Perception of audio quality in productions of popular music

Wilson, AD and Fazenda, BM

10.17743/jaes.2015.0090

OPEN ACCESS Freely available online

Journal of the Audio Engineering Society
Vol. 64, No. 1/2, January/February 2016 (© 2016)
DOI: http://dx.doi.org/10.17743/jaes.2015.0090

Perception of Audio Quality in Productions of Popular Music

ALEX WILSON, *AES Student Member,* **AND BRUNO M. FAZENDA,** *AES Member*
(a.wilson1@edu.salford.ac.uk)

Acoustics Research Centre, University of Salford, Greater Manchester, M5 4WT, UK

The quality of recorded music is often highly disputed. To gain insight into the dimensions of quality perception, subjective and objective evaluation of musical program material, extracted from commercial CDs, was undertaken. It was observed that perception of audio quality and liking of the music can be affected by separate factors. Familiarity with stimuli affected *like* ratings while *quality* ratings were most associated with signal features related to perceived loudness and dynamic range compression. The effect of listener expertise was small. Additionally, the sonic attributes describing *quality* ratings were gathered and indicate a diverse lexicon relating to timbre, space, defects, and other concepts. The results also suggest that while the perceived *quality* of popular music may have decreased over recent years, *like* ratings were unaffected.

0 INTRODUCTION

In the context of recorded sound there is great debate over which parameters influence the perception of quality or how quality should be defined. In the context of product development, sound quality has been defined as the "result of an assessment of the perceived auditory nature of a sound with respect to its desired nature" [1]. In order to assess the audio quality of a recording, the requirements for quality must be identified as well as the inherent characteristics of the audio signal. These characteristics must then be measured and used to estimate quality, which is then optimized subject to various constraints, e.g., the available budget, human resources, and projected time-to-market. This paper details the findings of a study into the perception of quality in commercial music productions, attempting to ascertain which objective and subjective parameters are involved as well as the relative importance of these parameters.

1 ASSESSMENT OF QUALITY

A variety of theories and methodologies exist for the assessment of quality in many different fields. A number of these can be applied to reproduced sound. In this context, quality judgments can be considered to be based on technical properties of the signal, such as bandwidth or distortion, or based on hedonic preference, which might be influenced by personal aspects of familiarity.

International standards exist regarding the measurement of audio quality based on determining the level of degradation from a reference [2]. These procedures are formulated under the assumption that a reference item exists, which can be used as an example of greatest quality, and test items are then compared against this reference. This usually applies to systems where the reference is formed from the original version of the program material and the test samples under evaluation are copies that have undergone some form of processing. The evaluation of systems such as audio codecs [3] is a good example of this type of approach. In these circumstances, it is not strictly the inherent quality of the program material that is being measured but rather the perceived degradation in quality of the signal, after being subject to destructive processes. In effect, the evaluation of the audio signal is being used as an intermediate step towards evaluating the algorithm, reproduction system, or other such device under test.

This approach to quality evaluation is difficult to apply to music productions as it is unlikely there exists a reference audio sample (a recording of a particular song), which represents the maximum quality rating, to which all other samples (other recordings of other songs) could be compared. Nonetheless, aspects of this approach can be useful. For example, a number of studies have pointed to the importance of distortion on the perception of audio quality. Often, non-linear distortion has been considered, where the intensity of the distortion has been shown to degrade the quality of both speech and music signals [4–8]. Similar considerations have been made regarding the use of dynamic range compression on music signals [9–11] as well as bandwidth and quantization distortion [12].

Additionally, a growing area of research is the assessment of quality/preference in music mixing practices [13–15]. In each of these studies the understanding of quality differs. Quality and preference are often conflated, as studies reporting on "quality" may have asked for "preference" ratings during testing [14]. It is clear that many possible explanations of the term "quality" can be applied to audio signals and that these descriptions have similarities as well as differences.

The work reported here presents an investigation into audio quality from both a technical as well as a hedonic preference approach. This has been done using a diverse set of audio stimuli within popular music and analyzed using multiple methodologies. In looking to investigate perception of technical quality and its differentiation from hedonic preference, or how much someone likes a song, it was ultimately decided not to directly define quality to the participants. This decision has allowed some ambiguity to remain and the consequences are discussed herein.

2 MOTIVATION

The aims of this work were, first, to investigate which attributes described the assessment of technical quality and how a subjective rating related to objective parameters extracted from the signal. Objective parameters have been used to characterize various forms of audio technology, such as loudspeakers, amplifiers, codecs, and recently music mixes (see Sec. 1). Correlations between objective measures of the audio signals and the subjective impressions provide insight into the perception of quality and of various music production techniques and signal processing procedures, such as equalization and dynamic range compression [13]. Earlier work indicated correlations between signal parameters and the quality ratings of music recordings [16]—the work herein expands on this.

The second aim was to quantify the hedonic appraisal of music samples—to understand the effect that familiarity might have on this "like" rating and whether like is distinct from quality. Ratings of pleasure when listening to music are related to emotional arousal [17] and an increase in blood oxygen level in regions of the brain related to emotion has been measured when listening to familiar music [18]. From this there is reason to believe that familiarity may play a role in preference ratings for music, as indicated by a number of studies [18–21]. Since the elements of preference, which may be more related to a hedonic assessment, are sometimes confounded with those of perceived quality (e.g., [14]), there is an interest in defining the interaction between these two methods of assessment.

Finally, as a demographic indicator, an investigation into the effect of expertise on quality-perception was undertaken. This is of interest since the use of expert listeners is commonly advocated for audio experiments [22] but, in contrast, studies have indicated that experts have unique behaviors and can be prone to biases that are not present in, or do not influence, non-experts [23, 24].

Being interested in the understanding of overall perceived quality, rather than the measurement of a specific, limited definition of quality, it is important to allow multiple interpretations, especially since the terms used to describe audio quality can be diverse [25]. Based on these motivations, as described in this section, the following research questions were devised for this study.

Q1. Are *quality* ratings related to objective measures of the music signal and if so, how?
Q2. Is the percept of *liking* a song distinct from that of assessing its *quality*?
Q3. What influence does *familiarity* with a song have on listener preference?
Q4. Does listener expertise have a significant influence on perception of quality?
Q5. Which words are used to justify *quality* ratings and is there significant variation in the words used to describe varying levels of *quality*?

3 METHODOLOGY

This section describes the test methodology that was implemented from the choice of audio stimuli and test participants, to the experimental set-up and analysis of audio stimuli using feature extraction.

