

Towards a better understanding of mix engineering

Brecht De Man

Submitted in partial fulfilment of the requirements
of the Degree of Doctor of Philosophy

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

January 2017

Statement of originality

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

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

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

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

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

Details of collaboration and publications: see Chapter 1: Introduction, Section 1.5.

Signature:

Date:

Abstract

This thesis explores how the study of realistic mixes can expand current knowledge about multitrack music mixing. An essential component of music production, mixing remains an esoteric matter with few established best practices. Research on the topic is challenged by a lack of suitable datasets, and consists primarily of controlled studies focusing on a single type of signal processing. However, considering one of these processes in isolation neglects the multidimensional nature of mixing. For this reason, this work presents an analysis and evaluation of real-life mixes, demonstrating that it is a viable and even necessary approach to learn more about how mixes are created and perceived.

Addressing the need for appropriate data, a database of 600 multitrack audio recordings is introduced, and mixes are produced by skilled engineers for a selection of songs. This corpus is subjectively evaluated by 33 expert listeners, using a new framework tailored to the requirements of comparison of musical signal processing.

By studying the relationship between these assessments and objective audio features, previous results are confirmed or revised, new rules are unearthed, and descriptive terms can be defined. In particular, it is shown that examples of inadequate processing, combined with subjective evaluation, are essential in revealing the impact of mix processes on perception. As a case study, the percept ‘reverberation amount’ is expressed as a function of two objective measures, and a range of acceptable values can be delineated.

To establish the generality of these findings, the experiments are repeated with an expanded set of 180 mixes, assessed by 150 subjects with varying levels of experience from seven different locations in five countries. This largely confirms initial findings, showing few distinguishable trends between groups. Increasing experience of the listener results in a larger proportion of critical and specific statements, and agreement with other experts.

Table of Contents

List of Tables	6
List of Figures	8
Acknowledgements	9
1 Introduction	11
1.1 Research questions	17
1.2 Objectives	19
1.3 Thesis structure	20
1.4 Applications	24
1.5 Related publications by the author	26
1.5.1 Journal articles	26
1.5.2 Conference papers	26
1.5.3 Book chapters	28
1.5.4 Patents	28
2 Knowledge-engineered mixing	29
2.1 System	31
2.1.1 Rule list	32
2.1.2 Measurement modules	33
2.1.3 Processing modules	35
2.2 Perceptual evaluation	52
2.2.1 Participants	52
2.2.2 Apparatus	52
2.2.3 Materials	52
2.2.4 Procedure	55
2.3 Results and discussion	56
2.4 Conclusion	59
3 Data collection	61
3.1 Testbed creation and curation	63
3.1.1 Content	65
3.1.2 Infrastructure	68
3.1.3 Mix creation experiment	70
3.2 Perceptual evaluation of mixing practices	74
3.2.1 Basic principles	74
3.2.2 Interface	75
3.2.3 Listening environment	83
3.2.4 Subject selection and surveys	85
3.2.5 Tools	86

3.2.6 Perceptual evaluation experiment	95
3.3 Conclusion	98
4 Single group analysis	100
4.1 Objective features	100
4.1.1 Features overview	100
4.1.2 Statistical analysis of audio features	103
4.1.3 Workflow statistics	112
4.1.4 Conclusion	114
4.2 Subjective numerical ratings	116
4.2.1 Preference rating	116
4.2.2 Correlation of audio features with preference	118
4.2.3 Correlation of workflow statistics with preference	121
4.2.4 Conclusion	122
4.3 Subjective free-form description	124
4.3.1 Thematic analysis	125
4.3.2 Challenges	129
4.3.3 Conclusion	134
4.4 Real-time attribute elicitation	136
4.4.1 System	136
4.4.2 Term analysis	141
4.4.3 Conclusion	148
5 Multi-group analysis	150
5.1 Experiments	152
5.2 Objective features	157
5.3 Subjective numerical ratings	159
5.3.1 Average rating	159
5.3.2 Self-assessment	159
5.4 Subjective free-form description	161
5.4.1 Praise and criticism	161
5.4.2 Comment focus	162
5.4.3 Agreement	164
5.5 Conclusion	167
6 Conclusion	169
Appendix — Case study: Use and perception of reverb	179
A.1 On reverb	179
A.2 Background	180
A.3 Problem formulation	182
A.4 Comment analysis	185
A.5 Relative Reverb Loudness	186
A.6 Equivalent Impulse Response	187
A.6.1 Process	187
A.6.2 Equivalent Impulse Response analysis and results	189
A.7 Multi-group analysis	190
A.8 Conclusion	193
Bibliography	195

List of Tables

1.1	Overview of systems that automate music production processes	14
2.1	Dynamic range compression rules	37
2.2	Equalisation rules	39
2.3	Spectral descriptors in practical sound engineering literature	40
2.4	Panning rules	50
2.5	Songs used in the perceptual evaluation experiment	53
3.1	Metadata fields per song, track, stem, and mix	67
3.2	Songs used in the experiment	71
3.3	Microphones under test	76
3.4	Subject groups	77
3.5	Existing listening test platforms	87
3.6	Selection of supported listening test formats	94
4.1	List of extracted features	101
4.2	Average values of features per instrument	104
4.3	Average change of feature values per instrument	106
4.4	Number of different individual subgroup types	113
4.5	Number of different multi-instrument subgroup types	114
4.6	Correlation coefficients of extracted features with perception	119
4.7	Spearman's rank correlation coefficient for different kinds of subgroups .	121
4.8	Top 25 most occurring descriptive terms over all comments	128
4.9	Features extracted from the audio before and after processing	139
4.10	Highest ranking terms	143
4.11	The first ten descriptors per processor, ranked by number of entries N_{dk}	143
5.1	Overview of evaluation experiments	153
5.2	Overview of mixed content	155
5.3	Effect of expertise on proportion of negative statements	161
5.4	Effect of expertise on generality of statements	163
6.1	Summary of confirmed, revised (crossed out), and new mixing rules . . .	171
A.1	Overview of studies on perception of reverberation of musical signals .	181
A.2	Logistic regression results	190

List of Figures

1.1	Example of a music production chain	12
2.1	Block diagram of knowledge-engineered automatic mixing system	31
2.2	Activity in function of audio level	34
2.3	Active audio regions highlighted as defined by the hysteresis gate	35
2.4	Dynamic range compressor input-output characteristic	37
2.5	Panning law: -3 dB, equal power sine-law	49
2.6	Transfer function of headphones used for the listening test	53
2.7	Box plot representation of the ratings per song and per system	57
2.8	Confidence intervals of ratings	58
3.1	Current search interface of the testbed	68
3.2	Current browse interface of the testbed	69
3.3	Example of linked data network	70
3.4	The Critical Listening Lab at CIRMMT	84
3.5	Frequency response of the Critical Listening Lab at CIRMMT	85
3.6	Online box and whisker plot	91
3.7	Web Audio Evaluation Tool timeline	93
4.1	Loudness of sources	108
4.2	Loudness of sources per song	109
4.3	Mean Stereo Panning Spectrum	110
4.4	Octave band energies for different instruments	112
4.5	Average octave band energies for total mix	113
4.6	Box plot of ratings per mix engineer	117
4.7	Box plot of ratings per mix engineer including their own assessment	117
4.8	Representation of instrument groups across statements	126
4.9	Representation of processors/features across statements	127
4.10	Proportion of negative, positive, and neutral statements	127
4.11	A schematic representation of the plugin architecture	137
4.12	Graphical user interfaces of the plugins	137
4.13	'Metadata' and 'Load' dialog boxes within the plugins	140
4.14	Generality of descriptor <i>thick</i>	142
4.15	Dendograms showing clustering based on feature space distances	145
4.16	Equalisation curves for two clusters of terms in the dataset	146
4.17	Biplots of the distortion and reverb classes	147
4.18	Vector-space similarity	148
5.1	Room frequency responses	154
5.2	Relative loudness of different sources in Lead Me and In The Meantime	157

5.3	Box plot showing the relative loudness of different sources, per song, for McGill and UCP. The bottom and top of the ‘box’ represent the 25% and 75% percentile, the inner horizontal line indicates the median, and the dashed vertical lines extend from the minimum to the maximum, not including outliers (filled circles), which are higher than the 75% percentile or lower than the 25% percentile by at least 1.5 the interquartile range.	158
5.4	Average rating as a function of level of expertise	159
5.5	Relative number of negative versus positive statements	162
5.6	Relative number of instrument-specific versus general statements	163
5.7	Relative agreement between different levels of expertise	165
5.8	Relative agreement r_{AB} between subjects from different groups	166
A.1	Reverb signal chains	182
A.2	Preference (0.0–1.0) per class: 95% confidence intervals	185
A.3	Relative reverb loudness versus perception	186
A.4	Perception of reverb amount as a function of relative reverb loudness . .	189
A.5	Preference as a function of perceived reverberation amount, across groups	191
A.6	Perceived amount of reverberation for different groups	192

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Josh Reiss, for striking the perfect balance between close guidance and trusting encouragement throughout this project. His generosity in time and knowledge cannot be overstated.

I am most grateful to my second supervisor, Marcus Pearce, and independent assessors, Mark Plumbley and Mark Sandler, for their sound advice at each evaluation of my research progression.

I owe a lot to the amazing mix of collaborators I've been lucky to work with over the course of this work, including Richard King, George Massenburg, Brett Leonard, and Matthew Boerum at McGill University; Kirk McNally at University of Victoria; Ryan Stables, Sean Enderby, Dominic Ward, Matthew Cheshire, and Nicholas Jillings at Birmingham City University; Pedro Pestana at the University of Porto; Alex Wilson and Bruno Fazenda at University of Salford; Mark Cartwright and Bryan Pardo at Northwestern University; Frank Duchêne at PXL University College; Alex Stevenson and Paul Thompson at Leeds Beckett University; Mariana Lopez at Anglia Ruskin University; Steven Fenton at University of Huddersfield; Masahiro Ikeda at Yamaha Corporation; Melissa Dickson at Oxford University; and David Moffat, Zheng Ma, David Ronan, György Fazekas, Mariano Mora-Mcginity, Thomas Wilmering, Mathieu Barthet, Giulio Moro, Chris Cannam, and Matthew White at Queen Mary University of London. In addition, I thank all members of the Centre for Digital Music since 2012, for making it the enjoyable and stimulating research environment that it is, and the staff of The Half Moon, for catering many inspiring discussions.

I am greatly indebted to my various sources of funding, without which I would indeed be greatly indebted. These excellent organisations are Yamaha Corporation, the Audio Engineering Society, Harman International Industries, the Engineering and Physical Sciences Research Council, the Association of British Turkish Academics, and Queen Mary University of London's School of Electronic Engineering and Computer Science.

Extra special thanks go to my parents, my family and my friends, who have all coped wonderfully — even worryingly — with my many periods of physical or mental absence, and never openly questioned my life choices. In particular, the unwavering support, motivation, and faith of my fantastic partner Yasmine, and the joy, drive, and welcome distractions brought by our children Nora and Ada were key factors in the successful and timely completion of this work.

*“A good recording is a combination of
the performance and the mix.
It’s necessary to use the technology to
emphasise certain parts of the score,
just as lighting is used to emphasise
colours on film.”*

JAMES LOCK (1939–2009)
Sound engineer at Decca
Studio Sound, April 1987 issue

Chapter 1

Introduction

The production of recorded and live music, from conception to consumption, consists of several stages of creative and technical processes. Compositions materialise as acoustic vibrations, which are then captured, sculpted, and eternalised or amplified as an electronic signal. Between the performance of the music and the commitment of these signals to the intended medium, the different recorded sources are transformed and merged into one consolidated signal, in a process known as the mix. Figure 1.1 shows a simplified depiction of such a music production chain. Mixing music is itself a complex task that includes dynamically adjusting levels, stereo positions, filter coefficients, dynamic range processing parameters, and effect settings of multiple audio streams [1]. Mix engineers are expected to solve technical issues, such as ensuring the audibility of sources, as well as to make creative choices to implement the musical vision of the artist, producer, or themselves [2]. As there are many viable ways to mix a given song, it may not be possible to compile a single set of rules underpinning this esoteric process [3]. However, some mixes are clearly favoured over others, suggesting there are ‘best practices’ in music production [4].

The democratisation of music technology has allowed musicians to produce music on limited budgets, putting decent results within reach of anyone who has access to a laptop, a microphone, and the abundance of free software on the web [5,6]. Similarly, at the distribution side, musicians can share their own content at very little cost and effort, also due to high availability of cheap technology (compact discs, the internet) and, more recently, the ubiquity of online publishing platforms like SoundCloud, Bandcamp,

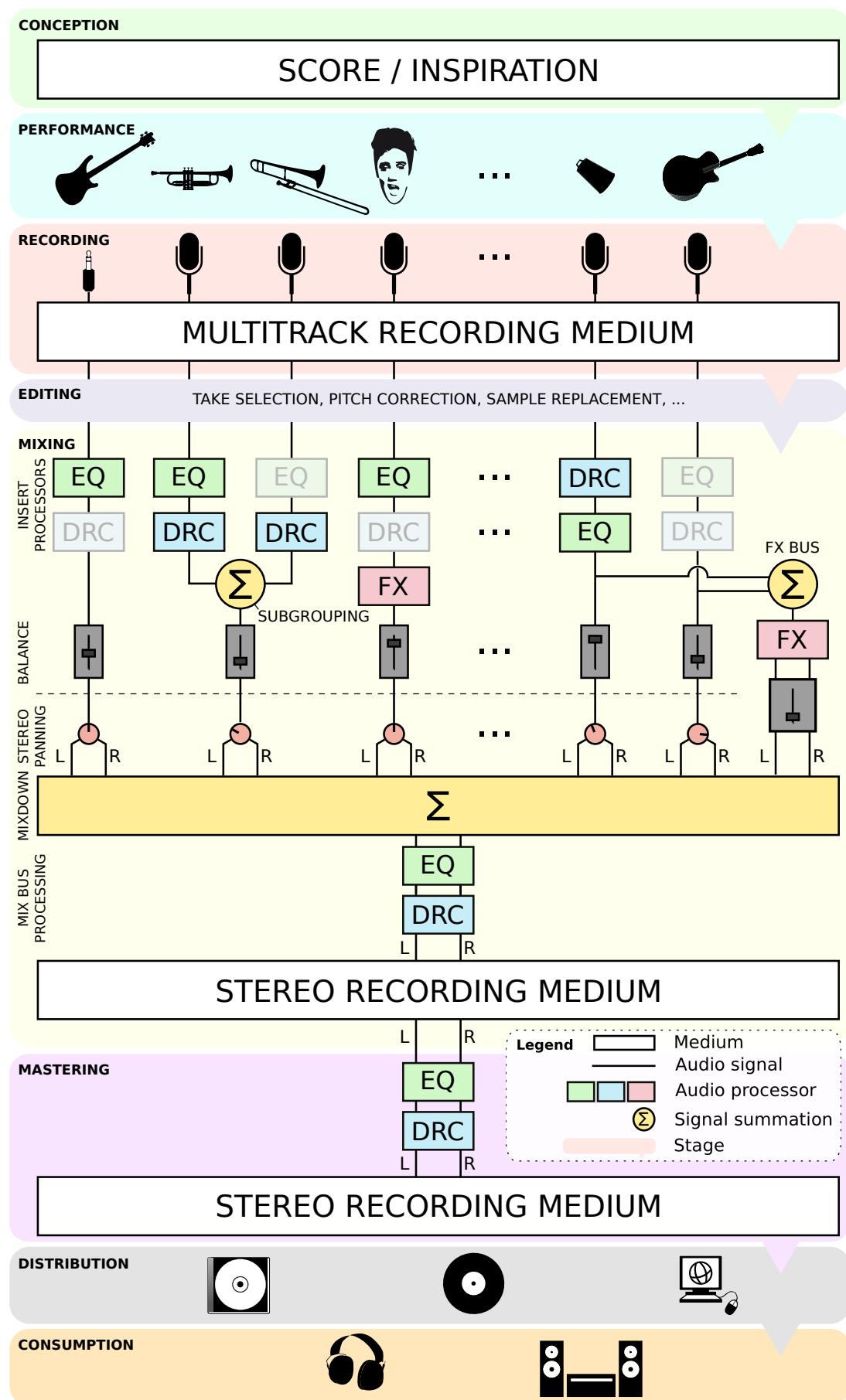


Figure 1.1: Example of a music production chain in the case of a stereo studio recording

and YouTube. Despite this, in order to deliver high quality material a skilled mix engineer is still needed [7]. Raw, recorded tracks almost always require a fair amount of processing before being ready for distribution, such as balancing, panning, equalising (EQ), dynamic range compression (DRC), and artificial reverberation, to name a few. Furthermore, despite the availability of reasonably high quality recording hardware on a budget, an amateur musician or inexperienced recording engineer will almost inevitably cause sonic problems while recording, due to less than perfect microphone placement, an unsuitable recording environment, or simply a poor performance or instrument. Such issues are a challenge to fix post-recording, which only increases the need for an expert mix engineer [8]. In live situations, especially in small venues, the mixing task is particularly demanding and crucial, due to problems such as acoustic feedback, room resonances, and poor equipment. In such cases, however, having a competent operator at the desk unfortunately is the exception rather than the rule. These observations indicate there is a clear need for systems that take care of the mixing stage of music production for live and recording situations. By obtaining a high quality mix quickly and autonomously, home recording becomes more affordable, smaller music venues are freed from the need for expert operators for their front of house and monitor systems, and musicians can increase their productivity and focus on the creative aspects of music production.

Meanwhile, professional audio engineers are often under tremendous pressure to produce high quality content quickly and at little cost [9]. While they may be unlikely to relinquish control entirely to autonomous mix software, assistance with tedious, time-consuming tasks through more powerful, intelligent, responsive, and intuitive algorithms and interfaces is beneficial to pro users as well [6, 10]. Throughout the history of technology, innovations have traditionally been met with resistance and scepticism, in particular from professional users who fear seeing their roles disrupted at best or made obsolete at worst. Music production technology may be especially susceptible to this kind of opposition, as it is notoriously characterised by a tendency towards nostalgia, skeuomorphisms, and analogue workflow, and concerned with aesthetic value in addition to technical excellence and efficiency. However, the evolution of music is inextricably linked to the development of new instruments and tools, and essentially utilitarian inventions such as automatic vocal riding, drum machines, electronic and

electromechanical keyboards, and digital pitch correction have been famously used and abused for creative effect. Already, these advancements have changed the very nature of the sound engineering profession from primarily technical to increasingly expressive. In other words, there is economic, technological, and artistic merit in exploiting the immense computing power and flexibility today’s digital technology affords, to venture away from the rigid structure of the traditional music production toolset.

Recent years have seen a steep increase in research on automatic mixing, where some of the tedious, routine tasks in audio production are automated to the benefit of the inexperienced amateur or hurried professional. Since the first automatic microphone mixer [11], many systems have been proposed to automate various processes, such as balancing levels [12–24], panning signals between channels [25–28], equalisation [29–32], dynamic range compression [33–39], reverberation [40,41], and harmonic distortion [42] (by the author). Other systems seek to mitigate artefacts that are often the result of poor recording practice, such as compensating for comb filtering [43,44], time-varying delays [45,46], popping [47], and interference [48–50] — such goals are not further considered in this work.

Table 1.1: Overview of systems that automate music production processes

	Obj. eval.	Subj. eval.	No eval.
Single track	[22, 31, 32, 38, 40]	[32–34, 41, 42]	[36, 39]
Multitrack	[11–19, 27–30]	[23–27, 29, 35, 37]	[20, 21]

Table 1.1 categorises the above as either systems analysing and processing a single ‘track’ (a stream of monaural or multichannel audio), or those manipulating each track based on features extracted from several tracks. The latter is required to accurately model most mix engineering processes, as each source’s desired level, spatial position, spectrum, and dynamic profile is highly context-dependent. The table further shows which systems have been evaluated objectively, e.g. measuring their performance based on example input using quantitative metrics, or subjectively, e.g. by comparing them to humans or other systems in a formal listening test. Perceptual evaluation validates the concept of the system and its underlying assumptions, and is therefore essential to further our understanding of the mix process.

Studies evaluating novel mixing systems, as well as mixes of human engineers [51,52], have thus far been concerned with a single processor at a time only, automating or

investigating one of the many interdependent tasks. While it is wise to approach a complex problem by tackling one of its components, this limits what can be learned about any processor’s usage and effect on perception in a realistic music production context, where the parameters of different processors on different tracks are ultimately related. For instance, when only faders are available, the level of a particular source may be excessively decreased because it is overly harsh or increased because it is too dull, instead of equalising it accordingly [52]. The measured fader levels may then differ from what they would be if all tools were available. Similarly, one might use a dynamic range compressor to address a perceived imbalance, which might otherwise be achieved by moving faders [35].

The focus of most of these studies is the production of technically correct mixes [53]. To allow the user to specify a creative goal of a mix or the desired effect of a constituent process, the relationship between relevant subjective adjectives and the corresponding objective, actionable audio features and parameters has to be defined. This also constitutes a challenge in gaining knowledge about mixing from sound engineers or listeners, as the translation from their evaluation to measurable quantities is missing. While such subjective terms do not allow accurate communication about sound properties [2, 54], they are prevalent among professionals and amateurs to effectively convey complex concepts. Previous studies have looked at perceptual descriptors (such as *bright*, *punchy*, and *church-like*) and corresponding audio production tool parameters (such as equaliser, compressor, and reverberation settings) [10, 55–64] but, again, these are concerned with the perceived effect of a single processor on an isolated signal. As a consequence, findings of these studies are not necessarily applicable to a multitrack music production context, where several sources are played back simultaneously. This further disregards the possibility that to fully achieve the sonic equivalent of a certain term, more than one type of traditional processors may be needed.

Other high level information, like instrumentation and genre, is also not considered in the above work, even though these are likely to have an impact on customary processing. A preliminary attempt at automatic, instrument-specific processing was made by [65], where a set of (ungrounded) assumptions determined the level, applied equalisation curve, and pan position of three drum tracks.

Finally, the human mixes on which these systems are based, or to which they are compared, are typically produced in lab environments, often by amateur operators, using a restricted and unfamiliar set of processors. While this leads to a high level of control, this data is not necessarily representative of commercial music production. To address this, a few studies have analysed the audio features extracted from a selection of commercially available songs [66–70] or of several realistic mixes of the same songs [71, 72], though without access to the individual tracks or their settings. Others have employed grounded theory, discourse analysis, and related qualitative approaches to describe roles, best practices, and language of sound engineers and related professions [51, 73, 74].

In conclusion, while mix engineering has been the subject of many important works in recent years, knowledge of practices, perception, and preference is still limited. Recurring challenges in this field include a lack of high-quality mixes in a realistic but sufficiently controlled setting, and tackling the inherently high cross-adaptivity and multidimensionality of the mix problem.

1.1 Research questions

The main question underpinning this work is

How can analysis of realistic mixes contribute to understanding of the process of mix engineering?

Prior work is mainly concerned with the emulation of the mix process through lab-based experiments and custom research software, sometimes with unskilled subjects. This maximises control and often allows higher numbers of participants, higher significance, and a more focused answer to the research question. However, the validity, transferability, and relevance of the results may suffer from this artificial context. The hypothesis considered here is that data gathered in a real-life, ecologically valid setting can be used to expand knowledge on mixing practices. While such experiments may be more expensive to organise, or lead to less significant results, they are unencumbered by the inevitable biases of a laboratory setting, and some contexts may allow one to readily collect mix features.

The following questions represent more concrete and tractable parts of this multi-faceted problem.

How can we address the challenges research on mixing is facing?

As discussed above, research on mixing multitrack music constitutes a recent, complex, and multidisciplinary field. Data on mixes and their perception is scarce and hard to produce. Furthermore, the problem of mixing is exceedingly multidimensional, as the perception of any one source is influenced by the sonic characteristics of other, simultaneously playing sources and their processing. Consequently, the various types of processing on the individual elements cannot be studied in isolation.

How can knowledge about mixes be obtained from poor examples?

If it is a challenge to collect a large amount of mixes, it is all but impossible to acquire many examples of a high quality, commercial grade mix. While the latter might indeed make it easier to infer rules about mixing, the cost of producing a sufficient number

of professional quality productions, including per-track settings and features, is simply prohibitive. Therefore, it will be necessary to explore what can be learned from mixed-quality data.

How can it be established how words used to describe sounds or mix processes correspond with objective features or process parameters?

In order to understand mix evaluations, the language used to subjectively evaluate music production practices needs to be translated to objective quantities. Conversely, defining these terms as a function of audio features or processor settings is an essential step towards designing intuitive, high-level metering and control interfaces.

To what extent do differences between sound engineers or listeners limit the generality of findings in music production?

The answers to previous questions may or may not hold across mix engineers or listeners. Even when studying a large number of realistic mixes produced by a group of expert practitioners in a representative setting, findings may be skewed due to that group's background, education, and location. Likewise, a particular group of listeners may have different tastes or expectations from other groups. The impact of background on mix practices, perception, and preference has not yet been assessed.

1.2 Objectives

The purpose of this work is to **develop a methodology to expand our understanding of the mechanics of mixing**. Systems based on the current knowledge and state-of-the-art algorithms will be tested to determine their limitations. The challenges faced by the field of mix engineering are **matched with the necessary tools and experiments**, which are then evaluated with regard to their ability to gather information about mixing tendencies and preference.

Realistic mixes will be produced by skilled engineers in such a way that the natural process of mix engineering is disturbed as little as possible, while still allowing for thorough analysis of all tracks and processes. Results from the analysis of these mixes are compared with findings from previous studies, where settings from a very limited set of tools (e.g. only faders) are considered.

With a large enough set of mixes and extensive perceptual evaluation of each, the **influence of low-level feature values on overall preference is measured**. Additionally, more in-depth assessment such as free-choice profiling will be used to **reveal preference for specific processing of specific instruments**.

Exploring how to make an **abstraction from low-level measures to the high-level terms used to describe musical signal processing**, a body of audio features, processor settings, and associated semantic descriptions is studied.

Finally, the generality of the findings in this work should be examined, by **assessing the influence of the song, genre, background of listener, or background of mix engineer**. To this end, the analysis is repeated using data collected at various sites.

As part of the aim of this work is to explore different approaches and assess their viability, by no means will the potential findings be exhausted. On the contrary, each approach can be utilised with different data and new research questions.

1.3 Thesis structure

Chapter 1 — Introduction

provides background on the topic of mixing, and outlines the intent and structure of the thesis.

Chapter 2 — Knowledge-engineered mixing

explores the potential of a knowledge-engineered approach to mixing, where decisions rely on high-level track metadata combined with best practices sourced from practical sound engineering literature, rather than low-level audio features. This serves to establish the limits of the current presumed knowledge, as well as the performance of the state-of-the-art, signal-dependent, instrument-independent methods from prior work. The results help identify the gaps in knowledge and data, and develop a suitable approach for further research.

Original contributions:

The first full knowledge-engineered mixing system, and perceptual comparison with other systems and humans.

A compiled glossary of terms used to describe spectral properties of sound, and the corresponding frequency ranges.

Chapter 3 — Data collection

discusses the lack of data to analyse in the domain of music production and proposes a solution in the form of a public multitrack audio and metadata repository, from which materials will be used — and via which materials will be shared for the sake of reproducibility and sustainability — throughout the remainder of the work. Several mixes are generated from a diverse selection of this source material, under controlled but ecologically valid circumstances.

It also presents a methodology regarding perceptual audio evaluation of differently processed musical content by skilled listeners, drawing from related literature as well as testing different interfaces within the context of this task, and describes a comparison of the mixes based on these principles.

Original contributions:

A growing Open Multitrack Testbed which addresses the need for large quantities of shareable, thoroughly annotated and diverse multitrack audio, including mixes and parameter settings.

A proposed set of principles for perceptual evaluation in the context of music production, grounded in a large body of literature and validated through use in subsequent chapters.

An open, web-based framework for efficiently designing listening tests.

Chapter 4 — Single group analysis

demonstrates several approaches towards gaining knowledge from the full, representative mixes created in the previous chapter.

The resulting mixes as well as their constituent elements are analysed with regard to low-level audio features. From this data, trends can be identified and variance of certain features is compared on a per-engineer, per-instrument, and per-song basis. Assumptions underpinning automatic mixing systems or observations in earlier literature can thus be confirmed or revised based on real-world data, and new rules are established.

Such features are then studied in relation to subjective ratings of these mixes, revealing which low-level signal characteristics correlate most strongly with preference.

Additional subjective comments are used to zoom in on specific aspects of the mixes and tendencies of the listeners. Opportunities and challenges emerging from the use of this type of unconstrained data are discussed.

Addressing the limited scalability of the presented approach, a system for attribute

elicitation from within a music production environment is proposed, and findings based on initial data are presented.

Original contributions:

Analysis of variance and identification of trends with regard to low-level features extracted from mixes, across different instruments, engineers, and source materials.

An evaluation of low-level features, their discriminatory power, and their correlation with perception in the context of several types of musical signal processing.

A novel effect plugin architecture allowing extensive data collection and collaborative filtering of parameter settings based on sonic descriptors, audio features, and source and user metadata.

A continually expanding dataset of descriptors and their associated parameter settings, absolute and differential feature values, and metadata.

Chapter 5 — Multi-group analysis

expands the study to multiple groups of content producers and listeners, from different countries and educational backgrounds. The influence of these parameters is shown, and some earlier findings are verified based on this larger and more diverse corpus.

Original contributions:

The largest set of mixes with multitrack audio, parameter settings, and subjective evaluations available, totalling 18 songs, 181 mixes, and 4873 evaluations.

A first comparison of music production practices, perception, and preference across groups from different countries and educational backgrounds.

Chapter 6 — Conclusion

offers concluding remarks and future perspectives.

Appendix — Case study: Use and perception of reverb

Combining the results from feature analysis and perceptual evaluation, the concept of a mix parameter space is proven in the context of perceived amount of reverberation, predicting subjective assessment using objective measures.

Original contributions:

Evidence for viewing the act of mixing as a movement within a parameter or feature space, characterised by boundaries corresponding to extremes of the acceptable range of values.

Introduction of a perceptually relevant feature quantifying the perceived reverberation time of a complete mix based on its reverberated and unreverberated components.

Identification of transition region between deficiency and excess of perceived reverberant energy, based on feature extraction and subjective evaluation.

Other contributions

Parameter automation techniques for amplitude distortion, adding the effect to the growing set of automated audio processors [42].

1.4 Applications

Knowledge about the process of mix engineering has many immediate applications, of which some are explored here. They range from completely autonomous automatic mixing systems, to more assistive, workflow-enhancing tools.

As suggested in several previous works, mix tasks could be fully automated so that no sound engineer is required to adjust parameters on a live or studio mix, or to quickly provide a starting point or sound check [14, 75, 76]. As such, a ‘black box’ device would be in control over the whole mix without the need — or option — for user interaction.

Adding control over high-level parameters such as targeted genre or sound shifts the potential of automatic mixing systems from corrective tools that help obtain a single, allegedly ideal mix, to creative tools offering countless possibilities and the user-friendly parameters to achieve them. For instance, an inexperienced user could then produce a mix that evokes a ‘classic rock’ sound, a ‘Tom Elmhirst’ sound, or a ‘1960’ sound. Even within a single processor, extracting relevant features from the audio and adjusting the chosen preset accordingly would represent a dramatic leap over the static presets commonly found in music production software [5].

Intuitive interfaces are likely to speed up music production tasks compared to traditional tools, but also facilitate new ways of working and spur creativity. Already, music software manufacturers are releasing products where the user controls complex processing by adjusting as little as one parameter. In addition, the stronger link between perceptual attributes and signal manipulation can be a significant advantage for educational purposes [77]. New research is needed to validate these relationships, uncover others, and confirm to what extent they hold across different regions and genres.

Intelligent metering constitutes another possible class of systems built on this new information, taking the omnipresent loudness meters, spectral analysers, and goniometers a step further, towards more semantic, mix-level alerts such as ‘reverb amount’, ‘punch’, or ‘muddiness’ [78]. By defining these high-level attributes as a function of measurable quantities, mix diagnostics become more useful and accessible to both experts and laymen. Furthermore, by looking at parameter settings or measured features of mixes

which were rated as either too much or too little of a certain quality, lower and upper bounds of what is perceptually pleasing can be identified. This opens up possibilities for alerts triggered by deviations from what is generally considered acceptable, or at least conventional. Once such perceptually informed issues have been identified, a feedback loop could adjust parameters until the problem is mitigated, for instance turning the reverberator level up or down until high-level attribute ‘reverb amount’ enters a predefined range.

1.5 Related publications by the author

This section lists where work presented in this thesis has previously appeared, with references to corresponding sections of the thesis. Where the author of this thesis is not first author of the publication, a breakdown of the author's contributions is given.

1.5.1 Journal articles

B. De Man, K. McNally, and J. D. Reiss, "Perceptual evaluation and analysis of reverberation in multitrack music production," *Journal of the Audio Engineering Society, Special Issue on Dereverberation and Reverberation of Audio, Music, and Speech*, vol. 65, pp. 108–116, January/February 2017.

Contains Appendix – Case study: Use and perception of reverb.

B. De Man and J. D. Reiss, "Analysis of peer reviews in music production," *Journal of the Art of Record Production*, vol. 10, July 2015.

Contains Chapter 4, Section 4.3 – Subjective free-form description.

Z. Ma, B. De Man, P. D. Pestana, D. A. A. Black, and J. D. Reiss, "Intelligent multi-track dynamic range compression," *Journal of the Audio Engineering Society*, vol. 63, pp. 412–426, June 2015.

The author developed the perceptual evaluation methodology, provided insight on automatic effect design, and edited the text.

B. De Man and J. D. Reiss, "A semantic approach to autonomous mixing," *Journal of the Art of Record Production*, vol. 8, December 2013.

1.5.2 Conference papers

Peer-reviewed

R. Stables, B. De Man, S. Enderby, J. D. Reiss, G. Fazekas, and T. Wilmering, "Semantic description of timbral transformations in music production," in *ACM International Conference on Multimedia*, October 2016.

Contains part of Chapter 4, Section 4.4 – Real-time attribute elicitation.

The author designed and implemented three of the four processors, investigated dataset statistics and the generality of terms, analysed the inter-transform similarity, and contributed to the text.

N. Jillings, B. De Man, D. Moffat, J. D. Reiss, and R. Stables, "Web Audio Evaluation Tool: A framework for subjective assessment of audio," in *2nd Web Audio Conference*, April 2016.

Contains part of Chapter 3, Section 3.2 – Perceptual evaluation of mixing practices.

The author proposed the tool and provided the initial interface design.

N. Jillings, D. Moffat, B. De Man, and J. D. Reiss, “Web Audio Evaluation Tool: A browser-based listening test environment,” in *12th Sound and Music Computing Conference*, July 2015.

As above.

B. De Man, B. Leonard, R. King, and J. D. Reiss, “An analysis and evaluation of audio features for multitrack music mixtures,” in *15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, October 2014.

Contains Chapter 4, Section 4.1 – Objective features.

R. Stables, S. Enderby, B. De Man, G. Fazekas, and J. D. Reiss, “SAFE: A system for the extraction and retrieval of semantic audio descriptors,” in *15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, October 2014.

Contains part of Chapter 4, Section 4.4 – Real-time attribute elicitation.

The author designed and implemented the DSP for three of the four processors, and contributed to the text.

B. De Man and J. D. Reiss, “Adaptive control of amplitude distortion effects,” in *53rd International Conference of the Audio Engineering Society: Semantic Audio*, January 2014.

Extended abstract peer-reviewed

B. De Man and J. D. Reiss, “The Open Multitrack Testbed: Features, content and use cases,” in *2nd AES Workshop on Intelligent Music Production*, September 2016.

Contains part of Chapter 3, Section 3.1 – Testbed creation and curation.

B. De Man, N. Jillings, D. Moffat, J. D. Reiss, and R. Stables, “Subjective comparison of music production practices using the Web Audio Evaluation Tool,” in *2nd AES Workshop on Intelligent Music Production*, September 2016.

Contains part of Chapter 3, Section 3.2 – Perceptual evaluation of mixing practices.

The author wrote the text, designed the tool, and implemented the analysis methods.

D. Ronan, B. De Man, H. Gunes, and J. D. Reiss, “The impact of subgrouping practices on the perception of multitrack mixes,” in *Audio Engineering Society Convention 139*, October 2015.

Contains part of Chapter 4, Section 4.1.3 – Workflow statistics, and Section 4.2.3 – Correlation of workflow statistics with preference.

The author organised the mix creation experiment, conducted the listening tests, processed the data, and edited the text.

B. De Man, M. Boerum, B. Leonard, G. Massenburg, R. King, and J. D. Reiss, “Perceptual evaluation of music mixing practices,” in *Audio Engineering Society Convention 138*, May 2015.

Contains Chapter 4, Section 4.2 – Subjective numerical ratings.

B. De Man, M. Mora-Mcginity, G. Fazekas, and J. D. Reiss, “The Open Multitrack Testbed,” in *Audio Engineering Society Convention 137*, October 2014.

Contains Chapter 3, Section 3.1 – Testbed creation and curation.

B. De Man and J. D. Reiss, “APE: Audio Perceptual Evaluation toolbox for MATLAB,” in *Audio Engineering Society Convention 136*, April 2014.

Contains part of Chapter 3, Section 3.2 – Perceptual evaluation of mixing practices.

B. De Man and J. D. Reiss, “A knowledge-engineered autonomous mixing system,” in *Audio Engineering Society Convention 135*, October 2013.

Corresponds to Chapter 2 – Knowledge-engineered mixing.

B. De Man and J. D. Reiss, “A pairwise and multiple stimuli approach to perceptual evaluation of microphone types,” in *Audio Engineering Society Convention 134*, May 2013.

Contains part of Chapter 3, Section 3.2 – Perceptual evaluation of mixing practices.

1.5.3 Book chapters

B. De Man and J. D. Reiss, “Crowd-sourced learning of music production practices through large-scale perceptual evaluation of mixes,” in *Innovation in Music II* (R. Hepworth-Sawyer, J. Hodgson, J. L. Paterson, and R. Toulson, eds.), Future Technology Press, 2016.

1.5.4 Patents

M. J. Terrell, S. Mansbridge, J. D. Reiss, and B. De Man, “System and method for performing automatic audio production using semantic data,” Mar. 5 2015. US Patent App. 14/471,758.

The knowledge-engineered system presented in Chapter 2 was published and patented, in combination with contributions from other authors.

Chapter 2

Knowledge-engineered mixing

To date, few mixing systems take semantic, high-level information into account. The applied processing is dependent on low-level signal features, but not on the instruments, recording conditions, listener playback conditions, musical genre, or target characteristics. This type of metadata, provided by an end user at little cost, could significantly increase the performance of such a semi-autonomous mixing system. Moreover, combined with instrument and even genre recognition, a fully autonomous mixing system could be designed [24].

An audio effect controlled by high-level features was proposed by [83], and used to compensate for listening conditions during playback [36], but this has not yet been realised within a music production context or as a multitrack implementation. In [65], a rule-based system for setting level, panning, and EQ parameters was proposed, but its assumptions were not backed up by perceptual data or expert knowledge. Listening test participants with varying levels of music production experience preferred the resulting mix over a monaural, unity level sum of the sources just 60% of the time.

Many audio engineering handbooks report standard settings for mixing for various instruments, genres, and desired effects. Some of these ‘rules’ are contradictory and very few have been validated. The very sources containing such mixing rules also state that mixing is highly nonlinear [3] and unpredictable [84], and that there are no hard and fast rules to follow [3], ‘magic’ settings [85], or even effective equaliser presets [84]. It should be noted that spectral and dynamic processing of tracks does indeed depend very much on the characteristics of the input signal [10], as will be

shown later. This work is by no means aiming to disprove that. Rather, it seeks to investigate to what extent suitable mixing decisions can be made based on semantic information about a project and its individual tracks, in combination with elementary low-level features.

As a proof of concept, considered here is an instrument-aware system that creates a stereo mix from raw audio tracks using balance, pan, compression, and equalisation rules derived from practical audio engineering literature [1,3,84–90], which are discussed in the following section. To this end, a framework is presented consisting of modules to read these rules; modules to measure basic, low-level features of audio signals; and modules to carry out elementary mixing tasks based on the rules. Its performance is assessed via a listening test, and compared to another automatic mixing system (not knowledge-based and without track labels) as well as human mix engineers. Thus, the limits of available rules and state-of-the-art mix systems are tested to identify suitable research directions guiding the remainder of this work.

Of interest here is finding the knowledge and logic underpinning the mixing process, which is different in both concept and procedure from designing a system capable of mixing. In the latter case, it would suffice to emulate the skill of a mix engineer, not (necessarily) the knowledge [10]. For the purposes of this work, however, a machine learning approach is beneficial only when it allows one to reverse engineer actionable rules from it.

2.1 System

Figure 2.1 shows a block diagram of the proposed system.

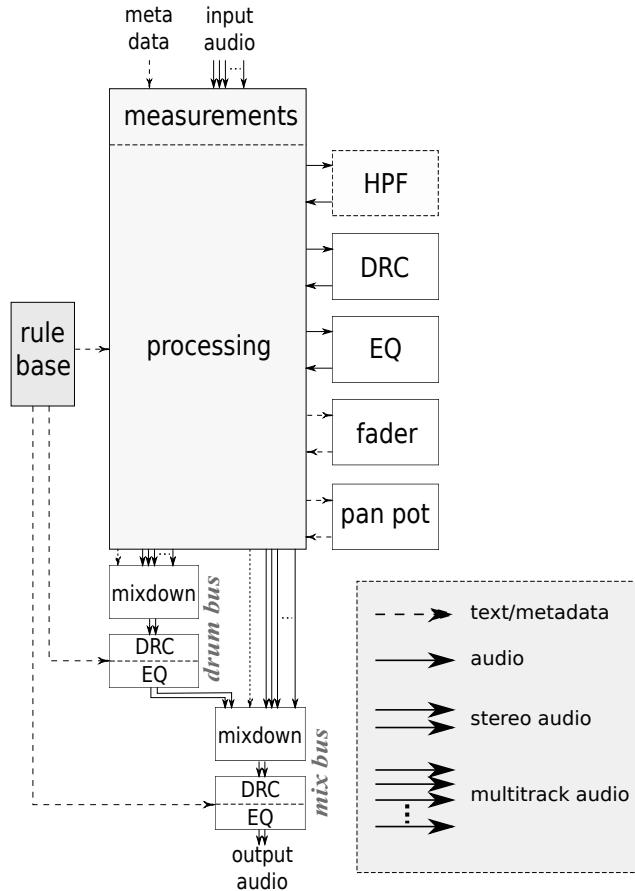


Figure 2.1: Block diagram of the system. Solid arrows represent audio input or output; dashed arrows represent textual information such as instrument names and rules.

The system's input consists of raw, multitrack audio (typically a mixture of mono and stereo tracks), and a text file specifying the instrument corresponding with every audio file (e.g. `Kick_D112.wav: kick drum`). Elementary features of every track are extracted at the measurement stage. For easy access within the system, the track number is automatically stored as an integer scalar or array named after the instrument (e.g. if channel 1 is a kick drum: `kickdrum = 1`, if channels 3 through 5 are toms: `tom = [3, 4, 5]`). The different track indices are also stored in subgroup arrays, e.g. `drums_g = [1, 2, 3, 4, 5, 7, 12]` allows access to all drum instruments at once. Then, rules are read from the rule base and, if applicable, applied to the respective input tracks. The rule specifies one out of five processors: high pass filtering ('HPF'), dynamic range

compression ('DRC'), equalisation ('EQ'), balancing ('fader'), and panning ('pan pot'). The order of the application of the rules is determined by the chosen order of the processors, i.e. first the knowledge base is scanned for rules related to processor 1, then processor 2, and so on.

After processing the individual tracks, the drum instruments (members of subgroup `drums_g`) are mixed down using the respective fader and panning constants, and equalised and compressed if there are rules related to the drum bus. Finally, the stereo drum bus is mixed down together with the remaining tracks, again with their respective fader and panning constants. The resulting mix is equalised and compressed if there are rules pertaining to mix bus processing.

In the current implementation, each extracted feature value and mix parameter is constant over the whole of the audio track. In case longer audio tracks should be processed, one may wish to calculate these features per song section (if sections are marked by the user or automatically), or have measures and settings that vary over time continuously.

2.1.1 Rule list

Each rule in the rule list consists of three parts:

- *tags*: comma-separated words denoting the source of the rule (sources can be included or excluded for comparison purposes), the instrument(s) it should be applied to (or 'generic'), the musical genre(s) it is applicable to (or 'all'), and the processor it concerns. Based on these tags, the inference engine decides if the rule should be applied, and on which tracks. The order and number of tags is arbitrary.
- *rules*: The 'insert' processors (high-pass filter, compressor, and equaliser) replace the audio of the track specified in the *tags* part with a processed version, based on the parameters specified in the *rules* part. This is done immediately upon reading the rule. The level and pan metadata manipulated by the rules, on the other hand, are not applied until the mixdown stage (see Section 2.1.3), after all rules have been read.

- *comments:* These are printed in the console to show which rules have been applied, and to facilitate debugging.

An example of a rule is as follows:

```
tags: authorX, kick drum, pop, rock, compressor
rules: ratio = 4.6; knee = 0; atime = 50; rtime = 1000;
threshold = ch{track}.peak - 12.5;
comments: punchy kick drum compression
```

Conversion of the rules to a formal data model and use of the Audio Effects Ontology [91] could facilitate exchanging, editing, and expanding the rule base, and enable use in description logic contexts. This is beyond the scope of the current experiment.

2.1.2 Measurement modules

For every incoming track, the following quantities are measured and added to the track metadata: the number of channels (mono or stereo), RMS level L_{rms} (Equation (2.1)), peak level L_{peak} (Equation (2.2)), crest factor C (Equation (2.3)) and ITU-R BS.1770 loudness [92].

$$L_{rms} = \sqrt{\frac{1}{N} \sum_{n=1}^N |x(n)|^2} \quad (2.1)$$

$$L_{peak} = \max(x) \quad (2.2)$$

$$C = L_{peak}/L_{rms} \quad (2.3)$$

with x the amplitude vector representing the mono audio file associated with the track.

For a stereo track $x = [x_L, x_R]$, these equations become:

$$\begin{aligned} L_{rms} &= \frac{\sqrt{\frac{1}{N} \sum_{n=1}^N |x_L(n)|^2} + \sqrt{\frac{1}{N} \sum_{n=1}^N |x_R(n)|^2}}{2} \\ &= \frac{L_{rms,L} + L_{rms,R}}{2} \end{aligned} \quad (2.4)$$

$$\begin{aligned} L_{peak} &= \max(\max(x_L), \max(x_R)) \\ &= \max(L_{peak,L}, L_{peak,R}) \end{aligned} \quad (2.5)$$

$$C = L_{peak}/L_{rms} \quad (2.6)$$

Additionally, a hysteresis gate or Schmitt trigger (see Figure 2.2) indicates which parts of the track are active:

$$a(n) = \begin{cases} 0, & \text{if } a(n-1) = 1 \text{ and } x(n) \leq L_1 \\ 1, & \text{if } a(n-1) = 0 \text{ and } x(n) > L_2 \\ a(n-1), & \text{otherwise} \end{cases} \quad (2.7)$$

where a is the binary vector indicating whether the track is active, $x(n)$ the track's audio at sample n , L_1 the level threshold when the gate is off (audio is active), L_2 the level threshold when the gate is on (audio is inactive), and $L_1 \leq L_2$. For stereo tracks, x is summed to mono (single channel) and divided by two. The example waveform in Figure 2.3 shows regions where the track is active highlighted in yellow.

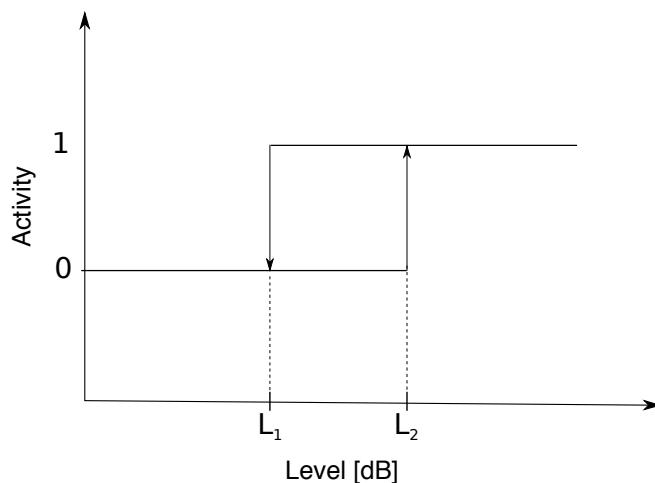


Figure 2.2: Activity in function of audio level (hysteresis gate) following Equation (2.7)

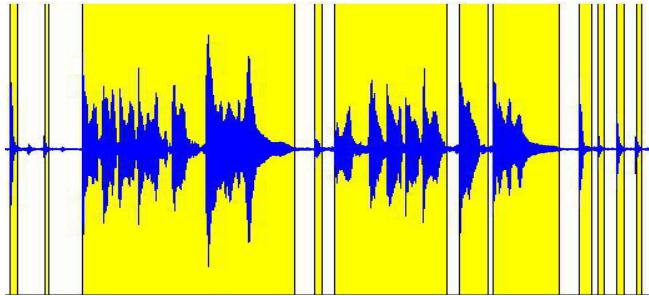


Figure 2.3: Active audio regions highlighted as defined by the hysteresis gate

Based on this definition, the following quantities are also included as metadata: the percentage of time the track is active, and the RMS level, peak level, crest factor, and loudness when active.

These measures can be accessed from within the rules, for instance to set a compression threshold relative to the RMS level. Note that at this point no spectral information is extracted.

2.1.3 Processing modules

Research about the suggested order of processing is ongoing, and most practical literature bases the preferred order on workflow considerations [1, 86]. In some cases, at least one EQ stage is desired before the compressor, because an undesirably heavy low end or a salient frequency triggers the compressor in a way different from the desired effect [1, 3, 84, 93]. In this experiment, the selected audio materials have no such spectral anomalies. Instead, a high-pass filter is placed before the compressor — preventing the compressor from being triggered by unwanted low frequency noise — and an equaliser after the compressor.

It is widely accepted that the faders and pan pots should manipulate the signal after the insert processors such as compressor and equaliser. The pan pots are placed after the faders as this is how mixing consoles are generally wired. Because of the linear nature of these processes and their independence in this system, the order is of no importance in this context.

Based on these considerations, the following order of processors is used for the assessment of this system: high-pass filter, dynamic range compressor, equaliser, fader and

pan pot — as in Figure 2.1.

