# Waveform Analysis for High-Quality Loop-Based Audio Distribution

Stuart Cunningham
*Centre for Applied Internet Research (CAIR), University of Wales / NEWI*
Plas Coch Campus, Mold Road, Wrexham, LL11 2AW, North Wales, UK
Tel: +44(0)1978 293583 Fax: +44(0)1978 293168
s.cunningham@newi.ac.uk

## Abstract

*Existing compression techniques for waveform audio still rely heavily on the removal of data from the stored representation – lossy compression. This reduces the quality and is particularly an issue when the audio file is to be distributed over low bandwidth communication media. Small file sizes can be achieved using compression procedures, but the quality of the audio suffers, to the detriment of the reproduced audio, and thus, the listening experience. Though acceptable audio quality can be achieved using these compression methods, the file sizes increase dramatically at acceptable levels and become unsuited for low bandwidth transmission purposes.*

*A possible solution to this problem is to use a compression procedure which maintains the original quality of the audio, whilst exploiting perceptual coding techniques. Using such a method would provide a high quality audio representation, and attempt to reduce the size of the file by making use of waveform analysis and comparison techniques.*

*By performing such analysis, and producing a representation of the audio in a small distributable file a high quality of audio could be maintained, whilst the only factor to be lost would be information that is deemed to be unperceivable to a degree. Sufficient audio chunks can be built up so that the entire original audio can be reproduced by using as few chunks as possible, to attain the minimum file size an accurate representation.*

*This document represents initial theories and research conducted into the possible development and application of such a compression method.*

## 1. Introduction

Over recent years the Internet has become a popular medium through which audio, and in particular music, has been distributed. The advent of MP3 compression technology coupled with the spread of high-speed home Internet access has caused a dramatic increase in the number of users downloading and uploading audio files, and consequently the amount of data being generated over networks.

The effort for improved compression algorithms is a continual process, with the goal of minimising the stored file size of an audio representation, whilst still maintaining a quality which is acceptable to the listener. The perfect solution would produce a minimal data representation that produced an inaudible difference, if any, in audio quality to the original sampled waveform. With the sharing and distribution of musical audio currently to the fore, an ideal scenario would be to achieve maximum compression of the file size, whilst still maintaining at least CD quality audio. The 'best' scenario would, in fact, be to employ electronic transfer of audio in CD quality format. However, it is widely accepted that such a format does create large amount of data [1] [2]. Such volumes of information are generally unsuitable for acceptable transmission times across low bandwidth networks. Therefore, compression is a necessity.

Through research and experimentation a new form of compression could be developed which would exploit these requirements for audio distribution, whilst minimising the data generated as much as possible, or to varying user-definable degrees. This particular method of compression should not affect, in any way, the quality of the audio, as it would employ the original sampled audio data, without altering the sampling rate or bit-depth; thereby preserving the audio quality.

This data reduction process will work by analysing existing information within a waveform and repeating sections of it as required, and disposing of the perceptually identical replicas contained in the file. By doing this the data is repeated or looped rather than storing two versions of the same thing. This particular method of reducing the data size will use a form of compression that is provisionally titled '*perceptual runtime compression*'.

## 2. The problem with existing compression techniques

Currently, all major and de-facto standard methods of audio compression designed with distribution in mind rely on lossy compression algorithms, which store the sampled waveform data in its entirety. By doing this, the full waveform of the original audio has to be encoded at some level or another. There are two main weaknesses which are identified with these encoding methods:

1. These compression processes perform no waveform analysis, so therefore will encode the audio regardless of whether or not the entire audio stream was simply the repeat of an audio sequence, or loop, a number of times. Encoding is performed without consideration for the perceivable audio content. Essentially the original sample waveform is encoded in its entirety, albeit in a manner where some degradation of the waveform's depiction occurs.

2. The actual encoding process involved in using these algorithms relies on lossy compression techniques to reduce the size of the stored file. Although the amount of compression can be controlled when creating audio files using these algorithms, some data is always thrown away because it is deemed irrelevant, resulting in a loss of quality. As the amount of compression is increased, the size of the file is reduced, but the quality of the audio is degraded. These changes occur exponentially.

