Evaluation of a Physiological Ear Model Considering Masking Effects Relevant to Audio Coding

Frank Baumgarte[1], *AES Member*

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

Abstract

A physiological ear model for improved simulation of masking effects was presented in a preceding publication. In order to demonstrate its applicability to audio coding the simulated masked thresholds are compared with subjective measurements under conditions as present in that application. This includes that narrow–band noise instead of a pure tone is employed as test signal.

0 INTRODUCTION

Audio data compression techniques for high sound quality and compression rates on the order of 1:10 compared to the compact disc (CD) format make use of a psychoacoustical model. This model is necessary for controlling the level of the introduced quantization noise according to auditory perception, so that the audibility of this distortion is minimized. A prominent example is the ISO/IEC MPEG-2 Advanced Audio Coding (AAC) standard [1]. At a data rate of 64 kbit/s per channel current implementations of this coding system achieve a sound quality comparable to CD for most assessed signals, but some very critical signals like speech remain distinguishable from the original in mono and stereo listening tests [2][3]. Sound quality improvements seem to be possible by introducing a more sophisticated perception model. The standard is open to such modifications because only the data format and decoding process is fixed but not the encoding process.

The AAC uses a time-to-frequency mapping also referred to as transform-coding technique which results in a subband signal representation. Adjacent subband signals are grouped and each group is individually quantized. The quantizer step size is controlled by the psychoacoustical model which estimates the masked threshold evoked by the audio signal. The masked threshold quantifies

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

the maximum allowable distortion power introduced by the quantizers, assuring that the distortion is just inaudible. This way, the data rate is reduced by limiting the number of quantization steps or by omitting subband signals with a power below the masked threshold.

Psychoacoustical models used in conjunction with AAC are usually based on the proposed model described in the standard [1]. The proposed model calculates the spectral signal power by means of a DFT. The spectral excitation, as approximation of the inner-ear excitation, is derived by a convolution of the power spectrum with a spreading function. The tonality of the spectral signal components is estimated by a prediction method from preceding DFTs and finally, the masked threshold is derived from the excitation level by reducing each component according to its tonality.

This threshold calculation procedure does not account for some basic properties of the auditory system which affect masking. The ear's frequency selectivity is continuously reduced for increasing frequencies, so that in turn, the temporal resolution increases with frequency. The uniform DFT spectral analysis deviates from this nonuniform filtering at low and high frequencies. The power spreading by convolution is based on the assumption of a linear system which does not account for the highly nonlinear properties of the auditory system. These nonlinear properties have relevant influence on masking in case of complex audio signals. The tonality measure is only useful for spectrally separated pure tones and it does not adequately take into consideration the temporal signal envelope fluctuations relevant to the auditory system.

Recently a physiological ear model was presented in [4][5] showing that these drawbacks of the psychoacoustical model proposed in the standard can be avoided by a more sophisticated approach on a physiological basis. First simulation results in these publications demonstrate that the model allows for a reproduction of physiological and psychoacoustical results from measurements. However, the masking simulations were only carried out at medium frequencies and with a pure-tone test signal. These results were achieved because the model offers adequate frequency selectivity and temporal resolution as well as a tonality measure based on the temporal signal envelope. A simulation of nonlinear masking effects is enabled due to the nonlinear elements contained in the ear model.

This paper is focused on a comparison of masking simulation results with psychoacoustical measurements for pure-tone maskers in a wide frequency range and test signals consisting of a pure tone or a narrow-band noise. For collecting masking data for the narrow-band noise a psychoacoustical test was carried out. This test signal was utilized because it models the quantization noise of a subband coder better than a pure tone and allows for an improved model verification for this application. The simulation results consider the detection of difference tones occurring in the inner ear and allow for the so called "off-frequency listening". Furthermore, the increased masked thresholds for combined narrow-band-noise maskers were investigated in two simple simulations which address the additivity of masking. Also simulation results for postmasking and for the detection threshold for amplitude modulation (AM) and frequency modulation (FM) are presented. The most relevant masking effects with respect to audio coding are simultaneous masking and the additivity of masking in case of complex audio signals. Temporal masking effects are important for signals with fast temporal variations, while AM or FM detection thresholds are only relevant to audio signals containing AM- or FM-like components.

