QUALITY, TIMBRE AND DISTORTION: PERCEIVED QUALITY OF CLIPPED MUSIC

TJ Cox Acoustics Research Centre, University of Salford, Manchester, UK
BM Fazenda Acoustics Research Centre, University of Salford, Manchester, UK
S Groves-Kirkby Acoustics Research Centre, University of Salford, Manchester, UK
IR Jackson Acoustics Research Centre, University of Salford, Manchester, UK
P Kendrick Acoustics Research Centre, University of Salford, Manchester, UK
F Li Acoustics Research Centre, University of Salford, Manchester, UK

1 INTRODUCTION

The Hearing Aid Speech Quality Index (HASQI)[1] has been shown to accurately predict the effect of noise, nonlinear distortion, and linear filtering on the perceived quality of speech.[1,2] A subsequent study by Arehart and colleagues found HASQI also performed reasonably well when assessing the same effects on quality in music.[3] That study however contained a number of limitations; primarily, the small number of audio samples used to develop the model. Only 3 samples of music were used in the test, and one of those contained only a solo voice. Furthermore, samples containing electronic or electronically amplified instruments were deliberately excluded as they "could inherently contain high levels of distortion". Those researchers were also investigating hearing-impaired participants and so the musical styles used were also chosen to be acceptable to older participants as well as the younger listeners in the experiment. Nonetheless, Arehart *et al*[3] did find differences in the prediction accuracy between the three types of music they used, and went on to conclude that that additional work is needed to optimize the index to a wide range of music genres.

To date, most efforts to validate models that predict quality for music have focussed on assessing the effect of a wide range and number of degradation conditions (including both linear and nonlinear processing) on very few samples (or even a single sample) of music.[4] In this paper we adopt an alternative approach and investigate the effect of a single distortion type, hard clipping, on the perceived quality of a comprehensive and systematically selected range of music samples.

Clipping occurs when an input signal exceeds a threshold. During recording, that threshold is set by the gain level of the microphone preamplifier. The effect of clipping distortion is to add extra harmonics. Depending on the sound in question this effect can serve to either enhance or impair the perceived quality.

2 MUSIC SELECTION

A key aim of the current work was to gather a test set of stimuli to include representative examples from a wide range of music. Our starting point for this goal was to refer to previous work investigating music preferences.[5] In a range of studies, Rentfrow and Gosling identified 14 main music genres from an initial pool of 80 genre and sub-genres. These 14 categories were: classical, jazz, blues, folk, alternative, rock, heavy metal, country, pop, religious, rap/hip-hop, soul, funk, and electronica/dance. For each of these 14 genres they then identified approximately 25 prototype exemplar songs, based on consultation with online music providers and music critics, number of units sold, and customer recommendations. From these groups of 25 songs the researchers then selected 10 songs for each genre which best represented "a broad array of styles, artists, and time periods". The outcome of this selection process was a set of 140 songs representing a comprehensive range of music.

CD copies of 117 songs from the list of 140 were obtained. The production of a manageable number of samples for our test set required selecting a subset from this initial pool. Rather than select the samples somewhat arbitrarily, based on their genre category, we opted for a more systematic approach by clustering samples according to timbre.

2.1 Genre vs Timbre

References to genre are often sufficient at an intuitive level but become problematic when attempting to apply the term in a formal manner.[6] Indeed, it is not necessarily clear which level of a hierarchy we should most appropriately apply the term to; the body of work of a group or an artist, a whole album, a particular song, or even particular segments within a song. While intuitively useful, we considered the subjective concept of genre to be too loosely defined, unconstrained, and lacking in universality to be used methodically as the basis for sample selection. Instead, our approach was to analyze features characterising timbre and to cluster the samples with similar timbre.

In perceptual terms, timbre can be considered the feature which allows us to perceive a difference between two sounds with the same pitch, loudness, and duration. Consider the ease with which we can tell apart, say, a flute and a piano playing the same note, for example. In terms of signal processing, timbre is characterised by perceptually weighted spectral measures, such as Mel Frequency Cepstrum Coefficients (MFCC). Extracting these features from a signal allows us to objectively compare the degree of similarity or difference in timbre between samples and thus group or separate samples them as required.

