# Examination and Analysis of Psychoacoustic Models in Transparent Perceptual Audio Compression

Angel R. Kanchev

*Abstract* – **Different psychoacoustic models for transparent perceptual audio compression concerning fastness and calculation complexity are examined. Encoders using these models are compared.**

*Keywords* – **Audio coding, filter bank, psychoacoustic model, calculation complexity, time delay.**

## I. Introduction

Masking properties of the human auditory system allow lossy audio encoding with no hearing quality loss (i.e. transparent encoding). The mathematical models using some of those properties to determine sound's audibility are called psychoacoustic models. In this article, a determination of computation complexity and time delay for encoders with different psychoacoustic models is made (delay caused by buffering - not by calculations). The analysis method is described in point II. The encoders to be compared are presented in point III. These encoders are chosen because they are optimal by some of the comparing parameters. The results of the examination and analysis are given in point IV.

## II. Examination and Analysis Description

Input audio signal is considered to be one-channel (mono), 16-bit/sample, sampled with frequency 44.1 kHhz.

The complexity is given as a number of DSP operation per one input sample for a generalized DSP. The number of DSP operations is calculated as the number of operations multiplied by the number of clock ticks for each operation to be done. Multiplication, addition and multiplication with addition are considered to be one clock tick operations. Division is 5..10 cycles, log – 5..15. Radix 4 FFT is considered to be 8000 cycles for 256-point input and 32000 for 1024 point input. Divided by the number of samples FFT is 31.25 cycles per one input sample in both cases.

The formulas for psychoacoustic model using Signal To Mask Ratio (SMR) are given in point IIIA. This model (with modifications) is used in all examined encoders. There are analytical models that are not presented here because they have too big computational cost or their parameters are not optimal (although the hearing quality given by some of them is much better). These are: model with perceived loudness (N') [1]; with specific partial loudness ($N_S$') [3]; with Just-Noticeable Level of Difference (JNLD) [1] and with Just-Noticeable Distortion (JND) [1]. Example encoders using such models are MASCAM, MUSICAM, OCF, PXFM and ASPEC – [11]. Actually, encoders using Filter Banks (FB) only are examined here (although not all encoders are using FB).

The delay is the biggest delay caused by buffering and filter bank processing. It has no connection with the time necessary for calculations. The second one is determined by the number of DSP operations (and DSP's clock frequency).

## III. Encoders with Filter Banks

### A. Common – NMR Calculation

The purpose of a psychoacoustic model is to calculate the Noise to Mask Ratio (NMR) using some masking model of the human auditory system. The encoding, which uses psychoacoustic models, is inaudible when the maximum NMR is negative or zero dB. Non-linear loudness scales used in the calculations are phon [1] and sone [5]. Non-linear frequency scales are Bark – $z(f)$ [4] and Equivalent Rectangular Bandwidth Scale – ERBS($f$) [2,3].

**1.** Frequency distribution of the signal level $L(k)$ determination. The input samples sequence $s(n)$ with number of bits per sample b is normalized (Eq. (1)) and transformed with FFT with length $N$ (Eq. (2)). To avoid spectral leak caused by the finite sum ($N < \infty$) Hann window – Eq. (3) is used.

$$x(n) = \frac{s(n)}{N\left(2^{b-1}\right)} \tag{1}$$

$$L(k) = PN + 10\log_{10}\left|\sum_{n=0}^{N-1} w(n)x(n)e^{-j\frac{2\pi kn}{N-1}}\right|^2, \ 0 \leqslant k \leqslant \frac{N}{2} \tag{2}$$

$$w(n) = \frac{1}{2}\left[1 - \cos\left(\frac{2\pi n}{N-1}\right)\right] \tag{3}$$

$PN$ – power normalization term = 90 dB. The index $k$ determines corresponding Fourier spectral line with frequency $f(k)$. Bark and ERB scales are indexed buffers: $z(k)$ and ERBS($k$).

**2.** The threshold in quiet $\mathrm{L}_{Tq}$ is taken from buffer $\mathrm{L}_{Tq}$(k).

**3.** Each masker's threshold level $\mathrm{L}_T$ calculation:

$$L_T(k, k_c) = L(k) + SF(k, k_c) + MI(k_c), \ \text{dB} \tag{4}$$

In Eq. (4) $SF(k, k_c)$ is a "Spreading" Function simulating the fall-off (in dB) of the masking curve for sine tone masker with frequency $k_c$. $MI(k_c)$ is a Masking "Index" (in dB) – correction caused by the frequency width of the masker.

