Audio Engineering Society

Convention Paper 9235

Presented at the 138th Convention
2015 May 7–10 Warsaw, Poland

Perceptual Evaluation of Music Mixing Practices

Brecht De Man[1], Matthew Boerum[2,3], Brett Leonard[4], Richard King [2,3], George Massenburg[2,3], and Joshua D. Reiss[1]

[1] *Centre for Digital Music, Queen Mary University of London*

[2] *The Graduate Program in Sound Recording, Schulich School of Music, McGill University*

[3] *Centre for Interdisciplinary Research in Music Media and Technology*

[4] *The School of Music, University of Nebraska at Omaha*

Correspondence should be addressed to Brecht De Man (`b.deman@qmul.ac.uk`)

ABSTRACT

The relation of music production practices to preference is still poorly understood. Due to the highly complex process of mixing music, few studies have been able to reliably investigate mixing engineering, as investigating one process parameter or feature without considering the correlation with other parameters inevitably oversimplifies the problem. In this work, we present an experiment where different mixes of different songs, obtained with a representative set of audio engineering tools, are rated by experienced subjects. The relation between the perceived mix quality and sonic features extracted from the mixes is investigated, and we find that a number of features correlate with quality.

1. INTRODUCTION

Mixing music is a complex, expert process, during which the mixing engineer is expected to solve technical issues (e.g. ensuring the audibility of sources) as well as to make important creative choices to implement the musical vision of the artist, producer and/or themselves (e.g. positioning the respective instruments in the sonic space) [1]. As such, even highly trained mixing engineers are likely to deliver significantly different mixes of the same song, indicating that there is no single best mix [2]. The search for underlying rules of the mixing process is further complicated since many versions of the same song might be considered commercially viable.

Very little is known about preference of music production practices and there have been few rigorous studies addressing this problem [3–5]. Therefore furthering the understanding of the multidimensional, nonlinear problem of crafting a mix from raw musical elements is an important research area, with applications in the design of intelligent audio processing tools. This type of testing helps to investigate the larger question of why mixing engineers make certain decisions while they work.

Previous studies have analysed how subjects set parameters of a common audio processor for different musical fragments, for one or more simultaneously playing tracks [6–13]. However, by constraining the number of parameters of the inherently cross-adaptive mixing process to those controlling a single processor, we acquire only limited insight into the use of this particular tool. Certain processors may be used quite differently when other parameters are also accessible. For instance, in a method of adjustment test where only dynamic range compression parameters are to be set for different tracks (see [13]), subjects may use these controls to change the relative level of tracks in the absence of faders, whereas they might use the compressor differently in a real life mixing environment.

In this work, we want to gain insight into the mixing process and how mixing decisions influence listener's preference, by studying realistic mixes. To this end, we collect multiple mixes for different songs and conduct a perceptual evaluation study to assess their perceived quality. By doing so we find out whether there is a tendency to prefer one's own mix, and whether more experienced mixing engineers produce mixes of higher perceived quality.

We also investigate how the perceptual evaluation results can be understood by looking at audio features extracted from the mixdown. Specifically, we assess which features, if any, correlate strongly with preference.

In the following section, we describe the mixing experiment which we set up to collect a large amount of realistic data, with several mixes for each song. The listening test design and results from the perceptual evaluation of these mixes are discussed in section 3. Furthermore, features were extracted from the mixes to correlate these with the results of the listening test, and findings are reported in section 4. Section 5 summarises the conclusions from these results and future work is outlined in section 6.

