**Queen Mary University of London**

# Intelligent Subgrouping of Multitrack Audio

by

David Ronan

A thesis submitted in partial fulfilment for the
degree of Doctor of Philosophy

in the

Centre for Intelligent Sensing

November 2018

# Declaration of Authorship

I, David Ronan, 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 partys 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.



Signed: _____

Date: _____

*“Waka, waka, waka!”*.

Fela Kuti

# *Abstract*

Subgrouping facilitates the simultaneous manipulation of a number of audio tracks and is a central aspect of mix engineering. However, the decision process of subgrouping is a poorly documented technique. This research sheds light on this ubiquitous but poorly defined mix practice, provides rules and constraints on how it should be approached as well as demonstrates its benefit to an automatic mixing system.

I first explored the relationship that subgrouping has with perceived mix quality by examining a number of mix projects. This was in order to decipher the actual process of creating subgroups and to see if any of the decisions made were intrinsically linked to mix quality. I found mix quality to be related to the number of subgroups and type of subgroup processing used. This subsequently led me to interviewing distinguished professionals in the audio engineering field, with the intention of gaining a deeper understanding of the process. The outcome of these interviews and the previous analyses of mix projects allowed me to propose rules that could be used for real life mixing and automatic mixing. Some of the rules I established were used to research and develop a method for the automatic creation of subgroups using machine learning techniques.

I also investigated the relationship between music production quality and human emotion. This was to see if music production quality had an emotional effect on a particular type of listener. The results showed that the emotional impact of mixing only really mattered to those with critical listening skills. This result is important for automatic mixing systems in general, as it would imply that quality only really matters to a minority of people.

I concluded my research on subgrouping by conducting an experiment to see if subgrouping would benefit the perceived clarity and quality of a mix. The results of a subjective listening test showed this to be true.

# *Acknowledgements*

First and foremost I would like to thank my supervisors Joshua Reiss, Hatice Gunes and Andrea Cavallaro. Without their expert guidance and wisdom, there would be no PhD. I would also like to thank all of my family (Mike, Helene, Marc and Cathal) for their love and support through the years. I thank all the researchers in C4DM, MMV and CIS I've interacted with, had banter with, stressed with, hiked with, partied with and had many interesting discussions with. No point in me naming names, there are too many. You know who you are.

A special thank you to Native Instruments for taking me on for an internship for six months. I learned as much about MIR as I did about Techno during that time. Everyone at AI Music for doing my experiments, putting up with my writing up crankiness and Sia for keeping the midnight oil burning.

Collectively I'd like to thank all the Flaxman gang, the Schwiftyhos, everyone in the P*ss Posse, the Berlin crew and the lads from the other island for the many many laughs, hangovers and support through the years. I'm sure I'm forgetting other crews and groups of friends, but you're all awesome.

Oh yes, thanks to Macca and Kate for convincing me to do a PhD in the first place, thanks to Pet for the 80's music days, tea and procrastination. Finally, thanks to Nora for putting up with me during the good times and the bad times. No mean feat.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Due to the advancements in computer processing power, we can now produce studio quality music with very inexpensive software. An amateur producer can use their own personal computer to get started quite easily. These advancements have lowered the bar of entry into music production and have made it more cost efficient than approaching a high-end studio to make a recording.

To use a studio, a musician is required to rent a studio space, pay for a qualified sound engineer to make a recording and subsequently mix the audio. Getting from a musical performance in a studio to a finished product that we can listen to at home is a lengthy and involved process. Firstly, the performer needs to be recorded correctly. This involves the recording engineer making sure to get a clean and balanced recording for all the instrumentation. This also requires making sure all the recording equipment is setup and working correctly. This is so that all recordings are free of artefacts such as hum, clicks, distortion and broadband noise induced by improper recording.

Once the recording stage is complete, it is up to the engineer to make the recordings sound as professional as possible through the mixing and editing of the audio tracks. This stage is called post-production. This is where you need an engineer who is skilled and experienced at what they do in order to get good results. Usually the greater the skill of the engineer, the greater the cost to avail of their services. Due to the length of time this process takes; consequently, it uses up most of the musicians production budget and usually takes twice as long as any of the other processes [1]. As soon as the engineer has achieved a final mix that they are happy with, the mixed audio is then sent to a mastering engineer. They then prepare the audio recording, so that it can be transferred to the desired media for mass distribution.

Amateur producers normally assume all the previously described roles. This is because the expensive equipment that would normally be found in a studio has been developed into software that can be bought at a fraction of the cost. Also, many of the mixing and editing processes can be learned from books and online tutorials. This has empowered people to start making professional sounding music without the overhead normally associated with a studio.

In recent years a number of systems have been developed to automate many of the processes required to deliver a successful mix [2, 3]. Some of these systems are for dynamic range compression, panning and equalisation [4–6]. These systems allow amateur producers to create professional sounding recordings at a fraction of the cost of going to a studio and could someday make the recording engineer redundant. However, not all the decisions made by the recording engineer during the mixing and editing stage can be automated as some of the decisions made are for artistic reasons.

It is these automatic mixing systems that are central to most of the research conducted as part of this thesis. In my case the mix concept being explored is called subgrouping, which is a mix technique used for control and effect processing [7]. This concept is expanded on further in the next section.

## 1.1 Motivation

At the early stages of the mixing and editing process, the engineer will typically group instrument tracks into subgroups depending on what family of instruments they belong to. This means grouping guitar tracks with other guitar tracks or vocal tracks with other vocal tracks. This is done, so that the engineer can treat each subgroup of instruments separately [7]. For example, the engineer can compress just the drums without affecting anything else in the mix or change the overall level of the drums without having to change the level for each individual drum track. An example of what the subgrouping process looks like can be seen in Figure 1.1. Typically, the producer groups these instruments into subgroups based on rule of thumb [7]. As explained previously, this is done normally based on instrument type.

In the literature reviewed as part of this thesis, there was currently no system that attempted to automate the subgrouping process. Also, as part of this thesis a survey interview was conducted on how professionals subgroup, this will be discussed in detail in chapter 4. It showed that all the professionals interviewed use subgrouping when mixing. A number of themes were also developed from the survey as to why they do so.

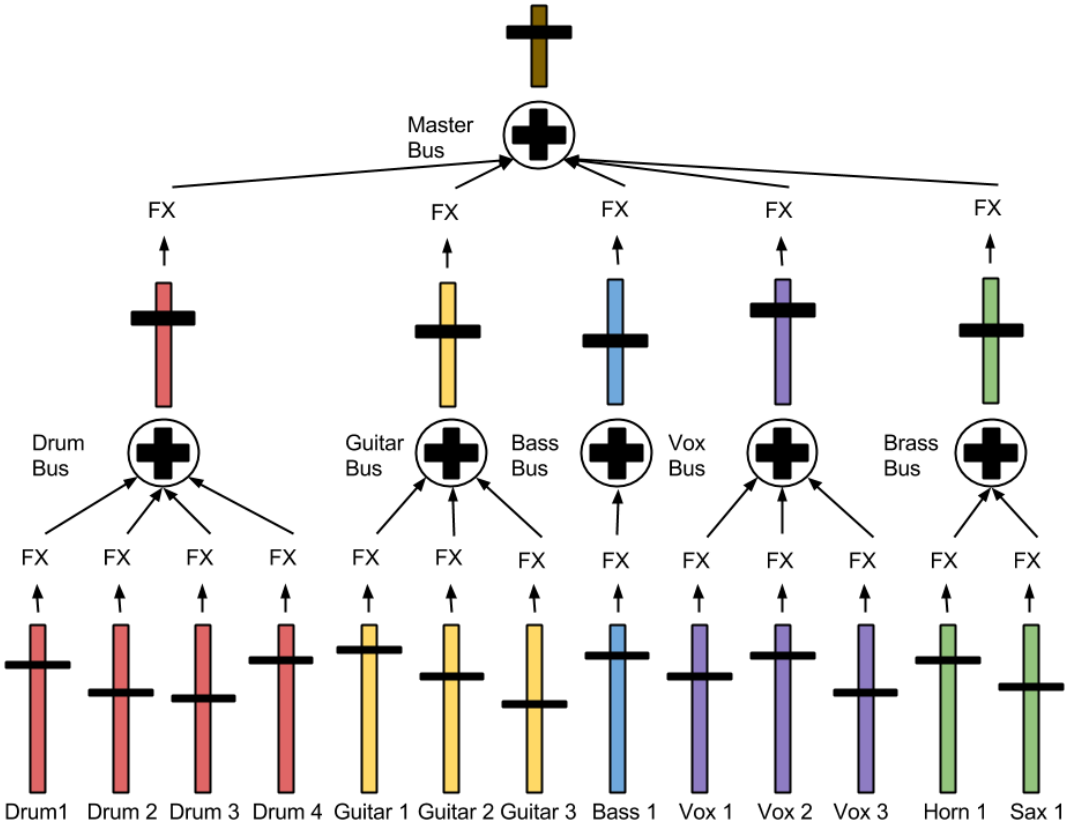


FIGURE 1.1: This is a typical subgrouping setup you might find in a studio. Each of the instrument types are summed together and processed as a group i.e. drums 1-4 are processed in a drum group

The results indicated that subgrouping is an important step in the mixing process and is something that is always done at a professional level.

If systems like these are designed to mimic the ability of a mix engineer to achieve a good mix, it could be argued that subgrouping is needed in current automatic mixing systems. In many of the papers looked at in the literature review based on automatic mixing, the instrument tracks were never subgrouped together and were treated individually [2, 8, 9]. Also, as part of this thesis, data that was collected from a mix experiment that showed a strong correlation between the number of subgroups used and mix quality. This experiment will be discussed in greater detail in chapter 3. The data also showed a correlation between subgroup processing and mix quality, specifically EQ and compression. Leading us further to believe that subgrouping is a necessary and overlooked mix process in the literature.

It is relatively easy to subgroup instrument tracks in the conventional sense. However, through the analysis of the spectro-temporal features of a number of multitracks I discovered that there are more intelligent ways to subgroup instrument tracks using state of the art machine learning techniques. An example output of this process, would be

that the more percussive instruments may be put in a subgroup together. In terms of musical instruments, the subgroup may consist of your traditional drum instruments and a bass guitar, but where the bass is played in a slapped style. Due to the subgroup now consisting of only percussively played instruments, this will have an effect on how the dynamic range compression for this subgroup will be applied and how the bass guitar would have normally been subgrouped.

Another possible outcome of the analysis of a multitrack is that it may be found over time that an instrument track may change and may become more similar to another instrument track in another subgroup. An example of how this may occur would be where the bass player suddenly switched from picking the bass guitar to playing in the style of slap bass. What was once subgrouped with the bass instruments could now be subgrouped with the percussive instruments. It may make sense at this point to split the single bass guitar instrument track into two individual tracks and have them designated to separate subgroups. How this could potentially be applied to the time series can be seen in Figure 1.2.

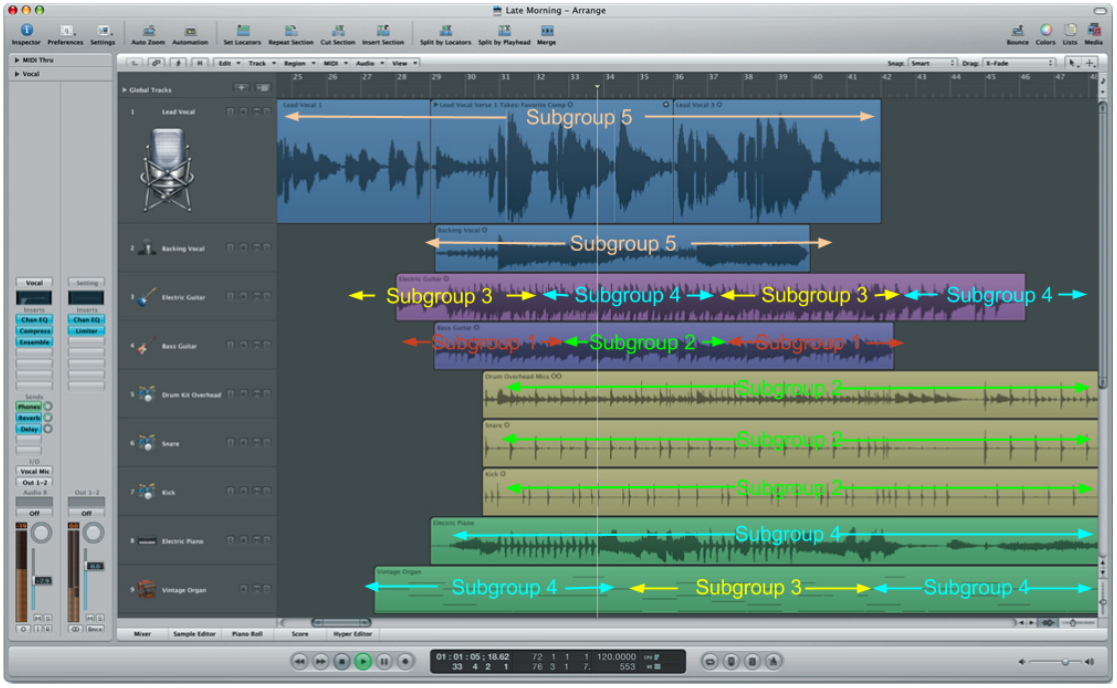


FIGURE 1.2: This is a screen shot from a DAW project. This illustrates how some instrument types might change their subgroup type over the course of a mix. In this example, the bass guitar changes between subgroups 3 and 4.

As mentioned previously, the subgrouping will affect how different audio effects will be applied to the subgroups such as dynamic range compression, panning and EQ [4–6]. These effects will have to adapt to the possibility of more spectrally diverse subgroups. The knock-on effect of this may be the way the overall balance of the mix may change and how the listener may perceive the emotional expression of the music. This might

be due to mix now having more "punchines" due to the amount of dynamic range compression being applied. The opposite and in most cases the undesired effect is that the mix becomes "flat" sounding and the listener may not feel as emotionally engaged by the music [7].

Another alternative subgrouping method could be that parts of the multitrack that are more melodic and harmonious may also get subgrouped together. Now the producer has some control of some of the more specifically emotive parts of the music and is free to process this subgroup differently [10].

## 1.2 Aims and Objectives

In light of the above, the main aims and objectives of this thesis were as follows:

- *Investigate why and how subgrouping is performed when mixing*: This was to gain a deeper understanding of the mix process, by examining the mix habits of professional and amateur mix engineers. This involved examining mix setups, which is detailed in chapter 3, interviewing professionals mix engineers which is detailed in chapter 4 and further mining of related literature. This was to establish how important the subgrouping process was when it came to mixing audio. It was also to generate rules or guidelines that could be applied in an automatic mix system, since they are so poorly defined in the literature. It involved finding what subgrouping decisions were made to improve the mix and how much improvement subgrouping meant, if any at all.

- *Investigate how to automatically subgroup multitrack audio*: The purpose of this was to investigate methods and techniques for automatically generating the subgroups that humans would create, but by using machine learning techniques. This involved performing feature selection using a Random Forest and then finally using agglomerative clustering to create the subgroups. I was interested to see what kind of audio features were useful for this process and what subgroups would be created based on these learned features. This is detailed in chapter 5.

- *Investigate if music production quality has an emotional impact on the listener*: I performed exploratory research to see if music production quality could have an impact on the perceived and induced emotions of a listener. I did this by performing a listening test where 10 critical listeners and 10 non-critical listeners evaluated 10 songs. There were two mixes of each song, the low quality mix and the high quality mix. Each participants subjective experience was measured

directly through questionnaire and indirectly by examining peripheral physiological changes, change in facial expressions and the number of head nods and shakes they made as they listened to each mix. The details of the experiment are provided in chapter 6. This research is related to automatic mixing systems, since I believe that current systems are not able to generate mixes that are on par with a human most of the time. However, if I were able to prove that mix quality was only important to those with critical listening skills then it further justifies the use of automatic mixing systems.

- *Determine if subgrouping can be used to improve automatic mixing systems*: With this aim I conducted an experiment in order to provide empirical evidence that using subgroups in an automatic mixing system could improve mix quality, perceived clarity as well as reduce mix complexity. I did this by creating automatic mixes using subgroups and automatic mixes without. I then conducted a listening test were the participants had to indicate which mix type they preferred as well as indicate which mixes had less inter-channel auditory masking. The details of the experiment are in chapter 7. The main purpose of this experiment was to test the hypothesis that subgrouping can be beneficial for automatic mixing.

The overarching aim of this thesis was to document and understand how subgrouping is used day to day in a studio since the literature on this widely used technique is sparse. With this understanding in place, I then wanted to know if it was beneficial to use this mix technique in an automatic mixing system.

## 1.3 Contributions

- **Chapter 3**: I showed that the number of subgroups and subgroup effect processing is correlated with mix quality. I was also able to observe some common mix decision patterns, which I later used to infer mix decisions in chapter 4
- **Chapter 4**: I proposed a number of recommendations on how subgrouping should be implemented in an automatic mixing system.
- **Chapter 5**: I determined a set of low level audio features that could be used to automatically subgroup multitrack audio. I determined these audio features using a Random Forest classifier for feature selection.
- **Chapter 6**: I present findings that suggest that having a high level of skill in mix engineering only seems to matter in an emotional context to those with critical listening skills.

- **Chapter 7**: I showed that subgrouping is beneficial to an automatic mixing system, which was an important result for this thesis.

## 1.4 Associated Publications

**Conferences:**

D. Ronan, B. De Man, H.Gunes and J. D. Reiss, "The impact of subgrouping practices on the perception of multitrack music mixes", in Audio Engineering Society Convention 139, September 2015. This is associated with chapter 3, where I examine common mix decisions and show that the number of subgroups and subgroup processing is correlated with mix quality.

D. Ronan, D. Moffat, H. Gunes, and J. D. Reiss, "Automatic subgrouping of multitrack audio", in Proc. 18th International Conference on Digital Audio Effects, DAFx-15, November 2015. This is associated with chapter 5, where I determine a set of candidate low level audio features to be used to automatically subgroup audio.

D. Ronan, H. Gunes, and J. D. Reiss, "Analysis of the subgrouping practices of professional mix engineers", in Audio Engineering Society Convention 142, May 2017. This is associated with chapter 4, where I analyse the survey responses of professional mix engineers. I test nine assumptions related to subgrouping and propose a number of recommendations on how subgrouping should be conducted.

**Journals:**

D. Ronan, J. D. Reiss and H. Gunes, "An empirical approach to the relationship between emotion and music production quality", Journal of the Audio Engineering Society (Under Review). This is associated with chapter 6, where I explore the relationship between music production quality and human emotion. I present findings that show mix engineering skill only matters to those with critical listening skills.

D. Ronan, H. Gunes and J. D. Reiss, "Automatic Minimisation of Masking in Multitrack Audio using Subgroups", IEEE/ACM Transactions on Audio, Speech, and Language Processing (Under Review). This is associated with chapter 7, where I prove that subgrouping is beneficial to automatic mixing systems.

**Other Contributions:** These conference publications are not directly related to this thesis, but are relevant to the fields of music information retrieval and sound synthesis.

D. Moffat, D. Ronan and J. D. Reiss, "An evaluation of audio feature extraction toolboxes", in Proc. 18th International Conference on Digital Audio Effects (DAFx-15),

November 2015. (Honourable mention for the best paper award). This is not related to the thesis, however audio feature extraction is important to chapter 5.

D. Moffat, D. Ronan and J. D. Reiss, "Unsupervised Taxonomy of Sound Effects", in Proc. of the 20th International Conference on Digital Audio Effects (DAFx-17), September 2017. This is an extension to the work carried out in chapter 5, but is applied to sound effects.

## 1.5 Thesis Outline

The remainder of this thesis is organised as follows:

- **Chapter 2** presents the background upon which this thesis will be developed. I outline the mix process and where subgrouping belongs in this process. I look at the physiology of the human ear, as well as critical bands, auditory filters and auditory masking. I look at other automatic mixing systems in the context of the main audio effects that are being automated. I also give an overview of emotion in music, where I show what the difference between perceived and induced emotions is, what the different psychological models of emotion are and how emotional responses to music are measured. Finally, I detail how Random Forest classifiers work and how they can be used for feature selection.

- **Chapter 3** analyses the impact that subgrouping practices have on the perception of quality in a dataset of multitrack mixes. I also analysed the multitracks in order to see if any decision patterns emerged, which I later used to infer mix decisions in chapter 4.

- **Chapter 4** presents a study I performed where I interviewed ten award winning mix engineers through an online questionnaire, where I asked questions related to subgrouping of a qualitative and quantitative nature. This was done to build on the data presented in chapter 3. I was able propose a number of recommendations on how subgrouping should be implemented in an automatic mixing system and this study gave us a deeper understanding of the mix process.

- **Chapter 5** investigate methods and techniques to automatically generate the subgroups that a human would create, but by using machine learning. I determined a set of low level audio features that could be used to automatically subgroup multitrack audio. I determined these audio features using a Random Forest classifier for feature selection.

- **Chapter 6** investigates if music production quality has an emotional impact on the listener. The findings suggest that having a high level of skill in mix engineering only seems to matter in an emotional context to those with critical listening skills. This is important in the context of automatic mixing algorithms, in the sense that the perceived quality of an automatically generated mix may not be that important to those without critical listening skills. Suggesting that automatically generated mixes may be good enough for the general public.

- **Chapter 7** investigates whether or not using subgroups in an automatic mixing system can improve the overall perceivable quality of a mix. I also investigated if using subgroups can have an impact on the perceived emotional response of a listener. I showed that participants always preferred the automatic mix that utilised subgrouping.

- **Chapter 8** concludes the thesis. Research findings are discussed and the prospects for future research are considered.

# Chapter 2

# Background

I start by discussing subgrouping since it is central to this thesis. I then discuss the machine learning methods I used in this thesis, where I discuss Random Forests, how feature selection is performed and agglomerative clustering. I also give the background of emotion in music, where I discuss the different types of musical emotions, the different psychological measures of emotion and how they can be measured. Finally, I discuss the physiology of the human hearing system with an emphasis on the concepts of masking, critical bands and auditory filters. Several psychoacoustic-inspired loudness and masking models as the perceptual basis of my intelligent mixing studies are then reviewed. I finally provide a review of the state of the art in automatic mixing systems.

## 2.1 Subgrouping

As mentioned previously, at the early stages of the mixing and editing process of a multitrack mix, the mix engineer will typically group instrument tracks into subgroups [7]. An example of this would be grouping guitar tracks with other guitar tracks or vocal tracks with other vocal tracks. Subgrouping can speed up the mix workflow by allowing the mix engineer to manipulate a number of tracks at once, for example by changing the level of all drums with one fader movement, instead of changing the level of each drum track individually [7]. Note that this can also be achieved by a Voltage Controlled Amplifier (VCA) group - a concept similar to a subgroup where a specified set of faders are moved in unison by one 'master fader', without first summing each of these channels into one bus. However, subgrouping also allows for processing that cannot be achieved by manipulation of individual tracks. For instance, when nonlinear processing such as dynamic range compression or harmonic distortion is applied to a subgroup, the processor will affect the sum of the sources differently than when it would

be applied to every track individually. An example of a typical subgrouping setup can be seen in Figure 1.1.

Subgrouping historically comes from the days of two-, four- and eight track tape recorders, when analogue recording and mixing devices were limited by the amount of inputs. Mix engineers back then would have recorded and mixed six drums tracks separately. Once the mix engineer was happy with the drum submix, they would then bounce it to stereo thus allowing the remaining four tracks to be used for other instruments such as vocal, guitars etc. to be mixed with the drums. This sounds like a tedious and potentially unforgiving process in comparison to what is possible in today's modern recording and mixing equipment. Nowadays, it is possible to have hundreds of tracks processed and mixed at the same time. However, it is not uncommon for mix engineers these days to create submixes in order to conserve processing power [1, 7, 11].

Very little is known about how mix engineers choose to apply audio processing techniques to a mix. There have been few studies looking at this problem and none of them specifically looked at subgrouping [12–14]. Subgrouping was touched on briefly in [12] when the authors tested the assumption *"Gentle bus/mix compression helps blend things better"* and found this to be true, but it did not give much insight into how subgrouping is generally used. In [15], the authors explored the potential of a hierarchical approach to multitrack mixing using instrument class as a guide to processing techniques. However, providing a deeper understanding of subgrouping was not the aim of the paper. Subgrouping was also used in [16], but similarly to [15] this was only applied to drums and no other instrument types were explored. The technique of subgrouping is to the best of my knowledge a poorly documented mix technique in audio engineering literature [1, 7, 17].

Although subgrouping is not well documented, it is used extensively in all areas of audio engineering and production. This would imply that there are basic unwritten rules that are carried out when a mix engineer makes use of subgrouping. These rules can be as simple as putting similar instruments together in the one subgroup [15, 16]. By investigating these practices I hope to develop these rules and generate constraints that may someday be used in intelligent mixing systems such as those described in [2, 3, 8, 15, 18].

One approach that already exists to subgrouping, is to subgroup by frequency bands. This mixing approach was developed by a famous mixing engineer called Michael Brauer (`http://www.mbrauer.com/qna2.asp`). In this approach, there are four subgroups, one for bass, one for mid-range, one for treble and finally another for distortion. This approach does not consider the traditional instrument approach and may be worth investigating as an alternative method to automatically subgrouping. This could also

be utilised in spatialisation of audio tracks, whereby everything in the bass group stays in the centre, vocals are used as the fourth group and everything else is split into the other two groups. This could then allow us to pan everything automatically, so as to minimise auditory masking.

## 2.2 The Physiology of the Human Hearing System

There are three main parts that constitute the human auditory system: the outer ear, the middle ear, and the inner ear. The outer ear is the fleshy part of the ear that is visible on the sides of the human head. This is known as the auricle. The purpose of the auricle is sound collection and spectral shaping, so that we can localise sound. Once sound reaches the auricle, it travels down the auditory canal to the eardrum. This is where the middle ear begins. The middle ear is an air-filled central cavity that consists of the three smallest bones in the body: malleus, incus and stapes (known collectively as the ossicles) [19]. The ossicles transmit the vibrations picked up by the eardrum to the inner ear. The inner ear consists of the cochlea and vestibular system. The cochlea is responsible for taking sound pressure patterns and converting these to electrochemical pulses that are passed to the auditory nerve. Inside the cochlea we also have the basilar membrane, where different parts of it resonate with respect to frequency. The vestibular system is responsible for providing balance [19].

## 2.3 Auditory Masking

Masking is a perceptual property of the human auditory system that occurs whenever the presence of a strong audio signal makes the temporal or spectral neighbourhood of weaker audio signals imperceptible [20, 21]. Frequency masking may occur when two or more stimuli are simultaneously presented to the auditory system. The relative shapes of the masker's and maskee's magnitude spectra determine to what extent the presence of certain spectral energy will mask the presence of other spectral energy.

Temporal masking is the characteristic of the auditory system where sounds are hidden due to a masking signal occurring before (pre-masking) or after (post-masking) a masked signal. The effectiveness of temporal masking attenuates exponentially from the onset and offset of the masker [22].

A simplified explanation of masking phenomena is when a strong noise or tone masker creates an excitation of sufficient strength on the basilar membrane. An excitation pattern is a neural representation of the pattern of resonance on the basilar membrane,

caused by a given sound [23]. The area around the characteristic frequency (referred to as the frequency bandwidth of the "overlapping bandpass filter" created by the cochlea) of the masker's signal location effectively blocks the detection of weaker signals [22]. Examples of frequency and temporal masking are shown in Figure 2.1 and Figure 2.2 respectively.

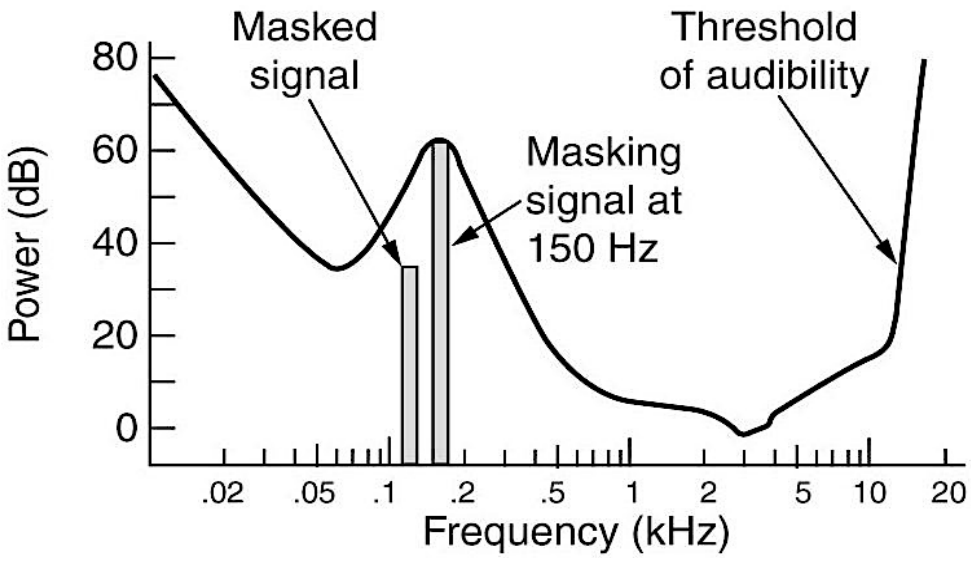


FIGURE 2.1: Frequency masking example of a 150 Hz tone signal masking an adjacent frequency tone by increasing the threshold of audibility around 150 Hz. pre-masking and simultaneous masking [22].

In the process of mixing, sound sources inevitably mask one another, which reduces the ability to fully hear and distinguish each sound source. Partial masking occurs whenever the audibility of a sound is degraded due to the presence of other content, but the sound may still be perceived. It is often partial masking that occurs within a mix. The mix can sound poorly produced or underwhelming, and have a lack of clarity as a result [24].

Masking reduction in a mix involves a trial and error adjustment of the relative levels, spatial positioning, frequency and dynamic characteristics of each of the individual audio tracks. In practice, the masking reduction process embodies an iterative search process similar to that of numerical optimisation theory [25, 26]. Masking reduction therefore can be thought of as an optimisation problem, which provides some insight to the methodology of automatic mixing in order to reduce masking. Given a certain set of controls for a multitrack, the final mix output can be thought of as the optimal solution to a system of equations that describe the masking relationship between the audio tracks in a multitrack recording.

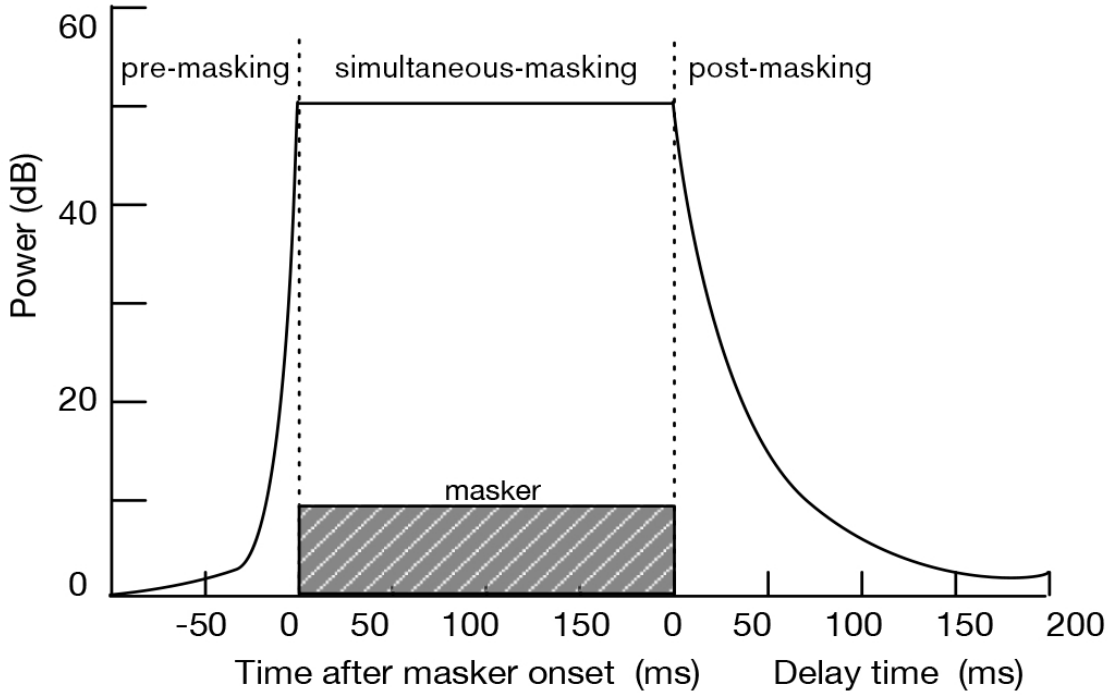


FIGURE 2.2: Schematic drawing to illustrate and characterise the regions within which pre-masking, simultaneous masking and post masking occur. Note that post-masking uses a different time origin than pre-masking and simultaneous masking [22].

Frequency processing, dynamics processing and subgrouping are the three main aspects of my masking minimisation investigation. Equalisation can effectively reduce masking by manipulating the spectral contour of different instruments so that there is less frequency domain interference between each audio track. Dynamic range processing is a nonlinear audio effect that can alter the dynamic contour of a signal in order to reduce masking over time. The classic operations of dynamics processing and equalisation control are two separate domains of an audio signal. The combined use of both filtering and dynamics processing implies a larger control space, and can reduce masking much more precisely and effectively in both frequency and time aspects than using either processor alone [7, 27]. Subgrouping allows us to localise the application of the frequency and dynamics processing to specific instrument types that would typically share similar timbre, dynamic range and spectral content.

The two principle aspects of automating a masking reduction process are the creation of a model of masking in multitrack audio that correlates well with human perception, and the development of audio techniques and algorithms to reduce masking without causing unpleasant audio artefacts.

### 2.3.1 Perceptual Models

Perceptual models capable of predicting masking behaviour have received much attention over the years, particularly in fields such as audio coding [28–32], where the masked threshold of a signal is approximated to inform a bit-allocation algorithm. [33] proposes a method for adjusting the masking threshold in audio coding to make the decoded signal robust to quantisation noise unmasking. Masking models are also often used in image and audio watermarking [34, 35]. Similar models are used in distortion measurement [36] and sound quality assessment [37–39], where nonlinear time-domain filter banks are used to allow for excitation pattern calculation whilst maintaining good temporal resolution. Another simple masking model is used in [40] to remove perceptually irrelevant time-frequency components. More advanced signal processing masking models that lie closer to the physiology of the human ear include a single-band model that accounts for a number of frequency and temporal masking experiments. A number of experiments were based on providing an internal Gaussian noise in order to model the nonlinear processing of the auditory system and to describe non-simultaneous masking [41]. In subsequent work, a 'modulation filter bank' was added to the previous model in order to analyse the temporal envelope at the output of a gammatone filter whose output is half-rectified and low pass filtered at 1kHz. This was to simulate the frequency to place transform across the basilar membrane, and receptor potentials of the inner hair cells [42]. Building upon the proposed 'modulation filter bank', a more complete masking model called the Computational Auditory Signal-Processing and Perception (CASP) model was presented that accounts for various aspects of masking and modulation detection. The experiments performed included intensity discrimination with pure tones and broadband noise, tone-in-noise detection, spectral masking with narrow-band signals and maskers, forward masking with tone signals and tone or noise maskers, and amplitude-modulation detection with narrow- and wideband noise carriers [43]. These account for various aspects of simultaneous and non-simultaneous masking in human listeners.

However, all mentioned models only output masked threshold as a measurement of masking, and only considered the situation when a signal (usually a test-tone signal) was fully masked. [44] explored partial loudness of mobile telephone ring tones in a variety of everyday background sounds e.g. traffic, based on the psychoacoustic loudness models proposed in [45, 46]. By comparing the excitation patterns (computed based on [45, 46]) between maskee and masker, [47] introduced a quantitative measure of masking in multitrack recording. Similarly, a Masked-to-Unmasked Ratio which related the original loudness of an instrument to its loudness in the mix was proposed in [48].

Previous attempts to perform masking reduction in audio mixing include [9, 18, 49, 50]. [49] aimed to achieve equal average perceptual loudness on all frequencies amongst all multitrack channels, based on the assumption that the individual tracks and overall mix should have equal loudness across frequency bands. However, this assumption may not be valid, and their approach does not directly address spectral masking. [18] designed a simplified measure of masking based on best practices in sound engineering and introduced an automatic multitrack equalisation system. However the simple masking measure in [18] might not correlate well with the perception of human hearing, as is evident in the evaluation. [50] applied a partial loudness model and [44] adjusts the levels of tracks within a multitrack in order to counteract masking. Similar techniques were investigated through an optimisation framework in [9]. However both [50] and [9] only performed basic level adjustment to tackle masking, which may have additional detrimental effects on the relative balance of sources in the mix [27].

### 2.3.2 Masking Metrics

There are a number of different multitrack masking metrics available that can be combined to perform a cross-analysis on multitracks. We can quantify the amount of masking by investigating the interaction between the excitation patterns of a maskee and a masker, where the maskee is an individual track and the masker is the combination of all the other tracks in a multitrack. This is done utilising the cross-adaptive architecture proposed in [2, 51]. All the masking metrics I discuss make use of this cross adaptive architecture. However, the first two masking metrics I will discuss are based on the perceptual loudness work of Moore [52, 53] and the final masking metric I discuss is based on spectral magnitude.

The procedure to derive loudness and partial loudness of each track in a multitrack is summarised as follows [50]. A multitrack consists of $N$ sources that have been pre-recorded onto $N$ tracks. Track $n$ therefore contains the audio signal from source $n$, given by $s_n$. The transformation of $s_n$ through the outer and middle ear to the inner ear (cochlea) is simulated by a fixed linear filter. A multi-resolution Short Time Fourier Transform (STFT), comprising 6 parallel FFTs, performs the spectral analysis of the input signal. Each spectral frame is filtered by a bank of level-dependent Roex filters whose centre frequencies range from 50Hz to 15kHz. A Roex filter is used to represent the magnitude response of the auditory filter found in the human ear [54]. Such auditory filtering represents the displacement distribution and tuning characteristics across the human basilar membrane.

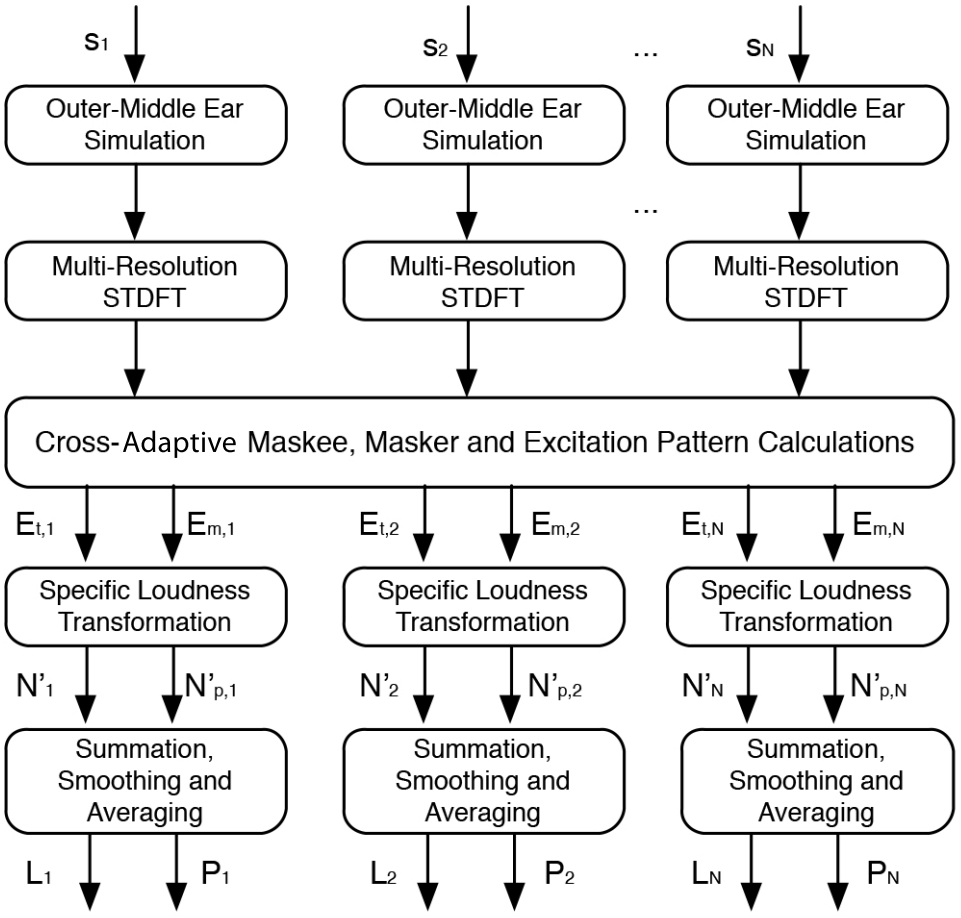


FIGURE 2.3: Flowchart of multitrack loudness model for $N$ input signals. This illustrates all the transformations applied to the audio and how each individual input signal is considered a maskee and the sum of all the other remaining tracks are the maskers.

The excitation pattern $E$ is calculated as the output of the auditory filters as a function of the centre frequency spaced at 0.25 equivalent rectangular bandwidth (ERB) intervals. ERB gives a measure of auditory filter width. The mapping between frequency, $f$ (Hz), and ERB (Hz) is shown in Equation 2.1.

$$\text{ERB} = 24.7(0.0437f + 1) \tag{2.1}$$

To account for masking, two excitation patterns, the target track (maskee) $E_{t,n}$ and the masker $E_{m,n}$, with respect to $s_n$ are calculated as described in [45, 46]. The masker here is the supplementary sum of the accompanying tracks related to the target track, as given by [48]

$$s'(n) = \sum_{i=1, i\neq 1}^{N} s_i \tag{2.2}$$

For a sound heard in isolation, the intensity represented in the excitation pattern is converted into specific loudness $N'_n$, which represents the loudness at the output of each auditory filter. In a partial masking scenario with concurrent masker $E_{m,n}$, partial specific loudness $N'_{p,n}$ is calculated. The detailed mathematical transformations to obtain specific and partial specific loudness can be found in [45].

The summation of $N'_n$, and $N'_{p,n}$ across the whole ERB scale produces the total unmasked and masked instantaneous loudness. All instantaneous loudness frames are smoothed to reflect the time-response of the auditory system, as described in [46], and then averaged into scalar perceptual loudness measures, loudness $L_n$ and partial loudness $P_n$. This is illustrated in Figure 2.3

Adapting the method of Vega et al [47], the masking measurement $M_n$ can be defined as the masker-to-signal ratio (MSR) based on an excitation pattern integrated across ERB scale and time. This is given by

$$M_n = \mathrm{MSR}(n) = 10 \log_{10} \frac{\sum_{\mathrm{ERB}} E_{m,n}}{\sum_{\mathrm{ERB}} E_{t,n}} \tag{2.3}$$

Wichern et al. [55] used a model based on loudness loss, $L_{loss}$, to measure masking. This can be defined as,

$$L_{loss} = L_{phon} - PL_{phon} \tag{2.4}$$

where $L_{phon}$ is the loudness of the maskee in isolation and $PL_{phon}$ is the partial loudness of the maskee when masked by the rest of the mix. The loudness unit here is phon as opposed to sones, which was used in Moore's original loudness model I discussed initially. The authors subsequently use a gating procedure to only measure masking when an instrument is actively playing.