2.2 Clustering by Timbre

Each of the 117 songs was stored as a stereo WAV file at a sampling rate of 44.1 kHz. Three 7-second segments were then identified for each song (representing key sections such as an intro, verse, chorus, etc). A sample length of 7 seconds was used as per the design of similar work conducted previously.[3] The resulting 351 samples and the 3 samples used by Arehart *et al*[3] were then distorted by hard clipping, using a threshold optimised to give a HASQI value of 0.5. All samples, clean and distorted, were then clustered according to timbre. The distorted version of the samples was included in the clustering to provide a better understanding of the effect of distortion on timbre, and to ensure that the timbral-space was representative of the proposed subjective study.

A modified version of the method of timbral clustering used by Autocoutrier and Pachet[7] was used. This technique discards temporal information, which has in any event been shown not to enhance the performance of a musical similarity predictor.[8] Furthermore, the most significant perceptual effects for hard clipping will arise from the changes in spectral content.

A time-frequency representation of the audio was extracted (12 MFCCs, using 20 ms windows, 50 % overlap). To quantify the stochastic variation of the spectra over time, a Gaussian Mixture Model (GMM) was fitted to the MFCCs for each song, treating the MFCCs as a random variable with 12 dimensions. To measure similarity between songs, the Likelihood that each set of MFCCs was generated by each GMM is used as a timbre similarity measure (the likelihood that a song's own MFCCs were generated by its own GMM was excluded). This yields a similarity vector, with 707 dimensions. Finally an unsupervised clustering of similarity vectors is carried out, also using a GMM. The Akaike information criterion is used as an indication of the best number of clusters, which in this case was found to be six.

Clustering samples according to timbre allows us to sidestep the problem of subjectivity in the category definitions of genre. In this approach output classification is neither specified nor labelled in advance but emerges from objective measures of similarity in the data. In this way, the clusters which emerge from the process can no longer be said to (necessarily) represent particular genres in the normal sense. However, by clustering according to features we can be confident that a test set drawn from each of the clusters is representative of the full range of timbre space present in the initial pool.

Two samples were drawn from each of the six clusters to form the final test set of stimuli (with the exception of one cluster which contained one sample and no others). Samples were drawn by selecting the two with the shortest Euclidian distance to the cluster centres, as defined by the GMM components parameters. Additionally, each of the three samples used by Arehart *et al*[3] were also included in the test set, regardless of which cluster they had grouped with. The 14 songs which the test set stimuli were taken from are presented in Table 1.

Table 1. The 14 songs the final test samples were taken from, by cluster number.

Cluster Number	Song Name	Artist/Composer
1	Riverboat Set: Denis Dillon's Square Dance Polka, Dancing on the Riverboat	John Whelan
	Crazy Train	Black Sabbath
	"Haydn" *	*
2	Ave Maria	Franz Schubert
	Packin' Truck	Leadbelly
	"vocalise" *	Tierney Sutton
3	Kalifornia	Fatboy Slim
	Brown Sugar	The Rolling Stones
4	The Four Seasons: Spring	Antonio Vivaldi
5	For What It's Worth	Buffalo Springfield
	The Girl From Ipanema	Stan Getz
6	Spoonful	Howlin' Wolf
	Nobody Loves Me But My Mother	B.B. King
	"jazz" *	*

Note. The 3 samples marked with an asterisk are those used in Arehart *et al's* study and appear here as labelled in that study. Additionally, the sample "jazz" was also the sole music sample investigated in an earlier study by Tan, Moore, and Zacharov.[9]

The sample "vocalise" is an extract of a cappella performance by a female jazz vocalist singing in a scat style. The sample "Haydn" is an excerpt from the second movement of Haydn's Symphony No. 82, and features the full orchestra. The sample "jazz" is an excerpt of a performance by a jazz trio consisting of piano, drums and double bass.