2. EXPERIMENT

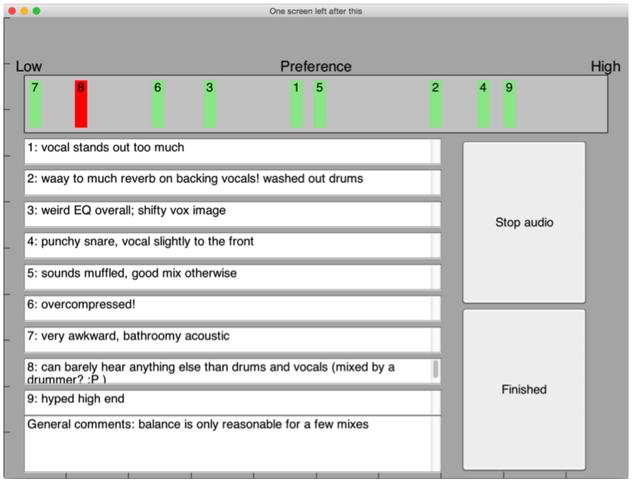
In the present study, groups of nine engineers were asked to mix ten different songs in a setting that was both realistic (as close as possible to their usual work flow so that the results are meaningful and relevant) and constrained (using a set number of tools and techniques so that the resulting mixes were sufficiently comparable and processing could be analysed after the fact).

The mixes were created by students of the MMus in Sound Recording at the Schulich School of Music, McGill University, and one professional mix was submitted by their teacher, who often also mixed the released version of the song. The experiments were spread over three terms: four songs were mixed in the Fall of 2013, four in the Spring of 2014, and two in the Fall of 2014. Each song was mixed by one of the two classes of eight students each, such that one group of students mixed five songs in total (over three semesters - four as first years and one more as second years), one group mixed four songs in total (over two semesters) and one group mixed just one song.

The reason for having each song mixed by just half of the available mixing engineers is to allow perceptual evaluation by a group of experienced subjects who are not familiar with the song, and therefore have no bias from having made mixing decisions during their own mix and generally being exposed to the song.

Each mixing engineer allocated up to six hours to each of their mix assignments in which they were to produce a stereo mix using an industry standard DAW with included plugins, and a high quality reverberation plugin. This corresponded well with the tools they were used to using, and constraining the toolset allowed us to reproduce and analyse their mixes in detail. Editing, rerecording, the use of samples or any other form of adding new audio was not allowed.

An automatic mix was rendered for eight out of ten mixes, directly based on automatic mixing algorithms from [8, 13–17].



1: vocal stands out too much

2: waay to much reverb on backing vocals! washed out drums

3: weird EQ overall; shifty vox image

4: punchy snare, vocal slightly to the front

5: sounds muffled, good mix otherwise

6: overcompressed!

7: very awkward, bathroomy acoustic

8: can barely hear anything else than drums and vocals (mixed by a drummer? :P)

9: hyped high end

General comments: balance is only reasonable for a few mixes

Fig. 1: Listening test interface

3. PERCEPTUAL EVALUATION

The mixes were evaluated in a listening test to infer the quality, as perceived by a group of trained listeners. The subjects were Sound Recording master students, their teachers, and other students and staff at the same institution. For each song, there were between 13 and 22 ratings per mix.

For the purpose of perceptual evaluation, we considered a fragment of the song only, consisting of the second verse and chorus. This reduces the strain on the subjects' attention, likely leading to more reliable listening test results. It also places the focus on a region of the song where the most musical elements are active.

We used the APE toolbox described in [18], developed specifically with evaluation of music production practices in mind, which allows rating of the nine to ten mixes on a single horizontal axis (to encourage careful ordering and rating between the respective mixes) and commenting using numbered text boxes (to gain more insight into the preferences and to facilitate the subject's process by taking notes), as in Fig. 1. From its position on the mix 'quality' axis, every mix received a rating between 0 and 100 (continuous). The order of songs and of the mixes within each song was randomised, aiming for an equal amount of participants for each permutation of songs. Per song, each mix was loudness normalised using [19] to avoid bias towards (or away from) mixes which were louder on average. One session consisted of between two and four songs, and subjects spent on average 17 min ± 8 min per song, well within the recommendations from [20].

