A Physiological Ear Model for Auditory Masking Applicable to Perceptual Coding

Frank Baumgarte[1], *AES Member*

Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung
Universität Hannover, Hannover, Germany

Abstract

The developed ear model covers an increased number of known masking effects compared to psychoacoustical models used so far in perceptual audio coders. It comprises a cochlear model which shows the same frequency selectivity and nonlinear behavior as observed from basilar membrane motion of the mammalian cochlea. The generated masked thresholds closely match classical psychoacoustics in terms of temporal and spectral masking.

0 INTRODUCTION

Data rate reduction techniques for high quality audio signals have rapidly advanced in the last years [1]. Sound quality comparable to the compact disk (CD) format is now achieved by the most advanced coding system at a data rate of 64 kbit/s for each channel of a stereo signal. This coding system was just standardized by ISO/IEC as MPEG–2 Advanced Audio Coding (AAC) [2][3]. At a data rate of 64 kbit/s some critical test signals like speech remain distinguishable from the original [4], so that for interesting applications which open up at this data rate a sound quality improvement is still desirable.

The basic structure common to perceptual audio coders for high sound quality – like the AAC – consists of a spectral analysis of the audio signal by a filterbank, as shown in Fig. 1. The resulting subband signals are quantized, coded and finally the bitstream is formatted. In parallel a psychoacoustical model generates an approximation of the masked threshold caused by the audio signal. The masked threshold are used to control the quantizers so that the quantizing distortions will be not audible. This simplified functional scheme is extended in case of the AAC–coder by several redundancy reduction tools like block companding, Huffman coding, subband signal prediction and techniques exploiting stereo redundancy. The spectral and temporal envelope shape of the allowable quantization noise is approximated according to the masked threshold spectrally by controlling the quantizer step size in each subband. The temporal shape is formed by a new adjustment of the step size for each block of the input signal and additional application of a temporal noise shaping technique.

1. Mail address: Institut für Theoretische Nachrichtentechnik; Appelstraße 9A; D–30167 Hannover; Germany
E–mail address: baumgart@tnt.uni–hannover.de

Psychoacoustical models used so far in perceptual coders are based on simplified assumptions which may result in significant errors of the approximated masked threshold. The psychoacoustical models proposed for the AAC–coder use a DFT of successive blocks of the audio signal which provides the associated spectral components of the blocks. For each spectral component an individual masked threshold is generated. The overall masked threshold follows from superposition of the individual thresholds, which is carried out by simply adding up the thresholds at corresponding frequencies. The main limitations of this method are the fixed blockwise signal analysis which determines temporal and spectral resolution and the assumption of linearity needed for the superposition of individual threshold square values. These limitations cause considerable deviations of the approximated masked threshold from the model and psychoacoustical results in case of complex audio signals, because sound analysis in the ear uses a non–uniform spectral analysis and a highly nonlinear processing.

To overcome these limitations a perceptual model can be improved by taking advantage of the increasing knowledge about sound processing of the ear. Therefore, the developed ear model is based not only on psychoacoustical measurements of masking but also on physiological measurements and understanding of procedures involved in auditory perception. This approach enables an accurate modeling of auditory sound analysis which is independent from signal complexity.

In order to summarize the basic measuremental data which should be considered by an improved perceptual model, the important basic physiological and psychoacoustical results are briefly reviewed in Chapter 1. The structure of the proposed ear model is described in Chapter 2. Modeling results are presented in Chapter 3 in comparison of physiological and psychoacoustical counterparts. In Chapter 4 results are discussed and related to the basic features of the model. Finally, conclusion are drawn in Chapter 5.

1 REVIEW OF BASIC PHYSIOLOGICAL AND PSYCHOACOUSTICAL RESULTS

A model for auditory masking must be finally verified by psychoacoustical measurements of masked thresholds for sound signals which cover masking effects that are expected to be relevant for the intended application. In case of audio coding arbitrary sound signals are possible so that relevant masking effects have to be identified by the amount of irrelevance reduction they allow. It can be shown for example that for most sound signals irrelevance reduction is mainly determined by simultaneous masking so that additional consideration of postmasking has only little effect.

Because the relevant masking effects can be only partly explained or modeled in a uniform approach by a direct mathematical approximation and extrapolation of psychoacoustically measured masked thresholds, a more sophisticated consideration of the sound processing in the ear is required. As the understanding of the ear's functions continuously grows, it seems possible to get more close to the physiological origins of masking by improving the accuracy of a physiological ear model with regard to both, psychoacoustical and physiological measuremental results.

1.1 Physiological Measuremental Results

The human ear can appropriately be divided into a peripheral and a central part. The peripheral ear consists of outer, middle and inner ear and processes sound mechanically. The interface of both parts is the cochlea of the inner ear, where the inner hair cells (IHCs) convert the mechanical signal of the peripheral part into a neural representation of action potentials for further processing in the central part. The central ear consists of the auditory pathways with the auditory nerve as first stage which is linked with higher neural processing centers in the brain up to the auditory cortex.

Outer and middle ear behave linearly for sound pressure levels (SPLs) of interest up to 100 dB. Therefore, they can be sufficiently examined by measurements of the frequency response which for the outer ear is described by head–related transfer functions [5][6], reflecting the dependency on sound incidence direction. Results of such measurements basically show a bandpass characteristic [7][8][9] which mostly has negligible influence on investigations of the following processing stages in the medium frequency range.

Sound pressure is coupled into the cochlea via the middle ear ossicles. The membrane of the oval window separates the fluid in the cochlea from the surrounding air in the middle ear. With sufficient accuracy the cochlea can be assumed as a mechanical system of two ducts which are coupled lengthwise by the basilar membrane (BM). If this system is excited by excursions of the oval window a traveling wave arises moving up the cochlea. The traveling wave of a pure–tone produces transverse vibrations of the BM which show continuously growing amplitudes with increasing distance from the oval window until a sudden breakdown occurs immediately after reaching the characteristic resonance location for the tone's frequency. The BM vibration is accompanied by motion of the adjoining fluids. Situated on the BM the organ of Corti contains the inner hair cells (IHCs) distributed over the whole length of the cochlea which are receptors for fluid velocity. The fluid velocity deflects the hair bundles of the IHCs. The IHCs transduce this stimulation signal into neural action potentials of the auditory nerve usually quantified by mean firing rates. The frequency response maximum measured in firing rate of each auditory nerve fiber is the characteristic frequency (CF) of that fiber. A meaningful frequency–place map of the BM can thus be defined by assigning the CF of each fiber to the place on the BM where the associated IHC is innervated.

Parallel to one row of IHCs the cochlea contains three rows of outer hair cells (OHCs) which are stimulated by BM and fluid motion. The OHCs convert the stimulation signal into mechanical force coupling back to the BM vibration [8][10]. At low sound pressure levels a drastic amplification of the traveling wave is caused by the active OHC–"motors" reaching amplification factors of 1000 or even more [11][12][13]. For increasing sound levels amplification is continuously reduced and completely disabled at very high levels. Precise measurements of BM velocity are available for locations at the base (near the oval window) [11][12][13][14][15] and the apex region (near the helicotrema at the cochlear tip) [16] of animal cochleae. Locations in between which are important

for investigations at medium frequencies are unfortunately difficult to explore so that no comparable measurements of the BM motion have been published. Recent results from the cochlear apex [16] suggest that not only the BM vibration determines the sound processing of the cochlea but also the tectorial membrane which is located at the apical side of the hair cells while the BM is located at the base of the hair cells. It is supposed that the tectorial membrane is a second mechanical resonator with an associated independent resonance frequency affecting cochlear frequency selectivity at low frequencies.

Since the observation of BM motion at medium distances from the oval window is very difficult the mean firing rate of auditory nerve fibers is usually taken as indirect measure of cochlear frequency selectivity in this region. However, the drawback of this method is the uncertain influence of the IHC transduction and the IHC synapse which show different types of level–dependent saturated input–output–functions each with a rather limited dynamic range [17][18]. Knowledge of the detailed sound processing in higher centers of the brain gets more and more vague because already the next stage following the auditory nerve, the cochlear nuclei, exhibits broadly different response properties which may have correspondingly different functions. The processing in the final stage, the auditory cortex, cannot be explored in an exact quantitative fashion but the local activity in response to special stimuli can be recorded in order to identify regions of the brain performing different feature detection tasks [19].

Disregarding direct measurements of nerve fiber potentials, electrical responses of the central ear are obtained from electrodes on the surface of the head. This method is also applicable for humans but it detects only the so called "summation potentials" or "evoked potentials" of the whole nerve population providing some indications of the brain location with maximum activity for a given stimulus. Summation potentials are well suited for exploring the different neural delay times occurring when a short sound signal is sequentially processed by neural centers of the auditory pathways. Additionally, this method allows for a control experiment verifying that frequency dependent masking originates from the cochlea because in this case a masked signal will not produce evoked potentials [20].

Another important method for investigations of the cochlea is made possible by otoacoustic emissions (OAEs) of the ear which are generated spontaneously or in response to acoustic stimulation. OAEs can be demonstrated by recording sounds coming out of the ear with a microphone. Their origin is expected in OHC force generation [21] so that OAEs provide means for investigations of OHC stimulation. In masking experiments it can be shown furthermore that emissions are suppressed by appropriate masker signals suggesting that most masking effects especially simultaneous spectral masking are already present in the cochlea [20].

1.2 Psychoacoustical Measuremental Results

Masked thresholds determined in classical psychoacoustical hearing experiments are the reference data to be approximated by the physiological model proposed here. In simple experiments a masker signal of constant spectral envelope and finite duration is presented to a test subject with the task of detecting a test signal in the masker, usually a pure tone, presented at a certain temporal position relative to the masker. The test signal level at the threshold of detection is the masked threshold for the chosen combination of masker and test tone [22] with respect to the signal durations or in general the temporal envelopes. Spectral representations of masked thresholds obtained for varying test tone frequencies are advantageously scaled and explored using critical–band rates in Bark units. The critical–band rate scale reflects the CF distribution on the BM dividing its length of 32 mm into 24 equally spaced partitions each having a bandwidth of 1 Bark. The scaling of the test tone frequency in critical–band rate results in masked threshold curves with a slope steepness approximately independent from masker frequency so that threshold patterns will have the same shape and only their position on the critical–band rate scale will vary with masker frequency.

Loudness is a psychoacoustical quantity which describes the sensation of sound intensity. This total loudness of sound is represented by one value which may change over time. Loudness may also be spectrally analyzed in critical bands resulting in specific loudness values for each of the 24 critical bands. Loudness is taken into consideration because masked thresholds can be interpreted as just noticeable changes in specific loudness. This idea is a fundamental assumption for the modeling approach which is described later. In the aforementioned simple masking experiment the masked threshold can thus be explained by the change in specific loudness [23][24] caused by the superposition of the test signal compared to the masker alone. This change is detectable according to the model of just–noticeable changes [22] if it exceeds a certain threshold value. A comprehensive description of procedures and results for masking and loudness measurements is given in [22].

