Improved Audio Coding Using a Psychoacoustic Model Based on a Cochlear Filter Bank

Frank Baumgarte

***Abstract*—Perceptual audio coders use an estimated masked threshold for the determination of the maximum permissible just-inaudible noise level introduced by quantization. This estimate is derived from a psychoacoustic model mimicking the properties of masking. Most psychoacoustic models for coding applications use a uniform (equal bandwidth) spectral decomposition as a first step to approximate the frequency selectivity of the human auditory system. However, the equal filter properties of the uniform subbands do not match the nonuniform characteristics of cochlear filters and reduce the precision of psychoacoustic modeling. Even so, uniform filter banks are applied because they are computationally efficient. This paper presents a psychoacoustic model based on an efficient nonuniform cochlear filter bank and a simple masked threshold estimation. The novel filter-bank structure employs cascaded low-order IIR filters and appropriate down-sampling to increase efficiency. The filter responses are optimized for the modeling of auditory masking effects. Results of the new psychoacoustic model applied to audio coding show better performance in terms of bit rate and/or quality of the new model in comparison with other state-of-the-art models using a uniform spectral decomposition. The low delay of the new model is particularly suitable for low-delay coders.**

***Index Terms*—Audio coding, filter bank, masked threshold, model of masking, perceptual model.**

I. Introduction

In PERCEPTUAL audio coding [1], the audio signal is treated as a masker for distortions introduced by lossy data compression. For this purpose, the masked threshold for the distortions is approximated by a psychoacoustic model. The masked threshold is the time and frequency-dependent maximum level that marks the boundary for distortions being inaudible if superimposed to the audio signal. The initial audio signal processing within the psychoacoustic model consists of a spectral decomposition to account for the frequency selectivity of the auditory system. However, the auditory system performs a nonuniform (nonequal bandwidths) spectral decomposition of the acoustic signal in the cochlea. This first stage of cochlear sound processing already determines basic properties of masking, e.g., the frequency spread of masking which is related to the frequency response of the human cochlear filters. Above 1 kHz, the cochlear filter bandwidths increase almost proportionally to the center frequency. These bandwidths determine both, the spectral width of energy integration associated with a band and the range of spectral components that can interact within a band, e.g., two sinusoids creating a beating effect. This interaction plays a crucial role in the perception of whether a sound is noise-like which in turn corresponds to a significantly more efficient masking compared with a tone-like signal [2]. The noise or tone-like character is basically determined by the amount of envelope fluctuations at the cochlear filter outputs which widely depend on the interaction of the spectral components in the pass-band of the filter.

Many existing psychoacoustic models, e.g., [1], [3], and [4], employ an FFT-based transform to derive a spectral decomposition of the audio signal into uniform subbands with equal bandwidths. The nonuniform spectral resolution of the auditory system is taken into account by summing up the energies of the appropriate number of neighboring FFT frequency subbands. Consequently, the phase relation between the spectral components of the different subbands within a cochlear filter band is not taken into account. Since the cochlear filter slopes are less steep than the subband slopes, they must be approximated by spreading the subband energies across several bands. This way of mapping the uniform subbands to cochlear filter bands produces envelopes of the output signal that are different from those measured at the output of the cochlea. The temporal resolution of the spectral decomposition is determined by the transform size, i.e., FFT length, and thus, is constant across all center frequencies. For high center frequencies this results in a significantly lower temporal resolution in comparison with that of the corresponding cochlear filters. All the described mismatches contribute to an inaccurate modeling of masking that causes suboptimal coder compression performance.

To overcome the mismatch between uniform filter banks and the spectral decomposition of the cochlea, a linear nonuniform cochlear filter bank was developed. A linear filter bank was chosen because it is computationally less complex than a nonlinear one [5], [6]. Furthermore, a psychoacoustic model based on a nonlinear filter bank generally approximates the masked threshold in an iteration process. Applied to audio coding, this involves encoding, decoding, and threshold computation for each iteration step, which can considerably increase the encoder complexity. The linear filter bank does not account for sound level-dependent effects. However, since the playback level of the decoded audio signal is usually unknown, this is considered a minor restriction only.

The cochlear filter bank is based on a novel structure that supports the time- and frequency resolution necessary to simulate psychophysical data closely related to cochlear spectral decomposition properties. It will be shown that this filter bank is able to closely mimic the spectral and temporal properties

Manuscript received June 20, 2001; revised July 18, 2002. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Peter Vary.

The author is with the Media Signal Processing Research Department, Agere Systems, Berkeley Heights, NJ 07922 USA (e-mail: fb@agere.com).

Digital Object Identifier 10.1109/TSA.2002.804536

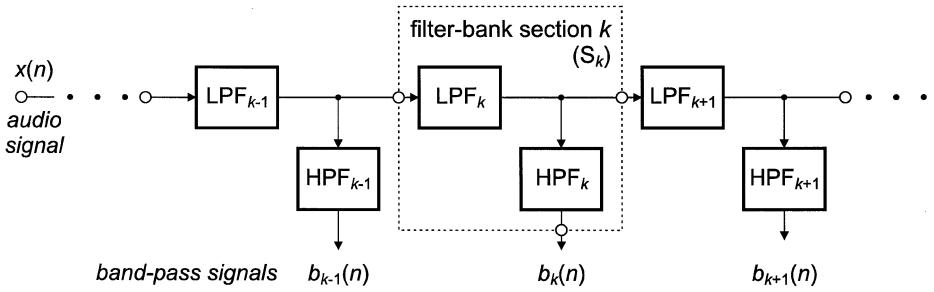

Fig. 1. Block diagram of the cochlear filter-bank structure.

