

## ABSTRACT

### ANALYSIS OF AUDIO FINGERPRINTING TECHNIQUES

SATISH KUMAR SIVA SANKARAN, MS  
Department of Electrical Engineering  
Northern Illinois University, 2017  
Lichuan Liu, Director

The goal of this thesis is to compare various audio fingerprinting algorithms under a common framework. An audio fingerprint is a compact content-based signature that uniquely summarizes an audio recording. In this thesis, acoustic fingerprints are based on prominent peaks extracted from the spectrogram of the audio signal in question. A spectrogram is a visual representation of the spectrum of frequencies in an audio signal as it varies with time. Some of the applications of audio fingerprinting include but are not limited to music identification, advertisement detection, channel identification in TV and radio broadcasts. Currently, there are several fingerprinting techniques that employ different fingerprinting algorithms. However, there is no study or concrete proof that suggests one algorithm is better in comparison with the other algorithms. In this thesis, some of the feasible techniques employed in audio fingerprint extraction such as Same-Band Frequency analysis, Cross-Band Frequency analysis, use of Mel Frequency Banks, and use of Mel Frequency Cepstral Coefficients (MFCC) are analyzed and compared under the same framework.

NORTHERN ILLINOIS UNIVERSITY  
DE KALB, ILLINOIS

MAY 2017

ANALYSIS OF AUDIO FINGERPRINTING TECHNIQUES

BY

SATISH KUMAR SIVA SANKARAN  
©2017 Satish Kumar Siva Sankaran

A THESIS SUBMITTED TO THE GRADUATE SCHOOL  
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE DEGREE  
MASTER OF SCIENCE

DEPARTMENT OF ELECTRICAL ENGINEERING

Thesis Director:  
Lichuan Liu  
Thesis Co - Advisor:  
Benedito Fonseca

## ACKNOWLEDGEMENTS

It is an immense pleasure to express my deepest gratitude to my thesis and academic advisor Dr. Benedito Fonseca for his thoughtful inspirations, prompt suggestions and generous attitude. His deep knowledge in the field of audio fingerprinting, enthusiasm towards this thesis and above all his dynamism triggered me to complete my thesis work in short time. I would also like to express my hearty gratitude to my thesis director, Dr. Lichuan Liu for her outgoing and liberal personality. Her warm support from the start of my graduate study has been a factor of encouragement to complete this thesis.

I am highly indebted to Dr. Donald Zinger, the program director at Northern Illinois University's department of Electrical Engineering for patiently listening to my goals and advising me on the best possible options for my graduate study. I would like to thank my entire committee for their valuable time and insights towards my thesis.

I would like to extend my thanks to my parents Mr. Siva Sankaran and Mrs. Pushpa Siva Sankaran for their love and support and not the least of all, the love of my life, Janani Jayabalakrishnan. Her motivation and support during pinch situations, encouraged me to complete my graduate studies in time.

## DEDICATION

*To my parents Mr. & Mrs. Siva Sanakran and to my girl friend Ms. Janani  
Jayabalakrishnan...*

# TABLE OF CONTENTS

	Page
LIST OF TABLES . . . . .	viii
LIST OF FIGURES. . . . .	ix
1 INTRODUCTION. . . . .	1
1.1 MOTIVATION. . . . .	1
1.2 OVERVIEW OF AUDIO FINGERPRINTING. . . . .	3
1.2.1 Filter Bank. . . . .	5
1.2.2 Mel Frequency Cepstral Coefficients (MFCC). . . . .	5
1.2.3 Peak Based Methods. . . . .	5
1.3 ISSUES IN CURRENT SYSTEMS. . . . .	6
1.4 THESIS STATEMENT. . . . .	8
1.5 SUMMARY OF CONTRIBUTION . . . . .	8
1.6 ORGANIZATION OF THE THESIS . . . . .	9
2 LITERATURE REVIEW. . . . .	11
2.1 GENERAL PRINCIPLES OF AUDIO FINGERPRINT EX- TRACTION AND MATCHING: . . . . .	11

Chapter	Page
2.2 SPECTROGRAM . . . . .	13
2.3 MEL FREQUENCY CEPSTRAL COEFFICIENTS (MFCC) . .	14
2.4 SAME - FREQUENCY BAND FEATURE EXTRACTION. . . .	17
2.5 CROSS - FREQUENCY BAND FEATURE EXTRACTION. . .	18
3 ISSUES CONTRIBUTING TO THESIS . . . . .	20
3.1 PROBLEM 1: NONEXISTENT STANDARD IN COMPAR- ING CURRENT APPROACHES . . . . .	20
3.2 PROBLEM 2: MFCC APPROACH TO MUSIC RETRIEVAL .	21
4 OVERVIEW OF EVALUATION PROCEDURE . . . . .	23
4.1 FRAMING AND SPECTROGRAM GENERATION. . . . .	23
4.2 BASELINE APPROACH WITHOUT THE USE OF FILTER BANK . . . . .	25
4.2.1 DATABASE GENERATION . . . . .	25
4.2.2 QUERYING AND SCORING . . . . .	26
5 PEAK BASED APPROACH. . . . .	30
5.1 FILTER BANK GENERATION . . . . .	30
5.2 NEED FOR ONE FRAMEWORK . . . . .	32
5.2.1 Double Kill Scoring Process. . . . .	33
6 EMPLOYING MFCC FOR AUDIO FINGERPRINTING . . . . .	38
6.1 STEPS INVOLVED IN MFCC GENERATION . . . . .	38
6.2 FILTER BANK VS MFC COEFFICIENTS . . . . .	39
7 CONCLUSIONS AND FUTURE WORKS . . . . .	43
7.1 CONCLUSIONS . . . . .	43
7.2 FUTURE WORKS . . . . .	44
REFERENCES . . . . .	46

	Page
APPENDIX A .....	49
APPENDIX B.....	53

## LIST OF TABLES

5.1	Performance analysis of traditional approaches against the new approach. . .	35
-----	--	----



## LIST OF FIGURES

1.1	Audio fingerprint generation and database setup . . . . .	3
1.2	Scoring process and audio identification using signatures commonly used in music identification application. . . . .	4
2.1	General Principles Of Audio Fingerprinting . . . . .	12
2.2	Spectrogram of a digital audio signal . . . . .	13
2.3	Relationship between Mel scale and Linear frequency scale . . . . .	15
4.1	Time Domain Audio Signal . . . . .	23
4.2	Framing a segment of the audio signal. . . . .	24
4.3	Spectrogram of the audio signal . . . . .	25
4.4	Spectrogram with threshold of 0.5 . . . . .	26
4.5	Fingerprint database setup. . . . .	27
4.6	Fingerprint Scoring . . . . .	28
4.7	Audio signature scoring results. . . . .	29
5.1	Relation between Mel Frequency Vs Linear Frequency in thesis. . . . .	31
5.2	Mel scale filter bank . . . . .	32
5.3	Results of single frequency band audio fingerprinting . . . . .	34
5.4	Results of cross frequency band audio fingerprinting for seven bands . . . . .	35
5.5	Results of cross frequency band audio fingerprinting for all fourteen bands . . . . .	36
5.6	Performance of different approaches at -10 dB SNR . . . . .	36
6.1	Result of using Euclidean distance scoring process on filter bank . . . . .	41

Figure		Page
6.2	Result of using Euclidean distance scoring process on MFCC . . . . .	42
1	Database Generation . . . . .	49
2	Peak Based Approach Scoring Process . . . . .	49
3	MFCC Based Approach Scoring Process . . . . .	50
4	Noise Cancellation In Double Kill Process . . . . .	50
5	Fingerprint Scoring . . . . .	51
6	Time Expense For Querying Using Same-Frequency Band Approach In This Thesis . . . . .	53
7	Time Expense For Querying Using 7 Cross-Frequency Band Approach In Thesis . . . . .	54
8	Time Expense For Querying Using All Cross-Frequency Band Approach In Thesis . . . . .	54
9	Time Expense For Querying Using 7 Cross-Frequency Band Approach In Thesis . . . . .	55
10	Time Expense For Querying Using All Cross-Frequency Band Approach In Thesis . . . . .	55

# CHAPTER 1

## INTRODUCTION

Most audio retrieval algorithms are fundamentally based on the extraction of salient features within the audio files, a technique called acoustic fingerprinting. The condensed summary or audio fingerprint consists of frequency features of the audio clip. These features do not change or undergo minimal change when the audio clip is hampered with noise. A good fingerprinting algorithm should take into consideration the perceptual characteristics of the audio clip which distinguish it from other audio clips. Some of these perceptual characteristics are spectral flatness, average spectrum, prominent tones across a set of frequency bands, zero crossing rate, and bandwidth. Prominent tones across frequency bands (same-band and cross-band frequency analysis) is the characteristic that is considered to define audio fingerprints in the hypothesis below. Frequency analysis is of much interest when it comes to acoustic fingerprints as researchers believe they are more robust, scalable, and retain features that prominently define the signal. Simply put, frequency characteristics retain the uniqueness of a signal.

### 1.1 MOTIVATION

A fingerprint is a unique identity to humans and the same theory can be applied to audio or video files. Audio fingerprinting is a concept of generating unique signatures for audio files. Currently there are more than one approach to achieve this goal. Consider being provided with several options to choose from among the multiple approaches to audio fingerprinting.

Which one should an engineer choose when provided with some of these said approaches? On what grounds do they choose one approach over the other? What if there was a novel method to bring out a comparison among the most commonly considered approaches? This here is one such attempt to provide an even ground to compare such approaches. To provide one such standard, a closed set of audio files are considered in this approach. The audio signals considered for the querying process are chosen from among the 60 audio signals in the database in this thesis. Hence the name closed set.

Some researchers may argue that closed set may be inadequate to the traditional application of music identification, but is important in many other applications such as:

#### **1. ADVERTISEMENT DETECTION:**

Consider a network provider such as NETFLIX, HULU, etc., who offer Network DVR, that re-telecasts saved shows over different regions of the world. The network provider needs an automated system that can track the position of the advertisements within the entire duration of the shows to change the advertisements from among its different sponsors that relates to viewers of each region. For this, the network provider would then employ a closed set audio fingerprinting system.

#### **2. CHANNEL DETECTION:**

Assume a user is watching a documentary on an exotic place to travel or a product that he hears over the radio and it intrigues him. An application in his smart phone or tablet could detect the channel in which the documentary is being aired in order to obtain information on the same.

### 3. COPYRIGHT ENFORCEMENT:

Audio signal feature extraction can help in meta-data cleanup by attaching the proper artist, album and track name to every track in a music collection. This will help music distributors, such as, bulk CD copying companies, can ensure they are not unknowingly duplicating audio for which a customer does not have a license to copy. Audio fingerprinting can also help sites like YouTube keep copyright violations out of their site.

## 1.2 OVERVIEW OF AUDIO FINGERPRINTING

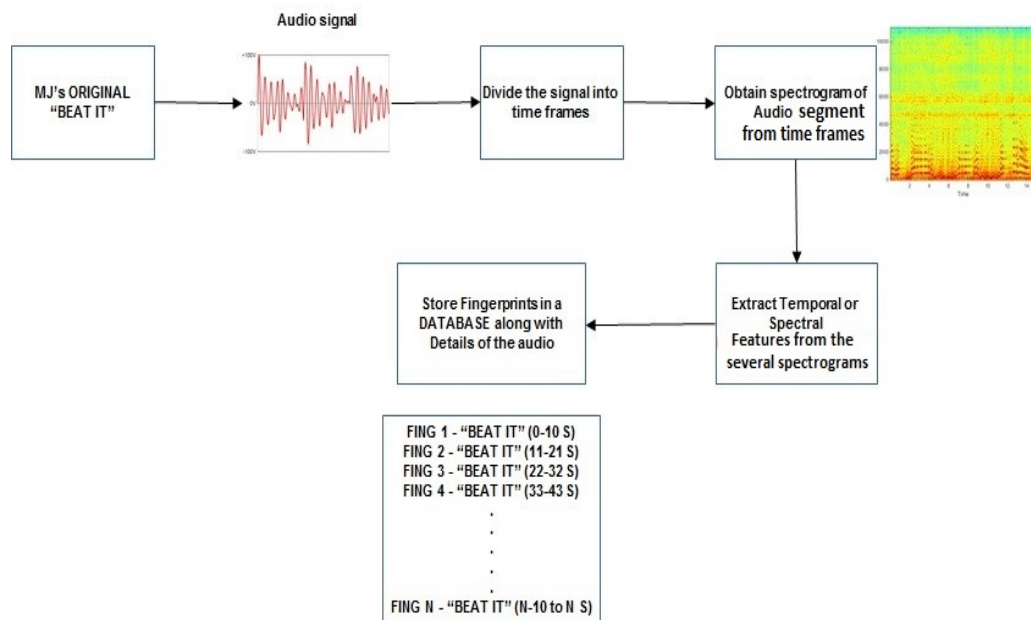


Figure 1.1: Audio fingerprint generation and database setup

At the present time, there are several approaches to building an audio fingerprinting system. Most of these approaches are derived from the fundamental concept termed **Spec-**

**rogram.** A Spectrogram of an audio signal is a time - frequency representation of the signal. However, there are several ways to generate audio fingerprint from the spectrogram of an audio signal. More details on the literature of audio fingerprinting are provided in chapter 2. The following figures, depict the general idea on audio fingerprint database setup and scoring process.

