OPEN ACCESS Freely available online
Journal of the Audio Engineering Society
Vol. 63, No. 9, September 2015 (© 2015)
DOI: http://dx.doi.org/10.17743/jaes.2015.0068

Perceived Audio Quality of Sounds Degraded by Non-linear Distortions and Single-Ended Assessment Using HASQI

PAUL KENDRICK, *AES Member*, **IAIN R. JACKSON, FRANCIS F. LI, TREVOR J. COX,** *AES Member*, **AND BRUNO M. FAZENDA,** *AES Member*

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

For field recordings and user generated content recorded on phones, tablets, and other mobile devices nonlinear distortions caused by clipping and limiting at pre-amplification stages, and dynamic range control (DRC) are common causes of poor audio quality. A single-ended method to detect these distortions and predict perceived degradation in speech, music, and soundscapes has been developed. This was done by training an ensemble of decision trees. During training, both clean and distorted audio was available and so the perceived quality could be gauged using HASQI (Hearing Aid Sound Quality Index). The new single-ended method can correctly predict HASQI from distorted samples to an accuracy of ±0.19 (95% confidence interval) using a quality range between 0.0 and 1.0. The method also has potential for estimating HASQI when other types of degradations are present. Subsequent perceptual tests validated the method for music and soundscapes. For the average mean opinion score for perceived audio quality on a scale from 0 to 1, the single ended method could estimate it within ±0.33.

0 INTRODUCTION

Modern technologies have enabled handy recording devices, large data storage, and diverse outlets of User Generated Content (UGC). Three hundred hours of video are uploaded to YouTube every single minute, and along with other online databases such as freesound.org and soundcloud.com, much user generated audio is widely available. UGC is now used extensively in news broadcasting: on average, a news agency adopts 11 pieces of UGC daily [1]. This necessitates a rapid assessment method to determine if the UGC is broadcast-worthy and so media asset management systems would benefit from automatically generated audio quality metadata. Furthermore, if audio problems can be detected while recording, feedback can be given to the operator of the device and many disappointing end results can be avoided. A survey of both amateur and expert recordists [2] found that the four most commonly reported errors were: background noise (59%), wind noise (46%), handling noise (31%), and other distortions (19%). Wind noise problems in recordings have been addressed recently by the authors [3]. Motivated by the need to tackle other recording errors, this paper develops a method that can predict the perceived quality of audio contaminated by distortion. Distortion problems also arise with other audio systems such as hearing aids, sound reinforcement, and public address systems, and consequently the method developed has a wider applicability than just UGC.

Three of the most common objective measures to quantify non-linear distortions are Total Harmonic Distortion (THD) [4], Inter-Modulation Distortion (IMD) [5], and Total Difference-Frequency Distortion TDFD [6] [7]. Lee and Geddes [8] [9] showed that there is a poor correlation between the perceived amount of distortion and the THD and IMD for a piece of music. They proposed an alternative measure with improved correlation based on integrating the 2^{nd} differential of the non-linear amplitude transfer function. A number of perceptual measures have been developed to better model the perceived quality after degradation. These include double-ended methods for speech [10]–[13] that have been standardized such as Perceptual Evaluation of Speech Quality (PESQ) [14] and the updated version POLQA [15]. Perceptual Evaluation of Audio Quality (PEAQ) [16] has also been developed to assess audio quality. PEAQ and PESQ are primarily used for assessing quality degradations caused by digital coding, complex audio processing, or transmission chains [17]. The Distortion Score (DS) [18], R_{nonlin} [17], and the Hearing Aid Sound Quality Index (HASQI) [19] are double-ended methods able to predict the degradation in quality caused by overloading of transducers and preamplifiers. Recent studies have shown that HAQSI generalizes well for normal

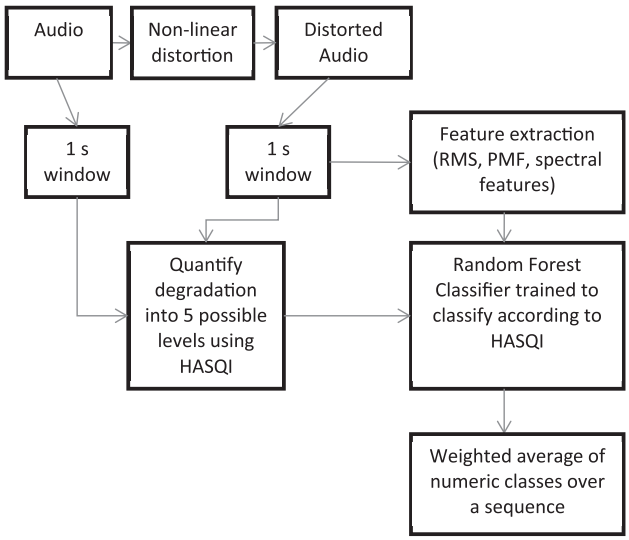

Fig. 1. A block diagram of the proposed method

hearing listeners [20] achieving good accuracy when predicting mean opinion scores. For music HASQI was found to be able to predict the perceived degradation in audio quality due to clipping effectively [21]. HASQI can therefore be used to assess distortion on transmission channels but only if both the original and degraded signals are available.

There are many occasions where the undistorted sound is unknown. UGC is a good example where a single-ended method is needed working just from the corrupted audio. An example of a single-ended method is ITU Recommendation P.563 [22] but this is restricted to narrow band speech. Maré [23] presented a method to detect clipping in audio signals using a supervised artificial neural network. The test set was not sufficiently distinct from the training set, however, raising doubts about the capability of the method to generalize to unknown sources.

The new method presented below exploits a different machine learning regime to map features extracted from the corrupted audio to predict human perceived quality monitored using HASQI. A broader database of samples is used, demonstrating the need for more features to achieve generalization.

1 METHOD

A machine learning regime is used to take features extracted from the distorted audio and predict human perceived quality. Fig. 1 gives an overview of the proposed method. Speech, music, and soundscape samples were artificially distorted in a controlled manner using a diverse range of non-linear processes. The distortion of each sample was quantified using HASQI to form a teacher value for the machine learning algorithm that is used during supervised training. Before passing the audio to the machine learning algorithm it is necessary to reduce the amount of data, and this is done by extracting key features.

1.1 Database Formation

The machine learning scheme will learn to map from audio features to HASQI using a large database of training examples. The inclusion of a sufficient number of cases in the dataset is vital. The cases need to represent the wide range of likely audio samples in terms of what might be recorded and also the distortion likely to be encountered.

1.1.1 Audio Database

Speech, music, and soundscape samples were used to represent all the most likely sources of recorded audio. An audio database was collected from a large collection of CDs, including speech, music of various genres, and soundscapes counting a range of geophonic, biophonic, and anthrophonic sound sources. The database contains 404 music files with an average length of 2 minutes 45 seconds, 182 speech files with an average length of 4 minutes 48 seconds, and 469 soundscape clips with an average duration of 1 minute 48 seconds. At least one 10-second excerpt was randomly taken from each of these files, resulting in 1500 10-second excerpts for each of speech, music, and soundscape, with about 500 of each type.

1.1.2 Distorting Samples

To create distortion algorithms to degrade the samples, it was necessary to better understand common recording problems and technologies. In microphones and preamplifiers, overloading can occur when the signals go beyond a device's dynamic range. This causes the peaks in a waveform to be clipped generating harmonics of the original signal. In addition, when the analogue signal exceeds the dynamic range of an AD converter, aliased distortions may also be introduced.

Many devices incorporate Dynamic Range Control (DRC) to protect against overloading. The DRC reduces the amplification gain when the peak or root mean square (rms) of the signal is likely to overload the circuit. Instead of reducing the gain instantaneously, the DRC often incorporates an integration period, characterized by an attack and release time, and the gain reduction is usually characterized by a compression ratio. Dynamic range control systems can inadvertently degrade perceived quality, and careful choice of parameters is important [24]: (i) Audible distortion may occur if the release time is too short and the amplitude gain is modulated too quickly. (ii) Dropouts are likely to happen if the release time is too long because the suppressed gain does not recover quick enough to handle subsequent weak signals. This produces a "pumping" effect that is obvious to the listener. (iii) When the attack time is too short, the transients are suppressed excessively resulting in a lack of punch and clarity. The effectiveness of the compression can also be compromised. In addition, the DRC system is a dynamic compressor and so it may also introduce other artifacts or nonlinear distortions and degrade the signal to noise ratio [25].

