Research Article

Subjective Evaluation of Music Compressed with the ACER Codec Compared to AAC, MP3, and Uncompressed PCM

Stuart Cunningham [1] and Iain McGregor [2]

[1]*Centre for Advanced Computational Science, Manchester Metropolitan University, Manchester M1 5GD, UK*
[2]*School of Computing, Edinburgh Napier University, Edinburgh EH10 5DT, UK*

Correspondence should be addressed to Stuart Cunningham; s.cunningham@mmu.ac.uk

Audio data compression has revolutionised the way in which the music industry and musicians sell and distribute their products. Our previous research presented a novel codec named ACER (Audio Compression Exploiting Repetition), which achieves data reduction by exploiting irrelevancy and redundancy in musical structure whilst generally maintaining acceptable levels of noise and distortion in objective evaluations. However, previous work did not evaluate ACER using subjective listening tests, leaving a gap to demonstrate its applicability under human audio perception tests. In this paper, we present a double-blind listening test that was conducted with a range of listeners (N=100). The aim was to determine the efficacy of the ACER codec, in terms of perceptible noise and spatial distortion artefacts, against de facto standards for audio data compression and an uncompressed reference. Results show that participants reported no perceived differences between the uncompressed, MP3, AAC, ACER high quality, and ACER medium quality compressed audio in terms of noise and distortions but that the ACER low quality format was perceived as being of lower quality. However, in terms of participants' perceptions of the stereo field, all formats under test performed as well as each other, with no statistically significant differences. A qualitative, thematic analysis of listeners' feedback revealed that the noise artefacts that produced the ACER technique are different from those of comparator codecs, reflecting its novel approach. Results show that the quality of contemporary audio compression systems has reached a stage where their performance is perceived to be as good as uncompressed audio. The ACER format is able to compete as an alternative, with results showing a preference for the ACER medium quality versions over WAV, MP3, and AAC. The ACER process itself is viable on its own or in conjunction with techniques such as MP3 and AAC.

1. Introduction

In this work, we evaluate the performance of the ACER (Audio Compression Exploiting Repetition) codec [1]. Audio compression has evolved dramatically over the last 25 years, enabling many notable advances within fields such as multimedia broadcast, content distribution, consumer entertainment, and video games. During this period, a series of psychoacoustic-oriented lossy codecs have led this change, most notably the introduction of MPEG 1/2 Audio Layer 3 (MP3) and its successor Advanced Audio Coding (AAC). The general trend in lossy compression techniques has continued to follow this approach, with enhancement of the underpinning psychoacoustic models as well as support for multiple

channels and streaming [2–4]. Fraunhofer, the creator of the MP3 codec, announced the termination of the license for MP3 technology in 2017 in favour of its successors AAC, MPEG-H, and Enhanced Voice Services (EVS), which has cast doubt upon the MP3's ability to compete with alternative audio coding schemes from Fraunhofer and other providers [5].

In a previous work, the ACER audio coding scheme was presented. ACER approached the task of audio compression differently from current methods by being able to exploit the musical structure contained in the audio file using a dictionary-based method. The ACER approach is unusual in the audio compression domain, where the more conventional approach is to exploit psychoacoustic models of human

hearing and reflect these in the way that bits are allocated across the frequency spectrum. This is primarily achieved by focusing upon listener perceived characteristics of music that can be identified in order to exploit redundancy and irrelevancy in underlying audio signals [1]. The ACER scheme was envisaged as either a standalone coding scheme or as an additional processing step that might precede other codecs, such as MP3, AAC, or Ogg Vorbis. However, existing evaluation of ACER focused only upon objective quality evaluation [1] and a pilot subjective evaluation, conducted in an uncontrolled environment [6].

In this study, we conducted a large-scale evaluation of the ACER scheme against two popular audio codecs (MP3 and AAC), as well as an uncompressed wave (WAV) version of the audio. Since we are interested, in this study, in the human perception of audio compression schemes, we focus upon evaluating key perceptual qualities. As such, we aim to investigate the following null hypotheses:

> H_1: The perceived differences in audio quality, in terms of noise and distortion, between uncompressed WAV, AAC, MP3, and ACER music samples are insignificant.
>
> H_2: The perceived differences in audio quality, in terms of audio stereo imaging, between uncompressed WAV, AAC, MP3, and ACER music samples are insignificant.

We propose that if these hypotheses are maintained, then use of the ACER codec can be considered an appropriate alternative method of audio coding in a stand-alone form or be integrated with an existing psychoacoustic coding technique to enhance the amount of data reduction that can be achieved. The use of the ACER codec has the potential to expand the range of audio compression technologies available and provide an alternate data reduction method in situations where psychoacoustic compression, and the reduction in spectral resolution, may not be appropriate, such as in certain audio analysis tasks or high-fidelity audio playback.

The remainder of the paper is organised as follows: the next section provides background to our work by providing a critical discussion of recent research in the field of audio compression and associated perceptual testing approaches. After that, an overview of the ACER compression scheme is presented. Section 4 describes the subjective listening test method and stimuli used before. Section 5 explores the results and analysis of the ACER scheme alongside the alternate audio codecs. Section 6 explores the qualitative descriptions of participants' experiences with each of the codecs. Finally, we provide conclusions, incorporating discussion of limitations of this study and areas of future Work.

2. Related Work

The development of audio compression schemes from their inception to evaluation is a domain that draws upon multiple disciplines, including computer science, audio engineering, and listening tests and evaluations. In this section, we aim to provide the reader with a broad, informative account of the pertinent aspects of audio data compression which contextualise and underpin the work that is presented in this paper.

2.1. Audio Coding. As with other forms of digital media information, audio has received significant attention with regard to ways to reduce the number of bits required for storage and transmission. The process of analogue-to-digital conversion (sampling) itself is one where decisions must be made as to the sample rate and bit depth of the subsequent audio that will reliably allow the desired frequencies and level dynamics of the original sound to be represented. This is typically done when creating a necessarily compressed Pulse Code Modulation (PCM) representation, which itself can be described as a form of data compression. The successful reproduction of frequencies and dynamics is paramount in order to provide listeners with high-fidelity (Hi-Fi) audio reproduction. However, the Human Auditory System (HAS) is not linear in its interpretation of the frequency and amplitude of sounds presented to it, meaning that human perception of sound does not always require that all of the potentially audible frequencies and dynamic qualities of sound are present when auditory stimuli are presented. The phenomena of frequency and temporal masking [7, 8] are often exploited in lossy approaches to audio compression. Most modern codecs are hybrids, augmenting semantic approaches, such as perceptual redundancies associated with the HAS, with traditional syntactic methods such as Huffman [9] and Rice [10] codes.

Lossless coding approaches to audio, whilst effective, have largely been stagnant in terms of the amount of data reduction obtainable [11]. One exception in the field of lossless audio coding is the Free Lossless Audio Codec (FLAC), which is able to achieve compressions ratios in the region of 2:1 with no loss of data through the use of predictive models [12]. The ability of FLAC to produce lossless audio is relatively novel amongst audio compression methods, although it is not able to yield similar compression ratios to its lossy contemporaries, which are typically between the range of 4:1 and 15:1. Other contemporary lossless techniques have expanded upon these principles of using linear predictors, with marginal increases in compression ratios being achieved [13, 14]. It is essential that any method of audio compression is efficient in the reduction of the number of bits used to represent sound. In lossless techniques, preservation of the original signal is paramount.

However, it is often necessary to employ lossy techniques to achieve higher ratios of compression, which generally operate by exploiting psychoacoustic properties and limitations of the HAS. It is crucial that the decoding process does not inhibit the fluid playback of the sound, requiring that it is fast, requires a small amount of CPU processing time, and produces relatively accurate results. Consequently, audio encoding techniques are asymmetric, with tolerable delays in compression, provided that the decompression process is as close as possible to real time [15]. Lossy techniques are commonplace within digital media, especially with regard to music, and are exemplified by methods such as Ogg Vorbis [16], MP3, and AAC [17]. The methods achieve scalable

data reduction, depending upon the usage application, and are able to achieve perceptually highly similar results to uncompressed audio [18–20].

More recent developments in the audio compression domain have seen work done to enhance the audio fidelity able to be produced by codecs operating at very low bit rates, such as 24, 48, 64, or 92 kbps [21, 22], whereas coding around 120 to 256 kbps might be considered typical, aiming to achieve extremely high "perceptually transparent" data-reduced coding. Work has also focused upon audio compression systems in high-quality telecommunications and in multichannel systems designed for spatial audio reproduction, which are typically 6 or 8 channels, but are easily expanded into larger numbers [23].

2.2. Perceptual Audio Evaluation. When dealing with audio, it is key to include perceptual evaluation when measuring the performance of a codec. The determination of how resultant audio sounds to listeners as a consequence of the data reduction process is essential if it is to be widely adopted. Perceptual evaluation can be conducted using either objective and/or subjective mechanisms.

Objective evaluations rely upon signal features of the audio being analysed and compared to a known reference or benchmark. This process can use simplistic mechanisms, such as Signal-to-Noise Ratio (SNR) or more complex algorithms, based upon models of the human auditory system, such as the Perceptual Evaluation of Audio Quality (PEAQ) metric [24]. Both of these approaches are usually quick and convenient to implement, enabling large numbers of audio samples to be processed and evaluated. However, simpler measures of audio quality may not necessarily reflect actual human perception of the signal. More complex models may not be fully generalizable due to the differences from person to person with regard to their unique auditory systems [25, 26].

