14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

OBJECTIVE AND SUBJECTIVE EVALUATION MPEG LAYER III PERCEIVED QUALITY

Giancarlo Vercellesi, Martino Zerbini

Laboratorio di Informatica Musicale (LIM)
Dipartimento di Informatica e COmunicazione (DICO)
Università degli Studi di Milano
Via Comelico, 39 - 20135 Milano - Italia
phone: + (39) 02 50316382, fax: + (39) 02 50316373,
email: vercellesi@dico.unimi.it, martino.zerbini@studenti.unimi.it

Andrea Lorenzo Vitali

STMicroelectronics
Centro Direzionale Colleoni - Palazzo "Dialettica"
20041 Agrate Brianza (MI) - Italia
phone: (+39) 039 603 7244, fax: (+39) 039 603 6129,
email: andrea.vitali@st.com

ABSTRACT

Artefacts due to MPEG layer III (MP3) coding are analyzed: non uniform quantization noise, hybrid filter bank aliasing, MDCT pre- and post-echoes. Both objective and subjective evaluation of several MP3 decoded audio files are presented. Finally, it is shown how to predict the subjective perceived quality based on objective measures.

1. INTRODUCTION

MPEG Layer III, also known as MP3, is one of the most popular algorithms to compress digital audio signals. It is used in many applications: web radio, voice over IP, multimedia streaming, music on-demand and P2P.

MP3 is lossy: to achieve its high compression ratios it removes some information which is believed not to be perceivable. This causes a quality loss which can be assessed in a subjective or objective way.

In case of subjective quality assessment, a session of listening tests must be conducted. The session must be set-up carefully in order to guarantee that the perceived quality loss is due only to compression artefacts. ITU provides some recommendation [5] [6] which address typical issues of session set-up. The quality is evaluated with respect the original reference signal. The resulting index is named SDG (subjective difference grade).

In case of objective quality assessment, a set of acoustic parameters like SNR (Signal to Noise Ratio) or NMR (Noise to Mask Ratio) is computed. Parameters should be chosen so that their value is strongly correlated to the perceived quality loss. ITU provides a flawed and underspecified recommendation [4] to evaluate the perceived audio quality with two methods: Basic PEAQ (Perceptual Evaluation of Audio Quality) and Advanced PEAQ. Both are full-reference quality indexes: they are computed with respect to the original reference signal. The resulting indexes are named ODG (objective difference grade).
There are three typical cases for quality assessment:

- *first coding* where the signal under test is the result of a single coding and decoding operation.
- *tandem coding* where the signal under test is the result of multiple coding and decoding operations, coding parameters (target bitrate) being equal.
- *bitrate scaling* where the signal under test is the result of multiple coding and decoding operations, coding parameters (target bitrate) being different: single step from 320 kbps to 128 kbps, double step from 320 kbps to 256 kbps and then from 256 kbps to 128 kbps.

In this paper we focus on the first case. Other cases will be discussed separately. The ODG was computed using the Basic PEAQ software available from [17].

The paper is structured as follow: in section 2 we present related works; in section 3 we provide a short overview of MP3 algorithm; in section 4 we describe specific artefacts that are caused by MP3-like compression; in section 5 we illustrate the procedures we have followed to perform both objective and subjective quality assessments; in section 6 we show the results; in section 7, we draw the conclusions and show how the subjective quality can be predicted from the objective quality.

