Priyanka Mishra, Suyash Agrawal / International Journal of Engineering Research and Applications (IJERA)
ISSN: 2248-9622 www.ijera.com Vol. 2, Issue 2,Mar-Apr 2012, pp.933-938

Recognition Of Voice Using Mel Cepstral Coefficient & Vector Quantization

Priyanka Mishra*, Suyash Agrawal**
** Department of Computer Science & Engineering Rungta college of Engineering & Technlogy, Kurud Bhilai, CSVT University, Chhattisgarh
** Department of Computer Science & Engineering ,Rungta college of Engineering & Technlogy, Kurud Bhilai, CSVT University, Chhattisgarh

Abstract
Human Voice is characteristic for an individual. The ability to recognize the speaker by his/her voice can be a valuable biometric tool with enormous commercial as well as academic potential. Commercially, it can be utilized for ensuring secure access to any system. Academically, it can shed light on the speech processing abilities of the brain as well as speech mechanism. In fact, this feature is being used preliminarily along with other biometrics including face and finger print recognition for commercial security products. Speaker recognition is the method of automatically identify who is speaking on the basis of individual information integrated in speech waves. There are two types of speaker recognition systems basically divided into two – classification: speaker identification and speaker verification. Speaker identification determines from which of the registered speakers a given utterance comes whereas speaker verification is the process of accepting or rejecting the claimed identity of a speaker .The fundamental difference between identification & verification modes is the number of decision alternatives. In the Identification mode the number of decision alternatives is equal to the size of the population, whereas in the verification mode there are only two alternatives, accept or reject the Identification claim, regardless of the size of population. Most applications of speaker recognition are actually speaker verifications. Speaker Recognition is of two types :Text Based and Text independent. In Text based approach, the speaker is identified by the utterance of some fixed piece of text while in the text independent approach the speaker is allowed to utter any text whatsoever.

Keywords – The Sound Wave, a Band Pass Filter of bandwidth, vector quantization techniques, Voice Recognition Algorithm, Mel Cepstral Coefficient & Vector Quantization (Mfcc).

I. INTRODUCTION
Speech is produced through the biological system of a larynx or sound box, which resides in the throat of the human beings .The larynx or the sound box is provided with some fibres which are capable to vibration when the air passes through them .The major organs that pump air through the larynx are the lungs attached to the larynx by the windpipe. The laryngeal fibres or vocal cords, as they are properly capable of vibrating at all frequencies. In this particular case, the range goes from the just audible 20 Hz up to about 11000 Hz. The higher frequencies found usually in the children and the women and the lower in the men. In terms of Signal Processing, the larynx is the source & the vocal tract the filter .The Source is capable of vibrating at many if not all frequencies .The sound thus produced is filtered by the vocal tract which includes the vocal cavity, the mouth cavity & the nasal cavity to resonate some frequencies & anti resonates others so that a special emanates from the lips in the form of speech. The shape and length of the vocal tract are not constantly changing but are stable over a scale of millisecond. Hence the characteristic of the filter are stable over the millisecond scale of the millisecond. Hence the voice of a person has a static nature on small scale and a dynamic nature on a larger scale. [1] The human speech contains numerous discriminative features that can be used to identify speakers. Speech contains significant energy from zero frequency up to around 5 kHz. The objective of automatic speaker recognition is to extract, characterize and recognize the information about speaker identity. The property of speech signal changes markedly as a function of time. To study the spectral properties of speech signal the concept of time varying Fourier representation is used. However, the temporal properties of speech signal such, as energy, zero crossing, correlation etc are assumed constant over a short period. That is its characteristics are short-time stationary .Therefore, using hamming window, Speech signal is divided into a number of blocks of short duration so that normal Fourier transform can be used. [2]

