A LOW-COMPLEXITY SPECTRO-TEMPORAL BASED PERCEPTUAL MODEL

Cees Taal and Richard Heusdens

Delft University of Technology
Dept. of Mediamatics
2628 CD Delft, The Netherlands
email: $\{$C.H.Taal, R.Heusdens$\}$ @TUDelft.nl

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. Up to now, these masking models are mostly based on spectral masking. In this paper, we propose a new perceptual model for audio and speech processing algorithms based on spectro-temporal masking. A sophisticated perceptual model is simplified, such that the eventual distortion measure can be written as a frequency-weighted l_2-norm. This yields the same computational complexity as conventional spectral-based methods, but with the preservation of the temporal fine structure of the clean signal. It is shown that the new model can successfully avoid pre-echoes and can correctly predict masking curves for various maskers.

Index Terms— psychoacoustics, audio coding, auditory masking

1. INTRODUCTION

It is well-known that the properties of the auditory system play an important role for various audio and speech processing algorithms. One common example is transparent audio coding where, by reducing the bit-rate, errors are introduced to a signal such that the distorted signal is perceptually indistinguishable from the original (e.g. [1, 2]). A typical approach is to shape the quantization error in the frequency domain, on a frame-by-frame basis, according to the so-called masking threshold. As long as the error signal is below this threshold, the original signal will act as a masker on the error signal.

In the MPEG perceptual model this masking threshold is found by first separating the signal in tonal and noise maskers, after which for each of these spectral components a spreading function is defined [1]. Then, by power addition of these spreading functions, a masking threshold can be obtained. This method is based on the assumption that the detectability of a specific frequency component is only determined by the auditory filter centered around that particular frequency. Since this is not in line with various results in literature (e.g. [3]), which suggests that the detectability of a specific frequency component is also determined by off-frequency auditory filters, van de Par et al. introduced a perceptual distortion measure based on spectral integration [4]. This method shows a better correspondence with data from psychoacoustic listening tests than the MPEG model, without separating the signal into tonal and noise maskers, and leads to better coding results for various fixed bit-rates [4]. In addition, the distortion measure is expressed as a mathematical norm, which allows for incorporating perceptual properties in least squares optimization algorithms, e.g. [5].

Both mentioned methods, like many other spectral-masking models, assume that the introduced error occurs simultaneously with the clean signal within one frame and do not take any temporal changes into account. The consequence is that if an error is introduced before an onset of a clean signal in the same frame, both models will consider the error to be masked, which is not the case. This leads to pre-echoes, which are unwanted perceptual artifacts [2]. There are several solutions available to prevent these pre-echoes (i.e. window switching [2], temporal noise shaping [6]). However, these methods are heuristic in nature and cannot be used to derive analytic solutions for, for example, least-squares optimization problems.

Although there are better perceptual models available which do take spectral and temporal information into account (e.g. [7]), their complexity is often too high. The main reason for this complexity is that a masking threshold for a given error signal can only be found using adaptive procedures [7]; an analytic expression is not available. This means that in a coding environment, for each newly introduced quantization level the model must be applied several times to find an estimation of its masking threshold, which is computationally highly demanding.

In this paper a new perceptual model is proposed that has the benefits of a spectro-temporal perceptual model, but has a computational complexity that is of the same order of magnitude as that of existing spectral-masking models. In the proposed method, time-consuming stages (e.g. auditory filter-bank) only have to be applied once to the clean signal. The perceptual distortion measure can then be expressed as a frequency-weighted l_2-norm.

This paper is organized as follows. First, in Section 2, a general approach of the perceptual model will be explained, after which more mathematical details are given for the eventual distortion measure. In Section 3 we describe experimental results obtained by computer simulation and compare the proposed method with a state-of-the-art spectral-masking method. Finally, in Section 4, we draw some conclusions.

2. PROPOSED PERCEPTUAL MODEL

Typically, in a spectro-temporal perceptual model, several stages of the auditory periphery are simulated, resulting in an internal time-frequency representation of a clean signal, say x, and a distorted version of the clean signal, say y. By assuming that these internal representations are degraded by additive internal noise, a signal detection approach can be used to express the discriminability between the two signals [7], where the internal noise represents the uncer-

The research is supported by Oticon and the Technology Foundation STW.

tainty in the detection process. If we specify the internal noise as a Gaussian i.i.d. stochastic process with zero mean and variance σ^2, and the detection of the difference between the two signals is optimal, the perceptual difference for an auditory filter can be expressed as

$$d_i' = \sigma^{-1}\sqrt{\sum_n \left(R_y\left(i,n\right) - R_x\left(i,n\right)\right)^2}, \tag{1}$$

where d_i' indicates the 'within-channel' sensitivity index [8] of the i^{th} auditory filter and $R_x\left(i,n\right)$ and $R_y\left(i,n\right)$ represent the internal representations of the clean and the distorted signal, respectively. Here, n denotes the time-sample index and i the index of the auditory filter.

