Evaluation and Modelling of Perceived Audio Quality in Popular Music, towards Intelligent Music Production

ALEX WILSON

A dissertation submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy
of the
University of Salford.

School of Computing, Science and Engineering

2017

Contents

List of Figures

List of Tables

Acknowledgements

This work could not have been completed without the support of a great many individuals. First and foremost, thanks to Dr. Bruno Fazenda for his supervision and support throughout the years. A special thanks must be made to Dr. Joshua Reiss for supervision during a six-month stay at QMUL, as well as David Ronan and Brecht De Man for the hospitality and the many interesting conversations in the years since. Thanks to Mike Senior for the useful discussions and permission to use the CMT database of mixes, which contributed greatly to the work in this thesis. It's unlikely this work would have taken the form it did without a great many tunes; thanks to Shane, Louise, Let's Set Sail and Weathervane Music for the initial inspirations. A special mention must go to my family for the support, and to all the G10/11 folk, Elle, April, Joe, Derry, Kerry, Róisín, Aoife, Kimberley and many others (Bill, Greg and Nige) for assorted acts of kindness and words of support over the years and generally being helpful when they didn't have to be.

Declaration

This is a declaration that the contents of this thesis are, except where due acknowledgement has been made, the work of the author alone. No portion of the work contained in this document has been submitted in support of an application for a degree or qualification of this or any other university or other institution of learning. All verbatim extracts have been distinguished by quotation marks, and all sources of information have been specifically acknowledged. A list of publications associated with this thesis can be found in Appendix A.

Signed:

Date:

Abstract

This thesis addresses three fundamental questions: What is mixing? What makes a high-quality mix? How can high-quality mixes be automatically generated? While these may seem essential to the very foundations of intelligent music production, this thesis argues that they have not been sufficiently addressed in previous studies. An important contribution is the questioning of previously-held definitions of a ‘mix’. Experiments were conducted in which participants used traditional mixing interfaces to create mixes using gain, panning and equalisation. The data was analysed in a novel ‘mix-space’, ‘panning-space’ and ‘tone-space’ in order to determine if there is a consensus in how these tools are used. Methods were developed to create mixes by populating the mix-space according to parametric models. These mixes were characterised by signal features, the distributions of which suggest tolerance bounds for automated mixing systems. This was complemented by a study of real-world music mixes, containing hundreds of mixes each for ten songs, collected from on-line communities. Mixes were shown to vary along four dimensions: loudness/dynamics, brightness, bass and stereo width. The variations between individual mix engineers were also studied, indicating a small effect of the mix engineer on mix preference ratings ($\eta^2 = 0.021$). Perceptual audio evaluation revealed that listeners appreciate ‘quality’ in a variety of ways, depending on the circumstances. In commercially-released music, ‘quality’ was related to the loudness/dynamic dimension. In mixes, ‘quality’ is highly correlated with ‘preference’. To create mixes which maximised perceived quality, a novel semi-automatic mixing system was developed using evolutionary computation, wherein a population of mixes, generated in the mix-space, is guided by the subjective evaluations of the listener. This system was evaluated by a panel of users, who used it to create *their* ideal mixes, rather than the *technically-correct* mixes which previous systems strived for. It is hoped that this thesis encourages the community to pursue subjectively motivated methods when designing systems for music-mixing.

1 Introduction

Generally-speaking, the music production process comprises of five steps: composition, performance, recording, mixing and mastering. In popular music[1], these five acts can be undertaken by completely separate actors, each motivated towards creating the best possible end product. Each of these actions requires a high level of creativity, technical proficiency and, ultimately, good critical listening skills.

The technical challenges faced at each step vary. To support their actions, the actor can employ the use of certain tools. For example, a composer may use specific notation software, performers take advantage of musical instruments and new music technologies for sound effects, recording engineers will choose microphones and recording environments with appropriate acoustics, mix engineers will consider many different editing and processing strategies in order to enhance the impact of the recording, while the mastering engineer might use specially-designed amplifiers and a cutting lathe to make the audio sound its best on a vinyl record.

Clearly, there is a highly-specialised use of tools, and each actor builds this knowledge over their education and subsequent career. However, there can be barriers. For the novice user, there can be a steep learning curve. For the visually-impaired user, these tools may place too much emphasis on visual feedback. For a musician with limited mobility, traditional instruments may present specific challenges.

This thesis addresses the novel research area of Intelligent Music Production (IMP). Research in this area has a variety of aims, such as improving productivity, increasing accessibility and furthering the study of psychoacoustics and music perception. IMP has been the subject of two recent workshops by the Audio Engineering Society in the UK, with a third planned for 2017[2]. It

[1] Described throughout as popular music, the audio samples used in this thesis are predominately of music featuring a consistent set of instruments and timbres, particularly vocals, guitars and drums. Where this limitation leads to potentially genre-specific analysis, it is noted (e.g. § 6.1.1).

[2] `http://www.semanticaudio.co.uk/events/wimp2017/`

is proposed that research in this exciting new area can benefit from returning to some fundamental questions about the nature of music production.

1.1 Scope of the thesis

Within these five stages of music production, this thesis is concerned with the fourth: mix engineering. To introduce this thesis, three fundamental questions are posed:

1. What is mixing?
2. What makes a good mix?
3. How can good mixes be automatically generated?

Questions 1 and 2 are fundamental and an exhaustive investigation into these questions is not possible, given the limited scope of the work. However, these three questions help to clarify the motivations behind the work. Each question receives sufficient attention within this thesis.

1.1.1 What is mixing?

When an engineer is mixing a recording, what is it that is being altered? An engineer can utilise a variety of tools to shape the multitrack audio recording and present the music in a variety of different ways. These tools include volume control, equalisation, panning and spatial effects, dynamic range compression and expansion, the addition of reverberation, delay, and a host of related tools using modulated delays.

As outlined in Chapter 2, there exists a large collection of literature which suggests how these tools could/should be used in certain situations, to mitigate certain technical issues and also for creative manipulation of sound. What this thesis seeks to address is the effect these tools have on the mix. What mixes are possible and how do they vary? These fundamental questions are addressed in Chapters 4, 5 and 6.

1.1.2 What makes a good mix?

One does not often have the opportunity to hear more than one mix of any given song. Typically, one only has access to the specific mix that was chosen by the artist/producer/engineer, was sent to the mastering engineer and was distributed to the public. Consequently, many studies of digital music signals are concerned with analysis across different songs, instruments, genres or artists, but not between mixes, due to this relative level of scarcity. In contrast, the artist will often compare many mixes of their own material. Furthermore, the mix engineer is constantly engaged with the task of comparing different mixes, different processing outcomes, different strategies for presenting the music. Finally, a producer may compare mixes from different mix engineers in order to decide which should be hired for the job of mixing further content. Because of this, developing a greater understanding of the psychoacoustics of mix-engineering is a worthwhile endeavour, yet one which has rarely received much attention.

Traditionally, occasions where the public get to hear multiple mixes of the same material are highly limited. These might include when an album has been completely remixed for re-release (comparisons between different masters are very common, as many re-releases are not remixed, only remastered). These comparisons are becoming more common, not only since the extensive catalogue of 20th century popular music allows for multiple re-releases but also because of emergent music distribution technologies. With the release of an album as an app, or with object-based broadcast technologies, more and more are listeners being exposed to alternate audio mixes. In order to further understand our perception of quality in music mixes, we wish to determine what it is that makes a mix ‘good’. This question is explored in Chapters 3 and 7 using psychoacoustic testing.

1.1.3 How can good mixes be automatically generated?

Existing automated music production tools have succeeded in generating mixes by addressing technical aspects of the mixing process (see Chapter 2). Rarely has subjectivity been so directly included in an automated mixing tool.

After the first two questions were investigated in the initial chapters, it was clear that while some degree of consensus may exist, there is not one optimal way to mix a given song. Rather, there are multiple good-quality mixes, identified according to the subjective tastes of the user. Having built this increased understanding of mixing and of quality perception in mixes, new methodologies must be explored in an attempt to incorporate human subjectivity into the production processes. This thesis addresses this final question in Chapters 8 and 9 by designing a novel music mixing system using interactive evolutionary computation.

1.2 Contributions made by this thesis

The following is a list of individual contributions to knowledge made in this thesis.

1. The 'mix space'

   (a) A definition for 'mix' and associated formulations for a space of mixes

   (b) The analysis of real-world mixes in this space

   (c) A method for creating mixes in this space by parametric models

   (d) A system for the user-guided creation of mixes, by interactive evolutionary computation, which produces a range of mixes comparable to that of a traditional fader-based interface.

2. Analysis of audio signal features in mixes

   (a) Creation of a large dataset of mixes

   (b) Analysis of audio signal features in this dataset

   (c) Factor analysis, revealing loudness/dynamics, brightness, stereo width and bass as the four dimensions of greatest variance

   (d) Classification of mix engineers using the signal features of their mixes

3. Understanding quality in mixes

   (a) Application of existing quality definitions to music mixes

   (b) The relationship between "quality" and simply liking an audio clip depends on song-familiarity

   (c) For mixes, "quality" and "like" are highly-correlated

   (d) Listeners can identify the mix engineer from the audio signal, in limited cases, although the effect of mix engineer on preference ratings is small.

In summarising these individual contributions, two macro-contributions are made.

- The development of objective techniques and approaches to mix analysis, leading towards large-scale simulations of music mixing practice

- The acknowledgement that, when mixes are to be generated for a specific user, the subjective elements of audio quality need to be incorporated into the mix-creation process.

This work has aimed to further the understanding of music mixing. This thesis provides insights into what can be achieved by mixing and the influence of audio signal processing tools on the outcome of the mixing process. These findings can be utilised to complement existing audio education pedagogy, for example, by illustrating to student mix engineers that mixes tend to vary along dimensions such as loudness, brightness and width. Examples of the extreme loud/quiet, bright/dull and wide/narrow mixes can help to illustrate what it is possible to achieve within the design space of a particular mix.

Knowing the extent of how these dimensions vary is useful for the design of automated or intelligent music production tools. With a defined distribution for each of a number of audio signal features, the system can be guided away from mixes that would be unlikely to be created by a human mix engineer. It would also be possible to guide the system towards mixes which are in line with the specified requests of the user, or explore less-expected areas of the mix-space to uncover more unusual mix results.

Ultimately, the availability of a large dataset of real-world mixes, as well as the ability to generate even greater quantities of random mixes, allows for further understanding of music informatics. It is hoped that the contributions in this thesis will aid further study in the analysis of audio signals and the generation of new audio signal features, for complex tasks such as onset detection, beat detection, genre prediction and the prediction of how a piece of music can induce emotions in a listener.

2
Review of literature

This chapter provides an outline of relevant literature at the time of writing. Many of the topics considered are still under active research. While not intending to be an exhaustive review of the related subject areas, this chapter is intended to clarify the motivations and intentions behind the research that is presented in subsequent chapters. These later sections of the thesis contain additional review of literature, as deemed appropriate.

The organisation of this chapter is as follows. First, the theory behind quality perception is explored, leading to a variety of definitions and perspectives. The application of these definitions to the case of audio quality is subsequently discussed. From here, a review of the psychoacoustics of music production is provided, detailing studies which have investigated human perception of audio processing typically found in music production. Automated music production techniques are discussed along with a review of the various system architectures used and the studies in which systems have been developed and subjectively evaluated. Finally, a brief introduction of evolutionary computing is provided.

2.1 Perception of quality

The following is a definition of the term *'quality'*, taken from the Oxford English dictionary [1]:

a. The standard of something as measured against other things of a similar kind; the degree of excellence of something, e.g. *"an improvement in product quality"*.

b. A distinctive attribute or characteristic possessed by someone or something, e.g. *"he shows strong leadership qualities"*.

Colloquially, these two meanings can give rise to some confusion when one is considering quality. An individual may become confused as to which particular quality is being assessed, or if an overall measure of goodness is the concept sought. After conducting a detailed review of available literature, a framework for quality assessment was provided by Reeves and Bednar [1], suggesting that the concept of quality can be considered from four points of view:

1. Quality as excellence or superiority
2. Quality as value
3. Quality as conforming to specifications
4. Quality as meeting or exceeding customer expectations

A summary of these four approaches, along with the strengths and weakness of each, is presented in Table 2.1. According to point #3, it is possible for anything to be of good quality if it conforms to specifications, while the specifications themselves may not be excellent, have value, or exceed customer expectations. The ISO-9000 series of standards [2] have been designed to address this point and guide the manufacturing industry towards high-quality production of goods. ISO-9000 defines quality as follows:

Definition 1. *The degree to which a set of inherent characteristics fulfil requirements*

This definition calls for product or service to have certain defined requirements, and a set of inherent characteristics that have been identified and demonstrated to influence quality. These characteristics can then be optimised in order to maximise quality. This optimisation may be subject to certain constraints, such as available resources — human, financial, temporal or otherwise.

Therefore, to aid this optimisation, there is great interest in understanding how the quality of the product will be perceived by the consumer. Consider the case of wine, which is one of the more well-studied examples in the field of food quality and preference. Seven dimensions related to quality in the specific case of red wine have been identified, namely 'origin', 'image', 'presentation', 'age', 'harvest', 'sensitivity' and 'acuteness of bouquet' [3].

A study by Thach and Olsen [4] has indicated that the primary reason a person does or does not drink wine is that they do or do not like the taste. However, notice how a number of the scales by Verdú Jover et al. [3] are related to perceptions other than taste — the presentation of the product and the image of the brand are significant. These factors provide expectations to the consumer, often conveyed through the choice of label and so the label on the bottle is important

[1] http://www.oxforddictionaries.com/definition/english/quality , accessed: 18/3/16

Table 2.1: Synopsis of the strengths and weaknesses of the four approaches to quality definition, as reproduced from Reeves and Bednar [1]

Definition	*Strengths*	*Weaknesses*
Excellence	Strong marketing and human resource benefits Universally recognisable — mark of uncompromising standards and high achievement	Provides little practical guidance to practitioners Measurement difficulties Attributes of excellence may change dramatically and rapidly Sufficient number of customers must be willing to pay for excellence
Value	Concept of value incorporates multiple attributes Focusses attention on a firm's internal efficiency and external effectiveness Allows for comparisons across disparate objects and experiences	Difficulty extracting individual components of value judgement Questionable inclusiveness Quality and value are different constructs
Conformance to specifications	Facilitates precise measurement Leads to increased efficiency Necessary for global strategy Should force disaggregation of consumer needs Most parsimonious and appropriate definition for some customers	Consumers do not know or care about internal specifications Inappropriate for services Potentially reduces organisational adaptability Specifications may become obsolete in rapidly changing markets Internally focussed
Meeting and/or exceeding expectations	Evaluates from customer's perspective Applicable industries Responsive to market changes All-encompassing definition	Most complex definition Difficult to measure Customers may not know expectations Idiosyncratic reactions Pre-purchase attitudes affect subsequent judgements Short-term and long-term evaluations may differ Confusion between customer service and customer satisfaction

in consumers' quality assessment [5]. A study by Bruwer et al. [6] has indicated that preference for wine can be influenced by a variety of additional factors, including the age and sex of the consumer. Additionally, these factors interacted with others, as consumers of varying age and sex were influenced by the bottle's label in varying ways [6].

Clearly, these scales would not be suitable for other food and drink items, and may not be accurate even for white wines, due to differences in colour, taste and odour. Babakus and Boller [7] suggested that quality is specific of a single good or service — while the term 'quality' is ubiquitous, the meaning must be carefully evaluated for each specific case. Nonetheless, the methodologies and concepts discussed in relation to food quality and preference can be important in the assessment of quality in other modalities, such as audio quality evaluation.

2.1.1 Perception of audio quality

Audio quality, generally, refers to the quality of an audio stream. However, due to the various ways in which audio can be experienced, a variety of meanings have been attributed to audio quality. The following is an overview of a number of quality-assessment concepts used in audio evaluation.

Section § 2.1 referred mainly to the perceived quality of products. Many products may be evaluated on their auditory nature, if that is perceived to be an important part of the experience of using that product (examples of products often evaluated on their auditory nature are numerous, but include vehicles, home appliances and even seating). In the context of product sound quality, Jekosch [8] has defined quality with the following statement:

Definition 2. *The result of an assessment of the perceived auditory nature of a sound with respect to its desired nature.*

This definition shares many characteristics with the model of Reeves and Bednar [1] and ISO 9000:2005 [2], in that it refers to quality with respect to a product's desired nature, something which may be unique to each product. Importantly, this definition refers to the *perceived* auditory nature, which implies that the subjective impression of the listener is being evaluated. Figure 2.1 shows how this quality judgement is made by a listener. Since the reflexion is unique to the observer, the perceived quality is also unique. However, since the result of reflexion is based on experiential, social and cultural factors, amongst others, groups of similar observers may reach a comparable understanding of quality in a given scenario.

The concept of Quality of Experience (QoE) differs somewhat from the definitions provided by Jekosch etc. as it not only considers the auditory elements of the item being evaluated but an overall quality. According to ITU-T P.10, 2008, QoE is defined as follows:

Definition 3. *The overall acceptability of an application or service, as perceived subjectively by the end user.*

Since this definition provides no information about what constitutes acceptability, the following definition from Qualinet [9] helps to clarify the concept of QoE.

Definition 4. *Quality of Experience is the degree of delight or annoyance of the user of an application or service. It results from the fulfilment of his or her expectations with respect to the utility and / or enjoyment of the application or service in the light of the user's personality and current state.*

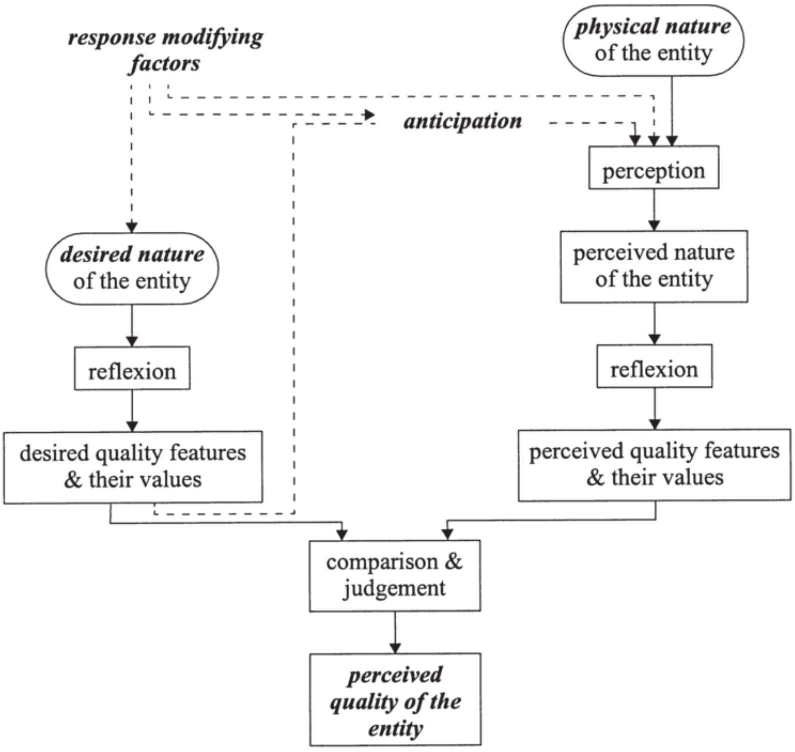

response modifying factors
physical nature of the entity
anticipation
perception
desired nature of the entity
perceived nature of the entity
reflexion
reflexion
desired quality features & their values
perceived quality features & their values
comparison & judgement
perceived quality of the entity

Figure 2.1: Judgement of quality, according to Jekosch [8]

In addition to introducing the idea of enjoyment, definition 4 refers to the user's personality and current state. This suggests that the subjective evaluation is not consistent but modulated by these factors — the same service may appear to have over-exaggerated or under-exaggerated quality depending on the mood of the user. The inclusion of emotion in a model of quality is an important consideration.

Thus far, the definitions of quality have pertained to products, applications and services. It is debatable whether an audio stream can belong to one, all or none of the categories. The answer is context-dependent. Clearly, music is marketed and sold as a product (such as a physical CD or record) but can also be delivered to a user by an application (such as audio streaming services like Spotify, iTunes etc.) which is concurrently providing a service (music-listening). For now, one can consider these definitions to have varying applicability to audio, particularly music.

A study by Blauert and Jekosch [10] has proposed a layer model of sound quality which was an attempt to structure the broad field of sound-quality evaluation and assessment on strictly perceptual grounds. Table 2.2 is taken directly from Blauert and Jekosch [10] and outlines four main categories on which quality can be perceived, along with examples of the methods which can be employed in their measurement.

2.1.2 Categorisation of sound attributes

Letowski [11] refers to audio quality as being comprised of timbral quality and spatial quality. Each of these categories is further divided into subcategories, as shown in Figure 2.2. Berg and Rumsey [12] is concerned with spatial quality and the development of scales by elicitation and structuring of verbal data, provided in response to auditory stimuli. Four categories of quality are determined from this study.

Table 2.2: Synopsis of the four identified conceptual layers of sound-quality, as reproduced from Blauert and Jekosch [10]

Conceptual Aspect	*Examples of Issues*	*Suitable Measuring Methods*
Auditive Quality (*Classical Psychoacoustics*)	Perceptual properties such as loudness, roughness, sharpness, pitch, timbre, spaciousness	*Indirect scaling:* thresholds, difference limens, points of subjective equality *Direct scaling:* category scaling, ratio scaling, direct magnitude estimation
Aural-scene Quality (*Perceptual Psychology*)	Identification and localization of sounds in a mixture, speech intelligibility, audio perspective incl. distance cues, scenic arrangement, tonal balance, aural transparency	*Discretic:* semantic differential, multi-dimensional scaling. *Syncretic:* scaling of preference, suitability, and/or appropriateness, benchmarking against target sounds
Acoustic Quality (*Physics*)	Sound-pressure level, impulse response, transmissions function, reverberation time, sound-source position, lateral-energy fraction, inter-aural cross correlation	Instrumental measurements with physical equipment for the measurement of elasto-dynamic vibrations and waves, including appropriate signal processing
Aural-communication Quality (*Communication Sciences*)	Product-sound quality, comprehensibility, usability, content quality, immersion, assignment of meaning, dialogue quality	Psychological (cognitive) tests, particularly in realistic use cases, e.g., the product in use, the audience in concert, etc., questionnaires, dialogue tests, comprehension test, usability tests, market surveys

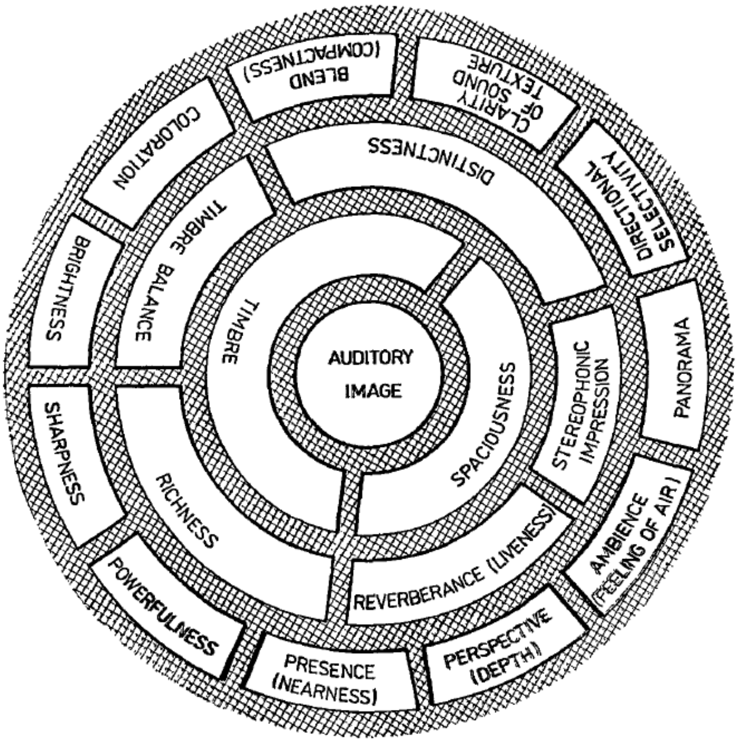


Figure 2.2: MURAL (**MU**ltilevel audito**R**y **A**ssessment **L**anguage), reproduced from Letowski [11].

- **Technical:** relating to distortion, hiss, hum, etc.
- **Spatial:** relating to the three-dimensional nature of the sound sources and environments
- **Timbral:** relating to the tone colour
- **Miscellaneous:** relating to the remaining properties

A decade later, a study by Le Bagousse et al. [13] categorised a corpus of words describing various sound attributes. While this test was a lexical study and did not have an auditory component, the categorisation of terms has much in common with the result of Berg and Rumsey [12] — four categories were obtained and are described as follows:

- **Defects:** these are interfering elements or nuisances present in a sound
- **Space:** refers to spatial impression-related characteristics
- **Timbre:** deals with the sound colour
- **Quality:** is made up of homogeneity, stability, sharpness, realism, fidelity and dynamics

This indicates a level of agreement in the ways in which audio quality is described. Interestingly, the final category of Le Bagousse et al. [13], referred to as ‘quality’, contains the terms which describe, in the language of Reeves and Bednar [1], ‘excellence’, as shown in Table 2.1 — the three other categories are more in reference to specifications.

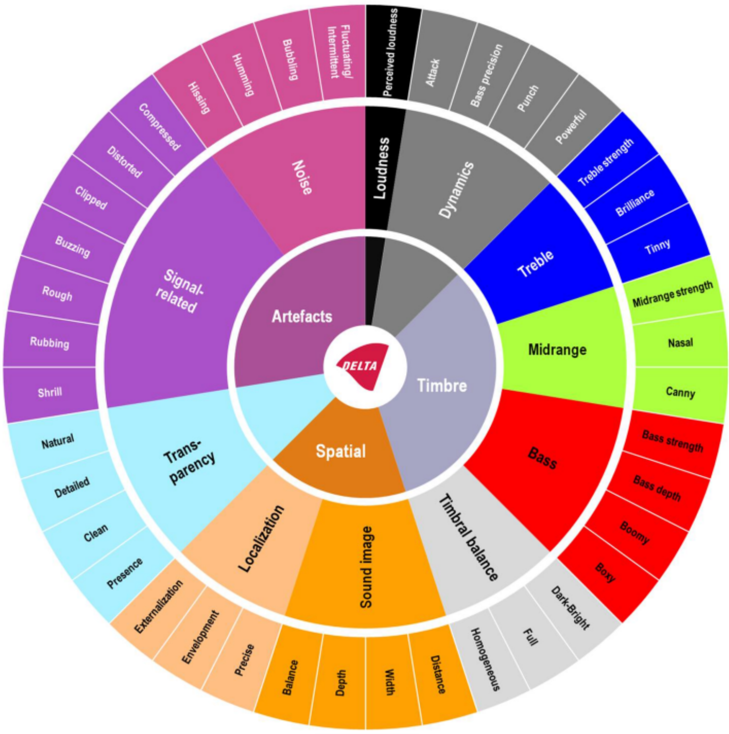


Figure 2.3: Sound wheel, for the development of a common lexicon for reproduced sound, taken from Pedersen and Zacharov [14].

Pedersen & Zacharov [14] proposed a "sound wheel", representing a lexicon of terms used to describe reproduced sound, as shown in Figure 2.3. Contained are many of the same categories that have been seen in earlier studies. Each of the terms in the outer ring has been defined and a scale provided for its evaluation.

2.1.3 Audio quality with respect to a reference example

Audio quality has an understood meaning when applied to the ability of a data compression system to reproduce audio signals at reduced bitrates. When signal information is lost, the perceived degradation of the audio experience is measured. Systems for which the perceived degradation is minimal are considered to have higher quality than those where the degradation can clearly be perceived. The following are examples of such audio quality evaluation procedures.

Perceptual evaluation of speech quality (PESQ) is a method for estimating speech quality in telecommunications systems [15]. It has been incorporated into the ITU-T recommendation P.862. PEAQ [16], or perceptual evaluation of audio quality, was originally released as ITU-R recommendation BS.1387. It is a method of predicting subjective responses to listening tests performed under ITU-R BS.1116 (methods for the subjective assessment of small impairments in audio systems). This is achieved by the use of a psychoacoustic model and audio signal feature extraction.

HASQI [17, 18], or Hearing Aid Speech Quality Index, is a measure of audio quality originally designed for the assessment of speech quality after processing by a hearing aid system. Beyond hearing aid users, HASQI has been shown to have predictive power comparable to PESQ [19]. As hearing aid processing often consists of linear filters, noise and non-linear distortions, HASQI has since been used as a measure of audio quality in a variety of non-speech sounds, even music signals [20].

This approach to quality evaluation assumes the existence of a reference. In the case of data compression systems (for these examples we can consider audio codec systems such as MP3 or AAC), a number of samples of audio are compared to one another. These samples may be created using the same codec but at varying bitrates or possibly different codecs at the same bit rate. The original programme material, from which all compressed versions were created, can be used as a reference, an example of the highest quality possible. Systems of testing in this style are described in various standards (including [21] and [22]), and include the MUSHRA (**MU**ltiple **S**timuli with **H**idden **R**eference and **A**nchor) method of audio evaluation, recently updated by Liebetrau et al. [23]. In addition to assuming a reference sample of highest quality, this method also utilises an anchor sample, of lowest quality. In MUSHRA, these samples are not explicitly revealed to test participants, as they are hidden.

It is important to consider that, in these circumstances, it is not strictly the inherent quality of the programme material that is being measured but rather the perceived degradation in quality of the signal, after being subject to destructive processes. In effect, the evaluation of the audio signal is being used as an intermediate step towards evaluating the algorithm, reproduction system, or other such device under test.

Other “MUSHRA-like” test interfaces offer variations on the theme, where the reference and anchor may be hidden, not hidden, or omitted entirely. MUSHRA was designed for the evaluation of audio codecs but has been re-imagined for other scenarios. These tests can be described using the term multi-stimulus audio evaluation and will feature in later chapters of this thesis.

2.1.4 Quality of audio programme material

Recall the statement of Babakus and Boller [7], that quality is specific of a single good or service. The approach to quality evaluation in Section § 2.1.3 is difficult to apply to music productions as it is unlikely there exists a reference audio sample (a recording of a particular song), which represents the maximum quality rating, to which all other samples (other recordings of other songs) could be compared. Nonetheless, aspects of this approach can be useful. This topic is discussed in detail in Chapter 3.

In the case of music, the perceived quality of the audio content depends on more than just the technical aspects of the signal — there are subjective and personal aspects to consider. If, as in Table 2.1, quality can be considered as value, then an audio signal representing such music that is of value to an individual may be perceived to have a high level of quality. Music that is highly liked can be considered to have high quality in these circumstances. Ratings of pleasure when listening to music are related to emotional arousal [24] and an increase in blood oxygen level in regions of the brain related to emotion has been measured when listening to familiar music [25]. Hargreaves indicated that as music becomes more familiar, it becomes liked more, although this effect can reach a point of saturation [26]. A number of studies have further investigated this relationship between familiarity and liking of music [27–29]. It is then hypothesised that this liking of the programme material influences the evaluation of more technical aspects of quality.

In ITU-R Recommendation BS.1534 [22] Basic Audio Quality is described as a “single, global attribute is used to judge any and all detected differences between the reference and the object”. While this is commonly used in tests as outlined in § 2.1.3, the discussion in this chapter thus far suggests that audio quality would be difficult to explain in one attribute. Chapter 3 will explore this in greater detail.

2.2 Psychoacoustics of music production

Music production is a diverse and complex topic. While the goal of the music production process may be to create an artistic work that delivers the required experience to the listener, the necessary skills to achieve this goal can be considered tacit knowledge of the artists, producers and engineers who work together towards this aim. Attempts have been made to represent this tacit knowledge as formal knowledge. To this end, lists of "best-practice" behaviours have been compiled, often based on extensive interviews with expert practitioners. Section § 2.3 details how these behaviours are often used as rules in the development of automated music production systems.

This section of the literature review outlines selected studies which have examined the psychoacoustics of various music production activities. Particular attention is paid to mixing practices, as this is the focus of the original work presented in later chapters. A more fundamental review of the psychoacoustics related to audio engineering can be found in a text by Zwicker and Zwicker [30]. Loudness-perception and other related topics are relatively well-studied compared to other aspects of the literature review. Consequently, there is little need for an in-depth review. An overview of loudness models can be found in Glasberg and Moore [31] and related standards [32].

2.2.1 Level-balancing

The modern day process of mix engineering began as level balancing. In the electrical era of recording, the relative levels of various microphones would be set by an engineer for recording onto the medium of choice. Rather than the term mix engineering, this task was referred to as balance engineering. Often, a mix engineer will attempt to balance the level of various instruments and sources to recreate the impression of a live performance. A microphone (or array of microphones) may be used to capture the overall sound of an instrument in a space, while additional microphones may be placed close to particular instruments, or locations surrounding an instrument. These are often referred to as "close mics" or "spot mics". It is then required to balance the relative level of these microphones to create the desired impression of space and tone. One instrument where this is often required is the drum kit. Typically, a pair of microphones is suspended over the kit, in some stereo configuration, and individual close mics are positioned to record the kick drum, snare drum and possibly other elements of the kit. While the overhead microphones will capture the sound of the kick and snare drums, the close mics are useful in helping those elements be heard clearly above other instruments, which is important in establishing rhythm. The choice of balance between overheads and close mics, or between kick and snare, is dependent on the desired sound, as influenced by the style of music.

Few scientific studies have been performed which have investigated level-balancing technique. Lembke et al. [33] tested a scenario analogous to the preceding paragraph, where the close mics of a horn and bassoon were blended, along with a microphone positioned further away in the space. Nineteen participants were asked to create a mix of these three sources which achieved "the highest degree of blend possible." The relative volume levels of these three sources was found to vary across a number of factors, including the simulated acoustic environment in which the performance took place. This concept of representing a mix as a series of inter-channel blends is similar to the work which forms the basis of Chapter 4.

Some of the earliest studies known to the author took place within the last decade and tested trivial level balancing examples. King et al [34] investigated the preferred levels within a simple mix scenario and the variance in this balance over time. Twenty-two participants were asked to balance the volume of a stereo backing track with a solo instrument or voice. This was repeated eight times for each of three different music samples — classical (soprano voice over orchestral backing), rock (vocalist over guitar/bass/drums backing) and jazz (solo trumpet over unknown backing track). Ten of the participants were tested twice, in two separate sessions, one week apart. When compared to the level set by the original engineer of each sample, the median levels found were -3.6 dB, +0.6 dB and 0 dB, for classical, rock and jazz samples respectively. Participants with a greater number of years experience displayed less variance. Since these levels are relative to those of the original mix engineer, one can observe that there is apparent consensus concerning the level of the voice in the rock sample and the level of the trumpet in the jazz sample. It appears as if the original mix engineer set the level of the soprano voice in the classical sample higher than average. With only one sample per genre, and only two solo instruments, it is difficult to generalise these results to the act of mixing as a whole. A follow-up study was conducted [35]. This investigation used three excerpts per genre (this time referred to as classical, jazz and pop). The classical and pop samples featured voice over instrumental backing, while each excerpt in the jazz category featured a different solo instrument (piano, trumpet or guitar). Results indicated that the median results for each category, when all trials were taken into account, were roughly 0 dB, 0 dB and +2 dB for the classical, jazz and pop categories respectively. This indicated a consensus for levels in classical and jazz settings but that the original mix engineer set the vocal level in the pop recordings lower than the consensus. Both excerpt and genre were found to be significant factors in the setting of level, using a repeated measures ANOVA. The level of the vocal in the pop category was suggested to be multi-modal, meaning that various optimal levels were identified. This finding suggests that there may be various levels at which to set the vocal, each acceptable to different listeners. One flaw in these studies is that the levels presented are relative to the levels set by the original mix engineer, assumed to be the ideal levels. The results are not presented in terms of an absolute measure of loudness, or a level relative to the combined mix, which would have been more insightful and repeatable.

There are few studies of complete, real-world, mixing scenarios. One study is that by De Man et al. [36] in which students were asked to mix a complete multitrack session, using a restricted but representative selection of processing options, including equalisation, panning, dynamic range compression, delay, reverberation and modulation effects. Each student had two hours to mix each session. Later, each student participated in a multi-stimulus audio evaluation test, in which they evaluated all mixes of each song, including their own mix. Preference ratings were given for each audio sample, along with comments. As the Digital Audio Workstation (DAW) session file for each audio sample was known, the relative levels of each instruments could be extracted. Figure 2.4 displays an average loudness of each of a number of instruments, relative to the combined mix, over all songs and participants. It is evident that vocals are set at higher levels than other instruments.

These studies have investigated the practice of level-balancing and indicated that some consensus can be found. As this area of study is in its relative infancy, no one test methodology has been established and utilised over a variety of studies, covering a large enough sample of

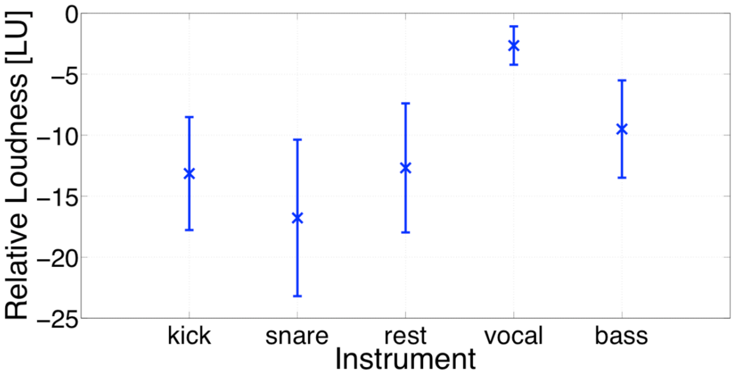


Figure 2.4: Average and standard deviation of loudness of sources relative to the total loudness of the mix, across songs and mixing engineers, taken from De Man et al. [36]. 'Rest' refers to the sum of the rest of the drums.

participants and music samples. Further studies are necessary.

2.2.2 Perception of other common processes

While a relatively large amount of research has been carried out in relation to perceived loudness and its influence on level balancing, a number of other technical aspects of the music production process have been investigated from a psychoacoustic point of view. Many of these processes, such as equalisation and panning, are themselves concerned with loudness, as equalisation is frequency-dependent loudness and panning is channel-dependent loudness.

2.2.2.1 Equalisation

Equalisation (EQ) is used to adjust the distribution of spectral energy in a sound and therefore is a vital tool in audio engineering. In applying EQ, tt is not uncommon for audio engineers to communicate with artists and other engineers using a language that contains many seemingly abstract terms. Cartwright et al. [37] investigated the ways in which equalisation is used to achieve certain auditory impressions. Each participant entered a word used to describe sound, such as "warm". Then the participant was asked to rate 40 audio samples in which the equalisation curve varied, on a scale from "not at all warm" to "very warm". From these responses, an equalisation curve relating to the supplied term ("warm") was determined. As the test was conducted on-line, the total number of training sessions included in the study was 731, where 324 unique descriptors were used.

Another study attempted to collect similar data but directly from a users DAW session. Stables et al. [38] describes the development of a series of plugins, known by the acronym SAFE (**S**emantic **A**udio **F**eature **E**xtraction). These plugins allow the current setting to be uploaded to a webserver, along with metadata, such as which instrument is being processed. A series of audio signal features are also extracted from the track, before and after processing. Upon upload, the user can describe the result by a semantic descriptor, such as "warm" or "bright". This allows other users to load settings from the webserver. A user can then process their audio tracks without direct adjustment of plugin parameters, if they so desire, but by simply choosing a semantic descriptor which best matches their desired result. Ideally, the system can learn how to associate the plugin settings to a given descriptor.

2.2.2.2 Panning

While surround sound formats are standard in cinema, and there is increasing interest in bringing surround and 3D audio systems into the home for broadcast, within music production, two-channel stereo can be considered the standard format. This has been the case for roughly 50 years. As such, the work in this thesis does not consider reproduction formats with more than two channels.

In this format, with careful set-up, it is possible create a realistic auditory scene by placing sources in phantom locations on an imaginary line between the two loudspeakers. This is typically achieved by adjusting the relative volume of the source in the two channels, resulting in the perception of inter-aural level difference (ILD) and inter-aural time difference (ITD) in the listener. A variety of laws exist for placing sources in such locations, one of the most common shown in Eqn. 2.1c. Here, θ_s is the azimuthal angle of the virtual source, θ_0 is the loudspeaker base angle (typically 30°), g_l and g_r are the normalised gains of the left and right loudspeakers and $p \in [-1,1]$, -1 indicating a pan position fully left, and 1 indicating a pan position fully right.

$$\frac{\sin\theta_s}{\sin\theta_0} = \frac{g_l - g_r}{g_l + g_r} \tag{2.1a}$$

$$g_l = \cos\left(\frac{(p+1)\pi}{4}\right) \tag{2.1b}$$

$$g_r = \sin\left(\frac{(p+1)\pi}{4}\right) \tag{2.1c}$$

The placement of sound sources in a stereo system can also be achieved by the use of delay, creating perceived inter-aural time difference (ITD) in the listener. A study by Lee and Rumsey [39] has produced the following expressions which can be used to determine the ILD and ITD required to place a source at an angle α.

$$ILD(\alpha) = \begin{cases} 0.425\alpha \text{ dB}, & \alpha \leq 20^\circ \\ 0.85\alpha - 8.5 \text{ dB}, & 20^\circ < \alpha \leq 30^\circ \end{cases} \tag{2.2a}$$

$$ITD(\alpha) = \begin{cases} 0.025\alpha \text{ ms}, & \alpha \leq 20^\circ \\ 0.05\alpha - 0.5 \text{ ms}, & 20^\circ < \alpha \leq 30^\circ \end{cases} \tag{2.2b}$$

2.2.2.3 Delay, reverberation and dynamic range compression

As previously described [39], panning can be achieved using very short delay times. More often, reverberation is used to give the impression of space or depth, to tell the listener what size space the sound was produced in. Artificial reverberation is also used for creative purposes.

According to Pestana and Reiss [40], excessive amounts of reverberation are strongly disliked. However, such a preference is context-dependant. The use of additional reverberation was commonplace in the popular music of the 1980s, at levels which may be considered excessive by modern standards. Recall from Table 2.1 that the perception of quality can change over time, as consumer expectations change. Additionally, certain styles of music are also linked to use of reverb. Certain artists and/or producers are known for the use of reverb, or lack thereof.

At the time of writing, there have been a number of recent publications on the perception of delay and reverberation: Pestana et al. [41] indicated that delay time preferences were linked to song tempo, De Man et al. [42] found that “too much reverb” was a comment often levied at mixes that received relatively low ratings of overall preference and Chourdakis and Reiss [43] proposed a method of semi-automated reverb application based on user-provided examples.

This thesis does not directly address issues relating to delay and reverberation and so this section has been added for completeness. Similarly, the scope of this thesis does not extend to models of dynamic range processing, although this topic has received attention in other works [44].

2.2.3 Effect of reproduction system / environment

The influence of the reproduction environment has been debated. This includes the influence of the room acoustics as well as the playback system being used. For example, a more reverberant control room might lead a mix engineer to add less artificial reverb than they might in a less reverberant space. Leonard et al [45] tasked 13 experienced mix engineers with adding artificial reverberation in a control room with adjustable acoustics. The programme material used was an orchestral recording made in a relatively dry hall and a soprano voice which was recorded separately. The room microphones and additional reverb, applied by the session’s original engineer, were printed to a separate track and participants were asked to set this track to their preferred level. In the more reflective condition, the level of the reverb was lower when compared to the less reflective condition and the variance was also lower.

It can be argued that, in 2017, headphone reproduction may be one of the most prevalent ways in which music is consumed. Subsequently, there is interest in knowing if mixing music on headphones, specifically for headphone reproduction, could produce high quality results. In the case of headphone reproduction, the acoustics of the room are bypassed and therefore assumed to be negligible[2]. The important factors are therefore the transfer function of the electroacoustic system and the mechanical coupling to the wearers head, the latter having an effect on low-frequency reproduction.

King et al [46] has compared the mixing behaviours of users in two conditions — loudspeaker and headphone reproduction. Ten participants were tested in both conditions, and asked to set the volume of a vocal performance relative to the instrumental backing track. Three music samples were used, representing three styles of music (jazz, classical and rock). For classical and rock samples, there was a significant difference in the balance set under the two conditions: vocals were set lower in the headphone condition for the rock sample, while, for the classical sample, vocals were set lower in the loudspeaker condition. It is hard to draw conclusions from this inconsistent result and it highlights the need for a more complete study.

[2] Virtual room acoustics and reproduction systems can, of course, be rendered over headphones using binaural technology, although this will not be considered here.

2.3 Intelligent music production

As the processes involved with producing music are often highly technical, highly creative and greatly time-consuming, there has been much research devoted to automating certain elements of the process. Often an audio engineer will approach a session from a standard point of view, each session beginning with the same fundamental tasks. An example of this may be to set the input gain of each channel, and then set the fader level to achieve a rough balance. The automation of this process would save valuable time, but also physical effort in the case of disabled users. For musicians, who may not have the technical skills to adequately record and mix their music, intelligent music production tools could be used to assist in this task. Considering this from another point of view, with more experienced engineers, the user could act as a guide to the intelligent system, allowing the system to improve over time.

As discussed in § 2.2.1, level-balancing has always been considered a fundamental aspect of music mixing. The balancing of instrument levels is often a first step in creating a mix, before the more creative processes begin. In the context of live sound, especially in amateur settings, this balancing of levels and some basic equalisation may be all that is done to the mix. Perhaps because of its relative importance, one of the first tasks attempted in automated music production was the setting of track fader levels and input gains.

Some of the earliest examples of the automated audio engineering in this context comes from the work of Duggan in the 1970's and 1980's [47]. These developments produced systems for the automatic adjustment of microphone levels, ideally for multiple speakers, as well as automated noise-gating, for feedback reduction. A summary of these developments was provided in 1992 [48]. Additional developments in this area were sporadic although an increasing number of authors referred to the concept of computers acting as assistants to mix engineers [49–52]. A renewed interest in the subject arose in the mid-2000's, spurred by advances in computer processing power and storage, machine learning, the prevalence of low-cost DAWs and the availability of multitrack digital audio on-line, among other factors.

2.3.1 Definitions of an audio 'mix'

Naturally, in order to implement automated mixing, the term 'mix' must first be defined. Izhaki [53] offers the following definitions.

> *A basic definition of mixing is: a process in which multitrack material - whether recorded, sampled or synthesized - is balanced, treated and combined into a multichannel format. Most commonly, two-channel stereo. But a less technical definition would be: a sonic presentation of emotions, creative ideas and performance.*

These definitions can be interpreted in a number of ways. The first definition is for 'mixing' (an action), but does not define 'mix' (an object). The second definition may apply to 'mix' but is not easy to implement in the form of an equation, as it is highly subjective.

The following are equations used to define a 'mix' according to various authors. Note that the nomenclature in the following equations has not been changed from the original texts. Equation 2.3 was used by Perez-Gonzalez and Reiss [54], stating simply that a mix is the sum of all channels.

$$\text{mix} = \sum_{n=1}^{N} \text{Ch}_n[t] \tag{2.3}$$

This definition seems logical, even trivial, and has become the foundation for a series of more elaborate definitions. Tsilfidis et al. [55] goes a step further in adding a gain vector, a to each track, allowing for time-dependent changes to the track gains.

$$y[n] = \sum_{k=1}^{K} a_k[n].x_k[n] \tag{2.4}$$

In a review paper from 2011, Equation 2.5 was used by Reiss [56], adding generic control vectors c which modulate the input signals x. These control vectors allow for a variety of results, such as polarity correction, delay correction, panning and source separation, depending on their implementation.

$$\mathrm{mix}_l(n) = \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} c_{k,m,l}(n) * x_m(n) \tag{2.5}$$

Each of these equations considers the mix as the sum of the input tracks, although there is little agreement on terminology or nomenclature in this general definition. It is shown in Chapter 4 that this definition is too broad for certain automated mixing tasks and, as such, a new definition is put forward in this thesis (see § 4.1).

2.3.2 System architectures

Table 2.3 refers to three types of system architecture used in automatic music production systems, namely knowledge engineering (KE), grounded theory (GT) and machine learning (ML). Knowledge engineering refers to the coding of expert knowledge, for use in expert systems. An

Table 2.3: List of selected automated music production literature broken into three categories — knowledge engineering, grounded theory and machine learning. Entries marked with * involved the development of a system. The table is expanded from that which was presented by Reiss [57] in 2015. Note that there is not much work featured from 2013 onwards: arguably, the field has moved slightly away from the development of systems and towards perceptual studies. For simplicity, studies from this thesis are not included.

Work	Refs.	Topic	KE	GT	ML
Chourdakis 2015	[43]	Reverberation			x
*Ma 2015	[58]	Dynamic range compression		x	x
Ma 2013	[59]	Equalisation			x
*De Man 2013	[60]	Various	x		
Pestana 2013	[61]	Various	x	x	x
*Ward 2012	[62]	Level balancing		x	
*Scott 2011	[63]	Level balancing			x
*Maddams 2012	[64]	Dynamic range compression	x		
*Mansbridge 2012	[65]	Level balancing	x		
Aichinger 2011	[66]	Inter-channel masking		x	
Bocko 2010	[67]	Various			x
Lopez 2010	[68]	Equalisation		x	
*Terrel 09-10	[69, 70]	Various		x	
Pardo 09-12	[37, 71, 72]	Equalisation			x
Heise 09-10	[73]	Reverberation			x
Barchiesi 09-10	[74, 75]	Various			x
Perez 07-10	[54, 76–78]	Level balancing	x		

expert system is a computer system that emulates the complex decision-making of an expert in some specific field of expertise. Knowledge engineered expert systems are well suited to problems which can be represented as decision trees i.e. where the application of a finite amount of rules can yield a decision. Applications have been found in medical diagnosis and mortgage approval. Music production, and mix engineering in this case, is arguably more nuanced than this due to its creative elements. Intelligent music production systems based on this architecture have therefore shown varied results, as discussed in § 2.3.3. Table 2.3 shows a number of studies that have used knowledge engineering in attempts to design expert systems for music mixing. This typically involves gathering "best-practice" rules from a variety of sources and implementing these rules in the system. Pestana & Reiss [40] refer to a number of these rules, which are listed in Table 2.4 — some are supported by subjective testing and some are not, demonstrating a lack of consensus on some "best-practice".

Grounded theory involves the analysis of experimental outcomes, leading to the formulation of hypotheses [79]. For music production systems, this can be achieved using psychoacoustic studies, particularly those which assess quality and preference. For this approach to be useful, the number of participants in such experiments needs to be sufficiently high and experiments must be carefully designed.

Machine learning is the field of study which exploits a computer's ability to learn without being programmed explicity, achieved by large-scale analysis of observations. Often a system is trained on a set of input data with known output and the rules learned in this training phase are applied to new inputs with unknown output. This process is known as supervised learning. Unsupervised learning is also commonly used, in situations where no labelled data is available. In this case, patterns in the data and the clustering of observations are used to infer information.

2.3.3 Subjective evaluation of systems

In order to determine if the developed system is operating as intended, and also establish whether its development can be considered a success, subjective evaluation is necessary. A number of papers listed in Table 2.3 include a subjective evaluation. This section outlines some of the issues with subjective evaluation of automated music production systems.

Scott & Kim [80, 81] have proposed a method of automatic mixing in which the instruments are identified and common instrument-specific processing is applied, based on best-practice. The processing consisted of gain adjustment, stereo panning and "coarse" equalisation. Figure 2.5 shows the result of a subjective evaluation with 15 participants. For only six of the ten songs evaluated is the proposed model preferred over the summed mix. From this result it is not clear that the system provides an advantage over the default condition, which is a mix where the gains of each track are each set to an equal, arbitrary value. This indicates the system has trouble adapting to different songs. This issue is possibly caused by the use of best-practice guidelines in the mixing process.

In the evaluation of an automatic dynamic range compression (DRC), Maddams [64] uses "no DRC", "expert manual" and four variations on their own settings, using four songs. Participants were asked to "*rate the following according to the overall quality of the mix*." The results indicate that the application of DRC did not noticeably improve the quality but that inappropriate application did reduce the overall quality, as shown in Figure 2.6.

Table 2.4: Best practice assumptions used for the design of intelligent music production systems, as reproduced from Pestana and Reiss [40]. The origin of the assumption can either be literature review (LIT), the interview process with professionals (INT), or the assumption made in previous implementations (PI). The method of testing is either through mixing exercises by professionals (EX), measuring number one hit singles for features (MM), subjective evaluation with a listening panel (SE) or a questionnaire sent to professionals (Q).

#	Title	Proven	Origin	Tested
01	All signals should be presented with equal loudness.	False	PI	SE; Q
02	The main element should be up by an understandable amount of loudness units.	True	INT	EX; MM; SE; Q
03	Vocals should be ridden above the backing track	True	INT; LIT	EX; Q
04	No element should be able to mask any of the frequency content of the vocals.	True	INT; PI	Q
05	Track panning affects partial loudness	True	LIT	EX; SE
06	Dynamic Range Compression affects relative loudness choices.	False	INT	SE
07	Low-end frequencies should be centrally panned.	True	LIT; INT; PI	MM; SE
08	The main track is always panned centrally.	True	LIT; INT; PI	MM
09	Remaining tracks are panned out of the center.	True	LIT; INT	EX; MM; Q
10	The higher the frequency content the more a track can be panned sideways.	False	LIT; PI	MM
11	Frequency balance should be kept between left and right.	True	LIT; INT; PI	MM; Q
12	Hard panning should be avoided.	False	LIT; PI	SE ; Q
13	Sources recorded with close (mono) and far (stereo) techniques simultaneously should have the mono source panned to the same perceived position featured in the stereo source.	True	INT	Q
14	Monophonic compatibility should be kept.	True	LIT, INT	MM; Q
15	Panning is mostly done audience-perspective.	False	LIT	Q
16	It is customary to apply temporal cues to panning.	False	PI	Q
17	Equalization is frequently done to avoid inter-track masking effects.	True	LIT; INT; PI	EX; Q
18	Salient resonant frequencies should be subdued.	True	INT	Q
19	High-pass filters should be used in all tracks with no significant low-frequency content.	False	LIT; PI	SE; Q
20	There is a specific low-mid region that can be attenuated to improve clarity.	False	LIT	SE ; Q
21	Expert mixers tend to cut more than boost.	False	LIT	Q
22	High Q-factors should be used when cutting and low Q-factors when boosting.	True	LIT; INT	Q
23	Equalization use should always be minimized.	False	LIT	Q
24	Every song is unique in its spectral/timbral contour.	True	INT	MM; Q
25	Reverb time is strongly dependent on song tempo.	False	INT	SE ; Q
26	Reverb time is strongly dependent to an autocorrelation measure.	True	-	SE
27	Delay times are typically locked to song tempo.	True	LIT; INT	SE ; Q
28	The pre-delay is timed as a multiple of the subdivided song tempo.	True	LIT; INT	SE ; Q
29	The level of the reverb returns is on average set to a specific amount of loudness lower than the direct sound.	True	-	SE
30	Low-end frequencies are less tolerant of reverb and delay.	True	LIT; INT	EX; Q
31	Transients are less tolerant of reverb and delay.	True	LIT; INT	EX; Q
32	The sends into the reverbs should be equalized.	True	INT	Q
33	Reverbs can be carefully substituted by delays to lessen masking effects.	True	INT	SE; Q
34	Compression takes place whenever a source track varies too much in loudness.	True	LIT; INT	EX; SE; Q
35	Compression takes place whenever headroom is at stake, and the low-end is usually more critical.	True	INT	MM; EX; SE; Q
36	Gentle bus/mix compression helps blend things better.	True	LIT; INT	SE; Q
37	There is an optimal amount of compression in terms of dB and it depends on sound source features.	True	LIT	EX; Q
38	Compression should not be overused and there are maximum values for it.	False	LIT	EX; Q
39	Compressor attack is set up so that only the transient goes through.	False	LIT	EX; Q
40	Compressor release is set up so that it is over when the next note is about to start.	False	LIT	EX; Q
41	It is acceptable to judiciously lop off some micro-burst transients to gain peak-to-RMS space.	True	-	SE ; Q
42	In deciding a tracks dynamic profile, an expert engineer will shift the focus of the listener by enhancing different tracks over time, with volume changes that may some times be quite big.	True	INT	EX; Q

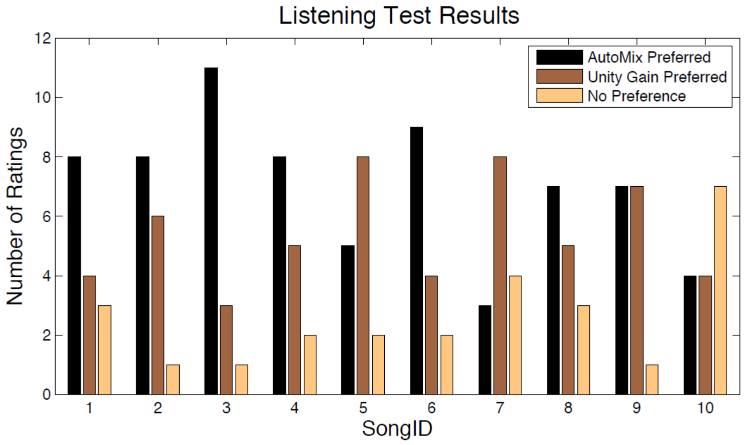


Figure 2.5: Subjective evaluation of automatic mixing system, taken from Scott and Kim [80].

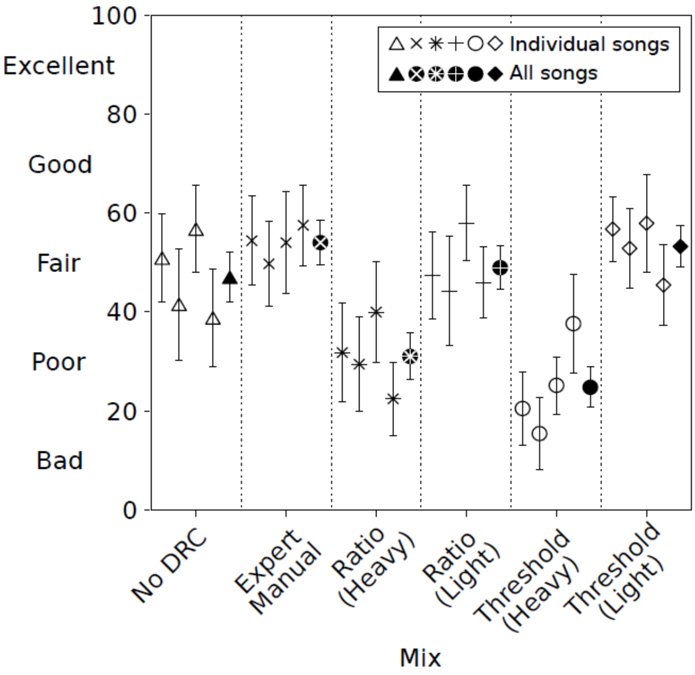


Figure 2.6: Subjective evaluation of automatic dynamic range compression system, taken from Maddams et al. [64].

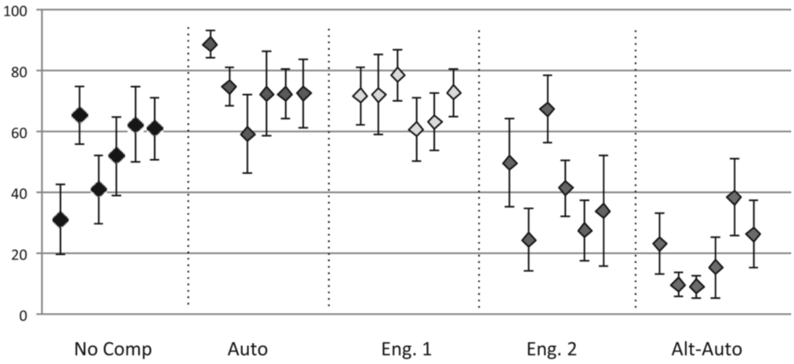


Figure 2.7: Subjective evaluation of automatic dynamic range compression system, taken from Ma et al. [58]. Y-axis shows overall preference and error bars indicate 95% confidence intervals.

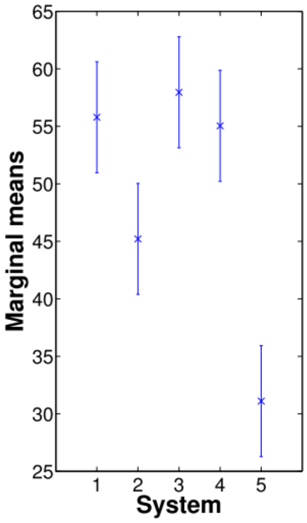


Figure 2.8: Subjective evaluation of automated mixing systems, taken from De Man and Reiss [60]. Y-axis shows overall preference and error bars indicate 95% confidence intervals. Systems are listed as follows (1: 'KEAMS', 2: 'VST', 3: 'pro 1', 4: 'pro2', 5: 'sum').

Ma et al. [58] describes a newer dynamic range compression system. In subjective evaluation, shown in Figure 2.7, the ratings of overall preference were found to be comparable to one out of two human engineers, slightly preferred to no compression and superior to a alternate implementation. Based on these results, the system is described as having an "*outstanding performance*". However, the alternate implementation is actually that of Maddams [64], which itself was reported on par with an experienced engineer, for certain settings. A previous question asked participants to rate each system according to the perceived amount of DRC applied and the perceived amount of DRC artefacts — the 'no DRC' condition scored middle of the range for both questions. This illustrates that participants were either not well trained in preparation for the test or that the concepts were not well-defined. This example illustrates the need for greater care when designing subjective evaluation experiments in this field.

A complete mixing system was implemented using the knowledge engineering approach [60]. This system was compared against two experienced engineers, an alternate implementation and an anchor condition, where the mix was created by summing all of the individual tracks, after normalisation. The alternative system was collection of VST processors developed for previous work (including [58, 64, 65]). Figure 2.8 indicates that the proposed system is preferred to the alternate implementation and the anchor, and is comparable to the two professionals.

In a later study [82], an automated system (believed to be the same as 'VST' used above) was compared against 26 mix-engineers (the same two professionals as [60] and 24 of their students). Each engineer had two hours to create a mix, using an identical array of tools. The automated system was out-performed by all 26 engineers, as shown in Figure 2.9. This suggests that there is much room for improvement in the development of future systems. However, the settings for the automated system in this study were not calibrated for each song, which may explain some of the poor performance. Additionally, many automated music production systems are designed with the live environment in mind — they operate in real-time, performing simple operations such as gain adjustment, equalisation, DRC and panning. In this study, the automated mix was being compared

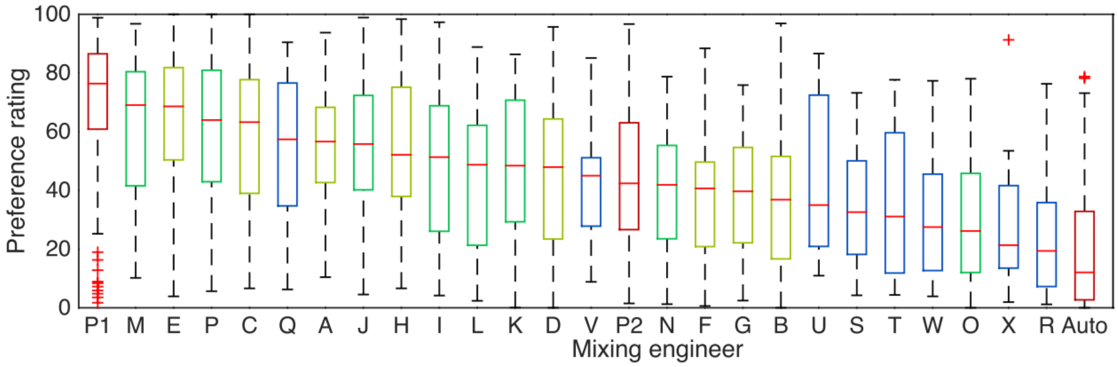


Figure 2.9: Box plot of ratings per mixing engineer, in decreasing order per median. A-H are first year students in 2013-2014 (4 songs), and second year students in 2014-2015 (1 song); I-P are second year students in 2013-2014 (4 songs), and Q-X are first year students in 2014-2015 (1 song). 'P1' and 'P2' are their teachers ('Pro'), 'Auto' denotes the automatic mix. Graph taken from De Man et al. [82].

to mixes generated in a studio environment, and it could be argued that the human engineers could exercise a greater level of creativity than the automated system was capable of.

This highlights an aesthetic consideration in the design of automated music production systems — should the system be capable of 'blindly' processing any audio material or should they require user input? If user input is not required, it may still be possible for a user to interact with the system, to further improve the mix or tailor it to their requirements. In summary, the following observations can be made:

- Many papers include a subjective evaluation.
- There is often comparison to at least one previous implementation.
- There is often comparison to some "real-world" mixes by mix engineers, although often only one or two.
- Audio samples are often rated on one attribute (such as clarity of sources, or audibility of compression artefacts) in addition to preference, yet the relationship between the two is not known

Subsequently, what is required for a detailed and fair evaluation of any proposed system is comparison with a number of other methods and comparison with a range of real-world mixes.

2.4 Evolutionary computing

One of the propositions in this thesis is the use of evolutionary computing to address intelligent music production challenges. A brief review of literature in this topic is therefore required. Evolutionary computing (EC) is a broad term referring to a series of algorithms and analysis methods often used in global optimisation problems. They are so-called as they utilise systems in which a population consisting of multiple potential solutions changes over time, *evolving* towards the optimal point in the solution space. Generally, this is achieved using some meta-heuristic derived from biological processes, noting that living systems have evolved towards optimal solutions to specific problems, such as adapting their collective behaviour in order to survive in new landscapes. This is in contrast with more traditional optimisation strategies which iterate one solution over the solution space. These methods include gradient-based or "hill-climbing" methods which require that the solution space be smooth and differentiable. EC is commonly used for problems which are non-deterministic, or non-linear, where the solution space may not be smooth and differentiable, such as logistics, scheduling, engineering and design.

2.4.1 Genetic algorithm

There is a great variety of biologically-inspired meta-heuristics used in optimisation. Perhaps the most well-known is the genetic algorithm (GA). The contemporary understanding of what constitutes a genetic algorithm owes much to the works of Holland [83] and Goldberg et al. [84], among other authors.

A genetic algorithm begins with an initial population of candidate solutions, which *evolve* towards an optimal solution in a manner akin to genetic evolution using Darwinian principles, particularly "survival of the fittest". Each solution is represented as a chromosome, a list of ordered genes. For binary GA, each gene is a value in the allele set $\{0,1\}$. For example, the chromosome $[0,1,1,0]$ contains four genes. The dimensions of the problem to be solved are represented within this chromosome, i.e. if x and y coordinates in a fixed range can each be represented as a 4-bit binary string, then the 8-bit chromosome $[1,1,1,1,0,0,0,0]$ represents a maximal value of x and a minimal value of y. Within the population of solutions, each is rated according to its quality as a solution. This is known as a fitness function. Individuals with fit solutions are chosen more frequently to "mate" with other fit solutions, thus producing offspring in the next generation of solutions. Done correctly, this allows the average fitness of the population to increase, converging on the optimal solution. The basic form of a genetic algorithm can be described as follows, illustrated in Figure 2.10.

1. Initialise population

2. Representation of population as chromosomes

3. Evaluation of population 'fitness'

4. Selection of fittest individuals for reproduction

5. Reproduction by genetic crossover and mutation

6. Repeat 3 $\rightarrow$ 5 until stop condition is met

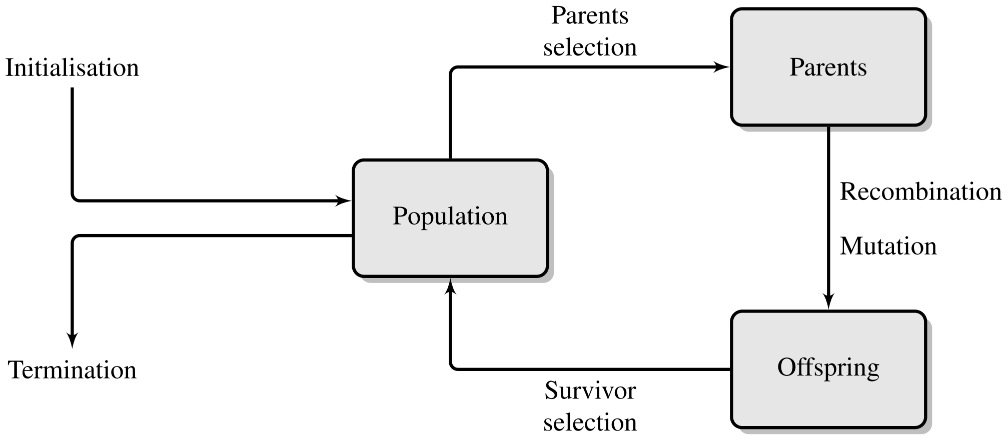


Figure 2.10: Flowchart of a basic Genetic Algorithm

For the sake of brevity at this point, further explanation of selection, crossover and mutation will be discussed in Chapters 7 and 8, within the context of the specific problem at hand.

2.4.2 Interactive Evolutionary Computation (IEC)

In problems that are highly subjective, EC methods are particularly suitable. IEC is a form of EC in which the fitness evaluation is not based on a clearly defined formula but on the subjective response of a user. IEC has been utilised in the solution of various problems which are subjective, such as fashion design [85], logo design [86] and sound design (see Takagi [87] for a detailed overview of applications). Notably, these examples all incorporate design problems in which aesthetics are important. In such applications relating to aesthetic design, there may not be a clearly defined optimal solution that is considered suitable for a range of users. Neither is the fitness landscape clearly defined. The fitness function depends greatly on what is asked of the user conducting the evaluation and their understanding of the question posed and the domain of the problem. For example, in the case of fashion design, users may be asked to rate the fitness of presented candidate solutions (outfits) where the target is a series of descriptions such as "warm, smart, casual, autumnal" etc. IEC is useful here since when attempting to solve such a problem "*...we cannot use the gradient information of our mental psychological space...*" [87].

2.4.3 Specific challenges of IEC

In IEC, the system generates solutions in the problems parameter space while the user evaluates the fitness of the solution in some psychological space, which may be unique to each user. The mapping between these two spaces may not be well-defined.

Considering that in IEC a user must evaluate the fitness of each solution, this can become a time-consuming activity, with potential for high levels of cognitive demand and eventual fatigue. This is especially a problem in audio, where each individual solution may take tens of seconds to evaluate, rather than in image evaluation, where a number of solutions can be compared side-by-side. In parallel to the emergence of IEC has been the development of hybrid methods in which a relatively small number of solutions is evaluated by the user and the remaining solutions are evaluated by extrapolation. This reduces the burden on the user for problem types where large populations are helpful. Approaches to reducing the user burden include clustering of solutions [88] and alternating user-evaluated generations with computer-evaluated generations.

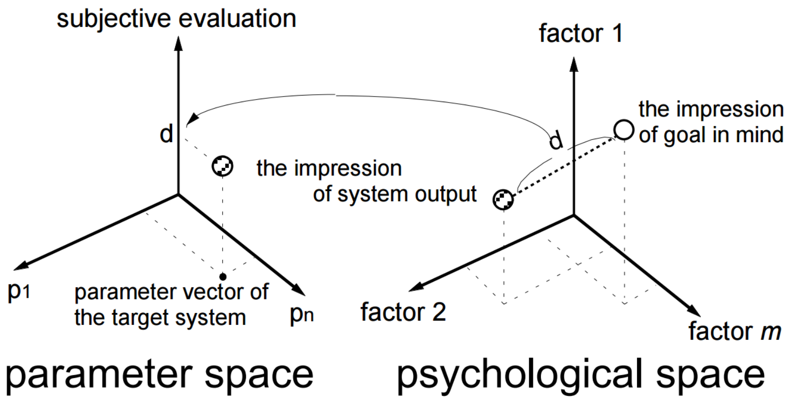

subjective evaluation
d
the impression of system output
p1
parameter vector of the target system
pn
parameter space
factor 1
d
the impression of goal in mind
factor 2
factor *m*
psychological space

Figure 2.11: Psychological distance between target in our psychological spaces and actual system output become the fitness axis of a feature parameter space where EC searches for the global optimum in an IEC system. Image taken from Takagi [87].

2.4.4 Suitability of EC to IMP problems

This thesis proposes that there exists a strong argument as to why EC is well-suited to IMP problems. This argument is based on the following.

Non-linearities — due to the perceptual nature of audio evaluation, the solution space may not be smooth and differentiable, making optimisation methods such as gradient descent difficult or impossible to apply. Additionally, as each user may have a different goal in mind, there may not exist a single global optimum. Each user may perceive a "*personal* global optimum" rather than every user agreeing on a "*universal* global optimum".

Large number of parameters — often there are a large number of parameters where the relationships between them are not well-understood. Furthering the understanding of these relationships helps construct more efficient search spaces. It is also important to establish the mapping between system parameters and perceptual factors.

Fitness functions — the definition of a "good" mix, or at least a desired mix, can be complex but is ultimately subjective. What is required is a numerical value for fitness. Quantities to be minimised include the distance to a desired target which is known in advance, or quantities thought to degrade audio quality such as inter-channel masking [66, 89]. However, if perceptual targets are being sought, such as "warmth" or "clarity", explicit subjective ratings can be used as a fitness function in place of a numerical approximation.

A synthesis of these three observations leads to the use of Interactive Evolutionary Computing. If "quality" is the variable to be optimised one must appreciate that quality can be considered as specific to a single product, good or service [7]. Recall the framework for quality proposed by Reeves and Bednar [1], repeated below. While definition #3 could possibly lead to an objective fitness function, the other perspectives suggest subjective evaluation, furthering the case for using IEC.

1. Quality as excellence of superiority
2. Quality as value

3. Quality as conforming to specifications

4. Quality as meeting or exceeding customer expectations

Many of the works in Table 2.3 were aimed at live-sound applications, i.e., real-time processing of incoming audio streams without prior knowledge, analysis of extracted features, heuristics used to guide optimisation etc. An EC-based approach may be more suited to studio environments, where processing is often applied after audio has been recorded, where there exists the time and the possibility to compare various processing decisions before arriving at the final settings. Here, there is no longer a need to analyse "live" audio as the entire audio track is known. Importantly, multiple audio tracks are known as are the relationships between them. This scenario increasingly allows for cross-adaptive effects and the temporal variation of parameters.

2.4.5 Previous work on EC in IMP

Much of the earliest applications of EC to this area are in subjects that may not be considered as intelligent music production in the modern context, but do relate to audio/acoustic engineering applications, such as filter optimisation in non-musical applications [90, 91], acoustic designs [92, 93] and binaural hearing [94, 95]. Synthesis and/or sound design is perhaps the area that has made most use out of EC-based techniques, where the parameter space of a synthesis engine is searched for optimal sounds [96–101].

Many of these prior works are based on matching a sound or mix to a target, using the distance from the target as a fitness function to be minimised. Of course, this target must be known in advance. Heise et al. [73] compared four techniques (including genetic algorithm and particle swarm optimisation) in the task of adjusting the parameters of a reverberation plug-in to best match a given room impulse response. Kolasinski [102] was concerned with matching a mix to a target, by adjusting tracks gains and using the Euclidean distance between spectral histograms as a similarity measure that was to be minimised using GA. Barchiesi and Reiss [74] also attempted matching to a given target mix, by optimising track gains and track EQ filters, using least-squares. This paper was critical of Kolasinski [102] and of GA in general for this application, stating

> *"... for the purpose of this application, the results are quite poor as the number of tracks increases and the algorithm is computationally expensive".*

These performance issues may not have been due to high-dimensionality *per se*, but rather the choice of an inefficient solution space. Chapter 4 shows that optimisation of track gains and EQ filters benefits from carefully designed solution spaces, in which each possible configuration exists only once. Additionally, computational expense is less of a problem now than in 2009. There are many more papers on various "matching to a target" applications [103–106]. What about when there is no target audio available? In place of explicit target audio there may still exist a target in some other domain, such as a perceptual target ("Make the mix sound bright/warm...etc"). Reed [49], while not using EC, does emphasise that IMP applications should be "assistants" rather than replacing the human operator. This is a philosophy that has been echoed by others [50–52] and is applied in this thesis.

2.5 Summary of literature review

In general, 'quality' can be considered as the degree to which a set of inherent characteristics fulfils requirements, as defined in Definition 1. In a more detailed fashion, quality (of experience) is described in Definition 4 and emphasises the importance of consumer expectations and emotional state in the perception of subjective quality. In the case of audio programme material (such as music), other factors are often considered such as perceived loudness, distortion, noise and other signal defects, frequency response, timbre and spatial impression. More subjectively, it is shown that, if quality can be considered as value (see Table 2.1), then liked music may be of high-quality. Familiarity is often associated with liked material, and so familarity may be related to quality.

When evaluating musical material, the impression of quality can depend on how one listens to elements of the production. For example, one may consider the quality of music to be reduced if the vocal, and thus lyrics, are unintelligible. As intelligibility of speech is also subjective, the mechanisms for this can be evaluated in different ways — is the overall level of the vocal too low, or simply too low in certain critical bands? — is the vocal masked by another instrument? An understanding of the psychoacoustics, mechanics and aesthetics of music production is important in understanding quality perception.

Automated music production systems have been developed to automate simple tasks yet the results have been mixed. This thesis is built on the following proposal: the reason for this is that the understanding of quality perception in music production is currently insufficient. Thus, by studying this specialised area of quality perception, alongside the psychoacoustics of music production, greater understanding can be reached and new systems can be developed. It is also proposed herein that evolutionary computing can be utilised to overcome some of the challenges brought on by perceptual evaluation.

3
Quality in commercially-released music

As noted by Izhaki [53, p. 7], rarely does one have the opportunity to compare more than one mix of the same song. This chapter is about the perception of audio quality in commercially-released music, where there is only one mix of each song available. The majority of the chapter refers to one experiment in particular, in which the audio stimuli were programme material that, being examples of commercially-released popular music, were familiar to participants (albeit to varying degrees of familiarity, including none). The following were the research questions which applied to the work in this chapter.

RQ-1. Are quality ratings related to objective measures of the music signal and if so, how?

RQ-2. Is the percept of liking a song distinct from that of assessing its quality?

RQ-3. What influence does familiarity with a song have on listener preference?

RQ-4. Does listener expertise have a significant influence on perception of quality?

RQ-5. Which words are used to justify quality ratings and is there significant variation in the words used to describe varying levels of quality?

Portions of the work in this chapter have been published in [107–110].

3.1 Dataset #1 — popular music, 1982 to 2016

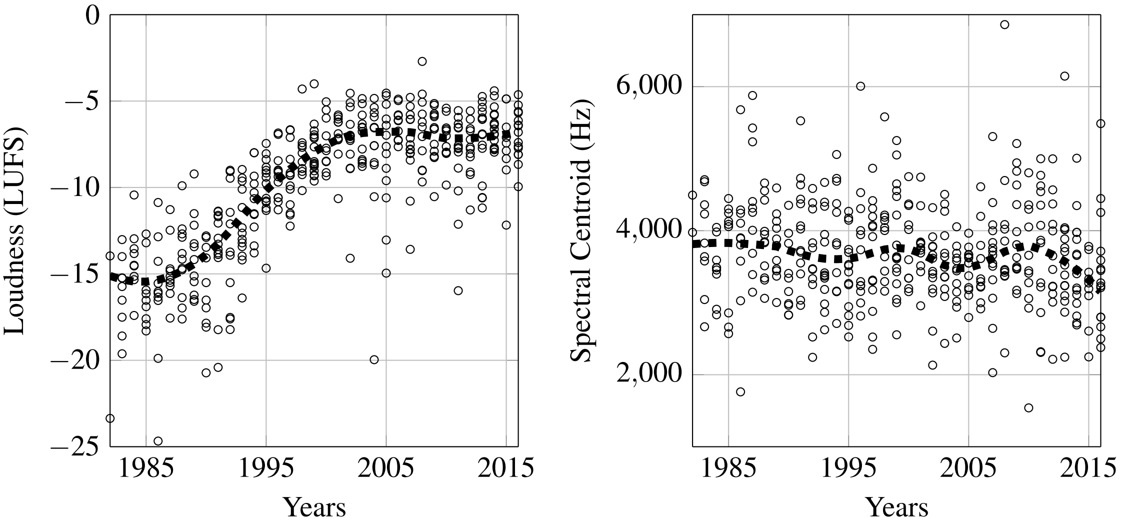
In order to provide a dataset of audio samples for study, 404 audio samples were collected from commercially released compact discs. As such, these files have a sampling rate of 44.1 kHz and a bit-depth of 16-bits. Each sample was 20 seconds in duration, centred about the second chorus of the song. This region was chosen for consistency, as a chorus is frequently a memorable centrepiece of the song. For songs without a chorus, or where the chorus does not feature the vocals, an alternative section was chosen based on audition. A one-second fade-in and fade-out were applied. In this dataset, care has been taken to include a wide variety of musical styles which were popular during the time period considered. There are at least ten audio samples from each calendar year, mostly covering pop, rock, electronic and hip-hop styles. All samples feature vocals.

The earliest samples in this dataset are from 1982. This date was chosen as it represents the commercial release of the CD format. Previous studies have studied large datasets of digital audio for feature-extraction [111], however these datasets have contained samples of music originally released long before the CD format was created. When these samples are included, they have been sourced from remastered releases, since 1982. Due to this inclusion of remastered audio samples, the results in these studies, describing the features of audio signals of music from the 1970s and earlier cannot be confidently stated. It is because of this that the current study only uses music originally released on digital media, since 1982. As with other datasets of popular music used in the literature there is a “western bias” [112], as nearly all of these samples feature vocals in English.This was due to a requirement of subjective testing, that all samples used feature vocals with lyrics in English, in order to maximise the likelihood of comprehension among test participants, the tests being conducted in the UK.

For two of the earliest samples in the dataset, the audio extracted from CD was subject to pre-emphasis, similar to the RIAA equalisation applied to vinyl records. This practice was occasionally implemented on some of the earliest commercially released CDs and was necessitated by the use of technologies originally designed for a 14-bit system, with a higher noise floor. The high frequencies are boosted at the mastering stage and a flag encoded onto the disc. This flag would engage a filter on-board the player so as to compensate for this boost, restoring the intended frequency response but with reduced noise. For the samples in this dataset with pre-emphasis, a de-emphasis filter was created according to specifications described by Galo [113].

The list of audio signal features used to characterise the dataset is shown in Table 3.3. Many of these tasks were aided by the use of the MIR toolbox [114] and additional references are shown in Table 3.3. Of all the audio datasets and feature extraction described in this thesis, chronologically, this dataset was the first to be examined. As such the set of features is not identical to those described in later chapters but those analyses were informed by the analysis in this chapter.

As shown in Figure 3.1b, the central value of spectral centroid is between 3.3 and 3.8 kHz throughout the period. This compares well with the distribution of spectral centroid in mixes, as shown in Chapter 6 (see Fig. 6.5, Table 6.6 and Fig. 6.9). It is also clear that the perceived loudness of the audio has generally increased (see Fig. 3.1a). In Figs. 3.1a and 3.1b, the smoothed lines were determined using a weighted linear least squares method, as implemented in the `smooth` function in Matlab. This method rejected outliers, defined in this case as being outside of six mean



(a) Loudness trend

(b) Spectral centroid trend

Figure 3.1: Trends in audio dataset #1. It is clear that the perceived loudness of digital music releases has increased over this timespan.

absolute deviations.

Plotting the change in a single feature over time is useful as it is repeatable and directly comparable to other studies, such as by Deruty [115]. However, a single feature does not fully explain the complex nature of loudness or brightness [1] [116–118], for example, let alone subjective impressions of audio quality. Later in this chapter, this type of plot will be re-visited using a factor-based approach, which can better reveal the combined effect of numerous signal features — *the bigger picture*.

[1] Brightness is typically considered to be well-approximated by spectral centroid alone, while § 6.1.1 indicates that additional features provide additional insights.

3.2 Experimental set-up

The test took place in the listening room at University of Salford, a room which conforms to appropriate standards set out in ITU-R BS 1116-1 [21]. In total, 63 songs were chosen for this listening test, from the dataset in § 3.1. These were chosen pseudo-randomly such that there was an even distribution over the 31 year period from 1982 to 2013. Being examples of popular music, these samples would be familiar to participants, to varying degrees. For each audio sample, the participant was asked to respond to the four questions/requests shown here;

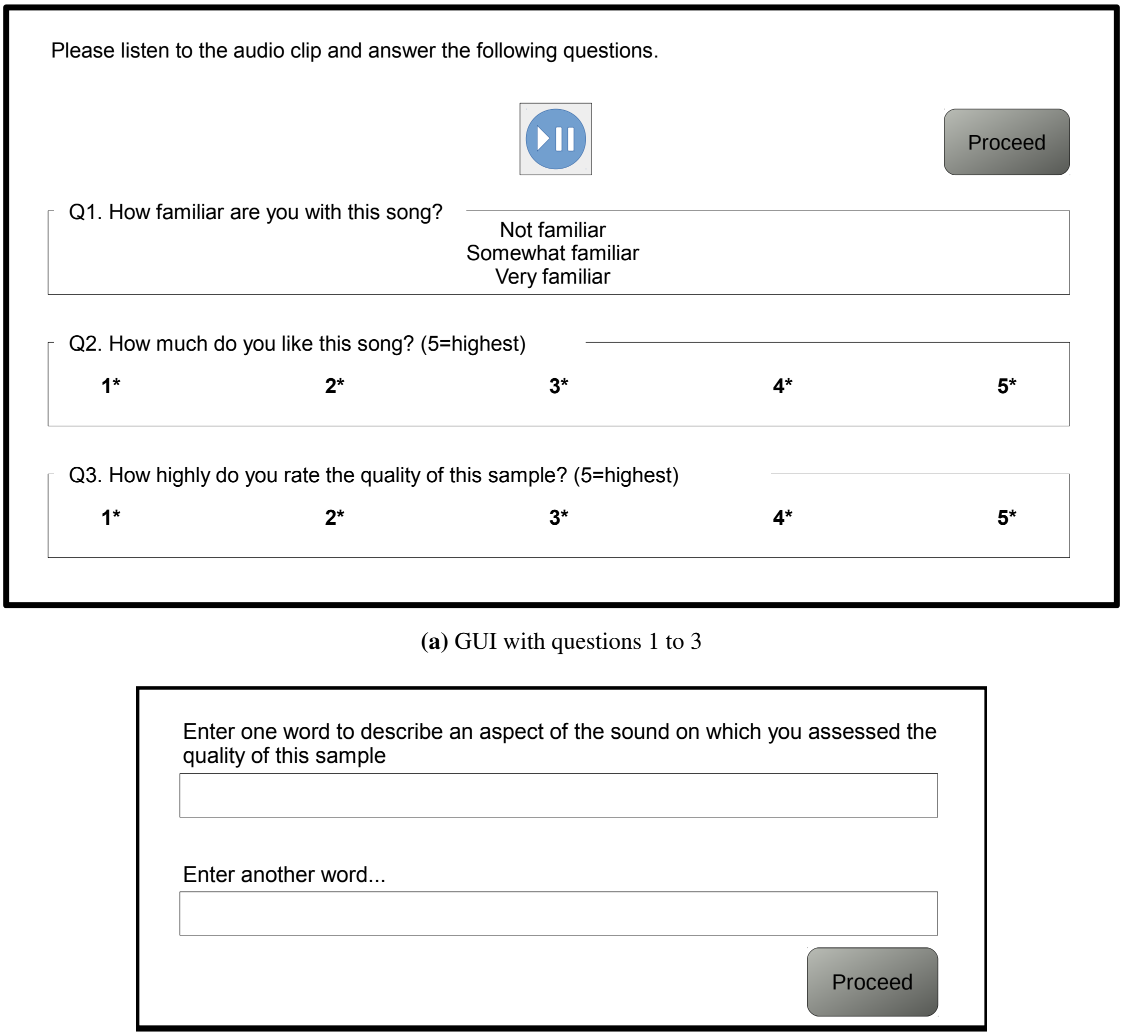
1. How familiar are you with this song?
2. How much do you like this song?
3. How highly do you rate the quality of this sample?
4. Choose two words to describe the attributes on which you assessed the audio quality

One clip was used at the beginning of each test to serve as a trial and from there on the order of playback was randomised. An optional break was automatically suggested when 40% of the trials were completed. Four questions were presented for each audio sample. The test interface for questions 1, 2 and 3 is shown in Figure 3.2a and for question 4 in Figure 3.2b. The interface also contained a play/pause button for controlling audio playback. The *like* and *quality* ratings were provided using a 5-star scale, as also used in other contemporary studies [119]. When assessing the familiarity of the sample, a 'not familiar' option was included for samples which were not familiar or previously unknown to the participant.

While *quality* was not strictly defined in this context, the request for a *like* rating in the same answer pane forces the participants into a deliberate distinction between the two. To investigate how *quality* was interpreted, the participant was asked for two words to describe attributes of the sample on which quality was assessed.

Audio was delivered via Sennheiser HD 800 headphones, the frequency response of which was measured using a Brüel & Kjær Head and Torso Simulator (HATS). Low-frequency rolloff in the response below 110 Hz was compensated using an IIR filter designed using the Yule-Walker method. As this compensation boosted the response at low frequencies, the addition of a notch filter at 0Hz was required to ameliorate the increased DC offset. To avoid clipping, audio was attenuated prior to equalisation.

The reproduction system consisted of the test computer, Focusrite Scarlett 2i4 USB interface and the headphones. The loudness of all audio samples was normalised according to BS.1770-3 [120], after the previously described headphone compensation had taken place. The target loudness for normalisation was -22 LUFS, providing ample headroom. The presentation level to participants was set to 82 dB LAeq, considered to be a suitably realistic level for headphone reproduction. This level was set by recording a 1 kHz calibration signal at 94 dB through the HATS microphone, onto the test computer. The loudness-normalised programme material was then played back over headphones situated on the HATS and recorded through the same signal chain.



Please listen to the audio clip and answer the following questions.

Q1. How familiar are you with this song?

Not familiar
Somewhat familiar
Very familiar

Q2. How much do you like this song? (5=highest)

1* 2* 3* 4* 5*

Q3. How highly do you rate the quality of this sample? (5=highest)

1* 2* 3* 4* 5*

(a) GUI with questions 1 to 3

Enter one word to describe an aspect of the sound on which you assessed the quality of this sample

Enter another word...

(b) GUI with question 4

Figure 3.2: Illustration of the graphical user interface which was used in the listening test

The total number of participants was 22 (4 female, 18 male), tested over a period of five consecutive days. Each participant was asked to choose their level of expertise, based on participation in previous listening tests. From this self-reported response, there were 13 experts and 9 non-experts. The median age of the participants was 23 years, ranging from 19 to 39 years. No participant reported any serious hearing impairment. Each participant chose two preferred musical genres as an open question — from these responses it was observed that the participants had diverse preferences, as the categories proposed by Rentfrow et al. [121] were represented (mellow, unpretentious, sophisticated, intense and contemporary). The overall test duration varied by participant, with median duration of 38 minutes, ranging from 22 to 69 minutes. As the test contained the option of a break, any effects of fatigue on the reliability of subjective quality ratings were considered to be negligible, in line with guidelines suggested in recent literature [122]. Participants were monitored from outside the room but were able to request assistance if needed.

3.3 Results of experiment

The data obtained from this experiment falls into one of two categories: subjective data gathered from test participants and signal features extracted from the audio stimuli. These are subsequently referred to by the shorthand “subjective parameters” and “objective parameters”.

3.3.1 Effect of subjective parameters

With 63 audio samples and 22 subjects, these 1386 auditions were gathered and analysis was performed on this dataset. In order to ascertain the importance of subjective measures in the assessment of *quality* and *like*, a 3-way multivariate analysis of variance (MANOVA) was performed (using IBM SPSS Statistics V.20), with independent variables of music *sample*, *expertise* and *familiarity*. The results are shown in Table 3.1. The assumptions for MANOVA were tested using Box’s test of equality of covariance matrices and using Bartlett’s test of sphericity [123]. Box’s M value of 686.15 was associated with a p-value of 0.802, which was interpreted as non-significant. Bartlett’s test yielded a significant result.

$$\chi^2(2, N = 1386) = 88.346, p < 0.001$$

These two test results indicate that the basic assumptions required for MANOVA are satisfactorily met. Using Wilks’ Λ, there was a significant effect of *sample*,

$$\Lambda = 0.597, F(124, 2144) = 5.082, p < 0.001$$

familiarity,

$$\Lambda = 0.721, F(4, 2144) = 95.313, p < 0.001$$

and *expertise*

$$\Lambda = 0.991, F(2, 1072) = 4.694, p = 0.009$$

on the ratings of *like* and *quality*. For Wilks’ Λ, the effect size is calculated as

$$\eta_p^2 = 1 - \Lambda^{1/s}$$

where $s = (\text{the number of groups} - 1)$ or the number of dependent variables, whichever is smaller. Effect sizes are shown in Table 3.1. None of the interactions were deemed to have a significant effect.

The multivariate test was followed-up by univariate analysis of variance (ANOVA), the results of which are shown in Table 3.2. For ANOVA, effect sizes are calculated according to the usual conventions [124]. Both η_p^2 and η^2 were calculated, using Eqns 3.1 and 3.2, and are shown in Table 3.2.

$$\eta_p^2 = \frac{SS_{\text{effect}}}{SS_{\text{effect}} + SS_{\text{error}}} \tag{3.1}$$

$$\eta^2 = \frac{SS_{\text{effect}}}{SS_{\text{total}}} \tag{3.2}$$

In ANOVA, as in MANOVA, none of the interactions were found to be significant, while all

Table 3.1: Results of 3-way MANOVA. Significant p-values (<0.05) are highlighted by an asterix.

Effect	Wilks' Λ	F	Hyp. df	Error df	p	η_p^2	Obs. power
Sample	.597	5.082	124	2144	.000*	.227	1.000
Familiar	.721	95.313	4	2144	.000*	.151	1.000
Expertise	.991	4.694	2	1072	.009*	.009	.788
S×F	.808	1.009	220	2144	.162	.101	1.000
S×E	.879	1.151	124	2144	.127	.062	1.000
E×F	.997	.672	4	2144	.611	.001	.221
S×F×E	.884	.937	146	2144	.689	.060	1.000

Table 3.2: Results of 3-way ANOVA follow-up. Significant p-values (<0.05) are highlighted by an asterix.

Source		df	F	p	η_p^2	η^2	Obs. power
Sample	Like	62	4.418	.000*	.203	.127	1.000
	Quality	62	5.542	.000*	.243	.201	1.000
Familiar	Like	2	201.927	.000*	.273	.187	1.000
	Quality	2	20.360	.000*	.037	.024	1.000
Expertise	Like	1	4.126	.042*	.004	.002	.528
	Quality	1	7.532	.006*	.007	.004	.783
S×F	Like	110	1.170	.121	.107	.060	1.000
	Quality	110	.977	.551	.091	.063	1.000
S×E	Like	62	1.167	.181	.063	.033	.998
	Quality	62	1.027	.422	.056	.037	.992
E×F	Like	2	1.230	.293	.002	.001	.269
	Quality	2	.230	.794	.000	.000	.086
S×E×F	Like	73	.907	.697	.058	.031	.990
	Quality	73	.992	.498	.063	.042	.995
Error	Like	1073					
	Quality	1073					
Total	Like	1386					
	Quality	1386					

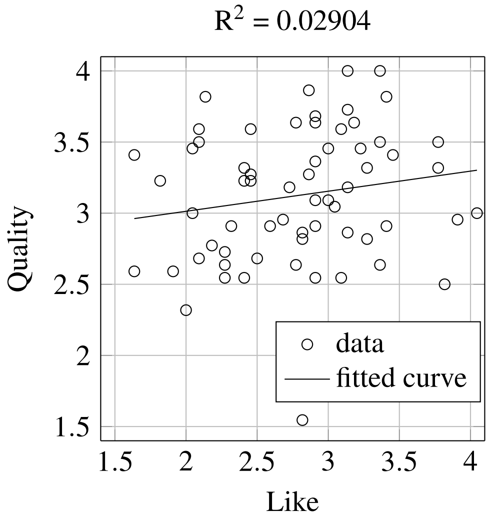


Figure 3.3: Scatterplot showing the correlation between like and quality ratings. Each point represents the mean rating for each audio sample.

main effects were significant. While the MANOVA test showed a correlation between raw *like* and *quality* ratings of $R^2 = 0.26$, when mean *like* and mean *quality* values are evaluated for each song, the value of $R^2 = 0.03$, a non-significant correlation. The mean *like* and *quality* ratings for each audio sample are shown in Figure 3.4, arranged in order of ascending *quality* illustrating the non-existing correlation. Figure 3.3 indicates the correlation between mean like and quality ratings for each sample.

Expertise does not appear to be as important a factor in this study as evidenced by the lower η^2 and observed power in Table 3.2. There is a large effect of the variable *familiarity* on *like* ratings (which will be discussed later) and a small effect of *familiarity* on *quality* ratings.

3.3.2 Effect of objective parameters

Features extracted from the signal were compared against *quality* and *like* ratings gathered by the subjective test. A linear function was fitted using the mean *like* and *quality* ratings for each song and the goodness-of-fit is shown by the coefficient of determination R^2 and associated p-values in Table 3.3. Features for which a significant correlation was found (where $p < 0.05$) are highlighted in bold. Since the value shown is R^2, which spans the range 0 to 1, arrows indicate positive ($\Uparrow$) or negative ($\Downarrow$) correlation, as determined by the polarity of Pearson's r.

From this data it can be seen that there is a difference between the *quality* and *like* ratings in terms of responsible parameters. *Like* ratings were generally correlated with spectral features while *quality* ratings were correlated with amplitude features. The correlations with *emotion factors* support this. *Quality* was correlated with both *RMS* and *roughness* while *like* was correlated with *spectral spread*. *Spectral flux* serves as both an indicator of amplitude and spectral characteristics — higher values indicate greater amplitudes and were negatively correlated with *quality*. In this study, there was no significant correlations found between spatial features or rhythmic features and either *like* or *quality* ratings.

In order to reduce the dimensions of the feature space, Principal Component Analysis (PCA) was used. This process attempts to construct a set of orthogonal features which are algebraic sums of the input vectors and explain as much variance as possible. This process can reduce the

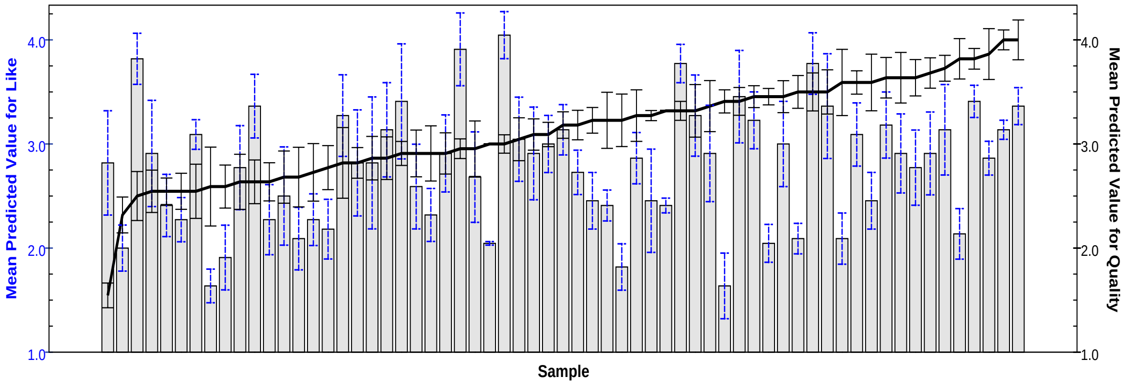
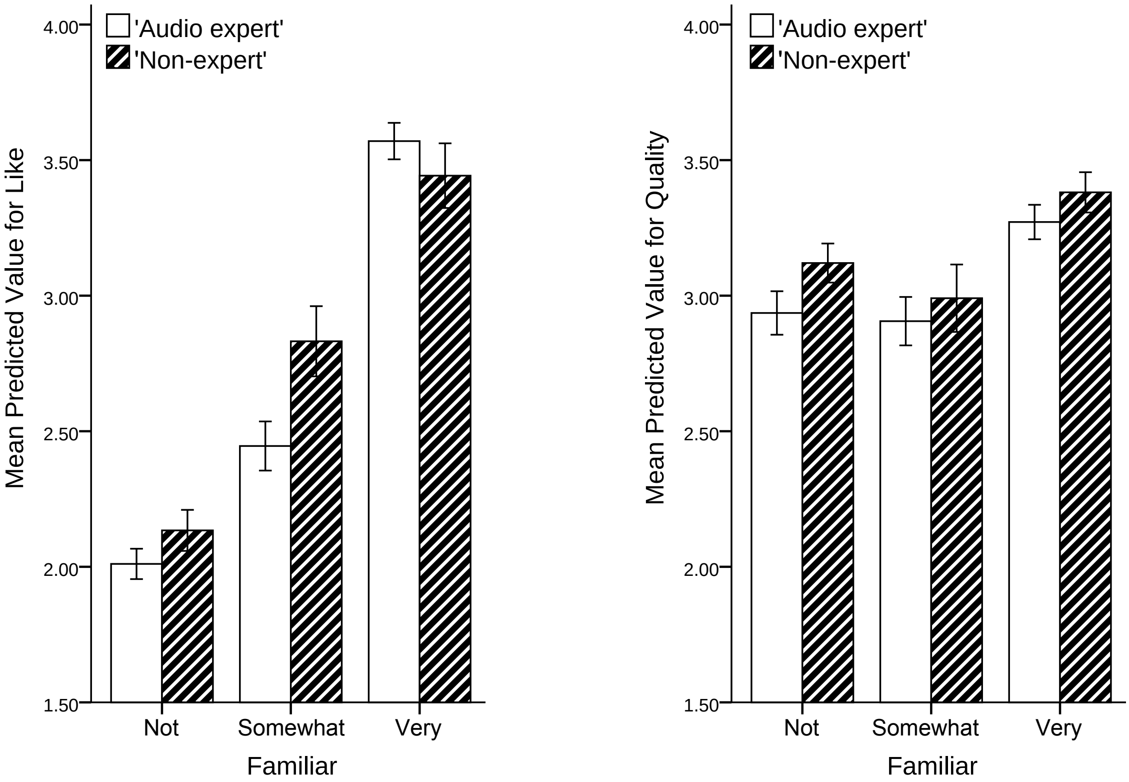


Figure 3.4: Average *like* (bar plot) and *quality* (line plot) ratings for each sample, with 95% confidence intervals.



(a) *Like* ratings

(b) *Quality* ratings

Figure 3.5: Mean and 95% confidence interval for *like* and *quality* ratings over each familiarity rating and expertise group

Table 3.3: Correlation of features with subjective results. Significant correlations (where $p < 0.05$) are highlighted in bold and considered for PCA. Features with KMO < 0.6, marked with an asterix, are not included in the PCA.

Type	Feature	Quality		Like		KMO
		R^2	p	R^2	p	
Amplitude	Crest factor	**.125**⇑	.004	.000		.842
	Loudness [120]	**.160**⇓	.001	.002		.915
	Top1db [125]	**.078**⇓	.028	.000		.833
	Gauss [107]	**.201**⇑	.000	.000		.835
	PMF Kurtosis	**.108**⇑	.009	.000		.646
	PMF Flatness	**.081**⇓	.025	.001		.951
	PMF Spread	**.155**⇓	.002	.002		.837
Spectral	Spectral Centroid	.000		.061		
	Rolloff85 [126]	.008		**.137**⇓	.003	.732
	Rolloff95	.039		**.086**⇓	.024	.663
	Harsh [107]	.058		**.201**⇑	.000	.363*
	LF Energy [107]	**.065**⇑	.046	.016		.489*
Spatial	Width-all (all freq.)	.000		.027		
	Width-band (200Hz-10k)	.013		.035		
	Width-low (0-200Hz)	.000		.006		
	Width-mid (200Hz-2kHz)	.037		.047		
	Width-high (2kHz-10kHz)	.008		.028		
Rhythm	Tempo	.000		.037		
	Event density	.000		.005		
	Pulse clarity	.021		.005		
Emo. Factors [127]	RMS	**.166**⇓	.001	.004		.829
	Max. summarised fluctuation	**.065**⇑	.045	**.079**⇓	.027	.493*
	Spectral spread	**.143**⇑	.002	**.076**⇓	.030	.804
	Avg. HCDF	.001		**.068**⇓	.040	.471*
	Roughness	**.289**⇓	.000	.036		.826
	Std.dev. roughness	**.153**⇓	.002	.006		.812
Spectral Flux [128]	Band 1 (<50Hz)	**.067**⇓	.043	.014		.858
	Band 2 (50-100 Hz)	.053		.002		
	Band 3 (100-200 Hz)	**.221**⇓	.000	.024		.910
	Band 4 (200-400 Hz)	**.132**⇓		.023		.844
	Band 5 (400-800 Hz)	**.153**⇓		.013		.884
	Band 6 (800-1600 Hz)	**.222**⇓	.000	.009		.900
	Band 7 (1.6-3.2 kHz)	**.277**⇓	.000	.049		.938
	Band 8 (3.2-6.4 kHz)	**.274**⇓	.000	.038		.851
	Band 9 (6.4-12.8 kHz)	**.179**⇓		.003		.886
	Band 10 (12.8-22.05 kHz)	**.071**⇓		.031		.831

Table 3.4: Calibration of Kaiser-Meyer-Olkin measure of sampling adequacy, from Dziuban and Shirkey [129], based on Kaiser and Rice [134].