An overview of the most important psychoacoustical facts according to the ear model will be given in the following which represent the basis of the modeling approach taken here.

1.2.1 Masked Threshold

Psychoacoustic masked thresholds for pure tones and critical–band wide noise maskers at 1 kHz are outlined in Fig. 2. The apparent features of these masked threshold patterns can be summarized as follows.

- For medium and high SPLs the masked threshold patterns show a steep slope for increasing frequencies up to a maximum and a decreasing threshold with shallower slope at high frequencies.
- The upper slope of the masked threshold which covers frequencies higher than the masker frequency shows a pronounced increasing steepness with decreasing masker SPL.
- Thresholds of noise and pure tone maskers show significant differences for equal masker SPLs.
- These threshold differences decrease for increasing frequencies higher than the masker frequency.

Compared to noise signals masked thresholds of pure tones are difficult to obtain in case of pure tone test signals. Interaction of the masking tone and a test tone of neighboring frequency results in audible beating, furthermore, the nonlinear sound processing of the cochlea generates audible difference tones in case of high SPLs [22]. In order to avoid false detection of perceived changes apart from the test tone sensation, experienced subjects or advanced test procedures preventing audible difference tones [25] are necessary. If these detection cues, beating and difference tones,

were allowed, the masked thresholds for tones will be up to 20 dB lower near the masker frequency than without these cues.

Replacing the pure tone by a critical–band noise, having a bandwidth of 1 Bark, as test signal for a masking tone according to Fig. 2 results in comparable threshold differences between tone and noise maskers of approximately 18 dB for the masker center frequency [26]. Pure tone test signals are usually applied because their spectral content is exactly determined and the temporal envelope is constant in contrast to the undeterministic characteristic of noise.

The noise masker used in Fig. 2 was Gaussian noise (GN) arbitrarily taken from a noise generator at different time instances for each masker presentation. At least two different other noise signals have to be distinguished from this ”running” noise. Repeated presentation of a GN masker by playback of one and the same signal results in a ”frozen” noise which shows lower masked thresholds because of adaptation and learning effects of the subject. Noise signals with minimum temporal fluctuations are known as low–noise noise (LNN) [27][28]. Masked thresholds of LNN are lower than for standard GN, if its bandwidth is less than the critical bandwidth of 1 Bark [29][30]. This is also true for noises with reduced envelope fluctuations compared to GN [31].

Variations in bandwidth of the noise masker with constant SPL yield to a 1.5 dB masking increment for every halving of bandwidths below the critical bandwidth of 1 Bark [30]. For increasing bandwidths larger than 1 Bark the masked threshold at the center frequency decreases faster [32] because only the part of the masker energy which falls into the critical band of the test signal is effective in masking.

Complex maskers consisting of combinations of several pure tones and narrow band noises mostly show ”excess” masking compared to the masked thresholds of the individual components. From psychophysics it is expected that the intensity values (square values) of the individual masked thresholds of each masker component alone will be added for the overall threshold. E.g. the addition of two thresholds of equal value results in a 3 dB higher overall threshold compared to the individual thresholds. Psychoacoustical measurements show much higher overall thresholds reaching from about 10 to 20 dB increment at the frequency where two individual thresholds have equal amount [33]. Higher numbers of overlapping individual thresholds yield to higher increments than 3 dB per additional masker component. Excess masking occurs in frequency bands having a minimum distance of 1 Bark from each masker component and for masker component levels well above threshold in quite [34][35][36].

For measurements of temporal masking effects a short test signal burst is used in order to obtain a well defined temporal position of the observed threshold for an appropriate temporal assignment to the time–varying masked threshold. The spectral selectivity of the masked threshold depends on the test tone duration and temporal envelope with shorter durations resulting in poorer spectral selectivity. A Gaussian window is often used providing a good compromise between temporal and spectral selectivity. Test signal durations shorter than about 200 ms show masked thresholds depending on the duration. In this range the threshold decreases by approximately 10 dB for each tenfold test signal duration [22]. Additionally, for durations shorter than 5 ms significantly higher thresholds are measured if the test tone is presented shortly after the masker is switched on. This phenomenon is known as "overshoot" effect which is most pronounced for broad band noise maskers.

For postmasking measurements the test signal is presented after the masker. Postmasking shows level dependent decay slopes with a transition to the threshold in quite after about 200 ms [22]. Higher SPLs yield steeper slopes. Increasing masker durations influence the immediate or delayed decay of the masked threshold. In premasking experiments the test signal is presented immediately before the masker onset. Postmasking shows longer durations than premasking which occurs during some milliseconds. The slopes of premasking are accordingly much steeper. Measurement of premasking is reported to be difficult because of large interindividual deviations, therefore, results are rarely available and not always reliable.

Masking–period patterns [22] are obtained by a continuous presentations of a modulated masker or unmodulated low–frequency masker with a short test signal occurring during a certain phase of the modulation period. This method allows temporal masking effects to be explored not only for pre– and postmasking but also for simultaneous masking using for example a square wave modulated masker. Masking by frequency modulated (FM) signals is a typical example for the measurement of masking–period patterns where the masker frequency is usually sinusoidally modulated. FM–maskers show different frequency dependent masked threshold patterns for the phases of increasing and decreasing masker frequency. In this case the test tone can be presented at the two phase positions where the instantaneous masker frequency is equal to the mean masker frequency for both directions of frequency alteration as shown in Fig. 3. The different masking curves observed for the two phases can be explained by postmasking which acts like a short–time memory on masking. Therefore, with FM–maskers the combined effect of simultaneous and postmasking can be observed.

For the masking effects described above a monaural or a binaural presentation of the same signal is assumed. If the masker or test signal is presented to one ear out of phase with respect to the other ear the masked threshold is significantly reduced for frequencies below 1.5 kHz. For a 180° phase shift of the test signal at one ear this binaural masking level difference (BMLD) amounts to typically 10 dB and up to 20 dB for special signal configurations [5]. Comparable results are obtained for a phase shifted masker and a test signal in phase at both ears.

1.2.2 Loudness

The perception of loudness is related to masking because masking can be partly explained by the detectability of changes in specific loudness which becomes obvious in the case of a pure tone masking a pure tone having the same frequency or the masking of noise by noise.

Measurements of loudness for a pure tone or a wide band noise as function of SPL shows the high dynamic compression of sound level found in loudness. The loudness N, quantified in unit sone, of a high level broadband noise can be assumed to be proportional to the compressed sound intensity I according to Equ. (1) [22].

$$N \sim I^{0.23} \tag{1}$$

The high amount of dynamic compression achieved by the auditory system enables the exceptional dynamic range covered by the auditory system. The compression was for a long time partly attributed to the nonlinearity of IHCs and the following synapses. Recent physiological results suggest that the compression observed in loudness is already achieved by the mechanical BM velocity signal [37]. More advanced measuremental techniques using a laser–velocimeter have shown a higher compression of BM velocity than supposed in the past from animal experiments. If these results are adaptable to man there is clear evidence for this compression to be fully achieved on the mechanical signal processing stage of the cochlea.

Specific loudness is defined according to Zwicker [22] as distribution of loudness on the critical–band scale so that specific loudness N' integrated over critical–band rate z results in total loudness N.

$$N = \int_0^{24} N' dz \tag{2}$$

Loudness is a perception quantity derived in psychoacoustical measurements which considers properties of the whole auditory system. Specific loudness is also assumed to cover the nonlinear properties as well as the frequency selectivity of the ear represented by the distribution over the critical bands. Using specific loudness makes analysis of frequency dependent changes in loudness possible which is important for masked threshold modeling approach used here. The loudness perception of temporally variable sounds has been extensively studied [38][39] and shows similar properties as temporal masking.

1.3 Implications for an Improved Perceptual Model

From physiological measurements of frequency selectivity achieved at the mechanical stage of sound processing in the cochlea and psychoacoustical measurements of masked thresholds it can be concluded that the cochlea mainly determines the frequency selectivity in masking. This is in line with a detailed study [40][41] with the aim of explaining the origins of masking with the outcome that spectral masking results from the cochlea while temporal masking originates mainly from the central ear. Therefore, a cochlear model as part of a perceptual model must provide appropriate filter responses in accordance with the cochlear frequency selectivity. Additionally, the high dynamic compression observed in BM velocity measurements and loudness should be considered in order to achieve a good approximation of specific loudness for the masking modeling.

The generation of masked thresholds can be developed from the principle of intensity discrimination which goes back to the well–known Weber's Law. Intensity discrimination is based on the assumption that relative changes of sound intensity I greater than a certain threshold value are detectable for a human observer. For the measure of change the ratio $(I + \Delta I)/I$ can be chosen where I denotes the masker intensity and $(I + \Delta I)$ denotes the masker plus test signal intensity [42]. Sound intensity should be replaced by specific loudness which is more closely related to masking as stated above. Applying a constant threshold to the derived changes in specific loudness leads to the problem of equal masked thresholds for pure tone and noise maskers in contradiction to psychoacoustics (see Fig. 2). Inspection of masker envelope fluctuations provides a solution to this problem. A pure tone shows no fluctuation so that the ear is able to detect small envelope changes due to a test signal after adaptation to the tone. In contrast, the envelope fluctuations of narrow band noise suggest a reduced sensitivity for envelope changes produced by an additional test signal. The utilization of masker envelope fluctuations by the ear is plausible and explains the different masking by noise compared to a pure tone of equal level as well as the reduced masking of LNN [29][30].

While the frequency–dependent distribution and the compressed dynamic range of specific loudness can be approximated by a cochlear model in accordance with physiological measurements, the time–dependency cannot be sufficiently explained without consideration of sound processing in the central ear. Adaptation and recovery processes found in neural processing mainly influence the temporal response decays of a short–time stimulus in several stages of the auditory pathways. These processes seem to be responsible for the longer duration of postmasking compared to premasking and for the similar time–dependency in loudness perception.

2 STRUCTURE OF THE EAR MODEL

The goal of the modeling approach taken here is to provide a reliable and realistic approximation of the masked threshold produced by arbitrary sounds. The model itself should cover the most relevant (known) features of sound processing of the ear but details not important for masking should be omitted in order to facilitate the understanding of cause and effect and to achieve a reduced model complexity. Signal processing by the model should be carried out in the time–domain because the temporal and spectral features relevant to auditory perception cannot easily be preserved in a frequency–domain representation for arbitrary sounds due to the inherent nonlinearity.

