International Journal of Scientific Engineering and Applied Science (IJSEAS) - Volume-1, Issue-5, August 2015
ISSN: 2395-3470
www.ijseas.com

Review On:Security Based On Speech Recognition Using MFCC Method

Surekha Rathod [1], Sangita Nikumbh[2]
[1] Yadavrao Tasgaonkar Institute Of Engineering & Technology, Karjat , India
[2] Yadavrao Tasgaonkar Institute Of Engineering & Technology, Karjat, India

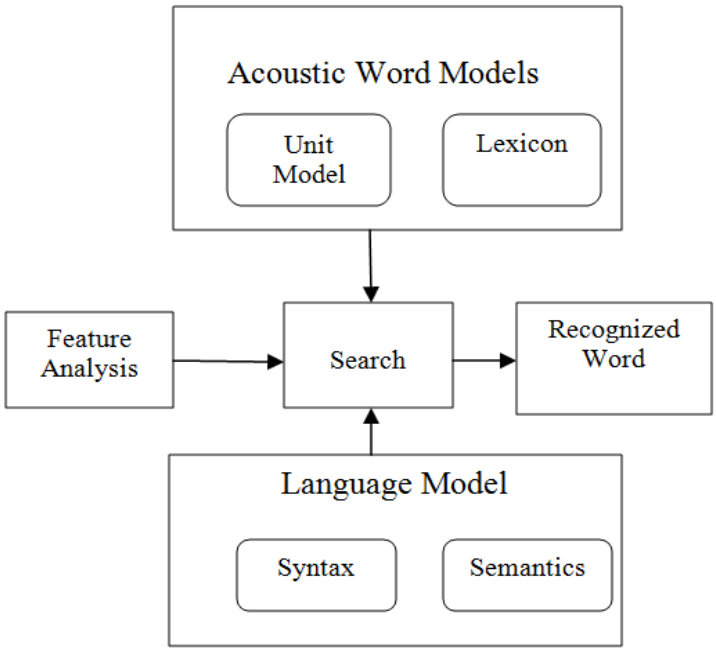
Abstract
Speech is most, effective and natural medium to exchange the information among people. People are comfortable with speech, so that any person would also like to interact with computer via speech. Speech recognition basically means talking to a computer, having it recognize what we are saying and lastly, doing this in real time. The main goal of speech recognition technologies is to allow machine to , "hear" , "understand" ,and "act upon" human spoken word. Speech recognition technology is used for security purpose. Security system means method by which something is secured through a system of interworking component and devices, that is protection from harm. There are many security system in IT realm like computer security, internet security, network security, information security. To prevent the information, the speech goes under the process of some software tools, provide useful and valuable service. In this paper we present review on security based on speech recognition using MFCC(Mel Frequency Cepstral coefficients) method.
Keywords: *Acoustic Word Model ,Feature Extraction, LPC, MFCC, Speech Recognition.*

1. Introduction

Speech processing is one of the exciting areas of signal processing. The goal of speech recognition area is to developed technique and system to enable computer to act upon human voice. Speaker recognition methods can be divided into text-independent and text-dependent methods. In a text-independent system, speaker models capture characteristics of somebody's speech which show up irrespective of what one is saying. In a text-dependent system, on the other hand, the recognition of the speaker's identity is based on his or her speaking one or more specific phrases, like passwords, card numbers, PIN codes, etc.

1.1 Basic Speech Recognition System:

Figure (1) shows basic block diagram of speech recognition system. In speech recognition system the feature analysis module provides the acoustic feature vectors used to characterize the spectral properties of the time varying speech signal. The input to speech recognizer is in the form of a stream of amplitudes, sampled at about 16,000 times per second. But audio in this form is not useful for the recognizer. Hence, Fast-Fourier transformations are used to produce graphs of frequency components describing the sound heard for 1/100th of a second. Any sound is then identified by matching it to it closest entry in the database of such graphs, producing a number, called the "feature number" that describes the sound.