For the proposed method we define the noisy signal as $y = x + \varepsilon w_k$, where ε is some additive error signal, windowed by the window function w_k (e.g. Hanning). Here, k refers to the k^{th} frame, which corresponds to the window start and stop positions located at $n = k$ and $n = k + N - 1$, respectively. We make the assumption that ε is statistically independent of x, and that the window function has a short support ($\approx 5 - 40$ ms). In order to include the filter tails of the internal representation, which may influence the eventual detectability, ε is zero-padded up to a length of M samples ($M \approx$ 80ms).

We are interested in the level of the error signal such that it is still masked (i.e.just not detectable) by the clean signal. Motivated by this, the assumption is made that the energy of the error signal is smaller than the energy of the clean signal. As an example, εw_k could be the quantization error introduced in a specific frame by an audio coder and x the original, clean signal. For notational convenience the window and frame index will be excluded and εw_k is simply denoted by ε.

The internal representation for the new method can be obtained by applying the stages illustrated in Fig. 1. First, to let the model correctly predict the absolute hearing threshold, an outer-middle ear filter is applied, where its magnitude response equals the inverse of the threshold in quiet. Then, a gammatone filter-bank is used to simulate the basilar membrane. To simplify the notation, the outer-middle ear filter and the i^{th} gammatone filter are together denoted by h_i. Next, a squared Hilbert envelope[1], followed by a smoothing low-pass filter, say h_s, with a cut-off frequency of 500 Hz is applied to simulate the haircell behavior. This gives the following expression at the output of the smoothing filter

$$x_i = \left|\left(x * h_i\right)_a\right|^2 * h_s. \tag{2}$$

Then, to introduce an absolute threshold, a constant c is added followed by a log-transform. The use of a log-transform instead of a non-linear gain control, as is done in more sophisticated perceptual models (e.g. [7]), has the consequence that neural adaptive properties for fast temporal fluctuations are not taken into account (e.g. forward masking). However, the model will still be able to predict the temporal fine structure of the input signal. Hence, sensitivity to errors introduced just before the onset of a signal is still preserved. Finally, to reduce the temporal resolution, all samples for a specific zero-padded frame k are set to its average value. This yields

$$R_x\left(i,n\right) = \frac{1}{M}\sum_{m=k}^{k+M-1} \log\left(x_i\left(m\right) + c\right), \tag{3}$$

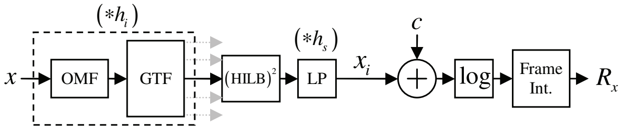

Fig. 1. *General structure for the internal representation.*

for $n = k, \ldots, k+M-1$. Note, that due to the temporal integration, *all* samples for frame k are set to the same constant value. By finding the internal representation of y, in the same manner as in (3), and by exploiting the linear properties of the Hilbert transform and the smoothing filter, (1) can be expressed as follows

$$d_i' = \frac{1}{\sigma\sqrt{M}}\sum_{m=k}^{k+M-1} \log\left(1 + \frac{\varepsilon_i\left(m\right) + v_i\left(m\right)}{x_i\left(m\right) + c}\right). \tag{4}$$

Here, ε_i denotes $\left|\left(\varepsilon * h_i\right)_a\right|^2 * h_s$, the hair-cell output when ε is fed into the system. The signal v_i contains various cross-terms between the clean and error signal and can be expressed as

$$v_i = 2h_s * \left(\left(x * h_i\right)\left(\varepsilon * h_i\right) + \widetilde{\left(x * h_i\right)}\widetilde{\left(\varepsilon * h_i\right)}\right), \tag{5}$$

where $\widetilde{(.)}$ indicates the Hilbert transform. Next, we approximate (4) by a first-order Taylor series expansion around $\left(\varepsilon_i + v_i\right)\left(x_i + c\right)^{-1} = 0$, which results in

$$d_i' \approx \frac{1}{\sigma\sqrt{M}}\sum_{m=k}^{k+M-1} \frac{\varepsilon_i\left(m\right) + v_i\left(m\right)}{x_i\left(m\right) + c}, \tag{6}$$

and is a good approximation for small ε. Since we assumed that ε and x are statistically independent, and we are integrating over a complete frame, the following term will be close to zero

$$\sum_{m=k}^{k+M-1} \frac{v_i\left(m\right)}{x_i\left(m\right) + c} \approx 0. \tag{7}$$

