PERCEPTION & EVALUATION OF AUDIO QUALITY IN COMMERCIAL MUSIC PRODUCTION

AD Wilson The University Of Salford, Salford, UK
BM Fazenda The University Of Salford, Salford, UK

ABSTRACT:

Audio quality is still an elusive concept in published music and broadcast. Most attentive listeners are able to distinguish good audio from bad. However, a conclusive correlation between this subjective assessment and objective factors measured directly from the signal has not been found. In this work, a dataset of audio clips was prepared and quality assessed by means of controlled subjective testing. Encoded as digital signals, a large amount of feature extraction on the audio data was carried out and correlations between these and their subjective assessments were obtained. Results show that subjective responses may indeed be correlated to some objective metrics in the audio signal. In particular, spectral and dynamic measures as well as distortion and tempo were found to be significantly correlated. The results of these analyses informed attempts at quality estimation, which show promising results.

1 INTRODUCTION

A single consistent definition for quality has not yet been offered, however, for certain restricted circumstances, `quality' has an understood meaning when applied to audio. Measurement techniques exist for the assessment of audio quality[1,2], however such standards typically apply to the measure of quality with reference to a golden sample; what is in fact being ascertained is the reduction in perceived quality due to destructive processes, such as the effects of the reproduction equipment being used. This applies well to systems of encoding, in which the audio being evaluated is a compressed version of the reference and the deterioration in quality is measured[3].

Such descriptions would not strictly apply to the evaluation of quality in musical recordings. This study is concerned with the audio quality of `produced' commercial music where there is no fixed reference and quality is evaluated by comparison with all other samples heard. This judgement is thought to be based on both subjective and objective parameters.

In systems where objective measurement is possible, there is still disagreement regarding which parameters contribute to quality and the manner of their contribution. With loudspeakers, even a measure as trivial as the on-axis amplitude response does not have a simple relationship to quality and there is evidence of many secondary factors influencing listener preference, even based on cultural context[4].

If quality is highly subjective the aspects of the subject which are of influence can be investigated. An expert group of music professionals was able to distinguish various recording media from one another, based on their subjective audio quality ratings of classical music recordings[5]. While a CD and cassette displayed a significant measurable difference, formats of higher fidelity than CD were not rated significantly higher than the CD. The expertise of the listener is therefore thought to be a factor in quality-perception, due to ability to detect technical flaws.

Additionally, social factors are thought to influence quality-perception, especially in music. For example, it has been suggested that some dislike of modern music, attributed to production methods, is in-fact misplaced nostalgia effects; “People tend to bond closely with the music they

heard in their pre-teen and teenage years. As society ages, listeners may blame hyper-compression for a loss of musical interest that may result from other factors such as changes in musical styles, age-related hearing loss and various lifestyle changes." [6]

In summation, audio quality, as applied to music recordings, is predicted to be based on both subjective and objective measures. This paper will investigate both aspects using independent methodologies, and subsequently attempt to determine what correlations exist and how the subjective and objective evaluations can be linked, to lead towards a quality-prediction model.

The following research questions are investigated.

- What is meant by `audio quality' in the context of music production?
- How much of what influences the perception of quality is objective and how much is subjective?
- What influencing parameters can be determined?

2 METHODOLOGY

The hypotheses under test are as follows:

1. There are noticeable differences in quality between samples
2. Listener training has an influence on perception of quality
3. Familiarity with a sample is related to how much it is liked
4. Quality is related to one or more objective signal parameters

2.1 Subjective Testing

To obtain subjective measures of the audio signals a listening test was designed in which subjects listened to a series of audio clips and answered simple questions on their experience. Basic information about the subject was gathered so that results could be analysed based on demographics. The age and sex of each subject was recorded. In addition, subjects were asked to identify themselves as either 'audio expert', 'musician' or 'none of the above'. This last category is used as a control group herein referred to as the naïve group. Subjects in this category would ideally not possess professional knowledge of acoustics or audio engineering and lack any above-average musical ability. Subjects were given a short briefing in order to ensure the questions were understood. For each audition the subject was asked the questions in Table 1, designed to investigate the hypotheses under test.

Question	Answers	Variable
How familiar are you with this song?	Not, Somewhat or Very Familiar	Familiar
Please rate this song?	1 to 5 (5 is highest)	Like
Please rate the Sound Quality of this sample?	1 to 5 (5 is highest)	Quality

Table 1 - Questions to be answered for each audio sample

All audio samples were 16-bit, 44.1kHz stereo PCM files. 55 samples were used, each a 20 second segment of the song around the second chorus (where possible) with a one second fade-in and fade-out. This forecast the test duration at 20-25 minutes. Based on the guidelines of previous studies listener fatigue was considered negligible[7]. The audio selection process was influenced by the test hypotheses. As familiarity was investigated, there needed to be a number that were unfamiliar to all and some familiar to all. To achieve this, six songs by unsigned Irish artists were used. The bulk of the samples used was from 1972 to 2012, with two 1960s samples, and was predominately pop and rock styles. For consistency, all samples contained vocals.

