A PROPOSED FRAMEWORK FOR FORENSIC AUDIO ENHANCEMENT

by

JAMES ZJALIC

BSc, Birmingham City University, 2015

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Master of Science

Recording Arts Program

2017

This thesis for the Master of Science degree by

James Zjalic

has been approved for the

Recording Arts Program

by

Catalin Grigoras, Chair

Jeffrey Smith

Lorne Bregitzer

December 1, 2017

Zjalic, James (MS, Recording Arts Program)

A Proposed Framework For Forensic Audio Enhancement

Thesis directed by Associate Professor Catalin Grigoras

ABSTRACT

Although there are many processes available to the forensic audio analyst seeking to enhance a recording, the order in which they are executed is vital in achieving the optimal enhancement, as each tool is used in isolation and modifies the signal in a unique way based on algorithms applied. The output of a processor in a sequence is determined by the output of the previous, and the order of this chain has a cumulative effect that provides differing quality enhancement results. There are currently various papers available detailing the individual methods of audio enhancement processes, but no document yet exists which provides a context for the interaction of these processes and how the sequence of methods can be ordered in a myriad of ways to produce differing results. Using both scientific reasoning and experimental procedures, this research will propose a framework to produce optimal results when performing audio enhancement for forensic purposes.

The form and content of this abstract are approved. I recommend its publication.

Approved: Dr. Catalin Grigoras

ACKNOWLEDGEMENTS

I would like to thank everybody at the National Centre of Media Forensics for their support, knowledge, and patience during my studies. Specifically, thank you to my professors Dr. Catalin Grigoras and Jeff Smith for the guidance, support, and knowledge. More explicitly, Catalin’s late night seminars provided me with an informal platform to discuss ideas freely, however rudimentary or implausible they may have been. Jeff’s enthusiasm for the subject matter is infectious and he was always highly responsive and helpful with any questions I had, however mundane. Thank you to Leah & Emma for keeping me on track with the clerical side of my studies, allowing me to focus on getting as much out of my Masters as possible. I would also like to thank my cohorts for the laughs, conversations, and support during my time at the NCMF.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

CHAPTER I

INTRODUCTION

The use of audio forensics as a collection of techniques to provide scientific evidence relating to audio recordings has been utilized since the 1960's, and although technology, the types of crime, and the methods of investigation have changed, the fundamental methodology behind forensic audio remains the same:

- Obtain an audio recording pertaining to a crime
- Perform scientific analysis on the recording
- Compile a report based on the analysis

Analysis branches out into several distinct directions, including authentication, enhancement, signal analysis and speaker comparison (Fig. 1.1).

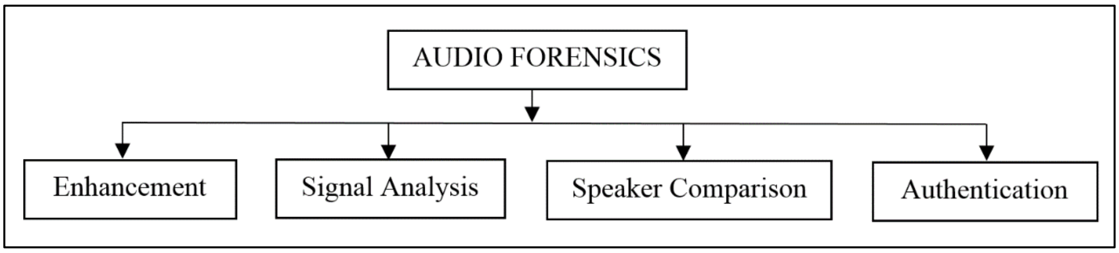


Figure 1.1: Audio forensics disciplines

This thesis will focus solely on the enhancement of forensic audio recordings and seek to propose a framework in the performing of such. The reasoning behind enhancement is generally focused on improving intelligibility and/or quality of speech, although there are also cases in which other signals may be enhanced, such as during gunshot analysis. The enhancement of speech can be problematic due to issues such as the merging of two non-stationary signals (the speech and background noise) of unknown distribution [1], auditory masking of phenomes, and perceptually encoded recordings, which sacrifice quality for file size. Add other speakers into the mix and it is clear why enhancement is an area which requires close attention to the fine details, and as is true of many disciplines, it is in the inches not the yards that differences are made.

As the final destination for forensic audio is to be judged subjectively by a judge, jury or prosecutor, a question must be raised as to how a particular enhancement can be deemed as better-quality than another. Science teaches us to be objective, to mitigate biases and allow for reproducibility, but can judging the result of an enhancement objectively and quantifiably correlate relatively to subjective judgments by humans? Although subjective listening remains the most accurate method of judging enhancements [2], objective algorithms (which predict MOS's (Mean Opinion Score's) provided by listeners), have shown correlations of upwards of 0.94. Enhancements themselves are judged by two factors, those of quality and intelligibility. Quality is a subjective measure which refers to an individual's preference, whereas intelligibility refers to the number of correctly identified words by the listener. It is entirely possible a recording may be both high in quality and intelligibility, but also possible that it may be rated highly in one area and poor in the other. It has been shown that current speech enhancement algorithms do not improve speech intelligibility, mainly due to the difficulty in achieving a good estimate of the background noise present within a recording. Enhancement algorithms are not created to improve speech intelligibility as they utilize a cost function that does not necessarily correlate with speech intelligibility [3].

When performing an enhancement, processing will often be applied to audio in a sequence analogous to a chain. These chains are formed by various processes (or links, to continue with the analogy) which will each enhance the audio in their own way and by varying degrees. For example, a chain may be composed of 3 processes: a filter, de-hum, and de-reverb. The sequence in which a chain is compiled can have a significant bearing on the final result of the enhancement, due to each process using the output of the previous process as an input.

At its most fundamental level, audio is composed of binary digits, and enhancement algorithms process these digits and produce an output based on how the algorithm transforms the input. As with all digital tasks, these processes are generally stable and repeatable (apart from the exceptions such as where random noise is added) and comparisons can, therefore, be drawn when processes are repeated and put into differing orders within a sequence. This is something that was not as easy in the era of analog audio due to the infinite nature of both analog sound and the electrical components used, which were susceptible to all kinds of environmental variables, including temperature, moisture, and equipment degrading.

The key limitation of an audio enhancement framework is that it cannot be followed religiously, step by step. This is because every new input produces a different output, and it can be difficult to predict the output without knowledge of the audio input being introduced. It is therefore of paramount importance that the proposed framework is conceived as a flexible structure in which pieces can be removed, repositioned or even replicated, based on the audio recording for enhancement. It is hoped that by providing information on the fundamentals of the processes, how they function, and when and why they should be applied, the analyst can use their own discretion in deciding how best to apply the framework. Strict rules are put in place where possible as to the placement within the sequence of certain processes of which their positioning is stationary and does not change from enhancement to enhancement.

With regard to the actual practice of enhancement, Koenig states [4]:

> "The "golden rule" of enhancement is that no audio signals are removed or attenuated that decrease speech intelligibility, even slightly. If a recording sounds better overall by the reduction of a particular masking noise, but the understandability is somewhat reduced in the process, the noise is left in or lesser attenuation corrections are tried"

It is with this that the distinction between enhancement for forensic purposes differs greatly from studio enhancement, where the goal is to increase the pleasure of the human listening experience. The products being enhanced are also worlds apart and will be discussed in the early chapters detailing why forensic audio requires enhancement, the possible legalities, and the science behind digital audio signals. From these foundation's the knowledge of individual processes can then be easily understood and proposals made for an audio enhancement framework.

Scope

Forensics are grounded in the fundamentals of science and while it deals in the business of either proving or disproving an individual’s role in a crime, the forensic expert must, in turn, provide “proof” of their scientific methods. The Daubert standard sets out a rule of evidence regarding expert witness testimony in US Federal Courts focusing on the scientific method. To this end, it is crucial that experts are able to validate any methods and procedures used. This principle is based on the landmark Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 US 579 where a precedent was set for expert testimony in federal court cases. It held that the opinion of an expert must be based on recognized scientific principles and not supported speculation. To form an opinion on this, the trial judge must review the following questions:

- If it is a theory, can it be tested?
- Has it been subject to peer review?
- Has it been published in the appropriate specialized literature?
- If it is a test or technique, has an error rate been determined?
- Are there existing standards for the test?
- Is there general acceptance of the theory, test or technique in the relevant scientific community?

This research aims to provide positive answers to the first half of the questions. The second half is answered through the proxy of past research into the various techniques and processes used as the basis for the research. If a forensic expert can cite a single document that utilizes concepts and principles through the use of a literature review and applicable references, it goes a long way to meeting these requirements [5].

As forensic digital audio enhancement is a relatively new field there is as of today no scientific journal article that provides a holistic, practical review of the interaction between digital audio enhancement process from a scientific point of view. There are multiple articles on various elements of audio enhancement, but it would be of an advantage for analysts, government bodies and corporate companies to have the option of referencing a complete work on audio enhancement and all its facets as the foundation for their work, particularly in relation to meeting the Daubert criteria.

Aside from the legal/expert witness aspect, this research would give a point of reference to audio examiners in allowing them to either follow or reference the research when problems arise. A Standard Operating Procedure created from these findings would provide quality assurance and ensure the framework and procedures are followed correctly.

It is hoped that it will provide an initial foundation for students of audio forensics to educate, inspire research ideas, and re-evaluate the proposed framework.

The research will follow best practices, make use of rigorous scientific testing and abide by forensic principles while making the framework as accessible as possible through the use of methods and principles, not particular pieces of software. The software used is irrelevant to this research and although there are certain software suites that may be used during the case studies and/or may even perform a certain task better than others, it is the intention of the author to be agnostic and endorse science, not software.

Another reason behind basing the framework on principles rather than software is future-proofing. Although technology is changing, and new research, practices, and methods are being constantly introduced to all disciplines of media forensics, the basic scientific principles will not change. Technology will always evolve but the fundamental science behind audio capture, digital signal processing, and the human auditory system will remain as it ever has. It can be thought of as a living document which can be updated in the event of a new process being developed by reviewing the scientific logic behind the process in relation to the framework and appending it accordingly.

In summary, the work will allow investigators, researchers, and students an efficient framework to guide enhancements, a single document to reference when problems arise and provide a scientific foundation for all work conducted, whether that be as an expert witness or not.

Audio Enhancement Characteristics

Although some enhancement techniques may be interchangeable between studio and forensic audio, the initial and final products differ immensely. Where studio engineers will often start with the best microphones, rooms, singers, and recording equipment, forensic audio examiners will deal with high noise levels from poor quality microphones and poor recording techniques [6].

Due to the covert nature of forensic audio recordings, small recording devices are used with limited storage capabilities. In an attempt to reach a balance between quality and recording time capability, sacrifices are made, and recordings are often bandwidth limited, compressed and captured at a less than the optimal sample and bit rate. Although digital technology has mitigated many issues with analog audio capture, the laws of nature determine most of the key issues involved. For example, although digital recording systems may have less system noise than analog systems, the fact remains that system noise is the least of an examiner's worries when there is a 2-ton truck driving past two individuals having a telling criminal conversation. This acoustic real-world noise deems any system noise inaudible and irrelevant. Key differences between studio and forensic recordings by Koenig and Lacey [7] are detailed in Table 1.1.

Table 1.1: Recording discipline differences

Characteristic	**Digital Studio**	**Digital Forensics**
Frequency Response	20Hz to 20kHz	20-400Hz to 3-8kHz
Signal to Noise Ratio	90 dB+	Negative to 30 dB
Distortion	Inaudible	1-10%
Equipment Operator	Trained technician	Investigator
Microphone	Large professional	Miniature
Recorder	Professional digital	Inexpensive to professional digital
Medium type	Removable, flash memory, hard drive	Removable, flash memory, hard drive
Noise reduction	Yes or digital	Usually not used
Microphone to speaker distance	Close	Varies
Microphone location	Open	Hidden
Transmission system	Usually none	Often telephone to low-power RF
Reverberation	Usually damped	Often high

Number of persons	Controlled access	In most cases not controlled
Speakers ID	Known	Sometimes unknown
Background voices	Inaudible	Usually no control
Background noise	Inaudible	Usually no control
Background signals (e.g music)	Inaudible	Usually no control
Digital lossy compression	Not used	Used

Digital data files also have the advantage of performing enhancements off-line on a digital working copy rather than risking damage to the original recording. Offline processing allows multiple passes through the data, use of iterative algorithms and the opportunity to evaluate the results subjectively [8].

Enhancement or manipulation?

Although during enhancement the digital audio samples are being changed, it is the intent behind the manipulation, the limitations to each approach and the documentation created that creates a juxtaposition between an enhancement for forensic purposes and a manipulation for nefarious resolves. If an individual has bad intentions, for example, removing a condemning phrase from a recording, they will have only the result in mind, thus not following any specific protocol. Creating documentation as to how the audio has changed would also be low on the list of priorities of a would-be manipulator, as that would only serve as evidence of the act later if discovered. If edits are made, they can be made clear by the placement of a short beep between edits. There are some explicit ways to ensure provenance of the original recording and prevent any accusations of a malicious manipulation. Firstly, always bit-stream the original and perform any enhancement on a working copy. This allows a copy of the unprocessed file to be available and submitted with the enhanced version for reference. Secondly, create written documentation of all processing performed with sufficient details to allow another forensic scientist, competent in the same field of expertise, to identify how the audio has changed and to repeat the procedure. Records during the enhancement should include at a minimum [9] :

- The examinations and analysis that have been carried out, when, in what order, where and by who (this could include screenshots of settings and processes).
- The version of any software tools used in the examination.
- All observations made, photographs taken, and data located.
- All draft final reports or statement generated administrative and technical reviews.

Finally, it is encouraged that the definition of an audio enhancement in SWGDE's Digital and Multimedia Evidence Glossary [10] is reviewed, and any processing of the audio which is not in keeping with the stated purposes in mind should not be carried out.

Audio Enhancement:

> "Processing of recordings for the purpose of increased intelligibility, attenuation of noise, improvement of understanding the recorded material and/or improvement of quality or ease of hearing"

Legalities and Best Practices

The legal guidelines and best practices surrounding audio enhancement often go hand in hand as it is usually the legal issues that could arise during an enhancement that will drive the production and content of best practices. As the destination for audio enhancements is acceptance as admissible evidence to the court, it is essential that processes used are not questionable and can later be refuted by an opposing counsel. To ensure evidence becomes admissible, there are guides for the forensic practitioner to follow, which are grouped within two distinct subsets: those with links, and those without links, to ISO (International Organization of Standardization). Those best practices and standards without links to ISO are often made by recognized scientific bodies such as NIST (The National Institute for Science and Technology), SWGDE (Scientific Working Group for Digital Evidence) and ENFSI (European Network of Forensic Science Institutes) who work with individuals from within the field with practical working knowledge of the processes, science and admissibility policies of the courtroom. There are no ISO's specific to audio forensics as many of the general digital evidence rulings capture the media forensics disciplines under their umbrella.

In terms of ISO, the following is a list of standards which should be reviewed by an individual undertaking enhancement of audio:

- ISO/IEC 27037:2012 ‘Guidelines for identification, collection, acquisition, and preservation of digital evidence’ [11]
- ISO 17025:2005 ‘General requirements for the competence of testing and calibration laboratories’ [12]
- ISO 9001:2008 ‘Quality management systems – requirements’ [13]
- ISO/IEC 27042:2015 ‘Guidelines for the analysis and interpretation of digital evidence’ [14]

And in terms of best practices:

- SWGDE Digital and Multimedia Evidence Glossary Version 3.0 [10]
- SWGDE ‘Best Practices for Forensic Audio’ Version 2.2 [15]
- SWGDE Core Competencies for Digital Audio Version 1.0 [16]
- ENFSI Guidelines for Best Practice in the Forensic Examination of Digital Technology Version 6.0 [9]

Standard Operating Procedures

Although these guidelines have some practical aspects, they are mostly theoretical. The practical aspect which stems from these guidelines arrive in the form of SOP’s (Standard Operating Procedures), which provide steps for forensic procedures to be performing correctly. SOP’s relating to audio enhancement will include receiving and processing of evidence, the enhancement itself and the deliverables prepared for the client. It is advised that these be created by referring to the best practice documentation available before any audio enhancement takes place, and followed rigorously during any enhancement. By doing so, accountability is created, allowing another expert of the same skill level with the same equipment to achieve equivalent results. It is the intention of this research to contribute to the pool of knowledge and assist in providing a reference document from which to create robust audio enhancement SOP’s.

CHAPTER II

SOUND FUNDAMENTALS

Auditory Systems

Although some elements of audio enhancement can be judged through computed quantitative results, the audience for the end-product enhanced audio recording is always a human. During enhancement, the audience is the examiner, who is constantly working to improve the recording through judgments made via critical listening (although other, visual cues such as FFT and waveforms are taken advantage of). Eventually, the audience becomes the transcribers, judges, juries, and lawyers. The determination of the problem, the result of enhancement processes and the final product are therefore all judged by a human in their effectiveness, so understanding auditory perception will thus be a necessity in gaining the most value from each individual process as well as the final creation. Only those areas which apply to forensic audio enhancement will be discussed.

Equal loudness contours

As the human auditory system has evolved over hundreds of thousands of years of evolution to be advantageous to us in adverse situations, it has peaks in areas which are most important to our survival. To serve as an example, the ear resonates at around 3 – 4 kHz, increasing sensitivity in that range, which can give increased intelligibility to speech. The obvious advantages to this are in alerting us as to when babies are distressed, or to when others are attempting to communicate with us. The sensitivity of this organ has been examined, tested and finally plotted in such graphs as the Equal-loudness contours. These curves show the sensitivity of the ear at different frequencies across the audible range and are based on experiments by Fletcher and Munson who derived their plots from tests of subjects asked to adjust the level of test tones until they appeared as loud as a reference tone with a frequency of 1Khz [17]. They were updated in 2003 based on new cross-correlated findings from research performed by various countries around the world (Fig. 2.1) and have become standardized as ISO 226:2003 [18].

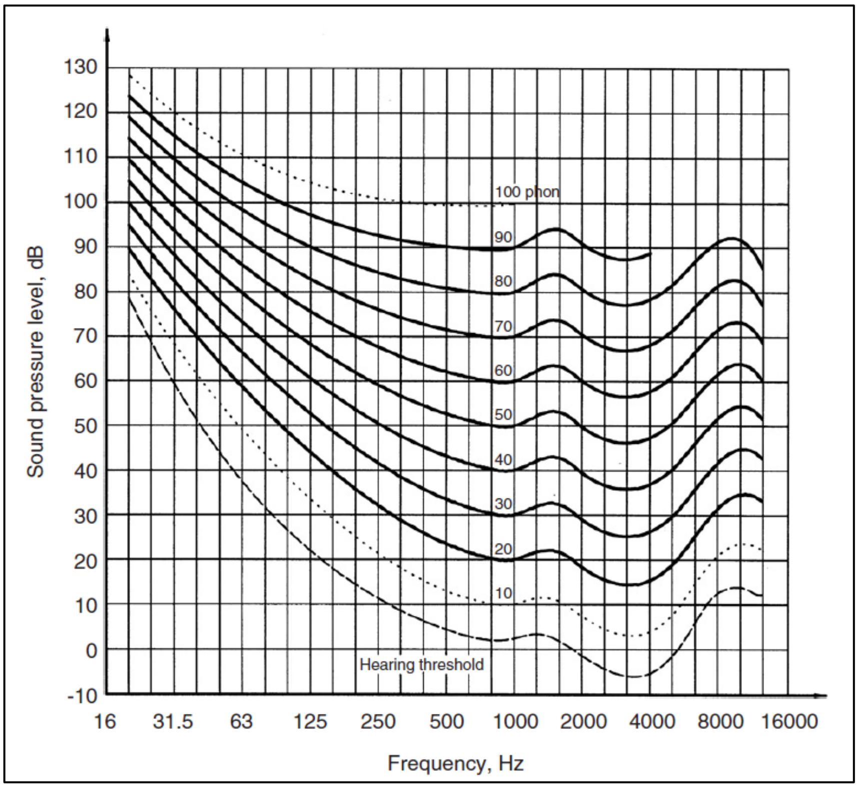


Figure 2.1: ISO 226:2003

In a practical sense and for the forensic examiner, the equal loudness contours can have many implications as to how processes should be applied when dealing with enhancements. For instance, although frequencies within the 2000Hz – 4000Hz range may appear to be of a lower amplitude than surrounding areas when performing visual analysis, considerations should be made to the amount of gain applied, knowing that the ear is extremely sensitive within that area. It should also be considered that although speech is audible and above a constant noise level, sounds within areas which are less sensitive to the human hearing system (but may be at the same amplitude as the speech), can go unheard due to the lack of ear sensitivity in those areas.

Critical bandwidth

Discrimination of sounds within the auditory system is dependent on the frequency being heard. At lower frequencies, tones that are only a few hertz apart can be discriminated but as the frequency increases, tones must differ by an order of hundreds of hertz to be differentiated. The hair cells within the ear respond to the strongest stimulation within their locality, or “critical bandwidth”. Experiments by Fletcher concluded that when a noise masks a pure tone, it is only frequencies within that region that will mask the intelligibility of that tone, this is known as intra-band masking. When shown in plots, the critical bandwidth at frequencies below 500 Hz are linear, and above this are logarithmic. Critical bands are vital when performing enhancements as they show that frequencies outside of a critical band are inconsequential when attempting to provide clarity within a particular band of frequencies [19].

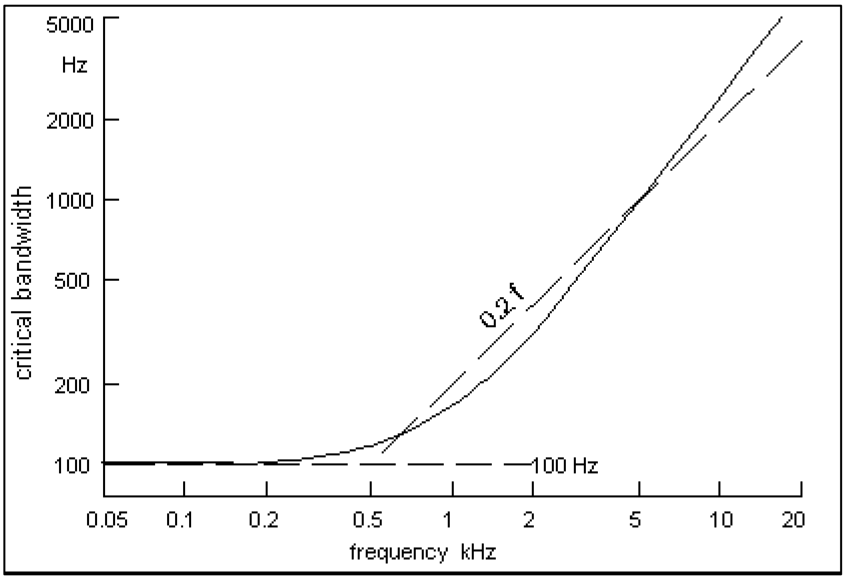


Figure 2.2: Critical masking curve

As can be seen in Fig. 2.2 [20], the band rate is linear from 0 – 500Hz and increases exponentially above that. The critical bandwidth can be estimated through the equation:

$$\text{Critical bandwidth} = 25 + 75[1 + 1.4(f/1000^{2}]^{0.69}\ \text{Hz}$$

Where f is the center frequency in Hz.