KMO	Interpretation
Above .90	Marvellous
In the .80s	Meritorious
In the .70s	Middling
In the .60s	Mediocre
In the .50s	Miserable
Below .50	Unacceptable

dimensions of the feature space to a small number of principal components which together explain most of the variance in the problem. In order to remove features which do not reveal information about the subjective parameters, only the statistically significant features from Table 3.3 were initially considered for use in the PCA. The appropriateness of PCA was tested as follows, based on a scheme proposed by Dziuban and Shirkey [129], and using R, a language and environment for statistical computing and graphics [130]. Using Bartlett's test of sphericity (using the `psych` package [131]), the null hypothesis that the correlation matrix of the data is equivalent to an identity matrix was rejected.

$$\chi^2(325, N = 62) = 2674, p < 0.001$$

This indicates that factor analysis can be performed. The Kaiser-Meyer-Olkin measure of sampling adequacy (*KMO*, see Eqn. 3.3 [132]) was calculated for the full feature set and returned a value of 0.837, above the recommended value of 0.6 suggested by Hutcheson and Sofroniou [133], and by Kaiser and Rice [134], who suggested a calibration of the index, shown in Table 3.4. This result suggested that such a factor analysis would be useful. The value of 0.6 was chosen as the cut-off, as it was both a more conservative and more contemporary value. The communalities were all above 0.3, further indicating that each variable shared some common variance with others. The *KMO* for each of the significantly correlated variables is shown in Table 3.3. Only variables with $KMO > 0.6$ were used as input variables for PCA.

$$\mathrm{KMO} = \frac{\sum\limits_{j \neq k}\sum r_{jk}^2}{\sum\limits_{j \neq k}\sum r_{jk}^2 + \sum\limits_{j \neq k}\sum q_{jk}^2} \tag{3.3}$$

In Eqn. 3.3 q_{jk}^2 are the squares of the off-diagonal elements of the anti-image correlation matrix $SR^{-1}S$, where R^{-1} is the inverse of the correlation matrix and S^2 is the diagonal matrix $(diagR^{-1})^{-1}$, and r_{jk}^2 are the squares of the off-diagonal elements of the original correlations [2].

PCA was performed using R and the `FactoMineR` package [135]. *Quality* and *like* ratings were considered as supplementary quantitative variables, meaning that they were not used as inputs for the calculation of principal components, only that they were included in the output data

[2] R code for the calculation of KMO can be obtained from `http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/en_Tanagra_KMO_Bartlett.pdf`

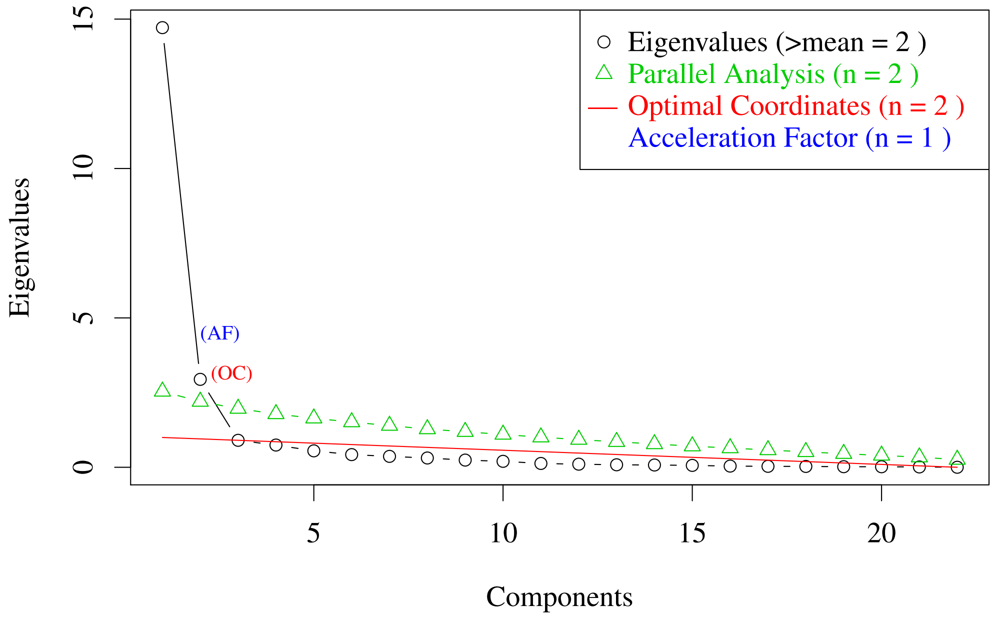


Figure 3.6: Scree plot with non-graphical solutions indicating two components be retained. These first two components account for 80.2% of the total variance of the input.

and compared against the components (see Figure 3.7).

In order to determine the number of components to retain from the analysis, a typical approach is to inspect the scree plot of eigenvalues and determine the “knee” in the curve by visual inspection. For a more quantitative approach, the following method was used. Using the `nFactors` package [136] a variety of methods were employed in order to determine the number of dimensions to keep for further analysis, shown in Figure 3.6. Kaiser’s rule [137] suggests retaining those dimensions with eigenvalues greater than 1, which in this case was the first two components. The acceleration factor (AF) [136] determines the knee in the plot by examining the second derivative. This method would retain only the first dimensions but is known to underestimate [138]. The optimal coordinates (OC) method [136] suggested that the first two dimensions be kept. Parallel analysis (PA) [139] also suggested that the first two dimensions were suitable to retain. Additionally, these two components have eigenvalues >1. Based on agreement suggested by three of the four methods, two dimensions were kept for the subsequent analysis. As all variables were significantly correlated with at least one of these two principal components, there was no reason to exclude any additional variables at this stage.

From Figure 3.7 it can be seen that the first principal component (*dim. 1*) represents variables associated with amplitude features, such as *crest factor*, *loudness*, *PMF kurtosis* and all *spectral flux* bands. The second principal component (*dim. 2*) describes high-frequency spectral features, such as *rolloff85* and *rolloff95*, along with the highest bands of *spectral flux*, all related to the positive values. The projection of *quality* along the negative direction of *dim. 1* indicates that higher ratings were associated with recordings with greater dynamic range, such as high *crest factor* or *PMF kurtosis*. *Quality* is also projected along the positive axis of *dim. 2*, although its loading on this dimension is comparatively low. *Like* ratings show no noteworthy correlation with

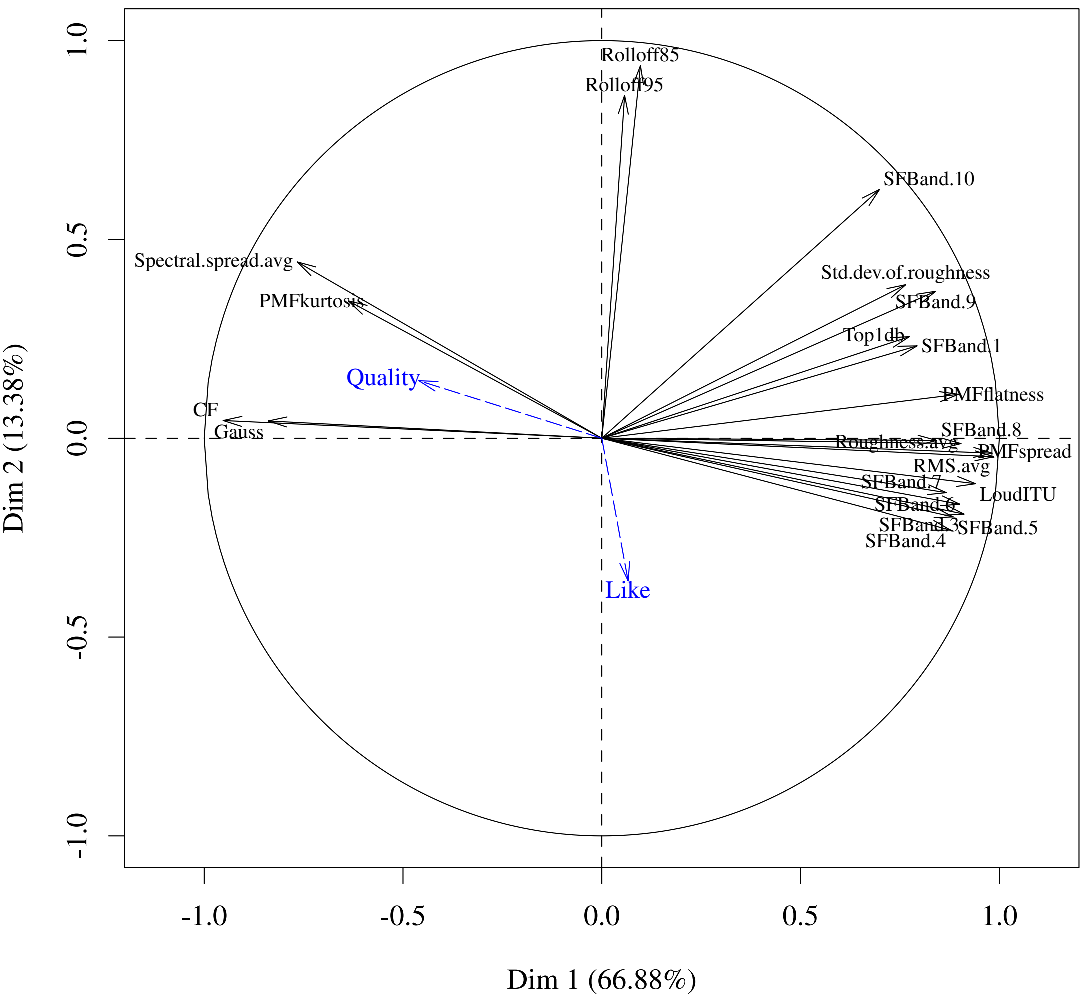


Figure 3.7: Correlation circle, showing components 1 and 2. *Dim. 1* can be explained by amplitude-based features and *dim. 2* by mostly spectral features.

dim. 1, indicating that amplitude-based features do not appear to play a strong part in listener hedonic preference. There was however, a preference for less treble frequencies, indicated by the low values of *rolloff* features. This negative correlation to *rolloff* (as shown in Table 3.3) supports the relation between *like* ratings and a peak in mid-range frequencies, or a simple disliking of samples with too great an emphasis on high-frequencies, also seen in other related studies (see § 6.2). These results for *like* are not surprising since the rating of how much a listener likes a song seems to be dependent on aesthetic and musical content and ultimately, *familiarity*, as indicated by Fig. 3.4 and discussed later.

Table 3.5 shows the R^2 values of linear fits of both *quality* and *like* ratings to the dimensions of the principal component analysis. From this it can be seen that *quality* is significantly and negatively correlated to *dim. 1* ($R^2 = 0.212$) but not *dim. 2* ($R^2 = 0.021$), and that *like* is significantly, but negatively, correlated to *dim. 2* ($R^2 = 0.129$) but not *dim. 1* ($R^2 = 0.004$).

Figure 3.8 shows the 63 audio samples plotted against the first two principal components. As the release year of each sample is known, the samples can be grouped by decade. The group

Table 3.5: Correlation of subjective response variables to principal components. Value shown is R^2. Significant correlations highlighted in **bold**.

	Dim. 1	Dim. 2
Like	.004	**.129**
Quality	**.212**	.021

centroid and 95% confidence ellipses for the population centroid are shown for the four categories of 1982-1989, 1990-1999, 2000-2009 and 2010-2013. The data shows that, even with relatively few audio samples per decade, there is an observable difference between the centroid of the 1980s, 1990s and 2000s categories along the first dimension. Due to the smaller size of the 2010s category, the confidence ellipse is relatively large.

The location of each decade centroid on *dim. 1*, which is negatively correlated to *quality*, increases chronologically. This result suggests that, according to the test panel and their definition, *quality* seems to have decreased over the decades, mainly due to a change in features associated with dynamic range, as addressed in other studies [108, 140]. This should be considered as an indicative result due to the relatively low number of audio samples and it is important to stress that *like* ratings were not influenced by this trend.

It should be noted that the use of the decade of release as a discrete qualitative variable is not without problems. Release date, as a variable, is effectively continuous and so one would expect to find little difference between 1989 and 1990 but a noticeable change from 1980 and 1999. Consequently, we see that the four decade categories in this study would not be easily separable in a multi-dimensional feature space, implying an upper limit to the success of decade-prediction tasks, and helping to explain why such attempts to categorise audio by decade have had limited success [125].

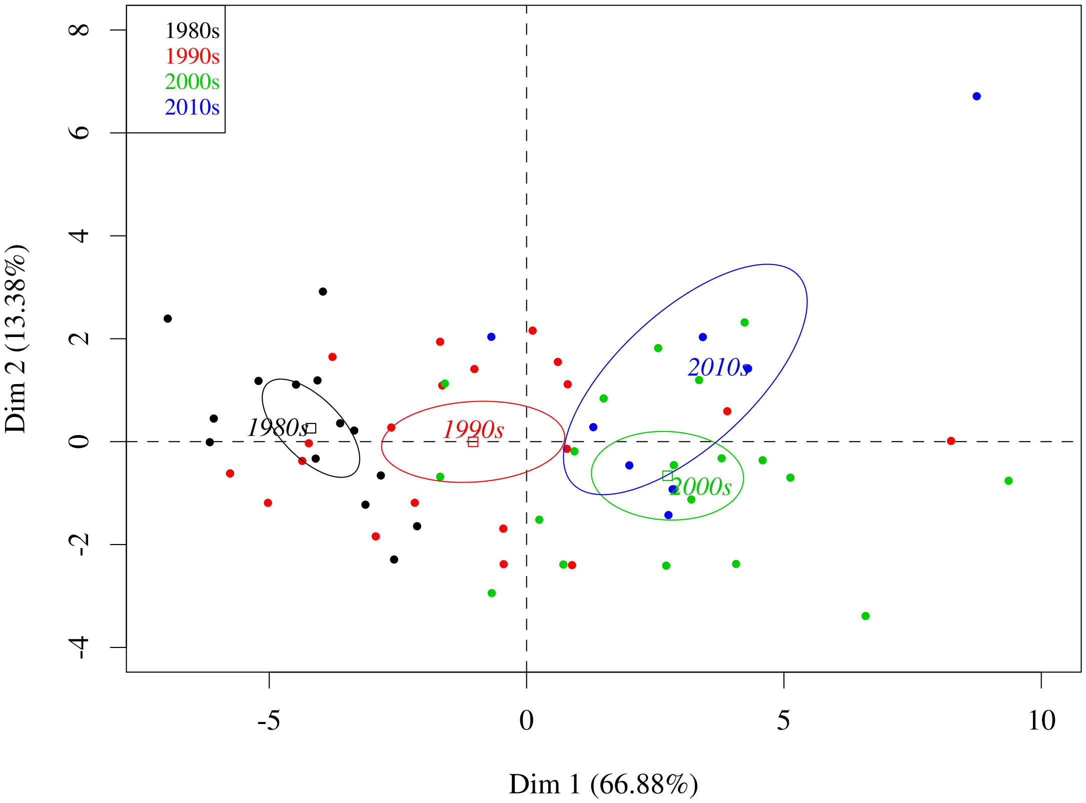


Figure 3.8: Individual samples plotted in PCA space, grouped by decade of release. The centroid of each group is marked by solid markers and the ellipses represent regions of 95% confidence in the population centroid of that group.

3.4 Words used to justify ratings

For each audition, each participant was asked to provide two words to describe attributes on which quality was assessed (see Fig. 3.2). This allowed for a larger corpus to be gathered than if a single word had been requested. This section describes the analysis methods which were applied to this data in an attempt to further understand the perception of quality.

3.4.1 Methodology

Once all data had been gathered, missing values were replaced with the term 'blank' which could then be removed from further analysis. Spelling was corrected and terms deemed to be equivalent were collated (such as 'compressed' and 'over-compressed'). This resulted in 255 unique terms, over 2669 instances. A term-frequency matrix was generated using the **R** statistical computing environment along with the `tm` package [141]. From this term-frequency matrix it can be seen that the 3 most frequently occurring words account for approximately 14% of all instances, while the top 20 terms account for approximately 54% of all instances. This shows that many terms are only used a small number of times. This relation between term-frequency and term-rank is found in larger examples of linguistic corpora [142] and will be exploited later to determine the most relevant words to analyse further.

In order to inspect the relationships between the words used and the individual audio samples, participants and quality ratings, a series of network graphs were constructed as follows. For each desired network a list of nodes and edges was created. This data was saved as a .CSV file and imported into Gephi, an open-source software for exploring and manipulating networks [143]. Graph layout, as shown in Fig. 3.9, 3.10, 3.11 and 3.12, used the ForceAtlas2 algorithm [144] to position the nodes relative to one another. Three types of graph were generated. For each graph, the size of each node is proportional to the degree of the node (the number of connections) and the thickness of lines between nodes indicates the weight of the edges (the number of times that connection is made by participants).

3.4.1.1 Term Network

Here edges are drawn between individual terms and so the list of edges is simply the list of the participants' responses. In other words, for a given audition, a certain participant may have used the terms 'compressed' and 'loud' to justify their quality rating. This is described by a single edge, between two nodes, labelled 'compressed' and 'loud'. As the complete graph contains 255 nodes (one node for each of the 255 terms, shown in Figure 3.9), a subset of this graph is shown in Fig. 3.10. This smaller graph displays only the nodes with degree greater than 10.

3.4.1.2 Term-Quality Network

Here edges are drawn between pairs of terms (as above) and also between terms and any of the five quality ratings which were awarded. For example, if the term 'distorted' is used to describe a sample which was rated $2/5$ by one participant and used to describe a sample rated $1/5$ by another, or for another sample, then edges are drawn from the node labelled 'distorted' to the nodes labelled '1' and '2'. In Figure 3.11 the quality ratings are shown in red, while words are shown in blue.

3.4.1.3 Term-Participant Network

This network shows the words used by each of the 22 participants. The users considered to be experts are shown in yellow, the non-experts are shown in red and the words are shown in blue.

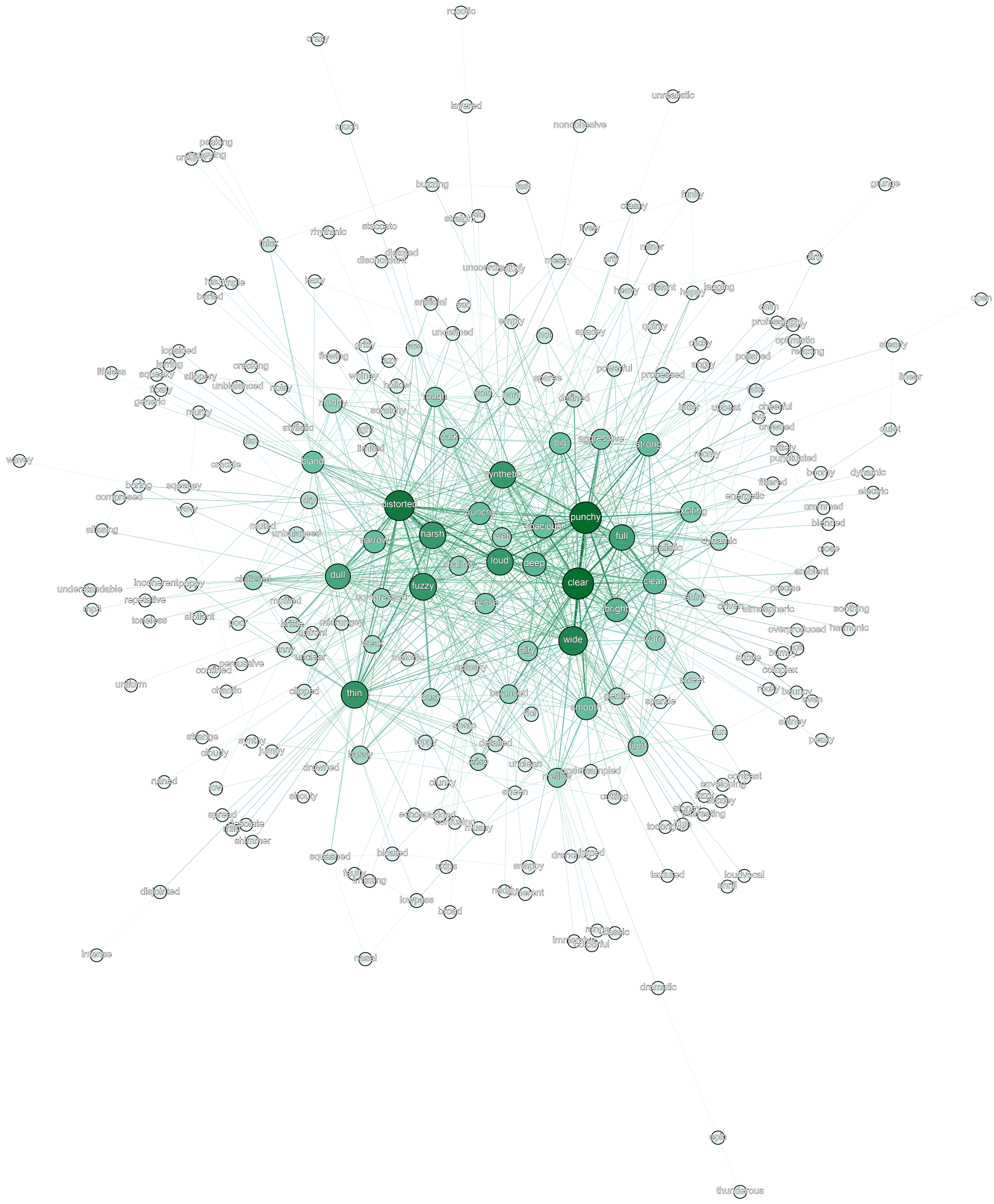


Figure 3.9: Term network, with 255 nodes (some are cropped out to fit on this page). By using the ForceAtlas2 layout algorithm, terms which are frequently used together are located closer to one another than terms which are rarely used together.

Edges are drawn between a participant and a word, the weight of the edge referring to how many times that participant used that word.

3.4.2 Metrics

In order to characterise each of the terms used in an objective manner, a series of metrics were introduced. Each term is scored based on the properties of each network allowing insights into how the terms were used in the experiment, and how the terms were organised by the participants.

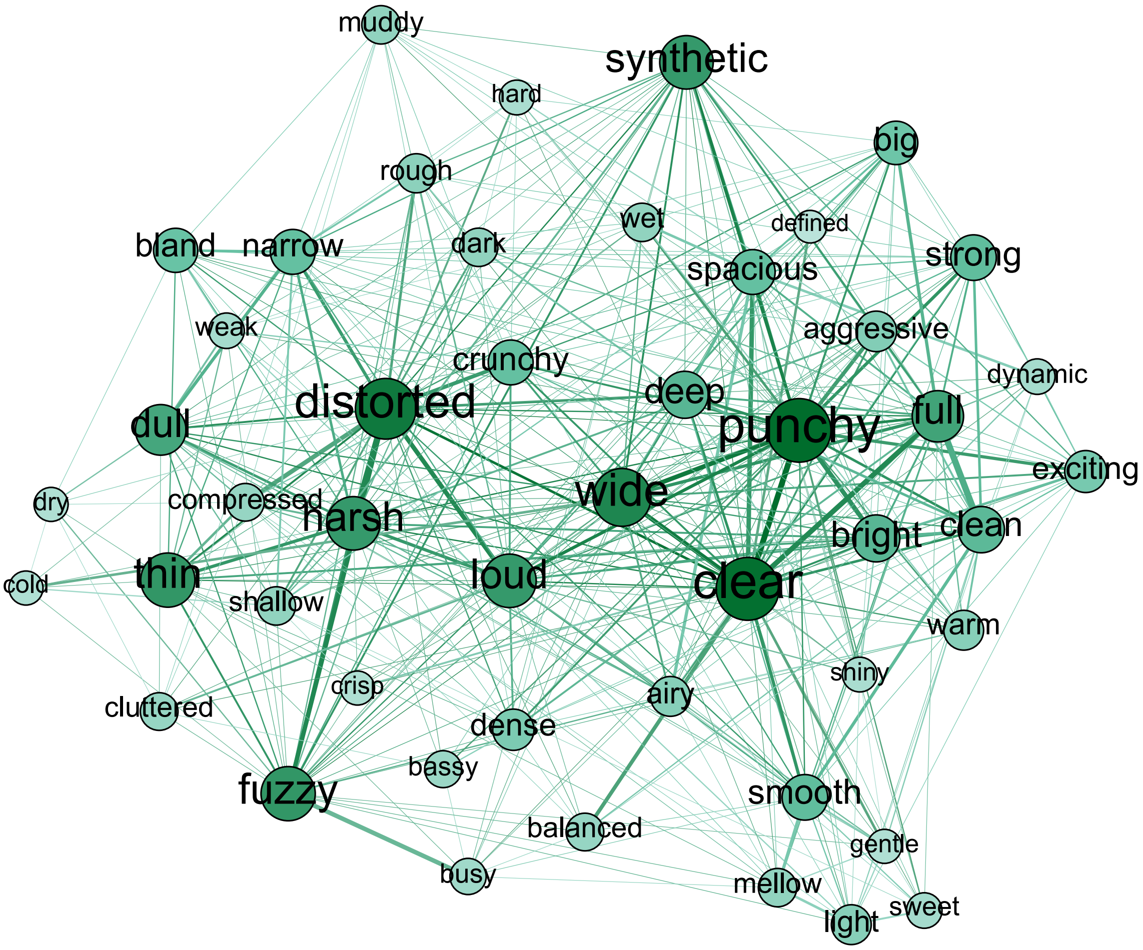


Figure 3.10: Term network, with nodes where degree >10. By using the ForceAtlas2 layout algorithm, terms which are frequently used together are located closer to one another than terms which are rarely used together.

3.4.2.1 Normalised quality-score

The normalised quality-score, $Z_{quality}$, of each word is given by the Eqn. 3.4, where N_Q is the number of times the word is used to describe a quality rating equal to Q and N_{total} is the total amount of times the word is used. All ratings are normalised to the range 1 to 5, the same range as the quality ratings.

$$Z_{\text{quality}} = \sum_{Q=1}^{5} \left(\frac{N_Q}{N_{total}} . Q \right) \tag{3.4}$$

3.4.2.2 Normalised expertise-score

Similarly, the normalised expertise-score, $Z_{\text{expertise}}$ of each word is given by Eqn. 3.5, where $S_i = 1$ for expert listeners and $S_i = -1$ for non-expert listeners. An expertise score of 1 indicates that a word has only been used by the expert group while a score of -1 indicates that a word has only been used by members of the non-expert group.

$$Z_{\text{expertise}} = \sum_{i=1}^{22} \left(\frac{N_{S_i}}{N_{total}} . S_i \right) \tag{3.5}$$

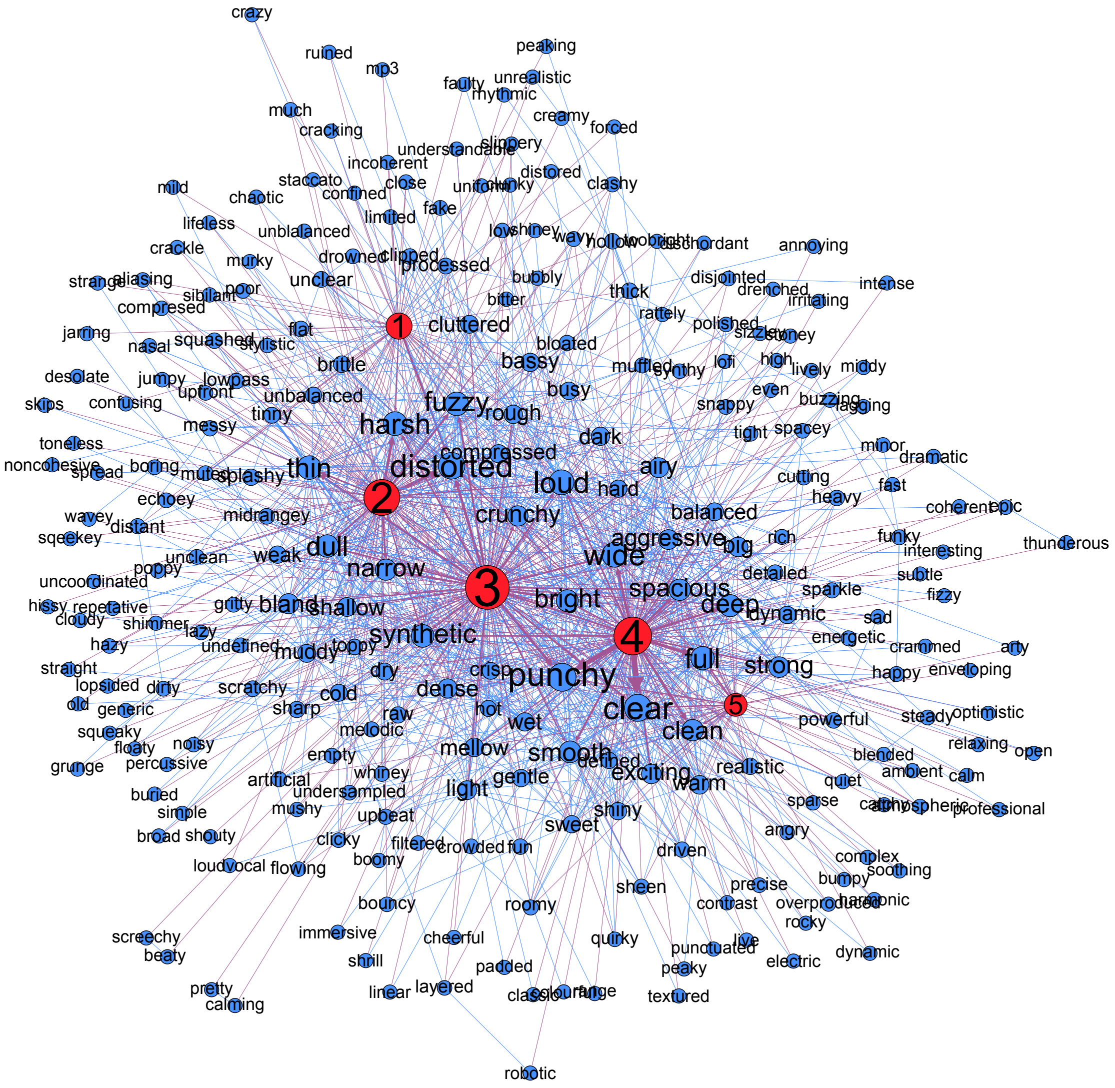


Figure 3.11: Term-Quality network. Terms used to describe specific quality ratings (in red) are shown close to those ratings.

3.4.2.3 PCA-score

This score investigates how certain words were used to describe certain songs, and determines a score based on the objective parameters of those audio signals. For all audio samples, a set of objective signal features was extracted which was then subject to PCA (see § 3.3.2). The first two dimensions explain 80.2% of the total variance in the extracted features. Dimension 1 can be described by amplitude-based features, with positive values referring to louder, more compressed samples and negative values referring to quieter, more dynamic samples. Dimension 2 describes signal bandwidth, where positive values have greater high-frequency extension.

For each term used, a score is obtained for each of these two dimensions, similar to the previous metrics. This allows all words to be positioned in the same feature-reduced space used for audio analysis, using the scores of all audio samples on each principal component. Here N_A is

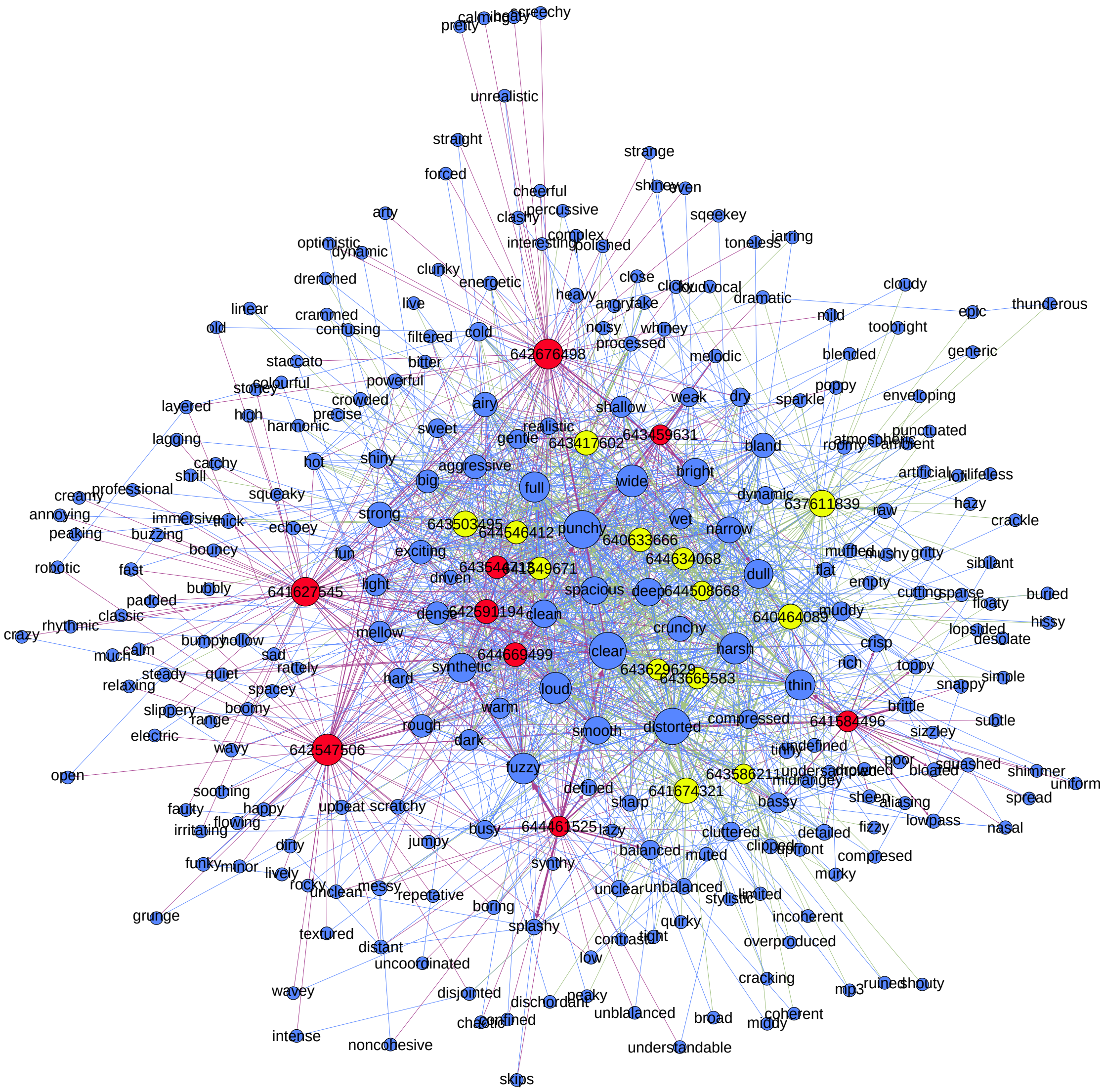


Figure 3.12: Term-Participant network, showing all words and participants. Experts are shown in yellow and non-experts in red. Frequently used terms are located towards the centre of the graph and infrequent terms are located at the exterior.

the number of times a word is used to describe sample A, and N_{total} is the total number of times the word is used. From the earlier PCA, $dim1_A$ and $dim2_A$ are the scores for sample A on each of the first two dimensions of the PCA space (see Fig. 3.8). For each word, the scores are determined as follows.

$$Z_{\text{dim1}} = \sum_{A=1}^{62} \left(\frac{N_A}{N_{total}} .dim1_A \right) \tag{3.6a}$$

$$Z_{\text{dim2}} = \sum_{A=1}^{62} \left(\frac{N_A}{N_{total}} .dim2_A \right) \tag{3.6b}$$

Table 3.6: Frequency count (Chi square test analysis) of 20 most used words

	Quality rating					
	1⋆	2⋆	3⋆	4⋆	5⋆	TOTAL
Distorted	**31**>	**43**>	37	**13**<	**2**<	126
Punchy	**1**<	**11**<	37	**63**>	13	125
Clear	**1**<	**4**<	**24**<	**77**>	**18**>	124
Full	0	**4**<	21	**41**>	**21**>	87
Harsh	**15**>	**38**>	23	**9**<	0	85
Wide	3	**5**<	28	**35**>	10	81
Loud	**10**>	18	25	22	4	79
Clean	**0**	0	**13**<	**36**>	**20**>	69
Fuzzy	7	**28**>	28	**4**<	0	67
Synthetic	**1**<	**18**>	21	20	4	64
Spacious	**1**<	0	20	**30**>	**10**>	61
Thin	6	**21**>	**29**>	**5**<	0	61
Bright	**1**<	9	**26**>	17	7	60
Dull	**8**>	**25**>	20	**7**<	0	60
Deep	**0**	**4**<	15	**29**>	9	57
Narrow	2	**25**>	23	**6**<	0	56
Smooth	**0**	**3**<	18	**27**>	7	55
Crunchy	**0**	10	**23**>	9	2	44
Strong	**0**	**2**<	10	**21**>	**9**>	42
Aggressive	2	5	8	**18**>	5	38
⋮	⋮	⋮	⋮	⋮	⋮	⋮
TOTAL	197	528	876	856	212	2669

3.4.3 Results

For all these metrics, words which are infrequently used would achieve scores heavily weighted by the few instances on which they were used. Therefore, the following discussion displays only a subset of the total set of words.

3.4.3.1 Quality scores

The 20 most frequently used words are shown in Table 3.6, along with the number of times used to describe each quality rating. A Chi-square test is used to determine whether there was significant variation in the usage of words across these five categories. The result of this test indicated that there were significant variations present, as different words were used to describe different quality ratings.

$$\chi^2(76, N = 1441) = 2131.26, p =< .001$$

Z_{quality} for each of the top 20 words is shown in Table 3.7. This data shows the importance of distortion in the perception of quality, as audio samples described as distorted are awarded low ratings of quality.

3.4.3.2 Expertise scores

$Z_{\text{expertise}}$ for all words was obtained. Words which were used by only a single participant were removed, leaving 96 words out of the initial 255. When used by only one participant a word has a score of either 1 or -1 and therefore would bias the interpretation of the following results. In

Table 3.7: Quality score of the 20 most frequently-occurring words.

Word	Quality score
Clean	4.10
Full	3.91
Strong	3.88
Clear	3.86
Spacious	3.79
Deep	3.75
Smooth	3.69
Punchy	3.61
Wide	3.54
Aggressive	3.50
Bright	3.33
Synthetic	3.13
Crunchy	3.07
Loud	2.90
Narrow	2.59
Thin	2.54
Dull	2.43
Fuzzy	2.43
Harsh	2.31
Distorted	2.30

order to keep the most agreed-upon words, all words used by only 2 or 3 participants were also removed, leaving 60 words which account for approximately 84% of all instances. The histogram in Fig. 3.13 shows the distribution of counts among the 60 remaining words. This distribution shows a skew towards higher scores, which suggests that the most agreed-upon terms are mostly used by the expert group, while the non-experts used more individual terms, with less agreement.

3.4.3.3 Feature-based scores

Figure 3.14 shows the 60 most agreed-upon terms positioned in the first two dimensions of the PCA space. The words ‘compressed’, ‘distorted’, ‘clipped’ and ‘loud’ have positive values on dimension 1 while ‘dynamic’ and ‘gentle’ have negative values. The words ‘bright’, ‘brittle’ and ‘harsh’ have positive values on dimension 2, which is related to high-frequency characteristics, while ‘dark’, ‘warm’ and ‘dull’ each have negative values. This shows agreement with the objective descriptions (see § 3.3.2).

The Euclidean distance between pairs of words in the full 22-dimensional PCA space is obtained (refer back to Fig. 3.6). These distances are used to perform multi-dimensional scaling, in which words are positioned to minimise the total strain in the graph. Positions of words in a two-dimensional MDS solution are shown in Figure 3.15.

3.4.4 Interpretation of scores

3.4.4.1 Quality ratings

The word ‘distorted’ is the most frequently occurring word and is used significantly more often than chance in describing audio samples that were rated 1 or 2 stars, while significantly less often than chance for 4 and 5 stars. This suggests that the participants very often judged the quality of

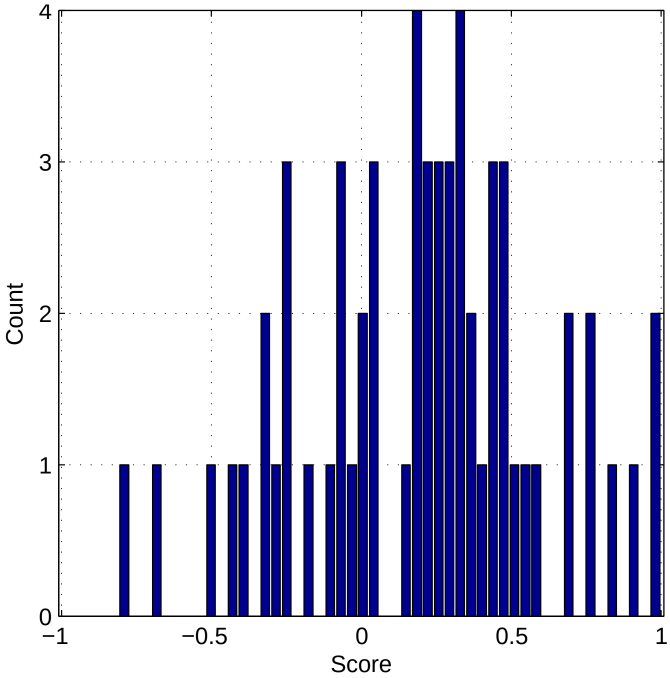


Figure 3.13: Histogram of normalised expertise score. When the words used by three or fewer participants are omitted, most remaining words have positive expertise scores, indicating they are favoured by experts.

audio samples based on the level of perceived distortion. Similarly, the word ‘clean’ is never used to describe ratings below 3 stars and achieves the highest quality score. The words ‘punchy’ and ‘clear’ are also frequently-occurring, suggesting that these words were familiar to participants and can be used to describe sound attributes of musical recordings which relate to audio quality. This result helps to justify recent research into the objective characterisation of these terms by Fenton et al. [145, 146].

3.4.4.2 Expertise

The distribution of the expertise scores suggests that the two expert groups used a different set of words to one another when assessing audio quality. The expert groups used a smaller, more agreed-upon set of words, while the non-experts used a larger variety of individual terms. This suggests that expert users have been trained to identify certain aspects of an audio signal and describe them in a way that is understandable to other expert users. The word-usage patterns of the non-experts shows that these participants were more likely to use words which were not used by other participants, terms for which the meaning may not be as universally understood. Only two words were used by all participants at least once — ‘distorted’ and ‘clear’ — further suggesting that this subjective dimension is generally important to listeners. Of the words used by over four participants, the five most associated with high expertise are ‘dynamic’, ‘muddy’, ‘cluttered’, ‘compressed’ and ‘tinny’. The five words most associated with low expertise are ‘busy’, ‘messy’, ‘mellow’, ‘brittle’ and ‘light’. While open to interpretation, the expert group appears to employ a subjectively more technical language, while, in contrast, the non-expert group refer to similar properties in a more abstract fashion. For example, one can consider ‘tinny’ to be equivalent to ‘brittle’ and ‘light’ as well as ‘busy’ to be comparable to ‘cluttered’ or ‘compressed’.

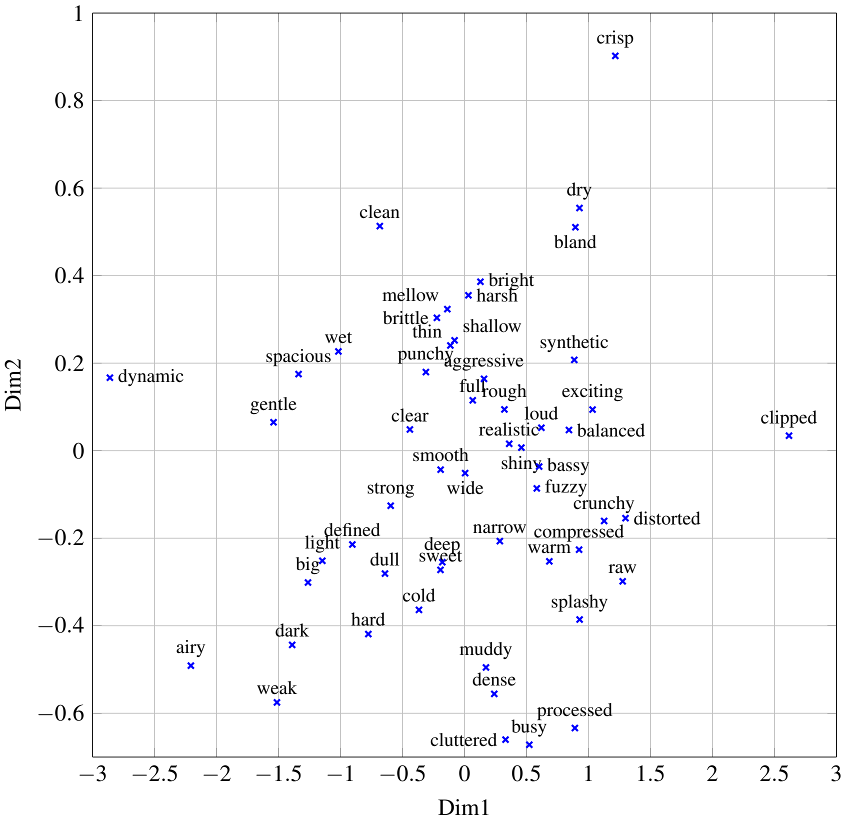


Figure 3.14: PCA scores of the top 60 words.

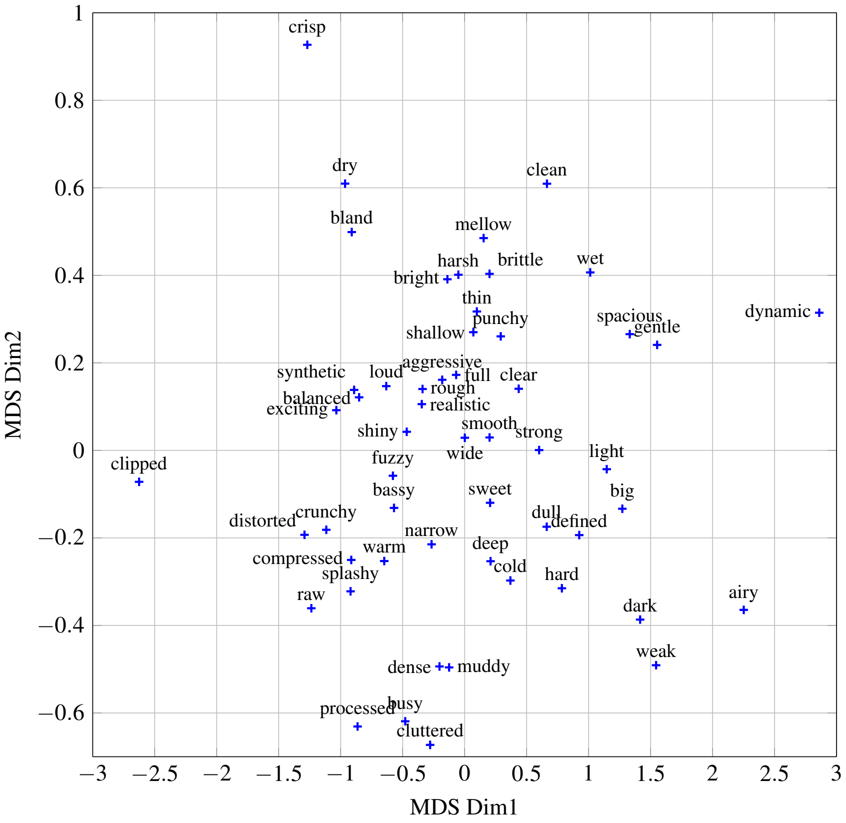


Figure 3.15: MDS of PCA scores of top 60 words.

3.4.4.3 Words in PCA space and MDS

The words were scored based on the principal components of the samples' signal features in order to gain insight into the meaning of each word. Figure 3.15 suggests that, when based on objective features, the differences and similarities between pairs of words can be seen, for example, 'cluttered' and 'busy' are similar, as are 'distorted', 'crunchy' and 'compressed', among other pairings. The words 'punchy', 'clear', 'full' and 'smooth', which all have high quality scores, are closely located in Fig. 3.15 which suggests that these words were used to describe songs which shared similar values of the objective features relating to high quality.

Of course, this does not mean that an absolute mapping between words and subjective variables exists, for example, that negative values of dim.2 are associated with high "like" scores (see Fig. 3.7). Recall that the correlation between like ratings and dim.2 is $R^2 = 0.129$ (see Table. 3.5) and so an absolute mapping between these words and the subjective variables would not be advisable.

3.5 Discussion

These results are now discussed in light of the initial research questions, RQ.1-5. Results indicated that the samples used in this test elicited different ratings and that, overall, the effect of sample was the largest contributor to the variance found in the subjective ratings, shown in Table 3.1, where $\eta_2^p = 0.227$. The effect size of the audio sample was large ($\eta^2 = 0.201$) for quality and medium ($\eta^2 = 0.127$) for like. This confirmed that the corpus of audio samples used was successful in triggering significant perceptual variation in ratings from the participants for both concepts.

There appears to be a stronger correlation between quality ratings and the objective features extracted from the signal than that found for like ratings (see Table 3.5). This suggests the former is a more reliable concept for the subjective evaluation of technical quality, related to modifications of the signal and distinct from hedonic perception. A meaningful correlation was found between like and quality ratings ($R^2 = 0.26$) using raw results pertaining to individual ratings of songs. This however, became non-significant when values were averaged over all participants ($R^2 = 0.02$), removing inter-subject variation. If the two concepts of like and quality are plotted in the space resulting from reducing signal features to a two dimensional space (see Fig. 3.7), they are nearly orthogonal, further supporting the idea that there is low correlation between them. Each concept is found to describe a different percept in the minds of listeners, where quality refers to technical aspects of the recording and production and like refers to hedonic perception that might be rooted in the musical style/genre or the actual song content itself. This is perhaps the most insightful finding in this study, that quality and like ratings can be considered as two percepts, explained by different factors. Participants elected their own definitions of quality in the experiment by justifying their ratings

3.5.1 Effects of expertise

While expert listeners, on average, provided slightly lower quality ratings than non-experts, the effect of expertise is observed to be small for both quality ($\eta^2 = 0.004$) and like ($\eta^2 = 0.002$). It appears that expertise is not a key factor in the appraisal of either technical quality or hedonic preference, under the conditions investigated here, although, after further investigation, it was observed that experts and non-experts typically used different words to justify their ratings (see Fig. 3.12).

3.5.2 Liking and familiarity

Participants were significantly more likely to award greater ratings of like and quality when they were more familiar with the music. However, it is clear that this effect is greater for like ratings, explaining 18.7% of the variance (see Fig. 3.5a), whereas for quality ratings it explains only 2.4% of the variance (see Fig. 3.5b). This relationship between familiarity and hedonic preference could be explained by two factors; one may like a song, subsequently choose to listen to it many times, becoming familiar with it, or one may hear a song many times, become familiar with it and grow to like it. This result suggests a clear differentiation between the concepts of preference (how much someone likes a song) and *technical* quality (how well a song has been produced), since familiarity does not seem to play a strong part in the latter.

3.5.3 Predictive power of signal features

Objective features extracted from the signal were reduced to two components: component 1 mainly describing aspects of amplitude and explaining 67% of the variance in the features considered, while component 2 describes aspects of the spectral content and explains 13% of the variance. Significant correlations were found between features and the subjective response variables (see Table 3.3 and 3.5).

Perceived quality is significantly correlated to amplitude features. Samples with higher dynamic range seem to elicit higher ratings of quality, while those with higher loudness seem to be associated with lower ratings. Recall that all samples have been presented at a normalized loudness level, thus effectively removing the differences in loudness but retaining the effect of reduced dynamic range that often ensues from production techniques which maximize loudness. This can explain why "louder" samples are perceived as lower quality in this context.

Measures of spectral flux and some of the underlying features in the MIRtoolbox used to develop emotional predictions are also found to be correlated to quality. Metrics for spectral content do not appear to have a significant effect on quality ratings.

Like ratings do not seem to be affected by amplitude features. As the presentation of audio to participants was normalized according to perceived loudness, as in modern on-line music streaming services such as Spotify and iTunes Radio, these results suggest instead that the effects of dynamic range compression arising from efforts to increase loudness do not appear to affect hedonic perception despite their degrading effects on perceived audio quality.

Like ratings appear to be correlated to spectral features although the strength of the correlation is about half of that observed between quality and component 1 (see Table 3.5). This low correlation suggests that ratings of like are more strongly affected by a listener's familiarity with a song than with objective features describing it.

These results further reinforce the idea that like and quality are separate aspects of an overall "preference" paradigm. When one simply asks participants for one of these concepts, like or quality, the result may be coloured by the participants impression of the other, which is not asked for, a phenomenon known as "dumping bias" [147].

3.5.4 Temporal variation in loudness/dynamics — the "loudness war"

The sample that scored the lowest mean rating for quality (see Fig. 3.3) was taken from an album whose perceived audio quality, due to production techniques, received negative attention in mainstream media at the time of release [148, 149]. Participants were possibly aware of this criticism and therefore open to bias. As shown in Fig. 5b, there is a difference in the mean value of dim. 1 for samples from each decade between the 1980s and 2000s. While the "loudness war" has been well-documented [108, 115, 150, 151] and has been observed by plotting individual amplitude-based variables over time, one can now see that the effect is visible on a factor level in a feature reduced space. The samples from the 1980s display more variation across dim. 2 than dim. 1, i.e., more variation in spectrum/timbre than loudness/compression. There is a greater range of loudness/compression in the 2000s since it is then possible to make louder but more compressed productions, while some content producers still choose to create dynamic productions. The greatest variation in loudness/compression in one decade is during the 1990s. This particularly significant period of the "loudness war" has been previously referred to by the term "loudness race" [108].

Table 3.8: Coefficients for fit shown in Eqn. 3.7.

Coefficient	Value	95% conf. interval
a0	1.181	(0.6643, 1.698)
a1	2.748	(-34.34, 39.84)
b1	-3.511	(-33.05, 26.03)
a2	-0.1323	(-2.962, 2.697)
b2	0.1495	(-3.04, 3.339)
a3	0.6051	(-13.28, 14.49)
b3	0.4349	(-18.49, 19.36)
w	0.09484	(0.0896, 0.1001)

A more detailed investigation was carried out in order to reveal more information about this trend. Rather than simply using the audio samples from the subjective test, an equivalent analysis was undertaken on the *entire* dataset of 404 samples. Figure 3.16 shows the value of the first PCA dimension for all audio samples [3] Again, as in Figs. 3.1a and 3.1b, the smoothed lines were determined using Matlab's `smooth` function. Rather than using numerical differentiation techniques, the smoothed line is represented by an analytical expression. This was obtained using the curve-fitting tool in Matlab. A Fourier series was used to fit a curve to the smoothed line. Three terms were used. This provided a near-perfect fit to the smoothed line, with $R^2 = 0.9999$. The general form of the equation is shown in Eqn. 3.7 and the coefficients are displayed in Table 3.8.

$$y = a_0 + a_1\cos(wx) + b_1\sin(wx) + a_2\cos(2wx) + b_2\cos(2wx) + a_3\cos(3wx) + b_3\sin(3wx) \quad (3.7)$$

The first and second derivatives of this expression are found using the symbolic math toolbox in Matlab. These functions are shown in Fig. 3.17. The minimum value of f' shows the point where f moves from concave to convex. The minimum and maximum values of f'' show the points of inflection. These points are used to describe the start and end of a specific period of time, when the change in loudness was most rapid. From the data, the period can be dated as 1989 to 2002. This has been indirectly referred to by Deruty [115], who stated that loudness levels reached a peak in 2007 and described "pre-1990" levels as being a target to return to in the future. By Fig. 3.17, there is no signs of change since 2007. This result may be more accurate as it used a factor-based approach rather than analysis of features in isolation.

[3] As with the subset used in the subjective test, the first component of this PCA explains loudness variables, in a highly similar fashion to what is shown in Fig. 3.7.

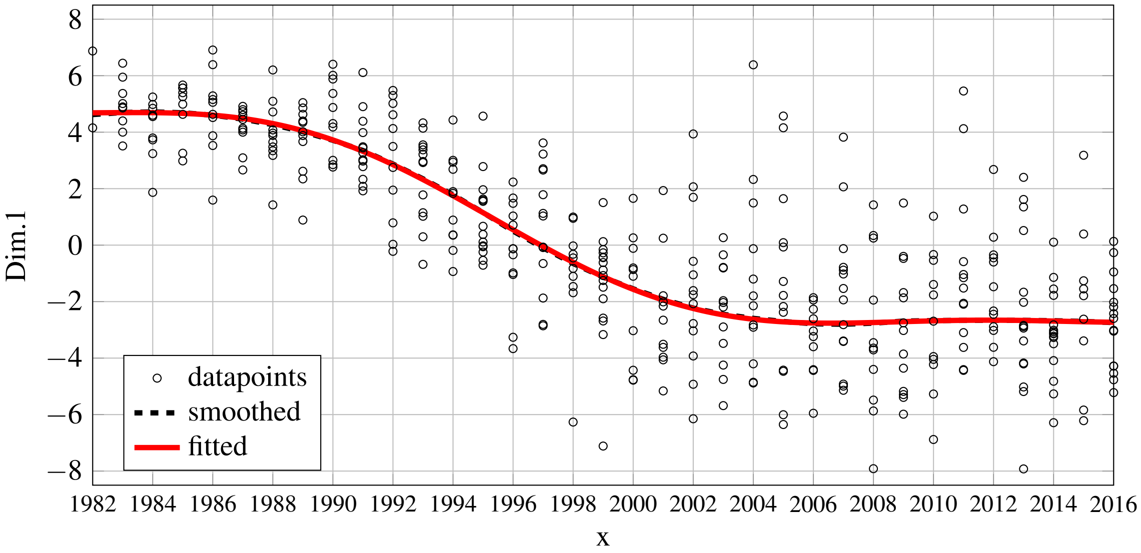


Figure 3.16: Fit of PC1 to 404 songs. Lower values indicate reduced dynamic range / increased loudness.

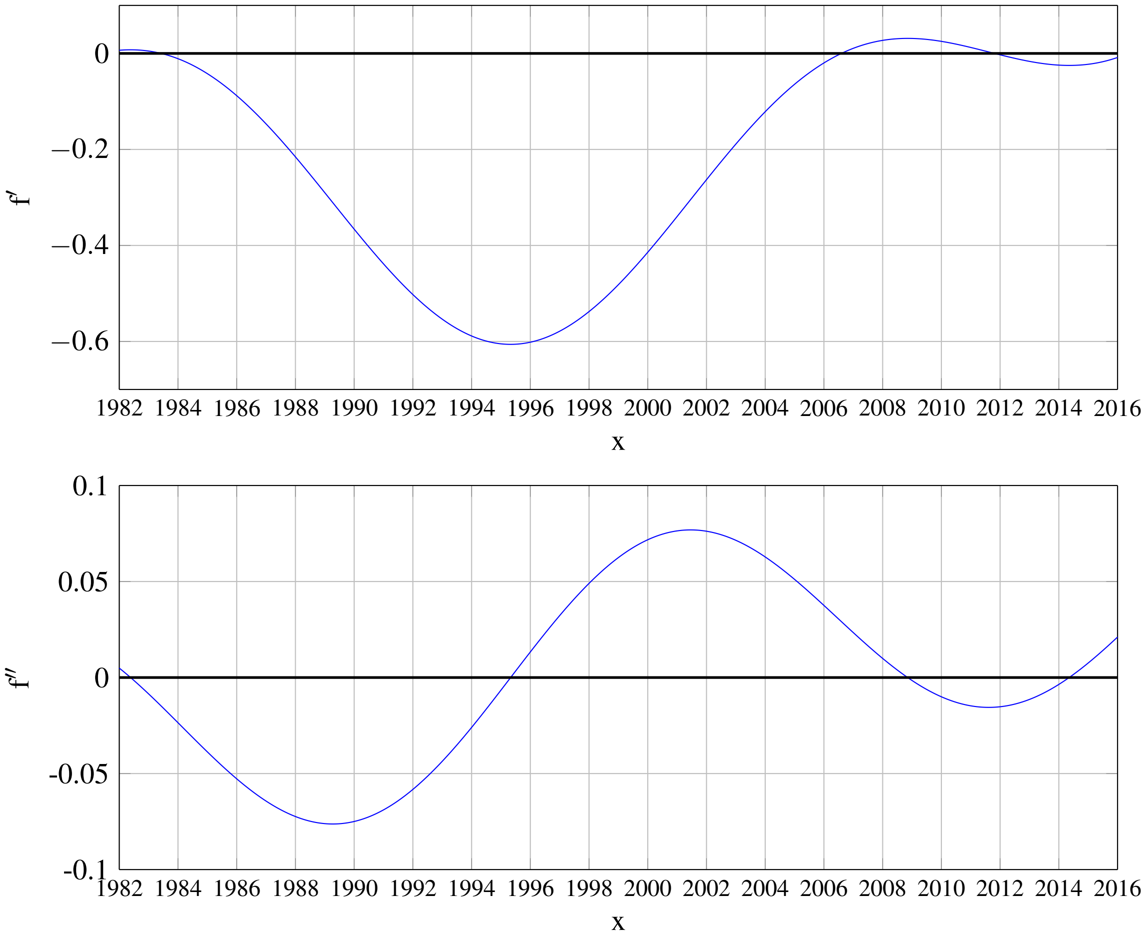


Figure 3.17: Derivatives of fit of PC1 to 404 songs. This result indicates that 1995 was the year in which loudness values were increasing at the greatest rate. It is also suggest that loudness values have not undergone any notable changes since 2002.

3.6 Chapter summary

The analysis of the experiment described in this chapter revealed that the perception of quality in mastered, commercially-released music samples is related to the perception of dynamic range and amplitude. The perception of the more hedonic qualities, which relate to liking of a music sample, do not relate to these measures in a significant way. These 'like' ratings were, however, strongly influenced by song-familiarity, implying instead that aspects of preference and liking are distinct from the interpretation of quality and might not be the best descriptors for studies where technical quality is the percept being sought.

The expertise of listeners, although significant, had a weak effect on the ratings of quality and like, suggesting, somewhat counter-intuitively, that a participant's expertise is not a strong factor in assessing audio quality or musical preference (see Figs. 3.5a and 3.5b).

It has been observed that the words used to describe sonic attributes of the audio signal on which quality was assessed were typically those words that describe perceived timbre, space, and defects. The frequency of word usage varied significantly depending on the rating being awarded, with words such as 'clean' and 'full' strongly associated with high ratings of quality, while 'distorted' and 'harsh' were associated with low ratings.

In summary, quality in music production is revealed as a perceptual construct distinct from hedonic, musical preference, which is more likely influenced by familiarity with the song. Audio quality can be predicted from objective features in the signal and be adequately and consensually described using verbal attributes. The work presented has implications in the music industry, particularly if issues such as the "loudness war" are being rendered moot by new loudness normalised broadcast standards. However, as this study dealt only with one particular dataset, containing multiple songs and only one mix per song, this separation of the two concepts may not be the case in other scenarios, such as music mixing (see § 6.2.4).

This chapter has introduced a number of concepts that will be re-visited later, such as audio signal feature extraction and a detailed procedure for principal component analysis. From this point onwards, the content of the thesis deals with music-mixing processes in a more explicit manner.

4
Exploring the “mix space”

In the process of mix-engineering, many complex actions are undertaken, such as level-balancing, equalisation, dynamic range compression and expansion, or the use of time-based effects such as delay and reverberation. Each of these types of processing can utilise a number of different parameters and can be applied in any particular sequence on any individual audio element in the project. Consequently, there exist a large number of possible mixes that can be produced from a given set of audio elements and tools, and the problem quickly becomes intractable.

This chapter deals with the fundamental question of “what is mixing”, or, more explicitly, “what can be achieved by mixing”? The mixing of music can be considered as an optimisation problem, as described by Terrell et al. [152], albeit one with a large amount of variables and a target which is not well defined. Two approaches can be taken.

1. maximise ‘quality’, ‘preference’ or some other concept or percept.
2. minimise technical issues/faults, the absence of which is believed to benefit the production.

This latter approach was discussed in Chapter 2 as a means of automating music production tasks. The former may be preferable but requires an in-depth understanding of what constitutes ‘quality’. Rather than to rely on “best practice” rules determined from interviews and other qualitative methods, the work presented in this Chapter used quantitative methods in the observation of music mixing practice. Beginning with a trivial example of the mixing of two audio elements and moving on to the study of more realistic mixing scenarios, this chapter presents representations of simple mixing practices and the analyses of data gathered by experiment. Thus far, one publication has been published based on portions of the work described in § 4.1, § 4.2 and § 4.3 [153].

4.1 Basic theory

Consider the trivial case where two audio signals are to be mixed, where only the absolute levels of each signal can be adjusted. This can be considered the most simple mixing exercise (it is shown later that, for tasks referred to as 'mixing', the number of signals must be more than one). In Figure 4.1, the gains of the two signals are represented by x and y. Assume both are positive-bound. Consider the point p as a configuration of the signal gains i.e. (p_x, p_y). From this point, the values of x and y are both increased in equal proportion, arriving at the point p'. The magnitude of p is less than that of p' ($\|p\| < \|p'\|$) yet since the ratio of x to y is identical, the angles subtended by the vectors from the y-axis are equal ($\angle p = \angle p'$). In the context of a mix of two tracks, what this means is that the amplitude of p' is greater than p, yet the blend of signals is the same.

At this point, consider what is meant by a 'mix'. Recall from § 2.3.1 that often-used definitions consider a mix as a sum of input channels. This definition is too broad, as numerous "mixes" are copies of one another but at different loudness levels. In the gain-space, if all the points on a line from the origin at a fixed angle are the same blend of tracks then they are perceptually very similar, just louder or quieter. Quite likely this would create a ridge or valley in the fitness landscape. Ridges and valleys are challenging obstacles for hill-climbing algorithms although gradient-descent can perform better. Since gradient descent requires the function be differentiable, it may not be the best approach for perceptually motivated fitness evaluations.

Alternately, a 'mix' can be thought of as the specific blend/balance/ratio of audio signals. From this definition, the point p and p' are the same mix, only p' is being presented at a greater volume. If the listener has control over the master volume of the system, then any difference between p and p' becomes ambiguous. From p, the level of fader y can be increased by Δ_y such as to arrive at the point r. In this particular case, the value of Δ_y was specifically chosen such that $\|r\| = \|p'\|$. However, for any $|\Delta_y| > 0$, $\angle r \neq \angle p'$. Therefore, the vector r clearly represents a different mix to either p or p'. Consequently, the definition of a 'mix' is clarified by what it is not: when two audio streams contain the same blend of input tracks but the result is at different overall loudness levels, these two outputs can be considered the same mix.

Definition 5. *mix: an audio stream constructed by the superposition of others in accordance with a specific blend/balance/ratio*

For this mixing example, where there are $n = 2$ signals, represented by n gain values, the mix is dependant on $n - 1$ variables, in this case, the angle to the vector. The norm of the vector is simply proportional to the overall loudness of the mix. Figure 4.2 shows a similar structure, with $n = 3$. Here, the point p' is also an extension of p. As in Figure 4.1, r is located by increasing the value of y from the point p and $\|r\| = \|p'\|$. Here, the values of each angle are explicitly determined and displayed. All three vectors share the equatorial angle of 60°. The polar angle of p and p' is 50°, while the polar angle of r is less than this, at $\approx 37°$. As in the two-dimensional case, it is the angles which determine the parameters of the mix and the norm of the vector is related to the overall loudness. While Figures 4.1 and 4.2 show a space of track gains there is redundancy of mixes in this space. What is ultimately desired is a space of mixes.

Definition 6. *Mix-space: a parameter space containing all the possible audio mixes that can be achieved using a defined set of processes.*

It becomes apparent that a euclidean space with track gains as basis vectors is not an efficient way to represent a space of *mixes*, according to Definition 6. If, in Figure 4.2, a set of m points randomly selected on $\mathbb{R}^3$ was chosen, the number of mixes could be less than m, as the same mix could be chosen multiple times at different overall volumes. A set of m randomly selected points on a $1/8^{\text{th}}$ sphere of any radius ($\mathbb{S}^2$) would result in a number of mixes equal to m. This surface is represented in Figure 4.3, which shows the portion of a unit-sphere in positively-unbounded $\mathbb{R}^3$, upon which exist all possible mixes of three tracks. This surface is a mix-space for the problem of three-track mixing, where the only available tool is gain adjustment. Figure 4.4 represents two mixes in this space using a ternary plot.

While both the 2-content of $\mathbb{S}^2$ (the 'surface area') and the 3-content of the enclosing $\mathbb{R}^3$, (the 'volume') both, strictly, contain an infinite amount of points, the reduced dimensionality of $\mathbb{S}^2$ makes it a more attractive content[1] to use in optimisation, as $\mathbb{S}^2$ is a subset of $\mathbb{R}^3$. As a consequence, the 'mix-space' is a more compact representation of audio mixes than the 'gain-space'. Such an optimisation is discussed in Chapter 8.

While the examples so far have used polar and spherical coordinates, for $n = 2$ and $n = 3$ respectively, to extend the concept to any n dimensions, generalised hyperspherical coordinates are used. The conversion from cartesian to hyperspherical coordinates is given below in Equations 4.1. The inverse operation, from hyperspherical to cartesian is provided in Equations 4.2 [154]. Here, g_j is the gain of the j^{th} track out of a total of n tracks. The angles are represented by ϕ_i. By convention, ϕ_{n-1} is the equatorial angle, over the range $[0, 2\pi)$ radians, while all other angles range over $[0, \pi]$ radians.

$$r = \sqrt{{g_n}^2 + {g_{n-1}}^2 + \cdots + {g_2}^2 + {g_1}^2} \tag{4.1a}$$

$$\phi_i = \arccos \frac{g_i}{\sqrt{{g_n}^2 + {g_{n-1}}^2 + \cdots + {g_i}^2}} \text{ , where } i = [1, 2, \ldots, n-3], i \in \mathbb{Z} \tag{4.1b}$$

$$\vdots$$

$$\phi_{n-2} = \arccos \frac{g_{n-2}}{\sqrt{g_n^2 + {g_{n-1}}^2 + {g_{n-2}}^2}} \tag{4.1c}$$

$$\phi_{n-1} = \begin{cases} \arccos \frac{g_{n-1}}{\sqrt{g_n^2 + {g_{n-1}}^2}} & g_n \geq 0 \\ 2\pi - \arccos \frac{g_{n-1}}{\sqrt{g_n^2 + {g_{n-1}}^2}} & g_n < 0 \end{cases} \tag{4.1d}$$

$$g_1 = r \cos \phi_1 \tag{4.2a}$$

$$g_j = r \cos \phi_j \prod_{i=1}^{j-1} \sin \phi_i \text{ , where } j = [2, 3, \ldots n-2], j \in \mathbb{Z} \tag{4.2b}$$

$$g_n = r \prod_{i=1}^{n-1} \sin \phi_i \tag{4.2c}$$

Figure 4.5 represents a comparable 4-track mixing exercise. The four audio sources are specifically

[1] In this context, content can be considered as "hypervolume". See Weisstein, Eric W. "Content." From MathWorld– A Wolfram Web Resource. `http://mathworld.wolfram.com/Content.html`

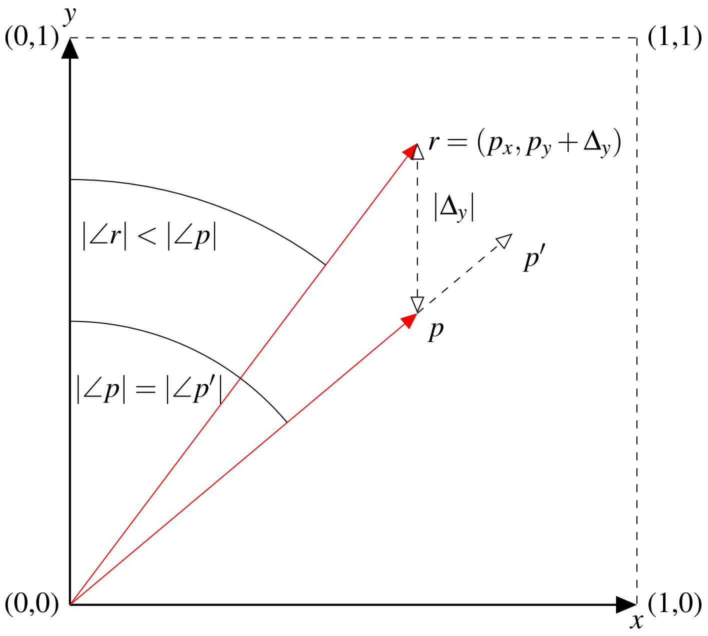

y
(0,1)
(1,1)
$r = (p_x, p_y + \Delta_y)$
$|\Delta_y|$
$|\angle r| < |\angle p|$
p'
p
$|\angle p| = |\angle p'|$
(0,0)
(1,0)
x

Figure 4.1: Points p, p' and r, in 2-track gain space. Note that the audio output at points p and p' are the same 'mix'.

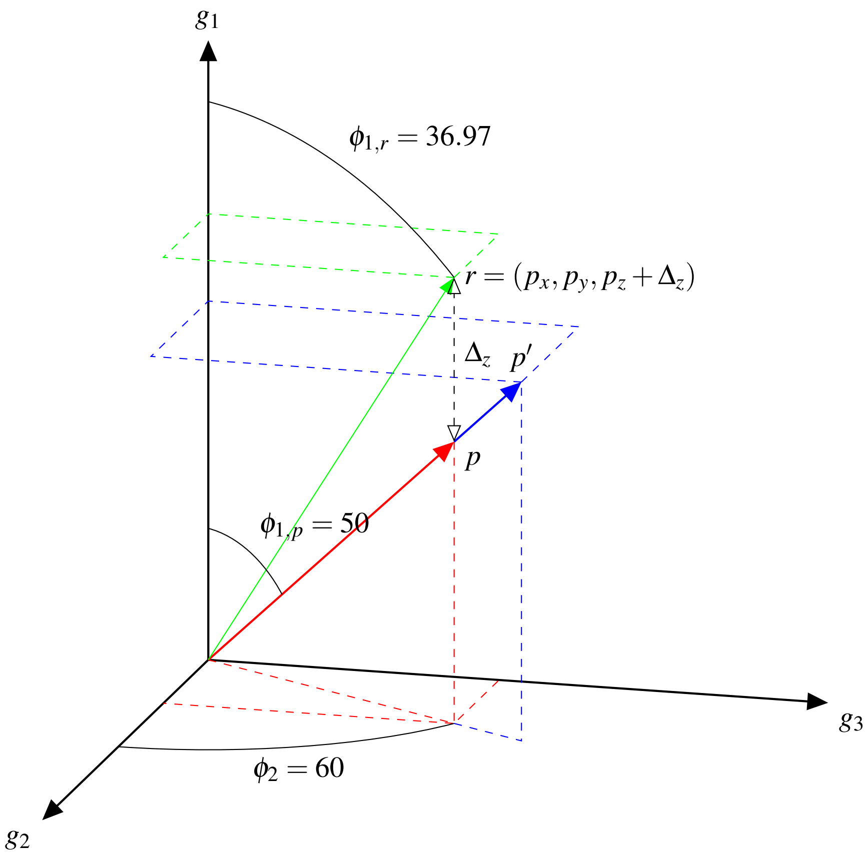


Figure 4.2: Mix at a point in 3-track gain space

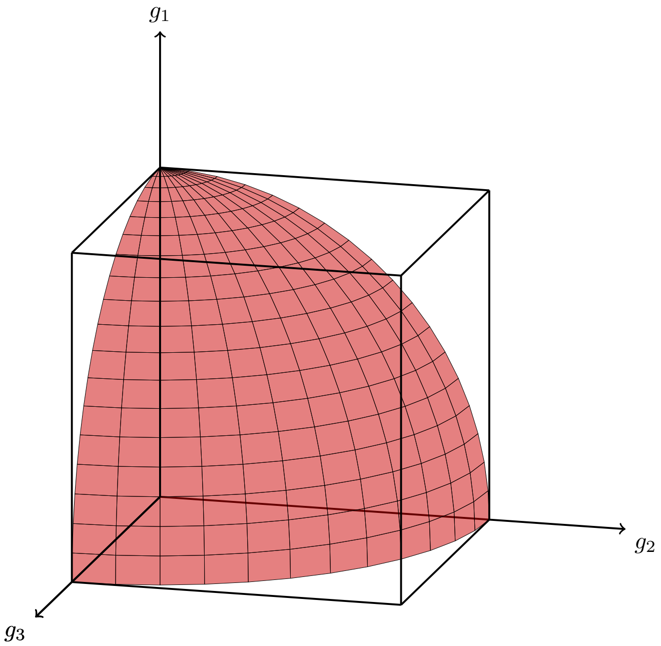


Figure 4.3: Surface containing all unique mixes of a 3-track mixture

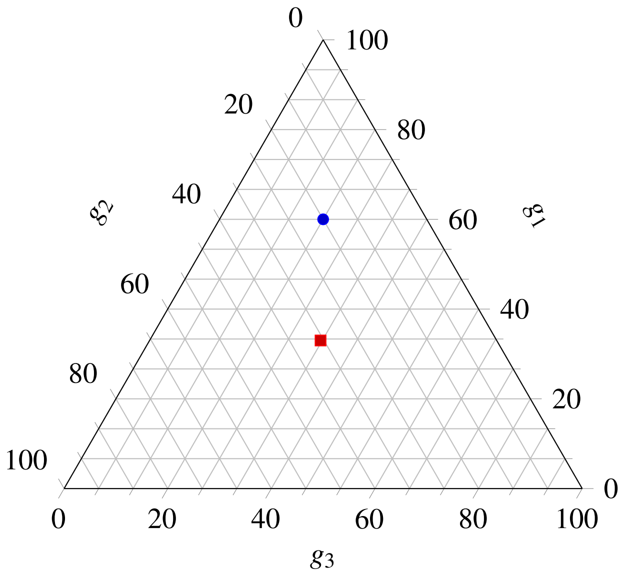


Figure 4.4: Ternary plot, where each point is a sum of three properties such that the sum is 100%. The square indicated the point where the mix is an equal blend of the three tracks. The circle has a higher level of vocals.

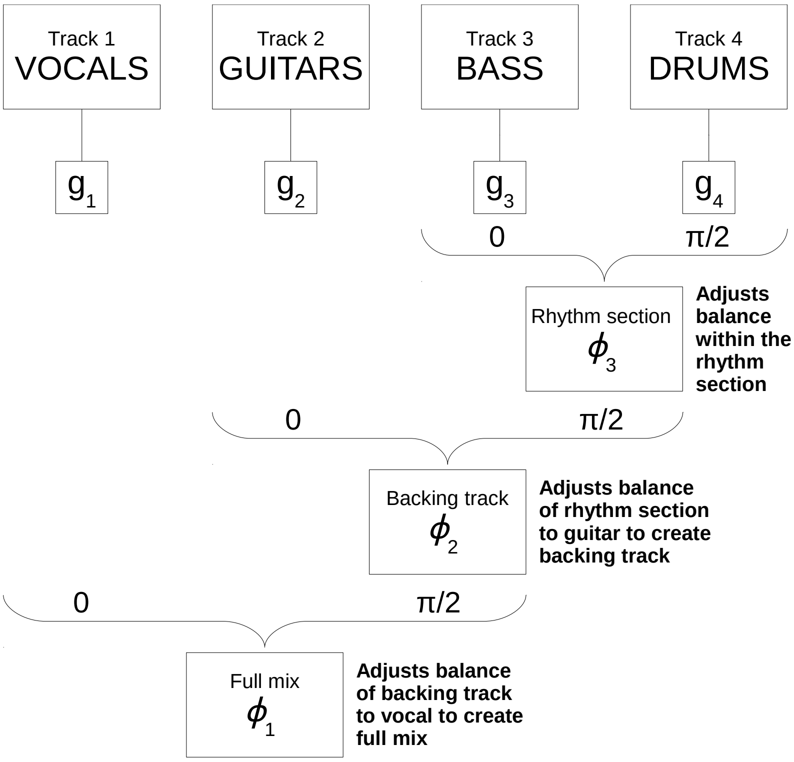
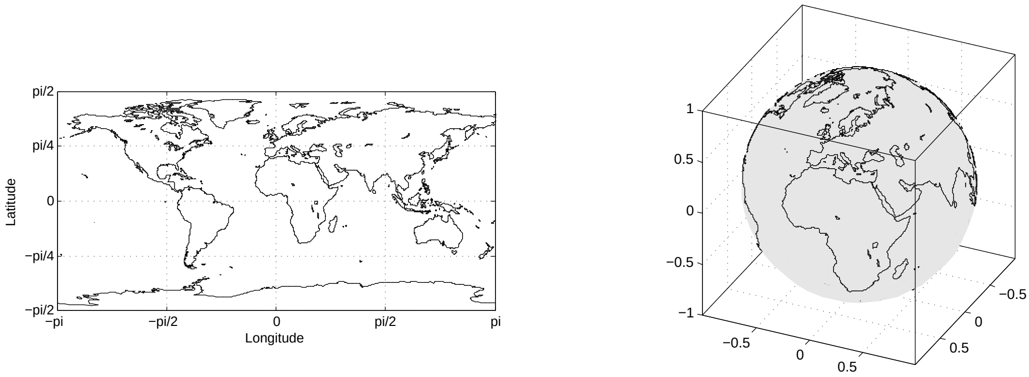

Track 1
VOCALS
Track 2
GUITARS
Track 3
BASS
Track 4
DRUMS
g1
g2
g3
g4
0
π/2
Rhythm section
ϕ3
Adjusts balance within the rhythm section
0
π/2
Backing track
ϕ2
Adjusts balance of rhythm section to guitar to create backing track
0
π/2
Full mix
ϕ1
Adjusts balance of backing track to vocal to create full mix

Figure 4.5: Schematic representation of a four-track mixing task, with track gains g_1, g_2, g_3, g_4, and the semantic description of the three ϕ terms, when adjusted from 0 to $\pi/2$.

chosen for this example, vocals, guitar, bass and drums, and assigned to g_1, g_2, g_3, g_4 respectively. Consequently, the set of mixes is represented by a 3-sphere of radius r. The coordinates ϕ_1, ϕ_2 and ϕ_3 represent a set of inter-channel balances which have musical importance. The value of ϕ_3 determines the balance of bass to drums, the rhythm section in this case. ϕ_2 describes the projection of this balance onto the g_2 axis, i.e. the blend of guitar to rhythm section. Finally, ϕ_1 describes the balance of the vocal to this backing track.

From herein, the parameter space comprising the $n-1$ angular components of the hyperspherical coordinates of a $(n-1)$-sphere in a n-dimensional gain-space, is referred to as a $(n-1)$-dimensional mix-space. More simply, this can be stated by saying the mix-space is the surface of a hypersphere in gain-space. In the case of music mixing, only the positive values of g are of interest. Subsequently, the interesting region of the mix-space is only a small proportion of the total hypersurface. This fraction is $1/2^n$. For this 4-track case in Fig. 4.5 the mix-space is 3-dimensional. However, as it represents a 3-sphere, it is not a Euclidean space. Consider the case of a world map. This map is a common 2-dimensional representation of the surface of the globe, a 2-sphere. A map displaying longitude and latitude coordinates will stretch the North and South poles from a single point on the 2-sphere to a line on the map. A variety of map projections have been proposed in order to represent the surface of the Earth as a flat map however each introduces some degree of distortion.

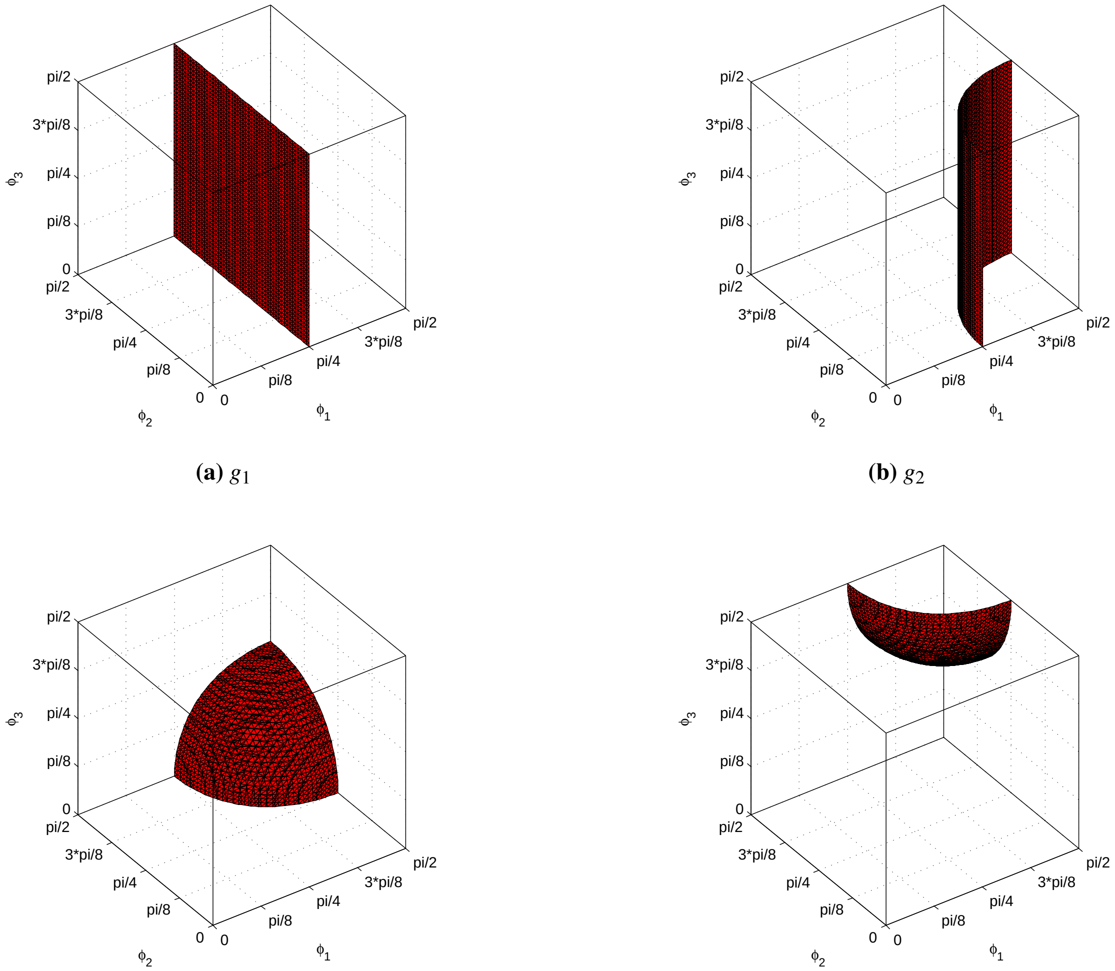
Figure 4.7 indicates the limitations of a Euclidean representation for the mix-space for the example in Figure 4.5. The ‘north’ pole of the 3-sphere is where $\phi_1 = 0$, the ϕ_2-ϕ_3 plane. In each subplot, the surface shown represents the mixes where a specific track is set to -3 dB. Figure 4.7a shows that half of the map has $\phi_1 < \pi/4$ and therefore $g_1 > 1/\sqrt{2}$. This surface area, and the enclosed volume decreases as j increases, as shown in Figures 4.7b, 4.7c and 4.7d. It is clear that a randomly selected point in *this* $\mathbb{R}^3$ would most likely contain loud vocals compared to drums and bass. This limitation is re-visited in Chapter 8, and is further discussed. For the purposes of visualising a 4-track mixing process, this representation can be useful. While a sphere is a non-euclidean space, *locally*, euclidean geometry is a good approximation.



(a) Coastline data placed on a unit 2-sphere

(b) Coastline data mapped by latitude and longitude, in radians

Figure 4.6: Illustration of spatial distortions introduced during mapping



(a) g_1

(b) g_2

(c) g_3

(d) g_4

Figure 4.7: Surfaces representing a gain of -3 dB for each of the g terms in the four-track mixing problem also shown in Figure 4.5

.

4.2 Mix-space concepts

As each point in this space represents a unique mix, the process of mixing can be represented as a path through the space. For example, consider a random walk in the mix-space. This path can be used to determine a random time-varying gain for each track. It is hypothesised that real mix engineers do not carry out a random walk but a guided and informed walk, from some starting point (“source”) to their ideal final mix (“sink”).

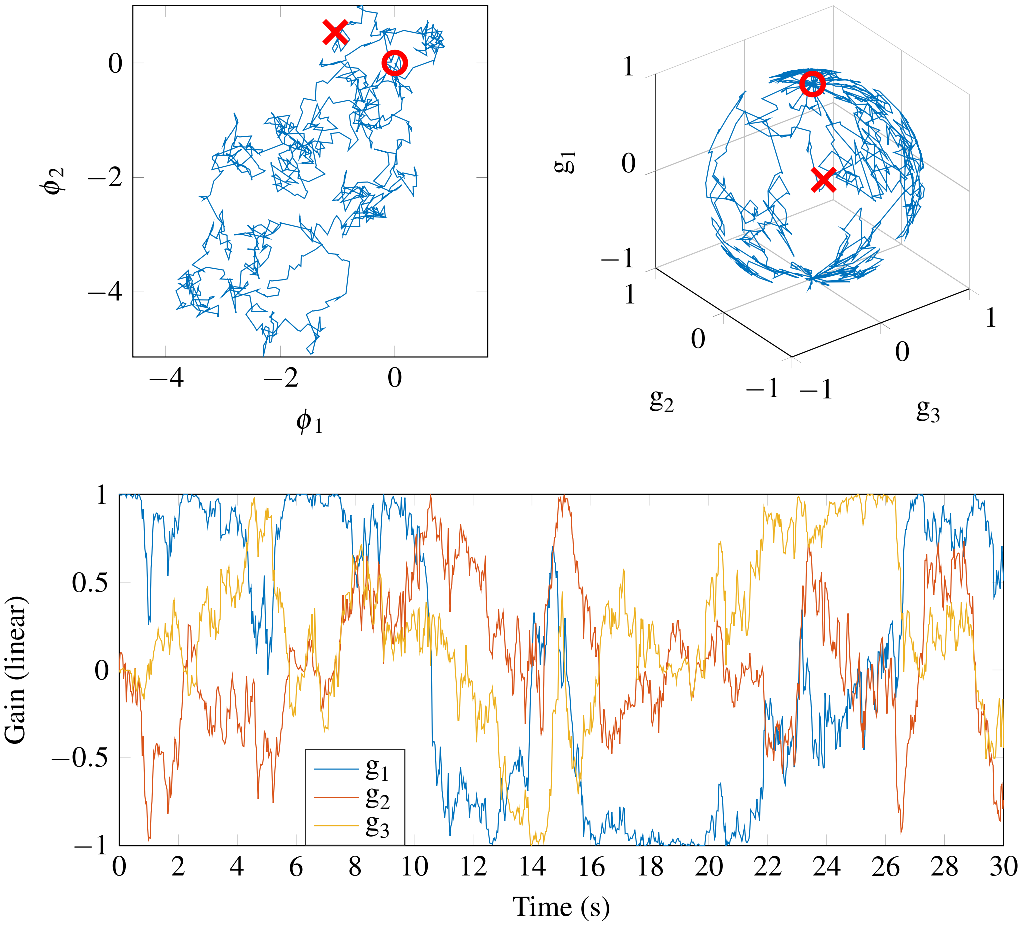


Figure 4.8: Random walk in mix-space

In Fig. 4.8 a random walk begins at the point marked ‘○’ in the 2D mix-space (the origin [0,0], which corresponds to a gain vector of [1,0,0]). The model for the walk is a simple Brownian motion [2]. After 30 seconds the final point reached is marked ‘×’. The gain values for each of the three tracks are shown and it is clear that the random walk is on a 2-sphere. The time-series of gain values is also shown. Note that $g \in [-1, 1]$, so for positive g the region explored is as Fig. 4.3.

4.2.1 The ‘source’

In a real-world context, on receiving a multitrack session and first loading the files into a DAW, each engineer will initially hear the same mix, a linear sum of the raw tracks [3]. This has been

[2] http://people.sc.fsu.edu/~jburkardt/m_src/brownian_motion_simulation/brownian_motion_simulation.html

[3] Here it is significant that a DAW typically defaults to faders at 0 dB, while a separate mixing console may default to all faders at $-\infty$ dB. This allows an experimenter to ensure that all mixers begin by hearing the same ‘mix’.

referred to in previous studies as a 'sum' or 'unmixed sum' [60, 81, 155]. While the term 'unmixed' can be misleading, it does reflect the fact that the artistic process of mixing has not yet begun. While each of these raw tracks can be presented in various ways, if we presume each track is recorded with high signal-to-noise ratio (as would have been more important when using analogue equipment) then, with all faders set to 0 dB, the perceived loudness of those tracks with reduced dynamic range (such as synthesisers, electric bass and distorted electric guitars) would be higher than that of more dynamic instruments (such as percussion or vocals). Much like the final mixes, this initial 'mix' can be represented as a point in some high-dimensional, or feature-reduced, space. It is rather unlikely that a engineer would open the session, hear this mix and consider it ideal, therefore, changes will most likely be made in order to move away from this location in the space. For this reason, this position in the mix-space is referred to as a '*source*'.

Definition 7. *Source: A point in the mix-space representing the initial configuration of tracks which is deemed not to be ideal by a significant proportion of mix engineers.*

In practice, the session, as it has been received by the mix engineer, may be an "unmixed sum" or may be a rough mix, as assembled by the producer or recording engineer. In a real-world scenario, the work may be received as a DAW session, where tracks have been roughly mixed. Alternatively, where multitrack content is made available online, such as in mix competitions [4], the unprocessed audio tracks are usually provided without a DAW session file. The latter approach is assumed in this study, in order for mix engineers to have full creative control over the mixing process. If mixers were to make unique changes to the initial configuration then that source can be considered to be radiating omni-directionally in the mix-space. However, it is possible that, for a given session, there may be some changes which will seem apparent to most mixers, for example, a single instrument which is louder than all others requiring attenuation. For such sessions, the source may be directional, or if a number of likely outcomes exist, there may exist a numerous paths from the source.

4.2.2 Paths in the mix-space

The path from the *source* to the final mix could be represented as a series of vectors in the *mix-space*, henceforth named '*mix-velocity*', and defined in Eqn. 4.3, for the three dimensions shown in Fig. 4.5. In this case the values of Φ are sampled at regular intervals.

$$u_t = \phi_{(1,t)} - \phi_{(1,t-1)} \tag{4.3a}$$

$$v_t = \phi_{(2,t)} - \phi_{(2,t-1)} \tag{4.3b}$$

$$w_t = \phi_{(3,t)} - \phi_{(3,t-1)} \tag{4.3c}$$

If all mixers begin at the same *source* then a number of questions can be raised in relation to movement through the *mix-space*, which help understand the nature of multi-track mixing.

- Moving away from the *source*, at what point do mix engineers diverge, if at all?

[4] `http://www.cambridge-mt.com/YoungGriffoCompetition.htm`

- How do mix engineers arrive at their final mixes? What paths through the *mix-space* do they take?
- Do mix engineers eventually converge towards an ideal mix?

4.2.3 The 'sink'

Complementary to the concept of a *source* in the *mix-space*, a '*sink*' would represent a configuration of the input tracks which produces a high-quality mix that is apparent to a sizeable portion of mix engineers and to which they would mix towards. This is similar to the goal displayed in Fig. 2.11.

Definition 8. *Sink: A point in the mix-space representing an ideal final configuration of tracks, as perceived by a significant proportion of mix engineers*

As the concept of quality in mixes is still relatively unknown there are a number of open questions in the field which can be addressed using this framework of sources, paths and sinks in the mix-space.

- Is there a single sink, i.e. one ideal mix for each multitrack session? In this case the highest mix-quality would be achieved at this point and this would be agreed upon by all mix engineers.
- Are there multiple sinks, i.e. given enough available mixes, are these mixes clustered such that one can observe a number of possible alternate mixes of a given multitrack session? These multiple sinks would represent mixes that are all of high mix-quality but audibly different, for example, the same song could be mixed in a number of different styles.
- Are there no sinks, i.e. does each mix engineer produce a unique mix, such that there is no discernible clustering of final mixes in the mix-space.

4.3 Mix-space experiment 1 — Mono

In order to examine how mix engineers navigate the mix-space a simple experiment was conducted. In this instance the mixing exercise was to subjectively balance the level of four tracks, using only a volume fader for each track, such that the participant achieves their own ideal mix. Importantly, the participants all began with a predetermined balance, in order to examine the influence of the source. This experiment aims to answer the following research questions:

RQ-6. Can the source be considered omni-directional or are there distinct paths away from the source?

RQ-7. Is there an ideal balance (single sink)?

RQ-8. Are there a number of optimal balances (multiple sinks)?

RQ-9. What are the ideal level balances between instruments?

4.3.1 Set-up

The multitrack audio sessions used in this experiment have been made available under a creative commons license[5] [6]. These files are also indexed in a number of databases of multitrack audio content [156, 157]. Three songs were used for this experiment, which consisted of vocals, guitar, bass and drums, as per Fig. 4.5, and as such the interpretations of Φ from here on are those in Fig. 4.5.

The following is a description of the audio stimuli used. The four tracks used from "*Borrowed Heart*"[7] are raw tracks, where no additional processing has been performed apart from that which was applied when the tracks were recorded [8]. The tracks from "*Sister Cities*"[9] also represent the four main instruments but were additionally processed using equalisation and dynamic range compression [10]. These can be referred to as 'stems', as the 11 drum tracks have been mixed down, the two bass tracks (a DI signal and amplifier signal) have been mixed together, the guitar track is a blend of a close and distant microphone signals and the vocal has undergone parallel compression, equalisation and subtle amounts of modulation and delay. In the case of "*Heartbeats*"[11], the tracks used are complete 'mix stems', in that the song was mixed[12] and bounced down to four tracks consisting of 'all vocals', 'all music' (guitars and synthesisers), 'all bass' and 'all drums'. For testing, all audio was further prepared as follows:

- 30-second sections were chosen, so that participants would be able to create a static mix, where the desired final gains for each track are not time-varying.

- Within each song, each 30-second track was normalised according to loudness. In this case, loudness is defined by BS.1770-3, with modifications to increase the measurements

[5] http://weathervanemusic.org/shakingthrough

[6] http://www.cambridge-mt.com/ms-mtk.htm

[7] https://weathervanemusic.org/shakingthrough/hezekiahjones

[8] This information can be found at https://s3.amazonaws.com/tracksheets/Hezekiah+Jones+-+Tracksheet.xlsx

[9] https://weathervanemusic.org/shakingthrough/hopalong

[10] This processing was performed by the author as part of a mix that was created prior to the conception of this study. That DAW session was opened and the four tracks to be used were exported.

[11] http://www.cambridge-mt.com/ms-mtk.htm#JulietsRescue_Heartbeats

[12] This mix was created by the author prior to the conception of the experiment.

suitability to single instruments, rather than full-bandwidth mixes [158]. This allows the relative loudness of instruments to be determined directly from the mix-space coordinates.

- For each song, two source positions were selected. The value of the vector Φ was selected using a random number generator, with two constraints:

    1. to ensure the two sources are sufficiently different, the pair of sources must be separated by unit Euclidean distance in the mix-space.
    2. to ensure the sources are not mixes where any track is muted, the values were chosen from the range $\pi/8$ to $3\pi/8$ (see Fig. 4.5).

4.3.2 Procedure

The experimental interface was designed using Pure Data, an open source, visual programming environment [13]. The GUI used by participants is shown in Fig. 4.9. Each participant listens to the audio clip in full at least once, then the audio is looped while mixing takes place and fader movement is recorded. The participant then clicks '`stop mix`' and the next session is loaded. For each session the user is asked to create their preferred mix by adjusting the faders. Since the number of dimensions in the mix-space is one less than the number of dimensions in the gain-space, by definition, more than one track must be active for a mix to exist. Consequently, the range of the faders was limited to ± 20dB from the source, to prevent solo-ing or muting any instrument, due to the uniqueness of the mix-space breaking down at boundaries. An initial trial was provided in order for participants to become familiar with the test procedure, after which the six conditions (3 songs, 2 sources each) were presented in a randomised order. The real-time audio output during mixing was recorded to a .WAV file at a sampling rate of 44,100Hz and a resolution of 16 bits. Fader positions were also recorded to .WAV files using the same format. As shown in Fig. 4.9, the true instrument levels were hidden from participants by displaying arbitrary, unmarked fader controls — the faders add ± 20 dB offset to the source position. This prevented participants from simply mixing *'visually'*, by recognising patterns in the fader positions.

4.3.3 Cohort A — Headphones

The first experiment using the mix-space concept and Fig. 4.9 took place in April 2015. This experiment was conceived as a pilot test and to collect data which could be used to verify the mix-space concepts before proceeding. In total, eight participants (two female, six male) took part in this mixing experiment. As staff and students within Acoustics, Digital Media and Audio Engineering at University of Salford, each of these participants had prior experience of mixing audio signals. The mean age of participants was 25 years and none reported hearing difficulties. The mean test duration was 14.2 minutes, ranging from 11 to 17 minutes.

Rather than use loudspeakers in a typical control room, the test set-up used a more neutral reproduction. The experiment was conducted in a semi-anechoic chamber at University of Salford, where the background noise level was negligible. Audio was reproduced using a pair of Sennheiser HD 800 headphones, connected to the test computer by a Focusrite 2i4 USB interface. Due to the nature of the task and the wide loudness range of the experiment, each participant adjusted the playback volume as required. Reproduction was monaural, presented equally to both ears.

[13] `https://puredata.info/`

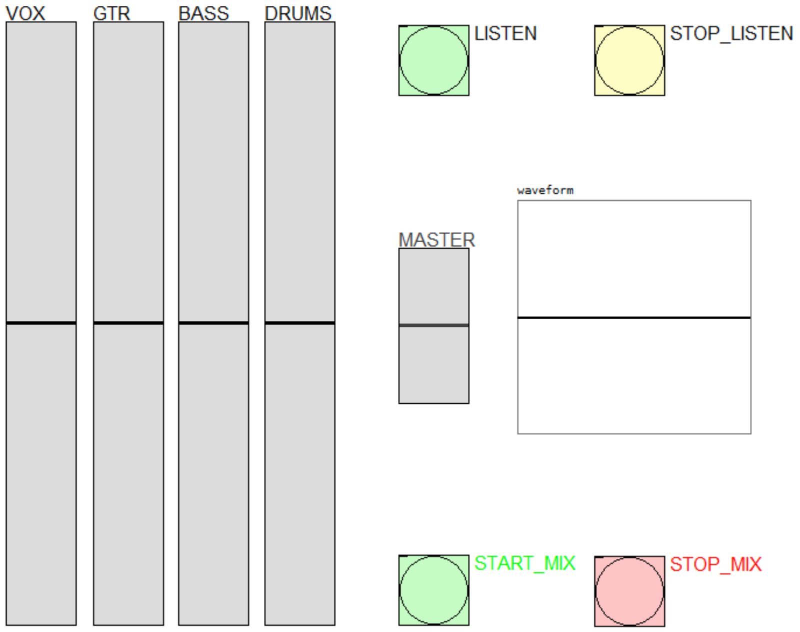


Figure 4.9: GUI of mixing test. The faders are unmarked and all begin at the same central value, which prevents participants from relying on fader position to dictate their mix.

While the choice between loudspeakers and headphones is often debated [46, 159], in this case, particularly as reproduction was mono, headphones were considered to be the choice with greater potential for reproducibility. Some results of this initial experiment were analysed and reported in [153].

4.3.4 Cohort B — Loudspeaker

A follow-up experiment was conducted in October 2015, after [153] was written, peer-reviewed and presented. One notable difference between the pilot test and the follow-up was the change in environment and reproduction system, from headphones to a single loudspeaker. The environment also changed from a semi-anechoic chamber to a BS.1116 listening room at the University of Salford, however, since the first test used headphones, the acoustic effect of that experiment's environment should be considered negligible. The decision was made to repeat the experiment with a loudspeaker in order to prepare for the stereo experiment, which is described in § 4.5. The loudspeakers used were Genelec 8020a, positioned in an LCR set-up, as shown in Fig. 4.11. While only the centre loudspeaker was used, playing back the mono signal, the left and right speakers were so positioned to provide continuity (both visual and acoustic) with the (future) experiment with stereo playback (see § 4.5). The measured room response is displayed in Fig. 4.10. A new test panel was recruited, consisting of 17 subjects who had not taken part in the pilot test. The median age of these participants was 27 years, ranging from 18 to 42. There were three female participants and 14 male participants.