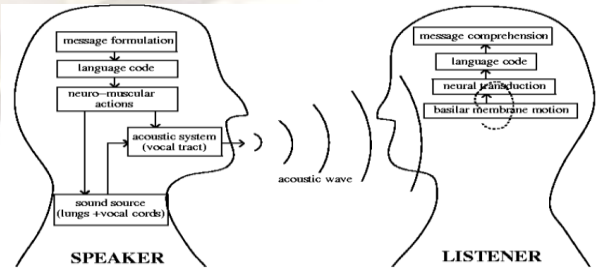


Fig : 1

II. HEADINGS

In the study, the effectiveness of combinations of cepstral features, channel compensation techniques, and different local distances in the Dynamic Time Warping (DTW) algorithm is experimentally evaluated in the text-dependent speaker identification task. The training and the testing has been done with noisy telephone speech (short phrases in Bulgarian with length of about 2 seconds) selected from the BG-SRD at corpus. The employed cepstral features are – Linear Predictive Coding derived Cepstrum (LPCC), Mel-Frequency Cepstral Coefficients (MFCC), Adaptive Component Weighted Cepstrum (ACWC), Post-Filtered Cepstrum (PFC) and Perceptually Linear Predictive coding derived Cepstrum (PLPC). Two unsupervised techniques for channel compensation are applied–Cepstral Mean Subtraction (CMS) and Relative Spectral (RASTA) technique. In the DTW algorithm two cepstral distances are utilized – the Euclidean and the Root Power Sum (RPS) distance. The experiments have shown that the best recognition rate for available noisy speech data was obtained by using the combination of the MFCC, CMS and the DTW-RPS distance. [3] At the highest level, all speaker recognition systems contain two main modules feature extraction and feature matching. Feature extraction is the process that extracts a small amount of data from the voice signal that can later be used to represent each speaker. Feature matching involves the actual procedure to identify the unknown speaker by comparing extracted features from his/her voice input with the ones from a set of known speakers. We will discuss each module in detail in later sections. Although voice authentication appears to be an easy authentication method in both how it is implemented and how it is used, there are some user influences that must be addressed, Colds. If the user has a cold which affects his or her voice that will have an effect on the acceptance of the voice-scanning device. Any major difference in the sound of the voice may cause the voice-scanning device to react in a negative way, causing the system to reject the user. Expression and volume. If a person is trying to speak with expressions on their face (i.e. smiling at the same time) their voice will sound different. The user of the device must also be able to speak loudly and clearly in order to obtain accurate results. Misspoken or misread prompted phrases. If the user is required to authenticate by speaking a prompted phrase and they mispronounce the phrase, they will be rejected by the system. Previous user activity may have an impact on the outcome of the voice scanning device. For example, if the user is out of breath and is unable to speak well. Background noises will interfere with the user who is trying to authenticate to the device. The environment in which the user is authenticating to the device must be free of any major background noise. [4] The mel-frequency cepstral coefficients (MFCCs) are frequently used as a speech parameterization in speech recognizers. Practical applications of speech recognition and dialogue systems bring sometimes a requirement to synthesize or reconstruct the speech from the saved or transmitted MFCCs. Presented paper describes an approach to the construction of a MFCC-based speech production system and discusses based speech production system and discusses various possibilities of its excitation.

