A NEW PSYCHOACOUSTICAL MASKING MODEL FOR AUDIO CODING APPLICATIONS

Steven van de Par, and Armin Kohlrausch

Philips Research Laboratories
Digital Signal Processing Group
Prof. Holstlaan 4, 5656 AA, Eindhoven,
The Netherlands

Ghassan Charestan, and Richard Heusdens

Delft University of Technology
Dept. of Mediamatics
Mekelweg 4, 2628 CD, Delft,
The Netherlands

ABSTRACT

The use of psychoacoustical masking models for audio coding applications has been wide spread over the past decades. In such applications, it is typically assumed that the original input signal serves as a masker for the distortions that are introduced by the lossy coding method that is used. Such masking models are based on the peripheral bandpass filtering properties of the auditory system and basically evaluate the distortion-to-masker ratio within each auditory filter. Up to now these models have been based on the assumption that the masking of distortions is governed by the auditory filter for which the ratio between distortion and masker is largest. This assumption, however, is not in line with some new findings within the field of psychoacoustics. A more accurate assumption would be that the human auditory system is able to integrate distortions that are present within a range of auditory filters. In this contribution a new model is presented which is in line with new psychoacoustical studies and which is suitable for application within an audio codec. Although this model can be used to derive a masking curve, the model also gives a measure for the detectability of distortions provided that distortions are not too large.

1. INTRODUCTION

Quantitative models describing auditory masking have proven to be a valuable tool in lossy audio coding algorithms. More specifically, these models have been used extensively in a wide variety of waveform coding based algorithms (cf. [1, 2]). In these audio coding algorithms it is assumed that the input signal serves as a masker for the distortions that are introduced by the lossy coding algorithm. Using a model that describes auditory masking enables these algorithms to adjust the distortion levels for each part of the spectrum in such a way that the distortion is just not detectable.

Typically, masking models include some assumptions with respect to the limited frequency selectivity of the human auditory system which can be attributed to the filtering properties of the basilar membrane that is placed within the cochlea. The concept of critical bands refers to these limitations in frequency resolution of the auditory system [3]. The frequency resolution is found to be better at low frequencies and to decrease at high frequencies. This frequency resolution can be modelled by assuming that the frequency selective behaviour of the basilar membrane can be described by a series of bandpass filters with narrow bandwidths at low frequencies and larger bandwidths at high frequencies.

Most existing masking models used in audio coding are based on a method where the input signal is analysed in the frequency domain and where each frequency component is assumed to result in a spreading function which captures the masking properties of that spectral component [1]. The total masking function is derived by power addition of the separate spreading functions. In such models the spectral components are usually classified as either being tonal or noise-like. Depending on the classification, the spreading functions are adjusted. This is necessary because it is known from psychoacoustical data that tonal maskers are much less effective maskers than noise-like maskers [4, 5].

The underlying physical explanation for using spreading functions is that they reflect the band-pass characteristic of auditory filters. It is assumed that the detectability of a distortion component of a particular frequency is determined only by the auditory filter that is spectrally centered around this distortion component. Since this auditory filter has a band-pass characteristic, the masker energy that is close to the centre frequency of this filter will contribute most to the masking within this filter while remote frequency regions will contribute less. This behaviour is captured by the addition of spreading functions.

The assumption that only the on-frequency filter (the filter centred around the distortion compenent that has to be masked) contributes to the masking is a fair assumption when the distortion is narrow-band, and in fact this is most often the case in psychoacoustical studies dealing with masking where the signal to be detected is often sinusoidal. However, when the distortion is wider in bandwidth, like what is typically the case in the context of audio coding, there is evidence that more auditory filters than only a single one contribute to the detectability of the distortion. In a study with a broadband masker and a complex of constant level tones that subjects had to detect, it appeared that with increasing bandwidth of the tonal complex the detectability improved even when the bandwidth of the tonal complex exceeded the critical bandwidth [6]. This result and various other studies suggest that the human auditory system is able to integrate information over a range of auditory filters in order to improve detectability of a distortion signal.

In this paper a new masking model will be presented (Section 2) that incorporates the across-frequency intergration that has been observed within the human auditory system. Similar to most of the masking models used in audio coding it operates on a frame-by-frame basis to predict masking curves and it is of low complexity. In Section 3, the calibration of this model will be discussed, in Section 4, an evaluation of this model will be given by comparing the model to some basic psychoacoustical data, and finally, in Section 5 we draw some conclusions.

