# Design of Psychometric Tests for the Evaluation of Audio Coding Systems

## Diseño de pruebas psicométricas para la evaluación de sistemas de codificación de audio

Recibido: 1 de marzo de 2014 • Aceptado: 15 de diciembre de 2014

**Para citar este artículo:** : M. Herrera y B. Rodríguez, «Design of Psychometric Tests for the Evaluation of Audio Coding Systems», *Ingenium*, vol. 16, n.° 32, pp. 118-125, junio, 2015.

Marcelo Herrera Martínez*
Belman Jahir Rodríguez**

## Abstract

The present paper focuses on the description of some subjective psychoacoustic evaluation of some chosen audio coding technologies by a set of psychometric methods. Key parts for the test implementation are the suitable selection of critical signals, testing methodologies and listeners in order to perform an effective test. Based on theoretical and experimental issues from psychoacoustics, the author tries to design and implement a set of tests, which output can be interpreted as the Quality of Service of a given compression algorithm. Three psychometric methods are applied. The Constant method, the pair comparison, and DBTS (Double-Blind Triple-Stimulus with hidden reference) method, recommended by the ITU-BS 1117. Results should then, in a further step, be processed statistically.

### Keywords

Psychoacoustic assessment, Audio Coding Technologies, Psychometry, Quantization noise, Compression artifacts, Masking.

[*] Ph.D, Acústica, M.Sc. Radioelectrónica, Universidad Técnica de Praga. Profesor Titular de la Facultad de Ingeniería de la Universidad de San Buenaventura, Bogotá. Líder del Semillero de Investigación de "Sistemas de Compresión Perceptual de Audio". Grupo de Investigación: Acústica Aplicada. E-mail: mherrera@usbbog.edu.co

[**] Ingeniero de Sonido de la Universidad de San Buenaventura, Sede Bogotá. Profesor Asistente de la Facultad de Ingeniería de la Universidad de San Buenaventura, Sede Bogotá. E-mail: brodriguez@usbbog.edu.co.

## Resumen

El presente trabajo se enfoca en la evaluación psicoacústica subjetiva de algunas tecnologías de codificación de audio con un conjunto de métodos psicométricos. Las partes clave de la implementación de los pruebas son la selección adecuada de señales críticas, metodologías de evaluación y personas-escucha, para poder realizar un test efectivo. Basado en claves teoréticas y experimentales de la psicoacústica, el autor diseña una serie de pruebas, cuyo resultado puede ser interpretado como la Calidad de Servicio del algoritmo de compresión dado. Tres métodos psicométricos son aplicados. El método Constante, comparación por pares, y el DBTS (Double-Blind Triple-Stimulus with hidden reference), recomendado por el ITU-BS 1117 [1]. Los resultados son procesados estadísticamente.

### Palabras clave

Psicoacústica, Tecnologías de codificación de audio, Psicometría, Codificadores perceptuales, Señales de audio.

## 1. Introduction

Compression algorithms are nowadays widely spread and preferred, because of their ability to compress a PCM signal (from an audio CD, SACD, DVD…) to a factor of even 1:10, making bandwidth saving possible when broadcasting and storaging. Since 1987, when the Fraunhoffer Institute released the known MPEG-1 Layer III standard (MP3) [2], a huge number of algorithms and formats became available having conceptually the same principle as MP3, but introducing new techniques for the effective bit saving. Therefore the evaluation of these systems became a necessity, and the present paper is a contribution to that aim.

## 2. Subjective procedures for the evaluation of coding systems

Objective procedures when evaluating coding systems fail, because there is a small correlation between the subjective/objective parameters. Objective parameters of a compressed signal are Signal-to-Noise ratio (S/N), Mask-to-Noise ratio (M/N), Transmitted bandwidth, frequency, amplitude and phase of the signal, among others [3]. When trying to correlate these parameters with the subjective ones, as loudness, timbre, pitch and sound fullness among others, problems arise, because it does not exist a simple manner to relate these two sets. Besides this fact, objective evaluation systems are built on the same principles as the compression algorithms. They include the same psychoacoustic model, as the evaluated compression system, and the same time-to-frequency transformation. So they cannot perform a judgment in an upper level than the technologies which are to be evaluated. Subjective tests are therefore the only way how to achieve this measurement.