4.3.5 Results

For each participant, song and source, the recorded time-series data was downsampled from 44.1 kHz to 10 Hz (an interval of 0.1 seconds), then transformed from gain to mix domains using Eqn. 4.1, with $r = 1$. From this data the vectors representing *mix-velocity*, described in Section § 4.2.2, were obtained using Eqn. 4.3.

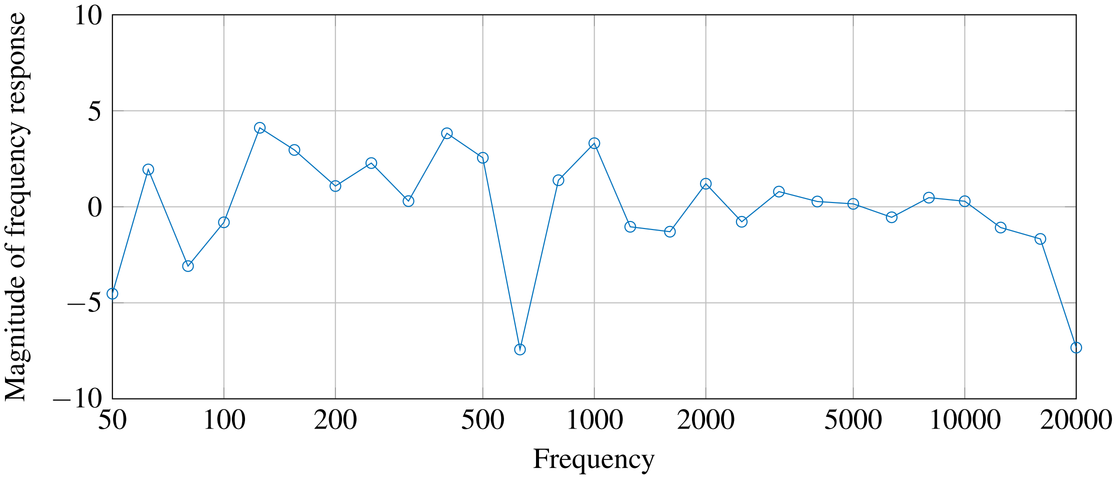


Figure 4.10: Magnitude of the frequency response at the listening position for mix-space after all furniture and equipment had been placed. Audio produced by Genelec 8020a in a BS.1116 listening room. Shown are the third-octave band levels, where 0 dB is the geometric mean from 50 to 20,000 Hz. The dips at ≈ 600 Hz may be caused by placement of furniture used in test (see Fig. 4.11).

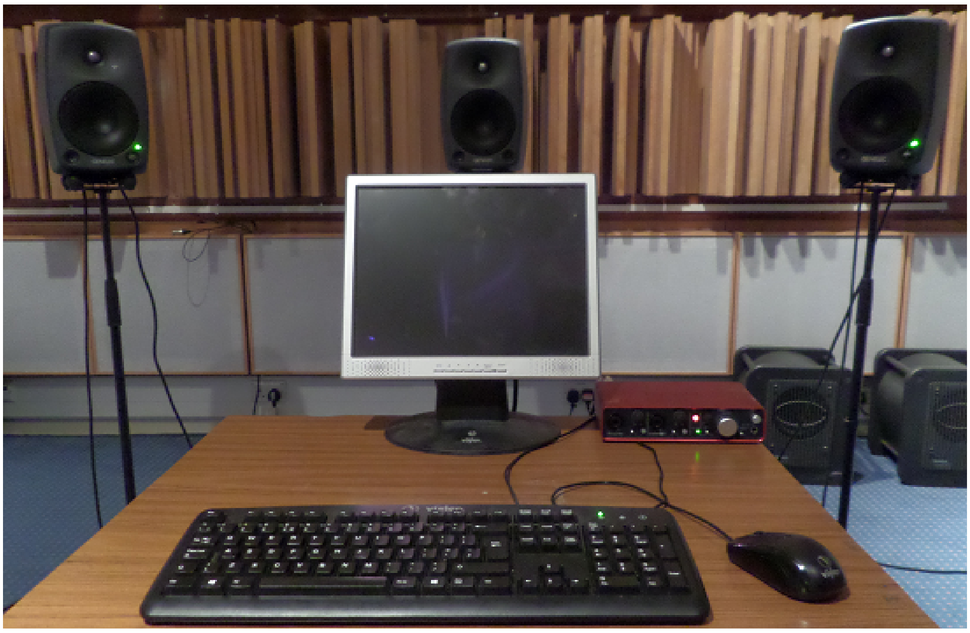


Figure 4.11: Mix-space test set-up in BS-1116 listening room.

Table 4.1: Median levels per group

System	Song	Vox	Guitar	Bass	Drums
Headphones	S1-Borrowed Heart	-3.51	-9.01	-7.68	-8.81
	S2-Sister Cities	-3.13	-7.66	-9.26	-7.63
	S3-Heartbeats	-2.13	-7.63	-11.69	-8.14
Loudspeaker	S1-Borrowed Heart	-3.28	-8.92	-6.40	-8.36
	S2-Sister Cities	-2.30	-8.89	-10.43	-8.33
	S3-Heartbeats	-2.66	-8.52	-12.66	-7.38

4.3.5.1 Instrument levels

Investigating research question RQ.9, the ideal loudness levels of each instrument in the mix was determined from the experimental data. In the boxplots which follow the median is marked at the central position and the box covers the interquartile range. The whiskers extend to extreme points not considered outliers and outliers are marked with a dot. Two medians are significantly different at the 5% level if their notched intervals do not overlap.

Since the experiment is concerned with relative loudness levels between instruments and not the absolute gain values which were recorded, normalised gains can be calculated from Eqn. 4.2, with $r = 1$. When all songs, sources and participants are considered, the distribution of normalised gains at the final mix positions is shown in Fig. 4.13, expressed in LU relative to the total mix loudness. Fig. 4.13 shows good agreement with previous studies, particularly a level of ≈ -3 LU for vocals [36, 40] and ≈ -10 LU for bass (see Fig. 1 of [36], which is shown in § 2.2.1 as Figure 2.4).

There was subtle variation in the levels of instruments over songs, summarised in Table 4.1. Figure 4.14 shows the variation in vocal level for each cohort and song. For each song there is no significant difference between headphone and loudspeaker groups. For the loudspeaker cohort there is no significant difference between songs. For the smaller cohort of headphone users there is a significant difference between the median level set for song 1 and song 3, although the sample size ($n = 8$) is a likely cause of this variation. The data in Fig. 4.15 indicates no significant difference in the median level set for the guitar track across cohort or song.

Figure 4.16 shows the distribution of final levels set for the bass track. Here, the median levels set are significantly different for song 1 and song 3. There is no significant effect of cohort in the median levels, although the variance is notably greater for the loudspeaker cohort, for songs 2 and 3. This could be partly explained by the larger sample size ($n = 17$) although room acoustics and reproduction system are expected to have played a part. In the loudspeaker group, the posture of the participant could have contributed to the perception of bass frequencies, due to room modes.

The distribution of drums level in the final mixes is displayed in Fig. 4.17. There is little variation observed across these six groups, although, as in Fig. 4.13, there are a number of outliers in the loudspeaker cohort who set the level of the drums quite low in the final mix.

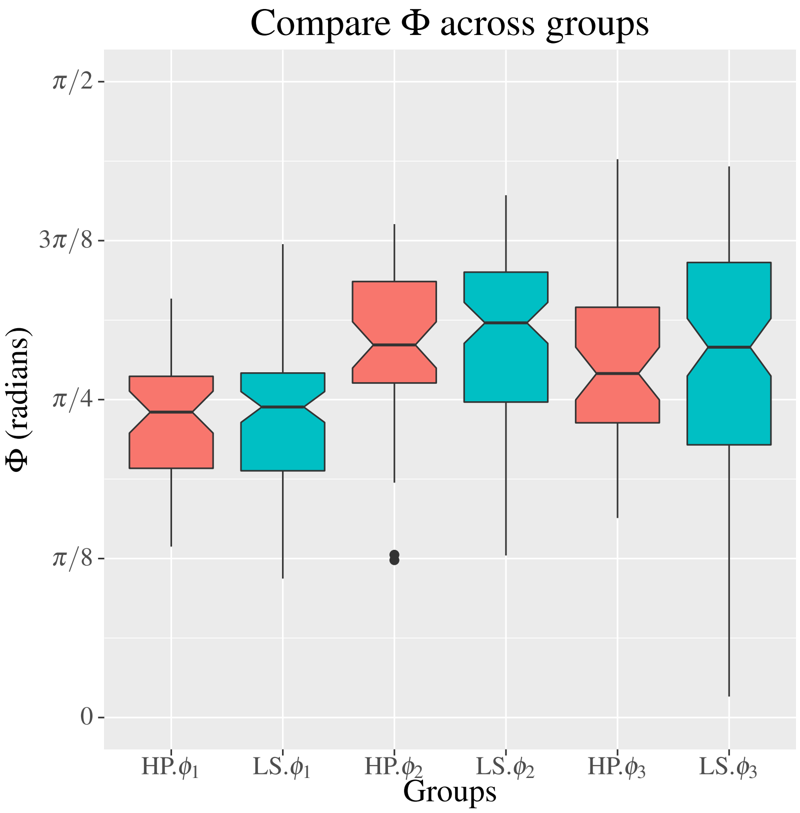


Figure 4.12: Boxplot of Φ for all songs and sources, grouped by cohort

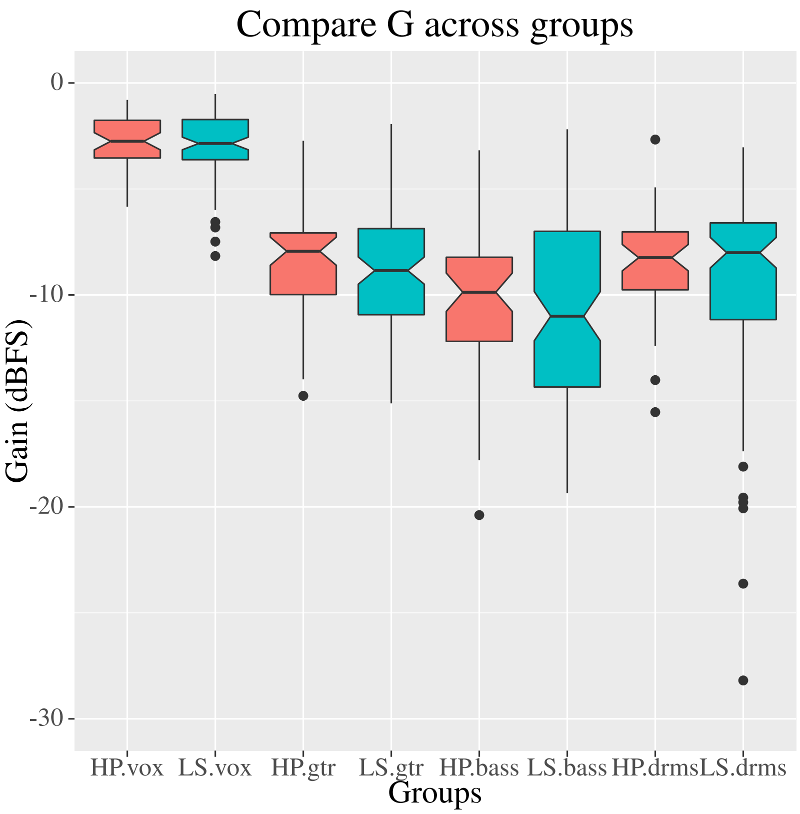


Figure 4.13: Boxplot of *G* for all songs and sources, grouped by cohort

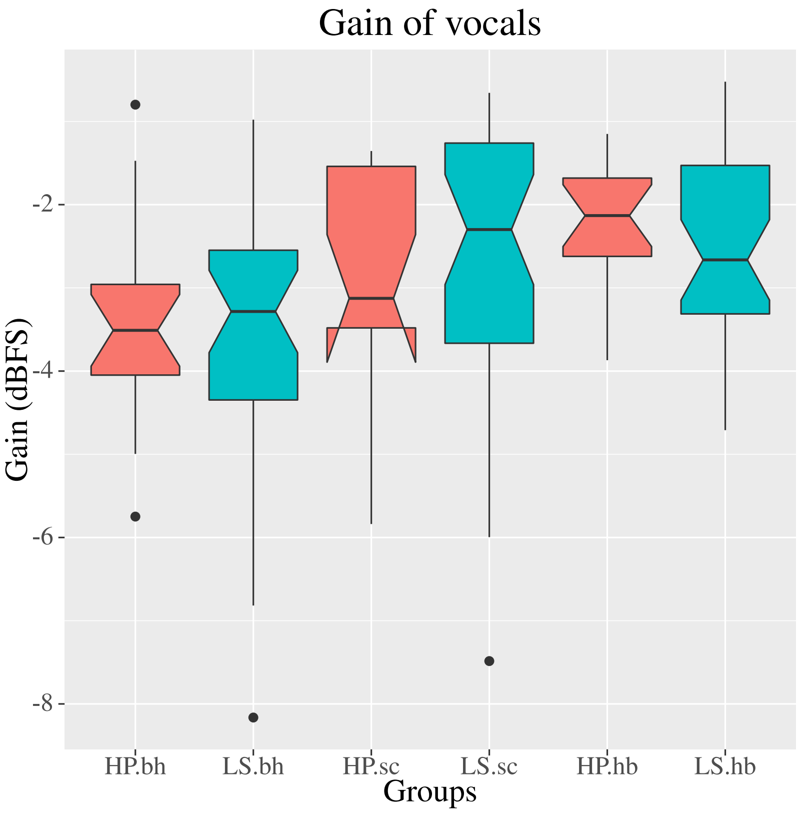


Figure 4.14: Boxplot of vocal level for all sources, grouped by cohort and by song.

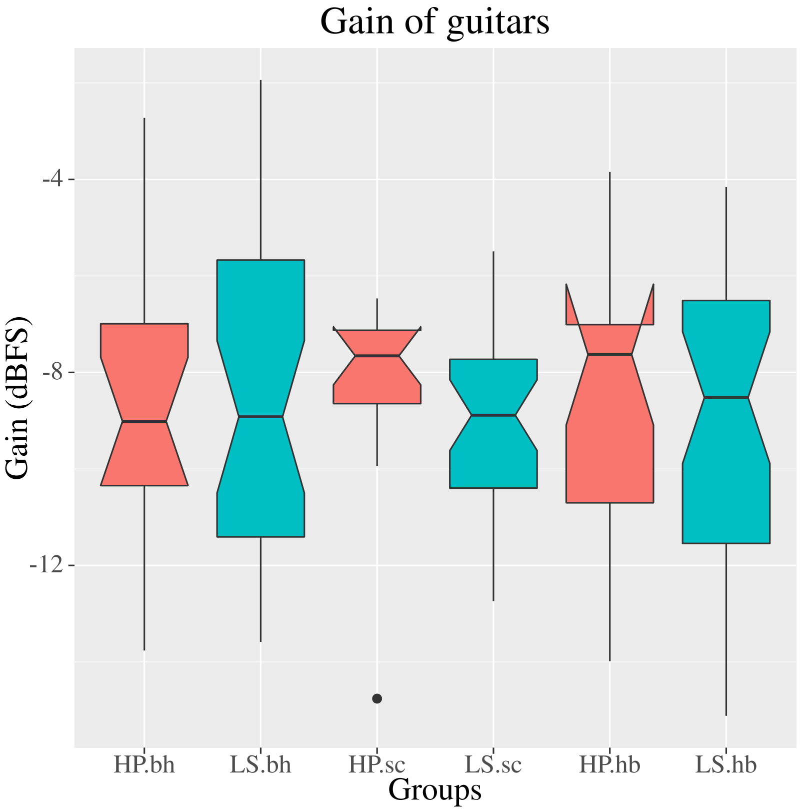


Figure 4.15: Boxplot of guitar level for all sources, grouped by cohort and by song

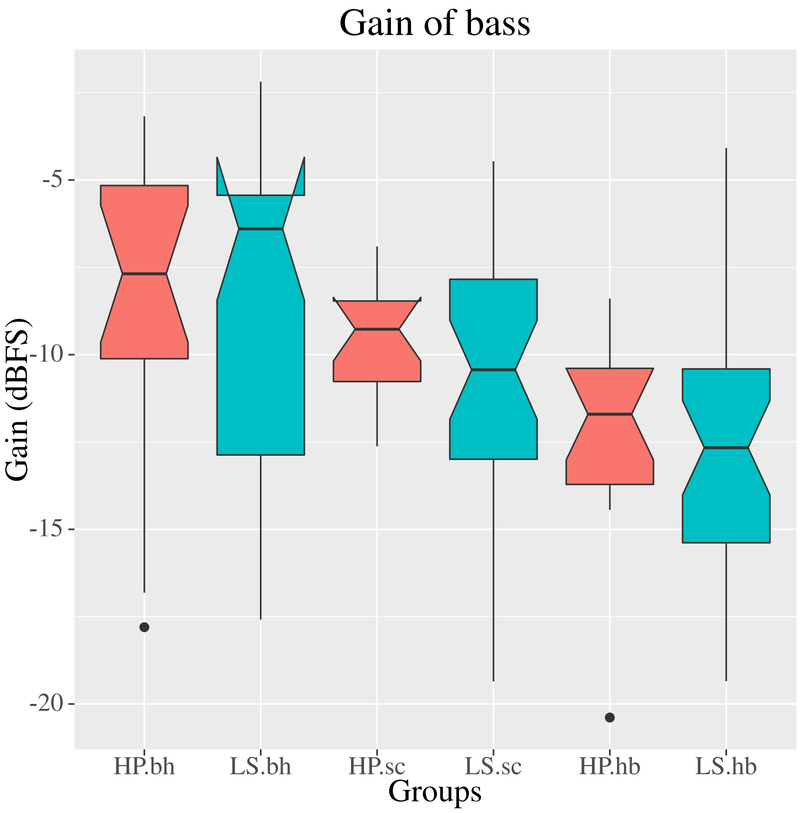


Figure 4.16: Boxplot of bass level for all sources, grouped by cohort and by song

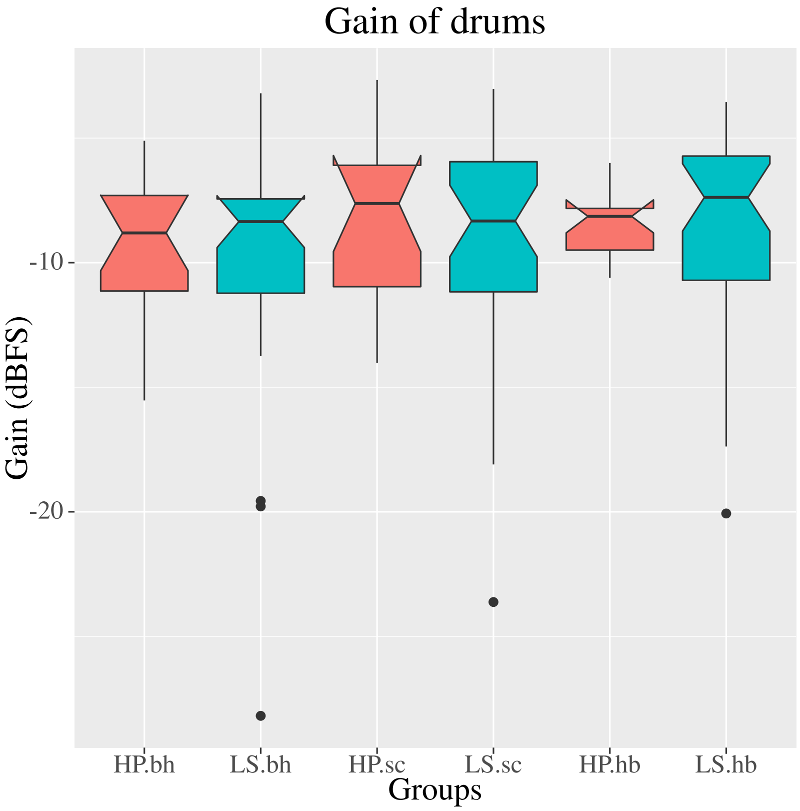


Figure 4.17: Boxplot of drums level for all sources, grouped by cohort and by song

4.3.5.2 Source-directivity

Since each participant was required to listen to the audio before mixing began, it was hypothesised that participants would make similar initial changes to the mix, such as when one instrument required a clear change in level. Movement away from the source is characterised by the first non-zero element of the mix-velocity triple u, v, w (see Eqn. 4.3). The displacement and direction of this step is used to investigate the source directivity. For each song and source, these vectors are plotted in Fig. 4.18 to 4.23. These vectors indicate the direction and step size of the first changes to the mix. As the participant had control over four faders there are only 2×4 possible initial actions that could be taken — to increase or decrease the level of each fader. However, this can produce a number of vectors in the mix space. One would not expect to observe anything approaching spherical radiation from the source with such low number of dimensions, only that each of the possible outcomes is equally likely. Figures 4.18 to 4.23 show the normalised vectors leaving the source for each participant. The similarity matrix is also shown, computed using the cosine distance metric. In each example there are at least two opposing vectors, which produces the maximum cosine distance of 2. Subsequently, darker similarity matrices indicate many similar vectors.

4.3.5.3 Mix-space navigation

Fig. 4.24 to 4.26 show the probability density function (PDF) of Φ_t when averaged over all 25 participants, where the solid line is from trials where the source was at position **A** and the dashed line, position **B**. The function is estimated using Kernel Density Estimation (KDE), using 100 points between the lower and upper bounds of each variable. This plot displays the mix configurations which the participants spent most time listening to and it is seen that all distributions are multi-modal. There are peaks close to the initial positions, the final positions and other interim positions that were evaluated during the mixing process.

There are a number of different approaches to multitrack mixing of pop and rock music, one of which is to start with one instrument (such as drums or vocals) and build the mix around this by introducing additional elements. Some participants were observed mixing in this fashion, shown in Figs. 4.24 to 4.26, where peaks at extreme values of ϕ_n show that instruments were attenuated as much as the constraints of the experiment would allow. For Song 1, ϕ_1 is reasonably well balanced and centered close to $\pi/4$. This indicates that mixers tended to listen in states where the relative loudness of the vocal and backing track were similar. This is also observed for song 3 but less so for song 2.

There are notable differences due to the source. The distributions for ϕ_3 in Song 1 suggest that exploration depended on the initial source configuration, with Source A leading to louder drums than Source B. Note that a value of $\pi/2$ for ϕ_1 or ϕ_2 simply indicates that the vocal or guitars were muted and so it is a frequently-occurring state. A value of 0 indicates that this track was solo-ed. For ϕ_3, $\pi/2$ indicates drums were muted and 0 indicates bass was muted.

In order to quantify the variation in mixes as they explored the space, the pairwise distance between mixes was calculated at each point in time. This data was used to create a dissimilarity matrix. The sum of all distances was used as a metric relating to mix-variation at a point in time. By converting the ϕ terms back to gain, using Eqn. 4.2, the normalised gains were obtained (where the norm of the G is equal to 1 at each point in time) when setting $r = 1$. The distance metric used

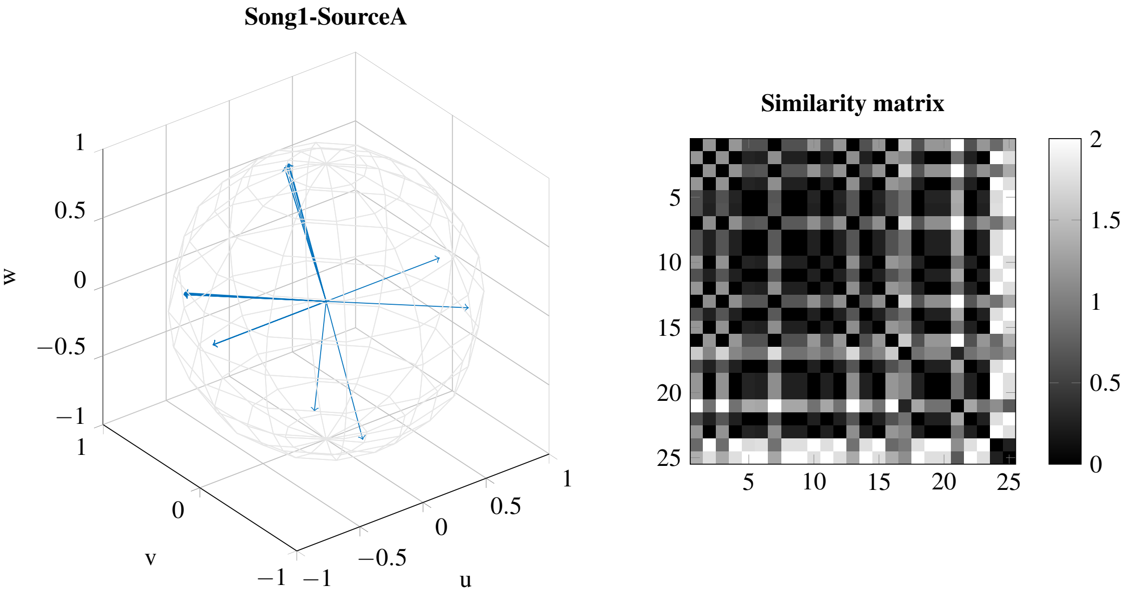


Figure 4.18: Source directivity — Song 1 Source A. This results shows good agreement between many participants.

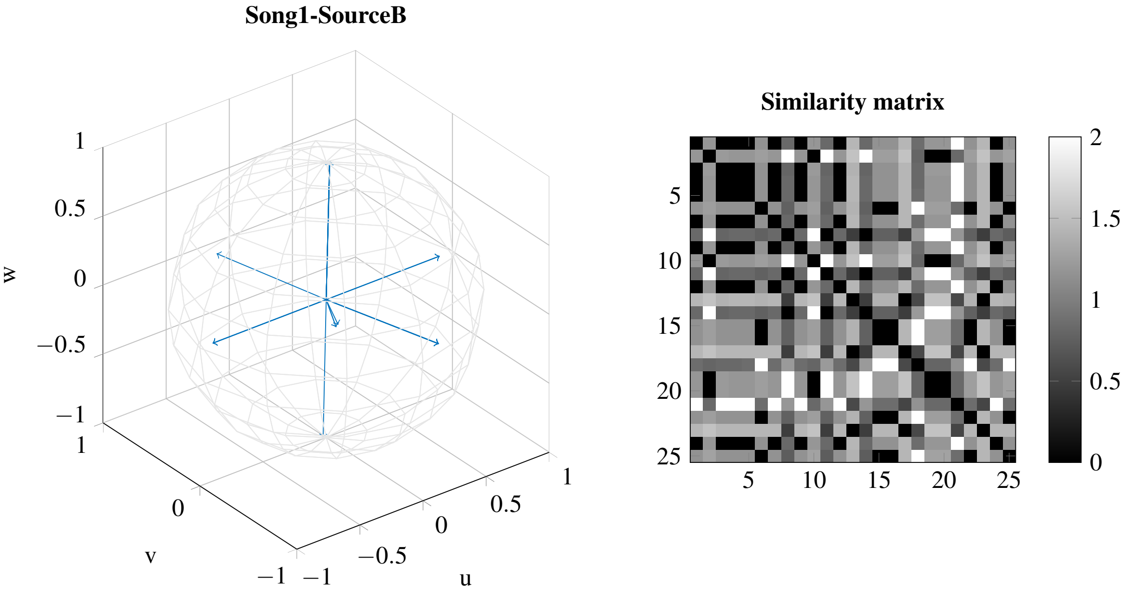


Figure 4.19: Source directivity — Song 1 Source B.

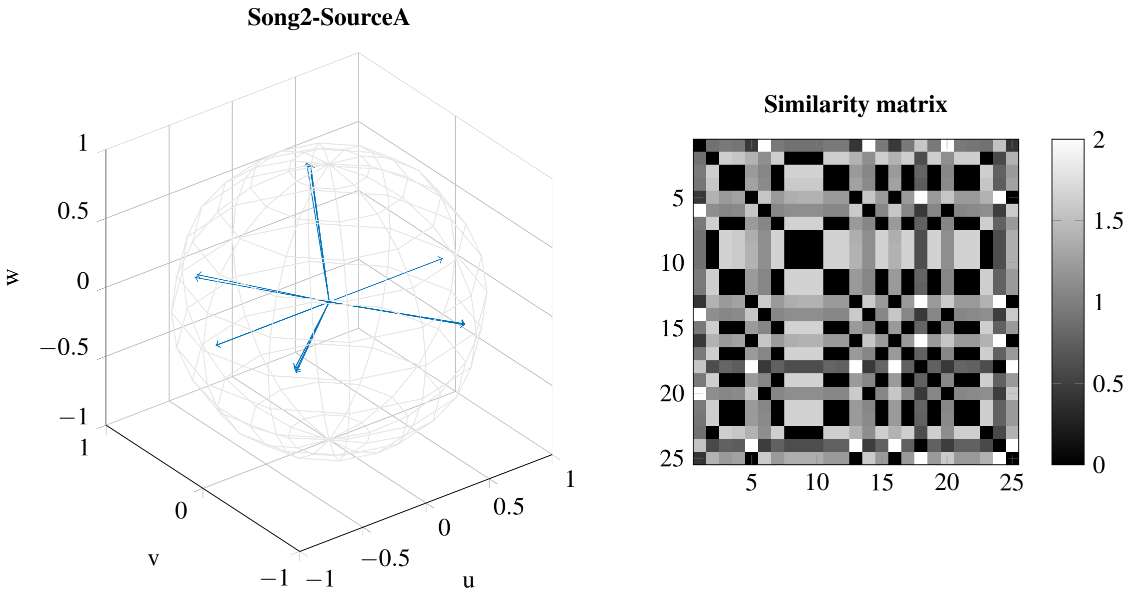


Figure 4.20: Source directivity — Song 2 Source A. Note the region of the space $(-w)$ in which there are no vectors.

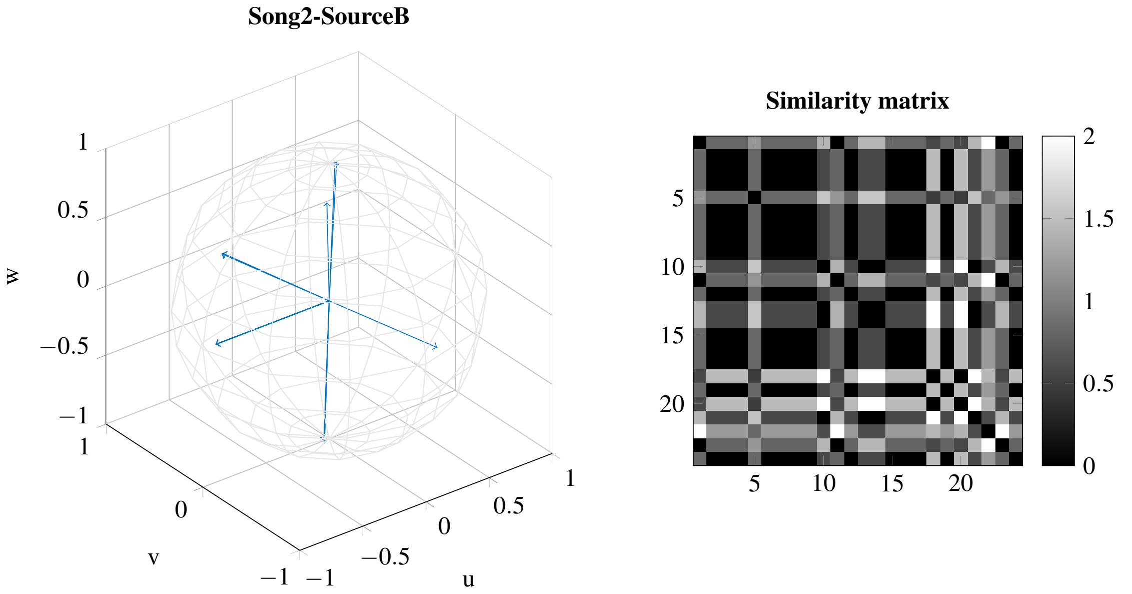


Figure 4.21: Source directivity — Song 2 Source B. Due to one corrupted entry only 24 data points are shown here. Note the region of the space $(-u)$ in which there are no vectors. Good agreement among many of the first 17 participants.

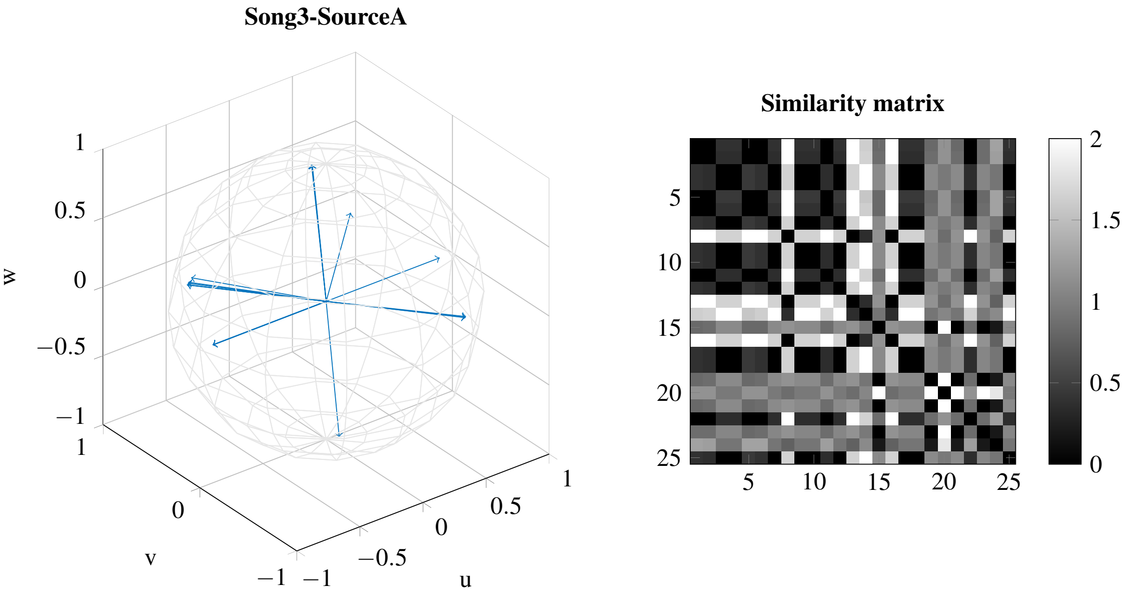


Figure 4.22: Source directivity — Song 3 Source A. High agreement between the first 7 participants.

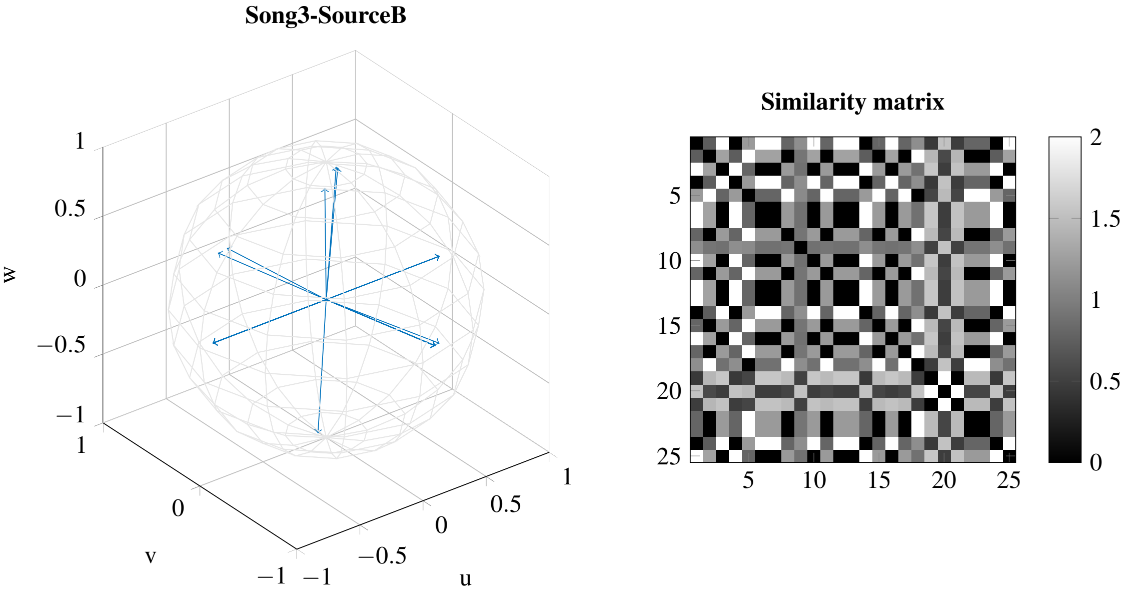


Figure 4.23: Source directivity — Song 3 Source B. The result shows a lack of agreement.

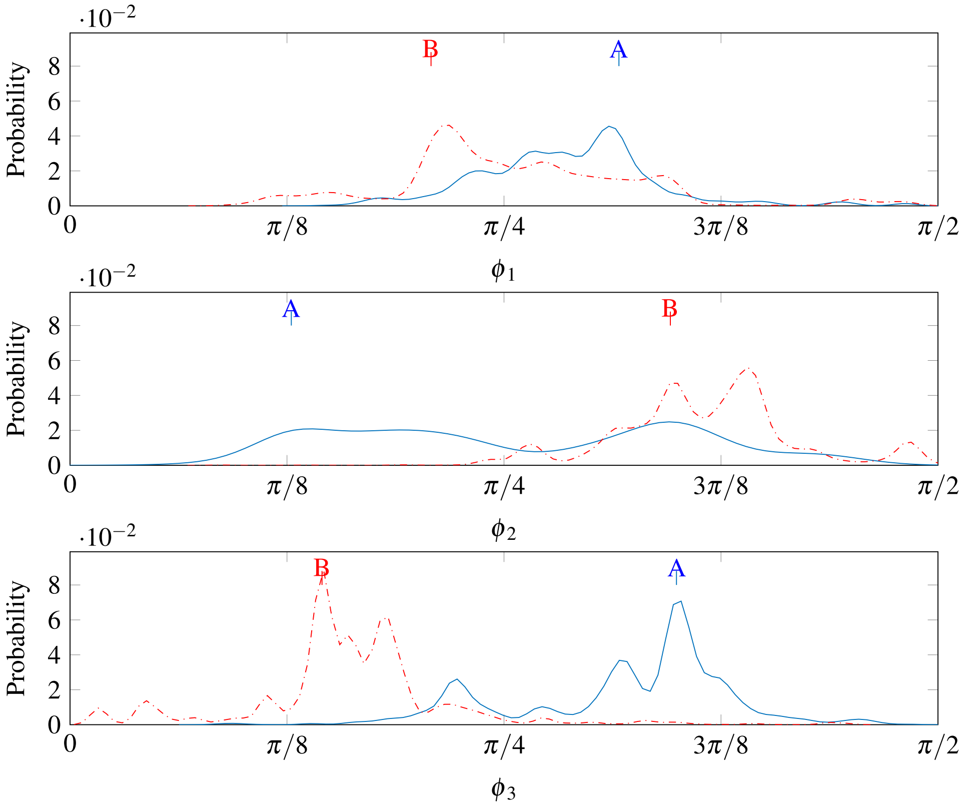


Figure 4.24: Estimated probability density functions of ϕ terms, for song 1, averaged over all mixers. Sources positions are highlighted with **A** and **B**.

was the cosine distance metric. This is standard for determining the distance between points on a sphere (in this case, a 3-sphere) [14]. The plots in Fig. 4.27a, 4.27b, and 4.27c, show the sum of this dissimilarity matrix at each point in time. Note that, for the sake of clarity in plotting, the number of points is reduced by a factor of 4, using the `decimate` function in MATLAB. A logarithmic axis scale is used, since most of the coarse mixing takes place in the earlier time periods, before settling down to fine adjustments towards the end.

As all participants begin at the same point, the initial value is equal to zero. In all three songs, the maximum levels of inter-participant variation take place after approximately 10-15 seconds, at which time a large region of the space is spanned by the 25 mixes. After this point, there is a slow convergence for the remaining duration. Note that, as mixing duration varied by participant, the final gain values for each mix was held until the final participant had completed mixing.

For songs 1 and 3, the amount of variation between mixes is less for source **B**. This suggests that this source was closer to an ideal mix than source **A** and, as a consequence, less exploration of the space was deemed necessary.

[14] Strictly speaking, it is not necessary to use the normalised gains when using the cosine distance metric. This metric is concerned with the difference in the angles between vectors. The length of these vectors is not important. Identical results are achieved when using the raw gain values or the normalised gain values.

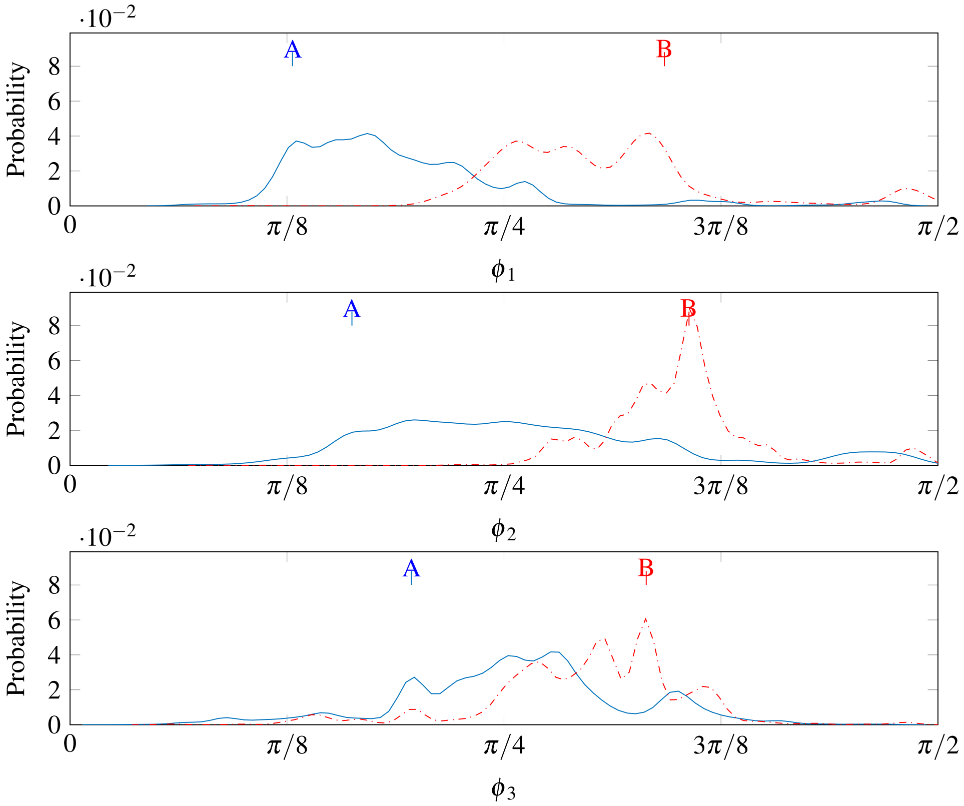


Figure 4.25: Estimated probability density functions of ϕ terms, for song 2, averaged over all mixers. Sources positions are highlighted with **A** and **B**.

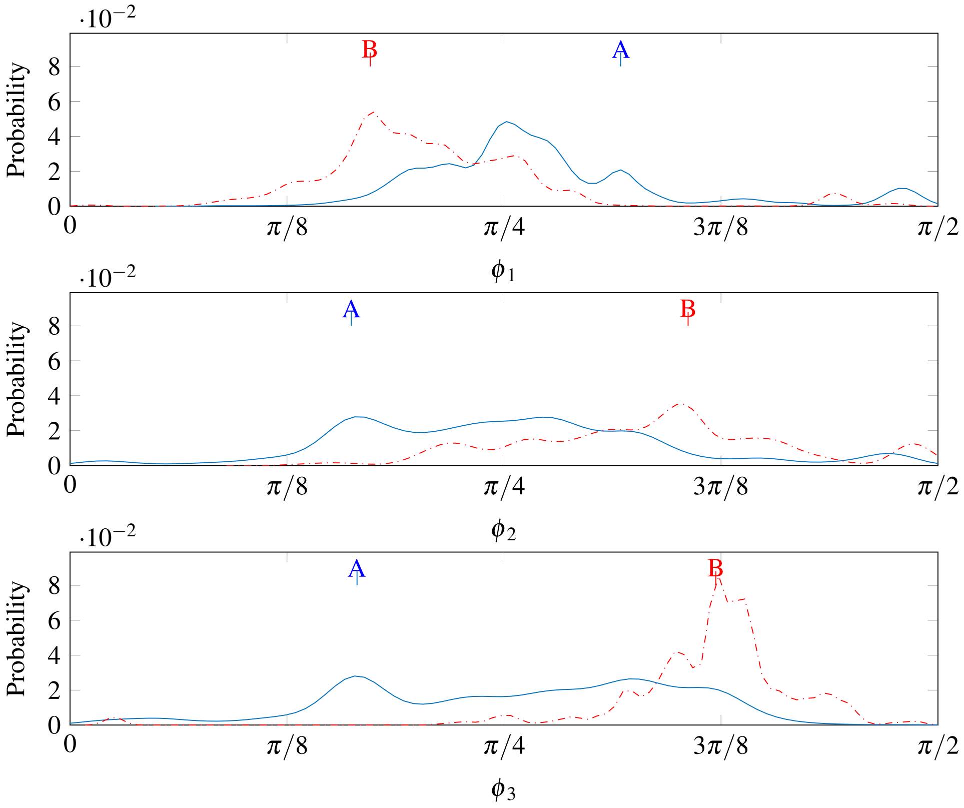
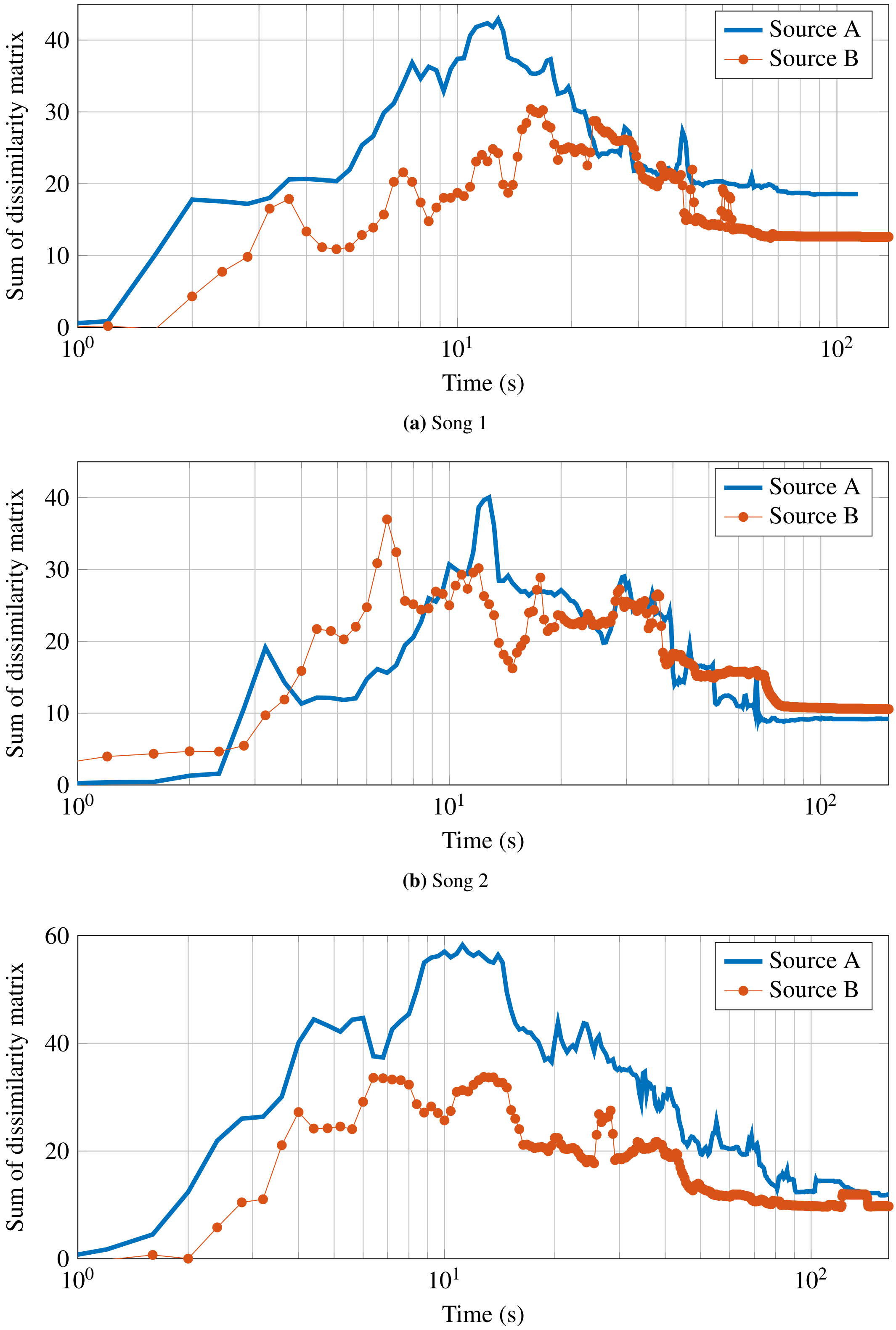


Figure 4.26: Estimated probability density functions of ϕ terms, for song 3, averaged over all mixers. Sources positions are highlighted with **A** and **B**.



(a) Song 1

(b) Song 2

(c) Song 3

Figure 4.27: Diversity in the set of mixes over time. As all trials begin at the same point, the value is zeros. Diversity then increases as each participant explores different regions in the space. Diversity then decreases over time, indicating a degree of convergence.

4.3.5.4 Sink convergence

Figures 4.27a, 4.27b, and 4.27c indicate that after an initial exploration phase, mixes begin to converge, with the distribution of final instrument levels already shown in Section 4.3.5.1. Final mixes created by participants show notable clustering due to the source position. For each song, the final mixes created after starting at source **A** can be clearly distinguished from those created after starting at point **B**. This is shown in Figures 4.28, 4.30 and 4.32. While these figures display the final mixes in the mix-space the clustering is determined differently, on the 3-sphere in the gain-space. Spherical k-means clustering [160] was used after the gains had been normalised onto the sphere (convert Φ back to G, using Eqn. 4.2, with $r = 1$, as shown in Fig. 4.13).

Clustering due to reproduction system was also investigated yet no apparent difference was determined between the loudspeaker and headphone cohorts. The greater variance in the loudspeaker cohort, as shown in Figure 4.12, is also observed in Figures 4.29, 4.31 and 4.33.

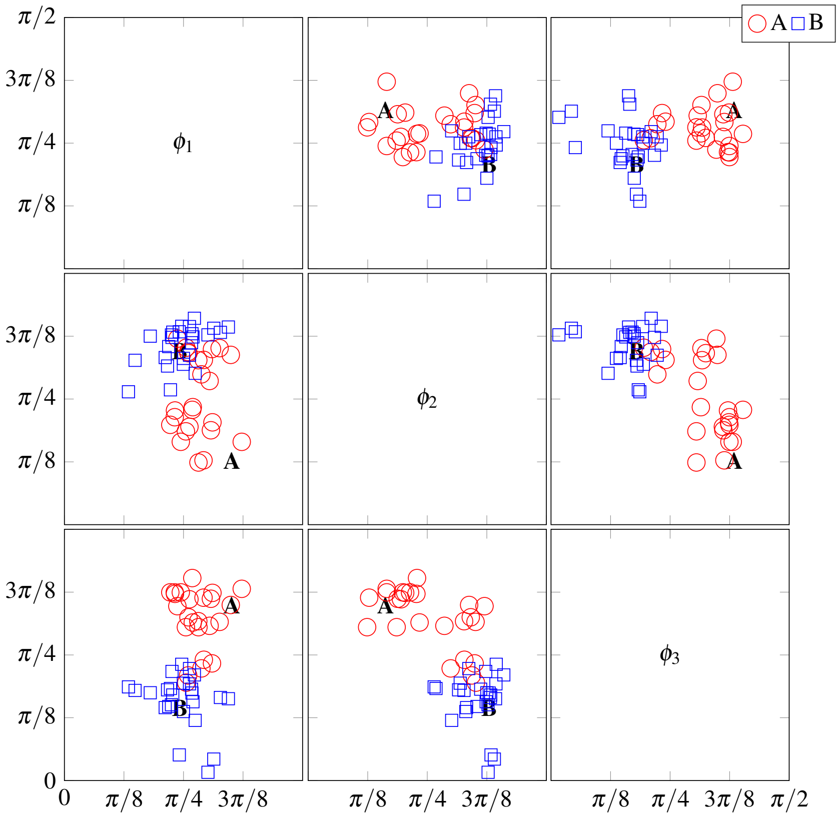


Figure 4.28: Final mixes, grouped by source position, for song 1

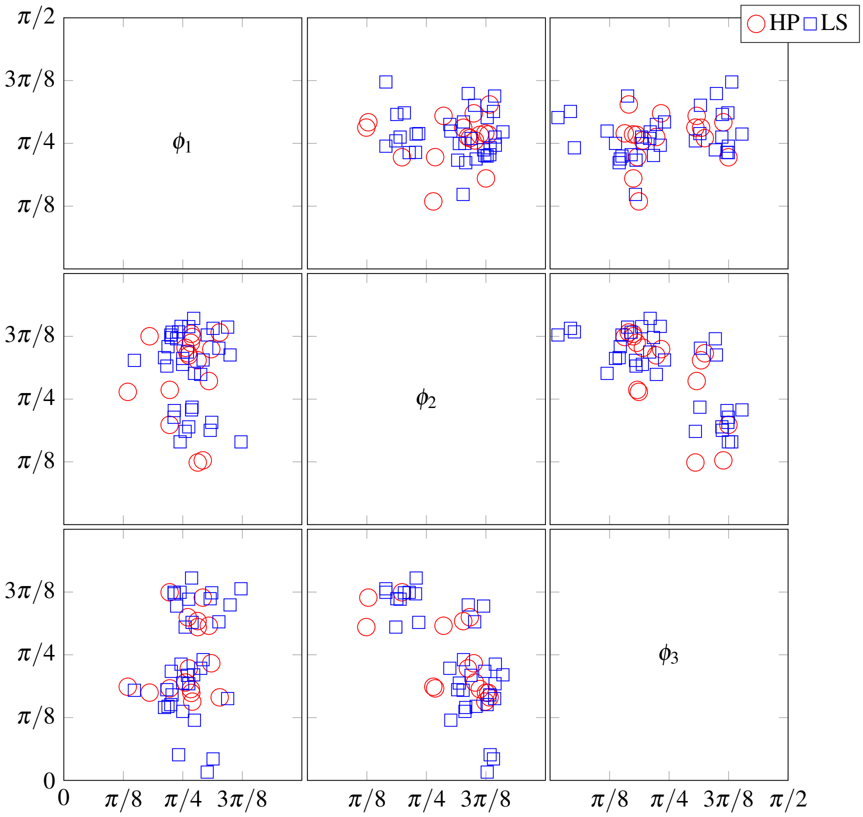


Figure 4.29: Final mixes, grouped by experimental group, for song 1

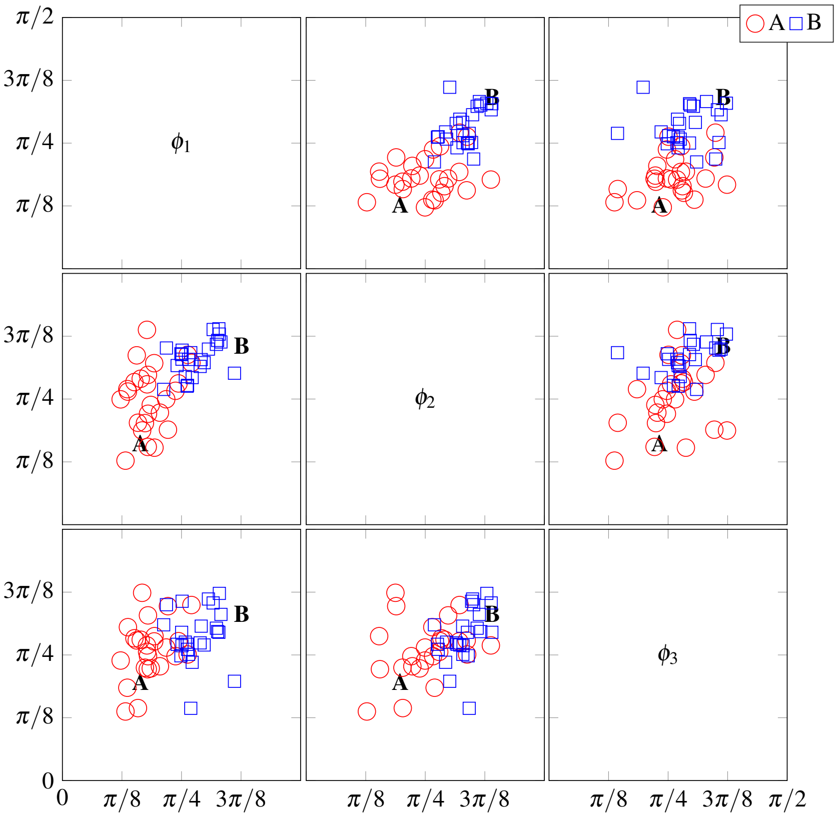


Figure 4.30: Final mixes, grouped by source position, for song 2

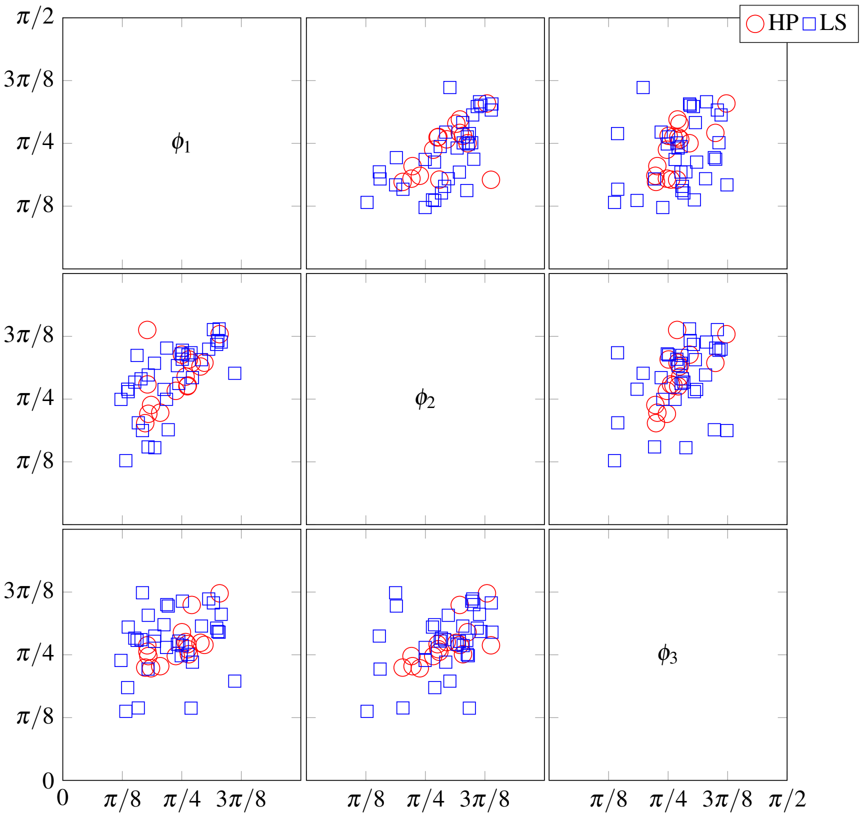


Figure 4.31: Final mixes, grouped by experimental group, for song 2

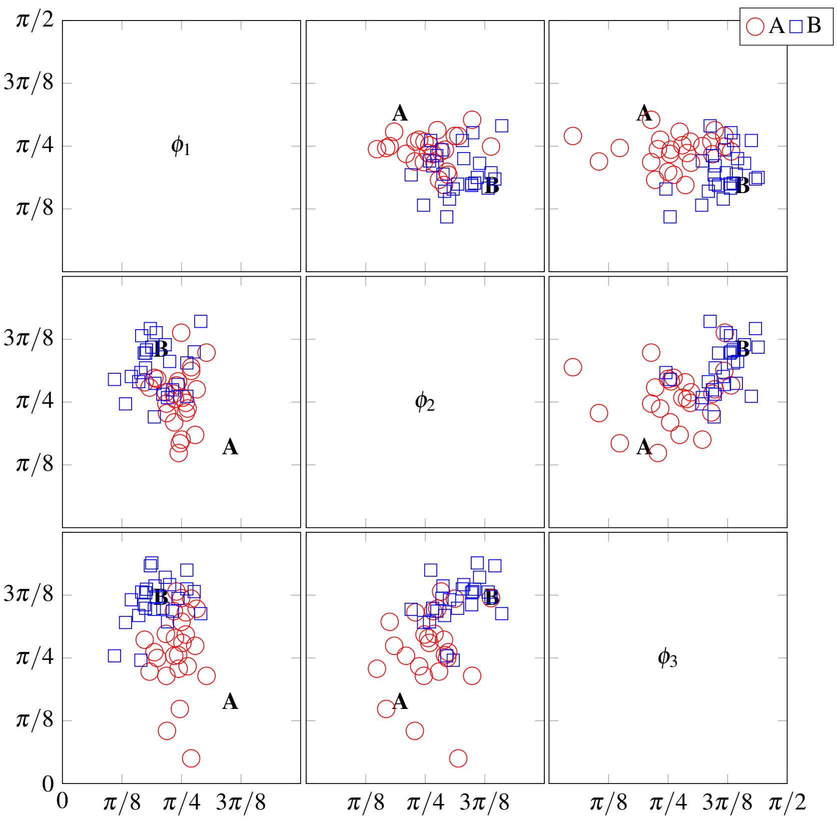


Figure 4.32: Final mixes, grouped by source position, for song 3

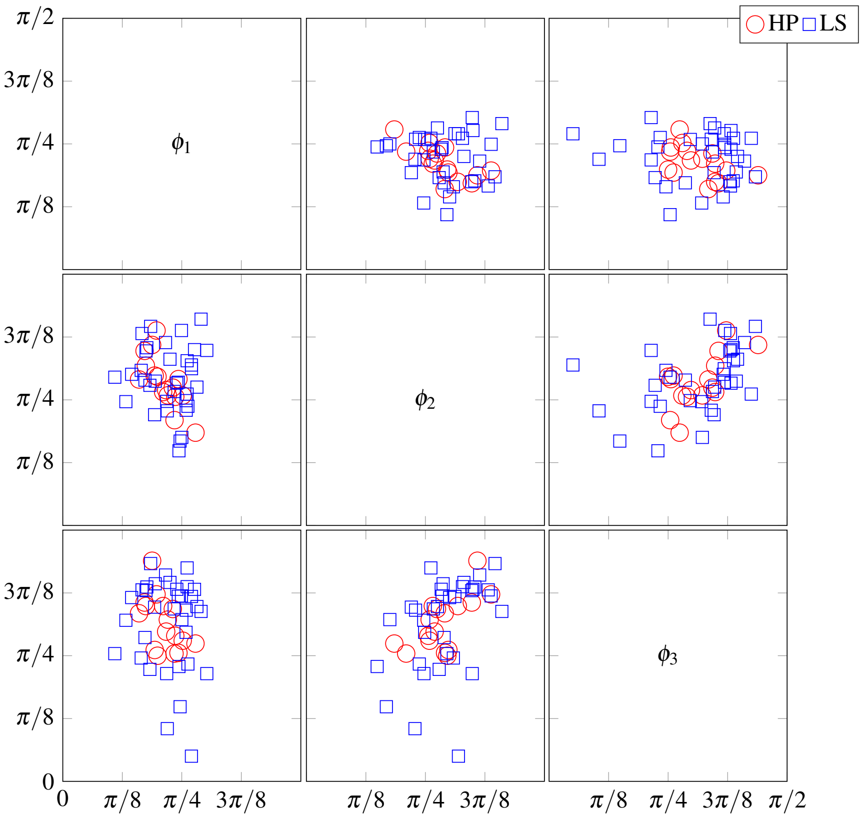


Figure 4.33: Final mixes, grouped by experimental group, for song 3

4.4 Further Theory

In order to extend the mix-space concept to a more realistic mixing scenario, equalisation and panning were added to the model. While only track gain has been considered thus far, equalisation is merely a frequency-dependent gain and panning a channel-dependent gain.

4.4.1 Equalisation

Similarly to how the mix can be considered as a series of *inter-channel* gains, when the frequency-response of a single audio track is split into a fixed number of bands, the *inter-band* gains can be used to construct a *"tone-space"* using the same formulae. With the gain of the low, middle and high bands in the filter being g_L, g_M and g_H respectively, then the problem is comparable to the 3-track mixing problem shown in Figure 4.2. Again, one can convert this to spherical coordinates (by Equations 4.1) and obtain $[r_{\text{EQ}}, \phi_{1\text{EQ}}, \phi_{2\text{EQ}}]$, yet, in this case, the values of $\phi_{n\text{EQ}}$ control the EQ filter applied, and r_{EQ} is the total loudness change produced by equalisation. As before, if all three bands are increased or decreased by the same proportion, then the tone of the instrument does not change apart from an overall change in presented loudness, r_{EQ}. Analogous to its use in track gains, the value of $\phi_{2\text{EQ}}$ adjusts the balance between g_M and g_H, while ϕ_{1EQ} adjusts the balance of g_L to the previous balance.

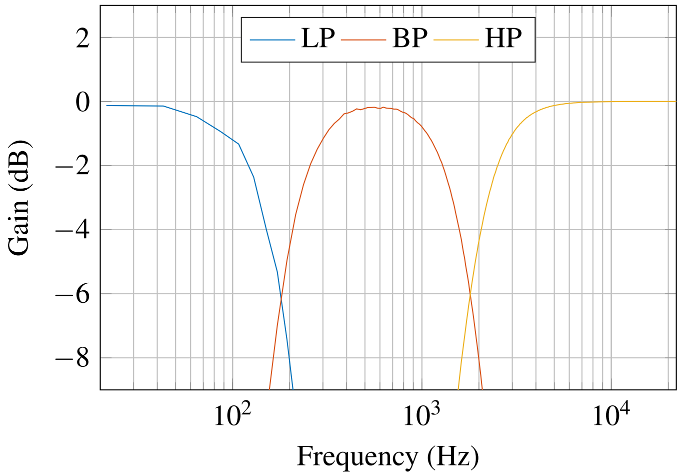


Figure 4.34: Example of a 3-band crossover filter, using 4th order Linkwitz-Riley filters, which can be used as a basic 3-band EQ

In Fig. 4.35, five points are randomly chosen in the "tone-space". These co-ordinates are converted to three band gains as before, except that, in order to center on a gain vector of $[1,1,1]$, $r_{EQ} = \sqrt{N_{\text{bands}}}$, which is $\sqrt{3}$ in this example. Of course, for this to work, one must assume an audio track has equal loudness in each band and this is rarely the case. When g_L is increased on a hi-hat track there may be little effect, compared to a bass guitar. Therefore, the loudness change r_{EQ} is a function of the spectral envelope of the track, prior to equalisation (it is shown later that this effect is negligible and so it is not considered herein).

4.4.2 Panning

Thus far, only mono mixes have been considered, where all audio tracks are summed to one channel. In creative music production, it is rare that mono mixes are encountered (although notable exceptions can still be found). The same mathematical formulations of the mix-space can be used

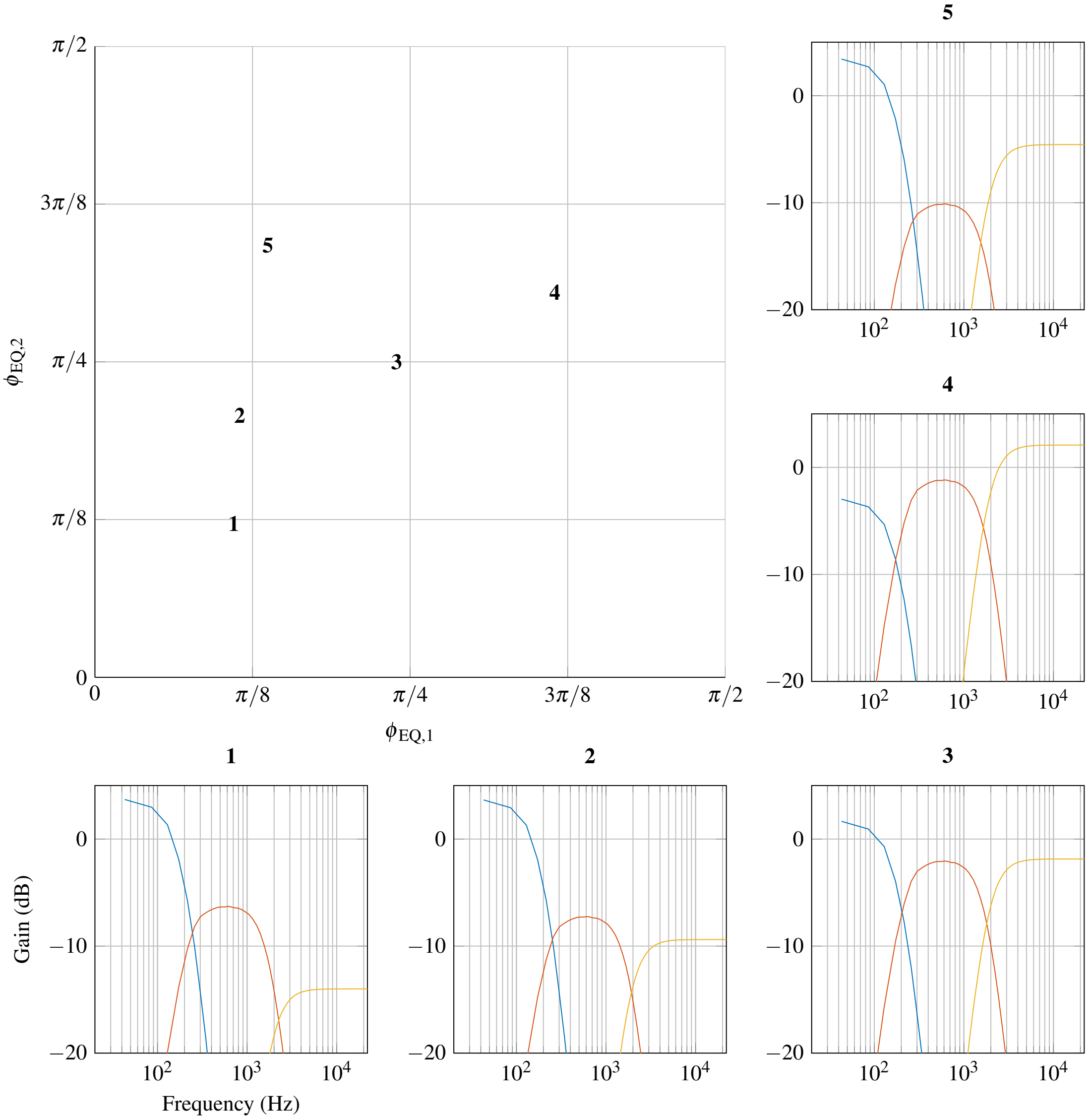


Figure 4.35: Five randomly-chosen examples of a 3-band EQ, chosen from 2D tone-space. As $\phi_{EQ,2}$ goes to zero, the gain of the high band decreases. As $\phi_{EQ,1}$ goes to zero, the gain of the low band increases at the expense of the other two bands, their balance determined by $\phi_{EQ,2}$.

to represent panning. Consider Fig. 4.8, which shows track gains in the range $[-1,1]$. Should these be replaced with track pan positions p_n then the mix-space (or "pan-space") can be used to generate a position for each track in the stereo field.

However, the mix-space for gains takes advantage of the fact that a mix (in terms of track gains only) is comprised of a series of inter-channel gain ratios, meaning that the radius r is arbitrary and represents a master volume. In terms of track panning one would obtain a series of inter-channel panning ratios, the precise meaning of which is not clear. Additionally the radial component would still be required to determine the *exact* pan position of the individual track.

For a simple example with only two tracks, the meaning of r_{pan} and ϕ_{pan} are simple to understand. Consider the unit circle in a plane where the cartesian coordinates (x,y) represent the pan positions of two tracks, as shown in Fig. 4.36. Mix A is at the point $(0.707, 0.707)$: both tracks are panned at the same position. As this is a circle with arbitrary radius, r_{pan}, then the radius controls how far

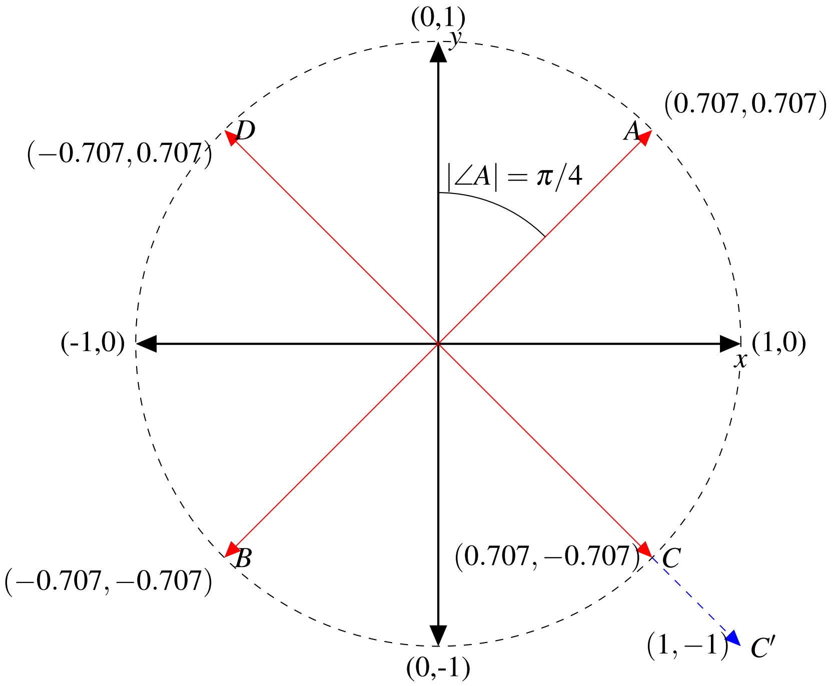


Figure 4.36: Panning of two tracks

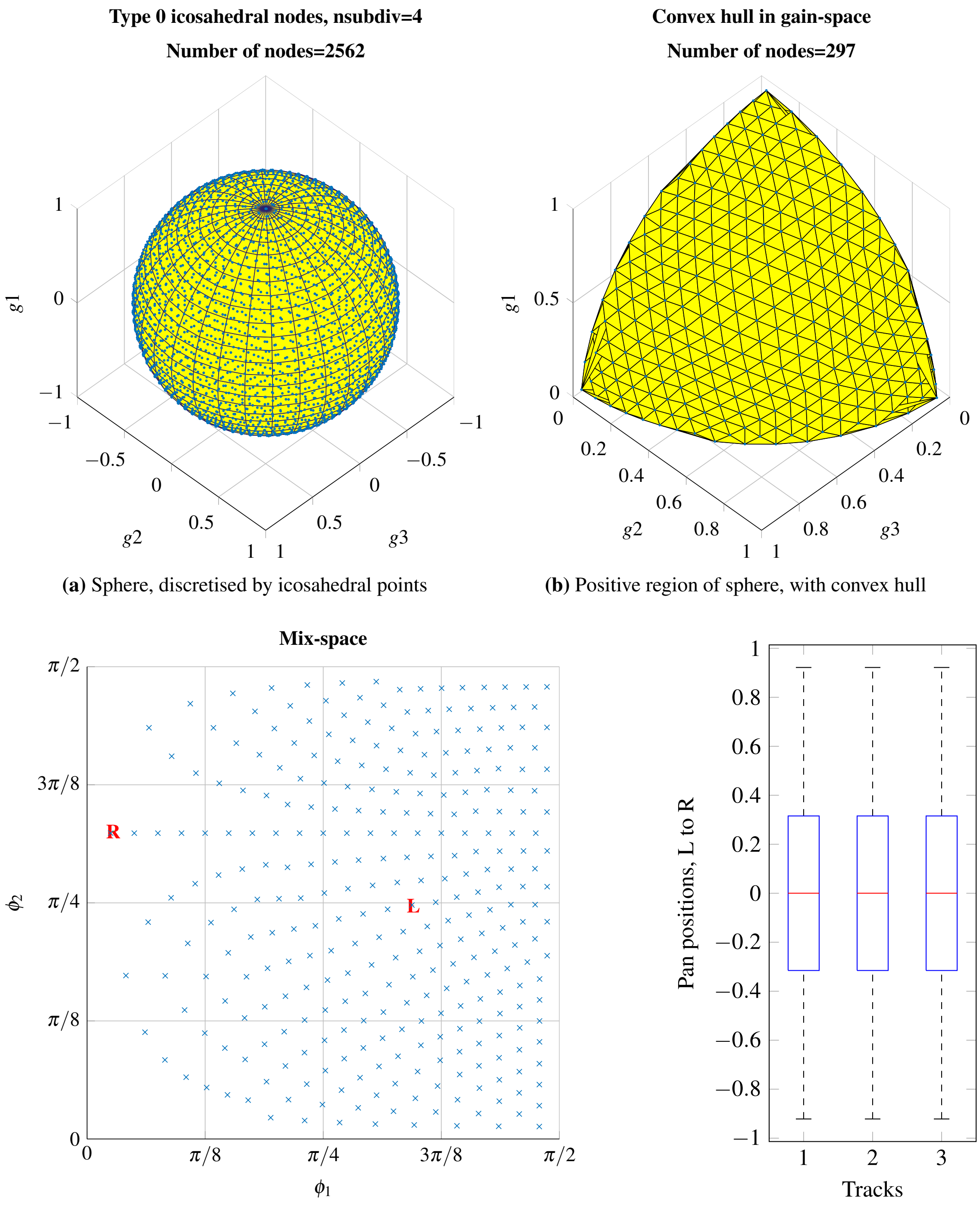
positive (right) the two tracks are panned, from 0 (centre) to $+1$ (far right). Mix B does the same but towards the left channel. Is this the same mix? Are A and B identical "panning-mixes", as p and p' were identical "gain-mixes"?

Now consider mix C, where one track is panned left and the other right. Mix D is simply the mirror image of this. Are *these* to be considered as the same mix? Here r_{pan} adjusts the distance between the two tracks, from both centre when $r_{\text{pan}} = 0$, to$(-1,1)$ when $r_{\text{pan}} = \sqrt{2}$ (as indicated by mix C'). Does a change in r_{pan} change the mix, or is it simply the same mix only wider/narrower? Here the term *mix* applies to panning mix (not a mix of gains, as it was in earlier sections). Overall, the angle ϕ_{pan} adjusts the panning mix and r_{pan} is used to obtain absolute positions in the stereo field, at a particular scale (i.e. to zoom in or zoom out).

Alternatively, if two points in the mix-space are chosen, one to represent the balance of instruments in the left channel and one for balance of instruments in the right channel (or as many as needed for a multi-channel system), then a stereo mix can be created. Figure 4.37a shows a 2-sphere representing all the mixes of three tracks. This space is discretised according to icosahedral subdivision [161][15]. The 297 points in the positive region of this space are shown in Fig. 4.37b along with the convex hull of the points. The precise number of points depends on the number of subdivisions used (here, $N_{\text{subdiv}} = 4$). Figure 4.37c shows these points in the 2D mix-space (as in earlier figures). Two points, marked **L** and **R** are chosen as the left and right mixes respectively.

It can be shown, however, that not all possible combinations of pan positions can be achieved. For example, given three tracks, a vocal, drums and guitar, then it would not be possible to pan the vocal centrally while simultaneously panning both other instruments hard left. This is due to the

[15] Using RBFSPHERE package for Matlab, available from `http://math.boisestate.edu/~wright/montestigliano/index.html`



(a) Sphere, discretised by icosahedral points

(b) Positive region of sphere, with convex hull

(c) Positive region of sphere, mapped as spherical coordinates. LEFT and RIGHT mixes are highlighted. The L mix contains a *roughly* equal blend of all tracks while R was chosen as it is mostly vocals.

(d) Boxplot showing pan positions of all possible mixes in Fig.4.37c.

Figure 4.37: Creation of a stereo mix by choosing two points in the mix-space. With 297 points, there are 297^2 (88,209) possible stereo mixes.

Table 4.2: Parameters of the mix selected from Fig. 4.37c

	Vox	Gtr	Drums
g_L (dBFS)	-6.6212	-3.9978	-4.1569
g_R (dBFS)	-0.0329	-26.8935	-22.5907
p	-0.3621	0.8663	0.7861

fact that, to be panned centrally, the vocal must be presented at the same level in both channels. While the right channel would contain only vocals, the introduction of other instruments in the left channel would necessitate a reduction in vocal level in order for both 'mixes' to be presented at equivalent loudness. As a result, the vocal would be louder in the right channel and appear panned towards the right. To pan centrally, either the right channel would need to be attenuated or the left channel amplified. Table 4.2 shows the gain and resultant pan positions of each of the three tracks, based on the points chosen in Fig. 4.37c.

The pan position p_i of a track i is a function of the left and right gains of that track, as shown in Equation 4.4. Figure 4.38 displays a plot of this equation.

$$p_i = \frac{(g_{L,i} - g_{R,i})}{(g_{L,i} + g_{R,i})} \tag{4.4}$$

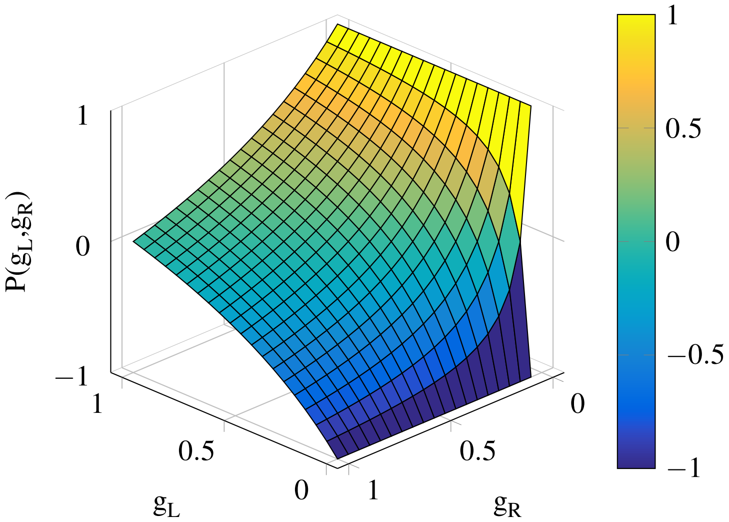


Figure 4.38: Pan position as function of left and right gains. Note that the function is not defined when both gains are set to zero.

As a mix is generated by choosing a pair of points in the mix-space, the pan positions generated by each pair of points can be obtained. A boxplot of all possible panning vectors is displayed in Fig. 4.37d, showing that all tracks have equal distributions of pan position, distributed about the centre position. Using this method, it would however be possible to achieve a panning vector of (0, -0.707 -0.707) as the L_2 norm (the "radius") of that vector is equal to 1. Increasing this to $\sqrt{2}$ would result in a panning vector of (0, 1, 1), in a similar fashion to mix C' in Fig. 4.36, but if *exact* pan positions are required then these methods may not be suitable: they only show relative pan positions. As such, there would need to be separate hyperspheres for each of the two reproduction channels, which could have different radii. The ratio of radii would be required to recreate a desired stereo mix using exact pan positions.

Ultimately, one must consider what panning operations are robust to “mixing”. For gains, one can consider scaler multiplication of all track gains an operation which does not change the mix, according to definition 5. With panning, it is not so clear. As an example, consider two tracks, one panned hard left and the other hard right. If they are swapped is the result the same mix, from a panning perspective? If the width of the panning is reduced, is *this* result the same mix?

4.5 Mix-space experiment 2 — Stereo w/ EQ

As adjusting track gains alone, and only having four tracks, are both highly simplified mixing scenarios, an experiment was devised in which participants could use panning, EQ and gain controls, of eight tracks. The experimental set-up of this experiment, which took place in April 2016, was identical to the earlier, mono experiment in terms of test location and audio reproduction (refer back to Fig. 4.11). The GUI for this experiment is displayed in Fig. 4.39. Each section of the GUI is subsequently described.

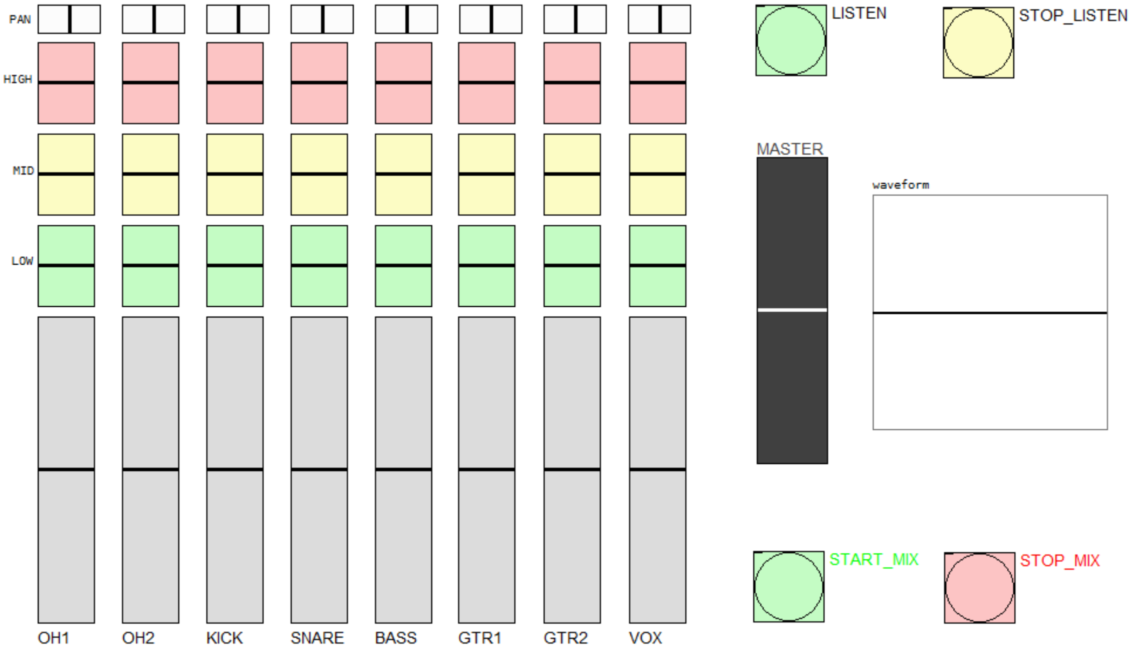


Figure 4.39: GUI used in mix-space test (stereo w/EQ). For this experiment, the initial listening phase was controlled by the experimenter rather than the participant.

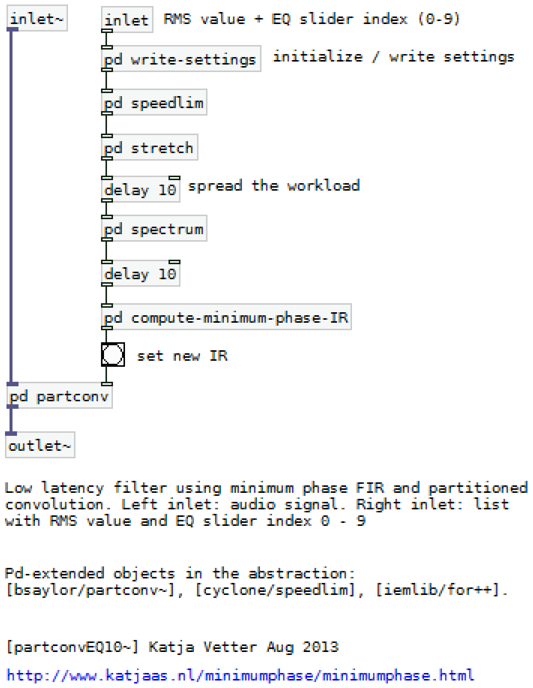
Again, the mixer was implemented using Pure Data, using a similar patch. Each audio file is first processed by EQ, the overall track gain is adjusted according to the GUI fader and the audio is placed in the stereo image according to the chosen pan position. A limiter was placed at the end of the signal chain, before the `dac` object, in order to prevent clipping. Care was taken to ensure the system had enough headroom such that the limiter was rarely engaged.

An equal power panning law was used to position each signal in the mix. The equaliser used in this patch is based on the patch shown in Fig. 4.40. This patch implements a low latency filter using minimum-phase FIR and partitioned convolution [16]. The patch was used to create a three-band EQ and the response was made to match that shown in Fig. 4.34.

4.5.1 Set-up

The experimental set-up is identical to that described earlier (see § 4.3.4), only this time the left and right loudspeakers were used while the centre loudspeaker was unused. Again, this allowed visual and acoustic continuity between the two experiments. In contrast to the mono experiment, for the listening phase of each trial the GUI was hidden from participants. This was done as some

[16] `http://www.katjaas.nl/minimumphase/minimumphase.html` — this page describes the filter design process, and cites the work of Damera-Venkata et al. [162].



Low latency filter using minimum phase FIR and partitioned convolution. Left inlet: audio signal. Right inlet: list with RMS value and EQ slider index 0 - 9

Pd-extended objects in the abstraction: [bsaylor/partconv~], [cyclone/speedlim], [iemlib/for++].

[partconvEQ10~] Katja Vetter Aug 2013

http://www.katjaas.nl/minimumphase/minimumphase.html

Figure 4.40: Patch used for EQ

participants in the previous experiments used the ‘start listen’ and ‘stop listen’ controls incorrectly. For the stereo experiment, these controls were hidden on a second monitor and controlled by the experimenter. Once the listening phase was completed, the GUI was revealed to the participant, without the ‘start listen’ and ‘stop listen’ controls.

4.5.2 Audio stimuli

Four songs were used. For each, only eight specific tracks were used, corresponding to the following instruments: drum overheads (split into two mono tracks), kick drum, snare drum, bass guitar, guitar 1, guitar 2 and lead vocals. Since the source hypothesis no longer needed to be tested, each song was tested one time only. This allowed the addition of a fourth song within the same approximate test duration. The songs used were as follows: “Borrowed Heart”, “Fighting, We Were”[17], “New Skin”[18] and “Sister Cities”. All of these sessions had more than two guitar tracks recorded. The choice of what should be “guitar 1” and “guitar 2” was based on choosing two similar tracks, e.g. two recordings of the same part, with different performers/guitars/amplifiers etc. In the case of “Borrowed Heart”, the guitar 2 track was in fact a banjo recording, as it was that track which best matched guitar 1, an acoustic guitar.

The audio from these sessions was as processed as little as possible, since participants would have control of a basic equaliser. In order to reduce the many raw tracks to a usable 8-track session it was necessary to merge some of these raw tracks, for example, bouncing the multiple snare drum channels to one, or combining overheads with room microphones and even close-mic’ed tom channels. These bounces and submixes were created by the author, using MATLAB rather than a DAW, for continuity and repeatability.

[17] https://bookclub.bandcamp.com/track/fighting-we-were

[18] https://weathervanemusic.org/shakingthrough/torres

4.5.3 Test panel

Fifteen participants (5 female, 10 male) were recruited for the stereo experiment. Eight of these participants had previously taken part in either one of the mono experiments. The median age was 25 years, in the range of 18 to 42. Again, none reported hearing difficulties.

4.5.4 Results

In the mono experiments (see § 4.3.5) the fader values were stored at the same sampling rate as the audio and later downsampled to 10 Hz. Since the number of controls in the stereo experiment was 10 times greater (and it was expected participants would take longer to mix each session on account of the increased level of control) these control values were saved directly to a .CSV file at a rate of sampling rate of 10 Hz. Figure 4.41 reveals that the time taken to complete each song did not appear to differ. There were, however, differences between participants. For example, participants 3, 11 and 13 took, on average, much less time to mix than most other participants.

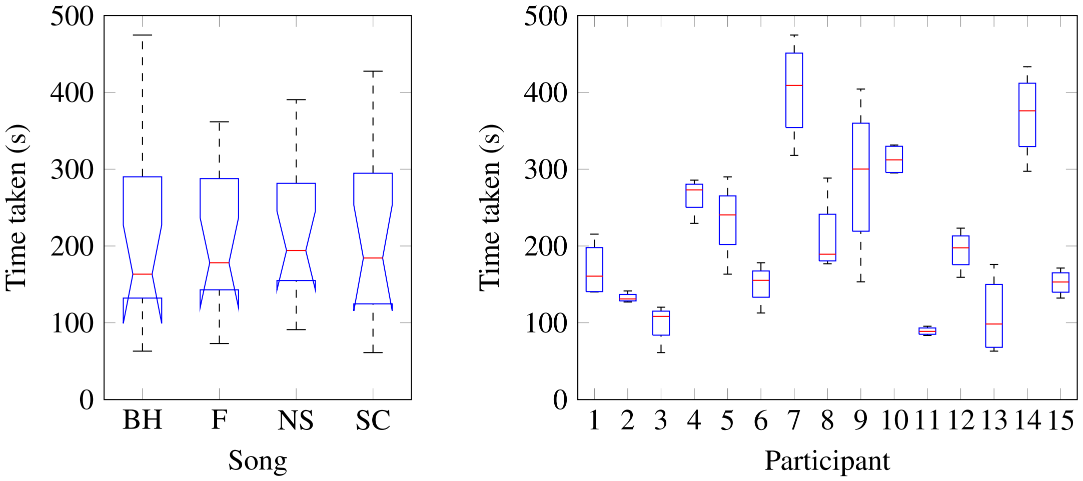


Figure 4.41: Boxplot showing time taken to complete mixes in stereo mix-space experiment. The distribution is similar for each song, while it is clear that different participants spend various amounts of time on the test.

4.5.4.1 Instrument levels

Figure 4.42 shows the distribution of relative loudness levels of each instrument within the final mix. Note that the level shown ignores both EQ and panning, and was simply determined using a method identical to the mono experiments with four tracks. One notable difference is that the level of vocals was lower. This is believed to be due to the spatial unmasking that takes place when the competing sounds (mostly guitars) can now be panned away from the vocals — vocals no longer need to be presented at such a high level. These instrument levels, shown in Table 4.4 can be compared against the results from the 4-track, mono experiment, as displayed in Fig. 4.13. The exact values cannot be compared, since the number of tracks is different in both experiments, however, by grouping the 8 tracks into the same 4 as the mono experiment (vox, guitars, bass and drums), comparisons can be made. When the levels of all four drums tracks are summed for each participant, and likewise for the two guitars, the results can be more easily compared to the 4-track mono experiment. This comparison is summarised in Table 4.3. In order to incorporate the loudness changes that are caused by the use of EQ, the following process was implemented.

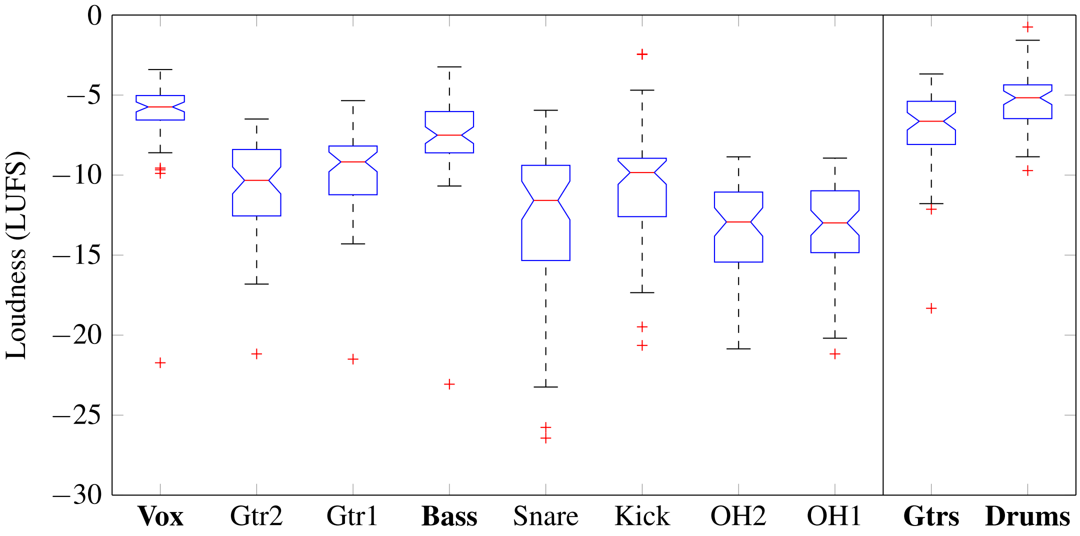


Figure 4.42: Levels of instruments, *ignoring* EQ. The four variables marked in **bold** are comparable to the four tracks shown in Fig. 4.13.

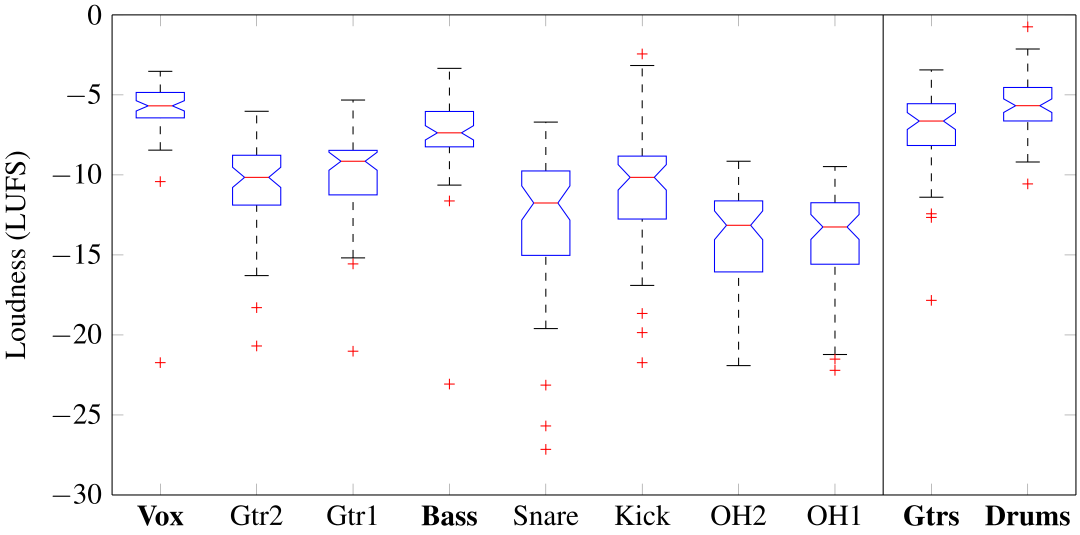


Figure 4.43: Levels of instruments, *with* EQ considered. The four variables marked in **bold** are comparable to the four tracks shown in Fig. 4.13.

Instrument	Median Level (LUFS)		
	stereo	*mono*	*difference*
DRUMS	-5.17	-7.96	+2.79
BASS	-7.51	-11.05	+3.48
GTRS	-6.63	-8.87	+2.24
VOX	-5.74	-2.85	-2.89

Table 4.3: Median values - mono/stereo comparison. Mono results are taken from the LS groups in Fig. 4.13

Instrument	Median Level (LUFS)		
	with EQ	*without EQ*	*difference*
OH1	-13.2535	-12.9870	- 0.2665
OH2	-13.1480	-12.9295	- 0.2185
KICK	-10.1527	-9.8416	- 0.3111
SNARE	-11.7547	-11.5837	- 0.1710
BASS	-7.3708	-7.5059	+ 0.1351
GTR1	-9.1465	-9.1760	+ 0.0295
GTR2	-10.1537	-10.3300	+ 0.1763
VOX	-5.6845	-5.7406	+ 0.0561

Table 4.4: Instrument levels, with and without EQ. The differences are small (< 0.33 LU)

- For each track, calculate inter-band Φ_{eq} and r_{eq} first
- then adjust track gains according to r_{eq}, giving $r_{\text{eq}} \times G$
- then work out inter-channel Φ_{tracks}, using $r_{\text{eq}} \times G$
- Final mixes are based on the values of Φ_{tracks} at t_{max}, the time when mixing stopped. The gains of the final mixes are referred to as $\mathbf{g}_{\text{final}}$.
- Final pan positions, $\mathbf{g}_{\text{final}}$ were also found. From this $\mathbf{g}_{\text{final,L}}$ and $\mathbf{g}_{\text{final,R}}$ were determined using Eqn. 2.1c.

The following is a brief summary of results, as displayed in Table 4.4. Typically, both overhead tracks are set to comparable loudness levels. This indicates that these two tracks are associated with one another and treated similarly by participants, possibly due to the high correlation between the two signals. The result is a balanced stereo image when the tracks are panned. The two guitars are set to similar levels. Again, these two signals are highly correlated, as both were (in 3 out of 4 songs) different recordings of the same musical part. The median level of the kick drum is 1.6 LU greater than the snare drum. However, as the snare drum can be heard in the overhead tracks (as too can the kick drum but to a lesser extent), the perceived loudness of the snare drum is based on the loudness of the overheads tracks in addition to the close-mic'ed signal. The vocals are quieter in the mix when mixing in stereo, due to spatial unmasking. Relatively speaking, all other instruments are louder, as seen in the boxplot.

4.5.4.2 EQ

The use of equalisation can be observed using a bagplot in the "tone-space". Extending the familiar concept of the univariate boxplot, the bagplot can be used for bivariate data (and also for multivariate data). The interested reader is referred to Rousseeuw et al. [163] for the precise details of the bagplot production. In summary, a 'bag' is drawn which contains 50% of the data points, a 'fence' (which has three times the area of the bag) separates inliers from outliers and a 'loop' region which contains points inside the fence but outside of the bag. The Tukey median is the point at which the minimum halfspace depth is found, analogous to a univariate median. It is the same point in each of the plots below, the starting point where $g = [1, 1, 1]$, and this simply shows that the space is appropriately normalised as desired.

In Fig. 4.44 and 4.45, each of these plots shows the distribution of EQ settings applied to each track. As there were 15 participants and four songs, there are 60 examples of EQ use for each of the eight different instrument tracks. The following is a brief interpretation of each plot, where the skewness of the distribution generally dictates the typical EQ applied. Plots were generated using the LIBRA matlab library [164].

Vocals (see Fig. 4.44a) The bag and fence both extend rather evenly in all directions. This suggests that the EQ applied to vocals was varied and there was no consensus as to a typical vocal EQ. It is worth noting that there were two male voices and two female voices and, with this variation, a lack of consensus is perhaps not surprising.

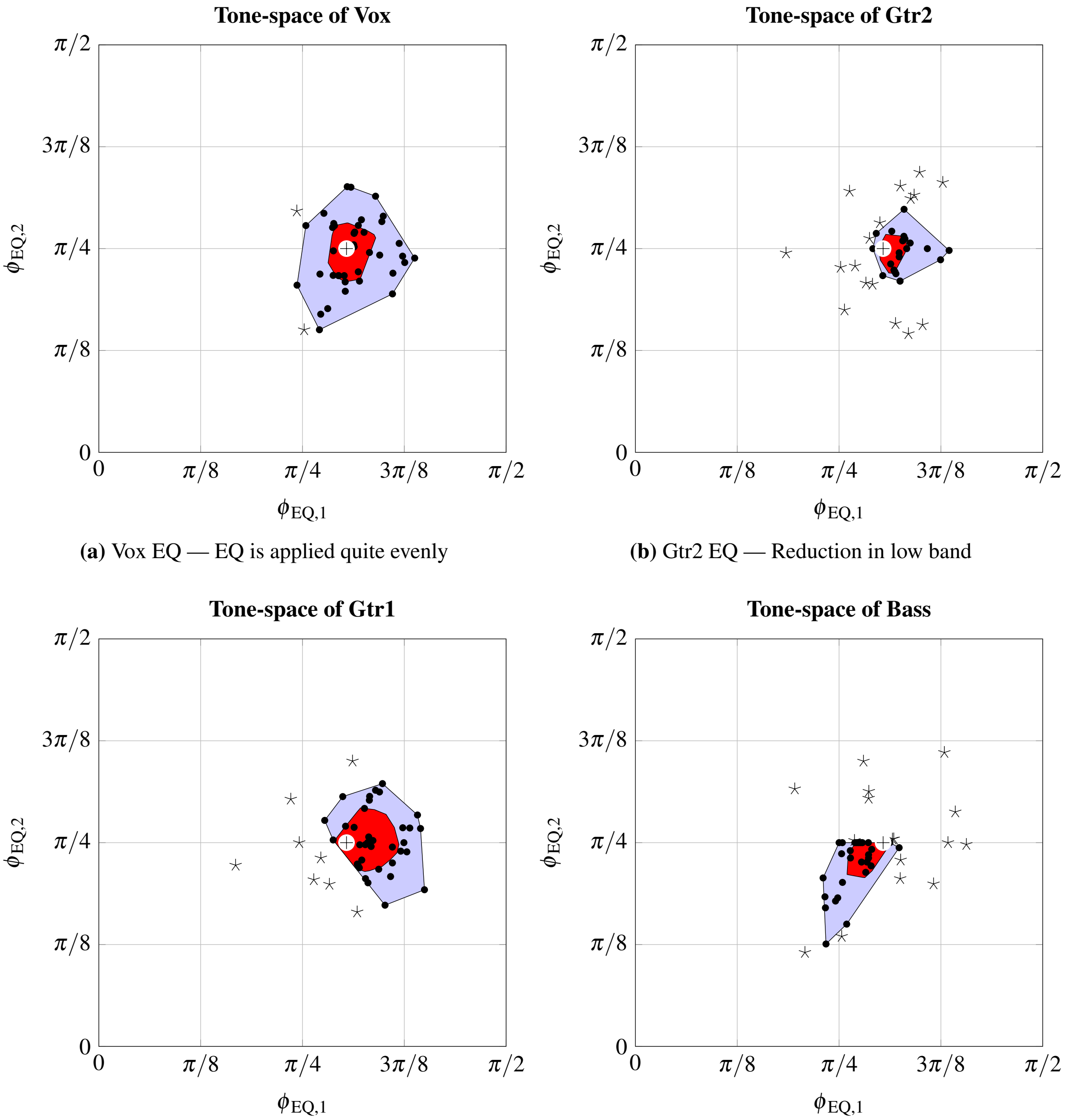
Guitars (see Figs. 4.44c and 4.44b) Both guitar 1 and guitar 2 display similar EQ use, i.e., a reduction in the low band, but alterations to the EQ of guitar 1 were, on average, more varied, while adjustments to guitar 2 were typically small reductions in the low band. Both plots show a relatively large number of outliers when compared to vocal EQ.

