ISSN: 2278 – 7798
International Journal of Science, Engineering and Technology Research (IJSETR)
Volume 1, Issue 6, December 2012

Recognition of Speaker Using Mel Frequency Cepstral Coefficient & Vector Quantization

Priyanka Mishra, Suyash Agrawal

Abstract— Voice recognition is basically divided into two-classification: Voice recognition and Voice identification and it is the method of automatically identify who is speaking on the basis of individual information integrated in speech waves. Voice 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. Another important application of Voice recognition technology is for forensic purposes.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).

A. *Index Terms*— The Sound Wave, a Band Pass Filter of bandwidth, vector quantization techniques, Voice Recognition Algorithm, Mel Cepstral Coefficient & Vector Quantization (Mfcc) .

II. Introduction

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.

Manuscript received Nov 15, 2012.
***Priyanka Mishra**, M.Tech Computer Technilogy, CSVTU/ RCET College/ Bhilai Chhattisgarh., Durg Chhattisgarh, India, 91- 9755931266.*
***Suyash Agrawal,**, Reader in CSE. Depatment,, CSVTU / RCET College / Bhilai , Durg, India,91 7828678782.,*

III. Procedure for Paper Submission

A. *Review Stage*

Voice recognition has been an interesting research field for the last decades, which still yields a number of unsolved problems. This paper aims to present a speaker recognition system which recognizes the speaker as opposed to what is being said by the speaker as in case of speech recognition. The methodology followed in this paper for Speaker recognition is using Feature Extraction process and then Vector Quantization of extracted features is done. At last the speaker is identified by comparing the data from a tested speaker to the codebook of each speaker and then measuring the difference. [14] Speech processing is emerged as one of the important application area of digital signal processing. Various fields for research in speech processing are speech recognition, speaker recognition, speech synthesis, speech coding etc.

Final Stage

studied for many papers in Voice Recognition. 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. But 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.

B. *Figures*

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.

IV. MATH

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.

V. UNITS

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

A. LPC

LPC (Linear Predictive coding) analyzes the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The process of removing the formants is called inverse filtering, and the remaining signal is called the residue. In LPC system, each sample of the signal is expressed as a linear combination of the previous samples. This equation is called a linear predictor and hence it is called as linear predictive coding .The coefficients of the difference equation (the prediction coefficients) characterize the formants.

B. MFCC

MFCC is based on the human peripheral auditory system. The human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency t measured in Hz, a subjective pitch is measured on a scale called the 'Mel Scale' .The mel frequency scale is a linear frequency spacing below 1000 Hz and logarithmic spacing above 1kHz.As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 Mels.

C. LDB

LDB is an audio feature extraction and a multi group classification scheme that focuses on identifying discriminatory time-frequency subspaces. Two dissimilarity measures are used in the process of selecting the LDB nodes and extracting features from them. The extracted features are then fed to a linear discriminant analysis based classifier for a multi-level hierarchical classification of audio signals. [2]

VI. HELPFUL HINTS

A. *Figures and Tables*

MEL FREQUENCY SPECTRAL COFFICIENTS (MFCC)

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]

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. 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] Speaker recognition is a generic term used for two related problems: speaker identification and verification. In the identification task the goal is to recognize the unknown speaker from a set of N known speakers. In verification, an identity claim (e.g., a username) is given to the recognizer and the goal is to accept or reject the given identity claim. In this work we concentrate on the identification task. The input of a speaker identification system is a sampled speech data, and the output is the index of the identified speaker. There are three important components in a speaker recognition system: the feature extraction component, the speaker models and the matching algorithm. [9]

1. Voice Input

(a) Data Set's (known)

(b) Run Time (unknown)

2. Convert Voice into .Wav Form
3. Window the signal
4. Apply Fast Fourier Transform (FFT)
5. Take the magnitude
6. Take logarithm of magnitude
7. Warp the frequencies according to the Mel scale
8. Take the inverse FFT

6.1 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 6.1 Modular Representation of the Training Phase of Speaker Verification System

Figure 6.2 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.

