Efficient Speaker Verification Algorithm using Spectral Characteristics

Vyshali V Nayak

***Abstract*—Speaker recognition is a recognition purpose that articulates words. The speaker recognition process relies on physical structure of an individual's person's vocal tract and the behavioral characteristics of the individual. Speaker verification is evolved with the technologies of speech recognition and speech synthesis because of the similar characteristics in the voice and challenges associated with it. Speaker recognition has two forms which is text dependent or text independent. In text dependent method a particular phrase or password is stored into the system, whereas in text independent method the speaker will not be aware that his voice is being collected. In the proposed algorithm, speech signal has been recorded in the database. And the speaker is verified using the input the speaker provides by comparing with the database. The time domain, frequency domain and power domain features of the speech is extracted. For validating the performance, a comparative analysis has been carried out with various other methods. These methods exhibit some unique behavior.**

***Index Terms*—Spectral Characteristics, Speech Recognition, Text Dependent, Text Independent**

I. INTRODUCTION

Speech contains the information of behaviour feature that is embedded in the signal which can be used for speaker recognition. The performance of a speaker recognition system mainly depends on the technique employed in the various stages of speaker recognition system such as Mel frequency Spectral coefficients (MFFCs), Gaussian mixture model (GMM) and feature extraction. The basic fundamental of speaker identification and verification system is feature extraction of speech signal. Speaker identification is a process of identifying the speech of different speakers. As shown in Figure 1 it has been observed that speaker identification process identifies speaker through the behavioural feature of speech.

Speaker verification system determines automatically whether a given segment of speech is actually spoken by the speaker who claims to be a certain person. It is further divided into text independent speaker verification (TISV) and text dependent speaker verification (TDSV). Speaker verification then verifies the speech of the claimed speaker as shown Figure 2.

Manuscript published on 30 April 2016.
*****Correspondence Author(s)

Vyshali V Nayak, B.E in E n C, Signal Processing, Cambridge Institute of Technology, VTU, Bangalore, Karnataka, India.

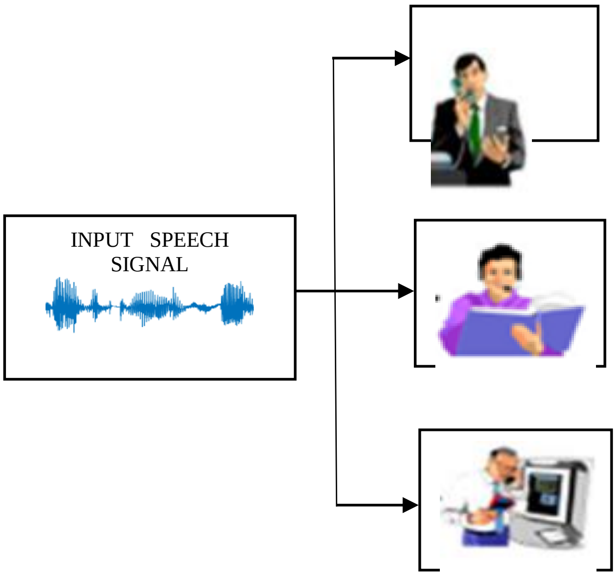

Fig. 1 Speaker Identification Model

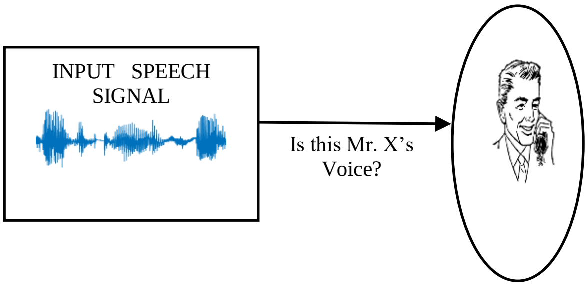

Figure 2 Speaker Verification of an Individual

The applications of speaker recognition technology uses the speaker's voice for verification. After verification it enables the controlled access to services like voice dialling and voice mail, tele-banking, telephone shopping, access to database related services, information services, security control for confidential information areas, forensic applications, and remote access to computers. Efficient speaker recognition algorithm can improve the technology of speaker verification that will make our daily lives more convenient.