All listening tests took place in CIRMMT's Critical Listening Lab (see Fig. 2). The impulse responses of the listening environment including playback system can be found on `www.brechtdeman.com/research.html`.

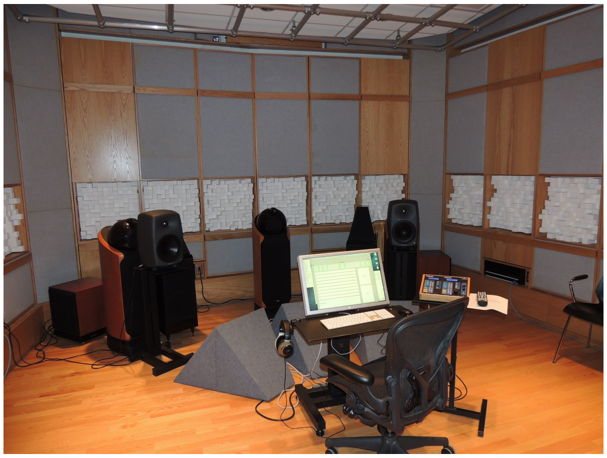


Fig. 2: The Critical Listening Lab at CIRMMT, where all listening tests took place.

Fig. 3 shows the ratings received by every mixing engineer in the test (for one or more songs) including the teachers ('P1' and 'P2', shown together as 'Pro') and the completely autonomous mix ('Auto'), as well as the combined ratings received by first year ('Y1') and second year ('Y2') students. While subjects did not agree on a clear order in terms of preference in this case, there is a definite tendency to prefer certain mixes over others. Mixes by second year students are only given a slightly higher preference rating on average than those by first year students, although it should be noted the two are never assessed at the same time.

Two songs were also evaluated by the group of engineers who mixed the song, so that each would also assess their own mix. Except for one engineer, who rated his own mix lowest, all rated their own mix

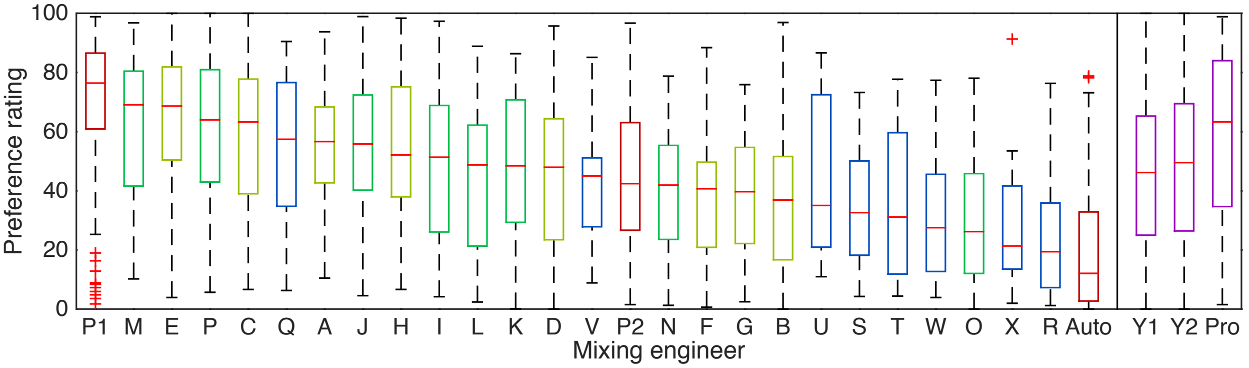


Fig. 3: Box plot of ratings per mixing engineer, in decreasing order per median. A-H are first year students in 2013-2014 (4 songs), and second year students in 2014-2015 (1 song); I-P are second year students in 2013-2014 (4 songs), and Q-X are first year students in 2014-2015 (1 song). 'P1' and 'P2' are their teachers ('Pro'), 'Y1' and 'Y2' are the results of mixes by first year and second year students, respectively, and 'Auto' denotes the automatic mix.

