bayesian Information Criterion (BIC) is one of the most popular technique for detecting speaker change point in an audio recording presented in[5].The input stream is a Gaussian process in the cepstral space. We present a maximum likelihood approach to detect turns of a Gaussian process the decision of a turn is based on the Bayesian Information criterion (BIC), a model selection criterion in the statistics literature[8]. The problem of the model identification is to choose one among a set of speaker models to describe a given set data set. We often have speaker of a series of models with different number of parameters. It is evident that when the number of parameters in the models is increased. The likelihood of the training data is also increased however when the number of parameters is too large, this might cause the problem of overtraining. Several criteria for model selection have been introduced in the statistics literature, ranging from non-parametric methods such as cross-validation to parametric methods such as the Bayesian Information Criterion (BIC). BIC is a likelihood-based criterion penalized by the model complexity.

4.2 Kullback –Leibler Distance Metric (KL2)

The KL distance between two segments A and B is an information theoretic measure equivalent to the additional bit rate accrued by encoding segment B with a code that was designed for optimal coding A. the distance is calculated between the probability density function (pdf) of two segments. The larger this value, the greater the distance between the pdf of two Random Variables[9].

The KL divergence of two random distributions is given as:

$$KL\ (AB) = E_A(\log P_A \backslash P_B) \quad (5)$$

E_A is the expected value with respect to the pdf of A. But in above case the distribution is not symmetric.
By symmetrising the KL to obtain the close form solution, KL2 divergence is obtained as:

$$KL2\ (AB) = KL\ (AB) + KL\ (BA) \quad (6)$$

The value of KL2 gives the distance between segments therefore we have to take the segments which are Gaussian in nature these type of segments help in capturing the speaker features easily. Larger the KL2 value larger the distance between segments. But performance of Kl2 is not as efficient as BIC.

4.3 T Test Distance Metrics

T test is applied to measure the similarity between two speaker models. The measure is evaluated by comparing with other distance metrics. The criterion deduces the number of speakers automatically by maximizing the separation between intraspeaker distances and interspeaker distances. It requires no development data and works well with various distance metrics. All of the distance measures are operating in the score space with the underlying assumptions that if λ_1 and λ_2 are models of the same speaker then the likelihood score value $L(X/\ \lambda_1)$ would be close to the likelihood score value$(X/\ \lambda_2)$ where $x= \{X_1,X_2 \ldots X_N)$, the observed features vectors.

V. Performance Evaluation

To measure the performance of algorithm Detection Error Trade-off (DET) & Receiver Operating Characteristics (ROC) curve is plotted. DET is tradeoff between false positive rate and false negative rate. It gives the miss error rate and false alarm rate. ROC is the curve between false positive rate and true positive rate. These curves evaluate the error rate and performance of the system.

VI. Conclusions

The paper is an overview of speaker recognition system that includes different methods of features extraction and feature matching. The focus is based on LPC, MFCC, TKEO, DWT, DPWT and Pyknogram for feature extraction. For feature matching algorithms BIC, KL2, T-Test are observed. The performance of this system can be evaluated by ROC and DET. Its applications have also been brought into light. It has been concluded that Pyknogram provides better results than other algorithms even in noisy environment. Speaker recognition system still has some drawbacks that can be reduced by carrying out research in those sub-domains**.**

References

[1] Bansal, P. 2017 .Review : Speaker Recognition Using Automated Systems. AGU International Journal of Engineering, 5.

[2] Sharma.V. and. Bansal, P. K.2013.A review on speaker recognition approaches and challenges. Int. J. Eng. Res. Technol., 2(5): 1581–1588.

[3] Chauhan, T. Soni H., and Zafar, S. 2013. A Review of Automatic Speaker Recognition System. Int. J. Soft Comput. Eng. ,3(4):132–135.

[4] Nijhawan G. and Soni, M. K. 2014. Speaker Recognition Using MFCC and Vector Quantisation. Int. J. Recent Trends Eng. Technol, 11(1) : 211–218,

[5] Kaur, S. 2017.Optimized Speaker Diarization System using Discrete Wavelet Transform and Pyknogram.International Journal on future revolution in computer science and engineering, 3(9) : 52–58.

[6] Wang, S. S. Lin, P.. Tsao, Y. Hung, J. W and Su, B.2018. Suppression by Selecting Wavelets for Feature Compression in Distributed Speech Recognition.IEEE/ACM Trans. Audio Speech Lang. Process, 26(3):564–579.

[7] Shokouhi, N. Ziaei, A. Sangwan, A. and Hansen, J. H. L.2015. Robust Overlapped Speech Detection And Its Application In Word-Count Estimation For Prof-Life-Log Data. Center for Robust Speech Systems (CRSS) The University of Texas at Dallas , 978: 4724–4728.

[8] Da Wu, J. and. Lin, B. F. 2009. Speaker identification using discrete wavelet packet transform technique with irregular decomposition. Expert Syst. Appl, 36(2): 3136–3143.

[9] Siegler, M. A. Jain, U. Raj, B. and Stern, R. M. 1997. Automatic Segmentation, Classification and Clustering of Broadcast News Audio. Proceeding. DARPA Speech Recognition Workshop, 4–6.