# 3. Basic principles of compression algorithms

Compression algorithms mainly consist of three procedures. At the first of them, a time-to-frequency transform is applied on the particular frame of the signal. Having the spectrum of the signal (obtained either by FFT, or DCT [4]), masking curves are calculated and then the quantization noise can be allocated so to be masked. The signal to mask ratio will depend on the bit-resolution of the codec, the more bits we represent the signal, the more we will mask the quantization noise, and the less the original signal will be degraded.

A typical scheme of an audio compression scheme is presented in the Fig. 1

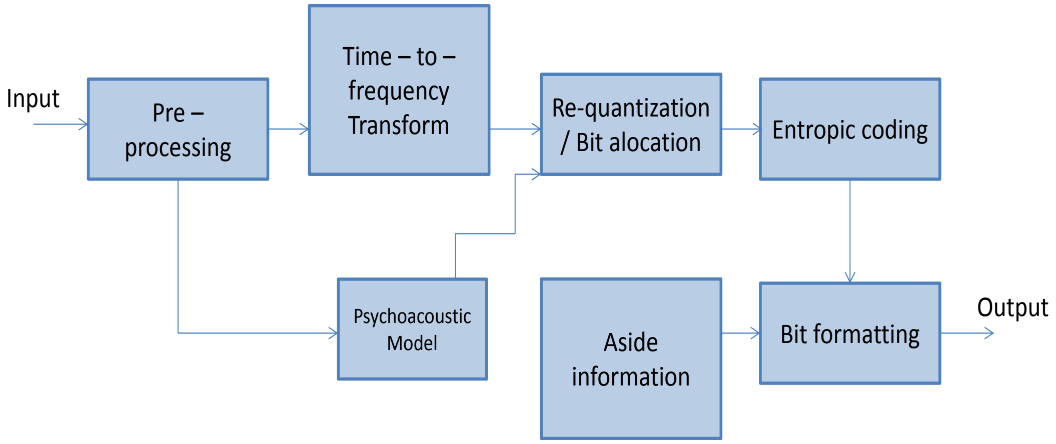

Fig. 1. Perceptual audio compression scheme [5]

## 3.1. Artifact as a quality descriptor

The evaluation of compression algorithms relies on the identification of some annoying elements on the compressed signal, referred by the audio-coding community as compression artifacts [6]. Due to (re)sampling and (re)quantization of the original PCM signal, quantization noise is introduced into the signal, and when this noise becomes audible, we call it „artifact“.

As a visual explanation of the “artifact” concept, let us observe the Figure No. 2.

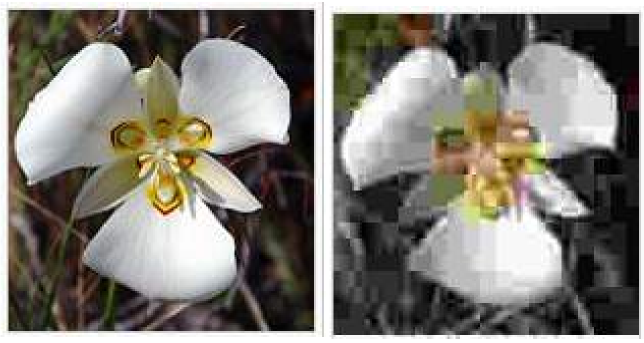

Fig. 2. Visual description of image artifacts (Comparison between original-compressed signals) [7]

As an analogy to image compression schemes, Figure 2, illustrates the idea of compression artifacts. The nature of audio and image signals is different, therefore the compression schemes for audio and image are also different; nevertheless they are based on the discarding of redundancy and irrelevance. One of the pictures is the original image, and the other is the compressed image with JPEG. In the compressed image, "blockiness" in busy regions, loss of edge clarity and tone "fuzziness" arise due to compression.

## 4. Realization of a Subjective Test for the evaluation of codecs

Designing and implementing a set of tests, which output can be interpreted as the Quality of Service (QoS) of a given compression algorithm requires the fulfillment of determined features and parameters. Generally, the BS-1114-1 Recommendation [1] states the methodology, time duration, transducers and room specifications which most serve for this purpose. Nevertheless, the author of the present paper purposes an alternative method, which lead to more reliable results, without enhancing the test time duration.

