

**REPRODUCE SOUND THROUGH
SUBTLE VIBRATIONS IN THE IMAGERY**

ONG EU-JEEN

**A project report submitted in partial fulfilment of the
requirements for the award of Bachelor of Engineering
(Hons.) Mechatronics Engineering**

**Lee Kong Chian Faculty of Engineering and Science
Universiti Tunku Abdul Rahman**

May 2016

DECLARATION

I hereby declare that this project report is based on my original work except for citations and quotations which have been duly acknowledged. I also declare that it has not been previously and concurrently submitted for any other degree or award at UTAR or other institutions.

Signature : _____

Name : Ong Eu-Jeen

ID No. : 12UEB06811

Date : 29/08/2016

APPROVAL FOR SUBMISSION

I certify that this project report entitled “**REPRODUCE SOUND THROUGH SUBTLE VIBRATIONS IN THE IMAGERY**” was prepared by **ONG EU-JEEN** has met the required standard for submission in partial fulfilment of the requirements for the award of Bachelor of Engineering (Hons.) Mechatronics Engineering at Universiti Tunku Abdul Rahman.

Approved by,

Signature : _____

Supervisor : _____

Date : _____

The copyright of this report belongs to the author under the terms of the copyright Act 1987 as qualified by Intellectual Property Policy of Universiti Tunku Abdul Rahman. Due acknowledgement shall always be made of the use of any material contained in, or derived from, this report.

© 2016, Ong Eu-Jeen. All right reserved.

ACKNOWLEDGEMENTS

I would like to thank everyone who had contributed to the successful completion of this project. I would like to express my gratitude to my research supervisor, Dr. Tee Yee Kai for his invaluable advice, guidance and his enormous patience throughout the development of the research.

In addition, I would also like to express my gratitude to my loving parents and friends who had helped and given me encouragement throughout the process of undertaking this research. I would like to show my appreciation towards Leong Chee Yong for his advices and suggestions as his research has similar concepts and theories and Jonathan Robert Tsen Tze Kian who is undertaking another research for much information has been shared and discussed.

REPRODUCE SOUND THROUGH SUBTLE VIBRATIONS IN THE IMAGERY

ABSTRACT

Subtle vibrations can be produced on the object's surface when sound hits the object. This research intends to recover sound that produces these subtle vibrations at the surface of the object by means of a video footage of the object of interest. A high speed video footage is preferable but the best available video camera will be used. The algorithm is based on the Eulerian phase-based approach of magnifying and visualising the small changes in motion and colour in videos which is hard to be perceived by the human naked eye. Using the same approach, phase information regarding these small changes in motion or also known as subtle vibrations can be utilised to recover sound. With this new remote sound acquisition technique introduced, many of the objects in normal surroundings can be turned into visual microphones. The algorithm constructed in this research has successfully recovered sound from a few objects such as guitar strings, diaphragm of a speaker, plastic bag and a bag of chips. The input of sounds are from guitar strings, a frequency sweep, self-constructed audio, amplified bass components of songs and piano music. The quality of the recovered sound is analysed by qualitatively assessing the audibility of the sound and also by evaluating visually the various graphs plotted such as the spectrogram, periodogram and the displacement graph of both the original and recovered sound. Several control factors are studied to understand the limitations of the algorithm and to identify ways of improving the algorithm. They are the resolution of camera, volume level of speaker, region of interest, motion magnification and digital filtering. The capability of the algorithm is still limited by the frame rate of the camera, the type of object used and the complexity of the input of sound.

TABLE OF CONTENTS

DECLARATION	ii
APPROVAL FOR SUBMISSION	iii
ACKNOWLEDGEMENTS	v
ABSTRACT	vi
TABLE OF CONTENTS	vii
LIST OF TABLES	xi
LIST OF FIGURES	xiii
LIST OF SYMBOLS / ABBREVIATIONS	xvi

CHAPTER

1	INTRODUCTION	1
	1.1 Background	1
	1.2 Aims and Objectives	2
	1.3 Motivation in Improving Surveillance Systems	3
2	LITERATURE REVIEW	6
	2.1 Theory and Characteristics of Sound	6
	2.1.1 General Information	6
	2.1.2 Sound Range of the Human Ear	7
	2.2 Traditional and Laser Microphones	9
	2.2.1 Traditional Microphone	9
	2.2.2 Laser Microphone	10
	2.3 Visualisation of Temporal Variations in Video	12
	2.3.1 Motion and Colour Magnification	12

	2.3.2	Eulerian vs. Lagrangian Perspectives	13
	2.3.3	Eulerian Video Magnification	14
2.4		Space-Time Video Processing	17
	2.4.1	Spatial Processing	17
	2.4.2	Temporal Processing	18
	2.4.3	Digital Image Processing	20
2.5		Computation of Motion Signal into Audio Signal	21
	2.5.1	Local Motion Signals	21
	2.5.2	Global Motion Signal	22
	2.5.3	Sound Recovery from Various Objects	23
2.6		Audio Denoising	25
2.7		Rolling Shutter Technique	26
3		METHODOLOGY	28
	3.1	Applied Theories	28
		3.1.1 Pre-processing Technique	28
		3.1.2 Process Algorithm	29
		3.1.3 Denoising Methods	29
		3.1.4 Program Flow	30
	3.2	Source of Sound	32
		3.2.1 Acoustic Guitar Strings	32
		3.2.2 Frequency Sweep Sound	33
		3.2.3 Self-Constructed Audio	34
		3.2.4 Amplified Bass Components of Songs	35
		3.2.5 Piano Sound	37
	3.3	Source of Object	37
	3.4	Sound Recovery Experimental Setup	39
		3.4.1 Workstation Specification	39
		3.4.2 Acoustic Guitar Strings	39
		3.4.3 Diaphragm (Resolution)	41
		3.4.4 Diaphragm (Volume, Digital Filtering, ROI)	42
		3.4.5 Plastic Bag (Volume, Motion Magnification)	44
		3.4.6 High Speed Video of Bag of Chips	46

3.5	Experimental Limitations	47
3.5.1	Frame Rate of Video Camera	47
3.5.2	Pixel Resolution of Video Camera	48
3.5.3	Video Camera Zoom Factor	48
4	RESULTS AND DISCUSSION	49
4.1	Preliminary Results – 40Hz Bass Sound (Water Ripple)	49
4.2	Experimental Results	52
4.2.1	Acoustic Guitar Strings	52
4.2.2	Diaphragm of Speaker	54
4.2.3	Plastic Bag	57
4.2.4	High Speed Video of Bag of Chips	60
4.3	Interpretation of Graphs	60
4.3.1	Displacement vs. Time	60
4.3.2	Power Spectral Density Estimate	61
4.3.3	Spectrogram	63
4.3.4	Waterfall Plot	64
4.4	Analysis of Control Factors	65
4.4.1	High Resolution vs. Low Resolution	65
4.4.2	Region of Interest (ROI)	66
4.4.3	Digital Filtering	68
4.4.4	Volume Level of Speaker	70
4.4.5	Implementation of Motion Magnification	72
4.5	Highlighted Challenges	73
4.5.1	Brightness of Video	73
4.5.2	Adjusting Volume Level of Speaker	74
4.5.3	Induced Vibration on Camera	74
4.5.4	Distance between Camera and Object	75
4.5.5	Environmental Effects	75
4.5.6	High Data Storage	76
4.5.7	Determination of Quality of Sound	76
5	CONCLUSION AND RECOMMENDATIONS	77

5.1	Conclusion	77
5.2	Recommendations	78

REFERENCES	80
-------------------	-----------

LIST OF TABLES

TABLE	TITLE	PAGE
1.1	Survey on the support of surveillance cameras of Americans	4
2.1	Difference between Eulerian vs. Lagrangian processing	13
2.2	Difference between Linear and Phase-Based Approach	16
2.3	Performance of Frequency Response of Various Objects	24
3.1	Guitar String Key Frequency	32
3.2	Selected and Filtered Songs	36
3.3	Source of Object	38
3.4	Summary of the Diaphragm(Resolution) Experimental Conditions	41
3.5	Summary of the Diaphragm(V,DF,ROI) Experimental Conditions	42
3.6	Summary of the Plastic Bag(V,MM) Experimental Conditions	45
4.1	Frequency Sweep Diaphragm Results	54
4.2	Self-Constructed Audio (Simple) Diaphragm Results	54
4.3	Self-Constructed Audio (Complex) Diaphragm Results	55
4.4	‘Justin Bieber-What Do You Mean?’ Bass Diaphragm Results	55

4.5	‘Black Eyed Peas Ft Justin Timberlake - Where Is The Love?’ Bass Diaphragm Results	56
4.6	‘Macklemore & Ryan Lewis- Downtown’ Bass Diaphragm Results	57
4.7	Frequency Sweep Plastic Bag Results	58
4.8	Self-Constructed Audio (Simple) Plastic Bag Results	58
4.9	Self-Constructed Audio (Complex) Plastic Bag Results	59
4.10	‘Black Eyed Peas Ft Justin Timberlake - Where Is The Love?’ Bass Plastic Bag Results	59
4.11	‘Mary had a little lamb’ Piano Music Bag of Chips Results	60
4.12	High Resolution vs. Low Resolution Results	66
4.13	ROI Comparison Results	67
4.14	Median Filtering Comparison Results	69
4.15	Optimum Volume Level of Speaker	70
4.16	Volume Level of Sound Comparison Results	71
4.17	Motion Magnification Comparison Results	73
5.1	Summary of Control Factors	78

LIST OF FIGURES

FIGURE	TITLE	PAGE
1.1	Number of CCTV cameras per 1,000 people	5
2.1	Basic Properties of Sound	7
2.2	Sound Pressure Levels	8
2.3	Frequency Range of Human Voice and Speech Formant	8
2.4	Dynamic Microphone	10
2.5	Laser Vibrometer Schematic	11
2.6	PPLV Schematic	12
2.7	Eulerian Linear Approach Framework	15
2.8	Eulerian Phase-Based Approach Framework	16
2.9	Levels in the Pyramid	17
2.10	Amplitude Spectrum of Ideal Band-Pass Filter	19
2.11	Butterworth and Chebyshev Band-Pass Filter	19
2.12	3x3 square kernel for mean filter	20
2.13	3x3 square kernel for median filter	21
2.14	Sound Recovery Experiment Setup	23
2.15	Motion vs. Sound Volume	24
2.16	Rolling Shutter Properties	27
3.1	Program Flowchart	31

3.2	Frequency Sweep (30Hz to 200Hz)	33
3.3	Self-Constructed Audio (Simple Version)	34
3.4	Self-Constructed Audio (Complex Version)	35
3.5	Acoustic Guitar Strings Experiment Setup	40
3.6	Camera View of Guitar Strings	40
3.7	Diaphragm(Resolution) Experiment Setup	41
3.8	Camera View of Diaphragm(Resolution)	42
3.9	Diaphragm(V,DF,ROI) Experiment Setup	43
3.10	Camera View of Diaphragm(V,DF,ROI)	44
3.11	Plastic Bag(V,MM) Experiment Setup	46
3.12	Camera View of Plastic Bag(V, MM)	46
3.13	Camera View of Bag of Chips	47
4.1	Water Ripple Experiment Setup	50
4.2	Amplitude Graph 40 Hz (a)Original (b)Recovered (c)Filtered	50
4.3	Power Spectral Density Graph 40 Hz (a)Original (b)Recovered (c)Filtered	51
4.4	Observed Guitar Sound from Video	52
4.5	Guitar Strings Frequency Spectrum	53
4.6	Guitar Strings Power Spectral Density Estimate	53
4.7	Displacement vs Time Graph	61
4.8	Power Spectral Density Estimate Graph	62
4.9	Power Spectral Density Estimate120Hz Limit Graph	62
4.10	Spectrogram	63
4.11	Spectrogram 120Hz Limit Graph	64
4.12	Waterfall Plot	65

4.13	ROI 1	67
4.14	ROI 2	67
4.15	Median Filtering (a) with (b) without	69

LIST OF SYMBOLS / ABBREVIATIONS

f_{max}	Maximum Frequency Input, Hz
ϕ	Local Phases
ϕ_v	Phase Variations
x	Row
y	Column
r	Scale
θ	Orientation
t	Frame Number
CCTV	Closed Circuit Television
PPLV	Photo-EMF Pulsed Laser Vibrometer
fps	Frames per second
ROI	Region of Interest
DIP	Digital Image Processing
SSNR	Segmental Signal-to-Noise Ratio
LLR	Log Likelihood Ratio

CHAPTER 1

INTRODUCTION

1.1 Background

Sound can be defined as the variation or fluctuation of air pressure in the air. Depending on the medium where sound travels, sound being compressed and expanded causes molecules to vibrate back and forth to generate a wave that is able to transfer sound energy. When sound hits on any surface or object, the surface or object produces a subtle vibration. The vibration mode varies with different objects but the pattern of its motion may be analysed and used to recover sound.

Recovering sound through imagery may be quite a recent technology, but there have been approaches on acquiring sound remotely such as the laser microphone which emits a laser beam projecting towards the vibrating surface of the object. The difference is that most of the existing approaches are of active nature whereas the method under this research approaches in a passive nature. By active nature, it means to apply an active illumination in the process of recovering sound.

An important element that has been taken into consideration is that most of the subtle vibrations of the object caused by sound provide enough visual signals to partially recover them. The essential criteria required to recover the sound is as basic as having a high speed video on the particular object. In further research, Davis, A., et al. (2014) stated that a normal video camera can recover sound as well using a technique which manipulates the camera's rolling shutter properties to increase sampling rate and recover sound frequencies beyond the camera's frame rate.

In brief, the method used in this research is to use a high-speed video or normal video camera to record videos of the object of interest that has been induced with subtle vibrations due to sound. Inputting these videos into the algorithm, the information regarding these subtle vibrations may be converted back into its audio signal. Such technique opens up more possibilities that our everyday surrounding objects can become potential microphones. A short term to describe this method for remote sound acquisition is also called the visual microphone. This microphone could possibly be implemented in surveillance systems in future.

1.2 Aims and Objectives

The main aim of undergoing this research project is to discover a new alternative in recovering and recording sound. With such discovery, many potential applications and possibilities are opened up such as implementing them in surveillance systems. Three internal objectives are to be carried out throughout the research project in order to achieve and extend the capabilities corresponding to the stated goal.

Recover audible sound through imagery. This is the general and minimal task to be accomplished in the project. It acts as a proof of concept to show that the method and the underlying principles are valid. The task is to recover any sound input from video and converting into sound output with a similar sound as the input. Similarity of the sound can be determined by qualitatively accessing the audibility of the sound and analysing the patterns of the amplitude and frequency spectrum of the sound output.

Improve similarity between recovered and original sound. The idea is to discover techniques and solutions to improve the similarity between the recovered and the original sound. This serves to increase the accuracy and quality of the sound recovered. Aspects such as analysing the frequency spectrum and amplitude vs. time relationship may be useful tools in finding ideas and solutions. Noise filtering would indefinitely improve the quality of the sound output.

Recover more sophisticated sound up to human speech. The research will start off by recovering simple sounds and then increasing the complexity of the sound. By simple sound, it means to be consisting of only a single or few frequencies at each instantaneous moment. As the sound complexity increases, the sound may become more significant and meaningful to be recovered. The ultimate accomplishment is to recover human speech which important message may be recognised and retrieved. Other sounds could be such as the human heartbeat, music, human snore, and water dripping. The approach is to start recovering the simplest sound possible and then proceed to more sophisticated sound, step by step.

1.3 Motivation in Improving Surveillance Systems

One of the motivations that inspire this research is to improve the existing surveillance system and eventually reducing the crime rate. The clearest path would be to apply on CCTV cameras. CCTV is an abbreviation for closed-circuit television, which is the use of video cameras that transmit signal to a set of monitors. Most video cameras fit into this definition but the main difference is that the term CCTV is commonly associated with video surveillance and monitoring of various places of interest.

Most existing CCTVs only provide video monitoring without any audio input or microphone. This opens up the opportunity for the algorithm of recovering sound through imagery where it gives user the ability to recover sound if necessary despite not having any audio input. The algorithm gives an added on value as it could be easily implemented on existing systems and provides another alternative in recovering sound. Besides the ease of implementation on existing systems, audio can be recovered only when desired and reduces the need of storage as a prerequisite as compared to the conventional built-in microphone which promotes flexibility in the method.

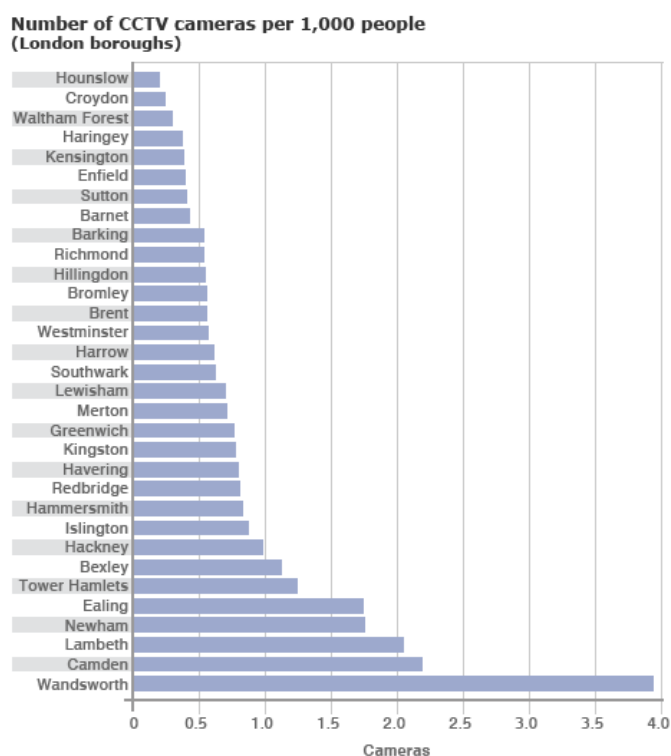
Another interesting application that traditional microphones are limited to is that the algorithm is theoretically able to recover sound from another room and perhaps an enclosed one as long as the visual sight of the room can be captured in footage. This is because the algorithm depends on the visual information even when no sound can be heard from the camera's position. The placement of the traditional microphones becomes an important criterion for conventional CCTV cameras to obtain the best quality and built-in microphones may not always have the best positioning for both visual and audio monitoring.

Looking into some of the statistics around the world, the British Security Industry (BSIA, 2013) estimated that there were 5.9 million CCTV cameras in Britain which leads to approximately 1 CCTV camera for 11 people in Britain at the maximum possibility. Without surprise, CCTV surveillance is also increasing in the United States of America. Cases such as the Boston bombing, Rodney King being assaulted by the police and the attacks on the World Trade Center has been aided by CCTV camera footage (BBC News Magazine, 2013). A survey conducted by HuffPost/YouGov on the April 22 and 23, 2013 shows the support on surveillance cameras from the American people.

Table 1.1: Survey on the support of surveillance cameras of Americans

<i>Americans</i>	<i>Decision</i>
<i>40%</i>	Need more surveillance cameras
<i>43%</i>	Number of cameras is about right
<i>12%</i>	Should have fewer cameras

Malaysia is also not far behind in the demand for increased surveillance cameras. Dewan Bandaraya Kuala Lumpur (2013) stated that there is plan to set up a total of 1200 CCTV cameras to monitor the traffic around Kuala Lumpur with an estimated cost of RM200 million. Narrowing down the scale, Sunway Pyramid has 400 CCTV cameras installed at strategic locations around the shopping centre for safety reasons (Chan, H. C., Chief Executive Officer of Sunway Pyramid Sdn. Bhd, 2012). Continuing on, the statistics for the number of CCTV cameras in cities of Britain is obtained from the BBC's Newsnight programme. It also mentions that Britain appears to be the highest number of CCTV cameras than any other countries.



**Figure 1.1: Number of CCTV cameras per 1,000 people
(Source: BBC News, 2012)**

From all these statistics given, it is apparent that the demand for an increased surveillance system is crucial and growing. One restriction to be taken care of is the law and regulations regarding audio recording in public places. It is common that the public should be notified that there is audio recording in the premise.

On top of that, another less common application but would often come into mind when using the algorithm for surveillance is the use of drones or as a cutting edge military device. Recently, Nadirah, H. R. and Natasha, J. (2016) written in the STAR news that police drones will be implemented as a crime prevention initiative. The drones would be the third unit in addition to the Motorcycle Patrol Unit (MPU) and Mobile Patrol Vehicle (MPV). The algorithm may work alongside with this new step taken. As drones normally have a certain distance away from their targets of interest, it would be much less practical to implement a built-in microphone on the drone. This is when the algorithm comes into place where it recovers sound from imagery itself. One criterion is that the visual signal has to be of top notch.

CHAPTER 2

LITERATURE REVIEW

2.1 Theory and Characteristics of Sound

2.1.1 General Information

As mentioned in the introduction, sound is the product of the vibrations and change of pressure in air. The speed of sound is dependent on the medium it is transmitted through and commonly the listening of human speech occurs through air with the speed of approximately 340 metres per second. Sound waves are also referred to as pressure waves due to the repeating pattern of high and low pressure regions and its amplitude is often measured in terms of pressure.

The simplest sound or type of vibration is known as the pure tone, where only a single frequency is present in the sinusoidal vibration. These vibrations are under the study of the simple harmonic vibrations. Figure 2.1 describes a simple sound waveform and basic properties of the wave.

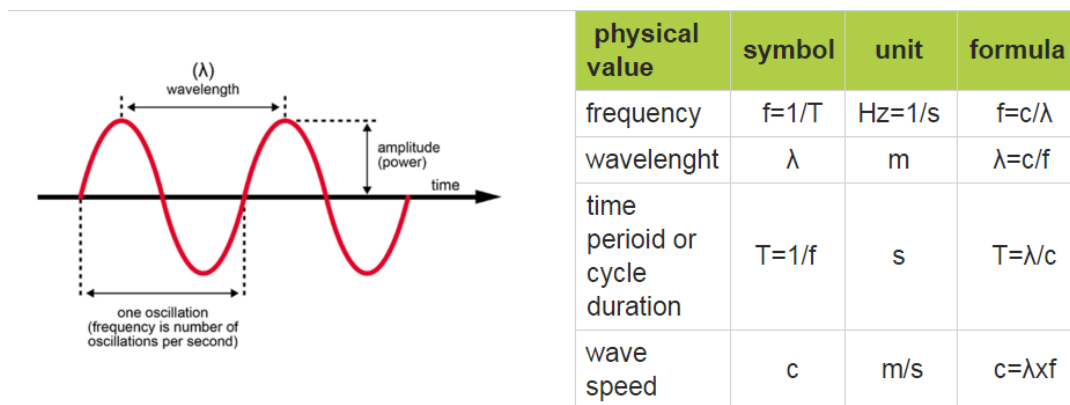


Figure 2.1: Basic Properties of Sound

(Source: Paroc)

2.1.2 Sound Range of the Human Ear

The human ear responds to sound pressure which has measurement units of Pascal. It is usually expressed in dB (logarithmic scale) because of such an extensive range in pressure. The power and intensity range that can be heard ranges from 0 dB to 120 dB. The upper limit is the threshold for pain and could contribute to permanent reduced hearing if exceeded. Figure 2.2 shows some examples of sources of sound with their respective sound pressure level.

For the frequency range that a standard healthy human ear can hear is from the range of 20 Hz to 20k Hz. That being said, the sensitivity of the ear towards this range differs and peaks at 3-4k Hz. The human ear is generally more sensitive around the frequency spectra of speech. The range of the human speech can be understood much clearer in Figure 2.3.