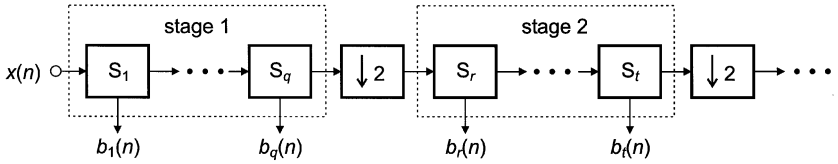

Fig. 2. Downsampling scheme of the cochlear filter bank.

of frequency decomposition of the human peripheral auditory system. The benefits of using this filter bank in a new psychoacoustic model are explained and evaluated for two different audio coders. An informal subjective quality assessment was carried out for both state-of-the-art coders. For this comparison the coders were used with their individual reference psychoacoustic model based on a uniform filter bank and with the new psychoacoustic model. Results show improved coder performance for the new psychoacoustic model.

The paper is organized as follows. The filter-bank structure is described in Section II. In Section III the filter-bank implementation using low-order IIR filters is presented. The filter responses are optimized for modeling of masked thresholds. A novel psychoacoustic model based on that filter bank is the subject of Section IV. The experimental setup of the coders used and of the subjective listening tests is outlined in Section V. Results are given in Section VI in terms of the subjective quality and data rate. Conclusions are drawn in Section VII.

II. Filter-Bank Structure

The peripheral auditory system performs spectral analysis of the input acoustic signal in the cochlea with spectrally highly overlapping band-pass filters. The nonuniform frequency resolution and bandwidth of these filters is approximated in the proposed structure by cascaded IIR filters. Fig. 1 shows the proposed filter-bank structure with low-pass filters (LPF) and high-pass filters (HPF). The LPFs in the cascade have a decreasing cutoff frequency from left to right (see Fig. 1). Each LPF output is connected to an HPF. The HPF cutoff frequency is equal to the cutoff frequency of the LPF cascade segment between the filter-bank input and the HPF input of the next section. Thus, the output of each HPF has a bandpass characteristic with respect to the filter-bank input signal. The basic block of an LPF connected to an HPF, as shown in Fig. 1, is called a filter-bank section. The filter-bank structure resembles a traditional pruned tree, however, it takes advantage of a novel frequency spacing.

The decreasing cutoff frequency of the LPF cascade permits a reduction of sampling rate, that reduces computational complexity. A simple and efficient way to implement a "stage-wise" sampling rate reduction is shown in Fig. 2, where a stage comprises a group of those cascaded filter-bank sections having equal sampling rate. The rate reduction by a factor of two is achieved by leaving out every other sample at the stage input. It is applied when the cutoff frequency of the LPF cascade output is below a given ratio with respect to the sampling rate in that stage to reduce aliasing. The number of sections covering the auditory frequency range is usually in the order of 100. It can be adapted to the desired frequency resolution for a specific application. The number of stages is typically chosen between five and nine.

All the high-pass filters have the same order. Also, all the low-pass filters have the same order. However, the LPF and HPF orders can be chosen independently and should be large enough to accurately model the spectral decomposition features found in relevant psychophysical data. After the orders are fixed, the filter coefficients can be determined by an optimization algorithm to minimize the difference between the responses of the desired and the proposed filter banks. The responses of the desired filters are generally derived from psychophysical measurements.

III. Filter Bank Implementation

In this section, the cochlear filter bank parameters are given and the derivation of the filter coefficients is described for the application in a psychoacoustic model. It turns out that an LPF order of $M_{\mathrm{LP}} = 2$ and an HPF order of $M_{\mathrm{HP}} = 4$ is sufficient to achieve a reasonable approximation of the desired frequency responses. The slopes of the desired magnitude frequency responses are chosen according to simple masking models that assume a constant slope steepness on a Bark [7] or an equivalent-rectangular-bandwidth (ERB) [8] scale. For center frequencies above 500 Hz the filter slope steepness is 30.4 dB/octave below the center frequency and 95 dB/octave above. The desired filter bandwidths and center-frequency spacing is based on the ERB scale. For simplicity, the ERB scale is approximated in the filter coefficient optimization by a constant bandwidth

below 500 Hz and bandwidths proportional to the center frequency above 500 Hz.

The first filter-bank stage is composed of 28 sections as opposed to 15 sections for all five subsequent stages. The increased number of sections in the first stage is necessary to sufficiently reduce aliasing due to the first down-sampling. The first stage has a larger input signal bandwidth with respect to the sampling rate than the other stages.

Due to the uniform linear spacing of the filter bands below 500 Hz, the filter bank is generally not scalable with sampling rate. However, the filter sections with center frequencies higher than 500 Hz are applicable independently from the sampling rate, since the filter bandwidths change proportionally to the center frequencies, i.e., their ratio remains constant. Only the linearly spaced filters for center frequencies below 500 Hz must be designed specifically for any given sampling rate. In the following, a sampling rate of 44.1 kHz is assumed. For lower sampling rates the number of sections at low center frequencies of the cochlear filter bank can be reduced. For a sampling rate of 32 kHz stage six has ten sections, for 16 kHz only five stages are necessary, with stage five consisting of ten sections.

A. Desired Frequency Responses

The filter bank covers the full range of audible frequencies. The desired center frequencies of the filter bands are uniformly distributed on a logarithmic scale above 500 Hz implying a proportional bandwidth increase. The center frequencies are chosen such that approximately two overlapping filter bands are available within one ERB. Below a center frequency of 500 Hz the filter-bank bandwidths are equal. Analytically the filter center frequency $f_c(k)$ of band $k+1$ is related to band k by (1) with $k \in \{1, 2, \ldots 102\}$

$$f_c(k+1) = \begin{cases} 0.5^{1/N_S} f_c(k), & f_c(k) \geq 500 \\ f_c(k) - 22.4, & f_c(k) < 500. \end{cases} \quad (1)$$

The first filter-bank section $(k = 1)$ has the highest center frequency. At a sampling rate of 44.1 kHz it is $f_c(1) = 20\,948$ Hz. The constant $N_S = 15$ determines the frequency resolution. It denotes the number of sections of the logarithmically-spaced stages with center frequencies larger than 500 Hz, except stage one. Increasing number of bands (sections) will yield a higher spectral resolution and a larger spectral band overlap since the filter bandwidths remain unchanged. For comparison, in the human cochlea we could assign one band to the output of each inner hair cell which amounts to about 3000 bands. The overlapping bands in the model not only result in a higher resolution than achieved by the minimum number of bands, e.g., 24 one-Bark-wide critical bands [7]. They also smooth the outputs of neighboring bands in case of frequency-modulated signal components moving into or out of the pass band.

