Available online www.ejaet.com

European Journal of Advances in Engineering and Technology, 2016, 3(1): 34-39

Research Article ISSN: 2394 - 658X

Performance Evaluation of Feature Extraction Algorithms Used For Speaker Detection

V Divyateja[1], P Aruna Kumari [1] and Dhiren Kumar Sethi[2]

[1] Department of Electronics and Communication, Gayatri Vidya Parishad College of Engg., Visakhapatnam, India
[2] Scientist-D, Naval Science & Technological Laboratory (NSTL), Visakhapatnam, India
dia.tez6@gmail.com

ABSTRACT

Speaker recognition is defined as the ability to recognize an individual based on the speech signal. The speech signals are recorded using microphones and the data stored in .wav format. Since the speaker recognition systems are mostly used in security based applications modeling, the distribution in the presence of noise and extracting the features efficiently with moderate or low signals is very much needed. The input speech signal is processed and converted in to machine readable format. The features are extracted using feature extraction algorithm. The identity of the speaker is then established with the matching of those feature vectors with that of the data available in database. Several models are utilized for the effective recognition of the speaker. In this paper using different techniques like MFCC, SDC and PCA, the speech is detected along with gamma distribution. The software used for implementing the above techniques is MATLAB.

Key words: MFCC, SDC, LPC

INTRODUCTION

Speech serves as a communication channel for the humans to interact each other and sharing the information. The transmitted speech signals in the wave forms help to understand the speech messages, such as, language of speech, text and speakers identify along with the ability to recognize a speaker (speaker Recognition), interpret the language (Language Interpreter) and the interpretation of text from the speech (speech Recognition) [1]. The rapid developments in the area of digital signal processing , influenced the usage of speech processing techniques which are widely used in areas not limited to speech enhancement, speech perception/ recognition and speech synthesis.. Lot of research is projected in the area of speech recognition, since it has a potential scope with the security related applications, where the individuals identity is established based on the speech signals. [2-5].

Basics of Speaker Recognition System

Speaker recognition is defined as the ability to recognize an individual based on the speech signal. Speaker recognition is classified into

Speaker Identification: The speaker identification is also called as 1: N mapping where the test speaker's identity is compared with that of the registered speakers or trained speakers in the data Base.

Speaker Verification: The speaker verification, also called 1:1 mapping, is a process of identifying whether the speaker is the one who claims to be [2].

Automatic Speaker Recognition

Automatic speech recognition (ASR) can be defined as the independent, computer-driven transcription of spoken language into readable text in real time. In a nutshell, ASR is technology that allows a computer to identify the words that a person speaks into a microphone or telephone and convert it to written text. Having a machine to understand fluently spoken speech has driven speech research for more than 50 years. Although ASR technology is not yet at the point where machines understand all speech, in any acoustic environment, or by any person, it is used on a day-to-day basis in a number of applications and services. Commercially available ASR systems usually require only a short period of speaker training and may successfully capture continuous speech with a large vocabulary at normal pace with a very high accuracy. Most commercial companies claim that recognition software can achieve between 98% to 99% accuracy if operated under optimal conditions. Optimal conditions' usually assume that users: have speech characteristics which match the training data, can achieve proper speaker adaptation, and work in a clean noise environment (e.g. quiet space). This explains why some users, especially those whose speech is heavily accented, might achieve recognition rates much lower than expected [1].

History of ASR

The earliest attempts to devise systems for automatic speech recognition by machine were made in the 1950s. Much of the early research leading to the development of speech activation and recognition technology was funded by the National Science Foundation (NSF) and the Defence Department's Defence Advanced Research Projects Agency (DARPA). Much of the initial research, performed with NSA and NSF funding, was conducted in the 1980s. Speech recognition technology was designed initially for individuals in the disability community. For example, voice recognition can help people with musculoskeletal disabilities caused by multiple sclerosis, cerebral palsy, or arthritis achieves maximum productivity on computers. During the early 1990s, tremendous market opportunities emerged for speech recognition computer technology. The early versions of these products were clunky and hard to use. The early language recognition systems had to make compromises: they were "tuned" to be dependent on a particular speaker, or had small vocabulary, or used a very stylized and rigid syntax. However, in the computer industry, nothing stays the same for very long and by the end of the 1990s there was a whole new crop of commercial speech recognition software packages that were easier to use and more effective than their predecessors [2].