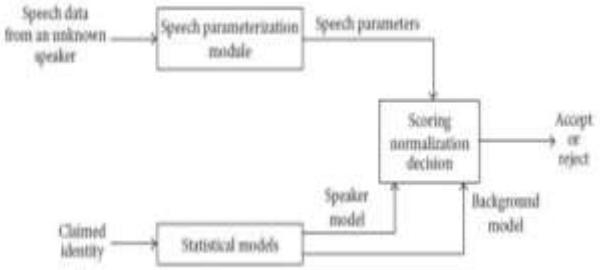

Figure 6.1 Modular Representation of the Test Phase of a Speaker Verification System

TRAINING and TESTING ANN

1 The feature vectors of many sentences from a person are considered at once and are clustered
2 This clustered data of different voice in speakers is used to train the ANN
3 Sample to be tested is processed and the feature vectors are clustered
4 Clustered test data is fed to ANN for classification
5 "One against one" classification method is used in this program
6 Classification is done for each vector from the test sample and the identities are ranked according to the classifier output
7 The one with maximum rank is the identity for given test sample

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]

6.2 APPROACHES TO SPEAKER RECOGNITION

Conventionally, there are two approaches to speaker recognition: the one employing the techniques of speech analysis like the formant analysis and pitch analysis with a good measure of artificial intelligence tools thrown in such as neural networks and HMM's etc; and the other is from the perspective of signal processing.

The Formants are actually the major frequencies, which are most pronounced .i.e. which have the highest amplified or power when a vowel is spoken. Vowels are periodic utterances that contain most of the energy in speech. For this reason formants are good tools to be used in speech analysis in finding out what has been spoken but literature indicates that these same are not too reliable in the case of speaker recognition.

The use of artificial intelligence tools tike neural networks promises to solve the problem with much less labour but does not does not shed any light on the essence of speaker recognition. The digital signal processing approach highlights the parameters, which make a speaker different from any other speaker. The significance of these parameters emerges by modeling the speech production mechanism of an individual as a digital system. The neural networks or HMM's, on the other hand, can only be used to arrive at a decision about the identity of the speaker, and not the reason or the process behind it. [1] Three speaker-modeling techniques: Dynamic Time Warping (DTW), Hidden Markov's Models (HMM) and VQ (Vector Quantization) are used in the text-dependent speaker recognition task. In the study the employed cepstral features are – LPCC, MFCC, Adaptive Component Weighted Cepstrum (ACWC), 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 [3]. Since many years, the two most common and successful approaches for speaker recognition, independently of the pronounced text, are based on modeling the speech by Gaussian Mixture Models, and Hidden Markov Models. These methods are attractive for their phonetic discrimination capacity. The acoustics analyses based on the MFCC, which represent the ear model, has proved good results in speaker recognition especially when a high number of coefficient is used. We use the MFCC extracted from the speaker phonemes as a discriminative features. The text independent speaker recognition is done by classifying these features by A multi-layer neural network.

6.3 **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). Let us consider a sequence of same-speech input vocal utterances to be recognized: {S_1 , ... , S_n }. The feature extraction process is then applied to them, the feature set {V(S_1) ,...,V(S_n) } being obtained. 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.

6.4 TEXT-INDEPENDENT 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]

reference

Human Voice recognition has been an interesting research field for the last decades, which still yields a number of unsolved problems. This paper aims to present a speaker recognition system which recognizes the speaker as opposed to what is being said by the speaker as in case of speech recognition. The methodology followed in this paper for Speaker recognition is using Feature Extraction process and then Vector Quantization of extracted features is done. At last the speaker is identified by comparing the data from a tested speaker to the codebook of each speaker and then measuring the difference. [14] Speech processing is emerged as one of the important application area of digital signal processing. Various fields for research in speech processing are speech recognition, speaker recognition, speech synthesis, speech coding etc. The objective of automatic speaker recognition is to extract, characterize and recognize the information about speaker identity. Feature extraction is the first step for speaker recognition.[1] overview of automatic speaker recognition technology, with an emphasis on text-independent recognition. Speaker recognition has been studied actively for several decades. We give an overview of both the classical and the state-of-the-art methods. We start with the fundamentals of automatic speaker recognition, concerning feature extraction and speaker modeling. We elaborate advanced computational techniques to address robustness and session variability.[2] 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.[4]

C Abbreviations and Acronyms

TESTING OF PROPOSED SYSTEM

Around fifty seconds of speech data of the intended speaker was collected for training the neural network. In testing phase, 10% tolerance is present for the intended speaker, i.e. if the output of the network is 10% less or greater than 10%, still the speaker is recognized as the intended speaker otherwise rejected.

The test data consists of fifty (50) speech samples of the speaker for whom network is trained and some samples of imposter speech. The imposter speech data was collected from other persons. Out of 50 samples of the intended speaker correct data and false rejects are having in proportionate form.

Experiment Samples:

Here we used the following samples of sentences for each speaker.

- We Engineers are best.
- I proud to be here.
- I like watching Movies.
- Wish you good luck.
- Knowledge is Power.