Thesis contribution is to propose an efficient model for speaker verification using the Spectral Characteristics of the input speech. For developing efficient model literature survey on existing techniques has been carried out. The Modelling of proposed algorithm is executed/implemented using Mat lab software. Finally, the algorithm proposed is validated with various test inputs. Proposed technique is providing a future scope for speaker verification challenges.

II. LITERATURE SURVEY

Speaker usually contains two main features. The two features are the low level features and the other one is high level feature. Low level features refer to or are associated with the perception of the speech by the brain. The feature that is hard to track is segmental. Especially the feature formants. Pitch periodicity which is the supra segmental feature is comparatively easy to extract. But the main issue is that it requires a voiced or either unvoiced detector. Detailed individual differences of these long term averages of the low level features may be measured. It may be used if it does not resolve any individual differences in detail. Central locations are usually spoken by the high level features. The perception of the words and its individual meaning and syntax as well as its prosody. It also depends on the dialect and the variety of language that is unique to a person. The features are harder relatively to extract compared to the low level features.

A Adaptation

In majority of the speaker recognition cases, the speech data that is available for the enrolment process is narrowly limited to train certain models. In the system where the password is fixed for certain speaker authentication systems, the data that is enrolled or the enrolment data could be recorded in a single cell. As a result, it ends up in the enrolment and also the test conditions may be severely mismatched. The different telephone handsets and the networks like landline or the cell network due the disturbance in the background noises. In text independent models the additional problems may be resulted from mismatches in the content of linguistics. Due to these reasons, adaptation techniques can be used to build models for a specific set of targets. This error can be reduced significantly by using a fixed password in the system. [5]

Low level features are usually of the short time spectra which are generally MFCCs. A speech recognition system must speaker independent. Also the speaker recognition system should speech independent. This actually suggests that the optimal acoustic features would be definitely different. What a Cepstral mean actually does is that is subtraction. It subtracts the average of the cepstral over a sufficiently long speech recording. Hence it removes the distortions of the convolution, which is slowly varying in the channels. The derivatives Δ and also the second derivatives Δ^2 of the dynamic information of the above mentioned features are also helpful. It is helpful for both the speech and for speaker recognition. A robust pitch extraction is hard because the pitch has the larger variation in the intra speaker. [5]

Speaker recognition models can be divided mainly into two classes. One is non-parametric model. The non-parametric model makes a few structural assumptions about the input data. It is effective when there is sufficient amount of enrolment data which is to be matched to the test data. These non-parametric models are based on techniques like nearest neighbor models and template matching. Parametric models on the other hand offers a representation of structural constraints or parsimonious representation. Which means it can make a very effective use of data if the constraints are chosen properly [5].

B. Models

Models are usually based on techniques like Vector quantization, Hidden Markov models (HMM), Gaussian mixture models (GMM), Support vector machines and something based on these models or newer models which are in research. [6] GMMs can be considered as a generalization of k means. Here the individual cluster can be allowed to have its own covariance matrix. The parameters of the model like mean, mixing coefficients and covariance are usually learned with the algorithm. GMM's are usually suitable for text independent speaker recognition. It does not model the temporal aspects of the input speech.

HMM's are shown to be much more effective for text dependent systems. HMMs may be trained through the phone, at the word or sentence level, depending on the password vocabulary such as digit sequence which are most commonly used. HMMs are used by training generally through maximum likelihood. Discriminative training techniques can be used if the examples from the competing speakers are readily available. Ergodic HMMs may be used for text independent systems. Ergodic HMMs usually allows all the transitions that are possible between the states unlike the left right HMMs which is generally used in ASR. Using this way the emission probabilities, there is a tendency to represent different spectral characteristics. At the same time transition probabilities allows some modelling of the temporal information. However, the experimental comparisons of the GMMs and the ergodic HMMs end up showing an additional issue of the transition probabilities. HMM's has very little effect on the performance. [5]