By using such lossy compression techniques, small file sizes can be created, but the important issue is the loss of audio quality when doing this, and the fact that there is no analysis of the sample to discern its quantifiable content. Even algorithms based on the MPEG-1 Audio Layer 3 (MP3) compression technique, which uses a form of perceptual compression, will suffer from a noticeable loss of audio quality once the ratio of compression goes beyond a certain threshold.

## 3. Perceptual runtime compression - Discernment

The application of perceptual compression methods has been widely recognised in the field of digital audio representation and storage [1] [2], but the main methods employed in such techniques to date, has focussed on the masking phenomena almost exclusively.

The way in which file data is reduced in this system does rely on the perception of the listener's ear, but does not work solely by using the limitations in frequency or dynamic characteristics of the human ear to reduce the size of an audio representation, in the way some compression methods like MP3 do. The perceptive encoding in this particular case will operate not only by the limits of the human ear, but by noticeable interpretation or difference between two similar sounds.

Though the theory of perceptual runtime compression could be applied to any audio waveform, its primary, and most efficient, use will be made in the compression and distribution of musical audio. This is the target area for the compression technique, mainly due to the common, structured, nature of music as well as the complex range of sounds that occur in musical recordings. The successful application of this technique to music will vary to extremes of the spectrum depending on the particular genre and style of a musical piece.

For example, imagine a piece of music recorded by a band where the same chorus will have been played twice by the group during the recording process. The music in the second chorus would be intended to be reproduced by the group using the same music and words, to the same tempo as the first time the chorus was played. It is unlikely though, that the recording of the group playing the second chorus would be 100% the same as the first chorus which had been played. There are many factors which could affect the performance of the second chorus during recording, many of which are attributable to human error, or to uncontrollable anomalies occurring. The guitarist could play a note which had a small amount of natural tremolo on it, a bass player may not hold a note for a few milliseconds longer, the drummer might hit the snare drum harder and not quite in the same spot. All of these occurrences may not have arisen in the first chorus, but we could not say that the second chorus was totally different to the first. It may not be exactly the same, but it would not be radically different. The important concept derived from this is: "*Would the listener be able to detect such subtle nuances and differences in a particular piece of music?*"

This method relies on the principle that human perception of audio signals is frequently inaccurate, prejudiced, and easily influenced, therefore perceptual compression operated by representing a signal to the listener 'as it sounds' rather than by encoding all of the data required to represent the signal fully [2]. The whole concept behind this theory is that a similar piece of audio could be used in place of a similar, or potentially identical, portion of waveform information, which would otherwise have to be encoded as part of the final file distribution. The perceptive aspect of the listener which is being exploited is primarily their ability to distinguish between two similar pieces of audio, if, indeed, the original sample contained two different pieces, rather than the same piece which had been identically encoded in the waveform. There are two main issues that have to be addressed if using such a technique, from the point of view of the psychoacoustics behind the listener's perception of a piece of audio:

1. Does the listener know that the same piece is being repeated?
2. Does the listener care if the same piece is being repeated?

Depending on the type of sample which is being processed, the perception of the audio may not even be affected in any way at all, if the exact same piece of audio

reappears at a later time for example. Of course, when employing this technique the feature would be provided to allow the creator of the compressed data to specify how accurate, and true to the original sample, these wave sections (chunks) that would be used are. But for types of musical composition where one sample has been used and the same waveform repeated or looped later in the composition, then there would be no detectable difference in the resulting representation created using perceptual runtime compression. This would make such a technique particularly suitable for certain genres of music, where repetition of the same piece is intended, and commonplace. Immediate examples of where this technique could be potentially highly successfully applied are in the representation of dance, trance, and electronic music.

## 4. Perceptual runtime compression - Application

It is not necessary to limit the use of the perceptual runtime compression theory purely to musical audio, although this is the desirable goal. The process would have application in a wide variety of non-musical audio information. For example, consider the recording of a radio interview, between an interviewer and an interviewee, which is to be recorded at CD quality. In a fifteen minute interview there will be a certain, measurable amount of silence, in the duration of that conversation. This silence will still be encoded at 44.1 kHz and 16-bit depth, regardless of how long any piece of silence lasts. The perceptual runtime encoding could be applied to this situation, where only a fraction of the silence would need to actually be stored (still at 44.1 kHz and 16-bit) and simply looped and repeated as necessary, wherever silence occurred in the original sample, rather than waste storage space by encoding redundant information.