Optimization formulas in calculation of $SF(k,k_c)$:

$$SF(k,k_c)=SF(\Delta z,z_c)= \begin{cases} 17(\Delta z+1)-(0.4L(z_c)+6), & -3\leqslant\Delta z<-1 \\ (0.4L(z_c)+6)\Delta z, & -1\leqslant\Delta z<0 \\ -17\Delta z, & 0\leqslant\Delta z<1 \\ -(\Delta z-1)(17-0.15L(z_c))-17, & 1\leqslant\Delta z<8 \end{cases} \quad (5)$$

In Eq. (5):

$$\Delta z=z(k_c)-z(k);\ z_c=z(k_c) \quad (6)$$

Optimized calculation of $MI(k_c)$:

$$MI(k_c)=\alpha MI_T(k_c)+(1-\alpha)MI_N(k_c),\ \text{dB} \quad (7)$$

$$MI_T(k_c)=-6.025-0.275z(k_c)\ ,\text{dB} \quad (8)$$

$$MI_N(k_c)=-2.025-0.175z(k_c)\ ,\text{dB} \quad (9)$$

In Eqs. (7)-(9): $MI_T$ is tonal index, $MI_N$ is noise index, $\alpha\in[0;1]$ – "tonality" factor (constant for each critical band [1])

$$\alpha=\min\left(\frac{SFM}{-60},1\right);\quad SFM=10\log_{10}\left(\frac{G_m}{A_m}\right) \quad (10)$$

$$\alpha=-0.3-0.43\log_{10}(e)\in[0;1] \quad (11)$$

$SFM$ – Spectral Flatness Measure; $e$ – relative prediction error (for models with prediction); $G_m$ and $A_m$ are geometric and arithmetic means of $L(f)$ in a critical band.

**4.** Temporal masking – backward masking is seldom taken into account so forward masking is examined here: small

$$L_{TF}(t,T_m,k,k_c)= =L_T(k,k_c)\left[1.0-\frac{1}{1.35}\arctan\left(\frac{13.47t}{T_m^{0.25}}\right)\right],\ \text{dB} \quad (12)$$

$T_m$ – masker duration, $s$; $t$ – time after the end of masker, $s$. For Eq. (12) to be correct $t<t_m=0.3307T_m^{0.25}$ should be satisfied. Up until $t_m$ seconds after each masker, $L_{TF}$ is summed to the current $L_T$ and the cumulative level $L_{T\Sigma}$ is determined. In most psychoacoustic models $L_{T\Sigma}\equiv L_T$.

**5.** Excitation level $E(k)$:

$$E(k)=10\log_{10}\left\{\left[\sum_{k'=0}^{N-1}\left(10^{L_{T\Sigma}(k',k)/10}\right)^p\right]^{1/p}\right\},\text{dB} \quad (13)$$

$p\in[0.2;0.3]$ for high quality models and $p$=1 for fast calculations. For optimization purposes in MPEG [6], [7] the sum is over "detected" masker frequency indexes only.

**6.** NMR calculation:

Global threshold level is $L_{TG}$:

$$L_{TG}(z(k))=10\log_{10}\left(10^{\frac{L_{Tq}(z(k))}{10}}+10^{\frac{E(z(k))}{10}}\right),\ \text{dB} \quad (14)$$

$$SMR(k)=L(k)-L_{TG}(z(k))\ ,\ \text{dB} \quad (15)$$

$$NMR(k)=SMR(k)-SNR,\ \text{dB} \quad (16)$$

### *B. Encoder with Linear Filter Bank with IIR Filters*

The encoder with cochlear filter bank gives the best quality [8]. It consists of Low-Pass (LPF) and High Pass (HPF) filter pairs – Fig. 1. The section $S_k$ is with center frequency of its amplitude response $f_c$(k). Numerous sections for one stage (concerning the decimation) are necessary (Fig. 2).