2.2 Description of the listening test

The listening test was delivered using a MATLAB script, which displays text, plays audio and receives user input. Controlled tests took place in the listening room at University Of Salford. The room has been designed for subjective testing and meets the requirements of ITU-R BS. 1116-1, with a background noise level of 5.7dBA[8]. Audio was delivered using Sennheiser HD800 headphones and the order of playback was randomised for each subject. While the test ran on a laptop computer subjects were seated at a displaced monitor and keyboard, minimising distractions as well as reducing fan noise from the computer. Additionally, the user interface was deliberately minimal and lighting was subdued in order to reduce visual influences.

Unlike methods described in Section 1, this experiment contains no reference for what constitutes highest or lowest quality. In this case, what is being tested is that which the subject does naturally, listening to music. In more rigorously-controlled testing with modified stimuli, there is a risk of gathering unnatural responses to unnatural stimuli. In order to test normal listening experiences, each audition was unique and the audio had not been treated in any way, other than loudness equalisation to mimic modern broadcast standards or programs such as Spotify. Loudness levels were calculated using the model described by Glasberg & Moore[9]. These predictions agreed well with in-situ measurements performed using a Brüel & Kjær Head And Torso Simulator (HATS) and sound level meter. For testing, samples were auditioned at an average listening level of 84dBA, measured using the HATS and sound level meter.

Additional subjects were tested in less controlled circumstances, outside of the listening room, using Sennheiser HD 25-1 II headphones. A small number of subjects were tested in both controlled and uncontrolled circumstances and displayed a high level of consistency, permitting further uncontrolled tests, most of which took place in quiet locations at Trinity College Dublin and National University of Ireland, Maynooth.

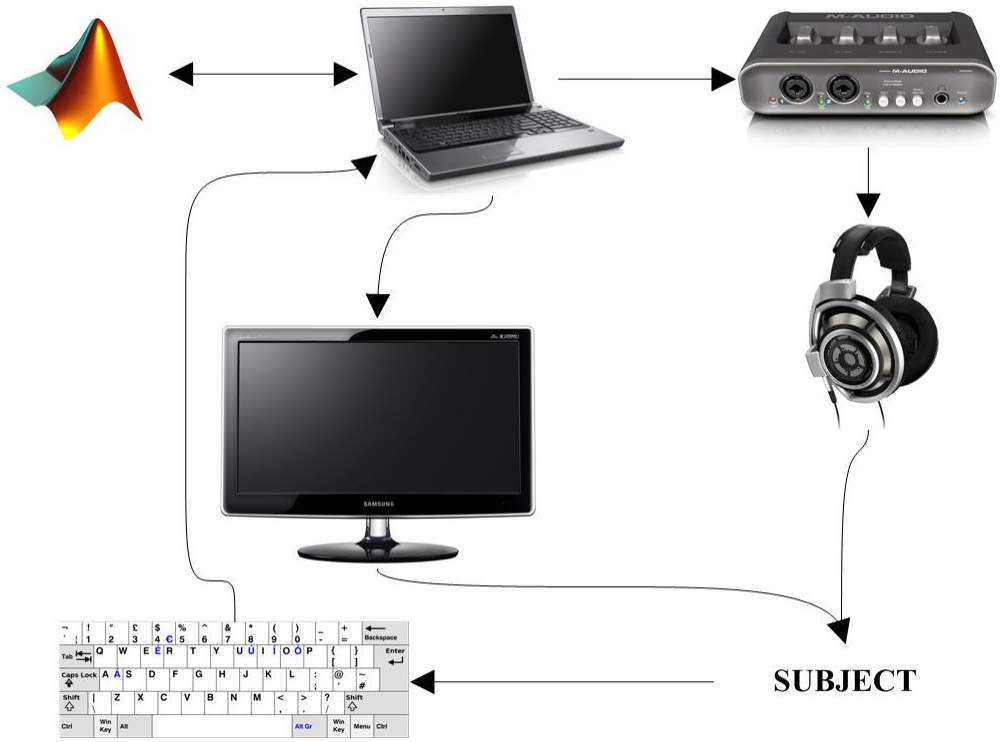

Figure 1 - Diagram of test set-up

2.3 Objective Measures

In order to characterise the audio signals the features in Table 2 were extracted from each sample. Feature-extraction was aided by the use of the MIRtoolbox[10].

