THE MIX EVALUATION DATASET

Brecht De Man *Joshua D. Reiss*

Centre for Digital Music,
School of Electronic Engineering and Computer Science,
Queen Mary University of London
London, UK
{b.deman,joshua.reiss}@qmul.ac.uk

ABSTRACT

Research on perception of music production practices is mainly concerned with the emulation of sound engineering tasks through lab-based experiments and custom software, sometimes with unskilled subjects. This can improve the level of control, but the validity, transferability, and relevance of the results may suffer from this artificial context. This paper presents a dataset consisting of mixes gathered in a real-life, ecologically valid setting, and perceptual evaluation thereof, which can be used to expand knowledge on the mixing process. With 180 mixes including parameter settings, close to 5000 preference ratings and free-form descriptions, and a diverse range of contributors from five different countries, the data offers many opportunities for music production analysis, some of which are explored here. In particular, more experienced subjects were found to be more negative and more specific in their assessments of mixes, and to increasingly agree with each other.

1. INTRODUCTION

Many types of audio and music research rely on multitrack audio for analysis, training and testing of models, or demonstration of algorithms. For instance, music production analysis [1], automatic mixing [2], audio effect interface design [3], instrument grouping [4], and various types of music information retrieval [5] all require or could benefit from a large number of raw tracks, mixes, and processing parameters. This kind of data is also useful for budding mix engineers, audio educators, and developers, as well as creative professionals in need of accompanying music or other audio where some tracks can be disabled [6].

Despite this, multitrack audio is scarce. Existing online resources of multitrack audio content typically have a relatively low number of songs, offer little variation, are restricted due to copyright, provide little to no metadata, or lack mixed versions and corresponding parameter settings. An important obstacle to the widespread availability of multitrack audio and mixes is copyright, which restricts the free sharing of most music and their components. Furthermore, due to reluctance to expose the unpolished material, content owners are unlikely to share source content, parameter settings, or alternative versions of their music. 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 therefore suffers from a severe lack of relevant material.

This impedes reproduction or improvement of previous studies where the data cannot be made public, and comparison of different works when there is no common dataset used across a wider community. It further limits the generality, relevance, and quality of the research and the designed systems. 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.

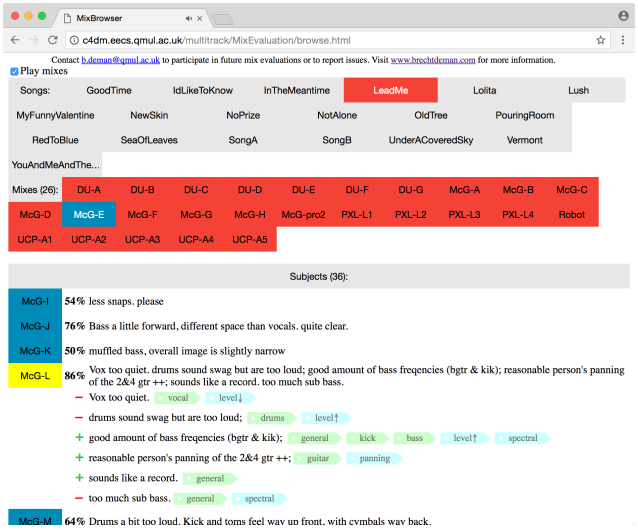


Figure 1: Online interface to browse the contents of the Mix Evaluation Dataset

The dataset presented here is therefore based on a series of controlled experiments wherein realistic, ecologically valid mixes are produced — i.e. by experienced engineers, in their preferred environment and using professional tools — and evaluated. The sessions can be recreated so that any feature or 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. While some of the data presented here has been used in several previous studies, the dataset is now consolidated and opened up to the community, and can be browsed on `c4dm.eecs.qmul.ac.uk/multitrack/MixEvaluation/`, see Figure 1.