Kendrick et al. examined the DRC systems for a number of portable audio devices [26]. The devices tested included mobile phones, portable audio recorders, cameras, and

Table 1. The range of DRC parameters measured for 9 devices

	Minimum	Maximum
Attack time	1 ms	17 ms
Release time	0 ms	400 ms
Compression ratio	1.4	Inf.

sound cards (Cannon 550D, Edirol r44, Neumann U87ai via Focusrite 2i4, Shure SM57 via Focusrite 2i4, Zoom H2, Zoom H4, Google Nexus 4, iPhone, and a Sony vx2000 camcorder). Table 1 describes the ranges of the three key parameters found in the devices that had DRC.

DRC may not completely eliminate overloading, in which case when the signal level is high the compression ratio would be inadequate. Therefore, to detect non-linear distortions in audio all three scenarios must be carefully considered in constructing the database of examples—overloading at the preamplifier; distortions due to the DRC system, and overloading during analogue to digital conversion.

Distortion was emulated using the method developed by De Man and Reiss [27] in which the following amplitude transfer function was used to generate non-linear distortions of different types,

$$f(x_B) = sgn(x_B)\frac{K|x|^3 - T\left(2K^{\frac{3}{2}} + K + 2K^{\frac{1}{2}}\right)|x|^2 + T^2\left(2K^2 + 2K^{\frac{3}{2}} + 4K\right)|x| - T^3\left(2K^{\frac{1}{2}} + 1\right)}{\left(K^2 + 2K^{\frac{3}{2}} - 2K^{\frac{1}{2}} - 1\right)T^2} \quad (1)$$

where $x_B = x + B$; $x$is the instantaneous value of the input signal (ranging between –1 and 1); T is the threshold (value between 0 and 1); K is the knee parameter ($K = 1$ for a hard knee, $K > 1$ for a soft knee) where a Hermite spline is used to connect the linear part (that ends where $|x| = T/\sqrt{K}$) and the non-linear part; and B is a bias parameter that adds a small DC offset to the signal. Components in the signal from 22050 to ∞ Hz, can be aliased. To simulate distortion without significant aliasing, the signal was up-sampled four times to 176.4 kHz prior to applying the amplitude transfer function and then down-sampled to 44.1 kHz afterwards. The oversampling rate was chosen by computing the signal power above 22050 Hz in the oversampled signal for typical sources and parameters. As the oversampling rate is increased the signal power above 22050 Hz in the digital domain converges towards the power in the analogue signal above 22050 Hz. This convergence indicates that above a certain oversampling level aliasing becomes insignificant; an oversampling rate of 4 was found to be sufficient.

The Dynamic Range Control was emulated using the method by Giannoulis et al. [28]. Peak level detection was chosen for its prevalence in DRC systems. Giannoulis et al. modeled four peak detection methods in DRC systems including branching, smoothed-branching, decoupled, and smoothed-decoupled.

Decoupling is where the peak level is measured using a separate circuit that ensures that the peak level measure is not dependent on the attack time. This is simulated by,

$$\begin{aligned} peak_1[n] &= \max\left(x_L[n], \alpha_R peak_1[n-1]\right) \\ peak_L[n] &= \alpha_A peak_L[n-1] \\ &\quad + (1-\alpha_A)\alpha_R peak_1[n] \end{aligned} \quad (2)$$

where $\alpha_A = e^{-1/(\tau_a Fs)}$ and $\alpha_R = e^{-1/(\tau_r Fs)}$; τ_a is the attack time; τ_r the release time; $peak_L[n]$ is the peak level at sample n; $x_L[n]$ is the absolute value of sample n; and Fs is sampling frequency. In this method the attack envelope is imposed on the release envelope, and therefore a branching simulation is also developed that ensures the attack and release envelopes are also decoupled. If the signal does not completely decay away after the compressor is released, the release envelope will decay at the prescribed rate and will meet a background plateau more quickly than expected. To ensure that the release time is always the same, the releaseenvelope can be smoothed so that it decays gently to the background level rather than silencing abruptly.

$$peak_L[n] = \begin{cases} \alpha_A peak_L[n-1] + (1-\alpha_A)x_L[n] & x_L[n] > peak_L[n-1] \\ \alpha_R peak_L[n-1] & x_L[n] \le peak_L[n-1] \end{cases} \quad (3)$$

Smoothing can be applied to both methods; for the branching method the peak detection becomes,

$$peak_L[n] = \begin{cases} \alpha_A peak_L[n-1] + (1-\alpha_A)x_L[n]x_L[n] > peak_L[n-1] \\ \alpha_R peak_L[n-1] + (1-\alpha_R)x_L[n]x_L[n] \le peak_L[n-1] \end{cases} \quad (4)$$

and the decoupled peak detection,

$$\begin{aligned} peak_1[n] &= \max\left(x_L[n], \alpha_R peak_1[n-1] + (1-\alpha_R)x_L[n-1]\right) \\ peak_L[n] &= \alpha_A peak_L[n-1] + (1-\alpha_A)\alpha_R peak_1[n] \end{aligned} \quad (5)$$

These four methods introduce varying levels of harmonic distortion [24].

A Monte Carlo simulation was carried out with each of the 10-second audio samples being distorted or compressed in six ways as shown in Table 2. As this is a system that learns from data, care was taken to ensure that the distribution of samples was well balanced in terms of the types of non-linear processing that may be encountered. For the clipping distortion, the parameters used for the simulation are described in Table 3 and for the DRC the parameters in

Table 2. Distortion types used to train detector

Distortion class	Distortion type
1	No Distortion
2	Clipping with reduced aliasing
3	Clipping with aliasing
4	DRC present
5	DRC present with clipping afterwards
6	DRC present with aliasing clipping afterwards

Table 4. These parameters are randomly generated but with rules applied to the generating functions to ensure balanced distribution of examples. The reasons for each choice are explained in more detail in Appendices 1 and 2.

1.1.3 Teacher Values

Supervised machine learning needs large quantities of labeled data for training. The massive number of samples due to the combination of distortion types, distortion levels, and huge number of original sources make labeling them by subjective testing impossible. Taking advantage of having both the original and distorted audio during the training phase, a double-ended method could be used to estimate HASQI [19] as the teacher values. The original and distorted audio samples were truncated using rectangular windows of one second. Fifty-percent overlap was used. Each window was normalized to the rms value of that window before estimating HASQI.

HASQI is a continuous value from 0 to 1 but is based on subjective tests that returned a five level quality score from *Bad* to *Excellent* as suggested by ITU-R BS.1284-1 [29]. As a supervised classifier was adopted to perform the prediction, HASQI is first quantized back to the five classes shown in Table 5. The class determined by HASQI over one second using the double-ended method will be referred to as $ClassD$, and the single-ended estimate of that class is referred to as $ClassS$. The reason for the non-uniform scale divisions is due to the definition of the ends of the HASQI scale, where $Bad = 0$ and $Excellent = 1$, spacing the other descriptors equally over the scale and then quantizing causes, *Good*, *Fair*, and *Poor* classes to have a width of 0.25, while Excellent and Bad have a smaller width of 0.125.

1.2 Machine Learning Algorithms

Support Vector Machines (SVMs), Artificial Neural Networks (ANNs), Hidden Markov Models (HMMs), and Gaussian Mixture Models (GMMs) are well-known machine learning algorithms in audio classification and pattern recognition. Decision trees have recently gained much attention in related applications and the authors have applied them to wind noise assessment [3]. Consequently, the random decision forest [30], also known as a random forest, was adopted. The Matlab class "TreeBagger" is used to train the random forest [31].

Machine learning is often tested using k-fold cross validation to test how well the trained system deals with cases that were not present in the training and is used in the study. In addition, perceptual experiments were carried out to more rigorously validate the method (see Sec. 3).