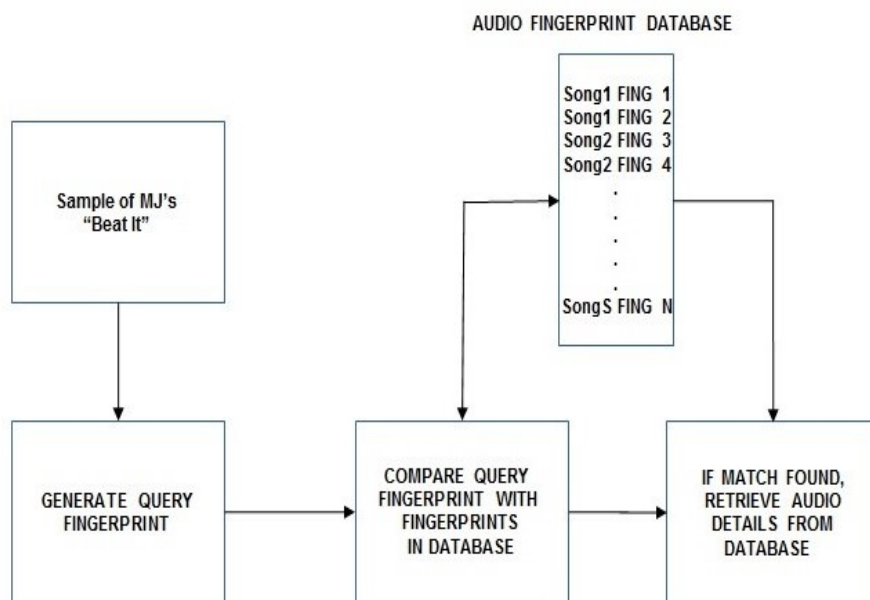


Figure 1.2: Scoring process and audio identification using signatures commonly used in music identification application

The concepts below are some of the most commonly employed manipulation of the spectrogram data of audio signals:

### 1.2.1 Filter Bank

The Spectrogram of an audio signal provides rich information on the frequency scale. However, such information cannot be directly employed as audio signatures, as they contain too high a resolution to be distinctively separated. In order to generate distinct signatures, most audio fingerprinting algorithms employ the use of band pass filter bank. Filtering the audio signal through filter bank reduces the resolution of the information, thereby enhancing robustness of audio signatures in the form of a more distinct spectrogram.

### 1.2.2 Mel Frequency Cepstral Coefficients (MFCC)

MFCC represent coefficients of the **Mel Frequency Cepstrum (MFC)**. MFC is different from the **Linear Frequency Cepstrum** in a sense that the MFC frequency bands are uniformly spaced in the Mel frequency scale unlike the non - uniform distribution of the linear frequency bands in the Mel scale. Because of this nature, the MFC coefficients, are able to approximate the human auditory system better than the linear frequency cepstrum. For this reason MFCC data is seen as interesting method method to extract audio fingerprint features. It is already being used extensively in speech recognition.

### 1.2.3 Peak Based Methods

This is another approach to utilize the spectrogram data to generate audio signatures. In these approaches, after passing the spectrogram through the filter bank, the peaks in each of the bands in the resultant data, is utilized to generate signatures for the audio files. There

are several algorithms that utilize the peak based approach and we can group them into two basic concepts:

- **Same Frequency Band Feature Extraction:**

In this approach, the audio signatures are generated by computing the time difference ( $\Delta t$ ) between the peaks in each individual frequency bands. Each  $\Delta t$  is a feature of the audio signal and it is more robust than just considering the peaks in isolation. In music identification features extracted, based on both location of peaks and their relative position with respect to other peaks ( $\Delta t$ ), are more potent than features extracted from just filter bank results.

- **Cross Frequency Band Feature Extraction:**

Cross - band analysis goes one step further and considers exploiting the relative position of peaks with respect to not only other peaks in the same frequency bands but also with respect to peaks in adjacent frequency bands. By doing so, the uniqueness of the signatures generated, has been increased further when compared to same - band analysis. In other words, feature extraction includes, computing both ( $\Delta t$ ) and ( $\Delta f$ ). It is generally considered that, signatures generated using cross bands have shown better resistance to impact of noise than those generated using same - bands.

### 1.3 ISSUES IN CURRENT SYSTEMS

This thesis is driven by the fact that, although there are several audio fingerprinting algorithms currently in use, there is **no fixed standard** to compare the efficiency of each approach against the other. Some basic problems in audio fingerprint extraction and matching include, the speed of execution of the algorithm and storage space for the database.



These are termed as cost incurred for each implementation. For example, cross-band analysis have a greater matching rate when compared to same-band analysis but results in more fingerprints being generated than same - band analysis. So a larger space is required when storing fingerprints resulting from cross-band analysis.

1. Consider the problem of building an audio fingerprinting system for one of the applications discussed in section 1.1. Developers would need concrete evidence to choose one existing audio feature extraction algorithm over the other. Surely, existing cross - band feature extraction show better results than same - band analysis. But the number of peaks and the window size differ in cross frequency band analysis when compared to same frequency band analysis. Is it really a worth while expense on cross frequency band analysis when there is no even standard to compare cross frequency band analysis against same frequency band analysis? What if cross frequency band analysis pales in performance when compared to same frequency band analysis when they are built on the same framework? Even if cross frequency band analysis proves to be better, can one really use all the frequency bands generated in the same window? What then is the optimum number of bands to consider in one window?
2. Some music retrieval system currently being employed define signatures based on the MFC coefficients described in section 1.2.2. There is a fundamental difference in the matching algorithm between peak based approach and the MFCC approach. Although MFCC seem to show promise, they fundamentally involve more complex computation over filter bank. Is it really worth while to compute MFCC or can filter bank show satisfactory results on comparison with MFCC? If they do, can we choose to deploy filter bank results in the MFCC matching algorithm to reduce expense of computation?

## 1.4 THESIS STATEMENT

This thesis is driven to provide a solution to the issues discussed in section 1.3. The following thesis statement is supported by genuine verification, which are incorporated in this thesis:

**If an agreeable framework can be defined to provide a platform to compare the various closed set audio retrieval algorithms, a system developer will then have a concrete evidence to design their system based on the available computational resources to achieve required or even better outcomes.**

## 1.5 SUMMARY OF CONTRIBUTION

As stated in section 1.4, this thesis is an attempt as a foundation for a framework that compares the various audio retrieval approaches currently in practice.

The results in this thesis establish an environment in which a system developer can develop an audio retrieval algorithm with least computational cost. This thesis provides the following contributions:

1. The foremost contribution from the results of the thesis establishes an original way of modelling cross - band approaches and same - band approaches under the same framework. Unlike the approaches depicted in [1, 4], the framework established in this thesis, does not directly compute  $\Delta t$  and  $\Delta f$ . Instead, a double scoring function is defined as against the traditional scoring algorithms explained in [1, 3, 4, 5, 7].
2. The results in this thesis shows a comparative analysis of MFC coefficients over the use of just filter bank to extract acoustic features. However, the scoring algorithm

for MFCC based audio retrieval algorithm is based on Euclidean distance between the coefficients, while that of the peak based approach is based on computing  $\Delta t$  and  $\Delta f$ . In order to make a comparative study, the features extracted from filter bank approach are also computed using Euclidean distance method.

## 1.6 ORGANIZATION OF THE THESIS

Chapter 2 highlights the audio fingerprinting generation and scoring process, as well as the various currently available approaches to achieve the same. Chapter 2 also provides a major motivation towards this thesis.

Chapter 3 discusses the problem statement and research questions that fuel this thesis. It outlays the need for a generalized framework to evaluate audio fingerprinting systems and also throws light on the efficiency of using MFCC for devising audio fingerprinting systems.

Chapter 4 establishes the simulation setup that is employed in this thesis, to define the framework that evaluates the different approaches of audio fingerprinting.

Chapter 5 defines a framework to evaluate both same - frequency band and cross - frequency band approaches to audio fingerprinting. It also discusses the results of comparative evaluation of the peak - based approaches devised on the common framework.

Chapter 6 discusses the implications of employing of MFCC in audio retrieval systems by evaluating the MFCC approach and comparing the results with just the filter bank approach without computing peaks. Incidentally, the filter bank approach also follows the same scoring

process as that of MFCC approach to provide a fair comparison.

Chapter 7 concludes this thesis by synthesizing the results discussed in chapter 5 and 6 along with the contributions of this thesis towards the research questions.

## **CHAPTER 2**

### **LITERATURE REVIEW**

The following sections clarify on some of the common approaches used to generate audio fingerprints.

#### **2.1 GENERAL PRINCIPLES OF AUDIO FINGERPRINT EXTRACTION AND MATCHING:**

Irrespective of the number of different ways an audio fingerprint is generated, for an audio signal, most implementations use the same general principles to build the skeleton for their algorithm. The flowcharts displayed in Figure 2.1 dictate the general principles behind audio fingerprint extraction and audio fingerprint scoring.

Figure 2.1(a) depicts a flowchart that explains the general principles behind audio fingerprint extraction while Figure 2.1(b) represents the basic idea of matching the sampled audio's fingerprint with the correct audio fingerprint stored in the database.

Each audio signal is segmented into a number of 10 second audio segments. Each segment is divided into a number of overlapping frames. Each frame is then treated with a hamming window. The FFTs of the multiple windowed audio frames provide the spectrogram of the audio segments. Typically an audio fingerprint is a set of information on all frames of the audio segment.

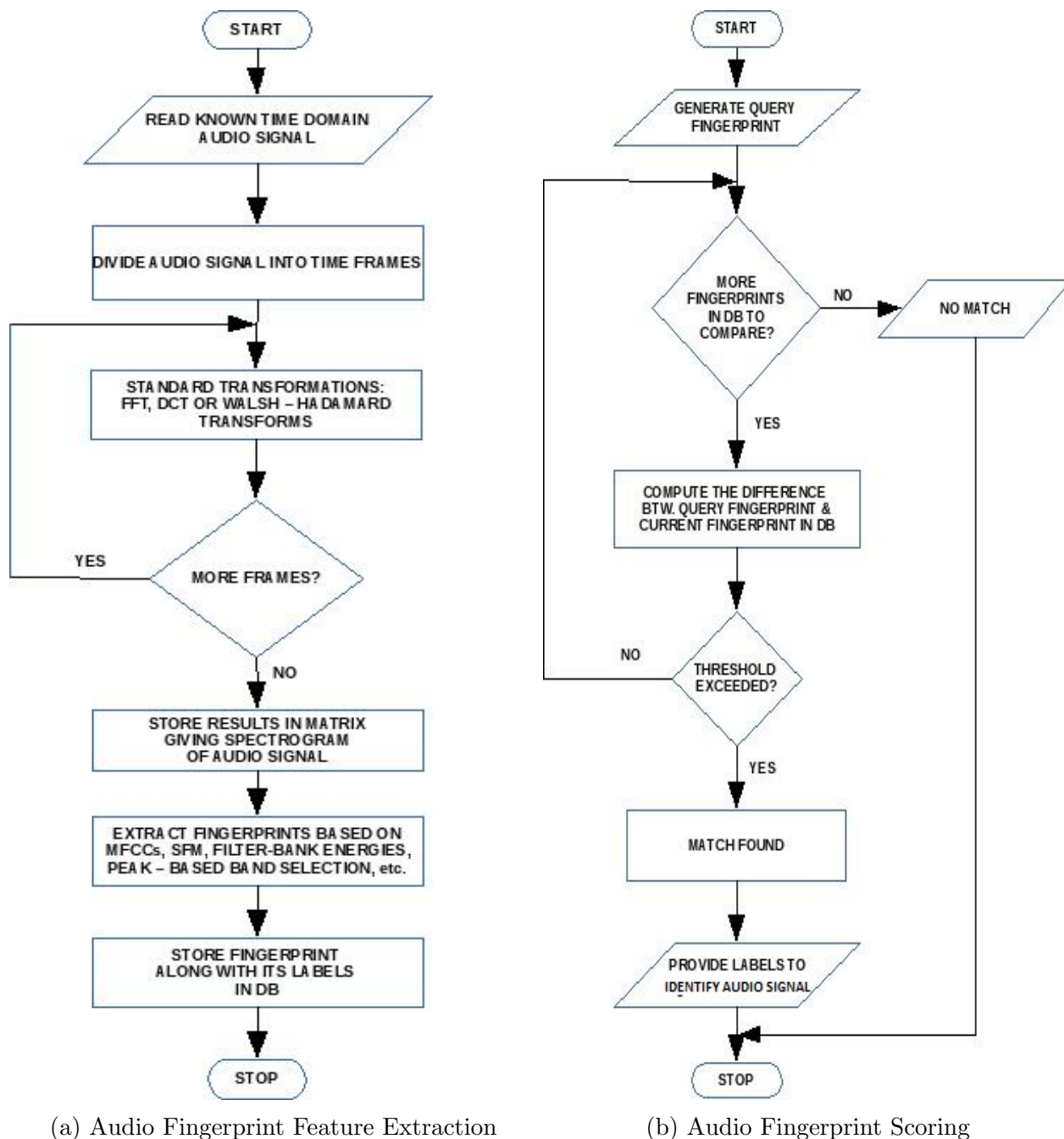


Figure 2.1: General Principles Of Audio Fingerprinting

Each fingerprint is stored in a database, along with two unique labels that represent the location of the audio signal as stored in the database and the audio signal name in the actual

folder containing all audio signals. Once a match is acquired, the corresponding audio signal can be retrieved using the above mentioned labels.

## 2.2 SPECTROGRAM

A spectrogram of an audio signal is the time - frequency analysis of that signal in the form of an image. In other words, a spectrogram provides a visual representation of the frequency spectrum of the time - domain signal. A spectrogram provides a more precise understanding of the otherwise unrecognizable spectral features in the signal. This is the fundamental reason to why researchers employ spectrogram analysis to extract features that distinguish each audio signal. A spectrogram can be generated at various time frames using Fourier transforms.

