Perceptual Filter Comparisons for Wideband and FM Bandwidth Audio Coders

Marcos Perreau Guimaraes[1], Nicolas Moreau[2], Madeleine Bonnet[1]

[1]Université René Descartes-Paris 5, UFR de Mathématiques et Informatique
45 rue des Saints Pères, 75270 Paris Cedex 06
email : perm,bonnet@math-info.univ-paris5.fr
[2]ENST/SIG, 46 rue Barrault, 75634 Paris Cedex 13
email : moreau@sig.enst.fr

ABSTRACT

High quality music coders commonly use auditory masked thresholds to account for the characteristics of the human ear. Perceptual filters (based upon linear signal prediction used in speech coders) are compared to filters using masked thresholds. Using listening tests, we have noticed that the second method does not provide better perceptual results. A natural way of proceeding would be to define a better psycho-acoustical model. However, an intermediate method is presented here which allows additional degrees of freedom in a standard technique. The roots of the whitening filter are treated individually.

1 Introduction

Existing speech and music coders take into account specific properties of the human ear to reduce the bit rate by eight without any loss of perceptual quality whereas when only the statistical redundancy of the signal is exploited, the bit rate is reduced by two. Spectral shaping is used in "perceptual" coders to keep the quantizing noise level below a "masking threshold" [2, 10, 7]. High bit rate (96 kbit/s) music coders for Hi-Fi bandwidths (f_s = 44.1 kHz) use a sophisticated ear model to apply additional bits to certain frequency bands where the bits are perceptually more effective. Low bit rate (8 kbit/s) speech coders for telephone-bandwidths (f_s= 8 kHz) use a perceptual filter [1], proposed by Atal and Schroeder. This filter is based on a simple but efficient auditory model.
For intermediate bit rate (32 kbit/s) wideband (f_s = 16 kHz) speech coders, Ordentlich and Shoham [11] show that the number of parameters involved in $W(z)$ (the perceptual filter expression) is not sufficient to obtain a good masking threshold approximation. These authors correct the spectral "tilt" by multiplying $W(z)$ by $T(z)$, a polynomial with very few coefficients. Chang and Wang [3] propose to directly evaluate the perceptual filter $W(z)$ from a masked threshold. Murgia and al. [9] have implemented this method in a speech and music coder for FM bandwidth (f_s = 32 kHz) and 64 kbit/s rate. They have found a noticeable improvement in the reconstructed signal quality.

The performances are compared for some perceptual filters which are obtained by different methods. The goal is the selection of a perceptual filter for use in a FM bandwidth, variable bit rate (24 to 64 kbit/s) coder of the TCX type [6]. The selected experimental conditions are relatively independent of any particular coder.

More precisely, each quantization procedure is assumed modelled by an additive white noise as shown in the schemes in figure 1.

Section 2 reviews basic psycho-acoustic principles. Four classical psycho-acoustical models are briefly described. Section 3 describes the standard perceptual filters commonly used in speech coding [1, 11]. Perceptual filters are defined based upon a masking threshold [3]. Finally, section 3.3 proposes a new evaluation method for the perceptual filter, by adding degrees of freedom to the perceptual filter in [11]. Section 4 defines a procedure for a subjective comparison of different perceptual filters.

2 Recall about psycho-acoustic

Reconstruction noise spectral shaping consists of finding an inaudible noise that minimizes the binary resources. Equivalently, for a fixed bit rate, find the noise that gives the best subjective reconstructed sig-

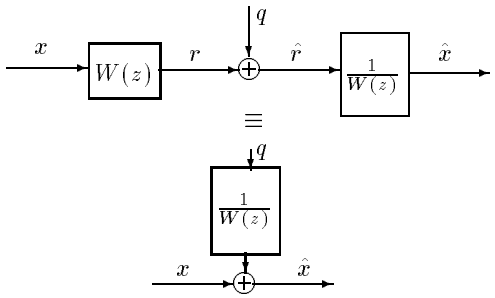

Figure 1: Fundamental scheme

nal quality. In this way the *masked threshold* is the noise power spectral density (psd) that minimizes the bit rate while achieving the transparency. This masked threshold reflects the human ear physiology explained and modelled by psycho-acoustics.