In the work by Sina et al. [18], the authors do not use an auditory model to measure masking. They based their measurement on spectral magnitude. Where the amount of masking that track A (masker) at frequency $f$ and time $t$ causes on track B (maskee)

at the same frequency and time is given by

$$M_{A,B}(f,t) = \begin{cases} X_A(f,t)X_B(f,t) & \text{if} \\ & R_B(f,t) \leq R_T < R_A(f,t) \\ 0 & \text{else} \end{cases} \tag{2.5}$$

where $X_N(f,t)$ and $R_N(f,t)$ are respectively the magnitude in decibels and the rank of frequency $f$, at time $t$ for track $N$. $R_T$ is the maximum rank for a frequency region to be considered essential.

The work discussed here provided the framework and inspiration on how to reduce masking in the system proposed in Chapter 7. It was decided that by measuring how much masking occurs cross-adaptively as used in [18, 50] and using this as a basis for optimisation was a sensible approach. However, the work in Chapter 7 uses a different masking metric than the approaches discussed here and uses subgrouping. This was what made it a novel approach.

## 2.4 Automatic Mixing Systems

In recent years a number of systems have been developed to automate many of the processes required to deliver a successful mix [2, 3]. These systems empower amateur producers to create professional sounding recordings at a fraction of the cost of going to a studio and in a sense make a professional recording engineer's skill somewhat redundant. These systems can give an amateur producer a good starting point when it comes to mixing, however they may never be able to provide the level of polish a professional can. In this section I will explore some of the existing work published around automatic mixing systems. These systems are essentially an extension of the research area of adaptive digital audio effects [56].

### 2.4.1 Level

There have been a number of systems proposed where the parameters being adjusted to achieve a desirable mix are the individual levels of each instrument track in a multitrack. This is not something I looked at automating directly, but is important to my proposed automatic mixing framework in Chapter 7.

In [49], the authors developed a real-time cross-adaptive mixing system for live music, where they optimised the loudness levels of each audio channel based on accumulated

loudness over time. Similarly in [50], the authors developed a cross-adaptive system, but it was offline. They used a psychoacoustic measure for loudness and partial loudness in order to optimise the gain settings for each track, with the aim of reducing inter-channel masking. Furthermore, they measured the loudness of each track when mixed with the combination of the other tracks in the multitrack. This is similar to the approach I took in Chapter 7.

[57] took a cross-adaptive approach similar to literature I have just discussed, however they used the EBU R-128 loudness measure. This is also a measure I have utilised in Chapter 7 as part of my mixing system. [9] developed an optimisation framework in order to adjust the levels of each audio track, which is an approach that influenced my work in Chapter 7.

The advantages of using just level based mixing are you that can get a relatively satisfactory mix using very little simplistic audio signal processing. However, inter-channel auditory masking may still be significant as this process does not allow for the spectral shaping of audio tracks that would be provided by tools such as equalisation and dynamic range compression. The optimisation framework used in [9] was used as an inspiration for the study performed in Chapter 7, where I used particle swarm optimisation to arrive at an optimal solution.

### 2.4.2 Equalisation

As well as looking at level adjustment, some other approaches to automatic mixing have been to adjust equalisation settings cross-adaptively. This is done to adjust the frequency content of each track, usually with the aim of reducing masking.

In [58], the authors proposed a system to automatically adjust equalisation settings with the aim of having equal average perceptual loudness on all frequencies amongst all multitrack audio channels. This system was designed to be used in a live context, which is not how my proposed system is designed to be used. However, this approach is still relevant.

In [59], the authors proposed a system where there was a target frequency spectrum and recursive IIR filters were set in order match an input signal to a desired target signal. This was done using the Yule-Walker algorithms [60]. In my work, I did not have a desired target frequency spectrum, but I did use an optimisation procedure in my work.

The paper that had the most similar implementation to the system I propose in Chapter 7 is [18]. They proposed a system for reducing inter-channel masking by using just equalisation, but they did not use an auditory model and instead based their measure

on spectral magnitude. I found the approach to inter-channel masking to be a useful approach as it allowed me to measure how much each individual track was being masked by all the other tracks in a multitrack.

### 2.4.3 Dynamic Range Compression

Dynamic range compression (DRC) while being an important tool in the arsenal of a mix engineer is also very useful for controlling the dynamic contour of audio over time. There have been a few publications that have used it an automatic mixing context.

Although, [61] and [6] do not describe how DRC could be used as part of a complete automatic mixing system. These publications are important with regard to how dynamic range compression works and how it can be used adaptively in the wider framework of automatic mixing.

In [5], the authors cross-adaptively set the parameters for DRC in a multitrack. The parameters were set based on loudness as well as loudness range (LRA). This was an interesting approach, but their motivation was to maintain equal loudness range between each of the audio tracks, where in my work I was looking at inter-channel auditory masking.

[8] proposed a system where DRC is applied based on audio features extracted from the sidechain, where the feature extraction process approach was derived from [6]. It was the first fully automated multitrack dynamic range compressor where all the parameters of a typical compressor were dynamically adjusted depending on extracted features and control rules. In relation to my work, I automated the DRC parameters differently and had no side chain feature extraction other than level. I also used optimisation with the intention of reducing masking.

### 2.4.4 Panning

Although I did not consider panning in my proposed automatic mixing system. It is still a very important part of the mix process and can be very effective at reducing auditory masking. Panning is useful as it allows a engineer to place instrument tracks at different points in a stereo field i.e. different instrument types can be placed left and right of the centre point of a mix. This is very useful especially if the different instruments types live in the same frequency range as each other. Some typical panning rules are to place bass instruments and lead vocals at the center of a mix [1, 7].

[4] proposed an adaptive digital audio effect for panning where a source is panned between two desired points based on the RMS of the signal. This is interesting, however it does require some user input and is not based on any spectral properties.

In [62], the authors propose a fully automated cross-adaptive system where each audio channel is panned based on loudness, spectral properties and is constrained based by typical panning rules. The audio channel pan positions are also updated over time, so the system is designed to work in real time.

In [63], the authors also had a fully automated cross-adaptive system, where the azimuth positions of the time frequency bins of each track are dynamically spread out with the aim of reducing auditory masking. They found that this approach reduces masking and could compete with a professional mix. This is an approach that I would like to explore in future work in conjunction with what is presented in Chapter 7.

## 2.5 Emotion in Music

### 2.5.1 Musically Induced vs. Perceived Emotions

In the study of emotion and music listening, induced emotions are those experienced by the listener and perceived emotions are those conveyed in the music, though perceived emotions may also be induced [64–66]. A listener's perception of emotional expression is mainly related to how they perceive and think about a musical process, in contrast to their emotional response to the music where someone experiences an emotion [66].

Perceived emotion in music can be provoked in a number of ways. It can be associated with the metrical structure of the music, or how a certain song might be perceived as happy or sad (valence) because of the chords being played [64]. Numerous studies have shown that any increase in tempo/speed, intensity/loudness or spectral centroid causes higher arousal. These studies have been summarised in [67]. In [67], tempo, loudness and timbre were shown to have an impact on how other typical 'musical' variables such as pitch and the major-happy minor-sad chord associations are perceived. Valence and arousal are two typical scales for measuring emotion in music. I discuss these scales in more detail further on in Section 2.5.2.

The most complete framework of psychological mechanisms for emotional induction is in [68] and its extensions [69, 70]. Until that point, most research in that area had been exploratory, but Juslin et al. posited a theoretical framework of eight different cognitive mechanisms known as BRECVEMA.

How both perceived and induced emotions in music relate to music production quality is an area of music and emotion that has not yet been explored. For both induced and perceived musical emotions I have proposed a number of ways in which a mix engineer may have a direct effect on these emotions. These are proposed with respect to BRECVEMA. The eight mechanisms and their potential relationship to music production quality are as follows:

- **Brain stem reflex** is a hard-wired primordial response that humans have to sudden loud noises and dissonant sounds. A reason given for the brain stem reflex reaction is the dynamic changes in music [70]. This particular mechanism might be related to music production in terms of a recording having good dynamics. A mix that has sudden large bursts in volume should arouse the listener more.

- **Rhythmic entrainment** is when the listener's internal body rhythm adjusts to an external source, such as a drum beat. This may relate to music production in a similar way as the brain stem reflex, i.e. if the drums in a musical production are loud and have a clear pulse, the listener may be more aroused.

- **Evaluative conditioning** occurs because a piece of music has been paired repeatedly with a positive or negative experience and an emotion is induced.

- **Emotional contagion** is when the listener perceives an emotional expression in the music and mimics the emotions internally [71]. This may mean that a better quality mix conveys the emotion in music in a clearer sense than a poorer quality mix, e.g. vocals or lead guitar is more audible in one mix over the other.

- **Visual imagery** may occur when a piece of music conjures up a particularly strong image. This could potentially have negative or positive valence and has been linked to feelings of pleasure and deep relaxation [70].

- **Episodic memory** is when music triggers a particular memory from a listener's past life. When a memory is triggered, so is an attached emotion [68]. A mix engineer might use a certain music production technique from a specific era, which may trigger nostalgia in the listener.

- **Musical expectancy** is believed to be activated by an unexpected melodic or harmonic sequence. The listener will expect musical structure to be resolved, but suddenly it is violated or changes in an unexpected way [71].

- **Aesthetic judgment** is the mechanism that induces 'aesthetic emotion' such as admiration and awe. This may play a part in music production quality by enhancing musically induced emotions. How well a song has been mixed can be

judged on the artistic skill involved as well as how much expression is in the mix. A poor mix is not typically going to be as expressive as a well constructed mix.

I seek to capture perceived and induced emotions from the listener with respect to music production quality through self-report, physiological measures, facial expression and body movement in chapter 6.

### 2.5.2 Psychological Models of Emotion

To describe musical emotions, three well known models may be employed; discrete, dimensional and music specific.

The discrete or categorical model is constructed from a limited number of universal emotions such as happiness, sadness and fear [72, 73]. One criticism is that the basic emotions in the model are unable to describe many of the emotions found in everyday life and there is not a consistent set of basic emotions [74, 75].

Dimensional models consider all affective terms along broad dimensions. The dimensions are usually related to valence and arousal, but can include other dimensions such as pleasure or dominance [76, 77]. Dimensional models have been criticised for blurring the distinction between certain emotions such as anger and fear, and because participants can not indicate they are experiencing both positive and negative emotions [66, 74, 75].

In recent years, a music-specific multidimensional model has been constructed. This is derived from the Geneva Emotion Music Scale (GEMS) and has been developed for musically induced emotions. This consists of nine emotional scales; wonder, transcendence, tenderness, nostalgia, peacefulness, power, joyful activation, tension and sadness [66, 78]. The scales have been shown to factor down to three emotional scales; calmness-power, joyful activation-sadness and solemnity-nostalgia [78, 79].

Empirical evidence [80, 81] suggests both discrete and dimensional models are suitable for measuring musically induced and perceived emotions [66]. [78] compared the discrete approach, the dimensional approach and the GEMS approach. It was found that participants preferred to report their emotions using the GEMS approach. Therefore, I adopted the GEMS approach as well as the dimensional model in my research.

### 2.5.3 Measuring Emotional Responses to Music

There are a number of different methods for measuring emotional responses to music. Here I discuss self-report, physiological measures, facial expression analysis and head nod-shake detection.

#### 2.5.3.1 Self-Report Methods

The most common self-report method to measure emotional responses to music is to ask listeners to rate the extent to which they perceive or feel a particular emotion, such as happiness. Techniques to assess affect are measured using a Likert scale or choosing a visual representation of the emotion the person is feeling. An example visual representation is the Self-Assessment Manikin [82] where the user is asked to rate the scales of arousal, valence and dominance based on an illustrative picture.

Another method is to present listeners with a list of possible emotions and ask them to indicate which one (or ones) they hear. Examples are the Differential Emotion Scale and the Positive and Negative Affect Schedule (PANAS). In PANAS, participants are requested to rate 60 words that characterise their emotion or feeling. The Differential Emotion Scale contains 30 words, 3 for each of the 10 emotions. These would be examples of the categorical approach mentioned previously [83, 84].

A third approach is to require participants to rate pieces on a number of dimensions. These are often arousal and valence, but can include a third dimension such as power, tension or dominance [74, 85].

Self-reporting leads to concerns about response bias. Fortunately, people tend to be attuned to how they are feeling (i.e., to the subjective component of their emotional responses) [86]. Furthermore, Gabrielsson came to the conclusion that self-reports are "the best and most natural method to study emotional responses to music" after conducting a review of empirical studies of emotion perception [64]. One caveat with retrospective self-report is 'duration neglect' [87], where the listener may forget the momentary point of intensity of the emotion attempted to be measured.

I chose self-report in my experiment due to it being the most reliable measure according to [64]. GEMS-9 was used for measuring induced emotion and Arousal-Valence-Tension for perceived emotion. I selected GEMS-9 to report induced emotions over a dimensional method due to it being a specialised measure for the self-report of musically induced emotions. I then chose to use Arousal-Valence-Tension due to it being a dimensional rather than categorical model like GEMS-9. This allowed me to use two different models of self-report.

#### 2.5.3.2 Physiological Measures

Measures for recording physiological responses to music include heart or pulse rate, galvanic skin response, respiration or breathing rate and facial electromyography. Such measures have been used in recent papers [71, 88, 89].

High arousal or stimulative music tends to cause an increase in heart rate, while calm music tends to cause a decrease [90]. Respiration has been shown to increase in 19 studies on emotional responses to music [90]. These studies found differences between high- and low-arousal emotions but few differences between emotions with positive or negative valence.

One physiological measure that corresponds with valence is facial electromyography (EMG). EMG measurements of cheek and brow facial muscles are associated with processing positive and negative events, respectively [91]. In [92], each participant's facial muscle activity was measured while they listened to different pieces of music that were selected to cover all parts of the valence-arousal space. Results showed greater cheek muscle activity when participants listened to music that was considered high arousal and positive valence. Brow muscle activity increased in response to music that was considered to induce negative valence, irrespective of the arousal level.

*Galvanic skin response* (GSR) is a measurement of electrodermal activity or resistance of the skin [93]. When a listener is aroused, resistance tends to decrease and skin conductance increases [94, 95]. I used ECG and skin conductance measurements in my experiment as it had been used extensively in previous studies related to music and emotion [71, 88–90]. I also felt it would be better to have as many measures as feasibly possible, since it is much easier to throw away data rather than re-run an experiment with more measurements.

#### 2.5.3.3 Facial Expression and Head Movement

The Facial Action Coding System (FACS) [96] provides a systematic and objective way to study facial expressions, representing them as a combination of individual facial muscle actions known as Action Units (AU). Action Units can track brow and cheek activity, which can be linked to arousal and valence when listening to music [92].

[97] examined how schizophrenic patients perceive emotion in music using facial expression, and [98] looked at the role of a musical conductors facial expression in a musical ensemble. I were unable to find anything directly related to my research questions.

People move their bodies to the rhythms of music in a variety of different ways. This can occur through finger and foot tapping or other rhythmic movements such as head nods and shakes [99, 100]. In human psychology, head nods are typically associated with a positive response and head shakes negative one [101]. In one study, participants who gauged the content of a simulated radio broadcast more positively were more inclined to nod their head than those who performed a negatively associated head shaking movement [100, 102]. But for music, a head shake might be considered a positive response as this might simply be a rhythmic response.

I examined facial expression in this experiment since it had not been attempted before in music and emotion or music production quality research. Facial expression analysis is somewhat similar to facial EMG, so we should be able to link results to previous findings [90].

## 2.6 Feature Learning and Classification

I discuss the background of some relevant machine learning topics here as they are important background for Chapter 5.

### 2.6.1 Decision Trees

Decision trees are a commonly used machine learning classifier that belong to the family of supervised learning algorithms. Decisions trees can be used for either classification or regression tasks, where these trees are Classification And Regression Tree's (CART). Decision trees build either a classification or regression model in a tree structure, where they take a dataset and recursively break the dataset down into smaller and smaller datasets using a technique called recursive partitioning. The dataset is broken down based on a feature value test, the test usually being Gini Diversity Index (GDI). GDI is calculated as

$$GDI = 1 - \sum_{i} p(i)^2 \tag{2.6}$$

where $i$ is the class and $p(i)$ is the fraction of objects within class $i$ following the branch. I refer the reader to [103] for a further discussion on CART. An example decision tree is shown in Figure 2.4

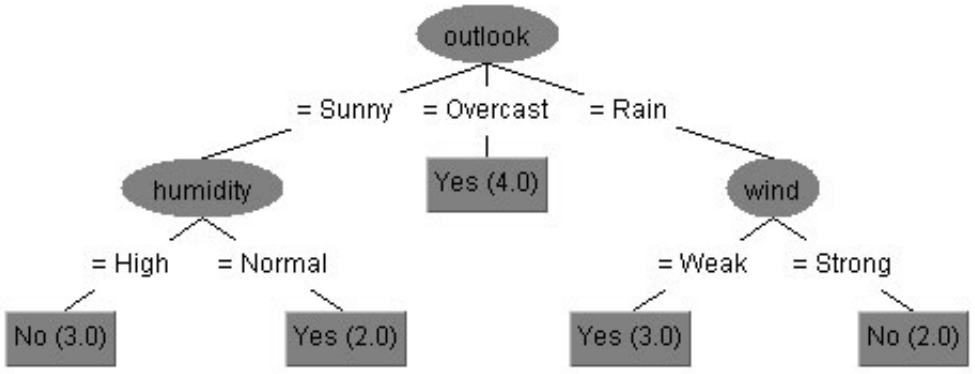


FIGURE 2.4: This is an example of how different decisions are arrived at based on certain features [104]. In my case this might be audio features such as RMS or Spectral Centroid.

### 2.6.2 Random Forest

Random Forest is a particular type of Ensemble Learning method based on growing decision trees. This can be used for either classification or regression problems, but can also be used for feature selection. After training has occurred on a dataset each decision tree that is grown predicts an outcome. For regression decision trees, the output is the average value predicted by all of the decision trees grown. For classification decision trees it is the classification outcome that was voted most popular by all of the decision trees grown [105]. Random Forest is based on the idea of bootstrap aggregating or more commonly know as bagging. Bagging in this instance is where each decision tree makes a decision and the majority decision is what is used to make a prediction. In the context of my work this could be a prediction of what type of subgroup an audio track belongs to.

Random Forest was chosen because it has been proven to work very well for feature selection in other fields such as bio-informatics and medicine [106, 107]. Also Random Forest is know to generalise well and tends to avoid over fitting due to the the cross-validation that is inherent in the algorithm.

The measure of a Random Forest's accuracy is known as its Out-of-bag error (OOB). I refer the reader to [105] for a more detailed explanation of the specifics of this classifier.

### 2.6.3 Feature Selection

When attempting to train a machine learning classifier, each data point you feed the classifier has features that represent it. It is these features that the classifier learns from

and ultimately makes decisions from. However, sometime the features provided to the classifier can be redundant or highly correlated with other features. When this occurs, there may by too many unimportant features trying to describe something, which wastes computation time and can reduce a trained classifier's discriminative power.

Feature selection is the iterative process of removing poorly performing features and selecting the features that give you the most discriminative power. There are a number of different approaches to going about this, which are out of the scope of this thesis. I refer the reader to [103] for a more detailed explanation.

In this work, the Random Forest classifier was used to perform feature selection. Random Forest was chosen as it is robust, easy to tune and requires very little feature engineering. It can also be setup to avoid biased variable selection by using subsampling without replacement [106]. It was also chosen as it was found to out-perform Naive Bayes and SVM's when used for another classification task explained in Appendix 9.1.

The Random Forest gives a Feature Importance Index (FII). This ranks all features in terms of importance by evaluating the OOB error for each tree grown with a given feature, to the overall OOB error. Random Forest feature importance can be defined for $X^i$, where the vector $X = (X^1, ...X^p)$, contains feature values and where $p$ is the number of audio features used. For each tree $\tau$ in the Random Forest, consider the associated $OOB_\tau$ sample (this is the out-of-bag data that is not used to construct $\tau$). $errOOB_\tau$ denotes the error of a single tree $\tau$ using the $OOB_\tau$ sample. The error being a measure of the Random Forest classifier's accuracy. If the values of $X^i$ are randomly permuted in $OOB_\tau$ to get a different sample denoted by $\widetilde{OOB^j_\tau}$ and we compute $\widetilde{errOOB^j_\tau}$. $\widetilde{errOOB^j_\tau}$ being the error of $\tau$ because of the different sample. The feature importance of $X^i$ is equal to:

$$FI(X^i) = \frac{1}{ntree} \sum_\tau (\widetilde{errOOB^j_\tau} - \widetilde{OOB^j_\tau}) \tag{2.7}$$

where the sum is over all trees $\tau$ of the Random Forest and $ntree$ is the number of trees in the Random Forest [108].

### 2.6.4 Hierarchical Clustering

Hierarchical clustering is a type of unsupervised data clustering. Generally in Hierarchical clustering a cluster hierarchy or a tree of clusters, also known as a dendrogram is constructed. An example of a dendrogram can be seen in Figure 2.5. Hierarchical clustering methods are categorised into agglomerative and divisive. The agglomerative

clustering method is what I used in thesis. The idea is that the algorithm starts with singular clusters and recursively merges two or more of the most similar clusters [109]. The reason why I chose agglomerative clustering is because the algorithmic process is similar to how a human would create subgroups in a multitrack. Initially, a human would find two audio tracks that belong together in a subgroup and then keep adding audio tracks until a subgroup is formed. An example would be pairing a kick track with a snare track and then pairing them with a hi-hat track to create a drum subgroup. It is also worth noting that Figure 1.1 which is a typical subgrouping setup can be likened to a tree structure, so it would make sense to attempt to cluster audio tracks in a tree like fashion. It also provides the benefit of providing cophonetic distances between different clusters, so that the relative distances between nodes of the hierarchy are clear.

The agglomerative clustering algorithm can be described as thus [110]. Given a set of $N$ audio feature vectors to be clustered.

1. Assign each audio feature vector $V_{audio}$ to its own singleton cluster and number the clusters 1 through $c$.
2. Compute the between cluster distance $d(r, s)$ as the between object distance of the two objects in $r$ and $s$ respectively, $r, s = 1, 2, ..., c$. Where $d(r, s) = \sqrt{\sum_c (r_c - s_c)^2}$ is the Euclidean distance function and let the square matrix $D = (d(r, s))$.
3. Find the most similar pair of clusters $r$ and $s$, such that the distance, $D(r, s)$, is minimum among all the pairwise distances, $d(c_i, c_j) = min\{d(r, s) : r \in c_i, s \in c_j\}$. This is what is known as the linkage function. A similar pair of clusters could be a snare track and a hi-hat track.
4. Merge $r$ and $s$ to a new cluster $u$ and compute the between-cluster distance $d(u, k)$ for any existing cluster $k \neq r, s$. Once the distances are obtained, remove the rows and columns corresponding to the old cluster $r$ and $s$ in $D$, since $r$ and $s$ do not exist any more. Then add a new row and column in $D$ corresponding to cluster $u$. Merging two clusters is like grouping two audio tracks together or else adding an audio track to an existing subgroup.
5. Iteratively repeat steps 3 to 5 a total of $c - 1$ times until all the data items are merged into one cluster.

## 2.7 Summary

The audio engineering concepts, affect analysis approaches, machine learning techniques and relevant computational background to this thesis were introduced in this chapter. I

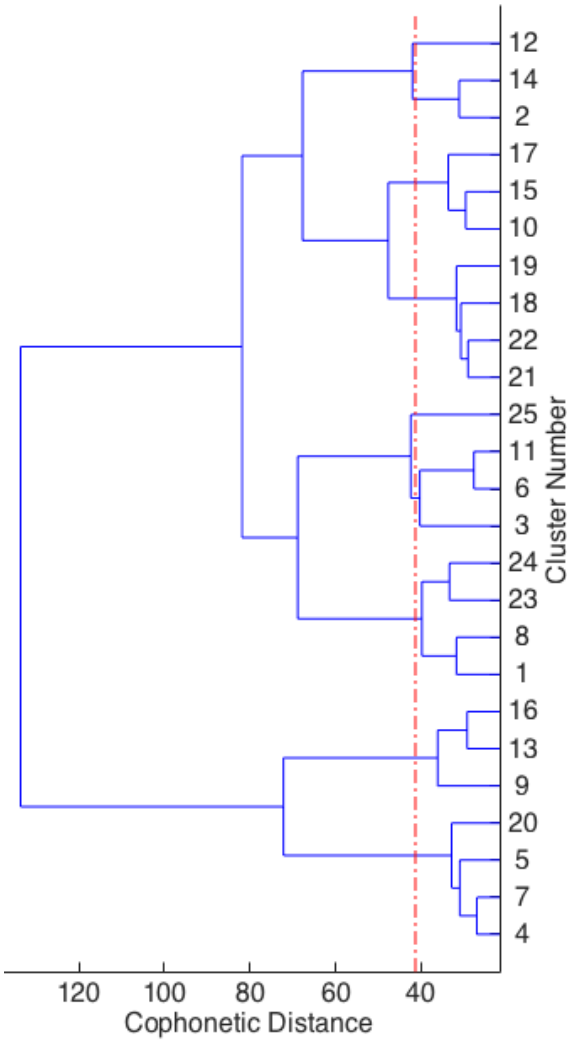


FIGURE 2.5: An example dendrogram

also reviewed existing automatic mixing systems, where I went through the individual audio effect types that were being automated.

I found there to be no automatic system where there was an emphasis on mixing in subgroups. I also found the literature around subgrouping to be quite limited. This is what inspired the work carried out in chapter 3 and chapter 4, since I needed to document and define the subgrouping process in greater detail. I also found no automatic system that makes use of both DRC and equalisation. Typically a mix engineer will make use of both DRC and equalisation when mixing as they are essential tools for frequency and dynamics processing. This is why I chose to use these effects in chapter 7.

I also found there were no studies that examined the relationship between music production quality and emotional response. This is something I investigated in chapter 6 and is important in the context of automatic mixing systems. It is important because I

do not know how good a mixing system needs to be, before it actually becomes a useful tool to a beginner producer.

# Chapter 3

# The impact of subgrouping practices on the perception of multitrack mixes

## 3.1 Introduction

The aim of this chapter is to investigate how different mix engineers perform subgrouping, and what kind of subgroup processing they use. Furthermore, I attempt to quantify what effect subgrouping has on the subjective quality of a mix. Section 5.2 provides the details of a mix experiment from which I gathered the subgrouping data. Section 5.5 provides the results obtained from the mix session files, which are analysed and discussed in Section 3.4. In Section 3.5, I summarise my findings and outline future work.

## 3.2 Dataset

### 3.2.1 Experiment

A dataset of mixes and mix projects obtained from the Open Multitrack Testbed [111] was examined to see how many subgroups were created by the mix engineers, what kind of subgroup processing they used and how the mix engineers created the subgroups. This dataset was the same data recorded from an experiment that had been previously conducted [14]. In this experiment, different mixes of different songs were rated by experienced subjects. These mixes were rated from (0-100) in terms of how much each participant preferred the mix quality. Each listener compared 8 mixes of each song and

| **Song name** | **Genre** | **Mix engineers** |
|---|---|---|
| Red To Blue (S1) | Pop-Rock | A - H |
| Not Alone (S2) | Funk | A - H |
| *My Funny Valentine* (S3) | Jazz | A - H |
| Lead Me (S4) | Pop-Rock | A - H |
| In The Meantime (S5) | Funk | A - H |
| - (S6) | Soul-Blues | I - P |
| *No Prize* (S7) | Soul-Jazz | I - P |
| - (S8) | Pop-Rock | I - P |
| Under A Covered Sky (S9) | Pop-Rock | I - P |

TABLE 3.1: Song title, genre and mix group. Songs in italics are not available online due to copyright restrictions.

then gave each one a rating. In the context of my research, this allowed me to investigate the relationship between subgrouping and how preferred a mix was.

The mix engineers in this experiment were students of the MMus in Sound Recording at the Schulich School of Music, McGill University. Each song was mixed by one of the two classes of eight students each, such that one group of students mixed five songs in total (over three semesters - four as first years and one more as second years), and one group mixed four songs in total (over two semesters) [14]. A breakdown of which songs were mixed by which group can be seen in Table 3.1.

Five out of nine songs are available on the Open Multitrack Testbed[1] [111] including raw tracks, the rendered mixes and the complete Pro Tools project files, allowing others to reproduce or extend the research.

### 3.2.2 Data Extraction

The data for each mix engineer's subgrouping setup was extracted manually from each of their Pro Tools session files. Information extracted from each session file included how many subgroups there were, if any subgroup processing such as equalisation (EQ), dynamic range processing (DRC) and reverb were used, and if subgroup send processing was used. Subgroup send processing is when the audio from a subgroup is sent to an auxiliary track or outboard device for audio processing.

I also logged the instruments in each subgroup, to determine on what basis different tracks are subgrouped, and whether the subgroups were hierarchical. I define a hierarchical subgroup as a type of subgroup that groups two or more subgroups together. An

[1] multitrack.eecs.qmul.ac.uk

| **Subgroup type** | **# subgroups** | **# tracks** |
|---|---|---|
| Vocals | 90 | 324 |
| Drums | 78 | 680 |
| Guitars | 69 | 371 |
| Keys | 56 | 164 |
| Bass | 47 | 88 |
| Other percussion | 17 | 43 |
| Brass | 12 | 33 |
| Strings | 10 | 24 |

TABLE 3.2: The number of different individual subgroup types and how many audio tracks of that type occurred in all the mixes.

example would be a guitar subgroup that contains a rhythm guitar subgroup and a lead guitar subgroup.

The overall preference score for each mix engineer on each mix was calculated by taking the median rating value given by the mix engineers and the mix professionals from the other group participating in the experiment. I used the median value as the mix preference ratings are not all normally distributed. However, I found that the difference between the median and mean mix preference ratings were not large enough to report separately. The distributions of the mix preference ratings for each mix engineer are presented in the results section.

## 3.3 Results

Table 3.2 shows a breakdown of the most commonly created individual subgroup types. The subgroup type indicates the main instrument type in that subgroup. I found there to be eight individual subgroup types and drums was the most common instrument type in all of the mix projects. Table 3.3 shows that a number of subgroups contained combinations of instruments. I also found that almost all mix engineers subgrouped audio tracks based on instrumentation and only four out of the 72 mixes had no subgroups at all, in which three out of the four mixes were of the same song.

Table 3.4 shows how many hierarchical subgroups I had in the mixes I examined. Drums and vocals were the only single instrument types that were hierarchically grouped and the rest were combinations of instrument types. The most hierarchically subgrouped instrument was drums. Furthermore, I found that hierarchical subgroups were present in 19 of the 72 mixes examined.

In Tables 3.5 and 3.6 I present the absolute amount of subgroups created by each mix engineer for each of the songs they mixed. The number in the parentheses is the number of audio tracks that each mix engineer used for each mix. The reason there is a variation in the audio track number for each mix is because some mix engineers duplicated audio tracks or else completely left them out of the mix.

Table 3.7 shows the different amount of track types available to each mix engineer before they began to mix. The subgroup types used in Table 3.2 are based on the different audio track types I found for each song.

In Tables 3.8 and 3.9 I present the correlations (Spearman's rank correlation coefficient) of the average amount of subgroups, EQ subgroups, DRC subgroups and EQ + DRC subgroups created per mix engineer with median mix preference as well as the correlation (Spearman's rank correlation coefficient) of the amount of subgroups, EQ subgroups, DRC subgroups and EQ + DRC subgroups created per mix with median mix preference.

| **Subgroup type** | **# subgroups** |
|---|---|
| Bass + Guitars + Keys + Vocals | 4 |
| Drums + Bass + Guitars + Keys | 4 |
| Bass + Guitars + Keys | 3 |
| Drums + Percussion | 3 |
| Guitars + Keys | 3 |
| Drums + Bass + Vocals | 1 |
| Drums + Bass | 1 |
| Bass + Guitars | 1 |
| Drums + Bass + Keys + Vocals | 1 |

TABLE 3.3: The number of different multi-instrument subgroup types that occurred in all the mixes.

| **Hierarchical subgroup type** | **No. of hierarchical subgroups** |
|---|---|
| Drums | 10 |
| Vocals | 3 |
| Bass + Guitar + Keys + Vocals | 2 |
| Drums + Bass + Guitars + Keys | 2 |
| Drums + Bass + Vocals | 1 |
| Bass + Guitar + Keys | 1 |
| Drums + Vocals | 1 |
| Drums + Bass + Keys + Vocals | 1 |
| Bass + Guitars | 1 |

TABLE 3.4: The number of different hierarchical subgroup types that occurred in all the mixes.

| | **S1** | **S2** | **S3** | **S4** | **S5** |
|---|---|---|---|---|---|
| **A** | 10 (44) | 10 (25) | 9 (17) | 9 (23) | 3 (26) |
| **B** | 2 (45) | 5 (28) | 8 (17) | 7 (22) | 6 (25) |
| **C** | 13 (42) | 8 (25) | 9 (17) | 6 (25) | 8 (25) |
| **D** | 4 (43) | 3 (25) | 0 (19) | 4 (23) | 3 (25) |
| **E** | 10 (45) | 7 (25) | 9 (19) | 10 (23) | 8 (25) |
| **F** | 2 (44) | 3 (25) | 0 (19) | 7 (23) | 4 (25) |
| **G** | 8 (43) | 8 (25) | 0 (19) | 6 (23) | 6 (25) |
| **H** | 6 (43) | 3 (25) | 9 (19) | 8 (23) | 6 (25) |

TABLE 3.5: The number of subgroups created for each song by each each mix engineer in mix group A - H. The number of audio tracks used in each mixing project is in parentheses.

| | **S6** | **S7** | **S8** | **S9** |
|---|---|---|---|---|
| **I** | 7 (18) | 3 (12) | 3 (16) | 5 (28) |
| **J** | 7 (25) | 4 (17) | 4 (25) | 7 (28) |
| **K** | 7 (26) | 0 (17) | 1 (28) | 5 (28) |
| **L** | 6 (25) | 6 (17) | 4 (20) | 3 (30) |
| **M** | 10 (25) | 7 (17) | 4 (25) | 3 (22) |
| **N** | 8 (25) | 3 (17) | 6 (25) | 4 (29) |
| **O** | 9 (25) | 5 (18) | 8 (26) | 8 (29) |
| **P** | 6 (14) | 6 (20) | 5 (29) | 6 (22) |

TABLE 3.6: The number of subgroups created for each song by each each mix engineer in mix group I - P. The number of audio tracks used in each mixing project is in parentheses.

| **Track type** | **S1** | **S2** | **S3** | **S4** | **S5** | **S6** | **S7** | **S8** | **S9** |
|---|---|---|---|---|---|---|---|---|---|
| Vocals | 17 | 9 | 1 | 6 | 9 | 4 | 1 | 4 | 10 |
| Drums | 11 | 10 | 9 | 9 | 10 | 10 | 8 | 10 | 9 |
| Guitars | 12 | 2 | 0 | 6 | 2 | 2 | 5 | 7 | 15 |
| Keys | 1 | 4 | 2 | 2 | 2 | 2 | 1 | 2 | 1 |
| Bass | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 |
| Other percussion | 1 | 0 | 4 | 1 | 0 | 1 | 0 | 0 | 0 |
| Brass | 0 | 0 | 1 | 0 | 0 | 3 | 0 | 0 | 0 |
| Strings | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 |

TABLE 3.7: The number of different audio track types in each song before they were mixed.

| **Ratio type** | $\boldsymbol{\rho}$ |
|---|---|
| *Subgroup - Audio Track Ratio* | 0.62 ($p < 0.01$) |
| *Subgroup EQ - Audio Track Ratio* | 0.67 ($p < 0.01$) |
| *Subgroup DRC - Audio Track Ratio* | 0.45 ($p < 0.05$) |
| *Subgroup EQ + DRC - Audio Track Ratio* | 0.59 ($p < 0.01$) |

TABLE 3.8: Average amount of subgroups, EQ subgroups, DRC subgroups and EQ + DRC subgroups created per mix engineer and its correlation (Spearman's rank correlation coefficient) with median mix preference.

| **Ratio type** | $\boldsymbol{\rho}$ |
|---|---|
| *Subgroup - Audio Track Ratio* | 0.32 ($p < 0.01$) |
| *Subgroup EQ - Audio Track Ratio* | 0.4 ($p < 0.01$) |
| *Subgroup DRC - Audio Track Ratio* | 0.35 ($p < 0.01$) |
| *Subgroup EQ + DRC - Audio Track Ratio* | 0.38 ($p < 0.01$) |

TABLE 3.9: Amount of subgroups, EQ subgroups, DRC subgroups and EQ + DRC subgroups created per mix and its correlation (Spearman's rank correlation coefficient) with median mix preference.

I chose the Spearman's rank correlation coefficient since it is non-parametric and my data was not normally distributed. The number of subgroups in the correlation scores is presented as the number of created subgroups relative to how many audio tracks the mix engineer used to create the final mix. I call this the *Subgroup - Audio Track Ratio*. This also applies to the different types of processing applied to each subgroup, so I have the *EQ Subgroup - Audio Track Ratio*, the *DRC Subgroup - Audio Track Ratio* and the *EQ + DRC Subgroup - Audio Track Ratio*. The *EQ + DRC Subgroup - Audio Track Ratio* is a measure of when a subgroup was created and both EQ and DRC processing are applied. Ratios were used because larger mixes with more instrumentation are likely to have more subgroups. This allowed us to compare the amount of subgroups created and the types of subgroup processing used on a mix by mix basis. This linear relationship is evident in Table 3.2 where we see that when more audio tracks are available there tends to be more subgroups created. In fact, the Spearman rank correlation coefficient for this relationship is very strong and significant with a value of 0.93 ($p < 0.01$).

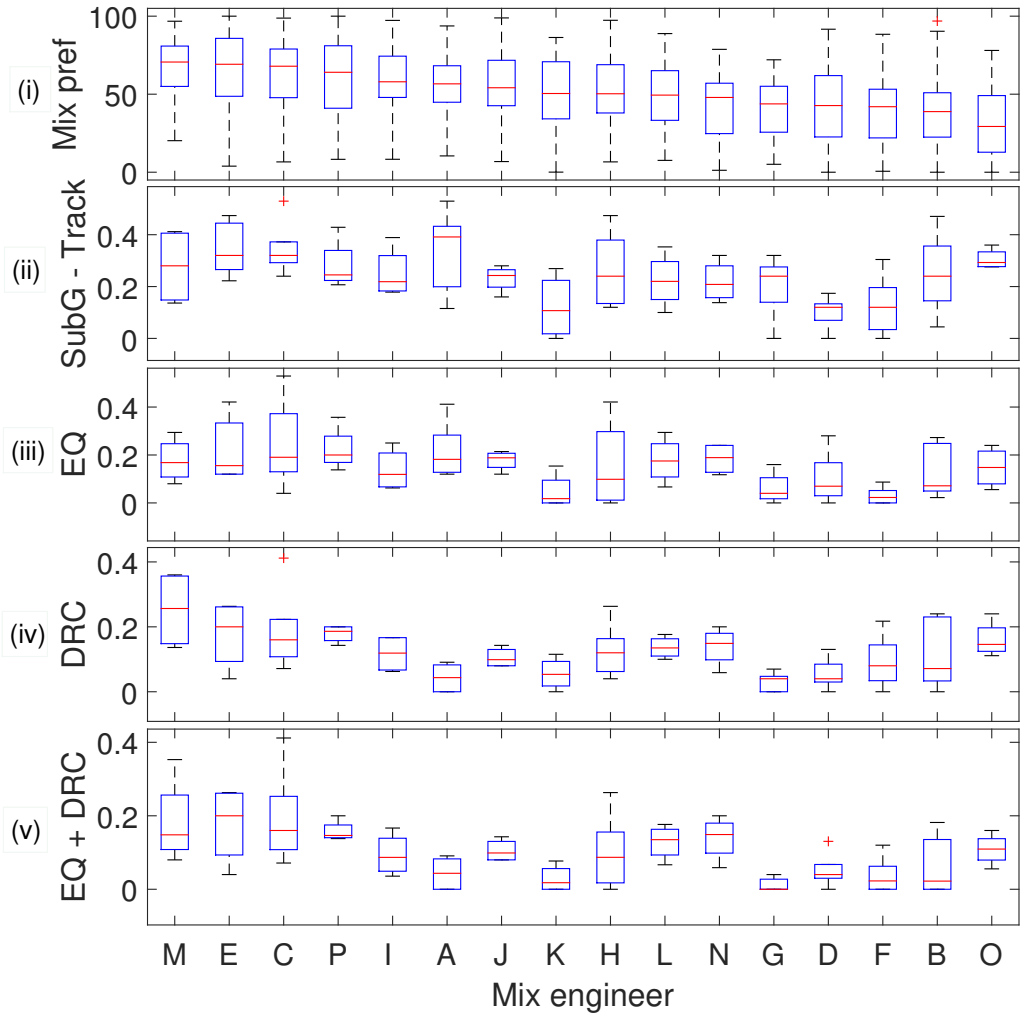


FIGURE 3.1: (i) shows each mix engineer's mix preference ratings ranked from highest to lowest median value. (ii - v) show the *Subgroup - Audio Track Ratio's*, the *EQ Subgroup - Audio Track Ratio's*, the *DRC Subgroup - Audio Track Ratio's* and the *EQ + DRC Subgroup - Audio Track Ratio's* for all the mixes created by each mix engineer.

## 3.4 Analysis and Discussion

In Tables 3.2-3.4 I summarised the different subgroup types that were created in all the mixes examined. I looked at standard subgroups and hierarchical subgroups. Table 3.2 shows that the top three standard subgroups were vocals, drums and guitars. In a mix there can be many different vocalist types. There may be a lead vocalist, a secondary vocalist and background vocalists. This would explain why it is the most subgrouped instrument type. The mix engineers may have wanted to control and process different subgroups of singers that are singing in different styles or singing different parts of each song. The song Red to Blue (S1) is a perfect example of when this occurs. Three of the eight mix engineers have split the vocal tracks into separate subgroups for processing.

One of the mix engineers was doing this for simple gain processing, but the other two mix engineers were doing it for gain processing as well as applying EQ and DRC processing. Also, vocals tend be the most important instrument type in a mix. In [112] it was shown that most of the listener's attention and about a third of the critical comments on the same mixes used in this paper were about vocals. It has also been shown that the vocals are consistently the loudest instrument type in the same mixes I examined [13].

The second most subgrouped instrument type was drums. Drums are an important part of a mix as they are the rhythm section that keeps the rest of the song in time, so it would be important to be able to control how loud they are in a mix. It is also worth mentioning that in [12], under testing the assumption *"Gentle bus/mix compression compression helps blend things better"*, it was found that some professional mix engineers like to apply DRC to the drums as a subgroup. Drums also have the most amount of instrument tracks in all of the mix projects, see Table 3.2.