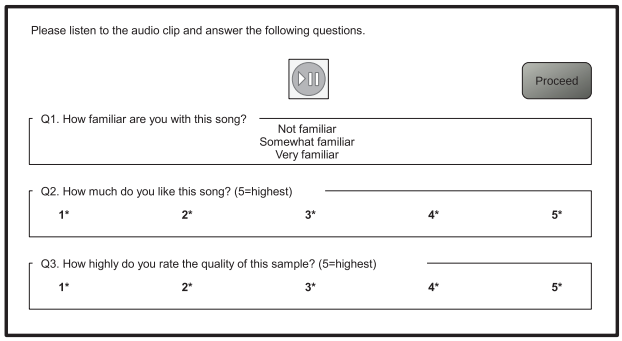
3.1 Audio Dataset

To provide a dataset of audio samples for study, audio was extracted from commercially released compact discs (stereo .WAV files with 16-bit resolution and sampling rate of 44.1 kHz). In previous studies regarding the analysis of music releases, samples were included from the 1950s [26] or 1960s [9, 11] onwards. In these studies, music samples that pre-date the commercial release of the CD (1982) would have been remastered for release in a digital format at a later date. As such, in order to be confident that the data obtained truly represents production trends of its year of release, this study only considers samples from the original digital sources, dating back to 1982. This dataset contained 321 songs by 229 artists, with an average of ten songs per year from 1982 to 2013, sourced from available CDs in a variety of genres (see [10]).

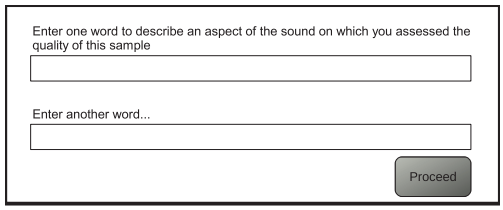
3.2 Test Design

In total, 63 songs were chosen from this dataset for the listening test. These were chosen randomly such that there was an even distribution over the 31-year period. Each sample was 20 seconds in duration, centered about the song's second chorus. This region was chosen for consistency, as a chorus is often a memorable part of the song. For songs without a chorus, or where the chorus does not feature lead vocals, an alternative section was chosen based on audition. A 1 s fade-in and fade-out were applied.

Being examples of popular music, these samples would be familiar to participants to varying degrees. A "not familiar" option was included for samples that were not familiar. One clip was used at the beginning of each test to serve as a trial and from there on the order of playback was



(a) GUI with questions 1 to 3



(b) GUI with question 4

Fig. 1. Illustration of the graphical user interface which was used in listening test.

randomized. An optional break was automatically suggested when 40% of the trials were completed.

Four questions were presented for each audio sample. The test interface for questions 1, 2, and 3 is shown in Fig. 1a and for question 4 in Fig. 1b. The interface also contained a play/pause button for controlling audio playback. The like and quality ratings were provided using a 5-star scale, as also used in other contemporary studies [12].

While quality was not strictly defined in this context, the request for a like rating in the same answer box forces the participants into a deliberate distinction between the two. To investigate how quality was interpreted, the participant was asked for two words to describe attributes of the sample on which quality was assessed. Commonly used words were provided (see Appendix A.1 and [25]).

The test took place in the listening room at University of Salford, a room that conforms to appropriate standards set out in ITU-R BS 1116-1 [2]. Audio was delivered via Sennheiser HD 800 headphones, the frequency response of which was measured using a Brüel & Kjær Head and Torso Simulator (HATS). Low-frequency rolloff in the response below 110 Hz was compensated using an IIR filter designed using the Yule-Walker method. As this compensation boosted the response at low frequencies, the addition of a notch filter at 0 Hz was required to ameliorate the increased DC offset. To avoid clipping, audio was attenuated prior to equalization.

The reproduction system consisted of the test computer, a Focusrite Scarlett 2i4 USB interface, and the headphones. The loudness of all audio samples was normalized according to BS.1770-3 [27], after the previously described headphone compensation had taken place. The target loudness for normalization was –22 LUFS, providing ample headroom. The presentation level to participants was set to 82 dB *LAeq*, considered to be a suitably realistic level for headphone reproduction. This level was set by recording a 1 kHz calibration signal at 94 dB through the HATS microphone onto the test computer. The loudness-normalized program material was then played back over headphones situated on the HATS and recorded through the same signal chain.

3.3 Test Panel

The total number of participants was 22 (4 female, 18 male), tested over a period of five consecutive days in February of 2014. Each participant was asked to choose their level of expertise based on participation in previous listening tests. From this self-reported response, there were 13 experts and 9 non-experts. The median age of the participants was 23 years, ranging from 19 to 39 years. No participant reported any serious hearing impairment. Each participant chose two preferred musical genres as an open question—from these responses it was observed that the participants had diverse preferences, as the categories proposed by [28] were represented (mellow, unpretentious, sophisticated, intense, and contemporary). The overall test duration varied by participant, with median duration of 38 minutes, ranging from 22 to 69 minutes. As the test contained the option of a break, any effects of fatigue on the reliability of subjective quality ratings were considered to be negligible, in line with guidelines suggested in recent literature [29]. Participants were monitored from outside the room but were able to request assistance if needed. All necessary ethical approval was obtained based on the policies of the University of Salford.

3.4 Feature Extraction

In order to compare attributes of the signal to subjective measures, various objective features of the audio were extracted consisting of amplitude, spectral, spatial, and rhythmic features. Many of these features are time-varying and can be calculated as such. In this case, as the samples were short in duration (20 seconds), the features were evaluated over the entire segment.

A number of feature extraction tasks were aided by the use of the MIRtoolbox [30]. The objective predictions of emotional response were also used [31]—these higher-level features have been shown to relate to audio quality in earlier studies [16], however, with the caveat that these features may not generalize to modern popular music [16, 32, 33]. This may be due to the fact that the original study used a dataset of film soundtrack music [31], which would rarely be as heavily processed as modern pop and rock music. Each of these emotional prediction features is calculated using a multiple linear regression model [31] and so the constituent factors of each prediction have also been evaluated. These are listed as emotion factors (the factors not found to have a significant correlation to either like or quality ratings are not reported).

The spatial features were based on the Stereo Panning Spectrogram (SPS) [34]. The SPS compares the left and

right channels of a given audio signal in the time-frequency plane and derives a two-dimensional map that can identify the panning gains associated with each time-frequency bin. In the current study, width values were obtained for each frequency by evaluating the standard deviation of the panning gains along each frequency slice in the SPS. The features used were created by obtaining the average of this function over different frequency bands.

The probability mass function (PMF) of the sample amplitudes of each audio clip was evaluated and subsequently reduced to a histogram, as described in the authors' previous work [10, 16]. The gauss feature, kurtosis, spread, and flatness of the PMF were thus extracted [30].

In order to characterize both amplitude and spectral characteristics of the audio signals, the Sub-band Spectral Flux was determined [35]. In this process the audio signal is processed by a bank of filters and, for each filtered output, the Euclidean distance between spectra of adjacent frames of audio is determined. In the original study [35], it was found that bands 1, 2, 3, 6, 7, and 8 were correlated to perceptual dimensions of polyphonic timbre (activity, brightness, and fullness), however all bands were used in the study reported herein. The list of features is shown in Table 3.

4 RESULTS

This section presents the results of the analysis of subjective responses, correlations of the signal features with subjective responses, an exploratory factor analysis of signal features, and a brief analysis of the words used to describe quality ratings.