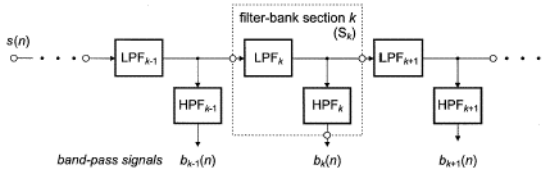


Fig. 1. Filter bank structure [8]

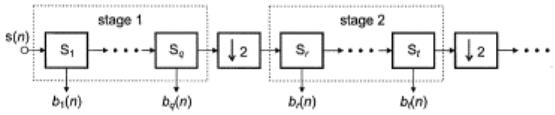


Fig. 2. Down sampling scheme [8]

Each section's center frequency can be determined by Eq. (17) for $k$=1,2,3,…:

$$f_c(k+1)=\begin{cases} 0.5^{1/N_s},\ f_c(k)\geqslant 500 \\ f_c(k)-22.4,\ f_c(k)<500 \end{cases},\ Hz \quad (17)$$

$f_c$(1) and $N_s$ depend on sampling frequency ($f_s$): for $f_s$=44.1 kHz, $f_c$(1)=20948 Hz, $N_s$=15. The desired amp120litude frequency response of one band centered at f $_c$ for $f_c\geqslant 500$ Hz is:

$$|H(f)|=\left|\frac{1}{1+\left(\frac{f}{f_c}\right)^{S_{LP}}}\cdot\frac{\left(\frac{f}{f_c}\right)^{S_{HP}}}{1+\frac{j}{4}\left(\frac{f}{f_c}\right)^{\frac{S_{HP}}{2}}-\left(\frac{f}{f_c}\right)^{S_{HP}}}\right| \quad (18)$$

$$S_{LP}=\frac{25}{20\log_{10}(1.2)};\quad S_{HP}=\frac{8}{20\log_{10}(1.2)} \quad (19)$$

$$j=\sqrt{-1} \quad (20)$$

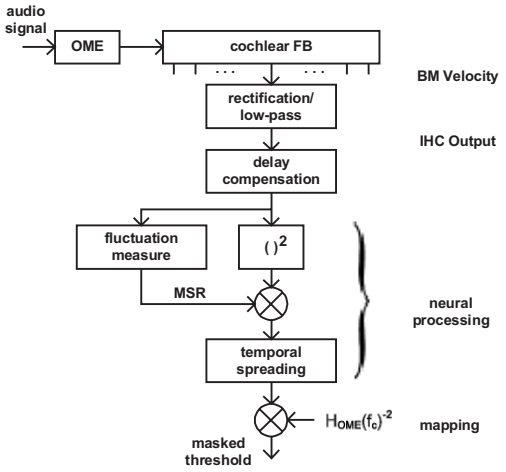


Fig. 3.

For $f_c$ <500 Hz the desired response is a replica of the filter response closest to, but not less than a center frequency of 500 Hz shifted on a linear frequency scale.

When using cochlear filter bank the psychoacoustic model is simplified (Fig. 3).

OME – Outer- and Middle Ear transfer filter with amplitude response $H_{OME}(f)$ – see Fig. 4 [8].

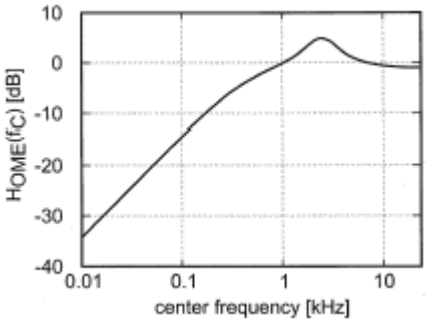


Fig. 4. OME filter amplitude response [8]

BM – Basilar Membrane; IHC – Inner-Hair Cells effect. The cut-off frequency of the second order low-pass filter is:

$$f_{LP} = \begin{cases} f_c, & f_c < 300 \\ 300\left(\frac{f_c}{300}\right)^{0.25}, & f_c \geqslant 300 \end{cases} \tag{21}$$

The delay compensation is at most 10 ms (Fig. 5) [8].

The fluctuation measure corresponds to unpredictable tonality index (1-$\alpha$) – Eq. (10). MSR – Masker to Signal Ratio. The temporal spreading is for backward and forward (Eq. (12)) masking.