3 METHOD

3.1 Participants

A total of 30 participants (mean age: 23.7 years; SD: 4.7 years) completed the experiment. None reported any known hearing impairments.

3.2 Test Materials

Each participant was presented with a total of 140 samples. These stimuli consisted of 10 processing conditions (including a clean condition) of a 7 second extract from each of the 14 songs listed in Table 1. All samples were presented at 72 dB over Sennheiser 650 HD headphones, via a Focusrite Scarlett 2i4 audio interface (having previously been calibrated using a dummy head). Samples were presented in stereo. Playback level was calibrated by setting the playback of the clean "jazz" excerpt to 72 dB (average of both channels), as this was the level used by Arehart *et al.*

To normalise playback level across the test set, the A-weighted level of each of the other samples was set to the same A-weighted level as the "jazz" sample.

To ensure that each song sample was presented across a wide range of quality, from clean to highly distorted, each sample was subjected to 9 levels of distortion. These levels were estimated using HASQI values. As the HASQI index values for quality run from 1 (clean) to 0 (poorest quality) we calculated thresholds for each clip which approximated every (nonlinear) HASQI value between 0.1 and 1, in 0.1 intervals. For the poorest quality samples, in instances where estimates approximating 0.1 could not be achieved using HASQI, the lowest available value was used instead. Using this procedure we obtained a set of 10 different stimuli (9 levels of distortion plus the original clean sample) for each of the 14 samples, creating the final test set of 140 samples.

3.3 Procedure

Participants were provided with information about the test and reminded that they were judging the overall quality of each sample, not how much they liked the music. Ratings were provided by participants moving a slider with a mouse. The slider was labelled "Bad" and "Excellent" at each endpoint (as per the scale endpoints used by Arehart *et al*[3]) but contained no other markers (actual output data were integer values from 0 to 100). The slider's initial position was at the "Bad" end of the scale on each trial. Progression from one trial to the next was conditional on listening to the sample in full and providing a rating but there were no limits on the number of times each sample could be listened to. There was no time limit for completion of the test and participants were prompted to take a short break at the half-way stage if required. Presentation order of the samples was fully randomised.

Before the test began participants were presented with 3 pairs of samples - not included in the test set – and informed that these samples represented examples of the best and worst audio quality they would hear (pre-test example samples were: "Fire and Rain" by James Taylor; "Mountain Song" by Jane's Addiction; "Summer in the City" by Quincy Jones). The test session typically lasted around 40 minutes and participants were financially reimbursed for their time.

4 RESULTS

Relationships in the data were explored in a number of ways. First, differences between clusters were examined using a repeated-measures analysis of variance (ANOVA) with distortion level and cluster as independent variables, and mean quality ratings for each cluster as the dependent variable.

A significant main effect was found for both cluster, $F(2.33, 67.48) = 42.43$, $p = <.01$, $\eta_p^2 = .59$, and distortion level, $F(4.97, 144.26) = 458.38$, $p = <.01$, $\eta_p^2 = .94$. A significant interaction between cluster and distortion level was also observed, $F(11.91, 345.41) = 6.98$, $p = <.01$, $\eta_p^2 = .19$, indicating that, while increases in distortion significantly reduced quality across all clusters, for some clusters the effect was more pronounced than in others. For distortion level, each consecutive increase in level of clipping was associated with a significant reduction in quality.

Analyses of cluster types showed that clusters 1, 2, and 6 did not significantly differ from each other, but did significantly differ from each of the other clusters. The same pattern was found for clusters 3 and 5 (all, *p*s < .01). Quality ratings in Cluster 4 were found to be consistently higher than all other cluster types. Mean quality ratings are displayed in Figure 1, shown by cluster.