Evaluation of a masked threshold requires two stages. First, the signal $x(t)$ is analysed by a filter banks, denoted *cochlear filters*. Various works [12, 5] present experimental results and analytical expressions for those cochlear filters. Using a semi-logarithmic frequency scale (the Bark scale), the squared modulus of the filter frequency responses, $|H_j(f)|^2$ are nearly triangular with regularly distributed central frequencies. The maximum of $|H_j(f)|^2$ is unity. Furthermore, the filter shapes vary according to the signal power. The output signal power of each cochlear filter defines a time-frequency signal analysis. This power *excitation* corresponds to the transversal vibration intensity along the ear basilar membrane:

$$E(j) = \int |H_j(f)|^2 S_X(f) df$$

where $S_X(f)$ is the dsp of the signal $x(t)$. The notation adopted now corresponds to discrete time and frequency axes:

$$E(j) = \sum_i |H_j(i)|^2 S_X(i)$$

The index i could represent a Hertz or a Bark scale. Some psycho-acoustical models use the *basilar membrane spreading function*. This function is defined as the fraction of the signal intensity at the frequency i which affects the auditory perception at j frequency. This function is obtained from the cochlear filter frequency response by $f_{etal}(i,j) = |H_j(i)|^2$.

The four psycho-acoustical models, used in this paper to calculate the perceptual filters, have very simple spreading functions. These functions are constant for the model 2 of MPEG, ASPEC and Mahieux and Petit's models. The excitation is calculated with Bark scaling with one spectral line/Bark resolution in model 1 of MPEG and ASPEC models. Two spectral lines/Bark are used for model 2 of MPEG. Mahieux and Petit's model uses the Hertz scaling with 512 spectrum lines.

This time-frequency analysis is followed by a second stage. The nervous system detects the information in the propagation wave along the basilar membrane. What is of interest in audio coding is how the subjective difference between two sounds is perceived. The resolution of the sound intensity perception is limited: a masked sound x_1 simultaneous with a masking sound x_2 is inaudible if and only if the ratio of their excitations $E_1(j)$ and $E_2(j)$ is less than a curve $av(j)$, the *masking rate* [12] :

$$\frac{E_1(j)}{E_2(j)} \leq av(j) \quad \forall j \tag{1}$$

Recent works [4] have corroborated that the masking rate depends on the tonality of the signal. The masking rate is weaker when the masking sound is a tone. It is greater when the masking sound is a narrow-band noise. Thus, it is necessary to estimate the tone-like or noise-like signal nature to evaluate the masking rate. The psycho-acoustical model number 1 of MPEG looks for tones in each critical band (one Bark bandwidth) and uses two different masking rates for tone-like and noise-like parts of the signal. The ASPEC model uses a spectral flatness measure (SFM) to calculate an index which is a global estimate of the signal tone-like nature. The index is then used to weight the masking rate value for a tone and a noise, resulting in a masking rate evaluation. The tone-like signal nature is estimated for each spectral line for the psycho-acoustical model number 2 of MPEG by calculating the non-predictibility vector. Mahieux and Petit's model uses a constant masking rate at $-30dB$.

Using the masked threshold definition, a cost function must be minimized which depends upon the bit rate and the constraint of equation (1). The audio coding models do not carry out this optimization because of its complexity. To simplify the problem, the *masked auditory threshold* is calculated instead. This is the threshold beyond which a narrow-band noise becomes inaudible in the presence of the original signal, it is simply obtained by $av(j)E_2(j)$.

However, a problem remains, in general. The coding noise is not a narrow-band noise and the masked auditory threshold does not provide conditions over the dsp of the masked signal. In order to take this into account, the model 1 of MPEG removes about $15dB$ of the threshold obtained in calculating the signal-to-mask ratios. This is done by taking the minimum value of the masked threshold and the maximum value of the dsp signal in each of the 32 sub-bands of the coder. ASPEC and MPEG 2 models divide the obtained auditory threshold by a normalization func-

tion that corresponds to the gains of the spreading functions. Mahieux and Petit's model accepts the weak value of the chosen masking rate.