4.1 Subjective Attributes

With 63 audio samples and 22 subjects, these 1386 auditions were gathered and analysis was performed on this dataset. In order to ascertain the importance of subjective measures in the assessment of quality and like, a 3-way multivariate analysis of variance (MANOVA) was performed (using IBM SPSS Statistics V.20), with independent variables of music sample, expertise, and familiarity. The results are shown in Table 1. The assumptions for MANOVA were tested using Box's test of equality of covariance matrices (the Box's M value of 686.15 was associated with a p-value of 0.802, which was interpreted as non-significant) and using Bartlett's test of sphericity, which is significant ($\chi^2(2, N = 1386) = 88.346, p < 0.001$).

Using Wilks' Λ, there was a significant effect of sample ($\Lambda = 0.597, F(124, 2144) = 5.082, p < 0.001$), familiarity ($\Lambda = 0.721, F(4, 2144) = 95.313, p < 0.001$), and expertise ($\Lambda = 0.991, F(2, 1072) = 4.694, p = 0.009$) on the ratings of like and quality. For Wilks' Λ, the effect size is calculated as follows: $\eta_p^2 = 1 - \Lambda^{1/s}$, where $s =$ (the number of groups $- 1$) or the number of dependent variables, whichever is smaller.

The multivariate test was followed-up by univariate analysis of variance (ANOVA), the results of which are shown in Table 2. For ANOVA, effect sizes are calculated according to the usual conventions [36].

None of the interactions were found to be significant, while all main effects were significant. While the MANOVA test showed a correlation between raw like and quality ratings of $R^2 = 0.26$, when mean like and mean quality values are evaluated for each song, the value of $R^2 = 0.02$, a non-significant correlation. The mean like and quality ratings for each audio sample are shown in Fig. 2, arranged in order of ascending quality illustrating the non-existing correlation.

Expertise does not appear to be as important a factor in this study as evidenced by the lower η^2 and observed power in Table 2. There is a large effect of the variable familiarity on like ratings (that will be discussed later) and a small effect of familiarity on quality ratings.

4.2 Objective Signal Features

Features extracted from the signal were compared against quality and like ratings gathered by the subjective test. A linear function was fitted using the mean like and quality ratings for each song and the goodness-of-fit is shown by R^2 and associated p-values in Table 3. Features for which a significant correlation was found (where $p < 0.05$) are highlighted in bold. Since the value shown is R^2, which spans the range 0 to 1, arrows indicate positive ($\Uparrow$) or negative ($\Downarrow$) correlation, as determined by the polarity of Pearson's r.

From this data it can be seen that there is a difference between the quality and like ratings in terms of responsible parameters. Like ratings were correlated with spectral features while quality ratings were correlated with amplitude features. The correlations with emotion factors support this. Quality was correlated with both RMS and roughness while like was correlated with spectral spread. Spectral flux serves as both an indicator of amplitude and spectral

Table 1. Results of 3-way MANOVA. Significant p-values (<0.05) are highlighted by an asterix.

Effect	Wilks' Λ	F	Hyp. df	Error df	p	η_p^2	Obs. power
Sample	.597	5.082	124	2144	.000*	.227	1.000
Familiar	.721	95.313	4	2144	.000*	.151	1.000
Expertise	.991	4.694	2	1072	.009*	.009	.788
S×F	.808	1.009	220	2144	.162	.101	1.000
S×E	.879	1.151	124	2144	.127	.062	1.000
E×F	.997	.672	4	2144	.611	.001	.221
S×F×E	.884	.937	146	2144	.689	.060	1.000

Table 2. Results of 3-way ANOVA follow-up. Significant p-values (<0.05) are highlighted by an asterix.

Source		df	F	p	η_p^2	η^2	Obs. power
Sample	Like	62	4.418	.000*	.203	.127	1.000
	Quality	62	5.542	.000*	.243	.201	1.000
Familiar	Like	2	201.927	.000*	.273	.187	1.000
	Quality	2	20.360	.000*	.037	.024	1.000
Expertise	Like	1	4.126	.042*	.004	.002	.528
	Quality	1	7.532	.006*	.007	.004	.783
S×F	Like	110	1.170	.121	.107	.060	1.000
	Quality	110	.977	.551	.091	.063	1.000
S×E	Like	62	1.167	.181	.063	.033	.998
	Quality	62	1.027	.422	.056	.037	.992
E×F	Like	2	1.230	.293	.002	.001	.269
	Quality	2	.230	.794	.000	.000	.086
S×E×F	Like	73	.907	.697	.058	.031	.990
	Quality	73	.992	.498	.063	.042	.995
Error	Like	1073					
	Quality	1073					
Total	Like	1386					
	Quality	1386					

characteristics—higher values indicate greater amplitudes and were negatively correlated with quality. There was no significant correlations found between spatial features or rhythmic features and either like or quality ratings.

4.3 Principal Component Analysis

In order to reduce the dimensions of the feature space Principal Component Analysis (PCA) was performed. Only the statistically significant features from Table 3 were initially considered for use in the PCA.

Using Bartlett's test of sphericity, the null hypothesis that the correlation matrix of the data is equivalent to an identity matrix was rejected ($\chi^2(325, N = 62) = 2674, p < 0.001$). This indicates that factor analysis can be performed, while a Kaiser-Meyer-Olkin measure of sampling adequacy (*MSA*) of 0.837, above the recommended value of 0.6 [39], suggests that such a factor analysis would be useful. The communalities were all above 0.3, further indicating that each variable shared some common variance with others. The *MSA* for each of the significantly correlated variables is shown in Table 3. Only variables with $MSA > 0.6$ were used as input variables for PCA.

PCA was performed using **R**, a language and environment for statistical computing and graphics (version 3.2.1), and the "FactoMineR" package (version 1.31.3) [40]. Quality and like ratings were considered as supplementary quantitative variables, meaning that they were not used as inputs for the calculation of principal components, only that they were included in the output data and compared against the components (see Fig. 5a).

In order to determine the number of components to retain from the analysis, a typical approach is to inspect the scree plot and determine the "knee" in the curve. A number of non-graphical methods of making this determination are implemented in the "nFactors" package (version 2.3.3) [41]. The output, shown in Fig. 4, suggests two principal components be kept. This decision was based on the agreement between the results of three of the four methods. As all variables were significantly correlated with at least one of these two principal components, there was no reason to exclude any variables at this stage.