Working of ASR

The goal of an ASR system is to accurately and efficiently convert a speech signal into a text message transcription of the spoken words independent of the speaker, environment or the device used to record the speech (i.e. the microphone). This process begins when a speaker decides what to say and actually speaks a sentence. The software then produces a speech wave form, which embodies the words of the sentence as well as the extraneous sounds and pauses in the spoken input. Next, the software attempts to decode the speech into the best estimate of the sentence. First it converts the speech signal into a sequence of vectors which are measured throughout the duration of the speech signal. Then, using a syntactic decoder it generates a valid sequence of representations.

FEATURE EXTRACTION

Feature Extraction is a process of extracting the features from a speech signal; it has its vital influence in effective speaker recognition. .It can also be defined as a transformation process of a speech signal into a sequence of Feature Vectors For the Speaker Independent / Speaker Verification lot of training and test data is needed for the classification of an individual speaker, this increases the dimension of the problem, in order to reduce the dimension, Feature Vector will be very much useful. The feature vectors used in this thesis include both acoustic features and prosodic features, such as MFCC, SDC and PCA [1].

In feature extraction the speech waves stored in wav format each converted to a parametric form. The speech signals remains stationary between the time intervals 5 ms to 100 ms. And the changes observed over long periods i.e. 0.2sec or more. Therefore to identify the speech variation in short time sequence, cepstral analysis is mostly preferred hence MFCC are considered**.** Linear prediction coding (LPC) coefficient helps to extract signal more effectively in the presence of noise and when the speech signal is of very short duration .so in this thesis we have exploited MFCC combined with LPC to have effective feature vector identification.

In speech analysis, significant information spread over few 100s of ms. There may be overlaps and the speech signals are not completely separated in-time. These overlaps may result in to ambiguities at the time of classification to overcome this it is assumed to extract the features between the frequencies 2 to 16 Hz, a maximum of 4 Hz. In order to distinguish these signals in the overlapping situations delta features are mostly preferred. In delta coefficients we obtained the derivative to estimate the differences in the speech trajectories. Delta-delta coefficients are also considered for every longer temporal context. But these features will be effective for short term speech samples, for long term features shifted delta coefficients (SDC) are well proffered. The features obtained from MFCC are converted to shifted delta coefficients. It is observed that the features obtained from MFCC followed by SDC outperform MFCC followed by delta. SDC reflects the dynamic cepstral features along with pseudo-prosodic feature behaviour [4].

SPEAKER RECONITION ALGORITHM

The following steps describe our approach of the speaker recognition, in this chapter of thesis.
Step1: Obtain the training set by recording the speech voices in a .wav form
Step2: Pre-emphasis the speech signals to remove silence and noise.
Step3: Identify the compound feature vectors feature vector of these speech signals by using MFCC, LPC, SDC, delta, delta-delta.

Step4: Generate the Probability Density Function (PDF) of the generalized gamma distribution for all the trained data set.
Step5: Same procedure is followed for test sequence.
Step6: Find the range of speech of test signal in the trained set.
Step7: Evaluation metrics such as Acceptance Rate (AR), false acceptance rate (far), and missed detection rate (mdr) are calculated to find the accuracy of speaker recognition.

MFCC- MEL-FREQUENCY CEPSTRAL COEFFICIENTS

Mel-Frequency Cepstral Coefficients (MFCC) is one of the most popular approaches of feature extraction technique in both speaker recognition and LID system. MFCCs are based on the known variation of the human ears critical bandwidths with frequency, filters spaced linearly at low frequencies and logarithmically at high frequencies that have been used to capture the phonetically important characteristics of speech. The characteristics are expressed on the mel-frequency scale, which is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. Therefore we can use the following approximate formula to compute the mels for a given linear frequency f in Hz. The Mel-Frequency Cepstral Coefficients (MFCC) features is the most commonly used features in speaker recognition. It combines the advantages of the cepstrum analysis with a perceptual frequency scale based on critical bands. MFCC is based on human hearing perceptions which cannot perceive frequencies over 1Khz. θIn other words, in MFCC is based on known variation of the human ear’s critical bandwidth with frequency. MFCC has two types of filter which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz [1], [3-5].