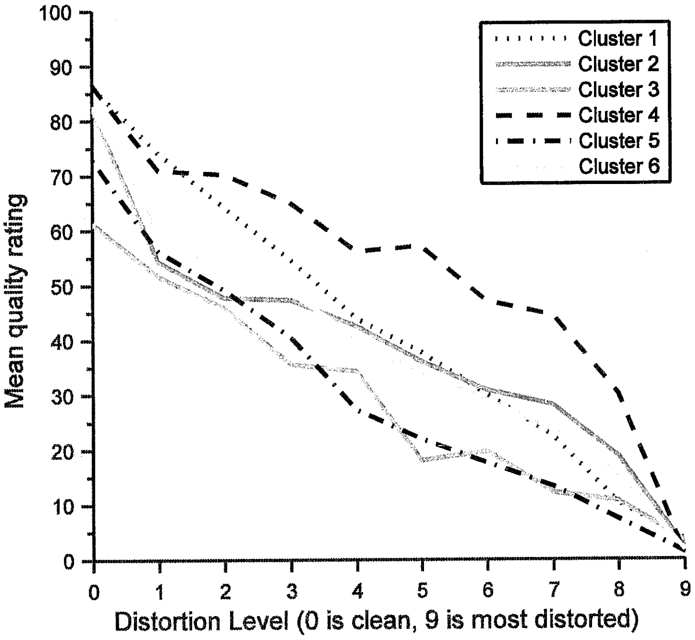

Figure 1. Mean quality ratings of each cluster, as a function of distortion level.

As can be seen in Figure 1, participants' mean ratings for the clean samples were found to vary considerably (by a range of 25 points on our 100 point scale). To assess the influence of these initial differences in quality the ANOVA was repeated with quality ratings normalised so that each participant's rating for each of the clean samples was equal to 1. Rates of degradation following normalisation of the quality ratings can be seen in Figure 2.

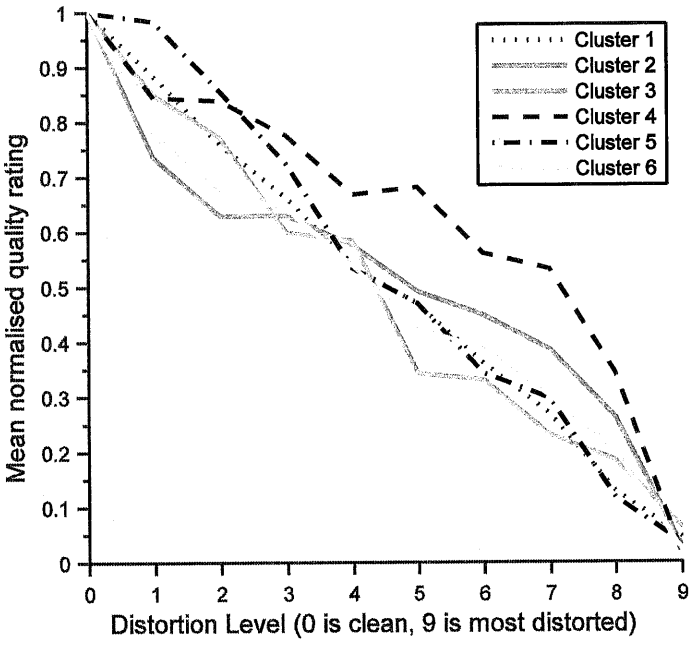

Figure 2. Mean normalized quality ratings of each cluster, as a function of distortion level.

Once normalised it was found that the previously significant main effect of cluster type was no longer present, suggesting that the rates of degradation do not significantly differ across cluster types once the initial quality rating of the unprocessed sample is taken into account. A significant main effect for distortion level remained, $F(3.32, 96.24) = 280.73$, $p = <.01$, $\eta_p^2 = .91$.

To assess the performance of the HASQI model, correlation coefficients were calculated between the models' predicted values for quality and participants' mean quality ratings in each cluster. Coefficients were calculated for both participants' actual ratings and the normalised ratings of quality, and also for values predicted by the HASQI Nonlinear model and HASQI combined (nonlinear and linear components) model. All correlation coefficients are listed in Table 2.

Table 2. Correlation coefficients for relationships between participant quality ratings and values predicted by the HASQI nonlinear and combined models for each timbre cluster.

