PSYCHOACOUSTIC MODEL IN A PERCEPTUAL COMPRESSION SYSTEM BASED ON THE WAVELET TRANSFORM

Herrera Martínez, Marcelo; Guzmán Palacios, Ana María
Universidad de San Buenaventura,
Carrera 8 H n.° 172-20 | PBX: (57) 1- 667
1090 Bogotá, Colombia
mherrera@usbbog.edu.co

ABSTRACT

In this paper, a perceptual audio compressor is developed with the use of the Wavelet Transform in an Embedded System. Former compressors were not able to achieve remarkable compression ratios while maintaining the same format type (in this case .wav). The present project makes an efficient use of a Transform which enables appropriate time-frequency tracking, without perceptual losses. As it is known, the FFT tracking is not suitable for the transient representation in audio signals, and signals containing highly variable spectral components from frame to frame. The present work addresses a more suitable representation with the use of the Daubechies-Wavelet Type 4, which solves satisfactorily the problem.

INTRODUCTION

Perceptual compression is one modern area which conjugates principles from psychoacoustics and information theory. It is based on the fact that the total size of an audio file can be reduced to a determined minor size, discarding components of the signal due to two independent phenomena: irrelevance and redundance. Irrelevance is associated to the discarding of signal components due to psychoacoustic phenomena as masking, and sound loudness curves. These two phenomena are originated from the limited capabilities of the human hair cells (inner and outer cilia) inside the Human Auditory System (H.A.S.). Redundance is achieved by making use of information technology procedures that reduce the data volume, taking into account symbol event probability, assigning longer data words to less probably events and the opposite. that our research group ("Semillero de Compresión de Audio") has developed a technique, which enables to process the audio signal with a more refined type of signal transform, called the Wavelet-Transform, concretely the Daubechies-4 Wavelet Transform with 5 levels of decomposition. This Transform enables to overcome the known "Uncertainty Principle of Heissenberg" within the time-frequency duality, when processing digitally audio signals. It enables to accurately localize events in both Time and frequency with a high degree of precision. The other advantage of the implementation of the DWT (Discrete Wavelet Transform) combined with the Psychoacoustic Model is the achievement of compression to half the original volume data, maintaining the same data structure, the .wav format.

This fact enables efficient storage and broadcasting with equipment that does not require format conversion or matching, and at the same time saving space when storing and/or bandwidth when transmitting.

GENERAL WAVELET COMPRESSOR MODEL

The main core of the perceptual coder is a Wavelet filter Bank. In comparison to the PQMF filter bank, used in ISO/IEC MPEG-1 [1], the wavelet filter bank does not have equally separated sub-band divisions. More accordingly to the human auditory perception, specially the critical band theory. [2, chapter 6] and inspired by [3] a decomposition tree was designed. The bandwidths of the outputs of the sub-bands of the second wavelet filter corresponds to the critical band rate [3] as it is shown in Figure 1.

Psychoacoustic model

The psychoacoustic model analyzes the input signal from a perceptual point of view and it should calculate a global masking threshold. The masking threshold varies with time, then it is calculated for each window of the signal separately. The psychoacoustic model is a Wavelet filter bank itself, based on the ISO/IEC MPEG-1. The Psychoacoustic Model Analysis 1 is described in [4] y [5].

