Intelligent Tools for Multitrack Frequency and Dynamics Processing

Zheng Ma

PhD thesis

School of Electronic Engineering and Computer Science
Queen Mary University of London

2016

Submitted to University of London in partial fulfilment of the requirements for the degree of Doctor of Philosophy
Queen Mary University of London
2016

Statement of Originality

I, Zheng Ma, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signature:

Date:

Abstract

This research explores the possibility of reproducing mixing decisions of a skilled audio engineer with minimal human interaction that can improve the overall listening experience of musical mixtures, i.e., intelligent mixing. By producing a balanced mix automatically musician and mixing engineering can focus on their creativity while the productivity of music production is increased. We focus on the two essential aspects of such a system, frequency and dynamics. This thesis presents an intelligent strategy for multitrack frequency and dynamics processing that exploit the interdependence of input audio features, incorporates best practices in audio engineering, and driven by perceptual models and subjective criteria.

The intelligent frequency processing research begins with a spectral characteristic analysis of commercial recordings, where we discover a consistent leaning towards a target equalization spectrum. A novel approach for automatically equalizing audio signals towards the observed target spectrum is then described and evaluated. We proceed to dynamics processing, and introduce an intelligent multitrack dynamic range compression algorithm, in which various audio features are proposed and validated to better describe the transient nature and spectral content of the signals. An experiment to investigate the human preference on dynamic processing is described to inform our choices of parameter automations. To provide a perceptual basis for the intelligent system, we evaluate existing perceptual models, and propose several masking metrics to quantify the masking behaviour within the multitrack mixture. Ultimately, we integrate previous research on auditory masking, frequency and dynamics processing, into one intelligent system of mix optimization that replicates the iterative process of human mixing. Within the system, we explore the relationship between equalization and dynamics processing, and propose a general frequency and dynamics processing framework. Various implementations of the intelligent system are explored and evaluated objectively and subjectively through listening experiments.

Acknowledgements

This work was supported by the China Scholarship Council. I'd like to thank everyone in the Centre for Digital Music at Queen Mary University of London. In particular my supervisor Josh Reiss, whose wide-ranging knowledge and enthusiasm has been invaluable in every aspect of my PhD work. A big thank to all my dear friends for their support. Finally, to mum, dad and sister, thank you. None of these could have been possible without you. I love you.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Music mixing is a process in which multitrack material is balanced, treated and combined into a multichannel format, most commonly two-channel stereo or single channel mono (Izhaki, 2013). Mixing is often regarded as a creative art form. However, mixing entails technical aspects too. Achieving balance in frequency and dynamics domains remains the most challenging, technical task, which requires adequate knowledge in acoustics, signal processing and years of practice. In fact, much of the initial, non-artistic mixing work follows established rules and best practices (Reiss, 2011). Some modern audio production tools are able to apply pre-sets to the signal. However they lack the ability to make intelligent mixing decisions (Reiss, 2011). The complexity of the software interface and mixing desk often discourage non-experts too.

On the other hand, nowadays amateur or bedroom musicians can create music using digital production tools with an access to a laptop. However, a mixing engineer is still needed in order to produce a well-balanced mix. Having a mixing engineer behind the mixing desk is essential to live performance due to problems such as feedback, imbalance, room resonances and poor equipment. Unfortunately, it is not always affordable especially for small venues. (Reiss & De Man, 2013)

To address these requirements, the concept of intelligent multitrack mixing systems is proposed (Moorer, 2000). The word “intelligent” suggests that such a system must be able to analyze the signals, dynamically adapt to audio signals, automatically derive mixing parameters based on best practices, subjective evaluation and perceptual criteria (Reiss, 2011). With the intelligent mixing system, musician and mixing engineering can focus on their

creativity while the productivity of music production is increased, and smaller music venues are free of hiring professional mixing engineers. (Reiss & De Man, 2013).

1.2 Scope of the Research

The thesis contributes to the field of intelligent mixing with a focus on frequency and dynamics aspects. Frequency equalization and dynamics processing dominate exclusive domains. Equalization influences amplitude in the spectral domain, while dynamics processing influences amplitude in the time domain. However, the operational nature of the two processors gives insight into a manner in which they may be combined into a general frequency and a dynamic processing framework. Such a general tool can act as an inclusive superset of an equalizer and dynamic processor, where the functionality of the two disparate processors is intuitively combined yet their standalone versatility is retained (Wise, 2009). It creates a larger control space and more detailed adjustments to the audio environment, providing invaluable advantages in intelligent mixing.

Furthermore, high levels of cognition dictate the way in which sound is perceived. For a true intelligent mixing system to prevail, it is rational to hypothesize that a signal analysis chain that considers properties of the hearing system would be beneficial. Therefore this thesis is also targeted towards a perceptual understanding of the mixing process and harnessing this understanding to optimize the auditory experience of the musical mixture.

Perceptual models establish a bridge between the objective physical domain and the subjective domain of human hearing. When equipped with auditory models capable of predicting psychoacoustic phenomena, an opportunity arises in which one can investigate auditory aspects of music production and employ them to perform automatic mixing operations that are influenced by perception, such as equalization and dynamics compression.

1.3 Aim and Objectives

The aim of this thesis is to develop a novel intelligent system for multitrack frequency and dynamics processing, exploiting the interdependence of the input audio features,

incorporating best practices in audio engineering, and driven by perceptual models and subjective criteria. This will be achieved by fulfilling the following objectives:

- Investigate spectral characteristics of the musical mixtures.
- Propose and evaluate intelligent mixing strategies for frequency manipulation to achieve spectral balance.
- Investigate audio features to describe the dynamic behaviour of musical signals.
- Develop and evaluate intelligent multitrack dynamic range compression algorithms.
- Evaluate existing computational hearing models in order to propose and apply perceptual models pertaining to properties that are fundamental to the context of mixing multitrack audio, such as auditory masking.
- Integrate previous findings in perception studies, frequency and dynamics processing into an intelligent system for mix optimization, and evaluate the system performance.

1.4 Thesis Structure

- **Chapter 2** presents the background upon which this thesis will be developed. The physiology of the human hearing system is discussed, with an emphasis on the concepts of masking, critical bands and auditory filters. Several psychoacoustics-inspired loudness and masking models are reviewed as the perceptual criteria basis of our intelligent mixing studies. The process of multitrack mixing with a focus on frequency and dynamics aspects was discussed. This chapter is concluded by reflecting upon how the state of the art in the field of intelligent mixing bears upon our choice of approaches.
- **Chapter 3** investigates the frequency aspect of intelligent mixing. A spectral characteristic analysis of popular commercial recordings is presented first. A consistent leaning towards a target equalization spectrum that stems from practices in the music industry is discovered. A new approach for automatically equalizing audio signals towards the observed target spectrum is then described and evaluated.
- **Chapter 4** investigates the dynamics aspect of intelligent mixing. A fully automated multitrack dynamic range compression algorithm is introduced, in which we investigate and propose various audio features to better describe the transient nature and spectral content of the signals. A method of adjustment experiment is described to investigate the relationship between human preference for ratio and threshold.

The results of this inform the choices for our intelligent algorithms. Subjective evaluation of the system is presented in the form of a multiple stimulus listening test. And lastly a personalized compressor, which can adapt to the real-time noise environment, is presented.

- **Chapter 5** focuses on the studies of masking in multitrack audio, offering a perceptual understanding of the mixing process. An equal loudness matching experiment is first described to evaluate the performance of existing loudness model on musical signals. Parameter modification of the loudness model that yields better compliance with the human perception of masking is proposed. The outcome of this experiment is then integrated into the development of several psychoacoustics-inspired, cross-adaptive multitrack masking metrics to describe the masking behaviour within the musical mixture.
- **Chapter 6** integrate all previous findings in spectral manipulation (Chapter 3), dynamic processing (Chapter 4) and auditory masking (Chapter 5), into one intelligent masking minimization system built upon an optimization framework that replicates the iterative process of human mixing. Within the system, we also explore the relationship between the two essential signal processing operations: equalization and dynamic processing, and proposes a general frequency and dynamics processing framework. Various implementations of the intelligent system are explored and evaluated objectively and subjectively through a listening experiment.
- **Chapter 7** concludes the thesis. Research findings are discussed and the prospects for future research were considered.

1.5 Contributions

- **Chapter 3**: A comprehensive spectral characteristic study of a large commercial recording dataset from 1950 to 2010.
- **Chapter 3**: A novel equalization algorithm based on Yule-Walker filter design to match any desired frequency response.
- **Chapter 4**: A fully automated multitrack dynamic range compression algorithm.
- **Chapter 4**: A novel Web Audio API based approach to compress an unprocessed broadcasting signal based on dynamically varying environmental noise level.
- **Chapter 5**: Novel psychoacoustics-inspired cross-adaptive masking metrics capable of quantifying the amount of masking occurring in multitrack audio.

- **Chapter 6**: An optimization-based approach to autonomous minimization of masking in multitrack audio.

1.6 Associated Publications

Conference

The spectral characteristics analysis of commercial recordings presented in Chapter 3 was published as:

- Pestana, P., et al. "Spectral characteristics of popular commercial recordings 1950-2010." *Audio Engineering Society Convention 135*. Audio Engineering Society, 2013.

This was done in a close collaboration with Pedro D. Pestana during the author's academic visit to Catholic University of Oporto, Porto, Portugal. The author of the thesis collected the large part of the dataset and did the core analysis of the spectral features in terms of yearly evaluation and genre differences. Pedro D. Pestana proposed the methodology to compare the spectrums and did the analysis of the overall average spectrum. The author was co-author on the paper. He wrote large part of Section 2, 3 of the paper. Pedro D. Pestana wrote the large part of the introduction, Section 1 and 4. All other authors had an editing and supervising role.

The novel approach to equalize audio signals toward a target spectrum curve described in Chapter 3 was published as:

- Ma, Z., et al. "Implementation of an intelligent equalization tool using Yule-Walker for music mixing and mastering." *Audio Engineering Society Convention 134*. Audio Engineering Society, 2013.

The author of the thesis wrote and did the main research. All other authors had an editing and supervising role.

The web-based personalized compression presented in Chapter 4 was published as:

- Mason, A., et al. "Adaptive Audio Reproduction Using Personalized Compression." *Audio Engineering Society Conference: 57th International Conference*. Audio Engineering Society, 2015.

The author's main contribution was the design and implementation of the core dynamic processing algorithm. He wrote the initial technical report of the research together with Nick

Jillings. Nick Jilling's main contribution is the HTML5 realization of the dynamic processing algorithms. The author of the thesis wrote the initial content in the introduction and section 2, 3 of this paper. Nick Jillings wrote the initial content in Section 1, 4. Andrew Mason, the first author of this paper, is the industrial supervisor on this project who wrote the final paper based on the technical report based on the technical report.

The loudness matching experiment described in Chapter 5 was published as:

- Ma, Z., et al. "Partial Loudness in Multitrack Mixing." *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*. Audio Engineering Society, 2014.

The author of the thesis wrote and did the main research. All other authors had an editing and supervising role.

Journal

The largest part of Chapter 4 on intelligent multitrack dynamic range compression algorithms was published as:

- Ma, Z., et al. "Intelligent multitrack dynamic range compression." *Journal of the Audio Engineering Society* 63.6 (2015): 412-426.

The author of the thesis wrote and did the main research. The second author Pedro D. Pestana provided insight into the rules used in the dynamic range compression automation in Section 2 of the paper. The third author Brecht D. Man provided insight into the testing methodology and the audio content in Section 5. All other authors had an editing and supervising role.

The largest part of Chapter 6, together with portions from Chapter 5 on masking modeling, were submitted as:

- Ma, Z., Reiss, J. D. "Autonomous Minimization of Masking in Multitrack Audio." Submitted to *IEEE Transactions on Audio, Speech, and Language Processing, 2015*.

The author of the thesis wrote and did the main research. All other authors had an editing and supervising role.

Patent

Intelligent multitrack mixing algorithms developed from Chapter 3 and 4 were published as:

- Reiss, J. D., Mansbridge, S., Clifford, A., Ma, Z., Hafezi, S. and Jillings, N. "System And Method For Autonomous Multi-Track Audio Processing." U.S. Patent 20,150,117,685, issued April 30, 2015.

Chapter 2

Background

We start by discussing the physiology of the human hearing system with an emphasis on the concepts of masking, critical bands and auditory filters. Several psychoacoustics-inspired loudness and masking models as the perceptual basis of our intelligent mixing studies are then reviewed. The process of multitrack mixing and related mixing techniques with a focus on the frequency and dynamic domains is also discussed. The chapter is concluded by reviewing the state of the art in the field of automatic mixing.

2.1 The Physiology of the Human Hearing System

There are three principal parts of the auditory system: the outer ear, the middle ear, and the inner ear. The auricle that has the responsibility of sound localizing, spectral shaping and overall loudness intensification, is located in the outer ear. Sound travels through the auditory canal and reach the eardrum that vibrates. Next to the eardrum are three smallest bones in the body, the malleus, incus and stapes (known collectively as the ossicles) (Moore, 2012). Ossicles are assisting the vibration transmission through middle ear to inner ear. The inner ear consists of the cochlea and the vestibular nerve. The cochlea is the sensory organ for hearing. Inside the cochlea is the basilar membrane which is tonotopic and each frequency has a characteristic place of resonance along it (Goldstein, 2013).

2.2 Critical Bands, Auditory Filters and Masking

The frequency resolution limitation of the human auditory system is often termed “frequency selectivity”. When presented with two sinusoidal stimuli of frequencies that are close enough. One believes to be hearing a single frequency that is the exact average of both, oscillating in amplitude at a rate that is equal to the absolute value of the frequency difference. As the

frequency difference grows apart, amplitude change discrimination starts becoming impossibility and a sense of roughness is heard instead. Raising the frequency difference even further will make it come to a threshold above which the two sinusoidal tones can be clearly distinguished (Howard & Angus, 2009).

The threshold mentioned defines the critical bandwidth for a certain central frequency. Harvey Fletcher (Fletcher, 1940), in a very famous experiment, measured the shift in threshold for detecting a sinusoidal signal for different bandwidths of band-pass noise maskers, where noise power density was constant. He found out that for small bandwidths the detection threshold would increase rapidly, but after a certain point it would completely cease to increase. These are now termed "Auditory Filters", and the idea of critical bandwidth defines he spectral length of an auditory filter. Fletcher provided a definition of critical bandwidth (CB) as "the bandwidth at which the signal threshold ceased to increase" (Fletcher, 1940).

The shape of the auditory filters can be determined in several different ways, all of them necessarily slightly flawed, as we are measuring auditory response to a signal in the presence of a masker, whereas the physiological auditory filter will respond to signal alone. As it is non-linear, the presence of the masker will necessarily bias our measurement.

Critical bands have different bandwidths up and down the spectrum. They are naturally smaller in absolute terms at low center frequencies, due to the log-lin behavior of frequency perception, but if one thinks of relative bandwidth (bandwidth divided by center frequency), they will actually be bigger at low frequencies, indicating worst tone-discrimination. In literature, critical bands are usually approximated by one-third octave filters or, alternatively, by what is called an Equivalent Rectangular Bandwidth (ERB), a rectangular function that covers the exact same area as a critical band would. The equation for ERB is shown in Equation (0.0).

$$ERB = 24.7(0.0437\,f + 1) \tag{0.0}$$

The distribution of activity evoked by that sound as a function of the characteristic frequency is called the excitation pattern (Moore, 2012). Excitation patterns are usually asymmetric,

being less steep on the high-frequency side. The asymmetry increases with increasing sound level.

Auditory masking is an auditory phenomenon we experience in our everyday life. In psychoacoustics masking is defined as "the process by which the threshold of audibility for one sound (the maskee) is raised by the presence of another sound (the masker)" (ANSI, 1994). Simultaneous masking or frequency masking occurs in the time domain while non-simultaneous masking or temporal masking occurs in the time domain.

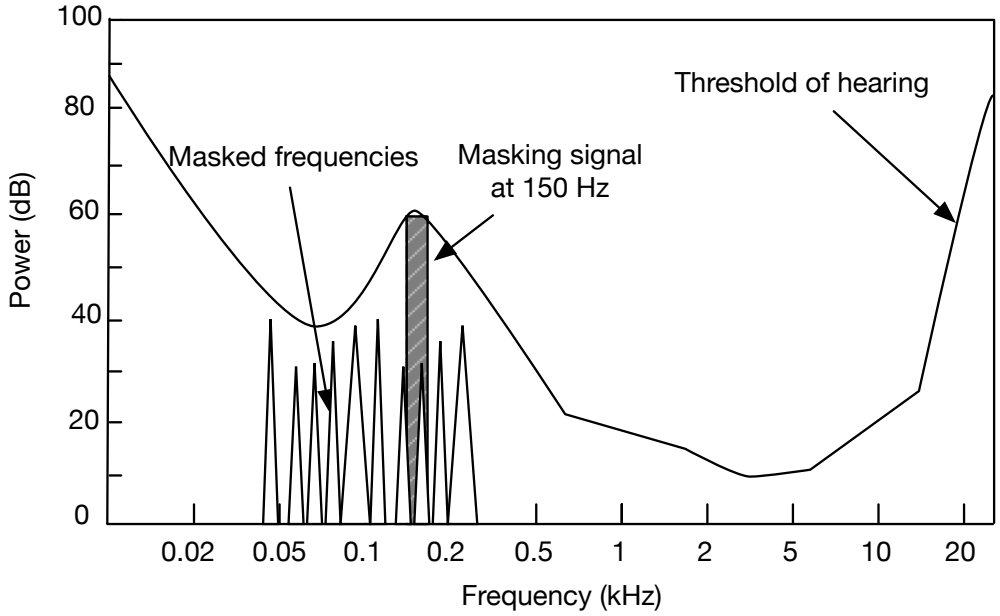


Figure 2.1 Simultaneous masking example of a 150 Hz tone signal masking an adjacent frequencies by increasing the threshold of audibility around 150 Hz.

Simultaneous masking may occur when two or more stimuli are simultaneously presented to the auditory system. An example of a 150 Hz tone signal masking adjacent frequencies is shown in Figure 2.1. A simplified explanation is that the presence of a strong signal creates a sufficient excitation strength on the basilar membrane to block the detection of the weaker signal at its CB location (Moore, 2012). Temporal masking happens when sounds are imperceptible due to maskers before or even after the presence of the sounds as shown in Figure 2.2.

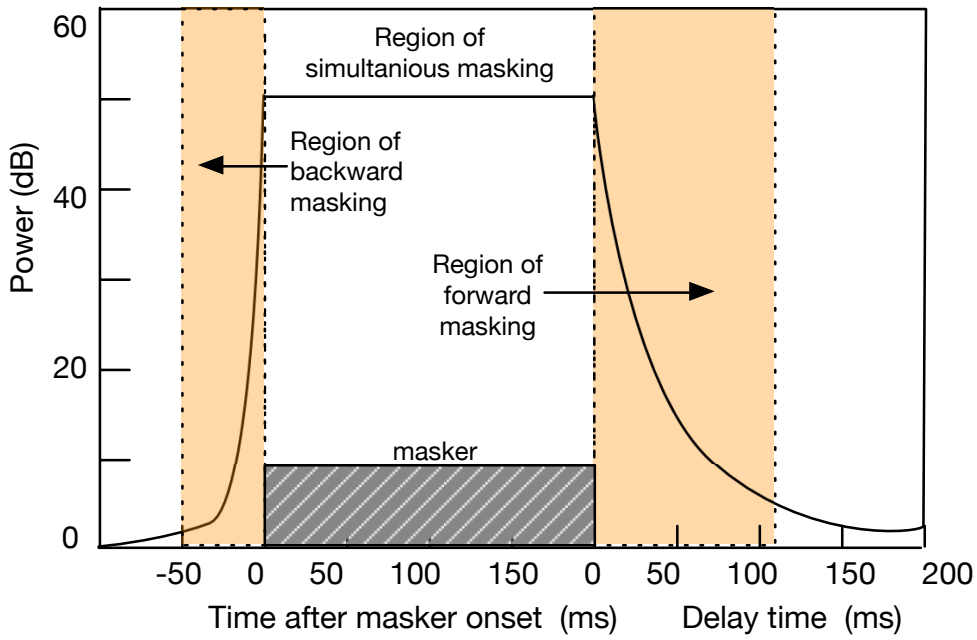


Figure 2.2 Regions of backward masking, simultaneous masking and forward masking. Note that backward masking uses a different time origin than forward masking and simultaneous masking.

2.3 Perceptual Models

The purpose of auditory modeling is to establish a bridge between the objective physical domain and the subjective domain of human hearing, which can offer a perceptual understanding of the mixing process. Perceptual models capable of quantifying both subjective loudness and auditory masking (two useful psychoacoustic properties that play an important role in mixing) are highly valued. For this reason, the development of popular loudness and masking models is presented next, followed by an overview of additional areas within the field of perceptual modelling.

2.3.1 Loudness Models

Loudness is defined as a psychological term used to describe the magnitude of an auditory sensation in (Fletcher & Munson, 1933).

Single band loudness models are very popular for describing program material e.g. broadcast material, primarily because of their practicality, and provide a general picture of loudness using direct measures of sound pressure. However, traditional volume unit meters such as the Peak Program Meter (PPM), as still used today, provide only a crude indication of loudness (Lund, 2005), requiring audio specialists to apply rule-based correction factors based on the type of input material (Emmett & Emmett, 2003). In response to this issue, (Soulodre, 2004) investigated the correlation between each of ten potential loudness meters and an additional two basic loudness algorithms, and the results of a series of loudness matching listening tasks involving typical program material. Though findings signified that a simple frequency weighted averaged energy measurement, known as $L_{eq}(RLB)$, outweighed the success of its competitors, companies TC Electronic and Dolby announced internal experiments, implying their own models were superior (Lund, 2005). $L_{eq}(RLB)$ became the BS.1770 standard (ITU, 2012a) which covers mono, stereo and 5.1 surround formats. Following this, TC Electronic proposed two supplementary loudness descriptors to the standard for characterizing properties of the audio material (Skovenborg & Lund, 2008). This led to three suggested descriptors as part of the EBU recommendation (EBU–Recommendation, 2011).

Multiband loudness models are often inspired by the psychoacoustics properties of human hearing system (Nielsen & Skovenborg, 2004). The fundamental models incorporating a common assumption that loudness is related to the total neural activity evoked by a sound (Moore, 2012) have been proposed by (Fletcher & Munson, 1933), (El Zwicker, 1958), (Eberhard Zwicker, 1977; Eberhard Zwicker & Scharf, 1965). It has been extended in the more recent work of (Glasberg & Moore, 2002, 2005; Moore & Glasberg, 1996; Moore, Glasberg, & Baer, 1997), as well as the Dynamic Loudness Model of (Chalupper & Fastl, 2002).

A true loudness model that accounts for influences of phase on loudness would require a time domain filterbank, such as the Gammatone (De Boer, 1975), compressive Gammachirp filter (Irino & Patterson, 2001), in either parallel or cascaded form (Unoki, Irino, Glasberg, Moore, & Patterson, 2006) or the dual resonance nonlinear (DRNL) filter (Lopez-Poveda & Meddis, 2001). (Chen, Hu, Glasberg, & Moore, 2011) demonstrated how excitation patterns

and loudness can be calculated directly via parallel filter structure i.e., the double-roex filter model. Though no time domain filter was designed, this research demonstrates a promising direction for the development of loudness models that lie closer to cochlear physiology.

Models for predicting the loudness of time-varying sounds have also been developed (Chalupper & Fastl, 2002; Glasberg & Moore, 2002; Eberhard Zwicker, 1977). The general goal is to model temporal integration in a way that corresponded with empirical data such as post-masking. Unlike other models, (Chalupper & Fastl, 2002) applied temporal smoothing to the specific loudness patterns prior to final integration. As demonstrated in (Rennies, Verhey, & Fastl, 2010), the slow decay of specific loudness patterns is necessary to account for loudness summation of non-synchronous tone pulses at different frequencies; the model of (Glasberg & Moore, 2002) cannot account for this. Though the model of Moore et al. (Moore et al., 1997) has been extended to account for the time-varying partial loudness of a signal in noise (Glasberg & Moore, 2005), it still does not account for temporal masking. On the other hand, since the partial loudness model calculates a masked threshold for a signal in noise, it can also be viewed as a model of simultaneous masking. Furthermore, Glasberg and Moore recently modified their previous work to account for binaural inhibition (Moore & Glasberg, 2007), but this was only verified for steady-state sounds.

2.3.1.1 Loudness and Partial Loudness Model of Glasberg and Moore

The loudness model of Glasberg and Moore (Glasberg & Moore, 2002; Moore et al., 1997) is one of the key perceptual models we evaluate, adapt and apply to our intelligent mixing system in the later chapters. The block diagram in Figure 2.3 illustrates the simplified stages involved in the model that account for three important processes in the human auditory system: the outer/middle ear transformations, basilar membrane processing and the cochlear hair cells firing signals to the brain. The procedure to derive the loudness and partial loudness of an audio signal (when presented with a masker signal) is described as follows. Equations used in this section are adapted from the original papers (Glasberg & Moore, 2002; Moore et al., 1997) and descriptions are adapted from (Simpson, Terrell, & Reiss, 2013).

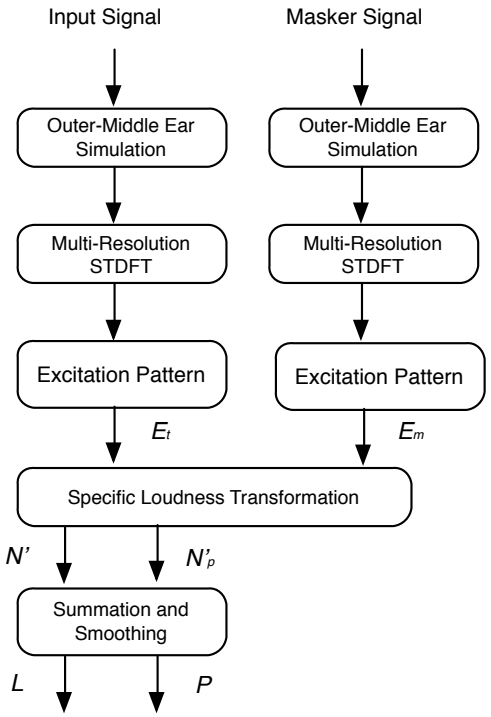


Figure 2.3 Block diagram of the model of Glasberg and Moore to derive loudness and partial loudness.

Stage 1: Outer and Middle Ear Transformations

The first stage of the model is to approximate the transformations that take place in the outer/middle ear. The signal is passed through an experimentally determined transfer function (implemented as a 4097 coefficient FIR filter) that models the frequency response of the sound pressure transmission through the outer and middle ear towards the cochlea.

Stage 2: Calculation of Running Spectrum and Excitation Pattern

The original model calculates six Hanning-windowed FFTs in parallel, using signal segment durations that decrease with increasing center frequency (Moore, 2012). Each spectral frame is filtered by a bank of level-dependent roex filters. Such spectral filtering represents the displacement distribution and tuning characteristics across the human basilar membrane.

The excitation pattern *E* is then calculated as the output of the auditory filters as a function of the center frequency spaced at 0.25 ERB intervals. Detailed excitation pattern calculation

can be found in (Moore & Glasberg, 1996). To account for partial masking when presented with a masker signal, two excitation patterns, the target input signal E_t and the masker signal E_m, are calculated.

Stage 3: Specific Loudness and Partial Specific Loudness

To reflect the production of neural signals in response to inner hair cell displacement caused by excitation of the basilar membrane, the excitation pattern is then transformed from excitation level into specific loudness N' (loudness per ERB) according to three possible conditions regarding the values of E_t and E_Q, which represents the threshold excitation in quiet and is frequency dependent. Detailed calculation can be found in (Moore et al., 1997) or (Simpson et al., 2013).

To account for partial masking due to the excitation pattern of the masker signal E_m, the model calculates partial specific loudness N_p' instead, by considering four conditions regarding the values of E_t, E_Q and E_m. Detailed calculation can be found in (Moore et al., 1997), or (Simpson et al., 2013).

Stage 4: Summation and Smoothing

The summation of N' and N'_p across the whole ERB scale produces the total unmasked and masked instantaneous loudness I, I_p respectively using Equation (0.0) and (0.0). ERB bands, $b_{ERB_{\min}}$ and $b_{ERB_{\max}}$ may be calculated from center frequencies of 50 and 15,000 Hz respectively, using Equation (0.0).

$$I = \sum_{b_{ERB_{\min}}}^{b_{ERB_{\max}}} N'(b_{ERB}) \tag{0.0}$$

$$I_p = \sum_{b_{ERB_{\min}}}^{b_{ERB_{\max}}} N'_p(b_{ERB}) \tag{0.0}$$

To account for the temporal integration of loudness due to the time-response of the auditory hearing system, the decaying value of loudness at time t is smoothed and calculated as short-term loudness, $I_{ST}(t)$ or long-term loudness, $I_{LT}(t)$ with an exponential sliding window:

$$I_{ST}(t) = (1-\alpha)I(t) + \alpha I_{ST}(t-\Delta t) \quad (0.0)$$

$$I_{LT}(t) = (1-\alpha)I_{ST}(t) + \alpha I_{LT}(t-\Delta t) \quad (0.0)$$

where α is the smoothing coefficient given by,

$$\alpha = e^{-\Delta t/\tau}. \quad (0.0)$$

Δt is the time step of the model and τ is the time constant that represents the decay of loudness. The value of τ is conditional depending on whether the functions are in the attack or release phase as shown in Equation (0.0) (Simpson et al., 2013):

$$\tau = \begin{cases} 22 & \text{for } I(t) > I_{ST}(t-\Delta t) \\ 50 & \text{for } I(t) < I_{ST}(t-\Delta t) \\ 100 & \text{for } I_{ST}(t) > I_{LT}(t-\Delta t) \\ 2000 & \text{for } I_{ST}(t) < I_{LT}(t-\Delta t) \end{cases}. \quad (0.0)$$

And finally, $I_{ST}(t)$ or $I_{LT}(t)$ is averaged across all time frames into scalar perceptual loudness measures, L (L_{ST} or L_{LT}). The same smoothing, summing and averaging operations are applied to $I_p(t)$ to derive the overall partial loudness of the input signal, P (P_{ST} or P_{LT}).

2.3.2 Masking Models

Perceptual models capable of predicting masking behavior have received much attention over the years, particularly in fields such as audio coding (Bosi et al., 1997; Gersho, 1994; Johnston, 1988b; Schroeder, Atal, & Hall, 1979), where the masking threshold of a signal was approximated to inform a bit-allocation algorithm. Similar models were used in sound quality assessment (Karjalainen, 1985; Thiede et al., 2000), where nonlinear time-domain filterbanks were used to allow for excitation patterns to be calculated whilst maintaining good temporal resolution. More advanced signal processing masking models that lie closer to physiology include (Dau, Püschel, & Kohlrausch, 1996). This initial single-band model accounts for a number of simultaneous and non-simultaneous masking experiments. A "modulation filterbank" was subsequently added to analyze the temporal envelope at the output of a gammatone filter whose output is half-rectified and lowpass filtered at 1kHz,

simulating the frequency to place transform across the basilar membrane, and receptor potentials of the inner hair cells (Dau, Kollmeier, & Kohlrausch, 1997). Building upon the proposed modulation filterbank, a masking model called the Computation Auditory Signal-Processing and Perception (CASP) model was presented that accounts for various aspects of masking and modulation detection (Jepsen, Ewert, & Dau, 2008).

However, all mentioned models only produce masking threshold as a measurement of masking, and only consider the situation when signal (typically, a test-tone) is fully masked. (Glasberg & Moore, 2005) explored partial loudness of mobile telephone ring tones in a variety of 'everyday background sounds' e.g. traffic based on previous psychoacoustic loudness models (Glasberg & Moore, 2002; Moore et al., 1997). By comparing the excitation patterns (computed based on (Glasberg & Moore, 2002; Moore et al., 1997)) between maskee and masker, (Vega & Janer, 2010) introduced a quantitative measure of masking in multitrack recording. Similarly, a Masked-to-Unmasked Ratio corresponding to the original loudness of an instrument to its loudness in the mix was proposed in (Aichinger, Sontacchi, & Schneider-Stickler, 2011). However, no temporal masking considered and no formal evaluations were provided in both (Aichinger et al., 2011; Vega & Janer, 2010).

(Plack, Oxenham, & Drga, 2002) incorporated the DRNL filter with a well-known model of temporal masking called the temporal-window model (Plack & Moore, 1990). In (Plack et al., 2002), the combination of a nonlinear filter based on response measurements of the basilar membrane with a leaky integrator was used to feed a decision device. (Hafezi & Reiss, 2015) introduced a simplified measure of masking based on best practices in sound engineering. However it might not correlate well with the perception of human hearing, as evidenced by the evaluation.

2.3.2.1 Psychoacoustic Model in MPEG Audio Coding

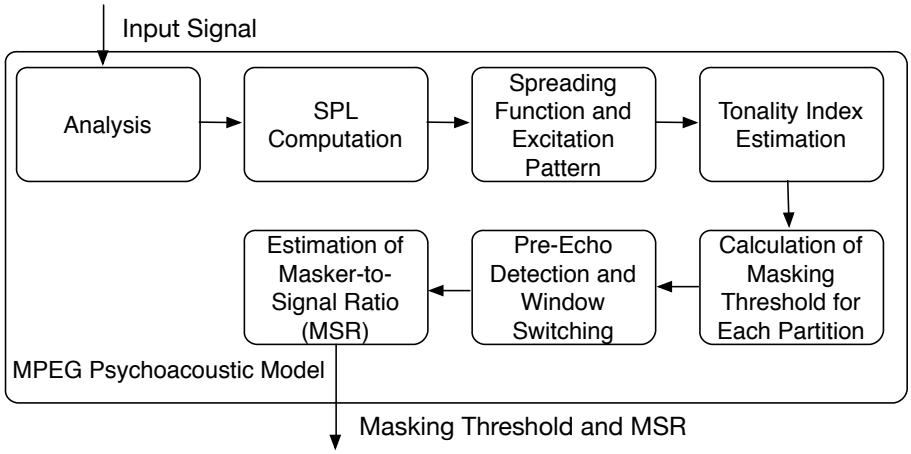


Figure 2.4 Simplified block diagram of the psychoacoustic model used in MPEG audio coding.

"The objective of audio coding algorithms is to represent the signal with a small number of bits while maintaining its perceptual quality such that it is indistinguishable from the original" (Thiagarajan & Spanias, 2011). The basic ideas behind perceptual audio coding involves first decompose a signal into separate frequency bands by using a filter bank; analyse the signal energy in different bands and determine the total masking threshold of each band because of signals in other band/time; quantise samples in different bands with accuracy proportional to the masking level. Any signal below the masking threshold does not need to be coded and signal above the masking threshold are quantized with a quantization step size according to the masking threshold and bits are assigned across bands so that each additional bit provides maximum reduction in perceived distortion.

The psychoacoustic model is the key element to the compression algorithm. The MPEG psychoacoustic model (ISO, 1993) computes the masking thresholds as a function of scaled frequency by analysing the signal and considering basic hearing properties. The simplified block diagram in Figure 2.4 illustrates the stages involved in the psychoacoustic model. The procedure and equations to derive masking thresholds adapted from (Thiagarajan & Spanias, 2011) are summarized as follows.

Step 1: Computation of Energy and Unpredictability in Threshold Partitions

A standard FFT is applied to the input signal to compute the complex spectrum. The polar representation of the spectrum is then used to compute the measure of unpredictability. The spectral components are grouped into threshold partitions (Thiagarajan & Spanias, 2011) to reduce the computational cost for following steps. The energy and unpredictability as functions of the threshold partitions are computed through integration.

Step 2: Computation of Spreading Function, Excitation Pattern and Tonality Index

The model applies a spreading function to account for the smearing effect of masking in the same critical band and neighbouring bands. The spreading function, s_f (measured in dB) used in this model is given by

$$s_f(i,j) = \begin{cases} 0 & B(z) \le -60 \\ 10^{\frac{(x+B(d_z))}{10}} & else \end{cases}, \qquad (0.0)$$

where the calculation of $B(d_z)$ can be found in (Bosi et al., 1997). d_z is the bar distance between maskee and masker (Thiagarajan & Spanias, 2011). Conversion between bar scale and frequency Hz can be approximated by

$$z(f) = 13\arctan(0.00076 f) + 3.5\arctan\left((f/7500)^2\right). \qquad (0.0)$$

The spreading function is then convolved with the partitioned, renormalized energy to derive the excitation pattern in threshold partitions. The unpredictability measure is convolved with the spreading function to take the spreading effect into account (Thiagarajan & Spanias, 2011). A tonality index to measure the degree of tone-like or noise-like is then derived from the energy and unpredictability of the signal in threshold partitions.

Step 3: Calculation of Masking Threshold in Threshold Partitions

The masking threshold is determined by providing an offset to the excitation pattern, where the value of the offset strongly depends on the nature of the masker (Thiagarajan & Spanias, 2011). The values for the offset are interpolated based on the tonality index of a noise masker to a frequency-dependent value defined in the audio coding standard (ISO, 1993) for a tonal masker (Thiagarajan & Spanias, 2011).

Step 4: Pre-echo Detection and Window Switching

Pre-echoes is a common artefact where the sound occurs before it happens due to the quantization errors in audio compression algorithm. Pre-echo is controlled by switching to shorter windows using perceptual entropy (Johnston, 1988a) as an indicator (Thiagarajan & Spanias, 2011).

Step 5: Estimation of MSR

The energy in each scale-factor band, $E_{sf}(sb)$ and the threshold in each scale-factor band, $T(sb)$ are calculated as described (Bosi et al., 1997) in a similar way. Thus the final Masker-to-Signal Ratio (MSR) in each scale-factor band is defined by

$$MSR(sb) = 10\log_{10}\left(\frac{T(sb)}{E_{sf}(sb)}\right). \tag{0.0}$$

2.3.3 Perceptual Models Summary

Most sophisticated, multiband models of masking and loudness are typically verified by experiments involving "laboratory stimuli" consisting of stationary sounds such as pure tones, broadband and narrowband noise, and time-varying sounds such as amplitude modulated sinusoids or sequences of noise bursts. The single band approach only loosely estimates auditory temporal integration, intensity scaling and approximations of outer-middle ear filtering, and thus cannot capture additional aspects of loudness perception such as spectral summation.

From reviewing the literature on perceptual models, it is clear that the incorporation of psychoacoustic principles into the signal analysis stage presents an attractive prospect. Various perceptual measures are being integrated as part of the analysis chain, and more established models are beginning to be explored, particularly those that predict masking phenomena and loudness perception of time-varying sounds. As discussed previously, under certain conditions, the models themselves are limited in performance, requiring either modifications

to account for psychophysical data or more sophisticated processing techniques based on human physiology to improve results if the application requires.

2.4 Multitrack Mixing

2.4.1 Mixing Process Overview

A typical (though not mandatory) signal processing workflow for mixing is shown in Figure 2.5.

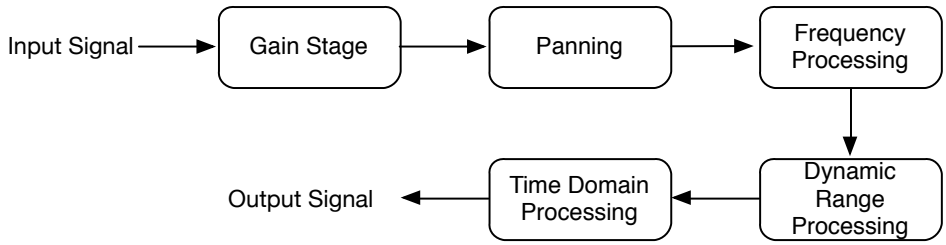


Figure 2.5 A typical (though not mandatory) signal processing workflow of mixing.

In general, the first stage of mixing process is to scale the input signal with a certain "gain" value. The fader is the most straightforward tool for coarse level adjustment in the mixing arsenal. Panning is the distribution of a sound signal into a new stereo or multichannel sound field through the use of amplitude differences between channels. Frequency domain processing involves using equalization and filtering to alter the spectral content of the audio signal. Equalization is one of the most important aspects of mixing (Izhaki, 2013). Dynamic domain processing as a nonlinear effect, involves the manipulation of the dynamic characteristics of the signals. Time domain processing, which is often classified into two classes: delays and artificial reverberation, is performed on the time axis (Izhaki, 2013).

Returning to the overall picture of the mixing process. Mixing can benefit from an iterative coarse-to-fine search (Izhaki, 2013) as illustrated in Figure 2.6. In a way, mixing is a equivalent optimization problem (Dennis Jr & Schnabel, 1996; Gill & Murray, 1974), which can shed some light on how to automate the mixing process (M. Terrell, Simpson, & Sandler, 2014). Given a certain set of controls of a multitrack, a mixing output can be

thought of as the optimal solution to a system of equations that describes the quality of the multitrack mixture, such as the amount of masking.

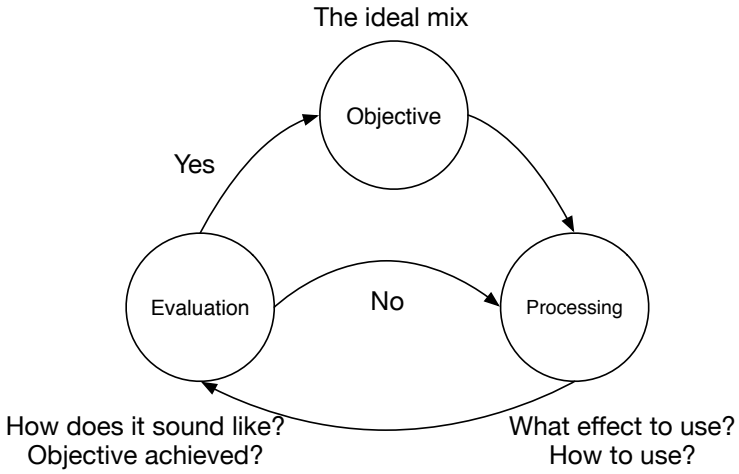


Figure 2.6 The iterative search process of mixing.

2.4.2 Frequency and Dynamic Domains

Achieving frequency balance is a prime challenge in most mixes (Izhaki, 2013). (Katz, 2007) proposed that the tonal balance of a symphony orchestra is an ideal reference for the frequency balance of music. The equalizer is the conventional tool to manipulate the spectral characteristics of the audio signal to achieve frequency balance. The filters used within equalizers is categorized as pass, shelving and parametric filters.

Dynamic range is often defined as the difference between the quietest and loudest sounds that an audio signal or a system can accommodate (Izhaki, 2013). Dynamic range processors including tools like compressors, limiters, gates, expanders and duckers, are tools to control the level variation and dynamic envelope of the signal. Among them, dynamic range compressor is one of the most important tools in mixing, which defines much of the sound of contemporary mixes.

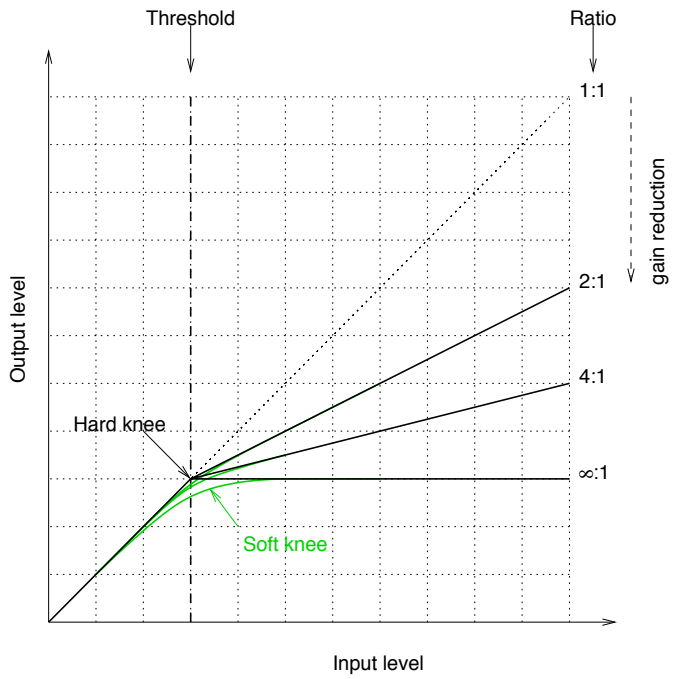


Figure 2.7 General form of compressor's transfer characteristic with different ratio values, hard or soft knee, and without make-up gain.