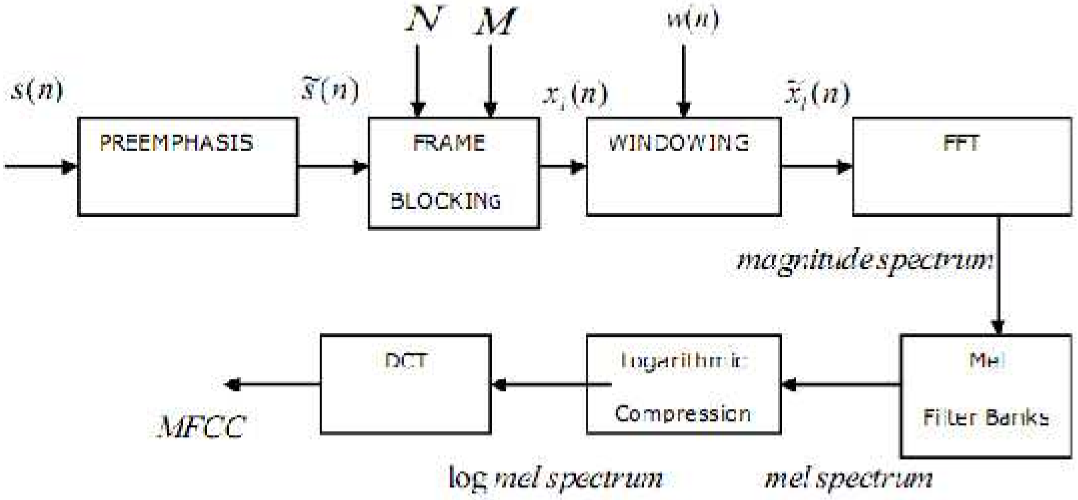

Fig.1 Feature Extractor schematic diagram for MFCC analysis[3]

Pre-emphasis

Pre-emphasis is needed because high frequency components of the speech signal have small amplitude with respect to low frequency components.

Framing

The width of the frames is generally about 30ms with an overlap of about 20ms (10ms shift). Each frame contains N sample points of the speech signal. Overlap rate of frames, between %30 and % 75 of the length of the frames. This is why we frame the signal into 20-40ms frames. If the frame is much shorter we don't have enough samples to get a reliable spectral estimate, if it is longer the signal changes too much throughout the frame. It is assumed that although the speech signal is non-stationary, but is stationary for a short duration of time.

Windowing

The window function is used to smooth the signal for the computation of the DFT. The DFT computation makes an assumption that the input signal repeats over and over. The discontinuity in the frame is prevented. If there is a discontinuity between the first point and the last point of the signal, artifacts occur in the DFT spectrum. By multiplying a window function to smoothly attenuate both ends of the signal towards zero, this unwanted artifacts can be avoided. The objective is to reduce the spectral effects.

Windowing functions commonly used: Hamming, Hanning, Blackman, Gauss, rectangular, and triangular. The hamming window is usually used in speech signal spectral analysis, because its spectrum falls off rather quickly so the resulting frequency resolution is better, which is suitable for detecting formants.

Fast Fourier Transforms (FFT)

To convert each frame of N samples from time domain into frequency domain. The Fourier Transform is to convert the convolution of the glottal pulse U[n] and the vocal tract impulse response H[n] in the time domain.

Mel-Filter Bank Processing

- Human hearing is not equally sensitive to all frequency bands.
- Less sensitive at higher frequencies, roughly > 1000 Hz i.e. human perception of frequency is non-linear
- Mel (melody) is a unit of pitch. Mel-frequency scale is approximately linear up to the frequency of 1 KHz and then becomes close to logarithmic for the higher frequencies.
- Human ear acts as filters that concentrate on only certain frequency components. Band-pass filters.
- These filters are non-uniformly spaced on the frequency scale, with more filters in the low frequency regions and less filters in the high frequency regions.