Objective testing is a convenient and resource-efficient way of measuring the efficacy of a particular audio codec. Especially since the typical barriers to conducting subjective tests are time, equipment resources, and obtaining a sufficient number of participants, there is limited evidence indicating that objective measures of higher bit rate audio codecs produce comparable results to subjective evaluations [27]. However, it is recognized that the introduction of any new coding technique should be complemented by subjective testing in order to obtain a fuller picture of the perceptual effect [24, 28].

In terms of the ideal number of participants to use in such audio quality evaluations, the International Telecommunication Union Radiocommunication (ITU-R) body advocates a minimum of 10, if using expert listeners, or minimum of 20, if using nonexpert listeners [29]. Existing subjective audio evaluation studies have tended to comply with this utilisation of small sample sizes, with 26 being an average number of participants [30–33].

2.3. Performance of Contemporary Codecs. In one subjective evaluation undertaken [22], it was found that, at low bit rates varying between 24 kbps and 64 kbps, MP3, high-efficiency AAC, low-complexity AAC, and five other coding schemes commonly used in broadcast applications received varying subjective quality scores from a group of 23 participants in terms of the degradations present in the audio. However, at higher bit rates, these schemes demonstrated greater consistency between scores and lower levels of degradation, "*... all codecs provide a near transparent audio quality*". This work indicates that, at relatively high bit rates, varying between 128 kbps and 320 kbps, the psychoacoustic codecs perform perceptually similarly.

Another study [20] evaluated MP3 music encodings at a series of bit rates, 96, 128, 192, 256, and 320 kbps, against uncompressed CD quality audio using a total of 13 trained listeners, with a range of backgrounds, including sound engineers and musicians. The five music samples in their study were drawn from two genres: rock and roll and classical. Each clip duration was between 5 and 11 seconds to encompass a distinct musical phrase from the respective song. Participants carried out a series of AB comparisons across the six representations of each music sample. Their findings, across all participants and music tracks, suggested that there was a statistically significant preference for the uncompressed CD quality audio when compared to the 96, 128, and 192 kbps MP3 versions. However, there were no significant differences identified when comparing CD quality audio to the 256 and 320 kbps MP3 versions. Participants of this study were also asked to provide qualitative descriptions of the artefacts and distortions they perceived in the audio. The authors identified the following categories of artefacts, in order of their instances of occurrence: high-frequency artefacts, general distortion, reverberation, transient artefacts, stereo image, dynamic range, and background noise. This work is of interest as it suggests that participants cannot easily distinguish between MP3 and uncompressed audio beyond a threshold of 256 kbps, as well as presenting a potential framework for measuring artefacts that might be perceived in coded audio samples.

3. Summary of the ACER Codec Approach

The main tenet of the ACER approach is to exploit the structural compositional redundancies present in contemporary music to achieve data reduction rather than to rely upon deficiencies with the HAS in its resultant perception. Popular music, in particular, utilises repetition as a conscious tool to engage listeners and bring form and structure to a piece. In a large number of cases, this means that identical content is repeated at several instances during music playback rather than a human performance of the same musical sequence, which would be prone to subtle differences in timing and dynamics. The presence of this repetition gives rise to the opportunity for redundancies to be detected and taken advantage of to achieve data compression. The ACER approach draws upon principles of lossless dictionary-based schemes [15] to achieve this. These principles can be easily exemplified by considering the short sequence of musical notation, in the key of C major, presented in Figure 1.

This example presents a simple musical melody over eight bars of music and using a total of thirty explicitly encoded
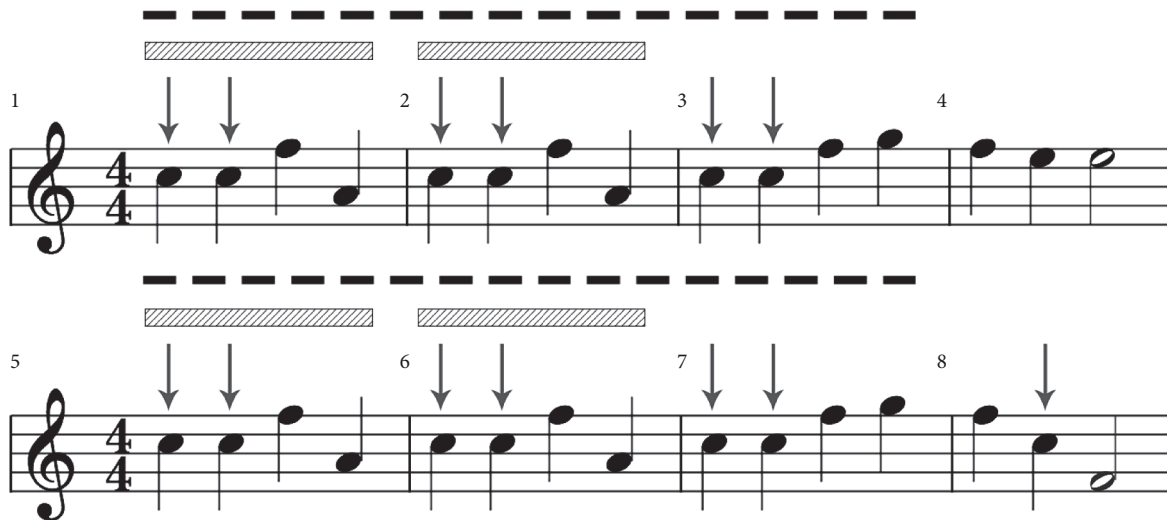
FIGURE 1: Simple eight-bar musical sequence. Arrows indicate a repeated note, shaded areas indicate a repeated 4-note/1-bar sequence, and the dotted areas indicate a repeated 12-note/3-bar sequence.

notes. It is evident that there are redundancies present in this representation, which could be exploited to achieve a reduced size representation of the piece and that these redundant objects may be detected with windows (durations) of different sizes. For instance, the first note in the sequence appears a total of thirteen times (each note is highlighted by an arrow in the diagram); however, the overhead of the dictionary index and symbol makes this inefficient. On a larger scale, the first complete bar of music appears four times (highlighted by the shaded rectangles), potentially providing saving of eight out of the thirty notes, plus a small coding overhead. The observation may also be made that, scaling up further, the first three bars of the piece are identical to bars five, six, and seven (highlighted by the dashed line), presenting another redundancy that saves twelve of the thirty notes, plus a small coding overhead, because the first line (bars 1 to 4) and second line (bars 5 to 6) differ only by the final two notes.

The ACER technique takes the approach outlined above and executes the same principles, as discussed on a symbolic level, but at signal level. This presents additional challenges due to a number of factors, such as noise, polyphony, and absence of quantisation, as well as performative and expressive factors. ACER performs searches within musical audio pieces to detect perceptually identical, or similar, sections of music that occur and extracts redundant segments.

The ACER coding process begins by establishing a *search* block, which has a size derived using the tempo of the music track to be coded. The tempo is trivial to obtain using metadata or, if there is no metadata available, through beat detection analysis of the track's signal. The track is then divided into consecutive *target* blocks of the same size and a linear search is performed to identify those blocks deemed perceptually similar. In comparing search and target blocks, a windowed Fourier Transform is taken of each and a difference spectrum calculated from the two. The mean value of this difference spectrum is then compared to a threshold

to determine if the two blocks are perceptually similar. The threshold is defined prior to the search and has the effect of manipulating the quality settings and compression amounts ACER will achieve [1]. When all current target blocks have been compared to the search block, the search block is incremented and the process repeated until the search space is exhausted. The index location of matching search and corresponding target blocks identified are stored so that they can later be removed from the track. Thus, when the ACER encoding stage is complete, the end user is left with a collection of audio blocks and indices, from which it is possible to reconstruct a representation of the original track. These steps of the algorithm are defined in more detail in our earlier work [1].

The perceptually similar definitions are based upon regression models developed using human listeners, which form part of an earlier technical description of the ACER compression processes and algorithms [1]. In that study, an objective quality evaluation of the ACER system was conducted where the Objective Difference Grade (ODG) [24] and Signal-to-Noise Ratio (SNR) were studied over five different levels of ACER audio quality (fidelity). Over the 43 tracks compressed, the mean bit rates achieved were as follows: 1037 kbps (lowest quality), 1118 kbps (low quality), 1218 kbps (medium quality), 1298 kbps (high quality), and 1352 kbps (highest quality). The two lowest levels of ACER quality were deemed to have performed poorly, on average falling between the ODG descriptors of "annoying" and "very annoying". In comparison, the top-quality ACER encoding scored between the descriptors of "imperceptible" and "perceptible, but not annoying", the second highest between "perceptible, but not annoying" and "slightly annoying", and the third highest between "slightly annoying" and "annoying". These findings were followed by a small-scale subjective evaluation of the ACER scheme, where each of its coding levels was investigated to determine the relative difference in quality between

each [6]. Hence, for the study to be undertaken here, only the upper three of the quality levels of the ACER scheme are employed, now renamed as follows: ACER high, ACER medium, and ACER low.