A typical set of dynamic range compressor parameters includes threshold, ratio, knee, attack, release and make-up gain. Figure 2.7 shows the basic transfer characteristic of the compressor. Threshold defines the level above which compression starts. Signals exceeding the threshold will be reduced in level. Ratio controls the amount of compression applied. It defines a drop in level above the threshold. The knee width controls the transfer characteristic around threshold. A sharp transition is called a "hard knee", and a smooth one, where the ratio gradually grows from 1:1 to a final value over a transition region spanning both sides of the threshold, is called a "soft knee" (Giannoulis, Massberg, & Reiss, 2012a). Softening the knee reduces the production of audible artefacts. The soft knee is shown in green in Figure 2.7. The attack and release times define how long it takes for the compressor to change its gain by 10dB towards the level determined by the ratio when the signal exceeds the threshold, and back again when it has stopped doing so.

Dynamic range compression (DRC) is commonly used in audio production, noise management, broadcasting, and live performance applications. However, it is arguably the most misused and overused effect in audio mixing (Izhaki, 2013). If used excessively, the dynamic range compressor suppresses musical dynamics, producing lifeless recordings

deprived of their natural character (Giannoulis et al., 2012a). Inappropriate parameter settings also produce artefacts such as pumping and breathing (Izhaki, 2013).

2.4.3 Equalization vs. Dynamic Processing

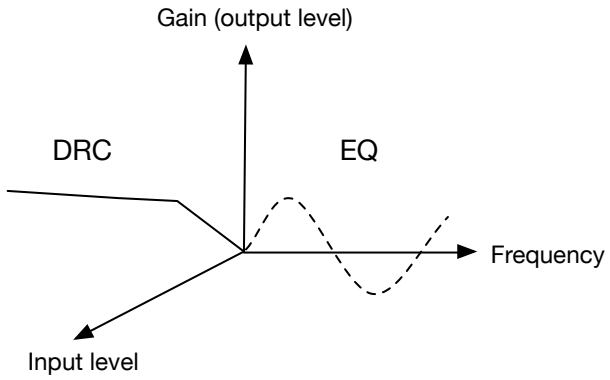


Figure 2.8 Separated control domains of equalization and dynamic range compression.

Equalization and dynamics processing are two essential signal processing operations in audio engineering. Equalization and dynamics processing often dominate exclusive domains, as shown in Figure 2.8. Equalization allows for the control of amplitude in the spectral domain, whereas dynamics processing allows for the control of amplitude in the time domain, especially in regard to the input level.

There have been many variants of systems combining the two operations, that is, time-domain control of amplitude over one or multiple spectral bands. Most of these variants address specific functionality such as gates, maximizers, or de-essers, and as such have limited configurability beyond their applications. Many problems, for example removing problematic frequencies, in audio production can be addressed by using combinations of filtering and dynamics processing. And previous research (Ma, 2015; Pestana, 2013; Pestana & Reiss, 2014) has shown that it is good practice to set compressor parameters based on the frequency content in the signal.

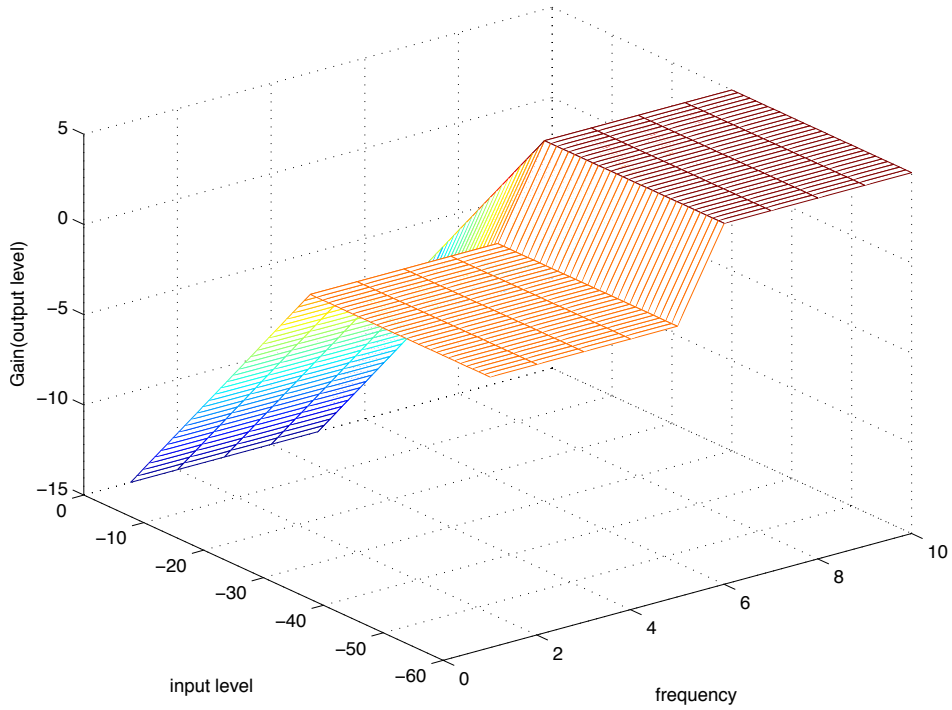


Figure 2.9 Control characteristics of a two-band compressor captured in the 3D space of frequency, input level and output (gain) level.

Multiband compressor operates differently and independently on different frequency bands of a signal, offering more precise adjustment of dynamics than single band compressor. Unwanted gain changes or artefacts (such as pumping and breathing) are avoided when applying compression on one frequency band. The crossover frequencies are often adjustable. The compression effect on each frequency band is controlled by its own compression parameters. The output signals of each frequency band are then combined as a final step. The control characteristics of multiband compression can be captured in the 3D space of equalization and dynamic processing as shown in Figure 2.9.

Dynamic equalizer (Reiss & McPherson, 2014) provides the ballistic control of a compressor like threshold, attack and release, to the conventional equalizer allowing time-varying adjustment of equalization curve. In other words, the equalization stage is able to respond dynamically to the input signal level. The control characteristic of a 3-band dynamic equalizer is shown Figure 2.10.

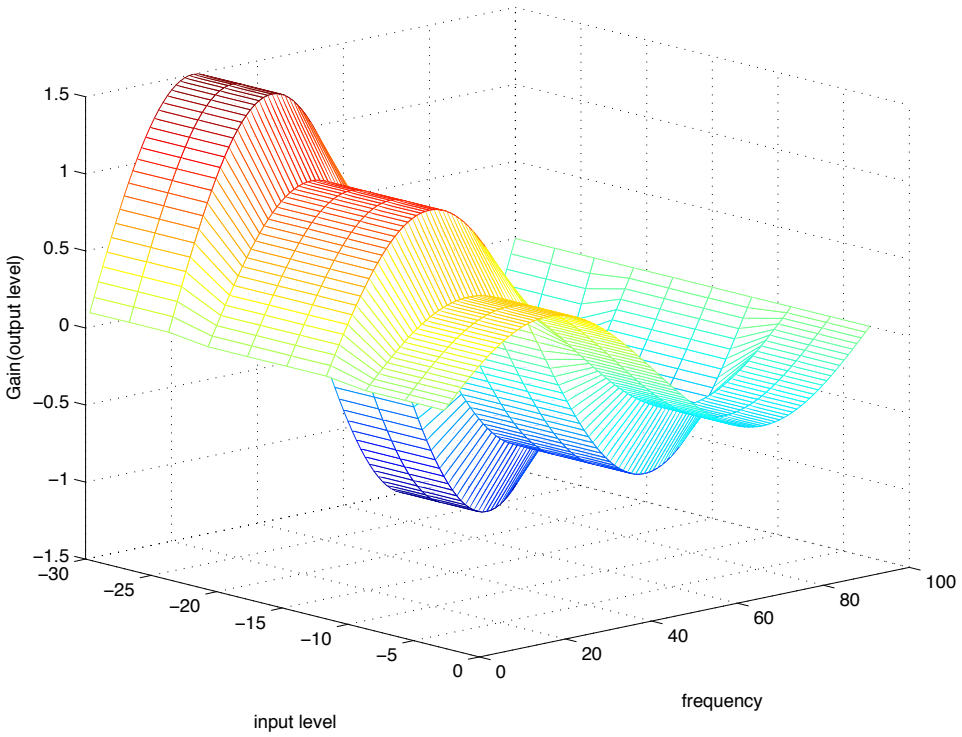


Figure 2.10 Control characteristics of a 3-band dynamic equalizer captured in the 3D space of frequency, input level and output (gain) level.

Many of these dynamic equalizer implementations are often used for noise reduction in audio restoration (Godsill, Rayner, & Cappé, 2002), hearing-loss correction (Lindemann, 1997), and compliance with broadcasting regulations. Other dynamic equalizers employ automatic gain adjustment of a fixed FIR or IIR filter. The modulation can be gated, as in de-hum and de-ess processors (Zolzer, 2011). Still other dynamic equalizers allow the filter to be configurable in the band it operates on. The dynamics that most of these systems offer to the engineer are constrained to the point that not all of the details are controllable. Yet dynamic equalizer is the closest design currently available to the concept of a general frequency and dynamics tool. Aassuming all parameters are configurable, the dynamic equalizer can be configured to a conventional equalizer, dynamic range compressor or multiband compressor.

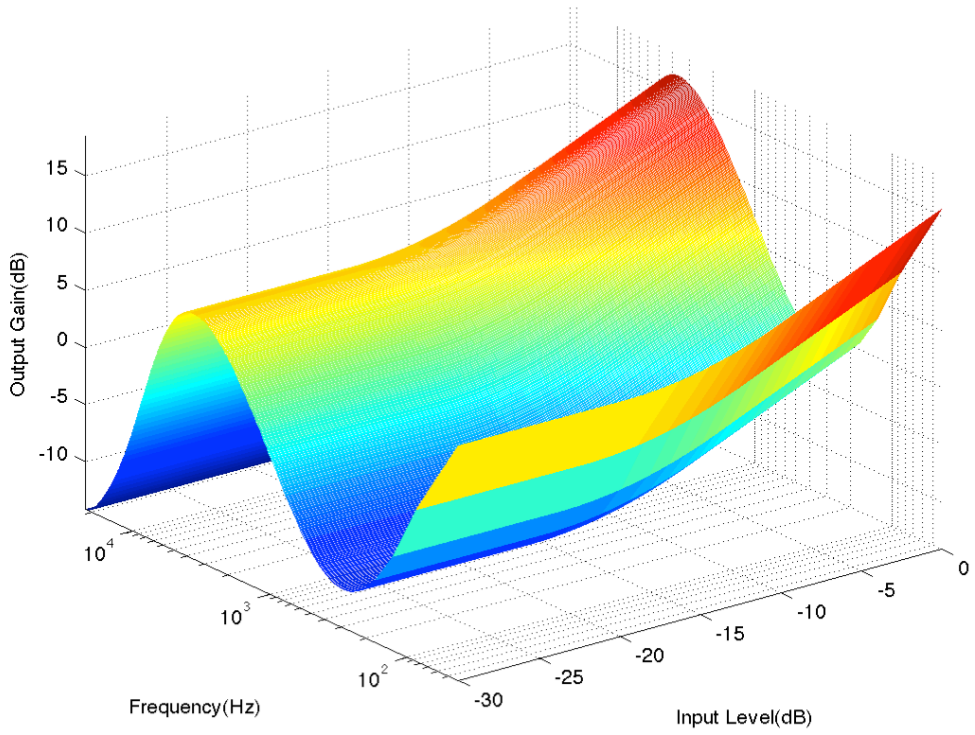


Figure 2.11 Control characteristics of a general frequency and dynamics processing tool in a 3D space of frequency, input level and output (gain) level.

"The operational nature of the equalizer and dynamic processors gives insight to a manner in which they may be combined into a general processor. This integrated processor can perform as the equivalent of a standalone dynamics processor or parametric equalizer, but can also modify the boost and/or cut of an equalizer stage over time following a dynamics curve" (Wise, 2009). Such idea of a general processor that utilizes the equalization and dynamic processing operations offers larger, unprecedented control over dynamics of specific frequencies of the audio as shown in Figure 2.11.

2.5 State of the Art: Intelligent Mixing

(Moorer, 2000) proposed the arrival of intelligent assistants, allowing computer programs to "take over the mundane aspects of music production, leaving the creative side to the professionals, where it belongs", in other words, intelligent mixing. Adaptive digital audio effects (A-DAFx), time-varying effects for controlling specific mix parameters automatically based on feature extraction (Zolzer, 2011), have been developed as processing devices

commonly employed by engineers to fulfill such requirements. However, the sound features propelling such effects are at the forefront of research for music mixing applications. In the pursuit of replicating human mixing, an important cross-adaptive digital audio effect (CA-DAFx) framework is proposed (Reiss, 2011; Zolzer, 2011), allowing for a more sophisticated system in which the sound features are extracted from multiple channels. In this section, we start with an introduction of the CA-DAFx framework, upon which most of our researches are built. And then we present a comprehensive review on the state of the art of the intelligent mixing techniques from various aspects: level, frequency, dynamics and beyond.

2.5.1 Cross-Adaptive Digital Audio Effects

Figure 2.12 depicts the aforementioned CA-DAFx framework, an important breakthrough for intelligent mixing systems. Both the feature extraction stage and cross-channel analysis feed information to a decision device, which subsequently processes each of the incoming channels, resulting in sonic improvement. As stated in (Reiss, 2011), a CA-DAFx is an inter-channel dependent effect and the signal processing of one individual source is the result of the relationships between all involved sources. The actual cross-adaptive processing can be informed and constrained by a set of constrained rules from mixing best practices, perceptual models and subjective evaluation (Reiss, 2011).

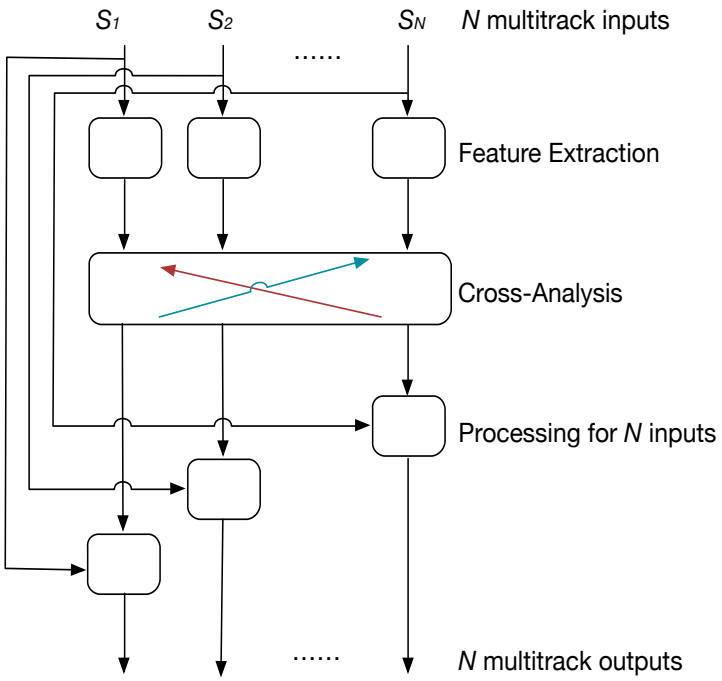


Figure 2.12 Block diagram of the cross-adaptive digital audio effect architecture with *N* multitrack inputs and outputs.

2.5.2 Level

In terms of the mixing domains of overall level or loudness, (Mansbridge, Finn, & Reiss, 2012) improved (Perez-Gonzalez & Reiss, 2009) by the use of the loudness measure, with a cross-adaptive process to bring each track to a time-varying loudness average measured by EBU loudness standard (EBU–Recommendation, 2011). A hysteresis loudness gate and selective smoothing were also introduced to prevent the unwanted artifacts. (Ward, Reiss, & Athwal, 2012) again adapted the equal loudness mixing concept, using a more sophisticated psychoacoustic loudness and partial loudness models (Glasberg & Moore, 2002; Moore et al., 1997). (M. J. Terrell & Reiss, 2009) presented a model to improve the monitor mix experienced by different musicians in a live performance tailored to their own listening condition and requirement where feedback prevention, SPL contraints were the main concerns.

(Kolasinski, 2008) introduced a method for balancing the multitrack level using timbral classification and genetic optimization. (J. Scott, Prockup, Schmidt, & Kim, 2011) developed a system can derive the mixing parameters through least-squares estimation. However the proposed system required prior knowledge of the instrumentation and was limited to very specific instruments. (J. J. Scott & Kim, 2011) improved the system with the introduction of acoustic features constraints.

(Ward et al., 2012) applied a partial loudness model (Glasberg & Moore, 2002; Moore et al., 1997) to adjust the levels of tracks within a multitrack in order to counteract masking. Based on the same model, (M. Terrell et al., 2014) developed an optimization theory treatment of the problem of level adjustment.

2.5.3 Frequency

Intelligent mixing techniques related to frequency processing are relatively unexplored. (Tsingos, 2005; Tsingos, Gallo, & Drettakis, 2004) applied perceptual audio coding to cull irrelevant sound sources to accelerate the rendering of complex virtual environments.

(A. Kleczkowski & Kleczkowski, 2006; P. Kleczkowski, 2005) proposed a novel multitrack mixing technique by removing non-dominant parts from the time-frequency space to improve the clarity of the multitrack mixture. (Tsilfidis, Papadakos, & Mourjopoulos, 2009) followed this idea and proposed a method to maintain only perceptually relevant elements of the audio signals according to the calculated minimum masking threshold. (Hafezi & Reiss, 2015) designed a simplified measure of masking based on best practice, and proposes an automatic multitrack equalization to reduce masking.

(Reed, 2000) proposed a simple machine learning based equalization system to replicate the human process. (Sabin & Pardo, 2009) described an equalizer with intuitive controls by mapping an individual's descriptive term onto equalization setting from a user's subjective preference. The method was extended and improved in (Pardo, Little, & Gergle, 2012) by training the system with a prior knowledge database of experts.

2.5.4 Dynamics

Automatic dynamic range compression research has a diverse history (Tyler, 1979). An RMS estimation was used to automate the release parameter in (McNally, 1984). In (Aichinger et al., 2011) the time constants were automated based on the difference between the peak and RMS levels of the signal fed into the side-chain. More relevant research can be found in (Giannoulis et al., 2012a; Giannoulis, Massberg, & Reiss, 2012b) where a series of DRC parameter automation methods derived from side-chain feature extraction were presented. However, in this system, the threshold was still manually chosen, with ratio set to infinity and an automated soft knee determining the amount of compression based on spectral flux. A new linear DRC technique that reduced the peak amplitude of transient signals using golden ratio allpass filters was introduced in (Parker & Valimaki, 2013). In (Wilmering, Fazekas, & Sandler, 2012) a new class of adaptive digital audio effects that mapped semantic metadata to control parameters was proposed. However, the system assumed that the metadata already

exists, either from a prior process or manual configuration and might be invoked on demand. The automation was performed using a fairly simple mapping between metadata and static compression presets. No subjective evaluation was provided.

Perhaps the most relevant previous work is (Maddams, Finn, & Reiss, 2012), which described an off-line method for automating multitrack DRC based on loudness and loudness range. The control strategy was to reduce the difference between the highest and lowest loudness range of the multitracks and sound sources where a higher loudness range requires greater amounts of DRC. However, the parameter automation of transforming the three controls (threshold, ratio and knee) into a single control could have a significant effect on the final result. The evaluation results in (Maddams et al., 2012) were inconclusive regarding the sonic improvement of the mixes.

2.5.5 Other Approaches

(Bocko, Bocko, Headlam, Lundberg, & Ren, 2010) proposed an automatic mixing system that applied probabilistic graphical model to best practices in audio engineering to produce mixing decisions based on audio features. (Sánchez, 2009) suggested that most mixing parameters can be derived from masking as they are all frequency dependent.

Reverse engineering offer another interesting aspect of intelligent mixing. Two different least squares optimization based methods were presented in (Barchiesi & Reiss, 2010) to derive the mixing parameters such as gains, delays, filters and panning setting when the unprocessed multitrack and the final mix are at hand. However, the system does not incorporate any form of perceptual analysis, as the goal was to retrieve effect parameters from a target mix.

Chapter 3

Frequency Processing

This chapter investigates the frequency aspect of intelligent mixing. We first present a spectral characteristic analysis of popular commercial recordings. We discover a consistent leaning towards a target spectrum that stems from practices in the music industry. A new approach for automatically equalizing audio signals towards the observed target spectrum is then described.

3.1 Introduction

Previous research on the frequency aspect of intelligent mixing has been reviewed in the Background Section 2.5.3. Evaluation results often appear to be inconclusive. Indeed, applying equalization to achieve a balanced spectral distribution is the most challenging task in mixing.

An efficient and stable filter design that can resemble any desired frequency response offers great value for the intelligent equalization techniques. Finite impulse response (FIR) filter design based on the least-squares method provides a quick solution (Ahmad & Wang, 1989; Algazi, Suk, & Rim, 1986; Friedlander & Porat, 1984; Kobayashi & Imai, 1990; Lim, Lee, Chen, & Yang, 1992; Pei & Shyu, 1994; Sunder & Ramachandran, 1994). But there are caveats with FFT convolution methods, to do with loss of precision, computational complexity, quantization, dither and the effects of the inevitable FIR windowing when filtering the input signal with FIRs. IIRs can avoid many of these disadvantages. But an IIR filter is a complex feedback network. There is a dearth of good methods to design these once moving away from classic filter transfer functions. (Lee, 2008) described a method of fitting infinite impulse response (IIR) filters to an arbitrary frequency response using Singular Value Decomposition (SVD). However, evaluation showed that this method also lacked accuracy in

the lower frequencies and not suitable for low-pass, high-pass and band-pass filters with classic response shapes. The Yule-Walker method of Autoregressive Moving Average (ARMA) spectral estimation (Friedlander & Porat, 1984) was found to provide a better spectral accuracy, where the computational cost was reasonable.

A few commercial plug-ins are capable of matching the spectrum of one piece of audio to another, such as Logic Pro's Match EQ, iZotope's Ozone, DUY's MagicSpectrum. However, none of them are truly real-time. A learning process of the spectral content of both input and source file is needed before actual filtering. In most cases, a single equalization curve (time-constant) is calculated and applied to the whole signal using either FIR filters or parametric filters to fulfil the roles.

In this chapter, we first present spectral characteristics analysis of popular commercial recordings in Section 3.2. The long-term spectral contours of a large dataset are analyzed. Overall spectrum trends, spectral feature evolution in years and in genres are analyzed. We discover that there is a consistent leaning towards a target spectrum. Based on the analysis, a new approach for automatically equalizing an audio signal towards the observed spectrum is presented in Section 3.3. The algorithm is based on the Yule-Walker method and designs recursive IIR digital filters using least-squares fitting to any desired frequency response. Objective evaluation is provided in Section 3.4, where the output frequency spectra are compared against the target spectrum and those produced by an alternative equalization method.

3.2 Spectral Characteristics of Popular Commercial Recordings

3.2.1 Dataset

The dataset contains almost half the number-one singles (either in the UK or US chart) over the last 60 years according to OCC, Billboard and Wikipedia. We chose these criteria in order to be consistent with public preference. It has a good representation of both genre and year of production.

All the songs in our dataset are uncompressed and, while we tried to find un-remastered versions, it was not always possible. This means that we gave extra prominence to current

standards of production and the differences we present should be even greater than that which our data suggests. Table 3.1 shows the number of songs we had, divided by decade and genre.

Table 3.1 Number of songs per decades in the dataset.

Years	Number of Songs	Genre	Number of Songs
50s	71	Pop	178
60s	156	Rock	102
70s	129	Electronic	64
80s	193	Hip-hop	79
90s	96	Folk	48
After 2000	127	Disco	52
		R&B	112
		Soul	89

3.2.2 Overall Average Spectrum of Commercial Recordings

Our main analysis focused on the monaural, average long-term spectrum of the aforementioned dataset. In order for spectra to be comparable, we first make sure that all songs are sampled at the same frequency (44.1 kHz being the obvious candidate for us, as most works stemmed from CD copies), and that we apply the same window length (4096 samples) to all contents, so that the frequency resolution is consistent (≈ 10 Hz). Let:

$$X(k,\tau)=\sum_{n=\tau w_{len}}^{(\tau+1)w_{len}-1} x(n)e^{-j2\pi k\frac{n}{N}},$$
$$k=\{0,1,...,2^{12}-1\},\tau=\{0,1,...,\left[\frac{x_{len}}{w_{len}}\right]\}, \quad (0.0)$$

where k is the frequency bin and τ the time window number. x_{len} and w_{len} are the song and window lengths, respectively. And we then consider the integrated spectral response to be the mean magnitude over τ:

$$\bar{X}(k) = \frac{\sum_{\tau} |X(k,\tau)|}{\left\lfloor \frac{x_{len}}{w_{len}} \right\rfloor + 1}. \qquad (0.0)$$

Equation (0.0) loses the 1 in the denominator whenever mod(x_{len},w_{len})=0.

It is still necessary to tackle the problem of different spectral distributions having potentially different overall power values. Strict normalization is not the answer, as spurious radical peaks in the frequency distribution might cause overall lower power levels, and the comparison would yield results that showed a variability that was greater than the real variability (one could take, as an example, a comparison between a white noise spectrum and one that adds a single sinusoid at 1000 Hz to the same white noise — if the sinusoid is greater in magnitude, a normalization process would bring all other bins in the second spectrum down and lead us to conclude that the spectra were very different, while in actuality they are not). There are several available solutions, but we opted to scale all spectral distributions so that the bin sum would be 1, followed by averaging the cumulative distribution function. This means normalizing according to:

$$\tilde{X}(k) = \frac{\bar{X}(k)}{\sum_{k} \bar{X}(k)}, \qquad (0.0)$$

and accumulating over the bins:

$$X_c(k) = \sum_{i=0}^{k} \tilde{X}(i). \qquad (0.0)$$

We then compute a mean calculation of each point in the cumulative distribution $X_c(k)$. The average spectrum is computed from the differences between adjacent bins, and multiplying by the average magnitude of all songs. This is basically an inversion of the process described above, and it is shown in the following equation:

$$\tilde{X}_{AV}(k) = \frac{\sum_{k} \bar{X}(k)}{N} (\overline{X_c(k)} - \overline{X_c(k-1)}), \qquad (0.0)$$

where N is the total number of songs.

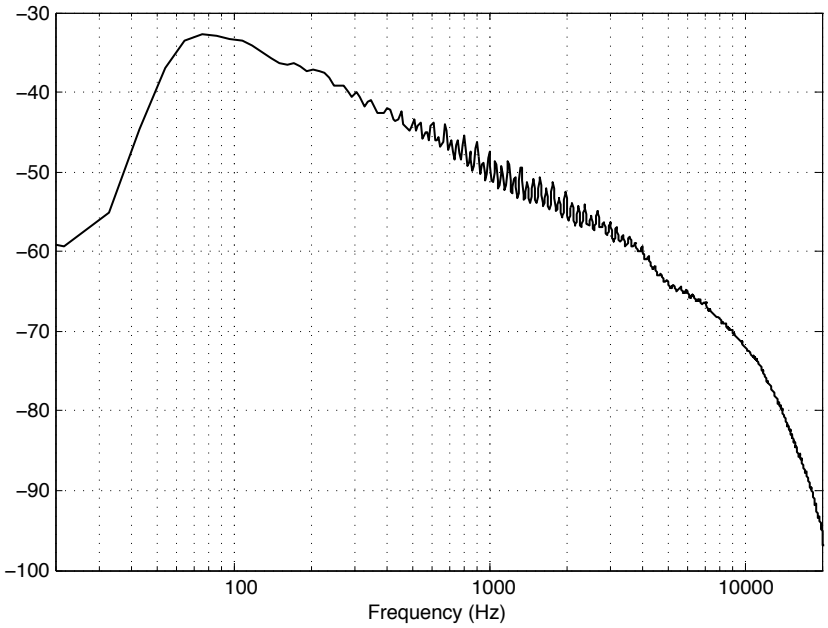


Figure 3.1 Average spectrum of all available data.

The result of averaging the spectra of all songs in the dataset is shown in Figure 3.1, along with a plot that overlaps all the individual distributions. The trend seen in the average spectrum is consistent with what can be observed for the individual distributions and the 95% confidence interval indicated are so narrow that they are not perceptible on the shown scale. The average standard deviation for the normalized cumulative values is 0.044, which is a well-behaved value across frequency bins (though averaging 2048 standard deviations drowns out the larger values in the low-end frequency region). All the subsequent analysis follows this averaging scheme.

3.2.3 Yearly Evolution of Spectra and Spectral Features

Figure 3.2 shows the average spectrum evolution through time, along with some decade-by-decade snapshots of revealing frequency ranges.

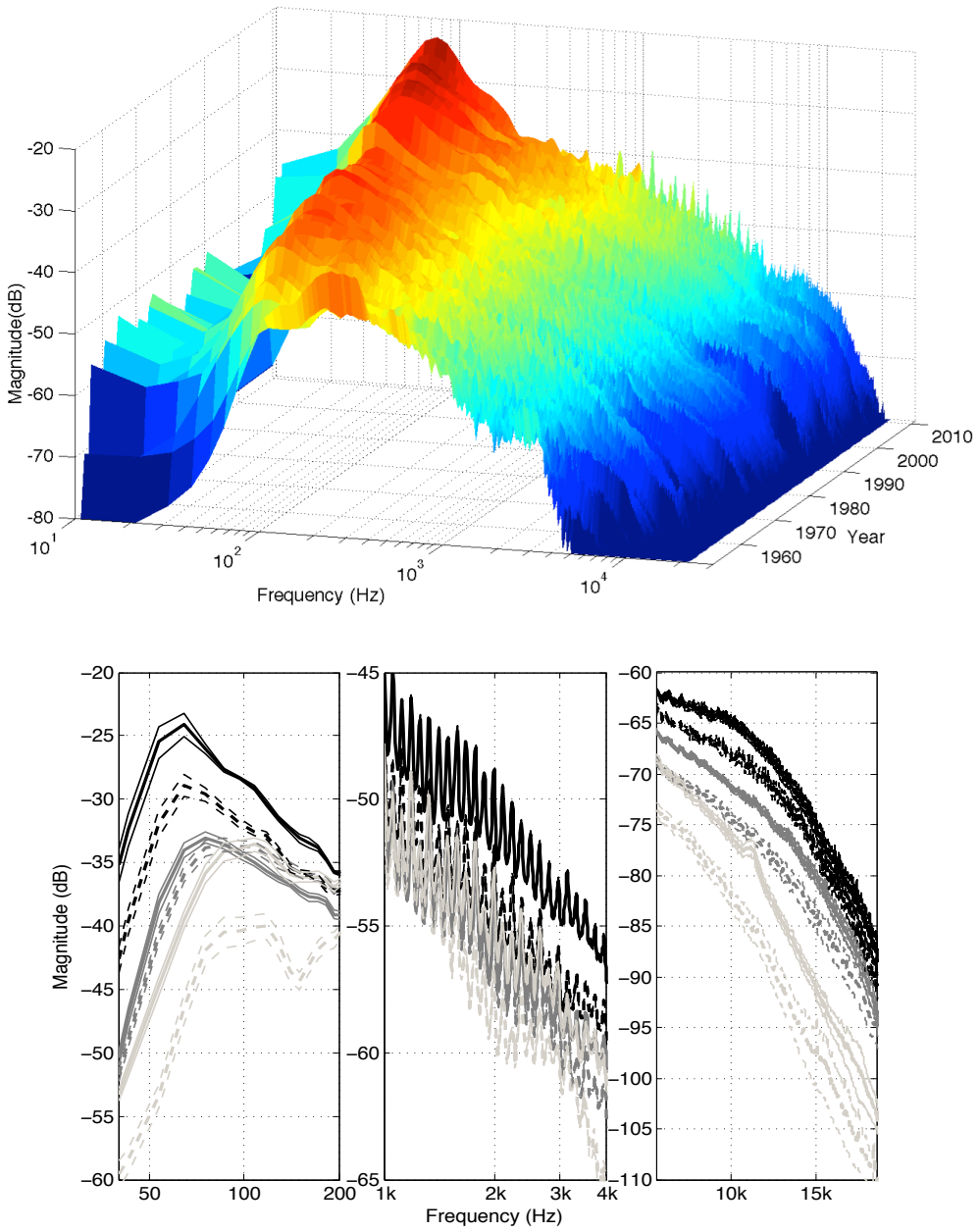


Figure 3.2 Average spectra on a yearly base (top) and frequency region details per decade (bottom), from left to right: 40–200 Hz, 1–4 kHz and 7–20 kHz. Darker colors represent later decades in the bottom plot.

An interesting feature is the raggedness of the mid-distribution (detailed in Figure 3.3), and particularly its evolution. When we look at the comb-like shape of the line representing the most recent decade, we are seeing peaks in every note of the dodecaphonic scale in equal-tempered western tuning. Looking back in time we see that raggedness emphasizes some notes over others, which may well indicate predominance of certain tonalities over others.

This is particularly clear during the 50s and 60s. While this is an interesting point, if we are concerned with equalization practices on the engineering and production side we should discard tonal features and concentrate on the broad spectral contour.

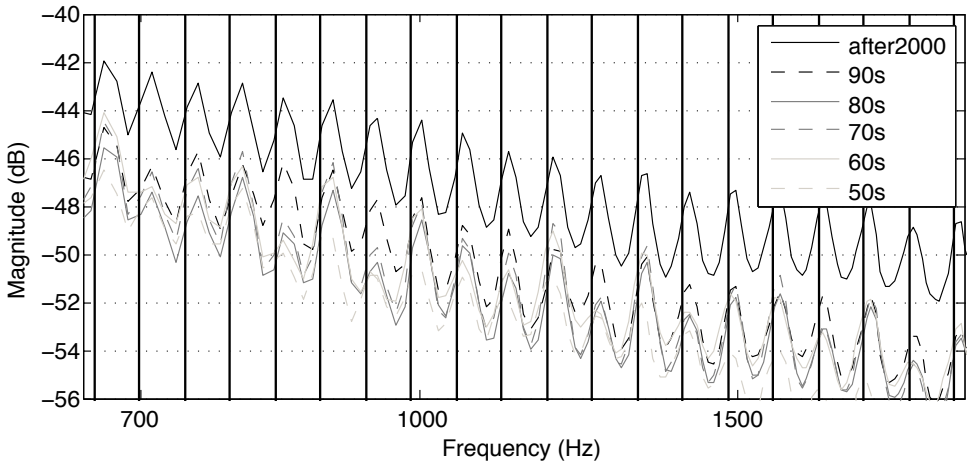


Figure 3.3 Detail of the emphasis on tonal frequencies for the decades where the difference is more accentuated. Actual fundamental frequencies are shown as vertical black lines.

There are some additional spectral features whose evolution might be interesting to look at, detailed in Figure 3.4. Spectral centroid is defined in (Peeters, 2004):

$$\mu = \sum_k kp(k), \tag{0.0}$$

where k is the frequency bins k of the DFT. And the magnitude of the normalized spectral envelope $p(k)$, is given by

$$p(k) = \frac{|X(k)|}{\sum_k |X(k)|}. \tag{0.0}$$

Spectral crest uses the equation given in (Krippendorff, 2012) defined in

$$\sigma^2 = \sum_k k^2 p(k). \tag{0.0}$$

We simplified the spectral slope measure, in that it is simply the slope of the log-log regression of the data points between 100 Hz and 10000 Hz ($i = k \in (100,1000)$ hz):

$$\lambda = \frac{1}{\sum_i p(i)} \frac{N\sum_i p(i)k(i) - \sum_i p(i)\sum_i k(i)}{N\sum_i k^2(i) - \left(\sum_i k(i)\right)^2}. \tag{0.0}$$

Finally, the spectral peak is purely a measure of the log magnitude of the bin whose value represents the global maximum.

Spectral centroid, as a common approximation of brightness, is maintained roughly around 900 Hz throughout all time. This suggests that popular commercial recordings have a dominant preference on the overall brightness no matter which decade the recordings were produced. The peak frequency decreases dramatically from 1950 to 1980 then slows down until 2003, from where it starts to increase slightly. Similar behavior (with different direction) can be seen from the results of peak magnitude. There is significant increase in peak magnitude from 1950 to around 1975. During the period of 1975 to 1990, the peak magnitude exhibits less sharp deviation. However from 1990 onward, it starts to increase dramatically again until it reaches the peak at 2005. Then it starts to show a trend of decreasing. The significant changes in peak frequency and magnitude could due to the audio world undergo the "switch" from analogue to digital. And during the modern digital era, the average magnitude peak and overall magnitude are increasing, and the spectrum (as spectral crest and spectral slope results suggest) tends to become flatter, partly due to the increasing amount of compression, see (Vickers, 2011).

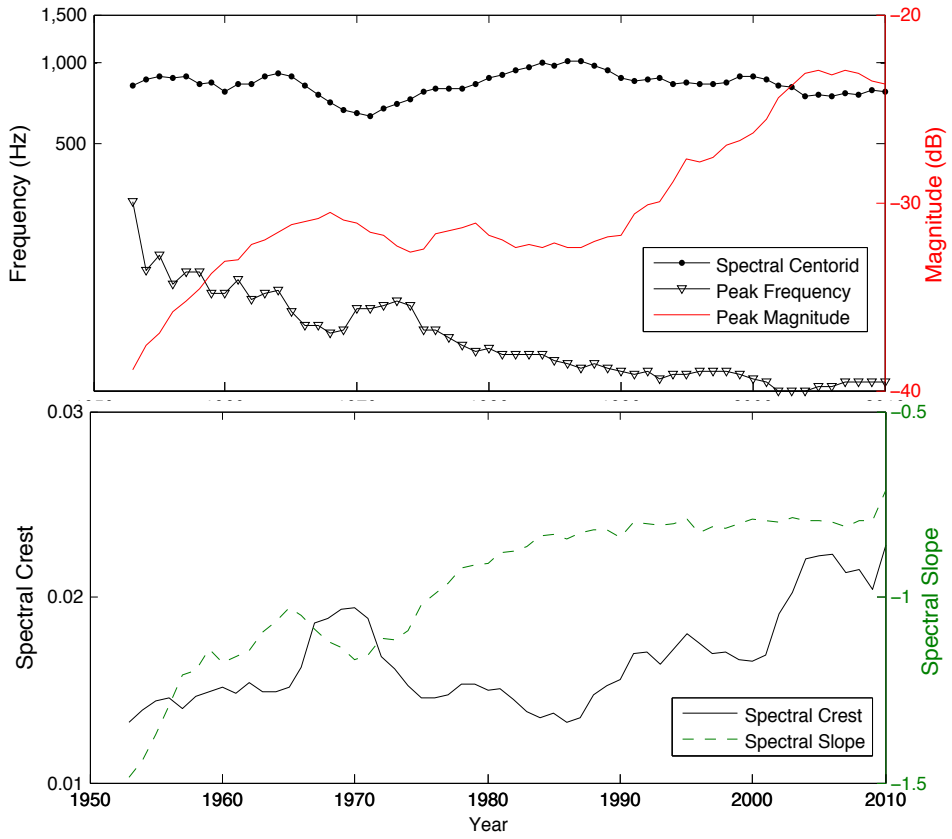


Figure 3.4 Yearly evolution of low-level spectral features: spectral centroid, peak frequency, peak magnitude, spectral crest and spectral slope.

3.2.4 Differences Stemming from Genre

Genre differences can also yield interesting results, and these are shown in Figure 3.5. We took our data from Wikipedia primarily, with tags from EchoNest and LastFM (when tags from Last.fm disagreed with Wikipedia, the data from Wikipedia is used). The extremely extended low-end response of electronica and hip-hop is unmistakable, whereas, as expected, R&B and jazz have a lighter bottom. The prominence of the top-end also yields differences in excess of 10 dB, which are meaningful even in the light of the overall magnitude increase of the brighter genres. The brightest mixes seem to be hip-hop ones, followed by electronic and disco. Here, however, this enhanced top end is negligible when considering that there is an overall enhancement (due to higher loudness specifications). On the dull side, folk and jazz genres suggest that there is natural top-end decay on more acoustic endeavours, whereas electronic ones allow and benefit from bigger frequency extensions.

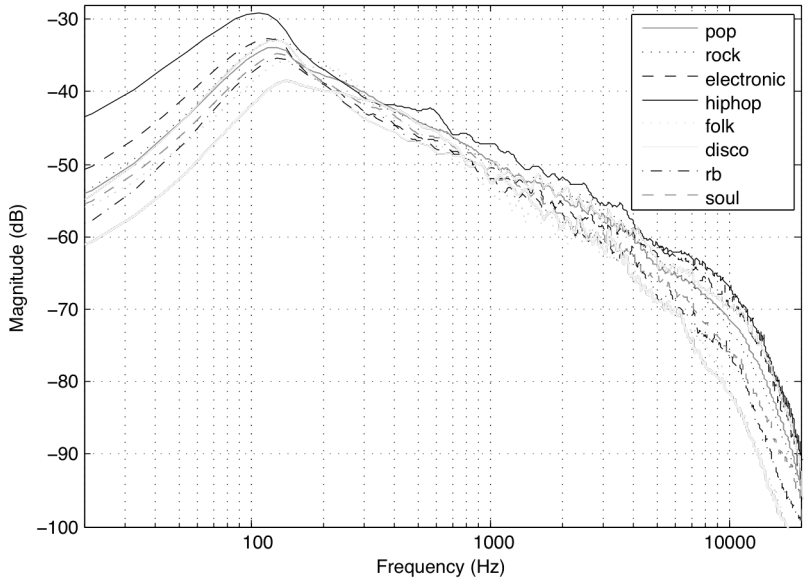


Figure 3.5 Average spectra by genre for a selection of genres.

On the middle-part of the spectrum, it is interesting to observe that pop and rock seem to be more openly harmonic in nature (again, raggedness in the frequency response), with no preference of tonality. Hip-hop in contrast, seems to have fewer harmonicas, which may be due to the prominence of rhythmic elements. Note that there might be a bias induced by the number of songs in each genre. The domination of pop and rock in the charts may possibly enhance a more even distribution of tonal content, as there are more songs in more varied keys. We chose not to go into sub-genres, as the academic consensus is very low in terms of genre definition, let alone sub-genres. The genre divisions are much less clear-cut, and the only region with no confidence interval overlap is the low-end.

Table 3.2 shows the difference in the low-level descriptors mentioned above. These reinforce the observations above in that genre differences are significant in terms of spectra. However, genre-popularity shifts over time. Thus, hip-hop's more prominent loudness and extended bass response is evidently related to the fact that post-2000 songs share the same tendency.

Table 3.2 Values of low-level spectral features compiled by genre.

Genre	Spectral Centroid (Hz)	Spectral Crest	Spectral Slope	Peak Magnitude (dB)
Pop	868	0.0158	-0.9433	-30.58
Rock	858	0.0153	-0.9793	-30.66
Electronic	845	0.0194	-0.7461	-27.7
Hip-hop	662	0.0265	-0.8141	-22.52
Jazz	785	0.0141	-1.2929	-35.58
Folk	603	0.0191	-1.1824	-32.54
Disco	963	0.0148	-0.8042	-30.31
R&B	811	0.0149	-1.0336	-33.87
Soul	760	0.0157	-1.0303	-32.94

3.3 Intelligent Equalization Algorithms

3.3.1 Target Equalization Spectrum

The average spectrum of all songs in the dataset is shown in Figure 3.1. Popular commercial recordings appear to share a consistent trend, which can be described as a linearly decaying distribution of around 5 dB per octave between 100 and 4000 Hz, becoming gradually steeper with higher frequencies, and a severe low-cut around 60 Hz.

The average spectra can be used as a frequency balance reference as a "best practices" approach. We use a smoothed version of the average spectra as the target equalization spectrum. A 17-point moving average filter is applied to the original spectra. We only perform the smoothing mechanism on frequencies higher than 200 Hz, so that the peak frequency and peak magnitude on lower frequency are preserved, while higher frequency bin values are smoothed to filter out the raggedness (comb-like shape) of the mid-distribution. The smoothed, target equalization curve is illustrated in Figure 3.6.