With close to 5000 mix evaluations, the dataset is by far the largest study of evaluated mixes known to the authors. MedleyDB, another resource shared with researchers on request, consists of raw tracks including pitch annotations, instrument activations, and metadata [7]. [8] analyses audio features extracted from a total of 1501 unevaluated mixes from 10 different songs. The same au-

thors examine audio features extracted from 101 mixes of the same song, evaluated by one person who classified the mixes in five preference categories [9]. In both cases, the mixes were created by anonymous visitors of the Mixing Secrets Free Multitrack Download Library [10], 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 in these works.

This paper introduces the dataset and shows how it allows to further our understanding of sound engineering, and is structured as follows. Section 2 presents the series of acquisition experiments wherein mixes were created and evaluated. The resulting data is described in Section 3. Section 4 then demonstrates how this content can be used to efficiently obtain knowledge about music production practices, perception, and preferences. To this end, previous studies and key results based on part of this dataset are listed, and new findings about the influence of subject expertise based on the complete set are presented. Finally, Section 5 offers concluding remarks and suggestions for future research.

2. METHODOLOGY

2.1. Mix creation

Mix experiments and listening tests were conducted at seven institutions located in five different countries. The mix process was maximally preserved in the interest of ecological relevance, while information such as parameter settings was logged as much as possible. Perceptual evaluation further helped validate the content and investigate the perception and preference in relation to mix practices.

Students and staff members from sound engineering degrees at McGill University (McG), Dalarna University (DU), PXL University College (PXL), and Universidade Católica Portuguesa (UCP) created mixes and participated in listening tests. In addition, employees from a music production startup (MG) and researchers from Queen Mary University of London (QM) and students from the institution's Sound and Music Computing master (SMC) took part in the perceptual evaluation stage as well.

Table 1 lists the songs and the corresponding number of mixes created from the source material, as well as the number of subjects evaluating (a number of) these mixes. Numbers between parentheses refer to additional mixes for which stems, Digital Audio Workstation (DAW) sessions, and parameter settings are not available. These correspond to the original release or analogue mixes, see Section 3.2. Songs with an asterisk (*) are copyrighted and not available online, whereas raw tracks to others can be found via the Open Multitrack Testbed[1] [11]. For two songs, permission to disclose artist and song title was not granted. Evaluations with an obelus (†) indicate that subjects included those who produced the mixes. Consistent anonymous identifiers of the participants (e.g. 'McG-A') allow exclusion of this segment or examination of the associated biases [12].

The participants produced these 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 toolset was restricted somewhat so that each mix could be faithfully recalled and analysed in depth later, with a limited number of software plugins available, typically consisting of those which come with the respective DAWs. All students used Avid Pro Tools 10, an industry standard DAW, except for the PXL group who used Apple Logic Pro X. Instructions explicitly forbade outboard processing, recording new audio, sample replacement, pitch and timing correction, rearranging sections, or manipulating audio in an external editor. Beyond this, any kind of processing was allowed, including automation, subgrouping, and multing.

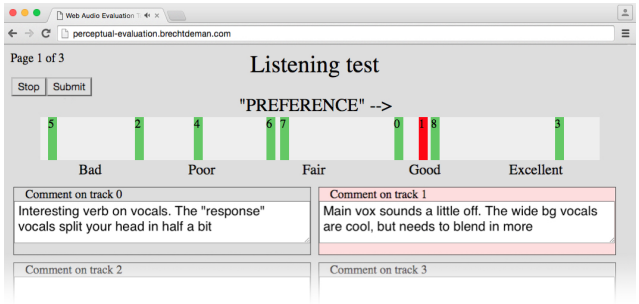


Figure 2: Example interface used to acquire subjective assessments of nine different mixes of the same song, created with the Web Audio Evaluation Tool

2.2. Perceptual evaluation

The different mixes were evaluated in a listening test using the interface presented in [13]. With the exception of groups McG and MG, the browser-based version of this interface from the Web Audio Evaluation Tool [14] was used, see Figure 2.

As dictated by common practices, this listening test was conducted in a double blind fashion [15], with randomised presentation order [16], minimal visual information [17], and free and immediate switching between time-aligned stimuli [18]. The interface presented multiple stimuli [19] on a single, 'drag-and-drop' rating axis [20], and without ticks to avoid build-up around these marks [21]. A 'reference' was not provided because it is not defined for this exceedingly subjective task. Indeed, even commercial mixes by renowned mix engineers prove not to be appropriate reference stimuli, as these are not necessarily rated more highly than mixes by students [12].

