CHAPTER 4

AUDIO FEATURE EXTRACTION FOR QMR

4.1 Introduction

This chapter presents an automatic domain-independent method, to detect relevant features from QMR audio files, associating them to concepts modelled in a background ontology. The key to AFE is choosing a set of features that allows the computation of mathematical spaces for sounds and semantics. As describes in previous chapter, AFE is used to extract features for melodious speech processing of QMR in the development of proposed MPS system. In this section, three methods of AFE will be presented followed with an existing approach of extracting spectral descriptors and similarity measures analysis in defining and analysing the features of QMR audio signals. The aim of both methods is to identify the unique properties based on the significant features contained in the selected QMR audio signals. This chapters provide the low-level spectral analysis for second phase of spectral extraction and analysis for similarity measures based on the unique spectral descriptors. The first section discussed on the spectral descriptors applied on the audio files and later section is focused on the analysis of similarity measures for each maqamat and discussed the performance for each descriptor.

4.2 Audio Feature Extraction

Audio features extraction is considered as commonly used method among researchers in speech recognition. The process of feature extraction is required for classification, prediction and recommendation algorithms. The basic idea of audio extraction is when an input audio is windowed and framed prior performing appropriate and suitable analysis as shown in Figure 4.1. In this study, two techniques have been proposed to perform the extraction process, which are the Spectral Descriptors as the low-level parameters and Mel-Frequency Cepstral Coefficient with frequency warping as enhanced technique. Both techniques will be presented and discussed in the next section.

4.2.1 Pre-processed data

The QMR audio was captured with studio parameters i.e. no background noise. However, the unvoiced region that need to be refined and trimmed using an open source software for audio editing tool called Wavepad Audio Editor by NCH Software as in Figure 4.2. The audio quality is maintained with sampling rate of 44100Hz as the highest standard rate for audio processing and analysis.

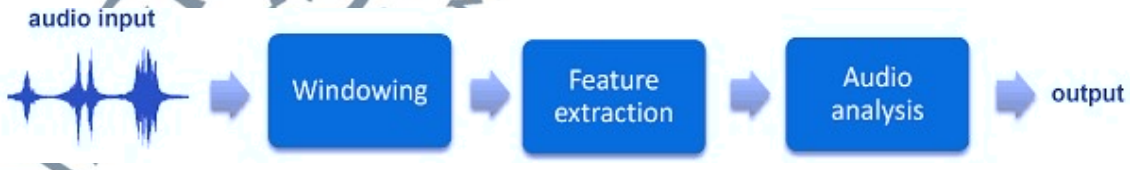

Figure 4.1 Basic Audio Feature Extraction

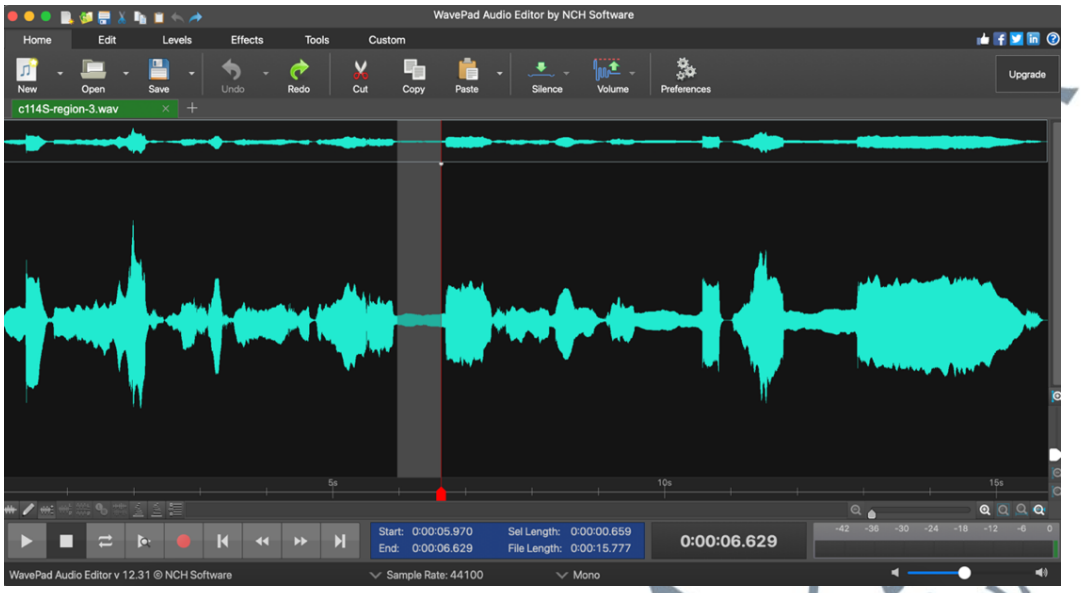

Figure 4.2 Screenshot of software audio editing tool

4.2.2 Spectral Descriptors

The Spectral Centroid is very widely used in Signal Processing to reveal the peculiar characteristics of a spectrum. It demonstrates where the “center of mass” of the audio signal spectrum is located at. It can be interpreted as having a direct relation with the brightness of the speech signal. The Spectral Centroid is basically the weighted average of the frequency components present in the audio signal. The frequency components are derived by computing the Fourier Transform and these components are further used, with the absolute value of magnitude as the weights. Spectral spread is the standard deviation around the spectral centroid. The spectral spread represents the "instantaneous bandwidth" of the spectrum. It is used as an indication of the dominance of a tone. For example, the spread increases as the tones diverge and decreases as the tones converge. The spectral rolloff point has been used to distinguish between voiced and unvoiced speech, speech/music discrimination, music genre classification, acoustic

scene recognition, and music mood classification. The features use MATLAB command as below:

```
centroid = spectralCentroid(speech,fs);
spread = spectralSpread(speech,fs);
rolloff = spectralRolloffPoint(speech,fs);
```

