CHAPTER 2

AUDIO FEATURE EXTRACTION IN SPEECH PROFILING FOR QURANIC SEMANTIC AUDIO

2.1. Introduction

Speech recognition has been widely studied for various type of languages, not only for English, but also for other languages including the Arabic recitations of the Holy Al-Quran. It is known that the recitation of Al-Quran is quite different from normal Arabic readings in terms of pronunciations and meanings. However without knowledge and proper learning of the Arabic language, misunderstanding or misinterpretation may occur when one reads and studies the Qur'an especially when the understanding is to derive from the context of the Qur'anic verses (Ramli et al., 2018). Recent study show that any attempt to comprehend the Quran using only linguistic meaning may lead to misreading and misinterpreting due to the high level of Arabic language of the Qur'an which consists of language properties, sound symbolism, linguistic forms, rhyme, word play, irony, and metaphors that exceed the standard Arabic language (Hedayat, 2013).

This chapter starts with an overview, past and present work of speech profiling. Then it provides a brief introduction of Quranic maqamat and sound element that relates to this work and later section described the basic concept of audio feature extraction and its application for complex spectrum. Then followed by techniques used in audio feature extraction in profiling the Quranic audio signal and ends with overview of semantic audio analysis for Quranic search.

2.2. Quranic Maqamat Recitation (QMR)

The Qur'an contains much scientific knowledge that ranges from basic arithmetic to the most complicated terms. On the other hand, QMR represents the emotions narrated by the reciters during reciting the Holy Quran. This section will elaborate the background, types and acoustic elements contained in QMR.

2.2.1. Overview of Quranic Maqamat

Reading or reciting the Qur'an is fundamental to all Muslims because the holy content contains comprehensive guidance to help daily routine of Muslims in all aspects of life. Therefore, the Muslims is urge to recite it with certain melody to beautify the voice in order to feel certain emotions during reciting them. In the study of maqamat, learning the Quranic recitation is beautifying the voice in reciting the Quran. It is originated from the technique of Arabic maqam which imposed from improvising the patterns, pitches and development of a musical art of Arabian culture. In recent study by Albakri (2020) shows the maqams as a center in Arabic Islamic musical culture. Some historical and academic references have been quoted, which are to prove the place of the maqams in the music of Islam and more specifically in the performance of the Holy Quran which meaningful accents have been pointed out over the tajweed and tartil in the performing art of the Holy Quran. (Albakri, 2020). Like these varied melodies, the verses of the Quran vary widely in term of topics and event to generate different feelings to the listener. Maqamat, plural from maqam, a set of pitches having characteristic of melodious elements, with a traditional pattern of their use.

2.2.2. Types of Maqamat

As previously mentioned, approximately 50 maqamat exist, only among the most widely used are the Maqamat of Bayyati, Hijaz Rast, Jiharkah, Nahawand, Soba, and Sikah. The maqam names originated from many Arabic and Turkish names, which mostly comes from throughout the Middle East in the formative period of Persian culture musical system (Nettl, 2007). Next section provides details description of each maqamat and its features and Table 2.1 summarizes the description based on features and suitable theme.

1. Maqam Bayyati

   Maqam Bayyati is the prime *maqam* in the recitation of maqamat Quranic. The function of this *maqam* served as an opening and closing of the rhythm in the Quranic recitation. In Malaysia currently, the *qaris* as well as the *qariahs* has popularised this *tarranum* during the recitation. They have placed the Maqamat Bayyati as the top *maqam* which acted as a stimulation during the recitation performance (M.Zaini et al. 2018). According to Drs. Muhsin Salim, the origin of the word *Bayyati* came from the Arabic word (بيت) which means a "house", subsequently used in a form of *mubalaghah* (بيات) and the letter ya (ي) was added, which then form the word *Bayyati (Lembaga* Bahasa Dan Ilmu Al-Quran,1987).

   In theory, the word maqam literally means as a "house", which is actually linked to several beliefs, referring *Bayyati* as a basis for any art form of Quranic rhythm. This is due to the fact that it has been used at the beginning and the end of each Quranic recitation. In addition, there are other theories which claimed that the word *Bayyati* itself sufficed from one of the place or area in Iraq. The letter *ya'ilah* was added to localize within

the area. This belief came from Mohd Ali b. Abu Bakar, which sourced from Dr. Sayid Agil Hussain al- Munawwar's theory, an expert in Quranic rhythm from Indonesian (A.Bakar,1997). However, both theories related to the origins of Bayyati cannot be officially recognized as accurate and valid. This is due to that the word is used as part of daily conversation of the locals in Malay Archipelago, not among Arabic community. (M.Zaini, 2009).

Bayyati has a wider scope in dividing and providing meaningful content of the Quranic recitation. This is because that *Bayyati* has twelve melodies style of reciting from four level of intonations known as *Qarar*, *Nawa, Jawab* and *Jawab al-Jawab*. Besides that, *Bayyati* can be combined with other melodies for tone similarity, in order to form a variation in Bayyati which is: *Syuri*, *Husainiy*, *'Ajam* and *Kurdi (*M.Munir, 2005). In general, *Bayyati* has a unique, soft, scarcely audible and low tone with a sharp pitch. Sometimes, these three tones; high or low and normal is used. By combining these two tones, this *maqam* is very flexible, easy acceptable and covering a much wider scope (Nik Jaafar, 1998).

2. Maqam Hijaz

The word *Hijaz* is referring to the one of the states in Arabic Territory. However, no other specific and detailed information regarding the responsible person who has provided the sources as well as the link of the name based on that particular place. However, there is a reliable source which stated the word *maqam* originated and evolved in Hijaz (Nik, 1998). Initially, based on the community who lived in the scarce land and desert, stated that Hijaz was introduced by their *qari*. Then, this melody was

brought to Egypt and localized in line with the greenery of Nil Valley (A.Bakar,1997).

The features of Hijaz are light and fast melody while the rhythm is quite strict, and the recitation is using a loud and clear voice. Apparently, it was changed to a softer and harmonious maqam called Hijaz *Misri.* As a result, the listener will feel more mesmerize with the beautiful melody of the tone. (M.Zaini, 2009). In addition, Hijaz can be combined with various type of maqamat which has similar in tone. This will lead to a variation which become part of Hijaz which is *Kard*, *Kurd*, *Kard Kurd* and *Nakriz (*M.Munir, 2005). There are several features of *Maqamat Hijaz*, such as reciting in a slow mode but very effective, strict and a hard rhythm. It can be adopted to any voice of *tabaqah* and used in command, assertiveness, and reminder verse effectively. (M.Zaini et al. 2018).

3. Maqam Jiharkah

Similar to Rast, Jiharkah also originated from Persian and has been modified by an expert in the field of Quranic art song from Hijaz and Egypt and blending in with their culture. This Maqam however was adapted by the *qaris* and *qariahs* from all over the world. There are other theories however, which claimed that it has originated from African Continent. (Nik, 1998). Jiharkah has a minor rhyme and rhythm. This *Maqam* depends on either the *qari* recites in slow or fast pace. Jiharkah only has one variation, namely *Kurdi* and it can be recited in two intonation which is *tabaqat Nawa* or *Jawab* and *Jawab al-Jawab* in four rhythm tones. *(*M.Munir, 2005). This *Maqam* has its own uniqueness, easy and fast, effective and suitable for a moderate *tabaqah.* Meanwhile, the roles

and functions for this *maqam* is to eliminate the tensions in any reciting, fluency in recitation, notably for sad and emotion sentences, provide an accuracy in pronouncing all letters, wordings or sentences as well as better in concentration and self-cautions. (A.Bakar,1997).