Bass (see Fig. 4.44d) The use of EQ on bass was typically to reduce the gain of the high band relative to the middle band, while boosting the low band, although there are notable outliers. However, in interpreting this result it is important to consider the spectrum of the instrument — most of the spectral energy would be contained in the lower two bands. There are outliers where the high band has been boosted but perhaps this did not produce enough of an audible difference for the participant to observe it as unpleasant and turn it back down.

Snare drum (see Fig. 4.45a) The snare drum EQ can be characterised as a less bass and more treble. As with the kick drum, boosting of the mid band was relatively rare and is indicated by outliers.

Kick drum (see Fig. 4.45b) Here, the bag and fence lean to the left of the graph, indicating higher bass, but also to the top of the graph, indicating greater treble. This can also be the result of reducing the middle band, and low-mid cuts are often used when equalising a kick drum.

Overheads (see Figs. 4.45d and Fig. 4.45c) Both overhead tracks showed similar use of EQ, namely a reduction the in low band. The shape of the bags are similar in both plots and the pattern of outliers was similar. This suggests that individual participants produced matching EQ for the two tracks.



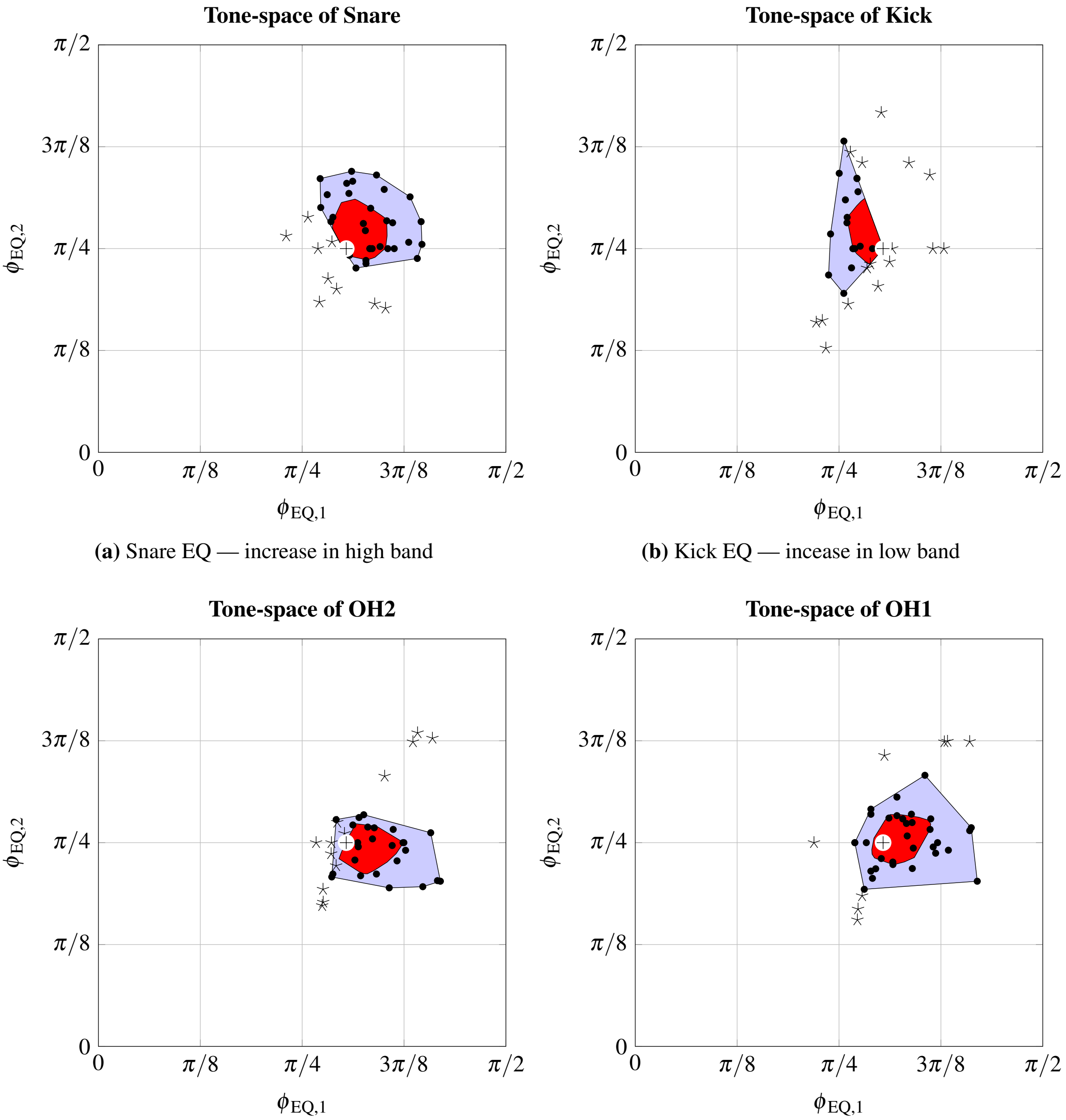
(a) Vox EQ — EQ is applied quite evenly

(b) Gtr2 EQ — Reduction in low band

(c) Gtr1 EQ — reduction in low band

(d) Bass EQ — reduction in high band, increase in low band

Figure 4.44: Tone-space of vocals, guitars and bass



(a) Snare EQ — increase in high band

(b) Kick EQ — incease in low band

(c) OH2 EQ — decrease in low band

(d) OH1 EQ — quite even but with general decrease in low band

Figure 4.45: Tone-space of drum tracks

4.5.4.3 Pan positions

Figure 4.46 shows the distributions of the final pan positions of each instrument, for each of the songs. It is immediately clear that the vocals are never panned far to either side. To further investigate the nature of the panning distribution, the density functions were estimated using KDE. The resulting estimations are displayed in Figures 4.47a to 4.47d. In each of Figures 4.47a to 4.47d, the kernel width used is a fraction of the default value, h, which is considered optimal for normal distributions. As there was no prior assumption of normality, this narrower kernel width results in more modal values being revealed. The width used was 1/5$^{\text{th}}$ of the default.

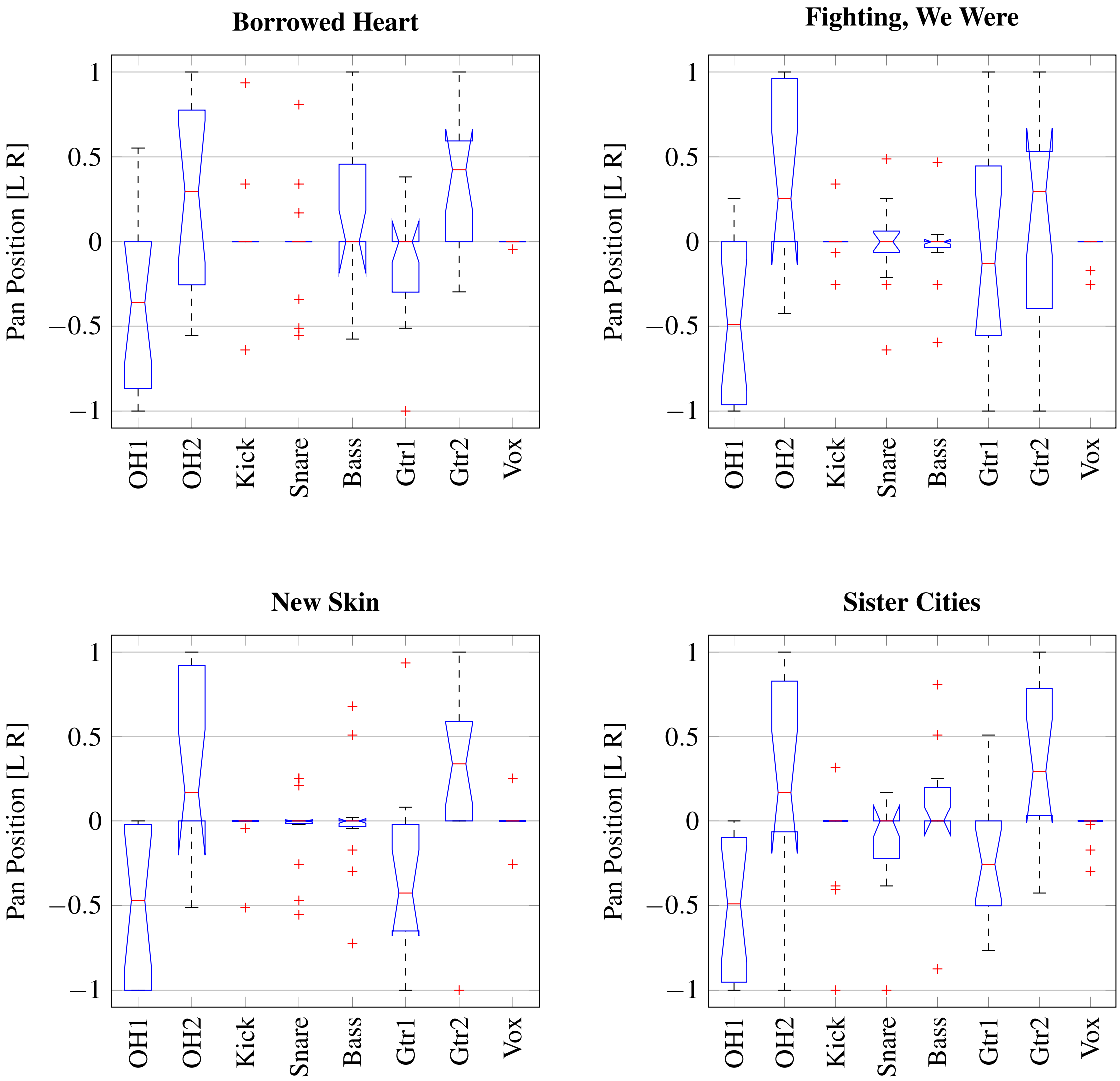
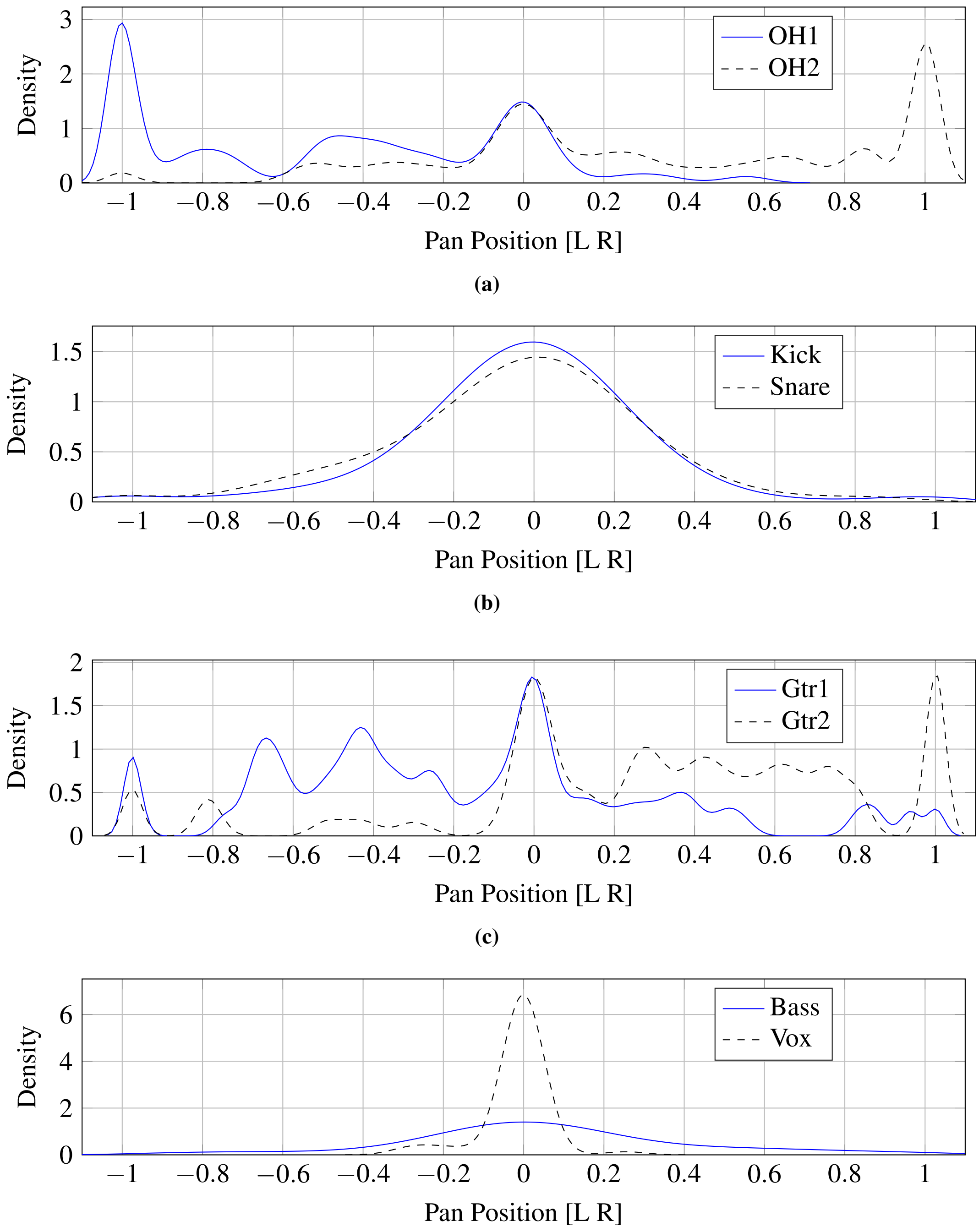


Figure 4.46: Distribution of pan positions in final mixes of each song. Any instance where vocals were panned off-centre is marked as an outlier.

In the case of kick drum, snare drum, bass guitar and vocals, the resulting density estimate is not dissimilar to a normal distribution. Guitar panning decisions are the most multi-modal: $-1, 0$ and 1 are commonly occurring values but there are also a number of modal values in-between. This shows that guitar panning is highly subjective and depends on the specific song (see Fig. 4.46). For the song “Borrowed Heart” the median pan position of Gtr1 is central, however, as mentioned



(a)

(b)

(c)

(d)

Figure 4.47: KDE of pan positions for each track. Kernel width = $h/5$, in order to reveal multiple modes, where they exist. Guitar and drum overhead panning functions were multi-modal, while kick, snare, bass and vocals followed more simplistic distributions.

in section § 4.5.2, Gtr1 and Gtr2 were most dissimilar in this song. This may explain why, in Fig. 4.47b, the density function indicates that Gtr2 was hard-panned more often than Gtr1, since Gtr2 was a banjo in the song ‘Borrowed Heart’, and it was likely to be hard-panned while the acoustic guitar remained close to the centre position, as indicated by the median positions in Fig. 4.46.

There was an effect of the track ordering on the pan positions, in the case of drum overheads and guitars. For both of these track groups, individual tracks were typically panned according to

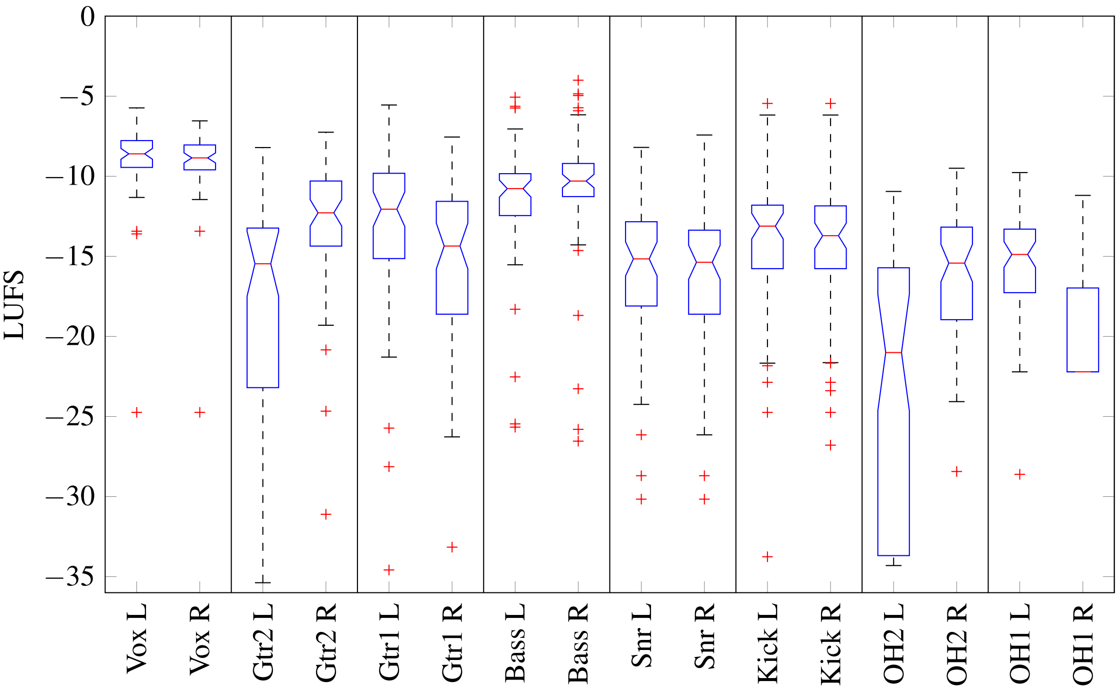


Figure 4.48: Distribution of final mix gains, taking into account pan positions.

the left-to-right positions of the tracks in the GUI. OH1 was panned left and OH2 was panned right, in most cases. The effect is less for the guitars but, in many cases, Gtr1 was panned left-of-centre and Gtr2 was panned right-of-centre.

Note that the position alone can only reveal so much information about the mix, as a track can be panned but at such a low volume as to not be heard. The data showed that one participant panned a single of the overheads far to one side but then greatly reduced the volume, giving a sense of space, without resorting to the conventional technique of hard panning both tracks. It is important to look at the final mix levels of both left and right channels, not just the combined sum. This is shown in Fig. 4.48, which is a recreation of Fig. 4.43 but where left and right gains are determined from the final pan positions, using Eqn. 2.1c.

4.6 Discussion

Since these experiments gathered data for only five songs, the results should be considered as specific rather than general. It is not known at this time how many songs would need to be studied to be able to generalise to mixing as a whole, however, these five songs are considered to be typical within pop/rock styles, due to their conventional instrumentation.

4.6.1 Effect of source position

The final mixes created depended on the initial mix presented, as when beginning with source **A** or **B** the final mixes are typically closer to this position (see Figs. 4.28, 4.30 and 4.32). This may be an example of an anchoring effect, in which the initially presented stimulus biases an individuals perception of the alternatives. A literature review of this effect is provided by Furnham and Boo [165]. This suggests that music mixing is influenced by the rough mix that is first presented. In mixing experiments care should be taken in choosing the initial conditions. Previous work had used randomised initial conditions [166], although this does make comparison difficult when one is interested in the precise mixing process, as in this chapter. This effect may also have implications in subjective testing of alternate mixes, as that which is presented first may be favoured, or those similar to that which is presented first. Subjective evaluation of alternative mixes is one of the main themes of this thesis and is discussed further in Chapters 6, 7 and 9.

4.6.2 Differences due to reproduction system

King et al. [46] had previously reported a statistically significant difference between the mixes created on headphones and loudspeakers. In that case, the task involved mixing only in one degree-of-freedom (balancing a lead instrument with a backing track). Additionally, that study reported on the 10 participants who took part in both the loudspeaker and headphone sessions and difference in these participants' mixes. In this chapter, with three degrees-of-freedom (see Figs. 4.12 and 4.13), there was not any statistically significant difference in the levels of the instruments within the mix, when comparing loudspeaker and headphone groups. The small sample size of the headphone group should be noted ($n = 8$), as well as the change in location. However, since the loudspeaker group was tested in a standardised room, this is not thought to be an important factor. It is hypothesised that the main factor explaining the difference between these two studies is the additional complexity and realism of the mixing process presented herein. Additionally, King et al. [46] found the largest inter-group difference for a classical music sample, which is a style of music not represented in this chapter.

4.6.3 Equalisation

The data gathered suggests that, when applying equalisation to a track, it was typical to boost frequencies that are salient in that track, i.e. boosting the low band on bass and kick drum, as shown in Figs. 4.44d and 4.45b. Recall that the crossover to the low band was set to ≈ 180 Hz: this band was generally attenuated for guitar tracks and drum overheads. Vocal EQ application did not appear to follow any particular pattern and has an even spread about the starting position with little observed skewness. These results also suggest that the use of equalisation on the individual channels within a mix does not have a notable effect on the inter-channel loudness differences (see Figs. 4.42 and 4.43). When EQ is applied to a signal, any loudness changes are compensated for by the main track fader.

4.6.4 Panning

Many suggest the panning of low-frequency instruments centrally [53, 61, 167, 168]. This pattern of behaviour was observed in these experiments, as kick drum, bass guitar and snare drum were typically panned close to centre. Panning decisions may have been influenced by track ordering, as similar tracks (drum overheads, two guitars) were typically panned opposite to one another as the tracks were read (the fader to the left was panned left and the fader to the right was panned right). No participant defied this convention (by panning the left track right and the right track left). This indicates the importance of GUI elements on music mixing, as in order to pan a pair of similar tracks far apart, their panning faders were moved to a greater visual displacement. The influence of visual information on music mixing is a topic of recent research, for both software [169] and hardware [170] user interfaces. There is evidence of an interaction between panning on level. Panning the guitars far from the centre position, while the vocals remain in the centre, results in a spatial unmasking effect. Consequently, the vocals do not need to be set so loud in order to compete with the guitars. The reduction in vocal level in Fig. 4.13 compared to Fig 4.42 illustrates this.

4.6.5 Importance of vocals

In both mono and stereo experiments, with 4-tracks or 8-tracks, vocals were typically set at the loudest level of all instruments. Additionally, the variance in the panning of vocals was smaller than any other track. Participants chose to place the vocals near the centre of the stereo image. These results highlight the importance of vocals within popular music. The spoken voice has great communicative power, which can be modified by singing. The recorded singing voice therefore has great affective potential and this can be exploited in the mixing process [171].

4.7 Chapter summary

The work in this chapter introduced the concept of the mix-space and a formulation for track gains, equalisation and panning. The formulation is based on representing the normalised track gains, band gains or pan positions, using hyperspherical coordinates. This parameter space contains all of the mixes that could be created with these tools and forms the basis for the efficient analysis of mixes. In this chapter, mixes were created by test participants in the conventional manner: with individual track faders for gain, 3-band EQ and panning. These mixes were then converted to the mix-domain for comparative analysis. It is perhaps a more simple task to directly generate points in this domain. This topic is explored in Chapter 5 as a means of creating random mixes for Monte Carlo simulation of music mixing and in Chapter 8 as a basis for automated and semi-automated music mixing. There is room for further work. The EQ analysis ("tone-space") was generated based on a 3-band volume adjustment. While this is generalisable to any number of bands, further work would be to incorporate more conventional EQ structures, such as parametric EQ. As illustrated in Fig. 4.6 and 4.7, most of this chapter considers a map of the mix-space, rather than the mix-space itself, i.e., the mix-space is a hypersphere in a gain-space but this chapter creates a Euclidean space from the angular components of the hyperspherical coordinates. It is possible to solve these problems directly on the sphere but this would increase the number of dimensions by 1 and circular statistics would be used in place of linear statistics. Some of these issues are explored in greater detail in Chapter 5 and 8.

5

Analysis of randomly-generated mixes

The previous chapter described a series of experiments in which participants used traditional mixing interfaces to generate mixes. These were constrained in such a way that their mixes could be then transformed into a simple mix-space, so that they could be compared to one another. Could mixes not simply be generated in the mix-space, directly? It would be advantageous to do so, as asking test participants to generate data is time-consuming and would be unlikely to create a large enough dataset for a robust statistical analysis. The ability to quickly generate a large set of mixes, covering the whole range of mixes that it is possible to make, has a number of uses.

a) Typically, feature-extraction is performed on only one mix of a given song, since only one mix exists. Having a set of alternate mixes for each song allows for a more in-depth testing of the robustness of a feature-extraction algorithm. Rather than gathering a large number of real mixes, which is not always possible, the distribution of features within mixes of a song can be estimated on an artificial dataset of random mixes.

b) Creating a population of mixes for use in optimisation (see Chapter 8).

While Chapter 6 discusses the variation in mixes created by real mix-engineers, a highly informative insight into the process of mix-engineering, it is also necessary to understand the baseline conditions to which these real distributions can be compared. To achieve this, the work presented in this chapter uses randomly generated mixes. These will be compared to the real-world mixes in Chapter 6. The research questions pertaining to this chapter are as follows.

RQ.10 Do mix engineers, collectively, produce mixes with feature distributions similar to randomly generated mixes? If not, how do mixes by real engineers differ from the randomly generated mixes?

RQ.11 Can randomly generated mixes be used to help test the performance and accuracy of feature extraction algorithms, such as onset detection and tempo estimation?

5.1 Generating randomised track gains

As described in § 4.1, for a given n tracks, all the unique mixes exist on a hypersphere in $\mathbb{R}^n$, i.e. an $(n-1)$-sphere. To generate random track gains, random points in this space were determined. For n tracks, and m mixes, m points on a unit $(n-1)$-sphere (denoted as $\mathbb{S}^{n-1}$) were generated. The n tracks were first normalised according to perceived loudness, as defined in BS.1770-3 [32] and modified by Pestana et al. [158]. A number of methods can be used to generate a distribution of mixes. Two such methods are detailed here.

5.1.1 Method 1: uniform mixes

An easy way to pick random points on a hypersphere of arbitrary dimension is to generate n Gaussian random variables $x_1,x_2,...,x_n$. Then the distribution of the vectors **g**, as defined by Equation 5.1, is uniform over the surface $\mathbb{S}^{n-1}$ [172, 173].

$$\mathbf{g} = \frac{1}{\sqrt{x_1^2+x_2^2+\ldots+x_n^2}} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \tag{5.1}$$

For sufficiently large number of points m, this method will return virtually all possible mixes of the n tracks. However, are uniformly generated mixes representative of real mixes? It was hypothesised that the generation of uniformly distributed mixes would likely produce many mixes that would not realistically be created by real mixers (see Chapter 6). As a consequence, the value of m would have to be very large in order to be comparable to the number of real mixes listed in Table 6.1 and constraints would need to be implemented in order to ensure that all instruments are presented with sufficient gain as to be audible.

5.1.2 Method 2: mixes close to arbitrary point

There are advantages to generating track gains according to some parametric distribution. For example, the value of m can be lower, greatly reducing the computation time required for feature-extraction and analysis. This method requires explicit parameters to be chosen. From § 2.2, assuming that the better mixes are *generally* the ones where the tracks are *roughly* equal in perceived loudness, this method can be used to generate mixes distributed about the equal-loudness mix. The equal-loudness mix is determined as follows. When the gains of all n tracks are equal, what **g** gives a point on $\mathbb{S}^{n-1}$, i.e. where the L_2 norm of **g** is equal to 1?

$$r = 1 = |\mathbf{g}| \tag{5.2a}$$

$$1 = \sqrt{\sum_{i=1}^{n} g_i^2} \tag{5.2b}$$

$$1^2 = \sum_{i=1}^{n} g_i^2 \tag{5.2c}$$

$$1 = ng^2 \tag{5.2d}$$

$$n^{-2} = g \tag{5.2e}$$

For example, when $n = 16, g = 0.25$. Applying this linear gain g to all n loudness normalised tracks would result in the equal-loudness mix, where the L_2 norm is equal to 1.

In selecting a suitable parametric distribution it is important to note that linear distributions, such as the normal distribution, are not appropriate as the domain in question is not linear but a spherical surface. Recall that a linear domain extends over the range $[-\infty + \infty]$, while a circular domain is wrapped over a smaller range such as $[0, 2\pi)$. The statistics of such distributions are described by a number of equivalent terms in the literature, such as circular, spherical or directional statistics. In order to generate random points close to the desired position on the $(n-1)$-sphere, points are generated from a von-Mises-Fisher distribution (vMF). The probability density function of the vMF distribution for a random n-dimensional unit vector $\mathbf{x}$ is given by

$$f_n(\mathbf{x}; \mu, \kappa) = C_n(\kappa) e^{\kappa \mu^T \mathbf{x}} \tag{5.3}$$

where $\kappa \geq 0, ||\mu|| = 1, n \geq 2$ and the normalisation constant $C_n(\kappa)$ is given by

$$C_n(\kappa) = \frac{\kappa^{n/2-1}}{(2\pi)^{n/2} I_{n/2-1}(\kappa)} \tag{5.4}$$

Here I_v is the modified Bessel function of the first kind at order v. The parameters μ and κ are called the mean direction and concentration parameter, respectively, and μ^T is the transpose of μ. The greater the value of κ the higher the concentration of the distribution around the mean direction μ. The distribution is unimodal for $\kappa > 0$ and is uniform on the $\mathbb{S}^{n-1}$ for $\kappa = 0$. Further details can be found in Fisher [174] and Mardia and Jupp [175]. To generate points according to a vMF distribution the SphericalDistributionsRand[1] code was used based on the work of Chen et al. [176]. In the context of audio mixes, μ represents the mix about which others are distributed, akin to the mean in a normal distribution. The κ term represents the diversity of mixes generated.

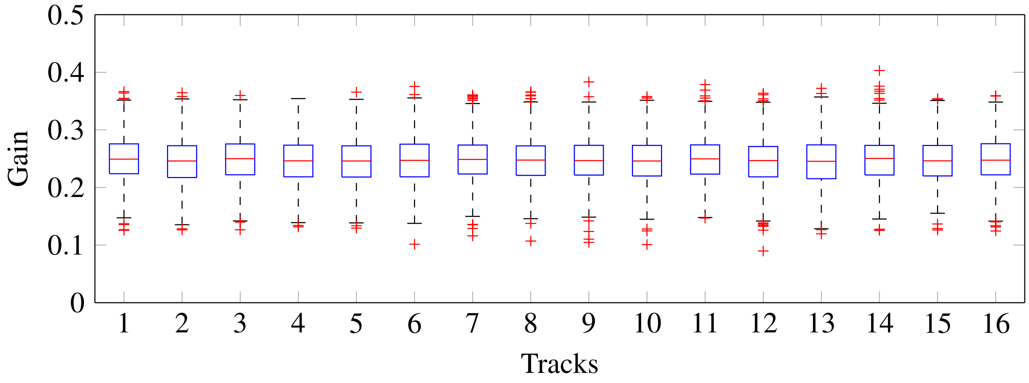


Figure 5.1: Boxplot of gain values for 1,000 mixes of 16 tracks, generated from vMF distribution, designed to produce mixes around the equal-loudness mix.

For $n = 8$ tracks, as in § 4.5, the gains required for the equal-loudness mix are distributed around

[1] https://github.com/yuhuichen1015/SphericalDistributionsRand

the following point, μ. This calculation is based on Equation 5.2.

$$\mu = [0.3536 \quad 0.3536 \quad 0.3536 \quad 0.3536 \quad 0.3536 \quad 0.3536 \quad 0.3536 \quad 0.3536]$$

Previous studies have indicated that, while a good initial guess, presenting each track at equal loudness is not an ideal final mix. As discussed in the literature review (see § 2.2.1) and also was shown in Chapter 4, vocals are often the loudest element in a mix (in particular, see Fig. 4.13 and Table 4.4). To this equal loudness configuration, a vocal boost is added according to p.157 of [61], i.e. a boost of 6.54 dB. A sanity check was performed by audition of mixes generated with this boost and it was decided that, while it may be more than the authors' own taste, such a boost is not unrealistic. This addition of 6.54 dB to the vocal track produces the following vector, where track 8 is vocals.

$$\mu = [0.3536 \quad 0.3536 \quad 0.3536 \quad 0.3536 \quad 0.3536 \quad 0.3536 \quad 0.3536 \quad 0.7507]$$

If the previous vector was, then it is clear that this point is no longer on the unit 7-sphere. To project the point back onto the unit 7-sphere, the vector is divided by it's L_2 norm, resulting in the following.

$$\mu = [0.2948 \quad 0.2948 \quad 0.2948 \quad 0.2948 \quad 0.2948 \quad 0.2948 \quad 0.2948 \quad 0.6259] \tag{5.5}$$

This vector is the new μ on the unit 7-sphere about which mixes will be generated. The result is shown in Figure 5.2. Each mix generated draws a gain value for each track such that the L_2 norm is equal to 1. Note that the median values closely match the vector μ, as expected. Of course, there may not exist a mix which has these median values. This specific value of κ was chosen to avoid generating negative gains, achieved through trial and error. Ignoring phase, a gain of g is perceptually equal to $-g$, meaning that the nature of the distribution would change if negative gains were included. Of course, for a distribution which produces negative gains the absolute value could be taken to avoid inverting the phase of the tracks.

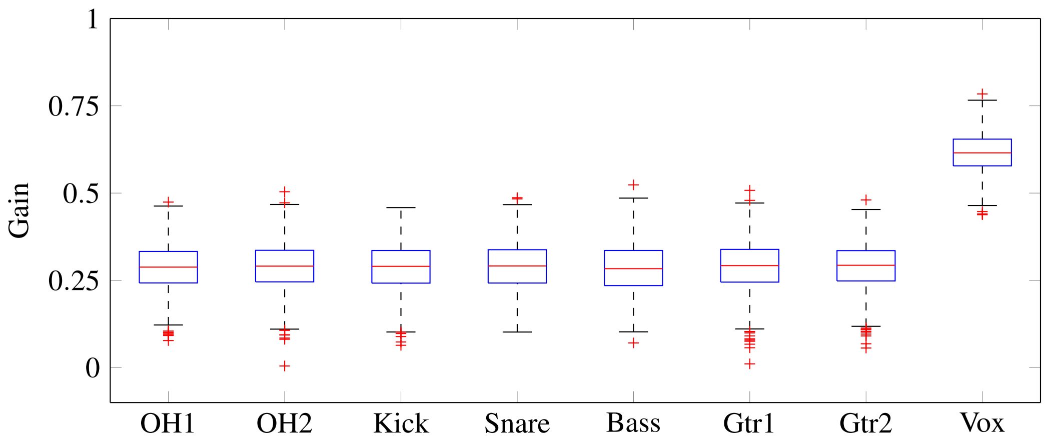


Figure 5.2: Boxplot of gain values for 1,000 mixes, generated from VMF distribution, with 6.54 dB boost to vocals (μ = Eqn. 5.5, $\kappa = 200$)

5.2 Generating randomised equalisation

As the tone-space, introduced in § 4.4.1, only modelled a very simple equalisation method, an alternate method was used for this chapter. In order to achieve as wide a range of equalisation as possible, using as few processing stages as possible, equalisation was applied as follows.

1. Analysis of how an equaliser is used
2. Create a random EQ curve
3. Create a filter from this curve

5.2.1 Principal component analysis of EQ

In order to determine an efficient way to represent the use of a realistic equaliser, the raw data[2] from the Social EQ study [37] was analysed. Briefly described in § 2.2, this raw data consists of 731 terms and a 40-point EQ curve describing the term, from 20 Hz to 19,682 Hz. Of the 731 examples there are 324 unique terms. For the purposes of the study presented herein, it is not particularly important that the EQ examples describe qualitative terms but merely that they are realistic examples of equalisation that is applied to individual instruments. The instruments in question are not always known, since participants had the ability to upload their own sounds, but most are likely to be guitar, piano and drums, as these were the default sounds supplied, according to Cartwright and Pardo [37].

With a 731×40 matrix of data, i.e. 731 observations of 40 variables, PCA was used to determine a set of basis vectors. Since the 40 variables are individual bands of an equaliser, the assumption can easily be made that there is correlation between all variables. This is confirmed by the data shown in Fig. 5.3, where it can be seen that nearby bands are positively correlated and distant bands are negatively correlated. The data was standardised prior to PCA. Since all data was in units of dB, and spanning a similar range, this standardisation may not be critical, but was done to provide consistency with other uses of PCA in the thesis. PCA revealed that the first two dimensions can be approximately described as a spectral tilt (with centre point near 850Hz) and a mid-boost (wide Q, around 1 kHz). These two components account for $\approx 70\%$ of the variance in a 10-band EQ. The Pareto chart is shown in Fig. 5.4. The first six components are required to explain over 95% of the variance. These first six basis functions are shown in Figure 5.5. Since the aim was the represent close to 100% of the variance in a reasonably low number of components, choosing 95% variance as the reason to keep six components is justifiable.

5.2.2 Choose random point in PCA space

For each track, a random EQ filter was generated as follows. A random position in this six-dimensional PCA-space was determined by generating a vector of six Gaussian variables. This resulted in an equalisation curve, by combination of the six basis functions. With units in dB, the mean of the distribution was chosen as 0 and the standard deviation $\sigma = 6$. The greater the value of σ, the greater the variance in gain, resulting in a greater chance of more pronounced and noticeable equalisation. The chosen value was the result of an informal trial in which a number of values were used and the results subjectively compared. A value of $\sigma = 6$ produced a noticeable amount of variation in the tone of individual instruments.

[2] Available from `http://socialeq.org/data/`

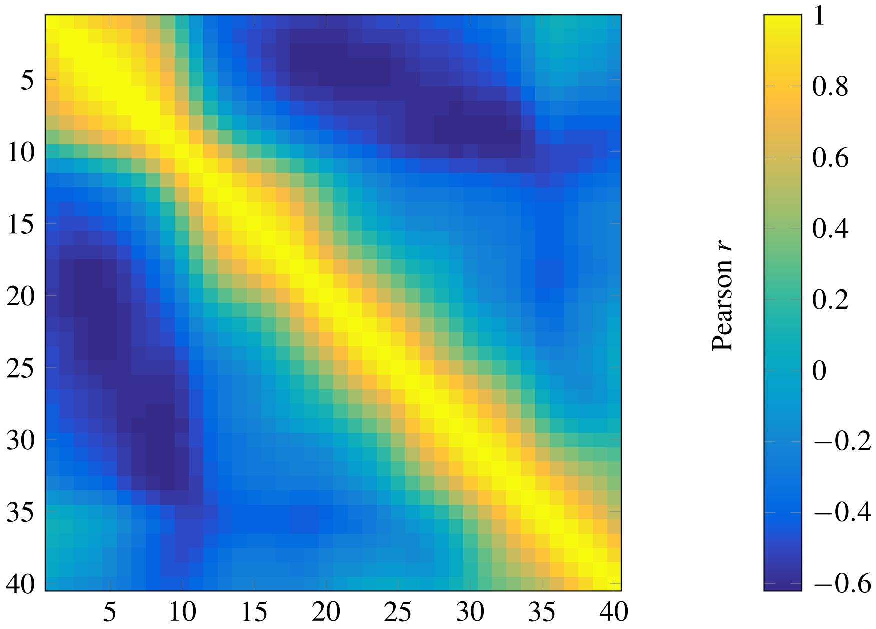


Figure 5.3: Correlation of EQ bands in “Social EQ” raw data

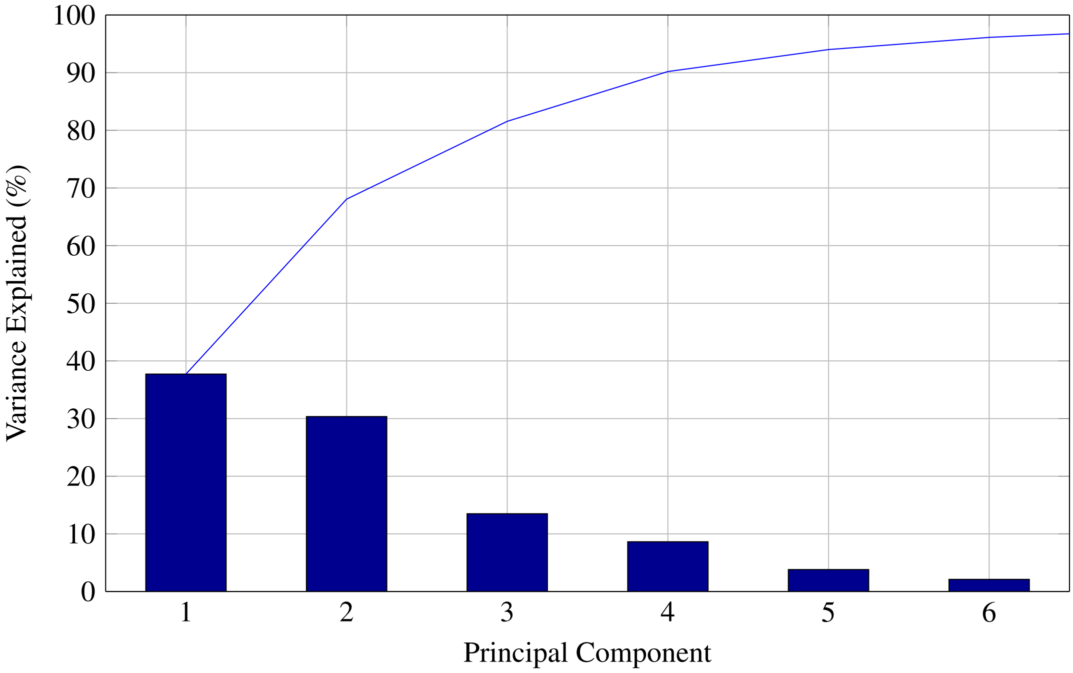
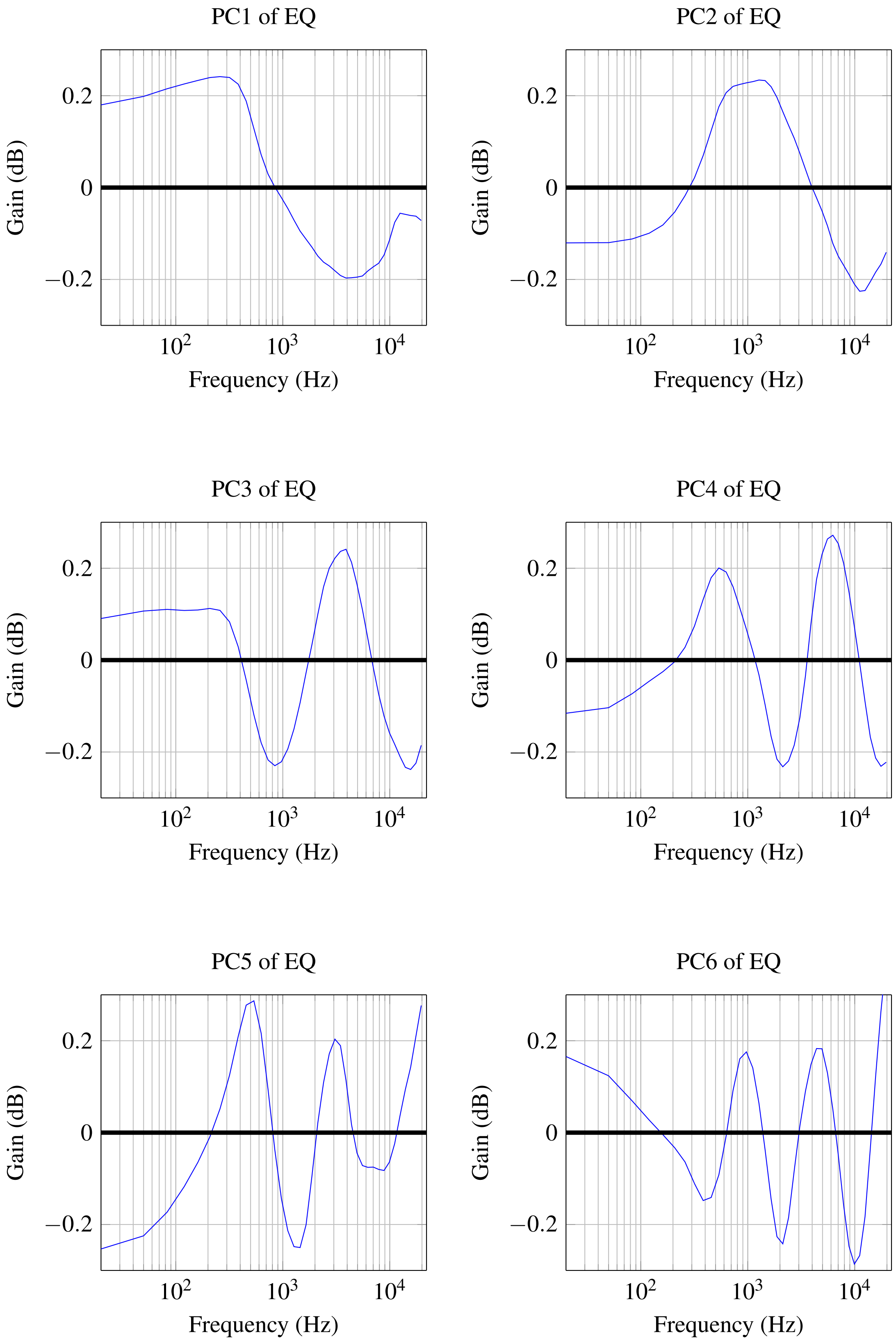


Figure 5.4: Pareto chart for PCA of “Social EQ” raw data (standardised)


PC1 of EQ

PC2 of EQ

PC3 of EQ

PC4 of EQ

PC5 of EQ

PC6 of EQ

Figure 5.5: First six basis functions of "Social EQ" raw data (standardised)

Of course, applying equalisation at random is only an approximation to how equalisation would be used by a human operator. For example, instruments with greater energy at low frequencies may not require any equalisation at the higher frequencies, and vice-versa. The application of random EQ in this case was intended to produce randomised variations in the measured audio signal features, consistent with the variations that would occur when equalisation is applied by a human operator under realistic circumstances.

5.2.3 Approximate EQ with IIR filter

With a randomised EQ curve generated, a filter with this approximate curve shape was determined using the Yule-Walker method [177]. Note that only data up to 19,682 Hz was available from the original study [37], and so the gain at this point is held until f_s to generate the target curve used to generate the filter coefficients. For each track in a mix, such a filter was produced and applied to the audio signal. Examples of these filters are shown in Figure 5.6. In order to generate 1,000 mixes of an 8-track session, 8,000 EQ settings were generated. Figure 5.7 shows the mean and standard deviation of a set of 8,000 such filters. This shows that the mean value is close to zero at all frequencies, as desired. There are a number of reasons why the standard deviation is not equal across all frequencies.

1. The filter produced by the Yule-Walker method is an approximation to the desired filter (see Figure 5.6), and so there is some error.

2. The first six dimensions of the PCA do not explain the entire variance (see Figure 5.4).

3. The participants in the original study [37] did not perceive equalisation equally across all frequencies. While they were not using an equaliser in an explicit sense, simply listening to generated EQ curves, the point is still valid.

The first and second points here are most likely to be trivial while it is expected that the third point contributes most to the result. Since the spectral distribution of music is not flat and given that we perceive different frequencies at different loudness levels, there is little reason to expect EQ usage to be equal across all frequencies.

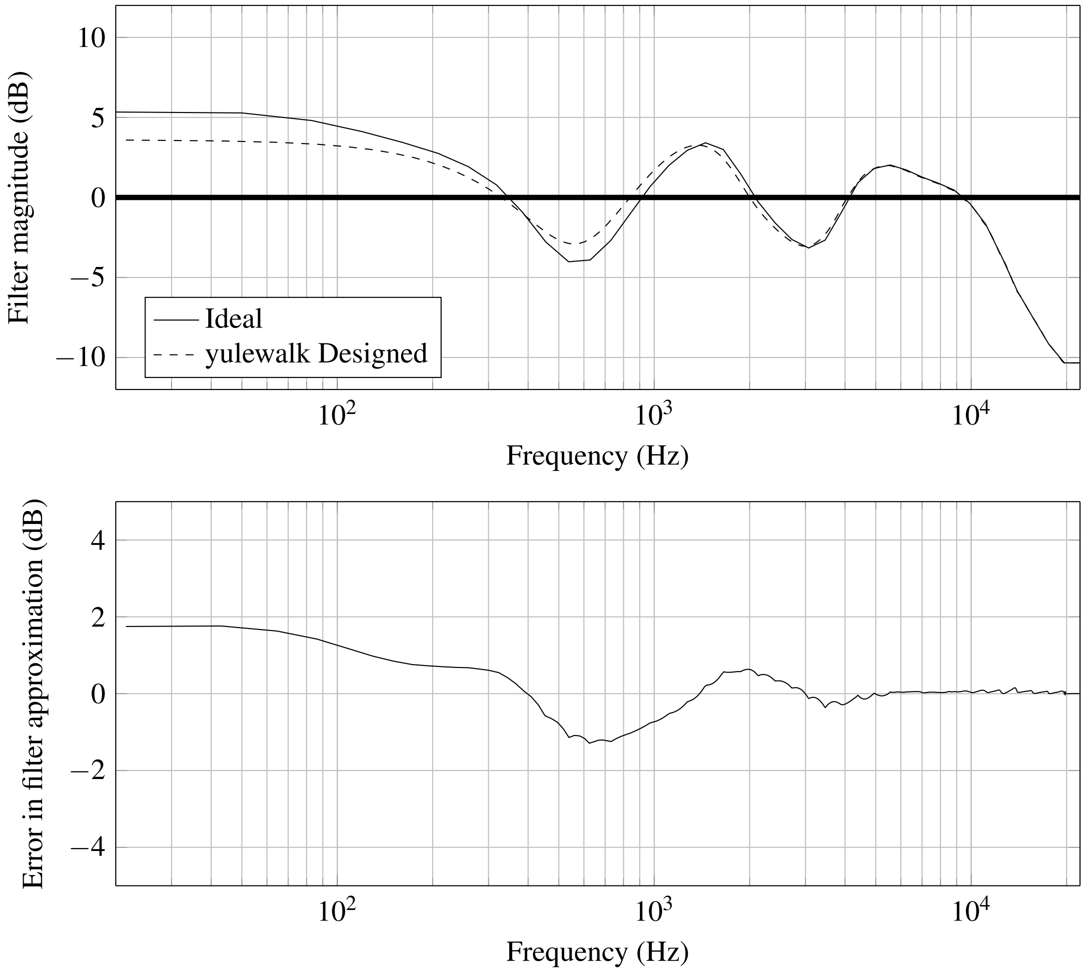


Figure 5.6: Random EQ filter chosen from PCA space

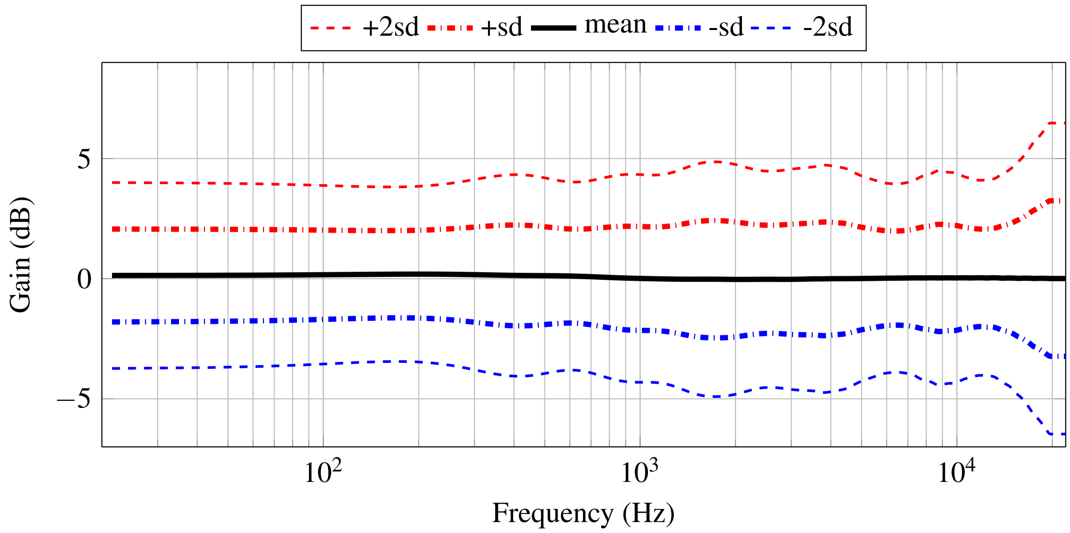


Figure 5.7: Mean and Std. Dev. of 8000 random EQ filters chosen from PCA space

5.3 Analysis of mono mixes

In order to create the random mixes, songs were split into 8-track sessions, featuring the same eight tracks as in § 4.5, namely two drum overheads, kick drum, snare drum, bass guitar, two similar guitars and one track of vocals. In order to achieve this, and for the sake of simplicity, only three songs were used. These were the three songs for which the most real-world mixes were available (see Table 6.1) and could easily be represented in this 8-track form. These three songs were “Burning Bridges”[3], “I’m Alright”[4] and “What I Want”[5]. For each song, the eight tracks were normalised according to loudness [32, 111], just as in § 4.3 and § 4.5.

For each song a set of 1,000 random mixes were generated. The same set of 1,000 gain vectors were used for each song, to set the levels of the eight tracks. These are shown in Fig. 5.2. Similarly, when EQ was applied, the same 1,000 EQ settings were used for each song (see Fig. 5.7). The i^{th} gain vector is paired with the i^{th} set of 8 filters, to generate the i^{th} mix. By generating the settings first and then applying these settings to each song, the 1,000 mixes of each song are comparable to one another, especially given that the tracks are loudness-normalised. A number of signal features were extracted from each set of random mixes. These included features previously used in Chapter 3 as well as additional signal features describing aspects of rhythm not addressed previously. These include the three related concepts of onset detection, tempo estimation and pulse clarity (which has also been referred to as “beat strength”[178]).

Table 5.1: Table of features — random mixes

Name	Description
Loudness	[32]
Spectral Centroid	[114]
LF energy	[107]
Pulse Clarity	[114, 179]
Onset detection	[114]
Tempo	[114]

5.3.1 Loudness/Amplitude

According to Eqn. 5.5, as each mix is on a unit hypersphere, the L_2 norm of the gain vector is equal to 1, for each mix. Theoretically, any variations in perceived loudness are due to differences in the spectral content of each track, limitations in the applicability of the modified BS.1770-3 to narrowband signals [111] (as used in the initial normalisation) or inaccuracy in the application of BS.1770-3 [32] to broadband signals (in the measurement of the mix loudness). The estimated probability density function of loudness values is shown in Fig. 5.8, estimated using KDE. From this result it was possible to confirm that the perceived loudness of all mixes is equal, to a small margin of error. Recall from Chapter 4 that h is the default kernel width, a width that assumes a normal distribution, and $h/3$ is used here to gain greater insight into modal values. The value of $h/3$ was considered as a compromise value that would allow for sufficient detail, based on informal experiments.

[3] http://www.cambridge-mt.com/ms-mtk.htm#DarkRide_BurningBridges

[4] http://www.cambridge-mt.com/ms-mtk.htm#AngelsInAmplifiers_ImAlright

[5] http://www.cambridge-mt.com/ms-mtk.htm#TheBrew_WhatIWant

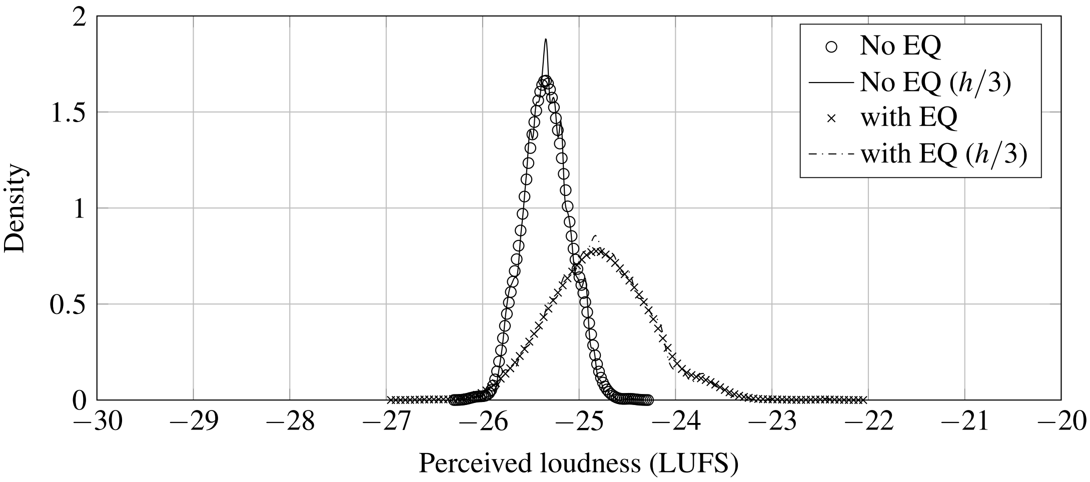


Figure 5.8: KDE of perceived loudness, for 1,000 vMF-distributed mixes of 'Burning Bridges', with vocal boost, before and after random equalisation and two different kernel smoothing values. vMF distribution: μ = Eqn. 5.5, $\kappa = 200$

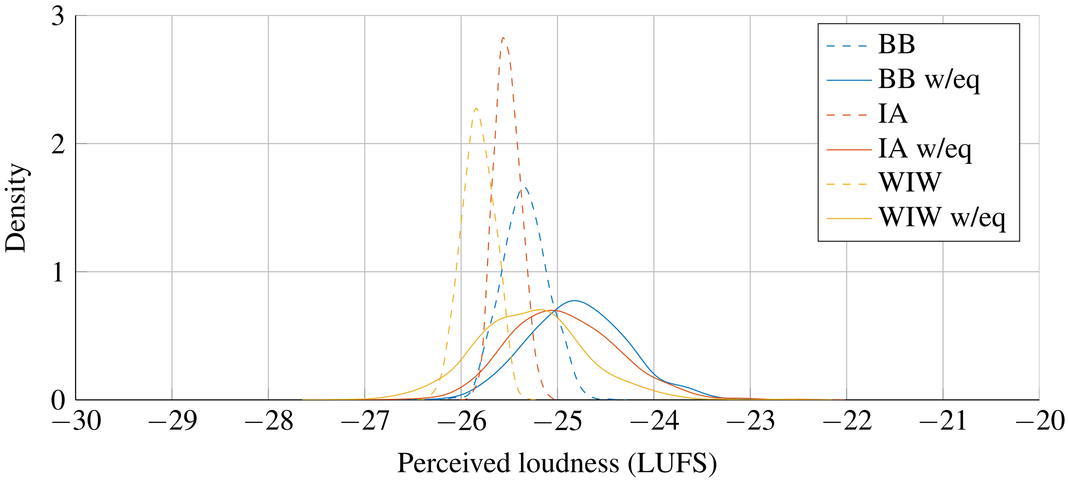


Figure 5.9: KDE of perceived loudness, for 1,000 vMF-distributed mixes of three songs, with vocal boost, before and after random equalisation. Each curve is drawn with the default kernel width. vMF distribution: μ = Eqn. 5.5, $\kappa = 200$

Figure 5.8 also shows that the process of adding random equalisation adds, on average, 0.51 LUFS to the perceived loudness of the mix, for "Burning Bridges". This value was obtained by measuring the difference between the peak of each $h/3$ curve. The variance in loudness values when EQ is added is greater than without EQ. However, since the majority of mixes lie within -25.5 and -24 LUFS, this can still be considered a small variation in perceived loudness. Note that, when using a more narrow kernel width of $h/3$, the presence of modal values is move evident. The overall shape of the curve is however well-estimated using the default kernel width of h.

Some additional insights were obtained by comparing the KDE curves for three different songs. Without EQ being added there are differences between the peak values, although they exist within a small range of less than 0.5 LU (see dashed lines in Fig. 5.9). The peak values for each song are at loudness values of -25.34, -25.56 and -25.83 LU and the variance is low, at 0.59, 0.30 and 0.37 LU respectively. Perceptually, this range of loudness values would be difficult to discriminate. Numerically, these differences are likely to be due to small inaccuracies in the loudness measurement algorithm. For all three songs, Fig. 5.9 indicates that the addition of EQ adds loudness to the mix but also broadens the range of loudness values in the set of 1,000 mixes (refer to solid lines in Fig. 5.9). This is not surprising as the addition of EQ increases the degrees of freedom in the mix. When EQ is added in a random fashion it may boost or attenuate the salient frequencies of a given track, giving rise to changes in perceived loudness.

5.3.2 Spectral features

Making changes to the spectral content of individual instruments is one of the primary tasks of a mix engineer, as identified during the literature review. This has been shown elsewhere in the thesis (see Chapter 6, where "brightness" and "bass" were two of the primary dimensions uncovered). It is therefore expected that alternate mixes vary greatly in terms of spectral characteristics, such as spectral centroid. While this is shown in Fig. 6.5, for real mixes, the extent of the variability in this feature is of interest and can be found in the study of random mixes.

Figure 5.10 shows the results of spectral centroid measurements on the random mixes, with and without the random equalisation being applied, for the song "Burning Bridges". For both cases, two plots are shown: one at the default kernel width and one at 1/3 the default width. The default width of h makes a good estimate and is therefore used in Fig. 5.11. It is clear that a wider range of values is attained after the equalisation has been applied. There is no noticeable difference in the peak value ($\approx$ 4200 Hz), indicating that the random equalisation process is fair, equally likely to raise as to lower the value of the spectral centroid. Figure 5.11 indicates that this effect is also true for the two other songs, as there is little difference in the mean spectral centroid before and after EQ.

While 1,000 mixes were generated, how can one be sure that this was a sufficiently large amount? The effect of sample size of the spectral centroid estimation is shown in Fig. 5.12. Note that when only 100 samples are generated, the modal value is over-estimated. Increasing past 300 samples does not seem to increase the accuracy of the estimation, for this default level of kernel smoothing. Interestingly, this is not much larger than the greatest amount of human mixes obtained for Chapter 6, which is 373, as shown in Table 6.1. Since the distribution does not change noticeably after $N = 300$, it is possible to state that 373 human-made mixes was a significant amount. Most of the songs did not have this many mixes (the mean amount was 150 and the median 127), however, it is also shown that when EQ was added the distribution did not change noticeably after $N = 100$. A sample size of 1,000 is therefore assumed to be an adequately large sample for the purposes of evaluating spectral features in a Monte Carlo simulation of music mixes.

5.3.3 Rhythm

Referring to Chapter 6, since the analysis of real mixes did not include features related to rhythm, a number of these features were included for the study of random mixes. In this analysis, since all of the mixing parameters are known, the variation of the features describing rhythm can be better examined, without confounding factors introduced by the mix engineers.

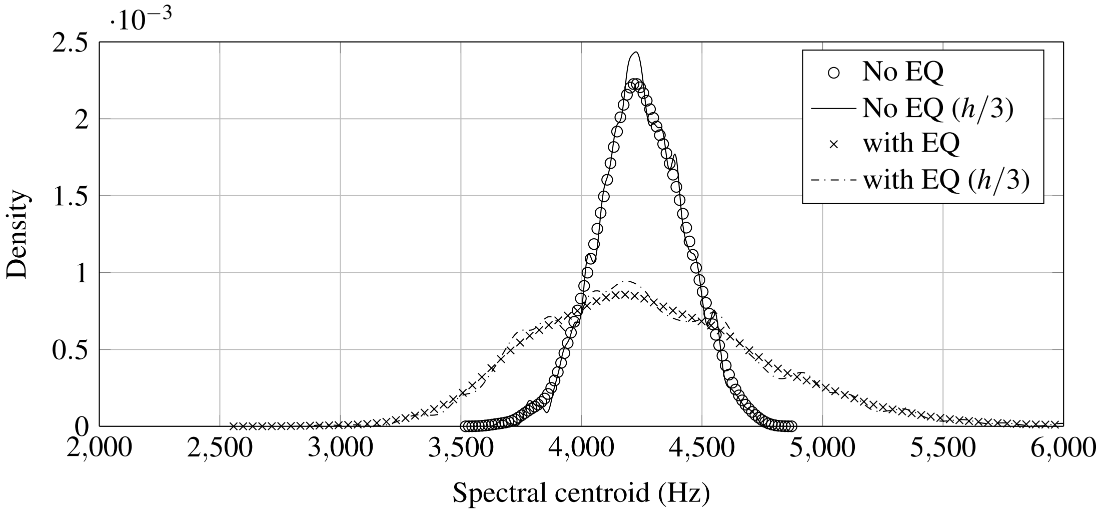


Figure 5.10: KDE of spectral centroid, for 1,000 vMF-distributed mixes of 'Burning Bridges', with vocal boost (μ = Eqn. 5.5, κ = 200), before and after random equalisation. h is the default kernel width.

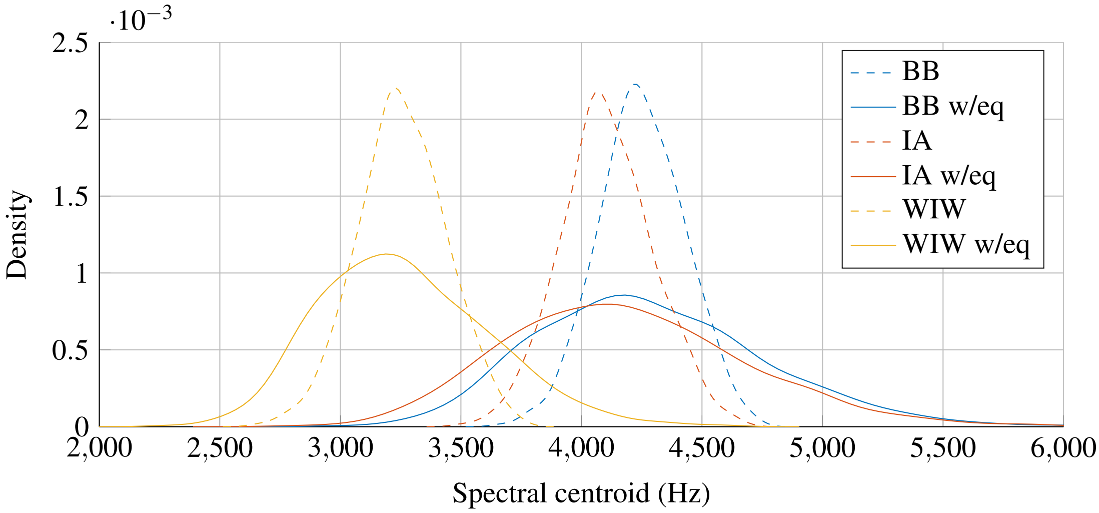


Figure 5.11: KDE of spectral centroid, for 1,000 vMF-distributed mixes of three songs, with vocal boost (μ = Eqn. 5.5, κ = 200), before and after random equalisation.

5.3.3.1 Note onsets

Onset detection is an active area of research seeking to identify when a note occurs in a piece of music and also to characterise the onset of the note according to parameters such as attack slope [180–182]. In the MIRtoolbox, onset detection can be performed using either an envelope-based method or spectral flux-based method. The envelope method was used here. Polyphonic onset detection is a greater challenge, particularly if multiple instruments are involved, where these different instruments have varying envelope characteristics, such as attack time and attack slope. Bello et al. [180] compared the performance of five methods of onset detection on a number of audio clips. The result showed that accuracy is reduced for a "complex mix" when compared to individual instruments, with the lowest true positive rate and highest false positive rate being achieved on this particular audio clip. As each of the instruments in a mix can have a different number of onsets, an onset detection algorithm should return varying results for a mixture of these

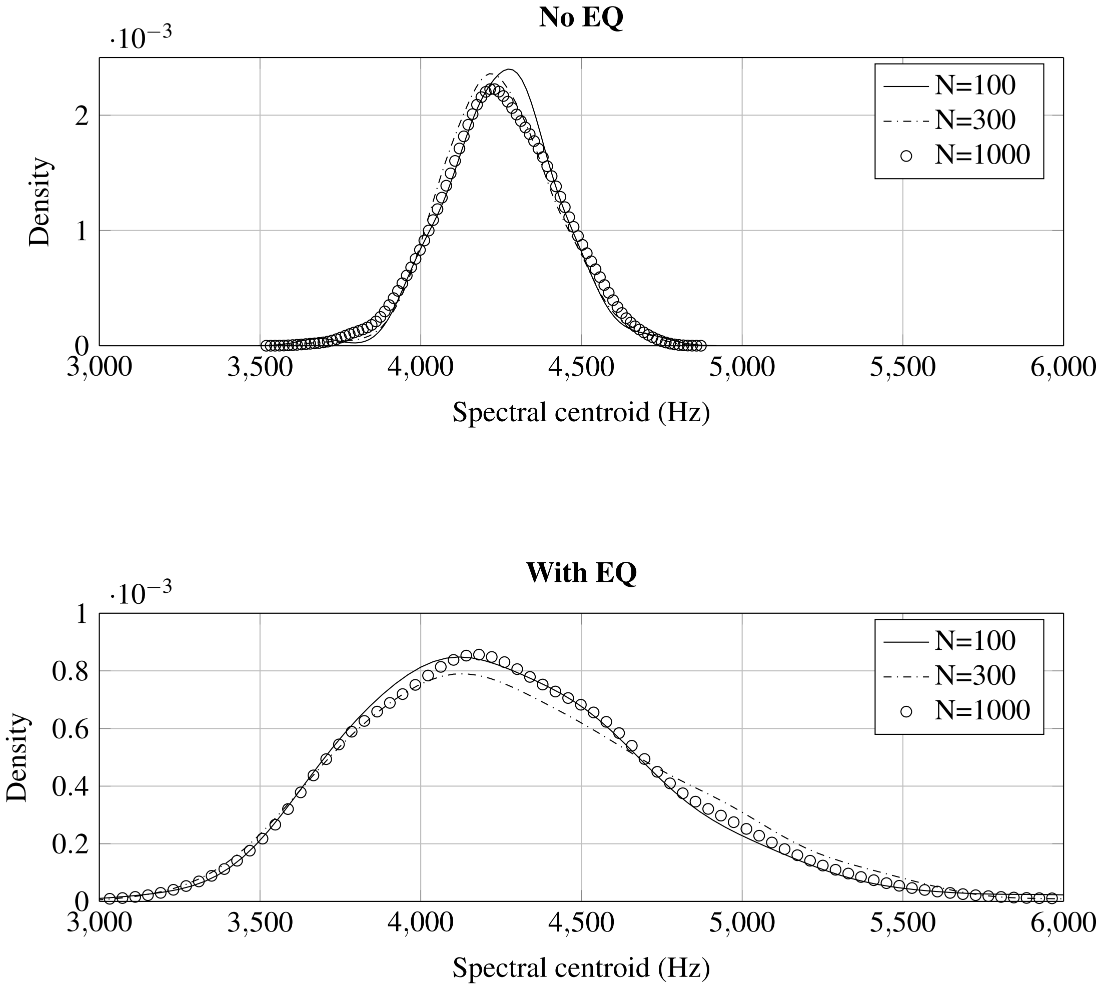


Figure 5.12: Effect of sample size on spectral centroid KDE, for "Burning Bridges"

instruments, depending on the relative volume of each instrument in the mix and the ease at the algorithm can pick out the individual onsets of quieter instruments.

5.3.3.2 Tempo

If the tempo of a song is 100 beats per minute (bpm) it follows that all mixes of the song are also at 100 bpm. This is to say that it is trivial to obtain the ground truth tempo values for all mixes of a song. However, current tempo estimation algorithms are imperfect. Classic methods of tempo estimation relied on detecting periodicities in the onset detection curve by means of autocorrelation. This method, and some derivatives, can be prone to "octave errors", where the estimated tempo is twice, or half the correct tempo. This can also produce other fractional errors. The MIRtoolbox includes two tempo estimation methods: 'classical', as above, and 'metre' [183]. The latter tracks the metrical structure of the audio, allowing a more consistent estimation of tempo. Of course, tempo estimation is a frequently attempted task and the subject of competitions, such as the annual MIREX (Music Information Retrieval Evaluation eXchange) challenges. Many of the more contemporary, and high-performing, algorithms have been entries/winners of these competitions [184, 185]. The issue of estimation accuracy has been addressed in some recent publications [186, 187].

By measuring the estimated tempo over the set of random mixes, for a number of songs, the

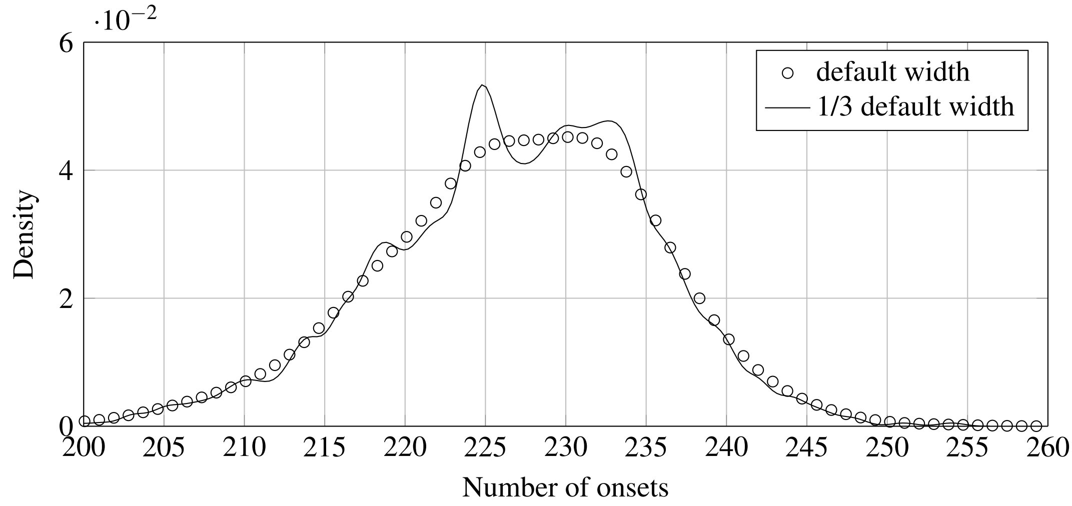
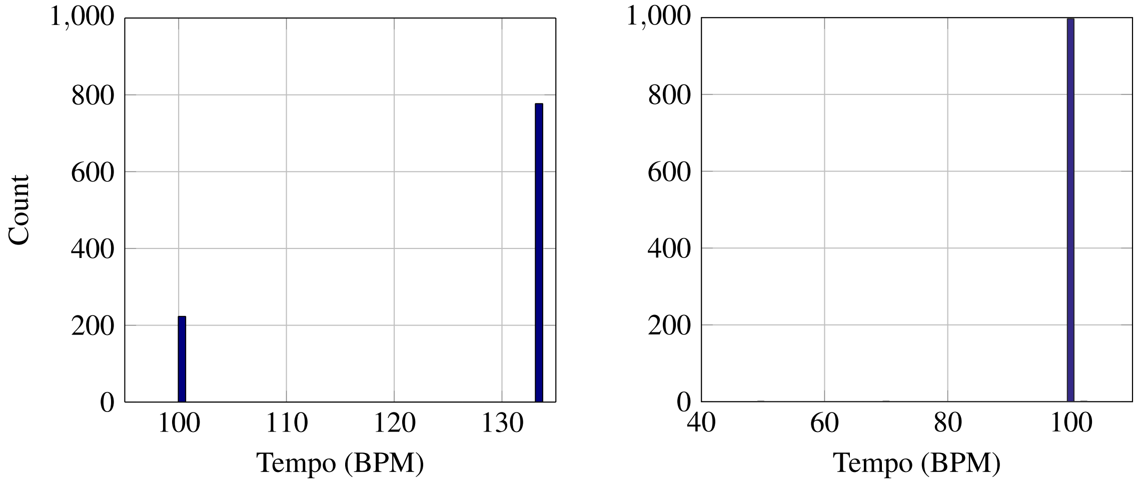


Figure 5.13: KDE of note onsets, for 1,000 vMF-distributed mixes of ‘Burning Bridges’ with random equalisation. Onset detection used the envelope method from MIRtoolbox.



(a) The tempo was estimated using the “classic” form of mirtempo, in the MIRtoolbox. The correct tempo (100 bpm) was only estimated in 223/1,000 cases, while the majority of mixes were estimated to be approximately 133 bpm.

(b) The tempo was estimated using the “metre” form of mirtempo, in the MIRtoolbox. The correct tempo (100 bpm) was estimated in 997/1,000 cases, while the remaining mixes were estimated to be approximately 50, 70 and 102 bpm.

Figure 5.14: Histogram of estimated tempo for 1,000 random mixes of “Burning Bridges” with random equalisation

Table 5.2: Tempo estimation accuracy results. Shown is the proportion of the 1,000 mixes for which the correct tempo was estimated.

Song	Ground truth	mirtempo(classic)	mirtempo(metre)
Burning Bridges	100 bpm	0.223	0.997
I’m Alright	96 bpm	0.098	1.000
What I Want	99 bpm	0.976	1.000

accuracy of a tempo estimation can be assessed. The results of the tempo estimation accuracy investigation are shown in Table 5.2. Only three songs were used, as this was enough to show the level of disagreement that can exist between algorithms and across songs, despite these songs have similar tempi ("Burning Bridges" is arguably performed at 200 bpm but 100 bpm is considered correct due to octave error confusion). From the results it is clear that the metre-based method is more robust to changes in the mix than the classic method. While the classic method did perform very well for "What I Want", detecting the true tempo in almost 98% of mixes, the performance for "I'm Alright" was a very poor 9.8%.

5.3.3.3 Pulse clarity measurement in alternate mixes

Figure 5.14a indicates the inaccuracy of the mirtempo(classic) tempo estimation algorithm, on the set of 1,000 mixes of "Burning Bridges". In order to better understand the reasons for this type of inaccuracy, other features must be investigated. Pulse clarity is defined as the ease with which listeners can perceive the underlying rhythmic or metrical pulsation in a piece of music [179]. Subsequently, it was hypothesised that the measured pulse clarity would vary for different mixes based on the relative contributions of the varying instruments. For example, louder drums may make tempo estimation easier if that drum pattern is one with clear note onsets being played in a predictable and stable pattern. Figures 5.15, 5.16 and 5.17 each show the relationship between the measurement of pulse clarity and the gains of each individual track in that specific mix. The track with which pulse clarity is most strongly correlated is vocals and this correlation is negative. This suggests that an increased level of vocals (and therefore a relative decrease in the level of all other instruments) results in an increased difficulty in a listener recognising the underlying pulse of the song as a whole. This is a logical finding, as the rhythm of a lead vocal is often less regular than an instrument such as a drum kit and a vocal performance may contain frequent periods of silence between phrases. Additionally, due to the reduced transients compared to drums, onset detection can be a greater challenge in vocals, which can add inaccuracies to tempo estimation. Supporting this conclusion is the positive correlation between both drum overhead tracks and pulse clarity, especially considering that the correlation is not strong for either kick drum or snare drum. This indicates that when listening to the drum kit *as a whole* the pulse of the music can be perceived with greater ease than individual components in isolation. This finding was observed in all three songs investigated.

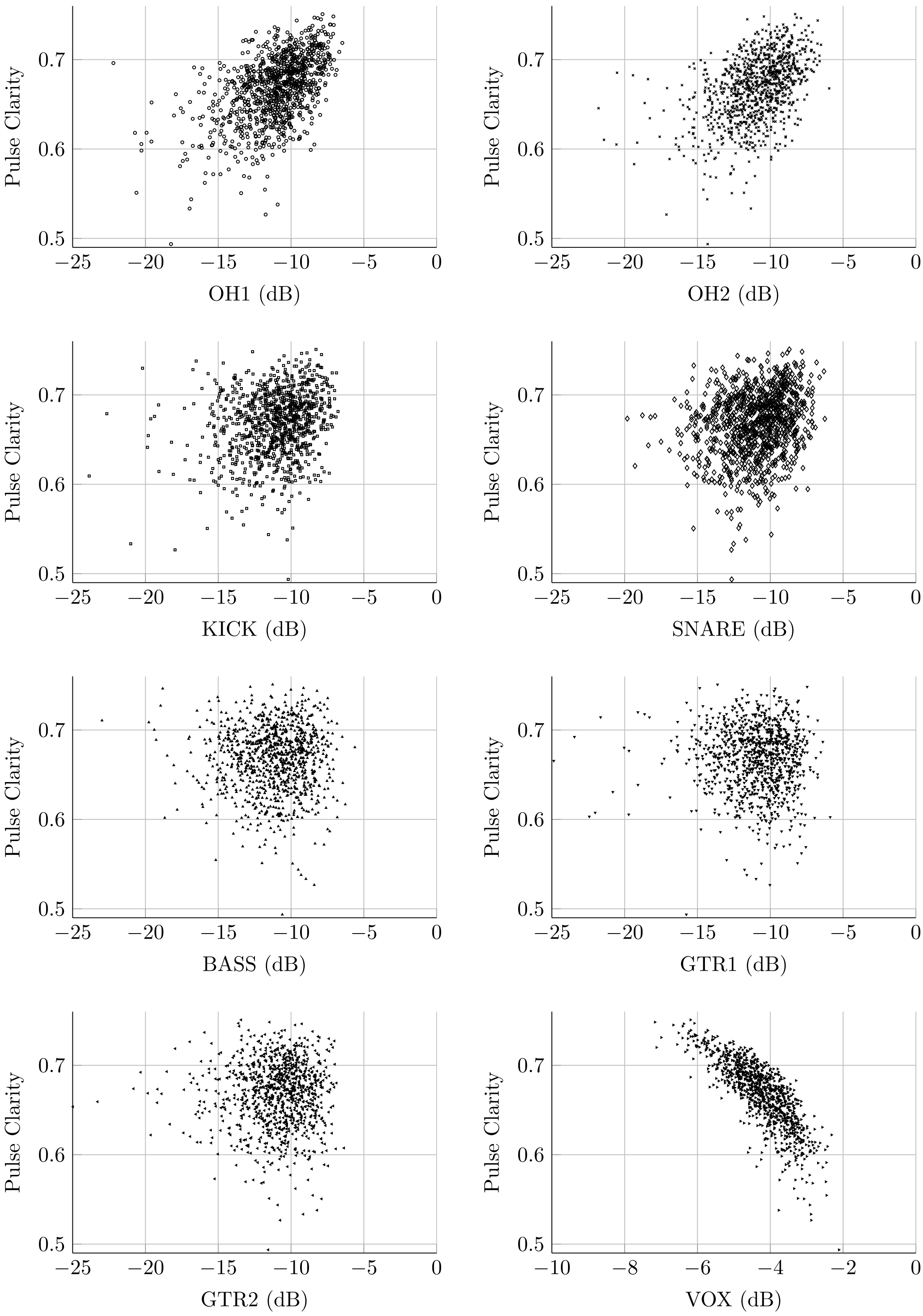


Figure 5.15: Variation in measured pulse clarity (before EQ) when compared to individual track gains, in dB, for the song "Burning Bridges". The track gain with which pulse clarity is most strongly correlated is vocals, followed by drum overheads.

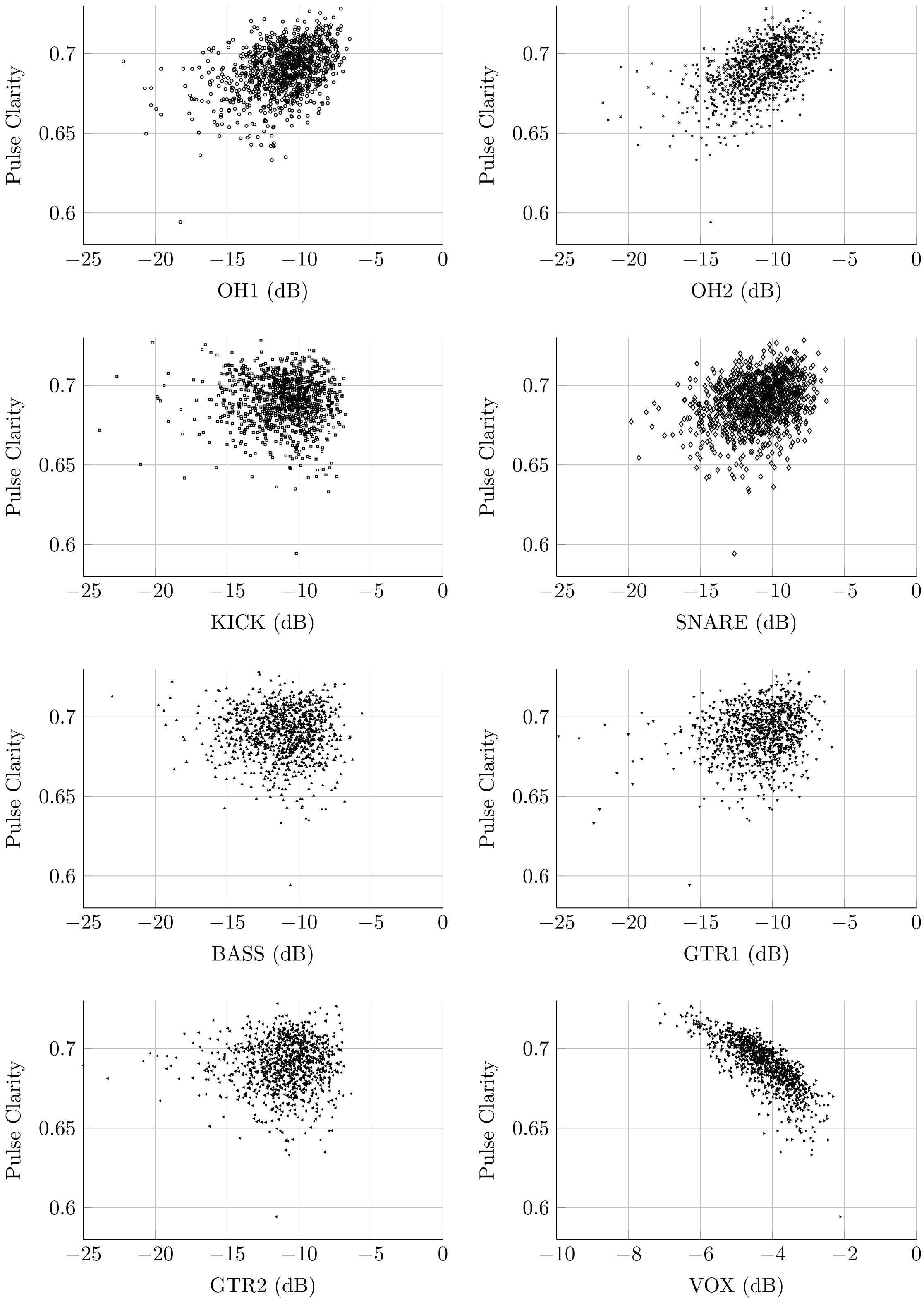


Figure 5.16: Variation in measured pulse clarity (before EQ) when compared to individual track gains, in dB, for the song "I'm Alright". The track gain with which pulse clarity is most strongly correlated is vocals, followed by drum overheads.

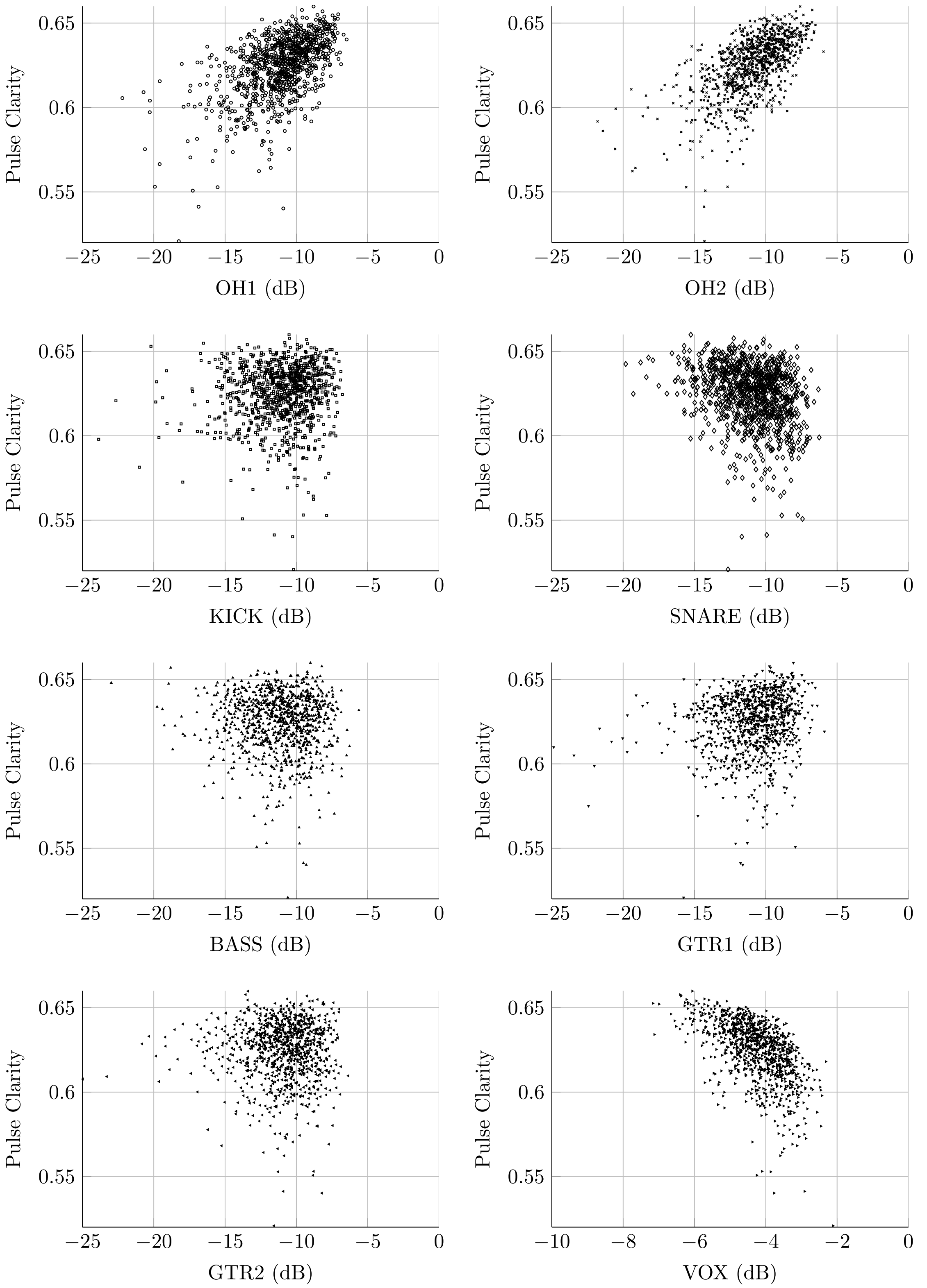


Figure 5.17: Variation in measured pulse clarity (before EQ) when compared to individual track gains, in dB, for the song "What I Want". The track gain with which pulse clarity is most strongly correlated is vocals, followed by drum overheads.

5.4 Mixes informed by experimental results

Of course, equal-loudness mixes with boosted vocals is a simplification of what levels real mix engineers actually use. From Chapter 4 we know that there is some degree of consensus when it comes to setting levels in a pop/rock mix. Consequently, a new distribution was made, with a value of μ directly informed by experiment. The results from § 4.5 were used, specifically the result shown in Fig. 4.43. By using the median levels for each track as a starting point for a new distribution, the new, informed, value of μ was as follows.

$$\mu_{\text{informed}} = [0.2174 \quad 0.2201 \quad 0.3107 \quad 0.2584 \quad 0.4280 \quad 0.3489 \quad 0.3107 \quad 0.5197] \tag{5.6}$$

Since the median values do not represent an observed data point, the L_2 norm of μ_{informed} is not necessarily equal to 1 (in fact, it is approximately 0.96 in this case). In order to use this as a mean direction in a vMF distribution, μ_{informed} was divided by the L_2 norm, resulting in the following.

$$\mu_{\text{informed}} = [0.2254 \quad 0.2282 \quad 0.3221 \quad 0.2679 \quad 0.4437 \quad 0.3616 \quad 0.3221 \quad 0.5387] \tag{5.7}$$

A concentration parameter $\kappa = 200$ was used. The result was a set of 1,000 mixes, the gains of which are shown in Fig. 5.18. Feature extraction was then undertaken in a manner identical to the naïve approach. The perceived loudness was not extracted for this set of informed random mixes, as the normalisation of the mixes was sufficiently demonstrated by the result in Fig. 5.9. When the spectral centroid of all mixes in this set had been obtained, the estimated probability density function was determined using KDE. The resultant distributions are shown in Fig. 5.19, for “Burning Bridges” (equivalent figures for the other songs are shown in Chapter 6). The mean value is approximately 3950 Hz, before and after EQ, as indicated by the peaks in the $h/3$ curves for “no EQ” and “w/EQ” conditions. As the informed mixes have proportionally higher vocal levels than the naïve mixes, as well as other characteristics such as attenuated drum overheads and boosted bass guitar, the distribution of given features was influenced by the feature values of these instruments and how they interact in the generated mixes. For example, drums overheads typically have a relatively high spectral centroid compared to the other instruments and so an attenuation results in a lower spectral centroid. The same effect is generated by an increase in the bass guitar. Consequently, it is expected that, when following μ_{informed}, the spectral centroid distribution of a set of generated mixes is lower than in the naïve case.

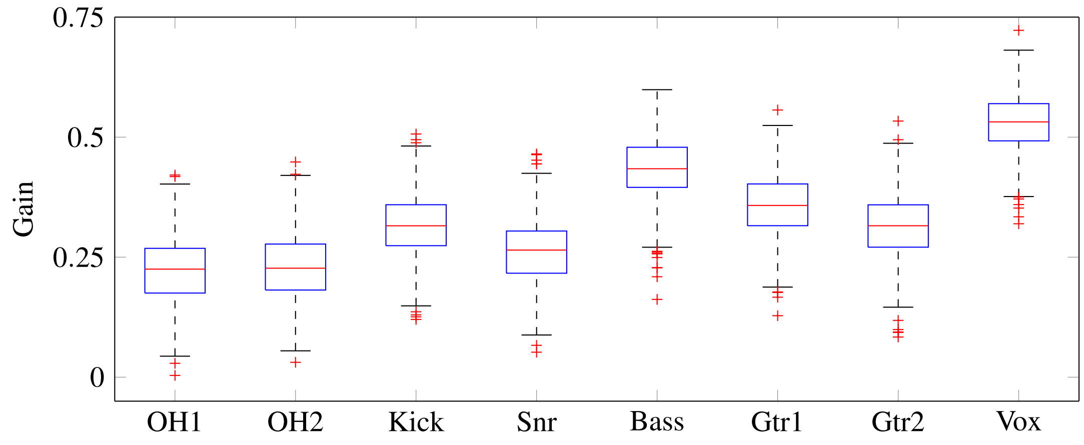


Figure 5.18: Boxplot of gain values for 1,000 mixes, generated from VMF distribution informed by Fig. 4.43 (μ = Eqn.5.7, $\kappa = 200$)

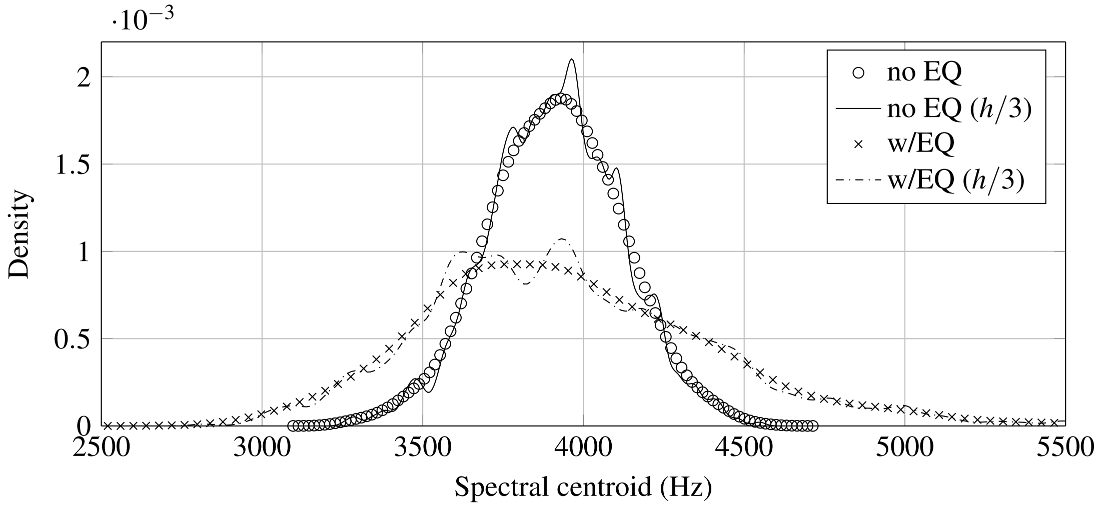


Figure 5.19: KDE of spectral centroid, for 1,000 vMF-distributed mixes of ‘Burning Bridges’, with gains informed by Fig. 4.43 (μ = Eqn. 5.7, $\kappa = 200$), before and after random equalisation. h is the default kernel width.

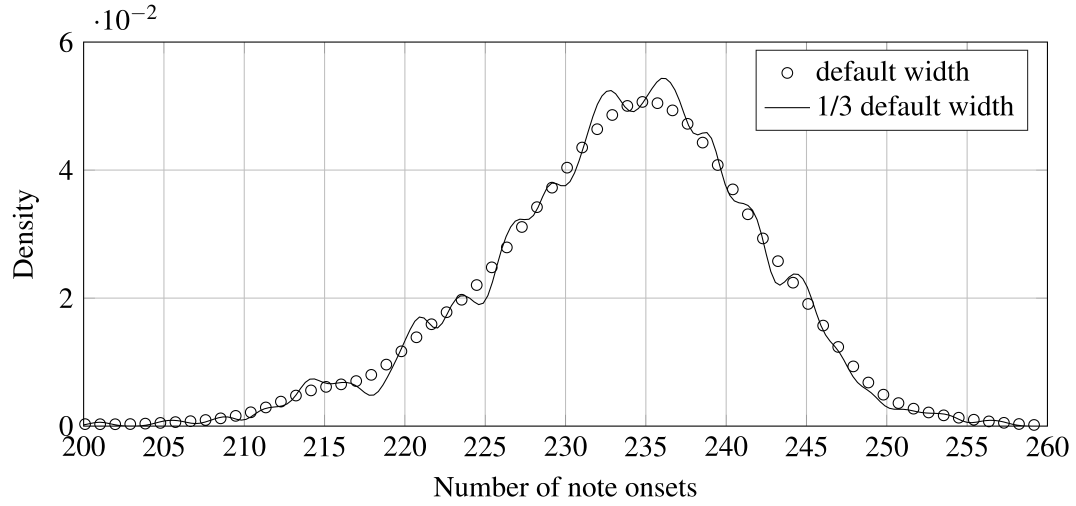


Figure 5.20: KDE of note onsets, for 1000 vMF-distributed mixes of ‘Burning Bridges’ with random equalisation and gains informed by Fig. 4.43 (μ =Eqn. 5.7, $\kappa = 200$).

5.5 Generating randomised panning

Up to now, all mixes considered were single-channel mixes, i.e., mono. For generating stereo mixes, a number of methods were trialled in attempting to create random panning.

5.5.1 Method 1 — separate left and right gains

The method for random gains was used to create separate mixes for the left and right channels of a stereo mix. Recall that hard panning only exists when the gain in one channel is zero. Since the vocal boost prevents any zero-gain on vocals, the panning of the vocals is much less wide than the other tracks. Additionally, since $\kappa = 200$ was chosen to prevent any negative gains, there are few zero-gain instances, therefore, a lack of hard panning. Fig. 5.21 shows the gain settings produced and a boxplot of the resulting pan positions — the inter-quartile range extends to ±0.4 for the seven instrument tracks and about ±0.2 for the vocals. The estimated density of pan positions for each track is shown, illustrating the relatively narrow vocal panning. As expected, these estimated density functions are Gaussian, to a good approximation.

5.5.2 Method 2 — separate gain and panning

This method involved generating random mono mixes as section § 5.1 and then generating pan positions separately. A μ_{pan} was created for a vMF distribution. This vector was based on the experimental results shown in Fig. 4.46. They showed that overheads and guitars were panned while kick, snare, bass and vocals were positioned centrally.

$$\mu = [-0.5 \quad 0.5 \quad 0 \quad 0 \quad 0 \quad -0.4 \quad 0.4 \quad 0] \tag{5.8}$$

This then needs to be a unit vector for it to be used in creating vMF-distributed points. Consequently, the precise values are not critically important, as it is the relative pan positions that are reflected in the normalised vector.

$$\mu = [-0.5522 \quad 0.5522 \quad 0 \quad 0 \quad 0 \quad -0.4417 \quad 0.4417 \quad 0] \tag{5.9}$$

Three different values for κ were used, which illustrates how this parameter controls the distribution of panning. The results are shown in Fig. 5.22.

5.5.3 Method 3 — informed left and right gains

While method 2 was informed by the *general* pan positions of the tracks, method 3 was informed by the median stereo gains which produced those pan positions, as shown in Fig. 4.48. Therefore, method 3 has the advantage that different instruments can have different variance of pan positions, with κ acting as a scaling variable for each variance. The vectors used are shown in Eqns. 5.10 and 5.11. To avoid negative track gains and resulting phase inversions, the absolute magnitude of the gain was used.

$$\mu_L = [0.27414 \quad 0.13544 \quad 0.33612 \quad 0.26565 \quad 0.4401 \quad 0.37959 \quad 0.25659 \quad 0.5651] \tag{5.10}$$

$$\mu_R = [0.11886 \quad 0.25966 \quad 0.31617 \quad 0.26118 \quad 0.4683 \quad 0.29354 \quad 0.37265 \quad 0.55311] \tag{5.11}$$

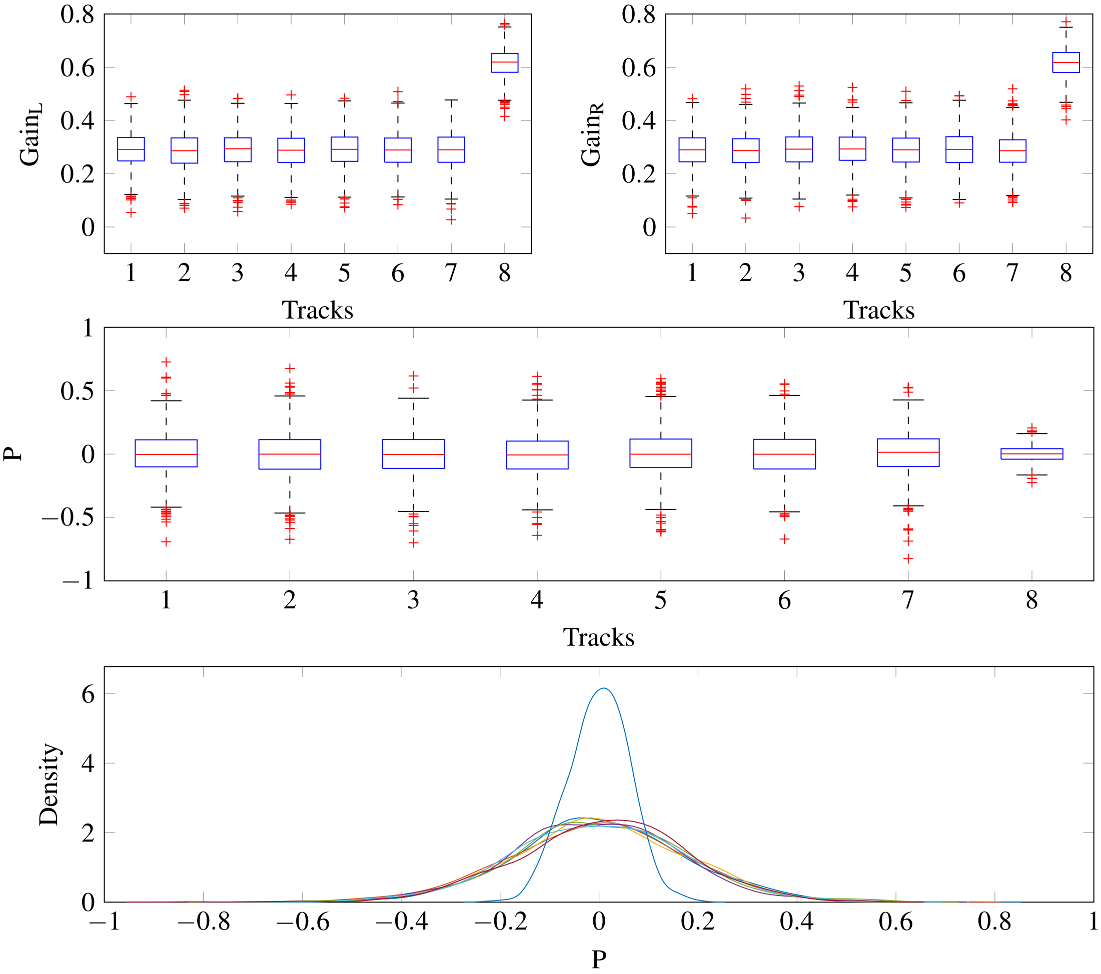


Figure 5.21: Panning method 1 — separate vMF distributions for g_L and g_R.