This suggests that v_i can be neglected from (6), without introducing any significant errors, which yields

$$d_i' \approx \frac{1}{\sigma\sqrt{M}}\sum_{m=k}^{k+M-1} \frac{\varepsilon_i\left(m\right)}{x_i\left(m\right) + c}. \tag{8}$$

Finally, to include the spectral integration property of the auditory system, the within channel detectabilities for all i are combined by an additive operation [4]

$$d \approx \sum_i d_i' \approx \frac{1}{\sigma\sqrt{M}}\sum_i\sum_{m=k}^{k+M-1} \frac{\varepsilon_i\left(m\right)}{x_i\left(m\right) + c}. \tag{9}$$

For some applications the distortion measure as defined in (9) can still be computational complex. Assume that this measure will be used in a rate-distortion loop to check whether an introduced quantization error is perceptible. Although the denominator can be pre-calculated, the auditory filter-bank and the hair-cell model still have to be applied for every new introduced error. In order to further simplify the model, we generalize the temporal structure of ε_i, by assuming that it is equally distributed over the frame. It can be expected that the energy of the error, within one auditory channel,

[1] The Hilbert envelope of any arbitrary signal x is defined as the absolute value of its analytic signal $x_a = x + j.\tilde{x}$, where $\tilde{x}$ denotes the Hilbert transform of x.

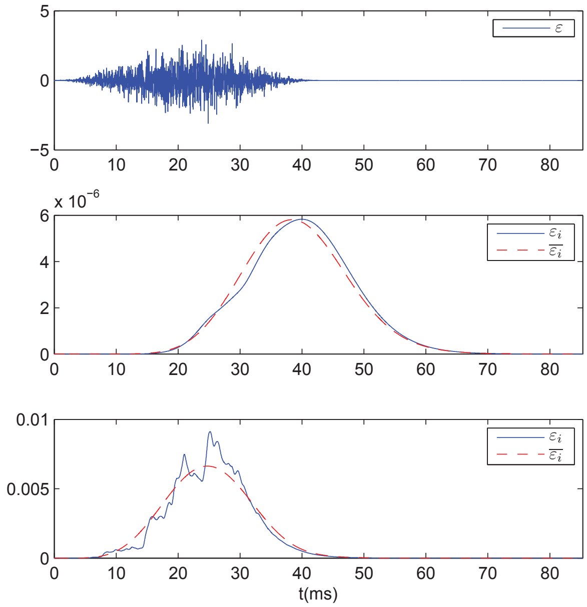

Fig. 2. *Normal, and generalized haircell output for a windowed noise signal. The middle and bottom plots indicate the gammatone filters with center frequencies 150 Hz, and 2000 Hz, respectively.*

will be centralized around the center frequency of the corresponding gammatone filter, say $\omega_{c(i)}$. Hence, certain frequency dependent filter properties (e.g. group delay) are mainly determined by $\omega_{c(i)}$. Motivated by this, we introduce the following generalized haircell output

$$\overline{\varepsilon_i}\left(n\right) = z_i\left(n\right)\left(\sum_{p=k}^{k+M-1} z_i\left(p\right)\right)^{-1}\sum_{l=k}^{k+M-1}\varepsilon_i\left(l\right), \qquad (10)$$

where z_i denotes the haircell output for $w_k\left(n\right)e^{j\omega_c(i)n}$; the windowed complex exponential centered at the center frequency of the corresponding gammatone filter. Note that z_i is normalized in order to preserve the mean value of the original haircell output. Fig. 2 illustrates an example where ε equals a 40 ms Hanning-windowed noise signal (top plot). The generalized, and the original haircell output are shown in the middle and bottom plot for the gammatone filters with center frequencies 150 Hz, and 2000 Hz, respectively. Note, that the generalized haircell output shows a good approximation for the rise and fall times for both gammatone filters.