This paper is organized as follows. Chapter 1 contains some basics of sound processing in the auditory system and may be skipped by readers familiar with this subject. In chapter 2 the structure of the physiological ear model is reviewed. Simulation results are presented together with the corresponding psychoacoustical measuremental results in chapter 3 to verify the model. Conclusions are drawn in chapter 4.

1 AUDITORY SOUND PROCESSING

The basic auditory masking effects can be explained by known properties of the sound processing in the auditory system. These properties are usually derived on the one hand by investigating the underlying physiological processes and on the other hand by psychoacoustical measurements which allow to describe the relationship between physical and perceptual quantities by mathematical models. While results from psychoacoustical measurements are given in chapter 3, some basics of auditory system physiology are summarized here.

The auditory system is functionally subdivided into a peripheral and a central part. The boundary is assumed to be located at the initial level of neural activity, the generation of electrical impulses (action potentials) in individual fibres of the auditory nerve by the inner hair cells (IHCs). This boundary represents not a purely anatomical division, it also differentiates between the two basic sound signal representations: a mechanical representation in the periphery and an electrical representation in the central part. The physiological properties of the periphery are known, even if underlying details are still subject to debate. In contrast, sound processing in the central auditory system is widely unknown with respect to a physiological point of view. This is why psychoacoustical measurements are necessary in order to derive input-output functions of the central part. A description of the central auditory system is omitted here, because anatomical details in that region are not relevant for the understanding of the ear model.

1.1 Peripheral Auditory System

The peripheral auditory system is further divided into outer, middle, and inner ear, as partly shown in Fig. 1. The outer ear consists of a cavity open to the external sound field and closed by the eardrum. The middle ear contains three coupled ossicles (small bones), which provide a mechanism for matching the impedance at the eardrum with that of the fluid of the inner ear. Pressure variations in air travel through the outer ear to the eardrum, where the vibration is transmitted via the ossicles to the oval window of the inner ear. The cochlea of the inner ear comprises the sound processing and transducing organ which provides a sound signal analysis resulting in a spectro-temporal activity pattern. The cochlea is a coiled, fluid filled tube, divided lengthwise by the cochlear partition into two mechanically effective chambers. This partition contains as main parts the basilar membrane (BM) and the organ of Corti attached to the BM, as shown in Fig. 2. The width and compliance of the BM increase with its length, with the widest and most flexible part being furthest away from the oval window at the helicotrema. Vibrations of the oval window create pressure variations in the fluid surrounding the BM and cause transverse vibrations of the BM. For a single frequency stimulation,

starting from the oval window the amplitude of BM velocity builds up to a maximum and then decreases rapidly, as shown in Fig. 3. The position of the maximum depends on the stimulation frequency. The resonance position for a high frequency is nearer to the oval window than for a lower frequency. In this way, the envelope of BM velocity represents a spectral analysis of the signal, with frequency converted to place information.

The organ of Corti converts the BM velocity into electrical impulses on the auditory nerve by means of the inner hair cells. In addition to the place of activity, the amount of activity in terms of the number of action potentials per time and the temporal envelope and fine structure of the impulse patterns makes up the coded signal information in each fiber of the auditory nerve.

The outer hair cells (OHCs) inside the organ of Corti support the adaptation of the input signal dynamic range to the limited dynamic range of the IHCs. The assumption is widely accepted that small BM vibrations are mechanically amplified by the OHCs while large vibrations are not influenced. The OHCs act more or less independently at each place along the BM, so that a frequency selective amplification of spectral components with low levels is achieved. This nonlinear mechanical amplification provides a uniform explanation for many nonlinear effects in hearing and especially in masking.