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 [1].

The headphones used were Beyerdynamic DT770 PRO for group MG and Audio Technica M50x for group SMC. In all other cases, the listening tests took place in dedicated, high quality listening rooms at the respective institutions, the room impulse responses of which are included in the dataset. This knowledge could be used to estimate the impact of the respective playback systems, although in these cases the groups differ significantly in other aspects as well.

Comments in other languages than English (DU, PXL, and UCP) were translated by native speakers of the respective languages, who are also proficient in English and have a good knowledge of audio engineering.

Table 1: Overview of mixed content, with number of mixes (left side) and number of subjects (right side) per song

ARTIST – SONG	GENRE	NUMBER OF MIXES				NUMBER OF SUBJECTS						
		McG	**DU**	**PXL**	**UCP**	**McG**	**MG**	**QM**	**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			
Artist Y – Song B*	blues	8 (2)				14		10	5			
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			
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

Table 2 shows additional details of the perceptual evaluation experiments.

3. CONTENT

3.1. Raw tracks

Raw tracks of the mixes can be found via the Open Multitrack Testbed[1]. The first two sections of Table 1 are newly presented here and were recorded by Grammy-winning engineers. Six of the ten songs are made available in their entirety under a Creative Commons BY 4.0 license. Raw tracks to songs from the last two sections of the table can be downloaded from Weathervane Music's Shaking Through[2] and Mike Senior's Mixing Secrets Multitrack Library [10], respectively. Many more mixes of these tracks are available on the forums of these websites, albeit without associated parameter settings or evaluations.

3.2. Mixes and stems

All stereo mixes are available in uncompressed, high resolution WAV format. Unique to this dataset is the availability of DAW session files, which includes all parameter setting of 'in-the-box' mixes. Where the mix and its constituent elements could be recreated, stems of the vocal, kick drum, snare drum, rest of the drums, and bass instrument are rendered. Similarly, the sum of all reverberation signals ('wet') and the rest of a mix ('dry'), as in [22], are shared as well.

The dataset also contains mixes which were produced mostly through analogue processing. While this makes detailed analysis more difficult, it increases the diversity and allows a wider range of possible research questions the data could answer. To mitigate this relative lack of control, approximate parameter settings can be derived from recall sheets, pictures of the devices, the parsed recall files from the SSL AWS900 console (DU), and a recording of a fragment of each channel as the engineer sequentially soloes each track (PXL).

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

[2] `weathervanemusic.org/shakingthrough`

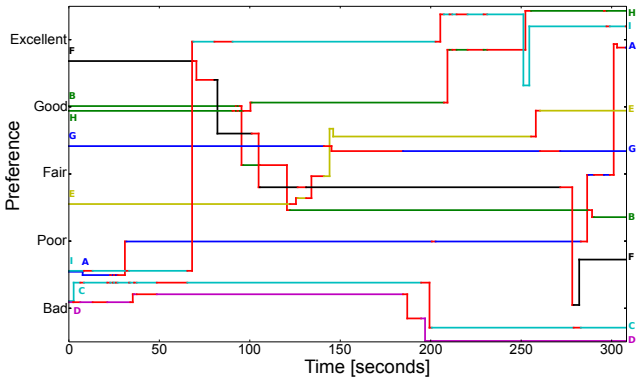


Figure 3: Built-in timeline visualisation of the Web Audio Evaluation Tool, showing playback (red) and movement of sliders of a single subject rating nine mixes of a single song

3.3. Preference ratings

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' [23] — is a more sensory judgement [24]. Indeed, preference and perceived quality are not always concurrent [25]: 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 [26]. Personal preference was therefore deemed a more appropriate attribute than audio quality or fidelity. Such a 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. Where the Web Audio Evaluation Tool was used, the positions of the sliders over time was registered as well, see Figure 3.