4. Maqam Nahawand

Maqam Nahawand is taken from the word Nahawand (نهاوند) which refers to a place in *Hamadan*, a district in Iran (Persian). Nahawand was adopted and modified by the Egypt *qari's* to make it more localize (Nik Jaafar, 1998). Nahawand has five types of rhythm *harakat* and two tone *tabaqah*, namely *Jawab* and *Jawab al-Jawab*. Maqamat Nahawand can be combined with four different types of songs which is Asli, *Nakriz*, *'usyaq* and *Murakkab (*M.Munir, 2005). The Nahawand rhythm is fast, light soft and harmonious tone, making it more appealing, mesmerizing and attractive rhyme. A high tone is a requirement should the *qaris* and *qariahs* intent to adopt this *maqam* in their recitations, making a moderate *tabaqah* change from lowest to the highest tone, making more vibration simultaneously as well as a good voice-control (M.Zaini, 2009). The features in Nahawand is considered as easy, soft and touching, suitable for a moderate *tabaqah*. There are several emotion's affects as a result from Nahawand which is more calming, focusing and awareness, make it more suitable to a happy and sad verse. (A.Bakar,1997).

5. Maqam Rast

There are several views in regard to the origins of Rast. The first view came from Dr. Sayid Agil Husain Munawwar who stated that the Rast was originated from a City called Rast (راس). Similar to the case of maqam *Hijaz* which eventually widely adopted, the letter *Ta* (ت) is added and become (راست) (A.Bakar,1997). For the second view which came from Haji Nik Ja'far b. Nik Ismail (1998), he stated that the Rast is sourced from the Persian's local or language itself whereby it was modified by the expert in Hijaz and Egypt art song based on their *lahjah* and culture. Nowadays, Rast is well-known among the *qaris* and *qariahs* all over the world. Unlike the above views, Dr. Muhsin Salim, an expert in a field of maqamat from Quranic Institution of Jakarta, in his opinion has stated that the letter Rast (راست) came from the Arabic word (ذا رست)or ((هذا رست which leads to the word *Rasydah* (رشدة) or Rast *(Lembaga* Bahasa Dan Ilmu Al-Quran,1987).

Rast is divided into seven types of *harakat* and two *tabaqat* intonation: *Jawab and Jawab al-Jawab*. Furthermore, Rast can be combined with three types of variation of maqamat which is *Usyaq*, *Zanjiran* and *Syabir 'Ala Rast* (M.Munir, 2005). According to Mohd Zaini (2009), the features of Rast should as consist of some characteristics such as an easy movement, fast and enthusiastic, can be applied with any *tabaqah* and suitable to be used in any types of maqamat. Basically, the functions of Rast are to provide the essence to the overall of maqamat and stimulate other maqamat which can be used in the future. To listeners, Rast can be comforting and give stimulation to our soul. It also helps the reciters in term of accuracy and fluency in a pronunciation as well as in pronouncing the letters.

6. Maqam Sikah

Maqam Sikah is a rare member of the Sikah family. Its scale starts with the root Jins Sikah on the tonic, followed by Jins Upper Rast on the 3rd degree (with its tonic on the 6th degree) then Jins Rast on the 6th degree (which is a secondary ghammaz). Similar to Jiharkah, maqam Sikah means "guitar strumming", which is also originated from Persian. This *Maqam* has been modified by the experts in the field of Quranic art song from Hijaz and Egypt according to their own culture. *Maqam* has made popularized by the expert's reciters around the world beginning from 7th century until 19th AD (Nik, 1998). A full concentration is required for any *qari* or *qariah* who wish to adopt this maqamat since it is difficult to perform in a perfect condition. Sikah can produce up to six types of *harakat* and two types of *tabaqah; Jawab* and *Jawab al-Jawab*. *Maqam* Sikah also can be combined with four different type of song variations such as *Asli*, *Turkiy*,*Raml* and *Iraqiy (*M.Munir, 2005). Among the features of this *maqam* is soft, graceful and harmonious, suitable for a higher *tabaqah*. Meanwhile the functions of Sikah is making the readings softer, more satisfaction to the reciters and listener, suitable for a desire, solace and reliance *verse* (A.Bakar,1997).

7. Maqam Soba

According to the expert in Quranic art song, *maqam* Soba is originated from one of the Syrian areas. This theory is supported by Dr. Sayid Agil Husain al- Munawwar. Whereby, Nik Ja'far b. Nik Ismail has claimed the it is possible that it has sourced from the Egypt melody, in which *Soba* is part of the *Bayyati* known as *Soba Mesir* (A.Bakar,1997). Soba is divided into five types of harakat's intonation and can be combined with three

types of melody's variations, among them are *'Ajami*, *Mahur or Muhur* and *Bastanjar (*M.Munir, 2005). This *Maqam* portrayed towards the sad feeling which described the desire for resolution and assistance. In comparison with Bayyati and Hijaz which has a high and low tones, Soba has a monotonous, high and fast rhythm. The uniqueness in Soba is in its harmonious, rhythmic and slow pitch. (Nik, 1998). Some of the features in Soba are soft and fast recitation, a lighter and rhythmic tones, suitable for a moderate *tabaqah.* Meanwhile, the role of *Soba* is to gain peace and release stress, thus a person can be more focused in life and realizing their mistakes, making it more appropriate for Quranic recitation which describe a self-happiness, sadness and tranquility. Therefore, the recitation will become more smoothly and fluently (A.Bakar,1997).

Table 2.1 Summary of maqamat types with its features and themes.

Maqam	**Features**	**Suitable Theme**
Bayyati	unique, soft, scarcely audible, and low tone with a sharp pitch.	Greetings, opening and closing remarks.
Hijaz	slow mode but very effective, strict and a hard rhythm	Command, assertiveness, and reminder.
Jiharkah	Soft, graceful, and harmonious	Awareness, happy and sad.
Nahawand	Easy, calm, soft and moderate.	Awareness, happy and sad.
Rast	Easy movement, fast and enthusiastic	All types.
Sikah	Soft, graceful, and harmonious	Desire, solace, and reliance
Soba	Harmonious, rhythmic, and slow pitch.	Self-happiness, sadness and tranquility

2.2.3. Characteristics and Sound element in QMR

The Quranic maqamat, or also named as '*tarannum'*, define a set of patterns of melody used by the reciters in their recitation adding in a proper articulation and rhythm *tajweed*. For instance, maqam Rast is said to evoke pride, power, soundness of mind, and masculinity, Bayyati relates with vitality and joy, Sikah more towards love, Saba is sadness and pain and Hijaz as reminder and assertiveness explained distant desert (M. Zaini, 2018). These emotions are said to be evoked in part through change in the size of an interval during a maqam presentation. Some references describe *maqam* moods using very vague and subjective terminology. However, there has not been any significant research using scientific methodology on a sample of reciters from different background or culture (whether Arab or non-Arab) to deduce the relationship between the emotions and the selected *maqam*.