From Fig. 5a it can be seen that the first principal component (*dim. 1*) represents variables associated with amplitude features, such as crest factor, loudness, PMF kurtosis, and all spectral flux bands. The second principal component (*dim. 2*) describes high-frequency spectral features, such as *rolloff85* and *rolloff95*, along with the highest bands of spectral flux, all related to the positive values. The projection of

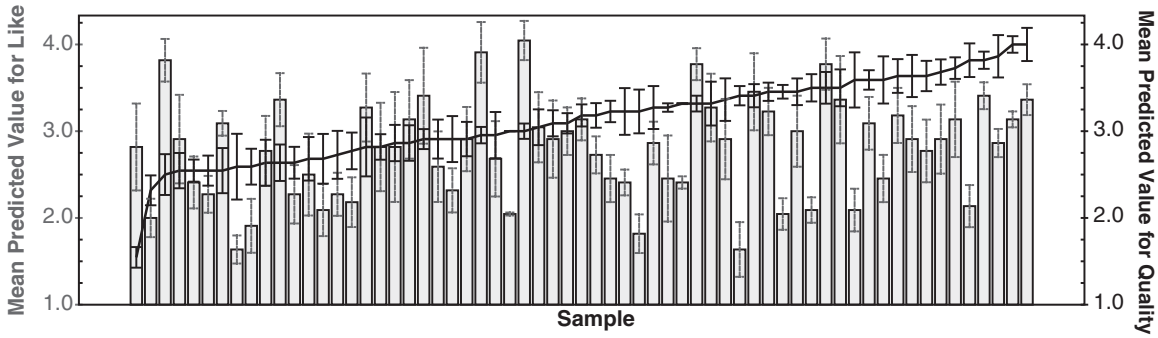
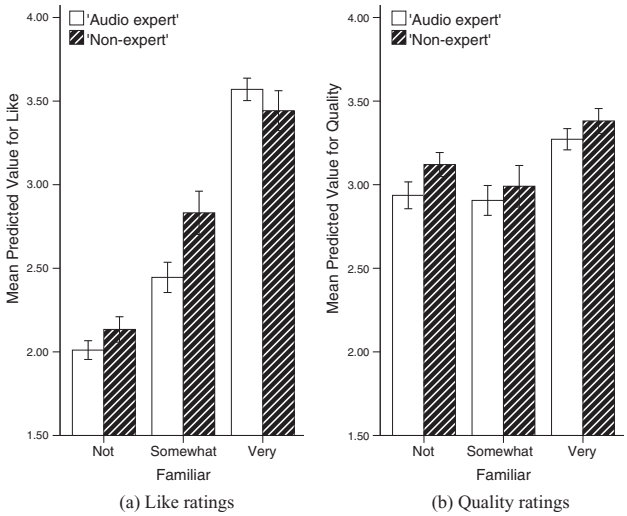


Fig. 2. Average like (bar plot) and quality (line plot) ratings for each sample, with 95% confidence intervals.



(a) Like ratings (b) Quality ratings

Fig. 3. Mean and 95% confidence interval for like and quality ratings over each familiarity rating and expertise group.

quality along the negative direction of *dim. 1* indicates that higher ratings were associated with recordings with greater dynamic range, such as high crest factor or PMF kurtosis. Quality is also projected along the positive axis of *dim. 2*, although its loading on this dimension is comparatively low.

Like ratings show no noteworthy correlation with *dim. 1*, indicating that amplitude-based features do not appear to play a strong part in listener hedonic preference. There was however, a preference for less treble frequencies indicated by the low values of rolloff features. This negative correlation to rolloff (as shown in Table 3) supports the relation between like ratings and a peak in mid-range frequencies, or a simple disliking of samples with too great an emphasis on high-frequencies, also seen in other related studies [13]. These results for like are not surprising since the rating of how much a listener likes a song seems to be dependent on aesthetic and musical content and ultimately, familiarity, as will be discussed later.

Table 3. Correlation of features with subjective results. Significant correlations (where $p < 0.05$) are highlighted in bold and considered for PCA. Features with MSA <0.6, marked with an asterix, are not included in the PCA.

		Quality		Like		
Type	Feature	R^2	p	R^2	p	MSA
Amplitude	Crest factor	**.125⇑**	.004	.000		.842
	Loudness[27]	**.160⇓**	.001	.002		.915
	Top1db[37]	**.078⇓**	.028	.000		.833
	Gauss[16]	**.201⇑**	.000	.000		.835
	PMF Kurtosis	**.108⇑**	.009	.000		.646
	PMF Flatness	**.081⇓**	.025	.001		.951
	PMF Spread	**.155⇓**	.002	.002		.837
Spectral	Spectral Centroid	.000		.061		
	Rolloff85[38]	.008		**.137⇓**	.003	.732
	Rolloff95	.039		**.086⇓**	.024	.663
	Harsh[16]	.058		**.201⇑**	.000	.363*
	LF Energy[16]	**.065⇑**	.046	.016		.489*
Spatial	Width-all (all freq.)	.000		.027		
	Width-band (200Hz–10k)	.013		.035		
	Width-low (0–200Hz)	.000		.006		
	Width-mid (200Hz–2kHz)	.037		.047		
	Width-high (2kHz–10kHz)	.008		.028		
Rhyth.	Tempo	.000		.037		
	Event density	.000		.005		
	Pulse clarity	.021		.005		
Emo. Factors [31]	RMS	**.166⇓**	.001	.004		.829
	Max. summarized fluctuation	**.065⇑**	.045	**.079⇓**	.027	.493*
	Spectral spread	**.143⇑**	.002	**.076⇓**	.030	.804
	Avg. HCDF	.001		**.068⇓**	.040	.471*
	Roughness	**.289⇓**	.000	.036		.826
	Std.dev. roughness	**.153⇓**	.002	.006		.812
Spectral Flux [35]	Band 1 (<50Hz)	**.067⇓**	.043	.014		.858
	Band 2 (50–100 Hz)	.053		.002		
	Band 3 (100–200 Hz)	**.221⇓**	.000	.024		.910
	Band 4 (200–400 Hz)	**.132⇓**		.023		.844
	Band 5 (400–800 Hz)	**.153⇓**		.013		.884
	Band 6 (800–1600 Hz)	**.222⇓**	.000	.009		.900
	Band 7 (1.6–3.2 kHz)	**.277⇓**	.000	.049		.938
	Band 8 (3.2–6.4 kHz)	**.274⇓**	.000	.038		.851
	Band 9 (6.4–12.8 kHz)	**.179⇓**		.003		.886
	Band 10 (12.8–22.05 kHz)	**.071⇓**		.031		.831

