A Review on User Identification Using Voice as a Biometric Feature

Sumit Srivastava
Dept. of CSE
BIT Mesra Ranchi

G. Sahoo
Dept. of CSE
BIT Mesra Ranchi

NamanLadha
Dept. of CSE
BIT Mesra Ranchi

Mahesh Chandra
Dept. of ECE
BIT Mesra Ranchi

ABSTRACT
In this paper, we provide a concise overview for the user identification using his biometric featurespeech. Voice processing has multiple fields of research and is widely used in many applications. Speaker recognition to identify user is a complex process in which various techniques (feature extraction, feature matching, and identification) is used to match varied characteristics of voice between training and testing data to identify the user. This paper aims to discuss efficient method to implement the identification of user on basis of their biometric feature- speech.

Keywords
LPC, MFCC, LFCC, PLP, VQ, GMM, HMM.

1. INTRODUCTION
Traditional security methods such as passwords are becoming obsolete. They can be hacked, they slow down the security process and comes with a fear that a user can forget their password at very crucial moment. Nowadays, biometric security is ubiquitous. We are familiar with fingerprint, retina scan but there is one more important biometric component, the voice. The voice of every human being is unique and consists of hundreds of different characteristics. Many organizations now use speech recognition to identify users. Unlike passwords, voice cannot be hacked and the system is designed in such a way that recorded voice does not grant success. Thus, user identification using voice recognition provides a secure, fast and efficient way to grant the authentic user access to all his important data.

Homo sapiens are evolved to communicate in various diversified languages with different characteristics. Humans propensity towards curiosity ranging from technology to the mechanics of human speech drives us to research more about the nature of human voice and how it can be recognized using machine learning. Speaker recognition[2,3] is biometric identification, derivative of general speech processing. It deals which the concept of natural language processing and neural networks. Biometric security is used in many aspects of our life (phone, automated identification, interacting with AI,

crime investigation) and which adumbrated the advent of voice security systems. Every individual has different physiological characteristics [3] of their voice (e.g. pitch) which can be distinguished with one another using cepstral coefficients [4,5,6]. Linguistic, articulatory, semantic, and acoustic are some different characteristics which tells us the emotional state of a speaker. Usually speaker recognition deals with two separate concepts: speaker identification and speaker verification. User identification refers to the process of discerning the speaker's identity by matching the extracted characteristics from the input voice against the database. Hence, it is 1: Nmatching process.

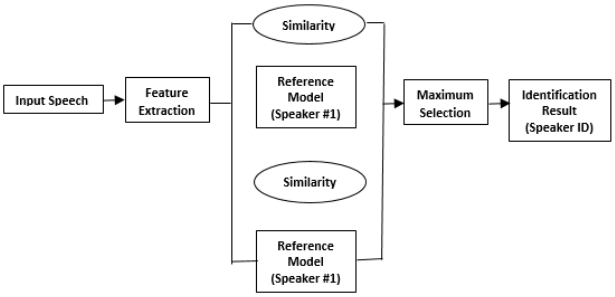

Fig 1: Speaker Identification

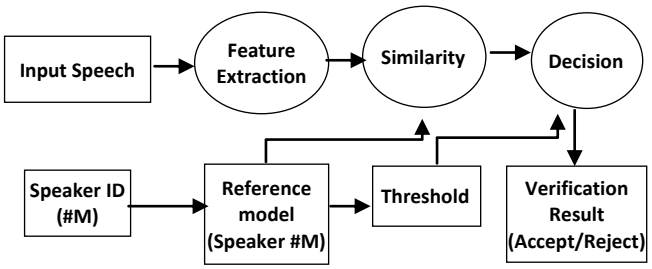

Fig 2: Speaker Verification

User verification refers to finding out whether a user's claim of his identity is true or not which makes it a 1:1 matching process.

There are two types of voice recognitions [7,8,9]: text-dependent and text-independent. In former case, the voice signals are matched against a particular sequence of texts whereas in text independent [10,11], the user can speak any words he shall like. User identification contains two components: feature extraction and feature classification [13,14]. The research embodies the constant effort to increase the efficiency of these algorithms.

1.1. Automatic Speaker Recognition Systems(ASR)
In an ASR system, voice is represented as an electrical signal which is a result of different vocal tract compression. Unique vocal tracts generate unique and the signal is stationary for each sound for about 10-20ms. During which we get the most basic sound called phoneme which are later combined to form complex words or sentences which helps us to distinguish between different speakers.