The model structure should be derived from the processing stages known from auditory physiology which provide means for inspections of intermediate physiological quantities. This step by step modeling offers the advantage of a broad range of measurements from physiology and psychoacoustics to be utilized for the tuning of free or not fully determined parameters so that the resulting model can be well–founded in this sense. The following description of the model is split into two parts covering the peripheral and the central ear model respectively. Because knowledge of peripheral ear physiology is more advanced compared to the central ear the modeling approach is based additionally on physiology for the peripheral ear while it is based mainly on psychoacoustics for the central ear.

2.1 Model of Peripheral Ear

Since the outer and middle ear have negligible influence on masking for medium frequencies they were not yet incorporated in the ear model. At the present development state of the model it seems reasonable to concentrate on a proper adjustment of the cochlear model which is currently the only part of the peripheral ear model. After the evaluation of several published cochlear models Zwicker's hardware model [43] with extensions from Peisl [44][45] was selected as most suitable approach

for the aforementioned demands. This cochlear model is based, as many other, on the correspondence of mechanical and electrical quantities. It was developed using some simplifying assumptions about cochlear hydromechanics which result in a one–dimensional mechanical model dividing the cochlear ducts longitudinally into small sections of finite length. The passive cochlear hydromechanics is usually modeled by an electrical network derived from equivalent mechanical quantities. The network shown in Fig. 4 contains in each section a parallel resonance circuit which rebuilds the elasticity, mass, and damping taking effect in each section by the vertical motion of the BM and surrounding fluids and a serial capacitance with parallel damping which account for the coupling of sections by fluid motion in a direction parallel to the BM. The local BM velocity is equivalent to the voltage across the parallel resonance of each section. The active amplification of OHCs can be incorporated into this model by amplifiers with nonlinear feedback acting on the BM velocity signal.

The underlying model, in the following named "original model", proposed by Peisl [44][45] is able to simulate a broad range of cochlear signal processing effects such as nonlinear distortions, dynamic compression, otoacoustic emissions, two tone suppression, and frequency selectivity. This model was implemented using wave digital filters which offer the advantage of time–domain processing and satisfactory stability properties. The most important extensions compared to Zwicker's hardware model due to Peisl consisted in a lateral coupling across 8 neighboring model sections. This coupling shows a stabilizing effect for high amplification factors and additionally an appropriate frequency selectivity compared to the critical bandwidth.

The proposed cochlear model, in the following named "extended cochlear model", is shown in Fig. 5. It consists of Peisl's model extended by a second selective amplification stage (amplification V_2), using the same parallel resonance as found in the associated BM section. This circuit ensures that only the feedback signal of the first stage is amplified outside the feedback loop and added to the output. The output from the extended cochlear model is assumed as IHC stimulation signal associated with the fluid velocity at the hair bundles of the IHCs. The second stage was added in order to achieve the dynamic compression observed in physiological measurements without driving the model out of stability. The original model achieves amplification factors up to 100 with one amplification stage in the feedback loop without instabilities or oscillations. V_2 increases the amplification approximating the desired factor of 1000 which is found in animal experiments. There is no direct physiological evidence for this second stage to exist in the cochlea but it may partly be justified on one hand by the up to now unclear influence of the tectorial membrane and the insufficient stability properties of todays simulation methods on the other hand. The tectorial membrane may effectively amplify the feedback forces of the OHCs for the fluid velocity signal reaching the IHCs without coupling back to BM vibration. The second reason accounts for the numerical and discretization problems in time–domain simulations with nonlinearities involved. Despite of using wave digital filters, which show the best stability properties for this application, they may not be as stable as the natural cochlea.

The coupling of neighboring sections of the extended cochlear model symbolically outlined in Fig. 5 includes 16 sections in each direction with decreasing coupling (higher resistances) for increasing distances (a detailed description can be found in [44]). The incorporation of lateral coupling is essential for achieving a realistic frequency selectivity of the model. It is motivated by the fact that neighboring hair cells may not act in perfect electrical isolation but show capacitive coupling. Additionally, cochlear mechanics shows coupling mechanisms in the longitudinal direction of the BM because vibration of adjacent BM sections is coupled by the BM stiffness apart from the fluids involved.

The simulation results were obtained using 241 model sections corresponding to a bandwidth of 0.1 Bark critical–band rate or 0.133 mm spatial resolution. This section count is sufficiently high in order to avoid discontinuities along the filter chain which otherwise may cause undesired signal reflections. The chosen sampling frequency was 100 kHz ensuring that aliasing, frequency warping, and the effects of the nonlinearity cause only negligible differences compared to a corresponding analog signal processing. The filter parameters for each BM section where extensively taken from [44], however, some parameters had to be adjusted in order to obtain a comparable characteristic of the extended cochlear model with the increased overall amplification.

2.2 Model of Central Ear

The central ear model is associated with auditory pathways and neural centers up to the auditory cortex. The influence of central processing on masking cannot be directly estimated by a physiological modeling approach because the detailed procedures involved are widely unknown. Therefore, a classical psychoacoustical approach is demanding were the system under consideration is often treated as "black box" by modeling its input/output relationship with simple mathematical functions.

The central ear model consists of a set of identical units which process the output from the extended cochlear model in parrallel. In the first stage of each unit outlined in Fig. 6, the squared IHC stimulation signal from the extended cochlear model is low–pass filtered by an RC–network with a time constant of 2 ms. The square value accounts for the intensity sensation observed in perception. Afterwards, "internal" noise is added accounting for the lower limit of perceptible sound levels known as the threshold of hearing in quiet. The resulting signal at this stage is in the following referred to as excitation. The fast rise times observed in premasking are likely to be influenced by different neural delays and the undeterministic time instances of nerve firing. This is incorporated into the model by a Gaussian–shaped temporal spreading function with 2.8 ms standard deviation. In the next part two exponential decay functions are applied. The weighted output of a slow decay function of 50 ms time constant is added to the weighted output of a faster decay function of a time constant in the range of 5 to 20 ms. The weighting is adapted according to the excitation level which amounts for the slow decay function relative to the fast decay function to 99.9% for low levels continuously reduced to a 50% weighting for high levels. The weighting adaptation provides means for the approximation of level–dependent decay slopes in postmasking. Additionally, the faster decay time constant is adapted according to the signal duration, yielding a slower decay for longer durations as observed in loudness perception and postmasking [22]. The central ear model unit described so far is similar to the proposed procedure in [39] if the noise generator and the temporal spreading function is removed. The output is now assumed as specific loudness which represents the desired quantity for subsequent masked threshold approximation.

In order to generate a total loudness approximation a short–time integration should be applied finally after the spectral integration over the critical bands of specific loudness [22]. For masking estimation, in contrast, this temporal integration is applied in each critical band after calculating the specific loudness ratio of the masker plus test signal divided by the masker alone. This method accounts for the dependency on test signal duration in masking. The ratio will be greater than unity approximately as long as the test signal is present so that increasing durations yield to higher integrator output values. The short–time integration is realized by two RC–networks of 50 and 2 ms time constant with the output weighted by 90% and 10% respectively. A threshold detector forms the final stage of each unit as part of each central ear model section with the detection threshold adapted according to the envelope fluctuation of the excitation signal in the associated critical band. Increased fluctuations yield to higher ratio thresholds accounting for the assumed reduced sensitivity for the test signal. The test signal is detected by the central ear model whenever the specific loudness ratio of any section exceeds the ratio threshold r_{thr} of that section.

The proposed ear model does not provide detection of inter–channel cues which require common post–processing of signals for different sections before applying a threshold detector. Furthermore, no binaural processing is considered at the present state.

The adaptation networks for the parameter adjustment in the central ear model are fed by the excitation signal. Only the fluctuation detector will be explained further because it is of essential influence for the modeling of simultaneous masking. The excitation level detector and the network for duration dependent adaptation are omitted. The fluctuation detector outlined in Fig. 7 performs envelope detection similar to an AM–demodulator. The output of one RC–network, the minimum detector, immediately follows the temporal minima while the other, the maximum detector, immediately follows the maxima of the input signal. For a constant input signal the outputs of both RC–networks converge to the input value at a rate determined by their time constant. The fluctuation is measured as relative change with respect to the signal value by using the ratio of the minimum and maximum detector outputs. A subsequent low–pass filtering prevents the evaluation of signal fine–structure for the adaptation of the threshold detector.

2.3 Masked Threshold Generation

A test signal is assumed to be detectable if the specific loudness ratio in any section exceeds the ratio threshold of the section. For the ratio calculation it is necessary to first process the masker signal in order to obtain the reference values of specific loudness over time. In each subsequent processing step the masker and test signal are applied together and the test signal level is iteratively adjusted so that for the first section where the ratio threshold is reached, the distance of the ratio and the ratio threshold is minimized. This represents the condition of the test signal at threshold and is the desired value of the masked threshold. The generation of masked threshold patterns is carried out by repeating the iteration for changed test signal parameters, e.g. frequency, duration, or delay according to the parameter of interest.

For investigations of narrow band or pure–tone maskers and test signals the spread of excitation over the model sections is highly concentrated around sections with CFs corresponding to the masker or test signal center frequency. Model sections with lower CFs may react sensitively to small changes due to the superimposed test signal. In order to avoid a detection of these low level "artefacts" in masked threshold generation the evaluation of threshold detectors is limited to sections with a CF in the range of –1 to +2.5 Bark around the test signal center frequency where the test signal detection is expected. This method is also applicable for the inclusion or exclusion of difference tone detection in masking by fitting the CF range of the section evaluation to the desired detection frequencies.

2.4 Model Parameter Adjustment

The filter quality factors (Q–values) of the extended cochlear model, which determine the relative filter bandwidth, had to be lowered by approximately a factor of two compared to the original model in order to achieve a comparable frequency selectivity in conjunction with the increased amplification due to the additional second amplification stage. The shape of the nonlinearity was also slightly modified in order to achieve a dynamic compression function which approximates the signal level dependency observed in loudness perception.

The parameters of the central ear model were adjusted according to temporal masking patterns. The adjustment is simplified by the fact that each stage of the model, to a great extend independently from the other stages, accounts for the effect of a certain signal parameter variation on temporal masking. The reference masking patterns are replicated as part of the presented results in the next section in order to demonstrate the degree of correspondence between psychoacoustical measurements and modeling results.

3 MODELING RESULTS

Selected simulation results from the ear model are summarized in this Chapter in order to illustrate the broad range of psychoacoustical and physiological measurements which can be covered by the proposed modeling approach. At the same time the limitations of the present ear model are outlined by a quantitative comparison of simulations and published psychoacoustical measurements which were taken as reference.

