IOSR Journal of VLSI and Signal Processing (IOSR-JVSP)
Volume 6, Issue 2, Ver. I (Mar. -Apr. 2016), PP 07-10
e-ISSN: 2319 – 4200, p-ISSN No. : 2319 – 4197
www.iosrjournals.org

Designing A Tool Using GUI In MATLAB For Comparing Front-End Acoustic Features For Speaker Diarization System

J. S. Sohal[1], Sukhvinder Kaur[2]
[1](Director, LCET, Ludhiana, Punjab, India)
[2](Ph.D. Research Scholar, PTU, Jalandhar, Punjab, India)

Abstract: *This paper presents a Tool, designed in MATLAB mostly used to find and compare various types of features of speech signals which are used in Speaker Diarization system. It includes fundamental frequency, spectrogram, formant frequency, Mel Frequency Cepstral Coefficient (MFCC) and Vector Quantized-MFCC. Two applications for which the tool has been used are presented: one for recordings of single speaker and other for multiple speakers using single distant microphone (SDM). At the end, using cosine correlation method, various features are compared and found Modified MFCC is the best.*

Keywords: *Fundamental Frequency, Formant Frequency, MFCC, Spectrogram, Speaker Diarization.*

I. Introduction

Speaker Diarization is one of the tasks in the NIST Rich Transcription (RT) meeting recognition evaluation. It automatically find the segments of time within a meeting in which each meeting participant is talking [1].The goal of Speaker Diarization is to segment an audio recording into speaker homogenous speech regions, where the number of participants, participant identities and amount of speech are not known a priori. The sources of audio recording can include particular speakers, background noise sources, music and other channel characteristics. Under Single Distant Microphone (SDM) conditions Speaker Diarization can be roughly divided into three stages: Speech activity detection (SAD), segmentation and clustering. In Multiple Distant Microphone (MDM) condition, for dealing with multiple microphones channels, one more stage i.e. Acoustic Beamforming is added as the first stage. Under both conditions speech activity detection (SAD) involves the labeling of speech, non-speech (background noises, music, silence) and overlapping segments. SAD can have significant impact on speaker diarization performance for two regions: the first step directly from the standard speaker diarization performance metric i.e. Diarization error rate (DER) [2], which takes into account both the False alarms and missed speaker error rates. The second follows from the fact that non-speech segments can disturb the speaker diarization process. Poor SAD performance will lead to an increased DER. After SAD is performed, the initial segmentation is created. This segmentation is achieved by first splitting the speech regions into one second segments then the acoustic features like LPC, Pitch, and MFCC etc. are extracted over these segments. These feature vectors are used as input signal to the GMMs. After segmentation, similar segments are grouped together and form a cluster. For clustering Bayesian Information Criterion (BIC) is used [1].

In the process of speaker diarization, there is need of some tool for extracting features of audio signal, segmentation and clustering. Already available tools like Sphinx 4 are based on Java or C/C++. We decided to make use of MATLAB which is easy and easily available. It has several features like Graphical User Interface (GUI) [3] design environment, built-in- functions and plays .wav files for the users.

The paper is organized as follows. Section II describes the datasets used in the present work. In Section III acoustic feature extraction techniques are introduced, in section IV we present the experimental results and finally the conclusion is given in section V.

II. Data-Sets Used

In this research work three data sets are used. The first data-set contains single word spoken by 8 speakers, second data-set is concatenated short utterances of eight speakers and third is testing data-set of 2.5 minutes of video clip free down loaded from youtube.com in MP4 format. Further it is converted into .wav form.

III. Acoustic Feature Extraction And Comparision

Feature extraction can be stood as a step to reduce the dimensionality of the input speech, a reduction which inevitably leads to some information loss [4]. During feature extraction we divide the speech signal into frames and extract features for each frame. It changes the speech signals into a sequence of feature vectors. It results information loss during the transition from speech signal to a sequence of feature vector. It must be kept low. Our Goal was to design a system that fully automates the process of feature extraction of speech signals for speaker recognition and speaker diarization system. We have developed a user friendly program that finds

various features of speeches spoken by single speaker and multiple speakers. We used the MATLAB GUIDE to develop our program's GUI. Using GUI we developed two hyperlinked figures: one is used to load .wav and .mat files as shown in figure 1 & Figure 2, second for extracting features. Finally all features of speech signals are compared.