Feature	Description
Crest Factor	Ratio of peak amplitude to rms amplitude
Width	1 – (cross-correlation between left and right channels of the stereo signal, at a time offset of zero samples)
Rolloff	Frequency at which 85% of the spectral energy is contained below[11]
Harsh Energy	Proportion of total spectral energy that is contained in the 2kHz to 5kHz band
LF energy	Proportion of total spectral energy that is contained in the 20Hz to 80Hz band
Tempo	Beats per minute
Gauss	See Section 2.3.1
Happy	Prediction of the listener's emotional response[12]
Anger	

Table 2 - Features used in objective analysis

Happiness and Anger were chosen from a set of five classes, including Sadness, Fear and Tenderness. The latter three were rejected due to weak correlation with quality ratings obtained. Rolloff describes the extent of the high frequencies and the overall bandwidth, especially when combined with LF energy, representing the band reproduced by a typical studio subwoofer. While this could extend to 100 or 120Hz in some units, various ranges were compared to quality ratings and the 20-80Hz range was found to be most highly correlated.

Harsh energy was based on the authors experience and the accounts of mix engineers, where this range was said to imbue a 'cheap', 'harsh', 'luxurious' or 'smooth' character, depending on the concentration of energy in this band. Again, various bands were compared to quality ratings and 2k-5kHz was found to be most highly correlated.

2.3.1 A Novel Measure of Audible Distortion

The classic model of the amplitude distribution of digital audio is a modified Gaussian probability mass function (PMF) with zero mean. While the terms 'histogram', 'probability density function' and 'probability mass function' are often used interchangeably in digital audio, the unique distinctions are used here. Most commercially-released music prior to the mid-1990s adheres to this model, particularly when there is sufficient dynamic range and the mix consists of many individual elements. Hard-limiting becomes a feature of the PMF with the onset of the 'loudness war', where the extreme amplitude levels assume higher probabilities, sometimes exceeding the zero-amplitude probability to become the most probable of all levels.

A more recent phenomenon has been the presence of wider, localised peaks in the interim values, (as seen in the Amy Winehouse example in Figure 2) often as a means of avoiding the sonic effects of clipping. This type of amplitude distribution can be caused by a number of issues, such as the mastering of mixes which have already been limited as well as the limiting of individual elements in the mix, such as the drums. It is not uncommon in modern audio productions to use multiple stages of limiting in the mix process to prepare for the limiting it will receive during the mastering process[13].

The PMF of each audio signal was analysed to provide features associated with audible distortion, in particular, why certain recordings are deemed 'unlistenable' by some audiophile reviewers. The effects of dynamic range compression and hard-limiting have been studied in relation to listener preference[14]. Since these parameters are encompassed by the PMF, this study attempts to gather them into a higher-level feature.

The histogram was evaluated using 201 bins, providing a good trade-off between runtime, accuracy and clarity of visualisation. In order to evaluate the shape of the distribution, particularly the slope and the presence of any localised peaks, the first derivate was determined. For the ideal distribution this had a Gaussian form (see Figure 2) so the goodness-of-fit to a Gaussian profile was calculated for each sample. This was used as a feature describing loudness, dynamic range and related audible distortions, referred to as 'Gauss'.