3.1 Results of Physiological Simulations

Simulation results from the original cochlear model for two–tone suppression, otoacoustic emissions, and nonlinear distortions [44][45] are expected to occur also in the extended cochlear model and are not replicated here. Different behavior of the extended model due to the increased amplification results in a higher dynamic compression which is coupled with a more locally concentrated distribution of IHC stimulation at low levels. The results for the IHC stimulation over the BM distance, which is equivalently given by the critical–band rate, is given in Fig. 8 for 1 kHz pure tones with different SPLs. The apparent features of the stimulation patterns are the dynamic compression around the CF of 1 kHz (8.5 Bark) which disappears at higher CFs and the reduced frequency selectivity at high SPLs. From the maxima of the stimulation patterns the so called "BM–input/output–function" for the CF location is derived in Fig. 9, showing the level–dependent shape of

compression at CF. This function is usually observable from physiological BM–measurements at a single observation point. As aforementioned direct comparisons of physiological BM–I/O–functions are not yet possible because measurements were published only for the low and high CF regions of the cochlea. The BM–I/O–function shows a compression of the dynamic range from 0 to 100 dB SPL into IHC stimulation range spanning over 45 dB with a minimum steepness of 0.27 dB/dB at 70 dB SPL. The observed sharp tuning of the stimulation patterns at low levels is connected with an increased temporal extension of the impulse responses. Therefore, the signal delay due to the filtering process is level dependent and shows similar values as observed from measurements of action potentials from the auditory nerve or evoked potentials [19][46].

3.2 Comparison to Loudness Perception

In the literature excitation is often associated with loudness perception and specific loudness. The comparison of the physiological BM–I/O–functions of different species with the shape of the level–dependent loudness perception function shows apparent similarities [37]. In Fig. 10 the excitation integral over critical–band rate for each excitation pattern of Fig. 8 is plotted together with the human loudness functions for a pure tone and noise. This integral represents the approximated total loudness corresponding to Equ. (2) which is adjusted so that it coincides with the 1 sone, 40 dB point of the pure tone loudness function, the reference value for the sone scale [22]. The comparison shows that approximately the same dynamic compression known from loudness perception is achieved by the critical–band integral of excitation.

3.3 Comparison to Psychoacoustic Masked Threshold

Generated masked thresholds for pure tones at 1 kHz and critical–band wide Gaussian noise centered at 1 kHz masking a pure tone are shown in Fig. 11 and 12 respectively. The masker duration was 700 ms while the test tone was switched on during the last 500 ms of the masker each with a 5 ms Gaussian start and stop window. The results show nearly the same level–dependency of the upper slope of the masked thresholds and a quantitative coincidence. It should be noted that it is not obvious for a sole model to generate different slopes for a pure tone compared to a noise masker of the same level as was already shown in Fig. 2. The reason can be found in nonlinear processing and will be explained in Chapter 4. The masked threshold of the pure tone with 100 dB SPL shows significant higher values than the psychoacoustical reference. This difference disappears if other psychoacoustical reference values are taken, which show comparable growth of masking functions of equal expansive character at the upper slope [31]. The steepness of the lower slope is always higher than observed in psychoacoustics but the property of shallower slopes for the noise maskers compared to the pure tone maskers can be demonstrated. Near the masker center frequency the generated masked threshold shows variations caused by interactions of masker and test signal. In case of the pure tone maskers this effect was considered by taking the mean value of the simulation results for 6 phase shifted test signals by multiples of 60° of the test tone period. In case of the noise masker the variations result partly from the fluctuating spectral envelope of the masker. The psychoacoustical masking patterns of Fig. 11 and 12 were used for parameter adjustment of the ratio threshold detector and the fluctuation detector. Therefore, both results demonstrate the best possible coincidence which can be obtained from the proposed ear model with respect to the psychoacoustical reference. All parameters of the ear model were kept constant for the simulation results presented in this paper.

Generated masked thresholds for narrow band noise maskers with minimum temporal envelope fluctuations (LNN) are outlined in Fig. 13 together with the generated masked thresholds for pure tones and GN of Fig. 11 and 12 and a GN masker having a bandwidth of only 10 Hz. The LNN maskers have the same bandwidths of 10 Hz and 160 Hz as the GN with the latter bandwidth corresponding to the critical–band width at 1 kHz. The LNN was generated by the algorithm described in [27]. Near the masker center frequency the generated masked threshold was obtained again by averaging 6 results for phase–shifted test tones by multiples of 60° in order to avoid large variations due to interactions of the masker and test tone. In general, the generated masked threshold of a LNN masker approximates the threshold of a pure–tone masker if the bandwidth is continuously reduced. The comparison of masking obtained from LNN and GN with equal bandwidth shows that LNN is more effectively masking than GN for frequencies higher than the masker center frequency. This contradicts the results published in [31] where lower LNN thresholds compared to GN were found at a critical–band rate of 3 Bark above the masker center frequency. In that study, the pure tone results in the highest masked threshold compared to LNN and GN. Masking of LNN compared to GN at the center frequency was investigated in [29] where a 5 dB smaller masked threshold was observed for LNN in case of a 100 Hz bandwidth at 60 dB SPL. Unfortunately, the simulation results from the ear model are not clearly interpreted at the center frequency which corresponds to 8.5 Bark. At 100 dB SPL the generated masked thresholds at 8.5 Bark are in line with the reduced threshold for LNN maskers of equal bandwidth. At 80 and 60 dB SPL the differences seem to disappear at 8.5 Bark but are still apparent at 9 Bark. At high levels the shallower upper slope of pure tone maskers is also obtained for the LNN maskers. The variation in bandwidth results in shallower masking slopes below the center frequency if the bandwidth is greater than the critical bandwidth.

Psychoacoustical tuning curves are measured using a test tone with fixed frequency and SPL while adjusting the masker SPL so that the test tone reaches the masked threshold. In this manner the tuning curves specify the frequency–dependent masker SPL needed for keeping the test tone at the masked threshold. This representation can be derived from classical masking patterns and is just another possibility in choosing the fixed and variable parameters of a graph. Such a graph allows for a direct comparison with the frequency selectivity observed in physiological tuning curves of single nerve fibers. In Fig. 14 the tuning curves for 1 kHz test tones of different SPLs are outlined. At low levels up to 40 dB SPL the generated tuning curves show significantly higher masker levels. For higher levels the reference tuning curves are well approximated by the ear model at the shallower slopes (for masker frequencies lower than the test tone frequency). The high frequency slopes are steeper for the model results compared to the reference which corresponds to the increased steepness of the lower slope of the generated masking patterns in Fig. 11. These results indicate that simultaneous masking is adequately modeled not only at a single masker frequency as shown in Fig. 11 and 12 but in the medium frequency range.

In contrast to LNN masking, the properties of temporal masking are well documented, e.g. [22]. In Fig. 15 psychoacoustical masking patterns of postmasking by white noise using a short test pulse are compared to the generated masked thresholds. The ear model parameters which influence

postmasking were adjusted according to these reference values, so that the best fit is achieved. This result shows that postmasking can be adequately approximated by the incorporation of a level–dependent weighting of the outputs of two different exponential decay functions as described above. This seems to be the most simple approach for the generation of the level–dependent decay slopes in postmasking. Two examples for generated premasking patterns together with reference data from [47][48] are given in Fig. 16. These results show that modeling of premasking is provided but it should not be overinterpreted regarding the quantitative agreement of both patterns. The left diagram shows premasking of a uniform masking noise from the reference [47] together with the generated masked threshold. The pronounced overshoot effect which is apparent for broadband noise maskers is not covered by the ear model, so that the generated threshold is significantly lower. The right diagram shows results for premasking of a critical–band wide noise masker centered at 8.5 kHz which shows a much reduced overshoot effect. The generated thresholds in both diagrams have a slope steepness comparable to the reference thresholds but are temporally shifted in the range of 2 to 8 ms.

The masking dependency on test signal duration is outlined in Fig. 17. The decay of masking at a rate of 10 dB per decade for durations shorter than 200 ms as observed in psychoacoustics is equally obtained from the ear model. The variations in the generated masked thresholds for different test tone frequencies are likely to occur due to the probabilistic masker signal, in this case a uniform masking noise which can be generated from white noise by attenuating the frequency range starting from 1 kHz with continuously increasing damping reaching 15 dB at 20 kHz.

The combined effect from simultaneous and postmasking can be explored by frequency modulated maskers. Psychoacoustical reference measurements for a sinusoidally modulated pure tone are plotted in Fig. 18 together with simulation results from the ear model. The instantaneous frequency of the FM–maskers can be read according to Fig. 3 with $f_{\mathrm{mean}} = 1.5$ kHz, $f_{\max} = 2.2$ kHz, and $f_{\min} = 800$ Hz. The frequency of modulation $1/T$ for the two maskers is 0.5 Hz and 8 Hz. The 5 ms test tone impulse was presented at four phase positions of the modulation period at $t_{\mathrm{T}} = 0,\ T/4,\ 2T/4,\ 3T/4$. At each phase position the masked threshold was determined for test tones in the range of 6 to 16 Bark. If no postmasking were present the same masking patterns for the positions $T/4$ and $3T/4$ are expected due to the same instantaneous masker frequency. This is approximately true for the slow 0.5 Hz frequency of modulation while postmasking is effective at 8 Hz modulation. The results from the ear model in general show more pronounced maxima and a comparable shift of the masking patterns in critical–band rate for these two phase positions. The other two phase positions are given in Fig. 18 for completeness.

The ear model provides adequate approximation of the most important masking effects for applications in audio coding. Therefore, it is seems suitable for the detection of audible distortions caused by audio data rate reduction. For this simulation the masker signal and the test signal for the masked threshold generation are replaced by the original audio signal and the distorted signal. Subsequent processing of both signals yields to specific loudness ratios in each model section which determine if the associated ratio threshold is exceeded as well as the distance of both values. The distances obtained from each section represent a measure of how far the distortion is away from being audible if the ratio threshold is not exceeded and how loud the distortion is perceived in the other case. In Fig. 19 the spectrogram of the original and coded signal of a castanet is shown. The coded signal shows apparent regions where no signal energy is present because they are assumed to be inaudible due to masking from neighboring signal components. After processing of both signals by the ear model the resulting distance of the specific loudness ratio and the ratio threshold in each section is derived and shown in Fig. 20 indicating that audible distortions occur at the third and fourth strike of the castanets. This result was confirmed by a listening test where these two strikes sounded somehow distorted without further characterization of the artefacts, e.g. a temporal preecho or a spectrally limited distortion. Such analysis of coded sound signals offers the possibility for coder improvements due to the precise locating of artefacts which can often be traced back to an insufficient coder parameter adjustment. Additionally, this method indicates spectro–temporal regions where significant higher distortions are tolerable due to the greater distance from the ratio threshold. In these regions a coarser quantization could be used for an improved data rate reduction.