0-7803-7402-9/02/$17.00 ©2002 IEEE

2. A NEW PSYCHOACOUSTICAL MASKING MODEL

The basic peripheral stages of the human auditory system are incorporated in the model as well as some very simple assumptions about the higher stages of processing in the auditory system. The precise properties of each of these stages is accurate only to a first order approximation, and is adapted such that predictions of the model are in line with some critical psychoacoustical data.

Since the model is based on the assumption that the essential properties of the basilar membrane can be modelled by a gamma-tone filterbank, this filterbank will first be defined. The transfer function of an n-th order gamma-tone filter as a function of f can be approximated quite well by

$$\gamma(f) = \frac{1}{\left(1 + (\frac{f-f_0}{k\mathrm{ERB}(f_0)})^2\right)^{\frac{n}{2}}}, \qquad (1)$$

where f_0 is the centre-frequency of the filter, $\mathrm{ERB}(f_0)$ is the Equivalent Rectangular Bandwidth of the filter centred at f_0 such as measured by Glasberg and Moore [7], n is the filter order which is commonly assumed to be 4, and $k = \frac{2^{(n-1)}(n-1)!}{\pi(2n-3)!!}$, a factor needed to ensure that the filter indeed has the specified ERB.

In Fig. 1, an outline of the model is given. It is assumed that an input signal, $X(f)$, is presented to the model which is the Fourier transform of a short windowed segment of a larger input signal. This input signal is first filtered by the outer and middle ear transfer function, $H_{om}(f)$. This filter incorporates the properties of, for example, the ear canal, and of the ossicles in the middle ear. For simplicity it is assumed here that this transfer function is equal to the inverse of the threshold-in-quiet function H_{tq}. Although this is only a rough approximation of the actual outer and middle ear function, it proves to be a good choice because it results in a very accurate prediction of the threshold in quiet by the model. The filter $H_{om}(f)$ is followed by the basilar membrane which is modelled by a series of gamma-tone filters, $\gamma_i(f)$, with centre frequencies, f_0, spaced linearly on an ERB scale, with i the index of the filter. The energy at the outputs of these filters is increased with a constant, C_a, which can be attributed to internal noise within the auditory periphery that limits the detectability of weak signals.

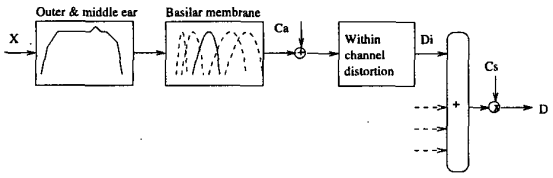

Fig. 1. *General structure of the masking model. See text for further details.*

The next stage derives the within-channel distortion-to-masker ratio, e.g. the distortion-to-masker ratio that can be observed within the i-th gamma-tone filter. It is assumed that this distortion can be derived by writing $X(f)$ as $X(f) = m(f) + s(f)$, where $m(f)$ is the masker, and $s(f)$ is the distortion that is introduced by the coder. The masker power within the i-th filter is given by

$$M_i = \frac{1}{N}\sum_f |H_{om}(f)|^2 |\gamma_i(f)|^2 |m(f)|^2, \qquad (2)$$

where N is the segment size. Equivalently, the distortion power within the i-th filter is given by

$$S_i = \frac{1}{N}\sum_f |H_{om}(f)|^2 |\gamma_i(f)|^2 |s(f)|^2. \qquad (3)$$

Note that $\frac{1}{N}|m(f)|^2$ denotes the power spectrum of the original, masking signal in Sound Pressure Level (SPL) per frequency bin, and similarly $\frac{1}{N}|s(f)|^2$ is the power spectrum of the distorting signal.

The 'specific' distortion detectability is then defined as the ratio between distortion power and masker power plus the absolute threshold noise power C_a:

$$D_i = \frac{S_i}{M_i + C_a}. \qquad (4)$$

Some important observations can be made from this equation. First of all, for $M_i \gg C_a$, the distortion detectability is equal to the distortion-to-masker-power ratio. This is in line with the so-called sensitivity index d' which is used in signal detection theory as a measure for the detectability of a signal close to its masking threshold [8]. This ratio also implies that if the powers of the distortion and the masker are increased by the same amount (in dBs) that the detectability, within one filter, does not change provided that the masker level is well above the threshold in quiet.

For $M_i \ll C_a$, the 'specific' distortion detectability is determined only by the distortion power. This occurs when the masker power is below the absolute threshold or threshold in quiet.