2. RELATED WORKS

In 2000 EBU [9] conducted a series of subjective listening tests to evaluate the quality of internet audio codecs [5]. The MP3 codec under test was the Opticom version at 16 kbps, 20 kbps, 32 kbps, 48 kbps, 64 kbps. Its quality was rated "Poor" at lowest bitrates, and "Good" at highest bitrates.
In 2003 EBU conducted another series of subjective tests [10]. Among the codecs there were the same MP3 codec as above as well as new low-bitrate codecs. As before the higher the bitrate the higher the quality with one exception: going from 16 kbps mono to 20 kbps stereo with the old generation codecs (such as MPEG-2 Layer 3) caused a noticeable loss of perceived quality.
Vanam in [11] compared the simple PEAQ (Basic and Advanced) with the EEA (Energy Equalization Approach) method, used alone or as an additional MOV of Advanced PEAQ with the SDG. Better performances were achieved with the last one.
In our test we compare the objective and subjective using a wider range of bit rates and the basic implementation of the

PEAQ and following the ITU-R BS.1116 (for the subjective).

3. OVERVIEW OF MPEG LAYER 3

A short overview of the MP3 audio compression algorithm is provided in order to better understand the rest of the paper.
MP3 is a standard for lossy audio compression [1] [2]. It uses non-uniform quantization in the frequency domain. The quantization is driven by a perceptual model.
A hybrid filterbank is used to process incoming PCM samples. It is made by a polyphase filterbank [1] and a cascaded Modified Discrete Cosine Transform (MDCT) [1]. The polyphase filterbank is made of 32 filters whose output is critically downsampled 32:1. PCM blocks are processed by the filterbank and converted into 32 frequency subbands. Overlapped blocks of frequency coefficients are windowed and transformed by the MDCT. The MDCT further splits each subband into 18 finer subbands, also known as frequency lines. Short and long blocks are adaptively selected to trade-off time-resolution (short blocks) and frequency resolution (long blocks). Short blocks are used during transients.
MDCT coefficients are quantized by a non-uniform quantization. The quantization step is chosen based on the signal-to-mask ratio (SMR). The SMR is computed by means of a psychoacoustic model based on a 1024-point FFT. Usable quantization steps have a quantization noise which is below the mask level. Hence, the signal-to-noise ratio (SNR) must be greater or equal to SMR minus the noise-to-mask ratio (NMR).
Quantized MDCT coefficients are compressed by a lossless entropy coder using fixed Huffman tables. Finally, the syntax is added and the MP3 bitstream is generated.

4. AUDIO ARTEFACTS INTRODUCED BY MP3 COMPRESSION

Artefacts are due to several reasons:

- Noise caused by non uniform quantization.
- Pre and post echoes due to the MDCT applied to overlapping and windowed blocks.
- Ripples due to aliasing between adjacent subbands of the polyphase filter bank.
- Loss of stereo image caused by compression.

In this paper we do not consider stereo compression artefacts.

4.1. Non Uniform Quantization

The quantization step is computed algorithmically. There are two nested loops: the inner loop, also known as the rate loop, and the outer loop, also known as the noise loop.
In the inner (rate) loop, the quantization step is computed following an optimal power-law formula. Frequency lines are grouped in regions coded with different Huffman tables. If regions cannot be coded or the target bitrate is exceeded, the quantization step is iteratively re-computed. The quantization step is increased by incrementing the global gain.
In the outer (noise) loop, the quantization noise is checked. For every frequency subband, if the quantization noise is not masked, then the quantization step is decreased by incrementing the local scale factor.
This should ensure that the first kind of artefacts, the quantization noise, is not audible. Other artefacts, such as birdies and bandwidth limitation, may be introduced. They may change some perceived parameters (e.g. the timbre), which are strongly related to spectrum shape.
Bandwidth limitation happens when the bitrate value is very low. In this case, the encoder favours low frequencies. Furthermore, high frequencies can be cut by time-domain filters applied before the coding process. In most cases, the audio is still recognizable but the perceived quality is low.
Birdies also happen at low bitrates. In this case, the slight variation of masking thresholds between two adjacent frames may lead to very different bit assignments. As a result, groups of spectral frequencies may appear and disappear. This kind of artefact causes a strong reduction of the global perceived quality. It has been reported for objective perceptual assessment methods [13].