Table 2: Overview of evaluation experiments

	McG	**MG**	**QM**	**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**
Loudspeakers/headphones	LS	HP	LS	HP	LS	LS	LS	

3.4. Comments

A single numerical rating does not convey any detailed information about what aspects of a mix are (dis)liked. For instance, a higher score for mixes with a higher dynamic range [12] may relate to subtle use of dynamic range compression (e.g. preference for substantial level variations), but also to a relatively loud transient source (e.g. preference for prominent snare drums). In addition, the probability of spurious correlation increases as an ever larger number of features is considered. Furthermore, subjects tend to be frustrated when they do not have the ability to express their thoughts on a particular attribute, and isolated ratings do not provide any information about the difficulty, focus, or thought process associated with the evaluation task.

For this reason, free-form text response in the form of comment boxes is accommodated, facilitating in-depth analysis of the perception and preference with regard to various music production aspects. The results of this 'free-choice profiling' also allow learning how subjects used and misused the interface. An additional, practical reason for allowing subjects to write comments is that taking notes on shortcomings or strengths of the different mixes helps them to keep track of which fragment is which, facilitating the complex task at hand.

Extensive annotation of the comments is included in the form of tags associated with each atomic 'statement' of which the comment consists. For instance, a comment *'Drums a little distant. Vox a little hot. Lower midrange feels a little hollow, otherwise pretty good.*' comprises four separate statements. Tags then indicate the instrument ('drums', 'vocal', ...), feature ('level (high)', 'spectrum', ...), and valence ('negative'/'positive').

The XML structure of the Web Audio Evaluation Tool output files was adopted to share preference ratings, comments, and annotation data associated with the content.

3.5. Metadata

3.5.1. Track labels

Reliable track 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 this dataset 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.

The available raw tracks and mixes are annotated on the Open Multitrack Testbed[1], including metadata describing for instance the respective instruments, microphones, and take numbers. This metadata further allows tracks and mixes to be found through the Testbed's search and browse interfaces.

3.5.2. Genre

The source material was selected in coordination with the programme's teachers from the participating institutions, because they fit the educational goals, were considered ecologically valid and homogeneous with regard to production quality, and were deemed to represent an adequate spread of genre. Due 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, providing a post hoc confirmation of the musical diversity. Each song's most often occurring genre label was added to Table 1 for reference.

3.6. Survey responses and subject demographics

The listening test included a survey 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.

4. ANALYSIS

4.1. Prior work

The McG portion of this dataset has previously been used in studies on mix practices and perception, as detailed below.

The mild constraints on tools used and the availability of parameter settings allows one to compare signal features between different mixes, songs, or institutions, and identify trends. A detailed analysis of tendencies in a wide range of audio features — extracted from vocal, drum (kick drum, snare drum, and other), bass, and mix stems — appeared in [27]. As an example, Figure 4 shows the ITU-R BS.1770 loudness [28] of several processed stems for two songs, as mixed by engineers from two institutions (McG and UCP). No significant differences in balance choices are apparent here.

Correlation between preference ratings and audio features extracted from the total mixes have shown a higher preference for mixes with relative higher dynamic range, and mixes with a relatively strong phantom centre [12].

In [29], relative attention to each of these categories was quantified based on annotated comments. Figure 5 shows the relative proportion of statements referring to detailed feature categories for the complete dataset (all groups).

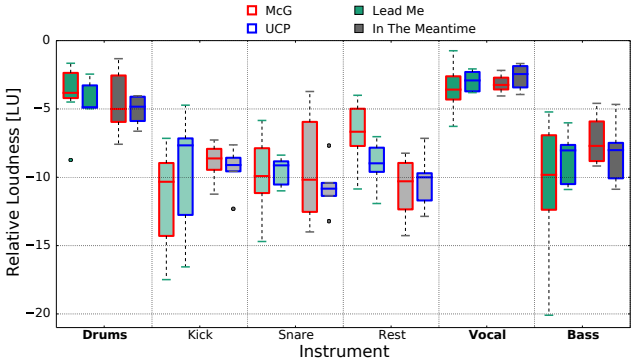


Figure 4: Stem loudness relative to the total mix of Fredy V's In The Meantime and The DoneFors' Lead Me, as mixed by students from the McG and UCP groups