An ASR system have four stages- pre-processing, feature extraction, classification and language model, shown in Fig. 3. The pre-processing stage optimizes the input speech signal for extracting the desired features from it. Then the feature extraction stage extracts features defined previously from the optimized signal. They must be robust to any external noise hence efficiency of this stage is very important. A language model is required to match against the syntax and semantics of the languages. Finally, in classification stage, the gathered features and language model helps us to recognize the input speech.

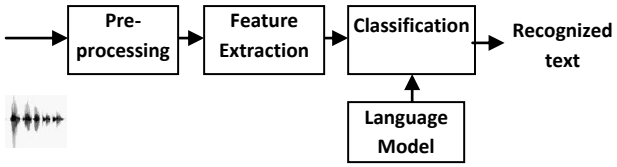

Fig 3: An ASR system

2. FEATURE EXTRACTION

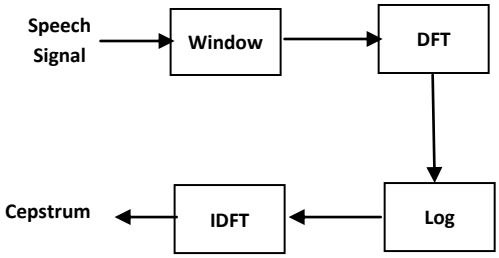

Fig 4: LPC Feature Extraction

Voice signal contains a variety of features but some which can be productively used for user identification should

- have big between speaker and little within speaker variability
- high robustness in opposition to noise signals and distortions
- have high frequency and be natural to the speech
- be easily measured from speech signal
- could not be easily mimicked
- speaker's health or variations due to age should have no effect

The most ideal features selected for speaker recognition are sub-divided as

- Dynamic features
- Suprasegmental features
- Source features
- Spectral features
- High-level features

2.1 Linear Predictive Coding (LPC)

$$\hat{s}[n] = \sum_{k=1}^{P} a_k s(n-k)$$

This technique works on the time domain of the signal where it mimics the resonance of the voice tract when we speak. It approximates every particular sample by calculatingsum of P occurred samples, given as

To procure this, frames from the input speech signal are generated and then they are windowed to remove the discontinuity that the frames could have retained on both ends. Lastly, autocorrelation among frames is estimated proceeding Durbin's method is used to perform LPC analysis on the derived autocorrelation coefficients.

2.2 Cepstral Analysis

This method is considered a good technique for modelling spectral energy admeasurement as it works in domains where glottal frequency and vocal tract resonance are kept separate. It extracts low order and high order coefficients: low order coefficients contain data about vocal tract whereas high order coefficients have information about voice excitation. Log of power spectrum being subjected to inverse Fourier transform results in a cepstrum. Both order coefficients are estimated over narrow frames in time. The initial N derived coefficients only are used for feature matching.

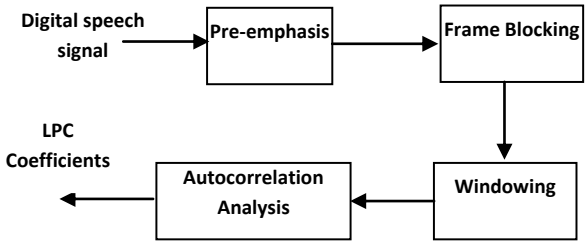

Fig 5: Cepstral Analysis

2.3 Mel-frequency Cepstral Coefficients

MFCC is widely used in software for feature extraction in the area of speech processing. It exactly mimics how a human ear works.Because of human perception behavior, which does not follow linear scale that is above 1000 Hz, we take log scale above 1000Hz and call it as Mel Scale. The MFCC algorithm works exactly like this. MFCC was designed to extract coefficients that makes most sense to humans, like pitch etc. The formula used to convert actual frequency to Mel is:

$$Mel\ (f) = 2595 * \log_{10}(1 + f/700) \quad (2)$$

At the end, we get the desired Mel-frequency cepstral coefficients.

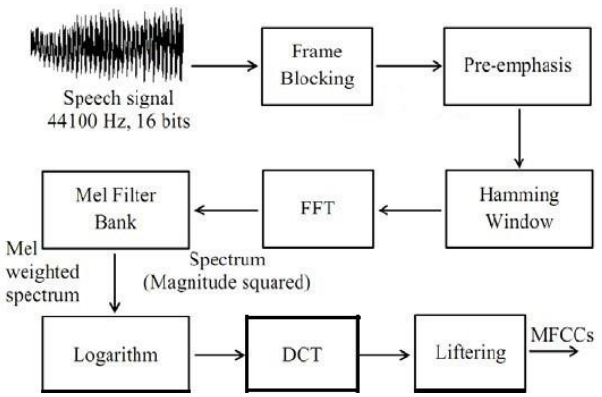

Fig 6: MFCC Feature Extraction

2.4 Linear Frequency Cepstral Coefficients (LFCC)

Linear frequency cepstral coefficients(LFCC)are same as MFCC in almost every aspect that of the filter. The Mel-frequency filter is replaced with Linear-frequency filter. The filter covers the frequency between 133-6857 Hz.