Now we take up the paper "Speaker verification based on the fusion of the speech acoustics and inverted articulatory signals" for the comparative study of that particular model with our proposal. We find a few other important things from the paper. Here the author has proposed the speaker verification system by using a single exemplar speaker only. It is shown in Figure 3 and 4. In this paper both Text Independent Speaker Verification (TISV) and Text Dependant Speaker Verification (TDSV) has been implemented. In this method a score level and feature level fusion is carried out. It combines both the acoustic and articulatory information for both the text dependent and the text independent speaker verification. From real life based point of view, the study on how to improve the performance of the speaker verification by combining the information characterizing of the articulatory trajectories in the speech production.

The author has gathered the information that for the text independent speaker verification concatenation of the articulatory features that are obtained from the measured amount of speech production data with the help of conventional MFCCs helps in improving the performance of the model dramatically. In the above mentioned paper the author has explored both the score level as well as the feature level fusion methods. It is also observed that the overall performance of the system is further enhanced by a significant amount.

The third paper that we have used for the survey is the Comparative Evaluation of Feature Normalization Techniques for Speaker Verification. In this paper a simple method called feature normalization method which is called STMSN was involved. Its performance can be in the content of an i- vector speaker verification system. This particular method was compared to STG and STMVN methods. Bothe of the methods STMVN and STMSN methods can provide comparable speaker verification results. It uses i-vectors compared to that of STG. The STG method is considerably more complex and it takes a much longer time to normalize the features of MFCC vectors when compared to STMSN and STMVN [7].

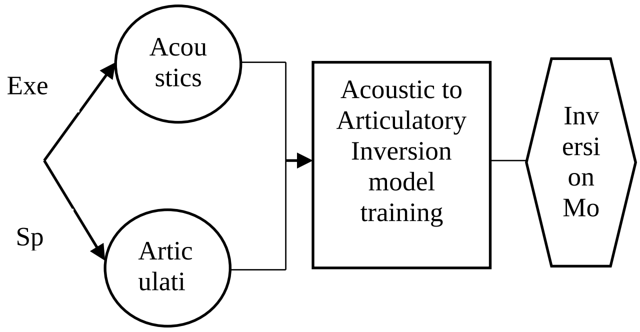

Fig. 3. System overview of the proposed speaker verification system with the help of a single exemplar System

By now we have realized the importance of the speech verification system. The Speech is most noticeable and essential method of Communication among of person. The correspondence among human PC cooperation is called human PC interface. Discourse has capability of being essential method of connection with the PC. A Review on the Speech Recognition Technique, paper gives an outline of major innovative point of view and energy about the key advancement of discourse acknowledgment furthermore gives diagram system created in every phase of discourse acknowledgment.

This paper helps in picking the speaker verification system alongside their relative benefits and its negatives. A similar investigation of various strategy is done according to the procedure that is to be followed. This paper closes with the choice on highlight course to develop procedure in human PC interface framework utilizing a regional Language. It likewise introduces the rundown of system with their properties for Feature extraction as show in Table. Through this survey it is found that MFCC is utilized broadly for highlight extraction of discourse and GHM and HMM is best among all demonstrating procedure [8]. As per the proposed block diagram flow chart has been developed. First voice has been recorded from acquisition system. After that feature extraction of recorded voice has been carried out same feature extraction is carried out for recorded database. Spectral behaviour of voice has been identified. After identification both are compared to verify the voice. If both the voice is matched, then speaker is verified else speaker is not verified. The flowchart of proposed algorithm is shown in Figure 5.

Signals are represented in different forms, one is time domain and the other one is frequency domain. In these domains different features of signals are reflected. In existing speaker recognition system, we mainly use segmental analysis such as Mel frequency Spectral coefficients (MFFCs), Gaussian mixture model (GMM) and feature extraction. The speech feature extraction could have a drawback with its categorization. Its main concern would be to reduce the dimensionality of the input vector whereas at the same time to maintain the discriminating power of the speech signal. The elementary formation of speaker identification and verification system should be to extract that particularly mentioned feature. Fundamentally that should be because of the training and because the number of training and test vector needed for the categorization problem increases with the dimension of the given input.