Our previous studies lacked any in-depth and sustained subjective, perceptual evaluation of the efficacy of the ACER scheme in comparison to uncompressed and compressed alternative formats (MP3 and AAC). This was due to a lack of time and access to a specialist listening suite resource. It is this deficiency that is addressed in this work.

4. Materials and Methods

4.1. Method. A listening test study was conducted to determine the perceived quality and performance of the ACER approach in comparison to uncompressed WAV, MP3, and AAC coded musical audio. Use of a listening test methodology such as ITU-R BS-1116 [34] or Multiple Stimulus Hidden Reference and Anchors (MUSHRA) [35] would have been a feasible approach. However, such approaches require study participants to be expert listeners who are proficient at detecting small differences in audio quality. Whilst the use of expert listeners is intended to ensure reliable results, it does not accurately reflect the broader population, which has a much greater level of variation with regard to their perception of audio quality. Based upon this, a custom approach was adopted and it was decided to use untrained listeners in the study.

Participants were provided with the opportunity to hear a short (20 s) sample from the 10 selected songs. Each was played back repeatedly until the participant completed their response or wished to move on. They were able to hear six versions of each song: uncompressed WAV, MP3 192 kbps CBR, AAC 192 kbps CBR, ACER low quality, ACER medium quality, and ACER high quality. Each sample was played back concurrently and fed in random order into a Canford Source Selector HG8/1 hardware switch, allowing participants to freely select which sample stream they were listening to using a simple rotary switch.

Enclosed Beyer Dynamic DT770M 80-ohm headphones were chosen for the study as they have a passive ambient noise reduction of 35 dB, according to the manufacturer's specification. A Rane HC6S headphone amplifier was set so that the RMS level was 82 dBC, broadly in accordance with the reference level recommended by the ITU-R [29, 34], and with a peak of 95 dBC. Music is the most popular media form for headphone use with high levels of adoption and regular use [36, 37]. Headphones are reported as being the second equal most popular method after computer speakers for the consumption of music [38].

The use of headphones also minimised the effect of any room acoustic colouration, which are known to affect listening studies [39]. They also potentially facilitate a greater level of detail due to driver proximity and minimal crosstalk. It is acknowledged that the stereo image experienced when using headphones will differ from that of loudspeakers. Nevertheless, when using headphones, the listener experiences the sound as being perceptually from the exterior world [40]. It has been found that there is little difference between studio loudspeakers and studio quality headphones in audio evaluation situations; both MUSHRA [41] and ITU-R standards for listening tests endorse use of either headphones or loudspeakers [29, 34].

With respect to each song, participants were invited to provide a response, using paper-based scoring sheets, to two questions. The first concerned the presence of any noise in the samples presented, and the second related to the quality of the stereo image they experienced. The wording used for these two questions was selected by considering the terminology recommended in ITU-R BS.1284 [29]. Each question on the scoring sheet clearly articulated the scoring criteria and the bipolar descriptors used at each end of the grading scale.

Participants were asked to rate each clip's audio quality with respect to noise and distortions using a five-point semantic differential scale as follows: *1 = imperceptible noise and distortions; 5 = perceptible noise and distortions*. This question would allow the participants to refer to any type of noise or artefact present within the sample, providing scope to capture both linear and nonlinear distortion factors. Participants were then asked to rate each clip in terms of its stereo image quality, using a five-point semantic differential scale as follows: *1 = narrow and imprecise; 5 = wide and precise*. Similarly, this question provided participants with the opportunity to describe the stereo spread and their ability to localise distinct sound sources within the music. As participants listened to the six codec variations of each of the ten song samples, they were asked to specify which of the six clips was their favourite and which was their least favourite.

4.2. Participants. A total of 100 participants engaged with the listening test and were recruited from the Merchiston campus at Edinburgh Napier University. With respect to background, 28% were students at the University, whilst 33% were academic or faculty staff and 39% were administrative and support staff. Participants were not offered any form of remuneration or any other form of inducement for their involvement.

In terms of other demographic details, 55 participants were female and 45 were male. The mean age was 40 (SD=12) with a minimum age of 20 and a maximum age of 68. All participants identified themselves as having what they considered to be normal hearing for their age. 17% identified that they had some form of professional audio training and 37% indicated that they had some form of musical training. Finally, participants were asked to give an indication of how much time they typically spent listening to music per day. 72% responded that they listened to music between 1 and 3 hours each day, and 8% did not listen to any music at all.

4.3. Test Materials. A total of 10 musical excerpts were used in the evaluation. These songs were chosen at random from a double-CD album compilation of contemporary pop music in the UK: *Now That's What I Call Music! 90* [42]. This was chosen as it represented a broad sample of contemporary, popular music in the sampled population. The tracks that were selected for use in the evaluation are shown in Table 1.

As the samples were taken from a commercial CD, each song was represented in CD audio quality (Red Book) [43]:

Table 1: Selected music tracks for perceptual testing.

Artist	Song
Mark Ronson feat. Bruno Mars	*Uptown Funk*
Sia	*Elastic Heart*
Take That	*These Days*
Alesso feat. Tove Lo	*Heroes (We could be)*
Marlon Roudette	*When the beat drops out*
David Guetta feat. Sam Martin	*Dangerous*
Flo Rida feat. Sage The Gemini & Lookas	*GDFR*
Charli XCX feat. Rita Ora	*Doing It*
Alex Adair	*Make Me Feel Better*
Florence + The Machine	*What Kind of Man*

two's complement binary 44.1 kHz sample rate, 16-bit word length, 2 channels (stereo), and PCM recording. From each song, a sample of 20 seconds in duration was extracted. The beginning of each sample had a linear fade-in of 1.5 seconds applied and an equivalent 1.5-second fade-out was applied to the end of each sample. This modification was intended to make the experience of hearing each clip less abrupt for participants and to make it easier to determine when each sample started and finished.

To create the compressed versions of each song, the clips were subjected to the respective compression processes and the same 20-second-long excerpt subsequently was extracted. The fade-ins and fade-outs were then applied, in line with ITU-R recommendations for the duration and presentation of musical samples [29]. Since the evaluation would be carried out in a double-blind manner, all samples were then resaved as CD quality PCM and allocated names of randomly generated four-character strings. The materials were then passed to the second author who conducted the listening evaluation.

The obtained bit rates for each of the six versions of the song are shown in Table 2. It is worth noting that, with the exception of the ACER approach, the other methods provide a fixed bit rate regardless of audio content. Over the ten tracks used in this experiment, the ACER high quality codec achieved a mean reduction in size of 12.60%; the ACER medium quality received a mean reduction in size of 19.92%; and ACER low quality received a mean reduction in size of 27.53%.

Since the ACER technique operates by removing redundancies in a particular piece of musical audio, the amount of compression (i.e., reduction in bit rate) is directly influenced by the sonic content of the audio file itself. For instance, music that features high amounts of repetition and small amounts of variation in musical performance, articulation, and orchestration will achieve much reduced bit rates with the ACER scheme, whereas music that may be considered more avant-garde, with unconventional structure or great variation in performance, articulation, and orchestration, will achieve less of a reduction in bit

rate. The quality settings of the ACER scheme throttle the amount of perceptual similarity tolerated by the coder: high-quality settings are strict about which sequences are considered to be a match, whilst lower-quality settings are less strict and more likely to give rise to perceptual anomalies.

5. Results: Quantitative Measures

Although 100 people took part in the listening test, it was not compulsory for them to provide a rating for each audio stimulus so as to accommodate listener uncertainty or inability to select a preference. This mandate of not forcing participants to provide responses is also a requirement of achieving ethical approval from the University (Edinburgh Napier) where the listening study took place. As such, not all participants provided a full set of ratings for all of the stimuli, making a complete, repeated-measures comparison of ratings impossible using the entire set of 100 participants. Those who did not provide a rating for every track have been excluded from the analysis presented in the subsequent subsections, which deal with the quantitative scoring of noise and stereo field factors being assessed from the listening test. However, if participants responded to the subsequent questions, relating to their most and least favourite versions of each songs, their responses have been included in the subsequent subsection and any qualitative feedback received has also been used. This was decided to be an appropriate strategy, since it is likely that participants may not have rated some versions of each track by mistake, given the relatively large number of comparisons ($6*10$) undertaken.

5.1. Perceptions of Noise and Distortion. A complete set of scores was provided by 68 of the 100 experiment participants (n = 68). A summary of the results obtained for each of the 10 songs used in the listening experiment is shown in Figure 2 (songs 1 to 5) and Figure 3 (songs 6 to 10). These graphs present the mean score for each codec with error bars illustrating one standard deviation from the mean.

As suggested by these figures, the mean and standard deviation (SD) scores for the six coding variations appear to be similar in terms of perceived noise and distortion. These descriptive statistics are specifically shown in Tables 3 and 4. The experiment contained two independent variables: the six methods used to encode the music and the ten music tracks that were encoded. In order to address the null hypothesis H_1, stated in Introduction of this article, a two-way repeated-measures ANOVA was performed upon the scores received from all 68 valid responses to the question related to noise and distortions. The expectation in doing so was that if each of the coding mechanisms is equivalent in terms of quality, there should be no significant difference in listening test participants' scores. A repeated-measures ANOVA with a Greenhouse-Geisser correction showed that scores of noise and distortions differed significantly between the six codecs $F(3.829, 256.516) = 5.988, p < 0.001$. Post hoc pairwise tests using the Bonferroni correction revealed that this result was due to the ACER low quality encodings, which yielded significantly different noise and distortion scores to all other

TABLE 2: Bit rates achieved for each combination of codec and song.