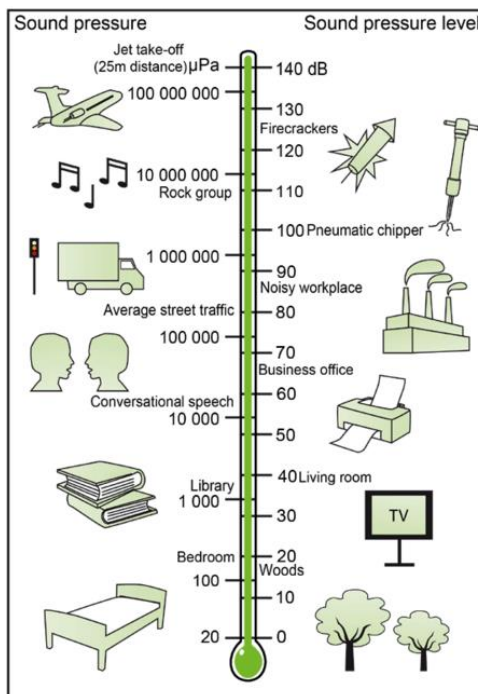


Figure 2.2: Sound Pressure Levels
(Source: Paroc)

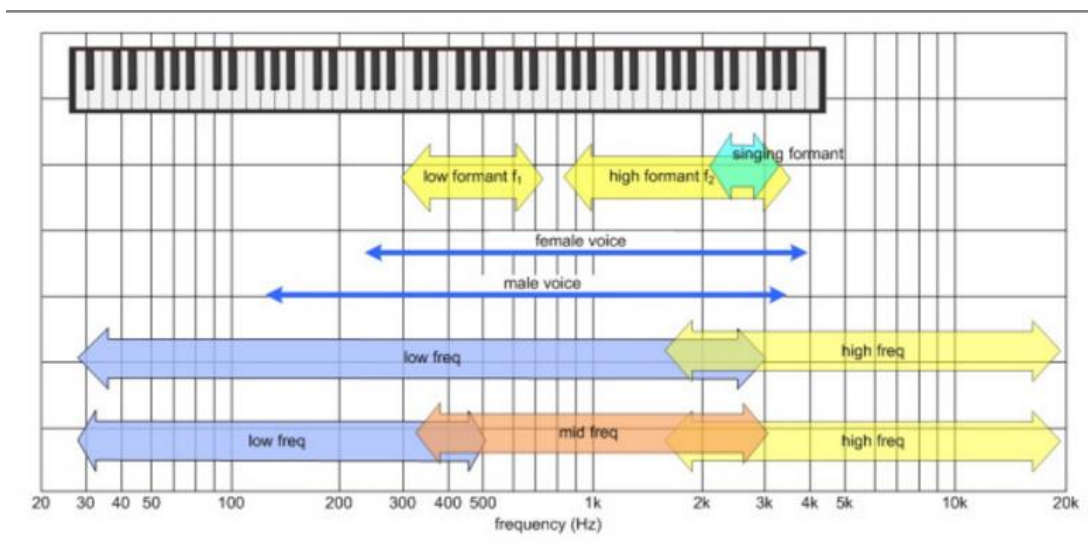


Figure 2.3: Frequency Range of Human Voice and Speech Formant
(Source: www.bnoack.com)

From Figure 2.3, it is observed that the typical frequency range of the male and female voice starts from about 150 Hz to over 3k Hz. The speech is characterized by formants. Speech formants describe the areas in the frequency

spectrum where they are energetically higher than the average energy. Only the first two formants are considered for vowel recognition. The lower speech first formant ranges from 300Hz to 750 Hz and the higher speech second formant ranges from 900Hz to 3000Hz. This range of frequencies would be of the project's interest to recover sound depending on its application.

The Nyquist Theorem. The Nyquist theorem is also known as the sampling theorem which states that for a given analog signal of f_{\max} , the sampling rate must be at least $2f_{\max}$ or higher in order for all the frequencies of the analog input signal to be correctly represented in the digitized output. This theorem plays a very important role in the algorithm which also becomes a limitation on the capability of the algorithm to recover sound input fully. The limitation may arise from the hardware itself which would be the sampling rate of the video camera. This problem will be put into thoughts in the later sections.

2.2 Traditional and Laser Microphones

2.2.1 Traditional Microphone

Traditional microphones are described because they have the basis concept of this research, where they detect sound signals, convert and reproduce them digitally. Microphones are known as transducers where sound signals are detected and transformed into electrical signal. There are various types of microphones such as dynamic, ribbon, condenser, crystal and electret condenser microphones. Only the dynamic microphone will be described in this report as an example on the basic mechanism of microphones.

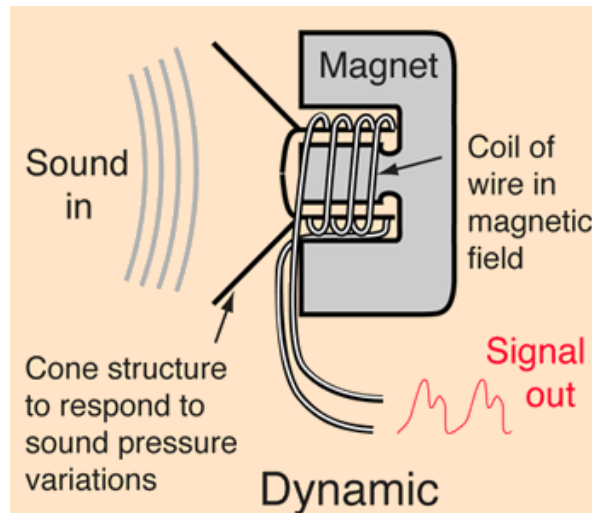


Figure 2.4: Dynamic Microphone

(R. Nave)

The cone acts as the diaphragm of the microphone. Its role is to move correspondingly to the sound input. This movement moves the coil of wire in the magnetic field and consequently generates a voltage which follows the variation in the sound pressure. The geometry of the dynamic microphone is similar with a tiny loudspeaker but works the opposite. The loudspeaker works by converting electrical signal to sound instead.

2.2.2 Laser Microphone

Laser microphones also work using a similar principle with the dynamic microphone to record sound. Instead, it uses the object of interest as the diaphragm by recording the motion and vibration through the reflection of the laser aimed on the surface of the object.

Rothberg, S.J. et al. (1989) introduced the laser Doppler vibrometer which measures the normal-to-surface vibration of an object. With its ability to determine the phase of the reflected laser, sound can be recovered from the object of interest. The principle of operation is that if there is a detection of a Doppler shift, frequency

of the vibration and thus the laser can be determined with the implementation of a reference beam and heterodyned on the detector surface since the photodetector would not be able to respond quick enough to detect the laser frequency.

Chen, C. W. et al. (2008) introduced a high sensitivity pulsed laser vibrometer to improve the laser microphone. This development is produced using the combination of photo-EMF sensors and pulsed light sources which proves a high sensitivity of 75 pm in terms of surface displacement. The similarity of the laser microphone with this research project, visual microphone is that both use a distant object as the diaphragm of its microphone. The difference arises where an active laser beam needs to be projected while this project only needs the visual input from the video. Figure 2.5 will display the schematic of a typical laser vibrometer while Figure 2.6 displays the schematic of the Photo-EMF Pulsed laser vibrometer (PPLV).

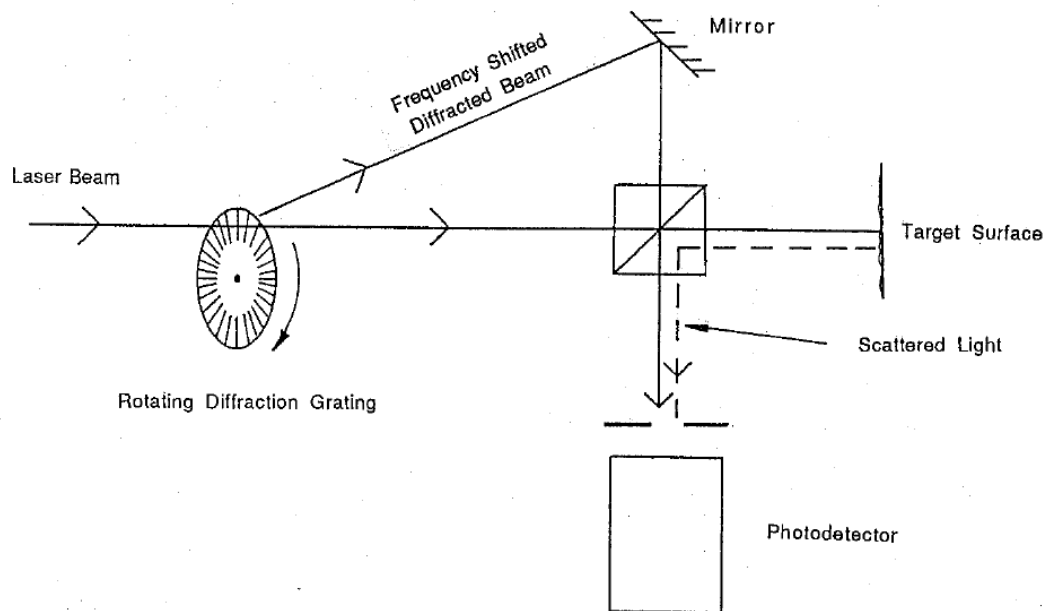


Figure 2.5: Laser Vibrometer Schematic
(Rothberg et al., 1989)

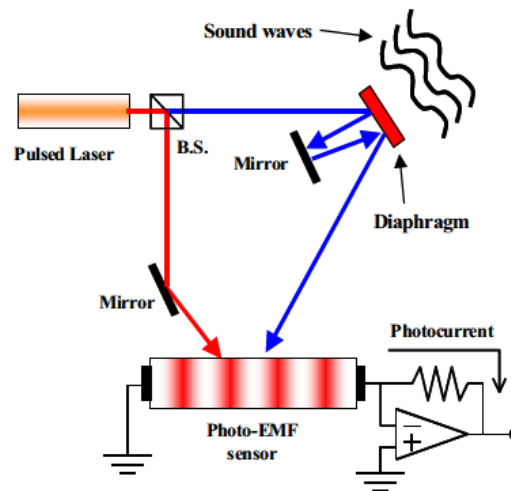


Figure 2.6: PPLV Schematic
(Chen, C. W. et al., 2008)

2.3 Visualisation of Temporal Variations in Video

2.3.1 Motion and Colour Magnification

Rubinstein, M. (2014) has introduced the technique to analyse subtle motions and colour changes in a video and then to magnify them and put it back into the video with the magnified properties for the ability to visualise such small motion and colour changes. It is termed as the “motion microscope” due to its said abilities.

The human eyes have limited spatial-temporal sensitivity which explains why there are subtle changes that could not be seen. Some information below this limit may provide interesting and meaningful information. A video camera is able to detect these small changes that the human eyes are not sensitive enough.

Some of the useful information that can be obtained is the change in skin colour due to blood circulation. The human heart rate could be measured by finding the peak time of these colour changes and also to magnify them for visualisation. Pulse rate can also be measured without any physical attachment to the patient. Analysis of motion on building and bridges structures can be done in a new approach

with an added on visualisation of these motions. One more application that would be highlighted in this research is the ability to use this information to recover sound that would be discussed in further sections.

2.3.2 Eulerian vs. Lagrangian Perspectives

The approach that Rubinstein, M. (2014) utilised is mainly inspired by the Eulerian perspective. There have been other previous approaches prior to this to analyse and amplify subtle motions which follow the Lagrangian perspective. A brief description of the Lagrangian perspective is that it mainly refers to properties of fluid dynamics where trajectory of particles is recorded over time. Due to accurate motion approximation, complicated motions, additional techniques to ensure high quality and artifact-free, the algorithm may be very complex and much more computationally expensive.

On the other hand, the Eulerian perspective is based on how velocity and pressure evolve over time and these properties of fluid are analysed. Optical flow algorithms are the basis of the approximated motions. It can be said that the Eulerian approach exaggerate motions with the magnification of temporal colour changes rather than explicitly estimating motions. Table 2.1 will summarise the main difference between these two approaches.

Table 2.1: Difference between Eulerian vs. Lagrangian processing

<i>Eulerian</i>	<i>Lagrangian</i>
<i>Exaggeration motions</i>	Explicitly estimates motions
<i>More suitable for smaller amplifications and smoother structure</i>	Supports bigger amplification factor and suitable for fine point features
<i>Has temporal characteristic of noise. Preferable for higher noise level.</i>	Has temporal and higher spatial characteristics of noise. Less preferable for higher noise level.
<i>Simple and faster computation</i>	Higher and more complex computation

It should be noted that the comparison done is with the linear approximation method of the Eulerian approach. An improved method of the Eulerian approach will be introduced in the next sub-section.

2.3.3 Eulerian Video Magnification

Eulerian Video Magnification is the method used by Rubinstein, M. (2014) in revealing subtle motions and colour changes which or are almost impossible to see with the human naked eye and then putting back into the video output for visualisation. The summarised steps for the method is to receive a video sequence as an input, spatial decomposition is applied on the input, and then temporal filtering is carried out on the decomposed frames. The output of the temporal filtering is then amplified to visualise the subtle motions.

Linear Approach. The approach focuses on the colour changes in any local fixed region. It can be deduced that changes in colour of these regions may either be due to the literal change in colour of the static object or the object may have moved. Amplification of these colour changes is the main concept of the method to amplify the motion or colour changes for visualisation. Figure 2.7 illustrates the flow of the method.

Phase-Based Approach. This approach is an improvement over the linear approach. The technique behind this approach is that the video input is spatially decomposed based on the complex-valued steerable pyramids whereas the linear approach is decomposed based on the Gaussian or Laplacian pyramid. Local motions can be retrieved from the phase variations of the complex steerable pyramids which these motions would undergo temporal processing and then amplification to before writing to the video output. Figure 2.8 illustrates the flow of the method.

The background of the phase-based approach is mainly constituted by the phase-based optical flow and the complex steerable pyramids. Fleet and Jepson (cited in Rubinstein, M., 2014) found that a good estimation of the motion field can be achieved by tracking the constant phase contours and then calculating the phase gradient of the spatial-temporal video. A similar technique was also achieved by Gautama, T. & Van Hulle, M. (2002). For the complex steerable pyramids, the fundamental functions may resemble Gabor wavelets in which a Gaussian envelope is used to window the sinusoids. The resemblance is also steerable but not exploited in their work. For more insight regarding Gabor wavelets, the works of Bařina, D. (2011) can be referred to as he shows the use of two-dimensional Gabor wavelets in image processing.

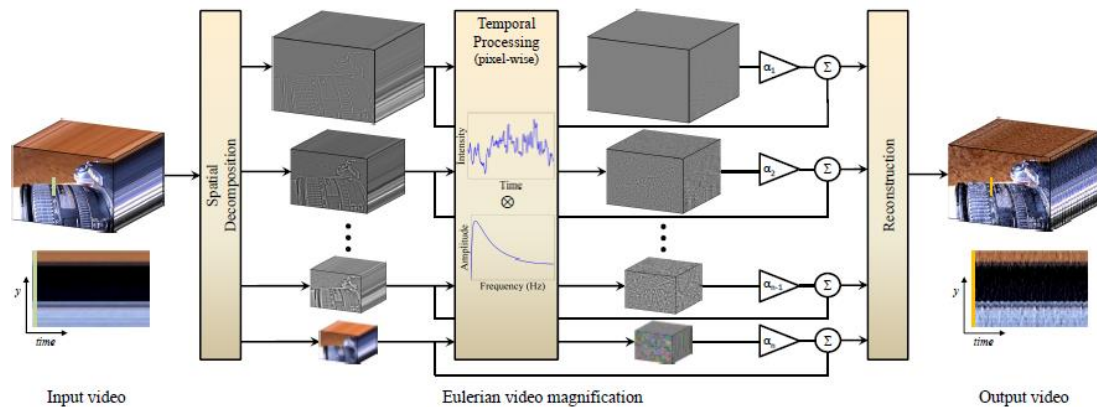


Figure 2.7: Eulerian Linear Approach Framework
(Rubinstein, M., 2014, p. 56)

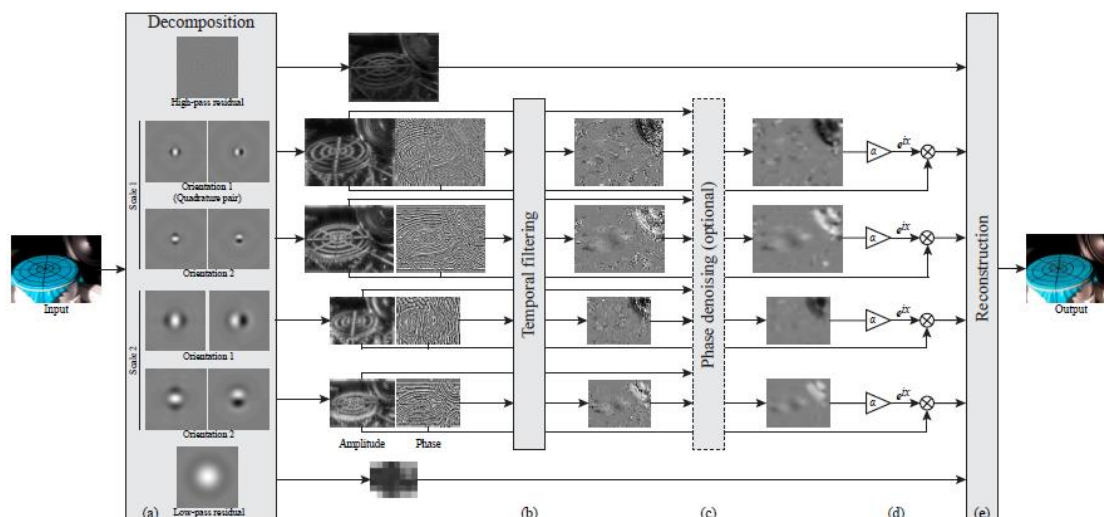


Figure 2.8: Eulerian Phase-Based Approach Framework
(Rubinstein, M., 2014, p. 76.)

Two main advantages of the Phase-Based Approach over the Linear approach is that it supports larger magnification and improved noise performance because amplification of motion does not amplify noise but only translated. A short summary of the different properties of the two approaches has been established in Rubinstein, M.'s thesis. Table 2.2 will display this comparison.

Table 2.2: Difference between Linear and Phase-Based Approach

	Linear	Phase-based
Decomposition	Laplacian pyramid	Complex steerable pyramid
Over-complete	$4/3$	$2k/(1 - 2^{-2/n})$
Exact for	Linear ramps	Sinusoids
Bounds	$(1 + \alpha)\delta(t) < \lambda/8$	$\alpha\delta(t) < \lambda n/4$
Noise	Magnified	Translated

(Rubinstein, M., 2014, p. 88.)

2.4 Space-Time Video Processing

2.4.1 Spatial Processing

This section serves to elaborate a little further on the functions used for processing. As mentioned in the previous section, a video input mainly undergoes spatial and temporal processing before the motion signal is amplified. For the linear approach, the work of the author stated two pyramids which have been applied for spatial decomposition of the video image. They are mainly the Gaussian and Laplacian pyramid.

The theory behind this spatial processing or filtering is that a video image is to undergo convolution by using a Gaussian kernel. The corresponding output of the convolution is a low-pass filtered image from the original image. Continuation process of this convolution will determine the number of levels in the Gaussian pyramid. The parameter that determines the cut-off frequency is commonly used as the symbol σ .

The Laplacian comes into role as it computes the difference between the original and low-pass filtered video image. The continuation process also forms the Laplacian pyramid where each level represents the difference between two levels of the Gaussian pyramid. It can be deduced that the Laplacian Pyramid is a set of band-pass filters. Figure 2.9 illustrates the levels of the pyramids.

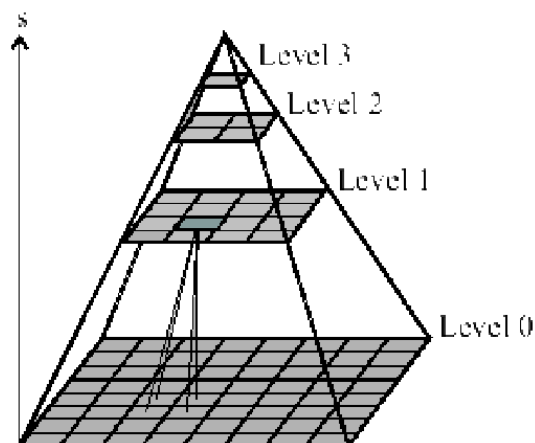


Figure 2.9: Levels in the Pyramid

(Source : <http://www.cs.utah.edu>)

For the Phase-Based approach, the pyramid used is the complex-valued steerable pyramid. Steerable pyramids can be viewed as the extension of Laplacian pyramids and are overcomplete transforms that may represent the image in terms of scale and orientation. As mentioned, the resemblance of the function is the Gabor wavelets. These transfer functions are applied with the Discrete Fourier Transform of the video image to build the steerable pyramid which will decompose the video image into different spatial frequency bands.

Consequently, the phase of each of the subband can be obtained from the complex-valued coefficients in the steerable pyramid. The real part of the coefficient indicates the even symmetric filter while the imaginary part represents the odd-symmetric filter.

2.4.2 Temporal Processing

After the video image undergoes the spatial processing, it is sent for temporal processing or filter where only desired frequency bandwidths are obtained at the output. The difference for the temporal filtering is that it filters frequencies with respect to time domain instead of space domain. These frequencies are essential for the recovery process of sound.

One of the simplest temporal filters that can be used is the ideal band-pass filter. The definition of an ideal filter is that it allows the entire signal power in the pass-band and completely blocks all signal power outside the pass-band. For a better explanation, amplitude spectrum for the ideal band-pass filter will be displayed in Figure 2.10.

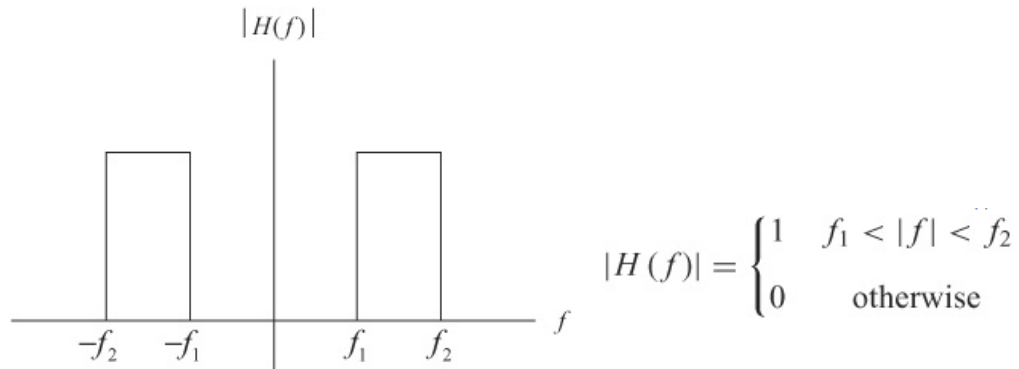


Figure 2.10: Amplitude Spectrum of Ideal Band-Pass Filter

(Source : <http://www.chegg.com>)

Another temporal filter which can be used is the IIR filter. IIR is an abbreviation for Infinite Impulse Response. A good example of IIR filters is the Butterworth filter or Chebyshev filter. These filters have a smoother attenuation at the cut-off frequencies as that they do not completely block all the signal power after cut-off frequency. It provides a smoother transition which may be useful in different applications and replicates a closer behaviour in real circuits where they cannot be infinitely attenuated at a particular cut-off frequency.

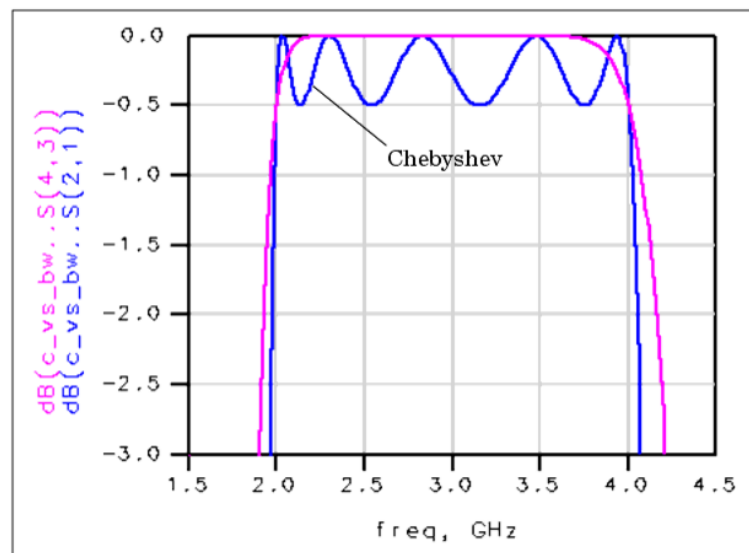


Figure 2.11: Butterworth and Chebyshev Band-Pass Filter

(Source : <http://cp.literature.agilent.com>)

From Figure 2.11, it can be seen the waveform of both the filters where the blue waveform represents the Chebyshev filter and the pink represents the

Butterworth filter. It can be observed that the Chebyshev has a better rate of attenuation than the Butterworth filter while the Butterworth filter has a good maximally flat magnitude response within its pass-band.

2.4.3 Digital Image Processing

Digital image processing (DIP) is similar to the previous sub-section of spatial processing as both are involved in the spatial domain. The methods mentioned earlier can be used in the main process while the methods introduced in this subsection may be used as pre-processing techniques before running the input videos into the main process. DIPs are useful for purposes such as smoothing, sharpening and edge detection. The two DIPs that will be introduced are the mean filter and median filter.

One of the simplest filters is the mean filter whereby it substitutes each pixel value with the average value of its neighbouring pixels. It is a method for smoothing images and is also known as a convolution filter which a kernel is utilised as its base. An example of a 3x3 kernel for mean filter is

$$\frac{1}{9} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

Figure 2.12: 3x3 square kernel for mean filter

(Source : www.bogotobogo.com)

Another common filter is the median filter. The difference is that the median filter is a non-linear operation. It uses an existing pixel value which is the middle pixel value among its neighbouring pixels rather than using the average value. The neighbouring pixels are first sorted out in numerical order before identifying the middle pixel value. Figure 2.13 illustrates in determining the median value as the 3x3 kernel is applied on the pixels.