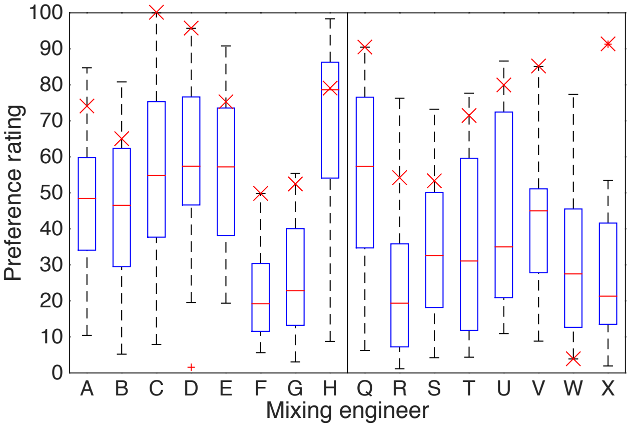


Fig. 4: Box plot of ratings per mixing engineer including their own assessment (red 'X') of one song.

higher than the average rating their mix received (see Fig. 4). 13 out of 16 participants also rated their mix higher than the average rating they attributed to other mixes of the same song. This suggests that engineers have a consistent taste whether they are mixing themselves or only listening, or they are biased by the way they have mixed this song themselves in the recent past, or they remember their own mix well (or a combination of these).

We found a strong correlation (Pearson's correlation coefficient .52, $p < 10^{-12}$) between the average rating of different mixes by the same mixing engineer, meaning that the perceived quality of a single mix is indicative of the general performance of the engineer.

4. FEATURE EXTRACTION AND ANALYSIS

A number of features were extracted from the 98 evaluated mixes (see Table 1). Only the perceptually evaluated fragment of the song was considered, to ensure the features were relevant to the evaluation. Whereas the audio was recorded, mixed and evaluated at 96 kHz/24 bit, we converted all audio to 44.1 kHz using SoX [21] to calculate spectral features based on the more perceptually relevant audible region.

As listed in Table 1, preference shows a positive linear correlation with microdynamics measure LDR [24] (.26, $p = .01$) and the 4th MFCC (.21, $p < .05$), and a negative linear correlation with the relative energy of the octave band centred at 4 kHz ($-.21$, $p < .05$) and spatial feature side-to-mid ratio [28] ($-.32$, $p = .001$).

The preference rating used to calculate these correlations is the average of all raw ratings for each mix, regardless of each subject's use of the scale. We investigated possible effects by post hoc scaling

Table 1: Pearson's (r), Spearman's (ρ) and Kendall's (τ) correlation coefficients (including p-values) of extracted features with perception (average of raw ratings).

