Acta Polytechnica Vol. 47 No. 1/2007

Evaluation of Audio Compression Artifacts

M. Herrera Martinez

This paper deals with subjective evaluation of audio-coding systems. From this evaluation, it is found that, depending on the type of signal and the algorithm of the audio-coding system, different types of audible errors arise. These errors are called coding artifacts. Although three kinds of artifacts are perceivable in the auditory domain, the author proposes that in the coding domain there is only one common cause for the appearance of the artifact, inefficient tracking of transient-stochastic signals. For this purpose, state-of-the art audio coding systems use a wide range of signal processing techniques, including application of the wavelet transform, which is described here.

Keywords: Audio-coding, Artifacts, Wavelet transform, Psychoacoustics, Orthonormal transforms.

This text was a part of the International Conference POSTER 2006 which was held in Faculty of Electrical Engineering CTU in Prague.

1 Introduction

Information Technology has seen big advances in the audio data storage and transmission field. In 1981, the CD (Compact Disc) was developed by Philips Corporation in the Netherlands, implementing a storage solution based on optical laser and digital representation. However, data transmission bounds have demanded lower transmission rates, and therefore compression algorithms for reducing the data information stream without significantly distorting the signal. In 1987, the Fraunhoffer Institute released a compression algorithm standard, based on perceptual models of the hearing system, using masking phenomena models. The quantization noise introduced by these coding systems, specially when coding at low bit rates, gave rise to audible distortion errors, known as artifacts. Subjective evaluation led to a blurred classification of these artifacts. One type of artifact – preecho – is dealt with in this paper. Preecho cancelation is discussed, and then a wavelet transform technique for this purpose is discussed, as well as some mathematic considerations about the transform. Hybrid coders, making use of FFT or DCT for the quasi-periodic components of the signal, and DWT for the transient attacks of the signal seem to be, in author's opinion, the right direction for further research.

2 Two psychometric methods for evaluating coding systems

DBTS and SR are two pscyhometric methods that have been tested for subjective evaluation of audio codecs. Results from these tests have been published in [1][2], together with a description of the tests, excerpts and results. The DBTS method is a psychometric method which introduces the reference signal. The listener compares the coded signal with the reference, while in SR the reference is not introduced.

Here we show the ANOVA tables which show that the DBTS method is stricter than SR.

Table 1: ANOVA results for DBTS methodology

Source of variation	Degr of Freed	Sums of Squares	Mean Square	Variance Ratio (F)	Probability
Factor A	5	109.8867	21.9773	59.0312	$p < 0.05$
Factor B	6	12.0919	2.0153	5.4131	$p < 0.05$
Factor A×B	30	47.5625	1.5854	4.2584	$p < 0.05$
Error	840	312.7176	0.3723		
Total	881	482.2587			

Table 2: ANOVA results for SR methodology

Source of variation	Degr of Freed	Sums of Squares	Mean Square	Variance Ratio (F)	Probability
Factor A	5	7.1488	1.4298	5.7146	$p < 0.05$
Factor B	6	9.6904	1.6151	6.4552	$p < 0.05$
Factor A×B	30	7.4262	0.2475	1.12	$p < 0.05$
Error	966	241.6617	0.2502		
Total	1007	265.9271			

3 Artifacts from audio compression

Subjective tests performed on coded-audio signals show that individual codecs vary considerably in performance (this is validated by the ANOVA method), and also differ in performance depending on the type of signal that is used for the test. Coding signals with a strongly aperiodic character, called "attack signals" or "signals with transient behaviour" lead to an artifact known as preecho. Similarly, speech signal coding introduces to the signal an artifact known as reverberation. Sometimes, when coding at low bit rates, variations in the masking threshold from one frame to the next may lead to different bit assignments, and as a result some groups of spectral coefficients can appear or disappear[3].

Preecho is analyzed here, and some techniques for canceling it are described.

When describing artifact generation, researchers explain that a pointed artifact originates because of incorrect bit assignment from frame to frame, due to dispersion of the signal energy, which spreads out to neighbouring frames and even subbands. The relations between these dispersion lengths give rise to various perceptual artifacts. In the time domain, it is signals with a transient-stochastic character, that are affected. Percussive signals such as castanets, cymbals, clicks, claps, drums, etc. give rise to preecho when coding.