### 4.1. Critical material

Codecs (or compression algorithms) can be applied on musical or speech signals. Musical and speech signals are signals which vary from stationary or quasi-stationary waveforms until sharp dynamic attacks. Typical musical signals which have a quasi-stationary character, and therefore can be well localized in frequency are tonal sounds as when playing a particular note during a few seconds from piccolo instrument. Vocals represent the analogy of this phenomenon for speech signals. Musical signals with strong attacks, are those from sharp dynamic changes in percussive instruments, castanets, sharp drum attacks, or even other types of instruments where loudness is suddenly decreased or increased, and at the same time (because of sharp signal changes during time) frequency coefficients suddenly appear and disappear in a period of time shorter than the frame length (in the case of MPEG 1- Layer 3, frame length is about 48 ms). [8]

Therefore, these two signal extremes (frequency-localized, and time-localized signals) have to be part of every test involving compression evaluation. State-of-the art codecs, either solve efficiently the first or the second type of signals, but implementations for a good resolution of the both are just arising. These implementations are so-called hybrid implementations, where the codec adapts to the type of signal running, and either applies FFT (when solving tonal-like signals) or chooses other transforms for solving transient-like signals.

Figure 1 shows consecutive arrangement of Fourier coefficients of a musical excerpt in a time interval less than 48 ms (time duration of MPEG frame). When these changes of coefficients are abrupt as the ones showed in the figure, codec resolution should be increased in order to avoid artifacts.

The table No.1 includes the final selection of ten musical excerpts which were chosen as critical material for the test based on the considerations named before.

Tab. 1: Critical Test Material

| Musical excerpt-Artist | Description |
|---|---|
| Aleluja-Adash | Choral formed by female singers |
| Castanets | Castanets and spanish guitar, Attack sound signal |
| Cardigans | Excerpt from rock-pop music |
| Bell + Guitar | Excerpt of a short interval bell immersed in a guitar playing track |
| Glockenspiel | Excerpt where tonal and sharp attacks are combined |
| Glockenspiel 2 | Excerpt where tonal and sharp attacks are combined |
| Harpsichord | Excerpt where tonal and sharp attacks are combined |
| Violin | Excerpt of tonal character |
| Mix | Excerpt formed bell excerpt, guitar, |
| Piano | Excerpt where tonal and sharp attacks are combined |

## 4.2. Psychometric methods applied

Three psychometric methods for the subjective evaluation of codecs were applied.

### 4.2.1. Constant method

Five levels of compression are performed to a musical excerpt, ranging from 64 kbit/s until 128 kbit/s. The observer listens to the sequence and he is asked to choose a point, in which the compression effect is no longer perceived, in order words, where there is no presence of artifact. This point is denoted as the limen. Figure 2 shows the basic arrangement of this method. For stating the reliability and the validation of the method one-sided t-test is applied.

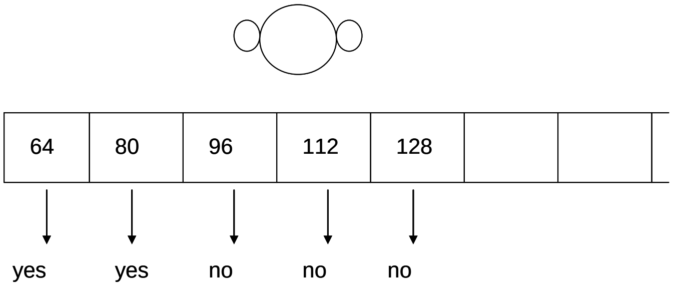

Fig. 1: Example of an evaluation during the Constant Method

### 4.2.2. Pair comparison method

Based on the results of the previous method, two levels of compression for each codec are selected: the two levels which are placed in the limen neighborhood. A direct comparison between the two codecs is performed. Fig. 3 shows the arrangement of this method. The Kendal coefficient and the coefficient of agreement are calculated for the validation.

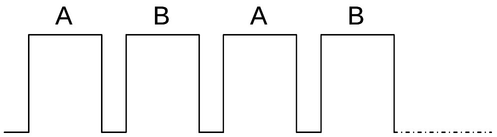

Fig. 2. Example of an evaluation during the Pair Comparison method

### 4.2.3. DBTS (Double-Blind Triple-Stimulus with hidden reference)

Based on the results of the previous methodology, one level of compression for each codec is selected. Then, the DBTS method is applied in the sense of the ITU-BS 1117-1 Recommendation [1]. Fig. 4 shows the arrangement of this method. For stating the reliability and the validation of the method, t-test and ANOVA (Analysis of Variances) is applied.