Feature	r	p	ρ	p	τ	p	**Ref.**
crest factor (whole)	.092	.369	.101	.323	.062	.371	max. amplitude over RMS
crest factor (1s)	.012	.905	.003	.973	.000	.998	max. amplitude over RMS
crest factor (100ms)	-.053	.607	-.084	.415	-.053	.445	max. amplitude over RMS
dynamic spread	.093	.364	.128	.211	.088	.205	[22]
peak-to-loudness ratio	.164	.109	.142	.165	.092	.183	[23, 24]
LRA	-.029	.779	.010	.919	.012	.859	[25]
TT DR	.052	.614	.029	.776	.017	.806	[26]
LDR	**.258**	**.011**	**.244**	**.016**	**.168**	**.015**	[24]
spectral centroid	-.146	.154	-.130	.204	-.086	.214	[27]
spectral flux	.141	.167	.169	.098	.111	.108	[27]
brightness	-.166	.105	-.181	.077	-.125	.069	[27]
spectral spread	-.143	.162	-.080	.433	-.055	.427	[27]
spectral skewness	.147	.150	.138	.178	.094	.175	[27]
spectral kurtosis	.128	.212	.136	.183	.092	.183	[27]
spectral rolloff 95%	-.159	.119	-.118	.248	-.076	.269	[27]
spectral rolloff 85%	-.150	.143	-.127	.216	-.089	.198	[27]
spectral entropy	-.156	.126	-.165	.107	-.113	.100	[27]
spectral flatness	-.129	.207	-.098	.340	-.070	.311	[27]
spectral roughness	-.019	.851	.007	.947	-.001	.988	[27]
spectral irregularity	-.028	.785	-.019	.854	-.010	.888	[27]
zero crossing rate	-.131	.201	-.168	.100	-.118	.088	[27]
low energy	.075	.464	.053	.606	.031	.651	[27]
cepstral flux	.011	.914	.087	.398	.054	.438	[27]
31.5 Hz octave band	.086	.400	.047	.649	.039	.577	energy of octave band divided by total power
63 Hz octave band	-.020	.845	-.044	.668	-.034	.625	
125 Hz octave band	-.152	.137	-.098	.339	-.064	.355	
250 Hz octave band	.110	.284	.044	.670	.030	.665	
500 Hz octave band	-.014	.890	-.072	.481	-.047	.499	
1 kHz octave band	-.023	.826	-.088	.389	-.057	.412	
2 kHz octave band	-.123	.231	-.146	.155	-.097	.161	
4 kHz octave band	**-.207**	**.042**	**-.135**	**.189**	**-.087**	**.209**	
8 kHz octave band	-.128	.212	-.054	.601	-.036	.601	
16 kHz octave band	-.003	.975	-.049	.634	-.038	.593	
MFCC1	.132	.197	.114	.266	.082	.235	
MFCC2	-.090	.382	-.031	.761	-.018	.801	
MFCC3	-.111	.278	-.120	.242	-.076	.274	
MFCC4	**.210**	**.039**	**.166**	**.105**	**.109**	**.116**	
MFCC5	.100	.328	.087	.394	.057	.412	
side-to-mid ratio	**-.320**	**.001**	**-.324**	**.001**	**-.223**	**.001**	[28]
left/right imbalance	-.007	.948	-.007	.948	-.006	.933	[28]
P_total	-.128	.211	-.138	.176	-.086	.214	[29]
P_low	.144	.158	.084	.410	.051	.464	[29]
P_mid	.036	.723	.011	.913	.024	.729	[29]
P_high	-.143	.164	-.158	.121	-.098	.154	[29]

of each subject's ratings for a given song between 0 and 100, subtracting the average rating for that song from each rating, using the median instead of the mean, and any combination of the above. In each of these cases, the correlations found were similar and not worth reporting separately, except for the peak-to-loudness ratio, which became significant for each of the modifications, and crest factor (over the whole file) in most of the cases.

Preference towards a higher amount of peak-to-loudness and short-time dynamic variation (recall that all mixes were evaluated at equal loudness) suggests that for a given loudness, a mix should have peaks of sufficient magnitude. While in many situations, a high loudness for a given peak amplitude seems to have a positive effect on the listener's relative preference [30,31], it seems that when the loudness is normalised instead of the peak amplitude, a relatively higher dynamic range is preferred over a lower one.

Mixes with a relatively higher proportion of energy in the 4 kHz octave band (and conversely a lower proportion of energy at other frequencies) tend to be rated lower. Correlation with the 4th MFCC further suggests that the spectral envelope could help predict the perceived quality.

The side-to-mid ratio is the ratio of the power of the side and mid channels [28], where side channel x_S and mid channel x_M relate to the left channel x_L and the right channel x_R as follows:

$$x_S = \frac{x_L - x_R}{2} \tag{1}$$

$$x_M = \frac{x_L + x_R}{2} \tag{2}$$

In other words, a stronger side channel means more monophonic sources are placed on either side of the stereo field, which is not preferred here. However, overly monophonic mixes (very low side-to-mid ratio) generally received low ratings as well.