2 STRUCTURE OF THE PHYSIOLOGICAL EAR MODEL

As the auditory system is divided into anatomical or sound processing units, this structure is reflected by the model as outlined in the block diagram of Fig. 4. The peripheral auditory system is represented by a model of the outer and middle ear (OME) which is connected to the cochlear model containing a basilar membrane (BM) model, outer hair cell (OHC) models, and inner hair cell (IHC) models. The central auditory system (CAS) models are fed by the output of the inner hair cell models. The masked threshold for a test signal in the presence of a masking signal is generated in the model by first calculating the specific loudness of the masker only, and in a second step for the masker with the test signal added. The specific loudness is defined as the spectral distribution of the perceived total loudness. The amount of change in specific loudness due to the test signal is assessed afterwards by the CAS model in order to decide whether the test signal is audible.

2.1 Model of Peripheral Auditory System

The model of the peripheral auditory system is specified in terms of an electrical network using the correspondence of velocities and forces in an acoustical or mechanical system to voltages and currents respectively in an electrical system. The electrical network is simulated in the time domain by wave digital filters [6].

The OME model is designed as a linear prefilter for the incoming sound signal with the electrical network structure given in Fig. 5. The magnitude frequency response shows increased damping at a rate of 20 dB/decade for decreasing frequencies below 1 kHz and a resonance at 3 kHz which rebuilds the low frequency damping of the middle ear and the first ear canal resonance of the outer ear. The acoustics of the outer and middle ear was not considered for frequencies higher than 3 kHz because numerous resonances occur in this range which deviate interinvidually so that an adaptation is only useful for a single individual. As an average for high frequencies a constant magnitude response is assumed.

The output signal of the OME model is the velocity of the oval window at the entrance of the cochlea corresponding to the voltage u_{OW}. For one section of the cochlear model the BM model and the OHC model are shown in Fig. 6, the IHC model is shown in the left part of Fig. 7. The BM model rebuilds the hydromechanics of the cochlear duct and the cochlear partition [4]. Each section of the BM model rebuilds the hydromechanics of one of the slices created by dividing the cochlear duct into parts of equal length. A number of 251 sections was used to represent the cochlear hydromechanics from the oval window to the helicotrema. The model in [4] was extended by 10 sections at the oval window to allow for simulations at frequencies higher than 16 kHz. Each model section contains a parallel resonance circuit and corresponds to a BM location with its characteristic resonance, the characteristic frequency. The additional capacitor and resistance in each section models the mechanical coupling and damping by the fluid. The voltage u_{BM} represents the transverse BM velocity which is amplified by g_1 in the OHC model inside a positive feedback loop with continuous saturation for increasing amplitudes. At low signal amplitudes the nonlinearity causes a high u_{BM} gain for frequencies near the section resonance frequency compared to a system without feedback. In addition, the anti-symmetric nonlinearity produces odd order distortions. A second amplification stage is added to achieve a resulting maximum amplification of 60 dB at the output of each section, which is a value found in BM measurements from animals [7][8][9]. The second stage amplifies the feedback current i_f by g_2. The current source is loaded by a parallel resonance circuit identical to the circuit in the BM model of the same section so that the same frequency dependence for the associated voltage is obtained by the current i_f and the amplified current $g_2 i_f$ flowing through the corresponding parallel resonance. The second stage outside the feedback loop was slightly modified compared to [4]. In this network structure the additional stage is necessary to maintain the stability of the system for such a high amplification. The amplifier in the feedback loop is only stable up to a gain of approximately 40 dB. Neighboring OHC models are coupled via resistors with increasing resistances for increasing section distances. Each section is coupled with 16 sections to either side which is only partly outlined in Fig. 6.

The second amplification stage generates the stimulation signal u_{IHC} for the IHC model in each section. The IHC model consists of a square value calculation and a low-pass filtering as shown in Fig. 7. These two elements approximate the half-wave rectification and the low-pass character as basic properties of the IHCs. Despite the very limited dynamic range of single IHCs a saturation is not included. The IHC model rather combines different types of IHCs in the cochlea which act at different signal levels and are assumed to cause no effective dynamic compression as a whole. The IHC model output signals represent the degree of neural activity of the auditory nerve fibers connected to the IHCs at the BM location corresponding to the model section. Neural activity can be interpreted as the spatial and short-term averaged number of action potentials (nerve firings) per time interval.