Figure(1):Basic Block Diagram Of Speech Recognition

In acoustic word model, the word level acoustic match module evaluates the similarity between the input feature vector sequence (corresponding to a portion of the input speech) and a set of acoustic word models for all words in the recognition task vocabulary to determine which words were most likely spoken. In unit model, Unit matching system provides likelihoods of a match of all sequences of speech recognition units to the input speech. These units may be phones, syllables or derivative units such as fenones and acoustic units. They may also be whole word units or units corresponding to group of 2 or more words. Each such unit is characterized by some HMM whose parameters are estimated through a training set of speech data. In lexicon, lexical decoding constraints the unit matching system to follow only those search paths sequences whose speech units are present in a word dictionary. In language model, language model to determine the most likely sequence of words. Language model contain syntax model and semantics model. Syntactic and semantic rules can be specified, either manually, based on task constraints, or with statistical models such as word and class N-gram probabilities. In syntax model, Apply a "grammar" so the speech recognizer knows what phonemes to expect. This further places constraints on the search sequence of unit matching system. A grammar could be anything from a context-free grammar to full-blown English.
In semantics model, this is a task model, as different words sound differently as spoken by different persons. Also, background noises from microphone make the recognizer hear a different vector. Thus a probability analysis is done during recognition. A hypothesis is formed based on this analysis. A speech recognizer works by hypothesizing a number of different "states" at once. Each state contains a phoneme with a history of previous phonemes. The hypothesized state with the highest score is used as the final recognition result. In search, Search and recognition decisions are made by considering all likely word sequences and choosing the one with the best matching score as the recognized sentence.

1.2 Use of MFCC Method In Speech Recognition:

MFCC stands for Mel Frequency Cepstral Coefficient. The speech signal consists of tones with different frequencies. For each tone with an actual Frequency, f, measured in Hz, a subjective pitch is measured on the 'Mel' scale. To capture the phonetically important characteristics of speech, signal is expressed in the Mel frequency scale. The Mel scale relates perceived frequency, or pitch, of a pure tone to its actual measured frequency. Humans are much better at discerning small changes in pitch at low frequencies than they are at high frequencies. Incorporating this scale makes our features match more closely what humans hear. The mel -frequency scale is a linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz. As a reference point, the pitch of a 1kHz tone, 40dB above the perceptual hearing threshold, is defined as 1000 mels. In sound processing, the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.

Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the speech. The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping can allow for better representation of sound.

2. Feature Extraction:

Feature extraction is the process that extracts a small amount of data from the voice signal that can later be used to represent each utterance Feature extraction involves analysis of speech signal. Broadly the feature extraction techniques are classified as temporal analysis and spectral analysis technique. In temporal analysis the speech waveform itself is used for analysis. In spectral analysis spectral representation of speech signal is used for analysis. There are different method used for feature extraction, such as Linear Prediction Coding(LPC), Mel-Frequency Cepstrum Coefficients(MFCC), and others.

2.1 Linear Prediction Coding(LPC):

One of the most powerful signal analysis techniques is the method of linear prediction. LPC of speech has

become the predominant technique for estimating the basic parameters of speech. It provides both an accurate estimate of the speech parameters and it is also an efficient computational model of speech. The basic idea behind LPC is that a speech sample can be approximated as a linear combination of past speech samples. Through minimizing the sum of squared differences (over a finite interval) between the actual speech samples and predicted values, a unique set of parameters or predictor coefficients can be determined. These coefficients form the basis for LPC of speech. The analysis provides the capability for computing the linear prediction model of speech over time. The predictor coefficients are therefore transformed to a more robust set of parameters known as cepstral coefficients. As human voice is nonlinear in nature, Linear Predictive Codes are not a good choice for speech estimation.