1.3 Audio Features

Features were extracted from the distorted audio to be used as the input to the random decision forest. Features were extracted within frames of 1024 samples (23 ms) and 50% overlap was used. Clipping and DRC are known to cause sample values to be redistributed. This can be captured by the probability mass function (PMF), which is the discrete form of the probability density function. Fig. 2 shows four example PMFs for the same one second of audio, one with no clipping and the others with clipping applied. Hard clipping ($K = 1$), causes an increase in the probability a sample will occur around a relative sample value of ± 1. Amplitude transfer functions with a soft knee also show a peak at ± 1 but with a smoother transition and a lower peak value. A bias causes translation of the PMF in the direction

Table 3. Clipping parameters for Monte Carlo simulation

Parameter	Parameter generating functions x is a random variable with a uniform probability density function between 0 and 1		
T (Threshold, linear)	$T = x^{1.5}$		
K (Knee type)	50 % chance $K = 1$ (hard clipping)	25 % chance $K = 1 + 100x$ (soft clipping)	25 % chance $K = 1 + x$ (soft clipping)
B (Bias)	50 % chance $B = 0$	50 % chance $B = x - 0.5$	

Table 4. DRC parameters for Monte Carlo simulation

Parameter	Parameter generating functions, x is a random variable with a uniform probability density function between 0 and 1			
T (Threshold, dB)	$T = -40x$			
τ_a (attack time, s)	$\tau_a = 0.02x + 0.0001$			
τ_r (release time, s)	$\tau_r = 0.5x$			
R (Compression ratio)	50 % chance $R = \infty$	50 % chance $R = 40x$		
DRC model	25 % chance branching model	25 % chance smoothed branching model	25 % chance decoupled model	25 % chance smoothed decoupled model

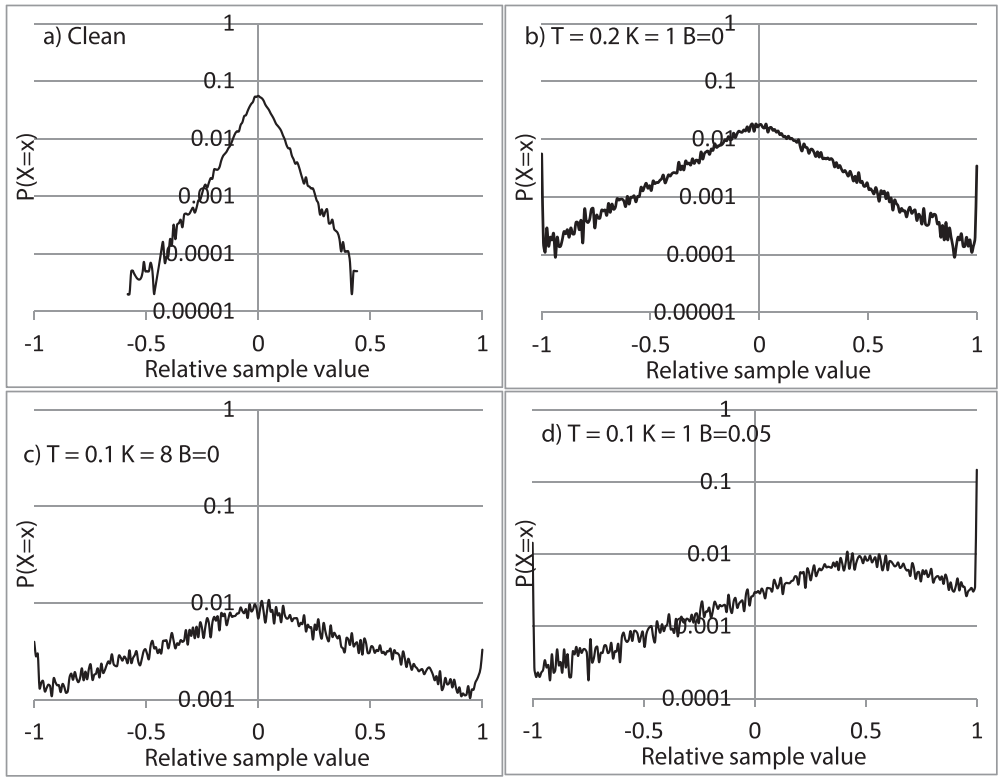

Fig. 2. Probability Mass Functions (PMF) for an audio sample comparing the clean (a) with three different levels of distortion hard clipping (b), soft clipping (c), and hard clipping with a DC bias (d).

Table 5. Quantization of HAQSI into five classes

Class D	HASQI range	Quality
5	$1 < \text{HASQI} \leq 0.875$	Excellent
4	$0.875 < \text{HASQI} \leq 0.625$	Good
3	$0.625 < \text{HASQI} \leq 0.375$	Fair
2	$0.375 < \text{HASQI} \leq 0.125$	Poor
1	$0 < \text{HASQI} \leq 0.125$	Bad

of the sign of the bias and reduces the peak at one extreme while increasing it at the other.

To compute the PMF, each audio frame was normalized to the maximum absolute sample value, the histogram was then computed using 255 equally spaced sample levels from –1 to 1. The normalization in each window ensured that the PMF was represented with an optimal resolution for that window.

Maré [23] showed how the PMF could be used to identify distortions. To achieve generalization to audio not seen in training, we found that more features are necessary to represent a wide range of signal properties including timbre, spectral features. These were calculated using the MIR toolbox [32] and are listed in Table 6. The mean for each feature was then computed over 1 second.

1.3.1 Feature Selection and Training

To identify which features should be presented to the random decision forest, a sequential forward feature selection was carried out using 2-fold cross validation. Random decision forests allow some integration of automatic feature selection within the learning process. This is particularly useful when handling empirical data with no explicit model or clue for heuristic feature selection.

The random decision forest is an ensemble learning method that uses bagging, whereby a number of classification decision trees are each trained on a bootstrap sampled (with replacement) subset of the data, and at each node a randomized subset of features are selected and used for classification. Brieman [30] suggested that an optimal size of the feature subset would be $\sqrt{m}$ (rounded to the nearest integer), where m is the total number of features.

Using $\sqrt{m}$ features for each split, greedy forward feature selection [33] (FFS) was carried out using a wrapper method, which means that the output error from the trained classifier is used to gauge the quality of the algorithm. Two-fold cross validation was carried out for every feature set, each time ensuring that the same source of audio did not appear in both training and validation tests.

The performance was quantified using the Matthews Correlation Coefficient (MCC), which takes a value between 0 and 1, where 1 represents optimal performance. The MCC is calculated from the confusion matrix [34]. The FFS was initialized by training a predictor using each feature separately. The best performing feature was the one that produced the highest MCC averaged over all folds. Having determined the first feature to be used, the second, third, fourth, etc., were then determined. The training was

Table 6. Features and their rank order in the feature selection process, definitions of the features are provided in [32]

Rank order	MIR toolbox features	Number of times feature was selected
1	PMF	12
1	Spectral Flux	12
3	Spectral Kurtosis	10
4	Spectral Entropy	8
4	Spectral Roughness	8
6	Spectral Skewness	7
6	Zero crossing rate	7
8	Spectral Irregularity	6
9	Attack Slope	5
10	Spectral Spread	4
11	MFCCs	2
12	dMFCCs	1
13	Spectral Flatness	1
–	rms level	0
–	Tempo	0
–	Spectral Centroid	0
–	Spectral Brightness	0
–	Spectral Rolloff 85%	0
–	Spectral Rolloff 95%	0
–	ddMFCCs	0
–	Low energy	0
–	Attack Time	0
–	Spectrum	0

undertaken with every possible additional feature added to the first feature with the best individual performance. If the added feature increased the MCC, then the feature was retained. This procedure was repeated until all the features under investigation were exhausted or there was no further improvement in performance. If a feature contained multiple values, such as the 255 values in the PMF, these were treated as a single feature, i.e., all 255 values were included or removed in one block.

The random forest is a stochastic method and will yield different results every training phase due to both the bootstrap sampling and the random selection of features at each node. By increasing the size of the forest the variance between the outputs from the trees is decreased, therefore there is a trade-off between variance and speed of processing. A rule of thumb, the number of trees in the forest needs to be sufficient so that the ranking of the features no longer changes as the number of trees is increased [35]. To determine the optimal forest size, a significance test of the performance improvement was carried out between two forest sizes after feature selection. The feature selection procedure was repeated for a number of forest sizes, increasing the number of trees by a factor of 2 starting at 12 (multiples of 12 was a convenient choice because the parallel code was running on a 12 core machine).

McNemar's hypothesis test was used to determine the significance [36]. A hypothesis test is defined where the null hypothesis is rejected (that there is no difference between predictors), if $\chi^2 > \chi^2_{1,0.05} = 3.851$ (significance level $p <$ 0.05) and if the MCC of the larger forest is greater than the smaller one where,

$$\chi^2 = \frac{(|M_{ab} - M_{ba}| - 1)^2}{M_{ab} + M_{ba}} \sim \chi^2_1 \tag{6}$$

where M_{ab} is the number of misclassifications made by the smaller forest, which were correctly classified by the larger forest, and M_{ba} is the number of misclassifications made by a larger forest, which were correctly classified by the smaller forest, $\sim\chi_1{}^2$ expresses that the function has a chi-square distribution with 1 degree of freedom. Table 7 presents the results from the forest size investigation showing no significant improvement in performance above a forest size of 96.