4.2.3 Cepstral extraction with CCA

The objective of CCA is to separate the speech into its source and system components without any a priori knowledge about source and/or system. As explain in section 2.3.2, the multiplication of the excitation and system component need to be deconvolved to vocal tract components in the time domain (Giacobello et al., 2010). For this purpose, cepstral analysis is used for transforming the multiplied source and system components in the frequency domain to linear combination of the two components in the cepstral domain. The result of CCA will be used for comparison and performance evaluation.

4.2.4 MFCC with W-DFT

MFCC among the well-known method for recognition. As explained in section 2.3.3, the aim is to mimics the frequency response of the human hearing, so the coefficients are depend on filterbank energies from mel-frequency spacing. According to steps from HMM model, the initial step is to windowed using is Hamming window in overlapping steps. In these experiments, each window is set to frame duration, T_w=25ms with overlapping time steps for frame shift, T_s = 10ms so this gives a total of 100 windows. The window width and overlap can be defined to optimize the visualizations. For each window, the log power spectrum is computed using a discrete

Fourier transform (DFT). According to Reynold, the log spectral coefficients are perceptually weighted by a non-linear map of the frequency scale which known as Mel-scale. This is to emphasizes the mid-frequency bands that proportionate to their perceptual importance. The process is further transforming the Mel-weighted spectrum to coefficients value. Only the first 12 order of coefficients are selected while remaining higher order are discarded. This will transform the log energy into 13- dimensional feature vectors (12 MFCCs plus energy) at a 100 Hz rate.

This framework is based on Dan Ellis' rastamat routines (Ellis, 2005). The emphasis is placed on closely matching MFCCs produced by HTK (Young et al., 2006) with simplicity and compactness as main considerations, but at a cost of reduced flexibility. This routine is meant to be easy to extend, and as a starting point for work with cepstral coefficients in MATLAB. MFCC returns mel frequency cepstral coefficients computed from speech signal given in vector S and sampled at FS (Hz). The speech signal is first reemphasized using a first order FIR filter with pre-emphasis coefficient ALPHA. The pre-emphasized speech signal is subjected to the short-time Fourier transform analysis with frame durations (ms), frame shifts (ms) and analysis window function given as a function handle in WINDOW. This is followed by magnitude spectrum computation followed by filterbank design with M triangular filters uniformly spaced on the mel scale between lower and upper frequency limits given in R (Hz). The filterbank is applied to the magnitude spectrum values to produce filterbank energies (FBEs) (M per frame). Log-compressed FBEs are then decorrelated using the discrete cosine transform to produce cepstral coefficients. Final step applies sinusoidal lifter to produce liftered MFCCs that closely match those produced by HTK. Figure 4.3 shows the MATLAB command lines for HTK routines.

```
%% PRELIMINARIES

% Ensure correct number of inputs
if( nargin~= 10 ), help mfcc; return; end;

% Explode samples to the range of 16 bit shorts
if( max(abs(speech))<=1 ), speech = speech * 2^15; end;

Nw = round( 1E-3*Tw*fs );    % frame duration (samples)
Ns = round( 1E-3*Ts*fs );    % frame shift (samples)

nfft = 2^nextpow2( Nw );     % length of FFT analysis
K = nfft/2+1;                % length of the unique FFT


%% FEATURE EXTRACTION

% Preemphasis filtering (see Eq. (5.1) on p.73 of [1])
speech = filter( [1 -alpha], 1, speech); % fvtool( [1 -alpha],1);

% Framing and windowing (frames as columns)
frames = vec2frames( speech, Nw, Ns, 'cols', window, false );

% Magnitude spectrum computation (as column vectors)
MAG = abs( fft(frames,nfft,1) );

% Triangular filterbank with uniformly spaced filters on melscale
H = trifbank( M, K, R, fs, hz2mel, mel2hz ); % size of H is M x K

% Filterbank application to unique part of the magnitude spectrum
FBE = H * MAG(1:K,:); % FBE( FBE<1.0 ) = 1.0; % apply mel floor

% DCT matrix computation
DCT = dctm( N,M );

% Conversion of logFBEs to cepstral coefficients through DCT
CC =  DCT * log( FBE );

% Cepstral lifter computation
lifter = ceplifter( N, L );

% Cepstral liftering gives liftered cepstral coefficients
CC = diag( lifter ) * CC;
% Hertz to mel warping function
hz2mel = @( hz )( 1127*log(1+hz/700) );
% mel to Hertz warping function
mel2hz = @( mel )( 700*exp(mel/1127)-700);
```