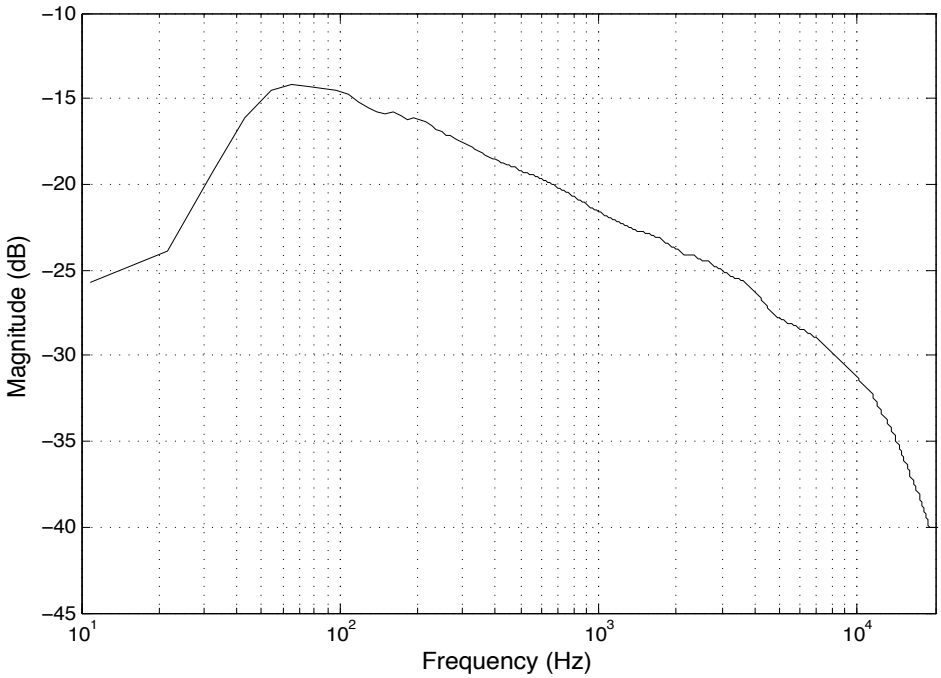


Figure 3.6 Smoothed target equalization spectrum.

3.3.2 System Workflow

After the target equalization spectrum has been found, the next stage is to design an algorithm to filter the input audio signal so that its spectrum matches the target. The overall block diagram of the full system is shown in Figure 3.7. In summary, the filtering process is first controlled by a noise gate, which determines from a frame's energy whether it can be considered to be active. Only active frames enter the filter design stage. For inactive frames, the filter curve is kept stationary. The spectrum of the active frame is then analysed and matched against the target, creating a filter curve using the Yule-Walker method. Filter curves are smoothed within and between frames to minimize the artifacts.

Implementations of the algorithm have been developed in both Matlab and C++. Both operate on a frame-by-frame basis, but the C++ implementation uses a sample-based approach to realize real-time, low latency processing for practical use. The C++ version deploys a host/plug-in structure, where the host defines the frame size.

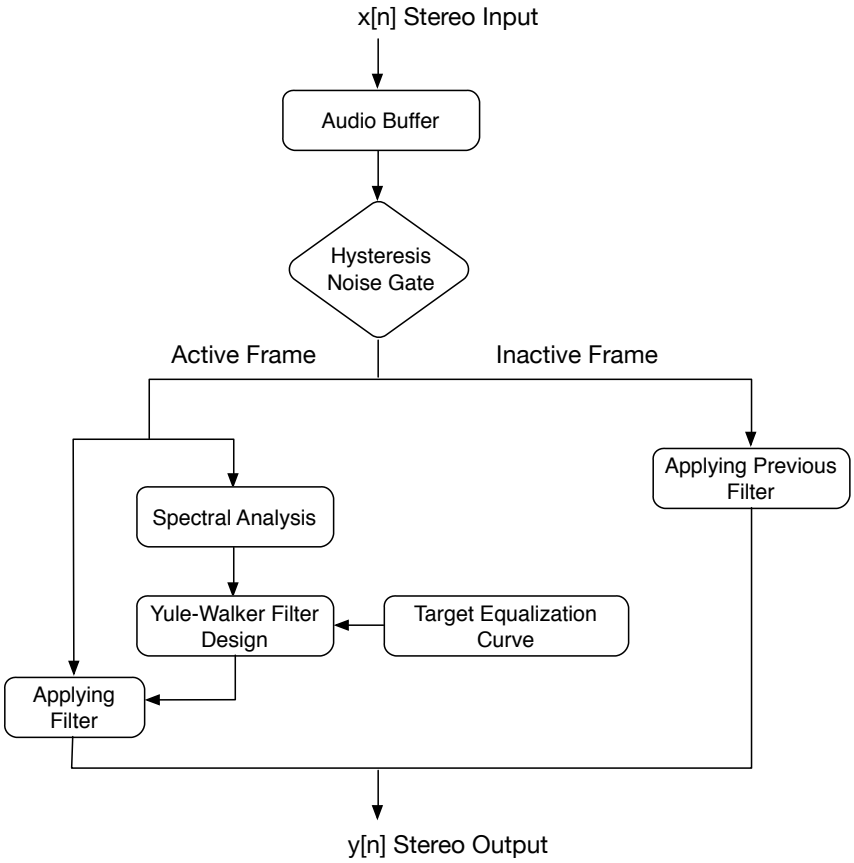


Figure 3.7 Block diagram of the intelligent equalization system.

3.3.3 Hysteresis Noise Gate

We adapt the noise gate with hysteresis algorithm in (Mansbridge et al., 2012) to classify the input frames to be either active or inactive based on their loudness estimated by the R-128 loudness measure (ITU, 2012a).

We assume that a frame must be active to contribute to the next stage, where the filter curve is created and applied to match the target equalization spectrum. If inactive, the same filter curve that was applied on the previous active frame will be applied. Hysteresis thresholds (Filanovsky & Baltes, 1994), T_{open}=-25 LUFS and T_{close}=-30 LUFS are chosen to prevent inappropriate state switching as shown in Figure 3.8.

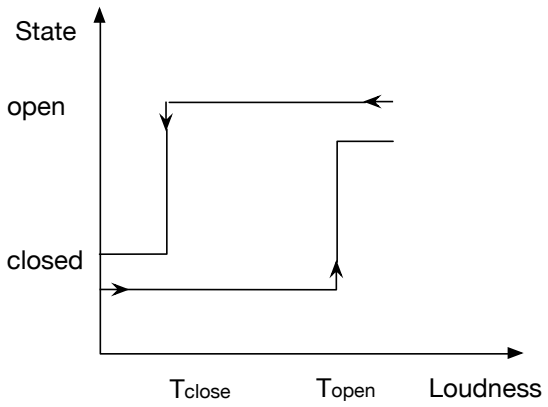


Figure 3.8 Noise gate with hysteresis operation.

3.3.4 Spectral Analysis

A 4096-point sliding FFT with Hanning windows was performed. Since the host itself defines the frame size and it's usually less than 4096 samples, we need to create a buffer of 4096 samples to store enough samples. The FFT was only performed on the active frames in order to cut down the computational cost and achieve a low latency.

Since the calculated magnitude spectrum will act as a denominator in a later analysis stage to obtain the desired transfer function (filter curve), a problem could arise if the amplitude values at one or more frequency components are too small. As a result, we end up with a transfer function with unreasonable peaks, which is difficult to estimate and produces unpleasant sound artifacts. It may also make the IIR filter highly unstable. To avoid this, a simple threshold technique is applied. We opt to use a threshold of -40 dB to filter the magnitude spectrum. Values less than -40 dB are usually found only at very high frequencies. So the thresholding mechanism will have an insignificant effect on the accuracy of the filter design. Later, we normalized the spectra by dividing the magnitudes by the maximum magnitude for spectrum comparison.

3.3.5 IIR Filter Design

The design of an IIR filter with arbitrary magnitude response using the Yule-Walker least-squares approach is described in this section.

Step 1: Obtain Desired Magnitude Response

Let $X(w)$ denote the thresholded, normalized magnitude spectrum of the active frame, and $T(w)$ denote the target equalization spectrum. Therefore, the desired transfer function $H_d(w)$ can be simply obtained from the equation:

$$H_d(w) = \frac{T(w)}{X(w)}. \tag{0.0}$$

The values of $H_d(w)$ are calculated at every 1/3 octave center frequency, which is closely approximate the perception of sound by human hearing system. 33 frequency bands are large enough to capture the transition of the impulse response with an arbitrary shape while the computational cost is reduced significantly compared with the one of using 2048 linear-spaced frequency points. Afterwards, we normalize $H_d(w)$ into the range (0,1) to prevent overshooting.

In the practical implementation, the actual values of $T(w)$ are weighted values between the target spectrum and magnitude spectrum of the processed frame defined as follow:

$$T'(w) = T(w)a + (1-a)X(w) \quad a \in [0,1]. \tag{0.0}$$

The weighting factor a is left as one of the user control parameters: increase the value of a to match the target spectrum more or decrease it to preserve the original spectral content more, based on their personal listening evaluation.

Step 2: Filter Curve Smoothing

As the algorithm produces time-varying filter curves operating on audio signals, variable smoothing on desired IIR filters' magnitude responses within a single frame and between adjacent frames is necessary to avoid sound artifacts. Since the intelligent equalization tool runs in real-time with a sample-based approach, an efficient and reliable long-term average measure is necessary throughout to produce useful and smoothly varying data variables. Exponential moving average (EMA) filters are used extensively to fulfill this role. The EMA filter is a first order IIR filter described by the general difference equation:

$$X'(t) = \alpha X'(t-1) + (1-\alpha)X(t), \tag{0.0}$$

$$\alpha = e^{-1/(\tau f_s)}. \tag{0.0}$$

f_s is the sample rate of the input signal, and α determines the degree of filtering between adjacent samples: the higher the value the less the rate of decay. τ corresponds to the time that takes the system to reach (1-1/e) of its final value. Let W denote the window size, which is decided by the host itself. Then the actual processing frequency rate becomes:

$$f_W = \frac{f_s}{W}. \tag{0.0}$$

EMA filters are first applied to the desired magnitude response $H_d(w_n)$ of the filter curve within one active frame as follows:

$$H'_d(w_n) = \alpha_1 H'_d(w_{n-1}) + (1-\alpha_1)H_d(w_n). \tag{0.0}$$

In this case, $\alpha_1 = e^{-1/(\tau_1 f_W)}$ and τ_1 is set to 0.5 (ms) based on empirical experiments.

EMA filters are also applied to smooth the overall variations of filtering curves between consecutive frames:

$$H'_m(w) = \alpha_2 H'_{m-1}(w) + (1-\alpha_2)H_m(w), \tag{0.0}$$

where $\alpha_2 = e^{-1/(\tau_2 f_W)}$ and $H'_m(w)$ corresponds to the new value of $H_m(w)$ for current frame m. $H'_{m-1}(w)$ denotes the transfer function value for previous frame (m-1). New filter curves are calculated once every frame. τ_2 is set to 1.28 (s) for a typical frame size W=64, to prevent filter curves from changing wildly from frame to frame.

The choices of the time constant (τ_1 and τ_2) for the filter curve smoothing mechanism within one active frame and between consecutive frames are particularly important to tackle the potential inter-frame spectral variation, which might produce undesired artefacts. Listening evaluation of the algorithm on various songs with different time constants suggests

that the time constant for consecutive frames smoothing (τ_2 > 1 s) can prevent such unpleasant artefacts in most cases. Furthermore, in the real-time implementation of the algorithm, time constants are set to the optimal values as mentioned. However, they are user-controllable. User can adjust the time constants to tailor the algorithm for each individual song.

Step 3: Obtain IIR Filter Coefficients Using Yule-Walker

We adapt the Yule-Walker method to perform a least-squares fitting to the desired frequency response $H_d(w)$ to find a causal stable rational function:

$$H(z) = \frac{B(z)}{A(z)}, \tag{0.0}$$

which best approximates $H_d(w)$. The Yule-Walker method finds the p-th order recursive filter coefficients B and A such that the filter:

$$\frac{B(z)}{A(z)} = \frac{b(0) + b(1)z^{-1} + ... + b(p)z^{-p}}{1 + a(1)z^{-1} + ... + a(p)z^{-p}}, \tag{0.0}$$

where $\{b(0), \ldots, b(p)\}$, $\{a(0), \ldots, a(p)\}$ are the denominator and numerator coefficients of the desirable IIR filter, and $a(0)$ equals to 1. The denominator coefficients are calculated by the modified Yule Walker equations (MathWorks, 2015) using correlation coefficients computed by inverse Fourier Transformation of the specified frequency response $H_d(w)$. The Yule Walker method is summarised as follow, the detailed calculation of the numerator can be found in (MathWorks, 2015):

- Step 1: A numerator polynomial corresponding to an additive decomposition of the power frequency response is computed.
- Step 2: The complete frequency response corresponding to the numerator and denominator polynomials is evaluated.
- Step 3: A spectral factorization technique is used to obtain the impulse response of the filter.

Step 4: The numerator polynomial is obtained by a least-squares fitting to this impulse response.

p is set to 16 based on listening evaluation of the quality of the mixes produced by the algorithm with different p values by the author, and objective evaluation (see Section 3.4). IIR filter of a slightly high order gives us a good approximation and does not cause any latency problems.

3.3.6 Filter Applying

We filter the audio samples with an IIR filter described by its denominator coefficients {$b(0)$, …, $b(p)$} and numerator coefficients {$a(0)$, …, $a(p)$}. The filtering process is implemented as a difference equation:

$$\begin{aligned} y(n) = {} & b(1)x(n) + b(2)x(n-1) + \ldots + b(p)x(n-p) \\ & - a(2)y(n-1) - \ldots - a(p)y(n-p), \end{aligned} \tag{0.0}$$

where $x(n)$ is the current input audio sample, $y(n)$ is the current output. Since the deployed host/plug-in structure operates on frame-by-frame basis. Two audio buffers, one to store previous values of $x(n\text{-}1)$ to $x(n\text{-}16)$, another to store previous values of $y(n\text{-}1)$ to $y(n\text{-}16)$, are needed to realize the filtering process across consecutive frames.

3.4 Results and Evaluation

A straightforward objective evaluation to compare of the before-and-after magnitude spectrums of the signal is presented.

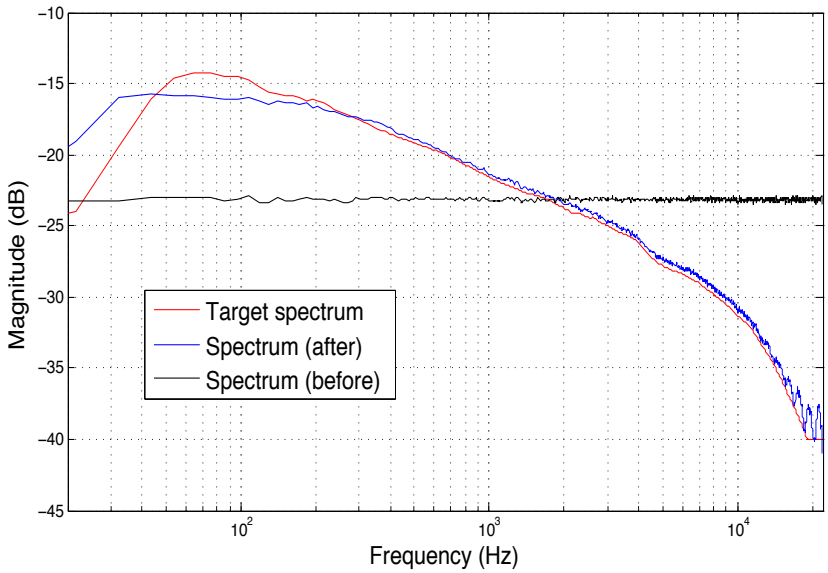


Figure 3.9 Before-and-after magnitude spectrums of a white noise signal compared with the target spectrum.

First, we applied our equalization algorithm to a white noise signal. The result is shown in Figure 3.9. Overall, it shows that the algorithm is able to match the spectrum of the white noise signal to the target equalization curve. However, notable errors are appeared at low frequencies.

We also tested on an uncompressed musical signal at a typical 44.1 kHz sampling rate. The musical signal (Elvis Presley's "It's Now or Never") is one of the commercial songs used in Section 3.2. 20s segment of the song was extracted and tested. The result is presented in Figure 3.10. The result shows the output spectrum matches to the target equalization curve roughly. The effect of the algorithm is particularly obvious at high frequencies. The order p of the IIR filter is set to 16, and τ_1, τ_2 are set to optimal values for both tests.

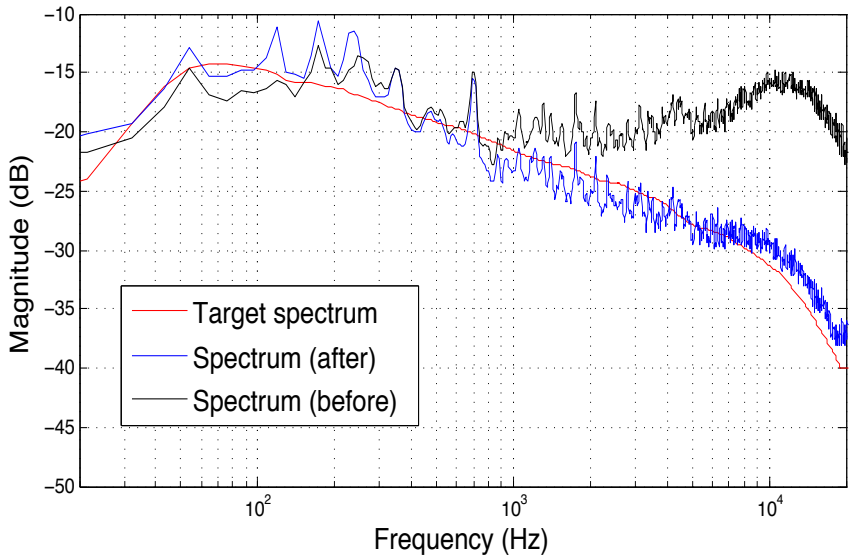
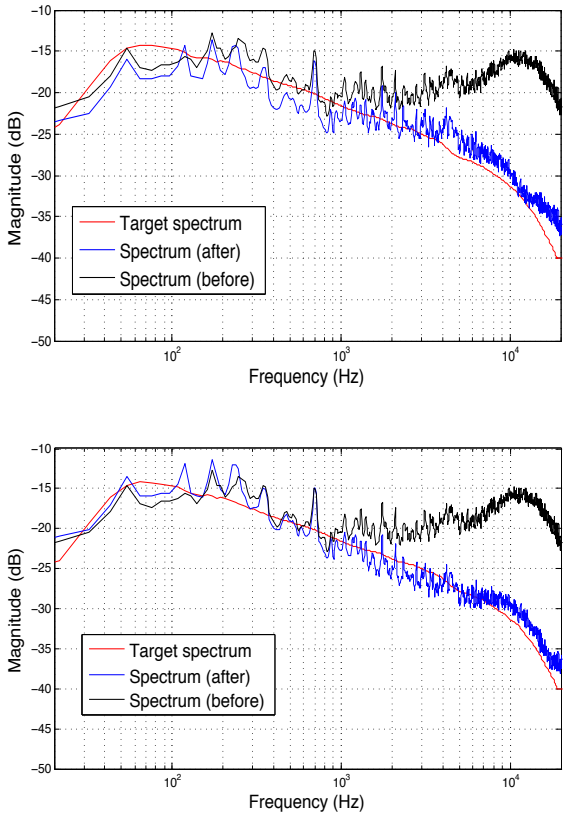


Figure 3.10 Before-and-after magnitude spectrums of a musical signal compared with the target spectrum.

To choose the optimal IIR order for real time implementation, we evaluate the performance of the algorithm briefly with different IIR orders. Same musical signal as previous experiment was used. The before and after spectrum results in terms of different IIR orders are shown in Figure 3.11.



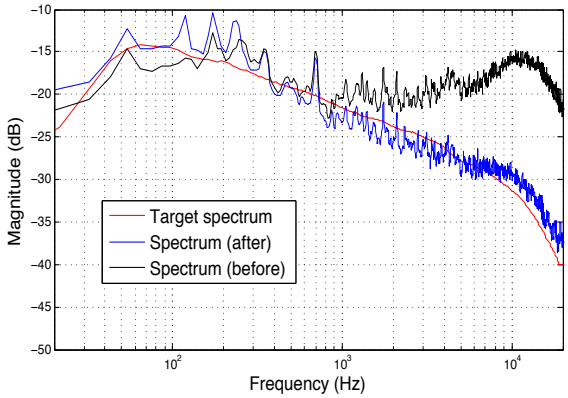


Figure 3.11 Results of IIR orders with 8, 16 and 32 respectively from top to bottom.

The results suggested that IIR order choice of 16 has similar performance as higher order 32, and much accurate spectrum matching ability comparing to low order of 8.

We also compared our approach against an alternative target equalization implementation provided by Landr, Ltd. In brief, the alternative equalization applies 9-band FIR filters with each sub-band gains computed by comparing the target spectrum and the input spectrum at each sub-band. The time-varying alternative algorithm is based on traditional fixed-band equalizers. First, the same white noise signal was fed into both plug-ins. All control parameters are set to optimal values. The results are depicted in Figure 3.12.

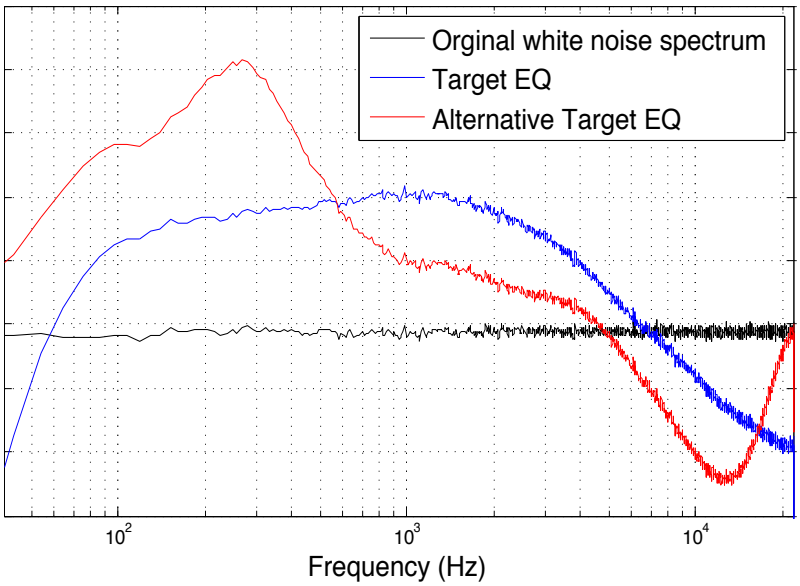


Figure 3.12 Output spectrums obtained from the proposed target equalization approach and an alterative equalization approach against the original spectrum of a white noise signal.

The spectrum curve obtained from the alternative target equalization shows relatively sharp peaks around 250 Hz and 10 kHz with an up-climbing slope at the high end possibly owing to the fact that it uses fixed frequency bands equalization method. The spectrum curve produced by our target equalization approach sustains a flat response at middle range and constant exponential decrease at both low and high end. Regarding spectrum matching toward a specific target, the Yule-Walker method shows its advantage over fixed frequency bands limitation.

Following the same process, a musical signal was also tested. The results are presented in Figure 3.13. The averaged output spectrums of the song after being processed by both approaches appear to lie close to each other in general. The zoom-in difference between these two approaches is depicted in Figure 3.14. We can see irregular variations across the whole frequency range.

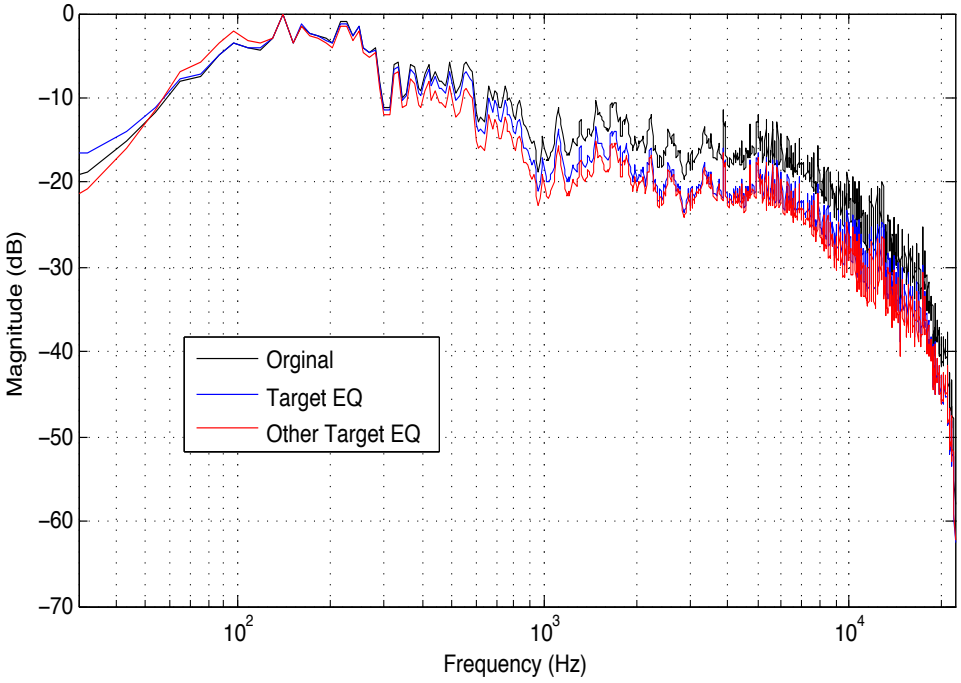


Figure 3.13 Output spectrums obtained from the proposed target equalization approach and an alterative equalization approach against the original spectrum of a musical signal.

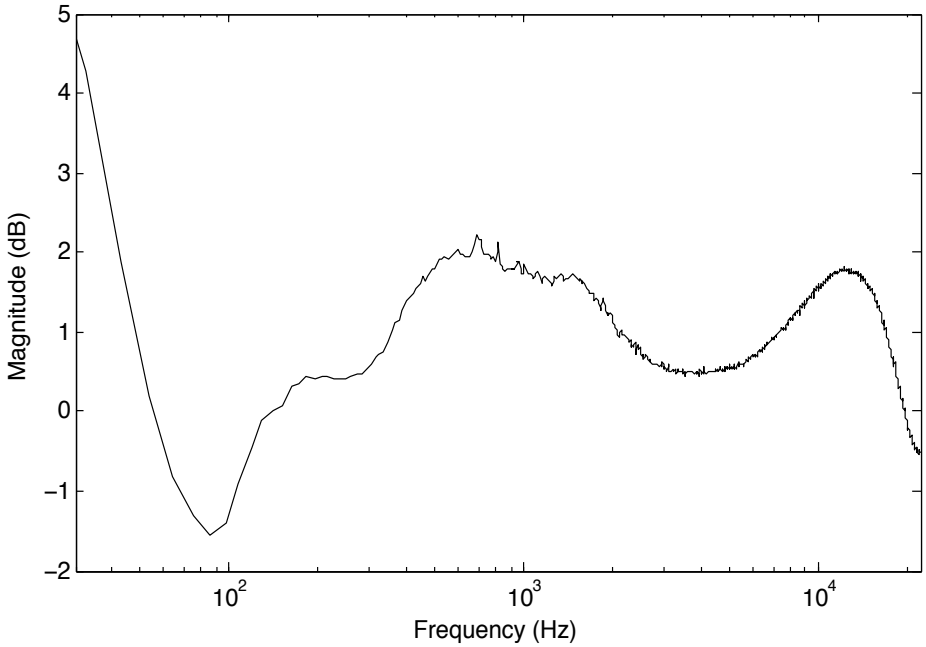


Figure 3.14 The difference between the spectrums obtained from the proposed target equalization approach and the alternative equalization approach.

3.5 Conclusions

A spectral characteristic analysis was performed on a dataset is comprised of almost half the number one recordings over the past 60 years. It showed that the spectra of these popular commercial recordings share a consistent trend, which can roughly be described as a linearly decaying distribution of around 5 dB per octave between 100 and 4000 Hz, becoming gradually steeper with higher frequencies, and a severe low-cut around 60 Hz. It also suggested that the shapes of the spectra are dependent on genre and on the year of production. However the analysis was performed on monaural content. The difference exhibited between the left and right channels is another interesting topic yet to be explored. Analysis on the spectral difference between the original songs and their modern remastered versions is another direction for future work, which can offer a comprehensive insight into modern mixing technique. In general, the broad statistical analysis of successful commercial recordings shows a lot of promise for knowledge that could be useful for intelligent mixing system.

We then proposed a novel time-varying equalization approach to match the spectral distribution of the input signal to a target equalization curve (such as the common curve

obtained from the spectral characteristic studies) or any desired frequency response, based on the Yule-Walker IIR filter design method. Objective evaluation of the algorithm showed that the algorithm is able to fulfill the objective with appropriate ballistics setting.

The limitation of this equalization approach is that it applies IIR filters on the mix rather than the individual tracks. Therefore, this approach is more applicable to audio mastering than mixing at its current state. Future work to explore how to apply Yule-Walker IIR filter to individual tracks to achieve a target spectrum is desirable.

Chapter 4

Dynamic Processing

4.1 Introduction

Chapter 3 dealt with the frequency aspect of intelligent mixing, in which we have proposed an intelligent equalization technique to manipulate the spectral content of audio signals to match a target spectrum discovered from analysis of a large dataset of successful commercial recordings. In this chapter we proceed to the dynamics aspect of intelligent mixing.

Dynamic range compression (DRC) in multitrack mixing has been discussed in Section 2.5.4. The rich history of automatic DRC research has been reviewed in Section 2.5.4. To a large extent, DRC defines much of the sound of contemporary mixes. However, it is arguably the most misused and overused effect in audio mixing (Izhaki, 2013). If used excessively, the dynamic range compressor suppresses musical dynamics, producing lifeless recordings deprived of their natural character. Inappropriate parameter settings also produce artifacts such as pumping and breathing. Furthermore, conventional use of a static set of compressor parameters might not be optimal when the dynamic characteristics of the signal vary significantly over time. Parameter automation of a dynamic range compressor using computerized signal analysis can provide advantages to audio amateurs or musicians who lack expert knowledge in signal processing. Such tools are capable of producing intelligent mixing decisions that speed up the routine work and the trial-and-error process of avoiding inappropriate sonic artifacts."

In this chapter, we propose a fully automated multitrack DRC algorithm exploiting the interdependence of the input audio features and incorporating best practices as control rules. Section 4.2 provides control assumptions to automate the system and the rationale as to why the proposed features explored in Section 4.3 are relevant. A method of adjustment experiment is described in Section 4.3 to explore the subjective preference for ratio and

threshold parameter setting, and multiple linear regression models are then applied to the results to derive the ratio and threshold automations. Finally, the intelligent multitrack DRC algorithms are presented in Section 4.4 followed by a subjective evaluation in the form of a listening test and discussion. The procedure is illustrated in

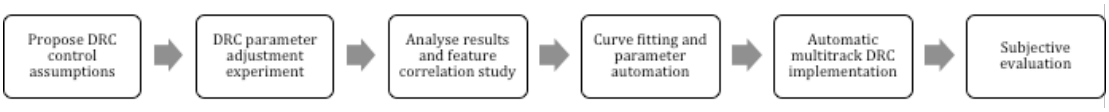


Figure 4.1 The development of the automatic multitrack DRC algorithm.

4.2 DRC Control Assumptions

DRC control assumptions, derived from the literature and analysis to automate the compressor parameters are listed and discussed.

- Assumption 1: A signal with a high degree of level fluctuations should have more compression.
- Assumption 2: A signal with more low frequency content should have more compression.
- Assumption 3: Attack and release time should be dependent on the transient nature of the signal.
- Assumption 4: Knee width should depend on the amount of compression applied.
- Assumption 5: Make-up gain should be set so that output loudness equals input loudness.
- Assumption 6: There is a maximum and optimal amount of DRC that depends on sound source features.

Regarding Assumption 1, in a survey about the main reasons to apply DRC (Pestana, 2013; Pestana & Reiss, 2014), most professional mixing engineers who participated stated that their main intention was to 'stabilise erratic loudness range'. They often compress instruments that have high note-to-note level variations, such as vocals or drum tracks, so that their relative levels are more consistent. A number of dynamics features have been proposed recently that measure the degree of level fluctuation, including EBU loudness range (ITU, 2012a) and dynamic spread (Vickers, 2001), which is simply the p-norm of the signal. Yet subjective listening test results in (Boley, Danner, & Lester, 2010) suggested none of the

metrics accurately predict the perceived dynamic range of a musical track. In (Pestana, Reiss, & Barbosa, 2013), the authors proposed some parameter alterations of (ITU, 2012a) that might yield better results for multitrack material. Alternatively, the crest factor, calculated, as the peak amplitude of an audio waveform divided by its RMS value, can also be a coarse measurement of dynamic range.

Assumption 2 is based on analysis of mixes in (Pestana, 2013; Pestana & Reiss, 2014), which showed that 'Compression takes place whenever headroom is at stake, and the low-end is usually more critical'. Thus spectral features of the source audio signal such as spectral centroid, spectral spread, and brightness are worth exploring to reveal the degree of frequency dependence and low-end sensitivity of DRC.

As for Assumption 3, attack times usually span between 5 ms and 250 ms and release times are often within the 5–3000 ms range. It is generally accepted that attack and release time parameters are employed to catch the transient nature of the sound (Izhaki, 2013; Kraght, 2000). Some commercial compressors offer a switchable auto-attack or auto-release, which are mostly based on measuring the difference between the peak and RMS levels of the side-chain signal. In academia, (Aichinger et al., 2011) automate attack and release times based on the crest factor of the multitrack. More recently, (Giannoulis et al., 2012b) improved the subject based on either modified crest factor or modified spectral flux. The outcome was used in (Maddams et al., 2012). Previous research shares a general idea: if a signal is highly transient or percussive, shorter time constants are preferred.

Regarding Assumption 4, a soft knee enables smoother transition between non-compressed and compressed parts of the signal, and thus yields a more transparent compression effect. In order to produce a natural compression effect in an automatic mixing system, the knee width should be adaptively configured based on the estimated amount of compression applied on the signal (Reiss, 2011). The amount of compression applied largely depends on the relationship between threshold and ratio.

Assumption 5 can be regarded as a direct consequence of the definition of make-up gain. Automatic make-up gain based on the average control-voltage is commonly used in commercial DRC products. However, (Izhaki, 2013) pointed out that this often produces a perceived loudness variation in practice. Subjective evaluation in (Giannoulis et al., 2012b)

showed that the EBU loudness-based make-up gain produced a better approximation of how professional mixing engineers would set the make-up gain.

As for Assumption 6, quantitative descriptions about the amount of compression that should be applied on different instruments can be found in the literature (Huber & Runstein, 2013; Thiele, 2005). Both (Giannoulis et al., 2012b) and (Foudi, 2012) separated between transient and steady state signals as they are assumed to need a different treatment.

4.3 Compressor Parameter Adjustment Experiment

Ratio and threshold are the most crucial parameters in determining the amount of DRC. Assumption 1 and 2 (Section 2.2) suggest that audio features that describe the dynamic and spectral content of the signal might have a high degree of correlation with the preferred amount of DRC. We propose a method of adjustment experiment to uncover how subjects set the ratio and threshold. Several feature candidates are proposed and their correlations with the subjective results are analysed. We apply a least-squares based multiple linear regression model to formulate the relationship between the identified features and the test results, and finally to derive the ratio and threshold parameter automation.

4.3.1 Method of Adjustment Experiment

Four multitrack songs of different genres (Song 1: Rock; Song 2: Pop; Song 3: Alternative; Song 4: Folk) were selected for testing. 20-second excerpts were extracted from the chorus of each song for use in the test. Each excerpts consisted 6 or 7 different instrument stems (a sub-mix of the tracks that represent the same instrument in the process of mixing), all in mono and running at a typical sampling rate of 44.1 kHz. The loudness of the songs were normalised manually based on subjective listening rather than objective loudness measurement, by a group of professional mixing engineers as suggested in (ITU, 2003). This is to ensure that all songs are perceived equally loud when they were played at the same playback system (around 80 dB SPL playback level) used in the adjustment experiment. In this case, gains were applied to the overall mix rather than each individual instrument to achieve equally loudness between songs. No peak normalisation processes were applied for each individual instrument track. Different gains values have to be applied to each track with

the objective to achieve peak normalization. However, this will produce unpleasant level balance between instruments (professional mixing engineers rarely perform peak normalise to achieve level balance). And ill-balanced mixes introduce psychological bias and interference for subjects to correctly perform subjective evaluation of the dynamics of the songs. The specifications such as instrumentation, RMS levels and etc. of the testing audios can be found in Table 4.3. All songs used in this experiment can be accessed from the Open Multitrack Testbed at multitrack.eecs.qmul.ac.uk (De Man, Mora-Mcginity, Fazekas, & Reiss, 2014). This experiment is to explore the participants' subjective preference of threshold and ratio settings when presented with various signals that have different feature characteristics. Therefore signals with wider range of audio feature (such as crest factors, RMS, dynamic spread, spectral centroid, spectral spread, brightness and etc. See Table 4.3) are selected for the tests.

Fifteen participants, all of whom have audio engineering experience, two of whom are professional mixing engineers, were recruited to perform a DRC ratio and threshold adjustment experiment. Related information about the participants is shown in Table 4.1.

Table 4.1 Related information about the participants.

Gender	Male	10
	Female	5
Critical listening skill	No experience	0
	Moderate	11
	Professional training	4
Hearing impairment?	No	15
	Yes	0
Mixing background	Audio Researcher	8
	Amateur mixing engineer	5
	Professional mixing engineer	2

The author is aware that audio experience does not guarantee best practice for compression tasks. Especially compression requires training and professional experience. The result can be improved with more recruitment of professional mixing engineers to validate the result as "best practice".

All tests were performed in a soundproof listening room with the same headphone set-up, where the environmental noise is minimized. Participants were allowed to adjust the playback level during the experiment in order to evaluate the dynamics efficiently. Participants were asked to adjust the ratio and threshold parameters for each instrument track of each song until they were satisfied with the amount of DRC applied to the mix. A solo function to play back an individual track with or without compression was provided in the experiments. However, subjects were advised to listen to the mix when setting the parameters for each individual track.

The digital compressor model design employed in the experiment is a feed-forward compressor with smoothed branching peak detector (Giannoulis et al., 2012a). The ratio and threshold values were hidden from participants to prevent bias resulting from common practices. Other compressor parameters were automated as described in Section 4.4. The interface used for this experiment is shown in Figure 4.2.

Figure 4.2 Interface for the ratio and threshold adjustment experiment.

The normality of the result is checked with Lilliefors test (Lilliefors, 1967), which has a statistic:

$$L = \sup_{x} \left| scdf(x) - cdf(x) \right|, \tag{0.0}$$

where *scdf* is the empirical sample-based estimation of the cumulative distribution function and *cdf* is the normal cumulative distribution function with mean and standard deviation equal to those of the sample. This is an appropriate approach for unknown specifications of the null distribution, which is our case.

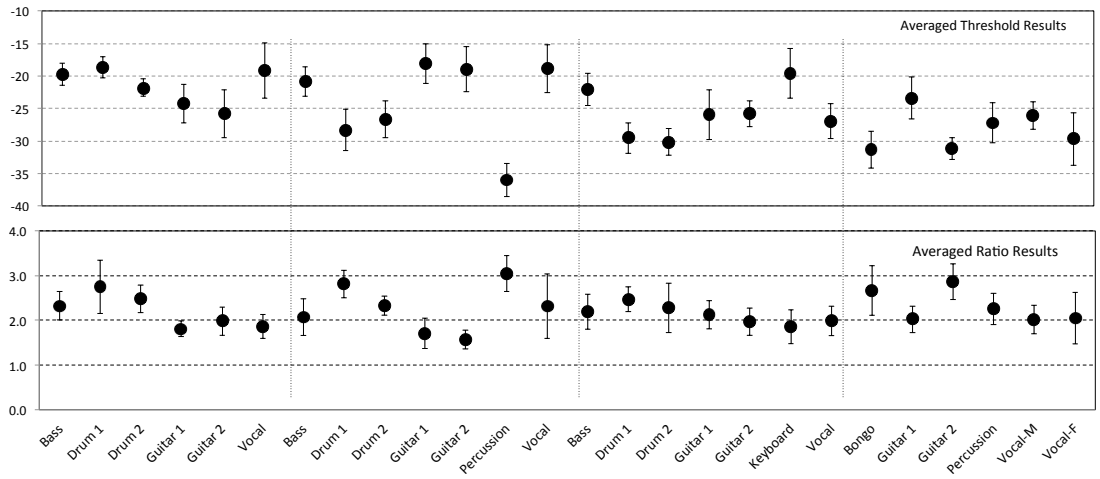
The results of the normality test for each instrument of each song, together with the *p*-value are shown in Table 4.2. The results suggest that more than half of test cases (37 out of 52) do not reject the null hypothesis (p>0.05).

Table 4.2 Normality test results for each instrument of each song with *p*-value included, h is the hypothesis test result (h = 1 to indicate rejection of the null hypothesis that the experiment results come from a distribution in the normal family, at the 5% significance level; h=0 to indicates a failure to reject the null hypothesis at the 5% significance level).

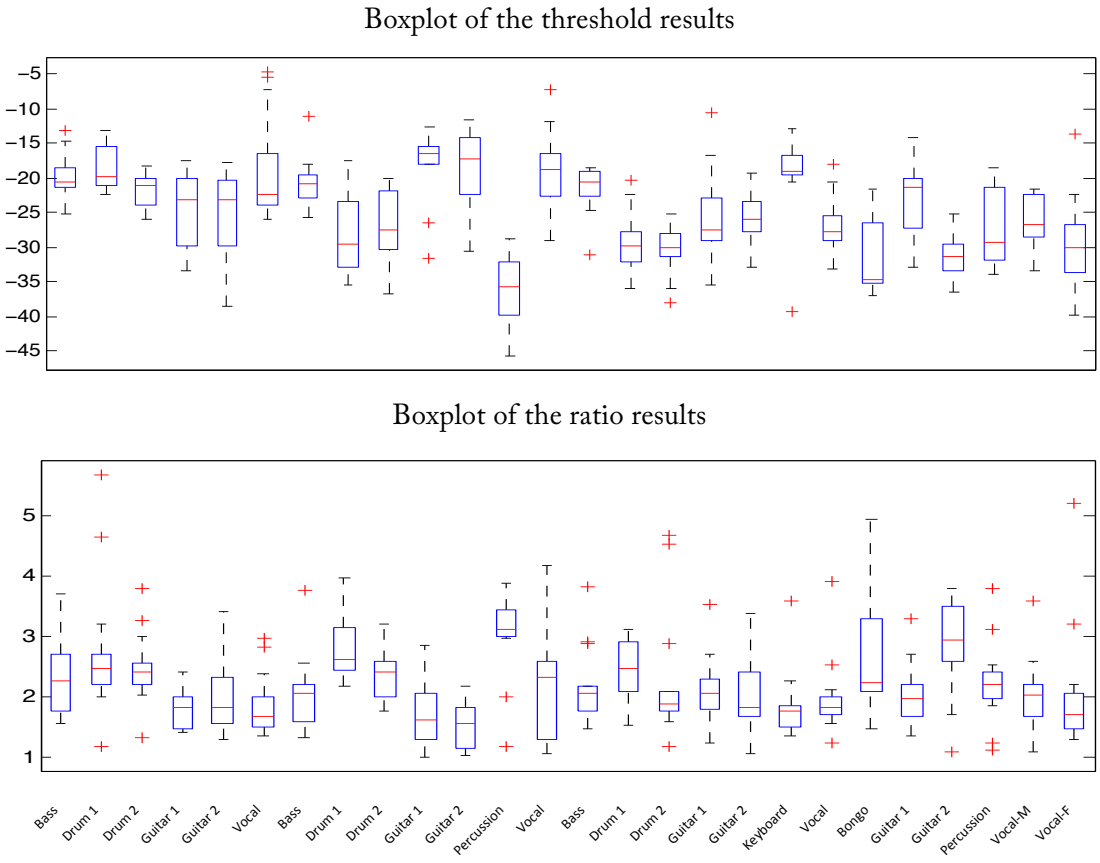
			Threshold		Ratio	
Song	Genre	Instrument	h	p-value	h	p-value
		Bass	0	0.2982	0	0.5
		Drum 1 (Drum set)	0	0.1997	1	0.0016
		Drum 2 (Drum set room)	0	0.2591	0	0.0506
		Guitar 1 (Electric)	0	0.3333	0	0.4212
		Guitar 2 (Electric)	0	0.118	1	0.0267
1	Rock	Vocal (Male)	1	0.0029	1	0.0329
		Bass	0	0.5	0	0.3491
		Drum 1 (Drum set)	0	0.1792	1	0.0055
		Drum 2 (Drum set room)	0	0.1214	0	0.5
		Guitar 1 (Acoustic)	1	0.002	0	0.288
		Guitar 2 (Acoustic)	0	0.2187	0	0.5
		Percussion	0	0.5	1	0.0015
2	Folk	Vocal (Female)	0	0.5	0	0.3174
		Bass	1	1.00E-03	1	0.0081
		Drum 1 (Drum set)	1	0.4133	0	0.4963
		Drum 2 (Drum set room)	0	0.1713	1	1.00E-03
		Guitar 1 (Electric)	0	0.0699	0	0.5
		Guitar 2 (Electric)	0	0.5	0	0.1731
		Keyboard	1	1.00E-03	1	0.0243
3	Indie	Vocal (Male)	0	0.1239	1	0.0057
		Bongo	1	0.0081	1	0.0088
		Guitar 1 (Electric)	0	0.394	0	0.5
		Guitar 2 (Acoustic)	0	0.5	0	0.5
		Percussion	0	0.1706	0	0.1587
4		Vocal-M	0	0.2254	0	0.3866

Pop	Vocal-F	0	0.4094	1	0.0011	

The specification for MUSHRA (ITU, 2003) codec quality tests, which are quite similar to ours, offers the simplification that no overlap in the confidence intervals for two conditions means one is significantly better than the other. We will follow this idea whenever it is clear, presenting the standard plots. The average mean results of the 15 participants for ratio and threshold for each track, along with 95% confidence interval, together with the standard boxplots results are shown in Figure 4.2. The small variations in results were unexpected since dynamic range compression is often assumed to be an art, with varying tastes in its application. However, we can also see from Figure 4.2 that different tracks have differing variation sizes, suggesting that DRC parameter setting is track dependent. We further note that half of the participants are from the same UK research group, and thus might share a similar taste in compression that could potentially bias the results.