Time-based effects such as reverb and delay are not incorporated in the current system. There is a notable lack of rules or guidelines with regard to these processors in practical literature, possibly because of the large number of parameters and implementations of such effects, or the absence of established best practices. Interestingly, in contrast with level, panning, EQ, and DRC, no automatic reverberation effects had been developed up until [40].

Dynamic range compression

A generic, downward compressor model is used, with a variable threshold layout (as opposed to for example a fixed threshold, variable input gain design), a quadratic knee and the following, standard parameters: threshold, ratio, attack and release ('ballistics'), and knee width [94], see Figure 2.4.

Make-up gain is not included since the levels are set at a later stage by the 'fader' module, rendering manipulation of the gain at the compressor stage redundant. The compressor processes the incoming audio sample by sample. Stereo files (such as an overhead microphone pair) are compressed in 'stereo link' mode, i.e. the levels of both channels are reduced by an equal amount, rather than independently.

Practical literature lists a considerable number of suggested compressor settings for various instruments and desired effects, see Table 2.1. Rules from different sources are combined when complimentary, averaged when different, and rejected when opposite to what the majority of sources asserts. Presets from Logic Pro 9, a digital audio workstation (DAW), are used to fill in the gaps.

EQ and filtering

A second essential processing step is the equalisation and filtering of the different tracks, or groups of tracks. Two tools take care of this task in the current system: a high pass filter (implementing rules such as "high pass filter with cutoff frequency of 100 Hz on every track but the bass guitar and kick drum") and a parametric equaliser (with high

¹Having access to the tempo (beats per minute) or the number of bars in the processed fragment, the time between snare hits on the backbeat is determined as two beats or half a bar.

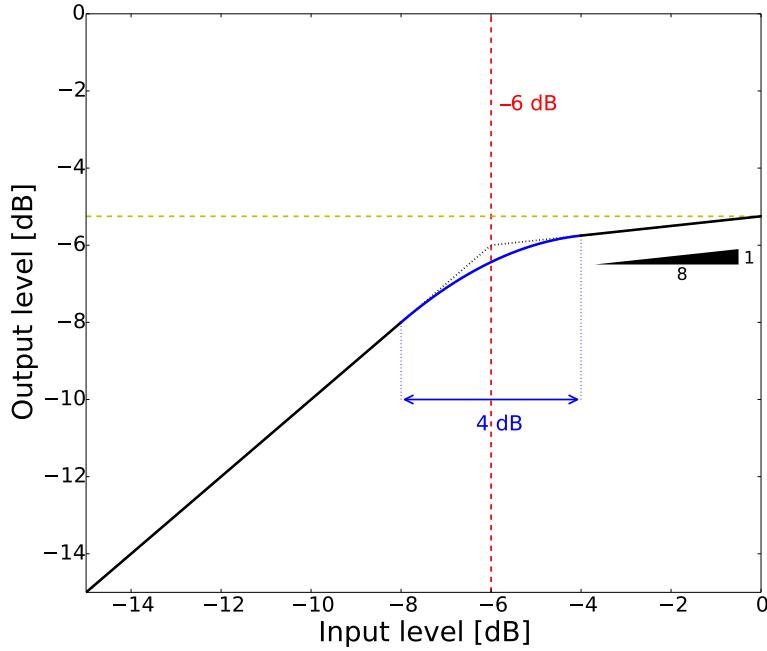


Figure 2.4: Dynamic range compressor input-output characteristic (with quadratic knee). Settings used in this example are: a 8:1 ratio, a -6 dB threshold, and a knee width of 4 dB.

Table 2.1: Dynamic range compression rules

Instrument	Rule	Ref.
Kick drum	5:1–10:1 ratio, hard knee, 5–15 dB reduction (peak), 1–5 ms attack, 200 ms release	[84, 90]
Snare drum	6:1 ratio, slow to medium attack, release until next snare hit ¹	[3, 86]
Drums overhead	4:1–6:1 ratio, 12 dB reduction, 10 ms attack, 20–100 ms release	[1, 89]
Bass guitar	5:1–infinite ratio, 3–4 dB reduction, hard knee, slow to medium attack, medium release	[3, 86, 89]
Acoustic guitar	4:1–8:1 ratio, hard knee, 5–10 dB reduction, attack 10–40 ms, release 100–500 ms	[89]
Distorted guitar	uncompressed	[84, 89]
Electric guitar	8:1 ratio, hard knee, 5–15 dB reduction, attack 2–10 ms, release 500 ms	[89]
Lead vocal	4:1 ratio, 4–6 dB reduction, soft knee (4 dB wide), medium attack and release (500 ms), RMS sensing	[86, 89]
Lead vocal	Clip peaks: infinity ratio, high threshold (low reduction)	[86]
Backing vocal	High (up to 10 dB reduction)	[88]
Lead vocal (rock)	4:1 ratio, 5–15 dB reduction, hard knee, fast attack, 300 ms release, RMS sensing	[86, 89]
Mix bus	2:1–4:1 ratio, 3–6 dB reduction, slow attack, slow release	[3, 88]
Mix bus	Limiter (infinite ratio) at 0.3 dB	[88]

shelving, low shelving, and peak modes). The parameters for the latter are *frequency*, *gain*, and *Q* (quality factor) [95].

Both the high-pass filter (12 dB/octave, as suggested by [84]) and the equaliser (second order filter per stage, i.e. one for every *frequency/gain/Q* triplet) are implemented as a simple biquadratic filter².

Again, practical literature offers a wide range of equalisation advice, and Table 2.2 lists recommended settings which pertain to a specific instrument. However, most of these rules leave a great deal of interpretation to the reader. Usually, an approximate frequency around which the track should be boosted or cut is given, but exact gain and quality factor values are absent. In this case, an estimated value is used for the gain and the quality factor. Unless it is explicitly specified that the cut/boost should be modest or substantial, ± 3 dB is a generic gain value that seemed to work well during informal pilot tests. As sources often suggest to cut/boost a frequency region, such as 1–2 kHz, the quality factor is chosen so that the width of the peak loosely corresponds with the width of this region.

When attempting to translate vague equalising suggestions into quantifiable mix actions, it helps to translate terms like *airy*, *muddy*, and *thump* into frequency ranges. This is possible because many sources provide tables or graphs that define these words in terms of frequencies [1, 86–88, 96–101], see Table 2.3. Due to the subjective nature of these terms, their definitions vary — sometimes within the same book — and are intended as an approximation. In addition, some frequency ranges are derived from figures where the precise lower and upper bounds are unclear. Several sources also suggest that the spectral band to which such a term refers may depend on the instrument in question [1, 88, 100]. In some cases, it needs to be made clear that the term signifies a lack of energy in this band. For instance, *dark* would denote a lack of high frequencies. In other cases, the term is simply associated with a certain frequency range: more or less *edge* depends on more or less energy in the corresponding frequency range. Note that some terms may always be positive, and some always negative, meaning a deficit or an excess of that quality is not possible [102].

²www.musicdsp.org/files/Audio-EQ-Cookbook.txt

Table 2.2: Equalisation rules

Instrument	Rule	Ref.
All tracks	low cut	[1, 3, 86]
Kick drum	cut below 30 Hz	[86]
	40–100 Hz boost	[1, 3, 85, 86]
	100–400 Hz, 6–12 dB cut, $Q > 4$	[3, 84, 88]
	300–600 Hz, 2–6 dB cut, $Q > 4$	[3, 86]
	60–100 Hz, boost	[86]
	1–4 kHz boost	[1, 86]
	10 kHz, shelf cut	[1]
Snare drum	120–240 Hz boost	[1, 3, 85, 86]
	1 kHz boost	[86]
	1–2 kHz cut	[3]
	5 kHz boost	[1, 3, 86, 88]
	10 kHz boost	[86]
Toms	shelf below 60 Hz	[86]
	200 Hz boost	[1]
	200–400 Hz cut, $Q > 4$	[3, 88]
	6 kHz boost	[1, 3, 86, 88]
Drums overhead	1 kHz cut	[86]
	1–5 kHz, <3 dB boost, low Q	[1, 3]
	8–10 kHz, 3–4 dB shelf boost	[1, 3]
Cymbals	cut below 500 Hz	[1]
	12 kHz, 3–6 dB boost	[88]
Bass guitar	cut below 50 Hz	[86]
	400 Hz boost	[84]
	1.5–3 kHz boost	[86, 88]
	5–7 kHz boost	[1, 86]
Acoustic guitar	100–300 Hz cut	[88]
	1–3 kHz cut	[88]
	6–10 kHz boost	[86, 88]
Electric guitar	240–500 Hz boost	[86]
	1 kHz cut	[86]
	1.5–3 kHz boost	[86, 88]
	6–10 kHz boost	[86, 88]
Keyboard	300 Hz cut	[88]
	1 kHz cut	[88]
	3–6 kHz boost	[86, 88]
Lead vocal	cut below 80 Hz	[1]
	250 Hz boost	[3, 86]
	1–6 kHz boost	[3, 86, 88]
	10–12 kHz boost	[3, 86]
Mix bus	80 Hz boost	[3, 85]
	10 kHz boost	[3, 85]

Table 2.3: Spectral descriptors in practical sound engineering literature

Term		Range	Reference
air ³		5–8 kHz	[88, p. 119]
		10–20 kHz	[87, p. 99]
		10–20 kHz	[1, p. 211]
		11–22.5 kHz	[86, p. 26]
		12–15 kHz	[96, p. 103]
		12–16 kHz	[99, p. 43]
		12–20 kHz	[87, p. 25]
		12–20 kHz	[97, p. 108]
anemic	lack of	20–110 Hz	[1, p. 211]
	lack of	40–200 Hz	[88, p. 119]
articulate		800–5000 Hz	[88, p. 119]
ballsy		40–200 Hz	[88, p. 119]
barrelly		200–800 Hz	[88, p. 119]
bathroomy		800–5000 Hz	[88, p. 119]
beefy		40–200 Hz	[88, p. 119]
big		40–250 Hz	[86, p. 25]
bite		2–6 kHz	[97, p. 106]
		2.5 kHz	[100, p. 484]
body		100–500 Hz	[87, p. 99]
		100–500 Hz	[1, p. 211]
		150–600 Hz	[87, p. 24]
		200–800 Hz	[88, p. 119]
		240 Hz	[100, p. 484]
boom(y)		20–100 Hz	[1, p. 211]
		40–200 Hz	[88, p. 119]
		60–250 Hz	[86, p. 25]
		62–125 Hz	[99, p. 43]

³In some books, ‘air’ is also used to denote a part of the audible frequency range, exceeding ‘highs’ [98, p. 86], [97, p. 108].

Table 2.3: Spectral descriptors in practical sound engineering literature (continued)

Term		Range	Reference
		90–175 Hz	[86, p. 26]
		200–240 kHz	[100, p. 484]
bottom		40–100 Hz	[88, p. 119]
		45–90 Hz	[86, p. 26]
		60–120 Hz	[100, p. 484]
		62–300 Hz	[99, p. 43] ⁴
boxy, boxiness		250–800 Hz	[1, p. 211]
		300–600 Hz	[86, p. 31]
		300–900 Hz	[99, p. 43]
		800–5000 Hz	[88, p. 119]
bright		2–12 kHz	[99, p. 43]
		2–20 kHz	[1, p. 211]
		5–8 kHz	[88, p. 119]
brilliant, brilliance		5–8 kHz	[88, p. 119]
		5–11 kHz	[1, p. 211]
		5–20 kHz	[100, p. 484]
		6–16 kHz	[86, p. 25]
brittle		5–20 kHz	[100, p. 484]
		6–20 kHz	[87, p. 25]
cheap	lack of	8–12 kHz	[88, p. 119]
chunky		800–5000 Hz	[88, p. 119]
clarity		2.5–4 kHz	[98, p. 86]
		2.5–5 kHz	[100, p. 484]
		3–12 kHz	[1, p. 211]
		4–16 kHz	[86, p. 26]
clear		5–8 kHz	[88, p. 119]
close		2–4 kHz	[100, p. 484]
		4–6 kHz	[86, p. 25]

⁴More specifically, [99] calls this ‘extended bottom’.

Table 2.3: Spectral descriptors in practical sound engineering literature (continued)

Term		Range	Reference
colour		80–1000 Hz	[1, p. 211]
covered	lack of	800–5000 Hz	[88, p. 119]
crisp, crispness		3–12 kHz	[1, p. 211]
		5–10 kHz	[100, p. 484]
		5–12 kHz	[88, p. 119]
crunch		200–800 Hz	[88, p. 119]
		1400–2800 Hz	[86, p. 26]
cutting		5–8 kHz	[88, p. 119]
dark	lack of	5–8 kHz	[88, p. 119]
dead	lack of	5–8 kHz	[88, p. 119]
definition		2–6 kHz	[97, p. 106]
		2–7 kHz	[1, p. 211]
		6–12 kHz	[86, p. 26]
disembodied		200–800 Hz	[88, p. 119]
distant	lack of	200–800 Hz	[88, p. 119]
	lack of	700–20 000 Hz	[1, p. 211]
	lack of	4–6 kHz	[86, p. 25]
	lack of	5 kHz	[100, p. 484]
dull	lack of	4–20 kHz	[1, p. 211]
	lack of	5–8 kHz	[88, p. 119]
	lack of	6–16 kHz	[99, p. 43]
edge, edgy		800–5000 Hz	[88, p. 119]
		1–8 kHz	[1, p. 211]
		3–6 kHz	[86, p. 26]
		4–8 kHz	[99, p. 43]
fat		50–250 Hz	[1, p. 211]
		60–250 Hz	[86, p. 25]
		62–125 Hz	[99, p. 43]
		200–800 Hz	[88, p. 119]

Table 2.3: Spectral descriptors in practical sound engineering literature (continued)

Term		Range	Reference
		240 Hz	[100, p. 484]
flat	lack of	8–12 kHz	[88, p. 119]
forward		800–5000 Hz	[88, p. 119]
full(ness)		40–200 Hz	[88, p. 119]
		80–240 Hz	[100, p. 484]
		100–500 Hz	[87, p. 99]
		175–350 Hz	[86, p. 26]
		250–350 Hz	[99, p. 43]
glare		8–12 kHz	[88, p. 119]
glassy		8–12 kHz	[88, p. 119]
harsh		2–10 kHz	[1, p. 211]
		2–12 kHz	[99, p. 43]
		5–20 kHz	[100, p. 484]
heavy		40–200 Hz	[88, p. 119]
hollow	lack of	200–800 Hz	[88, p. 119]
honk(y)		350–700 Hz	[86, p. 26]
		400–3000 Hz	[1, p. 211]
		600–1500 Hz	[87, p. 24]
		800–5000 Hz	[88, p. 119]
horn-like		500–1000 Hz	[100, p. 484]
		500–1000 Hz	[86, p. 25]
		800–5000 Hz	[88, p. 119]
impact		62–400 Hz	[99, p. 43]
intelligible		800–5000 Hz	[88, p. 119]
		2–4 kHz	[100, p. 484]
in-your-face		1.5–6 kHz	[87, p. 24]
lisping		2–4 kHz	[86, p. 25]
live		5–8 kHz	[88, p. 119]
loudness		2.5–6 kHz	[1, p. 211]

Table 2.3: Spectral descriptors in practical sound engineering literature (continued)

Term	Range	Reference
	5 kHz	[100, p. 484]
metallic	5–8 kHz	[88, p. 119]
mud(dy)	16–60 Hz	[86, p. 26]
	20–400 Hz	[1, p. 211]
	60–500 Hz	[97, p. 104]
	150–600 Hz	[87, p. 24]
	175–350 Hz	[86, p. 26]
	200–400 Hz	[99, p. 43]
	200–800 Hz	[88, p. 119]
muffled	lack of 800–5000 Hz	[88, p. 119]
nasal	400–2500 Hz	[1, p. 211]
	500–1000 Hz	[97, p. 105]
	700–1200 Hz	[99, p. 43]
	800–5000 Hz	[88, p. 119]
natural tone	80–400 Hz	[1, p. 211]
oomph	150–600 Hz	[87, p. 24]
phonelike	800–5000 Hz	[88, p. 119]
piercing	5–8 kHz	[88, p. 119]
point	1–4 kHz	[86, p. 27]
power(ful)	16–60 Hz	[86, p. 26]
	40–200 Hz	[88, p. 119]
	40–100 Hz	[1, p. 211]
presence	800–12k Hz	[88, p. 119]
	1.5–6 kHz	[87, p. 24]
	2–8 kHz	[99, p. 43]
	2–11 kHz	[1, p. 211]
	2.5–5 kHz	[100, p. 484]
	4–6 kHz	[86, p. 25]
projected	800–5000 Hz	[88, p. 119]

Table 2.3: Spectral descriptors in practical sound engineering literature (continued)

Term	Range	Reference
punch	40–200 Hz	[88, p. 119]
	62–250 Hz	[99, p. 43] ⁵
robustness	200–800 Hz	[88, p. 119]
round	40–200 Hz	[88, p. 119]
rumble	20–100 Hz	[1, p. 211]
	40–200 Hz	[88, p. 119]
screamin'	5–12 kHz	[88, p. 119]
searing	8–12 kHz	[88, p. 119]
sharp	8–12 kHz	[88, p. 119]
shimmer	7.5–12 kHz	[100, p. 484]
shrill	5–7.5 kHz	[100, p. 484]
	5–8 kHz	[88, p. 119]
sibilant, sibilance	2–8 kHz	[1, p. 211]
	2–10 kHz	[99, p. 43]
	4 kHz	[97, p. 120]
	5–20 kHz	[100, p. 484]
	6–12 kHz	[86, p. 26]
	6–16 kHz	[86, p. 25]
sizzle, sizzly	6–20 kHz	[1, p. 211]
	7–12 kHz	[97, p. 107]
	8–12 kHz	[88, p. 119]
slam	62–200 Hz	[99, p. 43]
smooth	5–8 kHz	[88, p. 119]
solid(ity)	35–200 Hz	[1, p. 211]
	40–200 Hz	[88, p. 119]
	62–250 Hz	[99, p. 43]
sparkle, sparkling	5–10 kHz	[86, p. 27]
	5–15 kHz	[1, p. 211]
	5–20 kHz	[100, p. 484]

⁵More specifically, [99] calls this ‘punchy bass’.

Table 2.3: Spectral descriptors in practical sound engineering literature (continued)

Term		Range	Reference
		8–12 kHz	[88, p. 119]
steely		5–8 kHz	[88, p. 119]
strident		5–8 kHz	[88, p. 119]
sub-bass		16–60 Hz	[86, p. 25]
subsonic		0–20 Hz	[1, p. 209]
		0–25 Hz	[98, p. 84]
		10–60 Hz	[97, p. 102]
sweet		250–400 Hz	[99, p. 43]
		250–2000 Hz	[86, p. 25]
thickness		20–500 Hz	[1, p. 211]
		40–200 Hz	[88, p. 119]
		200–750 Hz	[99, p. 43]
thin	lack of	20–200 Hz	[1, p. 211]
	lack of	40–200 Hz	[88, p. 119]
	lack of	60–250 Hz	[86, p. 25]
	lack of	62–600 Hz	[99, p. 43]
thump		40–200 Hz	[88, p. 119]
		90–175 Hz	[86, p. 26]
tinny		1–2 kHz	[100, p. 484]
		1–2 kHz	[86, p. 25]
		5–8 kHz	[88, p. 119]
tone		500–1000 Hz	[97, p. 105]
transparent	lack of	4–6 kHz	[86, p. 25]
tubby		200–800 Hz	[88, p. 119]
veiled	lack of	800–5000 Hz	[88, p. 119]
warm, warmth		90–175 Hz	[86, p. 26]
		100–600 Hz	[1, p. 211]
		200 Hz	[100, p. 484]
		200–800 Hz	[88, p. 119]

Table 2.3: Spectral descriptors in practical sound engineering literature (continued)

Term		Range	Reference
		200–500 Hz	[97, p. 105]
		250–600 Hz	[99, p. 43]
whack		700–1400 Hz	[86, p. 26]
wimpy	lack of	40–200 Hz	[88, p. 119]
	lack of	55–500 Hz	[1, p. 211]
woody		800–5000 Hz	[88, p. 119]
woofy		800–5000 Hz	[88, p. 119]
zing		4–10 kHz	[87, p. 99]
		10–12 kHz	[87, p. 24]
bass/low end/		25–120 Hz	[98, p. 84]
lows		20–150 Hz	[87, p. 23]
		20–250 Hz	[1, p. 209] ⁶
		20–250 Hz	[99, p. 43]
		20–250 Hz	[101, p. 72]
		40–200 Hz	[88, p. 119]
		60–150 Hz	[97, p. 103]
		60–250 Hz	[86, p. 25]
low mids/		120–400 Hz	[98, p. 85]
lower midrange		150–600 Hz	[87, p. 24]
		200–500 Hz	[97, p. 104]
		250–500 Hz	[99, p. 43]
		200–800 Hz	[88, p. 119]
		250–1000 Hz	[101, p. 73]
		250–2000 Hz	[86, p. 25]
		250–2000 Hz	[1, p. 209]
(high) mids/		250–6000 Hz	[99, p. 43] ⁷
upper midrange		350–8000 Hz	[98, p. 85] ⁸

⁶ [1] distinguishes between low bass (20–60 Hz), mid bass (60–120 Hz) and upper bass (120–250 Hz)⁷ [99] distinguishes between lower midrange (250–500 Hz), midrange (250–2000 Hz) and upper midrange (2–6 kHz).⁸ [98] distinguishes between midrange (350–2000 Hz) and upper midrange (2–8 kHz).

Table 2.3: Spectral descriptors in practical sound engineering literature (continued)

Term	Range	Reference
	600–1500 Hz	[87, p. 24]
	800–5000 Hz	[88, p. 119]
	1–10 kHz	[101, p. 73]
	1.5–6 kHz	[87, p. 24]
	2–4 kHz	[86, p. 25]
	2–6 kHz	[1, p. 209]
	2–6 kHz	[97, p. 106]
highs/high end/	5–12 kHz	[88, p. 119] ⁹
treble	6–20 kHz	[1, p. 209]
	6–20 kHz	[87, p. 24]
	6–20 kHz	[99, p. 43] ¹⁰
	7–12 kHz	[97, p. 107]
	8–12 kHz	[98, p. 86]
	10–20 kHz	[101, p. 74]

Panning

The panning value P is stored in the metadata of every track and initially set to zero. The value ranges from -1 (panned completely to the left) to $+1$ (panned completely to the right), and determines the relative gain of the track during mixdown in the left versus the right channel.

Although a variety of panning laws are implemented, here the -3 dB, equal power, sine/cosine panning law is used (see Figure 2.5 — different names can be found in literature), as it is the one that is most commonly used [1].

The gain of the left (g_{Li}) and right channel (g_{Ri}) for track i is then calculated as follows,

⁹ [88] distinguishes between highs (5–8 kHz) and super highs (8–12 kHz)

¹⁰ [99] distinguishes between lower treble or highs (6–12 kHz) and extreme treble (12–20 kHz).

with pan pot value $P \in [-1, 1]$:

$$g_{Li} = \cos\left(\frac{\pi(P+1)}{4}\right) \quad (2.8)$$

$$g_{Ri} = \sin\left(\frac{\pi(P+1)}{4}\right) \quad (2.9)$$

Note that constant power is in fact obtained, regardless of the value of p , as $g_{Li}^2 + g_{Ri}^2 = 1$ (see Figure 2.5).

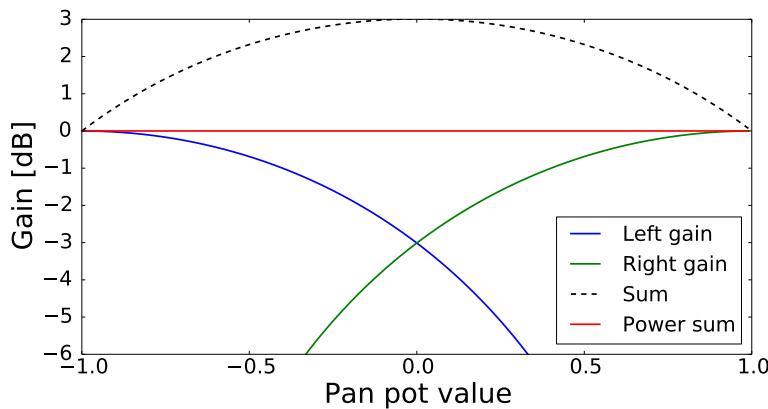


Figure 2.5: Panning law: -3 dB, equal power sine-law

There is a considerable amount of information available in practical literature on ‘standard’ panning for every common instrument, both in the form of exact panning values as well as rules of thumb (e.g. no two instruments at the exact same position [86]). Typically, the pan pot position is described in values ranging from 7:00 (or 7 o’ clock, i.e. fully left) to 17:00 (or 5 o’ clock, i.e. fully right), with 12:00 representing the centre of the stereo image [1]. Sometimes 8:00 and 16:00 are used instead. Rules for particular instruments as found in the considered textbooks are listed in Table 2.4.

Level

As with panning, the ‘level’ parameter is stored as metadata with the instrument track. All tracks have equal loudness initially, and are then brought up where literature suggests a level boost, e.g. lead vocal, or down if it should play a less prominent role, e.g. ambience microphones. The drum bus is regarded as one single instrument. Level adjustments can be specified in absolute or relative terms, i.e. ‘set level at x dB’ or

Table 2.4: Panning rules

Instrument	Rule	Ref.
Kick drum	centre	[1, 88]
Snare drum	same location as in overheads ¹¹	[1, 88]
Toms	at 10:00, 13:00 and 14:00	[1, 87]
Drums overhead	70% wide around centre	[1]
Cymbals	ride 15:00, crashes 9:00 and 14:00	[87]
Hi-hat	14:30–15:30	[87, 88]
Bass guitar	centre	[1]
Guitars	opposite sides if more than one	[1]
Keyboard	spread across stereo image (if stereo and no other harmony instruments)	[88]
Lead vocal	(very slightly off-)centre	[1, 3]
Backing vocal	spread across stereo image	[88]

‘increase/decrease level by x dB’, and are applied during mixdown.

Except for vague guidelines (“every instrument should be audible”, “lead instruments should be roughly x dB louder”), there is very little information available on exact level or loudness values from practical sound engineering literature. A possible reason for this is the arbitrary relationship between the fader level and the resulting loudness of a source, as the latter depends on the initial loudness and the subsequent processing [15]. Whereas a source’s stereo position is solely determined by a pan pot, and its spectrum is rather predictably modified by an equaliser, a fader position is meaningless without information on the source it processes. RMS level channel meters give a skewed view as they overestimate the loudness of low frequencies, and more sophisticated loudness meters are not common on channel strips in hardware or software. Even though balancing is regarded as one of the most basic elements of the mix process, it cannot be characterised by mere parameter settings and engineers are therefore typically unable to quantify their tendencies through self-reflection. Of course, other factors may contribute to the absence of best practices, such as a dependence on song genre and personal taste.

¹¹As a rudimentary solution to the requirement for the snare to approximately match its position in the stereo overhead microphone track, the snare is panned proportionally to the ratio of its correlation coefficient with the left and the right overhead microphone, respectively. In case the snare drum signal is equally correlated with the left and right overhead microphones, it is panned centre; in case it is predominantly in the left or right channel it will be panned accordingly.

Mixdown

The drum bus mixdown (Equations (2.10) and (2.11)) and the total mixdown (Equations (2.12) and (2.13)) then become:

$$d_L = \sum_{i=1}^{N_{drum}} 10^{\frac{L_i}{20}} \cdot g_{Li} \cdot x'_i \quad (2.10)$$

$$d_R = \sum_{i=1}^{N_{drum}} 10^{\frac{L_i}{20}} \cdot g_{Ri} \cdot x'_i \quad (2.11)$$

$$y_L = \sum_{j=1}^{N'} 10^{\frac{L_j}{20}} \cdot g_{Lj} \cdot x'_j + d'_L \quad (2.12)$$

$$y_R = \sum_{j=1}^{N'} 10^{\frac{L_j}{20}} \cdot g_{Rj} \cdot x'_j + d'_R \quad (2.13)$$

with x'_i the processed audio of track i after possible compression and equalisation, $d = [d_L \ d_R]$ the drum submix, N_{drum} the number of drum tracks, d' the processed drum submix after possible drum bus compression and equalisation, $y = [y_L \ y_R]$ the stereo output signal, N' the number of remaining tracks (i.e. non-drum sources), L_i the loudness of track i , and g_{Li} and g_{Ri} the left and right channel gain for track i . Note that after this mixdown stage, y can still be processed by the mix bus compressor and equaliser.

2.2 Perceptual evaluation

The performance of the proof-of-concept system described above was assessed through a listening test, where its output was compared to mixes by two human mix engineers; a plain, monaural sum of the normalised input audio; and a completely automatic mix by processors based on existing automatic mixing algorithms.

2.2.1 Participants

Of the 15 subjects who participated in the listening experiment, 7 had at least some practical audio engineering experience (mixing or recording). Two thirds of the subjects were male. All had previously participated in listening tests, and played musical instruments for at least five years — although neither of these were prerequisites to take part.

2.2.2 Apparatus

The listening tests in this chapter were carried out using the APE tool [103], using a multi-stimulus, single-axis rating scale with an optional comment box, and according to the principles put forward in Chapter 3.

The listening tests were conducted in a dedicated, well-isolated listening room, using an Apogee Duet audio interface and closed, circum-aural Beyerdynamic DT 770 PRO headphones (see Figure 2.6 for its transfer function), the most controlled and highest quality listening environment and system available.

2.2.3 Materials

The raw audio tracks are taken from Shaking Through, an online music project by Weathervane Music¹². The five songs used in this experiment (see Table 2.5) ranged from light pop-rock to heavier alternative rock (the author’s assessment). For every song, only one track was selected per instrument (two channels in the case of instruments recorded in stereo), even when multiple recordings of the same instrument were

¹²weathervanemusic.org/shakingthrough

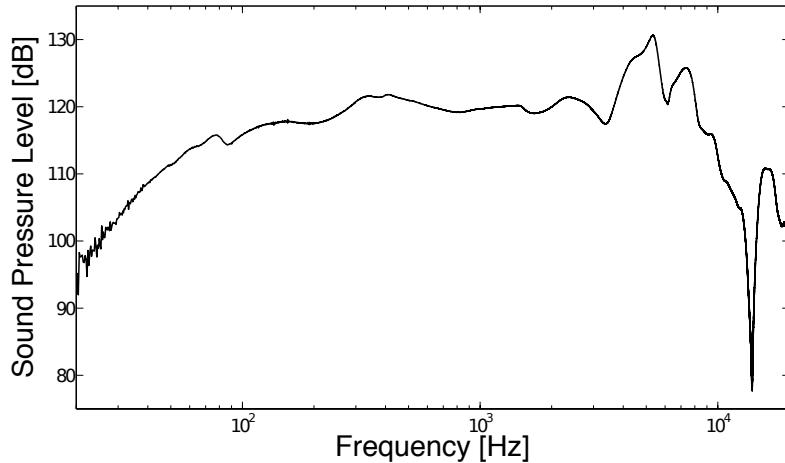


Figure 2.6: Transfer function of the Beyerdynamic DT 770 PRO headphones, as measured using a KEMAR artificial head and sine sweep excitation. It is an average of three left and right channel recordings, and shows the SPL in function of frequency.

available, because of multiple takes or simultaneous recording through different microphones and/or via direct injection.

Table 2.5: Songs used in the perceptual evaluation experiment

Artist	Title
A Classic Education	Night Owl
Auctioneer	Our Future Faces
Ava Luna	Water Duct
Big Troubles	Phantom
Strand Of Oaks	Spacestations

All songs were limited to just four bars to avoid drastic dynamic and spectral variations (since the applied mixing parameters are static in the current implementation, as described above) and to make the perceptual evaluation as well as the manual mixes not too demanding. This resulted in audio files between 11 and 24 seconds. The number of tracks varied from 10 to 22. Every song contained at least vocals, bass, kick drum, snare drum, drum overhead microphones, and one or more harmonic instruments like guitar or keyboards.

The rule-based mix ('KEAMS') was created by feeding these tracks through the system described above and depicted in Figure 2.1. The rule list (Section 2.1.1) consists of the rules given in the previous section, and Logic Pro 9 Channel EQ and Platinum Compressor presets are used to fill any gaps. While the sources leave much to interpretation with regard to specific values, the same set of rules was used throughout the experiment and independent of the song or instrumentation, preserving objectivity.

Mix engineer 1 and 2 ('Pro 1' and 'Pro 2') had professional experience spanning 12 years and 3 years, respectively. In this context, professional experience is defined as the time during which sound engineering is the primary source of income. For maximum comparability with the 'KEAMS' system, they were instructed to limit themselves to using a simple compressor, equaliser, pan pots, and faders, and not to use automation (static settings). They could also process the drum bus and mix bus with a simple compressor and equaliser. No time-based effects like reverb were used, to allow for better comparison with the automatic mixing systems that lack this. Each song was mixed within 45 minutes or less.

A monaural sum of the raw and peak-normalised tracks, 'Sum', served as a type of hidden anchor. However, it is possible that other mixes are perceived to be poor as well, or that the mono sum without processing is an acceptable mix for some songs.

The system consisting of existing automatic mixing algorithms comprised a multitrack compressor [35], equaliser [29], panner [26] and fader [23], and a single-track compressor and master EQ [31] on the drum bus and total mix bus. These processors are implemented in the form of VST (Virtual Studio Technology) effect plugins in Reaper, a DAW capable of accommodating multitrack plugins. Because the mix settings are adjusted during playback (real-time cross-adaptive audio effects), the audio was played back once before rendering the mix to allow the parameters to converge to suitable initial values. Note that this VST system ('VST') is unaware of the functions of the different tracks. It does not know which tracks are part of the drum set, or which are lead and which are background instruments. Instead, it extracts dynamic and spectral information in real-time and modifies the mix parameters based on these values.

The resulting mixes were set at equal loudness, according to the ITU-R BS.1770 loudness standard [92], to remove bias towards louder (or softer) samples during the listening test.

All stimuli are available on www.brechtdeman.com/research.html.

2.2.4 Procedure

Test participants were instructed to rate each ‘version of the same song’ from ‘Bad’ to ‘Excellent’, without any obligation to use the entire scale. The complete task took the subjects 15 minutes 52 seconds on average, with a standard deviation of 4 minutes 51 seconds, and total times ranging from 7 minutes 52 seconds to 26 minutes 34 seconds. The time per song did not depend much on which song was being assessed, but did decrease significantly from one page to the next (from 4 minutes 31 seconds for the first song to 2 minutes 31 seconds for the last song). The measured duration of the first page typically included a brief demonstration of the user interface.

After the test, overall impression and points of focus were determined during an informal chat with each subject.

2.3 Results and discussion

Figure 2.7 shows the ratings for each mixing system, for each song. A few trends are immediately apparent: the monaural sum is generally rated worse than the other mixes, as one would expect, and the fourth song is rated lower than the other songs. Overall, though, consistency among subjects is low, suggesting the task is difficult or subjective preference varies considerably, or both.

Calculation of the confidence intervals of the medians confirms that the normalised sum of the raw audio ('Sum') does perform notably worse than the other mixes. The same is true for the fourth song compared to all other songs. Furthermore, the automatic mix is rated lower than the human mixes and the rule-based system. No significant difference between the rule-based system and the human mix engineers is revealed by this experiment.

Via the interface's text box or during the subsequent conversation, all 15 subjects claimed to partly or entirely judge the different mixes based on the level balance, audibility, or masking of the sources. Examples of these issues include overpowering (backing) vocals, a barely audible lead vocal, and sometimes inaudible instruments like a guitar or a piano. In general, these remarks were caused by the 'Sum', as peak-normalising all sources without any other processing may cause a bad balance, and 'VST', making no distinction between lead and background instruments. It should also be noted mix engineer 'Pro 1' sometimes chose to omit (mute) an instrument as an artistic choice, an option mix engineers often gladly use [3] — more specifically a guitar in Song 4 and a piano in Song 5. This didn't always go unnoticed, although it seemed this was often rewarded in the ratings.

Many (9 out of 15) reported 'spacing', 'location', or 'panning' to be of influence in their ratings, sometimes referring to 'weird panning'. This was found to relate to the 'VST' system which sometimes panned the snare drum or lead vocals considerably to the left or right side, which is unconventional and rarely desired, and sometimes to the monaural 'Sum' where all instruments are 'centred'. The latter was often criticised, although some found this to work well with certain songs.

Other remarks included an overly harsh guitar sound with the 'KEAMS' version of

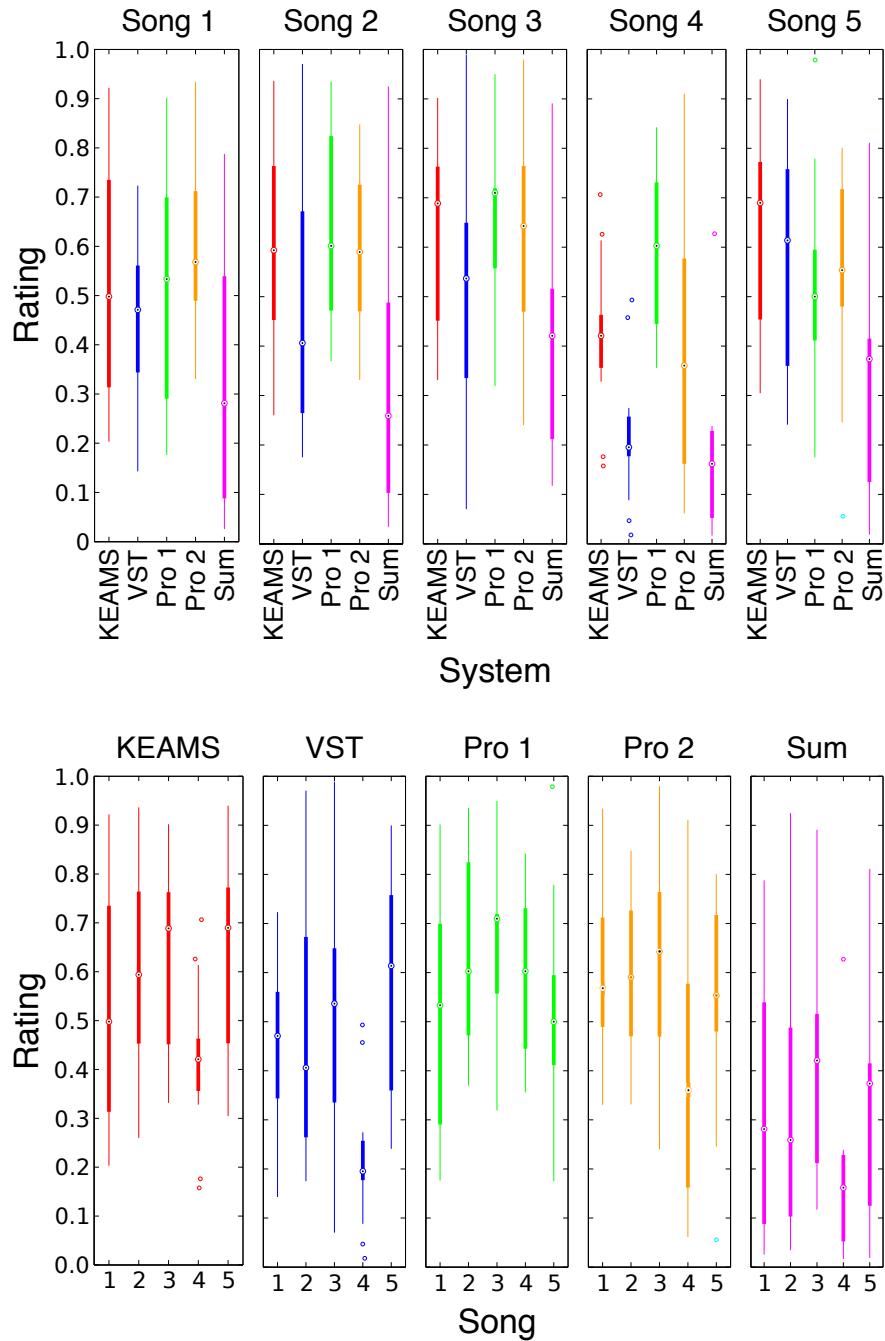


Figure 2.7: Box plot representation of the ratings per song and per system. Following the classic definition of a box and whisker plot, the dot represents the median, the bottom and top of the ‘box’ represent the 25% and 75% percentile, and the vertical lines extend from the minimum to the maximum, not including outliers, which are higher than the 75% percentile or lower than the 25% percentile by at least 1.5 the interquartile range.

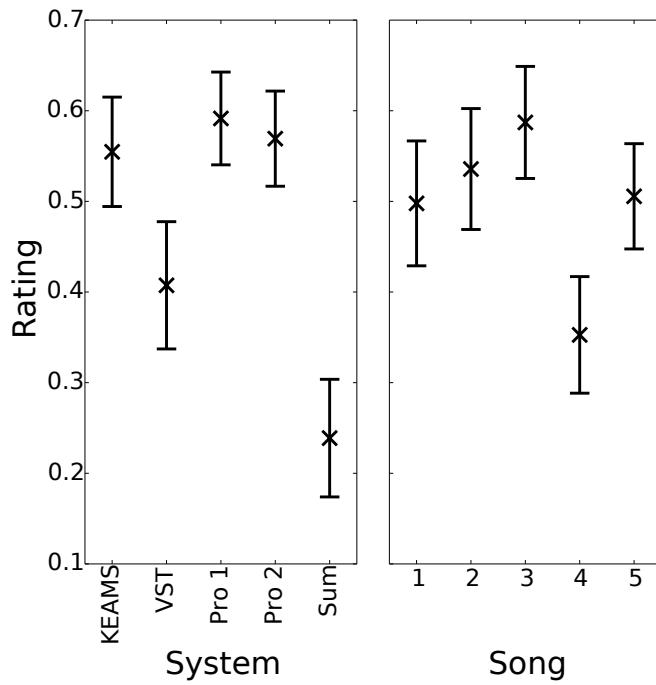


Figure 2.8: Confidence intervals of the median ratings, $p = .05$

Song 4, where default guitar EQ settings are applied to already quite bright guitars; a lack of blend, associated with the lack of reverb; and the absence of context, suggesting preferences may have been different had the fragment been part of a bigger whole. Overall, there seemed to be a tendency to focus on the vocals: 10 out of 15 explicitly mentioned the balance or spatial position of vocals.

2.4 Conclusion

The results of this experiment and the subsequent conversation with the subjects suggest a good performance of the knowledge-engineered system, with no significant difference in subjective preference from human mixes. This suggests incorporation of semantic metadata and rules based on best practices can improve new mixing systems. While the concept is demonstrated and validated here, perceptual motivation (or disproof) of the individual rules found in practical audio engineering literature is still necessary. In particular, knowledge about balance and time-based effects is scarce. A glossary of terms describing spectral properties was constructed from the same literature, but again the definitions have to be confirmed. On a higher level, the developed system proves to be a suitable framework for investigating user preferences of different mixing approaches and settings, as it allows for easy comparison of different sets of rules, different processor implementations and the order of processors. Formalisation of the rule list into a tractable knowledge base would further allow efficient handling in description logic contexts, facilitate the expansion and editing of the rule base, and enable sharing of rule sets.

Even though the knowledge-engineered system presented here uses less sophisticated feature extraction, it outperforms an example of a fully autonomous system based on state-of-the-art technology that does not take semantic information into account. An instrument-agnostic system may position and balance sources differently from what is traditionally expected.

However, an important shortcoming was highlighted during post-experiment discussion with the subjects: the knowledge-engineered system assumes particular spectral and dynamic characteristics, which causes problems when the recorded signals deviate from this. Similarly, while the raw audio tracks used for this test were of high quality, it is doubtful whether the system will perform well when the input audio is poorly recorded or has less conventional dynamic and spectral characteristics. For this reason, the system could likely be improved considerably by expanding the set of measurement modules, to allow for more enhanced listening and processing. The sonic characteristics of the original material need to be measured and taken into account when determining

the processing parameters [10]. This means effectively moving towards a more hybrid system, where semantic rules (processing dependent on high-level information such as instrument tags) and more advanced, cross-adaptive signal processing (processing dependent on signal features of the track itself as well as other tracks) are combined to obtain the highest possible performance.

The relatively low discrimination between different systems suggests that evaluation of different mixes may be challenging or heavily influenced by personal subjective preference. This stresses the importance of careful and rigorous perceptual evaluation practices for assessment of differences in music production, explored in the next chapter.

Some points of focus when listening critically to different mixes of identical source material were apparent, including balance, spatial positioning, and vocals. However, more detailed subjective evaluation of a large number of representative mixes is needed to quantify the attention to certain instruments and sonic attributes, and to ultimately understand what constitutes a good mix. The following chapter also details methods for the acquisition of such data.

Finally, in order to obtain acceptable mixes automatically, time-based effects such as reverberation and delay should be included in the system. Further research is necessary to include a viable autonomous reverb and delay processor, and to establish reverberation rules. Other types of processes, like level balance, need a higher number of more detailed rules as well.

Having assessed the limitations of common audio engineering knowledge and existing automatic mix systems, the following chapters describe an approach to generate and validate rules from real-world mixes, accounting for both high-level information and low-level audio feature measurements, evaluating the impact of several mix aspects on perception and preference, and incorporating time-based effects.

Chapter 3

Data collection

The mixing process is not easily studied in practice. Due to copyright considerations and reluctance to expose the unpolished material, content owners are unlikely to share source content, parameter settings or alternative versions of their music. Even when some mixes are available, extracting data from mix sessions is laborious at best. For this reason, existing research typically employs lab-based mix simulations, which means that its relation to professional mixing practices is uncertain.

This work is therefore based on a controlled experiment wherein realistic, ecologically valid mixes are produced and evaluated. The sessions can be recreated so that any feature and parameter can be extracted for later analysis, and different mixes of the same songs are compared through listening tests to assess the importance and impact of their attributes. As such, both high-level information, including instrument labels and subjective assessments, and low-level measures can be taken into account, as recommended in Chapter 2.

In the first section, the development of an online multitrack repository and associated database and front-end is discussed, and the selection of source material as well as the creation of mixes thereof is documented.

The second section describes a perceptual evaluation methodology developed specifically for the assessment of contrasting music production practices. As shown in Chapter 2, comparison of mixes can be a challenging task with low consistency, so good practices, rigour, and careful design are critical. Based on the proposed principles, two

listening test tools are developed, and a subjective evaluation experiment is conducted and discussed.

This chapter thus defines the parameters of a mix creation and evaluation experiment, the outcome of which is analysed in the next chapter.

3.1 Testbed creation and curation

Many types of audio and music research rely on multitrack audio for analysis, training and testing of models, or demonstration of algorithms. While there is no shortage of mono and stereo recordings of single instruments and ensembles, any work concerned with the study or processing of multitrack audio suffers from a severe lack of relevant material [5]. This limits the generality, relevance, and quality of the research and the designed systems. An important obstacle to the widespread availability of multitrack audio is copyright, which restricts the free sharing of most music and their components. It impedes reproducing or improving on previous studies, when the dataset cannot be made public, and comparing different works, when there is no common dataset used across a wider community.

Among the types of research that require or could benefit from a large number of audio tracks, mixes, and processing parameters, are music production analysis [51], automatic mixing [106], multitrack segmentation [107], and various types of music information retrieval [108, 109]. The availability of this type of data is also useful for budding mix engineers, audio educators, developers, as well as musicians or creative professionals in need of accompanying music or other audio where some tracks can be disabled [110]. Despite this, multitrack audio is scarce. Existing online resources of multitrack audio content have a relatively low number of songs, offer little variation, are restricted due to copyright, provide little to no metadata, lack mixed versions and corresponding parameter settings, or do not come with facilities to search the content for specific criteria.

The Structural Segmentation Multitrack Dataset [107] offers 104 songs including structural segmentation ground truth annotations. The MIXPLORATION Dataset¹ comprises 24 different stem mixes for three songs (four stems per song) [111]. TRIOS is a dataset of five score-aligned multitrack recordings of chamber music trio pieces [112]. BASS-dB is a database of 20 multitracks for evaluation of blind audio source separation, available under Creative Commons licenses [113]. Converse Rubber Tracks² contains royalty-free multitrack audio as well. For [68, 114], already processed stems

¹music.eecs.northwestern.edu/data/mixploration/

²www.conversesamplelibrary.com/

and mixes by a single engineer from the Rock Band video game were used, readily extracted from the game but not shareable as audio only, since its use is restricted by copyright. The Mixing Secrets Free Multitrack Download Library³ corresponding with [84] includes multitracks for about 180 primarily copyrighted songs, where forum users can submit their mixed versions of the song in MP3 format. Other copyrighted but freely available multitracks can be found on MixOff.org⁴, Ultimate Metal Forum⁵, and Telefunken Microphones⁶. Weathervane Music’s Shaking Through⁷, source of the multitrack recordings used in Chapter 2, contains over 50 multitracks with extensive educational materials documenting the recording and mixing process, and also encourages users to upload their own mixes. The content has a Creative Commons license for educational use, but the organisation relies on paid subscriptions and therefore does not allow sharing the source audio. Other paid resources include The Mix Academy⁸, David Glenn Recording⁹, and Dueling Mixes¹⁰. MedleyDB provides 122 royalty-free multitracks — some of them from Shaking Through — including melody annotation [115], to which access can be requested for non-commercial research. Many more multitrack resources cannot realistically be opened up to the public because of copyright restrictions, though some of them allow physical on-site access to researchers [110, 116, 117]. This overview is by no means exhaustive.

To address this need, an open testbed of multitrack material was launched accompanying this work, with a variety of shareable contributions and accompanying metadata necessary for research purposes. In order to be useful to the wider research community, the content should be highly diverse in terms of genre, instrumentation, and technical and artistic quality, so that sufficient data is available for most applications. Where training on large datasets is needed, such as with machine learning applications, a large number of audio samples is especially critical. Furthermore, researchers, journals, conferences, and funding bodies increasingly prefer data to be open, as it facilitates demonstration, reproduction, comparison, and extension of results. A single, widely used, large, and diverse dataset unencumbered by copyright accomplishes this. More-

³www.cambridge-mt.com/ms-mtk.htm

⁴mixoff.org

⁵www.ultimatemetal.com/forum/

⁶www.telefunken-elektroakustik.com/download/multi-track-session.php

⁷weathervanemusic.org/shakingthrough

⁸themixacademy.com

⁹www.davidglenncRecording.com

¹⁰www.duelingmixes.com

over, reliable metadata can serve as a ground truth that is necessary for applications such as instrument identification, where the algorithm’s output needs to be compared to the ‘actual’ instrument. Providing this data makes the testbed an attractive resource for training or testing such algorithms as it obviates the need for manual annotation of the audio, which can be particularly tedious if the number of files becomes large. In addition, for the testbed to be highly usable it is mandatory that the desired type of data can be easily retrieved by filtering or searches pertaining to this metadata. By offering convenient access to a variety of resources, the testbed aims to encourage other researchers and content producers to contribute more material, insofar licenses or ownership allows it.