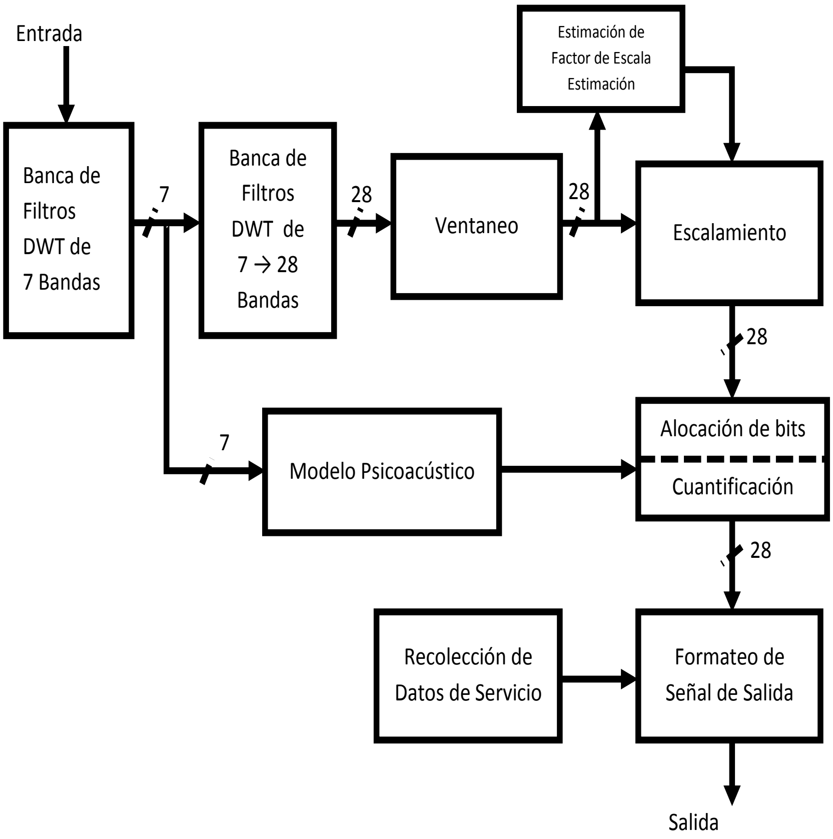


Fig. 1. Audio Compression System based on the Wavelet Transform

The block diagram of the psychoacoustic model of the Wavelet Filter bank is shown in Fig. 2. As it can be seen in it, the input of the psychoacoustic model is the output of the first stage of the Wavelet Filters of the audio perceptual compressor. In the block diagram of the psychoacoustic model in Fig. 2, this Wavelet filter bank is shown in the dashed line. An input signal of the psychoacoustic model consists of seven sub-band critically sampled signals with different simple frequencies. These seven signals are analyzed by the psychoacoustic model. The output of the model is a global threshold in dependence to the frequency or to the signal-to-masking ratio (SMR), respectively.

Windowing and FFT

The Hamming Window is used in the model, and it depends on the analyzed band. It can be a 32 or 64 sample window. For example, the 64 sample window in the first band (11-22 kHz) is a 2.6 ms long one (for a sampling frequency of 44.1 kHz), in the second band (5.5-11 kHz) is twice greater (5.3 ms) and in the last band, that is the seventh band (<343.75 Hz) is 85.3 ms long. The table No. 1 resumes the temporal lengths of the windows in the sub-bands.

The column “Equivalent Samples” compares the obtained resolution by the designed model with the designed model of “Equivalent Samples” of the Short-Time Fourier Transform (STFT) of the analyzed signal. The signal windowing is described by

$$s_{wn}(i,j) = s_n\left(i \cdot \frac{N}{2} + j\right) \cdot W(j) \qquad j = 0 \ldots N \ (1)$$

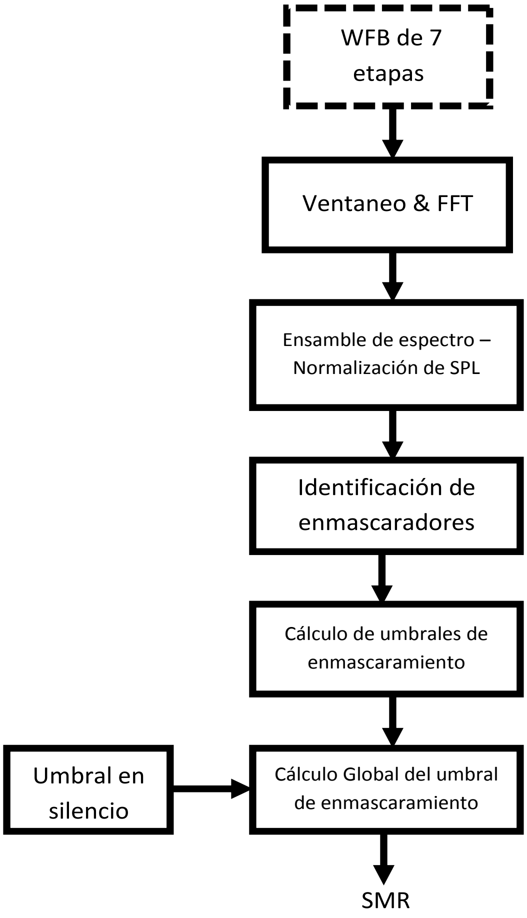

WFB de 7
etapas
Ventaneo & FFT
Ensamble de espectro – Normalización de SPL
Identificación de enmascaradores
Cálculo de umbrales de enmascaramiento
Umbral en silencio
Cálculo Global del umbral de enmascaramiento
SMR

