Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Voice Recognition Technique: A Review

Nisha
M. Tech Scholar, CSE Dept. DCRUST Murthal, India

Abstract: ***Voice Recognition is a biometric technology which is used to recognize a particular individual voice. The speech waves of particular voice form the basis of identification of speaker. We can use voice identification in multiple application areas such as telephone banking, shopping through telephone, access to database information and voice mail. One of the powerful applications of voice recognition is for security purpose where a person can enter his/her voice for authentication. Each type of voice has its unique characteristics called feature & the process of extracting these features from the individual voice is called feature extraction. The voice features which are extracted are compared with already saved voices in the database for matching. This paper provides review of various voice and speaker recognition systems.***
Keywords: Speaker Verification, Speaker Identification, Open and Close set

I. INTRODUCTION

Voice recognition or speaker identification is the process of identifying voice through its unique properties called acoustic properties [1]. The speaker voice characteristics such as expression in spoken words, emotion of the speaker and slowness or loudness [10] are represented by acoustic wave speech signal as shown in figure.1 below.

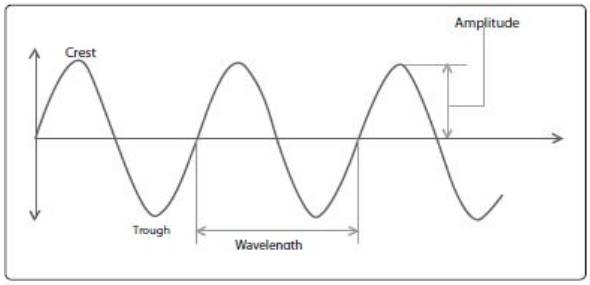


Figure 1: Sound wave of speech [10]

The speech or voice of any human being is his/her unique personal characteristics. No two have almost identical voice there are some features which must be present in one voice and missing from another voice. In order to identify this uniqueness a robust and efficient technique is required so that we can accurately identify the genuine voice from the bunch of fake voices.

The voice recognition system is broadly categorized in two ways

A. Speaker Identification

B. Speaker Verification

The process of identifying a voice of a given speech from the group of given speakers is called speaker identification. The speaker whose maximum voice characteristics are matches with the stored voice is identified & the speaker whose voice characteristics are not matched is eligible for new entry in the database (Figure 2).

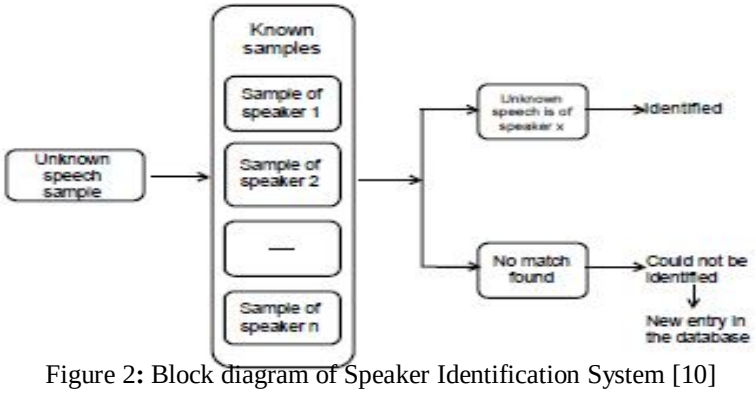


Figure 2**:** Block diagram of Speaker Identification System [10]

Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

The known sets of voices are cataloged into two parameters called Open set mode and the Close set mode. With Open set mode the speaker need not to be part of some known speakers. This is used in case of some criminal act where out of multiple suspects identify of main criminal is identified. In close set parameter mode the speaker is part of some known voices already available in database. This method is used for authentication purpose also called biometric security to identify the authorized person out of multiple claimed persons.

On the other hand, Speaker Verification is the process of accepting or rejecting the identity claim of a speaker. It is used for the verification of the person claiming for authentication.[3]

The Speaker or Voice Verification process is shown in figure 3 below.

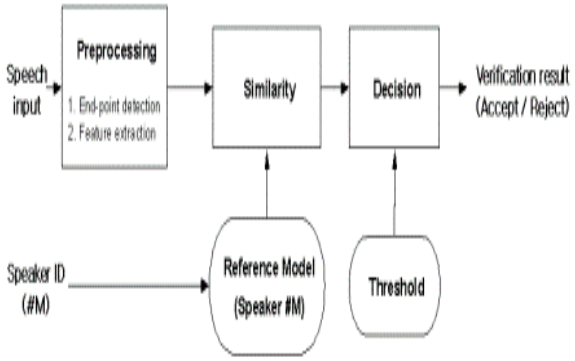


Figure 3: Block diagram of Speaker Verification System [4]

The Speaker Verification technique is generally referred as open set mode as it requires checking the authentication of voice from the set of speakers. It is most important part of voice recognition system software to verify the identity of any speaker who insists for authorization for some service.

