Initial Objective & Subjective Evaluation of a Similarity-Based Audio Compression Technique

Stuart Cunningham, Jonathan Weinel, Shaun Roberts, Vic Grout, and Darryl Griffiths
Creative & Applied Research for the Digital Society (CARDS)
Glyndŵr University
Plas Coch Campus, Mold Road, Wrexham, LL11 2AW, North Wales, UK
+44(0)1978 293583
{s.cunningham | j.weinel | roberts.s | v.grout | griffiths.d}@glyndwr.ac.uk

ABSTRACT

In this paper, we undertake an initial study evaluation of a recently developed audio compression approach; Audio Compression Exploiting Repetition (ACER). This is a novel compression method that employs dictionary-based techniques to encode repetitive musical sequences that naturally occur within musical audio. As such, it is a lossy compression technique that exploits human perception to achieve data reduction.

To evaluate the output from the ACER approach, we conduct a pilot evaluation of the ACER coded audio, by employing both objective and subjective testing, to validate the ACER approach. Results show that the ACER approach is capable of producing compressed audio that varies in subjective and objective and quality grades that are inline with the amount of compression desired; configured by setting a similarity threshold value. Several lessons are learned and suggestion given as to how a larger, enhanced series of listening tests will be taken forward in future, as a direct result of the work presented in this paper.

Categories and Subject Descriptors

E.4 [**Data**]: Coding and Information Theory - *Data compaction and compression.*

General Terms

Algorithms, Measurement, Performance, Experimentation, Human Factors, Verification.

Keywords

Compression, music, repetition, subjective, objective, evaluation.

1. INTRODUCTION

A great deal of work has been done in the field of music structure discovery [1], [2], [3], similarity between pieces of music [4], [5], [6], and classification of music [7], [8], [9]. However, although such research has advanced musicology and contributed to the improvement of processing in terms of music recommendation systems and dealing with large music libraries, little attention, aside from the work of Kirovski and Landau [10], has been given to what might be achieved in the field of compression by exploiting musical structure and repetitive characteristics found within musical audio. Indeed, fundamental, syntactic data compression can usually be achieved whenever source data exhibits patterns [11], [12].

ACER is a technique developed to allow data compression to take place by identifying and removing perceptually redundant and/or irrelevant passages of music within a larger piece. Although the technique has been implemented, it has not yet been fully evaluated and there are several options available to do so: objective methods and subjective methods. The main aim of this paper is to explore both techniques of validation and determine their suitability to be used in a large, full-scale evaluation of the ACER technique.

The remainder of the paper is structured as followed: in section 2 we provide a brief outline of the ACER method and background; in sections 3 and 4 we discuss the procedure used to obtain quality evaluations of the ACER compressed audio for objective and subjective techniques respectively; results are presented in section 5; and finally, in section 6, we present conclusions from these initial experiments and discuss our future work intentions.

2. AUDIO COMPRESSION EXPLOITING REPETITION (ACER)

The ACER approach achieves data reduction in musical audio by exploiting the presence of repetitive or similar sequences, which occur on multiple occasions in a piece or song. As such, it is a dictionary-based approach, where sequences deemed to be similar to one another are identified by their indices in the array representing the audio and a dictionary kept of their locations and the original sequence. Duplicates are then removed from the sound and the remaining copy of the sequence is inserted, at the decompression stage, to substitute for the original content. Figure 1 illustrates the principles of the ACER method in exploiting the presence of identical and perceptually similar sequences.

In this particular example, we can see that, within some defined threshold of acceptable similarity, it is possible to reduce an original piece of music of 9 frames in length to one that is 3 frames in length. Alternatively, should higher quality be required and only exact matches sought, the original audio can be represented by 4 frames (this example sees the inclusion of frame 4). Of course, the dictionary of match indices has to be stored as part of the ACER file format, but the size of these indices is negligible in comparison to the typical size of a frame of audio, which will often consist of hundreds, if not thousands, of audio samples.