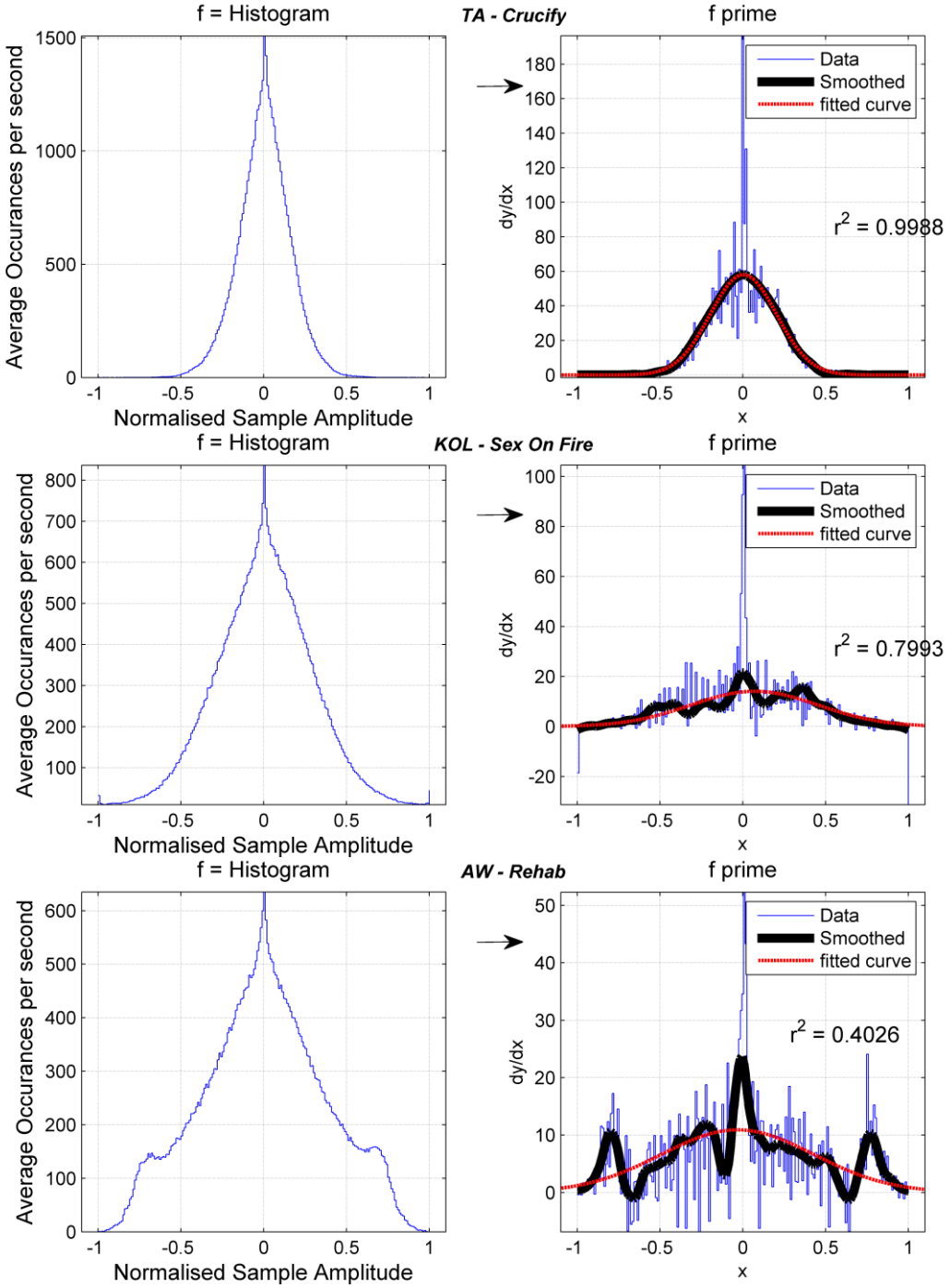

Figure 2 - Examples of Gauss values, derived from audio PMF. Samples are from albums by Tori Amos (1992), Kings Of Leon (2008) and Amy Winehouse (2006).

3 RESULTS

The total number of subjects tested was 24. Of this number 9 were female and 15 were male. Expertise was 12 expert, 5 musician and 7 neither, which can be considered an even split between expert and non-expert. The mean age was 27 years. With 55 audio samples and 24 subjects, 1320 auditions were gathered and analysis was performed on this dataset.

3.1 Subjective Test Results

The results of a 3-way ANOVA, shown in Table 3, show that each main effect is significant ($p < 0.05$), for quality and like, in addition to a significant second-level interaction between expertise and familiarity for quality and two second-level interactions for like (Sample/Familiar and Expertise/Familiar). To investigate further, one-way ANOVA tests were performed with post-hoc multiple comparison and Bonferroni adjustment applied.

Interaction level	Source	Degrees of freedom	Quality		Like	
			F	P	F	P
1	Sample (S)	54	7.78	0.00	7.74	0.00
1	Expertise (E)	2	4.50	0.01	3.95	0.02
1	Familiar (F)	2	17.62	0.00	204.47	0.00
2	S*E	108	0.94	0.65	0.85	0.87
2	S*F	94	1.08	0.29	1.30	0.03
2	E*F	4	3.16	0.01	3.20	0.01
3	S*E*F	106	1.06	0.34	0.95	0.61

Table 3 - Results of 3-way ANOVA

The mean quality ratings for the audio samples ranged from 2.12 to 4.29. The result in Table 3 supports test hypothesis #1, that certain samples are perceived as higher-quality than others.

Mean quality scores were significantly lower for the expert group than the naïve group ($F(1, 2) = 3.42$, $p = 0.03$, see Figure 3). This provides support for test hypothesis #2, that a listener's training has an influence on quality-perception.

Comparing the expert and musician groups shows that the groups agree on quality, with no statistically significant difference between ratings. The expert group were more critical of quality than the naïve group indicating that it is possible that factors such as distortion and dynamic range compression were more easily identified.

The mean time taken to evaluate each 20-second sample varied according to expertise, with the naïve group responding significantly quicker than the other two groups ($F(1, 2) = 12.16$, $p = 0.00$), shown in Figure 2. As their quality ratings were also higher, this indicates that the naïve group was less aware of what to listen for or simply less engaged in the experiment, further supporting hypothesis #2.