C. Applications of Speech Recognition

1. In car while a person is driving doing simple things like answering a call becomes difficult. Hence voice commands can be used to make a phone call or select a particular radio station or to play music from our smart phone which is compatible with the car system. Music might even be played with the MP3 player or a music loaded with the flash drive. The voice recognition capability might vary from the model of the automotive and therefore the automotive structure and facilities.
2. Health care is another domain where speech recognition are often enforced within the backend or the frontend of the documentation method within the medical system. Here the front end system refers to where the provider dictates into a speech recognition engine. The words which are recognized are displayed as they are spoken. The person who speaks is responsible for editing if the system doesn't recognize any words. And once edited should also sign off from the document.

   Back end systems or deferred speech recognition is when the provider/ speaker dictates into a digital dictation system. The voice then gets routed through a speech recognition machine. The words which are recognized is draft documented and is routed with the original recorded voice. It is rerouted with the original voice file to the editor. The editor in turn edits the report and a final document is finalized. Deferred speech recognition is a widely used toll in the medical industry today.
3. Speech recognition is also used widely in military applications. In the last few decades' speech recognition has been used in the evaluation and the test of speech recognition in fighter aircraft. Speech recognizers are operated quite success in these airplanes. The application such as setting up the radio frequency, functioning an autopilot system, setting the coordinate points for steering, release parameters for weapon in case of emergency and mainly controlling the flight display. In aviation particularly in the helicopters there is a complication in achieving high recognition preciseness owing to stress and noise. It also depends strongly on the environment around the helicopter. Especially during the jet fighter surrounding. The noise is high not just severe because of the environment, but there is also high level of noise which comes from the helicopter pilot since he doesn't wear a mask.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The acoustic noise is high due to the combination of both. The facemask would help in reducing the acoustic noise through the microphone.

4. Air Traffic Controllers (ATC) additionally represents a wonderful application for the speech recognition systems. ATC training systems needs an individual to act as a pseudo-pilot. The pseudo-pilot engages in a voice dialogue with the new trainee controller. They simulate the dialogue that the controller has to conduct with the pilots in a real time ATC scenario. Speech synthesis and speech recognition offers to eliminate the requirement of an individual who acts as the pseudo pilot. Thus it helps in reducing the coaching and support personnel.
5. Speech recognition is an excellent tool for learning a new second language. It helps us in learning the proper pronunciation. It also helps a person develop his/her fluency with the speaking skills. It helps in correcting the grammatical errors too.
6. Speech recognition is an excellent tool to the students who are visually impaired. The students who are blind or have a very low vision can benefit a lot from speech recognition technology. It helps them to convey words to the system and hear the system recite the same words. They can control the computer by using their voice as commands. They don't have to use the keyboard or the screen and strain their vision.

It is not just limited to the visually impaired students. It could also be used by physically disabled or the people who have had a strain or injury to the extreme. It could also be used by the people who are temporarily disable due to an accident. They need not worry about their handwriting or even typing. They can work on their school assignments by using speech to text programs. They could use speech recognition technology to search on the internet without the necessity of using a computer. They need not even use the mouse or the keyboard at home physically. All they have to do is say the words out loud. This can lead to an increase the productivity as the speed is enhanced. They also do not have to concern themselves regarding the punctuation or spelling or any other mechanics of grammar or writing.

Also students with learning disability can use speech recognition. Voice recognition software can be used along with the digital audio recorder or a personal computer. Software's like Microsoft word has proven to be helpful for individuals who have damaged short term memory especially in stroke and craniotomy (Brain defects).

Other people with disabilities can benefit from speech recognition. Individuals who cannot hear or are deaf or people who are hard of hearing with clarity. Here's where speech recognition comes into play. The software is used to automatically generate a closed captioning of the conversations or discussions in classroom lectures, conference/ business meetings, and religious get to gatherers.