2.2 Model of Central Auditory System

The physiological processes in the auditory pathways and cortex are widely unexplored, so that a psychoacoustical modelling approach is appropriate for the central auditory system. A simplified block diagram showing the most important functional units of the model is given in the right part of Fig. 7. Each block in the diagram accounts for the proper modelling of a specific masking effect, as explained in the following.

Internal noise is added to the input activity signal, limiting the audibility of signals below the threshold in quiet. A temporal delay spread function is applied afterwards to account for the limited masked threshold rise time in premasking. The delay spread may be attributed to different signal propagation times on the auditory pathways. A decay function, which is adapted to the masker level and duration, models the most important properties of postmasking and may be attributed to the recovery time of nerve cells after stimulation. The output signal at this stage is assumed to represent specific loudness [10]. A leaky integrator, realized as low-pass filter, is applied to the specific loudness change caused by the added test signal. This short-time integration reflects the dependency of masking on the test signal duration, e.g. a brief test signal is more difficult to detect than a test signal of the same level and longer duration. The threshold detector is fed by the integrator output. Each time its input value exceeds the threshold, the test signal is assumed to be audible. The change detection is independently applied in each section of the model in contrast to [4], where the detection was only applied in sections near the test signal resonance location. The threshold value in each section is adapted according to the envelope fluctuation of the neural activity signal with noise added. The adaptation is necessary for the consideration of the different masking effectivity of e.g. Gaussian noise with high fluctuation and a pure tone with disappearing fluctuation. More details of the central auditory system model are given in [4].

3 RESULTS

In comparison to the first modelling results mainly for medium frequencies reported in [4], loudness simulations and masking simulations for signals covering most of the audible frequency range will be presented here. This includes one experiment with narrow-band noise as test signal which better models quantization distortions in a subband coder than a pure tone usually employed in psychoacoustical tests. The properties of the additivity of masking for combined maskers were examined by two simple experiments which were setup corresponding to psychoacoustical tests described in the literature. Additional results for other signal configurations are included to demonstrate that the model rebuilds postmasking effects and is also applicable to modulation detection. The simulation results are presented graphically together with psychoacoustical data for the verification. The masked thresholds for narrow-band noise were collected in a separate psychoacoustical test carried out at our institute since data for this test signal type are rarely available from the literature.

3.1 Total Loudness of a Pure Tone and Wide-Band Noise

Total loudness is calculated by adding up the specific loudnesses in all model sections and normalizing the result so that a pure tone at 40 dB sound pressure level (SPL) yields a loudness of 1 sone.

Fig. 8 shows the total loudness for a pure tone and uniform-exciting noise, i.e. a colored wide-band noise defined in [10]. In correspondence to psychoacoustical measurements the simulation results show an increasing deviation of tone and noise loudnesses and a decreasing steepness of the loudness function with increasing levels. However, the steepness of the loudness function from the model is too high so that the dynamic compression appears to be insufficient for a precise loudness simulation. This deviation results from the fact that the model parameters were not adjusted to produce a minimized error in loudness simulations. An increased compression can be easily incorporated in the model by increasing the gain g_2 in the second amplification stage.

3.2 Simultaneous Spectral Masking

If the masker and test signal are presented simultaneously, nonlinear distortion products of the combined signals are created in the cochlea. These distortions are easily audible if their spectrum contains large components below the masker frequency because the masker is more effective at frequencies above the masker frequency than below. A test signal may be detected in the presence of a masker signal by listening at the frequency of the test signal or by listening at the frequency of a distortion product, which is called "off-frequency listening" in masking experiments. Off-frequency listening is usually avoided in masking experiments by an additional noise masker that masks the distortion products, or can be recognized by trained listeners who are able to differentiate between the sensation of a distortion product and the test signal.