123	125	126	130	140	Neighbourhood values: 115, 119, 120, 123, 124, 125, 126, 127, 150 Median value: 124
122	124	126	127	135	
118	120	150	125	134	
119	115	119	123	133	
111	116	110	120	130	

Figure 2.13: 3x3 square kernel for median filter

(Source : <http://homepages.inf.ed.ac.uk/rbf/HIPR2/median.htm>)

The advantages of the median filter over the mean filter is that it is more robust against extreme values such that if there is an extreme unrepresentative pixel value in the neighbourhood, the median value is less likely to be affected. Due to this, it is said to be more effective in reducing ‘salt and pepper’ noise. Also, as the median filter always uses one the pixel value in its neighbourhood, it is more likely to preserve sharp edges as compared to mean filter.

2.5 Computation of Motion Signal into Audio Signal

2.5.1 Local Motion Signals

This section will describe in more details on the algorithms and methods used in the previous sections to be applied on recovering sound from the motion information that is to be obtained. The explanations in this section are mainly the works of Davis, A., Rubinstein, M. et al. (2014).

Firstly, to obtain local motion signals, the video image is decomposed using the complex steerable pyramid representation. Each frame of the video is basically broken down into different sub-bands that have different scales and orientation. This complex image can be represented in a different form which consist of amplitude A and phase ϕ .

$$A(r, \theta, x, y, t)e^{i\varphi(r, \theta, x, y, t)} \quad (2.1)$$

The local phases, ϕ component from equation 2.1 is utilised and subtracted from a reference frame's local phase. The first frame is normally taken as the reference frame. The local phases consist of all points in a frame and also all frames in the video. Finally, the results would be the computation of $\phi_v(r, \theta, x, y, t)$ which are the phase variations, where x and y represents the row and column, r and θ represents the scale and orientation and t represents the frame number.

$$\varphi_v(r, \theta, x, y, t) = \varphi(r, \theta, x, y, t) - \varphi(r, \theta, x, y, t_0) \quad (2.2)$$

From the phase variations, the local motion signals can be approximated as the phase variations are approximately proportional to their displacements for small motions. (Gautama, T. & Van Hulle, M., 2002 cited in Davis, A. et al., 2014)

2.5.2 Global Motion Signal

Further processes on the local motion signals are carried out to obtain the global motion signal. A spatially weighted average of the local motion signals is computed to produce a single motion signal for each sub-band of the pyramid. A weighted average using the amplitude A for each local motion signal is used to measure the texture strength as low texture regions contribute to more noise. Equation 2.3 represents the single motion signals.

$$\Phi_i(r, \theta, t) = \sum_{x, y} A(r, \theta, x, y)^2 \varphi_v(r, \theta, x, y, t) \quad (2.3)$$

From the single motion signals, the global motion signal can be computed by adding each of the single motion signals from each sub-band of the pyramid. This

produces the global motion signal with only the parameter t , the number of frames. The global motion signal is then scaled and centered to the range of $[1,-1]$ to be converted into audio data.

$$\hat{s}(t) = \sum_i \Phi_i(r_i, \theta_i, t - t_i) \quad (2.4)$$

2.5.3 Sound Recovery from Various Objects

Davis, A., Rubinstein, M. et al. (2014) carried out an experiment that includes multiple different objects and materials to be tested. The equipment is consisting of a V10 high speed video camera, loudspeaker, and the object which is placed with a separate stand from the loudspeaker to avoid any external contact vibrations. The set up can be seen as in Figure 2.14.

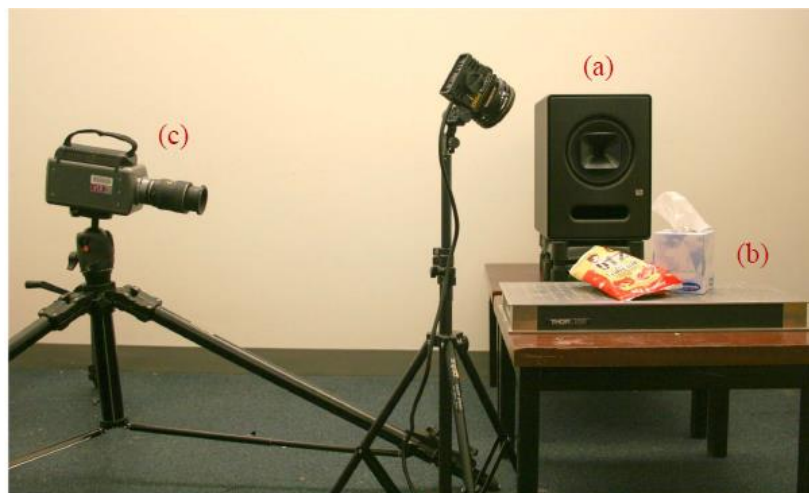


Figure 2.14: Sound Recovery Experiment Setup

(Davis, A., 2014, p. 4)

It is recorded that the typical processing time was about 2 to 3 hours with the processing specifications of 2 3.46GHz processors and 32GB of RAM. A ramp signal of pure tone was played from 100Hz to 1kHz in five seconds. Table 2.3

summarises the performance of the different objects on frequency response ranging from weakest to strongest.

Table 2.3: Performance of Frequency Response of Various Objects

Object	Frequency Response
Brick	Weakest
Water	'
Cardboard	'
Kitkat bag	'
Foil container	Strongest

Besides determining the performance of various objects, one observation that is noticeable is that the recovered higher frequencies are generally weaker in all the objects. This can be expected because higher frequencies would cause smaller displacements and higher attenuation by the object. The highest performance usually contributes to objects which are lighter and easier to move for sound recovery which in this case is the foil container.

Another controlled experiment test was the object response with the change in volume. A pure tone of 300Hz was played with an increasing volume from 57dB to 95dB. The results show that the motion is approximately linear with volume or also known as the sound pressure. This information is useful as the properties of the object can be predicted while carrying out the experiments.

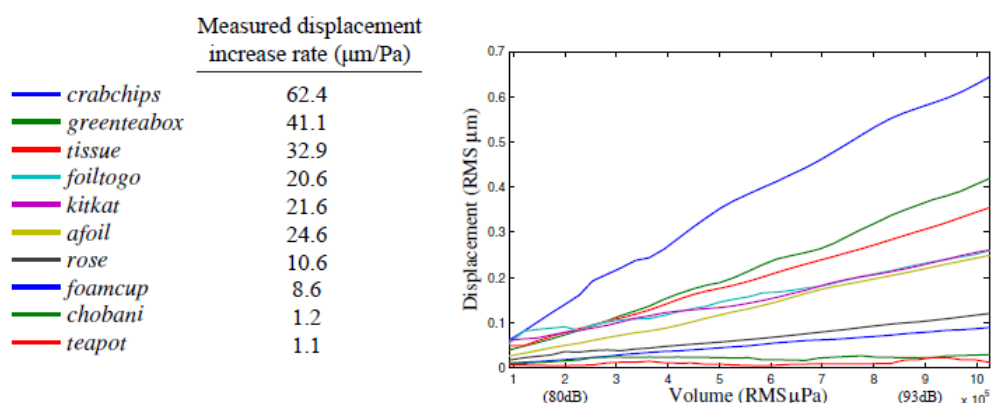


Figure 2.15: Motion vs. Sound Volume

(Davis, A., 2014, p. 7)

2.6 Audio Denoising

There have been a few solutions applied by Davis, A., Rubinstein, M. et al. (2014) to reduce the noise contained in the audio data. These solutions can be performed in the post-processing of the algorithm on the recovered audio. One solution was to apply a Butterworth filter with a cut-off of 20Hz to 100Hz as higher noise energy is found in lower frequencies. They also implemented other audio denoising algorithms such as in professional audio processing software which further improved their results.

Spectral subtraction technique (Boll, 1979 cited in Davis, A., 2014) was implemented for applications focusing on accuracy. For applications focusing on intelligibility, a perceptually motivated speech enhancement algorithm was utilised. (Loizou, 2005 cited in Davis, A., 2014)

During the main process itself, an additional step was also taken before converting the single motion signals into the global motion signal. All the scales and orientation pairs were temporally aligned to prevent destructive interference. The idea is that the phases are to be added constructively by shifting one of them in time.

Further research is done on one of the audio software, FL Studio12 which has the feature of a noise removal tool. One of the features is the 'Equalizer Envelope' which provides an interface to select regions of frequency in the frequency spectrum that are likely to be contributing to noise. Another feature which is the 'Denoiser' aims to reduce the constant background noise in recordings. A sample profile of the background noise is usually required in this feature.

One more feature available which may be applicable for this research is the 'Declipper'. The software detects clipped peaks and re-creates peaks which have exceeded the defined clipped level. All these various audio denoising techniques can be taken into high consideration in improving the quality of the sound recovered through the main algorithm.

2.7 Rolling Shutter Technique

The experiment conducted and the methods explained previously require the need of a high speed video camera. This is because the sampling rate of the video camera becomes the limitation of twice the maximum frequency of the sound it can recover according to Nyquist theorem. Once again, Davis, A., Rubinstein, M. et al. (2014) proposed a technique to recover sound using a normal video camera. Such discovery has allowed the possibilities of implementing the algorithm on normal cameras and CCTV cameras for surveillance. They took the opportunity of the rolling shutter properties in the CMOS sensors which can mostly be found in many mobile phones and DSLR cameras. It is actually a cheaper to implement design with lower power consumption as compared to uniform global shutter but produces skewing artifacts in recording images of moving objects.

The properties that became an interest is that the sensor pixels of the rolling shutter are exposed row by row from the top row to the bottom at a certain finite time. This allows each row to represent an audio signal instead of per frame. The researchers took the advantage to extend the sampling rate of the video camera beyond its frame rate to recover much higher sound frequencies.

The parameters required is the exposure time of the camera where it is the time taken for the shutter to open for each row, the line delay which is the length of time between consecutive rows, period of frame where it is the time taken from the capture of the first frame to the start of the second frame and the frame delay where there is a certain time interval between the end of the final row of the first frame and the initial row of the subsequent frame. Figure 2.16 will provide a much clearer illustration on the parameters.

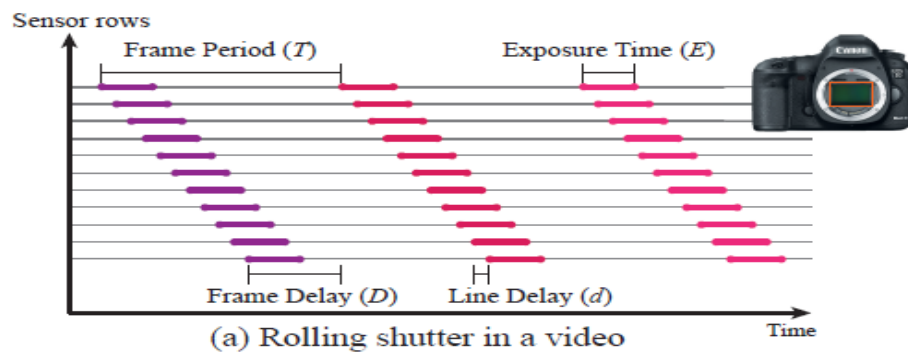


Figure 2.16: Rolling Shutter Properties
(Davis, A., 2014, p. 8)

CHAPTER 3

METHODOLOGY

3.1 Applied Theories

3.1.1 Pre-processing Technique

Based on the various literatures that have been reviewed, some of the theories and methods will be applied in constructing the algorithm. The pre-processing techniques mean they are to be applied before processing the video through the main processing algorithm. The reason for such is to reduce the noise present in the raw input videos. Reducing the noise in the early stage may prevent the noise not just from propagating but from amplifying in the later stages of the process.

Examples of noise present in the video that could be pre-processed are the flickering caused by the fluorescent light about 120Hz, two times the frequency of the AC current. Other noise that may be present is the ‘Gaussian’ or ‘salt and pepper’ noise. The technique that is applied in this research is the median filtering as the advantages over mean filtering have been discussed in Chapter 2 and is also more effective in reducing the ‘salt and pepper’ noise. Comparisons will be done for output results with and without the pre-processing filtering.

3.1.2 Process Algorithm

The Eulerian approach is selected due to a few reasons but mainly because the computational power is much lower than the Lagrangian approach. Also, since there is no necessity of motion magnification in recovering sound, the magnification factor would not be a major criterion and with a more prone to spatial noise, the Eulerian approach would be less sensitive to noise from the input video.

The Eulerian Phase-Based approach is chosen over the Eulerian Linear approach as it is an improved version of the Eulerian approach. On top of that, the phase variations in the phase-based approach contain the essential information that can be used to construct the audio data. This is done through applying the Discrete Fourier Transform of the images (frames of the video) on the complex steerable pyramid transfer functions.

This eventually means that the complex steerable pyramid is the spatial filter, decomposing the spatial information in terms of orientation and scale. At this point of time, no temporal filter was used within the main algorithm, thus the entire range of frequencies from the imagery would be retrieved.

The local phase variations of each pixel in the steerable pyramid are then summed and scaled to compute the global motion signal. Equations 2.1, 2.2, 2.3 and 2.4 are the fundamentals in writing the algorithm. The global motion signal can be used to construct the audio data.

3.1.3 Denoising Methods

Some of the concepts of audio denoising mentioned in Chapter 2.6 ‘Audio Denoising’ are applied in improving the audio data constructed and recovered from the process algorithm. However, they are implemented using MATLAB R2014a software instead of the professional audio software available in the market. Some of

the methods which are found useful are such as normalisation, band passing, amplification and threshold attenuation.

Normalisation is to readjust the values of the audio data in reference to the zero level without losing the frequency information of the sound. This would balance and produce a more steady sound. The raw audio is only scaled from -1 to +1 without reference to the zero level. Band passing filters are crucial in improving the sound quality significantly. Only the range of sound frequencies which are desired are band passed, thus reducing the noise outside this range. A Butterworth filter is used to realise this method. Flickering effect of the video can also be attenuated using similar temporal filters such as stop band filters.

Amplification may be useful if the volume of the sound recovered is too low. It is a good practice to apply the temporal filter first or any other noise reduction method before amplifying the audio data value as the amplification may amplify noise as well. In addition to that, attenuation below a certain threshold reduces noise while being able to amplify the volume of the actual sound. It can be done by only reducing any sound below a certain threshold while maintaining any values above the threshold. This is due to the assumption that majority of the noise are in the lower sound level at the same time realising that this method cannot eliminate noise level which are close to the actual sound level.

The original and denoised audio graphs are often compared closely with while performing these denoising methods to ensure the quality of the sound is improved.

3.1.4 Program Flow

All the algorithm and techniques are coded using the MATLAB R2014a software. Input video formats that have been tested are the .avi, .mp4 and .MOV format. The default audio format reconstructed is .wav and the audio data is of a double type monophonic sound. Other formats have not been tested to compare the difference in

quality as it is not the main objective of the research. The general flow of the latest algorithm and experimental process will be described in Figure 3.1.

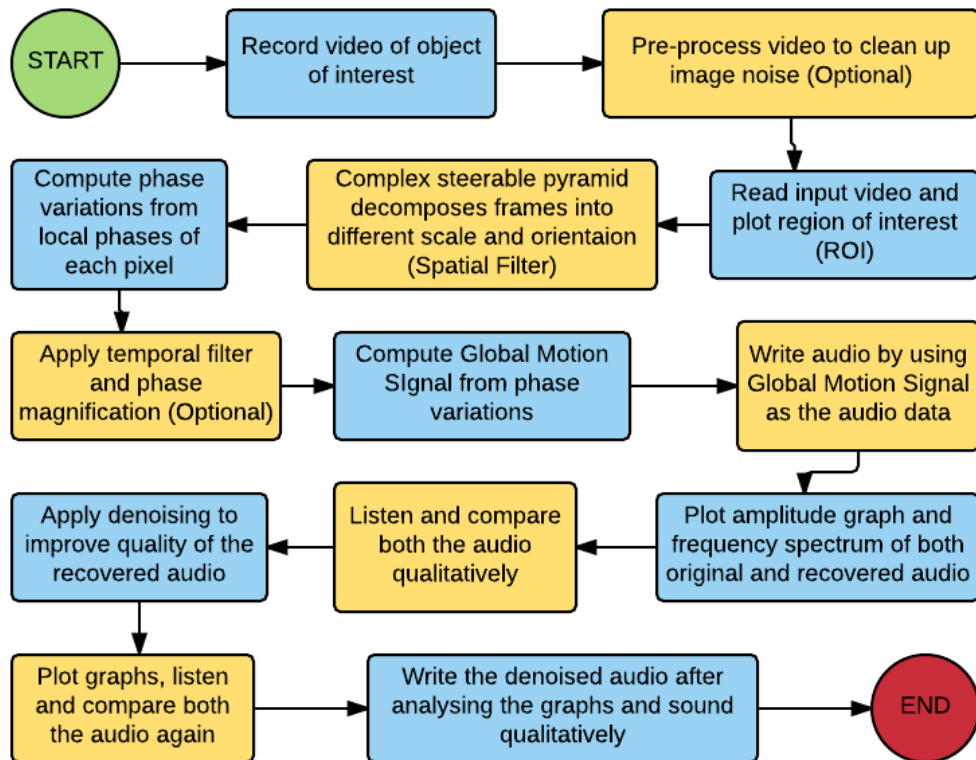


Figure 3.1: Program Flowchart

As a further explanation of the flowchart, prior to reading the input video, there are parameters that need to be defined. These parameters may consist of the sampling rate, magnification factor, cut-off frequencies for temporal filter if applicable, scaling of video and range of frames of interest. Region of interest is implemented so that only regions which give much higher sound information are chosen rather than the entire video which may contribute to more noise.

The temporal filter and phase magnification are optional because they are more useful in the application of visual motion magnification rather than the main objective in this project which is to recover audio. Various graphs are plotted to compare the original and recovered sound for a better analysis because it is difficult to solely determine qualitatively the performance of the recovered sound by hearing. Though, there is no denying that the sound being audible and recognised is important

and in fact more convincing to users rather than just visualising graphs. Hence, both are equally important in accessing the quality of the recovered sound. The procedure in applying audio denoising may be repetitive until the best audio quality is achieved.

3.2 Source of Sound

3.2.1 Acoustic Guitar Strings

To determine which source of sound would be suitable and practical, simple sounds are first carried out. In the preliminary experiment of this research, a single frequency sound have been tested and successfully recovered through a video footage of water ripples. The next idea of an experiment would be to carry out on real musical instruments rather than just from a speaker. With a little research on acoustic guitar, each guitar string key and sound frequency can be identified. A typical acoustic guitar string has 6 strings that are tuned in keys of E4, B3, G3, D3, A2 and E2. The frequencies of the guitar strings are listed in Table 3.1.

Table 3.1: Guitar String Key Frequency

Guitar String Key	Frequency (Hz)
E4	329.63
B3	246.94
G3	196.00
D3	146.83
A2	110
E2	82.41

From the Table 3.1, it is observed that the Key A2 and E2 falls below 120 Hz which theoretically should be able to be sampled by a 240 fps video camera commonly available in high-end smartphones. An observer could try plucking both of these guitar strings and record the subtle vibrations of the guitar strings on video to recover the sound. The sound produced is not too much of a difference as

compared to a single frequency sound if each of the strings are plucked independently, but would rather be interesting and convincing to verify the sound frequency of the strings through visual as well.

3.2.2 Frequency Sweep Sound

Single frequency sound has been proposed in the previous sub-section by independently plucking the guitar strings. The next step is to increase the complexity of the sound. A frequency sweep sound would be the next ideal sound as it covers a frequency range yet still produce a single frequency at a time. With a higher frame rate video camera, a larger frequency range can be recovered and an exponential curve would be observed at the spectrogram. The experiment would prove that the algorithm is able to process a range of frequencies rather than solely single frequencies. If the experiment is successful, the research can proceed with even more complex sounds.

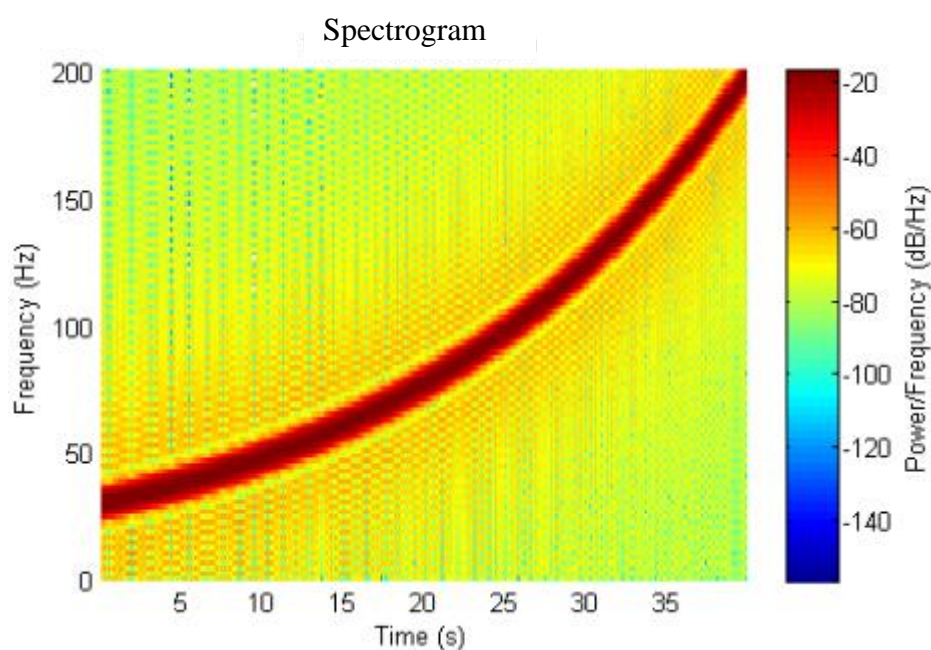


Figure 3.2: Frequency Sweep (30Hz to 200Hz)

3.2.3 Self-Constructed Audio

Other than recovering a certain range of sound frequencies through the frequency sweep, a self-constructed audio may be introduced for creativity and flexibility in manipulating different musical patterns. The main reason of doing so is because we can construct simpler musical patterns as compared to professionally recorded music in the entertainment industry. Recovering self-constructed audio would be the initial step before attempting to recover more sophisticated music.

In this research, 4 sound frequencies of pure tone (sinusoidal waveform) are selected for manipulation and construction of the audio. The 4 frequencies chosen are 70Hz, 80Hz, 90Hz and 110Hz. The construction of the audio is done using MATLAB R2014a as well. Basically, the 4 frequencies mentioned are arranged with various patterns that can be recognised when played. Two examples will be shown; the first version would be a simpler version and the second is a little more sophisticated. With the clear visual of the patterns in the spectrogram, it is hoped that there will be a clearer visual of the recovered sound compared to the more sophisticated music.