There are several terms used to describe different attributes of sound – pitch, tune, and rhythm. For example, men, with deeper voices, tend generally to have lower-pitched voices than women. The art of maqamat comes from varying the pitch. If the pitch did not vary at all, then the whole recitation would be completely flat and monotonous. For the pitch pattern as a whole is also known tune. In the context of Quranic recitation, this is what the maqamat are – tune patterns – an identifiable pattern of pitches progressing from one to another. With the maqamat, the patterns are loosely defined and not fixed, and there is a lot of improvisation in recitation, but nonetheless they are still recognizable enough to be categorized as one maqam or another. Rhythm refers to the time component of sound, which is the domain of tajweed in terms of Quranic recitation. This is the reason *tajweed* defines the rhythm of the Quran. Because tajweed determines the rhythm and is part of the Qur'an, one cannot apply an external rhythm to their recitation. All of these elements produce

certain profiles to the audio signal captured from the recitation of Quranic maqamat and contribute significant finding in acoustics analysis for speech profiling.

2.3. Audio Feature Extraction (AFE)

Theoretically, it should be possible to recognize speech directly from the digitized waveform. However, because of the large variability of the speech signal, it is better to perform some feature extraction that would reduce that variability. The audio data provided cannot be utilized by the models directly so they need to convert them into an understandable format. Thus, a process called Audio Feature Extraction (AFE) is used. It is a process that explains most of the data but in an understandable way. Feature extraction is required for classification, prediction and recommendation algorithms. The reason for computing the short-term spectrum is that the cochlea of the human ear performs a quasi-frequency analysis. The analysis in the cochlea takes place on a nonlinear frequency scale (known as the Bark scale or the mel scale). This scale is approximately linear up to about 1000 Hz and is approximately logarithmic thereafter. So, in the feature extraction, it is very common to perform a frequency warping of the frequency axis after the spectral computation.

Different taxonomies exist for the classification of audio features. Scaringella (2006) followed a standard taxonomy by dividing audio features used for genre classification into three groups based on timbre, rhythm, and pitch information, whereas Weihsetal (2007) has categorized the audio features into another four subcategories, namely short-term features, long-term features, semantic features, and compositional features. This section describes basic feature extraction techniques for audio as shown in Figure 2.1 that are in use today, or that may be useful in the future, especially in the speech recognition area.

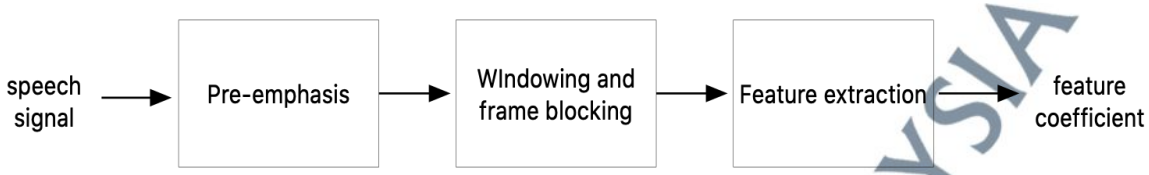

Figure 2.1: Basic concept of AFE

2.3.1. AFE for Complex Spectrum

The spectrum computation discussed in previous section is known as the *real* spectrum. As the real spectrum is computed from the log magnitude spectrum, the phase part is ignored. This will not enable the reconstruction of the sequence from the spectrum. However, the reconstruction can be done by preserving the Fourier phase and use it for reconstruction from the real spectrum. For the reconstruction of the sequence from the spectrum, *complex spectrum* is used. Instead of taking inverse Fourier transform of the log magnitude spectrum for the real spectrum, the Inverse Discrete Fourier Transform (IDFT) of the logarithm of complex spectrum is used for computing complex spectrum.

As the logarithm of all the spectral values are used, the phase is preserved in the complex cepstral sequence which can be used for reconstructing back the sequence. The methods for computing pitch and formant parameters from the complex spectrum remain same as that of the real spectrum as these parameters are obtained from the magnitude of the complex cepstral coefficients. The mathematical relation for computing complex spectrum in Eq. (2.1) as follow:

$$c_c(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\big(s(\omega)\big)\, e^{j\omega n} d\omega + j\frac{1}{2\pi}\int_{-\pi}^{\pi} s(\omega)\, e^{j\omega n} d\omega \qquad (2.1)$$

which also can be expressed as

$$c_c(n) = c_r(n) + jc_j(n) \tag{2.2}$$

where $c_r(n)$ is the real spectrum, and $jc_j(n)$ is the imaginary part of the complex spectrum.

2.3.2. Cepstral Analysis

The objective of cepstral analysis (CA) is to separate the speech into its source and system components without any a priori knowledge about source and/or system. The resulting of voiced and unvoiced speech can be considered as the convolution of respective excitation sequence and vocal tract filter characteristics. Based on the source filter theory of speech production, voiced sounds and unvoiced sounds are produced by exciting the time varying system characteristics with periodic impulse sequence and random noise sequence, respectively (Giacobello et al., 2012). The speech sequence $S(\omega)$, excitation sequence $E(\omega)$ and vocal tract filter $H(\omega)$ can be expressed and represented in frequency domain in Eq. (2.3) as follows:

$$S(\omega) = E(\omega) * H(\omega) \tag{2.3}$$

From the Eq. (2.3) the magnitude spectrum of given speech sequence can be represented as,

$$|S(\omega)| = |E(\omega)| * |H(\omega)| \tag{2.4}$$

To linearly combine the E(ω) and H(ω) in the frequency domain, logarithmic representation is used. Thus, the logarithmic representation of Eq. (2.4) will be,

$$log|S(\omega)| = log|E(\omega)| + log|H(\omega)| \tag{2.5}$$

$$c(n) = IDFT(log|S(\omega)|) = IDFT(log|E(\omega)| + log|H(\omega)|) \tag{2.6}$$

As indicated in Eq. (2.5), the magnitude speech spectrum is transformed by log operation. The log operation converted the multiplication of the speech spectrum ω into linearly summation of both excitation component and vocal tract component. IDFT is then used to separate the linearly combined log spectra of excitation and vocal tract system components. The IDFT of linear spectra transforms back to the time domain but the IDFT of log spectra transforms to inverse frequency domain or the cepstral domain which is similar to time domain. This is mathematically explained in Eq. (2.6). Figure 2.2 details the various steps involved in converting the given short-term speech signal to its cepstral domain representation. The output obtained at different stages of spectrum computation is the voiced frame considered and x(n) is the windowed frame. Here, s(n) multiplied by a hamming window to get x(n). |x(ω)| represent the spectrum of the windowed sequence x(n). As the spectrum of the given frame is symmetric, only one half of the spectral components is plotted. The log|x(ω)| represents the log magnitude spectrum obtained by taking logarithm of the |x(ω)|. c(n) represented the computed spectrum for the voiced frame s(n). The obtained spectrum contains vocal tract components which are linearly combined according to Eq. (2.6). As the spectrum is derived from the log magnitude of the linear spectrum, it is also symmetrical in the quefrency domain.