4 DISCUSSION

This chapter covers general aspects of the ear model which give further arguments for the proposed model structure including the inserted second amplification stage of the extended cochlear model. The increased amplification factor for low SPLs of the cochlear model becomes evident from the analysis of the upper slopes of the generated excitation patterns compared to the upper slopes of masking patterns for pure tones and critical–band noise.

In the following the cause of the different steepnesses of upper slopes in the generated masking patterns of Fig. 11 and 12 for corresponding levels of pure tone and critical–band noise maskers are explained by inspection of their excitation patterns. In Fig. 21 the excitation patterns of 1 kHz pure tone maskers at 100 dB and 60 dB are shown on the left and right graph respectively which are replicated from Fig. 8. The generated masked threshold is given by the lowest curve with two marked points at 10 and 13 Bark in each graph. For these marked critical–band rates and masked threshold levels the excitation patterns were calculated and plotted representing the excitation due to the test tone at the masked threshold without the masker signal. Assuming a linear model, even though this cannot be justified, the excitation patterns of the masker and test tone had to be added for obtaining the excitation pattern of their superposition. This property of superposition known from linear systems was shown to be applicable to a great extent also for the highly nonlinear original cochlear model [44]. From this property it follows for varying test tone frequencies that graphically the test tone excitation peak moves on the masker excitation pattern so that it just meets the curve without going far beyond. Therefore, the steepness of the masked threshold is determined mainly by the steepness of the masker excitation pattern and the level–dependent level difference between the test tone SPL and its associated excitation peak level. The level difference increases with decreasing test tone SPL so that the generated masked threshold usually shows increased steepness compared to the masker excitation pattern.

As shown in Fig. 2 the masked thresholds for critical–band noise are steeper than for pure tone maskers of equal SPL. This can be explained by the associated excitation patterns of this masker

analogous to the case of a pure tone masker. In case of critical–band noise maskers two additional effects have to be considered which consist of the reduced peak excitation level due to the spectral spread of the masker and the masker fluctuation. The reduced peak excitation results in a steeper upper slope of excitation. Additionally, the test tone is detected during a temporal minimum of the fluctuating masker envelope so that the instantaneous masker excitation pattern has even steeper slopes than expected from mean excitation. The reduced excitation does not yield to a reduced masked threshold compared to a pure tone masker because the fluctuation is evaluated so that in this case an increased test tone level is needed to meet the threshold criterion.

From this discussion it follows that the upper slope steepness of the generated masking patterns of pure–tone maskers and critical–band noise maskers is widely determined by the upper slope of the masker excitation pattern and the level difference between the test tone SPL and its excitation peak. The upper slope steepness of the excitation pattern far from the resonance location depends solely on the underlying cochlear hydromechanics independent from the usage of active amplification in each section. This property becomes obvious from Fig. 8 where all excitation slopes for critical–band rates greater than 20 have approximately equal steepness and shape. It is assumed that the extended cochlear model adequately reflects the passive hydromechanical properties so that the upper slope steepness far from the resonance location is fixed and cannot be subject to any parameter adjustment procedure. Therefore, the slope steepness of masking patterns has to be adjusted by an appropriate control of the nonlinear amplification which determines the level difference between the test tone SPL and its associated excitation peak. It should be noted that this conclusion is based on the assumption of a level–independent threshold criterion to be applied.

The nonlinear amplification function resulting from the slope adjustment turns out to be in accordance with amplification functions observed in animal experiments and with the dynamic compression measured in loudness perception. Even though the original cochlear model is driven out of stability for increased amplification factors in the feedback loop this approach is justified by the aforementioned correspondence which is not known up to now to be achievable with cochlear models containing lower amplification factors.

From the structure of the ear model some implications follow for the most important procedures involved in sound processing of the ear which may be inconsistent with other published physiological masking modeling approaches. The dynamic compression of the ear is fully present at the mechanical stage of sound processing in the cochlea so that the IHC population as a whole or following synapses are not likely to produce significant additional dynamic compression. Furthermore, the frequency selectivity in masking is determined by the extended cochlear model while temporal masking effects originate from the central ear model which is in line with psychoacoustical and physiological results [40][41].

The psychoacoustical modeling results can be assessed by inspecting the degree of conformance with the reference and a weighting according to the significance of the concerned masking effect for an application in perceptual audio coding. Simultaneous masking of stationary sounds is one of the most important effects to be exploited in audio coding. This effect is covered precisely by the ear model in case of critical–band GN maskers. Masking patterns of pure tones are less precisely met which may be attributed to reduced reliability and reproducibility of the reference data compared to GN maskers. Modeling results for the special LNN maskers show larger differences compared to the rarely available published reference data so that these results have still to be verified by additional psychoacoustical reference measurements. The proposed fluctuation detector which controls the threshold detector is a suitable uniform approach for considering the different masking effects of these maskers. Temporal masking is less important regarding its impact on irrelevance reduction in coding. Postmasking has longer duration than premasking but premasking is important for controlling the temporal spread of quantization noise due to the limited temporal resolution of the spectral decomposition used in audio coding so that the maximum possible irrelevance reduction following from the masked threshold is not achievable for a realized coder. While postmasking in its pure form is closely met in terms of level–dependent decay rates the results from FM–maskers are only qualitatively comparable. Nonetheless, this result indicates that the ear model structure is suitable for the reproduction of the FM masking effect. The slope steepness of premasking is in accordance with the modeling results but the maximum masked threshold values deviate significantly due to the overshoot effect in case of wide band noise maskers. This effect may be incorporated into the ear model by a detector for fast rising levels of a wide band masker which may be further extended by additional threshold detectors applied to commonly processed signals of neighboring model sections. In case of pure tone maskers the overshoot effect is covered by the adaptation time of the fluctuation detector yielding to an increased threshold for short test tone delays relative to the masker onset. The origin of different temporal shifts in premasking simulations compared to reference data may partly be explained by unreliable psychoacoustical data, other more systematic reasons are not known.

5 CONCLUSIONS

The proposed ear model for auditory masking is based on recent results from physiology and psychoacoustics. It is formed by an extended cochlear model which represents the peripheral ear and a central ear model. Outer– and middle ear are currently not incorporated into the model so that the results are only valid in the medium frequency range. The extended cochlear model is equivalent to simplified inner ear physiology due to an explicit approximation of hydromechanics and consideration of outer hair cell force generation. This model represents the cochlea by equally spaced sections for all of which the inner hair cell stimulation signals are generated. The central ear model consists of a set of identical units driven in parallel from the outputs of the extended cochlear model sections so that each unit is part of the associated ear model section. Each unit comprises short–time integrators and exponential decay functions as main stages for the modeling of temporal masking effects like postmasking and duration–dependent masking. The final stage in each unit is a threshold detector for changes in generated specific loudness which is taken as detection cue in masked threshold generation. These changes are derived from the ratio of specific loudnesses of masker plus test signal and masker alone. The masked threshold of the test signal is derived by an iterative adjustment of the test signal level until the ratio just reaches the threshold value in any unit. The simulations were carried out in the time domain simplifying the processing of temporally variable sounds and the analysis of temporal masking effects compared to other frequency–domain models.

Physiological results obtained from the extended cochlear model agree with the frequency selectivity and dynamic compression observed in measurements of basilar membrane motion. This agreement is achieved by the high amount of outer hair cell amplification incorporated into the extended cochlear model.

Masked thresholds for simultaneous masking of pure tone and critical–band wide noise maskers are closely met by the modeling results. This includes the level–dependent slope steepness of masking patterns which are not covered by known psychoacoustical models for coding applications. Different masking properties of pure tones and noise are considered by the analysis of masker envelope fluctuations which is also useful in case of frequency modulated signals or noise signals with flat envelope (low–noise noise) for which standard tonality estimation procedures in psychoacoustical models fail. Binaural masking effects are currently not covered by the ear model but should be additionally incorporated in order to account for the possible significant reduction of masking at low frequencies.

The generated temporal masking patterns are in qualitative agreement with psychoacoustics so that premasking and postmasking effects can be demonstrated with small quantitative deviations. Standard psychoacoustical models are restricted to simultaneous masking and exhibit unreliable results in case of signals showing mainly temporal masking.

At the present state the ear model is suitable for the automatic detection of audible distortions in coded signals. The intended application of the ear model for the quantization noise control in audio coding requires a simplified masked threshold generation avoiding the complex iteration loop. The most obvious approach for this future development step can be based on the idea that only the masker signal needs to be processed by the ear model. The masked threshold is generated afterwards by approximating the inverse ear model at the point of operation for calculating the necessary test tone level from the known threshold value for the specific loudness ratio. This procedure introduces only small deviations if the test signal level is small compared to the masker level.

6 ACKNOWLEDGMENT

The author wishes to thank Jon Pumplin who kindly provided the software for low–noise noise generation. This work was supported by the ”Deutsche Forschungsgemeinschaft (DFG)”.

7 REFERENCES

[1] Brandenburg, K.; Bosi, M.; "Overview of MPEG Audio: Current and Future Standards for Low–Bit–Rate Audio Coding", *J. Audio Eng. Soc.*, Vol. 45, No. 1/2, Jan./Feb. 1997.

[2] ISO/IEC JTC1/SC29/WG11; "Coding of moving pictures and audio – MPEG–2 Advanced Audio Coding AAC". *ISO/IEC 13818–7 International Standard*, 1997.

[3] Bosi, M.; Brandenburg, K.; Quackenbush, S.; Fielder, L.; Akagiri, K.; Fuchs, H.; Dietz, M.; Herre, J.; Davidson, G.; Oikawa Y.; "ISO/IEC MPEG–2 Advanced Audio Coding", *Presented at the 101st AES Convention*, Preprint no. 4382, Los Angeles, Nov. 1996.

[4] *ISO/IEC JTC1/SC29/WG11 Doc. N1279*; "NBC Reference Model 3 monophonic subjective tests: overall results". ISO/IEC, 1996.

[5] Blauert, J.; *Spatial Hearing*, MIT Press, 1983.

[6] Gardner, W. G.; Martin, K. D., "HRTF measurements of a KEMAR", *J. Acoust. Soc. Am.*, Vol. 97 (6), pp. 3907–3908, 1995.

[7] Nedzelnitsky, V.; "Sound pressures in the basal turn of the cat cochlea", *J. Acoust. Soc. Am.*, Vol. 68, pp. 1676–1689, 1980.

[8] Evans, E. F.; "Basic Physiology of the Hearing Mechanism", *Proc. AES 12th Int. Conf.*, Denmark, pp. 11–20, 1993.

[9] Gardner, M. B.; "Network representation of the external ear", *J. Acoust. Soc. Am.*, Vol. 52, pp. 1620–1628, 1972.

[10] Hallworth, R.; Evans, B. N.; Dallos, P.; " The Location and Mechanism of Electromitility in Guinea Pig Outer Hair Cells", *J. Neurophysiology*, Vol. 70, No. 2, pp. 549–558, Aug. 1993.