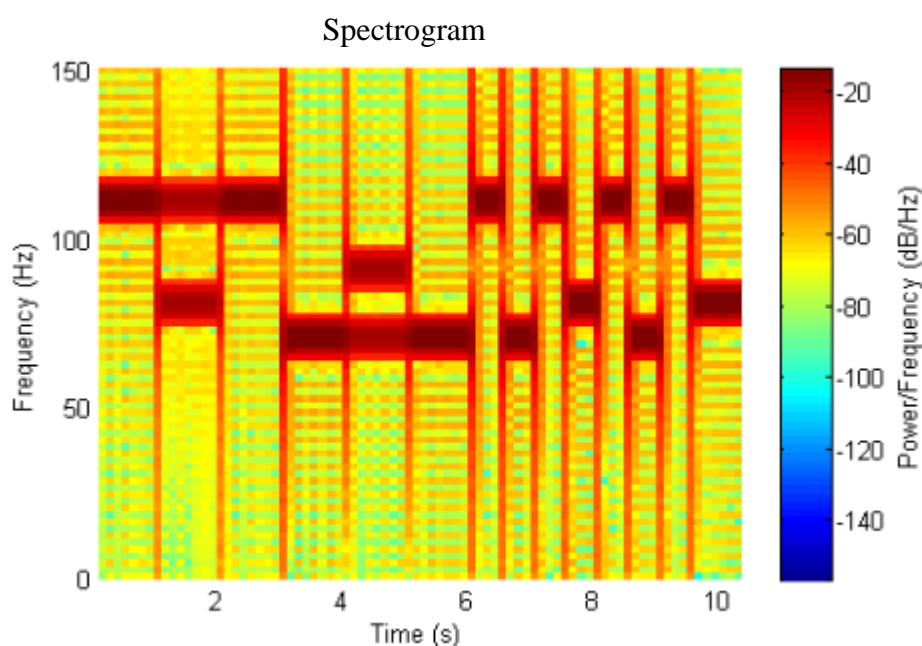


Figure 3.3: Self-Constructed Audio (Simple Version)

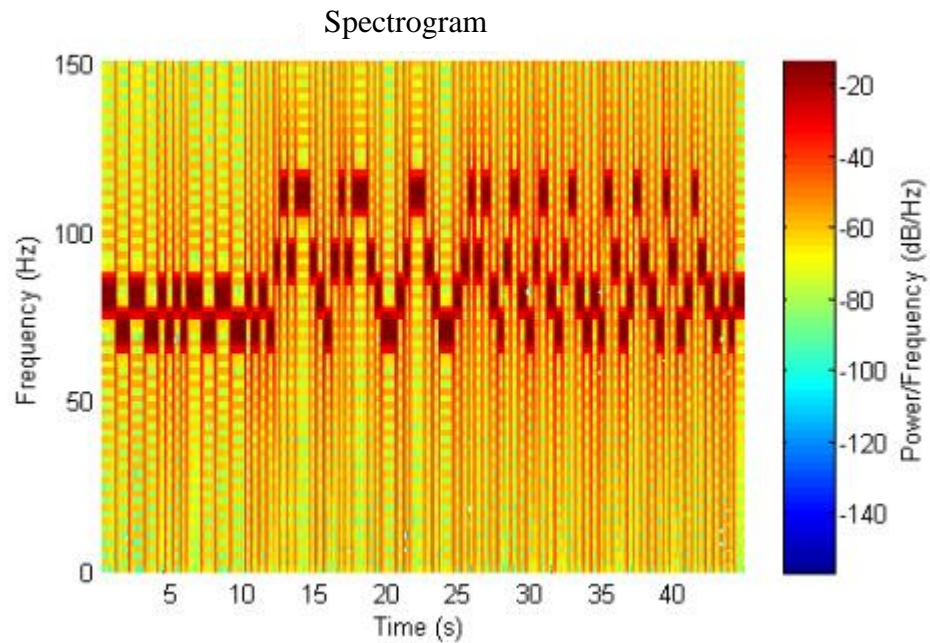


Figure 3.4: Self-Constructed Audio (Complex Version)

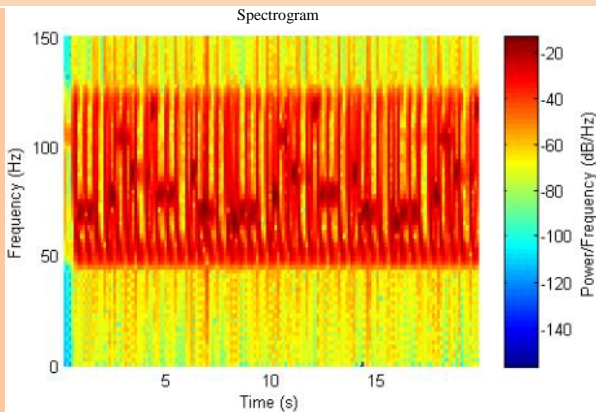
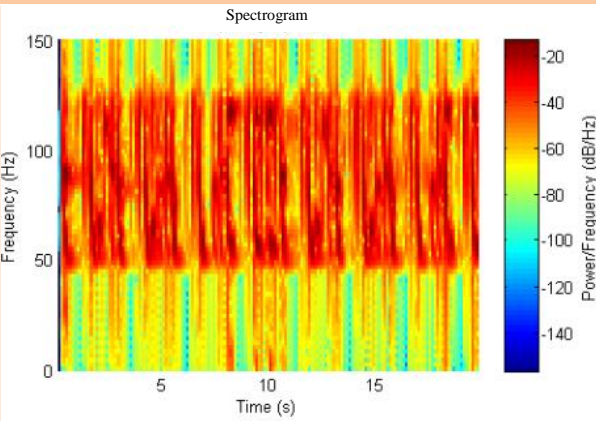
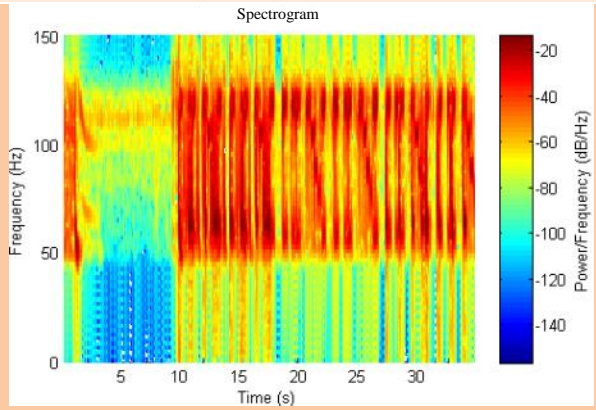
3.2.4 Amplified Bass Components of Songs

To step up the complexity of the audio to be recovered, the next attempt is to recover songs. In this research, the amplified bass components of the songs are used instead of the original song. The reason is due to the limitation of the equipment available which would be addressed in later sections. With the available video camera, it is most likely that the highest sound frequency that can be recovered is 120Hz with a frame rate of 240fps. Therefore, a few songs are selected but up to the limit of 120Hz sound only.

Again, this process is also done in MATLAB R2014a. A band pass filter is used to filter out sound frequencies of the song which are above 120Hz or too low which are redundant. Consecutively, the songs are then amplified so that the sound becomes more audible because the bass components of songs are usually softer than the melody or treble range. These songs eventually become just music with bass only as the range of the frequencies is within the bass range and out of the human voice range.

The sound may not be similar to the original song but importantly, its unique pattern can still be recognised. A good practice of selecting songs is to choose songs which have more sophisticated bass sound rather than very simple and repetitive beats. Table 3.2 will list the songs which are selected and filtered along with their frequency spectrum graph attached. A clearer visual and details of the frequency spectrum will be discussed in Chapter 4.

Table 3.2: Selected and Filtered Songs

Songs	Spectrogram
<p style="text-align: center;">Justin Bieber- What Do You Mean?</p>	
<p style="text-align: center;">Black Eyed Peas Ft Justin Timberlake - Where Is The Love?</p>	
<p style="text-align: center;">Macklemore & Ryan Lewis- Downtown</p>	

3.2.5 Piano Sound

Lastly, a piano sound playing ‘Mary had a little lamb’ is to be recovered from a high speed video camera. The range of the music covers a few hundred hertz up to around 500Hz. The high speed video footage of a bag of chips which the piano music is played towards can be retrieved from online sources uploaded by Davis, A. and his team. The video footage will be processed into the main algorithm prepared in this research. A spectrogram of the original music is plotted as shown in Figure 3.5.

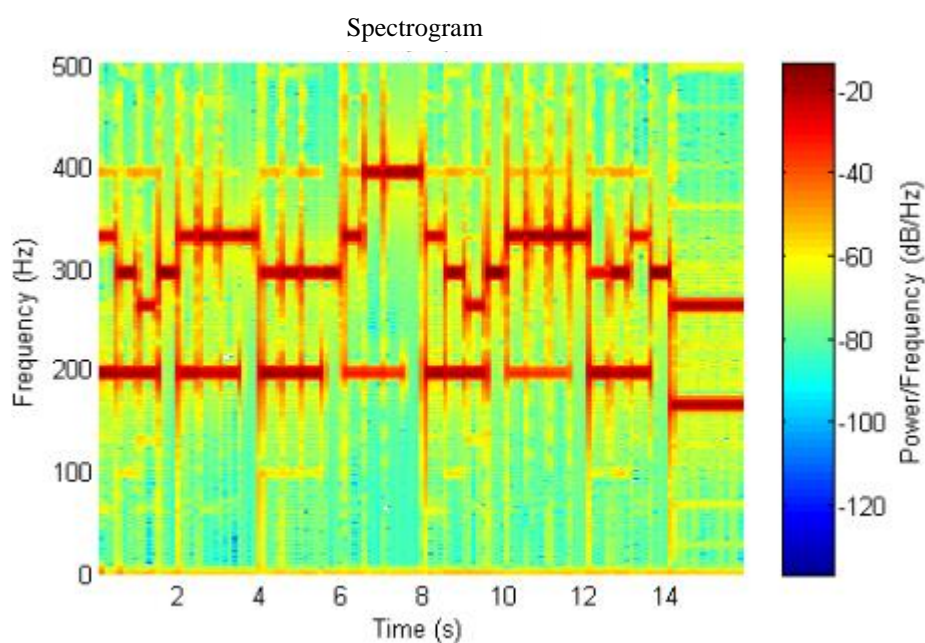


Figure 3.5: ‘Mary had a little lamb’ Piano Music

3.3 Source of Object

The source of object is where the object is excited by the source of sound and produces subtle vibrations which are difficult for the naked eye to observe. That object will then be recorded using a video camera. A few source of object is recorded in this research project which will be listed in Table 3.3

Table 3.3: Source of Object

<i>Source of Object</i>
<i>1. Acoustic Guitar Strings</i>
<i>2. Diaphragm of Speaker</i>
<i>3. Plastic Bag</i>
<i>4. Bag of Chips</i>

The acoustic guitar strings is a little bit more unique than the other source of objects. This is because it is a musical instrument thereby acting as both source of sound and source of object. The sound is produced by the vibration of its strings. Hence, the video footage of the source of the sound itself will be recorded as the source of object. The experimental setup will be explained in the next section.

The diaphragm of the speaker will be used as the primary objective to recover sound. As its name speaks for itself, the diaphragm would vibrate according to the sound played by the speaker. Such functionality allows various sound and magnitude of vibrations to be tested and hence, will be a good tool for proof of concept of this research method.

After successfully recovering sound from the two previous source of object, a plastic bag will be experimented to recover sound with the source of sound further from the source of object. This eventually means that the vibrations are even more subtle and presence of noise would be higher.

The high speed video footage of a bag of chips is actually retrieved from online sources as mentioned earlier whereby the piano music of ‘Mary had a little lamb’ acts as the source of sound. Therefore, it is also included as one of the source of objects as we aim to recover sound from its video footage as well.

3.4 Sound Recovery Experimental Setup

3.4.1 Workstation Specification

Before describing the experiments that were carried out, the workstation that is used for all the processing will be mentioned first as it is common to all the experiments. A high processing power workstation was managed to be attained during the period of this project. With the privilege addition of this equipment, the limitation of the experiments has been extended to higher limits, widening the potential applications that were not thought possible earlier such as recovering higher sound frequency ranges.

The specifications of the workstation are consisting of a dual 6 cores processor; Intel Xeon E5-2630v3 running at 2.6GHz/15MB, 48GB RAM memory, 2TB hard disk and graphics card NVIDIA QUADRO K420. Higher speed video footages are able to be processed with the large amount of memory available and on top of that significantly increasing the efficiency of the process due to shorter processing time because the amount of data to be processed is also very large for this research.

3.4.2 Acoustic Guitar Strings

The first step prior to recording the acoustic guitar strings is to ensure that the guitar strings are properly tuned to the right frequencies by using a guitar tuner. Only the first two strings are played as they are below 120Hz frequency. They are the E2 (82.41Hz) and A2 (110Hz). The two strings are either plucked independently or simultaneously from the other end of the guitar without interfering the region being recorded. Two video cameras with different resolutions are used to record the vibrations. One is an IphoneSE model with resolution of 1280x720 and the other is a digital camera Canon SX280HS model with a resolution of 320x240 with both having a frame rate of 240fps.



Figure 3.5: Acoustic Guitar Strings Experiment Setup

The experiment was carried out in the audio laboratory. From Figure 3.5, it can be noticed that a white paper was slipped under the strings for a better contrast between the strings and the background. As a good sound recorder is not available, the comparison of the recovered sound is done with the recorded video by observing the time when each string vibrates. The frequency can be verified by referring to the E2 and A2 values with the frequency spectrum obtained. Figure 3.6 will display the camera view of the strings and the estimation of the region of interest to be processed.

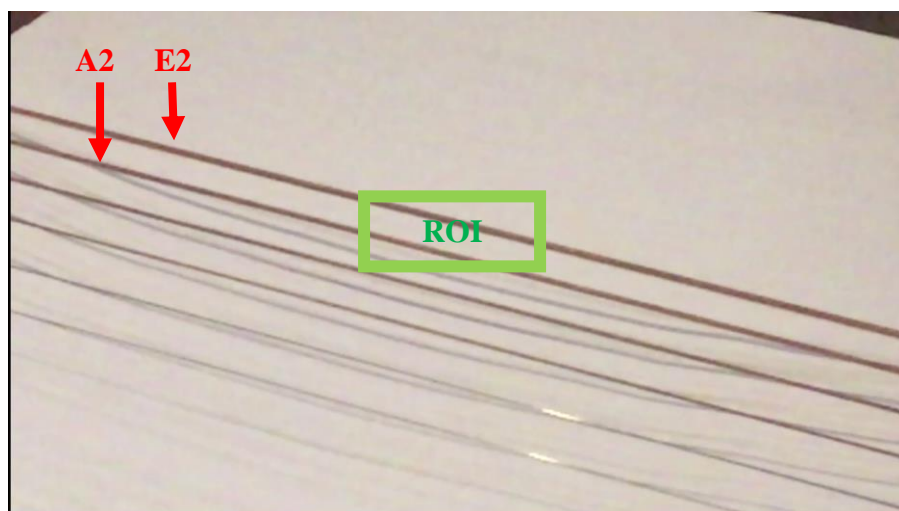


Figure 3.6: Camera View of Guitar Strings

3.4.3 Diaphragm (Resolution)

This section describes the experiment that is carried out by recording the diaphragm of the speaker which is also carried out in the audio laboratory. The control factor in this experiment is mainly the change in resolution for each source of sound recorded. A summary of the conditions of the experiment will be listed in Table 3.4.

Table 3.4: Summary of the Diaphragm(Resolution) Experimental Conditions

Location	Audio Laboratory
Resolution	1280x720 (IphoneSE) 320x240 (Canon SX280HS)
Frame Rate	240fps
Source of Sound	<ul style="list-style-type: none"> • Frequency Sweep • Justin Bieber – What Do You Mean? • Black Eyed Peas Ft Justin Timberlake – Where Is The Love?
Speaker	Audiobox A100-U



Figure 3.7: Diaphragm(Resolution) Experiment Setup

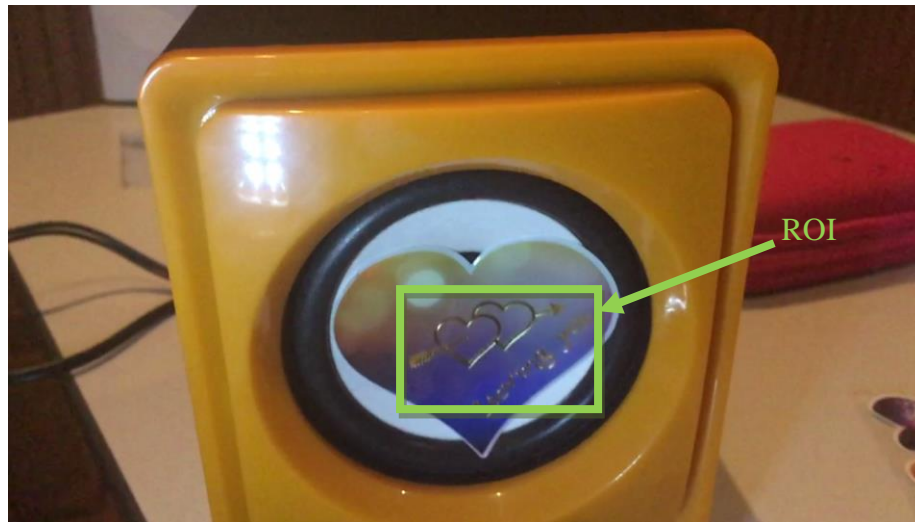


Figure 3.8: Camera View of Diaphragm(Resolution)

From Figure 3.8, it is observed that a sticker is attached on the diaphragm of the speaker. This is to provide some features and details on diaphragm as it may improve the algorithm in recovering the sound. Also in the experiment setup, a LED torchlight is used to provide enough illumination on the diaphragm of the speaker as the ceiling lights provided may not be adequate.

3.4.4 Diaphragm (Volume, Digital Filtering, ROI)

This experiment uses the same source of object but from a different speaker and different experimental conditions. The changes made are to investigate other control factors that may affect the results, to discover improvements for the experimental setup and to explore the capabilities in recovering more types of sound. Table 3.5 will help in observing the changes from the previous experiment.

Table 3.5: Summary of the Diaphragm(V,DF,ROI) Experimental Conditions

Location	Mechatronics Laboratory
Resolution	1280x720 (IphoneSE)
Frame Rate	240fps

Source of Sound	<ul style="list-style-type: none"> • Frequency Sweep • Self-Constructed Audio (Simple Version) • Self-Constructed Audio (Complex Version) • Justin Bieber – What Do You Mean? • Black Eyed Peas Ft Justin Timberlake – Where Is The Love? • Macklemore & Ryan Lewis- Downtown
Speaker	Salpido Tron 101
Speaker Volume	50%, 75%, 100%
Digital Image	With Median Filtering
Filtering	Without Median Filtering
Region of Interest	2 Regions of Interest (Refer to Figure 3.10)

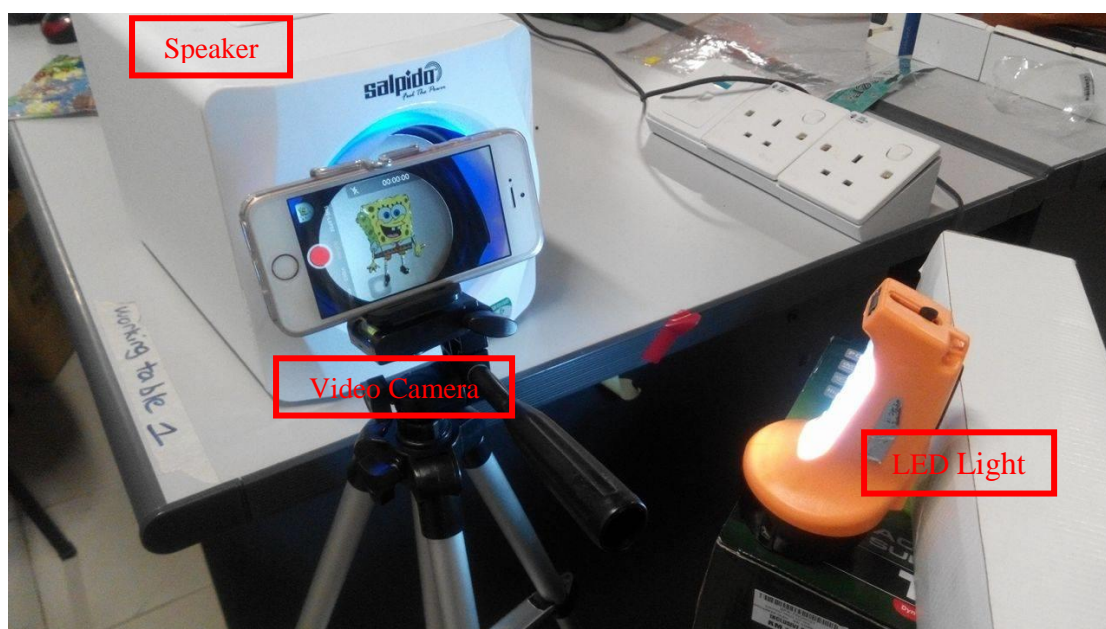


Figure 3.9: Diaphragm(V,DF,ROI) Experiment Setup

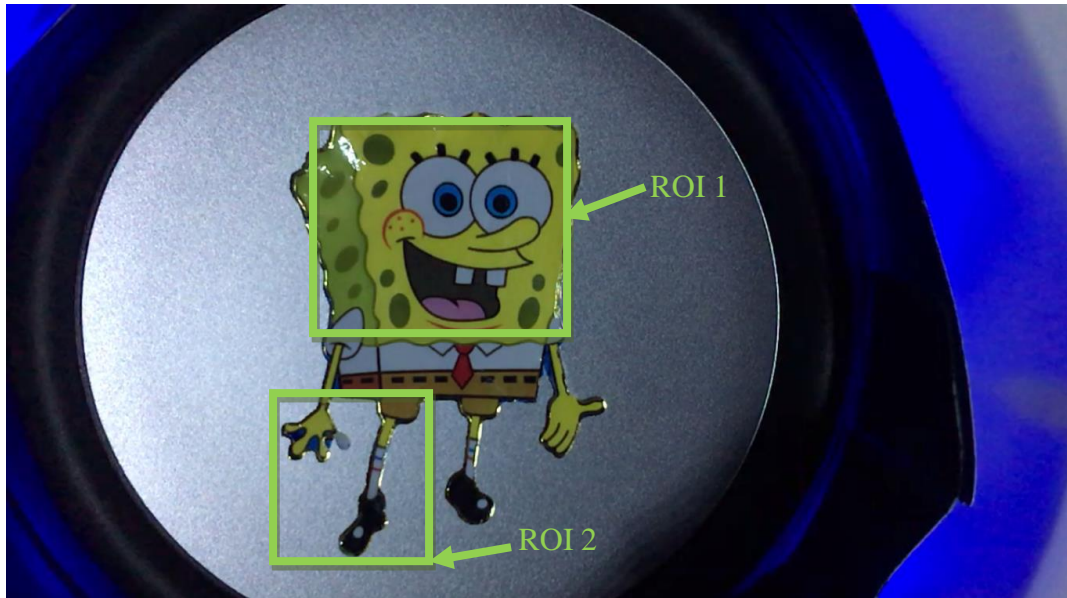


Figure 3.10: Camera View of Diaphragm(V,DF,ROI)

The location being changed to mechatronics laboratory is due to the unavailable space at audio laboratory rather than a better environmental conditioning reason. The source of sound is increased from 3 to 6 types of sound. The model of the speaker is also changed into a much more powerful bass speaker in hopes of improving the accuracy of the diaphragm's vibrations.

There are three control factors which are to be tested in the experiment. The volume of the speaker may be classified as the hardware changes as it needs to be manipulated during the experiment itself. The volume is controlled by fixing the mechanical potentiometer of the speaker at fixed position while adjusting the volume from the laptop with percentage levels of 50%, 75% and 100%. The software changes include the implementation with or without median filtering and the selection of ROI of the video recorded. Those changes may be manipulated after the recording of the videos. Only a single resolution is used throughout this experiment.

3.4.5 Plastic Bag (Volume, Motion Magnification)

Once the previous experiments are deemed successful, a different source of object is used to recover the same source of sound. The suggested object is a plastic bag as it

is further away from the source of sound and produces even more subtle vibrations. The limitation and capabilities of the process algorithm can be further discovered through this experiment. The volume of speaker is once again manipulated in the experiment. The motion magnification factor is a property which was discussed in Chapter 2 in magnifying subtle motions in video for visualisation. Using the same concept while recovering sound, the motion magnification factor can be embedded in the main algorithm to magnify the subtle vibrations before converting them into audio signal.

Table 3.6: Summary of the Plastic Bag(V,MM) Experimental Conditions

Location	Enclosed Room
Resolution	1280x720 (Iphone6s)
Frame Rate	240fps
Source of Sound	<ul style="list-style-type: none"> • Frequency Sweep • Self-Constructed Audio (Simple Version) • Self-Constructed Audio (Complex Version) • Justin Bieber – What Do You Mean? • Black Eyed Peas Ft Justin Timberlake – Where Is The Love? • Macklemore & Ryan Lewis- Downtown
Speaker	Salpido Tron 101
Speaker Volume	50%, 100%
Motion	Magnification Factor = 1
Magnification	Magnification Factor = 10

The equipment and location used are based on the availability during the period of time. The plastic bag is pasted on the wall while the speaker is played facing the plastic bag. The video camera records the plastic bag from another side which is away from the direction of where the speaker is facing. The setup can be seen in Figure 3.11.



Figure 3.11: Plastic Bag(V,MM) Experiment Setup



Figure 3.12: Camera View of Plastic Bag(V, MM)

3.4.6 High Speed Video of Bag of Chips

All the previous experiments that were dealt with are only recovering sound up to 120Hz. With a high speed video footage of a bag of chips which a piano music ‘Mary had a little lamb’ is played towards it, the main algorithm can process the

video to recover the piano music from the bag of chips. A region of interest of the bag of chips is selected as shown in Figure 3.13.