When performing audio enhancement, it is crucial to understand that at certain frequencies, a sound that is at a relatively large distance from the desired signal may be having a masking effect. The degree to which this is occurring can be determined by combining FFT analysis with critical masking/critical bandwidth plots, and informed decisions can then be made regarding the action that should be taken.

Masking

Masking is closely related to critical bandwidth and plays a significant role in the perception of sound by humans, in both the frequency and time domain. It can be perceived in two ways: simultaneous masking and non-simultaneous/temporal masking.

Simultaneous masking occurs when a noise masks a sound by occurring in the same temporal frame, for example, a truck driving past when a crucial word is spoken, masking the utterance. For the utterance to be masked, the noise of the truck must be at a certain level in relation to the sound, as shown in Fig. 2.3.

Non-simultaneous masking occurs when two sounds occur within a small-time interval and can occur as pre- or post-masking. Pre-masking is when the desired signal occurs after the masking sound in the time domain. Post-masking is if a signal precedes the masking sound in the time domain (Fig. 2.3 [20]). As humans take cues from both the attack and decay of a signal, this type of masking can cause non-intelligibility of the desired signal or even cause the brain to fill in absent material, potentially changing the perceived content [21].

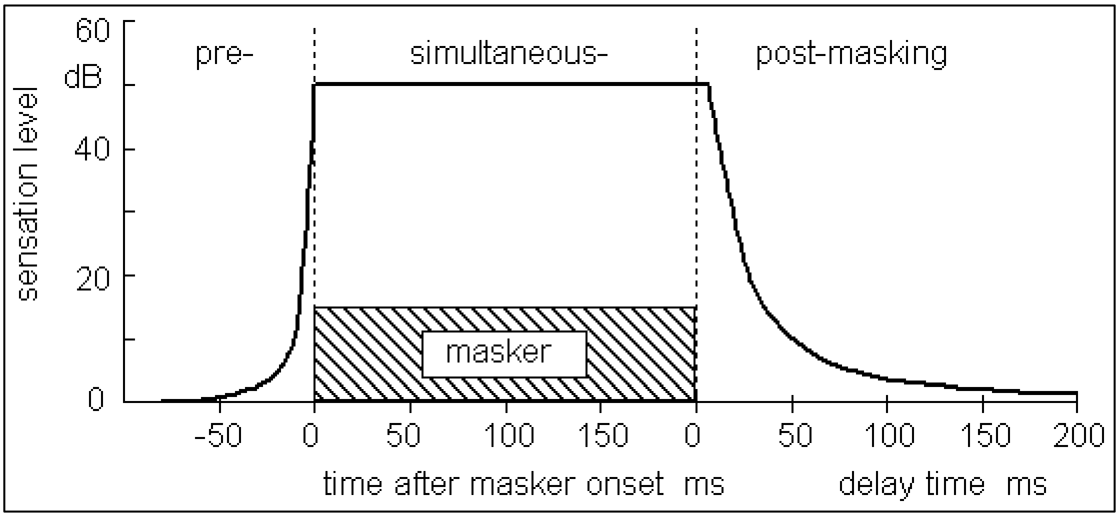


Figure 2.3: Masking thresholds

In the context of audio enhancement, as with critical bandwidth masking, decisions can be guided through the use of waveforms and information such as the masking thresholds plot.

By understanding the auditory system and how the loudness contours, critical bands, and masking affect each other allows us to recognize the type of masking when it occurs and act accordingly. It also offers the examiner an understanding of the limitations of enhancement, allowing explanations to be provided for clients and laymen alike. In relation to the framework, the masking effects on a desired signal can help decide in which order certain processes should occur.

DSP Basics

Some of the limitations of enhancement are due to limitations of digital recordings, for example, compressed audio recordings via perceptual encoding not capturing enough spectral information due or a poor SNR (Signal to Noise Ratio) courtesy of a low bit-rate. It is therefore vital to understand the foundations of digital signal processing to know the advantages and limitations before processes available to the audio examiner are detailed.

Sampling

For a digital system to capture an analog signal, it must first take samples of the continuous analog signal at discrete points in time (Fig. 2.4). The most common rates in forensic audio are 8kHz (telephone) and 44.1 kHz (CD). As the incoming signal is sampled at precise intervals, the signal is “held” while the converter stores the value (represented as a binary encoded word) with an accuracy dependent on the circuitry of the system. This prevents inaccuracies due to the ever-changing nature of a continuous sound. Without the “sample and hold” mechanism, the converter would begin storing a value but before it has completed the task the value would have changed, causing inaccuracies and distortion.

The Nyquist theorem states that to accurately represent a signal the sampling frequency must be twice the highest frequency being sampled by the system. If frequencies above this enter the chain then distortion occurs in the form of “aliasing”, which are ghost signals created by the folding back of signals in the recording into the audible range. To prevent aliasing, a low-pass filter is applied before the ADC stage. As filters require a roll-off period to be effective, sampling will often take place at a slightly higher frequency then double the highest frequency to allow for this roll-off period. For example, a sample rate of 44.1 kHz is used to effectively encode a bandwidth up to 20kHz [22].

Quantization

At each discrete sample point, the signal must be given a discrete level in relation to the amplitude, which is done through the process of quantization. For good quality audio, 16 bits are a good starting point (providing 65,536 graduations). Greater word-lengths lead to greater resolution due to the increased number of steps available for the signal to be digitally encoded [23]. The systems SNR also increases with the bit-depth, and once the recording has been captured, an increase in the bit-depth will not increase the SNR as the noise will scale upwards with the rest of the signal, and more noise will in-fact be introduced due to the re-quantization of the signal (which introduces further quantization error). It will, however, increase the resolution of further processing which can be of an advantage, dependent on the level of processing taking place.

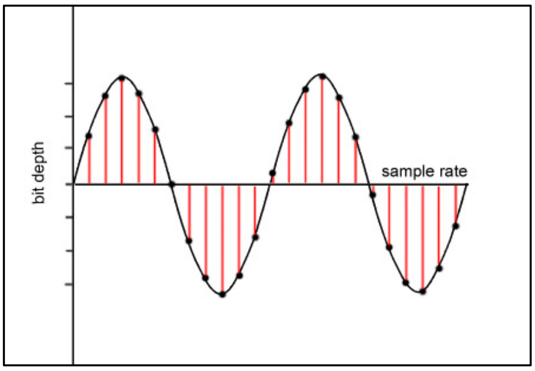


Figure 2.4: Sampling/Quantization plot [24]

Quantization error

Due to the discrete steps allocated for sampling, any value falling between these steps will be rounded either up or down, dependent on their value. For example, in a 1-bit system (which has 2 available values of 0 and 1), an input voltage of 0.8 would be rounded to 1. In a worst-case scenario, an input of 0.5 would be rounded to 0 or 1, losing ½ a quantization level worth of accuracy. This is known as quantization error and is equivalent to ½ the value of the least significant bit or LSB. This error can create quantization noise, which requires Dither to reduce. In terms of processing, every calculation performed increases the quantization noise, so care should be taken to only apply processes that are necessary and are increasing the clarity of the desired signal more than the quantization noise is decreasing it. If this is not monitored the final product could have more noise than the original.

Dither

As the quantization steps are discrete and can be very small, any values falling between two steps will be rounded to the closest step. Due to this phenomenon, quantization noise (which is audible distortion due to a long sequence of matching quantization levels at the LSB). It is audible due to the harmonic pattern which is distinguishable by the human auditory system. By adding a small amount of random noise (or “dither”), a probability curve will be added which allows the ADC to detect whether the signal is closer to the least significant 1 or 0, randomizing the values of the LSB, rendering them less perceptible to the human ear than the sequence of matching values.

SNR

In a general sense, signal to noise ratio or SNR refers to the number of dB between the reference level of a signal and the noise floor, but in a forensic sense, a more accurate description must be sought. The SNR can be stated as being the base-10 logarithm of the ratio between a wanted signal *s* and interfering noise vector *n*, measured in decibels.

$$SNR = 20\ log_{10}\left\{\frac{1}{N}\sum_{i=0}^{N-1}\left(\frac{s}{n}\right)\right\}$$

Segmental SNR relates to the measurement when analyzed against segments of speech, typically of 20-30ms in size, with some overlap. Although only minimally relevant due to hearing being in the frequency rather than time domain, they can still provide an examiner with a starting point from which to begin enhancement. A more accurate measurement would be to use spectral distortion, which makes use of the frequency domain through conversion of a section of audio using FFT and obtaining the power spectra for the required signal and interfering noise. It can then be enhanced further through the use of weightings so that the more audible frequencies have higher weightings [25].

$$SEGSNR(j) = 20\ log_{10} \left\{ \frac{1}{N} \sum_{i=jN}^{(j+1)N-1} \left(\frac{s}{n} \right) \right\}$$

SNR can be an important measure when attempting to determine the distortion present within a recording, but due to the nature of forensic recording is can often be difficult to calculate, as the desired signal and noise can often become indistinguishable. It also has little bearing on how well a signal has been enhanced for the same reasons, and the classic SNR measurement performs extremely poorly in predicting the perceived quality of an enhancement by a human listener.

Perceptual Encoding

Perceptual encoding is a form of lossy compression which uses a psychoacoustic model of the human auditory system to identify imperceptible content and code the incoming audio based on this phenomenon. The process is highly efficient in reducing the quantity of data required by removing both irrelevancy (imperceptible content) and redundancy (information that is repeated and unnecessary). Although this technique increases quantization noise, the previously discussed techniques regarding “masking” and critical bandwidth are implemented to hide it. Rather than a linear distribution of bits across the quantization range, a variable-bit-rate is used to selectively decrease the word-length based on the conditions, so that a high bit rate is used when the signal is information rich and a low bit rate when the signal is information poor. Also, consider that the areas of low signal information will be masked by those which contain more information, so even though a 2 bit/sample rate (12dB SNR) may be used for these low priority areas, the quantization noise associated with it will be inaudible.

Perceptual encoding can be an issue for forensic analysts and compressed recordings would never be recommended due to the amount of real-world information that is discarded during encoding. Before enhancement takes place is it, therefore, it is vital to convert any lossy encoded evidence into a lossless format. Although the information that was not captured due to the perceptual coding can never be recovered, conversion provides linear quantization levels and optimizes any enhancement processors applied thereafter.

Recording Limitations

The absolute nature of digital recordings means that the possibilities of enhancements are restricted. These limitations must be understood by the forensic audio analyst as to inform judges, juries, clients and other laymen that much of what is presented on television shows and within movies is not possible in the real world (the so-called CSI effect).

Firstly, if the digital system which has captured the recording hasn't captured the desired element of a signal (due to distance from source, bandwidth limitations, perceptual encoding etc), then it is impossible to recover. Unlike deleted files on a computer which can be recovered as they were, at some point, present on the system, if the recording process didn't capture an event it cannot be recovered as there is nothing to recover.

Secondly, if an event has been recorded but is completely inaudible (due to masking, reverberation etc), then the chances of recovery are very small. As the sampling process samples the incoming signal as a whole, the audio recorded is convoluted and although there are techniques that can de-convolute a signal based on estimations of noise against the desired signal, if the desired signal is inaudible then this indicates it is below the noise floor, a practical example of critical masking. Again, as in the first point, if the information was not captured, this time due to masking, it cannot be recovered.

As with many areas of life, prevention is always better than a cure. The capture of audio should be treated as the most important part of the forensic audio chain, as being the first event, it has the largest impact on the final result. Every decision after the recording has been made is based on the quality of the original recording. There can be an attitude of "fix it in the mix", where the technician capturing the recording believes any problems can be fixed later on, but this attitude is comparable to giving a chef poor ingredient's and asking him to make a great soup. The chef can only make the best with what he is given, as can the forensic audio examiner. To counter-act this "fix it in the mix" dogma, training should be made available to those in the field who are on the front-line and who create the recordings, with an emphasis on the correct settings for the recording device, optimal microphone placement and considered location choice. Some simple recommendations include:

Device Settings

- Use an uncompressed format such as WAV PCM
- Use as high a sample-rate and frequency as possible (minimum 44.1 kHz, 16 bit)

Device Setup

- Perform test recordings prior to event to optimize gain
- Use an external microphone
- If using 2 microphones or more, place the microphones at least 20cm apart
- Direct the microphone diaphragm towards the target

Recording Location

- Choose a location with low noise levels
- Choose a location with as little dynamic noise as possible
- AVOID locations with competing talkers where possible

If this is implemented and advice followed, it will offer more potential to the audio enhancement process, which in turn will provide improved results, which will then result in more useable forensic audio evidence within courtrooms.

Finally, desired events that are audible may be enhanced to improve the intelligibility of words for the listener, but in doing so this could, in fact, degrade the actual listening quality. A prime example of this is leaving, or even boosting high frequencies that can make recordings sound noisy or hissy as they contain voice information critical to intelligibility [26]. There is a distinction between quality and intelligibility and it is the forensic examiner's responsibility to improve intelligibility over quality. Intelligibility refers to the number of words which can be transcribed from a recording, whereas quality is a subjective measure as to the pleasure in listening to a recording. Sometimes it is possible to achieve both and/or sometimes improving quality naturally leads to an improvement in intelligibility, but in the cases that it doesn’t this is a crucial point to remember when performing enhancements and when providing clients with the enhanced recording.

It should also be made clear that some recordings are just impossible to enhance, and the practiced audio examiner should recognize this before attempting an enhancement which will leave the client disappointed.

CHAPTER III

SETTING THE STAGE

Evidence Preparation

The retrieval and processing of evidence are outside the scope of this paper, so it will be assumed that this has occurred, following digital evidence best practices and guidelines. It will also be assumed that the evidence is in a digital format on the examiner's workstation.

In preparing the evidence several factors must be considered to preserve the evidence, maintain a chain of custody, and optimise the enhancement processes. The first step is to make a digital bit-stream copy of the digital original which is then used as the working copy. Hash sums should then be calculated for both copies to ensure an exact match, thus preserving the integrity of the original recording. The details of the submitted audio should then be recorded within a document, including sample-rate, bit depth, and length. This ensures the appropriate computer software applications and settings are used to prevent any artifacts that may appear from improper playback combinations which could cause resampling of the audio [15]. If the audio being submitted is perceptually encoded, transcoding will be required to convert to an uncompressed format such as PCM to allow a wider range for processed samples to inhabit, improving the resolution. The original sample rate should be maintained, and bit depth maximised to allow the audio to be processed and clarify as much content as possible. Down-sampling without anti-aliasing should never be practiced as this can cause audible artifacts which are not part of the original signal [27].

Before performing any analysis, it can also be helpful to detail as much about the recording as possible and use this information to aid analysis and processing as a guide to conditions which must be mitigated or processed [28]. Some of this information may be provided by the individual who made the recording and can include but is not limited to [29] :

- Power source, e.g. batteries, AC
- Input, e.g. telephone, microphone

- Environment(s) e.g. restaurant, phone transmission
- Background noises, e.g. telephone, birds, conversations
- Foreground information, e.g. number of people conversing, gender of people conversing
- Recorder operations, e.g. if device was switched on and off

Listening Environment Optimization

Before critical listening and the following enhancement can begin, it must be ensured that the environment in which it is taking place is optimized for the procedure and that the system can perform the required enhancement processes without introducing distortions in the form of digital error. Recordings are often exposed to elevated levels of background noise and low SNR, so adding to these factors by not listening in the correct conditions is doing a disservice to the recording. Experiments by Bergfeld et al [30] show that audio hardware, compression, maximum output level and peripheral stimuli all have an adverse impact on the speech reception threshold. Recommendations include a setup which takes advantage of the following [31] :

- An external audio card;
- Specially equipped room with no distractions and good environmental hygiene;
- Quiet working environment (< 25dBA SPL);
- Computer-based playback system with low noise A/D and D/A;
- Audio editing, spectral analysis, and display software;
- Reliable, spectrally flat headphones with a built-in limiter (this is to prevent damage to headphones or the ears on recordings with a large dynamic range);

More advanced options, as recommended by the FBI in 2003 [32] include:

- Digital-adaptive enhancement processor to allow implementation of different filter algorithms including band-pass, adaptive and notch. All filtering systems should contain 16 bit or greater A/D and D/A, two input channels and sampling range extending to at least 32kHz;

- A separate compressor/limiter with minimum of 2 channels, adjustable, automatics compression ratios, attack-release times and gain reduction of at least 40dB
- A spectrum analyzer that includes an FFT with single channel capability. This should have a minimum of 16-bit resolution and frequency ranges adjustable from at least 0-100Hz up to 0-20kHz. It should also have a minimum of 800 lines of resolution in any frequency range.
- Non-real-time software programs for precise, nonlinear time and amplitude processing of audio recordings.

Initial Analysis

To allow both the examiner and listeners of the final product to understand exactly how the audio has been processed, it is vital to create an accurate account of all critical listening notes and processing activities that take place. If called to testify, having the ability to produce replicable results and describe/demonstrate the changes made to the signal is especially important in showing scientific methods are understood and applied.

Before any processes are to take place, it must first be determined as to what the underlying problems of the recording are and the aims of the enhancement, analogous to reviewing a map and plotting a journey before commencing a trip. An auditory assessment/critical listening review combined with FFT analysis is undertaken from which to compose a strategy regarding enhancement [33].

System Calibration

Before any enhancement procedure takes place, it is advisable to perform tests on controlled recordings to ensure the system is working correctly and as expected. These tests do not need to be performed for every enhancement, but for each tool, and then detailed in the organization's QA (Quality Assurance) documentation under the section relating to validation and verification of tools. Tests should be completed multiple times to ensure the system is working correctly and suggestions from the researcher include:

- Performing an overall gain increase on a sine wave using the enhancement software and objectively measuring the gain before and after.
- Addition of low-level white noise to the sine wave and perform noise reduction with identical settings. Take SNR measurements of the outputs for comparison.
- Apply a low-pass filter to white noise and measure results through FFT analysis.

CHAPTER IV

HYPOTHESIS

To provide a focus and create a proposed framework that is both scientifically sound and forensically practical it is reasoned a hypothesis must be created. To construct a hypothesis various factors must be considered:

- Are variables quantifiably measurable?
- Can variables be controlled?
- Do we have any knowns from previous research on which to construct a hypothesis?

The following premises can be considered as fact based on previous research:

- Processing an audio input affects the audio output;
- Algorithms react based on the audio input;
- Each processing stage will modify original information from the signal;
- Processing algorithms work as a standalone structure and are unaffected by the actual processing which occurred before or after, only by the audio product of said processing.

It is, therefore, a strong inference that:

- Cumulative processing of a recording will affect the final output.

And it can be then inferred from this inference and the fact that algorithms work as a standalone structure that:

- The output of one filter will affect the output of the next filter in the chain;
- The processing chain has a cumulative effect that can provide differing final outputs.

A reasonable hypothesis is, therefore:

> *"There is an ideal sequence of operations to produce the optimal results from audio enhancement for forensic purposes."*

This hypothesis alone cannot be tested without definitions put in place. The final aim has already been touched upon, but another definition from SWGDE [10] defines forensic audio enhancement as:

> “Processing of recordings for the purpose of increased intelligibility, attenuation of noise, improvement of understanding the recorded material and/or improvement of quality or ease of hearing.”

To produce optimal forensic results the stated purposes of enhancement must be focused on. Once the purpose has been defined, a scientific review of all techniques must be performed before a logical framework can be devised, based on the knowledge of how various techniques alter the input of a processor at a sample based level.

CHAPTER V

PRE-PROCESSING

Critical Listening

Once the environment has been prepared, critical listening can begin, which involves noting the issues that are audible when listening to the evidence. Recordings can contain many elements which are responsible for the inadequate quality and/or poor intelligibility of a recording and these factors are grouped into either noise (events captured when the recording was made) and distortions (changes to the content due to the transmission or recording system). The characteristics and causations of these features are documented below.

Noise

- *Convolutional changes:* a result of linear frequency alterations due to the recording system, microphone, transmission channel or acoustic environment.
- *Environmental noise:* any noise contributed by the environment in which the recording took place.
- *Large amplitude differences between talkers:* caused by individuals proximity to the microphone being of different distances and/or angles.

Distortion

- *Nonlinear distortion:* results from clipping of the signal, causing a square-wave effect and is seen by the production of odd and even harmonics within the frequency domain. Causes of these issues can be an input signal outside of the range of the system, overdriven electronic components, poor quality receiver/transmitters and component failures.
- *System noise:* any noise due to the system, microphone or transmission system
- *Signal loss:* Complete or partial loss of signal due to electronics failures, out of range transmitters and mobile phone dropouts.

- *Aliasing noise:* sound produced by the fold-back effect, creating mirror images of a tone reflected into the high frequencies due to disregard of the Nyquist law.
- *Transmission interference:* Noise resulting from effects of the transmission device.
- *Digital distortion and noise:* due to effects such as dither, lossy compression, and low bit rate encoding.

Matching the various problems that arise with forensic recordings and the processes available to the forensic examiner serves as a clear reference point for individuals undertaking enhancement. A table proposed by the NCMF (National Center for Media Forensics) [34] is used as reference (Table. 3.1).

Table 3.1: Common problems and solutions

Category	Problem	Solution
Distortion	Clicks/Crackle	De-click / De-pop
Distortion	Clipping	De-clip
Distortion	Mobile Phone Burst	Cell Phone (Noise) Filter
Distortion	Reverberates	De-reverberate
Periodic Noises	Tones	Notch Filter, Spectral Subtraction
Periodic Noises	Sirens	Spectral Editing
Non-Periodic Noise	Hiss	De-hiss / Adaptive Filters
Non-Periodic Noise	Broadband noise	Adaptive Filters
Non-Periodic Noise	Coughs, steps, pedals etc	Spectral Repair / Adaptive Filters
Source separation	Background music	Adaptive Filters with Reference Channel, Dynamic Spectral Subtraction (DSS)

Fast Fourier Transform

FFT (Fast Fourier Transform) processes the audio to display the information as a function of the frequency domain through a mathematical relationship based on the periodicity of sine and cosine functions, allowing a continuous, periodic signal to be represented as a sum of these functions and an imagery component (Fig. 3.1, 3.2). There are several types of plots which take advantage of FFT. Spectrograms display the audio as a function of time, frequency, and amplitude. The time is plotted on the x-axis, frequency on the y-axis, and the amplitude represented by the colors of the data. LTAS (Long Term Average Spectrum) provides an average across a number of bins, resulting in a stable and smooth plot in which the amplitude is represented as a function of frequency. The same data can often be found in both types of plots, but certain types may be easier represented by one type of plot than another. When using FFT, it is advisable to use high order for frequency measurements and a low order when taking time measurements.