For this reason, the testbed presented here

- can host a large amount of data;
- supports data of varying type, format, and quality, including raw tracks, stems, mixes, and digital audio workstation (DAW) files;
- contains data under Creative Commons license or similar (including those allowing commercial use);
- offers the possibility to add a wide range of meaningful metadata;
- comes with a semantic database to easily browse, filter, and search based on all metadata fields.

It can be accessed via multitrack.eecs.qmul.ac.uk.

3.1.1 Content

The Open Multitrack Testbed hosts a set of recorded or generated multitrack audio including stems and mixes thereof, without restrictions in terms of type (music, speech, movie soundtrack, game sound, ...), quality (professionally recorded as well as displaying interesting artefacts such as noise, distortion, reverberation, or interference), or number of tracks (from a single multi-microphone recording to a 96-track project with many takes).

In this context, a multitrack audio item, or *song*, is defined as a set of multiple streams

of audio (or *tracks*) which are meant to be played alongside each other. In addition to these *tracks*, some *songs* also contain *mixes* (processed sums of the raw *tracks*) and *stems* (processed sums of a subset of these *tracks*, e.g. only the drum part).

At the time of writing, the Testbed links to close to 600 multitracks, some of which have up to 300 individual constituent tracks from several takes, and others up to 400 mixes of the same source content. Uniquely, it features a number of songs with several mixes including DAW files containing all parameter settings. This content has already been proven useful in a wide range of research projects on various topics, including source separation and remixing using neural networks [118], location-based music selection and mixing [119], assessment of dereverberation methods [120], instrument recognition [121], training and testing machine learning algorithms [40], and evaluating automatic audio effects [29].

Extensive metadata is added manually to every song, track, stem, and mix, see Table 3.1. This allows searching for content that meets a set of specific criteria, sorting the entries based on a particular field, and filtering the displayed results.

Where licenses allow it, multitracks from the resources mentioned above are mirrored to the Testbed. For unclear or less liberal licenses, the metadata is still added to the database, but links point to the respective third party websites. By not imposing a specific license, and due to the variety of for instance Creative Commons licenses available, artists or institutions can share material with different restrictions, including whether or not the material can be used commercially. With the exception of CC0 or public domain material, however, the owner of the content is required to be properly attributed with every use of their work. Content creators who are not under a contract that prohibits them from releasing their intellectual property can benefit from sharing their work on this testbed with a wide community of researchers, students, educators, developers, and creative professionals. Audio shared in this way increases the exposure of the artist and all personnel involved in the production of the music and their affiliation, as this can be included in the metadata corresponding to every song. Furthermore, through dissemination of their work, artists can expect it to be reworked and used in creative applications. In case the owner judges sharing a song would damage record sales, tracks and stems can be shared through this platform while not releasing

Song metadata	Track metadata
Title	File name
Artist	Index
License	Instruments
Type	Number of channels
Composer	Microphone
Recording engineer	Processors
Recording date	Preamplifier
Recording studio	Converter
Location	Sampling rate
Issues	Bit depth
Comments	Duration
Testbed link	Take number
Number of raw audio tracks	
Number of stems	
Number of mixes	

Mix metadata	Stem metadata
Mix engineer	File name
File name DAW session	Index
DAW name	Name
DAW version number	Number of channels
File name mix	Sampling rate
Render format	Bit depth
Sampling rate	
Bit depth	
Duration	

Table 3.1: Metadata fields per song, track, stem, and mix

the final mix. Finally, some performers have chosen to contribute classical recordings anonymously.

3.1.2 Infrastructure



Figure 3.1: Current search interface of the testbed, allowing filtering/searching on various metadata fields

A triplestore database was chosen to store statements containing metadata related to the songs, raw tracks, stems, and mixes¹¹. Semantic databases allow the linking of data by storing subject-predicate-object structured triples [122]. One can then navigate the network formed by linked statements and for instance find more songs of the same artist, engineer, or contributing institution. The implementation features:

- A database which offers a SPARQL endpoint to query and insert data through HTTP requests.
- A REST web service, which receives JSON objects, parses them and stores the different elements in RDF format. These linked elements are then stored in the

¹¹franz.com/graph/allegrograph/

Title	Artist	Composer	Song Type	Recording Engineer	Number of Tracks	Number of Mixes	Number of Stems
Meet the Frownies	Twin Sister	Twin Sister	studio music	Amy Morrissey	34	3	34
Rise Up Singing	Jackie Greene	Jackie Greene	studio music	David Simon Baker	46	78	0
'La Lyra' Overture	King's College London Baroque Orchestra	Georg Philipp Telemann	studio music	Callum McGee	9	3	0
Night Owl	A Classic Education		studio music	Brian McTear	22	4	20
Waiting for GA	Faces on Film	Faces on Film	studio music	Joe Bisirri	15	4	19
Disturbing Wildlife	Invisible Familiars	Invisible Familiars	studio music	Joe Bisirri	28	22	14
Lush	The Tontons	The Tontons	studio music	Matt Schimelfenig	20	29	8
Save Me	Pattern is Movement	Pattern is Movement	studio music	Matt Schimelfenig	28	9	16
Prisoner's Cinema	The Dead Milkmen	The Dead Milkmen	studio music	Matt Schimelfenig	27	23	27
Stranger	Circuit des Yeux	Circuit des Yeux	studio music	Matt Schimelfenig	41	21	23
Green and Yellow	The Dove and the Wolf	The Dove and the Wolf	studio music	Matt Schimelfenig	26	21	27
Perfect Day	Cassandra Jenkins	Cassandra Jenkins	studio music	Matt Schimelfenig	35	29	15
You Always Look for Someone Lost	Peter Matthew Bauer	Peter Matthew Bauer	studio music	Matt Schimelfenig	30	28	29
New Skin	Torres	Mackenzie Scott	studio music	Matt Schimelfenig	21	13	10

Figure 3.2: Current browse interface of the testbed, allowing browsing/sorting on various metadata fields

database.

- A web application offering three functionalities:
 - An interface to insert metadata. Access to this interface is restricted to authorised users.
 - An interface to search for data based on a number of criteria, shown in Figure 3.1. The web application points at the SPARQL endpoint directly, dynamically building SPARQL queries without using a web service. Access to this interface is not restricted, although the data can be.
 - An interface to browse the data, sorted along one of a number of metadata fields, as shown in Figure 3.2.

Figure 3.3 presents a depiction of a scaled-down network formed by the linked data.

Classes are taken from existing ontologies^{12,13}, or extend classes from these ontologies. A ‘Track’, for example, is an instance of the <http://purl.org/ontology/studio/multitrack#AudioTrack> class defined in the Multitrack Ontology [123], from which the *Instrument* class is used as well; a ‘Song’ is an instance of <http://purl.org/ontology/>

¹²musiconontology.com

¹³motools.sourceforge.net/studio/multitrack

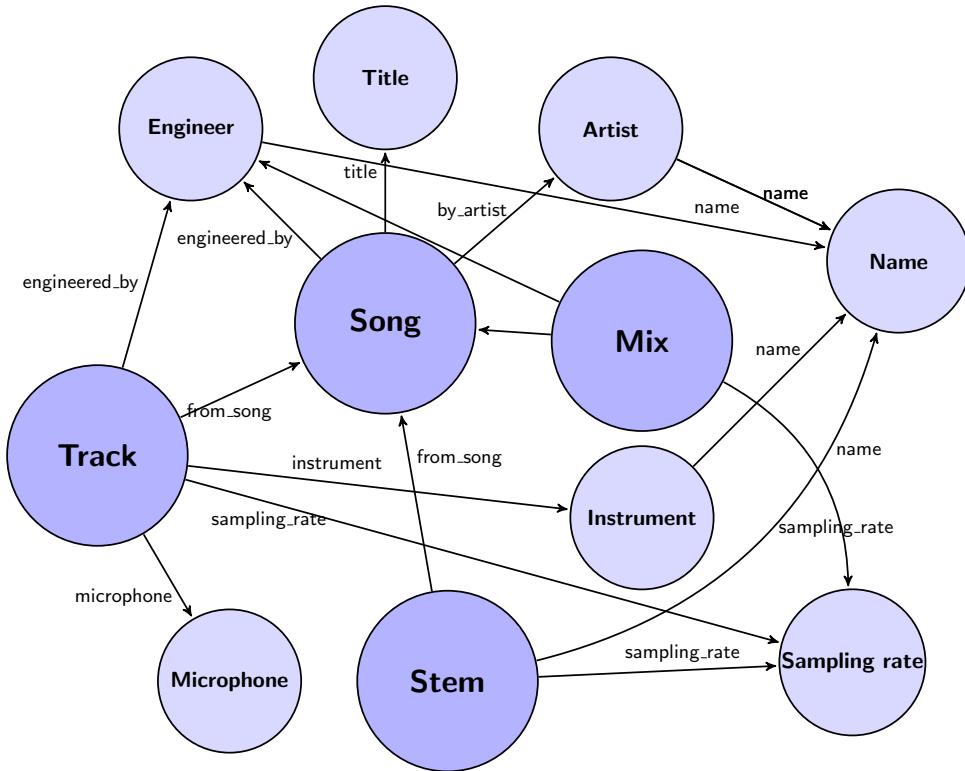


Figure 3.3: Example of linked data network, showing only a subset of the features, with class elements (larger nodes), other elements (smaller nodes), and connections through properties (edge labels)

mo/Composition from the Music Ontology [124], etc. These were extended with the classes ‘Stem’, ‘Mix’, and ‘Engineer’, as well as numerous properties, such as *engineered_by*, *from_song*, *bit_depth*, and *number_of_channels*. Hence, a ‘Track’ X can be *from_song* Y, which has the *name* Z and *by_artist* A, which in turn is a ‘MusicGroup’ with a list of members. The ‘Track’ itself is one of a number of tracks from that ‘Song’, and features an *instrument*, *bit_depth*, and *sampling_rate*, among others.

3.1.3 Mix creation experiment

A selection of the accumulated multitracks was given to skilled sound engineers in order to produce a range of mixes to analyse statistically and evaluate through subjective listening tests. The details of this mix experiment are described below.

Participants

The mix engineers in this experiment were students of the MMus in Sound Recording at the Schulich School of Music, McGill University. All of them were musicians with a Bachelor of Music degree. The average student was 25.4 ± 2.1 years old, with 5.2 ± 2.3 years of audio engineering experience. Of the 24 participants, 5 were female and 19 were male. Three groups of eight were each assigned a different set of songs to mix.

Materials

Table 3.2 lists the songs used in this experiment, along with the number of tracks T (mono and stereo) and the group of students (eight each, denoted by letters) that mixed each particular song. The study spanned two academic years, so first year students in Autumn 2013–Spring 2014 are identical to second year students in Autumn 2014. First year students in Autumn 2014 only mixed one song.

The raw tracks and all mixes (audio and Pro Tools session) of six of these songs are available on the Open Multitrack Testbed under a Creative Commons license (see ‘CC’ column in Table 3.2). The songs were played by professional musicians and recorded by Grammy award-winning recording engineers. The students were assumed to be unfamiliar with the content before the experiment.

Table 3.2: Songs used in the experiment

	Artist	Song	Genre	T	Group	Class	Term	CC
1	The DoneFors	Lead Me	country	23	A–H	1 st year	Autumn 2013	✓
2	Joshua Bell	My Funny Valentine	jazz	17	A–H	1 st year	Autumn 2013	
3	Artist X ¹⁴	Song A ¹⁴	blues	22	I–P	2 nd year	Autumn 2013	
4	Dawn Langstroth	No Prize	jazz	20	I–P	2 nd year	Autumn 2013	
5	Fredy V	Not Alone	soul	24	A–H	1 st year	Spring 2014	✓
6	Broken Crank	Red To Blue	rock	39	A–H	1 st year	Spring 2014	✓
7	Artist Y ¹⁴	Song B ¹⁴	blues	24	I–P	2 nd year	Spring 2014	
8	The DoneFors	Under A Covered Sky	pop	28	I–P	2 nd year	Spring 2014	✓
9	Fredy V	In The Meantime	funk	24	A–H	1 st year	Autumn 2014	✓
10	The DoneFors	Pouring Room	indie	19	Q–X	2 nd year	Autumn 2014	✓

These particular songs were selected in coordination with the programme’s teachers, because they fit the educational goals, were ecologically valid and homogeneous with regard to production quality (having been recorded by one of two Grammy winning recording engineers), and were deemed to represent an adequate spread of genre. Due

¹⁴For two songs permission to disclose artist and song name was not granted.

to the subjective nature of musical genre, a group of subjects were asked to comment on the genres of the songs during the evaluation experiments described below and in Chapter 5, providing a post hoc confirmation of the musical diversity. Each song’s most often occurring genre label was added to the table for reference only. Whereas two songs received the ‘blues’ label, these were considered quite different musically because of the instrumentation (the first has busy acoustic piano, brass, and backing vocal parts, whereas the second doesn’t feature these instruments at all) and tempo (100 BPM vs. 68 BPM). The two ‘jazz’ songs are also very different, as My Funny Valentine features a prominent violin, a harp, and a near-classical arrangement with fluctuating tempo, and No Prize is built on a rhythmic foundation of drums, bass, electric piano, and electric guitar.

Classical, electronic, electro-acoustic, and experimental music are purposely not considered in this work, as the production practices and the role of the mix engineer are substantially different from most pop, rock, folk, and jazz music [24, 102]. In classical music, the mix engineer likely strives for realism, recreating how it would be heard in a live setting [111]. In electronic music, the distinction between the roles of the performer, producer, and sound engineer is less defined.

Following standard practice at the institution where the mixes were created, the source audio’s resolution was maintained throughout the mixing process so that the resulting mixes have a sampling rate of 96 kHz and a bit depth of 24 bit. One exception, where the source material’s sample rate was 88.2 kHz, was printed at 88.2 kHz but later upsampled to 96 kHz using SoX¹⁵ to accommodate an uninterrupted listening test without adjusting the system’s sampling rate.

For comparison, one professional mix per song — often the original released version — was added as well. This allows examining whether the constrained student mixes are representative in terms of production value, and rated similarly during subjective evaluation. Furthermore, an automatic mix akin to the instrument-agnostic ‘VST’ mix in Chapter 2 was evaluated for songs 1 through 8. The only difference in this automatic mix system is that a manually tailored reverb was added to all tracks except bass instrument and kick drum, addressing the lack of time-based effects reported in

¹⁵SoX.sourceforge.net

Chapter 2. The reverb plugin was part of the series used by the students, and for the sake of objectivity the same, static setting was applied to each song.

Procedure

Each student allocated up to six hours to each of their mix assignments, and was allowed to use Avid's Pro Tools 10, its built-in effects, and the Lexicon PCM Native Reverb Plug-In Bundle. The toolset was restricted so that each mix could be faithfully recalled and analysed in depth later, with a limited number of software plugins available. The tools are considered to be ecologically valid as the students were used to using them in their courses. The participants produced the different mixes in their preferred mixing location, so as to achieve a natural and representative spread of environments without a bias imposed by a specific acoustic space, reproduction system, or playback level.

The students were simply tasked with the creation of a stereo mix from the source tracks, within six hours of total mixing time, and were not given any further directions. It was noted that this is unlike most real-life scenarios, where a mix engineer usually receives instructions and feedback from the artist or producer with regard to the desired sound and approach [74]. In this case, however, such artistic direction was not available and it was judged that fabricating any would limit the diversity and spontaneity. Editing the source material, rerecording, the use of samples, or otherwise adding new audio was prohibited, to tighten the scope and ensure maximum comparability between the different mixes. As the mastering process is typically separate from the mixing process to some degree, mix engineers are more often used to delivering mixes which leave room for processing by the mastering engineer. While professionally mixed as well as mastered songs are more representative of average music consumption, the effort required is substantially higher and extra dimensions would be added to the already highly multi-dimensional problem. Consequently, the participants in the presented studies were not asked to master their contributions.

3.2 Perceptual evaluation of mixing practices

For the subjective evaluation of audio engineering practices, central to this work, a suitable approach is needed. An effective methodology helps produce accurate results, with high discrimination, and minimal time and effort. In what follows the measures necessary to accomplish this are investigated.

To this end, the literature on listening test practices is examined, a number of principles are put forward, and tools currently available to conduct such tests are evaluated. Based on these considerations, and because the existing tools do not meet all criteria proposed in this section, a listening test tool has been developed and made available [103,126,127]. Finally, a perceptual evaluation experiment is conducted to assess the aforementioned mixes.

3.2.1 Basic principles

Certain principles are essential to any type of listening test, and supported by all known software tools [129]. All relate to minimising the uncontrolled factors that may cause ambiguity in the test results [130]. While some are a challenge to accommodate in an analogue setting, software-based listening tests fulfil these requirements with relative ease.

First, any information that could distract the subject from the task at hand should be concealed, e.g. any metadata regarding the stimulus [131]. In other words, the participant should be ‘blind’. Furthermore, the experimenter should not have any effect on the subject’s judgement, for instance by giving subconscious cues through facial expressions or body language. This is commonly referred to as the double blind principle, and is easily achieved in the case of a software-based test when the experimenter is outside of the subject’s field of view or even the room.

A subject may also be biased by the presence of other subjects taking the test at the same time. For this reason, it is advised that the test is conducted with one person in the room at a time, so as not to influence each other’s response.

Another potential bias is mitigated by randomising the order in which stimuli are

presented, as well as the order of the pages within a test, and the order of entire test sessions if there are several. This is necessary to avoid uneven amounts of attention to the (sets of) stimuli, and average out any influence of the evaluation sequence, such as subconsciously taking the first auditioned sample as a reference for what follows [130]. In case a limited number of subjects takes part, a pseudo-random test design can ensure an even spread over the different ‘blocks’, e.g. in the case of two sets of stimuli, 50% of the subjects would assess the first set last.

3.2.2 Interface

Multiple stimuli or pairwise

When selecting the appropriate interface, a first important distinction is between single stimulus interfaces, where one stimulus is evaluated at a time; pairwise interfaces, where the subject assesses how two stimuli compare to each other; and multi-stimulus interfaces, where more than two stimuli are presented ‘at the same time’, for the subject to compare in any order.

In case at least two differently processed versions of the same material are presented simultaneously, subjects are likely to focus on the contrasting sonic properties rather than the inherent properties of the source [102]. As this is the desired effect for the purposes of this work, a mix should not be considered in isolation, and the single stimulus approach is ruled out.

Many researchers have previously considered the performance of pairwise versus multi-stimulus tests, and judged that the latter are preferable as long as the number of conditions to be compared is not too large — preferably lower than 12 [132] or 15 [130] — as they are more reliable and discriminating than both pairwise and single stimulus tests [133]. In the case of attribute elicitation, multi-stimulus presentations enlarge the potential pool of descriptors, without the high number of combinations required in the case of pairwise comparison [134].

To evaluate how multi-stimulus tests compare to pairwise evaluation in the context of judging the sonic differences between audio engineering practices, both approaches were assessed for the comparison of microphones. A female singer was recorded using a

selection of six commonly used microphones, see Table 3.3. The human voice was chosen as a source because people are able to discriminate very subtle differences in the sound of the human voice [135]. The microphones were arranged closely together, each at approximately 30 cm from the singer’s mouth, allowing for simultaneous recording and thus minimising variations in timbre and phrasing [136]. Where available, a cardioid pickup pattern was used. The singer performed fragments of Black Velvet, a loud, high-pitched rock song, and No More Blues (Chega de Saudade) a softer, low-pitched jazz song. Two four-second fragments were chosen as stimuli, with the lyrics “Black velvet and that little boy’s smile” and “There’ll be no more blues”, respectively, in part due to the absence of ‘popping’ sounds.

Table 3.3: Microphones under test

	Microphone	Type	Directivity
1	Audio Technica AT2020	condenser	cardioid
2	AKG C414 B-XL II	condenser	cardioid
3	Coles 4038	ribbon	figure-of-eight
4	Shure SM57	dynamic	cardioid
5	Shure Beta 58A	dynamic	hypercardioid
6	Electro-Voice RE-20	dynamic	cardioid

The listening test was conducted in quiet rooms using Beyerdynamic DT 770 PRO headphones, of which the frequency response is shown in Figure 2.6. Each of the 36 listening test participants assessed both sets of stimuli and both types of interfaces. One set of stimuli was evaluated using a pairwise test, where each possible unordered pair of stimuli was presented and the preferred option, or neither, could be selected. In the interest of monitoring subject reliability, this also included pairs of the same microphone, for each microphone. The other set was presented on a multi-stimulus interface where all six stimuli could be arranged freely in order of preference on a single rating axis, the format of which is further described in Section 3.2.2. As in Table 3.4, a randomised block design was followed to control for the order of test types and the order of songs, with four groups of nine subjects each.

To be able to compare the outcome of the two interfaces, a score was attributed to A and B for each possible pair $\{A, B\}$ of microphones, equal to the number of times A was chosen over B by a subject, and vice versa. Microphone 3 was consistently disliked, regardless of the test type. The measured frequency responses show considerably less high frequency energy for this microphone, which may have caused its low score. In

Table 3.4: Subject groups

Group	Test type	Song
1.1	Pairwise	Black Velvet
	Multiple stimuli	No More Blues
1.2	Pairwise	No More Blues
	Multiple stimuli	Black Velvet
2.1	Multiple stimuli	Black Velvet
	Pairwise	No More Blues
2.2	Multiple stimuli	No More Blues
	Pairwise	Black Velvet

the multi-stimulus case, microphone 4 was also significantly preferred over microphone 6. Leaving out those who incorrectly labelled over 50% of different pairs as equal, or equal pairs as different, the remaining 21 subjects additionally preferred microphone 2 over microphone 1 in the multi-stimulus case.

Consistent with earlier findings, the multiple stimuli method was found to have a higher discriminative power, finding more differences between the microphones — between one and three more significantly different pairs, depending on which group of subjects was considered. Multi-stimulus evaluation was also found to be less time-consuming than pairwise evaluation, even with as few as six different stimuli per page. As the multiple stimuli responses here are interpreted as a ranking instead of a continuous rating, the additional advantage of expressing the magnitude of perceived differences is not taken into account here. Clearly, the task was very challenging and preference for certain microphones was not consistent, as very few significant differences were found, and 15 subjects were excluded for incorrectly identifying identical or different pairs of microphones. Further technical detail can be found in the associated paper [125].

Accordingly, only multi-stimulus interfaces are considered in what follows. The number of stimuli should be as high as possible without making the task too tedious (less than 12), as this elicits a richer response [137].

To MUSHRA or not to MUSHRA

The Multi Stimulus test with Hidden Reference and Anchor (MUSHRA) [130] is a well established type of test, originally designed for the assessment of audio codecs, i.e. the evaluation of (audible) distortions and artefacts in a compromised signal. Some of the

defining properties of the associated interface, set forth by Recommendation ITU-R BS.1534-1 (also referred to as the MUSHRA standard), are

- multiple stimuli, at least 4 and up to 15, are presented simultaneously;
- a separate slider per stimulus, with a continuous quality scale marked with adjectives ‘Bad’, ‘Poor’, ‘Fair’, ‘Good’, and ‘Excellent’;
- attributes to be rated can be one or more, but should include ‘basic audio quality’;
- a reference stimulus is provided;
- the reference is also included in the stimuli to be rated, as a ‘hidden reference’; and
- among the stimuli to be rated are one or more low-quality ‘anchors’.

Despite being developed with codec evaluation in mind, MUSHRA-style interfaces have been used for many other purposes, including evaluation of mixes [23,26,138,139]. It has the advantage of being well-known and well-defined, so that it needs little description in the context of research papers, and results from different studies can be compared. However, some question its suitability for other applications, and deviate from the rigorous MUSHRA specification to better address the needs of the research at hand. In this section the argument is made to employ a different test interface for the subjective evaluation experiments in this work.

First and foremost, a ‘reference’ is not always defined [66]. Even commercial mixes by renowned mix engineers prove not to be appropriate reference stimuli as they are not necessarily rated more highly than mixes by student engineers (see Section 4.2). The MUSHRA standard itself specifies that it is not suitable for situations where stimuli might exceed the reference in terms of subjective assessment. It is aimed at rating the attribute ‘quality’, by establishing the detectability and magnitude of impairments of a system relative to a reference, and not to measure the listeners’ preference [140].

In Chapter 2, where different mixes of the same multitrack source were compared, the ‘hidden anchor’ provided was an unprocessed, monaural sum of normalised audio. Without requirement to rate any of the samples below a certain value, the supposedly low quality anchor was not at the bottom of the ratings of some subjects, for some sets

of stimuli. On the other hand, the inclusion of a purposely very low quality sample tends to compress the ratings of the other stimuli, which are pushed to the higher end of the scale, as the large differences the anchor has with other stimuli distract from the meaningful differences among the test samples. An anchor serves to assess the ability of test participants to correctly rate the samples, to enforce a certain usage of the scale (i.e. to ‘anchor’ the scale at one or more positions), and to indicate the dimensions of the defects to identify. As the task at hand is a subjective one, and the salient objective correlates of the subject’s perception are not known, this is not applicable here. A final drawback of including anchors is the increased number of stimuli to assess, raising the complexity and duration of the task.

In the absence of anchors and references, and as listeners may or may not recalibrate their ratings for each page, resulting ratings cannot be compared across pages, though the average rating likely reflects their overall liking of all mixes and the song itself [141].

Furthermore, whereas MUSHRA-style multiple stimuli tests feature a separate slider for rating the quality of each individual sample versus a reference, here the perception between the different stimuli is of interest and no reference is provided. A single rating axis with multiple markers, each of which represent a different stimulus, encourages consideration of the relative placement of the stimuli with respect to the attribute to be rated. Such a ‘drag and drop’ type interface is more accessible than the classic MUSHRA-style interface, particularly to technically experienced subjects [142]. It also offers the possibility of an instantaneous visualisation of the ranking, helping the assessor to check their rating easily, and making the method more intuitive. Ordinal scales (rankings) have been proven to be preferable to interval scales (numerical ratings) [143], further strengthening the case for single-axis interfaces. As stated in the MUSHRA specification itself, albeit as an argument for multi-stimulus as opposed to pairwise comparison, the perceived differences between the stimuli may be lost when each stimulus is compared with the reference only [130]. In conclusion, while a true ‘multi-stimulus’ comparison of test samples, where each stimulus is compared with every other stimulus, is technically possible with MUSHRA even without a reference, it is probable that a subject may then not carefully compare each two similar sounding stimuli.

In the case of a discrete scale, it would for instance be possible for a subject to rate each sample in a test as ‘Good’, providing very little information and therefore requiring a high number of participants to obtain results with high discrimination. A plain ranking interface is not chosen either, as it would prevent learning which stimuli a subject perceives as almost equal, and which are considerably different. Thus, a continuous scale is appropriate for the application at hand, as it allows for rating of very small differences. Tick marks are omitted, to avoid a buildup of ratings at these marks [144].

Because of all of the above, a multi-stimulus, continuous, single-axis interface, without reference, anchors, or tick marks is used throughout the rest of this work. Following the original and adapted MUSHRA scales [130, 142], the scale is divided into five equal intervals with the basic hedonic adjectives ‘Bad’, ‘Poor’, ‘Fair’, ‘Good’, and ‘Excellent’.

Scales

Scale names used in listening tests often appear to have been defined by the experimenter, rather than derived from detailed elicitation experiments, and are therefore not necessarily meaningful or statistically independent of each other [145]. Scales associated with specific, fixed attributes further suffer from several biases, from a ‘dumping bias’ when ‘missing’ attribute scales impact the available scales [146], to a ‘halo bias’ when the simultaneous presentation of scales causes ratings to correlate [129]. Furthermore, the terms used may be understood differently by different people, particularly non-experts [147]. No established set of attributes exists for the evaluation of music production practices, whereas literature on topics like spatial sound includes many studies on the development of an associated lexicon [145, 148, 149]. As such, instead of imposed detailed scales, one general, hedonic scale is used here.

Evaluation of audio involves a combination of hedonic and sensory judgements. Preference is an example of a hedonic judgement, while (basic audio) quality — “the physical nature of an entity with regards to its ability to fulfill predetermined and fixed requirements” [150] — is primarily a sensory judgement [151, 152]. Indeed, preference and perceived quality are not always concurrent [66, 153, 154]: a musical sample of lower

perceived quality, e.g. having digital glitches or a ‘lo-fi’ sound, may still be preferred to other samples which are perceived as ‘clean’, but don’t have the same positive emotional impact. Especially when no reference is given, subjects sometimes prefer a ‘distorted’ version of a sound [135]. In this work, personal preference is deemed a more appropriate attribute than audio quality or fidelity.

This single, hedonic rating can reveal which mixes are preferred over others, and therefore which parameter settings are more desirable, or which can be excluded from analysis. However, it does not convey any detailed information about what aspects of a mix are (dis)liked. Furthermore, subjects tend to be frustrated when they do not have the ability to express their thoughts on a particular attribute [155]. For this reason, free-form text response in the form of comment boxes is accommodated. The results of this ‘free-choice profiling’ also allow learning how subjects used and misused the interface, whereas isolated ratings do not provide any information about the difficulty, focus, or thought process associated with the evaluation task. A final, practical reason for allowing subjects to write comments is that taking notes on shortcomings or strengths of the different mixes helps keep track of which fragment is which, facilitating the complex task at hand. The appropriate sliders and comment boxes are highlighted during playback so that it is clear which stimulus the subject is listening to, as recommended in [130].

As the purpose of these comments surpasses the goal of attribute elicitation, but also aims to evoke detailed descriptions of mix issues or strengths, standard approaches for the creation of semantic scales are not considered in the current work [102, 134, 147, 156]. At this early stage, it is unknown which instruments, processors, or sonic attributes draw most attention, and whether the salient perceptual differences between mixes can be expressed using descriptive terms (e.g. “drums are uneven”) or if more detailed descriptions are typical (e.g. “snare drum is too loud”). For this reason, a maximally free response format is chosen here. Undoubtedly, more focused studies aimed at constructing a vocabulary pertaining to specific processors or instruments will be useful for the successful development of high-level interfaces.

In the experiments described below, participants were able and encouraged to comment on all stimuli using first a single text box with numbers ‘1:’ through ‘10:’ (songs 1–4)

already present, and in later sessions a separate text box per stimulus (songs 5–8). For the same participants ($N = 21$), the percentage of comments on stimuli increased from 82.1% to 96.5%. When two participants who commented significantly less were excluded, the comment rate was even as high as 99.8%. Comments were also 47.2% longer in the case of separate boxes (88.3 rather than 60.0 characters per comment on average). The two tests were near identical otherwise, except for the stimuli.

To minimise the risk of semantic constraints on subjects [157] and elicit the richest possible vocabulary [102], all subjects should be allowed to use their native tongue. This necessitates tedious, high quality translation of the attributes [158], ensured in some cases by several bilingual experts on the topic [149]. However, it is understood that experienced sound engineers studying and working in an English-speaking environment are most comfortable using English terminology and sonic descriptors, regardless of their native language. Therefore, the issue did not present itself in the current experiment.

In conclusion, the preference rating task serves to determine the overall, personal appreciation of the mix, relative to other mixes of the same song. It further forces the subjects to consider which mix they prefer over which, so that they reflect and comment on the aspects that have an impact on their preference.

Visual distractions

A key principle in the design of auditory perception experiments is to keep visual distractions to a minimum. In the context of digital interfaces, this means only having essential elements on the screen, to minimally distract from the task at hand [159], and to avoid the need for scrolling, improving the subjects' accuracy and reaction times [160].

Apart from the necessary rating scale, comment boxes, and navigation buttons, a trade-off needs to be made between the value added and the attention claimed by interface elements like progress indicators, a scrubber bar, and additional explanatory text. For the experiments described here, only a page counter is deemed valuable enough, to allow the subjects to budget their time.

Free switching and time-aligned stimuli

Rather than playing all stimuli in a randomised but fixed sequence, allowing subjects to switch freely between them enhances the ability to perceive more delicate differences [147]. While this is fairly ubiquitous in digital listening test interfaces, some older experiments did not accommodate this.

The comparison of differently processed musical signals is further facilitated by synchronised playback of time-aligned audio samples, and immediate switching between them. This leads to seamless transitions where the relevant sonic characteristics change instantly while the source signal seemingly continues to play, directing the attention to the differences in processing rather than the intrinsic properties of the song. It also avoids excessive focus on the first few seconds of long stimuli, and makes toggling between them more pleasant.

3.2.3 Listening environment

Sound reproduction system

Headphones were not used to avoid sensory discrepancy between vision and hearing, as well as the expected differences in terms of preferred mix attributes between headphone and speaker listening [161]. With the exception of binaural audio, which is not considered here, most sources in stereo music are generally positioned ‘inside the head’ when listening to headphones [162]. While headphones represent an increasingly important portion of music consumption, they are usually regarded as a secondary monitoring system for mixing, when high quality loudspeakers in an acoustically superior space are available. For certain critical auditory tasks, listening over loudspeakers is similar to listening over headphones with regard to accuracy [163].

For this reason, high quality digital-to-analogue converters, amplifiers, and loudspeakers were used as available in the listening room.

Room

An important prerequisite for critical listening is a quiet, high quality, acoustically neutral listening environment [2]. Similar to the listening test interface, visual distractions in the room need to be reduced as well. This can be accomplished in part by dimming the lights, and covering any windows [159].

All listening tests took place in CIRMMT's Critical Listening Lab at McGill University (see Figure 3.4). The frequency response of the listening environment including playback system (left speaker) is shown in Figure 3.5.

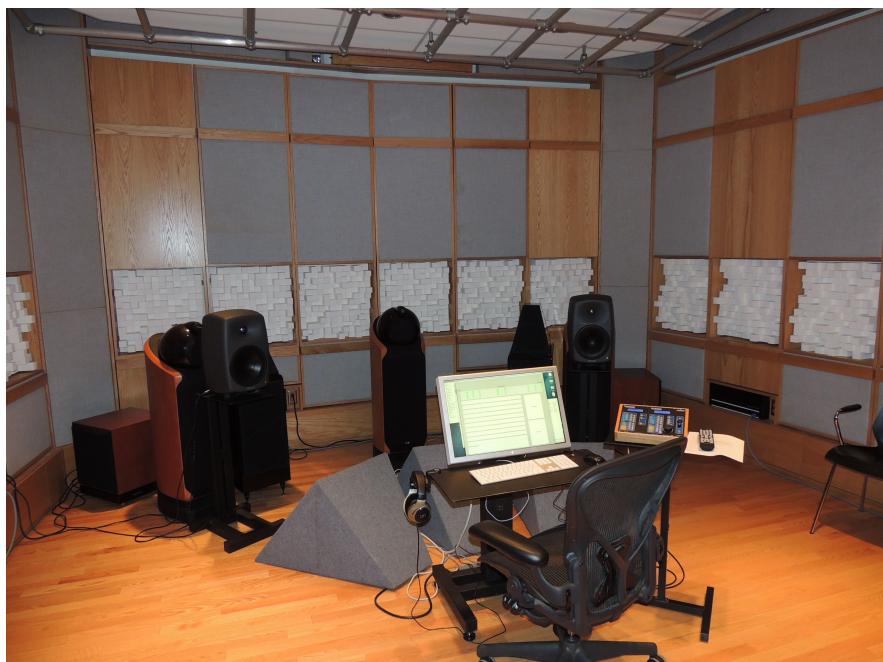


Figure 3.4: The Critical Listening Lab at CIRMMT, where all listening tests took place

Listening level

The playback level of each stimulus was adjusted to have the same integrated ITU-R BS.1770 loudness [92] — a universally accepted principle in listening test design whenever loudness should not have an influence on the rated attributes [129, 164, 165].

Subjects were instructed to first set the listening level as they wished, since their judgements are most relevant when listening at a comfortable and familiar level [166], and since many perceptual features vary with level, e.g. the perceived reverberation amount [167, 168]. Some studies allow only a ± 4 dB deviation from a reference level

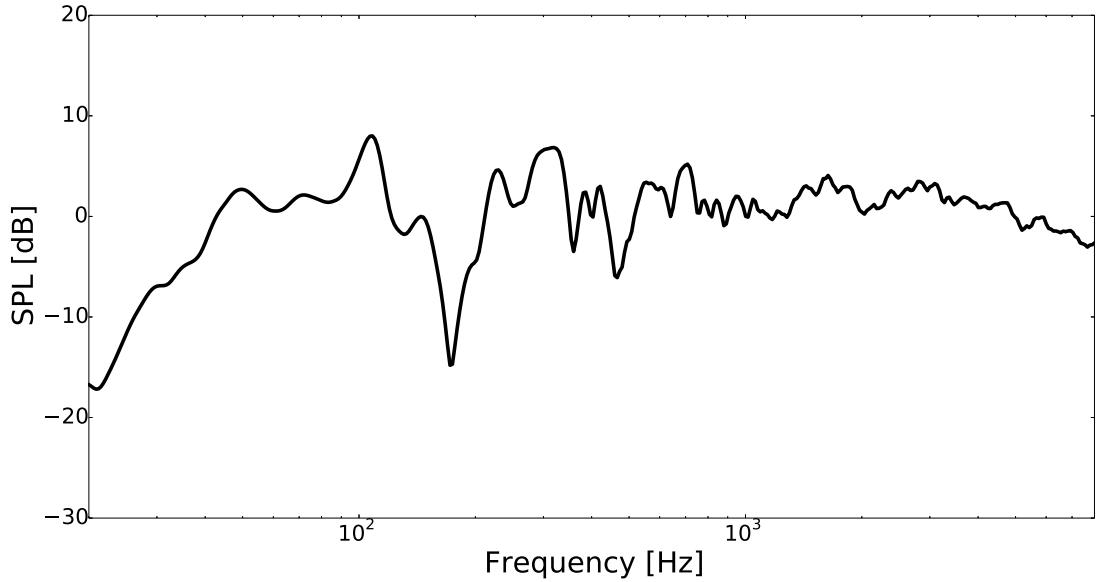


Figure 3.5: Frequency response of the Critical Listening Lab at CIRMMT at the listening position, relative to 0 dB at 1 kHz

[169], while others set a fixed level for all subjects [170]. No such constraints were deemed necessary here.

3.2.4 Subject selection and surveys

In the following, a distinction is made between skill, i.e. experience in audio or music, and training, i.e. preparing for a specific test, for instance by including a training stage from which results are not used for analysis. Therefore, a subject can be skilled on account of being an audio professional, but untrained due to lack of a training stage preceding the listening test.

Results from more skilled or trained subjects are likely to have higher discrimination [133, 171] and to be more reliable [130]. For this reason, the subjects selected for this task are all expert listeners. Training is not considered necessary due to the subjects' expertise, the low complexity of the task, and spontaneous nature of the preference rating and free-choice profiling. Furthermore, it is also costly both in terms of time and materials, as responses from a training phase are not usually regarded as valid results. Instead, the order of the pages is logged, so that it is possible to omit the results of the first part of the test if necessary.

Exclusion of a certain subject's results can also be deemed necessary based on self-

reported hearing problems, indication of misunderstanding the assignment, strong deviation from other subjects, failure to repeat ratings, or incomplete data. In some instances, one or more stimuli were not evaluated by a subject, in which case the ratings of the remaining stimuli are not necessarily comparable to the other ratings. All other results are used as none were deemed to be undesirable outliers.

Finally, a post-test survey was used to establish the subjects' gender, age, experience with audio engineering and playing a musical instrument (in number of years and described in more detail), whether they had previously participated in (non-medical) listening tests, and whether they had a cold or condition which could negatively affect their hearing [163]. Completing the survey was not mandatory due to the sensitive nature of some questions, yet none were left blank.

3.2.5 Tools

Existing tools

Listening tests require specialised software, usually custom-built, with meticulously designed interfaces and carefully formulated questions, and capable of playing back high quality audio with rapid switching between different samples. Several tools to run such tests exist: see Table 3.5 for a selection of free, publicly available applications. At present, HULTI-GEN [172] is the only example of a toolbox that presents the user with a large number of different test interfaces and customisation without requiring manual editing of configuration files or code, or knowledge of any programming language. While it was developed in Max, it does not require a copy to be run. With the exception of BeagleJS, which includes an example server side script for result collection, the tests have to be set up and conducted locally, and results are stored on the machine itself. In other words, remote deployment is not possible and the experimenter has to be present. As the single-axis, multi-stimulus interface described above is not supported by the available tools, this section presents two tools (*APE* and *WAET* in Table 3.5) which address this.

Other listening test software has been described in literature, but is not publicly avail-

Table 3.5: Existing listening test platforms and their features and supported interface types. ML stands for MATLAB, and JS stands for HTML/JavaScript. APE and WAET were developed by the author, and presented herein.

Toolbox	APE	BeagleJS	HULTI-GEN	MUSHRAM	Scale	WhisPER	WAET
Reference	[103]	[173]	[172]	[174]	[175]	[176]	[127]
Language	ML	JS	MAX	ML	ML	ML	JS
Remote		✓					✓
Time-aligned							✓
MUSHRA (ITU-R BS.1534)		✓	✓	✓			✓
Pairwise / AB Test	✓		✓				✓
Rank scale			✓				✓
Likert scale		✓			✓		✓
ABC/HR (ITU-R BS.1116)		✓					✓
-50 to 50 Bipolar with reference		✓					✓
Absolute Category Rating Scale		✓					✓
Degradation Category Rating Scale		✓					✓
Comparison Category Rating Scale		✓			✓		✓
9 Point Hedonic Category Rating Scale		✓			✓		✓
ITU-R 5 Continuous Impairment Scale		✓					✓
Multi-attribute ratings		✓					✓
ABX test	✓	✓					✓
Semantic differential					✓	✓	✓
Adaptive psychophysical methods						✓	
Repertory Grid Technique						✓	
n-Alternative Forced Choice					✓		
Single axis multi-stimulus		✓					✓

able at this time and therefore not considered here [177–180]. *MushraJS*¹⁶ is superseded by *BeagleJS*.

Audio Perceptual Evaluation (APE) tool for MATLAB

To accommodate multi-stimulus, single-axis rating with comments, a MATLAB tool was developed featuring both this interface and a pairwise evaluation mode. Multiple, simultaneously presented axes, with each axis corresponding to a certain attribute, are also supported and used by the author in [42]. The reference and hidden anchor are optional. Both stimulus and page order can be randomised. Configuration of a new test consists of a simple text file containing the number of scales, the scale names, number of stimuli, initial slider positions (randomised by default), scale marks, and quantisation of the scale. A separate text file lists the directory and names of the sound files. The results of the test are also returned as a text file, containing the subject ID, date, initial and final positions of the sliders per axis, comments, random mapping of stimulus numbers, elapsed time per page, and the sequence in which the different stimuli were played. The structure of this tool is based on an earlier MATLAB tool accompanying [181].

The software was published¹⁷ to help others set up similar listening tests without the need to develop an interface from scratch [182, 183]. This raised the bar with regard to software development as different operating systems and new versions of MATLAB had to be supported, and increased the quality and stability of the code as users reported problems.

This community usage, in addition to own experience and subject feedback during the experiment discussed below, inspired improvements to this software and eventually guided the development of a new tool. For instance, the following issues were identified:

- In the event of a MATLAB crash or other interruption of the test, it should be possible to keep the results and repeat the test from where the subject left off.
- Before continuing to the next page, asserting all stimuli were auditioned preserves

¹⁶github.com/akaroice/mushraJS

¹⁷code.soundsoftware.ac.uk/projects/ape

the validity of the results. If one or more fragments were not played, the results can turn out quite differently.

- Switching between time-aligned samples is possible, though a brief pause is heard.
- From the perspective of the experimenter, increasing numbers of test participants necessitate automated compilation, processing, and even visualisation of test results.
- While the order of playback is logged, no information is stored regarding the time and length of each audio playback, the corresponding positions in the audio file, or the time of slider movement events. Such metrics can help identify low-quality subjects and learn how the interface is used.

The main drawbacks of this tool were the tedious maintenance of the code as bugs were identified across different operating systems and new versions of MATLAB, the resulting difficult deployment and troubleshooting, and the requirement to have a MATLAB license on the listening room's computer.

The case for browser-based listening tests

In many situations, listening tests are run on one or more computers in dedicated listening rooms, sometimes in different cities or countries. As these computers may have different operating systems and versions of necessary software, developing an interface that works on all machines can be a challenge. Furthermore, as new versions of such software may not support the tool, it is best to reduce dependencies to a minimum. When the test is run locally, a problem with the machine itself can lead to loss of all results thus far — including tests from previous subjects if these were not backed up. Result collection from several computers, especially when they are remote, is tedious and can easily lead to lost or misplaced data. Similarly, installation, configuration, and troubleshooting can be a hurdle for participants or a proxy standing in for the experimenter.

All these potential obstacles are mitigated in the case of a web-based listening test: system requirements are essentially reduced to the availability of certain browsers, installation and configuration of software is not needed, and results could be sent over

the web and stored elsewhere. On the server side, deployment requirements only consist of a basic web server, with PHP functionality or similar if result collection and online access to results is desired. As recruiting participants can be very time-consuming, and as some studies necessitate a large or diverse number of participants, browser-based tests can enable participants in multiple locations to perform the test simultaneously [184].

Finally, any browser-based listening test can be integrated within other sites or enhanced with various kinds of web technologies. This can include YouTube videos as instructions, HTML index pages tracking progression through a series of tests, automatic scheduling of a next test through Doodle, awarding an Amazon voucher, or access to an event invitation on Eventbrite.

Naturally, remote deployment of listening tests inevitably leads to loss of control to a certain degree, as the experimenter is not familiar with the subjects and possibly the listening environment, and not present to notice irregularities or clarify misunderstandings. Some safeguards are possible, like assertions regarding proper interface use and extensive diagnostics, but the difference in control cannot be avoided entirely. Note, however, that in some cases the ecological validity of the familiar listening test environment and the high degree of voluntariness may be an advantage [185]. While studies have failed to show a difference in reliability between controlled and online listening tests [163, 186], these were not concerned with the assessment of music production practices.

In this work, all perceptual evaluation takes place in a dedicated, high quality listening room. For most of the experiments, the author was not in the same country, but a proxy filled in. The observations in this section were made during the process of conducting listening tests using the MATLAB-based interface described in the previous section.

Web Audio Evaluation Tool

To address the aforementioned concerns, a new, browser-based tool was created with which a wide variety of highly configurable tests can be designed, while keeping setup and result collection as straightforward as possible.

Whereas most available software still requires a substantial amount of programming or tedious configuration on behalf of the user, the Web Audio Evaluation Tool allows anything from test setup to visualisation of results to happen entirely in the browser, making it attractive to researchers with less technical backgrounds as well. To this end, all of the user modifiable options are set in a single XML document that can be written manually from scratch or from an existing document, or using the included test creator HTML GUI. The code itself only needs to be altered when advanced modifications need to be made.

There are several benefits to providing basic analysis tools in the browser: they allow immediate diagnosis of problems, with the interface or with the test subject; they may be sufficient for many researchers' purposes; and test subjects may enjoy seeing an overview of their results — or all results thus far — at the end of their tests. An example of such visualisations is shown in Figure 3.6.

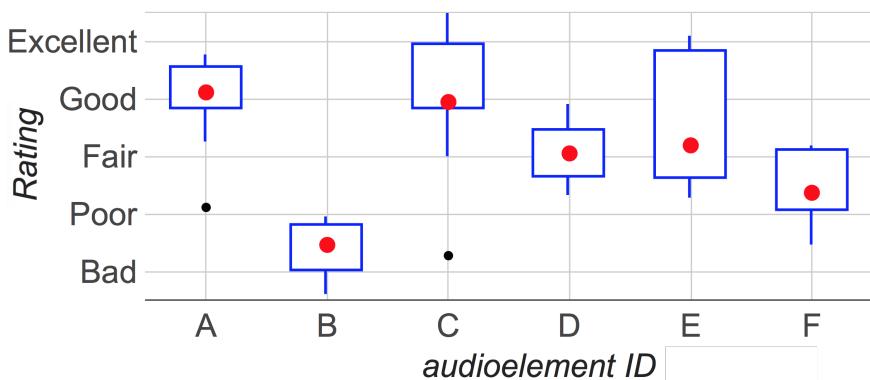


Figure 3.6: Online box and whisker plot showing the aggregated numerical ratings of six stimuli by a group of subjects

With the exception of optional remote result storage and access, the tool exclusively uses client side processing utilising the new HTML5 Web Audio API, supported by most major web browsers. This API allows for constructing audio processing elements and connecting them together to produce a high quality, real-time signal processor to manipulate audio streams on a sample level [184]. It further supports multichannel processing and has an accurate playback timer for precise, scheduled playback control. The Web Audio API is controlled through the browser's JavaScript engine and is therefore highly configurable. Because processing is all performed in a low latency thread separate from the main JavaScript thread, blocking due to real time processing does not occur.

Each audio sample is downloaded asynchronously into the JavaScript environment for further processing. This is particularly useful for the Web Audio API because it supports downloading of files in their binary form. Once downloaded, the file is decoded into a raw *float32* array using the Web Audio API offline decoder. The decoded audio is then ready for instant playback, making the interface very responsive. Immediate and seamless switching between time-aligned samples is made possible by playing back all samples at the same time, with all gains equal to zero except the currently playing sample. The integrated loudness of each sample is calculated and stored to enable on-the-fly loudness normalisation. Performing this in the browser obviates any need for pre-processing.

To address the problem of unevaluated stimuli, an optional assertion reminds the subject to play back all samples at least partially if they did not do so before submitting. In addition, safeguards are available to ascertain that every sample was auditioned in its entirety, that every slider was moved, that all commented boxes contain text, or that at least one slider is below or above a certain value.

Owing to the tool's stability, and warning messages when closing a window, incomplete tests are all but avoided. When a test is somehow interrupted, by human or machine, it can be resumed from where the subject had left off because of continual intermediate session saves.

To allow for extensive analysis, diagnostics, and subject selection, it is possible to track which parts of the audio fragments were listened to and when; at what point in the audio stream the participant switched to a different fragment; and how a fragment's rating was adjusted over time within a session. Using this data, the timeline of the test can be visualised as in Figure 3.7 for each subject and each page. Volume changes and failed submission attempts (when the conditions are not fulfilled) are logged with a timestamp as well.