In case of audio coding a listener usually cannot distinguish between quantization distortions and possible distortion products of quantization distortions and the audio signal. From this point of view, measured masked thresholds without off-frequency listening restrictions should actually be applied. This problem is less important in subband coding, if a sufficiently narrow bandwidth of the subbands is maintained so that the quantization distortion frequencies are always close to the frequency of the corresponding signal component. This assures that distortion products resulting from the interaction of spectral audio signal components with quantization distortions can be neglected in comparison to distortion products from the interaction between different spectral audio signal components, if the quantization noise level is below the signal level.

The prevention of off-frequency listening is rebuild in the CAS model by only evaluating the thresholds in model sections with characteristic frequencies near the test signal frequency, which was always the case for the results in [4]. The evaluation of the thresholds in all model sections enables off-frequency listening and is the default setting here.

3.2.1 Threshold in Quiet

Fig. 9 shows the simulated threshold in quiet for a pure tone and different psychoacoustical results. Its shape is widely determined by the OME model for frequencies up to 1 kHz (8.5 Bark). The increased threshold at high frequencies is caused by the cochlear model which shows a reduced effective amplification for very high frequencies. The absolute threshold value is determined by the variance of the internal noise, which is identical in each model section. It is adjusted so that the simulated threshold is always below the psychoacoustical reference values in order to assure that the threshold in quiet of a very sensitive subject will be not exceeded. The local maximum of the

measured threshold in quiet in the range of 6 kHz (20 Bark) and the minimum near 10 kHz (22 Bark) are not present in the simulation result because the outer and middle ear resonances in this range are not taken into account.

3.2.2 Pure Tone Masked by a Pure Tone

Masked thresholds for tone-on-tone masking are shown in Fig. 10 for the simulation and for psychoacoustical measurements. The masker frequencies were selected in order to evaluate the model performance in a wide range of audible frequencies. The lowest masker frequency is 250 Hz (2.5 Bark) while the highest amounts to 6 kHz (19.7 Bark). Apart from the 2-kHz (13 Bark) masker, three masker levels were used and compared with psychoacoustical data. It turns out that the simulated masked thresholds match closely the measured data near the masker frequency at masker SPLs of 60 and 40 dB. At 20 dB SPL the masked thresholds are underestimated partly due to the lower threshold in quiet of the simulation compared to the psychoacoustical data.

The simulated masking patterns have approximately symmetrical shape with respect to the masker frequency. The audibility of distortion products is responsible for the observed steep slope of the masked threshold for test signal frequencies higher than the masker frequency. The most intense distortion product in tone-on-tone masking is the cubic difference tone at $f_3 = 2f_\mathrm{M} - f_\mathrm{T}$ with the masker frequency f_M and test tone frequency f_T. Masking patterns measured with disabled off-frequency listening have a shallow slope for test-tone frequencies higher than the masker frequency and the same steep slope as in Fig. 10 for test-signal frequencies below. The difference-tone detection is overestimated by the model which results in lower masked thresholds than psychoacoustically measured in the frequency range where the difference tone is detected at a lower test signal level than the test signal itself. This is most obvious for the 2-kHz (13-Bark) masker and higher masker frequencies at 60 dB SPL.

3.2.3 Critical-Band-Wide Noise Masked by a Pure Tone

The verification of the model with respect to the intended application in masked threshold generation for quantization noise of a coder can be improved by using an appropriate test signal. In psychoacoustical tests usually a pure tone is chosen as test signal, which is not well suited for the approximation of the noise-like quantization distortions. Only one publication [11] is known to the author, where a narrow-band test noise is employed. For this reason, a psychoacoustical test was performed with the aim of measuring the masked thresholds for a narrow-band-noise test signal in the presence of a pure-tone masker in different frequency ranges. The bandwidth of the noise was adjusted equal to the critical bandwidth [10] at the center frequency of the noise and it was generated by band-pass filtering of white Gaussian noise. In the test, a sequence of masker, break, and masker plus test signal was presented twice via headphones. The durations were 1 s for each signal and 0.2 s for the break. The test signal level was adjusted randomly by the signal generator and the subjects had to judge whether any change between the first and second signal in the sequence was audible. Because no additional masker was added to disable the detection of distortion products, the result is also influenced by off-frequency listening.