The third most frequently subgrouped instrument type was the guitars. Guitars are similar to vocals because it is possible to have different styles of guitar playing in a single mix. An arrangement might contain lead guitars and rhythm guitars, distorted and clean guitars, and electric and acoustic guitars. All of these guitar types serve a different purpose in a mix, so it is easy to see how a mix engineer might want to control them or process them individually. An example might be that a mix engineer wants to apply more EQ to a particular group of guitars. Something like this occurred in two separate mixes for the song Red to Blue (S1). One mix engineer had a subgroup for 'Heavy' guitars which used EQ processing, while another mix engineer had a subgroup for 'Lead' guitars which used DRC processing. I also found that acoustic guitars were subgrouped separate to other guitar types in 13 of the mixes I examined. Furthermore, in five of the 13 mixes, EQ or DRC subgroup processing was being applied to the acoustic guitars.

Interestingly, only four out of the 72 mixes did not use any subgrouping at all and three of these were of the same song. On examination of the instrumentation of the song where three mix engineers did not create any subgroups, I found there were flute, harp, vibraphone, piano and violin tracks. There was also no guitar tracks and only one vocal track. It might have been through inexperience that the mix engineers may not have known how to approach creating subgroups for instruments such as flutes, harps and vibraphones. However, it was found in 4 that six out of the ten professional mix engineers that were interviewed created subgroups based on genre. This suggests there could have been a style or genre dependency on how the mix engineers in the experiment created the subgroups for this particular song.

Table 3.4 shows that the most hierarchically subgrouped instrument type was drums. It was found on examining the many different mixes, in eight of the mixes, the mix engineers chose to separate the overhead microphones from the rest of the drum recordings. As the overhead microphones are often treated as a stereo pair with left and right microphones, grouping these into one channel allows simultaneous processing. I also found that some mix engineers chose to group the kick, snare and hi-hats separately. The kick, snare and hi-hats are the most important instruments in a drum kit and I found seven mixes where this occurred. Furthermore, 19 out of the 72 mixes used some form of hierarchical subgrouping, so this shows that it is a style of subgrouping that is practised often.

Table 3.8 shows there is a strong significant Spearman correlation of 0.62 ($p < 0.01$) between the average *Subgroup - Audio Track Ratio* per mix engineer and the median mix preference rating. This implies that the more the mix engineer creates subgroups on average, the higher the mix preference rating they receive.

In Table 3.8 there is a strong significant Spearman correlation of 0.67 ($p < 0.01$) between the average *EQ Subgroup - Audio Track Ratio* per mix engineer and the median mix preference rating. The strong *EQ Subgroup - Audio Track Ratio* correlation implies that the more EQ subgroup processing that occurs the higher a mix preference rating the mix engineer receives. The strong correlation also gives us confidence that this type of subgroup processing is an important mixing technique. This subgroup processing technique might be done frequently by a mix engineer, so that they can apply EQ to a group of instruments as a whole and stop them from masking another group of instruments.

Table 3.8 shows there is a moderate significant Spearman correlation of 0.45 ($p < 0.05$) between the average *DRC Subgroup - Audio Track Ratio* per mix engineer and the median mix preference rating. I was surprised to see such a low correlation for the *DRC Subgroup - Audio Track Ratio* as I would have expected people to process a lot of their subgroups with DRC. This seems to go against the assumption made in [12], but this may be because the participants in the experiment do not have the same level of experience as the mix engineers interviewed in [12] or I simply have not examined enough mixes to see this trend.

Table 3.8 shows a moderate significant Spearman correlation of 0.59 ($p < 0.05$) between the average *EQ + DRC Subgroup - Audio Track Ratio* per mix engineer and the median mix preference rating. I also expected the relationship between subgroups created that use EQ + DRC processing and mix preference rating to be stronger, but it is probably not as strong as I hoped since it corresponds with the moderate correlation for DRC subgroup processing.

Table 3.9 show there is a weak significant Spearman correlation of 0.32 ($p < 0.01$) between the *Subgroup - Audio Track Ratio* per mix and the median mix preference rating. This implies that there is very little relationship between the amount of subgroups created and mix preference when I consider each mix individually. This suggests that the assumption that creating more subgroups leads to a higher mix preference does not apply to mixes universally, but is more specific to the mix engineer. What I mean by this is that there may be latent variables involved I am not yet considering.

Table 3.9 shows there is a moderate significant Spearman correlation of 0.40 ($p < 0.01$) between *EQ Subgroup - Audio Track Ratio* and mix preference over all the mixes created. This is not as strong as the result in Table 3.8. In Table 3.9 we see a weak significant Spearman correlation of 0.35 ($p < 0.01$) between *DRC Subgroup - Audio Track Ratio* and mix preference over all the mixes created. Table 3.9 also shows a weak significant Spearman correlation of 0.38 ($p < 0.01$) between *EQ + DRC Subgroup - Audio Track Ratio* and mix preference over all the mixes created. This shows that the correlations are not strong for subgroup processing when I consider each mix individually, but are stronger when we examine each mix engineer individually. This leads us to further believe that there are other factors that I am not considering and the results from Table 3.8 may not be generalisable. Subgrouping and subgroup processing may only work well for some mix engineers.

Figure 3.1 plots the distribution of all the variables I correlated and are ranked from left to right in descending median mix preference value for each mix engineer. The distributions of *Subgroup - Audio Track Ratio's* of the top three ranked mix engineers (M, E and C) show that overall, the median value are higher than 10 of the other mix engineers. It also shows that the amount of subgroups they created varied over each of their mixes if I include the outlier for mix engineer C. This implies that each mix engineer considers how many subgroups they will create for each mix as opposed to an arbitrary number of subgroups. If we look at the *EQ Subgroup - Audio Track Ratio*'s, the median results are similar for the top three mix engineers, but it varies more from left to right. The inverse seems to be true for the *DRC Subgroup - Audio Track Ratio* as the median decreases going from left to right, as well as the amount of variance. If I compare the results of the top three mix engineers with the rest of the mix engineers I do see a trend of higher *Subgroup - Audio Track Ratios, EQ Subgroup - Audio Track Ratio's, DRC Subgroup - Audio Track Ratios* and *EQ + DRC Subgroup - Audio Track Ratios* than the other mix engineers. This is not true in all cases, but is a general observation.

## 3.5 Conclusion

From the experimental results I found that subgroups are mainly made up of similar instrumentation, but in some cases can be a combination of different types of instrumentation. However, I found the former to occur much more often. I found that the three instrument types that were subgrouped together the most were drums, vocals and guitars. I also found that when hierarchical subgrouping occurred, it was usually applied to drums and to a lesser extent vocals. I was able to show there was a strong significant Spearman correlation when looking at the median mix preference score of all the mixes done by each mix engineer and the amount of subgroups this mix engineer created on average. I also found a strong significant Spearman correlation when looking at the median mix preference score of all the mixes done by each mix engineer and the amount of EQ subgroup processing this mix engineer used on average. There was also a moderate significant Spearman correlation when looking at the median mix preference score of all the mixes done by each mix engineer and the amount of DRC subgroup processing this mix engineer used on average.

The results provide an important insight into the relationship between mix preference and the ubiquitous, but poorly documented practice of subgrouping. There appears to be a very distinct relationship between the number of subgroups used and mix preference. This may be because the mix engineer is able to exercise greater control over the mix through subgrouping as well as being able to treat an entire instrument group with effects processing. However, I do not know whether these findings apply to every mix engineer, since I only examined the mixes of 16 mix engineers in one university. This makes the results difficult to generalise. There is also potential for bias due to how they may have been taught to mix by the instructor. Correlation does not necessarily imply causation either, and more subgroups may not necessarily imply higher mix preference in this case. As mentioned already, there are a number of confounding variables to consider such as the previous experience of each mix engineer as well as the song preference and genre preference of the raters. All these variables can add bias to the presented results and need to be considered.

Overall, this research contributes to a deeper understanding of this poorly documented mixing practice. Informed by these results, further research questions emerge that require a larger dataset, and which could be answered by collecting and analysing a larger and more diverse set of mixes. Future work will be to further examine the link between EQ subgroup processing, DRC subgroup processing and mix preference.

# Chapter 4

# Analysis of the subgrouping practices of professional mix engineers

## 4.1 Introduction

This chapter sheds light on the ubiquitous but poorly defined mix practice of subgrouping, and provides rules and constraints derived from a questionnaire that could be used in intelligent audio production tools. I prepared an online questionnaire consisting of 21 questions testing nine assumptions in order to identify subgrouping decisions, such as why a mix engineer creates subgroups, when they subgroup and how many subgroups they use.

Previously, I analysed a number of multitrack mixes to determine how mix engineers created subgroups, how they apply subgroup effect processing such as equalisation (EQ) and dynamic range compression (DRC), and if there was any link between subgrouping and mix preference [113]. I had access to actual multitrack project files and were able to analyse exactly how each participant had constructed subgroups and what effect processing had been applied. However, the mixes that were analysed were created by three separate groups of music production students, so their level of mix engineering experience was contentious [114].

Section 4.2 describes the methodology used in this chapter. I describe the questionnaire, my hypotheses, how I approached the qualitative and quantitative analysis. Following that, I present the results and my analysis in section 4.3. I discuss participants, coding and theme development, and then analyse each theme in the context of the questions in

the survey in section 4.4. In section 4.5 I discuss the results and analysis in relation to my hypotheses and make recommendations based on my findings.

## 4.2 Methodology

### 4.2.1 Survey Questionnaire

Before the survey was conducted I proposed a number of assumptions about how mix engineers subgroup, and many survey questions were designed to test these assumptions. The assumptions are listed in Table 4.1. These assumptions were developed from audio engineering literature [1, 7, 17], from discussions with other mix engineers, academics and from past experiences in the field. The questionnaire that I used to test these assumptions is provided in Appendix 9.3.

TABLE 4.1: Subgrouping assumptions

| Assumptions | Description |
|---|---|
| **A1** | Mix engineers subgroup to achieve subgroup effect processing |
| **A2** | Mix engineers subgroup to create individual submixes |
| **A3** | Mix engineers create their subgroups based on the genre being mixed |
| **A4** | Mix engineers subgroup to make the mix process less complicated |
| **A5** | Mix engineers create subgroups within subgroups (*Hierarchical subgrouping*) |
| **A6** | Mix engineers subgroup based on instrument family |
| **A7** | Mix engineers subgroup to maintain good gain structure |
| **A8** | Mix engineers subgroup to reduce auditory masking |
| **A9** | The most common subgrouping effect to apply is dynamic range compression |

The survey consisted of 21 questions that allowed the respondent to provide both qualitative and quantitative responses. Similar to [27], I sought to probe their knowledge based on the assumptions rather than lead the respondent with them. I also tried to identify subgrouping habits and how those habits changed over time. Quantitative analysis of survey results are summarised in tables and figures throughout this chapter. Assumptions 1, 2, 5, 6, 7 and 9 came from reading audio engineering literature, discussions within my research groups, discussions with audio engineers and initial analysis of the data gathered in the previous chapter [1, 7, 17]. Assumptions 3, 4 and 8 mainly came from having lengthy discussions with my audio engineering research group about the uses of subgrouping.

### 4.2.2 Thematic Analysis

Thematic analysis [115] was used to analyse qualitative survey data. It involves familiarisation with the data and then coding sentences, paragraphs or statements from each respondent. This allows themes to be formulated and concepts or repeated ideas to be identified. The thematic analysis used here is mostly deductive, where analysis is driven by my particular analytical interest in the area. Due to the lack of subgrouping literature, I employed inductive thematic analysis, where survey responses allowed us to develop themes not directly related to the questions. I also took a latent approach to my thematic analysis [116], where the analysis goes beyond the semantic content to look for underlying ideas or thought processes. I followed the six phases of thematic analysis [115] to guide the analysis. I was unable to find this specific type of analysis applied anywhere else in audio engineering literature. However, it is a well documented and established technique for doing qualitative data analysis [115, 116].

## 4.3 Results and Analysis

### 4.3.1 Survey Questionnaire Respondent Data

The survey was provided via a web form, where respondents could complete it in their own time and come back to it later if needed. To ensure high quality answers representative of skilled practice, all ten respondents were distinguished, professional mix or mastering engineers, and had received a recognised award such as a Grammy or achieved a number one hit in the commercial music charts. The mixing background varied in terms of genre. The most common responses for genre of music mixed was Pop, Rock and Electronic music, but some were also involved in Jazz, Classical, Techno/IDM and World Music. All the respondents were male and their average age was 49.3 (SD:

8.13) years. The least amount of mixing projects a respondent was involved in a year was 5, the most was 100 and average was 40.8 (SD: 46.15).

### 4.3.2 Coding

Figure 4.1 gives an example of the manual coding applied to each respondent's answers to question one of the survey questionnaire. It illustrates how I broke down each respondents answers in too individual codes, which subsequently led to developing themes. The coding process generated 72 codes in total for all the respondents answers.

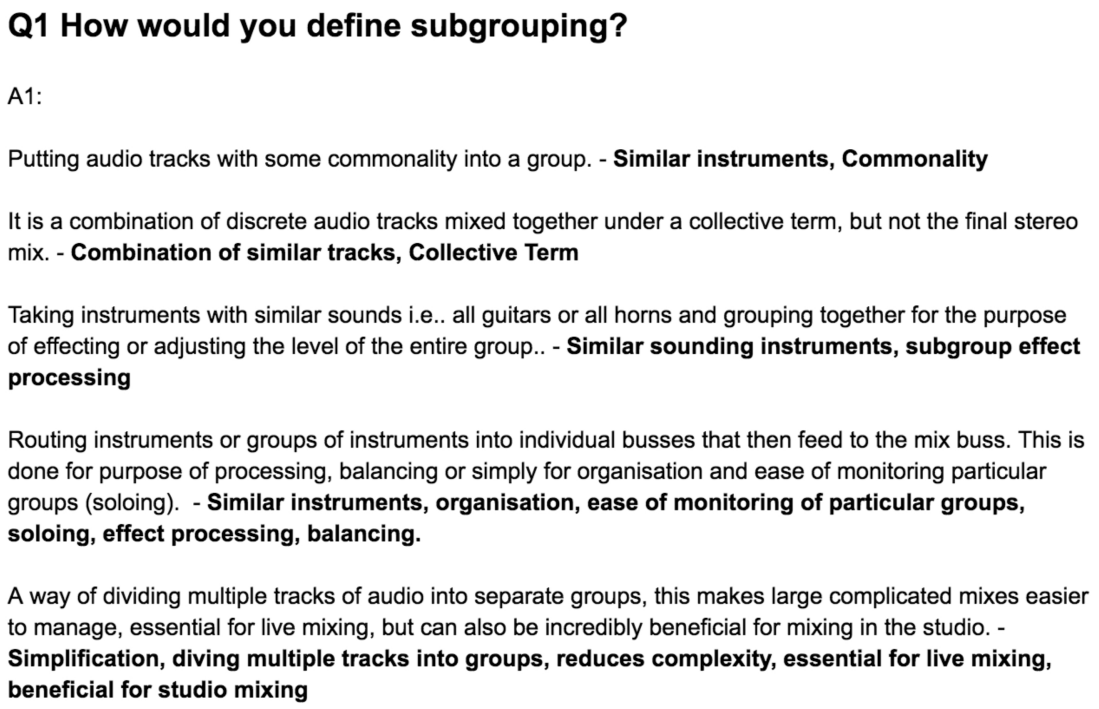


FIGURE 4.1: This is an example of coding a respondent's reply to a question. The sentence is summarised into as few words as possible.

### 4.3.3 Theme Development

Five main themes arose from the thematic analysis; *Decisions*, *Subgroup Effect Processing*, *Organisation*, *Exercising Control*, and *Analogue versus Digital*. They were developed by exporting coding details in the form of nodes and edges from QSR Nvivo[1], and visualised in Gephi[2]. Figure 4.2 illustrates one of the visualisations that were used to develop my thematic map, where each code is clustered based on the Pearson's correlation coefficient. The coding at the selected nodes is compared based on similarity of

[1]NVivo is a qualitative data analysis (QDA) software package.
[2]Gephi is an open source graph visualisation platform.

each of the coded text extracts with each other. Text extracts that have been coded similarly are clustered together on the cluster analysis diagram [117]. I used the graph in Figure 4.2 to decide what codes were related to each other and what codes had the most text references. The strength of Pearson's correlation coefficient is given in Figure 4.2 by how thick each graph edge is. Figure 4.3 shows the resultant thematic map with the main themes in red and one sub-theme in bold.

The theme Decisions arose mainly from responses to survey questions based on particular mix situations. This was the largest theme and was expected due to the types of questions I asked. It contained a *Genre* sub-theme because it became apparent from the data that many mix decisions have a genre dependency.

The Subgroup Effect Processing theme was expected since a number of survey questions were based around this theme. It was one of the largest themes and was mentioned often with respect to audio effects like EQ, DRC and to a lesser extent Reverb. In this theme I try to understand how and when subgroup effects are applied.

The Organisation theme covers what mix engineers would typically put in a subgroup, how many subgroups they would create relative to the amount of audio tracks available and why they would organise subgroups in a particular way. It is related to the themes of Decisions and Subgroup Effect Processing since a mix engineer needs to decide on how to organise a multitrack and this needs to be decided before any subgroup effect processing can be applied.

Exercising Control was not directly related to any of the questions on the questionnaire, but was foreseen. It relates to the mix engineer being able to control many audio tracks at once and simplifying the mixing process.

The final theme Analogue versus Digital, was not anticipated.Ie assembled this theme in the context of how subgrouping has changed for each respondent over a number of years. Since this was induced from the data itself I do not have an assumption related to it.

### 4.3.4 Survey response analysis and final theme analysis

Respondents were first asked how they would define subgrouping. Items mentioned included subgrouping tracks by similar instrumentation, combining tracks for subgroup effects processing and simplifying the mix process. Quotes used to define subgrouping were as follows,

*"Sub mixing different sets of audio (drums and percussion, strings, guitars etc.) in order to give them a global audio treatment, often compression and eq."*

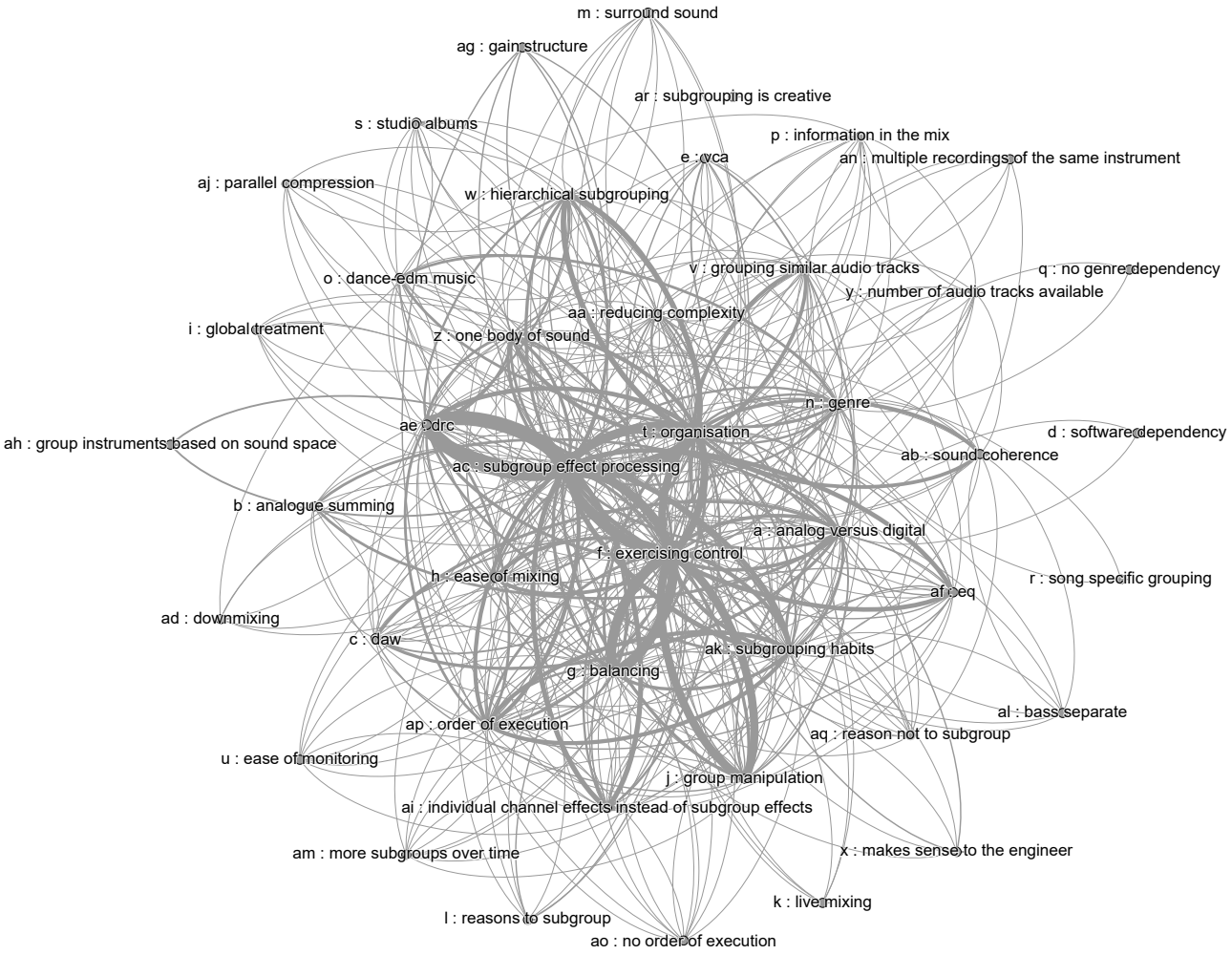


FIGURE 4.2: Codes clustered by word extract similarity.

*"Routing instruments or groups of instruments into individual busses that then feed to the mix bus... for purpose of processing, balancing or simply for organisation and ease of monitoring particular groups (soloing)."*

They were then provided with a definition of subgrouping and asked if they agreed;

*"Subgrouping can be defined as when you sum one or more audio tracks into a bus with the idea of creating a submix."*

All agreed, but some provided further alternate definitions. This implies that my proposed definition may have been too brief and did not capture all aspects of the subgrouping process.

Respondents were asked if specific reasons to subgroup applied to their workflow, depicted in Figure 4.4. Other reasons given for subgrouping included the need to create stereo stems from mono recordings, it being easier to fine-tune an instrument group

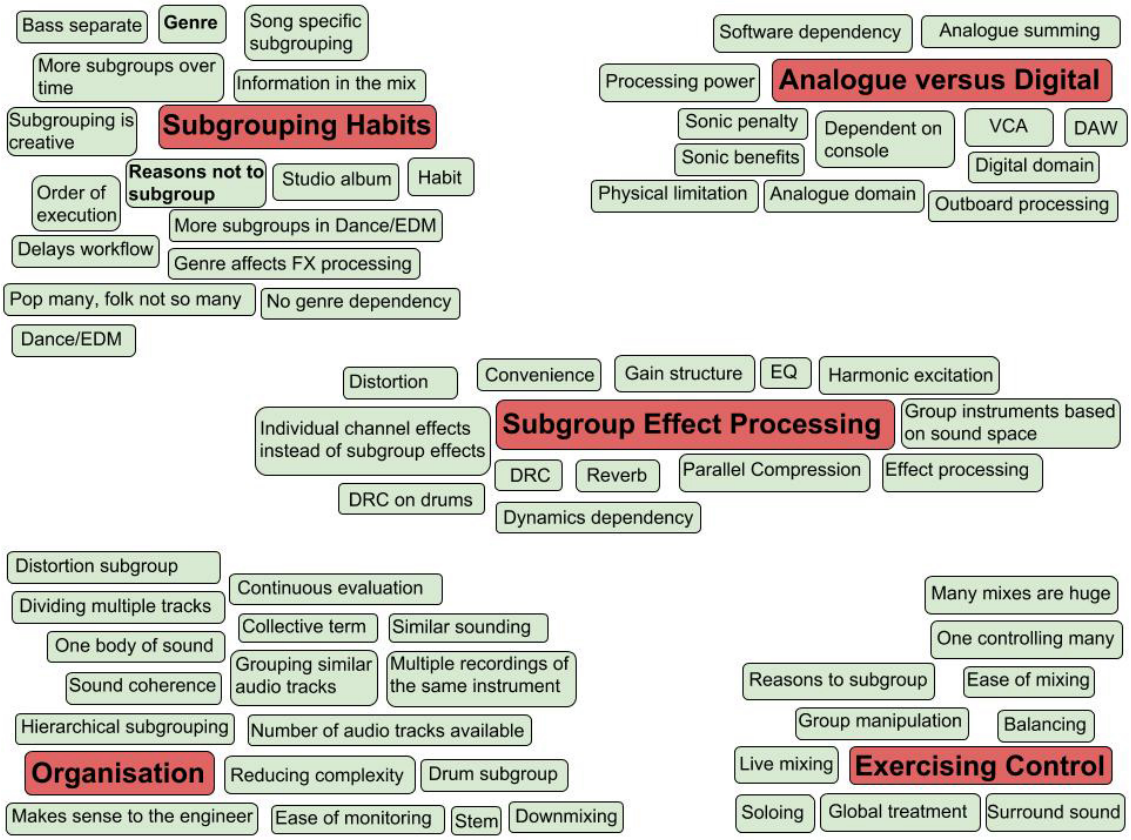


FIGURE 4.3: The thematic map. Themes are shown in red and codes are shown in green.

and combining large amounts of backing vocal tracks. Applying distortion was also mentioned and creating subgroups within subgroups (hierarchical subgrouping). One respondent stated that there should be no set rule and subgrouping should be used creatively. The respondent gave an example of how keyboardist Herbie Hancock has many subgroup routings for different types of keyboard modulation.

#### 4.3.4.1 Decisions

Decisions appeared to be the core theme as it is interlinked with all the other themes developed. Also, much of the data accumulated was based on how a mix engineer would act in certain mix situations, allowing us to determine patterns or habits typical of a professional mix engineer's workflow. Decisions was the only theme that had a sub-theme, the sub-theme being Genre.

I was interested to see at what point in the mix process the respondents normally consider putting audio tracks into subgroups. Table 4.2 summarises these results, where I used median ranking for each mix process over all respondents.

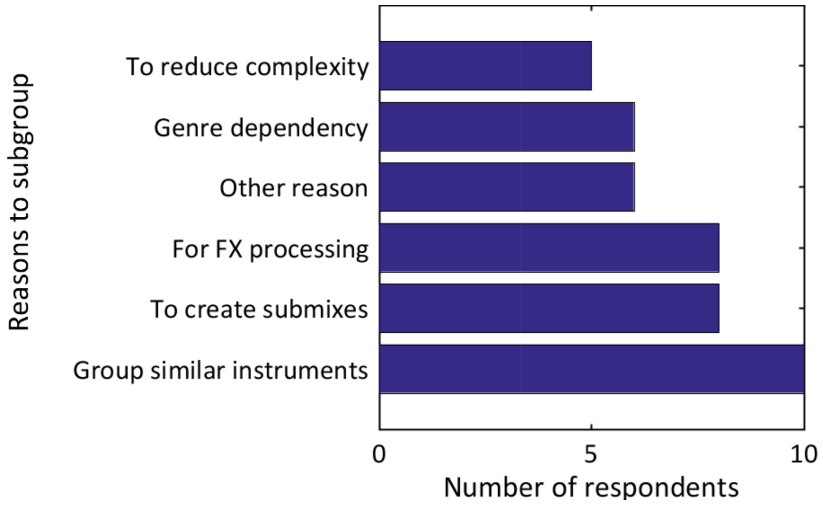


FIGURE 4.4: Respondent results based on how they subgroup.

TABLE 4.2: Rank order of execution in the mix process. This refers to question 9 in Appendix 9.3

| **Rank** | **Mix Process** |
|---|---|
| **1** | Panning |
| **2** | Subgrouping/EQ/DRC |
| **3** | Loudness/Level |
| **4** | Effects(temporal) |

Overall panning is most important, but subgrouping is considered as important as applying EQ and DRC. However, when I examined some of the statements provided in relation to this question I had a different representation;

*"Equalizing is first because I'm recording live instruments and it's important to clarify the spectral space of the recording and remove non-critical or distracting frequencies. Compression and effects further distinguish the recording. Then grouping, panning, and levelling are creative decisions most important in the final mix down, which must be made in the context of a full mix."*

*"I set level, panning, and compression/EQ on the individual tracks. Then submit usually by instrument. I apply FX to both the individual channels and the sub mixes."*

*"I progress from an organisational approach then to gain structure as primary focuses. Following that would be dynamics. Effects are 'sugar on top'. Loudness would be the last thing I would be thinking of, when the final balance is achieved... gain structure is probably the most important aspect to mixing in my opinion especially when mixing on an analog console ... The level out of the mix buss has a distinctive effect on how*

*the overall mix will sound. With digital you are more concerned with just simply not clipping."*

One respondent implies that subgrouping is creative while another suggests it is part of the organisational aspect and important for gain structure and another mentions that they subgroup by instrument type. In contrast, one respondent said there is no order of execution and that mixing is an organic process.

TABLE 4.3: The minimum, median, and maximum percentage of subgrouping decisions made by all the respondents in the last 100 mixes i.e. Respondent 1 subgrouped 10% of the last 100 mixes they did to maintain good gain structure. I present the minimum percentage for this question for all the respondents.

| **Mix Decision** | **Min %** | **Median %** | **Max %** |
|---|---|---|---|
| Subgroup to maintain good gain structure | 0 | 100 | 100 |
| Subgroup some or all of the audio tracks | 60 | 100 | 100 |
| Split drums into different subgroups | 0 | 35 | 100 |
| Change your subgroups partway through mixing | 0 | 23 | 80 |
| Subgroup to eliminate auditory masking | 0 | 5 | 100 |
| Subgroup to pan a group of instruments | 0 | 5 | 50 |

Respondents were asked to estimate how often various subgrouping related decisions were made over the last 100 mixes, see Table 4.3. "Subgrouping to maintain good gain structure" received 100% median percentage rating, which relates to Subgroup Effect Processing and will be discussed later. "Subgroup some or all of the audio tracks" indicated that there may be cases where subgrouping is not valid. However, the median percentage was 100, so this implies subgroups are used much more often than not. The median percentage for "Changed your subgroups partway through mixing" was 23. Two respondents said they would rarely change subgroups, but would further split them to create new subgroups, e.g.;

*"Goodness knows why I might change routing, but I change things all the time, it's often a refining process to achieve a better sound. I add subgroups more than change them but I might disband some that aren't working or I need more control into two separate subgroups, backing vocals being split up for example."*

The last two questions had a median of 5%. I assumed mix engineers might subgroup instruments together to reduce masking, since instruments in a subgroup often occupy the same spectral space and it would be useful to EQ all of the instruments together. However, I was surprised to see such a low score. In fact half of the participants gave a score of 0% and only one gave 100%.

Respondents were asked yes/no questions to decisions the mix engineer might make when mixing, summarised in Table 4.4. These types of questions were mainly related to instrument choices, especially drums and guitars. The two most polarising questions are related to auditory masking and to acoustic and lead guitar placement. The results to the auditory masking question tend to agree with the result in Table 4.3. Each of these questions was followed by 'can you please tell us why,' so that they could provide qualitative feedback. I did not test the knowledge of any of respondents with respect to masking. I assumed that since they were at such an advanced level in the field of mixing, they would already be quite knowledgeable in this area.

There was only one genre related question, but other questions generated genre related answers. Respondents noted genre-dependency in subgrouping, for instance;

*"I might submit 'strings' for a rock track, but for an orchestra I'll break this down into 'violins' and 'cellos'."*

One respondent mentions that some subgroups receive different effect processing based on genre, in particular DRC. Also, certain styles require effect processing using subgroup processing, while others benefit from a global treatment. A respondent noted that a guitar subgroup for reggae would be treated differently than in rock music. Other statements include;

*"The more compression required, the more subgroups necessary."*

*"Many genres of music need subgroups, it's not the genre, but the amount of information in the mix."*

The need for more subgroups when more compression is required indicates that there could be more need for gain staging, so as to correctly process the varying amounts of dynamic range. This suggests that a reason for creating subgroups is to reduce complexity.

TABLE 4.4: Answers to simple yes/no questions from online survey questionnaire

| **Mix Decision** | **Yes** | **No** |
|---|---|---|
| Do you create subgroups with subgroups (Hierarchical) | 6 | 4 |
| Subgroup kick drum separately | 4 | 6 |
| Subgroup snare drum separately | 3 | 7 |
| Subgroup bass guitar played percussively with percussion/-drums | 3 | 7 |
| Put rhythm guitar and lead guitar in the same subgroup | 6 | 4 |
| Put bass guitar and lead guitar in the same subgroup | 2 | 8 |
| Place acoustic guitar and lead guitar in the same subgroup | 8 | 2 |
| Subgroup to achieve a uniform tone | 6 | 4 |
| Subgroup to reduce auditory masking | 2 | 8 |

Dance and EDM music was mentioned separately by two different respondents. One statement being

*"Dance or EDM as a particular genre uses a vastly greater number of effect 'tricks' hence sub grouping with this genre is generally more focused on this as opposed to most other genres in which I am just concentrating on organisation and dynamics."*

An example relating the quantity of subgroups to genre is illustrated in the following ambiguous statement,

*"Pop=lots, folk=not so many."*

The respondent mentions that there are many subgroups when mixing 'Pop' music, but this could mean that there are more instruments to subgroup or that 'Pop' needs more subgroup processing.

Genre appears to be a significant deciding factor on how subgrouping is applied. However, at least two respondents claim that genre has no impact on their subgrouping

decision. One respondent stated that genre does not have much influence on their subgrouping decisions and is always song specific or depends on the information in the mix.

#### 4.3.4.2 Subgroup Effect Processing

Subgroup effect processing is where at all of the tracks in the group benefit from similar processing. This theme was formulated because the topic of effect processing was mentioned the most in responses (130 code references associated with this theme). It was also a major theme when I visualised the relationship between the coded references seen in Figure 4.2. The types of subgroup effect processing that respondents used is summarised in Figure 4.5. All respondents would apply DRC.

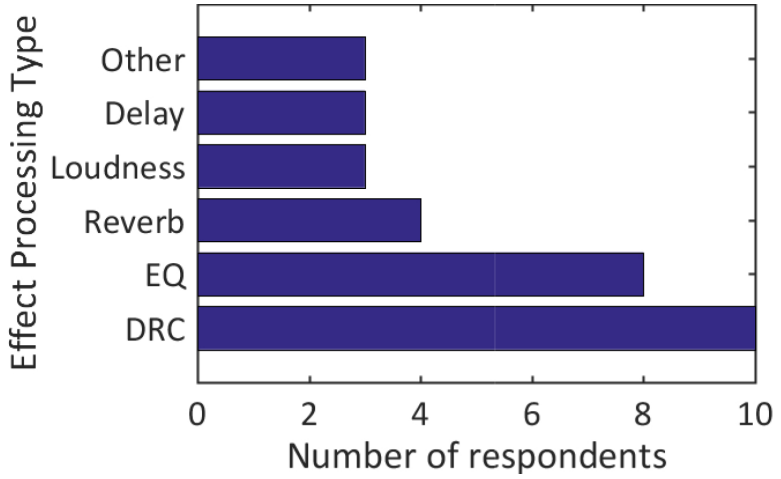


FIGURE 4.5: Summary of the different types of FX processing that each respondent would apply to a subgroup. This refers to question 3 in the questionnaire.

Other types of audio processing that were mentioned were enhanced stereo imaging, doubling, harmonic excitation, distortion and parallel compression. A statement from one respondent illustrating subgroup effect processing referred to the 'body of sound', which could be interpreted as a group of similar instrumentation;

*"Subgrouping drums, vocals, guitars etc. enables you to apply overall compression and FX so the body of sound can be treated as one, FX could be anything from as simple as reverb or more complicated like adding parallel compression."*

I asked respondents how likely they were to apply DRC to certain instrument subgroups, see Figure 4.6. Statements related to this question include

*"Elements such as drums, percussion and bass, need the most dynamic range compression because they create the groove. Legato instruments such as brass, pads or vocals are not as closely tied to the groove so they should be more free."*

*"I pretty much always use some form of compression on drums, lead vocals and bass, source, get rout and parallel compression."*

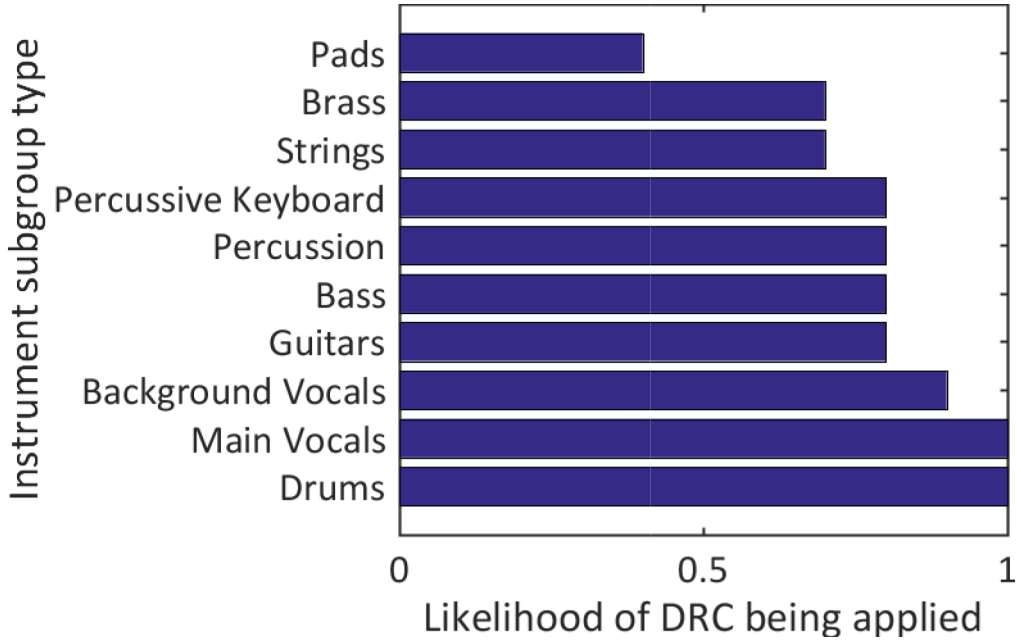


FIGURE 4.6: The subgroup types that are most likely to have DRC applied. This refers to question 8 in Appendix 9.3.

In [113], the most subgrouped and hierarchically subgrouped instrument types were drums and vocals. Many comments supported the view that drums, vocals and bass get more DRC because they have the greatest dynamic range and are the foundational part of a mix [17]. An interesting comment was

*"Drums and Vocals... always get a touch of compression in my mixes, even if it's one or two dB, this helps the master bus compressor focus on the overall mix and not be triggered by a subgroup."*

The rationale that some instruments may need to be removed from a subgroup because they adversely affect group compression leads to "Do you subgroup kick drum separately?" from Table 4.4. One stated that they would do so in order to compress it. Another related statement was that since it is such a powerful instrument it would affect compression on the other drums in the subgroup and need to be processed separately.

I asked "Do you subgroup instruments to achieve a uniform tone through EQ?" One respondent stated they would use subgroup EQ processing all the time, but not for

uniform tone. Two responses mentioned that they do it since it is easier on CPU, but this slightly contradicts earlier points about trying to treat a particular instrument type. Others noted convenience in achieving uniform tone, and the ability to make instruments sound like they are in the same room, which was the only response that discussed reverb in respect to subgroup effect was processing;

*"Primarily for convenience. I'm fascinated with gluing sounds together whether that's by creating a virtual soundstage or something more abstract."*

*"It can be handy to group the bass and drums when using ambience or reverb effects to make all instruments seem like they are in the same space/room."*

In Table 4.3 the "Do you subgroup to maintain good gain structure?" question had a median percentage score of 100;

*"...I have to note that gain structure is probably the most important aspect to mixing in my opinion especially when mixing on an analog console such as an SSL or Neve ..."*

*"Affects how subgroups get treated - some genres benefit from subgroup dynamic compression. Others just from the gain structure advantages."*

The second statement was in relation to genre and the respondent highlighted advantages of subgrouping to achieve good gain structure since it allows gain processing to be applied in a step by step instrument group process.

All respondents put strong emphasis on subgroup effect processing, but some referred to effect processing on individual tracks instead of subgroup effect processing. This mostly related to EQ, where a respondent might sculpt the sound of each instrument individually to reduce masking. In most cases this was in reference to guitars as in [113], where they were treated individually because they served different roles in the song e.g. distorted guitar, lead guitar.

#### 4.3.4.3 Organisation

Organisation directly relates to Exercising Control and Subgroup Effect Processing, since they cannot happen without first organising tracks in to sensible subgroups. It also relates to Decision, since the mix engineer has to decide how to organise their subgroups. Relevant statements include;

*"Putting audio tracks with some commonality into a group."*

*"Routing instruments or groups of instruments into individual busses that then feed to the mix buss. This is done for purpose of processing, balancing or simply for organisation and ease of monitoring particular groups (soloing)."*

*"It is a combination of discrete audio tracks mixed together under a collective term, but not the final stereo mix."*

The word organisation was only mentioned once above, but other words and phrases like 'commonality' and 'collective term' are organisational.

In Table 4.3, when I asked how often respondents split drums into different drum subgroups i.e. hierarchically subgroup, the median percentage was 35%. I previously found that when hierarchical subgrouping did occur, 12% of drum subgroups created were hierarchical [113]. When asked "did you modify the subgroups you had already created?" two respondents said they would rarely change subgroups, but would further split them to create new subgroups, an example of hierarchical subgrouping.

Two questions related to how many subgroups respondents used based on how many tracks were in a multitrack, and how many tracks were needed before they considered subgrouping. The minimum, average and maximum amount of subgroups the respondents would normally create in relation to the number of audio tracks can be seen in Figure4.7.

*"First if the subgrouping makes sense internally, and second if the group works in the context of a mix."*

One respondent would subgroup all guitars together simply for organisational purposes.