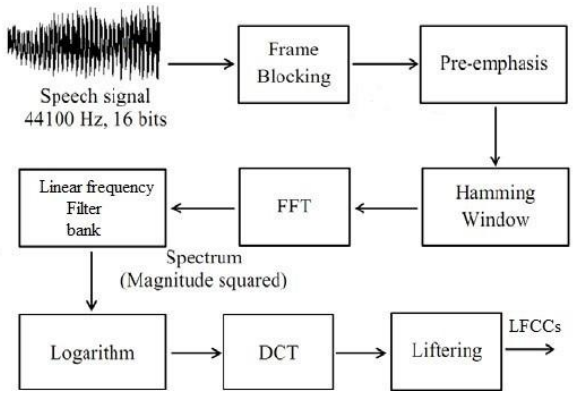

Fig 7: LFCC Feature Extraction

2.5 Perceptual Linear Predictive (PLP)

It consists of tri-essential elements: spectral resolution of crucial band, application of intensity-loudness power law, lastly equal loudness curve adjustment. PLP obtains its

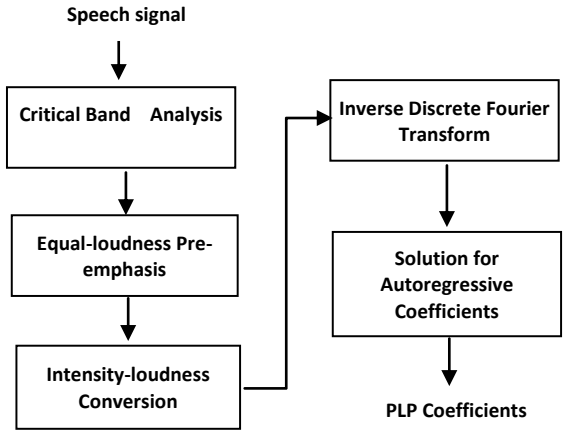

Fig 8: PLP Feature Extraction

coefficient by performing FFT on speech frame that is already windowed ahead of Bark-scale filtering. It reduces the disparity between voiced versus unvoiced speech. A disadvantage of PLP is that the resultant vectors are heavily dependent on whole spectrum of former amplitudes.

Table 1: Different Feature Extraction Techniques

S.No.	Method	Property
1.	LPC	It is a easy method which is mathematically precise.
2	Cepstral Analysis	It is a solid for extrapolating the basic frequencies of the signal but shows some aberrations on high frequency.
3	MFCC	It extracts less features from the voice signal so it is fast but have low robustness to noise.
4	LFCC	The linear filters consist of same bandwidth, homogenization becomes useless.
5	PLP	It works best in low dimensional analysis.

3. CLASSIFICATION TECHNIQUES

Classifiers approximate different feature coefficients into one to be matched against the stored data for user identification. Automatic speaker classification consists of two processes-training process and testing process. In training phase, the approximate vector from all the coefficients of the speaker is stored in the database to be matched against that same speaker in the future. In testing phase when an unknown speaker voice signal goes into the machine and a final vector is obtained and is matched against all the stored values in the database and the one that matches the testing signal vector comes out to be the identity of the user. Some of the classifiers are:

3.1 Hidden Markov models (HMM)

HMM is used in great degree in classification stage due to its ability to model the time distribution of voice signals. It is very efficient as it is very flexible and both the training and testing phases are simple to implement. It works on the likelihood that a dialect utterance was produced by how the user pronounce a specific phoneme.

There exist different states in the model and transition from one state to another takes place according to the above-mentioned probability. The probability is calculated from the training database. We cannot directly see whether a given word or sentence is spoken or not, that's why we have to rely on phoneme and hence, it is called hidden Markova model. The state can go back to itself also, so if you have three states then you have a total of nine transitions. The highest efficiency of this model is known to be 69.51%.

It is assumed that the system will go into a particular state by relying on its previous state hence it disregards long term reliance between the two states.