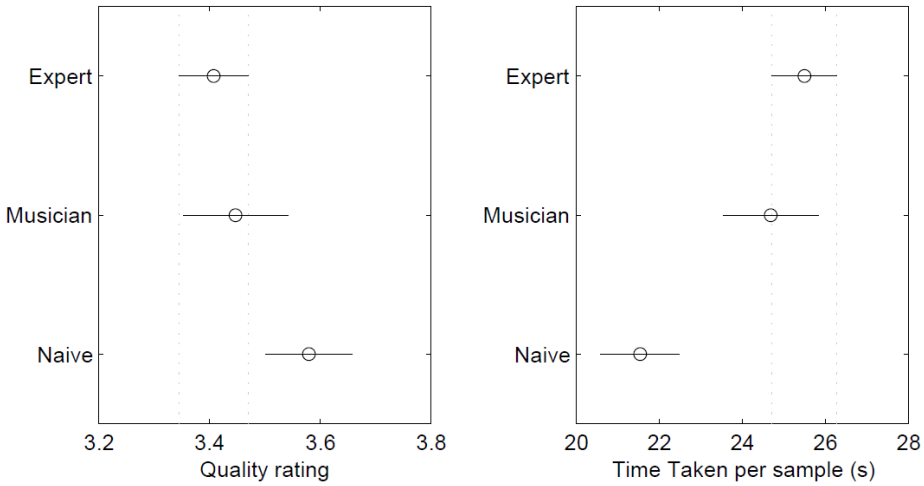

Figure 3 - ANOVA results, for Expertise

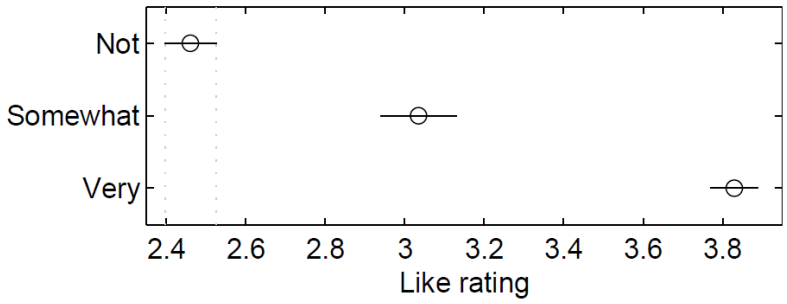

Figure 4 - 1-way ANOVA: Mean Like, grouped by Familiarity

Figure 4 shows that samples which were more familiar were liked more ($F(1, 2) = 283.62$; $p = 0.00$). This is the evidence to support test hypothesis #3 and reflects the idea that one is unlikely to become familiar with a song one does not enjoy listening to, or perhaps that one comes to like a song that is played to them more often. Previous work has suggested that in using commercially successful music there was an automatic assumption of high-quality by listeners[15]. This study supports this view, indicating that, on first listen, one's perception of quality is rather conservative and repeated listens allow quality to be better appreciated and a more realistic appraisal be made.

3.2 Objective Measures Compared To Quality

The expected output of the *miremotion* function is in the range 1 to 7, extending to a likely 0 to 8 range[12]. While happy scores ranged from 1.7 to 7.2, anger scores ranged from 2 to 32. Due to this distribution, the analysis is performed on a logarithmic scale. The Gauss feature values have a range of 0 to 1 but most lie between 0.9000 and 0.9999. To better approximate a linear plot, the data plotted is equal to $-\log_{10}(1 - \text{Gauss})$, which magnifies this upper range, producing a more even distribution of values. Little correlation was found between objective parameters and the like ratings. However, individual features were significantly correlated with subjective quality ratings. This is shown in Figure 5, where each point is the mean quality value for each sample over all subjects, and the trend lines shown are best fit lines ascertained using linear regression. r^2 values range from 0.0831 to 0.3532 and all correlations were found to be significant, with $p < 0.05$ (apart from the sample subset with above-optimal width, see Section 4.5). By these significant correlations, test hypothesis #4 is supported – objective parameters can be correlated to quality ratings.