4.2. Hybrid Filter Bank Aliasing

The hybrid filterbank introduces aliasing in the analysis phase, which is partially removed in synthesis phase. We have two different aliasing, one generated by polyphase filter bank and one by MDCT. In the decoding phase, only the MDCT aliasing is completely eliminated [14], while the polyphase aliasing is only reduced [15]. The polyphase aliasing introduces ripples of 0.07 dB in the decoded signal [12]. Generally they are inaudible; they can not be detected by subjective evaluation. However they may be detected by objective measures.

4.3. MDCT Pre and Post Echoes

Pre and post echoes may be introduced by the MDCT block processing during transients: transients' energy in time spread over subbands in frequency; if long blocks are used, there are many subbands and they have to be quantized roughly. Quantization noise which was supposed to be masked, may become un-masked. MP3 encoders reduce these kinds of artefacts switching to short blocks during transients. Echoes introduced in short blocks are still there but they may be completely masked thanks to temporal masking.

5. PERCEIVED QUALITY OF DECODED MP3 AUDIO

The methods used to evaluate objective and subjective perceived quality are described in [4], [5] and [7].
Objective quality: a short overview of PEAQ architecture is presented; the implementation illustrated in [4] has been used to compute the objective quality indexes.
Subjective quality: recommendations provided by [5] and [6] have been followed; a short description of the subjective

tests method is presented together with the description of the software and criteria used to select listeners.
Our results show that the subjective perceived quality can be predicted from the objective quality indexes.

5.1. Method for Objective Evaluation

The recommendation [4] defines a method to compare a reference signal and a signal under test. The output is called ODG (objective difference grade). It is an estimate of the perceived difference and it varies between -5 and 0. The estimate is based on:

- a sophisticated ear model comprising several intermediate steps; two different ear models are provided: one based on FFT and one based on filters bank
- the calculation of psychoacoustics variables named MOVs (Model Output Variables)
- a mapping from a set of MOVs to a single value (computed by a neural network) representing the perceived difference called ODG (objective difference grade) which is believed to be representative of the audio quality (no difference = good quality, high difference = bad quality)

In the case of stereo signals all computations are performed independently for the left and right channel and then a mean values is calculated, except where otherwise indicated.
We have selected a set of reference CD-quality signals. Many of them are derived from SQAM project [8]. Each audio file is aimed to test a specific set of compression artefacts (see table below).

Artefact	***Description***
Bandwidth limitation	Loss of bandwidth when device is under test by complex sounds. Generally, they generate roughness sounds or loss of brightness
Birdies	frequency distortion over the time
Extra sound	Sounds not related to materials (artificial artefacts , extra noise smearing, etc.) caused by hybrid filterbank and non linear quantization
Pre and post echo	Smearing of attacks or sometime time asynchronism of signal

Audio reference files are listed in the following table together with a short description of their content. Of course, compression artefacts are content dependent. These files are used in listening test and they have 5-15 seconds duration.
Files have been compressed and decompressed using a public domain codec known as LAME 3.96.1. Decoded audio files are called file under test or test file.

File Name	***Description***	***Content***
12471.wav	voice chorus	complex sound
12473.wav	plain voice English male	natural speech
12502.wav	Clarinet	Tonal
12521.wav	Castanets	Transients
12524.wav	Snare drum	Transients, complex sound
12545.wav	Piano	Tonal and Transients
10001.wav	Concerto Piano	Tonal, Complex Sound
10002.wav	Funky song	Complex Sound
10003.wav	Classical - Vivaldi, Spring	Complex Sound
10004.wav	440 Hz + 3 harm	Synth sound

The codec was configured as follow:

- short block and bit reservoir activated
- CRC, and emphasis algorithms deactivated
- no strict ISO bitstream
- Q2 algorithm

The bitrates have been selected with different application areas in mind:

- 32 kbps, 48 kbps, 64 kbps e 96 kbps: streaming applications (i.e. voice over IP, web radio, etc.)
- 128 kbps e 192 kbps: P2P and on-demand applications
- 256 kbps e 320 kbps: high quality audio applications

