KES Transactions on Innovation in Music: Vol 2
Innovation in Music II, Anglia Ruskin University, Cambridge, UK, 7–9 June 2015.

Crowd-sourced learning of music production practices through large-scale perceptual evaluation of mixes

Brecht De Man[1], Joshua D. Reiss[2]

Centre for Digital Music, School of Electronic Engineering and Computer Science,
Queen Mary University of London
[1]b.deman@qmul.ac.uk
[2]joshua.reiss@qmul.ac.uk

Abstract

Mixing music is a highly complex and important part of the music production process, with a variety of creative and technical challenges, few of which have established solutions. Consequently, several approaches are viable for each given recording, and evaluation of differences in music production practices is therefore highly subjective. However, the study of perception of music production processes reveals that there is some degree of consensus on which mixes or specific parameter settings are preferred over others.

In this paper, we give an overview of prior work based on a dataset consisting of songs mixed by at least eight different mixing engineers, with extensive perceptual evaluation in the form of preference ratings and free-form comments. In contrast with most previous work in the area, we investigate realistic mixes as opposed to considering a specific process in isolation, which disregards the cross-adaptive nature of the mixing process. Furthermore, detailed perceptual evaluation of each mix allows to distinguish if the complete song or specific components thereof received a treatment that was perceived as positive or negative. Finally, having access to the original, raw audio as well as the exact parameter settings used on each processor, thorough analysis of the mix is possible.

1. Introduction

1.1. Innovation in mixing music

At the most basic level, the main tools at the disposal of the mix engineer are a multitrack recorder and medium, gain and level controls, pan pots, dynamic range compression, parametric equalisers [1], effect units such as reverberators, and automation of the parameters above. Most mix environments, be it a top tier or bedroom studio, are essentially an arrangement of these elements, each of which have been invented well before 1980 - and most of them much earlier. Aside from more recent developments such as the advent and widespread adoption of digital audio, the architecture of the recording studio and live sound rig have changed surprisingly little in the last three to four decades, while other parts of the music industry have seen major disruptions fuelled by new technologies.

At the same time, the music production process has been tremendously democratised to a point where the contemporary bedroom producer has access to a track count, audio quality and diversity of tools unparalleled by any recording studio in the eighties. Cheap or free digital audio workstation (DAW) software and

inexpensive hardware have brought music production to the masses, but indeed the components of these studios are mostly identical in concept. Despite the immense analytical power of computers, few semantic technologies and intelligent tools have made it into the everyday arsenal of the sound engineer, meaning representation and manipulation of audio still happens at a very low level, through waveforms and filter gains. Recent years have seen a surge in research on automation of audio engineering tasks and even some commercial products with more high-level interface elements (such as a 'space', 'air' or even 'magic' control). However, for any of these technologies to be effective, both the meaning of such descriptors and the mix engineer process in general needs to be better understood.

1.2. Prior work

Earlier work on mixing practices and the perception thereof has heavily contributed to the understanding of the many tasks that constitute mix engineering. Each of these deviate to some extent from what we intend to study, to make it possible to make claims about mixing practices and how they are perceived. For instance, much of these works focus on a single processor, often varying the parameters while keeping other aspects of the mix constant [2-5]. As a result, the potential interdependence of the many variables in a mix is ignored and statements about how parameters should be set are not necessarily valid when other parameters and features change. Similarly, testing a hypothesis on a (very) limited amount of musical content limits the transferability of findings to other situations [2, 6]. In general, acquiring data in a lab environment wherein the typical workflow of a mix engineer is not preserved, it is possible for the mix process to differ from a real-life, commercially relevant scenario. Finally, some algorithms, hypotheses or experimentally obtained values are not tested through subjective evaluation [7-10], meaning poor mixes or less than ideal parameter settings are potentially skewing the results.

