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.

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/](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-
Proceedings of the 20th International Conference on Digital Audio Effects (DAFx-17), Edinburgh, UK, September 5–9, 2017

The authors examine audio features extracted from 101 mixes of the same song, evaluated by one person who classified the mixes in five preference categories [2]. 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 (QMI) 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 2.2. Songs with an asterisk (*) are copyrighted and available. These correspond to the original release or analogue mixes, see Section 2.2. Songs with an asterisk (*) are copyrighted and not available online, whereas raw tracks to others can be found via the Open Multitrack Testbed[14]. 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 multilayering.

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 proved 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 [11].

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.

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

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 2.2. Songs with an asterisk (*) are copyrighted and not available online, whereas raw tracks to others can be found via the Open Multitrack Testbed[14]. 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 multilayering.

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 proved 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 [11].

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

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

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](http://multitrack.eecs.qmul.ac.uk). 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](https://weathervanemusic.org/shakingthrough) and [Mike Senior’s Mixing Secrets Multitrack Library](http://www.mikemonitracklibrary.com), 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 recre-ated, 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. Where 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).

#### 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.

![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](image-url)
Table 2: Overview of evaluation experiments

<table>
<thead>
<tr>
<th>Country</th>
<th>McG</th>
<th>MG</th>
<th>QM</th>
<th>SMC</th>
<th>DU</th>
<th>PXL</th>
<th>UCP</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>#subjects</td>
<td>33</td>
<td>8</td>
<td>21</td>
<td>26</td>
<td>39</td>
<td>13</td>
<td>10</td>
<td>150</td>
</tr>
<tr>
<td>#songs</td>
<td>10</td>
<td>4</td>
<td>13</td>
<td>14</td>
<td>3</td>
<td>4</td>
<td>7</td>
<td>18</td>
</tr>
<tr>
<td>#mixes</td>
<td>98</td>
<td>40</td>
<td>111</td>
<td>116</td>
<td>805</td>
<td>236</td>
<td>310</td>
<td>4873</td>
</tr>
<tr>
<td>#evaluations</td>
<td>1444</td>
<td>310</td>
<td>1129</td>
<td>639</td>
<td>805</td>
<td>236</td>
<td>310</td>
<td>14206</td>
</tr>
<tr>
<td>#statements</td>
<td>4227</td>
<td>585</td>
<td>2403</td>
<td>1190</td>
<td>2331</td>
<td>909</td>
<td>1051</td>
<td>12096</td>
</tr>
<tr>
<td>#words/comment</td>
<td>13,39</td>
<td>11,76</td>
<td>11,32</td>
<td>12,39</td>
<td>18,95</td>
<td>31,94</td>
<td>25,21</td>
<td>15,25</td>
</tr>
<tr>
<td>Male/female</td>
<td>28/5</td>
<td>7/1</td>
<td>18/3</td>
<td>14/12</td>
<td>33/6</td>
<td>13/0</td>
<td>9/1</td>
<td>122/28</td>
</tr>
<tr>
<td>Loudspeakers/headphones</td>
<td>LS</td>
<td>HP</td>
<td>LS</td>
<td>HP</td>
<td>LS</td>
<td>LS</td>
<td>LS</td>
<td></td>
</tr>
</tbody>
</table>

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, 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 [13]. 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).
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].

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. Figure 6a. 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 \(r_{A,B}\) between two groups \(A\) and \(B\) is given by

\[
\frac{\text{Number of agreements}}{\text{Number of statements in } A \times \text{Number of statements in } B}.
\]
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 [11].

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


Proceedings of the 20th International Conference on Digital Audio Effects (DAFx-17), Edinburgh, UK, September 5–9, 2017

Figure 7: Level of agreement between groups of different expertise