*"Due to the physical limitations of an analog console... subgroup all the guitars anyway simply for organisational purposes. Any processing would be done individually."*

#### 4.3.4.4 Exercising Control

Exercising control and the simplification of the mixing task was an important theme in the data. By exercising control I mean that by subgrouping many audio tracks, the tracks can be collectively manipulated in terms of level and effect processing using a single fader or dial without losing control. Two definitions given by respondents on subgrouping are as follows,

*"... dividing multiple tracks of audio into separate groups, this makes large complicated mixes easier to manage, essential for live mixing, ... incredibly beneficial for mixing in the studio."*

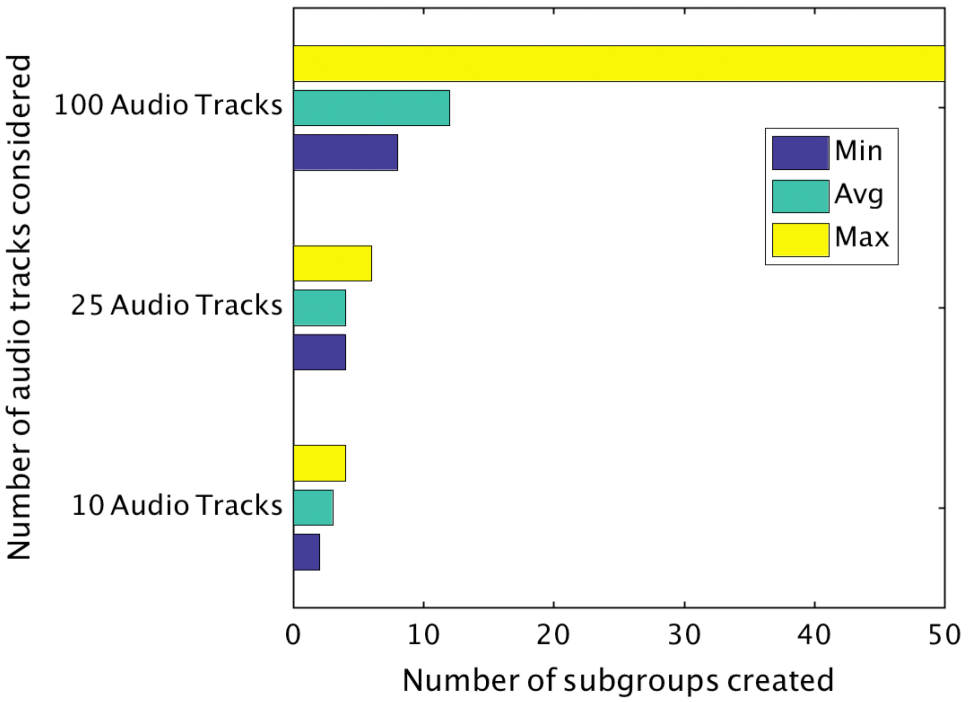


FIGURE 4.7: This shows the averaged results of all the respondents. I asked them to indicate the minimum (blue), average (green) and maximum (yellow) number of subgroups respondents create based on a given amount of audio tracks. This is in reference to question 6 in Appendix 9.3.

*"Whenever one controller is used to control more than one. Most commonly, it is fader grouping, and these take two forms: 1. Control groups (one fader controls other faders) and 2. Processing Groups (signals are combined into an ensemble/stem/group . . . )"*

One respondent mentions it being essential for live mixing, which is understandable considering that it simplifies continually adjusting gain levels. Another respondent refers to the subgroup as a control group and implies that subgrouping is used for control. Overall, only 50% of the respondents said they use subgrouping to simplify the process, but on further analysis, the terms control and simplification were mentioned throughout the responses. One statement summarising this was;

*"More complexity, more subgroups"*

Other statements say that the more tracks there are, the more subgroups needed to keep the mixing task as simple as possible while maintaining a good degree of control,

*"It makes it much easier to monitor groups of instruments or instruments that contain multiple sources (such as drums). One could describe this also as 'soloing'."*

*"When a mix gets beyond 24 tracks... it makes it easier to fine tune the overall mix if a group of instruments needs to be adjusted. Many mixes are 96 faders of information or more."*

*"Large track counts, e.g. 100+, subgrouping will be essential to retain control."*

*"If they're all too bright, it's easier and more DSP friendly to do them all at once."*

Exercising control was also mentioned in a surround sound context where the respondent states

*"Surround might need control over the centre channel, you may have many kick drums you want to compress as a whole etc. you may want to remove the kick drum from the main drum compression so it stops affecting the other drums via the compressor."*

#### 4.3.4.5 Analogue versus Digital

The modern Digital Audio Workstation (DAW) has revolutionised how mix engineers approach mixing, since they now rarely worry about physical limitations. The Analogue versus Digital theme became clear once I attempted to understand if subgrouping practice has changed over the last five years. Some respondents said that because of the modern DAW they now use more subgroups since they no longer have the physical limitation of an analogue desk and the amount of available subgroups is almost limitless. The Analogue versus Digital theme was not something I attempted to test with my assumptions. It was developed through thematic analysis and was not something I anticipated.

*"A big change... during the transition from all analog mixing to mixing in the box. Generally these were physical limitations due to the console... virtually unlimited subgrouping in DAWs."*

*"Subgrouping approach has only changed via computing power has grown, as I mix mostly within a DAW, the more power I have the more I can expand my mixer..."*

*"If I were using an analogue desk with only 8 groups, then maybe, but these days I don't need to constrain myself in that way."*

There were two statements made that illustrate why a mix engineer might not have used as many subgroups before they had access to a DAW.

*"In the analogue domain, I may not do this because the subgrouping requires an additional pass through a summing amp, which – depending on the console – might pay a sonic penalty."*

*"20 string mics are still one instrument and it is useful to be able to treat it as such. Pre DAW, these items would all have been bounced to stereo as part of the recording process."*

Finally, there was a statement that summarises what is meant by the theme of Analogue versus Digital in a subgrouping context.

*"Subgroup processing is part of the current sonic environment."*

## 4.4 Discussion and Assumptions

**(A1)** - eight out of the ten respondents agreed with this statement and subgroup effect processing was a major theme in this report. Also, statements were given that subgroup effect processing, especially DRC and EQ, is essential and is heavily used. DRC was referenced 31 times throughout the survey responses and EQ was referenced ten times. I therefore consider this assumption to be true.

**(A2)** - eight out of ten respondents agreed with this statement. This assumption was touched on under the themes of Organisation and Exercising Control where respondents mentioned putting similar instruments into the same subgroup in order to mix them as one. An example of this is when the mix engineer attempts to mix drums or is making a stem track. Based on the fact it is an obvious reason to create subgroups and so many respondents agreed, I would consider this assumption to be true.

**(A3)** - six out of ten respondents said that genre has an effect on how they create subgroups. There were many examples given by the respondents on when this would occur, particularly for EDM/Dance music. However, some respondents said it does not affect their subgroup choices and one respondent said it depends on the information in the mix. Based on the many examples given by the respondents on when genre affects subgroup choices and the overall majority of respondents agreeing with this assumption, I consider this assumption to be true.

**(A4)** - five out ten respondents said they create subgroups to reduce complexity. However, if I examine Figure 4.7 I see a trend where the more audio tracks there are, the more subgroups there are. This suggests that mix engineers create subgroups to reduce the amount of faders and effects they have to manage. Therefore, reducing complexity. There were many statements provided that fell under the themes of Organisation and Exercising Control that suggested that subgroups are created to make the mix engineers life easier. Despite that only half respondents agree with this statement, the volume of

qualitative data suggests otherwise. Therefore, I would consider this assumption to be true.

**(A5)** - six out of ten respondents said they hierarchically subgroup. The median percentage for respondents who split the drum subgroup up into smaller subgroups in their last 100 mixes was 35%. I also found this occurred in previous work mainly with respect to drums and vocals [113]. In relation to Table 4.3, two respondents both similarly said they would rarely change subgroups, but they would further split them and create new subgroups which is the same as hierarchical subgrouping. Based on these results I would consider this assumption to be true.

**(A6)** - All respondents agreed with this assumption. It was also found to be true in previous work [113]. The idea of subgrouping based on instrument family also came up under the themes of Decisions and Organisation. It could be argued that this was an obvious assumption. However, I have never seen it explicitly stated anywhere in the literature as a rule [1, 7, 17]. Consequently, I consider this assumption to be true.

**(A7)** - The median percentage for respondents who answered the question "in the last 100 mixes did you subgroup to maintain good gain structure" was 100%. One respondent mentioned this to be one of the most important aspects of mixing. They said that they would initially use subgrouping for organisational purposes and then for maintaining good gain structure. I consider this assumption to be true.

**(A8)** - The median percentage of respondents who answered the related question in Table 4.3 was 5%. Furthermore, when respondents answered in a simple yes or no context, only two out of ten respondents said yes. This is not a result I expected as I know that masking reduction is important to mix engineers and by treating instruments that share a similar spectral space together this would make masking reduction easier to achieve. Based on the results found, I consider this assumption to be false.

**(A9)** - All of the respondents said they would apply DRC to their subgroups. Furthermore, I also asked what instrument groups each respondent is most likely to apply DRC to and found this to be drums and vocals. These results agree with the findings in [12], where the authors tested the assumption "Gentle bus/mix compression helps blend things better" and it was found to be correct. I believe this assumption to be true.

## 4.5 Conclusion

From the analysis and discussion presented here, it is clear that subgrouping is not as simple as subgrouping all instruments that are similar to each other. There is more

of a thought process behind subgrouping and a number of different factors come into play when subgrouping decisions need to be made. For instance, genre has an impact on the type of subgrouping strategy used. It determines if and how subgroups should be broken down, what type of effects processing is to be used, what instrumentation subgroups contain, and how many tracks there will be in a subgroup.

The data gathered through the survey validates the majority of the assumptions that were made previously with regard to subgroup processing and organisation. It also uncovered underlying information that would otherwise be passed on from practitioner to practitioner, or learned through trial and error, but that would remain undocumented.

Many of the findings in this survey are of no surprise, such as subgrouping by instrument type, subgrouping for effects processing and subgrouping to make the mixing processing less complicated. However, these are often just stated in the literature as a reason to subgroup without discussion as to why [1, 7, 17]. Since many of these assumptions were obvious, but not clearly stated, the need was felt to clarify and test them as rigorously as qualitative analysis will allow. This explain why so many of the assumptions were found to be true. Although this process may not have been robust, this is the difficulty of working with qualitative data. Furthermore, results from the survey tend to agree with the results uncovered in chapter 3. However, it is worth mentioning that the participants in the previous study were students and not professional level mix/mastering engineers.

Considering these results in an intelligent audio production tool context, they indicate that subgrouping should be considered in developing these types of systems [2, 3, 15, 18]. If professionals perform subgrouping when mixing, then systems trying to mimic similar results may also benefit from this. I thus make the following seven recommendations for any intelligent mixing system that were to consider using subgrouping;

1. Subgrouping should be applied when there is more than one of any instrument type and should be applied to instruments that are similar to each other i.e. subgroup drums or guitars.

2. Subgrouping should be applied to maintain a good gain structure.

3. Based on the rankings in Table 4.2 I suggest that subgrouping be applied after panning and before DRC or EQ is applied. The reason for it being applied before DRC or EQ is because DRC or EQ will then be applied to each subgroup as well as individual channels.

4. Subgroups should be created based on the genre of the music being mixed. Genre should inform the types of effect processing applied to subgroups.

5. If hierarchical subgrouping is to be used, this should be applied to drums, vocals and guitars.

6. DRC subgroup processing should always be applied to drum and vocal subgroups and to a lesser extent EQ should be applied to all subgroups.

7. The number of subgroups should be created in proportion to the amount of audio tracks available as well as the genre of music being mixed in order to reduce complexity.

These recommendations are based on the analysis of 72 student mixes in chapter 3 and the detailed survey of ten award-winning professional mix and mastering engineers herein. They are by no means exhaustive, but it is hoped that they will be utilised and validated further in an automatic mixing system.

# Chapter 5

# Automatic subgrouping of multitrack audio

## 5.1 Introduction

In the literature reviewed, there is currently no proposals or discussions of a system that attempts to automate the subgrouping process [2, 3, 15, 118]. In this chapter, I suggest that this can be done autonomously using machine learning techniques. The motivation is two-fold. Firstly, not only would it be possible to subgroup the audio tracks in the conventional sense, but through analysis of each audio track's spectro-temporal audio features, I may discover in this study that there are more intelligent ways to create subgroups.

Secondly, the audio features that are determined to be important can be used to answer the research question are we putting the instruments in the correct subgroups? Whereby, if we have good audio features to determine subgroups, this may inform us that a certain audio track or even certain sections of an audio track should be subgrouped differently from how they would be typically subgrouped. An example of how this may work would be when we find over time that an audio track changes and may become more similar to another audio track in another subgroup. This could occur if the bass player suddenly switched from picking the bass guitar to playing in the style of slap bass. The audio track that was once in the bass subgroup could now be subgrouped with the percussive instrument audio tracks. At this point, it would make sense to split the single bass guitar audio track into two individual audio tracks and have them designated to their appropriate subgroups.

In light of the above discussion, the subgroup classification problem can been seen as somewhat similar to musical instrument identification, which has been done before for orchestral style instruments [119–122]. However, in subgrouping classification we are not trying to classify traditional instrument families, but defined groups of instrumentation that would be used for the mixing of audio from a specific genre. For example, in rock music the drum subgroup would consist of hi-hats, kicks and snares etc. while the percussion subgroup may contain tambourines, shakers and bongos. In practice, the genre of the music will dictate the type of instrumentation being used, the style in which the instrumentation will be played and what subgroup the instrument belongs to. It is also worth noting that typical subgroups such as vocals or guitars can be further broken down into smaller subgroups. In the case of vocals the two smaller subgroups might be lead vocals and background vocals. Furthermore, we can never assume that the multitrack recordings being used are good quality recordings. They may contain background noise, microphone bleed interference or other recording artefacts. All of these factors can affect the accuracy of a classification algorithm.

The purpose of this study is to determine the best set of audio features that can be extracted from multitrack audio in order to perform automatic subgrouping. In my particular case, I looked at multitracks that would be considered as Rock, Pop, Indie, Blues, Reggae, Metal and Punk genres, where the subgroups would typically be drums, bass, guitars, vocals etc. The rest of the chapter is organised as follows. Section 5.2 describes the dataset used for feature selection and testing. Section 5.3 provides a list of features used and describes how they were extracted. Section 5.4 explains how the experiments, classification and clustering were performed. Section 5.5 presents the results obtained. Section 5.6 discusses the results and then finally the chapter is concluded in section 5.7.

## 5.2 Dataset

The amount of data available for multitrack research is limited due to a multitrack being an important asset of a record label and the copyright issues that come with distributing them. The Open Multitrack Testbed contains multitrack audio, mixes of multitrack audio and corresponding metadata. I used this for my dataset because it is one of the largest of its kind (1.3 TB in size) and contained data that was available for public use [13]. A subset of data was selected from this.

The subset used for feature selection consists of 54 separate multitracks and 1467 audio tracks in total once all duplicate audio tracks were removed. The multitracks that were used span a wide variety of musical genres such as Pop, Rock, Blues, Indie, Reggae, Metal, and Punk. I annotated each track by referring to its filename and then listening

TABLE 5.1: *Details of the subset used for feature selection*

| Subgroup type | No. of tracks | Percentage of subset |
|---|---|---|
| Drums | 436 | 29.72% |
| Guitars | 365 | 24.88% |
| Vocals | 363 | 24.74% |
| Keys | 103 | 7.02% |
| Bass | 93 | 6.34% |
| Percussion | 80 | 5.45% |
| Strings | 19 | 1.30% |
| Brass | 8 | 0.55% |

to each file for a brief moment to confirm its instrument type. The labels used for each audio file were based on commonly used subgroup instrument types. These were drums, vocals, guitars, keys, bass, percussion, strings and brass. Table 5.1 shows the breakdown of all the multitrack data used for feature selection relative to what subgroup each audio track would normally belong to. It is worth noting the imbalance of label types in my dataset. This is because the most common instruments in my multitrack dataset are drums, vocals and guitars. Furthermore, the drum subgroup consists of many different types of drums such as kicks, snares, hi-hats etc. meaning it tends be the largest subgroup.

The subset used to test if the selected features were useful or not consists of five unseen multitracks. The breakdown of the different types of audio tracks for each test multitrack can be seen in Table 5.2.

TABLE 5.2: *Details of the subset used for testing*

| Subgroup type | MT 1 | MT 2 | MT 3 | MT 4 | MT 5 |
|---|---|---|---|---|---|
| Drums | 11 | 8 | 9 | 10 | 1 |
| Vocals | 17 | 11 | 6 | 9 | 3 |
| Guitars | 12 | 2 | 6 | 2 | 0 |
| Keys | 1 | 4 | 2 | 4 | 3 |
| Bass | 1 | 1 | 1 | 1 | 1 |
| Percussion | 1 | 0 | 1 | 0 | 0 |
| Strings | 0 | 0 | 0 | 0 | 6 |
| Brass | 0 | 0 | 0 | 0 | 0 |

## 5.3 Extracted Features

Each audio track in the dataset was downsampled to 22050 Hz and summed to mono using batch audio resampling software. The audio features were then extracted from the 30 secs of audio with the highest amount of total energy in each audio track [123]. This was done to speed up the feature extraction process as I did not see the need to extract features from long periods of silence that occur in multitrack recordings. 159 continuous low level audio features were extracted in total with a hamming window size of 1024 samples and a hop size of 512 samples. These window and hop size values were chosen as they were the most commonly used in all the literature I reviewed. A list of the audio features and the relevant references are in Table 5.3. Overall, there are 42 different low level audio feature types and the majority of these are frame based. Only three audio features were whole audio track features and not frame based. Since the whole track audio features were not frames like the others, no pooling was required. Pooling is a technique used in music information retrieval that allows for the summary of audio features over specific time frames i.e. the mean spectral centroid over 10 secs [124]. The mean, standard deviation, maximum and minimum values were taken of each framed audio feature over the 30 secs of audio used for feature extraction. This allowed the pooling of the framed features over the 30 secs of audio and is the reason why there was 159 audio features in total [124].

## 5.4 Experiment

Two experiments were conducted. The first experiment determined a reduced set of audio features from the 159 audio features that I extracted previously. This was done by performing feature selection. The goal of this experiment was to determine the best subset of the 159 original audio features that could be used for automatic subgrouping. A second experiment was conducted where five test multitracks were agglomeratively clustered using all of the 159 audio features extracted and then agglomeratively clustered using the reduced feature set for comparison. This was done to investigate how well the reduced audio feature set compared to the entire audio feature set when performing automatic subgrouping.

### 5.4.1 Feature Selection

Random Forest is a particular type of Ensemble Learning method based on growing decision trees. This can be used for either classification or regression problems, but can

TABLE 5.3: *Audio features*

| Category | Feature | Pooled | Reference |
|---|---|---|---|
| Dynamic | RMS | Y | |
| | Peak Amplitude | Y | |
| | Crest Factor | Y | |
| | Periodicity | N | [125] |
| | Entropy of Energy | N | [126] |
| | Low Energy | N | [127] |
| Spectral | Zero Crossing Rate | Y | [128] |
| | Centroid | Y | . |
| | Spread | Y | . |
| | Skewness | Y | . |
| | Kurtosis | Y | . |
| | Brightness | Y | . |
| | Flatness | Y | . |
| | Roll-Off (.85 and .95) | Y | . |
| | Entropy | Y | . |
| | Flux | Y | . |
| | MFCC's 1-12 | Y | . |
| | Delta-MFCC's 1-12 | Y | [128] |
| | Spectral Crest Factor | Y | [123] |

also be used for feature selection. Random Forest is based on the idea of bootstrap aggregating or more commonly know as bagging. After training has occurred on a dataset each decision tree that is grown predicts an outcome. For regression decision trees, the output is the average value predicted by all of the decision trees grown. For classification decision trees it is the classification outcome that was voted most popular by all of the decision trees grown [105]. Random Forest was chosen because it has been proven to work very well for feature selection in other fields such as bioinformatics and medicine [106, 107].

Determining the most salient features using the Random Forest classifier was performed as follows. 100 decision trees were grown arbitrarily and a feature importance index was calculated. It will be seen further on in Section 5.5 that this was an appropriate amount of decision trees to grow.

The feature importance index was calculated for each of the 159 audio features. The average feature index was then calculated and the audio features that performed below the average were eliminated. The use of the average importance index was found to give us the most satisfactory set of audio features.

I also tried eliminating the 20% worst performing audio features, then retraining on the new audio feature set and repeating the 20% worst performing audio feature elimination

process. This process would then stop once the out-of-bag error began to rise. However, I found that this was found to give us an unsatisfactory set of audio features. They were unsatisfactory because when I used these audio features to automatically create subgroups, the subgroups created were mostly incorrect e.g. drums in the same subgroup as guitars. This was the search method that was used in [108].

It should also be noted that when using the Random Forest classifier I set prior probabilities for each class based on my imbalanced dataset. The prior probabilities were set using the data in the *Percentage of subset* column in Table 5.1

### 5.4.2 Agglomerative clustering

In my case the similarity is found between every pair of audio feature vectors that represent the audio tracks in my dataset. This is normally calculated using a distance function such as Euclidean, Manhattan or Mahalanobis distance. I used Euclidean distance as I found it gave me more realistic clusters. It is also worth noting that I normalised each instance in my dataset using L2-normalisation, while each audio feature value was normalised between zero and one. This was done due to the Euclidean distance function being used. I then linked together audio feature vectors into binary pairs that were in close proximity to each other using a linkage function. I used the shortest distance measure as my linkage function, as this would make the most sense in my case as I am trying to subgroup similar audio tracks based on instrumentation. The newly formed clusters created through the linkage function were then used to create even larger clusters with other audio feature vectors. Once linkage has occurred between all the audio feature vector clusters, all the branches of the tree below a specified cut-off are pruned. This cut-off can be specified as an arbitrary height in the tree or else the maximum amount of clusters to create. A maximum number of eight clusters was specified in my case. This was due to there only being eight labels in the original dataset used for feature selection.

Figure 5.3 depicts that any two audio tracks in the dataset become linked together at some level of the dendrogram. The height of the link is known as the cophenetic distance and represents the distance between the two clusters that contain those two audio tracks. If the agglomerative clustering is suited to a dataset, the linking of audio tracks in the dendrogram should have a strong correlation with the distances between audio tracks generated by the distance function. A cophenetic correlation coefficient can be calculated to measure this relationship. The cophenetic correlation coefficient is measured from -1 to 1 and the closer the value is to 1 the more accurately the dendrogram reflects the dataset. Suppose that the previous example dataset $N_i$ has been modelled

using the above cluster method to produce a dendrogram $T_i$. The cophenetic correlation coefficient is calculated as such

$$c = \frac{\sum_{i<j}(d(i,j) - \bar{d})(t(i,j) - \bar{t})}{\sqrt{\left[\sum_{i<j}(d(i,j) - \bar{d})^2\right]\left[\sum_{i<j}(t(i,j) - \bar{t})^2\right]}} \tag{5.1}$$

where $d(i,j)$ is the ordinary Euclidean distance between the $i$th and $j$th observations of the dataset and $t(i,j)$ is the cophenetic distance between the dendrogram points $T_i$ and $T_j$. $\bar{d}$ is the average of the $d(i,j)$ and $\bar{t}$ is the average of the $t(i,j)$.

## 5.5 Results

In this section I present the results of the experiments conducted. I firstly show the results of the feature selection performed and then show the results of the agglomerative clustering. I also present the resulting dendrograms from the clustering.

### 5.5.1 Selected Features

Using the feature selection method mentioned in Section 5.4.1 I determined a subset of 74 audio features from the original 159. The average feature importance index was 0.421 with a standard deviation of 0.1569. The maximum value for feature importance index was 0.9086 and the minimum was -0.0135. The 20 most important features are depicted in Figure 5.1. This illustrates some of the audio features that would occur in an audio feature vector used during agglomerative clustering.

The cumulative out-of-bag error having grown 100 trees with the full audio feature set was 0.1384. Using the reduced feature set and growing 100 trees the cumulative out-of-bag error was 0.1431. Figure 5.2 shows that these results converge and start becoming very close after about 70 trees. This also supports my original choice to arbitrarily grow 100 decision trees for feature selection.

### 5.5.2 Agglomerative clustering

In Table 5.4 I present the results for each of the five multitracks that were agglomeratively clustered using the entire audio feature set and the reduced audio feature set. Also, I give the cophenetic correlation coefficients as described in Section 5.4.2. Also, I give the number of audio tracks in each multitrack as well as how many incorrect subgroups were

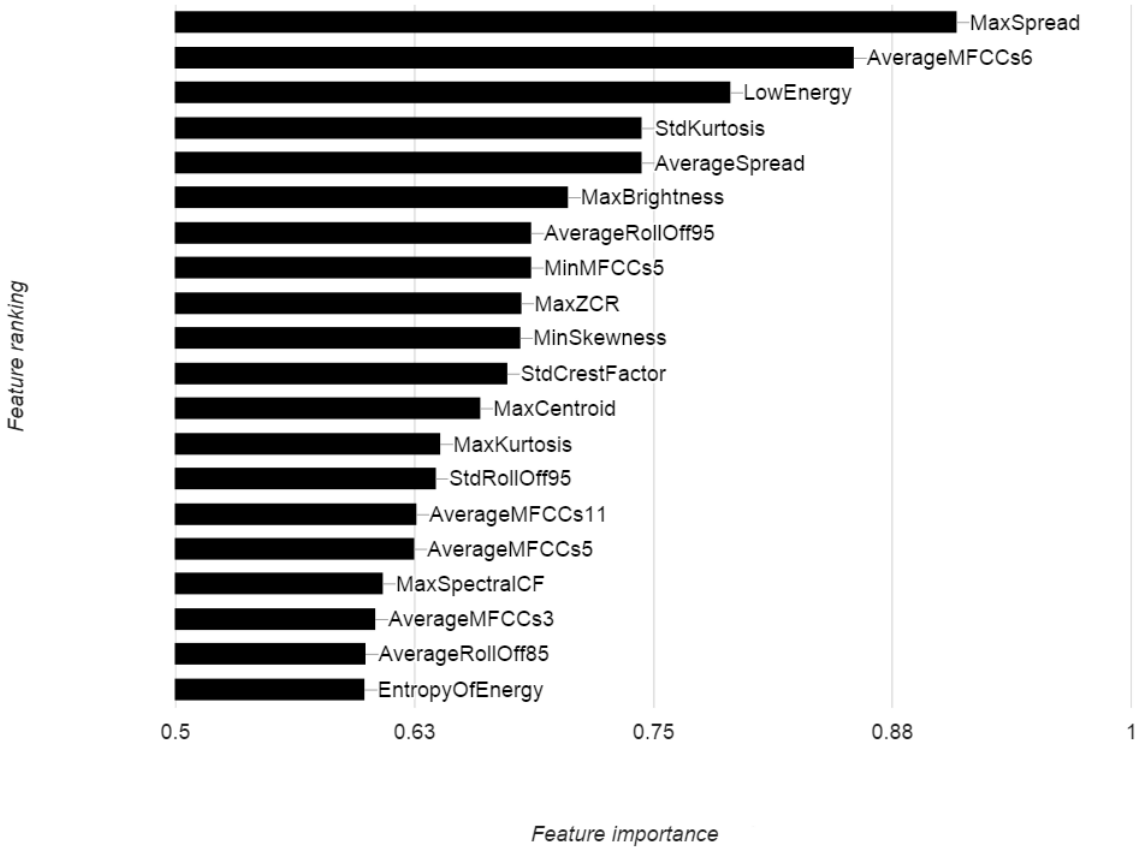


FIGURE 5.1: The 20 most important features

created to show how well the clustering is at creating meaningful subgroups. An incorrect subgroup would be where at least two different audio tracks with different instrument types are subgrouped together. An example of an incorrect subgroup would be if a subgroup consisted of drums, guitars and vocals. These three instrument types would normally be separate. There will always be eight subgroups due to the labels used in the training dataset, but these eight subgroups may not always be constructed correctly using agglomerative clustering. The number of incorrect audio tracks is measured by how many audio tracks were placed in a cluster where the majority of the instrument types where incompatible. An example being if I had a cluster of six guitars and two vocals. The guitars are the majority, so the incorrect audio tracks would be the vocal tracks. I also show this measure as a percentage of all the audio tracks in each multitrack.

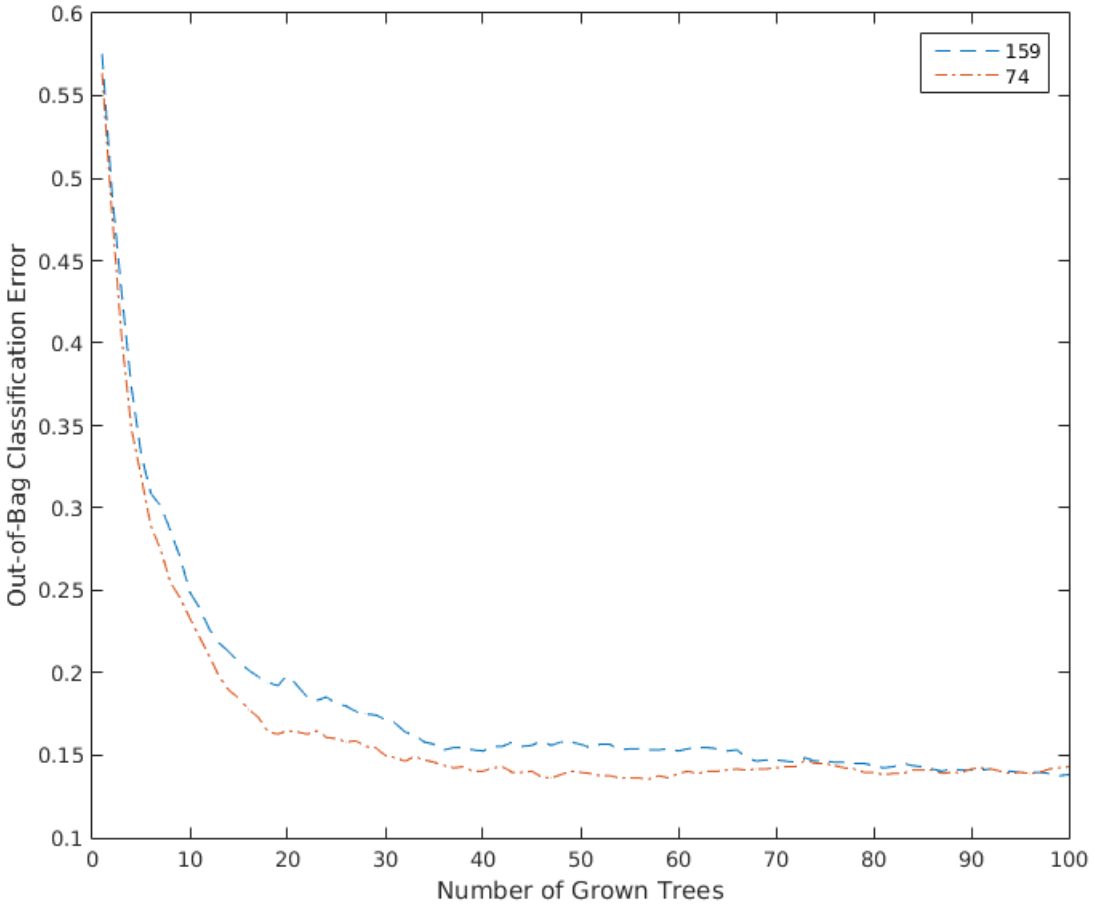


FIGURE 5.2: Cumulative out-of-bag classification errors for both feature sets

## 5.6 Analysis and Discussion

### 5.6.1 Selected Features

Looking at Figure 5.1 I can see the list is dominated by spectral features and has only three features related to dynamics. I was not surprised to see MFCC's in the 74 selected audio features as they have been proven before to perform quite well in speech recognition and audio classification tasks [129–131]. The Low Energy audio feature also plays a very significant role in classification. The Low Energy audio feature can be defined as the percentage of frames showing less than average RMS energy [127]. Vocals with silences or drum hits would have a high low energy rate compared to say a bowed string, so this may be one of the reasons it was so successful.

The maximum and average spectral spread as well as the standard deviation of kurtosis are also placed in top five ranked audio features. This suggests that the shape of the

TABLE 5.4: *Agglomerative clustering results using all features and the reduced feature set*

| **159 Features** | **MT 1** | **MT 2** | **MT 3** | **MT 4** | **MT 5** |
|---|---|---|---|---|---|
| Cophenetic C.C. | 0.799 | 0.844 | 0.751 | 0.894 | 0.814 |
| Audio tracks | 43 | 26 | 25 | 26 | 14 |
| Incorrect subgroups | 3 | 1 | 3 | 1 | 2 |
| Incorrect audio tracks | 19 | 7 | 5 | 7 | 2 |
| Percentage incorrect audio tracks | 44% | 26.9% | 20% | 26.9% | 14% |
| **74 Features** | **MT 1** | **MT 2** | **MT 3** | **MT 4** | **MT 5** |
| Cophenetic C.C. | 0.771 | 0.887 | 0.806 | 0.924 | 0.956 |
| Audio tracks | 43 | 26 | 25 | 26 | 14 |
| Incorrect groups | 2 | 0 | 1 | 0 | 2 |
| Incorrect audio tracks | 6 | 0 | 1 | 0 | 2 |
| Percentage incorrect audio tracks | 13% | 0% | 4% | 0% | 14% |

TABLE 5.5: *Agglomerative clustering results for all multitracks*

| | **159 Features** | **74 Features** |
|---|---|---|
| Avg. Cophenetic C.C. | 0.8203 | 0.8642 |
| Total no. audio tracks | 114 | 114 |
| Avg. no. audio tracks | 26.1 | 26.1 |
| Total incorrect subgroups | 10 | 5 |
| Total incorrect audio tracks | 40 | 9 |
| Percentage incorrect audio tracks | 35.08% | 7.89% |

audio spectrum for each audio track was one of the most important factors. The spectral centroid, brightness and roll off 95% also featured in the top 20, which are all spectral features.

I was expecting the Periodicity feature to perform much better, but it did not even make it into the subset of 74 audio features. I expected this to be important for drum and percussion classification. Ideally, this would be predictably high for drums, but low for vocals.

### 5.6.2 Agglomerative clustering

If we compare the results from agglomerative clustering using the entire audio feature set and the reduced audio feature set we can clearly see that the reduced audio feature set achieved a higher performance. The overall percentage of incorrectly clustered audio tracks changes from 35.08% for the entire audio feature set to 7.89% for the reduced audio feature set. I also found that the reduced audio feature set has a slightly higher average cophenetic correlation coefficient than the entire audio feature set. This suggests the

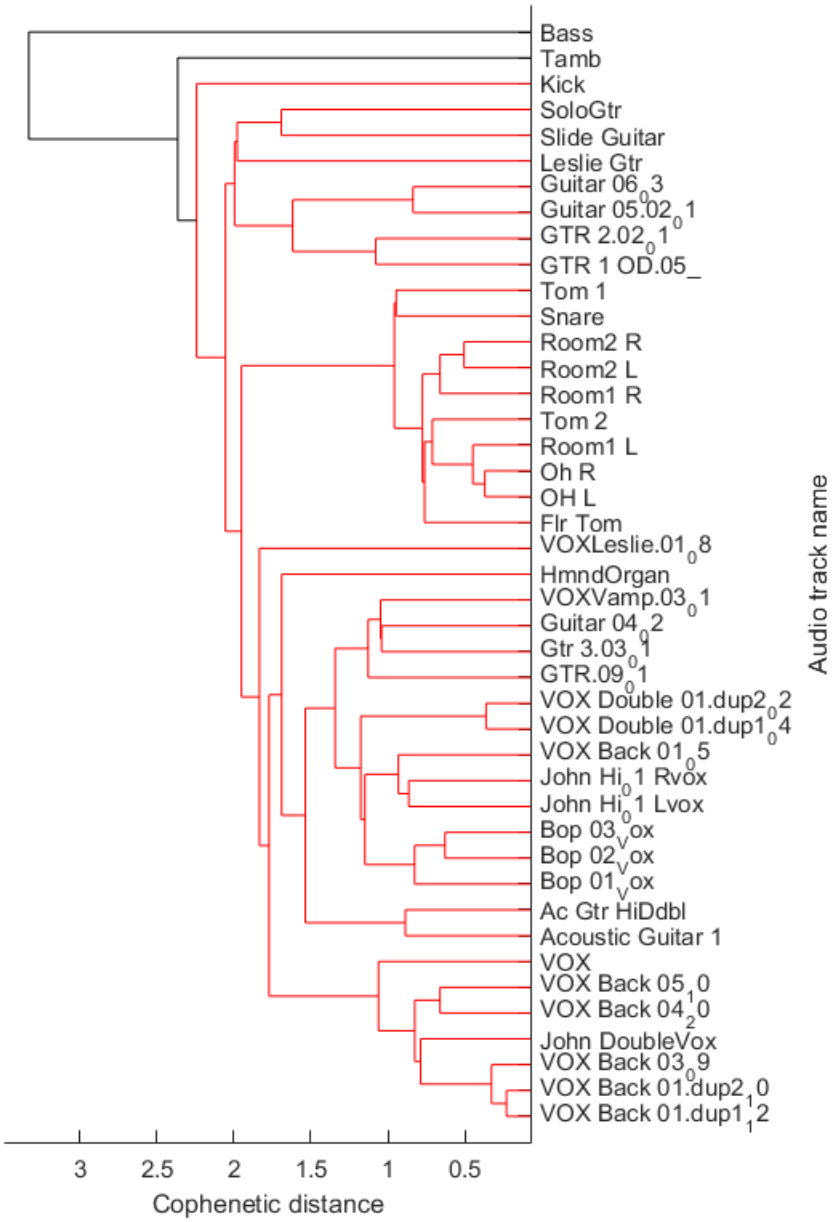


FIGURE 5.3: Dendrogram of MT 1 using the reduced feature set. Different stem types are shown to be close to each other. This is indicated by the cophenetic distance. The bottom of the part of this dendrogram has mainly vocals linked together, while the upper part has mainly drums and guitar lined together.

clustering better fits the reduced audio feature dataset. Furthermore, the total number of incorrectly created subgroups was halved when using the reduced audio feature set. Table 5.5 shows these results.

There is also an overall trend of higher performance for the reduced audio feature set

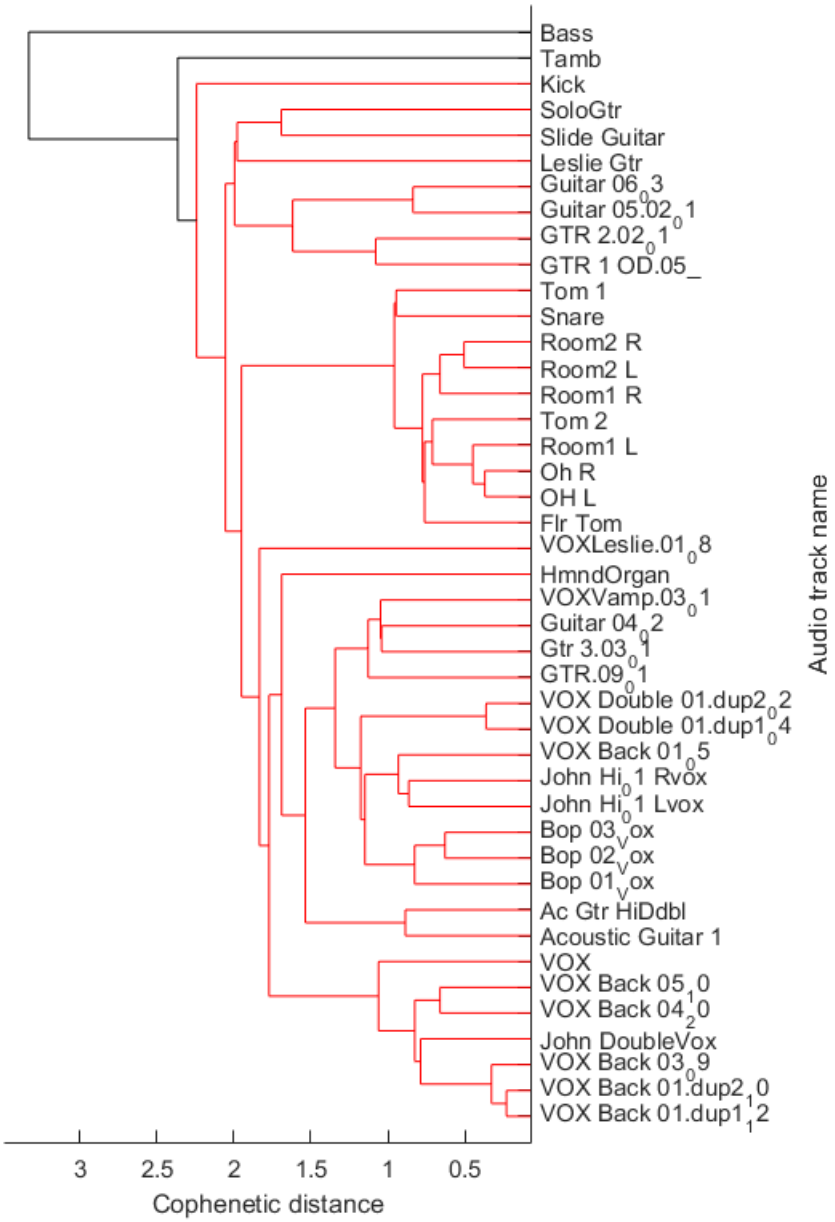


FIGURE 5.4: Dendrogram of MT 2 using the reduced feature set. Different stem types are shown to be close to each other. This is indicated by the cophenetic distance. The bottom of the part of this dendrogram has mainly vocals linked together, while the upper part has mainly drums and guitar lined together.

when we examine each multitrack separately. MT 1 was the worst performing multitrack for both the entire audio feature set and the reduced audio feature set. MT 1 when using the reduced audio feature set, had a lower misclassification measure than MT 1 using the entire audio feature set, but surprisingly has a slightly lower cophenetic correlation coefficient. Overall, MT 1 had the lowest cophenetic correlation coefficient for both sets

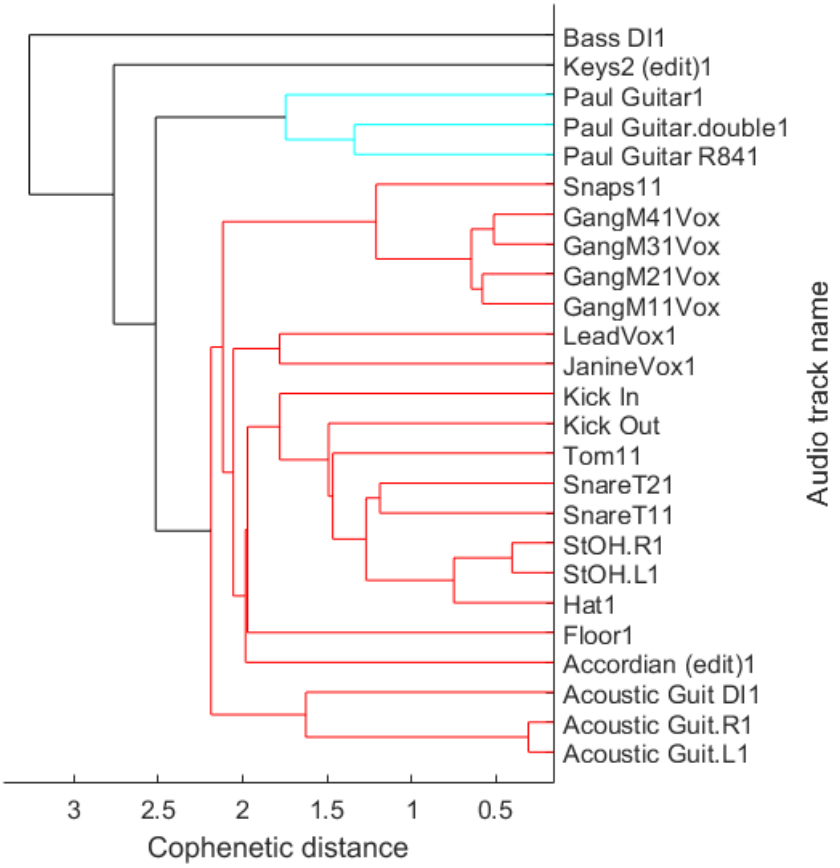


FIGURE 5.5: Dendrogram of MT 3 using the reduced feature set. Different stem types are shown to be close to each other. This is indicated by the cophenetic distance. The bottom of the part of this dendrogram has mainly acoustic guitar linked together, the middle part has mainly drums linked together, while the top part consists of vocal, keys and guitar linked together.

of audio features and this maybe because it also had the most amount of audio tracks to cluster. This may have been improved by using a varying maximum amount of clusters based on how many audio tracks are present. It is also worth mentioning that once the experiment was finished I listened back to the incorrectly subgrouped audio tracks for the reduced audio feature set and I found that these audio tracks suffered badly from microphone bleed. This is most likely the cause of the poor classification accuracy as two different instrument types can be heard on the recording. This problem could be addressed by using an automatic noise gate to reduce the microphone bleed [132].