Speech recognition can be used by people who find it difficult to use their hands. IT may range from repetitive stress injury (RSI) or people who have disability who cannot use conventional input devices of a computer system. In fact, people who used keyboard on a regular basis and developed RSI over a period of time there was an early urgent development for speech recognition.

Speech recognition is also used by people who have hearing impairment. Applications like voicemail to text conversion and captioned telephony. Another application for people with learning disability can use the thought to paper communication. For example, they think of an idea but it ends up different from their thought on the paper. They can benefit from the software but it comes with lot of bugs. And there is a lot of research required for this to grow up.

Also it has certain defects in the idea that the speech can be converted into text. Especially for the intellectually disabled because there are very few people who learns a technology to teach a person with disability to use it. Also people with dyslexia can find the product more difficult rather than useful. If the kid is not able to pronounce the word clearly the output maybe different from the input. Hence more work would be falling on the disabled person who is using the software.

D. Summary of the Literature Survey

Signals are represented in different forms, one is time domain and another on is frequency domain. In these domains different features of signals are reflected. In existing speaker recognition system mainly used segmental analysis like Mel frequency Spectral coefficients (MFFCs), Gaussian mixture model (GMM) and feature extraction.

The speech feature extraction in actually a categorization problem is about reducing the dimensionality of the input vector while maintaining the discriminating power of the signal. Fundamental formation of speaker identification and verification system needs feature extraction of speech signal because the number of training and test vector needed for the classification problem grows with the dimension of the given input.

III. PROPOSED ALGORITHM

The performance of a particular speaker recognition system depends on the technique employed in the various stages of speaker recognition system. In proposed algorithm the possibility of speaker verification is increased using spectral behavior of the speaker. Functional block diagram of the proposed algorithm is shown in Figure 5.

In this block diagram speech signal has been recorded using speech acquisition system. For speaker identification databases are identified. Feature extraction of both input speech and database speech has been carried out. After that comparison of both behaviors input as well as have been carried out for speaker verification. When the behavior got matched speaker is verified.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