To accommodate the widest possible variety of tests, other optional functionality includes cross-fades, fade-outs and fade-ins, pre- and post-silence, looping, a scrubber bar, a volume slider, a progress indicator, arbitrary per-sample gain, specific sample rate enforcement, an outside reference, a hidden reference, hidden anchors, an audiometric test where sine tones octaves apart are to be set at equal loudness, logging browser

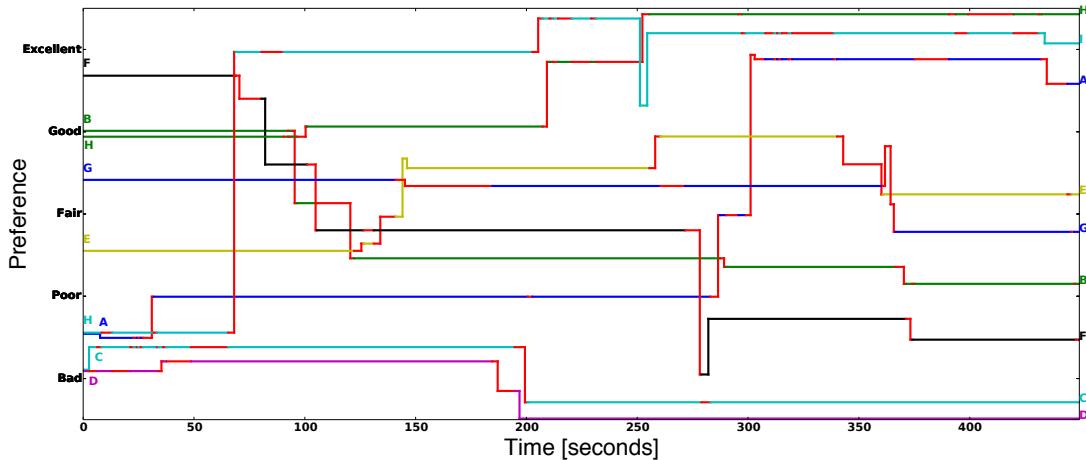


Figure 3.7: This timeline of a single subject’s listening test shows playback of fragments (red segments) and marker movements on the rating axis in function of time.

and display information, customisable marker labels, and built-in pre-/post-test and pre-/post-page surveys.

In an effort to open the tool to any kind of audio evaluation task, a wide range of highly customisable interfaces were implemented, such as AB(C...), ABX, vertical sliders (MUSHRA-style [130]), horizontal sliders, radio buttons (Likert-style), and waveform annotation. From these templates, all common, standardised listening test formats can be implemented — see Table 3.6.

Publishing and promoting the tool has led to extensive use¹⁸ and a large volume of feedback, as it has been used for studies on automatic audio effects [41], speech intelligibility [195], preference of commentary level [196], and realism of synthesised sound effects [197], among others. This has improved the code to a point where it is compatible with any browser supporting the Web Audio API and HTML 5, and sufficiently robust to handle the substantial challenges with which a cross-platform, web-based, user-facing, and critical piece of software has to cope. In addition to the time saved by using an off-the-shelf, feature-rich tool, researchers also benefit from an experimental apparatus that is well-documented and extensively tested by others, owing to its open character and versatility.

The code and documentation can be downloaded from the GitHub page¹⁹ (git) or

¹⁸github.com/BrechtDeMan/WebAudioEvaluationTool/wiki/Examples

¹⁹github.com/BrechtDeMan/WebAudioEvaluationTool

Table 3.6: Selection of supported listening test formats

Name	Ref.	Description
5 pt. Continuous Impairment	[187]	Same as ABC/HR but with a reference.
9 pt. Hedonic Category Rating	[188]	Each stimulus has a seven point scale with values: Like extremely, Like very much, Like moderate, Like slightly, Neither like nor dislike, Dislike extremely, Dislike very much, Dislike moderate, Dislike slightly. There is also a provided reference.
–50 to 50 bipolar w/ ref.		Each stimulus has a continuous scale –50 to 50 with default values as 0 in middle and a reference.
AB test	[135]	Two stimuli are presented simultaneously, participant selects a preferred stimulus.
ABC/HR	[169]	(Mean Opinion Score: MOS): each stimulus has a continuous scale (5–1), labelled as Imperceptible, Perceptible but not annoying, Slightly annoying, Annoying, Very annoying.
ABX test	[189]	Two stimuli are presented along with a reference and the participant has to select a preferred stimulus, often the closest to the reference.
ACR	[190]	Absolute Category Rating Scale. Like Likert but labels are Bad, Poor, Fair, Good, Excellent.
CCR	[190]	Comparison Category Rating. Like ACR & DCR, but 7 point scale, with reference, and labels are Much better, Better, Slightly better, About the same, Slightly worse, Worse, Much worse.
DCR	[190]	Degradation Category Rating. Like ABC & Likert, but labels are (5) Inaudible, (4) Audible but not annoying, (3) Slightly annoying, (2) Annoying, (1) Very annoying.
Likert	[191]	Each stimulus has a five point scale with values: Strongly agree, Agree, Neutral, Disagree, Strongly disagree.
MUSHRA	[192]	See Section 3.2.2.
Pairwise comparison	[193]	Every stimulus is rated as being either better or worse than the reference.
Rank	[194]	Stimuli ranked on single horizontal scale, where they are ordered in preference order.

SoundSoftware repository²⁰ (Mercurial). Further technical details can be found in the associated publications [126, 127].

3.2.6 Perceptual evaluation experiment

Design

The mixes were evaluated in a listening test to measure preference, as perceived by a group of trained listeners. The independent variables of the experiment were *mix* (or mix engineer) and *song*. The dependent variables consisted of preference rating and the free-choice profiling results.

Participants

For the perceptual evaluation experiment there were a total of 34 participants: 24 participants from the mix creation process and 10 instructors (all male) from the same sound recording program. Between 13 and 22 ratings were collected per mix. Each student received a small compensation for their time upon taking part in the listening test.

Materials

The source content and mix procedure was described in Section 3.1.3 of this chapter.

For the purpose of perceptual evaluation, a fragment consisting of the second verse and chorus was used. With an average length of one minute, this reduced the strain on the subjects' attention, likely leading to more reliable listening test results. It also placed the focus on a region of the song where the most musical elements were active. In particular, the elements which all songs have in common (drums, lead vocal, and a bass instrument) were all active here. A fade-in and fade-out of one second were applied at the start and end of the fragment [66].

²⁰code.soundsoftware.ac.uk/projects/webaudioevaluationtool

Apparatus

The listening tests in this chapter were carried out using the MATLAB-based APE tool [103], following the principles set forth above.

Procedure

The listening test was conducted with one participant at a time. After having been shown how to operate the interface, the participant was instructed — both written and verbally — to audition the samples as often as desired, to rate the different mixes according to their preference, and to write extensive comments in support of their ratings, for instance ‘why they rated a fragment the way they did’ and ‘what was particular or different about it’.

The instructions stated participants could use the preference rating scale however they saw fit, not requiring any sample to be rated at 0% or 100% of the scale. As such, the ratings were not anchored at any point except by the subjective adjectives on the rating scale, and reflected both the relative ratings of the stimuli with regard to one another, as well as a general appraisal of the stimuli. For instance, it was possible to rate no mixes as ‘Excellent’. In post-processing of the ratings, the effect of various forms of normalisation was studied, including stretching each subject’s ratings over the full scale, subtracting their mean or median, dividing by their standard deviation, and a combination of the aforementioned, but none were found to yield more significant or meaningfully different results.

Songs 1–8 (Table 3.2) were evaluated only by participants who did not take part in mixing that particular song. This reduced influence from having made mixing decisions during their own mix and generally being exposed to the song [66], while allowing to assess more content in less time. As a consequence, students assessed two songs per session (two in the Autumn semester of 2013, and two in the Spring semester of 2014), and others assessed four. To allow for analysis of self-assessment (see Section 4.2), songs 9 and 10 were analysed by the students who participated in mixing the respective songs, too.

Subjects were encouraged to take breaks between different pages if needed to prevent

listening fatigue. Subjects spent an average of $17 \text{ min} \pm 8 \text{ min}$ per song, well below the recommended duration limit found in literature [198]. The first evaluated song took $20 \text{ min} \pm 10 \text{ min}$, then $14 \text{ min} \pm 6 \text{ min}$, $13 \text{ min} \pm 5 \text{ min}$ and $12 \text{ min} \pm 3 \text{ min}$.

3.3 Conclusion

Examination of the resources available to researchers on quantitative analysis and perceptual evaluation of multitrack mix practices shows that improvements are possible, which this work addresses on multiple fronts.

First, a multitrack audio repository with semantic database was created in the form of the Open Multitrack Testbed, providing a centralised resource for raw streams of audio, combinations thereof, and accompanying metadata. Consisting of content that is readily obtainable online, often under liberal licenses, it promotes reproducibility and sustainability of this work and others, and continues to grow by welcoming contributions from the community.

Second, a dataset of realistic mixes was produced from high-quality music recordings, largely shared on said Testbed. In contrast with the type of data in most previous studies, these mixes are maximally representative of commercial music production, having been contributed by skilled mix engineers using professional tools in a familiar environment. Even so, in-depth analysis is possible as detailed parameter settings and raw audio are available.

Third, a methodology for the perceptual evaluation of music production practices was constructed, weighing different approaches and parameters for the task of rating and describing differently processed versions of musical source audio. Based on the proposed principles, listening test software was developed and a perceptual evaluation experiment was conducted to compare the different mixes. Results from this experiment are reported and analysed in the next chapter.

Finally, during the use of this listening test tool, further issues were identified and addressed in the implementation of a second, browser-based tool, that was shared as well, and used in Chapter 5.

The requirements of studies on this relatively recent and specialised topic, including multitrack audio and perceptual evaluation of subtle, highly subjective and multidimensional differences, are clearly different from other fields. As a result, it was not possible to rely on datasets and tools from related disciplines. Conversely, the assets

presented here are themselves proving useful in a variety of audio- and music-related domains [29, 40, 41, 118–121, 195, 197].

Chapter 4

Single group analysis

4.1 Objective features

To learn how people mix, low-level audio features are extracted from the mixes obtained in Chapter 3, as well as from their constituent elements. These can reveal trends and variances which further the understanding of mixing practices, and ultimately confirm, improve, or replace assumptions made in automatic mixing systems.

4.1.1 Features overview

The materials considered for this analysis are songs 1–8 (Table 3.2), each mixed by eight engineers whose recreated sessions are available. While some deviated slightly from the permitted set of tools, this was of no consequence to the elements under investigation. Three types of instruments — drums, bass, and lead vocal — are analysed here, as they are featured in all test songs, and are common elements in contemporary music in general. Furthermore, the drums are split into the elements kick drum, snare drum, and ‘rest’ which contains overhead, hi-hat, room microphones, and the occasional toms. Three out of eight songs had a male lead vocalist, and half of the songs featured a double bass (in one case part bowed) while the other half had a bass guitar for the bass part.

The audio was recorded and mixed at a sample rate of 96 kHz, but converted to 44.1 kHz

	Feature	Ref.
DYNAMIC	Loudness	[92]
	Crest factor (100 ms)	}
	Crest factor (1 s)	
	Activity	[23]
	Low energy	[202]
STEREO	SPS	}
	$P_{[band]}$	
	Side/mid ratio	
	Left/right imbalance	
SPECTRAL	Centroid	}
	Brightness	
	Spread	
	Skewness	
	Kurtosis	
	Rolloff 95%	
	Rolloff 85%	
	Entropy	
	Flatness	
	Roughness	
	Irregularity	
	Zero-crossing rate	
Octave band energies		

Table 4.1: List of extracted features

using SoX¹ to reduce computational cost and to calculate spectral features based on the mostly audible region. Sample rates are rarely higher in the domain of music information retrieval, from which most of the features were borrowed. The processed tracks are rendered from the digital audio workstation with all other tracks inactive, but with an unchanged signal path including send effects and bus processing².

The set of extracted features (Table 4.1) has been tailored to reflect dynamic, spatial, and spectral signal properties relevant to music production. Where applicable, the mean of the feature value over all frames is used. For the purpose of this investigation, only a fragment of the song consisting of the second verse and chorus is analysed, as most sources (including drums, bass, and lead vocal) are active here. When elements were muted (e.g. snare drum or kick drum track when deemed redundant by the mix engineer), the corresponding values are dropped from the analysis.

¹SoX.sourceforge.net

²When disabling the other tracks, nonlinear processes on groups of tracks (such as bus dynamic range compression) will result in a different effect on the rendered track since the processor may be triggered differently. While for the purpose of this experiment, the difference in triggering of bus compression does not affect the considered features significantly, it should be noted that for rigorous extraction of processed tracks, in such a manner that when summed together they result in the final mix, the true, time-varying bus compression gain should be measured and applied on the single tracks.

As a simple RMS level can be strongly influenced by high energy at frequencies the human ear is not particularly sensitive to, the perceptually informed ITU-R BS.1770 loudness measure of the processed source versus that of the complete mix is used instead [92]. More sophisticated, multi-band loudness features, which account for auditory masking by simultaneously playing sources, are not considered here as their performance is inferior to simpler, single band algorithms, particularly on broadband material [24, 205–207].

The crest factor over a window of 100 ms and 1 s with 50% overlap measures the short term dynamic range of the signal [201].

Gating, muting, and other processes that introduce silence are quantified as the percentage of time the track is active, with the activity state indicated by a Schmitt trigger (hysteresis gate) with thresholds at $L_1 = -25$ LUFS and $L_2 = -30$ LUFS, as in [23].

Spatial processing is measured using the Stereo Panning Spectrum (SPS), showing the spatial position of a certain time-frequency bin, and the Panning Root Mean Square ($P_{[band]}$), the RMS of the SPS over a number of frequency bins [203]. Specifically, the analysis includes the absolute value of the SPS, averaged over time, and the standard P_{total} (all bins), P_{low} (0–250 Hz), P_{mid} (250–2500 Hz), and P_{high} (2500–22050 Hz), also averaged over time.

Furthermore, two simple stereo measures are proposed. The side/mid ratio, calculated as the power of the side channel (average of left channel and polarity-reversed right channel, Equation (4.1)) over the power of the mid channel (average of left and right channel, Equation (4.2)), measures stereo width:

$$x_S = \frac{x_L - x_R}{2} \quad (4.1)$$

$$x_M = \frac{x_L + x_R}{2} \quad (4.2)$$

where x_L and x_R are the audio signals carried by the left and right channel, and x_S and x_M are the side and mid signal, respectively.

The left/right imbalance is defined as $|(R - L)/(R + L)|$, where L is the total power of the left channel, and R is the total power of the right channel. Thus, a centred track has low imbalance (≈ 0) and low side/mid ratio (≈ 0), while a hard panned track has high imbalance (≈ 1) and high side/mid ratio (≈ 1). Note that while these features are related, they do not mean the same thing. A stereo source could have uncorrelated or out-of-phase signals with equal energy in the left and right channel respectively, which would lead to a low left/right imbalance (≈ 0) and a high side/mid ratio (≈ 1 or $\rightarrow \infty$, respectively).

Finally, several features from the MIR Toolbox [204] (with the default 50 ms window length) as well as octave band energies describe the spectral characteristics of the audio.

4.1.2 Statistical analysis of audio features

Both the absolute values of the extracted features (showing the tracks' desired characteristics) as well as the change in features between raw and processed tracks (showing common manipulations) are considered. When taking only the manipulations into account, similar to blindly applying a software plugin's presets, the results would be less translatable to situations where the source material's properties differ from those in this work. Conversely, only examining absolute values would not reveal common practices that are less dependent on the source material.

Analysis of variance

Table 4.2 shows the mean values of the features, as well as the standard deviation between different mix engineers and the standard deviation between different songs, for the various considered instruments. Most features show greater variance across different songs for the same engineer, than over different engineers for the same song. Notable exceptions to this are the left/right imbalance and spectral roughness, which appear to be more dependent on the engineer than on the source content.

The change of features (difference before and after processing, where applicable), shown in Table 4.3, varies more between different songs than between different engineers,

Table 4.2: Average values of features per instrument, including average over all instruments and total mix, with standard deviation between different songs by the same mix engineer (top), and between different mixes of the same song (bottom). Bold figures indicate where variance across different engineers exceeds variance across different songs.

Feature	Kick drum	Snare drum	Rest drums	Bass	Lead vocal	Average	Mix
Loudness [LU]	-13.15 ± 4.05 3.89	-16.78 ± 6.17 4.57	-12.68 ± 5.46 2.80	-9.50 ± 3.51 2.86	-2.65 ± 1.52 1.31	-10.95 ± 4.14 3.09	N/A
Crest (100 ms)	3.599 ± 0.603 0.330	4.968 ± 0.998 0.469	4.510 ± 1.065 0.354	2.565 ± 0.443 0.166	3.315 ± 0.403 0.208	3.791 ± 0.634 0.274	3.332 ± 0.294 0.116
Crest (1 s)	9.824 ± 3.074 1.911	16.724 ± 6.458 3.135	12.472 ± 4.710 1.823	4.339 ± 1.098 0.449	5.283 ± 1.102 0.514	9.728 ± 2.907 1.398	5.315 ± 0.997 0.554
Activity	0.676 ± 0.250 0.122	0.861 ± 0.161 0.078	0.909 ± 0.115 0.029	0.958 ± 0.076 0.009	0.844 ± 0.089 0.044	0.850 ± 0.117 0.048	0.995 ± 0.009 0.004
Low energy	0.752 ± 0.113 0.081	0.723 ± 0.084 0.055	0.682 ± 0.047 0.034	0.507 ± 0.096 0.033	0.544 ± 0.065 0.048	0.641 ± 0.073 0.048	0.541 ± 0.035 0.038
L/R imbalance	0.075 ± 0.094 0.137	0.144 ± 0.153 0.227	0.361 ± 0.303 0.213	0.107 ± 0.135 0.176	0.045 ± 0.072 0.085	0.146 ± 0.139 0.152	0.088 ± 0.075 0.074
Side/mid ratio	0.036 ± 0.055 0.076	0.036 ± 0.040 0.043	0.242 ± 0.183 0.154	0.009 ± 0.013 0.015	0.022 ± 0.018 0.022	0.069 ± 0.060 0.059	0.101 ± 0.049 0.046
P_{total}	0.104 ± 0.102 0.090	0.108 ± 0.082 0.059	0.307 ± 0.028 0.027	0.075 ± 0.093 0.083	0.134 ± 0.022 0.027	0.145 ± 0.052 0.052	0.234 ± 0.027 0.027
P_{low}	0.066 ± 0.078 0.087	0.122 ± 0.102 0.073	0.243 ± 0.045 0.041	0.040 ± 0.063 0.059	0.147 ± 0.034 0.042	0.123 ± 0.061 0.056	0.188 ± 0.042 0.034
P_{mid}	0.066 ± 0.074 0.076	0.114 ± 0.090 0.064	0.290 ± 0.023 0.023	0.052 ± 0.082 0.067	0.177 ± 0.027 0.035	0.140 ± 0.054 0.048	0.248 ± 0.027 0.023
P_{high}	0.106 ± 0.104 0.091	0.105 ± 0.081 0.058	0.309 ± 0.029 0.028	0.076 ± 0.094 0.085	0.124 ± 0.022 0.028	0.144 ± 0.061 0.053	0.231 ± 0.033 0.029
Centroid [Hz]	2253.8 ± 1065.6 729.8	4395.3 ± 1448.6 554.2	4130.8 ± 1228.1 483.2	1046.5 ± 520.1 232.4	2920.2 ± 452.1 264.7	2949.3 ± 872.1 418.6	2478.8 ± 517.9 247.1
Brightness	0.306 ± 0.105 0.103	0.598 ± 0.156 0.069	0.557 ± 0.115 0.058	0.135 ± 0.082 0.031	0.455 ± 0.071 0.040	0.410 ± 0.100 0.056	0.362 ± 0.070 0.034
Spread	3250.1 ± 783.2 447.5	4363.6 ± 701.9 335.9	4422.1 ± 734.6 292.3	4246.6 ± 559.2 320.4	3369.9 ± 324.6 191.3	3566.5 ± 587.5 298.0	3453.2 ± 421.7 200.6
Skewness	3.649 ± 1.068 0.886	1.492 ± 0.663 0.301	1.665 ± 0.682 0.246	6.234 ± 1.885 0.630	2.470 ± 0.573 0.243	3.102 ± 0.912 0.427	2.779 ± 0.600 0.257

Table 4.2: Average values of features per instrument (continued)

Feature	Kick drum	Snare drum	Rest drums	Bass	Lead vocal	Average	Mix
Kurtosis	23.847 ± 11.997 9.164	5.965 ± 2.905 1.474	7.053 ± 3.449 1.263	58.870 ± 31.874 11.107	11.579 ± 4.267 1.784	21.463 ± 9.834 4.477	13.646 ± 4.511 2.073
Rolloff 95% [Hz]	8880.1 ± 3679.2 2151.2	13450.9 ± 3100.6 1582.2	13373.4 ± 2594.1 1007.4	4389.4 ± 2714.7 1244.5	9879.0 ± 1335.7 725.3	9994.5 ± 2498.0 1240.8	9679.0 ± 1563.8 734.3
Rolloff 85% [Hz]	4513.7 ± 2736.6 1788.8	8984.3 ± 3139.7 1348.5	8755.3 ± 2742.5 975.6	1625.5 ± 1205.0 594.3	5595.8 ± 1121.4 609.7	5894.9 ± 2047.2 986.1	5026.2 ± 1337.8 599.8
Entropy	0.655 ± 0.104 0.090	0.840 ± 0.084 0.057	0.832 ± 0.051 0.025	0.552 ± 0.073 0.026	0.735 ± 0.043 0.016	0.723 ± 0.066 0.038	0.744 ± 0.043 0.015
Flatness	0.148 ± 0.072 0.051	0.350 ± 0.142 0.056	0.337 ± 0.118 0.045	0.073 ± 0.035 0.020	0.167 ± 0.030 0.018	0.215 ± 0.074 0.035	0.174 ± 0.046 0.020
Roughness	84.72 ± 98.32	36.30 ± 41.16	67.57 ± 71.76	236.04 ± 160.38	247.00 ± 216.15	134.33 ± 247.36	1843.31 ± 1341.50
Irregularity	0.158 ± 0.098 0.063	0.235 ± 0.151 0.079	0.297 ± 0.135 0.069	0.502 ± 0.176 0.065	0.540 ± 0.165 0.094	0.346 ± 0.136 0.075	0.705 ± 0.090 0.078
Zero-crossing	584.7 ± 509.5 409.4	2222.0 ± 1183.3 604.7	1988.9 ± 944.1 466.1	246.6 ± 217.8 89.6	1177.5 ± 233.7 143.6	1243.9 ± 554.3 305.4	905.2 ± 237.4 118.8

Table 4.3: Average change of feature values per instrument, including average over instrument, with standard deviation between different songs by the same mix engineer (top), and between different mixes of the same song (bottom). Bold figures indicate where variance across different engineers exceeds variance across different songs.

Feature	Kick drum	Snare drum	Bass	Lead vocal	Average
Crest (100 ms)	0.155 ± 0.425 0.330	0.167 ± 0.511 0.469	0.041 ± 0.188 0.166	0.093 ± 0.219 0.208	0.204 ± 0.448 0.305
Crest (1 s)	0.785 ± 2.071 1.911	1.508 ± 3.301 3.135	-0.020 ± 0.483 0.449	0.143 ± 0.556 0.514	1.038 ± 1.735 1.566
Activity	0.047 ± 0.155 0.122	-0.013 ± 0.082 0.078	0.007 ± 0.014 0.009	0.053 ± 0.052 0.044	0.020 ± 0.069 0.056
Low energy	-0.036 ± 0.096 0.081	-0.027 ± 0.050 0.055	-0.049 ± 0.046 0.033	-0.078 ± 0.071 0.048	-0.038 ± 0.062 0.050
Centroid [Hz]	-11.105 ± 820.672 729.797	83.273 ± 725.828 554.199	-2.492 ± 291.139 232.377	178.879 ± 253.428 264.705	59.529 ± 537.441 452.850
Brightness	0.007 ± 0.106 0.103	0.023 ± 0.092 0.069	-0.002 ± 0.036 0.031	0.021 ± 0.040 0.040	0.020 ± 0.072 0.060
Spread	-27.453 ± 522.415 447.480	-38.985 ± 443.839 335.933	-69.772 ± 423.897 320.431	161.915 ± 171.962 191.288	-42.216 ± 405.750 317.488
Skewness	-0.135 ± 0.828 0.886	-0.054 ± 0.386 0.301	0.212 ± 0.896 0.630	-0.162 ± 0.265 0.243	-0.052 ± 0.537 0.461
Kurtosis	-1.114 ± 8.394 9.164	-0.139 ± 2.038 1.474	4.158 ± 17.096 11.107	-1.272 ± 2.042 1.784	0.236 ± 6.235 4.958
Rolloff 95% [Hz]	-821.018 ± 2744.058 2151.220	-399.719 ± 2062.781 1582.166	-138.752 ± 1647.534 1244.491	580.751 ± 646.491 725.289	-271.228 ± 1728.641 1342.108
Rolloff 85% [Hz]	-327.644 ± 2360.596 1788.805	-66.458 ± 1841.282 1348.496	65.825 ± 696.096 594.270	426.618 ± 609.147 609.656	10.633 ± 1343.638 1063.360
Entropy	-0.026 ± 0.111 0.090	-0.005 ± 0.064 0.057	0.003 ± 0.027 0.026	0.020 ± 0.016 0.016	0.002 ± 0.052 0.043
Flatness	-0.004 ± 0.061 0.051	-0.006 ± 0.079 0.056	-0.003 ± 0.025 0.020	0.019 ± 0.017 0.018	0.000 ± 0.047 0.038
Roughness	-0.845 ± 100.270 98.323	-10.266 ± 63.974 43.319	-21.177 ± 185.222 176.047	83.209 ± 207.064 247.355	21.865 ± 124.857 122.264
Irregularity	0.038 ± 0.053 0.063	-0.003 ± 0.081 0.079	0.007 ± 0.122 0.065	0.164 ± 0.113 0.094	0.034 ± 0.094 0.074
Zero-crossing	23.251 ± 600.192 409.434	12.038 ± 779.660 604.737	-14.411 ± 124.740 89.644	115.484 ± 125.34 143.638	116.182 ± 455.667 342.704

too, again with the exception of spectral roughness. Spatial features and loudness are not meaningful here as raw tracks are monaural and their level or loudness is inconsequential. The total mix and ‘rest’ are also not included, as these consist of several processed tracks.

Only the lead vocal has a larger spread across different engineers than across different songs for absolute spatial features, and for changes in half of the spectral feature values. This indicates that an individual mix engineer has a relatively consistent approach to processing the lead vocal, and the source material does not have a very strong influence. For other sources, the content, context, or musical genre is a more determining factor, as variation among mix engineers is smaller than among songs.

Consider the hypotheses that the different ‘treatments’ (different source material, mix engineer, or instrument) result in the same feature values, or the same change in feature values. An analysis of variance determines for which features these hypotheses can be rejected. For those features for which there is a significant effect ($p < .05$) in both groups, a multiple comparison of population means using the Bonferroni correction establishes which instruments, engineers, or songs cause a significantly lower or higher mean feature value compared to others. For individual instruments, the source material only causes the means of the feature to differ significantly for the zero-crossing rate of the snare drum track, and for the spectral entropy of the total mix. In other words, whereas some engineers would disagree on processing values, the source material has less impact on these decisions. The outcome of these tests is discussed in more detail in the following paragraphs.

Balance

The relative loudness of the bass ($p < .01$), snare drum ($p < .05$), and other drum instruments or ‘rest’ ($p < 5 \cdot 10^{-4}$) is highly dependent on the mix engineer.

From Figure 4.1, it is apparent that the lead vocal is significantly louder than the other elements considered here. Furthermore, the vocal spans a narrow range of loudness values, suggesting a near-universal agreement on a ‘target loudness’ of about -3 LU relative to the overall mix loudness. Pestana’s study of vocal level confirms this, concluding that on average vocals are as loud as the sum of all other tracks [51]. Note that

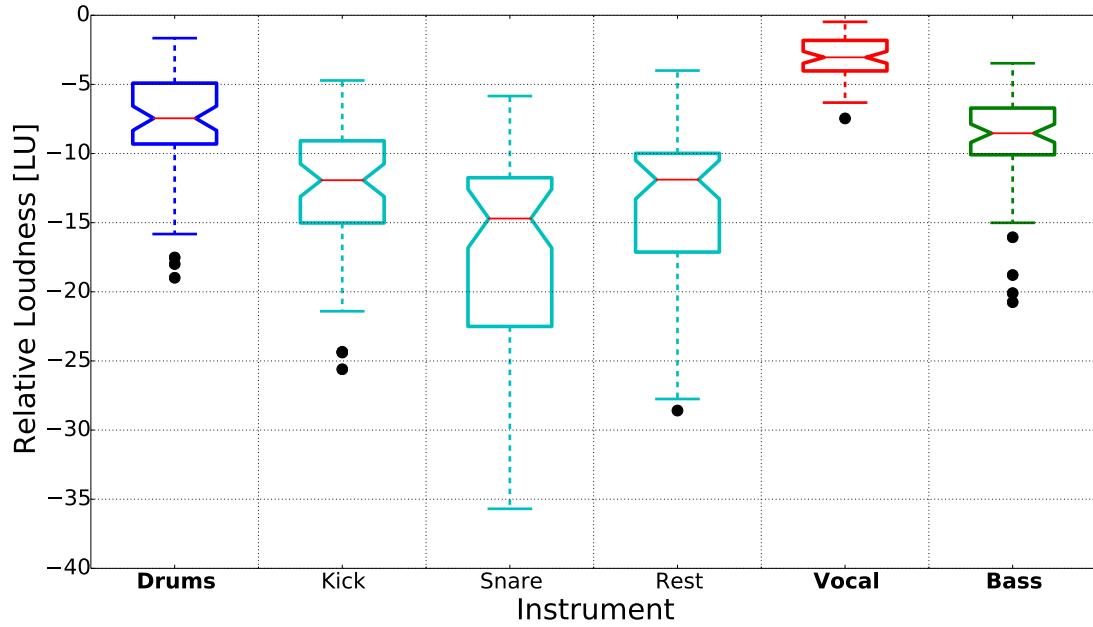


Figure 4.1: Box plot representing the loudness of the sources, across songs and mix engineers. The red horizontal line represents the median, the bottom and top of the ‘box’ represent the 25% and 75% percentile, and the dashed vertical lines extend from the minimum to the maximum, not including outliers (indicated by black dots), which are higher than the 75% percentile or lower than the 25% percentile by at least 1.5 the interquartile range. The notch spans the 95% confidence interval of the median.

the loudness shown here is relative to the whole mix, including vocals. Later work on the average relative loudness of sources further corroborated these ranges of values for vocal, bass, and drums, showing overlapping confidence intervals of the median [52]. Another study found an average vocal loudness below -6 LU relative to the total mix loudness [24], though no information is available about the mix engineers, the error is much larger, and the findings could not be reproduced and investigated further as the exact songs analysed have not been disclosed. In automatic mixing research, a popular assumption is that with the possible exception of the main element — usually the vocal — the loudness of the different tracks or sources should be equal [19, 21, 23, 53]. However, the results presented here directly contradict this hypothesis.

It should be noted that due to crosstalk between the drum microphones, and particularly overhead and room microphones, the effective loudness of the snare drum (and kick drum, albeit to a lesser extent) will differ from the loudness measured from the snare drum and kick drum tracks. Source separation methods could be employed to more accurately calculate the source loudness, and recent work on identifying overhead microphones in a multitrack session could further automate this process [121]. As a

result, the snare drum microphone loudness can be as low as -35 LU relative to the total mix loudness, though this is then compensated by a louder ‘rest’ of the drum set, and vice versa, as shown by the narrow spread of the complete drum stem loudness values. This confirms that two approaches exist with regard to mixing drums: using overhead microphones as the main signal and adding emphasis as needed with the close kick and snare drum microphones, or using the close microphones as primary signals and bringing up the more distant microphones for added ‘air’ or ‘ambience’ to taste [65].

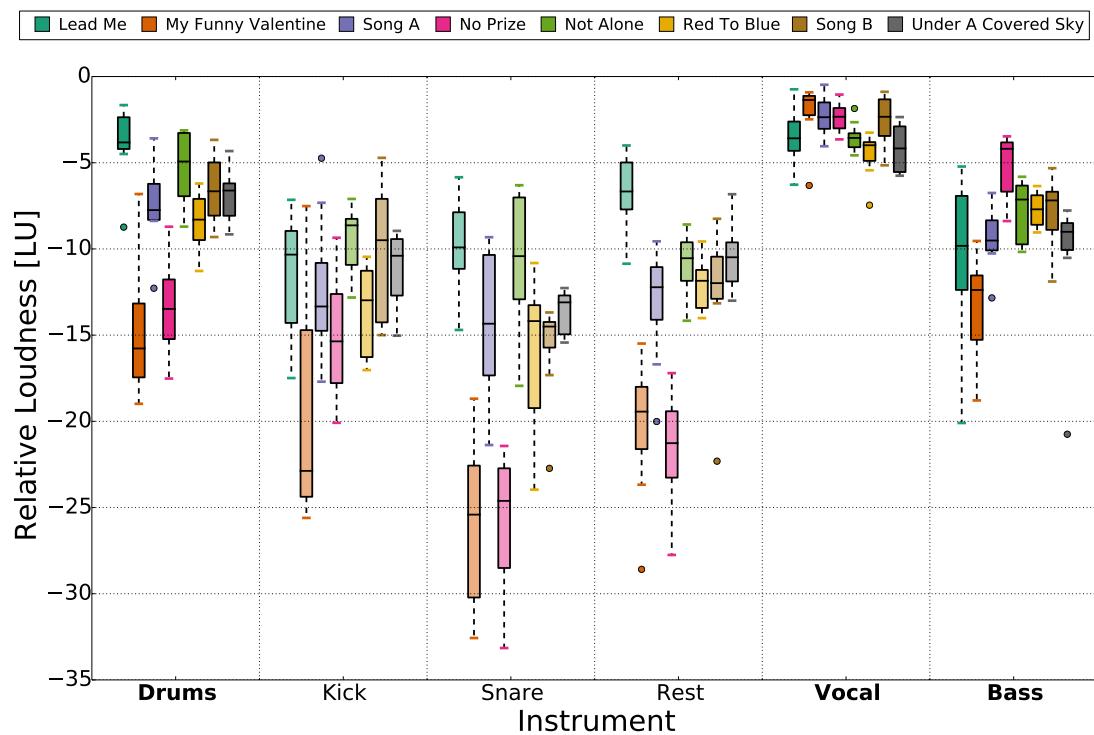


Figure 4.2: Box plot representing the loudness of the sources per song, across mix engineers. The black horizontal line represents the median, the bottom and top of the ‘box’ represent the 25% and 75% percentile, and the dashed vertical lines extend from the minimum to the maximum, not including outliers (filled circles), which are higher than the 75% percentile or lower than the 25% percentile by at least 1.5 the interquartile range.

A more detailed view of the loudness per instrument, broken down per song, is given in Figure 4.2. One obvious trend is the significantly lower drum loudness for the two jazz songs, My Funny Valentine and No Prize.

Dynamics processing

The crest factor is affected by the instrument ($p < .005$), and every instrument individually shows significantly different crest factor values for different engineers ($p < .005$). One exception to the latter is the kick drum for a crest factor window size of 1 s, where the null hypothesis was not disproved for one of the two groups of engineers and songs.

The percentage of the time a track is active depends on the mix engineer ($p < .01$), for instance the decision to gate the kick drum ($p < 10^{-4}$).

Stereo panning

The Panning Root Mean Square values (P_{band}) and side/mid ratio all show a proportionally large value for the total mix and for the ‘rest’ group, meaning these are meaningfully wider than the other, traditionally centred and ‘monaural’ sources, as can be expected. The difference is significant for all frequency bands but the lowest, where only the bass track is more central than the total mix and the drums. This confirms sound engineering textbooks and earlier research, stating that low-frequency sources as well as lead vocals and snare drums should be panned central [25, 26, 28, 65, 208].

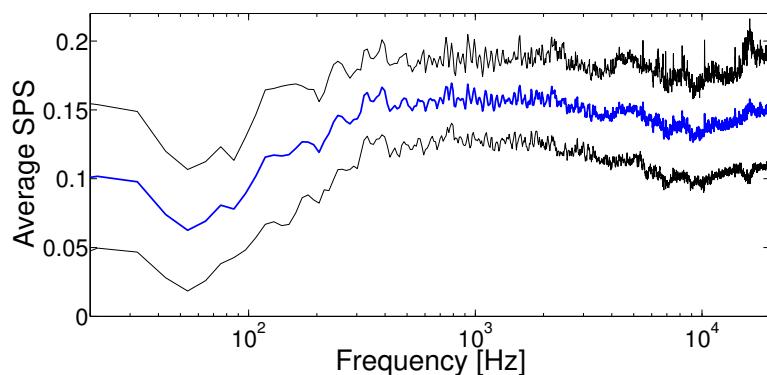


Figure 4.3: Mean Stereo Panning Spectrum (with standard deviation) over all mixes and songs

To further quantify the spatialisation for different frequencies, Figure 4.3 displays the panning as a function of frequency, using the average Stereo Panning Spectrum over all mixes and songs. From this figure, a clear increase in SPS with increasing frequency is apparent between 50 Hz and 400 Hz. However, in contrast to what is suggested by

literature [26, 208], this trend is not observed towards the very low (20–50 Hz) or higher frequencies (>400 Hz).

Equalisation

The spectral centroid of the whole (unmastered) mix varies strongly depending on the mix engineer ($p < 10^{-5}$). The centroid of the snare drum track is typically increased through processing, due to attenuation of the low frequency content, reduction of spill of instruments like kick drum, or the emphasis of frequency components above the original centroid. The brightness of each track except bass and kick drum is increased as well.

For a large set of spectral features (spectral centroid, brightness, skewness, roll-off, entropy, flatness, and zero-crossing), the engineers disagree on the preferred value for all instruments except kick drum. In other words, the values describing the spectrum of a kick drum across engineers are overlapping, implying a consistent spectral target (a certain range of appropriate values). For other features (spread, kurtosis, and irregularity) the values are different across engineers for all instruments. The roughness shows no significantly different means for any instrument except the ‘rest’ bus.

Analysis of the octave band energies of the different instruments reveals definite trends across songs and mix engineers, see Figure 4.4. The standard deviation does not consistently decrease or increase over the octave bands for any instrument when compared to the raw audio. Note that because the deviation can be skewed, some standard deviation intervals in this plot exceed 0 dB, while no (normalised) octave band can exhibit energy above 0 dB.

The suggested ‘mix target spectrum’ is in agreement with [67], which showed a ‘target spectrum’ that was more or less consistently aimed for, varying with genre and decade. Figure 4.5 shows the average spectrum of every number one hit after 2000 lies within standard deviation of the measured average mix spectrum.

The average relative change in energies is not significantly different from zero (no bands are consistently boosted or cut for certain instruments), but taking each song individually in consideration, a strong agreement of reasonably drastic boosts or cuts is

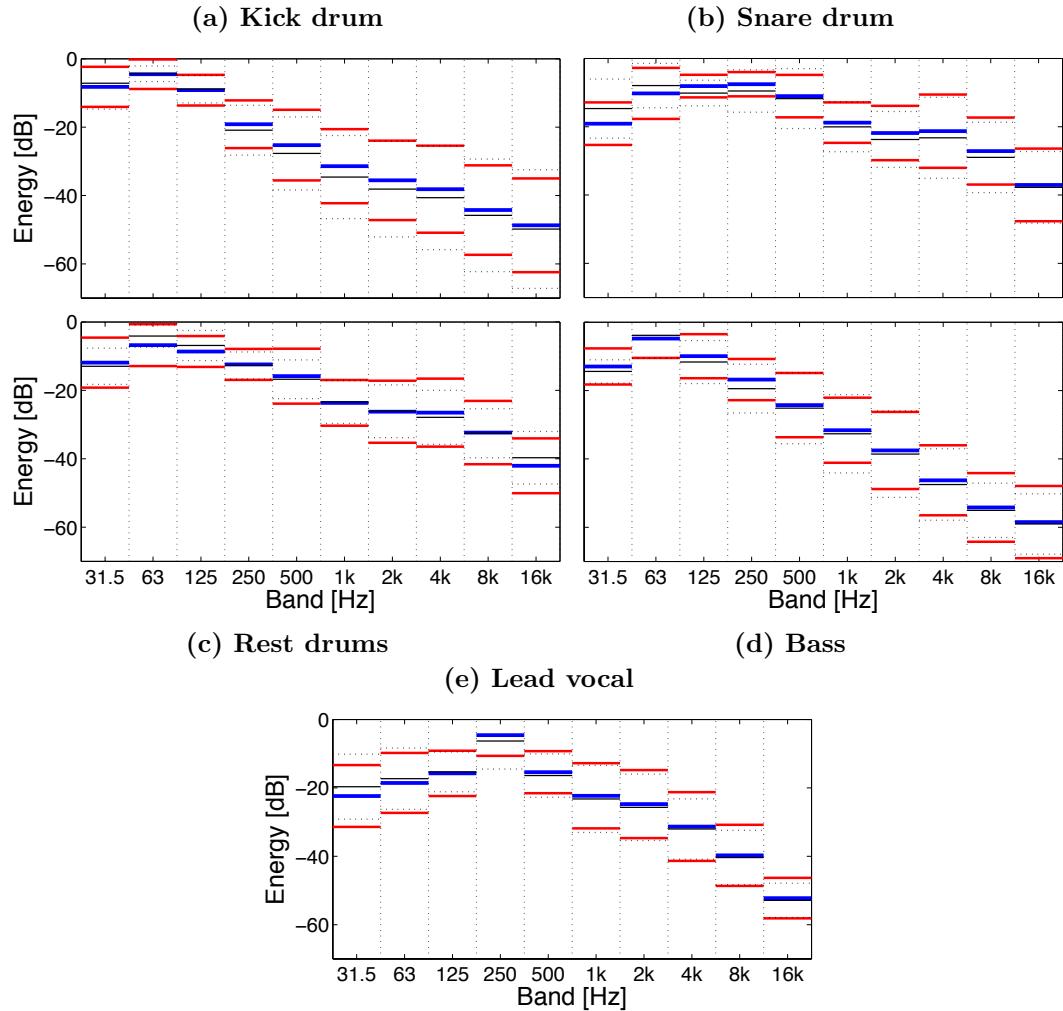


Figure 4.4: Normalised octave band energies for different instruments (average in blue and standard deviation in red) compared to raw signal (black)

shown for some songs. This confirms that the equalisation is highly dependent on the source material, and that engineers largely agree on the necessary treatment for source tracks showing spectral anomalies.

4.1.3 Workflow statistics

Beyond signal-level analysis of the processed tracks, having access to the DAW files also affords the opportunity to investigate the mixing workflow. In particular, the tendency to group tracks which exhibit a particular relationship (e.g. all guitar tracks) is considered in this section. This process, commonly referred to as subgrouping, allows faster or more convenient manipulation of several signals at once, and provides a better overview of the otherwise potentially overwhelming session.

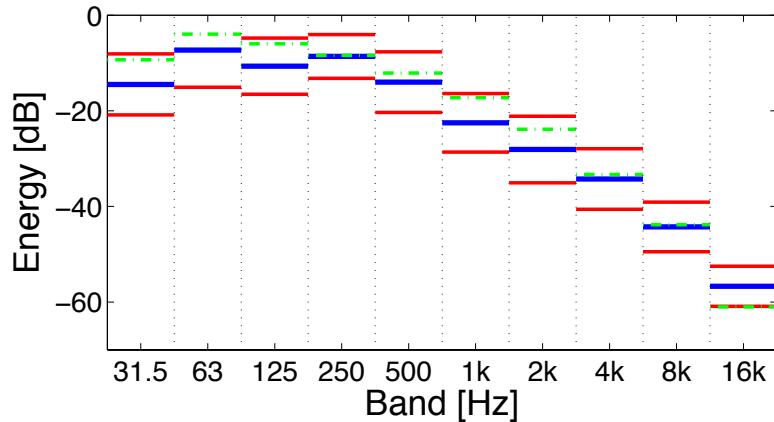


Figure 4.5: Average octave band energies for total mix, compared to ‘After 2000’ curve from [67] (green dashed line)

Subgroup type	# subgroups	# tracks
Drums	73	78
Vocals	71	48
Guitars	64	49
Keyboards	49	15
Bass	42	11

Table 4.4: Number of different individual subgroup types over all 64 mixes, and how many audio tracks of that type occurred across all 8 songs

Very little is known about how mix engineers choose to group sources. The problem was touched on briefly in [51] which showed gentle bus compression “helps blend things better”, but did not give much insight into how subgrouping is generally used. In [210], an automatic subgrouping algorithm learning from manually assigned instrument class labels is presented, but providing a deeper understanding of subgrouping by humans was not the aim of the paper.

Table 4.4 shows a breakdown of the most common instruments to be grouped together. It is clear that the likelihood of subgrouping depends on the number of tracks of that type. Indeed, the number of subgroups one creates is very strongly related to the number of tracks used in that final mix, with a Spearman rank correlation of $\rho = .93$ ($p < .01$).

Almost all mix engineers subgrouped audio tracks based on instrumentation, though Table 4.5 shows a number of subgroups containing combinations of instruments. Only 4 out of the 72 considered mixes had no subgroups at all, 3 of which were of the song *My Funny Valentine* — which has only one vocal part, one keyboard, and no guitars.

Subgroups sometimes contained other subgroups, often including several instruments

Subgroup type	# subgroups
Bass + Guitars + Keyboards + Vocals	4
Bass + Guitars + Keyboards	3
Drums + Percussion	3
Guitars + Keyboards	2
Drums + Bass + Guitars + Keyboards	2
Drums + Bass + Vocals	1
Bass + Guitars	1
Drums + Bass + Keyboards + Vocals	1

Table 4.5: Number of different multi-instrument subgroup types that occurred in all the mixes

(9 occurrences), but also consisting of only drums (10 occurrences) or vocals (3 occurrences). Upon closer inspection, it was found that in eight of the ‘nested’ drum subgroups, the overhead microphones were separated from other drum elements, so that they could be processed simultaneously as a stereo pair. In seven of these cases the kick drum, snare drum, and hi-hats, arguably the key instruments within a drum kit, were grouped together.

4.1.4 Conclusion

Sixty-four mixes from eight multitrack recordings by eight mix engineers each were analysed in terms of dynamic, spatial, and spectral processing of common key elements. This helped confirm or challenge assumptions from practical sound engineering literature and previous research, and identify consistent trends and notable points of disagreement. Most notably, the loudness of the lead vocal track relative to the total mix loudness was found to be significantly louder than all other tracks, with an average value of -3 LU. The amount of panning as a function of frequency was investigated, and found to be increasing with frequency up to about 400 Hz, above which it stays more or less constant. Lead vocal, snare drum, and low frequency elements are centrally panned. Spectral analysis has shown a definite target spectrum that agrees with the average spectrum of recent commercial recordings, even though the current content was not mastered. A greater variance of most features was measured across songs than across engineers, whereas the mean values corresponding to the different engineers were more often statistically different from each other.

The original DAW sessions of the mixes were examined to investigate subgrouping

practices, a markedly unexplored area. A strong tendency to group similar instruments together was noted, especially in the case of a high number of tracks.

Even if a limited selection of songs were studied, some genre-dependence was observed in the two only jazz songs of the set, in particular a lower drums loudness. A larger dataset is needed to make more authoritative claims.

An extrapolation to other instruments is also needed to validate the generality of the conclusions regarding the processing of drums, bass, and lead vocal at the mixing stage, and to further explore laws underpinning the processing of different instruments.

Finally, while the mixes were contributed by masters level students from a renowned sound engineering programme, perceptual evaluation is needed to determine whether they are truly representative of commercial music production. There is the possibility that unconventional or even poor mixes skewed the results and reduced their precision. At the same time, some of the chosen features may not be relevant to perception. In the next section, subjective ratings are studied in conjunction with these features to determine their importance and quantify the influence on preference. This shifts the question from “what makes a typical mix” to “what makes a good mix”.

4.2 Subjective numerical ratings

Perceptual evaluation of mixes is essential when investigating music production practices, as it reveals which processing corresponds with a generally favoured effect. In contrast, when mixes are studied in isolation, i.e. without comparison to alternative mixes or without feedback on the engineer's choices, it cannot be assumed that the work is representative of what an audience might perceive to be a good mix. Therefore, in this section, the subjective ratings from the perceptual evaluation experiment described in Section 3.2 are discussed in relation to low-level features extracted from the stereo mixes.

4.2.1 Preference rating

The preference ratings attributed to all mixes of songs 1–10 are considered (Table 3.2), including the additional professional mix and the machine-made mix. With the exception of songs 9 and 10, and the professional mixes, subjects only assessed songs they did not mix and which were therefore presumably unfamiliar to them.

Figure 4.6 shows the ratings received by every mix engineer in the test (for one or more songs) including the teachers ('P1' and 'P2', shown together as 'Pro') and the completely autonomous mix ('Auto'), as well as the combined ratings received by first year ('Y1') and second year ('Y2') students. While subjects did not agree on a clear order in terms of preference in this case, there is a definite tendency to favour certain mixes over others. Mixes by second year students are only given a slightly higher preference rating on average than those by first year students, although it should be noted the two are never assessed at the same time, i.e. each individual song was mixed by students from the same year.

Two songs (9 and 10) were also evaluated by the group of engineers who mixed the song, so that each would also assess their own mix. Except for one engineer, who rated his own mix lowest, all rated their own mix higher than the median rating their mix received (see Figure 4.7). Of these 16 participants, 13 also rated their mix higher than the average rating they attributed to other mixes of the same song. This suggests that engineers either have a consistent taste whether they are mixing themselves or

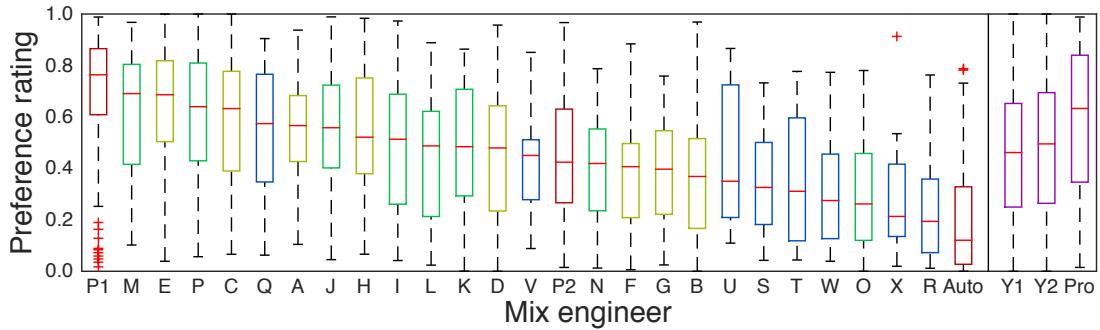


Figure 4.6: Box plot of ratings per mix engineer, in decreasing order of the median. A–H (yellow) are first year students in 2013–2014 (4 songs), and second year students in 2014–2015 (1 song); I–P (green) are second year students in 2013–2014 (4 songs), and Q–X (blue) are first year students in 2014–2015 (1 song). ‘P1’ and ‘P2’ are their teachers (‘Pro’), ‘Y1’ and ‘Y2’ are the results of mixes by first year and second year students, respectively, and ‘Auto’ denotes the automatic mix. The red horizontal line represents the median, the bottom and top of the ‘box’ represent the 25% and 75% percentile, and the vertical dashed lines extend from the minimum to the maximum, not including outliers (red pluses), which are higher than the 75% percentile or lower than the 25% percentile by at least 1.5 the interquartile range.

only listening, are subconsciously biased by the way they have recently mixed this song, outright recognise their own mix, or a combination of these. It also justifies the decision to avoid self-assessment for songs 1–8, out of concern for bias due to familiarity [66].