AM '13, September 18 - 20 2013, Piteå, Sweden

ACM 978-1-4503-2659-9/13/09…$15.00.

http://dx.doi.org/10.1145/2544114.2544116

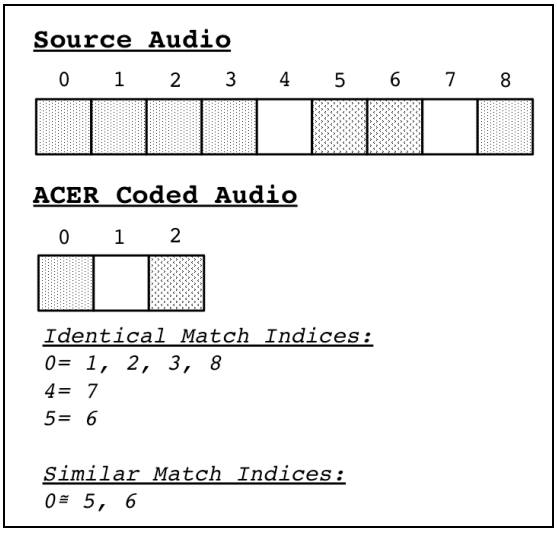

Figure 1: Example of ACER Principles

In practice, ACER functions by splitting the input audio into frames and carrying out a modified exhaustive search. Frames undergo a Hamming windowed Fast Fourier Transform and the difference between the two spectrums is calculated and then averaged to determine a mean difference value. If the difference value falls within a pre-determined similarity threshold, the location of the matching frames is noted and the duplicate(s) removed. The similarity threshold can be varied, depending upon the level of difference between sequences that are appropriate for any given scenario and there is a natural trade-off in the amount of compression obtainable. At present, ACER is configured with 5 levels of perceptual similarity, developed by human listener perceptions. The 5 levels currently used are: Identical; Very Similar; Similar; Dissimilar; and Different. It is anticipated that, in most normal use cases, Dissimilar and Different would not normally be used.

The focus of this work is in the quality evaluation of ACER compressed audio and not the technique itself. For further details on the processes and techniques associated with ACER, the reader is directed to other work from the authors [13], [14], [15].

A distinct advantage of the ACER approach is that the compressed version of the music is stored at the original sample rate and bit-depth, meaning that there is no down sampling or reduction in bit-rate in ACER audio. However, perceptual noise is introduced in the coding which is easiest to conceptualize as the difference between an original passage from the music and the sequence being used to replace it in the ACER version. The remainder of this paper examines and tests methods of evaluating ACER coded audio, in terms of listener satisfaction and quality perception.

3. OBJECTIVE EVALUATION

Objective audio quality evaluations typically take place using software tools, which analyze one or more features of the original (reference) audio and the coded (test) version. These features are normally extracted and then compared to a reference model, based on a psychoacoustic model, to determine the level of impairment or quality degradation. Several objective measures of audio quality are commonly employed in evaluating audio compression techniques. However, one such technique is formalized by an international organization and recommended as best practice, so for the purpose of this research the ITU-R BS.1387: Method for Objective Measurements of Perceived Audio Quality [16] approach is adopted. The ITU-R document behind the standard explains that the use of single, traditional measurements of impairment, such as signal-to-noise ratio (SNR) are ineffective for measuring quality in contemporary audio processing scenarios. As such, ITU-R BS.1387 makes use of a number of Model Output Variables (MOVs), which are calculated during the testing procedure (one of them is SNR) and it is these MOVs which are used to calculate the overall Objective Difference Grade (ODG) for the audio. The range of ODG values output by the software is shown in Table I.

Table I: ITU-R BS.1387 Objective Difference Grades (ODG)

ODG	Impairment Description
0.0	Imperceptible
-1.0	Perceptible, but not annoying
-2.0	Slightly annoying
-3.0	Annoying
-4.0	Very Annoying

Determining ODG for any given audio file is a relatively straightforward and fast process. Many signals can be processed within a short period of time. Another advantage of the approach is that multiple tests per song are not required, unlike the recommended practice in subjective listening tests.

