A COMPUTATIONALLY EFFICIENT COCHLEAR FILTER BANK FOR PERCEPTUAL AUDIO CODING

Frank Baumgarte

Multimedia Communications Research Laboratory
Bell Labs - Lucent Technologies, Murray Hill, NJ, U.S.A.

ABSTRACT

Many applications in auditory modeling require analysis filters that approximate the frequency selectivity given by psychophysical data, e.g. from masking experiments using narrow-band maskers. This frequency selectivity is largely determined by the spectral decomposition process inside the human cochlea. Currently used spectral decomposition schemes for masking modeling in audio coding generally do not achieve the non-uniform time and frequency resolution provided by the cochlea. These applications rather take advantage of the computational efficiency of uniform filter banks or transforms at the expense of coding gain.

This paper presents a suitable analysis filter-bank structure employing cascaded low-order IIR filters and appropriate down-sampling to increase efficiency. In an application example, the filter responses were optimized to model auditory masking effects. The results show that the time and frequency resolution of the filter bank matches or exceeds the masking properties. Thus, the filter bank enables improved masking modeling for audio coding at low computational costs.

1. INTRODUCTION

Human auditory frequency selectivity is largely determined by signal processing in the cochlea. The cochlea provides band-pass filtered versions of the input signal that are subsequently transduced into neural signals by the inner hair cells. The associated band-pass filters have increasing bandwidth with increasing center frequency and an asymmetric frequency response.

In perceptual audio coding, the audio signal is treated as a masker for distortions introduced by lossy data compression. For this purpose, the masked threshold is approximated by a perceptual model. Existing perceptual models, e.g. [1], employ an FFT-based transform to derive a spectral decomposition of the acoustic signal as first processing step. The non-uniform spectral resolution of the auditory system is taken into account by summing up the energies of the appropriate number of neighboring FFT bands. The phase relation between spectral components within an auditory filter band is not taken into account by the summation of energies. The temporal resolution of the spectral decomposition is determined by the transform size and is thus constant across all auditory bands. This results in a significantly lower temporal resolution at high center frequencies in comparison with the corresponding auditory filters. These deviations lead to inaccurate modeling of masking and suboptimal coding gain.

A higher temporal resolution is achieved by the non-uniform filter bank in the "Advanced Version" of the audio quality measurement standard [2]. Each of those 40 critical-band filter pairs is realized as FIR filter. The output of each filter pair is a critical-band signal and its (90 degrees phase-shifted) Hilbert transform which is down-sampled by a factor of 32. The appropriate auditory filter slopes are created by spectral convolution with a spreading function. This complex convolution increases the temporal resolution of the original filters, but the filter bank is computationally complex and the linear phase response is not in line with the auditory system. Furthermore, the down-sampling can create aliasing distortions in the high frequency bands.

In this paper, a novel filter-bank structure is proposed. This structure is suitable for achieving the time- and frequency resolution necessary to simulate psychophysical data closely related to cochlear spectral decomposition properties, and it overcomes the described drawbacks of known approaches. The filter-bank structure is outlined in Section 2. It consists of a cascade of low-order IIR filters. The cascade structure inherently supports sampling rate reduction due to the continuously decreasing cutoff frequency in the cascade.

In Section 3, an example is given in which the filter-bank coefficients are optimized for modeling of masked threshold patterns of narrow-band maskers. The generated thresholds are applied to perceptual audio coding.

Results and conclusions from this study are given in Section 4.

2. FILTER-BANK STRUCTURE

The peripheral auditory system performs spectral analysis of the input acoustic signal with spectrally highly overlapping band-pass filters. The non-uniform frequency resolution and bandwidths of these filters is approximated in the proposed structure by cascaded IIR filters. Figure 1 shows the proposed filter bank structure with low-pass (LPF) and high-pass (HPF) filters. 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. Thus, the output of each HPF has a band-pass 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.

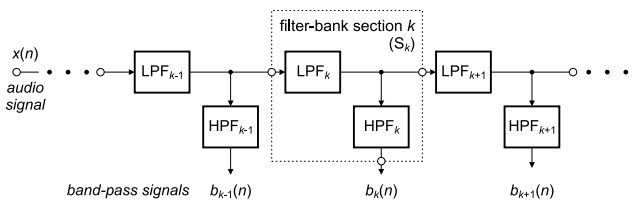

Figure 1: Block diagram of filter-bank structure.

The decreasing cutoff frequency of the LPF cascade permits a reduction of sampling rate, which 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 all cascaded filter-bank sections with equal sampling rate. The rate reduction by a factor of two is achieved by leaving out every second 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 frequency in that stage to limit aliasing.