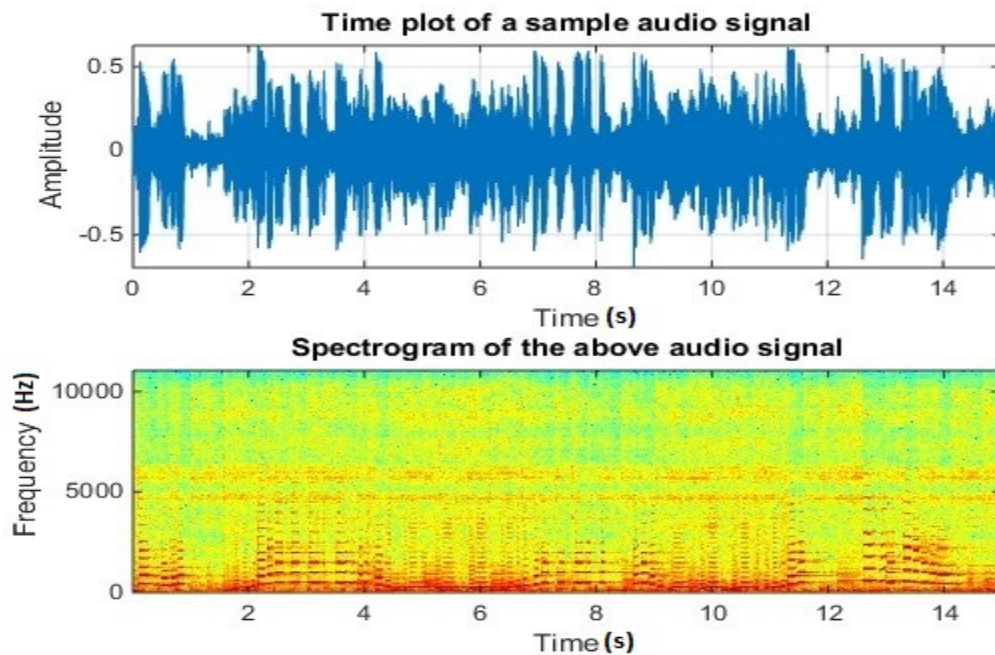


Figure 2.2: Spectrogram of a digital audio signal

While audio fingerprints can be directly computed from Fourier transforms, they can also be computed after passing the results of the Fourier transforms through a filter bank. A filterbank is a set of band-pass filters that extract spectral components of the audio signal on various frequency bands across time. This approach was the only means to generate spectrogram before the advent of digital signal processing. With the onset of digital signal processing, an alternative approach to spectrogram processing was defined using Fourier transforms. In general, the input signal is sampled and divided into time frames. The FFT of each time frame of the signal is then computed and put together to give the spectrogram of the input signal. Figure 2.2 is an example of spectrogram of a sampled audio signal.

### 2.3 MEL FREQUENCY CEPSTRAL COEFFICIENTS (MFCC)

The Mel scale for frequency domain was introduced to perceive **”pitch”** with a more understandable definition. By experiment in [21], it has been established that up to 1000 HZ, the pitch is perceived in the linear scale but above 1000 Hz, the pitch was found to be in logarithmic scale. Thus to define pitch in linear scale, the Mel frequency scale was introduced. A linear frequency  $F_{Mel}$  can be converted to a Mel scale frequency  $F_{Hz}$  using the following relation:

$$F_{Mel} = \frac{1000}{\log_{10}(2)} \cdot \left[ 1 + \frac{F_{Hz}}{1000} \right] \quad (2.1)$$

The figure 2.3 shows a relationship between Mel frequency scale and linear frequency scale based on equation 2.1. The sampling frequency is set to 8000 Hz for a frame size of 25 milliseconds.



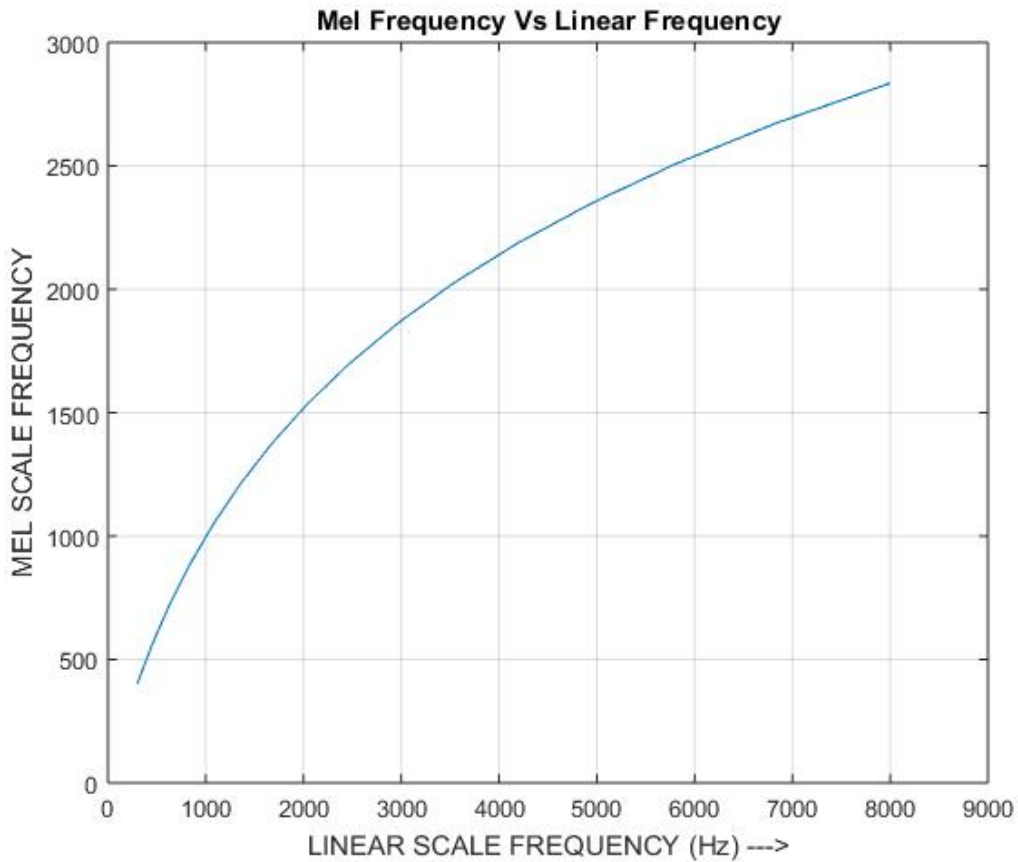


Figure 2.3: Relationship between Mel scale and Linear frequency scale

MFC coefficients can be calculated as follows:

- Pass the spectrogram of the digital input signal through a filter bank of band pass filters. The center frequency of the said filters are linearly distributed below 1000 Hz and for frequencies above 1000 Hz, equation 2.1 is used to compute the center frequency.
- The total energy in the output of filter bank is calculated using the following relation:

$$E(i) = \sum_{k=0}^{\frac{N}{2}} \log_{10} |S(n, k)| \cdot \left| H_i \left( \frac{2\pi}{N} \right) \right| \quad (2.2)$$

where,  $S(n,k)$  is the  $k^{th}$  bin of the FFT of the audio segment from the  $n^{th}$  frame and  $H_i\left(\frac{2\pi}{N}\right)$  is the frequency response of the filter bank.

- The computed energies in each band are then converted to  $N_{FB}$  MFC coefficients  $C(1), C(2), \dots, C(N_{FB})$  using the following relation:

$$C(i) = \frac{2}{N} \cdot \sum_{k=1}^{N_{FB}} E(k) \cdot \cos\left(k \frac{2\pi}{N} n\right) \quad (2.3)$$

The MFCC obtained can be further used to generate  $\Delta$ MFCC and  $\Delta^2$ MFCC to increase the robustness of the audio fingerprints extracted using the approach in this section.  $\Delta$ MFCC can be generated using the relation below:

$$\Delta_t MFCC = \frac{\sum_{i=1}^{N_t} i \cdot (C_{t+i} - C_{t-i})}{2 \cdot \sum_{i=1}^{N_t} i^2} \quad (2.4)$$

where,  $C_{t+i}$  refers to the MFC coefficient in the next time frame,  $i + 1$  and  $C_{t-i}$  refers to the MFC coefficient in the previous time frame,  $i - 1$ .  $N_t$  refers to the number of time frames being considered for the computation of the regression equation 2.4.

MFCC are mostly used in speech recognition, and for that purpose the lower cepstral coefficients are used while the higher cepstral coefficients are eliminated as is the case in [13, 21]. This is because the lower coefficients better reflect the transfer function of the human vocal tract. In a similar fashion, music retrieval systems consider the higher cepstral coefficients better reflect the harmonics of the background music. This is proven by work in [14, 15]

## 2.4 SAME - FREQUENCY BAND FEATURE EXTRACTION

One of the most significant same-band frequency analysis approach to audio signature extraction and scoring is established in [4]. Unlike [1] or any other cross-band frequency analysis algorithms that rely on spectral features in an audio file, [4] relies on temporal features such as onsets. An onset refers to the start of a musical note in an audio signal.

Firstly, to generate audio signatures, [4] runs the audio sample through the process of “whitening” to suppress any stationary resonances that may be present in the audio sample. “Whitening” flattens the spectrum of the signal so that it closely resembles white noise. This process is achieved by estimating the spectrogram of the audio sample and filtering it using a time – varying inverse filter that is calculated from the spectrogram of the audio sample. The whitened music sample is then passed through a band pass filter (MPEG – Audio 32 – band Filter Bank) to partition it into a number of frequency bands. The frequency bands range from 0 to 5500 Hertz. The output of the filters are termed as filtered music samples. Each filtered music sample is a series of time-domain samples representing the magnitude of the music sample within the corresponding frequency band.

Onsets can now be detected from within each filtered music sample along with a time-stamp that indicates the occurrence of onsets in the filtered sample with respect to a previous onset as a measure of time ( $\Delta t$ ). Detectors in each frequency band can be used for this purpose. Onset detection is achieved by comparing the magnitude of the corresponding filtered music sample with a fixed or time-varying threshold derived from the current and past magnitude within the respective band. These extracted data constitute the features for the audio

sample. This entire process is well explained in [5].

[4] provides a way to determine probability of error for its algorithm. However, as explained in section 1.3, [4, 5] consider their own guiding principles to justify its efficiency.

## 2.5 CROSS - FREQUENCY BAND FEATURE EXTRACTION

[1] explains one of many cross – band analysis approaches in audio signature extraction and scoring. [1] depicts a Fast Combinatorial Hashing technique to generate an audio signature database of known audio signals and generate audio signatures for audio samples (also termed queries). It also explains the concept of single – slice fingerprint extraction (function of one spectral peak), its drawbacks and how multi – slicing fingerprint extraction (function of more than one spectral peak) is better than the former in reducing false positives.[1] also claims that same – band analysis end up with high rate of false positives as compared to its cross – band analysis.

[1] reduces the rate of false positives by generating a constellation map of spectral peaks in the spectrogram of the audio signal. [1] then defines these constellation peaks as anchor points. Each anchor point in a frequency is associated with its own target zone which comprises of other constellation peaks in adjacent frequencies. Audio signatures are then generated from pairs of anchor points and associated target zone constellation peaks. Each pair of constellation peaks produce two components in the signature:

1. Frequency difference component ( $\Delta f$ ).
2. Time difference component ( $\Delta t$ ).

[1] claims that, given a multitude of different performances of the same audio signal, it is capable of identifying the correct audio even when the samples are virtually indistinguishable to the human ear. [1] fails to explain the increase in cost due to additional computation over same-band frequency analysis to achieve such said higher results.

Apart from the approaches described in this chapter, there are some audio retrieval algorithms that rely on Linear predictive Coding (LPC) [19], Hidden Markov Model (HMM) [17, 18], etc.

## **CHAPTER 3**

### **ISSUES CONTRIBUTING TO THESIS**

This thesis is motivated by the issues in the current audio signature extraction and scoring systems. It is an attempt to answer some of the questions in section 1.3. This chapter focuses on the issues that are considered to be influencing the decisions of system developers.

#### **3.1 PROBLEM 1: NONEXISTENT STANDARD IN COMPARING CURRENT APPROACHES**

As previously instituted in section 1.3, there is no established standard for comparison of the peak based approaches explained in sections 2.3 and 2.4. Same - frequency band feature extraction and cross - frequency band feature extraction have not been evaluated in the same framework. In other words, there is no definitive information to actually determine whether cross - frequency band feature extraction is better than same - frequency band analysis.

Are cross - frequency band analysis really better than same - frequency band analysis? Furthermore, even if cross - frequency band analysis tend to be better, can we use all frequency bands in the spectrogram to extract features? What could be the optimum number of frequency bands that could be employed in cross - frequency band analysis to attain maximum efficiency? Same - frequency band and cross - frequency band feature extraction algorithms have not been evaluated on the same framework before, until now.

What if there was a way to model the two approaches under the same framework? Is there a way for researchers to actually decide between same - frequency band and cross - frequency band feature extraction approaches, based on the same efficiency metric? Unfortunately, there is no current system for researchers to deploy, due to the unavailability of such metrics, to achieve the same. As of now, researchers can only go by the efficiency defined in each individual approach.

Establishing a metric to analyze same - frequency band and cross - frequency band feature extraction on the same framework could provide the following solutions. Researchers can choose between peak based approaches to meet their design requirements. Researchers can decide the number of frequency bands to consider while extracting features in cross - frequency band approach. Furthermore, this could also prove beneficial to alleviate the expense on scoring algorithm, due to the amount of data being traversed each time to identify a match in cross - frequency band approach. Therefore, this thesis attempts to compare the performance of two peak based approaches under a common framework.

## **3.2 PROBLEM 2: MFCC APPROACH TO MUSIC RETRIEVAL**

This thesis is motivated by the fact that MFCC provide better representation of the human auditory system. For this reason, by using the higher coefficients of the MFCC, they seem to show heightened efficiency. Traditional peak based approaches such as [1, 3, 4, 5] make use of  $\Delta t$  and  $\Delta f$  to record scoring hits. So can we really say it is worthwhile to compute MFCC?

Computing MFCC is a much complex process as just employing results of filter bank in peak based approaches. MFCC are computed from the results of filter bank as well but they require a longer computational time as compared to the generalized peak based approaches. This could be one of the reasons as to why MFCC are not as popular in music retrieval as they are in speech recognition since the lower coefficients of the MFC mostly contain information on the pitch of the human voice. This could be the reason as to why higher MFC coefficients are considered for music retrieval. Why do researchers have to move away from traditional norm of employing  $\Delta t$  and  $\Delta f$  when it comes to using MFCC for music retrieval?