(a)



(b)

Figure 4.3 (a) Ratio and threshold adjustment results with 95% confidence interval, dotted vertical lines separate results between songs. (b) Boxplots of the ratio and threshold adjustment results.

4.3.2 Feature Correlations

Several dynamic and spectral features are proposed, extracted and analysed based on the subjective results. First, the RMS level as a rough dynamic feature is defined by,

$$x_{RMS} = 20\log_{10}\sqrt{\frac{1}{N}\sum_{n=0}^{N-1}x^2(n)}, \tag{0.0}$$

where N is the window length.

EBU loudness range (LRA) is defined as the difference between the 10th percentile and the 95th percentile on the histogram after a dual gating process (ITU, 2012a).

Dynamic spread (Vickers, 2001) is given by

$$d = \frac{1}{N}\sum_{n=0}^{N-1}\left|x_{dB}(n) - x_{RMS}\right|, \tag{0.0}$$

where x_{dB} is the input signal in digital full scale (dBFS).

The spectral centroid is the barycentre of the spectrum, calculated by

$$\mu = \frac{\sum_{k=0}^{K-1} f(k)X(k)}{\sum_{k=0}^{K-1} X(k)}, \tag{0.0}$$

where $X(k)$ represents the spectral magnitude of signal $x(n)$, of bin number k, and $f(k)$ represents the centre frequency at that bin.

Spectral spread represents the spread of the spectrum around its mean value is defined as

$$\sigma_s^{\,2} = \sum_{k=0}^{K-1}(X(k) - \mu)^2 f(k). \tag{0.0}$$

The practical calculation of the features mentioned before can be found in (Lartillot, Toiviainen, & Eerola, 2008). We also propose two new, cross-adaptive audio features called percussivity weighting and low-frequency weighting.

Percussivity weighting describes the cross-adaptive relationship amongst all the input signals regarding the degree of level fluctuations and is based on the crest factor values. First, the average value of the crest factor over all tracks is computed as

$$\bar{x}_{crest} = \frac{1}{M}\sum_{m=1}^{M} x_{crest}(m) \tag{0.0}$$

where m is the index of the track number and M is the total number of input tracks. The average crest factor $\overline{x}_{crest}$ is then used as an adaptive threshold for the percussivity weighting $w_p(m)$ calculation. The mapping between $w_p(m)$ and $\overline{x}_{crest}$ is formulated using a modified Gaussian distribution centred around $\overline{x}_{crest}$ by

$$g(x) = ae^{\frac{(x_{crest}-\overline{x}_{crest})^2}{2\sigma^2}}, \tag{0.0}$$

where σ is the standard deviation controlling the width of the 'bell' shape. $w_p(m)$ is formulated empirically as follows,

$$w_p(m) = \begin{cases} e^{\frac{(x_{crest}(m)-\overline{x}_{crest})^2}{2\sigma^2}}, & x_{crest}(m) \le \overline{x}_{crest} \\ 2 - e^{\frac{(x_{crest}(m)-\overline{x}_{crest})^2}{2\sigma^2}}, & x_{crest}(m) > \overline{x}_{crest} \end{cases}, \tag{0.0}$$

σ is set to 2 based on informal testing. Equation (0.0) shows that $w_p(m) \in (0,2)$. The larger the $w_p(m)$ value, the more percussive the track m is. Equation (0.0) guarantees that most values of $w_p(m)$ are centred on the adaptive reference $\overline{x}_{crest}$.

Low-frequency weighting is introduced to describe the relative amount of low-frequency energy of each signal compared to the average low frequency ratio. A Fast Fourier Transform (FFT) with Hanning window is performed on each signal frame to obtain the spectral distribution, $X(m,k)$ of track m at frequency bin k. $X_{low}(m,k)$ is the spectral distribution of low-pass filtered version of input signals with cut-off frequency set to 1 kHz, the cross-adaptive low-frequency weighting, $w_f(m)$ is defined by Equation (0.0),

$$w_f = \frac{\sum_{k=0}^{K-1} \frac{X_{low}(m,k)}{X(m,k)}}{\frac{1}{M}\sum_{m-1}^{M}\sum_{k=0}^{K-1} \frac{X_{low}(m,k)}{X(m,k)}}. \tag{0.0}$$

The values of each described feature are extracted from each multitrack and shown in Table 4.3.

Table 4.3 Selected feature values of tested multitrack songs.

Song	Track	Percussivity weighting	EBU loudness range (LU)	Dynamic Spread	RMS (dB)	Low-Frequency weighting	Brightness	Spectral Centroid (Hz)	Spectral spread (Hz)
1	Bass	0.36	1.17	1.31	-15	1.37	0.026	373.1	1315.8
	Drum1	0.86	1.69	8.09	-17.5	1.38	0.539	3479.7	3906.4
	Drum2	1.52	2.09	3.39	-18.4	0.85	0.694	4394.7	3973.4
	Guitar1	1.15	0.47	0.88	-14.6	0.75	0.458	1762.1	1697.7
	Guitar2	1.07	0.77	0.97	-17.2	0.56	0.549	2140.9	1987.3
	Vocal-M	0.67	3.54	8.08	-19.8	0.72	0.592	4313.2	4219.6
2	Bass	0.52	1.88	2.47	-16.8	1.13	0.049	476.5	1331.6
	Drum1	0.93	3.91	10.65	-26.3	1.35	0.360	1927.8	2834.3
	Drum2	0.99	4.85	9.79	-25.6	0.93	0.440	2051.3	2665.5
	Guitar1	0.34	7.80	3.88	-13.6	1.03	0.196	836.5	1358.3
	Guitar2	0.36	0.65	2.02	-12.4	0.58	0.257	1087.3	1378.8
	Percussion	1.98	0.77	3.19	-31.9	1.29	0.997	3082.4	4115.7
	Vocal-M	0.63	6.08	8.38	-21.2	0.62	0.429	3354.7	4376.5
3	Bass	0.6	3.02	6.14	-17.2	1.38	0.105	683.7	1922.6
	Drum1	0.64	5.41	12.04	-25.1	1.63	0.488	3512.2	4245.3
	Drum2	0.97	6.83	10.58	-24.3	0.94	0.525	3563.6	4396.2
	Guitar1	0.99	2.00	1.26	-22.1	0.88	0.365	1458.2	1648.4
	Guitar2	1.13	3.67	1.96	-21.2	0.48	0.597	1987.3	1778.5
	Keyboard	1.01	3.91	2.67	-15.6	0.41	0.317	1573.1	2258.3
	Vocal-M	0.96	10.20	17.92	-21.6	0.79	0.305	1908.5	2628.4
4	Bongo	0.98	1.61	13.93	-26.2	0.92	0.222	1473.9	2406.1
	Guitar1	1.01	2.97	2.50	-30.5	0.93	0.170	1216.7	2691.3
	Guitar2	1.31	4.52	6.33	-18.3	1.25	0.390	2459.7	3695.1
	Percussion	0.76	3.22	26.98	-30.5	1.22	0.388	3082.4	4412.0
	Vocal-M	1.12	4.94	2.80	-25.4	0.88	0.224	1507.0	2798.0
	Vocal-F	0.7	9.30	10.44	-20.1	0.72	0.340	2581.7	4061.4

The cross-correlation coefficient between each feature and the averaged ratio and threshold values across 15 participants is calculated as follows,

$$r_{xy} = \frac{\sum_{i=1}^{M}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{M}(x_i - \bar{x})^2 \sum_{i=1}^{M}(y_i - \bar{y})^2}}, \tag{0.0}$$

where x_i is the feature value, y_i is the observed ratio or threshold value of each multitrack, and $\bar{x}$, $\bar{y}$ are the respective means. The coefficients are listed in Table 4.4.

Table 4.4 Feature correlations against the averaged ratio and threshold values.

	Feature	Ratio Correlation	Threshold Correlation
Dynamic Feature	Percussivity	0.4954	-0.6019
	LRA	-0.1499	-0.1275
	Dynamic Spread	0.2486	0.3294
	RMS level	-0.4871	0.6659
Spectral Feature	Low-Frequency	0.6351	-0.248
	Spectral Centroid	0.3592	-0.2031
	Spectral Spread	0.4996	-0.3571
	Brightness	0.3791	-0.3926

As Table 4.4 shows, spectral features generally exhibited higher correlation with ratio parameter than dynamic features. This is in agreement with Assumption 2. The proposed low-frequency weighting shows the highest correlation with the ratio parameter. The RMS level shows the strongest correlation with threshold. However, in the spectral feature subgroup, all correlations are relatively weak, indicating that dynamic features play a more significant role in setting the threshold parameter than spectral features.

Notice that the perception-based EBU LRA has the lowest correlation coefficients with both ratio and threshold. First, EBU loudness is designed for broadcast material rather than individual tracks in multitrack content. Second, the 3s integration window length is too long to capture small level fluctuations in terms of dynamics.

4.3.3 Curve Fitting

Multiple linear regression techniques are applied to model the relationship between the proposed features and the ratio and threshold experiment results (Lattin, Carroll, & Green, 2003). Combinations of different audio features and various modelling functions are investigated to obtain the best fit by assessing their Goodness-Of-Fit (GOF) statistics, confidence interval and residual plots with validation data.

We investigate various modelling functions with all the feature combinations considered. Ratio curve fits with significant Goodness-Of-Fit are presented in Table 4.5. Insignificant fits are not depicted in the table.

The Sum of Squares due to Error (SSE) is the total deviation of the response values from the fit to the response values, or simply the sum of squares of residuals, calculated as

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \tag{0.0}$$

where y_i is the i^{th} response value from the fit, $\hat{y}_i$ is response value and n is the number of observations. SSE is a measurement of the discrepancy between the data and an estimation model. Generally speaking, smaller SSE suggests a good model fit to the data.

Table 4.5 Ratio curve fitting results with Goodness-Of-Fit statistics.

Feature Selection		Modelling Functions *f(x,y)*	Coefficients	Goodness-Of-Fit			
X data	Y data			SSE	R-square	Adjusted R-square	RMSE
Low-Frequency		$p_1x + p_2$	p_1 =0.7411; p_2 =1.51	2.101	0.4034	0.3785	0.2959
Percussivity		$p_1x + p_2$	p_1 =0.5077; p_2 =1.762	2.657	0.2454	0.214	0.3327
Percussivity	Low-Frequency	$p_{10}x + p_{01}y$	p_{10} =0.969; p_{01} =1.342	2.45	0.3043	0.2753	0.3195
Percussivity	Low-Frequency	$p_{00} + p_{10}x + p_{01}y$	p_{00} =0.968; p_{10} =0.554 p_{01} =0.783	1.078	0.6939	0.6673	0.2165
Percussivity	Low-Frequency	$1 + p_{10}x + p_{01}y$	p_{10} =0.540; p_{01} =0.764	1.079	0.6935	0.6807	0.2121
Percussivity	Low-Frequency	$p_{00} + p_{10}x + p_{01}y + p_{11}xy$	p_{00} =1.108; p_{10} =0.122 p_{01} =1.257; p_{11} =1.108	1.84	0.4776	0.4063	0.2892
Percussivity	Low-Frequency	$p_{00} + p_{10}x + p_{01}y + p_{20}x^2 \ldots + p_{11}xy + p_{02}y^2$	p_{00} =0.933; p_{10} =0.222 p_{01} =1.325; p_{11} =-0.182 p_{20} =0.008; p_{02} =0.008	0.915	0.7401	0.6752	0.2139
Percussivity	Low-Frequency	$p_1 2^x + p_2 2^y$	p_1 =0.4508; p_2 =0.6694	1.337	0.6202	0.6044	0.2361
LRA	Low-Frequency	$p_{00} + p_{10}x + p_{01}y$	p_{00} =1.57; p_{10} =-0.013 p_{01} =0.731	2.07	0.4121	0.361	0.3
Dynamic spread	Low-Frequency	$p_{00} + p_{10}x + p_{01}y$	p_{00} =1.5; p_{10} =0.052 p_{01} =0.714	2.077	0.4101	0.3588	0.3005
Dynamic spread	Low-Frequency	$p_{00} + p_{10}x + p_{01}y + p_{20}x^2 \ldots + p_{11}xy + p_{02}y^2$	p_{00} =1.283; p_{10} =0.043 p_{01} =0.972; p_{11} =-0.003 p_{20} =0.017; p_{02} =-0.159	1.86	0.4719	0.3398	0.3049
RMS	Low-Frequency	$p_{00} + p_{10}x + p_{01}y$	p_{00} =1.117; p_{10} =0.023 p_{01} =0.6.34	1.724	0.5103	0.4677	0.2738
Percussivity	Spectral Spread	$p_{00} + p_{10}x + p_{01}y + p_{20}x^2 \ldots + p_{11}xy + p_{02}y^2$	p_{00} =1.516; p_{10} =0.3844 p_{01} =0.0001	2.196	0.3762	0.322	0.309
Percussivity	Spectral Spread	$p_{00} + p_{10}x + p_{01}\log(y)$	p_{00} =0.754; p_{10} =0.3546 p_{01} =0.7768	2.206	0.3735	0.319	0.309
Percussivity	Spectral Centroid	$p_{00} + p_{10}x + p_{01}\log(y)$	p_{00} =1.31; p_{10} =0.4502 p_{01} =0.1546	2.621	0.2557	0.191	0.3376
Percussivity	Brightness	$p_{00} + p_{10}x + p_{01}y$	p_{00} =1.761; p_{10} =0.4647 p_{01} =0.105	2.651	0.2472	0.1817	0.3395

The coefficient of determination R^2 provides a measure of how well the data are represented, as the proportion of variance, explained by the model. R^2 ranges from 0 to 1, with a value closer to 1 indicating that the model accounts for a greater proportion of variance. The general definition of the coefficient of determination is given by

$$R^2 = 1 - \frac{SSE}{SST}, \tag{0.0}$$

where SST is the total sum of squares proportional to the sample variance, defined as

$$SST = \sum_{i=0}^{n} (y_i - \overline{y}_i) \tag{0.0}$$

where $\overline{y}_i$ is the mean of y_i.

Degrees of Freedom $R^2_{adjusted}$ is generally the preferred indicator to compare two models that are nested. Like R^2, it ranges from 0 to 1, and is given by

$$R^2_{adjusted} = 1 - \frac{SSE(n-1)}{SST(v)}, \tag{0.0}$$

where $v=n-m$, v is the number of independent points involving the n data points that are required to calculate the sum of squares and m is the number of fitted coefficients estimated from the response values (Walker, 1940).

Root-Mean-Square Error (RMSE) estimates the standard error of the regression, as defined by

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}. \tag{0.0}$$

We found that the combination of percussivity and low-frequency weighting generates the best fit regardless of the modelling function evaluated. In general, the modelling functions using percussivity and low-frequency weighting yield an *SSE* smaller than 1.5 and RMSE

smaller than 2.5, while others show a *SSE* larger than 2 and RMSE larger than 3. The results agree with the feature correlation coefficients obtained in Section 4.3.2.

Four models, $f(x,y)=p_{00}+p_{10}x+p_{01}y$, $f(x,y)=p_{10}x+p_{01}y+1$, $f(x,y)=p_{00}+p_{01}y+\ldots+p_{02}y^2$ and $f(x,y)=p_1 2^x+p_2 2^y$ performed a better fit based on the Goodness-Of-Fit statistics. By comparing the Goodness-Of-Fit produced by the first order polynomial $f(x,y)=p_{00}+p_{10}x+p_{01}y$ with $f(x,y)= p_{10}x+p_{01}y$, both use the percussivity and low-frequency weighing features, we see that SSE decreases by more than half and RMSE decreases by roughly 0.1. This means the accuracy of the model improves. Moreover, since the two models are nested, the adjusted R^2 increases significantly from 0.2753 to 0.6673 when adding the additional constant term p_{00}, implying the latter performs better again. The model $f(x,y)=p_{00}+p_{10}x+p_{01}y$ also outdoes $f(x,y)=p_1 2^x+p_2 2^y$ with lower *SSE* and RMSE. The adjusted R^2 of the model $f(x,y)= p_{10}x+p_{01}y+1$ is larger than the one of $f(x,y)=p_{00}+p_{10}x+p_{01}y$, indicating that it excels the latter in the performance of model prediction.

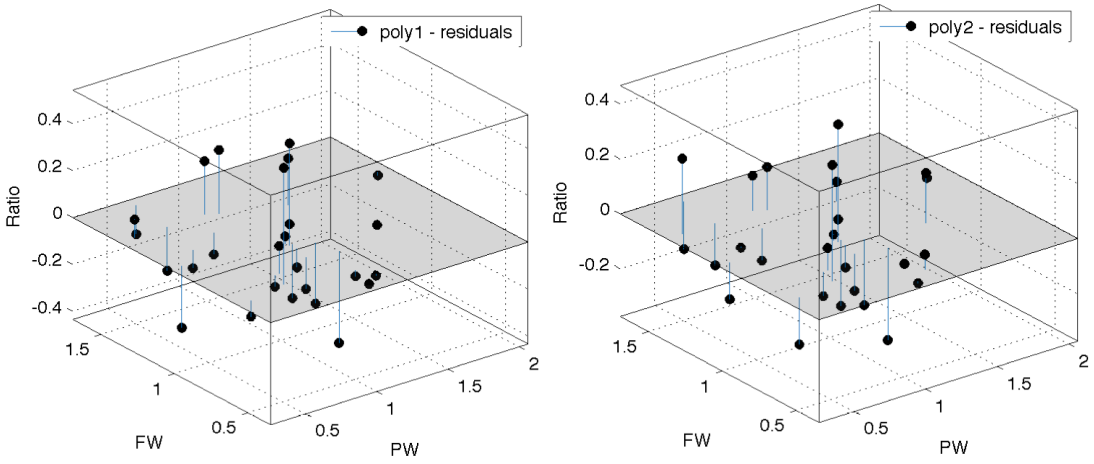


Figure 4.4 Residual plots of the first (left) and second (right) order polynomial models, where proposed low-frequency weighting and percussivity weighting feature are denoted as *FW* and *PW* respectively.

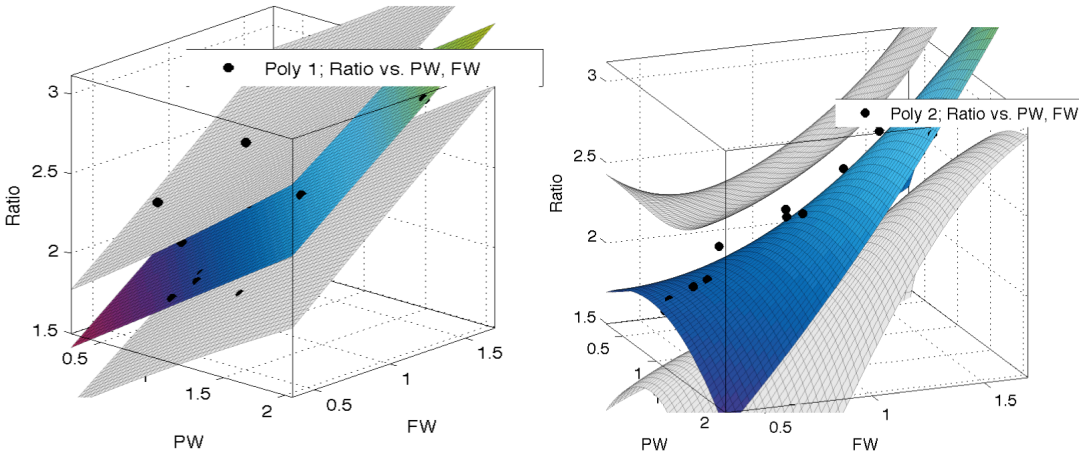


Figure 4.5 Prediction bounds (grey surface) with 95% confidence interval of the first (left) and second (right) order polynomial models.

Although the second-degree polynomial model has slightly larger RMSE, SSE is smaller and R^2 is larger. Since they are not nested, we cannot pick the best fit based on their adjusted R^2 coefficients. Therefore, we plot residuals and prediction bounds to assess both models graphically. The residual plots of the two models are shown in Figure 4.4. Neither residual plot provides exhibits structure, suggesting that both models fit the data to an acceptable extent. The prediction bounds with 95% confidence level are presented in Figure 4.5. The prediction bounds for the first-degree polynomial model with 1 as constant term indicate that the model can be predicted with a small uncertainty (less than 0.8). As for the case of the second-degree polynomial model, it has wider prediction bounds in the area where not enough data exists, suggesting that there is not enough data to estimate the second-degree polynomial terms accurately. In other words, a second order polynomial model overfits the data.

With all criteria considered, $f(x,y)= p_{10}x+p_{01}y+1$ using percussivity and frequency weighting performs the best curve fit. This modelling function is highlighted in Table 4.5.

Table 4.6 Threshold curve fitting results with Goodness-Of-Fit statistics.

Feature Selection		Modelling Functions $f(x,y)=\ldots$	Coefficients	Goodness-Of-Fit			
X data	Y data			SSE	R-square	Adjusted R-square	RMSE
RMS		p_1x+p_2	p_1 =0.5947; p_2 =-12.33	325.8	0.4434	0.4203	3.684
RMS		$p_1x^2+p_2x^1+p_3$	p_1 =-30.29; p_2 =1.341 p_3 =-4.584	319.4	0.4543	0.4069	3.727
RMS		$p1\cdot x^3+p2\cdot x^2$ $\ldots+p3\cdot x+p4$	p_1 =0.0007; p_2 =0.06706 p_3 =-2.409; p_4 =2.676	319.1	0.4549	0.3806	3.808
Percussivity		p_1x+p_2	p_1 =-7.954; p_2 =-17.66	373.3	0.3623	0.3358	3.944
Percussivity		$p_1x^2+p_2x^1+p_3$	p_1 =0.8318; p_2 =-9.69 p_3 =16.88	372.5	0.3637	0.3084	4.024
RMS	Percussivity	$p_{10}x+p_{01}y$	p_{10} *=0.8659;* p_{01} *=-6.59*	441.4	0.246	0.2145	4.289
RMS	Percussivity	$p_{00}+p_{10}x+p_{01}y$	p_{00} =-11.03; p_{10} =0.441 p_{01} =-4.987	259.6	0.5565	0.5179	3.36
RMS	Percussivity	$p_{00}+p_{10}x+p_{01}y+p_{20}x^2$ $\ldots+p_{11}xy+p_{02}y^2$	p_{00} =-0.951; p_{10} =1.684 p_{01} =1.267; p_{11} =0.032 p_{20} =0.199; p_{02} =-0.89	246.1	0.5796	0.4745	3.508
Dynamic Spread	Percussivity	$p_{00}+p_{10}x+p_{01}y$	p_{00} =-14.85; p_{10} =-0.3217 p_{01} =-8.614	276.8	0.5271	0.486	3.469
LRA	Percussivity	$p_{00}+p_{10}x+p_{01}y$	p_{00} =-15.22; p_{10} =-0.479 p_{01} =-8.662	334.9	0.4279	0.3782	3.816
RMS	Brightness	$p_{00}+p_{10}x+p_{01}y$	p_{00} =-11.72; p_{10} =-0.537 p_{01} =-4.691	303.3	0.4819	0.4369	3.631
RMS	Spectral Spread	$p_{00}+p_{10}x+p_{01}\log(y)$	p_{00} =-8.551; p_{10} =0.5706 p_{01} =-1.253	324.8	0.4451	0.3968	3.758
RMS	Spectral Centroid	$p_{00}+p_{10}x+p_{01}\log(y)$	p_{00} =-7.162; p_{10} =0.5589 p_{01} =-1.815	320.2	0.4529	0.4054	3.731

We perform the same analysis procedure for the model fitting of threshold. The Goodness-Of-Fit statistical results for the threshold curve fitting are presented in Table 4.4. Again, only modelling functions with relatively good degree of fit are listed here. Analysis based on Table 4.4 shows that models using a feature combination of RMS and percussivity weighting employing first and second order polynomial functions outperform others, with lowest SSE of 259.6, 246.1 and highest R^2 of 0.5565, 0.5796 respectively. Furthermore, second order polynomial models have a slightly better fit than first order in terms of SSE and R^2.

However, when comparing the $R^2_{adjusted}$ and RMSE values, first order appears to be the right choice.

Residual plots of each modelling are shown in Figure 4.6. Neither residual plot provides evidence for choosing the best fit. Therefore, prediction bounds with 95% confidence level are further considered, as shown in Figure 4.7. The second order polynomial model has a wider prediction bounds and tends to over-fit the data.

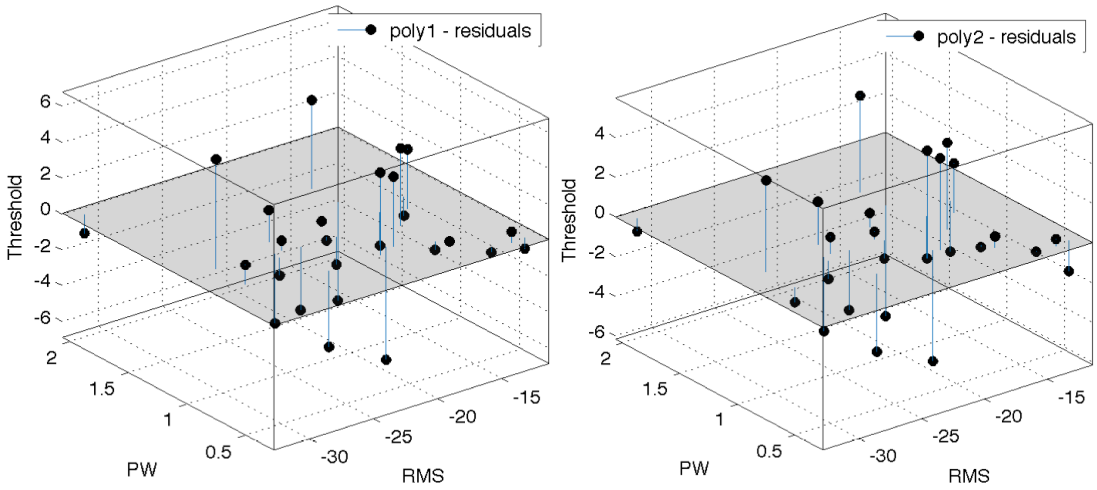


Figure 4.6 Residual plots for first (left) and second (right) polynomial models.

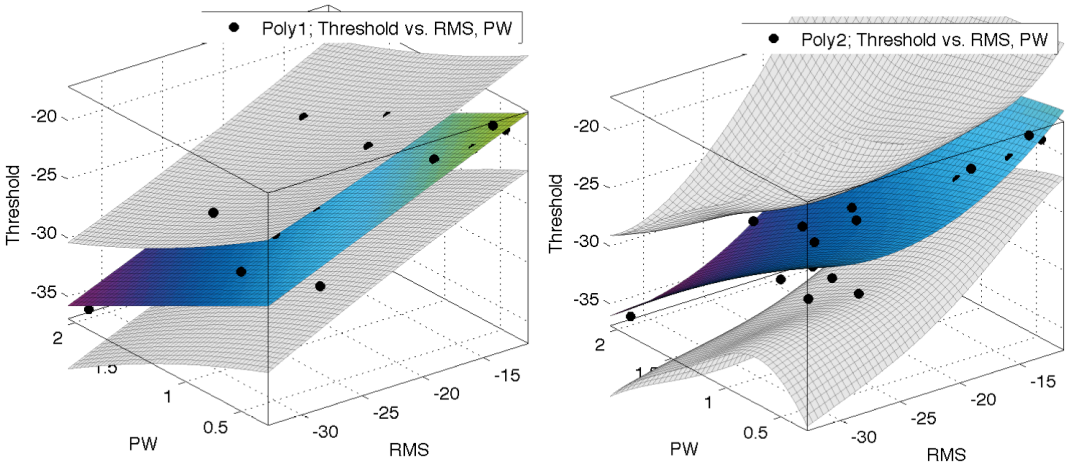


Figure 4.7 Prediction bounds (grey surface) with 95% confidence level of the first (left) and second (right) order polynomial models.

With all criteria considered, the first order polynomial model using RMS and percussivity weighting in Table 4.4 performs the best data fit.

4.4 Intelligent Multitrack Dynamic Range Compression Algorithm

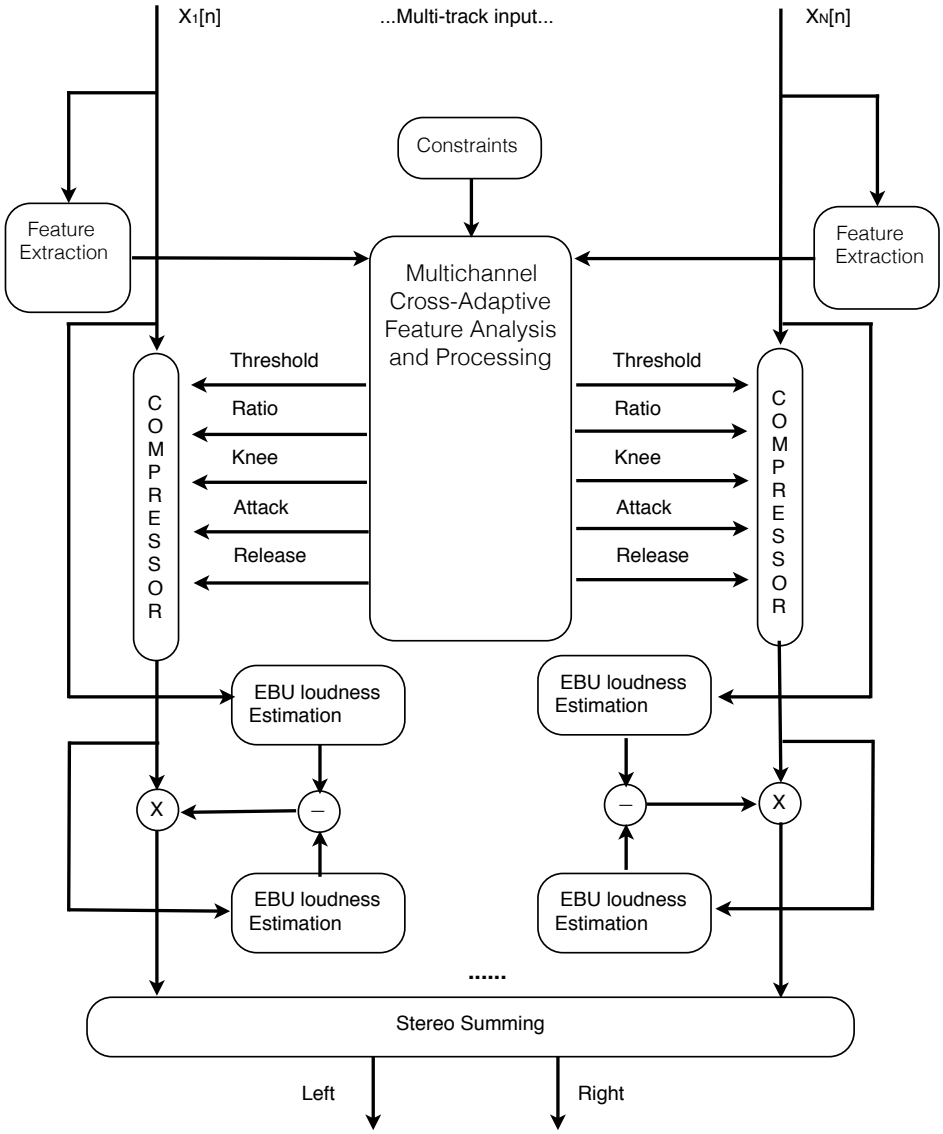


Figure 4.8 System block diagram of the cross-adaptive intelligent multitrack compressor.

The proposed intelligent multitrack compressor is based on the cross-adaptive digital audio effect architecture (Reiss, 2011; Zolzer, 2011). The system workflow is depicted in Figure 4.8.

The ratio and threshold automation is derived from the previous curve fitting process described in Section 4.3.

For ratio automation we choose the model of $f(x,y)= p_{10}x+p_{01}y+1$ with percussivity and frequency weighting features which performs the best curve fit. The final ratio automation for track m is

$$R(m) = 0.54w_p(m) + 0.764w_f(m) + 1. \quad (0.0)$$

Similarly, the threshold automation follows the first order polynomial model, $f(x,y)=p_{00}+p_{10}x+p_{01}y$ with RMS level and percussivity weighting. The threshold automation is

$$T(m) = -11.03 + 0.44x_{RMS}(m) - 4.987w_p(m). \quad (0.0)$$

Learning from Assumption 3, we adapt the algorithms for attack and release automation in (Giannoulis et al., 2012b) using crest factor as a short term signal measure to describe the transient nature of the input signals.

To obtain the average RMS values sample by sample, we apply an Exponential Moving Average filter,

$$x_{RMS}[n] = \sqrt{(1-\alpha)x^2[n] + \alpha x_{RMS}^2[n-1]}. \quad (0.0)$$

The sample-by-sample average peak magnitude of the signal is calculated as

$$x_{peak}[n] = \sqrt{\max(x^2[n], (1-\alpha)x^2[n] + \alpha x_{peak}^2[n-1])}. \quad (0.0)$$

Since the peak detector's and RMS detector's smoothing constants α are equal, the release envelopes of both detectors are guaranteed to be the same, and the peak detector's output is no less than the detected RMS output (Giannoulis et al., 2012b). The crest factor x_{crest} of the signal is defined as

$$x_{crest}[n] = \frac{x_{peak}[n]}{x_{RMS}[n]}. \quad (0.0)$$

Based on (Giannoulis et al., 2012b), the attack and release time constants are calculated by

$$\tau_A[n] = \frac{2\tau_{A-\max}}{x_{crest}^2[n]} \quad , \qquad \tau_R[n] = \frac{2\tau_{R-\max}}{x_{crest}^2[n]} \tag{0.0}$$

where the maximum attack time $\tau_{A-\max}$ is set to 80 ms and the maximum release time $\tau_{R-\max}$ is set to 1000 ms (Giannoulis et al., 2012b).

According to Assumption 4, we set the knee width to half the absolute value of the automated threshold value for a soft knee configuration as

$$W(m) = \frac{|T(m)|}{2}, \tag{0.0}$$

which ensures that a lower threshold results in a wider knee width.

Following Assumption 5, make-up gain is set so that output loudness equals input loudness. The make-up gain is simply the loudness difference between the input and output of the DRC, measured following the ITU/EBU loudness standard (ITU, 2012a),

$$G(m) = L_{in}(m) - L_{out}(m). \tag{0.0}$$

where L_{in} and L_{out} are the input and output loudness values of individual track m, before and after the compression block. In automatic mixing, the loudness setting is usually done post-compression.

4.5 Results and Evaluation

4.5.1 Evaluation Method

Subjective evaluation of the intelligent multitrack compression algorithm was performed in the form of a multiple stimulus (MUSHRA) listening test (ITU, 2003) to assess the

performance of the automatic DRC algorithm against raw mixes, two semi-professional mixes and an alternative automatic DRC implementation (Maddams et al., 2012).

Two mix engineers were master students of the MMus in Sound Recording at the Schulich School of Music at McGill University. They were asked to use Avid's Pro Tools with built-in dynamic range compression effect (with automatic make-up gain applied). Same headphone was used for both engineers. However they were allowed to mix the song with preferred playback level as their own. Editing, rerecording, the use of samples or any other form of adding new audio was not allowed. Analysis and evaluation of audio features of these semi-professional mixes used can be found in (Brecht De Man, King, & Reiss, 2014).

The automatic control strategy of the alternative approach (Maddams et al., 2012) is based on the *a priori* hypothesis that the fundamental role of DRC in multi-track audio mixes is to reduce the difference between the highest and lowest individual track LRA, and that sound sources with higher LRAs require greater amounts of DRC. This hypothesis was substantiated empirically by examining the post-DRC changes in LRA achieved when an experienced mix engineer chose the compressor settings manually.

We aim to evaluate the performance of the automatic algorithm regardless of the choices of genres, instrumentation and different loudness ranges. Therefore when selecting the songs for evaluation, the objective is to chose songs with various genres, different instrumentation and relatively wider range of loudness range (measured using the ITU/EBU loudness standard (ITU, 2012a)). Six different unprocessed multitrack songs (20 seconds segments, not used in the ratio and threshold adjustment experiment) in were selected. The specification of these songs is shown in Table 4.7. All songs used in this work can be accessed from the Open Multitrack Testbed at multitrack.eecs.qmul.ac.uk (De Man et al., 2014).

In all mixes, the only parameter modified was the dynamic range compression to minimise the perceptual bias caused by other audio effects as much as possible. The loudness of the final mixes were normalised manually by a group of professional mixing engineers (ITU, 2003), as done on the same playback system as the subjective evaluation. The order of mixing versions and songs presented to each participant was randomised by a pseudorandom number generator algorithm in Matlab. Participants were encouraged to take as much time as needed.

Table 4.7 The specification of the songs used in the evaluation.

Song	Number of track	Genre	Loudness Range (LU)	Instrumentation
1	3	Rock	5.8	Bass; Electric Guitar; Drum set
2	4	Jazz	11.1	Bass; Piano; Cello; Female vocal
3	6	Folk	14.9	Percussion; Bass; Drum; Acoustic guitar; Electric guitar; Keyboard
4	7	Pop	7.3	Bass; Keyboard; Retro synth; Pad; Female vocal; Piano; Electric drum set
5	7	Rock/Indie	11.1	Bell (synth); Bass; Male backing Vocal; Male lead vocal; Juno (synth); Piano; Drum set
6	5	Indie	16.5	Bass; Electric Guitar; Acoustic Guitar; Male vocal; Synth

All tests were performed in a soundproof listening room with the same headphone set-up, where the environmental noise is minimized. Participants were allowed to adjust the playback level during the experiment in order to evaluate the quality of the mixes efficiently. Sixteen participants with strong audio engineering experience, seven of whom were from the same group of people used in the previous ratio and threshold adjustment experiment, were asked to rate the mix versions according to four specific criteria/questions on a scale of five descriptors: "Bad (0 -20)", "Poor (20 - 40)", "Fair (40 - 60)", "Good (60 - 80)" and "Excellent (80 - 100)":

- Q1: According to the appropriateness of the amount of dynamic range compression applied to each individual sound source in the mix.
- Q2: In terms of the degree of any imperfection such as pumping, breathing artefacts, level imbalance etc.
- Q3: According to the ability to stabilise the erratic level fluctuation within the mix.
- Q4: According to participants' own overall preference.

Since DRC can be relatively subtle, we chose different songs for different questions to maximise the difference. Six songs were tested in Q1 and Q4 while four songs were tested in Q2 and Q3. For Q1, a no-compression mix of each song was also presented as a 'reference'.

However, it does not serve as objectively high quality reference or objectively low quality anchor, which was explained to the participants in advance. The order of the songs as well as the order of the versions of each individual song was randomised when presented to each participant for each question.

4.5.2 Evaluation Results

Lilliefors tests were used for normality check of the evaluation results. The specification for MUSHRA (ITU, 2003) codec quality tests, which are quite similar to ours, offers the simplification that no overlap in the confidence intervals for two conditions means one is significantly better than the other. We will follow this idea whenever it is clear, presenting the standard plots. In (Sporer et al. 2009) it suggests that box-and-whisker plots should be presented to look at possible skewness and outlier behavior. Though we have always followed their recommendation when looking at data, none of these plots are presented here. As for our cases, they proved not to give any new insights.

The Friedman test (Mosteller & Rourke, 1973) is an alternative to ANOVA with repeated measures when normal distribution of the data is not assured. It is appropriate in our cases to use such statistics method to evaluate whether there is significant difference between the different mix types. We perform this test on a song per song basis, and also under the hypothesis that all songs have similar behavior, so that a subject's evaluation of a condition can be averaged over the total number of songs. Furthermore we can the Wilcoxon signed-rank to evaluate the paired-wise difference between mix types of interests.

Q1: Appropriateness of the Amount Of DRC

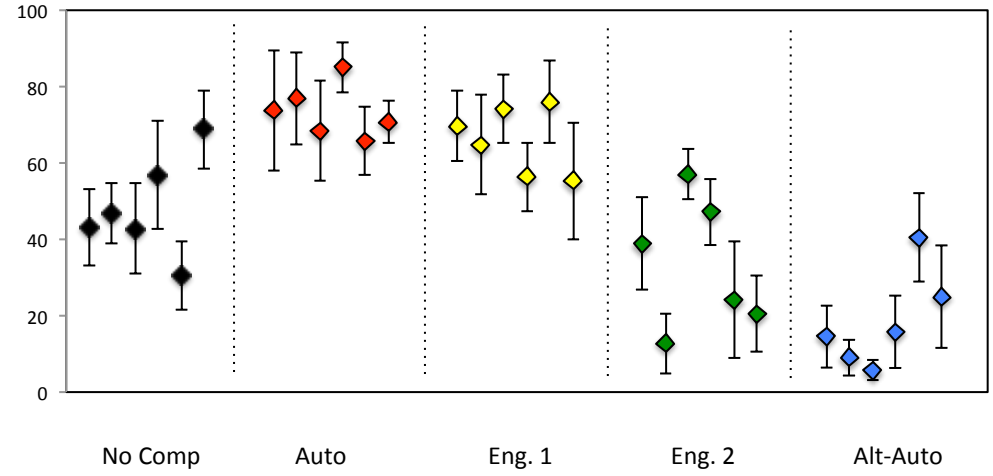
In Q1, participants were asked to rate the mixes in terms of the appropriateness of the amount of dynamic range compression applied in the mix. Table 4.8 shows the results of the Lilliefors normality tests. h=1 indicates non-normal distribution; h=0 indicates normal distribution.

Table 4.8 Normality test result for each song and mix type (Q1).

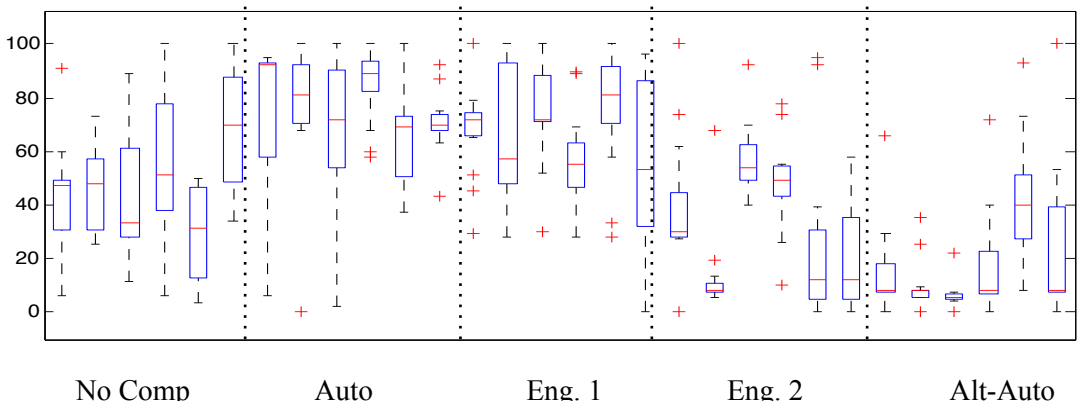
Mix type	Song	h	*p*-value
	1	0	0.1214
	2	0	0.4537
	3	0	0.0635
	4	0	0.5
	5	0	0.2286
No Comp	6	0	0.3903
	1	1	1.00E-03
	2	1	0.0017
	3	0	0.5
	4	1	0.0377
	5	0	0.5
Auto	6	0	0.0754
	1	0	0.103857362
	2	0	0.145547214
	3	1	0.030137838
	4	0	0.414332628
	5	0	0.140889223
Eng. 1	6	0	0.5
	1	1	0.002775508
	2	1	0.001
	3	0	0.096041157
	4	0	0.092483186
	5	0	0.052718565
Eng. 2	6	1	0.022573089
	1	1	0.001
	2	1	0.001
	3	1	0.001
	4	1	0.001
	5	0	0.259078408
Alt-Auto	6	1	0.004329057

Figure 4.9 showing the mean, grouped by mix type, with error bars displaying 95% confidence interval and standard boxplots of the results. No compression mix, automatic mix,

two semi-professional mixes and alternative automatic mix are notated as 'No Comp', 'Auto', 'Eng. 1', 'Eng. 2' and 'Alt-Auto' respectively. The 'Eng. 1' and 'Auto' mixes rate consistently high throughout. The 'Alt-Auto' mixes rate consistently low with the exception of Song 5 and 6.