Figure 4.3 Screenshot of MATLAB command lines for HTK routines

MFCC also returns FBEs and windowed frames with feature vectors and frames as columns. For inputs S is the input speech signal (as vector), FS is the sampling frequency (Hz), TW is the analysis frame duration (ms), TS is the analysis frame shift (ms), ALPHA is the preemphasis coefficient, WINDOW is a analysis window function handle, R is the frequency range (Hz) for filterbank analysis, M is the number of filterbank channels, N is the number of cepstral coefficients(including the 0th coefficient), L is the liftering parameter. As for Outputs, CC is a matrix of mel frequency cepstral coefficients, (MFCCs) with feature vectors as columns, FBE is a matrix of filterbank energies with feature vectors as columns, FRAMES is a matrix of windowed frames (one frame per column).

MFCC are popular features extracted from speech recognition signals. The vocal tract frequency response is relatively smooth, whereas the source of voiced speech can be modelled as an impulse train. As a result, the vocal tract can be estimated by the spectral envelope of a speech segment. The motivating idea of mel frequency cepstral coefficients is to compress information about the vocal tract (smoothed spectrum) into a small number of coefficients based on an understanding of the cochlea. The basic steps are outlined by the diagram as shown in Figure 4.4 in HMM model. The triangular filterbank equations are given in (Huang et al., 2001). They require no energy normalization as the bandwidths (and area) are nearly equal by construction. Results of the audio signals based on different Quranic maqamat using CCA, MFCC and conventional MFCC with W-DFT are compared and evaluated in the next section. Figure 4.5 shows the overall process of enhanced MFCC by applying W-DFT in filterbank energies.

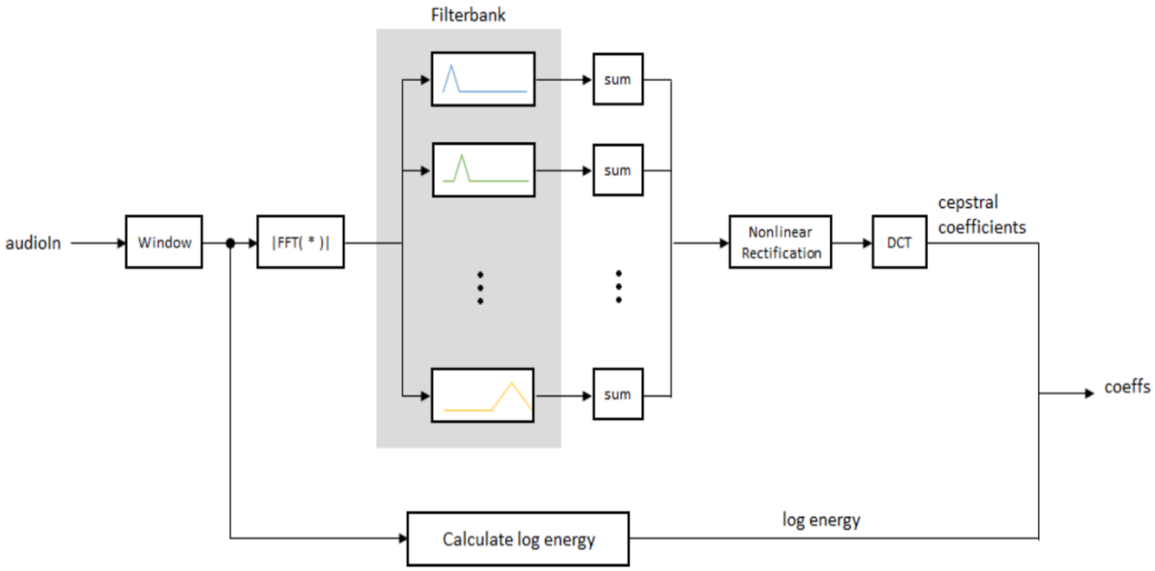

Figure 4.4 Feature extraction of MFCC in HMM model (Ellis, 2005)