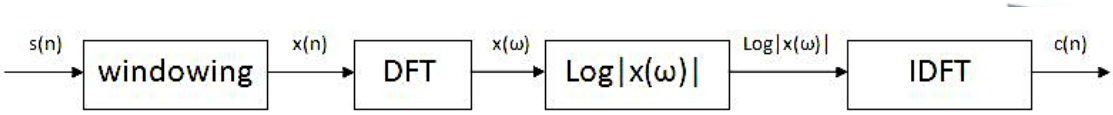

Figure 2.2 Block diagram representing computation of complex spectrum

2.3.3. Mel-Frequency Cepstral Coefficient (MFCC)

MFCC is considered amongst the widely used method in speech recognition due to its design that has been empirically determined to work well for speaker recognition (Reynold, 1994). The aim is to mimics the frequency response of the human hearing, the coefficients rely on a mel-frequency spacing of filterbank energies. The Mel-scaled feature extraction technique basically includes windowing the signal, applying the DFT, taking the log of the magnitude and then warping the frequencies on a Mel-scale. For final part, instead of using IDFT, it followed by applying the inverse Discrete Cosine Transform (IDCT) for better and faster processing due to its uses only cosine functions while DFT uses both cosine and sine (in the form of complex exponentials) while both operate on a function at a finite number of discrete data points. Next section are the steps involved in MFCC feature extraction.

1. Pre-Emphasis Filter

   The main goal of pre-emphasis filter is to emphasizes the high frequencies region that has been supressed during the sound mechanism production of vocal cord. It is able to amplify the significant formant frequencies that have important information in it. The most commonly used pre-emphasis filter is given by the following transfer function. In digitally speech waveform there has been occupied with highly range of additive noise.

Here pre-emphasis filter is used is applied to filter the additive noise. This is done by applying a first-order FIR high-pass filter.

In the time domain, with input $x[n]$ and $0.97 \leq a \leq 1.0$, the filter equation,

$$y[n] = x[n] - a.x[n-1] \tag{2.7}$$

and the transfer function of the FIR filter in z-domain is:

$$H(Z) = 1 - \alpha.z - 1, \quad 0.97 \leq \alpha \leq 1.0 \tag{2.8}$$

where α is the pre-emphasis parameter and coefficient a is adjusted according to time based on auto-correlation values of the audio signal. The aim of this stage is to amplify the amount of energy in high frequencies. The pre-emphasis filter is applied on the input signal prior next step of windowing.

2. Framing and Windowing

To avoid discontinuities and distortion in the underlying spectrum, the speech signal needs to be sliced in a short duration of time for better analysis. Therefore, the speech analysis have to be segmented in small segments called framing. The segmented signal is separated into several frames instead of analyzing the entire signal at once. For capturing accurate segments, the time window needs to be advanced every pre-selected time frame in millisecond (ms) to tracked the temporal characteristics of individual speech sounds. Slicing of pre-windowed time frame is usually sufficient to provide good spectral resolution of these sounds, and enough to resolve significant temporal characteristics. The overlapping analysis is

to centralized the input sequence of certain frame for each speech sound. The commonly used windows in mel analysis is Hanning or Hamming windows. These windows is applied on each frame a window to taper the signal towards the frame boundaries. This is performed to enhance the harmonics, smooth the edges and to reduce the edge effect while taking the DFT on the signal.

3. Discrete Fourier Transform (DFT)

   The input to the DFT is a windowed signal $x[n]...x[m]$, and the output, for each of N discrete frequency bands, is a complex number $X[k]$ representing the magnitude and phase of that frequency component in the original signal. Each windowed frame is converted into magnitude spectrum by applying DFT.

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{\frac{-j2\pi nk}{N}}; \quad 0 \le k \le N-1 \tag{2.9}$$

   where N is the number of points used to compute the DFT.

4. Mel-Scale Filterbank

   Mel refer to a unit measurement of human ears perceiving frequency. Mel-filterbank is a set of bandpass filter that use to filter the Fourier transformed signal in order to calculate the mel-spectrum value. Due to limitation of human ear does not perceive pitch linearly, the mel-spectrum does not correspond to the physical frequency of the tone linearly as well. The mel-scale is approximately a linear frequency spacing below 1kHz, and a logarithmic spacing above 1kHz. The approximation of mel from physical frequency can be expressed as

$$f_{mel} = 2595 \log\left(1 + \frac{f}{700}\right) \quad (2.10)$$

The mel-filterbanks can be implemented in both time domain and frequency domain. In order to mimic the human ears perception, warping the axis according to the non-linear function by applying the given function in Eq. (2.10). The triangular filter banks with mel frequency warping are displayed in Figure 2.3. The logarithm value is obtained by converting the values of DFT multiplication into an addition one. Mel Filter Bank values are reduced by replacing each value by its natural log. For MFCC, computing mel-filterbanks are commonly implemented in frequency domain. The center frequencies of the filters are normally evenly spaced on the frequency axis. The mel spectrum of the magnitude spectrum X(k) is computed by multiplying the magnitude spectrum by each of the of the triangular mel weighting filters. Using a logarithmic scale makes the feature estimates less sensitive to variations in input.

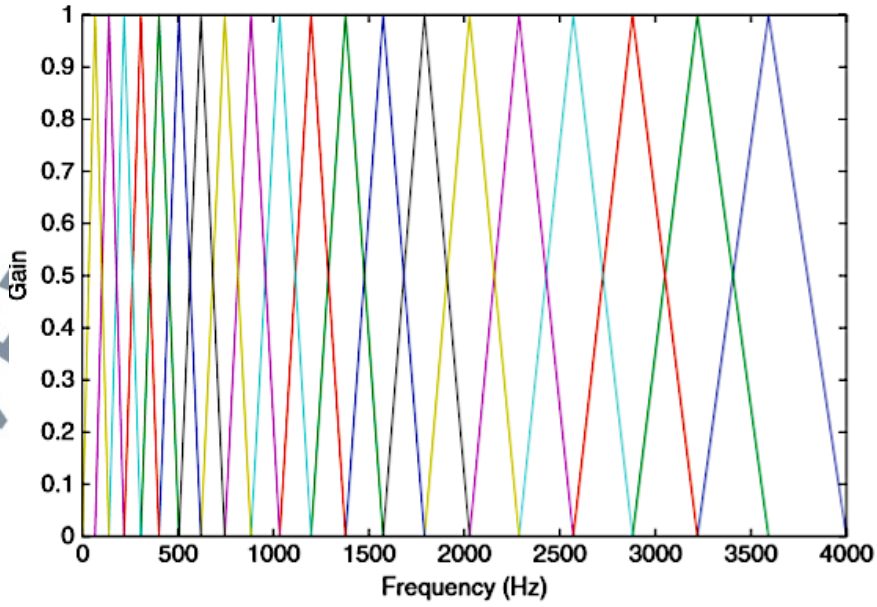

Figure 2.3. Mel-filter bank

$$s(m) = \sum_{k=0}^{N-1}[|X(k)|^2 H_m(k)]; \quad 0 \le m \le M-1 \qquad (2.11)$$

where M is total number of triangular mel weighting filters. $H_m(k)$ is the weight given to the kth energy spectrum bin contributing to the mth output band and is expressed as

$$\begin{cases} 0, & k < f(m-1) \\ \dfrac{2(k-f(m-1))}{f(m)-f(m-1)}, & f(m-1) \le k \le f(m) \\ \dfrac{2(f(m+1)-k)}{f(m+1)-f(m)}, & f(m) < k \le f(m+1) \\ 0, & k > f(m+1) \end{cases}$$

with m ranging from 0 to M−1.