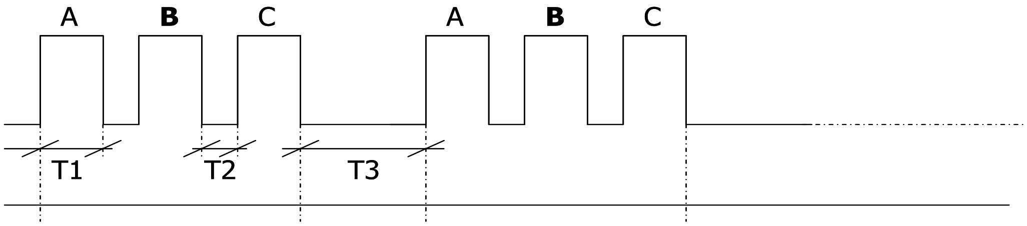

Fig. 3. Temporal organization in a sequence when evaluating by the DBTS psychometric method

### 4.3. Tested codecs, and other considerations about the test

The present test evaluates two coding technologies, MPEG-1 Layer III from the Fraunhoffer Institute, a codec supported by the program Cool Edit 2.1 Pro, and Ogg Vorbis 1.1.1, a free licensed codec, which can be downloaded from any home PC.

Time duration of the test does not exceed 25 minutes, which is a great advantage. It is the convenient time for the listener to keep the concentration for the present purpose. Time duration of each excerpt ranges from 7 seconds (in the case of the Castanets excerpt) to 15 seconds. The minimal time duration of an excerpt is associated with the minimum time for the listener to build in the auditory pathway (ear-brains) a complete auditory event. By an auditory event we understand a complete representation of the musical scene which is presented, without the necessity of presenting the whole track. When the musical excerpt is of short time duration, the listener is not able to recognize the musical reality transmitted by the artist, and a vague representation is formed on the listener's mind. Opposite to this, when the musical excerpt is of a very large time duration, the listener is not able to keep in memory the details of the first seconds of the excerpt, and therefore his evaluation will be only related to the last seconds which are stored in his mind.

## 5. Results and Discussion about the Results

The first method enables the calculation of the subjective limen, so an average value of the bit-rate/codec/excerpt combination will be the output of it. The second allows to identify an average preference of one codec above the another. The third method, the DBTS, allows to determine the codec transparency with respect to the original excerpt.

## 6. Conclusions

Compression systems are based on principles that come from psychoacoustic theory (irrelevance of some components), and from principles of information theory (redundancy algorithms). Moreover, phenomena described in psychoacoustics arise because of physiological principles.

Objective procedures fail to evaluate audio coding systems because their schemes rely on the same psychoacoustic basis as the designed perceptual audio coders. Therefore, the suitable manner to evaluate these algorithms is with the use of subjective procedures, with the use of an adecuate population, and then to take a significant and reliable sample from this population. The subjective procedure consists on a test, designed with the appropriate psychometric method, length, and signal excerpts. Afterwards, descriptive statistics are used, such as average value, standard deviation, and statistical procedures in order to test the reliability and/or validity of the test are performed (such as t-test, ANOVA).

## References

[1] RECOMMENDATION ITU-R BS.1116-1. Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. 1994-1997.

[2] MELKA, A. Zaklady experimentalní psychoakustiky. HAMU, 2005.

[3] AES. Tutorial CD-ROM, Perceptual Audio Coders, What to listen for. New York, 2002.

[4] HERRERA, M. “Evaluation of audio coding artifacts”. In Proceedings of the 10th International Student Conference on Electrical Engineering POSTER 2006. Prague (Czech Republic), 2006.

[5] RECOMMENDATION ITU-R BS.1534-1. Method for the subjective assessment of intermediate quality level of coding systems. 2001-2003.

[6] BECH, S., ZACHAROV, N. Perceptual Audio Evaluation. Ed. Wiley, 2006.

[7] SOULODRE, G. A., LAVOIE, M. C. ”Subjective Evaluation of Large and Small Impairments in Audio Codecs”. In 17th International AES Conference: High Quality Audio Coding, 1999.