3 Different methods for evaluating perceptual filters

3.1 Standard methods

Within the framework of a telephone bandwidth speech predictive coder, Atal and Schroeder [1] have proposed a spectral shaping of the reconstruction noise. This involves a "perceptual filter" defined by

$$W(z) = \frac{A(z)}{A(z/\gamma)} = \frac{\prod_{k=1}^{P}(1 - z_k z^{-1})}{\prod_{k=1}^{P}(1 - \gamma z_k z^{-1})}$$

where $A(z)$ is the whitening filter for the signal to be coded. It is obtained by an order P linear prediction. In $A(z/\gamma)$, all the roots of $A(z)$ are moved away from the unitary circle by a factor $\gamma \leq 1$. The effect is to reduce $|A(z/\gamma)|^2$ in the energetical areas of the signal. $|W(z)|^2$ then becomes < 1 in the neighbourhood of the energetical locations of the signal dsp. The smaller is γ, the more emphasized are the energetical locations of $|1/W(z)|^2$. In most speech coders for telephone bandwidth, $W(z)$ takes the form $A(z/\gamma_1)/A(z/\gamma_2)$ with $\gamma_1 \approx 0.9$ and $\gamma_2 \approx 0.4$.

This method is very efficient for telephone bandwidth speech coding. Some improvements have been proposed for use in wider pass bands. To encode wideband speech at 32 kbit/s, Ordentlich and Shoham [11] have proposed enhancing the general shape of $W(z)$, the "tilt", of the perceptual filter $W(z)$ by a $T(z)$ weighting filter

$$W(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)} T(z)$$

The advantage of this approach is the possibility to independently control 1) the fine shape of $W(z)$ with the parameters γ_1 and γ_2, and 2) the general shape with the $T(z)$ coefficients. They obtained good results for speech with

$$T(z) = \frac{1}{1 + \sum_{i=1}^{2} p_i \delta^i z^{-i}}$$

The coefficients p_i are obtained from an order two linear predictor. The $\delta = 0.7$ parameter is used to adjust the "tilt".

A better compromise with speech and music is obtained by

$$T(z) = \frac{A_K(z)}{1 - \mu z^{-1}} \qquad (2)$$

where $A_K(z)$ is the whitening filter with a low order K. Choosing $K = 2$, and $|A_K^{-1}(z)|^2$ gives the general shape of the signal spectrum. The parameter $\mu = 0.3$ allows control of the "tilt" of the perceptual filter.

All these perceptual filters use the auditory results very crudely. They only exploit the fact that it is better to inject noise into the energetical signal frequencies.

3.2 Perceptual filters evaluated from a masking threshold

Chang and Wang [3] have proposed to calculate the perceptual filter via a masking threshold to take better advantage of the psycho-acoustics. The idea is to consider the inverse of the auditory threshold as the frequency response of a filter. The dsp of the reconstruction noise, $\sigma_Q^2/|W(f)|^2$, must remain below the masked threshold.

Several methods can be used to synthesize a filter from its frequency response. The method used in [3] considers the masked threshold as being the dsp over the current window. It determines the autocorrelation function with the inverse discrete Fourier transform. An LPC analysis, done with the Levinson or Schur algorithms, determines the coefficients c_i of the filter $C(z)$. The perceptual filter is then given by

$$W(z) = \frac{C(z)}{C(z/\gamma)}$$

with $\gamma = 0.8$.

The model 1 of MPEG is the psycho-acoustical model used in [3]. For a very small delay with FM bandwidth and at 64 kbits/s bit rate, another implementation has been done [9] with the psycho-acoustical model of Mahieux and Petit. In order to extend the comparison, this method has been tested with model 2 of MPEG as well as with ASPEC model.

3.3 Proposition

The listening tests described in section 4 have shown surprising results especially in FM bandwidth : the standard method, with "tilt" enhancement made by (2), exhibits results as good as those using the masked threshold.

If a method of Ordentlich and Shoham type can control the general shape, the "tilt" of the perceptual filter, it treats simultaneously the heigth of the picks and the valleys of the shape of $1/W(z)$. Instead of multiplying every root of $A(z)$ by the same coefficient γ, each pair of complex conjugated roots is multiplied

with different γ_k. The γ_k coefficients are defined as a function of the phase of the corresponding roots :

$$\gamma_k = F_{att}(arg(z_k))$$

To obtain a smoother shape in the high frequency range, the function $F_{att}(arg(z_k))$ must decrease from 1, for low frequencies, towards a slightly smaller value for high frequencies. A piece-wise linear function is chosen which equals unity until $f_c \in [0, f_s/2]$, and then decreases to $att_{min} < 1$.