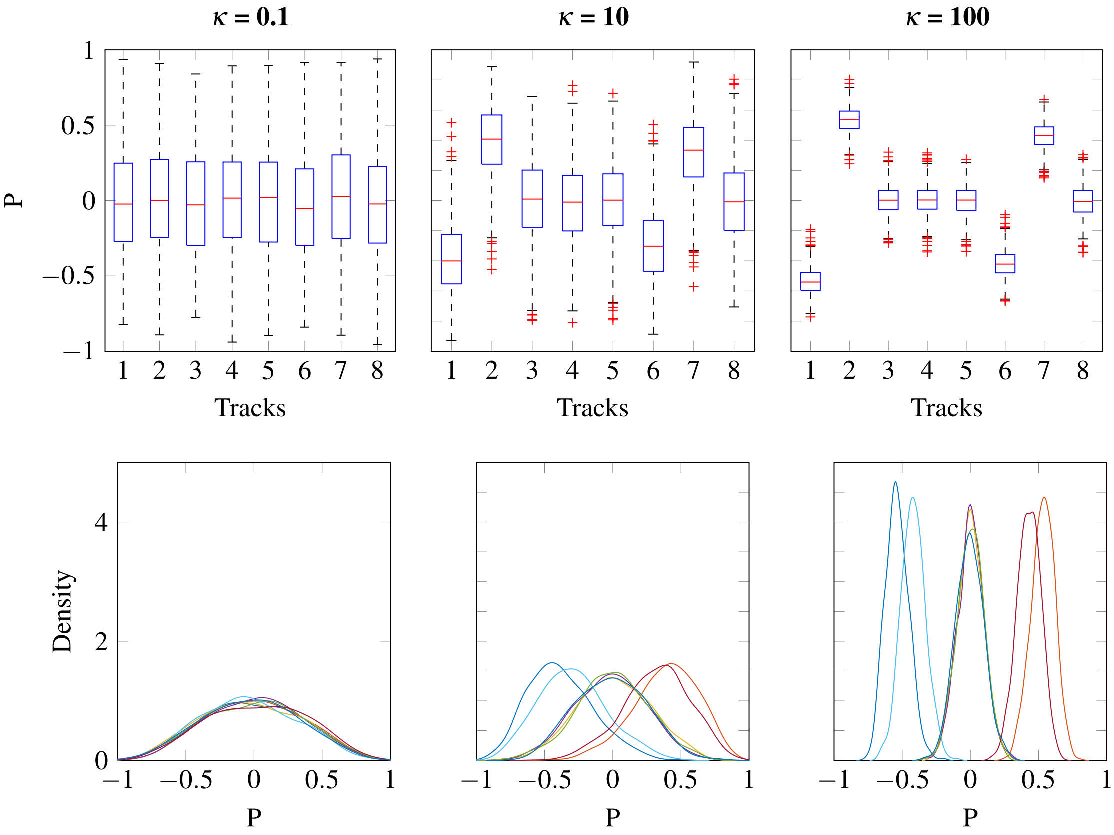


Figure 5.22: Panning method 2 — vMF distribution in panning space

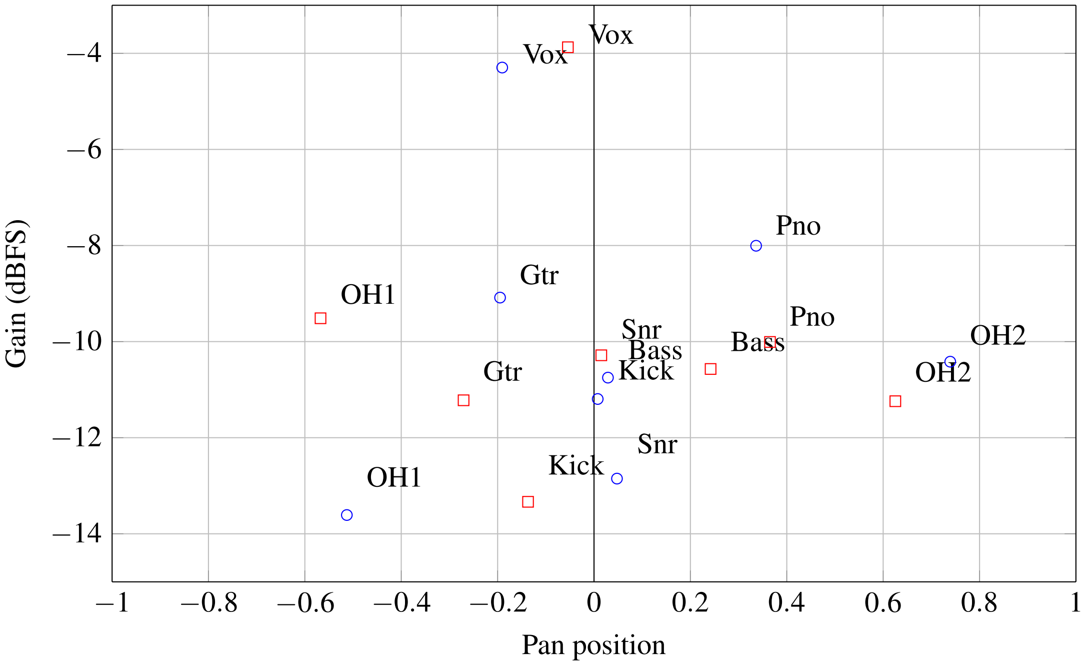


Figure 5.23: Two random mixes generated using panning method 2, shown as squares and circles. Each mix has a different gain vector and different pan vector (based on Eqn. 5.9).

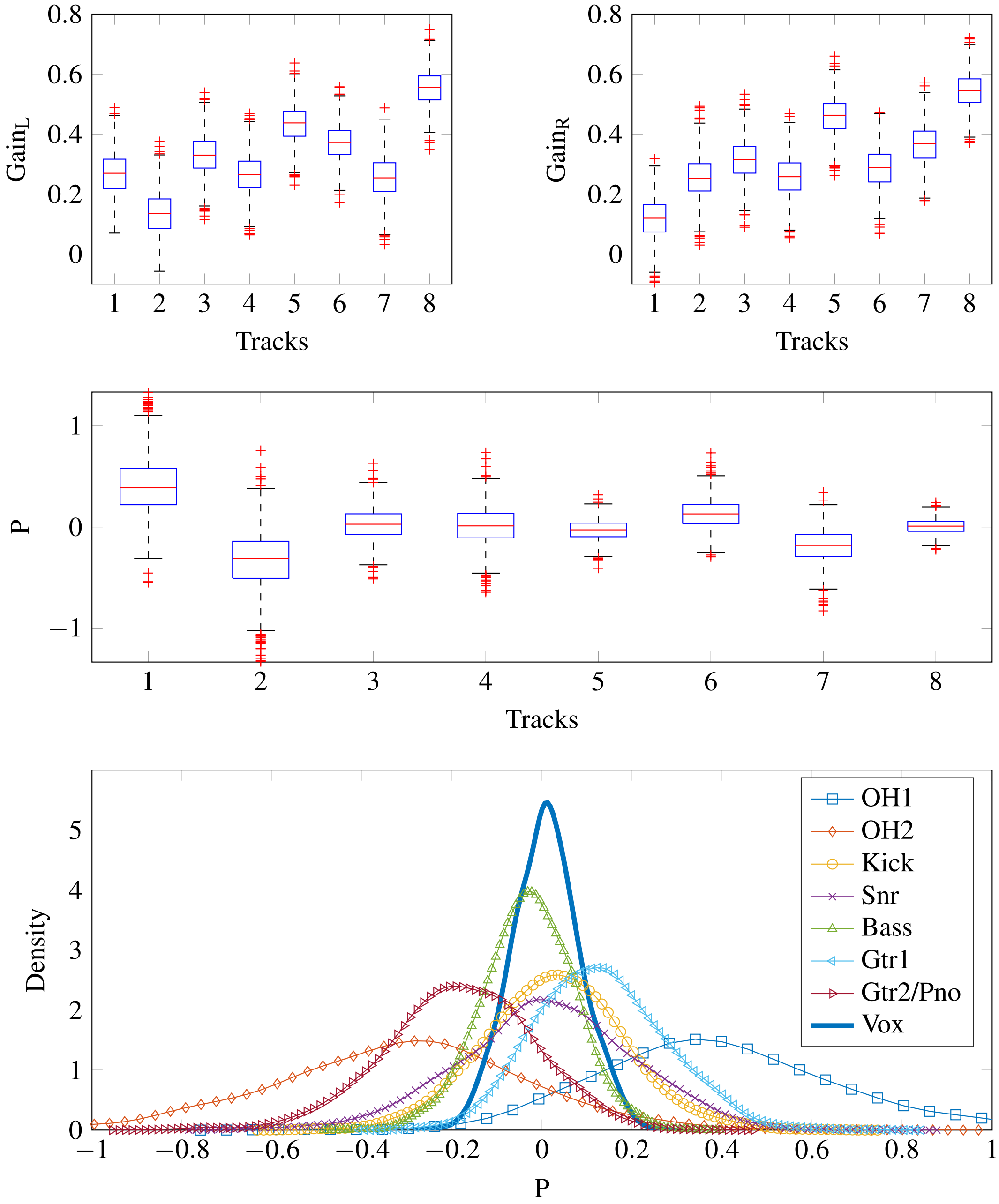


Figure 5.24: Panning method 3 — vMF distribution based on mix-space stereo result, shown in Fig. 4.48.

5.6 Analysis of stereo mixes

For each of the three methods, 1,000 stereo mixes were generated using the audio tracks from three songs, as was done for mono mixes. From these mixes the width was measured using the stereo panning spectrogram [188].

5.6.1 Method 1

When method #1 was used to create 1,000 random mixes with random stereo panning, Fig. 5.21 suggested that the range of pan positions would be relatively small, compared to the other two methods. Figure 5.25 shows the distribution of measured stereo width (using stereo panning spectrogram) from the 1,000 mixes, confirming that the perceived width is relatively low, generally below 0.1. Mixes of ‘I’m Alright’ typically produced wider mixes. This is possibly due to the fact that the two ‘guitar’ tracks, for this song, were actually guitar and piano — being that these two instruments are less similar than two guitars playing the same part, when they are panned left and right, the impression is that less of a phantom centre is created. For all three songs, the application of equalisation does not appear to significantly change the distribution of width measurements.

5.6.2 Method 2

For creating random mixes for measurement, the following parameters were used. This is in contrast to the example in § 5.5.2 but ensures that drum overhead panning is wider than guitar panning, on average.

$$\mu = [-1 \quad 1 \quad 0 \quad 0 \quad 0 \quad -0.5000 \quad 0.5000 \quad 0] \tag{5.12a}$$

$$\mu_{\mathrm{NRM}} = [-0.6325 \quad 0.6325 \quad 0 \quad 0 \quad 0 \quad -0.3162 \quad 0.3162 \quad 0] \tag{5.12b}$$

$$\kappa = 100 \tag{5.12c}$$

This method produces a reasonably narrow range of pan positions for each instrument (as shown in Fig. 5.22) and a relatively narrow range of measured pan values when random mixes are created (see Fig. 5.26). As the variance is low it is clear that the central values are quite dependent on the song in question. Unlike method #1, the application of equalisation does broaden the distribution, as clearly indicated by the lower values of maximum density, although the central values are only slightly increased.

5.6.3 Method 3

The use of method 3 to generate random stereo mixes results in the widest distributions of measured width, shown in Fig. 5.27, although the precise width, as in all methods, depends on the value of κ. As indicated by the two other methods, the mixes of “I’m Alright” are measured as wider than the two other songs. The addition of equalisation has results in no noticeable change in the distribution.

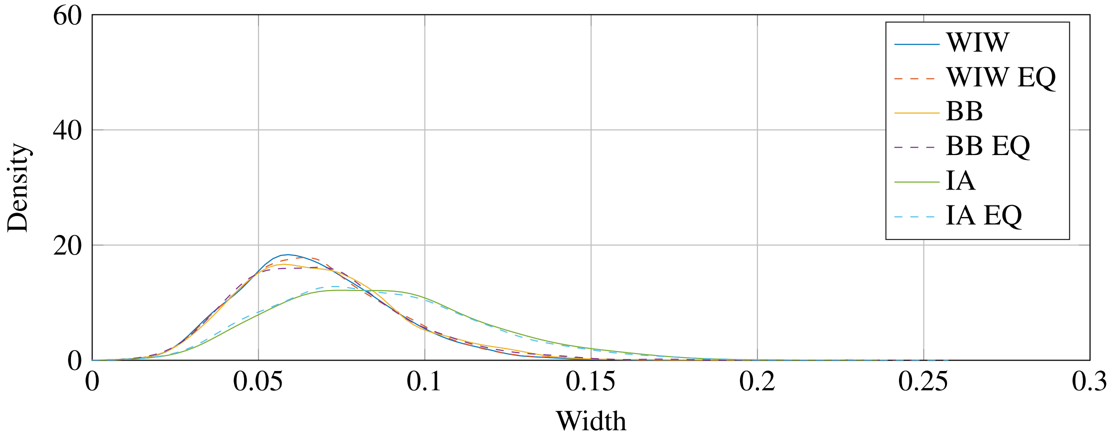


Figure 5.25: Method 1 — KDE of Width (all freq.), of 1000 mixes — GAIN vMF: $\kappa = 200, \mu =$ Eqn. 5.5

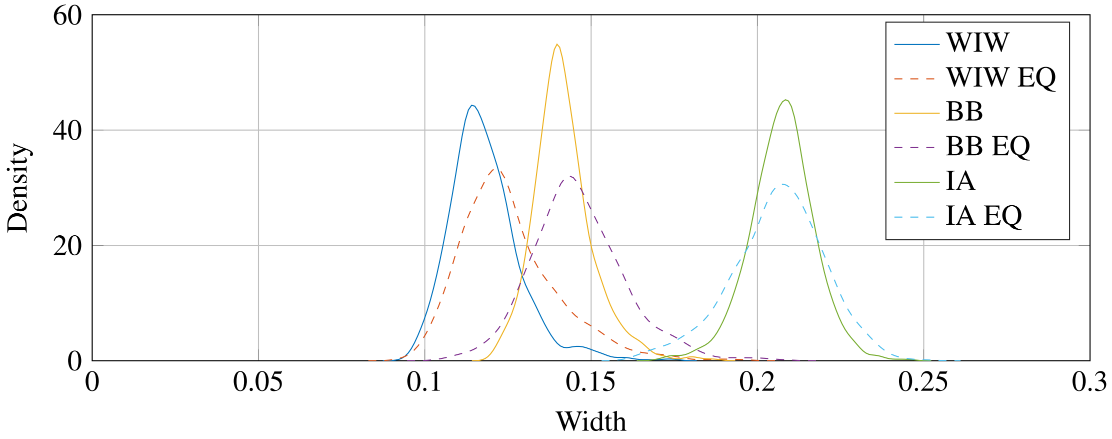


Figure 5.26: Method 2 — KDE of Width (all freq.), of 1000 mixes — GAIN vMF: $\kappa = 200, \mu =$ Eqn. 5.5, PAN vMF: $\kappa = 100, \mu =$ Eqn. 5.12b

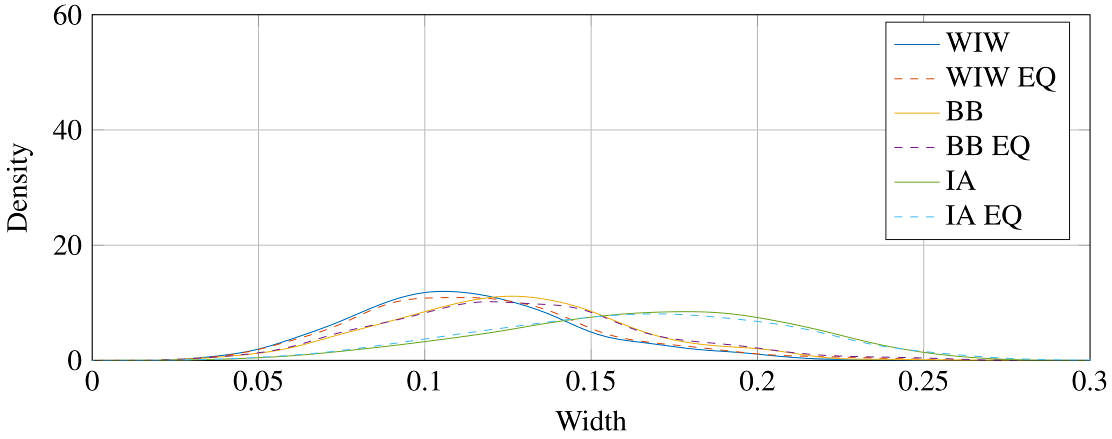


Figure 5.27: Method 3 — KDE of Width (all freq.), of 1000 mixes — GAIN L vMF:$\kappa = 200, \mu =$ Eqn. 5.10, GAIN R vMF: $\kappa = 200, \mu =$ Eqn. 5.11

5.7 Chapter summary

In this chapter, a method for generating random mixes was proposed, using a parametric model to populate the mix/tone/panning space described in Chapter 4. This model is based on a von-Mises-Fisher distribution, with a mean vector μ specifying a target mix, and a concentration parameter κ which specifies the variance in the distribution (uniform distrubitions can be achieved when $\kappa = 0$). By generating a large set of random mixes, these mixes can be characterised by audio signal features and the distribution of these features gives a good indication of their tolerance ranges when mixing.

How many random mixes are needed to fill the space? This study suggests that a value of 1,000 may have been much more than necessary, as feature distributions did not vary much in going beyond 300 mixes. The application of equalisation tends to broaden the distribution of features, unsurprising considering the additional degrees-of-freedom being introduced by this process.

The robustness of two tempo estimation methods to changes made during mixing was investigated by measuring the estimated tempo over all 1,000 mixes of three songs. This revealed that, even when a method fails to estimate the correct tempo for a given song, there is likely to be an alternate mix for which the correct tempo can be accurately measured. Additionally, it was shown that pulse clarity is increased when vocals are mixed at lower levels. These two findings could be useful for music information retrieval, as it suggests that various complex tasks, such as genre prediction, could be aided by re-mixing, where possible.

The techniques proposed in this Chapter are utilised later in this thesis, as the creation of a set of (pseudo-)random mixes is the first step in many evolutionary algorithms. Chapter 8 will continue where this chapter leaves off, in detailing such an example system.

It is important to challenge the idea that a particular song has specific values of signal features: all values are only specific to the mix of that song. This means that the analysis of signal features in sets of mixes can reveal novel insights. This is re-visited in Chapter 6, where a set of real-world mixes are subject to the same feature-extraction. The distributions can then be compared to the distributions of the random mixes in order to infer the motivations and actions of mix engineers in real-world conditions.

6
Analysis of real-world mixes

The diversity existing among music mixes has been discussed in previous literature, both qualitatively [189–191], quantitatively [33, 192, 193] and increasingly, both [37, 82] (refer back to literature review). Previous attempts to examine mixes using audio signal features have been limited to datasets which are too small to allow a detailed statistical analysis. One of the purposes of the work in this chapter is to make use of larger datasets in order to perform such analyses. The specific aims are as follows (aim #4 is addressed in Chapter 7):

1. To identify a source of audio mixes and create a large dataset for academic use
2. To objectively characterise such a dataset by means of audio signal feature extraction
3. To investigate the variation across all audio mixes
4. To investigate the variation across all mix engineers responsible for creating theses mixes

Portions of this chapter were published in 2015/2016 [194–196].

6.1 Variance in a large dataset of mixes

6.1.1 Dataset #2 — 1501 mixes

The data used in this study was collected directly from Cambridge Music Technology[1], which hosts multitrack content along with a forum where members can publicly post their mixes of that content. The database categorises multitrack content by genre. Of the ten most mixed sessions, eight belong to the "Rock/Punk/Metal" category. Table 6.1 shows the multitrack content which is used for the study in § 6.1.1. These mixes were gathered in late 2015[2]. The songs which have attracted the most mixes were specifically favoured. Due to the "Rock/Punk/Metal" category being preferred, this study focusses on these genres. Often-mixed songs from other categories are omitted in place of slightly less-often mixed songs from within this category. This allows the creation of a dataset which contains a consistent selection of instruments and sounds, including, but not limited to, drums, electric bass, guitars and vocals, as in Chapters 4 and 5.

Table 6.1: Audio samples obtained for this study.

Artist	Title	Tracks	Mixes
Angels in Amplifiers	I'm Alright	13	373
Dark Ride	Burning Bridges	32	183
Actions	Devil's Words	27	138
Young Griffo	Blood To Bone	23	135
The Brew	What I Want	37	129
Johnny Lokke	Promises and Lies	23	125
Hollow Ground	Ill Fate	21	118
Street Noise	Revelations	11	103
The Doppler Shift	Atrophy	22	100
Hollow Ground	Left Blind	16	97
TOTAL			1501

The majority of the mixes were only available in MP3 format, at bit-rates between 128kbps and 320kbps. All downloaded files were converted to PCM WAV format, at a sampling rate of 44.1kHz and a bit-depth of 16 bits. While lossy encoding, such as MP3, would have an effect on certain objective measures of the signal, such as reducing the value of *Spectral Centroid* and *Rolloff* features by the removal of some high frequency information (usually $>$ 16 kHz but dependent on settings [3]), this effect can be demonstrated to be negligible. Furthermore, Lee et al. [197] indicated that, for individual instrument tones, MP3 compression at 128 kbps "*caused almost no change in the timbre-spac*e", with relatively small changes in their spectral attributes (centroid, irregularity and incoherence).

For a given song, each mix was of a different length, due to varying amounts of silence at the start and end of each file and also occasional acts of creative re-arrangement such as the removal or duplication of certain bars. This made it difficult to use the entire audio in the analysis. To normalise the choice of audio segment, the audio was cut to short segments containing the second

[1] http://www.cambridge-mt.com/ms-mtk.htm

[2] Consequently, the number of available mixes is likely to have increased since that time.

[3] MP3 compression removes some high frequency content, although there is not typically a great deal of spectral energy above this point

chorus of the song, as in previous chapters. Each of these segments was then time-aligned, which was achieved by determining the peak in the cross-correlation vector when comparing one mix to all others. All of the mixes but one were zero-padded to align the files accordingly. Each mix was then trimmed to a 30-second length containing the chorus. This ensures that feature extraction tasks can be performed fairly on all mixes. This process was applied to each batch of mixes of each song. This processing assumes that tempo does not vary across mixes of the same song, as, if it were to vary, choosing the peak in the cross-correlation vector would not ensure that all mixes are in sync at all times. However, it was demonstrated by the success of this method that the tempo of all mixes of a particular song were identical. This was confirmed by audition.

6.1.2 Research questions

This dataset of mixes can be used to address a variety of challenges, a number of which are explored herein.

RQ-12 Which audio signal features vary most across mixes?

RQ-13 What are the dimensions of mix-engineering practice, across all songs and for a particular song?

RQ-14 How are the values of low-level features distributed in the dataset? What are their typical means and variance?

Direct subjective appraisal of these mixes, in the conventional sense of controlled listening tests, is not included in this thesis due to the overwhelming size of the dataset. However, as all mixes were created in real-world conditions, we assume each engineer produced their mixes to the best of their abilities and towards their desired targets. In this sense, subjective evaluation is implicit in the data itself. Additionally, a subset of this dataset forms the audio mixes that were entered into an on-line mix competition. Therefore, this subset does have some limited subjective evaluation and this is analysed in greater detail in § 6.2.

6.1.3 Feature-extraction

As many established audio signal features have been designed for Music Information Retrieval (MIR) tasks such as instrument recognition or genre classification, it is not widely understood which features would be best suited to categorising mixes of a given song. Features relating to the perception of polyphonic timbre were thought to be important (based on a chronologically earlier experiment which is described in Chapter 3) and so the sub-band spectral flux was determined, based on the work of Alluri and Toiviainen [128]. The statistical moments of the sample amplitude probability mass function (PMF) have been shown to categorise different types of distortion in mixing and mastering processes [108][4] and so these features are also used. Spatial features were derived from the stereo panning spectrogram (SPS) by Tzanetakis et al. [188]. Table 6.2 contains a full list of features. At this stage, features related to rhythm are not included, since the structure, form and meter of varying mixes should be identical if they are mixes of the same multitrack audio. Further discussion of rhythm can be found in Chapter 5.

[4] This paper reports on the experiment described in Chapter 3 but is not included in that chapter, as it is slightly outside the scope of this thesis.

Table 6.2: Audio signal features used in analysis. Features with KMO < 0.6, marked with an asterix, are not included in the PCA.

Feature	Label	Ref.	KMO
Spectral Centroid	SpecCent	[114]	0.758
Spectral Spread	SpecSpr	[114]	0.797
Spectral Skew	SpecSkew	[114]	0.851
Spectral Flatness	SpecFlat	[114]	0.898
Spectral Kurtosis	SpecKurt	[114]	0.852
Spectral Entropy	SpecEnt	[114]	0.639
Crest Factor	CF		0.967
LoudnessITU	LoudITU	[32]	0.834
Top1dB	Top1dB	[108]	0.900
Harsh	Harsh	[107]	0.633
LF Energy	LF	[107]	0.631
Rolloff85	RO85	[126]	0.819
Rolloff95	RO95	[126]	0.677
Gauss	Gauss	[107]	0.965
PMF Centroid	PMFcent	[108]	0.938
PMF Spread	PMFspr	[108]	0.890
PMF Skew		[108]	0.534*
PMF Flatness	PMFflat	[108]	0.962
PMF Kurtosis	PMFkurt	[108]	0.907
Width (all)	W.all	[110, 188]	0.966
Width (band)		[110, 188]	0.591*
Width (low)	W.low	[110, 188]	0.778
Width (mid)		[110, 188]	0.540*
Width (high)		[110, 188]	0.567*
Sides/Mid ratio			0.593*
LR imbalance		[36]	0.518*
Spectral Flux	sbflux1-10	[128]	All > 0.8

For these 1501 mixes, outlier detection was performed in the 36-dimensional feature-space (see Table 6.2). The Z-score of each point was determined by the Euclidean distance to the three nearest neighbours. Samples for which $Z > 2.5$ were deemed to be outliers. 35 such samples were found and once omitted there were 1466 audio samples remaining for further analysis.

6.1.4 Factor analysis

Principal Component Analysis (PCA) was used in order to reduce the dimensions of the feature-space. The appropriateness of PCA was tested as follows, based on a scheme proposed by Dziuban and Shirkey [129], and using R [130]. Using Bartlett's test of sphericity (using the `psych` package [131]), the null hypothesis that the correlation matrix of the data is equivalent to an identity matrix was rejected.

$$\chi^2(630, N = 1466) = 97162.75, p < 0.001$$

This indicated that factor analysis was a suitable analysis method. The Kaiser-Meyer-Olkin measure of sampling adequacy (KMO) was evaluated [132]. KMO for the full set of variables was 0.845. This value is above the value recommended by Hutcheson and Sofroniou [133] (0.6), and

Table 6.3: Eigenvalues of revised PCA.

	1st	2nd	3rd	4th
Eigenvalue	14.00	5.45	2.34	1.42
% variance	46.68	18.15	7.80	4.72
Cumulative % variance	46.68	64.83	72.62	77.34

by Kaiser (1974), who suggested a calibration of the index, shown in Table 3.4. The value of 0.6 was chosen as the cut-off, as it was both a more conservative and more contemporary value. Additionally, as there were no values below 0.5, such a cut-off would have had no extra benefit. This suggested that factor analysis would be both appropriate and useful. KMO for each individual variable was determined and any individual variables with a value less than 0.6 were excluded from analysis (see Table 6.2). Consequently, PCA was conducted with the remaining 30 variables. Each variable was standardised prior to PCA, i.e. rescaled such that mean $\mu = 0$ and standard deviation $\sigma = 1$. This initial PCA was not rotated and there was no limit on the number of components. The plot of eigenvalues is shown in Fig. 6.1.

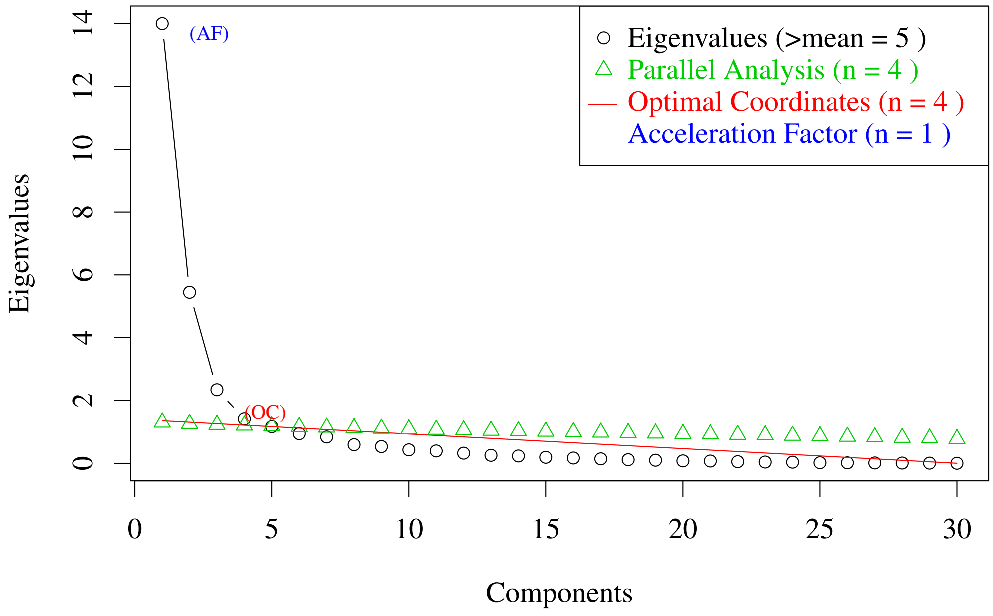


Figure 6.1: Scree plot for initial PCA, 1466 mixes. Also shown are the results of the *nFactors* analysis, demonstrating non-graphical solutions to the scree test.

As in Chapter 3, using the `nFactors` package [136], a variety of methods were employed in order to determine the number of dimensions to keep for further analysis, shown in Figure 6.1. This process was described in detail in § 3.3.2. Based on agreement suggested by three of the four methods, four dimensions were kept for the subsequent analysis. As before, 30 variables were used for a revised PCA, now limited to four dimensions and rotated using the varimax method [198]. This rotation was applied so that the resultant factors were easier to interpret, by ensuring variables had high loading on one dimension and low loadings on those remaining. The eigenvalues of this PCA are shown in Table 6.3, four dimensions accounting for $\approx 77\%$ of the variance. The aim

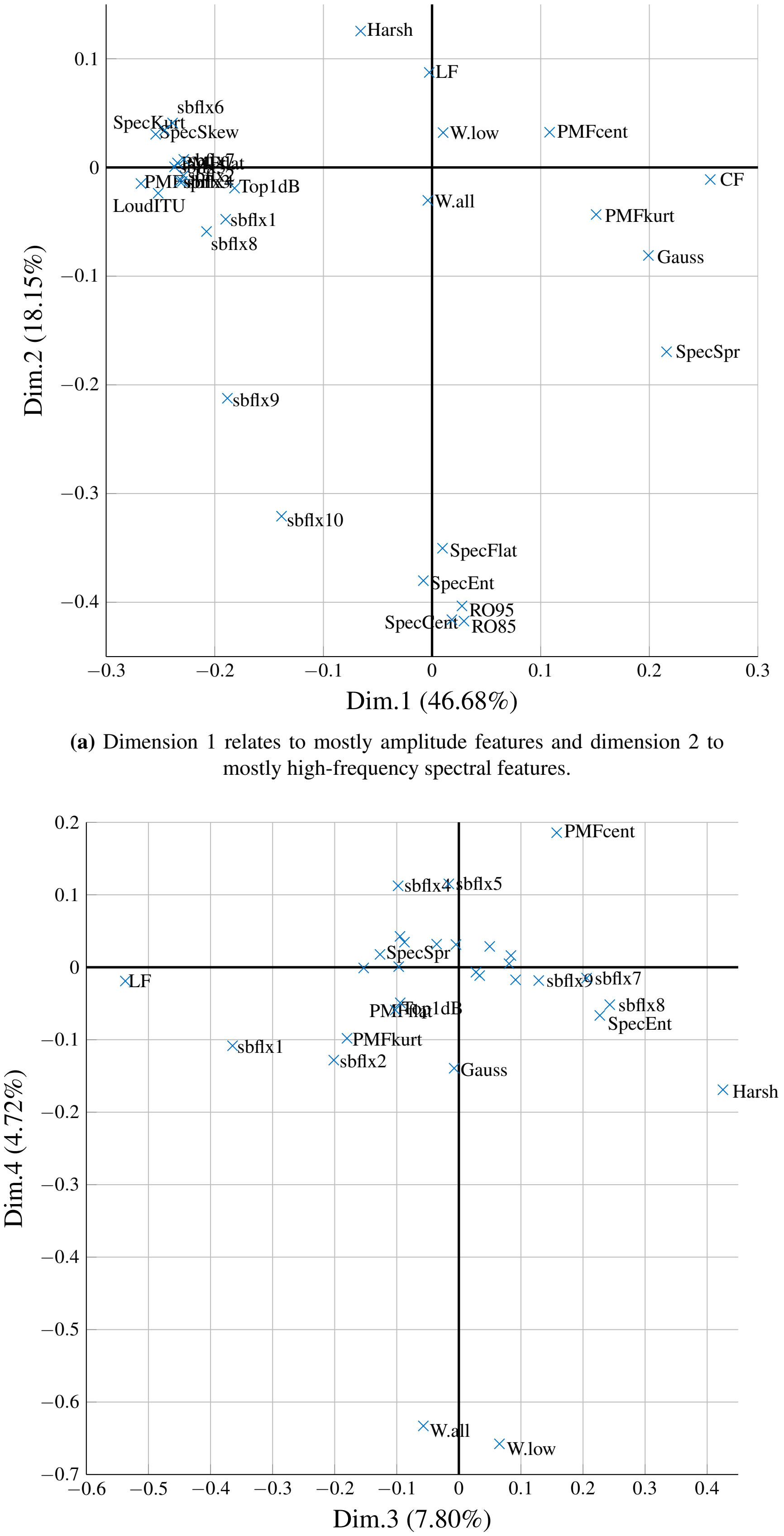
Table 6.4: Loadings of each variable to each component

Feature	**Loadings**			
	Comp1	Comp2	Comp3	Comp4
SpecCent	0.01825	-0.41606	0.09135	-0.017301
SpecSpr	0.21584	-0.16955	-0.12701	0.017751
SpecSkew	-0.25442	0.03050	0.04976	0.028781
SpecFlat	0.00963	-0.35029	-0.09688	0.000936
SpecKurt	-0.24626	0.03427	0.08393	0.016198
SpecEnt	-0.00800	-0.38014	0.22710	-0.066467
CF	0.25609	-0.01112	0.03358	-0.011442
LoudITU	-0.25212	-0.02367	0.02752	-0.006826
Top1dB	-0.18180	-0.01900	-0.10241	-0.058896
Harsh	-0.06573	0.12549	0.42542	-0.168963
LF	-0.00255	0.08740	-0.53723	-0.018975
RO85	0.02934	-0.41733	-0.03574	0.031890
RO95	0.02753	-0.40339	-0.09480	0.042657
Gauss	0.19914	-0.08094	-0.00757	-0.139339
PMFcent	0.10823	0.03238	0.15760	0.185700
PMFflat	-0.23436	0.00419	-0.15328	-0.000961
PMFspr	-0.26792	-0.01481	-0.08756	0.034592
PMFkurt	0.15104	-0.04328	-0.18015	-0.098049
W.all	-0.00388	-0.03029	-0.05717	-0.632903
W.low	0.01018	0.03206	0.06539	-0.657681
sbflx1	-0.18995	-0.04765	-0.36453	-0.108325
sbflx2	-0.22874	-0.00734	-0.20145	-0.128154
sbflx3	-0.23235	-0.01270	-0.09440	-0.049030
sbflx4	-0.22985	-0.01283	-0.09780	0.112366
sbflx5	-0.23700	0.00060	-0.01615	0.115090
sbflx6	-0.23885	0.04095	0.08085	0.004930
sbflx7	-0.22849	0.00740	0.20578	-0.014960
sbflx8	-0.20757	-0.05893	0.24309	-0.051601
sbflx9	-0.18835	-0.21232	0.12865	-0.018308
sbflx10	-0.13856	-0.32082	-0.00461	0.031369

of PCA was to reduce the set of features extracted to a small set of components which described the dimensions of the mixing process over which there was most variance. The following is an interpretation of each of the first four dimensions, based on the loadings of the individual features, as shown in Fig. 6.2a and 6.2b. This addresses research questions 12 and 13 from § 6.1.2.

1. Many of the input variables associated with signal amplitude, dynamic range and loudness are strongly correlated with the first principal component. Negative values indicate high amplitude mixes (see Fig 6.2a).

2. The second dimension can be described by the many strong correlations to spectral features with negative values denoting mixes that have a greater proportion of energy in higher frequencies (see Fig 6.2a).



(a) Dimension 1 relates to mostly amplitude features and dimension 2 to mostly high-frequency spectral features.

(b) Dimension 3 relates mostly to either low or high-frequency features and dimension 4 to spatial features. Labels for loadings < 0.1 are removed for clarity.

Figure 6.2: Results of PCA for 1466 audio samples. The variables factor maps, shown in (a) and (b), indicate loadings of variables on the varimax-rotated principal components.

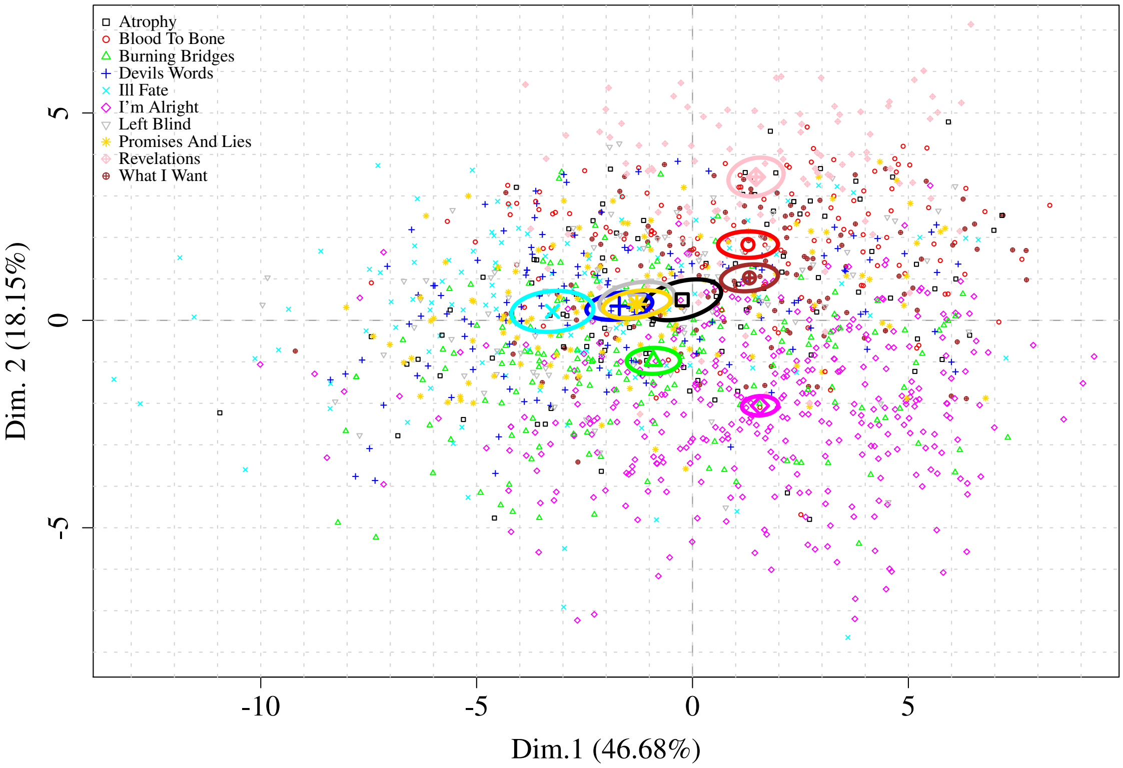


Figure 6.3: PCA individuals factor map — each point represents a single mix in the dataset and the colour/symbol represents which song it is a mix of. Group centroids are marked with a larger, bold symbol. Ellipses are drawn representing 95% confidence in the centroid. Mixes of a song vary more in dim.1 than dim.2, while songs differ from one another more along dim.2 than dim.1. The mixes of all songs overlap greatly in this feature-reduced space.

3. Features associated with low frequencies are more strongly loaded onto dimension 3 in the negative direction, while treble range features, such as “Harsh” and “sbflux” bands 7 & 8, are loaded with positive values (see Fig 6.2b).

4. Dimension 4 can be explained by the correlation of the spatial features to this dimension. As the value of this dimension decreases, the perceived width of the stereo image increases (see Fig 6.2b).

Figure 6.3 and Figure 6.4 show the dataset of mixes placed in the varimax-rotated PCA space. Each point represents a mix of a song, where the song is coded by a unique colour and symbol combination. We can see significant overlap between the range of mixes for all 10 songs. The estimated centroid of each group, and the 95% confidence ellipse of that centroid estimation, are also indicated in Figures 6.3 and 6.4.

6.1.5 Distribution of audio signal features

The density of each extracted feature was estimated using the `density` function in **R** with a Gaussian smoothing kernel. Figures 6.5, 6.6, 6.7 and 6.8 show the estimated density of four particular features extracted, features considered to be representative of the principal components

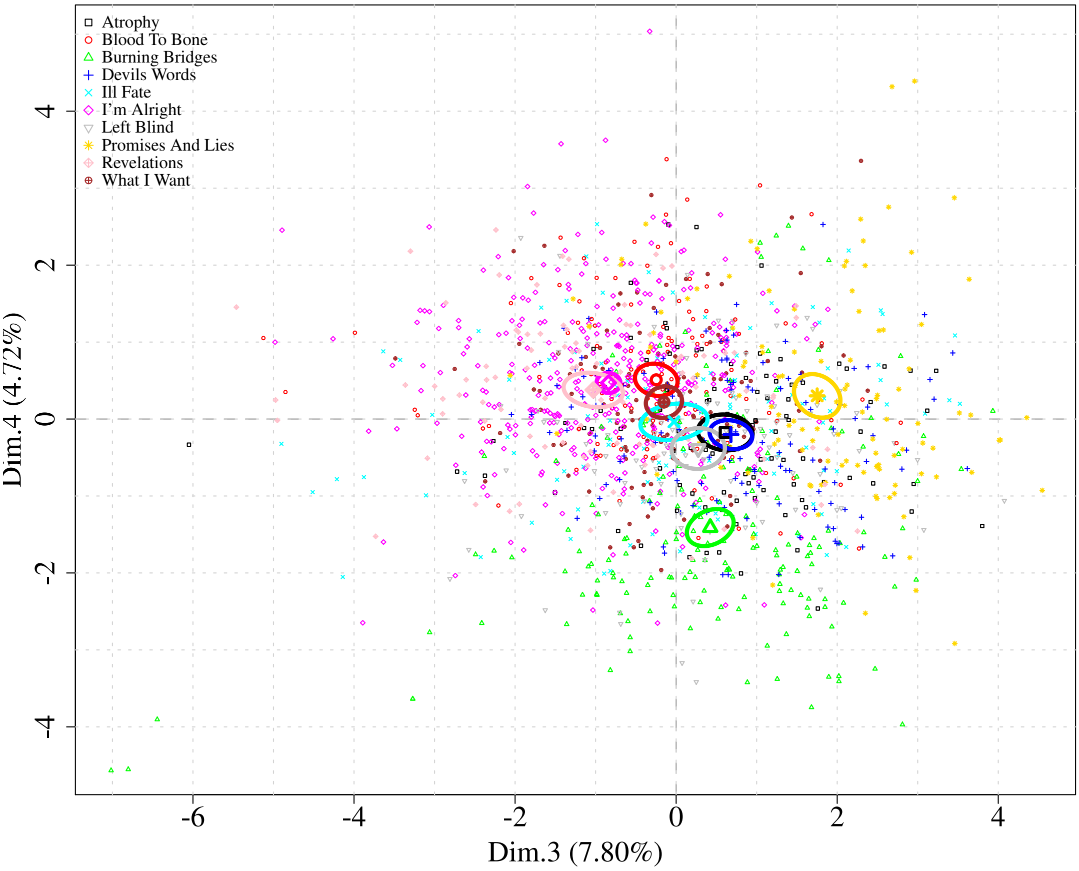


Figure 6.4: PCA individuals factor map — each point represents a single mix in the dataset and the colour/symbol represents which song it is a mix of. Group centroids are marked with a larger, bold symbol. Ellipses are drawn representing 95% confidence in the centroid.

due to their relatively high loadings. In each figure, estimated densities are shown for each song and also for all songs. The plots indicate that the distribution of features shows central tendency, whilst some curves display additional modes. A Shapiro-Wilk test of normality was carried out [199]. As this test is known to be biased for large sample sizes [200], the test was carried out not only on the raw data for each song but also the smoothed distributions shown in Figures 6.5, 6.6, 6.7 and 6.8, which contain fewer datapoints. The majority of these distributions tested were determined to be significantly different from a normal distribution: p-values are shown in Table 6.5.

A Gaussian Mixture Model (GMM) was used to determine how well the distribution over all mixes could be characterised by a sum of normal distributions. This was implemented using the `mixtools` package [201]. The function `normalmixEM` uses expectation maximisation for mixtures of normal distributions. The model parameters are shown in Table 6.6 and Figure 6.9, where λ_n is the mixing proportion (thus summing to 1), μ_n is the mean and σ_n is the standard deviation of each of the n Gaussian functions in the model. The coefficient of determination, R^2, is shown in

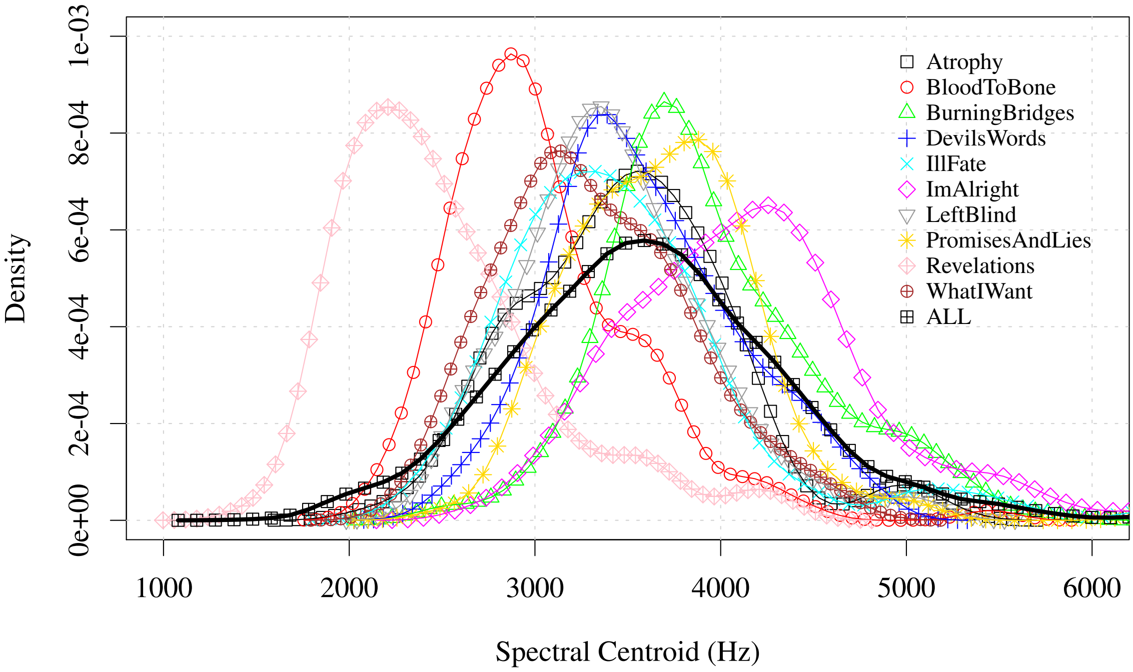


Figure 6.5: KDE of spectral centroid in 1466 mixes. The distributions shows distinct variation from song to song.

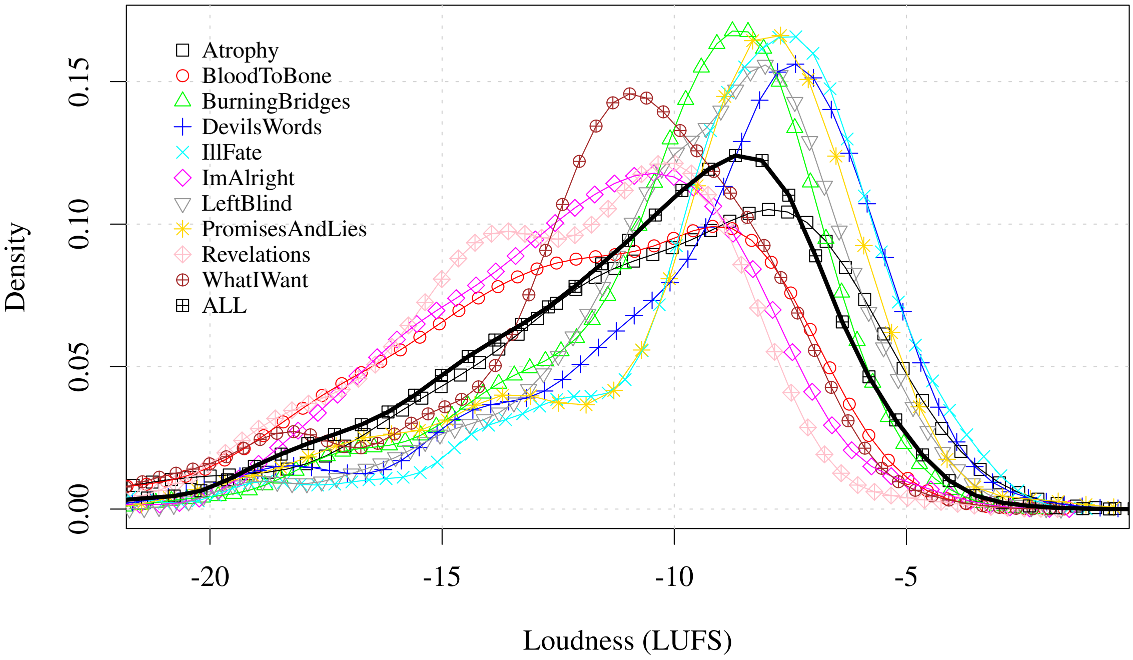


Figure 6.6: KDE of loudness in 1466 mixes. Many mixes were subject to mastering-style processing, resulting in high values of perceived loudness. Some songs, such as “Revelations” clearly show a bimodal distribution.

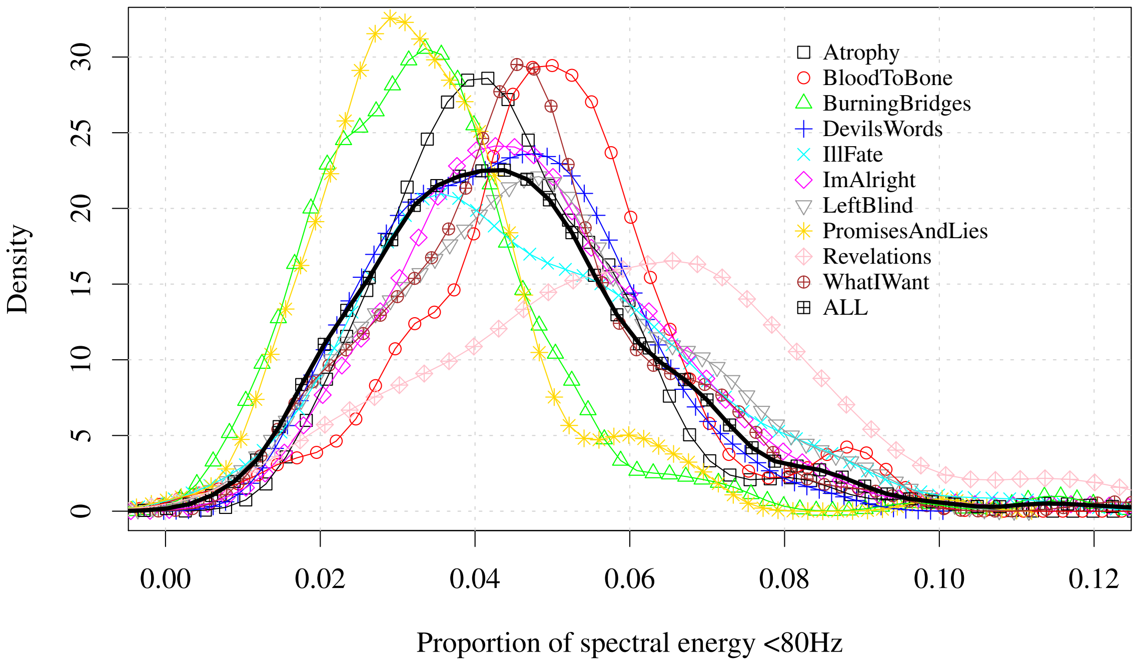


Figure 6.7: KDE of LF energy in 1466 mixes. Notable inter-song differences in LF energy

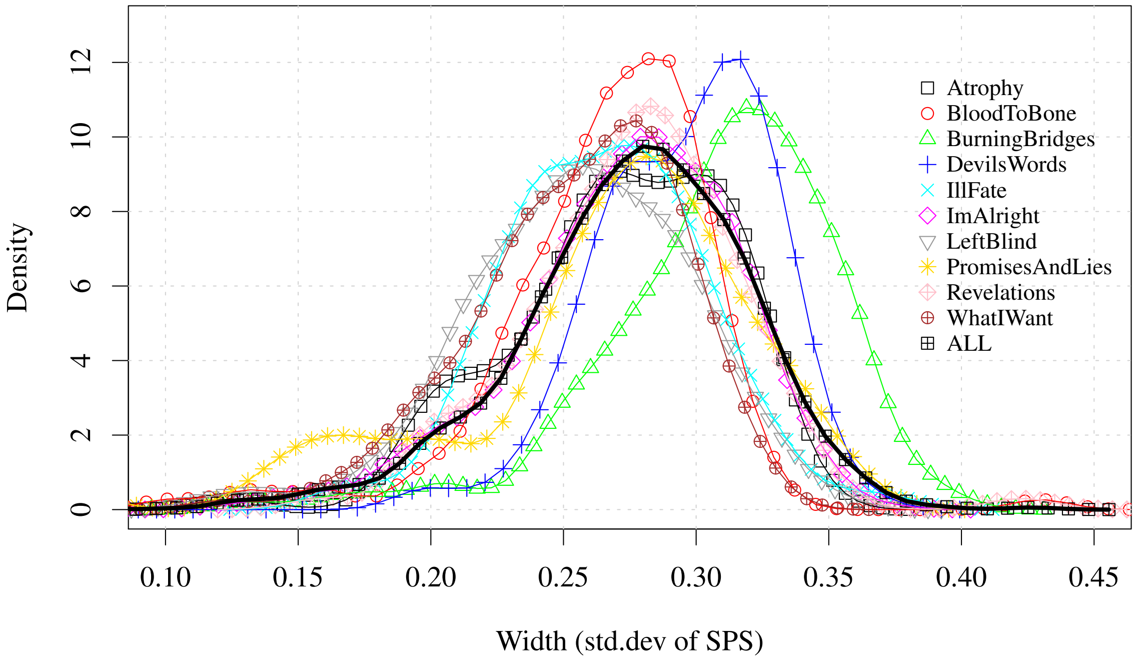


Figure 6.8: KDE of width in 1466 mixes. Most mixes occupy a narrow range of width values. Here the feature used is the value of width over all frequencies. Note that a value of 0 represents a mono mix.

Table 6.5: Results of Shapiro-Wilk test, where $p < 0.05$ indicates that the distribution is not normal. N is the number of samples in each group.

Group	AT	B2B	BB	DW	IF	IA	LB	P+L	RV	WI	ALL
N	100	101	183	138	118	373	97	125	103	129	1467
SpecCent	0.067	0.011	0.000	0.045	0.000	0.000	0.017	0.000	0.000	0.091	0.000
SpecSpread	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.049	0.000	0.000
SpecSkew	0.013	0.000	0.050	0.055	0.148	0.000	0.051	0.000	0.002	0.003	0.000
SpecFlat	0.008	0.082	0.543	0.008	0.000	0.004	0.128	0.001	0.001	0.116	0.000
SpecKurt	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
SpecEnt	0.591	0.666	0.450	0.673	0.023	0.781	0.289	0.214	0.079	0.688	0.000
CF	0.452	0.015	0.000	0.000	0.010	0.062	0.033	0.014	0.580	0.001	0.000
LoudITU	0.001	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.008	0.000	0.000
Top1dB	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Harsh	0.030	0.000	0.001	0.248	0.470	0.001	0.675	0.002	0.250	0.013	0.000
Sub80	0.000	0.000	0.000	0.168	0.009	0.000	0.355	0.000	0.004	0.000	0.000
RO85	0.004	0.012	0.020	0.482	0.000	0.361	0.010	0.008	0.000	0.290	0.000
RO95	0.440	0.007	0.076	0.121	0.010	0.010	0.223	0.793	0.014	0.596	0.000
Gauss	0.000	0.000	0.000	0.000	0.000	0.000	0.006	0.000	0.000	0.000	0.000
PMFcent	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
PMFflat	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
PMFspread	0.403	0.007	0.004	0.000	0.124	0.001	0.791	0.059	0.084	0.428	0.000
PMFskew	0.000	0.055	0.000	0.602	0.079	0.000	0.000	0.001	0.000	0.017	0.000
PMFkurt	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
W_all	0.002	0.000	0.000	0.000	0.936	0.000	0.258	0.000	0.095	0.000	0.000
W_band	0.000	0.000	0.000	0.000	0.096	0.000	0.069	0.000	0.001	0.000	0.000
W_low	0.196	0.000	0.012	0.013	0.206	0.000	0.039	0.009	0.001	0.072	0.000
W_mid	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.011	0.000	0.000
W_high	0.000	0.000	0.000	0.000	0.212	0.000	0.061	0.000	0.001	0.000	0.000
SMratio	0.000	0.003	0.017	0.000	0.061	0.000	0.000	0.049	0.000	0.000	0.000
LRimbalance	0.000	0.385	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
sbflux1	0.000	0.001	0.000	0.121	0.000	0.000	0.001	0.269	0.000	0.005	0.000
sbflux2	0.331	0.001	0.124	0.248	0.159	0.000	0.044	0.378	0.001	0.097	0.000
sbflux3	0.740	0.002	0.058	0.326	0.065	0.000	0.971	0.022	0.024	0.022	0.000
sbflux4	0.040	0.001	0.009	0.294	0.106	0.000	0.042	0.002	0.095	0.073	0.000
sbflux5	0.147	0.007	0.058	0.083	0.871	0.000	0.365	0.345	0.066	0.041	0.000
sbflux6	0.068	0.005	0.281	0.176	0.558	0.000	0.268	0.127	0.013	0.121	0.000
sbflux7	0.042	0.002	0.088	0.039	0.421	0.000	0.125	0.205	0.173	0.085	0.000
sbflux8	0.018	0.006	0.099	0.027	0.334	0.000	0.166	0.001	0.027	0.000	0.000
sbflux9	0.003	0.002	0.000	0.000	0.001	0.000	0.080	0.375	0.000	0.000	0.000
sbflux10	0.000	0.000	0.000	0.000	0.000	0.000	0.018	0.003	0.000	0.000	0.000

Table 6.6: GMM parameters for distributions of all 1501 mixes. R^2 is the coefficient of determination describing the fit of $(g1+g2)$ to the KDE curve. μ_1, μ_2, σ_1 and σ_2 are given in the units of the variable.

Feature	λ_1	λ_2	μ_1	μ_2	σ_1	σ_2	R^2
SpecCent	0.945	0.055	3532	4880	659	794	0.998
LoudITU	0.526	0.474	-12.910	-8.511	3.672	1.801	0.993
LF	0.877	0.123	0.042	0.071	0.015	0.026	0.989
Width	0.118	0.882	0.223	0.281	0.073	0.037	0.995

Table 6.6, according to Equation 6.1.

$$R^2 = 1 - \frac{SSR}{SST} \tag{6.1}$$

This indicates the proportion of the estimated density that can be explained by the model where $n = 2$. As this value is close to 1 in all cases it can be said that the sum of just two Gaussian functions well-approximates the estimated densities.

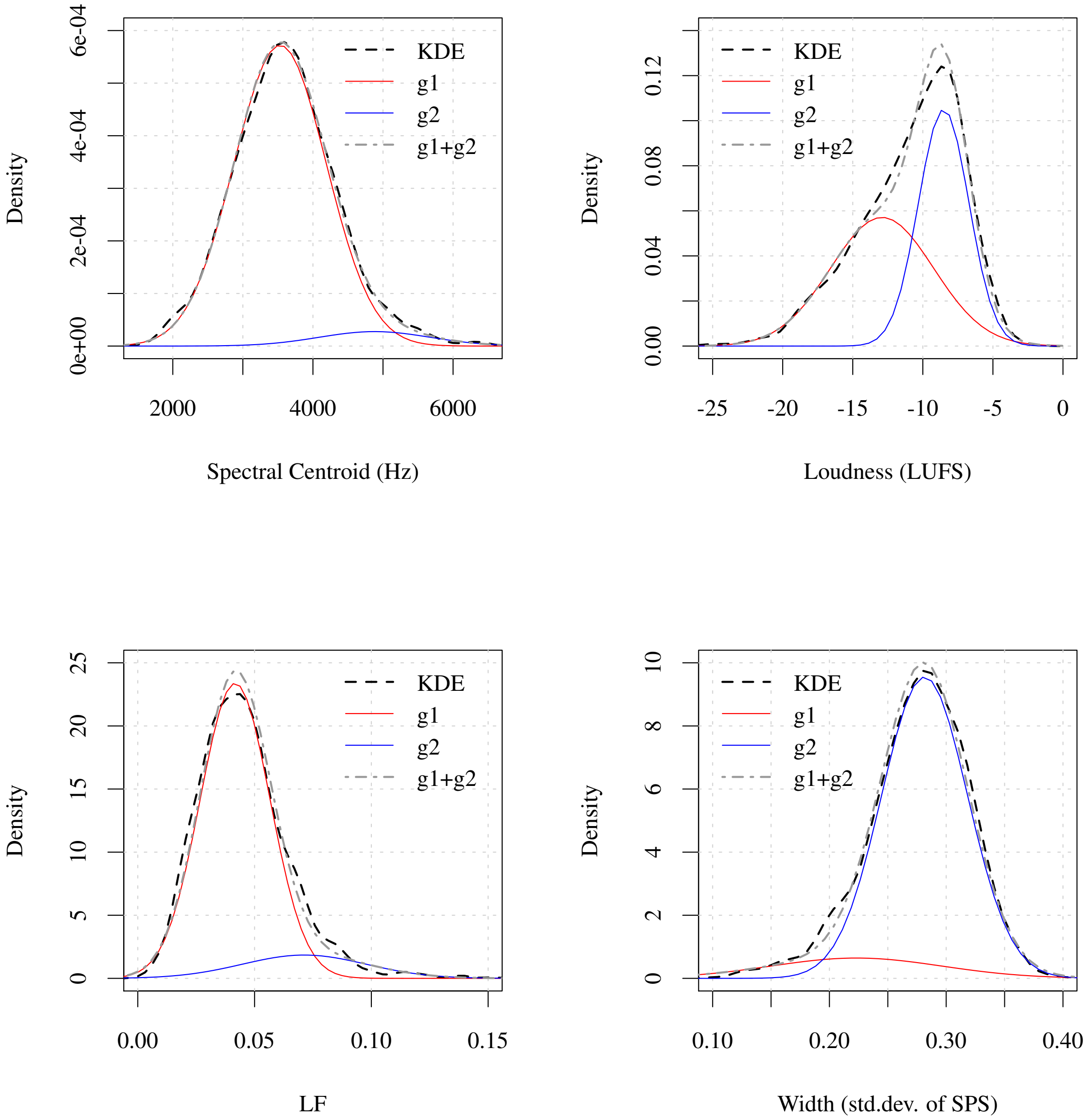


Figure 6.9: GMM parameters from Table 6.6. The filled, dashed curve represents the estimated density and the solid curves represent the GMM. While *Loudness* shows a bi-modal distribution, *Spectral Centroid*, *LF Energy* and *Width* are well characterised by a single Gaussian function.

6.1.6 Comparison with random mixes

For all three songs for which random mixes were created (see Chapter 5) the distribution of spectral centroid spans a lower range when informed by the mix-space results (when using Eqn. 5.7 instead of Eqn. 5.5). For “Burning Bridges”, as shown in Fig. 6.10a, the typical spectral centroid of the real mixes is noticeably lower than the random mixes. The distribution for real mixes is positively skewed, with a large number of mixes with higher spectral centroids. When informed by the mix-space result, the distribution of spectral centroid closely approximates the distribution of real mixes. Conversely, for both “I’m Alright” and “What I Want”, shown in Figs. 6.10b and 6.10c, the naive random mixes provide distributions closer to those of the real mixes, and it is the informed result that underestimates the mean of the real mixes.

This varied result highlights one specific issue with the data collection; there is no clear indication of mix-quality other than the assumption that all engineers created mixes in line with their intent. This intent may not necessarily be “best-practice” in the case of many amateur mixers. There is clearly a need for a specific study on the relationship between audio signal features in mixes and the perception of quality in those mixes. In Chapter 7 it is shown that one characteristic of mixes that are perceived to be low-quality is that they are also perceived to be particularly bright. This particular result may explain some of the inconsistency in the results above — that if there is a large number of poor mixes, or simply mixes of varying quality, then the distributions of real mixes may be hard to consistently predict.

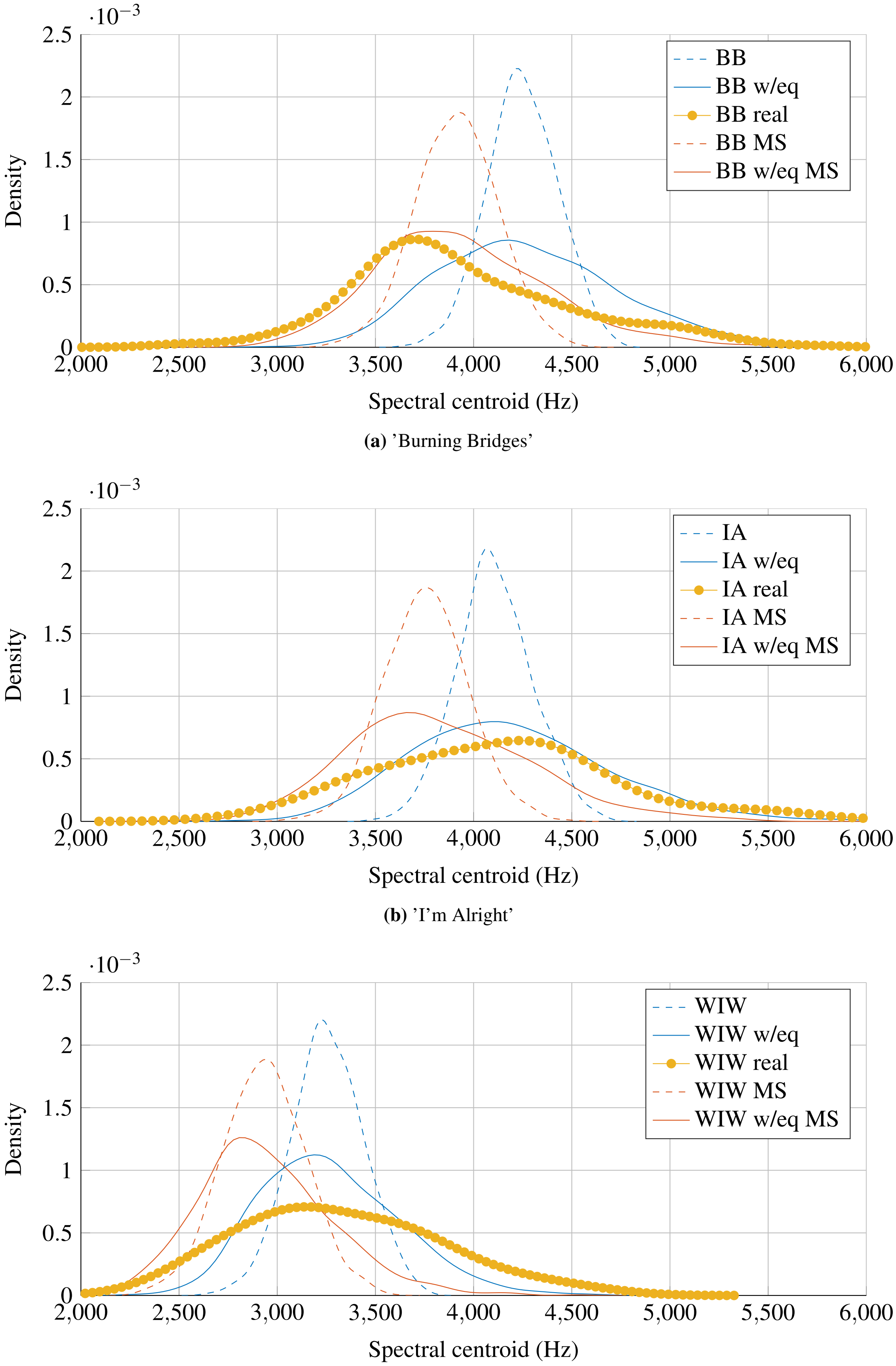
When comparing the KDE measurements, for all three different songs, it is clear that the addition of EQ results in a more similar distribution of spectral centroid, in terms of the breath of the distribution and the value of the maximum density, which, of course, are correlated since the area under the curves are equal.

The stereo width measurements of real mixes are compared to those of random mixes in Figs. 6.11a, 6.11b and 6.11c. From these comparisons it is shown that the variance of the distributions matches best for method #3. The central values of the distributions are, for all methods, lower than the real mixes. This is likely due to the fact that the methods for generating random pan positions do not readily allow for hard-panning of instruments.

6.1.7 Discussion

Before now there have not been any studies looking at feature variance over such a large number of alternative mixes of the same songs, and so this chapter makes a significant contribution to knowledge. In this study, the features extracted were amplitude-based, spectrum-based or spatial features. Over all 10 songs considered, the dimensions of variation revealed by the PCA were described as ‘amplitude’, ‘brightness’, ‘bass’ and ‘width’, in order of variance explained.

This shows that all songs, within their range of mixes, varied in terms of their perceived loudness and dynamics. Figure 6.3 shows certain songs with distinct dynamic range values when compared to other songs — the lowest values of dimension 1 (loud, low dynamic range) apply to songs in hard rock or metal styles, whereas the soft rock styles attain higher values along this dimension. As the data points in Figure 6.3 are spread out over the space, and not definitively grouped by song, it is observed that any one song can be mixed with the overall loudness/dynamics or brightness of any other song. Despite this, trends are apparent. The song ‘Revelations’ had the highest average value of *dim.2*, meaning the least amount of brightness. This may be due to the fact that the multitrack content was recorded in 1975, therefore the digital audio used here was sourced from an analogue tape. While little is known about the precise recording conditions, it is likely the reduced high-frequency content in mixes of this song was due to the limitations of the recording technology used at the time. Additionally, when creating mixes of this song, it is possible that engineers were inspired to use era-specific mixing techniques, either consciously or subconsciously, similar to the anchoring effect demonstrated in § 4.6.1. The song with the lowest values of *dim.2* (the brightest mixes) is ‘I’m Alright’, which features acoustic guitars and

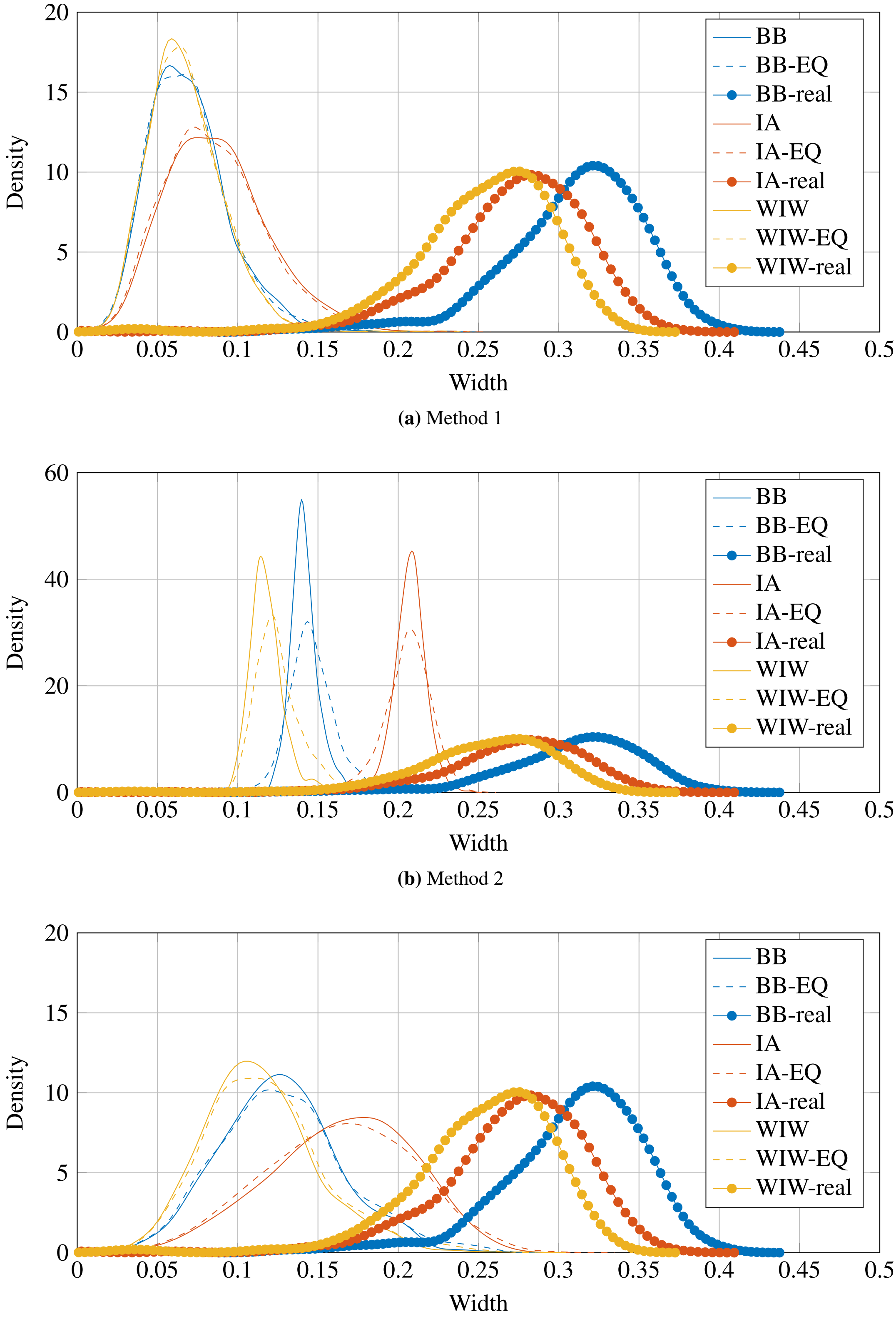


(a) 'Burning Bridges'

(b) 'I'm Alright'

(c) 'What I Want'

Figure 6.10: KDE of spectral centroid for 3 songs. Five conditions are shown: naive, naive w/eq, real mixes, informed and informed w/eq.



(a) Method 1

(b) Method 2

(c) Method 3

Figure 6.11: KDE of width for all three methods. Three conditions are shown for each of the three songs: without EQ, with EQ and real mixes.

shakers, both instruments with emphasis on high frequencies. *Dim.3* is difficult to interpret as it represents emphasis on bass or treble frequencies depending on the value, and there is little inter-song difference. Mixes of the song 'Promises and Lies' tended to have a higher concentration of spectral energy between 2 kHz and 5 kHz than other songs, or a lack of spectral energy below 80 Hz. There is little observed difference in the group centroids along *dim.4*, which represents stereo width, particularly at low frequencies, as expected.

Feature distributions suggest multi-modal behaviour, often dominated by one specific mode, which is dependent on the song. This distribution holds well for the songs considered, providing evidence for central tendency or even "optimal" values. In Figure 6.5, typical values of *Spectral Centroid* differ from song to song, suggesting each song has a range of possible values which can be tolerated, based on the arrangement, instrument timbre, key etc. The distribution of *Loudness* values in Figure 6.6 is quite similar from song to song. This is a possible side effect of the fact that many mixes were subjected to mastering-style processing, particularly heavy dynamic range processing. Figure 6.7 indicates that the proportion of spectral energy below 80 Hz is reasonably consistent from song to song, with some variation. This is possibly dependent on the key of the song, the precise arrangement and the relationship between bass guitar and kick drum performances. *Width* distributions shown in Figure 6.8 are similar for each song, occupying a narrow range of values. It was found that songs were mixed with a very wide range of panning conditions, from mono to wide stereo. However, central tendencies can be observed with clear distributions around them. This result indicates that panning conventions are applied similarly in all songs, restricted by the medium of two-channel stereo reproduction, and that a central tendency is observed.

6.1.7.1 Implications for intelligent music production

By examining a large dataset of mixes, from hundreds of individual mix-engineers of varying skill levels, the results here indicate the dimensions over which mixes vary and the amounts by which they vary in these dimensions. This could help to inform targets and bounds for intelligent mixing tools. For example, Figure 6.9 and Table 6.6 suggest that values of *Spectral Centroid* are normally distributed with a mean of $\approx$ 3.5 kHz and standard deviation of $\approx$ 660 Hz. Consequently, and also shown by Figure 6.5, few mixes would have a *Spectral Centroid* value below 2 kHz, although there may exist specific, context-dependent productions where this is possible, such as when analogue recording media are utilised. The results in Table 6.6 could inform a system which monitors the mix, in an automatic or human-operated system, and offers advice when the values of certain features deviate strongly from expected values (a version of this is described in Chapter 8).

Interestingly, while the average distribution of features for *Spectral Centroid*, *LF* and *Width* were all well described by a single Gaussian distribution, *Loudness* was best described by a combination of two Gaussians. This might be explained by the fact that some engineers tend to maximise the loudness of their mixes whilst others will be more concerned with maintaining a greater dynamic range. These differing strategies appear to be revealed by this GMM statistical analysis.

6.1.7.2 Implications for music information retrieval

In a number of tasks in Music Information Retrieval (MIR), feature-extraction is used as a means of characterising audio data, so that each data point, representing a song or instrument, can be described in a meaningful way. For example, when attempting to train a classifier to perform genre prediction, each song is labelled as belonging to a specific genre and features are extracted from each song. The assumption is that the features can be used to represent useful attributes of that song, and thus, its genre. However, perhaps the features only represent attributes of the recording of the song and not the song itself.

In this study, where there are hundreds of alternate mixes of a given song, we can see that these features do not clearly distinguish between songs. What are the implications then for tasks such as genre prediction? If a classifier was developed with α songs in genre A and β songs in genre B, how would the performance of the classifier change if alternate mixes were substituted for all $(\alpha + \beta)$ songs, or for all possible permutations of classifier that could be made from hundreds of alternative mixes?

Of course, this problem is simplified should estimated tempo be included, as the tempo of a song does not typically change with mix. However, as it has been shown in Chapter 5, the ability to correctly estimate tempo can depend on the mix. A detailed study on rhythm in multitrack mixes would be useful in furthering our perception of why certain music mixes are created. This is left to further work.

6.2 Case study: an on-line mix competition

Of the list of mixes in Table 6.1, a particular subset has been evaluated qualitatively. From this a quantitative evaluation can be inferred. This section describes a dataset containing 101 mixes of a multitrack session, which is Dataset #3 within this thesis.

6.2.1 Dataset #3 — 101 mixes

During March and April of 2011, an on-line mix-competition was held in which entrants were asked to mix a provided 23-track session, for the song 'Blood To Bone' by the band 'Young Griffo'. Along with the one original mix, created with input from the artists, 100 submitted mixes are currently hosted on-line[1]. In total, 73 individuals took part. When a contestant submitted a mix, a review was provided by a mix-engineer who, having authored a number of texts on the subject [190, 191], can be considered as an expert. After reading this review, a number of participants then decided to submit a second mix. Once the deadline was passed, a number of mixes were *shortlisted*, while others were given an *honourable mention*. From the *shortlisted* mixes a poll was then created for forum members to vote on their favourite mix. The *winner* and two *runner-up* mixes were chosen by the band. As a result, the 101 mixes can be classified into five categories which represent the level of success the mix attained in the competition, shown in the first group of Table 6.7. These 101 mixes make up 101 of the 135 mixes for this song that are shown in Table 6.1. As such, pre-processing of the audio and subsequent feature-extraction is described in § 6.1.1.

Table 6.7: Categorisation of 101 mixes, showing the number of mixes in each category

Category	*Number of mixes*
Winner	1
Runner-up	2
Shortlisted	6
Honourable Mentions	18
Rejected	74
Original mix	1
Mixes with reviews	64
Mixes without reviews	36
Only mix	49
First mix	26
Second mix	26

Additionally, the spectrum of each segment was determined using a constant-Q transform with q points per octave. Each spectrum was normalised with respect to energy, i.e. each magnitude is divided by the euclidean norm (root sum of the squared magnitude). The mean and standard deviation at each frequency along the vector F were determined. The standard deviation of the spectra is shown in figure 6.12, along with a smoothed curve determined by a moving average filter with a length L calculated according to equation 6.2. The value of q was set to 24, which

[1] As of January 2017, the data can be found at `http://www.cambridge-mt.com/YoungGriffoCompetition.htm`

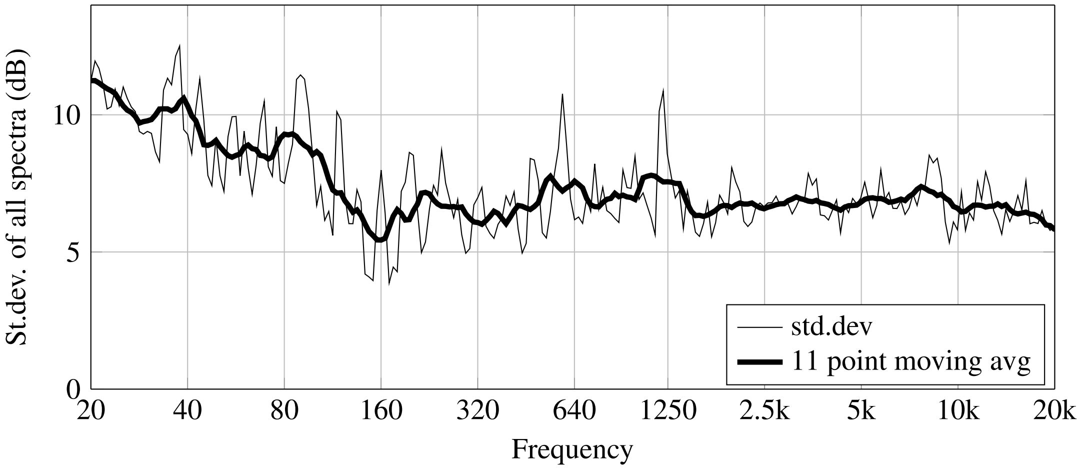


Figure 6.12: Analysis of frequency response. The standard deviation of the 101 spectra is displayed, showing increased variance at lower frequencies.

produced moving average filter with a length $L = 11$ calculated according to equation 6.2.

$$L = \lceil \frac{\text{length}(F)}{q} \rceil \tag{6.2}$$

From Fig. 6.12 it can be seen that the variation in spectra is reasonably consistent from 200 Hz and above, between 6 and 8 dB. There are a number of mid-range frequencies which show higher variance — since the vocal melody is rather simple, containing very few notes, variation in vocal level is likely to be the cause. The increased variance in the spectrum at lower frequencies is likely to be due to variation in reproduction equipment and room acoustics.

6.2.2 Factor analysis

The mix competition acted as a qualitative and subjective analysis of the set of mixes, as reviews were written and mixes were ranked as shown in Table 6.7. The following is a quantitative, objective analysis of the dataset. The methodology here is almost identical to § 6.1.4, however, only the 101 mixes of this one song are included in the analysis. This allows the resulting components to be directly compared to the subjective rating (the competition outcome).

Outlier detection was performed in the 36-dimensional feature-space. The Z-score of each point was determined using the Euclidean distance to the three nearest neighbours and those where $Z > 2.5$ were deemed outliers. This led to the removal of three mixes, all members of the lowest quality group. The total amount of signal features was 36, while there were 98 individual mixes remaining after removal of outliers. The appropriateness of PCA was tested as follows, using SPSS. Using Bartlett's test of sphericity, the null hypothesis that the correlation matrix of the data is equivalent to an identity matrix was rejected.

$$\chi^2(561, N = 101) = 7002, p < 0.001$$

This indicates that factor analysis can be performed, while a Kaiser-Meyer-Olkin measure of sampling adequacy of 0.808, above the recommended value of 0.6 [133], suggests that factor analysis would be useful. The communalities were all above 0.3, further indicating that each variable shared some common variance with others. As a result of these tests, PCA was conducted with all

36 variables. Each variable was standardised prior to PCA, i.e. mean $\mu = 0$ and standard deviation $\sigma = 1$. This initial PCA is unrotated and there was no limit on the number of components. The plot of eigenvalues is shown in Fig. 6.13. Using the `nFactors` package for R [136] a variety of methods were employed in order to determine the number of components to keep in further analysis. Kaiser's rule [137] suggests retaining those components with eigenvalues greater than 1, which in this case was the first seven components. The acceleration factor [136] determines the knee in the plot by examining the second derivative — due to the fact that components 2 and 3 have similar eigenvalues much lower than component 1, this method chose to retain only the first component. The optimal coordinates method [136] suggested that the first four components be kept, as indicated by Fig. 6.13. Parallel analysis [139] agreed that the first four components were suitable to retain, also shown in Fig. 6.13. Additionally, these four components have eigenvalues greater than one. As a result, the first four components were considered in the subsequent analysis.

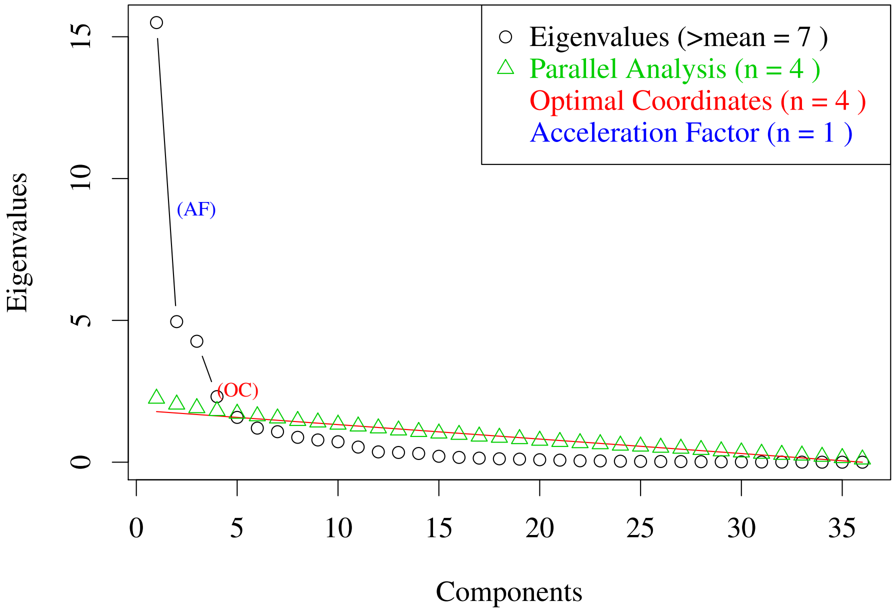


Figure 6.13: Eigenvalues of first PCA. The first four components account for approximately 75% of the total variance

Any variables which were not significantly correlated with any of the first four components (where $p < 0.05$) were removed from analysis. Subsequently, the features '*Harsh*' and '*LRimbalance*' were removed. This left 34 variables for a second PCA, this time with only four dimensions kept and rotated using the varimax method [198]. The eigenvalues of this PCA are shown in Table 6.8.

Table 6.8: Eigenvalues of revised PCA, also displayed as percentage of explained variance. Four components account for approximately 79% of the total variance.

	1st	2nd	3rd	4th
Eigenvalue	15.46	4.95	4.26	2.31
% variance	45.48	14.55	12.52	6.79
Cumulative % variance	45.48	60.03	72.55	79.33

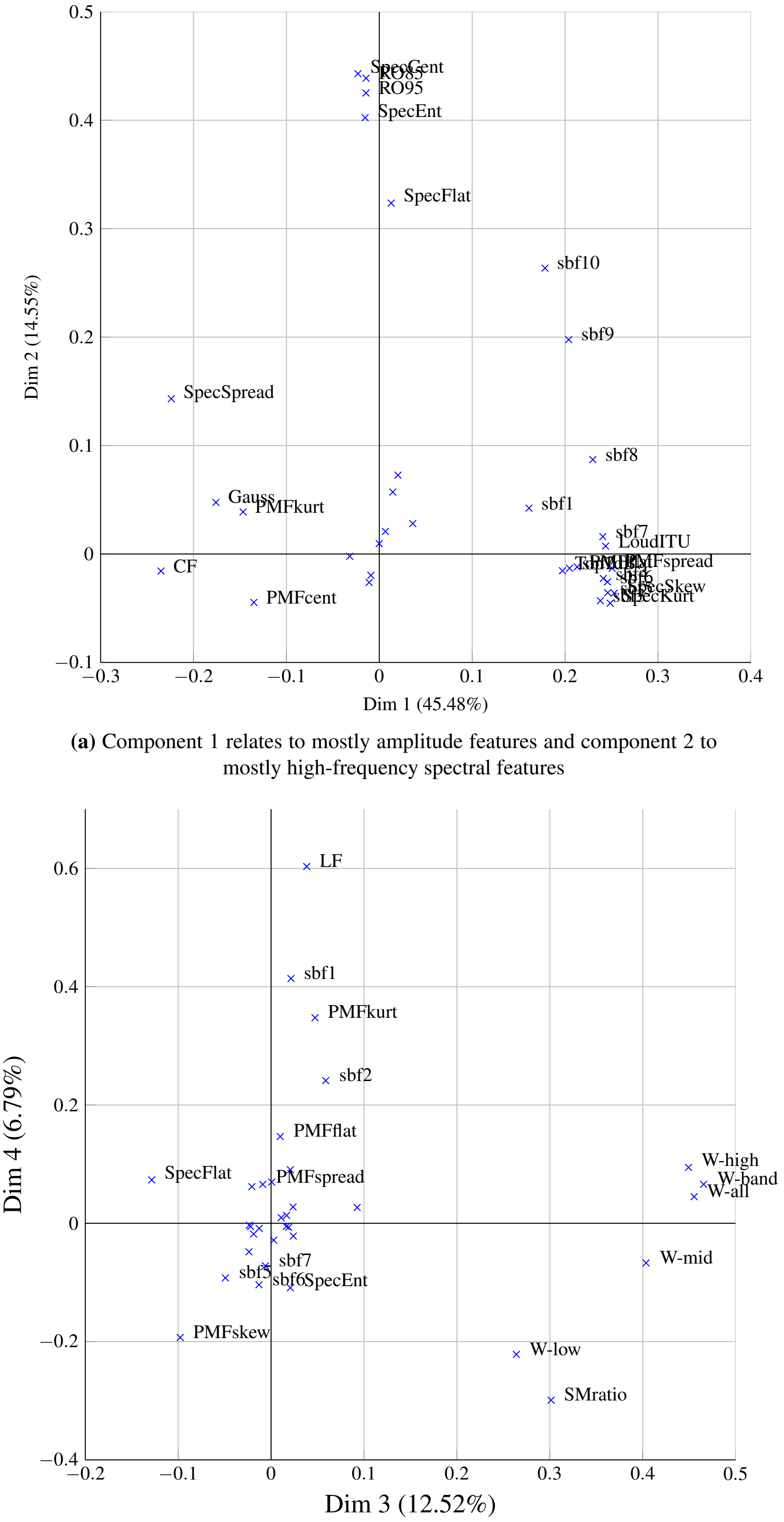
Table 6.9: Loadings of each variable to each component

Feature	**Loadings**			
	Comp1	Comp2	Comp3	Comp4
SpecCent	-0.0230	0.4429	0.0188	-0.0068
SpecSpread	-0.2239	0.1431	-0.0208	0.0619
SpecSkew	0.2529	-0.0361	-0.0221	-0.0054
SpecFlat	0.0128	0.3236	-0.1285	0.0734
SpecKurt	0.2485	-0.0454	-0.0234	-0.0033
SpecEnt	-0.0153	0.4023	0.0207	-0.1091
CF	-0.2350	-0.0157	-0.0129	-0.0087
LoudITU	0.2436	0.0073	0.0239	-0.0217
Top1dB	0.1971	-0.0154	0.0208	0.0906
LF	-0.0318	-0.0022	0.0384	0.6032
RO85	-0.0143	0.4388	0.0109	0.0095
RO95	-0.0143	0.4251	0.0006	0.0698
sbf1	0.1611	0.0423	0.0214	0.4137
sbf2	0.2048	-0.0132	0.0588	0.2412
sbf3	0.2380	-0.0431	0.0236	0.0276
sbf4	0.2411	-0.0225	-0.0238	-0.0482
sbf5	0.2458	-0.0357	-0.0492	-0.0923
sbf6	0.2456	-0.0258	-0.0130	-0.1039
sbf7	0.2406	0.0160	-0.0061	-0.0720
sbf8	0.2298	0.0870	0.0164	-0.0054
sbf9	0.2038	0.1977	0.0028	-0.0285
sbf10	0.1784	0.2636	0.0167	0.0134
Gauss	-0.1759	0.0476	-0.0188	-0.0182
PMFcent	-0.1350	-0.0446	0.0926	0.0268
PMFflat	0.2129	-0.0119	0.0098	0.1467
PMFspread	0.2506	-0.0132	-0.0090	0.0658
PMFskew	0.0145	0.0571	-0.0977	-0.1931
PMFkurt	-0.1466	0.0388	0.0473	0.3476
W-all	0.0066	0.0208	0.4554	0.0450
W-band	-0.0091	-0.0197	0.4655	0.0659
W-low	0.0360	0.0281	0.2640	-0.2216
W-mid	-0.0001	0.0096	0.4036	-0.0670
W-high	-0.0110	-0.0262	0.4493	0.0944
SMratio	0.0201	0.0727	0.3014	-0.2990

The following is an interpretation of each of the first four components and is based on the loadings of the individual features, as shown in Fig. 6.14a and 6.14b.

1. Many of the input variables associated with signal amplitude dynamic range and loudness are strongly correlated with the first principal component, with positive values indicating louder, more compressed mixes (see Fig 6.14a).

2. The second component can be described by the many strong correlations to spectral features with positive values denoting mixes that have a greater proportion of energy in higher



(a) Component 1 relates to mostly amplitude features and component 2 to mostly high-frequency spectral features

(b) Component 3 relates to mostly spatial features and component 4 to mostly low-frequency spectral features Labels for loadings < 0.1 are removed for clarity.

Figure 6.14: PCA variables factor map for 98 mixes, showing loadings of variables on varimax-rotated principal components.

frequencies (see Fig 6.14a).

3. Component 3 can be explained by the correlation of the spatial features to this component: as the value of this component increases, so to does the perceived width of the stereo image (see Fig 6.14b).

4. Features associated with low frequencies are more strongly loaded onto component 4 (see Fig 6.14b).

It can be seen from Figure 6.15 that, when the rotated principal components 1 to 4 are considered, the winning mix lies close to the centre of this space, at the position [0.72, 0.43, 2.14, 0.16]. This shows that the winning mix is, when compared to all other mixes, an example of a mix in which the concepts of loudness and spectral balance are each well-balanced, while having one of the widest stereo images. It is worth noting that one of the runner-up mixes is located very close to the winning mix, having a similar balance between loudness, spectrum and width. This suggests a consistency in the decision-making process which selected the "best" mixes.

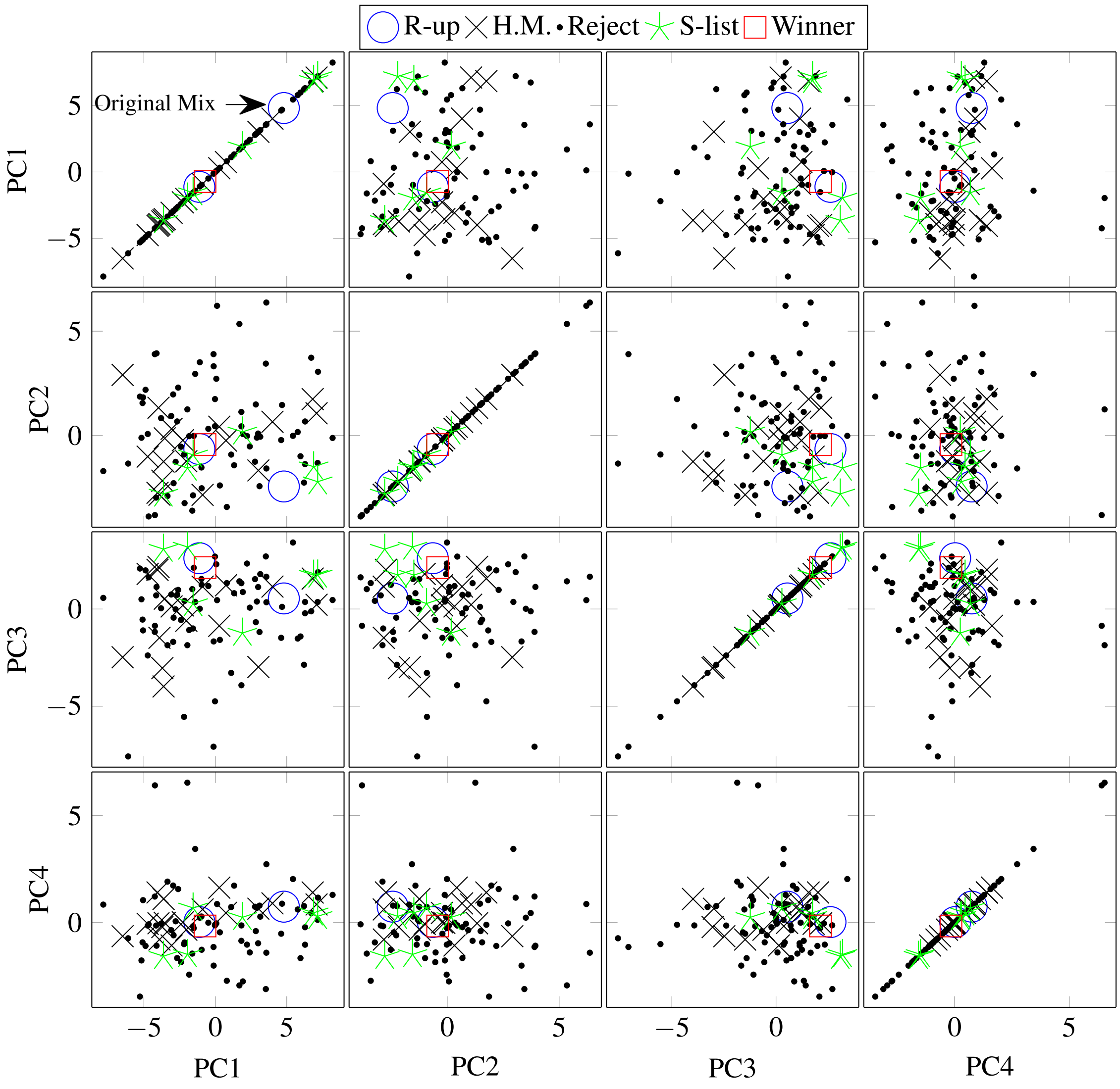


Figure 6.15: Individuals factor map — matrix of scatter plots showing each mix in the space of the first four rotated principal components, with mixes grouped by quality.

6.2.3 Ordinal logistic regression

It can be seen from Fig. 6.15 that the more highly ranked mixes have lower values of PC2 and higher values of PC3. To test the influence of these dimensions on the competition outcome an Ordinal Logistic Regression model was used. Rather than use the five categories shown in Table 6.7, the winning mix and the runner-up mixes were re-combined with the 'shortlisted' category. This forms three groups, as shown in Table 6.10. Table 6.11 shows the β and p values of the model, along with the *Odds Ratio*.

Table 6.10: Categorisation of 98 mixes into three groups

Quality	*Number of mixes*
High (q=3)	9
Middle (q=2)	18
Low (q=1)	71

Table 6.11: Parameter estimates for ordinal logistic regression model, with significant results in **bold**.

Type	*Var*	β	p	*Odds Ratio*
Threshold	$(q=2)/(q=1)$	.487		
	$(q=3)/(q=2)$	2.064		
Location	PC1	-.156	.569	0.856
	PC2	-.786	**.043**	0.456
	PC3	.659	**.034**	1.933
	PC4	.757	.149	2.132
	$PC1^2$	.141	.560	1.151
	$PC2^2$	-.519	.179	0.595
	$PC3^2$	.185	.276	1.203
	$PC4^2$	-1.518	**.023**	0.219

Behind the OLR model is the proportional odds assumption, which can be summarised as follows: the difference between groups is the same and the chance of moving from one group to the adjacent group is the same. As the task of assigning each mix to a group is perceptual, it is difficult to test this assumption. This assumption may limit the accuracy of the model. In addition to each of the principal components, the squared components were also used in the model, which also tests whether quality changes when moving away from a value of zero.

Components 2, 3 and 4 are shown to have significance in the model. As the *Odds Ratio* for $PC4^2$ is ≈ 0.22 this suggests a 78% chance of a drop in quality being observed for each unit step away from a value of 0. This suggests an optimal level at which to balance the low-frequency content of this song. The values of PC2 and PC3 indicate an approximate halving or doubling of the chance of a change in quality being observed with each unit increase of their respective values. By considering these quality groups as ordinal categories, each point in the space can be assigned a quality value by means of interpolation. Figure 6.16 shows the result of interpolating quality values across the space of PC2 and PC3. These two dimensions were chosen for illustrative purposes as they were both significant predictors in the OLR model. The interpolation is a two-dimensional cubic interpolation, as such there are some regions where *quality* is less than 1.

Rather than simply looking at the individual dimensions of the PCA, additional insights were obtained by reducing the four dimensions down to two by means of multi-dimensional scaling (as in Chapter 3). This is displayed in a surface plot of the reduced, three-level model is shown in Fig 6.17. These surfaces represent the fitness landscape for mixes of this song, as perceived by the competition judge. There is a noticeable region of the landscape (positive values of both dimensions) in which all mixes are of low quality. Yet there are multiple peak and ridges which represent high quality, which have lower quality 'valleys' in between. Of course, as this surface was generated using cubic interpolation, it is smooth and differentiable (with the exception of the boundaries). Since the axes used are MDS dimensions it is difficult to directly interperate their meaning. This is, however, not dissimilar to the nature of the psychological space in which mixes are evaluated, which is a series of perceptual factors (see Fig. 2.11). Due to the subjective nature of audio perception (refer back to Chapter 3) this is but one of many possible fitness landscapes. Ultimately, when evaluating a series of possible mixes, one is performing the evaluation based on one's own personal fitness landscape — this task is the the focus of Chapter 8 and 9.

From this ordinal logistic regression model it is shown that a mix had greater chance of scoring well in the competition if the spectral balance was not overly bright and the bass frequencies were well-balanced in level. It is also clear that preference correlates well with the width of the mix. Overall, amplitude-based features did not significantly influence the decision. At this stage it is important to note the following comment from the competitions review-writer:

> *"I also stipulated that the loudness of the mix would not be a contributing factor in the competition judgement."*

Further discussion can be found in Wilson and Fazenda [194].