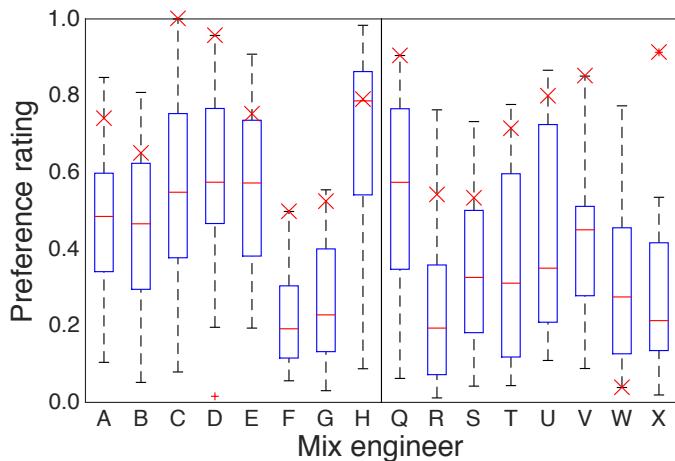


Figure 4.7: Box plot of ratings per mix engineer including their own assessment (red ‘X’) of one song. The red horizontal line represents the median, the bottom and top of the ‘box’ represent the 25% and 75% percentile, and the vertical dashed lines extend from the minimum to the maximum, not including outliers (red pluses), which are higher than the 75% percentile or lower than the 25% percentile by at least 1.5 the interquartile range.

Finally, the positive correlation (Pearson’s correlation coefficient $\rho = .52, p < 10^{-12}$) between the average rating of different mixes by the same mix engineer means that

the measured preference of a single mix is indicative of the general performance of the engineer.

4.2.2 Correlation of audio features with preference

A number of features were extracted from the 98 evaluated stereo mixes (see Table 4.6). In addition to the 33 features considered in the previous chapter, 18 new features were introduced, including more specialised dynamic range features, spectral and cepstral flux, and 12 MFCCs. As listed in Table 4.6, preference shows a positive linear correlation with microdynamics measure LDR [212] ($\rho = .26, p = .01$), and a negative linear correlation with the side/mid ratio ($\rho = -.32, p = .001$).

The preference rating used to calculate these correlations is the average of all raw ratings for each mix, regardless of each subject's use of the scale. Alternative agglomerated ratings were considered, such as post hoc scaling of each subject's ratings for a given song between 0 and 1, subtracting the average rating for that song from each rating, using the median instead of the mean, and any combination of the above. In each of these cases, the correlations found were similar and not worth reporting separately, except for the PLR (peak-to-loudness ratio), which became significant for each of the modifications, and crest factor (over the whole file) which became significant in the majority of the cases. This strengthens the confidence that increased dynamic range, as quantified in different ways by LDR, PLR, and crest factor, correlates positively with preference.

This preference towards a higher dynamic range, for musical stimuli compared at equal loudness, suggests that a mix should have peaks of sufficient magnitude. While in many situations, a high loudness for a given peak amplitude typically has a positive effect on the listener's relative preference [201, 216], it seems that when the loudness is normalised instead of the peak amplitude, a relatively higher dynamic range is preferred over a lower one. This confirms that it is better to err on the lighter side when applying dynamic range compression [37, 66].

A negative correlation between side-to-mid ratio and preference means a stronger mid channel is generally preferred. However, upon closer inspection, overly monaural mixes (very low side-to-mid ratio) generally received low ratings as well.

Table 4.6: Spearman’s rank correlation coefficient ρ (including p-values) between the extracted features and preference (average of raw ratings)

	Feature	ρ	p	Ref.
DYNAMIC RANGE	Crest factor (100ms)	−.084	.415	[201]
	Crest factor (1s)	.003	.973	
	Crest factor (whole)	.101	.323	
	Dynamic spread	.128	.211	[39]
	PLR	.142	.165	[213]
	LRA	.010	.919	[214]
	TT DR	.029	.776	[215]
STEREO	LDR	.244	.016	[212]
	Low energy	.053	.606	[202]
	Side/mid ratio	−.324	.001	
	L/R imbalance	−.007	.948	
	P_{total}	−.138	.176	[203]
	P_{low}	.084	.410	
	P_{mid}	.011	.913	
SPECTRAL	P_{high}	−.158	.121	
	Centroid	−.130	.204	[204]
	Brightness	−.181	.077	
	Spread	−.080	.433	
	Skewness	.138	.178	
	Kurtosis	.136	.183	
	Rolloff 95%	−.118	.248	
	Rolloff 85%	−.127	.216	
	Entropy	−.165	.107	
	Flatness	−.098	.340	
	Roughness	.007	.947	
	Irregularity	−.019	.854	
	Zero crossing rate	−.168	.100	
OCTAVE BAND ENERGY	Spectral flux	.169	.098	energy of octave band divided by total energy
	Cepstral flux	.087	.398	
	31.5 Hz	.047	.649	
	63 Hz	−.044	.668	
	125 Hz	−.098	.339	
	250 Hz	.044	.670	
	500 Hz	−.072	.481	
	1 kHz	−.088	.389	
	2 kHz	−.146	.155	
	4 kHz	−.135	.189	
MEL-FREQUENCY CEPSTRUM	8 kHz	−.054	.601	
	16 kHz	−.049	.634	
	MFCC1	.114	.266	
	MFCC2	−.031	.761	
	MFCC3	−.120	.242	
	MFCC4	.166	.105	
	MFCC5	.087	.394	
	MFCC6	.051	.622	
	MFCC7	.039	.707	
	MFCC8	−.109	.288	
	MFCC9	.021	.838	
	MFCC10	−.021	.837	
	MFCC11	.005	.962	
	MFCC12	−.020	.844	

Overall, these results suggest that dynamic and spatial features extracted from the audio can be predictive of preference, as confirmed by [66, 72].

4.2.3 Correlation of workflow statistics with preference

Subgrouping, as discussed in Section 4.1.3, is a technique primarily aimed at enhancing the mixing workflow, by allowing to change properties such as the level, spectrum, or effect send amount of multiple tracks. It technically does not add any new functionality, but merely saves the engineer from repeating the same operation several times, while retaining a better overview of the session. One exception to this is nonlinear processing, such as compression, which acts differently when applied to several sources at once instead of to each source individually.

However, an impact of subgrouping practices on preference ratings may be observed if the number of subgroups is indicative of the experience or performance of the engineer, if it relates to the time and effort spent on a mix, or if subgroups simply allow good mixes to be made more easily. This is quantified here by looking at the correlation between preference and the relative number of subgroups, as well as those including particular types of processing.

Specifically, the Spearman’s rank correlation coefficient is considered between the median preference for a mix, and the number of subgroups divided by the number of tracks used in that mix. It is calculated for each mix separately, and per mix engineer (four mixes per engineer), for different types of subgroups. The results of this analysis are shown in Table 4.7.

Table 4.7: Spearman’s rank correlation coefficient for different kinds of subgroups.
 $p < .01$ for all correlations except the DRC subgroup per engineer, where $p < .05$.

Subgroup type	ρ per engineer	ρ per mix
Any	.62	.32
With EQ	.67	.40
With DRC	.45	.35
With EQ & DRC	.59	.38

These results imply that the higher the number of subgroups an engineer typically creates, the higher the preference ratings they receive. The effect is even stronger when considering only subgroups with EQ processing applied to them, suggesting it is an important and effective mix technique. A similar but weaker trend can be observed with regard to dynamic range compression. When considering each mix individually, the correlation is markedly weaker, too.

4.2.4 Conclusion

Preference ratings of different mixes show there is a strong tendency for engineers to like their own mix better, possibly because of personal preferences that affect both the mixing process and the assessment of other mixes. Furthermore, mixes from the same engineer are likely to receive a similar rating, suggesting a consistent overall performance across mixes of different songs.

Studying the correlation between these preference ratings and features extracted from the mix, it appears some features can help predict the preference for this mix. Specifically, the relations demonstrated here point to very concrete, practical issues mixes may have, such as a limited dynamic range, or a weak centre stage in the stereo image. However, no spectral features were found to correlate with preference, in contrast to e.g. [66, 72].

Further work is required to understand exactly how objective features relate to preference for musical stimuli. The present work can be expanded by looking at extracted features of the different tracks, and relations between different tracks, to further understand what effect different mix actions have. The correlation with preference may also be stronger for more sophisticated, perceptually motivated features, or a combination of the features above. As the difference between the mixes can be subtle, the current dataset may not span a large enough range in the various feature dimensions to learn about their influence on perception — for instance, even if there is a universal dislike for mixes with low ‘brightness’, this can only be measured if examples of both high and low ‘brightness’ mixes are evaluated.

Correlation analysis is limited in the sense that it only shows a general, unidirectional trend, and provides no information about a potentially favourable ‘middle ground’ in the provided data. Detailed analysis is required to establish more definitive tendencies.

While some mixes are clearly preferred over others, no obvious ranking emerged from the subjective ratings. This can be due to differences in taste. However, it is also probable a mix has several positive and negative attributes, which are not conveyed through a one-dimensional preference scale. Analysis of comments on the different

mixes can help zoom in on specific processors and instruments.

Another shortcoming of the presented approach is that spurious correlation is bound to occur if an increasingly large number of features are analysed. Although the correlations found here are quite strong, any interpretation is speculative unless the relationship is confirmed by corresponding comments.

4.3 Subjective free-form description

Ratings or rankings of different mixes can indirectly indicate which mix qualities or settings are likely detrimental or favourable, but it requires a large amount of data to reliably discover a direct relation between preference and a complex objective feature. In contrast, a limited number of short reviews by experienced listeners can show a general dislike for an excessive reverb on the lead vocal, an overall lack of low frequency content, or a poorly controlled bass line. More generally, by collecting assessments for a number of mixes of a number of songs, it becomes possible to discover overall tendencies in perception of mix engineering, and the relative influence of genre and instrumentation on these tendencies. Comments accompanying mix ratings can further reveal what type of processing or instruments are likely to draw attention in a poor or excellent mix, help find examples of a good or bad treatment of a particular instrument, and expose differences in perception between listeners of varying degrees of expertise.

In this section, the comments from the perceptual evaluation experiment described in Section 3.2 are studied. Initial analysis of these annotated comments allows quantifying focus on different instruments and processing, and the ratio between positive and negative comments. Furthermore, challenges associated with the interpretation of comments are explored and, where possible, solutions are proposed to facilitate in-depth analysis.

The double-blind reviews, in the form of a compilation of anonymous comments for each engineer, provide a unique type of feedback that is especially useful in an educational setting. This has been an important stimulus for educators and students to get involved and contribute the valuable data studied here. Through an informal survey, educators from seven institutions in five countries confirmed that this type of detailed evaluation is insightful, and unlike any form of conventional assessment where generally only a teacher comments on a student's mix. On the other hand, subjective evaluation participants enjoy an interesting critical listening exercise that also has educational value. By making the tools and multitrack material available to the public, other institutions are able to use this approach for evaluating recording and mixing exercises as well as practising critical listening skills.

4.3.1 Thematic analysis

In total, 1326 comments were collected from 1498 mix evaluations: nine to ten mixes of ten songs evaluated by between 13 and 22 trained listeners.

These comments are a sequence of atomic *statements*, critiquing or praising a particular instrument (or the whole mix) and an aspect of its processing. Each statement is labelled as referring to a certain instrument or group of instruments (vocals, drums, bass, guitars, keyboards, or the mix as a whole) and a certain processor or feature (balance, space, spectrum, dynamics), as well as classified as ‘positive’, ‘negative’, or ‘neutral’. The drums are split up into ‘kick drum’, ‘snare drum’, ‘cymbals’, or the drums in general, and the space-related mix features into panning, reverb, and other.

For instance, the comment

“Drums a little distant. Vox a little hot. Lower midrange feels a little hollow, otherwise pretty good.”

consists of the four separate statements “Drums a little distant.”, “Vox a little hot.”, “Lower midrange feels a little hollow”, and “otherwise pretty good.”. The first statement relates to the instrument group ‘drums’ and the sonic feature group ‘space’, and is a ‘negative’ comment or criticism. The second statement is labelled as ‘vocals’, ‘level³’, and ‘negative’. The third pertains to the spectral properties of the mix in general (negative) and the fourth is a general, positive remark, again about the whole mix.

The 1326 comments thus resulted in a total of 4227 statements. On average, one comment consisted of 3.2 ± 1.8 statements (median 3). The maximum number of statements within one comment was 11.

As shown in Figure 4.8, 33% of the statements were about the mix in general (or an undefined subset of instruments), 31% regarded vocals (lead or backing), 19% were related to drums and percussion, 7% to guitar, 6% to bass, and 4% to keyboard instruments. Within the drums and percussion category, 24% referred specifically to the snare drum, 22% to the kick drum, and 4% to the hi-hat or other cymbals. As in Chapter 2, it can be inferred that in the considered genres the treatment of the vocals

³‘Hot’ means ‘high in level’, from electronic engineering jargon [217].

is crucial for the overall perception of the production, as trained listeners clearly listen for it and comment on positive and negative aspects in roughly a third of all statements made.

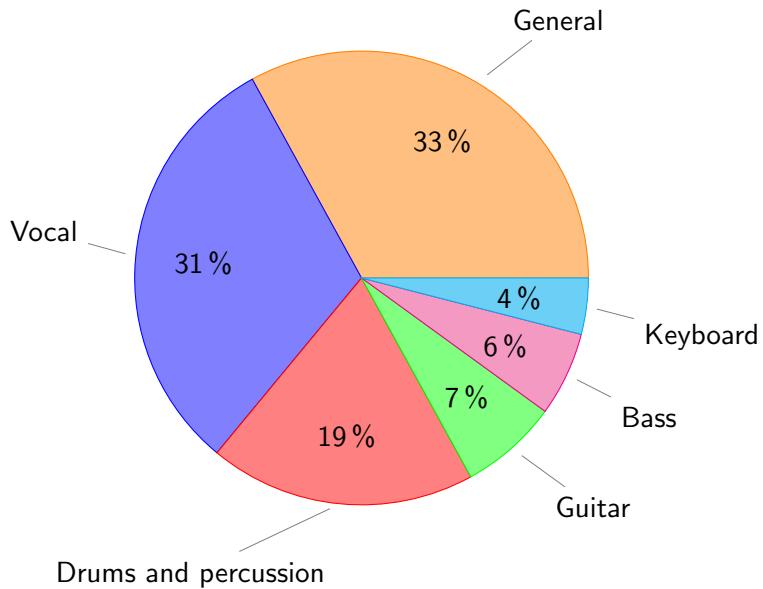


Figure 4.8: Representation of instrument groups across statements

Figure 4.9 further shows 35% of all statements concerned levels or balance, 29% space, 25% spectral qualities, and 11% dynamic processing (including automation and dynamic range compression). This again confirms more informal observations in Chapter 2, where level was cited as a strong influence on mix perception by all subjects, and spatial aspects by most. Within the category ‘space’, 58% of the statements were related to reverb, and 16% to panning.

Three out of four statements were some form of criticism on the mix. Of the 23% positive statements, many were more general (“good balance”, “otherwise nice”, “vocals sound great”). In the remaining cases, the statement had no clear positive or negative implication. The difference between the number of positive and negative comments showed some correlation (Spearman’s rank correlation coefficient $\rho = .20$) with the numeric preference rating values, meaning a relatively high proportion of negative comments indicates a higher probability the mix was less preferred by this subject.

Finally, Table 4.8 lists the 30 most frequently occurring descriptive terms across all comments. Derivations of the same root have been collapsed to one word. Other common words include forms and synonyms of ‘vocals’ (constituting 8% of all words),

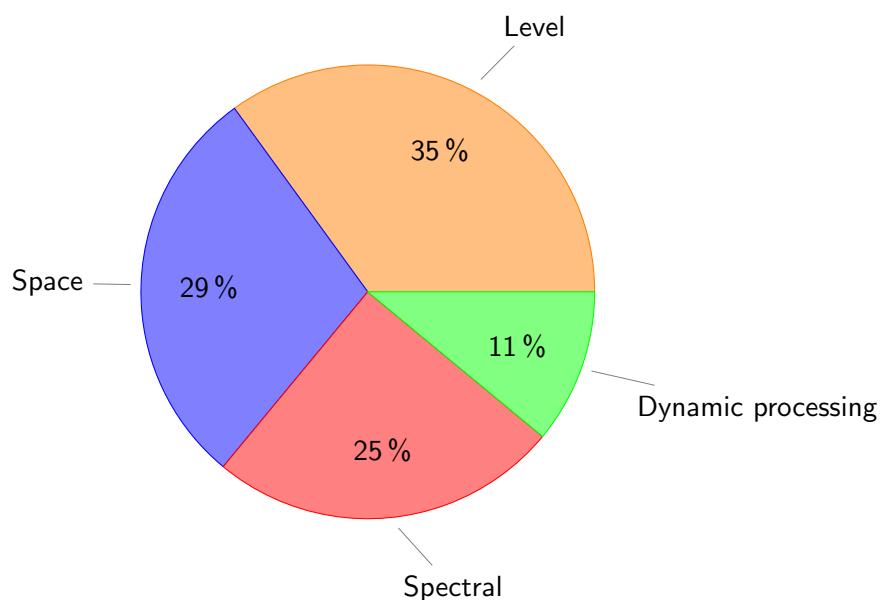


Figure 4.9: Representation of processors/features across statements

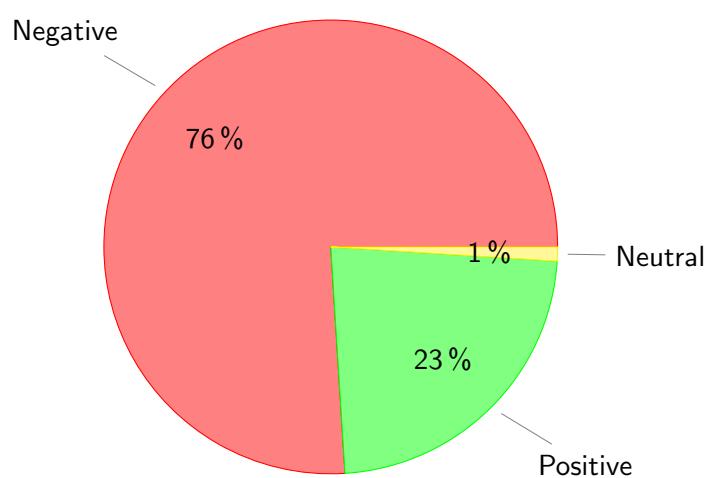


Figure 4.10: Proportion of negative, positive, and neutral statements

‘reverb’ (3%), ‘drums’ (3%), and ‘balance’ (2%).

Table 4.8: Top 25 most occurring descriptive terms over all comments

	Term	#	%
1	loud	234	1.227%
2	dry	115	0.603%
3	bright	99	0.519%
4	thin	89	0.467%
5	dark	79	0.414%
6	weird	77	0.404%
7	compressed	49	0.257%
8	present	47	0.246%
9	punch	47	0.246%
10	soft	46	0.241%
11	far	44	0.231%
12	muddy	44	0.231%
13	wide	44	0.231%
14	harsh	41	0.215%
15	room	41	0.215%
16	quiet	39	0.204%
17	hot	32	0.168%
18	clear	31	0.163%
19	big	30	0.157%
20	close	29	0.152%
21	mono	29	0.152%
22	defined	27	0.141%
23	cool	27	0.141%
24	strange	24	0.126%
25	forward	21	0.110%
26	heavy	21	0.110%
27	narrow	21	0.110%
28	small	21	0.110%
29	weak	20	0.105%
30	pumping	17	0.089%

4.3.2 Challenges

While the annotation of these comments is usually relatively straightforward, there were a number of cases where interpretation was more difficult. In this section the different types of challenges are discussed.

Which processor or sonic feature does this relate to?

The main challenge with interpreting the comments in this study is that it is often unclear to what processors or objective features the comment relates. Because of the multidimensionality of the complex mix problem, many perceived issues can be attributed to a variety of processors or signal properties.

This is further complicated by the subjects' use of semantic terms to describe sound or treatment which do not have an established relationship with sonic features or processor parameters — even if they are agreed upon by many subjects assessing the same mix and frequently used in a sound engineering context.

“Drums are flat and lifeless, no punch at all.”

“Snare is clear, but kick is lugubrious...”

“Too much ‘poof’ in kick. Not enough ‘crack’ in snare.”

“Thinking if it had a bit more ‘oomf’ in the lows it would be great.”

“Punchy drums.”

“I need some more meat from that snare.”

The term *present*, which could relate to level, reverb, EQ, dynamic processing, and more [54], is but one example of this.

“Electric guitar is a little too present.”

“Vox nice and present.”

“Hi-hat too present.”

“Lead vocals sometimes not present enough, other times too present.”

Some terms are associated with a lack, presence, or excess of energy in a certain frequency band in mixing handbooks, but even then this is not rigorously investigated, and the corresponding frequency range varies between and even within sources, see Table 2.3.

“*Vocal a little thick.*”

“*Piano a little muddy.*”

“*Kick is a bit tubby sometimes.*”

“*Drums sound a little thin to me.*”

“*Very bright guitars.*”

“*Vocal sounds dull.*”

“*Guitars have no bite.*”

“*Bass is dark.*”

“*Nasal vocals.*”

“*Guitars are woofy and too dark.*”

However, this usage of descriptive terms presents an opportunity to define them, when paired with low-level, objective features of the corresponding tracks or mixes.

Some statements are more generic and offer even less information on which of the mix properties, instruments, or processing the subject was listening for or (dis)pleased by.

“*Nice balance.*”

“*Best mix.*”

“*Lead vocal sounds good.*”

“*Nice vocal treatment.*”

“*Bad balance.*”

“*Guitars sound horrible.*”

“*This is awful.*”

“Everything is completely wrong.”

On the other hand, such a general assessment of a complete mix, a certain aspect of the mix, or the treatment of a specific instrument can be useful when looking for examples of appropriate or poor processing.

Good or bad?

In other instances, it is not clear whether a statement is meant as positive (highlighting a strength), negative (criticising a poor decision), or neither (neutral).

“Pretty dry.”

“Lots of space.”

“Round mix.”

“Wide imaging.”

“Big vocals.”

“This mix kind of sounds like Steely Dan.”

Fortunately, many of these can be better understood by considering other comments of the same person (if a similar statement, or its opposite, was made about a different mix of the same song, and had a clear positive or negative connotation), other statements in the same comment (e.g. two statements separated by a conjunction like ‘but’ will mostly be a positive and a negative one), comments by other subjects on the same mix (who may express themselves more clearly and remark similar things about the mix), or the rating attributed to the corresponding mix by the subject (e.g. if the mix received one of the lowest ratings from the mix, the comment associated with it will most likely consist of mostly negative statements).

Another statement category consisted of mentions of allegedly bold decisions, that the subject condoned or approved of despite sounding unusual.

“A lot of reverb but kind of pulling it off.”

“Horns a bit hot, but I kind of like it except in the swells.”

“Hated the vocal effect but in the end got used to it, nice one.”

“Most reverb on the vocals thus far, but I like it.”

This highlights a potential bias in comparative perceptual evaluation studies to reveal ‘best practices’ in mixing: there may be a tendency towards (or away from) more conventional sounding mixes and more mundane artistic decisions when several versions are judged simultaneously. Commercial music is typically available as one mix only, meaning bold mix moves may not be questioned to the same extent by the listener. Indeed, in [71] the outliers in terms of spectral and dynamic features are not rated highly — though likely because they are genuinely poor mixes even when auditioned in isolation.

In the context of an ‘acceptable mix space’, bounded by ranges of suitable parameter or feature values, these extreme settings could be considered outliers. This appreciation of unconventional mix practices once again underlines the creative nature of mixing, and suggests understanding, predicting, and emulating it entirely is a hard or perhaps impossible task.

Cryptic comments

It takes at least a basic background in music production to interpret the following statements.

“Kick has no punch.”

“Lots of drum spots.”

“Vocals too wet.”

A sound engineer will know to connect the use of the word *punchy* to the dynamic features of the signal [64], that ‘spots’ refers to microphones at close distance [218], and that the term *wet* is used here to denote an excessive amount of reverberation [219].

On the other hand, some comments are hard to understand even with years of audio engineering expertise, possibly because the subject forgot to complete the sentence, or because they are meant mainly to remind the subject which mix was which.

“Vocals.”

“Reverb.”

“Get the music.”

Scaling these experiments to substantially higher numbers of evaluations could prompt automated processing of comments, using natural language processing (NLP) or similar. However, due to the lack of constraints, many comments are near impossible to interpret by a machine. In the following cases, it would be challenging at best to automatically and reliably extract instrument, process or feature, and whether the statement is meant as criticism or highlighting a strength, especially when humour is involved:

“Why is the singer in the bathroom?”

“Where are the drums? 1 800 drums? Long distance please come home in time for dinner...”

“Is this a drum solo album instead of a lead female group?”

“Do you hate high frequencies?”

“Lead vocal, bass, and drum room does not a mix make.”

“No bass. No kick. No like.”

“If that was not made by a robot, that person has no soul.”

It would also take an advanced algorithm to understand these speculations about the mix engineer’s main instrument, suggesting the high level of these instruments is caused by the engineer’s bias towards their own instrument:

“Sounds like drummer mixed it...”

“Mixed by a drummer?”

“Guitar player’s mix?”

or the following comic book references (each from a different participant):

“Holy hi-hat!”

“Holy high end Batman!”

“Holy reverb, Batman!”

“Holy noise floor & drum compression!”

As all subjects were affiliated with the same institution, it is also likely that such a particular turn of phrase was shared among students or taught by teachers custom, serving as a reminder of the potential bias and limited generality of the findings.

At this point, it seems a trade-off has to be made between processing large amounts of machine-readable feedback, by imposing constraints on the feedback, or a free form text field so as not to interrupt or bias the subject's train of thought. If feedback were collected with a limited vocabulary (for instance borrowing from the Audio Effects Ontology [91], Music Ontology [124], and Studio Ontology [123]), or via user interface elements such as checkboxes and sliders instead of text fields, almost effortless acquisition of unambiguous information on the processing of different sources in different mixes would be possible. This data could then readily be processed without the need for manual annotation. On the other hand, studying free-form text feedback allows learning how listeners naturally react to differences in music production, and even what exactly these ill-defined terms and expressions mean and how they relate to different aspects of the mix. Which approach to choose therefore has to be informed by the research questions at hand. As both approaches have merit, and few attempts have been made in either direction, they should each be pursued.

4.3.3 Conclusion

Over 4200 statements describing different aspects of the mixes were annotated and the distribution of references to instruments, processors, and sonic features was studied. This data allowed quantification of the attention paid to different instruments, types of processing, and categories of features. Most of the statements were criticising aspects of the mix rather than praising them. Some challenges in the interpretation of these statements were considered and, where possible, solutions were proposed.

The main challenge when deriving meaningful information about the mix from its reviews, is to understand to which process or objective feature a statement relates. The wealth of subjective terms used in the assessments of mixes is an important obstacle in this regard.

Furthermore, reliably inferring whether a short review is meant as positive or negative is not always possible. However, considering numerical rating or ranking of the same

mix as well as comments by others on the same mix, or by the subject on other mixes, often provides additional insight in this matter. Interestingly, some unconventional or daring mix decisions were praised, suggesting outliers are not necessarily disliked, and the mixing problem is likely a complex one with several local optima.

Finally, due to the rich vocabulary and at times cryptic expressions used to describe various aspects of the mix, the tedious annotation process could only be automated if feedback were more constrained. Alternatively, translation of the free-form text responses into actionable rules and trends requires a better understanding of sound-related words.

In the following sections, a scalable approach to defining subjective terms in a multi-track music production context is developed, and mixing knowledge is produced from annotated comments combined with extracted low-level features, respectively.

4.4 Real-time attribute elicitation

The analysis and evaluation of real-world mixes offers a unique perspective on music production practices and their impact on perception. Unconstrained feedback allows objective correlates of the typical descriptors used to communicate sonic concepts in a sound engineering context to be defined. However, because of the time and effort required to conduct these controlled tests, the approach is only moderately scalable.

To address this, a novel data collection architecture was developed for the elicitation of semantic descriptions of musical timbre, deployed within the digital audio workstation. By embedding the data capture system into the music production workflow, the return of semantically annotated music production data is maximised, whilst mitigating against issues such as musical and environmental bias. Users of freely downloadable DAW plugins are able to submit semantic descriptions of their own music, whilst utilising the continually growing collaborative dataset of musical descriptors. In order to provide more contextually representative timbral transformations, the dataset is partitioned using metadata, obtained within the application.

Each plugin consists of a standard interface augmented with a free-text field, allowing input of one or more text labels. As the descriptors are entered, they are uploaded anonymously to the server along with a time-series matrix of audio features extracted both pre- and post-processing, a static parameter space vector, and a selection of metadata tags. To motivate the user base to provide this data, semantic profiles can also be loaded from the server, setting the parameters automatically based on accumulated knowledge, current audio features, and metadata (see Figure 4.11).

4.4.1 System

Digital audio effects

Four audio effect plugins have been implemented in VST, Audio Unit, and LV2 formats: an amplitude distortion effect with tone control, an algorithmic reverb based on the figure-of-eight technique proposed by Dattorro [222], a dynamic range compressor with variable threshold layout and attack and release parameters, and a parametric EQ with

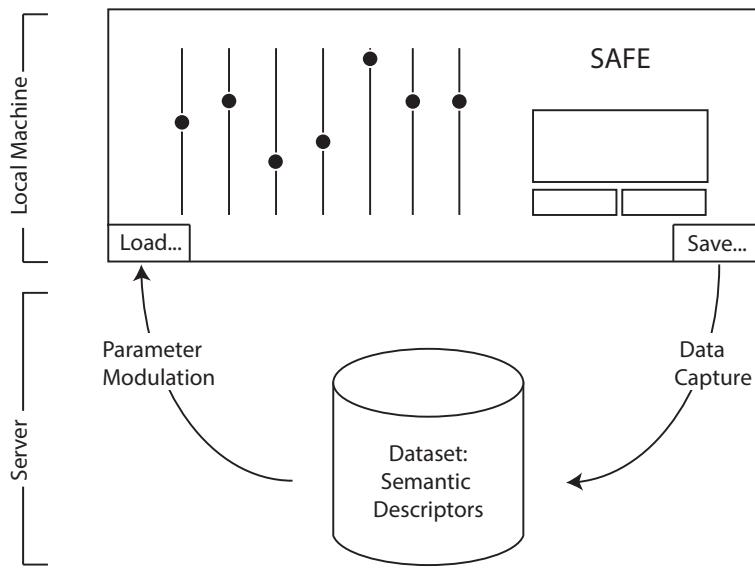


Figure 4.11: A schematic representation of the plugin architecture, providing users with load and save functionality

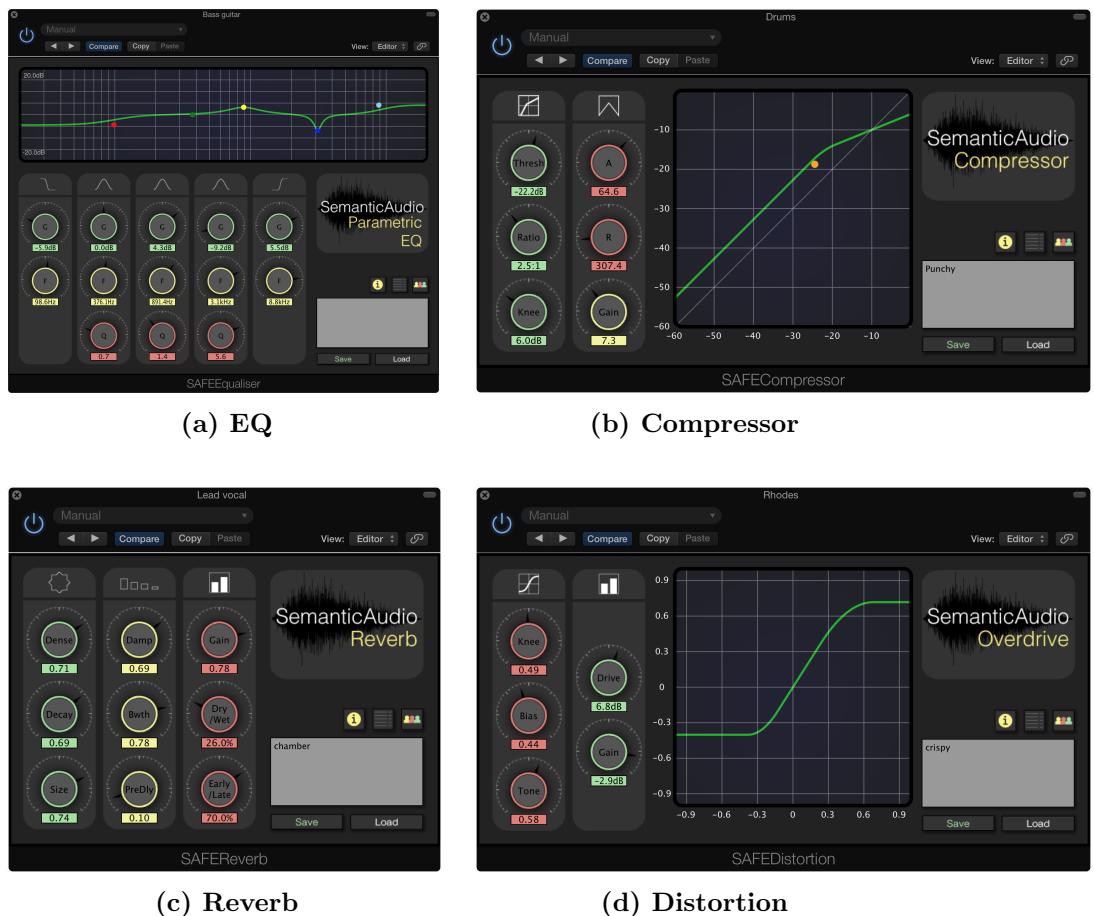


Figure 4.12: Graphical user interfaces of the equalisation, compression, and reverberation plugins

three peaking filters and two shelving filters. All visible parameters are included in the parameter space vector, and can be modulated via the text input field. The plugins can be downloaded from www.semanticaudio.co.uk/projects/download/. Their GUIs are shown in Figure 4.12.

To encourage third party expansion of the set of processors or integration of the presented functionality in existing software, a plugin template⁴ was published.

Data collection

In addition to traditional controls and visualisation, the plugin interface features a text box which allows the user to describe the perceived effect of the processor with the current parameter settings. To store one or more term descriptors, the user is prompted to play a representative section of the audio, and click ‘Save’ to record an excerpt of the audio spanning a few seconds.

To characterise the signal associated with each descriptor, an $N \times M$ matrix of audio features is stored, where N is the number of recorded frames and M is the number of audio features. These are extracted using the libxtract library [223], an audio feature extraction framework. Here $M = 85$ different features are considered, taken from 10 different input representations, see Table 4.9. To capture the timbral transformation imposed by the audio effect, the feature matrix is computed before and after the processing occurs, and differential measurements are taken.

Along with the feature matrix, a $1 \times P$ parameter vector is stored, where P is the number of UI parameters. In the current implementation, P ranges from 6 to 13. Furthermore, an optional metadata window is provided to store user and context information, see Figure 4.13a. This metadata currently consists of the user’s age, location, and production experience, the genre of the song, and musical instrument of the track, as these were deemed to be potentially significant factors explaining the variance of semantic terminology.

⁴github.com/semanticaudio/SAFE

Table 4.9: Features extracted from the audio before and after processing

Category	Feature
Time domain	Mean
	Variance
	Standard Deviation
	RMS amplitude
	Zero crossing rate
Spectral	Spectrum
	Centroid
	Variance
	Standard deviation
	Skewness
	Kurtosis
	Irregularity J
	Irregularity K
	Fundamental (f_0)
	Smoothness
	Rolloff
	Flatness
	Tonality
	Crest
	Slope
Peak spectral	Spectrum
	Centroid
	Variance
	Standard Deviation
	Skewness
	Kurtosis
	Irregularity J
	Irregularity K
	Tristimulus (1s, 2s, 3s)
	Inharmonicity
Harmonic spectral	Spectrum
	Centroid
	Variance
	Standard Deviation
	Skewness
	Kurtosis
	Irregularity J
	Irregularity K
	Tristimulus (1 s, 2 s, 3 s)
	Non-zero count
	Noisiness
Other	Parity ratio
	Bark coefficients (25)
	MFCCs (13)

Parameter modulation

Users can modulate the processor's parameters by searching for existing descriptors, loading associated semantic profiles from the server, and applying them to their own audio signals, see Figure 4.13b. Each semantic profile is updated in real-time, meaning they change dynamically based on new input to the server. To provide users with a more reliable representation of their semantic term, the terms are hierarchically partitioned into metadata categories when they are available. This allows users to load instrument-, genre-, and location-specific terms, as opposed to generic terms that cover a wide range of musical conditions. Additionally, transformations from nonlinear effects are applied relative to the signal's RMS to ensure timbral modifications are applied independently of signal level. Awaiting further data collection and analysis, the current implementation simply loads an average of the parameter settings associated with the chosen term.

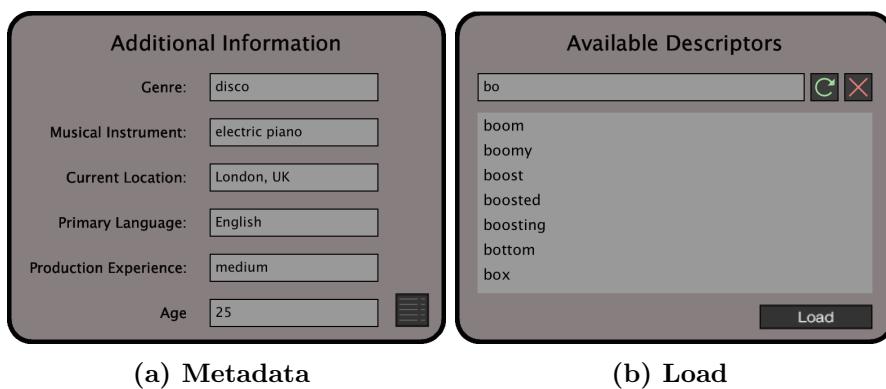


Figure 4.13: ‘Metadata’ and ‘Load’ dialog boxes within the plugins

Missing data approximation

Users frequently omit metadata tags, providing only audio data, the parameter space, and text descriptors. In these cases, missing data can be approximated using a number of techniques, thus improving the reliability of the semantic parameter settings. The user's location can be approximated from geolocation data relating to the IP address, and musical instrument and genre tags are estimated using an unsupervised machine learning algorithm, applied to a reduced-dimensionality representation of the audio feature set.

4.4.2 Term analysis

Statistics

The dataset considered here comprises 2694 transforms, split into four groups according to their transform class (processor). Overall, 454 were applied using the compressor, 303 using distortion, 1679 using the equaliser, and 258 using reverb. These transformations were described using 618 unique terms taken from 263 unique users, all of whom were music producers who participated by using these plugins within their workflow. Already, this data clearly surpasses the terms extracted from the mix comments in both size and diversity.

To group terms with shared meanings and variable suffixes, a Porter Stemmer [224] reduces them to their ‘stems’. This allows for the automated unification of terms such as *warm*, *warmer*, and *warmth* into parent category *warm*.

The *confidence* C_d of a descriptor d is defined as its average variance in feature space summed over all occurrences $n = \{1, \dots, N_d\}$, where each feature m is mapped to an M -dimensional space using Principal Component Analysis (PCA) in order to remove redundancy, whilst retaining $\geq 95\%$ of the variance with $M = 6$:

$$C_d = \sum_n^{N_d} \frac{1}{M} \sum_m^M (PC_{nm} - \mu_m)^2 \quad (4.3)$$

where PC_{nm} is the m^{th} principal component corresponding to the n^{th} occurrence of descriptor d , and μ_m is the mean of PC_{nm} over all occurrences n .

To quantify the *popularity* P_d of a descriptor, Equation (4.3) is then weighted with the logarithm of the relative occurrence of the descriptor in the given dataset:

$$P_d = \frac{\ln N_d}{C_d} \quad (4.4)$$

Finally, *generality* G_d measures the extent to which the descriptor is applicable across a range of transform classes, and is defined as the centroid of the density function over

transform classes sorted in decreasing order of number of occurrences.

$$G_d = \frac{2}{K-1} \sum_{k=0}^{K-1} k \operatorname{sort} \left(\frac{N_{dk}}{N_d} \right) \quad (4.5)$$

where N_{dk} is the number of occurrences of descriptor d in class $k \in [0, K-1]$. Figure 4.14 visualises the calculation of the generality of descriptor *thick*.

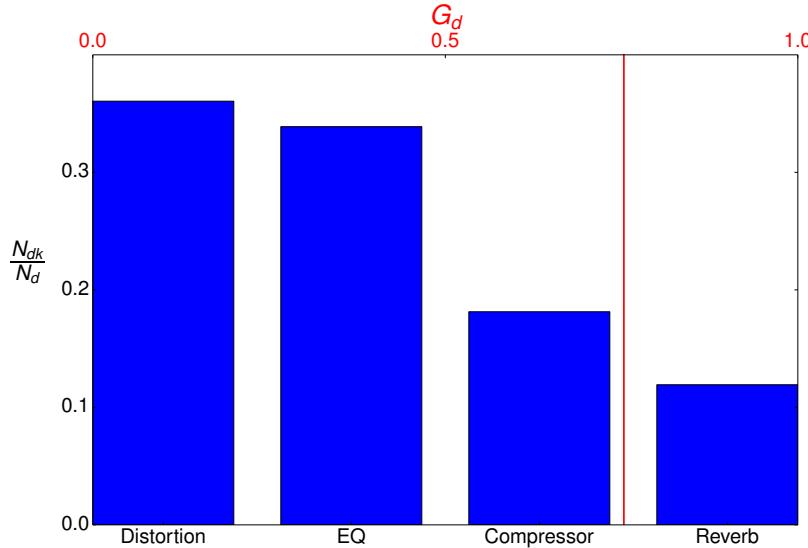


Figure 4.14: Generality of descriptor *thick*

Using these metrics, the database is sorted and the top 10 descriptors are shown in Table 4.10. Of the most often occurring terms, *bright*, *punch*, *room*, and *clear* are also extensively used in the subjective evaluation of mixes (see Table 4.8). However, only six occurrences of *warm* and *smooth*, and two of *air* and *crunch* were registered in the previous analysis. *Fuzz*, associated with distortion, was not used at all, likely because this creative effect was not among the tools used for the creation of mixes.

Table 4.11 shows the most commonly used descriptors for each individual transform class.

Feature space representation

Hierarchical clustering can be applied to differences between unprocessed and processed signals in feature space to find term similarities within transform classes. The mean of the audio feature vectors from each unique descriptor is computed and PCA is applied, reducing the number of dimensions, whilst preserving $\geq 95\%$ of the variance.

		# Instances	Confidence		Popularity		Generality	
1	warm	193	boxed	.250	warm	.0019	sharp	.828
2	bright	153	splash	.250	bright	.0014	deep	.819
3	punch	34	wholesome	.250	crunch	.0006	boom	.809
4	air	31	pumping	.247	room	.0005	thick	.806
5	crunch	29	rounded	.247	fuzz	.0004	piano	.696
6	room	28	sparkle	.247	crisp	.0004	strong	.596
7	smooth	22	atmosphere	.244	clear	.0004	soft	.575
8	vocal	21	balanced	.244	cut	.0004	bass	.555
9	clear	20	bass	.244	bass	.0004	gentle	.525
10	fuzz	19	basic	.244	low	.0004	tin	.483

Table 4.10: Terms ranking highest in number of instances N_d , confidence C_d , popularity P_d , and generality G_d

Compressor	Distortion	EQ	Reverb
27: punch	23: crunch	155: warm	30: room
17: smooth	20: warm	144: bright	13: air
15: sofa	6: fuzz	16: air	11: big
14: vocal	6: destroyed	16: clear	10: subtle
12: nice	5: cream	12: thin	9: hall
9: controlled	5: death	11: clean	9: small
9: together	5: bass	11: crisp	8: dream
9: crushed	5: clip	10: bass	7: damp
8: warm	5: decimated	9: boom	7: drum
7: comp	5: distorted	9: cut	6: close

Table 4.11: The first ten descriptors per processor, ranked by number of entries N_{dk}

The resulting clusters, shown in Figure 4.15, are intended to retain perceived latent groupings, based on underlying semantic representations. Earlier studies produced similar visualisations using data from subjective evaluation only [156, 225, 226], instead of the proximity of feature values. Terms with less than eight entries are omitted for readability and the distances between datapoints are calculated using Ward distance [227].

From these term clusters, groups of semantically similar timbral descriptions emerge. Among the compressor terms, groups tend to exhibit correlation with the extent to which gain reduction is applied to the signal. *Loud*, *fat*, and *squashed* generally refer to extreme compression, whereas *subtle*, *gentle*, and *soft* typically describe minor adjustments to the amplitude envelope. Distortion features mainly group based on the perceived dissonance of the transform, with terms such as *fuzz* and *harsh* clearly separated from *subtle*, *rasp*, and *growl*. Equalisation comprises a wide selection of description-categories, although terms that refer to specific regions of spectral energy such as *bass*, *mid*, and *air* tend to fall into separate partitions. Finally, reverb term clusters seem to relate to size of the space and magnitude of the effect: *hall* and *room* occupy similar feature spaces, as do *soft*, *damp*, and *natural*.

The meaning of terms like *sofa* or *sorry*, if any, is unclear, demonstrating the limits of a system without constraints, control, or subsequent interaction with the subject.

Parameter space representation

Further illustrating the relevance of the within-class feature groups presented above, it can be shown that terms within clusters maintain similar characteristics in their parameter spaces. For instance, Figure 4.16 shows curves corresponding to two groups of descriptors taken from opposing clusters in the equaliser’s feature-space: one consisting of *warm*, *bass*, *boom*, *box*, and *vocal*, and one consisting of *thin*, *clean*, *cut*, *click*, and *tin*. Curves in the first cluster generally add emphasis around 500 Hz with a high-frequency roll-off, whereas those in the latter have a boost in high-frequency energy around 5 kHz and attenuated lows.

Next, the organisation of terms based on their position in a parameter space is evaluated, using PCA to reduce the dimensionality of each space and overlay the parameter

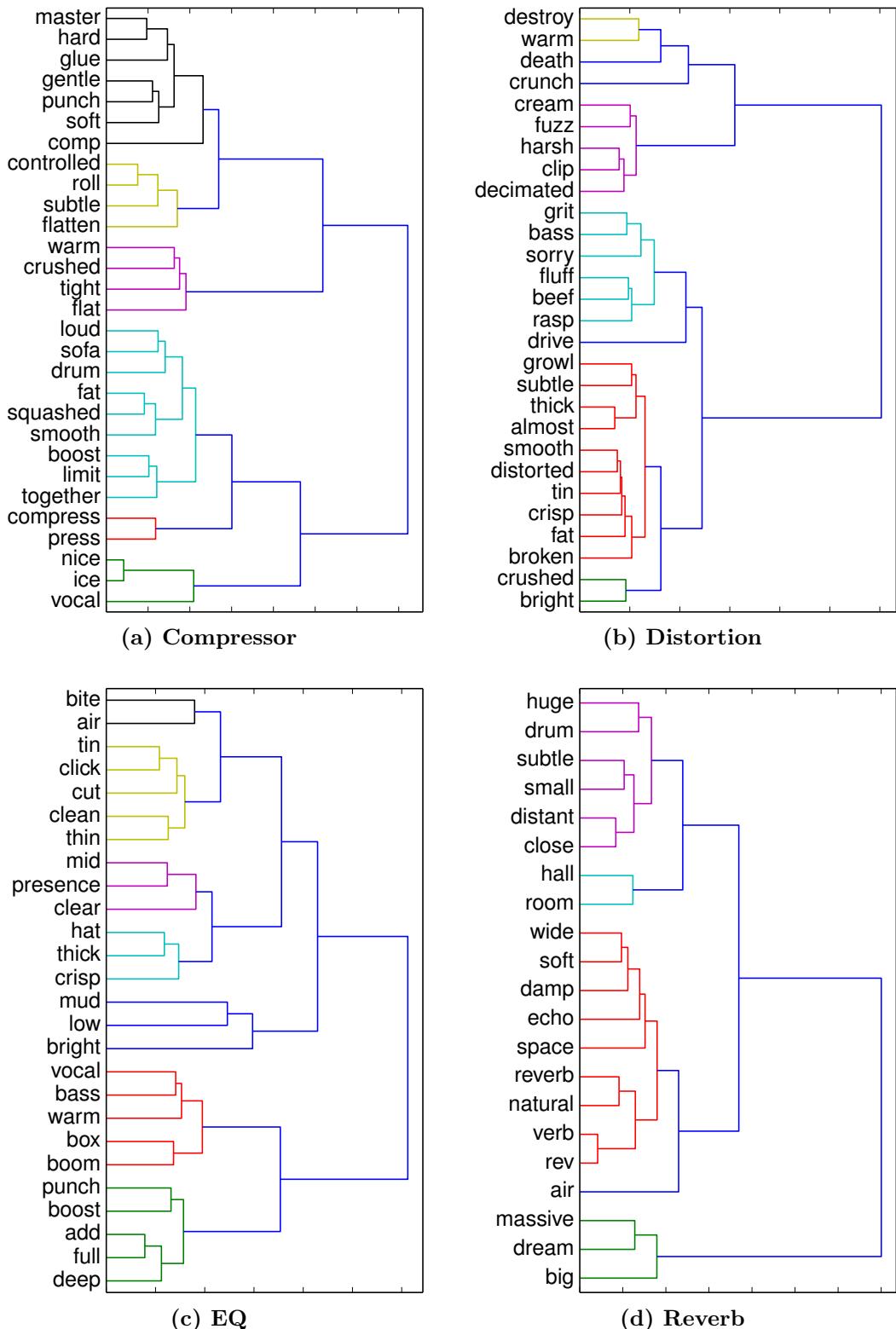


Figure 4.15: Dendrograms showing term clustering based on feature space distances for each transform class

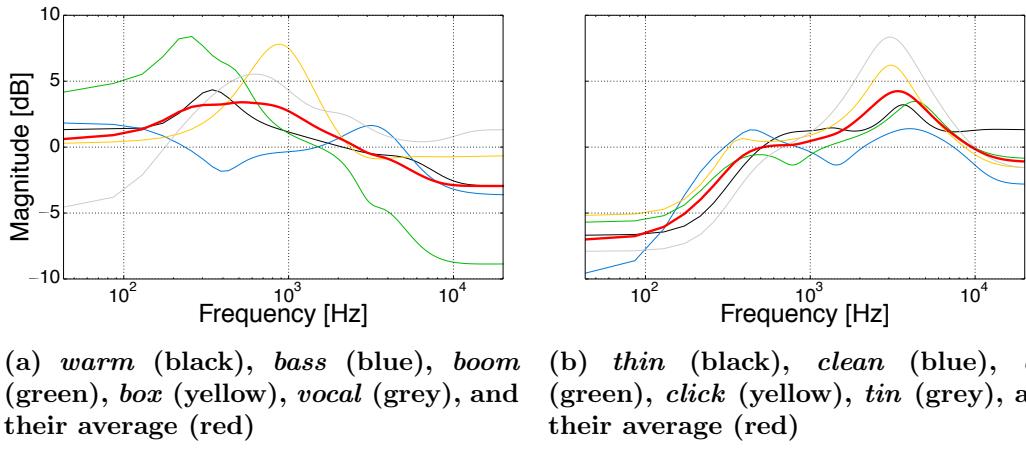


Figure 4.16: Equalisation curves for two clusters of terms in the dataset

vectors. Figure 4.17a shows this for distortion, where the bias is highly correlated with PC2, which tends to organise descriptors based on dissonance. Figure 4.17b shows the parameters of the reverb class where various parameter-term trends are apparent.