Table 7. Random forest size vs MCC

Trees	MCC	χ^2
12	0.56	N/A
24	0.58	18.20
48	0.60	47.76
96	0.61	10.72
192	0.61	0.03
384	0.61	0.92
768	0.61	0.18
1536	0.61	3.76

The feature selection algorithm produces a different permutation of features every time. Therefore to select the best set of features, the FFS was run repeatedly and the features most frequently selected were used. The FFS was repeated until the rank order of the top N features in the rank order stabilized (no change after two FFS repeats). On an average, 7 features were selected and stability occurred after 12 runs. The rank order and the frequency a feature was selected is shown in Table 6. PMF being joint top supports the work done by Maré [23]. Alongside this was spectral flux, which is the mean Euclidian distance of the spectra between successive frames. Other important features were Spectral Kurtosis, Spectral Entropy, Spectral Roughness (average of all the dissonance between all possible pairs of peaks [37]), Spectral Skewness, and the Zero crossing rate.

Much of the information contained in the spectral and timbral features is already available from the PMF. This indicates that in a lower computational power environment (e.g., a smart phone) where a compact algorithm may be required, the PMF might be sufficient.

2 RESULTS

Table 8 shows a confusion matrix from a system averaged over 2-folds using the 7 chosen features and 96 trees. The MCC was 0.616. Fig. 3 illustrates the performance for different signal and distortion types. Aliasing had little effect on performance of the algorithm, therefore non-aliasing and aliasing cases were pooled for each distortion type. Fig. 3 shows that the performance is generally similar for both soft and hard clipping, but there are small differences between source types, with the estimation being best for music and worst for speech. The relatively poor

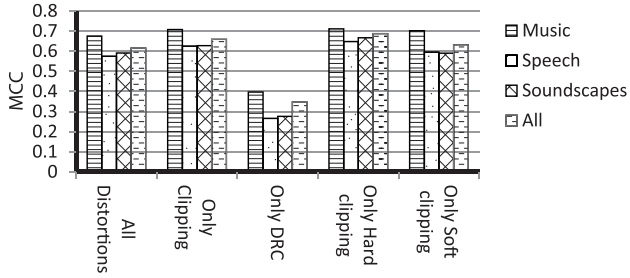

Fig. 3. Mathews Correlation Coefficient (MCC) as a measure of classification accuracy for different audio sources and distortion types.

performance occurs when the degradation to quality is due to DRC alone. The confusion matrix for DRC-only cases in Table 9 shows 96% were rated good or excellent—DRC is not degrading the audio as badly as the other types of distortion. While there appears to be confusion between the two highest quality classes, very rarely will a sample be mislabeled more than two classes above or below its true class.

2.1 Aggregation Over Longer Samples

Human judgments of audio quality are usually made over periods longer than one second, therefore a method to aggregated the results over a longer time period is needed. A similar judgment of temporally varying phenomena has been studied in soundscapes research and VoIP speech quality. Dittrich and Oberfeld [38] showed primacy (first sound heard) and recency (last sound heard) effects for annoyance from broadband noises. Västfjäll showed that listeners consistently preferred in-flight soundscapes with a better ending [39]. The peak-end rule hypotheses states that the most recent and the most extreme affective event are most salient for retrospective judgments. While in some studies this was found to explain the variance of the judgments [40], other researchers disagree [41]. It is suggested by Ariely and Carmon [42] that this was due to the recent exposure to affective peaks moderating the judgments. Recent work by Steffens and Guastavino on soundscape pleasantness [41] suggested that the best predictors might be a combination of the average instantaneous rating and the trend over the same judgments (modeled by a linear regression). The rationale is that the linear regression models the expectation of how the soundscapes will evolve in the future.

In summary, there is no agreement about exactly how best to model how humans aggregate sensory judgments over longer periods of time, and consequently this study simply averages the results from each one-second window over the whole sample.

Comparing a HASQI value formed from the whole 10 second sample, with the average of the one-second HASQI values reveals a 95% confidence limit of ±0.16. By weighting the one second HASQI values according to the rms over the one second window reduces the error to ±0.13. Consequently, the weighting by frame rms is adopted to give $bHASQI_A$, the aggregated single-ended HASQI estimate. The formulation is:

$$bHASQI_A = \frac{1}{4}\left(\frac{\sum_{i=1}^{M}(ClassS_i \cdot rms_i)}{\sum_{i=1}^{M}(rms_i)} - 1\right) \qquad (7)$$

where M is the total number of windows, $ClassS_i$ is the single-ended estimate of the HASQI class over window i and rms_i is the root mean square value over window$i$.

Fig. 4 compares $bHASQI_A$ with HASQI integrated over the whole 10-second clip. This dataset was computed using 10-fold cross validation and each of the 10 folds of the cross-validation are overlaid in Fig. 4 (all types of audio and distortion). The Pearson correlation coefficient is 0.97

Table 8. Confusion matrix for all results in one-second windows. Correct HASQI (ClassD) verses single-ended estimation (ClassS).

		Correct ($ClassD$)				
		Bad	Poor	Fair	Good	Excellent
Single-ended estimate ($ClassS$)	Bad	678	77	4	0	0
	Poor	89	502	138	24	7
	Fair	13	148	412	156	31
	Good	2	5	103	427	222
	Excellent	4	2	5	141	609

Table 9. Confusion matrix for DRC cases. One-second windows. Correct HASQI (ClassD) verses single-ended estimation (ClassS).

		Correct ($ClassD$)				
		Bad	Poor	Fair	Good	Excellent
Single-ended Estimate ($ClassS$)	Bad	0	0	0	0	0
	Poor	0	0	0	0	1
	Fair	1	1	1	10	10
	Good	0	1	1	95	94
	Excellent	1	2	3	65	386

Table 10. Single-ended aggregated estimate of quality, ($bHASQI_A$), verses correct, 10 second, value of HASQI. Aggregation over ten-seconds ($HASQI_{10s}$).

		$HASQI_{10s}$				
		Bad	Poor	Fair	Good	Excellent
Single-ended Estimate ($bHASQI_A$)	Bad	234	19	1	0	0
	Poor	4	40	11	2	0
	Fair	0	9	25	9	2
	Good	0	1	15	71	53
	Excellent	0	0	3	63	355

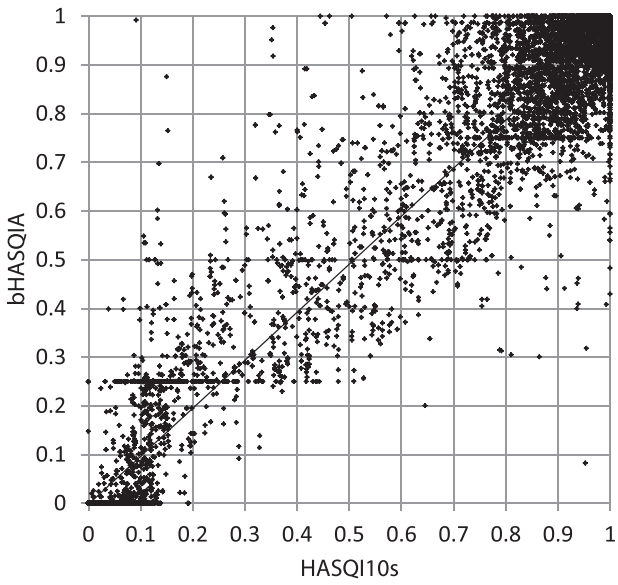

Fig. 4. Estimate of single-ended aggregated HASQI (bHASQIA) verses HASQI calculated using a double-ended method over 10 second (HASQI10s)

and 95% of the estimates are within ± 0.19 of HASQI, with previous results indicating much of this error is due to the aggregation. If $bHASQI_A$ is quantized into five classes, using the specifications in Table 5, the MCC is 0.7; Table 10 displays the averaged confusion matrix for this result. Seventy-nine percent of HASQI classes are correctly identified by the single-ended method, and for those incorrectly identified 95% of those are wrong by a single class. The Pearson correlation coefficient is likely inflated due to the presence of clusters of data near the origin and the top right corner of Fig. 4. The MCC, however, is a balanced measure of classifier performance and is immune to this inflation. Fig. 4 exhibits some quantization of the $bHASQI_A$ results around 0 0.25, 0.5 and 0.75 and 1, this is due to all windows in a sample having the same estimated $ClassS$.

3 SUBJECTIVE VALIDATION

For the single-ended method, HASQI was an intermediate tool to generate a large number of training and testing samples. How does this relate to perceived quality? Since HASQI has been extensively validated on speech, the focus of the subjective validations in this project has been music and soundscapes. Excerpts of music and soundscapes were distorted by varying amounts of hard clipping and then presented to subjects for subjective quality ratings. The perceptual results were compared with correct HASQI value and the single-ended estimate, $bHASQI_A$.

