A COMPUTATIONALLY EFFICIENT ALGORITHM FOR CALCULATING LOUDNESS PATTERNS OF NARROWBAND SPEECH

Markus Hauenstein

University of Kiel, Institute for Network and System Theory, D-24143 Kiel, Germany
mh@techfak.uni-kiel.de

ABSTRACT

Loudness patterns are closer to the human perception of sound waves than spectrograms. This paper describes how loudness patterns can be efficiently calculated with an allpass-transformed polyphase-filterbank based on a mixed radix FFT and three subsequent non-linear stages that model masking effects in the frequency and time domain as well as loudness compression.

1. INTRODUCTION

We can learn by psychoacoustic experiments that our ear and the subsequent sound processor, the human brain, do not view a sound event as a simple superposition of sinusoids. For example, a sine may be heard or not depending on the additional sounds that surround it in the frequency or time domain. These effects are known as frequency or time *masking*. Other experiments show that spectral components are packed together while being analyzed, this leads to the concept of *critical bands*. The bandwidth of a critical band increases with frequency. Furthermore, we know that our auditory system performs a *loudness compression*: For sines at medium to high levels we find a law stating that a sine with a tenfold intensity is only heard twice as loud.

2. OUTER TO INNER EAR FILTER

Before sound waves are analyzed by the nerve cells in the cochlea, they have to pass the outer and middle ear. We can model this transfer with a linear time-invariant filter. The filter was designed to approximate an analytical expression for the outer to inner ear attenuation function a_0 found in [4]. A fourth-order IIR filter can give a very good approximation.

3. CRITICAL-BAND FILTERBANK

3.1. Polyphase Filterbank

We start with a linear-phase prototype lowpass impulse response $h_0(k)$ of finite even length M:

$$h_0(k) = \begin{cases} h_0(M-1-k) & : \quad k = 0 \ldots M-1 \\ 0 & : \quad \text{elsewhere} \end{cases}$$

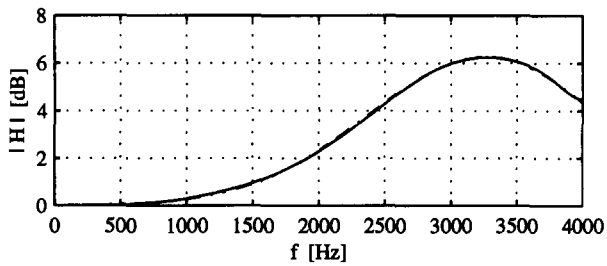

Figure 1: *Outer to inner ear transfer function. Solid: 4th order IIR-filter, dashed: analytical expression very close to measured data*

We can now modulate this lowpass function with two complex harmonic exponentials, i.e. a cosine function:

$$h_{BP,\mu}(k) = 2\, h_0(k) \cdot \cos\left[\frac{2\pi}{M}\left(\mu+\frac{1}{2}\right)\left(k-\frac{M-1}{2}\right)\right]$$

For $\mu = 0 \ldots M/2-1$ we can derive $M/2$ different real bandpass functions $h_{BP,\mu}(k)$ thus constructing a filterbank with $M/2$ channels. The filterbank output signals $y_\mu(k)$ can then be calculated as:

$$y_\mu(k) = \Re\left\{ 2\, \exp\left[j\,\pi\,\frac{M-1}{M}\left(\mu+\frac{1}{2}\right)\right] \cdot \sum_{m=0}^{M-1}\left[h_0(m)\, \exp\left(-j\,\frac{\pi}{M}m\right)\, v(k-M+1+m)\right] \cdot \exp\left(-j\,\frac{2\pi}{M}\mu m\right)\right\}$$

Basically, this is just the Discrete Fourier Transformation (DFT) of the windowed last M samples of the input signal $v(k)$. These samples are weighted by a modulated version

$$h_{mod}(k) = h_0(k)\, \exp\left(-j\,\frac{\pi}{M}k\right)$$

of the prototype-filter impulse response $h_0(k)$. The DFT output has to be multiplied with the factors

$$2\, \exp\left[j\,\pi\,\frac{M-1}{M}\left(\mu+\frac{1}{2}\right)\right]$$

and the real part must be extracted. This filterbank is closely related to a Discrete Cosine Transformation (DCT, see for example [7]).