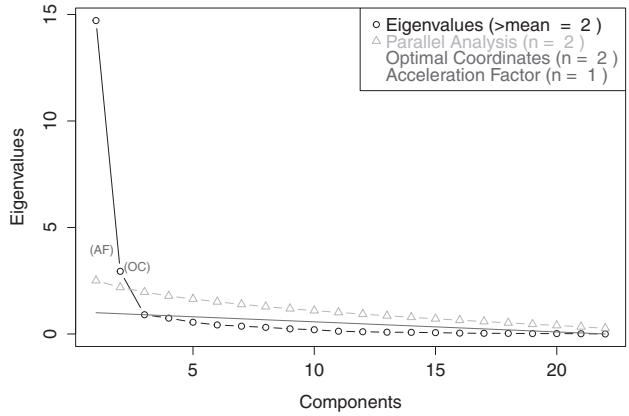
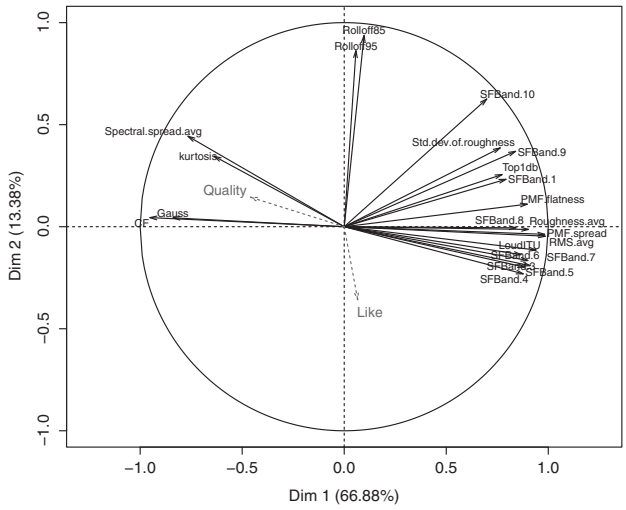
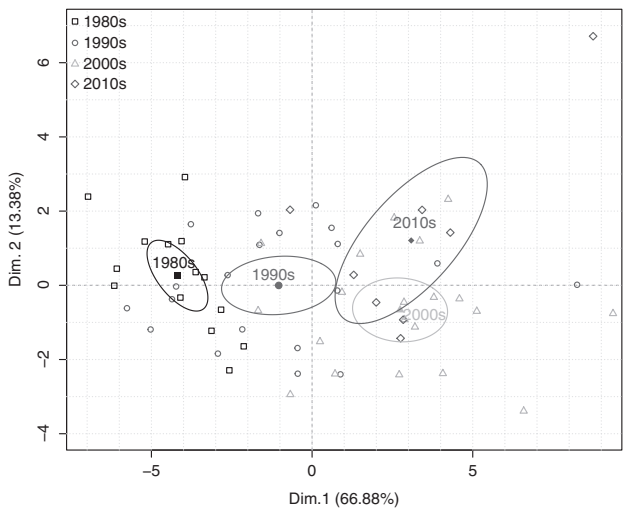


Fig. 4. Scree plot with non-graphical solutions indicating two components be retained. These first two components account for 80.2% of the total variance of the input.



(a) Correlation circle, showing components 1 and 2. *Dim. 1* can be explained by amplitude-based features and *dim. 2* by mostly spectral features.



(b) Individual samples plotted in PCA space, grouped by decade of release. The centroid of each group is marked by solid markers and the ellipses represent regions of 95% confidence in the population centroid of that group.

Fig. 5. Results of Principal Component Analysis, with variables factor map (a) and individuals factor map (b).

Table 4. Correlation of subjective response variables to principal components (Value shown is R^2. Significant correlations highlighted in **bold**).

	Dim. 1	Dim. 2
Like	.004	**.129**
Quality	**.212**	.021

Table 4 shows the R^2 values of linear fits of both quality and like ratings to the dimensions of the principal component analysis. From this it can be seen that quality is significantly and negatively correlated to *dim. 1* ($R^2 = 0.212$) but not *dim. 2* ($R^2 = 0.021$), and that like is significantly, but negatively, correlated to *dim. 2* ($R^2 = 0.129$) but not *dim. 1* ($R^2 = 0.004$).

Fig. 5b shows the 63 audio samples plotted against the first two principal components. As the release year of each sample is known, the samples can be grouped by decade. The group centroid and 95% confidence ellipses for the population centroid are shown for the four categories of 1982–1989, 1990–1999, 2000–2009, and 2010–2013. The data shows that, even with relatively few audio samples per decade, there is an observable difference between the centroid of the 1980s, 1990s, and 2000s categories along the first dimension. Due to the smaller size of the 2010s category, the confidence ellipse is relatively large.

It should be noted that the use of the decade of release as a discrete qualitative variable is not without problems. Release date, as a variable, is effectively continuous and so one would expect to find little difference between 1989 and 1990 but a noticeable change from 1980 and 1999. Consequently, we see that the four decade categories in this study would not be easily separable in a multi-dimensional feature space, implying an upper limit to the success of decade-prediction tasks [37].

The location of each decade centroid on *dim. 1*, which is negatively correlated to quality, increases chronologically. This result suggests that, according to the test panel and their definition, quality seems to have decreased over the decades, mainly due to a change in features associated with dynamic range, as addressed in other studies [10, 42]. This should be considered as an indicative result due to the relatively low number of audio samples and it is important to stress that like ratings were not influenced by this trend.

4.4 Analysis of Quality Descriptions

As shown in Fig. 1b, participants were asked to provide two words to describe the attributes on which quality was assessed for each sample. In total 255 unique words were gathered, after spelling had been corrected and equivalent words collated (such as "compressed" and "over-compressed" or "exciting" and "excited"). As there were some blank entries the total number of instances is slightly less than the full complement of $2 \times 22 \times 63$.

The descriptors were ranked according to the frequency of their usage. To achieve this, a term-frequency matrix was generated using **R** and the text-mining package "tm"

Table 5. Frequency count (Chi square test analysis) of 20 most used words.

	Quality rating					
	1⋆	2⋆	3⋆	4⋆	5⋆	TOTAL
Distorted	**31**>	**43**>	37	**13**<	**2**<	126
Punchy	**1**<	**11**<	37	**63**>	13	125
Clear	**1**<	**4**<	**24**<	**77**>	**18**>	124
Full	0	**4**<	21	**41**>	**21**>	87
Harsh	**15**>	**38**>	23	**9**<	0	85
Wide	3	**5**<	28	**35**>	10	81
Loud	**10**>	18	25	22	4	79
Clean	**0**<	0	**13**<	**36**>	**20**>	69
Fuzzy	7	**28**>	28	**4**<	0	67
Synthetic	**1**<	**18**>	21	20	4	64
Spacious	**1**<	0	20	**30**>	**10**>	61
Thin	6	**21**>	**29**>	**5**<	0	61
Bright	**1**<	9	**26**>	17	7	60
Dull	**8**>	**25**>	20	**7**<	0	60
Deep	**0**<	**4**<	15	**29**>	9	57
Narrow	2	**25**>	23	**6**<	0	56
Smooth	**0**<	**3**<	18	**27**>	7	55
Crunchy	**0**<	10	**23**>	9	2	44
Strong	**0**<	**2**<	10	**21**>	**9**>	42
Aggressive	2	5	8	**18**>	5	38
⋮	⋮	⋮	⋮	⋮	⋮	⋮
TOTAL	197	528	876	856	212	2669

(version 0.6-2) [43]. The top 3 words account for 14% of all instances, while the top 20 account for 54%. In order to determine if there was significant variation in the frequency of each term across the 5 categories of quality rating, a Chi-Square analysis was performed. Only the top 20 words are shown in Table 5, although all 255 words were used to calculate the expected values.

The words chosen to describe the quality of each discrete quality rating differed significantly ($\chi^2(76, N = 1441) = 2131.26$, $p = <.001$). This data provides evidence that can be used to answer research question 5 from Sec. 2. In Table 5, frequencies highlighted in bold (with ">" or "<") are either significantly greater than (>) or less than (<) the expected counts. Further discussion is presented in Sec. 5. Additional analyses, beyond the scope of this paper, can be found in other publications [10, 44].