6.2.4 Explicit ratings of Like/Quality

The ratings provided in this competition indicate a coarse categorisation of 'quality'. A subjective listening test was undertaken in order to obtain a finer grading of quality. As the total complement of 101 mixes would make a listening test impractically long, only the highest-scoring 27 mixes were used, by omitting the lowest-scoring category in Table 6.10. This experiment was designed to be analogous to the experiment in Chapter 3. In this case, ratings of like and quality were provided on a 5-star scale but short descriptions were requested for *both* like and quality ratings. This test design forces participants to consider their responses, while allowing an experimenter to examine the meaning behind the ratings provided. Similarly to the experiment in Chapter 3, it was hypothesised that like and quality ratings are correlated yet explained by separate factors. For this reason, in this experiment, descriptions of both like and quality ratings were obtained. The test interface was designed to be similar to the interface used in Chapter 3 (see Figure 3.2). There was no need to assess familiarity in this case, as all audio samples represent the same song. Here, for each audio sample, four questions were posed.

Q1. How much do you like this mix?

Q2. Describe an aspect of the sample on which you assessed the LIKE RATING of this sample.

Q3. How highly do you rate the quality of this sample?

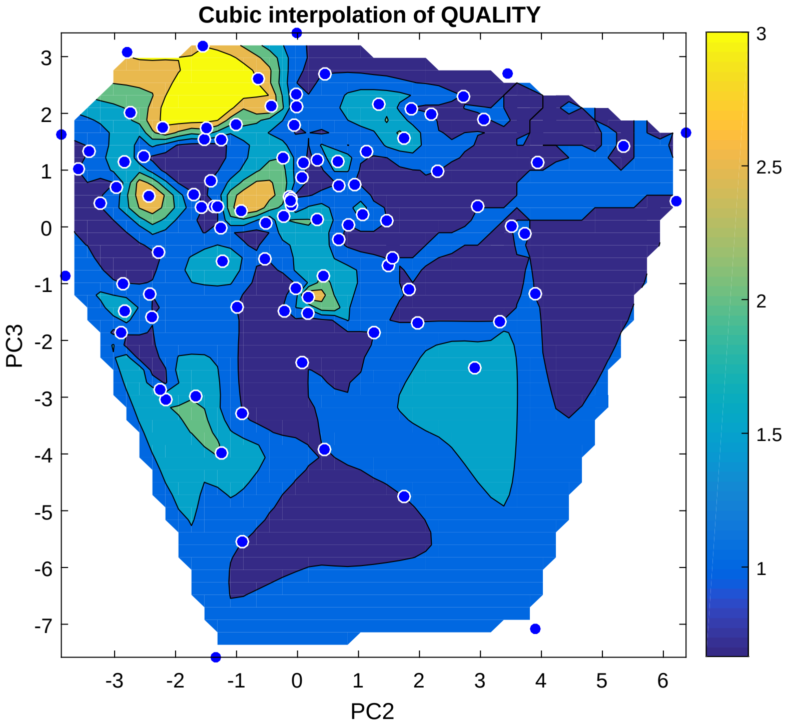


Figure 6.16: Individuals factor map, with interpolated quality values.

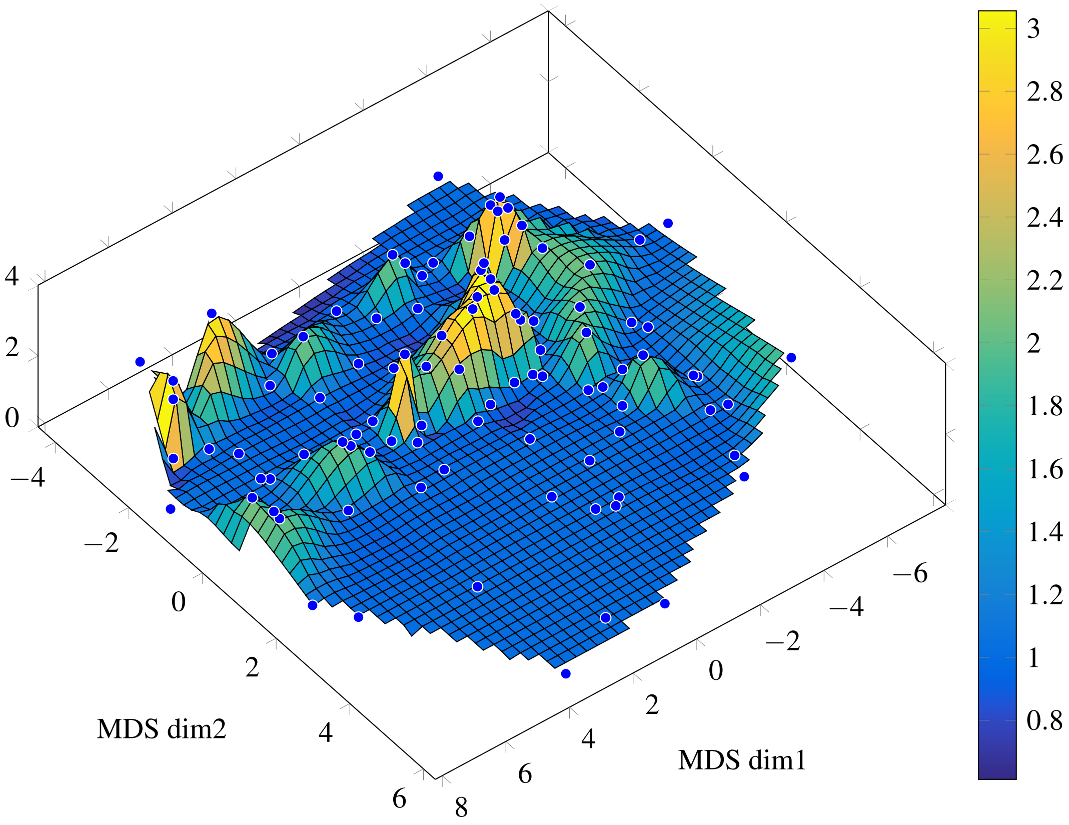


Figure 6.17: MDS of PCA individuals factor map - with interpolated quality contours. This is based on the 3-level model used in the ordinal logistic regression.

Q4. Describe an aspect of the sample on which you assessed the QUALITY RATING of this sample.

The test location and audio reproduction system was identical to the experiment in Chapter 3. A brief summary is as follows. The test took place in a BS.1116 listening room, while audio was reproduced using headphones (Sennheiser HD 800) connected to the test computed by a Focusrite 2i4 USB audio interface. Headphone equalisation was as Chapter 3. One clip was used at the beginning of each test to serve as a trial and from there on the order of playback was randomised. For the listening test a one-second fade-in and fade-out were applied and each sample was loudness-normalised, according to [32]. A break was automatically suggested when 40% of the trials were completed. Ultimately, the median duration of the experiment was 44 minutes, not including the scheduled break. As the test contained this option of a short break, any effects of fatigue on the reliability of subjective quality ratings were considered to be negligible [122].

6.2.4.1 Test Panel

The total number of participants was 13 (5 of whom had participated in the experiment in Chapter 3, although 24 months had passed since that participation). The age of participants ranged from 19 to 41 years, with a median of 25 years. Participants were asked how many previous listening tests they had participated in. From these responses seven participants were classed as experienced listeners (having completed over 10 similar listening tests) and six participants as not-experienced (having completed less than 10 similar listening tests). No participants reported any hearing difficulties. In a post-test question, participants were asked if the playback level was "louder", "about the same" or "quieter" compared to the level at which they would normally listen to similar music over headphones. From the responses (5 louder, 4 same and 4 quieter) it can be observed that the playback volume was suitable for the test.

6.2.4.2 Results

The influence of the audio sample on the assessment of quality and like ratings was measured using a multivariate analysis of variance (MANOVA). The assumptions for MANOVA were tested using Box's test of equality of co-variance matrices (the Box's M value of 51.801 was associated with a p-value of 0.996, interpreted as non-significant) and using Bartlett's test of sphericity, which was significant

$$\chi^2(2, N = 351) = 173.978, p < 0.001$$

Using Wilks' Λ, there was a significant effect of audio sample on the ratings of like and quality

$$\lambda = 0.745, F(52,646) = 1.974, p < 0.001$$

For Wilks' Λ, the effect size is calculated as follows:

$$\eta^2 = 1 - \Lambda^{1/s}$$

where s = (the number of groups$-$1) or the number of dependent variables, whichever is smaller. The effect size is 0.137, which can be considered as a medium effect [202, 203]. The remaining variance is accounted for by variables not measured. This may include musical taste or experience as an audio engineer, however the small number of participants makes this further analysis difficult.

In a follow-up univariate analysis of variance (ANOVA) the following results were obtained. There was a significant main effect of the audio sample on like ratings

$$F(1,26) = 3.45, p =< 0.001, \eta^2 = 0.217$$

and also on quality ratings.

$$F(1,26) = 2.09, p = 0.002, \eta^2 = 0.143$$

These effect sizes can be considered to be medium. Figure 6.18 shows a scatterplot of the mean like and mean quality ratings for each audio sample, when averaged over all participants. In this experiment, it can be seen that there is significant correlation between these two ratings ($R^2 = 0.82$). Furthermore, a significant correlation is found between the like and quality ratings of each individual participant, as shown in Table 6.12.

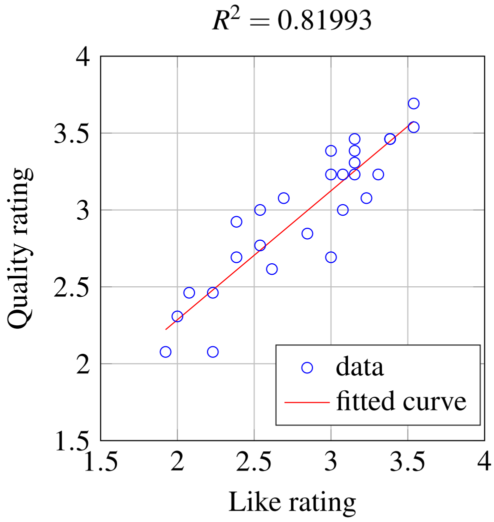


Figure 6.18: Correlation between like and quality ratings for 27 mixes of 'Blood To Bone'. Each point represents the mean like and quality rating of each audio sample.

In order to investigate the relative difficulty of each test question, the time taken to respond was measured. Figure 6.19 shows a boxplot of the results, where the marker represents the median value of the distribution while the whiskers extend to 1.5 times the interquartile range. Beyond this, outliers are marked by circles. Based on Figure 6.19 there is strong evidence to suggest that the time taken to provide a quality rating was less than the time taken to provide a like rating. There are a number of possible explanations for this:

- The time recorded for Q1 included an initial period of listening, resulting in an overestimation.

- Since quality was rated after like, the participants were familiar with the sample at this point and already had an idea about the quality rating they would give. This could have been avoided by randomising the order of the test questions, however, due to the similarity of both questions, this may have led to confusion in the participant, introducing error.

Table 6.12: Correlation between like and quality ratings, for each participant.

Participant	Pearson r	R^2	p-val
1	0.757	0.573	<0.001**
2	0.875	0.766	<0.001**
3	0.680	0.462	<0.001**
4	0.690	0.476	<0.001**
5	0.480	0.230	0.011**
6	0.907	0.823	<0.001**
7	0.741	0.549	<0.001**
8	0.729	0.532	<0.001**
9	0.847	0.717	<0.001**
10	0.429	0.184	0.026**
11	0.459	0.210	0.016**
12	0.853	0.727	<0.001**
13	0.917	0.842	<0.001**

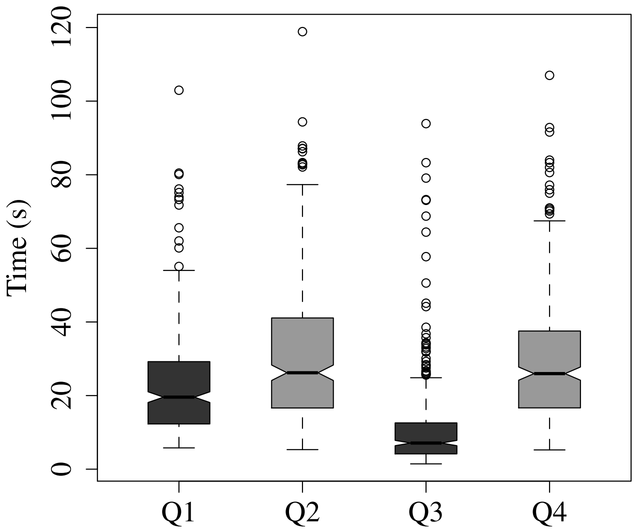


Figure 6.19: Boxplot showing the time taken in answering each of the four questions

- Like and quality ratings were explained by similar concepts, and so having already rated like, the participant could quickly rate quality.

The increased amount of time taken to provide descriptions, compared to ratings, suggests that the task required a greater level of effort. However, the time taken to provide descriptions of like ratings was comparable to the time taken to provide descriptions for quality ratings — there does not appear to be any notable difference between the effort required in providing like and quality descriptions.

The descriptions offered by participants were gathered into two corpora: one for like ratings and one for quality ratings. Text mining operations were performed using the `tm` package for R [141]. Punctuation and stopwords were removed, and stemming was performed. The word-frequencies were determined from a term-document matrix. The relative frequencies of the top 10 words, for both like and quality ratings, are shown in Figure 6.20. The terms used in the descriptions of ratings were similar for both like and quality. This further suggests that the two concepts are related, as they are explained using similar terms.

Table 6.13: Frequency count (Chi square test analysis) of comments used to describe ratings.

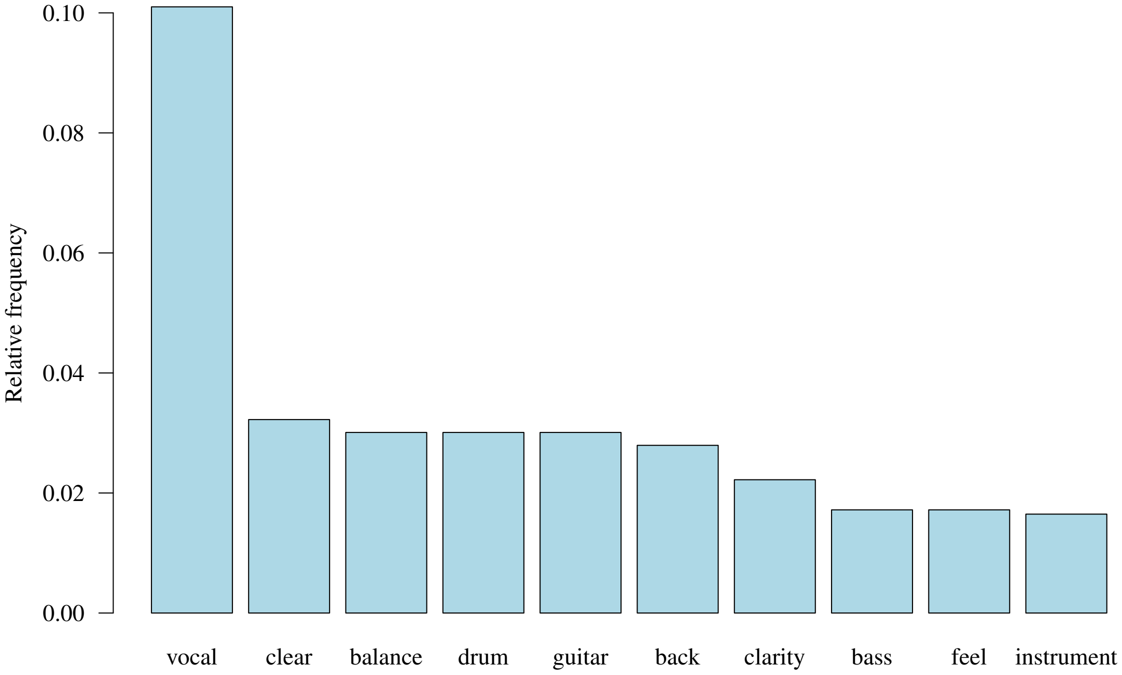
Subject of comment	Like	Quality	Total
Balance	116	152	268
Tone	**103<**	**155>**	258
Vocals	**119>**	**100<**	229
Drums	**32<**	**57>**	89
Bass	26	36	62
Panning	**14<**	**38>**	62
Guitars	27	33	60
Reverb	**29>**	**24<**	53
Dynamics	**5<**	**27>**	32
Pitch	11	9	20

There were, however, variations in how these terms were used revealed by a more detailed analysis. All subject responses were coded as being either concerned with the following subjects: “vocals”, “drums”, “guitars”, “bass”, “reverb”, “balance”, “tone”, “panning”, “dynamics” or “pitch”. For example, the comment “the reverberation of the vocal is too much” is coded as a negative comment concerned with vocals, reverb and balance.

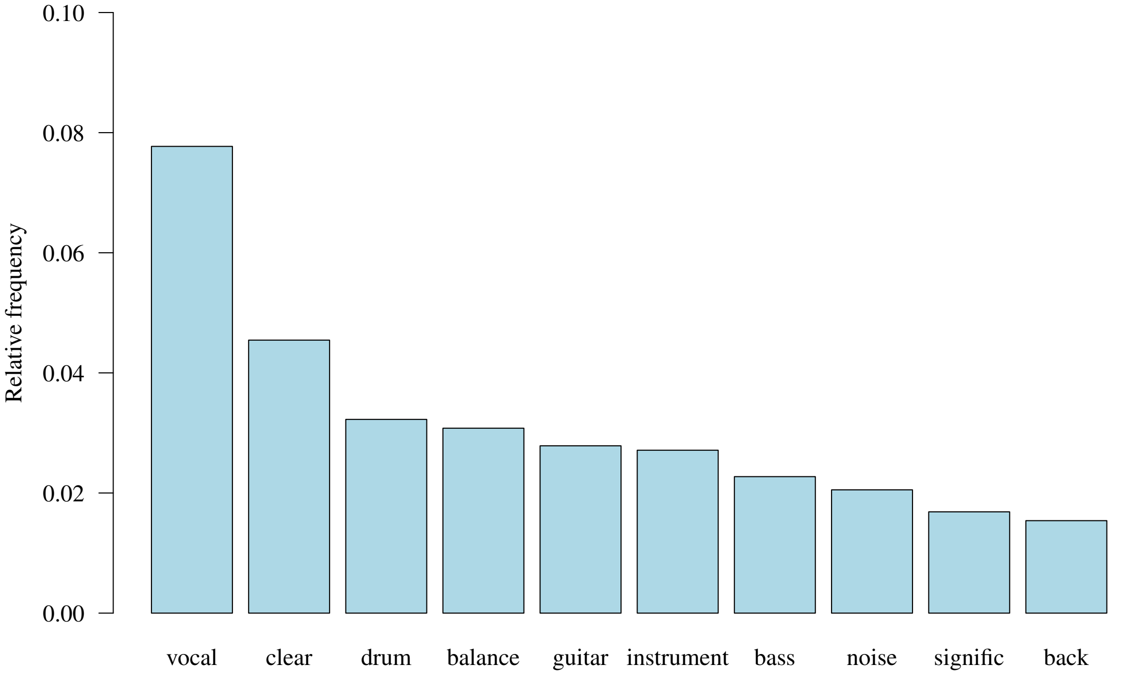
Table 6.13 shows the number of comments which fell into each category, for justifications of like and quality ratings. Frequencies highlighted in bold (with > or <) are either significantly greater than (>) or less than (<) the expected counts. From this it can be seen that the number of comments relating to “balance”, “tone” and “vocals” was far greater than other categories. This data indicates that the reasons for awarding quality ratings were more likely to be due to issues of tone, dynamics and panning when compared to like ratings. Additionally, like ratings were more often influenced by the perception of vocals and reverb than quality ratings. While not conclusive, this does appear to suggest an association of quality ratings with technical parameters and an association of like ratings with more aesthetic considerations.

6.2.5 How do features relate to subjective ratings?

With the subjective evaluation of the mixes available at a more fine grading than the simple five-level classification, it was possible to check the correlation of the subjective responses to the audio signal features. The Pearson r and coefficient of determination of a linear fit R^2, for each variable, are shown in Table 6.14. With 72 correlations (36 features and 2 subjective responses) only three are significant — Spectral Centroid to both Like and Quality, and RO85 to Like. Note that Spectral Centroid and RO85 are generally correlated in music mixes, with a Pearson r of 0.9648 over these 27 specific mixes. Of these 27 mixes evaluated here, considered the best 27 mixes in the competition, all have relatively central values of spectral centroid compared to the full set of mixes, which had a central value close to 2900 Hz. Consequently, if explicit subjective ratings were found for all 101 mixes in the competition, or all 135 mixes analysed in § 6.1.1, this relationship between Spectral Centroid and Like/Quality would likely be upheld. This suggests that the plots in Fig. 6.5



(a) Top 10 most frequently used words, when describing like ratings. The presence of both 'clear' and 'clarity' highlights a limitation in this word-stemming based approach.



(b) Top 10 most frequently used words, when describing quality ratings

Figure 6.20: Most frequent words for Like and Quality ratings. In both cases, the importance of vocals is indicated.

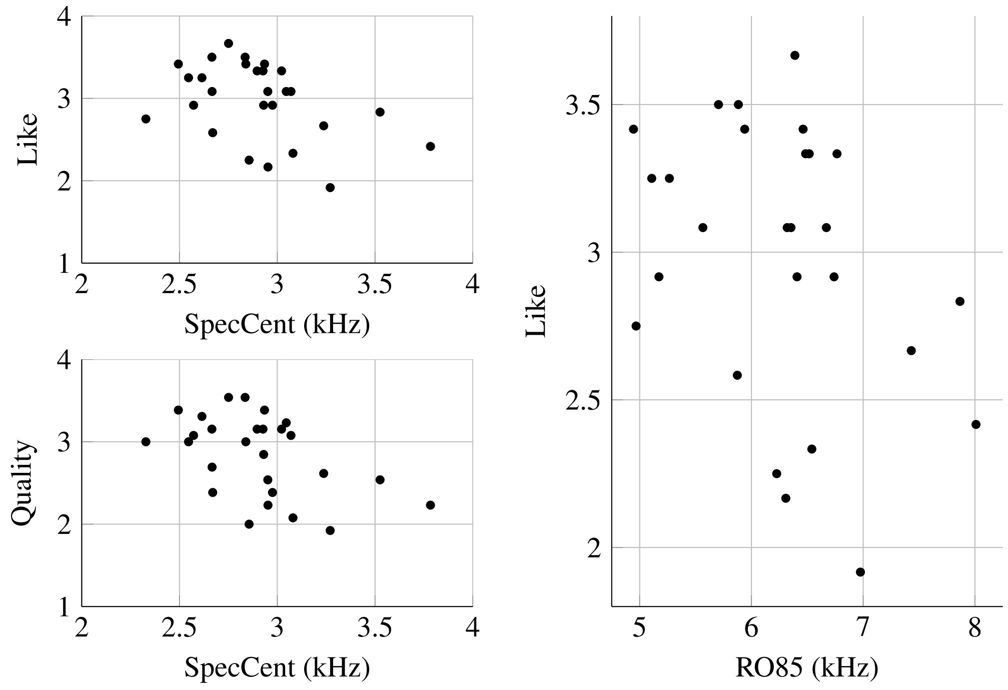


Figure 6.21: Correlation between like and quality ratings for 27 mixes of 'Blood To Bone' and features.

can be considered a good approximation of quality over mixes of each song.

As the KDE of spectral centroid values was well approximated by the sum of Guassian functions, the same technique was tested for the spectral centroid vs. like/quality relationships. This was achieved using the curve-fitting toolbox in Matlab. For like, shown in Fig. 6.22a, the use of two Gaussian functions produces a local maximum point near 3.6 kHz. However, this may be due to the influence of the few points in this region, which are possible outliers. Nonetheless, the global maximum near 2.8 kHz is based on more reliable data. For quality, a single Gaussian function performed better (as the Gauss2 fit was overfitting to the data), although Fig. 6.23a indicates that it is close to linear over the range of datapoints.

This finding can be related back to the PCA results for the 10-song analysis (see Fig. 6.3 and 6.4). In that case, as the points in the individuals factor map for mixes of this song ('Blood To Bone') overlap considerably with the other songs, we know that the mixes of this song have a varied range of feature values. This is also demonstrated by the distrubutions of features, shown in Figures 6.5, 6.6, 6.7 and 6.8, and the fact that the curves shown overlap. However, from a perceptual basis, it is clear that two mixes will sound more different if they are from two different songs, as opposed to two mixes from the same song, even if the values of features are identical in both cases. This is to say that feature values alone do not explain why mixes sound different. This also relates back to the overall competition judgement and the nature of PC2, where it was shown that brighter sounding mixes were less preferred (see Figs.6.15 and Table 6.11.)

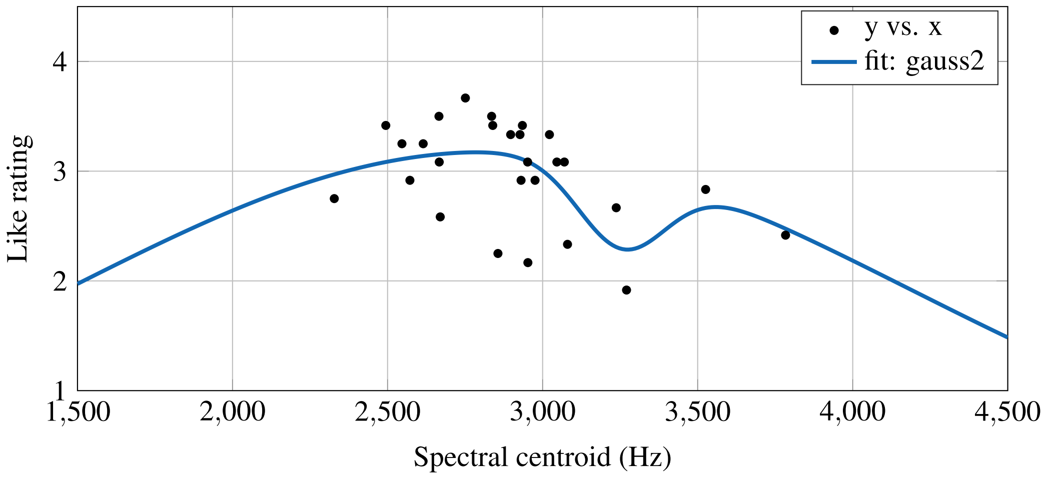
6.2.6 Discussion

The results and discussion from § 6.1.7 can be further interpreted with the addition of subjective evaluation. For example, while Fig. 6.5 shows the distribution of spectral centroid for 135 mixes of 'Blood To Bone', it is now clear that the best mixes are not necessarily at the central value.

Table 6.14: Linear fit of features to mean subjective ratings, for 27 mixes and 13 subjects' ratings. Entries in **bold** are statistically significant.

Feature	Like		Quality	
	Pearson r	R^2	Pearson r	R^2
SpecCent	**-0.409**	**0.167**	**-0.468**	**0.219**
SpecSpread	-0.048	0.002	-0.091	0.008
SpecSkew	-0.036	0.001	-0.046	0.002
SpecFlat	-0.281	0.079	-0.331	0.110
SpecKurt	-0.038	0.001	-0.052	0.003
SpecEnt	-0.294	0.086	-0.338	0.114
CF	-0.004	0.000	0.023	0.001
LoudITU	-0.055	0.003	-0.041	0.002
Top1dB	-0.211	0.045	-0.284	0.081
Harsh	-0.176	0.031	-0.232	0.054
LF	-0.230	0.053	-0.318	0.101
RO85	-0.348	0.121	**-0.416**	**0.173**
RO95	-0.287	0.083	-0.347	0.121
sbflux1	-0.221	0.049	-0.290	0.084
sbflux2	-0.114	0.013	-0.175	0.031
sbflux3	-0.086	0.007	-0.153	0.023
sbflux4	0.028	0.001	0.009	0.000
sbflux5	-0.015	0.000	0.019	0.000
sbflux6	-0.027	0.001	0.016	0.000
sbflux7	-0.122	0.015	-0.090	0.008
sbflux8	-0.223	0.050	-0.247	0.061
sbflux9	-0.283	0.080	-0.291	0.085
sbflux10	-0.272	0.074	-0.302	0.091
Gauss	-0.115	0.013	-0.108	0.012
PMFcent	0.254	0.065	0.170	0.029
PMFflat	0.011	0.000	-0.090	0.008
PMFspread	-0.070	0.005	-0.087	0.008
PMFskew	-0.073	0.005	-0.034	0.001
PMFkurt	-0.134	0.018	-0.115	0.013
W-all	0.152	0.023	0.174	0.030
W-band	0.185	0.034	0.210	0.044
W-low	0.166	0.026	0.193	0.037
W-mid	0.195	0.038	0.214	0.046
W-high	0.181	0.033	0.207	0.043
SMratio	0.088	0.008	0.164	0.027
LRimbalance	-0.179	0.032	-0.228	0.052

The spectral centroid values of the 27 most highly-rated mixes are all below the central value. This discussion is of interest since spectral centroid was the only audio signal feature extracted which was correlated to quality and like ratings in mixes of that song. The amount of variance explained is greater for like ratings than quality. In contrast to the results from Chapter 3 (as shown in Fig. 3.3), this investigation did not reveal any meaningful difference between like and quality concepts. This is suspected to be due to the absence of any inter-song variation and, therefore, any differences in song-familiarity, which was seen to be a predictor of like ratings in that study.



(a) $R^2 = 0.3209$.

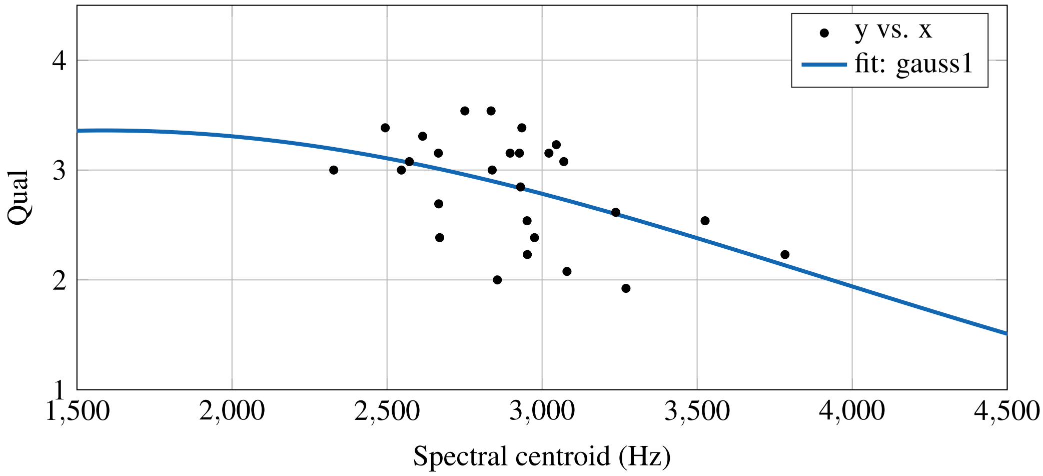
General model	$f(x) = a_1 e^{-\frac{x-b_1}{c_1}^2} + a_2 e^{-\frac{x-b_2}{c_2}^2}$	
Coefficients:		(with 95% confidence bounds)
	a_1	-0.7297 (-1.553, 0.09376)
	b_1	3254 (2975, 3532)
	c_1	200.3 (-128.3, 528.9)
	a_2	3.176 (2.82, 3.533)
	b_2	2825 (2122, 3528)
	c_2	1920 (134.4 3706)
Goodness of fit:		
	SSE	3.769
	R^2	0.3209
	Adjusted R^2	0.1592
	RMSE	0.4237

(b) Fit result: like ratings vs. spectral centroid

Figure 6.22: Relationship between like ratings and spectral centroid for 27 mixes of ‘Blood To Bone’.

Datasets of alternate music mixes are scarce in the literature. Evaluation of these datasets is perhaps even more so. On-line mix competitions provide an opportunity to examine an evaluated dataset of mixes. Since this analysis was undertaken, a second mix competition has also taken place, using the same format of winner, runner-up, shortlisted, honourable mentions and others [5]. In this case, the total number of mixes was 57. This presents an opportunity for a second case study, although too late for inclusion in this thesis. Additionally, the audio mixes of the CMT community have been indexed in the Open Multitrack Testbed [156], which will hopefully lead to subjective evaluations of these mixes becoming available in the future.

[5] `http://www.cambridge-mt.com/Diesel13Competition.htm`



(a) $R^2 = 0.2221$.

General model	$f(x) = a_1 e^{-\left(\frac{x-b_1}{c_1}\right)^2}$	
Coefficients:		(with 95% confidence bounds)
	a_1	3.361 (0.4189, 6.302)
	b_1	1588 (-5234, 8409)
	c_1	3256 (-4728, 1.124e+04)
Goodness of fit:		
	SSE	4.716
	R^2	0.2221
	Adjusted R^2	0.1573
	RMSE	0.4433

(b) Fit result: quality ratings vs. spectral centroid

Figure 6.23: Relationship between quality ratings and spectral centroid for 27 mixes of 'Blood To Bone'.

6.3 Chapter summary

A dataset was prepared containing 1501 audio files representing the mixes of 10 songs. The number of mixes of each song ranged from 97 to 373. A variety of objective signal features were extracted and principal component analysis was performed, revealing four dimensions of mix-variation for this collection of songs, which can be described as 'amplitude', 'brightness', 'bass' and 'width'. Feature distribution suggests multi-modal behaviour dominated by one specific mode. This distribution appears to be robust to the choice of song, with variation in modal parameters. This has provided insight into the creative decision making processes of mix engineers.

Subjective quality ratings were obtained for subsets of this dataset in order to examine the relationship between audio signal features and the perception of audio quality and mix-preference. This was done for 101 mixes of one song, with evaluation in the form of the mix's ranking in an on-line competition, and the highest ranking 27 mixes were evaluated under laboratory conditions. In contrast to the results from Chapter 3, like and quality ratings were strongly correlated.

For future work, as the study presented here only considered features relating to amplitude, spectrum and stereo panning, an in-depth study using rhythmic and metrical features is required. It is anticipated that this dataset of mixes can be used to test the robustness of algorithms used in MIR, for tasks such as tempo estimation, genre prediction and music structure analysis.

Real mix engineers do not apply random EQ or random track gains. The distribution of real mixes is also wider. This suggests that, in real mixes, the engineers choose from a wider variety of values than the random methods which were employed. When combined with the results from the mix-space experiments, this suggests that real mix engineers have intentions which they can realise. As trivial as this may sound this is an important point, since it is these intentions that an engineer will want to realise in any automated/intelligent mixing system, and these intentions relate to their own impression of quality.

Consequently, furthering the understanding of mix-variation will be necessary for the design of future intelligent/automated music production systems. However, this incipient study shows that relatively basic measures of central tendency and distribution are useful targets for such systems. Under higher level human supervision, this concept could be used to achieve sonic qualities which approximate current accepted practices, or as a creative contrast, to challenge current trends and exploit results which may lie at the boundaries of the feature spaces studied. This is explored in Chapters 8 and 9.

7
Analysis of mix engineers

Chapter 6 dealt with the variation in a set of mixes, analysing how hundreds of examples of a given song could vary, in terms of audio signal features, and how these variations were related to quality in specific case studies. The following chapter expanded on these findings and investigated the effect of individual mix engineers on the variation in audio signals. This chapter is divided into two main sections, § 7.2 and § 7.3, covering two experiments on the same dataset: one investigated the objective variation in signal features across six mix engineers and one study which sought to measure the subjective preference listeners had for the mixes of each engineer.

7.1 Introduction

In addition to the variation across mixes, it is important to understand that the mixes created by an individual mix engineer may vary compared to the mixes created by another. How can differences between mix engineers be explained? They may be using different DAWs, different reproduction equipment, different rooms etc. Some of these factors may 'leave an impression' on the mix, which can be measured using certain audio signal features. This impression can be referred to as a sonic signature.

Definition 9. *Sonic signatures are the audible traces of particular types of social activity involved in the production of recorded music, where social activity is interaction between people or between a person and a form of technology [204].*

In audio engineering, this term has been applied to a number of systems, such as dynamic range compressors [205]. By extension of Definition 9, and for the purposes of the investigation in this Chapter, "sonic signature" is specifically defined as follows.

Definition 10. *The audible traces of a mix engineer's creative and technical decisions on their produced mix, as observed over a series of their productions.*

7.1.1 Research questions

After considering the work of mix-engineers and the definition of a sonic signature, the following research questions were formed and are addressed in this chapter.

RQ-15 Is there a measurable difference in signal features between the mixes of mix engineers?

RQ-16 Can the mix engineer be predicted from the audio signal?

RQ-17 Is there a measurable subjective difference between the mixes of mix engineers?

RQ-18 Are the samples from one engineer typically preferred to those of another?

These four research questions pertain to this dataset of mixes. Questions 1 and 2 are addressed in § 7.2, while questions 3 and 4 are addressed in § 7.3.

7.1.2 Dataset #4a — 190 mixes

In order to investigate the measurable objective variation from one mix engineer to the next, it was necessary to compile a dataset of mixes by various mix engineers. As in § 6.1.1, the mixes used here were gathered from the CMT database [1]. In addition to being a collection of multitrack sessions, this website also functions as a forum where registered members can discuss a variety of topics. By retrieving the list of all members and arranging by amount of posts and threads started it was possible to determine which individuals have contributed the most mixes in total. This is due to the fact that when a member has created a mix and wishes to share it with the community, he/she most often starts a new discussion thread. Subsequently, a list of the contributed mixes from the most prolific members was compiled. By cross-referencing the entries for each mix engineer there were found to be 18 songs which six engineers had each mixed (as of October 2015 when this search was undertaken). In some cases, the mix engineer had contributed more than one mix

[1] `http://www.cambridge-mt.com/ms-mtk.htm`

of a given song and, as such, the total number of audio samples is greater than 6 × 18, with the final number of samples in the dataset equal to 190. The number of audio samples belonging to each mix engineer ranges from 21 to 44. The specific number of mixes produced for each song by each mix engineer is shown in Table 7.2.

Table 7.1: List of songs used in Sonic Signatures dataset

Index	Artist	Title
S1	Moosmusic	Big Dummy Shake
S2	Young Griffo	Blood To Bone
S3	Bill Chudziak	Children Of No One
S4	The Abletones Big Band	Corine
S5	Banned From The Zoo	Encore
S6	James Elder & Mark M Thompson	English Actor
S7	Ben Carrigan	Hey Carrie Anne
S8	Angels In Amplifiers	I'm Alright
S9	Bruks	Kak Tvoi Dela, Vova?
S10	Selwyn Jazz	Much Too Much
S11	The Wrong 'uns	Rothko
S12	Arise	Run
S13	Jokers, Jacks & Kings	Sea Of Leaves
S14	Sven Bornemark	Stop Messing With Me
S15	Rod Alexander	Tears In The Rain
S16	Signe Jakobsen	What Have You Done To Me
S17	The Brew	What I Want
S18	Street Noise	You Are The One

7.2 Variation in audio signal features across mix engineers

In order to objectively characterise the audio signals a number of signal features were extracted. The choice of features was identical to those used in Section § 6.1.1 (see Table 6.2). This analysis is conducted in order to answer the first and second research questions in Section § 7.1.1.

7.2.1 Preliminary investigations

As an initial investigation into the data, the distribution of four particular features was plotted and is shown in Fig. 7.1. These particular features were chosen as they are representative of the first four dimensions of the PCA in Chapter 6, as in Table 6.6. The distribution of individual signal features reveals some significant differences between mix engineers. For example, half of the mix engineers exhibit high loudness levels, compared to the other half, presumably due to the use of dynamic range compression applied to the overall mix. This is interesting as it recalls the result shown in Table 6.6, that among 1,501 mixes, the distribution of *loudness* values followed two Gaussian functions, with means of approximately -13 and -8.5 LU. This behaviour is replicated in Fig. 7.1, with similar values.

However, these differences only indicate limited, low-level, effects. The more high-level, perceptual differences between the mix engineers is not clear from these summary statistics. With 190 samples, over 6 classes, there was not a sufficient number of samples for machine learning.

Table 7.2: Sonic signatures dataset: table of mixers and songs

		Mixers						TOTAL
		M1	M2	M3	M4	M5	M6	
Songs	S1	3	1	2	1	1	2	10
	S2	5	1	2	2	2	4	16
	S3	1	1	1	1	2	2	8
	S4	1	1	4	1	3	1	11
	S5	1	1	2	1	1	1	7
	S6	2	2	1	1	1	2	9
	S7	1	1	5	1	1	1	10
	S8	2	2	1	1	3	1	10
	S9	1	2	1	1	1	2	8
	S10	2	2	3	1	1	1	10
	S11	1	1	1	1	2	2	8
	S12	1	1	4	1	1	3	11
	S13	3	2	2	1	1	1	10
	S14	2	1	3	3	1	2	12
	S15	6	1	1	1	1	1	11
	S16	3	2	4	1	1	1	12
	S17	3	2	4	1	3	1	14
	S18	6	2	1	1	1	2	13
TOTAL		44	26	42	21	27	30	190

A number of statistical classifications were attempted. Figure 7.2 shows these 190 samples positioned in the PCA space from Fig. 6.2a and 6.2b, which was derived from the larger study of 1,501 samples. While it was shown that there was some clustering due to song, there does not appear to be any noticeable effect due to the mix engineer that is visible in this space.

7.2.2 Optimised linear projection

In multivariate data analysis, one often encounters the so-called "curse of dimensionality", which describes how higher-dimension spaces become increasingly sparse [206]. One way to overcome this is to reduce the number of dimensions, omitting those which do not offer the power to discriminate between the different class. Consider the artificial data shown in Fig. 7.3. Each data point has an X, Y and Z coordinate. The images shown are both clearly only two-dimensional, as they are projected onto the page, yet both images show different 'views' of the data. As the two classes only differ along the X-axis, then the X-Y or X-Z axis view reveals the difference.

This is the principal behind projection pursuit, wherein an *interesting* linear projection of the dataset is sought. For the current dataset, there are 190 audio examples, across a class variable with six discrete values, measured over 36 audio signal features. Finding a linear projection of these 36 dimensions which shows the difference between the six mix engineers is non-trivial, assuming a difference were to exist at all. The remainder of this section describes a method of optimised projection pursuit.

While all 36 of these features could be used, in order to generalise to any number of features (which may be quite large) and not use too many variables and risk the curse of dimensionality, the following algorithm aims to select a subset of the total feature set which creates an *interesting*

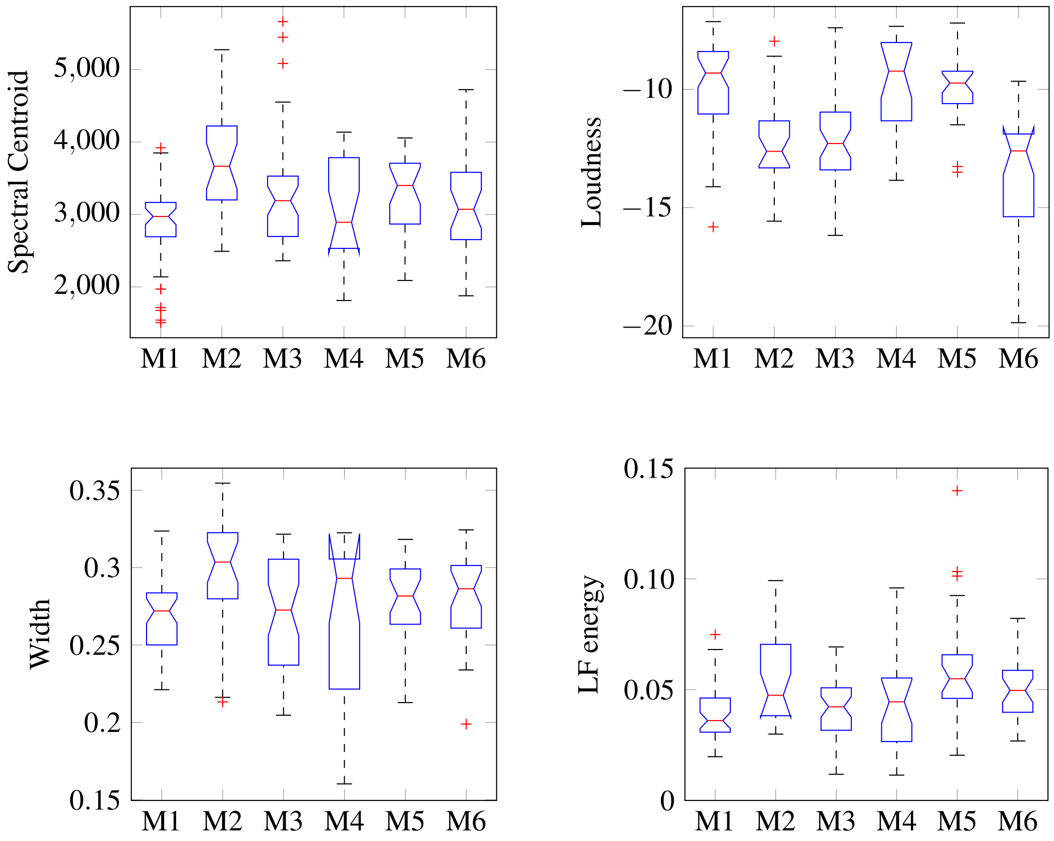


Figure 7.1: Boxplots of features, grouped by mix engineer

projection (one which reveals the difference between the different mix engineers). To obtain such a linear projection, a similar method to that of VizRank [207] was implemented. First, the ReliefF measure [208] is obtained for all n features. The result is displayed in Table 7.3. Once ranked according to ReliefF, a subset containing m of the n features was obtained by random sampling using a gamma probability distribution. This distribution was created by generating a large number (10^6) of gamma distributed random numbers, X_γ, using the shape parameter $k = 1$ and scale parameter $\theta = 2$. These parameter values were selected so that features with a high reliefF would be chosen much more often than those which score lower (see thick line in Fig. 7.4a). X_γ is normalised to the range [0,1]. A histogram is then obtained using n bins, which provides the probability of each of the n features being selected according to this particular gamma distribution. The result is shown in Fig. 7.4b. The first m probabilities are used as weights in the selection of m features.

This provides a subset of features which then must be scored according to its ability to distinguish between the various classes (the individual mix engineers in this case). The scoring metric is based on a k-Nearest Neighbours classifier (kNN). As the class with the least amount of observations has 21, the value chosen was $k = 20$ (see Table 7.1). For each point, the k nearest neighbours are discovered, based on the Mahalanobis distance metric in the m-dimensional feature space. This metric was used as it is unitless, scale-invariant and considers the correlations of the features [209]. The proportion of the k nearest neighbours which are members of the same class was obtained. This value was obtained for all points and the average proportion of same-class membership was recorded as *'kNN score'*, or S_{knn}. Subsets of m out of n features are randomly selected, based

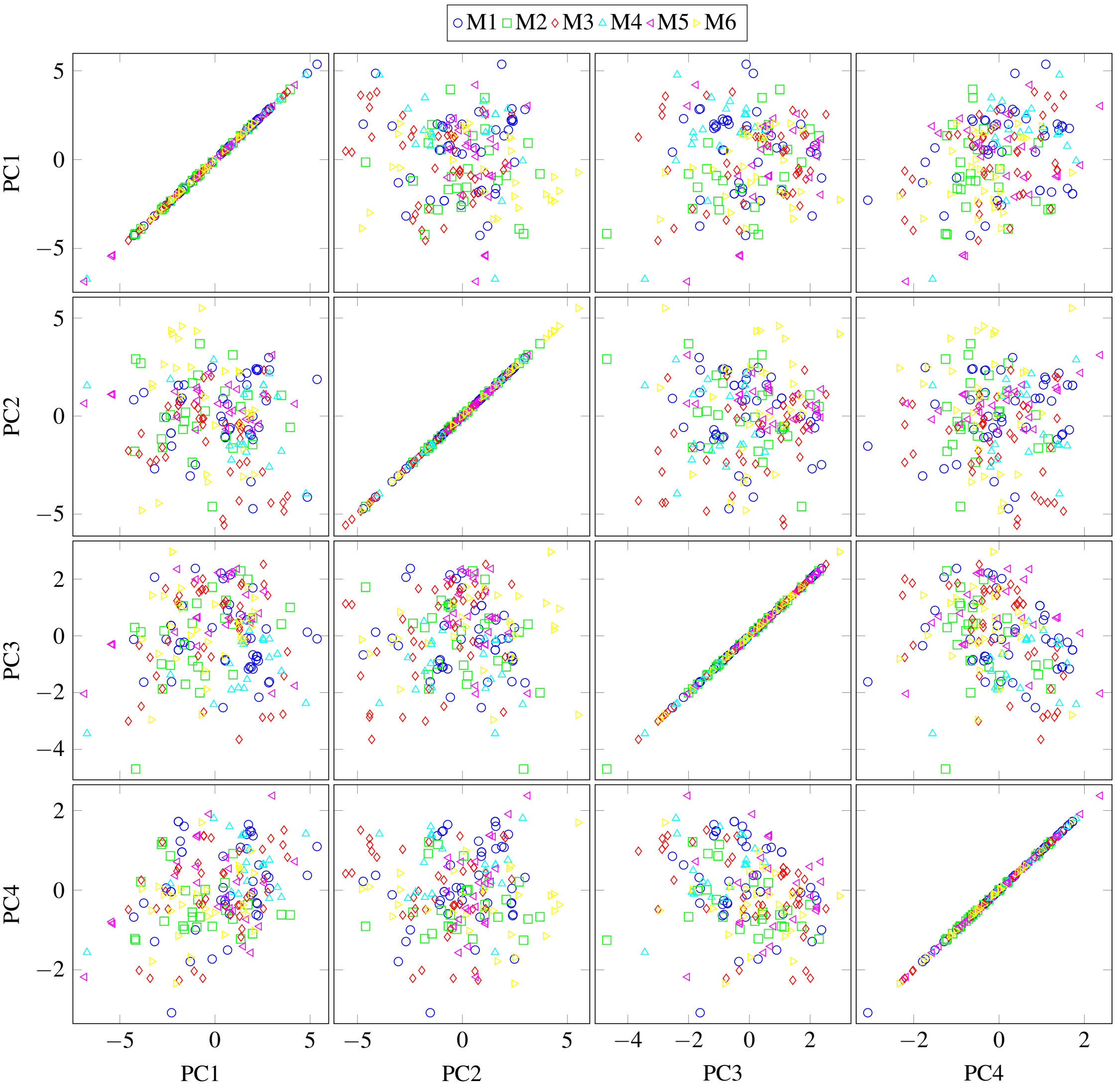


Figure 7.2: 190 mixes, displayed in PCA space from § 6.1.4

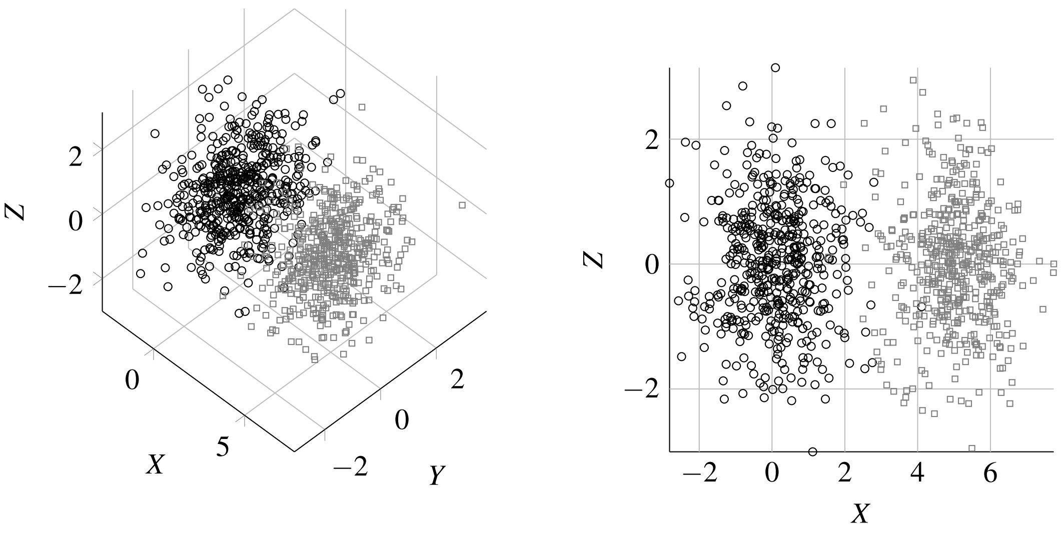


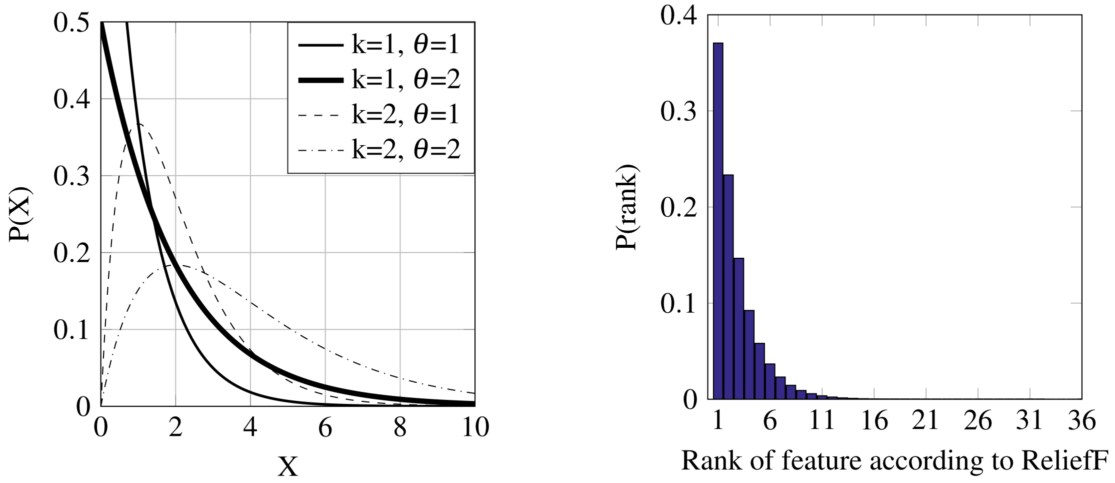
Figure 7.3: Illustration of the principle of linear projection, using artificial data.

Table 7.3: Results of Kruskal-Wallis test and ReliefF scores for 36 audio features

Name	Kruskal-Wallis test			ReliefF
	p-value	χ^2	η^2	
SpecCent	0.0008	21.1332	0.1124	0.0107
SpecSpread	0.0000	62.0574	0.3301	0.0297
SpecSkew	0.0000	53.2415	0.2832	0.0506
SpecFlat	0.0000	56.9879	0.3031	0.0427
SpecKurt	0.0000	56.2966	0.2995	0.0584
SpecEnt	0.0011	20.3214	0.1081	0.0082
CF	0.0000	63.5265	0.3379	0.0524
LoudITU	0.0000	64.8359	0.3449	0.0423
Top1dB	0.0000	62.6229	0.3331	0.0161
Harsh	0.9316	1.3317	0.0071	-0.0026
LF energy	0.0001	27.0581	0.1439	0.0064
RO85	0.0027	18.1930	0.0968	0.0106
RO95	0.0012	20.1270	0.1071	0.0110
Gauss	0.0000	31.0596	0.1652	0.0240
PMFcent	0.0000	93.1469	0.4955	0.0579
PMFflat	0.0000	58.7915	0.3127	0.0911
PMFspread	0.0000	66.3267	0.3528	0.0535
PMFskew	0.0001	26.7394	0.1422	0.0152
PMFkurt	0.0000	50.2646	0.2674	0.0172
W-all	0.0150	14.0907	0.0750	0.0091
W-band	0.2626	6.4764	0.0344	0.0072
W-low	0.0000	34.7328	0.1847	0.0229
W-mid	0.0002	24.1631	0.1285	0.0110
W-high	0.6788	3.1377	0.0167	0.0072
SMratio	0.0237	12.9632	0.0690	0.0103
LRimbalance	0.0623	10.4989	0.0558	0.0056
sbf1	0.1811	7.5771	0.0403	-0.0008
sbf2	0.0351	11.9740	0.0637	0.0027
sbf3	0.0004	22.8772	0.1217	0.0126
sbf4	0.0040	17.2536	0.0918	0.0046
sbf5	0.0160	13.9407	0.0742	0.0044
sbf6	0.0460	11.2853	0.0600	0.0027
sbf7	0.0748	10.0149	0.0533	0.0007
sbf8	0.0694	10.2121	0.0543	0.0039
sbf9	0.0032	17.8301	0.0948	0.0068
sbf10	0.0120	14.6378	0.0779	0.0023

on the probabilities in Fig. 7.4b, and subsequently scored up to a maximum number of iterations (which was set to be 500). The subset with the highest S_{knn} is the subset to be optimised.

To investigate the objective variation between different mix engineers, and determine how best to classify them, the method of projection pursuit is used. This method transforms an m-dimensional system to a 2D map. With a matrix $p \times m$, containing p observations of m variables, we seek the matrix containing the X-anchors and Y-anchors, such that the resulting x and y coordinates separate the different classes (mix-engineers) as best as possible.



(a) PDF of gamma distributions

(b) PMF of gamma distribution for 36 features

Figure 7.4: A gamma distribution was used to select a subset of the feature set. This ensures more highly ranked features were more likely to be chosen.

$$\begin{bmatrix} a_{1,1} & \cdots & a_{1,m} \\ \vdots & \ddots & \vdots \\ a_{p,1} & \ldots & a_{p,m} \end{bmatrix} \times \begin{bmatrix} X_1 & Y_1 \\ \vdots & \vdots \\ X_m & Y_m \end{bmatrix} = \begin{bmatrix} x_1 & y_1 \\ \vdots & \vdots \\ x_p & y_p \end{bmatrix} \tag{7.1}$$

Figure 7.6a shows a set of initial anchors. For simplicity, the anchors are equally spaced around a unit circle. Figure 7.6b displays the datapoints in this linear projection. Each mix engineer, referred to as $M1, M2, \ldots, M6$, is indicated by a colour/symbol combination. It is clear that the individual classes are not easily separable in this plot. We seek such a projection wherein the classes are most clearly distinguished from one another.

To find the matrix of anchors by optimisation a genetic algorithm was used. This has previously been referred to as evolutionary pursuit [210]. The goal was to determine the choice of anchors such that a high S_{knn} is achieved on the transformation result (the 2D representation) which those anchors yield. Consequently, the number of variables to optimise is $n_{\text{vars}} = 2 \times n_{\text{axes}}$, and the fitness function to be minimised is $1 - S_{knn}$.

For this chapter, the genetic algorithm was implemented using the global optimisation toolbox in Matlab. The initial population of solutions was uniformly chosen within the range [-1, 1], for all dimensions. The algorithm used rank fitness scaling and roulette selection. The mutation function used was “Adaptive Feasible”, which is the default mutation function when constraints are implemented [2]. A more complete discussion of genetic operators is presented in Chapter 8, wherein that work required the algorithm to be written from scratch.

This process was completed a total of ten times. The mean and best fitness value at each generation and each run is shown in Fig. 7.5. This indicates that, perhaps due to the complexity of the problem and its high dimensionality, that the fitness of the optimal solution found after each run varies — there are a number of possible optimal solutions and the algorithm does not converge

[2] https://uk.mathworks.com/help/gads/genetic-algorithm-options.html#f6633

Table 7.4: Settings used in the following example of IGA mixer

Parameter	Description	Value
N_{features}	Number of audio signal features	10
N_{vars}	Number of variables/dimensions in solution space	$2 \times N_{\text{features}}$
Population size	Number of candidate solutions per generation	100
Elite fraction	Proportion of children generated as clones of fittest parents	0.025
Crossover fraction	Proportion of children generated by crossover of two parents	0.90
Stop condition	Condition which, when met, causes evolution to cease	100 generations

towards a best solution. In all ten runs, it takes, at most, 75 generations for population diversity to become fatally low, from which point the population no longer evolves. Fig. 7.7a displays one such optimised set of anchors where the fitness function has a value of 0.4288 (and, as such, $S_{knn} = 0.5712$), meaning that, on average 57.12% of the 20 nearest neighbours to a given point are members of the same class.

In Fig. 7.7a, the anchors with the greatest length, and therefore most influence on this 2D projection, are PMFcent, Spectral flatness, RO85 and Spectral Skewness. These also roughly align with the X and Y axes. Therefore one can interpret mixes with high X-coordinates as having greater values of Spectral Flatness, often considered a measure of the amount of correlation structure existing in the audio signal, i.e. whether it is more tone-like or noise-like [211]. Since all mixers mixed the same songs, we cannot simply say that the more noisy songs are at one end of the graph. This "noise" or broadening of the spectrum, must have been caused by the mixing process. It is hypothesised that this is a result of increased distortion, caused, for example, by dynamic range compression.

It is important to note that this is not a factor map in the sense of what is produced by exploratory factor analysis or PCA. This is simply a map of anchors which produces an interesting linear projection. The result, from genetic optimisation, is of course based on the randomness inherent in that method of optimisation and, as shown in Fig. 7.5, multiple optimal solutions are possible. The point being, it is not a contradiction for Crest Factor (greater values representing *greater* dynamic range) and Spectral Flatness (greater values representing *reduced* dynamic range, in the context of mixes) to be pointing in the same direction. Neither does it mean that opposing vectors must represent opposing percepts, such as brightness and lack of brightness. Care must be taken when interpreting the resultant scatter plots.

In order to ascertain the degree of separation between the six classes, the centroid of each class was determined along with the 95% confidence ellipses. This calculation is made using the `FactoMineR` package [135] in **R**, as in earlier chapters. Note that this ellipse is a confidence estimate in the centroid itself and *not* an ellipse containing 95% of the data points in that class. The group centroids and confidence ellipses are plotted anew in Figure 7.8. This shows that the degree of separation between classes that might be expected with a much larger dataset (provided

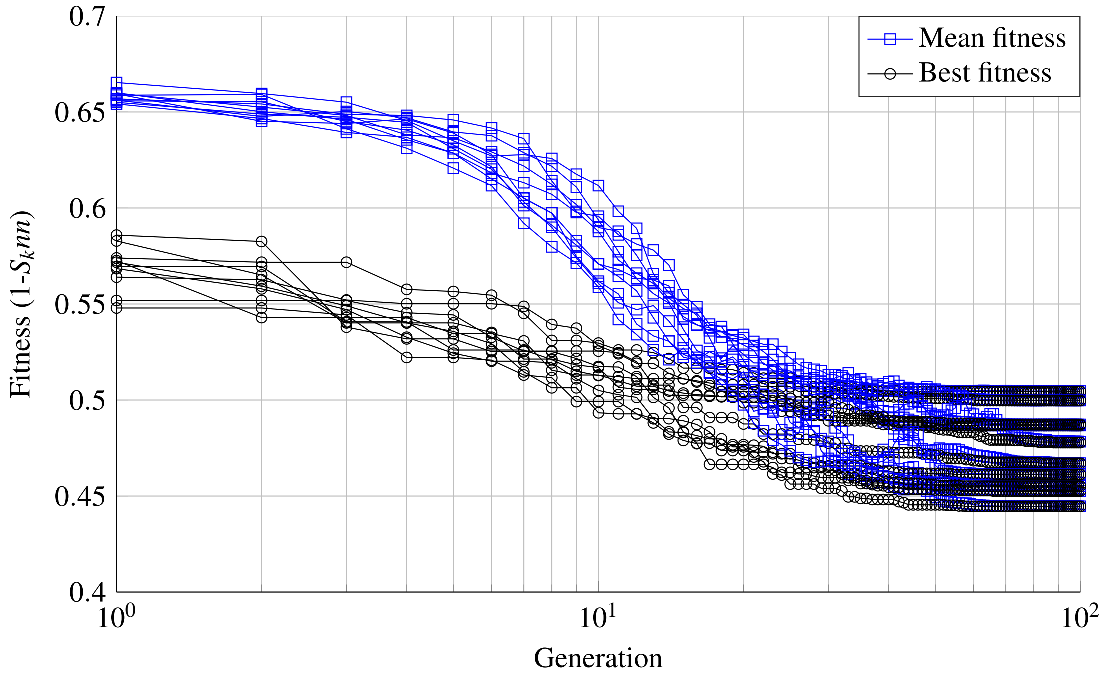


Figure 7.5: GA performance over 10 runs. For each run, the mean fitness eventually meets the best fitness, showing convergence on a solution. However, a consistent solution is not being found.

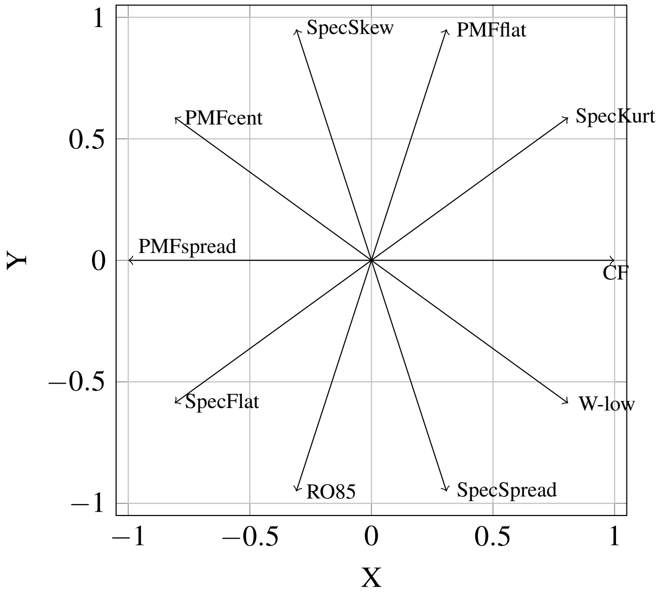
the same projection was used).

7.2.3 Discussion

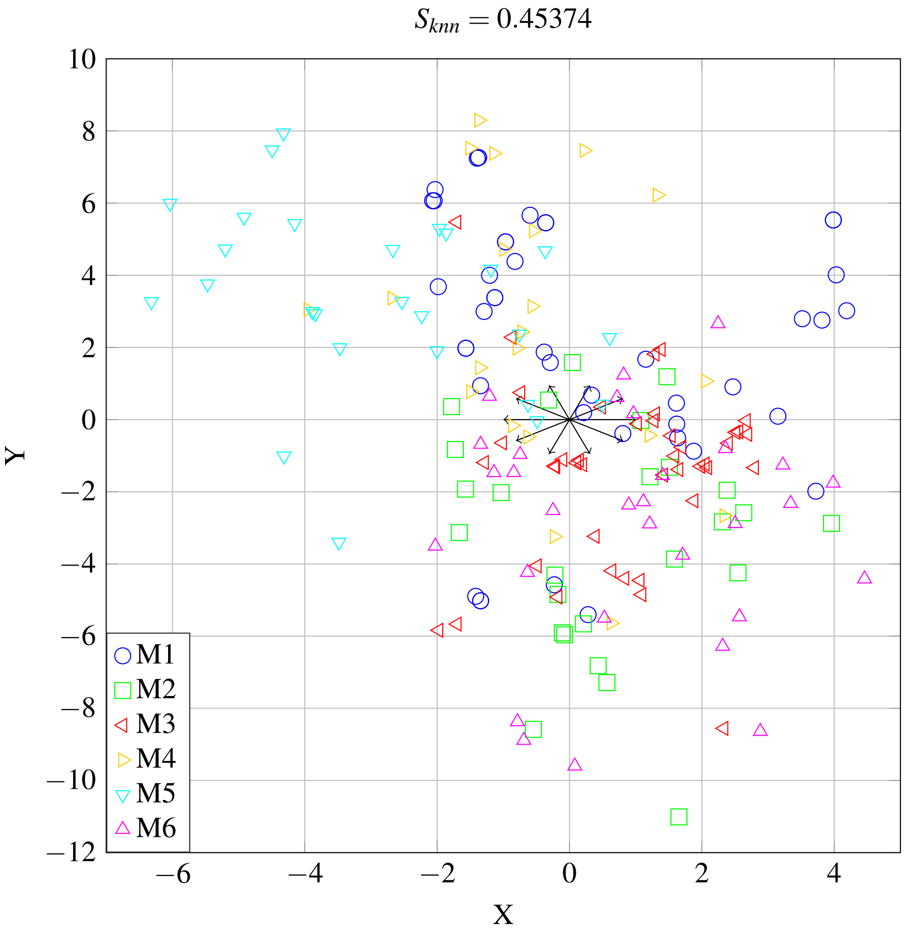
When ranked according to the value of the ReliefF measure (see Table 7.3), the most salient features in this classification strategy appear to be those associated with the sample amplitude PMF. More generally, it is seen that the most highly ranked features are associated with the loudness of the audio sample, both in terms of perceived loudness and the dynamic range of the signal.

As it is possible to measurably distinguish the output of some mix engineers from others, as shown in Fig. 7.7b, it can be suggested that this analysis has provided evidence in support of the sonic signature concept being applied to mix engineers. When an engineer creates a mix they invariably leave traces in the signal which can be used to identify them later. Figure 7.7b suggests that this may not be the case for all mix engineers, or that the style of some mix engineers is more identifiable from the signal features than others. In particular, the result in Fig. 7.8, which shows that the confidence ellipses of M2 & M6 overlap, as do M1 & M4 and M3 & M4, suggests that of these six different mix engineers, five belong to one of two groups. From inspecting the anchors it can be observed that the group of M1, M3 and M4 typically produces mixes that are subjectively brighter, as the have higher values along the RO85 variable. This group also produces mixes that are less loud and more dynamic, compared to the other three. M5 may be considered an outlier, as the values of the PMF centroid is notable different for this mixer only. This suggests there may have been some asymmetrical clipping of the output signals, or some slight DC offset.

This analysis has been completely based on audio signal features. While some of these may be perceptually-based features there is no explicit measure of subjective response. Section § 7.3 will focus on this topic — the subjective perception of quality in music mixes.

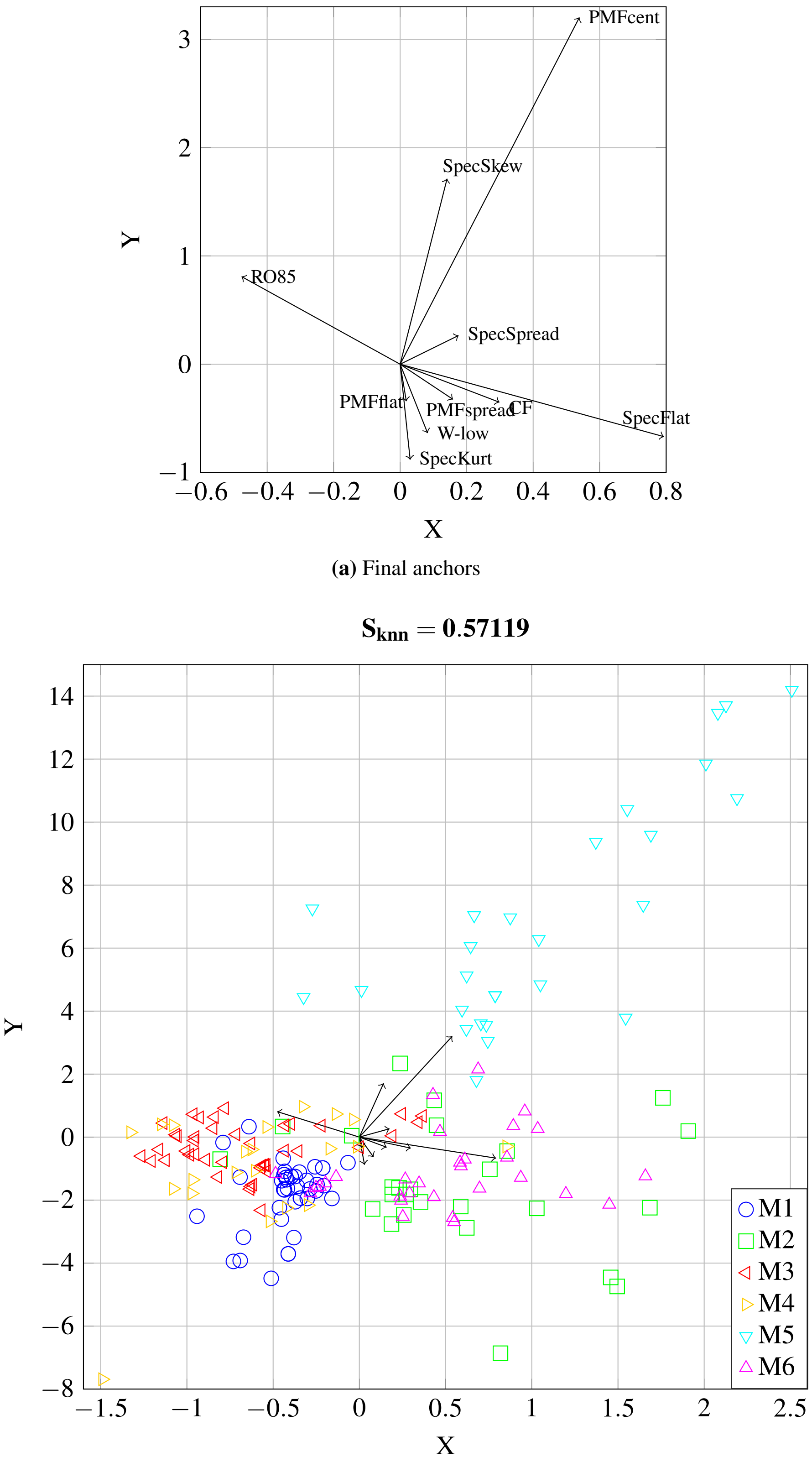


(a) Initial anchors, as placed on the unit circle



(b) Initial scatter plot, with initial anchors

Figure 7.6: Initial configuration of anchors, after feature selection but before optimisation



(a) Final anchors

(b) final scatter plot

Figure 7.7: Final configuration, after optimisation. Note that the value $1 - 0.57119 = 0.4288$ is better than any other of the 10 runs shown in Fig. 7.5, since those were only performed after this result was obtained.

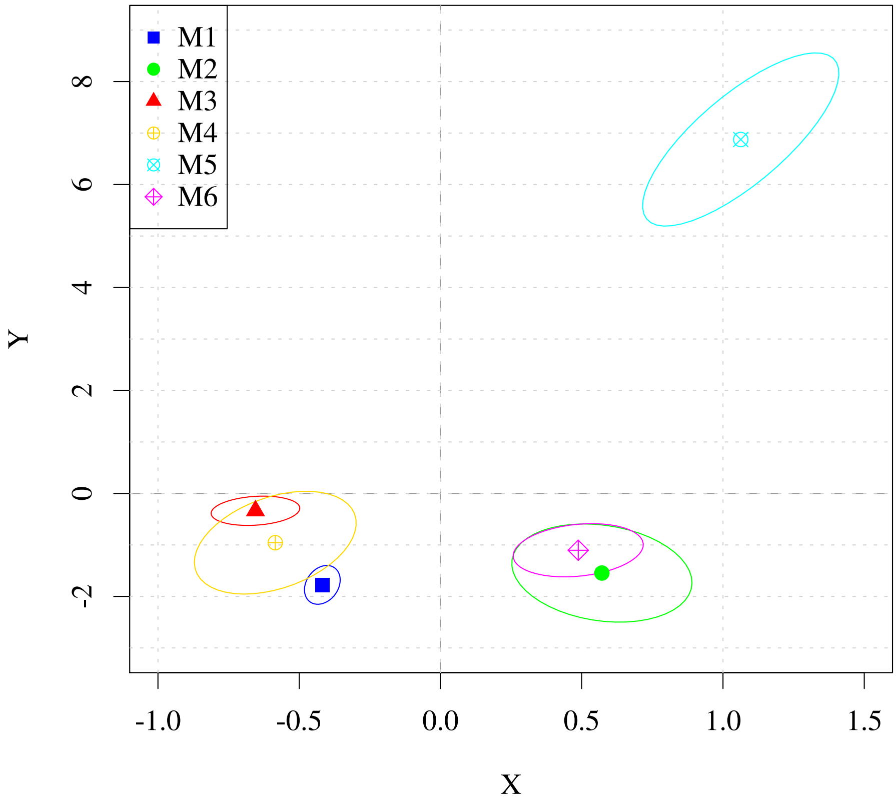


Figure 7.8: Clustering of mix-engineers in optimised linear projection space. Group centroids are highlighted by point markers and the coordinates of the ellipse around the centroid/barycentre of individuals are calculated (95% confidence) and displayed.

7.3 Sonic Signatures

In § 7.2 a dataset containing multiple mixes of multiple songs was described. Importantly, the same six engineers mixed every song. This allowed an investigation into the audio signal features of the mix engineers. What was absent at that point was any explicit rating of quality/preference, of how good the mixes are in comparison to one another. As such, only two of the research questions posed were answered. The following section of the thesis aims to answer the final two questions, namely,

RQ.17 Is there a measurable subjective difference between the mixes of mix engineers?

RQ.18 Are the samples from one engineer typically preferred to those of another?

An important aspect of the "sonic signature" of a mix engineer are these subjective and perceptual attributes of their mixes. While the feature-based analysis reported that different engineers produced mixes with significantly different audio signal features, in order to address these final two questions, explicit subjective evaluations of the audio stimuli were required.

7.3.1 Dataset #4b — 108 mixes

Recall that most engineers produced multiple mixes of each song. The final mix, chronologically, from each mix engineer was chosen for evaluation. This assumes that the final mix created by an individual is the one which they would be most happy with and most in-line with their vision for the song. This creates dataset #4b, a subset of dataset #4a, containing 108 (6×18) audio samples.

7.3.2 Test design

The listening test was designed as a multi-stimulus task, with all sliders co-located, as shown in Fig. 7.9. This test was deployed using the Web Audio Evaluation Tool [212], which allowed the test to be conducted in a web browser using the Web Audio API. All of the audio samples used in the test were normalised in perceived loudness, according to BS.1770-3 [32]. In order to reduce the duration of the test to a manageable length, each participant evaluated mixes of only four songs, chosen at random from the entire set of 18 songs. The order of playback was randomised. The initial positions of all sliders were randomly chosen.

7.3.3 Results

The test was launched in June 2016. Test results were compiled after a period of 6 months. Incomplete trials, where participants did not complete all tasks were excluded. Also excluded were trials in which participants only made minimal moves to the sliders, in order to simply advance the test. After unusable data was excluded there remained data from 56 individual participants. Each of these trials was saved as a separate .XML file. These files were combined and the data parsed using the `scores_parser.py` script from Web Audio Evaluation Tool [212]. This resulted in 18 .CSV files being created, one for each of the 18 songs used. Each of these files contained an $n \times 6$ matrix of scores: n is the number of participants who encountered and evaluated that song, and the columns are the six different mix engineers. The data from these .CSV files was imported into Matlab where it was reshaped to form a 224×6 matrix of scores (56 participants $\times$ 4 songs each $= 224$). Alongside this were created a 224×1 vector for each of the following labels: song titles, participant names and engineer names.

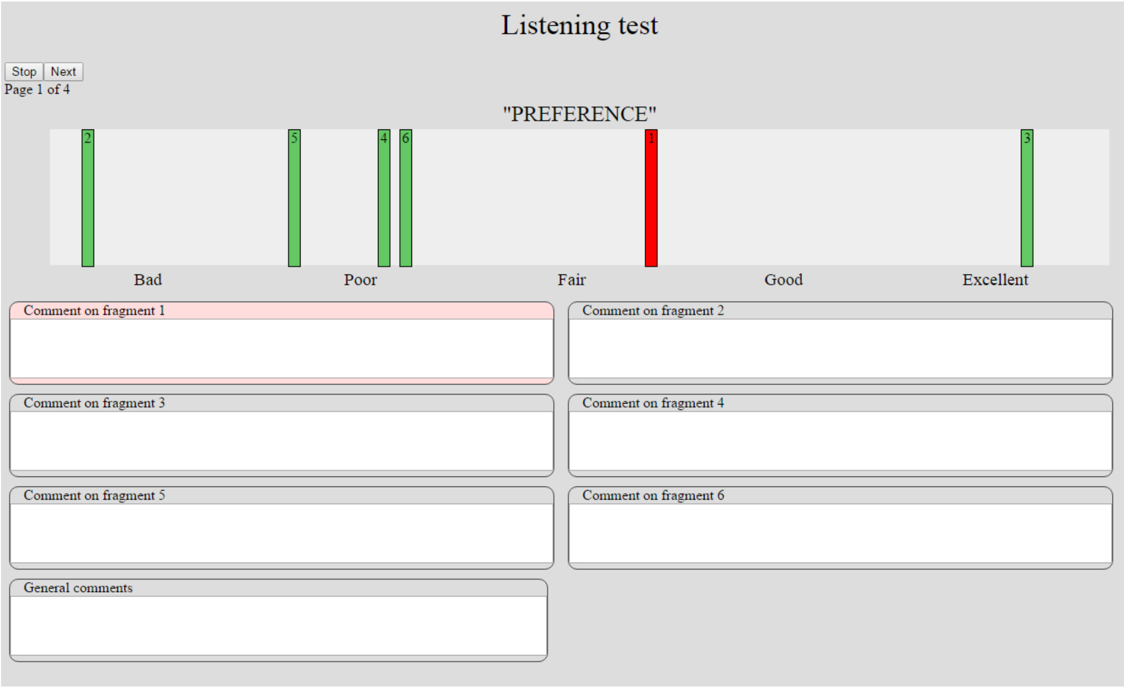


Figure 7.9: Sonic Signatures on-line test shown in Google Chrome. The mix being played is highlighted in red and this slider should then be dragged to the appropriate position on the scale. In each test the participant completes four such screens, representing a random four out of the total 18 songs.

Table 7.5: Table of Kruskal-Wallis test results, over all songs

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Groups	1.2999e+06	5	2.5999e+05	8.6294	0.1248
Error	2.0101e+08	1338	1.5023e+05		
Total	2.0231e+08	1343			

In the test design, each screen consisted of six mixes of a given song. Those mixes were rated on a scale from 0 to 1. Since there was no reference sample or anchor sample, nor the requirement that samples be placed at extreme ends of the scale, it was possible for various methods of rating to be employed. For example, a participant may, for one song, rate all mixes on the lower end of the scale, while, for the next song, rate all mixes at the higher end. Consequently, the scores were normalised. Considering the scores from one particular screen as a six-dimensional vector, these vectors were normalised according to their L_2 norm. This ensures that the contribution of each vector, to the total matrix, is equal.

For the combined data for each song, a Kruskal-Wallis (KW) test was performed. This is a test for non-parametric data, similar to ANOVA, which checks the medians of grouped data for equivalence [213]. The result of this test is shown in Table 7.5 and Fig. 7.10. Since $p > 0.05$ it can be said that was no significant effect of mix engineer on the ratings of preferences, across all songs.

Consequently, a KW test was undertaken for each song. For individual songs, the results are shown in Figs 7.11, 7.12 and 7.13. Each of these boxplots shows the distributions of the normalised preference ratings for each mix engineer. The number of participants who rated the

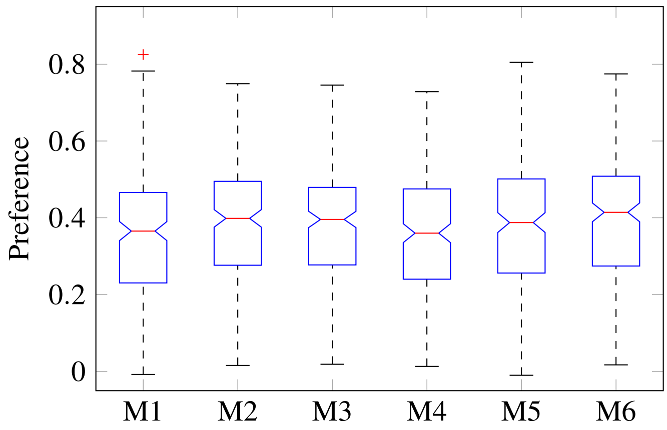


Figure 7.10: Boxplot of Kruskal Wallis test results, on entire dataset

song is indicated, as is the p-value of the KW test: where $p < 0.05$ this suggests that the null hypothesis, that the data for the different groups are drawn from the same distribution, be rejected.

Ten out of 18 songs have $p < 0.05$ indicating that, in the remaining eight songs, there was no consensus as to any observable difference between mix engineers. In some cases, this is likely due to the low number of times a particular song appear in trials. Data relating to songs for which non-significant results were obtained were removed and a KW test performed on this reduced dataset. These results are shown in Table 7.6, for the normalised scores. With $p < 0.05$, this indicated that, for songs where differences were observed, there was an observed effect of the mix engineer on the preference ratings. The effect size was calculated as follows:

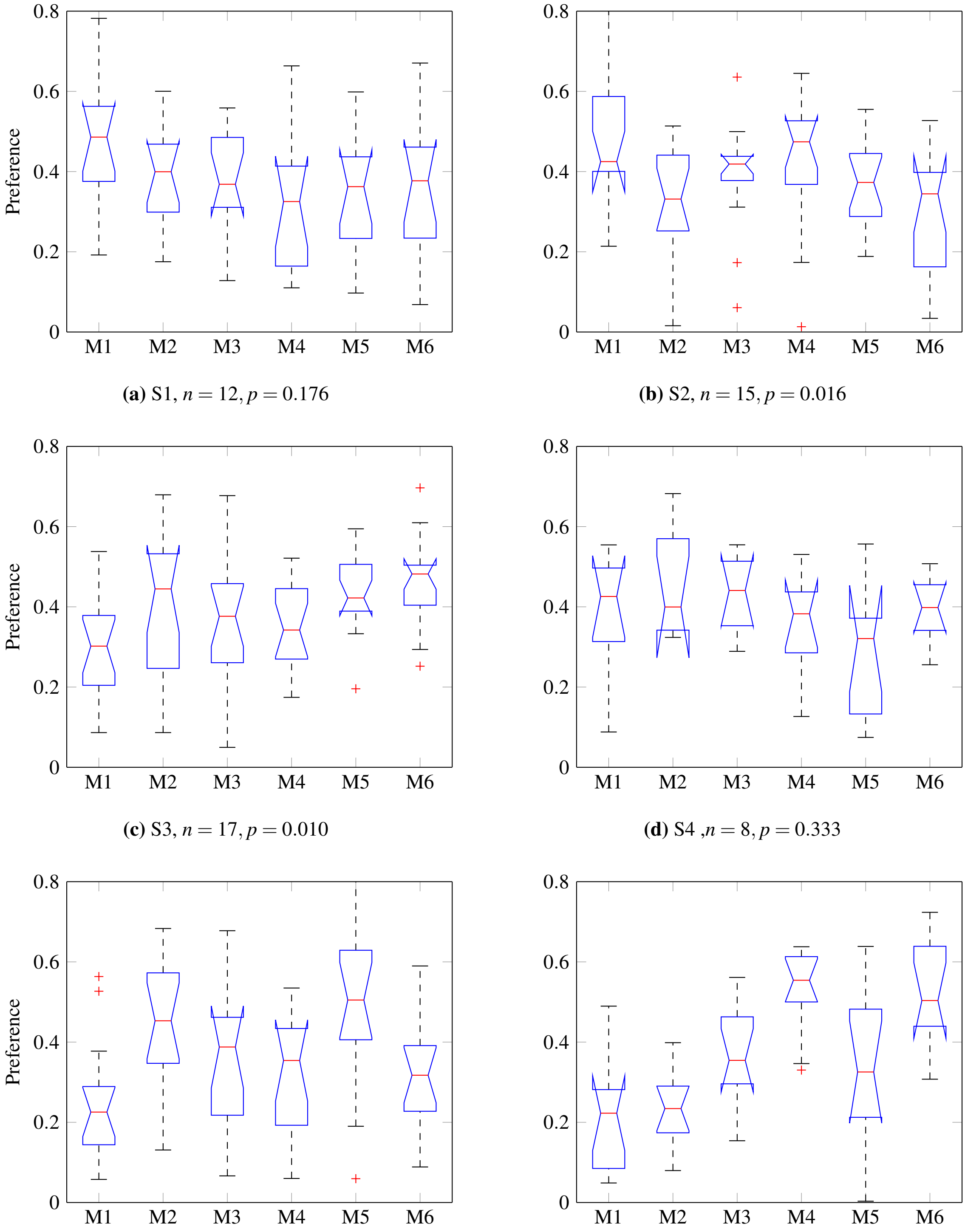
$$\eta^2 = \frac{\chi^2}{\mathrm{df}_{\mathrm{total}} - 1} = \frac{17.03}{833 - 1} = 0.021$$

This indicates that, for a collection of 10 songs for which an effect could be perceived, the amount of the variance in preference ratings that could be explained by the mix engineer was 2.1%.

Table 7.6: Kruskal-Wallis test results, for 10/18 songs

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Groups	988517.4	5	197703.5	17.03	0.0044
Error	47352555.1	828	57189.1		
Total	48341072.5	833			

The results of a multiple comparison test are shown in Table 7.7. Each row shows a comparison of one mixer (grp_1) to another (grp_2). The difference in the mean ranksum of the groups is denoted by $\Delta\mu$. The range of the ±95% confidence interval is also shown. Where this range includes zero, there is a high probability that there is no significant difference between groups. The rightmost column displays the p-value of a hypothesis test that the corresponding mean difference is equal to zero. As two rows have $p < 0.05$ it can be said that the mean ranksum of M1 differs significantly from both M3 and M6. Refer back to Fig. 7.8, where the confidence ellipses of M1 and M3 do not overlap and nor do M1 and M6. What can be seen here is that there are notable subjective *and* objective differences between these pairs of mix engineers.



(a) S1, $n = 12, p = 0.176$

(b) S2, $n = 15, p = 0.016$

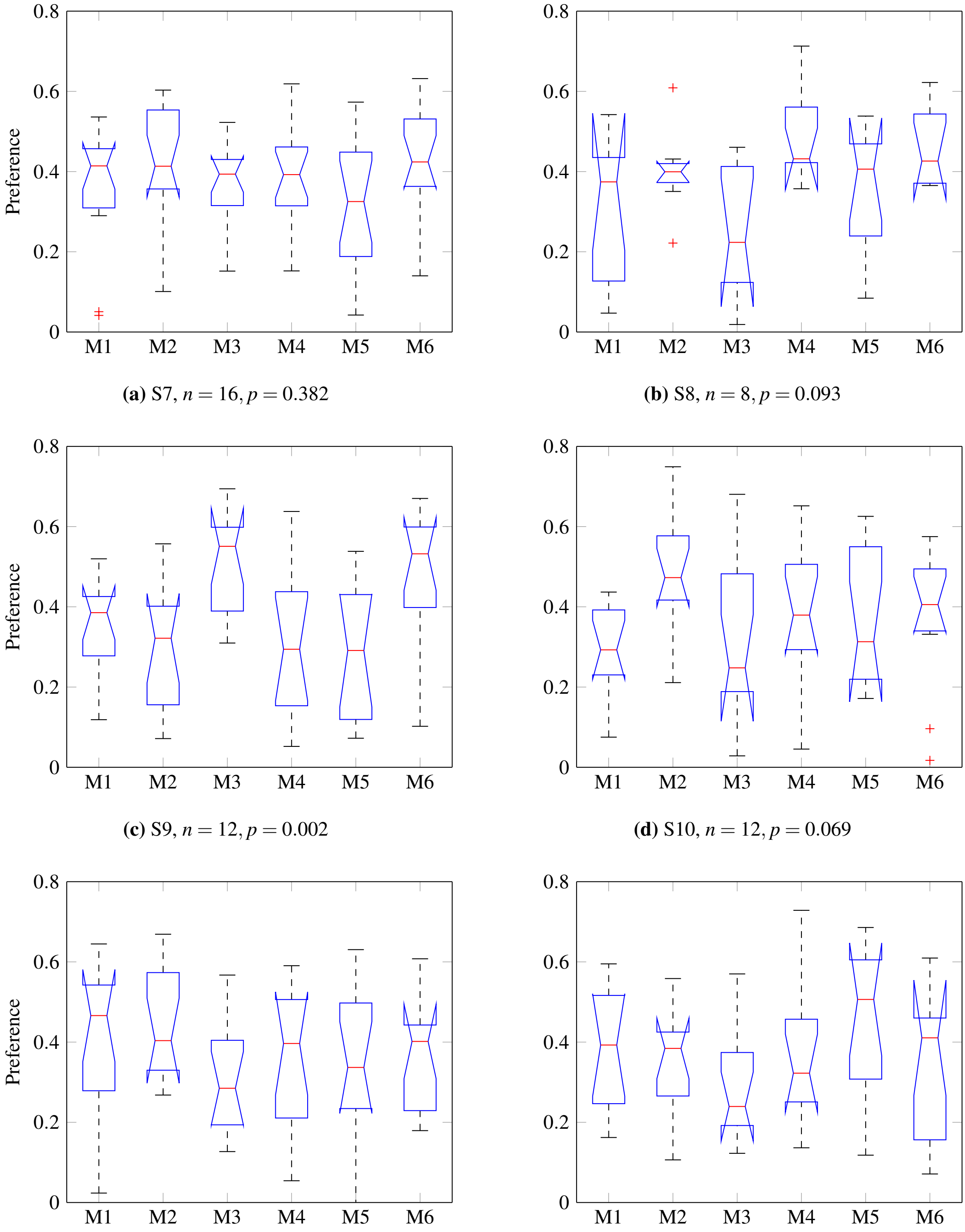
(c) S3, $n = 17, p = 0.010$

(d) S4 ,$n = 8, p = 0.333$

(e) S5, $n = 14, p = 0.004$

(f) S6, $n = 11, p < 0.001$

Figure 7.11: Kruskal-Wallis test, for songs 1 to 6



(a) S7, $n = 16, p = 0.382$

(b) S8, $n = 8, p = 0.093$

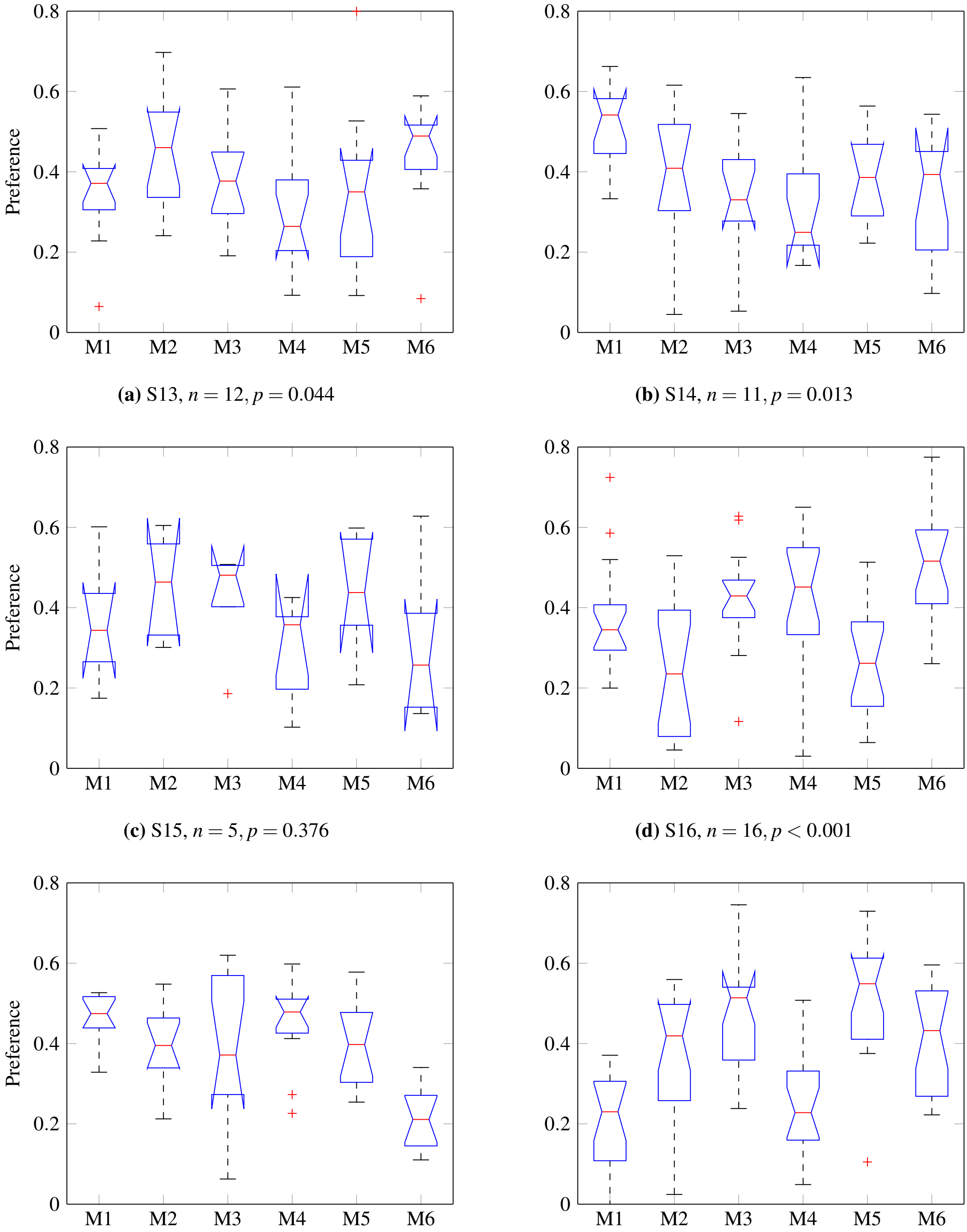
(c) S9, $n = 12, p = 0.002$

(d) S10, $n = 12, p = 0.069$

(e) S11, $n = 13, p = 0.307$

(f) S12, $n = 11, p = 0.306$

Figure 7.12: Kruskal-Wallis test, for songs 7 to 12



(a) S13, $n = 12, p = 0.044$

(b) S14, $n = 11, p = 0.013$

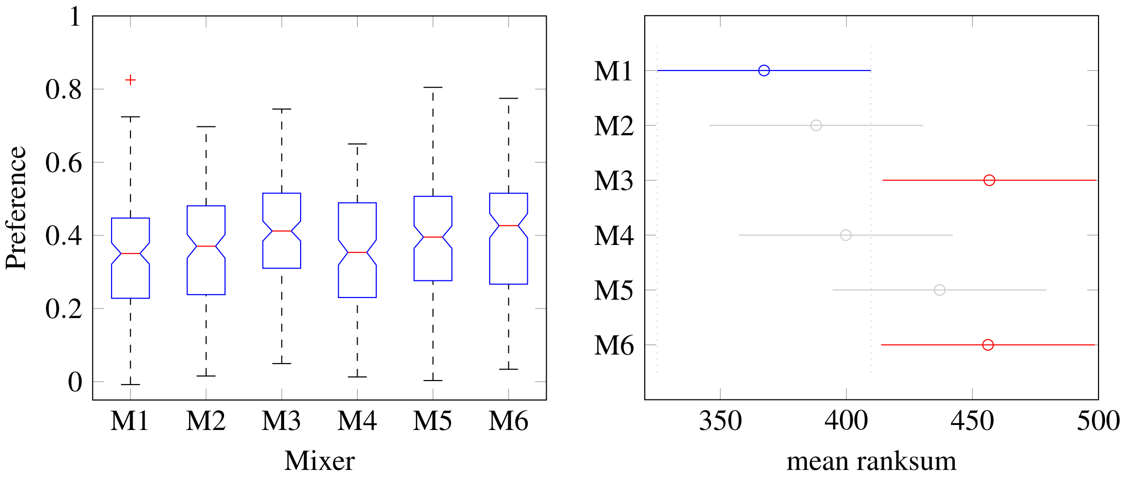
(c) S15, $n = 5, p = 0.376$

(d) S16, $n = 16, p < 0.001$

(e) S17, $n = 12, p < 0.001$

(f) S18, $n = 19, p < 0.001$

Figure 7.13: Kruskal-Wallis test, for songs 13 to 18



(a) KW test

(b) Multiple comparisons

Figure 7.14: Results of Kruskal Wallis test, on subset of 10/18 songs

Table 7.7: Table from multiple comparisons test. Bold type indicates where $p < 0.05$. There is a significant difference between the preference scores of M1 and both M3 and M6.

grp_1	grp_2	$-95\%\text{ci}$	$\Delta\mu$	$+95\%\text{ci}$	$1 - p(\Delta\mu = 0)$
1	2	-105.4713	-20.6547	64.1619	1.0000
1	3	-174.1763	-89.3597	-4.5431	**0.0298**
1	4	-117.2626	-32.4460	52.3706	1.0000
1	5	-154.4569	-69.6403	15.1763	0.2393
1	6	-173.6439	-88.8273	-4.0107	**0.0317**
2	3	-153.5216	-68.7050	16.1116	0.2614
2	4	-96.6080	-11.7914	73.0252	1.0000
2	5	-133.8022	-48.9856	35.8310	1.0000
2	6	-152.9893	-68.1727	16.6439	0.2747
3	4	-27.9029	56.9137	141.7303	0.7333
3	5	-65.0972	19.7194	104.5360	1.0000
3	6	-84.2842	0.5324	85.3490	1.0000
4	5	-122.0108	-37.1942	47.6224	1.0000
4	6	-141.1979	-56.3813	28.4353	0.7656
5	6	-104.0037	-19.1871	65.6295	1.0000

7.3.3.1 Relationship to features

Once subjective ratings were obtained from the test participants, these preference scores were compared against the audio signal features of the mix, in order to examine whether or not the features can explain why one mix engineer may be preferred over another. First, for each of the 36 extracted audio signal features, a linear fit to preference scores was made. The Pearson r and associated p-values of these fits are shown in Table 7.8. this indicates that only *Spectral Flatness* and *sbflux7* were significantly correlated to preference, in this way.

In order to gain a greater insight, the now-familiar method from Chapters 3 and 6 was used, to inspect the PCA dimensions and compare against the subjective rating. The dataset was inspected for outliers using the Z-score method. This revealed two outliers, which left 106 audio samples once removed. Using Bartlett's test of sphericity, the null hypothesis that the correlation matrix of

Table 7.8: Correlation of each variable to median preference scores. Bold type indicates where $p < 0.05$.

Variable	Pearson r	p-val
SpecCent	0.119	0.221
SpecSpread	0.153	0.114
SpecSkew	-0.062	0.522
SpecFlat	0.206	**0.033**
SpecKurt	-0.078	0.424
SpecEnt	0.103	0.287
CF	0.091	0.351
LoudITU	-0.081	0.405
Top1dB	0.073	0.451
Harsh	-0.076	0.433
Sub80	0.076	0.434
RO85	0.138	0.153
RO95	0.135	0.164
sbflux1	0.116	0.232
sbflux2	-0.051	0.597
sbflux3	0.012	0.904
sbflux4	-0.056	0.564
sbflux5	-0.038	0.700
sbflux6	0.094	0.332
sbflux7	0.213	**0.027**
sbflux8	0.129	0.184
sbflux9	0.006	0.950
sbflux10	0.042	0.663
Gauss	0.141	0.146
PMFcent	0.067	0.491
PMFflat	-0.090	0.355
PMFspread	-0.052	0.593
PMFskew	0.017	0.865
PMFkurt	0.047	0.628
W-all	0.154	0.111
W-band	0.085	0.384
W-low	0.102	0.295
W-mid	0.052	0.593
W-high	0.014	0.884
SMratio	-0.038	0.693
LRimbalance	-0.093	0.340

the data is equivalent to an identity matrix was rejected.

$$\chi^2(630, N = 106) = 5212.23, p < 0.001$$

This indicates that factor analysis can be performed, while a Kaiser-Meyer-Olkin measure of sampling adequacy of 0.722, above the recommended value of 0.6 [133], suggests that factor analysis would be useful. When the KMO of each variable was obtained, eight variables had values below the cut-off value of 0.6. These eight features (Harsh, PMFcent, PMFskew, Wband, Wlow,

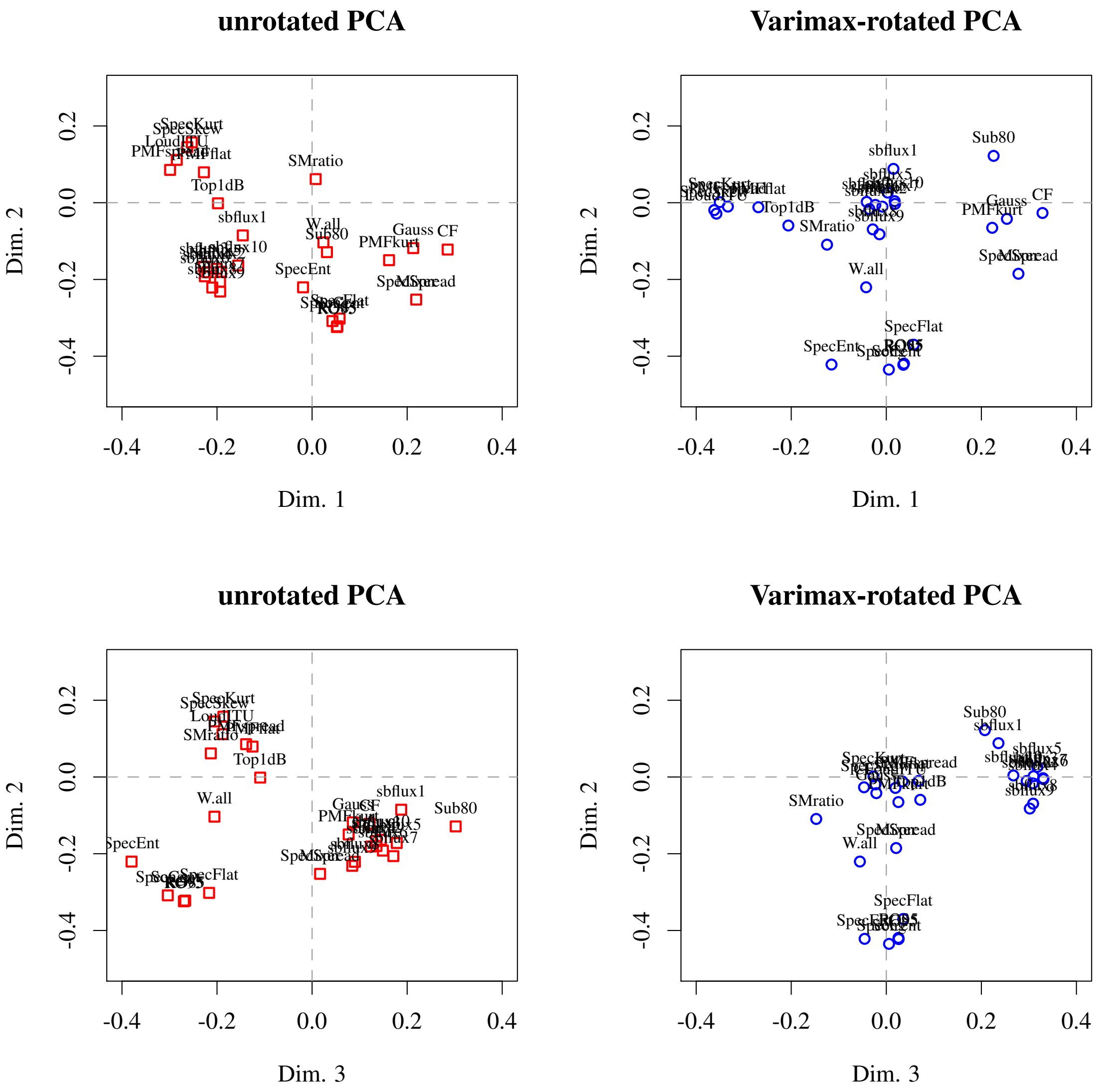


Figure 7.15: PCA for 108 mixes rated in online test. This shows the importance of rotation.