It is now assumed that the human auditory system is able to combine information from a range of filters to improve the detectability of distortion. This property is modelled by assuming that

$$D(m,s) = C_s \hat{L} \sum_i D_i$$

$$= C_s \hat{L} \sum_i \frac{\sum_f |H_{om}(f)|^2 |\gamma_i(f)|^2 |s(f)|^2}{\sum_f |H_{om}(f)|^2 |\gamma_i(f)|^2 |m(f)|^2 + C_a}, \qquad (5)$$

where $D(m,s)$ is the total distortion detectability as it is predicted for a human observer given an original signal m and a distortion signal s. It is assumed that C_s is chosen such that for $D = 1$, the distortion is at the threshold of detectability. To incoorporate the dependence of $D(m,s)$ on the segment duration we define an effective duration $\hat{L}$:

$$\hat{L} = \min(\frac{L}{L_{300ms}}, 1), \qquad (6)$$

where L is the duration of the segment, and L_{300ms} resembles a 300-ms segment in line with the temporal integration properties of the human auditory system [9]. Alternatively, with the definition $\hat{L} = \frac{L}{L_{tot}}$, with L_{tot} being the total duration of the excerpt to be

encoded, we can assume unlimited temporal integration of distortions. With this latter assumption the distortions can be added over all segments to result in a total distortion over the complete excerpt to be encoded, where $D = 1$ would be equivalent to the threshold of detectability. Whereas this last assumption is not supported by perceptual masking data, it can still lead to very satisfying results when applied within the context of an audio codec [10].

Equation 5 is the most general expression of the model. From this expression we can derive the masked threshold of any arbitrary distortion signal given the original, masking signal m. Let us assume that the arbitrary distortion signal is given by $A\epsilon(f)$, where A is the masked threshold of the distortion signal and where $\sum_f \epsilon(f^2) = 1$, which corresponds to a sound pressure level of 0 dB. Using Eq. (5) we can derive the masked threshold A:

$$\frac{1}{A^2} = C_s \hat{L} \sum_i \frac{\sum_f |H_{om}(f)|^2 |\gamma_i(f)|^2 |\epsilon(f)|^2}{\sum_f |H_{om}(f)|^2 |\gamma_i(f)|^2 |m(f)|^2 + C_a}. \quad (7)$$

In many applications for audio coding, a masked threshold *function* is derived. It can be defined as the masked threshold of a sinusoidal distortion as a function of the frequency of this sinusoid. Parallel to the way Eq. 7 was derived, we can obtain a masked threshold function. We assume that in the presence of a given masker m, we have to detect a sinusoidal distortion with an unknown amplitude v and a frequency f_m; thus $s(f) = v(f_m)\delta(f_m - f)$. Substituting this function in Eq. 5 together with the assumption that $D = 1$ we can derive the amplitude $v(f_m)$ such that the sinusoid is just masked. Thus, $v(f_m)$ defines the threshold of detectability of a sinusoidal distortion as a function of frequency; cq. the masking curve. It can be shown that the masking curve at f_m is given by

$$\frac{1}{v^2(f_m)} = C_s \hat{L} \sum_i \frac{|H_{om}(f_m)|^2 |\gamma_i(f_m)|^2}{\sum_f |H_{om}(f)|^2 |\gamma_i(f)|^2 |m(f)|^2 + C_a}. \quad (8)$$

Note that by using this masking curve and Eq. 5, we can show that the distortion detectability is equal to

$$D(m, s) = \sum_f \frac{|s(f)|^2}{v^2(f)}. \quad (9)$$

In this last expression it is clear that once v is calculated, the distortion detectability, D, can be computed highly efficiently. This is a very useful property for audio coding algorithms that optimize for the least possible perceptual distortion during the encoding process.

3. CALIBRATION OF THE MODEL

For the calibration of the model we will use two findings from psychoacoustical literature: the threshold in quiet at 1 kHz, and the Just Noticeable Difference (JND) of 1 dB at 70 dB SPL. The calibration constants, C_a and C_s, have to be chosen such that the model correctly predicts these findings.