Misalignments may be introduced by the codec: e.g. null samples at the beginning or at the end of the test file. These misalignments affect ODG computation. We therefore manipulated the test files in order to have a perfect alignment.
ODG can be computed using PEAQ only when the sampling frequency is 48 kHz but test material is sampled at 44.1 kHz. Hence we resampled the reference file and the decoded test file just before ODG computation. The resampler in the PEAQ s/w package by McGill University was used. We have verified that a noise floor is introduced by the resampling process.

5.2. Method for Subjective Evaluation

ITU-R provides many methods to perform subjective listening tests. The most relevant are described in [5], [6] and [16].
[5] describes a method for the subjective assessment of small impairments using the double blind triple-stimulus with hidden reference. It employs a 5-point impairment scale with anchor: 1 – Very Annoying, 2 – Annoying, 3 – Slightly Annoying, 4 – Perceptible but not Annoying, 5 – Imperceptible.
[6] defines a more general guide for evaluation of subjective perceived quality. It is based on [5] and generally is indicated in detection of large impairments.
Finally, [16] describes the so called MUSHRA test. It is aimed to handle intermediate audio quality using the Multi Stimulus Hidden Reference and Anchor method. It employs a 5-point impairment scale with anchor: 1 – Bad, 2 – Poor, 3 – Fair, 4 – Good, 5 – Excellent
In our tests we analyze MP3 files compressed with bitrate from 32 kbps to 320 kbps. It means we have to detect large, medium and small impairments. For this reason we decided to apply the method [5], aimed to detect small impairments.
We have selected 20 people, most of them with experience in audio and music. Only one subject at a time is involved. They can play three different stimuli (A, B, and C), at their discretion. The known reference is always available and it is indicated as stimulus A. The hidden reference and the test signal are also available but they are randomly assigned to B

and C. These audio files have 5-15 seconds durations, as previously mentioned, in order to rely only on short-term memory [5] [6].
The subject is asked to assess the impairments of B and C with respect to A according to the five grade impairment scale with anchors. At least one stimulus, B or C, should be indiscernible (grade 5) from the reference A; the other stimulus should reveal some impairment (grade <5). The experience of listening test has duration of about 30-60 minutes.
In order to help subjects, we have translated anchors in their mother language. We have also provided a written document with the rules to be followed during listening tests. Of course, each listener had a period of training, in order to get familiar with the test methodology and the use of the interface software.
The listening tests were conducted in an isolated and silent room using a personal computer equipped with an integrated audio board, a pair of headphones and the software. We have not performed listening test using loudspeakers in an open environment. Subjects were left alone and could complete the listening tests at their discretion following the rules.

6. ANALYSIS OF RESULTS

Figure 1 shows ODG plotted versus bitrate for audio file n.12524. As can be seen, ODG decreases with bitrate, as it should be: the lower the bitrate, the greater the objective difference.

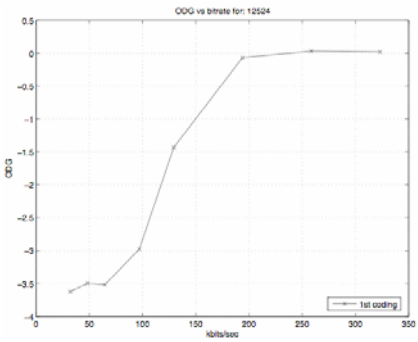

Figure 1: example of ODG vs bitrate for audio file 12524

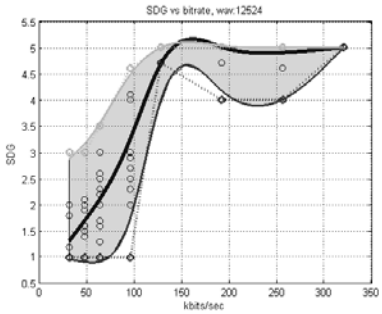

Figure 2: example of average SDG (bold line), maximum SDG and minimum SDG (thin lines) vs bitrate for audio file 12524