Wmid, Whigh, LRimbalance) were therefore removed. PCA was performed with the remaining 28 variables. Using the nFactors package, three components were kept from this initial PCA result. The revised PCA used only the first three components and varimax rotation was applied. These first three components explains 64.78% of the variance in the features. The subjective preference values were then compared directly to the rotated PCA scores.

While preference scores were significantly correlated to *dim.2*, it's hard to say that, overall, less bright mixes have lower preference. What may be happening is that they are lower preference if less bright than what it considered typical for that particular song. Recall that different songs can occupy different regions of the PCA-space (as in Fig. 6.2a). For each song, the mean value along each dimension was calculated, then the difference from the mean was recorded form each sample. This new variable is plotted against preference scores in Fig. 7.18. However, as are only six mixes for each song, care must be taken in interpreting the data, as the mean may not be

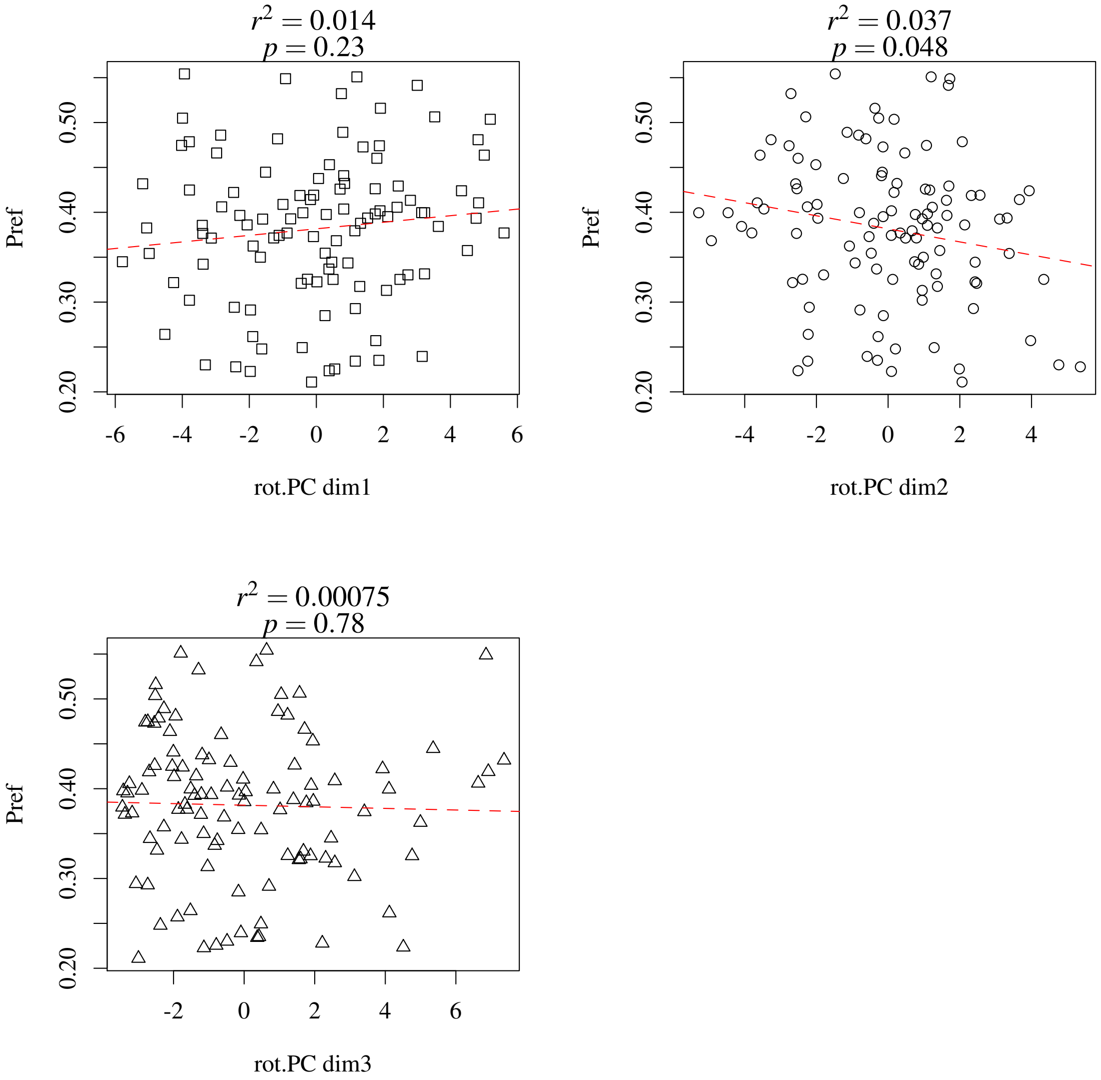
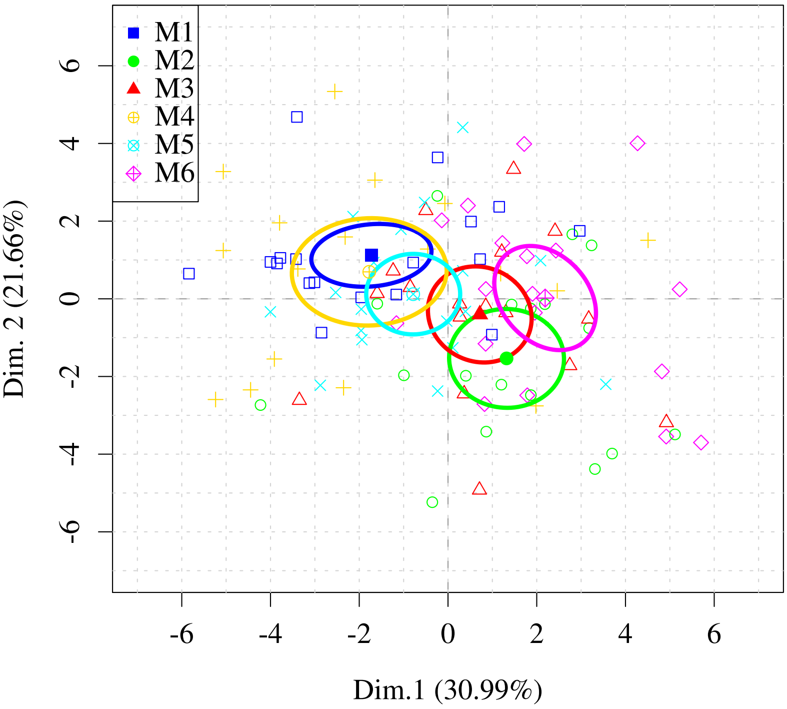


Figure 7.16: Preference plotted against rotated PCA dimensions. There is a statistically significant linear fit for median preference ratings against dimension 2, indicating that brighter sounding mixes were preferred.

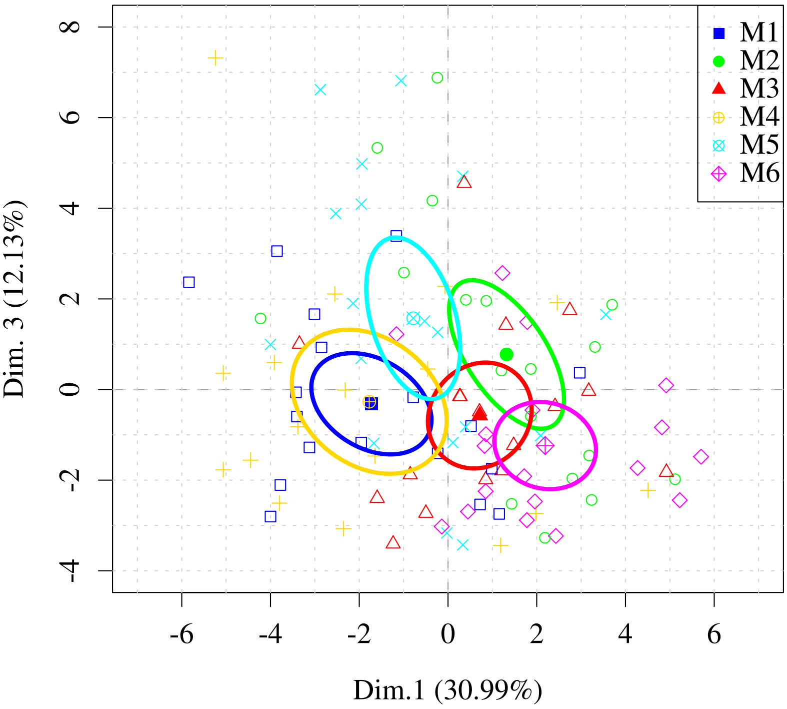
reliable.

Figure 7.18b shows an increased level of correlation when compared to Fig. 7.16. Interestingly, when the same principle is applied for *dim.1*, a relationship between scores and preference is revealed. Figure 7.18a shows the fit of a fourth-order polynomial to the data. This suggests that when mixes were louder and less dynamic than the average for that song they were preferred, up to a point. However, the opposite is also true, that more dynamic than average mixes were preferred, up to a point. Data for *dim.3* is not shown here, as no additional insights were revealed by this analysis. The finding that brighter-sounding mixes were preferred is at odds with other findings within this thesis.

- Table 6.14 showed that mixes with greater *spectral centroid* and greater *rolloff* were less preferred. However this was for just 27 mixes of one song.

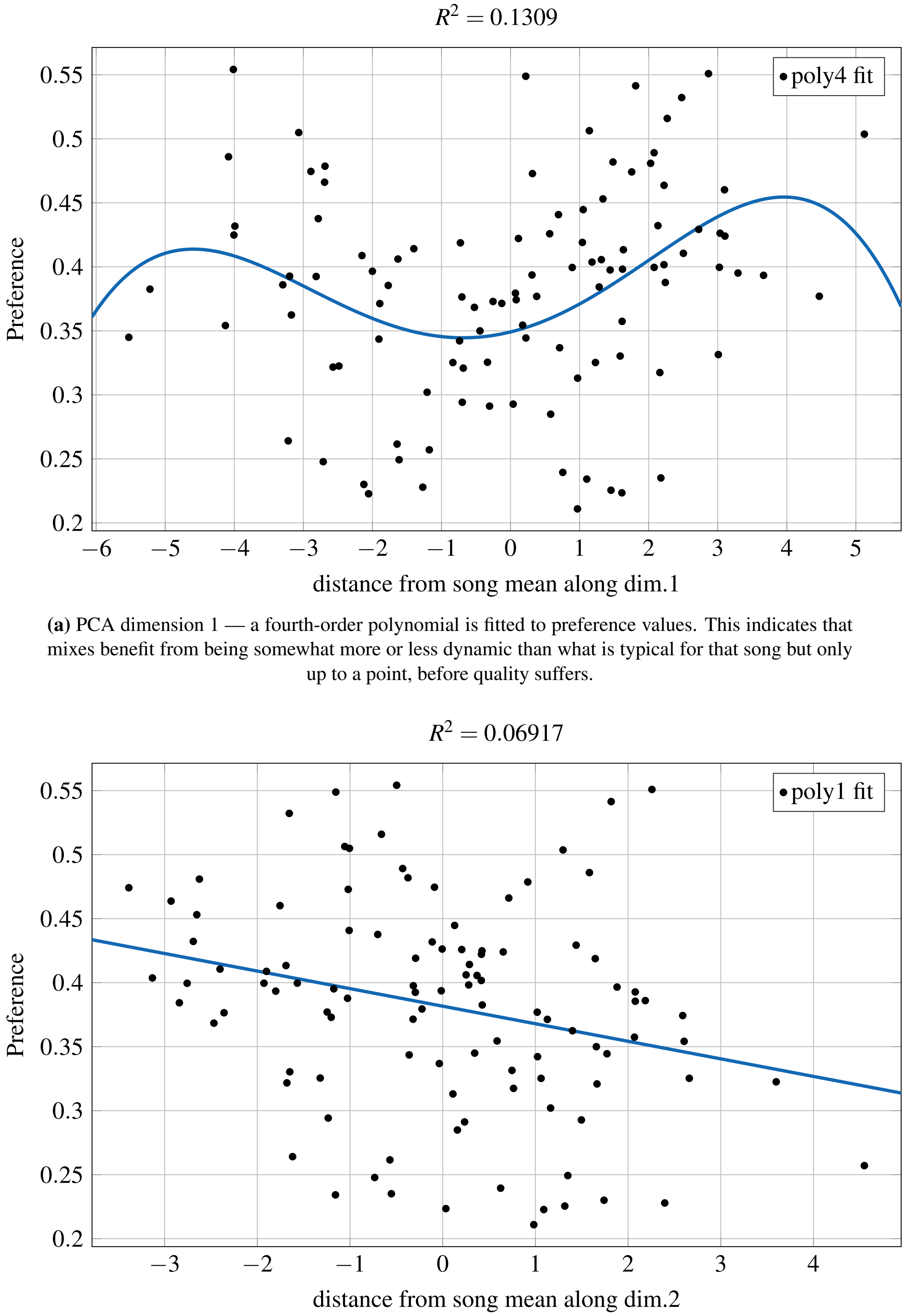


(a) PCA dimensions 1 and 2



(b) PCA dimensions 1 and 3.

Figure 7.17: Individuals factor plot for sonic signatures data. The group centroids are plotted along with the 95% confidence ellipses. This indicates that, among certain pairs of engineers, there is evidence to suggest the mixes they create are significantly different, on average



(a) PCA dimension 1 — a fourth-order polynomial is fitted to preference values. This indicates that mixes benefit from being somewhat more or less dynamic than what is typical for that song but only up to a point, before quality suffers.

(b) PCA dimension 2 — a first-order polynomial is fitted to preference scores. This indicates that mixes benefit from increased focus on high-frequency content.

Figure 7.18: Relationship between preference scores and PCA dimensions.

- Table 6.11 showed that brighter-sounding mixes were less likely to do well in a particular mixing competition. However this was for 98 mixes of the same song as above.

- Figure 3.7 showed that greater values of *rolloff* indicated songs that were less liked. However, this was not for mixes but for 63 different songs.

It is difficult to reconcile these seemingly-conflicting findings, however, the result in this chapter is the only one which includes multiple mixes of multiple songs. As such, it is not critical that all of these findings support one another. The preference for reduced brightness in mixes of "Blood To Bone" may be song specific — mixes of this song did display some of the lowest spectral centroid values when compare to nine other songs in Fig. 6.5. According to Table 6.6, the average spectral centroid over all 1501 mixes was 3.5 kHz.

As shown in Fig. 7.1, only M2 and M5 had a median spectral centroid value close to this, while the other four mix engineers had median values below this. Perhaps a reason why brighter-sounding mixes were preferred here is that, in *this* dataset, *brighter* actually means closer to the global average. Additionally, as these six mix engineers were some of the most regular contributors to the forum, we know that they have produced hundreds of mixes, while, in Fig. 6.5, many of the mixes may have been created by less experienced mix engineers.

Of course, as these results come from a data-driven study, care should be taken when trying to generalise the findings within this chapter to the art of mixing as a whole.

7.4 Chapter summary

Out of six mix engineers, creating mixes for 18 songs, the results suggest that they can not strictly be classified from one another at this stage but that they are arranged into two clusters: one group of bright and "toneful" mixes and one group of darker, "noisier" mixes. In § 7.3, the subjective nature of audio perception was incorporated into the model. A subjective test was undertaken which revealed that the effect of the mix engineer on the preference score of a mix is only a small effect ($\eta^2 = 0.021$) and was only observed in 10/18 songs used. Of the two objective methods (evolutionary pursuit and PCA) and the subjective test results, there was agreement that certain pairs of mix engineers had sufficiently varied styles: M1 was measurably distinct from both M3 and M6.

In addition to variance among mixes, as shown in Chapter 6, variance among mix engineers was also observed, in this chapter. Both of these findings are novel and important. We now know that while mixes, on the whole, differ from one another in some predictable way, and that features vary based on simple parametric models, additionally, individual mix engineers are shown to vary, in a purely feature-based model. If it is true that the audio signal features can tell us something interesting about the audio signal, then it can be said that quantifiable evidence now exists to suggest that mix-engineers do have a measurable style, which has been suggested anecdotally for some time.

While it is hard to draw definitive conclusions, this study has illustrated that a weak effect of mix engineer can be measured using these methodologies. Further work is encouraged, exploring alternate test methods and datasets. Ultimately, the differences between alternate mixes can be subtle and further attempts to uncover the differences between mix engineers will benefit from novel signal features, specifically developed for measuring these subtle variations.

8

Design of an evolutionary music mixing system

As introduced in § 2.4, an evolutionary algorithm can be described as a search or optimisation algorithm which utilises mechanisms inspired by biological processes. Algorithms have been inspired by genetic reproduction and mutation [83, 84], bees searching for pollen [214, 215] and animal flocking behaviours [216], to name a few. These methods, in general, are not deterministic, meaning a solution is rarely determined outright but rather it is approached from a variety of directions. This makes such methods particularly suitable to problems related to design and aesthetics and they have been used in a number of studies where aesthetic choices are to be made by an algorithm, such as music composition [217], sound design [218] or the production of logos and other graphical art [86].

Throughout these studies there is the notion that individual design problems have individual design spaces of a defined topography. This is an idea upon which the mix space study is based. If the creation of a mix from multitrack audio can be considered as a design problem, combining aesthetic considerations with technical limitations, then the exploration of such a space using EC methods could provide a novel contribution to the field.

Typically, in implementing evolutionary algorithms, a fitness function is required in order to determine which solution (music mixes, in this context) should be considered as the best (as in § 7.2.2). By contrast, in aesthetic problems, the user often selects the best solutions in a given generation. This second approach has been referred to as an interactive evolutionary algorithm with a human-in-the-loop acting as the fitness function [86]. As this can be time-consuming, especially for large populations of candidate solutions, automatic methods of establishing fitness have been proposed for certain tasks [85, 88, 219].

For the task of comparing alternate mixes, a hybrid approach is proposed in this chapter: a human evaluator offers explicit ratings for a subset of mixes and the fitness of the unrated population is estimated using heuristic rules obtained from earlier studies (such as preference for mixes with certain spectral characteristics). It may also be possible for the system to be trained by the user, so that over time, this estimation process is improved as the system learns the preferences of that specific user.

In addition to providing a novel method for the study of music mixing such an algorithm could also function as an interface for musical expression. Whereas many automated/intelligent music production tools aim to conduct tasks in place of a user, the proposed system could require human input to guide the mixing process; the goal is not to find the 'best' mix, but the best mix for that specific user. Such a system could be of particular use to the visually impaired, or user with reduced mobility, for whom the conventional approach to music mixing might be problematic.

In summarising the thesis thus far, the motivation for the work in this chapter becomes clear. From Chapter 3 we know that the quality of a mix is dependent on subjective impressions as well as objective measures. Additionally, from Chapter 7, there is evidence to suggest that listeners can perceive the different styles of mix engineers. These points suggest that it is important to allow the user to guide the system. The proposed intelligent mixing system must satisfy the following requirements.

- Explore a space that is representative of the mixing process.
- Approach the solution from more than one direction
- Acknowledge that more than one optimal solution may exist
- That the optimal solution(s) may vary from user to user.

The theory from Chapter 4 provides a space in which to generate mixes. Chapter 5 describes a method of generating a random population of mixes, which is the first step in an evolutionary algorithm. Chapter 6 suggests that mixes exhibit central tendency, therefore providing some rules to help guide the system, in addition to the guidance of the user. Consequently, all of the necessary critical and theoretical framework in developing the proposed system has been outlined.

8.1 Method

The flowchart in Figure 8.1 demonstrates the method used in the design of an IGA-based music mixing program. The important steps in the flowchart are summarised as follows and are each described further in the following subsections.

1. Import audio and normalise
2. Initialise population
3. Choose sub-population
4. Evaluate sub-population
5. Allocate fitness
6. Genetic operations
7. Stop criteria
8. Choose best mix

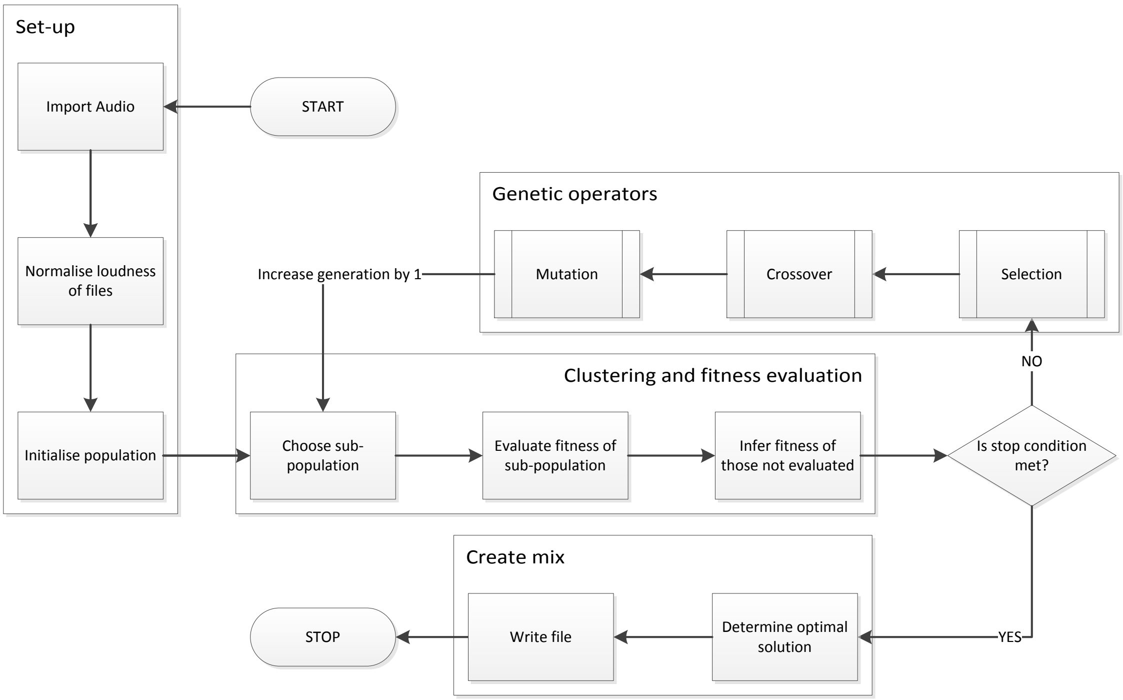


Figure 8.1: Flowchart of IGA mixer

8.1.1 Import audio and normalise

The audio which was used for the developed system is of the form listed below: there are a total of six tracks where each is a single-channel .WAV file, PCM encoded at a sampling rate of 44.1 kHz and a bit depth of 16-bits. The six tracks represent the following six instruments: vocals, guitar, bass guitar, snare drum, kick drum, drum overhead. Most of this audio was prepared for the experiments in § 4.5 and Chapter 5. This precise choice of track numbering allows the five coordinates in the mix-space to have some clear semantic meaning. As in Chapter 4, ϕ_1 indicates the balance of the vocal to the backing tracks, ϕ_2 is the balance of the guitar to the "rhythm section" of drums and bass, and so on, as displayed in Fig. 8.2. This ordering of tracks could be also be random, or some arbitrary order. The ordering of tracks does have some influence on the performance of the system, as will be discussed later in § 8.2.

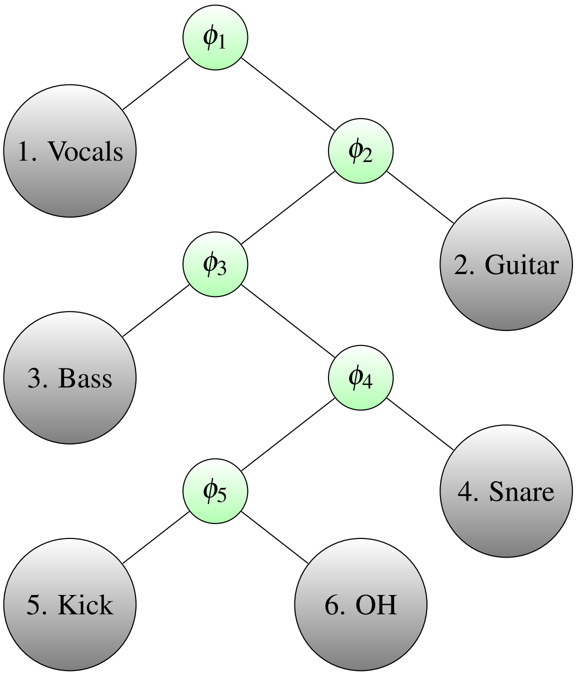


Figure 8.2: Representation of ϕ terms in a session of six audio tracks. Each of these terms decribes a specific balance between instruments or sets of instruments, as illustrated here.

As before, in § 4.3, when dealing with narrowband content, such as the individual tracks in a multitrack session, loudness was normalised according to a modified form of ITU-BS.1770 [158]. This ensures that the loudness of each track in a mix can be retrieved directly from the gain vector and, more crucially, that all points in the mix-space have the same perceived loudness (as shown in Fig. 5.9).

8.1.2 Initialise population

The initial population of mixes, the population that will be optimised, is created using the method described in Chapter 5. Recall the two methods proposed:

- Uniform selection on unit $(n-1)$-sphere
- von-Mises-Fisher distribution around an assumed good mix (where assumption is based on mix-space results).

Being points of the surface of a unit $(n-1)$-sphere ensures that the norm of the gain vector is equal to 1. This has the advantages that each mix is presented at roughly equal loudness (as demonstrated in Fig. 5.9) while also having sufficient headroom to avoid clipping.

While the vMF method proved to be useful in Chapter 5, in this case, it is desirable to begin with no assumptions as to what mix would be the ideal mix. If this system is to be used to create not only alternate mixes (mixes where all of the original mixes are present but with an alternative balance) but ‘remixes’ (mixes wherein some elements may be omitted in order to more radically change the presentation of the song) then the system must begin as a blank slate, with no assumptions. The equal-loudness with vocal boost, which formed the initial assumption behind the random mixes in Chapter 5 (the vector μ in Eqn. 5.5) is therefore not used here. With no estimate for μ in a vMF distribution, the first method is employed instead, and mixes are randomly chosen from the unit $(n-1)$-sphere by uniform distribution [1].

8.1.3 Choose sub-population

Since the population may be quite large, direct evaluation of each point can be fatiguing to the user. Rather than directly evaluate the entire population, the user only rates a sub-population of size c. This greatly reduces the level of user burden. To achieve this, the total population is divided into c clusters and a single representative mix is taken from each cluster. There are two questions which need to be addressed.

- In which domain would it be best to create clusters — the mix-space ($\mathbb{S}^{n-1}$) or the ambient gain-space ($\mathbb{R}^n$)?
- Knowing this, which clustering algorithm and/or distance measure is most suitable?

In k-means clustering, where c is the centroid and x is the feature vector, the aim is to minimise some measure of distance between each point and a cluster centroid, for some chosen number of clusters. A simple, common measure is a Euclidean or squared Euclidean distance.

$$d(x,c) = |x-c|^2 \tag{8.1}$$

To measure the similarity between two vectors the cosine similarity measure can be used. For use as a distance measure Cosine *dissimilarity* is defined as follows. This is particularly useful for clustering points on an hypersphere, as that surface is not Euclidean.

$$d(x,c) = 1-\cos(x,c) = 1-\frac{\langle x,c\rangle}{||x||\,||c||} \tag{8.2}$$

A simple test was conducted to determine the most suitable clustering technique. The results are shown in Fig. 8.3 and Fig. 8.4. The number of clusters was chosen to be five. The population size is deliberately large so that the convex hull of the population approximates a sphere. Figure 8.3 shows the result of squared Euclidean-based clustering on $\mathbb{S}^2$. This outcome shows that the clustering near the “north pole” (where $g_1 \approx 1$) is not correct, as two clusters merge approaching this point. Figure 8.4 displays the clustering due to the cosine metric on $\mathbb{S}^2$. Since clustering is based on distances between vectors projected from the origin, the resulting clusters do not translate well to $\mathbb{R}^3$. It is clear that the mixes with high values of g_1 and therefore low values of g_2 and g_3 are assigned to a variety of clusters, despite their obvious perceptual similarity.

Figure 8.5 and 8.6 show the results of clustering in $\mathbb{R}^3$, based on the squared Euclidean and cosine metrics respectively. For both of these cases, the points close to each corner belong to a unique cluster. The cosine metric, being based on the distance between vectors drawn from the

[1] A uniform distribution could also be obtained using μ as before and simply setting the concentration parameter $\kappa = 0$

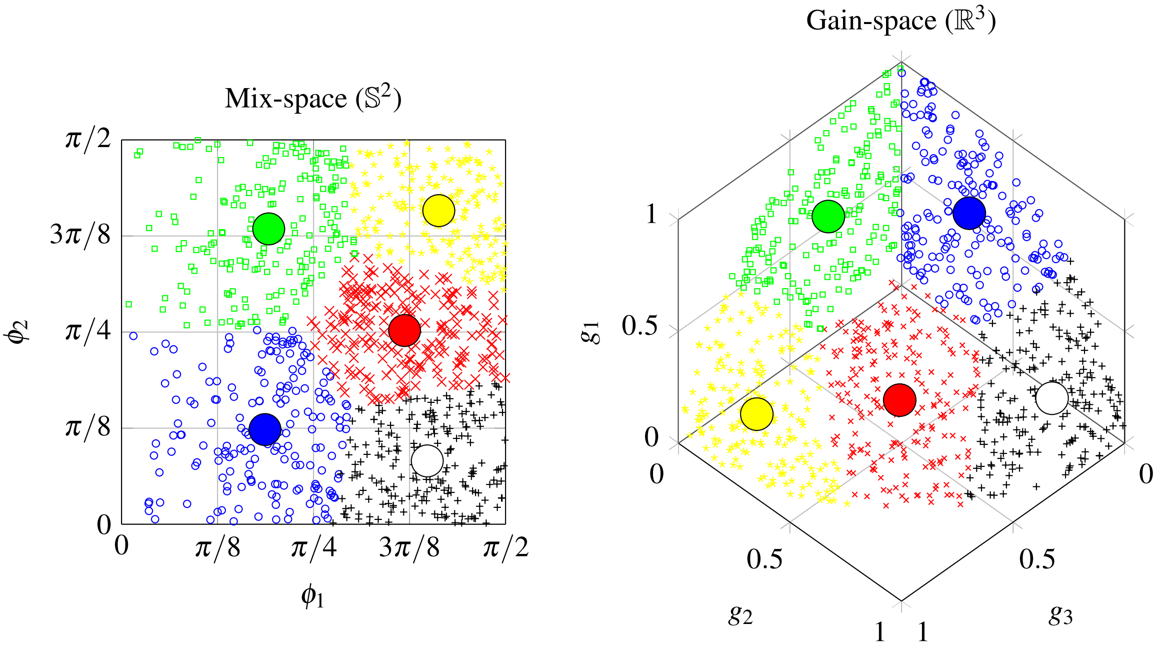


Figure 8.3: *k*-means in mix-space, with squared euclidean distance metric

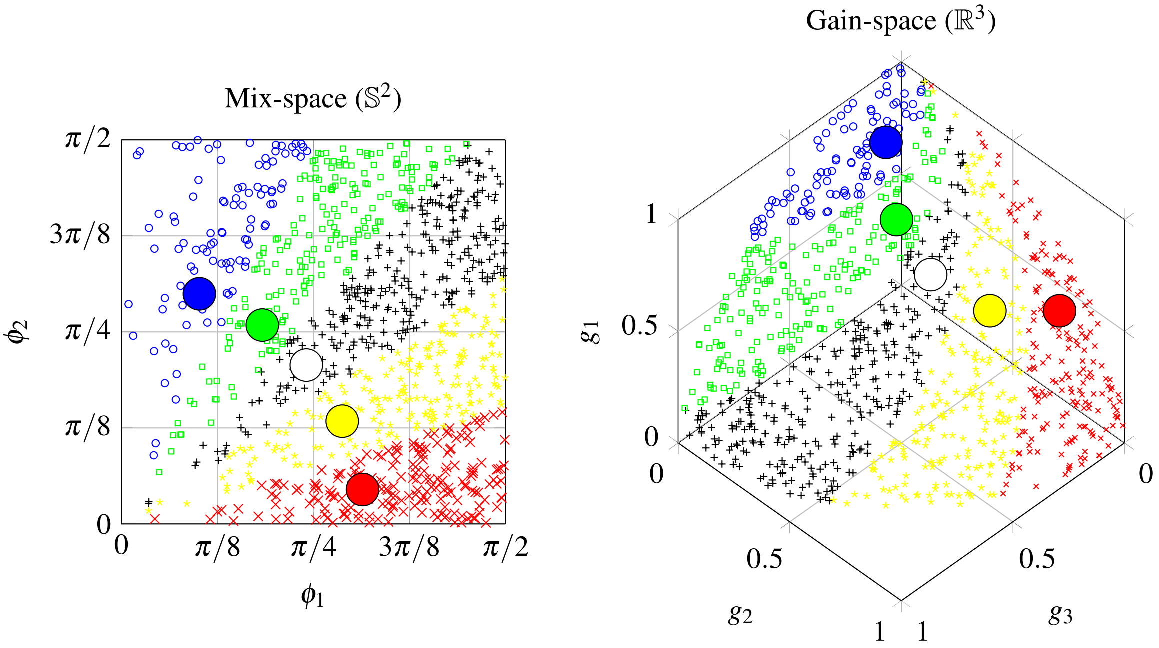


Figure 8.4: *k*-means in mix-space, with cosine distance metric

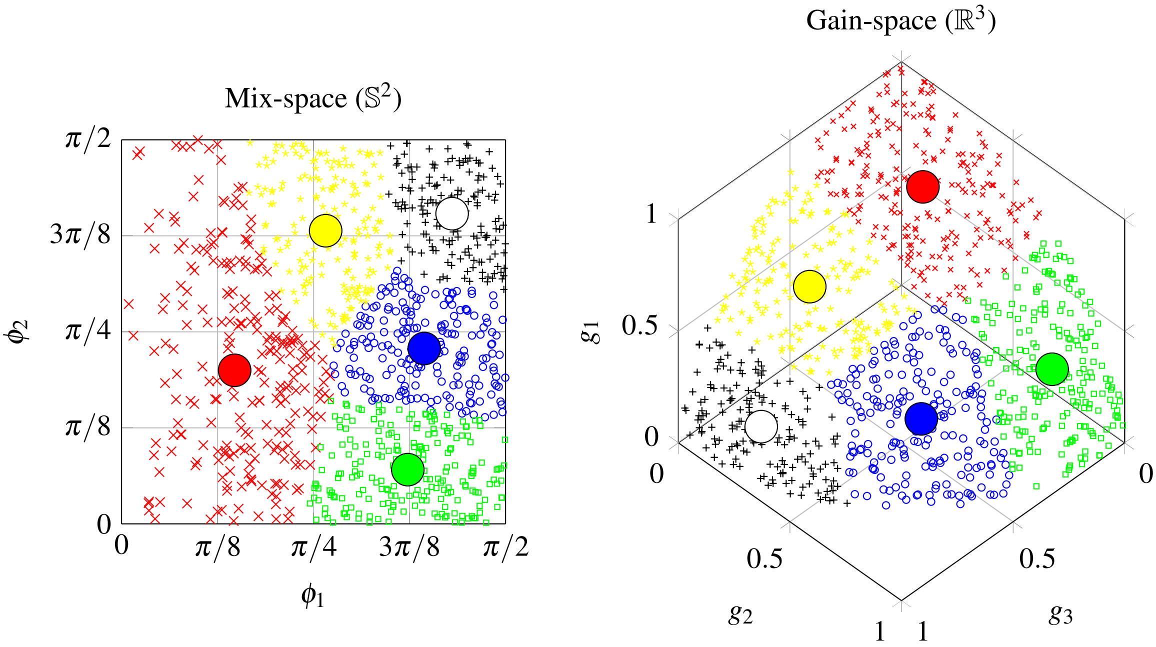


Figure 8.5: *k*-means in gain-space, with squared euclidean distance metric

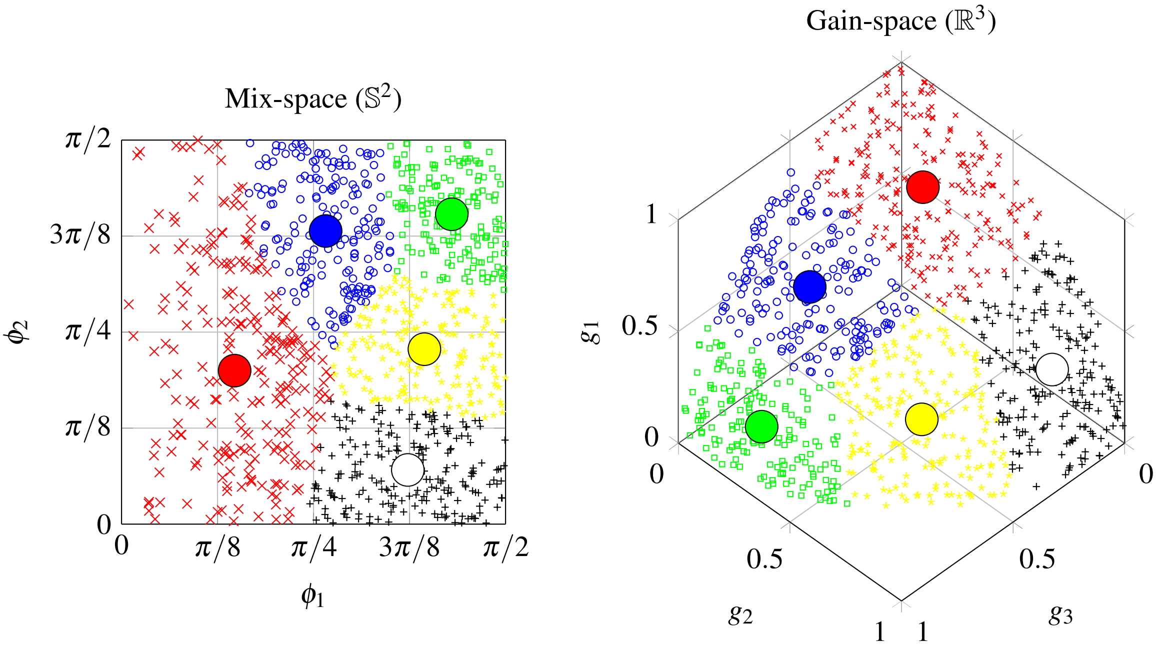


Figure 8.6: *k*-means in gain-space, with cosine distance metric (spherical *k*-means)

origin, is the more appropriate choice for spherical data. The use of this metric in k-means clustering is often referred to as spherical k-means clustering [160]. The chosen domain for clustering was therefore $\mathbb{R}^n$ and the cosine metric was used. It is worth noting, that for any sufficiently large population of uniformly-distributed random points on $\mathbb{S}^{n-1}$, the locations of the centroids would be comparable, for a particular distance metric. Since the sub-population is comprised of the individual solutions closest to the centroid, the initial sub-population can be well-predicted in advance.

8.1.4 Evaluate sub-population

Once the sub-population is determined, the fitness of each solution is evaluated. How this is achieved depends on the fitness function. In a standard EC approach, this function is well-defined. For IEC applications, the fitness is evaluated by the user (see Fig. 2.11) but can be augmented by an objective function [87]. In this system, each mix in the sub-population is generated and played back to the user. The user then directly evaluates each mix, independently, according to the desired criteria, and is prompted to assign an explicit rating that can be collected by the system.

8.1.5 Allocate fitness

Since only a sub-population is evaluated, the fitness of the remaining population must be estimated. This was done based on the assumption that nearby mixes share many common attributes and are perceptually similar. The primary method of inferring fitness of an unevaluated mix was to use the distance to the evaluated mix (the mix closest to the cluster centroid). Each mix within a cluster is awarded the same fitness as the evaluated representative and then an offset is subtracted, proportional to the distance from the centroid. Refer to Lee and Cho [88] and Kim and Cho [85] for a description of this approach. There are also a number of more recent papers which summarise this type of fitness estimation, reviewed by Takagi [87].

Generally speaking, audio signal features of the mixes can also be used in the fitness estimation of unevaluated mixes. This was not included in this working example program, which is the feature of this chapter and the next, but is discussed in § 8.4.2.

8.1.6 Genetic operations

In this example, while clustering takes place in $\mathbb{R}^n$, all genetic operations are performed in $\mathbb{S}^{n-1}$. This ensures that the offspring produced by crossover and mutation are always on the hypersphere in $\mathbb{R}^n$. Prior to genetic operations, the real-valued coordinates on $\mathbb{S}^{n-1}$ were first converted to binary strings as follows. When the values of g are positive, the range of Φ is from 0 to 2π. To convert to a binary representation, first the range is re-scaled to [0, 1] then multiplied by $2^q - 1$, where q is the number of bits used in the binary representation. This has a range of [0, $2^q - 1$]. These values are converted to binary strings using the Matlab function `dec2bin`. In this example, $q = 7$, allowing 128 levels for each variable. As an individual in the population is comprised of $n - 1$ coordinates, the values of each individual dimension were converted to a q-bit binary string and then concatenated. The individual is then represented, finally, as a $q(n-1)$-bit binary string.

8.1.6.1 Selection

To aid selection, fitness values were scaled according to their rank in the population [84]. Raw fitness values are scaled according to Eqn. 8.3, where r is the rank of the individual, when sorted by fitness. The results is a set of scaled fitness values in the range [0,1]. This has the following

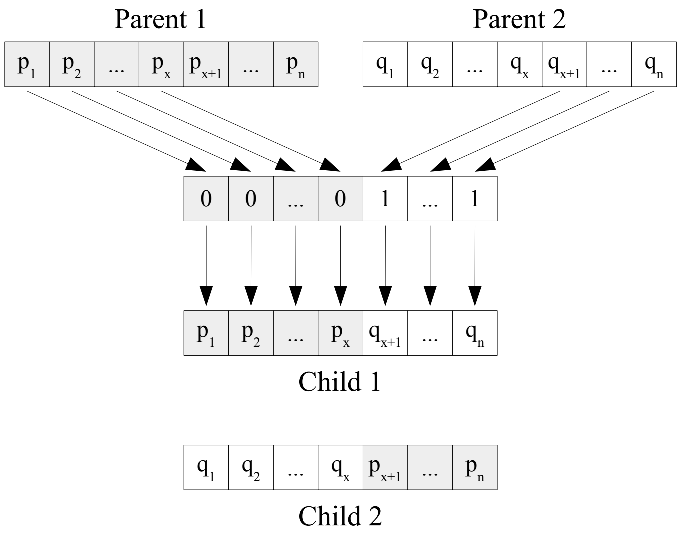

Parent 1
p_1 p_2 ... p_x p_{x+1} ... p_n
Parent 2
q_1 q_2 ... q_x q_{x+1} ... q_n
0 0 ... 0 1 ... 1
p_1 p_2 ... p_x q_{x+1} ... q_n
Child 1
q_1 q_2 ... q_x p_{x+1} ... p_n
Child 2

Figure 8.7: Example of a single-point crossover.

advantages.

- Ensures that fitness values are positive.
- Ensures that the range of fitness in each generation is equal.
- Prevents the emergence of “superindividuals”, whose fitness is so much higher than others as to dominate the competition in breeding.

$$f_{\text{scaled}} = \frac{1}{\sqrt{r}} \tag{8.3}$$

8.1.6.2 Elites

A proportion of the population automatically survives to the next generation. These individuals are referred to as ‘elites’ or ‘elite children’. In this case, the individuals with highest fitness are carried forward. This ensures that high-fitness solutions are not lost by the processes of crossover and mutation.

8.1.6.3 Crossover

The crossover function (XO) is important because it promotes diversity in the population of solutions, helping to prevent the algorithm getting stuck in local minima. A number of alternative crossover functions were tested in order to choose the most suitable for this problem.

Single-point XO: The single point XO is perhaps the most simple to visualise and implement. A single point along the bitstring is selected at random and the strings of each parent are spliced together at this point. This can be thought of as passing each parent string through a binary mask and joining the resulting sections. In some implementations, a second child is generated using the inverse mask. This is depicted in Fig. 8.7. For this application, with

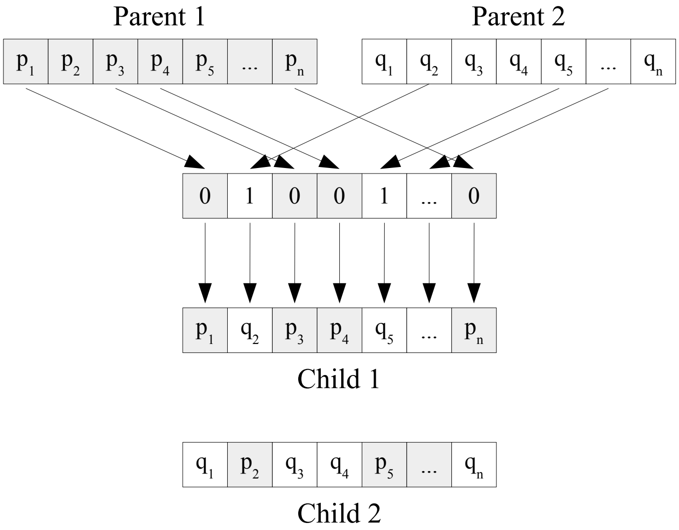


Figure 8.8: Example of a uniform crossover.

the strings typically being rather long (the length is $q(n-1)$ bits, so roughly 35, assuming 7 bits and 6 tracks, as used herein), a single crossover does not provide enough diversity. This can be thought of as a child having the left arm of the mother and the right arm of the father and a sibling with the opposite — it is possible that neither child will adapt and survive. A double-point XO is similar except two points are chosen at random and the mask alternates from zeros to ones to zeros again. This type of crossover was not tested for this application.

Uniform XO: A uniform crossover works in a similar way to the single-point and double point XO functions except the binary mask is generated as a random string. In this case, each bit in each parent has an equal chance of being put forth to the child string. When using this function, an inverse mask could be used to make a sibling or another random string could be created instead. The latter approach was used here. In informal tests, the performance of the uniform XO was improved over the single-point XO, as the diversity of the population was greater. This allowed the population to better explore the space and increases the likelihood of convergence towards an optimal solution.

Multi-parent XO: It is possible to expand the uniform XO to more than two parents. For example, if creating a child string from three parent strings, what is needed is simply a random ternary mask. This can be further expanded to an n-parent uniform XO, using an n-ary mask. Other, more sophisticated, n-parent implementations of genetic algorithms are discussed in the literature and show promise in multi-objective optimisation problems [220, 221]. For the single-objective application in this chapter, uniform XO was deemed to provide sufficient diversity.

8.1.6.4 Mutation

Individual solutions also undergo mutation, which promotes diversity in the population. In this case, a fraction of the total bits in each solution is randomly chosen to undergo mutation. The greater this fraction the more noticeable the mutation. For each of these randomly-selected indices in the bitstring the value is changed from a 0 to a 1, or vice-versa.

8.1.7 Stop criteria

A variety of stop criteria can be used in this type of genetic algorithm. The most simple would be to stop after a fixed number of generations. Alternatively, evolution could cease once the population has converged towards a sufficiently small region of the solution space. It was decided that it would be more appropriate to use a fixed number of generations, as this would keep the duration of subjective tests to a predictable timescale. It is also possible that, by using the latter method, the system would not always converge.

8.1.8 Choose best mix

Typically, in evolutionary algorithms, the best solution is considered to be the solution with the highest fitness. There are a number of reasons why this approach is not suitable here.

1. Since only a sub-population were directly evaluated, the fitness of the majority of the population is only estimated. Since fitness was subtracted in proportion to distance from the evaluated individuals, the individual with the highest fitness will always be one of the directly-evaluated sub-population.

2. Many problems that can be addressed by IEA are perceptual and as such do not require *exact* solutions but rather seek to identify an area of the solution space in which many good solutions exist which are perceptually similar [87]. In a music mixing problem there is a limit to the precision required when determining gain values, as small adjustments in the gain of individual tracks will not be perceived. To determine this precisely would involve using the spectrum of each track to determine the inter-channel perceptual masking. This is left to further work.

Assuming the population converges on a small region of the solution space, the centroid would be the most appropriate choice for the optimal solution, or ‘best’ mix. Determining this point employed kernel density estimation (KDE) methods. Two methods were tested here:

- Multiple univariate KDE, where the density of the population is evaluated for each dimension individually.

- Multivariate KDE, where the density of the population is determined in the multivariate space.

The multivariate approach is more scientifically sound but also a more complex, thus slower, calculation. The results from both methods were compared. Univariate KDE was determined using the `ksdensity` function in Matlab, as used in previous chapters. Figure 8.9 shows the univariate KDE result. The peaks in the density function are determined using the `findpeaks` function in Matlab. It is important to recognise that there may be multiple peaks. Therefore, a

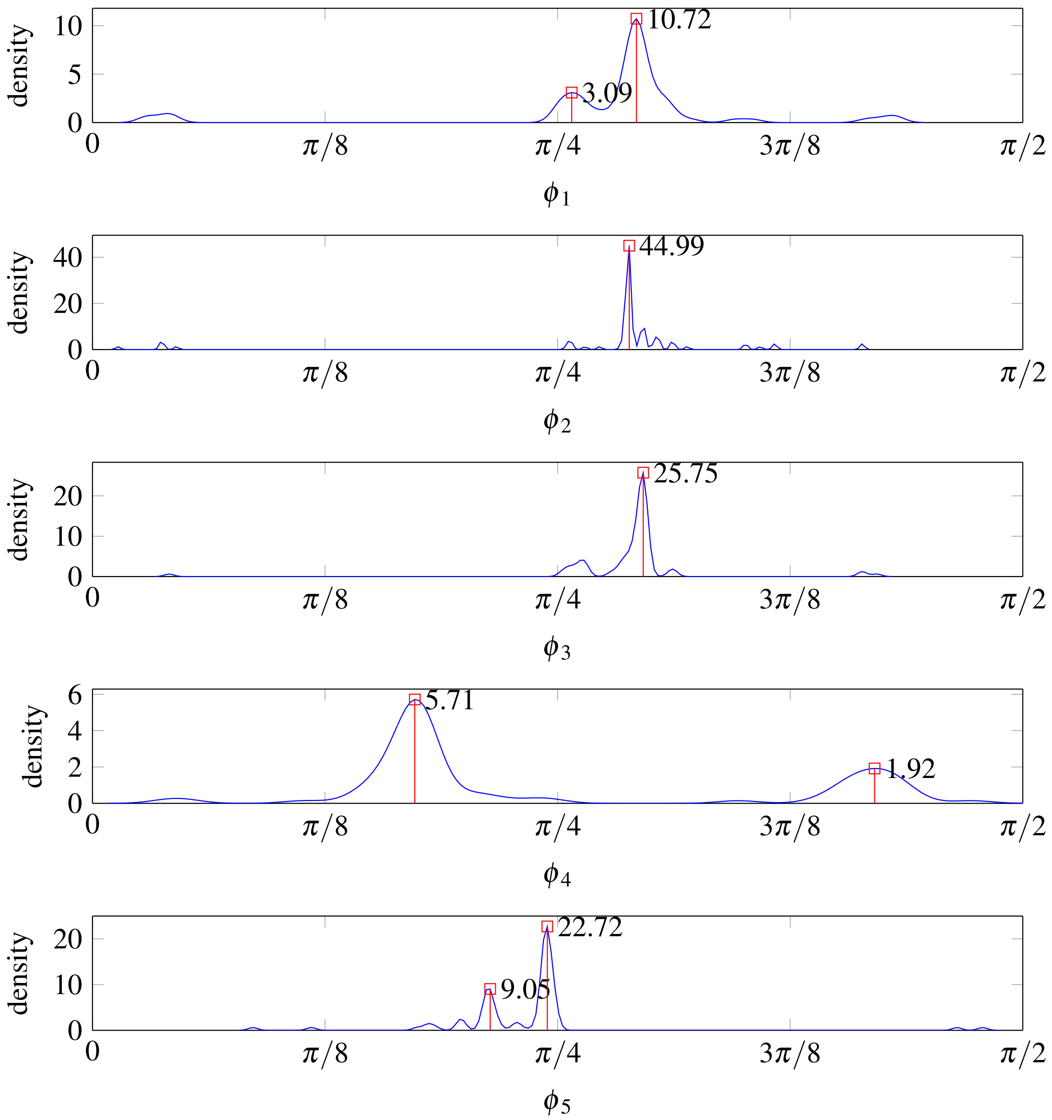


Figure 8.9: Univariate KDE: the point of maximal density is estimated separately for each dimension. As this example featured 6 tracks, there are 5 coordinates in the mix-space.

minimum peak value is set to 1/4 of the maximum value. The peaks are marked and labelled with the function value at that point. This shows that, for this specific trial, the user had strong preferences for certain values of ϕ_2 and ϕ_3 yet, for the remaining values, there are multiple peaks. This suggests the possibility of multiple mixes, although, the relative strength of peaks suggests that simply picking the maximum values should create the most preferred mix. For ϕ_1 and ϕ_5, the two peaks found are closely located, indicating that switching from one value to the other would create only a subtle change to the mix. The greatest variation exists for ϕ_4 which sets the balance of the close-mic'ed snare drum to the combined kick drum & overhead balance. Switching from one peak value of ϕ_4 to the other would result in a vastly different drum sound. It can be said, that in the mixing of these tracks, the ambience of the drums was the main factor that varies between this user's preferred mixes, in this particular example.

Strictly speaking, one does not seek to find the peak by simply concatenating each of the $n-1$ peaks but rather the single peak in the $n-1$-dimensional space. This can be achieved using multivariate KDE. Estimating the density of a multivariate sample is a challenging task and has only recently reached a level of maturity on par with univariate density estimation. In this implementation, the Maggot toolbox (v 3.5) was used [222, 223]. Of course, for the purposes of visualisation, the univariate method was favoured within this chapter.

8.2 Example of a human-guided genetic mixing session

This section describes a single mixing session using the developed IGA-based system. The settings used are shown in Table 8.1.

Table 8.1: Settings used in the following example of IGA mixer

Parameter	Description	Value
N_{tracks}	Number of audio tracks being mixed	6
N_{vars}	Number of variables/dimensions in solution space	$N_{\text{tracks}} - 1$
Population size	Number of candidate solutions per generation	100
N_{clusters}	Number of solutions to be auditioned/evaluated in each generation	5
N_{bits}	Number of bits used to represent the value of each variable	7
Elite fraction	Proportion of children generated as clones of fittest parents	0.05
Crossover fraction	Proportion of children generated by crossover of two parents	0.85
Mutation fraction	Amount of bits to be mutated in the remaining children	$\lfloor (N_{\text{bitss}} \times N_{\text{vars}})/3 \rfloor$
Stop condition	Condition which, when met, causes evolution to cease	10 generations

The initial population was created by uniform selection of points on the hypersphere and the distribution of gains is shown in Fig. 8.10, indicating a fair selection of random points. After conversion to hyperspherical coordinates, the initial population is displayed in Fig. 8.11. After 10 generations of evolution the final population is shown in Fig. 8.12. At this stage it is apparent that there is a region where many solutions lie.

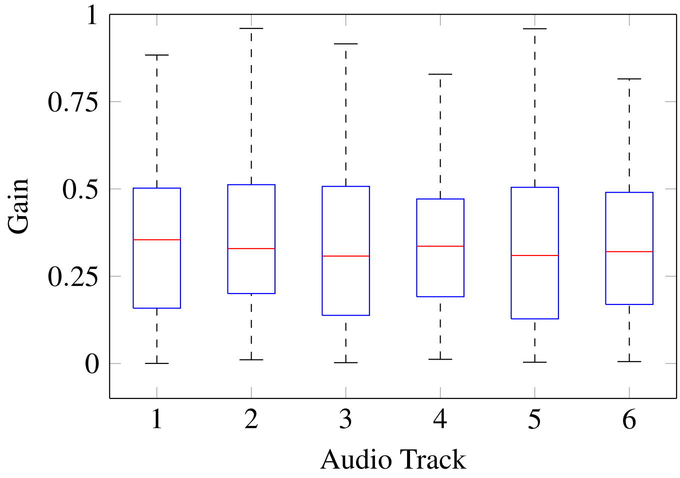


Figure 8.10: Gain values of initial population

Figure 8.13 shows the distribution of *raw* fitness scores at each generation. There were some negative fitness values, which were the result of an individual receiving fitness penalties which, when

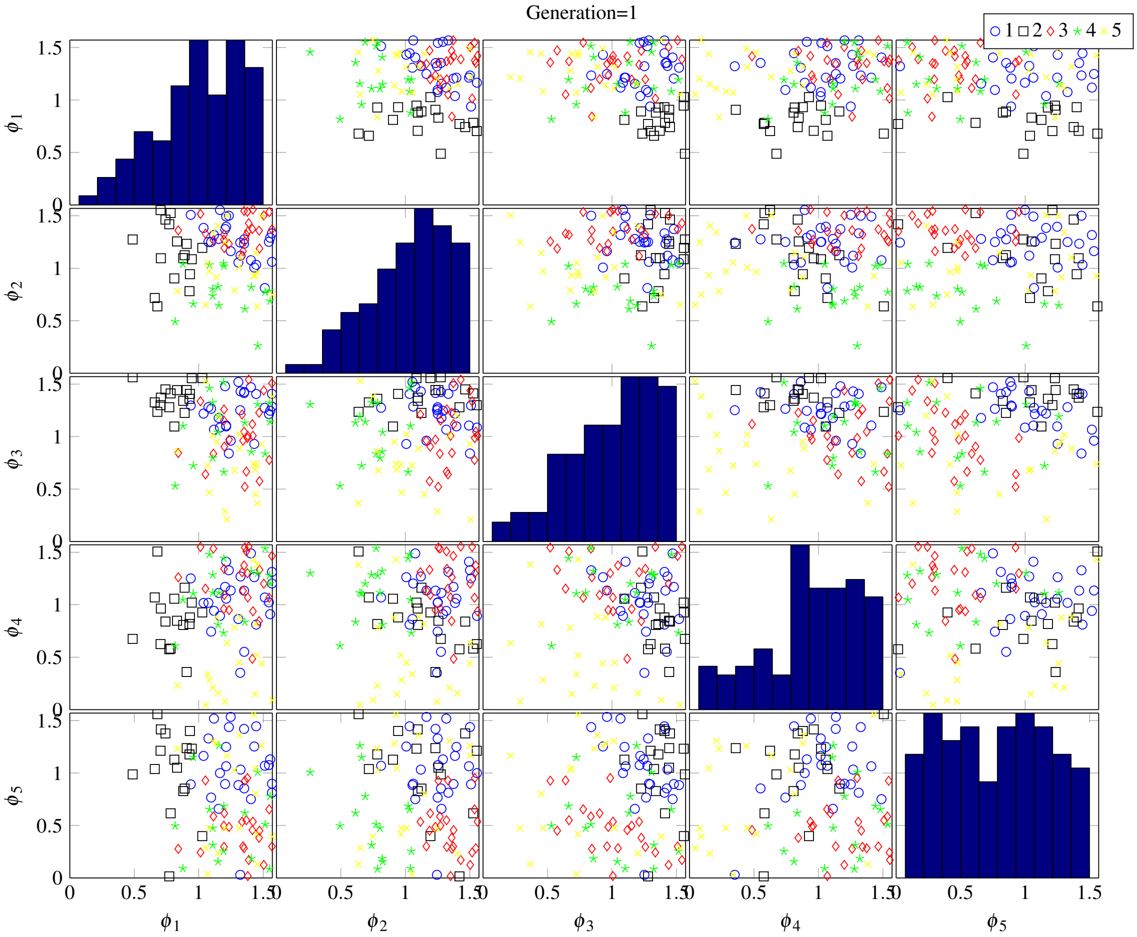


Figure 8.11: Population at generation 1

summed, had a greater magnitude than the fitness awarded to their cluster representative. Figure 8.13 illustrates the increase in fitness values as the evolution progressed. This is an indication that the user perceived the quality of mixes to increase during the course of evolution, as per the desired nature of the system. The median fitness appears to reach a plateau after the seventh generation.

Note that the peak fitness was achieved at generation #4 but that this value is not maintained by the system. While the peak value is being passed on to generation #5, as one of the elite children, the fitness value of this mix is being overwritten. Since only the cluster centroids are evaluated and other members of that cluster are awarded reductions in fitness, it is clear that the peak value in generation #4 was a cluster centroid (as are the peak values in all generations). The fitness value assigned to this elite individual in generation #5 is not necessarily the same as the value it was awarded in generation #4 as it is most likely no longer a cluster centroid. This behaviour suggested modifications to the algorithm were necessary in order to pass on the fitness ratings of the elite parents to the (identical) elite children. This is, of course, only a small correction to implement.

Both univariate and multivariate KDE methods were employed. The result of the univariate method is displayed in Fig. 8.14 and the comparison of the two methods is shown in Fig. 8.15. From this it is clear that the two methods show a high level of agreement, in this specific example.

It can be shown that the ordering of tracks does affect the outcome of the mixing session.

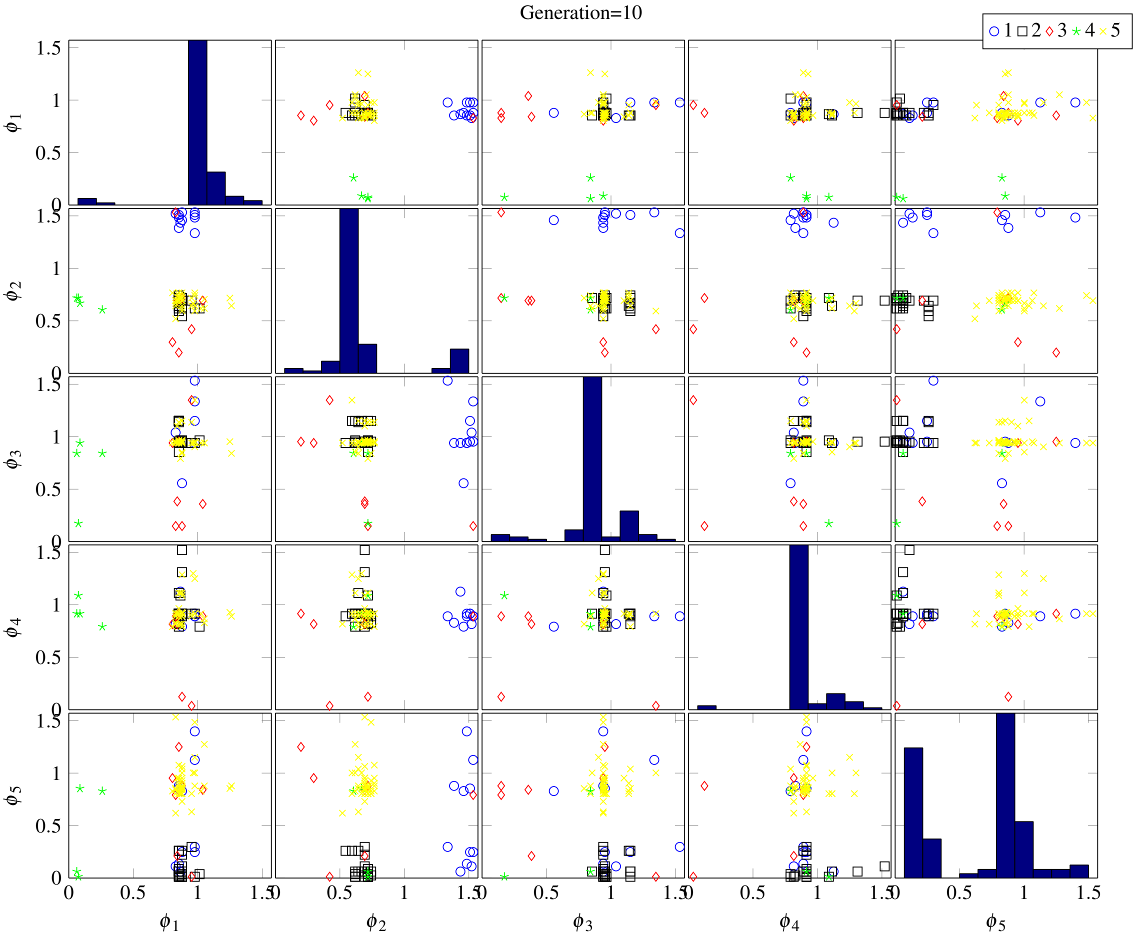


Figure 8.12: Population at generation 10

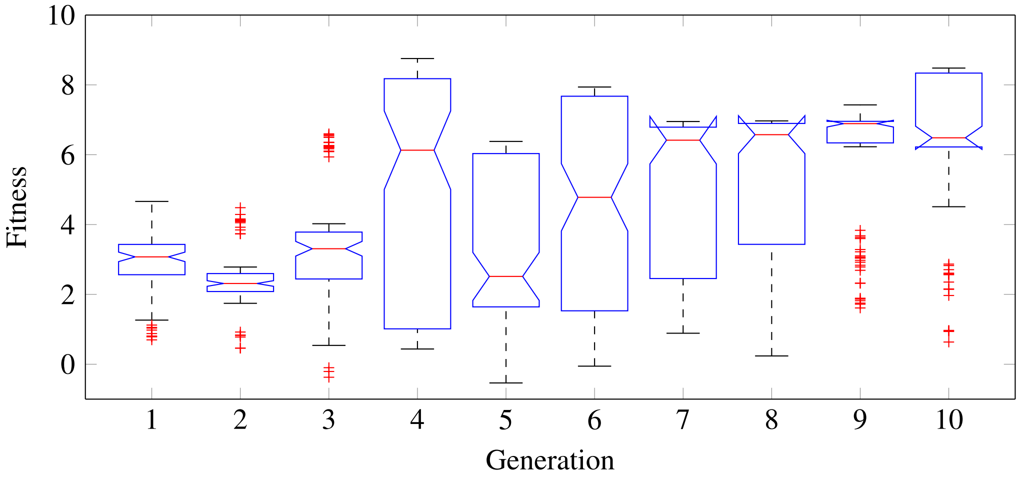


Figure 8.13: Fitness distribution at each generation

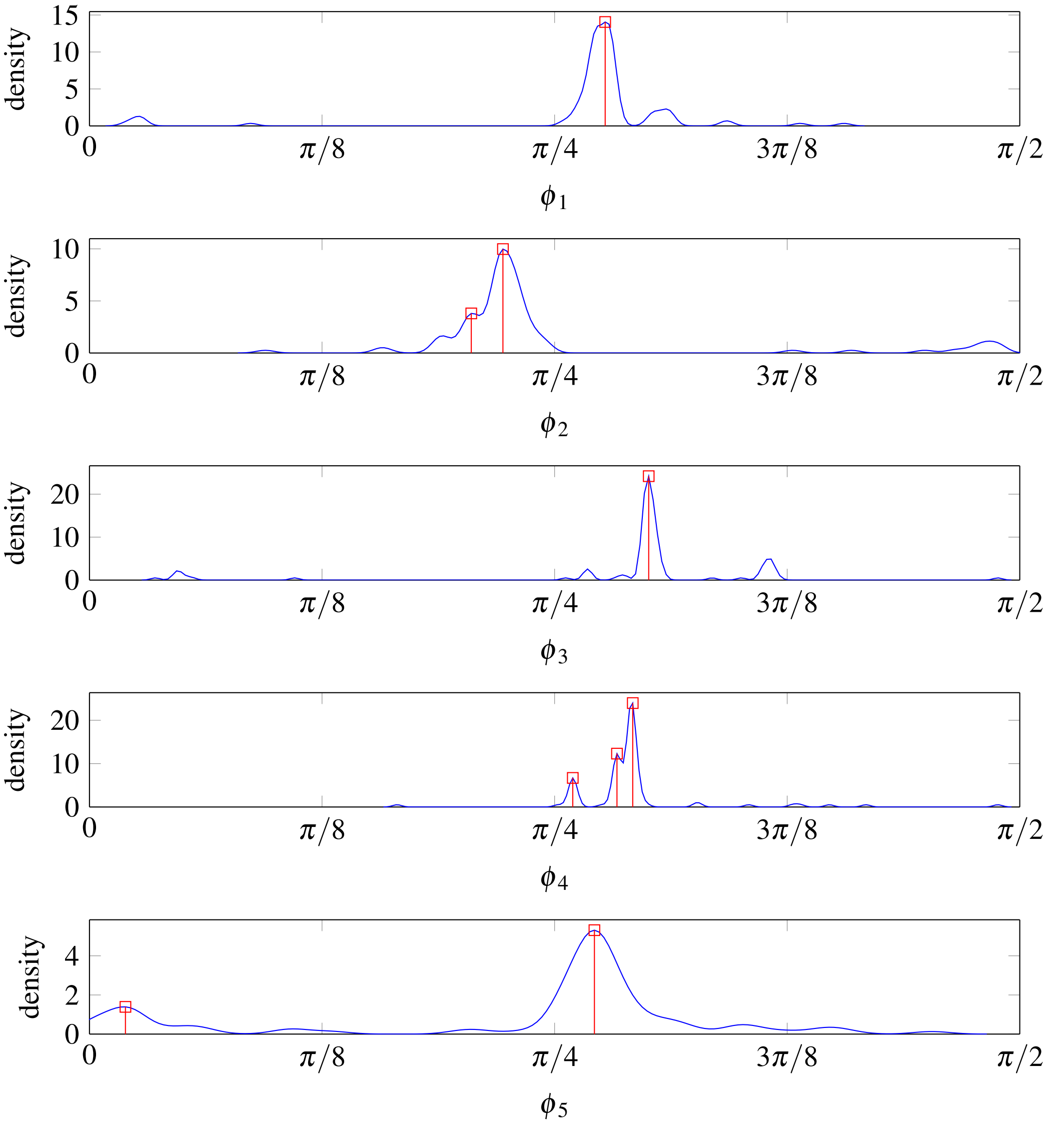


Figure 8.14: Univariate KDE result

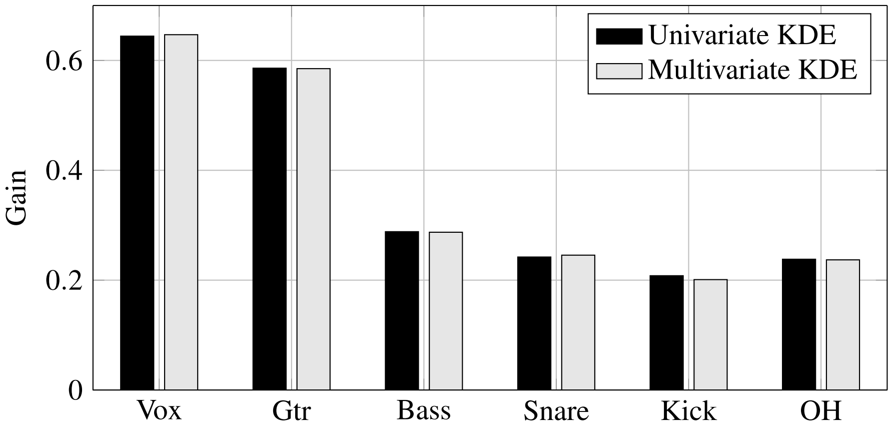


Figure 8.15: Comparison of mixes produced by each KDE method. The differences between the two are deemed imperceptible, ranging from 0.01 to 0.3 dB.

Consider that each variable is denoted by the same number of bits, 7 in this case. For each single variable, there is the familiar concept of a least significant bit (LSB) and a most significant bit (MSB) — that a change in state of the MSB has a much greater effect on the value of the variable than any other bit, and an change in the LSB has very little effect. Now, we also have a most different levels of significance for each variable due to the formulation of the mix-space, as ϕ_n is a function of all ϕ_i when $i > n$. In other words, changing the value of ϕ_1 changes the balance between track 1 and the mix of all other tracks (see Fig. 8.2). This is the most significant variable (MSV). As ϕ_{n-1} is the balance between tracks $n-1$ and n, the effect on the total mix of changing this variable is less than other variables. This is then the least significant variable (LSV). It is partly for this reason that the tracks are ordered as they are, with vocals as track 1, signifying the relative importance of vocals in the mixing process, as identified throughout the previous chapters. This formulation also means, that while there are 2^7 discrete levels for each variable in isolation, there exist various amounts of levels for different instruments: as few as 2^7 for track 1 (vocals) and many more for track $n-1$ and n (kick drum and drum overhead in this example). However, even 2^7 levels for vocal gain is sufficient to allow the gain to be finely adjusted in a mix.

This issue is partially a consequence of what has been referred to as the "Hamming cliff" problem, as the Hamming distance between binary-encoded values of adjacent numbers can be large. One possible solution is the use of Gray-encoded binary values (more specifically a binary-reflected Gray code). This method has been shown to improve performance in a number of studies [224, 225] but these improvements are not guaranteed [226]: it is still necessary to tune the genetic operators and parameters to the problem-at-hand. While Gray encoding can solve the issues associated with MSB/LSB, the MSV/LSV issue remains.

8.3 Example of IGA system used for panning

The IGA mixer was easily adapted to optimise pan positions instead of monaural track gains. In this case the gain vector was fixed such that all tracks had equal perceived loudness. It is the pan position P that was then optimised and then used to get g_L and g_R, using Eqn. 2.1c.

This system was trialled by the author. The aim was to create a mix wherein the vocal and guitar were panned as far apart as possible (direction not important) and all other tracks were panned centrally. From Fig. 4.36, it is clear that, on a unit circle, the maximum symmetrical separation would be (0.707, -0.707).

To modify the IGA mixer to the task of panning, the range of Φ was changed from $[0, \pi/2]$ to $[0, \pi]$. All other GA parameters were the same. The Matlab implementation was identical, although the fitness function had to be changed to create and output stereo mixes based on the panning variables.

Figure 8.16 shows the distribution of the initial population, which is concentrated on the central pan position (where $g_L = g_R$). As before, mixes were rated in terms of solution quality, from 1 to 10. A mix of 10 would be one where the objective is well satisfied, i.e. vocals and guitar are panned far apart but other tracks panned centrally. Figure 8.17 shows that fitness increased notably over the ten generations. The optimal solution is depicted in Fig. 8.18. A perfect result would be $\phi_1 = \pi/2$ and $\phi_2 = 0$. The precise pan positions of each track are shown in Fig. 8.19 to be -0.57 for vocals and 0.81 for guitar. In the same way that the gain optimisation favours solo vox, the panning system favours hard panned vox and central others. This is the most-significant-variable effect, as discussed in § 8.2.

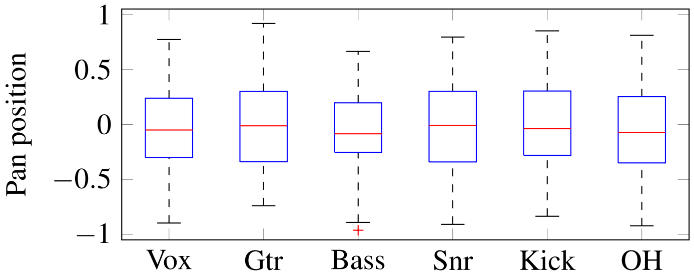


Figure 8.16: Pan positions of initial population

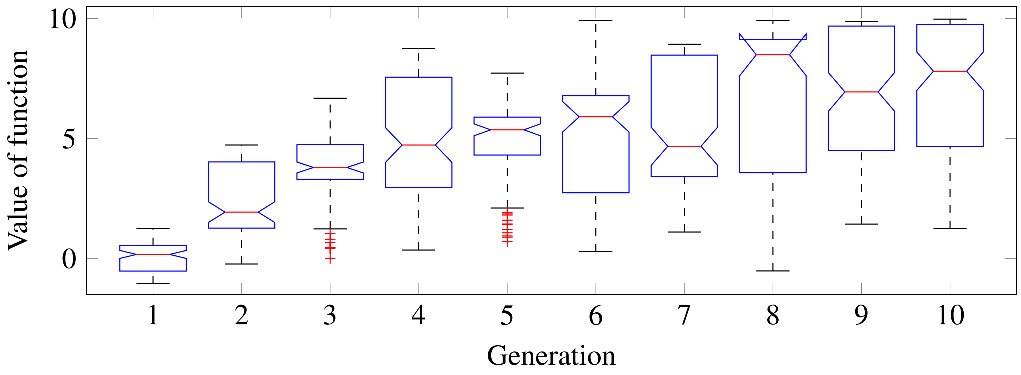


Figure 8.17: Fitness distribution at each generation. The fitness generally improves over time, as desired

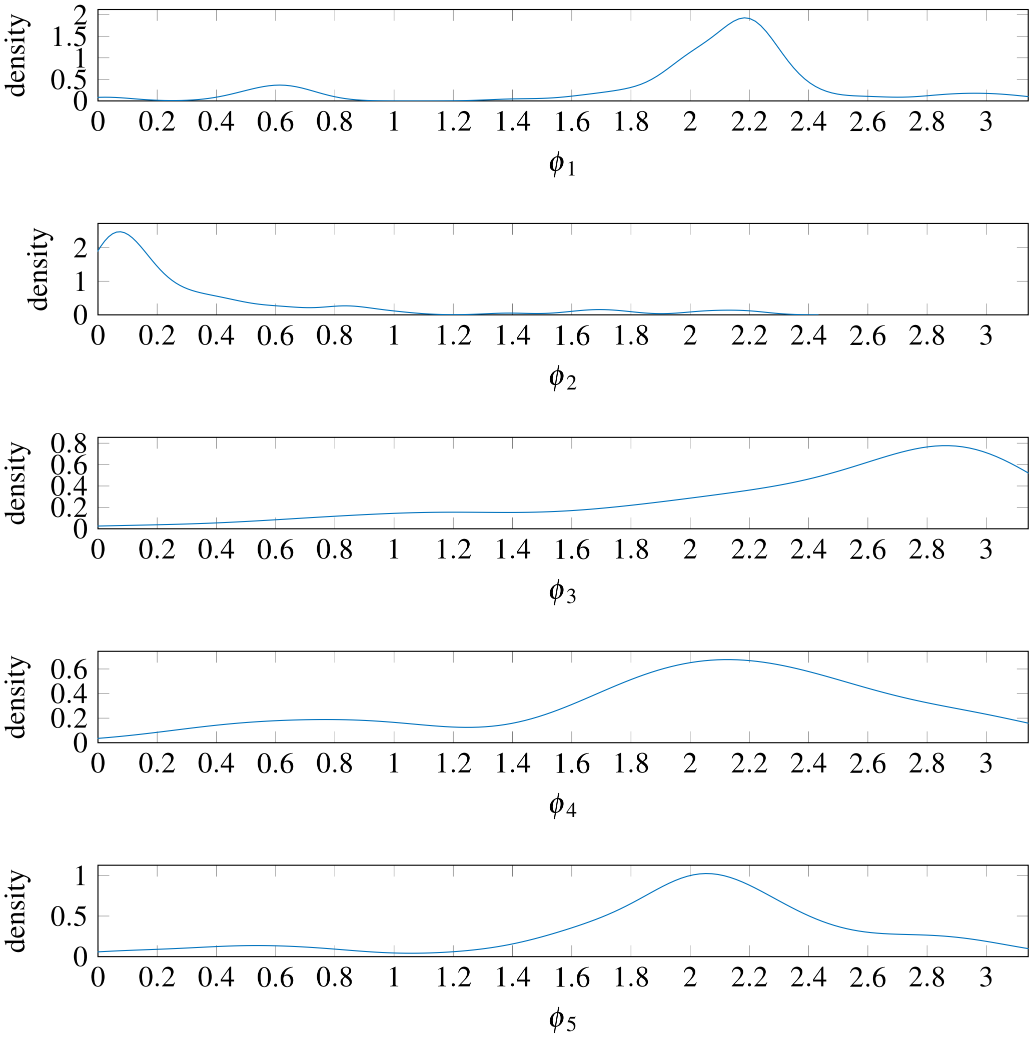


Figure 8.18: Univariate KDE result. To achieve the desired result of vox and guitar panned far apart would require the following result — ϕ_1 would be $\pi/2$ and ϕ_2 would be 0. Other values would have no impact.

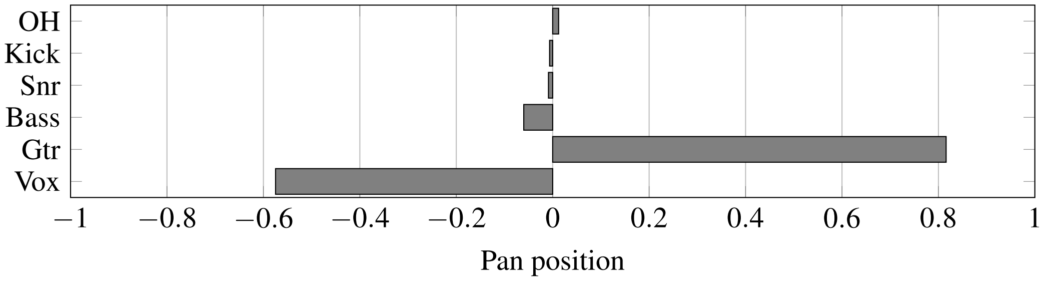


Figure 8.19: Bar graph showing pan positions of the optimal mix, after 10 generations, using the maximum points from Fig. 8.18.

8.4 Improved fitness estimation

This section describes improvements that can be made to the system but were not included in time for the evaluation in Chapter 9.

8.4.1 Inferring fitness based on past populations

As noted in Fig. 8.13, the maximum fitness is being lost in subsequent generations. One possible solution is to re-use the fitness of previous generations in the estimation of fitness of the current generation. Currently, the fitness of an individual is estimated using its distance to the nearest of the rated points. Of course, it may be closer to a previously rated point. The fitness of an individual can then be represented as a weighted average of the fitness it would have been granted as a member of each previous generation. This prevents previous rated points from being "forgotten" in the process of evolution. The fitness of an individual i at generation G can be given by the

$$\text{fitness}_{i,G} = \frac{\sum_{g=1}^{G} w_g \text{fitness}_{i,g}}{\sum_{g=1}^{G} w_g} \tag{8.4}$$

Here, $\text{fitness}_{i,g}$ is the fitness the individual i would have recieved in generation g, i.e. of the cluster centroids in generation g, the fitness of the closest minus the distance to it. Weights can be normalised, making the denominator in Eqn. 8.4 equal to 1. Weights can be equal for all generations, or could be greater for more recent generations.

Additionally, in this example, the number of explicitly rated solutions increases by five per generation. This suggests that more accurate fitness estimation should be achieved over time. After each generation, the rated subpopulation could be used to estimate a fitness landscape, by fitting a simple surface, either by interpolation or polynomial fitting. For the remaining population, their fitness could be estimated from the value of this fitted function. By the end of generation #10, 50 rated solutions exist. A surface could be fitted to these solutions, producing an estimated fitness landscape. An example is shown in Fig. 8.20. Of course, the interpolation should be done over all dimensions: only two are shown here. The maximum point on this surface could be chosen as the optimal mix.

In Fig. 8.20 it is clear that mixes where ϕ_1 is too high or low are rated poorly, as these represent mixes where the vocal level either dominates the backing track, or is lost beneath the backing track. Similarly, when $\phi_2 \to 0$ the drums and bass tracks are almost muted and so there are low fitness ratings here. As $\phi_2 \to \pi/2$ the guitar is almost muted, resulting in low fitness. As indicated by the KDE plots in Fig. 8.14, each variable has optimal points, rarely at the extremes.

8.4.2 Using features to help fitness evaluation

Figure 8.21 illustrates how a set of audio signal features can be used as an additional means of inferring the fitness of the population. Consider a feature, X, with a known (or assumed) probability distribution as shown. In this example it is a standard normal distribution but realistic distributions are shown in § 6.1.5. For each mix the value of the feature is measured and located on the curve. The distance from the mean value is indicated by δ. From § 6.1.5, the assumption is made that the better mixes are found close to the mean values. Consequently, the greater the value of δ the lower the fitness of that mix. By adding δ to the already-determined distance from the evaluated mix, D, a combined fitness penalty can be found. This method can be used for a number of audio signal features, yielding $\delta_X, \delta_Y, \ldots$ for features $X, Y \ldots$ etc. These distances can be weighted as desired, using a series of weighting coefficients β, as shown in Eqn. 8.5 for m features.

$$\text{fitness penalty} = \alpha D + \sum_{i=1}^{m} \beta_i \left|\delta_i\right| \tag{8.5}$$

$$\text{fitness} = \text{fitness of representative} - \text{fitness penalty} \tag{8.6}$$

It is possible to completely remove the user-evaluation from the system, and simply use the audio signal features to guide the mixing process. For example, the user can specify properties of the desired mix, such as values of the features. Figures 8.22 and 8.23 shows the result of a purely-objective GA run, in which the fitness function to be minimised was the distance to a target spectral centroid. Of course, there are many mixes which can have the same spectral centroid. In fact, if any of the individual instrument tracks has a spectral centroid close to the target, then this track will feature heavily in some of the optimal solutions found. As such, constraints would need to be imposed on the system, or multiple features could be used, making it a multi-objective genetic algorithm. Since even the measurement of the signal features, for so many mixes, can be time-consuming, the advantages of this approach, over the interactive genetic algorithm, are not clear.

Alternatively, the features can be used to aid evolution in a different way, using a hybrid genetic algorithm, sometimes referred to as a memetic algorithm (MA) [227]. Such an algorithm has a dual-phase evolution strategy, wherein both genes and memes are evolved. Similar to the gene being the basic unit of biological information, a meme is a basic unit of societal information. Take the example of two twins raised in opposite corners of the globe: while they will share a lot of genetic information they will inherit a different set of memes. Unlike genes, which remain constant over the course of a lifetime, memes can change, and allow an individual solution to adapt, learn and better its position in the solution space.

A genetic algorithm is good at exploring a large solution space but has limited success in "zooming-in" to the best solutions, according to Hart et al. [228]. This is where the hybrid approach can help. In the context of an interactive audio mixing system, the genetic part remains the same but the societal/cultural layer of the algorithm could be based on audio signal features. Often, in a hybrid algorithm, a proportion of the population can, after fitness evaluation, undergo a heuristic-based local search. This allows individuals to move to more optimal solutions. For example, after the user has auditioned and evaluated the subpopulation, these individuals could undergo a local search based on the desired values of audio signal features. One potential issue with this approach is that it relies on heuristics, which, as indicated in Chapter 2, are based on fallible domain knowledge. Here, a variety of approaches are proposed, based on the findings in Chapters 6 and 7. While the "genes" are the inter-channel balances between instruments, the "memes" in the population could be any of the following strategies:

bright: mixes should sound "brighter", which can be achieved by higher spectral centroid

warm: mixes should sound "warmer", which can be achieved by lower spectral centroid

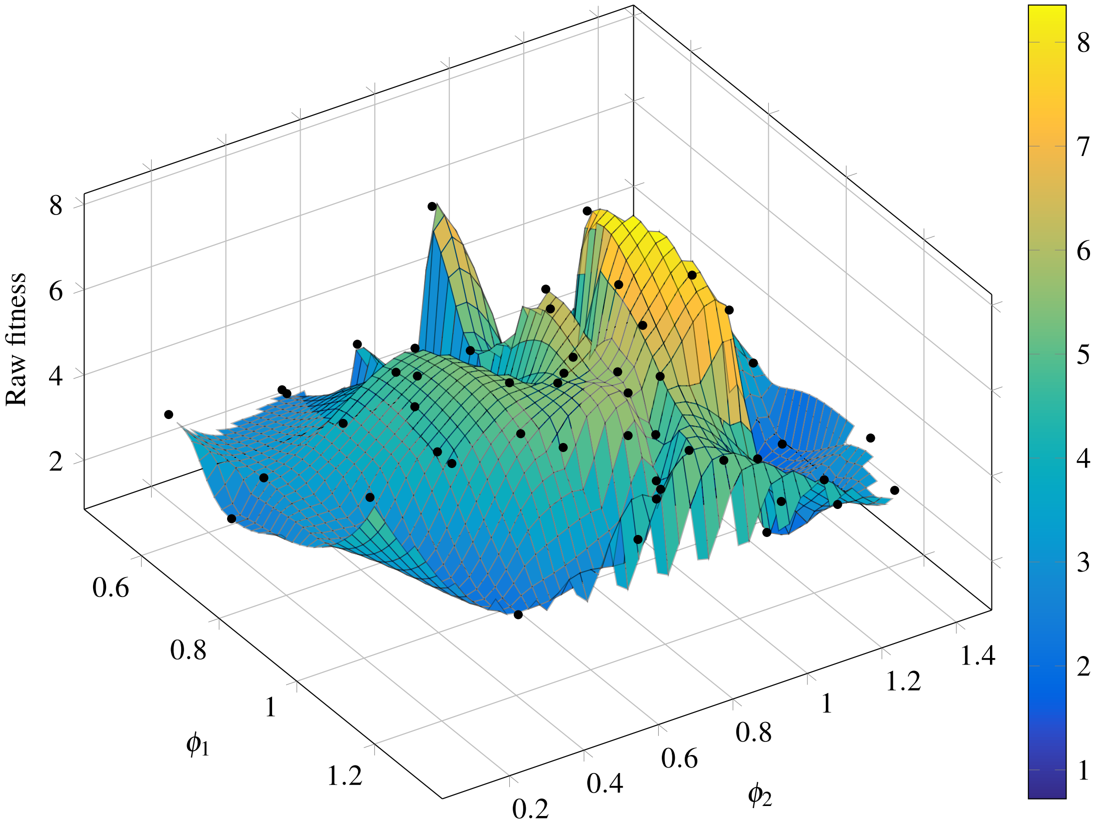


Figure 8.20: Estimated fitness landscape of 50 explicitly rated mixes, obtained using cubic interpolation

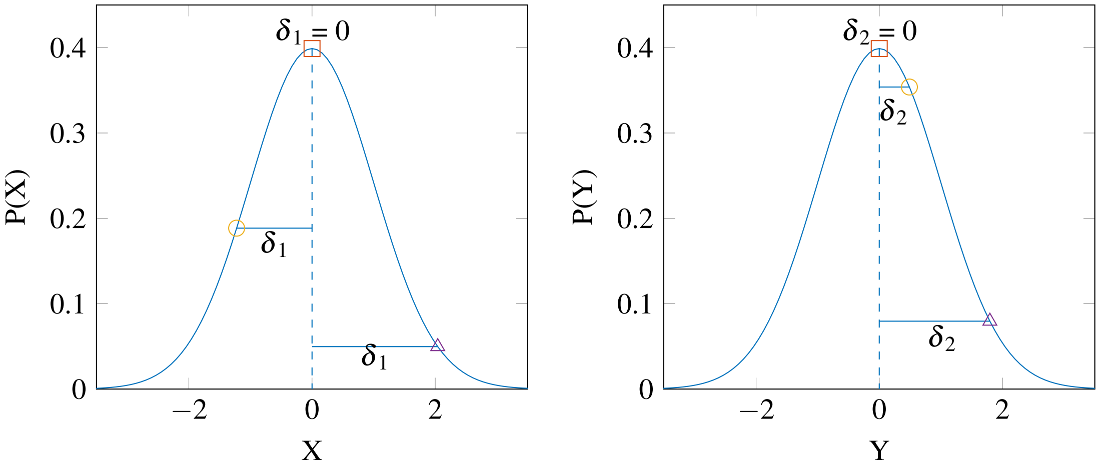


Figure 8.21: Using features for fitness evaluation. Assuming a normal distribution of audio signal features (see Fig. 6.9), the distance δ from the mean μ can be used to help infer the fitness of the population. Here X and Y are two audio signal features and each shape depicts a different mix. The mix indicated by □ has a mean value of both X and Y and is therefore seen as the fittest mix of the three. Similarly, △ is considered the least fit. In the case of a memetic algorithm, alternative points on these curves would be considered optimal, as indicated by the point m. Under this meme, △ is considered the fittest solution, as it is closest to the desired point m, on both curves.

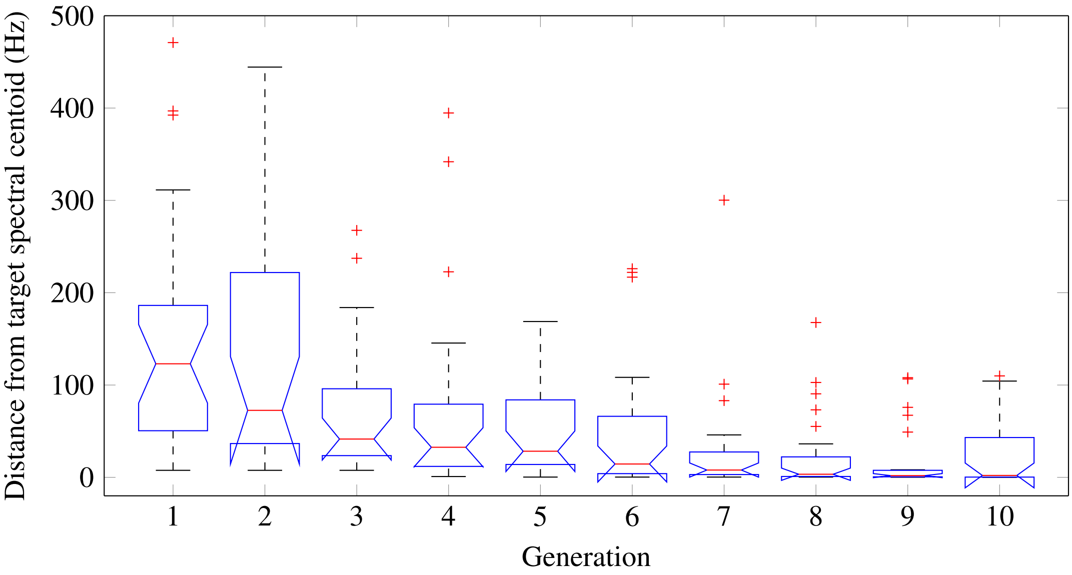


Figure 8.22: Objective GA. Distribution of raw fitness scores at each generation

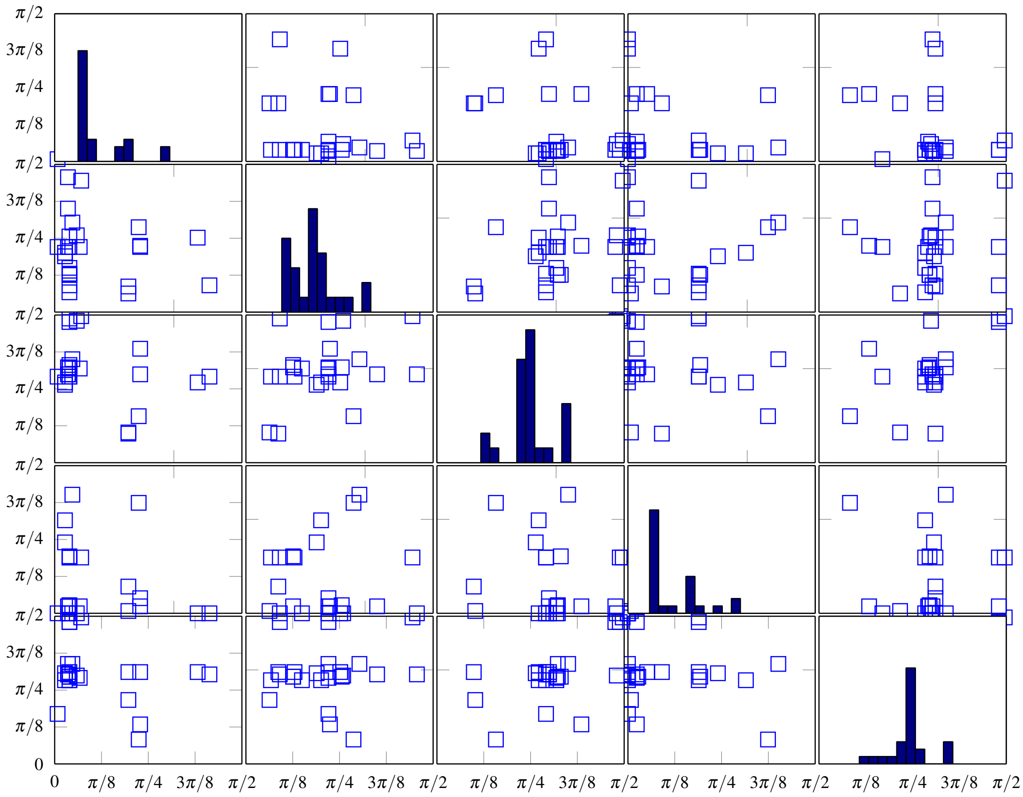


Figure 8.23: Population after 10 generations, using spectral centroid based GA. It is clear that the population has not converged on one optimal solution but that many optimal solutions exist.

wide: mixes are considered better if they exhibit wide stereo impressions, achieved by panning and equalisation, and measured using audio signal features such as the stereo panning spectrogram [188].

punchy: preference for mixes that are punchier (having short periods of significant change in power), as determined by audio signal features [229].

The different symbols in Fig. 8.21 can be understood to represent different memes, i.e. different target values of the signal features. This use of memes within the population allows certain assumptions to be placed into the system initially, such as “brighter mixes are better”, only for the user to validate or reject these assumptions by their fitness ratings. Any specific quality can be introduced as a meme provided that quality can be measured or approximated from the mix. This method shows great potential to be used in an improved version of the mixing system described in this chapter and is left to further work beyond this thesis.

8.5 Chapter summary

In this chapter, a novel mixing system was presented. The system is based on an interactive genetic algorithm, an evolutionary optimisation method which relies on human evaluation. This inclusion of the user at the very core of the algorithm is one aspect which makes this proposed system different to earlier attempts at automated music mixing. Rather than being an expert system, operated by a novice user (a listener with no particular music mixing experience), this system begins with no prior knowledge of music mixing and learns from the user. Therefore, both experienced and inexperienced users should be able to obtain satisfactory performance from the system, while also allowing for it to improve over time. While section § 8.2 demonstrated one isolated instance of the system being used to mix a 6-track session, the output of this instance could be used to inform future use of the system. Over time, the algorithm could adapt to a user in a more general sense, predicting which mixes are likely to be rated highly by that specific user. There is no reason that the system could not learn mixing generally enough to adapt to multiple users: this has been referred to as collaborative evolutionary computation in recent literature [230–232].

With each mixing session, the system has the potential to adapt further. By associating the evolution of the solution with the measured signal features of the input audio tracks, the system could further learn general traits of music mixing. Whether or not this is desired is another issue. In this chapter, the aesthetic proposed is one where the system makes no assumptions of the process. Earlier technologies have perhaps had an over-reliance on prior assumptions and so-called best-practice mixing techniques. Combining both strategies — adapting to a specific user while also learning best-practice from a collection of users — will be a challenge in further development of this and related systems.

Meanwhile, the system as it is proposed in this chapter, requires evaluation from a panel of users. This evaluation forms the basis of Chapter 9.

9 Evaluation of an evolutionary music mixing system

With Chapter 8 having described the design of an interactive music mixing system, the aim of the work in this chapter is to ascertain how users interact with the system and whether or not it can be considered useful. The following are the research questions pertaining to this chapter.

1. What are the median loudness levels of instruments when mixed using this system?
2. How does this compare to a more traditional, fader-based approach, as in Chapter 4?
3. How is the user experience evaluated, qualitatively, by the user?
4. How well does the optimal mix of one song translate to other songs?
5. How do users rate their own mixes?

The first two questions relates to the results found in Chapter 4. What median levels are found for this new system and how do they compare to a more traditional mixing interface? Should they both yield similar levels and distributions of track gain then it could be said that the new system does not prohibit the user from finding the type of mix they would create with a traditional system. This was a desired outcome of the experiment.

In addition to finding the types of mixes that are created with the system, it is important to determine the nature of the user-experience. The third question seeks to identify if a user is likely to encounter difficulty with using the system, and establish the difficulty with which one creates their desired mix.

The fourth question relates to the ability of the system to generalise to other songs, which would be desired. In order for the system to learn the style of the user, and be useful over a number of mixing sessions, an optimal mix for one song should be, at the very least, a good first guess for other songs. The effectiveness of this approach may well depend on how similar the style of music is, the instrumentation, and other factors.

The final question relates to the psychoacoustics of the mix-engineer, specifically their impression of their own mixes. Previous work by De Man et al. [82]suggested that a mix engineer, in later subjective evaluation of their mixes and the mixes of their peers, has a preference for their own mixes, even when presented blindly. Possible explanations for this effect are that they explicity recognised a mix they had created or that they implicitly recognised their style of mix, thinking "I like this mix — it sounds like what I would do", not realising that it *was*.

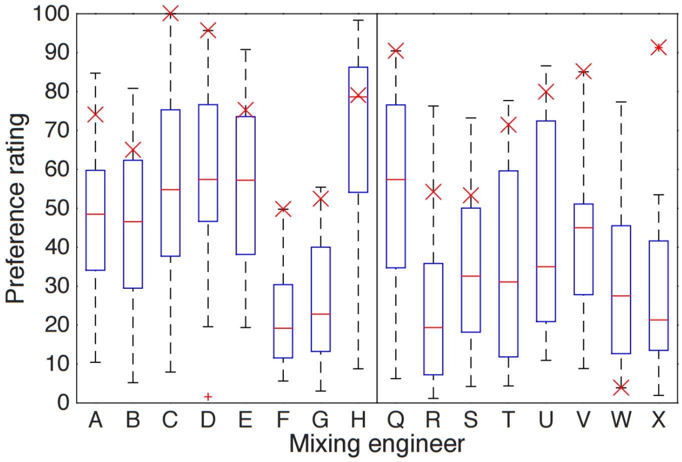


Figure 9.1: Box plot of ratings per mixing engineer including their own assessment (red 'X') of one song, reproduced from De Man et al. [82].

Herein, this has been investigated in a more indirect way. Since the output of the IGA mixer is the gain vector that was applied to loudness-normalised tracks, this vector can be applied to another song in that same form (same number of tracks, in the same order and loudness-normalised). In this case, the mixes being evaluated later are of unfamiliar songs, with the mixes being created in the style of the mix-engineers, using their previously made mix as a template.

To answer these questions, two experiments were devised. The first gave a number of participants the chance to use the system to create their desired mix of a specific song, and to report on their experience of the system. The second experiment took this mix and used it as a template: the optimal gain vector generated in the first experiment is used to generate mixes of other songs which were subsequently evaluated in the second experiment.

9.1 IGA-Expt.1 — Gather mixes

This experiment provided participants with the opportunity to trial the system. Each participant was asked to create a mix using the system, in accordance with their own preferences. The experiment took place in October 2016, in the BS.1116 compliant listening room at the University of Salford. The test set-up was comparable to that of experiments in previous chapters (see Fig. 4.11). Only a single loudspeaker (Genelec 8020a) was used, positioned centrally, at a distance of 1.4 metres from the listening position. Participants were free to adjust the playback level during their evaluation of generation #1 but not thereafter.