Fig. 2. Psychoacoustic Model of the Audio Compression

TABLE I. Band division, sampling frequencies and temporal window lengths

Band No.	Frequency [Hz] Low - High	Equivalent samples [-]	Window lenght [ms]
1	11 k - 22 k	128	2.7 ms
2	5.5 k - 11 k	256	5.3 ms
3	2750 - 5500	512	10.7 ms
4	1375 - 2750	1024	21.3 ms
5	687.5 - 1375	2048	42.7 ms
6	343.75 - 687.5	4096	85.3 ms
7	< 343.75	4096	85.3 ms3

Where sn represents the signal in the n-th band, W represents the used windowing function (the Hamming Window), i is a time position and N is the window length.

The windowed signal is then transformed by the Fourier Transform as can be seen,

$$S_n(i) = FFT(s_{wn}(i)) \qquad (2)$$

The output of this block is a partial frequency representation of the windowed signals of the seven sub-bands under test.

Spectrum assembling and SPL normalization
The global spectrum representation of the signal under test is formed by partial contributions of the spectrum calculated in each frequency band.
The vector of the spectral components S is then transformed to sound pressure levels by eq. 3 (from [8]).

$$S_{SPL}(i) = 90.302 + 20 \cdot \log_{10}(|S(i)|) \quad [dB] \quad (3)$$

Identification of tonal maskers and noise maskers

The tonal masking curves and the noise masking curves have different shapes [1], therefore there is the need to separate them. According to [4], the spectral components that exceeds its neighborhood in a difference of 7 dB minimally are tonal. When finding the tonal components it is necessary to find the local maxima first and then to compare them with the components of the neighborhood. This action is described in eq. 4.

$$S_{SPL}(i) - S_{SPL}(i \pm \Delta_i) \geq 7 \qquad (4)$$

where ∆_i represents the examined neighborhood. Every spectral component that satisfy the eq. 4 are tonal. Δ_i is a model parameter and it is usually Δ_i= $\{1,2,3\}$.

According to the psychoacoustic analysis model of the audio standard ISO/IEC MPEG-1 [1], the sound pressure level of the tonal maskers are computed by the eq. 5 as a sum of the masker spectral density and its neighborhoods.

$$\mathrm{X_{TM}(i)} = 10 \cdot \log_{10} \sum_{\mathrm{j}=-1}^{1} 10^{\frac{\mathrm{S_{SPL}(i+j)}}{10}} \quad [\mathrm{dB}] \qquad (5)$$

The sound pressure level of the noise maskers is computed according to eq. 6 as a sum of the sound pressure levels of every spectral component in the correspondent critical band.

$$\mathrm{X_{NM}(b)} = 10 \cdot \log_{10} \sum_{\mathrm{i}} 10^{\frac{\mathrm{S_{SPL}(i+j)}}{10}} \quad [\mathrm{dB}] \qquad (6)$$

where b represents the critical band, i the spectral component index localized in the correspondent critical band. The noise maskers are localized in the middle of the correspondent critical band.

Calculation of the masker thresholds

When the tonal and the noise maskers are identified, the masker threshold of each masker is determined. As it is defined in the Psychoacoustic Model in the audio standard ISO/IEC MPEG 1, the masker curve of the tonal masker can be calculated by the expression 7.

$$\mathrm{M_{TM}(i+j)} = \mathrm{X_{TM}(i)} + \mathrm{MF(i,j)} - 0.275z(\mathrm{j}) - .025[\mathrm{dB}] \qquad (7)$$

where X_TM is the SPL of the masker tone, z(j) is the position of the masker curve in the bark axes, MF(i,j) is a masking function defined by the eq. 8 (from [8]) and the constant 6.025 represents the SPL distance between the masker and the top of the masking curve.

$$\mathrm{MF(i,j)} = \begin{cases} 17\Delta_z - 0.4\mathrm{X_{TM}(i)} + 11 & \Delta_z \in < -3, -1) \\ (0.4\mathrm{X_{TM}(i)} + 6)\Delta_z & \Delta_z \in < -1, 0) \\ -17\Delta_z & \Delta_z \in < 0, 1) \\ -(\Delta_z - 1) \cdot \left(17 - 0.15\mathrm{X_{TM}(i)}\right) - 17 & \\ & \Delta_z \in < 1, 3 > \end{cases} \qquad (8)$$

where Δ_z represents the bark distance from the masker. Note that outside from the interval <-3,3> it is MF= - ∞.