Log Energy Computation

- Logarithm compresses dynamic range of values.
- Human response to signal level is logarithmic. Humans are less sensitive to slight differences in amplitude at high amplitudes than low amplitudes.
- Makes frequency estimates less sensitive to slight variations in input (power variation due to speaker's mouth moving closer to mike)
- Phase information not helpful in speech.

Discrete Cosine Transform

- This is the process to convert the log Mel spectrum into time domain using Discrete Cosine Transform (DCT).
- The result of the conversion is called Mel Frequency Cepstrum Coefficient.
- The set of coefficient is called acoustic vectors.
- Therefore, each input utterance is transformed into a sequence of acoustic vector.

SDC- SHIFTED DELTA CEPSTRA

The GMM LID system improved its performance combining high order (512-1024-2048) mixture models with shifted delta cepstra (SDC) feature vectors are an extension of delta-cepstra coefficients. In language identification, SDC feature is widely used.SDC feature vectors are created by stacking delta cepstra computed across multiple speech frames. SDC coefficients are based on four parameters namely written as N-d-P-k. For each frame of data, first MFCCs are computed based on N; (i.e. c0, c1, c2, c3… cN-1).

PCA- PRINCIPAL COMPONENT ANALYSIS

PCA extracts the orthogonal principal components from this matrix and the dimensionality is reduced by calculating the 2nd order moments for the low frequency values and decor relating these values. The PCA is based on the approximation of karhunen-loeva transformation (KLT). This process highlights the principal components by calculating the Eigen values which are of high dimension and convert then to low dimensional values by retaining the original values.

LINEAR PREDICTIVE CODING (LPC)

Linear predictive coding (LPC) method was developed in the 1960s by and being used for speech vocal tracing because it represents vocal tract parameters and the data size are very suitable for speech compression. This method gives encoding good quality speech at low bit rate and provides accurate estimates of speech parameters by describing the intensity, the residue signal. Linear predictive coding (LPC) method was developed in the 1960s by and being used for speech vocal tracing because it represents vocal tract parameters and the data size are very suitable for speech compression. This method gives encoding good quality speech at low bit rate and provides accurate estimates of speech parameters by describing the intensity, the residue signal. [1]

ALGORITHM FOR SPEAKER IDENTIFICATION

Step 1: Obtain the speech signals, using Microphone and store in .wav format.
Step 2: Calculate MFCC and obtain numeric coefficients.
The speech signal is dependent of tone and to understand the spectral properties of these signals, obtains the Fourier transform.
Map these coefficients using Mel scale with triangular overlapping windows. Take the logarithm.
Obtain DCT of the Mel log sequences.
Obtain the amplitude sequences of MFCC coefficients.
Step 3: Apply PCA, to reduce dimensionality.
Step4: Apply Generalized gamma distribution, to model the parameters.

RESULTS AND DISCUSSION

This section illustrates different simulation results from MATLAB. The Fig.2 shows the recorded speech signal which is saved in database for comparison with the obtained results. The recorded signal is obtained by using the MATLAB software. Fig.3 is simulation result is the spectrum of MFCC obtained by following the algorithm above mentioned.The spectrum is obtained by following each and every step in the MFCC technique. MATLAB software is used. Fig.4 is the simulation result is obtained for a speech spectrum by following the MFCC speech detection algorithm with the SDC technique. The above speech is analogus speech. Fig.5 is the simulation result of MFCC-SDC-LPC obtained by the speaker detection algorithm. Fig .6 shows the probability density function of the speech signal. Finally the spaker speech is identified by the above algorthim.