Apart from the difference in the scoring algorithm, as mentioned already, MFCC computation are more expensive than other approaches. Furthermore, MFCC computation complexity increase further when additional information are considered to be part of the system. Such information include  $\Delta$ MFCC and  $\Delta^2$ MFCC. This is so as to compensate for the excessive data loss due to omission of the lower MFC coefficients. Hence an MFCC computation is complete only when all MFCC,  $\Delta$ MFCC and  $\Delta^2$ MFCC are computed as explained in [13, 14, 15, 21]. Do researchers really need to increase computation cost to such heights to attain better results? What if the results of the filter bank are the same when the scoring algorithm is modified to that of the MFCC approach? If so, then the additional complexity can be avoided. This thesis aims to answer this particular question in researchers.



## CHAPTER 4

### OVERVIEW OF EVALUATION PROCEDURE

This thesis made use of MATLAB [10], for implementing and testing of the proposed framework. Furthermore, a closed set of 60 music signals were used in the evaluation. The following sections discuss the fundamental assumptions and simulations involved in this thesis:

#### 4.1 FRAMING AND SPECTROGRAM GENERATION

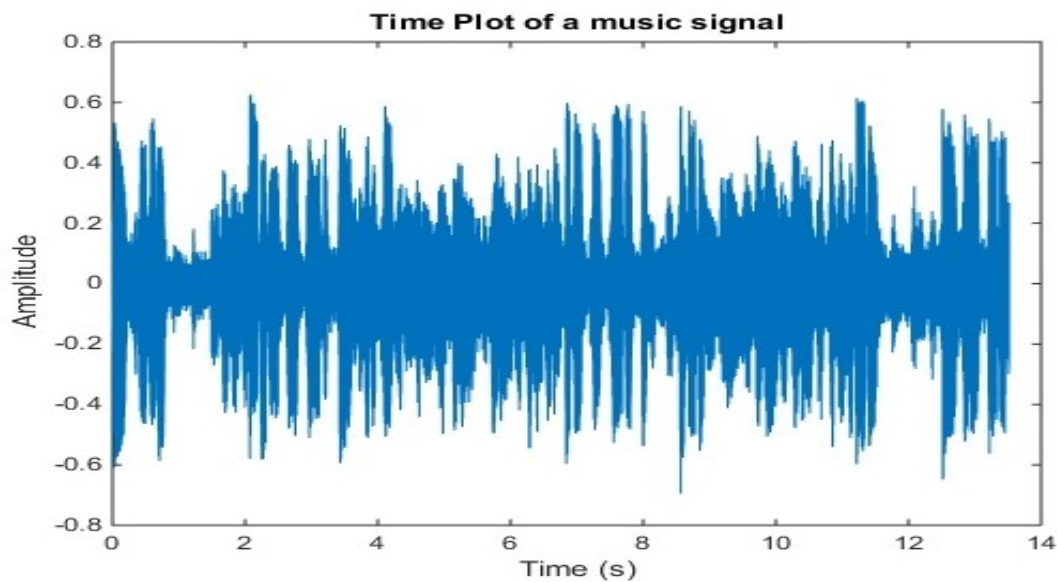


Figure 4.1: Time Domain Audio Signal

The music signals were first converted from continuous analog signals to sampled digital signals. The sampling rate of these music files are set to 22050 Hz. The sampled signals

are then framed using a window size of 40 milliseconds. Therefore, based on the sampling frequency and the frame size, the number of samples in each frame is given as  $22050 \times 0.040 = 882$  samples/frame. Framing window is then made to shift for every 20 milliseconds. The framing process of an audio signal shown in figure 4.1 is illustrated in figure 4.2. Each

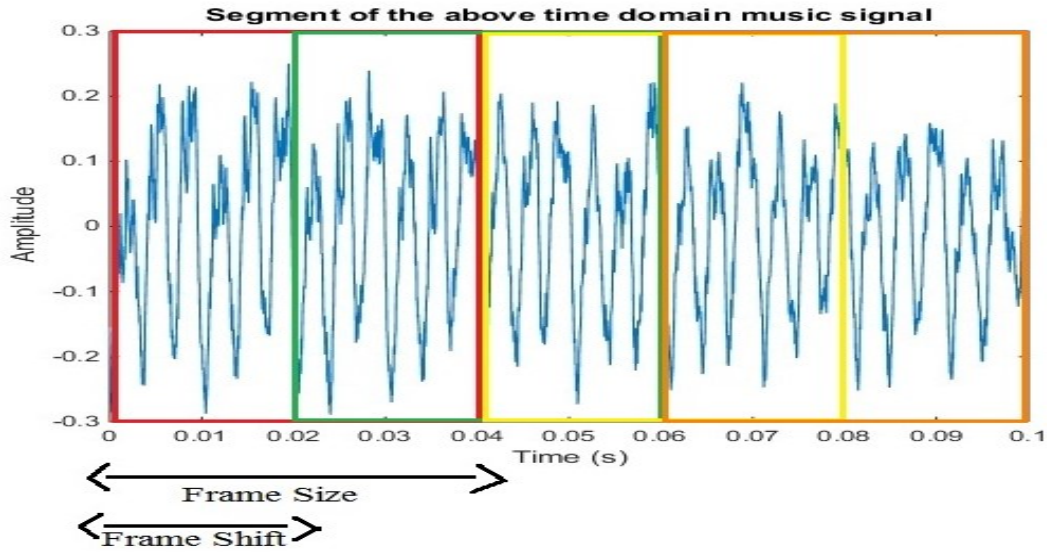


Figure 4.2: Framing a segment of the audio signal

frame is passed through a hamming window and FFT of the same is computed. Hamming window is of interest to reduce the side lobes caused by the FFT of a time limited signal. Employing hamming window results in smooth wrapping of audio segments or frames.

The resulting vector is stored as a column in a matrix  $\mathbf{M}$ . This process is repeated for each frame after a 20 milliseconds shift. The resulting matrix is equivalent to the spectrogram of the music file. The final spectrogram of the audio signal is shown in figure 4.3.

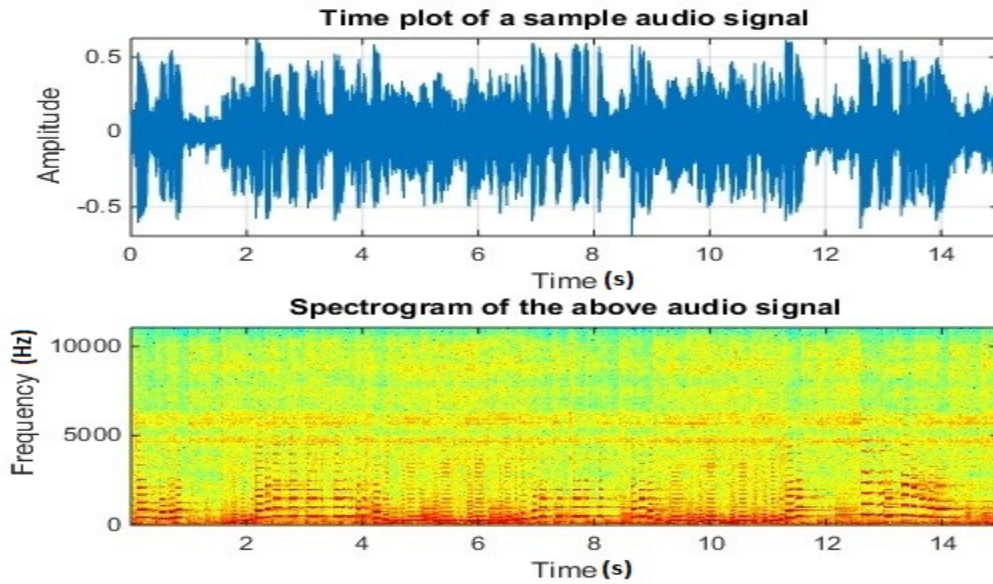


Figure 4.3: Spectrogram of the audio signal

## 4.2 BASELINE APPROACH WITHOUT THE USE OF FILTER BANK

### 4.2.1 DATABASE GENERATION

The spectrogram is further transformed into a binary matrix  $\mathbf{M}_b$ . To generate the binary matrix, for values greater than a predefined threshold, a value of 1 is assigned to the same position of the said value in the spectrogram. For positions of values below the threshold  $\mathbf{th}$ , the new value is set to 0. The figure 4.4 shows the binary spectrogram for a threshold value of 0.5.

$$M_b(i, j) = \begin{cases} 1, & \text{if } M(i, j) \geq \mathbf{th} \\ 0, & \text{otherwise} \end{cases} \quad (4.1)$$

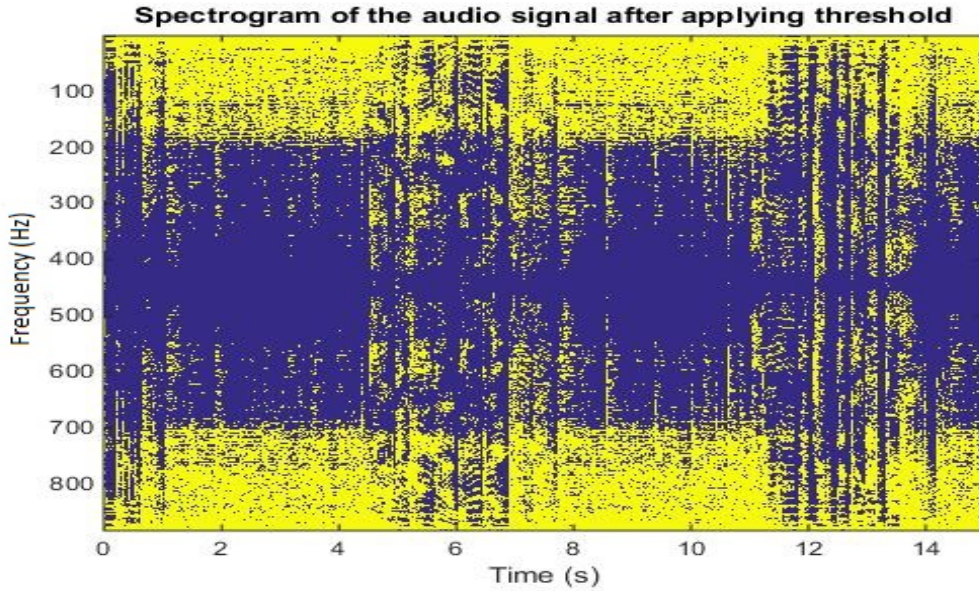


Figure 4.4: Spectrogram with threshold of 0.5

After computing the modified spectrogram of each audio signal, they are set up in one matrix  $\mathbf{M}_{DB}$  that serves as the database. For the setup in this thesis, 60 audio signals, each 15 seconds long are used to build the closed set. Each signal is then divided into time frames of size 40 milliseconds. For a query signal with length of 10 seconds, this framing process leads to 6 audio segments in each audio signal, thereby resulting in 6 audio fingerprints per each audio signal. During this process each frame is shifted by 20 milliseconds thereby creating an overlap of 20 milliseconds. For such a setup the resulting database is of the size 441 X 44880. The figure 4.5 depicts the Fingerprint database setup.

## 4.2.2 QUERYING AND SCORING

The query signal is a random 10 seconds of any one of the 60 music files that was assimilated in the database. Since the thesis only considers a closed set, only samples within the 60 set music database is chosen as queries. A random white noise is then generated and

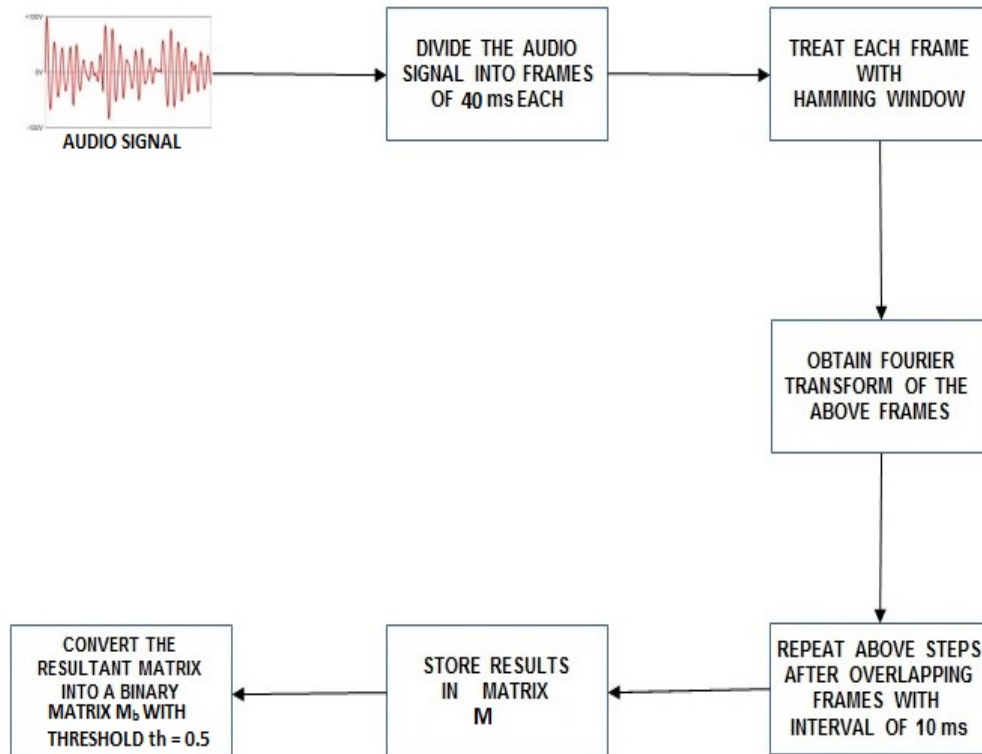


Figure 4.5: Fingerprint database setup

added to the selected 10 second signal. The query signal is then passed through the same process as the database. Due to the length of the query, there are 6 possible signatures for each of the music files. Scoring of the query with one of the music files in the database is a simple process. The query is matched over each of the signature per music in the database.