4. SUBJECTIVE EVALUATION

Attempting to measure human perception of anything, and especially audio, is a distinct challenge. Indeed, as Bech and Zacharov [17] point out, being able to directly measure a person's perception of audio stimuli is currently an impossible task, for which technology does not exist, since it would require some kind of direct connection with the neurological workings of the human brain and also the knowledge of how to translate these signals into an understanding of perceptual response. Therefore, it becomes necessary to determine the listener's perception by querying the listener to explain and express their perceptions as a result of a set of tests.

To complement the use of the ITU-R method for objective evaluations in the previous section, here we attempt to broadly comply with the ITU-R BS.1284-1: General Methods for the Subjective Assessment of Sound Quality recommendations [19]. The ITU-R guidelines advise the use of comparative testing where a reference signal is compared to a coded version (AB testing) or comparing a reference signal to a hidden reference and a coded version (ABC testing). In full testing, the ITU-R recommend that a minimum of 10 expert or 20 non-expert listeners are used, audio samples used should be no more than around 15 or 20 seconds in length and not abruptly begin or end, a testing phase is implemented prior to the recorded results being taken, the listening environment should be controlled, and that each testing session should last no more than 15 or 20 minutes. In terms of recording the results of participants, BS.1284-1 recommends the use of an appropriate impairment scale, where the use of either a Subjective Grade (SG) in AB testing or Subjective Difference Grade (SDG) can be attached to each test. The subjective impairment grades are shown in Table II.

However, as an interval scale, the numeric value assigned to each impairment description is somewhat arbitrary, provided the numeric values follow the same linear progression. Conveniently, the impairment descriptions are the same as those used in BS.1397; the objective evaluation recommendations.

Table II: ITU-R BS.1284-1 Subjective Impairment Grades

SG	SDG	Impairment Description
5	0.0	Imperceptible
4	-1.0	Perceptible, but not annoying
3	-2.0	Slightly annoying
2	-3.0	Annoying
1	-4.0	Very Annoying

For the subjective testing reported here, the experiment recruited students from a class being held on sound design. A brief introduction to the purpose of the experiment was given in class, before students were invited to participate. This included an explanation of the nature of the audio compression and the purpose of testing subjective human experience. Listening tests took place on an individual basis in a controlled environment of the University usability lab and were conducted using a pair of Samson studio monitoring speakers. The audio samples were cued by a researcher and played from a laptop computer and played with the VLC media player. For ease of playback, the ACER material was converted to wave, so there was no decoding time. The playback of samples was randomized.

Participants were individually briefed on the testing process and the scale being used. They were then played each A extract (original) followed by each B extract (ACER compressed version). After each AB test the participant was verbally asked to indicate their choice on the scale. All responses were recorded on a spreadsheet using the SG.

5. RESULTS

5.1 Experimental Materials

A small selection of songs was taken from disc 2 of 2 on a popular music compilation album: '*Now That's What I Call Music! 80'* [19]. The compilation was selected since it represents a collection of tracks that have been commercially successful and popular with listeners in the UK. The tracks shown in Table III were selected, at random, from the compilation to be used in this evaluation. Table IV also shows the amount of data reduction (the percentage of the original data that has been removed) achieved by the ACER method per similarity threshold employed.

Table III: Tracks Used in Objective and Subjective Testing

Track	Artist	Song
03	Gym Class Heroes	*Stereo Hearts (Feat. Adam Levine)*
09	Leona Lewis	*Collide (Feat. Avicci)*
13	Sak Noel	*Loca People (UK Radio Edit)*

Table IV: Data Reduction Achieved per ACER Similarity Threshold (SIM)

Track	SIM5	SIM4	SIM3	SIM2	SIM1
03	6.3%	10.0%	23.8%	30.1%	41.4%
09	3.2%	16.1%	23.3%	21.7%	31.3%
13	18.2%	25.2%	33.5%	37.7%	41.9%

It is worth noting that the amount of data reduction ACER achieves varies across the tracks being processed. This is a particular trait of the ACER approach. Since it is searching for repeating musical sequences, the amount of data reduction that can be achieved varies depending upon the amount of repetition present in the actual musical composition that is being analyzed. This has been illustrated in Figure 2, where, using Track 03, the sections ACER has detected as being similar have been highlighted using a lighter shading for each of the five SIM search values.