Term frequency analysis

Term similarity can also be measured independently of timbral or parameter space representations, using a term's association to a given transform class. Four-dimensional term frequency vectors define the distributions across classes, e.g. $\mathbf{t} = [0.0, 0.5, 0.5, 0.0]$ has equal association with the distortion and equaliser, but no entries in the compressor or reverb classes. Representing these using a Vector Space Model, the similarity between any two terms ($\mathbf{t}_1, \mathbf{t}_2$) is measured using cosine similarity:

$$\text{sim}(\mathbf{t}_1, \mathbf{t}_2) = \frac{\mathbf{t}_1 \cdot \mathbf{t}_2}{\|\mathbf{t}_1\| \|\mathbf{t}_2\|} = \frac{\sum_{i=1}^N t_{1,i} t_{2,i}}{\sqrt{\sum_{i=1}^N t_{1,i}^2} \sqrt{\sum_{i=1}^N t_{2,i}^2}} \quad (4.6)$$

In order to better capture the true semantic relations of the terms and the transforms they are associated with, Latent Semantic Indexing is applied [228]. This involves reducing the term-transform space from rank four to three by performing a singular value decomposition of the $N_{terms} \times 4$ occurrence matrix $\mathbf{M} = \mathbf{U}\Sigma\mathbf{V}^*$, and setting the smallest singular values to zero before reconstructing it using $\mathbf{M}' = \mathbf{U}\Sigma'\mathbf{V}^*$. This eliminates noise caused by differences in word usage, for instance due to synonymy and polysemy, whereas the ‘latent’ semantic relationships between terms and effects are preserved. Figure 4.18a shows the resulting pairwise similarities of the high-generality

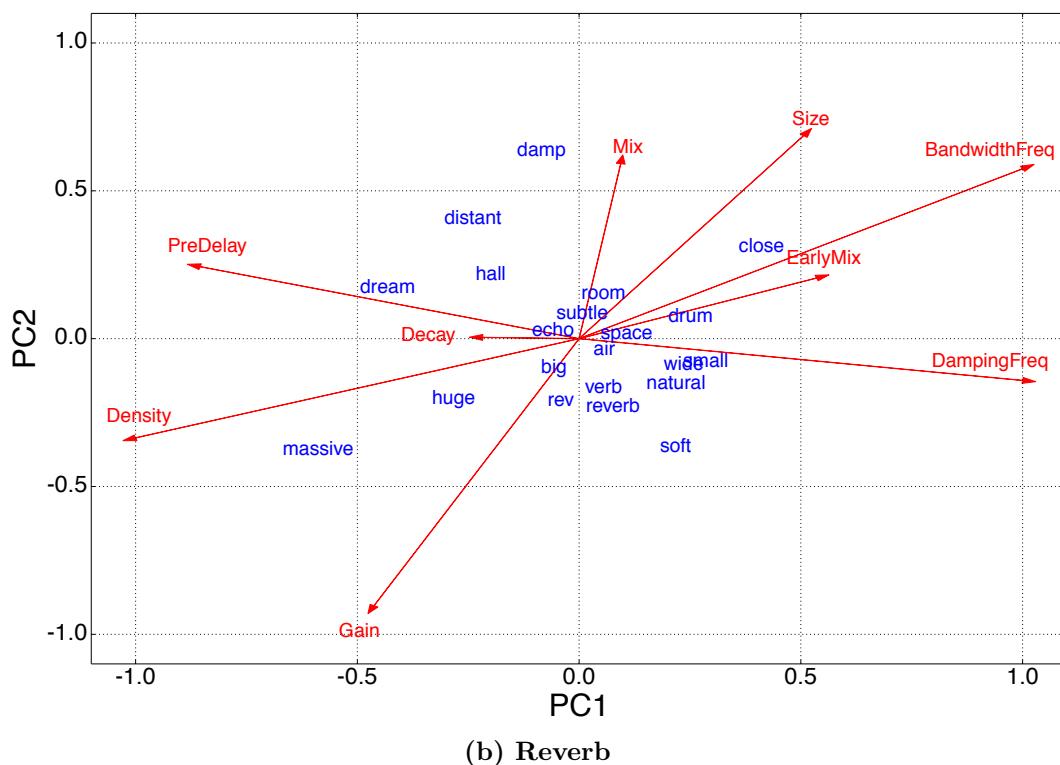
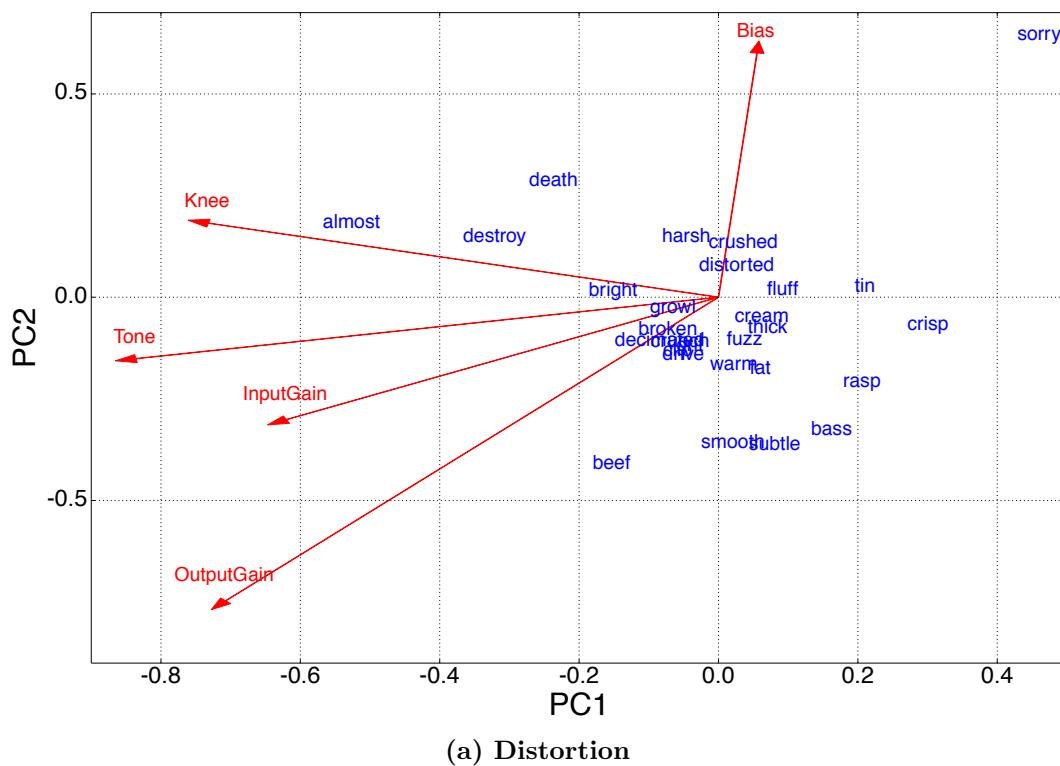


Figure 4.17: Biplots of the distortion and reverb classes, showing terms mapped onto two dimensions with overlaid parameter vectors

terms.

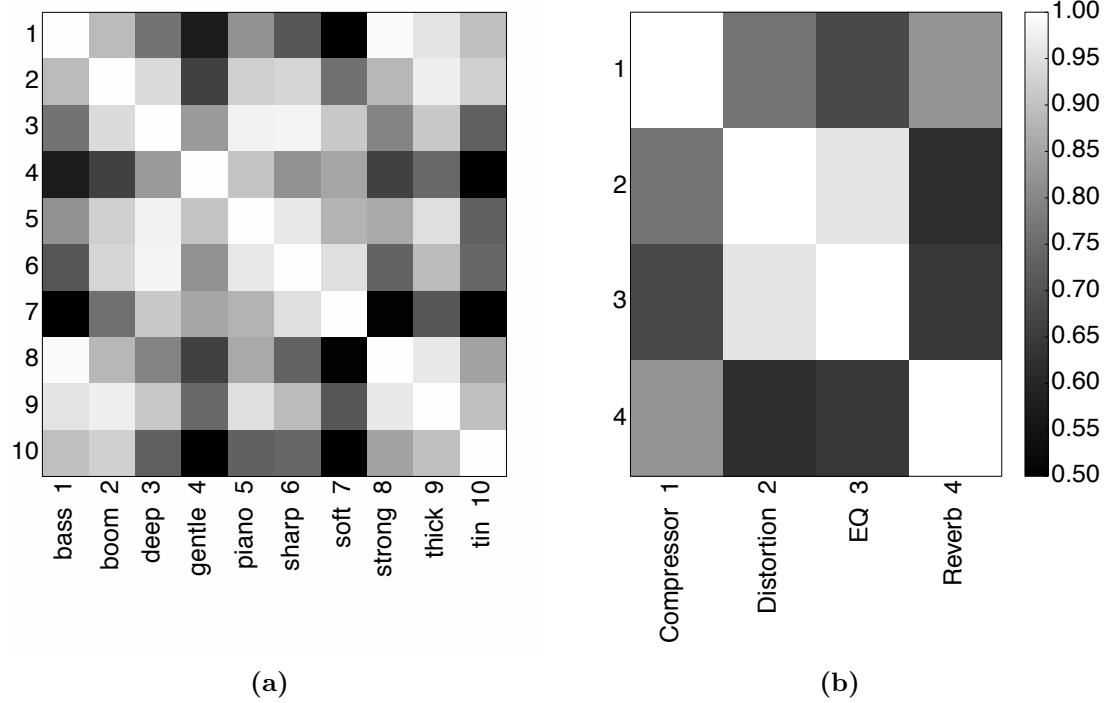


Figure 4.18: Vector-space similarity with regard to (a) high-generality terms and (b) transform-classes

Here, the most similar terms are *bass* and *strong*, *deep* and *sharp*, and *boom* and *thick* (all with a cosine similarity of 0.99).

Conversely, the similarity of transform types based on their descriptive attributes can be calculated by transposing the occurrence matrix in the VSM. This is illustrated in Figure 4.18b, showing terms used to describe equalisation transforms are similar to those associated with distortion (0.95), while equalisation and compression vocabulary is more disjunct (0.641).

4.4.3 Conclusion

The presented data collection architecture offers an effective way to gather objective correlates of user-defined sonic descriptors, from which definitions and semantic groupings of terms can be studied. Even though the descriptors can be chosen freely, the response format is constrained and therefore easy to analyse without the need for manual annotation. Because the elicitation itself is automated, from the experimenter's point of view, the concept is exceedingly scalable. Furthermore, participants are intrinsically motivated to contribute data, as the system proves useful as a software tool, compatible

with most DAWs, and its functionality is enhanced by entering descriptions.

On the other hand, only minimal control is retained as content producers are unknown and information about source content is reduced to a set of extracted features. This leads to noise and unexplained data points, possibly mitigated through large numbers of entries. In addition, despite being used in a realistic, multitrack music production environment, the system is inherently single track and unaware of any context. As of yet, commercial DAWs do not provide support for multitrack plugins, and extraction of data from within the mix session is limited to the features, parameters, and metadata pertaining to one source.

Analysis of the accumulated dataset has shown that meaningful within- and between-processor groupings of these semantic descriptions can be identified from associated feature and parameter data. Conversely, the amassed terms demonstrated how similar the processors themselves are to each other with regard to the vocabulary they share.

All anonymised user data and terms with related features and descriptors can be visualised and downloaded on www.semanticaudio.co.uk/datasets/. An API to access these datasets from within other applications is available on <https://github.com/semanicaudio/safe-api>.

Chapter 5

Multi-group analysis

There is evidence that the sound of recorded and mixed music can be influenced by the studio’s location and the engineer’s background, to a point where its origin can be reliably determined solely based on sonic properties — specifically for British and American recordings [86, 229, 230]. The same sources suggest that these differences have all but disappeared today due to the increased mobility of information, people, and equipment. Still, as the findings in Chapter 4 are based on subjects from a single institution, renowned as it may be, their ecological relevance can be questioned. For instance, it has been shown that definitions of sound-related adjectives vary between different countries [231, 232]. The effects of nationality or level of expertise, among others, have not been assessed with regard to mixing practices and their perception.

In this chapter, a range of experiments demonstrate the similarity of different mixes and their evaluations, based on a number of diverse datasets, produced and evaluated by participants from various countries and with varying degrees of audio engineering experience.

This cross-analysis also serves to prove the concept of the methodology developed in the previous chapters, and the presented tools (particularly the Open Multitrack Testbed and Web Audio Evaluation Tool) without which the extent of this study would not have been possible.

Finally, a selection of findings from the preceding chapters are tested, to confirm or challenge their validity beyond the initial experiment. This increases the significance

and relevance of the earlier conclusions if the findings are supported by these new datasets, and offers the potential to explore the influence of subject- or engineer-related factors if they differ.

Naturally, the ‘null hypothesis’ — i.e. the assumption that background does not have an effect on perception, preference, or the sonic signature of one’s mix — can only be disproved, namely when analysis of the various considered datasets leads to different results. Conversely, concurrence of the findings further strengthens the support for earlier conclusions, but does not rule out the possibility that data from other groups would contradict them.

5.1 Experiments

Additional experiments were organised at different institutions, in which mixes of both previous and newly added material were created and evaluated. In some instances, mixes from the initial experiment were evaluated as well. In each of the cases the participating educators selected materials from the Open Multitrack Testbed. The Web Audio Evaluation Tool facilitated the subjective evaluation and results collection procedure, sometimes without the author being present. With close to 5000 mix evaluations, it is by far the largest study of evaluated mixes known to the author. One other work analyses audio features extracted from a total of 1501 unevaluated mixes from 10 different songs [72]. A study by the same author examines audio features extracted from 101 mixes of the same song, evaluated by one person who classified the mixes in five preference categories [71]. In both cases, the mixes were created by anonymous visitors of the Mixing Secrets Free Multitrack Download Library, and principal component analysis preceded by outlier detection was employed to establish primary dimensions of variation. Parameter settings or individual processed stems were not available.

The main differences between the seven datasets are as follows:

McGill The initial dataset, studied in Chapter 4, consisting of students and lecturers from the MMus in Sound Recording programme at McGill University. Commenting on every mix was not enforced yet.

MG (*Perceptual evaluation only*) Employees from a Montréal-based startup, working on automatic music production tools. Commenting on every mix was not enforced yet. Beyerdynamic DT 770 PRO headphones were used (see Figure 2.6 for its frequency response). They were primarily amateurs with regard to music production, and one professional sound engineer.

QMUL (*Perceptual evaluation only*) Audio researchers from the Centre for Digital Music at Queen Mary University of London. All were amateurs with the exception of three professional sound engineers.

SMC (*Perceptual evaluation only*) Students from the MSc in Sound and Music Computing at Queen Mary University of London, a few weeks into a module on sound

engineering but otherwise amateur. These students used Audio Technica M50x headphones and participated simultaneously in a classroom. Only a minority had English as their first language but all but one (mixed Chinese/English) chose to answer in English. One subject's results were excluded from further analysis on account of not rating any fragments.

DU Sound engineering bachelor students at Dalarna University, Sweden. Mixes were produced in pairs, and in one case by a group of three, using an analogue Solid State Logic AWS900 console and analogue outboard equipment. Commenting on every mix was not enforced yet. As all were native Swedish speakers, comments were translated to English by the programme teacher who is a native Swedish speaker.

PXL Sound engineering bachelor students from the PXL University College music programme in Hasselt, Belgium, plus their teacher (an acclaimed professional sound engineer). As all were native Dutch speakers, comments were translated to English by the author who is a native Dutch speaker.

UCP Sound engineering bachelor students from the Universidade Católica Portuguesa in Porto, Portugal. As all were native Portuguese speakers, comments were translated to English by the programme teacher who is a native Portuguese speaker.

Table 5.1 lists the different sites where the evaluation experiments were conducted, and basic statistics associated with each of the resulting datasets.

Table 5.1: Overview of evaluation experiments

	McGill	MG	QMUL	SMC	DU	PXL	UCP	TOTAL
Country	Canada		United Kingdom		Sweden	Belgium	Portugal	
#subjects	33	8	21	26	39	13	10	150
#songs	10	4	13	14	3	4	7	18
#mixes	98	40	111	116	21	23	42	181
#evaluations	1444	310	1129	639	805	236	310	4873
#statements	4227	585	2403	1190	2331	909	1051	12696
#words/comment	13.39	11.76	11.32	12.39	18.95	31.94	25.21	15.25
Male/female	28/5	7/1	18/3	14/12	33/6	13/0	9/1	122/28

Unless noted otherwise, the listening tests took place in dedicated, high quality listening rooms at the respective institutions, using loudspeakers. The measured frequency responses, where available, are shown in Figure 5.1. It should be noted that differences between the respective rooms may affect the perception of spectral and temporal

properties of the stimuli.

All participants were given the choice to use their native language, but in Canada and the United Kingdom most non-native English speakers preferred to comment in English, supposedly because they were fluent, and more accustomed to describing sonic and musical properties in English. At the educators' request, all interfaces were in English with the exception of the Portuguese one.

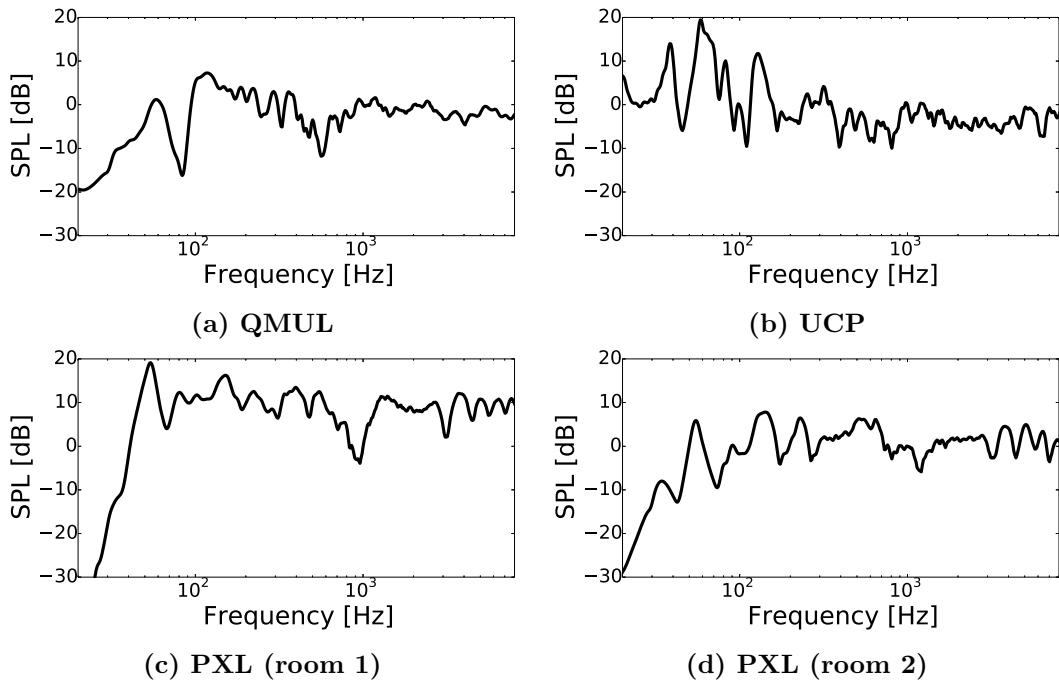


Figure 5.1: Combined frequency response of room and reproduction system, measured at listening position with only left speaker active, for all groups but DU (unavailable at present)

Table 5.2 lists the number of mixes of the different songs produced at the various sites, and the number of subjects assessing the respective mixes. With the exception of the previously considered McGill dataset, only students contributed mixes.

Lead Me and In The Meantime were mixed by mix engineers from all participating institutions, to accommodate comparison of mix practices and perception using the same source material. Each of these mixes were evaluated only by subjects from the institution where the mix was produced, with the exception of the McGill mixes studied in Chapter 4, which were also evaluated by subjects from the MG, QMUL, and SMC groups. In addition, the subjects at UCP evaluated their group's mixes alongside a selection of 5 McGill mixes.

To increase the diversity of the dataset, other mixes were created from new source

Table 5.2: Overview of mixed content, with number of mixes (left side) and number of subjects (right side) per song. Numbers between parentheses indicate (additional) mixes for which stems, DAW sessions, and parameter settings are not available (e.g. the original release, or analogue mixes). The most often occurring genre labels were determined post-hoc by surveying the perceptual evaluation participants. Evaluations with an asterisk (*) indicate that subjects included those who produced the mixes.

ARTIST	SONG	GENRE	NUMBER OF MIXES				NUMBER OF SUBJECTS						
			McGill	DU	PXL	UCP	McGill	MG	QMUL	SMC	DU	PXL	UCP
The DoneFors	Lead Me	country	8 (2)	(7)	(4)	5	15	8	10	4	39*	6*	10*
Fredy V	In The Meantime	funk	8 (1)	(7)	(7)	5	22*	10	5	38*	8*	10*	
Joshua Bell	My Funny Valentine	jazz	8 (2)				14	7	10	5			
Artist X	Song A	blues	8 (2)				14	8	10	9			
Dawn Langstroth	No Prize	jazz	8 (2)				14	8	10	5			
Fredy V	Not Alone	soul	8 (2)				13		10	5			
Broken Crank	Red To Blue	rock	8 (2)				13		10	4			
Artist Y	Song B	blues	8 (2)				14		10	5			
The DoneFors	Under A Covered Sky	pop	8 (2)				13		10	4			
The DoneFors	Pouring Room	indie	8 (1)				22*		9	6			
Torres	New Skin	indie		(7)					9	6	38*		
Filthybird	I'd Like To Know	pop rock		7					11	5	13*		
The Districts	Vermont	pop rock		2	5				11	5	13*		
Creepoid	Old Tree	indie rock				5						5	
Purling Hiss	Lolita	hard rock				5						5	
Louis Cressy Band	Good Time	rock				4						5	
Jokers, Jacks & Kings	Sea Of Leaves	pop rock				4						5	
Human Radio	You & Me & the Radio	pop rock				4						5	

material, made available by Weathervane Music’s Shaking Through and the Mixing Secrets Free Multitrack Download Library. As before, all metadata as well as the newly created mixes can be found on the Open Multitrack Testbed.

As some of the mixes at PXL and DU were created using a partly analogue setup, shown between parentheses in Table 5.2, recreating these sessions is tedious or impossible. Additional mixes from McGill (between parentheses) are the ‘professional’ mix and — for most songs — the machine-made mix which was not evaluated by subjects from QMUL and SMC. For these mixes, access to parameter settings or isolated tracks is unavailable as well.

In what follows, the considered factors are institution, level of expertise (amateur, student of a sound engineering programme, and professional sound engineer), and gender.

Usage and definition of terms is not formally investigated here because of the differences in native tongue, and the success of the tools presented in Section 4.4 in collecting data from a wide range of subjects across different countries.

5.2 Objective features

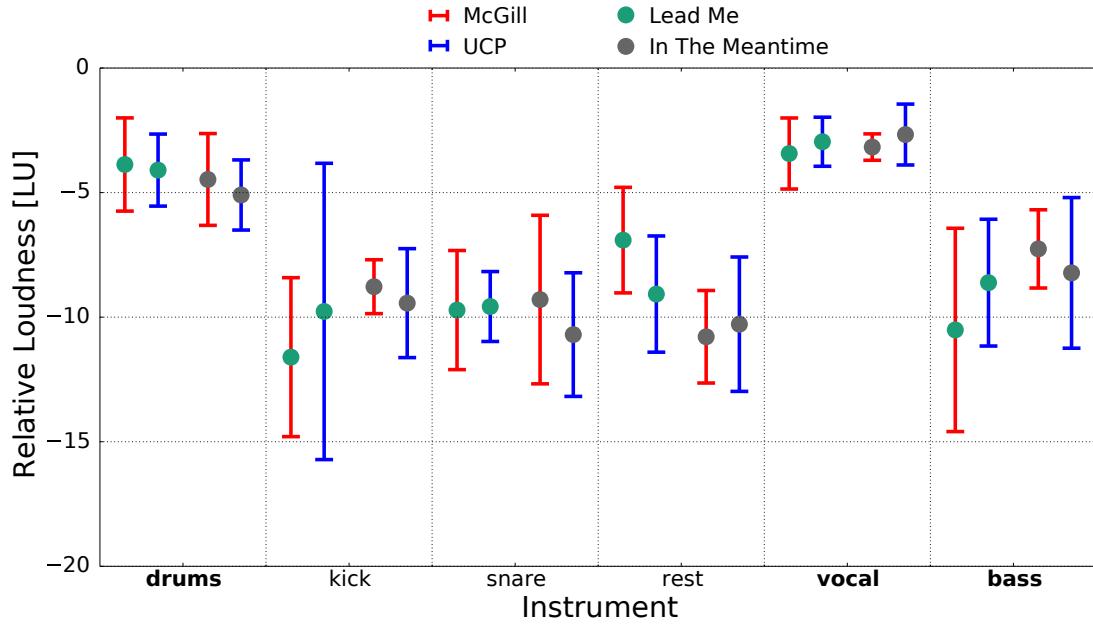


Figure 5.2: 95% confidence intervals of the mean relative loudness of different sources in Lead Me and In The Meantime, for McGill and UCP

Lead Me and In The Meantime were mixed by both McGill and UCP, for which the DAW session files are available. As such, the stems could be recreated and analysed separately, and the loudness analysis in Section 4.1.2 was repeated. The loudness of the drums, bass and lead vocal stems, shown in Figure 5.2, are not meaningfully different as the confidence intervals overlap ($p = .05$).

When considering all McGill songs and these two UCP songs, as in Figure 5.3, it is evident that the average drums loudness of these two common songs is higher than average. Therefore, the apparent differences between the two groups can be ascribed to the songs under investigation, and not to a definite tendency of the UCP group to mix drums louder. Agreement from engineers across different backgrounds on this relatively high drums loudness further supports the earlier hypothesis that the typical loudness of drums is song-dependent.

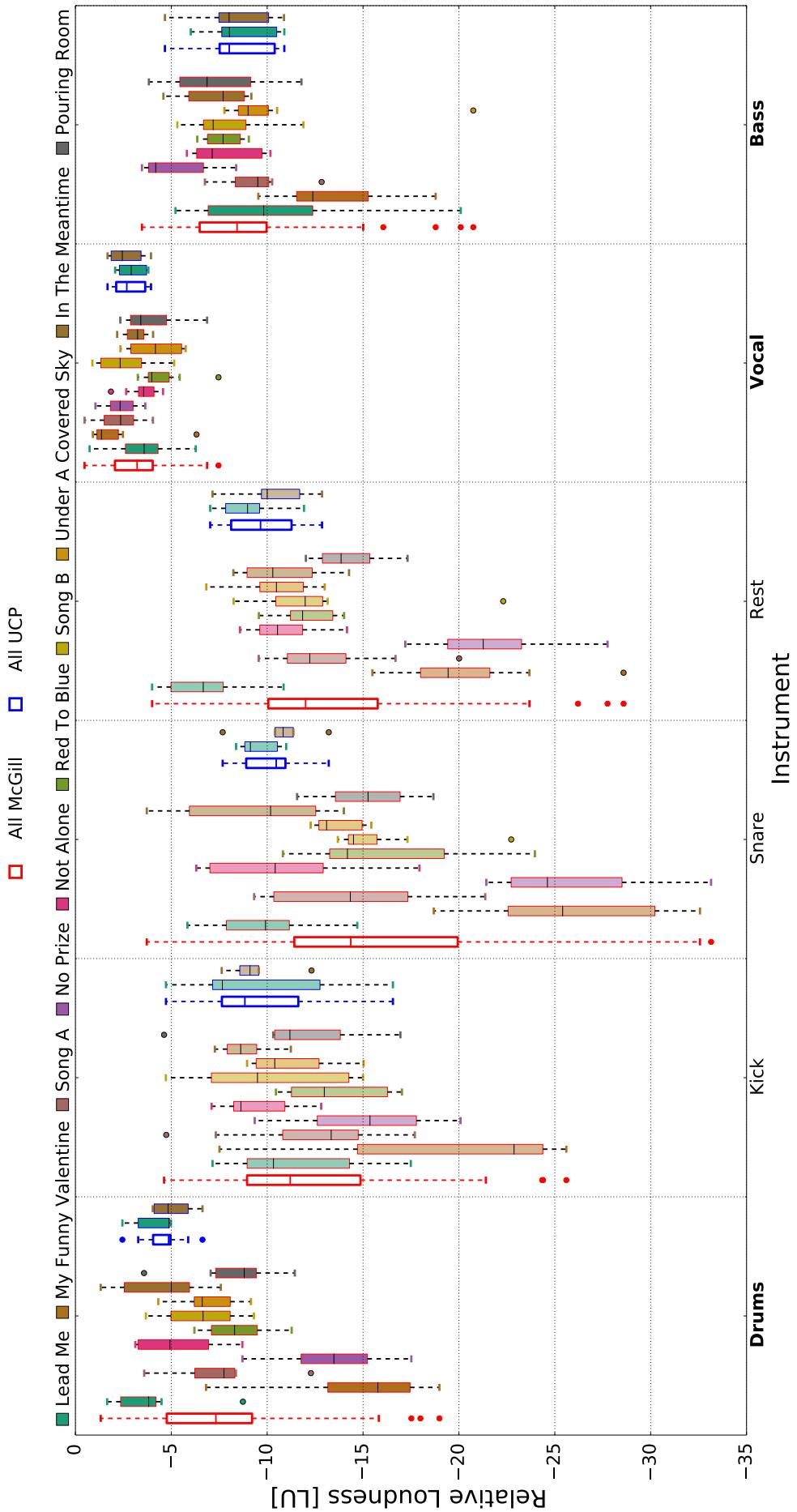


Figure 5.3: Box plot showing the relative loudness of different sources, per song, for McGill and UCP. The bottom and top of the ‘box’ represent the 25% and 75% percentile, the inner horizontal line indicates the median, and the dashed vertical lines extend from the minimum to the maximum, not including outliers (filled circles), which are higher than the 75% percentile or lower than the 25% percentile by at least 1.5 the interquartile range.

5.3 Subjective numerical ratings

5.3.1 Average rating

For different levels of expertise, the average rating from professionals (teaching and/or practising sound engineering professionally) is lower than from amateurs (no formal training in sound engineering) and students (currently training to be a sound engineer, and contributing mixes to the experiment), as expected [66, 140].

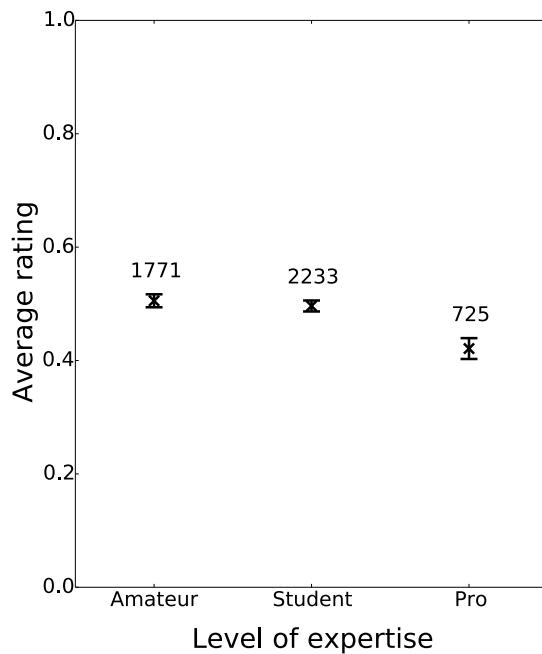


Figure 5.4: Average rating as a function of level of expertise (95% confidence intervals)

The effect of gender is non-existent when considering each level of expertise separately. While in this experiment, female subjects tend to award a slightly higher rating on average ($p = .05$), upon closer inspection this difference is more likely an effect of level of expertise. Note that 30% of the amateur listening test participants, only 14% of the students, and none of the professional engineers were women.

5.3.2 Self-assessment

In contrast to Section 4.2.1, where 15 out of 16 engineers rated their own mix higher than the median rating their mix received, this trend seems to be less universal when considering self-assessments from other institutions. In groups DU, PXL, and UCP,

ratings of one's own mix exceeded the median only slightly more than half the time. Curiously, In The Meantime does receive high self-assessment ratings for all institutions, with McGill (9/9 including professional), DU (6/7), PXL (6/7), and UCP (4/5) students awarding themselves an above-median score.

To determine whether the McGill group considered earlier is more likely to prefer their own mixes, a Pearson's chi-square test was conducted with categories 'self-rating below median' and 'self-rating above median'. From this, no significant difference was found between any two groups. Thus, this is a case where the larger study presented here contradicts the earlier conclusion that engineers would rate their own mixes higher than others. However, self-assessment may still lead to biased results.

5.4 Subjective free-form description

5.4.1 Praise and criticism

The proportion of negative statements among the comments is strongly influenced by the level of expertise of the subject. Dividing all subjects in groups of amateurs, students, and professionals, there is a significant tendency to criticise more, proportionally, with increasing experience. This is demonstrated in Table 5.3 using a Pearson's chi-squared test, for all levels of expertise, and for each pair of expertise levels individually.

Table 5.3: Chi-squared contingency table showing the observed total number of negative and positive statements for each level of expertise, the expected totals (between parentheses), the Pearson's chi-squared statistics for each cell (between square brackets), and the total chi-squared statistic and corresponding p -value for all groups, as well as each pair of groups

	Negative			Positive			Row totals
Amateur	2379	(2491.35)	[5.07]	1090	(977.65)	[12.91]	3469
Student	4889	(4898.67)	[0.02]	1932	(1922.33)	[0.05]	6821
Pro	1674	(1551.98)	[9.59]	487	(609.02)	[24.45]	2161
<i>Column Totals</i>	<i>8942</i>			<i>3509</i>			<i>12451 (Grand Total)</i>
							$\chi^2 = 52.09, p < 10^{-11}$
Amateur	2379	(2450.21)	[2.07]	1090	(1018.79)	[4.98]	3469
Student	4889	(4817.79)	[1.05]	1932	(2003.21)	[2.53]	6821
<i>Column Totals</i>	<i>7268</i>			<i>3022</i>			<i>10290 (Grand Total)</i>
							$\chi^2 = 10.63, p = .001$
Amateur	2379	(2497.31)	[5.60]	1090	(971.69)	[14.41]	3469
Pro	1674	(1555.69)	[9.00]	487	(605.31)	[23.12]	2161
<i>Column Totals</i>	<i>4053</i>			<i>1577</i>			<i>5630 (Grand Total)</i>
							$\chi^2 = 52.13, p < 10^{-12}$
Student	4889	(4983.99)	[1.81]	1932	(1837.01)	[4.91]	6821
Pro	1674	(1579.01)	[5.71]	487	(581.99)	[15.5]	2161
<i>Column Totals</i>	<i>6563</i>			<i>2419</i>			<i>8982 (Grand Total)</i>
							$\chi^2 = 27.94, p = 10^{-7}$

Independent of level of expertise, the proportion of negative statements is also significantly different per group ($\chi^2 = 127.49, p = 10^{-24}$), see Figure 5.5. Looking at the significant pairwise differences, McGill is more critical than all groups except UCP; UCP, in turn, more critical than all except MG; MG more critical than DU and SMC; and QMUL more critical than DU. All other pairwise differences are insignificant.

Gender, however, does not play a role in the ratio of positive versus negative statements ($\chi^2 = 1.70, p \approx .19$).

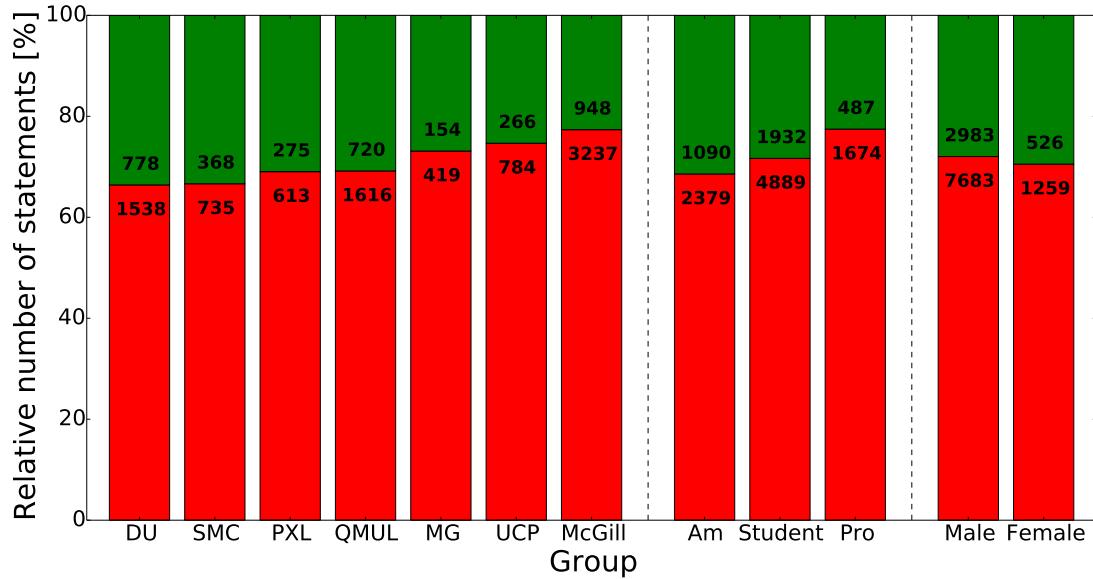


Figure 5.5: Relative number of negative (red) versus positive (green) statements, per group, level of expertise, and gender

Note that the simple statistical approach followed here and further on is flawed, as the statements are not truly independent, a requirement for such tests: many statements come from the same person, and many statements are describing the same phenomenon. A possible effect of violating the independence assumption is an increased likelihood of Type I errors, since the number of responses appears higher than it actually is. However, considering the very large number of responses and relatively low repetition, the effect should be minimal. In further work, it would be useful to conduct analysis with stronger independence by preprocessing the data, or to employ more advanced statistical methods where this independence is not required.

5.4.2 Comment focus

Likewise, it is clear that amateurs tend to give more ‘general’ comments, not pertaining to any particular instrument. This accounts for 55% of their statements. For students and professionals this proportion is 46% and 42%, respectively. The significance of these differences is demonstrated in Table 5.4 using a Pearson’s chi-squared test, for all levels of expertise, and for each pair of expertise levels individually.

The different groups also meaningfully differ with regard to the proportion of statements that discuss the mix as a whole, from 25% at UCP to 63% at DU, see Figure 5.6. As these two groups consisted of bachelor students only, the level of expertise is pre-

Table 5.4: Chi-squared contingency table showing the observed total number of instrument-specific and general statements for each level of expertise, the expected totals (between parentheses), the Pearson's chi-squared statistics for each cell (between square brackets), and the total chi-squared statistic and corresponding p -value for all groups, as well as each pair of groups

	Specific			General			Row totals
Amateur	1618	(1874.80)	[35.17]	1970	(1713.20)	[38.49]	3588
Student	3734	(3594.41)	[5.42]	3145	(3284.59)	[5.93]	6879
Pro	1261	(1143.79)	[12.01]	928	(1045.21)	[13.14]	2189
<i>Column Totals</i>	<i>6613</i>			<i>6043</i>			
							<i>12656 (Grand Total)</i>
							$\chi^2 = 110.17, p = 10^{-24}$
Amateur	1618	(1834.62)	[25.58]	1970	(1753.38)	[26.76]	3588
Student	3734	(3517.38)	[13.34]	3145	(3361.62)	[13.96]	6879
<i>Column Totals</i>	<i>5352</i>			<i>5115</i>			<i>10467 (Grand Total)</i>
							$\chi^2 = 79.64, p < 10^{-18}$
Amateur	1618	(1788.1)	[16.18]	1970	(1799.9)	[16.08]	3588
Pro	1261	(1090.9)	[26.52]	928	(1098.1)	[26.35]	2189
<i>Column Totals</i>	<i>2879</i>			<i>2898</i>			<i>5777 (Grand Total)</i>
							$\chi^2 = 85.13, p < 10^{-19}$
Student	3734	(3789.22)	[0.80]	3145	(3089.78)	[0.99]	6879
Pro	1261	(1205.78)	[2.53]	928	(983.22)	[3.10]	2189
<i>Column Totals</i>	<i>4995</i>			<i>4073</i>			<i>9068 (Grand Total)</i>
							$\chi^2 = 7.42, p = .006$

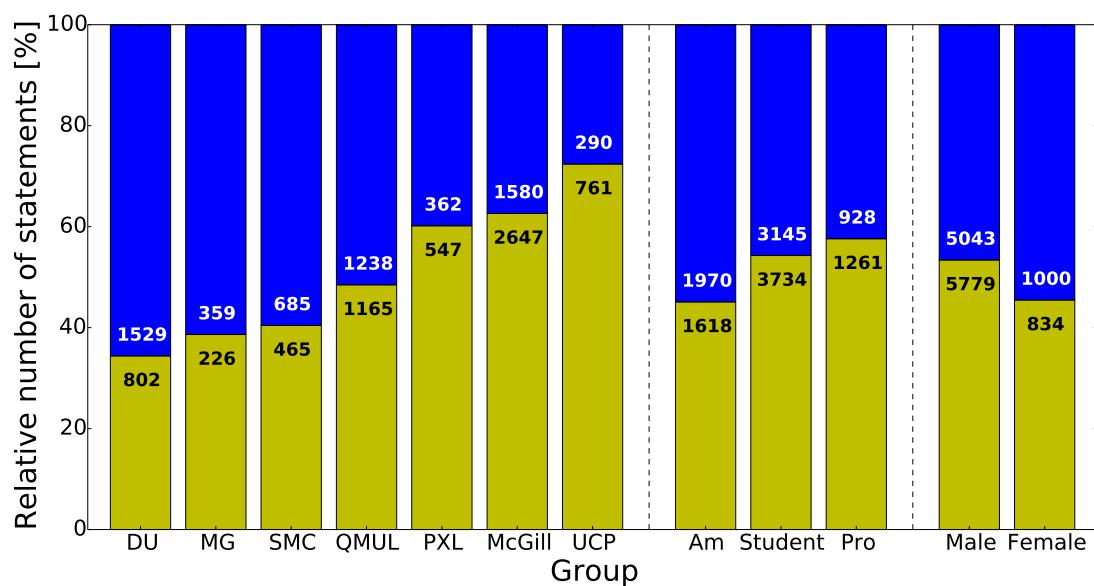


Figure 5.6: Relative number of instrument-specific (yellow) versus general (blue) statements, per group, level of expertise, and gender

sumably similar and other factors must be at play. The professional engineers from the PXL, QMUL, and MG groups contributed a comparable proportion of general statements, which was significantly higher than that from the professional engineers from McGill.

Among the amateurs only, male participants were more likely to discuss specific instruments than their female counterparts. As the division in levels of expertise was rough and based on self-reported years and type of experience, it is unclear whether this tendency is based on differences in training, intrinsically linked to gender, or caused by something else.

5.4.3 Agreement

Finally, the agreement within as well as between the groups is quantified, showing the relative number of statements which confirm each other.

In this context, a (dis)agreement is defined as a pair of statements related to the same instrument-processing pair and mix (e.g. each discussing ‘vocal, level’ for mix ‘McGill-A’ of the song ‘Lead Me’), with one statement confirming or opposing the other, respectively, with regard to either valence (‘negative’ versus ‘positive’) or value (‘low’ versus ‘high’). Only the processing categories ‘level’, ‘reverb’, ‘distance’, and ‘width’ have been assigned a value attribute. The ratio of agreements r_{AB} between two groups A and B is given by

$$r_{AB} = \frac{a_{AB}}{d_{AB} + a_{AB}} \quad (5.1)$$

where a_{AB} and d_{AB} are the total number of agreeing and disagreeing pairs of statements, respectively, where a pair of statements consists of a statement from group A and a statement from group B on the same topic.

Between and within the different levels of expertise, agreement increases consistently from amateurs over students to professionals, see Figure 5.7. In other words, skilled engineers are less likely to contradict each other when evaluating mixes (high within-group agreement). Conversely, amateur listeners tend to make more statements which are challenged by other amateurs, as well as more experienced subjects (low within-

group and between-group agreement). As the within-group agreement of amateurs is lower than any between-group agreement, this result does not indicate any consistent differences of opinion between two groups. For instance, there is no evidence that “amateurs want the vocal considerably louder than others”. Such distinctions may exist, but revealing them requires in-depth analysis of the individual statement categories.

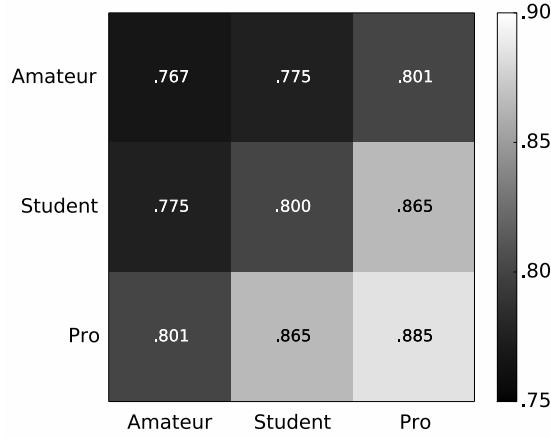


Figure 5.7: Relative agreement r_{AB} between subjects with different levels of expertise

Figure 5.8 shows the level of agreement between the various datasets, where the mixes evaluated by each overlap. Most of the differences can be explained by each group’s relative proportion of amateur, student, and pro subjects, though this does not account for all variance. As an example, responses from sound engineering students from DU are considerably less coherent ($r_{\{DU,DU\}} = .785$) than from UCP ($r_{\{UCP,UCP\}} = .843$). However, as different mixes were evaluated at these institutions, one cannot conclude that UCP has more reliable subjects. For instance, higher consistency between comments could also be caused by mixes with glaring issues: few subjects would then disagree on these aspects.

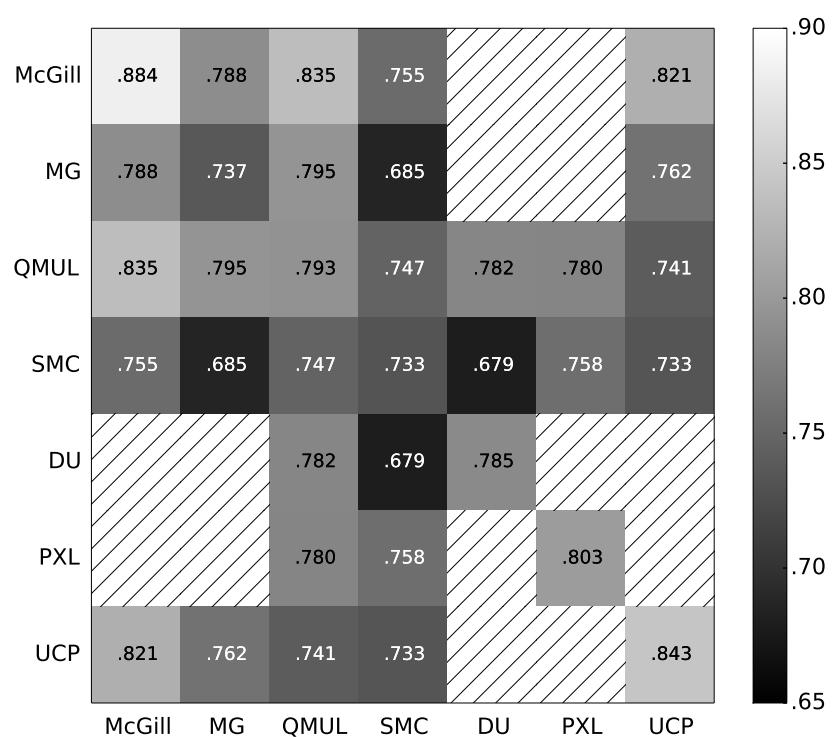


Figure 5.8: Relative agreement r_{AB} between subjects from different groups

5.5 Conclusion

The large set of mixes and mix evaluations presented here offers many opportunities for extension and reproduction of the previous work. The high degree of consistency between the considered populations suggests that there are universal best practices, and that these go beyond the straightforward technical aspects of eliminating recording artefacts. Repeating earlier experiments with this larger and more diverse agglomerated dataset generally confirmed the findings, strengthening original conclusions. Interestingly, though, one of the findings was refuted as it could not be reproduced using the new content. The results also provide a more nuanced view of earlier statistics, showing to what extent these values vary across different sections of the data.