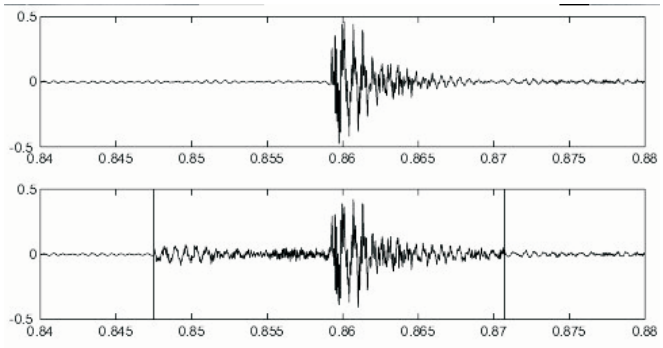

Fig. 1: A pre-echo artifact in a castanet excerpt [3]

Plosive phonemes are stochastical speech signals with a noisy character arising from turbulent air streaming in the formation of some consonants. When coding these signals, which of course occur together with vowel sounds of quasi-periodic character, reverberation is perceived.

When coding a signal which consists not only of the components explained above, but which has a frecuency representation that gives strong variations of the masking threshold from one frame to the next, the birdies artifact is perceived.

3.1 Origin of compression artifacts

The general structure of an audio-coder is given in [4]. There are three types of audio-coding systems, which differ according to the way they feed the input signal into the psychoacoustic model. The first type are transform coders, where samples from the input signal are transformed to the frequency domain. The second type are subband coders, where the transformation is performed, and then the masking thresholds are calculated for each subband. The third type are so-called parametric coders, in which a definite type of parametrization is observed.

Some authors observe that subband coders give better results when tracking transient signals, but the fixed window length that they apply does not track these signals accurately. For this purpose a wide range of techniques have been implemented, as will be described below.

4 Audio critical material selection

During this work, the author designed a program in the Matlab environment to describe the energy of the signal in each of the subbands that subband coders use.

Signals with a transient character show dispersion of their energy through the neighbouring subbands.

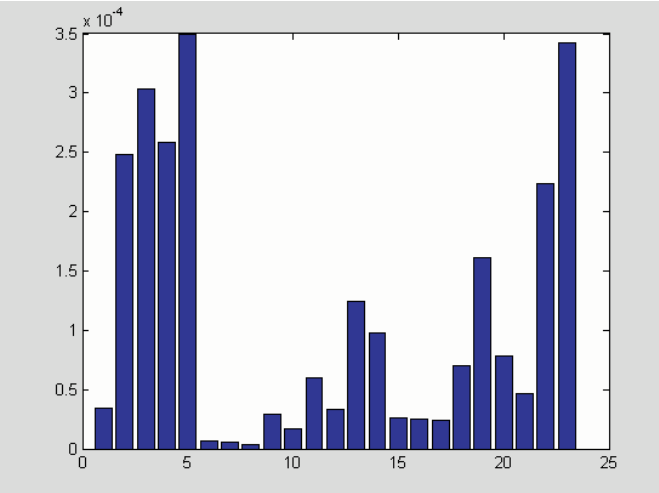

Fig. 2: Power spectrum density of a castanet audio signal

Therefore, for selecting audio material suitable for the subjective assessment of audio-codecs the program provides an estimation of which signals will behave critically and which not. Further research needs to be done to determine the relations between the subband representation of the particular signal and the artifact produced while compressed with a definite algorithm for transient tracking.

Stating the relations between the power spectrum levels inside each subband should give a cue to further research.

Figure 2 shows the energy allocation of the power spectrum density of signal castanets.