One more classification of Voice Recognition Scheme is Text dependent and Text independent Voice Recognition. This classification is anchored in the text which is spoken by the speaker. If the text spoken by the speaker is similar to the text stored during training phase is called Text dependent Voice Recognition System. On the other hand if any random text spoken by speaker for voice identification is called Text independent voice recognition scheme/system.

Hence there are three ways of classification for Voice Recognition System are – Text dependent and Text independent, Open set and Close set and Voice Identification and Voice Verification.

II. APPLICATIONS OF VOICE RECOGNITION

There are several applications of speaker recognition; the applications are in

A. Forensic Department

Forensics is an important claim for voice recognition. If a voice sample of a criminal was traced during the obligation of crime, the suspect's voice can be compared with this, in order to give a sign of similarity of two voices.

B. Telephony and other Domain

The use of telephony in Automatic Speaker Recognition is very common in today's life whereas in the field of computer gaming and simulation it is becoming more widespread. It is also used in voice mail, e-banking and in various voice control systems.

C. Access Control

Originally various physical facilities like token, finger print, password etc. are used to access the data. But due to the advancement of technologies it is also used in voice recognition system. Voice Recognition system provide access control to various services like automation in cars, homes and mobile phones by voice command, e-banking and telephone shopping etc.

D. Transaction Authentication

Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Transaction using telephone requires higher level of authenticity to provide access control to various accounts. It is widely used in user verification for e-commerce and m- commerce.

III. LITERATURE SURVEY

Various types of voice and speaker recognition techniques are available. In this section, we provide the literature review of work done in this field.

Anusuya M. A. et. al. (2009) [8] in their paper presented a review of various types of speech recognition systems. They discussed various enhancement performed in speech recognition from the beginning till date. They provided discussion about different methods with their pros and cons.

Fook C.Y et. al. (2012) [9] The main aspire of this paper is to evaluate and encapsulate the recognized speech recognition techniques which is used by numerous researchers.

Mathur S. et. al. (2013) [10] in their paper presented the introductory information regarding speaker recognition. They also presented the various application areas where such systems are applicable.

They also developed a parameter with the help of which accurate results can be obtained from speaker recognition system.

Nandyala S. P. et. al. (2014) [11] in their paper described a new loom of hybrid HMM/DTW. In this approach, kernel adaptive filters are used for speech analysis and for speaker identification. They described that filtration of noise improve that accuracy of voice recognition and identification. Their approach gave superior experimental outcome as compare to conventional results and proved better than traditional approaches.

Chaudhary P. J. et. al. (2015) [4] in their paper described about Speaker Recognition as an ordinary process whereas Speaker Identification and Speaker Verification refer to definite tasks which are associated with this process. Here, Speaker Recognition is the computing task of authenticating the uniqueness of a claimed person by means of features extracted from the database of various voices. For the areas in which security is a foremost concern, Speaker Recognition method is one of the most practical and admired biometric recognition techniques. Various techniques for feature extraction like MFCC, RCC, LPC, LPCC, and PLPC are discussed in their paper.

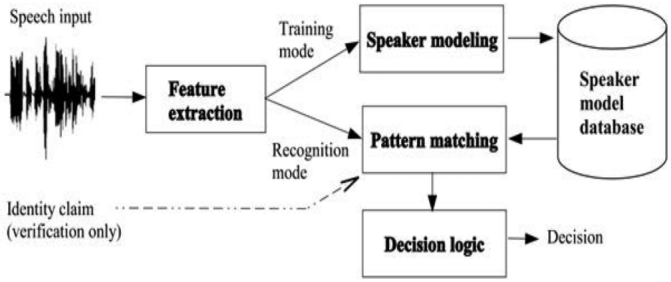
Karpagavalli S et. al. (2016) [13] described in their paper that voice is the most natural mode of communication for individual persons. The job of Speech Recognition is to renovate speech into a string of words by a computer program. Speech recognition technique helps people to use voice or speech an alternative mode of input for interaction with the machines such as computer or mobile devices. It helps the illiterate/semi-literate people to use the technology with ease. They provided detailed study on multiple Automatic Speech Recognition.

IV. AUTOMATIC SPEECH RECOGNITION SYSTEM

A. The Automatic Speech Recognition (ASR) system [2]

Composed of two approaches/modes i.e. Training mode and Testing mode. First, the input speech signal is pre-processed and then features are extracted. From the extracted features, reference samples are created from which the comparison and recognition is made. The block diagram of Automatic Speech Recognition (ASR) system is shown in fig.4.