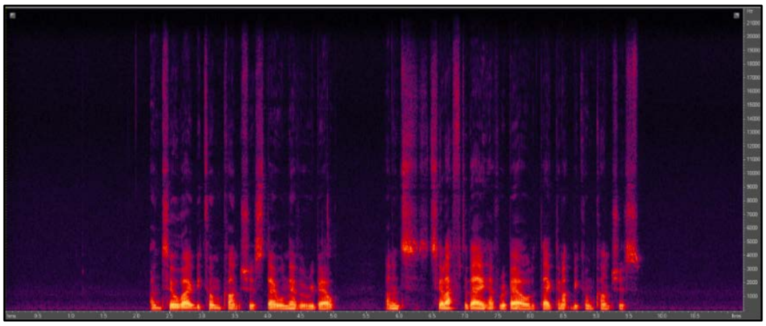


Figure 3.1: FFT spectrogram

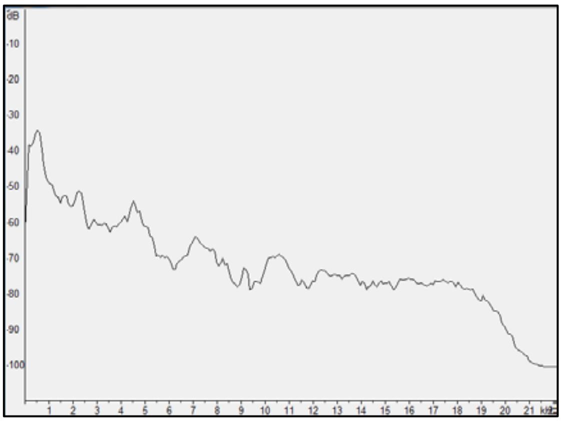


Figure 3.2: FFT Long Term Average Spectrum

An overall review of the audio using FFT analysis [7] related to the following pieces of information should be appended to the notes taken during critical listening.

- *The range of speech:* The lowest and highest frequencies in which speech is present are documented, resulting in a clear bandwidth which contains speech to ensure the enhancement process can be focused;
- *Speech to noise ratio:* A ratio regarding the level of speech information in relation to noise within the recording;

- *Discrete Tones:* Frequency, amplitude, and stability of all high-level tones;
- *Banded noise:* Bandwidth, amplitude, and stability of wide noise bands;
- *Convolutional effects:* Determination of whether speech is consistent with known LTAS information through comparison with an exemplar speech recording with no noise.

CHAPTER VI

REVIEW OF AUDIO ENHANCEMENTS

De-Click

Clicks and pops in audio are generally defined as undesired audible transients [35]. They are perceived by the listener in many ways, but in the realm of digital audio they are often heard as tiny tick sounds (Fig. 6.1) and are caused by poorly concealed digital errors and timing problems within the ADC (Analogue to Digital Converter) of the recording device [36].

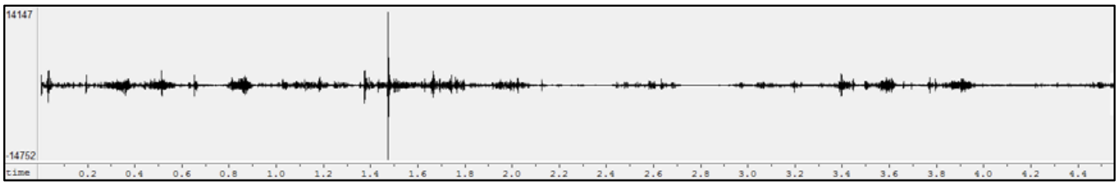


Figure 6.1: Audio click present at 1.5s

These digital clicks and pops can not only be distracting but also cause masking of phenomes, which are the elements of speech which determine the meaning of sound. For example, the C or B in “Cat” or “Bat”. Although interpretation is an unavoidable aspect of forensics, limiting the use of it should also be pursued where possible, and unmasking phenomes is a vital step in providing increased intelligibility to a word rather than an individual using subjective interpretation as to what the word may be.

There are various techniques used to remove clicks. Sample and Hold works by sampling the signal before the click and holding until afterward, assuming a plateau will be closer to the valid signal than a click. It detects the clicks through a high pass filter which detects transients above a certain amplitude threshold. This does work to an extent and removes the clicks, but leaves audible bumps and pops in the waveform.

Another method used is linear interpolation, correcting the clicks through the use of two good samples instead of one, and drawing a straight line between the sample before the click and the sample after the click. The audible result is less offensive than the sample and hold technique but causes low-frequency artifacts and reduction in bandwidth over the area. Dither of the interpolated area is applied by the algorithm to reduce quantization error of the corrected region.

One of the more recent methods is to reconstruct the area in which the click is present through an analysis of the areas before and after the click through a method coined “Spectral Repair” (Fig. 6.2, 6.3). At a high order of interpolation, this method is undetectable to the human ear [37] but runs the risk of changing phenomes, leaving potentially erroneous transcriptions.

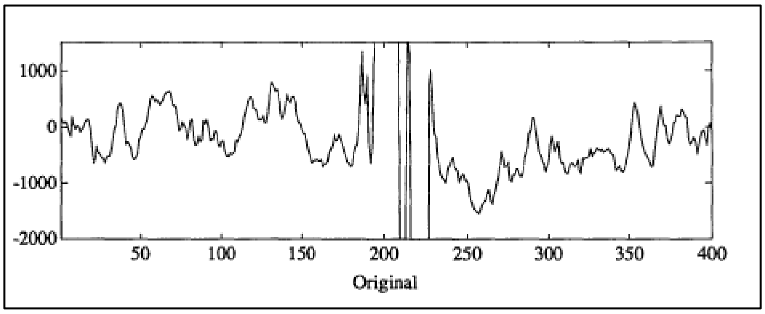


Figure 6.2: Audio click pre-repair [49]

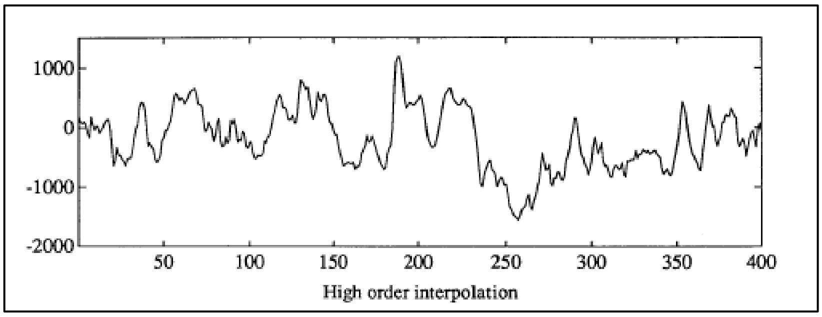


Figure 6.3: Audio click post-repair [49]

De-Clip

Clipping is caused when there are no more quantization levels available for the ADC to store a higher amplitude signal, due to limitations in the systems dynamic range. For example, if the maximum quantization level is 65,536 in an unsigned 16-bit system, any signals that are above this upper limit will be represented as 65,536 as there are no more levels available [17]. This causes a square wave and a loss of information which cannot be retrieved (Fig. 6.4, 6.5).

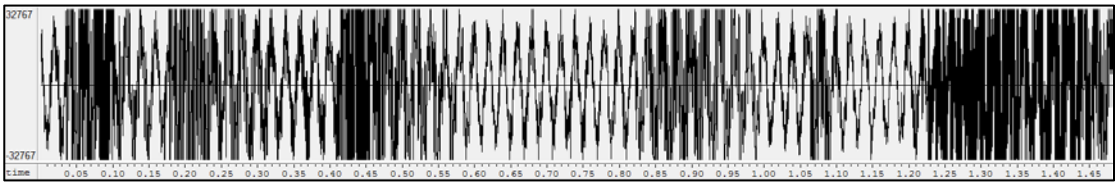


Figure 6.4: Audio clipping example

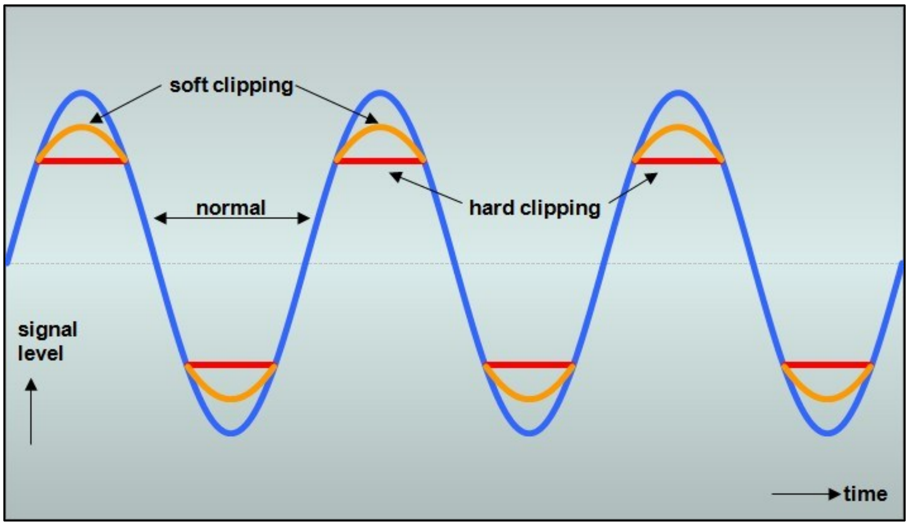


Figure 6.5: Macro-level audio clipping [38]

Common reasons for clipping are setting the record input level too high, an unexpected loud signal or using a bit depth that is too low to accommodate the type of recording. It is difficult to prepare for such events that are unexpectedly loud such as gunshots, but with proper training, clipping can be prevented in all but the most unpredictable situations. Care should also be taken when performing any enhancement as if there is not enough headroom available, a process which increases the signal level could cause clipping. Even processes such as attenuation of a certain frequency range with a filter can cause boosts in other regions as filters can ‘ring’ and change peak levels if the balance is skewed [39].

The reason for clipping removal is square waves created can cause distortion which reduces fidelity and impairs audio quality. Processing that takes place with the clipping present may produce new distortions, as an input of a sequence of repeated samples may produce an output sequence of repeated processed results if the algorithm is not feedback based.

Clipping is removed in 3 stages. First, the areas which are clipped must be identified. This is accomplished through statistical analysis. Some methods compare the waveform against the peak level where sections closest to the peak level are more likely to be damaged than those closer to the center of the waveform. Other algorithms define clipping as an area of 3 consecutive samples at the maximum quantization level, although this is very conservative as clipping under 2 milliseconds is likely inaudible. Next, the frequencies present before and after the clipped area are analyzed and the section bridged by the recreation of the waveform accordingly. Finally, the audio is divided into bands and analyzed through the peak-to-RMS average ratio for each band. These dynamic detectors set the ratio of multiband upward expanders (which cause increased expansion when fewer dynamics are present) while the threshold is kept at a fixed offset from the peak level of the original audio in each band. This "decompression" prevents peak-limiting and compression in the densest areas of the recording [40].

There has been debate over whether removing clipping when performing enhancement is scientifically ethical, as the process to remove clipping estimates the signal, so is not a true representation of the original. Proponents of clipping removal claim it creates a closer representation of the original signal than the signal with clipping present. Following research by Koenig & Lacey [41], it was found that clipping produced few improvements and introduced noise that wasn't present in the original. The caveat is that these tests were performed in isolation, not within a framework in which future processes are to take place. It is recommended that tests are performed using the de-clipping software before any processing takes place to determine the amount of distortion created and understand how the process is affecting the signal. De-clipping should never be used for an enhancement which is to be used for speaker recognition tasks as areas which are lost are being interpolated and may be misinterpreted during the comparison, leading to inaccurate results.

Spectral Repair

Various companies now offer software that features “spectral repair”, which allows the time-frequency domain precision removal of areas of the spectrum [42]. As much of the software is proprietary, the sample level processing that occurs is undocumented, but it can be logically deduced that the unwanted area is removed and then resynthesized using the surrounding areas, as is implemented when performing de-clipping. Spectral Repair can improve the quality of recordings by removing such events as birds chirping, dogs barking by interpolating samples from either side of the distracting area.

With regards to forensic audio enhancement, there are several issues arising from the use of spectral repair. Firstly, prominent artifacts can be introduced due to the unnatural re-synthesization of samples, which increase proportionally with the length of the area being repaired. The worst artifacts are introduced when attempting to repair audio which is flanked by other areas prominent areas, as it is through analyzing and synthesized those that the new samples will be created. Finally, it is a “black box” process of which little is known as to how the audio is being edited, although it is certainly similar to both de-click and de-clipping. When performing spectral repair, as with the previously mentioned processes, the area to be processed should be accurately selected, ensuring minimal introduction of artifacts and minimal change of the audio.

Stereo Source Separation

Due to the nature of forensic recordings and the uncontrollable aspects of the number of voices being recorded, background voices and background noise, the final product can often be a convolution of sounds. This is often referred to as the “cocktail party effect” and in a live situation, the human auditory system has various mechanisms to deal with isolating the signal pertinent to them through such phenomena as interaural level differences and interaural time differences. These are the acute differences in sounds reaching each ear respective of level and time delay. In a forensic situation, the recording is dealt with after the fact and due to the nature of a digital recording being a linear, 2-dimensional representation of a 3-dimensional world, much of the data regarding delays and reflections is lost through convolution.

One of the key areas of research regarding this is BSS (Blind Source Separation) which seeks to estimate the sources of the original signals using only the information from the mixed signal. As the signal is convoluted by nature, this involves the transformation of the signal to the time-frequency representation through windowed transform, where separation is carried out at each frequency bin [43]. This can be extremely computationally expensive as source separation must be carried out on such a large number of bins and methods have been proposed to only process the frequencies which lie within the speech domain (up to around 4kHz). The role of high frequencies with regard to speech intelligibility in blind source separation was investigated and shown that the quality of separation does not deteriorate significantly [44].

The performance of BSS can be highly limited in situations in which there is excess reverberation due to differences in the frame size versus the length of the room impulse. If the frame size is larger than the room impulse, the lack of data causes a collapse of the assumption of independence between the original signals frequency bins [45].

Other methods of BSS have since been proposed, including separation through independent component analysis. As relating to audio forensics processing this technique is of little consequence as it requires at least as many microphones and there are sources, which is generally not feasible in a forensics situation [46].

Reference cancellation

Reference cancellation is the process of removing or “cancelling” elements of a signal in relation to another recording or “reference”. To serve as a simple example, if there is a musical recording featuring an artist singing, if an exact copy of the music without the vocal is available, the music can be removed from the original track, leaving the isolated vocal (Fig. 6.6) [47].

Vocal = Original track (music + vocals) – Reference track (music only)

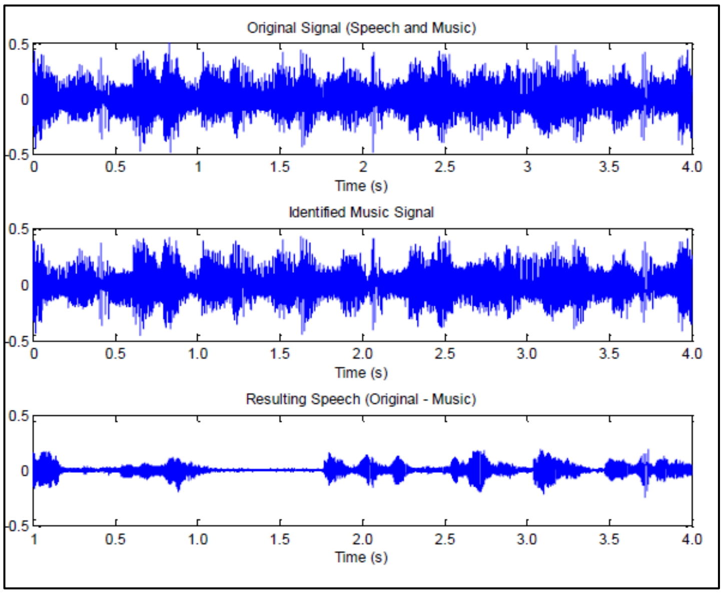


Figure 6.6: Reference cancellation waveforms

[47]

In reality, it is not as simple because the signal received in a forensic recording will have spectral additions such as reverberation from the room and the DAC responsible for playback of the music and the ADC responsible for capture. These issues are mitigated through the use of a noise signal cancellation algorithm, but it can often be painstaking in its implementation as the reference recording must first be located and then the recordings must be aligned precisely. Both the sample rates and average spectra of the reference recording must also be corrected to match the evidence recording. It is crucial that down-sampling rather than up-sampling is used as the recording being up-sampled would have no new components above the “original sample rate/2”, causing differences between the recordings in the high-frequency domain. As reference recordings are often richer in low-frequency content due to not having being convolved with the room and the capture microphone, it is required to correct this using manual filters or frequency equalization algorithms [48]. Recent years have seen an increase in software which can identify musical recordings (known as acoustic fingerprinting), such as Shazam [49], which makes the task of finding the musical track easier, rather than having to search through a myriad of records or recollect from memory the name of background track. The process is applicable to any recording in which it is possible to gain the reference track for, for example, music from a CD player, a television advertisement or a pre-recorded subway announcement. It must be made clear that the recording has to be an exact bit for bit replica for the method to be successful, so live music and speech are not applicable in this situation unless the music is being recorded through a multitrack device.

It is often the case in law enforcement that background music or television will be encountered when dealing with covert recordings. Individuals may purposely increase the amplitude of background devices to mask speech, or at least make it difficult to transcribe.

One method of removing the reference recording is through a landmark-based fingerprinting algorithm. It firstly analyses frequency peaks and the time difference between them and applies this to the original track to identify the song and start position of the music. [50]. Signal cancellation is then performed using a normalized least mean square (LMS) algorithm [51].

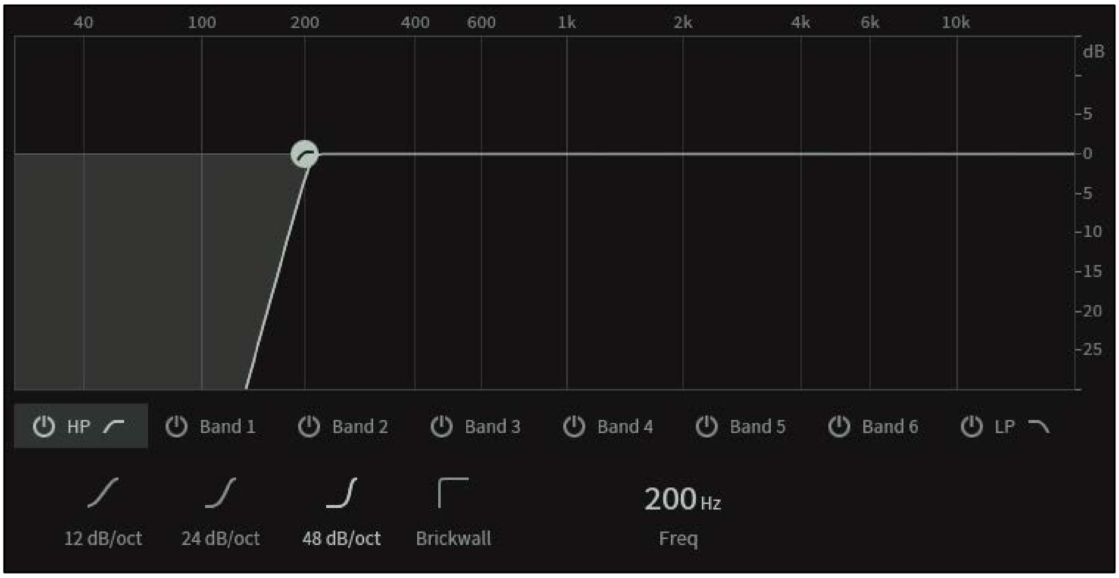
Another method is that of an adaptive filter which has 2 channels, one containing the recording and another containing a reference, possibly recorded on site. These filters compare both channels and determine which elements of the surveillance recording are from the reference and which are not. Similar to a standard adaptive filter, the process works through the prediction of future samples but does so from the reference track instead of the past samples of the surveillance recording [52].

Equalization

Equalization is a term used for processes which target the spectral content of a recording with the intention of changing the level of frequencies in relation to one another. It is so named as it was originally created to boost high frequencies that were lost over long telephone lines to “equal” the signals between the sending and receiving transmitters. Nowadays they are employed to manipulate the frequency spectrum, and in forensics specifically, they are used in attenuating areas that are of detriment to the desired signal and boosting areas which are required.

Static Filters

Technically, filtering is a form of equalization, and static filters work by attenuating or removing frequencies specified by the filter settings for the entire length of the recording, and are ideal for use on signals which remain constant throughout or are superfluous to the recording. The type of filter required for each recording may differ, but in general will be either a low pass, high pass or notch filter. Low pass filters work in attenuating the level of the signal above a user-defined cut-off frequency (the point at which -3dB attenuation occurs), leaving the spectrum below that frequency un-touched (Fig 6.7). High pass filters are the opposite and work to attenuate signals below a defined frequency (Fig. 6.8). Both allow the user to define the slope (the gradient between the cut-off frequency and the maximum attenuation). These are expressed in dB/octave and common values of multiples of 6, with a 6db/oct filter making a gentle transition and a 30dB/oct filter taking only 2 octaves to reduce the signal by 60dB, which is effectively muting [53]. A balance should be sought between the steepest gradient possible and the prevention of any artifacts that may be introduced in doing so, referred to as “ringing”, as they create a ringing type sound around the cut-off frequency.



48 dB/oct

200 Hz
Freq

Figure 6.7: High pass static filter

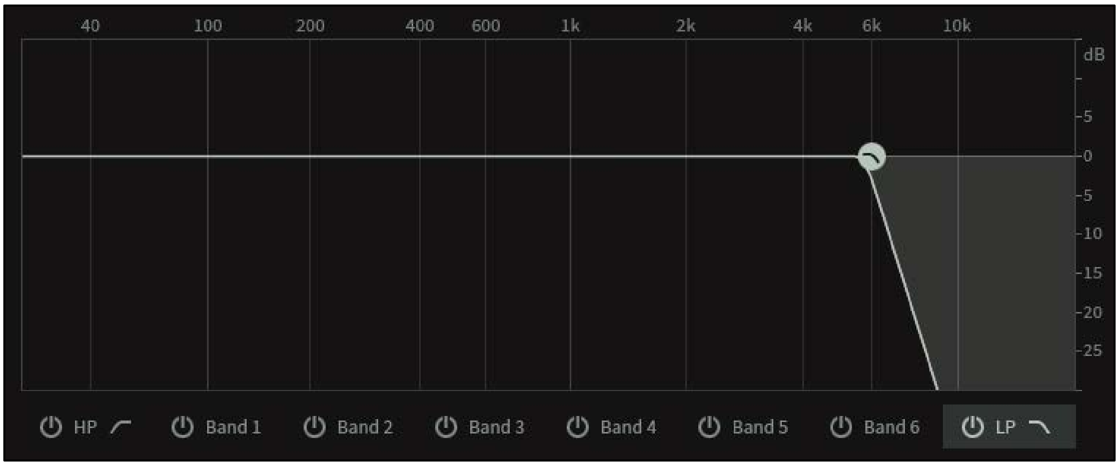


Figure 6.8: Low pass static filter

Notch pass filters are much akin to a surgeon's scalpel, allowing removal of small frequency ranges, ideal for static sounds (Fig. 6.9). Settings available to the user are the same as the comb-filter referred to in the section regarding the removal of hum (which is technically a series of notch filters).