Figure 2 shows SDG plotted versus bitrate for the same audio file. Dot circles represent the subjective test results. Three fitted curves put in evidence the maximum, the minimum and the best-fit for any given bitrate. As can be seen, SDG decreases with bitrate, as expected: the lower the bitrate, the greater the subjective difference.
We may conclude that ODG is significant and can be related to SDG for high and low bitrates. In order to confirm the correlation, in Figure 3 we plotted both SDG and ODG vs bitrates. Thin lines represent differences between ODG and average SDG for every audio file. Bold line shows the average of these differences. Audio files 12502 and 12473 have the greatest difference between SDG and ODG. We can note the lowest correlation is located around 96 kbps.
By figures from 4 to 7 we further show that ODG is more reliable at high bitrate than at low bitrate. In these images we plotted both, ODG and SDG, versus bitrate for most meaningfully audio file. ODG value has been translated to 0-5 scale. The grey area put in evidence the range of values for SDG, from maximum down to minimum.
As can be noticed, the reliability of SDG depends on the audio file under test. In some case, the SDG has little variance around its typical values (Figure 5 where grey area is wide). In some other case, the SDG has great variance (i.e. Figure 4 where grey area is narrow)".
Analyzing results we can deduce that:

- at low bitrates ODG is not reliable; the grey area is generally wide.
- at high bitrates, ODG is reliable and can be used to predict the SDG value; the grey area is generally narrow.

We can verify these deductions looking at Figure 8. It shows 10 curves, one for each audio file under test: ODG, SDG maximum, minimum and the respective averages (bold lines). Analyzing this graphic we can deduce the high confidence of ODG among 192 Kbps and 320 Kbps. Further, we observe we have a more dispersion of minimum SDG in respect of maximum SDG. It means there a more subjectively opinion in evaluation of bad MP3 files.

7. CONCLUSION AND FUTURE WORKS

An overview of audio artefacts introduced by MP3 compression has been done. Objective (ODG by PEAQ) and subjective (SDG by listening tests following ITU Rec.) audio quality assessments have been presented. ODG and SDG values have been matched, showing that they are highly correlated at high bitrates.

Future works will concern subjective listening tests done by means of loudspeakers or high quality playout systems that may affect the subjective perceived quality.

Also, audiometric tests will be done to check how hearing capabilities affect the perceived quality.

ACKNOWLEDGMENT

Thanks to professor Goffredo Haus, director of the audio laboratory of the University of Milan, for having provided

the instrumentation required. Thanks to professor Mancuso for its assistance in recruiting people for subjective assessments.

REFERENCES

[1] ISO/IEC International Standard IS 11172-3 "Information Technology - Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5 Mbits/s - Part 3: Audio"

[2] ISO/IEC International Standard IS 13818-3 "Information Technology - Generic Coding of Moving Pictures and Associated Audio, Part 3: Audio"

[3] M. Erne, "Perceptual Audio Coders: What to Listen for", presented at the 111th Convention of AES, preprint 5489, New York, 2001

[4] ITU Recommendation, ITU-R BS.1387-1, "Method for Objective Measurements of Perceived Audio Quality", 2001

[5] ITU Recommendation, ITU-R BS.1116-1, "Methods For The Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound System"

[6] ITU Recommendation, ITU-R BS.1284-1, "General Methods For The Subjective Assessment of Sound Quality"

[7] P. Kabal, "An Examination and Interpretation of ITU-R BS.1387: Perceptual Evaluation of Audio Quality", McGill University, 2002

[8] EBU-SQAM, "Sound Quality Assessment Material, Recordings for Subjective Tests"

[9] G. Stoll, F. Kozamernik, "EBU Listening tests on Internet Audio Codecs", EBU Technical Review, 2000