1) *Training Mode:* In Speaker dependent and Speaker independent system, the system has to be trained by the speaker. It means that samples have to be collected from different speakers using microphone as an input device.
2) *Testing Mode:* The test sample is analyzed for its acoustic/audible characteristics and the important features are extracted from the input speech sample. These feature vectors are used to generate an input pattern and stores it in form of matrix. This unknown pattern is compared against known reference pattern, element by element. Once the best match is found, the appropriate action or decision is enabled.[2]



Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Figure 4: Block Diagram of Automatic Speech Recognition System [2]

V. FEATURE EXTRACTION TECHNIQUES

Feature extraction is essential to the recognition of the speaker as the input to the speaker model and pattern matching process. Two admired steps, framing and windowing, are predecessor to the feature extraction methods. Framing is the process where the speech signal is made static by separating it into overlapping fixed duration segments called frames whereas Windowing is the process in which each frame is multiplied by a window function which smoothes the effect of using a finite segment. In any speaker recognition system, it is necessary to extract features from each frame which can confine the speaker-specific characteristics.[14]

A. Linear Predictive Coding (LPC)

It is a dominant, vigorous, perfect, consistent and trendy tool for voice recognition. It is linear combination of previous samples. The main purpose of LPC is frame-based examination of the input speech signal to produce experimental vectors. Each sample in LPC can be estimated as precedent samples in linear combination. To implement LPC and to produce the features, the input speech signals requires to surpass through pre-emphasizer. The production of pre-emphasizer performs as the input to frame blocking where the signal is blocked into frames of N samples. In the next step windowing is done where each frame is windowed in such a manner to reduce signal disruption at the starting and end of each frame. After this step each windowed frame is auto correlated and the maximum autocorrelation value provides the order of LPC analysis and finally the resultant are LPC coefficients.[6][15]

B. Perceptual Linear Prediction (PLP)

PLP model was given by Hermansky**.** It models the human speech which is based on the perception of psychophysics of hearing**.** It rejects inappropriate information of the speech and thus modifies the speech recognition rate. It is similar to LPC excluding its spectral features have been changed to match features of human auditory system. The PLP speech analysis method is more modified to human hearing, in comparison to the traditional Linear Prediction Coding (LPC).[16]

C. Relative Spectral Filtering (RASTA)

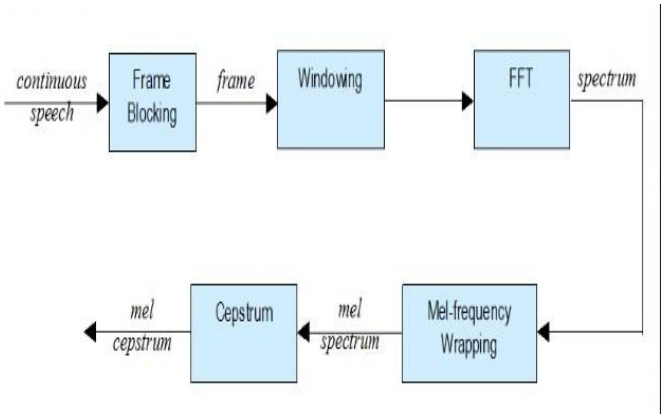
This method was initially introduced with the purpose of dropping the unnecessary noise in Automatic speech recognition. The RASTA technique not only improves the effect of noise in speech signal but it also improves the worth of speech with surroundings noise. It also includes linear filtering of trajectory of power spectrum in the case of noisy speech. RASTA filter band passes each feature coefficient. This method can be improved by combining it with PLP for better performance.[6]

D. Mel Frequency Cepstral Coefficient (MFCC)

Mel Frequency Cepstral Coefficient has a huge achievement in speaker recognition system. The MFCC is best acknowledged and most extensively used for both speech and speaker recognition. When the frequency bands are placed logarithmically in MFCC, it estimates the human system response more carefully than any other system. The method of processing MFCC is based on the short-term analysis, and thus from each frame a MFCC vector is computed. In order to obtain the coefficients, the speech samples is taken as the input and hamming window is applied to reduce the disruption of a signal. Then Discrete Fourier Transform (DFT) will be used to produce the Mel filter bank. MFCC can be calculated by using the formula [17]-

Mel (f) = 2595*log10 (1+f/700)

The following figure.5 shows the steps involved in MFCC feature extraction [17].



Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Figure 5: Block diagram of Mel frequency Cepstral Coefficient [17]

VI. COMPARISION OF VARIOUS FEATURE EXTRACTION TECHNIQUES

S.NO	TECHNIQUES	MERITS	DEMERITS
1.	Linear Predictive Coding (LPC)	Low resources required, easy implementation.	Disable to distinguish words with same vowel sounds, useful for only single speaker and single language and it is reliable for small vocabulary size.
2.	Perceptual Linear Prediction (PLP)	Discards irrelevant information of the speech and thus improves speech recognition rate.	It gives less recognition rate than MFCC and RASTA techniques.
3.	Relative Spectral Filtering (RASTA)	Useful for multi- speakers and multi-languages, and reliable for moderate size vocabulary.	It requires moderate hard implementation
4.	Mel Frequency Cepstral Coefficient (MFCC)	Useful for multi-speaker and multi-languages, reliable for moderate high size vocabulary and it is easy to implement.	Surrounding noise can influence and obstruct the quality of MFCC results.

VII. SPEAKER MODELLING APPROACHES/CLASSIFIER

After the acoustic features are extracted from the speech/voice signals, features are used to train a classifier so that it can classify the words which are spoken by the subject. Various classifiers which are used in voice recognition system are Hidden Markov Model, Neural Network Model, Dynamic Time Warping and Vector Quantization .

Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

A. Hidden Markov Model(HMM)

HMM is stochastic approach. It is easy, computationally practical and can be trained routinely, so they are very trendy model in speaker recognition. This model is characterised by a finite state Markov model and set of output distributions. It is based on huge vocabulary speech recognition systems and trained automatically on large speech data for hours. The advantage of HMM is that it decrease the complexity and time for training the huge vocabulary in recognition system [18]. The limitation of HMM is that it is often complex to examine the errors of an HMM scheme in an effort to enhance its performance.

B. Neural Network (NN) Modelling Approach

Neural Network have also been used for speaker recognition system. They are being used in solving complex identification tasks. The advantage of this approach is that they can control low quality, noisy data and are speaker independent. The disadvantage of NN approach is that optimal configuration selection is not easy to select. NN based approach used in phoneme recognition. System using NN approach based system provides better accuracy as compare to HMM which is used for limited training data and vocabulary. The NN-HMM hybrid use the NN part for phoneme identification and the HMM part for language modelling.[18]

C. Dynamic Time Warping (DTW)

DTW is an algorithm, which is used for calculating similarity between two series that may differ in time or speed. It has been useful to video, audio, graphics and any data that can be twisted into a linear representation and analysed with DTW. The optimization process is performed by using dynamic programming, hence it is named as Dynamic Time Warping. Continuity is less in DTW as compare to other approaches.

D. Vector Quantization (VQ)

VQ is a function of mapping the voice samples from a large vector space to a finite number of regions in that space. Each region is called as a Cluster and that cluster can be denoted by its center called as Centroid/Code words. The Codebook is a collection of code words. For each speaker a codebook is generated in VQ method. The Codebook then serves as prerecorded words for the speaker and it is used when speaker is tested in the system. In voice recognition system, to attain high speaker recognition rate it uses VQ as a parameter and this parameter is used for accuracy rate, processing time, number of speakers and size of training database. It is beneficial for speaker identification by using Euclidean distance for training data. The benefit of this technique is that it saves lots of time throughout the testing phase and decreases the storage and computation effort in resolving the resemblance of spectral analysis vectors.[18][19]

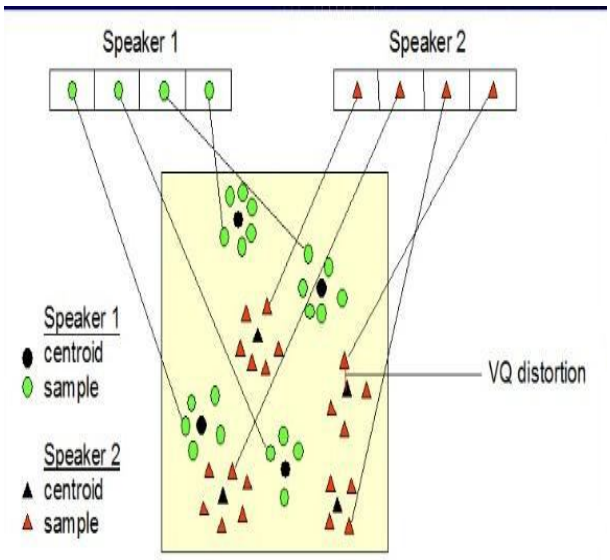


Figure 6: Conceptual diagram illustrating vector quantization code book formation. One speaker can be distinguished from another

based on the location of centroids [19]

VIII. CONCLUSION

Voice recognition is computer analysis of the human voice, particularly for the target of translating words and phrases and routinely identifying who is speaking on the foundation of individual information incorporated in speech waves. This procedure makes it feasible by using the voice of presenter and it is easy to authenticate their individuality. It provides control access to various applications such as google voice, e-commerce, window speech recognition, m-commerce, vehicle automation, home automation and security control etc. This paper provides review of various voice and speaker recognition systems.