The desired magnitude frequency response $|H(f)|$ of one band centered at f_c for $f_c \geq 500$ is defined in (2)

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

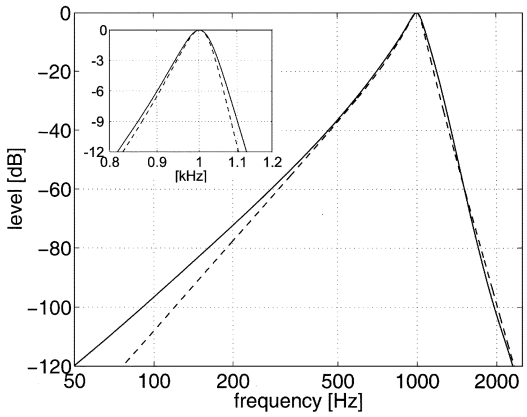

Fig. 3. Desired (dashed) and achieved (solid) magnitude response of the filter-bank channel at $f_c = 1002$ Hz. The inset shows in detail the response near the center frequency. The input audio sampling frequency is 44.1 kHz.

with

$$S_{LP} = \frac{25}{20\log_{10}(1.2)}; \quad S_{HP} = \frac{8}{20\log_{10}(1.2)}$$
$$q = 4; \quad j = \sqrt{-1}.$$

For $f_c < 500$ 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. The first term in (2) describes the steep filter slope toward high frequencies with a steepness of S_{LP}. The low-frequency slope is determined by the second term and has a steepness of S_{HP}. Both slopes are constant on a logarithmic scale with a bandwidth-to-center-frequency ratio of about 20%. Compared to a Bark scale the filter slopes are approximately 8 and -25 dB/Bark. The transition between the two slopes is controlled by a resonance quality factor q. The center frequencies f_c in (2) deviate slightly from the actual maximum of the frequency response function. Therefore, the center frequencies to be used in an application must be computed from those maxima.

B. Filter-Bank Responses

Given the desired frequency responses, the filter coefficients can be optimized using standard techniques, e.g., the damped Gauss–Newton method for iterative search [9] available in MATLAB. Fig. 3 shows the desired and the resulting magnitude frequency response of the filter at a center frequency of 1002 Hz. Near the center frequency, the deviation is small. At low frequencies, the deviation reaches about 10 dB at 100 Hz. However, this deviation is considered to have only minor effects for applications in audio coding, because the attenuation is high in this frequency range far from the center frequency. The distribution of the approximation error can be controlled by using a frequency-dependent weighting function for the error in the optimization algorithm.

Fig. 4 shows the resulting filter-bank responses of the first stage. The responses have basically the same shape on a logarithmic scale. They are shifted according to their center frequency and have large overlap. The frequency responses of stages two to four are nearly identical to the response of section $14 \cdots 28$ of stage one except they are shifted to 1/2, 1/4th,

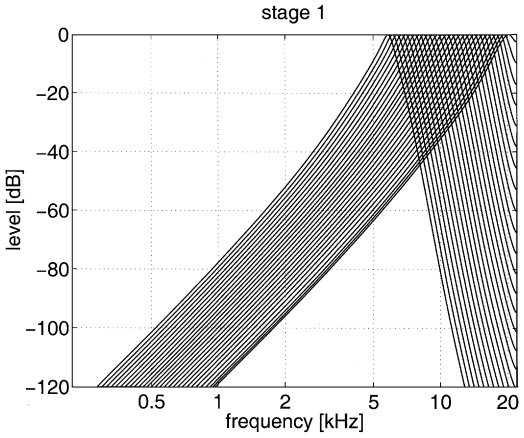

Fig. 4. Magnitude frequency responses of the filter-bank channels in stage one.

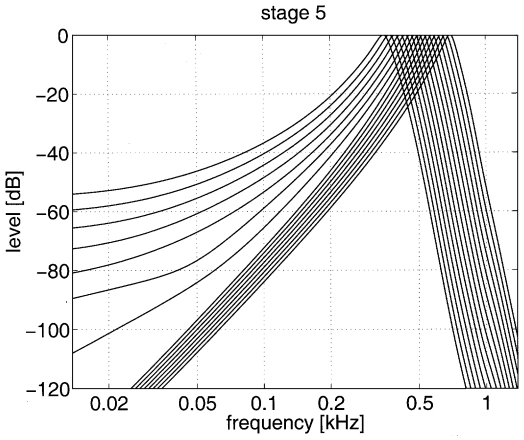

Fig. 5. Magnitude frequency responses of the filter-bank channels in stage five.

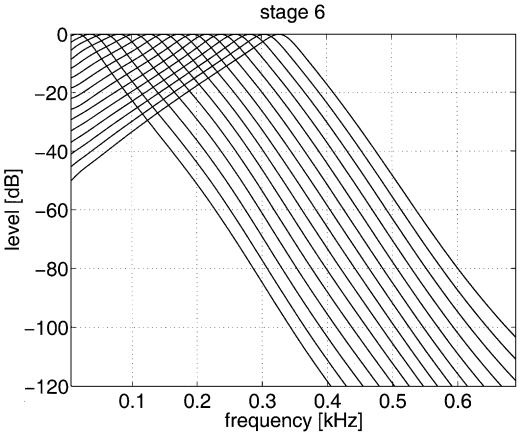

Fig. 6. Magnitude frequency responses of the filter-bank channels in stage six (note linear frequency scale).