By replacing ε_i by $\overline{\varepsilon_i}$ in (9), a new distortion measure D can be defined as

$$D = \sum_i \sum_{l=k}^{k+M-1} \varepsilon_i\left(l\right) g_i, \qquad (11)$$

where g_i indicates a weighting function, only dependent on i, specified by

$$g_i = \frac{1}{\sigma\sqrt{M}} \sum_{m=k}^{k+M-1}\left(\sum_{l=k}^{k+M-1} z_i\left(l\right)\right)^{-1} \frac{z_i\left(m\right)}{x_i\left(m\right)+c}. \qquad (12)$$

Regarding the summation in (11) over l for a specific i, it can be concluded that its outcome equals the DC-coefficient of ε_i, weighted by g_i. Since the (normalized) smoothing filter of the hair-cell model will not affect this outcome, it can be discarded. Hence

$$D = \sum_i \sum_{l=k}^{k+M-1} \left|\left(\varepsilon * h_i\right)_a\left(l\right)\right|^2 g_i. \qquad (13)$$

Then by applying Parseval's theorem, the following result can be obtained

$$D = \sum_{f=0}^{M-1} \left|\hat{\varepsilon}\left(f\right)\right|^2 a\left(f\right), \qquad (14)$$

where $\widehat{(.)}$ indicates the discrete Fourier transform and a is a weighting function denoted by

$$a\left(f\right) = \frac{4}{M}\sum_i g_i \left|\hat{h}_i\left(f\right)\right|^2 u^2\left(f\right). \qquad (15)$$

Here, $u\left(f\right)$ represents the unit step function where $u\left(0\right) = \frac{1}{2}$, which origins from the derivation of the analytic signal.

A couple of interesting conclusions can be drawn from (15). The weighting function a is independent of ε and can be pre-calculated for each frame, stored and reused. The result is that, in order to evaluate (14) for *any* ε, only one FFT has to be applied followed by a simple linear weighting. This has the same computational complexity as the spectral integration method used in [4], but now also based on temporal changes of the clean signal. Another important property is that, for the frequency range $f = 0, ..., M/2$, the weighting function a is real and positive so that, in fact, the perceptual distortion measure defines a norm for all real input signals, assuming that $\varepsilon w_k \neq 0$ for all $\varepsilon \neq 0$.

In many audio applications a masking curve is used; the masking threshold for a particular frequency component. Using the proposed model, we can define a masking curve by computing, for each frequency, the amount of distortion that is just not detectable, e.g., for which $D = 1$. By setting $\left|\hat{\varepsilon}\left(v\right)\right|^2 = \left|\hat{\varepsilon}\left(f\right)\right|^2 \delta\left(v - f\right)$, that is, all the energy of the distortion is concentrated at one single frequency, we can define a masking curve mc as follows:

$$mc\left(f\right) = \left(\frac{4}{M}\sum_i g_i \left|\hat{h}_i\left(f\right)\right|^2 u^2\left(f\right)\right)^{-1}. \qquad (16)$$

3. EXPERIMENTAL RESULTS

To evaluate the new method, a comparison is made with the perceptual model developed by van de Par et al. [4]. The degrees of freedom for both models are calibrated such that they correctly predict the threshold in quiet and the 1 dB just noticeable level difference for a 1 kHz, 70 dB SPL tonal masker (see [4] for more details).

Fig. 3 shows the masking curves for both methods, for a 30 dB/Hz noise masker just before and after the masker onset, indicated by frame1 and frame2 in the top plot, respectively. The length of the window equals 200 ms, including 10 ms fade times. The masking curves, for both frames, are showed in the bottom two plots together with the threshold in quiet.

Since the pre-masking property of the auditory system only occurs as from 10 ms before the onset of the masker [9], the masking curve for the first frame should be close to the threshold in quiet, which is in correspondence with the results predicted by the new

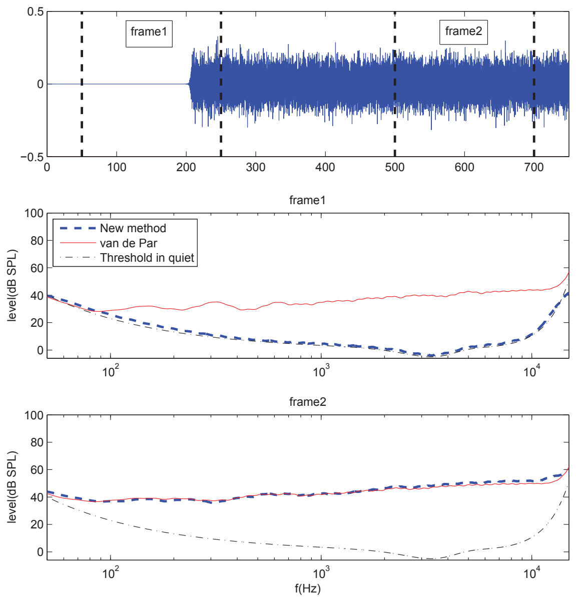

Fig. 3. *Input signal and masking curves for the proposed, and the van de Par-model, for a 30 dB/Hz noise masker.*

method. This is not the case for the van de Par-model, which overestimates the masking curve severely. Hence, the new model is significantly more sensitive to errors introduced before the masker onset than the van de Par-model. However, both models do agree on the stationary signal in frame 2, which is in line with data from real listening tests, as was already shown in [4]. Experiments with tonal maskers showed the same behavior; the new method successfully detects the silence before the onset and for stationary frames both models predict the same masking curve.