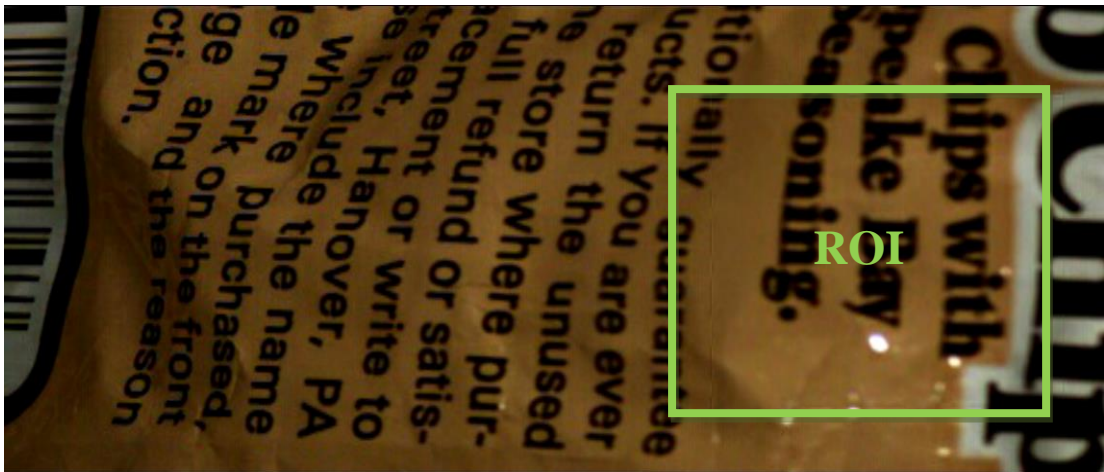


Figure 3.13: Camera View of Bag of Chips

3.5 Experimental Limitations

3.5.1 Frame Rate of Video Camera

Throughout the experiments, the choice and selection of the source of sound has always been limited by the frame rate of the video camera. The best available frame rate is a 240fps which allows sound frequency to be recovered up to 120Hz which is half of the sampling rate (fps) according to Nyquist Theorem. However, it is only the minimum requirement and much higher sampling rates are preferable in accurately recovering the desired frequencies. Acquiring a high speed camera may be too costly and over the budget given for this project. Another solution to extend this limitation is to implement the rolling shutter technique which has been reviewed in Chapter 2.7.

3.5.2 Pixel Resolution of Video Camera

The pixel resolution is another aspect of the video camera which is essential in this method. It is to be highlighted as it is usually an inverse relationship between the frame rate and resolution due to the hardware limitations of the current technology. In other words, the higher the frame rate of the video, the higher the tendency of the resolution to decrease. The resolution corresponds to the number of pixels, thus the number of samples available for each frame to determine the subtle vibrations. If the resolution is too low, the signal received would not be sufficient despite of having the required frame rate. A balance between these two aspects is required to find the optimum values.

3.5.3 Video Camera Zoom Factor

Besides the frame rate and the resolution of the camera, the zoom capability is also very useful in determining the maximum possible distance of the camera from the source of object. The useful zooming method would be to utilise an optical zoom lens rather than the digital zoom found in most digital cameras. Digital zooming merely crops the image and then readjusts it to the ratio and dimensions of the original image. This would reduce the actual resolution, sharpness and quality of the image which reduces the quality of the sound recovered. Optical zooming maintains the resolution and sharpness of the images or video while zooming into the region of interest. With the availability of optical zoom lenses, the distance between the camera and the object can be varied without deteriorating the sound quality.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Preliminary Results – 40Hz Bass Sound (Water Ripple)

This sub-section will provide a little recap on the experiment carried out during the first part of this project and also some of the improvements made. The early experiment instilled critical thinking and thoughts in improving the process and algorithm that finally led into the main experiments being carried out as mentioned in the previous chapter of this report. The following content will summarise the details of the experiment, results and the improvements made to obtain a better recovered sound quality.

Due to not having a higher speed video camera during that period of time, the slow-mo video recording of the HTC One M7 mobile phone was utilised, which after converting back to normal speed may reach up to 92 fps. This allows the possibility to sample bass sound frequencies up to 46 Hz which is still audible to the human ear. Hence, a 40Hz bass sound was chosen and played with certain pause intervals. Orange juice on the speaker diaphragm is used as the source of object which will create water ripples as sound is played and will be recorded by video.



Figure 4.1: Water Ripple Experiment Setup

In the current part of the project, a band pass filter was applied to clean up the sound recovered and to eliminate sound frequencies which are not around 40Hz. The quality of the sound increased tremendously. The graphs of the original, recovered and filtered version will be displayed subsequently in the figures below.

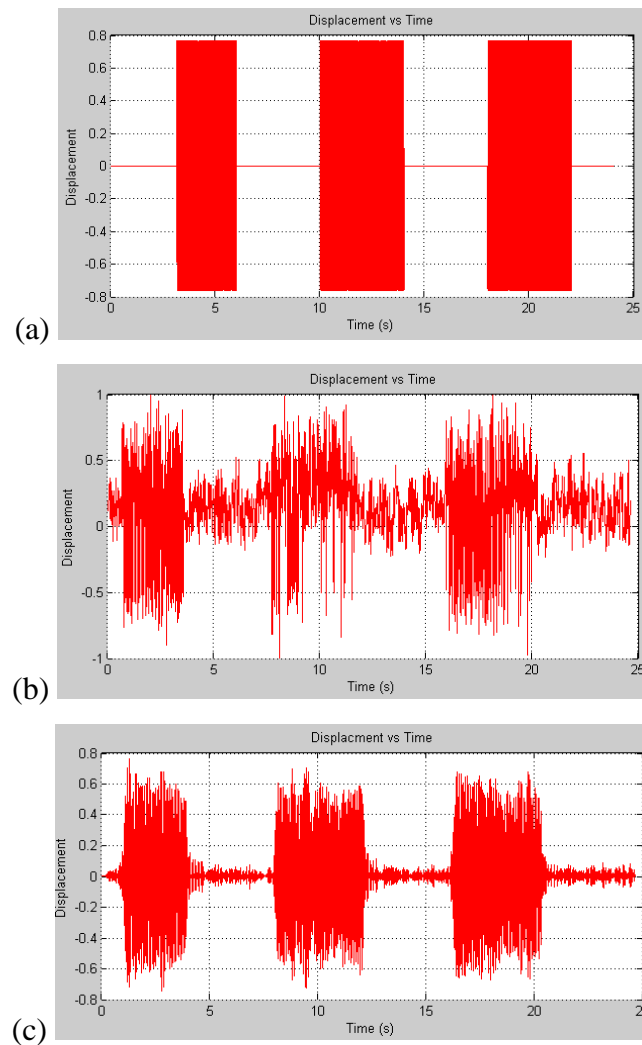


Figure 4.2: Amplitude Graph 40 Hz (a)Original (b)Recovered (c)Filtered

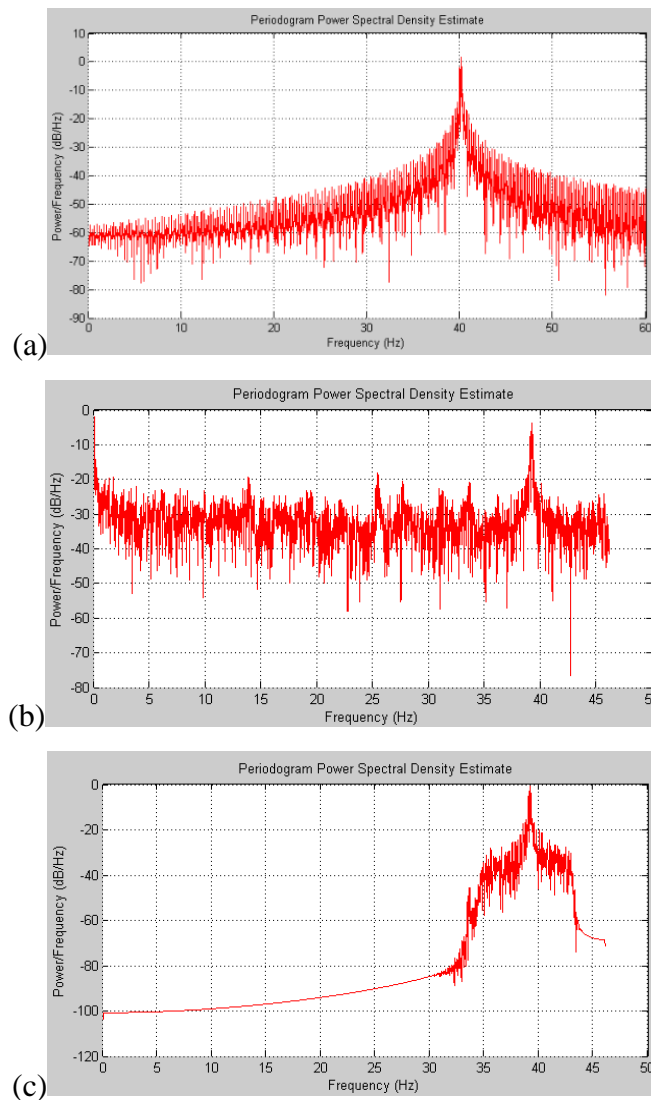


Figure 4.3: Power Spectral Density Graph 40 Hz (a)Original (b)Recovered (c)Filtered

From the graphs, it can be observed that the filtering applied is very effective for single frequency signals but the case might not be the same for more complex sounds which will be evaluated later. There are several reasons that further experiments were not conducted using this setup due to its unsuitability. Firstly, there is not much texture variation on the surface of the liquid. The reflection and opaqueness of the liquid are difficult to control during the experiment and becomes an obstacle to determine the best experimental condition.

The most crucial element is that as the water ripples reach the edges of the diaphragm, there is a tendency that there will be destructive and additional ripples which are not directly produced from the sound. Also, after observing the video recorded in slow motion, it is noticed that there is a delay in the forming of the water ripples as the plastic region vibrates earlier than the ripples. It is deduced that the time delay is due to the time required for enough energy to excite the liquid to start producing the ripples. Since there are many hindrances in this experiment, the main experiments are carried out with the replacement of other sources of objects as mentioned in Chapter 3.

4.2 Experimental Results

4.2.1 Acoustic Guitar Strings

The strings which are plucked in the experiment are the E2 and A2 strings which respectively would vibrate in 82.41Hz and 110Hz. There is no original sound to be compared since the sound is directly played from the guitar and a good sound recorder is not available. Hence, the verification of the results will be accomplished by observing the vibrations of the strings and sound from the video and compared with the frequency spectrum of the sound recovered. The sound of the string can be verified by analysing the frequency of the sound recovered. There will be two major frequencies in the frequency spectrum and periodogram.

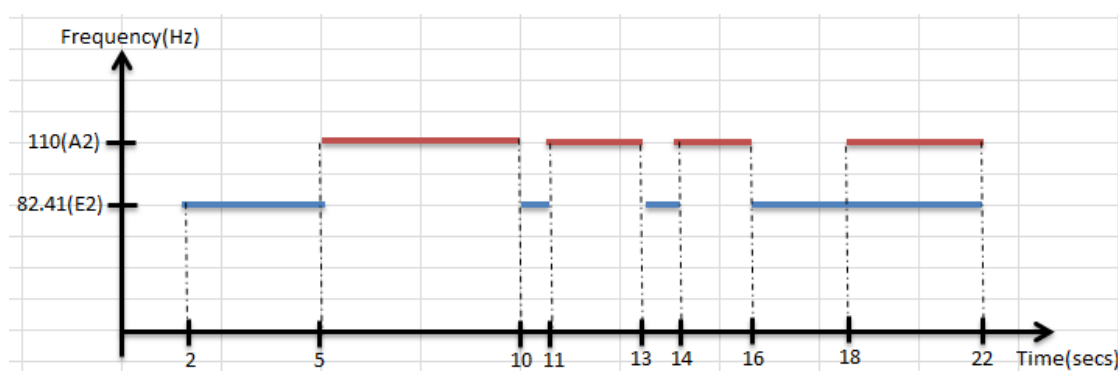


Figure 4.4: Observed Guitar Sound from Video

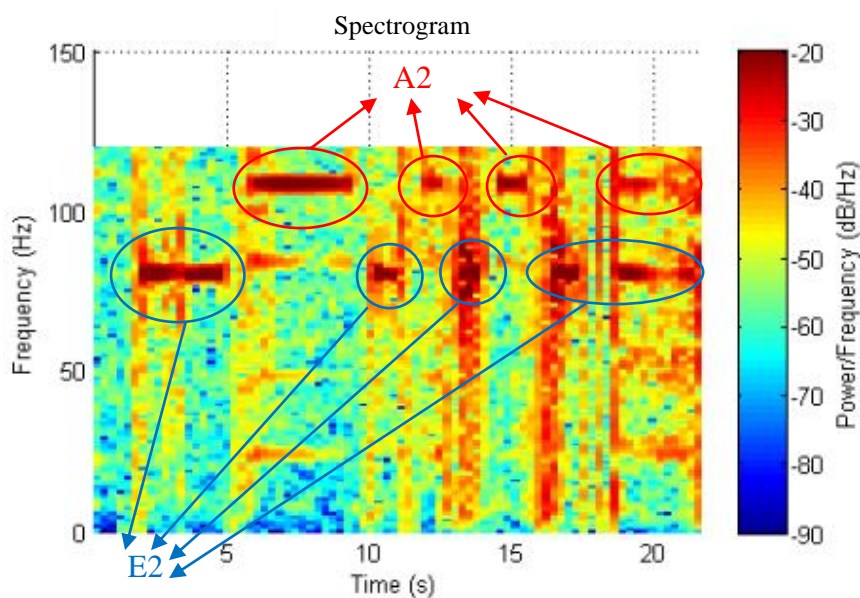


Figure 4.5: Guitar Strings Frequency Spectrum

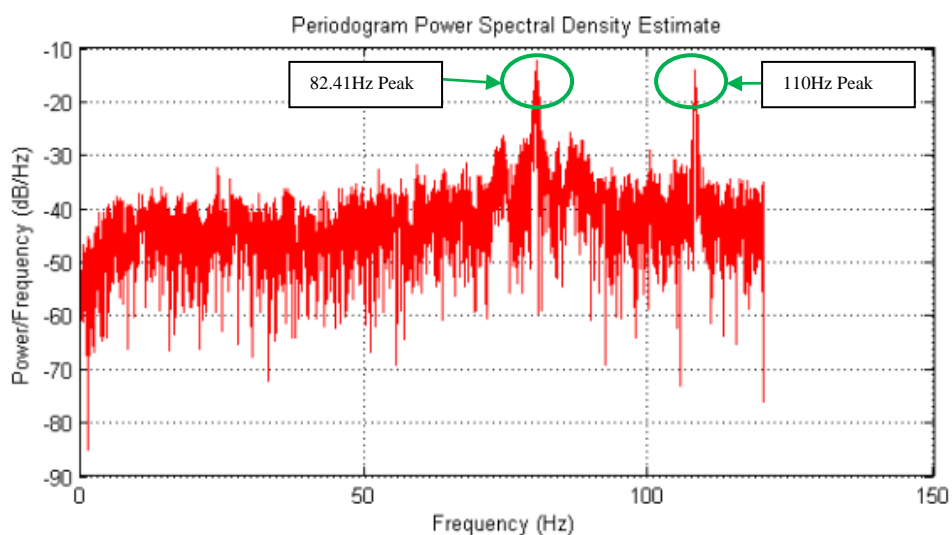


Figure 4.6: Guitar Strings Power Spectral Density Estimate

From the graphs plotted, the recovered sound is not as clear as the expected sound but reveals similar patterns which become more apparent to observers when the sound is played. The power spectral density estimate also shows two peaks which represent both of the guitar strings sound which are the dominant frequencies in the video.

4.2.2 Diaphragm of Speaker

The plotted graphs of the best results of each source of sound which are recovered from a diaphragm of the speaker will be displayed in this sub-section.

Table 4.1: Frequency Sweep Diaphragm Results

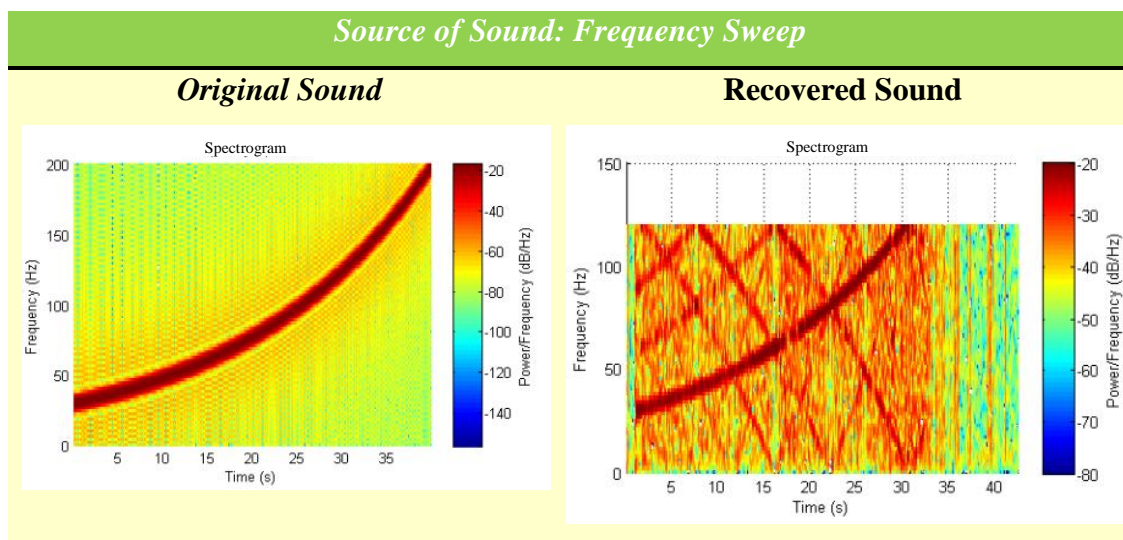


Table 4.2: Self-Constructed Audio (Simple) Diaphragm Results

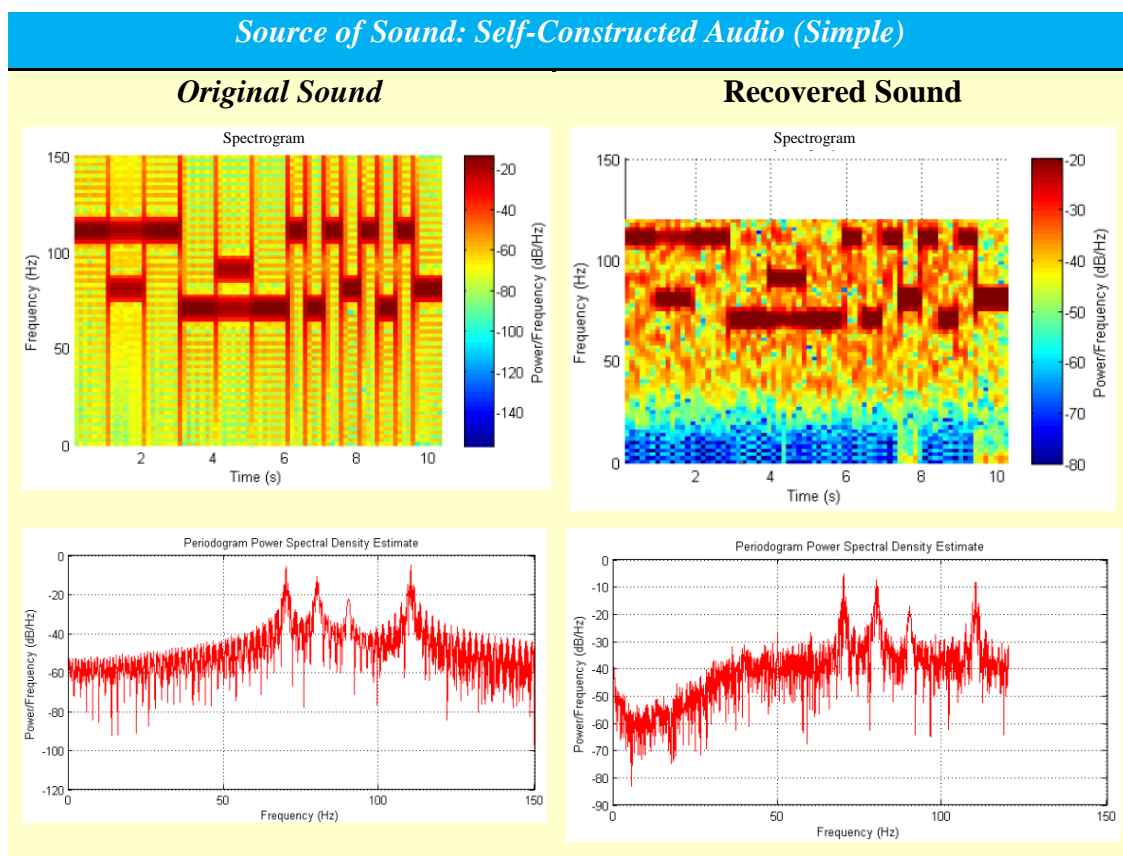


Table 4.3: Self-Constructed Audio (Complex) Diaphragm Results

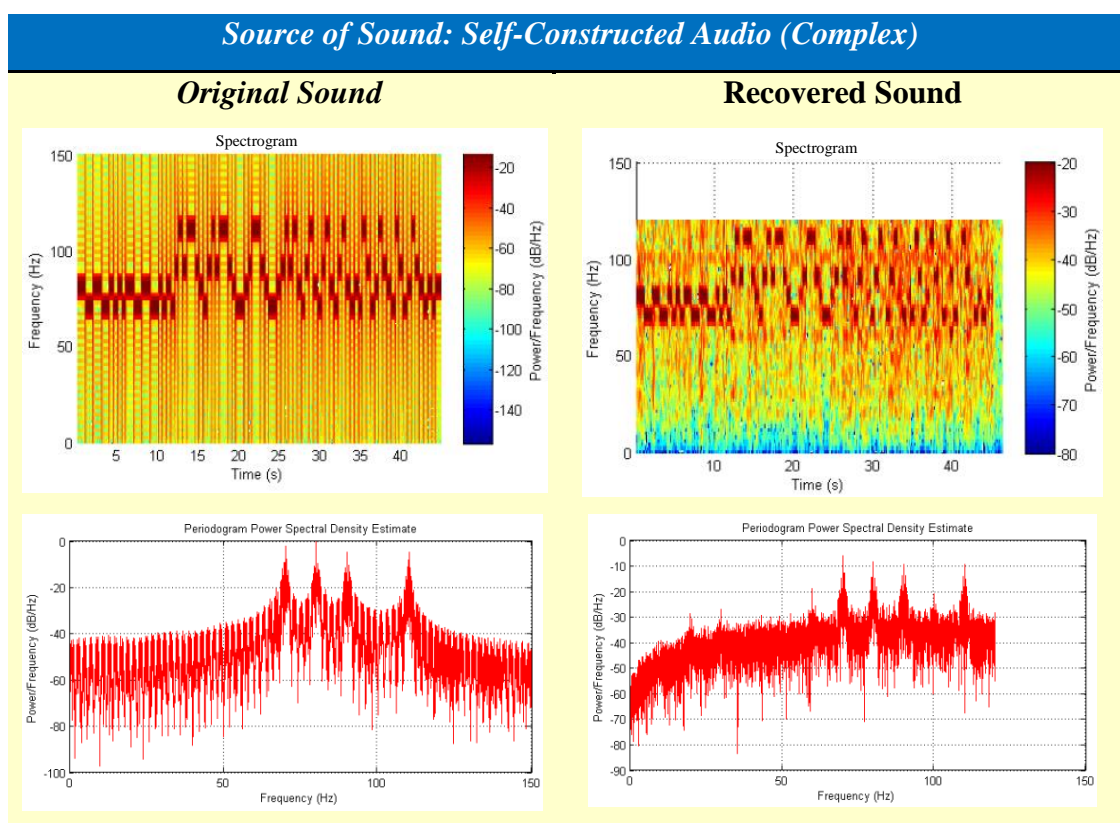
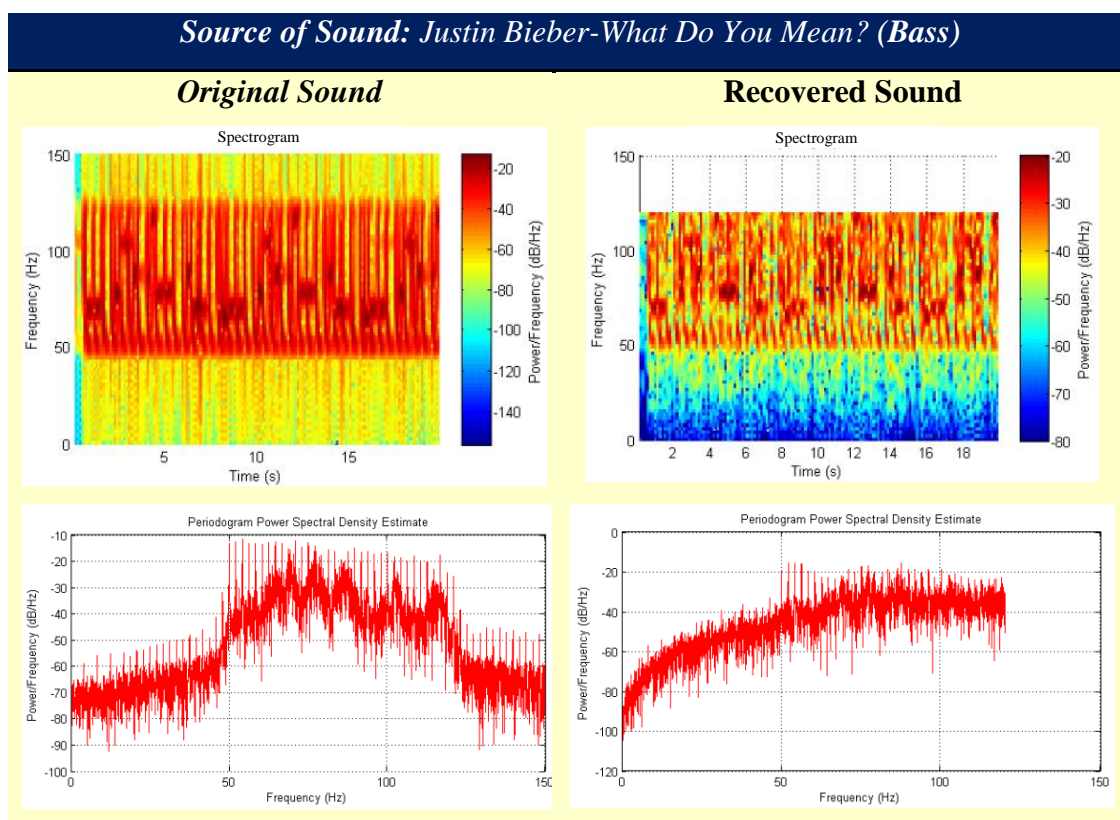