The four other multitracks had greater success than MT 1 when clustered, but this may be due to them having fewer audio tracks to cluster. When we compare the results of the entire audio feature set versus the reduced audio feature set we can see a big improvement in results. Especially in MT 2 and MT 4 where the misclassification measure dropped to 0% in both cases. In MT 3, when using the reduced audio feature set we see that

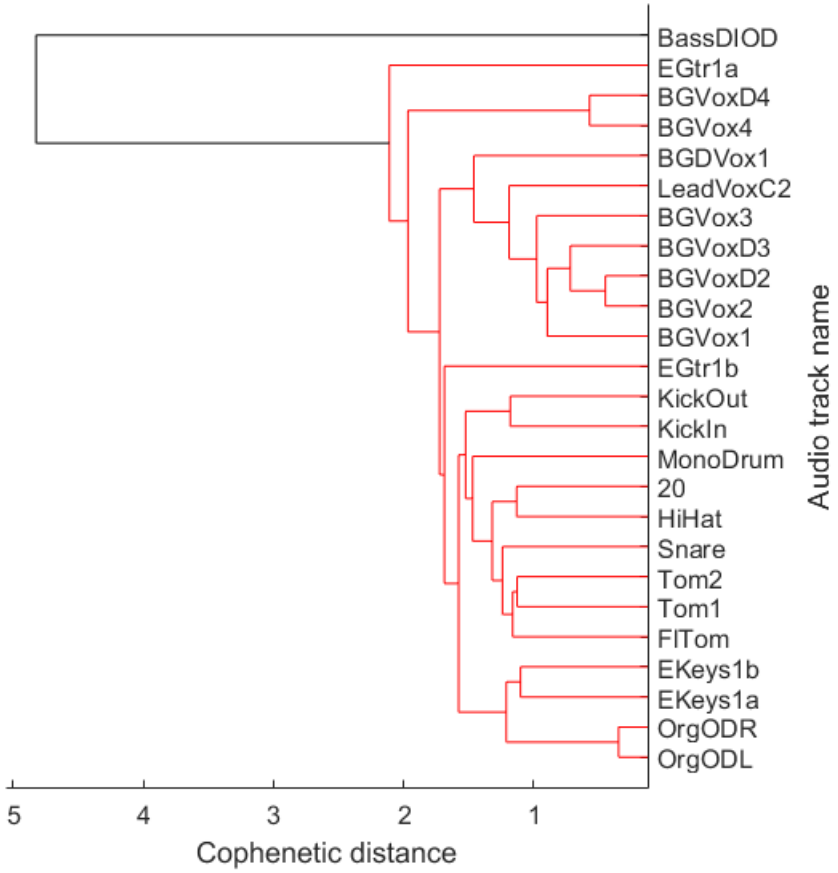


FIGURE 5.6: Dendrogram of MT 4 using the reduced feature set. Different stem types are shown to be close to each other. This is indicated by the cophenetic distance. The bottom of the part of this dendrogram has mainly keys linked together, the middle part has mainly drums linked together, while the top part consists of vocal, guitar and bass linked together.

we had only one misclassification. This was the 'Snaps' audio track being subgrouped with the 'GangM' vocal tracks and is depicted in Figure 5.5. There is a small amount of microphone bleed on the 'GangM' vocal tracks, so this may be the reason why we are seeing this misclassification. In MT 5 the misclassification is more difficult to explain as there does not seem to be any audible microphone bleed. This may be because the synthesiser has a similar timbre to the lead vocalist. Figure 5.7 shows that 'Synth21' is further away from the violins than the Synth11' is from the vocals, suggesting that 'Synth11' is similar to the vocal audio tracks.

When looking at Figure 5.3, Figure 5.4, Figure 5.5, Figure 5.6 and Figure 5.7 generally the lower parts of the trees tend to cluster the audio tracks together correctly. It is very easy to pick out drum, vocal and guitar clusters especially. The best examples are shown in Figure 5.4, Figure 5.5 and Figure 5.6. Interestingly, the 'Bass' audio track is the furthest distance from any other audio track in each of the multitracks. This most

likely has to do with this instrument occupying the lower frequency bands and the rest of the instruments tending to be in mid and upper frequency ranges.

## 5.7 Conclusion

In this chapter, I determined a set of audio features that could be used to automatically subgroup multitrack audio using a Random Forest for feature selection. I took a set of 159 low level audio features and reduced this to 74 low level audio features using feature selection. I selected these features from a dataset of 54 individual multitrack recordings of varying musical genre. I also showed that the most important audio features tended to be spectral features. I used the reduced audio feature set to agglomeratively cluster five unseen multitrack recordings. I then compared the results of the agglomerative clustering using the entire audio feature set to the agglomerative clustering using the reduced audio feature set. I was able to show that the overall misclassification measure went from 35.08% using the entire audio feature set to 7.89% using the reduced audio feature set. Thus indicating that my reduced set of audio features provides a significant increase in classification accuracy for the creation of automatic subgroups. Part of the novelty of this approach was that I was trying to classify audio tracks of entire multitrack recordings. Whereby, multitracks have the issue where recordings may contain artefacts such as microphone bleed. This did cause us problems in some cases, but I was easily able to identify the cause by listening to the problematic audio tracks.

In future work, automatic subgrouping could be applied to music from the Dance or Jazz music genres. In this case I only applied automatic subgrouping to Pop, Rock, Indie etc. However, it would seem that currently the subgroups for the Dance or Jazz music genres are not very well defined, so further research would be needed on best practices in subgrouping for music production of this kind. It would also be interesting to see how automatic subgrouping could be used in current automatic mixing systems like [9, 49, 133], where each automatic mixing algorithm is used on each subgroup of instruments individually to create a submix. Then once all the subgroups are automatically mixed, the automatic mixing algorithm would be used to mix each individual subgroup. In this work I inspected the correctness of the automatically generated subgroups manually, in further work I would like to test the validity of this technique automatically by using cross validation.

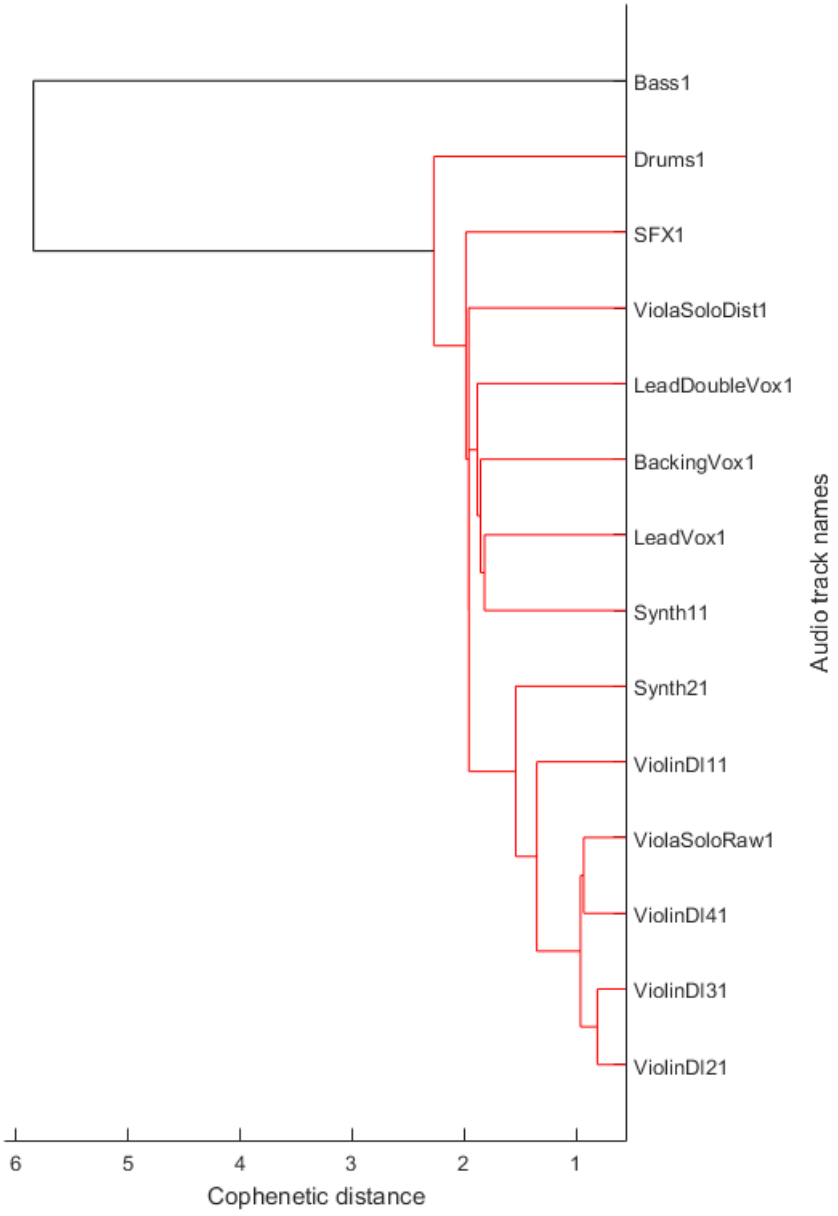


FIGURE 5.7: Dendrogram of MT 5 using the reduced feature set. Different stem types are shown to be close to each other. This is indicated by the cophenetic distance. The bottom of the part of this dendrogram has mainly strings linked together, the middle part has mainly vocals and synths linked together, while the top part consists of drums and bass linked together.

# Chapter 6

# An empirical approach to the relationship between emotion and music production quality

## 6.1 Introduction

There have been several studies that have looked at why people prefer certain mixes over others. [13, 134] conducted a mix experiment where groups of nine mix engineers were asked to mix 10 different songs. The mixes were evaluated in a listening test to infer the quality as perceived by a group of trained listeners. Mix preference ratings were correlated with a large number of low level features in order to explore if there was any relationship, but the findings indicated in this particular case was that there were no significantly strong correlations. The details of this study are described in Chapter 3.

In Chapter 3, the same tracks used in [13, 134] were used to ascertain the impact of subgrouping practices on mix preference. The quantity of subgroups and the type of subgroup effect processing used was looked at for each mix. Then these findings were correlated with mix quality preference ratings to see the extent of the relationship [113].

In a somewhat related study, [135] claimed that audio production quality is linked to perceived loudness and dynamic range compression. It also demonstrated that a participant's expertise is not a strong factor in assessing audio quality or musical preference. However, the relationship between music production quality was not explored in this study.

To my knowledge, there have been no previous studies that examined the relationship between music production quality and emotional response. This represents a new area

of research in music perception and emotion that I intend to explore. In [12], three of the mix engineers that were interviewed mentioned the importance of emotion in the context of mixing and producing music. This indicates that emotion plays a significant role in how a mix engineer tries to achieve a desired mix. [136] states that dynamic contrast in a piece of music has been heralded as one of the most important factors for conveying emotion.

The purpose of the current study is to determine the extent of the link between music production quality and musically induced and perceived emotions. The participants in this study listened to low and high quality mixes (rated in [13, 134]) of the same musical piece. These were participants I recruited separately for this experiment and had no relation to the studies detailed in [13, 134]. I then measured each participant's subjective experience, peripheral physiological changes, changes in facial expressions and head nods, and shakes as they listened to each mix.

The rest of the chapter is organised as follows. Section 6.2 provides the methodology used to conduct this experiment. Section 6.3 presents the results obtained and the subsequent analysis. Section 6.4 discusses the results, the chapter is concluded in section 6.5. Finally, section 6.6 proposes future work.

## 6.2 Methodology

### 6.2.1 Research questions and hypotheses

My original hypothesis was that music production quality had a direct effect on the induced and perceived emotions of the listener. However, before I proceeded to the main study, I conducted a short pilot study.

#### 6.2.1.1 Pilot Study

The pilot study consisted of us running the experiment for six participants, where three had critical listening skills and the other three did not. I measured each participant's subjective experience, peripheral physiological changes and changes in facial expressions as they listened to each mix. The feedback from the pilot study indicated that training was required in order for participants to become familiar with the adjectives used to describe induced emotions. I also decided to track head nods and shakes, a typical response to musical enjoyment, based on a review of the recorded videos. I found participants were moving their heads a lot in time with the music. Observation of

potential differences between critical and non-critical listeners led us to revise my original hypothesis.

#### 6.2.1.2 Main hypothesis

The main hypothesis was refined to be that music production quality has more of an effect on the induced and perceived emotions of critical listeners than those of non-critical listeners. Thus, implying that the null hypothesis is that critical and non-critical listeners experience the same induced and perceived emotions regardless of music production quality. This is what I tested using statistical analysis in the later sections.

### 6.2.2 Participants

Twenty participants were recruited from within the university. 14 were male, 6 female and their ages ranged from 26 to 42 ($\mu = 30.4, \sigma^2 = 4.4$). 10 participants had critical listening skills, i.e, knew what critical listening involved and had been trained to do so previously or had worked in a studio, while the other 10 did not i.e., no music production experience and not trained in how to critique a piece of music. A pre-experiment questionnaire established the genre preference of participants, shown in Table 6.1, since some participants may have bias towards certain genres.

TABLE 6.1: Genre preference for participants

| **Genre** | **No. of Participants** |
|---|---|
| Rock/Indie | 15 |
| Dance/Electronic | 11 |
| Pop | 8 |
| Jazz | 6 |
| Classical | 4 |

### 6.2.3 Stimuli

Ten different songs were used, each with nine mixes (90 mixes in total). Songs were split into three study groups, where mixes for songs within a study group were created by 8 student mix engineers and their instructor, who was a professional mix engineer (the same professional mix engineer participated in Groups 1 and 2). These mixes were obtained from the experiment conducted in [134] and the same ones used in Chapter 3. Mixes of a song had been rated for mix quality preference by all the members of the other study groups, so no one rated their own mix. Further details on how the stimuli was obtained can be seen in [134] and in Chapter 3. For my experiment, I selected

the lowest and highest quality rated mixes of each song. Table 6.2 shows the names of each song, the song genre and which group mixed each song. Some song names had to be removed due to copyright issues, but the rest are available on the Open Multitrack Testbed [137]. All mixes were loudness normalised using ITU-R BS. 1770-2 specification [138] to avoid bias towards loud mixes.

TABLE 6.2: Song titles, song genres and mix groups. Songs in italics are not available online due to copyright restrictions.

| **Song Name** | **Genre** | **Mixed By** |
|---|---|---|
| Red to Blue - (S1) | Pop-Rock | Group 1 |
| Not Alone - (S2) | Funk | Group 1 |
| *My Funny Valentine* - (S3) | Jazz | Group 1 |
| Lead Me - (S4) | Pop-Rock | Group 1 |
| In the Meantime - (S5) | Funk | Group 1 |
| - (S6) | Soul-Blues | Group 2 |
| *No Prize* - (S7) | Soul-Jazz | Group 2 |
| - (S8) | Pop-Rock | Group 2 |
| Under a Covered Sky - (S9) | Pop-Rock | Group 2 |
| Pouring Room - (S10) | Rock-Indie | Group 3 |

### 6.2.4 Measurements

#### 6.2.4.1 Physiological Measures

To measure skin conductance I used small (53mm x 32 mm x 19 mm) wireless GSR sensors developed by Shimmer Research. The GSR module was placed around the wrist of their usually inactive hand, and electrodes strapped to their index and middle finger. ECG measurements were attempted but discarded due to extreme noise levels in the data, at least partly since participants moved in the rotatable chair provided.

#### 6.2.4.2 Facial Expression and Head Nod-Shake

To record video for facial expression and head nod/shake detection, I used a Lenovo 720p webcam that was embedded in the laptop used to perform the experiment. In Figure 6.1 we can see the automatic facial feature tracking for one of my participants.

#### 6.2.4.3 Self-Report

After listening to each piece of music, participants used GEMS-9 to rate the emotions induced while listening. This was done using a 5-point Likert scales ranging from 'Not at all' to 'Very much' based on 9 adjectives; wonder, transcendence, power, tenderness, nostalgia, peacefulness, joyful activation, sadness and tension. Each participant also

FIGURE 6.1: Facial features tracked for detecting facial action units during music listening.

rated the emotions they perceived in each song using three discrete (1-100) sliders for arousal, valence and tension. They were also asked to indicate how much they liked each piece of music they heard based on a 5-point Likert scale ranging from 'Not at all' to 'Very much'.

#### 6.2.4.4 User Interface

The physiological measurements, self-report scores and video were recorded into a bespoke software program developed for the experiment. It was designed to allow the experiment to run without the need for assistance, and the graphical user interface was designed to be as aesthetically neutral as possible.

#### 6.2.4.5 Pre- and Post-Experiment Questionnaires

I provided pre- and post-experiment questionnaires. The pre-experiment questionnaire asked simple questions related to age, musical experience, music production experience, music genre preference and critical listening skills. There was also a question clarifying each participant's emotional state as well as how tired they were when they started the study. If any participant indicated that they were very tired, I asked them to attempt the experiment at a later time once rested.

The post-experiment questionnaire asked questions such as could they hear an audible difference between the two mixes of each song, was there any difference in emotional content between the two mixes of each song and was there any difference in the induced emotions between the two mixes of each song. These were all asked on a 5-point Likert scale ranging from 'Not at all' to 'Very much'.

### 6.2.5 Setup

The experiment took place in a dedicated listening room at the university. The room was very well lit, which was important for facial expression analysis and head nod/shake detection. Each participant was sat at a studio desk in front of the laptop used for the experiment. The audio was heard over a pair of studio quality loudspeakers, where the participant could adjust the volume of the audio to a comfortable level. Figure 6.2 shows the room in which the experiment was conducted.

FIGURE 6.2: Studio space where the experiment was conducted.

### 6.2.6 Tasks

After the pre-experiment questionnaire, I trained each participant in how the interface worked. They were supervised while they listened to two example songs and they were asked if they understood all the adjectives and terms used in the experiment. If they did not understand any adjective or term, they were referred to a dictionary where the adjective or term was subsequently explained to them.

Each participant was then asked to relax and listen to the music as they would at home for enjoyment. Next, three minutes of relaxing sounds were played to each participant in order to get an emotional baseline. They then had to click play in order for one of the mixes to be heard, where the order in which mixes were presented was randomised. While the music was playing, GSR measurements and facial and head movements were recorded. Once the music finished, each participant rated the induced emotions using GEMS-9. They then rated perceived emotions on the Arousal-Valence-Tension scale and

rated how much they liked each mix. Once answers were submitted, there was another 30 seconds of relaxing sounds played for an emotional baseline and the same procedure repeated for the next mix. The participant was updated on their progress throughout the experiment via the software. Finally, the participant filled out the post-experiment questionnaire and the experiment was concluded. This whole process is illustrated in Figure 6.3.

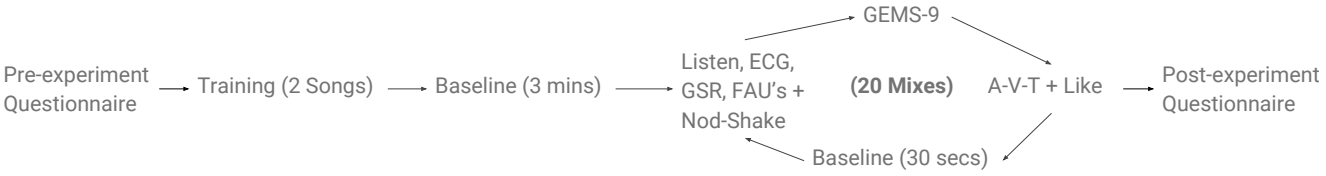


FIGURE 6.3: Tasks involved in the experiment.

### 6.2.7 Data Processing

Skin conductance response (SCR) has been shown to be useful in analysis of GSR data [139, 140]. I used Ledalab 5 to extract the timing and amplitude of SCR events from the raw GSR data (sampled at 5Hz) using Continuous Decomposition Analysis (CDA) [141]. Interpolation was performed and the mean, standard deviation, positions of maxima and minima, and number of extrema divided by task duration, were calculated from the SCR amplitude series for each mix [139, 142]. GSR data of one critical listener was discarded due to poor electrode contact.

I extracted head nod events, head shake events and dimensional measures such as arousal, expectation, intensity, power and valence from each video clip using the classification method introduced in [143]. This method captures each of these events and dimensional emotion values from every 20 frames (0.8 sec) of video. Head nod and head shake events are binary values, while the rest of the features are continuous values. The classification method used for this was trained to capture head nod, head shake events and variations of these using Hidden Markov Models and Support Vector Machines. I extracted the total head shake and head nod events and took average and standard deviation values for the rest of the features for each video clip.

Intensity values (0-1) of eight AUs, see Table 6.3, were extracted every five frames (0.2 sec) for each video, using the method of [144]. I calculated the average and standard deviation values of each AU for each video clip.

TABLE 6.3: Extracted Action Units

| **AU Number** | **FACS Name** |
|---|---|
| AU1 | Inner brow raiser |
| AU2 | Outer brow raiser |
| AU4 | Brow lowerer |
| AU12 | Lip corner puller |
| AU17 | Chin raiser |
| AU25 | Lip raiser |
| AU28 | Lip suck |
| AU45 | Blink |

## 6.3 Experiment and Results

Table 6.4 summarises the conditions tested in my experiment. In conditions C1, C2, C5 and C6, I constrained listener type and tested if there was a statistical difference in emotional response ratings and scores based on mix quality. In conditions C3, C4, C7 and C8 I constrained mix quality type and tested if there was a statistical difference in emotional response ratings and scores based on critical listening skills.

I used two types of weightings for ratings and scores, similar to the approaches in [145–147]. The audible difference weighting was used in conditions C1 - C4. It weighted participant results by how much they indicated they could hear an audible difference between the high and low quality mix types. The perceived emotional difference weighting was used in conditions C5 - C8, based on how much participants could perceive an emotional difference between the high and low quality mixes. Weights were calculated based on each participant's response to questions asked in the Post-Experiment questionnaire. Each participant indicated on a Likert scale how much they could perceive an audible difference between the two mixes of each song and to what extent they could perceive an emotional difference between the mixes of each song. Weighting was applied as $W_R = \frac{O_R D_X}{N}$, where $O_R$ is the original and $W_R$ the weighted result, $D_X$ is the Likert value for either perceived audible difference or perceived emotional difference, and $N$ is the number of points used in the Likert scale.

In conditions C1, C2, C5 and C6 I used the Wilcoxon Signed Rank non-parametric statistical test because my data is ordinal and I have the same subjects in both datasets. In conditions C3, C4, C7 and C8 I used the Mann-Whitney U non-parametric statistical test because my data is ordinal and I am comparing the medians of two independent groups. In each table in this section the results shown are p-values from the statistical tests for rejecting the null hypothesis, where the numbers in bold are significant ($p < 0.05$). I have not used the Bonferroni correction because the method is concerned with

the general null hypothesis. In this instance, I am investigating how emotions and reactions vary along the many different dimensions tested [148].

The data used for this analysis can be accessed online[1] for further examination.

TABLE 6.4: Different types of conditions tested

| Condition | Constrained | Varied | Weighting | Statistical Test |
|---|---|---|---|---|
| C1 | Critical Listener | High Quality Mix vs Low Quality Mix | Audible Difference | Wilcoxon Sign Rank |
| C2 | Non-critical Listener | High Quality Mix vs Low Quality Mix | Audible Difference | Wilcoxon Sign Rank |
| C3 | High Quality Mix | Critical Listener vs Non-Critical Listener | Audible Difference | Mann-Whitney U |
| C4 | Low Quality Mix | Critical Listener vs Non-Critical Listener | Audible Difference | Mann-Whitney U |
| C5 | Critical Listener | High Quality Mix vs Low Quality Mix | Emotional Difference | Wilcoxon Sign Rank |
| C6 | Non-critical Listener | High Quality Mix vs Low Quality Mix | Emotional Difference | Wilcoxon Sign Rank |
| C7 | High Quality Mix | Critical Listener vs Non-Critical Listener | Emotional Difference | Mann-Whitney U |
| C8 | Low Quality Mix | Critical Listener vs Non-Critical Listener | Emotional Difference | Mann-Whitney U |

### 6.3.1 GEMS-9

Table 6.5 compared the ratings for each of the GEMS-9 emotional adjectives on a song by song basis for conditions C1 to C4. I have removed any p-values that were not significant in order to make the tables easier to read. There are four statistically significant p-values for C1 in contrast to C2 where there are no statistically significant p-values. This occurred for two songs and happened for the emotions transcendence, tenderness, joyful activation and tension. I see a lot more significant p-values for C3 and C4 than for C1 and C2. I have 47 significant p-values out of a possible 90 for C3 and 43 significant p-values out of 90 for C4. The most amount of significant p-values occur for the emotions of nostalgia, peacefulness, joyful activation and sadness.

### 6.3.2 Arousal-Valence-Tension

Table 6.6 compares the ratings for Arousal-Valence-Tension dimensions on a song by song basis for Conditions C1 to C4. For C1, there are four statistically significant p-values for arousal, two for valence, and two for tension. This is in contrast to C2 where there is one significant p-value for arousal and one for valence. The significant p-values for C1 are related to six songs in contrast to C2 where they are only related to one song. For both C3 and C4, there are six significant p-values for arousal, all ten for valence and four for tension. p-values for both are similar in terms of distribution over the dimensions, but they differ by song.

[1] https://goo.gl/EA86K2

TABLE 6.5: GEMS-9 - Audible Difference Weighting for Conditions C1 to C4.

| C1 | **Wonder** | **Trans** | **Power** | **Tender** | **Nostal** | **Peace** | **Joyful** | **Sadness** | **Tension** |
|---|---|---|---|---|---|---|---|---|---|
| **S4** | | **0.031** | | | | | **0.031** | | |
| **S7** | | | | **0.031** | | | | | **0.031** |

| C2 | **Wond** | **Trans** | **Power** | **Tender** | **Nostal** | **Peace** | **Joyful** | **Sadness** | **Tension** |
|---|---|---|---|---|---|---|---|---|---|

| C3 | **Wonder** | **Trans** | **Power** | **Tender** | **Nostal** | **Peace** | **Joyful** | **Sadness** | **Tension** |
|---|---|---|---|---|---|---|---|---|---|
| **S1** | | **0.030** | **0.042** | | | **0.014** | **0.043** | **0.011** | **0.023** |
| **S2** | | | **0.039** | **0.028** | | **0.007** | | **0.034** | |
| **S3** | | | | | **0.024** | **0.005** | | **0.041** | |
| **S4** | **0.022** | | | **0.018** | **0.038** | **0.028** | **0.007** | **0.027** | |
| **S5** | **0.042** | | | | **0.031** | | | **0.035** | |
| **S6** | | | **0.039** | | **0.028** | **0.041** | **0.014** | | |
| **S7** | | | | | | **0.006** | **0.038** | **0.038** | |
| **S8** | **0.035** | **0.036** | **0.038** | **0.031** | **0.013** | | **0.027** | | |
| **S9** | **0.027** | | **0.043** | | **0.030** | **0.008** | **0.035** | **0.042** | **0.017** |
| **S10** | **0.017** | | | **0.022** | **0.031** | **0.020** | **0.027** | | |

| C4 | **Wonder** | **Trans** | **Power** | **Tender** | **Nostal** | **Peace** | **Joyful** | **Sadness** | **Tension** |
|---|---|---|---|---|---|---|---|---|---|
| **S1** | **0.010** | **0.033** | | | **0.029** | **0.006** | | **0.025** | **0.049** |
| **S2** | | | **0.011** | | | | **0.023** | **0.009** | |
| **S3** | | | **0.028** | | **0.014** | **0.005** | **0.026** | | |
| **S4** | | | | | | **0.039** | | | |
| **S5** | **0.042** | | | | | **0.024** | | | |
| **S6** | **0.034** | | | | **0.010** | **0.018** | | **0.028** | **0.020** |
| **S7** | | | | **0.020** | **0.034** | **0.004** | **0.023** | | |
| **S8** | **0.017** | **0.015** | **0.045** | **0.021** | **0.007** | **0.006** | | | |
| **S9** | | | **0.049** | | **0.018** | | | **0.039** | **0.031** |
| **S10** | **0.004** | **0.016** | **0.041** | **0.006** | **0.011** | **0.008** | **0.007** | **0.032** | |

### 6.3.3 GSR

I compared the mean, standard deviation, positions of maxima and minima and frequency of event values for each participant's GSR data on a song by song basis. However, since there were few significant p-values I did not present the results in a table. This was also the only part of the experiment where I tested conditions C1 to C4 as well as conditions C5 to C8, as it was the only time these conditions gave a noticeable amount of significant p-values.

When I tested C1 and C2, there were only 3 out of 50 statistically significant p-values for critical listeners and 3 out of 50 statistically significant p-values for non-critical listeners. Similar results occurred when I tested conditions C5 and C6. C3 gave 5 out of 50 statistically significant p-values for two songs, and there were 4 out of 50 for C4.

TABLE 6.6: Arousal-Valence-Tension - Audible Difference Weighting for Conditions C1 to C4.

| C1 | **A** | **V** | **T** |
|---|---|---|---|
| **S1** | | | |
| **S2** | | | |
| **S3** | **0.021** | | |
| **S4** | **0.002** | | |
| **S5** | | | |
| **S6** | | | |
| **S7** | **0.039** | | |
| **S8** | | **0.035** | |
| **S9** | **0.027** | | **0.016** |
| **S10** | | **0.016** | **0.031** |

| C3 | **A** | **V** | **T** |
|---|---|---|---|
| **S1** | **0.019** | **0.013** | **0.045** |
| **S2** | | **0.011** | |
| **S3** | | **0.004** | **0.021** |
| **S4** | | **0.018** | |
| **S5** | **0.008** | **0.017** | |
| **S6** | **0.021** | **0.009** | **0.049** |
| **S7** | | **0.002** | |
| **S8** | **0.008** | **0.006** | **0.038** |
| **S9** | **0.009** | **0.002** | |
| **S10** | **0.019** | **0.004** | |

| C2 | **A** | **V** | **T** |
|---|---|---|---|
| **S1** | | | |
| **S2** | **0.047** | **0.039** | |
| **S3** | | | |
| **S4** | | | |
| **S5** | | | |
| **S6** | | | |
| **S7** | | | |
| **S8** | | | |
| **S9** | | | |
| **S10** | | | |

| C4 | **A** | **V** | **T** |
|---|---|---|---|
| **S1** | **0.026** | **0.011** | |
| **S2** | **0.007** | **0.005** | |
| **S3** | **0.038** | **0.006** | |
| **S4** | | **0.014** | |
| **S5** | **0.004** | **0.010** | **0.010** |
| **S6** | | **0.005** | **0.026** |
| **S7** | | **0.006** | |
| **S8** | **0.011** | **0.021** | **0.041** |
| **S9** | **0.007** | **0.015** | **0.028** |
| **S10** | | **0.013** | |

When I tested condition C7, there were 9 out of 50 statistically significant p-values. This is in contrast to C8 where there were 2 out of 50 statistically significant p-values.

### 6.3.4 Head Nod and Shake

I compared Head Nod and Shake scores on a song by song basis. There were no statistically significant p-values for condition C1, and only 2 out 70 p-values for C2 were statistically significant. The results for conditions C3 and C4 are summarised in Table 6.7. For C3, I have 31 significant p-values out of a possible 70. The most amount of significant p-values occurred for shake, expectation and power. C4 gave 35 significant p-values out of 70. The largest amount of significant p-values occur for shake, arousal and power.

### 6.3.5 Facial Action Units

I compared the standard deviation for each participant's Facial Action Unit scores on a song by song basis. I saw 3 out of 80 statistically significant p-values for condition C1,

TABLE 6.7: Head Nod and Shake - Audible Difference Weighting for Conditions C3 and C4.

| C3 | Nod | Shake | Arousal | Expectation | Intensity | Power | Valence |
|---|---|---|---|---|---|---|---|
| **S1** | **0.023** | **0.041** | **0.006** | **0.006** | **0.006** | | |
| **S2** | | **0.017** | | | | **0.034** | |
| **S3** | | **0.002** | **0.009** | **0.004** | **0.017** | **0.000** | |
| **S4** | | | | | | **0.009** | **0.002** |
| **S5** | **0.026** | | **0.006** | | **0.006** | | |
| **S6** | | **0.013** | | **0.026** | | **0.038** | |
| **S7** | | | | **0.014** | | | |
| **S8** | | **0.011** | **0.021** | **0.009** | | **0.031** | |
| **S9** | | **0.005** | **0.001** | **0.001** | **0.002** | **0.003** | |
| **S10** | | | | **0.002** | | | |

| C4 | Nod | Shake | Arousal | Expectation | Intensity | Power | Valence |
|---|---|---|---|---|---|---|---|
| **S1** | | **0.005** | | | | **0.026** | |
| **S2** | **0.006** | | **0.010** | **0.009** | **0.010** | | |
| **S3** | **0.028** | | | | | **0.014** | |
| **S4** | | | | | | **0.045** | **0.007** |
| **S5** | **0.034** | **0.036** | **0.017** | | | | |
| **S6** | | **0.007** | **0.038** | **0.031** | **0.045** | **0.031** | |
| **S7** | **0.005** | | **0.007** | **0.011** | | **0.005** | |
| **S8** | | **0.001** | **0.017** | **0.000** | **0.021** | **0.004** | |
| **S9** | **0.017** | **0.017** | **0.021** | | | **0.034** | |
| **S10** | | **0.006** | **0.028** | **0.023** | **0.013** | | |

whereas C2 gave 7 out of 80 statistically significant p-values. Results for conditions C3 and C4 are summarised in Table 6.8. There were 23 significant p-values out of a possible 80, mainly for AU1, AU4 and AU45. For condition C4, I have 20 significant p-values out of 80, mostly from AU4 and AU45.

I also examined which AUs had the highest intensity throughout the experiment. I checked every mix that each participant listened to, to see if any of their average AU intensities was >= 0.5. If the average AU intensity was >= 0.5 I marked the AU for that particular mix with a 1, otherwise a 0. I summarised the results as a percentage of all the mixes listened to for critical listeners and non-critical listeners in Table 6.9. AU1 and AU4 gave the greatest amount of average AU intensities >= 0.5. The results for AU12 and AU17 were omitted since all the results were 0. Critical listeners experienced a greater number of average AU intensities >= 0.5 than non-critical listeners for all AUs except AU28. However, the difference in the case of AU28 is 0.005, which is negligible.

TABLE 6.8: FACS - Audible Difference Weighting for Conditions C3 and C4.

| C3 | AU1 | AU2 | AU4 | AU12 | AU17 | AU25 | AU28 | AU45 |
|---|---|---|---|---|---|---|---|---|
| **S1** | **0.011** | | **0.021** | | | | | |
| **S2** | **0.038** | | **0.006** | | | | | |
| **S3** | | | **0.026** | | | | | **0.038** |
| **S4** | **0.026** | | **0.014** | | | **0.038** | | |
| **S5** | **0.045** | | | | **0.007** | | | |
| **S6** | | | **0.004** | | | | | **0.045** |
| **S7** | **0.038** | | **0.006** | | **0.011** | | | **0.031** |
| **S8** | **0.038** | | **0.004** | | | | | **0.026** |
| **S9** | | | **0.038** | | | | | **0.014** |
| **S10** | | | | | | | | **0.021** |

| C4 | AU1 | AU2 | AU4 | AU12 | AU17 | AU25 | AU28 | AU45 |
|---|---|---|---|---|---|---|---|---|
| **S1** | | | **0.009** | | | | | **0.045** |
| **S2** | **0.031** | | **0.045** | | | | | **0.045** |
| **S3** | **0.003** | | **0.007** | | | | | |
| **S4** | | | **0.009** | | | | | **0.038** |
| **S5** | | | | | | | | **0.006** |
| **S6** | | | **0.002** | | **0.031** | | | **0.011** |
| **S7** | | | **0.021** | | | | | |
| **S8** | **0.045** | | **0.014** | | | | | **0.011** |
| **S9** | | | **0.009** | | | | | |
| **S10** | | | **0.006** | | | | | **0.026** |

## 6.4 Discussion

### 6.4.1 Findings

#### 6.4.1.1 GEMS-9

With GEMS-9 I investigated if there was a significant difference in the distribution of induced emotions of each listener type. Table 6.5 results indicate that the critical listeners were the only group where there was significant differences in the distribution of induced emotions between the two mix types. This suggests that my hypothesis is true. However, since there are so few p-values in comparison to the amount of tests I can not draw a strong conclusion from this.

Table 6.5 results also indicate that high quality mixes had a greater significant difference on the distribution of induced emotions between the two listener types. These results support my hypothesis, in that the high quality mix had more of an impact emotionally on one listener type over the other. They also imply that there was a greater difference in the indicated levels of joyful activation and sadness between critical and non-critical

TABLE 6.9: Percentage of mixes where average AU intensity was >= 0.5. (i) Non-critical listeners (ii) Critical listeners

| (i) | **AU1** | **AU2** | **AU4** | **AU25** | **AU28** | **AU45** |
|---|---|---|---|---|---|---|
| **A** | 0.9 | | | | | |
| **B** | 0.85 | | 0.85 | | | |
| **C** | 0.55 | | 0.7 | | | |
| **D** | | | | | | |
| **E** | | 0.05 | 0.95 | | | 0.05 |
| **F** | | | 0.75 | 0.05 | | |
| **G** | 0.25 | | 0.55 | | | |
| **H** | 1 | | 0.75 | | 0.05 | |
| **I** | 0.75 | | 0.7 | 0.05 | | |
| **J** | | | 0.85 | | | |
| **Total %** | 0.43 | 0.005 | 0.61 | 0.01 | 0.005 | 0.005 |
| (ii) | **AU1** | **AU2** | **AU4** | **AU25** | **AU28** | **AU45** |
| **K** | 1 | | 1 | | | |
| **L** | 0.95 | 0.05 | 0.25 | 0.25 | | 0.1 |
| **M** | 0.1 | | 0.95 | | | 0.2 |
| **N** | 0.55 | | 0.35 | | | |
| **O** | | | 1 | | | |
| **P** | 0.9 | | 1 | | | |
| **Q** | 0.45 | | 1 | | | |
| **R** | 0.2 | | 0.35 | 0.1 | | 0.15 |
| **S** | 0.75 | 0.45 | 1 | | | |
| **T** | 0.8 | | 0.05 | | | |
| **Total %** | 0.57 | 0.05 | 0.695 | 0.035 | 0 | 0.045 |

listeners for the high quality mixes (C3). Joyful activation and sadness would be synonymous with the positive and negative valence, implying that the quality of the mix may have an impact on how happy or sad a critical listener may feel.

#### 6.4.1.2 Arousal-Valence-Tension

I investigated if there was a significant difference in the distribution of emotions perceived by each listener type along Arousal-Valence-Tension dimensions. Table 6.6 indicates that for critical listeners there are more examples of where there are significant differences in the distribution of perceived emotions, especially with respect to arousal. This was the only time a noticeable difference in the amount of significant p-values occurred when I compared the critical listener's high quality mixes to critical listener's low quality mixes. This also occurred in the case of non-critical listeners (C2), but to a lesser extent. These results support my hypothesis, in that critical listeners were able to perceive an emotional difference between the two mixes much more so than non-critical listeners and this was mostly with respect to arousal and tension.

Table 6.6 showed a lot of significant p-values for Conditions C3 and C4 in comparison to C1 and C2. Interestingly, I have the same amount of significant values in each dimension for both conditions C3 and C4. This implies that there are the same amount of significant differences in the distribution of emotions for both listener types due to mix quality, but it varies by song. The two listener types are perceiving different levels of arousal and tension, but on different songs. However, this may have something to do with the participant's genre preference. These results are similar to those seen in Table 6.5 (iii) and (iv), in the respect that joyful activation corresponds to positive valence and sadness corresponds to negative valence.

#### 6.4.1.3 GSR

Overall GSR gave largely inconclusive results except when I examined the responses of critical and non-critical listeners to high quality mixes (C3, C7). There is also a trend when I compare the results for C3 and C7, against the results for critical and non-critical listeners low quality mixes (C4, C8). There are more significant results when I do this comparison as opposed to comparing responses of critical listeners to high and low quality mixes (C1, C5), against responses of non-critical listeners to high and low quality mixes (C2, C6). I also saw this for GEMS-9 and Arousal-Valence-Tension. Thus testing critical versus non-critical listener responses to high versus low quality mixes supported my hypothesis.

#### 6.4.1.4 Head Nod and Shake

Head nod/shake results proved to be conclusive and supported my hypothesis. The difference in nodding is far more apparent for low quality mixes (C4) than high quality mixes (C3). Notably, on low quality mixes, non-critical listeners nodded their heads more than critical listeners. This could mean that non-critical listeners might enjoy the mix regardless of mix quality. I also see something similar for arousal and power where there are slightly more significant p-values for the low quality mixes than for the high quality mixes.

Power, expectation and arousal seem to be divisive features when comparing the types of listeners. Power is based on the sense of control, expectation on the degree of anticipation and arousal on the degree of excitement or apathy [143]. These are features based on tracking emotional cues when conversing with someone, so it is interesting to see them having such an effect during music listening. Having examined the participant's videos I found that since they were sitting in a chair that could rotate, they sometimes moved the chair in time with the music. The classifier detected this as a head shake, which

would normally be viewed as a negative response [101], but in this case it could indicate that the participant is engaged with the music and most likely enjoying it. It is also worth noting that music is very cultural and certain individuals might react differently than others with respect to head nods and shakes.

#### 6.4.1.5 Facial Action Units

Table 6.8 results indicated that the high quality mixes had a greater effect than low quality mixes on the distribution of AU1 and AU4 between the two listener types. AU1 corresponds to inner brow raiser and AU4 corresponds to brow lowering, so this is similar to research on Facial EMG and music, where the brow is associated with the processing of negative events [91, 92]. AU45 corresponds to blinking. There is one more significant AU45 result for condition C4 than there is condition C3, which might imply that there is a difference in intensity of blinking for critical and non-critical listeners.

The percentage total of average AU intensities $>= 0.5$ for AU45 is small, but provided a large amount of significant p-values in Table 6.8. This suggests that the differences in blink intensity between listener type may have been very subtle.

This is the first experiment of its kind that has looked at automatic facial expression recognition and tracking head nod/shakes in a music production quality context. By inspecting the videos I found that some participants were much more expressive in their face than others or might be a lot more inclined to nod and shake their head than use facial expressions. Some critical listeners gazed left or right of the camera, closed their eyes while listening for a prolonged duration, placed their hand under their chin, looked down, looked up, moved their head back and forth, tilted their head or sucked their lip. For non-critical listeners, there were not as many AU's activated, except in one case where the participant was looking away, moving their body on the chair left and right, moving their head back and forth and moving their head left and right. Some stills from the videos can be seen in Figure 6.4, where the top two participants are critical listeners and the bottom two are non-critical listeners.