[11] Johnstone, B. M.; Patuzzi, R.; Yates, G. K.; "Basilar membrane measurements and the travelling wave", *Hear. Res.*, Vol. 22, pp. 147–153, 1986.

[12] Nuttall, A. L.; Dolan, F. D.; "Steady–state sinusoidal velocity responses of the basilar membrane in guinea pig", *J. Acoust. Soc. Am.*, Vol. 99 (3), pp. 1556–1565, Mar. 1996.

[13] Ruggero, M. A.; Rich, N. C.; Recio, A.; Narayan, S. S.; Robles, R.; "Basilar–membrane responses to tones at the base of the chinchilla cochlea", *J. Acoust. Soc. Am.*, Vol. 101 (4), pp. 2151–2163, Apr. 1997.

[14] Robles, L.; Ruggero, M. A.; Rich, N. C.; "Basilar membrane mechanics at the base of the chinchilla cochlea. I. Input–output functions, tuning curves, and response phases", *J. Acoust. Soc. Am.*, Vol. 80 (5), pp. 1364–1374, Nov. 1986.

[15] Sellik, P. M.; Patuzzi, R.; Johnstone, B. M.; "Measurement of basilar membrane motion in the guinea pig using the Mössbauer technique", *J. Acoust. Soc. Am.*, Vol. 72 (1), pp. 131–141, Jul. 1982.

[16] Gummer, A. W.; Hemmert, W.; Zenner, H.–P.; "Resonant tectorial membrane motion in the inner ear: Its crucial role in frequency tuning", *Proc. Natl. Acad. Sci. USA*, Vol. 93, pp. 8727–8732, Aug. 1996.

[17] Javel, E., "Basic Response Properties of Auditory Nerve Fibers", in *Neurobiology of Hearing: The Cochlea*, Eds.: Altschuler R. A., et. al., 1986.

[18] Zagaeski, M.; Cody, R. A.; Russell, I. J.; Mountain, D. C.; "Transfer characteristic of the inner hair cell synapse: Steady–state analysis", *J. Acoust. Soc. Am.*, Vol. 95 (6), Jun. 1994.

[19] Pickles, J. O.; *An Introduction to the Physiology of Hearing*, 2nd ed., Academic Press, London, 1988.

[20] Zwicker, E.; Stecker, M.; Hind, J.; "Relations between Masking, Otoacoustic Emissions, and Evoked Potentials", *Acustica*, Vol. 64, pp. 102–109, 1987.

[21] Nuttall, A. L.; Tianying, R.; "Electromotile hearing: evidence from basilar membrane motion and otoacoustic emissions", *Hear. Res.*, Vol. 92, pp. 170–177, 1995.

[22] Zwicker, E.; Fastl, H; *Psychoacoustics: Facts and Models*, Springer, New York, 1990.

[23] Gralla, G.; "Modelle zur Beschreibung von Wahrnehmungskriterien bei Mithörschwellenmessungen", (German), *Acustica*, Vol. 78, pp. 223–245, 1993.

[24] Zwicker, E.; "Dependence of post–masking on masker duration and its relation to temporal effects in loudness", *J. Acoust. Soc. Am.*, Vol. 75 (1), Jan. 1984.

[25] Schöne, P.; *Über den Zusammenhang zwischen Erregungsmuster und Schwankungsstärke eines Sinustons*, Ph.D. thesis (Dissertation, German), Technical University of Munich, 1977.

[26] Hellman, R. P.; "Asymmetry of masking between noise and tone", *Perception & Psychophysics*, Vol. 11 (3), pp. 241–246, 1972.

[27] Pumplin, J.; "Low–noise noise", *J. Acoust. Soc. Am.*, Vol. 78 (1), pp. 100–104, Jul. 1985.

[28] Hartmann, W. M.; Pumplin, J.; "Periodic signals with minimal power fluctuation", *J. Acoust. Soc. Am.*, Vol. 90 (4), pp. 1986–1999, Oct. 1991.

[29] Hartmann, W. M.; Pumplin, J.; "Noise power fluctuations and the masking of sine signals", *J. Acoust. Soc. Am.*, Vol. 83 (6), pp. 2277–2289, Jun. 1988.

[30] Kohlrausch, A.; Fassel, R.; v. d. Heijden, M.; Kortekaas, R.; v. d. Par, S.; Oxenham, J. A.; Püschel, D.; "Detection of tones in low–noise noise: Further evidence for the role of envelope fluctuations", *Acta Acustica / Acustica*, to appear in May/June issue, 1997.

[31] v. d. Heijden, M.; Kohlrausch, A.; "The role of envelope fluctuations in spectral masking", *J. Acoust. Soc. Am.*, Vol. 97 (3), pp. 1800–1807, 1995.

[32] Bos, C. E.; de Boer, E.; "Masking and Discrimination", *J. Acoust. Soc. Am.*, Vol. 39 (4), pp. 708–715, 1966.

[33] Lutfi, R. A.; "Additivity of simultaneous masking", *J. Acoust. Soc. Am.*, Vol. 73 (1), pp. 262–267, Jan. 1983.

[34] Moore, B. C. J.; ”Additivity of simultaneous masking, revisited”, *J. Acoust. Soc. Am.*, Vol. 78 (2), pp. 488–494, Aug. 1985.

[35] Humes, L. E.; Lee, L. W.; ”Two experiments on the spectral boundary conditions for non-linear additivity of simultaneous masking”, *J. Acoust. Soc. Am.*, Vol. 92 (5), pp. 2598–2606, 1992.

[36] Cokely, C. G.; Humes, L. E.; ”Two experiments on the temporal boundaries for the nonlinear additivity of masking”, *J. Acoust. Soc. Am.*, Vol. 94 (5), pp. 2553–2559, 1993.

[37] Yates, G. K.; Winter, I. M.; Robertson, D.; ”Basilar membrane nonlinearity determines auditory nerve rate–intensity functions and cochlear dynamic range”, *Hear. Res.*, Vol. 45, pp. 203–219, 1990.

[38] Zwicker, E.; ”Dependence of post–masking on masker duration and its relation to temporal effects in loudness”, *J. Acoust. Soc. Am.*, Vol. 75 (1), Jan. 1984.

[39] Zwicker, E.; ”Procedure for calculating loudness of temporally variable sounds”, *J. Acoust. Soc. Am.*, Vol. 62 (3), pp. 675–682, Sept. 1977.

[40] Scherer, A.; *Erklärung der spektralen Verdeckung mit Hilfe von Mithörschwellen– und Suppressionsmustern*, Ph.D. thesis (Dissertation, German), Technical University of Munich, 1987.

[41] Scherer, A.; ”Erklärung der spektralen Verdeckung mit Hilfe von Mithörschwellen– und Suppressionsmustern”, (German), *Acustica*, Vol. 67, pp. 1–18, 1988.

[42] Grantham, D. W.; Yost, W. A.; ”Measures of intensity discrimination”, *J. Acoust. Soc. Am.*, Vol. 72 (2), pp. 406–410, Aug. 1982.

[43] Zwicker, E.; ”A Hardware Nonlinear Preprocessing Model with Active Feedback”, *J. Acoust. Soc. Am.*, Vol. 80 (1), pp. 146–152, Jul. 1986.

[44] Peisl, W.; *Beschreibung aktiver nichtlinearer Effekte der peripheren Schallverarbeitung des Gehörs durch ein Rechnermodell*, Ph.D. thesis (Dissertation, German), Technical University of Munich, 1990.

[45] Zwicker, E.; Peisl, W.; ”Cochlear preprocessing in analog models, in digital models and in human inner ear”, *Hear. Res.*, Vol. 44, pp. 209–216, 1990.

[46] Baumgarte, F.; ”A physiological ear model for specific loudness and masking”, to appear in *Proc. Workshop on Applications of Sig. Proc. to Audio and Acoustics*, New Paltz, Oct. 1997.

[47] Fastl, H.; ”Temporal Masking Effects: I. Braod Band Noise Masker”, *Acustica*, Vol. 35, No. 5, pp. 287–302, 1976.

[48] Fastl, H.; ”Temporal Masking Effects: II. Braod Band Noise Masker”, *Acustica*, Vol. 36, No. 5, pp. 317–331, 1976/77.

8 FIGURES

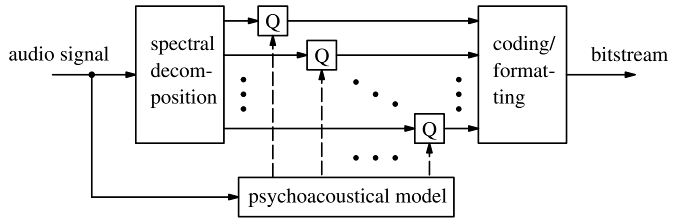

Figure 1. Simplified general block diagram of a perceptual encoder. The subband quantizers are denoted with ”Q”.

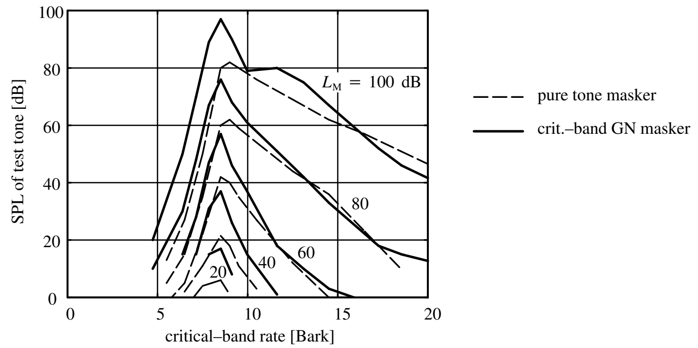

Figure 2. Psychoacoustic reference masked thresholds of critical–band wide Gaussian noise (GN) and pure tone maskers at 8.5 Bark (1 kHz) for test tones (adapted from [22]).

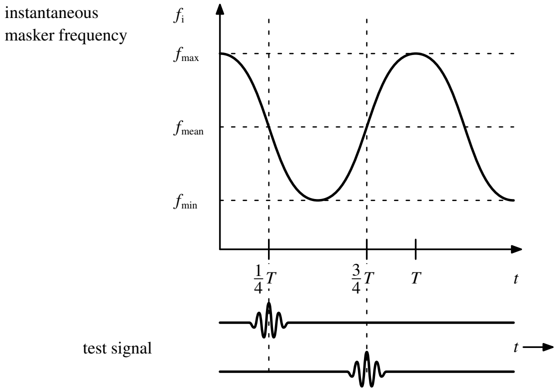

Figure 3. Paradigm for the measurement of masking–period patterns of an FM–masker.
Upper diagram: instantaneous masker frequency.
Lower part: test tone positions for equal instantaneous masker frequency f_{mean} on the falling and rising masker frequency slope (symbolic waveform).