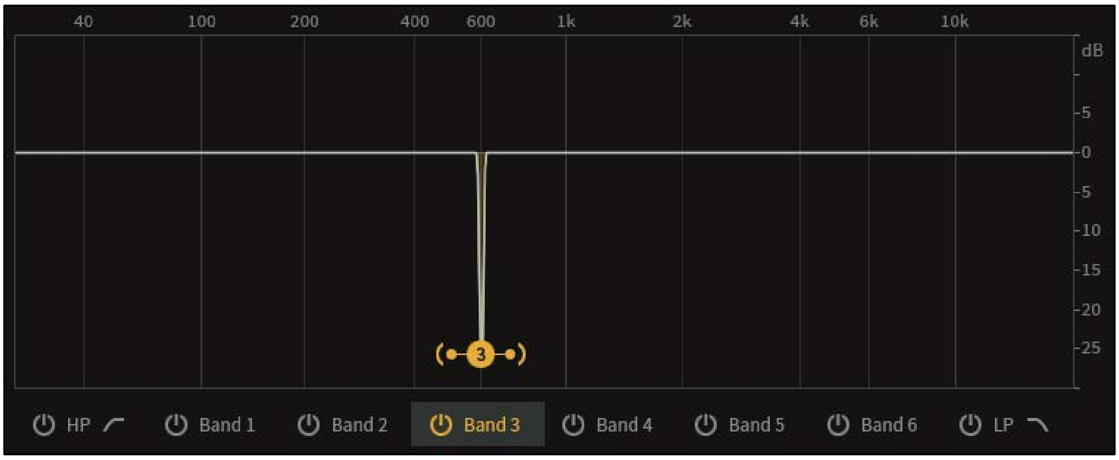


Figure 6.9: Notch static filter

Filters are ideal for the removal of any superfluous frequencies which have no relation to the desired signal, for example, a low pass filter removing everything above 6kHz if it is speech which is desired. Not only does this enhance the desired signal by removing noise, it also serves to refine the processes which follow by allowing them to focus on the desired signal and not waste resources on unwanted noise, resulting in optimal processing.

A digital filter is an algorithm that accepts input samples and applies an impulse response resulting in output samples. The input spectra are then multiplied with an ideal filter characteristic in the frequency domain (for example low pass filter), which is equivalent to convolving the time-domain signal with the impulse response of the time-domain. The filter characteristics take advantage of the fact that high-frequency signals contain larger differences between each sample than those of lower frequencies, due to a steeper gradient towards the peak amplitude point. Low pass filter characteristics work to reduce the range of frequencies, thus causing the higher frequencies to reduce. High pass filters work by increasing the range of lower frequency elements, causing the low frequencies to increase.

<u>Shelving Filters</u>

Shelving filters allow frequencies to not only be cut but also to be boosted. The reference frequency of a shelving filter leaves one side of the spectrum undisturbed while applying attenuation or boost to the other side of the spectrum, defined as gain by the user (Fig. 6.10). These filters are useful in subtlety changing the spectrum of a recording, for example in attenuating low frequencies caused by the proximity effect, or applying a small amount of boost to higher range vocal frequencies to improve clarity, and thus, intelligibility.

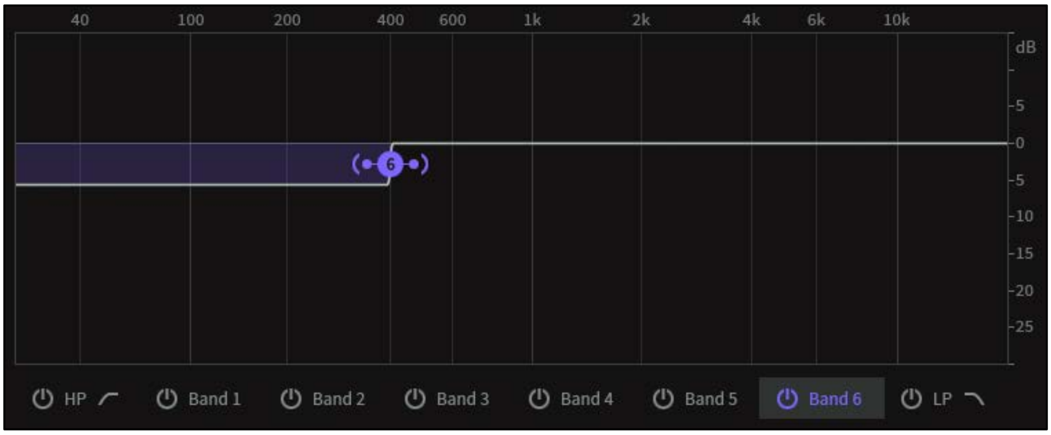


Figure 6.10: Shelving static filter

Dynamic Equalization

Where conventional equalization processes samples linearly based on the settings, independent of the audio input, dynamic equalization reacts to the audio based on both settings and the audio input, much akin to a compressor. Rather than performing compression on selective frequency bands like a multiband compressor, dynamic equalization uses traditional EQ in place of the gain reduction which is applied to a whole band using a multiband compressor, giving much more control over the signal [54]. Within forensic audio, there may be signals that appear only occasionally through a recording at a certain pitch. By using a dynamic EQ, the rest of the signal is unprocessed while the noise is not present, so only the noise is removed, and the process is non-destructive, which should always be chosen over a destructive process when working with forensic audio as it is of utmost importance to ensure as little of the original signal is removed as possible.

Adaptive Filters

When noise is static (statically constant), a static filter is capable of removing the noise to some degree to improve the quality of the desired signal. If, however, the noise is varying rapidly or has a complex spectrum, a filter is required which can "adapt" to the changing noise spectrum. Adaptive noise filters work by predicting future samples from previous samples, which can be especially effective with low-frequency repetitive noises such as wind, hums, low-frequency background speech and motors. Speech that competes with the desired signal is more difficult due to the unpredictability of speech. The number of taps is usually user-definable, but obviously, the more taps used, the more processing power required. Adaptive filters are also no use with unpredictable transient sounds such as clicks as these are also impossible to predict [52]. Even with good material, the filters can sometimes make the speech thin, requiring post-processing to enhance the speech.

De-Hum

Hum is one of the more common noises present on a forensic recording and is caused by electrical sources. One of the most well-known causes is that of ENF (Electrical Network Frequency), which can be inadvertently captured on a recording if the power source is mains driven or in close proximity of a mains powered device, through electromagnetism. This phenomenon is caused by the alternating current (around 50Hz in the UK, 60 Hz in the US) and has harmonics, albeit, at a lower amplitude, that can exist all the way into the bandwidth occupied by speech [55], [56]. Other causes of hum are lights and powerlines.

The frequencies present from this noise can affect the quality of the recording as they occupy an area of low frequencies and can be distracting to the listener. The low-level frequencies also contain a significant amount of power so will de-optimize the algorithms used by processing if they are not removed. In terms of speech intelligibility, the harmonics can extend into the frequencies occupied by the human voice, so it is a necessity that they are removed to prevent any simultaneous masking from taking place (Fig. 6.11).

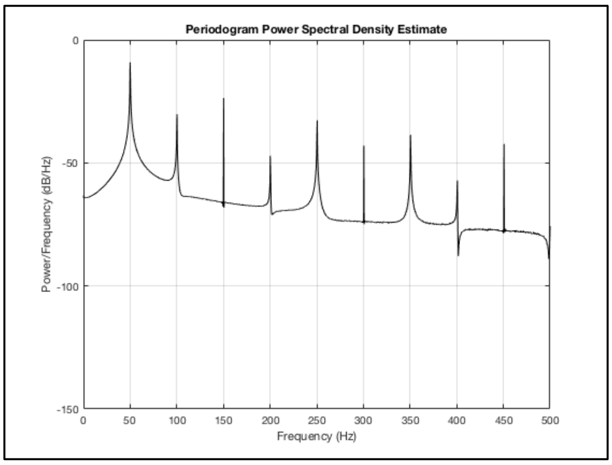


Figure 6.11: Hum FFT

To remove hum, a comb filter is generally the most effective means. It operates by applying notch filters at linearly spaced intervals to remove the fundamental and the integer harmonics of a sound (Fig. 6.12). They may not need to extend beyond 360Hz as the harmonics of hum often fall below the average signal level above this, so will be masked and inaudible [26]. If the frequencies are not linearly spaced, then adaptive filtering or manual notch filtering of multiple frequencies are the preferred options. Care should always be taken, especially in the higher order harmonics which share bandwidth space with speech as to keep the desired signal. Through the use analysis, high "Q" settings and gentle attention, the desired signal should remain untouched.

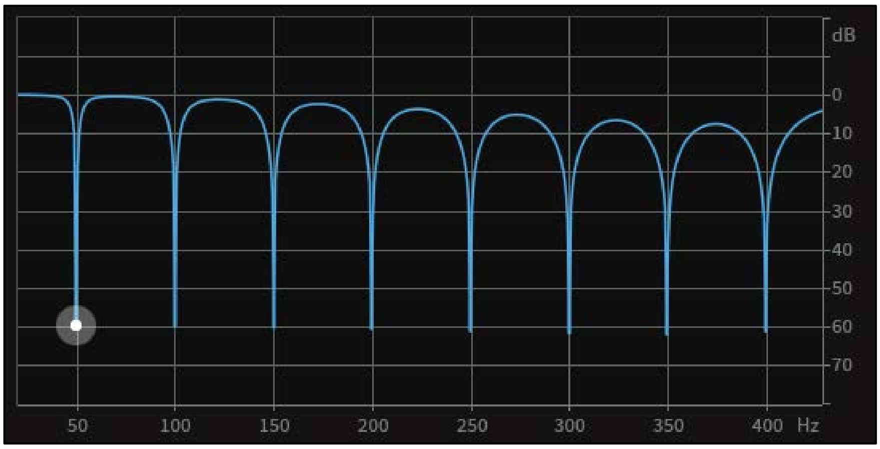


Figure 6.12: Comb Filter

Broadband Noise Reduction

Broadband sound refers to an effect which adds or subtracts a random amplitude throughout the recording across all frequencies within the audio spectrum. Wind is an example of a broadband noise, which appears primarily in the low-frequency range [57], but extends across the spectrum. Broadband noise reduction is a crucial element of many forensic audio recordings due to the high noise levels which can mask the desired signal, and improving speech quality is the most critical aspect of a noise reduction system. The spectrum of a noisy signal is the summation of the desired signal and any undesired noise and can be represented as:

$$X(n,k) = S(n,k) + N(n,k)$$

Where $S(n,k)$ represents the spectral coefficients of the speech and $N(n,k)$ the noise at frequency bin n, frame k [58]. The difficulty lies in having a single known $X(n,k)$, and 2 unknowns. It is required then to use a method to determine which parts of the known signal are noise and which are the desired signal. The methods for achieving this generally fall into 2 distinct categories, those which occur in the time domain and those in the frequency domain.

Time Domain

Time-based techniques work by attenuating the signal whenever a portion is detected that contains only noise, motioned by a drop below a user determined threshold. If the signal is above the threshold, it is presumed to contain noise and the desired signal, if below, it is presumed to contain only noise. This reduction in gain creates a perception that the noise level across the recording is lower than the original recording. This is generally referred to as a noise gate, and features settings such as threshold (the level at which the signal is attenuated), attack (the length of time it takes for the signal to be reduced), release (the length of time it takes for the signal to recover) and range (the amount of attention applied to the signal once it drops below the threshold). Some feature a look-ahead option which will analyze the audio before the area being processed to ensure no samples go undetected at the onset of the processing. The limitations of this method are that if the signal to noise ratio is high, even when the desired signal is present, it will do little to enhance the recording [8], and is also of no use if the desired signal is continuous in nature with no breaks.

Spectral Domain

Spectral domain processing is a ubiquitous technique in restoring a signal that has been corrupted by broadband noise and there are multiple methods to achieve the desired outcome. One of the more established techniques for the removal of broadband noise is that of spectral subtraction, first proposed by Boll in 1979. The technique utilizes estimation of the short-term spectral magnitude as a function of frequency and then subtracts the magnitude spectrum of noise from the noisy speech, assuming the noise is uncorrelated with the speech [59]. Since then various improved iterations based on spectral subtraction have been proposed and implemented, including critical-band spectral subtraction [60], multi-band spectral subtraction [61] and dynamic spectral subtraction, all which make use of the masking phenomena of the human auditory system [62].

The pitfalls of this technique are that any of the desired signal elements falling below the noise threshold will be removed with the noise, thus rendering this technique ineffective in extremely high signal to noise situations. It can also cause various artifacts due to statistical variance of short-time spectral estimates. The calculated gains can contain random oscillations leading to bursts of energy in the processed signal, known as musical noise [63].

Spectral subtraction can be a very effective technique, but this is dependent on the particular situation and requirements [8]. All currently known approaches in monaural recordings (which is often the format of audio analyzed within forensic audio) add distortion to the desired signal. There is therefore always a compromise between the noise reduction and speech distortion and various methods to assess the results should be used to ensure the best possible outcome for the presented recording have been achieved [64]. These include SNR measurements, critical listening, and FFT analysis.

Gain

Gain applies to increasing or attenuating the level of audio within a recording. There are various methods available to implement this enhancement and changes can be made locally or globally and based on pre-defined settings by the user. The various options will now be detailed.

Compression

It is a common factor of forensic recordings to be of a relativity low volume, particularly when using a portable device. This can result in only a fraction of the available quantization levels being occupied, specifically at the LSB (Least Significant Bit). Care should be taken as to the placement of compression within the chain as it is at the LSB level which quantization error is most prevalent, so increasing the level of the signal too early in the chain can result in unwanted artifacts becoming audible. By applying selective forms of gain in the right order, the number of levels used can be increased to increase the desired signal level.

Through the careful use of dynamic range compression, it is possible to increase the mean signal level while keeping the peak signal level. This allows lower volume signals that may be present such as speech and background noises of interest to be audible, but in doing so increases the SNR by bringing the level of the noise floor upwards. It can also change the dynamics of the signal, which could render the speech signal of a lower quality for speaker recognition tasks or the characteristics of a signal such as a gunshot for comparative analysis.

In terms of automated methods, the most effective is that of upwards expansion which will provide increased amplification to areas of low level than those of a higher level (Fig. 6.13, 6.14, 6.15). A downward compressor will provide the same result but through the diminution of higher levels relative to those that are lower. Typical characteristics of a compressor are the range, ratio and envelope settings. The range determines the level at which the compressor acts upon the signal. The ratio is the amount of gain reduction applied to the input and the envelope settings allow the time it takes for the compressor to act (attack) and stop attacking (release) upon the signal to be set. Fast attack times can smooth out increases in signal level and longer release times are required to prevent excessive distortion of the speech envelope [65].

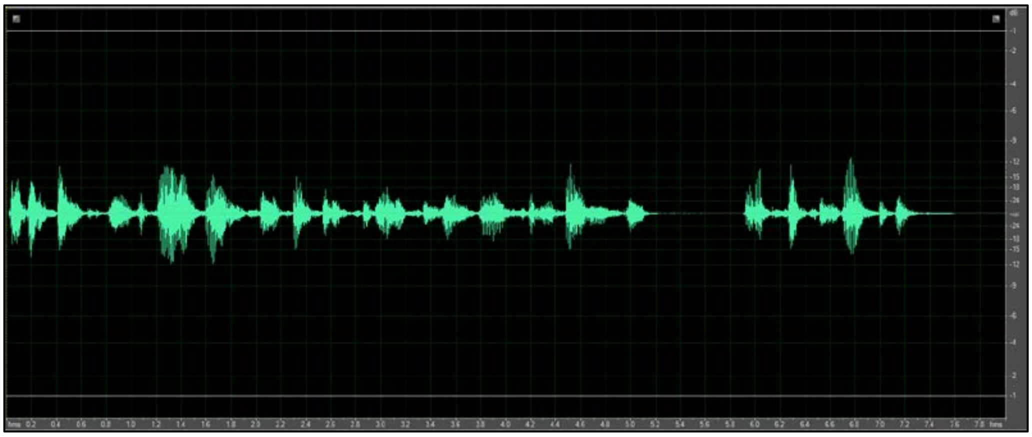


Figure 6.13: Pre-expansion waveform

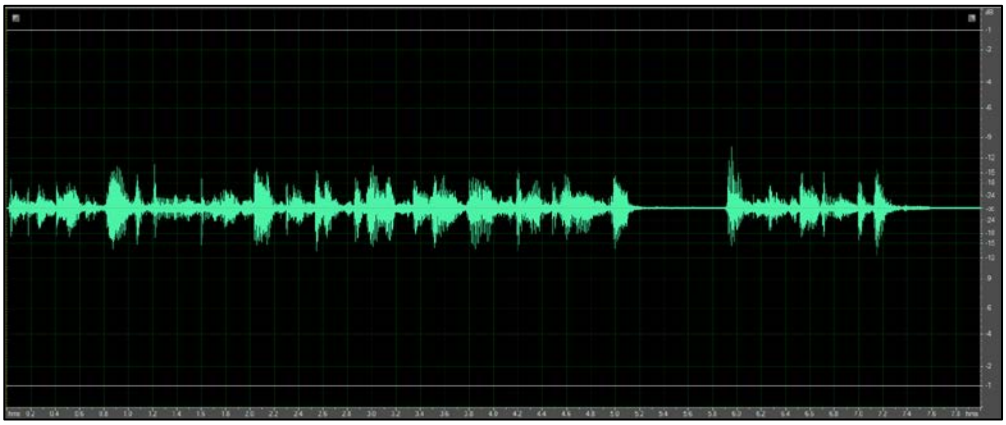
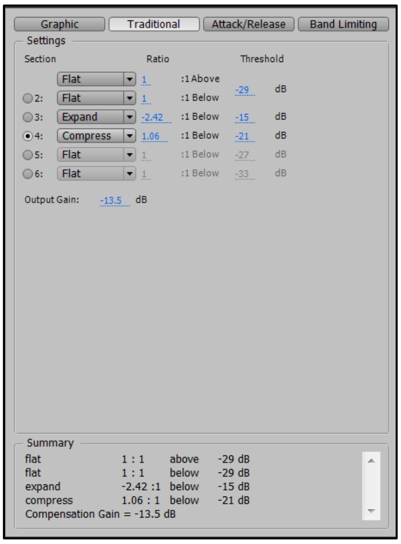


Figure 6.14: Post-expansion waveform



Settings

Section		Ratio		Threshold	
	Flat	1	:1 Above	-29	dB
○ 2:	Flat	1	:1 Below		
○ 3:	Expand	-2.42	:1 Below	-15	dB
● 4:	Compress	1.06	:1 Below	-21	dB
○ 5:	Flat	1	:1 Below	-27	dB
○ 6:	Flat	1	:1 Below	-33	dB

Output Gain: -13.5 dB

Summary

```
flat        1 : 1     above   -29 dB
flat        1 : 1     below   -29 dB
expand      -2.42 :1  below   -15 dB
compress    1.06 : 1  below   -21 dB
Compensation Gain = -13.5 dB
```

Figure 6.15: Expander settings

A more specific, time consuming, but precise method is that in which the examiner manually increases or decreases sections of the recording based on whether there is an area in which the level is too high or low relative to the entire recording. This is often the case in forensic audio, especially when recordings involve multiple speakers at differing distances from the microphone. For example, in the case of an informant wearing a wire, the individual wearing the microphone will always have the highest signal level, so the suspect's voice will need amplifying or the informants reducing. In a studio mixing environment, it is often encouraged to reduce the level of the loudest signal rather than bring the quiet levels up as this brings up the noise floor. As reducing the level will result in a loss of information from the recording, it makes more sense in forensic audio to bring the lower levels up as maintaining as much of the original signal as possible is vital when performing enhancements. Applying some light compression once this has been achieved can smooth out any level changes, analogous to smoothing the icing on the top of a cake to remove any slight bumps or ripples. This is the recommended technique, but time constraints can render it impractical for some longer recordings.

Limiters

Limiters are compressors which operate at a ratio of around 10:1 or above. At this ratio, the signal which passes over the threshold is being attenuated at such a high ratio (for every 10dB over the output is attenuated by 9dB) that it is “limiting” the signal (Fig. 6.16). Limiters are useful in situations when there may be short bursts of high-level sound amongst a relatively low-level recording (such as gunshots in a quiet suburban environment), but manual attenuation is always preferred as it provides increased control over the audio.

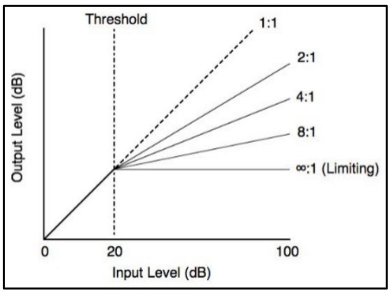


Figure 6.16: Audio limiter ratio [66]

De-reverberation

Forensic recordings are often recorded in less and ideal spaces. Add to this the lack of control of microphone placement and there is a culmination in highly reverberant recordings, which has adverse effects on speech intelligently through the loss of transient’s due to early and late reflections. The acoustic reverberation of speech can be described mathematically as:

$$x(n) = s(n) * h(n)$$

Where *s(n)* represents the clean signal and *h(n)* the impulse response of the environment [67]. The characteristics of reverberation are modeled as linear systems and the impulse response is dependent on the rooms acoustic properties as well as factors such as microphones position, direction and polar pattern [68]. To remove the impulse response, various techniques that fall under the umbrella of “de-reverberation” are used. This is the identification and removal of reverberation from the desired signal, using digital signal processing [69].

In a forensic context, there are often a few limitations caused by the recording process. Firstly, it is usually the case that the only information available is from a single microphone, which can mean cues such as the Interaural Time Difference (ITD) and Interaural Level Difference (ILD) are lost. Secondly, if no visual information such as video footage is available, clues that can be garnered from lip-sync reading are also lost. Finally, and most crucially, forensic examiners always work blind, that is to say, that the clean signal information is never available and all that remains is an equation with two unknowns and a single known (the evidence recording). De-reverberation has many elements similar to that of blind source separation and spectral subtraction, where in this instance it is the acoustical impulse of the room and the speech signal for which separation is required.

The effects of reverberation on speech can deem speech unintelligible, regardless of other factors such as distortion or noise. The main reason for this is a type of pre-masking known as overlap-masking, which refers to the energy of the previous phonemes being smeared over time and overlapping the following phenomes [70]. The process behind the smearing is due to the multiple reflections and diffusions of sound waves on boundaries and obstacles in a room, corresponding to late reverberation. This causes the reverberated energy to decay exponentially, with a time constant dependent on the room. This results in the reverberant tails for each phoneme having an exponential decay similar to that of the room impulse response [71].

De-reverberation is a form of convolved noise removal and there are three specific areas of approach to dereverberation. The first is speech enhancement, which exploits the characteristics of speech and the effect of reverberation on speech. The second is spatial processing, which uses multiple microphones placed in specific locations. This area can be disregarded as it is not applicable to forensic recordings. Finally, blind de-convolution uses channel identification to determine the room impulse response between the source and receiver and use the information to equalize the channel [70].

Most methods are a two-step procedure: the room response must first be estimated, then an inverse filter applied, either through least-squared error or cepstral separation techniques.