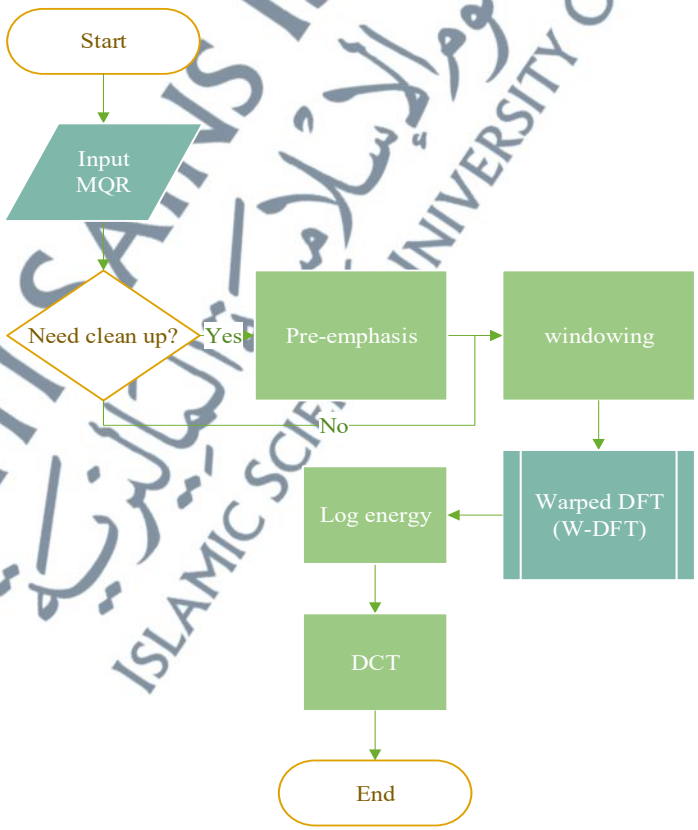

Figure 4.5 Flowchart of MFCC with W-DFT

4.3 Result & Discussions

In this section, results of the audio signals based on different Quranic maqamat using Spectral Descriptors and MFCC method with W-DFT are compared and evaluated. For MFCC, a set of 26 uniformly spaced filterbanks have been extracted within a dynamic range of 300-3700 Hz. All the simulation work has been done using MATLAB. This framework is based on Dan Ellis' RASTAMAT package. RASTA is a separate technique that applies a band-pass filter to the energy in each frequency sub band in order to smooth over short-term noise variations and to remove any constant offset resulting from static spectral coloration in the speech channel. The emphasis is placed on closely matching MFCCs produced by HTK with simplicity and compactness as main considerations, but at a cost of reduced flexibility. Meanwhile for Spectral Descriptors a straightforward MATLAB function is simulated on QMR audio files as input.

4.3.1 Spectral Descriptors

Here is example of captured SDs i.e., centroid, spread and roll-off point for Maqamat Bayyati of Surah Al-Ikhlas as shown in Figure 4.6. Then these three features are plotted on the same time-domain to find the correlation between the SDs. Figure 4.7 shows the spectral signal and spectrum of SDs in the same time domain. Observed the overlapping peaks of certain time event which highlighted in red. This shows that there are certain unique audio features correlate to each other. Next steps is to calculate the peak-magnitude-to-RMS ratio, mean and standard deviation and the overall performance will be presented in the next section.

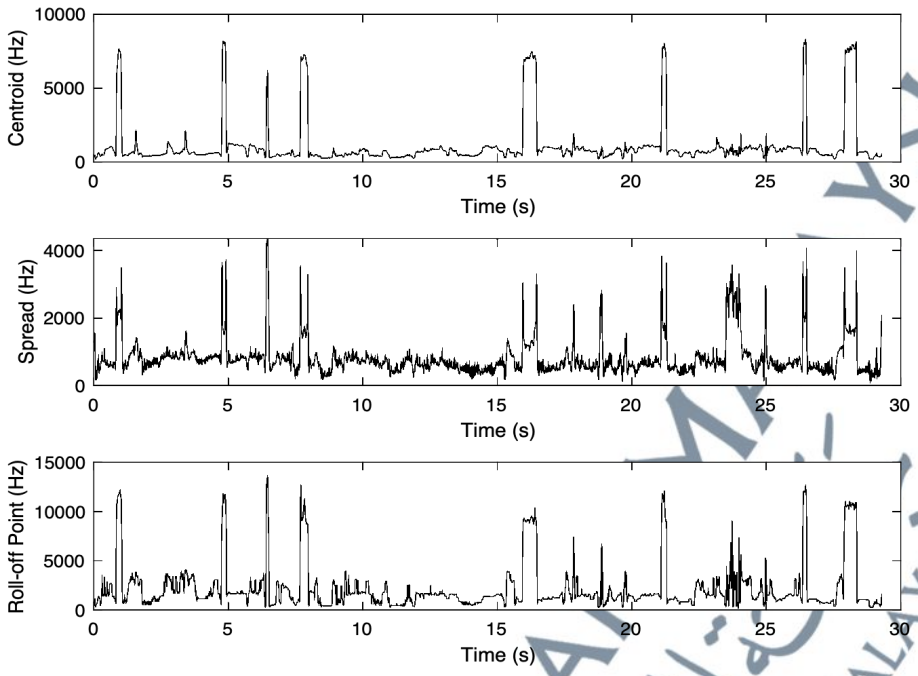

Figure 4.6 Example of plotted graph for SD for centroid, spread and roll-off point