(a)



(b)

Figure 4.9 (a) Averaged results of Q1: amount of DRC with 95% confidence interval, grouped by mix type. (b) Boxplots of Q1 results.

Table 4.9 shows the results of the Friedman test within each song for the overall data. 'SS' indicates the Sum of Squares (SS) due to each source; 'df' indicates the degrees of freedom (df) associated with each source; 'MS' indicates the Mean Squares (MS), which is the ratio SS/df; 'Chi-sq' indicates Friedman's chi-square statistic; 'Prob>Chi-sq' indicates the p value for the chi-square statistic. All p-values are extremely small, confirming that the mix type affects the evaluation scores significantly.

Table 4.9 The results of Friedman test (Q1).

Song 1: Friedman's ANOVA Table					
Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Columns	76.1333	4	19.0333	30.4533	3.96E-06
Error	73.8667	56	1.319		
Total	150	74			
Song 2: Friedman's ANOVA Table					
Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Columns	103.0667	4	25.7667	41.5034	2.11E-08
Error	45.9333	56	0.82024		
Total	149	74			
Song 3: Friedman's ANOVA Table					
Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Columns	93.7333	4	23.4333	37.4933	1.43E-07
Error	56.2667	56	1.0048		
Total	150	74			
Song 4: Friedman's ANOVA Table					
Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Columns	97.4	4	24.35	39.2215	6.27E-08
Error	51.6	56	0.92143		
Total	149	74			
Song 5: Friedman's ANOVA Table					
Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Columns	68.5667	4	17.1417	27.5184	1.56E-05
Error	80.9333	56	1.4452		
Total	149.5	74			
Song 6: Friedman's ANOVA Table					
Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Columns	74.6	4	18.65	30.0403	4.80E-06
Error	74.4	56	1.3286		
Total	149	74			
All songs: Friedman's ANOVA Table					
Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Columns	90432.9667	4	22608.2417	190.636	3.87E-40
Interaction	24609.8	20	1230.49		

Error	95579.7333	420	227.5708		
Total	210622.5	449			

Furthermore, the results for the paired Wilcoxon signed rank test comparing 'Auto' against 'No Comp' and 'Eng. 1' respectively are shown in Table 4.10. h=1 (when comparing with 'No Comp') indicates the test rejects the hypothesis that evaluation data for 'Auto' and 'No Comp' have no significantly difference. h=0 (when comparing with 'Eng. 1') confirmed again that automatic mixes can compete with professional mixing engineer 1.

Table 4.10 The Results of the Wilcoxon signed rank test when comparing 'Auto' against 'No Comp' and "Eng. 1" (Q1).

'Auto' against	h	*p*-value
No Comp	1	2.74E-10
Eng. 1	0	0.0524

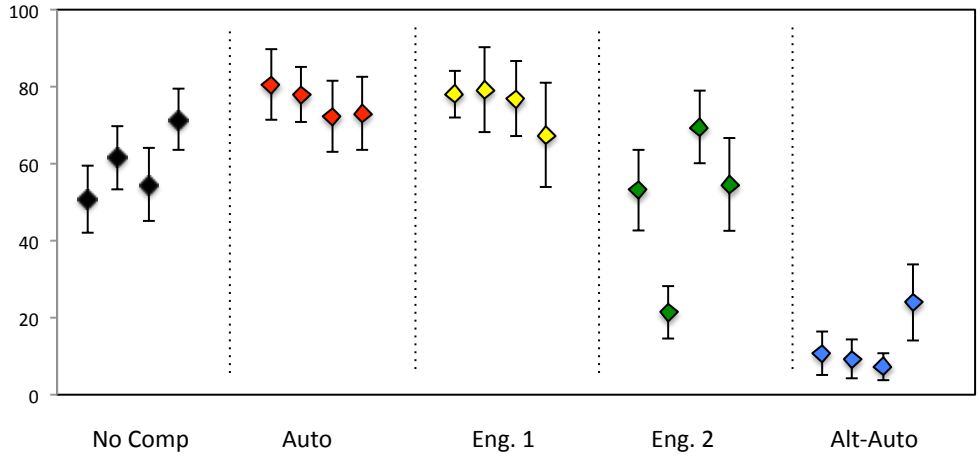
Q2: Degree of Imperfection

Table 4.11 Normality test result for each song and mix type (Q2).

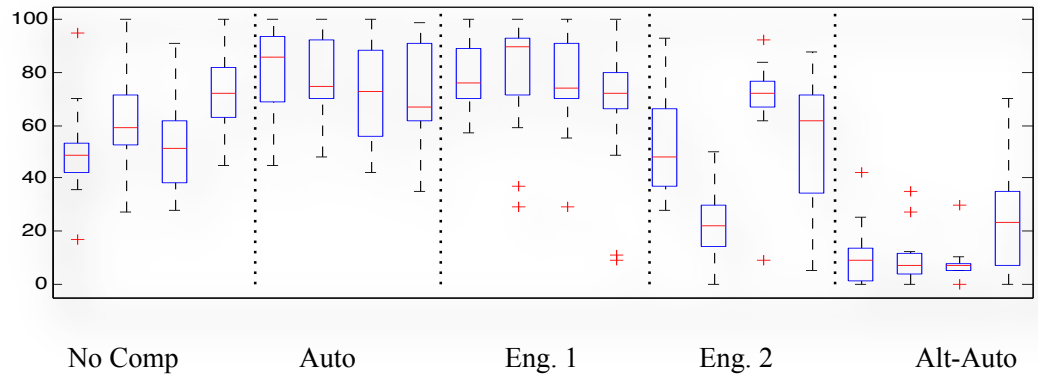
Mix type	Song	h	*p*-value
No Comp	1	1	0.037121269
	2	0	0.5
	3	0	0.5
	4	0	0.208110736
Auto	1	0	0.219525994
	2	0	0.335537323
	3	0	0.5
	4	0	0.374052005
Eng. 1	1	0	0.5
	2	1	0.007376429
	3	0	0.33209403
	4	1	0.00387977
Eng. 2	1	0	0.34033747
	2	0	0.5

	3	1	0.001474582
	4	0	0.374931973
Alt-Auto	1	0	0.177734237
	2	1	0.009146054
	3	1	0.001782132
	4	0	0.137639666

Q2 investigates the degree of sound artefacts or imperfection. Table 4.11 shows the results of the Lilliefors normality tests. The results of the evaluation are presented in Figure 4.10. It can be seen that all 'No Comp', 'Auto', 'Eng. 1', 'Eng. 2' mixes are all rated above the middle score, with only the exception of 'Eng. 2' in Song 2, which suggests these mixes do not have obvious artefacts. 'Alt-Auto' rates the lowest (<20 for most cases) implying significant artefacts are produced in the mixes.



(a)



(b)

Figure 4.10 (a) Averaged results of Q2: degree of imperfection with 95% confidence interval, grouped by mix type. (b) Boxplots of Q2 results.

Table 4.12 shows the results of the Friedman test within each song for the overall data. All *p*-values are extremely small, confirming that the mix type affects the evaluation scores significantly in Q2.

Table 4.12 The results of the Friedman test (Q2) for mix types within each song and across all songs.

	Chi-sq	p-value
Song 1	41.4411	2.18E-08
Song 2	50.4482	2.91E-10
Song 3	36.3525	2.45E-07
Song 4	27.3579	1.68E-05
All songs	153.8026	3.12E-32

Similarly, the results for the paired Wilcoxon signed rank test comparing 'Auto' against 'No Comp' and 'Eng. 1' respectively are shown in Table 4.13. h=1 (when comparing with 'No Comp') indicates that 'Auto' and 'No Comp' have significantly difference. h=0 (when comparing with 'Eng. 1') confirms again that automatic mixes can compete with professional mixing engineer 1 (no significant difference).

Table 4.13 The Results of the Wilcoxon signed rank test when comparing 'Auto' against 'No Comp' and "Eng. 1" (Q2).

'Auto' against	h	*p*-value
No Comp	1	1.3145e-04
Eng. 1	0	0.7829

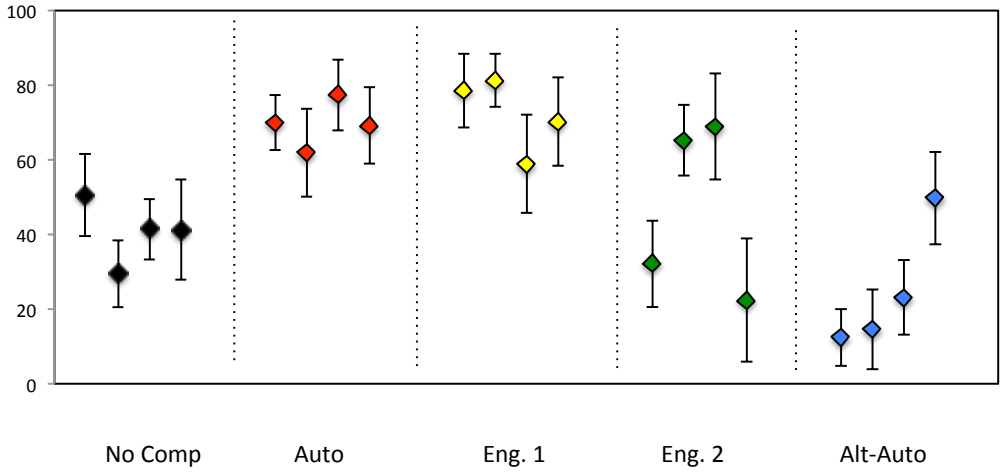
Q3: Ability To Stabilize Erratic Level Fluctuation

Table 4.14 Normality test result for each song and mix type (Q3).

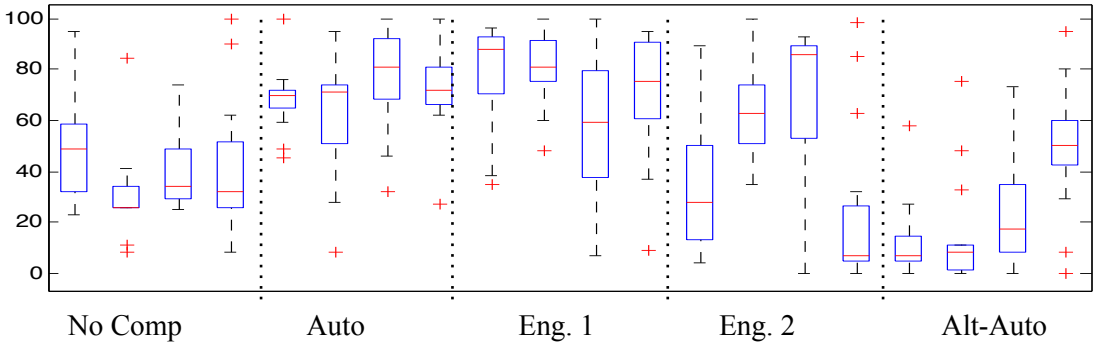
Mix type	Song	h	*p*-value

No Comp	1	1	0.041580352
	2	1	0.043873169
	3	0	0.065945109
	4	0	0.289679973
Auto	1	1	0.014310713
	2	1	0.001
	3	1	0.019291953
	4	1	0.023391155
Eng. 1	1	1	0.002069075
	2	0	0.5
	3	0	0.5
	4	0	0.214834354
Eng. 2	1	0	0.5
	2	0	0.5
	3	1	0.007914252
	4	1	0.001
Alt-Auto	1	1	0.00424392
	2	1	0.001
	3	0	0.365520367
	4	0	0.239473609

This question was designed to make the participants focus on how well the mixes can stabilise the level fluctuations. Table 4.6 shows the results of the Lilliefors normality tests. The results of the evaluation are shown in Figure 4.11. 'Eng. 1' performs best followed by 'Auto' except for Song 3. 'Eng. 2' performs well in Song 2 and 3, while it is the worst in Song 4.



(a)



(b)

Figure 4.11 (a) Averaged results of Q3: level stabilising with 95% confidence interval, grouped by mix type. (b) Boxplot of Q3 results.

Table 4.16 shows the results of the Friedman test within each song for the overall data. All *p*-values are extremely small, confirming that the mix type affects the evaluation scores significantly in Q3.

Table 4.16 The results of the Friedman test (Q3) for mix types within each song and across all songs.

	Chi-sq	*p*-value
Song 1	39.3067	6.02E-08
Song 2	38.8629	7.44E-08
Song 3	23.9467	8.19E-05

Song 4	102.8627	2.42E-21
All songs	152.445	2.12E-32

Similarly, the results for the paired Wilcoxon signed rank test comparing 'Auto' against 'No Comp' and 'Eng. 1' respectively are shown in Table 4.17. h=1 (when comparing with 'No Comp') indicate the test rejects the hypothesis that evaluation data for 'Auto' and 'No Comp' have no significantly difference. h=0 (when comparing with 'Eng. 1') confirmed again that automatic mixes can compete with professional mixing engineer 1.

Table 4.17 The results of the Wilcoxon signed rank test when comparing 'Auto' against 'No Comp' and "Eng. 1" (Q3).

'Auto' against	h	*p*-value
No Comp	1	1.1132e-07
Eng. 1	0	0.2992

Q4: Overall Preference

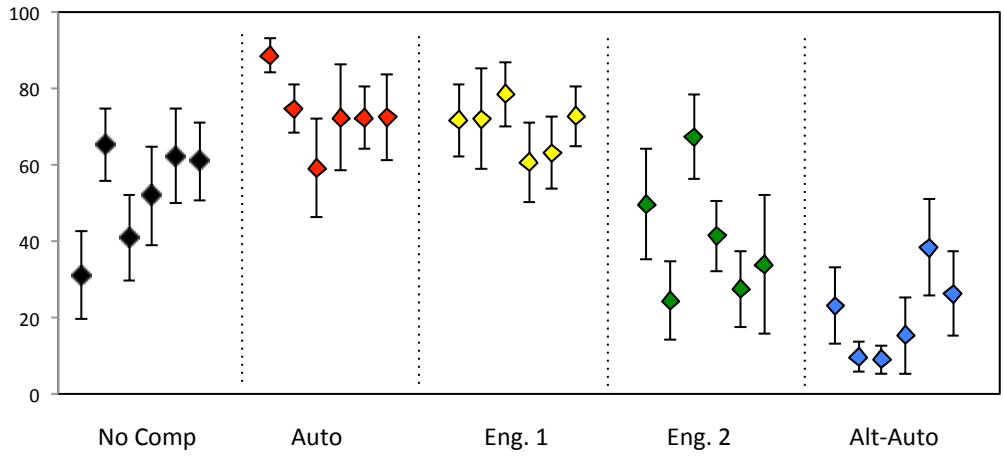
Q4 was designed to study participants' overall preference for the DRC. Normality test results are shown Table 4.17.

Table 4.18 Normality test result for each song and mix type (Q4).

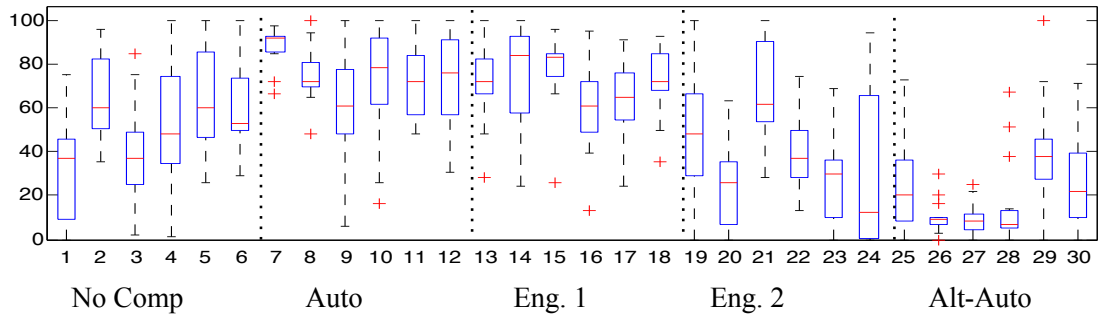
Mix type	Song	h	*p*-value
No Comp	1	0	0.5
	2	0	0.5
	3	0	0.365017838
	4	0	0.5
	5	0	0.5
	6	0	0.116165672
Auto	1	1	1.00E-03
	2	0	0.157743218
	3	0	0.5

	4	0	0.107853639
	5	0	0.5
	6	0	0.263648177
Eng. 1	1	1	0.034947922
	2	1	0.026026837
	3	1	0.0115311
	4	0	0.5
	5	1	0.03110281
	6	1	0.043838815
Eng. 2	1	0	0.5
	2	0	0.11399382
	3	0	0.200539442
	4	0	0.5
	5	0	0.5
	6	1	0.006440728
Alt-Auto	1	0	0.5
	2	1	0.001809799
	3	0	0.182683108
	4	1	0.001
	5	0	0.060639306
	6	0	0.107412693

The results are shown in Figure 4.12. Participants generally prefer 'Auto' and 'Eng. 1' mixes throughout all the songs. 'Eng. 2' has a strongly varying rating depending on the songs.



(a)



(b)

Figure 4.12 (a) Averaged results of Q4: overall preference with 95% confidence interval, grouped by mix type. (b) Boxplots of Q4 results.

Table 4.19 The results of the Friedman test (Q4) for mix types within each song and across all songs.

	Chi-sq	*p*-value
Song 1	38.7023	8.03e-08
Song 2	41.8255	1.81e-08
Song 3	37.1505	1.68e-07
Song 4	27.84	1.34e-05
Song 5	30.9699	3.11e-06
Song 6	26.6133	2.38e-05
All songs	178.0795	1.93e-37

Table 4.19 shows the results of the Friedman test within each song for the overall data. All *p*-values are extremely small, confirming that the mix type affects the evaluation scores significantly in Q3.

Table 4.20 The results of the Wilcoxon signed rank test when comparing 'Auto' against 'No Comp' and "Eng. 1" (Q4).

'Auto' against	h	*p*-value
No Comp	1	1.4204e-06
Eng. 1	0	0.2944

Similarly, the results for the paired Wilcoxon signed rank test comparing 'Auto' against 'No Comp' and 'Eng. 1' respectively are shown in Table 4.20. h=1 (when comparing with 'No

Comp') indicate the test rejects the hypothesis that evaluation data for 'Auto' and 'No Comp' have no significantly difference. h=0 (when comparing with 'Eng. 1') confirmed again that automatic mixes can compete with professional mixing engineer 1.

Overall performance

To give a clearer depiction of the overall performance of each mix type, the averaged mean results with 95% confidence interval across all participants and songs are displayed in Figure 4.13. Normality tests for different mix types suggest non-normal data distribution considering inter-song differences to be irrelevant.

.

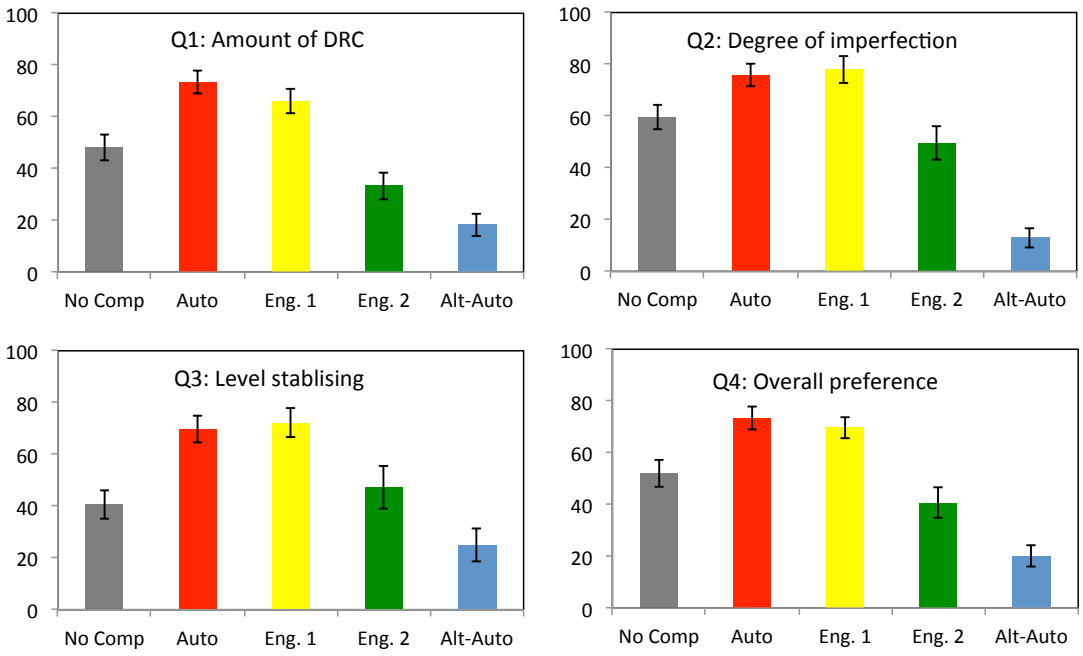


Figure 4.13 Overall mean results with 95% confidence interval for Q1-Q4 grouped by mix type.

Figure 4.13 shows the proposed 'Auto' performs best in Q1: the appropriate amount of DRC. 'Auto' also performs quite well in terms of stabilising the erratic level fluctuation in Q3. More importantly, the proposed automatic mixes are the participants' favourite in Q4, the overall preference. The no compression version is preferred to 'Eng. 2' in Q1, Q2 and Q4, suggesting that people sometimes dislike the use of compression. The 'Alt-Auto' is clearly the worst overall performer. The mean results in Q2 show that the Alt-Auto causes

obvious sound artefacts or unpleasant effects. 'Alt-Auto' uses pre-processing to equalise the loudness (measured by EBU loudness standard) of multitracks before compression, which is likely to make the percussive instrument much louder than other instruments, resulting in an unpleasant listening experience. The same issue was addressed in (Mansbridge et al., 2012). This could be the reason that the 'Alt-Auto' performs poorly. Notice that although Auto and Eng. 1 perform similarly in the subjective evaluation, results also show participants' preference is song dependent. For example in Figure 4.11, Song 3 is rated the highest for Auto, while the same song is rated the lowest for Eng. 1.

Overall, the results show that the proposed automatic compression has very good performance based on various criteria. However, since the algorithm uses relative measure of the audio feature such percussivity factors, it might tend to overlook the amount of compression needed when every track in the mix has a similar percussivity factors. It should be regarded as the limitation of the automatic multitrack dynamic range compression and to be investigated as future work.

Furthermore, when performing feature correlation analysis to perform curve fitting for threshold and ratio automations, results (see Table 4.4) show most features seem to exhibit low to medium correlation against the subjective results. Although the automatic algorithm combines two or three features that have the highest correlation based on control assumptions to tackle the relatively low correlation (see Section 4.2). It would be beneficial to extend the selection of features to include high-level or more perceptual features as future work to improve the performance of the algorithm.

4.6 Conclusions

In this chapter we have proposed a novel intelligent multitrack dynamic range compression algorithm. The algorithm utilises the CA-DAFX processing architecture (Reiss, 2011; Zolzer, 2011), exploits the interdependence of the input audio features and incorporates best practices as well as subjective evaluation results to produce the optimal amount of dynamic range compression for multitracks.

To the best of the authors' knowledge, this presented the first fully automated multitrack dynamic range compressor where all classic parameters of a typical compressor (ratio,

threshold, knee, attack and release) are dynamically adjusted depending on extracted features and control rules.

In the pursuit of intelligent algorithms, two new audio features, namely percussivity weighting and low-frequency weighting, were proposed to describe the transient nature and spectral content of the signal. A method of adjustment experiment was conducted to investigate the relationship between human preference for ratio and threshold. We applied multiple linear regression models to the subjective results to formulate the ratio and threshold automations that follow the choices of the human operators.

The output mix produced by the proposed algorithm has an outstanding performance in the final subjective evaluation when compared against a raw mix, two semi-professional mixes and a previous automatic compression approach. The results showed that the algorithm is able to compete with or outperform the semi-professional mixes in terms of four different perceptual criteria: the appropriateness of the amount of DRC applied, the degree of imperfection, ability to stabilise the erratic level fluctuations and overall preference. Subjective evaluation results also have shown that spectral content plays an important role in the pursuit of an intelligent solution to dynamic processing.

Chapter 5

Multitrack Masking Metrics

5.1 Introduction

Chapter 3 and 4 investigated the frequency and dynamics aspects of intelligent mixing, and proposed various intelligent mixing algorithms to achieve optimal balance in spectral and dynamics characteristics. However, no perception models were applied to the system to inform the mixing decisions.

For a true intelligent mixing system to triumph, it will be beneficial to equip the signal analysis chain with perceptual models that considers properties of the hearing system. Therefore this chapter explores the auditory aspects of intelligent music production.

Masking remains one of the most challenging issues entailed in the mixing process. The mix can sound confusing or underwhelming, and have a lack of clarity as a result of untreated masking. Previous perceptual models capable of predicting auditory masking have been discussed in Section 2.3.2. The loudness model of Glasberg and Moore (Glasberg & Moore, 2002; Moore et al., 1997) and the psychoacoustic model used in MPEG audio coding (ISO, 1993; Johnston, 1988a, 1988b) are the main concerns of this chapter. As the loudness model of Glasberg and Moore has the ability to predict the partial loudness, it can be viewed as a model of masking. However, the model has never been evaluated with musical signal.

In this chapter, we first present an equal loudness matching experiment to evaluate the performance of existing loudness models Moore (Glasberg & Moore, 2002; Moore et al., 1997) on musical signals in Section 5.2. We analyze the underlying features and propose a parameter modification of the model that can yield better compliance with the human perception of masking (Moore, 2012). The outcome of this experiment is then integrated into the development of several psychoacoustics-inspired, cross-adaptive multitrack masking

models to quantify the masking behaviour within the musical mixture in Section 5.3 and 5.4. Overall discussion and conclusion are outlined in Section 5.5.

5.2 Loudness Matching Experiment

5.2.1 Evaluated Multitrack Loudness Model

The evaluated multitrack loudness model adapts the loudness models of Glasberg and Moore (Glasberg & Moore, 2002; Moore et al., 1997) to estimate the loudness and partial loudness of multitrack where each track may be masked by the combination of every other track. The structural overview of the model is depicted in Figure 5.1. System calibration is crucial and performed by measuring the sound pressure level of a 1kHz full-scale tone at eardrum. The same headphone was used during all experiments.

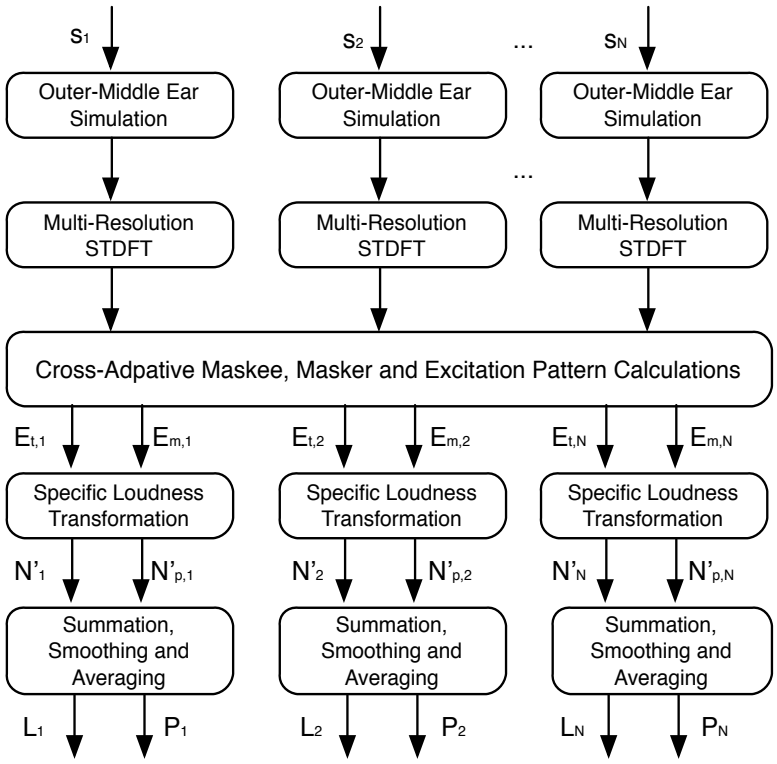


Figure 5.1 Block diagram of the cross-adaptive multitrack loudness model with *N* input signals, adapting the loudness models of Glasberg and Moore.

The procedure to derive the loudness L_n and partial loudness P_n of track n from a multitrack with N tracks is similar to the model of Glasberg and Moore (Glasberg & Moore, 2002; Moore et al., 1997) as described in Section 2.3.1.1, but adapting a cross-adaptive architecture (Zolzer, 2011) to address the multitrack scenario. To account for partial masking occurring in every track, two excitation patterns, the target track $E_{t,n}$ with respect to each track s_n, are computed. The masker s'_n here, related to track s_n is the supplementary sum of the other tracks in the multitrack mixture:

$$s'_n = \sum_{i=1, i\neq n}^{N} s_i. \tag{0.0}$$

The transformations from the excitation pattern $E_{t,n}$ to the specific loudness N'_n and partial specific loudness $N'_{p,n}$ are based on Section 2.3.1.1.

And then the operations of summation, smoothing and averaging described Section 2.3.1.1 are performed on N'_n and $N'_{p,n}$ to obtain the final loudness measures of input signal s_n: loudness L_n and partial loudness P_n (due to the presence of other tracks in the multitrack mixture).

5.2.2 Stimuli

Four multitrack songs of different genres were selected. 10s segments of each song were extracted from the uncompressed waveform signals. Each consisted 4 or 5 different instrument stems (a sub-mix of the tracks that represent the same instrument in the process of mixing), all in mono and running at a typical sampling rate of 44.1 kHz. The specifications of the testing samples are presented in Table 5.1.

Table 5.1 The specification of the testing samples in terms of genre, instrumentation and RMS level. The reference level for the RMS measurement is the lowest possible sample is for 16 bit audio in digital full scale: 96 dBFS.

	Genre	Instrumentation	Level (dB)
Song 1	Classical	Bassoon	64
		Clarinet	64
		Saxophone	67
		Violin	68
Song 2	Metal	Bass	67
		Electric Guitar	70
		Drum set	65
		Vocal	70
Song 3	Punk	Bass	60
		Electric Guitar	73
		Drum set	54
		Vocal	67
Song 4	Alternative rock /Electronic	Bass	52
		Drum set	65
		Acoustic Guitar	64
		Vocal	71
		Piano	62

The author chose these audio samples that vary from different genres and different instruments to test whether the proposed masking model can correctly describe the amount of masking perceived by the subjects. The temporal and spectral characteristics of the songs are taken into account regarding to the degree of masking. We select songs with various amount of masking that can be perceived.

5.2.3 Subjects

In total 12 participants whose age ranged from 21 to 32 had taken part in the experiment. Before commencing, subjects were asked to complete a personal information questionnaire.

The summary is displayed in Table 5.2. The results show that the majority of subjects had at least some experience in critical listening, and no one has hearing impairment.

Table 5.2 Results of the informational questionnaire.

Gender	Male	9
	Female	3
Critical listening skill? / Listening tests experience?	No	2
	Some	2
	Yes	8
Hearing impairment?	No	12
	Yes	0

5.2.4 Procedure

A preliminary listening test was performed before the actual loudness matching experiment. Subjects were required to listen to all the mixes and identify every instrument contained in each mix. Subjects need to pass this preliminary test in order to continue to the next formal experiment.

All tests were performed in a soundproof listening room with the same headphone set-up, where the environmental noise is minimized. Participants were allowed to adjust the playback level during the experiment in order to evaluate the masking efficiently. For each loudness matching trial, both solo stem (stem that is played separately) and mixed stem (stem that is played in a mixture together with other stems) were presented in a regular alternation with two seconds silent intervals between successive sounds played through the same calibrated headphone. The order of the trials was randomized for every subject to minimize the bias that subjects become familiar with the song and judge the loudness based on memory. Within a given trial, either the solo stem or the mixed stem level was fixed as the reference stem, and the level of the other as the target stem, was varied to reach the level corresponding to equal loudness in perception. By varying the level of the mixed stem, we mean subjects were only allowed to adjust the same instrumental stem in the mix while the levels of other

stems in the mix were kept unchanged. The starting level of the variable stem was chosen randomly from within a range of ±10 dB, around the level of the reference stem.

The loudness matching experiment was designed using the method of adjustment methodology used in (Moore, Vickers, Baer, & Launer, 1999). The difference between the target stem and the reference stem was recorded after each trial, which was expressed as the Root-Mean-Square level (RMS). The average difference for each stem across subjects was then calculated as a measure of partial masking. Model predictions were computed in both conditions in a similar way.

5.2.5 Subjective Results

All 12 subjects successfully passed the preliminary tests suggesting that subjects were able to identify and judge the partial loudness of an instrument stem when mixed with other stems. To present the results, the level difference between the solo stem and the mixed stem at the point of equal loudness are calculated as follows:

$$\Delta R = R_m - R_s, \tag{0.0}$$

where R_m, R_s are the RMS levels of the mixed stem and the solo stem respectively. Positive level difference ΔR indicates that mixed stem require a larger level increment to reach the point of equal loudness as the solo stems. This agrees with the concept of partial masking: the loudness of an audio signal is generally reduced in the presence of other sounds. However, unusual negative values of ΔR are also found and considered an error due to subjects' mistakes in the experiment or the sensitivity limit of human ears, which is generally within ± 2 dB.

The Lilliefors tests are performed for normality check of the subjective results for each instrument cases of each song in both mix varied and solo varied cases. Results (see Table 5.3) suggest the evaluation results are mostly normal (28 out of 34).

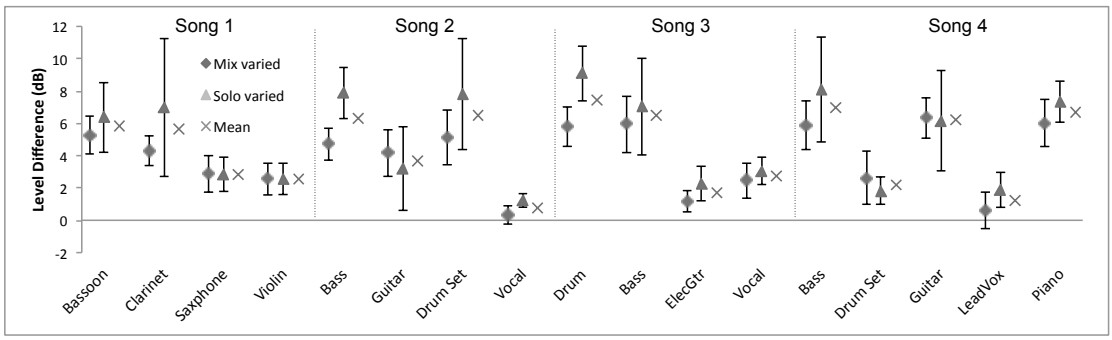
Table 5.3 Normality test result, h=0 indicates normal; h=1 indicate non-normal data.

	Instrument	h (Mix)	h (Solo)
Song 1	Bassoon	0	0
	Clarinet	0	1
	Saxophone	0	0
	Violin	0	0
Song 2	Bass	0	0
	Guitar	0	0
	Drum Set	0	0
	Vocal	0	0
Song 3	Drum	0	0
	Bass	0	1
	Guitar	0	0
	Vocal	0	0
Song 4	Bass	0	1
	Drum	0	0
	Guitar	0	1
	Vocal	1	1
	Piano	0	0

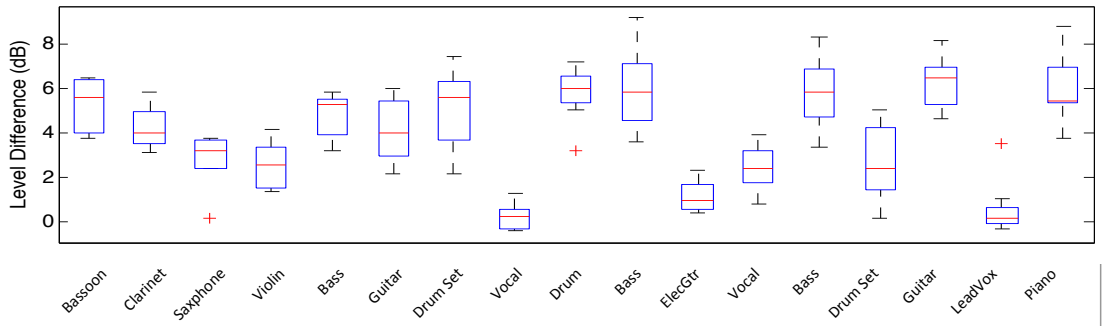
The mean subjective results of the loudness matching experiments across all the subjects, as well as the standard boxplots of the same results are shown in Figure 5.2. Results are plotted separately for the case where the mixed stem is varied and the case where the solo track is varied.

As Figure 5.2 shows, the evaluation results for both cases share a good degree of consistency (p=2.51e-4). There is a very small bias related to whether the mixed stem or the solo stem was varied, indicating that subjects tend to assign a lower level to the solo stem when matching loudness against the mixed stem. Also larger standard errors are observed for the cases of varying the solo stems comparing to the case of varying the mixes, suggesting that it's

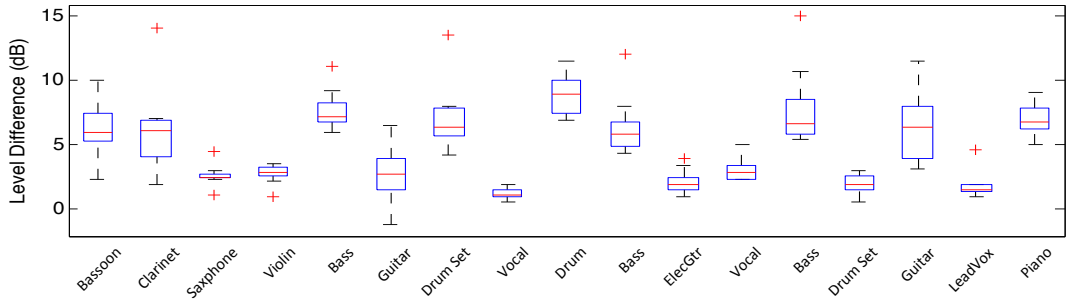
more difficult for participants in such condition. This can possibly due to the difficulty to evaluate the loudness of an individual instrument out of the mixture correctly. The mean of the consistent bias across all conditions and subjects is about +1.2 dB.



(a)



(b)



(c)

Figure 5.2 (a) The measured results plotted separately for the case where the mixed stem is varied (with 95% confidence intervals), the case where the solo track is varied, and the mean values of both cases. (b) Boxplot for the case where the mix stem is varied. (c) Boxplot for the case where the solo stem is varied.

Discounting the bias by looking at the mean for both cases. All values are positive, which means that at the point of equal loudness the RMS level of mixed stems are higher than the

solo stems. It implies that partial masking occurs. The level difference R_{Δ} at the point of equal loudness could be seen as a measurement of partial loudness.

We can also observe some variations across different instrument stems within each song. The drum set stem in song 3 scored the highest level-difference of 7.4 dB while the vocal tracks in song 2 and song 4 have the lowest average of variation of 0.8 dB and 1.2 dB respectively. It means some instruments suffer less partial masking while other instruments suffer significant partial masking resulting larger loudness reduction. It also confirmed that masking is source dependent. The level and frequency interactions between the masker and masked sounds decide the degree of simultaneous masking.

5.2.6 Model Prediction

We apply the proposed multitrack loudness model on the testing signals to obtain the level difference at the point of equal loudness predicted by the model, in a similar way as in the previous experiment. Theoretically, the point of equal loudness for the model prediction is the point when the loudness L_n equals to its partial loudness P_n in the mix:

$$L_n = P_n. \tag{0.0}$$

Model predictions are calculated for both cases as in the loudness matching experiments. For instance, the optimization-like process to derive the model prediction for the case of varying the level of the solo stem: the partial loudness of the mixed stem, P_n is first calculated as a loudness reference. A loudness of the solo stem is then calculated and compared against P_n. Iterations of applying boost or attenuation (in dB scale) to the solo stem are conducted. The iteration process continues until the equal loudness condition is fulfilled as given by:

$$\left| P_n - L_n(\Delta R') \right| \leq e, \tag{0.0}$$

where the tolerance of error e, equals to 1.5 phons. $L_n(\Delta R')$ is the new loudness value of the solo stem with the boost or attenuation, $\Delta R'$ (dB). The value of $\Delta R'$ is then recorded as the model prediction for the level different at the point of equal loudness. A similar scenario is performed for the case of varying the mixed stem.

Table 5.4 presents the level difference predicted by the proposed model, compared against the measured mean results from the loudness matching experiments. The final column lists the prediction errors ($\Delta R' - \Delta R$) of the model.

As Table 5.4 shows, although the model prediction values correlate well with the overall trend of the subjectively perceived level differences, the model predictions are much higher than the observed subjective results. Prediction errors are significantly larger than the minimum perception sensitivity of human hearing system of loudness variations.

Table 5.4 Level differences predicted by the proposed model compared against the measured results from the loudness matching experiments with prediction errors.

	Instrument	Measured Level Difference (dB)	Model Prediction (dB)	Prediction Error (dB)
Song 1	Bassoon	5.7	11.5	5.8
	Clarinet	5.2	12	6.8
	Saxophone	2.7	9.5	6.8
	Violin	2.7	5.5	2.8
Song 2	Bass	6.1	13	6.9
	Guitar	3.5	7	3.5
	Drum Set	6.7	13	6.3
	Vocal	0.7	5	4.3
Song 3	Drum	7.3	15	7.7
	Bass	6.2	16	9.8
	Guitar	1.6	5	3.4
	Vocal	2.8	9	6.2
Song 4	Bass	6.8	12	5.2
	Drum	2.3	8	5.7
	Guitar	6.3	12	5.7

Vocal	1.2	6	4.8
Piano	7.1	13	5.9

Overall, evaluation results suggest that the proposed multitrack loudness model over-estimated the loudness reduction due to partial masking. One possible reason could be that the loudness models of Glasberg and is not well applicable to music signals as discussed in Section 2.3.1. Unlike laboratory stimuli such as tones and noises, music signals contain distinct spectral components, rhythm pattern and melody structures, which could make it easier to distinguish from other sound sources. As a result, it reduces the effect of partial masking in the mix. The errors could also arise from the partial loudness calculation. The partial loudness estimation in the model of Glasberg and Moore does not take into account the fact that the audibility of a signal may be improved when the masker contains amplitude fluctuations that are correlated in different frequency regions. Assuming L_n corresponds well to perception, all errors are positive indicating that the partial loudness P_n, predicted by the model is lower than the actual loudness subjects perceived. That is, the partial loudness model underrates the loudness of musical signal in the presence of other sounds.

5.2.7 Modification of the Loudness Model

Following the evaluation results and discussion about the possible causes of the significant prediction errors, we investigate the partial loudness estimation of the model and search for possible modifications to the model, in order to have a better compliance with the human perception.

K Parameter in the Partial Loudness Estimation

A parameter *K,* defined as the signal-to-noise ratio at the output of the auditory filter required for threshold at high masker levels, was introduced in the process of transformation of the excitation pattern to a specific partial loudness pattern in (Moore et al., 1997). The parameter *K* has a crucial influence on the calculation on partial loudness. The lower the values of *K*, the higher the predicted partial loudness value. However, the values of *K* as a function of frequency were estimated by pooling data from relatively old research work

(Moore et al., 1997). Nevertheless, there were no estimates of *K* for centre frequencies below 100 Hz, *K* values from 50 to 100 Hz were based on extrapolation.

Adjustment of the *K* Parameter

In (Aichinger et al., 2011), threshold detection experiments using an adaptive two-alternative forced-choice task to adjust the partial loudness model were performed. The results showed that if *K* was reduced by 5 dB the compliance of the prediction and the measurement is improved. However, the stimuli used in the experiment were laboratory tones and noise rather than musical signal. For this reason, model adjustment based on *K* is further explored on musical signal here. We perform the same model prediction process using different partial loudness calculations with different *K* values of 0 dB, -5 dB, -10 dB and -15 dB attenuation. The results of the different model predictions are compared against the evaluation results obtained from the loudness matching experiments, as shown in Figure 5.3.