Normalization

Normalization can apply to both amplitude and frequency, but for the purpose of enhancement it is applied to amplitude and is the process of boosting the signal level to a maximum without causing clipping. For instance, if the maximum peak signal of a recording is -20dB, the signal can be increased by up to 20dB (although in reality a small amount of headroom is advised to be on the safe side). Rather than applying gain in various areas of the signal, such as a compressor does, the increase in level is applied to the entire signal indiscriminately. The issue with this process is that quantization noise will increase as the recording has already been quantized, especially on recordings of 16 bit or less. Increasing the level will leave the original SNR unchanged and add more even quantization noise as the audio is being re-quantized, so the sum will render the recording noisier than before. If normalization is applied at the beginning of the enhancement process, it will be reducing headroom which may be required for future processing. Normalization is required to allow an optimal level of playback for the listener.

There are several distinct methods for normalization, based on the signal peaks (peak volume detection) or the overall level of a signal (RMS volume detection). Peak volume detection only considers the peaks of the waveform and is used when bidding to achieve the loudest volume possible from the signal (Fig. 6.17, 6.18), but is safest as it can be ensured that the peak element of the signal will not exceed the normalization setting. RMS volume detection performs an RMS calculation on the audio file, which is mathematically represented as:

$$\text{RMS} = \text{Voltage Peak} \times 0.707$$

RMS does not consider the equal loudness contours of the human ear, which could result in bandwidths that are insensitive to the human auditory system being given more weight than is necessary, resulting in unnatural level changes. Due to these issues, a final method is available, EBU R-128 Volume Detection, which does consider the way in which the human ear performs, so normalized files are more consistent in their volume. When using either the RMS or EBU methods, care should be taken so that no peaks extend beyond 0 dBFS, which will result in clipping (Fig 6.19). If the previous processing stages have been carried out correctly, then all normalization methods available are acceptable, so long as the disadvantages of each are taken into account [72] and the signal is visually analyzed post-processing to confirm no clipping is taking place.

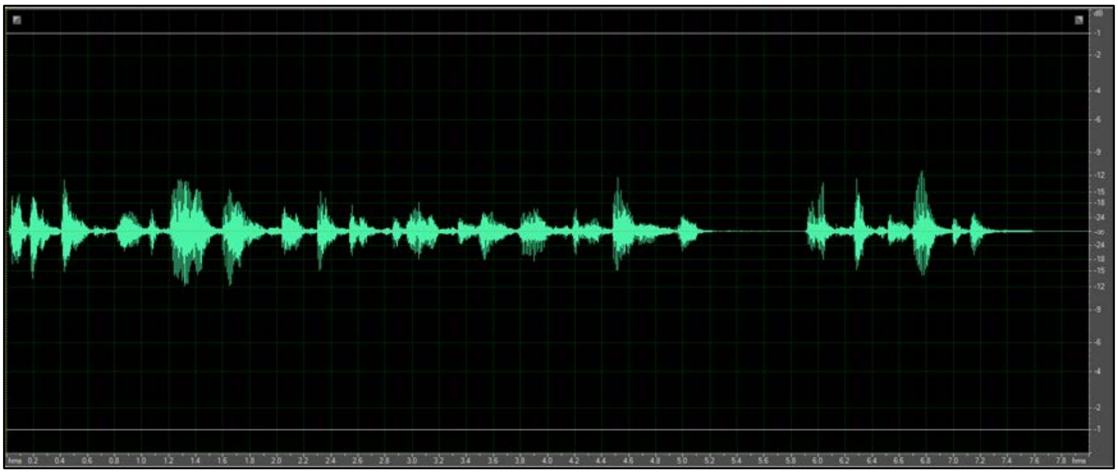


Figure 6.17: Normalization pre-processing

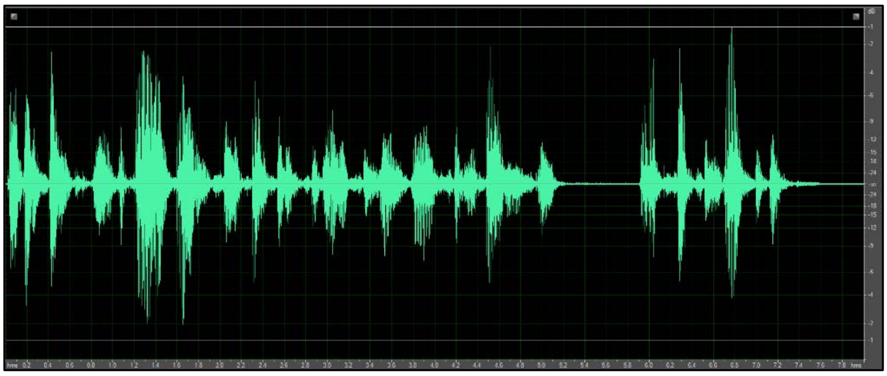


Figure 6.18: Peak level Normalization

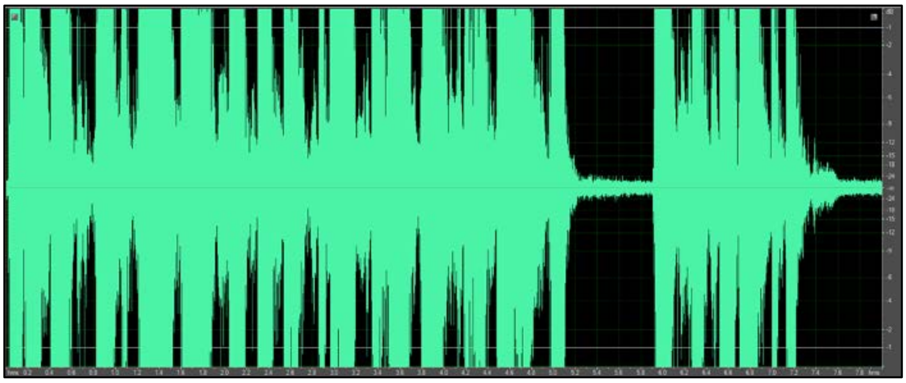


Figure 6.19: RMS Normalization (Resulting in clipping)

Time-stretching

Time-stretching is a method in which the speed of the recording is reduced or accelerated. In audio forensics, it is reduced to increase intelligibility while taking precautions to maintain the pitch. The voice can also be slowed through down-sampling, which spreads the samples linearly. As the samples are spread, the waveforms extend, causing the speed to decrease, but also the pitch to change. This is not recommended as it could mask the identity of a speaker, so software which allows time-stretching should be used rather than down-sampling. The software interpolates the samples and changes the start and end point of the audio in relation to this interpolation. Time-stretching is not available in all audio editing software's so considerations should be made by the analyst to ensure this tool is available to them. Care should also be taken as to limit how far the audio is stretched the as more artifacts are introduced the longer the audio samples are stretched. A safe recommendation is to stay within a 75% limit of the playback speed of the original audio.

CHAPTER VII

PROPOSED FRAMEWORK

The structure for the proposed framework model will be based on the work of Ledesma [74], who provided a similar framework relating to forensic image enhancement for his thesis project. The initial basis for the sequence of Source Separation, fix distortions, EQ/filter, noise reduction, de-reverberation and amplitude correction (in that relative order) was acquired from an NCMF lecture on audio enhancement [75]. From this foundation, the framework has been populated and built upon by the research. A color code has been developed to determine the likelihood of application at each stage when performing an enhancement, as some stages will only be required when the situation fits, while others will be required every single time. For ease of viewing, the entire table will be displayed on the following page.

🟩 Required

🟧 When situation requires

Table 7.1: Proposed framework

Evidence Preparation	
1	Working Notes
2	Ensure Playback
3	Level Optimization
Analysis	
4	Critical listening
5	FFT analysis
6	Define strategy
Processing	
7	Fix distortions - De-Click - De-Clip - Spectral Repair
8	Source separation - Blind source separation - Reference cancellation
9	Remove continuous noise - De-Hum - Static Filters - EQ
10	Remove dynamic noise - Adaptive Filters - Dynamic Filters
11	Remove noise - Blind spectral subtraction
12	De-reverberation
13	Gain correction - Equalization - Compression - Expander - Manual Gain - Limiter
14	Normalization
15	Time-Shifting
Output	
16	File output
17	A/B Comparison
18	Preparation of deliverables

Framework Rationale

In an attempt to disprove the hypothesis, testing will be composed of 3 stages. Firstly, the framework will be compiled from a logical scientific standpoint, based on the background research of all enhancement processes. During this stage, the changes made at the sample based level will be considered, in order to understand how the signal is being manipulated at a micro level. The macro-level will also be considered to gain an overall and practical perspective of the changes made to a signal during processing. Secondly, case studies will be created in which audio recordings are applied to the framework and enhanced using the processes in various sequences with identical settings. Finally, results will be judged through the use of objective, quantitative analysis. In summary:

1) Generate a logical framework based on the scientific research of enhancement processes.
2) Enhance case study recordings using processes with same settings but in different sequences.
3) Objectively judge and compare the results of the enhancements.

Evidence Preparation

The preparation of evidence is the crucial first step on which all enhancements are built, as if this stage is performed incorrectly it could at best be the causation of a less than optimal enhancement and at worst see the exhibit prepared deemed inadmissible by the court. It is assumed that the evidence has been received and properly processed according to documents such as SWGDE Best Practices for Forensic Audio [15] or ENFSI Guidelines for Best Practice in the Forensic Examination of Digital Media [9]. There are three steps within the framework that relate to evidence preparation.

1) Working Notes

A document should be created in which detailed notes on the evidence, all processes performed, and the settings for each enhancement process are stored. This not only shows proper procedures were followed but allows repeatability of the enhancement by another trained individual, which displays both an understanding and a transparency of techniques and methods used to achieve the enhancement. This file should be a living document to be updated throughout and requests may be made for the notes to be included within the deliverables once the enhancement has been performed, so should be formatted in a manner that is easily accessible to others, not just as a reference point for the examiner carrying out the enhancement. Initial notes should include sample rate, bit rate and recording length amongst other findings. The creation of a working notes document as a first step in the enhancement process ensures every process is documented from the very beginning, maintaining an extremely strong chain of custody.

2) Ensure playback

Before any analysis can be executed, proper playback must be ensured. If evidence is submitted as a proprietary format it will most likely not be playable by all software required during an enhancement so conversion to a format that allows this is required. For uncompressed audio, the sample rate should be maintained and a minimum bit depth of 16 utilized, preferably higher. Perceptually encoded audio should be transcoded to a linear, uncompressed format of a reasonably high sample and bit-depth (at a minimum, WAV PCM, 44.1 kHz, 24 bit). In general, operating at 32 bits with iterative saves before exporting at 24 bits for the final product should be practiced. Performing this procedure at any other stage other than at the beginning of the process makes no sense as all processing before conversion will have occurred at a less than optimal standard, as it would have been performed on recordings of limited sample frequency and quantization levels. The signal can also be decimated to the desired bandwidth at this stage, for example, by reducing the bandwidth to 11025Hz when enhancing speech. This removes parts of the spectrum that are unnecessary for the specific enhancement. This stage may be skipped if the recording submitted is already in a ubiquitous uncompressed format.

3) Level Optimization

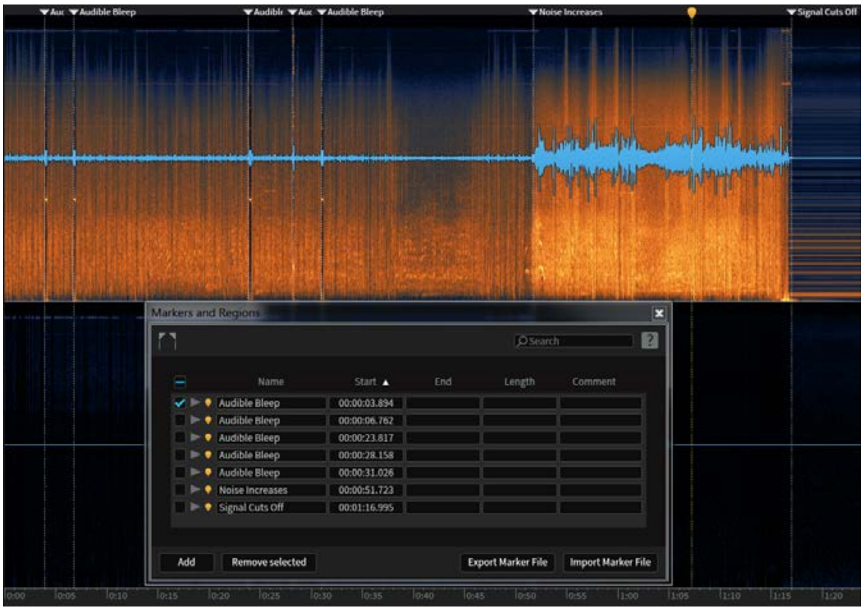
As processing audio can increase quantization levels, the audio should also be reviewed to ensure there is enough headroom for processing to take place without introducing artifacts such as clipping. To prevent this, ensure the file is at a maximum of -3dB dBFS before any processing takes place. If the overall amplitude of the recording is very low, it may be necessary to increase the levels before enhancement. This can be done through normalization or manual gain and should be applied across the entire length of recording and not in sections.

Analysis

A strategy is required before processing of the audio can take place and is informed by both critical listening and FFT analysis.

4) Critical Listening

Critical listening to document the occurrences of such features as clicks, clipping, and changes in the overall signal (e.g. amplitude, reverberation, noise level) should be performed on the entire length of the recording at the original playback speed. The times of each occurrence should also be noted. Some software does allow the use of ‘markers’ which are encouraged (Fig. 7.1), but more detailed word-processed notes should also be taken as to the critical listening analysis for clarity and in the preparation of request from clients. Spreadsheets can also be of use during this stage to tabulate the temporal and qualitative data. Any findings from this stage should be included in both the working notes and forensic report. This stage is placed before the FFT analysis so findings from the critical listening can be visually investigated further via plots such as spectrograms and LTAS.



	Name	Start ▲	End	Length	Comment
☑	Audible Bleep	00:00:03.894			
□	Audible Bleep	00:00:06.762			
□	Audible Bleep	00:00:23.817			
□	Audible Bleep	00:00:28.158			
□	Audible Bleep	00:00:31.026			
□	Noise Increases	00:00:51.723			
□	Signal Cuts Off	00:01:16.995			

Figure 7.1: Notetaking markers

5) FFT analysis

Notes should again be taken on the results of the FFT analysis, including the bandwidth of speech, signal to noise ratio and prominent outliers in the frequency domain. Settings such as the FFT resolution and window type used should be recorded to provide lucidity as to how the findings were reached, as all processes hereafter will be directed by these discoveries. Findings from the critical listening stage can here be appended with more detailed quantitative knowledge relating to the level and frequency of the offending signals.

6) Define Strategy

Once analysis notes from have been taken, decisions can then be made (based on the processing order of operations (Table 7.1) and the ‘Common problems and solutions table’ (Table 3.1)) as to how enhancements will be made. A strategy is vital to success and gives the opportunity to make informed decisions regarding the sequence of processes that will be applied. Considerations should be made as to the context, for instance, would removing every single click make the speech more audible?

Processing

The key to ensuring the best result possible depends on the application and optimization of each processor within the sequence. To do this, the user should understand every process applied and order the process correctly to assist the algorithms in focusing on only the desired signal by removing superfluous elements at each prior stage. An example would be reducing the bandwidth of a signal to increase the usefulness and flexibility of a filter [52]. Enhancement processes should be considered carefully and only be applied when essential, as unnecessarily using techniques will only increase quantization noise and produce a final product that is worse than the original. To this end, enhancement is a skill in which there is no "one size fits all", or applying every possible enhancement process expecting results. Through constant referral to the pre-composed strategy, analysis notes, A/B comparisons and applying careful consideration and reasoning to each decision, the final enhanced product will be cleaner than the original.

7) Fix Distortions

Certain noises are unsuitable for processing through complex algorithms such as adaptive filters and spectral subtraction, including stationary noises such as clicks due to their unpredictable nature. If these are not removed before processing they can cause the processors to react in ways which reduce its overall effectiveness as it cannot predict the distortions and will also use singular events such as a click to predict the behavior of future samples. Therefore, any transient distortions should be removed as a first step within the enhancement framework to optimize the processing stages which follow. When fixing distortions, the area of disruption should be tightly focused as to prevent the unnecessary removal of surrounding information. Zoom tools and careful selection of the start and end points will ensure this is achieved.

8) Stereo source separation

Source separation can be the first stage of processing, provided there are no distortions. It may also be skipped if the technique isn't applicable, such as if the recording is a mono format without any backing music which may be removed using reference cancellation. As these algorithms rely on the information within the recordings to be as closely matched to the original as possible to be efficient, any processing larger than the removal of distortions will change too many samples within the evidence recording that are required for processing by the reference. Consider that reference cancellation requires the evidence signal to match sample for sample as closely as possible to the reference signal. If the relationships between samples are excessively changed within the evidence signal it will render the process impossible. In terms of the amount of redundant data removed, source separation removes the largest amount. So, performing this technique as soon as possible also allows the processing which occurs afterward to focus on the desired signal and the remaining noise surrounding it.

9) Remove continuous noise

As the desired signal is often limited to a specific range of frequencies relative to the entire signal spectrum, the surrounding data can be removed easily through the use of filters. For example, if speech is the desired signal, the frequencies over 5kHz may be removed to provide further clarity to the desired area, analogous to removing the weeds to expose the flowers. The decimation of the signal during the initial preparation should have removed much of the unnecessary areas of the spectrum, but any other areas should be removed at this stage. Performing this procedure before complex adaptive processes such as de-reverberation and adaptive noise reduction saves on processing power, allowing the algorithms to be more flexible and optimized by focusing on only the area that is desired. There is little logic in performing complex processing techniques on any elements of the signal which are unrequired by the desired final enhancement.

10) Remove dynamic noise

Now that all signals which may reduce both the processing power and efficacy of complex processors have been removed, the remaining redundant data that can be manually determined through FFT analysis and critical listening must be removed. Processors which follow this are determined by machine learning, so by performing all processing that is possible through manual means first gives them the best chance of functioning at their optimum level. Care should always be taken as to not lose any information that is part of the desired signal. By listening to both the audio which will be removed and the audio which is being kept, these decisions can be made with confidence. This step can be seen as cleaning up the unwanted areas that the static removal stage missed. It occurs after the first gain correction stage as it does require some machine learning which may become confused by high-level signal bursts and low SNR signal areas.

11) Remove Noise

Automatic techniques for noise removal are employed at this stage as all manual methods have been exhausted. This stage must be placed before de-reverberation as reverb involves a specific convolution which smudges transients and any noise which is masking these areas will confuse the de-reverberation algorithm, making it less effective. If de-reverberation is performed before noise removal, it is possible that the algorithm will not be able to differentiate between noise and reverb tails and so will not be effective in performing the task of removing reverberation. For optimal performance when using de-noising, an area with a similar noise profile to the area in which the algorithm is being applied should be chosen for machine learning. Adaptive filters are also applied here to remove clean up any unrequired areas of the signal that were not removed by static filtering. Segmental processing should be considered if there are changes to the environment during the recording, for example, indoor to outdoor. In this case two specific noise removal instances will most likely be required. It is not uncommon to perform multiple passes of de-noising at different settings to achieve optimal results.

12) De-Reverberation

The removal of reverberation should be applied once all other noise has been removed so the signal transients and reverberation tails are exposed. Reverberation is highly correlated to the signal so any noise that coexists will cause erroneous processing of the reverberation, which could result in both increased noise and/or less than optimal results.

13) Gain Correction

A final gain correction is employed at this stage to enhance both the level and balance of the desired signal. Manual level changes or an expander to match levels of two speakers should be considered at this stage, now that all noise has been removed and the final levels can be accurately judged. Careful adaptions to the frequency spectrum can now be employed to enhance clarity and listenability, such as adding brightness through shelving filters or attenuating signals to remove the effects of the low-proximity effect during recording. Equalization should also be applied after filters as if applied too early any wideband filter boosts can make specific equalized frequencies hissy.

14) Normalization

The final stage of enhancement processing is to ensure the overall amplitude of the audio is at a level which guarantees listenability on all systems. Performing this before any processing takes place will result in an increase in the level of the noise floor, hence making it more difficult for the processor to differentiate between noise and the desired signal. It will also reduce signal headroom, leaving no room for processing. When performed on the final “clean” signal, the increase in noise is minimised as it has previously been removed by all other processes within the framework. It is recommended to normalize to -1 dBFS using the peak detection algorithm for optimal amplitude without clipping.

15) Time-Stretching

On occasions when intelligibility of speech is paramount, time-shifting may be required. This is performed after all other stages, as although it may be semantically seen as an enhancement, the desired signal is not being enhanced in terms of quality. To put this stage anywhere other than as a final step, therefore, would make no sense.

Output

16) File Output

The final stage of an audio enhancement is to output the media in a non-propriety format to allow playback on all media players, as assurance that it will be playable for the final audience and within the courtroom. This may be in the form of an Audio CD or Data CD, dependant on the client's request. The format should be in a standard lossless, uncompressed encoding format to guarantee audio is not lost in the process. Caution should be used as to the sample rate chosen as down-sampling from the original rate will result in a loss of information. This is something that may be of use if the enhancement contains speech only (down-sampling to 11025Hz), but considerations should also be made whenever the sample rate is being modified. It is recommended that an appropriate sample level and bit depth is maintained with regards to the content of the recording. The hash sum should be calculated, preferably as SHA-256 as it has been shown that SHA-1 and MD5 have had "collisions" in recent years, meaning two files have shown to have the same hash [76]. Although the level of effort to compromise these hashes is high, by using an uncompromised hash such as SHA-256, it prevents any questions over the integrity of a file. Utilizing more than one Hash sum will only serve to give the results more confidence as the likelihood of compromising multiple hashes is incredibly small.

17) A/B Comparison

Although comparisons should be made throughout between the working copy pre-processing and the effects of processing, a final comparison should always be made to ensure the processing hasn't made the recording less intelligible, of lower quality, added audible artifacts and is generally listenable. As previously stated, although higher quality doesn't mean increased intelligibility, it is rarely the case that a recording will become lower in quality but more intelligible and vice-versa. If there are problems present when performing the final A-B comparison, steps should be taken to discover why and to remedy those issues.

18) Preparation of deliverables

Deliverables that should be included with the enhanced audio file/s include the original version of the recording (to allow for comparison by the client) and a forensic report. The report should detail the client’s requests, hash values of the original file, analysis findings, processes which took place to enhance the recording and the file name, hash values and file size of the enhanced recording. Working notes may not be required but can be delivered on request. This is always the final step of enhancement.

CHAPTER VIII

CASE STUDIES

All case studies were created to simulate real-world situations, and in keeping with this, real-world sounds were recorded and implemented rather than artificial sounds such as white noise. As one of the intentions of this study is for the framework to be adopted within working forensic laboratories, results gained from artificial lab conditions may not be applicable in the real world. Although testing real forensic cases would ensure these conditions are met, access to the original clean speech (a pre-requisite to determine the degree of improvement performed by the processing) is vital. A clean reference recording quoting "Until they became conscious they will never rebel, and until after they have rebelled they cannot become conscious." was created on the same recording device and using the same settings for which to compare FFT data from the case study "evidence" recordings for analysis, as should be done in real-world cases (Fig. 8.1, 8.2, 8.3). This helps to distinguish the profile of the desired speech signal from any surrounding noise.

