COMPARISON OF AUDITORY MASKING MODELS FOR SPEECH CODING

*M. Lynch, E. Ambikairajah and A. Davis**
Speech Research Group, Department of Electronic Engineering,
Regional Technical College, Athlone, Ireland.
* BT Laboratories, Martlesham Heath, Ipswich IP5 7RE, U.K.
Tel.: +353 902 24542, FAX: +353 902 24493, E-mail: eambi@server1.rtc-athlone.ie

ABSTRACT

In this paper various auditory masking models recently developed for audio coding are compared and evaluated for telephone bandwidth speech coding applications. Four such models are outlined and their performance evaluated using a Wavelet Packet Transform based subband coder. The models are compared on the basis of the resulting perceptual speech quality and bit rate requirements. Results show that masking models 3 and 4 outlined in this paper provide near transparent quality at the lowest bit rates.

1. INTRODUCTION

Despite increasing activity and advances in the provision of high bit rate channels and networks, low bit rate speech coding remains important for cost effective transmission and storage of speech over limited bandwidth channels as in mobile communications and for the integration of voice with other services. Noise masking models [1], [2], [3], [4] which attempt to emulate the signal processing carried out by the human auditory system allow the noise inevitably introduced by any compression scheme to be shaped so that it remains below the level of just noticeable distortion.

Each model investigated in this study follows the fundamental principle that auditory perception is influenced by the critical band analysis performed in the human auditory system, though the models described differ in their approach to inter-band and intra-band masking effects, where inter-band describes the masking of signals in one critical band by those in another, and intra-band describes the masking effects of steady state tones and narrowband noise within each critical band. The output of each masking model, a signal-to-mask ratio (SMR) for each of 18 critical bands covering the range 0-4kHz is used in the Wavelet Packet Transform (WPT) [7], [8] based subband coder to adaptively allocate bits to each subband so that the quantisation noise within each subband is masked by the speech signal.

2. SUBBAND CODER

The input 8kHz PCM speech signal is analysed in 32ms (256 sample) frames and decomposed into 32 subbands of 125Hz bandwidth according to the tree structure shown in Fig. 1

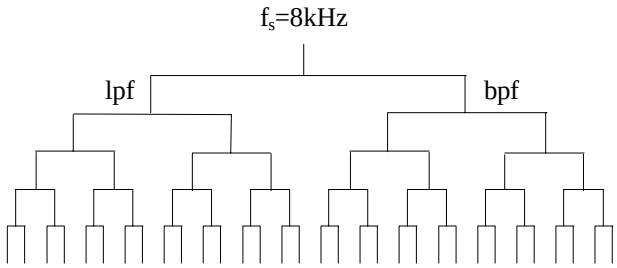

Fig. 1: WPT decomposition structure

Daubechies 16-tap orthogonal wavelet filters [9] are used as the quadrature mirror filter pairs for decomposition and reconstruction of the approximation (low-pass) and detail (band-pass) coefficients at each level. Bits are adaptively allocated to each subband based on the SMR information for each critical band received from the psychoacoustic model, where as close a match as possible is achieved between subband and critical band bandwidths.

3. AUDITORY MASKING MODELS

3.1 Masking Model 1

The psychoacoustic model proposed by Terhardt (1979) approximates the masking pattern produced by a pure tone as triangular in shape on a critical band rate scale. The upper slope of this masking pattern, or basilar membrane spreading function depends on the SPL of the masker and is given by:

$$S_{vh} = -\left[24 + 0.23\left(f_V / kHz\right)^{-1} - 0.2\,L_V / dB\right] \quad dB / bark$$

where f_v and L_v are the frequency and SPL of the

masking component, while the lower slope is taken to be independent of SPL as: $S_{vl} = 27 \quad dB/bark$.