Song	Bit rate (kbps)					
	WAV	MP3	AAC	ACER High	ACER Med	ACER Low
Uptown Funk	1411	192	192	1174	1086	1023
Elastic Heart	1411	192	192	1287	1174	896
These Days	1411	192	192	1174	1063	965
Heroes (We Could Be)	1411	192	192	1178	998	855
When the Beat Drops Out	1411	192	192	1395	1348	1178
Dangerous	1411	192	192	1244	1171	1153
GDFR	1411	192	192	1081	1019	945
Doing It	1411	192	192	1341	1184	1098
Make Me Feel Better	1411	192	192	1060	901	813
What Kind of Man	1411	192	192	1398	1356	1300
Mean bit rate (kbps):				1233	1130	1023
Standard deviation bit rate (kbps):				115	139	149

FIGURE 2: Noise and distortions results (songs 1 to 5). A score of 1 represents imperceptible noise and distortions and a score of 5 represents perceptible noise and distortions.
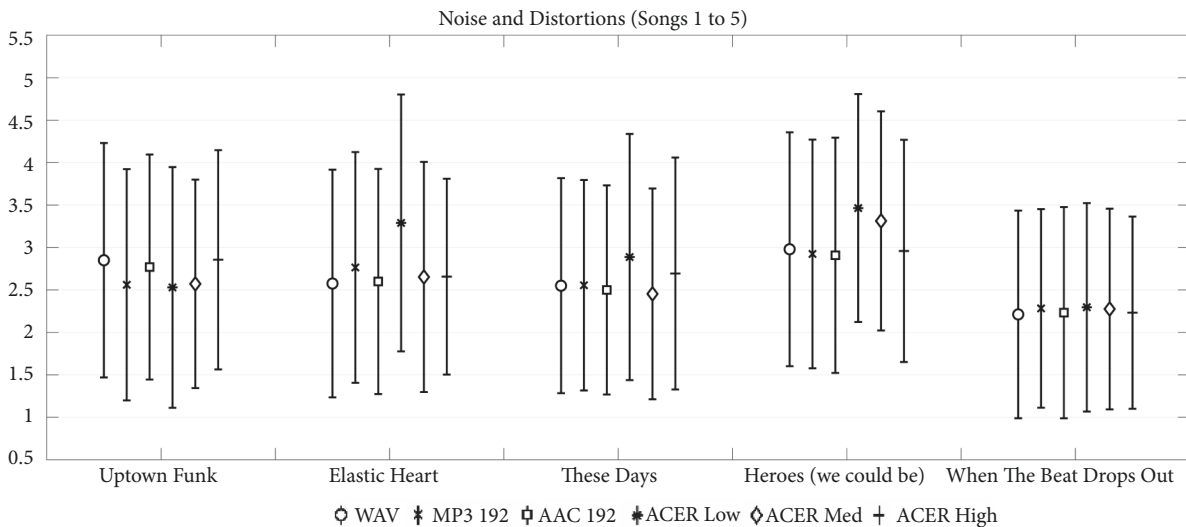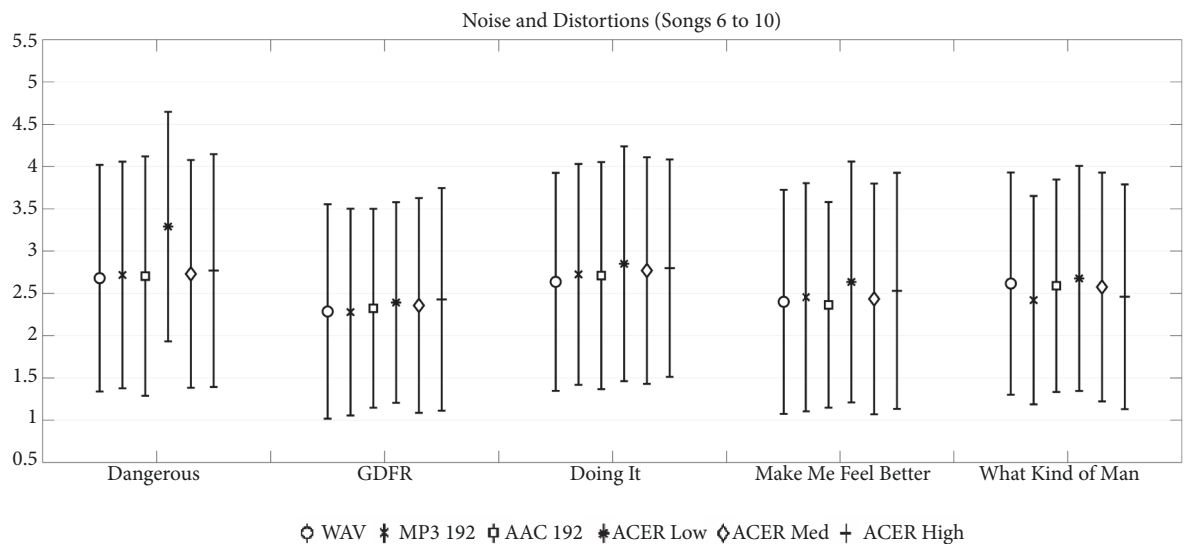

FIGURE 3: Noise and distortions results (songs 6 to 10). A score of 1 represents imperceptible noise and distortions and a score of 5 represents perceptible noise and distortions.

TABLE 3: Summary noise and distortions scores: WAV, MP3 192, and AAC 192 (n=68). A score of 1 represents imperceptible noise and distortions and a score of 5 represents perceptible noise and distortions.

Song	WAV		MP3 192		AAC 192	
	Mean	SD	Mean	SD	Mean	SD
Uptown Funk	2.85	1.38	2.56	1.36	2.77	1.32
Elastic Heart	2.58	1.34	2.77	1.36	2.60	1.33
These Days	2.55	1.27	2.56	1.24	2.50	1.23
Heroes (We Could Be)	2.98	1.38	2.92	1.35	2.91	1.39
When the Beat Drops Out	2.21	1.22	2.28	1.17	2.23	1.24
Dangerous	2.68	1.34	2.72	1.34	2.70	1.42
GDFR	2.29	1.27	2.28	1.22	2.32	1.18
Doing It	2.64	1.29	2.72	1.31	2.71	1.34
Make Me Feel Better	2.40	1.33	2.45	1.35	2.36	1.22
What Kind of Man	2.62	1.31	2.42	1.23	2.59	1.26
Grand Mean	2.58	1.31	2.57	1.29	2.57	1.29

TABLE 4: Summary noise and distortions scores: ACER low, ACER medium, and ACER high (n = 68). A score of 1 represents imperceptible noise and distortions and a score of 5 represents perceptible noise and distortions.

Song	ACER Low		ACER Med		ACER High	
	Mean	SD	Mean	SD	Mean	SD
Uptown Funk	2.53	1.42	2.57	1.23	2.86	1.29
Elastic Heart	3.29	1.51	2.65	1.36	2.66	1.15
These Days	2.89	1.45	2.45	1.24	2.69	1.37
Heroes (We Could Be)	3.46	1.34	3.31	1.29	2.96	1.31
When the Beat Drops Out	2.30	1.23	2.28	1.18	2.23	1.13
Dangerous	3.29	1.36	2.73	1.35	2.77	1.38
GDFR	2.39	1.19	2.36	1.27	2.43	1.32
Doing It	2.85	1.39	2.77	1.34	2.80	1.29
Make Me Feel Better	2.64	1.42	2.43	1.36	2.53	1.40
What Kind of Man	2.68	1.33	2.58	1.35	2.46	1.33
Grand Mean	2.83	1.36	2.61	1.30	2.64	1.30

TABLE 5: Post hoc pairwise codec comparisons (p values < 0.05 are highlighted in bold).

Codec	AAC 192	ACER High	ACER Med	ACER Low	MP3 192	WAV
AAC 192		*1.000*	*1.000*	***0.018***	*1.000*	*1.000*
ACER High	*1.000*		*1.000*	*0.110*	*1.000*	*1.000*
ACER Med	*1.000*	*1.000*		***0.006***	*1.000*	*1.000*
ACER Low	***0.018***	*0.110*	***0.006***		***0.002***	***0.005***
MP3 192	*1.000*	*1.000*	*1.000*	***0.002***		*1.000*
WAV	*1.000*	*1.000*	*1.000*	***0.005***	*1.000*	

codecs, with the exception of the ACER high quality codec scores.

There were no statistically significant differences between the remaining five codecs. This is illustrated in the obtained p value for the pairwise comparisons of each codec, shown in Table 5, with significant values ($p < 0.05$) highlighted in bold. The results from this part of the listening test demonstrate that, with the exception of the ACER low quality codec, the other codecs performed as well as the uncompressed WAV music samples in terms of noise and distortions perceived by participants.