By obtaining the sum of the values within the resultant matrix, the signature with the highest match score is returned as the matching signature and music. The figure 4.7 shows the end result of querying random music files from among the music database using the process described in this process. The results depict probability of error ( $P(e)$ ) against standard

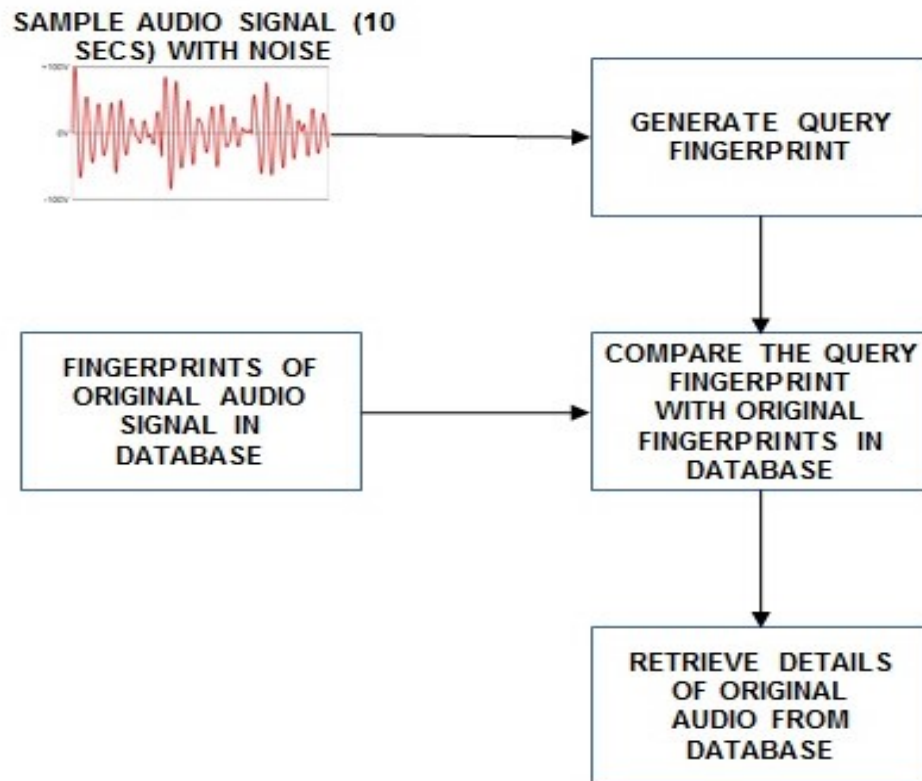


Figure 4.6: Fingerprint Scoring

deviation of white noise ( $\sigma$ ) added to the music sample. The process is repeated for 500 samples of one value of standard deviation of noise. The results are found to be satisfactory and are found to converge as expected.

It is important to highlight that, the procedure demonstrated in section 4.2 is not employed in further evaluation. This is only a baseline approach to setup the foundations for using a closed set of 60 audio signals in a robust environment that defines a single framework to evaluate most known approaches to audio fingerprinting.

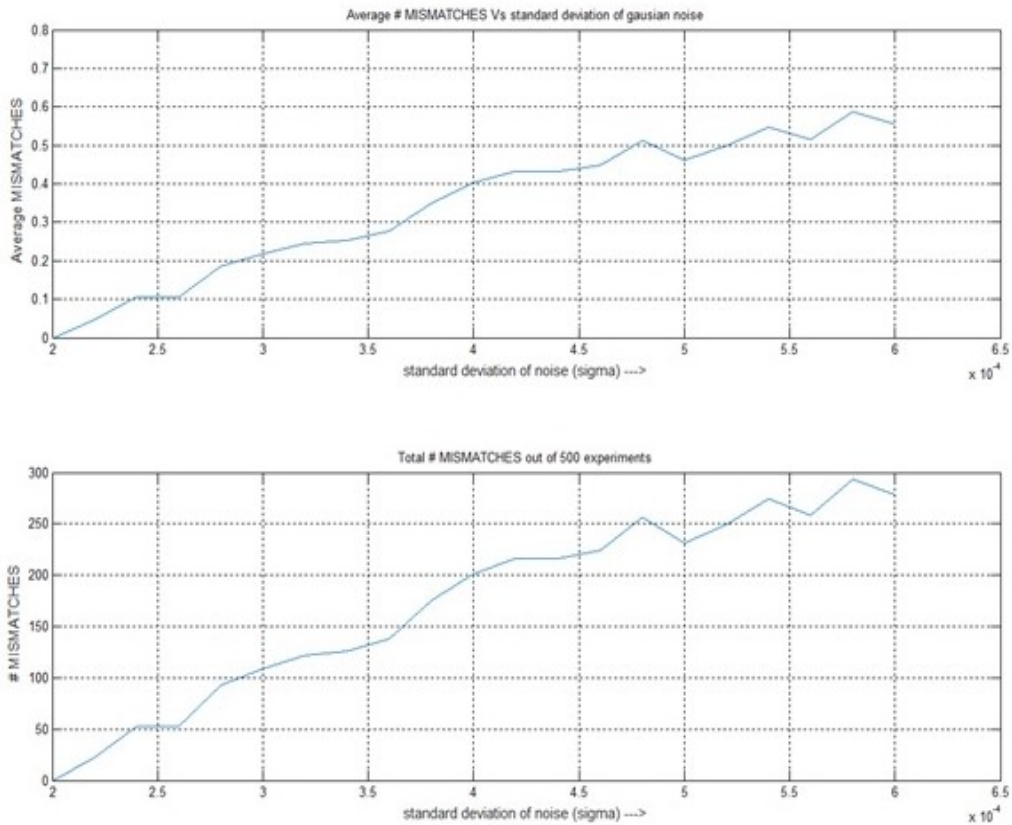


Figure 4.7: Audio signature scoring results

For the remaining chapters that discuss the results of the thesis, the standard deviation of the white noise is computed to satisfy a predefined Signal to Noise Ratio (SNR). The SNR is given by the following relation:

$$SNR = \frac{1}{N} \sum_i \frac{|x(i)|^2}{\sigma^2} \quad (4.2)$$

All results in the following chapters are based on the SNR computed using the equation 4.2. Therefore all results in the following chapters will show probability of error ( $P(e)$ ) against SNR.

## CHAPTER 5

### PEAK BASED APPROACH

Peak based approaches to audio fingerprinting rely on the results of passing the FFT of the audio frames through a filterbank the result would be a more robust spectrogram as represented in section 4.2. Be it same - frequency band approach or cross - frequency band approach, the first and foremost process is to find the peaks in each frequency band of the resultant spectrogram of the audio signal. The following sections explain the detailed process of peak based approach being simulated to compare same - frequency band and cross - frequency band approach.

#### 5.1 FILTER BANK GENERATION

The filterbank for this thesis is generated using the MEL scale. The relation being employed for computation MEL frequency  $F_{Mel}$  from linear frequency  $F_{Hz}$  is given in equation 5.1.

$$F_{Mel} = 1125. \ln \left( 1 + \frac{F_{Hz}}{700} \right) \quad (5.1)$$

To retrieve linear frequencies from MEL scale frequencies, the thesis employs the following relation in equation 5.2:

$$F_{Hz} = 700. \left( \exp \left( \frac{F_{Mel}}{1125} \right) - 1 \right) \quad (5.2)$$

The relation between the linear frequencies and the associated Mel frequencies being employed in the filter bank generation is shown in Figure 5.1. The generated filterbank is then



applied to the framing and spectrogram generation process discussed in section 4.1. As soon

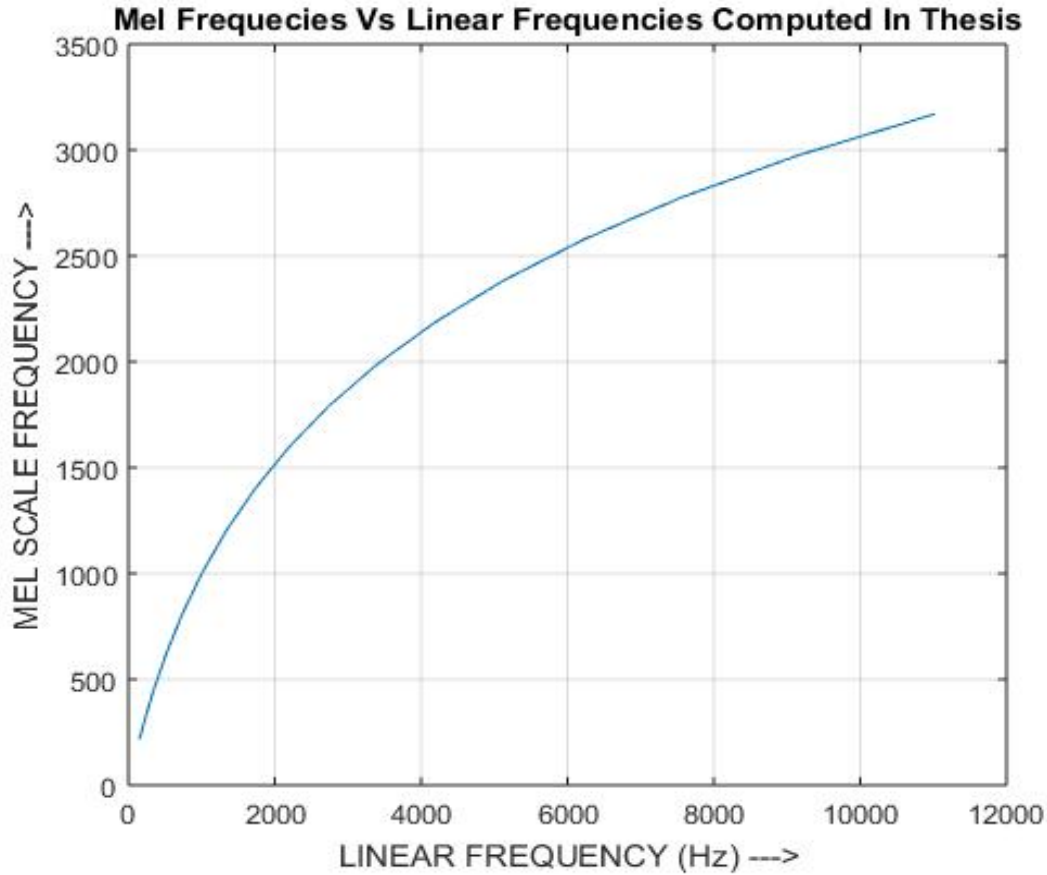


Figure 5.1: Relation between Mel Frequency Vs Linear Frequency in thesis

as the FFT of each frame is computed, it is passed through the filterbank resulting in a modified lower dimensional spectrogram. In this thesis we term the resulting matrix  $M_{DB}^{FB}$  as the filter bank results of the database. The filter bank employed in this thesis is depicted in Figure 5.2. As shown, 14 filter bands are being considered between 150 Hz and 11025 Hz. There is no specific reason as to the number of frequency bands being considered for the filter bank. The general norm is to consider at about 25-28 frequency bands however, this thesis proves that it is not a mandatory requirement.

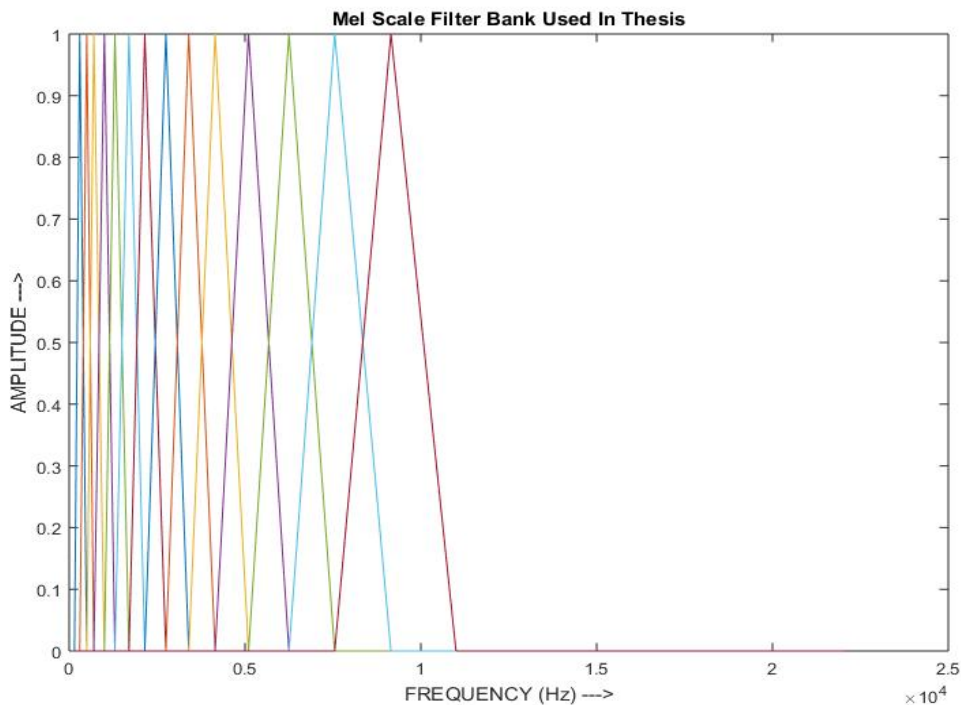


Figure 5.2: Mel scale filter bank

## 5.2 NEED FOR ONE FRAMEWORK

The primary research question this thesis is attempting to answer is the need for a generalized framework to compare for all peak based approaches to audio fingerprinting. For each of these approaches, the general idea is to compute the peaks in each of the frequency bands in the spectrogram of the audio signal. Since the length of the database is unchanged even after passing through the filter bank, there are only 6 possible fingerprints per music file in the database.

The scoring algorithm still follows a matrix dot product but with a slightly modified setup. Instead of passing the query over the original database  $\mathbf{M}_{\text{DB}}^{\text{FB}}$ , it is passed over a mask

of the original database  $\mathbf{M}_b^{\text{FB}}$ . The mask is basically a binary matrix that has a value of 1 in the position of peaks in the database while the other positions have a value of 0. This thesis defines a slightly different scoring process than the ones discussed in traditional peak based approaches such as [1, 4].