5 Current state-of-the-art for transient audio signal detection

Digital Signal Processing clearly has some potential for transient detection. This includes modifications of the Discrete Cosine Transform, DCT, with block variable lengths, tracking transient signals more accurately. The discrete wavelet transform, DWT, is also a powerful tool for transient tracking. Some implementations use a hybrid DWT/DCT. Other approaches combine non-linear transform coding and structured approximation techniques, together with hybrid modeling of the signal class under consideration. Techniques with non-uniform lapped transforms are also used. Here, a non-uniform filter bank is obtaining by joining uniform cosine modulated filter banks using a transition filter. Audio

watermarking, in which a watermark signal modifies the statistical characteristics of audio signals, in particular its stationarity, is also used [5].

5.1 Application of the wavelet transform while tracking transients

The representation of the signal in the frequency domain in earlier coders, such as MPEG-1 layer III, Ogg Vorbis and others was based on FFT, or DCT. Nowadays, applications aimed at transient tracking, use hybrid DCT, DWT among others.

Discarding the noise component, an audio signal can be represented in the following way [6],

$$x_{\text{ton}} = \sum_{\delta \in \Delta} \beta_\delta \, \omega_\delta, \quad x_{\text{tran}} = \sum_{\lambda \in \Lambda} \alpha_\lambda \, \psi_\lambda, \tag{1}$$

where $\{\psi_n, n = 0, \ldots, N-1\}$ is a wavelet basis, and
$\{\omega_m, m = 0, \ldots, N-1\}$ is an MDCT basis.

The resulting signal is

$$x = x_{\text{tran}} + x_{\text{ton}} + r \tag{2}$$

Daudet et al. [6] describe Λ and Δ as subsets of the index sets, termed *significance maps*. Residual signal r is not sparse with respect to the two bases considered here.

The main idea is that DCT, FFT and the other algorithms usually implemented in audio compression are very suitable for analysing and tracking the sinusoids or the quasi-stationary components of the signal. Transient tracking is more convenient with DWT. DWT transformation, and its ability to localize sharp attacks in time comes from the Fourier-Plancharel transformation and the uncertainty principle.

Further work is being done to apply these algorithms in improving codec performance.

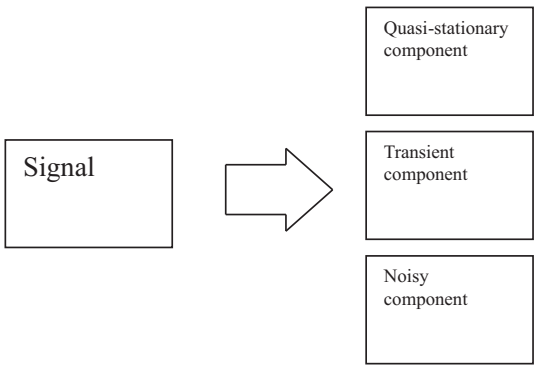

Fig. 3: Signal decomposition used in state-of-the art codecs

5.2 Demonstration of the wavelet transform when solving a transient signal

When castanets, one of the critical material excerpts, is processed by FFT or DCT with fixed window length, the spectrum disperses in such a manner that the bit-assignment derived from the psychoacoustic model is non-efficient and therefore an audible artifact known as preecho originates.

The following figure shows the original castanet signal, DWT, FFT, DCT and the other orthonormal transforms perform signal decomposition of the signal to the decomposition basis. In the case of FFT, the decomposition orthogonal basis is the set of all functions,

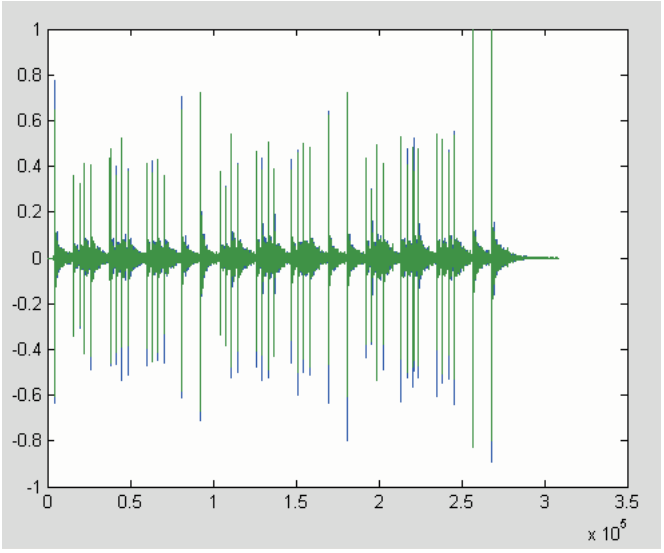

Fig. 4: Original castanet signal, critical material excerpt

$$t \to \frac{1}{\sqrt{N}} e^{j\omega t \frac{2\pi}{N}}, \quad \begin{matrix} t = \{0, 1, 2, \ldots, N-1\} \\ \omega = \{0, 1, 2, \ldots, N-1\} \end{matrix} \tag{3}$$