Overall, these results suggest that dynamic, spectral and spatial features extracted from the audio can be predictive of the subjective quality.

5. CONCLUSION

We have described a mixing experiment and perceptual evaluation thereof in which 9 different mixes of 10 songs were evaluated by between 13 and 22 subjects each.

There is a strong tendency for engineers to like their own mixes better, possibly because of personal preferences that inform both the mixing process and the assessment of other mixes. Mixes from the same engineer are likely to receive a similar rating.

We have found evidence that features extracted from the mix audio can provide insight into the perceived quality of that mix. This suggests that by taking a combination of specialised features, of the mix audio or of the individual processed tracks [28], some insight can be gained into the relation between the subjective mix quality and objective, measurable features.

Specifically, the relations shown in this work point to very concrete, practical issues mixes may have, such as a limited dynamic range, or sources panned overly left or right. This could inform semantic and perceptually motivated mix metering where a user would be notified of features that are outside of a generally preferred area, using a simple measurement or machine learning method like anomaly detection.

6. FUTURE WORK

Further mixing experiments with mixing engineers at different locations will be necessary to investigate the effect of different schools and geographic locations and compensate for this potential bias. A larger number of experiments will also help understand and partially compensate for the influence of genre on mixing practices, generally improve the accuracy and provide additional insight into this complex problem. A larger number of subjects, with and without music production background, will further increase the relevance and reliability of our findings.

A forthcoming study will discuss the comments on the respective mixes, which help clarify subjects' ratings and enable us to investigate strengths and shortcomings of specific tracks. Similarly, the present work should be expanded by looking at extracted features of the different tracks, and relations between different tracks, to further understand what effect different mix actions have. The correlation with preference could also be higher for more sophisticated, perceptually motivated features, or a combination of the features above. Further work is required to understand exactly how objective features relate to the perceived quality of music.

A possible bias inherent to a subjective comparison of mixes is that bold mix decisions, as they are taken even or especially by the best engineers, are subconsciously penalised on account of being an outlier in the test, even if they could work in the context of a single commercial mix which is not traditionally compared to alternative mixes. At this point this is only speculative as the described perceptual evaluation does not take this bias into account.

Of the 10 songs used in this work, 6 can be downloaded[1] - including all raw tracks, rendered mixes and DAW files - from the Open Multitrack Testbed at `multitrack.eecs.qmul.ac.uk` [32]. The MATLAB listening test interface source code used for the perceptual evaluation, described in [18], can be downloaded from `code.soundsoftware.ac.uk/projects/ape`.

7. REFERENCES

[1] R. Izhaki, *Mixing audio: Concepts, Practices and Tools.* Focal Press, 2008.

[2] A. Case, *Mix Smart: Professional Techniques for the Home Studio.* Focal Press, 2011.

[3] P. Pestana and J. D. Reiss, “Intelligent audio production strategies informed by best practices,” in *53rd Conference of the Audio Engineering Society*, January 2014.

[4] S. Fenton, B. Fazenda, and J. Wakefield, “Objective measurement of music quality using inter-band relationship analysis,” in *130th Convention of the Audio Engineering Society*, May 2011.

[5] S. Fenton, B. Fazenda, and J. P. Wakefield, “Objective quality measurement of audio using multiband dynamic range analysis,” in *Institute of Acoustics (IOA) Conference 2009 - Reproduced Sound*, November 2009.

[6] R. King, B. Leonard, and G. Sikora, “Variance in level preference of balance engineers: A study of mixing preference and variance over time,” in *129th Convention of the Audio Engineering Society*, November 2010.

[7] J. Bitzer, J. LeBoeuf, and U. Simmer, “Evaluating perception of salient frequencies: Do mixing engineers hear the same thing?,” *124th Convention of the Audio Engineering Society*, May 2008.