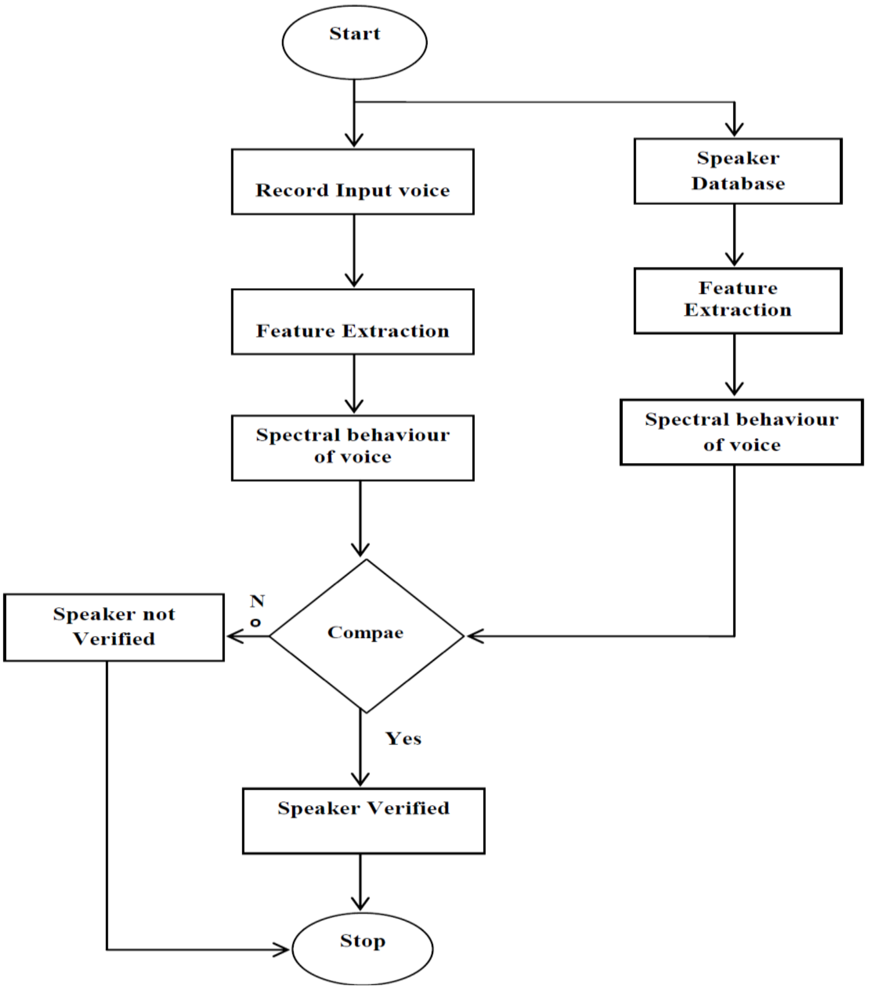

Fig. 4. System overview of the proposed speaker verification system with the help of a single exemplar speaker.

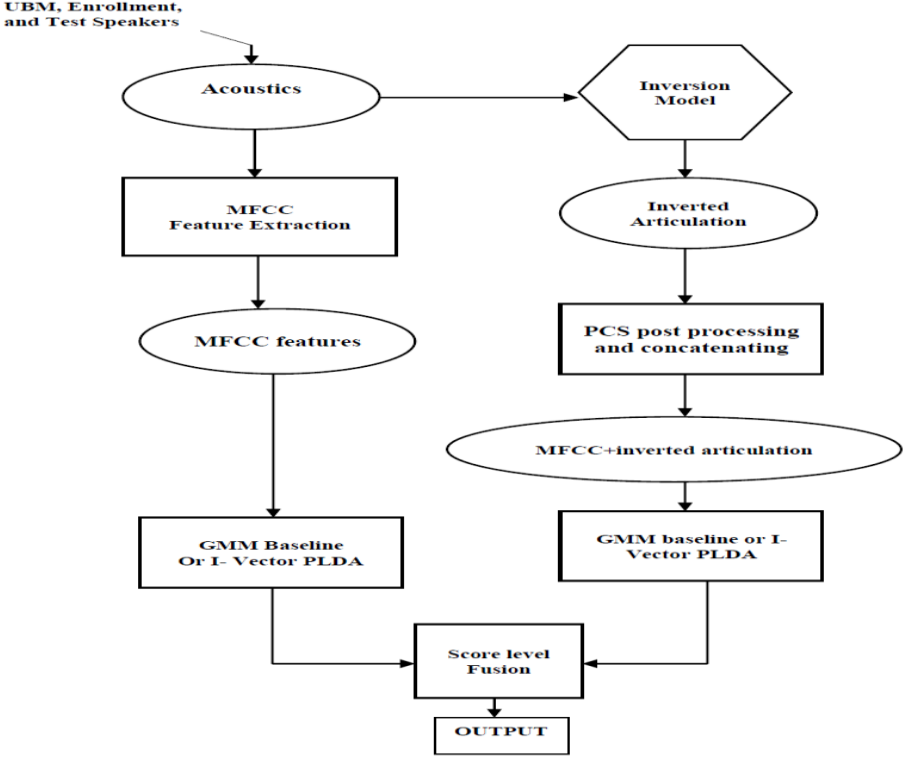
UBM, Enrollment, and Test Speakers
Acoustics
Inversion Model
MFCC Feature Extraction
Inverted Articulation
MFCC features
PCS post processing and concatenating
MFCC+inverted articulation
GMM Baseline Or I- Vector PLDA
GMM baseline or I- Vector PLDA
Score level Fusion
OUTPUT

Fig 5: Functional Block diagram of the proposed algorithm

In proposed algorithm the possibility of speaker verification is increased using spectral behaviour of the speaker. Functional block diagram of the proposed algorithm is as shown in Figure 5.2. In this block diagram speech signal has been recorded using speech acquisition system. For speaker identification databases are identified. Feature extraction of both input speech and database speech has been carried out. After that comparison of both behaviours have been carried out for speaker verification. When the behaviour got matched speaker is verified.