5.2. Perceptions of Stereo Image. A complete set of scores was provided by 63 of the 100 experiment participants (n = 63). A summary of the results obtained for each of the 10 songs used in the listening experiment is shown in Figure 4 (songs 1 to 5) and Figure 5 (songs 6 to 10). These graphs present the mean score for each codec with error bars illustrating one standard deviation from the mean. An initial visual inspection of this descriptive information shows general consistency within each of the songs analysed and no particular trend in terms of the performance of each of the codecs under scrutiny. This suggests that there were no significant differences between
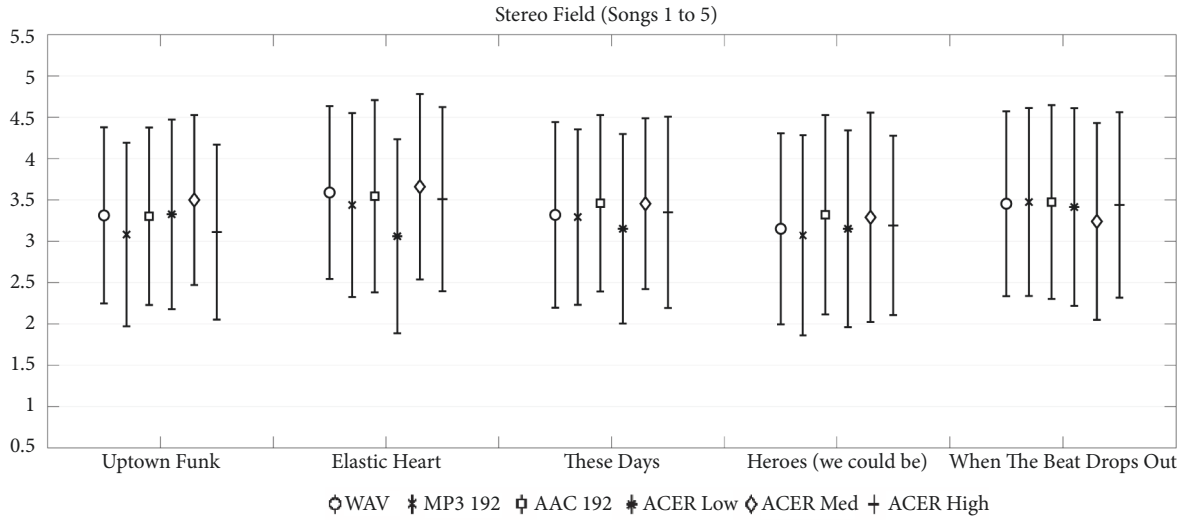
FIGURE 4: Stereo field results (songs 1 to 5). A score of 1 represents narrow and imprecise and a score of 5 represents wide and precise.
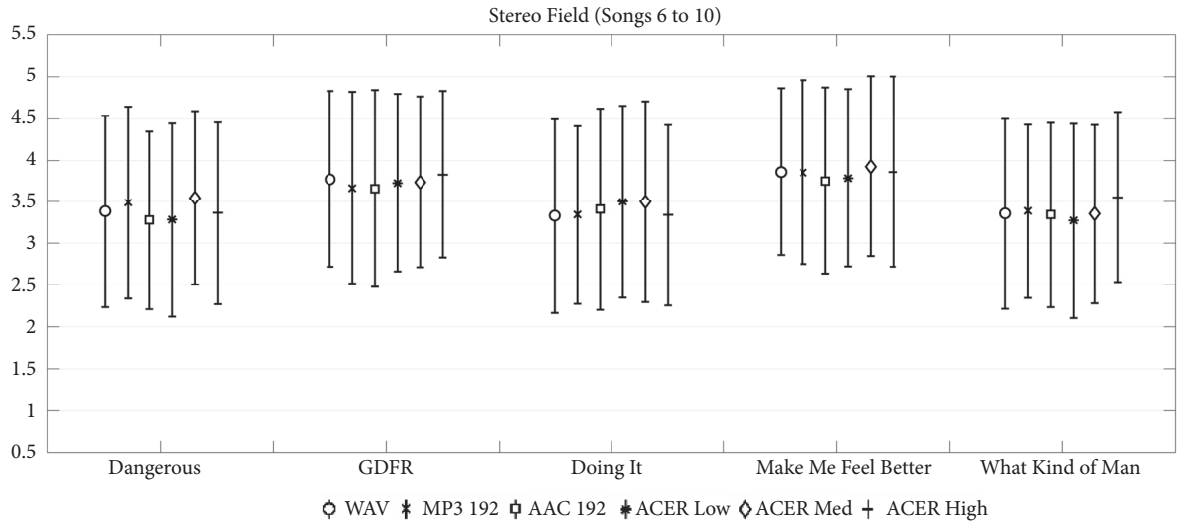

FIGURE 5: Stereo field results (songs 6 to 10). A score of 1 represents narrow and imprecise and a score of 5 represents wide and precise.

each of the coding approaches in terms of their perceived stereo image.

As suggested by these figures, the mean and standard deviation (SD) scores for the six coding variations seem to be similar in terms of perceived stereo image. These descriptive statistics are specifically shown in Tables 6 and 7.

The experiment contained two independent variables: the six methods used to encode the music and the ten music tracks that were encoded. In order to address the null hypothesis H_2, stated in Introduction of this article, a two-way repeated-measures ANOVA was performed upon the scores received from all 63 valid responses to the question related to stereo image. The expectation in doing so was that if each of the coding mechanisms is equivalent in terms of quality, there should be no significant difference in listening test participants' scores. A repeated-measures

ANOVA with a Greenhouse-Geisser correction showed no significant differences in scores of stereo image between the six codecs $F(4.097, 254.019) = 1.116, p > 0.05$. The results from this part of the listening test demonstrate that all of the codecs performed as well as the uncompressed WAV music samples in terms of the stereo image quality perceived by the experiment participants.

5.3. Audio Codec Preferences. Engagement with this part of the test was high, with almost all participants specifying a favourite coded version for at least one of the 10 songs presented to them (97 participants expressed 936 out of a possible 1000 preferences) and least favourite version (96 participants expressed 907 out of a possible 1000 preferences). 50 participants provided a favourite for every song, whilst 46 provided incomplete sets of favourites. Given the repetitive

TABLE 6: Summary stereo image scores: WAV, MP3 192, and AAC 192 (n=63). A score of 1 represents narrow and imprecise and a score of 5 represents wide and precise noise and distortions.

Song	WAV		MP3 192		AAC 192	
	Mean	SD	Mean	SD	Mean	SD
Uptown Funk	3.31	1.07	3.08	1.11	3.30	1.07
Elastic Heart	3.59	1.05	3.44	1.11	3.55	1.16
These Days	3.32	1.12	3.29	1.06	3.46	1.07
Heroes (We Could Be)	3.15	1.15	3.07	1.21	3.32	1.21
When the Beat Drops Out	3.45	1.12	3.47	1.14	3.47	1.17
Dangerous	3.38	1.15	3.49	1.15	3.28	1.07
GDFR	3.77	1.06	3.67	1.15	3.66	1.18
Doing It	3.33	1.16	3.34	1.07	3.41	1.21
Make Me Feel Better	3.86	1.00	3.85	1.10	3.75	1.12
What Kind of Man	3.36	1.14	3.39	1.04	3.34	1.11
Grand Mean	*3.45*	*1.10*	*3.41*	*1.11*	*3.45*	*1.14*

TABLE 7: Summary stereo image scores: ACER low, ACER medium, and ACER high 192 (n=63). A score of 1 represents narrow and imprecise and a score of 5 represents wide and precise noise and distortions.

Song	ACER Low		ACER Med		ACER High	
	Mean	SD	Mean	SD	Mean	SD
Uptown Funk	3.33	1.15	3.50	1.03	3.11	1.06
Elastic Heart	3.06	1.17	3.66	1.12	3.51	1.11
These Days	3.15	1.15	3.45	1.03	3.35	1.16
Heroes (We Could Be)	3.15	1.19	3.29	1.27	3.19	1.08
When the Beat Drops Out	3.41	1.20	3.24	1.19	3.44	1.12
Dangerous	3.28	1.16	3.55	1.04	3.36	1.09
GDFR	3.73	1.07	3.74	1.03	3.83	1.00
Doing It	3.50	1.15	3.50	1.20	3.34	1.08
Make Me Feel Better	3.79	1.07	3.93	1.08	3.86	1.14
What Kind of Man	3.27	1.17	3.35	1.07	3.56	1.02
Grand Mean	*3.37*	*1.15*	*3.52*	*1.11*	*3.46*	*1.09*

TABLE 8: Favourite and least favourite codec across all songs (largest values are highlighted in bold).

Codec	Favourite % (n=936)	Least Favourite % (n=907)
Uncompressed WAV	18.27	14.66
MP3 192 kbps	13.78	13.12
AAC 192 kbps	17.63	14.44
ACER Low Quality	14.21	**26.24**
ACER Medium Quality	**19.23**	16.43
ACER High Quality	16.88	15.10

nature of this question and to make best use of the data obtained, it was decided to include participants who had expressed a favourite on one or more occasion rather than to exclude any data that was not 100% complete. These scores were aggregated over all ten song samples to produce a distribution of scores for the six codec audio samples. Table 8 shows the proportions of favourite and least favourite codecs obtained.