### 5.2.1 Double Kill Scoring Process

Firstly, the query is passed over the sub - matrix, of the mask  $\mathbf{M}_b^{\text{FB}}$ , representing the signature in the database. Since  $\mathbf{M}_b^{\text{FB}}$  is a binary matrix, this process would kill off any unnecessary noise components in the query. This process has proven to aid in getting rid of noise components that may appear as false peaks in the spectrogram data of the query. This is the first kill in the scoring process.

Since the end goal is to define a single framework for both same - frequency band and cross - frequency band approaches, the second kill process should typically consider same number of peaks (approximately) per frame within the signature of each of the two approaches. To satisfy this requirement, the same - frequency band approach is set up with a frame length of 2 seconds, with each time frame beginning at the position of each peak in one band and then the window length of the cross - band is set such that they both have approximately same number of peaks.

Considering, most same - frequency band approaches such as [4, 5] depend on computing  $\Delta t$ , in other words the relative position of peaks within each band, this thesis computes the number of peaks within each time frame. This is assumed to be another method to hold information on relative position of peaks in the database. this assumption is proved right

based on the results shown in figures below. The same process is applied to the original database mask. If the number of peaks in the query time frame is close to the same number of peaks in the database mask time frame, a match score is incremented. This process is repeated for every peak in the signature. This process is termed the second kill in the scoring process. This approach is found to be a close approximation to same - frequency band approaches to audio fingerprinting and hence can be employed to define the standards required. Figure 5.3, depicts the same - frequency band music scoring results.

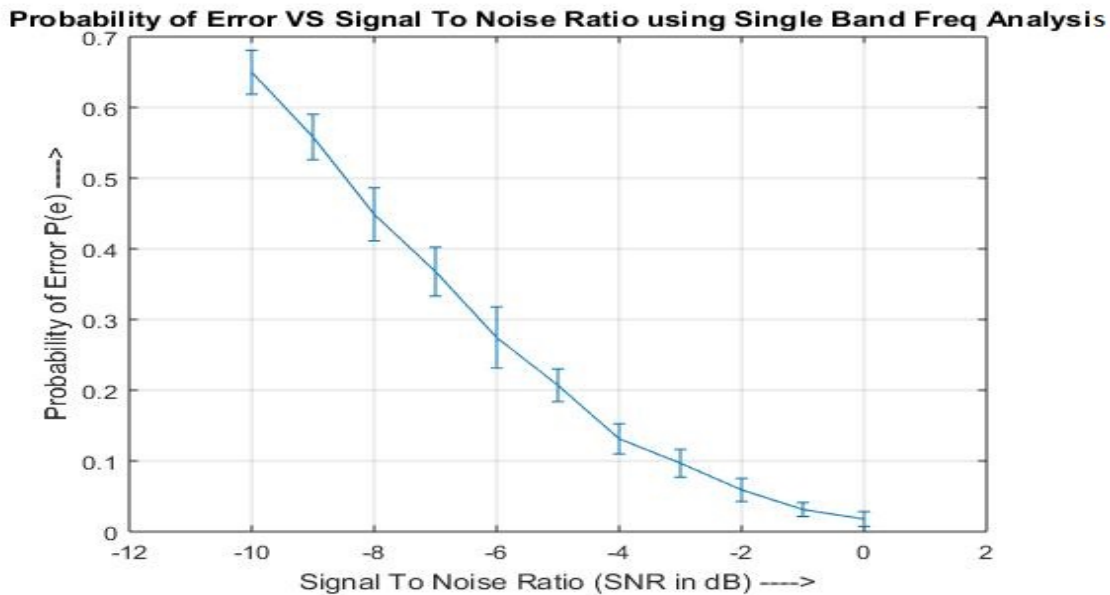


Figure 5.3: Results of single frequency band audio fingerprinting

As discussed earlier, to develop cross - frequency band audio fingerprinting setup on the same framework as same - frequency band approach, the area of the frame employed in both approaches has to be identical. By deploying the same frame size on both approaches, we considerably bring down the computation time and use up approximately the same number of peaks in each frame. This considerable difference is shown in the table below. The table 5.1 below shows the time it takes to compute different peak based approaches for one iteration

Time Frame	Number of bands	Time Consumed
Fixed Frame Size	1 (Same - Band)	2.017 s
	7 (Cross - Band)	3.458 s
	14 (All - Band)	5.071 s
Variable Frame Size	1 (Same - Band)	2.017 s
	7 (Cross - Band)	2.138 s
	14 (All - Band)	2.173 s

Table 5.1: Performance analysis of traditional approaches against the new approach

based on a fixed frame size as well as for variable frame sizes based on number of frequency bands used.

Based on the new cross - frequency band approach discussed above, the following Figures 5.4 and 5.5 show results from the approach for two setups of the cross - frequency band audio fingerprinting approach discussed in this thesis. Figure 5.4 shows the result of the cross -

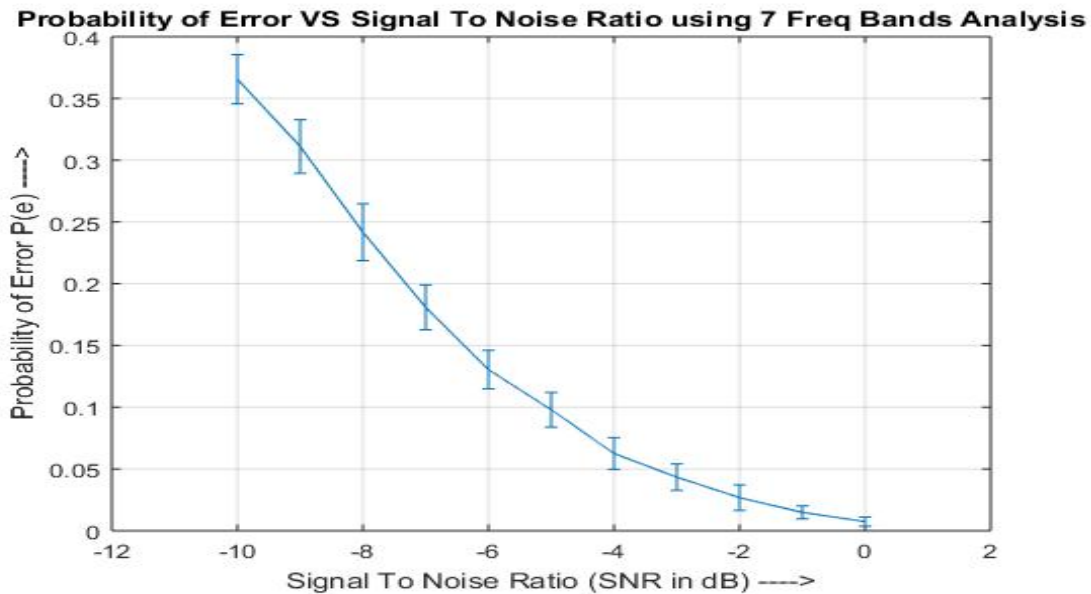


Figure 5.4: Results of cross frequency band audio fingerprinting for seven bands

frequency band approach that considers up to seven frequency bands for each frame. Figure

5.5 shows the results of the cross - frequency band approach considering all 14 frequency bands of the music file.

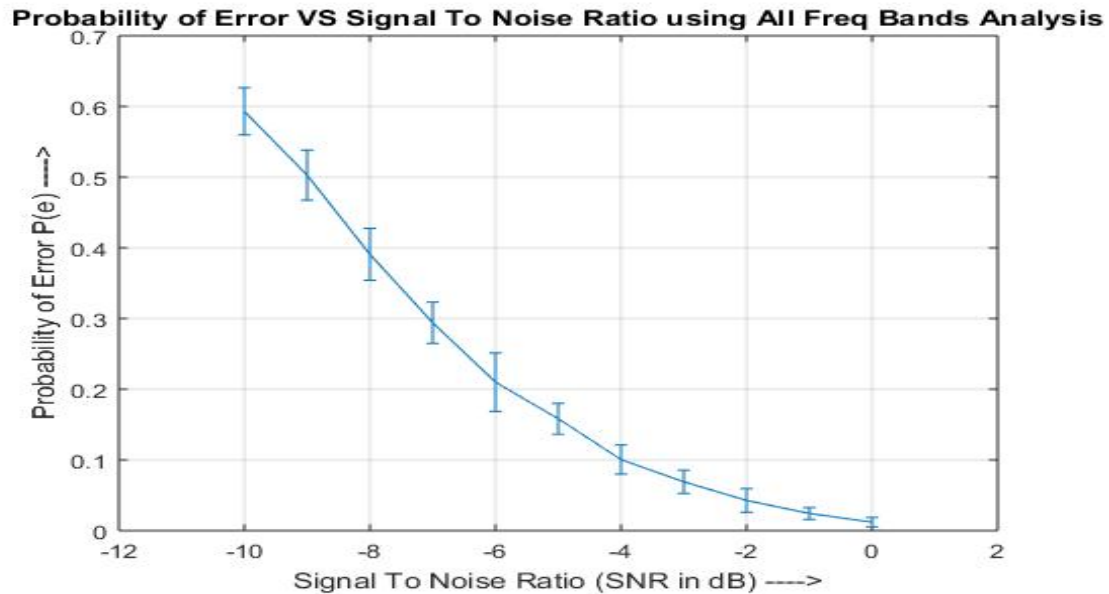


Figure 5.5: Results of cross frequency band audio fingerprinting for all fourteen bands

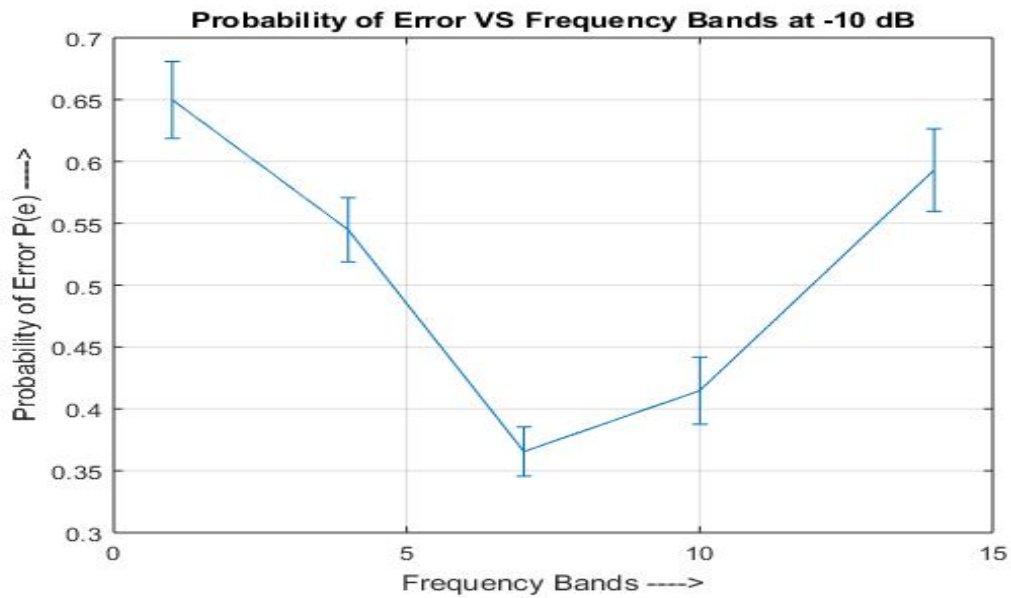


Figure 5.6: Performance of different approaches at -10 dB SNR

The results in Figures 5.3, 5.4, 5.5 show that, even in the same framework, cross - frequency band approaches still tend to show better results than same - frequency band approaches. However, at the same time, the traditional cross - frequency band approaches tend to create a burden on the scoring process. Same - frequency band approaches, especially [4], tend to ease the scoring process. Figure 5.6 depicts the performance of the framework based on considering different frequency band approaches at an SNR of -10 dB. For the setup used in this thesis, the cross - frequency band approach for audio fingerprinting using seven frequency bands is found to show the best performance when compared to other approaches.

## **CHAPTER 6**

### **EMPLOYING MFCC FOR AUDIO FINGERPRINTING**

This particular chapter aims to determine if MFC coefficients are suitable for music retrieval systems as discussed in section 3.2. Traditionally, MFCC are employed in speech recognition such as [13] for their close approximation to the human auditory system. [21] shows that the lower coefficients of the MFCC show a better approximation of the human vocal tract, and hence show highly efficient results in speech recognition. This thesis is based on the assumption that if the lower MFC coefficients resemble the pitch of the human voice, then the higher coefficients could resemble pitch of the background sound (in this case music). This assumption is at least proven to be not harmful in [11]. The following sections establish the MFCC approach to music retrieval being employed in this thesis.

#### **6.1 STEPS INVOLVED IN MFCC GENERATION**

This thesis performs the following procedures to generate MFC coefficients for the music database.

1. Employing equations 5.1 and 5.2, a triangular filter bank is first generated. To generate the filter bank, this thesis makes use of the sampling frequency of 22050 Hz and a minimum frequency of 150 Hz.
2. The audio signals are then framed as explained in section 4.1.
3. The framed samples are then treated with a hamming window as discussed in section 4.1.



4. After treating each audio frame with the window function, the FFT of the frames are computed. The absolute value of the FFT of all the frames form the spectrogram.
5. The spectrogram is then passed through the filter bank computed using equations 5.1 and 5.2. The filter bank is a set of 14 triangular band pass filters. This is achieved by typically multiplying the FFT magnitude response with the band pass filters. The main goal here is to reduce size of features involved.
6. To obtain the MFC coefficients, the **Discrete Cosine Transform (DCT)** of 20 times logarithm of the energies obtained in the previous step is computed. The DCT is obtained using equation 2.3. This results in 14 MFC coefficients for that particular audio frame.

So far in this thesis, the above discussed steps from step 1 through step 5 have been established and discussed in detail in chapters 2, 4 and 5. This thesis implements step 6 from the results of the filter bank discussed in chapter 5.

## 6.2 FILTER BANK VS MFC COEFFICIENTS

In this section, we address the following questions: is it really worth while to compute MFC coefficients? Can we employ Euclidean distance scoring process to filter bank results instead of computing MFC coefficients?. What if the modified scoring process (Euclidean distance) is the actual reason for heightened efficiency of MFC coefficients?