5 DISCUSSION

These results are now discussed in light of our initial hypotheses as listed in Sec. 2. Results indicate that the samples used in this test elicit different ratings and that, overall, the effect of sample is the largest contributor to the variance found in the subjective ratings, shown in Table 1, where $\eta_p^2 = 0.227$. The effect size of the audio sample is large ($\eta^2 = 0.201$) for quality and medium ($\eta^2 = 0.127$) for like. This confirms that the corpus of audio samples used was successful in triggering significant perceptual variation in ratings from the participants for both concepts.

There appears to be a stronger correlation between quality ratings and the objective features extracted from the signal than that found for like ratings (see Table 4). This suggests the former is a more reliable concept for the subjective evaluation of technical quality, related to modifications of the signal and distinct from hedonic perception. A meaningful correlation was found between like and quality ratings ($R^2 = 0.26$) using raw results pertaining to individual ratings of songs. This however, became non-significant when values were averaged over all participants ($R^2 = 0.02$), removing inter-subject variation. If the two concepts of like and quality are plotted in the space resulting from reducing signal features to a two dimensional space (Fig. 5a), they are nearly orthogonal, further supporting the idea that there is low correlation between them. Each concept is found to describe a different percept in the minds of listeners, where quality refers to technical aspects of the recording and production and like refers to hedonic perception that might be rooted in the musical style/genre or the actual song content itself. This is perhaps the most insightful finding in this study, that quality and like ratings can be considered as two percepts, explained by different factors. Participants elected their own definitions of quality in the experiment by justifying their ratings.

5.1 Effects of Expertise

While expert listeners, on average, provided slightly lower quality ratings than non-experts, the effect of expertise is observed to be small for both quality ($\eta^2 = 0.004$) and like ($\eta^2 = 0.002$). It appears that expertise is not a key factor in the appraisal of either technical quality or hedonic preference, under the conditions investigated here, although, in a study from the authors that further investigates this aspect, it was observed that experts and non-experts typically used different words to justify their ratings [44].

5.2 Liking and Familiarity

Participants were significantly more likely to award greater ratings of like and quality when they were more familiar with the music. However, it is clear that this effect is greater for like ratings, explaining 18.7% of the variance (see Fig. 3a), whereas for quality ratings it explains only 2.4% of the variance (see Fig. 3b). This relationship between familiarity and hedonic preference could be explained by two factors; one may like a song, subsequently choose to listen to it many times, becoming familiar with it, or one may hear a song many times, become familiar with it and grow to like it. This result suggests a clear differentiation between the concepts of preference (how much someone likes a song) and (technical) quality (how well a song has been produced), since familiarity does not seem to play a strong part in the latter.

5.3 Predictive Power of Signal Features

Objective features extracted from the signal were reduced to two components: component 1 mainly describing aspects of amplitude and explaining 67% of the variance in the features considered, while component 2 describes aspects of the spectral content and explains 13% of the variance. Significant correlations were found between features and the subjective response variables (see Table 3 and 4).

Perceived quality is significantly correlated to amplitude features. Samples with higher dynamic range seem to elicit higher ratings of quality, while those with higher loudness seem to be associated with lower ratings. Recall that all samples have been presented at a normalized loudness level, thus effectively removing the differences in loudness but retaining the effect of reduced dynamic range that often ensues from production techniques to maximize loudness. This can explain why "louder" samples are perceived as lower quality in this context.

Measures of spectral flux and some of the underlying features in the MIRtoolbox used to develop emotional predictions are also found to be correlated to quality. Metrics for spectral content do not appear to have a significant effect on quality ratings.

Like ratings do not seem to be affected by amplitude features. As the presentation of audio to participants was normalized according to perceived loudness, as in modern on-line music streaming services such as Spotify and iTunes Radio, these results suggest instead that the effects of dynamic range compression arising from efforts to increase loudness do not appear to affect hedonic perception despite their degrading effects on perceived audio quality.

Like ratings appear to be correlated to spectral features although the strength of the correlation is about half of that observed between quality and component 1 (see Table 4). This low correlation suggests that ratings of like are more strongly affected by a listener's familiarity with a song than with objective features describing it.

These results further reinforce the idea that like and quality are separate aspects of an overall "preference" paradigm. When one simply asks participants for one of these concepts, like or quality, the result may be colored by the participants impression of the other, which is not asked for, a phenomenon known as "dumping bias" [45].

5.4 Attributes Describing Quality

Table 5 shows the 20 most used quality descriptors. These terms describe sound by perceptual timbre, defects, space, and other descriptions. These categories of sound attributes were also found in a solely lexical study [25].

The most commonly used term was "distorted." This indicated that distortion is frequently associated with quality—it was shown that the word is used more than expected for low quality and less than expected for high quality. The term "clean" is never used to describe any rating lower than 3★, indicating the importance of "cleanliness" on the perception of high quality.

"Punchy" and "clear" are the next two most used terms. This result validates the importance of punch and clarity in assessment of audio quality and recent attempts to objectively profile these characteristics [46]. Both terms are associated with high quality ratings.

Participants often used words such as "wide," "narrow," "deep," and "spacious" describing quality ratings, yet, no correlation between perceived quality and spatial measures has been determined in this study (see Table 3). This suggests a need for further work into the extraction of spatial measures of stereo signals that correlate to perceptual attributes, particularly in the case of headphone reproduction as was used here.

There are also examples of the ambiguity that can arise when participants are free to define quality on their own terms. While the term "harsh" is associated with low quality ratings this could simply be due to connotations of the term itself, as there may not be many cases where harshness is a desirable characteristic. Similarly, "dull" may mean "not bright" or "boring/uneventful."

In summary, music samples described by higher quality ratings were typically referred to by terms such as punchy, clear, full, and clean, while they were not likely to be referred to as distorted, harsh, thin, or dull. Further work is presented in [44].

5.5 Insight into Music Production Trends

The sample that scored the lowest mean rating for quality (see Fig. 2) was taken from an album whose perceived audio quality received negative attention in mainstream media at the time of release [47]. Participants were possibly aware of this criticism and therefore open to bias.

As shown in Fig. 5b, there is a difference in the mean value of *dim. 1* for samples from each decade between the 1980s and 2000s. While the "loudness war" has been well-documented [9–11, 42] and has been observed by plotting individual amplitude-based variables over time, one can now see that the effect is visible on a factor level in a feature reduced space. The samples from the 1980s display more variation across *dim. 2* than *dim. 1,* i.e., more variation in spectrum/timbre than loudness/compression. There is a greater range of loudness/compression in the 2000s since it is then possible to make louder but more compressed productions, while some content producers still choose to create dynamic productions. The greatest variation in loudness/compression in one decade is during the 1990s. This particularly significant period of the "loudness war" has been previously referred to as a "loudness race" [10]. Future studies may wish to concentrate on this specific period of time.

6 CONCLUSIONS

The study described in this paper has been an investigation into the perception of quality in music productions. It was found that ratings of quality varied for different musical samples and these ratings were found to correlate to objective variables. The results indicate a difference in the way like and quality concepts were rated. Analysis using PCA indicated that quality ratings were significantly correlated with measures of signal amplitude, loudness, dynamic-range-compression, while like ratings were, on average, not affected by these parameters but instead correlated, less strongly, to measures of signal spectrum.