Closer inspection with a Chi-Square test revealed that the distribution of favourite codecs was distributed in a nonuniform way $\chi^2(5) = 13.744$, $p < 0.02$, as was the distribution of participants' least favourite codec $\chi^2(5) = 62.956$, $p < 0.00001$. To provide a balanced analysis of favourite versus least favourite, Figure 6 shows an analysis of the difference between the two sets of results to help illustrate the overall direction (positive or negative) of codec preference and the strength of this preference.

The data presented in Figure 6 indicates that the uncompressed WAV, MP3 192 kbps, AAC 192 kbps, medium-quality ACER, and high-quality ACER codecs all received positive preferences with the uncompressed WAV performing marginally the best, followed by the AAC and
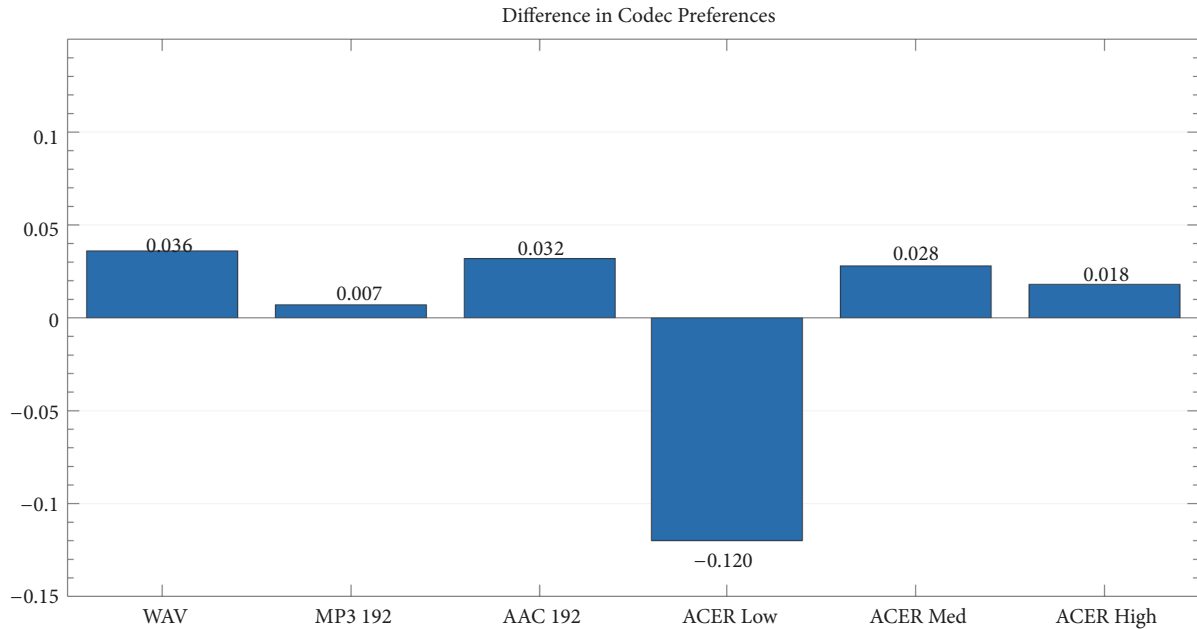
Difference in Codec Preferences

FIGURE 6: Difference between favourite and least favourite codecs.

medium-quality ACER. The most notable outcome from this analysis is the strong disliking for the low-quality ACER codec, the only one to have an overall negative preference. This outcome supports the findings from participants' ratings of noise and distortions, which demonstrated that only the low-quality ACER codec was statistically different from the others and that the remaining five codecs were similar in terms of their perceived audio quality.

6. Results: Qualitative Measures

The quantitative measures outlined previously provide strong and reliable indicators of the listeners' perceptions and preferences for each of the coding schemes under investigation. As explained earlier, such an approach is a common way of evaluating audio quality in controlled situations. To enhance the validity of these findings, as well as provide a more detailed exploration and understanding of the listeners' experience, a thematic analysis [44] was undertaken of the free text comments provided in response to the statement at the end of the listening test: "Please could you describe any noise or anomalies that you heard in any of the audio clips."

The use of these qualitative indicators is helpful in understanding some of the reasoning behind the quantitative values assigned by participants during the listening test, especially since the ACER scheme had not previously undergone such a detailed evaluation. Since the ACER approach does not reduce the resolution of the audio that is retained during compression, there should not be any added distortion or background noise. However, it was expected that, in some cases, especially at lower bit rates, ACER may produce a "skipping" or "jumping" effect at playback because of the reduction in similarity threshold between matching blocks in the music.

6.1. Approach. The use of thematic analysis and qualitative investigation in audio evaluation is encountered in a range of scenarios. It allows researchers to gain a better insight into the exact nature of audio artefacts and other perceptual objects that may be experienced by their listeners. For example, recent research [45] undertook a thematic analysis of listeners' comments whilst evaluating a media device orchestration approach to immersive spatial audio experiences. This allowed the authors to categorise specific positive and negative traits in their devised system. Other works in the field have utilised qualitative processes to identify salient features in audio distractors [46] or to validate the design of sound synthesis techniques [47].

Thematic analysis was carried out using the Nvivo 11 [48] software, which was used to code and organise themes as they emerged during the process. An initial study of all of the comments was carried out, followed by the formation of initial, high-level themes (distortion and noise), into which an initial set of coding was applied. Following this, the data, which had been coded using these two initial themes, were reread, resulting in increased granularity emerging, where more specific types of noise and distortion were identified, leading to subthemes and producing one additional top-level theme (timing). This was an iterative process, until no additional distinct themes could be identified.

6.2. Analysis. The resultant themes, and subthemes, are described in Table 9, where participant numbers accompany each statement in the example response column. These demonstrate the formation of three main themes related to description of impairments, along with a small number of associated subthemes.

To provide a broader context of the three themes and the descriptions elicited from the listeners, Figures 7, 8, and 9

TABLE 9: Summary of thematic analysis results.

Theme	Definition	n	Example Response
1. Distortion	Manipulation or processing of the original signal, altering it from its true state	88	"Some of the clips had a kind of "buzz" vibrating sound that other clips of the same piece didn't." (P4)
1.1 Amplitude	General presence of distortion or clipping	50	"A lit bit of distortion made the sound a little fuzzy." (P43)
1.2 Spectrum	Enhancement or diminution of frequency bands in the music	27	"…cutoff on highs - some tracks specifically those with live instrument seemed washed out/underwater when distorted (Take That)." (P16)
1.3 Vocal Clarity	Enhancement or diminution of the vocal in the music	8	"Sometimes the voice become clearer, purer." (P54)
1.4 Arrangement	Addition or removal of instruments or musical components in the music	3	"…choosing the wrong instruments in some clips, like horns. etc." (P83)
2. Noise	Presence of additional sounds that are not desirable	32	"Very much 'white noise' - sometimes 'echoy'. Some sounded like a record rather than digital." (P60)
2.1 Unwanted	General hiss, popping, cracking, etc., being present	29	"Noise could be heard in the background of some clips." (P05)
2.2 Echoes	Delay or reverberation effects that are noticeable or not in the true state of the music	3	"Echoy [sic] noises" (p91)
3. Timing	Temporal anomalies in the audio, where sequences or timing is not correct	29	"Clip 6 of one of the tracks seemed to "jump" and repeat." (P18)

FIGURE 7: Word cloud of participants' distortion theme descriptions.

FIGURE 8: Word cloud of participants' noise theme descriptions.

provide word cloud representations, created using Nvivo 11, up to a maximum of the 100 most frequently used words in each. In producing these graphical depictions, the stop words (irrelevant words used in descriptions such as "that", "seemed", and "sounded") were removed. Word stemming was also adopted, so that related words like "fuzz" and "fuzziness" are considered to fall into the same descriptor.

The size of each word represents its relative frequency of occurrence.

6.3. Results. The majority of responses received describe the presence of distortion, specifically amplitude-related effects, such as harmonic distortion, as well as the manipulation of frequency bands. This is not surprising, given the nature of the psychoacoustic codecs evaluated alongside ACER, where

FIGURE 9: Word cloud of participants' time theme descriptions.

the approach of splitting the frequency-domain transform of each frame of audio into subbands and allocating bits is commonplace. This explains many of the commonly occurring words in Figure 7, such as "distorted" and "fuzziness". However, it is useful to note that several of the songs used in the experiment make use of distortion as an artistic device, which may account for some of the descriptive feedback that has been elicited. This can be exemplified by a statement from one of the participants who appears to identify this fact:

> "I found it difficult to know if it was distortion or style of music. I found I may have said it was distorted on first hearing the music. Distortion was more the tone rather than a noise that shouldn't be there. So by listening more - the distortion wasn't there." (P79)

Whilst it is the case that distortion may be purposively present in the songs, presence of this technique should have been mitigated by the fact that it would be present in each codec's representation of the music to some extent.

The experience of unwanted noise reported by participants is likely to stem from similar issues as distortion, where variable allocation of bits between frames can result in a higher noise floor. This outcome was surprising because 192 kbps audio clips were used. It is especially interesting to note the set of responses in Figure 8 related to "crackle" and transients, unlikely to have been introduced by any of the codecs scrutinised.