For example, take a 1 minute sample from the interview, encoded at 44.1 kHz / 16-bit depth, in stereo. The file size on disk is approximately 10,584,064 bytes (10,336 KB or 10.09 MB)

If the total amount of silence in the sample was added and there was 20 seconds worth of silence in the file, this would mean that approx. 3445 KB of the file was used, in total, to encode silent audio.

Using the perceptual runtime compression method, only a small amount of silence would need to be encoded, and simply repeated and looped where required to produce a faithful representation of the original audio. If half a second of silence was encoded, and repeated and looped as required, then we could dispose of the other 19.5 seconds of silence which occurred in the original sample. From a mathematical point of view, this has the effect of deducting 20 seconds of silence from the waveform, and adding 0.5 seconds worth of silence back into the file. To illustrate this, it is useful to look at the file again, in terms of its stored size:

Half a second (0.5 seconds) of CD quality audio requires approx. 86 KB of storage space.

Twenty seconds (20 seconds) of CD quality audio requires approx. 3445 KB of storage space.

Given theses values, and that the original size of the audio file was 10,336 KB, we can perform the following calculation to determine how much the file size would be reduced by removing 20 seconds of audio and encoding 0.5 seconds of audio in its place:

$$(10336 - 3445) + 86 = 6997$$

$$\text{Total Saving} = 10336 - 6997$$

$$= \mathbf{3339\ (32.3\%)}$$

Therefore, by reducing the amount of silence contained in the waveform, and replacing it with a reusable half a second if silence, we have cut the size of the file by approximately 3,339 KB (3.26 MB). By simply replacing only the silent elements of the original waveform, which is perhaps one of the simplest uses for this method of compression, we have reduced the size of the original file by 32.3% without having altered the quality of the audio in terms of the bit-depth or sampling rate.

By the same rationale, this could also highlight a weakness in using this particular technique to encode non-musical (in this case speech) audio. It's highly possible in this scenario mentioned above that the word "and" would be used by both the interviewer and the interviewee on more than one occasion. This is something that could be exploited using the perceptual runtime compression method, so that the word "and" would only need to be recorded and saved once for each person speaking. However, unless the waveform was carefully trimmed, and even manipulated, it is likely that the listener would become aware that the flow of speech was unnatural and may even become choppy or stuttering. Due to this fact it is obvious that some kind of control mechanism would be required over the waveform analysis both to prevent such unwanted events, and to allow control over the amount of compression that will occur in the final file.

However, due to the creative nature of music, and the flexibility that must be permitted when using such a compression technique, the limitations of use should be decided only by the creator of a distributable file. This will allow the end result to perform to their requirements. The main areas of control which the user must have over the compression process, and degree of change of the audio, are:

1. Minimum chunk length (seconds)
2. Maximum chunk length (seconds)

3. Minimum number of chunks
    a. Number of each type of chunk
    b. Number of chunks in the entire representation
4. Maximum number of chunks
    a. Number of each type of chunk
    b. Number of chunks in the entire representation
5. Similarity between existing chunk and section to be deleted (percentage)
    a. Minimum similarity between waveforms (percentage)
    b. Maximum similarity between waveforms (percentage)
    c. Variable threshold settings (e.g. between 40 and 80%)
6. Bit depth of chunks (binary bits)
7. Sample rate of chunks (Hz)

Settings 4 and 5 are at the discretion of the file creator, they may wish to keep the original sample rate and bit depth (one of the main advantages of this compression method), although providing these features provides a far more adaptable and flexible mechanism for keeping file sizes to a minimum, and tailoring output specifically to the users needs. Options 4 and 5 could be provided by standard methods such as PCM, or could employ third party codec to supply compression methods such as MP3 or RealAudio. With the coupling of this proposed compression technique, as well as an existing compression method, such as MP3, the amount of overall compression achievable could feasibly reach levels previously unattainable whilst still maintaining an acceptable level of perceivable quality.