[8] S. Mansbridge, S. Finn, and J. D. Reiss, “An autonomous system for multi-track stereo pan positioning,” in *133rd Convention of the Audio Engineering Society*, October 2012.

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

[10] M. Cartwright and B. Pardo, “Social-EQ: Crowdsourcing an equalization descriptor map,” *14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, November 2013.

[11] R. King, B. Leonard, and G. Sikora, “Consistency of balance preferences in three musical genres,” in *133rd Convention of the Audio Engineering Society*, October 2012.

[12] B. Leonard, R. King, and G. Sikora, “The effect of acoustic environment on reverberation level preference,” in *133rd Convention of the Audio Engineering Society*, October 2012.

[13] D. Giannoulis, M. Massberg, and J. D. Reiss, “Parameter automation in a dynamic range compressor,” *Journal of the Audio Engineering Society*, vol. 61, pp. 716–726, October 2013.

[1] Available under a CC BY 3.0 license: “Lead Me”, “Pouring Room” and “Under A Covered Sky” by The Donefors, “In The Meantime” and “Not Alone” by Fredy V, and “Red To Blue” by Broken Crank.

[14] S. Mansbridge, S. Finn, and J. D. Reiss, "Implementation and evaluation of autonomous multitrack fader control," in *132nd Convention of the Audio Engineering Society*, April 2012.

[15] 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.

[16] S. Hafezi and J. D. Reiss, "Autonomous multitrack equalisation based on masking reduction," *submitted to Journal of the Audio Engineering Society*, 2015.

[17] Z. Ma, B. De Man, P. D. Pestana, D. A. A. Black, and J. D. Reiss, "Intelligent multitrack dynamic range compression," *submitted to Journal of the Audio Engineering Society*, 2015.

[18] B. De Man and J. D. Reiss, "APE: Audio Perceptual Evaluation toolbox for MATLAB," in *136th Convention of the Audio Engineering Society*, April 2014.

[19] ITU, "Recommendation ITU-R BS.1770-3: Algorithms to measure audio programme loudness and true-peak audio level," *Radiocommunication Sector of the International Telecommunication Union*, 2012.

[20] R. Schatz, S. Egger, and K. Masuch, "The impact of test duration on user fatigue and reliability of subjective quality ratings," *Journal of the Audio Engineering Society*, vol. 60, no. 1/2, pp. 63–73, 2012.

[21] "Sox - sound exchange," *webpage (http://SoX.sourceforge.net)*, Accessed on 12 February 2015.

[22] E. Vickers, "Automatic long-term loudness and dynamics matching," in *111th Convention of the Audio Engineering Society*, November 2001.

[23] ITU, "Recommendation ITU-R BS.1771-1: Requirements for loudness and true-peak indicating meters," *Radiocommunication Sector of the International Telecommunication Union*, 2012.

[24] E. Skovenborg, "Measures of microdynamics," in *137th Convention of the Audio Engineering Society*, October 2014.

[25] EBU, "Loudness Range: A measure to supplement loudness normalisation in accordance with EBU R 128," *European Broadcasting Union*, August 2011.

[26] Pleasurize Music Foundation, "TT Dynamic Range Meter," *webpage (http://dynamicrange.de)*, 2013.

[27] O. Lartillot and P. Toiviainen, "MIR in Matlab (II): A toolbox for musical feature extraction from audio," in *8th International Society for Music Information Retrieval Conference (ISMIR 2007)*, September 2007.

[28] 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.

[29] G. Tzanetakis, R. Jones, and K. McNally, "Stereo panning features for classifying recording production style," in *8th International Society for Music Information Retrieval Conference (ISMIR 2007)*, September 2007.

[30] E. Vickers, "The loudness war: Background, speculation, and recommendations," *129th Convention of the Audio Engineering Society*, November 2010.

[31] E. Skovenborg, "Loudness Range (LRA) - Design and evaluation," in *132nd Convention of the Audio Engineering Society*, April 2012.

[32] 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.