We work on total 05 speakers. All the data samples are collected from a database. Following performance curves show the training process of one Voice in Speaker :

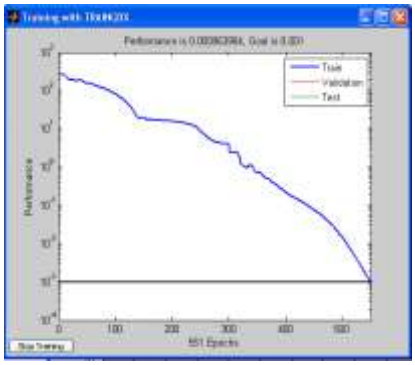

Figure 6.4: Performance Curve for speaker 1 sentence 1st

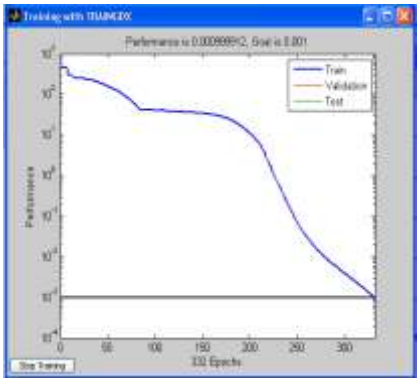

Figure 6.4: Performance Curve for speaker 1 sentence 2nd

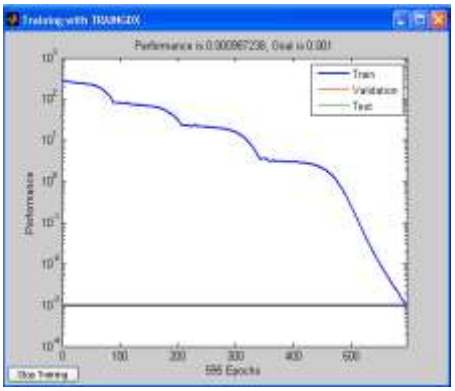

Figure 6.4: Performance Curve for speaker 1 sentence 3rd

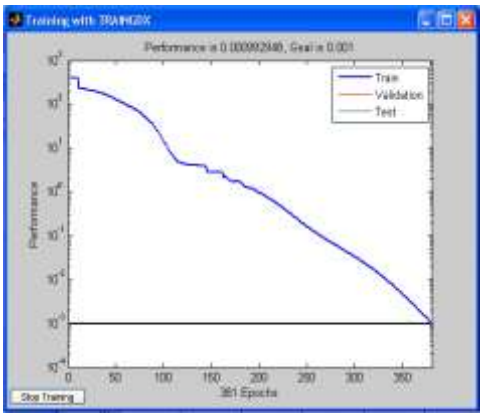

Figure 6.4: Performance Curve for speaker 1 sentence 4th

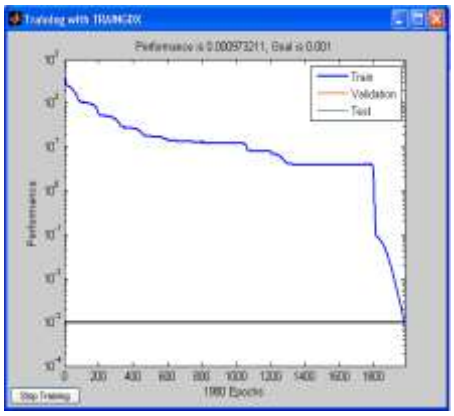

Figure 6.4: Performance Curve for speaker 1 sentence 5th

E. Equations

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.

Program code

```
n=size(x);
j=1;k=0;m=0;p=0;

for i=1:1:n
   p=p+1;
   if k==0
      if x(i)>0.050 || x(i)<-0.050
         k=1;
      end
   end

   if k==1
      if x(i)<0.050 && x(i)>-0.050
         m=m+1;
      end

      if m<p
         m=0;p=0;
      end

      if m<10
         w(j)=x(i);
         j=j+1;

      end

   end
   end
subplot(2,1,1)
plot(x)
subplot(2,1,2)
 plot(w)
```

F. Other Recommendations

SCOPE OF FUTURE WORK

In this proposed work I worked on MFCC . The growth of speech recognition technology in the past five years is amazing. We have come from a market with high-priced products that relied on discrete dictation, or speaking in a r-o-b-o-t-i-c way, to a market where speech recognition technology is common both in the office and at home. People can speak in a natural voice to interact with their computers. This, combined with affordable pricing, and increased consumer demand, is leading to the evolution of transparent computing, where human/machine interaction is to natural that it is almost invisible. In addition to the telephone and mobile devices, speech recognition is making Web-based information accessible, thanks to recent innovations such as Voice XML. Voice extensible markup language (Voice XML) will open enterprise applications for voice access, in the same way that HTML has enabled the development of graphical user interfaces. For example, Voice XML will let a person use a smart phone to access applications on the network by voice, touch, or key input as appropriate, see the results on the smart phone's display, and hear the results read back. As convenient as a desktop browser, Voice XML lets multiple devices access a company's information because data access is controlled by voice and accessed from a single point.