If we are only interested in every Kth of the $M/2$ filterbank channels (i.e. $L/2$ in total), we can write with $\mu = K \cdot \nu$, $\nu = 0 \ldots L/2 - 1$ and $K \cdot L = M$:

$$y_{K\nu}(k) = \Re \left\{ 2 \exp\left[j\pi \frac{M-1}{M} \left(K\nu + \frac{1}{2} \right) \right] \cdot \sum_{l=0}^{L-1} w(k,l) \exp\left(-j \frac{2\pi}{L} \nu l \right) \right\}$$

with

$$w(k,l) = \sum_{\kappa=0}^{K-1} h_{mod}(l + \kappa L)\, v(k - M + 1 + l + \kappa L)$$

and $l = 0 \ldots L-1$. In this case a DFT of size L is sufficient. If $h_0(k)$ is properly designed (for example with Parks' and McClellan's algorithm [5]), we can smoothly cover the whole frequency range from $\Omega = 0$ to $\Omega = 2\pi$ with this so called *polyphase filterbank* consisting of L/2 FIR-filters. $w(k,l)$ describes the output of the *polyphase network*. The network output is fed into a DFT that can be efficiently computed with FFT algorithms. The size of the FFT is only L although the length of the filter impulse responses is $M = KL$. A downsampling of the filterbank output is allowed, and because of the non-recursive structure of the filterbank we only have to calculate output samples at the lower sampling rate thus reducing the computational load according to the downsampling factor [8].

3.2. Allpass Transformation

Our ear does not analyze sound events in evenly spaced frequency bands. Laws that relate the auditory filter bandwidth (critical bandwidth or ERB) to frequency were analytically formulated by Zwicker [1] and Moore and Glasberg [2]. Zwicker finds 18 and Moore and Glasberg find 27 just not overlapping filters in the frequency range from 0 to 4000 Hz. In both cases we find a non-linear but monotonic warping of the frequency axis to a Bark or ERB scale. If we submit our filter functions to an allpass transformation, we should be able to model this frequency scale warping. We content ourselves with a real first order allpass defined by its transfer function

$$H_A(z) = \frac{-\alpha + z^{-1}}{1 - \alpha z^{-1}}$$

If we choose for example $\alpha = 0.4$ (Zwicker) or $\alpha = 0.55$ (Moore and Glasberg) we can sufficiently approximate the frequency warping characteristic of the human auditory system (see Figure 2). The cascade of $M-1$ delays that was needed for the storage of the last $M-1$ input samples is now replaced by a cascade of $M-1$ allpass filters. Thus the original FIR-filterbank is transformed to an IIR-filterbank. The phase characteristic is no longer linear and the group delay of the filters is now a function of frequency:

$$\tau_{gr}(\Omega) = \frac{M-1}{2} \frac{1-\alpha^2}{1 - 2\alpha\cos(\Omega) + \alpha^2}$$

We get a different (mean) time delay in each channel which must be balanced for the correct calculation of frequency

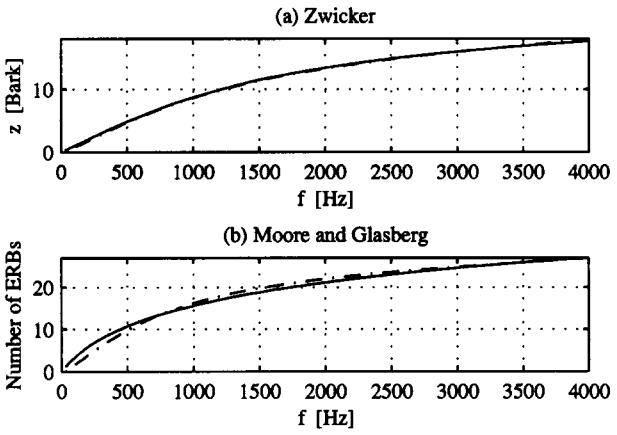

Figure 2: *Frequency warping of the auditory system (solid: exact formula, dashed: allpass approximation)*