[8] WALKER, R. HOEG, W., CHRISTENSEN L. Subjective assessment of audio quality – the means and methods within the EBU. EBU Technical Review Winter 1997.

[9] EBU-TECH 3339. EBU Evaluations of Multichannel Audio Codecs. Phase 3. Geneva, 2010.

[10] RECOMMENDATION ITU-R BS.1284-1. General methods for the subjective assessment of sound quality. 1997-2003.

[11] ZIELIN´SKI, S., RUMSEY F., BECH S. “On Some Biases Encountered in Modern Audio Quality Listening Tests—A Review”. In: J. Audio Eng. Soc., Vol. 56, No. 6, 2008 June.

[12] STOLL, G., KOZAMERNIK, F. “EBU Listening Tests on Internet Audio Codecs,” Tech. Rev. 283, European Broadcasting Union, Geneva, Switzerland (2000).

[13] ITU-T. P.800. “Methods for Subjective Determination of Transmission Quality”. International Telecommunications Union, Geneva, Switzerland (1996).

[14] ITU-R. Rep. BT.1082-1. “Studies toward the Unification of Picture Assessment Methodology”. International Telecommunications Union, Geneva, Switzerland, 1990.

[15] HANDS, D. “Multimodal Quality Perception: The Effects of Attending to Content on Subjective Quality Ratings”, in Proc IEEE 3rd Workshop on Multimedia Signal Processing, 1999.

[16] ZIELINSKI, S., RUMSEY, F., BECH, S., de BRUYN, B., KASSIER, R., “Computer Games and Multichannel Audio Quality – The Effect of Division of Attention between Auditory and Visual Modalities”, presented at the AES 24th Int. Conf., 2003.

[17] BIRNBAUM, M. H., “Controversies in Psychological Measurements”, in Social Attitudes and Psychophysical Measurement. B. Wegener, Ed., 1982.

[18] MEILGAARD, M., CIVILLE, G. V., CARR, B. T. Sensory Evaluation Techniques. CRC Press, New York, 1999.

[19] TOOLE, F. E. “Turning Opinion into Fact”. J. Audio Eng. Soc. Vol 30, 1982.

[20] GROSS, L., CHATEAU, S., BUSSON, S. “Effects of Context on the Subjective Assessment of Time-Varying Speech Quality: Listening/Conversation, Laboratory/Real Environment”. Acustica/Acta Acustica, vol. 90, 2004.

[21] ALDRIDGE, R., DAVIDOFF, M., GHANBARI, M., HANDS, D., PEARSON D. “Recency Effect in the Subjective Assessment of Digitally-Codec Television Pictures”, in Proc. 5th IEE Int. Conf. on Image Processing and its Applications. Edinburgh, 1995.

[22] SEFERIDIS V., GHANBARI, M., PEARSON, D. E. “Forgiveness Effect in Subjective Assessment of Packet Video”. Electron. Lett., vol. 28. 1992.

[23] HAMBERG, R., de RIDDER, H. “Continuous Assessment of Perceptual Image Quality”. J. Opt. Soc. Am., vol. 12, 1995.

[24] WATSON, A. “Assessing the Quality of Audio and Video Components in Desktop Multimedia Conferencing”, Ph. D. Thesis, University College London, 1999.

[25] LIPSHITZ, S. P., VANDERKOOY, J. “The Great Debate: Subjective Evaluation”. J. Audio Eng. Soc., vol. 29, 1981.

[26] BECH, S. “Selection and Training of Subjects for Listening Tests on Sound-Reproducing Equipment”. J. Audio Eng. Soc., vol. 40, 1992.

[27] BLAUERT, J. Communication Acoustics. Springer, 2005.

[28] BECH, S. “Training of Subjects for Auditory Experiments”. Acta Acustica, vol. 1, 1993.

[29] NEHER, T. “Towards a Spatial Ear Trainer”. Ph. D. Thesis, Institute of Sound Recording, 2004.

[30] FASTL, H. “Psycho-Acoustics and Sound Quality”. In Communication Acoustics. Ed. Springer, 2005.

[31] LETOWSKI, T. “Sound Quality Assessment: Cardinal Concepts”, presented at the 87th Convention of the AES, 1989.

[32] BLAUERT, J., JEKOSCH, U. “Sound Quality Evaluation – A Multi-Layered Problem”. Acustica/Acta Acustica, vol. 83., 1997.