In addition, this more diverse dataset has shown an increased proportion of negative, instrument-specific, and reverberation-related statements from mix engineers, when compared to inexperienced listeners. One could suppose a typical goal of the mixing process is to be imperceptible, so that the layman consumer does not think about it [53]. As such, they would lack the vocabulary or previous experience to formulate detailed comments about unfavourable aspects, instead highlighting features that tastefully grab attention and stand out in a positive sense. This would mean a successful mix should not just be technically sound, but also creatively outstanding. Experts, by contrast, are trained to spot and articulate problems with a mix, which would explain their tendency to make more negative and more specific statements.

Moreover, analysis of agreement has demonstrated a higher consistency in the mix evaluations of more experienced subjects, further supporting the notion that perceptual evaluation is more reliable and efficient when the participants are skilled [140, 171]. As experts converge on a common view, the hypothesis that universal best practices do exist is supported. The type of agreement analysis proposed here can be instrumental in comparing the quality of (groups of) subjects, on the condition that the evaluated stimuli are largely the same.

This comparison also showed that the initial dataset exhibited high agreement, as well as a relatively high number of negative and specific statements, indicative of a relevant and competent population.

Note that the subjects in this study all worked with audio, and most had experience with music education and performance. The classification in three levels of expertise was rudimentary and based on self-reported length and type of experience with music production. To reliably assess the effect of experience with music and sound, more work is needed with a larger and more diverse set of subjects. For instance, experience with playing a musical instrument could well have an influence on preferred loudness of particular instruments, as suggested by comments in Section 4.3.2.

Even with a study of this size, there is still a relatively small number of songs and genres, and only a handful of mixes of each song. This limits what can be said about the influence of song genre and engineer background.

Chapter 6

Conclusion

This thesis has presented a new approach to furthering knowledge of multitrack music mixing practices and perception. Revisiting the research questions that guided the work, this concluding chapter demonstrates how and to what extent each was addressed, and discusses what challenges remain.

The overarching research question of the work is how and to what extent the analysis of realistic mixes, produced by skilled engineers in a typical environment and using familiar tools, can be used to generate knowledge about mixing. This contrasts with other research in the field which is primarily based on controlled, lab-based experiments where a single component of the multi-faceted mix process is studied. An obvious advantage of minimally disrupting the natural workflow is the increased ecological validity of any findings the analysis may provide. The potential drawback, then, is that the relative lack of control could limit the number of significant results that can be produced. However, this was mitigated by ensuring a number of different mixes were produced from a specified set of songs, constraining the processors that could be used, and subsequently conducting rigorous subjective evaluation with expert listeners.

The various experiments presented here show that this true-to-life material can indeed meaningfully contribute to our understanding, as results with statistical significance are obtained in the context of each main category of mix processes (balance, panning, equalisation, dynamic range compression, and artificial reverberation). Moreover, in some instances the variance was lower than in prior studies on the same topic, despite the reduced level of control. This is likely due to the proficiency of both content creators

and listening test participants.

In addition, some questions can only be answered by studying realistic mixes. This includes any investigation into how mix engineers work, which by its very definition requires at least a close emulation of the mix process by actual mix engineers. An example of this was given in the form of a study of subgrouping practices (Sections 4.1.3 and 4.2.3). A quantitative analysis of subgrouping had not previously been published, likely because of the difficulty associated with collecting realistic sessions to examine.

An overview of the confirmed, contested, adjusted, and newly revealed mixing rules is given in Table 6.1, along with the number of the corresponding section in Chapter 4 (Sec.) and references to studies addressing the same topic (Ref.). Where values are mentioned, only overlapping confidence intervals are regarded as an agreement. Rules with an asterisk (*) were tested and confirmed based on a larger, more diverse dataset in Chapter 5.

The following strategies were used to obtain these rules:

4.1.2 Statistical analysis of audio features, extracted from processed instrument stems and the total mix of 64 mixes from 8 songs, as produced by 16 mix engineers (8 per song)

4.1.3 Statistical analysis of the number of tracks and subgroups counted in digital audio workstation session files of 64 mixes from 8 songs, by 16 mix engineers (8 per song)

4.2.2 Correlation between audio features extracted from 98 stereo mixes of 10 songs, by 26 mix engineers and one automatic system, and preference ratings by 34 skilled listeners

A.4 Analysis of preference ratings with associated comments, categorised as indicating excess or deficiency of reverberation, or neither, as assessed by 34 skilled listeners in response to 98 mixes of 10 songs, by 26 mix engineers and one automatic system

A.5 Measurement of total reverberation signal loudness, for mixes exhibiting a perceived excess and deficiency of reverberation, respectively, from 71 mixes of 10 songs, by 24 mix engineers (8 per song)

Table 6.1: Summary of confirmed, revised (crossed out), and new mixing rules

Sec.	Rule	Ref.
LEVEL	4.1.2* Vocal loudness should be around -3 LU relative to the total mix.	[51, 52], [24]
	4.1.2 Median total drums loudness is between -6.5 LU and -8.5 LU.	[24, 52]
	4.1.2 Total drums loudness is song- or genre-dependent.	
	4.1.2 Use overhead microphones as main signal and add emphasis with close microphones, or use close microphones as main signal and add ‘air’ or ‘ambience’ with overhead microphones.	[65]
PANNING	4.1.2* Median bass loudness is between -8 LU and -9 LU.	[24, 52]
	4.1.2 Lead vocals, snare drums, and low-frequency sources should be panned central.	[25, 26, 28, 65, 208]
	4.1.2 Higher frequencies are panned increasingly further <i>between 50 Hz and 400 Hz</i> .	[26, 208]
	4.2.2 The centre of the stereo image should be sufficiently strong.	
EQ	4.2.2 A mix should not be overly monaural.	
	4.1.2 The total mix tends to a particular target spectrum.	[67]
DRC	4.2.2 A mix should have sufficient dynamic range: it is better to err on the side of applying (too) little compression.	[37, 66]
REVERB	A.4* It is better to err on the side of too little reverb, rather than too much.	[51, 144]
	A.5* The total reverb loudness should be about -14 LU.	[51]
	A.6* For a higher perceived amount of reverberation, increase the reverb loudness and/or reverb time.	[233–235]
SG	4.1.3 The number of subgroups created strongly depends on the number of tracks.	
	4.1.3 Tracks are typically grouped per instrument type.	

A.6 As above, including measurement of Equivalent Impulse Response

The following discusses findings pertaining to each additional subquestion.

How can we address the challenges research on mixing is facing?

Analysis and evaluation of mix engineering is a specialised field requiring specialised tools and datasets. Despite this, the topic is characterised by a lack of suitable, shareable data, and borrows practices and software from neighbouring disciplines in the absence of established methods and purpose-built applications. An attempt at providing these was made in Chapter 3, in the form of a growing repository of multitrack audio

and mixes, offering centralised access to existing and new resources, and an evaluation tool based on a meticulously developed listening test methodology. This enabled the study of various aspects of the mixing process, as detailed below.

Previous mixing systems have largely ignored high-level information such as instrument labels. Because of their disregard for the semantic structure of music, their results can be unsatisfactory compared to mixes by human engineers or even an elementary rule-based system (Chapter 2). Without the incorporation of this type of metadata, traditional expectations such as centring lead vocals and snare drums cannot be met. However, low-level information is essential too, to account for deviations in spectral and dynamic characteristics of the source. This has prompted analysis of objective audio features, extracted from the audio of several instruments in different songs, to establish common instrument-specific processing practices.

Furthermore, an absence of established mixing guidelines was noted (Chapter 2), particularly in relation to balance and reverberation. Therefore, particular attention was dedicated to these processors in the ensuing analysis (Chapter 4 and 5, and Appendix), demonstrating the viability of the proposed approach with regard to unearthing new rules in relatively unexplored areas.

How can knowledge about mixes be obtained from poor examples?

Even with high quality source material, procuring such ‘realistic’ mixes of professional quality is a resource-intensive endeavour, not least because professional mix engineers are less likely to donate their time and skills than the average listening test participant. Sound engineering students and educators, on the other hand, evidently see didactic value in the mix exercise which can be implemented as part of relevant courses. This yields content of considerable quality from motivated engineers. Invariably, though, some of the contributions will be less than stellar, and not representative of typical or commercial-grade mix practices. Moreover, even mixes from acclaimed engineers are not always universally appreciated, and can be surpassed by student mixes (Section 4.2.1). Analysis of audio features and workflow statistics in Section 4.1, while insightful, is therefore inevitably based on at least a few ‘poor’ examples, affecting the results by adding noise at best or bias at worst.

To account for this, perceptual evaluation must be used to learn the preference for the respective mixes from a group of skilled subjects. Analysis of these ratings in combination with the aforementioned audio features can then determine the impact of such objective measures on preference (Section 4.2). Note that in this case, a wide range of sonic characteristics as well as preference ratings is actually beneficial, as examining these relationships is only possible if there are instances of both favourable and unfavourable feature values. In other words, ‘poor’ examples are not only unavoidable, but also useful.

Zooming in on the perception of particular attributes of the mixes, analysis of subjective comments allows one to reveal specifically which treatments of which instruments are liked or disliked. Here again, high variation in processing and ratings is advantageous as this permits learning how mix characteristics affect preference, which is impossible when only similar, equally preferred stimuli are available. This detailed subjective feedback of mixes is largely ‘negative’, pointing out aspects of the mix which are considered flaws (Section 4.3). As such, audio feature values from particular tracks can be categorised according to whether or not the processing is deemed appropriate, as discussed in the next section.

These supposedly poor examples are crucial for the exploration of a parameter space or feature space, if boundaries are to be found beyond which values are widely considered unacceptable. After all, establishing the limits of such a surface requires examples of mixes on both sides of it, with associated perceptual evaluation. In the event that subjective feedback consistently praises an attribute’s value when it is within a certain range, and criticises mixes in which this is not the case, a mixing space can indeed be defined for the correlated dimensions. Appendix provides evidence for the existence of such a space, in the relatively uncharted area of reverberation. As an example, the perceptual construct ‘amount of reverberation’ is mapped as a function of two objective features, from mix evaluations where a perceived relative lack or abundance of reverberation was reported. In-depth analysis, possibly from the same dataset, is required to establish which other parameters and features exhibit a range of widely accepted values. From this information, new tools can be devised which display alerts to the user, scale the control bounds, or automatically manipulate the signals when the limits are exceeded.

How can it be established how words used to describe sounds or mix processes correspond with objective features or process parameters?

Chapter 2 presents a first attempt at understanding these descriptive terms, by simply compiling a list of words used to denote an absence, presence, or excess of energy in particular frequency bands, as defined in practical sound engineering literature. While this provided some guidance when interpreting best practices from audio textbooks, the definitions are based on the opinion and experience of the respective authors rather than rigorous perceptual evaluation, vary between and within the different sources, and consist of a loosely defined frequency range only.

Determining the relation between sonic descriptors and actionable processing parameters or measurable audio features requires subjective evaluation of several stimuli which exhibit differences along the perceptual dimension in question. The produced dataset contains terms describing various types of processing of various instruments, the most common of which are listed in Section 4.3.

As collection and manual annotation of mix reviews proved to be tedious, a framework for collecting such descriptive terms is proposed in Section 4.4. Obviating arduous perceptual evaluation experiments, this approach focuses on the elicitation of descriptors pertaining to parameter settings and audio features of single sources, in a multitrack mixing environment. The large dataset accumulated thus far proved the viability of this system as a data collection method, and initial analysis demonstrated its suitability for investigating the definition and similarity of collected terms.

To what extent do differences between sound engineers or listeners limit the generality of findings in music production?

Through the use of the Web Audio Evaluation Tool and the Open Multitrack Testbed introduced in Chapter 3, it was possible to repeat the analysis with a vastly expanded dataset in Chapter 5. These tools proved to be essential to conduct mix creation and evaluation experiments on an unprecedented scale, enabling thorough reproduction and extension of the initial findings.

Differences between listening test subjects revealed the impact of differences in sound

engineering experience on the numerical rating, subjective comments, and degree of discrepancy in evaluations. Analysis showed more experienced subjects produce lower ratings, a higher proportion of negative and specific statements, more comments related to reverb, and fewer disputed judgements.

Including a wide range of data, content creators, and listening test participants has proven useful to demonstrate the relative variation of certain measures across different populations. One finding was revised, as it was no longer found to be significant when considering the larger dataset (Section 5.3.2).

Despite the differences in reliability, consistency, and style of responses, no contrasting tendencies were found between participants with different levels of expertise, locations, or genders. Consequently, the results corroborated earlier findings with regard to both audio features and perception, on topics as basic as balance and as unexplored as reverb. This supported the conclusions from the preceding chapters, suggesting they may be widely applicable regardless of the background of the sound engineer or listener. It also indicates that results of related studies are relevant beyond the scope of the considered subjects or data.

Future work

This work merely scratches the surface of what can be investigated with the combination of the source audio, parameter settings, preference ratings, and subjective descriptions of these mixes. For instance, only a fraction of the rules found in practical sound engineering literature were tested. It should also be noted that the findings herein do not necessarily apply to music from genres not considered here, live mixing, or non-musical audio, although most of the methods can be employed for such content. Instead, the current study sought to prove the concept of real-world data capture as a means to collect relevant knowledge about music production. Through publishing the data, sharing the tools, and documenting the methodology, it allows other researchers to extend the work.

It is possible that some of the premises on which the analysis relies are false, invalidating certain findings. For instance, despite the differences between the considered

datasets, the multi-group analysis presented in Chapter 5 is based mostly on the assessments of audio experts, using mostly the same source material. Furthermore, it was acknowledged that parts of the presented statistical methods were overly simplified, incorrectly assuming independence of evaluations from the same person or about the same phenomena. More advanced methods such as more formal qualitative analysis of mix reviews, borrowing from fields like grounded theory, would further improve the rigour of the work.

The current work, and almost all related work thus far, only considers mixes with at most two channels. Expanding to surround sound, object-based audio, and similar formats, would generate knowledge that is more relevant to the increasingly important domain of VR systems, as well as game and film audio.

Adding to the effort and time required for mix analysis, current commercial music production software severely obstructs automated extraction of features and parameter settings. As a result, even when content producers are willing to contribute data, the cost of post-processing can be prohibitive. Fortunately, the success and extension of the real-time attribute elicitation architecture discussed in Section 4.4 [236–243] and recent efforts towards a web-based digital audio workstation [244] indicate integrated data collection from a multitrack audio processing environment will soon be a reality. These developments have the potential to considerably increase the efficiency of data collection and expand the scope of possible analysis. As an example, recording all mix actions and corresponding timestamps, similar to the event logging functionality of the Web Audio Evaluation Tool (Figure 3.7), will allow studying the entire mixing process over time, instead of just the end result. Additionally, if online tests prove to be a viable method for subjective evaluation of musical signal processing, services like Amazon Mechanical Turk may be utilised to significantly increase the scale of the dataset. A larger, more diverse corpus of mixes could reveal differences in practices and perception between song genres and listener or engineer backgrounds, but also span a wider range of feature values and processing parameters, and hence be more successful at exposing correlations with preference. This may also open the door to new research directions, thus far impossible due to high data requirements, such as advanced automatic mixing powered by machine learning techniques.

As this work has shown real-life mix analysis to be a valid method to expand music production knowledge, manufacturers of mixing tools may choose to mine settings, features, and user data to create systems that learn from the actions of any individual user, or from all users combined.

Finally, with analysis of high volumes of data it may become possible to uncover the rules that govern not just mix engineering in general, but particular mixing styles. From an application point of view, a target profile can thus be applied to source content to mimick the approach of a certain engineer, to fit a specific musical genre, or to achieve the most suitable properties for a given medium.

Concluding remarks

Furthering the understanding of the complex mix process, even with its creative and esoteric nature, is a meaningful and attainable goal. Its pursuit, however, can easily continue for many more studies. This thesis has taken essential steps in this direction, by introducing the aforementioned dataset, evaluation methodology, analysis methods, and the conclusions that have already been drawn from these. The presented data and approaches have thus far enabled statistical analysis of audio features, investigation of best practices and workflow, measurement of the correlation between preference and features, quantification of the perceptual importance of various mix aspects, definition of descriptive terms, delineation of preferred parameter regions, and assessment of the variation across different groups of subjects and mixes.

The creation and assessment of realistic mixes is a laborious process, where a certain level of control is sacrificed for high ecological validity and preservation of the possible interplay of mix processes. This led to a particularly wide scope, as all common mix processes are studied in this thesis, rather than focusing on a specific aspect. Manual comment annotation of free-form text responses was also deemed necessary, as the vocabulary and salient perceptual dimensions of the mixing process are largely unknown. As a consequence, the approach is only moderately scalable, even when the data, tools, and methods presented here can be repurposed. However, exploratory studies such as this one can ultimately inspire more focused and controlled experiments, zooming in on a single aspect with proven relevance and known perceptual attributes.

Over the course of this work (2012–2016), the field of intelligent music production has witnessed massive growth in breadth, depth, and popularity. Neither dynamic range compression [33–38], reverberation [40, 41], or harmonic distortion [42] had any automated implementations before this time. Machine learning approaches to automatic mixing have received particular attention [20, 33, 40, 118], and the considered genres have been expanded, with some systems even specifically catering to jazz [118, 139]. In that same time, an increasing number of companies — from startups to major players — have released new products featuring high-level control of audio features, automatic parameter setting, and full ‘black-box’ music production services. Many of the recent conferences, conventions, and workshops by the Audio Engineering Society, some of which the author chaired or participated in, have featured dedicated sessions on topics like semantic music production or intelligent sound engineering. All of this leaves little doubt about the importance, in both academia and industry, of the wider field of analysis and automation of music production processes, and the relevance of understanding multitrack music mixing in particular.

Appendix

Case study: Use and perception of reverb

Bringing together the elements of the proposed methodology, objective features extracted from the multitrack mixes are combined with the annotated subjective comments to generate new knowledge about mixing practices and their impact on perception. The case of the reverb effect is considered as a proof of concept, to determine whether the collected data can be used to establish feature space boundaries corresponding to a perceived excess and deficiency of reverberation amount.

As previously demonstrated, few best practices are known with regard to reverberation in music production (Chapter 2), despite having an important impact on the perception of a mix (Section 4.3). With the exception of the recent work in [40], there have been no attempts at automatic reverb effects. The ability to predict the desired amount of reverberation with a reasonable degree of accuracy would also have applications in novel music production interfaces [57], compensation of listening conditions [168], and intelligent metering.

A.1 On reverb

Reverberation is one of the most important tools at the disposal of the audio engineer. Essential in any recording studio or live sound system [246], the use of artificial reverb (simply referred to as ‘reverb’ in this section) is widespread in most musical genres and it

is among the most universal types of audio processing in music production [247].

Beyond simulating an acoustic space, reverb can be used to bind sources together by seemingly placing separately recorded sources in the same environment, or to create distance and contrast. It can enhance audibility by sustaining certain elements such that they may become more noticeable, or blur and mask between sources [85] and fill gaps between sound events [246]. The addition of stereo reverb to a single source can add width [86]. Reverb further affects timbre [246], loudness [57], sound quality [248], and depth perception [1].

Reverberation is often generated artificially to avoid the practical and financial cost of recording on location, to retain control over the reverberation type and amount in post-production, to compensate for close microphone positions — maximising isolation from other sources or capturing a specific part of an instrument — or to create unnatural effects that are not constrained by the laws of physics [85, 249]. While early examples of artificial reverberation were generated by re-recording the signal through a speaker in an echo chamber, or emulated using electromechanical devices such as a spring or plate reverb, contemporary reverberation effects are mostly implemented digitally [1, 85, 250].

Despite its prominence in music production, there are few studies on the usage and perception of artificial reverberation relevant to this context. This scarcity may relate to a lack of universal parameters and interfaces, whilst algorithms across the available reverb units vary wildly. In comparison, typical EQ parameters are standardised and readily translate to other implementations.

A.2 Background

In contrast to other important mix engineering tools, such as level [11, 16], panning [25], EQ [29] and dynamic range compression [34, 35], to date only one attempt at automatic control of reverberators has been made [40, 41]. Very little work is available on novel, more intuitive interfaces for reverb [60, 251] and mapping terms to its parameters [57]. A number of studies have looked at perception of reverberation in musical contexts [51, 167, 168, 170, 233–235, 248, 252–263], see Table A.1.

Table A.1: Overview of studies concerning perception of reverberation of musical signals. Test method: PE or DA (Perceptual Evaluation or Direct Adjustment of reverb settings); participants Skilled or Unskilled in audio engineering. Reverberator properties: Stereo or Mono; Early Reflections or No Early Reflections.

		Stereo		Mono	
		ER	No ER	ER	No ER
PE	Skilled	[51, 233, 258, 259]	■■■■■	[253, 263]	[254]
	Unskilled	[234, 235, 255–257]	[248]	[168]	[167, 252]
DA	Skilled	[51]	[170]	[260]	[261]
	Unskilled	[257, 262]	■■■■■	[167]	

The focus of this proof-of-concept study is the perception of artificial reverberation of multi-source materials taken from examples of fully-realised, professional music productions. The present case thus stands apart from the work cited above, where the effect of reverb parameters on the subject's preference or perception is under investigation, as applied to a single source, and typically isolated from any musical, visual or sonic context. As reverberation is a complex and multifaceted matter, controlled experiments are often required. Several of these studies involved only a single, simple and potentially unpleasant and unfamiliar reverberator [256, 259], sometimes without the use of early reflections [248, 252] or stereo capabilities [168, 253]. In some cases, the number of reverberator parameters were limited, often taking a restricted range or set of values [234, 254, 255], and applied to a single (type of) source sample [235, 257, 258]. In [41, 167] the parameter values were set by unskilled participants using unfamiliar tools in inferior listening environments. Finally, the results of several parameter adjustment tests were not validated through perceptual evaluation [170, 260, 261].

It has not yet been investigated whether the perception of reverberation amount and time of a single source in isolation has any relevance within the context of multitrack music production, inherently a multidimensional problem, where different amounts and types of reverb are usually applied to different sources, which are then combined to form a coherent mixture. Thus, while relevant for the respective studies, these works may not offer insight into how an audio professional might apply reverb in a commercial music production environment. In order to better understand the use, perception, and preference with regard to reverberation in music, it is therefore deemed necessary to study its application by skilled engineers using familiar, professional grade tools in the

context of a complete, representative mix.

A.3 Problem formulation

In what follows, the perceived amount of reverberation is predicted based on objective features extracted from both the combined reverb signal and the remainder of the mix. These signals will be referred to as wet (\mathbf{s}_{wet}) and dry (\mathbf{s}_{dry}), respectively. They are complex and laborious to extract in practice, even when all source audio and DAW session files, including all parameter settings, are available. This is due to the following conditions:

1. different amounts and types of reverb are applied to the different sources in the mixture; and
2. post-reverb, nonlinear processing (dynamic range compression, fader riding, automation of parameters) as well as linear processing (balance, EQ) are applied to the individual sources as well as the complete mix or subgroups thereof.

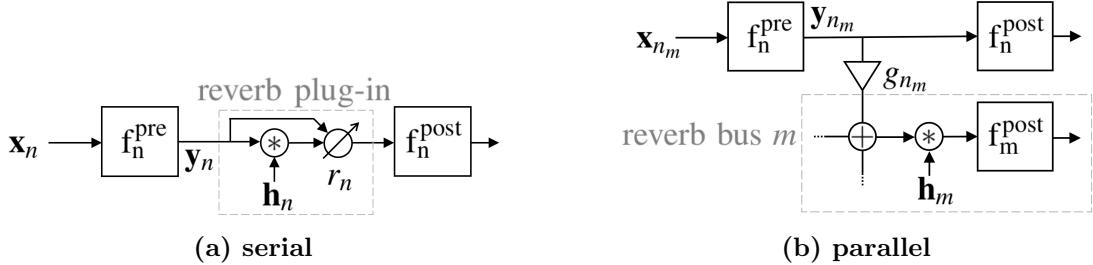


Figure A.1: Reverb signal chains

Omitting time arguments for readability, tracks $n = 1, \dots, N$ carry the source signals \mathbf{x}_n which are often already processed before any reverb is applied, giving $\mathbf{y}_n = f_n^{\text{pre}}(\mathbf{x}_n)$. Reverb (with impulse response \mathbf{h}_n) can be added to the processed tracks \mathbf{y}_n using serial processing, with the reverb plugin inserted ‘in-line’, where the gain ratio $r_n \in [0, 1]$ between the wet and dry signal is set within the plugin (Figure A.1a). Alternatively, reverb is added through parallel processing, with tracks scaled by a gain factor g and sent to a reverb plugin on a separate bus. Typically, several tracks $n_m = 1, \dots, N_m$ are sent to the same reverb bus m (Figure A.1b). In both cases, further processing $f_n^{\text{post}}(\cdot)$ is then applied to the respective tracks and buses, i.e. post-reverb.

The wet and dry part of the mix can therefore be expressed as:

$$\mathbf{s}_{wet} = \sum_n f_n^{\text{post}}(r_n \mathbf{h}_n * \mathbf{y}_n) + \sum_m f_m^{\text{post}}(\mathbf{h}_m * \sum_{n_m} g_{n_m} \mathbf{y}_{n_m}) \quad (\text{A.1})$$

$$\mathbf{s}_{dry} = \sum_n f_n^{\text{post}}((1 - r_n) \mathbf{y}_n) \quad (\text{A.2})$$

With $\mathbf{h}'_n = (r_n \mathbf{h}_n + (1 - r_n) \delta)$ as the total impulse response of the in-line reverb, reverberant ratio r_n included, where δ is the unit impulse, the total mix \mathbf{s}_{tot} then becomes

$$\mathbf{s}_{tot} = \sum_n f_n^{\text{post}}(\mathbf{h}'_n * \mathbf{y}_n) + \sum_m f_m^{\text{post}}(\mathbf{h}_m * \sum_{n_m} g_{n_m} \mathbf{y}_{n_m}) \quad (\text{A.3})$$

which is equal to $\mathbf{s}_{dry} + \mathbf{s}_{wet}$ as long as the condition $f_n^{\text{post}}(a + b) = f_n^{\text{post}}(a) + f_n^{\text{post}}(b)$ is satisfied. For this to be true, post-reverb nonlinear processing $f_n^{\text{post}}(\cdot)$ was applied to both the wet and dry signal in such a way that their sum still equals the original mix. Any gain changes applied by a dynamic range compressor are dependent on its side-chain signal (equal to the input signal by default). As such, the original mixed signal is used for this side-chain signal when processing the dry or wet signal. In other words, in Equations (A.1) and (A.2), $f_n^{\text{post}}(\cdot) = f_n^{\text{post}}(\cdot, \mathbf{h}'_n * \mathbf{y}_n)$, with the extra argument representing the side-chain signal, so that $\mathbf{s}_{tot} \equiv \mathbf{s}_{dry} + \mathbf{s}_{wet}$. For simplicity of the expressions, it is assumed that this post-processing is applied per track, though in practice it can be applied to groups of sources simultaneously.

The interest herein is how the perceived excess or lack of reverberation amount is influenced by the difference between the loudness of the reverb signal and the dry signal (see [51, 168, 248]), as well as the overall reverberation time (see [248, 258, 259]).

The first considered feature, *relative reverb loudness* (RRL), is defined as

$$\text{RRL} = \overline{\text{ML}(\mathbf{s}_{wet}) - \text{ML}(\mathbf{s}_{dry})} \quad (\text{A.4})$$

where ML is the *Momentary Loudness* in loudness units (LU) as specified in [264]. The difference of the momentary loudness of the wet and dry signal is calculated for each

measurement window, and the average (\bar{x}) is taken over each window. It should be noted that (forward) masking and binaural dereverberation are not taken into account with this measure. More advanced partial loudness features were used in [248] to predict the perceived amount of reverb. However, such features were not used in this work because they did not perform well on the considered content, showing weak correlation with perception, and more work is needed to establish the applicability of multi-band loudness models [205], specifically to multi-source music [265]. Furthermore, the simple filtered RMS measure used here is far less computationally expensive, and suitable for real-time applications.

The second feature, reverberation time, is usually derived from the reverberation impulse response (RIR). In the context of this study, however, the RIR is not readily defined, due to conditions 1) and 2) above. As such, the transformation between the mix *without* reverb and the mix *with* reverb is not a linear one, and it cannot be defined by an impulse response, even if the reverberator used is applying a linear transformation (which is also not always the case [250]). However, an *Equivalent Impulse Response* (EIR) \mathbf{h}_{eq} can be estimated in which temporal and spectral aspects of the total reverb are embedded:

$$\mathbf{s}_{wet} \approx \mathbf{h}_{eq} * \mathbf{s}_{dry} \quad (\text{A.5})$$

From such an impulse response, traditional acoustic reverberation parameters can be extracted, which describe the overall reverberation in universally defined terms such as reverberation time, along with clarity, IR spectral centroid and central time, which can then be translated to other reverberators [57].

In this section, student mixes from songs 1–10 are considered (Table 3.2). For each individual song, between 12 and 16 subjects assessed the different mixes, as evaluations of one's own mix were excluded (see Section 4.2.1). Only 71 of the 80 mixes are considered, where all parameters were accessible and the mix could be perfectly recreated.

In the other cases, participants used more than the permitted set of tools.

A.4 Comment analysis

Of the 1326 collected comments, 35.44% mention reverberation, and reverberation is not commented on by anyone for only 4 of the 98 mixes. Furthermore, every subject commented on reverberation for at least 11 percent of the mixes they assessed. The comments were classified into three classes: “*Too much reverb*”, “*Not enough reverb*”, and — when unrelated to the perceived amount of reverberation — *Neither*.

Participants disagreed on whether there was too much or too little reverberation in only 4 of the 525 comments which mention reverberation. This supports the idea that mix engineers have a consistent judgement on the ‘correct’ reverberation amount for a given mix. The low discrepancy may be explained by the fact that test participants are skilled listeners [167]. Going forward, only comments regarding the subjective excess or shortage of reverberation of the whole mix are considered, i.e. not specific to any particular instrument.

Figure A.2 shows the mean preference ratings associated with comments from the different classes. As previously observed in [51, 144], the preference rating for a mix the subject found too reverberant is significantly lower than if it was considered too dry.

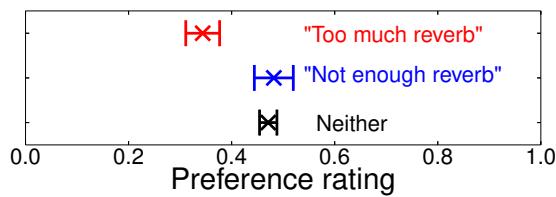


Figure A.2: Preference (0.0–1.0) per class: 95% confidence intervals

Note that the subjects were not instructed to focus on any particular aspect of the mix, including reverberation, so that the collected preference ratings and comments relate to any characteristic the subject deemed worthy to report. In other words, the experiment is not ‘forced choice’ with regard to reverberation, but each comment related to it is the result of a spontaneous reaction to the mix, on its own or in relation to other

mixes.

A.5 Relative Reverb Loudness

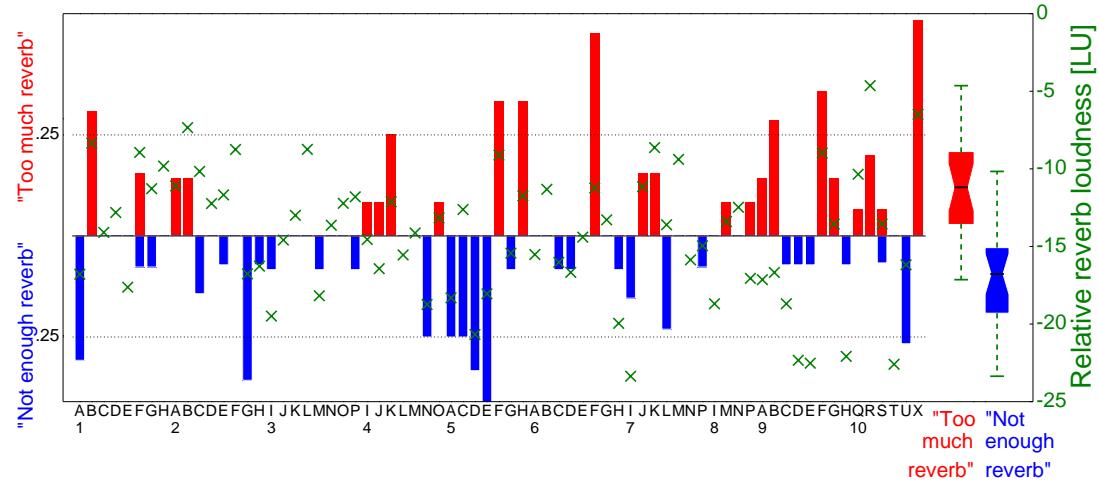


Figure A.3: Proportion of subjects who noted an excess (red) or deficit (blue) of reverberation, versus the relative loudness of the reverb signal (green Xes). Letters denote different mix engineers, numbers denote different songs (see Table 3.2). The box plots show the relative loudness values for mixes collectively found to be too ‘wet’ and ‘dry’, respectively.

The relative reverb loudness is shown for each mix in Figure A.3, along with the number of subjects who indicated the mix was perceived as too reverberant or not reverberant enough, divided by the total number of subjects for that song. As expected, the majority of the mixes labelled ‘too reverberant’ have a significantly higher relative reverb loudness than those labelled ‘not reverberant enough’.

Overall, the preferred reverb loudness differs from [51], where the optimal reverb return loudness is estimated to be at -9 LU relative to the total mix loudness. In the current experiment, every mix with a relative reverb loudness of -9 LU or higher was judged to be too reverberant, and -14 LU appears to be a more desirable loudness as it is in between 95% confidence intervals of the medians of either labelled group.

The differences in reverb loudness are mostly subtle, with the just-noticeable difference (JND) of direct-to-reverberant ratio estimated at 5–6 dB [266], proof of the critical nature of the engineer’s task [1]. Despite this, there is a large level of agreement with regard to what mixes have a reverb surplus or deficit. The variance of preferred reverb level is considerably larger in [167], possibly due to the unskilled listeners.

There are some cases where despite a relatively high reverb loudness, subjects agreed that there was not enough reverberation (e.g. mix 2C or 5C in Figure A.3), or where mixes with a perceived excess of reverb did not exhibit a significantly higher-than-average measured loudness (e.g. 8P, 9B). Closer study of these outliers, through informal listening and analysis of parameter settings, revealed that mixes with a high perceived amount of reverberation but low measured reverb loudness typically have a long reverberation tail. Those marked as too dry have a strong, yet short and clear reverb signal, to the point of sounding similar to the dry input. As in [248], it would seem relative loudness of the reverb signal alone is generally insufficient to predict the perceived or preferred amount of reverberation. It is therefore believed that measuring the reverberation time will help explain the perceived amount of reverberation [233–235].

A.6 Equivalent Impulse Response

A.6.1 Process

For the practical measurement of the EIR \mathbf{h}_{eq} (see Equation (A.5)) it is not possible to use sine sweep or maximum length sequence (MLS) methods due to condition 1) from Section A.3 above. In the frequency domain, if $f_n^{post}(\cdot)$ is a linear filter with frequency response $\mathbf{F}_n^{(post)}$, spectral division of the Fourier transforms of Equations (A.1) and (A.2) yields an equivalent frequency response

$$\begin{aligned}\mathbf{H}_{eq} &= \frac{\mathbf{S}_{wet}}{\mathbf{S}_{dry}} \\ &= \frac{\sum_n \mathbf{F}_n^{(post)} r_n \mathbf{H}_n \mathbf{Y}_n + \sum_m \mathbf{F}_m^{(post)} \mathbf{H}_m (\sum_{n_m} g_{n_m} \mathbf{Y}_{n_m})}{\sum_n \mathbf{F}_n^{(post)} (1 - r_n) \mathbf{Y}_n}\end{aligned}\tag{A.6}$$

In this case, the equivalent frequency response \mathbf{H}_{eq} is a frequency- and gain-weighted version of the various reverb frequency responses \mathbf{H}_n and \mathbf{H}_m , dependent on the pre-processed input signals, the post-processing, and the wet to dry ratios. This interpretation is violated to the extent that $f_n^{post}(\cdot)$ is not a linear function, see condition 2) from Section A.3. In the case it is approximately linear but not stationary, the equivalent frequency response can describe the total reverb with reasonable accuracy as a function of time.

Neglecting any nonlinearities, the EIR is obtained by division of the signals (\mathbf{s}_{wet} and \mathbf{s}_{dry}) in the spectral domain, an approach known as ‘dual channel FFT analysis’ [267]. Following Welch’s method, complex averaging is performed on both the dry signal’s power spectrum or auto spectrum ($\mathbf{G}_{dry,dry}^{(i)}$) and the cross spectrum ($\mathbf{G}_{dry,wet}^{(i)}$), taken from signal segments $i = 1, \dots, I$, with 50% overlap and a Hann window:

$$\begin{aligned}\mathbf{G}_{dry,dry}^{(i)} &= \mathbf{S}_{dry}^{*(i)} \mathbf{S}_{dry}^{(i)} \\ \mathbf{G}_{dry,wet}^{(i)} &= \mathbf{S}_{dry}^{*(i)} \mathbf{S}_{wet}^{(i)} \\ \mathbf{H}_{eq} &= \frac{\frac{1}{I} \sum_{i=1}^I \mathbf{G}_{dry,wet}^{(i)}}{\frac{1}{I} \sum_{i=1}^I \mathbf{G}_{dry,dry}^{(i)}} \equiv \frac{\mathbf{G}_{dry,wet}}{\mathbf{G}_{dry,dry}} \\ \mathbf{h}_{eq} &= \text{iFFT}(\mathbf{H}_{eq}) = \text{iFFT}\left(\frac{\mathbf{G}_{dry,wet}}{\mathbf{G}_{dry,dry}}\right)\end{aligned}\tag{A.7}$$

where iFFT is the inverse Fast Fourier Transform.

The window length has been empirically obtained to produce the impulse response with the lowest noise floor while still being sufficiently long compared to the reverberation times.

In contrast to most work on impulse response estimation and room impulse response inversion, in this case there is no reference or error measure to objectively evaluate the quality of the obtained impulse response. Convolving the dry signal with the EIR will rarely approximate the wet signal, due to condition 1).

While stereo reverberation generated from a monaural source is generally defined by two impulse responses (one for each channel), and stereo reverberation of a stereo source by four ($\mathbf{h}_{L \rightarrow L}$, $\mathbf{h}_{L \rightarrow R}$, ...), for the purpose of this study a single impulse response is extracted from the spectral division of the wet and dry signal, each summed to mono. It has been shown that with identical reverberation times and level, mono and stereo reverberation signals are perceived as having equal loudness regardless of the source material [144].

From this impulse response, it is possible to extract reverberation time measures such as the Early Decay Time (EDT). This is a particularly suitable feature as the calculated impulse responses are noisy. Furthermore, the EDT is more closely related to the conscious perception of reverberation, especially while the source is still playing during

the reverberation decay, as is the case here [233, 251, 268].

A.6.2 Equivalent Impulse Response analysis and results

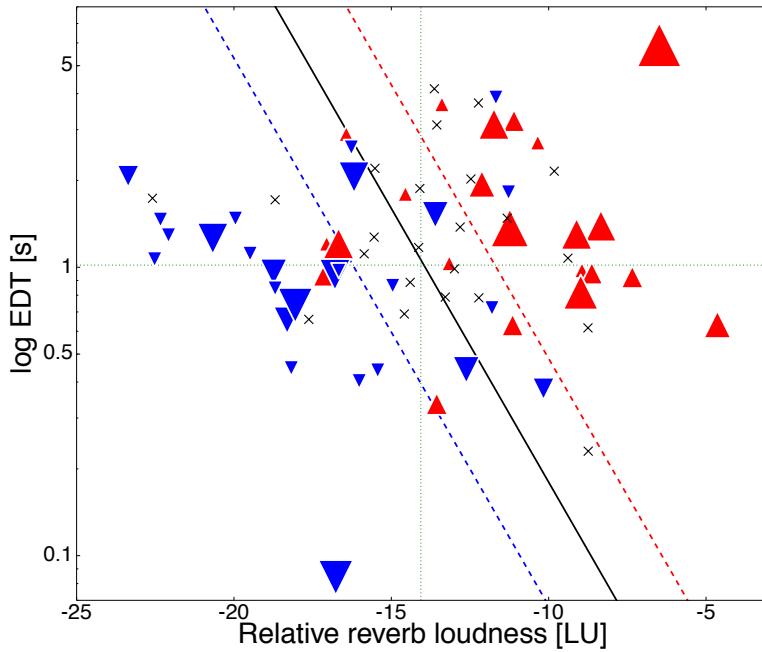


Figure A.4: Mixes where subjects noted an excess (red upwards triangle) or deficit (blue downwards triangle) of reverb, or neither (black X), as a function of the relative reverb loudness and the EDT of the reverb signal. Marker size is scaled by net number of subjects, and logistic regression decision boundaries are shown.

Figure A.4 shows all mixes as a function of their reverb loudness and reverb time, and labeled according to the net number of subjects who classified them as either “*Too much reverb*”, “*Not enough reverb*”, or *Neither*. The relative reverb loudness is as computed in Section A.5, and the EDT is calculated from the EIR using the decay method, as six times the time it takes for the decay curve to reach -10 dB [269]. The logarithm of the EDT is taken to better visualise a few large values, and this also makes the distribution normal.

As the dependent variable is a binary classification into ‘too reverberant’ or ‘not reverberant enough’, a logistic regression is performed based on the measurements of relative reverb loudness and EDT, for each assignment to either category by a subject, the results of which are shown in Table A.2. Comparing this to a restricted model with only the relative reverb loudness (RRL) as a predictor variable, a statistically significant increase is seen in the model fit (likelihood ratio $-2 \ln L_{both}/L_{RRL} = 7.749$, i.e. $p = .005$ on a χ^2 distribution). This shows the EDT is indeed helpful in explaining

the perception of the reverberation amount. As in [132], the perceived level of reverberation is more heavily influenced by the loudness than by the reverberation time. The decision boundaries at .25, .50 and .75 are shown in Figure A.4, along with the .50 decision boundaries for the individual predictor variables.

Table A.2: Logistic regression results

	Coeff.	SE	$P > z $	95% CI
RRL	0.4866	0.089	0.000	0.312 – 0.662
EDT	2.5619	1.043	0.014	0.519 – 4.605
Intercept	6.7767	1.282	0.000	4.263 – 9.290

Such a sharp transition between what is considered too reverberant and too dry again emphasises the importance of careful adjustment of reverb parameters. This is further supported by the observation in [262] that masking causes reverberation audibility to decrease by 4 dB for every dB decrease in reverberant level. The differences in reverberation time between the different mixes are mostly of the order of the JND [253], as was the case with the differences in relative reverb loudness.

It is customary in the analysis of acoustic impulse responses to perform octave analysis [233, 270, 271] and report the conventional measures for different octave bands. Furthermore, one could expect the reverberation time of certain octave bands to be more perceptually relevant than the broadband reverberation. However, none of these measures showed a notably better performance in accounting for the difference in perception of reverberation amount. Following similar reasoning, K-filtering was applied on the impulse response to account for the varying sensitivity of the auditory system with frequency [92], but again no significant improvement was measured.

A.7 Multi-group analysis

The importance of reverberation to the perception of mixes, measured as the proportion of comments which mention it at all, is similarly high when looking at other groups than McGill (as in Chapter 5). It is mentioned in 15% of all comments by subjects without music production experience, but this increases to 33% for mix engineers.

As before, preference ratings are lower on average for mixes which are deemed too reverberant, and significantly so for groups QMUL and UCP, see Figure A.5. At QMUL,

lack of reverberation also leads to a lower than average rating, which in turn is not demonstrably different from ratings associated with excessive reverb use. Subjects from the SMC group, by contrast, note a slightly average higher preference for mixes which they find ‘too dry’. Considering subjects of each expertise category separately, a consistent dislike for overly reverberant mixes can be seen across all levels.

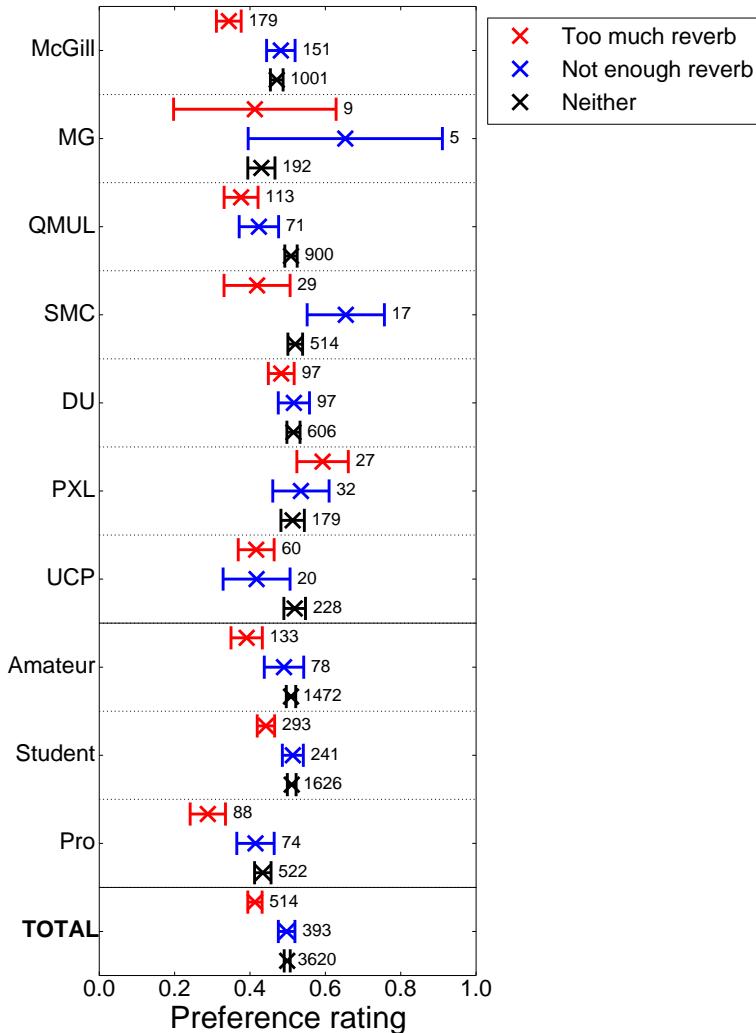


Figure A.5: Preference as a function of perceived reverberation amount, for different groups and levels of expertise

Figure A.6 shows the proportion of subjects per mix who reported an excess or shortage of perceived reverb, broken down by institution.

The net verdict on perceived amount of reverb (i.e. ‘too much’ or ‘not enough’) is consistent between groups, once again indicating that there is a general consensus on when a mix shows an excess or deficiency of reverb. An exception to this is mix 4J (No Prize), which one QMUL participant found to be overly dry, contesting one McGill

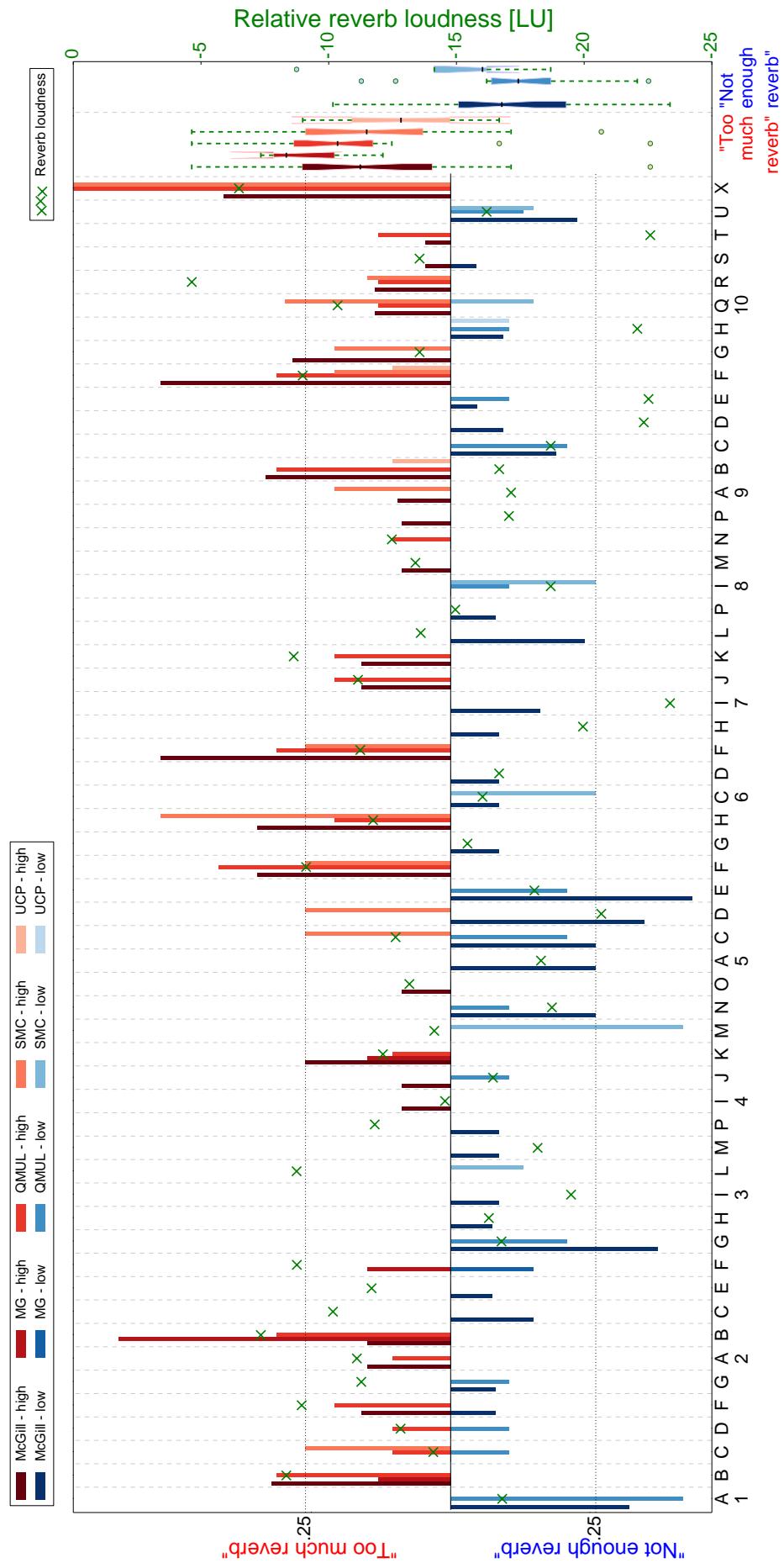


Figure A.6: Perceived amount of reverberation for different groups, compared to relative reverb loudness

participant. A surplus of reverb is also noted in mix 5C (Not Alone) by one SMC participant, whereas two QMUL and three McGill subjects agree that there is a deficit. The same SMC participant marks mix D of the same song as being too reverberant, challenging four McGill subjects. This could indicate an interpretation error, language-related confusion, or an unreliable participant.