5. Discrete Cosine Transform (DCT)

Since the vocal tract is smooth, the energy levels in adjacent bands tend to be correlated. The DCT is applied to the transformed mel frequency coefficients produces an array of cepstral coefficients. Prior to computing DCT the mel spectrum is usually represented on a log scale. This results in a signal in the cepstral domain with a que-frequency peak corresponding to the pitch of the signal and a number of formants representing low que-frequency peaks. Since most of the signal information is represented by the first few MFCC coefficients, the system can be made robust by extracting only those coefficients ignoring or truncating higher order DCT components. Frequency vector is applied with DCT using the equation below:

$$s(m) = \sum_{m=0}^{M-1} log\big(s(m)\big) \cos\left(\frac{\pi n(m-0.5)}{M}\right); \quad n = 0,12, \dots\dots, c-1 \qquad (2.12)$$

where c(n) are the cepstral coefficients and c is the number of MFCCs. Conventional MFCC systems use only 8–13 cepstral coefficients. The 0th coefficient is often excluded since it represents the average log-energy of the input signal, which only carries little speaker-specific information. Since the cepstral coefficients only contain information from a given frame, it commonly referred to as static features. To get the extra information about the temporal dynamics of the signal, it needs to compute the first and second derivatives of cepstral coefficients. These are known as delta coefficients, and delta-delta coefficients. Delta coefficients contains information about speech rate, and delta-delta coefficients provide information similar to acceleration of speech. The commonly used definition for computing dynamic parameter is

$$\Delta c_m(n) = \frac{\sum_{i=-T}^{T} k_i c_m(n+i)}{\sum_{i=-T}^{T} |i|} \qquad (2.13)$$

where $c_m(n)$ $d$enotes the mth feature for the nth time frame, k_i is the ith weight and T is the number of successive frames used for computation. In general, T is taken as 2. The delta-delta coefficients are computed by taking the first order derivative of the delta coefficients. Figure 2.4 summarizes the overall process of MFCC feature extraction that have been explained previously in details.

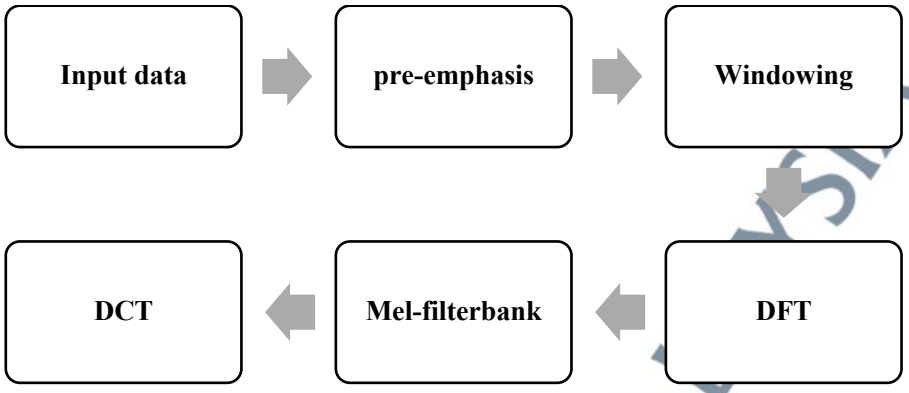

Figure 2.4 MFCC feature extraction

2.3.4. Frequency warping

In audio signal processing, frequency warping is a technique that are commonly used for spectral analysis. Its ability to mimic the human hearing due to its function as a non-uniform scaling of frequency. The basic concepts of frequency warping are by applying the unitary warping operator to the function. This warping procedure will drastically change the local density as well as the spectral density. Among well-known such warping a filterbank functions is the MFCC. This will be explained more details in chapter 4.

The warping function defines how the frequency components and frequency ranges are individual mapped on the new scale. There are several ways to create a frequency warped in such domain. Among them are frequency warping of the Fourier spectra, Fourier transforming the frequency warped time signal and non-uniform resolution filterbank. Warped frequency also defines the allocation of the new resolution which ranges in the original representation how they are compressed and expanded. The first method when combined to one operation, is equivalent to spectral domain processing with a nonuniform resolution filterbank made of warped (frequency domain) sine functions. The method is chosen depends on the actual

practical limitations, e.g., in the implementation of the warping process. Frequency warping is used as the main enhancement method for MFCC throughout the profiling system. Detailed method on how it will be implied on the frequency warping strategy by direct warping of the spectrum, rather than the filterbank will be discussed further in Chapter 5.

2.3.5. Spectral Descriptors

This section provides a set of functions that describe the timbre of audio known as spectral descriptors (SD). SD defines the equations used to determine the spectral features, common usage of each feature, and provides examples to describes the spectral descriptors more intuitively. SD are widely used in machine and deep learning applications, and perceptual analysis. Spectral descriptors have been applied to a range of applications, including speaker identification and recognition (Murthy, 1999), music genre classification (Li et al., 2005), mood recognition (Tsang, 2000) and voice activity detection (Scheirer et al.,1997).

1. Spectral Centroid

The spectral centroid is the frequency-weighted sum normalized by the unweighted sum (Peeters, 2004). The algorithm state that, μ_1, as the spectral centroid:

$$\mu_1 = \frac{\sum_{k=b_1}^{b_2} f_k s_k}{\sum_{k=b_1}^{b_2} s_k} \tag{2.14}$$

where f_k is the frequency in Hz corresponding to bin k, s_k is the spectral value at bin k and b_1 and b_2 are the band edges, in bins, to calculate the spectral centroid. The spectral centroid represents the central energy of the spectrum.

In audio analysis, it often uses for music or genre classification due to its function as brightness indicator. The spectral centroid is also commonly used to classify voiced or unvoiced speech. Figure 2.5 shows example of spectral centroid of human speech. Observed the centroid jumps in regions of unvoiced speech.