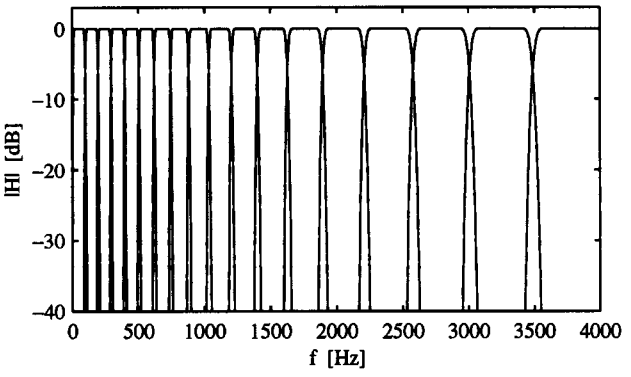

Figure 3: *Warped polyphase filterbank with 18 channels*

masking. The downsampling can no longer be performed in that easy way described above. Since the recursive structure of the transformed filterbank is limited to the cascade of allpass filters, we only need to calculate the cascade at the original sampling rate, and the subsequent stages like the FFT can be calculated at the lower sampling rate.

3.3. Mixed-Radix FFT

Most commonly used FFTs are only capable of transforming input vectors when the length is a power of two (Radix-2-FFT). To cover the whole narrowband frequency axis up to 4 kHz we need 18 filters in the Zwicker case and 27 filters if we follow Moore and Glasberg. This would require a DFT of length 36 or 54 respectively. For that reason we cannot build up a psychoacoustically correct filterbank with a common Radix-2-FFT. A prime factor decomposition yields $36 = 2 \cdot 2 \cdot 3 \cdot 3$ and $54 = 2 \cdot 3 \cdot 3 \cdot 3$. In both cases we only have the prime factors 2 and 3. An FFT based on these two radices increases thus the computation speed. The gain in computation speed is even higher if we try to calculate the critical band powers of more than 18 or 27 points along the basilar membrane. For example a quadrupling of these numbers would result in 72 or 108

filters which now are overlapping. We would have to calculate DFTs of length 144 or 216 which could also be performed with a combined Radix-2-Radix-3-FFT.

3.4. Power Estimation

The critical-band powers $P_\nu(k)$ have to be estimated. This task can be accomplished by squaring and subsequent lowpass filtering of the filterbank output signals $y_{K\nu}(k)$. A simple recursive first-order lowpass possessing the transfer function

$$H_{LP,\nu}(z) = \frac{1-\beta(\nu)}{1-\beta(\nu)\,z^{-1}}$$

is suitable for the filtering of $y^2_{K\nu}(k)$ and not wasting to much computation time. The integration time is controlled by the parameter $\beta(\nu)$ which should be a function of frequency and thus channel number $\nu = 0 \ldots L/2-1$ to get an appropriate time-frequency resolution.

4. EXCITATION

The excitation $E_\nu(k)$ of the auditory nerves is calculated by a frequency smearing of the critical band powers $P_\nu(k)$. Masking curves mirror the excitation produced by a masking sound. The slopes of the masking curves of narrowband noise have been analytically described by Terhardt [3]; these curves show how power is spread in the inner ear. The ascending slope is constant on the Bark scale: $S_1 =$ 27 dB/Bark. The descending slope depends both on sound pressure level SPL and frequency f:

$$S_2 = \left[24 + 0.23\left(\frac{f}{\text{kHz}}\right)^{-1} - 0.2\,\frac{SPL}{\text{dB}}\right]\frac{\text{dB}}{\text{Bark}}$$

Each critical band power $P_\nu(k)$ is treated separately and smeared with these slopes, so we get a total of $L/2$ interim excitation vectors of length $L/2$. The combined excitation at one specific point (described by the index ν) on the basilar membrane is here defined as the maximum of all interim excitations that stimulate that point. Thereby no computationally expensive add-up rules have to be obeyed. Since the slopes are linear on a logarithmic scale we can stepwise calculate the smeared power (interim excitation) evoked by one channel with a repeated multiplication by a constant factor. This factor remains constant and channel independent for the smearing toward lower channels and depends on frequency (channel number) and channel power for upward direction.

5. SPECIFIC LOUDNESS

The law derived by Zwicker for the transformation of excitation to specific loudness is in principle a power law with exponent 0.23:

$$N'_\nu(k) = 0.08\left(\frac{E_{TQ}(\nu)}{E_0}\right)^{0.23}\left[\left(\frac{1}{2}+\frac{1}{2}\frac{E_\nu(k)}{E_{TQ}(\nu)}\right)^{0.23}-1\right]$$

$N'_\nu(k)$ denotes the specific loudness, $E_\nu(k)$ the excitation, and E_0 is the excitation corresponding to the intensity normalization value $I_0 = 10^{-12} W/m^2$. $E_{TQ}(\nu)$ describes the excitation at the threshold of hearing. The influence of $E_{TQ}(\nu)$ disappears when only medium or higher levels are considered. In this case the following simplification holds:

$$N'_\nu(k) \sim E_\nu(k)^{\,0.23}$$

Since the exponent 0.23 is very close to 0.25, it may be even allowed to substitute Zwicker's relatively complicated compression law by simply twice calculating a square root.