3.2 Vector Quantization (VQ)

VQ also called "block quantization" or "pattern matching quantization" is very efficient in data reduction (compression) so it is very often used in ASR systems. VQ forms small clusters of vectors from big data by forming groups of vector closest to them. It then calculates the average of the clusters and assign them a value. When a testing data is fed, the distance between average of that testing data and training data is calculated and the minimum distant set results into the user identity.

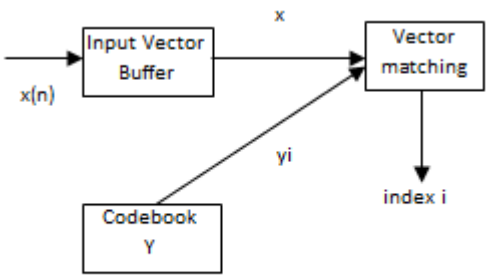

Fig 9: VQ technique

3.3 Gaussian Mixture Model (GMM)

It is used to generate parts of spectral information from short time frames of speech. It tells us about speaker's vocal physiological characteristics and it text-independent because it does not rely on phonetic content. It is frequently interfused with SVM classifier and thus problem of high dimensional super-vectors is tackled by using GMM super-vectors.

4. CONCLUSION

This paper considered different techniques that can be implemented in speaker recognition systems to identify user using biometric feature. Advantages and disadvantages of different methods are discussed to make appropriate choice while building such a system. Methods used for feature extraction include LPC, Cepstral coefficient, MFCC, LFCC, PLP among which MFCC is found to give maximum efficiency among them. For the next stage of classification, VQ, HMM, GMM is discussed and HMM is found to be most useful. Further research is required to make these algorithms more efficient and create a high quality automatic speaker recognition system.

5. REFERENCES

[1] Dr. Mahesh S. Chavan,Mrs. Sharada V. Chougule, "Speaker Features And Recognition Techniques: A Review",International Journal Of Computational Engineering Research / ISSN: 2250–3005

[2] Michelle Cutajar, Edward Gatt, Ivan Grech, Owen Casha, Joseph Micallef, University of Malta, Tal-Qroqq, Msida, MSD 2080, Malta," Comparative study of automatic speech recognition Techniques",ietdl,2012

[3] Parvati J.Chaudhary1, Kinjal M. Vagadia ,"A Review Article on Speaker Recognition with Feature Extraction",Ijetae, ISSN 2250-2459, ISO 9001:2008 Certified Journal,, Volume 5, Issue 2, February 2015

[4] J.P. Campbell, Jr., "Speaker Recognition: A Tutorial," Proceedings of IEEE, vol. 85, no. 9 , Sept. 1997

[5] Varsha Singh, Vinay Kumar Jain, Dr. Neeta Tripathi," A Comparative Study on Feature Extraction Techniques for Language Identification",International Journal of Engineering Research and General Science Volume 2, Issue 3, April-May 2014

[6] Alahari.Neelima, Gattupalli.Deepti," A Review on Speech Analysis and Automatic Speaker Recognition",International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 2, February 2015

[7] Sandeep Joshi, Parag Parandkar," A Review of Feature Extraction Technique for Automatic Speech Recognition",International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064,2012

[8] B.S. Atal, "Automatic recognition of speakers from their voices, " Proc. IEEE , vol. 64(4), pp. 460-75, Apr. 1976

[9] G. Doddington, "Speaker recognition -identifying people by their voices, " Proc. IEEE , vol. 73, pp. 1651-64, 1985

[10] D.A. Reynolds, ―Experimental Evaluation of Features for RobustSpeaker Identification‖, IEEE Trans. on Speech and Audio Processing, vol. 2, issue-4, pp. 639-643, October, 1994

[11] Ghahramani Z., ―An Introduction to Hidden Markov Models andBayesian Networks‖, International Journal of Pattern Recognition and Artificial Intelligence, vol. 5, issue-1, pp. 9-42, 2001

[12] Markowitz J (2007) The many roles of speaker classification in speaker verification and identification. Springer Berlin Heidelberg 4343: 218-225

[13] Sumit Srivastava, Mahesh Chandra, G Sahoo, Phase Based Mel Frequency Cepstral Coefficients for Speaker Identification",DOI: 10.1007/978-81-322-2757-1_31 In book: Information Systems Design and Intelligent Applications, pp.309-316

[14] Sumit Srivastava, Pratibha Nandi, G. Sahoo, Mahesh Chandra, "Formant Based Linear Prediction Coefficients for Speaker Identification", 2014 International Conference on Signal Processing and Integrated Networks (SPIN), 978-1-4799-2866-8/14/$31.00 ©2014 IEEE.