The example given of using silence would work particularly well as a sampled chunk, because very little analysis would be required to employ it, whereas more complex waveforms would almost certainly require longer sampled chunks. Such a technique, which was employed on silence in audio, could have a considerable impact on the amount of storage space saved. This would be particularly pertinent in multi-channel recordings, where sounds were panned only to certain channels at a time, and the other remained silent.

## 5. Testing the applicability of reusable audio samples

Some initial testing was required to assess the usefulness and scope of such a proposal for distributing music and audio which has been processed in such a way. To this extent a small set of tests were constructed to determine the possibility of further research and development into this field of audio manipulation. The main aim of these tests were to assess if a listener could distinguish between audio samples that had been directly recorded and samples which had sections replaced by previously encoded audio chunks and repeated to compensate for the lack of 'original' information.

It should be noted that tests which were conducted were considered to be a pilot study, and so were not implemented so under laboratory conditions or supervised. The tests were conducted by placing a form on the World Wide Web and asking users to visit the web site, listen to sound samples, and complete the associated questions. Currently there are future plans to perform far more detailed lab studies and tests under controlled conditions.

The test was placed online [3] and emails inviting participation were issued. The test involved ten questions, and was split into two sections, each containing five questions. In section one, each question asked the user to listen to, and compare, two sounds, but the user was only instructed to listen to each pair of sounds once. Each question then asked the user to decide if the two sounds were different or not. If the user stated that they believed the two sounds were different, then they should provide detail as to what they thought was different. Section two used the same five questions and the same pair of audio clips, as in section 1, for each question. This time however, the user could listen to each pair of sounds as many times as they wanted to. The questions in section two also contained an additional field, to allow the user to enter how many times they had compared each pair of sounds. Certain questions also had additional follow-up enquiries in section two. The first section forces the listener to only compare the sounds once, to try to emulate passive listening of audio, whilst the second section allowed the user to listen to the audio as many times as desired, to emulate methodical listening.

The first question used two recordings of a phrase being played on an acoustic guitar. The first audio clip was a direct recording of the guitarist playing the same phrase twice. The second audio recording contained the phrase the first time it had been played, and was trimmed and looped, so that that phrase was again, played twice. The aim of the question was to assess if the listener could tell the difference between the two sounds. The main difference is when the phrase is played the second time in both recordings.

The sounds in the second question used a recording of a complete song, which contained multiple instruments and singing. The first audio clip was several bars of the song, with several lines of vocals. The second audio clip was the same phrases of the song, but which had been extracted form another part of the original recording, this time the same vocals were sung to similar music, but the most notable addition was an additional vocal harmony, "*oohs*", in the background. This question aimed to see if

the listener would notice the addition of the background vocals in an otherwise similar piece of audio. Additionally, would the fact that the listener heard the background vocals prevent them from noticing that there were subtle differences in the accompanying music?

The third question was placed as essentially a placebo. Both sounds in this question were identical and there was no difference whatsoever between the audio samples. This was placed simply to attempt to catch the user out, and to see what sort of responses were received if they thought that the audio was different.

The fourth question presented the user with the same audio recording, a sample of several phrases being played on an electronic synthesiser, but the first clip contained a version of the recording which had a band-pass applied to it. All frequencies below 700 Hz had been cut, as had everything above 12.5 kHz. There was also a notch of cutting between 1850 Hz and 3100 Hz. The second sound was unchanged, and retained a full frequency spectrum. This test was placed to ascertain if the listener would detect the changes in frequency present, and how this affects their perception of this audio sample.

The fifth question contained two distinct audio samples of a distorted electric guitar playing a riff. Both audio samples started on the same note, but after that noted had been played the riff changed towards the end of the phrase. The first sample followed more strict timing changes and progressions, whilst the second sample was more ad lib and played freely. This would test to see if, what were two individual sounds were considered to be similar, because of the same tonality and timbre of the guitar, and both riffs having started in a very similar manner.

Thirty users had responded to the test at the time of writing and the responses that were received are summarized and presented in Table.1 for section one questions and Table.2 for section two questions.