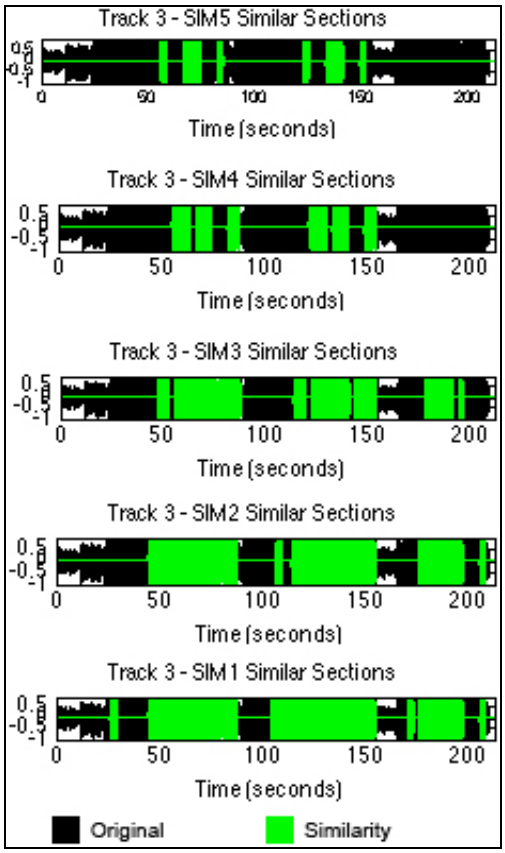

Figure 2: Similar Regions for Track 03 (SIM5-SIM1)

5.2 Objective Evaluation

An implementation of the Perceptual Evaluation of Audio Quality (PEAQ) software tool, based upon the specifications of BS.1397, was used to analyze the test tracks. This particular software tool is a Matlab implementation created by researchers at McGill University [20]. A compressed version was created for each ACER similarity level and then compared to the original, unaltered, reference signal, using the software tool. In the objective evaluations the entire song was used, the stereo track was summed to mono, and the audio was up-sampled from 44100 Hz to 48000 Hz to ensure compatibility with the PEAQ software. The results are shown in Figure 3, Figure 4, and Figure 5.

A separate PEAQ objective difference grade is given for each of the 5 levels of similarity threshold within the ACER system and these are shown in the graphs.