To illustrate the properties of the new model, it is included in a simple overlap-add DFT-based coding scheme and compared with the van de Par-model. An optimal flexible segmentation algorithm is used [10], where for each segment certain DFT-coefficients are set to zero, such that the total distortion of the complete signal is minimized. A constraint was set on the total amount of preserved DFT-coefficients. The segmentation algorithm was performed with possible segment sizes of 5, 10, 20 and 40 ms with a fixed overlap of 5 ms. To indicate the difference between the two methods, a percussive sound example with strong transients is used.

Fig. 4 shows the results (500 ms) for both methods, given the constraint that on average 20% of the DFT-coefficients is preserved. For each method, on top of the clean signal, the error signal is plotted together with the corresponding optimal segmentation. For visual clarity, the error signal is amplified. For the new method, the top figure clearly indicates that the segmentation algorithm tries to avoid encountering transients within one frame. This is due to the sensitivity to errors introduced before the masker onset, which was already shown in Fig. 3. In this manner, pre-echoes are avoided, which is not the case for the segmentation found for the van de Par-model, where at $t = 500$ ms clearly a pre-echo can be observed.

Preliminary listening tests indicate that listeners prefer the proposed method for signals with transients, while no perceptual degradation was observed for more stationary signals. Currently, formal listening tests are performed to test the method extensively.

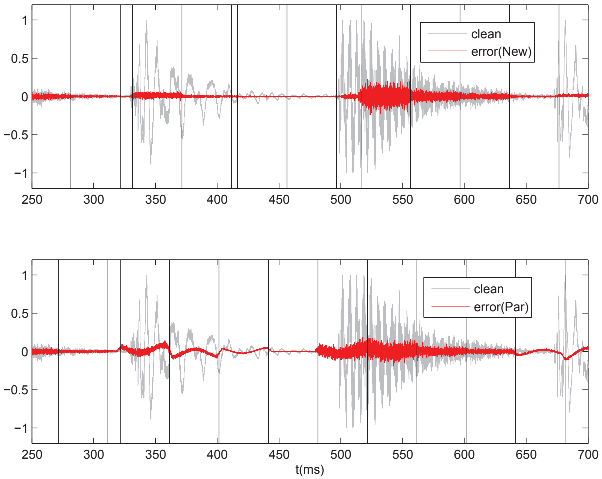

Fig. 4. *Optimal segmentations for the new method and the van de Par-Model.*

4. CONCLUSIONS

A new perceptual model for audio and speech processing algorithms is proposed, based on spectro-temporal masking. The eventual distortion measure is defined as a frequency-weighted l_2-norm, which yields the same complexity as spectral-based models, but with the preservation of the temporal fine structure of the clean signal. It is shown that the new model can successfully avoid pre-echoes and can correctly predict masking curves for various maskers.

5. REFERENCES

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

[2] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proc. of the IEEE*, vol. 88, no. 4, pp. 451–515, 2000.

[3] S. Buus, E. Schorer, M. Florentine, and E. Zwicker, "Decision rules in detection of simple and complex tones," *JASA*, vol. 80, pp. 1646, 1986.

[4] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP JASP*, vol. 2005, no. 9, pp. 1292–1304, 2005.

[5] R. C. Hendriks, R. Heusdens, and J. Jensen, "Perceptual linear predictive noise modelling for sinusoid-plus-noise audio coding," *IEEE ICASSP*, vol. 4, pp. 189–192, 2004.

[6] J. Herre and J. D. Johnston, "Enhancing the performance of perceptual audio coders by using temporal noise shaping (tns)," *Proc. 101st Conv. AES*, 1996.

[7] T. Dau, D. Puschel, and A. Kohlrausch, "A quantitative model of the effective signal processing in the auditory system. i. model structure," *JASA*, vol. 99, no. 6, pp. 3615–3622, 1996.

[8] D. McNicol, *A Primer of Signal Detection Theory*, 2005.

[9] E. Zwicker and H. Fastl, "Psychoacoustics: Facts and models," 1990.

[10] P. Prandoni, M. Goodwin, and M. Vetterli, "Optimal time segmentation for signal modeling and compression," *IEEE ICASSP*, vol. 3, 1997.