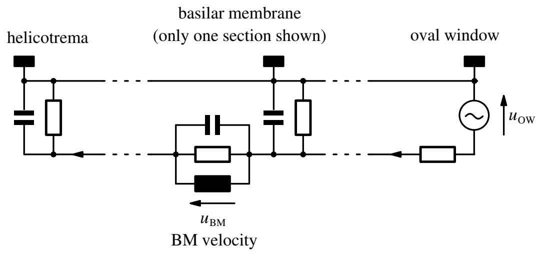

Figure 4. Equivalent electrical network of the simplified passive cochlear hydromechanics. All sections representing the basilar membrane have the same structure. The characteristic frequencies of the sections (resonance frequencies) increase from the helicotrema up to the oval window.

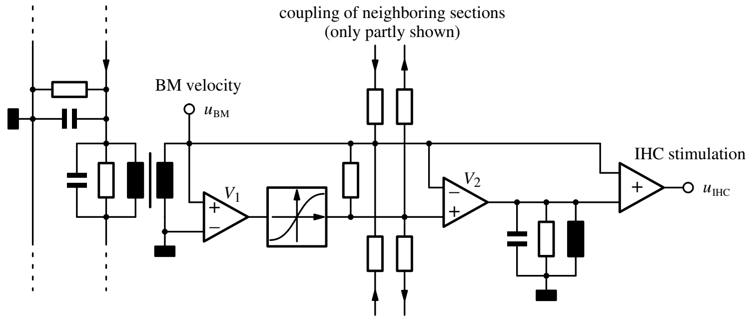

Figure 5. Electrical circuit of one section of the extended cochlear model. On the left side of the transformer the associated section of the passive model in Fig. 4 can be found. The circuit extension compared to Peisl's model (original cochlear model) [44][45] is formed by the second amplifier (V_2), the following parallel resonance, and the adder. The coupling spans over 16 neighboring sections in each direction using increasing resistances for increased distances which is only briefly indicated by 4 resistances instead of 64 in the model.

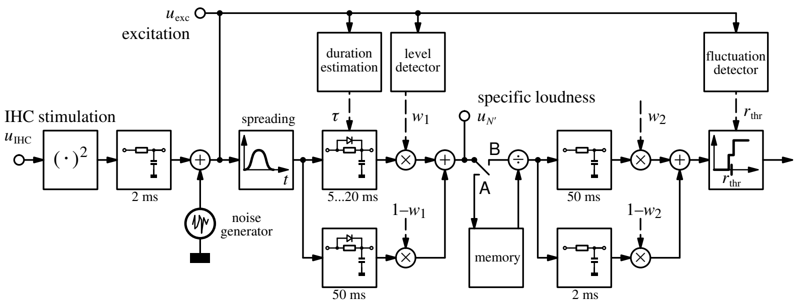

Figure 6. One unit of the central ear model. For the masked threshold generation the masker signal is first processed with the switch in position A for storing the specific loudness values. Afterwards, masker plus test signal are processed with the switch in position B so that the ratio of specific loudnesses can be evaluated. The constant weighting factor amounts to $w_2 = 0.9$.

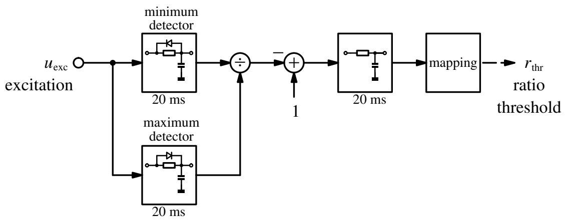

Figure 7. Fluctuation detector of each model section.

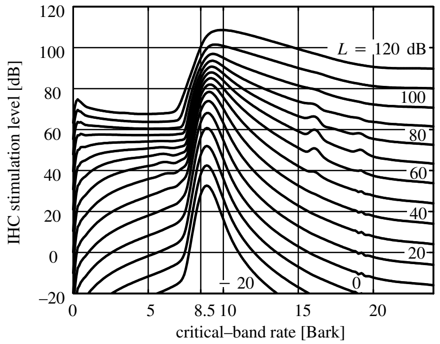

Figure 8. IHC stimulation level over critical–band rate for 1 kHz (8.5 Bark) pure tones with level *L*. The diagram equally represents the excitation patterns for the steady–state condition.

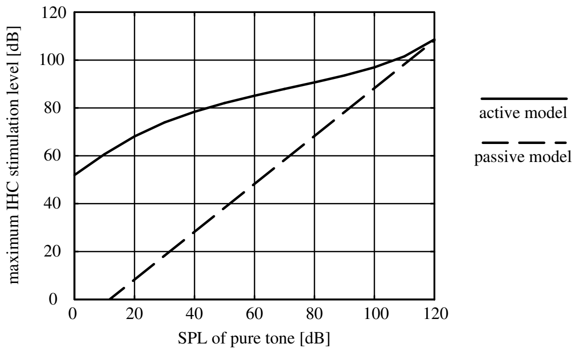

Figure 9. BM–I/O–function derived from the maxima of the IHC stimulation levels in Fig. 8. For comparison, the IHC stimulation for the passive model without amplification is shown as dashed line which also reflects the slope of a linear model.

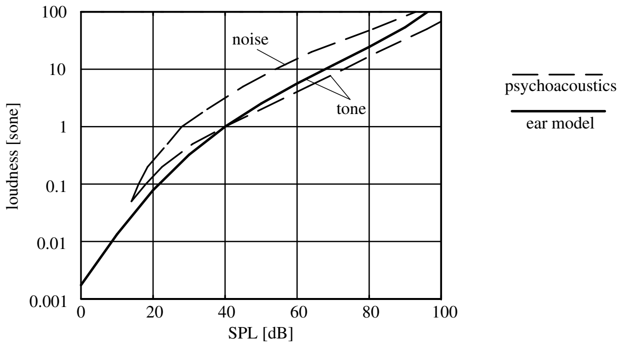

Figure 10. Generated total loudness of a 1 kHz pure tone compared to psychoacoustics (adapted from [22]).

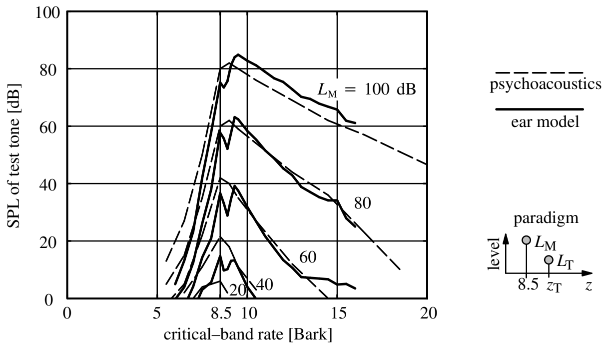

Figure 11. Generated masking patterns of 1 kHz (8.5 Bark) pure tone maskers compared to psychoacoustics (adapted from [22]).

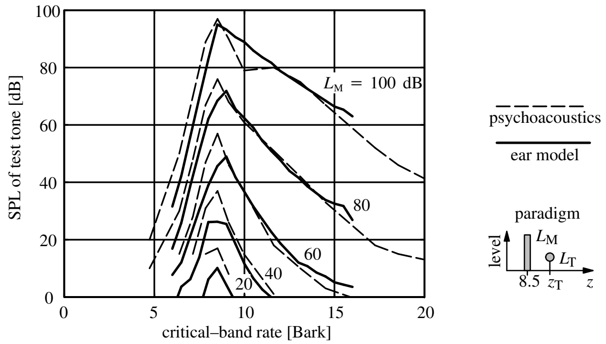

Figure 12. Generated masking patterns of critical–band wide Gaussian noise maskers centered at 1 kHz (8.5 Bark) compared to psychoacoustics (adapted from [22]).

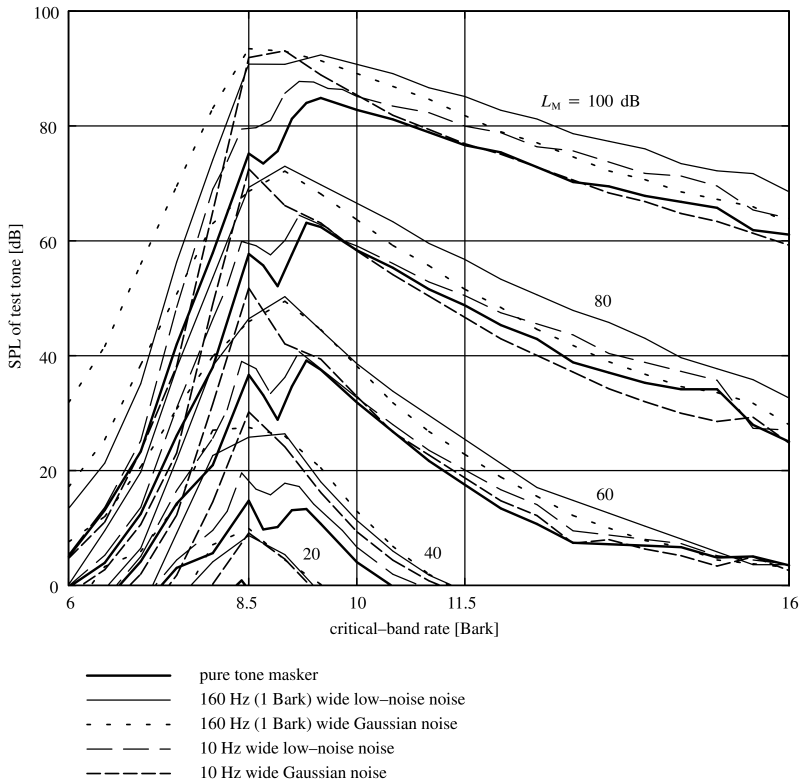

Figure 13. Generated masking patterns of pure tone, Gaussian noise, and low–noise noise maskers centered at 1 kHz (8.5 Bark).

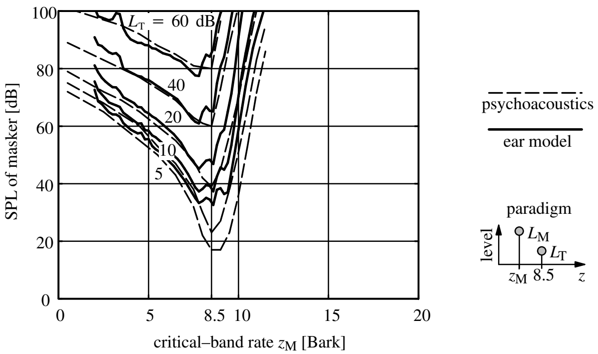

Figure 14. Generated psychoacoustical tuning curves for test tones at 1 kHz (1 Bark) compared to psychoacoustics (adapted from [22]).