The masking curves for the noise maskers are defined by the Psychoacoustic Model from ISO/IEC MPEG-1 [8] defined in a similar way to the tonal maskers by eq. 9.

$$\mathrm{M_{NM}(i+j)} = \mathrm{X_{NM}} + \mathrm{MF(i,j)} - 0.175z(\mathrm{j}) - 2.025 \quad [\mathrm{dB}] \qquad (9)$$

where X_NM is the SPL of the noise masker, z(j) is the position of the masker curve in the bark axes, MF(i,j) is a masking function defined by the eq. 8 with X_TM changed to X_NM and the constant 2.025 represents the distance SPL between the masker and the top of the masking curve. Outside of the interval <-3,3>, MF = -∞, again.

Calculation of the global masking threshold

When individual masker curves from tonal maskers and noise maskers are determined, the global masking threshold can be calculated. According to the Psychoacoustic Model of the ISO/IEC MPEG-1 [1], the global masking threshold can be calculated with Eq. 10. This expression shows the addition of the masking and takes into account the threshold in silence.

$$G_{th}(i) = 10 \cdot \log_{10} \sum_j \left(10^{0.1 \cdot M_{tm}(i,j)} + 10^{0.1 \cdot M_{nm}(i+j)} + 10^{0.1 \cdot T(j)}\right) [dB] \qquad (10)$$

where T(j) represents the threshold in silence at one particular frequency .

SMR Calculation

The signal-to-mask ratio (SMR) is calculated with the Eq. 11 as a substraction between the Sound Pressure Level and the global masking threshold of a given spectral component.

$$SMR(i) = S_{SPL}(i) - G_{th}(i) \quad [dB] \qquad (11)$$

A positive SMR value indicates that the signal is above the masking threshold, while a negative SMR value indicates that the signal is below the masking threshold and it can be excluded from transmission.

Wavelet Filter Bank Module in junction with the Psychoacoustic Model

The psychoacoustic model gives us the number of bits with which we can realize the requantization of the signal. This number of bits is calculated with the signal-to-mask ratio (SMR), an analogue relation to signal-to-noise ratio (SNR).