Judging from the distribution of relative reverb loudness of these ‘wet’ and ‘dry’ mixes, -14 LU relative to the total mix loudness is still a value that separates most of the respective categories.

With this larger dataset, the EDT of the newly proposed Equivalent Impulse Response offers an even bigger increase in the logistic regression model fit (likelihood ratio $-2 \ln L_{both}/L_{RRL} = 10.600$, i.e. $p = .001$ on a χ^2 distribution), further supporting the utility of this measure.

A.8 Conclusion

Annotated subjective comments were analysed to determine the importance of reverberation in the perception of mixes, as well as to classify mixes as having too much or too little overall reverberation. Objective features believed to be relevant to reverberation were proposed and evaluated in terms of their ability to estimate this perceived amount of reverb. This study is different from previous work in that it examines reverb in a relevant music production context, where reverb is applied to multiple tracks in varying degrees and types. As such, it proves the proposed concept of obtaining knowledge from realistic mixes and perceptual evaluation thereof, utilising both objective and subjective data.

Although the perceptual evaluation experiment purposely did not mention reverberation as a feature to consider, it is commented on in 35% of the cases, confirming that differences in reverb use have a large impact on the perceived quality of a mix [144], as assessed by skilled listeners. To a large extent, the relative reverb loudness gives a suitable indication of how audible or objectionable reverberation is. These subjective judgements are better predicted by also considering reverb decay time, derived from a newly proposed *Equivalent Impulse Response* which captures reverberation character-

istics for a mixture of sources with varying degrees and types of reverb. Both measures are suitable for real-time applications such as automated reverberators or assistive interfaces.

The results further support the notion that upper and lower bounds of a set of mix parameters or features can be identified with reasonable confidence. The importance of careful parameter adjustment is evident from the limited range of acceptable feature values with regard to perceived amount of reverberation, when compared to the just-noticeable differences in both relative reverb loudness and the Equivalent Impulse Response's EDT. This study confirms previous findings that a perceived excess of reverberation typically has a more detrimental effect on subjective preference than when the reverberation level was indicated to be too low, suggesting it is better to err on the 'dry' side. Notwithstanding the less controlled nature of the presented approach, variance in its findings is significantly narrower than in similar work, likely due in part to proficiency of participants in both the mix experiment and subsequent perceptual evaluation.

Future implementations should take into account how reverberant the 'dry' signal is, particularly when the original tracks contain a significant amount of natural reverberation. Source separation or dereverberation could help achieve a more accurate estimation of the dry and wet sound.

Artificial reverberation is defined by far more attributes, objective and perceptual, than those covered in this section. Further features and parameters to consider include predelay [262], echo density [250], autocorrelation [51], and more sophisticated loudness features [248].

Bibliography

- [1] R. Izhaki, *Mixing Audio: Concepts, Practices and Tools*. Focal Press, 2008.
- [2] W. Moylan, *Understanding and Crafting the Mix: The Art of Recording*. Focal Press, 2nd ed., 2006.
- [3] A. Case, *Mix Smart: Pro Audio Tips For Your Multitrack Mix*. Focal Press, 2011.
- [4] F. Rumsey, “Mixing and artificial intelligence,” *Journal of the Audio Engineering Society*, vol. 61, pp. 806–809, October 2013.
- [5] B. Kolasinski, “A framework for automatic mixing using timbral similarity measures and genetic optimization,” in *Audio Engineering Society Convention 124*, May 2008.
- [6] G. Bocko, M. Bocko, D. Headlam, J. Lundberg, and G. Ren, “Automatic music production system employing probabilistic expert systems,” *Audio Engineering Society Convention 129*, November 2010.
- [7] A. T. Sabin and B. Pardo, “2DEQ: An intuitive audio equalizer,” in *ACM Creativity and Cognition*, October 2009.
- [8] R. Toulson, “Can we fix it? – The consequences of ‘fixing it in the mix’ with common equalisation techniques are scientifically evaluated,” *Journal of the Art of Record Production*, vol. 3, November 2008.
- [9] A. Pras, C. Guastavino, and M. Lavoie, “The impact of technological advances on recording studio practices,” *Journal of the American Society for Information Science and Technology*, vol. 64, pp. 612–626, January 2013.
- [10] D. Reed, “A perceptual assistant to do sound equalization,” in *Proceedings of the 5th International Conference on Intelligent User Interfaces*, pp. 212–218, January 2000.
- [11] D. Dugan, “Automatic microphone mixing,” *Journal of the Audio Engineering Society*, vol. 23, July/August 1975.
- [12] M. J. Terrell, A. Simpson, and M. Sandler, “The mathematics of mixing,” *Journal of the Audio Engineering Society*, vol. 62, pp. 4–13, January/February 2014.
- [13] M. J. Terrell and J. D. Reiss, “Automatic monitor mixing for live musical performance,” *Journal of the Audio Engineering Society*, vol. 57, pp. 927–936, November 2009.
- [14] M. J. Terrell and M. Sandler, “An offline, automatic mixing method for live music, incorporating multiple sources, loudspeakers, and room effects,” *Computer Music Journal*, vol. 36, pp. 37–54, May 2012.

- [15] M. J. Terrell, A. J. R. Simpson, and M. Sandler, “A perceptual audio mixing device,” in *Audio Engineering Society Convention 134*, May 2013.
- [16] E. Perez Gonzalez and J. D. Reiss, “Automatic gain and fader control for live mixing,” *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2009.
- [17] E. Perez Gonzalez and J. D. Reiss, “Improved control for selective minimization of masking using inter-channel dependency effects,” *Proceedings of the 11th International Conference on Digital Audio Effects (DAFx-08)*, September 2008.
- [18] J. Scott, M. Prockup, E. M. Schmidt, and Y. E. Kim, “Automatic multi-track mixing using linear dynamical systems,” in *Proceedings of the Sound and Music Computing Conference*, July 2011.
- [19] D. Ward, J. D. Reiss, and C. Athwal, “Multitrack mixing using a model of loudness and partial loudness,” in *Audio Engineering Society Convention 133*, October 2012.
- [20] A. Wilson and B. Fazenda, “An evolutionary computation approach to intelligent music production, informed by experimentally gathered domain knowledge,” in *2nd AES Workshop on Intelligent Music Production*, September 2016.
- [21] R. B. Dannenberg, “An intelligent multi-track audio editor,” in *Proceedings of the International Computer Music Conference*, August 2007.
- [22] E. Perez Gonzalez and J. D. Reiss, “An automatic maximum gain normalization technique with applications to audio mixing,” *Audio Engineering Society Convention 124*, May 2008.
- [23] S. Mansbridge, S. Finn, and J. D. Reiss, “Implementation and evaluation of autonomous multi-track fader control,” in *Audio Engineering Society Convention 132*, April 2012.
- [24] G. Wichern, A. Wishnick, A. Lukin, and H. Robertson, “Comparison of loudness features for automatic level adjustment in mixing,” in *Audio Engineering Society Convention 139*, October 2015.
- [25] E. Perez Gonzalez and J. D. Reiss, “A real-time semiautonomous audio panning system for music mixing,” *EURASIP Journal on Advances in Signal Processing*, May 2010.
- [26] S. Mansbridge, S. Finn, and J. D. Reiss, “An autonomous system for multi-track stereo pan positioning,” in *Audio Engineering Society Convention 133*, October 2012.
- [27] P. D. Pestana and J. D. Reiss, “A cross-adaptive dynamic spectral panning technique,” in *Proceedings of the 17th International Conference on Digital Audio Effects (DAFx-14)*, September 2014.
- [28] E. Perez Gonzalez and J. D. Reiss, “Automatic mixing: Live downmixing stereo panner,” in *Proceedings of the 10th International Conference on Digital Audio Effects (DAFx-07)*, September 2007.
- [29] S. Hafezi and J. D. Reiss, “Autonomous multitrack equalization based on masking reduction,” *Journal of the Audio Engineering Society*, vol. 63, pp. 312–323, May 2015.

- [30] E. Perez Gonzalez and J. D. Reiss, "Automatic equalization of multi-channel audio using cross-adaptive methods," *Audio Engineering Society Convention 127*, October 2009.
- [31] Z. Ma, J. D. Reiss, and D. A. A. Black, "Implementation of an intelligent equalization tool using Yule-Walker for music mixing and mastering," in *Audio Engineering Society Convention 134*, May 2013.
- [32] S. I. Mimalakis, K. Drossos, A. Floros, and D. Katerelos, "Automated tonal balance enhancement for audio mastering applications," in *Audio Engineering Society Convention 134*, May 2013.
- [33] S. I. Mimalakis, K. Drossos, T. Virtanen, and G. Schuller, "Deep neural networks for dynamic range compression in mastering applications," in *Audio Engineering Society Convention 140*, May 2016.
- [34] D. Giannoulis, M. Massberg, and J. D. Reiss, "Parameter automation in a dynamic range compressor," *Journal of the Audio Engineering Society*, vol. 61, pp. 716–726, October 2013.
- [35] Z. Ma, B. De Man, P. D. Pestana, D. A. A. Black, and J. D. Reiss, "Intelligent multitrack dynamic range compression," *Journal of the Audio Engineering Society*, vol. 63, pp. 412–426, June 2015.
- [36] A. Mason, N. Jillings, Z. Ma, J. D. Reiss, and F. Melchior, "Adaptive audio reproduction using personalized compression," in *Audio Engineering Society Conference: 57th International Conference: The Future of Audio Entertainment Technology – Cinema, Television and the Internet*, March 2015.
- [37] J. A. Maddams, S. Finn, and J. D. Reiss, "An autonomous method for multi-track dynamic range compression," in *Proceedings of the 15th International Conference on Digital Audio Effects (DAFx-12)*, September 2012.
- [38] M. Hilsamer and S. Herzog, "A statistical approach to automated offline dynamic processing in the audio mastering process.,," in *Proceedings of the 17th International Conference on Digital Audio Effects (DAFx-14)*, September 2014.
- [39] E. Vickers, "Automatic long-term loudness and dynamics matching," in *Audio Engineering Society Convention 111*, November 2001.
- [40] E. T. Chourdakis and J. D. Reiss, "Automatic control of a digital reverberation effect using hybrid models," in *Audio Engineering Society Conference: 60th International Conference: DREAMS (Dereverberation and Reverberation of Audio, Music, and Speech)*, January 2016.
- [41] E. T. Chourdakis and J. D. Reiss, "A machine learning approach to application of intelligent artificial reverberation," *Journal of the Audio Engineering Society*, vol. 65, January/February 2017.
- [42] B. De Man and J. D. Reiss, "Adaptive control of amplitude distortion effects," in *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*, January 2014.
- [43] E. Perez Gonzalez and J. D. Reiss, "Determination and correction of individual channel time offsets for signals involved in an audio mixture," *Audio Engineering Society Convention 125*, October 2008.

- [44] A. Clifford and J. D. Reiss, “Reducing comb filtering on different musical instruments using time delay estimation,” *Journal of the Art of Record Production*, vol. 5, July 2011.
- [45] A. Clifford and J. D. Reiss, “Calculating time delays of multiple active sources in live sound,” in *Audio Engineering Society Convention 129*, November 2010.
- [46] N. Jillings, A. Clifford, and J. D. Reiss, “Performance optimization of GCC-PHAT for delay and polarity correction under real world conditions,” in *Audio Engineering Society Convention 134*, May 2013.
- [47] A. Clifford and J. D. Reiss, “Proximity effect detection for directional microphones,” in *Audio Engineering Society Convention 131*, October 2011.
- [48] S. Julstrom and T. Tichy, “Direction-sensitive gating: A new approach to automatic mixing,” in *Audio Engineering Society Convention 73*, March 1983.
- [49] A. Clifford and J. D. Reiss, “Microphone interference reduction in live sound,” in *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx-11)*, September 2011.
- [50] M. J. Terrell, J. D. Reiss, and M. Sandler, “Automatic noise gate settings for drum recordings containing bleed from secondary sources,” *EURASIP Journal on Advances in Signal Processing*, December 2010.
- [51] P. Pestana and J. D. Reiss, “Intelligent audio production strategies informed by best practices,” in *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*, January 2014.
- [52] A. Wilson and B. M. Fazenda, “Navigating the mix-space: Theoretical and practical level-balancing technique in multitrack music mixtures,” in *Proceedings of the 12th Sound and Music Computing Conference*, July 2015.
- [53] E. Deruty, “Goal-oriented mixing,” in *2nd AES Workshop on Intelligent Music Production*, September 2016.
- [54] R. J. Burgess, *The Art of Music Production: The Theory and Practice*. Oxford University Press, 2013.
- [55] A. T. Sabin and B. Pardo, “Rapid learning of subjective preference in equalization,” in *Audio Engineering Society Convention 125*, October 2008.
- [56] A. T. Sabin and B. Pardo, “A method for rapid personalization of audio equalization parameters,” in *ACM International Conference on Multimedia*, October 2009.
- [57] Z. Rafi and B. Pardo, “Learning to control a reverberator using subjective perceptual descriptors,” in *10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, October 2009.
- [58] A. T. Sabin, Z. Rafi, and B. Pardo, “Weighting-function-based rapid mapping of descriptors to audio processing parameters,” *Journal of the Audio Engineering Society*, vol. 59, pp. 419–430, June 2011.
- [59] M. Cartwright and B. Pardo, “Social-EQ: Crowdsourcing an equalization descriptor map,” *14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, November 2013.

- [60] P. Seetharaman and B. Pardo, “Crowdsourcing a reverberation descriptor map,” in *ACM International Conference on Multimedia*, November 2014.
- [61] P. Seetharaman and B. Pardo, “Reverbalize: A crowdsourced reverberation controller,” in *ACM International Conference on Multimedia*, November 2014.
- [62] T. Zheng, P. Seetharaman, and B. Pardo, “SocialFX: Studying a crowdsourced folksonomy of audio effects terms,” in *ACM International Conference on Multimedia*, October 2016.
- [63] S. Fenton and J. Wakefield, “Objective profiling of perceived punch and clarity in produced music,” in *Audio Engineering Society Convention 132*, April 2012.
- [64] S. Fenton, H. Lee, and J. Wakefield, “Elicitation and objective grading of punch within produced music,” in *Audio Engineering Society Convention 136*, April 2014.
- [65] J. Scott and Y. E. Kim, “Instrument identification informed multi-track mixing,” in *14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, 2013.
- [66] A. Wilson and B. M. Fazenda, “Perception of audio quality in productions of popular music,” *Journal of the Audio Engineering Society*, vol. 64, pp. 23–34, January/February 2016.
- [67] P. D. Pestana, Z. Ma, J. D. Reiss, A. Barbosa, and D. A. A. Black, “Spectral characteristics of popular commercial recordings 1950–2010,” in *Audio Engineering Society Convention 135*, October 2013.
- [68] E. Deruty, F. Pachet, and P. Roy, “Human-made rock mixes feature tight relations between spectrum and loudness,” *Journal of the Audio Engineering Society*, vol. 62, pp. 643–653, October 2014.
- [69] E. Deruty and F. Pachet, “The MIR perspective on the evolution of dynamics in mainstream music,” in *16th International Society for Music Information Retrieval Conference (ISMIR 2015)*, October 2015.
- [70] E. Deruty and D. Tardieu, “About dynamic processing in mainstream music,” *Journal of the Audio Engineering Society*, vol. 62, pp. 42–55, January/February 2014.
- [71] A. Wilson and B. Fazenda, “101 mixes: A statistical analysis of mix-variation in a dataset of multi-track music mixes,” in *Audio Engineering Society Convention 139*, October 2015.
- [72] A. Wilson and B. Fazenda, “Variation in multitrack mixes: Analysis of low-level audio signal features,” *Journal of the Audio Engineering Society*, vol. 64, pp. 466–473, July/August 2016.
- [73] T. Porcello, “Speaking of sound language and the professionalization of sound-recording engineers,” *Social Studies of Science*, vol. 34, pp. 733–758, October 2004.
- [74] A. Pras and C. Guastavino, “The role of music producers and sound engineers in the current recording context, as perceived by young professionals,” *Musicae Scientiae*, vol. 15, pp. 73–95, March 2011.
- [75] P. White, “Automation for the people,” *Sound On Sound*, October 2008.

- [76] J. D. Reiss, “Intelligent systems for mixing multichannel audio,” in *17th International Conference on Digital Signal Processing (DSP)*, July 2011.
- [77] S. Mecklenburg and J. Loviscach, “subjEQt: Controlling an equalizer through subjective terms,” in *CHI*, pp. 1109–1114, April 2006.
- [78] J. Ford, M. Cartwright, and B. Pardo, “MixViz: A tool to visualize masking in audio mixes,” in *Audio Engineering Society Convention 139*, October 2015.
- [79] B. De Man and J. D. Reiss, “Analysis of peer reviews in music production,” *Journal of the Art of Record Production*, vol. 10, July 2015.
- [80] B. De Man and J. D. Reiss, “A semantic approach to autonomous mixing,” *Journal of the Art of Record Production*, vol. 8, December 2013.
- [81] B. De Man and J. D. Reiss, “A knowledge-engineered autonomous mixing system,” in *Audio Engineering Society Convention 135*, October 2013.
- [82] M. J. Terrell, S. Mansbridge, J. D. Reiss, and B. De Man, “System and method for performing automatic audio production using semantic data,” Mar. 5 2015. US Patent App. 14/471,758.
- [83] T. Wilmering, G. Fazekas, and M. Sandler, “High-level semantic metadata for the control of multitrack adaptive digital audio effects,” in *Audio Engineering Society Convention 133*, October 2012.
- [84] M. Senior, *Mixing Secrets*. www.cambridge-mt.com/ms-mtk.htm: Taylor & Francis, 2012.
- [85] A. Case, *Sound FX: Unlocking the Creative Potential of Recording Studio Effects*. Taylor & Francis, 2012.
- [86] B. Owsinski, *The Mixing Engineer’s Handbook*. Course Technology, 2nd ed., 2006.
- [87] K. Coryat, *Guerrilla Home Recording: How to Get Great Sound from Any Studio (No Matter How Weird or Cheap Your Gear Is)*. MusicPro guides, Hal Leonard Corporation, 2008.
- [88] D. Gibson, *The Art Of Mixing: A Visual Guide to Recording, Engineering, and Production*. Thomson Course Technology, 2005.
- [89] P. White, *Basic Effects & Processors*. The Basic Series, Music Sales, 2000.
- [90] P. White, “Compressors: Exploration,” *Sound On Sound*, April 1997.
- [91] T. Wilmering, G. Fazekas, and M. Sandler, “Towards ontological representations of digital audio effects,” in *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx-11)*, September 2011.
- [92] Recommendation ITU-R BS.1770-4, “Algorithms to measure audio programme loudness and true-peak audio level,” *Radiocommunication Sector of the International Telecommunication Union*, October 2015.
- [93] M. Senior, “Sound advice: Should I EQ first or compress first?,” *Sound On Sound*, October 2007.
- [94] D. Giannoulis, M. Massberg, and J. D. Reiss, “Digital dynamic range compressor design – A tutorial and analysis,” *Journal of the Audio Engineering Society*, vol. 60, pp. 399–408, June 2012.

- [95] G. Massenburg, “Parametric equalization,” in *Audio Engineering Society Convention 42*, May 1972.
- [96] P. White, *Basic Mixers*. The Basic Series, Music Sales, 1999.
- [97] M. Cousins and R. Hepworth-Sawyer, *Practical Mastering: A Guide to Mastering in the Modern Studio*. Taylor & Francis, 2013.
- [98] G. Waddell, *Complete Audio Mastering: Practical Techniques*. McGraw-Hill Education, 2013.
- [99] B. Katz, *Mastering Audio*. Focal Press, 2002.
- [100] D. M. Huber and R. Runstein, *Modern Recording Techniques*. Taylor & Francis, 2013.
- [101] Mixerman, *Zen and the Art of Mixing*. Hal Leonard Corporation, 2010.
- [102] E. Skovenborg, “Development of semantic scales for music mastering,” in *Audio Engineering Society Convention 141*, September 2016.
- [103] B. De Man and J. D. Reiss, “APE: Audio Perceptual Evaluation toolbox for MATLAB,” in *Audio Engineering Society Convention 136*, April 2014.
- [104] B. De Man, M. Mora-Mcginity, G. Fazekas, and J. D. Reiss, “The Open Multitrack Testbed,” in *Audio Engineering Society Convention 137*, October 2014.
- [105] B. De Man and J. D. Reiss, “The Open Multitrack Testbed: Features, content and use cases,” in *2nd AES Workshop on Intelligent Music Production*, September 2016.
- [106] E. Perez Gonzalez and J. D. Reiss, “Automatic mixing,” in *DAFX: Digital Audio Effects, Second Edition*, pp. 523–549, John Wiley, 2011.
- [107] S. Hargreaves, A. Klapuri, and M. Sandler, “Structural segmentation of multitrack audio,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 2637–2647, December 2012.
- [108] J. D. Reiss and M. Sandler, “Audio issues in MIR evaluation,” in *5th International Society for Music Information Retrieval Conference (ISMIR 2004)*, October 2004.
- [109] J. Salamon, “Pitch analysis for active music discovery,” in *33rd International Conference on Machine Learning*, June 2016.
- [110] T. Seay, “Primary sources in music production research and education: Using the Drexel University Audio Archives as an institutional model,” *Journal of the Art of Record Production*, vol. 5, July 2011.
- [111] M. Cartwright, B. Pardo, and J. D. Reiss, “MIXPLORATION: Rethinking the audio mixer interface,” in *International Conference on Intelligent User Interfaces*, February 2014.
- [112] J. Fritsch and M. D. Plumbley, “Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP’13)*, pp. 888–891, May 2013.
- [113] E. Vincent, R. Gribonval, C. Fevotte, A. Nesbit, M. D. Plumbley, M. E. Davies, and L. Daudet, “BASS-dB: The blind audio source separation evaluation database,” in Available via <http://www.irisa.fr/metiss/BASS-dB>, 2010.

- [114] J. Scott and Y. E. Kim, “Analysis of acoustic features for automated multi-track mixing,” in *12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, October 2011.
- [115] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. Bello, “MedleyDB: a multitrack dataset for annotation-intensive MIR research,” in *15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, October 2014.
- [116] K. McNally, “Music archives in higher education: A case study,” *Journal of the Art of Record Production*, vol. 10, July 2015.
- [117] K. McNally, T. Seay, and P. Thompson, “What the masters teach us: Multitrack audio archives and popular music education,” in *Perspectives and Practices in Popular Music Education* (Z. Moir, B. Powell, and G. D. Smith, eds.), 2017.
- [118] S. I. Mimalakis, E. Cano, J. Abeßer, and G. Schuller, “New sonorities for jazz recordings: Separation and mixing using deep neural networks,” in *2nd AES Workshop on Intelligent Music Production*, September 2016.
- [119] C. Greenhalgh, A. Hazzard, S. McGrath, and S. Benford, “GeoTracks: Adaptive music for everyday journeys,” in *ACM International Conference on Multimedia*, pp. 42–46, October 2016.
- [120] D. Moffat and J. D. Reiss, “Implementation and assessment of joint source separation and dereverberation,” in *Audio Engineering Society Conference: 60th International Conference: DREAMS (Dereverberation and Reverberation of Audio, Music, and Speech)*, January 2016.
- [121] K. Arimoto, “Identification of drum overhead-microphone tracks in multi-track recordings,” in *2nd AES Workshop on Intelligent Music Production*, September 2016.
- [122] G. Fazekas and T. Wilmering, “Semantic web and semantic audio technologies,” in *Audio Engineering Society Convention 132*, April 2012.
- [123] G. Fazekas and M. Sandler, “The Studio Ontology framework,” in *12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, October 2011.
- [124] Y. Raimond, S. A. Abdallah, M. Sandler, and F. Giasson, “The Music Ontology,” in *8th International Society for Music Information Retrieval Conference (ISMIR 2007)*, pp. 417–422, September 2007.
- [125] B. De Man and J. D. Reiss, “A pairwise and multiple stimuli approach to perceptual evaluation of microphone types,” in *Audio Engineering Society Convention 134*, May 2013.
- [126] N. Jillings, B. De Man, D. Moffat, J. D. Reiss, and R. Stables, “Web audio evaluation tool: A framework for subjective assessment of audio,” in *2nd Web Audio Conference*, April 2016.
- [127] N. Jillings, D. Moffat, B. De Man, and J. D. Reiss, “Web Audio Evaluation Tool: A browser-based listening test environment,” in *12th Sound and Music Computing Conference*, July 2015.

- [128] B. De Man, N. Jillings, D. Moffat, J. D. Reiss, and R. Stables, "Subjective comparison of music production practices using the Web Audio Evaluation Tool," in *2nd AES Workshop on Intelligent Music Production*, September 2016.
- [129] S. Bech and N. Zacharov, *Perceptual Audio Evaluation - Theory, Method and Application*. John Wiley & Sons, 2007.
- [130] Recommendation ITU-R BS.1534-3, "Method for the subjective assessment of intermediate quality level of coding systems," *International Telecommunication Union*, 2003.
- [131] F. E. Toole and S. Olive, "Hearing is believing vs. believing is hearing: Blind vs. sighted listening tests, and other interesting things," in *Audio Engineering Society Convention 97*, November 1994.
- [132] B. Cauchi, H. Javed, T. Gerkmann, S. Doclo, S. Goetze, and P. Naylor, "Perceptual and instrumental evaluation of the perceived level of reverberation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'16)*, pp. 629–633, March 2016.
- [133] S. Olive and T. Welti, "The relationship between perception and measurement of headphone sound quality," in *Audio Engineering Society Convention 133*, October 2012.
- [134] J. Francombe, R. Mason, M. Dewhirst, and S. Bech, "Elicitation of attributes for the evaluation of audio-on-audio interference," *Journal of the Acoustical Society of America*, vol. 136, pp. 2630–2641, November 2014.
- [135] S. P. Lipshitz and J. Vanderkooy, "The Great Debate: Subjective evaluation," *Journal of the Audio Engineering Society*, vol. 29, pp. 482–491, July/August 1981.
- [136] A. Pearce, T. Brookes, and M. Dewhirst, "Validation of experimental methods to record stimuli for microphone comparisons," in *Audio Engineering Society Convention 139*, October 2015.
- [137] A. Pearce, T. Brookes, M. Dewhirst, and R. Mason, "Eliciting the most prominent perceived differences between microphones," *Journal of the Acoustical Society of America*, vol. 139, no. 5, pp. 2970–2981, 2016.
- [138] P. D. Pestana, J. D. Reiss, and A. Barbosa, "Loudness measurement of multi-track audio content using modifications of ITU-R BS.1770," in *Audio Engineering Society Convention 134*, May 2013.
- [139] D. Matz, E. Cano, and J. Abeßer, "New sonorities for early jazz recordings using sound source separation and automatic mixing tools," in *16th International Society for Music Information Retrieval Conference (ISMIR 2015)*, October 2015.
- [140] S. E. Olive, "Some new evidence that teenagers and college students may prefer accurate sound reproduction," in *Audio Engineering Society Convention 132*, April 2012.
- [141] C. Pike, T. Brookes, and R. Mason, "Auditory adaptation to loudspeakers and listening room acoustics," in *Audio Engineering Society Convention 135*, October 2013.
- [142] C. Völker and R. Huber, "Adaptions for the Multi Stimulus test with Hidden Reference and Anchor (MUSHRA) for elder and technical unexperienced participants," in *DAGA*, March 2015.

- [143] S. Bech, "Listening tests on loudspeakers: A discussion of experimental procedures and evaluation of the response data," in *Audio Engineering Society Conference: 8th International Conference: The Sound of Audio*, May 1990.
- [144] J. Paulus, C. Uhle, and J. Herre, "Perceived level of late reverberation in speech and music," in *Audio Engineering Society Convention 130*, May 2011.
- [145] F. Rumsey, "Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm," *Journal of the Audio Engineering Society*, vol. 50, pp. 651–666, September 2002.
- [146] K. Hermes, T. Brookes, and C. Hummersone, "The influence of dumping bias on timbral clarity ratings," in *Audio Engineering Society Convention 139*, October 2015.
- [147] J. Berg and F. Rumsey, "Spatial attribute identification and scaling by repertory grid technique and other methods," in *Audio Engineering Society Conference: 16th International Conference: Spatial Sound Reproduction*, March 1999.
- [148] J. Berg and F. Rumsey, "Identification of quality attributes of spatial audio by repertory grid technique," *Journal of the Audio Engineering Society*, vol. 54, pp. 365–379, May 2006.
- [149] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkman, and S. Weinzierl, "A spatial audio quality inventory (SAQI)," *Acta Acustica united with Acustica*, vol. 100, pp. 984–994, September/October 2014.
- [150] U. Jekosch, "Basic concepts and terms of 'quality', reconsidered in the context of product-sound quality," *Acta Acustica united with Acustica*, vol. 90, pp. 999–1006, November/December 2004.
- [151] S. Zielinski, "On some biases encountered in modern listening tests," in *Spatial Audio & Sensory Evaluation Techniques*, April 2006.
- [152] S. Zielinski, "On some biases encountered in modern audio quality listening tests (part 2): Selected graphical examples and discussion," *Journal of the Audio Engineering Society*, vol. 64, pp. 55–74, January/February 2016.
- [153] F. Rumsey, "New horizons in listening test design," *Journal of the Audio Engineering Society*, vol. 52, pp. 65–73, January/February 2004.
- [154] A. Wilson and B. M. Fazenda, "Relationship between hedonic preference and audio quality in tests of music production quality," in *8th International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, June 2016.
- [155] S. Olive, "Workshop on Perceptual Evaluation Interface Design (chair Brecht De Man)," in *Audio Engineering Society Convention 140*, June 2016.
- [156] W. L. Martens and A. Marui, "Constructing individual and group timbre spaces for sharpness-matched distorted guitar timbres," in *Audio Engineering Society Convention 119*, October 2005.
- [157] J. Berg and F. Rumsey, "In search of the spatial dimensions of reproduced sound: Verbal protocol analysis and cluster analysis of scaled verbal descriptors," in *Audio Engineering Society Convention 108*, February 2000.
- [158] N. Zacharov and T. H. Pedersen, "Spatial sound attributes – development of a common lexicon," in *Audio Engineering Society Convention 139*, October 2015.

- [159] J. Berg and F. Rumsey, "Validity of selected spatial attributes in the evaluation of 5-channel microphone techniques," in *Audio Engineering Society Convention 112*, April 2002.
- [160] J. Mycroft, J. D. Reiss, and T. Stockman, "The influence of graphical user interface design on critical listening skills," *Proceedings of the Sound and Music Computing Conference*, 2013.
- [161] B. Leonard, R. King, and G. Sikora, "The effect of playback system on reverberation level preference," in *Audio Engineering Society Convention 134*, May 2013.
- [162] N. Sakamoto, T. Gotoh, and Y. Kimura, "On out-of-head localization in headphones listening," *Journal of the Audio Engineering Society*, vol. 24, pp. 710–716, November 1976.
- [163] M. Schoeffler, F.-R. Stöter, H. Bayerlein, B. Edler, and J. Herre, "An experiment about estimating the number of instruments in polyphonic music: A comparison between internet and laboratory results," in *14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, November 2013.
- [164] Recommendation ITU-T P.830, "Subjective performance assessment of telephone-band and wideband digital codecs," *International Telecommunication Union*, 1996.
- [165] A.-R. Tereping, "Listener preference for concert sound levels: Do louder performances sound better?," *Journal of the Audio Engineering Society*, vol. 64, pp. 138–146, March 2016.
- [166] F. E. Toole, "Listening tests – Turning opinion into fact," *Journal of the Audio Engineering Society*, vol. 30, pp. 431–445, June 1982.
- [167] J. Paulus, C. Uhle, J. Herre, and M. Höpfel, "A study on the preferred level of late reverberation in speech and music," in *Audio Engineering Society Conference: 60th International Conference: DREAMS (Dereverberation and Reverberation of Audio, Music, and Speech)*, January 2016.
- [168] C. Bussey, M. J. Terrell, R. Rahman, and M. Sandler, "Metadata features that affect artificial reverberator intensity," in *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*, January 2014.
- [169] Recommendation ITU-R BS.1116-3, "Methods for the subjective assessment of small impairments in audio systems," *Radiocommunication Sector of the International Telecommunication Union*, February 2015.
- [170] W. G. Gardner and D. Griesinger, "Reverberation level matching experiments," in *Proceedings of the Sabine Centennial Symposium*, June 1994.
- [171] J. D. Reiss, "A meta-analysis of high resolution audio perceptual evaluation," *Journal of the Audio Engineering Society*, vol. 64, pp. 364–379, June 2016.
- [172] C. Gribben and H. Lee, "Toward the development of a universal listening test interface generator in Max," in *Audio Engineering Society Convention 138*, May 2015.
- [173] S. Kraft and U. Zölzer, "BeagleJS: HTML5 and JavaScript based framework for the subjective evaluation of audio quality," in *Linux Audio Conference*, May 2014.

- [174] E. Vincent, M. G. Jafari, and M. D. Plumbley, "Preliminary guidelines for subjective evaluation of audio source separation algorithms," in *UK ICA Research Network Workshop*, 2006.
- [175] A. V. Giner, "Scale – A software tool for listening experiments," in *AIA/DAGA Conference on Acoustics*, March 2013.
- [176] S. Ciba, A. Włodarski, and H.-J. Maempel, "WhisPER – A new tool for performing listening tests," in *Audio Engineering Society Convention 126*, May 2009.
- [177] J. Berg, "OPAQUE – A tool for the elicitation and grading of audio quality attributes," in *Audio Engineering Society Convention 118*, May 2005.
- [178] J. Wingstedt, M. Liljedahl, S. Lindberg, and J. Berg, "REMUPP: An interactive tool for investigating musical properties and relations," in *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME05)*, pp. 232–235, May 2005.
- [179] J. Hynninen and N. Zacharov, "GuineaPig - A generic subjective test system for multichannel audio," in *Audio Engineering Society Convention 106*, May 1999.
- [180] S. Olive, "A new listener training software application," in *Audio Engineering Society Convention 110*, May 2001.
- [181] V. Rioux, *Sound Quality of Flue Organ Pipes - An Interdisciplinary Study on the Art of Voicing*. PhD thesis, Department of Applied Acoustics, Chalmers University of Technology, Sweden, 2001.
- [182] G. Durr, L. Peixoto, M. Souza, R. Tanoue, and J. D. Reiss, "Implementation and evaluation of dynamic level of audio detail," in *Audio Engineering Society Conference: 56th International Conference: Audio for Games*, February 2015.
- [183] A. A. Osses Vecchi, A. Chaigne, and A. G. Kohlrausch, "Assessing the acoustic similarity of different pianos using an instrument-in-noise test," in *International Symposium on Musical and Room Acoustics*, September 2016.
- [184] M. Schoeffler, F.-R. Stöter, B. Edler, and J. Herre, "Towards the next generation of web-based experiments: A case study assessing basic audio quality following the ITU-R recommendation BS. 1534 (MUSHRA)," in *1st Web Audio Conference*, January 2015.
- [185] U.-D. Reips, "Standards for internet-based experimenting," *Experimental psychology*, vol. 49, no. 4, pp. 243–256, 2002.
- [186] M. Cartwright, B. Pardo, G. J. Mysore, and M. Hoffman, "Fast and easy crowdsourced perceptual audio evaluation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'16)*, March 2016.
- [187] Recommendation ITU-R BS. 562-3, "Subjective assessment of sound quality," *International Telecommunication Union*, 1997.
- [188] D. R. Peryam and N. F. Girardot, "Advanced taste-test method," *Food Engineering*, vol. 24, no. 7, pp. 58–61, 1952.
- [189] D. Clark, "High-resolution subjective testing using a double-blind comparator," *Journal of the AES*, vol. 30, no. 5, pp. 330–338, 1982.
- [190] Recommendation ITU-T P. 800, "Methods for subjective determination of transmission quality," *International Telecommunication Union*, 1996.

- [191] R. Likert, "A technique for the measurement of attitudes," *Archives of Psychology*, 1932.
- [192] Recommendation ITU-R BS.1534-1, "Method for the subjective assessment of intermediate quality levels of coding systems," *International Telecommunication Union*, 2003.
- [193] H. A. David, *The method of paired comparisons*, vol. 12. DTIC Document, 1963.
- [194] G. C. Pascoe and C. C. Attkisson, "The evaluation ranking scale: A new methodology for assessing satisfaction," *Evaluation and program planning*, vol. 6, no. 3, pp. 335–347, 1983.
- [195] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "SampleRNN: An unconditional end-to-end neural audio generation model," in *5th International Conference on Learning Representations*, April 2017.
- [196] T. Walton, M. Evans, D. Kirk, and F. Melchior, "Does environmental noise influence preference of background-foreground audio balance?," in *Audio Engineering Society Convention 141*, September 2016.
- [197] L. Mengual, D. Moffat, and J. D. Reiss, "Modal synthesis of weapon sounds," in *Audio Engineering Society Conference: 61st International Conference: Audio for Games*, February 2016.
- [198] R. Schatz, S. Egger, and K. Masuch, "The impact of test duration on user fatigue and reliability of subjective quality ratings," *Journal of the Audio Engineering Society*, vol. 60, pp. 63–73, January/February 2012.
- [199] B. De Man, B. Leonard, R. King, and J. D. Reiss, "An analysis and evaluation of audio features for multitrack music mixtures," in *15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, October 2014.
- [200] B. De Man and J. D. Reiss, "Crowd-sourced learning of music production practices through large-scale perceptual evaluation of mixes," in *Innovation in Music II* (R. Hepworth-Sawyer, J. Hodgson, J. L. Paterson, and R. Toulson, eds.), Future Technology Press, 2016.
- [201] E. Vickers, "The loudness war: Background, speculation, and recommendations," *Audio Engineering Society Convention 129*, November 2010.
- [202] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, pp. 293–302, July 2002.
- [203] G. Tzanetakis, R. Jones, and K. McNally, "Stereo panning features for classifying recording production style," in *8th International Society for Music Information Retrieval Conference (ISMIR 2007)*, September 2007.
- [204] O. Lartillot and P. Toivainen, "MIR in Matlab (II): A toolbox for musical feature extraction from audio," in *8th International Society for Music Information Retrieval Conference (ISMIR 2007)*, September 2007.
- [205] E. Skovenborg and S. H. Nielsen, "Evaluation of different loudness models with music and speech material," in *Audio Engineering Society Convention 117*, October 2004.
- [206] G. A. Soulodre, "Evaluation of objective loudness meters," in *Audio Engineering Society Convention 116*, May 2004.

- [207] B. G. Crockett, A. Seefeldt, and M. Smithers, “A new objective measure of perceived loudness,” in *Audio Engineering Society Convention 117*, October 2004.
- [208] P. Pestana, *Automatic mixing systems using adaptive digital audio effects*. PhD thesis, Catholic University of Portugal, 2013.
- [209] D. Ronan, B. De Man, H. Gunes, and J. D. Reiss, “The impact of subgrouping practices on the perception of multitrack mixes,” in *Audio Engineering Society Convention 139*, October 2015.
- [210] D. Ronan, H. Gunes, D. Moffat, and J. D. Reiss, “Automatic subgrouping of multitrack audio,” in *18th International Conference on Digital Audio Effects (DAFx-15)*, November 2015.
- [211] B. De Man, M. Boerum, B. Leonard, G. Massenburg, R. King, and J. D. Reiss, “Perceptual evaluation of music mixing practices,” in *Audio Engineering Society Convention 138*, May 2015.
- [212] E. Skovborg, “Measures of microdynamics,” in *Audio Engineering Society Convention 137*, October 2014.
- [213] Recommendation ITU-R BS.1771-1, “Requirements for loudness and true-peak indicating meters,” *Radiocommunication Sector of the International Telecommunication Union*, 2012.
- [214] EBU, “Loudness Range: A measure to supplement loudness normalisation in accordance with EBU R 128,” *European Broadcasting Union*, August 2011.
- [215] Pleasurize Music Foundation, “TT Dynamic Range Meter,” *webpage* (<http://dynamicrange.de>), 2013.
- [216] E. Skovborg, “Loudness Range (LRA) – Design and evaluation,” in *Audio Engineering Society Convention 132*, April 2012.
- [217] S. H. Nielsen and T. Lund, “Overload in signal conversion,” in *Audio Engineering Society Conference: 23rd International Conference: Signal Processing in Audio Recording and Reproduction*, May 2003.
- [218] C. Faller and M. Erne, “Modifying stereo recordings using acoustic information obtained with spot recordings,” in *Audio Engineering Society Convention 118*, May 2005.
- [219] M. Barthet and M. Sandler, “On the effect of reverberation on musical instrument automatic recognition,” in *Audio Engineering Society Convention 128*, May 2010.
- [220] R. Stables, S. Enderby, B. De Man, G. Fazekas, and J. D. Reiss, “SAFE: A system for the extraction and retrieval of semantic audio descriptors,” in *15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, October 2014.
- [221] R. Stables, B. De Man, S. Enderby, J. D. Reiss, G. Fazekas, and T. Wilmering, “Semantic description of timbral transformations in music production,” in *ACM International Conference on Multimedia*, October 2016.
- [222] J. Dattorro, “Effect design, part 1: Reverberator and other filters,” *Journal of the Audio Engineering Society*, vol. 45, pp. 660–684, September 1997.
- [223] J. Bullock, “libxtract: A lightweight library for audio feature extraction,” in *Proceedings of the International Computer Music Conference*, August 2007.

- [224] M. F. Porter, “An algorithm for suffix stripping,” *Program*, vol. 14, no. 3, pp. 130–137, 1980.
- [225] T. H. Pedersen and N. Zacharov, “The development of a sound wheel for reproduced sound,” in *Audio Engineering Society Convention 138*, May 2015.
- [226] M. E. Altinsoy and U. Jekosch, “The semantic space of vehicle sounds: Developing a semantic differential with regard to customer perception,” *Journal of the Audio Engineering Society*, vol. 60, pp. 13–20, January/February 2012.
- [227] J. H. Ward Jr., “Hierarchical grouping to optimize an objective function,” *Journal of the American Statistical Association*, vol. 58, pp. 236–244, March 1963.
- [228] T. A. Letsche and M. W. Berry, “Large-scale information retrieval with latent semantic indexing,” *Information sciences*, vol. 100, pp. 105–137, August 1997.
- [229] S. Zagorski-Thomas, “The US vs the UK sound: Meaning in music production in the 1970s,” in *The Art of Record Production: An Introductory Reader for a New Academic Field* (S. Frith and S. Zagorski-Thomas, eds.), Ashgate, 2012.
- [230] H. Massey, *The Great British Recording Studios*. Hal Leonard Corporation, 2015.
- [231] A. C. Disley and D. M. Howard, “Spectral correlates of timbral semantics relating to the pipe organ,” in *Proceedings of the Joint Baltic-Nordic Acoustics Meeting*, June 2004.
- [232] A. Zacharakis, K. Pastiadis, and J. D. Reiss, “An interlanguage study of musical timbre semantic dimensions and their acoustic correlates,” *Music Perception: An Interdisciplinary Journal*, vol. 31, pp. 339–358, April 2014.
- [233] E. Kahle and J.-P. Jullien, “Some new considerations on the subjective impression of reverberance and its correlation with objective criteria,” in *Proceedings of the Sabine Centennial Symposium*, pp. 239–242, June 1994.
- [234] S. Hase, A. Takatsu, S. Sato, H. Sakai, and Y. Ando, “Reverberance of an existing hall in relation to both subsequent reverberation time and SPL,” *Journal of Sound and Vibration*, vol. 232, pp. 149–155, April 2000.
- [235] G. A. Soulodre and J. S. Bradley, “Subjective evaluation of new room acoustic measures,” *Journal of the Acoustical Society of America*, vol. 98, pp. 294–301, July 1995.
- [236] S. Stasis, R. Stables, and J. Hockman, “A model for adaptive reduced-dimensionality equalisation,” in *Proceedings of the 18th International International Conference on Digital Audio Effects (DAFx-15)*, December 2015.
- [237] S. Stasis, J. Hockman, and R. Stables, “Descriptor sub-representations in semantic equalisation,” in *2nd AES Workshop on Intelligent Music Production*, September 2016.
- [238] S. Enderby, T. Wilmering, R. Stables, and G. Fazekas, “A semantic architecture for knowledge representation in the digital audio workstation,” in *2nd AES Workshop on Intelligent Music Production*, September 2016.
- [239] N. Jillings and R. Stables, “JSAP: Intelligent audio plugin format for the Web Audio API,” in *2nd AES Workshop on Intelligent Music Production*, September 2016.

- [240] S. Stasis, R. Stables, and J. Hockman, “Semantically controlled adaptive equalisation in reduced dimensionality parameter space,” *Applied Sciences*, vol. 6, p. 116, April 2016.
- [241] M. Uwins and D. Livesey, “A further investigation of echo thresholds for the optimization of fattening delays,” in *Audio Engineering Society Convention 140*, May 2016.
- [242] T. Wilmering, G. Fazekas, A. Allik, and M. Sandler, “Audio effects data on the semantic web,” in *Audio Engineering Society Convention 139*, October 2015.
- [243] T. Wilmering, G. Fazekas, and M. Sandler, “AUFX-O: Novel methods for the representation of audio processing workflows,” in *15th International Semantic Web Conference (ISWC 2016)*, pp. 229–237, October 2016.
- [244] N. Jillings and R. Stables, “A semantically powered digital audio workstation in the browser,” in *Audio Engineering Society International Conference: Semantic Audio*, June 2017.
- [245] B. De Man, K. McNally, and J. D. Reiss, “Perceptual evaluation and analysis of reverberation in multitrack music production,” *Journal of the Audio Engineering Society*, vol. 65, January/February 2017.
- [246] B. A. Blessler, “An interdisciplinary synthesis of reverberation viewpoints,” *Journal of the Audio Engineering Society*, vol. 49, pp. 867–903, October 2001.
- [247] B. Leonard, R. King, and G. Sikora, “The effect of acoustic environment on reverberation level preference,” in *Audio Engineering Society Convention 133*, October 2012.
- [248] C. Uhle, J. Paulus, and J. Herre, “Predicting the perceived level of late reverberation using computational models of loudness,” in *17th International Conference on Digital Signal Processing (DSP)*, pp. 1–7, July 2011.
- [249] M. R. Schroeder and B. F. Logan, “Colorless artificial reverberation,” *IRE Transactions on Audio*, vol. 9, pp. 209–214, November/December 1961.
- [250] V. Välimäki, J. D. Parker, L. Savioja, J. O. Smith, and J. S. Abel, “Fifty years of artificial reverberation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 1421–1448, July 2012.
- [251] J.-M. Jot and O. Warusfel, “Spat: A spatial processor for musicians and sound engineers,” in *CIARM: International Conference on Acoustics and Musical Research*, May 1995.
- [252] I. Frissen, B. F. G. Katz, and C. Guastavino, “Effect of sound source stimuli on the perception of reverberation in large volumes,” in *Auditory Display: 6th Int. Symposium*, pp. 358–376, May 2010.
- [253] Z. Meng, F. Zhao, and M. He, “The just noticeable difference of noise length and reverberation perception,” in *International Symposium on Communications and Information Technologies (ISCIT’06)*, pp. 418–421, October 2006.
- [254] P. Luizard, B. F. Katz, and C. Guastavino, “Perceived suitability of reverberation in large coupled volume concert halls.,” *Psychomusicology: Music, Mind, and Brain*, vol. 25, pp. 317–325, September 2015.

- [255] A. H. Marshall, D. Gottlob, and H. Alrutz, “Acoustical conditions preferred for ensemble,” *Journal of the Acoustical Society of America*, vol. 64, pp. 1437–1442, November 1978.
- [256] Y. Ando, M. Okura, and K. Yuasa, “On the preferred reverberation time in auditoriums,” *Acta Acustica united with Acustica*, vol. 50, pp. 134–141, February 1982.
- [257] M. Barron, “The subjective effects of first reflections in concert halls – The need for lateral reflections,” *Journal of Sound and Vibration*, vol. 15, pp. 475–494, April 1971.
- [258] M. R. Schroeder, D. Gottlob, and K. F. Siebrasse, “Comparative study of European concert halls: Correlation of subjective preference with geometric and acoustic parameters,” *Journal of the Acoustical Society of America*, vol. 56, pp. 1195–1201, October 1974.
- [259] A. Czyzewski, “A method of artificial reverberation quality testing,” *Journal of the Audio Engineering Society*, vol. 38, pp. 129–141, March 1990.
- [260] D. Lee and D. Cabrera, “Equal reverberance matching of music,” in *Proceedings of Acoustics*, November 2009.
- [261] D. Lee, D. Cabrera, and W. L. Martens, “Equal reverberance matching of running musical stimuli having various reverberation times and SPLs,” in *Proceedings of the 20th International Congress on Acoustics*, August 2010.
- [262] D. Griesinger, “How loud is my reverberation?,” in *Audio Engineering Society Convention 98*, February 1995.
- [263] W. Kuhl, “Über Versuche zur Ermittlung der günstigsten Nachhallzeit großer Musikstudios,” *Acta Acustica united with Acustica*, vol. 4, pp. 618–634, January 1954.
- [264] EBU Tech 3341, “Loudness metering: ‘EBU Mode’ metering to supplement loudness normalisation in accordance with EBU R128,” *European Broadcasting Union*, January 2016.
- [265] D. Ward and J. D. Reiss, “Loudness algorithms for automatic mixing,” in *2nd AES Workshop on Intelligent Music Production*, September 2016.
- [266] P. Zahorik, “Direct-to-reverberant energy ratio sensitivity,” *Journal of the Acoustical Society of America*, vol. 112, pp. 2110–2117, November 2002.
- [267] H. Herlufsen, “Dual channel FFT analysis (part I),” Tech. Rep. 1, Brüel & Kjær Technical Review, 1984.
- [268] ISO 18233:2006, “Application of new measurement methods in building and room acoustics,” *International Organization for Standardization*, 2006.
- [269] D. H. Griesinger, “Quantifying musical acoustics through audibility,” *Journal of the Acoustical Society of America*, vol. 94, pp. 1891–1891, September 1993.
- [270] IEC 61260:1995, “Electroacoustics – Octave-band and fractional-octave-band filters,” *International Electrotechnical Commission*, August 1995.
- [271] ANSI S.1.11-2004, “Specification for octave-band and fractional-octave-band analog and digital filters,” *American National Standards Institute*, February 2004.