In the Fourier basis, frequency localization is precise, but time localization is poor.

The Euclidean orthonormal basis, which has the form

$$(1, 0, 0, \ldots, N-1), (0, 1, 0, \ldots, N-1), \ldots \tag{4}$$

unlike FFT, performs precise localization in time, but is poor in frequency. STFT represented a possible solution to the problem. It windows the signal, and therefore gives the possibility to separate the signal into frames and get the frequency representation of these frames separately. However, it still faced the problem that because of the fixed window length, transient attack signals were non-efficiently tracked.

DWT represents a compromise between these two limit representations, and performs good localization either in frequency or in time.

Signal decomposition into a particular basis can be viewed as a scalar product of the signal with the corresponding coefficient of the basis. Mathematically,

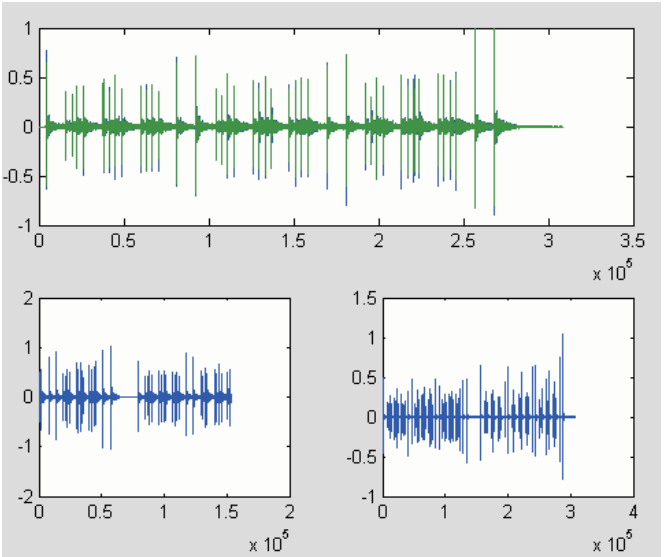

Fig. 5: 1-step decomposition of the signal using the wavelet transform

$$(f, g) = \int_{-\infty}^{\infty} f(x)\, \bar{g}(x)\, \mathrm{d}x \qquad (5)$$

representing how similar function f is to the corresponding coefficient of the orthonormal basis g.

Signal decomposition, mathematically expressed, is a set-mapping from the set of complex numbers to the set where the decomposition is described,

$$C^n \rightarrow (z(0), z(1), \ldots, z(n-1)). \qquad (6)$$

Let us perform a one-step decomposition of a castanet signal, with DWT. After one-step decomposition we achieve two signal components, depicted in Fig. 5.

Let us reconstruct the signal with the coefficients that arose after one-step decomposition. Fig. 6 gives the reconstructed signal.