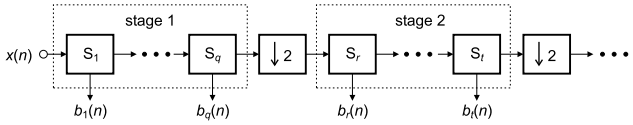

Figure 2: Down-sampling scheme of filter-bank.

The filter orders of all HPFs as well as the orders of all LPFs are the same. The LPF and HPF order can be chosen independently and should be large enough to accurately model the spectral decomposition features found in relevant psychophysical data. The example below shows that an LPF order of 2 and an HPF order of 4 is sufficient to model masking. After the orders are fixed, the filter coefficients can be determined by an optimization algorithm, which minimizes an error function of the responses of the desired filters and the proposed filter bank. The responses of the desired filters are generally derived from psychophysical measurements.

3. APPLICATION EXAMPLE

This section outlines the performance of the filter bank for an application example. The desired magnitude frequency responses of the filters are derived from psychophysical masking data. The filter orders of the HPFs and LPFs determine the achievable accuracy of the desired frequency-response approximation. They were chosen as low as possible to minimize computational complexity.

A simplified block diagram of the masked threshold model is given in Fig. 3. It is based on a psychophysiological model described in [3]. The cochlear filters of that model are replaced by the proposed filter bank. The input acoustic signal is processed by an outer- and middle-ear (OME) filter, which approximates the filter characteristic of these parts of the auditory system. The output signal is spectrally decomposed by the filter bank, which approximates the frequency-dependent spread of masking. The envelope of each band-pass signal is approximated by rectification and low-pass filtering. The amount of envelope fluctuation is estimated and used to adjust the masked threshold level by subtracting a fluctuation-dependent offset from the envelope level. For high fluctuations the masked threshold is assumed to have a higher level than for low fluctuations at the same envelope level. This property is related to the asymmetry of masking [4], which other models take into account by a tonality estimation. Temporal smearing is applied to the offset-adjusted thresholds in order to take properties of temporal masking, e.g. pre- and post-masking, into account. The smearing is motivated by the fact that temporal masking is mainly created in the auditory system after cochlear filtering.

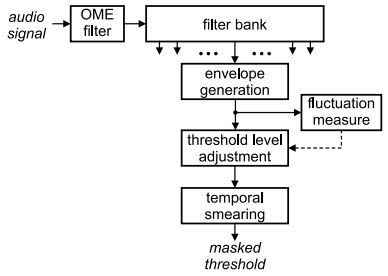

Figure 3: Block diagram of masked threshold model.

The aim of the model is to derive the masked threshold level at the output of each channel for an assumed probe at the center frequency of that channel. The desired frequency responses of the filter bank are derived from masking patterns of narrow-band noise maskers. For this type of masker, the envelope fluctuation at the filter outputs is assumed to be at the upper bound. Due to the stationary masker, temporal masking effects can be neglected and the output masked threshold of the model depends mainly on the filter-bank and OME-filter characteristic.

Due to the asymmetric frequency spread of masking, a probe at a higher frequency than the masker frequency is exposed to a larger masking effect than a probe at a lower frequency. This asymmetry can be modeled by a filter that produces more attenuation for a masker above the center frequency than for a masker below the center frequency. Thus, the band-pass filter slopes should be asymmetrical with a more shallow slope towards lower frequencies. In simple masking models, which are adopted here, masking patterns are often described by two constant slopes on a level vs. Bark scale. These slopes are chosen to be 8 dB/Bark and -25 dB/Bark. For simplicity, the Bark scale is approximated by a logarithmic frequency scale. This approximation is in good agreement with psychophysical data for frequencies above 1 kHz.

3.1. Desired Frequency Responses

The desired filter-bank center frequencies are uniformly distributed on a logarithmic scale, covering the full range of audible frequencies. The spacing is a quarter of a critical band and the critical-band width is assumed to be equal to 20% of the center frequency. Thus, the filter center frequency $f_c(k)$ of channel k is related to channel $k-1$ by (1). The desired magnitude frequency response $|H(f)|$ of one channel with the cutoff at f_c is defined in (2).

$$f_c(k) = 1.2^{-\frac{1}{4}} f_c(k-1) \tag{1}$$

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

with

$$S_{\mathrm{LP}} = \frac{-25}{20\log_{10}\left(\frac{1}{1.2}\right)} \; ; \quad S_{\mathrm{HP}} = \frac{-8}{20\log_{10}\left(\frac{1}{1.2}\right)}$$

$$q = 4 \; ; \quad j = \sqrt{-1}$$

The first term in (2) describes the steep filter slope towards 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}. The transition between the two slopes is controlled by a resonance quality factor q.