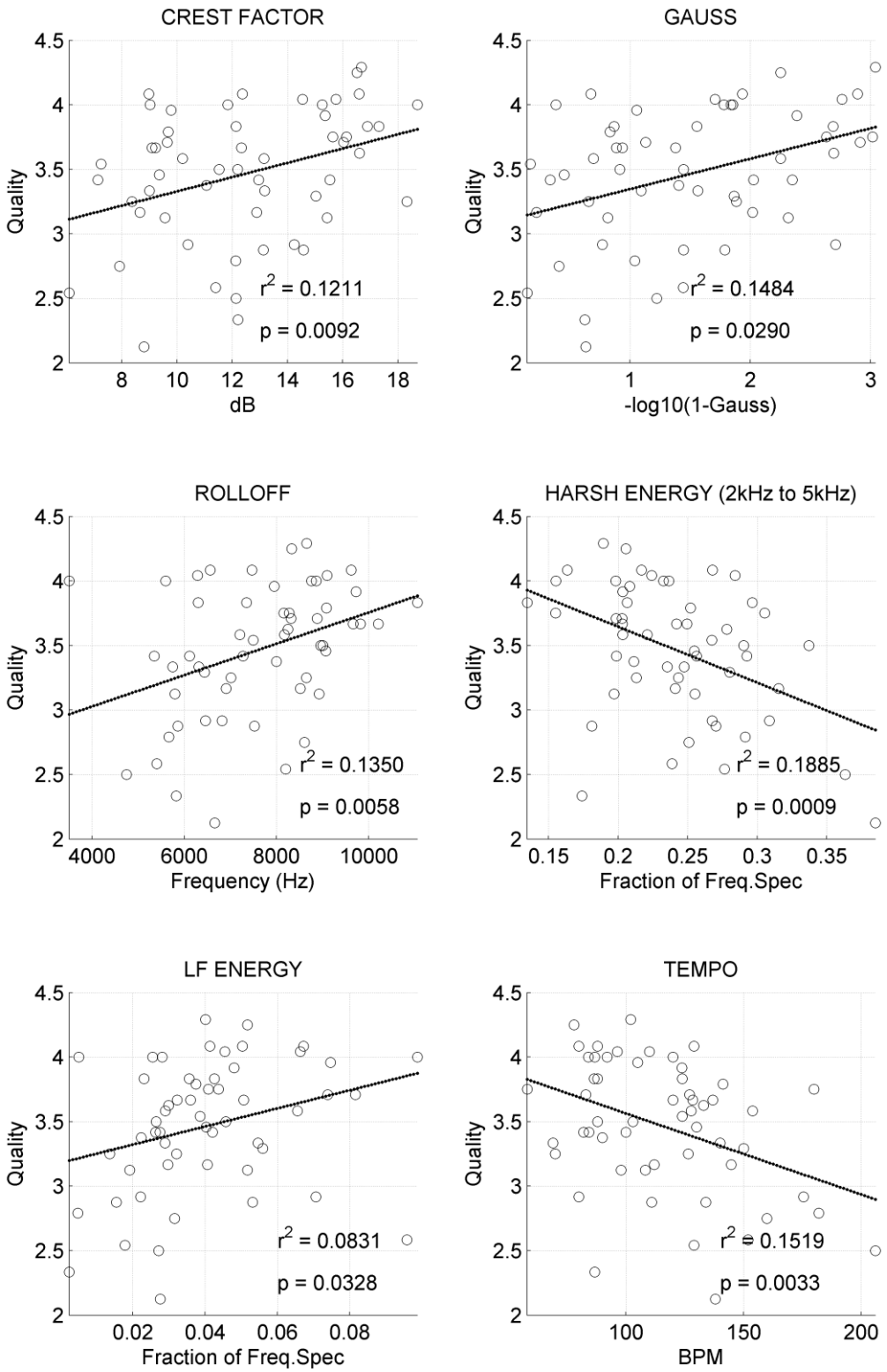

Figure 5a - Objective parameters compared to subjective quality

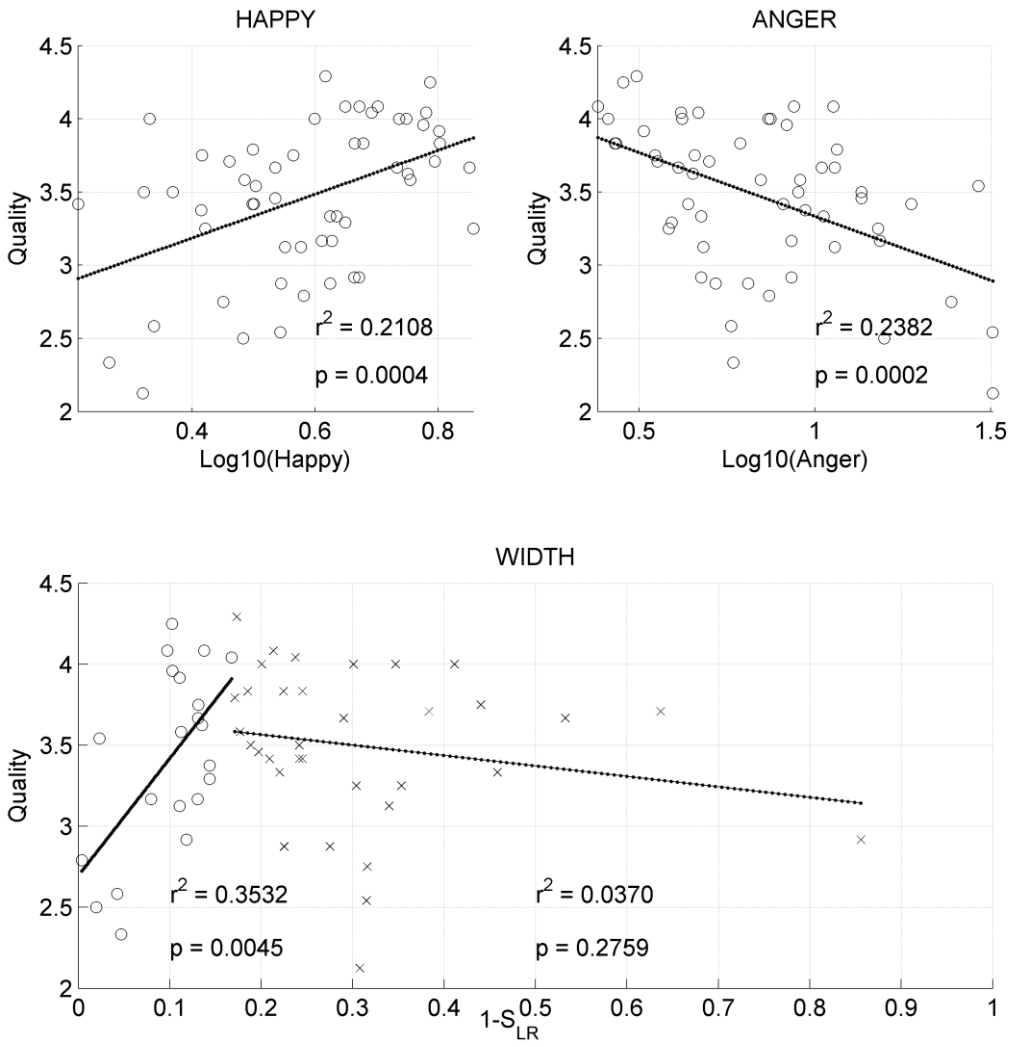

Figure 5b – Objective parameters compared to Subjective Quality

4 DISCUSSION

The most significant correlations are found for emotional features and spectral features. This allows aesthetic considerations to be made for quality. The emotional intent of the artist and the timbre of the instrumentation are important considerations, as well as the choice of tempo. Additionally, these three considerations are influenced by musical genre.

This connection to genre is not altogether unsurprising, as perceived audio quality can also be a genre-defining aesthetic, as in the cases of punk or garage rock as well as lo-fi genres associated with 8-bit computing or circuit-bending.

4.1. Emotional Response Predictions

The *miremotion* features used (happy and anger) were highly correlated with quality, yielding some of the highest r^2 values of the features in Figure 5. This suggests that the perception of quality is linked to the 'emotion class' in which the sample belongs and the listener's emotional reaction to the sample; high quality rating were awarded in instances of high happy coefficient and low quality ratings for high anger coefficient.

Moreover, this result indicates a connection between quality and a unified set of subjective and objective parameters, as the *miremotion* features are objective measures designed to predict subjective responses.

It should be noted that the algorithm used for prediction was originally trained by audio entirely from film soundtracks[12]. It is likely that these samples were classical and electronic styles and recorded with ample dynamic range. The samples used here yield a range of values suggesting a weakness in the assumptions of the original methodology - the predictions may not be reliable for commercial music from a wide timespan.