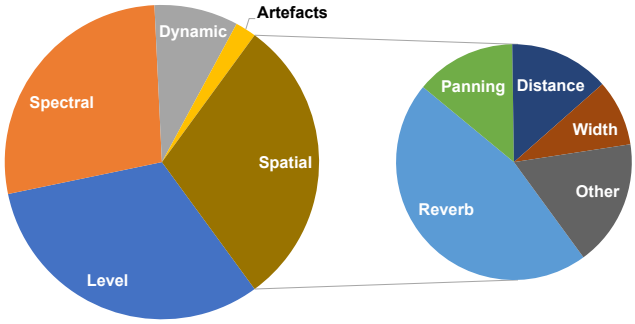


Figure 5: Relative proportion of statements describing mix aspects

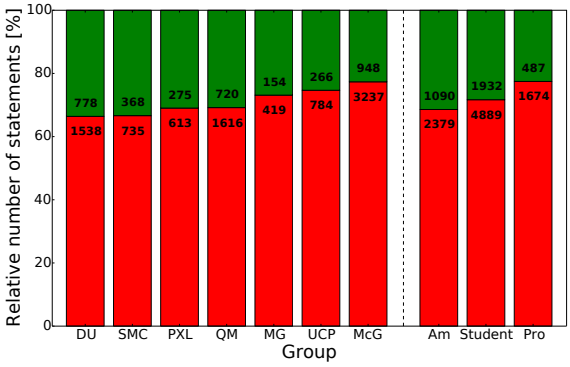
Finally, through combination of the comment annotations with preference ratings and extracted audio features, more focused research questions about music production can be answered. Proving this concept, [22] showed a notably lower preference rating for mixes tagged as overly reverberant than for those which have an alleged lack of reverberant energy, and determined that the optimal reverb loudness relative to the total mix loudness is close to −14 LU.

In addition to being able to render the entire mix or any part thereof, availability of DAW session files also presents a unique opportunity to study workflow and signal routing practices from working mix engineers in a realistic setting. As an example, the process of subgrouping has been studied in [30], where a strong correlation was shown between the number of raw tracks used and the number of subgroups that was created, as well as a medium correlation between the number of subgroups which were processed by EQ and the average preference rating for that mix.

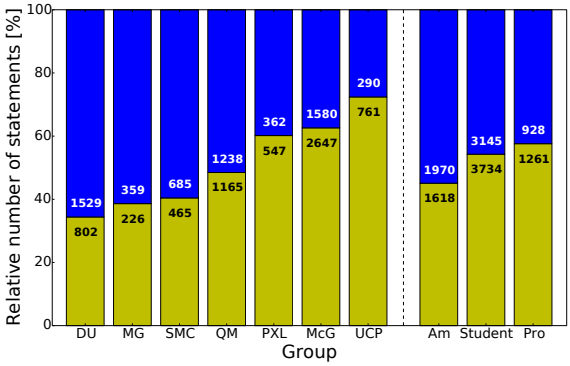
4.2. Effects of subject background

Access to the subject's level of experience, institution, and demographics makes it possible to determine the influence of these factors on subjective preference and perception.

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 [1].



(a) Proportion of negative (red) vs. positive (green) statements



(b) Proportion of instrument-specific (yellow) vs. general (blue) statements

Figure 6: Statement categories as a function of subject data

The proportion of negative statements among the comments is strongly influenced by the level of expertise of the subject as well: there is a significant tendency to criticise more, proportionally, with increasing experience, see Figure 6a. Independent of level of expertise, the proportion of negative statements is also significantly different per group.

Likewise, it is clear that amateurs tend to give more 'general' comments, not pertaining to any particular instrument, as shown in Figure 6b. This accounts for 55% of their statements. For students and professionals this proportion is 46% and 42%, respectively. 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. As these two groups consisted of bachelor students only, the level of expertise is presumably similar and other factors must be at play.