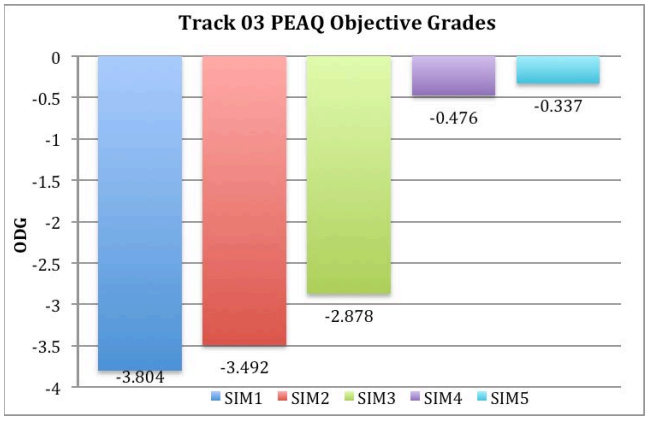

Figure 3: PEAQ Objective Grades - Track 03

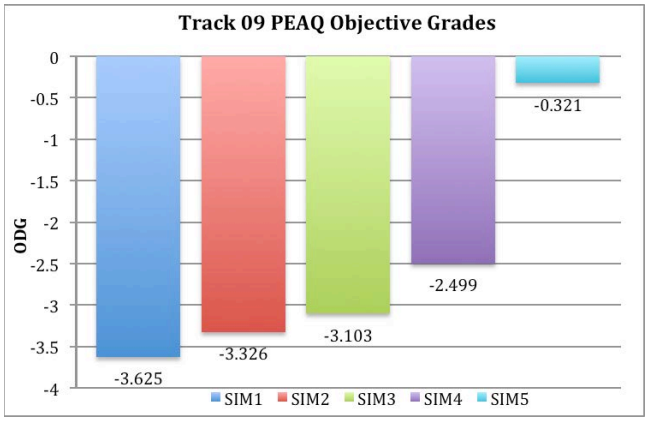

Figure 4: PEAQ Objective Grades - Track 09

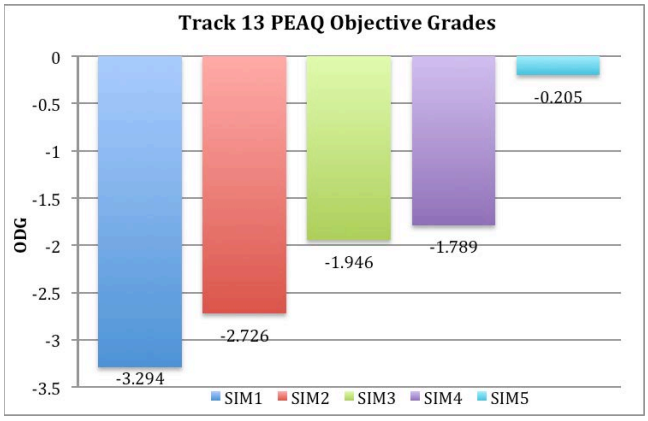

Figure 5: PEAQ Objective Grades - Track 13

5.3 Subjective Evaluation

Participants were played a total of 15 pairs of audio samples, which encompassed the three songs and 5 levels of ACER compression to be evaluated. Each sample was 30 seconds in length and contained a 4 second linear fade in/out at the start and end of the sample. Stereo tracks were summed to mono. As the ACER approach does not necessarily apply across the entirety of an input audio stream, samples had to be carefully selected, ensuring that each sample included at least one sequence that had been processed. The sample from Track 03 was taken between 1m52s – 2m22s; Track 09 sample was between 0m55s – 1m25s; and the sample from Track 13 was taken between 0m54s – 1m24s.

A set period of time of approximately 3 hours was allocated to conduct the experiments, and during this time a total of 8 participants were able to engage with the test. The playlist used was the same for each participant. Each AB pair presented cycled between different songs from the three chosen tracks, (e.g. song 1, song 2, song 3, song 1, song 2...) but with a random level of compression in each instance (produced by manually sorting the tracks into an order that seemed arbitrary). This was done so that participants did not find the evaluation repetitive or irritating and so there was no pattern that might influence the participants' judgment. A separate mean subjective difference grade is given for each of the 5 levels of similarity threshold within the ACER system and these are shown in the graphs.

Although the participant grades were gathered on a reverse SG scale, due to their interval nature, they are represented using SDG values, so as to make direct comparison with the objective measurements more straightforward. The mean results for all participants are presented, per test track, in Figure 6, Figure 7, and Figure 8.