Fig.3 Perceptual audio coder in Beagle-Board XM

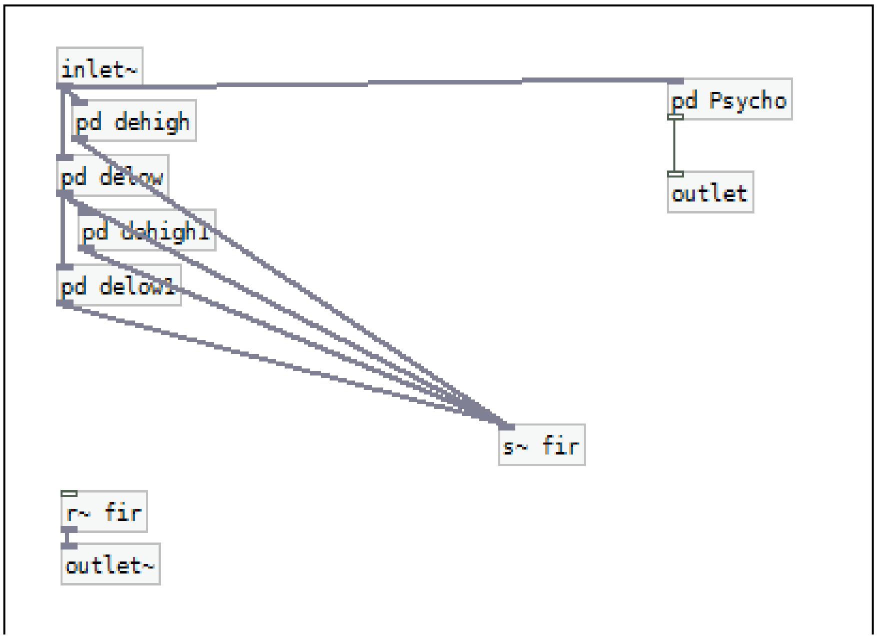
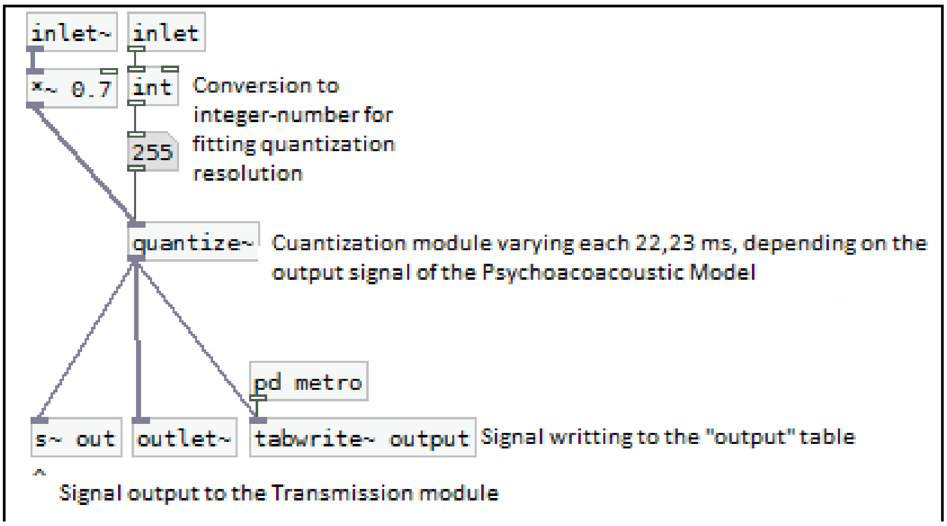


Fig. 4. Wavelet Filter Bank with the Psychoacoustic Model

Fig. 5. Psychoacoustic Model of the perceptual coder

Quantization noise module

The decomposition with the Wavelet Transform, consists, en each level, of a decomposition in low-pass and high-pass filters, as it has been mentioned lately. After these filters, decimation is performed by a factor of two (2). This is due to the fact that if the original signal contains n samples, after the filtering through the high-pass and low-pass sections, the signal at the output of these filters, will have the same number of samples that the original signal.



Conversion to integer-number for fitting quantization resolution

Cuantization module varying each 22,23 ms, depending on the output signal of the Psychoacoacoustic Model

Signal writting to the "output" table

Signal output to the Transmission module

Fig. 6. Quantization noise module

In this way, while retrieving in one single vector the output signal, it can be observed that the number of samples has been duplicated. For this reason, the decimation block (by a factor of two (2)) is implemented at the output of each one of the decomposition filters. This decomposition is known in the literature as Dyadic Decomposition or Multirresolution Analysis.

This filtering process gives as a result the details of the signal. If a higher level decomposition is performed, a more exact detail set can be obtained.

The figure shows a set of approximations and details of a signal.

Psychoacoustic Model Module

The perceptual compression system performs the effective compression of musical and voice signals based on two independent aspects: Irrelevance and Redundance.

The perceptual irrelevance is understood as a phenomenon that describes, that in despite of the entire reproduction of the audio signal, not all of the components will be perceived, and therefore it is not necessary to realize the codification of some of them.

In the audio domain, the irrelevance is explained from the psychoacoustic field, explained before, and more concretely in the phenomena that are present in the human auditory system, in the inner ear, as the frequency masking and the temporal masking.

In this way, the implemented module, analyzes the 1024-sample frames, renews the spectrum and after a sub-band analysis, it originates as a result, the number of necessary bits for signal requantization.

After this, within this spectrum, the signal frame is analyzed, through each sub-band, and the tonal and noisy thresholds are calculated, in order to locate within these masking models, the quantization noise.

ANALYSIS OF RESULTS

Comparison between the spectrum of the original signal and the compressed one

A comparison between the spectrum of the original signal and the compressed one is presented. Compression is performed with the Wavelet-Daubechies 4 Transform. The testing signal is “Castanets”, a signal with abrupt changes in time, which spectral representation is spread over various frequency bands. For this reason, the presented coder has serious problems while coding. Nevertheless, differences between the designed perceptual coder and the original signal are minimal.