### 6.4.2 Measures

Self-report measures proved to be the most revealing when comparing mixes and when comparing listener types. I expected the GSR results to be more telling, but found them to be mostly inconclusive. This might have been due to noise in the data as a result of poor electrode contact which is similar to what happened in [89].

FIGURE 6.4: Still images of four participants from the videos made during the experiment. Top two rows are critical listeners and the bottom two are non-critical listeners.

The values for the AUs only became interesting when I looked at the standard deviation. This is expected since someone that is more excited by music tends to be more expressive in their face as the music is played. Head nod/shake detection proved to be

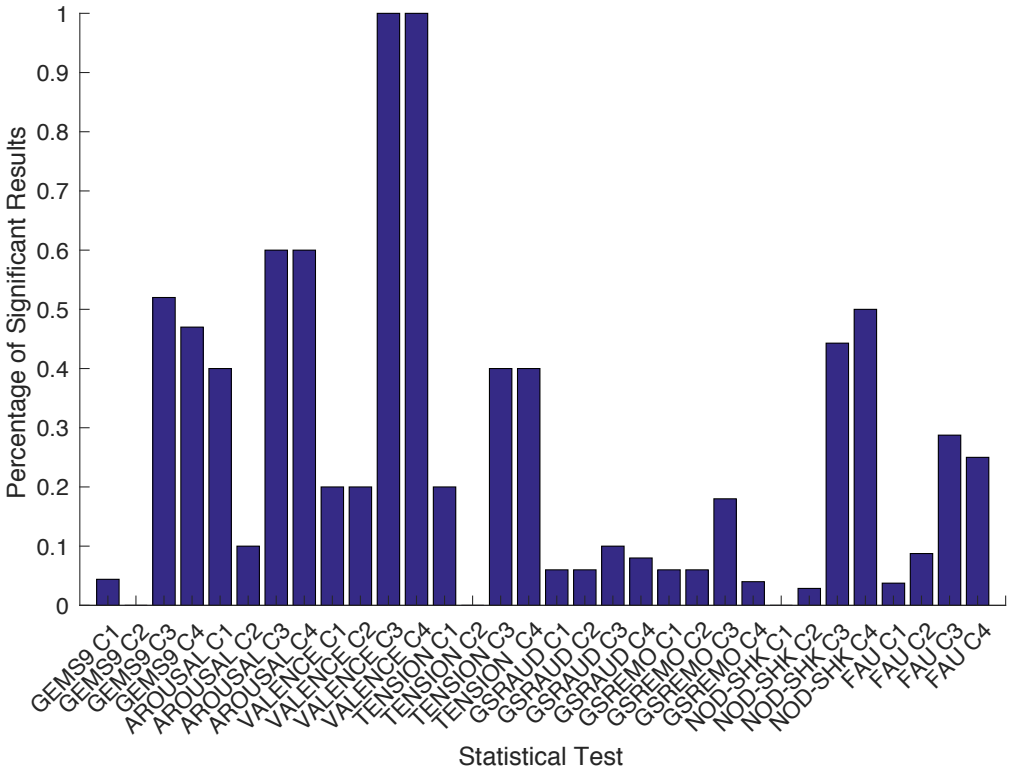


FIGURE 6.5: The percentage of significant results for each statistical test performed for each condition. The highest percentage of significant results occurred for GEMS9 (Induced emotion), Arousal-Valence-Tension (Perceived emotion), Head Nod/Shake and Facial Action Units.

very interesting when comparing the types of listeners. Non-critical listeners nodded their heads more than critical listeners when listening to the poor quality mix, which was something I decided to analyse based on my initial findings in the pilot study.

### 6.4.3 Design

As beneficial as it was to have a pilot study, I learned a lot about experimental design from the main part of the experiment, which could be used to help future studies. One participant reported that most of the emotions that music induces for them comes from the lyrics. They reported that if they disliked the lyrics, then they tended to dislike the song, thus potentially meaning a negative or lack of emotional response. This aspect of music listening may have had an impact on the emotional responses of non-native English speakers. Ten of the participants were non-native speakers and may not have fully understood all lyrics, so this is a confounding variable I had not considered.

Recent research on perceptual evaluation of high resolution audio found that providing training before conducting perceptual experiments greatly improved the reliability of results [149]. In my experiment I provided two training songs, but this was to become

familiar with the experimental interface. However, it could be argued that training would have blurred the distinction between critical and non-critical listeners.

Ideally I would have used songs in the experiment that came from a wider variety of genres. A number of participants were dissatisfied with the songs because they simply did not like the genre. But this was out of my control since I used songs rated in a previous experiment [13]. I would have also liked to have had a bigger sample size for my experiment, to further generalise the results.

I would also suggest that each participant be made sit on a chair that does not rotate or have wheels. When some participants were enjoying a song they tended to move around, which sometimes caused sensors to become dislodged and rendered the acquired data unusable.

## 6.5 Conclusion

My exploratory study provides an insight into the relationship between music production quality and musically induced and perceived emotions. I highlighted some of the challenges with working with physiological sensors and conducting listening tests when trying to measure emotional responses in a musical context. I conducted the first experiment of its kind using facial expression analysis and head nod-shake detection in conjunction with a perceptual listening test.

When I tested to see if critical listeners and non-critical listeners had different emotional responses based on the difference in music production quality, the results were inconclusive for GSR, facial expression and head nod-shake detection. Results strongly agreed with my hypothesis only when I looked at the self-report of perceived emotion.

When I examined just high quality mixes and looked at the difference in emotions of critical and non-critical listeners I found significant p-values in most cases. This was most evident for self-report, head nods/shakes and facial expression. When I examined low quality mixes and looked at the difference in emotions of critical and non-critical listeners I also found a lot of significant p-values, but to a lesser extent than that of the high quality mixes. This was also most evident for self-report, head nods/shakes and facial expression.

The results implied that emotion in a mix, whether induced or perceived, mattered the most to those with critical listening skills, which agrees with my hypothesis. This was most evident from the GEMS-9, Arousal-Valence-Tension, Head Nod/Shake Detection and Facial Action Unit results since they had the most amount of significant p-values.

If one was to take a cynical view, it could be said that using a more professional and experienced mix engineer to mix a piece of music only really matters to those who have been trained to listen for mix defects, and mix quality has little bearing on the layperson emotionally. This is a very important result for audio engineers, specifically in the context of automatic mixing systems and this thesis. The results imply that the perceived quality of an automatically generated mix may not be important to those without critical listening skills. It suggests that automatically generated mixes may be good enough for the general public and casual music listeners. This is something I touch on in the following chapter, where I compare automatically generated mixes with human made mixes. However, all the participants in Chapter 7 had critical listening skills. It would have been interesting to have had some non-critical listener participants to do a comparison.

## 6.6 Future Work

It would be interesting to perform pair-wise ranking between the two mix types, as Likert scales may not be the best tool for affect studies since the values they ask people to rate may mean different things to each participant [150]. However, one argument against pairwise testing is that it is time consuming, e.g. for 10 samples, one might need 10*9/2 comparisons [151, 152].

It would also be interesting to see if I would get similar results when non-critical listeners are provided with training before the experiment i.e. trained to spot common mix defects. This would help identify if the trained non-critical listeners exhibited emotions based on what they think is expected of them due to the training.

I would like to track if a participant is singing along to the music being played, as this could be regarded as a measure of engagement and potential enjoyment of the music. This could be achieved by tracking the Action Units that correspond to the mouth as well as having a microphone near the participant to verify if they were actually singing or not. I would also recommend looking at tracking foot or finger tapping as this is a common form of movement to music [99]. This could be achieved by attaching accelerometers to the participant's feet and placing small piezo contact microphones on their fingertips.

I hope this work will inspire future research. In particular there is a need to use more varied genres of music for evaluation and to see if emotional measures correlate well with low to high level audio features. This could potentially be used in automatic mixing systems such as [2, 18, 133, 153].

# Chapter 7

# Automatic Minimisation of Masking in Multitrack Audio using Subgroups

## 7.1 Introduction

The iterative process of masking minimisation when mixing multitrack audio is a challenging optimisation problem, in part due to the complexity and non-linearity of auditory perception. In this chapter, I first present a multitrack masking metric inspired by the MPEG psychoacoustic model. I investigate different audio processing techniques to manipulate the frequency and dynamic characteristics of the signal in order to reduce masking based on the presented metric. I also investigate whether or not automatically mixing using subgrouping is beneficial or not to perceived quality and clarity of a mix. Evaluation results suggest that the masking metric when utilised in an automatic mixing framework reduces inter-channel auditory masking as well as improves the perceived quality and perceived clarity of a mix. Furthermore, my results suggest that using subgrouping in an automatic mixing framework can also improve the perceived quality and perceived clarity of a mix.

It was shown in Chapter 4 that none of the professional mix engineers created subgroups with the aim of reducing masking. However, the results did show they subgrouped to apply effects such as DRC to many instruments at the same time and to maintain good gain structure. Also, since masking reduction is one of many goals when mixing audio [154]. It was decided to see if combining these two techniques could be used together to mix effectively.

The structure of this chapter is summarised as follows. In Section 7.2 describes the methodology of how I formed an automatic multitrack masking minimisation system and how I conducted the subsequent listening test. In section 7.3 performance evaluations are presented and finally in section 7.4 I discuss the most interesting aspects of the research and outline future directions.

## 7.2 Methodology

### 7.2.1 Research Questions and Hypotheses

The main hypothesis I aim to test is *can my proposed automatic mixing system be used to reduce the amount of auditory masking that occurs in a multitrack mix and subsequently improve its perceived quality.* I also tested two further hypotheses, *can using subgroups when generating an automatic mix improve the perceived quality and clarity of a mix* and *can the use of subgroups in an automatic mixing system have an impact on the perceived emotions of the listener over automatic mixes that do not use subgroups.* These hypotheses were evaluated through examination of the objective performance and subjective listening tests.

### 7.2.2 Automatic Mixing System

There were two types of automatic mixes generated for this experiment, one which made use of subgrouping and one which did not. The mix process is illustrated in Figure 7.1.

### 7.2.3 Audio Processing and Control Parameters

#### 7.2.3.1 Subgrouping

In the multitrack of each song I used for the experiment, I created subgroups based on typically grouped instrumentation such as vocals, drums and guitars etc. This is similar to the approach I developed in chapter 4. This allowed us to use the optimisation mixing technique presented here to create a number of submixes and then create a final mix by mixing each of the submixes together. This essentially gave us a multi-layer optimisation framework. When subgrouping was not used in an automatic mix, the optimisation mixing technique was applied to all the audio tracks at once.

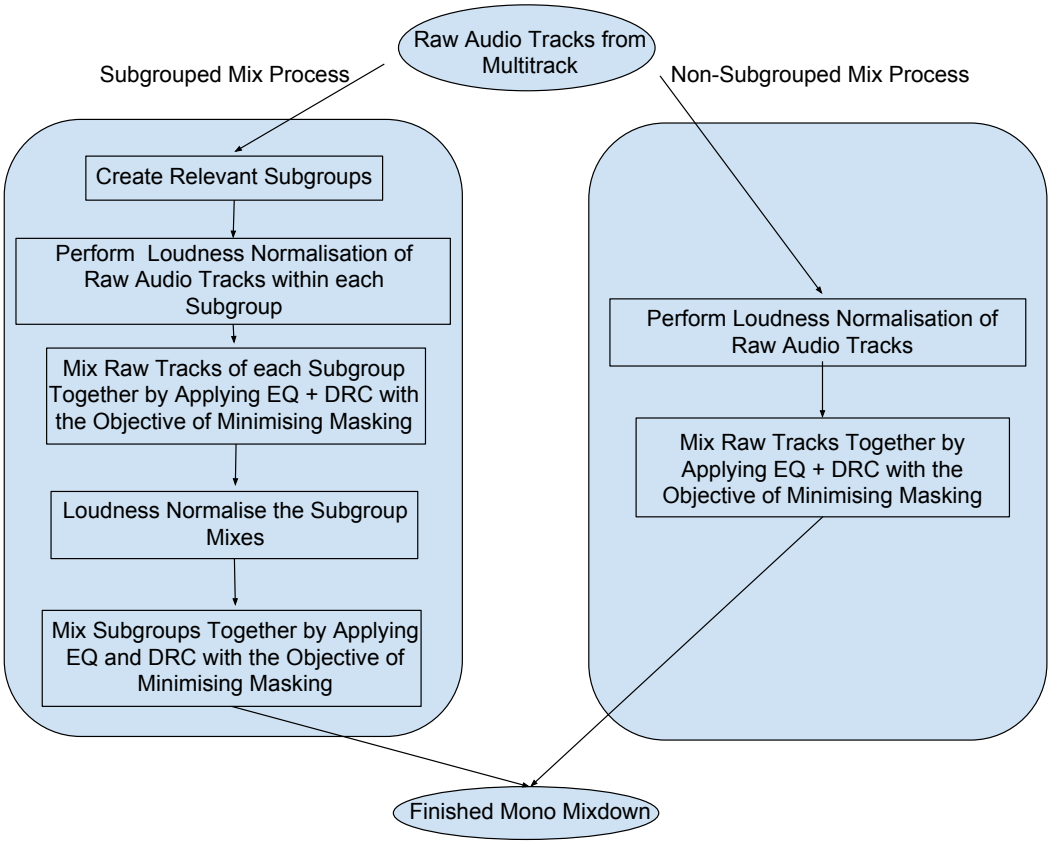


FIGURE 7.1: Automatic mixing process.

#### 7.2.3.2 Loudness Normalisation

Before I applied the optimisation mixing technique I employed loudness normalisation on each audio track in each multitrack. I performed loudness normalisation on all of the audio tracks using the ITU-R BS. 1770-2 specification [138]. Each audio track was loudness normalised to -24 LUFS except in the case of a lead vocal, where it was loudness normalised to -18 LUFS. I made the lead vocal louder than everything else as it is usually the most important audio track within a mix [17]. Once a subgroup had been mixed, it was also loudness normalised to -24 LUFS except in the case of vocal subgroups, which would be set to -18 LUFS. One of the caveats of using this loudness normalisation process is the potential for it to bring up the noise floor and thus may not be the best solution for an automatic mixing system. This effect could potentially be mitigated with an automatic gating system such as the one described in [132]. However, I did not include this in my experiment and is something worth considering in future work.

#### 7.2.3.3 Equalisation

I designed a six-band equaliser to be applied in the optimisation process. Six different cascaded second-order IIR filters were designed to cover the typical frequency range used when mixing. The filter specification is shown in Table 7.1

TABLE 7.1: Six band equaliser filter design specifications

| Band No. | Centre 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 equaliser filter for each track are selected as the control parameters to be obtained through the optimisation procedure. The control parameters in the equalisation cases are given by

$$\boldsymbol{x}_{EQ} = [\boldsymbol{g}_1 \quad \boldsymbol{g}_2 \quad \dots \quad \boldsymbol{g}_n], \tag{7.1}$$

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

$$\boldsymbol{g}_i = [g_{1i} \quad g_{2i} \quad \dots \quad g_{6i}], \tag{7.2}$$

contains the six gain controls for each track.

#### 7.2.3.4 Dynamic Range Compression

The digital compressor model employed in my approach was a feed-forward compressor with smoothed branching peak detector [61]. A typical set of parameters of a dynamic range compressor includes the Threshold, Ratio, Attack and Release Times, and Make-up gain. In the case of adjusting the dynamic of the signal to reduce masking through optimisation, the values of threshold ($T$), ratio ($R$), attack ($a$) and release ($r$) are control parameters to be optimised. Since dynamics are my main focus here rather than the level, the make-up gain of each track is set to compensate the loudness differences (measured by EBU loudness standard [138]) before and after dynamic processing. The make-up gain for each track is given by

$$g_{\Delta i} = L_{EBUi} - L'_{EBUi}, \tag{7.3}$$

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

$$\boldsymbol{x}_{DRC} = [\boldsymbol{d}_1 \quad \boldsymbol{d}_2 \quad \dots \quad \boldsymbol{d}_n] \tag{7.4}$$

Similarly, every $\boldsymbol{d}_i$ is constituted of four standard DRC control parameters denoted as, threshold ($T_i$), ratio ($R_i$) attack ($a_i$), release ($r_i$).

$$\boldsymbol{d}_i = [T_i \quad R_i \quad a_i \quad r_i] \tag{7.5}$$

#### 7.2.3.5 Control Parameters

The notation of the final control parameters to be optimised in the multitrack masking minimisation process is given by

$$\boldsymbol{x}_C = [\boldsymbol{c}_1 \quad \boldsymbol{c}_2 \quad \dots \quad \boldsymbol{c}_n], \tag{7.6}$$

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

$$c_i = \begin{pmatrix} g_{1,i} & \dots & g_{6,i} & T_i & R_i & a_i & r_i \end{pmatrix} \tag{7.7}$$

### 7.2.4 Masking Metric

#### 7.2.4.1 MPEG Psychoacoustic Model

Audio coding or audio compression algorithms compress the audio data in large part by removing the acoustically irrelevant parts of the audio signal. The MPEG psychoacoustic model [155] 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 analyses the signal, and computes the masking thresholds as a function of frequency [29, 155, 156]. The block diagram in Figure 7.2 illustrates the simplified stages involved in the psychoacoustic model.

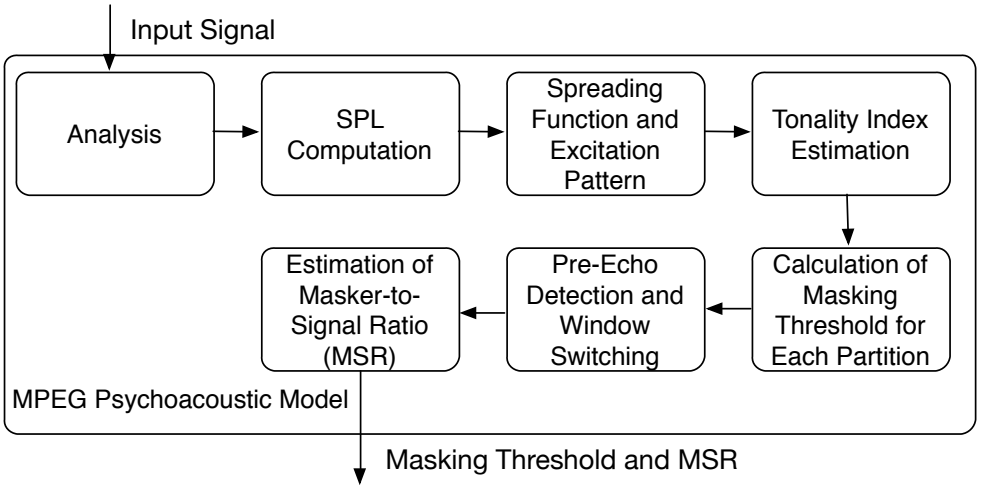


FIGURE 7.2: Flowchart of the MPEG psychoacoustic model [155].

The procedure to derive masking thresholds is summarised as follows. The complex spectrum of the input signal is calculated using a standard forward FFT. A tonality index as a function of frequency is calculated based on the local peaks of the audio power spectrum. This index gives a measure of whether a component is more tone-like or noise-like. This index is then interpolated between pure tone-masking-noise and noise-masking-tone values. The tonality index is based on a measure of predictability, where tonal components are more predictable and thus will have higher tonality indices [157].

A strong signal component reduces the audibility of weaker components in the same critical band and also the neighbouring bands. The psychoacoustic model emulates this by applying a spreading function to spread the energy of a critical band across other bands. The total masking energy of the audio frame is derived from the convolution of the spreading function with each of the maskers. The spreading function, $s_f$ (measured in dB) used in this model is given by

$$s_f(i,j) = \begin{cases} 0 & B(z) \leq 0 \\ x^{\frac{x+B(d_z)}{10}} & \text{else} \end{cases} \tag{7.8}$$

where the calculation of $B(d_z)$ can be found in [31]. $d_z$ is the bark distance between maskee and masker. Conversion between bark scale and frequency Hz can be approximated by

$$z(f) = 13 \arctan(0.00076f) + 3.5 \arctan((f/7500)^2). \tag{7.9}$$

The spreading function is then convolved with the partitioned, re-normalised energy to derive the excitation pattern 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. The tonality indices evaluated for each partition are used to determine the offset of the re-normalised convolved signal energy [155], which converts it into the global masking level. The values for the offset are interpolated based on the tonality index of a noise masker to a frequency-dependent value defined in the standard for a tonal masker. The interpolated offset is compared with a frequency dependent minimum value, *minval*, defined in the MPEG-1 standard and the larger value is used as the signal to noise ratio. In the standard, Noise Masking Tone is set to 6 dB and Tone Masking Noise to 29 dB for all partitions. The offset is obtained by weighting the maskers with the estimated tonality index. The partitioned threshold derived for the current frame is compared with that of the two previous frames and the threshold in quiet. The maximum of three values is chosen to be the actual threshold.

The energy in each scale-factor band, $E_{sf}(sb)$ and the threshold in each scale-factor band, $T(sb)$ are calculated as described in [31], in a similar way. Thus the final masker-to-signal ratio (MSR) in each scale-factor band is defined as

$$\mathrm{MSR}(sb) = 10\log_{10}(\frac{T(sb)}{E_{sf}(sb)}) \tag{7.10}$$

#### 7.2.4.2 Cross-adaptive MPEG Masking Metric

I adapt the masking threshold algorithm from MPEG audio coding into a multitrack masking metric based on a cross-adaptive architecture [2, 51]. The flowchart of the system is illustrated in Figure 7.3.

To account for the masking that is imposed on an arbitrary track by the other accompanying tracks rather than by itself, I replace $T(sb)$ with $T'(sb)$, 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 to derive the masking threshold. I thus can compute $T'(sb)$ as

$$T'_n(sb) = H(\sum_{i=1, i\neq n}^{N} s_i) \tag{7.11}$$

$E_{sf,n}(sb)$ denotes the energy at each scale-factor band of track $n$. I assume masking occurs at any scale-factor band where $T'_n(sb) > E(sb)$. The masker to signal ratio in multitrack content becomes

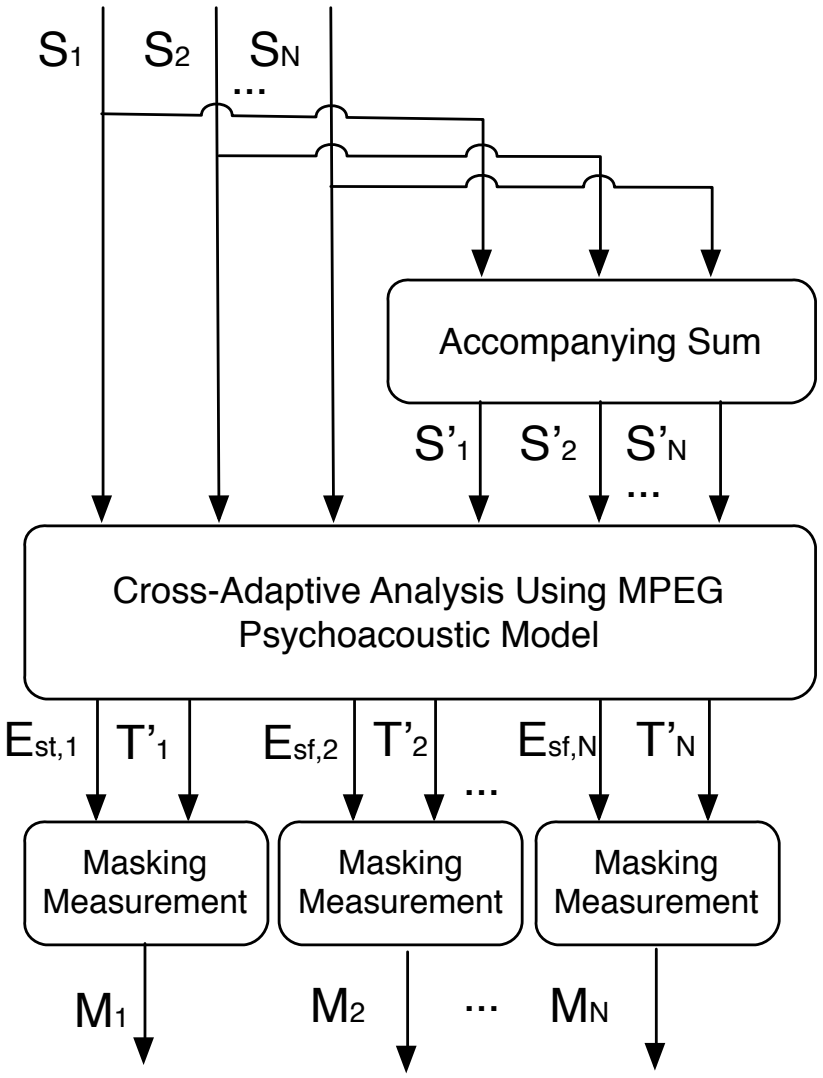


FIGURE 7.3: System flowchart of proposed cross-adaptive multitrack masking model. The multitrack consists of $N$ sources that have been pre-recorded onto $N$ tracks. Track $n$ therefore contains the audio signal from source $n$, given by $s_n$ and $s'(n) = \sum_{i=1, i\neq 1}^{N} s_i$. $T'_n$ is defined in Eq. 7.11 and $E_{st,n}$ is the energy in each scale-factor band. These are subsequently used to calculate $M_n$ in Eq. 7.13

$$\mathrm{MSR}_n(sb) = 10 \log_{10} \frac{T'_{sb}}{E_{sf,n}(sb)} \tag{7.12}$$

I then can define a cross-adaptive multitrack masking, $M_n$ as

$$M_n = \sum_{sb \subset E_{sf,n} < T'_n} \frac{\mathrm{MSR}_n(sb)}{T_{max}} \tag{7.13}$$

where $T_{max}$ is the predefined maximum amount of masking distance between $T(sb)$ and $E_{sf}(sb)$ for each scale-factor band, which is set to 20 dB.

### 7.2.5 Numerical Optimisation Algorithm

The multitrack masking minimisation process is treated as an optimisation problem concerned with minimising a vector-valued objective function described by the masking metric. 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 the error of the objective function is within a tolerance value (0.05), reaches the maximum number of iterations or the masking metric is reduced to zero.

#### 7.2.5.1 Function Bounds

The minimum and maximum values I used for the 6-band equaliser and the dynamic range compressors were set based on audio engineering literature and having consulted a professional practitioner in the audio engineering field [1, 7, 17, 133]. These are detailed in Table 7.2.

TABLE 7.2: The minimum and maximum values used for the different types of audio processing used during the optimisation procedure.

| **Audio Process** | **Min Value** | **Max Value** |
|---|---|---|
| Instrument EQ Gain Bands 1- 6 | -6 db | + 6 db |
| Subgroup EQ Gain Bands 1- 6 | -3 db | + 3 db |
| Instrument DRC Ratio | 1 | 6 |
| Subgroup DRC Ratio | 1 | 6 |
| Instrument DRC Threshold | -30 db | 0 db |
| Subgroup DRC Threshold | -30 db | 0 db |
| Instrument DRC Attack | 0.005 secs | 0.25 secs |
| Subgroup DRC Attack | 0.005 secs | 0.25 secs |
| Instrument DRC Release | 0.005 secs | 3 secs |
| Subgroup DRC Release | 0.005 secs | 3 secs |

I used smaller minimum and maximum equalisation gains when I was mixing the subgroups together, since the majority of the inter-channel auditory masking would have been removed when mixing the individual instrument tracks.

#### 7.2.5.2 Objective Function

A numerical optimisation approach was used in order to derive an optimal set of inputs which would result in a balanced mix. Before defining the objective functions a number of parameters are defined which were used with the optimisation algorithm.

Let $A$ denote the total number of tracks in the multitrack and $K$ denote the total number of the control parameters. The masking metrics are given by $M_i(\boldsymbol{x}_C)$, for $i = 1, \ldots, n$. These describe the amount of masking in each track as a function of the control parameters $\boldsymbol{x}_C$. Note that $\boldsymbol{x}_C$ represents the whole set of the control parameters for all tracks. The values of $\boldsymbol{x}_C$ tend to have multitrack influences, due to the complexity and non-linearity of the perception of masking. Changes in the control parameter for one track not only affect the masking of that particular track itself but also masking of all other tracks.

The total amount of masking, $M_T(\boldsymbol{x}_C)$, can be expressed as the sum of squares of $M_i(\boldsymbol{x}C)$, for $i = 1, \ldots, n$,

$$M_T(\boldsymbol{x}_C) = \sum_{i=1}^{A} M_i^2(\boldsymbol{x}_C) \tag{7.14}$$

It is desired to minimise the sum of the masking across tracks and so (7.14) can be used as the first part of the objective function.

The second objective is that the masking is balanced, i.e., there is not a significant difference between masking levels. Here a maximum masking difference based objective is formed as follows:

$$\begin{aligned} M_d(\boldsymbol{x}_C) = \max(\parallel M_i(\boldsymbol{x_C}) - M_j(\boldsymbol{x_C}) \parallel), \\ \text{for } i = 1, \ldots, n, j = 1, \ldots, n, i \neq j \end{aligned} \tag{7.15}$$

This allows this second part of the objective to be used within a min-max framework, similar to that used in [158].

Combining the two objective functions, the following optimisation problem is solved to give $\boldsymbol{x}_C$:

$$\boldsymbol{x}_C = \min_{\boldsymbol{x}_C} M_T(\boldsymbol{x}_C) + M_d(\boldsymbol{x}_C) \tag{7.16}$$

The optimisation problem is a nonlinear, non-convex formulation, and the only information available to the optimisation routine were returns of the function values. Thus a Particle Swarm Optimisation (PSO) approach was used to guide the optimisation routine about the solution space. The Levenberg-Marquardt algorithm was considered for this optimisation since the problem was non-linear. However, it was found that using it

was much slower and did not always give a global optimal solution. A similar optimisation approach to mixing was used in [9], where they used the *Gauss Newton* optimisation method.

### 7.2.6 Experiment Setup

#### 7.2.6.1 Participants

Twenty four participants, all of good hearing, were recruited. 20 were male, 4 were female and their ages ranged from 23 to 52 ($\mu = 30.09, \sigma^2 = 6.2$). All participants had some degree of critical listening skills, i.e, the participant knew what critical listening involved and had been trained to do so previously or had worked in a studio.

#### 7.2.6.2 Stimuli

There were five songs used in the experiment, where there were five different 30 sec. mono mixes of each song. Two of the mixes were automatically generated using my proposed mix algorithm, where one mix used subgroups and the other did not. There was one mix that was just a straight sum of all the raw audio tracks. Finally, there were two human mixes, where I selected the low quality mix and high quality mix of each song as determined from a previous experiment. The human mixes were created using standard audio processing tools available in Pro Tools, where I was able to get each mix without the added reverb [13]. The mixes were created with the intention of producing the best possible mix. The songs were sourced from the Open Multitrack Testbed [137]. I loudness normalised all of the mixes using the ITU-R BS. 1770-2 specification [138] to avoid bias towards mixes which were louder than others. The song name, genre, number of tracks, number of subgroups and how many of each instrument type there were is shown in Table 7.3

TABLE 7.3: The audio tracks names, genre types, total number of tracks mixed, number of subgroups mixed and the total number of individual instrument tracks mixed.

| **Track Name** | **Genre** | **Tracks** | **Subgroups** | **Drums** | **Vox** | **Bass** | **Keys** | **Guitars** |
|---|---|---|---|---|---|---|---|---|
| In the Meantime | Funk | 24 | 5 | 10 | 6 | 1 | 4 | 2 |
| Lead Me | Pop-Rock | 19 | 5 | 9 | 2 | 1 | 2 | 5 |
| Not Alone | Funk | 24 | 5 | 8 | 9 | 1 | 4 | 2 |
| Red to Blue | Pop-Rock | 14 | 4 | 9 | 1 | 1 | 0 | 3 |
| Under a Covered Sky | Pop-Rock | 25 | 5 | 9 | 5 | 1 | 2 | 8 |

#### 7.2.6.3 Pre-Experiment Questionnaire

I provided a pre-experiment questionnaire. The pre-experiment questionnaire asked simple questions related to age, hearing, musical experience, music production experience, music genre preference and each participant's confidence in their critical listening skills. There was also a question with respect to how tired they were when they started the study. If any participant indicated that they were very tired, I asked them to attempt the experiment at a later time once they were rested.

#### 7.2.6.4 Tasks

I explained to each participant how the experiment would proceed. They were also supervised during the experiment in the event a participant was unsure about anything.

There were two experiment types, where half the participants did experiment type 1 (E1) and the other half did experiment type 2 (E2). Each experiment type had two parts, where the second part was common to both. In E1 (i), I required the participants to rate each of the five mixes of each song they listened to in terms of their preference. In E2 (i), I required the participants to rate each of the five mixes of each song they listened to in terms of how well they could distinguish each of the sources present in the mix (Mix Clarity). In E1 (ii) and E2 (ii) each participant had to listen and compare the automatically generated mixes. They then had to each rate mix for their perceived emotion of each mix along three scales. The scales were Arousal, Valence and Tension (A-V-T). All the songs and mixes used in the experiment were presented in random in order.

After all mixes were rated, participants were asked to provide some feedback on how the experiment was conducted and what their impressions were of the mixes they heard.

#### 7.2.6.5 Setup and User Interface

The experiment either took place in a dedicated listening room at the university or at an external music studio environment. Each participant was sat at a studio desk in front of the laptop used for the experiment. The audio was heard over either a pair of PMC AML2 loudspeakers or Sennheiser HD-25 headphones, where the participant could adjust the volume of the audio to a comfortable level.

Mix preference and self-report scores were recorded into a bespoke software program developed for this experiment. The software was designed to allow the experiment to

run without the need for assistance, and the graphical user interface was designed to be as aesthetically neutral as possible, so as not to have any effect on the results.

In this section I present the results related to the optimisation procedure used to generate the automatic mixes. Furthermore, I present the results of the subjective evaluation of the automatic mixes, where the mixes were rated for preference, clarity and the participant's perceived emotion. I have placed all the mixed and unmixed audio used in this experiment in an online repository at `https://goo.gl/U2F3ed`.

## 7.3 Results of Optimised Automatic Mixing

In Figure 7.4 I present the results of the optimisation process used to mix "In the Meantime", for mixing each of the different subgroups, mixing the subgroups and mixing all the tracks together as one. The $x$-axis on the graph indicates how many iterations of the optimisation process occurred before a solution was found. The $y$-axis indicates masking was present. The results for the other four songs analysed follow a similar trend.

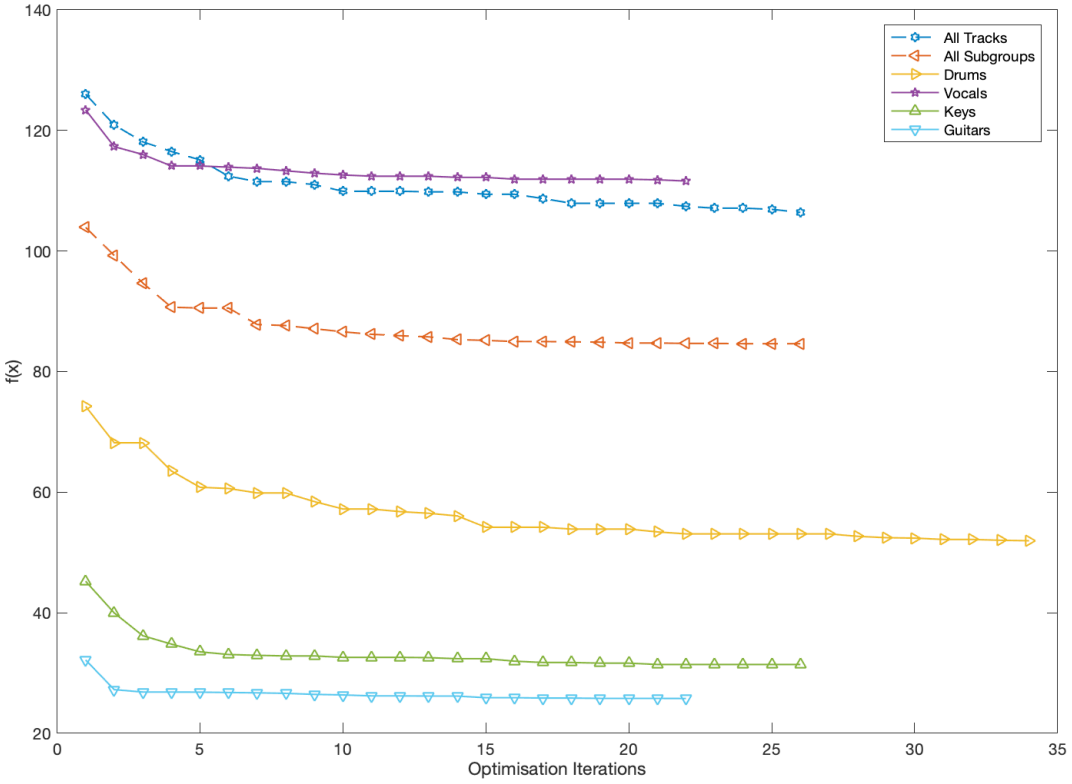


FIGURE 7.4: Cost function value ($f(x)$) for "In The Meantime" plotted against the number of optimisation function iterations. "All Tracks" is the optimisation process when mixing all the tracks together at once. "All Subgroups" is the optimisation process when mix all the individual subgroup types together. The different instrument types such as "Drums", "Vocals", "Keys" and "Guitars" are the instrument submixes.

When the vocal tracks (Vocals) were being mixed, the amount of inter-channel masking that occurred was similar to that of all the tracks being mixed (All Tracks), but took less time to find an optimal solution. This suggests that a lot of the inter-channel masking occurred among the vocalists.

As expected, subgroups with fewer tracks generally took less iterations to converge. Drums were the instrument type which took the most iterations to converge, with the exception of "Lead Me". This is only partly explained by the number of sources in the drums subgroup, since it often took more iterations than when mixing all raw tracks.

I summarise these results in Figure 7.4. In this table I present how many iterations were required to mix each type of each song, the change in masking that occurred and the average amount of masking that remained. The numbers in parentheses are the number of tracks used to do the average calculation. It is clear that applying subgroups to generate stems rather than raw tracks results in fewer iterations and a greater overall reduction in masking.

TABLE 7.4: Number of optimisation iterations required, the change in masking $M$, and the average masking $M$ where the number of tracks mixed is in brackets.

| | No. Iter | $\Delta M$ | $\mu M$ |
|---|---|---|---|
| In the Meantime - All Tracks | 26 | 19.6 | 4.43 (24) |
| In the Meantime - Subgroups | 25 | 19.28 | 16.92 (5) |
| Lead Me - All Tracks | 31 | 35.3 | 6.37 (19) |
| Lead Me - Subgroups | 25 | 16.98 | 18.66 (5) |
| Not Alone - All Tracks | 26 | 27.1 | 6.81 (24) |
| Not Alone - Subgroups | 24 | 19 | 20.56 (5) |
| Red to Blue - All Tracks | 37 | 39.6 | 7.7 (14) |
| Red to Blue - Subgroups | 24 | 17.6 | 26.13 (4) |
| Under a Covered Sky - All Tracks | 51 | 45.4 | 25 (4.82) |
| Under A Covered Sky - Subgroups | 25 | 18.57 | 19.85 (5) |

### 7.3.1 Subjective Evaluation Results

#### 7.3.1.1 Mix Preference

I asked half of the participants to rate each mix based on their preference (E1). The results are illustrated in Figure 7.5.

In Figure 7.5 we see the results for each of the five songs used in the experiment, where they are organised by mix type. The figure shows the mean values across all participants, where the red boxes are the 95% confidence intervals and the thin vertical lines represent 1 standard deviation. The songs are ordered for each mix type as follows:

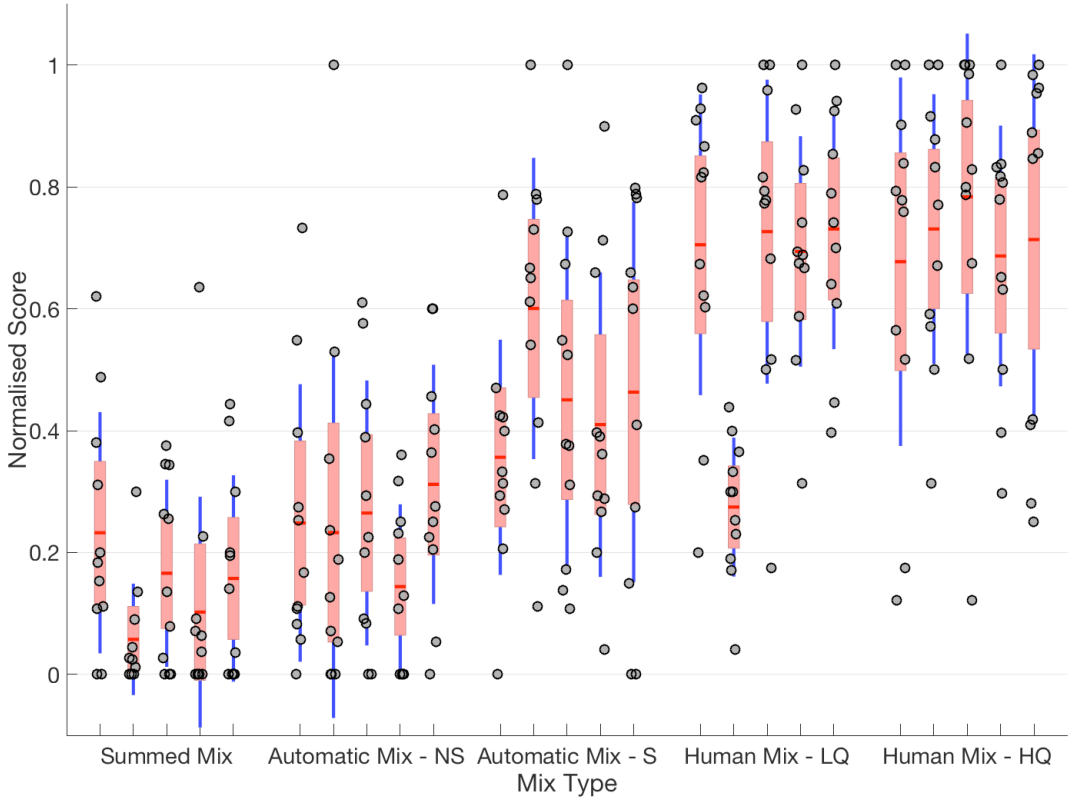


FIGURE 7.5: Results for mix preference based on mix type for each of the individual songs (E1). The songs are ordered for each mix type as follows: "In the Meantime", "Lead Me", "Not Alone", "Red to Blue" and "Under a Covered Sky".

"In the Meantime", "Lead Me", "Not Alone", "Red to Blue" and "Under a Covered Sky".

The mean scores for the summed mixes hover around 0.2, and were never greater than any of the corresponding automatic mixes. However, we see overlapping confidence intervals for all the summed mixes and the automatic mixes without subgroups. Furthermore, there is also some slight overlap with the automatic mixes that use subgroups, but it is not prevalent.