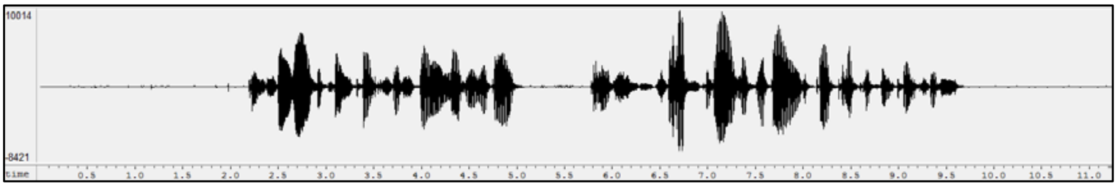


Figure 8.1: Clean speech reference waveform

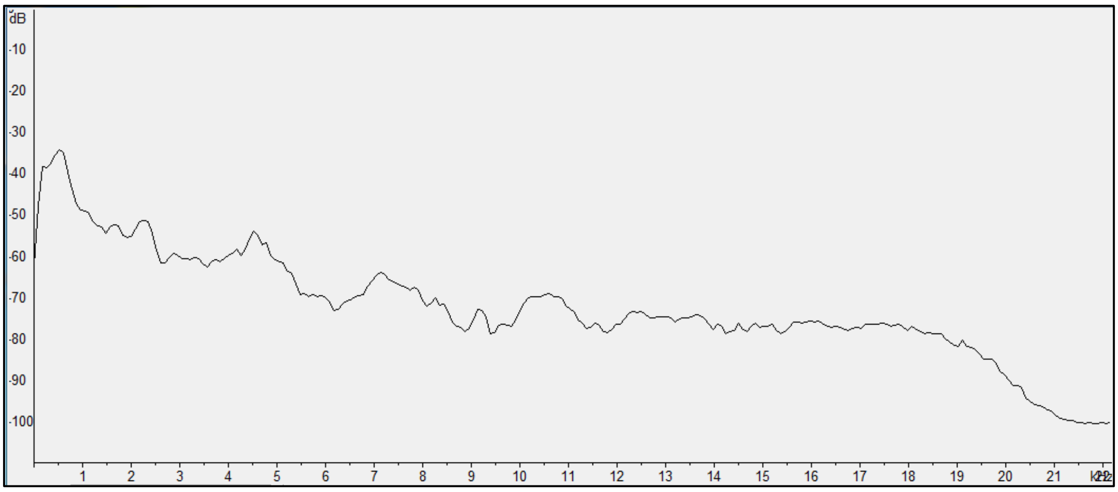


Figure 8.2: Clean speech reference LTAS

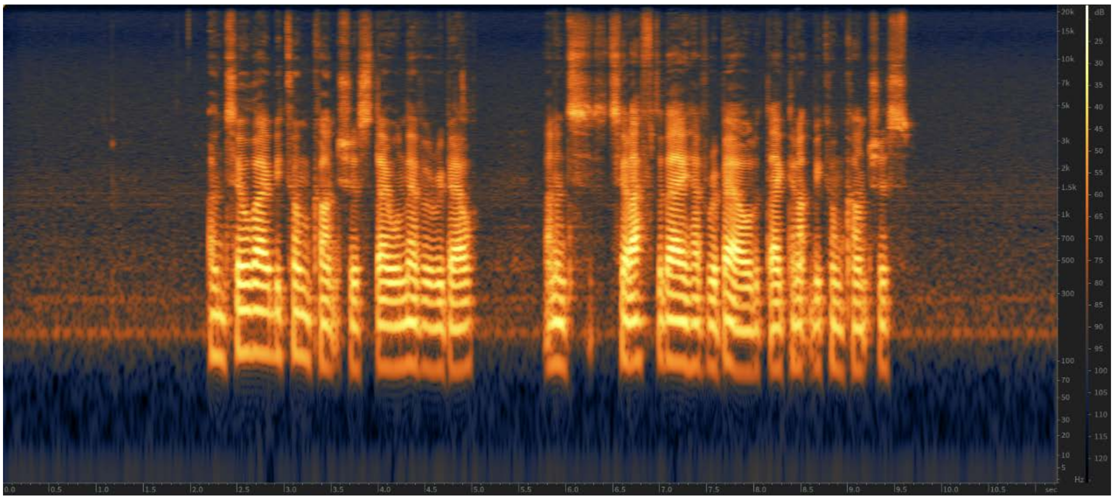


Figure 8.3: Clean speech reference spectrogram

To evaluate the performance of the enhancement, there are several methods which will be used. Although results of an enhancement are judged subjectively by parties such as transcribers, judges, and juries, these types of judgment can be expensive, time-consuming and require access to trained listeners. Although more reliable, since they are human interpretations, when there are several very similar results, the intricacies may not translate to the human ear, whereas objective algorithms will extract the finest details. Several objective methods have been developed which do predict accurately the results obtained from subjective human listeners. All objective algorithms work in a similar manner, by first segmenting the speech into short time windows of 10-30ms and then computing the difference in distortion between the original and processed signals. An average is then taken from the results of the segmented frame results. Of all the algorithms available, only those with the highest correlation to human subjective results are used, which are detailed below.

PESQ

PESQ (Perceptual Evaluation of Speech Quality) is a measure of objective speech distortion within a recording, by giving estimates of subjective mean opinion scores (MOS's) from humans. This is achieved through simulating our auditory system and making comparisons between the degraded speech and a clean reference of the same recording (Fig. 8.4). The method has been tested against a variety of distorted speech including experiments with background noise. The average correlation with subjective MOS was found to be 93.5% and is the recommended test for objective assessment of speech quality [77].

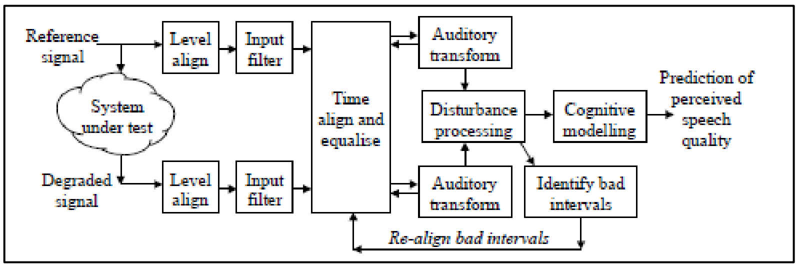


Figure 8.4: PESQ flow [78]

LLR

LLR (Log-Likelihood Ratio) is a LPC (Linear Predictive coding) based measure which represents the degree of correlation between smoothed spectra of the evidence and reference recordings [79]. It is calculated from the mean of the smallest 95% of the LLR distances at each frame. Smaller values indicate better speech quality.

Frequency weighted segmental SNR

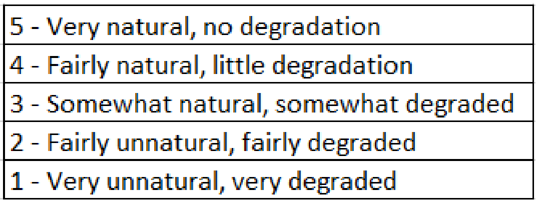
Frequency-weighted segmental SNR (fwSNRseg) is the SNR average for each frame, where the SNR is computed as the weighted-average of the SNR in K critical bands [80].

Coherence Speech Intelligibility Index

Coherence Speech Intelligibility Index (cSII) is used to evaluate the *intelligibility* of the speech present within a recording rather than the quality (as calculated by the previous algorithms), and uses the Speech Index as a base level and computes the signal to distortion ratio for every critical band. It then separates the signal across three level regions, those of high (containing primarily vowels), mid-level (containing mostly vowel/consonant transitions) and low containing weak consonants [81]. cSII shows significantly better results than other objective speech intelligibility predictors [82]. The mean of these results was then calculated for an overall review of the cSII. Although intelligibility is not expected to be improved due to reasons previously stated, it will be tested none the less in aiding a thorough and exhaustive investigation into how the enhancement has affected the evidence.

Composite

As the average of multiple differing tests is likely to be more reliable and accurate than those of single tests, a composite method was introduced by Hu and Loizou by linearly combining the PESQ, LLR and WSS (weighted slope spectral distance) to test the distortion of speech within a signal [83]. It was shown that this type of analysis can greatly improve the correlations of existing techniques. The rating system correlates with that of MOS, which allows the listener to score the speech signal on a scale of 1-5 (Fig. 8.5 [84]).



5 - Very natural, no degradation
4 - Fairly natural, little degradation
3 - Somewhat natural, somewhat degraded
2 - Fairly unnatural, fairly degraded
1 - Very unnatural, very degraded

Figure 8.5: Composite/MOS scoring chart

It is vital to have both a reference and the input and output results to determine whether an improvement in quality and/or intelligibility has been achieved [64]. Tests involved comparing the clean speech against itself to obtain the highest result possible from the algorithm and the audio file created for the case study (named 'Evidence') was then compared to the original clean speech file, followed by the 6 files which were processed (6 files due to the 6 variations in sequence available from 3 different processors). The enhanced results were then sorted based on the reference result, for example, if the reference result was 5, the results were sorted highest to lowest. If the reference result was 0, the results were sorted lowest to highest. In doing this it enables the highest rated enhancement sequence to appear at the top with the lowest at the bottom. To obtain results, files were processed using algorithms developed by Philipos C. Loizou [81].

To create audio for testing, three case studies were designed, consisting of various problems which are found in forensic audio recordings. They revolved around a recorded speech sample, quoting "Doublethink means the power of holding two contradictory beliefs in one's mind simultaneously, and accepting both of them", from George Orwell's dystopian novel Nineteen Eighty-Four [85]. In creating forensic type audio recordings rather than using actual cases, it allows for all elements of the recording to be carefully measured, and access to the clean speech, something which is vital if measurements are to be made relating to how the audio has changed. All case study files were output to 44100 Hz, 16 bit, mono, uncompressed WAV format to simulate the type of evidence received in real-world situations.

Case One – Clicks, noise, and reverb

The first study was created to determine the optimal sequence for the removal of noise, reverberation, and clicks. Reverberation was present as the recording chosen was taken 4 meters from the microphone in a domestic environment, adding authentic reverberation. Noise was captured from the London Subway system in the form of a 44,100Hz, 16 bit, mono recording and added using Colea software [86] at an SNR of -5dB. Clicks were then recorded and added to the recording, at levels which protruded significantly from the original waveform (Fig. 8.6).

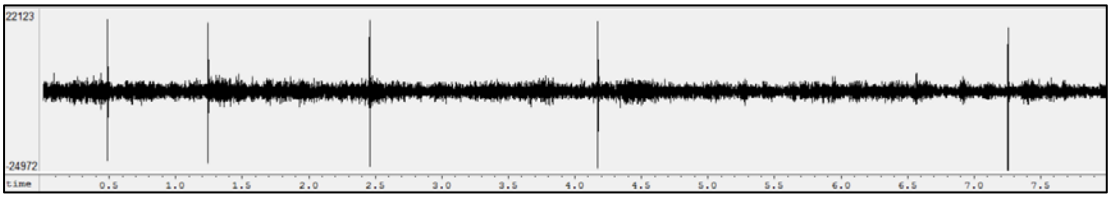


Figure 8.6: Case 1 waveform

Critical listening

Upon listening the problems were obvious and are summarized below:

- Large-scale, transient broadband subway noise masking the speech;
- Small level of reverberation causing poor speech intelligibility;
- 5 distinct clicks, spread sporadically across the length of the recording.

FFT Analysis

- Spectrum extends from 0– 20kHz range;
- - 40dB Clicks occur at approximately 0.5s, 1.25s, 2.5s, 4.25s, 7.25s;
- Speech spectrum is hidden by noise.

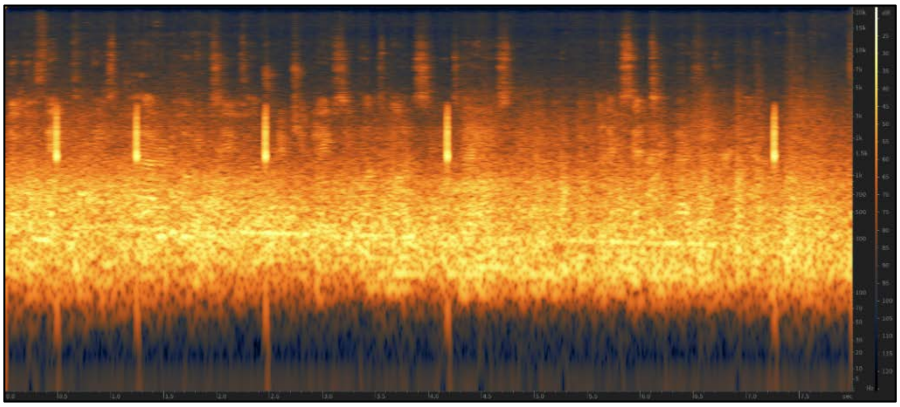


Figure 8.7: Case 1 spectrogram

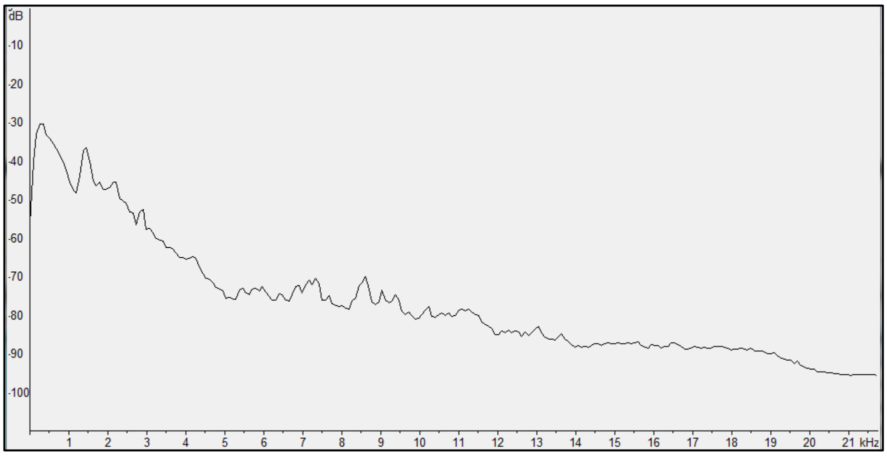


Figure 8.8: Case 1 LTAS

Processing applied

According to the proposed framework, the optimal sequence for enhancement of the issues within this recording is shown in Fig. 8.9

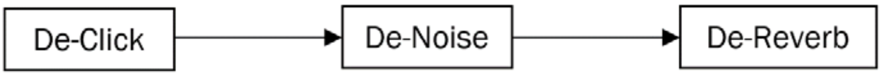
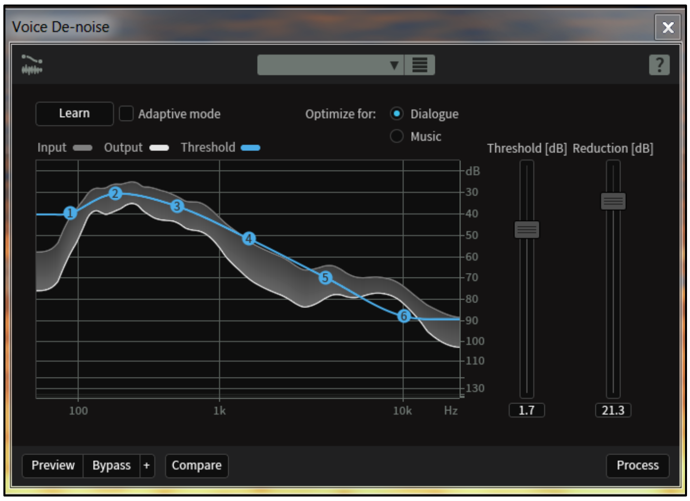


Figure 8.9: Case 1 optimal sequence

De-Click: Remove Clicks.

De-Noise: Remove Broadband Noise. A selection was made in which no speech exists and learn mode was enabled to allow the algorithm to learn the characteristics of the noise present within the recording (Fig. 8.10).



Voice De-noise

□ Adaptive mode

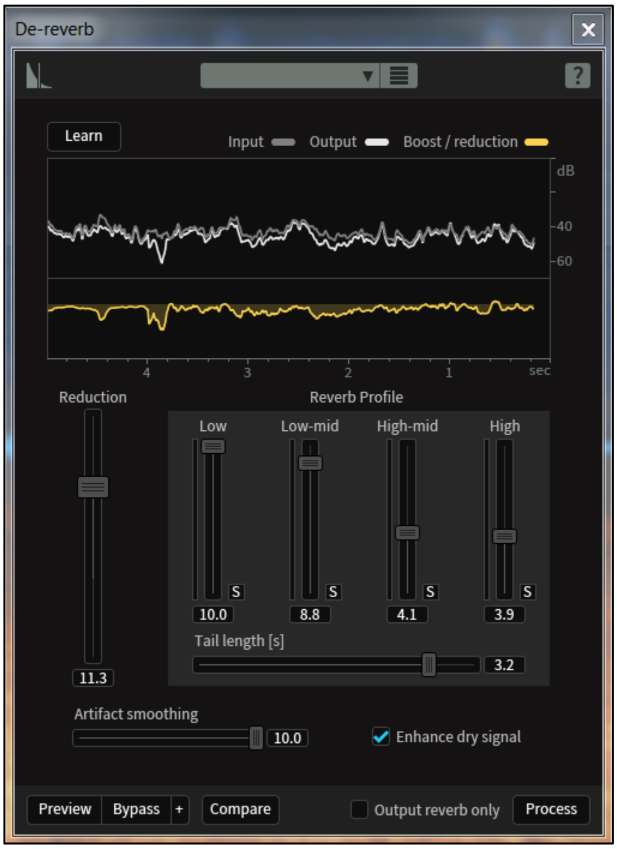
Optimize for: ● Dialogue ○ Music

Threshold [dB] 1.7

Reduction [dB] 21.3

Figure 8.10: Case 1 De-Noise settings

De-Verb: Remove Reverb. Learn mode was applied to the same selection as the noise to allow the algorithm to understand the character of the recording (Fig. 8.11).



De-reverb

Reduction 11.3

Reverb Profile: Low 10.0, Low-mid 8.8, High-mid 4.1, High 3.9

Tail length [s] 3.2

Artifact smoothing 10.0

☑ Enhance dry signal

□ Output reverb only

Figure 8.11: Case 1 de-reverb settings

Files were processed in the sequence in which elements were removed, as on the following page.

Table 8.1: Case 1 processing sequences

Sequence	Process 1	Process 2	Process 3
DC,DN,DR	*De-Click*	*De-Noise*	*De-Reverb*
DC,DR,DN	De-Click	De-Reverb	De-Noise
DR,DN,DC	De-Reverb	De-Noise	De-Click
DR,DC,DN	De-Reverb	De-Click	De-Noise
DN,DC,DR	De-Noise	De-Click	De-Reverb
DN,DR,DC	De-Noise	De-Reverb	De-Click

Results were then obtained through the comparison of results against the clean speech.

Results

Table 8.2: Case 1 PESQ results

Test	Sequence	Result	Rank
PESQ	Reference	4.6439	-
PESQ	Evidence	1.0701	-
PESQ	*DC,DN,DR*	*1.1214*	*1*
PESQ	DC,DR,DN	1.1204	2
PESQ	DN,DC,DR	1.1173	3
PESQ	DN,DR,DC	1.1109	4
PESQ	DR,DC,DN	1.1141	5
PESQ	DR,DN,DC	1.111	6

Table 8.3: Case 1 LLR results

Test	**Sequence**	**Result**	**Rank**
LLR	Reference	0	-
LLR	Evidence	0.9502	-
LLR	*DC,DN,DR*	*0.7686*	*1*
LLR	DN,DR,DC	0.7877	2
LLR	DC,DR,DN	0.7934	3
LLR	DN,DR,DC	0.8202	4
LLR	DR,DC,DN	0.8571	5
LLR	DR,DN,DC	0.8666	6

Table 8.4: Case 1 fw SNR Seg results

Test	Sequence	Result	Rank
fw_SNR_Seg	Reference	35	-
fw_SNR_Seg	Evidence	3.1714	-
fw_SNR_Seg	*DC,DR,DN*	*4.8479*	*1*
fw_SNR_Seg	DR,DN,DC	4.6944	2
fw_SNR_Seg	DR,DC,DN	4.6188	3
fw_SNR_Seg	DN,DC,DR	4.5889	4
fw_SNR_Seg	DC,DN,DR	4.534	5
fw_SNR_Seg	DN,DR,DC	4.2505	6

Table 8.5: Case 1 speech composite results

Test	Sequence	Result	Rank
Composite	Ref. Test	5	-
Composite	Vs. Ev	2.2932	-
Composite	*DC,DN,DR*	*2.5118*	*1*
Composite	DC,DR,DN	2.4954	2
Composite	DN,DC,DR	2.4863	3
Composite	DN,DR,DC	2.4457	4
Composite	DR,DC,DN	2.4207	5
Composite	DR,DN,DC	2.409	6

Table 8.6: Case 1 mean cSII results

Test	Sequence	Result	Rank
CSII_Mean	Reference	1	-
CSII_Mean	Evidence	0.1233	-
CSII_Mean	*DC,DN,DR*	*0.20913*	*1*
CSII_Mean	DC,DR,DN	0.20527	2
CSII_Mean	DN,DC,DR	0.19927	3
CSII_Mean	DR,DC,DN	0.1899	4
CSII_Mean	DN,DR,DC	0.18803	5
CSII_Mean	DR,DN,DC	0.17907	6

Discussion

All tests performed resulted in the evidence signal being of both poorer quality and intelligibility when compared to the clean speech from which it was derived, which is to be expected, and is a good indication that the algorithm is working correctly. All tests also showed that the enhancement process produced higher quality and more intelligible speech than the degraded evidence recording.

The recommended test (PESQ, Table. 8.2) for measuring speech quality concluded with the sequence of click removal, then de-noise, followed by de-reverb, as the optimal chain when processing recordings containing such problems. Both the LLR (Table 8.3) and composite tests (Table 8.5) also resulted in this conclusion, which is consistent with the logical framework suggested within this research. Interesting, an anomaly was presented when processing the audio enhancements through the frequency weighted segmental SNR algorithm (Table 8.4). This may be related to the algorithm basing its final values on signal to noise ratio rather than the quality of the speech, as the optimal result according to that test was to remove clicks, reverb and then noise. It can be argued that some noise was removed during the reverb processing as the algorithm could not determine the difference between reverb and noise, and then the noise was further removed by processing the noise last.

In the PESQ, LLR and composite testing, both the enhancements which removed reverb first were placed at the bottom of the table. This is a clear indication that performing de-reverb before first removing any transient distortions and noise will provide less than optimal enhancement, likely due to the issue of the processing algorithm not being able to determine the difference between noise and reverberant energy. This will result in clean speech being removed unnecessarily.

The results of the speech intelligibility test (Table 8.6) disclosed the sequences which begin with the removal of distortions at the top of the table. This is most likely as in doing this the optimal performance of following processes are optimized to focus on the cleansing of speech without interruption of high amplitude, transient broadband spikes. The removal of reverberation as the first chain of enhancement again placed in the bottom half of the results table.

Based on the results it can, therefore, be concluded that with regards to the removal of clicks, noise, and reverberation, the logical framework provided is correct, especially when seeking to obtain a higher *quality* recording.