3.1 Music

A small number of music samples, which somehow had to represent the diversity of all music, were needed. As the primary effect of distortion is to change the timbre, it was decided to select the test samples based on music with contrasting timbre. First a large number of music samples were gathered. Three-hundred-fifty-one music extracts were taken from an exemplar set of music samples suggested by Rentfrow and Gosling [43]. For each of the 117 pieces for which high quality recordings could be obtained, three 7-second excerpts representing key sections such as an intro, verse, and chorus were extracted. Additionally, each of the three music samples used by Arehart et al. [44] to develop HASQI were also included in the test set.

Then a method was devised to extract contrasting timbre examples from the hundreds of excerpts. The samples were distorted by hard clipping, using a threshold set to give a HASQI value of 0.5 for each sample. Each stereo example was sampled at 44.1 kHz (all HASQI values averaged over both channels). All samples, clean and distorted, were clustered according to their timbre using the method by Autocoutrier and Pachet [45]. Two samples were drawn from each of the six clusters. They were drawn by selecting the two with the shortest Euclidian distance to the cluster centres. Additionally, each of the three music samples used by Arehart et al. [44] were also included, regardless of which cluster they had grouped within. The 14 pieces from which the test stimuli were taken are listed in Table 11.

3.1.1 Perceptual Test Design

A total of 30 participants (mean age: 23.7 years; SD: 4.7 years) completed the experiment. None reported any known hearing impairments. Each participant was presented with 140 7-second clips that consisted of 9 different thresholds of hard clipping distortion and 1 clean for each of the 14 music pieces. All samples were presented in stereo at the same A-weighted sound pressure level, integrated over 7 seconds and both channels, over Sennheiser 650 HD headphones, via a Focusrite Scarlett 2i4 audio interface (this having

Table 11. The 14 music pieces the final test samples were taken from by cluster number '*' denotes sample used to develop HASQI

Cluster Number	Song Name	Artist/Composer	Publisher / product code
1	Riverboat Set: Denis Dillon's Square Dance Polka, Dancing on the Riverboat	John Whelan	Narada Lotus – ND-61060
	Crazy Train	Ozzy Osborne	Sony Music - 88697738182
	"Haydn" - Symphony in C Major, Hob. I: 82, The Bear: III. Menuet – Trio *	* Haydn	Sony Music – SX10k89750
2	Ave Maria	Franz Schubert	Phillips – 412 629-2
	Packin' Truck	Leadbelly	Saga – 982 076-7
	"vocalise" * Ding Dong the witch is dead	Tierney Sutton	Telarc Jazz – cd 83548
3	Kalifornia	Fatboy Slim	Skint – Brassic 11CD
	Brown Sugar	The Rolling Stones	Polydor – lc000309. 0602527015620
4	The Four Seasons: Spring	Antonio Vivaldi	EMI – 7243 5 56253 2 8
5	For What It's Worth	Buffalo Springfield	Acto 7567 90389 2yg
	The Girl From Ipanema	Stan Getz	Verve lc 00383
6	Spoonful	Howlin' Wolf	Universal – 329 375-2
	Nobody Loves Me But My Mother	B.B. King	Geffen records b0003854-02
	"jazz" * Corcovado	*	Verve lc 00383

previously been calibrated using a dummy head). Playback level was calibrated by setting the playback of the clean *Jazz* excerpt to 72 dB (linear, average of both channels), which meant samples were reproduced at an A-weighted L_{eq} of 62 dB, as this was the level used by Arehart et al. [44].

To ensure that the distortion applied to each music sample covered a wide range of quality degradations, nine thresholds for each clip were computed by setting target HASQI values between 0.1 and 1. A participant training session was held before the actual testing with three pairs of samples not included in the test. Participants were reminded that they were judging overall quality not any musical preference.

Ratings were entered via a mouse using a continuous slider labelled "Bad" and "Excellent" at each endpoint with no other markers based on the ITU-R BS.1284-1 [29] recommendations adopted in the development of HASQI [29]. Participants were asked to make absolute quality judgments on individual samples with no reference. The use of relative judgments of quality using a reference sample was not adopted for the following of reasons;

1) HASQI was also developed using absolute category ratings and a direct comparison was important.
2) One of the research questions in [21] from which some of this data is based was: is there any link between the underlying quality of a sample and the degradation due to amplitude clipping?
3) A high priority was placed on maximizing the number of music pieces and soundscapes to increase the validity of the resulting algorithm performance analysis. The large number of samples made the use of an impairment scale time prohibitive.

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. There were no limits on the number of times each sample could be repeated. 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 randomized. The test session typically lasted around 40 minutes and participants were financially reimbursed for their time.

Table 12. The 12 examples of soundscapes [46] used along with their crest factors.

Tag and ID	Crest Factor
'ambience_28252'	4.39
'beach_48412'	13.9
'car_50378'	6.95
'church_151381'	5.84
'crowd_160041'	6.92
'crowd_25522'	4.81
'forest_184201'	18.3
'machine_146211'	14.7
'nature_150888'	19.1
'rain_55512'	8.62
'thunder_169255'	5.05
'zoo_104483'	5.75

3.2 Validations with Soundscapes

Twelve sound samples (field recorded soundscapes) were selected from the freefield1010 database [46], which was a selection of ten-second audio clips uploaded to the freesound.com database and tagged as "field-recording." First, the 20 most popular tags were identified and all files with those tags were used. Then, the crest factors were computed. The crest factor is the ratio of the peak to the rms level. A signal with a low crest factor will exhibit fairly constant levels of clipping while a signal with a high crest factor might have some highly distorted regions while other regions may remain relatively clean. Four examples closest to the 10^{th}, 50^{th}, and 90^{th} percentiles of the crest factor distribution were selected and are listed in Table 12. The perceptual test procedure was the same as that used for the music clips—18 subjects participated in the test.

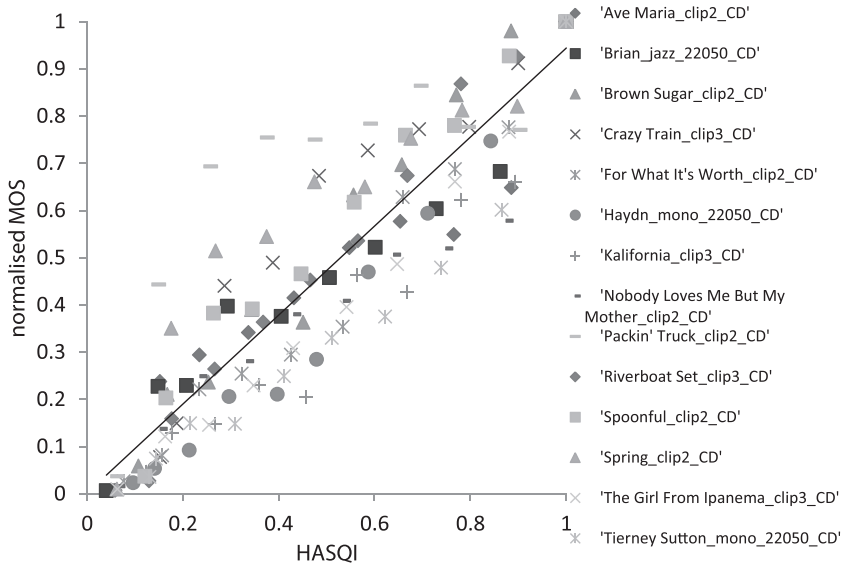

Fig. 5. Double-ended HASQI verses normalized MOS of quality for 14 pieces of music degraded by hard clipping at different thresholds

3.3 Results

For the music clips, Cox et al. [21] found that the MOS (Mean Opinion Score) of even the clean samples varied considerably because of different styles of audio production for the originals. As the interest is in distortions that degrade the quality, the MOS scores were normalized to the averaged MOS calculated from all subjects for the clean undistorted signals of a particular audio file. The standard deviation of the opinion scores for each clip and distortion condition provides a gauge of the intersubject variability of opinion; the average standard deviation for all conditions was 0.17.

Fig. 5 shows relationship between double-ended HASQI (x-axis) and the normalized MOS (y-axis); the Pearson's correlation coefficient is 0.916. The results seem to be more promising than Arehart et al. [44] report. Their correlation between HASQI and the MOS for three pieces of music was 0.838. The better correlation found in our experiments might be attributed to the fact that only clipping and DRC were considered. Ninety-five percent of the HASQI estimates are within ± 0.24 of the normalized MOS.