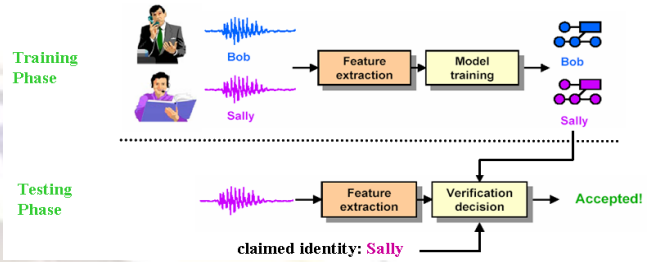


Fig : 2 Claimed identity is also fed to the system

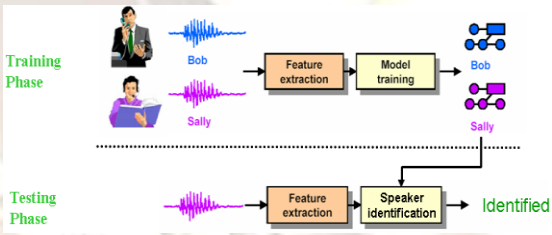


Fig : 2 Claimed identity is not known

III. INDENTATIONS AND EQUATIONS

The problem identification work depends solely on literature survey. Problem identification is a process of identifying the problem and it define the problem clearly. In the literature survey it is found that various author used various methods for speaker identification but still there is a scope for future work according to my shown method. Now in problem identification we will show the method used in literature survey and how i can improve these methods to get accurate speaker identification using the proposed work. Speaker recognition is basically divided into two-classification: speaker recognition and speaker identification and it is the method of automatically identify who is speaking on the basis of individual information integrated in speech waves. Speaker recognition is widely applicable in use of speaker's voice to verify their identity and control access to services such as banking by telephone, database access services, voice dialing telephone shopping, information services, voice mail, security control for secret information areas, and remote access to computer AT and T and TI with Sprint have started field tests and actual application of speaker recognition technology; many customers are already being used by Sprint's Voice Phone Card. Speaker recognition technology is the most potential technology to create new services that will make our everyday lives more secured. Another important application of speaker recognition technology is for forensic purposes. Speaker recognition has been seen an appealing research field for the last decades

which still yields a number of unsolved problems. The main aim of this project is speaker identification, which consists of comparing a speech signal from an unknown speaker to a database of known speaker. The system can recognize the speaker, which has been trained with a number of speakers. [11] Techniques of Feature Extraction: The general methodology of audio classification involves extracting discriminatory features from the audio data and feeding them to a pattern classifier. Different approaches and various kinds of audio features were proposed with varying success rates. Some of the audio features that have been successfully used for audio classification include Mel-frequency cepstral coefficients (MFCC), Linear predictive coding (LPC), Local discriminant bases (LDB). Few techniques generate a pattern from the features and use it for classification by the degree of correlation. Few other techniques use the numerical values of the features coupled to statistical classification method.

Equation
Coding of Algorithm used for Change Detection Filtering the Sound Wave The sound wave under consideration is filtered with a Band Pass Filter of bandwidth 80 Hz-8000 Hz.

Program code:
[b,a]= butter(4, [80/22050 8000/22050]);

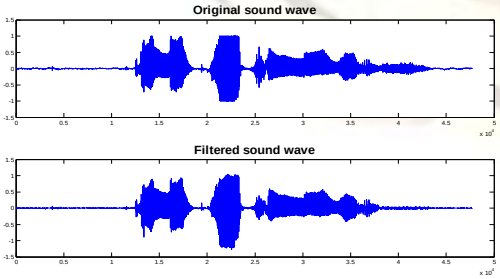
x=filter(b,a,y);

Where

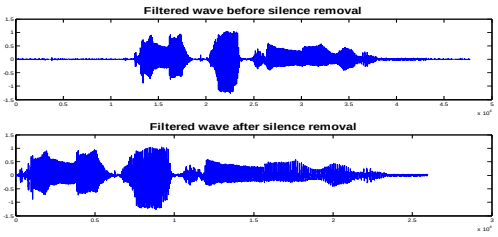
y= wavread('sample.wav')

22050=sampling frequency

Silence Removal Silence present before and after the voiced part is removed to improve the performance of classifier. FILTERING THE SIGNAL The sound wave under consideration is filtered using a Band Pass Filter of bandwidth (80hz-8000hz) to eliminate noise.



REMOVING SILENCE **Silence present in the sound wave is removed to emphasize only on the voiced part of the wave**



Feature Vectors
Now, mfcc's are calculated for the processed sound waves. Each frame of wave will have a mfcc vector and the frame rate for the present program is 100 frames/sec. First row of mfcc vectors gives the energy content of corresponding frames. Energy information is not a speaker specific feature and hence it is eliminated from the feature vector. The size becomes (12x1). Derivatives and double derivatives are found by the difference between consecutive vectors and second immediate vectors respectively.