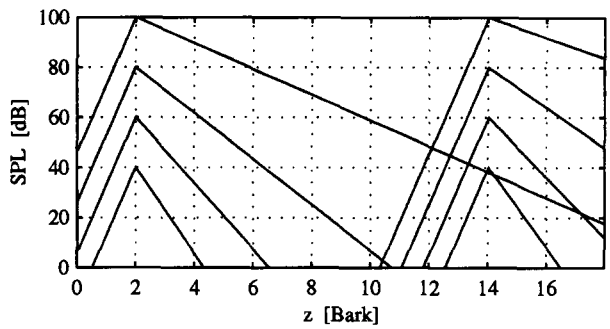

Figure 4: *Calculated excitation patterns for two sines at z=2 Bark and z=14 Bark and sound pressure levels = 40, 60, 80 and 100 dB*

6. TEMPORAL MASKING EFFECTS

Since premasking plays only a secondary role, we can restrict ourselves to the modeling of postmasking. Suppose a temporal masker of duration T_m ends at time t. Using continuous time notation, we can directly calculate a temporarily smeared specific loudness:

$$N'_\nu(t+\tau) = D(\tau, T_m)\cdot N'_\nu(t)$$

Kapust [4] has formulated an expression for $D(\tau, T_m)$ that fits well to experimental data given by Zwicker:

$$D(\tau, T_m) = 1 - \frac{1}{1.35}\arctan\left[\frac{\frac{\tau}{\text{ms}}}{13.2\left(\frac{T_m}{\text{ms}}\right)^{0.25}}\right]$$

An algorithm based on this formula (the arctangent may be substituted by a simpler third order polynomial) should estimate the masker duration T_m and find out if a temporarily smeared specific loudness component of prior time steps exceeds the specific loudness in the current time step. If the smeared component is larger, it replaces the physical specific loudness calculated for the current time step.

The method here proposed to determine the masker duration is to calculate the ratio between the mean of the specific loudness and its maximum in an interval of 200 ms (after approximately 200 ms the influence of postmasking vanishes [1]). This ratio multiplied by the interval length (200 ms) can be regarded as an effective masker duration.

Figure 5 shows the resulting specific loudness function when a test signal is fed into the implemented postmasking model. We can see how the decrease of the smeared specific loudness is dependent on the signal history.

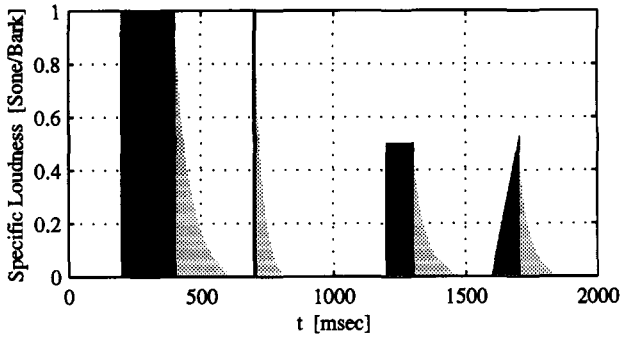

Figure 5: *Modeling postmasking*

7. APPLICATIONS

Figure 6 shows the loudness pattern of a speech sample which has been computed with an algorithm combining all the stages described in the previous sections (M=64*18, L=16*18, α = 0.4). This algorithm may enhance applications that rely on the computation of loudness patterns and can substitute conventional FFT-based methods. For instance speech recognizers and hearing aids could benefit from an efficient algorithm to calculate loudness patterns.