2.2 Mel Frequency Cepstral coefficients(MFCC):

MFCC method is best and more popular used for feature extraction in speech recognition. In MFCC method, the drawbacks present in LPC Method is reduced. MFCCs being considered as frequency domain features are much more accurate than time domain features. Mel-Frequency Cepstral Coefficients (MFCC) is a representation of the real cepstral of a windowed short-time signal derived from the Fast Fourier Transform (FFT) of that signal. Figure(2) shows Steps involved in MFCC feature extraction. MFCC consists of seven computational steps:

Step 1:The Speech Input:

The speech input is recorded at a sampling rate higher than 10khz.This Sampling Frequency is chosen to minimize the effects of aliasing in the analog-to-digital conversion process.

Step 2: Framing:

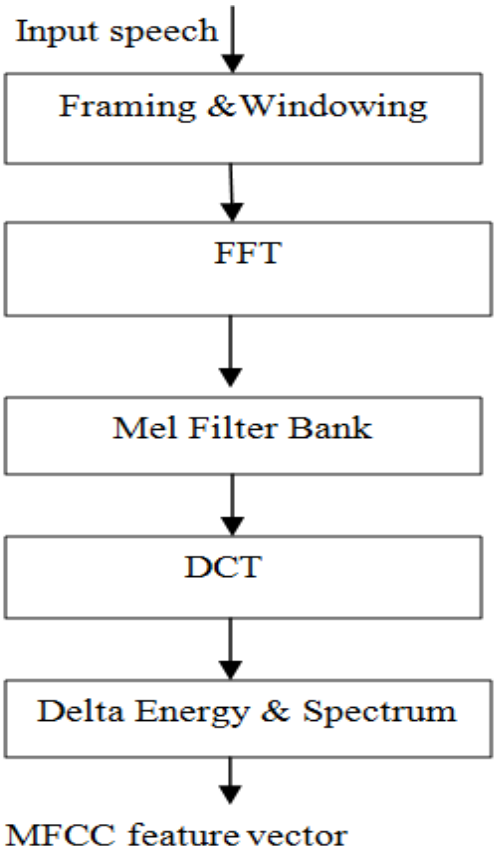
Speech samples obtained from analog to digital conversion (ADC) are segmented into a small frame with the length within the range of 20 to 40 msec.

Step 3: Hamming windowing:

The windowing process is the act of multiplying N sample of the signal by a window defined as,

hn = hnw(n)…….(1)

n€{0,1,….,N-1}



Figure(2):Steps Involved in MFCC Method

The Hamming window is by far the most popular window used in speech processing. Equation (2) presents the N-point Hamming window,

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right)\text{……...(2)}$$

N-Sample period of a frame

Hamming window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines. Hamming window is applied to minimize the discontinuities of a signal.

Step 4: Fast Fourier Transform:

The next step in the processing of the speech data to be able to compute its spectral features is to take a Discrete Fourier Transform of the windowed data. This is done using the FFT algorithm. Each frame of

N samples was converted from time domain into frequency domain. The Fourier Transform is used to convert the convolution of the glottal pulse and the vocal tract impulse response in the time domain into frequency domain.

Step 5: Mel Filter Bank Processing

The human auditory perception is based on a scale which is somewhat linear up to the frequency of 1000 Hz and then becomes close to logarithmic for the higher frequencies. This was the motivation for the definition of Pitch in the Mel-scale. Mel filter-bank to model the auditory system. The frequencies range in FFT spectrum is very wide and voice signal does not follow the linear scale. The bank of filters according to Mel scale is then performed. One approach to simulating the subjective spectrum is to use a filter bank, one filter for each desired mel frequency component. The filter bank has a triangular bandpass frequency response, and the spacing as well as the bandwidth is determined by a constant mel-frequency interval.