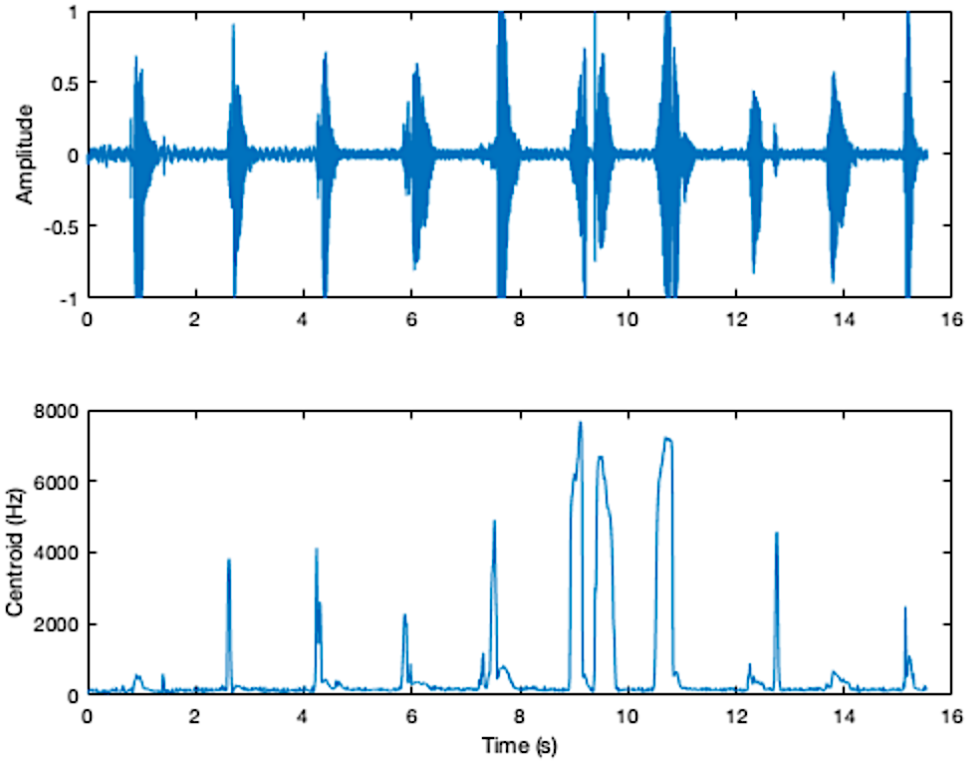

Figure 2.5 Example of spectral centroid from human speech.

2. Spectral Spread

Spectral spread is the standard deviation around the spectral centroid (Peeters, 2004). It is the second degree of the spectral centroid as stated below algorithm:

$$\mu_2 = \sqrt{\frac{\sum_{k=b_1}^{b_2}(f_{k-}\mu_1)^2 s_k}{\sum_{k=b_1}^{b_2} s_k}} \tag{2.15}$$

where, f_k is the frequency in Hz corresponding to bin k, s_k is the spectral value at bin k, b_1 and b_2 are the band edges, in bins, to calculate the spectral centroid and μ_1 is the spectral centroid. The spectral spread depicts the "instantaneous bandwidth" of the spectrum. It indicates the domination of a tone. For example, the spread increases as the tones diverge and decreases as the tones converge.

3. Spectral Roll-off Point

The spectral roll-off point measures the bandwidth of the audio signal by determining the frequency bin under which a given percentage of the total energy exists (Scheirer,1997).

$$\sum_{k=b_1}^{i} |s_k| = k \sum_{k=b_1}^{b_2} s_k \tag{2.16}$$

The spectral roll-off point has been used to distinguish between voiced and unvoiced speech, speech/music discrimination, music genre classification, acoustic scene recognition, and music mood classification.

2.4. Automatic Speech recognition for Speech Profiling

Automatic speech recognition (ASR) is the area of research that allows machines such as microphone or telephone to accept vocal input from humans and intelligently interpret them to the maximum degree of accuracy. It has a wide area of applications which makes life easier and very promising. This section will start with the overview of speech recognition over the decades and the developments throughout the years in ASR.

2.4.1. An overview of speech recognition

The study of speech recognition has been an area of research for more than fifty decades. The aim is to have the ability to capture, understand and react on the captured information. In Figure 2.6, Santosh classified that a speech recognition system includes four main stages which are further classified into subsystem (Santosh, 2010). These basic stages will be used extensively throughout this research. It involved analysis, extraction, modelling, and matching. Many models have developed by far, over the years, to produce an accurate system that can benefited by many (Pahwa et al., 2020). Speech recognition software use the various such as natural language processing. For instance, ASR system makes use of natural language processing techniques based on grammars (Reshamwala et al., 2013). It uses the context free grammars for representing syntax of that language presents a means of dealing with spontaneous through the spotlighting addition of automatic summarization including indexing, which extracts the gist of the speech transcriptions to deal with Information retrieval and dialogue system issues.

Communication has almost fully been using keyboards and screens, but speech is the most widely used, natural and the fastest means of communication for people. Many parameters affect the accuracy of the recognition system. These parameters are dependence or independence from speaker, discrete or continuous word recognition, vocabulary, environment, acoustic model, language model, and many more. Problems such as noisy environment, differentiating one word by two different speakers, incompatibility between train and test conditions led to made system without complete recognition.

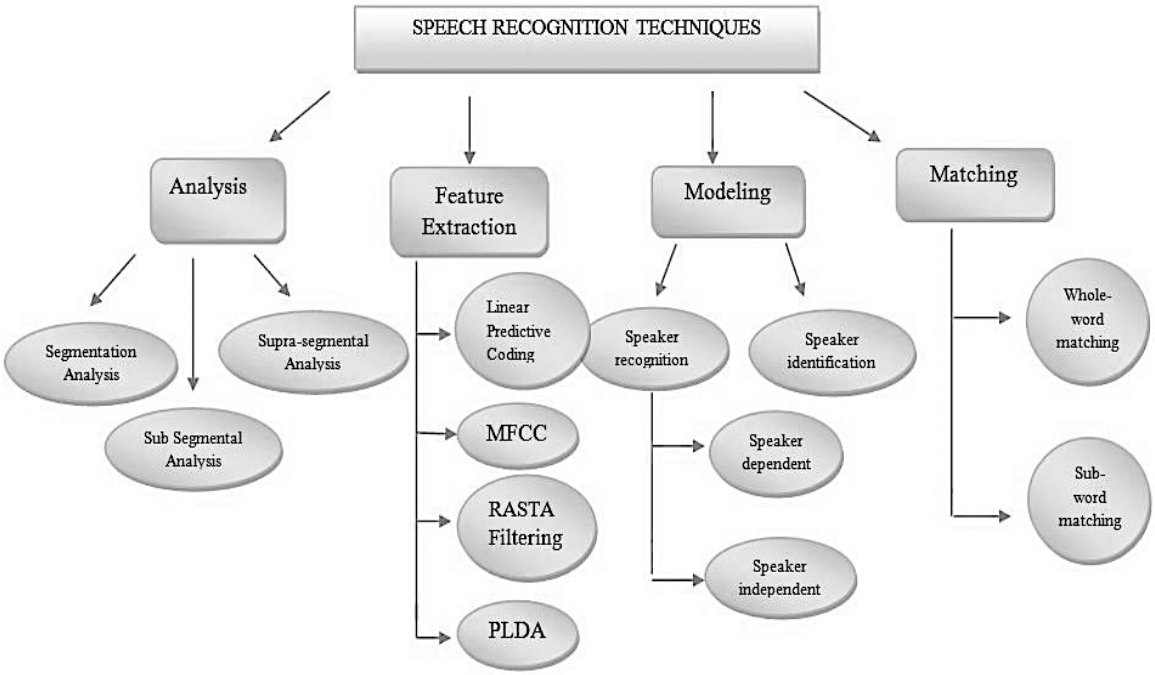

Figure 2.6 Four basic stages of Speech recognition (Santosh, 2010)

In this research, an open source model is used which is based on Hidden Markov Models. A Hidden Markov Model (HMM) is originated and created by Huang's team in 1990. For this, HMM toolkit is designed for speech recognition. Hidden Markov Model toolkit (HTK) is developed in 1989 by Steve Young at the Speech Vision and Robotics Group of the Cambridge University Engineering Department. HTK training tools are used to train HMMs using Training utterances from a speech corpus. HTK recognition tools are used to transcribe unknown utterances and to evaluate system performance. A method using Gaussian Mixture Model or statistical pattern classification is suggested to reduce computational load. This model is a statistical model where the system being modelled based on the assumption on Markov process with unknown parameters. The challenge is to determine the hidden parameters, from the observable parameters, based on this assumption. The extracted model parameters can then be used to perform further

analysis, for example for pattern recognition applications. Its extension from English as the standard into foreign languages, in this case of Arabic, its represent a real research challenge area.