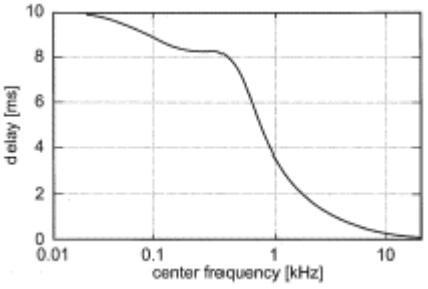


Fig. 5. Filter bank time delay [8]

### C. *Encoder with Linear Filter Bank with $\gamma$ – Tone Filters*

An easier implementation (and optimal encoder according quality/complexity ratio) is achieved with $\gamma$ –tone filter banks [9].

$$\begin{aligned} |H\gamma(f, f_c)| &= \frac{1}{\left(1 + \left(\frac{f - f_c}{kERB(f_c)}\right)^2\right)^{n/2}}; \\ k &= \frac{2^{n-1}(n-1)!}{\pi(2n-3)!!} \end{aligned} \tag{22}$$

$H\gamma(f, f_c)$ – amplitude frequency response; $f_c$ – center frequency; $n$ – filter order (usually is 4).

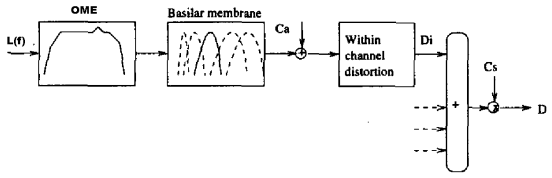


Fig. 6. General structure of the model [9]

In Fig. 6 Basilar membrane is simulated via $\gamma$ – tone filter bank, $C_a$ is the absolute threshold noise power, $C_s$ – integral ear detectability, $D_i$– "specific" and $D$ – total distortion detectability ($L_T$ analogue).

Lets present the input signal L(f) as

$$L(f) = m(f) + r(f) \tag{23}$$

$m(f)$ is the masker and $r(f)$ – the distorted signal.

$$D(m, r) = \sum_f |r(f)|^2 v^2(f) \tag{24}$$

$v(f)$ represents the masking curve in linear scale (like SF+MI in dB scale):

$$v^2(f_m) = C_s \hat{T} \sum_i \frac{|H_{OME}(f_m)|^2 |H\gamma(f_m, f_i)|^2}{\sum_f |H_{OME}(f)|^2 |H\gamma(f, f_i)|^2 |m(f)|^2 + C_a} \tag{25}$$

$f_i$ is the center frequency of the $i$-th filter.

$$\hat{T} = min\left(\frac{T}{T_{300ms}}, 1\right) \tag{26}$$

$\hat{T}$ – effective duration; $T_{300\ ms}$ represents 300 ms segment duration, $T$ – relevant segment duration.

Coefficients $C_a$ and $C_s$ satisfy:

$$\begin{aligned} & C_a = C_s \hat{T} \sum_i |H\gamma(f_{1kHz}, f_i)|^2 \\ & \frac{1}{C_s} = \hat{T} \sum_i \frac{|H_{OME}(f_{1kHz})|^2 |H\gamma(f_{1kHz}, f_i)|^2 A_{53}^2}{|H_{OME}(f_{1kHz})|^2 |H\gamma(f_{1kHz}, f_i)|^2 A_{70}^2 + C_a} \\ & A_{53} = 2.10^{-7}\ \text{W / m}^2\ (8, 93.10^{-3}\ \text{Pa}) \\ & A_{70} = 10^{-5}\ \text{W / m}^2\ (6, 325.10^{-2}\ \text{Pa}) \end{aligned} \tag{27}$$

### D. *Encoder with Wavelet Packet Decomposition Via FIR Filters*

The optimal encoder considering complexity and delay is the encoder with wavelet filter bank[10].

In [10] encoder with a wavelet packet filter bank is presented. Fig. 7 depicts encoder's decomposition tree (the decoder part is analogous).

Each lattice section is a $N$-th order FIR filter realizing Daubechies wavelet (Fig. 8).

$$A_m(z) = A_{m-1}(z) - \frac{\gamma_m}{z} B_{m-1}(z) \tag{28}$$

$$B_m(z) = \gamma_m A_{m-1}(z) + \frac{1}{z} B_{m-1}(z) \tag{29}$$

$$A_0 = \left(1 - \frac{\gamma_0}{z}\right) x \tag{30}$$

$$B_0 = \left(\gamma_0 + \frac{1}{z}\right) x \tag{31}$$

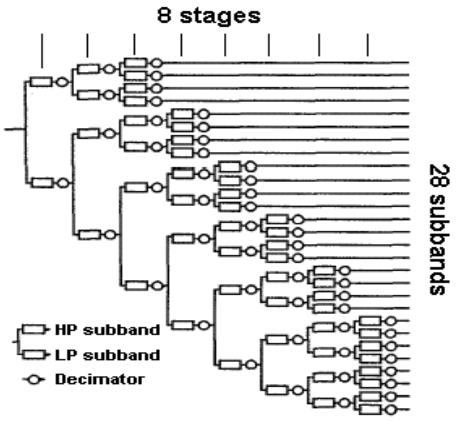


Fig. 7. Wavelet packet decomposition tree [10]