Table 4.4: 'Justin Bieber-What Do You Mean?' Bass Diaphragm Results



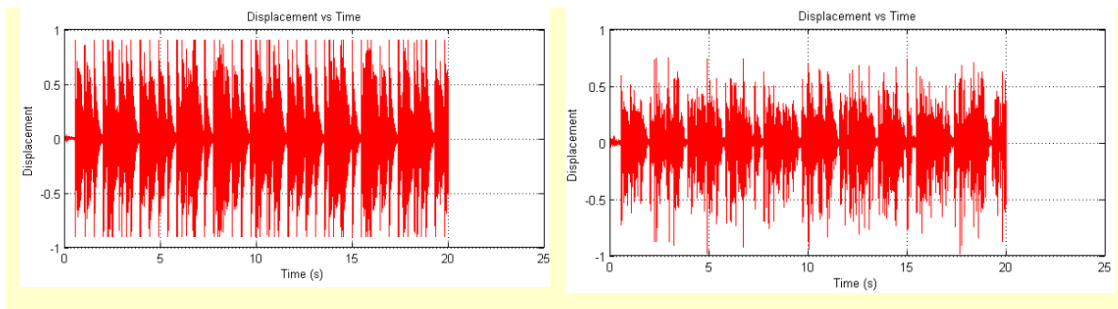


Table 4.5: ‘Black Eyed Peas Ft Justin Timberlake - Where Is The Love?’ Bass Diaphragm Results

Source of Sound: Black Eyed Peas Ft Justin Timberlake-Where Is The Love? (Bass)

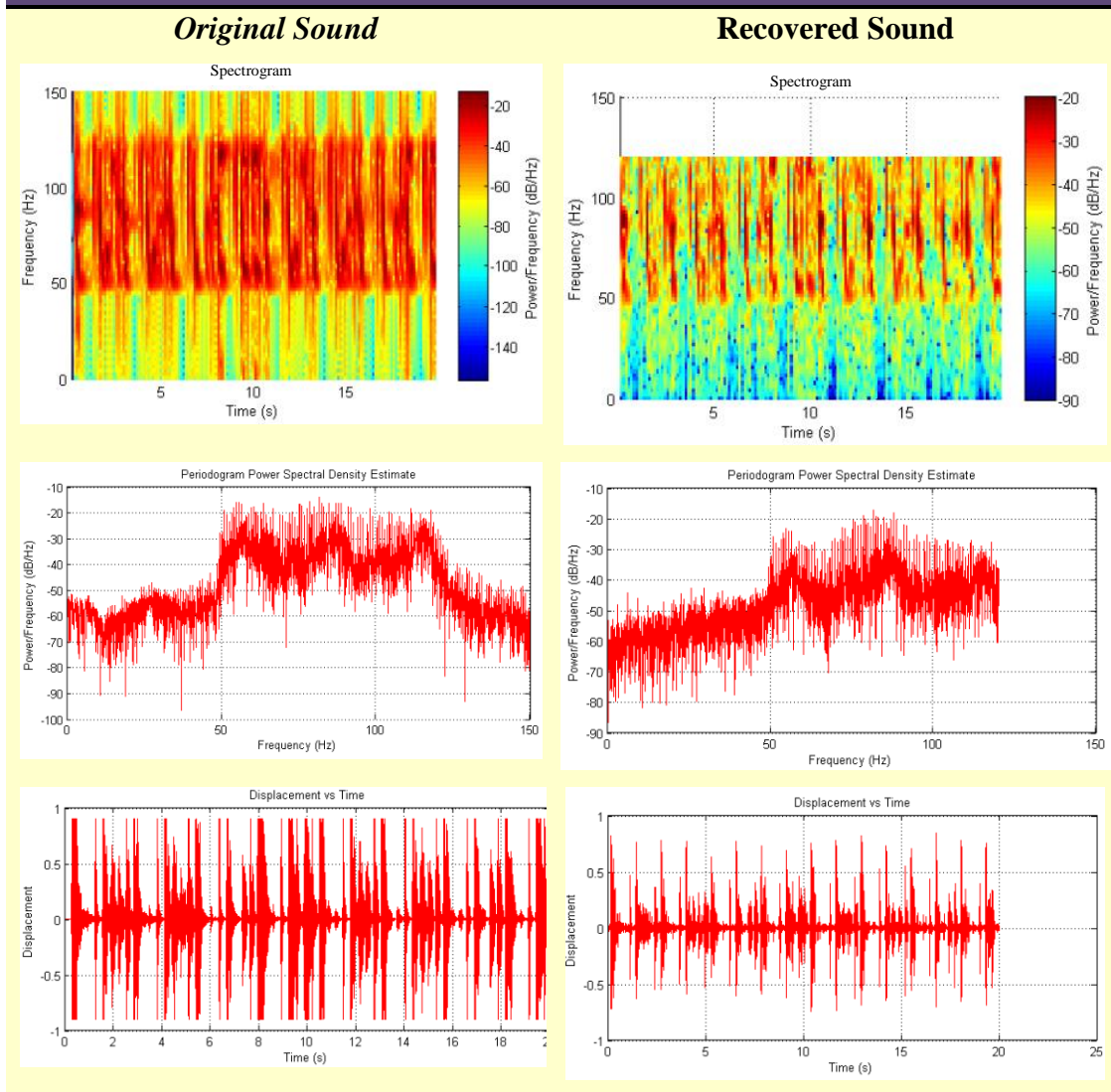
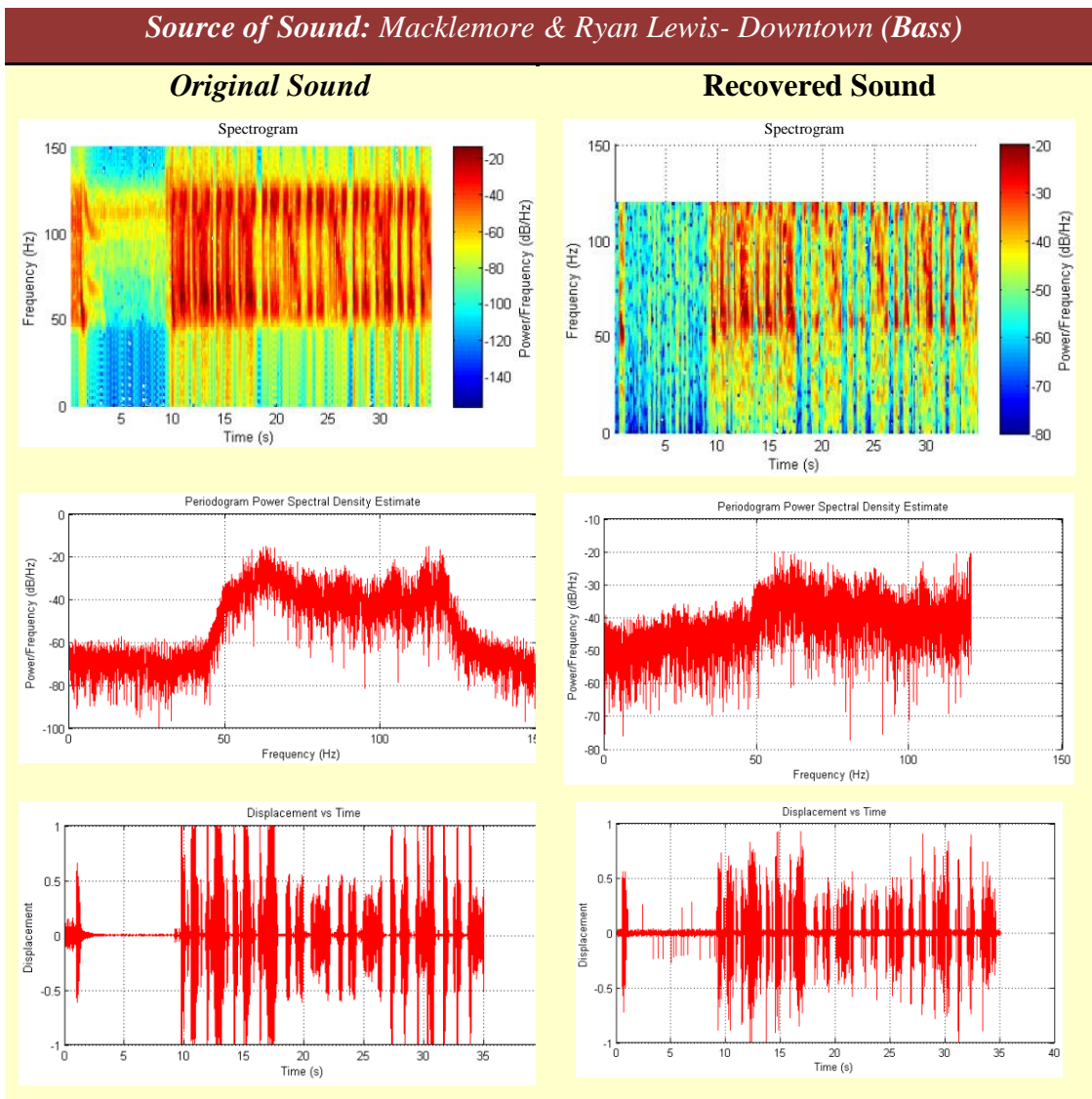


Table 4.6: ‘Macklemore & Ryan Lewis- Downtown’ Bass Diaphragm Results



4.2.3 Plastic Bag

As not all the sources of sound were managed to be recovered; only the graphs of the successful ones will be displayed.

Table 4.7: Frequency Sweep Plastic Bag Results

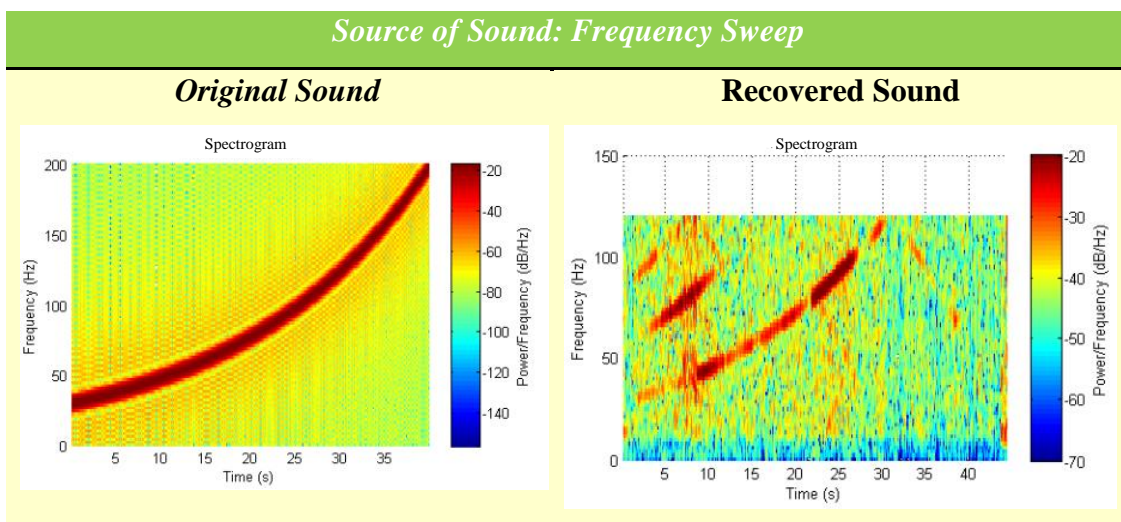


Table 4.8: Self-Constructed Audio (Simple) Plastic Bag Results

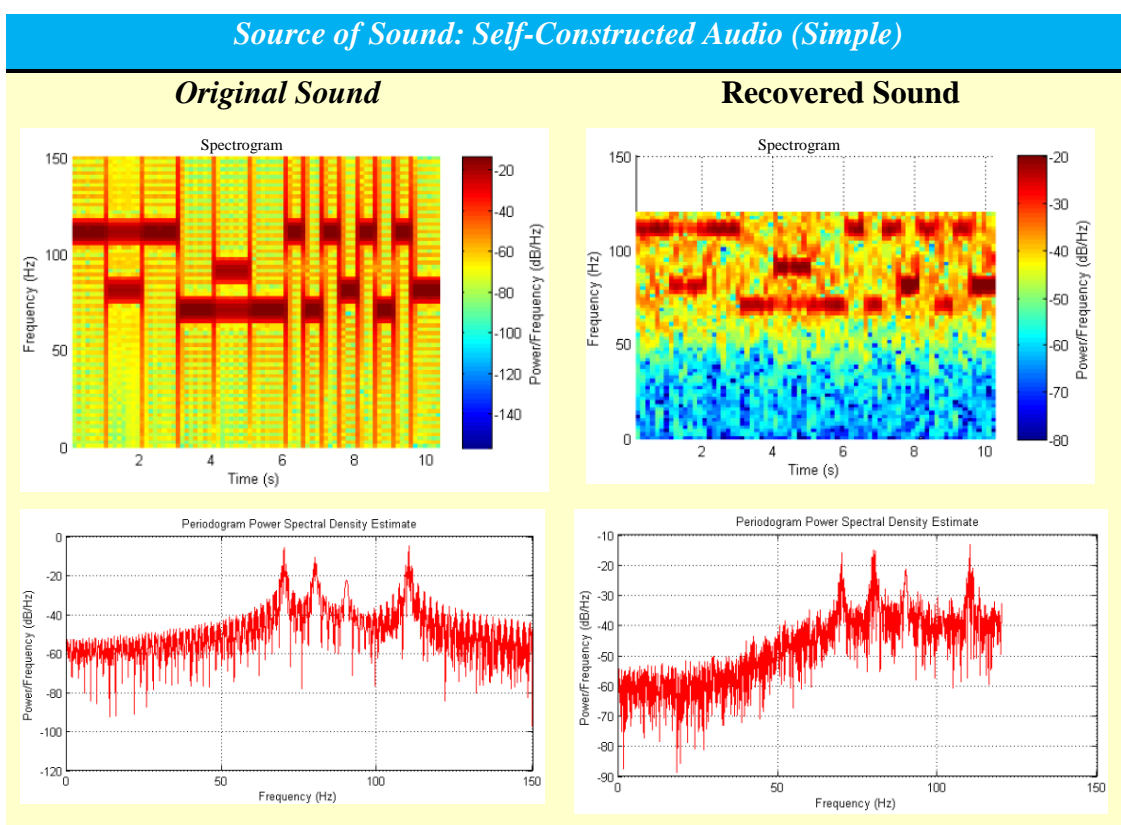


Table 4.9: Self-Constructed Audio (Complex) Plastic Bag Results

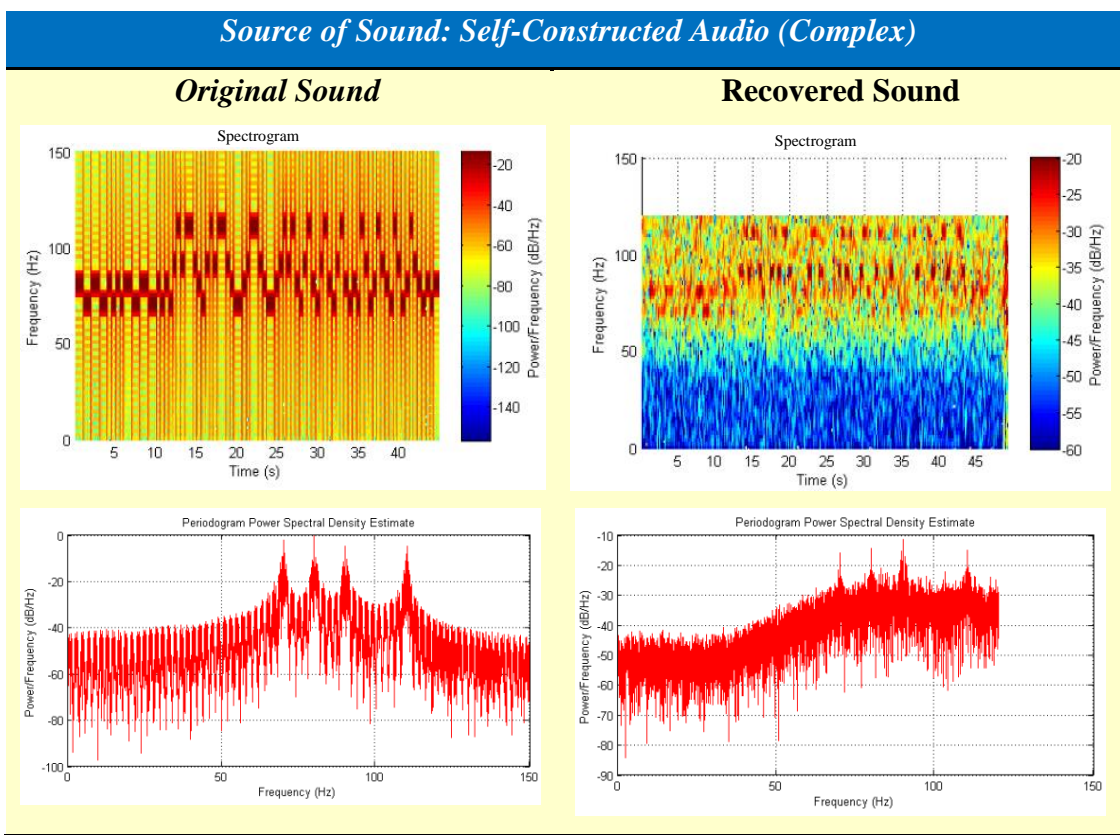
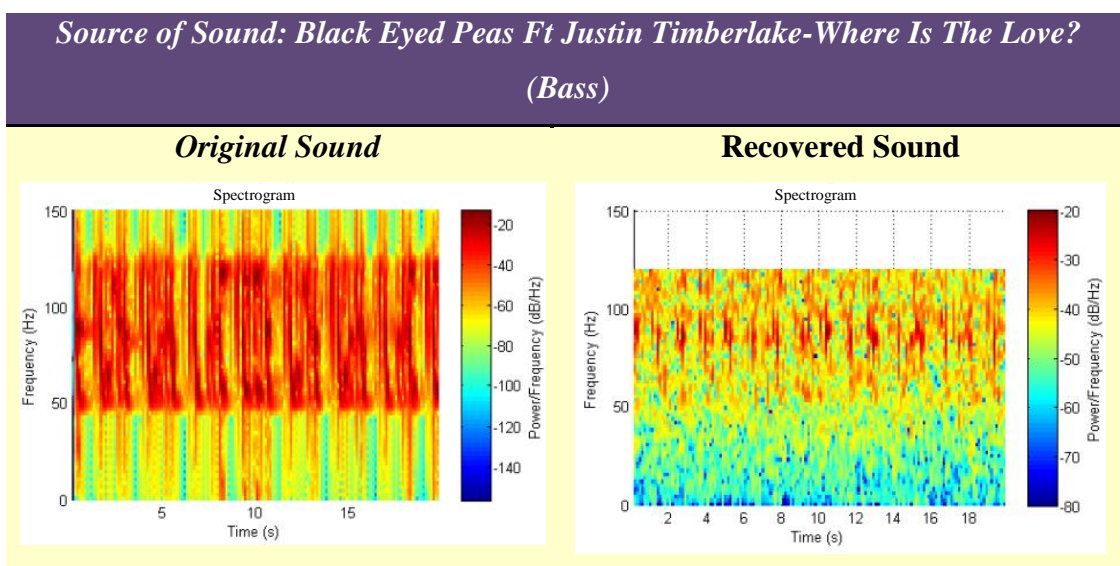
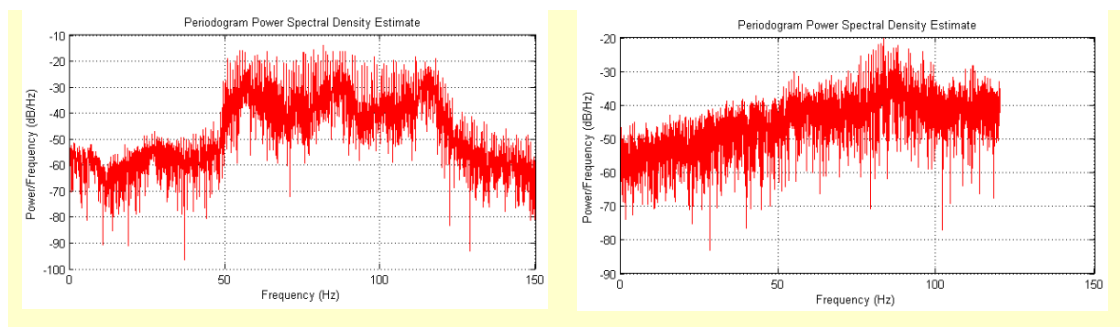


Table 4.10: ‘Black Eyed Peas Ft Justin Timberlake - Where Is The Love?’ Bass Plastic Bag Results

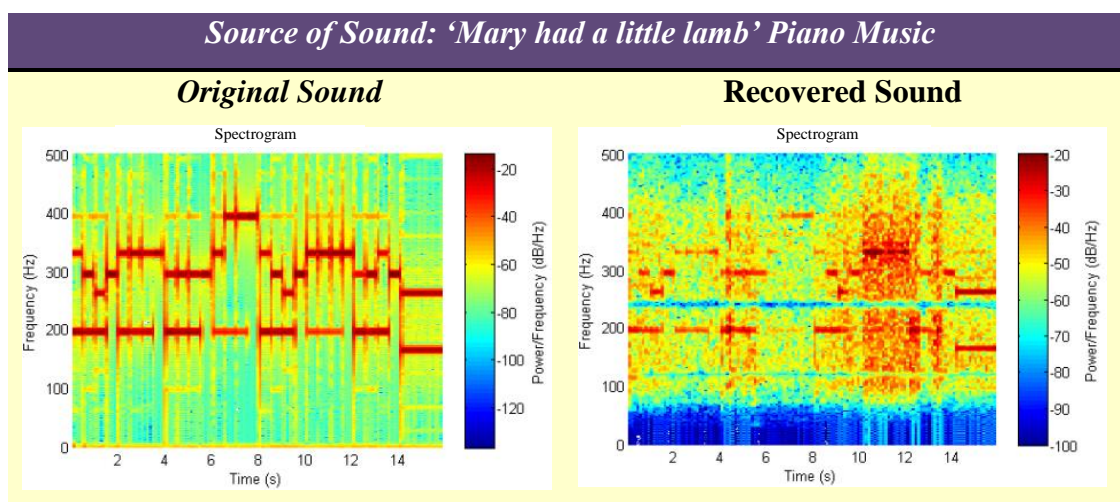




4.2.4 High Speed Video of Bag of Chips

Lastly, a high speed video footage is processed and the recovered sound was successfully retrieved. The table below shows the comparison in the frequency spectrums between the original and the recovered sound.

Table 4.11: ‘Mary had a little lamb’ Piano Music Bag of Chips Results



4.3 Interpretation of Graphs

4.3.1 Displacement vs. Time

The displacement vs. time graph basically represents the volume of the sound. Other representations are usually pressure as sound waves are also a form of pressure waves. For convenience, the term displacement is used since the signal obtained is

derived from small vibrations of the object. The larger the displacement range at a certain period of time, the louder the sound will be. Loud and quiet parts of the sound can be determined through this graph.

Also, it is a useful tool when filtering the sound to ensure that the signal does not get clipped as it reaches its maximum value. However, the displacement vs time graph does not give information regarding the pitch or frequencies of the sound. Figure 4.7 will provide some useful information from the graph.