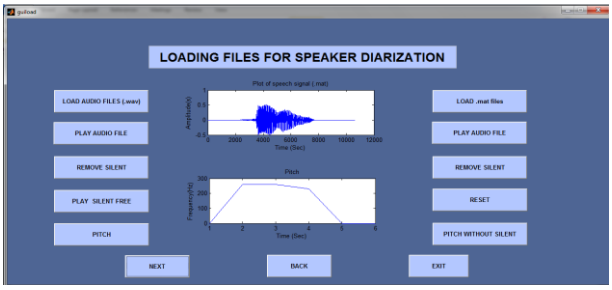

Figure1 the GUI. Waveform and pitch of recorded speech of single speaker

As discussed in [5] there are various important properties of speech used in speech-processing applications. These properties include voicing, which indicates whether or not the vocal cords are vibrating during production of sound; the fundamental frequency, which is the frequency of vocal-cord vibration; and the status of speech/non-speech activity.

In this research work we extract various features [3] of speech like pitch, formant frequency, spectrogram, Mel Frequency Cepstral Coefficient (MFCC) [4] and Vector Quantization-MFCC [6].

- Pitch is correlated with the physical feature of fundamental frequency. To estimate the fundamental frequency for a voiced sound, the repetitiveness in the signal is measured. The repetitiveness can be measured by apparent peak in the autocorrelation function at a lag corresponding to the pitch period. In autocorrelation a window of the signal is auto correlated with itself at growing intervals. The interval with the highest autocorrelation is said to be the base period and is used to determine base frequency (F0). If autocorrelation is at period '0', it implies that the frame is unvoiced.
- Formants are frequency peaks which have, in the spectrum, a high degree of energy. They are especially prominent in vowels. Each formant corresponds to a resonance in the vocal tract and can be estimated from speech spectrum with a certain level of precision.

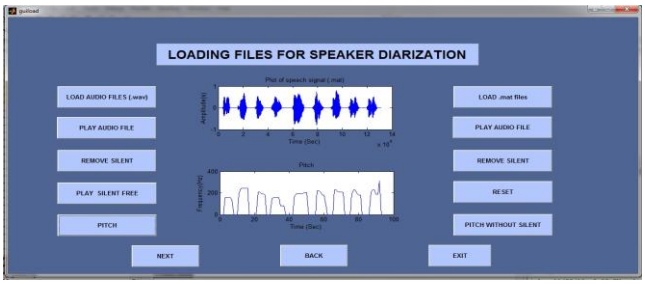

Figure 2 The GUI. Waveform and pitch of recorded speech of concatenated short utterances of eight speakers.

- Spectrogram/Specgram is a built-in- function which is used to obtain spectrograph of speech signal as shown in Figure 3. It includes three parameters: the horizontal axis is the time axis, the vertical axis is the frequency axis and the color represents the amplitude of the signal at given time and frequency.
- Cepstrum and Mel frequency cepstral coefficient (MFCC) are very important properties of speech. The cepstrum is the inverse Fourier transform of the log spectrum. Regularity in the log spectrum such as repetitive spikes at essentially equal spacing produced by excitation of voiced speech results in a clear spike in the cepstrum at a location corresponding to the pitch period.

We followed the following steps to compute MFCC:

1) Speech signal is blocked into frames of N samples. Frame size is M,(M<N), N=256, M=100.
2) Windowing of each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame
3) FFT, which converts each frame of N samples from the time domain into the frequency domain.
4) Scale of frequency is converted from linear to mel scale.
5) Logarithm is taken.
6) Log Mel scale is converted back to time domain. Using DCT. The result is called Mel frequency cepstral

coefficient.

7) Further Delta MFCC and Delta-Delta MFCC is taken. Delta MFCC is the difference between MFCs of two consecutive frames and delta-delta MFCCs corresponds to difference between Delta MFCCs of two consecutive frames.
8) At last vector quantization of MFCCs is computed using vqlbg function from mathworks.com.

- Vector quantization-MFCC: Vector quantization of MFCC uses Linde-Buzo-Gray (LBG) algorithm [6]. The acoustic vectors (MFCC) extracted from input speech of each speaker provide a set of training vectors for that speaker. The LBG algorithm designs an M-vector codebook in Stages. It starts first by designing a 1-vector codebook, then uses a splitting technique on the code-words to initialize the search for a 2-vector codebook, and continues the splitting process until the desired M-vector codebook is obtained.
- Finally we compared all the features using cosine correlation formula.