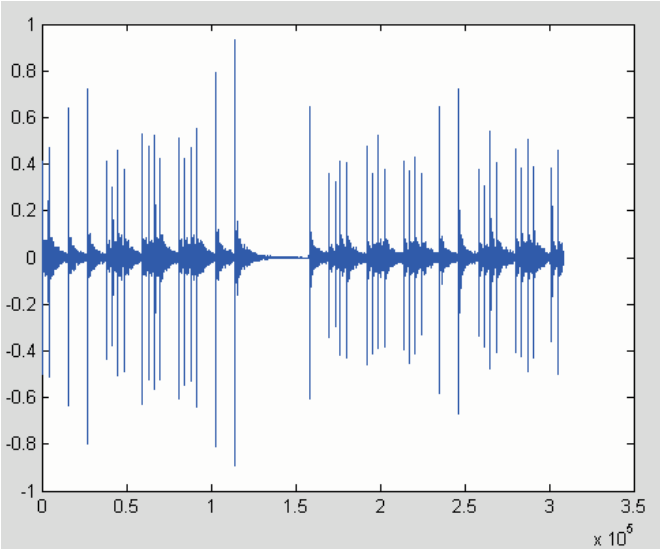

Fig. 6: Invert direct decomposition of a signal using coefficients

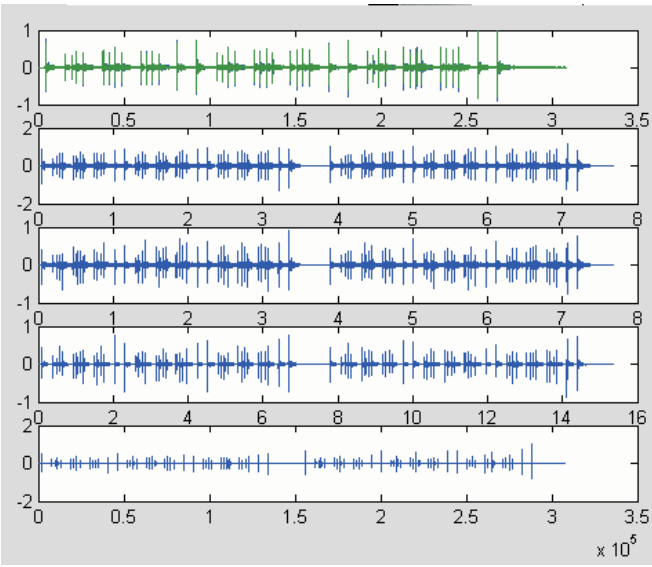

Fig. 7: Detailed coefficients at levels 1, 2 and 3 from the wavelet decomposition structure. Original signals, ca3, cd3, cd2 and cd1.

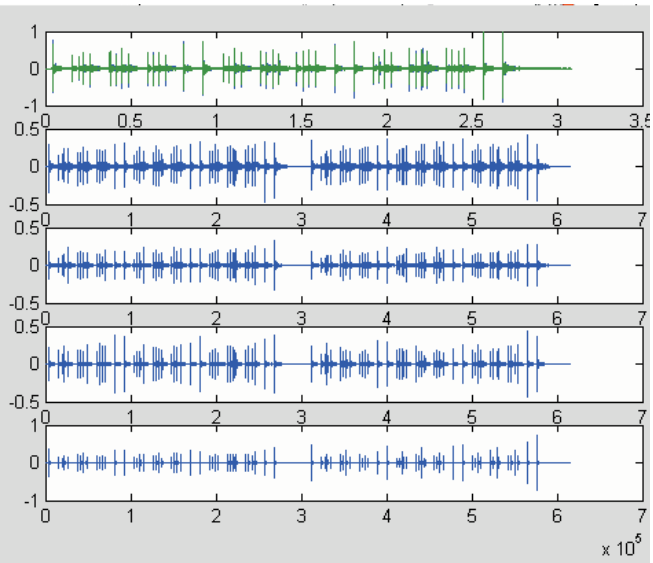

Fig.8: Reconstructed detailed coefficients at levels 1, 2 and 3, from the wavelet decomposition structure. The upper figure is the original signal, followed by the reconstructed signal, and then the coefficients.

Higher levels of signal decomposition, of course, will give more accurate representations of the audio signal, in a similar manner as higher frequency resolution improves the accuracy of the frequency representation of the signal in FFT.

DWT, then, has a hierarchical structure in which the higher the level that the decomposition affords, the longer the hierarchical DWT tree.

Comparing Figures 4 and 6, we see that the reconstruction was succesfully performed.

Now, let us perform a 3-step decomposition. A finite set of coefficients is obtained. Coefficient extraction is then performed, and this is presented in Fig. 7.

Finally we reconstruct an approximation at level 3 from the wavelet decomposition structure. We perform reconstructions of detailed coefficients at levels 1, 2 and 3 from the wavelet decomposition structure (Fig. 8).