1.3. This work

Our goal is to understand mixing, to define descriptive terms frequently used to describe sound, and to help develop tools for the analysis and manipulation of sound at a higher, more abstract level than low-level features (e.g. level) and parameters (e.g. filter coefficients) allow us to do. In this work, we aim to provide an overview of our previous work in which collection and analysis of realistic mixes and evaluations thereof has allowed us to answer some of these questions, as well as to highlight some of the many questions that remain unanswered, and hypothesise how we may do so. By realistic mixes, we mean mixes that were produced using a commercially relevant set of tools, by someone who is used to using these, in a natural environment such as their home studio, a professional studio or institution's facilities.

We propose a methodology to derive the bounds of what are considered 'preferred' or 'acceptable' ranges of values of processing parameters and mix properties, from subjective evaluation of different mixes of identical and/or different songs. As such, we systematically learn from mixes of arbitrary quality to zone in on production practices that are perceived as good, obviating the need for exemplary productions

to study the mixing process. On the contrary, a limited spread of parameter settings and mix features would impede a sufficiently accurate estimation of the lower and upper bounds of acceptable parameter ranges. As we learn how the bounds of these processing parameters or extracted audio features vary between different songs or genres, we identify zones wherein these values should lie in a majority of productions according to general perception.

In Section 2, we describe the process of collecting data for this research, from acquiring raw audio to be mixed, over producing mixes of this content, to the perceptual evaluation of said mixes. Analysis of this data is discussed in Section 3, where we explore the questions that the various types of data we collect can and - more importantly - cannot answer. Concluding remarks and a brief outline of future work are presented in Section 4. Finally, the reader is invited to use the resulting data for their own research, as well as to contribute to our dataset in Section 5.

2. Data collection

2.1. Multitrack audio content

A first challenge we encountered when attempting to collect and analyse realistic mixes is the acute lack of available raw tracks, specifically those that have an open license that allows sharing the data for the purpose of sustainability and reproducibility. For this reason, we created the Open Multitrack Testbed[1] [11], a platform for hosting multitrack audio, including stems and mixes thereof. Rich metadata allows for searching and filtering of the content (e.g. 'multitrack containing bagpipes', 'track recorded with a Shure SM58 microphone', ...) that can be hosted locally or at the original website, depending on permissions. While several projects have already made use of the data collected through and referenced from this platform, more contributions are still appreciated to ensure a large and diverse pool of multitrack audio.

2.2. Mixes

Having collected raw audio tracks of songs from a variety of genres and sources, we then had sound engineering students mix these in a specified digital audio workstation (DAW) using a restricted list of plugins. As such, we were able to analyse audio features and parameters from individual tracks and processors, allowing for extensive dissection of every mix.

Since then, engineers from sound engineering programs at schools in different countries have been added to the list, and we are still in the process of repeating these experiments with new songs, different engineers, and more subjects. As such, we are able to average out or indeed investigate the influence of location and background

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

2.3. Perceptual evaluation

For reasons explained in Section 3, trained listeners provided ratings and comments of the different mixes per song that we collected.

We devised a listening test tool (MATLAB[2]- and browser-based[3] [12, 13]) for assessing these different mixes relative to each other, such that they are to be rated on a single axis, and with the possibility to write comments on each mix. The goal of providing text boxes for comments is

- to facilitate the tedious task of comparing as many as ten different mixes against each other by allowing the subject to take notes;
- to provide feedback for the respective mix engineers; and
- to give us more insight into the subjects' perception of the different mix decisions.

3. Analysis

In this section, we present an overview of prior and future work based on the data presented in Section 2. We show (or hypothesise) what we can learn from analysing the mix audio [14], subjective ratings [15], and ultimately comments [16], and how the combination of each of these provides more insight into the mix process than the individual elements.

3.1. Audio and settings

Having access to not only the rendered mixes, but also the raw tracks and the DAW session files, it is possible to extract features from separate, processed tracks and look at the parameters of the different processors. As such, we are able to conduct a much deeper analysis of these mixes than what is possible with stereo audio files only [17]. In [14], we looked at a set of features extracted from lead vocal, bass, and various drum tracks (important tracks that were present in each of the considered songs) from mixes of eight different songs. For instance, Figure 1 shows the average loudness, relative to the total mix loudness, of the lead vocal, bass, kick drum, snare drum and rest of the drums. These results were since confirmed by [18] and correspond with [19].