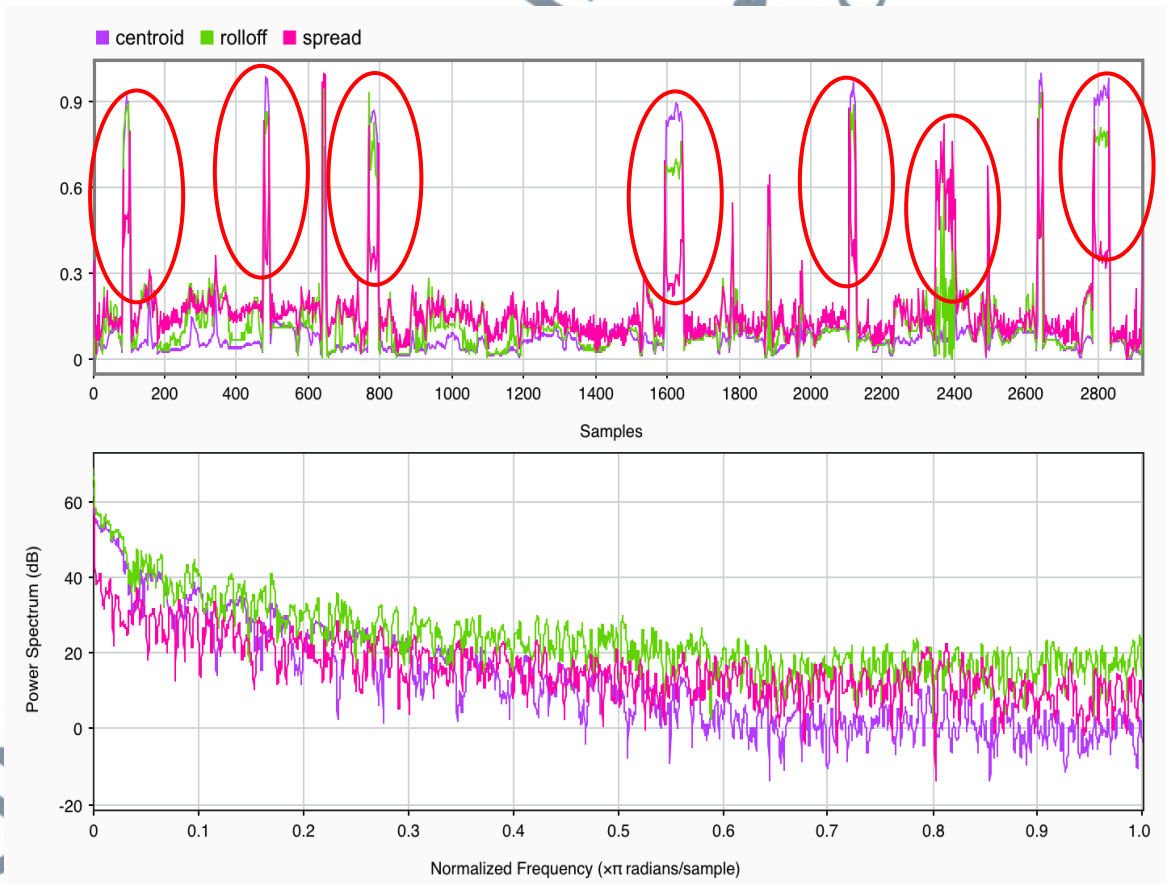

Figure 4.7 Example of plotted graph for the three SD in same time-domain

4.3.2 CCA, MFCC and W-DFT

In audio feature analysis, the cepstral coefficients of the maqamat speech signal are used as parameters to extract significant features of the maqamat content. Typically, in cepstral analysis, the signals are divided into frames and extract features from each frame. The overall performance for the three methods is shown in Fig. 4.8. Based on the signal pattern it can be observed that the formant frequencies using MFCC algorithm are higher than the ones using the cepstral coefficient method. It may be caused by the robustness of MFCC features in having the filterbank in its algorithm. CCA does not contain any filterbanks that makes the power signal less significant. Furthermore, higher accuracy of MFCC can be obtained with less number of filters used. Therefore, the computational complexity of the subsequent stages will be reduced.

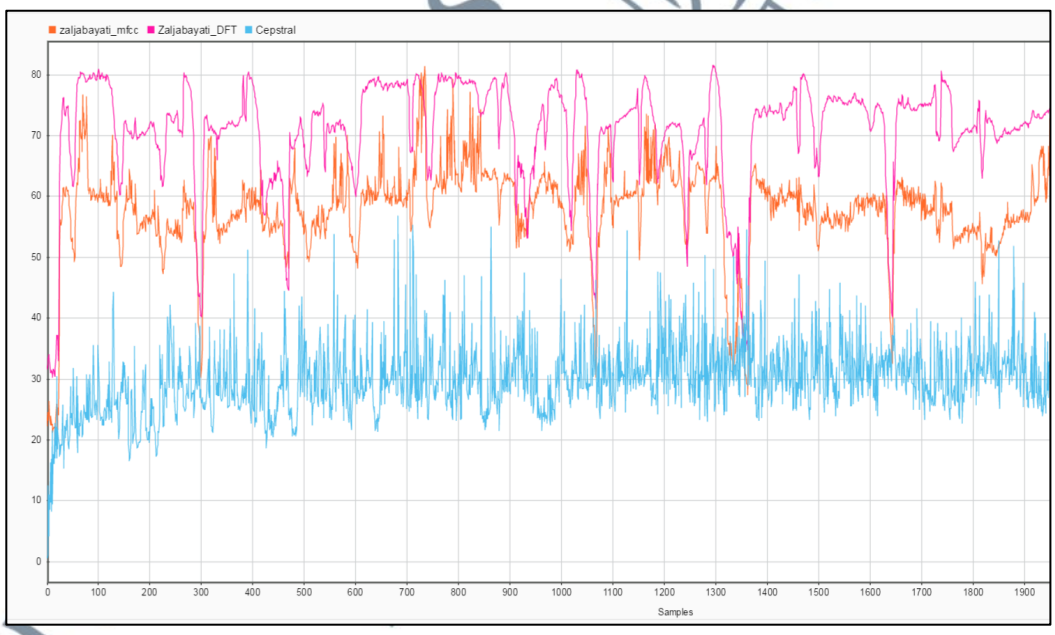

Figure 4.8 Sample of the overall performance for spectral envelope (formants) extracted using cepstral analysis (bottom graph), conventional MFCC (middle graph), and W-DFT (upper graph) for Maqamat Bayyati.