Table 9.1: Set-up for IGA mixer evaluation

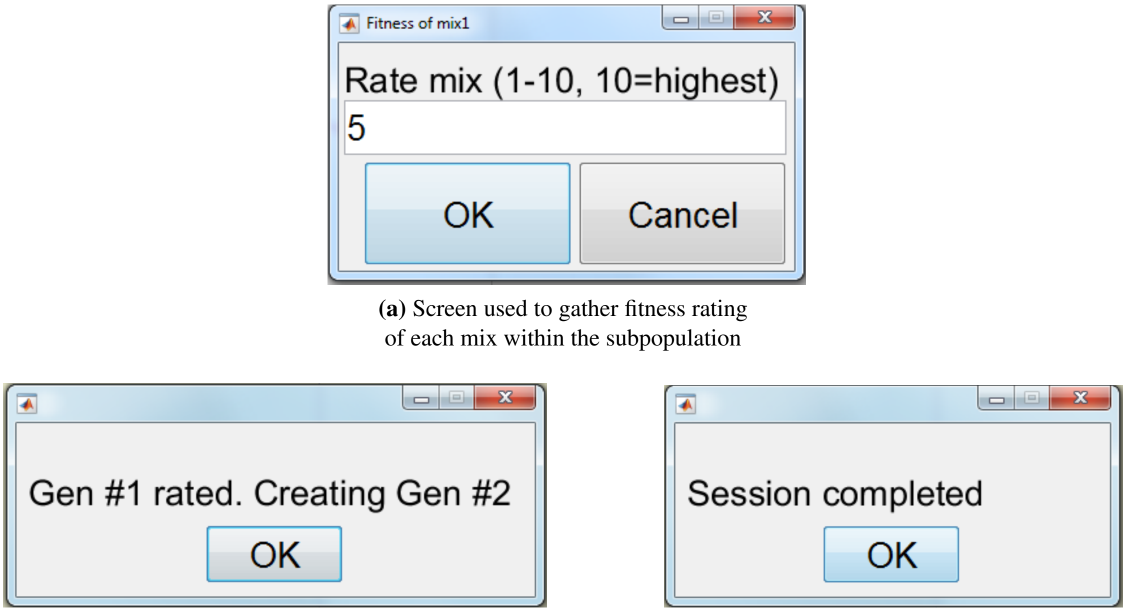
Audio stimuli	Multitrack content with 6 mono tracks (PCM .WAV, 16-bit, 44100 Hz): Vox, Guitar, Bass, Snare, Kick, OH
Song for expt 1	Sister Cities
Songs for expt 2 (see § 4.3 and § 6.1.1)	Burning Bridges, Borrowed Heart, Fighting (We Were), Heartbeats, I'm Alright, New Skin, Revelations, What I Want
Set-up	1 x Genelec 8020a, Focusrite 2i4 interface

The number of participants who took part in this experiment was 14 (13 plus the author), most of whom had previously participated in at least one of the experiments in Chapter 4 and were considered to be sufficiently familiar with the concepts of the task, namely the balancing of a number of audio signals. Furthermore, all were either postgraduate or undergraduate students in audio-based courses.

The task of each participant followed the same structure as the example in § 8.2[1]. None of the graphs were presented to the user, to prevent the introduction of a visual bias or the mixing of the music based on the visual information displayed. Consequently, the user needed to rely solely on audition. These graphs were saved to disk during each run in order to act as a diagnostic tool, and were visible to the experimenter during the session, on a second monitor. The only visual information presented to the user is a simple GUI to gather ratings of mixes (Fig. 9.2a) and to provide a progress update at the end of each generation (Figs. 9.2b and 9.2c). This represented a minimal amount of visual stimulus, however such a system could surely be implemented with no visual stimulus, e.g. using a numeric keypad for data entry. When rating mixes, participants were advised that a rating of 10/10 represented their ideal mix, while a rating of 1/10 is a mix most far from ideal, in any of the many ways that this might be possible.

Upon completing 10 generations the optimal mix was estimated using the univariate KDE method described in § 8.1.8. This mix was then played back to the user for qualitative evaluation but was not rated quantitatively. At this stage, the user was provided with a questionnaire in order to assess the interaction between the user and the system. The first 10 questions were those of the System Usability Scale (SUS), a short survey designed to gather information of a systems usability [233]. Additional questions were devised by the author as more directly related to audio mixing systems and this particular experiment. The list of statements is shown in Table 9.2. For each the

[1] i.e. this algorithm does not include Gray coding or fitness estimation using previous generations or audio features



Fitness of mix1

Rate mix (1-10, 10=highest)

5

(a) Screen used to gather fitness rating of each mix within the subpopulation

Gen #1 rated. Creating Gen #2

Session completed

(b) Screen shown after a generation was rated

(c) Screen after final generation was rated

Figure 9.2: Buttons used within IGA experiment.

user choose a response on a 5-point Likert scale, marked at the extremes by “strongly disagree” and “strongly agree.”

Table 9.2: Survey questions for IGA mixer

number	statement
1	I think that I would like to use this system frequently.
2	I found the system unnecessarily complex.
3	I thought the system was easy to use.
4	I think that I would need the support of a technical person to be able to use this system.
5	I found the various functions in this system were well integrated.
6	I thought there was too much inconsistency in this system.
7	I would imagine that most people would learn to use this system very quickly.
8	I found the system very cumbersome to use.
9	I felt very confident using the system.
10	I needed to learn a lot of things before I could get going with this system.
11	I felt in control of the mixing process.
12	I thought the loudness of samples was consistent.
13	I felt the mixes got better over time.
14	I found the interface to be physically demanding.
15	I thought the loudness of samples was suitable.
16	I found the interface to be mentally demanding.
17	I felt the test environment was comfortable.

9.2 Results from IGA-Expt.1

Over all 14 participants, the median the amount of time taken to evaluate 10 generations (50 mixes) was 11 minutes 17 seconds (see Fig. 9.3). This amounts to a mean of 13.34 seconds per mix (recall that each mix was 30 seconds long and no repeats were possible). As a mix deemed to be poor can be evaluated rather quickly, this short duration was not unexpected.

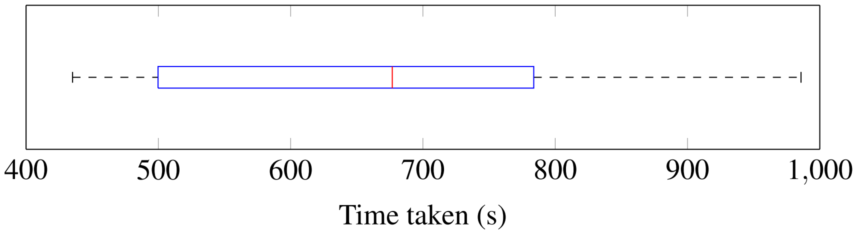


Figure 9.3: Time taken by participants to complete ten generations

Figure 9.4 shows the distribution of raw fitness scores per generation when all participant's data is combined. As desired, the fitness of the population typically increases as the system evolves. A few additional observations can be made from this plot.

1. Decrease at gen 2 — as the initial population is uniformly distributed on the sphere, there is likely to be a variety of mixes, rated good and bad. As mentioned in § 8.1, given a large enough population, the position of the evaluated mixes (closest to the centroids of the clusters) is predictable. Since gen #2 represents the first evolved generation, after a first generation of random mixes, it is credible that the fitness may drop initially.

2. Increase from gen 3→7 — as anticipated, the fitness increases over the duration of the session but mostly between generations 3 and 7. This indicates that once the system has identified an optimum point based on user ratings, after a few generations of searching it slowly begins to converge.

3. No significant change after gen 7 — the aforementioned convergence, however, seems to reach a saturation point at generation 7, as no significant change is observed from here on.

It is important to note that while the best mixes in a given generation are passed on to the next generation (as 'elite' children), they may not survive another generation. As mentioned in § 8.2, this is due to the fact that the inferred fitness is always based on subtracting an offset from the rated subset. The best mix in a given generation is therefore one which was part of the rated subset. It is unlikely that it would be form of the next generations subset, once the clusters are re-calculated on the new population.

Once the system completed 10 generations of user-evaluation and evolution, the univariate KDE method was used to determine that participant's supposed ideal mix (see § 8.1.8). Figure 9.5 shows the distribution of gain levels for each track. As with similar experiments in § 4.3 and § 4.5, vocals are set as the loudest track in the mix. This further justifies the use of a vocal boost in the creation of random mixes in Chapter 5. Vocals were also considered one of the most important elements in the mix as mentioned in § 6.2.4.

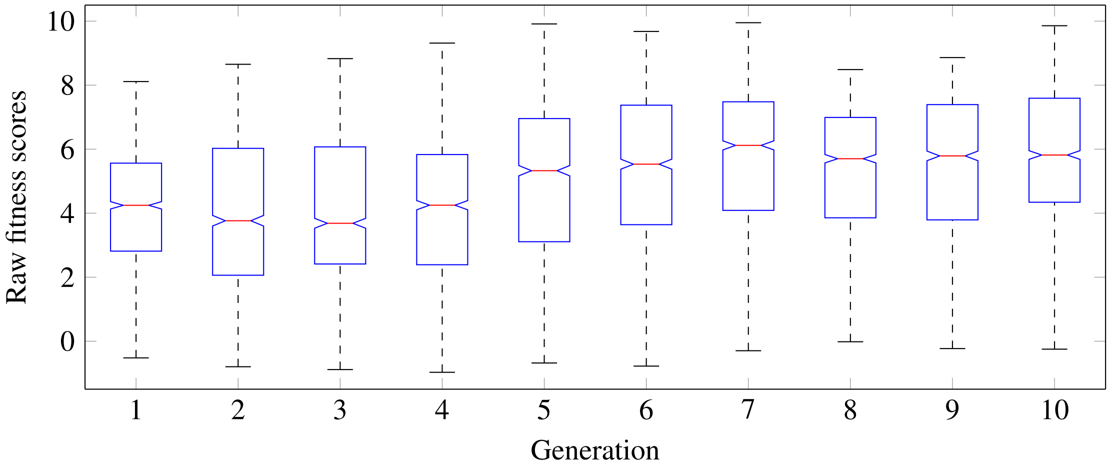


Figure 9.4: Boxplot showing the raw fitness scores per generation, for all 14 participants' sessions (1,400 mixes per generation).

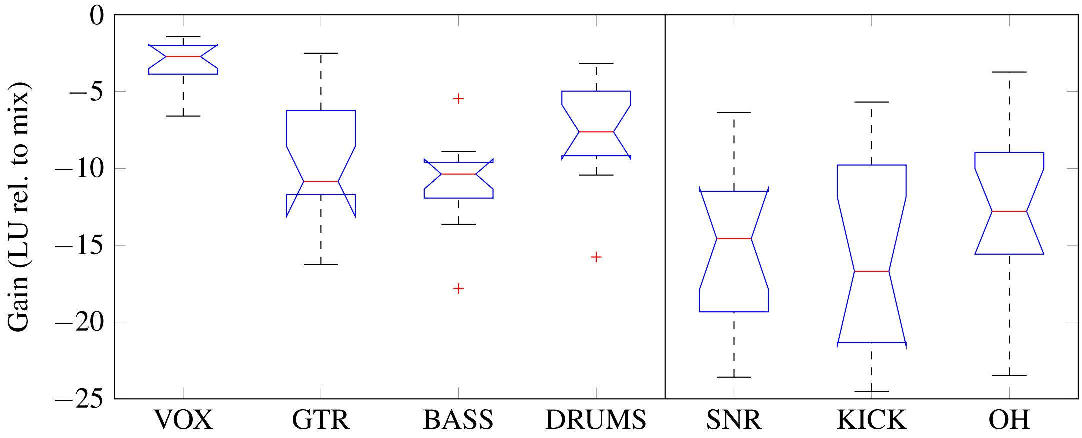


Figure 9.5: Boxplot of gains in final mixes (14 participants)

Table 9.3: Comparision of levels. Fader results are from § 4.3.5, where Faders(all) pertains to the entire experimental data and Faders(LS.sc) is the subset of results for the same conditions as the IGA (using loudspeakers and the song "Sister Cities")

Track	**Median Level (LUFS)**		
	IGA	Faders(LS.sc)	Faders(all)
Vox	-2.72	-2.30	-2.85
Gtr	-10.84	-8.89	-8.56
Bass	-10.37	-10.43	-10.46
Drums	-7.62	-8.33	-8.11
Snare	-14.57		
Kick	-16.69		
OH	-12.79		

9.2.1 Comparison with fader-based experiment

A comparison between median levels in various experiments is shown in Table 9.3. This reveals that there are only small differences between experiments. The largest difference is the fact that the guitar was typically set quieter using the IGA system, by about 2 LU. The level of the vocals in the IGA experiment is closer to the Faders(all) level than Faders(LS.sc), indicating that this level may generalise well to other songs, as is the basis for § 9.3. A precise match between experiments would have been surprising, especially considering the IGA method only approximates the user's ideal mix in the final KDE stage. That said, the close match for vocals, bass and drums (to a slightly lesser extent) indicates the success of the IGA method. From this it may be claimed with some confidence that the IGA method is capable of creating a range of mixes similar to that which would be created using the conventional fader-based approach.

9.2.2 Survey responses

Figure 9.6 shows histograms of the raw scores from the first ten questionnaire items. High scores on odd numbered questions indicate a positive impression of system usability, as do low scores on even-numbered questions. Scoring of the questionnaire results is as follows:

- For odd items: subtract one from the user response.
- For even-numbered items: subtract the user responses from 5
- This scales all values from 0 to 4 (with four being the *most positive* response).
- Sum the converted responses for each user and multiply the total by 2.5. This converts the range of possible values from 0 to 100.

Table 9.4 shows the mean of the converted scores for each item. Note that in Table 9.4, the score shown for items 1 to 10 is the mean positivity (from 0 to 4), not the mean of the raw scores (i.e. not the level of agreement with the statement). For items 11 to 17 the score shown *is* the mean level of agreement with the statement. The boxplot of overall scores for the system (from 0-100) is shown in Fig. 9.8a. The median score is 90 while the range was from 75 to 95. This score by itself does not offer much insight without other systems to compare to. Bangor et al. [234] analysed the SUS scores from a variety of different systems and found the average SUS score from over 200 studies to be 70. This score of 70 can therefore be considered an average score to which new studies can be compared. Figure 9.8b shows a normalised curve from which SUS scores can be interpreted [2].

The statement which received the least positive response was #1 ("I think that I would like to use this system frequently"). Initially, this particular observation seems to contradict the overall high score that users awarded the system. However, while it is the least positive response, the mean score is 2.92 on a scale of 0 to 4, suggesting a result that is still rather positive. However, it is important to realise that the users would have been comparing the system to a more conventional audio mixing system. The next least positive statement was #6 ("I thought there was too much inconsistency in this system"). This indicates that difficulties experienced by users were due to lack of direct, explicit control over the parameters of the mix, as also indicated by statement #11 ("I felt in control of the mixing process"). When asked whether the system was either physically

[2] `http://www.measuringu.com/sus.php`

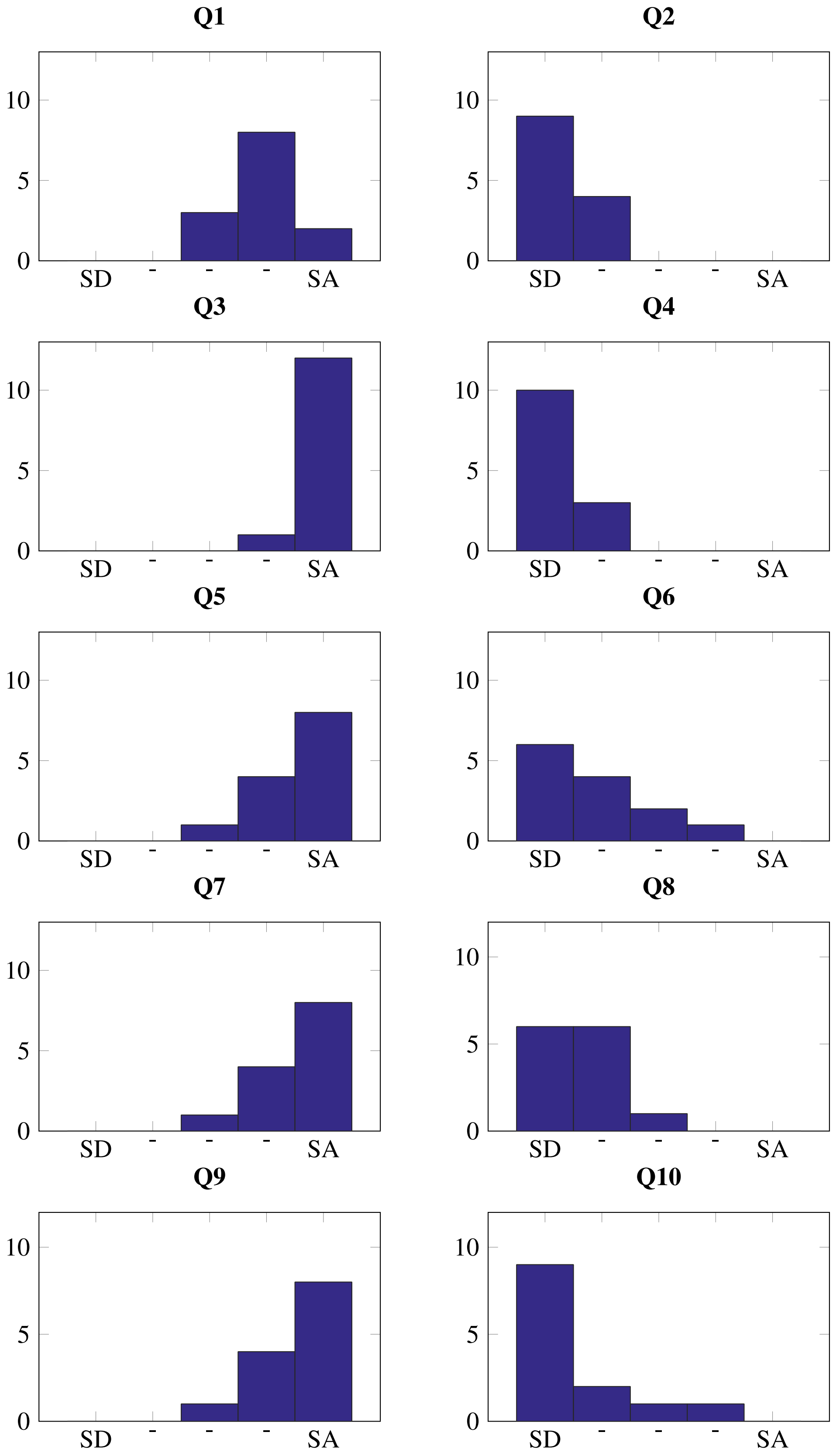
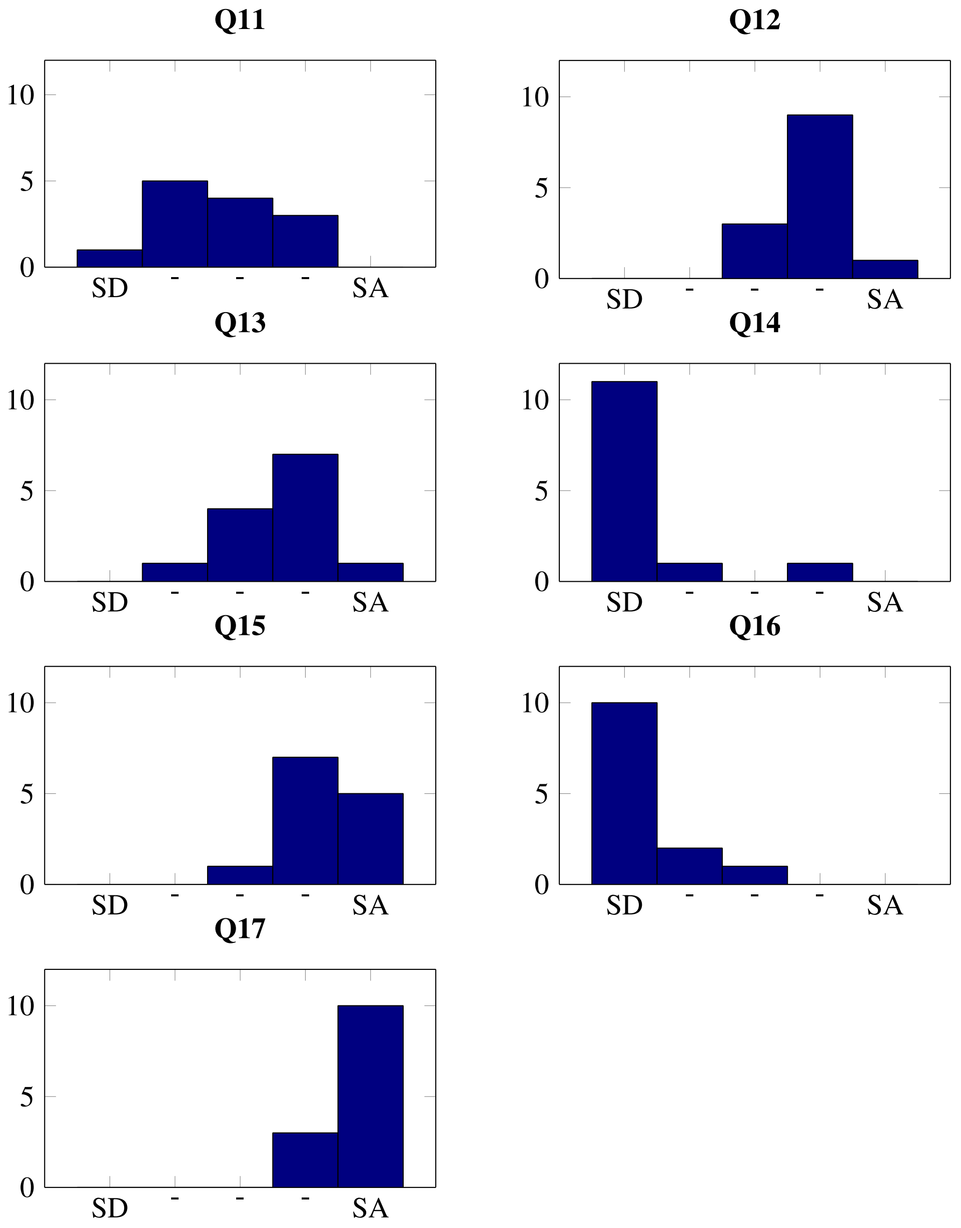


Figure 9.6: Histograms of raw responses to survey questions 1 to 10. SD and SA are "strongly disagree" and "strongly agree" respectively.


Q11

Q12

Q13

Q14

Q15

Q16

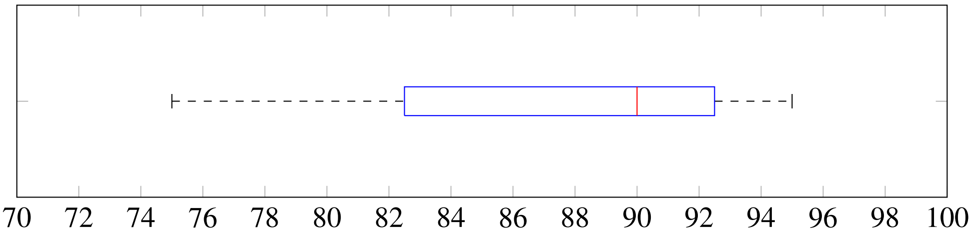
Q17

Figure 9.7: Histograms of raw responses to survey questions 11 to 17. SD and SA are “strongly disagree” and “strongly agree” respectively.

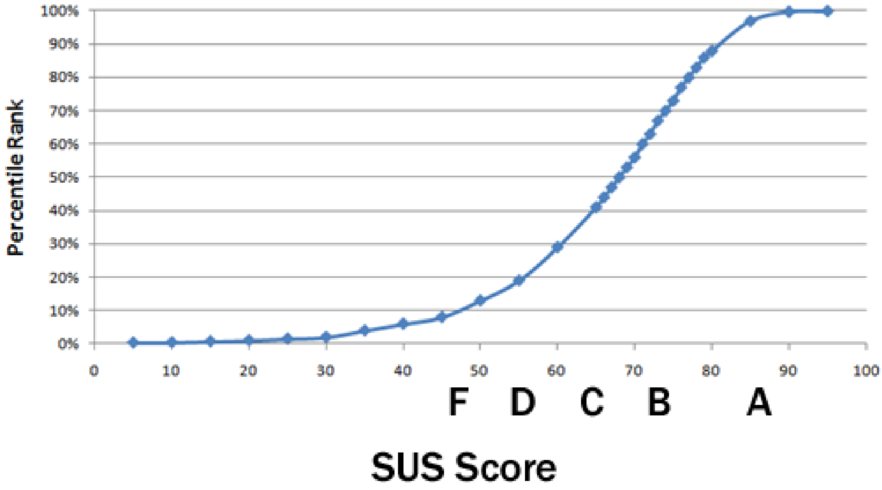
or mentally demanding, users typically responded that neither was the case. This indicates that the system has a low level of user-burden. The system also achieves its goal of not being a physical burden, suggesting a high level of accessibility. From the SUS items, the statement obtaining the most positive response was #3 ("I thought the system was easy to use"). Importantly, users generally felt that mixes got better over time, as desired. Overall, the impression of the system was positive, considering the results shown in Fig. 9.8.

Table 9.4: Survey results for IGA mixer. This table summarises the results shown in Figs. 9.6 and 9.7 by showing the mean and standard deviation of the data.

number	statement	mean positivity	std.dev
1	I think that I would like to use this system frequently.	2.92	0.64
2	I found the system unnecessarily complex.	3.69	0.48
3	I thought the system was easy to use.	3.92	0.28
4	I think that I would need the support of a technical person to be able to use this system.	3.77	0.44
5	I found the various functions in this system were well integrated.	3.54	0.66
6	I thought there was too much inconsistency in this system.	3.15	0.99
7	I would imagine that most people would learn to use this system very quickly.	3.54	0.66
8	I found the system very cumbersome to use.	3.38	0.65
9	I felt very confident using the system.	3.54	0.66
10	I needed to learn a lot of things before I could get going with this system.	3.46	0.97
		mean score	
11	I felt in control of the mixing process.	2.69	0.95
12	I thought the loudness of samples was consistent.	3.85	0.55
13	I felt the mixes got better over time.	3.62	0.77
14	I found the interface to be physically demanding.	1.31	0.85
15	I thought the loudness of samples was suitable.	4.31	0.63
16	I found the interface to be mentally demanding.	1.31	0.63
17	I felt the test environment was comfortable.	4.77	0.44



(a) Boxplot of SUS scores. The median score is 90, with the range being 75 to 95. This result indicates the system is highly usable.



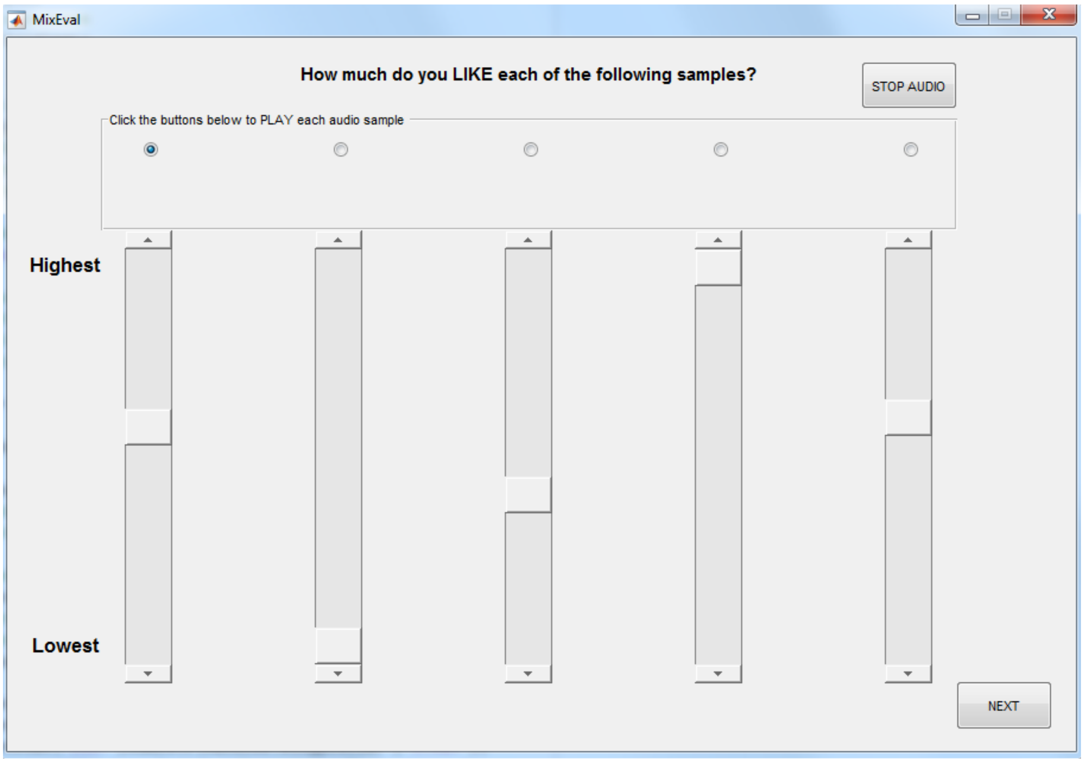
(b) SUS curve. Based on this curve, a score of 90 suggests the system is highly usable.

Figure 9.8: Overall usability score of the system, based on SUS questionnaire

9.3 IGA-Expt.2 — Subjective evaluation of peer mixes

After the first 12 participants completed experiment 1, their final mixes were used as a template from which mixes of eight other songs were created (the eight songs listed in Table 9.1). All 96 of these mixes were evaluated by the author and the mixes of five users were chosen for use in experiment 2. In order to decide which five were to be used, firstly a number of participants mixes were excluded due to particularly noticeable, or song-specific, mix decisions (such as any one instrument being especially low in the mix). Any participants who had previously notified of their unavailability for experiment 2 were also excluded. Ultimately the five users whose mixes were chosen were accepted as those mixes were considered to sound credible over all eight new songs (they did not produce noticeable undesired effects such as near-muted instruments), as well as sounding different enough from one another. These five participants are herein referred to as MixerA to MixerE.

The experiment also took place in October 2016, in the BS.1116 compliant listening room at the University of Salford, and overlapped with experiment 1, using an identical set-up. The playback level was set to 79dB(A). There was no need to explicitly normalise the perceived loudness of these audio stimuli as, being points in the mix-space, each mix was generated at the same loudness (see Fig. 5.8).



MixEval

How much do you LIKE each of the following samples?

Click the buttons below to PLAY each audio sample

● ○ ○ ○ ○

Figure 9.9: GUI used for evaluation of IGA mixes.

This experiment utilised a multi-stimulus audio evaluation. Each screen, as shown in Fig. 9.9 represents one song and displays all five mixes, one in the style of each mix engineer. These mixes are assigned to sliders randomly. Sliders range from 0 to 1 and the initial slider location is set to 0.5. Clicking the 'NEXT' button advances to the next song and songs are presented in a random order. The 'NEXT' button is only made visible once four conditions have been met.

1. All samples must have been played

2. All faders must have been moved

3. At least one fader must be set to the maximum value

4. At least one fader must be set to the minimal value

Only six participants took part in this experiment. While deliberate efforts were made to have each of the participants from experiment 1 return for experiment 2, of the five participants whose mixes were chosen only four completed experiment 2. Consequently, the aim of having each participant evaluated *their own* mixes (mixes of other songs made from their experiment 1 result) was not met entirely.

9.4 Results from IGA-Expt.2

Figure 9.10 displays the distribution of preference ratings awarded to the mixes by each mixer. Recall that each participant evaluated five mixes each for eight different songs. From this data, the mixes by MixerD were significantly preferred over those of MixerA, MixerB and MixerC. A number of other significant differences also exist between mixers. Unfortunately, the one set of mixes that was deemed to be most preferred (MixerD) was also the one participant not to return from experiment 1. Consequently there is no knowledge of MixerD's rating of their own mixes in Fig. 9.10.

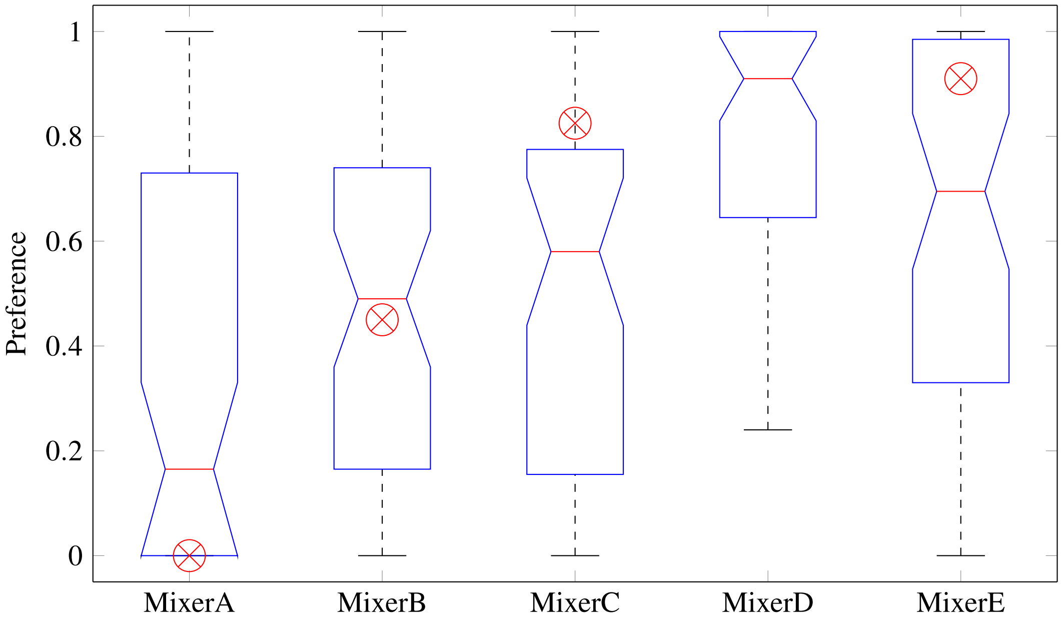


Figure 9.10: Boxplot of preference ratings for each mixer. The red mark indicates the participants median rating of the mixes made with their own mix setting from expt1. MixerD was not available to take part.

The initial hypothesis, formed from the result shown by De Man et al. [82], was that mix engineers prefer their own mixes. The result in Fig. 9.10 suggests that this may not be the case generally. That being said, in the earlier work of De Man et al. [82], not all participants preferred their own mixes, as mix engineer **'W'** in Fig. 9.1 appears to have been subject to the opposite psychological effect and rated their own mix poorly. There are a number of differences between these two studies and their results.

- Here, mix engineers were not told that their result from part 1 acted as a template for the mixes in part 2, i.e. they did not necessarily recognise the mixes in part 2 as being *"their"* mixes.
- Here, mix engineers rated *"their"* mixes of eight previously unknown songs.

The result from Fig. 9.10 is difficult to interpret. For two mixers, they rate the mixes in their style to be better than average, and on-par with the highest rated mixer. However, MixerA rated the mixes drawn from their template to be very poor (although close to the consensus rating). MixerB considered *their* mixes to be average, as too did the other participants. The overall trend seems to be that an individual's impression of *their* mixes is an exaggerated form of the consensus. As shown in Fig. 9.11, the mixes in the style of MixerD achieved the highest median score for 5/8 songs, while the mixes in the style of MixerA achieved the lowest rating for 5/8 songs.

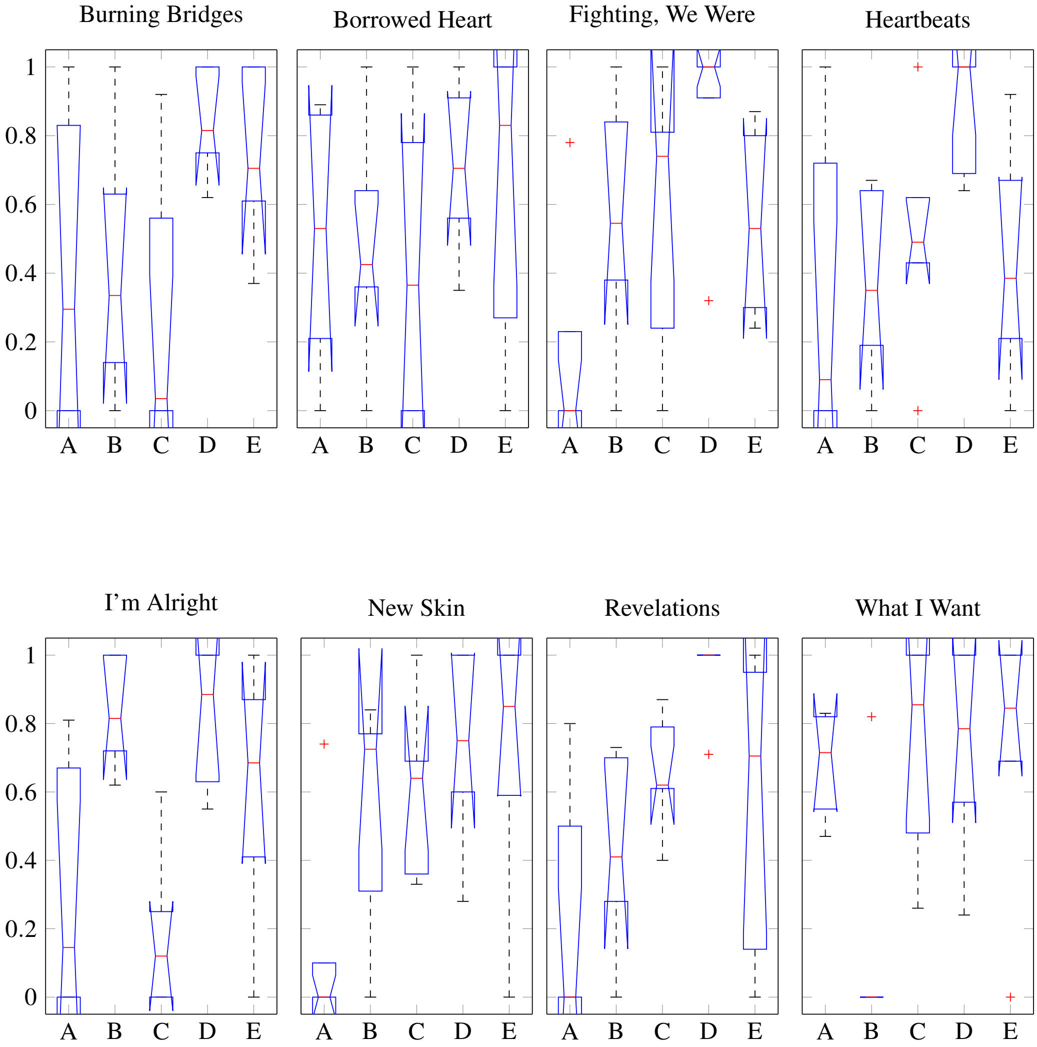
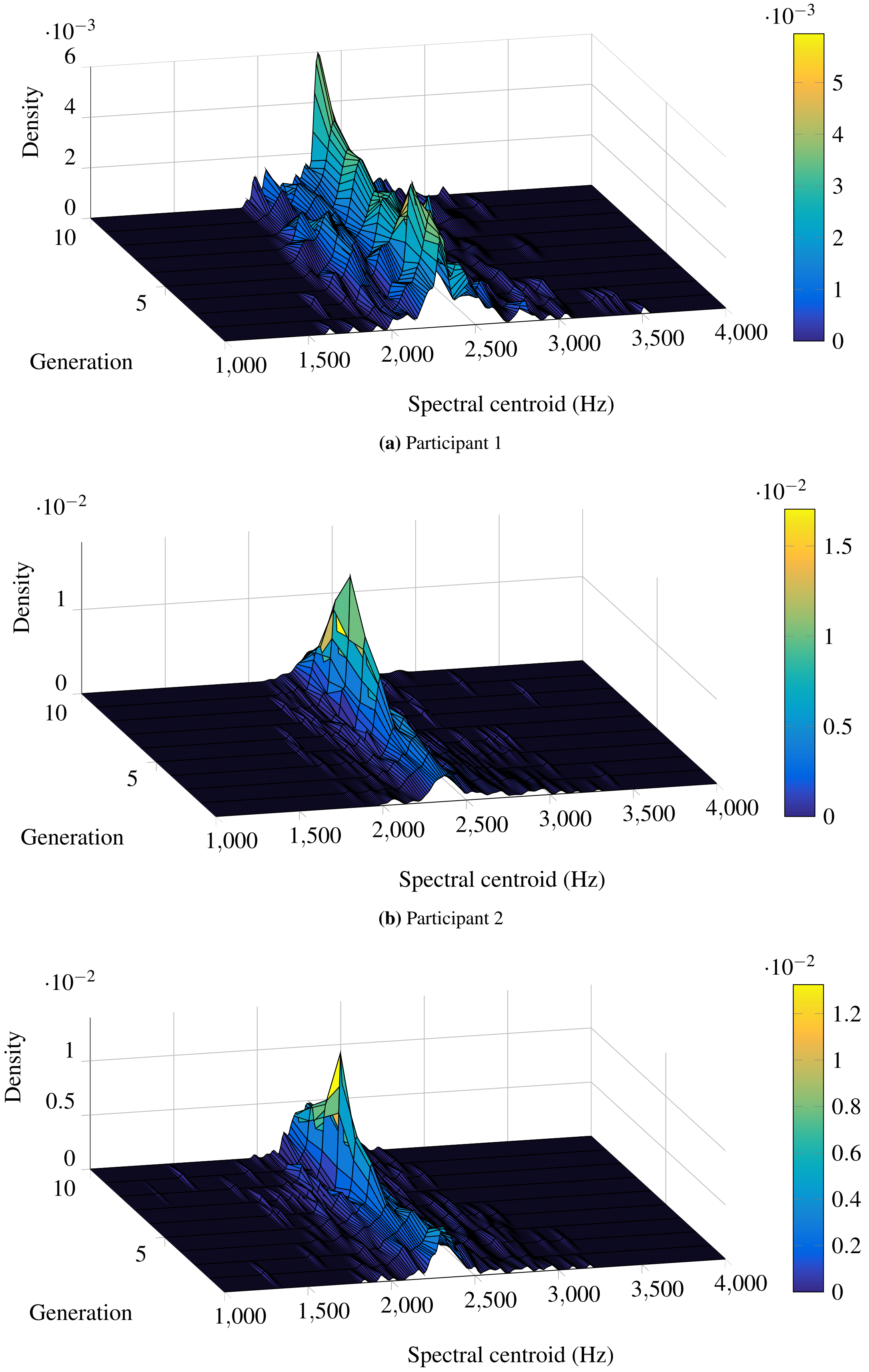


Figure 9.11: Boxplot of ratings, shown for each song

9.5 Features of generated mixes

Overall, the number of mixes created was equal to $n_{par} \times n_{pop} \times n_{gen}$, which is $14 \times 100 \times 10 = 14,000$. As the population at each generation was recorded, audio signal features can be extracted from the mixes of each participant. This allowed an investigation into the variation of certain features over the duration of the trial.

Below, in Figs. 9.12a to 9.13c, the estimated PDF of spectral centroid is displayed at each generation. To avoid unnecessary repetition, only data from these first six participants are displayed. In all cases shown, the maximum value is at a late generation, usually the final generation. This indicates that the distribution of mixes converges. Some participants, like P1 and P5, consistently created mixes with a variety of spectral centroid values, as evidenced by the multi-modal nature of the estimated PDF. This suggests that some participants did not base their ratings purely on this feature. In contrast, other participants, such as P2 and P4, show much smoother surfaces, indicating that the degree of convergence was higher, at least in terms of spectral centroid values. Note that the peak at each generation is always close to 2.4 kHz. Compared to the ten songs analysed in Fig. 6.5, this is a relatively low value. Clearly, a purely objective GA, aiming for the average value of 3.6 kHz (see Table 6.6), would produce overly bright-sounding mixes.

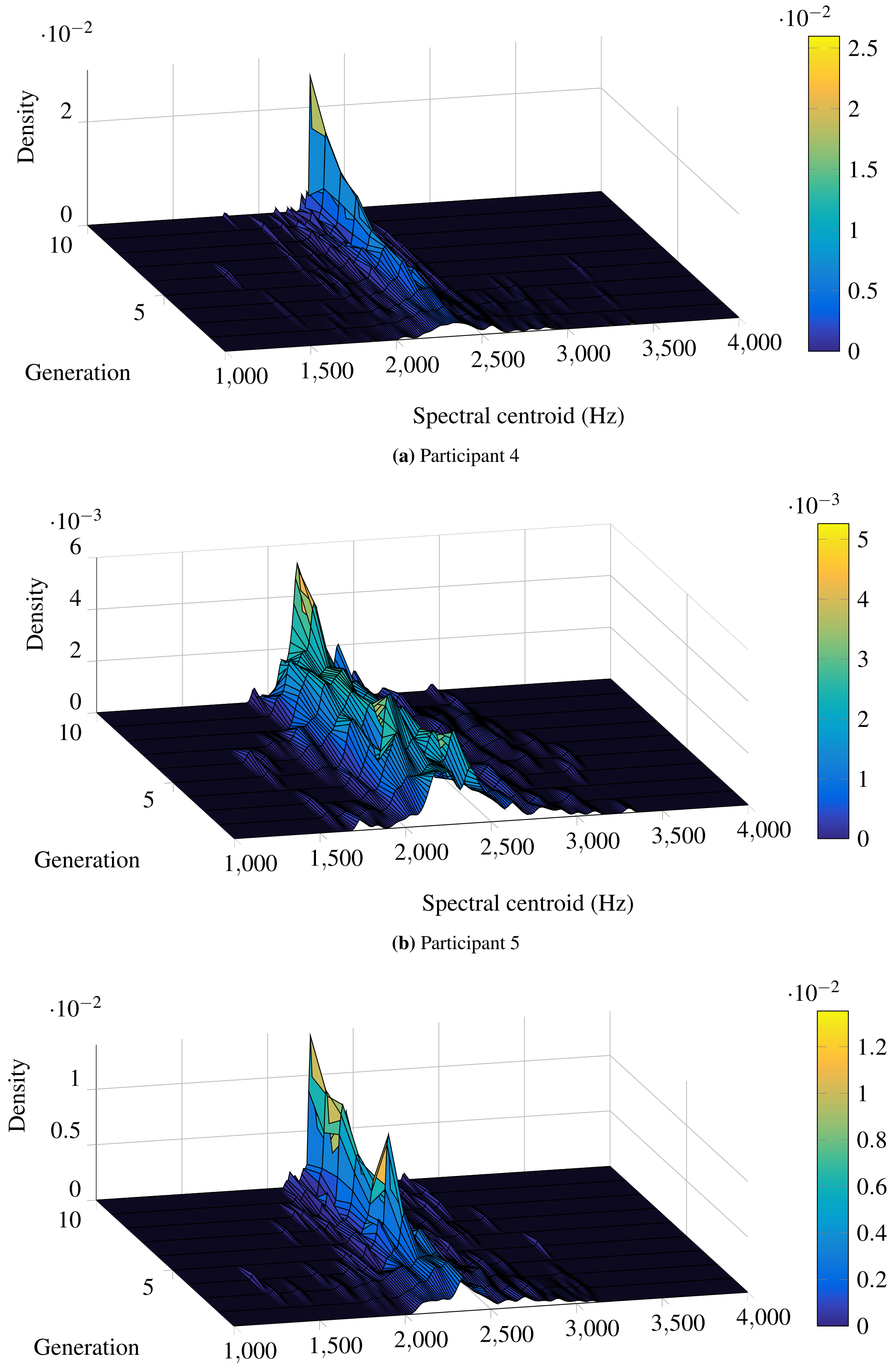


(a) Participant 1

(b) Participant 2

(c) Participant 3

Figure 9.12: Evolution of spectral centroid, in mixes of participants 1 to 3.



(a) Participant 4

(b) Participant 5

(c) Participant 6

Figure 9.13: Evolution of spectral centroid, in mixes of participants 4 to 6.

9.6 Discussion

The following is a response to each of the research questions.

- The median levels of instruments in the mixes generated by the IGA method showed that vocals were, once again, set as the loudest element in the mix. Other instruments were also set to similar levels.

- The diversity of mixes compares well to the traditional fader-based approach. This suggests that the system is working, as it is capable of creating the types of mixes desired by the users.

- From the survey results one can see that the users were pleased with usability of the system. Users rated the system as easy to use and that it was neither physically or mentally demanding. They did, however, state that they did not feel so in control of the process. This might be as expected, since the purpose of the system is to act as an assistant, thereby assuming some level of control from the user. In the case of a user with particular disabilities, this might be more welcome than an able-bodied user.

- Users rated the mixes of other songs, created by their template, in complex ways. Whether or not the mix template, derived from the optimal mix of one song, can be applied to other songs is still debatable. The results indicate that, perhaps, some users' optimal mixes were more generalisable than others.

- The users rate *their own* mixes in various ways, however, it is clear that there was not a consistent preference for ones own mix.

9.6.1 Does the system make music mixing more accessible?

Part of the motivation behind the development of such a system (as described in Chapter 8) was to imagine a music production system that could be operated by a user with severe visual impairment. Many contemporary systems, such as the Digital Audio Workstation (DAW) present a large amount of visual information to the user on a monitor, and the primary interface to this information is through a mouse, trackpad or similar peripheral. These systems present accessibility issues for the visually-impaired. A variety of solutions have been implemented over the years, such as text-to-speech systems which help the user to navigate long menus. Often, a human assistant is required. The IGA-based system could be used to take the place of this assistant. There is room for further work in which the proposed system is evaluated by visually-impaired users and directly compared to existing solutions.

9.6.2 What else can it be used for?

Creating mixes based on some perceptual dimension (such as warmth or brightness, for example) can help our understanding of such concepts. The proposed system therefore has potential for use in perceptual audio evaluation in general. Rather than a user comparing multiple pre-defined stimuli and rating the value of each, the user could start with a "blank slate" and evolve the system towards the optimal solution. Then the solution found, and the audio signal features of the solution, can be used to learn about the perceptual characteristic. Ultimately, this is what has been done in this chapter, with musical audio as the stimuli and "preference" as the perceptual dimension.

This type of system can also be used for the deployment of object based audio broadcast, in which audio objects need to be mixed at delivery stage. Such a system could learn quickly from the user and be less based on fixed best practices, which may not meet the personal requirements of each user. Furthermore, it has been shown that IEC is a suitable method for designing systems based on subjective audio evaluation. For example, this can be used in designing rooms for a specific level of subjective speech intelligibility (rather than using an objective approximation), or designing products for a specific sound quality using virtual acoustic prototypes.

9.7 Summary of chapter

When using the proposed IGA-based music mixing system, participants were able to create a range of mixes comparable to those made using the conventional fader-based system. This suggests that the system is not an obstacle to the creation of mixes, and does not impose noticeable limits on what mixes can be created. The system was considered to be highly usable. Both physical and mental demands were reported to be low. Consequently it is predicted that the system would be suitable for a variety of applications where physical interaction is to be kept low.

In light of the participants ratings, certain improvements could certainly be made. The most-significant-variable problem (see § 8.2) reduces the level of control the user has over the system. Further work would include solving this problem and removing the bias it introduces. Recall that only five mixes were presented to the user, per generation. If none of these five are especially good then the system would experience a setback and may fail to evolve towards the desired goal. Thus far, the system seems fairly robust to this problem, which has been referred to as the "hamming cliff" problem. Gray encoding of the binary values is a potential solution which has shown promise in previous studies [224, 225]. Although not formally tested, Gray encoding was added to the IGA algorithm and shows promising results.

Participants were able to detect an effect of the mix engineer, as the mixes in the style of MixerD were significantly preferred over those of all others. Participants' preference for their own mixes was varied: this suggests that the system generalises well for some users and less so for others but also that, being trained on only one song, the training data may not have provided a sufficiently general template for mixes.

10 Conclusions

The aims of this project were to address fundamental questions concerning music mixing and to incorporate these at the heart of a novel framework for audio analysis and intelligent music production. These questions were as follows:

- What is mixing, i.e. what can be achieved by mixing?
- What makes a good mix?
- How can good mixes be generated?

Each of these questions has been directly addressed within this thesis. As these questions are fundamental, exhaustive answers would lie far beyond the scope of this thesis. In fact, it has been shown that the answers to the first two questions are highly subjective and must take into account aesthetic concerns.

This is one of the main contributions of the thesis — that if audio quality is to be a motivating factor in the automatic creation of mixes, then the question must be asked, for whom is the mix being created and what is their impression of audio quality? One simple way to address this issue is to directly include this individual in the process, so they can specify to the system their preferences, guiding the system towards the desired output.

10.1 Main findings

10.1.1 Perception of quality

Based on the results of a listening test, using commercially available audio stimuli, it was indicated that the perception of quality was a percept distinct from hedonic preference. However, in a test comparing a set of alternate mixes, this distinction was no longer observed. From this, it can be said that preference is a motivating factor in maximising the quality of a mix.

10.1.2 Mix-space

It has been indicated that previously-held definitions, wherein an audio mix is described as the sum of individual input channels, have led to sub-optimal solution spaces for optimisation, leading to varied levels of performance in applications relating to intelligent music production. This thesis has proposed an alternative definition, that a mix can be considered as a series of inter-channel balances. This solution space contains all possible mixes that can be created using a finite set of tools, such as gain, panning and equalisation.

10.1.2.1 ...as a framework for analysis

This proves to be a useful framework for the comparative analysis of mixes created using conventional means. Based on a series of experimental studies, it has been suggested that...

- ...there exists a consensus among test participants when choosing the loudness levels at which to set instruments
- ...the rough mix presented to test participants influences their final mix, providing evidence of an anchoring effect
- ...for mono mixes, there was little difference between the mixes generated via headphone reproduction or loudspeaker reproduction
- ...for stereo mixes, the panning of instruments provides a degree of spatial unmasking, changing the consensus on loudness levels.
- ...there exists a consensus on the choice of pan positions, with vocals, bass, kick drum and snare drum being placed in the centre of the stereo field, while drum overheads and paired guitars were panned much wider
- ...test participants were influenced by the layout of tracks, panning drum overheads and guitars according to their relative positions on the mixing console.
- ...the application of equalisation typically boosted salient frequencies in the instrument
- ...there was no consensus on the application of equalisation to vocals.

10.1.2.2 ...as a means of generating mixes

In addition to providing a framework for mix analysis, the mix-space allows for the creation of mixes. As it is costly to gather a large set of mixes, created by mix engineers as the basis of experiment and analysis, the creation of a set of mixes, either randomly or in accordance with parametric models, has allowed for the study of mix-diversity. For a number of songs, sets of

1,000 mixes were generated according to various parameters and the audio signal features of these mixes were obtained. This revealed tolerance regions beyond which such a feature is unlikely to stray.

Generating a set of mixes in a defined solution space is the first step in a genetic algorithm, and this was used to create a novel mixing system. Users of the system were able to generate mixes in accordance with any perceptual property, such as “quality”. The mixes made were comparable to those made using the conventional fader-based approach, in terms of consensus in instrument level. Users found the system easy to use, in terms of physical and mental burden. Such a system has applications in assistive technologies, allowing greater access to music production.

At the core of the mix-space concept is the notion that a creative task can be represented using a design space and that this space can be explored, algorithmically, with an artist acting as director. The great potential of this approach should be explored further in future work, in the realm of audio production, acoustic design and beyond.

10.1.3 Analysis of real mixes

Typically, when listening to music, a listener is only ever exposed to one version of any given recording session: to only one mix. The mix engineers themselves are, of course, exposed to countless possible variations of the recording session. While the mix-space studies showed the paths that engineers take through a limited but well-defined solution space, the mixes created under real world conditions contain too many variables to study in detail.

While previous studies have gathered perhaps a dozen, or few dozen, mixes of a song, this thesis has examined the variation in larger collections of mixes, up to 373. This has provided a more detailed picture of the possible variation in mixes that mix engineer can explore.

This has indicated realistic upper and lower bounds for many audio signal features, features often used to guide optimisation processes in automated mixing tasks. This knowledge can be used to detect if an automatically generated mix is one that would be unlikely to have been created by a real mix engineer. If so desired, it can then be downgraded, or vetoed, by the system.

Of course, only low-level features were obtained from the signals and there is an increasing question as to whether or not the “bag-of-frames” approach is a useful representation for music signals. The work in this thesis has indicated that while these signal features may be useful in many information retrieval tasks up to now, when comparing different songs to one another, that they are of limited practical use when analysing the often-subtle differences between multiple mixes of the same song.

Gathering such a large set of real mixes involved scraping the largest known database of mix content, the cambridge multitracks website. As the studies on this thesis only focussed on the ten songs that were most often mixed, there exist thousands more mixes for further study.

In the majority of cases, such mixes lack any form of explicit subjective evaluation. Since this form of direct evaluation would be practically impossible to gather for a set of thousands of audio samples, the implicit subjective evaluation is of great importance. This is to say that, if a mix has been created by an engineer, after some period of time exploring the many possible mixes that exist, then this final mix can be considered to have value, and quality can be considered as being related to the value of an object. This observation is important as it opens up such studies to the world of big data, beyond the confines of a limited laboratory-based study.

Where subjective evaluation is available, the data indicates that a good mix is one which reaches a compromise between a number of factors, such as loudness/dynamics, brightness, stereo width and bass. This has only been shown here for one song as vast resources would be required to extend this study to other songs, for an unknown payoff in regards to this thesis. We invite other authors to study mixes of other songs as further work. A second mix competition was carried out, using a similar format, in November 2016.

10.1.4 Analysis of mix engineers

In addition to an investigation into mixes themselves, this thesis outlined an investigation into the individual mix-engineers. When comparing six engineers, who had each created mixes for 18 songs, it was found that there were differences in terms of the audio signal features of their mixes but only a very small perceptual effect of the mixers “style” on preference ratings.

After creating mixes using the IGA system, test participants then rated mixes of other songs created in their style, using their initial mix as a template. In this subjective evaluation, the mixes of one participant were clearly preferred. Interestingly, participants did not appear to identify their own style of mixes, perhaps indicating that more than one initial mix is required to create a template.

In regards to the variation in mixes and in mix-engineers, both of these findings are novel and important, as trivial as they may appear. We now know that while mixes, on the whole, differ from one another in some predictable way, and that features vary based on simple parametric models, additionally, individual mix engineers are shown to vary, in a purely feature-based model. If it is true that the audio signal features can tell us something interesting about the audio signal, then it can be said that quantifiable evidence now exists to suggest that mix-engineers do have a measurable style, which has been suggested anecdotally for some time.

10.2 Further work

There is the possibility of the models presented in this thesis to be expanded to encompass more complex music production tasks, such as time-varying fader control (which was briefly discussed in § 4.2) and dynamic range processing. However, this section describes new projects which can be undertaken as a consequence of the research presented in this thesis.

10.2.1 Emotion in mixes

Using the concept of a mix-space as a framework for testing, there are numerous further works that can be undertaken as a consequence of this thesis. One recent project that was undertaken but not included in this thesis investigated the psychoacoustics of emotions in music signals, specifically, does the mix influence a listeners perception of the song's mood? This used a set of alternate mixes, created by different mix engineers, as audio stimuli. This project could be continued using the mix-space framework, allowing for psychophysical time-series data to be mapped onto the mix-space as the user creates their mix. Ideally, this would be expected to reveal regions of the mix-space where highly emotional mixes exist. This aim could also be accomplished using the IGA mixer, where instead of optimising the "quality" of a mix, one is asked to optimise the degree of "happiness" or some other mood. This would be a challenging and ambitious project with potential for unique applications in the music technology industries.

10.2.2 Audio evaluation methodology

Ultimately, it is hoped that this work could lead to new methods of psychoacoustic evaluation. Often, in such a test, a participant is provided with a fixed number of ready-made audio stimuli and these are rated on some scale. It would be interesting to develop methods where the audio stimuli are generated as part of the test. Of course, this would require careful statistical analysis, as no two participants will have taken the same test — although, already, no two participants are the same, and so perhaps this type of subjective testing will become more commonplace in the future.

10.2.3 Quality

The relationship between quality and audio signals is a broad topic with a vast amount of applications. Many studies have reported on technical issues in the recording of the signal, such as wind noise or handling noise, or artefacts in the signal representation and storage, which can be introduced by perceptual coding. What is still required is an ambitious study into the interaction between subjective experience and audio signals in music.

10.2.4 Mix-analysis

In April of 2017, the audio attachments uploaded to the forums of the Cambridge Music Technology website were archived for further study. This archive contains over 15,000 mixes created from the multitrack sessions of over 300 different songs.

Further work would included developing new metrics which approximate the quality of a mix. One such avenue to explore would be relating to the musical structure of mixes. The quality of mix may be related to how interesting it is to the user, which in turn may be explained by entropy in the structure, i.e. a mix with contrasting verse and chorus may be more interesting than one in which these sections are more homogeneous. With dozens or hundreds of available real-world mixes (or thousands of artificially generated mixes), a large-scale statistical analysis of music structure could be undertaken.

A
Publications

The following journal and conference papers have been published as part of this project.

- Wilson, A. and Fazenda, B. (2013). Perception & evaluation of audio quality in music production. In Proc. of the 16th Int. Conference on Digital Audio Effects (DAFx-13), Maynooth, Ireland.

- Wilson, A. and Fazenda, B. (2014). Characterisation of distortion profiles in relation to audio quality. In Proc. of the 17th Int. Conference on Digital Audio Effects (DAFx-14), Erlangen, Germany.

- Wilson, A. and Fazenda, B. M. (2016). Perception of audio quality in productions of popular music. J. Audio Eng. Soc, 64(1/2):23–34.

- Wilson, A. and Fazenda, B. M. (2015). A Lexicon of Audio Quality. In Proceedings of the 9th Triennial conference of the European Society for the Cognitive Sciences of Music (ESCOM 2015), Manchester, UK.

- Wilson, A. and Fazenda, B. M. (2015). Navigating the mix-space: theoretical and practical level-balancing technique in multitrack music mixtures. In Proceedings of the Sound and Music Computing Conference, Maynooth, Ireland.

- Wilson, A. and Fazenda, B. (2016). Variation in multitrack mixes: Analysis of low-level audio signal features. J. Audio Eng. Soc, 64(7/8):466–473.

- Wilson, A. and Fazenda, B. M. (2015a). 101 mixes: a statistical analysis of mix-variation in a dataset of multitrack music mixes. In 139th AES Convention, pages 1–10, New York, New York, USA.

- Wilson, A. and Fazenda, B. (2016). An evolutionary computation approach to intelligent music production, informed by experimentally gathered domain knowledge. In 2nd AES Workshop on Intelligent Music Production.

Bibliography

[1] Carol A Reeves and David A Bednar. Defining Quality: Alternatives and Implications. *Academy of Management Review*, 19(3):419–445, 1994. ISSN 0363-7425. doi: 10.5465/AMR.1994.9412271805.

[2] ISO 9000:2005. ISO 9000:2005 Quality management systems – Fundamentals and vocabulary, 2009.

[3] Antonio J Verdú Jover, Francisco Javier Lloréns Montes, and María Del Mar Fuentes Fuentes. Measuring perceptions of quality in food products: the case of red wine. *Food Quality and Preference*, 15(5):453–469, jul 2004. ISSN 09503293. doi: 10.1016/j.foodqual.2003.08.002.

[4] Elizabeth C Thach and Janeen E Olsen. Market Segment Analysis to Target Young Adult Wine Drinkers. *Agribusiness*, 22(3):307–322, 2006. ISSN 0742-4477. doi: 10.1002/agr.

[5] Beth A Benjamin and Joel M Podolny. Social Order in the California Wine Industry. *Administrative Science Quarterly*, 44:563–589, 1999.

[6] Johan Bruwer, Anthony Saliba, and Bernadette Miller. Consumer behaviour and sensory preference differences: implications for wine product marketing. *Journal of Consumer Marketing*, 28(1):5–18, 2011. ISSN 0736-3761. doi: 10.1108/07363761111101903.

[7] Emin Babakus and Gregory W Boller. An empirical assessment of the SERVQUAL scale. *Journal of Business Research*, 24(3):253–268, 1992. ISSN 01482963. doi: 10.1016/0148-2963(92)90022-4.

[8] Ute Jekosch. Basic concepts and terms of "quality", reconsidered in the context of product-sound quality. *Acta Acustica united with Acustica*, 90(6):999–1006, 2004.

[9] Patrick Le Callet, Sebastian Möller, and Andrew Perkis. Qualinet White Paper on Definitions of Quality of Experience (2012). Technical report, University of Zurich, Department of Informatics, March 2013.

[10] Jens Blauert and Ute Jekosch. A layer model of sound quality. *Journal of the Audio Engineering Society*, 60(1), 2012.

[11] Tomasz Letowski. Sound quality assessment: Concepts and criteria. In *Audio Engineering Society Convention 87*, Oct 1989.

[12] Jan Berg and Francis Rumsey. Systematic evaluation of perceived spatial quality. In *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality*. Audio Engineering Society, 2003.

[13] Sarah Le Bagousse, Mathieu Paquier, and Catherine Colomes. Categorization of sound attributes for audio quality assessmenta lexical study. *Journal of the Audio Engineering Society*, 62(11):736–747, 2014.

[14] Torben H Pedersen and Nick Zacharov. The development of a sound wheel for reproduced sound. In *Audio Engineering Society Convention 138*, May 2015.

[15] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 749–752. IEEE, 2001.

[16] 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.

[17] James M Kates and Kathryn H Arehart. The hearing-aid speech quality index (HASQI). *Journal of the Audio Engineering Society*, 58(5):363–381, 2010.

[18] James M Kates and Kathryn H Arehart. The hearing-aid speech quality index (HASQI) version 2. *Journal of the Audio Engineering Society*, 62(3):99–117, 2014.

[19] Abigail A Kressner, David V Anderson, and Christopher J Rozell. Evaluating the generalization of the hearing aid speech quality index (HASQI). *IEEE transactions on audio, speech, and language processing*, 21(2):407–415, 2013.

[20] Paul Kendrick, Francis Li, Bruno Fazenda, Iain Jackson, and Trevor Cox. Perceived audio quality of sounds degraded by nonlinear distortions and single-ended assessment using HASQI. *Journal of the Audio Engineering Society*, 63(9):698–712, 2015.

[21] ITU. ITU-R BS.1116-1 Methods for the subjective assessment of small impairments, 1997.

[22] ITUR Recommendation. Bs. 1534-1. method for the subjective assessment of intermediate sound quality (mushra). *International Telecommunications Union, Geneva*, 2001.

[23] Judith Liebetrau, Frederik Nagel, Nick Zacharov, Kaoru Watanabe, Catherine Colomes, Poppy Crum, Thomas Sporer, and Andrew Mason. Revision of Rec. ITU-R BS. 1534. In *Audio Engineering Society Convention 137*. Audio Engineering Society, 2014.

[24] Valorie N Salimpoor, Mitchel Benovoy, Gregory Longo, Jeremy R Cooperstock, and Robert J Zatorre. The rewarding aspects of music listening are related to degree of emotional arousal. *PloS one*, 4(10):e7487, jan 2009. ISSN 1932-6203. doi: 10.1371/journal.pone.0007487.

[25] Carlos Silva Pereira, João Teixeira, Patrícia Figueiredo, João Xavier, São Luís Castro, and Elvira Brattico. Music and emotions in the brain: familiarity matters. *PloS one*, 6(11): e27241, jan 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0027241.

[26] David J Hargreaves. The effects of repetition on liking for music. *Journal of Research in Music Education*, 32(1):35–47, 1984.

[27] Isabelle Peretz, Danielle Gaudreau, and Anne-Marie Bonnel. Exposure effects on music preference and recognition. *Memory & Cognition*, 26(5):884–902, 1998. ISSN 0090-502X. doi: 10.3758/BF03201171.

[28] Karl K Szpunar, E Glenn Schellenberg, and Patricia Pliner. Liking and memory for musical stimuli as a function of exposure. *Journal of experimental psychology. Learning, memory, and cognition*, 30(2):370–81, mar 2004. ISSN 0278-7393. doi: 10.1037/0278-7393.30.2.370.

[29] Patrick G Hunter and E Glenn Schellenberg. Interactive effects of personality and frequency of exposure on liking for music. *Personality and Individual Differences*, 50(2):175–179, jan 2011. ISSN 01918869. doi: 10.1016/j.paid.2010.09.021.

[30] Eberhard Zwicker and U Tilmann Zwicker. Audio Engineering and Psychoacoustics: Matching Signals to the Final Receiver, the Human Auditory System. *Journal of the Audio Engineering Society*, 39(3):115–126, 1991.

[31] Brian R Glasberg and Brian C J Moore. A model of loudness applicable to time-varying sounds. *Journal of the Audio Engineering Society*, pages 331–342, 2002.

[32] ITU. ITU-R BS.1770-3 Algorithms to measure audio programme loudness and true-peak audio level, 2012.

[33] Sven-Amin Lembke, Scott Levine, Martha de Francisco, and Stephen McAdams. The use of microphone level balance in blending the timbre of horn and bassoon. In *Audio Engineering Society Convention 139*, pages 1–8, 2015.

[34] Richard King, Brett Leonard, and Grzegorz Sikora. Variance in Level Preference of Balance Engineers: A study of mixing preference and variance over time. *Audio Engineering Society Convention 129*, 2010.

[35] Richard King, Brett Leonard, and Grzegorz Sikora. Consistency of balance preferences in three musical genres. *Audio Engineering Society Convention 133*, pages 1–6, 2012.

[36] 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)*, pages 137–142, 2014.

[37] Mark Brozier Cartwright and Bryan Pardo. Social-EQ: Crowdsourcing an Equalization Descriptor Map. In *14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, pages 395–400, 2013.

[38] Ryan Stables, Sean Enderby, Brecht De Man, Gyorgy Fazekas, and Joshua Reiss. SAFE: a System for the Extraction and Retrieval of Semantic Audio Descriptors. In *15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, pages 1–2, 2014.

[39] Hyunkook Lee and Francis Rumsey. Level and Time Panning of Phantom Images for Musical Sources. *Journal of the Audio Engineering Society*, 61(12), 2013.

[40] Pedro Pestana and Joshua D Reiss. Intelligent Audio Production Strategies Informed by Best Practices. *AES 53rd International Conference: Semantic Audio*, pages 1–9, 2014.

[41] Pedro Duarte Pestana, Joshua D Reiss, and Álvaro Barbosa. User preference on artificial reverberation and delay time parameters. *Journal of the Audio Engineering Society*, 65 (1/2):100–107, 2017.

[42] Brecht De Man, Kirk McNally, and Joshua D Reiss. Perceptual evaluation and analysis of reverberation in multitrack music production. *Journal of the Audio Engineering Society*, 65 (1/2):108–116, 2017.

[43] Emmanouil T Chourdakis and Joshua D Reiss. A machine-learning approach to application of intelligent artificial reverberation. *Journal of the Audio Engineering Society*, 65(1/2): 56–65, 2017.

[44] Zheng Ma. *Intelligent Tools for Multitrack Frequency and Dynamics Processing*. PhD thesis, Queen Mary University of London, 2016.

[45] Brett Leonard, Richard King, and Grzegorz Sikora. The effect of acoustic environment on reverberation level preference. *Audio Engineering Society Convention 133*, 2012.

[46] Richard King, Brett Leonard, and Grzegorz Sikora. The Effects of Monitoring Systems on Balance Preference: A Comparative Study of Mixing on Headphones Versus Loudspeakers. *Audio Engineering Society Convention 131*, pages 1–7, 2011.

[47] Dan Dugan. Automatic microphone mixing. *Journal of the Audio Engineering Society*, 23 (6):442–449, 1975.

[48] Dan Dugan. Tutorial: Application of automatic mixing techniques to audio consoles. *SMPTE journal*, 101(1):19–27, 1992.

[49] Dale Reed. A Perceptual Assistant to do Sound Equalization. In *Proceedings of the 5th international conference on Intelligent user interfaces*, pages 212–218, 2000.

[50] James A Moorer. Audio in the new millennium. *Journal of the Audio Engineering Society*, 48(5):490–498, 2000.

[51] François Pachet and Olivier Delerue. On-the-fly multi-track mixing. In *Audio Engineering Society Convention 109*. Audio Engineering Society, 2000.

[52] Haruhiro Katayose, Akio Yatsui, and Masataka Goto. A mix-down assistant interface with reuse of examples. In *Automated Production of Cross Media Content for Multi-Channel Distribution, 2005. AXMEDIS 2005. First International Conference on*, pages 8–pp. IEEE, 2005.

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

[54] Enrique Perez-Gonzalez and Joshua Reiss. Improved control for selective minimization of masking using inter-channel dependancy effects. In *Proceedings of the 11th International Conference on Digital Audio Effects (DAFx-08)*, pages 1–7, Espoo, Finland, 2008. ISBN 9789512295173.

[55] Alexandros Tsilfidis, Charalambos Papadakos, and John Mourjopoulos. Hierarchical perceptual mixing. In *Audio Engineering Society Convention 126*. Audio Engineering Society, 2009.

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

[57] Joshua D Reiss. Intelligent music production: Challenges, frontiers and implications. In *Proceedings of the 1st AES Workshop on Intelligent Music Production*, 2015.

[58] Zheng Ma, Brecht De Man, Pedro D L Pestana, Dawn Black, and Joshua D Reiss. Intelligent Multitrack Dynamic Range Compression. *Journal of the Audio Engineering Society*, 63(6):412–426, 2015.

[59] Zheng Ma, Joshua D Reiss, and Dawn Black. Implmentation of an intelligent equalization tool using Yule-Waker for music mixing and mastering. *Audio Engineering Society Convention 134*, 2013.

[60] Brecht De Man and Joshua D Reiss. A knowledge-engineered autonomous mixing system. *Audio Engineering Society Convention 135*, 2013.

[61] Pedro Pestana. *Automatic mixing systems using adaptive audio effects*. PhD thesis, Universidade catolica portuguesa, 2013.

[62] Dominic Ward, Joshua D Reiss, and Cham Athwal. Multi-track mixing using a model of

loudness and partial loudness. In *Audio Engineering Society Convention 133*, San Francisco, USA, 2012.

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

[64] 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, York, UK, 2012.

[65] Stuart Mansbridge, Saoirse Finn, and Joshua D Reiss. Implementation and evaluation of autonomous multi-track fader control. In *Audio Engineering Society Convention 132*, Budapest, Hungary, April 2012. Audio Engineering Society.

[66] Phillip Aichinger, Alois Sontacchi, and Berit Schneider-Stickler. Describing the transparency of mixdowns: The Masked-to-Unmasked-Ratio. In *Audio Engineering Society Convention 130*, pages 1–10, London, UK, 2011. ISBN 9781617829253.

[67] Gregory Bocko, Mark F Bocko, Dave Headlam, Justin Lundberg, and Gang Ren. Automatic Music Production System Employing Probabilistic Expert Systems. In *Audio Engineering Society Convention 129*, 2010.

[68] Sebastian Vega and Jordi Janer. Quantifying masking in multi-track recordings. *Proceedings of SMC Conference 2010, Barcelona*, 2010.

[69] Michael Terrell and Joshua D Reiss. Automatic monitor mixing for live musical performance. *Journal of the Audio Engineering Society*, 57(11):927–936, 2009.

[70] 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.

[71] Andrew T Sabin and Bryan Pardo. A method for rapid personalization of audio equalization parameters. *Proceedings of the 17th ACM international conference on Multimedia - MM '09*, page 769, 2009. doi: 10.1145/1631272.1631410.

[72] Andrew T Sabin and Bryan Pardo. 2Deq: an intuitive audio equalizer. *Proceedings of the 7th ACM conference on Creativity and cognition - C&C '09*, page 435, 2009. doi: 10.1145/1640233.1640339.

[73] Sebastian Heise, Michael Hlatky, and Joern Loviscach. Automatic Adjustment of Off-the-Shelf Reverberation Effects. *Audio Engineering Society Convention 126*, pages 1–8, 2009.

[74] Daniele Barchiesi and Joshua Reiss. Automatic target mixing using least-squares optimization of gains and equalization settings. In *Proceedings of the 12th International Conference on Digital Audio Effects (DAFx-09)*, pages 1–8, Como, Italy, 2009.

[75] Daniele Barchiesi and Joshua Reiss. Reverse engineering of a mix. *Journal of the Audio Engineering Society*, 58(7-8):563–576, 2010. ISSN 15494950.

[76] Enrique Perez-Gonzalez and Joshua Reiss. Automatic mixing: live downmixing stereo panner. In *Proceedings of the 10th International Conference on Digital Audio Effects (DAFx-07)*, pages 1–6, Bordeaux, France, 2007. ISBN 9788890147913.

[77] Enrique Perez-Gonzalez and Joshua Reiss. Automatic equalization of multichannel audio using cross-adaptive methods. *Audio Engineering Society Convention 127*, 2009.

[78] 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(1), 2010. ISSN 16876172. doi: 10.1155/2010/436895.

[79] Patricia Yancey Martin and Barry A Turner. Grounded theory and organizational research. *The journal of applied behavioral science*, 22(2):141–157, 1986.

[80] Jeffrey Scott and Youngmoo E Kim. Instrument identification informed multi-track mixing. In *14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, 2013.

[81] Jeffrey Scott. Automated Multi-Track Mixing and Analysis of Instrument Mixtures. *Proceedings of the ACM International Conference on Multimedia - MM '14*, pages 651–654, 2014. doi: 10.1145/2647868.2654859.

[82] 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*, Warsaw, Poland, May 2015. Audio Engineering Society.

[83] John H Holland. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence.* U Michigan Press, 1975.

[84] David Goldberg, Kalyanmoy Deb, and Bradley Korb. Messy genetic algorithms: Motivation, analysis, and first results. *Complex systems*, (3):493–530, 1989.

[85] Hee-Su Kim and Sung-Bae Cho. Application of interactive genetic algorithm to fashion design. *Engineering Applications of Artificial Intelligence*, 13(6):635–644, 2000. ISSN 09521976. doi: 10.1016/S0952-1976(00)00045-2.

[86] Michael O'Neill and Anthony Brabazon. Evolving a logo design using lindenmayer systems, postscript & grammatical evolution. In *IEEE Congress on Evolutionary Computation*, pages 3788–3794. IEEE, 2008.

[87] Hideyuki Takagi. Interactive evolutionary computation: Fusion of the capabilities of ec optimization and human evaluation. *Proceedings of the IEEE*, 89(9):1275–1296, 2001. ISSN 00189219. doi: 10.1109/5.949485.

[88] Joo-Young Lee and Sung-Bae Cho. Sparse fitness evaluation for reducing user burden in interactive genetic algorithm. *IEEE International Fuzzy Systems Conference Proceedings (FUZZ-IEEE'99)*, pages 998–1003 vol.2, 1999. doi: 10.1109/FUZZY.1999.793088.

[89] 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.

[90] Thomas Görne and Martin Schneider. Design of digital filters with evolutionary algorithms. In *Artificial Neural Nets and Genetic Algorithms*, pages 368–374. Springer, 1993.

[91] Andrew Rimell and Malcolm Hawksford. The application of genetic algorithms to digital audio filters. In *Audio Engineering Society Convention 98*, Paris, France, 1995.

[92] Shin-ichi Sato, Keisuke Otori, Atsushi Takizawa, Hiroyuki Sakai, Yoichi Ando, and Hiroshi Kawamura. Applying genetic algorithms to the optimum design of a concert hall. *Journal of Sound and Vibration*, 258(3):517–526, 2002.

[93] Marine Baulac, Jérôme Defrance, and Philippe Jean. Optimisation with genetic algorithm of the acoustic performance of t-shaped noise barriers with a reactive top surface. *Applied*

Acoustics, 69(4):332–342, 2008.

[94] Anton Schlesinger and Marinus M Boone. Evolutionary optimization for hearing aids of computational auditory scene analysis. In *Audio Engineering Society Convention 126*. Audio Engineering Society, 2009.

[95] Eric S Schwenker and Griffin D Romigh. An evolutionary algorithm approach to customization of non-individualized head related transfer functions. In *Audio Engineering Society Convention 137*. Audio Engineering Society, 2014.

[96] Andrew Horner, James Beauchamp, and Lippold Haken. Machine Tongues XVI - Genetic Algorithms and their Applications to FM Matching Synthesis. *Computer Music Journal*, 17(4):17–29, 1993. ISSN 0148-9267.

[97] Colin G Johnson. Exploring the sound-space of synthesis algorithms using interactive genetic algorithms. In *Proceedings of the AISB'99 Symposium on Musical Creativity*, pages 20–27. Society for the Study of Artificial Intelligence and Simulation of Behaviour, 1999.

[98] Janne Riionheimo and Vesa Välimäki. Parameter estimation of a plucked string synthesis model using a genetic algorithm with perceptual fitness calculation. *EURASIP Journal on Applied Signal Processing*, 2003(8):791–805, 2003. ISSN 11108657. doi: 10.1155/S1110865703302100.

[99] James McDermott, Niall Griffith, and Michael O'Neill. Target-driven genetic algorithms for synthesizer control. In *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx 06)*, pages 1–15, Montreal, Canada, 2006.

[100] James McDermott, Niall J L Griffith, and Michael O'Neill. Evolutionary computation applied to sound synthesis. *The Art of Artificial Evolution*, pages 81–101, 2008. doi: 10.1007/978-3-540-72877-1{_}4.

[101] Gordan Krekovi and Davor Petrinoviç. Intelligent Exploration of Sound Spaces Using Decision Trees and Evolutionary Approach. In *Proceedings of ICMC/SMC*, pages 1263–1270, Athens, Greece, September 2014.

[102] Bennett A Kolasinski. A Framework for Automatic Mixing Using Timbral Similarity Measures and Genetic Optimization. In *Audio Engineering Society Convention 124*, pages 1–8, Amsterdam, The Netherlands, 2008.

[103] Ken Sharman and Anna Esparcia-Alcázar. Evolutionary methods for designing digital filters. *Contemporary Music Review*, 22(3):5–19, 2003.

[104] Marcelo Caetano and Xavier Rodet. Evolutionary spectral envelope morphing by spectral shape descriptors. In *International Computer Music Conference*, pages 1–1, 2009.

[105] Sayan Ghosh, Debarati Kundu, Kaushik Suresh, Swagatam Das, and Ajith Abraham. Design of optimal digital iir filters by using a bandwidth adaptive harmony search algorithm. In *World Congress on Nature & Biologically Inspired Computing (NaBIC 2009)*, pages 481–486. IEEE, 2009.

[106] Marcelo Freitas Caetano and Xavier Rodet. Independent manipulation of high-level spectral envelope shape features for sound morphing by means of evolutionary computation. In *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx-10)*, pages 11–21, Graz, Austria, 2010.

[107] Alex Wilson and Bruno Fazenda. Perception & evaluation of audio quality in music production. In *Proceedings of the 16th International Conference on Digital Audio Effects (DAFx-13)*, pages 1–6, Maynooth, Ireland, 2013.

[108] Alex Wilson and Bruno Fazenda. Characterisation of distortion profiles in relation to audio quality. In *Proceedings of the 17th International Conference on Digital Audio Effects (DAFx-14)*, pages 1–8, Erlangen, Germany, 2014.

[109] Alex Wilson and Bruno M Fazenda. A Lexicon of Audio Quality. In *Proceedings of the 9th Triennial conference of the European Society for the Cognitive Sciences of Music (ESCOM 2015)*, Manchester, UK, August 2015.

[110] Alex Wilson and Bruno M Fazenda. Perception of audio quality in productions of popular music. *Journal of the Audio Engineering Society*, 64(1/2):23–34, 2016. http://dx.doi.org/10.17743/jaes.2015.0090.

[111] Pedro D Pestana, Zheng Ma, and Joshua D Reiss. Spectral Characteristics of Popular Commercial Recordings 1950-2010. *Audio Engineering Society Convention 135*, pages 1–7, 2013.

[112] Xavier Serra. A multicultural approach in music information research. In *12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, Miami, Florida (USA), October 2011. International Society for Music Information Retrieval (ISMIR).

[113] Gary Galo. A De-Emphasis Test CD, 2009.

[114] Olivier Lartillot and Petri Toiviainen. A matlab toolbox for musical feature extraction from audio. In *Proceedings of the 10th International Conference on Digital Audio Effects (DAFx-07)*, pages 1–8, Bordeaux, France, 2007.

[115] Emmanuel Deruty. The MIR Perspective on the Evolution of Dynamics in Mainstream Music. In *16th International Society for Music Information Retrieval Conference (ISMIR 2015)*, pages 722–727, 2015.

[116] Gottfried von Bismarck. Timbre of steady sounds: A factorial investigation of its verbal attributes. *Acta Acustica united with Acustica*, 30(3):146–159, 1974.

[117] John M. Grey and John W. Gordon. Perceptual effects of spectral modifications on musical timbres. *The Journal of the Acoustical Society of America*, 63(5):1493–1500, 1978. doi: 10.1121/1.381843.

[118] Stephen McAdams, Suzanne Winsberg, Sophie Donnadieu, Geert Soete, and Jochen Krimphoff. Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological research*, 58(3):177–192, 1995.

[119] Michael Schoeffler and Jürgen Herre. About the Impact of Audio Quality on Overall Listening Experience. In *Proceedings of the Sound and Music Computing Conference*, pages 48–53, Stockholm, Sweden, 2013.

[120] ITU-R BS.1770-3. Algorithms to measure audio programme loudness and true-peak audio level. Technical report, International Telecommunications Union, 2012.

[121] Peter J Rentfrow, Lewis R Goldberg, and Daniel J Levitin. The structure of musical preferences: A five-factor model. *Journal of personality and social psychology*, 100(6): 1139–1157, 2011. doi: 10.1037/a0022406.The.

[122] 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*, pages 63–73, 2012.

[123] George W Snedecor and William G Cochran. Statistical methods, 8th edn. *Ames: Iowa State Univ. Press Iowa*, 1989.

[124] Timothy R Levine and Craig R Hullett. Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research*, 28(4):612–625, 2002.

[125] Damien Tardieu, Emmanuel Deruty, Christophe Charbuillet, and Geoffroy Peeters. Production effect: audio features for recording techniques description and decade prediction. In *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx-11)*, pages 441–446, Paris, France, 2011.

[126] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

[127] 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.

[128] Vinoo Alluri and Petri Toiviainen. Exploring perceptual and acoustical correlates of polyphonic timbre. *Music Perception*, 27(3):223–242, 2010.

[129] Charles D Dziuban and Edwin C Shirkey. When is a correlation matrix appropriate for factor analysis? some decision rules. *Psychological bulletin*, 81(6):358, 1974.

[130] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[131] William Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois, 2015. R package version 1.5.8.

[132] Henry F Kaiser. A second generation little jiffy. *Psychometrika*, 35(4):401–415, 1970.

[133] Graeme D Hutcheson and Nick Sofroniou. *The multivariate social scientist: Introductory statistics using generalized linear models*. Sage, 1999.

[134] Henry F Kaiser and John Rice. Little jiffy, mark iv. *Educational and psychological measurement*, 1974.

[135] Sèbastien Lê, Julie Josse, and François Husson. FactoMineR: An R package for multivariate analysis. *Journal of statistical software*, 25(1):1–18, 2008.

[136] Gilles Raîche, Theodore A Walls, David Magis, Martin Riopel, and Jean-Guy Blais. Non-Graphical Solutions for Cattells Scree Test. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 9(1):23–29, jan 2013. ISSN 1614-1881. doi: 10.1027/1614-2241/a000051.

[137] Henry F Kaiser. The application of electronic computers to factor analysis. *Educational and psychological measurement*, 1960.

[138] John Ruscio and Brendan Roche. Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. *Psychological assessment*, 24(2):282, 2012.

[139] John L Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185, 1965.

[140] Earl Vickers. The loudness war: Background, speculation, and recommendations. In *Audio Engineering Society Convention 129*, Nov 2010.

[141] David Meyer, Kurt Hornik, and Ingo Feinerer. Text mining infrastructure in R. *Journal of Statistical Software*, 25(5):1–54, 2008.

[142] George Kingsley Zipf. *Human behavior and the principle of least effort.* addison-wesley press, 1949.

[143] Mathieu Bastian, Sebastien Heymann, Mathieu Jacomy, et al. Gephi: an open source software for exploring and manipulating networks. *ICWSM*, 8:361–362, 2009.

[144] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS one*, 9(6):e98679, jan 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0098679.

[145] Steven Fenton and Jonathan Wakefield. Objective profiling of perceived punch and clarity in produced music. In *Audio Engineering Society Convention 132*. Audio Engineering Society, 2012.

[146] Steven Fenton and Hyunkook Lee. Towards a perceptual model of punch in musical signals. In *Audio Engineering Society Convention 139*, Oct 2015.

[147] Søren Bech and Nick Zacharov. *Perceptual audio evaluation-Theory, method and application*. John Wiley & Sons, 2007.

[148] Sean Michaels. Death magnetic: ‘loudness war’ rages on. *The Guardian*, October 2008. accessed: 18 March 2014.

[149] Ethan Smith. Even heavy-metal fans complain that today’s music is too loud. *Wall Street Journal*, September 2008. accessed: 18 March 2014.

[150] Emmanuel Deruty and Damien Tardieu. About Dynamic Processing in Mainstream Music. *Journal of the Audio Engineering Society*, 62(1), 2014.

[151] Earl Vickers. The loudness war: Background, speculation, and recommendations. *Audio Engineering Society Convention 129*, pages 1–27, 2010.

[152] Michael Terrell, Andrew Simpson, and Mark Sandler. The Mathematics of Mixing. *Journal of the Audio Engineering Society*, 62(1), 2014.

[153] Alex Wilson and Bruno M Fazenda. Navigating the mix-space: theoretical and practical level-balancing technique in multitrack music mixtures. In *Proceedings of the Sound and Music Computing Conference*, Maynooth, Ireland, July 2015.

[154] L. E. Blumenson. A Derivation of n-Dimensional Spherical Coordinates. *The American Mathematical Monthly*, 67(1):63, 1960. ISSN 0002-9890. doi: 10.2307/2308932.

[155] Sina Hafezi and Joshua D Reiss. Autonomous Multitrack Equalization Based on Masking Reduction. *Journal of the Audio Engineering Society*, 63(5):312–323, 2015.

[156] Brecht De Man, Mariano Mora-Mcginity, György Fazekas, and Joshua D Reiss. The Open Multitrack Testbed. In *Audio Engineering Society Convention 137*. Audio Engineering Society, 2014.

[157] Rachel Bittner, Justin Salamon, Mike Tierney, and Matthias Mauch. MedleyDB: a multitrack dataset for annotation-intensive mir research. In *15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, 2014.

[158] Pedro Duarte Pestana, Joshua D Reiss, and Alvaro Barbosa. Loudness measurement of multitrack audio content using modifications of itu-r bs. 1770. In *Audio Engineering Society Convention 134*, Rome, Italy, May 2013. Audio Engineering Society.

[159] Richard King, Brett Leonard, and Grzegorz Sikora. Loudspeakers and headphones: The effects of playback systems on listening test subjects. *Proceedings of Meetings on Acoustics*, 19(1):035035, 2013. doi: 10.1121/1.4799550.

[160] Kurt Hornik, Ingo Feinerer, Martin Kober, and Christian Buchta. Spherical k-Means Clustering. *Journal of Statistical Software*, 50(10):1–22, 2012. ISSN 1548-7660.

[161] John R Baumgardner and Paul O Frederickson. Icosahedral discretization of the two-sphere. *SIAM Journal on Numerical Analysis*, 22(6):1107–1115, 1985. ISSN 00361429.

[162] Niranjan Damera-Venkata, Brian L Evans, and Shawn R McCaslin. Design of optimal minimum-phase digital fir filters using discrete hilbert transforms. *IEEE Transactions on Signal processing*, 48(5):1491–1495, 2000.

[163] Peter J Rousseeuw, Ida Ruts, and John W Tukey. The bagplot: a bivariate boxplot. *The American Statistician*, 53(4):382–387, 1999.

[164] Sabine Verboven and Mia Hubert. Matlab library libra. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):509–515, 2010.

[165] Adrian Furnham and Hua Chu Boo. A literature review of the anchoring effect. *The Journal of Socio-Economics*, 40(1):35–42, 2011. ISSN 10535357. doi: 10.1016/j.socec.2010.10.008.

[166] Mark Cartwright, Bryan Pardo, and Josh Reiss. Mixploration: Rethinking the audio mixer interface. In *Proceedings of the 19th international conference on Intelligent User Interfaces*, pages 365–370. ACM, 2014.

[167] 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.

[168] Alex Case. *Mix smart: Professional techniques for the home studio*. Taylor & Francis, 2012.

[169] Josh Mycroft, Joshua D Reiss, and Tony Stockman. The influence of graphical user interface design on critical listening skills. *Proceedings of the Sound and Music Computing Conference*, 2013.

[170] Steven Gelineck and Stefania Serafin. A quantitative evaluation of the differences between knobs and sliders. In *NIME*, pages 13–18, 2009.

[171] Duncan Williams. On the affective potential of the recorded voice. *Journal of the Audio Engineering Society*, 64(6):429–437, 2016.

[172] Mervin E Muller. A note on a method for generating points uniformly on n-dimensional spheres. *Communications of the ACM*, 2(4):19–20, 1959.

[173] George Marsaglia. Choosing a point from the surface of a sphere. *The Annals of Mathematical Statistics*, 43(2):645–646, 1972.

[174] Nicholas I Fisher. *Statistical analysis of circular data*. Cambridge University Press, 1995.

[175] Kanti V Mardia and Peter E Jupp. *Directional statistics*, volume 494. John Wiley & Sons, 2009.

[176] Yu-Hui Chen, Dennis Wei, Gregory Newstadt, Marc DeGraef, Jeffrey Simmons, and Alfred Hero. Statistical estimation and clustering of group-invariant orientation parameters. In *Information Fusion (Fusion), 2015 18th International Conference on*, pages 719–726. IEEE, 2015.

[177] Benjamin Friedlander and Boaz Porat. The modified yule-walker method of arma spectral estimation. *IEEE Transactions on Aerospace and Electronic Systems*, AES-20(2):158–173, March 1984. ISSN 0018-9251. doi: 10.1109/TAES.1984.310437.

[178] George Tzanetakis, Georg Essl, and Perry Cook. Human perception and computer extraction of musical beat strength. In *Proceedings of the 5th International Conference on Digital Audio Effects (DAFx-02)*, volume 2, Hamburg, Germany, 2002.

[179] Olivier Lartillot, Tuomas Eerola, Petri Toiviainen, and Jose Fornari. Multi-feature modeling of pulse clarity: Design, validation and optimization. In *9th International Society for Music Information Retrieval Conference (ISMIR 2008)*, pages 521–526. Citeseer, 2008.

[180] Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5):1035–1046, 2005. ISSN 10636676. doi: 10.1109/TSA.2005.851998.

[181] Matthew E P Davies and Mark D Plumbley. Context-dependent beat tracking of musical audio. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3):1009–1020, 2007. ISSN 15587916. doi: 10.1109/TASL.2006.885257.

[182] Jan Schluter and Sebastian Böck. Musical Onset Detection with Convolutional Neural Networks. *International Workshop on Machine Learning and Music (MML)*, pages 1–4, 2013.

[183] Olivier Lartillot, Donato Cereghetti, Kim Eliard, Wiebke J Trost, Marc-André Rappaz, and Didier Grandjean. Estimating tempo and metrical features by tracking the whole metrical hierarchy. In *Proceedings of the 3rd International Conference on Music & Emotion (ICME3), Jyväskylä, Finland, 11th-15th June 2013. Geoff Luck & Olivier Brabant (Eds.). ISBN 978-951-39-5250-1*. University of Jyväskylä, Department of Music, 2013.

[184] Sebastian Bock, Florian Krebs, and Gerhard Widmer. Accurate Tempo Estimation based on Recurrent Neural Networks and Resonating Comb Filters. *Proceedings of the 16th International Society for Music Information Retrieval Conference*, pages 625–631, 2015.

[185] Peter Knees, Ángel Faraldo, Perfecto Herrera, Richard Vogl, Sebastian Böck, Florian Hörschläger, and Mickael Le Goff. Two Datasets for Tempo Estimation and Key Detection in Electronic Dance Music Annotated from User Corrections. In *16th International Society for Music Information Retrieval Conference (ISMIR 2015)*, pages 364–370, 2015.

[186] Florian Hörschlager, Richard Vogl, Sebastian Böck, and Peter Knees. Addressing Tempo Estimation Octave Errors in Electronic Music by Incorporating Style Information Extracted from Wikipedia. *Proceedings of the Sound and Music Computing Conference*, 2015.

[187] Frederic Font and Xavier Serra. Tempo Estimation for Music Loops and a Simple Confidence Measure. In *17th International Society for Music Information Retrieval Conference (ISMIR 2016)*, pages 1–7, 2016.

[188] George Tzanetakis, Randy Jones, and Kirk McNally. Stereo Panning Features for Classifying Recording Production Style. In *International Conference on Music Information Retrieval*, 2007.

[189] Bobby Owsinski. *The Mixing Engineer's Handbook*. Delmar, 2013. ISBN 9781285420875.

[190] Mike Senior. *Mixing secrets for the small studio*. Taylor & Francis, 2011.

[191] Mike Senior. *Recording Secrets for the Small Studio*. CRC Press, 2014.

[192] Richard King, Brett Leonard, and Grzegorz Sikora. Variance in level preference of balance engineers: A study of mixing preference and variance over time. In *Audio Engineering Society Convention 129*, San Francisco, USA, Nov 2010. Audio Engineering Society.

[193] Richard King, Brett Leonard, and Grzegorz Sikora. Consistency of balance preferences in three musical genres. In *Audio Engineering Society Convention 133*, San Francisco, USA, October 2012. Audio Engineering Society.

[194] Alex Wilson and Bruno M Fazenda. 101 mixes: a statistical analysis of mix-variation in a dataset of multitrack music mixes. In *Audio Engineering Society Convention 139*, pages 1–10, New York, New York, USA, 2015.

[195] Alex Wilson and Bruno Fazenda. Variation in multitrack mixes: Analysis of low-level audio signal features. *Journal of the Audio Engineering Society*, 64(7/8):466–473, 2016.

[196] Alex Wilson and Bruno Fazenda. Relationship between hedonic preference and audio quality in tests of music production quality. In *Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2016.

[197] Chung Lee, Andrew Horner, and Bin Wu. The effect of mp3 compression on the timbre space of sustained musical instrument tones. *Journal of the Audio Engineering Society*, 61 (11):840–849, 2013.

[198] Henry F Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200, 1958.

[199] Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611, 1965.

[200] Nornadiah Mohd Razali and Yap Bee Wah. Power comparisons of Shapiro-Wilk , Kolmogorov-Smirnov , Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, 2(1):21–33, 2011. ISSN 9789673631575.

[201] Tatiana Benaglia, Didier Chauveau, David R Hunter, and Derek Young. mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1–29, 2009.

[202] Jacob Cohen. *Statistical power analysis for the behavioural sciences. Hillside*. NJ: Lawrence Earlbaum Associates, 1988.

[203] Jeremy Miles and Mark Shevlin. *Applying regression and correlation: A guide for students and researchers*. Sage, 2001.

[204] Simon Zagorski-Thomas. Sonic signatures — the social construction of technological systems, distributed creativity and sonic cartoons, 2014. Sonic Signatures Symposium.

[205] Austin Moore, Rupert Till, and Jonathan Wakefield. An investigation into the sonic signature of three classic dynamic range compressors. In *Audio Engineering Society Convention 140*, Paris, France, 2016. Audio Engineering Society.

[206] Eamonn Keogh and Abdullah Mueen. Curse of dimensionality. In *Encyclopedia of Machine Learning*, pages 257–258. Springer, 2011.

[207] Gregor Leban, Blaž Zupan, Gaj Vidmar, and Ivan Bratko. VizRank: Data Visualization Guided by Machine Learning. *Data Mining and Knowledge Discovery*, 13(2):119–136, may 2006. ISSN 1384-5810. doi: 10.1007/s10618-005-0031-5.

[208] Igor Kononenko, Edvard Šimec, and Marko Robnik-Šikonja. Overcoming the myopia of inductive learning algorithms with relieff. *Applied Intelligence*, 7(1):39–55, 1997.

[209] Prasanta Chandra Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55, 1936.

[210] Chengjun Liu and Harry Wechsler. Evolutionary pursuit and its application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(6):570–582, 2000.

[211] Augustine Gray and John Markel. A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 22(3):207–217, 1974.

[212] Nicholas Jillings, Brecht De Man, David Moffat, and Joshua D Reiss. Web audio evaluation tool: A browser-based listening test environment. In *Proceedings of the Sound and Music Computing Conference*, 2015.

[213] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

[214] DT Pham, Afshin Ghanbarzadeh, Ebubekir Koc, Sameh Otri, S Rahim, and Muhamed Zaidi. The bees algorithm. Technical note. *Manufacturing Engineering Centre, Cardiff University, UK*, pages 1–57, 2005.

[215] DT Pham, Afshin Ghanbarzadeh, Ebubekir Koc, Sameh Otri, S Rahim, and Muhamed Zaidi. The bees algorithm–a novel tool for complex optimisation. In *Intelligent Production Machines and Systems-2nd I* PROMS Virtual International Conference 3-14 July 2006*, page 454. Elsevier, 2011.

[216] Russ C Eberhart and James Kennedy. A new optimizer using particle swarm theory. In *Proceedings of the 6th international symposium on micro machine and human science*, volume 1, pages 39–43. New York, NY, 1995.

[217] John Biles. Genjam: A genetic algorithm for generating jazz solos. In *Proceedings of the International Computer Music Conference*, pages 131–131. International Computer Music Association, 1994.

[218] James McDermott and Niall J L Griffith. Toward user-directed evolution of sound synthesis parameters. In *Applications of Evolutionary Computing, volume 3449 of LNCS*, pages 517–526. Springer-Verlag, 2005.

[219] Róisín Loughran, James McDermott, and Michael O'Neill. Grammatical Evolution with Zipf's Law Based Fitness for Melodic Composition. In *Proceedings of the Sound and Music Computing Conference*, Maynooth, Ireland, July 2015.

[220] Agoston E Eiben, P-E Raue, and Z Ruttkay. Genetic algorithms with multi-parent recombination. In *International Conference on Parallel Problem Solving from Nature*, pages 78–87. Springer, 1994.

[221] Joanna Lis and Agoston E Eiben. A multi-sexual genetic algorithm for multiobjective optimization. In *IEEE International Conference on Evolutionary Computation*, pages 59–64. IEEE, 1997.

[222] Matej Kristan, Aleš Leonardis, and Danijel Skočaj. Multivariate online kernel density estimation with gaussian kernels. *Pattern Recognition*, 44(10):2630–2642, 2011.

[223] Matej Kristan and Aleš Leonardis. Online discriminative kernel density estimator with gaussian kernels. *IEEE transactions on cybernetics*, 44(3):355–365, 2014.

[224] Richard A Caruana. Representation and hidden bias: Gray vs. binary coding for genetic algorithms. In *Proceedings of the Fifth International Conference on Machine Learing*, pages 153–161, Ann Arbor, Mich., 1988.

[225] Darrell Whitley. A free lunch proof for gray versus binary encodings. In *Proceedings of the 1st Annual Conference on Genetic and Evolutionary Computation-Volume 1*, pages 726–733. Morgan Kaufmann Publishers Inc., 1999.

[226] Uday K Chakraborty and Cezary Z Janikow. An analysis of gray versus binary encoding in genetic search. *Information Sciences*, 156(3):253–269, 2003.

[227] Pablo Moscato. On evolution, search, optimization, genetic algorithms and martial arts: Towards memetic algorithms. *Caltech concurrent computation program, C3P Report*, 826: 1989, 1989.

[228] William E Hart, Natalio Krasnogor, and James E Smith. *Recent advances in memetic algorithms*, volume 166. Springer Science & Business Media, 2004.

[229] Steven Fenton and Hyunkook Lee. Towards a Perceptual Model Of 'Punch' In Musical Signals. In *Audio Engineering Society Convention 139*, pages 1–10, New York, New York, USA, 2015.

[230] Anca Gog, D. Dumitrescu, and Béat Hirsbrunner. Community detection in complex networks using collaborative evolutionary algorithms. In *Advances in Artificial Life: 9th European Conference, ECAL 2007, Lisbon, Portugal, September 10-14, 2007. Proceedings*, pages 886–894, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-74913-4. doi: 10.1007/978-3-540-74913-4_89.

[231] Jimmy Secretan, Nicholas Beato, David B D'Ambrosio, Adelein Rodriguez, Adam Campbell, Jeremiah T Folsom-Kovarik, and Kenneth O Stanley. Picbreeder: A case study in collaborative evolutionary exploration of design space. *Evolutionary Computation*, 19(3): 373–403, 2011.

[232] Mario García-Valdez, Leonardo Trujillo, Francisco Fernández de Vega, and Gustavo Merelo Guervós, Juan Juliánand Olague. Evospace-interactive: A framework to develop distributed collaborative-interactive evolutionary algorithms for artistic design. In *Evolutionary and Biologically Inspired Music, Sound, Art and Design: Second International Conference, EvoMUSART 2013, Vienna, Austria, April 3-5, 2013. Proceedings*, pages 121–132, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-36955-1. doi: 10.1007/978-3-642-36955-1_11.

[233] John Brooke et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.

[234] Aaron Bangor, Philip T Kortum, and James T Miller. An empirical evaluation of the system usability scale. *International Journal of Human–Computer Interaction*, 24(6):574–594, 2008.