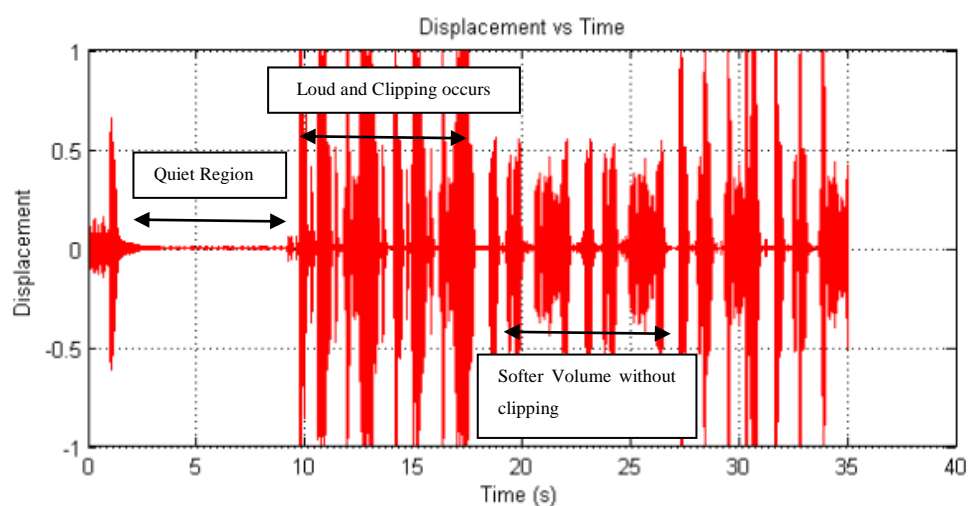


Figure 4.7: Displacement vs Time Graph

4.3.2 Power Spectral Density Estimate

The periodogram power spectral density estimate graph estimates the intensity of the range of frequencies from the audio data. It is able to detect and display major frequencies which are present throughout the entire length of the audio input. This tool is useful and simpler than the spectrogram for sound with single or several frequencies. As the audio becomes more complex, the spectrogram will perform better for analysis instead. The unit for the density measure is power per unit frequency (dB/Hz). The graph will be demonstrated using the self-constructed audio consisting of 4 frequencies as the source of sound.

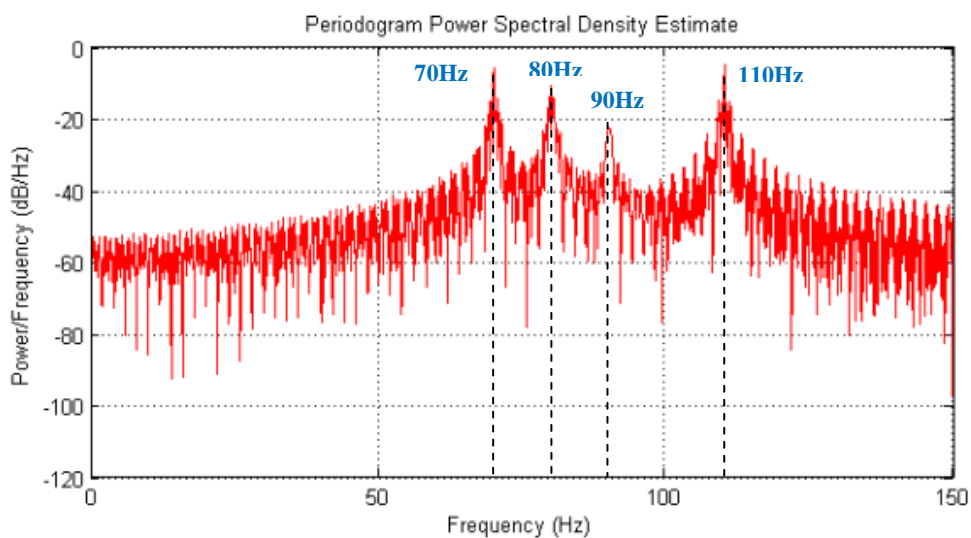


Figure 4.8: Power Spectral Density Estimate Graph

From most of the recovered sound graphs plotted in the experimental results, it can be noticed that there is a cut off after 120Hz in the graph. This is because of the limitation of the camera specification and algorithm which is only able to recover sound up to 120Hz due to sampling rate of 240 per second despite the source of sound having signal above that value. Figure 4.9 shows the recovered sound version from the original sound in Figure 4.8

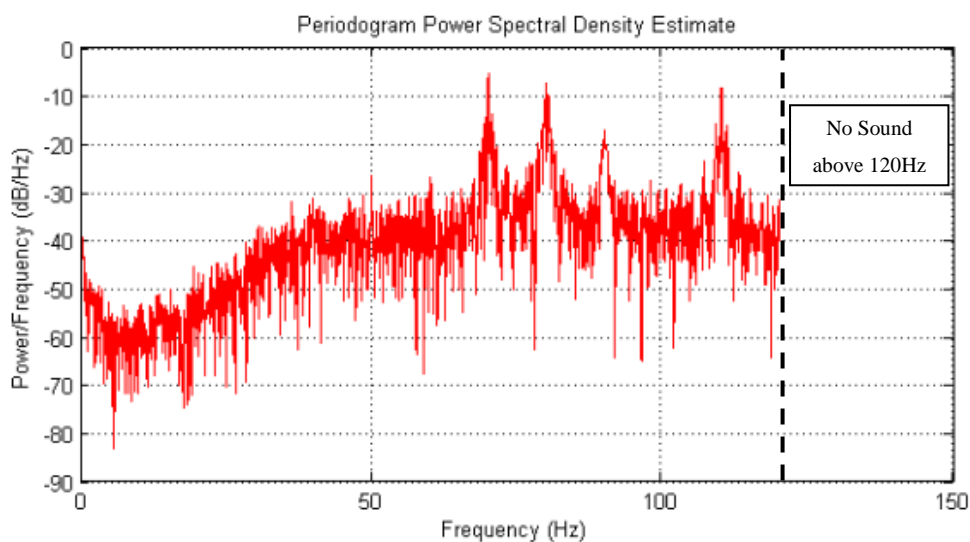


Figure 4.9: Power Spectral Density Estimate 120Hz Limit Graph

4.3.3 Spectrogram

The spectrogram is similar with the power spectral density estimate as both measure the frequency information of the sound. The major difference is that an additional dimension is present in the spectrogram which is time. The density of the frequencies is distinguished with colour while the other two axis are time and frequency respectively. The same unit for density is used which is dB/Hz. A colour bar on the right shows the values for the range of colours. Comparison between original and recovered sound are analysed much better for more sophisticated sound as the spectrogram represents more details than the displacement graph or periodogram.

Looking at Figure 4.10, by comparing with the periodogram in Figure 4.8, the spectrogram can also view the 4 dominant frequencies at any particular time. The identification of which sound frequency is played at a certain time can be determined and observed.

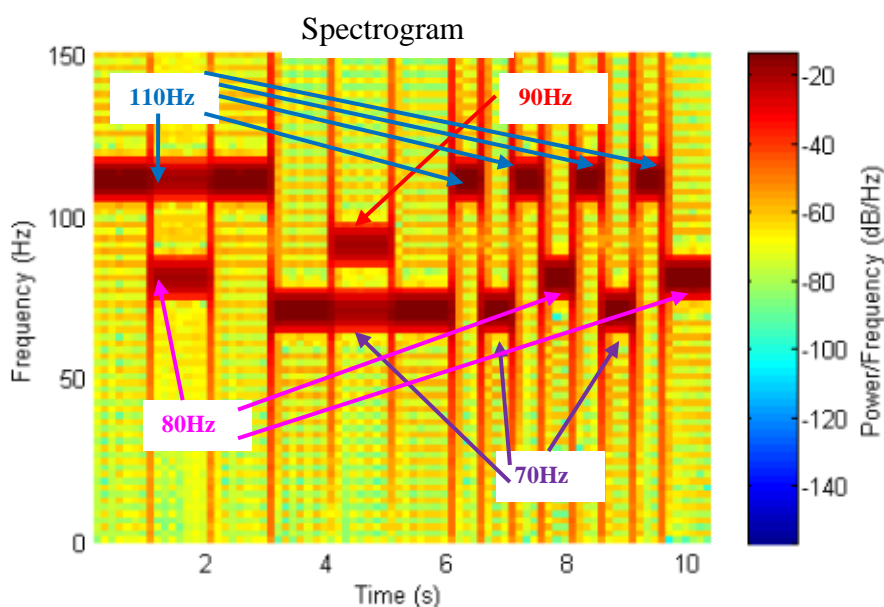


Figure 4.10: Spectrogram

Similar to Figure 4.9, most of the recovered sound's spectrogram does not display any values above the 120Hz frequency due to the same reason. A recovered sound of the frequency sweep is taken as an example. The original sound increases exponentially from 30Hz to 200Hz. However, the recovered sound only reproduces

the signal from 30Hz to 120Hz due to the very same reason. Another noticeable observation is that other sound patterns can be seen other than the major frequency sweep. It can be deduced that these sound patterns are produced due to harmonics as their frequencies are positive multiples of the original frequency sweep. The other decreasing pattern is due to the alias frequency based on the folding diagram if the frequencies exceed the Nyquist frequency.

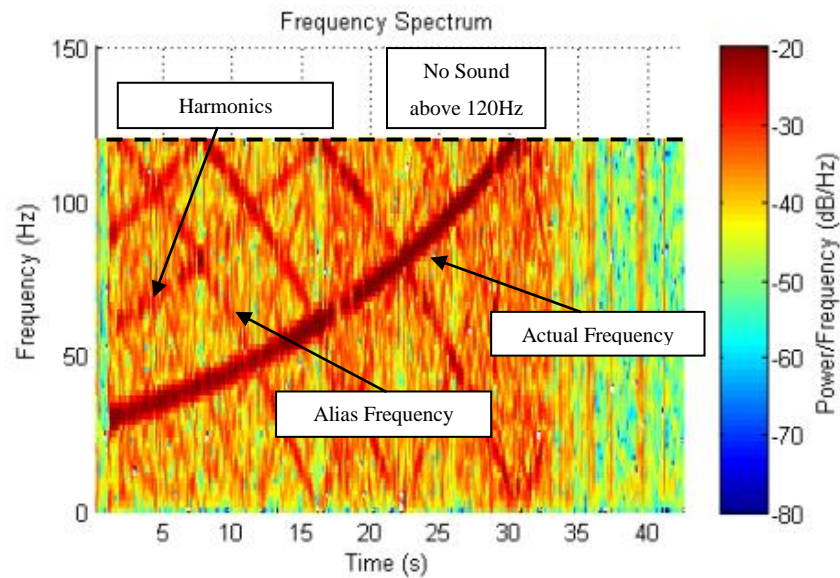


Figure 4.11: Spectrogram 120Hz Limit Graph

4.3.4 Waterfall Plot

The waterfall plot is a higher dimension view of the spectrogram. In other words, the spectrogram is a 2-D dimension while the waterfall plot can be viewed in 3-D. It provides a different perspective and better visualisation in the view of the densities of the frequencies besides the colour representation. Figure 4.10 is a waterfall plot representing the same signal as in Figure 4.9.

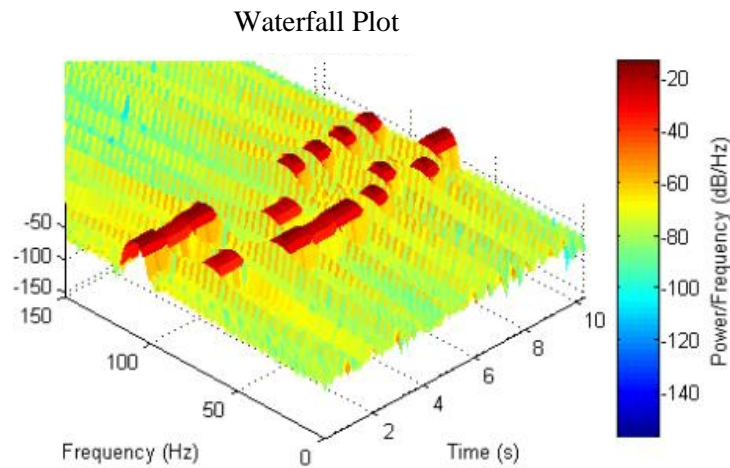


Figure 4.12: Waterfall Plot

4.4 Analysis of Control Factors

4.4.1 High Resolution vs. Low Resolution

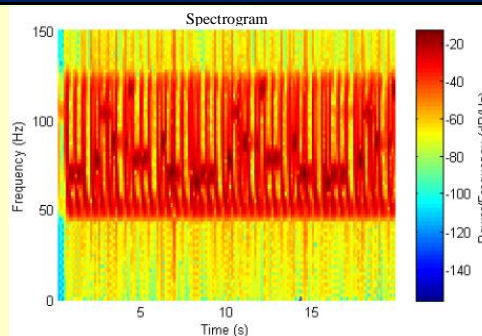
Based on the results of the sound recovered from the diaphragm of the speaker with different resolutions, one using the IphoneSE with resolution of 1280 x 720 and the other using a digital camera Canon SX280HS with resolution of 320 x 240, it is discovered that the higher resolution tends to produce a better sound quality than the low resolution. This agrees with the theory that as there are more pixels on the object which contribute to more samples, then the quality of the sound recovered would increase as well.

It should be noted that the results obtained are not the best with the resolution as the control factor but still audible. This is due to other factors such as the experimental setup where illumination and distance of the object from the camera also play an important role which will be highlighted in the next section. Just by comparing the same environmental setup with only the change in resolution, the graphs will be displayed to compare how the resolution can affect the results. The lower resolution has fewer details than the higher resolution.

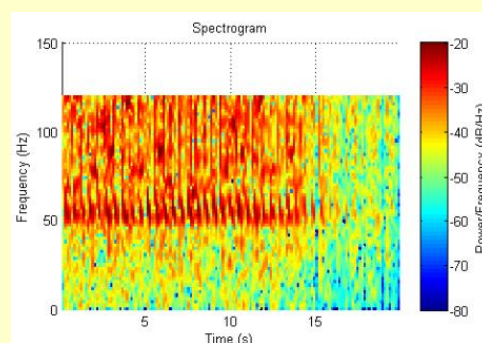
Table 4.12: High Resolution vs. Low Resolution Results

Source of Sound: Justin Bieber-What Do You Mean? (Bass)

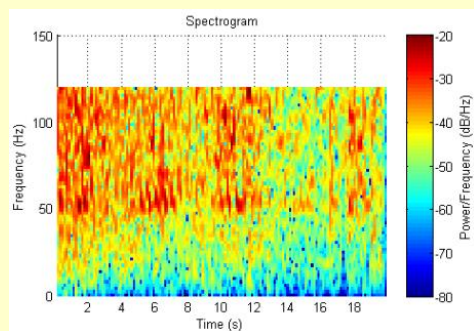
Original Sound



Resolution: 1280x720



Resolution: 320x240



4.4.2 Region of Interest (ROI)

There are mainly two ROIs used from the diaphragm of the speaker. A SpongeBob sticker is used to provide texture to the diaphragm surface. The first ROI selects the head of the SpongeBob while the other selects the hand and leg of the SpongeBob. The ROIs can be seen as shown in Figure 4.13 and 4.14.



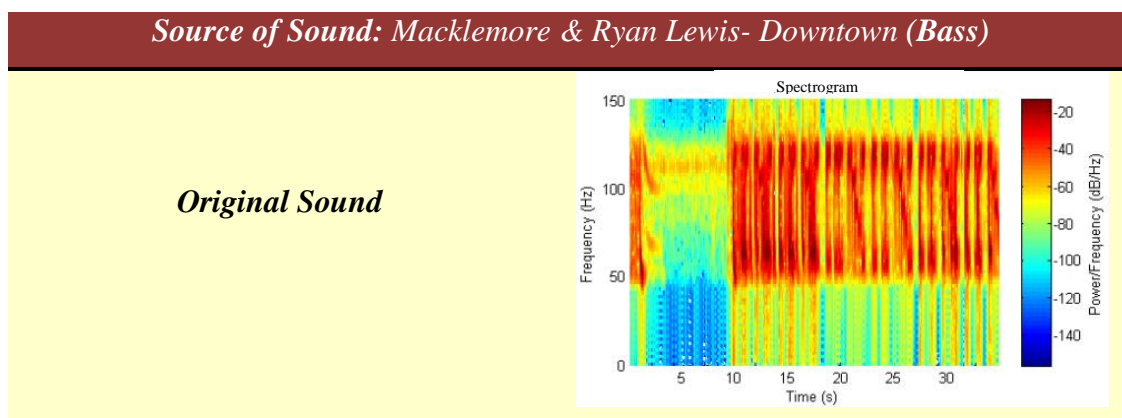
Figure 4.13: ROI 1

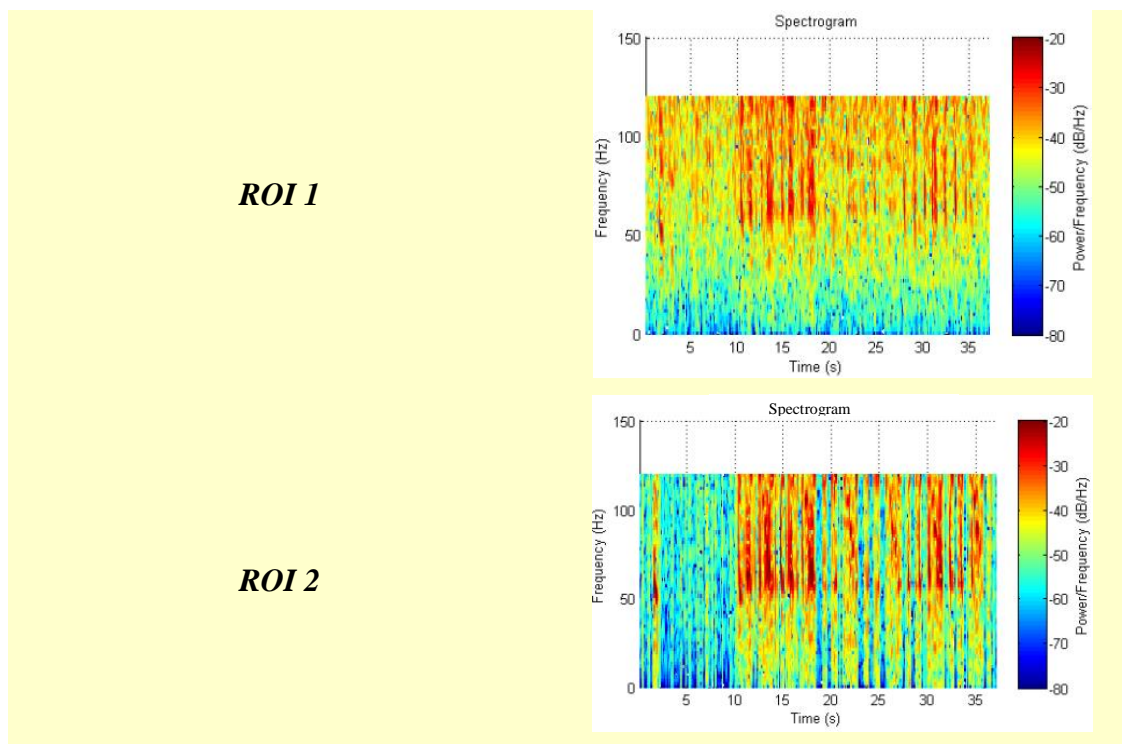


Figure 4.14: ROI 2

The results show that ROI 2 produces better sound quality than ROI 1. We can infer that a ROI with more contrast between the background and the image will estimate a better sound recovery. Consequently, if the region selected is mostly filled with the same background or colour, then it will recover sound poorly because the change in colour would be small. This makes sense as the algorithm takes the weighted average of all the pixels in the region of interest in a frame for each sample.

Table 4.13: ROI Comparison Results





4.4.3 Digital Filtering

Before processing the video into the algorithm, the video is pre-processed using median filter to produce another version of the video to be process and compared. The median filter used is a 5 by 5 kernel. The effect of this median filter did not bring any significant changes to the results when compared without the median filter. The size of the filter kernel can be altered to find the most suitable size but up to the latest results, the video without median filtering will be preferred despite the filter not reducing the quality, it does not increase the quality either.

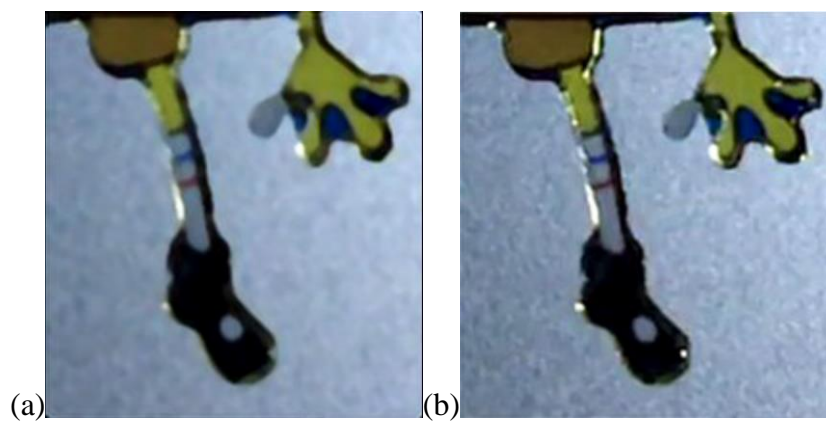
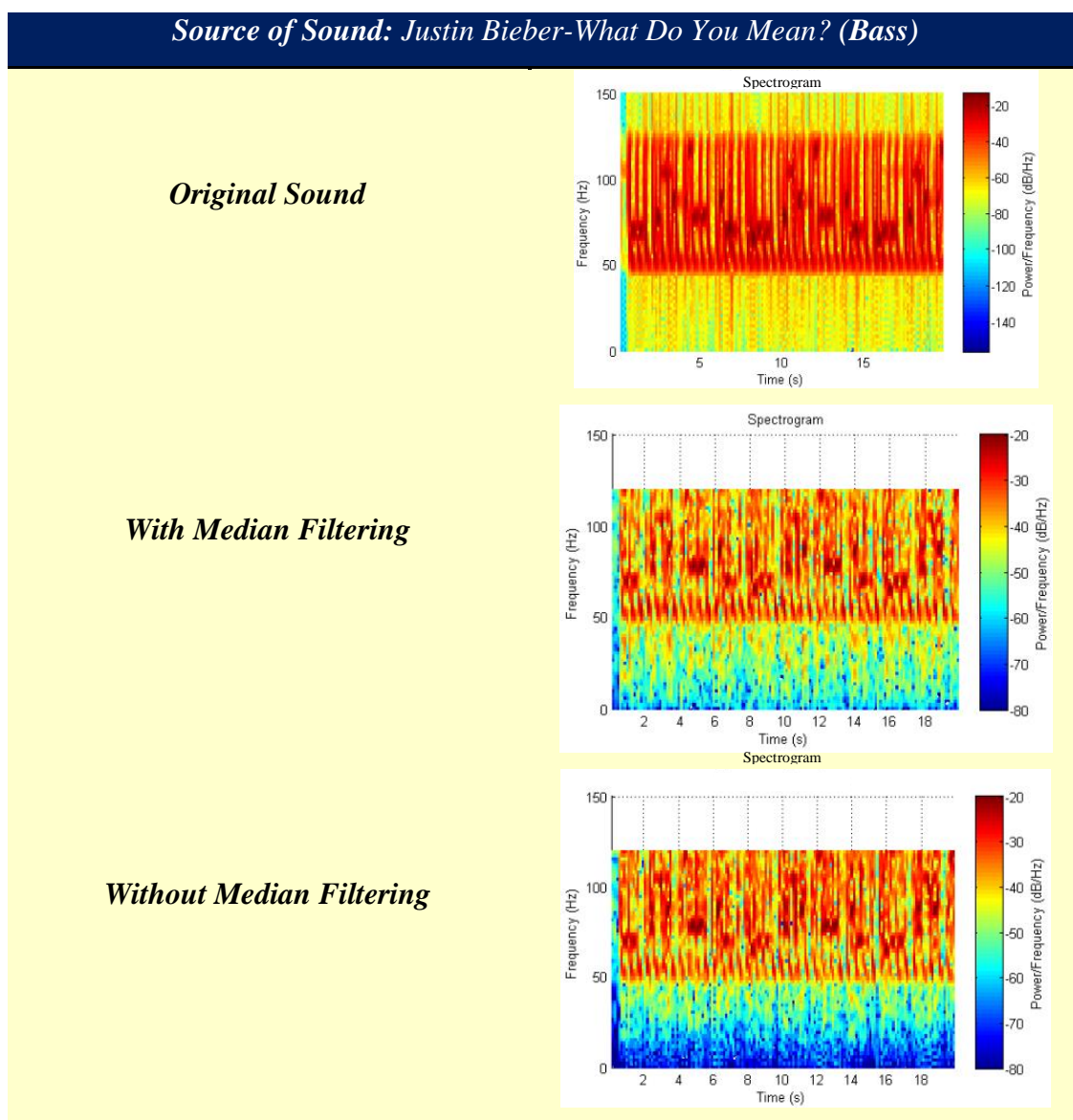


Figure 4.15: Median Filtering (a) with (b) without

Table 4.14: Median Filtering Comparison Results

Source of Sound: Justin Bieber-What Do You Mean? (Bass)