1) 5.1. Mathematical Model of Proposed Algorithm

Speech Signals are mathematically represented in time domain as well as frequency domain. In time domain it consists of speech samples in discrete times. In frequency domain it consists of frequency components of speech signal. Fast Fourier transform is used to convert time domain to frequency domain and inverse Fast Fourier transform (FFT) is used for frequency to time domain conversion. In Figure 5.3 the transformation has been shown.

DFT (FFT):

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\left(\frac{2\pi}{N}\right)nk} \qquad (k= 0, 1, \ldots\ldots N-1)$$

IDFT (IFFT):

$$x(n) = \frac{1}{N}\sum_{k=0}^{N-1} X(k) e^{j\left(\frac{2\pi}{N}\right)nk} \qquad (n=0, 1, \ldots. N-1)$$

Frequency domain representation of speech signal consists of frequency components that carry the behaviour of speech signal. That behaviour of each speech signal is unique for individuals which can help us to identify or verify the speaker.

IV. Result

As similar the power spectrum of a speaker contains unique behavior of that speaker. For understanding the unique behavior of speaker two speech signals has been taken and power spectrum of two different speeches have been plotted in Figure 7.

In text independent recognition system, it does not know text spoken by person, which could be user-selected phrases or conversational speech. It is unsuitable for security applications. It is suited for identification of uncooperative speakers. It is flexible system but it has more difficult problem. In proposed algorithm power spectrum based feature extraction technique is adopted because power spectrum of speech contains unique behaviour of speech. For differentiating different speaker's speech power gain varies, therefore in speaker verification algorithm power spectrum based feature extraction is giving better performance rather than time domain features or frequency domain features.

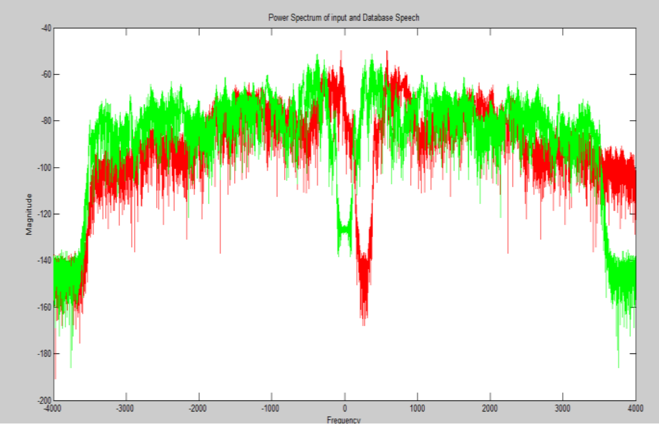

Fig 7 : Comparison of the power spectrum of two speakers

In speaker verification, user makes a claim as to his/her identity, and the goal is to determine the authenticity of the claim. In this particular project we use the voice samples which are compared only with the speaker model of the so called claimed identity. Speaker is generally assumed to be cooperative because only one comparison is made. Performance is independent of the size of the speaker population. In text dependent recognition system knows the text spoken by the person who knows the either a phrase like his name or a password. These systems assume that the speaker is cooperative and suited for security applications. To keep impostors from playing back a secret key that could be recorded from approved speakers, arbitrary provoked expressions can be utilized.

V. Conclusion

In speaker verification, the user makes a claim on to his/her identity, and also the goal is to work out the legitimacy of the claim. During this explicit project, we tend to use the voice samples that are measured and compared solely with the speaker model to that of the claimed identity. The speaker is usually assumed to be reciprocal. As a result, just one comparison is made. because only one comparison is made. Performance is totally independent of the dimensions of the speaker population. In text dependent recognition system, the user already knows what input has to be given in. It can either be a phrase or a password that is constantly used on a regular basis or can be changed depending on the application. These systems work under the principle that the speaker is reciprocal and hence it may be used as a security measure. In order to avoid thieves or an unrecognized user, any arbitrary phrase can be utilized by the same security application. The imposters might have a recording of the speaker and use it for acquiring the information.