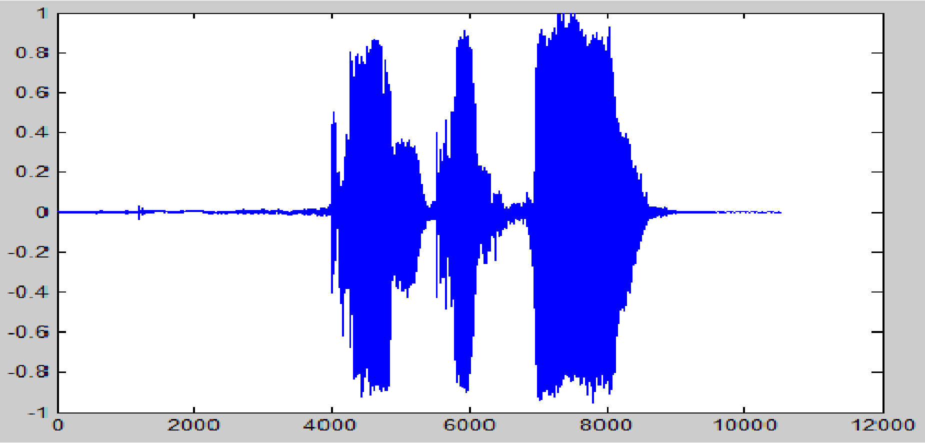

Fig.2 Specch signal of the speaker after recorded from microphone

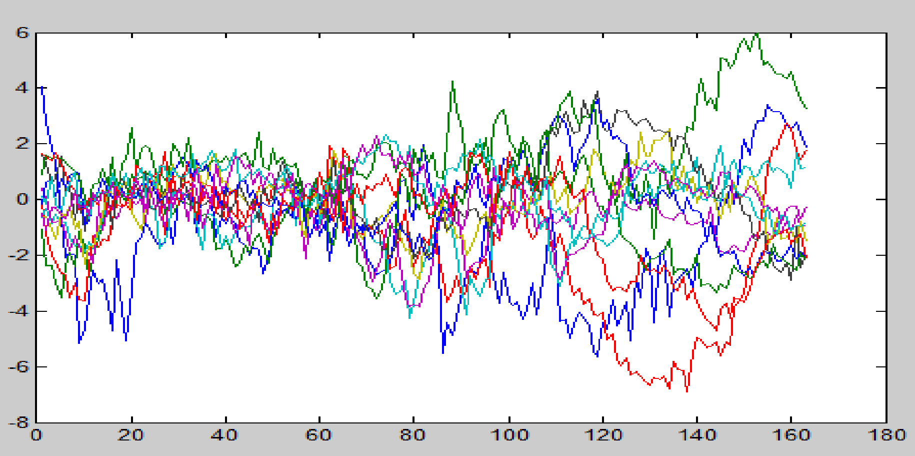

Fig.3 MFCC spectrum of speech

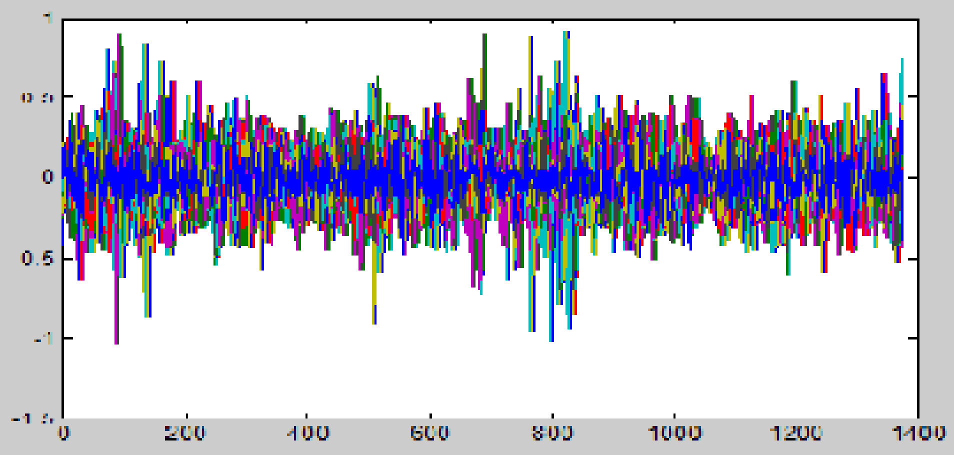

Fig.4 MFCC-SDC spectrum of speech

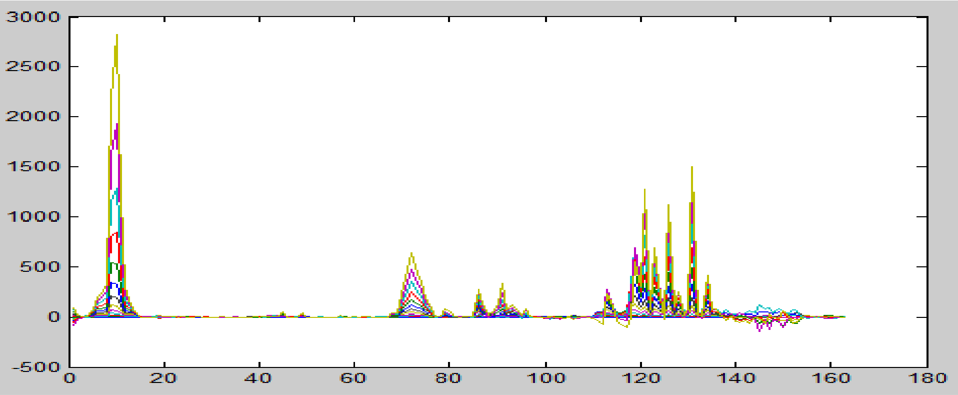

Fig.5 MFCC-SDC-LPC spectrum of speech

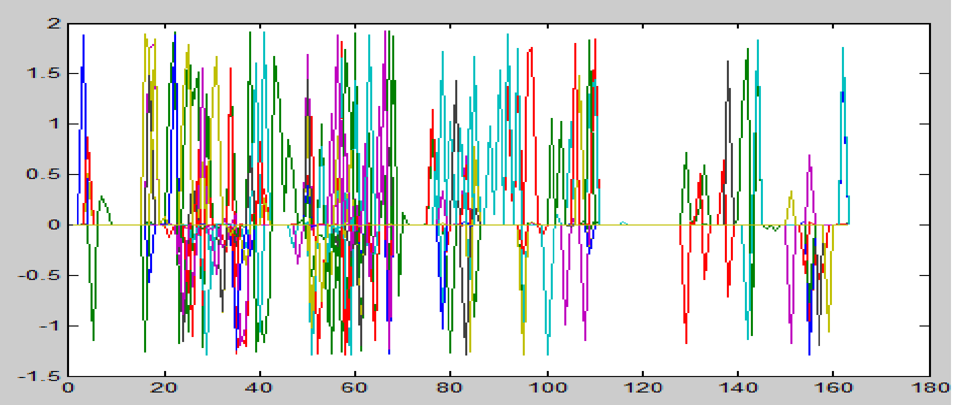

Fig.6 probability density function (PDF) of speech signal

CONCLUSION

This proposed paper is on the speaker identification and detection by recording the speech in database and then comparing it with every speech being given as input. The factors discussed in this paper are the algorithms for speaker detection and speaker identification of speech spectrum. The techniques used in this paper are MFCC, SDC, LPC and PCA. Using MATLAB and considering these factors the results are obtained.

REFERENCES

[1] Shanthi Therese S and Chelpa Lingam, Review of Feature Extraction Techniques in Automatic Speech Recognition, *International Journal of Scientific Engineering and Technology*, **2013**, 2 (6), 479-484.
[2] Suma Swamy and KV Ramakrishnan, An Efficient Speech Recognition System, *Computer Science & Engineering: An International Journal (CSEIJ)*, **2013**, 3 (4), 21-27.
[3] Kashyap Patel and RK Prasad, Speech Recognition and Verification Using MFCC & VQ, *International Journal of Advanced Research in Computer Science and Software Engineering*, **2013**, 3 (5), 478-483.
[4] Parwinder Pal Singh and Pushpa Rani, An Approach to Extract Feature using MFCC, *IOSR Journal of Engineering*, **2014,** 4 (8), 21-25.
[5] Nilu Singh, RA Khan and Raj Shree, MFCC and Prosodic Feature Extraction Techniques: A Comparative Study, *International Journal of Computer Applications*, **2012,** 54 (1), 9-13.