and 1/8th the center frequency, respectively. Fig. 5 shows the magnitude responses of stage four where the transition of the uniform filter spacing from logarithmic to linear occurs. Since the frequency scale is logarithmic, the linear shift of the lower seven responses appears distorted. In Fig. 6 the linearly shifted replicas of the filter response closest to 500 Hz are shown on a linear scale for stage six. A comparison of the filter bank

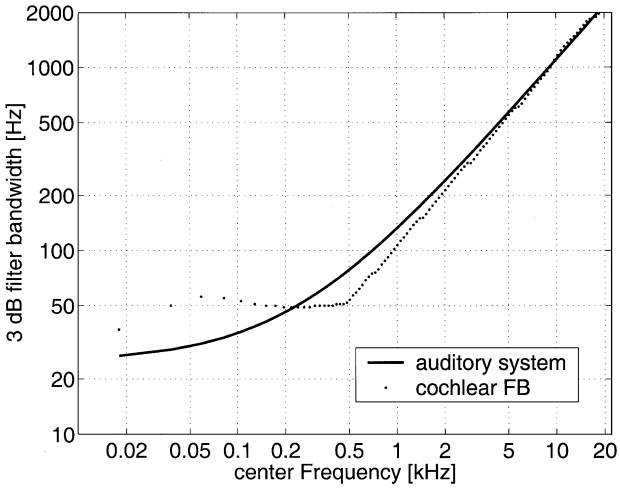

Fig. 7. Bandwidths of auditory system according to the ERB scale (solid) and of cochlear filter bank (dots). Each dot represents one filter bank output.

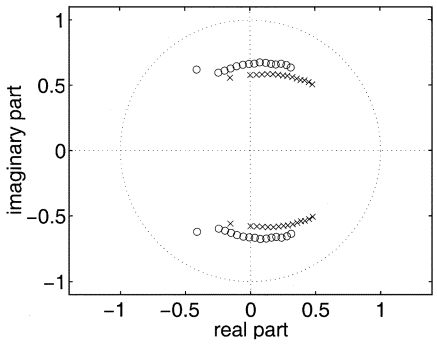

Fig. 8. Pole-zero plot of the LPF cascade in stage two (∘ zero, × pole).

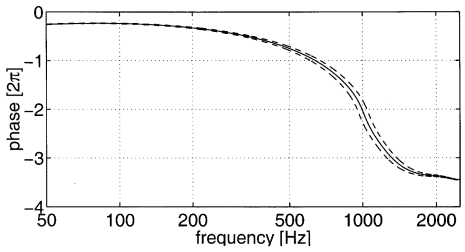

Fig. 9. Phase responses of the filter-bank channel at $f_c = 1002$ Hz and neighboring channels.

bandwidths dependent on their center frequencies with the ERBs are shown in Fig. 7. The deviation at low frequencies are assumed to be tolerable. They are caused by the simple ERB approximation of (1) and thus no inherent property of the filter bank structure.

Fig. 8 shows the location of the LPF poles and zeros in stage two. Due to their distance from the unit circle, implementation problems caused by limited arithmetic precision are unlikely. As an example, the filter bank responses will change indiscernible if the filter coefficients are quantized to 16 bits. Except the maximum HPF attenuation decreases to about 80 to 90 dB.

The phase responses of the filter-bank band in Fig. 3 and its neighbors are shown in Fig. 9. These phase responses are determined by the minimum-phase design of all LPFs and HPFs, which was chosen in accordance with models of cochlear hydro-

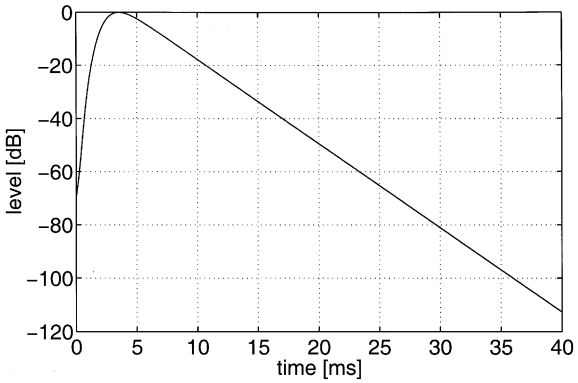

Fig. 10. Envelope of impulse response of the filter-bank channel at $f_c = 1002$ Hz.

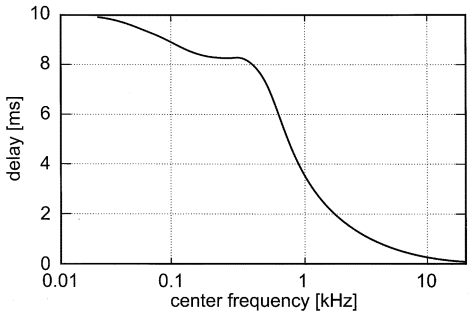

Fig. 11. Delay of cochlear filter bank (44.1 kHz sampling rate).

mechanics, for example see [5], [6]. Thus, the phase characteristic qualitatively agrees with measurements of basilar membrane velocity in the cochlea [10].

Fig. 10 shows the impulse-response envelope for the filter centered at 1002 Hz using decibel ordinate units. The modeling of temporal masking requires that the temporal spread of a filter that is reflected by its impulse response does not exceed the limits of premasking and postmasking. Premasking is generally considered to last for a few milliseconds before a masker is switched on. The temporal filter response is in the same time range, since it reaches the maximum after 3 ms. Postmasking can last for about 200 ms after a masker is switched off [7]. Since the temporal filter response shows an attenuation of more than 100 dB after 36 ms from the maximum, it fulfills the conditions above.

The time needed for the envelope to fall below a given threshold decreases with increasing filter center frequency. This duration is approximately inversely proportional to the center frequency. Thus, the filter responses above 1002 Hz do not exceed the limits of temporal masking. The time for reaching the impulse response maximum exceeds 3 ms at center frequencies well below 1002 Hz down to 500 Hz. It is assumed here that premasking duration increases at lower frequencies as well, so that the premasking duration is not exceeded. This assumption is motivated by the little amount of psychoacoustic premasking data available today.