	Quality		**Normalised Quality**	
Cluster	**Nonlinear**	**Combined**	**Nonlinear**	**Combined**
1	.805	.828	.746	.770
2	.675	.689	.568	.595
3	.694	.693	.695	.695
4	.675	.671	.647	.644
5	.794	.801	.359	.358
6	.748	.755	.645	.649
Mean (SD)	.721 (.059)	.732 (.065)	.544 (.136)	.533 (.141)

Note. All correlations significant at p <.01.

5 DISCUSSIONS AND CONCLUSIONS

The aim of this paper was to investigate the effect of one form of distortion, hard clipping, on a comprehensive range of music. Samples from 117 songs representing a wide range of music genres were clustered according to timbral features using Gaussain Mixture Models. From the 6 clusters that emerged we drew 14 samples to process with 9 levels of distortion, and obtained participant ratings of overall quality for each.

Overall, it was found that the presence of distortion in music affected perception of quality differently in music with different timbral features. Normalising the quality ratings to remove the influence of initial differences in quality ratings of the clean samples removed the significant differences between the timbre clusters but also reduced the overall accuracy of the HASQI predictions.

When assessing the predictive performance of HASQI against the actual quality ratings obtained in the lab the strongest overall correlation we observed (collapsing across cluster types) was .732, for the combined HASQI model prediction of quality. In a similar test, but using a fewer samples of music, Arehart *et al*[3] found a correlation coefficient of .838. By contrast, the same model's performance for prediction of speech quality is considerably stronger,[1] with a correlation coefficient of .942. Poorer performance when applied to a wider range of music styles suggests that we cannot assume the model is equally accurate across all genres when predicting quality for hard clipping. The range of correlation coefficients observed within our timbre clusters (from .671 to .828) is notable and reflects the fact that HASQI currently over-estimates the impact of distortion in some instances (cluster 4, for example) while under-estimating the effect of distortion in others (such as cluster 3). These findings imply the accuracy of HASQI could be improved by factoring in timbral features of samples prior to computing quality predictions.

Future challenges will include moving beyond hard clipping to investigate other forms of distortion, each of which will likely generate different clustering of samples.

6 ACKNOWLEDGMENT

This work was supported by the EPSRC (Project EP/J013013/1). The authors wish to thank James M. Kates and Brian C.J. Moore for sharing stimuli from their work cited in this paper.

7 REFERENCES

1. J.M. Kates and K.H. Arehart. The Hearing-Aid Speech Quality Index (HASQI). J. Audio Eng. Soc. 58(5): 363–381. (2010).

2. A. Kressner, D. Anderson, and C. Rozell. Evaluating the generalization of the Hearing Aid Speech Quality Index (HASQI). IEEE Trans. Audio. Speech. Lang. Processing. 21(2): 407–415. (2013).

3. K.H. Arehart, J.M. Kates and M.C. Anderson. Effects of noise, nonlinear processing, and linear filtering on perceived music quality. Int. J. Audiol. 50(3):177–190. (2011).

4 B.C.J. Moore, C-T, Tan, N.Zacharov and V-V. Mattila. Measuring and predicting the perceived quality of music and speech subjected to combined linear and nonlinear distortion. J. Audio Eng. Soc. 52(12): 1228–1244. (2004).

5. J.P. Rentfrow and S.D. Gosling. The Do Re Mi's of everyday life: The structure and personality correlates of music preferences. J. Pers. Soc. Psychol. 84(6): 1236-56. (2003)..

6. N. Scaringella, G. Zoia, and D. Mlynek. Automatic genre classification of music content: a survey. Signal Process. Mag. IEEE. 23(2):133–141. (2006).

7. J.J. Aucouturier and F. Pachet. Music similarity measures: What's the use?. Proc. ISMIR. (2002).

8. J.J. Aucouturier and F. Pachet. Improving timbre similarity: How high is the sky? J Neg Res in Speech and Audio Sciences. 1(1): 1-13. (2004)

9. C-T. Tan C-T, B.C.J. Moore and N. Zacharov. The effect of nonlinear distortion on the perceived quality of music and speech signals. J. Audio Eng. Soc. 51(11): 1012–1031. (2003).