Finally, the agreement within as well as between the groups is quantified, showing the relative number of statements which are consistent with 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 'McG-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

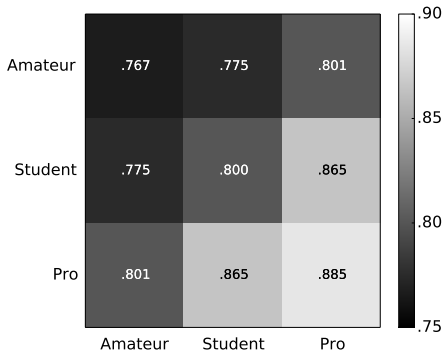


Figure 7: Level of agreement between groups of different expertise

$$r_{AB} = \frac{a_{AB}}{d_{AB} + a_{AB}} \tag{1}$$

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 7. In other words, skilled engineers are less likely to contradict each other when evaluating mixes (high within-group agreement), converging on a 'common view'. This supports the notion that universal 'best practices' with regard to mix engineering do exist, and that perceptual evaluation is more reliable and efficient when the participants are skilled. 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. 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.

Further analysis is necessary, for instance to establish which of these differences are significant and not spurious or under influence of other factors.

5. CONCLUSION

A dataset of mixes of multitrack music and their evaluations was constructed and released, based on contributions from five different countries. The mixes were created by skilled sound engineers, in a realistic setting, and using professional tools, to minimally disrupt the natural mixing process. By including parameter settings and limiting the set of software tools, mixes can be recreated and analysed in detail. Previous studies using this data were listed, and further statistical analysis of the data was presented.

In particular, it was shown that expert listeners are more likely to contribute negative and specific assessments, and to agree with others about various aspects of the mix. This is consistent with the expectation that they are trained to spot and articulate problems with a mix. Conversely, one could suppose amateur subjects 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.

The dataset and potential extensions offer interesting opportunities for further cross-analysis, comparing the practices, perception, and preferences of different groups. At this point, however, the dataset is heavily skewed towards Western musical genres, engineers, and subjects, and experienced music producers. Extension of the acquisition experiments presented here, with an emphasis on content from countries outside of North America and Western Europe, can mitigate this bias and help answer new research questions. In addition, a substantially larger dataset can be useful for analysis which requires high volumes of data, such as machine learning of music production practices [31].

6. ACKNOWLEDGEMENTS

The authors would like to thank Frank Duchêne, Pedro Pestana, Henrik Karlsson, Johan Nordin, Mathieu Barthet, Brett Leonard, Matthew Boerum, Richard King, and George Massenburg, for facilitating and contributing to the experiments. Special thanks also go to all who participated in the mix creation sessions or listening tests.

7. REFERENCES

[1] Alex Wilson and Bruno M. Fazenda, "Perception of audio quality in productions of popular music," *Journal of the Audio Engineering Society*, vol. 64, no. 1/2, pp. 23–34, Jan/Feb 2016.

[2] Enrique Perez Gonzalez and Joshua D. Reiss, "Automatic mixing: Live downmixing stereo panner," *Proc. Digital Audio Effects (DAFx-07)*, Sep 2007.

[3] Spyridon Stasis, Ryan Stables, and Jason Hockman, "A model for adaptive reduced-dimensionality equalisation," *Proc. Digital Audio Effects (DAFx-15)*, Dec 2015.

[4] David Ronan et al., "Automatic subgrouping of multitrack audio," *Proc. Digital Audio Effects (DAFx-15)*, Nov 2015.

[5] Justin Salamon, "Pitch analysis for active music discovery," *33rd Int. Conf. on Machine Learning*, June 2016.

[6] Chris Greenhalgh et al., "GeoTracks: Adaptive music for everyday journeys," *ACM Int. Conf. on Multimedia*, Oct 2016.

[7] Rachel Bittner et al., "MedleyDB: a multitrack dataset for annotation-intensive MIR research," *15th International Society for Music Information Retrieval Conf. (ISMIR 2014)*, Oct 2014.

[8] Alex Wilson and Bruno Fazenda, "Variation in multitrack mixes: Analysis of low-level audio signal features," *Journal of the Audio Engineering Society*, vol. 64, no. 7/8, pp. 466–473, Jul/Aug 2016.