After applying warping function on MFCC it can be observed that W-DFT method outperformed the other two methods by removing most of harmonic peaks and contained more significant features, which is different from spectrum. It can be observed that the smooth curve connecting the formants is the cepstral envelope and the task of extracting MFCC coefficients from spectral details is to separate the envelope. The formants of a voice signal show the unique properties of a voice that makes the profiles for each maqamat. In other words, Quranic maqamat identification uses the concept of profiling the formant descriptions which are extracted from the cepstral envelope. For spectral analysis, this deduce that the enhanced MFCC method with W-DFT appears promising and effectively give better significant features and coefficient.

Based on the extraction features in Figure 4.9, the pattern similarities between each surahs' signals can be obtained and detected for each maqamat. These similarities are extracted based on matched frequencies which will be used to propose a set of standard profiling matrix for the rest of maqamat features.

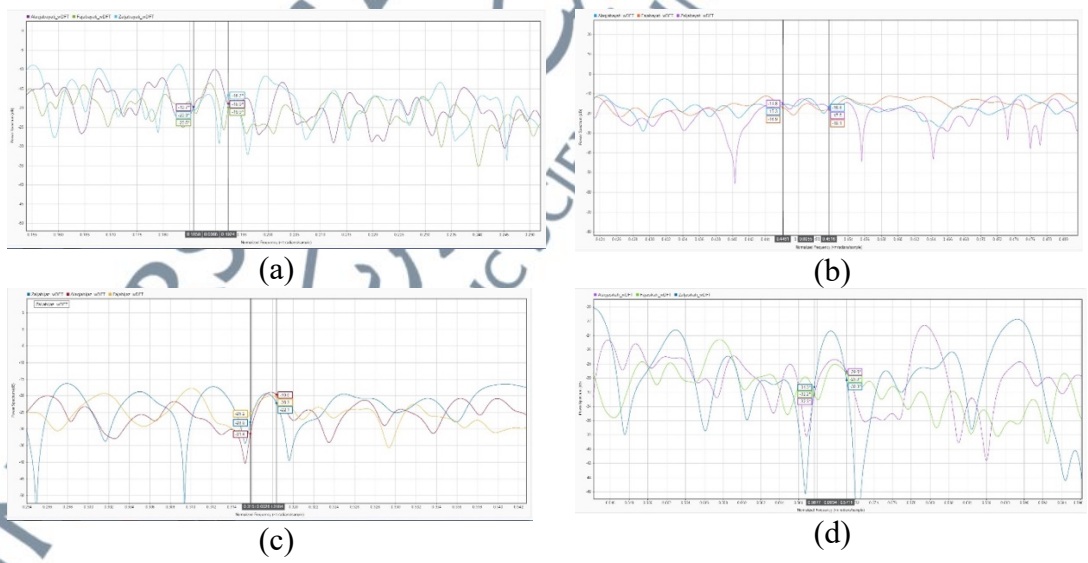

Figure 4.9 Example of frequency spectrum for 3 surahs using W-DFT for maqamat (a)Bayyati, (b)Hijaz, (c)Jiharkah and (d)Soba

4.3.3 Spectral envelope of W-DFT

The spectral envelope conveys sound quality and the time envelope is responsible for speech intelligibility (Mikio, 2020). It is the envelope curve of the amplitude spectrum which describes one point in time or in this case in one window. Having too many peaks in WDFT audio features, a smoothly varying estimate the highs-lows of WDFT features by taking the envelope function to connect maximum value of highs and lows detected over the time frames.

In this experiment, at least 10ms is considered between each extreme high and extreme low. The mean value extracted from the high and low will determined the mean spectral envelope as shown in example in Figure 4.10. The algorithm was simulated with initial frame duration, T_w = 25ms and a longer frame duration, T_w = 250ms, for comparison, as shown in Figure 4.11 and 4.12, respectively. The total time frame for both signals is remained 30s as changes is only effect on the number of frames and signal patterns. Based on the figures observed that the pattern was even smoother and significant for shorter time frame duration. The aim is to recognize the signal pattern significantly for each maqamat type based on their maqamat features and theme.