Derivatives of mfcc's provide information regarding the rate of speech. By augmenting derivatives, the size of feature vector becomes (36x1)

IV. FIGURES AND TABLES

The Mel-frequency cepstral coefficients (MFCCs) are frequently used as a speech parameterization in speech recognizers. Practical applications of speech recognition and dialogue systems bring sometimes a requirement to synthesize or reconstruct the speech from the saved or transmitted MFCCs. [7]

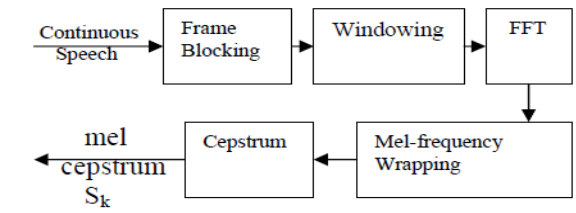


Figure 4 Block Diagram of the Mfcc Processor

The speech input is recorded at a sampling rate of 22050Hz. This sampling frequency is chosen to minimize the effects of ***aliasing*** in the analog-to-digital conversion process.
Figure 4.1 shows the block diagram of an MFCC processor. [8]. In this work, the Mel frequency Cepstrum Coefficient (MFCC) feature has been used for designing a text dependent speaker identification system. The extracted speech features (MFCC's) of a speaker are quantized to a number of centroids using vector quantization algorithm. These centroids constitute the codebook of that speaker. MFCC's are calculated in training phase and again in testing phase. Speakers uttered same words once in a training session and once in a testing session later. The Euclidean distance between the MFCC's of each speaker in training

phase to the centroids of individual speaker in testing phase is measured and the speaker is identified according to the minimum Euclidean distance. The code is developed in the MATLAB environment and performs the identification satisfactorily. [2]

IV.I SPEAKER VERIFICATION SYSTEM

A speaker verification system is composed of two distinct phases, a training phase and a test phase. Each of them can be seen as a succession of independent modules.

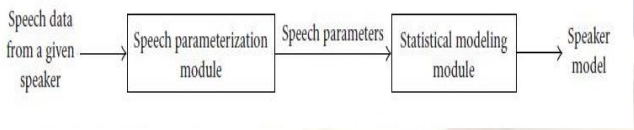


Figure 4.1 Modular Representation of the Training Phase of Speaker Verification System

Figure 4.2 shows a modular representation of the training phase of a speaker verification system. The first step consists in extracting parameters from the speech signal to obtain a representation suitable for statistical modeling as such models are extensively used in most state-of-the-art speaker verification systems. The second step consists in obtaining a statistical model from the parameters. This training scheme is also applied to the training of a background model.

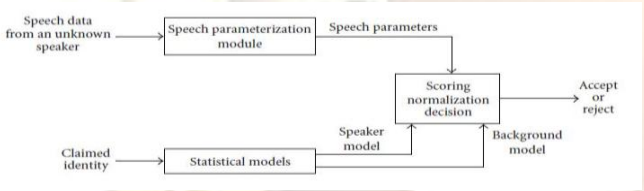


Figure 4.2Modular Representation of the Test Phase of a Speaker Verification System

Figure 4.3 shows a modular representation of the test phase of a speaker verification system. The entries of the system are a claimed identity and the speech samples pronounced by an unknown speaker. The purpose of a speaker verification system is to verify if the speech samples correspond to the claimed identity. First, speech parameters are extracted from the speech signal using exactly the same module as for the training phase. Then, the speaker model corresponding to the claimed identity and a background model are extracted from the set of statistical models calculated during the training phase. Finally, using the speech parameters extracted and the two statistical models, the last module computes some scores, normalizes them, and makes an acceptance or a rejection decision. The normalization step requires some score distributions to be estimated during the training phase or/and the test phase. [10]