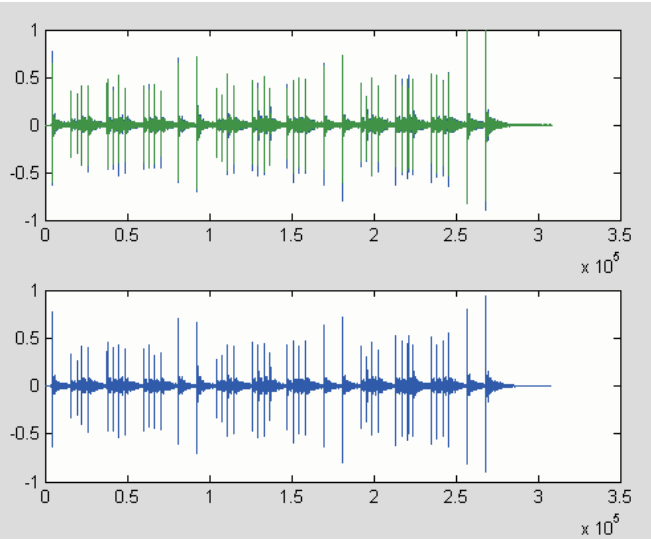

Fig. 9: Original and reconstructed signal

The last step is signal reconstruction from the wavelet decomposition structure (Fig. 9).

Transient signal reconstruction shows that DWT is a suitable method for decomposing transient signals, even performing just a 3-level decomposition. This result shows that a hybrid codec implementing FFT for extracting and processing quasi-stationary signals and DWT for extracting and processing transient signals is a more suitable algorithm for sound-coding than formerly-used codecs, which tracked signals with fixed window length DCT or FFT transforms.

6 Conclusions

Psychometric methods were used to evaluate audio-coding systems. DBTS and SR were the methods chosen to perform the evaluation. From these tests, the ANOVA validation of results shows that not only the codec performance but also the characteristics of the signal have a strong impact on the evaluation. Signals with a percussive character, such as castanets, cymbals, claps and others, when coded by algorithms which implement DCT and FFT for frequency representation of the signal, show preecho as an auditory artifact produced due to compression. The two other artifacts, while appearing to differ from preecho in the auditory domain, in the author's opinion, they have the same origin: the incorrect bit-allocation of the masking coefficients. This is because the critical signal has a power spectrum which spreads out not only to two neighbouring frames, but to the neighbouring bands.

The signal criticality can be checked by the program. Finally, some state-of-the-art techniques are discussed in order to efficiently track these critical audio signals, giving special attention to the wavelet transform.

Acknowledgments

This work has been supported by research project MSM 6840770014 "Research in the Area of Prospective Information and Communication Technologies" and by National Science Foundation grant No. 102/05/2054 "Qualitative aspects of Audiovisual Information Processing in Multimedia Systems".

References

[1] Herrera, M.: Summary of the subjective evaluation of audio-coding testing at the CVUT during the period 2003–2005. In: *XI. International Symposium of Audio and Video*, Krakov (Poland), 2005.

[2] Husnik, L., Herrera, M.: Comparison of Two Methods Used for the Subjective Evaluation of Compressed Sound Signals. In: *Forum Acousticum*. Budapest, 2005.

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

[4] Herrera, M., Dolejsi, P.: Subjective Evaluation of Audio-Coding Systems. In: *INTERNOISE 2004*. Prague, 2004.

[5] Larbi, S., Jaidane, M.: Audio Watermarking: A Way to Stationnarize Audio Signals. In: *IEEE Transactions of Signal Processing*, Vol. **53** (2005), No. 2, February 2005.

[6] Daudet, L., Molla, S., Torresani, B.: Towards a Hybrid Audio Coder. In: *Proceedings of the International Conference on Wavelet Analysis and Applications*. February 2004.

[7] http://www.mathworks.com/access/helpdesk/help/toolbox/wavelet/wavelet.htm

Marcelo Herrera Martinez
e-mail: herrerm@feld.cvut.cz

Department of Radioelectronics

Czech Technical University in Prague
Technická 2
166 27 Prague, Czech Republic