When we compare the two automatic mix types for each song, we see that the automatic mixes that used subgroups were preferred more on average than the automatic mixes that did not use subgroups. This supports my main hypothesis about subgroups improving the perceived mix quality of an automatic mix. However, we see overlapping confidence intervals for "In the Meantime", "Not Alone" and "Under a Covered Sky".

On comparing the automatic mixes to the human mixes, we see the human mixes outperforming the automatic mixes in nearly all cases except for "Lead Me". In the case of "Lead Me", the automatic mix with subgrouping scores 0.6 on average, while the

human low quality mix scores 0.27. There are also overlapping confidence intervals between "Lead Me" for mix types Automatic Mix - S and Human Mix - HQ, "Not Alone" for mix types Automatic Mix - S and Human Mix - LQ and "Under a Covered Sky" for mix types Automatic Mix - S and Human Mix - HQ.

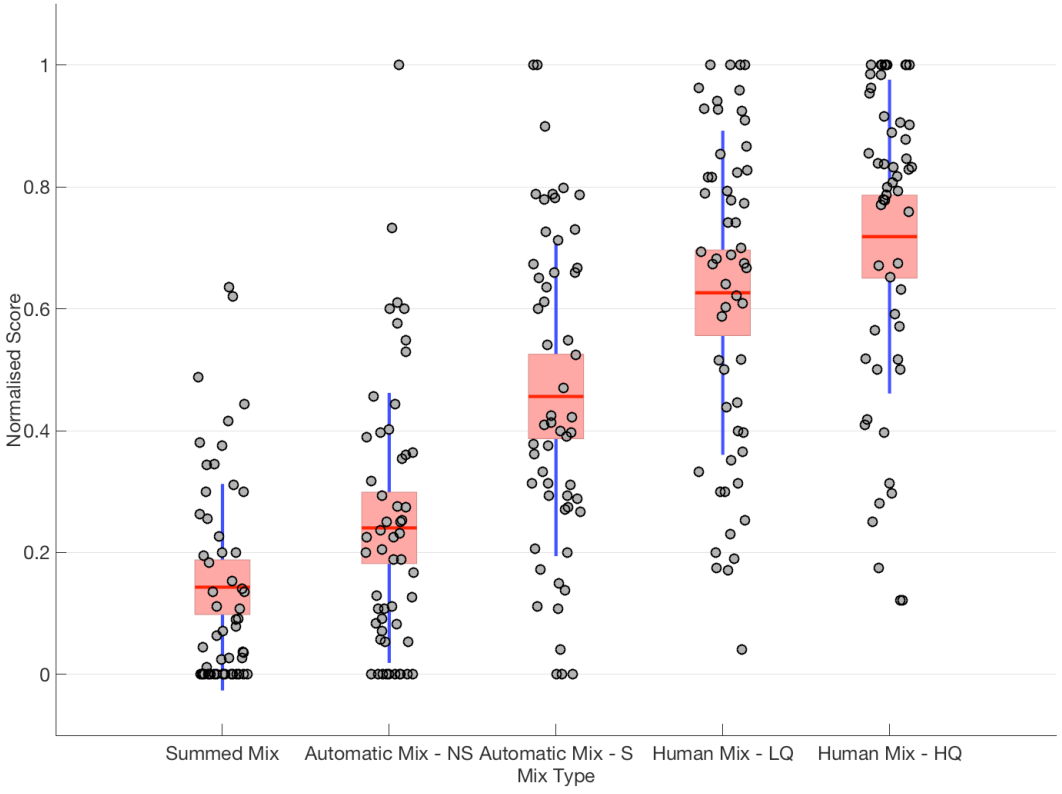


FIGURE 7.6: Results for mix preference based on mix type for all songs (E1).

In Figure 7.6 we see the results for each of the individual mixes, but where we have taken mean across all the different songs. The red boxes are the 95% confidence intervals and the thin vertical lines represent 1 standard deviation. We see there is a trend in increasing means going from Summed mix all the way to Human Mix - HQ. It is apparent that the automatic mixes have performed better than the summed mixes, which supports my main hypothesis. However, there is very slight confidence interval overlap between Summed Mixes and Automatic Mix - NS. In support of my second hypothesis we can clearly see that there is a preference for the mixes that use subgroups. However, we do not see any confidence interval overlap with either of the human mix types.

#### 7.3.1.2 Mix Clarity

I also asked the other half of all the participants to rate the mixes in terms of perceived clarity (E2). The results are illustrated in Figure 7.7.

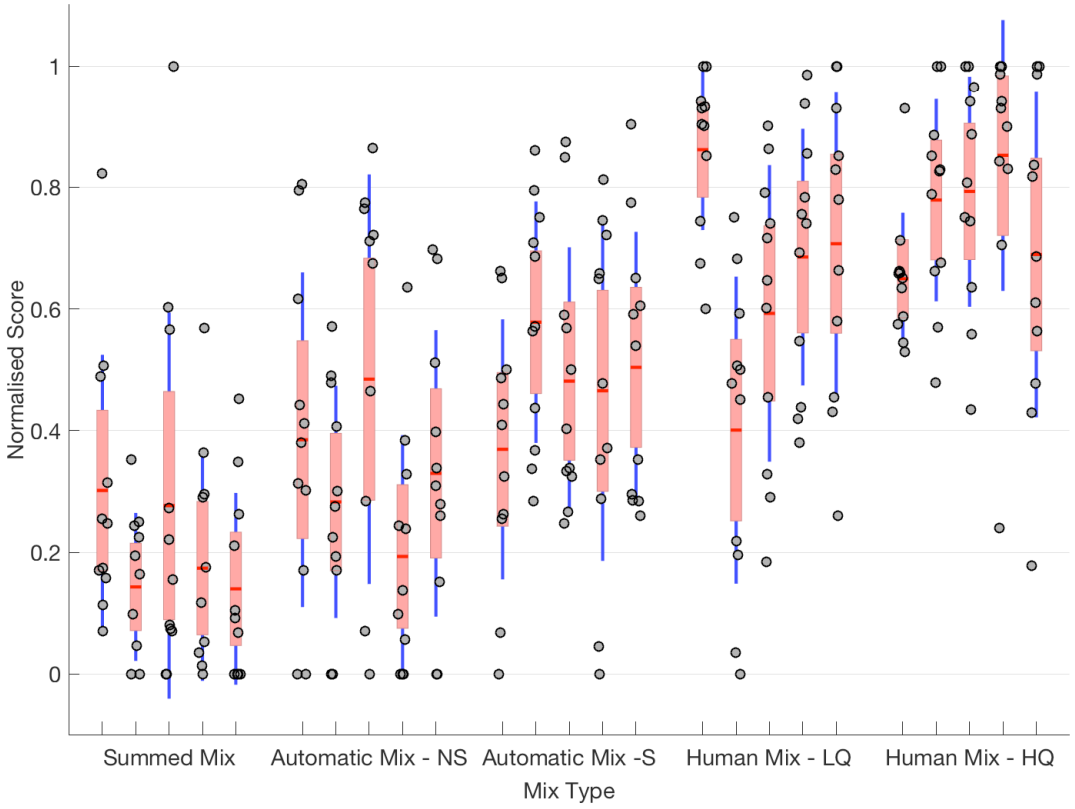


FIGURE 7.7: Results for mix clarity based on mix type for each of the individual songs (E2). The songs going from left to right for each mix type are "In the Meantime", "Lead Me", "Not Alone", "Red to Blue" and "Under a Covered Sky".

In Figure 7.7 we see the results for each of the five songs used in the experiment, where they are organised by mix type. The results are illustrated similarly to Figure 7.5.

As in Figure 7.5, the mean scores for the summed mixes are never greater than any of the corresponding automatic mixes. This indicates that the automatic mixes were perceived to have greater clarity on average than the summed mixes. However, we do see overlapping confidence intervals for all the summed mixes and the automatic mixes without subgroups. Furthermore, this also occurred for the songs "In the Meantime" and "Red to Blue" when we compared Summed mix to Automatic Mix - S.

When we compare the two automatic mix types for each song, we see that the automatic mixes that used subgroups had a better clarity rating on average than the automatic mixes that did not use subgroups in only three of the five songs. We also see overlapping confidence intervals for four of the five songs.

On comparing the automatic mixes to the human mixes, we see the human mixes outperforming the automatic mixes in nearly all cases except for "Lead Me". In the case of "Lead Me", the automatic mix with subgrouping scores 0.58 on average, while the low quality mix scores 0.4. There are also overlapping confidence intervals between "Lead

Me" for mix types Automatic Mix - NS and Human Mix - LQ, "Lead Me" for mix types Automatic Mix - S and Human Mix - HQ and "Under a Covered Sky" for mix types Automatic Mix - S and Human Mix - HQ.

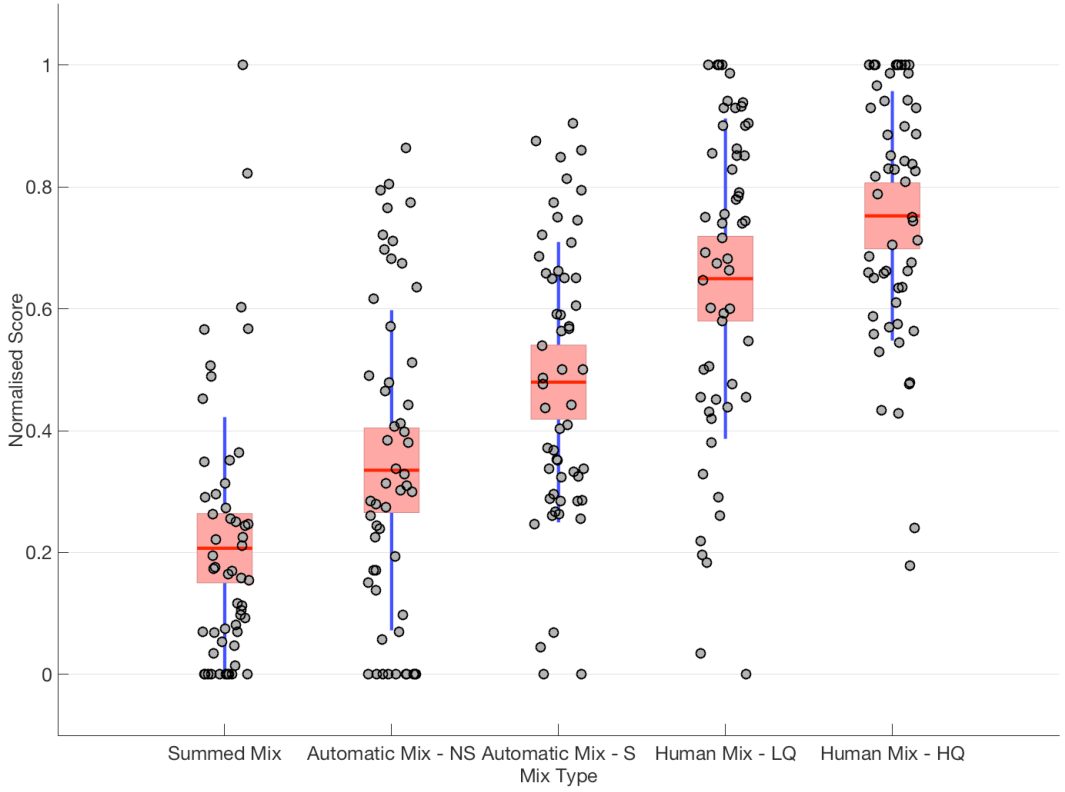


FIGURE 7.8: Results for mix clarity based on mix type for all songs (E2).

Again we see in Figure 7.8 there is a trend in increasing means going from Summed mix all the way to Human Mix - HQ. It is apparent that the automatic mixes have performed better than the summed mixes in terms of clarity. This supports my main hypothesis that I am reducing auditory masking as per Eq. 7.16, which reduces the masking in each individual track while keeping the masking reduction balanced between each track. And in support of my second hypothesis, there is a preference in terms of clarity for the mixes that use subgroups.

#### 7.3.1.3 Perceived Emotion

I asked each of the participants to listen to all the the automatic mixes with subgroups and without subgroups side by side. This was so that they could indicate if they could perceive an emotional difference between each of the two mixes along the three affect dimensions: arousal, valence and dominance. I used the results to test the hypothesis that using subgroups can have an emotional impact on the perceived emotions of the listener. I found my hypothesis to be true in only 1 out of 15 cases (5 songs measured

along 3 affect dimensions). The one significant result I found is illustrated in Figure 7.9. I tested all of the data using the Wilcoxon signed rank statistical test.

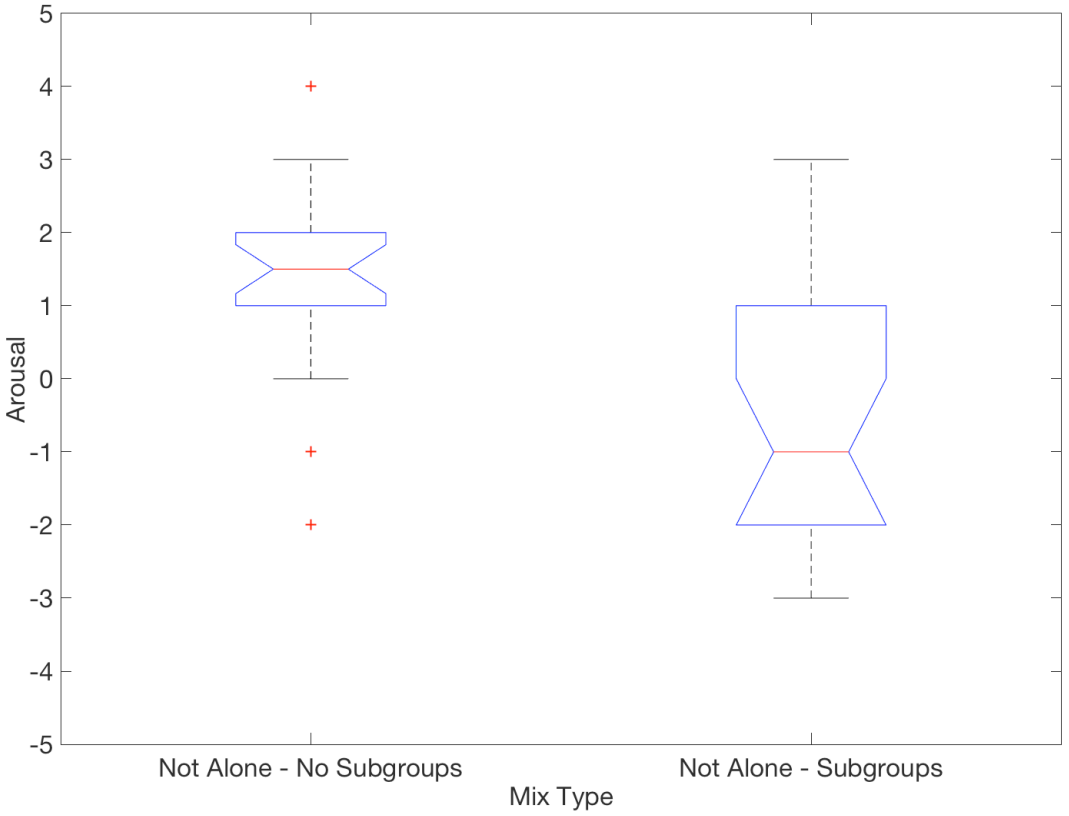


FIGURE 7.9: Box plot of perceived arousal for "Not Alone". This plot illustrates that there was a significant difference in perceived arousal for the two different mix types of this song. One mix was created using subgroups, the other did not.

### 7.3.2 Summary

Table 7.4 and Figure 7.4 objectively show that my proposed intelligent mixing system is able to reduce the amount of inter-channel auditory masking that occurs by changing the parameters of the equaliser and dynamic range compressor on each audio track. In all mixing cases it was able to reduce the amount of inter-channel masking after a few iterations of the optimisation procedure. Table 7.4 shows that the reduction in masking was significantly less in four out of the five songs when mixing Subgroups versus All Tracks. This suggests a lot of the masking had been reduced when mixing the subgroups, where the instrumentation would have been similar.

In Figure 7.10 I present the mean score for each mix type for each of the participating groups, where group 1 evaluated each mix for preference and group 2 evaluated the mixes for clarity. We see that the automatic mixes were preferred more on average than

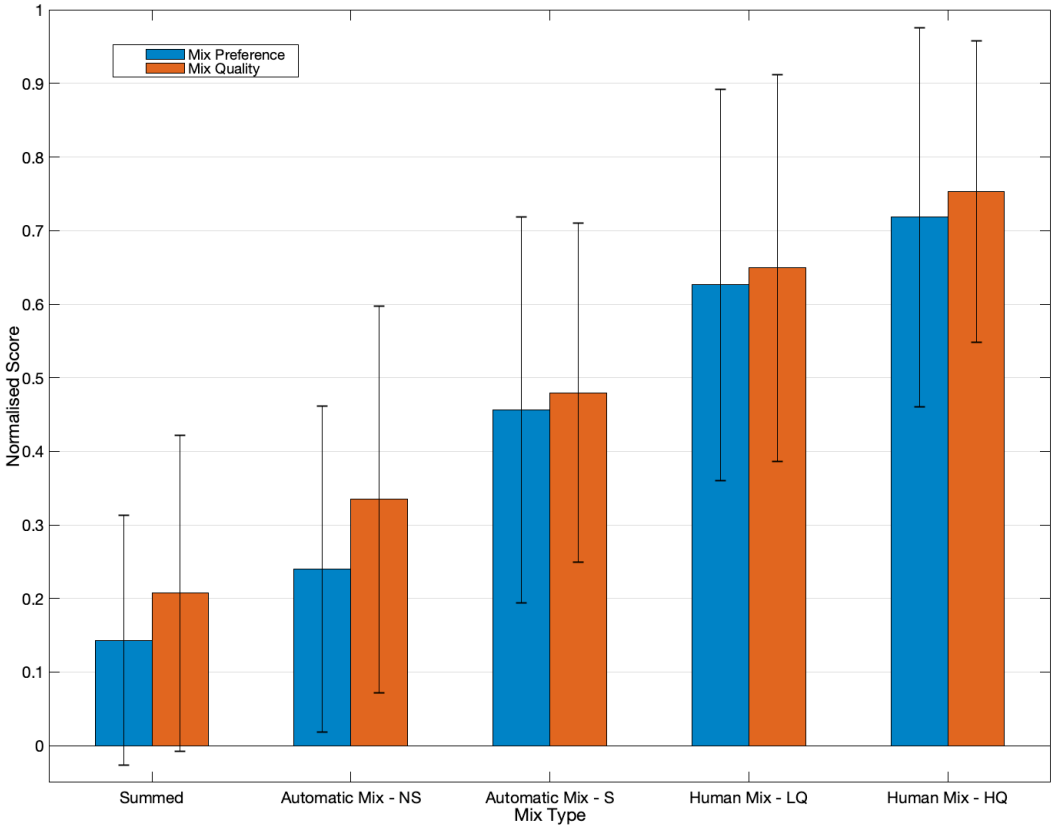


FIGURE 7.10: Mean and standard deviation scores of each mix type for each group, where the blue bars represent mix preference and the red bars represents mix clarity

the summed mixes, which agrees with my main hypothesis. However, the automatic mixes never outperformed the human mixes. We also see that the automatic mixes that used subgroups were preferred more on average than the automatic mixes that did not use subgroups. This supports my second hypothesis. However, there were three cases of overlapping confidence intervals. Figure 7.10 does not show any evidence my second hypothesis is true.

When we examine the results for Group 2, which are denoted by the light coloured bars in Figure 7.10, we see that the automatic mixes were preferred more on average than the summed mixes for clarity, which agrees with my main hypothesis. The results do not show any evidence my proposed de-masking method provides any more clarity to a mix than a human can on average. However, one automatic mix with subgroups performed better than human mix. Also, there were overlapping confidence intervals for two automatic mixes and two human mixes with respect to clarity. We see that the automatic mixes that used subgroups had better perceived clarity on average than the automatic mixes that did not use subgroups. This supports my second hypothesis. However, when we examined the clarity results for the individual songs this only occurred for three songs and there were overlapping confidence intervals for four songs.

The results for the mix clarity group are higher on average than the mix preference group. This might suggest that the technique presented here might be better just as a de-masking technique than an overall mixing technique or just that people are more likely to give higher marks for the word "Clarity" than for the word "Preference".

I was only able to show there was a significant difference in perceived emotions for 1 out of the 15 cases tested. This suggests out third hypothesis cannot be accepted to be true.

## 7.4 Conclusion

This chapter described the automation of loudness normalisation, equalisation and dynamic range compression in order to improve the overall quality of a mix by reducing the inter-channel auditory masking. I adapted and extended the masking threshold algorithm of the MPEG psychoacoustic model in order to measure inter-channel auditory masking. Ultimately, I proposed an intelligent system for masking minimisation using a numerical optimisation technique. I tested the hypothesis that my proposed intelligent system can be used to generate an automatic mix with reduced auditory masking and improved perceived quality. This paper also tested the hypothesis that using subgroups when generating an automatic mix can improve the perceived mix quality and clarity of a mix. I further tested to see if using subgrouping or not affects the perceived emotion in an automatic mix. I evaluated all my hypotheses through a subjective listening test.

I was able to show objectively and subjectively that the novel intelligent mixing system I proposed reduced the amount of inter-channel auditory masking that occurred in each of the mixes and it improved the perceived quality. However, the results did not match the results of the human mixes in most cases.

Furthermore, the results of the subjective listening test implied that subgrouping improves the perceived quality and perceived clarity in an automatic mix over automatic mixes that do not use subgroups. However, the results suggested that using subgroups had very little effect if any on the perceived emotion in any of the mixes. It was only shown to be true in 1 out of the 15 cases.

## 7.5 Future Work

It is clear that my proposed intelligent mixing system has scope for improvement. One way in which this could be improved is if the equalisation and dynamic range compression settings changed on a frame by frame based on the inter-channel auditory masking

metric. Currently the equalisation and dynamic range settings are static for the entire track. One of the more experienced participants in the subjective listening test mentioned that they could hear this.

I also believe the optimisation procedure could be improved by having a larger optimality tolerance, where once this tolerance has been reached another nonlinear solver begins, using the PSO results as initial conditions. If we examine Figure 7.4 we see that many of the optimisation procedures find a satisfactory solution in less than ten iterations.

I would also like to see this intelligent system used in combination with panning. I would have liked to have implemented panning, but I believe this would have removed the majority of the masking present in the mix and would have made it difficult to demonstrate the effectiveness of the inter-channel auditory masking metric.

The process of applying the correct gain, equalisation and dynamic range settings in a multitrack is a challenging and time consuming task. I believe the framework I proposed here could be useful in developing systems for beginner and amateur music producers where it could be an assistive tool, giving initial settings for compressors and EQs on all tracks, that are then refined by the mix engineer.

# Chapter 8

# Conclusions, Limitations and Future Work

I first summarise what contributions were made to the fields of audio engineering and automatic mixing systems. I relate these contributions to my aims and objectives. Finally, I discuss the limitations I encountered and propose future directions where this work could potentially be taken.

## 8.1 Conclusion

In fulfilment of my aim to further understand the practice of subgrouping, how subgrouping affects an automatic mixing system and the importance of emotion in mixing, there have been four main contributions;

- Development of rules and guidelines on how subgrouping should be approached.
- A technique for automatically creating subgroups.
- A deeper understanding of the importance of emotion when mixing.
- Evidence to show that subgrouping is beneficial to automatic mixing systems.

Overall, I have shown that subgrouping is a poorly understood and generally undocumented part of the mix process. There seems to be no formal approach on how to create different types of subgroups, but through examination of mix data and the interview of practitioners in the field I have shown to a certain degree that there is a documented

process and thus have made recommendations based on this. This deeper understanding of subgrouping has allowed us to show that it can be beneficial to the mix process, whether it be a mix created by a human or an automatically generated mix.

The main aim of this work was to highlight the importance subgrouping plays when mixing audio as this mix technique is often taken for granted. It was also to help the audio engineering community to have a better understanding of the underlying processes and concepts associated with it. The main contribution of this work was to document all the knowledge and understanding around subgrouping and present it in an easy to follow piece of literature. The novelty of this work was take a mix technique that is normally performed by a human, to automate it and demonstrate how beneficial it can be to an automatic mixing system.

In chapter 3 I analysed the impact that subgrouping practices had on the perception of quality in a number of multitrack mixes. I also analysed the multitracks in order to see if any decision patterns emerged, which I later used to infer mix decisions in chapter 4. The experimental results in chapter 3 showed that subgroups are mainly made up of similar instrumentation, but in some cases can be a combination of different types of instrumentation. However, I found the former to occur more often than the latter. I found that the three instrument types that were subgrouped together most frequently were drums, vocals and guitars. I also found that when hierarchical subgrouping occurred, it was usually applied to drums and to a lesser extent vocals. I was able to show there was a strong significant Spearman correlation when looking at the median mix preference score of all the mixes done by each mix engineer and the amount of subgroups this mix engineer created on average. I also found a strong significant Spearman correlation when looking at the median mix preference score of all the mixes done by each mix engineer and the amount of EQ subgroup processing this mix engineer used on average. There was also a moderate significant Spearman correlation when looking at the median mix preference score of all the mixes done by each mix engineer and the amount of DRC subgroup processing this mix engineer used on average. These results provided an insight into some of the typical subgroup processes that occur when creating a mix. However, it is worth bearing in mind the subjects in this experiment were audio engineering students and may have been inherently biased by their instructor. This is the problem with analysing mix habits of mix engineers from the same group. It would be interesting to see how these results compare to those of another unrelated group of mix engineers. Generally, the results agreed with my intuition on how subgroups are created and processed.

In chapter 4, I interviewed ten award winning mix engineers through an online questionnaire, where I asked questions related to subgrouping of a qualitative and quantitative

nature. This was done to further understand the process of subgrouping and get a practitioners perspective. The questionnaire consisted of 21 questions, where I tested nine assumptions related to subgrouping. The nine assumptions were based on identifying subgrouping decisions, such as why a mix engineer creates subgroups, when they subgroup and how many subgroups they use. I then used thematic analysis to analyse the responses from each participant. This allowed us to develop five themes; (i) Decisions, (ii) Subgroup Effect Processing, (iii) Organisation, (iv) Exercising Control, and (v) Analogue versus Digital. Four of these five themes were somewhat expected, however Analogue versus Digital was something I had overlooked in my development of the survey. The analysis of the themes allowed us to show that eight out of the nine assumptions could be accepted to be true. Furthermore, by also taking the results of chapter 3 into consideration along with the thematic analysis results, I was able to propose a number of recommendations on how subgrouping should be implemented in an automatic mixing system and gave us a deeper understanding of the mix process. It was these recommendations that I utilised in my automatic mixing system in chapter 7.

In chapter 5, I determined a set of low level audio features that could be used to automatically subgroup multitrack audio. I determined these audio features using a Random Forest classifier for feature selection. I took 159 low level audio features and reduced this to 74 low level audio features using a feature selection process. I selected these audio features from a dataset of 54 individual multitrack recordings of varying musical genre, but mainly Pop, Rock and Indie. I was able to show that the most important audio features tended to be spectral features. I also performed agglomerative clustering on five unseen multitrack recordings using the original and the reduced audio feature set in order to compare their performance. I was able to show that the overall mis-classification measure went from 35.08% using the entire audio feature set to 7.89% using the reduced audio feature set. Thus indicating that my reduced set of audio features provides a significant increase in classification accuracy for the creation of automatic subgroups. This potentially could be a useful tool for a mix engineer, where if they had say 100 audio channels to deal with. This would allow them quickly apply control to relevant audio groups and avoid the time consuming task of assigning audio channels to groups. I was happy with the overall results of this experiment, but I do not believe my selected features would generalise well to other genres of music. However, this method could be reapplied with a larger dataset of more varying genres of music. I took a similar approach to this in Appendix 9.1, where I applied the feature selection process mainly to music of an electronic style. Furthermore, an alternative approach to my method could be to use convolutional neural networks (CNN) in order to see what interesting features could be learned from the data. CNN's have been used successfully in the last few years in the fields of vision and music information retrieval.

In chapter 6, I investigated the relationship between music production quality and musically induced and perceived emotions. A listening test was performed where 10 critical listeners and 10 non-critical listeners evaluated 10 songs. There were two mixes of each song, the low quality mix and the high quality mix. Each participants subjective experience was measured directly through questionnaire and indirectly by examining peripheral physiological changes, change in facial expressions and the number of head nods and shakes they made as they listened to each mix. I showed that music production quality had more of an emotional impact on critical listeners. Also, critical listeners had significantly different emotional responses to non-critical listeners for the high quality mixes and to a lesser extent the low quality mixes. The findings suggest that a higher quality mix only seems to matter in an emotional context to a subset of music listeners. This is important in the context of automatic mixing algorithms, in the sense that the perceived quality of an automatically generated mix may not be that important to those without critical listening skills. This suggests that automatically generated mixes may be good enough for the general public. However, I should remain somewhat sceptical of these results since I had a small sample size and many of the sensors used were noisy.

In chapter 7, I investigated different audio processing techniques to manipulate the frequency and dynamic characteristics of the signal in order to reduce masking based on a proposed MPEG metric. I also investigated whether or not automatically mixing using subgroups is beneficial or not to perceived quality and clarity of a mix. Evaluation results suggest that my proposed masking metric when utilised in an automatic mixing framework reduces inter-channel auditory masking and improves the perceived quality and perceived clarity of a mix. Furthermore, my results suggest that using subgrouping in an automatic mixing framework can also improve the perceived quality and perceived clarity of a mix. These results were important in the context of this thesis. However, there is still a lot of work to be done in terms of algorithms matching the skills of a human. It is also worth pointing out that this system was mixing the audio monaurally and the results may have been very different if it were mixed in stereo with panning either manually or automatically applied beforehand.

The wider impact of this work in the field of automatic mixing, is that it establishes a new approach to automatic mixing. Usually automatic mixes are created by mixing all tracks at once, where in the approach presented here, the mixing is done in smaller separate stages [2, 18]. This method is essentially a divide and conquer approach, where different mixing rules can be applied to different subgroups depending on genre and instrumentation. This could also potentially allow for the parallelisation of mixing tasks in order to speed up computation time. Outside of the field of automatic mixing, the work in this thesis was extended to automatically generating taxonomies for SFX libraries. This was primarily based on the work in Chapter 5 and was published in [159].

## 8.2 Limitations and Future Work

The field of automatic mixing is a relatively new field and as such many avenues are left to explore. A brief description of the limitations I encountered and possible improvements that could be made to further the understanding of subgrouping, the mix process as a whole, and automatic mixing systems are presented here.

One of the struggles I had with analysing the subgrouping structure, was that I had to manually open each Pro Tools session file and hand annotate all the data. The issues with this were that it took quite a long time and because it is done by hand, fatigue became an issue and therefore double checking was required. If this process were to be automated, much more data could be analysed at once and it would be much less error prone. If it were possible to develop a tool to assist with, and automate the collection of data, this could give far more data. If enough data were to be collected rapidly, this data could then subsequently be mined and used with machine learning classifiers.

When I looked at the automatic creation of subgroups, the dataset I used only represented music from the genres of Pop, Rock and Indie. I would like to extend the technique to other genres like Dance and Jazz, where different subgrouping structures occur. It might be possible to guide an extension of this technique using heuristics based on some of the responses given by the professional mix engineers wI interviewed.

When I examined the importance of emotion in mixing, I would also have liked to do pair-wise ranking between the two mix types, since research has shown that Likert scales may not be the best tool for affect studies, since the values they ask people to rate may mean different things to each participant [150]. I also think it would have been interesting to see if similar results occurred if the non-critical listeners had been provided with some training before the experiment i.e. trained to spot common mix defects. This would then mean that all the participants would have a more clearly defined idea on how a mix should sound. The non-critical listeners may not exhibit the same emotions as they did without training as they now know what to listen for. In future studies similar to this, I would encourage researchers to try and track foot and finger tapping as this is a common form of movement to music and is something I overlooked when designing my experiment [99].

When I assessed how much subgrouping could improve an automatic mixing system I chose to use a static EQ. However, based on the feedback from one of the more technically experienced listening test participants, they claimed they could hear that the EQ had fixed parameter settings. They said that both of the automatic mixes could be improved greatly if the parameters were to dynamically change over the course of each mix based on the measured amount of auditory masking. If I were to re-implement my proposed

automatic mixing system I used, I would attempt to optimise the EQ and dynamic range compressor parameters on a frame by frame basis. I would use the optimised solution for each frame to inform the initial search conditions of each subsequent frame in order to kick start the optimisation process.

If the system presented in Chapter 7 were to be computed on a platform where there were no limitations on processing power, the results suggest that the subgrouped audio would still be preferred. However, if the complexity of the audio processing being applied were to be increased there maybe a difference in quality and subgrouping might not be as advantageous. Future work could look at using equalisation with more bands and multi-band compression. This would increase the amount of control parameters to be optimised hence the need for more processing power.

One of the avenues of research that I was unable to explore due to time constraints was multi-subgroup mixing. This was where any particular instrument track does not necessarily need to belong to the same subgroup throughout the whole mix. It may belong to two or more. The idea being that particular instruments dynamics or timbre may change over time and might be better suited in another group. An example being where a bass guitar went from being plucked to suddenly being played using a slap technique.

I believe that a lot more research can be conducted in relation to subgrouping, since it is still a relatively unexplored area of audio engineering. As more mix data becomes available, more interesting and concrete recommendations can be inferred, which subsequently can be used to improve the mix process whether it be a human made mix or an automatically generated one.

# Chapter 9

# Appendices

## 9.1 Appendix A

### 9.1.1 Native Instruments Internship

#### 9.1.1.1 Introduction

In July 2014, the author spent six months working as a member of the Music Information Retrieval Research team at the Native Instruments head office in Berlin. Native Instruments was founded in 1996 and are a leading manufacturer of software and hardware for computer based audio production and DJing.

During the six month period spent at Native Instruments, research and development was conducted in creating a stem analysis tool. The software allowed a user to drop a folder of audio stems on to the GUI. The software would then analyse and classify each audio stem to determine if it is was either drum/percussion, bass, vocal, lead synthesizer or a pad/fx audio file. Once all the audio stems had been analysed and classified, the user was then able play each group individually or play them altogether. Each audio file that was classified was also assigned a colour. This colour indicated to the user how confident the classifier was in determining what group it belonged too. Green being the most confident and red being the least. A screen shot of the software can be seen in Figure 9.1.

The development of this tool was a direct continuation of research previously done on the automatic subgrouping of pop/rock music as seen in chapter 5, but in this case the subgrouping was applied to electronic music stems.

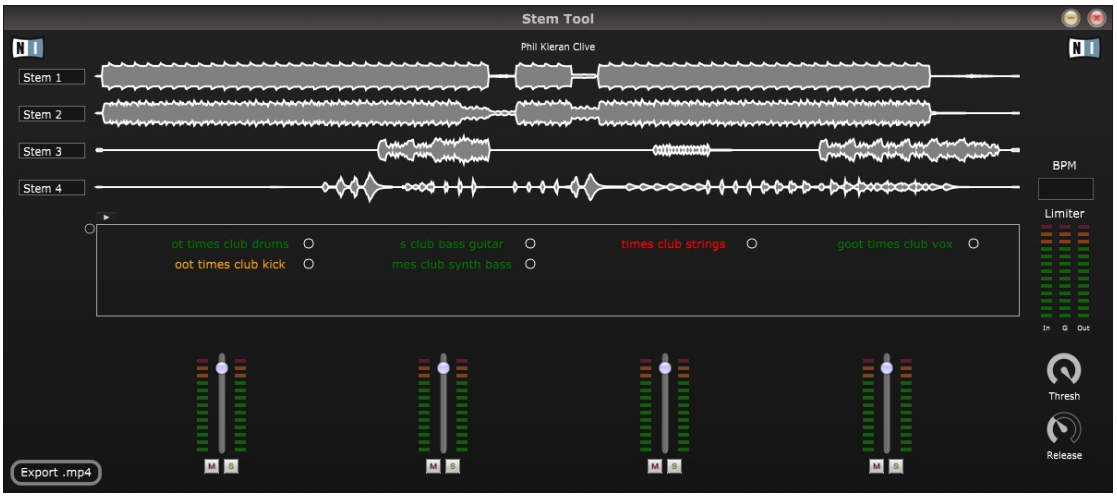


FIGURE 9.1: Native Instruments Stem Tool

This presented a new challenge because of the varying timbre of the instrumentation used in electronic music. An example of this would be that in pop/rock music, the drums normally tend to come from a recorded drum kit and will generally have a similar timbre throughout the genre, but in electronic music the percussive elements in a song could be something as simple as clicks and pops, but structured to give the music a pulse. An example of this type of sound can be found in the works of artists Ryoji Ikeda and Alva Noto [160, 161].

Addressing this problem, an audio feature that is normally used for tempo estimation and based on autocorrelation was adapted to determine if the stem had a periodic signal or not [125]. The next section discusses the approach used when classifying the stems.

#### 9.1.1.2 Waterfall Approach

Originally, it was decided to use a multi-class classifier for this problem. After realising the difficulty the varying timbre of electronic music presented, it was decided to use four binary classifiers instead and the classifier type that was used was Random Forest [105]. The binary classifiers were used in the way a number of waterfalls in succession would have different stages and pour into each other.

First, the percussive stems are separated from the harmonic stems. Then, the bass stems are separated from what is left over from the previous stage. This then happens to the vocals and ideally, what is supposed to be finally left over, is synthesised sounds that can be either lead synthesisers or pads/fx. This waterfall process is demonstrated in Figure 9.2.

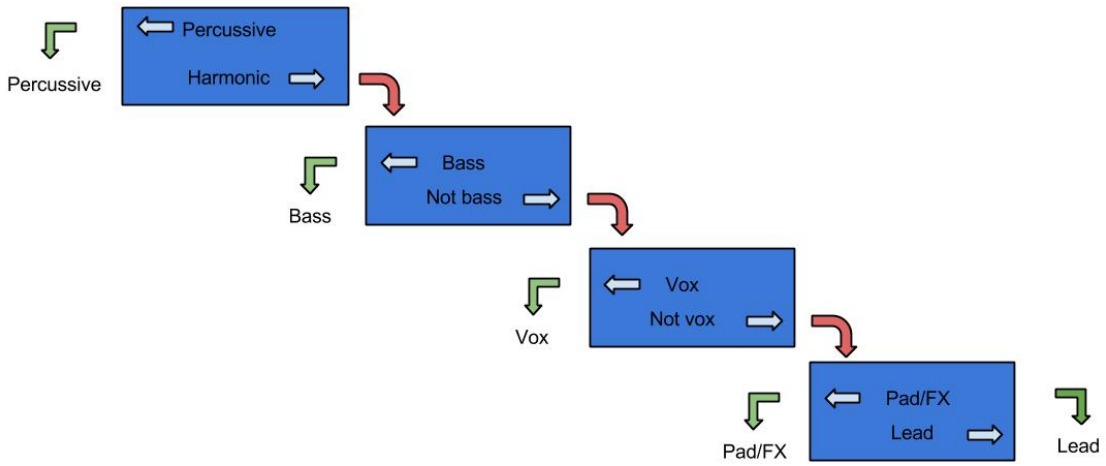


FIGURE 9.2: Waterfall Approach

As mentioned in the last section, the percussive-harmonic classifier was the most difficult to train, due to the varying timbre of electronic music.

#### 9.1.1.3 Dataset

The dataset consisted of 96 different songs by many different artists. All the songs used were of the Techno/House/IDM musical genre. These were provided by Native Instruments, where they had been licensed from the original artists to be used for the purpose of remixing. This consisted of 1496 audio stems sampled at 44100 Hz. For each stem that was used, the 30 secs of audio that had the most energy was used for feature extraction. This was also down-sampled to 22050 Hz in order speed up the feature extraction process. The breakdown of this data by label can be seen in Table 9.1

| Track Type | No. of Tracks | Mins. of Audio |
|---|---|---|
| Drums/Percussion | 514 | 257 |
| Lead/Synth | 364 | 182 |
| Vox | 275 | 137 |
| Pad/FX | 221 | 111 |
| Bass | 122 | 61 |

TABLE 9.1: Data Type Breakdown

The audio stems were annotated by the author using a very simple annotation tool. A .csv file that had a list of audio paths was opened and then a five second snippet of each audio file was listened to. The five second snippet that was heard was the five seconds of the stem with the most energy. The label for each stem was selected by the user and then this was appended to an output .csv file containing the file path and its chosen label.

#### 9.1.1.4 Feature Extraction Tool

Due to the Stem Tool and the extracted feature data needing to agree on the exact same values and calculations. It was decided to develop a batch feature extraction tool which would share a common code base with the Stem Tool.

The tool allowed the user to provide it with a .csv file that contains a column of file paths and another column corresponding to the audio file classification label. The user then specified how much audio to extract features from and a pooling time [124]. The feature extraction tool extracted audio with a window size of 1024 samples and a hop size of 512 samples. It extracted 159 audio features in total. The majority of these were frame based features, but some were whole track features. A list of the audio features and the relevant references are in Table 9.2 and Table 9.3

| Category | Feature | Reference |
|---|---|---|
| Dynamic | RMS | |
| | Peak Amplitude | |
| | Crest Factor | [161] |
| Spectral | Zero Crossing Rate | [128] |
| | Centroid | |
| | Spread | |
| | Skewness | |
| | Kurtosis | |
| | Brightness | . |
| | Flatness | . |
| | Roll-Off (.85 and .95) | |
| | Entropy | |
| | Flux | |
| | MFCC's 1-12 | |
| | Delta-MFCC's 1-12 | [128] |
| | Crest Factor | [123] |

Table 9.2: Pooled features

| Category | Feature | Reference |
|---|---|---|
| Dynamic | Periodicity | [125] |
| | Entropy of Energy | [126] |
| | Low Energy | [127] |

Table 9.3: Whole track features

#### 9.1.1.5 Classifier and Feature Selection

The Random Forest classifier was chosen for this project, due to its ability to perform feature selection and the ease at which it could be implemented into native code. It

also showed more favourable results when it was compared to k-NN and Support Vector Machine classifiers.

Determining the most salient features for each classifier was performed as follows. When training each Random Forest classifier, 100 trees were grown and feature importance was calculated. For any feature, the feature importance measure is the increase in prediction error if the values of that feature are permuted across the out-of-bag observations. This measure is computed for every tree, then averaged over the entire ensemble and divided by the standard deviation over the entire ensemble [108].

Once training was complete, a search method used to determine the better features. Any feature that performed under the average importance index of all the other features was eliminated.

A new Random Forest would then be trained with the new features. The overall performance of the classifier was evaluated by training at least 100 a trees and plotting the average F-Score as the number of trees increased. F-Score is a standard metric used to evaluate the performance of machine learning models. It is the harmonic average of the precision and recall scores of the model after it has been used to predict data from a test dataset. The number of trees used in the final classifier was determined from when the average F-Score was maximum when predicting a validation dataset. The final result of the classifier was determined by using a test dataset.

#### 9.1.1.6 Results

The results for each of the four classifiers are discussed in this section. Each classifiers performance will be presented as well as the features that were important for each.