4.4.4 Volume Level of Speaker

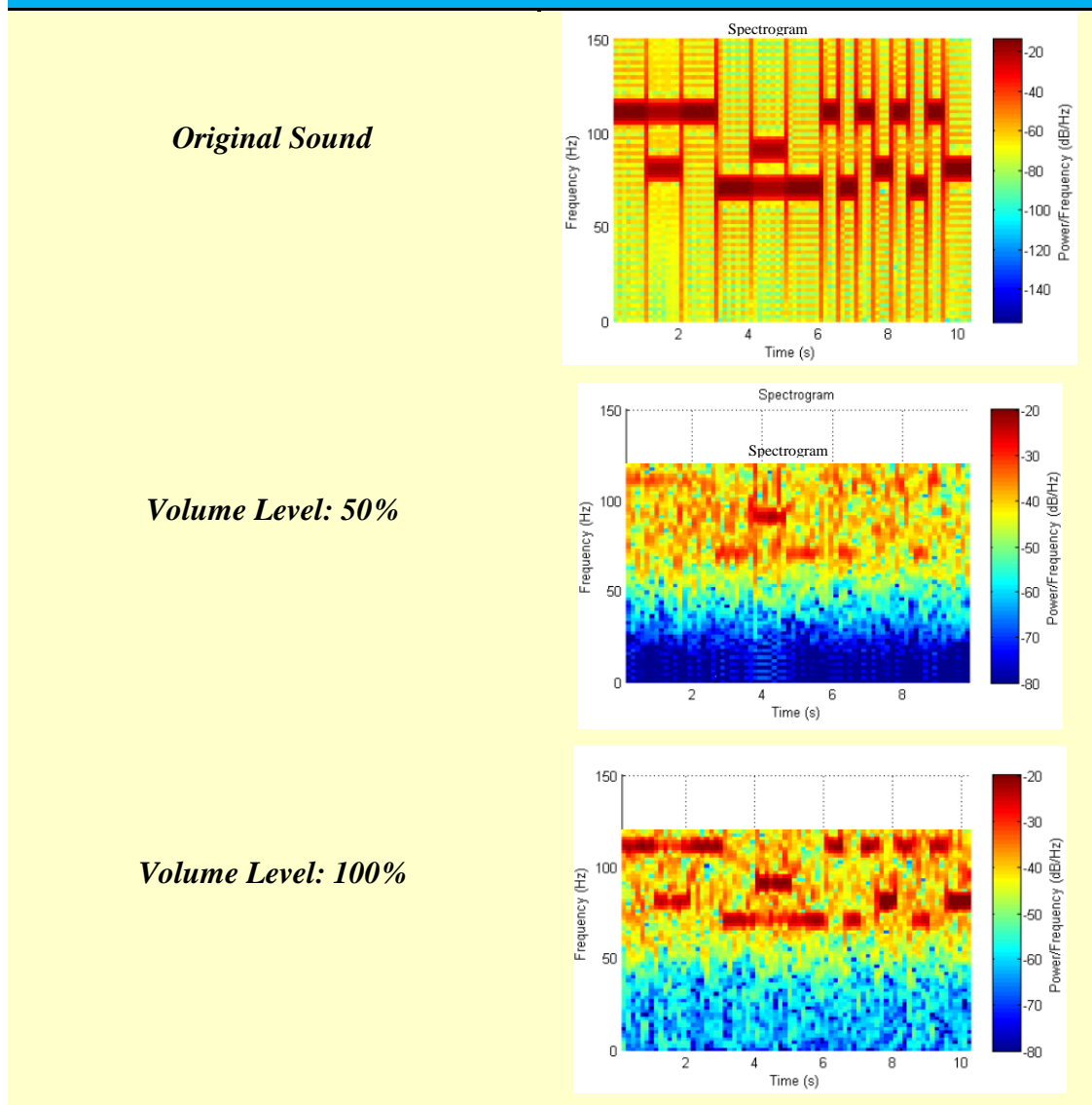
There is no clear relationship between the volume level of speaker and the quality of the recovered sound. In recovering the sound from the diaphragm of the speaker, certain sounds are better when a volume level of 50% but others are better when a volume level of 100% is used instead. For the sound recovered from the plastic bag, there is a direct relationship where all the conditions with volume level of 100% produce a clearer recovered sound.

The volume of the sound has an approximate linear relationship with the displacement that it may cause. The reason being could be that the optimum condition requires the vibrations to be not too small and not too large. In the case of the diaphragm, the vibrations may be apparent enough such that the lowest volume is suitable but not for the case of the plastic bag. If the vibrations are too large, the algorithm does not work as efficient as detecting more subtle vibrations. For the case of the plastic bag, the volume can still be increased because the vibrations are still subtle.

Table 4.15 will list the source of sound from the diaphragm and its corresponding volume which responds to the best results. Table 4.16 will display an example of the recovered sound from the plastic bag with two levels of volume used as comparison.

Table 4.15: Optimum Volume Level of Speaker

Source of Sound	Volume Level
Frequency Sweep	50%
Self-Constructed Audio (Simple Version)	50%
Self-Constructed Audio (Complex Version)	50%
Justin Bieber – What Do You Mean?	100%
Black Eyed Peas Ft Justin Timberlake – Where Is The Love?	100%
Macklemore & Ryan Lewis- Downtown	50%

Table 4.16: Volume Level of Sound Comparison Results*Source of Sound: Self-Constructed Audio (Simple)*

Hence, the optimum volume is very dependent on the particular source of sound and source of object and yet to be able to be directly determined.

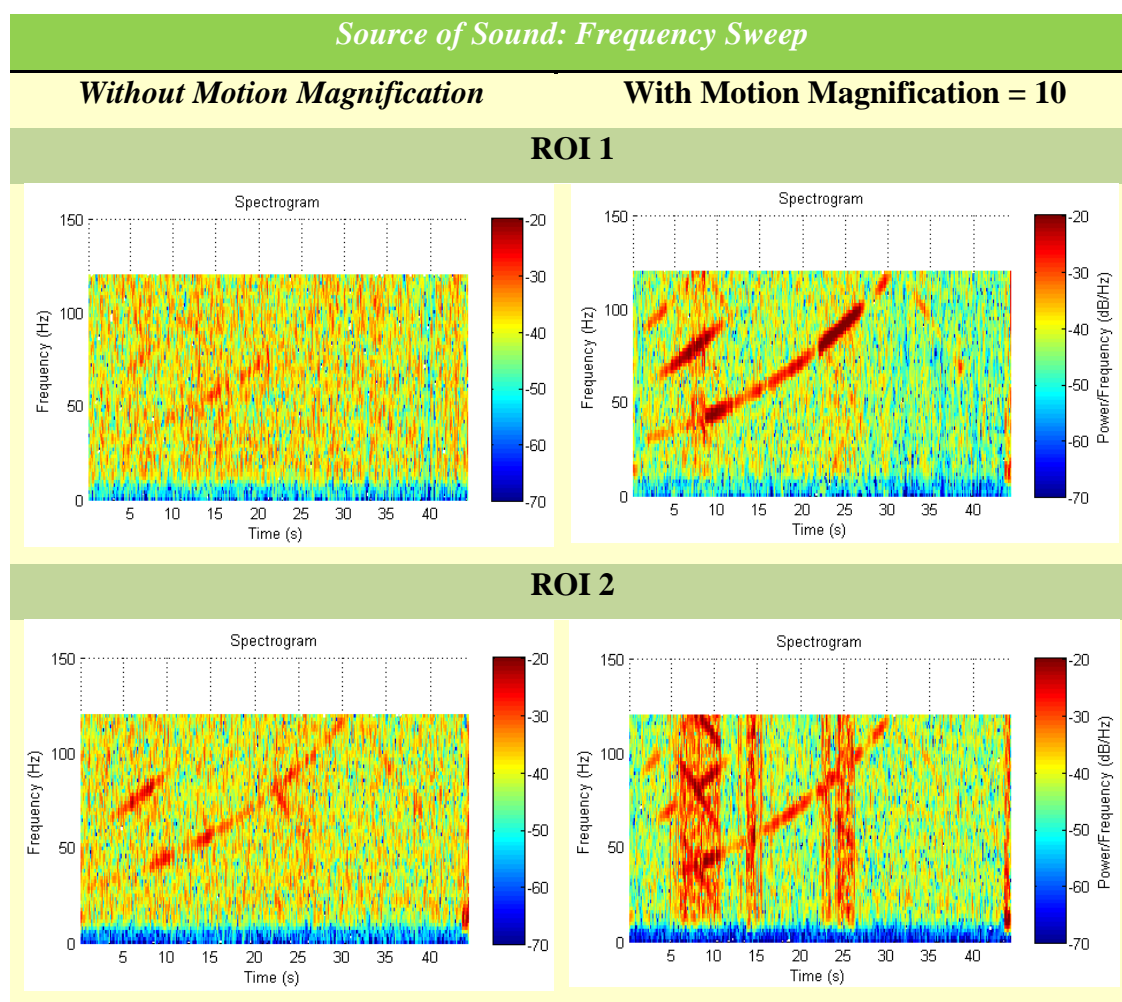
4.4.5 Implementation of Motion Magnification

The presence of the motion magnification behaves similarly with the change in the volume level of speaker. It is found that the implementation of motion magnification is also uncertain as it improves the sound for certain results but worse in others. Hence, an optimum magnification factor has to be identified by varying the value and determining the parameter that corresponds to the best result.

The similarity between the motion magnification and the volume level of the speaker is that they have a linear relationship with the displacement of motion. The difference is that the alteration is done in different steps of the process. Volume level of the speaker affects motion right from the source before processing the video whereas motion magnification is a property that amplifies motion properties inside the algorithm itself. Theoretically, it simulates how the subtle vibrations would behave when the vibrations are magnified. It gives a sense that the vibrations are large before the algorithm converts it into audio data.

Table 4.17 will show an example each of when the motion magnification factor improves and deteriorate the recovered sound. Both the recovered sound is retrieved from the same video of the plastic bag but with different ROIs. The magnification factor applied is 10.

ROI 1 without magnification shows a weak frequency sweep signal which is barely visible. Once the motion magnification is applied, the frequency sweep becomes more distinct and improves the sound despite having some harmonics. In the case of ROI 2, its recovered sound without magnification shows a more apparent frequency which is still not as strong as ROI 1 with magnification. Unfortunately when applied with the same magnification factor, the signal is covered with more harmonics and noise instead of its actual sound. Hence, the magnification factor is to be chosen wisely.

Table 4.17: Motion Magnification Comparison Results

4.5 Highlighted Challenges

4.5.1 Brightness of Video

One of the challenges encountered in setting up the experiment is the brightness of the video. It is noticeable that the higher frame rate video gives a lower overall brightness than the typical 30 or 60 frame rate video. This is probably due to the less time exposure for each frame as the number of frames per second increases tremendously. As a solution, an additional illumination is provided to the source of object for an increased illumination on the object. The illumination should not be too low such that the video becomes too dark and the features cannot be seen but should not be too much until the light reflections cover the features of the object.

Another challenge is that if the experiment is carried under the presence of fluorescent light, flickering may be present in the video at about 120Hz which is twice the frequency of the alternating current. Hence, a constant LED flashlight is used to provide the illumination instead.

4.5.2 Adjusting Volume Level of Speaker

One of the most difficult challenges is to determine the volume level of the speaker. First of all, it is already difficult to determine the optimum volume which produces the best result as discussed in the previous section. The other challenging aspect is to measure the actual sound level as each of the sound may have different levels. The current method is to fix the potentiometer at one position and only adjust the volume percentage digitally through laptop. A better approach is to purchase a decibel meter or also known as the sound level meter to measure the sound in decibels quantitatively which provides a much better reference.

4.5.3 Induced Vibration on Camera

In one of the attempts of the experiments, it is discovered that the camera vibrated due to the vibration induced from the speaker to the floor, to the camera stand and eventually to the camera. The vibrations of the speaker were so strong that it was able to transmit the vibrations significantly through the floor as a medium. The vibrations induced on the camera were significant because the entire video footage can be observed shaking overcoming the vibrations on the plastic bag. Surprisingly, there was still some sound recovered from the experiment which shows that sound can possibly be recovered visually in other mediums as well. This brings up other potential studies in the field of recovering sound visually but will not be in the scope of this project.

As the desired medium is through air from the speaker to the object, a solution to reduce the vibrations induced through the floor is to use materials that have good vibration absorption such as putting carpet on the floor and a stiffer table.

4.5.4 Distance between Camera and Object

Due to the unavailability of zoom lenses, the distance between the camera and object has become a limitation. The camera is set up as close as possible to the object to ensure the best quality of sound is recovered because more pixels can represent the object and thus, more samples are available. Such setup might sometimes be a little troublesome as the equipment have to be cramped and to ensure that none of them are having physical contact to each other to reduce the induced vibration as mentioned. With a zoom lens, distance between the camera and object can be varied and the limits can be tested.

4.5.5 Environmental Effects

When it comes to very subtle vibrations, the signal becomes even more sensitive to noise. This criterion is more crucial for the plastic bag rather than the diaphragm as any slight movement may cause the plastic bag to move. For example, the air conditioner in the enclosed room could cause the plastic bag to move and thus, it had to be switched off. Therefore, the ideal condition should be in an enclosed room, no wind factor, preferably soundproof which the most suitable would be the audio laboratory.

4.5.6 High Data Storage

As many videos with different combinations of parameter, sound and object, the amount of data storage required is very high. The high frame rate video also occupies a larger space than a typical frame rate video. It is suggested that an external hard disk is purchased solely for this application so that proper documentation can be prepared without the concern of insufficient storage data.

4.5.7 Determination of Quality of Sound

In this research project, the two methods used to determine the quality of sound is by qualitatively accessing the audibility of the sound and by evaluating visually the similarity of the graphs such as the displacement vs. time, periodogram and spectrogram between both the original and recovered sound. The graphs are a good tool to assist in applying filtering and denoising techniques while re-evaluating them. After that, the methods applied can be verified by accessing the audibility of the sound.

However, these methods instil some subjectivity as the quality cannot be quantified. Some of the quantitative methods that can be recommended in future works are the Log Likelihood Ratio (LLR) by Quackenbush et al. (1998) and Segmental Signal-to-Noise Ratio (SSNR) by Hanson & Pellom (1998). These methods will be described briefly under the recommendations section in Chapter 5.

CHAPTER 5

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

Based on the results obtained with the assessment of the audibility of the sound and evaluation of the graphs, the main aim of discovering a new alternative in recovering and recording sound is fulfilled using the algorithm constructed in this project to process the input videos. The minimum objective which is to recover audible sound through imagery is accomplished and proves that the concept of converting visual information into audio information can be made possible.

The other objectives are also initiated and in progress throughout the period of the research. Tasks are taken to improve the similarity between the recovered and original sound in different areas including median filtering in pre-processing, selection of proper ROI and motion magnification in the main algorithm, and post-processing audio denoising techniques are variously applied.

Recovering more sophisticated sound is also in progress and yet to be completed as the human speech has not been recovered. Sound recovery has been made from guitar strings, simple bass audio to more complex bass music and finally a simple piano music at a higher frequency range. Songs at their typical frequency range and human speech would be the next aim to be recovered. Also, since the sound recovered is also dependent on the source of object, a future objective could be to recover sound from potentially more difficult objects.

In addition, some of the control factors were experimented while recovering sound. They can be distinguished into two categories; hardware and software. The hardware control factors are the volume level of the speaker and resolution of the camera which helps to understand the limitations and capabilities of the algorithm better. The software control factors are the ROI selection, motion magnification and digital filtering which can be manipulated to improve the sound recovery process. Table 5.1 will be tabulated to summarise the optimum values of the control factors.

Table 5.1: Summary of Control Factors

Control Factor	Optimum Condition
Resolution of Camera	High resolution
Region of Interest	High contrast between image and background
Median Filtering	No effect for kernel size of 5x5
Volume Level of Speaker	<ul style="list-style-type: none"> - Audio dependent for diaphragm - Higher volume for plastic bag
Motion Magnification	Good to implement but too high magnification factor may cause more noise.

The limitation of the algorithm is noticed in using different sources of object. With the diaphragm of speaker, the algorithm performs better than the plastic bag. The limitation in recovering more complex sound is also apparent in recovering sound from the plastic bag. The algorithm performs weakly as it recovers the more complex bass music. Sound frequency limit that can be recovered by the algorithm is also limited to the frame rate of the video camera. Thus, several recommendations mentioned in the next section may be able to overcome so of the limitation of the algorithm.

5.2 Recommendations

For the purpose of future improvements, *noise reduction through software* should be a continuous process in improving the quality of sound. Reducing noise at the early stage of the process may be more effective, hence it will be a good approach to

find techniques in the pre-processing and processing of the algorithm. The median filtering may be adjusted with different kernel sizes and test for suitability. Scale and orientation pairs in the main algorithm can be temporally aligned to prevent destructive interference before constructing the audio data.

It would be of no surprise that upgrading the equipment would also improve sound recovery. If the cost is within the budget, a *high speed video camera, zoom lens and data storage* can be considered to be purchased.

The *rolling shutter technique* introduced by Davis, A., Rubinstein, M. et al. (2014) allows normal camera with 60fps to be utilised as well. If this method can be successfully implemented in the algorithm, then the cost to purchase a high speed video camera can be saved while increasing more possible applications.

Another interesting possible application is to utilise the *phase vocoder algorithm* in MATLAB introduced by Ellis, D. (2002). With the phase vocoder, pitch shifting is made possible while maintaining the same duration of time. If the video camera has a limited frame rate, the phase vocoder can be used to shift the pitch of the original sound to a lower frequency range as the source of sound. Once the sound is recovered, the recovered sound pitch is shifted back to the original frequency range to obtain the equivalent original sound. This approach is merely a suggestion as it has not been experimented. The application is more specific as the input has to be intentional but could be useful for video cameras with limited frame rate.

Lastly, a quantitative analysis will be a very good verification for the qualitative assessment approach that has been used and provides more objectivity. The *Segmental Signal-to-Noise Ratio* is a speech quality measure that measures the signal-to-noise ratio of many short segments usually between durations of 15-20ms and then finds the average of all these ratios. This value determines the accuracy of the recovered speech comparing with the original. The *Log Likelihood Ratio* measure calculates how similar the spectral shapes of the recovered sound are with the original sound. These speech quality measures could be useful even for lower frequency range music.

REFERENCES

- Bařina, D., 2011. *Gabor Wavelets in Image Processing*. Proceedings of the 17th Conference STUDENT EEICT ..., (2), pp.2–6. Available at: <http://www.fit.vutbr.cz/research/pubs/all.php?file=/pub/9598&id=9598>.
- Davis, A. et al., 2014. *The Visual Microphone: Passive Recovery of Sound from Video*. Siggraph 2014, pp.1–10. Available at: <http://dl.acm.org/citation.cfm?id=2601119>.
- Gautama, T. & Van Hulle, M. a, 2002. *A phase-based approach to the estimation of the optical flow field using spatial filtering*. IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council, 13(5), pp.1127–1136.
- Rothberg, S.J., Baker, J.R. & Halliwell, N.A., 1989. *Laser vibrometry: Pseudo-vibrations*. *Journal of Sound and Vibration*, 135(3), pp.516–522. Available at: <http://linkinghub.elsevier.com/retrieve/pii/0022460X89907050>.
- Kamble, K. et al., 2015. *A Review: Eulerian Video Motion Magnification*. , pp.2384–2390.
- Rubinstein, M., 2014. *Analysis and Visualization of Temporal Variations in Video*. PhD thesis, Massachusetts Institute of Technology.
- Wang, C.C. et al., 2008. *A new kind of laser microphone using high sensitivity pulsed laser vibrometer*. Conference on Quantum Electronics and Laser Science (QELS) - Technical Digest Series, pp.0–1.
- Wadhwa, N. et al., 2012. *Phase-Based Video Motion Processing*.
- Barrett, D., 2013. *One surveillance camera for every 11 people in Britain, says CCTV survey*. Available at: <http://www.telegraph.co.uk/technology/10172298/One-surveillance-camera-for-every-11-people-in-Britain-says-CCTV-survey.html> [Accessed: 10 April 2016].
- Dailey, K., 2013. *The rise of CCTV surveillance in the US*. Available at: <http://www.bbc.com/news/magazine-22274770> [Accessed: 10 April 2016].

- Swanson, E., 2013. *No clear support for more surveillance since Boston bombing*. Available at: http://www.huffingtonpost.com/2013/04/24/boston-bombings-surveillance-poll_n_3149576.html [Accessed: 9 April 2016].
- admin-s, 2013. *RM190 Juta Untuk 1, 200 CCTV di Kuala Lumpur?* Available at: <http://www.malaysia-today.net/rm190-juta-untuk-1200-cctv-di-kuala-lumpur/> [Accessed: 8 April 2016].
- Gan, P.L., 2012. *Question of safety*. Available at: <http://www.selangortimes.com/index.php?section=insight&permalink=20120727105013-question-of-safety> [Accessed: 9 April 2016].
- BBC, 2009. *The statistics of CCTV*. Available at: <http://news.bbc.co.uk/2/hi/uk/8159141.stm> [Accessed: 8 April 2016].
- Chaffin, T., 2011. *CCTV with audio: 3 things to know*. Available at: http://www.2mccctv.com/blog/2011_06_28-3-things-to-know-about-cctv-systems-with-built-in-microphones/ [Accessed: 10 April 2016].
- Nadirah, H.R. and Natasah, J., 2016. *'High' tech way to fight crime - nation | the Star Online*. Available at: <http://www.thestar.com.my/news/nation/2016/04/14/high-tech-way-to-fight-crime-police-drones-to-help-cops-monitor-crime-on-the-ground/> [Accessed: 14 April 2016].
- Paroc, G. *General information about sound*. Available at: http://www.paroc.com/knowhow/sound/general-information-about-sound?sc_lang=en [Accessed: 16 April 2016].
- Acoustics. *Sound theory*. Available at: http://acoustics.no/sound_measurement/sound_theory/ [Accessed: 18 February 2016].
- ProAV. *Data and information, lists, tables and links*. Available at: <http://www.bnoack.com/index.html?http&&www.bnoack.com/audio/speech-level.html> [Accessed: 14 March 2016].
- Matsuda, D., 2005. *What is Nyquist theorem?* Available at: <http://whatis.techtarget.com/definition/Nyquist-Theorem> [Accessed: 20 February 2016].
- Nave, R. *Microphones*. Available at: <http://hyperphysics.phy-astr.gsu.edu/hbase/audio/mic.html> [Accessed: 24 February 2016].
- Utah. *Gaussian and Laplacian pyramids*. Available at: <http://www.cs.utah.edu/~arul/report/node12.html> [Accessed: 3 February 2016].

- Irwin, D.J., 2003. *Ideal Band Pass Filter*. Available at: https://www.chegg.com/homework-help/definitions/ideal-band-pass-filter-4?adobe_reloaded=true [Accessed: 14 February 2016].
- Agilent Technologies, 2000. *Butterworth Response*. Available at: <http://cp.literature.agilent.com/litweb/pdf/ads15/esyn/es057.html> [Accessed: 1 March 2016].
- Roberts, M.J., 2010. *Frequency Response Analysis*. Available at: http://web.eecs.utk.edu/~husheng/ECE316_2015_files/Chapter11.pdf [Accessed: 1 March 2016].
- Pro Guitar Tuner, 2012. *Standard guitar tuning – EADGBE*. Available at: <https://www.proguitartuner.com/guitar-tuning/standard-tuning/> [Accessed: 2 April 2016].
- Malchaire, J., 2001. *Sound measuring instruments. Occupational exposure to noise: Evaluation, prevention and control*, pp.125–140.
- Taal, C.H. et al., 2009. *An evaluation of objective quality measures for speech intelligibility prediction, Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp.1947–1950.
- Hong, K., 2016. *Matlab Tutorial: Digital image processing 6 - smoothing: Low pass filter*. Available at: http://www.bogotobogo.com/Matlab/Matlab_Tutorial_Digital_Image_Processing_6_Filter_Smoothing_Low_Pass_special_filter2.php [Accessed: 26 July 2016].
- Fisher, R., Perkins, S., Walker, A. and Wolfart, E., 2003. *Spatial filters - median filter*. Available at: <http://homepages.inf.ed.ac.uk/rbf/HIPR2/median.htm> [Accessed: 28 July 2016].
- Dambrin, D. *Audio editor - noise removal tool*. Available at: https://www.image-line.com/support/FLHelp/html/plugins/editortool_clean.htm [Accessed: 22 June 2016].
- Kate, T., Sehgal, P., Bansal, R. and Jasuja, N., 2016. *Digital zoom vs optical zoom - difference and comparison*. Available at: http://www.diffen.com/difference/Digital_Zoom_vs_Optical_Zoom [Accessed: 15 August 2016].
- Ellis, D., 2002. *A Phase Vocoder in Matlab*. Available at: <http://labrosa.ee.columbia.edu/matlab/pvoc/> [Accessed: 24 August 2016].

Mohamed, S., 2003. *Objective speech quality measures*. Available at: <http://www.irisa.fr/armor/lesmembres/Mohamed/Thesis/node94.html> [Accessed: 25 August 2016].