One reason why this problem is of interest is because the computation of MFCC,  $\Delta$ MFCC and  $\Delta^2$ MFCC involves additional processing steps. If the assumption of MFCC increasing efficiency in music retrieval, as they do in speech recognition, is not verified, MFCC compu-

tation for music retrieval becomes unnecessary, thereby saving computational expense.

The solution to this problem could be to compare the filter bank results against the MFCC results while employing the Euclidean distance approach given by the relation in equation 6.1.

$$D = \sqrt{\sum_i \sum_j (p(i,j) - q(i,j))^2} \quad (6.1)$$

where  $p(i,j)$  refers to the elements of the audio signatures in the MFCC database while  $q(i,j)$  refers to the elements of the query fingerprint generated and  $D$  refers to the distance between the two elements.

Since chapter 5 is a discussion about peak based approaches, the filter bank results from chapter 5, which is just the spectrogram obtained after filtering the audio frames using triangular band pass filters, can be compared against the MFCC computed from them. Instead of computing peaks or employing the *double kill scoring process* established in chapter 5, the modified spectrogram of the music database can then be queried by employing scoring process using Euclidean distance in equation 6.1. The results of querying on the crude filter bank music database is shown in figure 6.1.

From the results in figure 6.1 filter bank with Euclidean distance is not sufficient to device a good music retrieval system.

To obtain the MFC coefficients from the filter bank results, the discrete cosine transform of the 20 times logarithm of the energies of the filter bank results are computed. Thus each audio frame is represented by 14 MFC coefficients. Since the lower MFC coefficients mostly represent the human vocal tract and audio signals of our closed set contains songs, this thesis

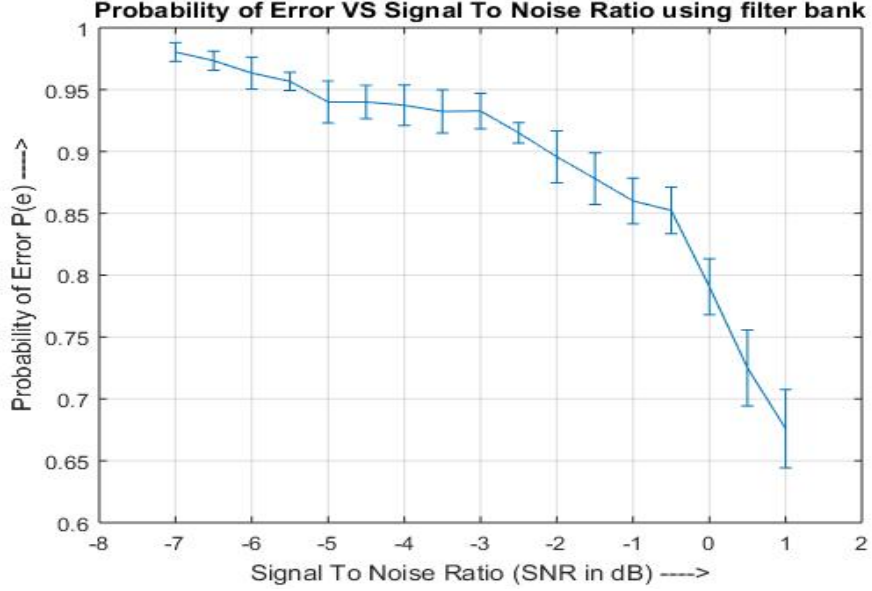


Figure 6.1: Result of using Euclidean distance scoring process on filter bank

just makes use of the higher 9 MFCC coefficients of the music database.

$\Delta$ MFCC and  $\Delta^2$ MFCC can also be computed using the higher MFCC coefficients using the equation 2.4. However, computing  $\Delta$ MFCC and  $\Delta^2$ MFCC increases complexity and so this thesis considers only the initial higher MFCC coefficients of the music database.

Comparing figure 6.1 with 6.2, it is fairly straight forward to see that MFCC show higher efficiency when deploying Euclidean distance scoring process. Furthermore, by computing  $\Delta$ MFCC and  $\Delta^2$ MFCC using the relation in equation 6.2 and including them in the database, the potency of MFCC should be able to reach a much greater efficiency as shown in [21].

$$\Delta_t MFCC = \frac{\sum_{i=1}^{N_t} i \cdot (C_{t+i} - C_{t-i})}{2 \cdot \sum_{i=1}^{N_t} i^2} \quad (6.2)$$

where,  $C_{t+i}$  refers to the MFCC coefficient in the next time frame,  $i + 1$  and  $C_{t-i}$  refers to the MFCC coefficient in the previous time frame,  $i - 1$ . Typical value for  $N_t$  could be 2.  $\Delta^2$ MFCC

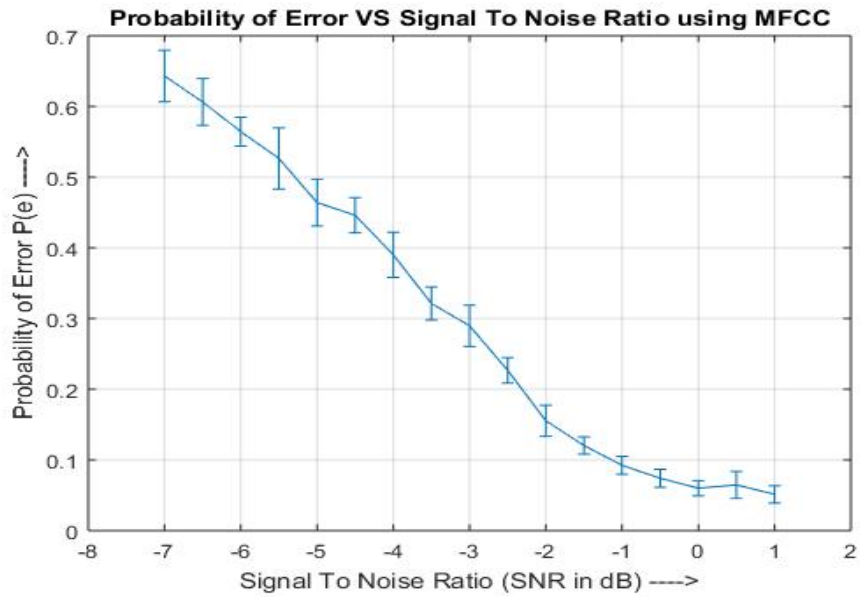


Figure 6.2: Result of using Euclidean distance scoring process on MFCC

can be generated using the same equation above using  $\Delta$ MFCC computed above.

## CHAPTER 7

### CONCLUSIONS AND FUTURE WORKS

#### 7.1 CONCLUSIONS

This thesis presented a comparative evaluation of peak - based approaches to audio fingerprinting along with an analysis on the value of employing MFCC in audio fingerprinting systems. While chapter 3 laid the foundations to this thesis, chapter 5 and 6 established what could be solutions to the research questions set forth in chapter 3.

The first and the most significant conclusion of this thesis is that, under set conditions, a single framework can be designed to evaluate all peak based audio fingerprinting approaches. Such a framework crucially helps establish a standard that can aid researchers and system designers to choose between same - frequency band and cross - frequency band audio fingerprinting approaches.

The second conclusion of this thesis is that it is possible to implement both cross - band and single - band approaches in a single evaluation framework by using the *double kill scoring process* defined in chapter 5. This conclusion could help researchers to tackle noise affecting the audio signals. This could prove necessary as current cross - frequency band based approaches tend to have more signatures in database, thereby increasing expense in scoring process.

The third conclusion of this thesis is that, even under the same design framework and same average number of peaks, cross -frequency band approach still show a better performance than same -frequency band approach. It was also determined that, there is an optimal number of frequency bands that could be employed in each signature being extracted. This conclusion is supported by the results depicted in figure 5.6.

The fourth and final conclusion of thesis is that, based on results depicted in chapter 6, MFCC audio fingerprinting approach shows better performance when compared to filter bank under scoring based on Euclidean distance. It should however be said that, filter bank showed similar performance as MFCC when using scoring based on peaks.

## 7.2 FUTURE WORKS

Although the closed set scenario has important applications, as described in section 1.1, an improvement to this thesis would be to consider an open set of music database. In other words, querying music that is not part of the original database. Currently the thesis considers a closed set database, wherein music being queried are part of the original music database. By allowing queries that are not part of the database, the analysis could be extended to other applications, such as music identification.

Another case to extend the thesis to real world problems would be, to consider employing queries that represent real world environments. Currently, this thesis makes use of white noise in the query algorithm before the scoring process. However, to evaluate real world audio fingerprinting systems using this approach, it would prove beneficial to deploy music data that are recorded from environments such as restaurants, schools, radio broadcasts,

television broadcasts, etc., using microphones that further add distortion to the query.

The third endeavour that could prove fruitful would be to consider deploying MFCC approach in the same framework as the traditionally prominent peak based approaches. MFCC have already proved to show some level of prudence in audio retrieval. If they could be somehow manipulated to efficiently deploy the double kill scoring process as the peak based approach in this thesis, it would be an interesting study and a new avenue for research in audio fingerprinting.

## REFERENCES

- [1] Wang, Avery. "An Industrial Strength Audio Search Algorithm." International Society for Music Information Retrieval(ISMIR). pp. 7-13. Oct. 2003.
- [2] Wang, Avery Li-Chun, and Julius O. Smith III. "System and methods for recognizing sound and music signals in high noise and distortion." U.S. Patent No. 6,990,453. 24 Jan. 2006.
- [3] Cheng Yang, "MACS: music audio characteristic sequence indexing for similarity retrieval," Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575), New Platz, NY, 2001, pp. 123-126. doi: 10.1109/ASPAA.2001.969558
- [4] Ellis, Daniel PW, Brian Whitman, and Alastair Porter. "Echoprint: An open music identification service." Proc. International Society for Music Information Retrieval(ISMIR). 28 Oct. 2011.
- [5] Whitman, Brian, Andrew Nesbit, and Daniel Ellis. "Musical fingerprinting." U.S. Patent No.8,492,633. 23 Jul. 2013.
- [6] Haitsma, Jaap, and Ton Kalker. "A highly robust audio fingerprinting system with an efficient search strategy." Journal of New Music Research 32.2 (2003): 211-221.
- [7] Haitsma, Jaap, and Ton Kalker. "A highly robust audio fingerprinting system." International Society for Music Information Retrieval(ISMIR). Vol. 2002. pp. 107-115. Oct. 2002.
- [8] Cano, Pedro, et al. "A review of algorithms for audio fingerprinting." Multimedia Signal Processing, 2002 IEEE Workshop on. IEEE, 2002.
- [9] Zhang, Tong. "System and method for spectrogram analysis of an audio signal." U.S. Patent Application No. 10/465,640.
- [10] MATLAB. (2012). MATLAB (Version 8.0) and Statistics Toolbox (Version 8.1) [Software]. Natick, MA: The MathWorks, Inc. Available from <http://www.mathworks.com/products/matlab>
- [11] Singh, Parwinder Pal, and Pushpa Rani. "An Approach to Extract Feature using MFCC." IOSR Journal of Engineering 4.8. pp. 21-25. Aug. 2014.



- [12] Porter, Alastair. Evaluating musical fingerprinting systems. PhD. Diss. McGill University, 2012.
- [13] Ittichaichareon, Chadawan, Siwat Suksri, and Thaweesak Yingthawornsuk. "Speech recognition using MFCC." International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012) July. 2012.
- [14] Guo, Yina, et al. "Optimized phase-space reconstruction for accurate musical-instrument signal classification." Multimedia Tools and Applications (2016): 1-19.
- [15] Muda, Lindasalwa, Mumtaj Begam, and Irraivan Elamvazuthi. "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques." arXiv preprint arXiv:1003.4083 (2010).
- [16] Bilobrov, Sergiy. "Extraction and matching of characteristic fingerprints from audio signals." U.S. Patent No. 7,516,074. 7 Apr. 2009.
- [17] Fierrez, Julian, et al. "HMM-based on-line signature verification: Feature extraction and signature modeling." Pattern recognition letters 28.16 (2007): 2325-2334.
- [18] Batlle, Eloi, et al. "Scalability issues in an HMM-based audio fingerprinting." Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on. Vol. 1. IEEE, 2004.
- [19] Moussallam, Manuel, and Laurent Daudet. "A general framework for dictionary based audio fingerprinting." Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014.
- [20] Hossan, Md Afzal, Sheeraz Memon, and Mark A. Gregory. "A novel approach for MFCC feature extraction." Signal Processing and Communication Systems (ICSPCS), 2010 4th International Conference on. IEEE, 2010.
- [21] Thakur, Akanksha Singh, and Namrata Sahayam. "Speech recognition using Euclidean distance." International Journal of Emerging Technology and Advanced Engineering 3.3 (2013): 587-590.