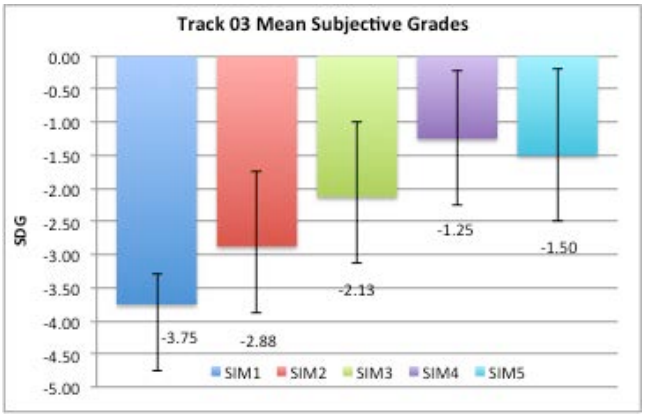

Figure 6: Mean Subjective Difference Grades - Track 03

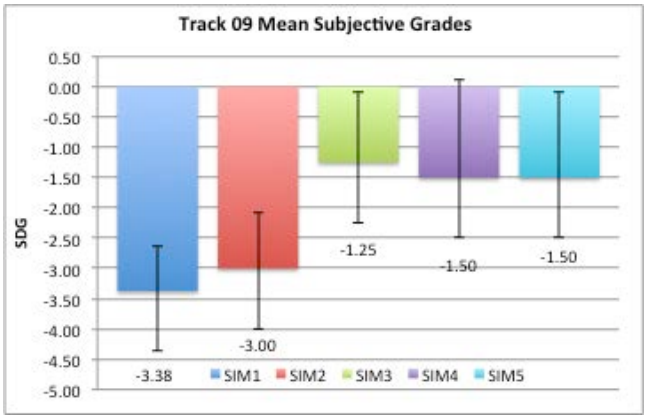

Figure 7: Mean Subjective Difference Grades - Track 09

It is worth noting that one of the eight participants in the subjective test consistently rated all test pairs with a much higher perception of impairment than the other participants and this naturally skews the results somewhat. After testing, the participant was questioned to ensure they had understood the task effectively, which they stated they did (for the sake of benchmarking, the researchers then played this participant two identical pieces, a placebo, and the participant rated this sound as being moderately impaired). In a larger test, it might have been worth removing this participant's data as an outlier. Such use of placebos in a formal testing methodology might be an effective way of highlighting

potential outliers in larger data sets and/or as a benchmark error value.

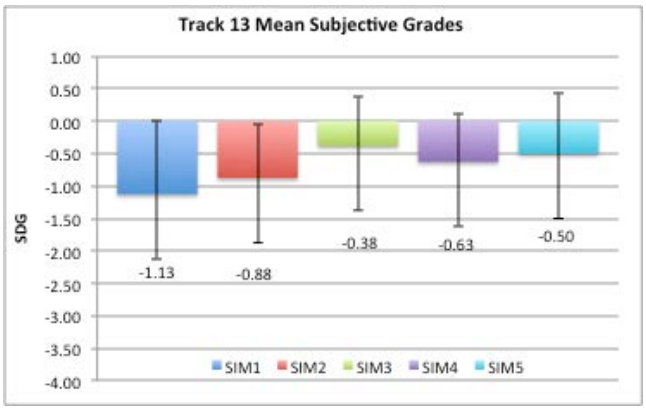

Figure 8: Mean Subjective Difference Grades - Track 13

5.4 Objective and Subjective Comparison

Both the objective and subjective evaluations demonstrate a change in difference grading that is inline with the reduction in audio quality, or proportion of data reduction, implemented by applying the various levels of ACER similarity threshold.
In order to compare the results of the PEAQ objective evaluation and the mean values from subjective testing,

Table V (which appears on the final page of this paper, due to size) tabulates the results per track and per ACER similarity threshold.

To examine the relationship between the subjective and objective results for each of the three experimental tracks, a t-Test was conducted. In doing so, we establish that there is no significant difference in the objective and subjective grading given to Track 03 (*paired t(4)= 0.27, ns, two-tailed*) and Track 09 (*paired t(4)= -0.90, ns, two-tailed*). However, we find that a statistically significant difference exists between the objective and subjective grading given to Track 13 (*paired t(4)= -3.00, p<0.05, two-tailed*). This difference in Track 13 may be explained by the fact that its musical composition is much more repetitive and has likely been produced by the use of looped musical samples; meaning that passages seen as 'similar' are actually identical. Therefore, the objective measures have been comparing signals that have no difference in spectra, phase, SNR, and so on. However, this being the case, it is interesting that the human listeners perceived a difference significant enough for them to identity some kind of degradation over the reference.