The combined masking effect of several spectral components is assumed to be additive and the resulting threshold at component μ due to N-1 spectral components is:

$$T(z_\mu) = 20\log_{10} \sum_{\mu \neq \nu}^{N} 10^{\left[L_\nu - S_\nu(z_\nu - z_\mu)\right]/20} \quad dB$$

The estimated offset, derived from psychoacoustic experiments, for a tone masking noise is -[14.5+i]dB, where i is the critical band index and for noise masking a tone is -5.5dB [3], [6]. This model assumes a simplistic approach to the calculation of intra-band masking effects in that signal components in lower critical bands are considered inherently more tone-like in nature and those in higher critical bands as inherently more noise like. Within each critical band $T(z_\mu)$ is offset to give the global masking threshold as follows:

$$T'(z_\mu) = \begin{cases} T(z_\mu) - k(14.5+i) & dB \quad 0 \le i \le 14 \\ T(z_\mu) - k(42.5-i) & dB \quad 15 \le i \le 17 \end{cases}$$

The factor k (e.g. k=0.8) and the conservative estimate of the offset at higher frequencies were introduced to compensate for the lack of an accurate estimate of tonality. The SMR in each critical band is evaluated as the ratio of the maximum signal component to the global threshold in the critical band.

3.2 Masking Model 2

The psychoacoustic model outlined by Veldhuis et al. (1989) assumes that the shape of the masking pattern produced by a pure tone is independent of both the frequency and SPL of the tone, and is approximated by:

$$T(f_m, f) = \begin{cases} T_{\max}(f_m)\left(\frac{f}{f_m}\right)^{28} & f \le f_m \\ T_{\max}(f_m)\left(\frac{f}{f_m}\right)^{-10} & f > f_m \end{cases}$$

where f_m and f represent the frequencies of the masking and masked components respectively. This expression represents both the inter-band and intra-band masking estimates where $T_{max}(f_m)$ is defined as the relative masking threshold at the masking frequency and depends on the frequency and tonality of the signal. This relative threshold is calculated based on results from [3], and as with model 1 is applied on the assumption that signals in lower critical bands are more tone-like than those in higher bands. The masked power at frequency f due to a component at frequency f_m is obtained by the multiplication of $T(f_m,f)$ and the power of the component. The total masked power at f is estimated as the sum of the contributions from all components and the minimum level within each critical band is chosen as the masked power for the band.

3.3 Masking Model 3

The third psychoacoustic model investigated is that proposed by Johnston (1988). A similar approach to model 1 is taken for the estimation of masking effects across critical bands. However the spreading function used has constant lower and upper slopes of +25dB and -10dB respectively per critical band, and has been expressed in [5] as:

$$B(x) = 15.81 + 7.5(x + 0.474) - 17.5\left(1 + (x + 0.474)^2\right)^{1/2} \quad dB$$

where x represents the relative separation in critical bands, and is calculated for $|j - i| \le 18$ where j and i are the masking and masked bark frequencies. A Toeplitz matrix S_{ij} is formed and the summed energy in each critical band, $B(\omega)$, is convolved with the spreading function as a matrix multiplication to yield the spread critical band spectrum C_i.

Model 3 offers a more accurate determination of the noise masking threshold than any of the models so far outlined. Indeed the noise masking effect calculations of the previous models are largely approximations to the method outlined here. For tone masking noise effects the threshold is estimated as (14.5+i)dB below C_i, where i is the bark frequency and for noise masking tone effects it is estimated as 5.5dB below C_i uniformly across the critical band spectrum. To determine the noise-like or tone-like nature of the signal the spectral flatness measure (SFM) is used:

$$SFM = 10\log_{10}\frac{GM}{AM} \quad dB$$

GM and AM are the geometric and arithmetic means of the power spectrum. A coefficient of tonality α is calculated as:

$$\alpha = \min\left(\frac{SFM_{dB}}{SFM_{dB_{\max}}}, 1\right)$$

A maximum SFM of -60dB describes a signal that is entirely tone-like while an SFM of 0dB indicates an entirely noise-like signal. Speech signals in the range 200 to 3200 Hz have typical SFM's between -20dB and -30dB. The spread threshold in each band is offset according to a geometric weighting of the two threshold offsets:

$$O_i = \alpha(14.5 + i) + (1 - \alpha)5.5 \quad dB$$

The deconvolution of the spread spectrum is modelled as a renormalisation by multiplying the offset spectrum by the inverse of the energy gain. This accounts for increased energy effects due to the spreading function.