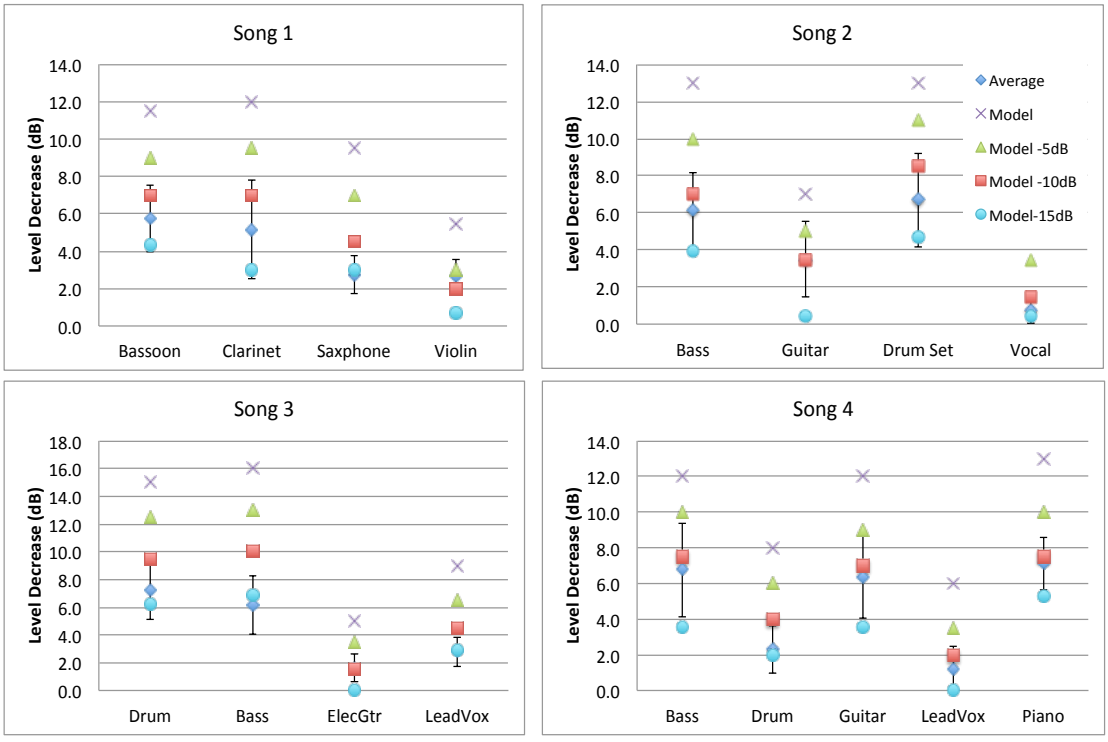


Figure 5.3 Comparison of different model predictions of different *K* parameter values against subjective results plotted with standard deviation.

In Figure 5.3, the blue diamond indicates the mean result obtained from the loudness matching experiment with error bars corresponding to the standard deviation across all subjects. Blue circle, red square, green triangle and purple cross indicate the model predictions with -15 dB, -10 dB, -5 dB, 0 dB attenuation respectively.

The model predictions of the original *K* values, -5 dB *K* values are all above the upper standard deviation of the obtained subject's data, implies that these two model modifications still overestimate the effect of partial masking. The -15 dB *K* modification, however, underrates the effect of partial masking as its results fall below the subjective average. Overall, the -10 dB *K* modification has the best compliance, as most model predictions values (19 out of 21) are within the standard deviation range of the empirical results.

Detailed comparison of the -10 dB *K* modification with the subjective results is shown in Table 4. It shows that the prediction errors are within 0 - 1.5 dB variation for most cases (17 out of 21), which are barely perceivable by the human hearing system, suggesting that the -10 dB *K* modification is appropriate.

Table 5.5 Level differences predicted by the -10 dB *K* modification, compared against the results from the loudness matching experiments with prediction errors.

	Instrument	Measured Level Difference (dB)	-10 dB Model Prediction (dB)	Prediction Error (dB)
Song 1	Bassoon	5.7	7	1.3
	Clarinet	5.2	7	1.8
	Saxophone	2.7	4.5	1.8
	Violin	2.7	2	-0.7
Song 2	Bass	6.1	7	0.9
	Guitar	3.5	3.5	0.0
	Drum Set	6.7	8.5	1.8
	Vocal	0.7	1.5	0.8
Song 3	Drum	7.3	9.5	2.2
	Bass	6.2	10	3.8
	Guitar	1.6	1.5	-0.1

	Vocal	2.8	4.5	1.7
Song 4	Bass	6.8	7.5	0.6
	Drum	2.3	4	1.7
	Guitar	6.3	7	0.7
	Vocal	1.2	2	0.8
	Piano	7.1	7.5	0.4

In summary, with a -10 dB modification to the *K* parameter in the calculation of partial loudness, the proposed multitrack loudness model based on (Glasberg & Moore, 2002; Moore et al., 1997) yields a better model compliance with the human perception of masking.

5.3 Masking Metrics Based on Glasberg and Moore's Loudness Models

We propose two cross-adaptive masking metrics adapting the multitrack loudness model described in Section 5.2.1, incorporating the parameter modification we discovered from the loudness matching experiment.

Metric I: Cross-Adaptive Multitrack Masking Metric

Metric I is a cross-adaptive multitrack masking metric that makes use of the loudness and partial loudness estimations of the multitrack loudness model directly. It quantifies the amount of masking as the loudness deduction due to the presence of the accompanying tracks in the mix. Let M_n denote the approximated amount of masking of track n. M_n therefore can be calculated using Equation (0.0):

$$M_n = \frac{L_n - P_n}{L_n}. \tag{0.0}$$

Unlike previous masking models discussed in Background Section 2.3.2, which only consider the situation when audio signal is completed masked by using masking threshold as a measurement of masking, Metric I is able to take partial masking into account.

Table 5.6 The amount of masking occurred in each instrument track of a 7-track song, measured by the masking Metric I. The masker signal is listed in the first row, the maskee

signal is listed in the first column. So each value (apart from the last "Mix" columns) can be read as the amount of masking occurring in each instrument track masked by a related masker signal (0 - no masking; 1 – fully masked). The last column is the standard M_n regarding the accompanying sum as the masker signal.

	Bass	Beat	Cocotte	Guitar 1	Guitar 2	Keyboard	Voice	Mix
Bass	*	0.49	0.31	0.44	0.55	0.29	0.48	0.79
Beat	0.16	*	0.11	0.24	0.26	0.13	0.26	0.48
Cocotte	0.32	0.41	*	0.47	0.55	0.31	0.48	0.70
Guitar 1	0.25	0.44	0.23	*	0.50	0.33	0.58	0.87
Guitar 2	0.17	0.26	0.15	0.27	*	0.19	0.33	0.55
Keyboard	0.22	0.43	0.24	0.60	0.52	*	0.59	0.82
Voice	0.20	0.36	0.18	0.40	0.40	0.24	*	0.66

We apply Metric I on a selected multitrack song to informally evaluate its performance. The amount of masking occurring among a 7-track multitrack song estimated by Metric I, is listed in Table 5.6. We can see that the Guitar 1 track has the largest masking problem, M_n=0.87. Beat, Guitar 2, and Voice generate more masking effect on others since they are set to have higher sound levels and themselves have smaller masking values of 0.48, 0.55, 0.66. It is reasonable because they are the most important tracks in the mix. When we investigate the table horizontally, we can see that the Bass track is masked by Guitar 2 and Beat tracks by 0.55 and 0.49, respectively. However, the Beat track is masked only a small amount by the Bass track, 0.16.

The multitrack was mixed manually by the author with the objective to minimize the amount of masking. The masking model was then applied to the processed multitrack, to evaluate whether the masking behaviour captured by the model can reflect manual processing of masking reduction informally. The author processed the mix in Logic Pro with built-in fader and equalizer. More specifically: a -2.5 dB gain was applied to the bass track; a -3dB cut at centre frequency of 1200 Hz and Q-factor of 4.5. A 3.2 dB; 1.7 dB boost on 670 Hz and 2000 Hz respectively on Guitar 1; A overall gain of 1.2 dB applied on Keyboard track. The new masking result of the processed multitrack is shown in Table 5.7

Table 5.7 The amount of masking occurring in every instrument track of the "re-mixed" 7-track song measured by Metric I.

	Bass	Beat	Cocotte	Guitar 1	Guitar 2	Keyboard	Voice	Mix
Bass	*	0.45	0.29	0.38	0.52	0.29	0.45	0.77
Beat	0.17	*	0.11	0.26	0.24	0.14	0.26	0.49
Cocotte	0.34	0.41	*	0.46	0.53	0.33	0.48	0.70
Guitar 1	0.24	0.40	0.20	*	0.39	0.31	0.54	0.79
Guitar 2	0.21	0.28	0.17	0.30	*	0.22	0.34	0.59
Keyboard	0.22	0.39	0.21	0.56	0.43	*	0.54	0.76
Voice	0.22	0.36	0.18	0.44	0.37	0.27	*	0.69

As Table 5.7 shows, there is no masking value larger than 0.8 in the "Mix" column, which means that the masking effect has been decreased in the overall mix. Furthermore, the amount of masking is generally smaller than pervious results, especially for the Guitar 1 track where the value drops from 0.87 to 0.79.

Metric II: Masking Metric Adapting the Method Of Vega Et Al.

Alternatively, we can quantify the amount of masking by investigating the interaction between the excitation patterns of the track, $E_{t,n}$ and the supplementary sum of the other tracks $E_{m,n}$, adapting the method of Vega et al (Vega & Janer, 2010). The masking measurement, M_n thus can be defined as the masker-to-signal ratio (MSR) based on excitation pattern integrated across ERB scale and time, is given by

$$M_n = \frac{\sum_{b_{ERB}} E_{m,n}}{\sum_{b_{ERB}} E_{t,n}}. \tag{0.0}$$

As Equation (0.0) suggests, Metric II is based on excitation patterns of the masker and maskee rather than more perceptual measurement such as loudness.

5.4 Masking Metrics Based on MPEG Psychoacoustic Model

As discussed in Background Section 2.3.2, The MPEG psychoacoustic model plays a central role in the compression algorithm. This model produces a time-adaptive spectral pattern that

emulates the sensitivity of the human sound perception system. The model analyzes the signal, and computes the masking thresholds as a function of frequency (ISO, 1993; Johnston, 1988a, 1988b). The procedure to derive masking thresholds has been summarized in Section 2.3.2.1 together with a block diagram (see in Figure 2.4) illustrated the estimation stages involved in the psychoacoustic model. The mechanism behind the psychoacoustic model gives insight into a manner in which it can be adapted into masking metrics to describe the masking behavior for multitrack audio.

In addition to the Glasberg and Moore's Loudness Model based Metric I and II, we also propose two other masking metrics adapting and expanding the psychoacoustic model in MPEG audio coding.

Metric III: MPEG Masking Metric Derived From the Final Mix

We can measure the amount of masking by looking at the masking threshold of the final stereo mix directly. This approach assumes that when there is more masking in the multitrack, there will be more masking within the final mix, and more efficient MPEG audio coding can be applied to the final mix. The masking measurement of the mixture, M_{mix} then becomes

$$M_{mix} = \sum_{sb \subset E_{sf} < T} \frac{MSR(sb)}{T_{\max}}, \tag{0.0}$$

where T_{max} is the predefined maximum amount of distance between the energy of the mix in each scale-factor band $E_{sf}(sb)$, and measured masking threshold in each scale-factor band, $T(sb)$. $MSR(sb)$ is the Masker-to-Signal Ratio (MSR) in each scale-factor band. $E_{sf}(sb)$, $T(sb)$ and $MSR(sb)$ are derived from the psychoacoustic model used in MPEG audio coding as described in Section 2.3.2.1.

As Equation (0.0) suggests, Metric III is not a cross-adaptive masking metric as it derives the masking measurement directly from the summed mix rather than the relationship between multitrack. Notice the notation M_{mix} is used for the set of a multitrack, rather than M_n for each track.

Metric IV: Cross-Adaptive Multitrack MPEG Masking Metric

Next, we adapt the masking threshold algorithm from MPEG audio coding into a multitrack masking metric based on a cross-adaptive architecture (Reiss, 2011; Zolzer, 2011). The flowchart of the system is illustrated in Figure 5.4.

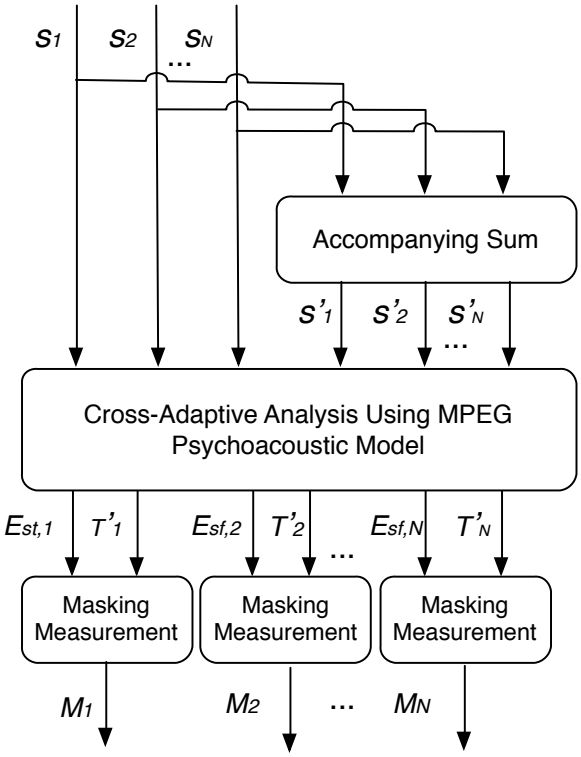


Figure 5.4 System flowchart of the proposed MPEG cross-adaptive multitrack masking model of *N* input signal.

To account for the masking that is imposed on an arbitrary track by the other accompanying tracks rather than by itself, we replace $T(sb)$ with $T_n'(sb)$, which is the masking threshold of track n caused by the sum of its accompanying tracks. Let H denote all the mathematical transformations of the MPEG psychoacoustic model (see Section 2.3.2.1) to derive the masking threshold. We thus can compute $T_n'(sb)$ as

$$T_n'(sb) = H\left(\sum_{i=1, i\neq n}^{N} s_i\right). \tag{0.0}$$

$E_{sf,n}(sb)$ denotes the energy at each scale-factor band of track n. We assume masking occurs at any scale-factor band where $T_n'(sb) > E_{sf,n}(sb)$. The Masker-to-Signal Ratio in multitrack content becomes

$$MSR_n(sb) = 10\log_{10}(\frac{T_n'(sb)}{E_{sf,n}(sb)}). \tag{0.0}$$

We then can define a cross-adaptive multitrack masking measurement for each track, M_n, as

$$M_n = \sum_{sb \subset E_{sf,n} < T_n'} (\frac{MSR_n(sb)}{T_{\max}}). \tag{0.0}$$

5.5 Conclusions

First, a loudness matching experiment on musical signals using a method of adjustment was conducted to evaluate the performance of proposed partial loudness model. Empirical results suggested the proposed loudness model over-estimated the loudness reduction due to partial masking. An adjustment of the parameter K in the partial loudness implementation was proposed that yields a better compliance between model predictions and subjective evaluation.

We incorporated the K parameter modification into the multitrack loudness model (Glasberg & Moore, 2002; Moore et al., 1997). We then adapted this model into two cross-adaptive multitrack masking metrics to describe the amount of masking in multitrack content. We also adapted and extended the masking threshold algorithm of the psychoacoustic model (ISO, 1993; Johnston, 1988a, 1988b) used in MPEG audio coding into another two masking metrics. However, objective and subjective evaluations of the proposed masking metrics are presented in Chapter 6, where they are integrated into an autonomous masking minimization system built upon a typical optimization framework.

Chapter 6

General Processing

6.1 Introduction

So far, we have investigated and explored the intelligent equalization techniques in Chapter 3, intelligent multitrack dynamic range compression in Chapter 4 and a perceptual study on masking in Chapter 5, in which we proposed several masking metrics for multitrack mixing. In this chapter, we aim to integrate previous findings into one intelligent system of masking minimization, built upon an optimization framework that replicates the iterative process of human mixing.

Equalization can effectively reduce masking by manipulating the spectral contour of different instruments to reduce interference in frequency domain. Dynamic range processing can alter the dynamic contour of the signals to reduce the masking over time. As discussed in Section 2.4.3, the operational nature of the equalizer and dynamic processor gives insight into a manner in which they may be combined into a general frequency and a dynamic processing framework. It can create a larger control space and more detailed adjustments to the audio environment, providing invaluable advantages in our intelligent mixing system.

Previous attempts to perform masking reduction for audio mixing have been discussed in Section 2.5. Following the discussion in Section 2.4.1, we saw that mixing is a quintessential optimization problem which benefits from an iterative coarse-to-fine search. This provides some insight regarding the methodology of automating the mixing process to perform masking reduction. Given a certain set of controls of a multitrack, a mixing output can be thought of as the optimal solution to a system of equations that describe the masking behaviour within the multitrack mixture.

In this chapter, we investigate how to use different audio processing techniques to manipulate the frequency and dynamics characteristics of the signal in order to reduce masking. An optimization framework (Section 6.2) is employed, in which we introduce a general frequency and dynamic processing processor. Ultimately, we propose an autonomous masking minimization system, where the aforementioned masking metrics (presented in Chapter 5) are employed to describe the objective function in Section 6.3. The automated audio effects proposed in this chapter are not time-varying compared to previous research. Various implementations of the system are explored and evaluated objectively and subjectively through a listening experiment.

6.2 Audio Effects and Control Parameters

We first investigate how to use different audio processing techniques to manipulate the frequency and dynamics characteristics of the signal to reduce masking in an optimization framework. The extracted control parameters optimized iteratively through the system are described in this section.

6.2.1 Equalization

A six-band equalizer is explored in the optimization process. Six different second-order IIR filters are connected in cascade to equalize the audio signal over the typical frequency range. The filter specification is shown in Table 6.1.

Table 6.1 Six-band equalizer filter design specifications.

Band No.	Center Frequency (Hz)	Q-factor
1	75	1
2	100	0.6
3	250	0.3
4	750	0.3
5	2500	0.2
6	7500	1

The gains of the six-band equalizer filter for each track are varied through the optimization procedure. The control parameters are thus given by

$$\mathbf{x} = [\mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_N], \tag{0.0}$$

in which for each $\mathbf{g}_i$ (vector-valued)

$$\mathbf{g_i} = [\ g_{1,i} \quad g_{2,i} \quad \cdots \quad g_{6,i}\ ], \tag{0.0}$$

contains the six gains control for each track.

6.2.2 Dynamic Range Compression

The digital compressor model design employed in our approach is a feed-forward compressor with smoothed branching peak detector (Giannoulis et al., 2012a). A typical set of parameters of a dynamic range compressor (DRC) includes the threshold, ratio, knee width, attack, release, and make-up gain. In the case of adjusting the dynamic of the signal to reduce masking through optimization, the values of threshold, ratio, knee, attack and release are control parameters to be optimized. Since dynamics are our main focus here rather than the level, make-up gain of each track is set to compensate the loudness differences (measured by the ITU 1770 loudness standard (ITU, 2012a)) before and after dynamic processing. The make-up gain for each track is given by

$$g_{\triangle,i} = L_{ITU,i} - L'_{ITU,i}, \tag{0.0}$$

where $L_{ITU,i}$, $L'_{ITU,i}$ represent the measured loudness before and after the dynamic range compression respectively. The control parameters in the dynamic case are given by

$$\begin{aligned} \mathbf{x} &= [\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_N] \\ \mathbf{d_i} &= [\ T_i \quad R_i \quad K_i \quad a_i \quad r_i\ ]\ , \end{aligned} \tag{0.0}$$

where $\mathbf{d}_i$ is constituted of the five standard DRC control parameters for the i^{th} track; threshold (T_i), ratio (R_i), knee (K_i), attack (a_i) and release (r_i).

6.2.3 General Frequency and Dynamics Processing

We adapt the integrated processor concept proposed in (Wise, 2009). Conventional multiband compressors compress frequency bands differently through band-pass filters or crossover filters. The general processor utilizes this concept, but replaces crossover filters with parametric equalizer filters. It offers larger control over the dynamics of specific frequencies of the audio.

The general processor can adjust the frequency, gain, and bandwidth of a filter, with controls common dynamic range compression controls. The attack and release determine how fast the dynamic EQ acts towards the defined amount of boost or cut. The characteristic of the processing on each frequency band (j is the frequency band index) is controlled by 4 parameters: EQ gain ($g_{j,i}$), threshold ($T_{j,i}$), attack ($a_{j,i}$) and release ($r_{j,i}$). The functionality of DRC's ratio is replaced by the EQ gain, and knee is set to zero as default. If 6 equalization filters are used, then there will be 4-by-6 parameters to be optimized for each instrument track. The notation of the final control parameters to be optimized in the general processing tool is given by

$$\mathbf{x} = [\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_N]. \tag{0.0}$$

In this case, for each $\mathbf{c}_i$:

$$\mathbf{c}_i = \begin{pmatrix} g_{1,i} & g_{2,i} & \cdots & g_{6,i} \\ T_{1,i} & T_{2,i} & \cdots & T_{6,i} \\ a_{1,i} & a_{2,i} & \cdots & a_{6,i} \\ r_{1,i} & r_{2,i} & \cdots & r_{6,i} \end{pmatrix}. \tag{0.0}$$

6.3 Optimization Method and Implementations

The multitrack masking minimization process is treated as an optimization problem concerned with minimizing a vector-valued objective function described by the masking metrics. It systematically varies the input variables, which are the control parameters of the audio effect to be applied, and computes the value of the function until its error is within a tolerance value or a maximum number of iterations is reached.

6.3.1 Objective Function

Let N denote the total number of tracks in the multitrack and K denote the total number of control parameters, which depends on the effects to be applied. The objective function, $M(\mathbf{x})$ can be expressed by the masking metrics as a vector-valued function of the control parameters, $\mathbf{x}$, for each individual track:

$$M(\mathbf{x}) = \begin{bmatrix} M_1(\mathbf{x}) \\ M_2(\mathbf{x}) \\ \cdots \\ M_N(\mathbf{x}) \end{bmatrix}, \tag{0.0}$$

Each component of objective function $M_n(\mathbf{x})$, describes the amount of masking occurring in each track as a function of the control parameters $\mathbf{x}$. Note that $\mathbf{x}$ represents the whole set of the control parameters for all tracks. Changes in the control parameter in one track not only affect the masking of that particular track but also masking of all other tracks.

The derivation of M_n is from the masking metrics proposed in Chapter 5, namely, Metric I, II, III, and IV, as summarized in Equation (0.0):

$$\text{Metric I:} \quad M_n = \frac{L_n - P_n}{L_n}$$

$$\text{Metric II:} \quad M_n = \frac{\sum_{b_{ERB}} E_{m,n}}{\sum_{b_{ERB}} E_{t,n}} \qquad (0.0)$$

$$\text{Metric III:} \quad M_{mix} = \sum_{sb \subset E_{sf} < T} \frac{MSR(sb)}{T_{\max}}$$

$$\text{Metric IV:} \quad M_n = \sum_{sb \subset E_{sf,n} < T'_n} \left(\frac{MSR_n(sb)}{T_{\max}}\right)$$

Detailed descriptions of each metric to produce the masking measurements for multitrack mixing are presented in Section 5.3 and 5.4. Since Metric III is a non cross-adaptive masking metric, it measures the amount of masking occurring in the final mix instead of the multitrack. Therefore when using Metric III to describe the objective function, $M(\mathbf{x})$ becomes:

$$M(\mathbf{x}) = \left[M_{mix}(\mathbf{x}) \right]. \qquad (0.0)$$

6.3.2 Numerical Optimization Algorithms

Numerical optimization theory is employed to find the optimal set of control parameters $\mathbf{x}$ that is a local minimizer to $M(\mathbf{x})$ as shown in Equation (0.0).

$$\min_{\mathbf{x}} \left|\left| M(\mathbf{x}) \right|\right|_2^2 = \min_{\mathbf{x}} \left(M_1(\mathbf{x})^2 + M_2(\mathbf{x})^2 + \ldots + M_N(\mathbf{x})^2 \right) \qquad (0.0)$$

We chose to use the Levenberg-Marquardt Algorithm (LMA) (Marquardt, 1963; Pujol, 2007) to solve this nonlinear least-squares problem. LMA lies between the Gauss-Newton algorithm and the method of gradient descent. LMA as a local optimization algorithm is more suitable than other global optimization algorithms (such as genetic and pattern search algorithms) for the masking minimization problem. It finds the smallest objective function value in some feasible neighbourhood rather than all space. Use of local minima can avoid possible extreme control values (which may include is the global minimum) that might cause unpleasant sound artifacts in the mix.

LMA embeds an iterative procedure like other optimization algorithms. To start the masking optimization process, an initial guess for the control parameters, $\mathbf{x}_0$ has to be provided. A search direction $\mathbf{d}_q$ in the control parameters where the error is decreasing most rapidly, is computed at each iteration, q. In each iteration step, the control parameter, $\mathbf{x}$, is replaced by a new estimate, $\mathbf{x}+\mathbf{d}_q$. To determine the search direction, the new value of the objective function is approximated by the following linearization,

$$M(\mathbf{x}+\mathbf{d}_\mathbf{q}) \approx M(\mathbf{x}) + J\mathbf{d}_\mathbf{q}. \tag{0.0}$$

The Levenberg-Marquardt method obtains a search direction that is a solution of the linear set of equations:

$$(J(\mathbf{x}_\mathbf{q})^T J(\mathbf{x}_\mathbf{q}) + \lambda_q \mathbf{I})\mathbf{d}_\mathbf{k} = -J(\mathbf{x}_\mathbf{q})^T M(\mathbf{x}_\mathbf{q}). \tag{0.0}$$

The damping factor λ controls both the magnitude and direction of $\mathbf{d}_q$ and $\mathbf{I}$ is the identity matrix. The first-order partial derivatives of the objective function give the Jacobian N-by-K matrix,

$$J = \begin{bmatrix} J_{11} & J_{12} & \cdots & J_{1k} \\ J_{21} & J_{22} & \cdots & J_{2k} \\ \vdots & \vdots & & \vdots \\ J_{n1} & J_{n2} & \cdots & J_{nk} \end{bmatrix}, \tag{0.0}$$

where

$$J_{nk} = \frac{\partial M_n(\mathbf{x})}{\partial x_k}. \tag{0.0}$$

λ is adjustable at each iteration. The value of λ decreases with a rapid reduction of M (similar to the Gauss–Newton algorithm), though if iteration gives insufficient reduction in the residual, λ can be increased (similar to the gradient descent method).

6.3.3 Optimization System Variations

When applying Metric I-IV (as described in Chapter 5) to the optimization system, different optimization constraints have to be considered in order to avoid sound artifacts. We propose several implementation variations (listed in Table 6.2) to investigate the best approaches for further evaluation.

Table 6.2 List of different optimization implementations paired with different optimization constraints. Selected implementations (bolded and shaded) are further analysed and evaluated in the following section. The last column gives the notations used in the following section to indicate applied masking metrics.

Masking Metric (IMP. ID)	Constraints	Notation with different effects
I (a)	-	-
I (b)	$L_n^{'} = L_n$ (Maintain loudness)	-
I (c)	$P_n^{'} = P_n$ **(Maintain partial loudness)**	**EQ: EQ-GM** **DRC: DRC-GM** **GE: GE-GM**
II (a)	-	-
II (b)	$L_n^{'} = L_n$ (Maintain loudness)	-
II (c)	$P_n^{'} = P_n$ (Maintain partial loudness)	-
III (a)	-	-
III (b)	$L_{ITU,mix}^{'} = L_{ITU,mix}$ (Maintain ITU loudness of the mix)	-
IV (a)	-	**EQ: EQ-MPEG** **DRC: DRC-MPEG** **GE: GE-MPEG**
IV (b)	$L_{ITU,n}^{'} = L_{ITU,n}$ (Maintain ITU loudness)	-

As shown in Table 6.2, Implementation I (a) applies the Glasberg and Moore based Metric I without any constraint. Implementation I (b) employs Metric I with a before-and-after loudness constraint. That is, the control parameters are optimized at every optimization iteration to reduce masking and comply with a loudness condition, as given by

$$L_n^{'} = L_n, \tag{0.0}$$

where L_n, $L_n^{'}$ are the loudness of track n before and after applying the optimized audio effects respectively. Implementation I (c) also uses Metric I but constrained with a before-and-after partial loudness condition.

$$P_n^{'} = P_n \tag{0.0}$$

Implementation II follows the same logic for its constraints as Implementation I, but using Metric II instead. Implementation III (a) is based on MPEG Metric III with no constraint applied. In Implementation III (b), an equal before-and-after ITU loudness constraint of the mix is added, as given by

$$L_{ITU,mix}^{'} = L_{ITU,mix}, \tag{0.0}$$

where $L_{ITU,mix}$, $L_{ITU,mix}^{'}$ are the ITU loudness of the mix at every optimization iteration, measured by (ITU, 2003). Implementation IV follows the same logic as III but using Metric IV instead. For IV (b), the optimization constraints are given by

$$L_{ITU,n}^{'} = L_{ITU,n}, \tag{0.0}$$

where $L_{ITU,n}$, $L_{ITU,n}^{'}$ are the ITU loudness of every track at every optimization iteration.

The mix quality produced by some of the implementation variations is significantly better than others in terms of sound artifacts and the ability to reduce perceptual masking through listening evaluation.

After informal listening, Implementation II (a)(b)(c) were rejected from further study since they produced mixes with obvious sound artifacts due to sharp EQ, heavy compression, compared to Implementations I (a)(b)(c). Implementation I (a) often applied attenuation on all frequency bands, which could be seen as decreasing the volume of each track. As a result, it created final mixes with quiet level, yet masking still persists in the mix. This perhaps is due

to the level-dependent nature of the loudness model of Glasberg and Moore. The model underestimates the perceived amount of masking within signals of lower level. Furthermore, Implementation I (c) was favoured slightly but consistently over I (b) in terms of overall mix quality.

Implementation III exhibited limited and inconclusive masking reduction. This could due to the mechanism of Metric III since it only quantifies the amount of masking within the final mixture rather than the masking relationship between tracks. Therefore III (a)(b) were also rejected from further use. Informal listening also suggested that Implementation IV (a) consistently had better masking reduction performance than IV (b).

Thus, Implementation I (c) and IV (a) were kept for further study, as indicated in Table 6.2. These implementations were each paired with 3 different audio effects (described in Section 6.2, namely, equalization (EQ), dynamic range compression (DRC) and general processing (GE)) in the optimization algorithm to perform masking reduction. The notations for every audio effect case are given in the last column of Table 6.2. Finally, DRC-GM was rejected since it produced significant pumping and breathing artifacts (Izhaki, 2013).

6.4 Results and Evaluation

6.4.1 Optimization Results

In this section, optimization results of a 4-track multitrack recording of 20 seconds are presented and discussed. The rock multitrack song is the 'Song 2' used in the following subjective evaluation as shown in Table 6.3. It has 4 instrument tracks/stems (track 1: bass; track 2: drum set; track 3: electric guitar; track 4: Synth). Specification of the song is shown in Figure 6.1.

Figure 6.1 Specification of the test song (the reference level is the lowest possible sample is for 16 bit audio in digital full scale: 96 dBFS).

Genre	Instrument	Level (dB)
Rock	Bass	68
	Drum set	57
	Electric guitar	61
	Synth	65

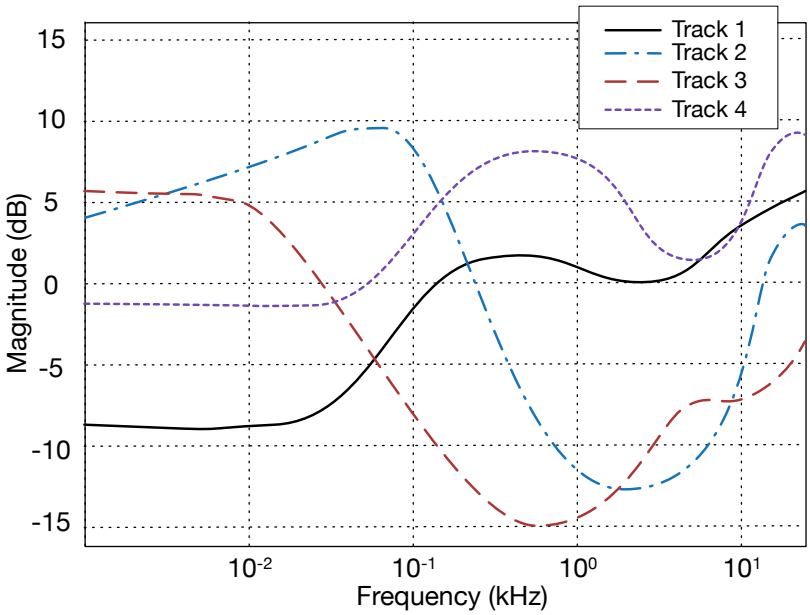


Figure 6.2 EQ curves of each track using EQ-GM, on a 4-track multitrack song.

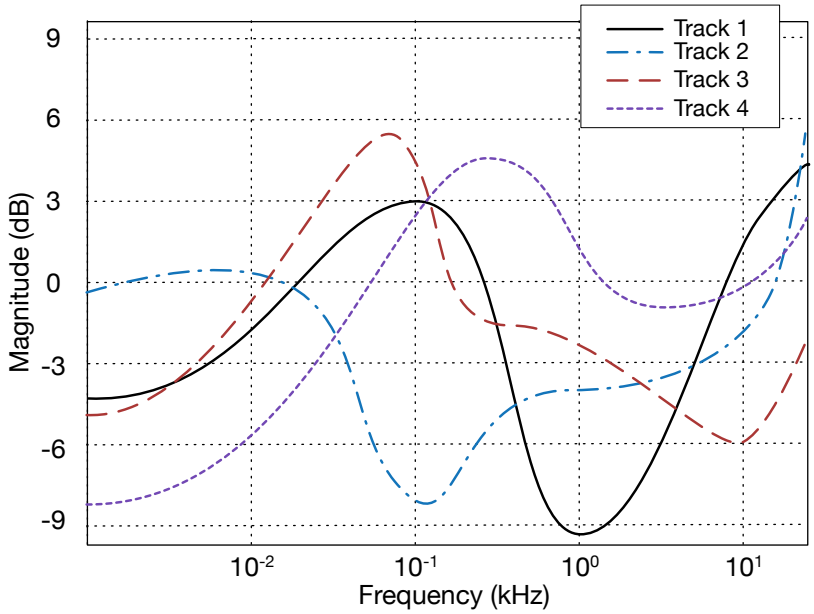


Figure 6.3 EQ curves of each track using EQ-MPEG, on a 4-track multitrack song.

The optimized EQ curves based on the EQ-GM and EQ-MPEG optimization methods on the same 4-track multitrack song are shown in Figure 6.2 and Figure 6.3 respectively. It shows that different masking metrics produce significantly different EQ results on the same instrument track. In particular, EQ-GM produces relatively sharper EQ curves than EQ-MPEG.

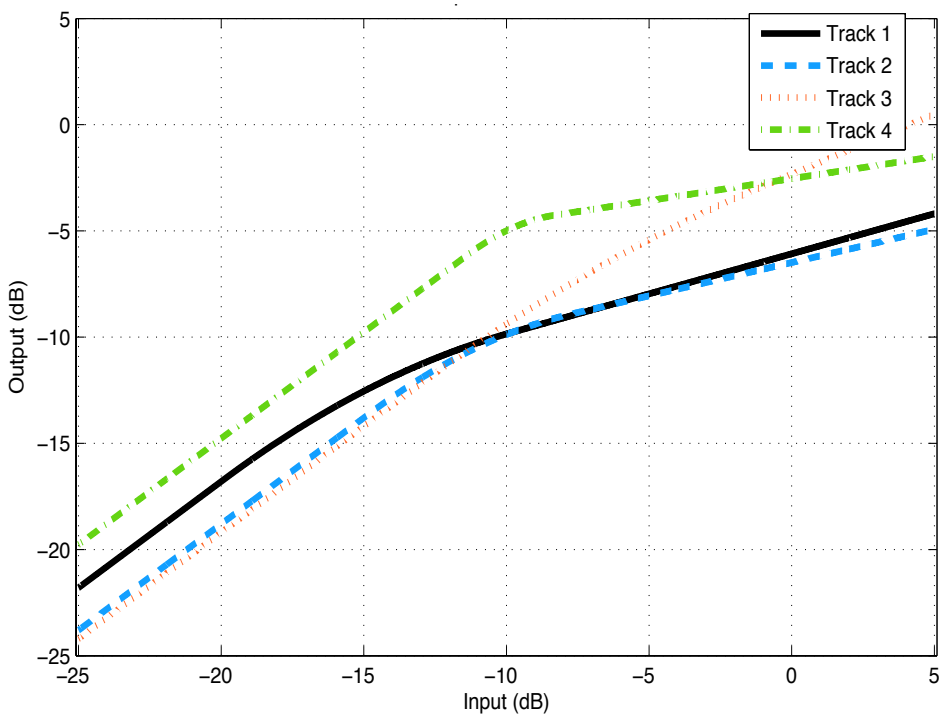


Figure 6.4 Static DRC curves of each track using EQ-GM, on a 4-track multitrack song.

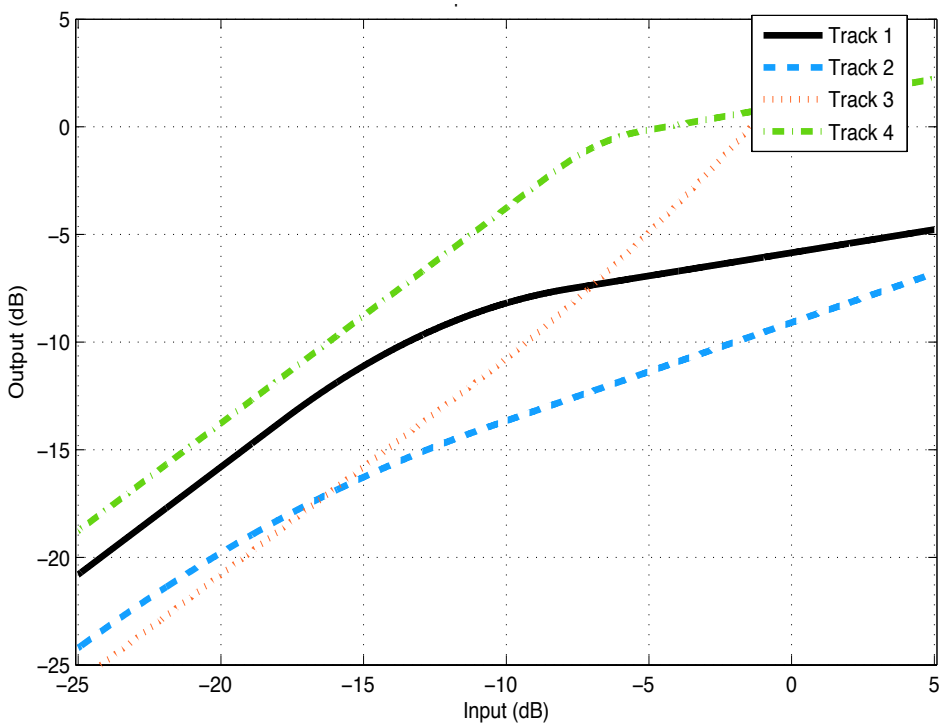
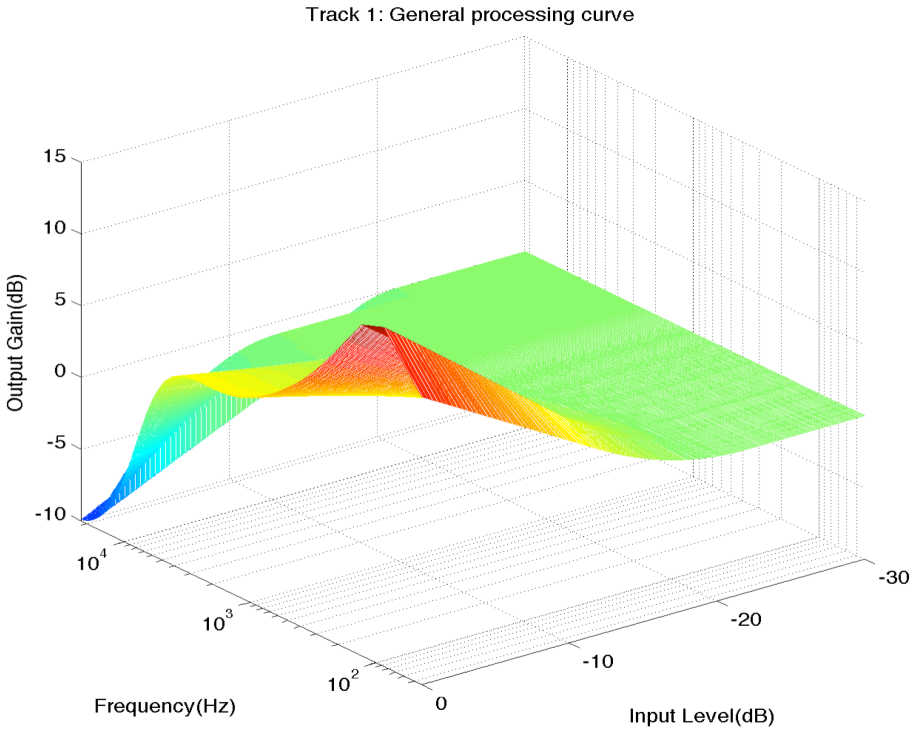
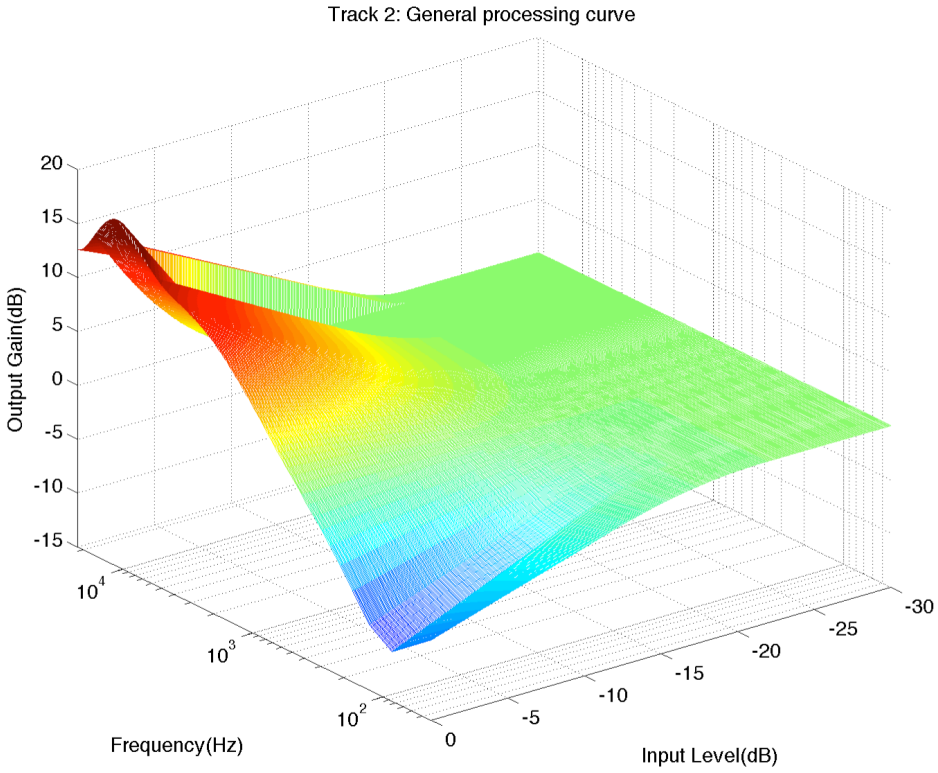


Figure 6.5 Static DRC curves of each track using EQ-MPEG, on a 4-track multitrack song

The optimized DRC characteristics for DRC-GM and DRC-MPEG are shown in Figure 6.4 and Figure 6.5 respectively. As the Figures suggest, optimized threshold and ratio values are dependent on the masking metric. Notably, for track 3, the GM metric produces a small

amount of downward compression (ratio = 1.69) while MPEG metric generates an upward compression (ratio = 0.86) instead.