Fig. 11 shows the filter bank delay computed from the time delay of the impulse response envelope maximum. Below a center frequency of 500 Hz, a filter delay between 8 and 10 ms exists. Above 500 Hz the delay decreases exponentially.

C. Complexity

The computational complexity and memory requirements of the cochlear filter bank can be calculated with the following equations:

$$N_{\mathrm{Mul}} = 2(M_{\mathrm{LP}} + M_{\mathrm{HP}} + 1) \sum_{l=1}^{L} Q_l R_l^{-1} \tag{3}$$

$$N_{\mathrm{Add}} = 2(M_{\mathrm{LP}} + M_{\mathrm{HP}}) \sum_{l=1}^{L} Q_l R_l^{-1} \tag{4}$$

$$N_{\mathrm{Coef}} = 2(M_{\mathrm{LP}} + M_{\mathrm{HP}} + 1) \sum_{l=1}^{L} Q_l \tag{5}$$

$$N_{\mathrm{State}} = (M_{\mathrm{LP}} + M_{\mathrm{HP}}) \sum_{l=1}^{L} Q_l. \tag{6}$$

The formulas are based on the direct form I or II filter structure. The number of multiplications, N_{Mul}, and additions, N_{Add}, relates to the processing of one input audio sample. The requirements for coefficient ROM, N_{Coef}, and state RAM, N_{State}, is calculated in words. The word length used for the described implementation is 32 bits. The complexity depends on the filter order of the LPF, M_{LP}, and HPF, M_{HP}, moreover on the number of sections per stage, Q_l, and the number of stages, L with the stage index l. The ratio R_l is the sampling rate quotient of the filter bank input and the stage output. For a cochlear filter bank as given in the previous subsection with an input sampling rate of 44.1 kHz the complexity results in $N_{\mathrm{Mul}} = 595$, $N_{\mathrm{Add}} = 510$, $N_{\mathrm{Coef}} = 1442$, and $N_{\mathrm{State}} = 618$.

For applications requiring a lower complexity the number of high frequency bands in the first filter-bank stage can be reduced by omitting the corresponding high-pass filter calculation. The coarser frequency resolution at high frequencies significantly reduces the complexity but might be still tolerable for many applications. Reducing the number of sections is another way to achieve lower complexity. But the filter slopes will be affected (become less steep) so that it might be necessary to increase the low-pass filter orders to maintain the slope steepness. Thus, a reduced number of sections does not necessarily provide a proportional complexity reduction. The coefficient ROM requirement N_{Coef} can be reduced by about 50% if only one set of coefficients is used for identical filters in different stages.

IV. Psychoacoustic Model

In this section, a psychoacoustic model is proposed based on the cochlear filter bank. It is designed for applications in audio coding that use an approximated masked threshold in the encoding process. Fig. 12 shows a simplified block diagram of the proposed psychoacoustic model. The input audio signal is processed by a filter approximating the smoothed average frequency transfer function of the outer- and middle ear (OME). It is implemented as a fifth-order IIR filter with a magnitude frequency response as shown in Fig. 13. Subsequently a spectral decomposition by the cochlear filter bank (FB) is performed as described in Section III.

The following processing steps are applied to all output bands. They are shown in Fig. 12 for one band only. After

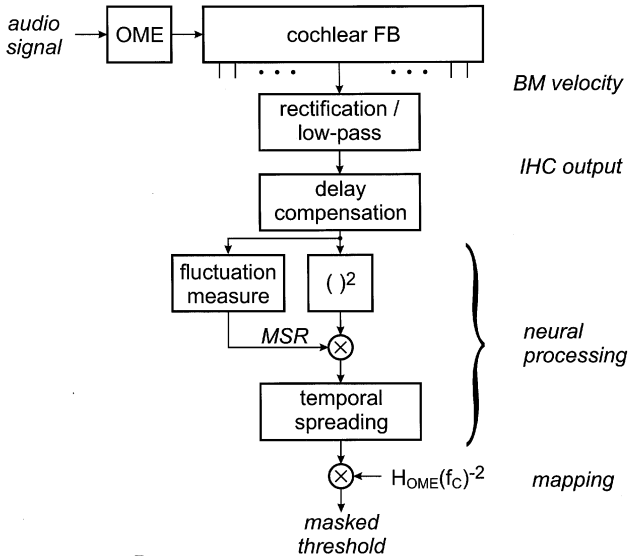

Fig. 12. Block diagram of psychoacoustic model. (BM: basilar membrane; IHC: inner hair cell).

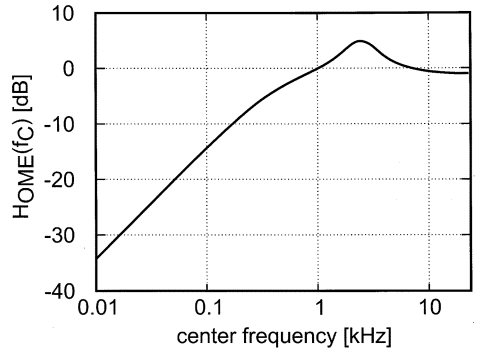

Fig. 13. Magnitude frequency response of OME filter.

cochlear filtering the effect of the inner hair cells (IHC) is taken into account by rectification and second order low-pass filtering with the cutoff frequency f_{LP} in hertz

$$f_{\mathrm{LP}} = \begin{cases} f_c, & f_c < 300 \\ 300\left(\dfrac{f_c}{300}\right)^{0.25}, & f_c \geq 300 \end{cases} \tag{7}$$

where f_c is the center frequency of the corresponding band in hertz. The cutoff frequencies are in the order of physiological data (for example see animal data in [11]). The dependency on center-frequency is not physiologically motivated. It was introduced to optimize the overall performance of the model.