REFERENCES

[1] Ramachandran, R. P., Farrell, K. R., Ramachandran, R., & Mammone, R. J. (2002). Speaker recognition—general classifier approaches and data fusion methods. Pattern Recognition, 35(12), 2801-2821.
[2] Karthikeyan, V., & Vijayalakshmi, V. J. (2016). PERFORMANCE COMPARISON OF SPEECH RECOGNITION FOR VOICE ENABLING APPLICATIONS-A STUDY. American Journal of Engineering and Technology Research Vol, 16(1).
[3] Petrovska-Delacrétaz, D., El Hannani, A., & Chollet, G. (2007). Text-independent speaker verification: state of the art and challenges. In Progress in nonlinear speech processing (pp. 135-169). Springer Berlin Heidelberg.
[4] Chaudhary, P. J., & Vagadia, K. M. (2015). A Review Article on Speaker Recognition with Feature Extraction. International Journal of Emerging Technology and Advanced Engineering, IJETAE, 5(2).
[5] Wang, Y., Han, K., & Wang, D. (2013). Exploring monaural features for classification-based speech segregation. IEEE Transactions on Audio, Speech, and Language Processing, 21(2), 270-279
[6] [6] Gupta, K., & Gupta, D. (2016, January). An analysis on LPC, RASTA and MFCC techniques in Automatic Speech recognition system. In Cloud System and Big Data Engineering (Confluence), 2016 6th International Conference (pp. 493-497). IEEE.
[7] Deng, L. (2014). IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) Volume 22 Issue 6.
[8] Anusuya, M. A., & Katti, S. K. (2010). Speech recognition by machine, a review. arXiv preprint arXiv:1001.2267.
[9] Fook, C. Y., Hariharan, M., Yaacob, S., & Adom, A. H. (2012, February). A review: Malay speech recognition and audio visual speech recognition. In Biomedical Engineering (ICoBE), 2012 International Conference on (pp. 479-484). IEEE
[10] Mathur, S., Choudhary, S. K., & Vyas, J. M. (2013). Speaker Recognition System and its Forensic Implications. Open Access Scientific Reports, 2(4)
[11] Nandyala, S. P., & Kumar, T. K. Hybrid HMM/DTW based Speech Recognition with Kernel Adaptive Filtering Method. Int. J. on Computational Sciences & Applications (IJCSA).–2014.–4, (1), 11-21
[12] Singh, P. P., & Singh, E. B. (2012). Speech recognition as emerging revolutionary technology. International Journal of advanced research in computer science and software engineering, 2(10), 410-3.
[13] Karpagavalli, S., & Chandra, E. (2016). A Review on Automatic Speech Recognition Architecture and Approaches. International Journal of Signal Processing, Image Processing and Pattern Recognition, 9(4), 393-404.
[14] Sapijaszko, G. I., & Mikhael, W. B. (2012, August). An overview of recent window based feature extraction algorithms for speaker recognition. In Circuits and Systems (MWSCAS), 2012 IEEE 55th International Midwest Symposium on (pp. 880-883). IEEE.
[15] Jung, Y. G., Han, M. S., & Lee, S. J. (2007). Development of an optimized feature extraction algorithm for throat signal analysis. ETRI journal, 29(3), 292-299.
[16] Dave, N. (2013). Feature extraction methods LPC, PLP and MFCC in speech recognition. International journal for advance research in engineering and technology, 1(6), 1-4.
[17] Kumar, J., Prabhakar, O. P., & Sahu, N. K. (2014). Comparative Analysis of Different Feature Extraction and Classifier Techniques for Speaker Identification Systems: A Review. International Journal of Innovative Research in Computer and Communication Engineering, 2(1), 2760-2269.
[18] Singh, A., Panchal, T., & Saharan, M. (2013). Review on Automatic Speaker Recognition System. International Journal of Advanced Research in Computer Science and Software Engineering, 3(2), 350-354.
[19] Sunitha, C., & Chandra, E. (2015). Speaker Recognition using MFCC and Improved Weighted Vector Quantization Algorithm. International Journal of Engineering and Technology(IJET), 7(5), 1685-1692.