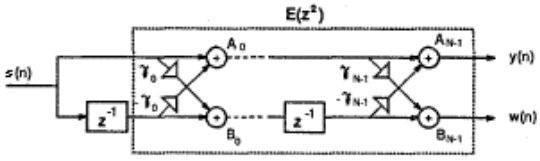


Fig. 8. Lattice element (filter) [10]

An optimization when $N$ is even can be made because $\gamma_m = \gamma_{m-1}$, $m$=1,3,..,$N$-1. The number of sections in Fig. 8 will be cut by half and decimation position will change – see [10].

The psychoacoustic model uses Eqs. (4)–(9) with $MI_N$ only. Determination of noise maskers is the same as in MPEG Psychoacoustic model I.

## IV. Results and Conclusions

As a basis for comparison, MPEG Layer I with Psycho model I encoder is used (the filter bank is 32-band, polyphase). In Table 1 the encoders presented in chapter III are compared.

Table 1. Perceptual encoders comparison

| $Q$ | Encoder | DSP op. | delay, ms | bands |
|---|---|---|---|---|
| 1 | Cochlear FB | 550 | 21.6 | 103 |
| 2 | $\gamma$-tone FB | 200-250 | $\approx 310$ | 48 |
| 3 | Wavelet packet FB | 120-140 | 93 | 28 |
| 4 | MPEG | 400-500 | 11.6 | 32 |

The first column is a Quality order value (determined by descriptions in [8–10]). It is determined by the closeness of the FB bands to the critical bands and model's precision – no subjective hearing quality test is made. The third column contains number of DSP operations per one input sample (complexity value). "Delay" is the sum of the maximum buffer size in samples divided by 44100 Hz and FB time compensations. Delay under 100ms allows interactive real-time encoding.

For conclusions see Table 2.

Table 2. Conclusions

| Encoder | Comment |
|---|---|
| Cochlear FB | Best quality; optimum quality / delay |
| $\gamma$-tone FB | Optimum quality / complexity |
| Wavelet packet FB | Optimum complexity + delay |
| MPEG | Least delay |

## References

[1] Zwicker, E.; Fastl, H.: Psychoacoustics, Facts and Models. Berlin; Heidelberg: Springer Verlag, 1990.

[2] Moore, B.C.J.; Glasberg, B.R.: Suggested Formulae For Calculating Auditory-Filter Bandwidths And Excitation Patterns. Journal of the Acoustical Society of America, Vol. 74 (3), September 1983, pp. 750-753.

[3] Moore, B.C.J.; Glasberg, B.R.; Baer, Th .: A Model For The Prediction Of Thresholds, Loudness, And Partial Loudness. Journal of the Audio Engineering Society, Vol. 45 (4), April 1997, pp. 224-240.

[4] Zwicker, E.; Terhardt, E.: Analytical Expressions For Critical Bandwidth As A Function Of Frequency. Journal of the Acoustical Society of America, Vol. 68 (5), November 1980, pp. 1523-1525.

[5] Stevens, S.S.: A Scale For The Measurement Of A Psychological Magnitude: Loudness. Psychological Review, Vol. 43, 1936 pp. 405-416.

[6] ISO/IEC 11172-3 – MPEG 1 Audio part, Psychoacoustic models 1 and 2

[7] ISO/IEC 13818-3 – MPEG 2 Audio part, Psychoacoustic models 1 and 2 extensions

[8] Baumgarte F.: Improved Audio Coding Using a Psychoacoustic Model Based on a Cochlear Filter Bank, IEEE, Vol. 10, No. 7, October 2002.

[9] Van de Par, S.; Kohlrausch A.: A New Psychoacoustical Masking Model For Audio Coding Applications, IEEE 0-7803-7402-9/02, 2002

[10] Black M.; Zeytinoglu M.: Computationally Efficient Wavelet Packet Coding Of Wide-Band Stereo Audio Signals, IEEE 0-7803-2431-5/95, 1995

[11] Painter T.; Spanias A.: A Review Of Algorithms For Perceptual Coding Of Digital Audio Signals, IEEE 0-7803-4137-6/97, 1977