**Table 1: APT01 Results – Section 1**

| SECTION 1 | | | |
|---|---|---|---|
| | | | |
| ***Question 1*** | | YES | NO |
| Is there a difference? | | **20.0%** | **80.0%** |
| | | | |
| ***Question 2*** | | | |
| Is there a difference? | | **90.0%** | **10.0%** |
| | | | |
| ***Question 3*** | | | |
| Is there a difference? | | **16.7%** | **83.3%** |
| | | | |
| ***Question 4*** | | | |
| Is there a difference? | | **96.7%** | **3.3%** |
| | | | |
| ***Question 5*** | | | |
| Is there a difference? | | **70.0%** | **30.0%** |

**Table 2: APT01 Results – Section 2**

| SECTION 2 | | | |
|---|---|---|---|
| | | | |
| ***Question 1*** | | YES | NO |
| Is there a difference? | | **16.7%** | **83.3%** |
| Ave comparisons: | **1.8** | | |
| | | | |
| ***Question 2*** | | | |
| Is there a difference? | | **96.7%** | **3.3%** |
| Ave comparisons: | **1.4** | | |
| | | | |
| ***Question 3*** | | | |
| Is there a difference? | | **10.0%** | **90.0%** |
| Ave comparisons: | **1.8** | | |
| | | | |
| ***Question 4*** | | | |
| Is there a difference? | | **96.7%** | **3.3%** |
| Ave comparisons: | **1.4** | | |
| | | | |
| ***Question 5*** | | | |
| Is there a difference? | | **86.7%** | **13.3%** |
| Ave comparisons: | **1.7** | | |

These figures gained from this provide an insight into the perception ability of listeners. The difference between the results in section one and section two are not significantly different, except in the case of question five. Otherwise the variance in answers between the two sections in statistically minimal. This could be attributable to many various factors, such as ability of listeners hearing, or even due to the limits of the number of participants in the test.

The results of question one were as expected, with few people detecting that there was a difference between the two sounds, which had been considered very similar, though different, when setting up the test.

Question two went a small way to show that more detailed listening would provide the listener with an opportunity to notice the background vocals, and showed that the majority of listeners detected this difference in the first instance of listening.

Question three provided an interesting result. It showed that some listeners, although a minority, believed that the two identical sounds used were in fact different. This indicates at another area of possible research into the perception of hearing.

Question four produced expected results, with the vast majority of listeners noticing that the two sounds were very different in terms of the frequencies being reproduced. Interestingly however, certain listeners commented that this changed the overall perception of the sound, to the extent that they though it was a different instrument being played.

Question five produced the results which showed the most significant difference between sections one and two. Further listening had proved that a difference in the sounds became apparent after carefully listening to the sounds again a number of times.

Detailed analysis of the additional feedback generated by individual users was also gathered during this testing process, but analysis of these comments is beyond the scope of this paper.

Overall the results gained were generally expected. Indicating that listeners had difficulty determining well-constructed loops from the original recordings, and that pieces of audio which were different, but had many qualities in common, could also take more careful listening to determine the difference.

## 6. Waveform analysis

Initial testing was carried out by taking a piece of music which was approximately 8 minutes 35 seconds in length, and sampled at 44 kHz with 16-bit depth. The size of this file was 88.6 MB (90,886,144 bytes). This file was then manually divided into sections and 12 distinct chunks, which were listened to and deemed to be either very similar to other sections repeated in the song, or chunks which were not repeated but were necessary to produce a proper rendition of the original. These chunks were extracted from the original. The theory being that with these 12 chunks the original musical piece could be recreated accurately. All chunks were saved at the same quality as the original sample.

These 12 chunks would be used to produce a copy of the original piece of music, by re-using and repeating the chunks where necessary. As this was done manually and was a time-intensive task, the chunks which were extracted were quite large, and varied in size from 761 KB to 26.8 MB, with the statistical average chunk size being in the region of 6.2 MB. This demonstrates that the test was initial, rough, and far from being optimised. These chunks were then manually re-assembled and repeated to form a representation of the original file. The total disk space used by the stored chunks was 74.8 MB. This gives a difference of 11.8 MB from the original waveform (90,886,144 bytes - 78,438,400 bytes = 12,447,744 bytes), a saving of 13.7% on the original file size. Though this saving is not of an exceptional proportion in comparison to the original file, it is important to remember that CD quality audio was maintained throughout, and the difference in perceivable quality, and true to the original, was not easily detectable, although no official tests were conducted. As previously stated, some of the chunk sizes were large, and manually edited. With more time this process could have been further optimised, and with the assistance of a software application, as previously discussed, to perform such tasks, the saving which could be made to the same file would be, theoretically, heavily improved.