[10] F. Kozamernik, "EBU Subjective Listening Tests on Low-Bitrate Audio Codecs", 2003

[11] Rahul Vanam, Charles D. Creusere, "Evaluating Low Bitrate Scalable Audio Quality Using Advanced Version of PEAQ and Energy Equalization Approach", IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2005

[12] Davis Pan, "A Tutorial on MPEG/Audio Compression", IEEE Multimedia, Volume 2, Issue 2, Pages 60-74, 1995

[13] J.Beerends and J.Stemerdink, "Modelling a cognitive aspect in the measurement of the quality of music codecs", presented at the 96th AES Convention, preprint 3800, Amsterdam, 1994

[14] J. Princen, A. Johnson, A. Bradley, "Subband / Transform Coding Technique Based on Time Domain Aliasing Cancellation," Proc. of the Int. Conf. IEEE ICASSP, pp. 2161-2164, 1987

[15] J.H.Rothweiler, "Polyphase Quadrature Filters - a New Subband Coding Technique," Proc of the Int. Conf. IEEE ASSP, 27.2, pp 1280-1283, Boston, 1983

[16] ITU Recomendation, ITU-R BS.1534, "Method for the subjective assessment of intermediate quality level of coding systems"

[17] http://www.peaq.org/

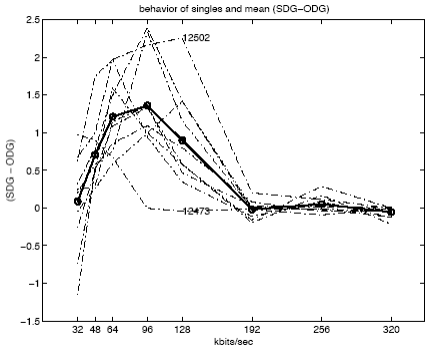

Figure 3: behavior of single and mean difference between ODG and average SDG

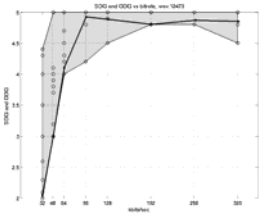

Figure 4: ODG (bold line), maximum SDG and minimum SDG (thin lines) vs bitrate for audio file 12473

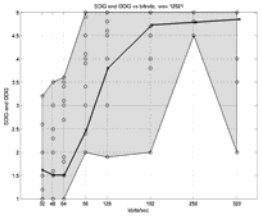

Figure 5: ODG (bold line), maximum SDG and minimum SDG (thin lines) vs bitrate for audio file 12521

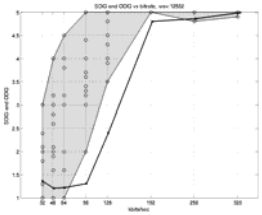

Figure 6: ODG (bold line), maximum SDG and minimum SDG (thin lines) vs bitrate for audio file 12502

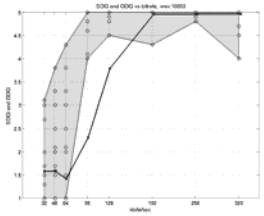

Figure 7: ODG (bold line), maximum SDG and minimum SDG (thin lines) vs bitrate for audio file 10003

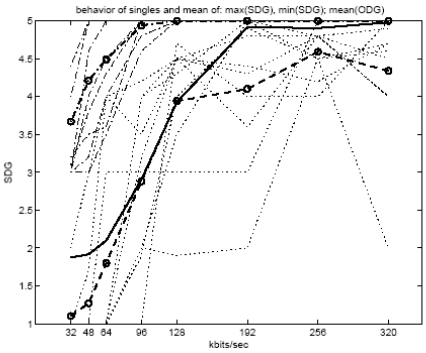

Figure 8: average of ODG (bold line), maximum SDG and minimum SDG (thin lines) vs bitrate for all audio files