3.2. Filter-Bank Response and Model Output

In order to minimize computational complexity, the LPFs and HPFs are realized as IIR filters. Additional advantages of IIR over FIR filters consist of the reduced group delay and a phase response better matched with the auditory system. Given the desired frequency responses, their filter coefficients can be optimized using standard techniques, e.g. the damped Gauss-Newton method for iterative search [5] available in MATLAB$^{\mathrm{TM}}$. A reasonably good approximation of the desired responses is already achieved by an HPF order of 4 and an LPF order of 2. Figure 4 shows the desired and the resulting magnitude frequency response of the filter at 1002 Hz center frequency. Near the center frequency f_c, the deviation is small. At low frequencies, the deviation reaches about 10 dB at 100 Hz. However, due to the high damping in this frequency range far from the center frequency, this deviation is considered to have only minor effects for applications in audio coding. The distribution of the approximation error can be controled by using a frequency-dependent weighting function for the error in the optimization algorithm.

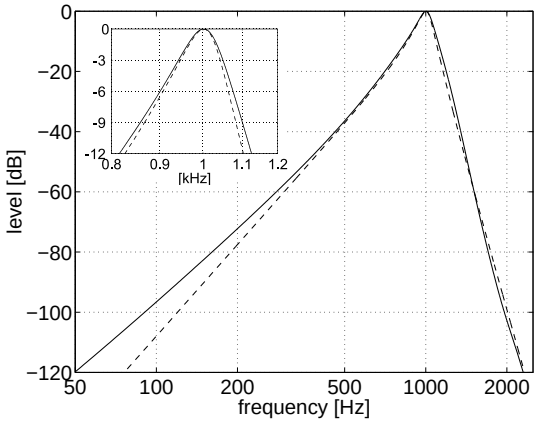

Figure 4: Desired (dashed) and achieved (solid) magnitude response of the filter-bank channel at $f_c = 1002$ Hz. The inset shows in detail the response near the center frequency. The input audio sampling frequency is 44.1 kHz.

Figure 5 shows the resulting filter-bank responses of stage 2. The frequency scale is normalized by half the sampling frequency of that stage. The responses have basically the same shape on a logarithmic scale. They are shifted according to their center frequency and are highly overlapping.

The phase responses of the filter-bank channel in Fig. 4 and its neighbors are shown in Fig. 6. These phase responses are determined by the minimum-phase design of all LPFs and HPFs, which was chosen in accordance with known models of cochlear hydromechanics. Thus, the phase qualitatively agrees with measurements of basilar membrane motion in the cochlea [6].

Figure 7 shows the location of the LPF poles and zeros in stage 2. Due to their distance from the unit circle, implementation problems caused by limited arithmetic precision are unlikely.

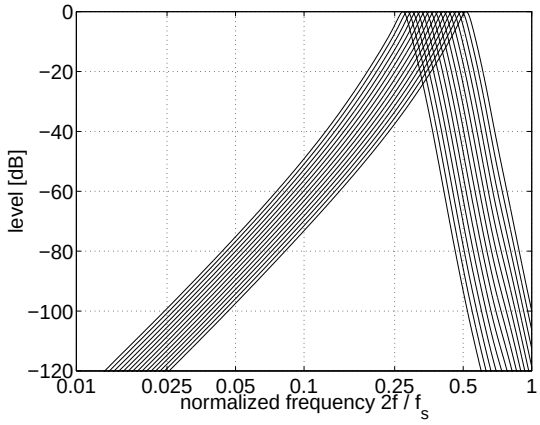

Figure 5: Magnitude frequency responses of the filter-bank channels in stage 2.

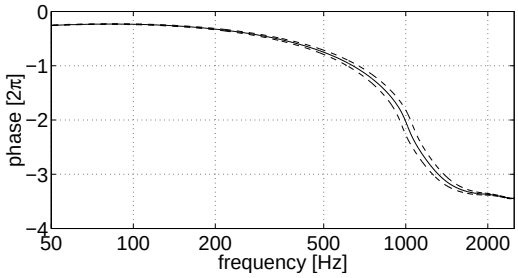

Figure 6: Phase responses of the filter-bank channel at $f_c = 1002$ Hz and neighboring channels.

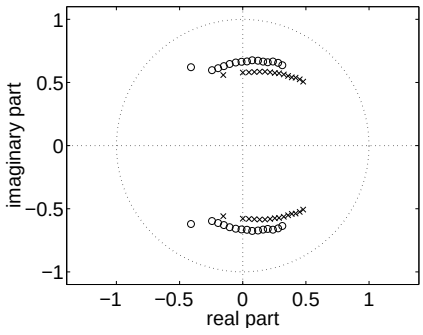

Figure 7: Pole-zero plot of the LPF cascade in stage 2 ($\circ$ Zero, $\times$ Pole).