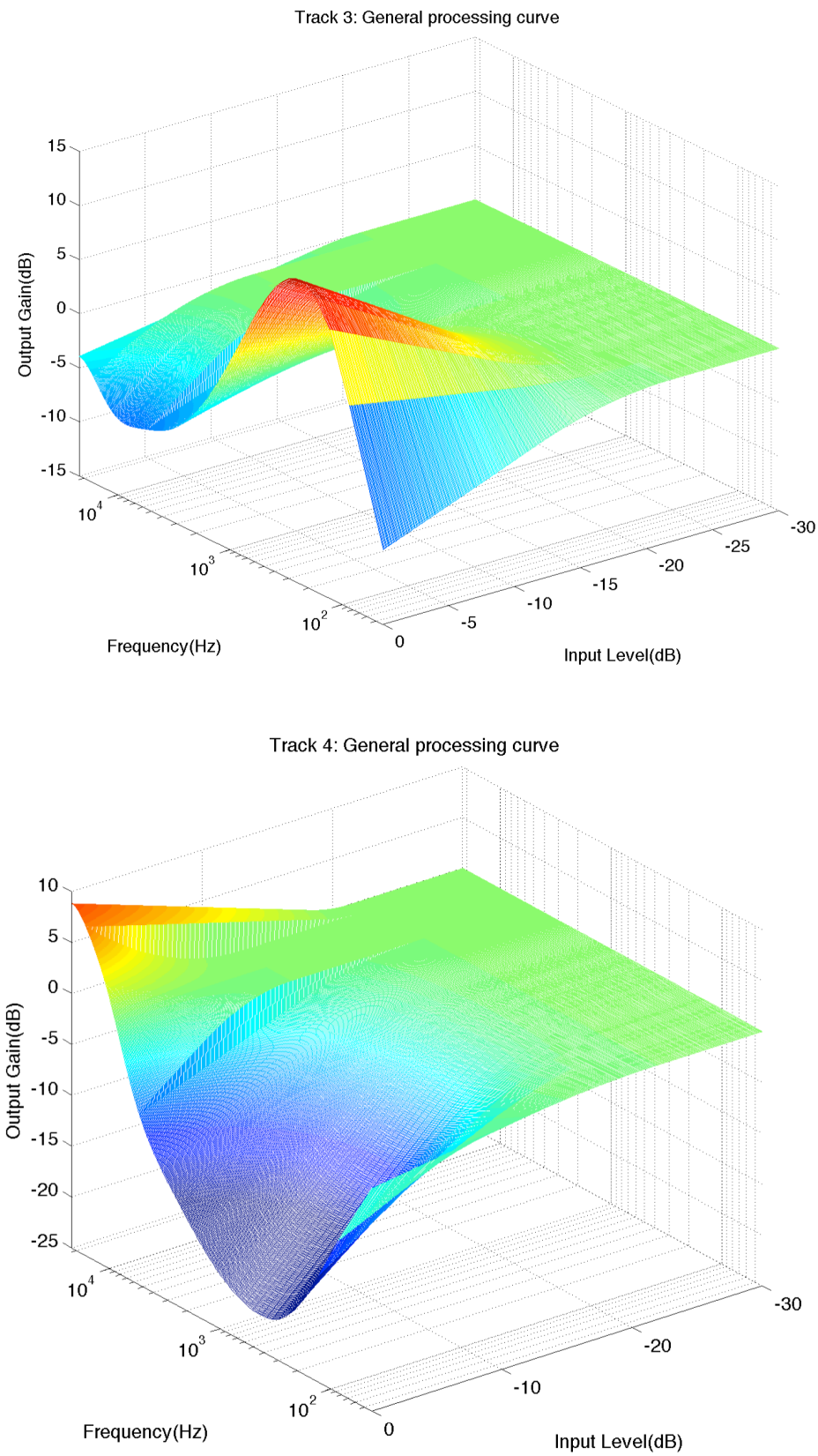
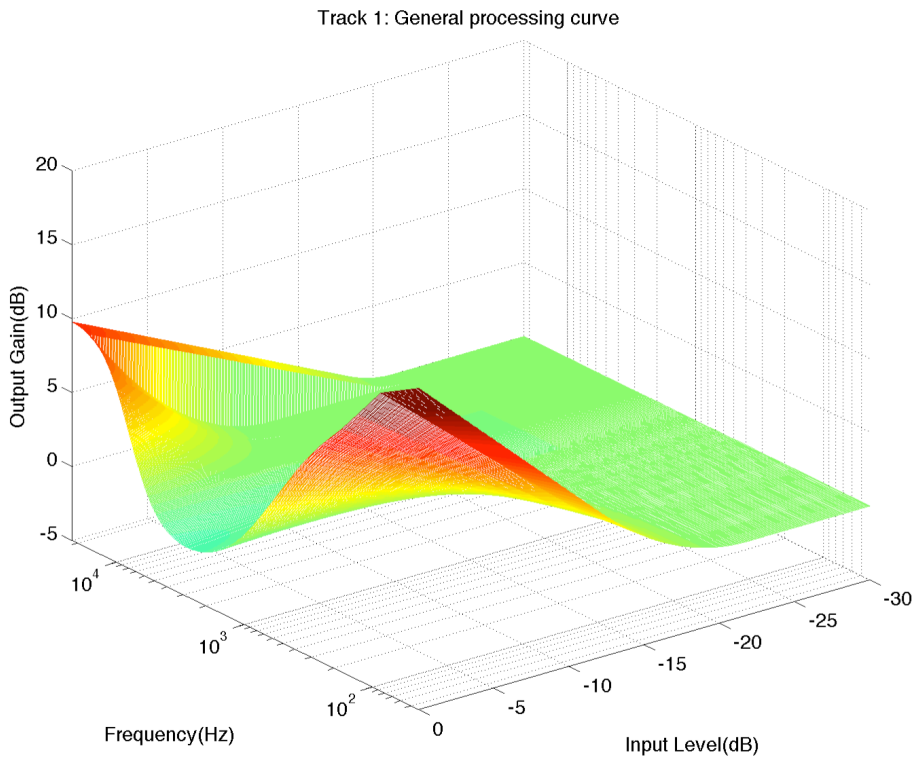
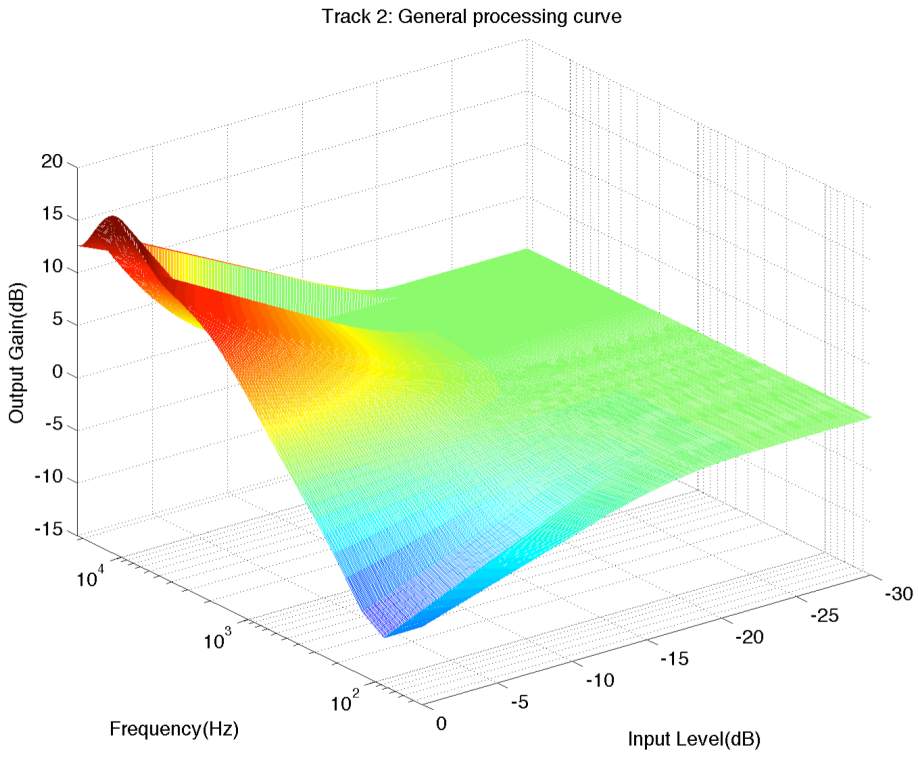


Figure 6.6 General processing curves based GM masking metric.


Track 1: General processing curve
Output Gain(dB)
20
15
10
5
0
-5
10^4
10^3
10^2
Frequency(Hz)
0
-5
-10
-15
-20
-25
-30
Input Level(dB)


Track 2: General processing curve
Output Gain(dB)
20
15
10
5
0
-5
-10
-15
10^4
10^3
10^2
Frequency(Hz)
0
-5
-10
-15
-20
-25
-30
Input Level(dB)

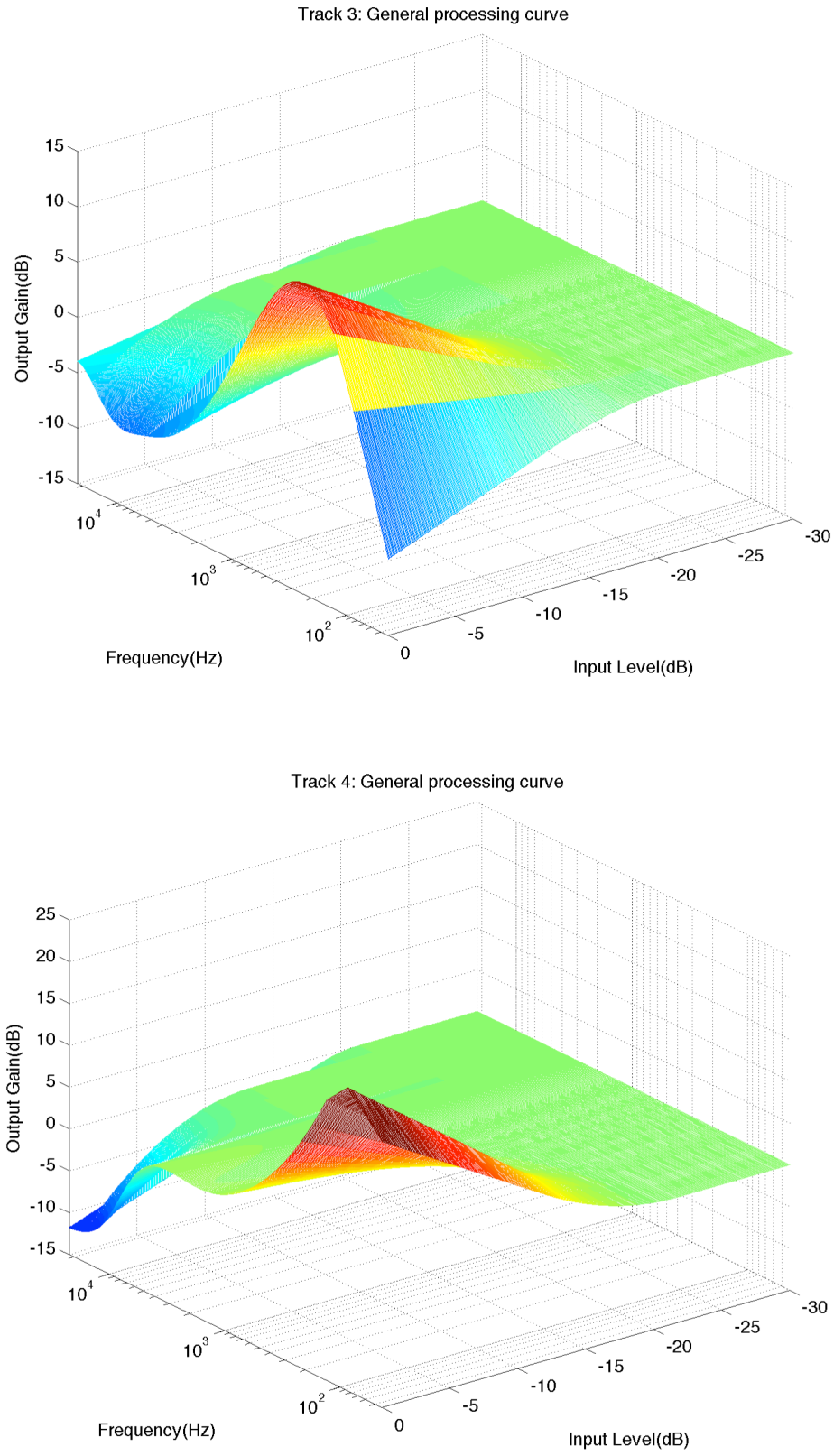


Figure 6.7 General processing curves of track 1 based MPEG masking metric.

Selected examples of the optimized GE parameters of track 1 are visualized in a three dimensional space in Figure 6.6 and Figure 6.7.

Together, these results show that the optimized parameters of the three audio processors (EQ, DRC, GE) are significantly different dependent on the masking metrics (GM, MPEG) used. The nature of these two masking metrics could shed light on why the difference occurs. The MPEG metric is based on masking threshold, where the final masking value is a function of frequency band. But the GM metric is a function of loudness, which is based on the overall loudness reduction. In other words, the GM metric is only indirectly frequency dependent. In order to achieve the same amount of masking reduction as the MPEG metric, the GM metric might have to apply more severe audio effects.

The computational complexity of the optimization algorithms depends greatly on which masking metric is used. MPEG metric requires much shorter processing time than GE metric (at least, 100 times less). This is due the complexity of the algorithm behind the Glasberg and Moore's loudness model (mainly due the calculation of excitation pattern). When using the same metric, the computational performance is influenced by a number of factors such as the number of iterations, the numbers of variables to be optimized and the number of the tracks within the mix. At the current stage of the research, these optimization algorithms are not efficient enough to be embedded into commercial product, real-time processing at least.

6.4.2 Subjective Evaluation

Method

We conducted a formal subjective evaluation in the form of a multiple stimulus listening test, similar to MUSHRA (ITU, 2003), to assess the performance of the five selected implementations against raw mixes and professional mixes. However, unlike MUSHRA, no fixed reference was available, and thus it can be considered a semantic differential test. Raw mix is the direct sum of the unprocessed tracks. For the professional mix, a mix engineer with 3-year professional mixing experience and 5-year musician experience was asked to create his own mix with the objective of reducing the masking, using Apple's Logic Pro software. He

was instructed to only use built-in dynamic range compression and equalizer. And he was allowed to mix the songs with preferred playback level as his own. However editing, rerecording, the use of samples or any other form of adding new audio was not allowed.

Five multitrack recordings (20s segments) in various genres, selected from the Open Multitrack Testbed (http://multitrack.eecs.qmul.ac.uk) (De Man et al., 2014), were used in the test. The specification of the tested songs is shown in Table 6.3.

Table 6.3 Specification of tested songs.

No.	Number of Tracks	Genre	Instrumentation
1	3	Jazz	Bass; Drum; Piano
2	4	Rock	Bass; Drum; Electric Guitar; Synth
3	6	Rock	Percussion; Bass; Drum; Guitar 1; Guitar 2; Keys
4	7	Hip Hop	Bell(synth); Bass; Backing Vocal; Leading Vocal; Juno(synth); Piano; Drum
5	9	Punk	Bass; Drum; Electric Guitar; Leading Vocal; Percussion; Sub-bass; Acoustic Guitar; Vibes; Backing Vocal

The loudness of the final mixes was normalized manually by a group of professional mixing engineers, using the same playback system as used for the subjective evaluation. The orders of mix variations and songs presented to participants were randomized. All tests were performed in a soundproof listening room with the same headphone set-up, where the environmental noise is minimized. Participants were allowed to adjust the playback level during the experiment in order to evaluate the quality of the mixes efficiently.

Eighteen participants with moderate audio engineering experience from two different audio research groups (Queen Mary University of London and Goldsmiths University of London) were recruited. Related personal information about the participant is displayed in Table 6.4, based on the results of the questionnaire, which was given to all participants before commencing the test.

Table 6.4 Results of preliminary questions to test participants.

Gender	Male	12
	Female	6
Audio Group	Queen Mary	8
	Goldsmiths	10
Hearing Impairment	No	18
	Yes	0
Age Range	20 – 36	

Participants were asked to rate the mixes according to two criteria on a full scale of 0 to 100, split up into five descriptors: "Bad (0 -20)", "Poor (20 - 40)", "Fair (40 - 60)", "Good (60 - 80)" and "Excellent (80 - 100)" as shown in Figure 6.8. The average time that participants spent on this experiment is about 50 minutes, including a suggested 5 minutes break between the two questions.

- Q1: Rate the following mixes in terms of the ability to distinguish the sources (i.e., the lack of masking).
- Q2: Rate the following mixes in terms of your own overall preference.

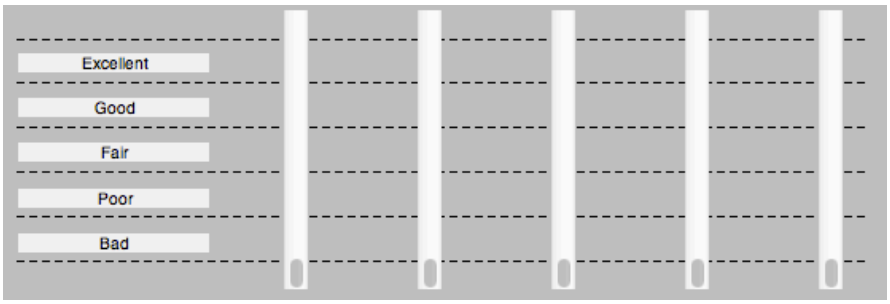


Figure 6.8 The evaluation interface used in the experiment.

Evaluation Results

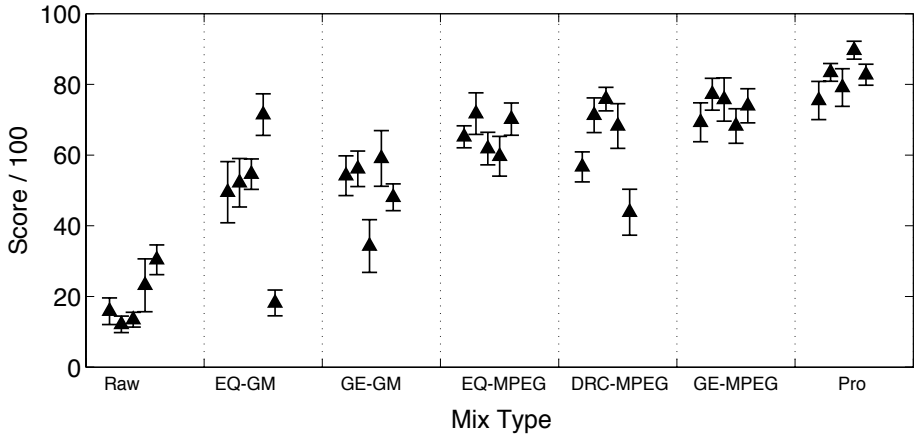
The result analysis follows the idea used in Section 4.5.2, using means with confidence intervals data visualization. We use the Lilliefors test for normality check, Friedman test and Wilcoxon signed rank test for significance check. However, we decide to follow the specification for MUSHRA (ITU, 2003) to visualize the date with mean and confidence intervals: no overlap in the confidence intervals for two conditions means one is significantly better than the other. The normality tests are performed for each song and mix types for both Q1 and Q2. Results are shown in Table 6.5.

Table 6.5 Results of the Lilliefors tests for Q1 and Q2 (h=0 indicate normal, h=1 indicate non-normal).

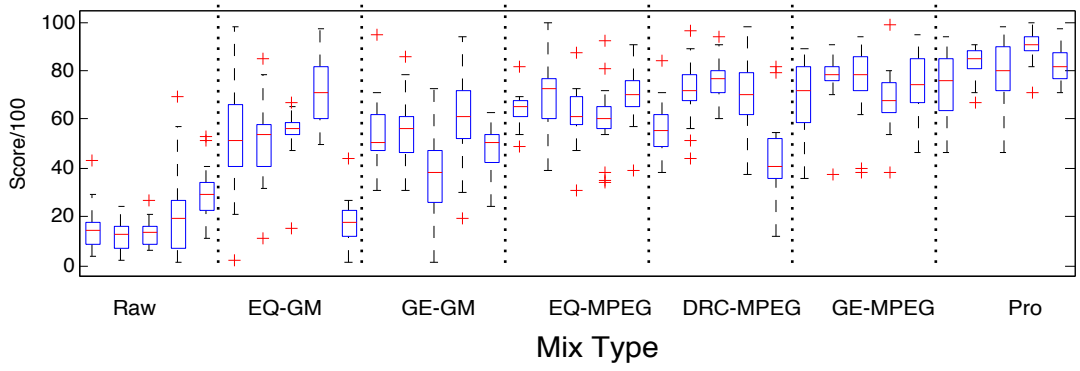
Mix Type	Song	Q1		Q2	
		h	*p*-value	h	*p*-value
Raw	1	0	0.082105143	0	0.499190162
	2	0	0.17677317	0	0.5
	3	0	0.5	0	0.416346307
	4	0	0.055816825	0	0.408682066
	5	0	0.354705565	0	0.227950773
EQ-GM	1	0	0.5	0	0.383944551
	2	0	0.1643441	0	0.399519164
	3	1	0.041908747	0	0.45201578
	4	0	0.5	0	0.385541187
	5	0	0.5	0	0.053795024
GE-GM	1	0	0.1643441	1	0.036102532
	2	1	0.041908747	0	0.5
	3	0	0.5	0	0.338227581
	4	0	0.5	0	0.5
	5	0	0.5	0	0.088990345
EQ-MPEG	1	1	0.041908747	0	0.111509779
	2	0	0.5	0	0.139270692
	3	0	0.5	0	0.5
	4	0	0.5	0	0.257015986
	5	0	0.17677317	1	0.001
DRC-MPEG	1	0	0.5	0	0.236873714
	2	0	0.5	0	0.079271012
	3	0	0.5	0	0.5

	4	0	0.17677317	0	0.5
	5	0	0.34511909	0	0.316400138
GE-MPEG	1	0	0.5	0	0.283357946
	2	0	0.5	1	0.014403988
	3	0	0.17677317	0	0.484454298
	4	0	0.34511909	0	0.5
	5	0	0.267617225	0	0.122612895
Pro	1	0	0.5	0	0.313817463
	2	0	0.17677317	0	0.37541535
	3	0	0.34511909	0	0.5
	4	0	0.267617225	0	0.5
	5	0	0.355108799	0	0.088901724

Subjective evaluation results are summarized in Figure 6.9 and Figure 6.12. The raw mix and professional mix are denoted as 'Raw' and 'Pro'. Notations for optimized mixes are listed in Table 6.2.



(a)



(b)

Figure 6.9 (a) Evaluation results of Q1, which are organized by mix type, showing the mean values (of each song) across all participants with errors bars displaying 95% confidence interval (t-distribution). (b) Boxplot of the same Q1 results.

Figure 6.9 plots the results for Q1. As expected, 'Pro' performs the best on every song. Almost all mixes are rated higher than 'Raw' except song 5 where 'Raw' rates higher than EQ-GM.

7 out of 10 'GM' mixes are rated 'Fair' at masking reduction. Comparison of 'EQ-GM' with 'GE-GM' within each song shows that the mix using the general frequency and dynamic processing technique reduces the masking more effectively.

'MPEG' mixes rate consistently within the 'Good' scale. This suggests that the masking metric based on the MPEG perceptual model (EQ-MPEG, DRC-MPEG, GE-MPEG) has better performance than metrics based on Glasberg and Moore's loudness models when describing the multitrack masking. Results also show that whether 'EQ-MPEG' rates higher than 'DRC-MPEG' is song dependent, and there is no clear preference.

Statistical tests for the significance in terms of mix types and songs are performed are performed as the results are shown in Table 6.6 and Table 6.7.

Table 6.6 Results of the one-way ANOVA of mix types within each song (Q1).

Song	*p*-value
1	5.05e-24
2	6.60e-36
3	3.10e-35
4	5.91e-22
5	9.27e-39

Table 6.7 Results of the one-way ANOVA for song choices within each mix type (Q1).

Mix type	*p*-value
Raw	1.2796e-05
EQ-GM	2.0499e-14
GE-GM	6.9453e-05
EQ-MPEG	6.9453e-05
DRC-MPEG	3.8140e-10
GE-MPEG	0.1897
Pro	0.0015

Table 6.6 indicates there is strong statistical evidence that mix types have significant effect on the evolution scores (*p*-values are all extremely small). Table 6.7 also suggests song choices might also have certain degree of affect on evaluation scores. However *p*-value equals to 0.1897 and 0.015 for 'GE-MPEG' and 'Pro' respectively indicate the otherwise.

Two-way ANOVA test is then performed to investigate the interaction between these two factors (mix types and songs). Table 6.8 indicate that there is some degree of interaction effect between these two factors.

Table 6.8 Two-way ANOVA result table (Q1).

ANOVA Table					
Source	SS	df	MS	F	Prob>F
Column: Mix type	228892.7048	6	38148.7841	213.1661	1.23E-144
Row: Song	8734.5937	4	2183.6484	12.2017	1.53E-09
Interaction	46516.9619	24	1938.2067	10.8302	1.67E-33
Error	106482.8333	595	178.9627		
Total	390627.0937	629			

Therefore multiple comparison tests are performed to see if different mix types and song choice yield significantly different evaluation scores.

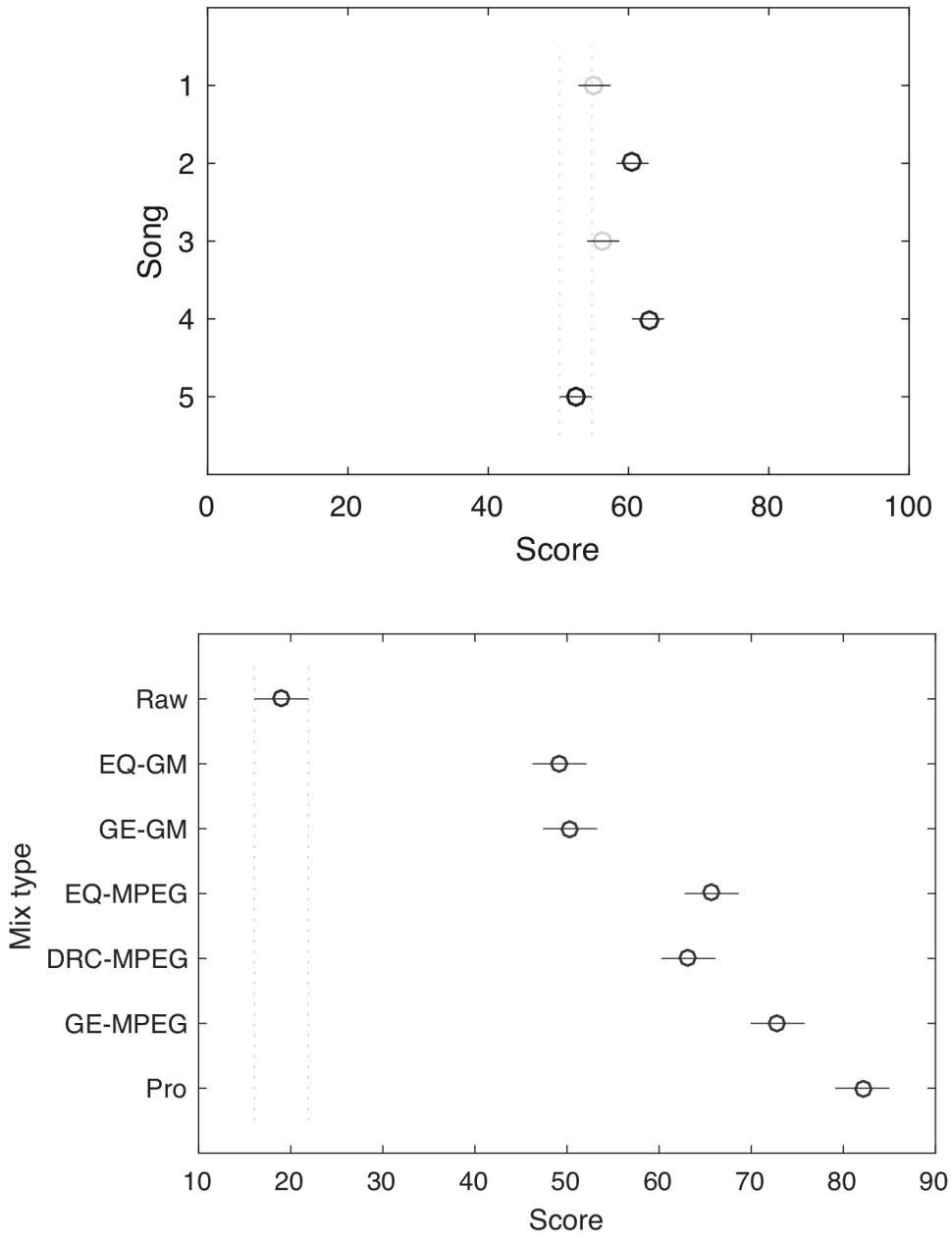


Figure 6.10 The result plots shows multiple comparison of the means with 95% confidence intervals for both mix types and songs.

Results from Figure 6.10 confirm again (see Figure 6.9) that there is significant difference between mix types with 'GE-MPEG' outperform all other mix types apart from the 'Pro'. Mixes produced with 'GM' model are rated lower than mixes produced with "MPEG" model. Histograms of the evaluation scores for 'GE-MPEG' and 'Pro' are shown as in the figure below. 70% of the time, participants evaluated the quality of 'GE-MPEG' as 'Good'

(score: 60-80). As for 'Pro', about 65% of the time, participants think the quality of 'Pro' is "Excellent" (90-100).

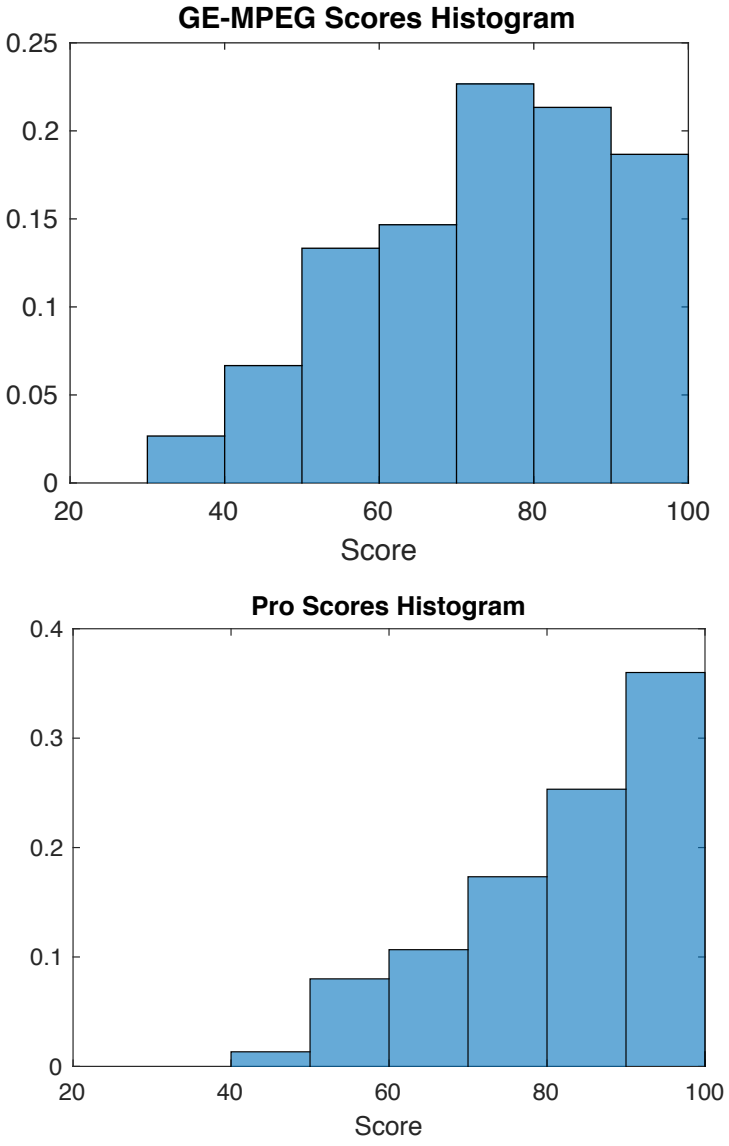
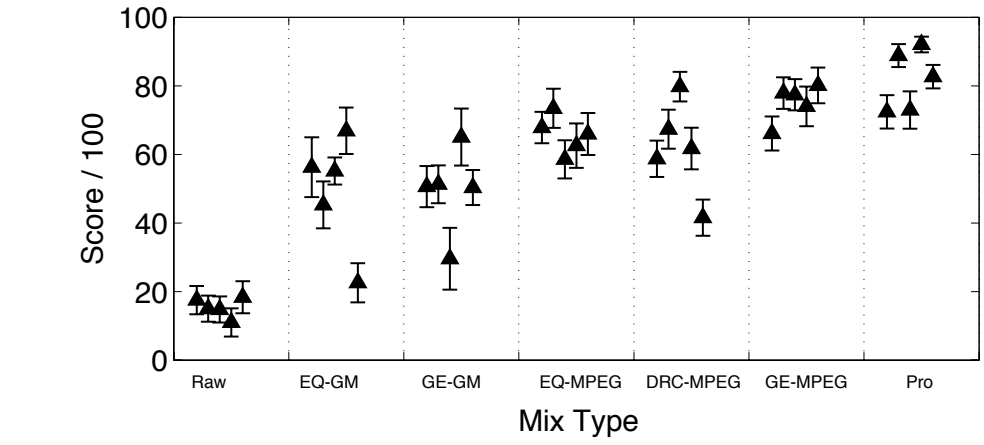
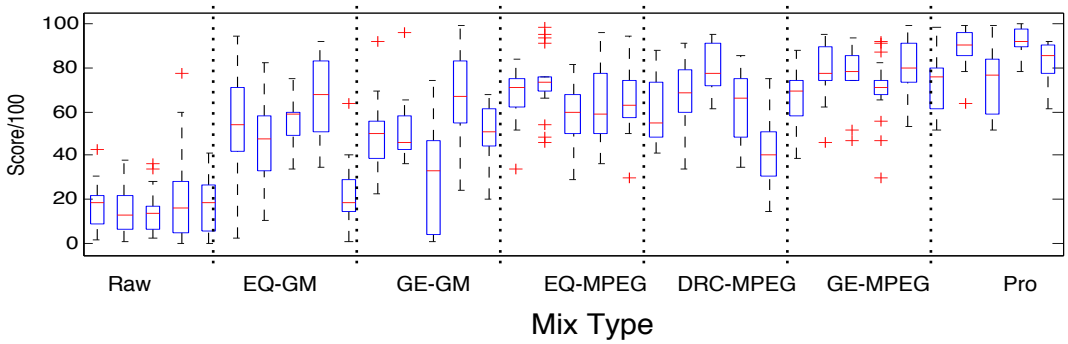


Figure 6.11 Score histograms for GE-MEPG and Pro (Q1).



(a)



(b)

Figure 6.12 (a) Evaluation results of Q2, organized by mix type, showing the mean values (of each song) across all participants with errors bars displaying 95% confidence interval (t-distribution). (b) Boxplot of the same Q2 results.

Evaluation results for overall preference (Q2) of the mixes are shown in Figure 6.12. Although the rankings of the mixes for Q1 and Q2 share a similar pattern, the particular ranking within each song is different. It implies that the amount of masking in the multitrack is strongly related to the overall preference of participants.

Further statistical tests for the significance are performed. Results of the one-way ANOVA tests for mix types and song choices for Q2 are very similar to the result for Q1 (see Table 6.6 and Table 6.7).

Table 6.9 Results of the one-way ANOVA test within each song (Q2).

Song	*p*-value
1	8.2275e-21

2	3.1603e-34
3	3.2366e-32
4	1.3495e-20
5	2.2093e-36

Table 6.10 Results of the one-way ANOVA test within each mix type (Q2)

Mix type	*p*-value
Raw	0.5863
EQ-GM	1.6319e-10
GE-GM	4.3256e-06
EQ-MPEG	0.0426
DRC-MPEG	7.7225e-11
GE-MPEG	0.0069
Pro	1.6828e-08

Table 6.9 indicates there is strong statistical evidence that mix types have significant effect on the evolution scores (*p*-values are all extremely small). Table 6.10 suggest it's hard to conclude whether there is significant different in song choices. Two-way ANOVA test is then performed. Results are shown in Table 6.11. Multiple comparison tests are performed to see if different mix types and song choices yield evaluation scores.

Table 6.11 Two-way ANOVA result table (Q2).

ANOVA Table					
Source	SS	df	MS	F	Prob>F
Column: Mix type	242936.5714	6	40489.4286	191.504	2.22E-135
Row: Song	9389.6603	4	2347.4151	11.1026	1.08E-08
Interaction	47507.2063	24	1979.4669	9.3623	1.48E-28
Error	125800.0556	595	211.4287		
Total	425633.4937	629			

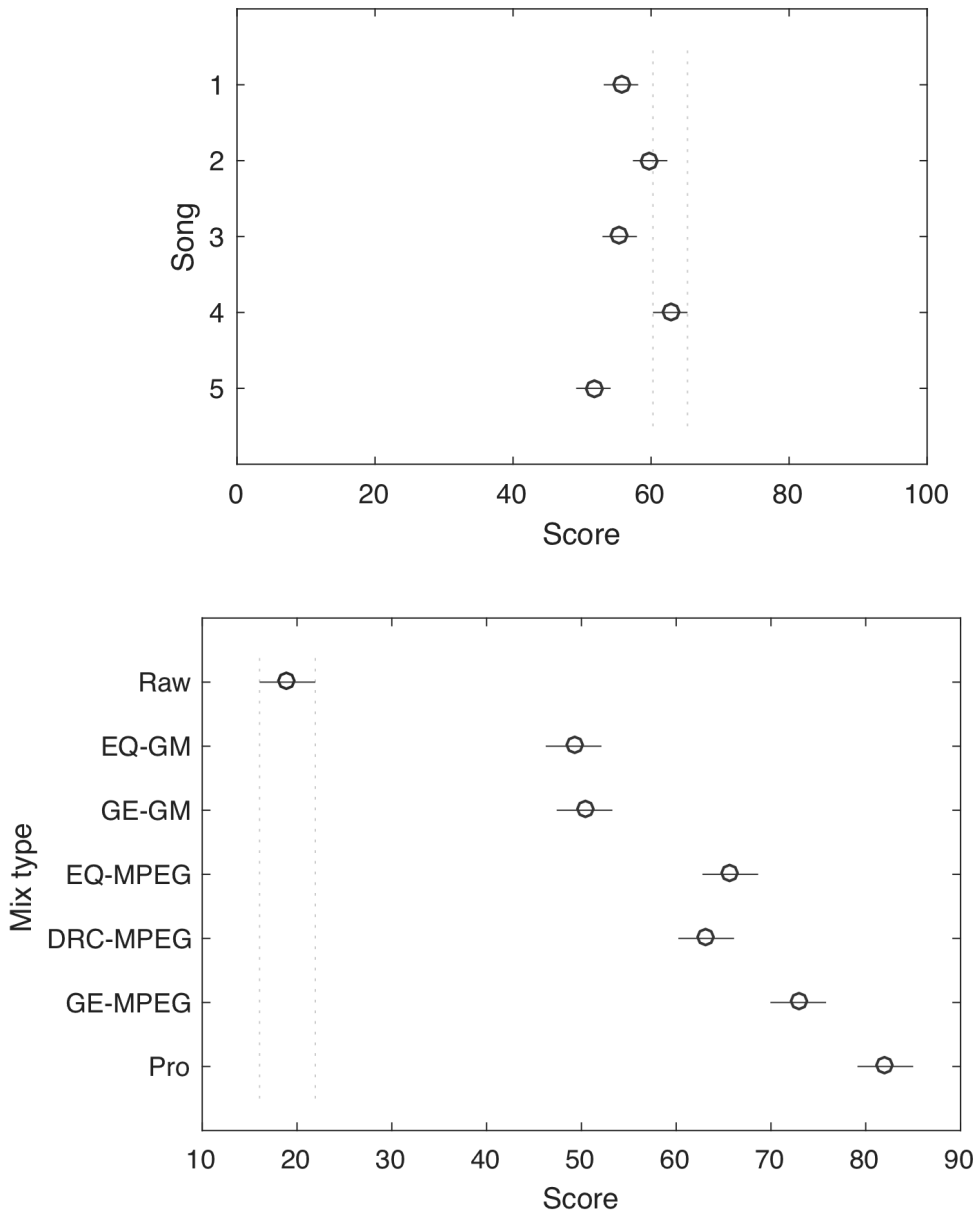


Figure 6.13 The result plots shows multiple comparison of the means with 95% confidence intervals for both mix types and songs.

Results indicate that there is significant difference between mix types with 'GE-MPEG' outperform all other mix types apart from the 'Pro'. Score histograms for 'GE-MEPG' and 'Pro' are shown in Figure 6.14. Mixes produced with 'GM' model are rated lower than mixes

produced with "MPEG" model. No significant difference between 'DRC-MPEG' and 'EQ-MPEG'. Also there is no significant difference between 'EQ-GM' and 'GE-GM'.

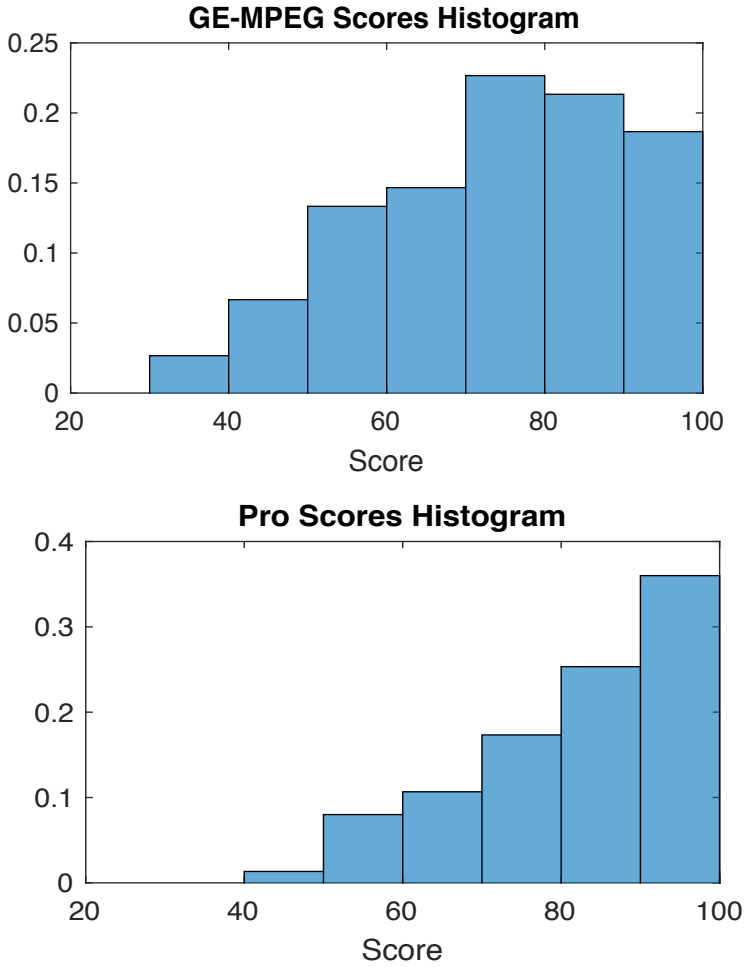


Figure 6.14 Score histograms for GE-MEPG and Pro (Q2).

To give a clearer depiction of the general performance of each mix type, the mean results across all participants and songs are displayed in Figure 6.15. The professional mix is clearly the best assessed in both criteria, and mixes using the masking metric based on the MPEG perceptual model are preferred over those using Glasberg and Moore's loudness models. This is an unexpected result, since Glasberg and Moore's loudness model is considered more advanced than the simple MPEG psychoacoustic model. A possible explanation could be found by comparing the nature of these two models. The GM masking metric is based on the overall loudness reduction but the MPEG masking metric is defined as a function of masked frequency bands. That is, the GM masking metric might not be able to capture masking

behaviour in the higher frequency range, since it may not decrease the overall loudness as much as masking in the low frequency range.