VII. EDITORIAL POLICY

In this proposed work I worked on MFCC . The growth of speech recognition technology in the past five years is amazing. We have come from a market with high-priced products that relied on discrete dictation, or speaking in a r-o-b-o-t-i-c way, to a market where speech recognition technology is common both in the office and at home. People can speak in a natural voice to interact with their computers. This, combined with affordable pricing, and increased consumer demand, is leading to the evolution of transparent computing, where human/machine interaction is to natural that it is almost invisible. In addition to the telephone and mobile devices, speech recognition is making Web-based information accessible, thanks to recent innovations such as Voice XML. Voice extensible markup language (Voice XML) will open enterprise applications for voice access, in the same way that HTML has enabled the development of graphical user interfaces. For example, Voice XML will let a person use a smart phone to access applications on the network by voice, touch, or key input as appropriate, see the results on the smart phone's display, and hear the results read back. As convenient as a desktop browser, Voice XML lets multiple devices access a company's information because data access is controlled by voice and accessed from a single point.

VIII. CONCLUSION

This work successfully presents an approach based on neural networks for voice recognition and user identification. This approach based on MFCC- domain by using k means clustering method. We have found a good result after testing the voice dependent system. By using one against one classification method every feature vector from the test sample is put for classification and ranking is done among the identities.

The one with the maximum rank is the identity of the test sample and finally we are getting the claimed speaker identity.

APPENDIX

Recognition Of Voice Using Mel Cepstral Coefficient & Vector Quantization
1.Priyanka Mishra, Suyash Agrawal / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.c
Issue 2,Mar-Apr 2012, pp.933-938

2. Recognotion of Speaker Useing Mel Frequency Cepstral Coefficient & Vector Quantization for Authentication International Journal of Scientific & Engineering Research Volume 3, Issue 8, August-2012 1 ISSN 2229-5518

ACKNOWLEDGMENT

I find myself lacking in expression for extending my profound sense of respect and deepest gratitude to my project guide under whose precise guidance and gracious encouragement I had the privilege to work. Had it not been for his incredible help coupled with valuable suggestions, relentless efforts and constructive criticism, this would never have become an interesting task. Moreover, his optimistic attitude, vision and appreciation was such as to give impetus to my own thoughts and understandings, making me believe that, all that was accomplished was of my own efforts for which I will ever remain indebted to him. Lastly, I feel immensely moved in expressing my indebtedness to my revered parents whose sacrifice, guidance and blessings helped me to complete my work.

REFERENCES

1] . MFCC and Its Applications in Speaker Recognition Vibha Tiwari IJET pp 19-22, 2010.
[2] Cepstral Features and Text-Dependent Speaker Identification –A Comparative Study Atanas Ouzounov C& Vol 10. No 1 2010.
[3] An Overview of Text-Independent Speaker Recognition: from Features to Supervector, Tomi Kinnunen, July 1, 2009.
[4] 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.
[5] A Review on Speech Recognition Technique, Santosh K.Gaikwad, International Journal of Computer Applications (0975 – 8887) Volume 10– No.3, November 2010.
[6] Text Independent Speaker Identification using Integrated Independent Component Analysis with Generalized Gaussian Mixture Model, Dr.V Sailaja, International Journal of Advanced Computer Science and Applications, Vol. 2, No. 12, 2011.
[7] Speech Recognition by Machine: A Review, M.A.Anusuya, International Journal of Computer Science and Information Security,Vol. 6, No. 3, 2009.
[8] HINDI SPEECH RECOGNITION SYSTEM USING HTK, Kuldee Kumar, International Journal of Computing and Business Research ISSN (Online) : 2229-6166 Volume 2 Issue 2 May 2011.
[9] Wavelet Formants Speaker Identification Based System via Neural Network, K. Daqrouq, International Journal of Recent Trends in Engineering, Vol 2, No. 5, November 2009
[10] Text Independent Speaker Recognition Using the Mel Frequency Cepstral Coefficients and A Neural Network Classifier Hassen Seddi AmelRahmouni and Mounir Sayad IEEE pp 631-634, 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 Gerhar 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 May10,2006.citeseerx.ist.psu.edu/viewdoc/download/doi=10.1.1.138...