[2] code.soundsoftware.ac.uk/projects/ape
[3] code.soundsoftware.ac.uk/projects/webaudioevaluationtool

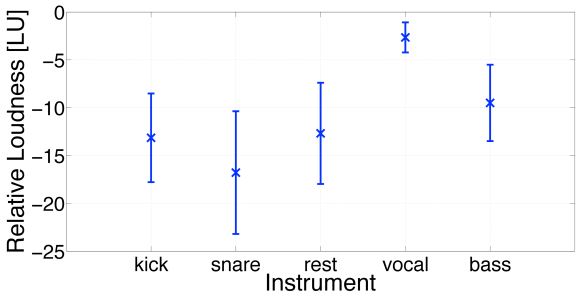

Figure 1: Average loudness of kick drum, snare drum, rest of drums, lead vocal and bass, relative to the total mix loudness (0 LU) - from [14].

However, analysing the audio or settings of the different collected mixes in isolation means that poor, unconventional or otherwise irrelevant mixes may add noise or skew our findings towards feature and parameter values that are less pleasing or typical of commercial mixes.

3.2. Ratings

Through perceptual evaluation in the form of ranking and/or rating of different mixes of the same song, it is possible to weed out poor mixes and investigate only mix settings and audio features of good ones, which are supposedly more similar to commercial mixes and more relevant to furthering our understanding of mix engineering. It even allows us to look at differences between highly and poorly rated mixes, and to some extent to answer the question of what makes a good mix.

Looking at ratings only, we have learned that mix engineers who also take part in the blind subjective evaluation of the different mixes tend to favour their own mix over others - see Figure 2 [15]. This suggests the mix engineer has a distinct preference for a certain mixing style, both when producing or assessing content, but it can also be influenced by the possible bias due to having mixed the song a couple of weeks earlier, or even downright recognising the mix as their own.

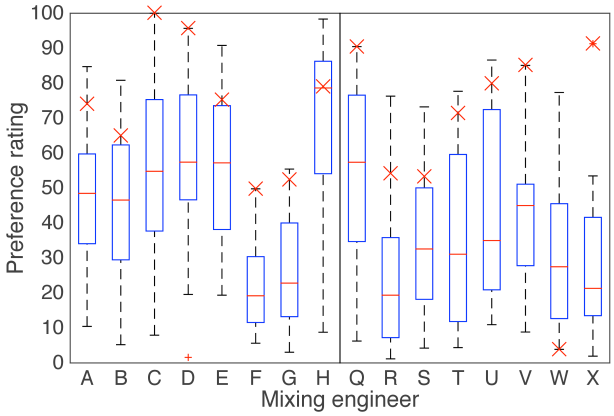

Figure 2: Ratings of own mix (red 'X') versus ratings by all participants of this mix (box plot) - from [15].

When combining the rating data with the extracted audio features [15] or workflow aspects [20], some correlations with preference ratings can be found that indicate a preference towards a higher dynamic range, a stronger central component, or a higher degree of grouping similar tracks together. However, to reliably infer which mix decisions are favoured over others, we would need to look at a very large and diverse set of evaluated mixes, and/or use specialised, complex, and perhaps perceptually motivated features. We do not know which specific features of which tracks to look at to understand which mixes or practices are perceived to be better, and the task of investigating the correlation of preference with every imaginable feature of every audio track is not only enormous, but comes with a near-certainty of overfitting the problem. After all, the chance of any of a million features (or a combination thereof) spuriously correlating highly with a limited number of preference scores is very high.

3.3. Comments

In [16], we investigated the comments assigned to the different mixes in the perceptual evaluation part of the experiment, and found the representation of different instruments and processors or features shown in Figure 3.