3.4 Masking Model 4

The MPEG/audio standard [4] provides information on two psychoacoustic model implementations which are defined for audio sampling rates of 32, 44.1 and 48kHz. Model 4 is based on the MPEG psychoacoustic model 1 which has been adapted for 8kHz speech. The masking function used with this model, v, varies with the masker SPL, X(z(j)), and the relative spread between the maskee and the masker, dz, as follows:

$$v = 17(dz+1) - \left(0.4X\left[z(j)\right]+6\right) \quad dB \quad -3 \le dz < -1 \quad Bark$$

$$v = \left(0.4X\left[z(j)\right]+6\right)dz \quad dB \quad -1 \le dz < 0 \quad Bark$$

$$v = -17dz \quad dB \quad 0 \le dz < 1 \quad Bark$$

$$v = -(dz-1)\left(17-0.15X\left[z(j)\right]\right)-17 \quad dB \quad 1 \le dz < 8 \quad Bark$$

where $dz = z(i) - z(j)$.

The model identifies the separate tonal and non-tonal components of the speech signal and calculates a masking index, a_t or a_n for each, where:

$$a_t = -1.525 - 0.275z(j) - 4.5 \quad dB$$

$$a_n = -1.525 - 0.175z(j) - 0.5 \quad dB$$

Tonal components are identified from an analysis of the local peaks of the power spectrum and the remaining components within a critical band are summed to form a single non-tonal component for each critical band. The individual masking thresholds of every tonal and non-tonal component are computed as the sum of the SPL, the masking index and masking function. The powers of these individual thresholds are added to the absolute threshold to form a global masking threshold estimate at every frequency. The minimum global threshold and the sound pressure level, L_i in each critical band provide the SMR information to be input to the coder, where:

$$L_i = 10\log_{10}\left(\sum_k 10^{x(k)/10}\right) \quad dB$$

4. RESULTS

Figure 2 plots the computed masking threshold shapes and original signal energy of a typical speech frame for each of the models outlined. The diagrams illustrate the different estimation of the masking threshold produced by each model.

The performance of each model is compared and evaluated in terms of the resulting perceptual quality of the decoded speech signal and the bit rate requirements. Three principle tests were performed for each masking model and listeners were asked to rate using the Mean Opinion Score (MOS) scale the perceptual quality of the reconstructed speech signal with that of the original.

The WPT coefficients are quantised in blocks of 8 subband samples using a linear mid-tread quantiser. Bits are allocated per block to satisfy the theoretical noise masking requirements which are quantified by the noise-to-mask ratio (NMR), where:

$$NMR = SMR - SNR \quad dB$$

In the first test case the total bit rate was unconstrained and bits were allocated per subband block to completely satisfy the noise masking requirements, i.e. an $NMR \le 0$ in each subband. All models performed well in this test with each scoring over 4 on the MOS scale. However the typical average bit requirements varied widely with models 1, 2, 3 and 4 requiring approximately 5, 5, 2.5 and 2.8 bits/subband sample respectively.

In the second test the total bit rate for the WPT coefficients was limited to 16kbps. Models 3 and 4 were rated slightly higher than models 1 and 2 with MOS scores between 3.0 and 3.5.

The third test specified a total WPT coefficient bit rate of 8kbps. All models performed poorly at this lower rate.

5. CONCLUSION

Four noise masking models have been implemented and evaluated for telephone bandwidth speech. Results indicate that the noise masking threshold derived from any of these models may be used as part of a subband coder, which in this case is based on the WPT, to maintain the perceptual quality of the original speech. However a choice of model for a particular application would depend on bit rate and timing constraints, with masking models 3 and 4 outlined in this paper providing near-transparent quality at lower bit rates at the expense of computational complexity. Future work involves the incorporation of these models in a CELP coder .

ACKNOWLEDGEMENT

This work has been carried out with funding provided by Forbairt, the Irish Science and Technology Agency, to the Regional Technical College Athlone under the Applied Research Programme 1994/1996.