In the case of a text- independent recognition system, it doesn't recognize text spoken by the person that may well be any word uttered by the user. It is generally unsuitable for security applications. It can be used in applications where the user would not be reciprocal. It is a much more difficult approach compared to the text dependent one. It's a versatile system with a tougher drawback.

*Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.*

In proposed algorithm power spectrum based feature, extraction technique is adopted because power spectrum of speech contains unique behavior of speech. For differentiating different speaker’s speech power gain varies, therefore in speaker verification algorithm power spectrum based feature extraction is giving better performance rather than time domain features or frequency domain features.

ACKNOWLEDGMENT

I take this opportunity to convey my gratitude to all those who have been kind enough to offer their advice and provide assistance when needed, which has led to the successful completion of this project. I would like to thank **Dr. L Suresh, Principal**, Cambridge Institute of Technology, for giving permission to carry out project with the facilities provided to us in our college campus.I express my immense gratitude to **Dr. G. Indumathi**, Professor and Head, Department of Electronics & Communication Engineering (M. Tech), Cambridge Institute of Technology, for providing right ambience for carrying out my project work.I express sincere thanks to my internal guide **Mr. Girish H**, **Assistant Prof., Dept. of ECE** Department of Electronics & Communication Engineering, Cambridge Institute of Technology, for his valuable suggestions, guidance, advice, encouragement and reviews. I take this opportunity to thank all the teaching and non-teaching staff of Cambridge Institute of Technology, for all kind of help, suggestion and support in the successful completion of this project. I would like to thank my Parents, my baby sister, my in laws and my husband for all their support and encouragement. Last but not the least I thank almighty for having given me encouragement and faith with which I could finish my thesis.

REFERENCES

1. Douglas A. Reynolds and Larry P.Heck, “Automatic Speaker Recognition: Recent Progress, Current Applications and Future Trends”, 19 February 2000, http://www.ll.mit.edu/IST/pubs/aaas00-dar-pres.pdf
2. Joseph P. Campbell, “Speaker Recognition”, Identification in Networked Society, 1999
3. Samudravijaya K, “Speech and Speaker Recognition: A Tutorial”, 2001
4. Bojan Imperl, “Speaker recognition techniques”, Maribor, Slovenia, 2000
5. Rosenberg, “L16: Speaker recognition”, Benesty, 2008
6. W. M. Campbell, D. E. Sturim, and D. A. Reynolds, “Support Vector Machines Using GMM Supervectors for Speaker Verification”, IEEE Signal Processing Letters, vol. 13, no. 5, may 2006
7. Md Jahangir Alam, Pierre Ouellet, Patrick Kenny, Douglas O’Shaughnessy, “Comparative Evaluation of Feature Normalization Techniques for Speaker Verification”, Springer, 2011
8. Santosh K.Gaikwad, Bharti W.Gawali, Pravin Yannawar, “A Review on Speech Recognition Technique”, International Journal of Computer Applications (0975 – 8887), Volume 10– No.3, November 2010
9. Zhang Wanli, Li Guoxin, “Application of Improved Spectral Subtraction Algorithm for Speech Emotion Recognition”, IEEE Fifth International Conference, 2015
10. Luciana Ferrer, Yun Lei, Mitchell McLaren, and Nicolas Scheffer, “Study of Senone-Based Deep Neural Network Approaches for Spoken Language Recognition” IEEE/ACM Transactions, 2015
11. S. K. Singh, “Features and Techniques for Speaker Recognition”, 2003
12. W. M. Campbell, D. E. Sturim, and D. A. Reynolds, “Support Vector Machines Using GMM Supervectors for Speaker Verification”, IEEE Signal Processing Letters, vol. 13, no. 5, may 2006