A few samples showed relatively large prediction errors. For example, "Packin' Truck" has HASQI overestimating the MOS by up to 40%. This track was recorded in 1935 and the recording quality is poor with noise and distortion already present. There appears to be some leniency in quality ratings of degraded audio when the expected technical quality of the original audio is already low.

For the soundscape samples there was an increase in the variability of the opinion scores compared with music, the standard deviation of the opinion scores was 0.29; this can be seen in Fig. 6. This increase in variability may be due to the smaller number of listeners (18 rather than 30). Despite this increase in the variability of opinion, the correlation of HASQI and the normalized MOS yields a correlation coefficient of 0.85 with 95% of HASQI estimates within ± 0.29 of the normalized MOS.

For soundscapes, HASQI over-estimated the level of degradation for two clips in particular. These two clips contained mainly high frequency bird and insect sounds. There were also cases where HASQI under-estimated the degradation, such as thunder, rain, and machinery sounds. These clips differentiate themselves from the others as they do not contain harmonic sounds. It is likely that the reason for the lower performance with soundscapes is that HASQI was primarily aimed at speech quality during development and naturally performs better on such cases.

Next, the proposed single-ended algorithm was trained using every sample from the audio library described in Sec. 1.1.1 excluding those used in the perceptual studies. Figs. 7 and 8 show the relationship between the normalized MOS and the single-ended estimates, $bHASQI_A$, for music and soundscapes. For music the correlation coefficient between $bHASQI_A$ and the normalized MOS is 0.861 and 95% of the single-ended estimates of $bHASQI_A$ are within ± 0.3 of the MOS. For the soundscapes, similar results are found, with the correlation coefficient between HASQI_A and the normalized MOS being 0.802 and 95% of the estimates are within ± 0.33 of the MOS.

As previously mentioned, the average standard deviation of the opinion scores for each clip gives an estimation of the intersubject variability. This was 0.17 for music and 0.29 for soundscapes. The intersubject variability and the error in the single-ended estimation of quality can be compared using the standard deviation of the error in the MOS estimation using $bHASQI_A$. This was 0.17 for both music and soundscapes. This shows that on average the error in the single-ended estimate of quality for a single clip is of the same order, or lower than, the intersubject variability of opinion.

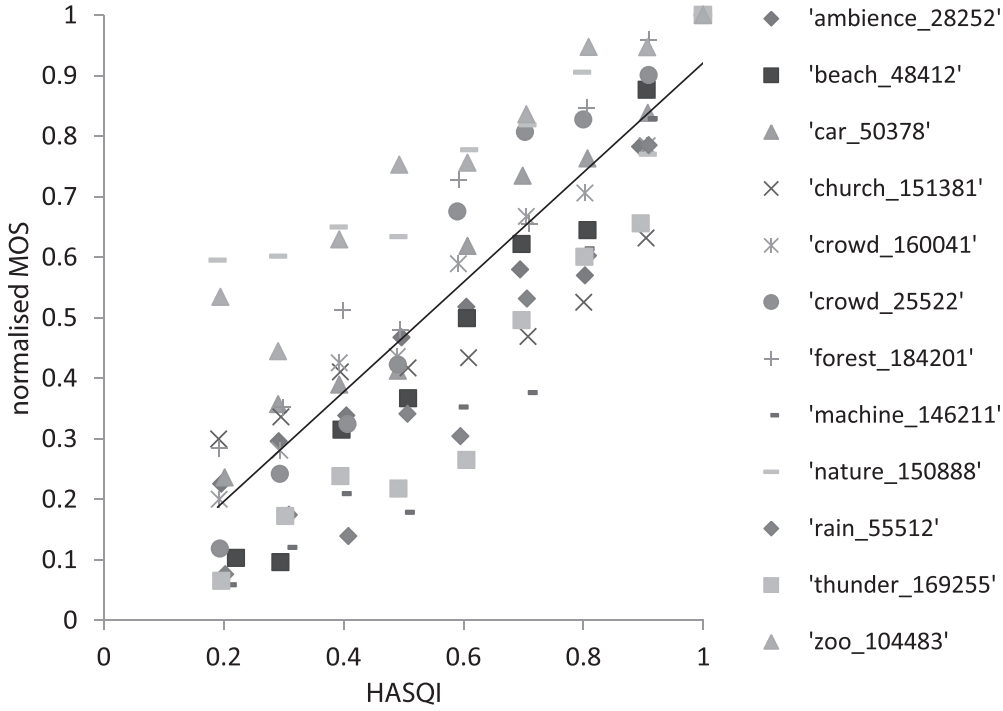

Fig. 6. Double-ended HASQI verses normalized MOS of quality for 12 soundscape clips degraded by hard clipping at different thresholds

4 CONCLUSION

A single-ended method to quantify perceived audio quality in the presence of non-linear distortions has been developed and presented in this paper. This single-ended method estimates HASQI (Hearing Aid Sound Quality Index). The model uses machine learning to learn from examples and generalize. Validations on a set of music and soundscapes not seen during training, yield single-ended estimates within ±0.19 of HASQI, using a quality range

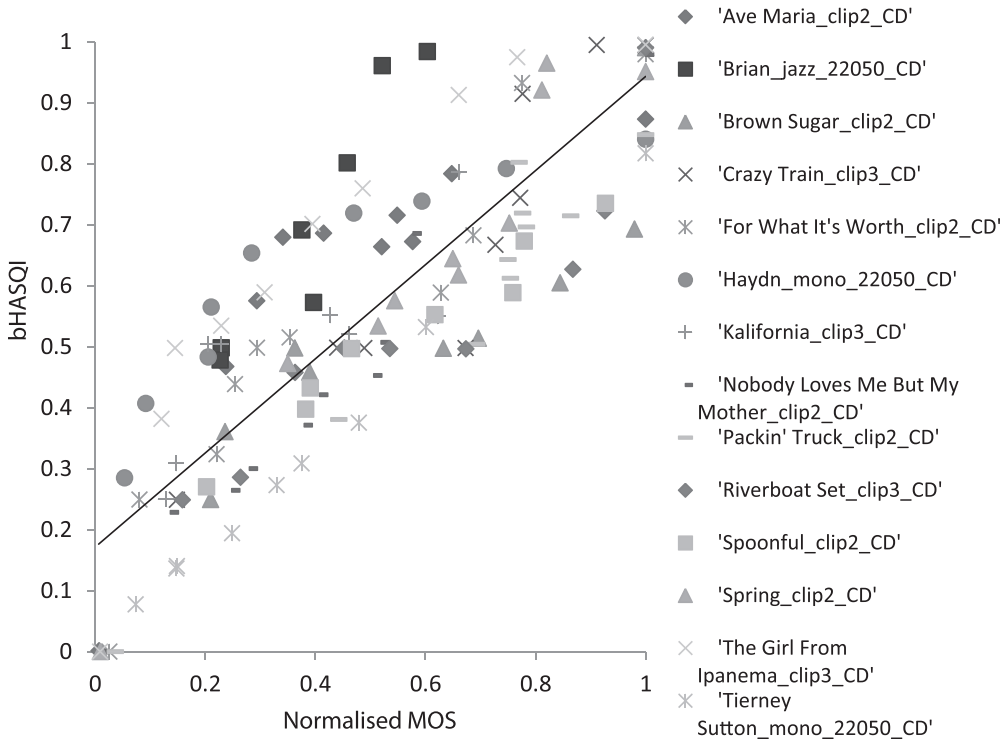

Fig. 7. Single-ended quality estimate (bHASQIA) verses normalized MOS of quality for 14 pieces of music degraded by hard clipping at different thresholds