## 7. Distribution

An important issue for users of such a compression mechanism will be how the final product or audio piece will be ultimately presented and distributed. Clearly, the proposition of splitting an original audio stream into reusable chunks, which have to be reassembled, could cause problems in the case of storage and distribution. Fragmentation and accidental loss of these chunks would cause problems, and result in the representation to be un-true to the original piece. This, coupled with the need for an instructional document which re-arranges the chunks, means multiple files and information chunks. Therefore; it would be desirable to have one single file, which contained all of the necessary audio chunks and the re-assembly information in order to make distribution and storage reliable, and perhaps more importantly user friendly. This could be likened to the use of popular computer compression tools, where multiple files can be kept within one file on the computer such as Zip files, and is also a similar technique employed when storing multiple Java program class files and resources within a Java Archive (JAR) file. The only difference required for such an implementation and grouping of these audio files, would be that no compression of the files would be required, and indeed would not be wanted.

This would be an addition to the actual encoding process and could be looked upon as a sideline to the development of the compression technique itself. However, it is an important issue which has to be considered whilst developing the main functionality of such an audio manipulation system. This would be the desired delivery method for the final audio chunks which had been produced.

It is also likely that some form of software application plug-in or codec would have to be written to read these files into standard wave and music playback software such as Microsoft Windows Media Player or Nullsoft's Winamp. Although these applications have the playback facility for uncompressed wave based files, the structure and ordering of the command or instruction file would have to be interpreted in some way. This could also be extended to such applications as CD, SACD, or DVD burning, should the resulting audio product need to be placed on an audio disc.

## 8. Conclusions

Although discussion has centred on the principle of encoding CD-quality audio, this technique could feasibly be applied to any quality of audio sample. The procedure could theoretically be applied to both the lower and upper ends of the quality of standard PCM wave files. However, there may be issues relating to the rate of success at which useful audio analysis could be performed, depending on the sampling rate and bit depth of low and high quality audio. This presents a further sphere or research to be carried out.

The way that this encoding method exploits the similarities that occur in music, the amount of data reduction which could take place is potentially huge. Particularly with applications of music audio, and varying musical genres, the file size of a high quality waveform representation could be reduced to a level only limited by the degree of perception where the audio becomes untrue to the original, whilst still maintaining the original quality (sampling rate and bit depth) of the original waveform.

The initial concept shows promise, and the preliminary tests that were performed indicate that there is a great deal of scope in this field, both in terms of academic research, and product development. The ramifications of a successful implementation of such a system go much further than purely a data saving method, but also impinge on many other aspects of the music and audio industries.

Content-based searching is a field where the use of small, searchable, clips of audio could prove tremendously useful, saving search time, and diversifying content. The waveform analysis and comparison techniques could be used in legal situations, to settle differences about how similar or possibly copied; one piece of music was to another.

The concept of such an analysis and encoding system opens up many avenues for future research and development activity, which will cross multidisciplinary fields.

## 9. References


[1] Rumsey, F., *The Audio Workstation Handbook*, Focal Press, Oxford, UK, 1996

[2] Sellars, P., *Behind the Mask: Perceptual Coding-How MP3 Compression Works*, Sound on Sound Magazine, SOS Publications, Cambridge, UK, May 2004
Available on-line:
http://www.soundonsound.com/sos/may00/articles/mp3.htm

[3] Cunningham, S., *Audio Perception Test – APT01*, University of Wales / NEWI, UK, 2004,
Available from:
http://www.newi.ac.uk/cunninghams/research/


## 10. Acknowledgments


Thanks to: Dr Rich Picking (NEWI), Dr Vic Grout (NEWI), all of my students who were test subjects in the online prototyping stages, and everyone who participated in the final online perception test.