Although some discrepancy is present regarding Track 13, the use of objective testing, and the validity of its results, certainly appears to be accurate enough to give an initial indication as to the likely subjective grading that test samples would receive in subjective evaluations involving human listeners.

It would certainly appear that the ACER technique, in higher quality similarity thresholds of SIM3, SIM4, and SIM5, shows potential for implementation that would produce results that are acceptable to listeners. However, further, deeper, evaluation from both a subjective and objective perspective is necessary, to provide a more granular evaluation of this postulate.

6. DISCUSSION & FUTURE WORK

On reflection, it is felt that introducing the experimental set-up to participants as a small group and with some visual aides, such as a slideshow, to illustrate the concepts under evaluation, could enhance the training/briefing session and reduce the time taken individually introducing the experiment. However, it is recognized that may not always be practicable as participants would have to be kept waiting following the briefing.

Another issue arose, in that the playback volume of the VLC media player was set at over 100% during some of the subjective tests, which caused distortion in some of the playback. This is likely to have affected some of the results, although from a theoretical perspective both the reference and test signal would have been subject to any distortion and so this should not have skewed the participant comparison between samples.

A major issue during the subjective evaluation is in the amount of time it took to conduct the listening tests. The time it took to complete the experiment with 15 extracts was approximately 25 minutes per person. This is too long to keep the participant's concentration and goes outside of the ITU-R subjective evaluation guidelines. Shortening of clips would help in this as well as reducing the range of ACER compressed versions being tested. For example, future tests may only make use of a reduced range of compressed samples or to focus upon the higher-quality end of the ACER outputs. Another option would be to perform group testing, ideally using separate headphones to ensure parity of listening experience among participants. A more feasible option, though much harder to control, would be to use an online or electronic testing mechanism, although this presents a number of factors that then become almost impossible to control, such as listening equipment, background noise, volume of playback, distractions, timescales, and so on.

Another area for enhancement, as advised in [18] is the use of a training phase. As this was a pilot study, such a phase was not used, although a researcher gave a short briefing. In a larger scale study, such a session will be employed.

Particularly in the subjective evaluation, the impairment descriptions of the scale used may need to be reconsidered. Firstly, since there is no audio fidelity difference between the samples, it is questionable as to whether quality artifacts are being evaluated or whether 'trueness' to the original is a more accurate description of what is being evaluated. The other issue with the scale is that the use of 'annoying' could be misleading. These are too easily corrupted if the participant finds the original piece 'annoying'. There is also a tangible case to argue that the compressed versions are musically interesting, since there is an overlap between the results and some musical aesthetics found in pop composition. So the compression may add novelty or interest, essentially. We would be better off focusing on the difference from the original, potentially using some different semantic.

There are some clear disparities in the difference grades assigned from the objective and subjective evaluations. In some ways, this is an expected realization, since it exemplifies the divide that can exist between model-based and human-based testing systems; highlighting the importance of applying both approaches and triangulating the results to gain a better understanding of the real situation.

It is our intention to refine the experimental approach and conduct a subsequent series of evaluations using a larger test corpus (likely to be in the region of 10 to 20 tracks) and utilizing a larger number of listeners in the subjective testing, ideally with a group of around 10 expert listeners and another test group with around 20 non-expert listeners.

7. ACKNOWLEDGMENTS

Our thanks to participants of the pilot subjective evaluation study.

8. REFERENCES

[1] Martin, B., Robine, M., and Hanna, P. 2009. Musical structure retrieval by aligning self-similarity matrices. *In Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, 483-488.

[2] Paulus, J., Müller, M., and Klapuri, A. 2010. State of the art report: Audio-based music structure analysis. *Proceedings of ISMIR 2010*, 625-636.