Like ratings were, however, strongly influenced by song familiarity, implying instead that aspects of preference and liking are distinct from the interpretation of quality and might not be the best descriptors for studies where technical quality is the percept being sought.

The expertise of listeners, although significant, had a weak effect on the ratings of quality and like, suggesting, somewhat counter-intuitively, that a participant's expertise is not a strong factor in assessing audio quality or musical preference (see Figs. 3a and 3b).

It has been observed that the words used to describe sonic attributes of the audio signal on which quality was assessed were typically those words that describe perceived timbre, space, and defects. The frequency of word usage varied significantly depending on the rating being awarded, with words such as 'clean' and 'full' strongly associated with high ratings of quality, while 'distorted' and 'harsh' were associated with low ratings.

In summary, quality in music production is revealed as a perceptual construct distinct from hedonic, musical preference, which is more likely influenced by familiarity with the song. Audio quality can be predicted from objective features in the signal and be adequately and consensually described using verbal attributes. The work presented has implications in the music industry, particularly if issues such as the "loudness war" are being rendered moot by new loudness normalized broadcast standards.

7 ACKNOWLEDGMENTS

We wish to thank Trevor Cox, Paul Kendrick, and Jamie Angus at the University of Salford for their comments on an earlier version of this paper, as well as the anonymous reviewers for their detailed feedback.

8 REFERENCES

[1] U. 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 (2004).

[2] ITU-R BS.1116-1, "Methods for the Subjective Assessment of Small Impairments in Audio Systems including Multichannel Sound Systems," Tech. Rep., International Telecommunications Union (1997).

[3] J. Liebetrau, F. Nagel, N. Zacharov, K. Watanabe, C. Colomes, P. Crum, T. Sporer, and A. Mason, "Revision of Rec. ITU-R BS.1534," presented at the 137th Convention of the Audio Engineering Society (2014 Oct), convention paper 9172.

[4] N. Croghan, K. Arehart, and J. Kates, "Quality and Loudness Judgments for Music Subjected to Compression Limiting," *J. Ac. Soc. of Am.*, vol. 132, no. 2, pp. 1177–1188 (2012 Aug.).

[5] C. Tan, B. Moore, and N. Zacharov, "The Effect of Nonlinear Distortion on the Perceived Quality of Music and Speech Signals," *J. Audio Eng. Soc.*, vol. 51, pp. 1012–1031 (2003 Nov.).

[6] C. Tan, B. Moore, N. Zacharov, and V. Mattila, "Predicting the Perceived Quality of Nonlinearly Distorted Music and Speech Signals," *J. Audio Eng. Soc.*, vol. 52, pp. 699–711 (2004 Jul./Aug.).

[7] B. Moore, C. Tan, N. Zacharov, and V. Mattila, "Measuring and Predicting the Perceived Quality of Music and Speech Subjected to Combined Linear and Nonlinear Distortion," *J. Audio Eng. Soc.*, vol. 52, pp. 1228–1244 (2004 Dec.).

[8] P. Kendrick, F. Li, B. Fazenda, I. Jackson, and T. Cox, "Perceived Audio Quality of Sounds Degraded by Nonlinear Distortions and Single-Ended Assessment Using HASQI," *J. Audio Eng. Soc*, vol. 63, pp. 698–712 (2015 Sep.).

[9] E. Deruty and D. Tardieu, "About Dynamic Processing in Mainstream Music," *J. Audio Eng. Soc.*, vol. 62, pp. 42–55 (2014 Jan./Feb.), http://dx.doi.org/10.17743/jaes.2014.0001.

[10] A. Wilson and B. Fazenda, "Characterisation of Distortion Profiles in Relation to Audio Quality," *Proc. of the 17th Int. Conference on Digital Audio Effects (DAFx-14)*, Erlangen, Germany (2014), pp. 1–8.

[11] E. Deruty and F. Pachet, "The MIR Perspective on the Evolution of Dynamics in Mainstream Music," ISMIR, Malaga, Spain (2015 Oct.).

[12] M. Schoeffler and J. Herre, "About the Impact of Audio Quality on Overall Listening Experience," *Proceedings of the Sound and Music Computing Conference 2013*, Stockholm, Sweden (2013), pp. 48–53.

[13] A. Wilson and B. Fazenda, "101 Mixes: A Statistical Analysis of Mix-Variation in a Dataset of Multitrack Music Mixes," presented at the *139th Convention of the Audio Engineering Society* (2015 Oct.), convention paper 9398.

[14] B. De Man, M. Boerum, B. Leonard, R. King, G. Massenburg, and J. Reiss, "Perceptual Evaluation of Music Mixing Practices," presented at the *138th Convention of the Audio Engineering Society* (2015 May), convention paper 9235.

[15] E. Deruty, F. Pachet, and P. Roy, "Human Made Rock Mixes Feature Tight Relations between Spectrum and Loudness," *J. Audio Eng. Soc*, vol. 62, pp. 643–653 (2014 Oct.).

[16] A. Wilson and B. Fazenda, "Perception and Evaluation of Audio Quality in Music Production," Proc. of the 16th Int. Conference on *Digital Audio Effects (DAFx-13)*, Maynooth, Ireland (2013), pp. 1–6.

[17] V. Salimpoor, M. Benovoy, G. Longo, J. Cooperstock, and R. Zatorre, "The Rewarding Aspects of Music Listening Are Related to Degree of Emotional Arousal," *PloS one*, vol. 4, no. 10, pp. e7487 (2009 Jan.), http://dx.doi.org/10.1371/journal.pone.0007487.

[18] C. Pereira, J. Teixeira, P. Figueiredo, J. Xavier, S. Castro, and E. Brattico, "Music and Emotions in the Brain: Familiarity Matters," *PloS one*, vol. 6, no. 11, pp. e27241 (2011 Jan.), http://dx.doi.org/10.1371/journal.pone.0027241.

[19] I. Peretz, D. Gaudreau, and A. Bonnel, "Exposure Effects on Music Preference and Recognition," *Memory & Cognition*, vol. 26, no. 5, pp. 884–902 (1998), http://dx.doi.org/10.3758/BF03201171.

[20] K. Szpunar, E. Schellenberg, and P. Pliner, "Liking and Memory for Musical Stimuli as a Function of Exposure," *J. Experimental Psychology: Learning, Memory, and Cognition*, vol. 30, no. 2, pp. 370–381 (2004 Mar.), http://dx.doi.org/10.1037/0278-7393.30.2.370.

[21] P. Hunter and E. Schellenberg, "Interactive Effects of Personality and Frequency of Exposure on Liking for Music," *Personality and Individual Differences*, vol. 50, no. 2, pp. 175–179 (2011), http://dx.doi.org/10.1016/j.paid.2010.09.021.

[22] S. Olive, "Differences in Performance and Preference of Trained versus Untrained Listeners in Loudspeaker Tests: A Case Study," presented at the *114th Convention of the Audio Engineering Society* (2003 Mar.), convention paper 5728.