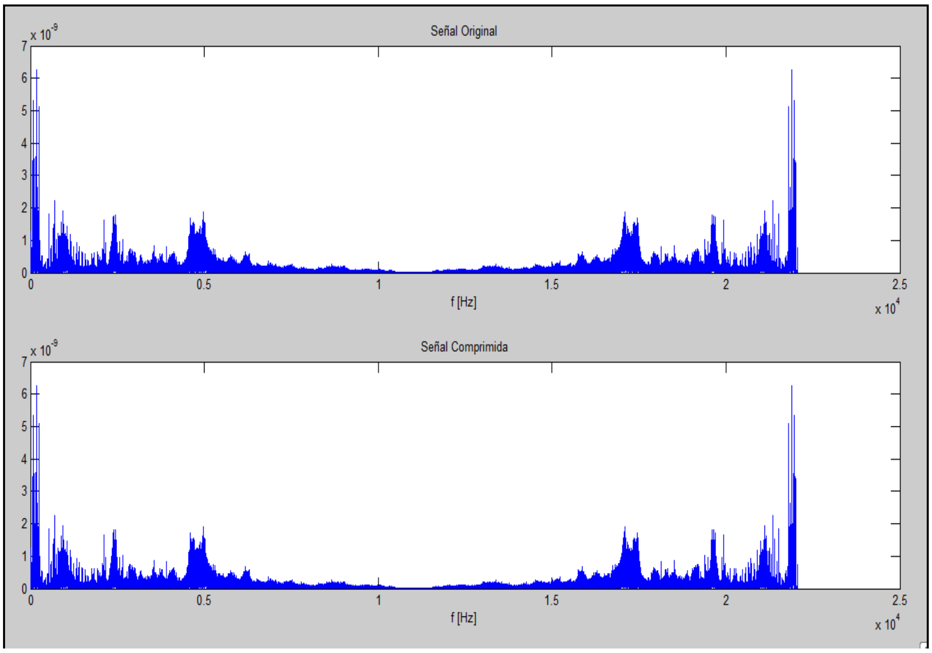


Fig. 7. Comparison between the spectrum of the original signal and the compressed one

Further projects could be related to enhancing this model with the introduction of lossless entropic coding as the Huffman coding or the Lempel-Ziv algorithm, as it is the case of .FLAC coding.

CONCLUSIONS

A perceptual audio coder is implemented based on the Wavelets-Daubechies 4 Transform. This type of Wavelet enables to simulate the psychoacoustic model without the introduction of the classical Fourier representation in the model.

REFERENCES

ISO/IEC 11172-3: Information Technology - Coding of moving pictures and associate audio for digital storage media at up to about 1,5 Mbit/s – Part 3: Audio, 1993.Psychoacoustics, Springer-Verlag, Berlin, 1990, ISBN 0-387-52600-5.

Jhang-Liang Lin, Yan-Chen Lu, Hsueh-Ming Hang, “Scalable Audio Coding Using Wavelet Packet Transform”, Proc. ISCE 1998, Taipei 1998.

M. Bosi, R. E. Goldberg, “Introduction to digital audio coding and standards”, Kluwer Academic Publishers, Boston, 2003, ISBN 1-4020-7357-7.

T. Painter, A. Spanias, “Perceptual Coding of Digital Audio”, Proceedings of the IEEE, Volume 88, Issue 4, Pages: 451 - 515, April 2000.

S. Mallat, “A Wavelet tour of signal processing”, Academic Press, San Diego, 1998, ISBN 0-12-466605-1.

R. M. Rao, A. S. Bopardikar, “Wavelet transforms: introduction to theory and applications”,Addison Wesley, Reading, Massachusetts, 1998, ISBN 0-201-63463-5 .

Jhang-Liang Lin, Yan-Chen Lu, Hsueh-Ming Hang, “Scalable Audio Coding Using Wavelet Packet Transform”, Proc. ISCE 1998, Taipei 1998.

K. Brandenburg,“OCF – A new coding algorithm for high quality sound. signals”, in: Proc. of the 1987 Int. Conf. IEEE ASSP.

L. Daudet, S. Molla, B. Torresani, “Towards a hybrid audio coder”. In Proceedings of the International Conference Wavelet Analysis and Applications. February 2004.