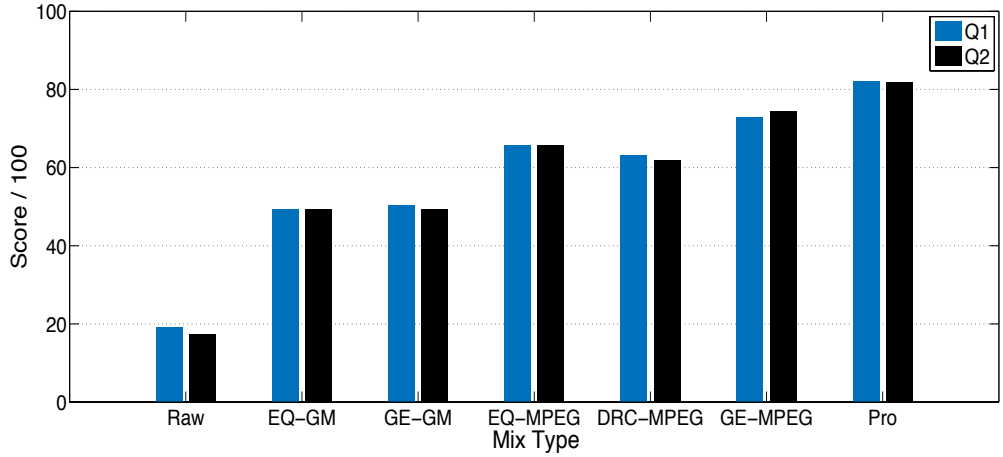


Figure 6.15 Overall mean results across all songs and participants for Q1 and Q2.

Results also suggested that applying dynamic range compression could reduce masking as efficiently as applying equalization. However, general frequency and dynamic processing that intuitively integrates both EQ and DRC functionality can achieve better results in masking reduction than the alternatives, and it had a high overall preference among participants.

Overall, the general frequency and dynamic processing used with the MPEG masking metric (GE-MPEG) performed best among the proposed autonomous masking reduction algorithms. Subjective evaluation showed that this approach could result in a mix that can compete with the mix produced by professional engineer. Future study on whether the instrumentation within the mix or music genre has significantly effect on the performance of the optimization methods (especially GE-MPEG) can provide further evaluation of the proposed automatic mixing algorithm.

6.5 Conclusions

We investigated different audio effects that are commonly used to minimize masking. By exploring the control mechanisms and operating spaces for equalization and dynamic range compression, we presented an integrated, general frequency and dynamic processing algorithm that acts within a higher dimensional control space. We proposed an intelligent

system for masking minimization using numerical optimization technique. Different masking metrics were paired with different audio effects, whose control parameters were obtained iteratively through the optimization process. Finally, formal evaluation of the system was described. The results of the subjective listening experiment implied that our novel MPEG-based metric is able to describe multitrack masking better than more advanced psychoacoustic models (Glasberg & Moore, 2002, 2005; Moore et al., 1997). A general frequency and dynamics processing algorithm was shown to be more powerful in masking minimization than equalization or dynamic range compression in this context. The best masking minimization performance was achieved by incorporating the general tool with the MPEG masking metric.

It would be beneficial to further investigate sophisticated masking models that can appropriately modelling partial masking with 'real world' content. Furthermore, perception of temporal masking is seldom considered in established masking models, and thus this offers a promising future research direction. Since the proposed autonomous masking minimization algorithm only considers local optimization, an interesting extension would be to explore other global optimization algorithms such as pattern/direct search, genetic algorithm and etc.

Chapter 7

Conclusions and Future Work

To conclude the thesis, we first summarize the contributions made in the fields of intelligent mixing, perception modelling and beyond. We then reflect upon possible improvements that can be made to improve the intelligent multitrack frequency and dynamics processing system. Finally we consider some potential avenues for future work.

7.1 Conclusions

In fulfilment of our aim to develop intelligent methods for multitrack frequency and dynamics processing, there have been four main contributions; developments in frequency manipulation, dynamics processing, auditory masking and finally, an integration of previous findings into a general, intelligent system of mix optimization based on masking reduction.

Overall, we have shown that by using a cross-adaptive architecture, feature extraction and analysis, optimization techniques, embedding best practices as control rules and utilizing perceptual models, it is possible to generate intelligent mixing choices of similar quality to those of a skilled audio engineer. This can be achieved with minimal or no human intervention, and can improve the overall experience of listening to musical mixtures.

Chapter 3 investigated the frequency aspect of intelligent mixing. We presented a spectral characteristic analysis of a large commercial recording dataset. We found that the spectra of commercial successful recordings share a consistent trend, which can roughly be described as a linearly decaying distribution of around 5 dB per octave between 100 and 4000 Hz, becoming gradually steeper with higher frequencies, and a severe low-cut around 60 Hz. We then proposed a novel time-varying equalization approach to match the spectral distribution of the input signal to a target equalization curve (such as the common curve obtained from the spectral characteristic studies) or any desired frequency response, based on the Yule-

Walker IIR filter design method. Objective evaluation of the algorithm showed that the algorithm is able to fulfill the objective with appropriate ballistics setting.

Chapter 4 explored the dynamics aspect of intelligent mixing. We proposed a novel intelligent multitrack dynamic range compression algorithm. The algorithm utilises the cross-adaptive digital audio effect architecture again (Reiss, 2011; Zolzer, 2011), exploits the interdependence of the input audio features and incorporates best practices as well as subjective evaluation results to produce the optimal amount of dynamic range compression for multitracks. We presents a fully automated multitrack dynamic range compressor where all classic parameters of a typical compressor (ratio, threshold, knee, attack and release) are dynamically adjusted depending on extracted features and control rules. In the pursuit of better descriptors to characterize the transient nature and spectral content of the signal, two new audio features, namely percussivity weighting and low-frequency weighting, were proposed. A method of adjustment experiment was conducted to uncover how subjects set the ratio and threshold parameters. We applied multiple linear regression models to the subjective results to formulate the ratio and threshold automations that follow the preference of mixing engineers. The output mix produced by the proposed algorithm has an outstanding performance in the final subjective evaluation when compared against a raw mix, two semi-professional mixes and a previous automatic compression approach. The results showed that the algorithm is able to compete with or outperform the semi-professional mixes in terms of four different perceptual criteria: the appropriateness of the amount of DRC applied, the degree of imperfection, the ability to stabilise the erratic level fluctuations and overall preference. Additionally, we described a demonstration system that has shown personalized dynamic range control can easily be achieved in a web browser (using Web Audio API), responding to the environment around the listener. Demonstrations to listeners showed that the processing was unobtrusive and very effective at adapting to changes in environment noise.

Chapter 5 contributed to the field of auditory masking. We proposed several masking metrics for quantifying masking behaviour within the multitrack mixture, adapting the cross-adaptive digital audio effect architecture (Reiss, 2011) and expanding existing psychoacoustics models of Glasberg and Moore (Glasberg & Moore, 2002; Moore et al., 1997) and MPEG audio coding (Bosi et al., 1997; ISO, 1993). First, an equal loudness matching experiment using the method of adjustment (Glasberg & Moore, 2005) was conducted to evaluate the performance of the proposed multitrack loudness model on musical signals against human perception. We

found that the model over-estimated the partial masking occurring in the multitrack audio. We then analyzed the underlying features and proposed a modification of the *K* parameter in the implementation of the partial loudness model. Evaluation results showed that model with the proposed modification yields better perceptual compliance for musical signals. The outcomes of the experiment were then integrated into the development of the multitrack masking metrics, which offer a perceptual understanding of the mixing process. Evaluations of masking metrics were presented later in Chapter 6, where the metrics were integrated into an autonomous masking minimization system built upon a typical optimization framework.

In Chapter 6 we incorporated previous research outcomes in frequency manipulation (Chapter 3), dynamic processing (Chapter 4) and auditory masking (Chapter 5), into one intelligent multitrack masking minimization system. We first explored the relationship between the two essential signal-processing operations in mixing, equalization and dynamic processing. By investigating the control mechanisms and operating spaces of these two operations, we presented a general frequency and dynamic processing tool, capable of modifying the boost and/or cut of an equalization stage over time, following a dynamics curve. We then investigated how to employ different audio techniques (equalization, dynamic processing and proposed general processing) to manipulate the spectral and dynamic characteristics of the signals to perform masking reduction. We proposed an autonomous masking minimization system based on an optimization framework, where the aforementioned masking models (Chapter 5) were employed to describe the objective function. Various implementations of the system were explored and evaluated objectively and subjectively through a listening experiment. The results implied that our novel MPEG-based masking metric is able to predict the multitrack masking better than the more advanced psychoacoustic models based on Glasberg and Moore (Glasberg & Moore, 2002; Moore et al., 1997). And the general frequency and dynamics processing algorithm proved to be more efficient and powerful in masking reduction than using equalization or dynamic range compression alone.

Most of these research outcomes were represented in international peer-reviewed conference and journal articles, as listed in Section 1.6.

7.2 Future Directions

The concept of intelligent mixing is not new, but it is still relatively unexplored. Therefore there are numerous directions the future research could take. A compendium of possible improvements to the intelligent methods for frequency and dynamics processing, and relevant future research directions, are presented here.

As for our spectral characteristic studies of successful commercial recordings, additional analysis of the difference between the original version and re-mastered of the same recording is a fascinating direction to achieve a better understanding of the progression of modern mixing techniques, as well as the evaluation of music appreciation. Subjective evaluation of the intelligent equalization method should be conducted as future work in the form of a listening test to assess and validate whether the algorithm can improve the listening experience of the musical mixture by matching the spectral content of the audio signal to the common curve pattern of successful commercial recordings.

The parameter automations of the intelligent multitrack dynamic range compression algorithm can be improved by more sophisticated use of audio features to describe the spectral and dynamic characteristics of signal. Audio features proposed in the fields of instrument identification (Eronen, 2001) and genre recognition (Tzanetakis & Cook, 2002) are worth exploring. In general, the more we know about the input signals, the more assumptions we can make based on the best practices in audio engineering and perceptual criteria. As a result, the system is able to generate intelligent mixing choices that are closer to how professional engineers operate.

There are a few limitations to the proposed multitrack masking models. Metric I & II adapt the loudness and partial loudness models of Glasberg and Moore calculate a short-term spectrum to derive an excitation pattern via a bank of level-dependent overlapping filters. This approach might not accurately represent the way that excitation patterns are evoked in the human auditory system. In particular, the model does not take into account the fact that the auditory filters have a phase characteristic with significant curvature. Because of this curvature, harmonic complex sounds with identical power spectra can give rise to waveforms on the basilar membrane with very different peak factors (ratio of peak amplitude to RMS amplitude), depending on their phase spectra. This in turn may lead to differences in loudness. Furthermore, the proposed Masking Metric I & II only deploy a simple smoothing mechanism (which resembles the way that a control signal is generated in an automatic gain control circuit), using conditional filter coefficients based on whether the sound is in an

attack or release phase, to account for the temporal integration of loudness. This means that forward and backward masking may not be well quantified. Masking models based on the psychoacoustic model of MPEG audio coding (Metric III & IV) also have the similar limitation on capturing temporal masking. Therefore it would be beneficial to further investigate auditory masking models that are more applicable to musical signals, and account for temporal masking as well. Research on informational masking (Moore, 2012) offers another interesting research direction. However it lies closer to the area of music cognition. And such informational masking may be wanted in a mix, i.e., a saxophone, trumpet and trombone are intended to be heard as a 'horn section'.

Masking reduction was performed using frequency and dynamics processing. However, (Wakefield & Dewey, 2015) recently showed that stereo panning of sources is often a preferred masking reduction technique, when compared against frequency-based alternatives. Hence, more effective intelligent masking reduction might be achieved by incorporating spatial aspects into the masking metrics and incorporating panning into the multitrack processing tools.

These improvements mentioned above can be applied to enhance the performance of our final work on the autonomous minimization of masking multitrack audio. Additionally, since the proposed system only considers local optimization technique, an interesting extension would be to explore other global optimization algorithms such as pattern/direct search or genetic algorithms.

A semantic approach (Reiss & De Man, 2013) to autonomous mixing offers another interesting future research direction. High-level semantic knowledge can be used to inform the mixing decisions. Applying machine learning techniques to intelligent mixing is also promising, as shown in (Pardo et al., 2012; J. Scott et al., 2011; J. J. Scott & Kim, 2011). However, this approach is currently limited due to the rarity of available multitrack and mixing settings as training data.

As a final conclusion, the author believes that research on intelligent mixing has the potential to result in fascinating applications that can change the way we record, produce, and reproduce music.

Appendix A

Appendix: BBC Web-Based Compression

8.1 Web-Based Personalized Compression

Research presented in this section was performed during an internship at BBC. It has close ties with our research on the intelligent multitrack dynamic processing but approaches it from a different angle. The parameter automation described in previous Section 4.4, informs this work on a web-based personalized compression that adapts the dynamic range of the audio being played according to the environmental noise around the listener. This Section also indirectly addresses the problem of masking, since it is concerned with how to process the broadcast signal (maskee) in the presence of background noise (masker).

8.1.1 Introduction

Dynamic range compression (DRC) has been employed to solve the problem of loudness variability, which affects audibility, intelligibility, comfort and overall satisfaction with the programme material and its delivery in broadcast audio for decades (Skovenborg & Lund, 2009). Digital audio broadcasting (DAB) and digital television (DTV) both attempt to solve the problem by including a dynamic range control mechanism at the receiver side (Hoeg, IRT, & Jünger, 1994). However, particularly in DAB, not all receivers support the technique. There is no standard for calculating the compression control data, and the systems are under-used.

An unconventional approach is to hand over the control the dynamic range of the sound to the listener. This approach gains increasing support amongst broadcasters since it the listener who knows what, when and where they are listening, and what is listening environment around them.

Personalized compression that adapts the dynamic range of the audio being played according to the environmental noise around the listener is a relatively new field. However, it has its roots in automatic DRC research. Previous research on automatic DRC has been discussed in Section 2.5.4. However, none of these automation approaches were "environment aware". That is, automation of parameters was made solely based on the audio to be compressed, independent of listening level and independent of any additional sounds in the environment.

This section describes personalised compression algorithm that adapts the dynamic range of the audio being played according to the environmental noise around the listener, and offers simple control of the process to the listener. Environmental noise is picked up by the microphone in or attached to the phone, tablet, laptop, or PC being used, and a graphical user interface provides information and control. The web audio API is used as the basis of a player implemented in a web browser. Internet delivery of content allows much easier experimentation, and potentially quicker and cheaper deployment of this type of adaptation. The web audio API allows deployment of new techniques for audio processing without requiring software installation, and with independence of the platform being used.

8.1.2 Automatic Dynamic Range Compression

The web audio API (HTML5) provides a dynamic range compression node with threshold, ratio, knee width, attack time, and release time controls (Smus, 2013). We only apply downward compression in this application.

The parameter values of knee, attack and release time of the compressor were set to optimal values that have been defined by informal listening, as follows:

- knee width = 15 dB - a soft knee allowing smooth transition at the threshold level
- attack time = 8 ms - short attack time to catch the transients in the audio signal
- release time = 80 ms - moderate release time to give a smooth compression recovery

The threshold and ratio parameters are continuously adjusted to adapt to changing programmes and environments.

Compressor Threshold Automation

A simple RMS calculation is performed on blocks of 4096 samples, at a sampling rate of 48kHz, on a mono down-mix (L+R)/2 of the audio signal, as shown in Equation :

$$x_{rms} = \sqrt{\frac{\sum_{T} \left(\frac{x_l(t) + x_r(t)}{2}\right)^2}{4096}}, \tag{0.0}$$

where $x_l(t)$ and $x_r(t)$ are the left and right channel sample values at time t, respectively.

Due to the short block-based processing in the algorithm, an efficient and reliable long-term averaging process is needed to produce smoothly varying data, removing rapid changes that would lead to artefacts being introduced. An EMA filter is used to smooth the x_{rms} values, with a time-constant of approximately 0.8 s.

Listening in an environment with a high level of environmental noise requires more compression to make the quieter parts of the audio audible whilst not making the louder parts too loud. When the environmental noise level is very low less compression is needed, and therefore listeners can enjoy a wider dynamic range.

This implies that the compressor threshold should be lower than the RMS value when environment noise is high, and vice versa. Based on this, the threshold value is adapted by weighting the RMS of the audio signal with a value that is a function of the environment noise level. The threshold weighting factor, c_T is an altered Gaussian function of the environment noise level, as shown in Equation .

$$\begin{aligned} c_T &= e^{-\frac{(\frac{L_{K(E)}(t)}{60} - b)^2}{2c^2}} && \text{for } L_{K(E)}(t) \le 60dB(SPL) \\ c_T &= -e^{-\frac{(\frac{L_{K(E)}(t)}{60} - b)^2}{2c^2}} && \text{for } L_{K(E)}(t) > 60dB(SPL) \end{aligned} \quad . \tag{0.0}$$

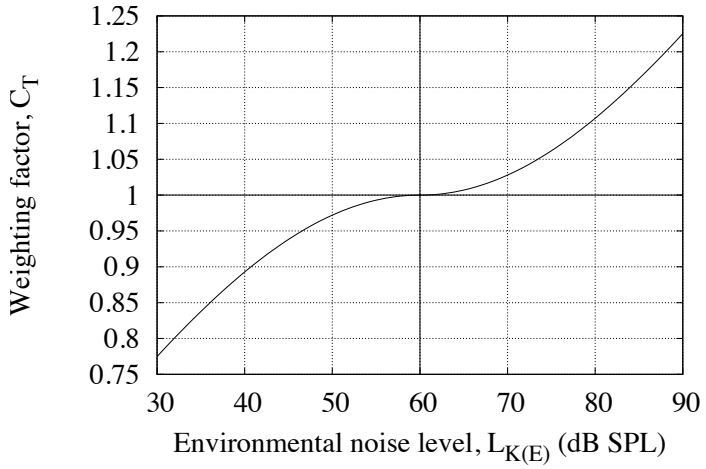


Figure 8.1 Weighting function applied to compressor threshold.

The ideal shape of the function was determined by informal listening and is shown in Figure 8.1, where *b* has a value of 1 and *c* a value of 0.7.

$$T(t) = x_{rms(dB)} c_T, \tag{0.0}$$

The threshold is weighted as shown in Equation .

$$T(t) = x_{rms(dB)} c_T, \tag{0.0}$$

where $x_{rms(\mathrm{dB})}$ is the RMS value from Equation converted to dBFS.

The result of applying this weighting is that the threshold is slightly lower than the RMS when the environment noise is higher than 60 dB (SPL), and slightly larger when it is less than 60 dB (SPL). When, for example, the RMS value of the audio signal is -25 dBFS, the threshold varies as a function of environment noise level as shown in Figure 8.2.

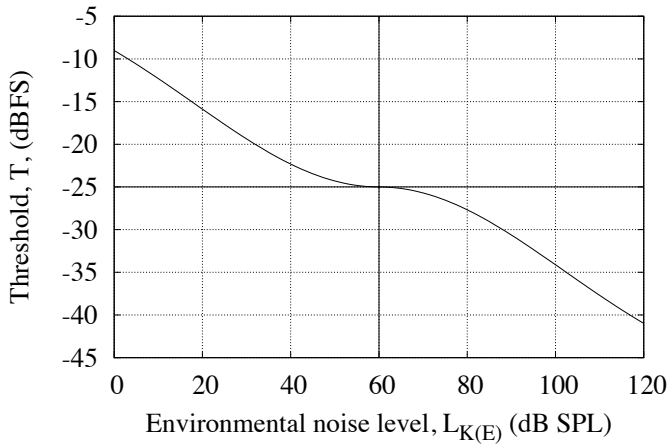


Figure 8.2 Compressor threshold as a function of environment noise level.

In the real world, the environment noise typically ranges from 30 dB (SPL) to 90 dB (SPL), being representative of a quiet room and traffic on a busy road. As shown in Figure 8.2, within that range, the threshold value is set close to the RMS. The purpose of the Gaussian curve is so that the threshold varies slowly around the RMS value within the anticipated environmental noise level range.

When compression is being applied with a time-varying threshold, intensive variation of the threshold in a short time causes audible artefacts, so another EMA smoothing, with α=0.95, is used prior to the actual setting of the compressor threshold. This is done every 3 ms, so the time constant is approximately 60ms.

Compressor Ratio Automation

The adaptation of the ratio is similar to that of the threshold, but based only on the environment noise level. In general, higher environment noise level demands a higher ratio. No compression is applied when the noise level is less than 30 dB (SPL), and the compression increases monotonically, but nonlinearly, in a way that matches human perception of the compression effect. Here, the ratio, *R*, is calculated as shown in Equation ,

$$R = c(L_{K(E)} - 30)^2 + 1, \qquad (0.0)$$

where c has been chosen through informal listening experiments to be 0.003265. This gives the curve shown in Figure 8.3.

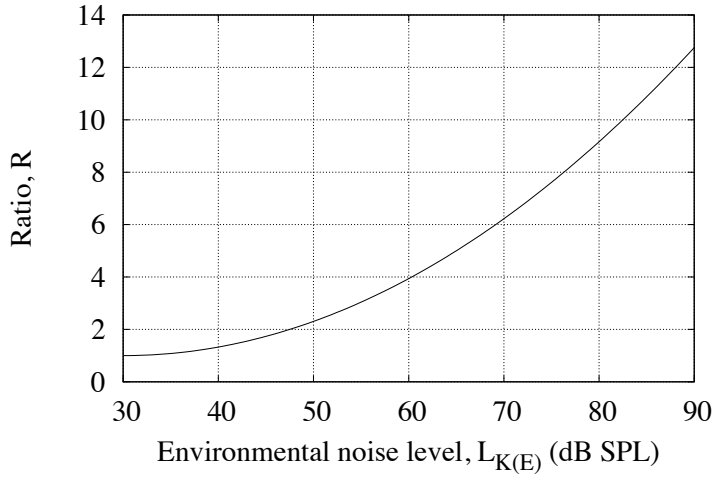


Figure 8.3 Compressor ratio as a function of environmental noise level.

As with the other parameters, an EMA filter, with α=0.95 (a time-constant of 60ms), is used prior to setting the compressor ratio. Although very large values of ratio would not normally be used, no limit is applied. Above a ratio of 10:1, the effect is that of a limiter, and, furthermore, the physiological effects of dangerously high environmental noise levels might become a problem before one needs to worry about the audible effects of extremely large values of ratio.

8.1.3 Automatic Volume Control

In addition to the automatic dynamics control, an automatic volume adjustment is also applied to the audio signal. The system measures the loudness of the audio signal and of the environmental noise and adjusts the gain applied to the audio signal to maintain a 6 LU (loudness unit) signal-to-noise ratio. The rationale is that when the noise level is high and applying dynamic range compression is no enough to compensate the noise, an overall gain adjustment is needed to further improve the listening experience.

The loudness $L_{K(E)}$ of the environment noise, and $L_{K(P')}$ of the audio signal (after compression) are measured according to the Recommendation ITU-R BS.1770 (ITU, 2012a)

with a 3 second integration time, as for a "short term" measurement according to Recommendation ITU-R BS.1771-1 (ITU, 2012b).

The algorithm makes the measurements every block of 4096 samples (at a sampling rate of 48kHz). The gain being applied is updated every 3ms to adapt to changes in measured values. The changes in gain are smoothed using an EMA filter to avoid jumps in response, but not make the adaptation to slow:

$$L_K'[n] = (1-\alpha)L_K[n] + \alpha L_K'[n-1], \tag{0.0}$$

where $L_K[n]$ is the most recent loudness value of the $L_{K(E)}$ or $L_{K(P')}$, and $L_K'[n]$, $L_K'[n-1]$ are the new smoothed value and the previous smoothed value, respectively. The value of the smoothing factor α is set to 0.998 and to 0.9 for gain increases and decreases, respectively, with corresponding time constants of 1.5 s and 20 ms.

The maximum gain applied is limited to 10 dB, in order to minimise the risk of damaging the hearing of the listener when the environmental noise is very loud.

The implementation relies on microphone calibration using a white noise source at 65 dBA. It is anticipated that this explicit requirement will be engineered out of the system, either by finding reasonable assumptions, or by learning from the listener's use of any controls provided.

To adapt the automatic gain control further the listener may indicate that they are using a particular style of headphone, with a corresponding typical attenuation of environmental noise. Measurements made on a small selection of headphones suggest that attenuation of 10 dB might be expected for circum-aural closed-back headphones, about 8 dB for supra-aural closed-back ones, and less than 1 dB for open-backed ones. Again, manual intervention by the listener might be engineered out, for example in future generation of devices, which potentially will automatically detect the type of headphone being used.

8.1.4 Evaluation

Feedback from listeners in an informal listening test conducted in the lab using open-backed headphones, a set of test programme material, and environmental noise from the BBC sound effects library played back over loudspeakers, suggested that the system was working quite well already: listeners reported that the system was doing very much what they wanted, and that its operation was unobtrusive.

The choice of operating parameters appeared to have been made well, and listeners sometimes did not realise just what the processing had been doing until it was turned off. A few comments about excessive compression being apparent on one of the items could be addressed by simple adjustment of the "More/Less" slider

8.1.5 Section Summary

This section has described a demonstration system of personalised dynamic range control in a browser, responding to the environment around the listener. Demonstrations to listeners showed that the processing was unobtrusive and very effective at adapting to changes in environment noise. This research project indirectly addressed the problem of masking. It leads into the following two chapters, which address multitrack masking reduction

Bibliography

Ahmad, M. O., & Wang, J. (1989). An analytical least square solution to the design problem of two-dimensional FIR filters with quadrantally symmetric or antisymmetric frequency response. *Circuits and Systems, IEEE Transactions on, 36*(7), 968-979.

Aichinger, P., Sontacchi, A., & Schneider-Stickler, B. (2011). *Describing the Transparency of Mixdowns: The Masked-to-Unmasked-Ratio.* Paper presented at the Audio Engineering Society Convention 130.

Algazi, V. R., Suk, M., & Rim, C.-S. (1986). Design of almost minimax FIR filters in one and two dimensions by WLS techniques. *Circuits and Systems, IEEE Transactions on, 33*(6), 590-596.

ANSI. (1994). American National Standard Acoustical Terminology. *ANSI SI*, 1-1994.

Barchiesi, D., & Reiss, J. (2010). Reverse engineering of a mix. *Journal of the Audio Engineering Society, 58*(7/8), 563-576.

Bocko, G., Bocko, M. F., Headlam, D., Lundberg, J., & Ren, G. (2010). *Automatic music production system employing probabilistic expert systems.* Paper presented at the Audio Engineering Society Convention 129.

Boley, J., Danner, C., & Lester, M. (2010). *Measuring Dynamics: Comparing and contrasting algorithms for the computation of dynamic range.* Paper presented at the Audio Engineering Society Convention 129.

Bosi, M., Brandenburg, K., Quackenbush, S., Fielder, L., Akagiri, K., Fuchs, H., & Dietz, M. (1997). ISO/IEC MPEG-2 advanced audio coding. *Journal of the Audio Engineering Society, 45*(10), 789-814.

Brecht De Man, B. L., King, R., & Reiss, J. D. (2014). *An analysis and evaluation of audio features for multitrack music mixtures*. Paper presented at the 15th International Society for Music Information Retrieval Conference, Taipei.

Chalupper, J., & Fastl, H. (2002). Dynamic loudness model (DLM) for normal and hearing-impaired listeners. *Acta Acustica united with Acustica, 88*(3), 378-386.

Chen, Z., Hu, G., Glasberg, B. R., & Moore, B. C. (2011). A new method of calculating auditory excitation patterns and loudness for steady sounds. *Hearing research, 282*(1), 204-215.

Dau, T., Kollmeier, B., & Kohlrausch, A. (1997). Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. *The Journal of the Acoustical Society of America, 102*(5), 2892-2905.

Dau, T., Püschel, D., & Kohlrausch, A. (1996). A quantitative model of the "effective"signal processing in the auditory system. I. Model structure. *The Journal of the Acoustical Society of America, 99*(6), 3615-3622.

De Boer, E. (1975). Synthetic whole - nerve action potentials for the cat. *The Journal of the Acoustical Society of America, 58*(5), 1030-1045.

De Man, B., Mora-Mcginity, M., Fazekas, G., & Reiss, J. D. (2014). *The Open Multitrack Testbed.* Paper presented at the Audio Engineering Society Convention 137.

Dennis Jr, J. E., & Schnabel, R. B. (1996). *Numerical methods for unconstrained optimization and nonlinear equations* (Vol. 16): Siam.

EBU–Recommendation. (2011). Loudness normalisation and permitted maximum level of audio signals.

Emmett, J., & Emmett, J. (2003). Audio levels-in the new world of digital systems. *EBU Technical Review, 2003*.
Eronen, A. (2001). *Comparison of features for musical instrument recognition.* Paper presented at the Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the.
Filanovsky, I., & Baltes, H. (1994). CMOS Schmitt trigger design. *Circuits and Systems I: Fundamental Theory and Applications, IEEE Transactions on, 41*(1), 46-49.
Fletcher, H. (1940). Auditory patterns. *Reviews of modern physics, 12*(1), 47.
Fletcher, H., & Munson, W. A. (1933). Loudness, Its Definition, Measurement and Calculation*. *Bell System Technical Journal, 12*(4), 377-430.
Foudi, N. (2012). *A Preliminary Framework For Automatic Cross-Adaptive Compression.* (Master), University of Massachusetts Lowell.
Friedlander, B., & Porat, B. (1984). The modified Yule-Walker method of ARMA spectral estimation. *Aerospace and Electronic Systems, IEEE Transactions on*(2), 158-173.
Gersho, A. (1994). Advances in speech and audio compression. *Proceedings of the IEEE, 82*(6), 900-918.
Giannoulis, D., Massberg, M., & Reiss, J. D. (2012a). Digital dynamic range compressor design—A tutorial and analysis. *Journal of the Audio Engineering Society, 60*(6), 399-408.
Giannoulis, D., Massberg, M., & Reiss, J. D. (2012b). Parameter Automation in a Dynamic Range Compressor. *Journal of the Audio Engineering Society*.
Gill, P. E., & Murray, W. (1974). *Numerical methods for constrained optimization* (Vol. 1): Academic Press London.
Glasberg, B. R., & Moore, B. C. (2002). A model of loudness applicable to time-varying sounds. *Journal of the Audio Engineering Society, 50*(5), 331-342.
Glasberg, B. R., & Moore, B. C. (2005). Development and evaluation of a model for predicting the audibility of time-varying sounds in the presence of background sounds. *Journal of the Audio Engineering Society, 53*(10), 906-918.
Godsill, S., Rayner, P., & Cappé, O. (2002). *Digital audio restoration*: Springer.
Goldstein, E. (2013). *Sensation and perception*: Cengage Learning.
Hafezi, S., & Reiss, J. D. (2015). Autonomous Multitrack Equalization Based on Masking Reduction. *Journal of the Audio Engineering Society, 63*(5), 312-323.
Hoeg, W., IRT, H. T., & Jünger, H. (1994). Dynamic Range Control (DRC) and music/speech control (MSC). *EBU Technical Review*, 56-70.
Howard, D. M., & Angus, J. (2009). *Acoustics and psychoacoustics*: Taylor & Francis.
Huber, D. M., & Runstein, R. E. (2013). *Modern recording techniques*: CRC Press.
Irino, T., & Patterson, R. D. (2001). A compressive gammachirp auditory filter for both physiological and psychophysical data. *The Journal of the Acoustical Society of America, 109*(5), 2008-2022.
ISO. (1993). IEC 11172-3 Information technology-coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s-Part3: Audio *Motion Picture Experts Group*.
ITU. (2003). ITU Recommendation BS.1534 Method for the subjective assessment of intermediate quality levels of coding systems.
ITU. (2012a). ITU-R Recommendation BS.1770-3, Algorithms to measure audio programme loudness and true-peak audio level.
ITU. (2012b). ITU-R Recommendation BS.1771-1, Requirements for loudness and true-peak indicating meters.
Izhaki, R. (2013). *Mixing audio: concepts, practices and tools*: Focal Press.
Jepsen, M. L., Ewert, S. D., & Dau, T. (2008). A computational model of human auditory signal processing and perception. *The Journal of the Acoustical Society of America, 124*(1), 422-438.

Johnston, J. D. (1988a). *Estimation of perceptual entropy using noise masking criteria.* Paper presented at the IEEE International Conference on Acoustics, Speech and Signal Processing.
Johnston, J. D. (1988b). Transform coding of audio signals using perceptual noise criteria. *Selected Areas in Communications, IEEE Journal on, 6*(2), 314-323.
Karjalainen, M. (1985). *A new auditory model for the evaluation of sound quality of audio systems.* Paper presented at the IEEE International Conference on Acoustics, Speech and Signal Processing.
Katz, B. (2007). *Mastering Audio: The Art and the Science. 2002*: Burlington, MA: Focal Press.
Kleczkowski, A., & Kleczkowski, P. (2006). *Advanced methods for shaping time-frequency areas for the selective mixing of sounds.* Paper presented at the Audio Engineering Society Convention 120.
Kleczkowski, P. (2005). *Selective Mixing of Sounds.* Paper presented at the Audio Engineering Society Convention 119.
Kobayashi, T., & Imai, S. (1990). Design of IIR digital filters with arbitrary log magnitude function by WLS techniques. *Acoustics, Speech and Signal Processing, IEEE Transactions on, 38*(2), 247-252.
Kolasinski, B. (2008). *A framework for automatic mixing using timbral similarity measures and genetic optimization.* Paper presented at the Audio Engineering Society Convention 124.
Kraght, P. H. (2000). Aliasing in digital clippers and compressors. *Journal of the Audio Engineering Society, 48*(11), 1060-1065.
Krippendorff, K. (2012). *Content analysis: An introduction to its methodology*: Sage.
Lartillot, O., Toiviainen, P., & Eerola, T. (2008). A matlab toolbox for music information retrieval *Data analysis, machine learning and applications* (pp. 261-268): Springer.
Lattin, J. M., Carroll, J. D., & Green, P. E. (2003). *Analyzing multivariate data*: Thomson Brooks/Cole Pacific Grove, CA, USA.
Lee, R. (2008). *Simple Arbitary IIRs.* Paper presented at the Audio Engineering Society Convention 125.
Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association, 62*(318), 399-402.
Lim, Y. C., Lee, J.-H., Chen, C.-K., & Yang, R.-H. (1992). A weighted least squares algorithm for quasi-equiripple FIR and IIR digital filter design. *Signal Processing, IEEE Transactions on, 40*(3), 551-558.
Lindemann, E. (1997). *The continuous frequency dynamic range compressor.* Paper presented at the Applications of Signal Processing to Audio and Acoustics, 1997. 1997 IEEE ASSP Workshop on.
Lopez-Poveda, E. A., & Meddis, R. (2001). A human nonlinear cochlear filterbank. *The Journal of the Acoustical Society of America, 110*(6), 3107-3118.
Lund, T. (2005). Realtime Loudness Control for Broadcast. *TC Electronic A/S, Denmark.*
Ma, Z. a. D. M., Brecht and Pestana, Pedro D. L. and Black, Dawn A. A. and Reiss, Joshua D. (2015). Intelligent Multitrack Dynamic Range Compression. *J. Audio Eng. Soc, 63*(6), 412--426.
Maddams, J. A., Finn, S., & Reiss, J. D. (2012). *An Autonomous Method for Multi-track Dynamic Range Compression.* Paper presented at the Proceedings of the 15th Int. Conference on Digital Audio Effects (DAFx-12).
Mansbridge, S., Finn, S., & Reiss, J. D. (2012). *Implementation and Evaluation of Autonomous Multi-Track Fader Control.* Paper presented at the Audio Engineering Society Convention 132.
Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial & Applied Mathematics, 11*(2), 431-441.
MathWorks. (2015). Recursive digital filter design - MATLAB yulewalk. from http://uk.mathworks.com/help/signal/ref/yulewalk.html

McNally, G. W. (1984). Dynamic range control of digital audio signals. *Journal of the Audio Engineering Society, 32*(5), 316-327.
Moore, B. C. (2012). *An introduction to the psychology of hearing*: Brill.
Moore, B. C., & Glasberg, B. R. (1996). A revision of Zwicker's loudness model. *Acta Acustica united with Acustica, 82*(2), 335-345.
Moore, B. C., & Glasberg, B. R. (2007). Modeling binaural loudness. *The Journal of the Acoustical Society of America, 121*(3), 1604-1612.
Moore, B. C., Glasberg, B. R., & Baer, T. (1997). A model for the prediction of thresholds, loudness, and partial loudness. *Journal of the Audio Engineering Society, 45*(4), 224-240.
Moore, B. C., Vickers, D. A., Baer, T., & Launer, S. (1999). Factors affecting the loudness of modulated sounds. *The Journal of the Acoustical Society of America, 105*, 2757.
Moorer, J. A. (2000). Audio in the new millennium. *Journal of the Audio Engineering Society, 48*(5), 490-498.
Nielsen, S. H., & Skovenborg, E. (2004). *Evaluation of different loudness models with music and speech material.* Paper presented at the Audio Engineering Society Convention 117.
Pardo, B., Little, D., & Gergle, D. (2012). *Building a personalized audio equalizer interface with transfer learning and active learning.* Paper presented at the Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies.
Parker, J., & Valimaki, V. (2013). Linear dynamic range reduction of musical audio using an allpass filter chain. *Signal Processing Letters, IEEE, 20*(7), 669-672.
Peeters, G. (2004). {A large set of audio features for sound description (similarity and classification) in the CUIDADO project}.
Pei, S.-C., & Shyu, J.-J. (1994). Design of arbitrary FIR log filters by weighted least squares technique. *Signal Processing, IEEE Transactions on, 42*(9), 2495-2499.
Perez-Gonzalez, E., & Reiss, J. (2009). *Automatic gain and fader control for live mixing.* Paper presented at the Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA'09. IEEE Workshop on.
Pestana, P. D. (2013). *Automatic Mixing Systems Using Adaptive Audio Effects.* (PhD.), Universidade Catolica Portuguesa.
Pestana, P. D., & Reiss, J. (2014). *Intelligent Audio Production Strategies Informed by Best Practices.* Paper presented at the Audio Engineering Society Conference: 53rd International Conference: Semantic Audio.
Pestana, P. D., Reiss, J. D., & Barbosa, A. (2013). *Loudness Measurement of Multitrack Audio Content Using Modifications of ITU-R BS. 1770.* Paper presented at the Audio Engineering Society Convention 134.
Plack, C. J., & Moore, B. C. (1990). Temporal window shape as a function of frequency and level. *The Journal of the Acoustical Society of America, 87*(5), 2178-2187.
Plack, C. J., Oxenham, A. J., & Drga, V. (2002). Linear and nonlinear processes in temporal masking. *Acta Acustica united with Acustica, 88*(3), 348-358.
Pujol, J. (2007). The solution of nonlinear inverse problems and the Levenberg-Marquardt method. *Geophysics, 72*(4), W1-W16.
Reed, D. (2000). *A perceptual assistant to do sound equalization.* Paper presented at the Proceedings of the 5th international conference on Intelligent user interfaces.
Reiss, J. D. (2011). *Intelligent systems for mixing multichannel audio.* Paper presented at the Digital Signal Processing (DSP), 2011 17th International Conference on.
Reiss, J. D., & De Man, B. (2013). A Semantic Approach To Autonomous Mixing.
Reiss, J. D., & McPherson, A. (2014). *Audio Effects: Theory, Implementation and Application*: CRC Press.
Rennies, J., Verhey, J. L., & Fastl, H. (2010). Comparison of loudness models for time-varying sounds. *Acta Acustica united with Acustica, 96*(2), 383-396.

Sabin, A. T., & Pardo, B. (2009). *A method for rapid personalization of audio equalization parameters.* Paper presented at the Proceedings of the 17th ACM international conference on Multimedia.

Sánchez, S. M. (2009). *Extending Automatic Audio Mixing with Dynamics and Equalization Effects.* (Master), Universitat Pompeu Fabra.

Schroeder, M. R., Atal, B. S., & Hall, J. (1979). Optimizing digital speech coders by exploiting masking properties of the human ear. *The Journal of the Acoustical Society of America, 66*(6), 1647-1652.

Scott, J., Prockup, M., Schmidt, E. M., & Kim, Y. E. (2011). *Automatic multi-track mixing using linear dynamical systems.* Paper presented at the Proceedings of the 8th Sound and Music Computing Conference, Padova, Italy.

Scott, J. J., & Kim, Y. E. (2011). *Analysis of Acoustic Features for Automated Multi-Track Mixing.* Paper presented at the ISMIR.

Simpson, A. J., Terrell, M. J., & Reiss, J. D. (2013). *A Practical Step-by-Step Guide to the Time-Varying Loudness Model of Moore, Glasberg, and Baer (1997; 2002).* Paper presented at the Audio Engineering Society Convention 134.

Skovenborg, E., & Lund, T. (2008). *Loudness descriptors to characterize programs and music tracks.* Paper presented at the Audio Engineering Society Convention 125.

Skovenborg, E., & Lund, T. (2009). *Loudness Descriptors to Characterize Wide Loudness-Range Material.* Paper presented at the Audio Engineering Society Convention 127.

Smus, B. (2013). *Web audio API*: " O'Reilly Media, Inc.".

Soulodre, G. A. (2004). *Evaluation of objective loudness meters.* Paper presented at the Audio Engineering Society Convention 116.

Sunder, S., & Ramachandran, V. (1994). Design of recursive differentiators with constant group-delay characteristics. *Signal processing, 39*(1), 79-88.

Terrell, M., Simpson, A., & Sandler, M. (2014). The Mathematics of Mixing. *Journal of the Audio Engineering Society, 62*(1/2), 4-13.

Terrell, M. J., & Reiss, J. D. (2009). Automatic monitor mixing for live musical performance. *Journal of the Audio Engineering Society, 57*(11), 927-936.

Thiagarajan, J. J., & Spanias, A. (2011). Analysis of the MPEG-1 Layer III (MP3) algorithm using MATLAB. *Synthesis Lectures on Algorithms and Software in Engineering, 3*(3), 1-129.

Thiede, T., Treurniet, W. C., Bitto, R., Schmidmer, C., Sporer, T., Beerends, J. G., & Colomes, C. (2000). PEAQ-The ITU standard for objective measurement of perceived audio quality. *Journal of the Audio Engineering Society, 48*(1/2), 3-29.

Thiele, N. (2005). Some Thoughts On The Dynamiics Of Reproduced Sound. *Journal of the Audio Engineering Society, 53*(1/2), 130-132.

Tsilfidis, A., Papadakos, C., & Mourjopoulos, J. (2009). *Hierarchical perceptual mixing.* Paper presented at the Audio Engineering Society Convention 126.

Tsingos, N. (2005). *Scalable perceptual mixing and filtering of audio signals using an augmented spectral representation.* Paper presented at the 8th International Conference on Digital Audio Effects (DAFx 2005).

Tsingos, N., Gallo, E., & Drettakis, G. (2004). *Perceptual audio rendering of complex virtual environments.* Paper presented at the ACM Transactions on Graphics (TOG).

Tyler, L. B. (1979). *An above threshold compressor with one control.* Paper presented at the Audio Engineering Society Convention 63.

Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *Speech and Audio Processing, IEEE Transactions on, 10*(5), 293-302.

Unoki, M., Irino, T., Glasberg, B., Moore, B. C., & Patterson, R. D. (2006). Comparison of the roex and gammachirp filters as representations of the auditory filter. *The Journal of the Acoustical Society of America, 120*(3), 1474-1492.

Vega, S., & Janer, J. (2010). *Quantifying Masking in Multi-track Recordings*. Paper presented at the Proceedings of SMC Conference.
Vickers, E. (2001). *Automatic long-term loudness and dynamics matching.* Paper presented at the Audio Engineering Society Convention 111.
Vickers, E. (2011). The loudness war: Do louder, hypercompressed recordings sell better? *Journal of the Audio Engineering Society, 59*(5), 346-351.
Wakefield, J., & Dewey, C. (2015). *An Investigation into the Efficacy of Methods Commonly Employed by Mix Engineers to Reduce Frequency Masking in the Mixing of Multitrack Musical Recordings.* Paper presented at the Audio Engineering Society Convention 138.
Walker, H. M. (1940). Degrees of freedom. *Journal of Educational Psychology, 31*(4), 253.
Ward, D., Reiss, J. D., & Athwal, C. (2012). *Multitrack Mixing Using a Model of Loudness and Partial Loudness.* Paper presented at the Audio Engineering Society Convention 133.
Wilmering, T., Fazekas, G., & Sandler, M. (2012). *High level semantic metadata for the control of multitrack adaptive audio effects.* Paper presented at the 133rd Convention of the AES, San Francisco, USA.
Wise, D. K. (2009). Concept, Design, and Implementation of a General Dynamic Parametric Equalizer. *Journal of the Audio Engineering Society, 57*(1/2), 16-28.
Zolzer, U. (2011). Adaptive digital audio effects. *DAFX: Digital Audio Effects*, 321.
Zwicker, E. (1958). Über psychologische und methodische Grundlagen der Lautheit. *Acta Acustica united with Acustica, 8*(Supplement 1), 237-258.
Zwicker, E. (1977). Procedure for calculating loudness of temporally variable sounds. *The Journal of the Acoustical Society of America, 62*(3), 675-682.
Zwicker, E., & Scharf, B. (1965). A model of loudness summation. *Psychological review, 72*(1), 3.