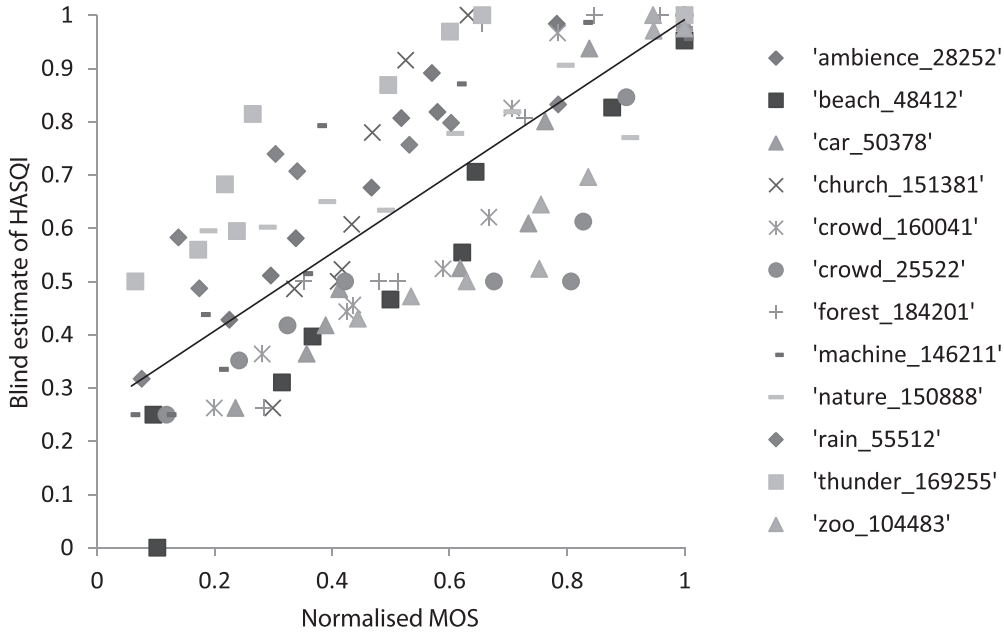

Fig. 8. Single-ended quality estimate (bHASQIA) verses normalized MOS of quality for 12 soundscapes degraded by hard clipping at different thresholds

between 0.0 and 1.0. HASQI has also been shown to predict quality degradations for processes other than non-linear distortions including additive noise, linear filtering, and spectral changes. By including these other causes of quality degradations, the current model for non-linear distortion assessment might be expanded, although additional features and validation would be required.

A series of perceptual measurements on music and soundscapes were undertaken. The subjective testing provided more data that shows that HASQI can be used to quantify perceived non-linear distortion for normal hearing listeners. The new single-ended method was used to estimate quality and compared to the Mean Opinion Scores (MOS) from the subjective tests. The standard deviation of the error in the single-ended MOS estimations was 0.17. This is of a similar order to the standard deviation of human subjects: the standard deviation of the MOS from the perceptual tests was for music was 0.17 and 0.29 for music and soundscapes respectively.

The code to estimate bHASQI is freely available for download at [47] for non-commercial purposes under an Attribution-NonCommercial 4.0 International (CC BY-NC) license. The databases used to develop the algorithm are not available due to copyright issues with the audio samples.

5 ACKNOWLEDGMENTS

This project is funded by Engineering and Physical Science Research Council, UK (EPSRC EP/J013013/1) and carried out in collaboration with the BBC R&D and the British Library Sound Archive. The perceptual tests were carried out by Stephen David Groves-Kirkby. This work is published under a CC-BY license (http://creativecommons.org/licenses/by/3.0/).

6 REFERENCES

[1.] C. Wardle, S. Dubberley, and P. Brown, "Amateur Footage: A Global Study of User-Generated Content in TV and Online-News Output" (2014). [Online]. Available: http://towcenter.org/wp-content/uploads/2014/04/80458_Tow-Center-Report-WEB.pdf. [Accessed: 26-Nov-2014].

[2.] I. Jackson, "What You Told Us about Recording Audio: An Overview of Our Web Survey," The Good Recording Project Blog (2012). [Online]. Available: http://www.goodrecording.net/211/. [Accessed: 20-Nov-2012].

[3.] I. R. Jackson, P. Kendrick, T. J. Cox, B. M. Fazenda, and F. F. Li, "Perception and Automatic Detection of Wind-Induced Microphone Noise," *J. Acous. Soc. Am.*, vol. 136, no. 3, p. 1176 (2014). http://dx.doi.org/10.1121/1.4892772

[4.] "Sound System Equipment—Part 5 Loudspeakers," BS EN 60268-5 (2009).

[5.] "Measurement of Intermodulation Distortion in Audio Systems," SMPTE Recommended Practice RP 120:2005 (2005).

[6.] R. Small, "Total Difference-Frequency Distortion: Practical Measurements," *J. Audio Eng. Soc.*, vol. 34, no. 6, pp. 427–436 (1986 June).

[7.] "Sound System Equipment—Part 3 Amplifiers," BS EN 60268-3 (2001).

[8.] L. Lee and E. Geddes, "Auditory Perception of Non-linear Distortion," presented at the *115th Convention of the Audio Engineering Society* (2003 Oct.), convention paper 5891.

[9.] E. Geddes and L. Lee, "Auditory Perception of Non-linear Distortion-Theory," presented at the *115th Convention of the Audio Engineering Society* (2003 Oct.), convention paper 5890.

[10.] A. W. Rix, M. P. Hollier, A. P. Hekstra and J. G. Beerends, "PESQ, the New ITU Standard for Objective Measurement of Perceived Speech Quality, Part I—Time Alignment," *J. Audio Eng. Soc.*, vol. 50, pp. 755–764 (2002 Oct.).

[11.] J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier, "PESQ, the New ITU Standard for Objective Measurement of Perceived Speech Quality, Part II—Perceptual Model," *J. Audio Eng. Soc.*, vol. 50, pp. 765–778 (2002 Oct.).

[12.] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullman, J. Pomy and M. Keyhl, "Perceptual Objective Listening Quality Assessment (POLQA), The Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part I—Temporal Alignment," *J. Audio Eng. Soc.*, vol. 61, pp. 366–384 (2013 June).

[13.] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullman, J. Pomy and M. Keyhl, "Perceptual Objective Listening Quality Assessment (POLQA), The Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part II—Perceptual Model," *J. Audio Eng. Soc.*, vol. 61, pp. 385-402 (2013 June).

[14.] "Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks," ITU P. 862 (2001).

[15.] "Perceptual Objective Listening Quality Assessment," ITU-T P.863 (2011).

[16.] T. Thiede, W. Treurniet, and R. Bitto, "PEAQ—The ITU Standard for Objective Measurement of Perceived Audio Quality," *J. Audio Eng. Soc.*, vol. 48, pp. 3–29 (2000 Jan./Feb.).

[17.] 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.).

[18.] 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.).

[19.] J. Kates and K. Arehart, "The Hearing-Aid Speech Quality Index (HASQI)," *J. Audio Eng. Soc.*, vol. 58, pp. 363–381 (2010 May).

[20.] A. A. Kressner, D. V Anderson, and C. J. Rozell, "Evaluating the Generalization of the Hearing Aid Speech Quality Index (HASQI)," *IEEE Trans. Audio. Speech Lang. Processing*, vol. 21, no. 2, pp. 407–415 (2013). http://dx.doi.org/10.1109/tasl.2012.2217132

[21.] T. Cox, B. Fazenda, S. Groves-Kirkby, I. Jackson, P. Kendrick, and F. Li, "Quality Timbre and Distortion: Perceived Quality of Clipped Music," in *30th Anniversary Conference Reproduced Sound Oct. 14–15* (2014).

[22.] "Single-Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony Applications," ITU P. 563 (2004).

[23.] S. Mare, "Detection of Nonlinear Distortion in Audio Signals," *Broadcast. IEEE Trans.*, vol. 48, no. 2, pp. 76–80 (2002). http://dx.doi.org/10.1109/tbc.2002.1021270

[24.] M. Massberg, "Investigation in Dynamic Range Compression," MSc dissertation, Queen Mary University of London (2009).

[25.] P. E. Souza, L. M. Jenstad, and K. T. Boike, Measuring the Acoustic Effects of Compression Amplification on Speech in Noise," *J. Acoust. Soc. Am.*, vol. 119, no. 1, p. 41 (2006). http://dx.doi.org/10.1121/1.2108861

[26.] P. Kendrick, S. Groves-kirkby, I. Jackson, T. Cox, and B. Fazenda, "Measuring a Portable Audio Device's Response to Excessive Sound Levels," Internal report, Salford (2013). Available: http://usir.salford.ac.uk/29371/

[27.] B. De Man and J. D. Reiss, "Adaptive Control of Amplitude Distortion Effects," presented at the *53rd AES International Conference: Semantic Audio* (2014 Jan.), conference paper P2-9.

[28.] D. Giannoulis, M. Massberg, and J. Reiss, "Digital Dynamic Range Compressor Design—A Tutorial and Analysis," *J. Audio Eng. Soc*, vol. 60, pp. 399–408 (2012 June).

[29.] "General Methods for the Subjective Assessment of Sound Quality," ITU 1284-1 (1997).

[30.] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 1–33 (2001).

[31.] Matlab, *Matlab:2013b* (Natick, MA, The MathWorks Inc. 2013).

[32.] O. Lartillot, P. Toiviainen, and T. Eerola, "MIRtoolbox" (2014).

[33.] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection 1 Introduction," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182 (2003).