4 Experimental results

The simulation of the reconstructed noise of a coder that uses a perceptual filter $W(z)$ is obtained by splitting up the signal in windows of N samples. To avoid artefacts due to cutting up in windows, an $N/2$ samples over-lapping is used. For each window m, a white noise with power $\sigma_Q^2(m)$, filtered by $1/W(z)$, simulates the reconstructed noise $r(m)$. The quantization noise $q(m)$ depends on the quantizer used but it is assumed that the noise is white with power

$$\sigma_Q^2(m) = c\ \sigma_X^2(m)\ 2^{-2b}$$

with the general assumption that the resolution b is high [8]. The signal-to-noise ratio is fixed at a 15 dB. This procedure is executed on a sound corpus. It contains speech samples from men and women speakers, classical music with different dominating instruments and pop singing music. The files of eigh seconds duration are sampled at 16 kHz and 32 kHz. The results are listened in with headphone. The degraded files are compared with the original.

The Atal method generates an important noise in high frequencies. This is in agreement with the "tilt" being not-sufficient with bands wider than the telephone bandwidth. It is better to use the psychoacoustical model 2 of MPEG with Wang and Chang method. Ordentlich and Shoham method gives results that are near-like those obtained with number 2 MPEG model. The results are slightly better with very harmonic music pieces. With the perceptual filter that we propose, the best results are obtained with the values $f_c = 2\pi/3$ and $att_{min} = 0.98$. As compared with Ordentlich and Shoham's method, the high frequency whistlings are decreased by smooothing the picks of the filter shape in high frequencies.

5 Conclusion

Perceptual filters based upon linear signal prediction are compared to filters using masked thresholds appearing in music coders. It has been noticed that those last filters do not give better perceptual results. A method is presented which allows additional degrees of freedom in a standard technique. The roots of the whitening filter are treated individually. The best results are obtained for a shape-like Ordentlich and Shoham filter in which $|W(z)|^2$ has been smoothed only in high frequencies. The perceptual results are slightly enhanced, whisthling sounds, audible with Ordentlich and Shoham filter, have vanished.

References

[1] B.S. Atal and M.R. Schroeder. Predictive coding of speech signals and subjective error criteria. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-27, June 1979.

[2] K. Brandenburg, H. Herre, J. Johnston, Y. Mahieux, and E. Schroeder. ASPEC : Adaptive perceptual entropy coding of high quality music signals. *Proceedings of the 90th AES convention*, pages 1–11, 1991.

[3] W.W. Chang and C.T. Wang. Audio coding using masking-threshold adapted perceptual filter. *Proc. IEEE Workshop on Speech Coding for Telecommunications*, pages 9–10, October 1993.

[4] J. L. Hall. Asymmetry of masking revisited: Generalization of masker and probe bandwidth. *J. Acoust. Soc. Am.*, 101:1023–1033, 1997.

[5] T. Irino and R. D. Patterson. A time domain, level dependant auditory filter: the gammachirp. *J. Acoust. Soc. Am.*, 101:412–419, 1997.

[6] R. Lefebvre, R. Salami, C. Laflamme, and J.P. Adoul. High quality coding of wideband of wideband audio signals using transform coded excitation (TCX). *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pages I–193–196, 1994.

[7] Y. Mahieux and J.P. Petit. High-quality audio transform coding at 64 kbps. *IEEE Trans. on Communications*, Vol. 42, No. 11:3010–3019, November 1994.

[8] N. Moreau. *Techniques de compression des signaux.* Masson, Collection technique et scientifique des télécommunications, 1995.

[9] C. Murgia, G. Feng, C. Quinquis, and A. Le Guyader. Very low delay and high quality coding of 20 Hz - 15 kHz speech at 64 kbit/s. *4th Europ. Conf. on Speech Comm. and Technol.*, pages 37–40, September 1995.

[10] Norme internationale ISO/CEI 11172. *Codage de l'image animée et du son associé pour les supports de stockage numérique jusqu'à environ 1,5 Mbit/s*, 1993.

[11] E. Ordentlich and Y. Shoham. Low-delay code-excited linear-predictive coding of wideband speech at 32 kbps. *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pages 9–12, 1991.

[12] E. Zwicker and E. Feldtkeller. *Psychoacoustique, l'oreille récepteur d'information.* Masson, Collection technique et scientifique des télécommunications, Traduit de l'allemand par C. Sorin, 1981.