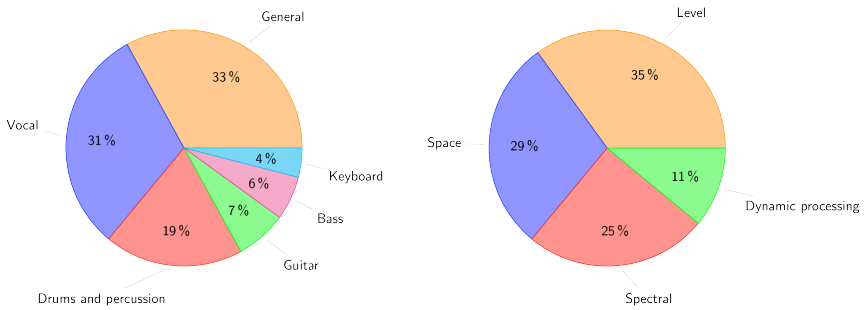

Figure 3: Representation of instruments and types of processing in statements about mixes - from [16].

At a lower level, we can look at which specific words occur most often, such as shown in Table 1. Looking exclusively at terms used to describe characteristics of sounds (‘dry’, ‘bright’, ‘thin’, ...), we learn how the processing on certain instruments in certain mixes is perceived - and, conversely, we may learn what these subjective words mean exactly by looking at the features of the sounds they were meant to describe - see Table 2.

Table 1: Top 25 most frequently occurring words over all mix comments.

Term	**#**	**%**
vocal(s,ist)+vox+voc(s)+voice +singer	1082	7.87%
rev(erb)(s,y)+reverber(ant,ated) +verb(y)	412	3.00%
good	350	2.54%
mix	328	2.38%
balance+balanced+balances +balancing	327	2.38%
drum(s)	345	2.51%
loud	264	1.92%
bass	258	1.88%
nice	246	1.79%
low	171	1.24%
like(d)	168	1.22%
snare	164	1.19%
guitar	160	1.16%
kick	150	1.09%
compress(ed,ing,ion,or)	141	1.03%
dry(ness)	127	0.92%
lead	116	0.84%
bright(er), brightness	107	0.78%
thin(ner,ness)	94	0.68%
weird(ly,ness,-sounding)	87	0.63%
chorus(es)	86	0.63%
instruments	84	0.61%
dark	79	0.57%
eq+eq'd+eq'ed+eqd+eqed+eqs	79	0.57%
well	78	0.57%

Table 2: Top 25 most frequently occurring descriptive terms over all mix comments.

Term	#	%
dry(ness)	127	0.9233%
bright(er), brightness	107	0.7779%
thin(ner,ness)	94	0.6834%
weird(ly,ness,-sounding)	87	0.6325%
dark	79	0.5743%
far	52	0.3780%
soft	52	0.3780%
muddy+mud+muddiness	50	0.3635%
harsh(ness)	47	0.3417%
room(-y,y)	47	0.3417%
punch(y,ier,iness)	46	0.3344%
quiet(er)	43	0.3126%
wide	41	0.2981%
big+bigger	37	0.2690%
hot	37	0.2690%
flat	32	0.2326%
big	31	0.2254%
mono(ish,-ish)	30	0.2181%
definition+defined	28	0.2036%
present	28	0.2036%
presence	26	0.1890%
narrow	24	0.1745%
small	23	0.1212%
weak	23	0.1212%
forward	21	0.1107%

For instance, ‘bright’ (and derivatives) and ‘dark’ occur 186 times combined over ca. 1400 mix ‘reviews’ - indicating these are popular terms to describe sound and in particular to note why a mix or the processing of a particular instrument is (dis)liked. Investigating what this term means may therefore be instrumental towards gaining understanding of sound, in particular in a music production context. Looking at the spectrum of the mix, or the instrument in relation to which it is used, we can then reveal e.g.

- what the average spectrum of a ‘bright’ and ‘dark’ sound looks like, as well as of sounds which elicit neither of these responses;
- if they are really used as each other’s opposite (on average, or for the same person using the term);
- if they are indeed referring to the spectrum only, or if other features correlate with the description.

As an example, Figure 4 shows the objective features Brightness and Spectral Centroid of each of the mixes of ten different songs. Mixes which over 50% of the subjects found to be 'too bright', as understood from the comments, are above the upper dashed line, while mixes which they found 'too dark' are below the lower dashed line. This shows that with a large enough number of mixes and a large enough variance of the considered parameter or feature, it is possible to find the upper and lower bounds of its range of values deemed acceptable by the subjects - as long as there is enough consensus. As such, we can develop a 'mix space' of various parameters or features within which one can move while still maintaining a good sounding mix, and outside of which some mix decisions would be considered bold at best, and possibly poor. One of many challenges lies in taking the interdependence of some of these features into account.