Speaker recognition can also divide into two methods, text-dependent and text -independent methods. In text dependent method the speaker to say key words or sentences having the same text for both training and recognition trials. Whereas in the text independent does not rely on a specific text being speak. Formerly text dependent methods were widely in application, but later text independent is in use. Both text dependent and text independent methods share a problem however. By playing back the recorded voice of registered speakers this system can be easily deceived. There are different techniques used to cope up with such problems. Such as a small set of words or digits are used as input and each user is provoked to thorough a specified sequence of key words that is randomly selected every time the system is used. Still this method is not completely reliable. This method can be deceived with the highly developed electronics recording system that can repeat secrete key words in a request order. [11]

IV.II TEXT DEPENDENT VOICE RECOGNITION

The speech-dependent recognition techniques discriminate the users based on the same spoken utterance. Text-dependent recognition methods are usually based on template-matching techniques. Many of them use *Dynamic time warping* (*DTW*) algorithms or *Hidden Markov models* (*HMM*). Speaker classification represents the next stage of this pattern recognition process. We use a supervised classifier for our voice identification system, proposing a *minimum mean distance classification* approach. A set of registered (advised) speakers is set first. Next, a *training set* is obtained as a collection of spoken utterances, corresponding to the same speech, provided by these speakers and filtered for noise removal. Each speech signal of the training set constitutes a *vocal prototype*. The feature vectors computed for these prototypes make the *feature training set*. Then, we apply an extended version of the minimum distance classification procedure. We consider N classes, each class corresponding to an advised speaker. Our algorithm introduces each input vocal sequence i S in the class of the speaker corresponding to the smallest mean distance between the feature vector of the input signal and his prototype vectors.

IV.III TEXT-IN DEPENDENT VOICE RECOGNITION

The speech-independent recognition systems involve impressing volumes of training data ensuring that the entire vocal range is captured. Thus, it is useful for not cooperative subjects, for example like those in the surveillance systems. The most successful speech-independent recognition methods are based on Vector Quantization (VQ) or Gaussian Mixture Model (GMM). The VQ-based methods are parametric approaches which use VQ codebooks consisting of a small number of representative feature vectors, while the GMM-based methods represent non-parametric techniques using K Gaussian distributions. We

utilize the same delta mel cepstral analysis for the feature extraction part of this recognition system. [16]

IV.IV MFCC EVALUATION

The MFCC parameterization follows common requirements imposed on a speech parameterization for speech recognition purposes. Its main features are aimed above all at: 1. to capture an important information presented in a speech signal for recognition purposes .2 to handle as little data as necessary and 3 to use any quick evaluation algorithm. Moreover the benefit of MFCCs is also in their perceptually scaled frequency axis. The mel-scale offers higher frequency resolution on the lower frequencies in the same way as a sound is percepted by the human auditory organ. In addition, the MFCCs offer through their cepstral nature abilities to model both poles and zeros.

V. CONCLUSION

In this proposed work the methods used Speaker Recognition based on their results are as follows. The methods used for Speaker Recognition are classified into two parts

1) **Training Phase**
2) **Testing Phase**

This thesis project describes an enhanced Mel frequency Cepstral coefficient (MFCC) Technique for speaker recognition & analysis by the Computer .The computer records the voice pattern of the speaker during training phase. During the testing phase ,a speaker speaks into the microphone, and the computer analyses it by using Mat Lab software Two pattern matching algorithm are used in recognition mode.The algorirhm can be used in security devices that uses voice recognition technology for identification. Most important parts of speaker recognition system are (i) Feature Extraction (ii) Classification method & Feature matching , The aim of the feature extraction step is to strip unnecessary Information from the sensor data and convert the properties of the signal which are important for the pattern recognition task to a format that simplifies the distinction of the classes. The goal of the classification step is to estimate the general extension of the classes within feature space from a training set. According to Martens there are various feature extraction techniques including Linear Predictive Coding (LPC), Perceptual Linear Prediction (PLP) and MFCC, However MFCC is frequently used method for Speaker Recognition & Speaker Verification. In MFCC, the main advantage is that it uses mel frequency scaling which is Very approximate to the human auditory system. The basic steps are: Cepstrum Preprocessing and FFT. The MFCC technique has been applied for speaker identification. VQ is used to minimize the data of the extracted feature. The study reveals that as number of centroids increases, identification rate of the system increases. It has been found that combination of Mel frequency and Hamming window gives the best performance. It also suggests that in order to obtain satisfactory result the number of centroids has to be increased as the number of speakers increases The study shows that the linear scale can also have a reasonable identification rate if a comparatively higher number of centroids is used. However, the recognition rate using a linear scale would be much lower if the number of speakers