[3] Sapp, C. S. 2011. *Computational Methods for the Analysis of Musical Structure*. Doctoral Thesis, Stanford University.

[4] Urbano, J. and Schedl, M. 2013. Minimal Test Collections for Low-Cost Evaluation of Audio Music Similarity and Retrieval Systems. *International Journal of Multimedia Information Retrieval*, 2, 1 (March 2013), 59 – 70.

[5] Aucouturier, J. J., and Pachet, F. 2002. Music similarity measures: What's the use. In *Proc. ISMIR* (Vol. 2).

[6] Foote, J. 1999. An overview of audio information retrieval. Multimedia Systems, 7, 1, 2-10.

[7] Harb, H. and Chen, L. 2007. A general audio classifier based on human perception motivated model. Multimedia Tools and Applications, 34, 3, 375-395.

[8] Fu, Z., Lu, G., Ting, K. M., and Zhang, D. 2011. A survey of audio-based music classification and annotation. *Multimedia, IEEE Transactions on*, 13, 2, 303-319.

[9] Wülfing, J., and Riedmiller, M. 2012. Unsupervised learning of local features for music classification. In *Proc. ISMIR*.

[10] Kirovski, D. and Landau, Z. 2007. Generalized Lempel–Ziv Compression for Audio. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15, 2, 509-518.

[11] Shannon, C. E. 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27,379 – 423, pp.623 – 656, July, October.

[12] Sayood, K. 2000. *Introduction to Data Compression*, Second Edition. Academic Press, London, UK.

[13] Cunningham, S. 2005. *Waveform Analysis for High-Quality Loop-Based Audio Distribution*. In Proceedings of ISCA 20th International Conference on Computers and Their Applications, New Orleans, USA, 16th-18th March.

[14] Cunningham, S. and Grout, V. 2007. *Advances in Similarity-Based Audio Compression*. In SEIN 2007: Proceedings of the Third Collaborative Research Symposium on Security, E-Learning, Internet and Networking (p. 129).

[15] Cunningham, S. and V. Grout. 2009. *Audio Compression Exploiting Repetition (ACER): Challenges and Solutions*, Proc. of Third International Conference on Internet Technologies and Applications, Glyndwr University, Wrexham, Wales, UK.

[16] ITU-R. 2001. *Recommendation ITU-R BS.1387-1, Method for objective measurements of perceived audio quality*. International Telecommunication Union – Radio communication Sector (ITU-R), Geneva, Switzerland.

[17] Bech, S. and Zacharov, N. 2006. *Perceptual Audio Evaluation – Theory, Method and Application*. Wiley-Blackwell.

[18] ITU-R, 2003. *Recommendation ITU-R BS.1284-1, General methods for the subjective assessment of sound quality*. International Telecommuncation Union – Radio communication Sector (ITU-R), Geneva, Switzerland.

[19] Various Artists. 2011. *Now That's What I Call Music! 80*. Compilation [Double Audio CD]. EMI TV.

[20] Kabal, P. 2004. *TSP Lab Software*. Electrical & Computer Engineering Department, McGill University, Canada. Available at: http://www-mmsp.ece.mcgill.ca/Documents/Software/index.html [Last accessed 6th May 2013]

Table V: PEAQ Objective Difference Grades (ODG) vs. Mean Subjective Difference Grades (SDG) for ACER Compressed Tracks

Track	ODG	SDG	ODG	SDG	ODG	SDG	ODG	SDG	ODG	SDG
03	-3.80	-3.75	-3.49	-2.88	-2.88	-2.13	-0.48	-1.25	-0.34	-1.50
09	-3.63	-3.38	-3.33	-3.00	-3.10	-1.25	-2.50	-1.50	-0.32	-1.50
13	-3.29	-1.13	-2.73	-0.88	-1.95	-0.38	-1.79	-0.63	-0.21	-0.50
	SIM1		**SIM2**		**SIM3**		**SIM4**		**SIM5**	