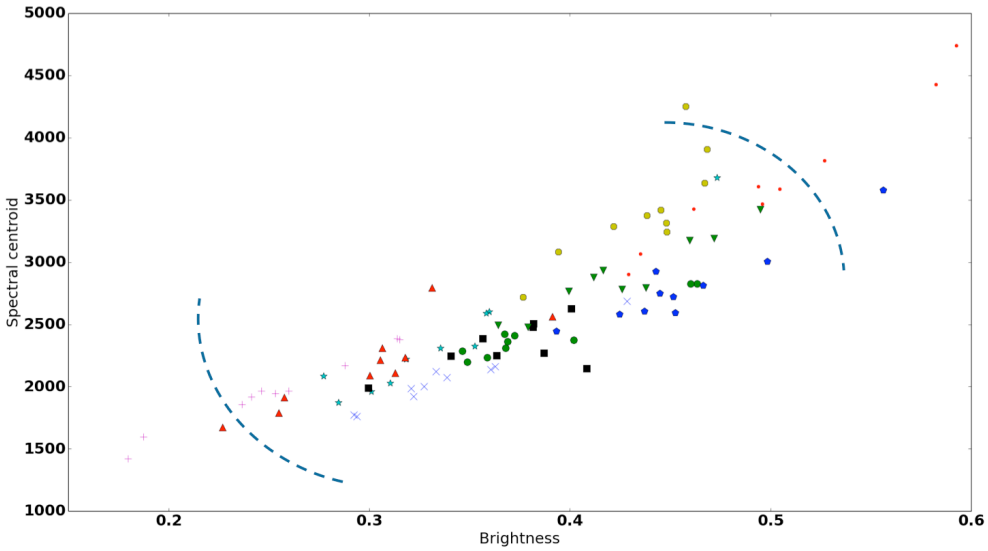

Figure 4: The evaluated mix fragments (each marker represents a different song) plotted as a function of the objective features Brightness (x-axis) and Spectral Centroid (y-axis). The dashed lines separate mixes which are labelled as too bright or too dark, respectively, by at least 50% of the subjects.

Looking at the comments that were written as part of this perceptual evaluation reveals which aspects of the mix prompted the subject to rate it highly or poorly, or in other words what problems and strengths the various mixes have. By combining assessments of different subjects, and preserving the salient comments, we can zone in on particular aspects of the mix that we assume to be true because of this consensus, and investigate them in isolation. This has the advantage that we don't need a mix to be 'good' in every way to learn what settings or features are perceived to be appropriate, but that we can learn even - or especially? - from poor mixes, where positive aspects ('punchy snare') denote appropriate settings but where negative features ('vocal too loud') can help find upper and lower bounds of the corresponding parameters. Taking preference ratings into account further helps understand which comments are likely positive or negative in nature, when it is not clear from the comment itself or its context.

Still, a thorough understanding of mix practices and their perception for arbitrary songs in different genres requires a very large dataset with a high number of highly diverse songs, mixed and evaluated by people from various backgrounds and locations.

4. Concluding remarks

In this paper we have provided an overview of recent work on mix practices and the perception thereof, based on a dataset consisting of realistic mixes of high quality multitrack audio, and subjective evaluation of these mixes. We have shown that the combination of process parameters, audio features extracted from the mixes and individual processed tracks, subjective ratings, and/or comments on the mixes is more effective than each of those separately when deriving information from this data, and explored how it may be used for further research. This data is mostly available (insofar the license allows it) for other researchers to use. Furthermore, contributions to this data of any kind are also appreciated - see Section 5.

At this point, the number and diversity of songs, engineers and subjects limits the generality of findings inferred from this data. Therefore, we are in the process of repeating these experiments at various institutions, with new raw audio. This allows us to assess whether the discovered (and not yet discovered) trends in mix practices and perception of music production apply to engineers and listeners with different backgrounds and levels of experience, or, if not, what influence these factors have.