[9] Alex Wilson and Bruno Fazenda, "101 mixes: A statistical analysis of mix-variation in a dataset of multi-track music mixes," *Audio Engineering Society Convention 139*, Oct 2015.

[10] Mike Senior, *Mixing Secrets*, Taylor & Francis, `www.cambridge-mt.com/ms-mtk.htm`, 2012.

[11] Brecht De Man et al., "The Open Multitrack Testbed," *Audio Engineering Society Convention 137*, Oct 2014.

[12] Brecht De Man et al., "Perceptual evaluation of music mixing practices," *Audio Engineering Society Convention 138*, May 2015.

[13] Brecht De Man and Joshua D. Reiss, "APE: Audio Perceptual Evaluation toolbox for MATLAB," *Audio Engineering Society Convention 136*, Apr 2014.

[14] Nicholas Jillings et al., "Web Audio Evaluation Tool: A browser-based listening test environment," *12th Sound and Music Computing Conf.*, July 2015.

[15] Floyd E. Toole and Sean Olive, "Hearing is believing vs. believing is hearing: Blind vs. sighted listening tests, and other interesting things," *Audio Engineering Society Convention 97*, Nov 1994.

[16] S. Bech and N. Zacharov, *Perceptual Audio Evaluation - Theory, Method and Application*, John Wiley & Sons, 2007.

[17] Jan Berg and Francis Rumsey, "Validity of selected spatial attributes in the evaluation of 5-channel microphone techniques," *Audio Engineering Society Convention 112*, Apr 2002.

[18] Neofytos Kaplanis et al., "A rapid sensory analysis method for perceptual assessment of automotive audio," *Journal of the Audio Engineering Society*, vol. 65, no. 1/2, pp. 130–146, Jan/Feb 2017.

[19] Sean Olive and Todd Welti, "The relationship between perception and measurement of headphone sound quality," *Audio Engineering Society Convention 133*, Oct 2012.

[20] Christoph Völker and Rainer Huber, "Adaptions for the MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA) for elder and technical unexperienced participants," *DAGA*, Mar 2015.

[21] Jouni Paulus, Christian Uhle, and Jürgen Herre, "Perceived level of late reverberation in speech and music," *Audio Engineering Society Convention 130*, May 2011.

[22] Brecht De Man, Kirk McNally, and Joshua D. Reiss, "Perceptual evaluation and analysis of reverberation in multitrack music production," *Journal of the Audio Engineering Society*, vol. 65, no. 1/2, pp. 108–116, Jan/Feb 2017.

[23] Ute Jekosch, "Basic concepts and terms of 'quality', reconsidered in the context of product-sound quality," *Acta Acustica united with Acustica*, vol. 90, no. 6, pp. 999–1006, November/December 2004.

[24] Slawomir Zielinski, "On some biases encountered in modern listening tests," *Spatial Audio & Sensory Evaluation Techniques*, Apr 2006.

[25] Francis Rumsey, "New horizons in listening test design," *Journal of the Audio Engineering Society*, vol. 52, pp. 65–73, Jan/Feb 2004.

[26] Stanley P. Lipshitz and John Vanderkooy, "The Great Debate: Subjective evaluation," *Journal of the Audio Engineering Society*, vol. 29, no. 7/8, pp. 482–491, Jul/Aug 1981.

[27] Brecht De Man et al., "An analysis and evaluation of audio features for multitrack music mixtures," *15th International Society for Music Information Retrieval Conf. (ISMIR 2014)*, Oct 2014.

[28] Recommendation ITU-R BS.1770-4, "Algorithms to measure audio programme loudness and true-peak audio level," Oct 2015.

[29] Brecht De Man and Joshua D. Reiss, "Analysis of peer reviews in music production," *Journal of the Art of Record Production*, vol. 10, July 2015.

[30] David Ronan et al., "The impact of subgrouping practices on the perception of multitrack mixes," *Audio Engineering Society Convention 139*, Oct 2015.

[31] Stylianos Ioannis Mimilakis et al., "New sonorities for jazz recordings: Separation and mixing using deep neural networks," *2nd AES Workshop on Intelligent Music Production*, Sep 2016.