Step 6: Discrete Cosine Transform

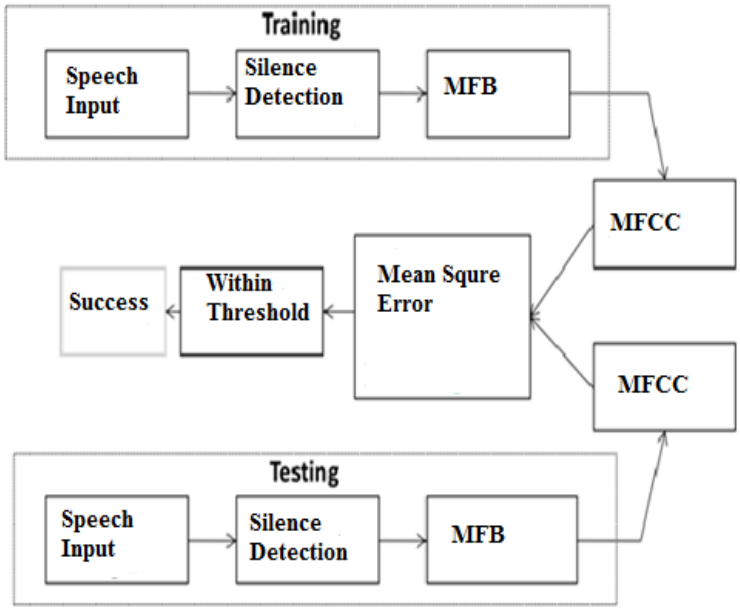
Converting the log Mel spectrum into time domain using Discrete Cosine Transform(DCT).The result of the conversion is called Mel Frequency Cepstrum Coefficient.

Step 7: Delta Energy and Delta Spectrum

The voice signal and the frames changes over time, such as the slope of a formant at its transitions. Therefore, there is a need to add features related to the change in cepstral. By applying the procedure, for each speech frame of about 30 ms with overlap, a set of mel-frequency cepstrum coefficients is computed. This set of coefficients is called an acoustic vector. These acoustic vectors can be used to represent and recognize the voice characteristic of the speaker. Therefore each input utterance is transformed into a sequence of acoustic vectors.

3. Proposed System

The analysis of various feature extraction techniques signifies the MFCC in regards of the efficient speech recognition system. The implementation is speech recognition system with single utterance. In the experimentation, the results are analyzed for the single utterance in MATLAB environment. Figure (3) show block diagram for speech recognition system with single utterance.



Figure(3) Speech Recognition System With Single Utterance

Following steps involved in block diagram:

Step 1 : Recording of input speech during training phase:

The speech input is typically recorded with a microphone at a sampling rate above 10000 Hz. This sampling frequency was chosen to minimize the effects of aliasing in the analog to digital conversion.

Step 2: Silence Detection:

Silence detection will remove the non-speech segments from the utterance for the faithful processing of the speech recognition system.

Step 3: Mel Filter Bank:

Mel filter banks the Mel frequency cepstral coefficients vector with certain dimension will be obtained during the training and testing sessions.

Step 4:Mean Square Error:

If the mean square error is well below the threshold then the success will be asserted.

Step 5 : Testing Phase:

During the testing phase the same word is uttered with approximate same energy. The uttered signal is compared with the one which was uttered during the training phase. The mean square error is determined between the two signals. If the mean square error is within the threshold determined by the user the word is said to be detected and the 'Access granted' signal is displayed in the command window and 'thank you' audio file will be played in the background. A user can save important data in a folder, the name of folder is "LOCKER". When access granted then a folder is open .So that, only user can open the folder. when third person try to open the folder then the mean square is not below the threshold the 'Access Denied' signal is displayed in the command window and the 'Try again' wave file is played in the background. the window will be closed.

4. Performance Parameter Evolution

The comparison between different speech recognition systems are based on: Mean Square Error (MSE), Recognition accuracy and Power Spectral density resolution.

4.1 Mean Square Error (MSE) :

It is defined as the square of error between training input speech & testing input speech signal. The distortion in the speech signal can be measured using MSE. The error is the amount by which the value implied by the estimator differs from the quantity to be estimated. It is calculated as follows:

$$\text{MSC} = \frac{1}{2}\sum_{n=0}^{\infty}[y(n) - x(n)]\text{^}2$$

where, x(n) : training input speech signal value and

y (n) : testing input speech signal value

4.2 Recognition Accuracy :

It is a measure used in science and engineering that compares the level of a desired signal to the level of false signal. It is defined as the ratio of number of correctly recognized words to the total number of words uttered

$$\text{Recognition Accuracy} = a(n) \div b(n)$$

where, a(n) : Correctly Recognized utterances

b(n) : total number of utterances

4.3 Time and Frequency Resolution of PSD :

The degree to which the finer details of the power spectrum can be achieved is usually measured in terms of the frequency and time resolution. The resolution always boosts importance of analysis in the speech recognition technique. It allows us to highlighting the energy content of the utterances according to time and frequency scale.