Case Two – Clipping, hum, background speech

The second study was created to determine the optimal sequence for the removal of clipping, hum and background speech. A recording, exactly the same in phrasing as the first case study, was used but recorded at a distance of 0 meters from the microphone. Noise was captured from a small gathering of people to form of a 44,100Hz, 16 bit, mono recording and added using Colea [86] at an SNR of 10dB. 50 Hz hum was then taken from a recording and added at an SNR of 0dB. Clipping was finally induced by amplification of the entire recording followed by a holistic reduction in gain.

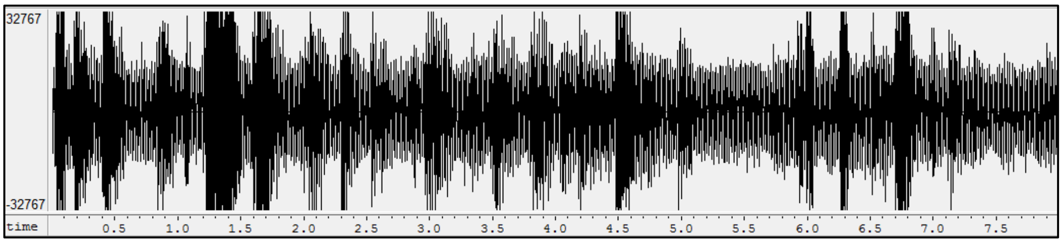


Figure 8.12: Case 2 waveform

Critical listening

Critical listening was first applied, and problems summarized below:

- Clipping causing audible distortion;
- Low-frequency hum masking formant frequencies;
- Broadband background speech competing with the desired speech.

FFT Analysis

- 0dB hum occurs at approximately 30 – 70Hz for the duration of the recording (Fig. 8.13);
- Clear speech extending from 150Hz to 4kHz (Fig. 8.13);
- Spectrum extends from 0 – 20kHz range (Fig. 8.14).

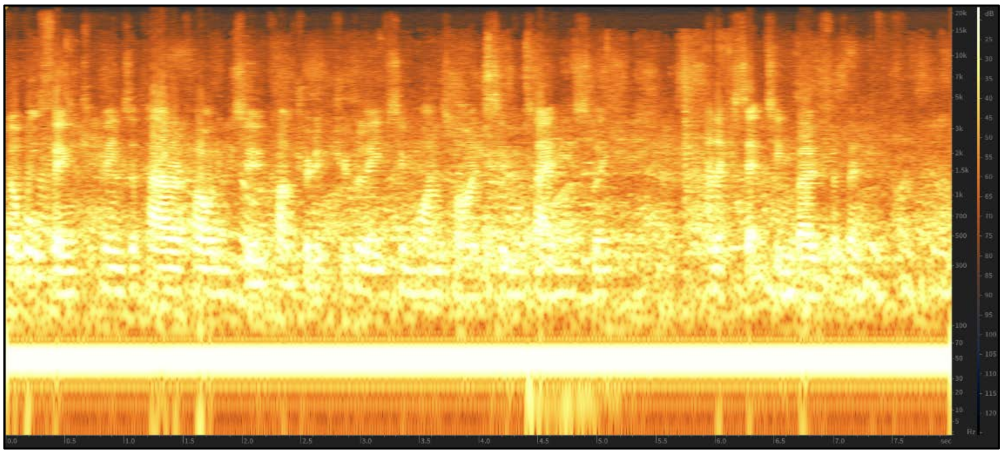


Figure 8.13: Case 2 spectrogram

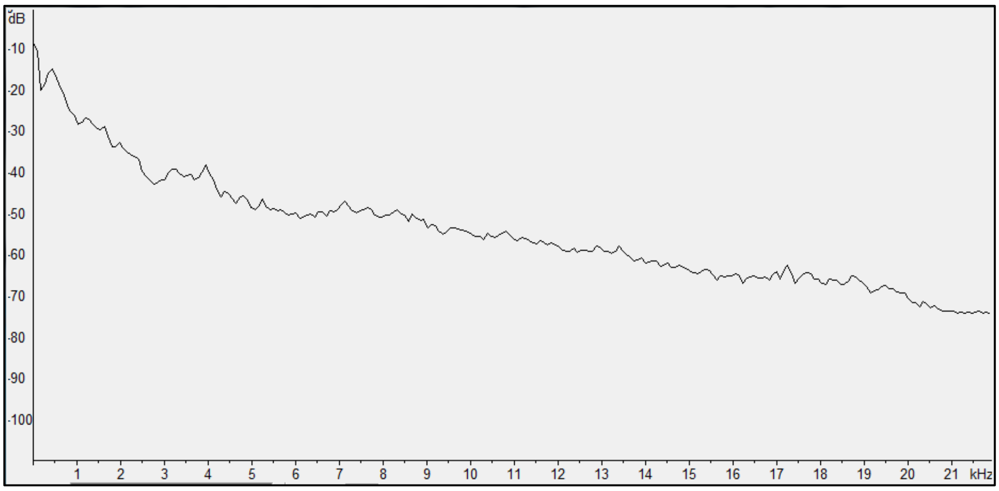


Figure 8.14: Case 2 LTAS

Processing applied

According to the proposed framework, the optimal sequence for enhancement of the issues within this recording is shown in Fig. 8.15

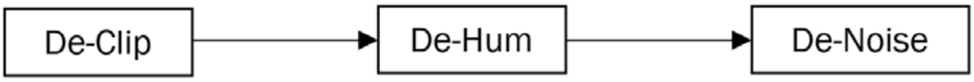
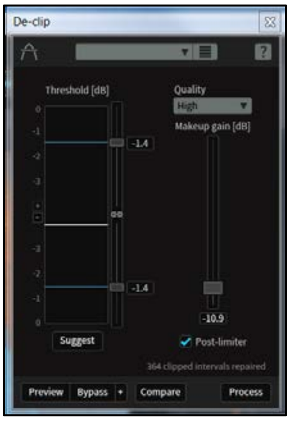


Figure 8.15: Case 2 Optimal sequence

De-Clip: Remove areas of clipping (Fig. 8.16).



De-clip

Threshold [dB] -1.4 / -1.4

Quality High

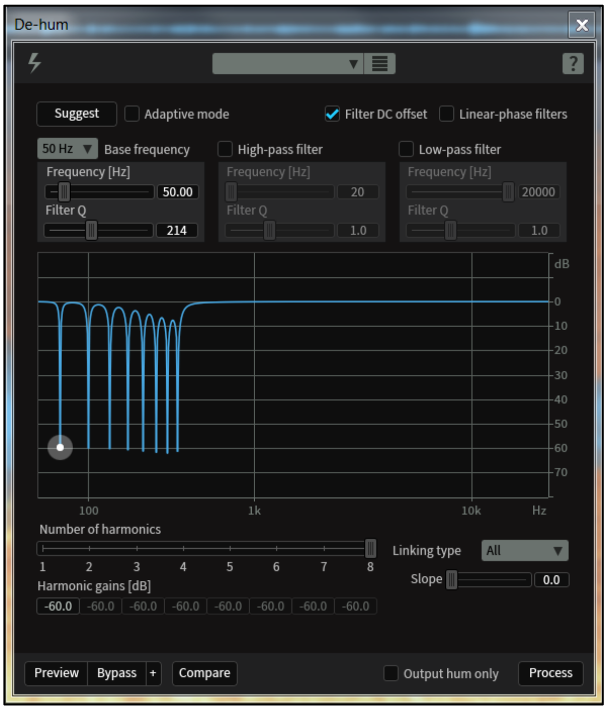
Makeup gain [dB] -10.9

☑ Post-limiter

364 clipped intervals repaired

Figure 8.16: Case 2 de-clip settings

De-Hum: Remove hum. From FFT analysis it was clear that the base frequency of the noise created by hum was at 50Hz, so this was selected within the GUI (Graphical User Interface). Harmonics were included automatically (Fig. 8.17).



De-hum

□ Adaptive mode ☑ Filter DC offset □ Linear-phase filters

50 Hz Base frequency

Frequency [Hz] 50.00

Filter Q 214

□ High-pass filter

Frequency [Hz] 20

Filter Q 1.0

□ Low-pass filter

Frequency [Hz] 20000

Filter Q 1.0

Number of harmonics 8

Linking type All

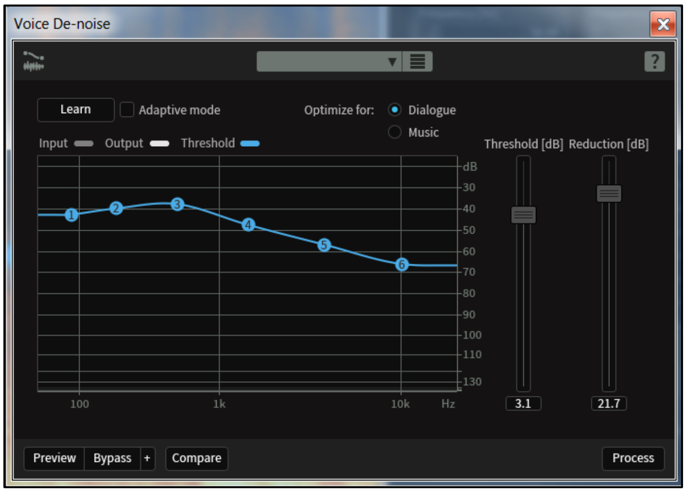
Slope 0.0

Harmonic gains [dB] -60.0 -60.0 -60.0 -60.0 -60.0 -60.0 -60.0 -60.0

□ Output hum only

Figure 8.17: Case 2 de-hum settings

De-Noise: Remove background speech. De-noise specifically designed for speech was used, with learning mode employed on a selection of speech which contained none of the desired signal (Fig. 8.18).



Voice De-noise

Learn □ Adaptive mode

Optimize for: ● Dialogue ○ Music

Threshold [dB] 3.1

Reduction [dB] 21.7

Figure 8.18: Case 2 de-noise settings

Files were processed in the sequence in which elements were removed, as below:

Table 8.7: Case 2 processing sequences

Sequence	Process 1	Process 2	Process 3
DC,DH,DN	*De-Clip*	*De-Hum*	*De-Noise*
DC,DN,DH	De-Clip	De-Noise	De-Hum
DN,DC,DH	De-Noise	De-Clip	De-Hum
DN,DH,DC	De-Noise	De-Hum	De-Clip
DH,DC,DN	De-Hum	De-Clip	De-Noise
DH,DN,DC	De-Hum	De-Noise	De-Clip

Results were then obtained through the comparison of results from the sequential processing and the initial evidence file against the clean speech.

Results

Table 8.8: Case 2 PESQ results

Test	Sequence	Result	Rank
PESQ	Reference	4.6439	-
PESQ	Evidence	1.2598	-
PESQ	*DC,DH,DN*	*1.8301*	*1*
PESQ	DC,DN,DH	1.7752	2
PESQ	DH,DC,DN	1.7663	3
PESQ	DH,DN,DC	1.5848	4
PESQ	DN,DC,DH	1.5678	5
PESQ	DN,DH,DC	1.5678	6

Table 8.9: Case 2 LLR results

Test	Sequence	Result	Rank
LLR	Reference	0	-
LLR	Evidence	0.6613	-
LLR	*DC,DN,DH*	*0.5098*	*1*
LLR	DC,DH,DN	0.5423	2
LLR	DH,DC,DN	0.5649	3
LLR	DN,DC,DH	0.6111	4
LLR	DN,DH,DC	0.6281	5
LLR	DH,DN,DC	0.6677	6

Table 8.10: Case 2 fw SNR seg results

Test	Sequence	Result	Rank
fw_SNR_Seg	Reference	35	-
fw_SNR_Seg	Evidence	6.3735	-
fw_SNR_Seg	*DC,DH,DN*	*7.1895*	*1*
fw_SNR_Seg	DH,DN,DC	7.1407	2
fw_SNR_Seg	DN,DC,DH	7.1324	3
fw_SNR_Seg	DH,DC,DN	7.1267	4
fw_SNR_Seg	DC,DN,DH	7.0934	5
fw_SNR_Seg	DN,DH,DC	7.0707	6

Table 8.11: Case 2 mean cSII results

Test	Sequence	Result	Rank
CSII_Mean	Reference	1	-
CSII_Mean	Evidence	0.4981	-
CSII_Mean	*DN,DC,DH*	*0.4666*	*1*
CSII_Mean	DN,DH,DC	0.4643	2
CSII_Mean	DH,DN,DC	0.4613	3
CSII_Mean	DC,DN,DH	0.4515	4
CSII_Mean	DC,DH,DN	0.4441	5
CSII_Mean	DH,DC,DN	0.4318	6

Table 8.12: Case 2 speech composite results

Test	Sequence	Result	Rank
Composite	Reference	5	-
Composite	Evidence	2.3637	-
Composite	*DC,DH,DN*	*3.2124*	*1*
Composite	DH,DC,DN	3.1396	2
Composite	DC,DN,DH	3.0765	3
Composite	DN,DC,DH	2.9743	4
Composite	DN,DH,DC	2.9339	5
Composite	DH,DN,DC	2.9282	6

Discussion

All quality tests again showed that the enhancement process produced higher quality and more intelligible speech than the evidence recording, but in this case study the intelligibility results showed all enhancements rendered the audio of poorer quality than the evidence signal. This is a clear reminder that quality and intelligibility are not always in correlation with each other and that enhancement processes aren't designed to improve intelligibility. An enhancement of good listening quality could provide poorer transcription results, and an enhancement could also improve intelligibility as a by-product of improving the quality.

The PESQ test (Table 8.8) for determination of speech quality concluded with the sequence of clipping removal, then hum, followed by noise, as the optimal chain when processing recordings containing such problems. Both the frequency weighted segmented SNR (Table 8.10) and desired signal composite tests (Table 8.12) also resulted in this conclusion, which is consistent with the logical framework. The LLR (Table 8.9) resulted in this sequence placing 2nd to the removal of clipping, then noise, and finally hum, which is still a positive result. As hum will be removed with noise if the noise profile captured includes the hum, it is understandable why the process of hum removal and noise removal could be interchangeable in terms of optimal placement. Removal of hum is placed before the removal of noise within the framework as it is a static, bandwidth limited noise which is simple to remove with a comb-filter and allows the optimization of the noise removal algorithm afterwards, as the processing power which would have been used to remove the hum can now be focused on the removal of transient noise.

The removal of clipping as the final process placed at the bottom of many of the quality tests, meaning in doing this the audio will be of poorer listenability. This makes sense in light of the fact that if the noise is causing the clipping, and noise and hum are removed first, then de-clipping will only affect the speech signal, removing speech data which does not need to be removed. This will obviously result in a poorer quality audio signal than if clipping was processed first.

The results of the speech intelligibility test (Table 8.11) show the sequences which begin with the removal of noise at the top of the table. It is likely that this process would remove both the hum and noise as the hum is present throughout, leaving only clipping to have a real effect on the audio. Clipping is generally a phenomenon that has little effect on the intelligibility of the desired speech, more the quality in the form of a compressed type sound where the audio has no values available with which to represent the incoming signal. These intelligibility results are of little consequence considering all enhancements provided intelligibility of lower quality than the original evidence recording.

Based on the results it can be concluded that with regards to the removal of clipping, hum, and noise, the logical framework provided is correct for improving the quality of the recording.

Case Three – Birds chirping, wind, changing speech levels

The third and final study was created to determine the optimal sequence for the removal of transient signals (in this case birds chirping), wind and changes in the level of the desired speech. The same original clean speech recording as case study two was used and changes were made to the level of speech by adding 6dB gain to some words and -6dB to others. Broadband wind was captured and added to the clean speech using Colea [86] at an SNR of 0dB. Chirping birds were finally added at random placements through the recording (Fig. 8.19).

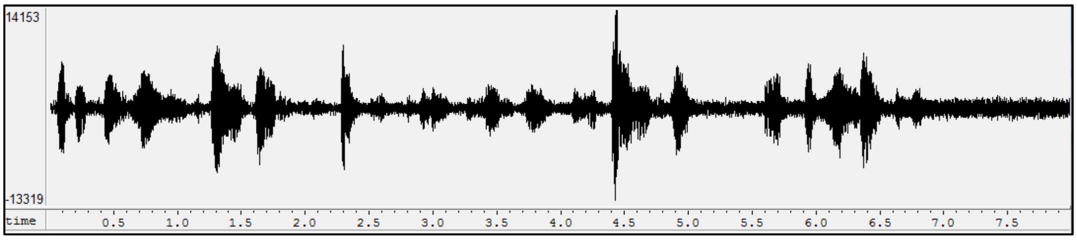


Figure 8.19: Case 3 waveform

Critical listening

The recording was first listened to and problems summarized below:

- Audible broadband noise from wind, masking the speech;
- Audible changes to the level of the desired speaker;
- Bird chirps, which do not affect the intelligibility of the words, but do create a distraction, reducing the quality.

FFT Analysis

- The spectrogram shows the location of the chirps at 0.5s and 6.0s (Fig. 8.20);
- Clear speech extending from 250Hz to 3.5kHz (Fig. 8.20);
- The spectrum extends from 0Hz – 20kHz range (Fig. 8.21);
- LTAS shows -40dB spike occurring at approximately 7kHz (Fig. 8.21). Can be reasoned that this is not speech as it is not in the correct range, and is too high for a non-whistling wind. It must, therefore, be the frequency of the chirping birds.

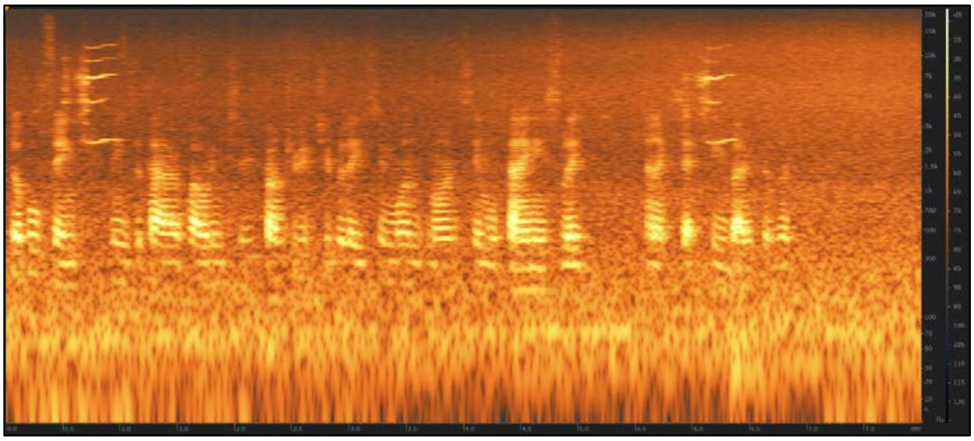


Figure 8.20: Case 3 spectrogram

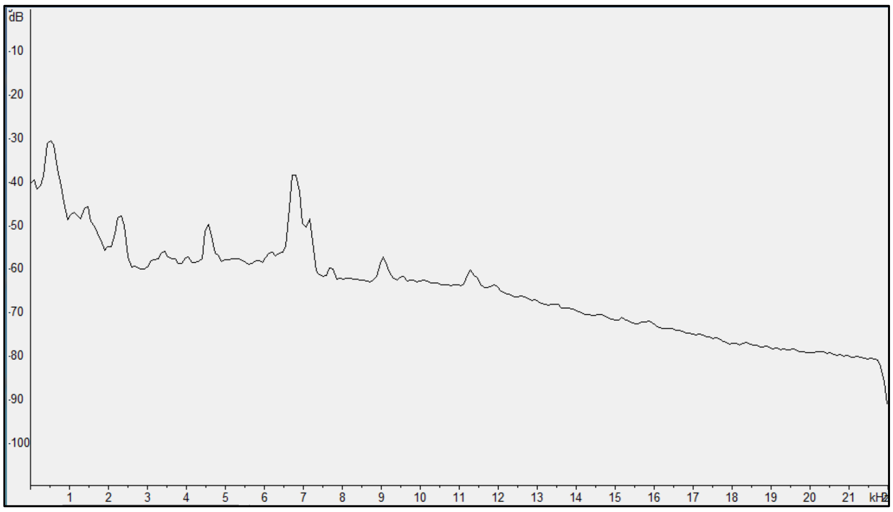


Figure 8.21: Case 3 LTAS

Processing applied

According to the proposed framework, the optimal sequence for enhancement of the issues within this recording is shown in Fig. 8.22

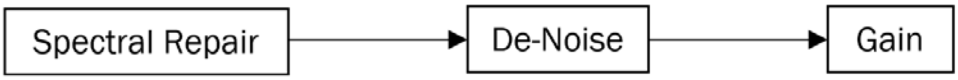
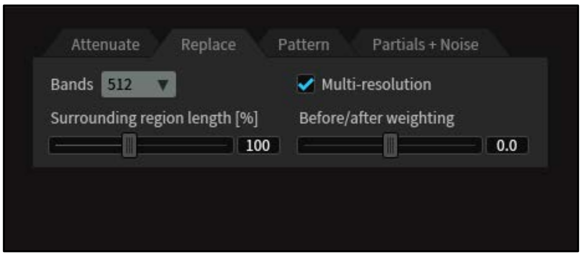


Figure 8.22: Case 3 optimal sequence

Spectral Repair: Remove areas of birds chirping within the spectrogram (Fig. 8.23).



Bands 512

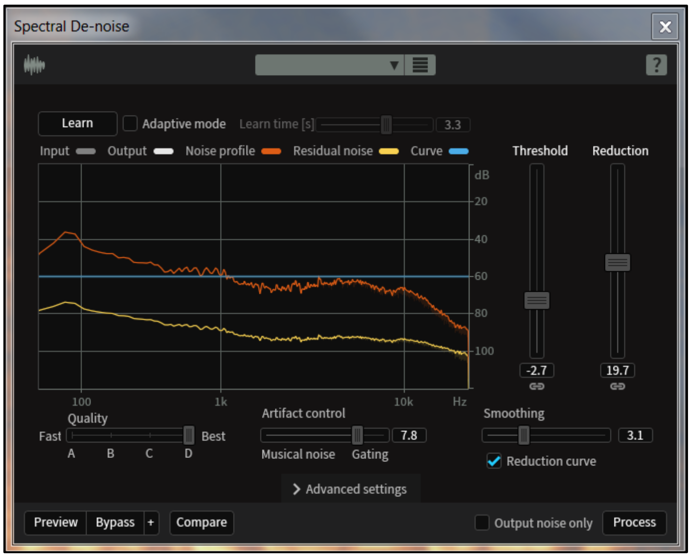
☑ Multi-resolution

Surrounding region length [%] 100

Before/after weighting 0.0

Figure 8.23: Case 3 spectral repair settings

De-Noise: As the wind was present in a large bandwidth extending across the speech, noise removal was employed rather than a static filter. Learning mode was used to capture the profile of the wind noise (Fig. 8.24).