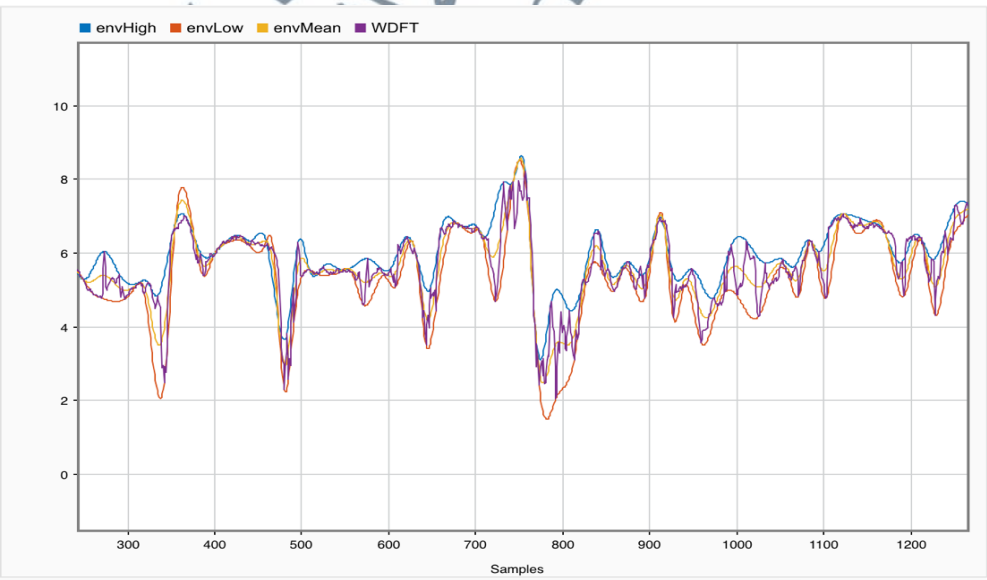

Figure 4.10 Spectral envelope of WDFT with high-low and extracted mean value

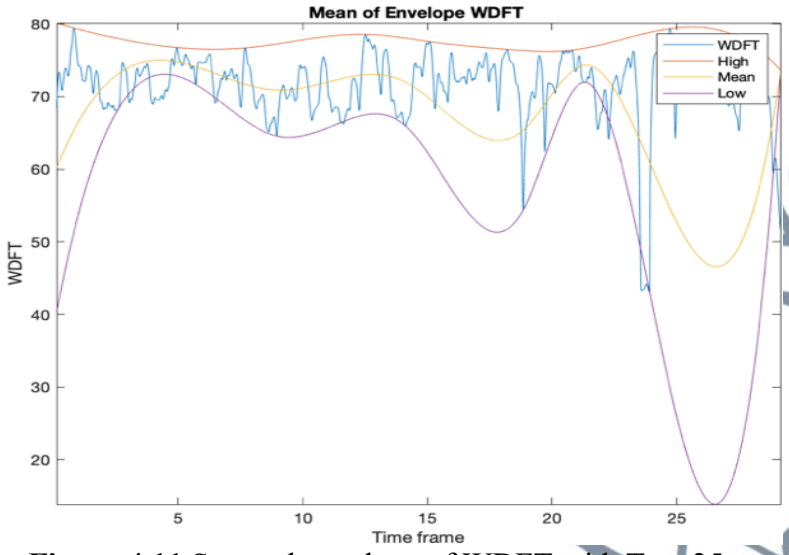

Figure 4.11 Spectral envelope of WDFT with Ts = 25ms

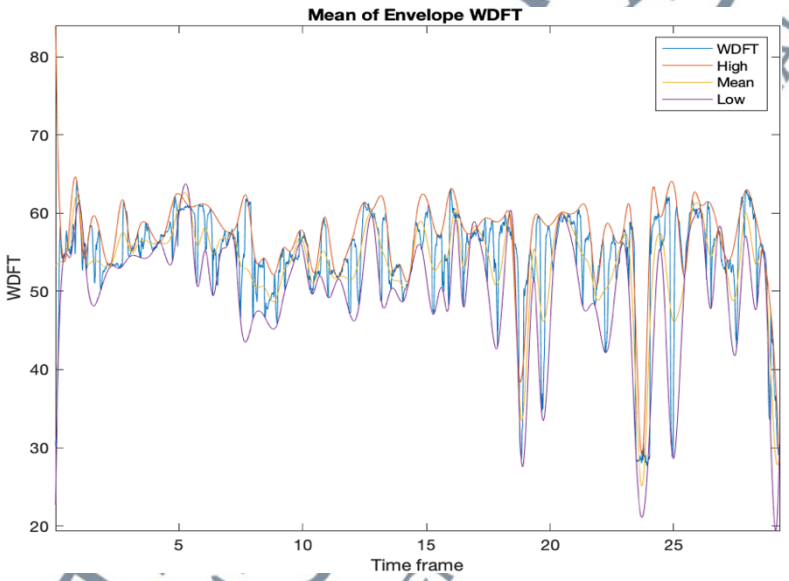

Figure 4.12 Spectral envelope of WDFT with Ts=250ms

The overall output for W-DFT is shown in Figure 4.13 for filterbank energy, mean spectral envelope and image spectrum. It clearly shows that the mel frequency cepstrum value of each frame. These images graphically depict the similarity between two-time regions in an audio file which is represented as a square. Here the dark region indicates the formats or peaks in the spectrum, in this region high amplitude is occurs. Meanwhile, the overall pattern of W-DFT features for each maqamat are shown in

Figure 4.14. However, these patterns only represent speech signals for designated *qari*. Each maqamat pattern defined certain acoustic elements that will be classified and analysed in next chapter.