4.4 Testing

MATLAB is a numerical computing environment and fourth generation programming language created by Math Works. One of the reasons for selecting MATLAB is to fit perfectly in the necessities of an speech processing research due to its inherent characteristics and helpful to solve problems with matrix and vector formulations.

5.Conclusion

It is concluded that the proposed research uses the technique of MFCC to extract unique and reliable human voice feature pitch in the form of Mel frequency. Because we have discussed LPC and MFCC extraction method. LPC parameter is not so acceptable because of its linear computation nature. As human voice is nonlinear in nature, LPC are not a good choice for speech estimation. MFCC is derived on the concept of logarithmically spaced filter bank, clubbed with the concept of human auditory system and hence had the better response. So we get more & more security based on speech recognition using MFCC method.

Acknowledgments

I would like to take this opportunity to express my heartfelt thanks to my guide Mrs. Sangita Nikumbh for her esteemed guidance and encouragement, especially through difficult times. Her suggestions broaden my vision and guided me to succeed in this work. I am also very grateful for his guidance and comments while designing part of my project and learnt many things under his leadership.

I would like to express my appreciation for the wonderful experience while completions of this project work.

References

[1] Seiichi Nakagawa " speaker identification and verification by combining MFCC and phase information," IEEE transaction on audio,speech ,& language processing may 2012.

[2] Vimala.C,"A Review On Speech Recognition Challlenges & Approaches"World Of Computer Science & Information Technology Journal(WCSIT),2012

[3]Jeevanesh.J.Chavathe,P.V.Thakre,"Speech Operated System Using DSP:A Review"IJESRT,Dec 2013

[4] Md.Rashidual Hasan,Mustafa Jamil,"Speaker Identification Using MFCC"Internationl Conference On Electrical & Computer Engineering ICECE-2012 ,IEEE transaction on speech processing may 2012

[5] F. Bimbot et al., "A tutorial on text-independent speaker verification,"EURASIP J. IEEE transactionon Appl. Signal Process., pp. 430–451, 2004.

[6] L. Liu, J. He, and G. Palm, "Effects of phase on the perception of intervocalic

stop consonants," Speech Commun., vol. 22, pp. 403–417,1997.

[7] K. K. Paliwal and L. Alsteris, "Usefulness of phase spectrum in human speech perception," in Proc. Eurospeech'03, 2003, pp. 2117–2120.

[8] G. Shi et al., "On the importance of phase in human speech recognition," IEEE Trans. Audio, Speech, Lang. Process., vol. 14, no. 5, pp. 1867–1874, Sep. 2006.

[9] R. Schluter and H. Ney, "Using phase spectrum information for improved speech recognition performance," Proc. ICASSP, vol. 1, pp.133–136, 2001.

[10] P. Aarabi et al., Phase-Based Speech Processing. Singapore: World Scientific, 2005.

[11] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker verification," IEEE Signal Process. Lett., vol. 13, no. 1, pp. 52–5.

[12] R. M. Hegde, H. A. Murthy, and G. V. R. Rao, "Application of the modified group delay function to speaker identification and discrimination,"in Proc. ICASSP, 2004, pp. 517

First Author

Surekha Baban Rathod, M.E student of Electronics and telecommunication Dept,. Yadavrao Tasgaonkar Institute Of Engineering & Technology, Karjat , India

Second Author

Prof. Sangita Nikumbh, Professor of Electronics Dept,. Yadavrao Tasgaonkar Institute Of Engineering & Technology, Karjat , India