For example, 'Sober' by Tool (1993) scores 3.9 for anger, compared to 'Teenage Dream' by Katy Perry (2010) scoring 10.5. Additionally, 'Raining Blood' by Slayer (1986) scores similarly to 'With Or Without You' by U2 (1987), rated 4.5 and 4.6 respectively. In this case the similar release time rules out extreme production differences in the previous example. It is suggested that anger shows good correlation with quality due to the features which make up the prediction, although in this case it may not be a good prediction of the actual emotional response of the listener, due to differences in pop/rock music to the original training set of film scores.

4.2. Spectral Features

Quality ratings were higher in cases of high rolloff and high LF energy, relating to wider bandwidth. LF energy also relates to production trends and advancements in technology as the ability to capture and reproduce these low frequencies has improved over time due to a number of factors, including the use of synthesisers and developments in loudspeaker technology - the use of stronger magnet materials allows smaller cabinet volumes, which are more easily installed in the home. That high harsh energy is related to low quality shows the sensitivity of the ear at these frequencies, and this measure displays one of the highest correlations.

4.3. Amplitude Features

The relationship between crest factor (as a measure of dynamic range) and quality suggests that listeners can identify reduced dynamic range as a determinant of reduced quality. Despite a different methodology this supports recent studies which refute that hyper-compressed audio is preferred or achieves greater sales[14].

The newly derived Gauss metric works well as a means of classifying the most distorted tracks from those less so, by encoding fine structure in the signal's PMF. With issues relating to loudness and dynamic range compression receiving much attention in the community this new feature can be used to gain insight into the perceptual effects of loudness-maximisation and the history of the "loudness war".