The next stage compensates for a possible delay mismatch between the psychoacoustic model and the audio coder used. The delay introduced by the model is dominated by the cochlear filter bank. Fig. 11 shows that the delay decreases with increasing center frequency and has a maximum of 10 ms for a sampling rate of 44.1 kHz. For most audio coders a delay of more than 10 ms is appropriate, so that the masked threshold can be perfectly synchronized with the coder. For special low-delay coders a shorter delay may be required. In this case a perfect synchronization can still be achieved for the high frequency bands that have a sufficiently low delay. In this case, a slight temporal shift of the threshold at low frequencies can usually be tolerated and is assumed to be much less critical than a temporal mismatch at higher frequencies.

The signal representation in medium and high center frequency bands at this model stage consists of the band-pass envelopes corresponding to the inner hair cell outputs. Due to the rectification and low-pass filtering, the sampling rates of these signals can advantageously be reduced at medium and high center frequencies which creates only negligible aliasing.

The estimation of masked thresholds from this representation involves the following steps, motivated by the auditory neural processing: a square function to calculate the energy, a level offset (masker-to-signal ratio—MSR), and temporal spreading of the threshold to account for temporal masking effects. Additionally, inverse OME-filtering is necessary to map the "internal" energy representation to the external audio signal domain. The level offset is adjusted according to the amount of envelope fluctuation in the same band. The fluctuation measure corresponds to the "tonality" measure applied in other models, like [3]. Here it is based on the maximum-to-minimum ratio of the envelope as described in [5], [6]. Larger fluctuations result in a larger MSR (less tonality) and thus in a higher masked threshold level. A temporal spreading or smearing is applied afterwards to account for that part of temporal premasking and postmasking that is assumed to be created by auditory neural processing. The cochlear filter bank already introduces temporal smearing according to the shape of the filter impulse responses, but its amount is too small to fully model temporal masking effects. The inverse OME-filtering is finally done by applying a constant gain to each band. The resulting masked threshold level is valid for the band center frequency and varies in time with a maximum sampling rate corresponding to the possibly downsampled filter bank output signal.

Traditional psychoacoustic models, e.g., the basic MPEG-2 AAC model [3], have a computationally more efficient uniform filter bank compared to the cochlear filter bank. However, due to the uniform spectral resolution a mapping to cochlear filter bands and a spectral energy spreading must be done as described above which is not necessary in the new model. The mapping and spreading contribute significantly to the complexity of traditional models.

Fig. 14 shows masked thresholds produced by the model in Fig. 12 for a 160-Hz-wide Gaussian noise masker centered at 1 kHz. The different masking curves are randomly selected samples from different time instances and reflect the fluctuating nature of the masker. The masked threshold at the output of each model band is assigned to the band center frequency. For example, a probe signal at a band center frequency is assumed to be inaudible, if its level is below the calculated masked threshold. As expected, the masked threshold resembles the filter response at 1 kHz on a reversed frequency scale.

V. Experiments

To evaluate the performance of the new psychoacoustic model it was implemented in two state-of-the-art audio coders by replacing the original (reference) model with the new one.

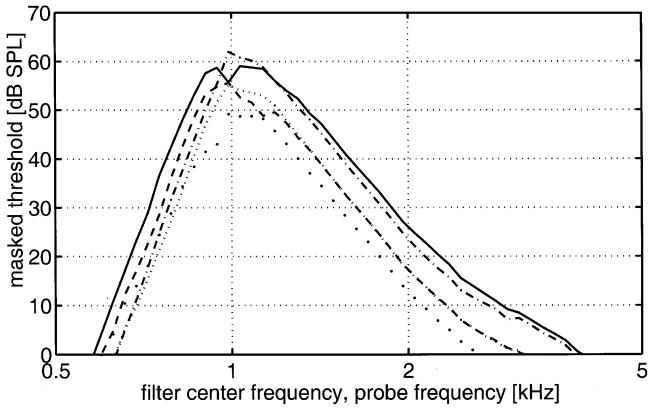

Fig. 14. Simulated masked threshold for 160-Hz-wide 60 dB SPL Gaussian noise centered at 1 kHz. The threshold patterns were generated by the model in Fig. 12 at four randomly selected times.

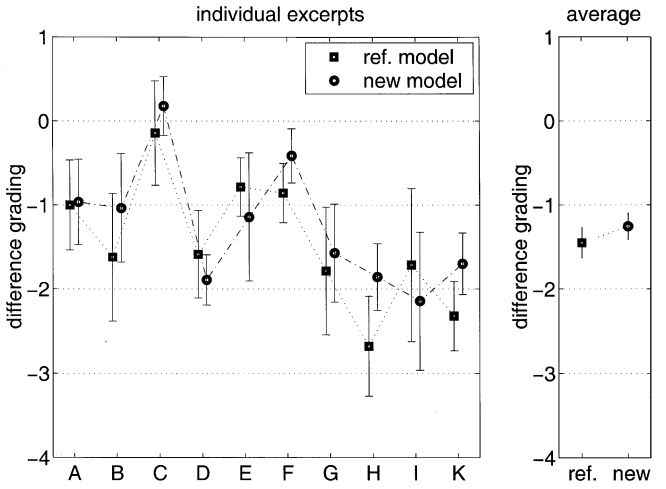

Fig. 15. Subjective difference gradings and 95%-confidence intervals of PAC with reference or new psychoacoustic model for seven subjects. Results for individual excerpts (left) and average over all excerpts (right).

The reference models of both coders are based on a uniform filter bank. The performance of each coder was compared for both, the reference and the new psychoacoustic model in terms of bit rate and audio quality. The aim was to achieve transparent quality of audio material with respect to its compact disc (CD) version downmixed to mono. The psychoacoustic model parameters were tuned individually for each coder to achieve the target quality at minimum bit rate for a small representative set of audio excerpts. These excerpts were not included in the evaluation.

The first coder is a traditional subband coder, the perceptual audio coder (PAC) [12]. The original psychoacoustic model of this coder is based on a complex uniform filter bank. The basic underlying algorithm of this model is similar to the model in [3] but it uses an improved tonality measure. PAC block-wise encodes successive windowed input samples. It uses one long or a group of eight short windows in case of transient input signals. A pre-echo control mechanism is applied to avoid audible distortions due to the temporal noise spreading within each window. A separate pre-echo control is not necessary with the new model since the minimum masked threshold from the high time-resolution model output is applied in each block.