## APPENDIX A

## AUDIO FINGERPRINTING PROCESS IN THESIS

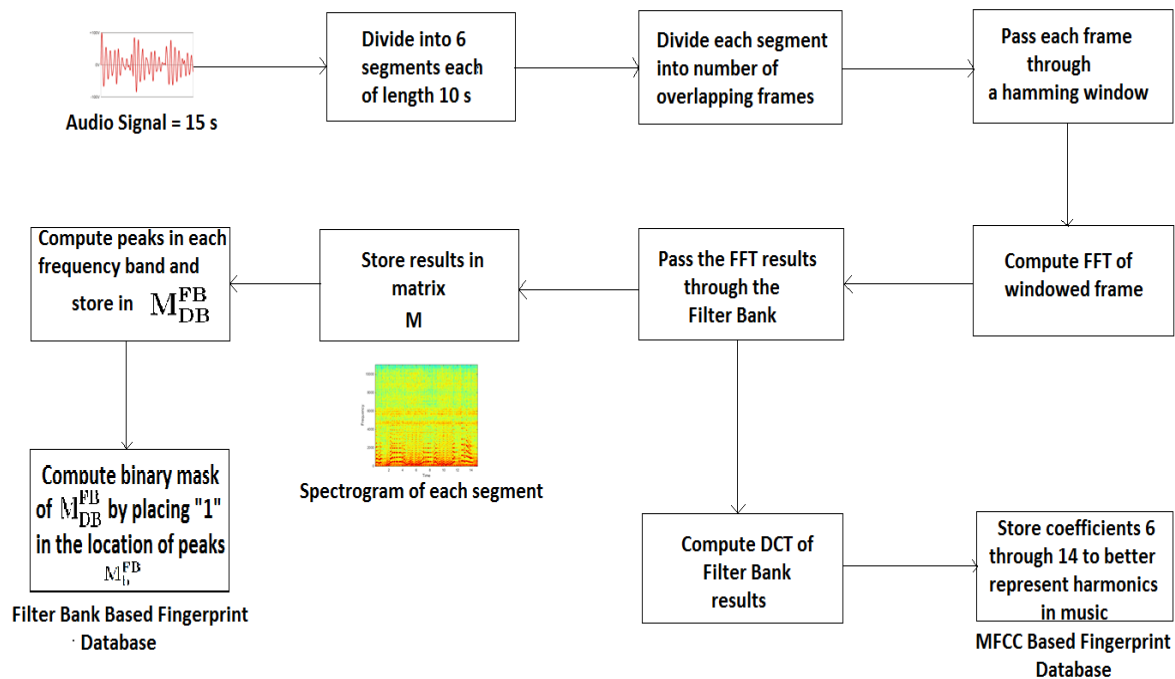


Figure 1: Database Generation

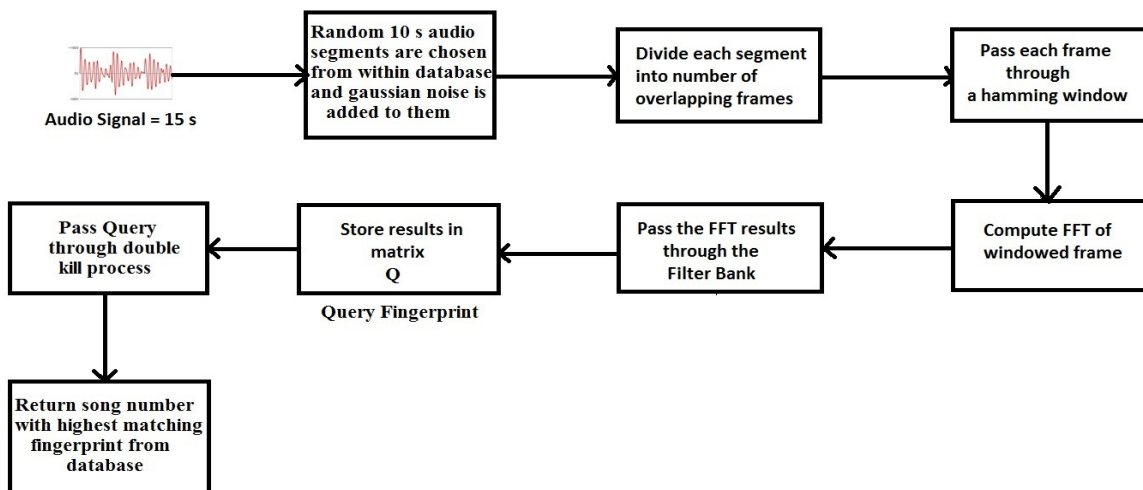


Figure 2: Peak Based Approach Scoring Process

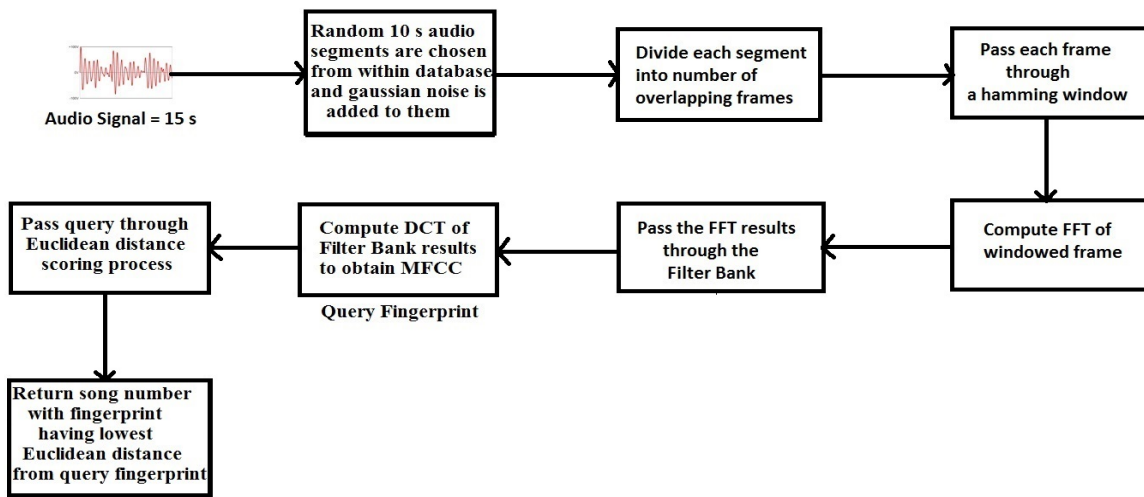


Figure 3: MFCC Based Approach Scoring Process

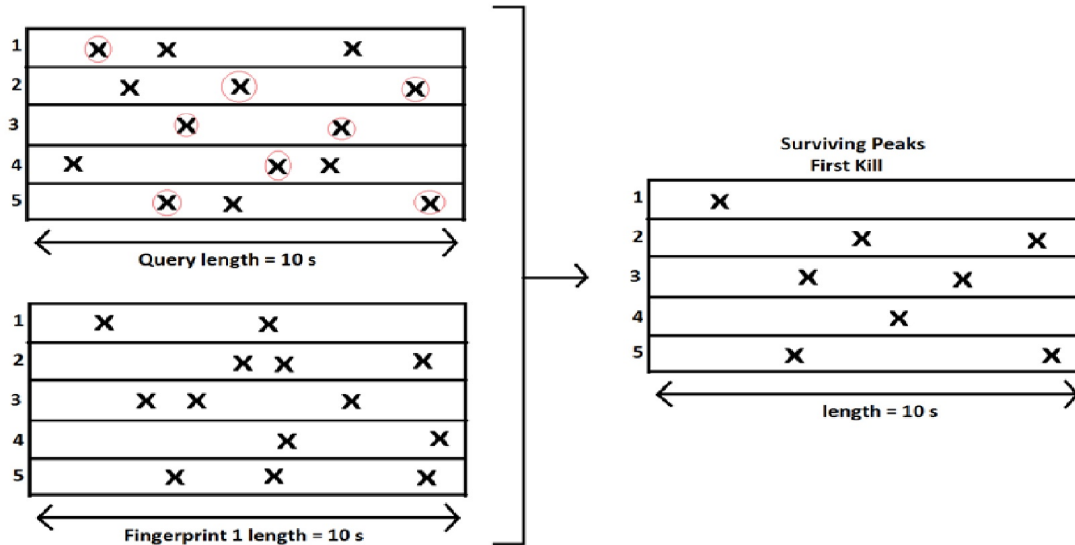


Figure 4: Noise Cancellation In Double Kill Process

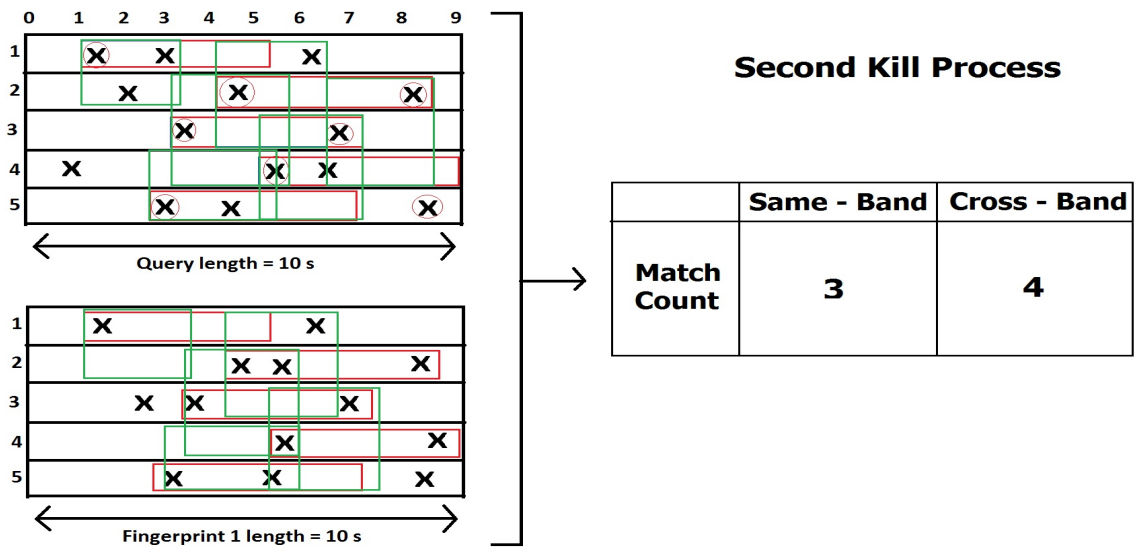


Figure 5: Fingerprint Scoring

## APPENDIX B

## COMPUTATIONAL EXPENSE EVALUATION

Figure 7.6 represents the time expense for same-frequency band analysis while figures 7.7 and 7.8 represent time expense for cross-frequency band analysis using the same sized time frame and figures 7.9 and 7.10 represent the same cross-frequency band analysis using the varying sized time frame deployed in this thesis. The function `fing_match` is used to match the query fingerprint with the fingerprint in the database. The time expense of this function is referenced in table 5.1.






Children (called functions)					
Function Name	Function Type	Calls	Total Time	% Time	Time Plot
<a href="#">fing_match</a>	function	1	2.017 s	77.5%	
<a href="#">raw_fingerprint</a>	function	1	0.192 s	7.4%	
<a href="#">fpks</a>	function	1	0.092 s	3.5%	
<a href="#">wavread</a>	function	1	0.079 s	3.0%	
<a href="#">close</a>	function	1	0.009 s	0.3%	
<a href="#">strcat</a>	function	1	0.008 s	0.3%	
<a href="#">std</a>	function	1	0.006 s	0.2%	
<a href="#">mean</a>	function	2	0.005 s	0.2%	
Self time (built-ins, overhead, etc.)			0.194 s	7.5%	
Totals			2.602 s	100%	

Figure 6: Time Expense For Querying Using Same-Frequency Band Approach In This Thesis

**Children** (called functions)






Function Name	Function Type	Calls	Total Time	% Time	Time Plot
<a href="#">fing_match</a>	function	1	3.458 s	85.6%	
<a href="#">raw_fingerprint</a>	function	1	0.197 s	4.9%	
<a href="#">fpks</a>	function	1	0.095 s	2.4%	
<a href="#">wavread</a>	function	1	0.075 s	1.9%	
<a href="#">strcat</a>	function	1	0.009 s	0.2%	
<a href="#">std</a>	function	1	0.006 s	0.1%	
<a href="#">close</a>	function	1	0.006 s	0.1%	
<a href="#">mean</a>	function	2	0.004 s	0.1%	
Self time (built-ins, overhead, etc.)			0.192 s	4.8%	
Totals			4.042 s	100%	

Figure 7: Time Expense For Querying Using 7 Cross-Frequency Band Approach In Thesis

**Children** (called functions)






Function Name	Function Type	Calls	Total Time	% Time	Time Plot
<a href="#">fing_match</a>	function	1	5.071 s	89.9%	
<a href="#">raw_fingerprint</a>	function	1	0.189 s	3.3%	
<a href="#">fpks</a>	function	1	0.095 s	1.7%	
<a href="#">wavread</a>	function	1	0.074 s	1.3%	
<a href="#">std</a>	function	1	0.010 s	0.2%	
<a href="#">close</a>	function	1	0.008 s	0.1%	
<a href="#">strcat</a>	function	1	0.007 s	0.1%	
<a href="#">mean</a>	function	2	0.005 s	0.1%	
Self time (built-ins, overhead, etc.)			0.184 s	3.3%	
Totals			5.643 s	100%	

Figure 8: Time Expense For Querying Using All Cross-Frequency Band Approach In Thesis



**Children** (called functions)






Function Name	Function Type	Calls	Total Time	% Time	Time Plot
<a href="#">fing_match</a>	function	1	2.138 s	79.0%	
<a href="#">raw_fingerprint</a>	function	1	0.189 s	7.0%	
<a href="#">fpks</a>	function	1	0.094 s	3.5%	
<a href="#">wavread</a>	function	1	0.075 s	2.8%	
<a href="#">close</a>	function	1	0.008 s	0.3%	
<a href="#">strcat</a>	function	1	0.007 s	0.3%	
<a href="#">std</a>	function	1	0.006 s	0.2%	
<a href="#">mean</a>	function	2	0.003 s	0.1%	
Self time (built-ins, overhead, etc.)			0.187 s	6.9%	
<b>Totals</b>			<b>2.707 s</b>	<b>100%</b>	

Figure 9: Time Expense For Querying Using 7 Cross-Frequency Band Approach In Thesis

**Children** (called functions)






Function Name	Function Type	Calls	Total Time	% Time	Time Plot
<a href="#">fing_match</a>	function	1	2.173 s	79.1%	
<a href="#">raw_fingerprint</a>	function	1	0.188 s	6.8%	
<a href="#">fpks</a>	function	1	0.093 s	3.4%	
<a href="#">wavread</a>	function	1	0.082 s	3.0%	
<a href="#">close</a>	function	1	0.009 s	0.3%	
<a href="#">strcat</a>	function	1	0.008 s	0.3%	
<a href="#">std</a>	function	1	0.006 s	0.2%	
<a href="#">mean</a>	function	2	0.004 s	0.1%	
Self time (built-ins, overhead, etc.)			0.183 s	6.7%	
<b>Totals</b>			<b>2.746 s</b>	<b>100%</b>	

Figure 10: Time Expense For Querying Using All Cross-Frequency Band Approach In Thesis