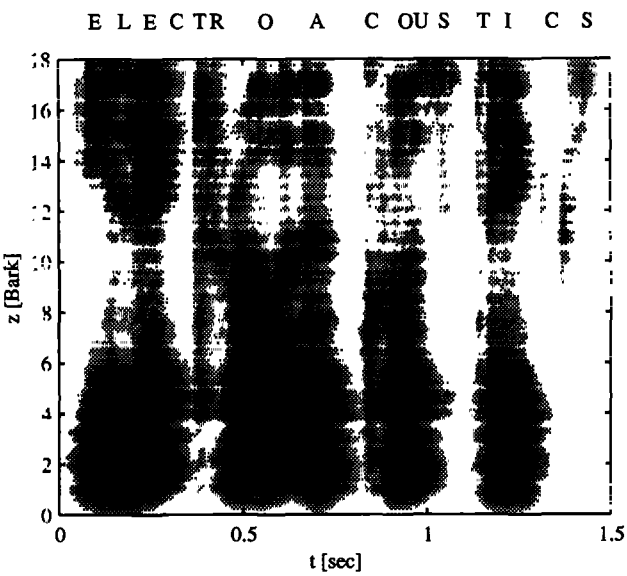

Figure 6: *Loudnes pattern of the word "electroacoustics". Dark (light) areas indicate high (small) specific loudness.*

The field of instrumental speech-quality assessment can also be cited as an example. Instrumental (objective) methods try to replace the expensive subjective codec tests by an instrument (usually a computer program) that is fed with the processed and the original speech material then calculating an objective score indicating the quality of the codec. An objective measure based on the comparison of loudness patterns was implemented, and the loudness patterns were calculated with the algorithm presented above. The measure was applied to a test that was conducted to characterize the subjective performance of the ITU-T 8 kbit/s codec (G.729). In Figure 7 the instrumental results are plotted versus the subjective scores (MOS). The correlation is very high.

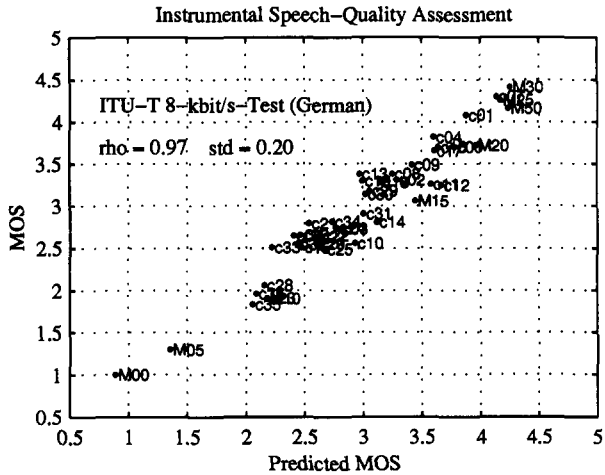

Figure 7: *Application of an instrumental speech-quality measure to a test of the ITU-T 8 kbit/s-codec*

8. IMPLEMENTATION

The algorithm was implemented in MATLAB. All time consuming parts have been coded as C-subroutines.

9. REFERENCES

[1] E. Zwicker and H. Fastl, *Psychoacoustics - Facts and Models*, Springer-Verlag Berlin Heidelberg, 1990.

[2] B.C.J. Moore, B.R.Glasberg, "A Revision of Zwicker's Loudness Model", *ACUSTICA / acta acustica*, Vol. 82, 1996.

[3] E. Terhardt: "Calculating Virtual Pitch", *Hearing Research*, 1, 1979, Elsevier / North Holland Biomedical Press.

[4] R. Kapust, *Qualitätsbeurteilung codierter Audiosignale mittels einer BARK-Transformation*, Dissertation, Universität Erlangen-Nürnberg, 1993.

[5] J. McClellan, T. Parks and L. Rabiner, "A computer program for designing optimum FIR linear phase digital filters", *IEEE Trans. Audio Electroacoust.*, Vol. AU-21, No. 6, pp. 506-526, 1973.

[6] M. Kappelan, B. Strauß and P. Vary, "Flexible Nonuniform Filter Banks using Allpass Transformation of Multiple Order", Proc. EUSIPCO-96, pp. 1745-1748, 1996.

[7] R. Gluth, *Beiträge zu Beschreibung und Realisierung digitaler, nichtrekursiver Filterbänke auf der Grundlage linearer diskreter Transformationen*, Dissertation, Ruhr-Universität Bochum, 1992.

[8] P.Vary, *Ein Beitrag zur Kurzzeitspektralanalyse mit digitalen Systemen*, Dissertation, Universität Erlangen-Nürnberg, 1978.