The second coder used for a comparison of model performance is a prefilter-based coder [13]. In the coder configuration used here, the encoder comprises a time-varying prefilter cascaded with a subband coder. PAC [12] was applied as the subband coder using the same constant quantizer step size in all subbands with disabled psychoacoustic model. Thus, at the output of the subband decoder white quantization noise is superimposed to its input signal. The postfilter in the decoder shapes the noise according to the masked threshold such that it is just inaudible. The prefilter is inverse to the postfilter. It distorts the input audio signal in order to compensate for the postfilter effect on the audio signal.

The reference psychoacoustic model of the prefilter coder is based on ideas described in [14]. The spectral decomposition of that model is also based on a complex uniform filter bank. The cochlear-filter band energies are calculated by merging the corresponding subband energies. The resulting energies are dynamically compressed and spectrally spreaded. The masked threshold is calculated by uncompressing and applying a tonality measure to adjust the threshold level.

All coders were operated without rate control in order to ensure the availability of a sufficient number of bits for encoding. Thus, the quantization noise level does not exceed the masked threshold due to an insufficient number of bits. The resulting different bit rates for the audio excerpts are considered in the evaluation.

The coder performance in terms of audio quality was evaluated with an informal subjective listening test. The test design and procedure was based on ITU-R BS.1116 [15]. Most of the subjects were expert listeners. The excerpts were mono audio signals of about 10 s duration sampled at 44.1 kHz and presented via headphones (Stax) in a sound booth. The ten most critical excerpts for the listening test of each coder were blindly preselected by two subjects from a set of more than forty critical excerpts that were not used for coder tuning.

VI. Results

The experimental results for PAC with reference or new psychoacoustic model are summarized in Fig. 15 and Table I. A difference grading of zero means no perceptual difference between reference and coded signal. Smaller difference gradings correspond to increasing degradations with respect to the unprocessed reference signal. The overall performance with the new model is slightly but not statistically significantly better (confidence intervals overlap; see Fig. 15 right). The gradings for the individual excerpts are not significantly different either. However, significant differences could have been detected from a test with more subjects since the confidence intervals usually get smaller with an increasing number of subjects. The average bit rate for PAC with new psychoacoustic model is 9% less than the reference PAC (see Table I). In summary, the new model in PAC achieves the same quality as the reference at a lower bit rate.

The experimental results for the prefilter-based coder with reference or new psychoacoustic model are summarized in

TABLE I
TEST ITEMS AND BIT RATES FOR PAC WITH REFERENCE OR NEW PSYCHOACOUSTIC MODEL. THE RIGHT-MOST COLUMN SHOWS THE RATIO OF BIT-RATES FROM THE NEW AND REFERENCE MODEL

excerpt	source	reference [bit/sample]	new/ref. [%]
A	fire alarm bell	1.291	100
B	hit on glass	0.616	127
C	table tennis	1.445	96
D	clarinet	1.157	82
E	male speech	1.976	95
F	oboe	1.153	86
G	bag pipes	0.934	101
H	glockenspiel	1.105	76
I	male vocal	1.601	71
K	triangle	1.242	79
	average:	**1.252**	**91**

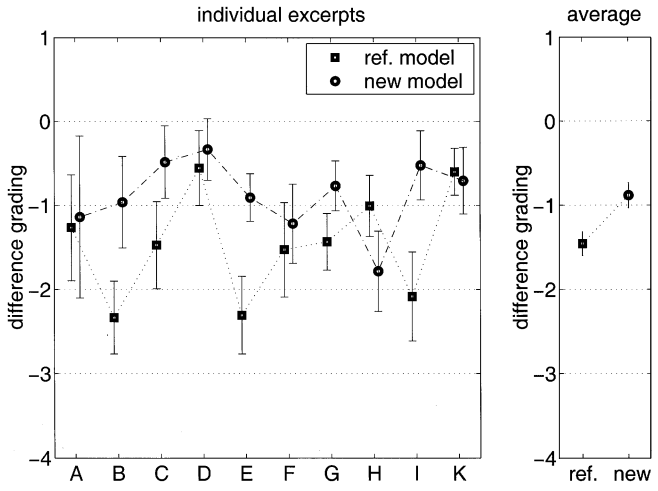

Fig. 16. Difference gradings and 95%-confidence intervals of the prefilter-based coder with reference or new psychoacoustic model for nine subjects. Results for individual excerpts (left) and average over all excerpts (right).

Fig. 16 and Table II. The average difference gradings for all excerpts (Fig. 16 right) show a significant quality improvement for the new psychoacoustic model. 50% of the individual excerpts are significantly better (nonoverlapping confidence intervals). The average bit rate for all excerpts for the coder with the new model is 7% less than the reference (see Table II). Thus, the new model improves the performance of the prefilter-based coder in terms of bit rate and audio quality.

Results from both experiments suggest that the new psychoacoustic model is able to improve coders independent of the type of core technology used. A performance comparison between the prefilter-based coder and PAC based on these results is not relevant and meaningful since the quality scales as applied by the subjects in the two listening tests was probably significantly different.

VII. CONCLUSIONS

A psychoacoustic model is proposed that overcomes the mismatch of the spectral analysis of traditional models based on a uniform decomposition and the properties of the human auditory system. This is achieved by introducing an efficient cochlear filter bank that closely approximates cochlear time-frequency resolution. In contrast to the uniform transform used in traditional models, the filter bank achieves a phase response in better agreement with human cochlear filters and preserves the phase-related interaction of frequency components in each band. The postprocessing for the masked threshold estimation is less complex than in traditional models. The generated masked thresholds appear to be more accurate since the performance of state-of-the-art coders increases with the new model. The new model is a good candidate for low-delay coders, since the delay is dominated by the cochlear filter bank that has the low delay of minimum-phase IIR filters with a maximum of 10 ms at very low center frequencies.