IV. Experiments And Results

In this section, we describe the experiments performed on different data sets using graphical user interface (GUI) in MATLAB. Fig.1 and Fig. 2 represents first GUI figures in which we can load .wav and .mat files of recordings respectively. We have successfully loaded and processed single utterances, concatenated short utterances and small video clips in .wav form. Fig. 1 displays the single speaker speech and its pitch. Fig. 2 presents multiple speaker speech and its pitch value. Fig.3 and Figure 4 represents all the features extracted from speech signal of single and multiple speakers respectively.

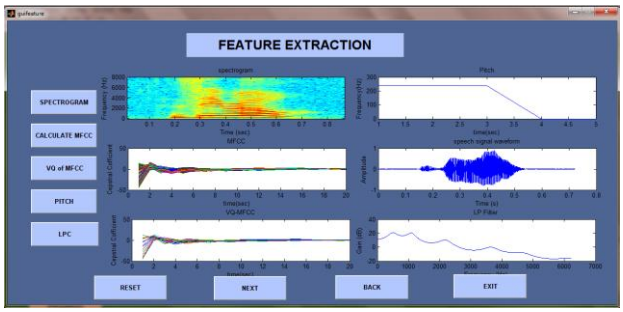

Figure 3 Showing various features of loaded speech of single speaker

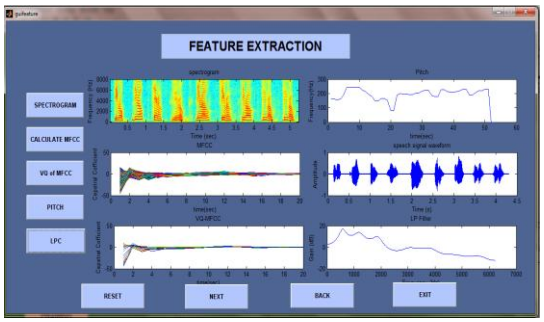

Figure 4 showing various features of loaded speech of multiple speakers

Next, we compared every non-noise vector (A) of speaker one to every non-noise vector (B) of second speaker with various extracted features using cosine correlation method:

$$\cos\theta = \frac{\left|\vec{a}\,\vec{b}\right|}{\|\vec{a}\|\|\vec{b}\|}$$

Their results are shown below in the table 1.

TABLE 1 Recognition Rate

Features	Same speaker (spoken same word)	Different speakers (spoken same word)	Same speaker (spoken different words)
Spectrogram	100%	53.20%	43.36%
Pitch	100%	86.73%	30.99%
MFCC	100%	87.36%	60.46%
VQ-MFCC	100%	87.83%	60.81%

V. Conclusion

Using Graphical User Interface (GUI) design environment in MATLAB we developed a tool to present various features of speech signals and their comparison. We described Pitch, Formant frequency, spectrogram and MFCC of speech signal. We also modified the MFCC by vector quantization method. After applying correlation function to match speeches of two speakers, Modified MFCC gives best results.

Further work will be separated into two research areas: Improved segmentation and clustering of audio recording in speaker diarization system.

References

[1] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization : A review of recent research," IEEE Transactions on Audio, Speech and Language Processing., 20(2), February 2012, 356–370.

[2] M.T. Knox, N. Mirghafori, G. Friedland, "Exploring methods of improving accuracy for speaker diarization",in Interspeech Conf., Lyon, France, August 25-29, 2013, 2783-2787.

[3] S. Engelberg, Y. Saidoff, and Y. Israeli, "Voice identification through spectral analysis," IEEE Instrumentation and Measurement Magazine.October 2006, 52-55.

[4] Lee Chulhee, Hyun D., Choi Euisun, Lee Chungyong "Optimization Feature Extraction for Speech Recognition," IEEE Transactions on Speech and Audio Processing.11 (1), 2003, 80-87.

[5] Rabiner , L. and Schafer, R., Digital Processing of Speech Signals(Prentice Hall, Inc., Englewood Cliffs, New Jersey, 1978).

[6] Y. Linde, A. Buzo & R. Gray, "An algorithm for vector quantizer design", IEEE Transactions on Communications, 28, 1980, 84-95,