2.4.2. ASR for Quranic Semantic Search

The research on Arabic ASR has focused on developing recognition system for modern standard Arabic. Since 2004, most issues faced in developing highly accurate ASRs for Arabic are the predominance of non-discretised text material, the enormous dialectal variety, and the morphological complexity (Hussein et al., 2021). The study uses a morphology-based language model at different stages in a speech recognition system for conversational Arabic and the automatic discretising Arabic text for use in acoustic model training for ASR. Quranic Arabic is the form of Arabic in which the Quran is written. With the similar language notation, The Quranic text is considered as predominance of non-discretised text material in Arabic language. Domain specific ontologies are created and inferred to one and another has leads researchers to Semantic search. Recently, the state-of-the-art ASR in the Arabic language comes from modular Hidden Markov Model Deep Neural Network systems (P.Smit et al., 2017). According to Hussein, there are various major challenges need to be faced in dealing with the language complexity. The best ASR results on the modern standard Arabic data were reported by the Aalto University team. To deal with morphological complexity in Arabic language, the character-level language model was suggested by (A.Ahmed, 2018).

2.4.3. Ontologies for Semantic Audio Analysis

An ontology describes problem entities, operations, relations and structures. In the context of semantic audio tools, the entities may be sounds or sound objects, while relations and structures are described by their organization. Operations describe the available tools and their context. A discussion on information management and knowledge representation requirements of these tools can be facilitated by a model for building semantic audio tools. Utilising the design principles, a set of ontologies has been developed for describing the process of audio recording. The proposed ontology detailed in this section is closely related to the information management framework for semantic audio tools outlined in the next section. It is designed to satisfy some of its requirements, for instance, the need for collection information about production, and uses the technologies deemed to be most appropriate for managing heterogeneous information in an open ended way. The proposed ontology allows for describing QMR production in more detail than what was possible using previously published ontologies.

2.4.4. Ontology design principles

In this section summarises the features which make the proposed ontology more suitable for QMR audio files. For instance, there are many audio feature ontologies exist in music world (Allik et al., 2016). All of these audio ontologies are based on music arrangement rather than dealing with acoustic features contained in frequency domain. These are the design principles and their advantages as follows:

1. Time frame and temporal entities can be used to localise events.
2. The proposed ontology is published as a modular ontology library whose components may be reused or extended outside of its framework.

3. Ease of use. The proposed ontology provides only the terms required for descriptive knowledge representation without more foundational elements.
4. Adaptation to existing and future applications in industry and academia.

The models provide the basis for content annotation as well as the decomposition of events in complex workflows. While elements of these models can also be found in other ontologies, they are not present all at once in a single unified framework. Thus, the design of proposed ontology will fill this gap. It provides a model to describe the production workflow from composition to delivery, including QMR recording, provided with very basic concepts to do so in detail. In the next following section will provide an overview and detailed of relevant audio feature extraction techniques related to the semantical analysis for knowledge base construction.

2.5. Semantic Audio Analysis (SAA) for Knowledge base construction

Since 1962, semantics is regarded as the study of meaning of human expression through language (Ullmann, 1962), whereas computer science studies regard semantics as a knowledge representation issue (Guarino and Giaretta, 1995). Semantic audio represent sound or feature that is meaningful or contained some information related to production of the audio. The motivation of this work is to design software systems that enable to support well-structured information for audio editing, and can facilitate data collection, and audio engineering knowledge base of QMR for future used.

Many applications that have been developed using semantic information to support the user in identifying, organizing, and exploring and manipulating audio signals. Speech recognition is an important SAA application. It includes language identification, speaker identification or gender identification. SAA involves the process of understand the audio information and incorporate machine learning, digital signal processing, speech processing, source separation, perceptual models of hearing, and ontologies.

2.5.1. Utilities and applications

The concept of semantic audio is designed in this work such that the technologies involved should enable the analysis of audio content in order for meaningful associations between the content and the acoustic elements are represent and manageable associations in a digital computer. Two crucial components of semantic audio applications are the capability of representing and structuring information of audio element, and the capability to analyse the association of these concepts with a representation of the recording. Extracting information from audio recordings is requisite for building semantic audio applications. It is important to review the basic categorical distinctions in audio features, and the relationships of these features in the physical, perceptual and audio domains. According to Olson's taxonomy of audio dimensions (Olson, 1952) provides an insightful parallel view on how the qualities of sound and audio are interpreted in various disciplines, and provides a basic terminology related to the concepts of physical and psychological qualities. Following this line of thought enables us to resolve ambiguities that often appear in relation to acoustical, perceptual and audio quantities.

In Table 2.2 illustrates physical quantities used to describe elementary sounds and the most related perceptual and audio concepts. It is crucial to see that if the basic physical quantities and concepts become more complex, and it becomes more difficult to establish the correspondence between categories. A sound may be classified by growth and decay, or the attack and release times related to timbre, harmonicity and inharmonicity, regular or irregular spacing of frequency components, frequency and amplitude modulation. Obtaining audio features corresponding to simple physical quantities, such as the fundamental frequency of a sound, is a question of measurement involving simple mathematical transformations.

In semantic audio applications, recognising more complex acoustic elements such as audio note or an instrument, will require more complex processing, such as pattern recognition and classification, or knowledge-based processing. All of these relies on logical inference using contextual information alongside directly measured physical quantities. There are underlying needs for the design of ontologies and some basic pre-cautions that should be addressed in designing the ontologies. information management solutions discussed in the following chapters.

Table 2.2 Principal dimensions of elementary audio sounds (Olson, 1952)

Physical Quantity (or concept)	**Perceptual Quality**	**Musical Category**
Frequency (fundamental)	Perceived pitch	Musical note
Amplitude (intensity)	Perceived loudness	Dynamics
Duration (time)	Perceived duration	Beat and tempo
Waveform (or complex spectrum)	Perceived timbre	Tone quality

2.5.2. Semantic Audio Tool

A system for integrating components that allow the implementation of the ideas mentioned so far in this work may be modelled as shown in Figure 2.8. This model has three analysis layers corresponding to audio feature extractors, three information layers corresponding to ontologies for describing tools and the results of audio analysis, and three application layers corresponding to tools that can be built using this information and their descriptions. In the following, we outline the role of the three layers and the components that may be utilised in the model.