[23] I. Dror, D. Charlton, and A. Péron, "Contextual Information Renders Experts Vulnerable to Making Erroneous Identifications," *Forensic Science Int.*, vol. 156, no. 1, pp. 74–78 (2006 Jan.), http://dx.doi.org/10.1016/j.forsciint.2005.10.017.

[24] I. Dror and R. Rosenthal, "Meta-Analytically Quantifying the Reliability and Biasability of Forensic Experts," *J. Forensic Sciences*, vol. 53, no. 4, pp. 900–903 (2008 July), http://dx.doi.org/10.1111/j.1556-4029.2008.00762.x.

[25] S. Le Bagousse, M. Paquier, and C. Colomes, "Categorization of Sound Attributes for Audio Quality Assessment—A Lexical Study," *J. Audio Eng. Soc*, vol. 62, pp. 736–747 (2014 Nov.).

[26] P. Pestana, Z. Ma, and J. Reiss, "Spectral Characteristics of Popular Commercial Recordings 1950–2010," presented at the *135th Convention of the Audio Engineering Society* (2013 Oct.), convention paper 8960.

[27] ITU-R BS.1770-3, "Algorithms to Measure Audio Programme Loudness and True-Peak Audio Level," Tech. Rep., International Telecommunications Union (2012).

[28] P. Rentfrow, L. Goldberg, and D. Levitin, "The Structure of Musical Preferences: A Five-Factor Model," *J. Personality and Social Psychology*, vol. 100, no. 6, pp. 1139–1157 (2011), http://dx.doi.org/10.1037/a0022406.The.

[29] R. Schatz, S. Egger, and K. Masuch, "The Impact of Test Duration on User Fatigue and Reliability of Subjective Quality Ratings," *J. Audio Eng. Soc.*, vol. 60, pp. 63–73 (2012 Jan./Feb.).

[30] O. Lartillot and P. Toiviainen, "A Matlab Toolbox for Musical Feature Extraction from Audio," International Conference on *Digital Audio Effects (DAFx-07)* (2007), pp. 1–8.

[31] T. Eerola, O. Lartillot, and P. Toiviainen, "Prediction of Multidimensional Emotional Ratings in Music from Audio Using Multivariate Regression Models," *ISMIR*, pp. 621–626 (2009).

[32] S. Beveridge and D. Knox, "A Feature Survey for Emotion Classification of Western Popular Music," 9th International Symposium on Computer Music Modeling and Retrieval, *CMMR2012* (2012), pp. 19–22.

[33] T. Eerola, "Are the Emotions Expressed in Music Genre-Specific? An Audio-Based Evaluation of Datasets Spanning Classical, Film, Pop, and Mixed Genres," *J. New Music Res.*, vol. 40, no. 4, pp. 349–366 (2011 Dec.), http://dx.doi.org/10.1080/09298215.2011.602195.

[34] G. Tzanetakis, R. Jones, and K. McNally, "Stereo Panning Features for Classifying Recording Production Style," *ISMIR* (2007).

[35] V. Alluri and P. Toiviainen, "Exploring Perceptual and Acoustical Correlates of Polyphonic Timbre," *Music Perception*, vol. 27, no. 3, pp. 223–242 (2010), http://dx.doi.org/10.1525/mp.2010.27.3.223.

[36] T. Levine and C. Hullett, "Eta Squared, Partial Eta Squared, and Misreporting of Effect Size in Communication Research," *Human Communication Research*, vol. 28, no. 4, pp. 612–625 (2002).

[37] D. Tardieu, E. Deruty, C. Charbuillet, and G. Peeters, "Production Effect: Audio Features for Recording Techniques Description and Decade Prediction," in *Proc. of the 14th Int. Conference on Digital Audio Effects (DAFx-11)*, Paris, France (2011).

[38] G. Tzanetakis and P. Cook, "Musical Genre Classification of Audio Signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302 (2002), http://dx.doi.org/10.1109/TSA.2002.800560.

[39] G. Hutcheson and N. Sofroniou, *The Multivariate Social Scientist: Introductory Statistics Using Generalized Linear Models* (Sage, 1999).

[40] S. Lê, J. Josse, and F. Husson, "FactoMineR: An R Package for Multivariate Analysis," *J. Statistical Software*, vol. 25, no. 1, pp. 1–18 (2008).

[41] G. Raîche, T. Walls, D. Magis, M. Riopel, and J. Blais, "Non-Graphical Solutions for Cattells Scree Test," *Methodology: European J. Research Methods for the Behavioral and Social Sciences*, vol. 9, no. 1, pp. 23 (2013).

[42] E. Vickers, "The Loudness War: Background, Speculation, and Recommendations," presented at the *129th Convention of the Audio Engineering Society* (2010 Nov.), convention paper 8175.

[43] D. Meyer, K. Hornik, and I. Feinerer, "Text Mining Infrastructure in R," *J. Statistical Software*, vol. 25, no. 5, pp. 1–54 (2008).

[44] A. Wilson and B. Fazenda, "A Lexicon of Audio Quality," Proc. 9th Triennial Conference of the *European Society for the Cognitive Sciences of Music (ESCOM 2015)*, Manchester, UK (2015 Aug.).

[45] S. Bech and N. Zacharov, *Perceptual Audio Evaluation: Theory, Method and Application* (John Wiley & Sons, Chichester, West Sussex, UK, 2006).

[46] S. Fenton and H. Lee, "Towards a Perceptual Model of Punch in Musical Signals," presented at the *139th Convention of the Audio Engineering Society* (2015 Oct.), convention paper 9381.

[47] E. Smith, "Even Heavy-Metal Fans Complain that Today's Music Is Too Loud," *Wall Street Journal*, September 2008, accessed: 18 March 2014.

A.1 LIST OF AUDIO DESCRIPTORS PROVIDED TO PARTICIPANTS

Bright, dark, loud, quiet, mellow, clear, clean, punchy, dull, bland, dense, exciting, weak, strong, sweet, shiny, fuzzy, wet, dry, distorted, realistic, spacious, narrow, wide, deep, shallow, aggressive, light, gentle, cold, hard, synthetic, crunchy, hot, rough, harsh, smooth, thin, full, airy, big.

THE AUTHORS

Alex Wilson

Bruno Fazenda

Alex Wilson is currently a Ph.D. student at the University of Salford, investigating the perception of quality in sound recordings, focussing on music productions. He obtained a B.Sc. in experimental physics from NUI Maynooth in 2008 and a B.Eng. in audio technology from University of Salford in 2013, which included a year of industrial experience in the area of studio monitor R&D. He maintains interests in digital audio processing, psychoacoustics, and the art of record production.

Bruno Fazenda is a senior lecturer and researcher at the Acoustics Research Centre, University of Salford. His research interests span room acoustics, sound reproduction, and psychoacoustics, in particular, the assessment of how an acoustic environment, technology or psychological state impacts on perception of sound quality. He is a researcher in a number of research council funded projects. He is also a keen student on aspects of human evolution, perception, and brain function.