increases. Mel scale is also less vulnerable to the changes of speaker's vocal cord in course of time. The present study is still ongoing, which may include following further works HMM may be used to improve the efficiency and precision of the segmentation to deal with crosstalk, laughter and uncharacteristic sounds. A more effective normalization algorithm can be adopted on extracted parametric representations of the acoustic signal, which would improve the identification rate further. Finally, a combination of features (MFCC, LPC, LPCC, Formant etc) may be used to implement a robust parametric representation for speaker identification

REFERENCES

[1]. MFCC and Its Applications in Speaker Recognition Vibha Tiwari IJET pp 19-22, 2010.

[2] Speaker Identification Using Cepstral Analysis Muhammad Noman Nazar IEEE pp 139-143, 2002.

[3]. Cepstral Features and Text-Dependent Speaker Identification – A Comparative Study Atanas Ouzounov C& Vol 10. No 1 2010..

[4] Artificial Neural Network & Mel-Frequency Cepstrum Coefficients-Based Speaker Recognition Adjoudj Réda , Boukelif Aoued March 27-31, 2005.

[5] Voice Recognition Algorithms Using Mel Frequency Cepstral Coefficient (Mfcc) and Dynamic Time Warping (Dtw) Techniques lindasaiwa Muda,Mumtaj Begam and I. Elamvazuthi Journal of Computing Vol 2, Issue 3 pp 138-143, 2010.

[6]. A Viable Technique: Speaker Recognition Manjot Kaur Gill.

[7]. Speech Production Based on the Mel-Frequency Cepstral Coefficients Zbynìk Tychtl and Josef Psutka.

[8]. Speaker Identification Using Mel Frequency Cepstral Coefficients Md. Rashidul Hasan,Mustafa Jamil, Md. Golam Rabbani Md. Saifur Rahman ICECE pp 565-568, 2004.

[9]. Comparison of Clustering Algorithms in Speaker Identification Tomi Kinnunen, Teemu Kilpeläinen and Pasi Fräntic .

[10] A Tutorial on Text-Independent Speaker Verification Fr ´ed´ eric Bimbot, Jean- Franc¸ois Bonastre, Corinne Fredouille, And others Journal on Applied Signal Processing pp 430-451, 2004.

[11]. Speaker Recognition Project Report speaker-recognition.googlecode.com/files/Finally_version1.pdf.

[12]. Text Independent Speaker Recognition Using the Mel Frequency Cepstral Coefficients and A Neural Network Classifier Hassen Seddik AmelRahmouni and Mounir Sayadi IEEE pp 631-634, 2004.

[13]. Speaker Discriminative Weighting Method for VQ-Based Speaker Identification Tomi Kinnunen and Pasi Fränti IEEE pp 150-156, 2001.

[14]. Pitch Extraction and Fundamental Frequency: History and Current Techniques David Gerhard Technical Report TR-CS 2003-06 pp 1-22, November, 2003.

[15] Speaker Identification System Using HMM and Mel Frequency Cepstra Coefficient Dr. Yingen Xiong , Seonho Kim May 10, 2006. citeseerx.ist.psu.edu/viewdoc/download/doi=10.1.1.138...

[16.] A Toolbox For Ann Learning Poliana Magalhães Reis and Carlos Alberto Ynoguti IEEE pp 130-133, 2005.

[17] L R Rabiner & R W Schafer, "DIGITAL PROCESSING OF SPEECH SIGNALS", Low Price Edition.