The timing theme, as is postulated, arose because of the ACER clip versions. During the development of the technique, such artefacts were encountered and it is a known aspect of lower bit rate ACER audio that it can make music sound glitchy. With the exception of a small number of

descriptors in this theme, related to phase and frequencies, the majority of terms elicited are consistent with our experiences, evident through terms in Figure 9, such as "skipped" and "stuttered".

Of course, in terms of each of these three top-level themes and their respective subthemes, there is the possibility that the descriptions produced were because of subject-expectancy effect [49]. This is the phenomenon where subjects subconsciously articulate impairments in the audio because the questions posed have specifically asked about noise and anomalies. Whilst this may be true for the distortion and noise themes, there was no specific wording when asking about the temporal aspects of the clips. This analysis leads us to conclude that where ACER is able to perform comparably with its contemporaries, its limitations at lower-quality levels can be perceived and the constructs produced by our participants are valid.

7. Conclusions and Future Work

The ACER medium- and high-quality approaches not only perform as well as the contemporary psychoacoustic codes, MP3 and AAC at 192 kbps CBR, but also yield comparable scores to uncompressed WAV PCM audio. The low-quality ACER codec showed significant differences from the others in terms of noise and distortions, though not in terms of the quality of stereo image that it portrayed. These findings were supported by providing an analysis of participants' preference of codec, where the majority of negative preferences expressed were towards the low-quality ACER codec. This secondary approach of appraising the codecs assures and adds to the reliability of these conclusions. The results highlight that there was consistency across participants who were able to perceive differences between the ACER low-quality version and each of the others using an alternative method of assessment, which is a common practice of demonstrating interitem accuracy.

All codecs performed similarly in terms of the perceived stereo image presented to listeners. This demonstrates that the stereo field was maintained successfully in all versions of the music. Given that the songs used come from a compilation of popular music, where stereo panning is a common mixing technique used to add width to recordings, this is a notable finding. Any errors or anomalies incurred during the coding process should have been noticeable and easily perceived by the listeners, especially since they were using headphones and the stereo image they perceived will not have been influenced by factors in the room or due to their own head movements.

Although the ACER low-quality version resulted in poor evaluation results, in terms of noise and distortions, the outcome is beneficial in the wider context of the research. It contributes to the reliability of the overall results, since it demonstrates that the group of listeners who took part were able to perceive and articulate quality differences between ACER low quality and the other codecs. By contrast, if the results had shown complete homogeneity, this could have indicated success of the ACER low-quality version but would also have raised questions about the ability of the listeners to

tell the difference between the audio samples, bringing the credibility of the results into question. 37% of the participants indicated that they had some form of musical training and 17% had some professional audio training, with an overlap between the two groups of 14%, meaning that the majority were nonexpert listeners. These listener numbers more than comply with ITU-R guidelines [29] and demonstrate the effectiveness of nonexpert listeners. The subsequent round of development to the ACER codec would be a suitable time to perform more listening tests. This would be particularly appropriate in light of the results with untrained listeners that have been reported in this work. The use of expert listeners could provide a more critical appraisal of any differences in audio quality that may have gone undetected. Such future investigations would afford the use of methods such as ITU-R BS.1116 [34] or MUSHRA [35].

A perceived constraint of this study could be the choice of 192 kbps bit rate for the MP3 and AAC codecs. The decision was made to utilise this bit rate to reflect the de facto standard practice in the consumer audio market. As such, the non-ACER compression of each song in the study from uncompressed WAV to MP3 and AAC formats was undertaken using Apple's iTunes software, which describes MP3 192 as "higher quality", hence selecting it as the compressed benchmark bit rate. Our finding of no differences between the ACER high- and medium-quality versions, in terms of noise, distortions, and stereo field, leads to the conclusion that these ACER versions produce musical audio that is of a perceptually comparable quality to the 192 kbps compressed versions. More interesting still is the outcome that the 192 kbps MP3 and AAC versions, and the ACER high- and medium-quality songs, exhibited similar results against uncompressed WAV versions. This result is in contrast to the work of [20], discussed earlier, which found that MP3 bit rates had to be greater than, or equal to, 256 kbps to elicit such a result. However, the sample size (n = 13) used in [20] is much smaller than that in our study, which may explain this outcome. Further, homogeneity in ratings of MP3 and AAC coding variations of 192 kbps or more is consistent with the findings of [22]. This suggests that the comparison of ACER to higher bit rate MP3 and AAC would be a redundant exercise.

A limitation in the qualitative evaluation of the codecs was that listeners were not asked to leave comments about noise and artefacts specifically for each of the codecs they listened to. Due to the double-blind nature of the experiment, this would have necessitated asking participants to leave a comment about every audio sample they heard. As a result, it is not possible to know which of the codecs were unequivocally related to each of the themes that were devised from the qualitative feedback. Such an analysis would have added significant time and completion overheads to conducting the existing study; therefore it is proposed that this kind of enquiry would be suitable for a separate piece of future work. In such an investigation, participants could be asked to describe the qualities they perceive in a range of coded audio samples, without necessarily having to produce quantitative scores or to listen to so many clips. This would further validate the tentative conclusion presented here, which suggests that MP3- and AAC-coded audio presents distortion and noise-based impairments, whilst ACER compression introduces temporal glitches.

The ACER codec could be used for auditory interface cues that have a perceived musical element such as earcons [50]. Whilst earcons are not intended to be musical, they share many of the same properties and as such would be suitable candidates for this form of compression. Other forms of auditory interface cues that have repetitive elements such as spearcons [51] might also be suitable. Although the compression method was originally designed for longer audio files, the principles should still be suitable for short clips. Long form audio such as audio books might also benefit from this technique, as many vocal elements and especially pauses and breaths often exhibit similarities. The technique could also be used in noise-reduction software and games audio software to highlight differences and emphasise them to retain sonic interest.

The outcomes of this research indicate that the ACER codec, at medium- and high-quality settings, is highly functional as an alternative approach to contemporary techniques of MP3 and AAC, potentially making it suitable as a stand-alone codec, with moderate data reduction, or as a potential partner to psychoacoustic approaches to achieve even lower bitrates. The results demonstrate that the novel approach of ACER, which seeks out redundancies in music structure and pattern, is a viable technique and that listeners were not able to detect significant differences between it, other codecs, and uncompressed audio. Although there are artefacts and impairments introduced during ACER coding, which manifest themselves in the temporal domain rather than as amplitude-related distortions or noise, ACER audio retains a full frequency spectrum and resolution, making it distinct from MP3 and AAC.

The bitrates achieved using the ACER codec provide marginal gains over those achieved using WAV. This may be appropriate in situations where reduced data rates are desirable but losses in absolute audio fidelity, as a result of frequency manipulations and quantisation, are not permitted. This may be true in scenarios such as audio analysis tasks, computer game sound, forensic analysis, and multichannel formats where highly repetitive elements are confined to a single channel such as LFE in 5.1, 7.1, or Atmos systems or in the archival audio. Further, performance of ACER is dependent upon the level of musical repetition in the musical composition to be coded. This means that highly repetitive music will yield greater reductions in bitrate at the same ACER settings. With this in mind, it is possible that the ACER settings themselves can be tuned specifically to the piece of music being compressed, something which has not been attempted at this time. Ultimately, however, we propose that the most suitable application of ACER is as a preprocessing step before music is compressed using a psychoacoustic method, such as MP3 or AAC, providing an enhancement of current state of the art [52]. This would enhance the compression ratios already obtainable using these techniques on their own and is likely to have little impact upon the perceived quality of the listening experience.

Next stages in the development of ACER will focus upon refining the regression model used to determine the quality

of ACER files using the similarity between audio segments within songs. Creating a refined model will involve a series of focused listening tests, allowing us to determine the point at which these differences are perceived and when they become problematic or distracting. It is anticipated that a refined model will be able to achieve greater bit rate reduction and to improve the quality of perceptual similarity between clips, which may lead to the ACER low-quality version being able to compete with the medium- and high-quality versions, along with MP3, AAC, and uncompressed WAV.

Data Availability

The listening test data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] S. Cunningham and V. Grout, "Data reduction of audio by exploiting musical repetition," *Multimedia Tools and Applications*, vol. 72, no. 3, pp. 2299–2320, 2014.

[2] G. G. Rogozinsky, D. R. Fadeyev, and D. A. Podolsky, "Adaptation of Psychoacoustic analysis to wavelet domain in lossy audio coding," in *Proceedings of the 2017 Systems of Signal Synchronization, Generating and Processing in Telecommunications (SINKHROINFO)*, pp. 1–5, Kazan, Russia, July 2017.

[3] T. S. Gunawan, S. A. Rashid, and M. Kartiwi, "Investigation of various algorithms on multichannel audio compression," in *Proceedings of the 2017 IEEE 4th International Conference on Smart Instrumentation, Measurement and Application (ICSIMA)*, pp. 1–5, Putrajaya, Malaysia, November 2017.

[4] M. Sandler and D. Black, "Scalable audio coding for compression and loss resilient streaming," *IEE Proceedings—Vision, Image and Signal Processing*, vol. 153, no. 3, pp. 331–339, 2006.

[5] I. I. S. Fraunhofer, "Alive and Kicking – mp3 Software, Patents and Licenses — Fraunhofer Audio Blog," 2017, http://www.audioblog.iis.fraunhofer.com/mp3-software-patents-licenses/.