TABLE II
TEST ITEMS AND BIT RATES FOR PREFILTER-BASED CODER WITH REFERENCE OR NEW PSYCHOACOUSTIC MODEL. (BIT RATES DO NOT INCLUDE SIDE INFORMATION FOR POSTFILTER ADAPTATION.) THE RIGHT-MOST COLUMN SHOWS THE RATIO OF BIT-RATES FROM THE NEW AND REFERENCE MODEL

excerpt	source	reference [bit/sample]	new/ref. [%]
A	hit on glass	0.946	109
B	clarinet	1.394	87
C	female speech	1.772	90
D	child's speech	1.757	116
E	oboe	1.304	96
F	pitch pipe	1.615	85
G	bag pipes	1.321	107
H	male vocal	1.596	65
I	triangle	1.454	83
K	female vocal	1.962	95
	average:	**1.512**	**93**

The cochlear filter bank used in the psychoacoustic model covers a range of center frequencies from 20 Hz to 20 kHz by 103 filter bands. The low-pass and high-pass filters of the filter bank are realized as IIR filters. Compared to FIR filters, the low-order IIR filters lead to lower complexity, a reduced group delay, and a phase response better matched with the auditory system.

The filter bank can be adapted to applications that require frequency responses different from the example above. This flexibility also permits different frequency spacings or bandwidths, e.g., according to a Bark or ERB scale [8], [16] by defining the appropriate desired frequency response $H(f)$ for each filter band. Thus, the proposed filter-bank structure provides a flexible framework for approximating the auditory time and frequency resolution in different applications.

ACKNOWLEDGMENT

The authors would like to thank C. Faller for supporting the implementations into PAC. B. Edler provided the prefilter-based coder and support. Many colleagues provided time and patience as participants in the listening tests. C. Faller, O.

Ghitza, P. Kroon, Y. Shoham, and two anonymous reviewers provided helpful comments on earlier versions of this paper.

REFERENCES

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

[2] J. L. Hall, "Asymmetry of masking revisited: Generalization of masker and probe bandwidth," *J. Acoust. Soc. Amer.*, vol. 101, no. 2, pp. 1023–1033, 1997.

[3] ISO/IEC JTC1/SC29/WG11, *Coding of Moving Pictures and Audio—MPEG-2 Advanced Audio Coding*, ISO/IEC 13 818-7 Int. Std., Geneva, Switzerland, 1997.

[4] K. Brandenburg and G. Stoll *et al.*, "ISO-MPEG-1 Audio: A generic standard for coding of high-quality digital audio," *J. AES*, vol. 42, no. 10, pp. 780–791, Oct. 1994.

[5] F. Baumgarte, "A physiological ear model for auditory masking applicable to perceptual coding," in *103rd AES Conv.*, New York, 1997. Preprint 4511. [Online]. Available: http://www.uni-hannover.de/~baumgarte/publications.html.

[6] ——, "Ein physiologisches Gehörmodell zur Nachbildung von Wahrnehmungsschwellen für die Audiocodierung," Dissertation (in German), University of Hannover, Germany, 2000. [Online]. Available: from http://www.uni-hannover.de/~baumgarte/publications.html.

[7] E. Zwicker and H. Fastl, *Psychoacoustics*, 2nd ed. New York: Springer, 1999.

[8] B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, pp. 103–138, 1990.

[9] J. E. Dennis, Jr. and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Englewood Cliffs, NJ: Prentice-Hall, 1983.

[10] M. A. Ruggero, N. C. Rich, S. S. Narayan, A. Recio, and L. Robles, "Basilar-membrane responses to tones at the base of the chinchilla cochlea," *J. Acoust. Soc. Amer.*, vol. 101, no. 4, pp. 2151–2163, 1997.

[11] P. X. Joris and T. C. T. Yin, "Response to amplitude-modulated tones in the auditory nerve of the cat," *J. Acoust. Soc. Amer.*, vol. 91, no. 1, pp. 215–232, 1992.

[12] D. Sinha, J. D. Johnston, S. Dorward, and S. R. Quackenbush, "The Perceptual Audio Coder (PAC)," in *The Digital Signal Processing Handbook*, V. Madisetti and D. B. Williams, Eds. New York: IEEE Press, 1998, pp. 42-1–42-18.

[13] B. Edler and G. Schuller, "Audio coding using a psychoacoustic pre- and postfilter," in *Proc. ICASSP 2000*, June 2000, pp. 1881–1884.

[14] F. Baumgarte, "A nonlinear psychoacoustic model applied to the ISO MPEG Layer 3 coder," in *99th AES Conv.*, 1995. Preprint 4087. [Online]. Available: from http://www.uni-hannover.de/~baumgarte/publications.html.

[15] ITU-R, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," in *Rec. ITU-R BS.1116-1* Geneva, Switzerland, 1997.

[16] B. C. J. Moore, *An Introduction to the Psychology of Hearing*. San Diego, CA: Academic, 1997.

Frank Baumgarte received the M.S. and Ph.D. (Dr.-Ing.) degrees in electrical engineering from the University of Hannover, Germany, in 1989 and 2000, respectively.

During the studies and as independent consultant he implemented real-time signal processing algorithms on a variety of DSPs including a speech coder and an MPEG-1 Layer 3 decoder. His dissertation includes a nonlinear physiological auditory model for application in audio coding. In 1999, he joined the Acoustics and Speech Research Department, Bell Labs, Lucent Technologies, Murray Hill, NJ, where he was engaged in objective quality assessment and psychoacoustic modeling for audio coding. He became a Member of Technical Staff of the Media Signal Processing Research Department, Agere Systems, Berkeley Heights, NJ, a Lucent spinoff, in 2001, focusing on advanced perceptual models for multichannel audio coding and spatial hearing. His main research interests in the area of acoustic communication include the understanding and modeling of the human auditory system physiology, psychophysics, audio and speech coding, and quality assessment.