5. Call for contributions

The Open Multitrack Testbed [11] welcomes any kind of multitrack audio, with or without mixes, which is already available for download on the internet or which has a license that allows for us to host it. As most licenses stipulate, content owners (as well as contributors) are acknowledged on the website[4].

Listening test tools developed for this type of research[5] [12, 13] are freely available including source code, and feedback as well as contributions are highly appreciated.

Furthermore, we are encouraging sound engineers and sound engineering students of various backgrounds, levels and locations to participate in further mix experiments and/or listening tests, to contribute (anonymously) to the

[4] `multitrack.eecs.qmul.ac.uk`

[5] APE Audio Perceptual Evaluation toolbox (MATLAB): `code.soundsoftware.ac.uk/projects/ape` and Web Audio Evaluation Tool (browser-based): `code.soundsoftware.ac.uk/projects/webaudioevaluationtool`

aforementioned research as well as to be part of a unique critical listening exercise and to receive a wealth of feedback from a diverse and unbiased audience.

Further info and data is available at `www.brechtdeman.com/research.html`.

6. References

[1] G. Massenburg, "Parametric equalization," in 42nd Convention of the Audio Engineering Society, May 1972.
[2] S.-I. Mimilakis, K. Drossos, A. Floros, and D. Katerelos, "Automated tonal balance enhancement for audio mastering applications," in 134th Convention of the Audio Engineering Society, Audio Engineering Society, 2013.
[3] S. Hafezi and J. D. Reiss, "Autonomous multitrack equalisation based on masking reduction," to appear in Journal of the Audio Engineering Society, 2015.
[4] 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.
[5] E. Perez Gonzalez and J. D. Reiss, "A real-time semiautonomous audio panning system for music mixing," EURASIP Journal on Advances in Signal Processing, 2010.
[6] 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.
[7] J. Scott, M. Prockup, E. Schmidt, and Y. Kim, "Automatic multi-track mixing using linear dynamical systems," in Proceedings of the 8th Sound and Music Computing Conference, Padova, Italy, 2011.
[8] 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.
[9] E. Perez-Gonzalez and J. D. Reiss, "Automatic mixing: Live downmixing stereo panner," in 10th International Conference on Digital Audio Effects (DAFx-10), 2007.
[10] E. Perez-Gonzalez and J. D. Reiss, "Automatic equalization of multi-channel audio using cross-adaptive methods," 127th Convention of the Audio Engineering Society, October 2009.
[11] B. De Man, M. Mora-Mcginity, G. Fazekas, and J. D. Reiss, "The Open Multitrack Testbed," in 137th Convention of the Audio Engineering Society, October 2014.
[12] B. De Man and J. D. Reiss, "APE: Audio Perceptual Evaluation toolbox for MATLAB," in 136th Convention of the Audio Engineering Society, April 2014.
[13] 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.
[14] 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.
[15] B. De Man, M. Boerum, B. Leonard, G. Massenburg, R. King, and J. D. Reiss, "Perceptual evaluation of music mixing practices," in 138th Convention of the Audio Engineering Society, May 2015.

[16]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.
[17]P. D. Pestana, Z. Ma, J. D. Reiss, A. Barbosa, and D. A. A. Black, “Spectral characteristics of popular commercial recordings 1950-2010,” in 135th Convention of the Audio Engineering Society, October 2013.
[18]A. Wilson and B. M. Fazenda, “Navigating the mix-space: Theoretical and practicallevel-balancing technique in multitrack music mixtures,” in 12th Sound and Music Computing Conference, July 2015.
[19]P. Pestana and J. D. Reiss, “Intelligent audio production strategies informed by best practices,” in 53rd Conference of the Audio Engineering Society, January 2014.
[20]D. M. Ronan, B. De Man, H. Gunes, and J. D. Reiss, “The impact of subgrouping practices on the perception of multitrack mixes,” in 139th Convention of the Audio Engineering Society, October 2015.