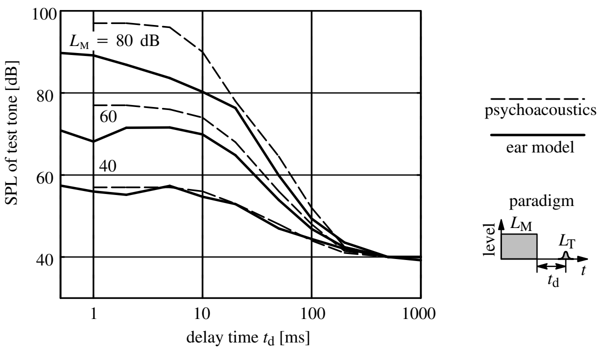

Figure 15. Generated postmasking patterns for white noise maskers and 20 µs Gaussian test pulses with peak level L_T (adapted from [22]).

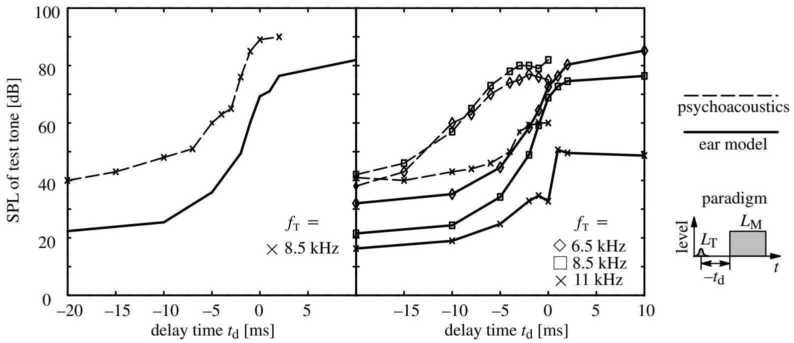

Figure 16. Generated premasking patterns compared to psychoacoustics (adapted from [47][48]).
Left: uniform masking noise masker. L_M = 60 dB SPL, test tone duration T_T = 1 ms.
Right: critical–band wide noise masker. L_M = 70 dB SPL, T_T = 1 ms.

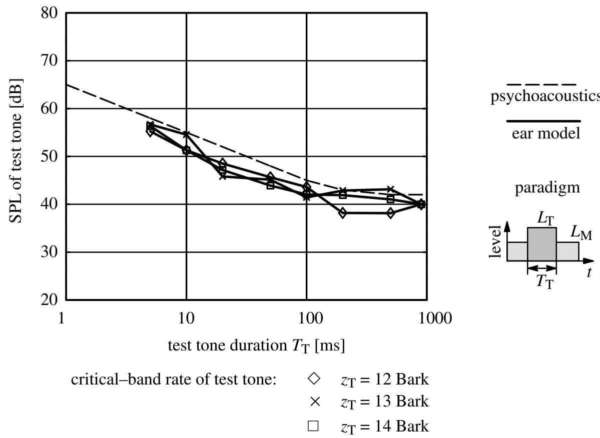

Figure 17. Generated masked thresholds for a uniform masking noise masker with L_M = 60 dB and test tones with variable duration compared to psychoacoustics (adapted from [22]).

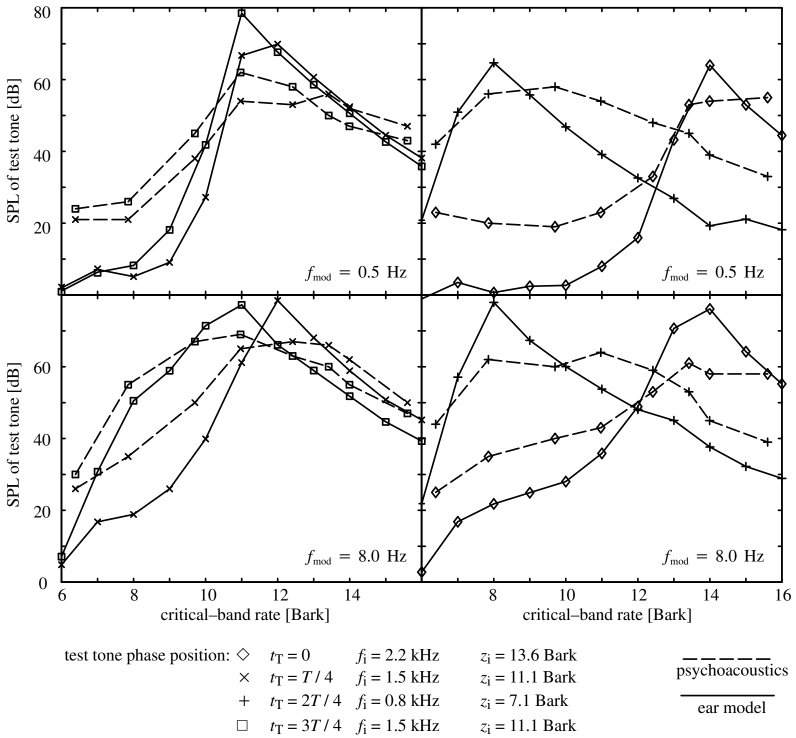

Figure 18. Masking of a frequency modulated pure tone (70 dB SPL) generated from the ear model compared to psychoacoustics (adapted from [22]). The frequency of modulation $f_{mod} = 1/T$ is 0.5 Hz for the upper and 8 Hz for the lower diagrams. The definition of the test tone temporal phase position relative to the modulation phase is outlined in Fig. 3. Instantaneous frequencies f_i and critical–band rates z_i of the masker for each phase position are listed with the associated symbol.

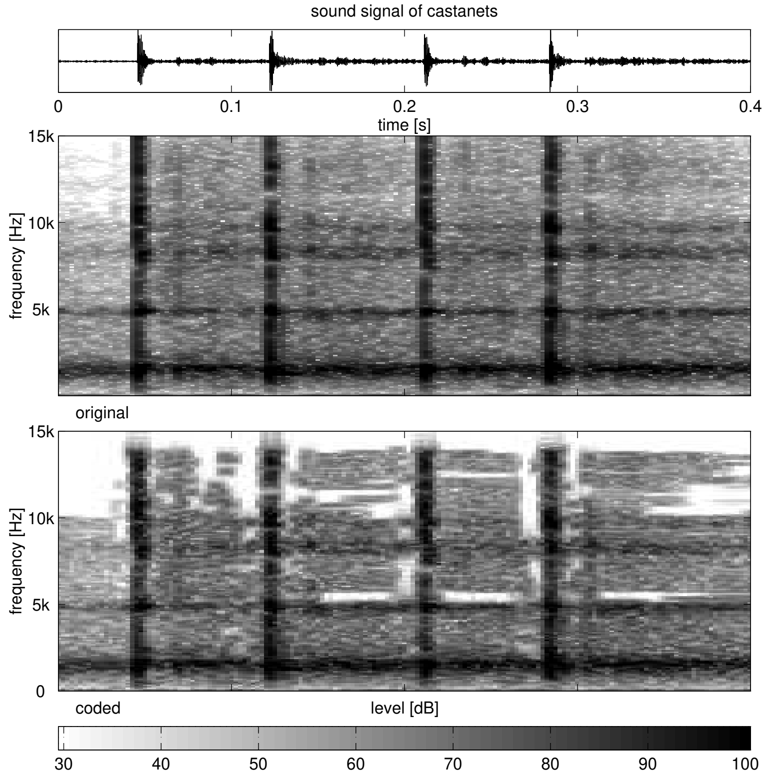

Figure 19. Spectrogram of original and coded castanet signal from the NBC–coder reference model 2, an older version of AAC, at 64 kbit/s. White areas in the lower spectrogram indicate where spectral components of the signal are omitted because they are assumed to be inaudible.

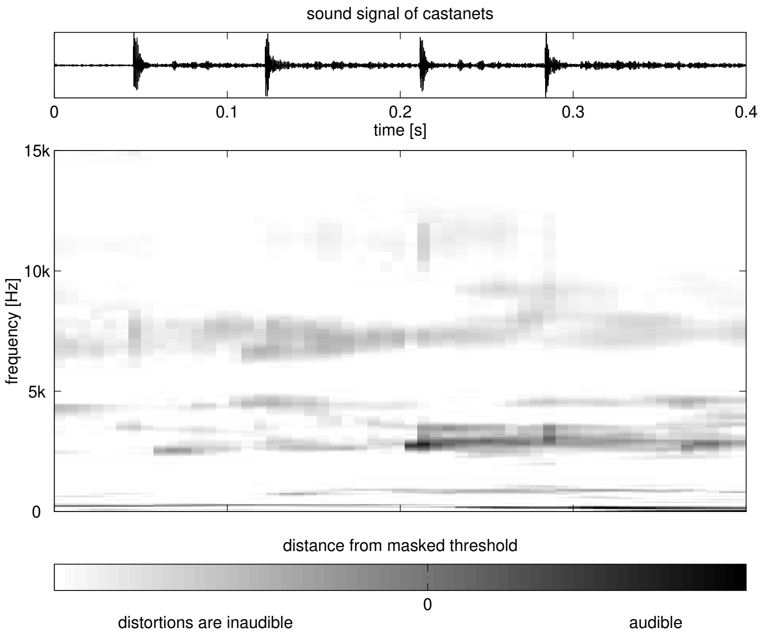

Figure 20. Distance from masked threshold of the coded signal compared to the original from Fig. 19. Audible distortions are detected by the ear model near 3 kHz for the third and fourth strike of the castanets.

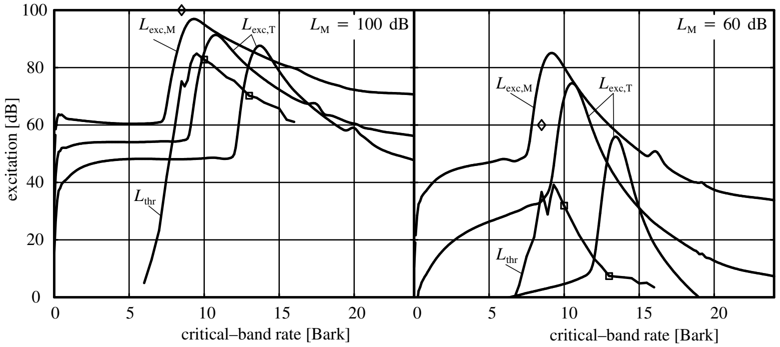

Figure 21. Relation of excitation patterns and masked thresholds. Excitation pattern $L_{\mathrm{exc,M}}$ of a pure tone masker at 1 kHz (8.5 Bark) ”◇”. Masked threshold L_{thr} and two excitation patterns $L_{\mathrm{exc,T}}$ for the indicated test tone levels ”□” at the masked threshold. *Left*: 100 dB masker SPL. *Right*: 60 dB masker SPL.