For the threshold-in-quiet (absolute threshold) we can assume that $m(f) = 0$ and that $s(f) = H_{tq}(f)\delta(f - f_{1\,\mathrm{kHz}})$. Substituting this in Eq. 5, and assuming that $D = 1$, we obtain

$$C_a = C_s \hat{L} \sum_{i=1} |\gamma_i(f_{1\,\mathrm{kHz}})|^2. \quad (10)$$

For the 1-dB JND in level, we need to assume that this situation corresponds to a masking situation where a 1-kHz sinusoidal distortion has to be detected in the presence of a 1-kHz sinusoidal masker. For this configuration the distortion has to be 17 dB lower in level as compared to the masker, assuming that the masker and distortion are added in-phase. Thus we assume that $m(f) = A_{70}\delta(f - f_{1\,\mathrm{kHz}})$ and $s(f) = A_{53}\delta(f - f_{1\,\mathrm{kHz}})$, with A_{70} and A_{53} the amplitudes for a 70-, and 53-dB SPL signal, respectively. This leads to the expression

$$\frac{1}{C_s} = \hat{L} \sum_i \frac{|H_{om}(f_{1\,\mathrm{kHz}})|^2 |\gamma_i(f_{1\,\mathrm{kHz}})|^2 A_{53}^2}{|H_{om}(f_{1\,\mathrm{kHz}})|^2 |\gamma_i(f_{1\,\mathrm{kHz}})|^2 A_{70}^2 + C_a}. \quad (11)$$

By substituting Eq. 10 into Eq. 11, we get an expression where only C_s is unknown and which has a unique solution for $C_s > 0$. With, e.g., a bisection method, we can easily find a solution that satisfies this expression [11].

4. MODELLING RESULTS

In this section some basic psychoacoustical masking data obtained with synthetic signals will be compared to the predictions of the new model. In Fig. 2 masking curves are shown for white-noise maskers with a spectrum level of 30 dB/Hz (top panel) and of a 50-dB SPL 1-kHz sinusoidal masker (lower panel) together with data from some relevant psychoacoustical studies [4, 12, 5].

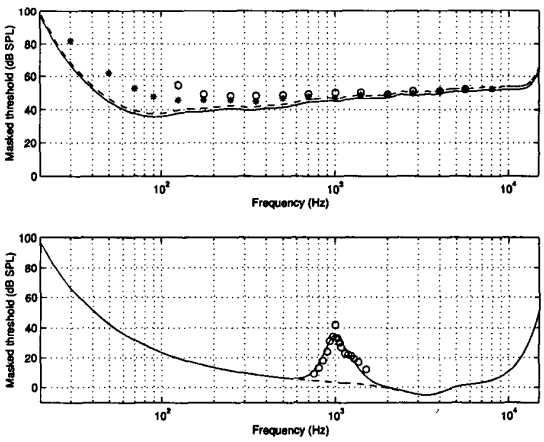

Fig. 2. *The top panel shows masking curves predicted by the model for a white noise with a spectrum level of 30-dB/Hz for a long duration signal (solid line) and a 200-ms signal (dashed line) with corresponding data represented by the circles [4] and astrices [12], respectively. The lower panel shows masking curves for a 1-kHz 50 dB-SPL sinusoid. The dashed line is the threshold in quiet. Circles in the lower panel show data of [5]*

It can be seen that the model predicts the masking by tonal and noise maskers quite well without the need for a separate tonality detector such as is typically included in masking models for audio coding to account for the weaker masking power of tonal versus noisy signals [1]. There is some discrepancy at low frequencies between literature data and model predictions although less so for the more recent study [12]. A reason for the discrepancy may be that the *data* were obtained with a running-noise masker while the *model* is based on the assumption that the masker is always deterministic. This is actually the most proper assumption for the situation where audio excerpts are encoded by an audio coder. An important property of running noise is that upon each different realization of the noise, within each separate auditory filter, the masker energy will vary. This variability has been shown to result in an extra masking effect as compared to the situation where a fixed (frozen) noise masker is used [13]. Specifically, at low frequencies where the auditory filter bandwidth is small, this effect would be expected to be greatest [14].

As noted, the model is in line with literature data that show a weaker masking power for tonal signals than for noisy signals. In this model, the relatively weak masking power, or high distortion sensitivity for tonal maskers is caused by the fact that for a tonal masker and signal, a range of filters centered around the tonal masker contributes to the detectability of the distortion. For a noise masker with a tonal distortion, only the filter centered around the tonal distortion contributes to the detectability of the distortion. This better detectability for tonal maskers as compared to noisy maskers can be translated into a lower masking threshold for a tonal masker than for a noisy masker.