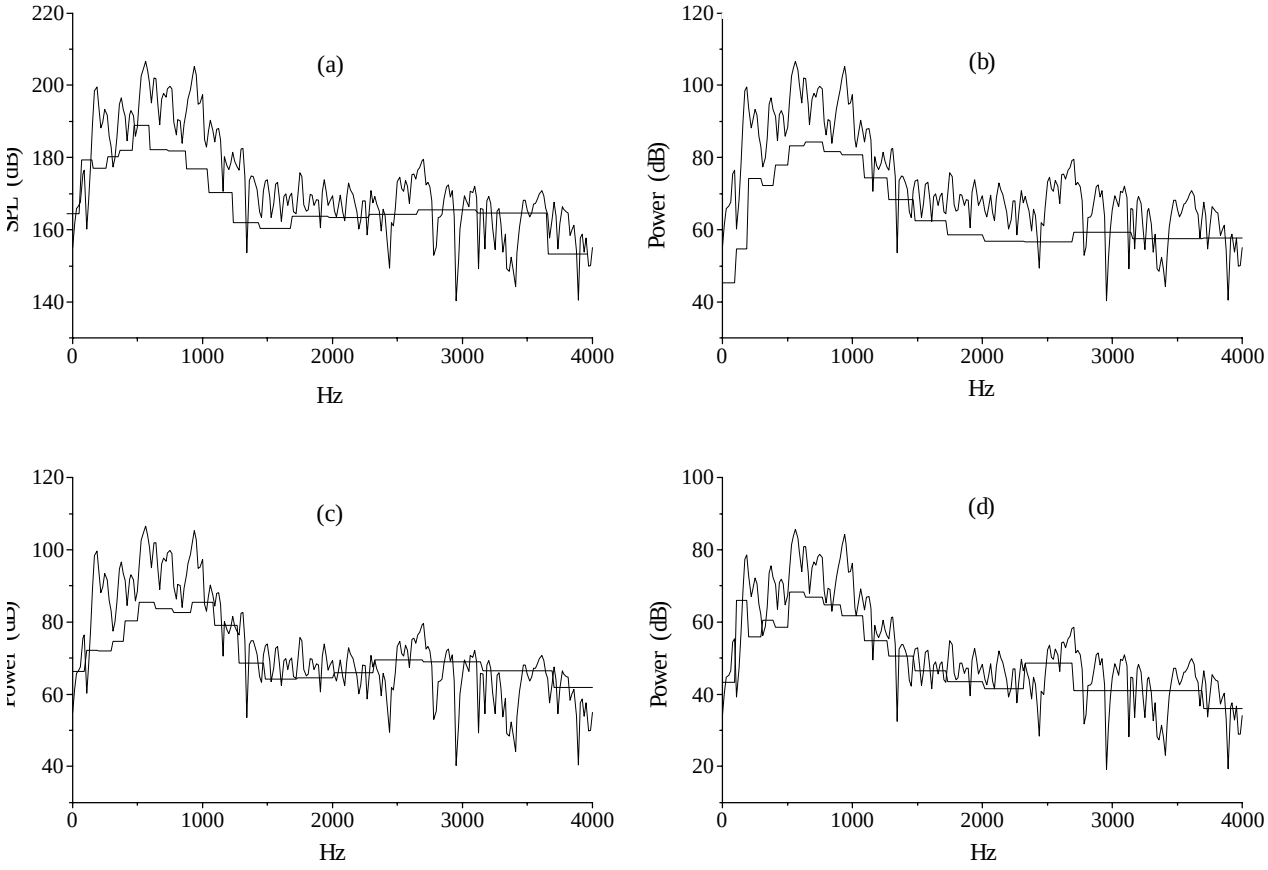

Fig. 2: Computed masking thresholds for (a) model 1, (b) model 2, (c) model 3 and (d) model 4

REFERENCES

[1] Terhardt, E. "Calculating virtual pitch", Hearing Research, pp. 155-182, 1979.

[2] Veldhuis, R.N.J., Breeuwer M. and van der Waal, R.G. "Subband coding of digital audio signals" Philips J. of Res., vol. 44, no. 2-3, pp. 329-343, 1989.

[3] Johnston, J.D. "Transform coding of audio signal using perceptual noise criteria", IEEE J. Select Areas Commun., vol. 6, no. 2, pp. 314-323, 1988.

[4] ISO/IEC IS11172-3, "Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s -Part 3: Audio", 1993.

[5] Schroeder, M.R., Atal, B.S. and Hall, J.L., "Optimizing digital speech coders by exploiting masking properties of the human ear", J. Acoust. Soc. Am., vol. 66, no. 6, pp. 1647-1651, 1979.

[6] Hellman, R.P., "Asymmetry of masking between tone and noise", Percept. Psychophys., vol. 11, no. 3, pp. 241-246, 1981.

[7] Cody, M.A. "The wavelet packet transform", Dr. Dobb's Journal, pp. 44-54, April 1994.

[8] Strang, G. and Nguyen, T. "Wavelets and filter banks", Wellesley-Cambridge Press, 1996

[9] Daubechies, I. "Orthonormal bases of compactly supported wavelets", Comm. Pure Appl. Math. vol. 41, pp. 909-996, 1988.