[34.] G. Jurman, S. Riccadonna, and C. Furlanello, "A Comparison of MCC and CEN Error Measures in Multi-Class Prediction," *PLoS One*, vol. 7, no. 8, p. e41882 (2012 Jan.). http://dx.doi.org/10.1371/journal.pone.0041882

[35.] C. Strobl, T. Hothorn, and A. Zeileis, "2009 Party On! A New, Conditional Variable Importance Measure for Random Forests Available in the Party Package," Technical Report Number 050, Department of Statistics, University of Munich (2009).

[36.] P. Latinne, O. Debeir, and C. Decaestecker, "Limiting the Number of Trees in Random Forests," in *Proceedings of MCS, LNCS 2096* (2001), pp. 1–10. http://dx.doi.org/10.1007/3-540-48219-9_18

[37.] W. A. Sethares, *Tuning, Timbre, Spectrum, Scale* (Springer-Verlag, 1998). http://dx.doi.org/10.1007/b138848

[38.] K. Dittrich and D. Oberfeld, "A Comparison of the Temporal Weighting of Annoyance and Loudness," *J. Acoust. Soc. Am.*, vol. 126, no. 6, pp. 3168–78 (2009 Dec.). http://dx.doi.org/10.1121/1.3238233

[39.] D. Västfjäll, "The 'End Effect' in Retrospective Sound Quality Evaluation," *Acoust. Sci. Technol.*, vol. 25, no. 2, pp. 170–172 (2004). http://dx.doi.org/10.1250/ast.25.170

[40.] A. Rozin, P. Rozin, and E. Goldberg, "The Feeling of Music Past: How Listeners Remember Musical Affect," *Music Percept.*, vol. 22, no. 1, pp. 15–39 (2004). http://dx.doi.org/10.1525/mp.2004.22.1.15

[41.] J. Steffens and C. Guastavino, "(Tr-)end Effects of Momentary and Retrospective Soundscape Evaluations," *Acta Acust. united with Acust.*, vol. 98 (2014).

[42.] D. A. N. Ariely and Z. I. V Carmon, "Gestalt Characteristics of Experiences: The Defining Features of Summarized Events," *J. Behav. Dec. Mak.*, vol. 201, pp. 191–201 (2000). http://dx.doi.org/10.1002/(sici)1099-0771(200004/06)13:2<191::aid-bdm330>3.3.co;2-1

[43.] P. J. 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.*, vol. 84, no. 6, pp. 1236–1256 (2003). http://dx.doi.org/10.1037/0022-3514.84.6.1236

[44.] K. H. Arehart, J. M. Kates, and M. C. Anderson, "Effects of Noise, Nonlinear Processing, and Linear Filtering on Percieved Music Quality," *Int. J. Audiol.*, vol. 50, pp. 177–190 (2011). http://dx.doi.org/10.3109/14992027.2010.539273

[45.] J. Aucouturier and F. Pachet, "Music Similarity Measures: What's the Use," *Proc. Conference of the International Society for Music Information Retrieval (ISMIR)* (2002).

[46.] D. Stowell, and M. D. Plumbley, "An open dataset for research on audio field recording archives: freefield1010", submitted. http://arxiv.org/abs/1309.5275.

[47.] P. Kendrick, "Distortion and Clipping, C ++program for Automatic Detection and Metering" (2015). [Online]. Available: http://usir.salford.ac.uk/35954/.

APPENDIX 1 DESCRIPTION OF PARAMETER DISTRIBUTIONS FOR CLIPPING FUNCTION

The parameters using in the clipping model described in Eq. (1), T(Threshold), K(Knee type), and B (Bias) were randomly generated for every example. To ensure that the distribution of examples in the resulting database was representative, a number of rules were applied to the generating functions:

- A nonlinear distribution was chosen for the threshold T so that there was a roughly even distribution of samples along the HASQI scale. $T = x^{1.5}$ was used where x is a uniformly distributed random number between 0 and 1.
- Half of all the examples were assigned a hard knee ($K = 1$) and the other half a soft knee ($K > 1$) to simulate the different types of clipping that may occur.
- When a soft knee is selected, half of these were generated using a modest smoothing parameter, where K varies uniformly between 1 and 2, effectively this smooths just the transition region in the amplitude transfer function. For the other half $K$was varied uniformly between 1 and 101, to ensure some extreme examples were present.
- Bias is avoided in mobile devices but may occur in some poorly designed devices; for this reason half of all examples had no bias ($B = 0$). To ensure some more extreme examples were present, the other half was generated so that B was uniformly distributed between –0.5 and 0.5.

APPENDIX 2 DESCRIPTION OF PARAMETER DISTRIBUTIONS FOR DRC FUNCTION

The parameters using in the DRC models described in Eqs. (2)–(5) are: T(Threshold dB), τ_a (attack time, s), τ_r (release time, s), R (Compression ratio), and the DRC model type. These were randomly generated for every example. To ensure that the distribution of examples in the resulting database is representative, a number of rules are applied to the generating functions.

- The threshold T was varied uniformly between 0 dB and –40 dB; this represents a range of realistic cases as well as some extreme examples.
- For the attack and release times, Table 1 describes the range of parameters commonly found in mobile devices; therefore the attack time (τ_a) is varied uniformly between 0.1 ms and 20.1 ms. The release time (τ_r) is varied uniformly between 0 and 500ms.
- For the Compression ratio$R$, half of examples used a value of infinity to represent limiting, and the other half used a finite value to represent compression, for compression examples R was varied uniformly between 0 and 40.
- Finally, equal numbers of each of the four different models of compression was ensured.

THE AUTHORS


Paul Kendrick


Francis Li


Bruno Fazenda


Iain Jackson


Trevor Cox

Dr. Paul Kendrick is a lecturer in broadcast engineering at the University of Salford. One of his primary research themes is in Machine Audition, or the use of algorithms to analyze and extract meaning from recorded audio. After receiving a B.Eng. in electronic engineering from the University of Manchester in 2001, Paul then completed an M.Sc. in audio acoustics from the University of Salford in 2003. Paul received a Ph.D. from the University of Salford in 2009 that developed ways to estimate reverberation times using only speech or music with no reference to the clean signal. Paul has interests in signal processing, machine learning, and bioacoustics.

•

Dr. Francis Li embarked on an academic career in 1986 at Shanghai University. He then moved to the UK and completed a Ph.D. in statistical signal processing and machine learning applied to architectural acoustics. Appointed senior lecturer in computing at Manchester Metropolitan University in 2001, he then joined Salford as a senior lecturer in acoustical signal processing in 2006. His major research areas are signal processing and computational intelligence applied to acoustics, speech and audio technology, broadcast engineering, machine audition, and biomedical engineering. Francis has published over 100 research papers. He is an Associate Technical Editor for the *JAES*.

•

Bruno Fazenda received a B.Sc. (1st Hons.) degree in audio technology in 1999 and a Ph.D. degree in 2004 for his thesis on the perception of room modes, both from the University of Salford, UK. He worked for a short while as a research fellow with a Marie Curie research fellowship at the Danish Technical University before becoming a lecturer at the Universities of Glamorgan and then Huddersfield. He now lectures in the acoustics and audio area at the University of Salford. His research interests span from room acoustics, particularly the perception in critical listening spaces, to sound quality, condition monitoring, and archaeo-acoustics. He is a member of the Audio Engineering Society.

•

Dr. Iain Jackson is a research associate in the School of Psychological Sciences at the University of Manchester. He is a member of the ESRC International Centre for Language and Communicative Development (LuCiD) research group, investigating relationships between infants' visual attention and language development. Iain completed his Ph.D. in 2010 at the University of Manchester, investigating infant perception and cognition using eye tracking and pupillometry. In 2012 he joined the Acoustics Research Centre in the University of Salford as part of the EPSRC-funded Good Recording Project, exploring the influence of common recording errors on the perception of quality in recorded audio. Iain's research interests include psychoacoustics, visual perception and cognition, and cognitive development.

•

Trevor Cox is professor of acoustic engineering at the University of Salford and a past president of the UK's Institute of Acoustics (IOA). He was award the IOA's Tyndall Medal in 2004. One major strand of his research is room acoustics for intelligible speech and quality music production and reproduction. Trevor's diffuser designs can be found in rooms around the world. Trevor is also a BBC radio presenter and author. He was given the IOA award for promoting acoustics to the public in 2009. He has presented 21 documentaries for BBC radio including: Life's Soundtrack, Green Ears, and The Physicist's Guide to the Orchestra. His popular science book is *Sonic Wonderland* (Bodley Head, UK) (The Sound Book, W W Norton in USA).