1. Analysis layers (Audio feature extraction)

   As mentioned that AFE is the main technology used in this layer. This steps is very crucial which would give the biggest impact to the next layer. Some of the signal processing components required for extracting information from audio content are well researched and may be adapted. Basic feature extraction techniques standardised in speech recognition are successfully applied to audio. For example, segmentation audio semantically had done many research this past years (Theodorou, 2014, Aggarwal, 2022). Many techniques of DSPs techniques applicable to these problems are describe in (Aggarwal, 2022). For instance, high and mid-level feature extraction are the focus of MIR research. High-level segmentation of audio recordings played by a single instrument and the analysis of master recordings however were not considered by previous research.

2. Information layers (Audio Features Ontology)

   The audio ontology and its extensions is defined as frames of reference for describing the domain. It can be used to give a reference point for the

information management layer considered here. This ontology is also considered to be useful to represent for instance, an audio elements and its relation to corresponding signals. Its basic components allow for associating entities with the event occurred time-based domain. This is crucial in representing audio features, and serve as the basis for the Audio Features Ontology and pre-designed Maqamat Ontology in the next chapter. It also provide the relation between features, which considered as important elements for efficient feature extraction. Ontologies for describing audio analysis algorithms, ideally, including even their low-level digitally components, and ontologies that allow for describing audio processing tools are equally important in building intelligent audio processing environments. Currently there are no ontologies describing specific signal models or performance related data. These shall be developed in the future as the need arises.

3. Application layers (Interaction and navigation)

   A well-defined and structured representation of the audio and its various representations is the key for development a semantic audio tool. The ontological needs of describing applications include the ability to create a knowledge base. This information is very useful to retrieve data, feature extraction if needed, or ask for user interaction. Figure 2.7 and 2.8 shows the knowledge representation model between analysis, information and application layers (Fazekas, 2012).

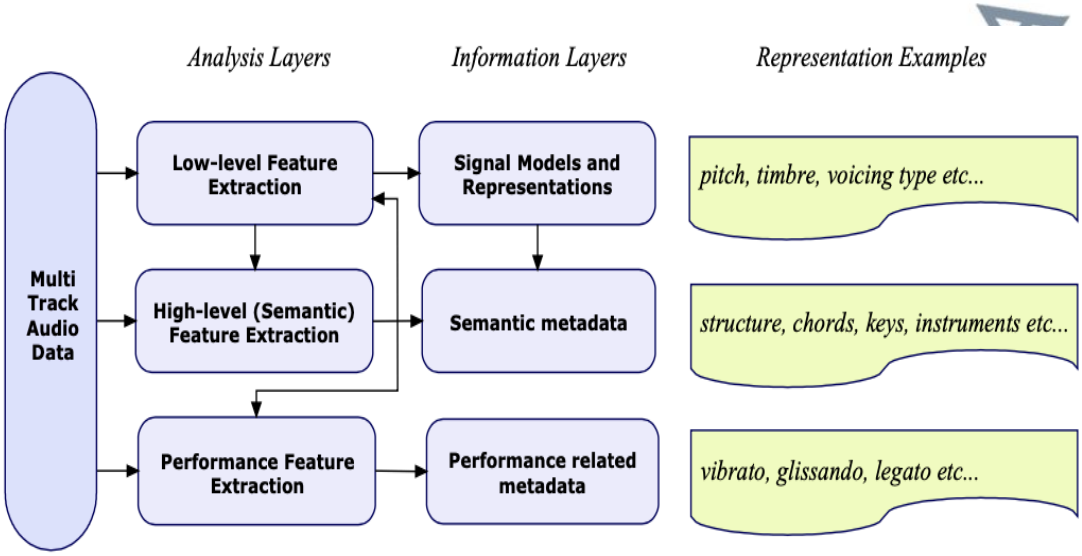

Figure 2.7: Knowledge representation model for analysis and information layers (Fazekas, 2012)

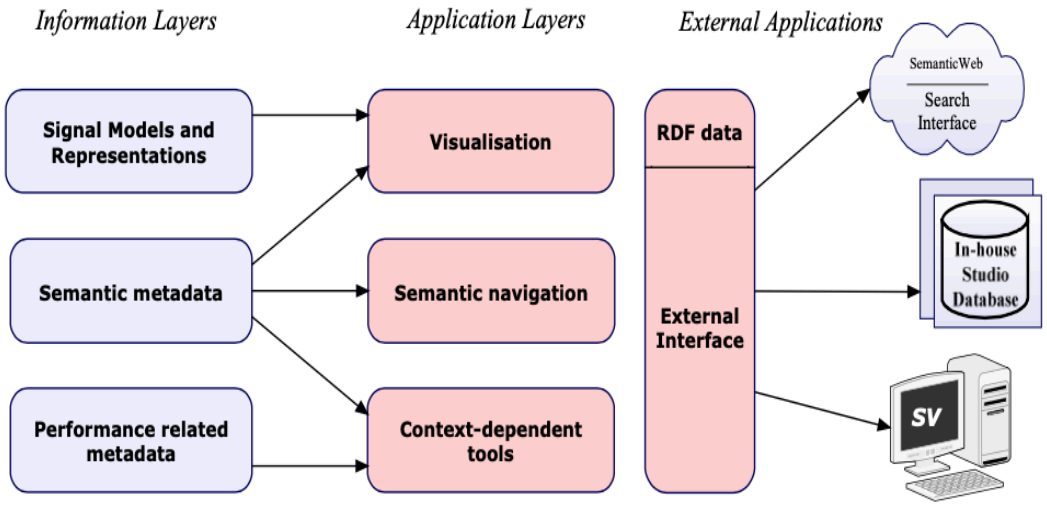

Figure 2.8: Knowledge representation model for information and application layers (Fazekas, 2012)

2.6. Summary

Speech recognition has been widely studied for various type of languages, including the Quranic recitation which contained acoustics features. These properties contained uniquely in formants which define as a concentration of acoustic energy around a particular frequency that corresponds to a resonance in the vocal tract. Like these varied melodies, the verses of the Quran vary widely in term of topics and event to generate different feelings to the listener. These acoustic features contained in the QMR are considered as a complex speech signal that need to be extract and analyse for the purpose of understanding the characteristics of the signal components.

This chapter provides a brief overview of Quranic maqamat and sound element that relates to this work. The acoustic features contained in the Quranic recitation are considered as a complex speech signal that need to be extracted and analyzed. These techniques are crucial for understanding the characteristics of phonological and morphological elements from QMR audio files to determine any correlation between acoustics properties and Quranic rhetoric elements contained in QMR audio features. Then later section described the basic concept of speech recognition, AFE and its application for complex spectrum and followed by cepstral analysis in complex spectrum. For complex spectrum, two types of algorithms will be presented in cepstral analysis method. Those are the typical cepstral analysis with warping frequency function and mel-scaled features, which also known as MFCC. The section also outlines techniques used in audio feature extraction in profiling the Quranic audio signal. The final part briefly explains the semantic analysis and its utilities and tool in audio for the development of well-structured database system that support studio environment which facilitate Quranic semantic audio search based on ontological audio features.