[6] S. Cunningham, J. Weinel, S. Roberts, V. Grout, and D. Griffiths, "Initial objective & subjective evaluation of a similarity-based audio compression technique," in *Proceedings of the the 8th Audio Mostly Conference*, pp. 1–6, Piteå, Sweden, September 2013.

[7] W. Yost, *Fundamentals of Hearing: An Introduction*, 5th edition, 2013.

[8] D. M. Howard and J. A. S. Angus, *Acoustics and Psychoacoustics*, Focal Press, 5th edition, 2017.

[9] D. A. Huffman, "A method for the construction of minimum-redundancy codes," *Proceedings of the IRE*, vol. 40, no. 9, pp. 1098–1101, 1952.

[10] R. F. Rice, "Some practical universal noiseless coding techniques," Tech. Rep., NASA Technical Report, Pasadena, Calif, USA, 1979.

[11] M. Hans and R. Schafer, "Lossless compression of digital audio," *IEEE Signal Processing Magazine*, vol. 18, no. 4, pp. 21–32.

[12] J. Coalson, *FLAC – Free Lossless Audio Codec*, Xiph.Org Foundation, 2014, https://xiph.org/flac/index.html.

[13] F. Ghido and I. Tabus, "Sparse modeling for lossless audio compression," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 1, pp. 14–28, 2013.

[14] H. Huang, H. Shu, and R. Yu, "Lossless audio compression in the new IEEE standard for advanced audio coding," in *Proceedings of the ICASSP 2014 - 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6934–6938, Florence, Italy, May 2014.

[15] D. Salomon and G. Motta, *Handbook of Data Compression*, Springer Science & Business Media, 5th edition, 2010.

[16] J. Moffitt, "Ogg Vorbis—open, free audio—set your media free," *Linux Journal*, vol. 81, no. 9, 2001.

[17] K. Brandenburg, "MP3 and AAC explained," in *Proceedings of the Audio Engineering Society Conference: 17th International Conference: High-Quality Audio Coding*, Audio Engineering Society, 1999.

[18] K. Brandenburg and R. Henke, "Near-lossless coding of high quality digital audio: first results," in *Proceedings of ICASSP '93*, pp. 193–196 vol.1, Minneapolis, MN, USA, April 1993.

[19] G. A. Soulodre, T. Grusec, M. Lavoie, and L. Thibault, "Subjective evaluation of state-of-the-art two-channel audio codecs," *Journal of the Audio Engineering Society*, vol. 46, no. 3, pp. 164–174, 1998.

[20] A. Pras, R. Zimmerman, D. Levitin, and C. Guastavino, "Subjective evaluation of mp3 compression for different musical genres," in *Proceedings of the 127th Audio Engineering Society Convention 2009*, pp. 459–465, USA, October 2009.

[21] A. Hines, E. Gillen, D. Kelly, J. Skoglund, A. Kokaram, and N. Harte, "ViSQOLAudio: an objective audio quality metric for low bitrate codecs," *The Journal of the Acoustical Society of America*, vol. 137, no. 6, pp. EL449–EL455, 2015.

[22] P. Pocta and J. G. Beerends, "Subjective and objective assessment of perceived audio quality of current digital audio broadcasting systems and web-casting applications," *IEEE Transactions on Broadcasting*, vol. 61, no. 3, pp. 407–415, 2015.

[23] K. Brandenburg, C. Faller, J. Herre, J. D. Johnston, and W. B. Kleijn, "Perceptual coding of high-quality digital audio," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 1905–1919, 2013.

[24] ITU-R, "Method for objective measurements of perceived audio quality," in *Proceedings of the International Telecommunication Union Recommendation*, 2001.

[25] M. Bodden, "Instrumentation for sound quality evaluation," *Acta Acustica united with Acustica*, vol. 83, no. 5, pp. 775–783, 1997.

[26] D. Campbell, E. Jones, and M. Glavin, "Audio quality assessment techniques—A review, and recent developments," *Signal Processing*, vol. 89, no. 8, pp. 1489–1500, 2009.

[27] G. Vercellesi, M. Zerbini, and A. L. Vitali, "Objective and subjective evaluation MPEG layer III perceived quality," in *Proceedings of the 14th European Signal Processing Conference, EUSIPCO 2006*, 5, 1 pages, Florence, Italy, 2006.

[28] W. Hoeg, L. Christensen, and R. Walker, "Subjective assessment of audio quality - The means and methods within the EBU," *EBU Technical Review, European Broadcasting Union*, no. 274, pp. 40–50, 1997.

[29] ITU-R, "General methods for the subjective assessment of sound quality," in *Proceedings of the International Telecommunication Union Recommendation*, 2003.

[30] B. Defraene, T. van Waterschoot, M. Diehl, and M. Moonen, "Subjective audio quality evaluation of embedded-optimization-based distortion precompensation algorithms," *The Journal of the Acoustical Society of America*, vol. 140, no. 1, pp. EL101–EL106, 2016.

[31] J. C. Garcia-Alvarez, S. E. Aguirre, and P. C. Diaz-Solarte, "Perceptual audio quality assessment for coder evaluation," in *Proceedings of the 2014 IEEE Fourth International Conference on Consumer Electronics – Berlin (ICCE-Berlin)*, pp. 408–410, Berlin, Germany, September 2014.

[32] L. Gaston and R. Sanders, "Evaluation of HE-AAC, AC-3, and E-AC-3 codecs," *Journal of the Audio Engineering Society*, vol. 56, no. 3, pp. 140–155, 2008.

[33] J. Villegas, T. Stegenborg-Andersen, N. Zacharov, and J. Ramsgaard, "Effect of presentation method modifications on standardized listening tests," in *Proceedings of the 141st Audio Engineering Society Convention*, Los Angeles, Calif, USA, 2016.

[34] ITU-R, "Methods for the subjective assessment of small impairments in audio systems," in *Proceedings of the International Telecommunication Union Recommendation*, 2015.

[35] A. J. Mason, "The MUSHRA audio subjective test method," R&D White Paper WHP 038, Research & Development, British Broadcasting Corporation (BBC), 2002.

[36] Statista GmbH, "Headphone usage in the US 2017 — Usage frequency of headphones in the United States 2017," 2018, https://www.statista.com/statistics/283620/us-consumer-purchase-plans-smartphone-accessories-2010/.

[37] Statista GmbH, "Headphone usage in the US 2017 — Purposes headphones are used for in the United States 2017," 2018 https://www.statista.com/statistics/696862/uses-of-headphones-in-the-us/.

[38] D. Watkins, "Computer speakers now most popular way people listen to music," in *Strategy Analytics*, 2019, https://www.strategyanalytics.com/strategy-analytics/news/strategy-analytics-press-releases/2015/12/17/computer-speakers-now-most-popular-way-people-listen-to-music.

[39] S. Bech, "Timbral aspects of reproduced sound in small rooms. I," *The Journal of the Acoustical Society of America*, vol. 97, no. 3, pp. 1717–1726, 1995.

[40] M. Velmans, *Understanding Consciousness*, Routledge, 2009.

[41] A. Hines, J. Skoglund, E. Gillen, A. Kokaram, D. Kelly, and N. Harte, "Perceived audio quality for streaming stereo music," in *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 1173–1176, 2014.

[42] Various Artists, ""Now that's what I call music! 90". Compilation [Double Audio CD]. Now! Music," 2015.

[43] IEC RB, "Audio Recording-Compact Disc Digital Audio System, IEC 60908," 1999-2002.

[44] G. Guest, K. M. MacQueen, and E. E. Namey, *Applied Thematic Analysis*, SAGE Publications, Thousand Oaks, Calif, USA, 2012.

[45] J. Francombe, J. Woodcock, R. J. Hughes et al., "Qualitative evaluation of media device orchestration for immersive spatial audio reproduction," *Journal of the Audio Engineering Society*, vol. 66, no. 6, pp. 414–429, 2018.

[46] J. Francombe, R. Mason, M. Dewhirst, and S. Bech, "A model of distraction in an audio-on-audio interference situation with music program material," *Journal of the Audio Engineering Society*, vol. 63, no. 1-2, pp. 63–77, 2015.

[47] S. Conan, O. Derrien, M. Aramaki, S. Ystad, and R. Kronland-Martinet, "A synthesis model with intuitive control capabilities for rolling sounds," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 8, pp. 1260–1273, 2014.

[48] QSR International, "NVivo qualitative data analysis software — QSR International," 2018, https://www.qsrinternational.com/nvivo/home.

[49] D. J. Stang, "On the relationship between novelty and complexity," *The Journal of Psychology: Interdisciplinary and Applied*, vol. 95, no. 2, pp. 317–323, 1977.

[50] M. Blattner, D. Sumikawa, and R. Greenberg, "Earcons and icons: their structure and common design principles," *Human–Computer Interaction*, vol. 4, no. 1, pp. 11–44, 1989.

[51] B. N. Walker, J. Lindsay, A. Nance et al., "Spearcons (speech-based earcons) improve navigation performance in advanced auditory menus," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 55, no. 1, pp. 157–182, 2013.

[52] V. Rao and K. Pohlmann, "Audio compression using repetitive structures," *U.S. Patent Application*, 2006.