Figure 8 shows the logarithm of the impulse response envelope for a filter center frequency of 1002 Hz. The modeling of temporal masking requires that the temporal spread of a filter which is reflected by its impulse response does not exceed the limits of pre- and post-masking. 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. Post-masking can last for about 200 ms after a masker is switched off [7]. Since the temporal filter response

shows a damping 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. It is assumed here that premasking duration increases at lower frequencies as well, so that the premasking duration is not exceeded.

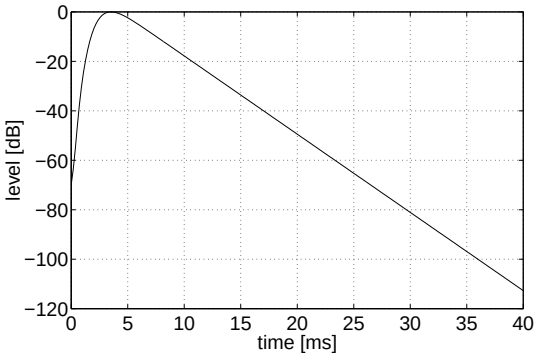

Figure 8: Envelope of impulse response of the filter-bank channel at $f_c = 1002$ Hz.

Preliminary results from the model shown in Fig. 3 for the masked threshold of a 160 Hz-wide Gaussian noise masker centered at 1 kHz are outlined in Fig. 9. 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 channel is assigned to the channel center frequency. E.g., a probe signal at a channel center frequency is assumed to be inaudible, if its level is below the calculated masked threshold.

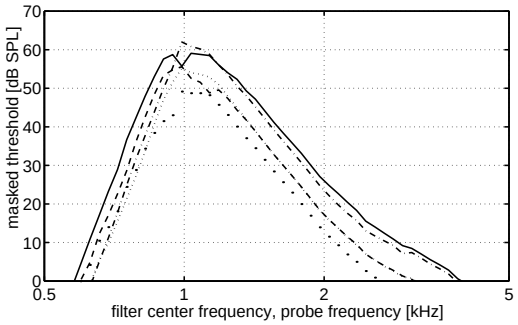

Figure 9: Simulated masked threshold for 160-Hz wide 60 dB SPL Gaussian noise centered at 1 kHz. The threshold patterns were generated by the model in Fig. 3 at four randomly selected times

4. RESULTS AND CONCLUSIONS

The model of Fig. 3 was applied to a pre-filter-based audio coder [8]. This coder is designed to support quantization noise shaping with non-uniform spectral resolution according to the auditory system. Coding results with the aim of transparency were subjectively compared with the reference coder [8], which is controlled by a perceptual model based on a uniform spectral decomposition.

An informal subjective assessment of coded signals indicates an overall improved quality and a significantly higher quality for the most critical speech and music material at the same average bitrate. This result suggests a superior performance of the perceptual model with the proposed filter bank.

In the application example, the proposed filter bank needs a total of 517 multiply-accumulate instructions for the processing of one input sample at a sampling rate of 44.1 kHz. A range of center frequencies from 20 Hz to 20 kHz is covered by 150 filter channels or sections. This is in contrast to the filter bank in [2], which has 265% the complexity and only 27% the number of channels.

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 [9], by defining the appropriate desired frequency response $H(f)$ for each filter channel. Thus the proposed filter-bank structure provides a flexible framework for approximating the auditory time- and frequency resolution in different applications. In contrast to a uniform transform or FIR filters, it achieves a phase response in better agreement with cochlear filters and preserves the phase-related interaction of frequency components in each critical band. It has significantly less computational complexity than the filter bank in [2].

5. REFERENCES

[1] ISO/IEC JTC1/SC29/WG11 “Coding of moving pictures and audio – MPEG-2 Advanced Audio Coding”. *ISO/IEC 13818-7 International Standard*, 1997.

[2] ITU-R “Method for objective measurement of perceived audio quality,” *Rec. ITU-R BS.1387*, Geneva, 1998.

[3] Baumgarte F. “Evaluation of a Physiological Ear Model Considering Masking Effects Relevant to Audio Coding”. *105th AES Convention*, San Francisco, CA, September 1998, Preprint 4789.

[4] Hall J.L. “Asymmetry of masking revisited: Generalization of masker and probe bandwidth”. *J. Acoust. Soc. Am.*, 101(2), 1023–1033, 1997.

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

[6] Ruggero M.A., Rich N.C., Narayan S.S., Recio A., and Robles L. “Basilar-membrane responses to tones at the base of the chinchilla cochlea”. *J. Acoust. Soc. Am.*, 101(4), 2151–2163, 1997.

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

[8] Edler B. and Schuller G. “Audio Coding using a psychoacoustic pre- and postfilter”. *Proc. ICASSP*, Istanbul, 1881–1884, June 2000.

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