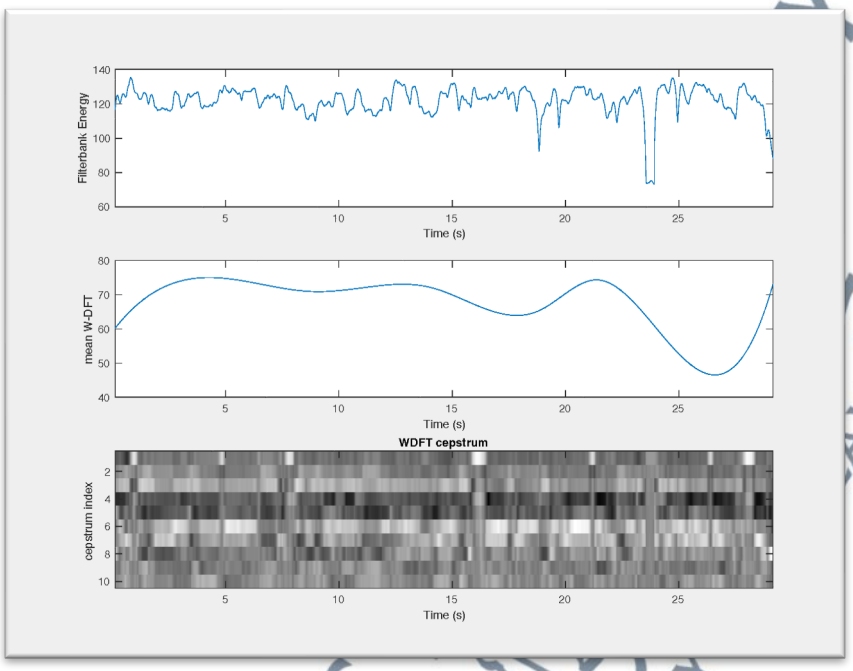

Figure 4.13 Filterbank energy, spectrum, and spectral envelope of WDFT.

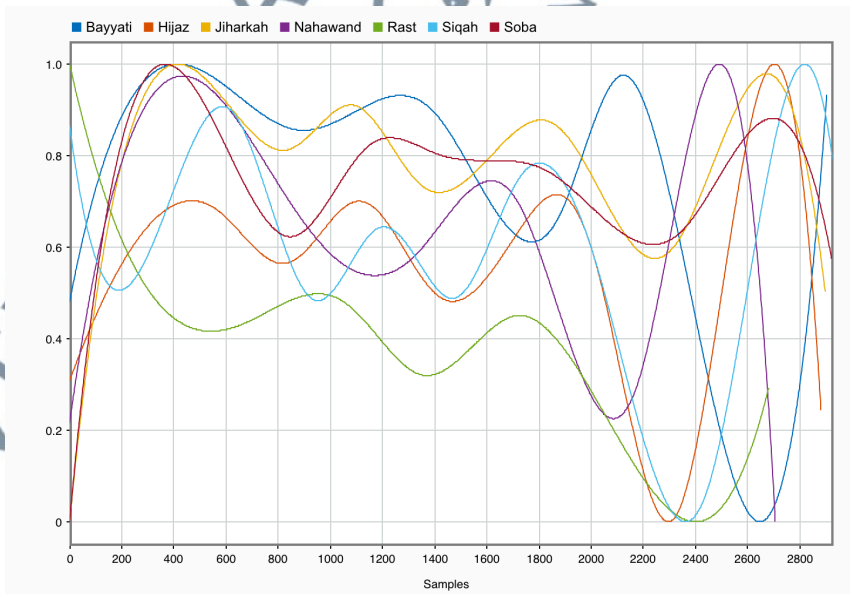

Figure 4.14 The overall W-DFT features for each maqamat

4.4 Summary

Audio feature extraction technique identify strong features that characterise the complex nature of audio signals especially in melodious speech analysis based on pitch detection. This chapter focused on feature construction and selection for acoustic profiles. Mel Frequency Cepstral Coefficient (MFCC) technique has been used and optimised to extract acoustic features in the complex speech signal of the seven selected Quranic maqamat recitations. An alternative of Mel-frequency warped feature representations using a DFT direct warping function is presented. MFCCs computed through directly warped spectrum showed improvement slightly over conventional MFCCs. Overall, alternative warping variants show some promising result and produce better acoustic features with the proposed spectral method. It is shown that the MFCC with W-DFT technique able to capture and extract the significant features of the complex speech signal as compared to conventional method of cepstral coefficient and MFCC. Three features of Spectral Descriptors i.e., Spectral centroid, Spectral Spread and Spectral Roll-off Point as a low-level signal for additional acoustic features. The overall output of both techniques has been analyzed and will be used for speech recognition process in the next chapter based on the power signal extracted from the formant frequencies of the spectral details for each maqamat. The acoustic features can be derived from the maqamat parameterization based on formant frequency.