4.4. Rhythmic Features

Slower tempo is associated with higher quality, possibly due to higher production values that can be applied to slower music, such as the addition of string orchestra or layers of backing vocals that can be found in ballads. Also, with a lower event density, there is more space between notes to hear detail in the instruments and better evaluate spaciousness. The correlation between tempo and harsh energy should be noted – faster songs tended to have more cymbal crashes as well as harsh instrument tones to cut through the mix.

4.5. Spatial Features

While one linear model was not appropriate for width, an optimal value is found close to 0.17, where quality ratings reach a peak. This precise value was likely influenced by headphone playback, where sensitivity to width is enhanced. Due to this relationship, the width plot in Figure 5 shows two linear fits, with the dataset divided into values above and below 0.17. The data indicates that there is an increase in perceived quality in going from monaural presentation to an ideal stereo width. However, wider-still samples saw no significant change in quality. For the reference of the reader, the samples used in the test which measured closest to this optimal width are 'Sledgehammer' by Peter Gabriel (1986), 'Superstition' by Stevie Wonder (1972) and 'Firestarter' by Prodigy (1996).

The optimal width was narrower than the mean width (0.25). This may be due to recent attempts in popular music production to produce wider mixes, where this modern width may be presenting as lower-quality due to coincidence with modern dynamic range reduction.

5. CONCLUSIONS AND FURTHER WORK

Correlations between objective measures of digital audio signals and subjective measures of audio quality have been found for the open-ended case of commercial music productions. Dynamics, distortions, tempo, spectral features and emotional predictions have shown correlation with perceived audio quality.

A new objective signal parameter is proposed to unify crest factor, clipping and other features of the audio PMF. This feature works better than crest factor alone for identifying quality, and an analysis of how the feature varies with release year provides an insight into production trends and the evolution of the much-discussed loudness war. Some further work is needed to optimise performance, identify a more robust feature or to test the use of a histogram itself as a feature vector.

Due to the correlations between objective measures and quality perception it is anticipated that quality scores can be predicted by means of the extracted signal parameters. A number of possible implementations are being explored at the time of writing.

With only a relatively small test panel, the results are indicative rather than conclusive. Additional subjective testing would be required to increase confidence in the findings. While concepts have been proven a greater number of audio samples and subjects would be needed for future development. Such a listening test would be well-suited to a mass-participation experiment, conducted on-line. The robustness of feature based quality predictions would benefit from this larger dataset, towards the goal of automatic quality-evaluation and subsequent enhancement.

5 REFERENCES

1. ITU-R BS.1534-1, “Method for the subjective assessment of intermediate quality levels of coding systems”, Tech. Rep., International Telecommunication Union, 2003
2. ITU-T P.800, 1996, “Methods for objective and subjective assessment of quality”, Tech. Rep., International Telecommunications Union, 1996
3. Pras et al, “Subjective evaluation of mp3 compression for different musical genres”, Audio Engineering Society Convention 127, 2009
4. F. Toole, “Loudspeaker measurements and their relationship to listener preferences: Part 1”, J. Audio Eng. Soc. Vol. 34, no. 4, pp227-235, 1986
5. R. Repp, “Recording quality ratings by music professionals”, Proc. Intl. Computer Music conf, New Orleans, USA, Nov. 6-11 2006, pp468-474, 2006
6. E. Vickers, “The loudness war: background, speculation and recommendations”, Audio Engineering Society Convention 129, 2010
7. R. Shatz et al, “The impact of test duration on user fatigue and reliability of subjective ratings”, J. Audio Eng. Soc. Vol. 60, no. 1/2, pp 63-73, 2012
8. 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, Vol. 1, pp. 1-11, 1997
9. B. Glasberg and B. Moore, “A model of loudness applicable to time-varying sound”, J. Audio Eng. Soc. Vol. 50, no. 5, pp. 331-342, 2002
10. O. Lartillot and P. Toiviainen, “A matlab toolbox for musical feature extraction from audio”, Proc. Digital Audio Effects (DAFx-07), Bordeaux, France, Sept. 10-15 2007, pp. 237-244, 2007
11. 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.
12. T. Eerola et al, “Prediction of multidimensional emotional ratings in music from audio using multivariate regression models”, International Conference on Music Information Retrieval, Kobe, Japan, Oct. 26-30, 2009, pp. 621-626, 2009
13. D. Pensado, “Into the lair #47 – Creating loud tracks with EQ and compression”, http://www.pensadosplace.tv/2012/09/18/into-the-lair-47-creating-loud-tracks-w-eq-and-compression/ accessed 4/4/13, 2012
14. N. Croghan et al, “Quality and loudness judgments for music subjected to compression limiting”, J. Acoust Soc Am, Vol. 132, no. 2, pp. 1178-1188, 2012
15. S. Fenton et al, “Objective measurement of music quality using inter-band relationship analysis”, Audio Engineering Society Convention 130, 2011