To conclude this section we present model predictions for a distortion signal with a varying bandwidth in the presence of a wideband masker (0-2 kHz) of 80-dB SPL to demonstrate that the model correctly describes spectral integration in auditory masking conditions. The distortion signal consists of a series of equal-level sinusoidal components centred around 400 Hz with intercomponent spacings of 10 Hz. When the series consists of a few components only, the distortion-signal components fall within a single auditory filter. For a large number of components, the components are distributed over a range of auditory filters. Masking thresholds were obtained using Eq. (7). In Fig. 4 model predictions (solid line) versus data (circles) are shown [6]. The model predictions show a good correspondence with the psychoacoustical data. Such a good prediction would not have been obtained with conventional masking models used in audio coding which assume that wide band distortion signals are not detectable provided that there is no component that exceeds the masking curve e.g. [1].

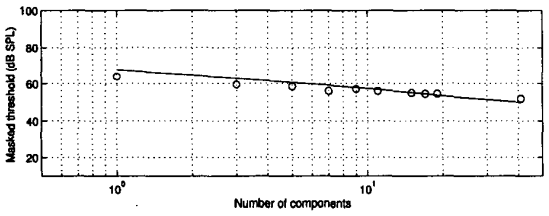

Fig. 3. *Masking thresholds predicted by the model (solid line) and psychoacoustical data (circles) [6]. Masked thresholds are expressed in dB SPL per component.*

5. CONCLUSIONS

The model presented here gives a computationally efficient description of auditory spectral masking that is very useful for the application within lossy audio coders. In contrast to existing models used for this purpose, this model describes spectral integration of distortion detectability in a proper way. In addition, this model predicts the different masking power of noisy versus sinusoidal maskers, thereby precluding the need of a separate tonality detector to account for the differential masking power of both signal types.

6. REFERENCES

[1] ISO/MPEG Committee, *Coding of moving pictures and associated audio for digital storage media at up to about 1.5Mbit/s - part 3: Audio*, 1993, ISO/IEC 11172-3.

[2] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proceedings of the IEEE*, vol. 88, pp. 451–513, 2000.

[3] H. Fletcher, "Auditory patterns," *Reviews of Modern Physics*, vol. 12, pp. 47–65, 1940.

[4] J.E. Hawkins and S.S. Stevens, "The masking of pure tones and of speech by white noise," *J. Acoust. Soc. Am.*, vol. 22, pp. 6–13, 1950.

[5] E. Zwicker and A. Jaroszewski, "Inverse frequency dependence of simultaneous tone-on-tone masking patterns at low levels," *J. Acoust. Soc. Am.*, vol. 71, pp. 1508–1512, 1982.

[6] A. Langhans and A. Kohlrausch, "Spectral integration of broadband signals in diotic and dichotic masking experiments," *J. Acoust. Soc. Am.*, vol. 91, pp. 317–326, 1992.

[7] B.R. Glasberg and B.C.J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, pp. 103–138, 1990.

[8] J.P. Egan W.A. Lindner and D. McFadden, "Masking-level differences and the form of the psychometric function," *Perception & Psychophysics*, vol. 6, pp. 209–215, 1969.

[9] G. van den Brink, "Detection of tone pulse of various durations in noise of various bandwidths," *J. Acoust. Soc. Am.*, vol. 36, pp. 1206–1211, 1964.

[10] R. Heusdens and S. van de Par, "Rate-distortion optimal sinusoidal modeling of audio and speech using psychoacoustical matching pursuits," Orlando, Florida, USA, May 2002, Submitted to ICASSP'02.

[11] G. Charestan R. Heusdens and S. van de Par, "A gammatone based psychoacoustical modeling approach for speech and audio coding," in *Proceedings ProRISC/IEEE: Workshop on Circuits, Systems and Signal Processing*, Veldhoven, the Netherlands, 2001.

[12] A.J.M. Houtsma, "Hawkins and Stevens revisited at low frequencies," *J. Acoust. Soc. Am.*, vol. 103, pp. 2848, 1998.

[13] A. Langhans and A. Kohlrausch, "Differences in auditory performance between monaural and diotic conditions. I: Masked thresholds in frozen noise," *J. Acoust. Soc. Am.*, vol. 91, pp. 3456–3470, 1992.

[14] S. van de Par and A. Kohlrausch, "Dependence of binaural masking level differences on center frequency, masker bandwidth and interaural parameters," *J. Acoust. Soc. Am.*, vol. 106, pp. 1940–1947, 1999.