Harmonic Percussive Classifier

This was the most challenging classifier to design due to the wildly varying timbre of electronic music and the difficulty in labelling some of the data. The idea was to use features that would most importantly capture the periodicity of drums and percussion. Using feature selection it was determined the four most important features were Periodicity, Entropy of Energy, Crest Factor and the Low Energy feature [125, 126].

The dataset was split up into 'DRUMS' and 'NOTDRUMS' for this classifier, so that meant the dataset was 34% Percussive data and 66% harmonic. Unfortunately, this imbalance in the dataset was unavoidable due to lack of data. The classifier required 14 trees to be grown to reach the highest average F-Score on the validation set. The classifier results on the test set are presented in Table 9.4.

| | Drums | Not Drums |
|---|---|---|
| Drums | **82.09%** | 17.91% |
| Not Drums | 6.11% | **93.89%** |
| Precision | 0.95 | 0.77 |
| Recall | 0.82 | 0.94 |
| F-Score | 0.88 | 0.85 |

TABLE 9.4: Test Data Results

Bass Classifier

This classifier had the worst imbalance out off all the classifiers. The data was labelled 'BASS' and 'NOTBASS'. The data was split 12% bass and then 88% not bass. This suffered from difficulty in labelling as sometimes it was hard to decide when a synthesizer could be considered a bass synthesizer or not just by listening. The classification rate is quite high for such an imbalance, but this is most likely due to bass having a lower spectral centroid than, say, vocals or pads. The most important features were Periodicity, Low Energy and Spectral Centroid. The classifier required 25 trees to be grown to reach the highest average F-Score on the validation set. The classifier results on the test set are presented in Table 9.5.

| | Bass | Not Bass |
|---|---|---|
| Bass | **81.58%** | 18.42% |
| Not Bass | 3.56% | **96.45%** |
| Precision | 0.78 | 0.97 |
| Recall | 0.82 | 0.97 |
| F-Score | 0.80 | 0.97 |

TABLE 9.5: Test Data Results

Vox Classifier

The vox classifier also experienced its own difficulties. This was because a lot of the vocals used in electronic music are heavily processed and barely recognisable. The human ear can perfectly discern that the audio is somewhat vocal, but it is difficult to train a classifier to do so. The most important features for this were Periodicity, MFCC's, Delta MFCC's and Spectral Flatness. The data was labelled 'NOTVOX' and 'VOX'. The data was split 28% vox and then 72% not vox. The classifier required 35 trees to be grown to reach the highest average F-Score on the validation set. The classifier results on the test set are presented in Table 9.6.

Pad/Fx Synth Classifier

| | Not Vox | Vox |
|---|---|---|
| Not Vox | **92.42%** | 7.58% |
| Vox | 4.57% | **95.43%** |
| Precision | 0.91 | 0.96 |
| Recall | 0.92 | 0.95 |
| F-Score | 0.91 | 0.96 |

TABLE 9.6: Test Data Results

The Pad/FX Synth classifier was the last classifier that was worked on during the internship, so it had the least amount of time dedicated to it. This classifier suffered the most from data labelling. It was very difficult at times to label some of audio stems as they would fall somewhere in between Pad/FX or Synth. The data was labelled 'PADFX' and 'SYNTH'. The data was split 37.7% Pad/FX and then 62.3% Synth. The classifier required 21 trees to be grown to reach the highest average F-Score on the validation set. The classifier results on the test set are presented in Table 9.7.

| | Synth | Pad/FX |
|---|---|---|
| Synth | **88.85%** | 11.18% |
| Pad/FX | 31.11% | **68.89%** |
| Precision | 0.82 | 0.79 |
| Recall | 0.89 | 0.69 |
| F-Score | 0.85 | 0.74 |

TABLE 9.7: Test Data Results

#### 9.1.1.7 Discussion

The Periodicity feature proved itself to be one of the most important features in the classification tasks as well as the Entropy of Energy feature. There is definitely scope to improve the vox classifier, as this was suffering poor classification on processed vocals. Analysis of attempts to recognising vocals in polyphonic music mixtures maybe a good research direction for this.

What could improve the Stem Tool is a transfer learning and active learning approach to classification, due to the fact a lot of the data being used on the tool would be completely unseen and the training data was difficult to label a lot of the time.

## 9.2 Appendix B

### 9.2.1 Ethics Approval and Pro Forma for "An empirical approach to the relationship between emotion and music production quality"

**Queen Mary, University of London**
Room W117
Queen's Building
Queen Mary University of London
Mile End Road
London E1 4NS

**Queen Mary Ethics of Research Committee**
Hazel Covill
Research Ethics Administrator
Tel: +44 (0) 20 7882 7915
Email: h.covill@qmul.ac.uk

c/o Dr Hatice Gunes
Eng 211
Department of Electronic Engineering
Queen Mary University of London
Mile End Road
London

12th October 2015

To Whom It May Concern:

**Re: QMREC1441 – The relationship between musically induced emotions and music production quality.**

I can confirm that Mr David Ronan has completed a Research Ethics Questionnaire with regard to the above research.

The result of which was the conclusion that his proposed work does not present any ethical concerns; is extremely low risk; and thus does not require the scrutiny of the full Research Ethics Committee.

Yours faithfully

Ms Hazel Covill – QMERC Administrator

**Pro forma information sheet and consent form**

Queen Mary
University of London

## Information sheet

### Research study "The relationship between musically induced emotions and production quality" information for participants

We would like to invite you to be part of this research project, if you would like to. You should only agree to take part if you want to, it is entirely up to you. If you choose not to take part there won't be any disadvantages for you and you will hear no more about it.

Please read the following information carefully before you decide to take part; this will tell you why the research is being done and what you will be asked to do if you take part. Please ask if there is anything that is not clear or if you would like more information.

If you decide to take part you will be asked to sign the attached form to say that you agree.

You are still free to withdraw at any time and without giving a reason.

Details

The purpose of this study is to determine the extent of the link between musically induced emotions and music production quality. In order to investigate the link between emotion and music production quality, subjective feeling will be measured through self-report, where each participant will indicate their emotions using the Geneva Emotional Music Scale (GEMS-9) throughout the listening experience. A multivariate approach will then be used for psychophysiological response, where it is planned to measure skin conductance (GSR) and heart rate (ECG). We will also record each participant's facial expressions and analyse these for affect, so it is important that each participant looks into the camera.

It is up to you to decide whether or not to take part. If you do decide to take part you will be given this information sheet to keep and be asked to sign a consent form.

If you have any questions or concerns about the manner in which the study was conducted please, in the first instance, contact the researcher responsible for the study.

If this is unsuccessful, or not appropriate, please contact the Secretary at the Queen Mary Ethics of Research Committee, Room W117, Queen's Building, Mile End Campus, Mile End Road, London or research-ethics@qmul.ac.uk.

Queen Mary
University of London

## Consent form

Please complete this form after you have read the Information Sheet and/or listened to an explanation about the research.

Title of Study: _____
Queen Mary Ethics of Research Committee Ref: _____

- Thank you for considering taking part in this research. The person organizing the research must explain the project to you before you agree to take part.
- If you have any questions arising from the Information Sheet or explanation already given to you, please ask the researcher before you decide whether to join in. You will be given a copy of this Consent Form to keep and refer to at any time.
- *I understand that if I decide at any other time during the research that I no longer wish to participate in this project, I can notify the researchers involved and be withdrawn from it immediately.*
- *I consent to the processing of my personal information for the purposes of this research study. I understand that such information will be treated as strictly confidential and handled in accordance with the provisions of the Data Protection Act 1998.*

**Participant's Statement:**

I _____ agree that the research project named above has been explained to me to my satisfaction and I agree to take part in the study. I have read both the notes written above and the Information Sheet about the project, and understand what the research study involves.

Signed: Date:

**Investigator's Statement:**

I _____ confirm that I have carefully explained the nature, demands and any foreseeable risks (where applicable) of the proposed research to the volunteer

### 9.2.2 Ethics Approval and Pro Forma for "Automatic Minimisation of Masking in Multitrack Audio using Subgroups"

Queen Mary
University of London

**Queen Mary, University of London**
Room W117
Queen's Building
Queen Mary University of London
Mile End Road
London E1 4NS

**Queen Mary Ethics of Research Committee**
Hazel Covill
Research Ethics Administrator
Tel: +44 (0) 20 7882 7915
Email: h.covill@qmul.ac.uk

c/o Dr Josh Reiss
Eng E305
EECS
Mile End
London

17th July 2017

To Whom It May Concern:

**Re: QMREC2034a - Automatic Minimisation of Masking in a Multitrack using Subgroups.**

I can confirm that David Ronan has completed a Research Ethics Questionnaire with regard to the above research.

The result of which was the conclusion that his proposed work does not present any ethical concerns; is extremely low risk; and thus does not require the scrutiny of the full Research Ethics Committee.

Yours faithfully

Mr Jack Biddle – Research Approvals Advisor

**Pro forma information sheet and consent form**

Queen Mary
University of London

## Information sheet

### Research study "Automatic Minimisation of Masking in Multitrack Audio": information for participants

We would like to invite you to be part of this research project, if you would like to. You should only agree to take part if you want to, it is entirely up to you. If you choose not to take part there won't be any disadvantages for you and you will hear no more about it. [*If appropriate:* Choosing not to take part will not affect your access to treatment or services in any way].

Please read the following information carefully before you decide to take part; this will tell you why the research is being done and what you will be asked to do if you take part. Please ask if there is anything that is not clear or if you would like more information.

If you decide to take part you will be asked to sign the attached form to say that you agree.

You are still free to withdraw at any time and without giving a reason.

"The aim of this study is to conduct a listening test, where you will listen to a number of different mixes of the same song. Each mix will have been either created by a human or by using an automatic mixing algorithm. We will require you to rate each mix in terms of your preference or your ability to distinguish the individual sources (i.e. the lack of masking). In the second part of the experiment, we require you to compare two different mixes of each song and rate each mix for perceived emotion along three different emotional dimensions. You will hear five different songs in this experiment, where we will be using five different mixes of each song. You are required to have critical listening skills in order to take part in this experiment."

It is up to you to decide whether or not to take part. If you do decide to take part you will be given this information sheet to keep and be asked to sign a consent form.

If you have any questions or concerns about the manner in which the study was conducted please, in the first instance, contact the researcher responsible for the study. If this is unsuccessful, or not appropriate, please contact the Secretary at the Queen Mary Ethics of Research Committee, Room W104, Queen's Building, Mile End Campus, Mile End Road, London or research-ethics@qmul.ac.uk.

Queen Mary
University of London

## Consent form

Please complete this form after you have read the Information Sheet and/or listened to an explanation about the research.

Title of Study: **Automatic Minimisation of Masking in Multitrack Audio**
Queen Mary Ethics of Research Committee Ref:

. • Thank you for considering taking part in this research. The person organizing the research must explain the project to you before you agree to take part.

. • If you have any questions arising from the Information Sheet or explanation already given to you, please ask the researcher before you decide whether to join in. You will be given a copy of this Consent Form to keep and refer to at any time.

. • *I understand that if I decide at any other time during the research that I no longer wish to participate in this project, I can notify the researchers involved and be withdrawn from it immediately.*

. • *I consent to the processing of my personal information for the purposes of this research study. I understand that such information will be treated as strictly confidential and handled in accordance with the provisions of the Data Protection Act 1998.*

**Participant's Statement:**

I _____ agree that the research project named above has been explained to me to my satisfaction and I agree to take part in the study. I have read both the notes written above and the Information Sheet about the project, and understand what the research study involves.

Signed: Date:

**Investigator's Statement:**

I _____ confirm that I have carefully explained the nature, demands and any foreseeable risks (where applicable) of the proposed research to the volunteer

## 9.3 Appendix C

# Subgrouping survey

Save my progress and resume later | Resume a previously saved form


1 – How would you define subgrouping?

2 If subgrouping can be defined as when you sum one or more audio tracks into a bus with the idea of creating a submix. Do you subgroup your audio tracks during the mixing process?

- Yes
- No

2 (i) – If you chose no, can you please state why you don't use this approach?

2 (ii) – If you chose yes, please select which of these statements apply to you

| | | |
|---|---|---|
| I subgroup based on instrument family | Yes | No |
| I subgroup with FX processing in mind | Yes | No |
| I subgroup so that I can create individual submixes | Yes | No |
| I subgroup depending on the genre being mixed | Yes | No |
| I subgroup because it makes the mixing process less complicated | Yes | No |
| I subgroup for other reasons not stated here | Yes | No |

Please state the other reason(s)

3 – If you subgroup with FX processing mind, please check the box next to the FX this applies to

- [ ] Reverb
- [ ] Dynamic Range Compression
- [ ] EQ
- [ ] Loudness
- [ ] Delay
- [ ] Other

Please state:

4 – How does the genre of music being mixed affect your subgrouping?

5 – How many tracks are usually required before you need to start considering subgrouping?

If possible, please state why:

6 – What is the maximum, minimum and average amount of subgroups you would have in a mix project depending on the number of audio tracks present?

| | Minimum | Average | Maximum |
|---|---|---|---|
| 10 Tracks | | | |
| 25 Tracks | | | |
| 100 Tracks | | | |

7 – Over the course of the last five years, how has your subgrouping approach changed over time?

8 – Given the following subgroups, please rank these instruments in order of how likely you are to apply dynamic range compression? (if you feel there is no order, please assign the same number)

| | |
|---|---|
| Drums | Rank (1–10) |
| Guitars | Rank (1–10) |
| Pads | Rank (1–10) |
| Bass | Rank (1–10) |
| Main Vocals | Rank (1–10) |
| Background Vocals | Rank (1–10) |
| Strings | Rank (1–10) |
| Brass | Rank (1–10) |
| Percussion | Rank (1–10) |
| Percussive Keyboard Instruments | Rank (1–10) |

If possible, can you please explain why you have chosen these rankings?

9 – Rank, from one to six, your order of execution of each of the following aspects of mixing (if you feel there is no order, please assign the same number)

| | |
|---|---|
| Loudness/Level | Rank (1–6) |
| Subgrouping | Rank (1–6) |

Panning ☐ Rank (1-6)

Equalizing ☐ Rank (1-6)

Effects (temporal) ☐ Rank (1-6)

Dynamic Range Compression ☐ Rank (1-6)

Can you please explain why you take this order of execution?

10 - Over the course of your last 100 individual mixing projects, approximatley...

How often did you subgroup to maintain good gain structure?

☐ (percentage: 0 - 100)

How often did you subgroup some or all the audio tracks?

☐ (percentage: 0 - 100)

How often have you used subgrouping in order to pan a group of instruments?

☐ (percentage: 0 - 100)

How often did you split your drums into different subgroups?

☐ (percentage: 0 - 100)

How often did you subgroup instruments to eliminate auditory masking?

☐ (percentage: 0 - 100)

11 - How often in your last 100 individual mixes have you decided part way through mixing that some or all of the subgroupings are incorrect and made changes to them?

☐ (percentage: 0 - 100)

What were your reasons for changing them?

12 – How do you evaluate if the subgroupings you have made are correct?

13 – Would you subgroup a bass guitar being played percussively i.e. slap technique, with a drum or percussion subgroup?

Yes No

If possible, please state why:

14 – Do you ever create a subgroup containing two or more subgroups?

Yes No

If possible, please state why:

15 – Do you ever put the kick drum in its own separate subgroup?

Yes No

If possible, please state why:

16 – Do you ever put the snare drum in its own separate subgroup?

Yes No

If possible, please state why:

17 – Do you ever put rhythm guitars and lead guitars in the same subgroup?

Yes No

If possible, please state why:

18 – Do you ever put bass guitar and lead guitars in the same subgroup?

Yes No

If possible, please state why:

19 – Do you ever put acoustic guitars and electric guitars in the same subgroup?

Yes No

If possible, please state why:

20 – Do you ever subgroup instruments in order to achieve a uniform tone through EQ?

Yes No

If possible, please state why:

21 – Do you ever subgroup instruments in order to eliminate auditory masking? i.e. vocal masking

- ( ) Yes
- ( ) No

If possible, please state why:

Name:

Email:

Age:

Average number of mixing projects you are involved in per year (over the course of the last 5 years):

What genres do you usually mix?

- [ ] Rock
- [ ] Pop
- [ ] Metal
- [ ] Country
- [ ] R'n'B
- [ ] Electronic
- [ ] Jazz
- [ ] Other(s)

Please specify other genres:

Any further comments you would like to add:

Submit

Save my progress and resume later | Resume a previously saved form

Need assistance with this form?

# Bibliography


[1] Bobby Owsinski. *The mixing engineer's handbook*. Nelson Education, 2013.

[2] Joshua D Reiss. Intelligent systems for mixing multichannel audio. In *17th International Conference on Digital Signal Processing (DSP)*, pages 1–6. IEEE, 2011.

[3] Jeffrey Scott, Matthew Prockup, Erik M Schmidt, and Youngmoo E Kim. Automatic multi-track mixing using linear dynamical systems. In *Proceedings of the 8th Sound and Music Computing Conference, Padova, Italy*, page 12, 2011.

[4] Martin Morrell and Joshua Reiss. Dynamic panner: An adaptive digital audio effect for spatial audio. In *Audio Engineering Society Convention 127*. Audio Engineering Society, 2009.

[5] Jacob A Maddams, Saoirse Finn, and Joshua D Reiss. An autonomous method for multi-track dynamic range compression. In *Proceedings of the 15th International Conference on Digital Audio Effects (DAFx-12)*, pages 1–8, 2012.

[6] Dimitrios Giannoulis, Michael Massberg, and Joshua D Reiss. Parameter automation in a dynamic range compressor. *Journal of the Audio Engineering Society*, 61 (10):716–726, 2013.

[7] Roey Izhaki. *Mixing audio: concepts, practices and tools*. Taylor & Francis, 2013.

[8] Zheng Ma, Brecht De Man, Pedro DL Pestana, Dawn AA Black, and Joshua D Reiss. Intelligent multitrack dynamic range compression. *Journal of the Audio Engineering Society*, 63(6):412–426, 2015.

[9] Michael Terrell, Andrew Simpson, and Mark Sandler. The mathematics of mixing. *Journal of the Audio Engineering Society*, 62(1/2):4–13, 2014.

[10] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull. Music emotion recognition: A state of the art review. In *11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, pages 255–266. Citeseer, 2010.

[11] Joshua D Reiss and Andrew McPherson. *Audio effects: theory, implementation and application*. CRC Press, 2014.

[12] Pedro Pestana and Joshua Reiss. Intelligent audio production strategies informed by best practices. In *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*. Audio Engineering Society, 2014.

[13] Brecht De Man, Brett Leonard, Richard King, and Joshua D. Reiss. An analysis and evaluation of audio features for multitrack music mixtures. In *15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, October 2014.

[14] Brecht De Man, Matt Boerum, Brett Leonard, George Massenburg, Richard King, and Joshua D. Reiss. Perceptual evaluation of music mixing practices. In *138th Convention of the Audio Engineering Society*, May 2015.

[15] Jeffrey J Scott and Youngmoo E Kim. Instrument identification informed multitrack mixing. In *14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, pages 305–310, 2013.

[16] Brecht De Man and Joshua D. Reiss. A knowledge-engineered autonomous mixing system. In *135th Convention of the Audio Engineering Society*, October 2013.

[17] Alexander U Case. *Mix smart*. Focal Press, 2011.

[18] Sina Hafezi and Joshua D Reiss. Autonomous multitrack equalization based on masking reduction. *Journal of the Audio Engineering Society*, 63(5):312–323, 2015.

[19] Brian CJ Moore. *An introduction to the psychology of hearing*. Brill, 2012.

[20] Brian R Glasberg and Brian CJ Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing research*, 47(1):103–138, 1990.

[21] Andrew J Oxenham and Brian CJ Moore. Modeling the additivity of nonsimultaneous masking. *Hearing research*, 80(1):105–118, 1994.

[22] Eberhard Zwicker and Hugo Fastl. *Psychoacoustics: Facts and models*, volume 22. Springer Science & Business Media, 2013.

[23] Brian CJ Moore and Brian R Glasberg. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *The Journal of the Acoustical Society of America*, 74(3):750–753, 1983.

[24] Zheng Ma, Joshua D Reiss, and Dawn AA Black. Partial loudness in multitrack mixing. In *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*. Audio Engineering Society, 2014.

[25] John E Dennis Jr and Robert B Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*. SIAM, 1996.

[26] Philip E Gill and Walter Murray. *Numerical methods for constrained optimization*. Academic Pr, 1974.

[27] Pedro Duarte Leal Gomes Pestana. *Automatic mixing systems using adaptive digital audio effects*. PhD thesis, Universidade Católica Portuguesa, 2013.

[28] Manfred R Schroeder, Bishnu S Atal, and JL Hall. Optimizing digital speech coders by exploiting masking properties of the human ear. *The Journal of the Acoustical Society of America*, 66(6):1647–1652, 1979.

[29] James D Johnston. Transform coding of audio signals using perceptual noise criteria. *IEEE Journal on selected areas in communications*, 6(2):314–323, 1988.

[30] Allen Gersho. Advances in speech and audio compression. *Proceedings of the IEEE*, 82(6):900–918, 1994.

[31] Marina Bosi, Karlheinz Brandenburg, Schuyler Quackenbush, Louis Fielder, Kenzo Akagiri, Hendrik Fuchs, and Martin Dietz. Iso/iec mpeg-2 advanced audio coding. *Journal of the Audio engineering society*, 45(10):789–814, 1997.

[32] Ted Painter and Andreas Spanias. Perceptual coding of digital audio. *Proceedings of the IEEE*, 88(4):451–515, 2000.

[33] Michael M Goodwin, Aaron J Hipple, and Brian Link. Predicting and preventing unmasking incurred in coded audio post-processing. *IEEE Transactions on Speech and Audio Processing*, 13(1):32–41, 2005.

[34] Arnaud Robert and Justin Picard. On the use of masking models for image and audio watermarking. *IEEE Transactions on Multimedia*, 7(4):727–739, 2005.

[35] Charfeddine Maha, Elarbi Maher, and Ben Amar Chokri. A blind audio watermarking scheme based on neural network and psychoacoustic model with error correcting code in wavelet domain. In *3rd International Symposium on Communications, Control and Signal Processing, 2008, ISCCSP 2008*, pages 1138–1143. IEEE, 2008.

[36] Jan H Plasberg and W Bastiaan Kleijn. The sensitivity matrix: Using advanced auditory models in speech and audio processing. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):310–319, 2007.

[37] Matti Karjalainen. A new auditory model for the evaluation of sound quality of audio systems. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'85*, volume 10, pages 608–611. IEEE, 1985.

[38] Thilo Thiede, William C Treurniet, Roland Bitto, Christian Schmidmer, Thomas Sporer, John G Beerends, and Catherine Colomes. Peaq-the itu standard for objective measurement of perceived audio quality. *Journal of the Audio Engineering Society*, 48(1/2):3–29, 2000.

[39] Antony W Rix, John G Beerends, D-S Kim, Peter Kroon, and Oded Ghitza. Objective assessment of speech and audio qualitytechnology and applications. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):1890–1901, 2006.

[40] Peter Balazs, Bernhard Laback, Gerhard Eckel, and Werner A Deutsch. Time–frequency sparsity by removing perceptually irrelevant components using a simple model of simultaneous masking. *IEEE Transactions on Audio, Speech, and Language processing*, 18(1):34–49, 2010.

[41] Torsten Dau, Dirk Püschel, and Armin Kohlrausch. A quantitative model of the effectivesignal processing in the auditory system. i. model structure. *The Journal of the Acoustical Society of America*, 99(6):3615–3622, 1996.

[42] Torsten Dau, Birger Kollmeier, and Armin Kohlrausch. 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, 1997.

[43] Morten L Jepsen, Stephan D Ewert, and Torsten Dau. A computational model of human auditory signal processing and perception. *The Journal of the Acoustical Society of America*, 124(1):422–438, 2008.

[44] Brian R Glasberg and Brian CJ Moore. 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, 2005.

[45] Brian CJ Moore, Brian R Glasberg, and Thomas Baer. A model for the prediction of thresholds, loudness, and partial loudness. *Journal of the Audio Engineering Society*, 45(4):224–240, 1997.

[46] Brian R Glasberg and Brian CJ Moore. A model of loudness applicable to time-varying sounds. *Journal of the Audio Engineering Society*, 50(5):331–342, 2002.

[47] Sebastian Vega and Jordi Janer. Quantifying masking in multi-track recordings. In *Proceedings of SMC Conference*, 2010.

[48] Philipp Aichinger, Alois Sontacchi, and Berit Schneider-Stickler. Describing the transparency of mixdowns: The masked-to-unmasked-ratio. In *Audio Engineering Society Convention 130*. Audio Engineering Society, 2011.

[49] Enrique Perez-Gonzalez and Joshua Reiss. Automatic gain and fader control for live mixing. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA'09.*, pages 1–4. IEEE, 2009.

[50] Dominic Ward, Joshua D Reiss, and Cham Athwal. Multitrack mixing using a model of loudness and partial loudness. In *Audio Engineering Society Convention 133*. Audio Engineering Society, 2012.

[51] Udo Zölzer. *DAFX: digital audio effects*. John Wiley & Sons, 2011.

[52] Brian CJ Moore. Masking in the human auditory system. In *Audio Engineering Society Conference: Collected Papers on Digital Audio Bit-Rate Reduction*. Audio Engineering Society, 1996.

[53] Brian CJ Moore. An introduction to the psychology of hearing. *San Diego*, 2003.

[54] Masashi Unoki, Toshio Irino, Brian Glasberg, Brian CJ Moore, and Roy D Patterson. 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, 2006.

[55] Gordon Wichern, Hannah Robertson, and Aaron Wishnick. Quantitative analysis of masking in multitrack mixes using loudness loss. In *Audio Engineering Society Convention 141*. Audio Engineering Society, 2016.

[56] Vincent Verfaille, Udo Zolzer, and Daniel Arfib. Adaptive digital audio effects (a-dafx): A new class of sound transformations. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1817–1831, 2006.

[57] Stuart Mansbridge, Saoirse Finn, and Joshua D Reiss. Implementation and evaluation of autonomous multi-track fader control. In *Audio Engineering Society Convention 132*. Audio Engineering Society, 2012.

[58] Enrique Perez-Gonzalez and Joshua Reiss. Automatic equalization of multichannel audio using cross-adaptive methods. In *Audio Engineering Society Convention 127*. Audio Engineering Society, 2009.

[59] Zheng Ma, Joshua D Reiss, and Dawn AA Black. Implementation of an intelligent equalization tool using yule-walker for music mixing and mastering. In *Audio Engineering Society Convention 134*. Audio Engineering Society, 2013.

[60] Benjamin Friedlander and Boaz Porat. The modified yule-walker method of arma spectral estimation. *IEEE Transactions on Aerospace and Electronic Systems*, (2): 158–173, 1984.

[61] Dimitrios Giannoulis, Michael Massberg, and Joshua D Reiss. Digital dynamic range compressor designa tutorial and analysis. *Journal of the Audio Engineering Society*, 60(6):399–408, 2012.

[62] Stuart Mansbridge, Saorise Finn, and Joshua D Reiss. An autonomous system for multitrack stereo pan positioning. In *Audio Engineering Society Convention 133*. Audio Engineering Society, 2012.

[63] Pedro D Pestana and Joshua D Reiss. A cross-adaptive dynamic spectral panning technique. In *Proceedings of the 17th International Conference on Digital Audio Effects (DAFx-14)*, pages 303–307, 2014.

[64] Alf Gabrielsson. Emotion perceived and emotion felt: Same or different? *Musicae Scientiae*, 5(1 suppl):123–147, 2002.

[65] Tuomas Eerola and Jonna K Vuoskoski. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 2010.

[66] Yading Song, Simon Dixon, Marcus T Pearce, and Andrea R Halpern. Perceived and induced emotion responses to popular music. *Music Perception: An Interdisciplinary Journal*, 33(4):472–492, 2016.

[67] Alf Gabrielsson and Erik Lindström. The role of structure in the musical expression of emotions. *Handbook of music and emotion: Theory, research, applications*, pages 367–400, 2010.

[68] Patrik N Juslin and Daniel Västfjäll. Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and brain sciences*, 31(05):559–575, 2008.

[69] Patrik N Juslin, Simon Liljeström, Daniel Västfjäll, and Lars-Olov Lundqvist. How does music evoke emotions? exploring the underlying mechanisms. In *Handbook of music and emotion*, pages 605–642. Oxford Press, 2010.

[70] Patrik N Juslin. From everyday emotions to aesthetic emotions: towards a unified theory of musical emotions. *Physics of life reviews*, 10(3):235–266, 2013.

[71] Patrik N Juslin, László Harmat, and Tuomas Eerola. What makes music emotionally significant? exploring the underlying mechanisms. *Psychology of Music*, 2013.

[72] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4): 169–200, 1992.

[73] Jaak Panksepp. *Affective neuroscience: The foundations of human and animal emotions*. Oxford university press, 1998.

[74] Tuomas Eerola, Olivier Lartillot, and Petri Toiviainen. Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. In *10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, pages 621–626, 2009.

[75] John A Sloboda and Patrik N Juslin. Psychological perspectives on music and emotion. 2001.

[76] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[77] Albert Mehrabian and James A Russell. *An approach to environmental psychology.* the MIT Press, 1974.

[78] Marcel Zentner, Didier Grandjean, and Klaus R Scherer. Emotions evoked by the sound of music: characterization, classification, and measurement. *Emotion*, 8(4): 494, 2008.

[79] Marcus T Pearce and Andrea R Halpern. Age-related patterns in emotions evoked by music. *Psychology of Aesthetics, Creativity, and the Arts*, 9(3):248, 2015.

[80] Gunter Kreutz, Ulrich Ott, Daniel Teichmann, Patrick Osawa, and Dieter Vaitl. Using music to induce emotions: Influences of musical preference and absorption. *Psychology of music*, 2007.

[81] Sandrine Vieillard, Isabelle Peretz, Nathalie Gosselin, Stéphanie Khalfa, Lise Gagnon, and Bernard Bouchard. Happy, sad, scary and peaceful musical excerpts for research on emotions. *Cognition & Emotion*, 22(4):720–752, 2008.

[82] Margaret M Bradley and Peter J Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59, 1994.

[83] Carroll E Izard. Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspectives on psychological science*, 2(3):260–280, 2007.

[84] David Watson, Lee A Clark, and Auke Tellegen. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54(6):1063, 1988.

[85] Oliver Grewe, Frederik Nagel, Reinhard Kopiez, and Eckart Altenmüller. Emotions over time: Synchronicity and development of subjective, physiological, and facial affective reactions to music. *Emotion*, 7(4):774, 2007.

[86] Patrick G Hunter and E Glenn Schellenberg. Music and emotion. In *Music perception*, pages 129–164. Springer, 2010.

[87] Emery Schubert. Continuous self-report methods. *Handbook of music and emotion: Theory, research, applications*, 2:223–253, 2010.

[88] Hauke Egermann, Marcus T Pearce, Geraint A Wiggins, and Stephen McAdams. Probabilistic models of expectation violation predict psychophysiological emotional responses to live concert music. *Cognitive, Affective, & Behavioral Neuroscience*, 13(3):533–553, 2013.

[89] Evan Morgan, Hatice Gunes, and Nick Bryan-Kinns. Using affective and behavioural sensors to explore aspects of collaborative music making. *International Journal of Human-Computer Studies*, 82:31–47, 2015.

[90] D Hodges. Psychophysiological measures. *Handbook of music and emotion*, pages 279–312, 2010.

[91] Gary E Schwartz, Serena-Lynn Brown, and Geoffrey L Ahern. Facial muscle patterning and subjective experience during affective imagery: Sex differences. *Psychophysiology*, 17(1):75–82, 1980.

[92] Charlotte VO Witvliet and Scott R Vrana. Play it again sam: Repeated exposure to emotionally evocative music polarises liking and smiling responses, and influences other affective reports, facial emg, and heart rate. *Cognition and Emotion*, 21(1):3–25, 2007.

[93] John L Andreassi. *Psychophysiology: Human behavior & physiological response*. Psychology Press, 2013.

[94] Lars-Olov Lundqvist, Fredrik Carlsson, Per Hilmersson, and Patrik Juslin. Emotional responses to music: experience, expression, and physiology. *Psychology of Music*, 2008.

[95] Mathieu Roy, Jean-Philippe Mailhot, Nathalie Gosselin, Sébastien Paquette, and Isabelle Peretz. Modulation of the startle reflex by pleasant and unpleasant music. *International Journal of Psychophysiology*, 71(1):37–42, 2009.

[96] Paul Ekman and Wallace V Friesen. *Facial action coding system: Investigator's guide*. Consulting Psychologists Press, 1978.

[97] Anne Weisgerber, Nicolas Vermeulen, Isabelle Peretz, Séverine Samson, Pierre Philippot, Pierre Maurage, D'Aoust Catherine De Graeuwe, Aline De Jaegere, Benoît Delatte, Benoît Gillain, et al. Facial, vocal and musical emotion recognition is altered in paranoid schizophrenic patients. *Psychiatry research*, 2015.

[98] Brian A Silvey. The role of conductor facial expression in students evaluation of ensemble expressivity. *Journal of Research in Music Education*, 2012.

[99] Nils Lennart Wallin and Björn Merker. *The origins of music*. MIT press, 2001.

[100] Peter Sedlmeier, Oliver Weigelt, and Eva Walther. Music is in the muscle: How embodied cognition may influence music preferences. *Music Perception: An Interdisciplinary Journal*, 28(3):297–306, 2011.

[101] Gail Tom, Paul Pettersen, Teresa Lau, Trevor Burton, and Jim Cook. The role of overt head movement in the formation of affect. *Basic and Applied Social Psychology*, 12(3):281–289, 1991.

[102] Gary L Wells and Richard E Petty. The effects of over head movements on persuasion: Compatibility and incompatibility of responses. *Basic and Applied Social Psychology*, 1(3):219–230, 1980.

[103] Kevin Murphy. Machine learning: a probabilistic approach. *Massachusetts Institute of Technology*, pages 1–21, 2012.

[104] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[105] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[106] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1):25, 2007.

[107] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC bioinformatics*, 9(1):307, 2008.

[108] Robin Genuer, Jean-Michel Poggi, and Christine Tuleau-Malot. Variable selection using random forests. *Pattern Recognition Letters*, 31(14):2225–2236, 2010.

[109] Pavel Berkhin. A survey of clustering data mining techniques. In *Grouping multidimensional data*, pages 25–71. Springer, 2006.

[110] Stephen P Borgatti. *How to explain hierarchical clustering*. Citeseer, 1994.

[111] Brecht De Man, Mariano Mora-Mcginity, György Fazekas, and Joshua D. Reiss. The Open Multitrack Testbed. In *137th Convention of the Audio Engineering Society*, October 2014.

[112] Brecht De Man and Joshua D. Reiss. Analysis of peer reviews in music production. *Journal on the Art of Record Production*, 10, July 2015.

[113] David Ronan, Brecht De Man, Hatice Gunes, and Joshua D Reiss. The impact of subgrouping practices on the perception of multitrack mixes. In *Audio Engineering Society Convention 139*. Audio Engineering Society, 2015.

[114] Brecht De Man, Matthew Boerum, Brett Leonard, Richard King, George Massenburg, and Joshua D Reiss. Perceptual evaluation of music mixing practices. In *Audio Engineering Society Convention 138*. Audio Engineering Society, 2015.

[115] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[116] Richard E Boyatzis. *Transforming qualitative information: Thematic analysis and code development*. Sage, 1998.

[117] URL `https://help-nv.qsrinternational.com/12/win/v12.1.55-d3ea61/Content/vizualizations/cluster-analysis.htm`.

[118] Enrique Perez-Gonzalez and Joshua D Reiss. Automatic mixing. *DAFX: Digital Audio Effects, Second Edition*, pages 523–549, 2011.

[119] Philippe Hamel, Sean Wood, and Douglas Eck. Automatic identification of instrument classes in polyphonic and poly-instrument audio. In *10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, pages 399–404, 2009.

[120] Judith C Brown, Olivier Houix, and Stephen McAdams. Feature dependence in the automatic identification of musical woodwind instruments. *The Journal of the Acoustical Society of America*, 109(3):1064–1072, 2001.

[121] Antti Eronen and Anssi Klapuri. Musical instrument recognition using cepstral coefficients and temporal features. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000. ICASSP'00.*, volume 2, pages II753–II756. IEEE, 2000.

[122] Keith D Martin and Youngmoo E Kim. Musical instrument identification: A pattern-recognition approach. *The Journal of the Acoustical Society of America*, 104(3):1768–1768, 1998.

[123] Geoffroy Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. *Institut de Recherche et Coordination Acoustique/Musique (IRCAM), Tech. Rep.*, 2004.

[124] Philippe Hamel, Simon Lemieux, Yoshua Bengio, and Douglas Eck. Temporal pooling and multiscale learning for automatic annotation and ranking of music audio. In *12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, pages 729–734, 2011.

[125] Christian Uhle. Tempo induction by investigating the metrical structure of music using a periodicity signal that relates to the tatum period. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, 2005.

[126] Theodoros Giannakopoulos and Aggelos Pikrakis. *Introduction to Audio Analysis: A MATLAB Approach.* Academic Press, 2014.

[127] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

[128] Olivier Lartillot and Petri Toiviainen. A matlab toolbox for musical feature extraction from audio. In *International Conference on Digital Audio Effects*, pages 237–244, 2007.

[129] Zhouyu Fu, Guojun Lu, Kai Ming Ting, and Dengsheng Zhang. A survey of audio-based music classification and annotation. *IEEE Transactions on Multimedia*, 13 (2):303–319, 2011.

[130] Tim Pohle, Elias Pampalk, and Gerhard Widmer. Evaluation of frequently used audio features for classification of music into perceptual categories. In *Proceedings of the Fourth International Workshop on Content-Based Multimedia Indexing*, volume 162. Citeseer, 2005.

[131] Todor Ganchev, Nikos Fakotakis, and George Kokkinakis. Comparative evaluation of various mfcc implementations on the speaker verification task. In *Proceedings of the SPECOM*, volume 1, pages 191–194, 2005.

[132] Michael Terrell, Joshua D Reiss, and Mark Sandler. Automatic noise gate settings for drum recordings containing bleed from secondary sources. *EURASIP Journal on Advances in Signal Processing*, 2010:10, 2010.

[133] Zheng Ma, Brecht De Man, Pedro Duarte Pestana, Dawn A. A. Black, and Joshua D. Reiss. Intelligent multitrack dynamic range compression. *Journal of the Audio Engineering Society*, 2015.

[134] Brecht De Man, Matthew Boerum, Brett Leonard, Richard King, George Massenburg, and Joshua D. Reiss. Perceptual evaluation of music mixing practices. In *138th Convention of the Audio Engineering Society*, May 2015.

[135] Alex Wilson and Bruno M Fazenda. Perception of audio quality in productions of popular music. *Journal of the Audio Engineering Society*, 2015.

[136] Alex Ross. *The rest is noise: Listening to the twentieth century*. Macmillan, 2007.

[137] Brecht De Man, Mariano Mora-Mcginity, György Fazekas, and Joshua D. Reiss. The Open Multitrack Testbed. In *137th Convention of the Audio Engineering Society*, October 2014.

[138] ITU-R BS. 1770-2. Algorithms to measure audio programme loudness and true-peak audio level. *International Telecommunications Union, Geneva*, 2011.

[139] Jonghwa Kim and Elisabeth André. Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(12):2067–2083, 2008.

[140] Kyung Hwan Kim, SW Bang, and SR Kim. Emotion recognition system using short-term monitoring of physiological signals. *Medical and Biological Engineering and Computing*, 42(3):419–427, 2004.

[141] Mathias Benedek and Christian Kaernbach. A continuous measure of phasic electrodermal activity. *Journal of neuroscience methods*, 190(1):80–91, 2010.

[142] Ian Daly, Asad Malik, James Weaver, Faustina Hwang, Slawomir J Nasuto, Duncan Williams, Alexis Kirke, and Eduardo Miranda. Towards human-computer music interaction: Evaluation of an affectively-driven music generator via galvanic skin response measures. In *Computer Science and Electronic Engineering Conference (CEEC), 2015 7th*, pages 87–92. IEEE, 2015.

[143] Hatice Gunes and Maja Pantic. Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners. In *Intelligent virtual agents*, pages 371–377. Springer, 2010.

[144] Shashank Jaiswal and Michel F Valstar. Deep learning the dynamic appearance and shape of facial action units. *Winter Conference on Applications of Computer Vision (WACV), 7-9 March 2016, Lake Placid, USA.*, 2016.

[145] Michael Grimm and Kristian Kroschel. Evaluation of natural emotions using self assessment manikins. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005*, pages 381–385. IEEE, 2005.

[146] Enrique Perez-Gonzalez and Joshua D. Reiss. A real-time semiautonomous audio panning system for music mixing. *EURASIP Journal on Advances in Signal Processing*, 2010.

[147] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105, 2011.

[148] Thomas V Perneger. Whats wrong with bonferroni adjustments. *BMJ: British Medical Journal*, 316(7139):1236, 1998.

[149] Joshua D Reiss. A meta-analysis of high resolution audio perceptual evaluation. *Journal of the Audio Engineering Society, vol. 64 (6), June*, 2016.

[150] Georgios N Yannakakis and John Hallam. Ranking vs. preference: a comparative study of self-reporting. In *International Conference on Affective Computing and Intelligent Interaction*, pages 437–446. Springer, 2011.

[151] Raimund Schatz, Sebastian Egger, and Kathrin Masuch. The impact of test duration on user fatigue and reliability of subjective quality ratings. *Journal of the Audio Engineering Society*, 60(1/2):63–73, 2012.

[152] Phillip L Ackerman and Ruth Kanfer. Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied*, 15(2):163, 2009.

[153] David Ronan, David Moffat, Hatice Gunes, and Joshua D. Reiss. Automatic subgrouping of multitrack audio. In *Proc. 18th International Conference on Digital Audio Effects (DAFx-15)*, 2015.

[154] Emmanuel Deruty. Goal-oriented mixing. In *Proceedings of the 2nd AES Workshop on Intelligent Music Production*, volume 13, 2016.

[155] Karlheinz Brandenburg and Gerhard Stoll. Iso/mpeg-1 audio: A generic standard for coding of high-quality digital audio. *Journal of the Audio Engineering Society*, 42(10):780–792, 1994.

[156] James D Johnston. Estimation of perceptual entropy using noise masking criteria. In *1988 International Conference on Acoustics, Speech, and Signal Processing, 1988. ICASSP-88.*, pages 2524–2527. IEEE, 1988.

[157] Davis Pan. A tutorial on mpeg/audio compression. *IEEE Multimedia magazine*, 2(2):60–74, 1995.

[158] Paul McNamara and Seán McLoone. Hierarchical demand response for peak minimization using dantzig–wolfe decomposition. *IEEE Transactions on Smart Grid*, 6(6):2807–2815, 2015.

[159] David Moffat, David Ronan, Joshua D. Reiss, et al. Unsupervised taxonomy of sound effects. 20th International Conference on Digital Audio Effects (DAFx-17), 2017.

[160] Ryoji Ikeda. Test pattern. Audio CD on Raster-Noton, Feb 2008.

[161] Alva Noto. Xerrox vol. 2. Audio CD on Raster-Noton, Jan 2009.