□ Adaptive mode

Learn time [s] 3.3

Threshold -2.7

Reduction 19.7

Quality: Fast A B C D Best

Artifact control 7.8 (Musical noise – Gating)

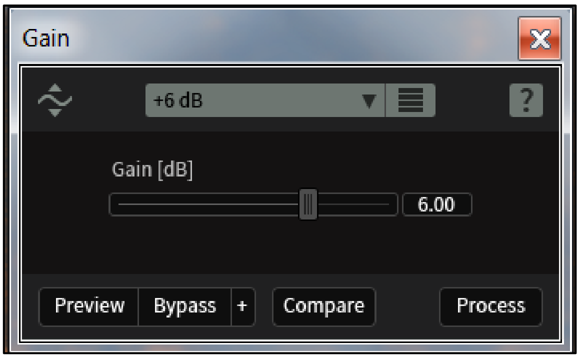
Smoothing 3.1

☑ Reduction curve

□ Output noise only

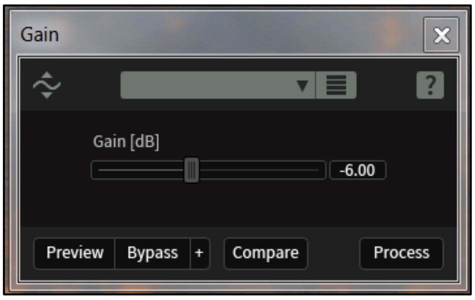
Figure 8.24: Case 3 de-noise settings

Manual Gain: Rather than the use of compression or expansion, manual gain provides a more focused and concise approach, especially useful in shorter audio recordings. Gain was applied at +6dB where the desired signal was low and -6dB to areas which required attenuation. Reduction should always be used instead of gain as to not increase the level of noise present in the recording but is at times necessary if the speech is of extremely low levels in comparison to the average level of the speech present (Fig. 8.25, 8.26).



Gain [dB] 6.00

Figure 8.25: Case 3 gain settings



Gain [dB] -6.00

Figure 8.26: Case 4 attenuation settings

Files were processed in the sequence in which elements were processed, as below:

Table 8.13: Case 3 processing sequences

Sequence	Process 1	Process 2	Process 3
SR,DN,FL	*Spectral repair*	*De-Noise*	*Fix levels*
SR,FL,DN	Spectral repair	Fix levels	De-Noise
DN,FL,SR	De-Noise	Fix levels	Spectral repair
DN,SR,FL	De-Noise	Spectral repair	Fix levels
FL,SR,DN	Fix levels	Spectral repair	De-Noise
FL,DN,SR	Fix levels	De-Noise	Spectral repair

Results were then obtained through the comparison of results from the sequential processing and the initial evidence file against the clean speech.

Results

Table 8.14: Case 3 PESQ results

Test	Sequence	Result	Rank
PESQ	Reference	4.6439	-
PESQ	Evidence	1.2471	-
PESQ	*SR,DN,FL*	*2.1283*	*1*
PESQ	DN,SR,FL	2.1066	2
PESQ	DN,FL,SR	2.1039	3
PESQ	FL,SR,DN	2.0855	4
PESQ	SR,FL,DN	2.0606	5
PESQ	FL,DN,SR	2.036	6

Table 8.15: Case 3 LLR results

Test	Sequence	Result	Rank
LLR	Reference	0	-
LLR	Evidence	1.1755	-
LLR	*SR,DN,FL*	*0.7104*	*1*
LLR	SR,FL,DN	0.7195	2
LLR	FL,SR,DN	0.7312	3
LLR	DN,FL,SR	0.7415	4
LLR	FL,DN,SR	0.7576	5
LLR	DN,SR,FL	0.7576	6

Table 8.16: Case 3 fw SNR seg results

Test	Sequence	Result	Rank
fw_SNR_Seg	Reference	35	-
fw_SNR_Seg	Evidence	8.2269	-
fw_SNR_Seg	*SR,DN,FL*	*10.9843*	*1*
fw_SNR_Seg	DN,SR,FL	10.9713	2
fw_SNR_Seg	DN,FL,SR	10.9316	3
fw_SNR_Seg	SR,FL,DN	10.5609	4
fw_SNR_Seg	FL,DN,SR	10.5478	5
fw_SNR_Seg	FL,SR,DN	10.5384	6

Table 8.17: Case 4 mean cSII results

Test	Sequence	Result	Rank
CSII_Mean	Reference	1	-
CSII_Mean	Evidence	0.5412	-
CSII_Mean	*FL,SR,DN*	*0.8084*	*1*
CSII_Mean	SR,FL,DN	0.6503	2
CSII_Mean	FL,DN,SR	0.6495	3
CSII_Mean	SR,DN,FL	0.6488	4
CSII_Mean	DN,SR,FL	0.6477	5
CSII_Mean	DN,FL,SR	0.6455	6

Table 8.18: Case 3 speech composite results

Test	Sequence	Result	Rank
Composite	Reference	5	-
Composite	Evidence	2.1945	-
Composite	*SR,DN,FL*	*3.2958*	*1*
Composite	DN,FL,SR	3.2491	2
Composite	FL,SR,DN	3.2481	3
Composite	SR,FL,DN	3.2457	4
Composite	DN,SR,FL	3.2316	5
Composite	FL,DN,SR	3.1927	6

Discussion

All tests showed that the enhancement process produced higher quality and more intelligible speech than the evidence recording.

All tests for determination of speech quality (Table 8.14, 8.16, 8.16, 8.18) concluded with the sequence of spectral repair, then noise removal, followed by the matching of speech levels, which is consistent with the logical framework. The opposite sequence to this (fix levels, remove noise and then perform spectral repair) performed consistently poorly on all quality tests, proving that fixing speech levels too early in the process will also increase noise levels (if applying gain), or remove speech data (if applying attenuation).

The results of the speech intelligibility test (Table 8.17) showed the sequence which performed worst in improving quality performed best for intelligibility. It is possible that changing the levels before removing noise may cause increased intelligibility even though the level of noise has increased with any gain applied. Again, this highlights the difference in quality and intelligibility of recordings.

Based on the results it can be concluded that when applying spectral repair, removing broadband noise such as wind and balancing levels of speech, the logical framework provided is correct.

CHAPTER IX

CONCLUSION

The enhancement of audio recordings for forensic purposes is a discipline that should not be taken lightly, due to both the effect results can have on individual lives within the criminal justice and civil litigation systems, but also due to the complexity of the process itself. For these reasons it is essential that decisions related to the process should be deeply considered before any action is taken, as the effects of a poor choices are irreversible once the final product has been delivered. Not only should the type of processing, amount of processing and area of a recording which is to be affected by the processing be taken into account, but as discovered in this research, the order in which the processing is applied must also be carefully measured. To this end, audio enhancement can be seen as a balance between many fundamentals such as noise and the desired signal, quality and intelligibility, and the level and order of the processes.

To perform audio enhancement not only should an individual recognize what the tools available to them do, they must also understand how these tools are manipulating the signal. Only in doing this does it allow the individual to create optimal enhancements, which is vital when dealing with poor quality recordings of which every little counts towards the final product. If the reason for applying an enhancement process as well as the inner workings of that process are known, then the logical framework can be used as a reference document in which the order is not fixed, but malleable dependent on the audio at hand. If every process is applied without being necessary or understood, there is a good chance of the result being worse than the original evidence. Training is therefore essential, as is continuing education through the reading of scientific journals and attending conferences. The research presented in this study should be supplemented through education of these forms to ensure that a deep understanding of when and how to deviate from the framework is second nature to the examiner. Processing of audio for enhancement can often involve retracing certain steps, for example applying gain at 2 separate stages, or applying multiple instances of one process with different settings to achieve the best results from a certain process.

Rather than considering the benefits of applying a certain tool, it is important that the negative aspects are weighted with the same considerations. Each and every process applied adds quantization noise, so if the process is not required, it is not enhancing the audio, but destroying it. For this reason, it is always best to urge on the side of caution. It is much better to not lose any vital elements of the desired signal and keep the noise than to lose parts of the desired signal when removing the noise.

The framework proposed within this research was tested against recordings with issues that are found within real forensic casework. The results showed that the framework is correct and that the scientific logic transcends into the real-world for its application to audio enhancement within federal, local and private laboratories around the world.

Although already known from previous studies, but not widely tested, a key finding from the case studies of the research related to the difference between the quality and the intelligibility of a recording. The framework was based on scientific logic with the aim of enhancing an audio recording through the removal of noise and distortions (therefore increasing the quality), which may provide increased intelligibility as a by-product. Findings from the case studies show that an enhancement framework does not improve intelligibility, as unexpected sequences outperformed the proposed frameworks order of operations. The fact that intelligibility is not improved through audio enhancement is re-enforced through these findings, although it is possible that in some cases the aural/human/subjective intelligibility may be increased, but not the mathematical one.

Future work could seek to develop an algorithm which can provide objective results, correlated to subjective results, without the need for the original clean signal. This would allow analysts to then test processes in real time to help guide their subjective decisions as to whether the processing they are applying is improving or degrading the recording.

Enhancement of forensic audio is a complex topic in which each decision affects the final result of the recording, and in turn, could affect the final result of a criminal case. It is therefore vital that all measures available are implemented in helping to make the correct decisions at each fork in the road, as optimal enhancement means choosing the correct path every single time, whereas a poor-quality enhancement can result from making just one wrong decision. Analogous to a roadmap, it is hoped that this research serves as a reference which can be studied at every turn, increasing the chances of consistently making the correct decision. But remember, that some roads are bumpier than others and roadworks can occur at any time, so deviating from the map is allowed. But only if that decision is informed, considered and necessary.

BIBLIOGRAPHY

[1] L. Singh and S. Sridharan, “Speech enhancement using critical band spectral subtraction.,” in *ICSLP*, 1998.

[2] K. Naruka and O. P. Sahu, “Objective Quality and Intelligibility Evaluation for Speech Enhancement Algorithms.”

[3] P. C. Loizou and G. Kim, “Reasons why Current Speech-Enhancement Algorithms do not Improve Speech Intelligibility and Suggested Solutions,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 1, pp. 47–56, Jan. 2011.

[4] B. E. Koenig, “Enhancement of forensic audio recordings,” *J. Audio Eng. Soc.*, vol. 36, no. 11, pp. 884–894, 1988.

[5] D. A. Bronstein, “Law for the Expert Witness,” in *Law for the Expert Witness*, FL: CRC Press, 2012, p. 86.

[6] P. Manchester, “An Introduction To Forensic Audio,” *Sound on Sound*, Jan-2010.

[7] B. E. Koenig, D. S. Lacey, and S. A. Killion, “Forensic enhancement of digital audio recordings,” *J. Audio Eng. Soc.*, vol. 55, no. 5, pp. 352–371, 2007.

[8] R. C. Maher, “Audio Enancement using Nonlinear Time-Frequency Filtering,” in *Audio Engineering Society Conference: 26th International Conference: Audio Forensics in the Digital Age*, 2005.

[9] ENFSI, “Guidelines for Best Practice in the Forensic Examination of Digital Media.” Apr-2009.

[10] SWGDE, “Digital and Multimedia Evidence Glossary Version 3.0.” 20-Jun-2016.

[11] International Standards Organisation, “ISO/IEC 27037.” 2012.

[12] International Standards Organisation, “ISO/IEC 17025.” 2005.

[13] International Standards Organisation, “ISO 9001.” 2008.

[14] International Standards Organisation, “ISO/IEC 27042.” 2015.

[15] SWGDE, “Best Practices for Forensic Audio.” 08-Oct-2016.

[16] SWGDE, “Core Competencies for Digital Audio.” 15-Sep-2011.

[17] Francis Rumsey and Tim McCornmick, *Sound and Recording*, Fifth Edition. London: Elsevier Ltd, 2006.

[18] International Standards Organisation, “ISO 226.” 2003.

[19] Ken C. Pohlmann, *Principles of Digital Audio*, 5th Editon. New York: McGraw-Hill, 2005.

[20] Eberhard Zwicker and Hugo Fastl, *Pscychoacoustics, Facts and Models*. Berlin: Springer-Verlag, 1990.

[21] E. Ambikairajah, A. G. Davis, and W. T. K. Wong, “Auditory masking and MPEG-1 audio compression,” *Electron. Commun. Eng. J.*, vol. 9, no. 4, pp. 165–175, 1997.

[22] David Miles Huber and Rober E. Runstein, *Modern Recording Techniques*, Sixth Edition. UK: Elsevier Inc, 2005.

[23] Micheal Talbot-Smith, *Sound Engineering Explained*, 2nd Edition. Oxford, UK: Focal Press, 2005.

[24] Dario Brandt, “How to Find the Sample Rate and Bit Depth of an Audio File,” *Ear Monk*, 2015. Retrieved Oct 1, 2017, from the EarMonk website: http://earmonk.com/find-sample-rate-bit-depth-audio-file/

[25] Ian McLoughlin, *Applied Speech and Audio Processing*. United Kingdom: Cambridge University Press, 2009.

[26] D. Luknowsky and J. Boyczuk, “Audio processing in police investigations,” *Can. Acoust.*, vol. 32, no. 3, pp. 154–155, 2004.

[27] C Grigoras and JM Smith, “Audio Enhancement and Authentication,” in *Encyclopedia of Forensic Sciences*, Second Edition., Elsevier Ltd, 2013, pp. 315–326.

[28] Harry Hollien, *The Acoustics of Crime*. New York and London: Plenum Press, 1990.

[29] Bruce Koenig and Catalin Grigoras, “Digital Audio Authentication Workshop,” presented at the 2017 AES Audio Forensics Conference, Arlington, VA, Jun-2017.

[30] D. Bergfeld and K. Junte, “The Effects of Peripheral Stimuli and Equipment Used on Speech Intelligibility in Noise,” in *Audio Engineering Society Conference: 2017 AES International Conference on Audio Forensics*, 2017.

[31] R. Maher, “Overview of audio forensics,” *Intell. Multimed. Anal. Secur. Appl.*, pp. 127–144, 2010.

[32] FBI, “Equipping the Modern Audio-Video Forensic Laboratory,” *Forensic Sci. Commun.*, vol. 5, Apr. 2003.

[33] Anthony T.S Ho and Shujun Li, *Handbook of Digital Forensics and Multimedia Data Devices*. UK: John Wiley & Sons, Ltd, 2015.

[34] Catalin Grigoras, “Forensic Audio Analysis Introduction,” presented at the MSc Recording Arts with a focus on Media Forensics, NCMF, University of Colorado Denver, Fall-2016.

[35] K. Christman, “An Objective Method of Measuring Subjective Click-and-Pop Performance for Audio Amplifiers.” *in Audio Engineering Society Conference: 125th Convention,* 2008

[36] Simon J. Godsill, “Digital Audio Restoration - a statistical model based apporach.” CEDAR, 21-Sep-1998.

[37] D. Betts, G. Reid, and D. Chan, “Application of Digital Signal Processing to Audio Restoration,” in *Audio Engineering Society Conference: UK 7th Conference: Digital Signal Processing (DSP)*, 1992.

[38] GoofyDawg, “Overdrive vs. Distortion,” *Guitar Gear*, 05-Nov-2009. Retrieved Oct 29, 2017, from the GuitarGear website: https://guitargear.org/2009/11/05/overdrive-vs-distortion/

[39] Bob Katz, *Mastering Audio: The Science and the Art*. Focal Press, 2002.

[40] L. Claesson, “Making Audio Sound Better One Square Wave at a Time (Or How an Algorithm Called ‘Undo’ Fixes Audio),” in *Audio Engineering Society Convention 137*, 2014.

[41] B. E. Koenig and D. S. Lacey, “Evaluation of clipped-sample restoration software,” *Forensic Sci. Commun.*, vol. 12, no. 2, p. N_A, 2010.

[42] SoundBlade, “Spectral Repair Tool User Manaul.” Retrieved Oct 5, 2017, from the SonicStudio website: http://help.sonicstudio.com/srt/srt_um.html)

[43] N. Murata, S. Ikeda, and A. Ziehe, “An approach to blind source separation based on temporal structure of speech signals,” *Neurocomputing*, vol. 41, no. 1, pp. 1–24, 2001.

[44] M. G. Jafari and M. D. Plumbley, “Convolutive blind source separation of speech signals in the low frequency bands,” in *Audio Engineering Society Convention 123*, 2007.

[45] Shoko Araki, Shoji Makino, Ryo Mukai, Tsuyoki Nishikawa, and Hiroshi Saruwatari, “Fundamenal limitation of frequency domain blind source separation for convolved mixture of speech,” in *2001 IEEE International Conference*, 2001, vol. 5.

[46] T.-W. Lee, M. S. Lewicki, M. Girolami, and T. J. Sejnowski, “Blind source separation of more sources than mixtures using overcomplete representations,” *IEEE Signal Process. Lett.*, vol. 6, no. 4, pp. 87–90, 1999.

[47] Anil Alexander and Oscar Forth, “‘No, thank you, for the music’: An application of audio fingerprinting and automatic music signal cancellation for forensic audio enhancement,” 2011.

[48] P. Ignatov, M. Stolbov, and S. Aleinik, “Semi-Automated Technique for Noisy Recording Enhancement Using an Independent Reference Recording,” in *Audio Engineering Society Conference: 46th International Conference: Audio Forensics*, 2012.

[49] Shazam Entertainment, *Shazam*. Retrieved 28 Sept, 2017 from Shazam website: https://www.shazam.com/

[50] Dan Ellis, “Robust Landmark-Based Audio Fingerprinting,” 2009. Retrieved Oct 21, 2017, from the Labrosa website: https://labrosa.ee.columbia.edu/matlab/fingerprint/

[51] A. Alexander, O. Forth, and D. Tunstall, “Music and noise fingerprinting and reference cancellation applied to forensic audio enhancement,” in *Audio engineering society conference: 46th international conference: audio forensics*, 2012.

[52] D. Betts, A. French, C. Hicks, and G. Reid, “The Role of Adaptive Filtering in Audio Surveillance,” in *Audio Engineering Society Conference: 26th International Conference: Audio Forensics in the Digital Age*, 2005.

[53] Roey Izhaki, *Mixing Audio: Concepts, Practices and Tools*. Oxford: Elsevier Ltd, 2008.

[54] Izotope, “How to Use Dynamic EQ in Mastering.” Retrieved 26 Sept, 2017 from Izotope website: https://www.izotope.com/en/community/blog/tips-tutorials/2014/11/how-to-use-dynamic-eq-in-mastering.html

[55] J. Zjalic, C. Grigoras, and J. Smith, “A Low Cost, Cloud Based, Portable, Remote ENF System,” in *Audio Engineering Society Conference: 2017 AES International Conference on Audio Forensics*, 2017.

[56] C. Grigoras, “Applications of ENF analysis method in forensic authentication of digital audio and video recordings,” in *Audio Engineering Society Convention 123*, 2007.

[57] E. B. Brixen and R. Hensen, “Wind Generated Noise in Microphones-An Overview-Part 1,” in *Audio Engineering Society Convention 120*, 2006.

[58] M. Stolbov and P. Ignatov, “Speech Enhancement Technique for Low SNR Recording,” in *Audio Engineering Society Conference: 46th International Conference: Audio Forensics*, 2012.

[59] Steven F. Boll, “Suppression of Acoustic Noise in Speech using Spectral Subtraction,” *IEEE Trans. Acoust. Speech Signal Process.*, Apr. 1979.

[60] L. Singh and S. Sridharan, “Speech enhancement using critical band spectral subtraction,” in *Fifth International Conference on Spoken Language Processing*, 1998.

[61] S. Kamath and P. Loizou, “A multi-band spectral subtraction method for enhancing speech corrupted by colored noise.,” in *ICASSP*, 2002, vol. 4, pp. 44164–44164.

[62] J. Poruba, “EXPERIMENTS WITH DYNAMIC SPECTRAL SUBTRACTION USING THE HUMAN EAR MASKING CHARACTERISTICS.”

[63] A. Lukin and J. Todd, “Suppression of musical noise artifacts in audio noise reduction by adaptive 2-D filtering,” in *Audio Engineering Society Convention 123*, 2007.

[64] Jacob Benesty, Jingdog Chen, Yiteng Huang, and Israel Cohen, *Noise Reduction in Speech Processing*, vol. 2. Springer, 2009.

[65] S. Hoare, P. Hughes, and R. Turnbull, “Audio Enhancement for Portable Device Based Speech Applications,” in *Audio Engineering Society Convention 124*, 2008.

[66] Rich Williams, “What an audio limiter does in the studio,” *Practical Music Production*.

[67] D. R. Cole, M. P. Moody, and S. Sridharan, “Robust Enhancement of Reverberant Speech,” in *Audio Engineering Society Convention 5r*, 1995.

[68] Y. Mahieux and C. Marro, “Comparison of dereverberation techniques for videoconferencing applications,” in *Audio Engineering Society Convention 100*, 1996.

[69] A. E. S. Staff, “Reverberation and Dereverberation,” *J. Audio Eng. Soc.*, vol. 55, no. 3, pp. 189–194, 2007.

[70] E. A. Habets, “Single-channel speech dereverberation based on spectral subtraction,” in *Proceedings of the 15th Annual Workshop on Circuits, Systems and Signal Processing (ProRISC)*, 2004.

[71] K. Lebart, J.-M. Boucher, and P. N. Denbigh, “A new method based on spectral subtraction for speech dereverberation,” *Acta Acust. United Acust.*, vol. 87, no. 3, pp. 359–366, 2001.

[72] “How to Normalise Audio - Why do it?,” *Learn Digital Audio*. Retrieved 1 Nov, 2017 from Learn Digital Audio Website: http://www.learndigitalaudio.com/normalize-audio

[74] Spencer Ledesma, “A Proposed Frame for image enhancement,” University of Colorado, Denver, 2015.

[75] Jeff Smith and Catalin Grigoras, “Forensic Audio Enhancement Section 2,” National Center for Media Forensics, 2017.

[76] Dave Shackleford, “SHA-1 Has Been Broken: Now What?,” *IANS*, 24-Feb-2017.

[77] B. C. Bispo *et al.*, “EW-PESQ: a quality assessment method for speech signals sampled at 48 kHz,” *J. Audio Eng. Soc.*, vol. 58, no. 4, pp. 251–268, 2010.

[78] A. W. Rix, M. P. Hollier, J. G. Beerends, and A. P. Hekstra, “PESQ-the new ITU standard for end-to-end speech quality assessment,” in *Audio Engineering Society Convention 109*, 2000.

[79] K. Kinoshita *et al.*, “The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech,” in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*, 2013, pp. 1–4.

[80] A. A. Kressner, D. V. Anderson, and C. J. Rozell, “Robustness of the hearing aid speech quality index (HASQI),” in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, 2011, pp. 209–212.

[81] Philipos c. Loizou, *Speech Enhancement*, 2nd Edition. CRC Press, 2013.

[82] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, and U. Kjems, “An evaluation of objective quality measures for speech intelligibility prediction,” in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[83] N. Harlander, R. Huber, and S. D. Ewert, “Sound quality assessment using auditory models,” *J. Audio Eng. Soc.*, vol. 62, no. 5, pp. 324–336, 2014.

[84] Y. Hu and P. C. Loizou, “Evaluation of objective quality measures for speech enhancement,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 1, pp. 229–238, 2008.

[85] George Orwell, *Nineteen Eighty-Four*. 1949.

[86] Philip Loizou, *COLEA*. Retrieved 4 Nov, 2017 from UT Dallas Website: http://ecs.utdallas.edu/loizou/speech/colea.htm