For each sequence the masker and test signal were generated by appropriate level adjustment from the same stored reference signals. In order to avoid an additional variance of the measured masked thresholds due to different test noise signals, the frequency dependent masking pattern was measured by varying the masker frequency while using a test noise signal fixed in frequency. This way, masking patterns were obtained for critical-band test noises at frequencies of 250 Hz (2.5 Bark), 1 kHz (8.5 Bark), 2.5 kHz (14.5 Bark), and 7 kHz (20.5 Bark). The measured masked thresholds of maskers with 80 dB SPL from 8 subjects with normal hearing are given as median values and interquartile ranges in Fig. 11. The median indicates the value below which half of the measured data points lie. The interquartile range encompasses 50% of the data points so that 25% of the measured points have higher values and 25% have lower values than the interquartile range.

The perception of distortion products appears again in this test for test signal frequencies higher than the masker frequency. In contrast to the tone-on-tone masking in Fig. 10 the reduced masked thresholds due to off-frequency listening are located on the left side of each masking pattern in Fig. 11. The corresponding simulation results show comparable thresholds for masker frequencies near and above the test signal frequency. As in tone-on-tone masking the detection of distortion products is overestimated by the model. This overestimation is no major problem, if the masked thresholds are applied in subband coding, as explained above.

Fig. 12 shows the same psychoacoustical data as Fig. 11 together with simulation results with disabled distortion-product detection. As expected, the simulated masked thresholds in the frequency range of possible distortion-product detection are considerably increased compared to Fig. 11. For low and medium test signal frequencies the measured masked thresholds in this range are significantly below the simulation results. For comparison, the masked threshold generated by the proposed basic psychoacoustical model of the AAC is also shown in Fig. 12. This threshold was calculated without using a DFT analysis directly by applying the spreading function to the tonal masker at the known masker frequency. It should be noted, that this threshold may significantly differ from the thresholds which are calculated and used in current optimized AAC encoder implementations.

The psychoacoustical test with the results shown in Fig. 11 was additionally performed with the same conditions for masker levels of 60 and 40 dB SPL in order to investigate the level-dependency of masking. At 60 dB 3 subjects and at 40 dB 2 subjects participated. The results are plotted in Fig. 13. The deviation of the simulation and measuremental results appear similar to data in Fig. 11 for the 80-dB maskers so that the interpretation of the 80-dB data holds also for lower masker levels. However, at low masker levels off-frequency listening can be used to a small extent since the distortion products are near or below the threshold in quiet. For this reason, the masked threshold patterns in Fig. 13 have no marked minima in contrast to Fig. 11. It is also apparent from Fig. 13 that the simulated masked thresholds for the 2.5-Bark maskers deviate from the psychoacoustical measurement to a larger extent than for a masker at 80 dB SPL. Further simulations are necessary in order to explain this deviation.

3.2.4 Additivity of Simultaneous Masking

The masked threshold of more than one simultaneous masker is compared to the masked threshold of each single masker in experiments with the aim of investigating the properties of masking additivity. For example, if an intensity summation model is assumed for the additivity of masking,

the predicted masked threshold increase at a frequency where two single maskers produce the same masked threshold is 3 dB for the threshold of the combined maskers. In psychoacoustical measurements it is observed, that the threshold from two combined maskers can be 13 to 20 dB higher than the maximum threshold of one of the two single maskers [12]. This masking increase is near to the higher value (20 dB) for maskers which have a spectral distance of more than one Bark [10] and vice versa.

The first simulation concerning the additivity of masking was carried out for two different conditions which are outlined in Fig. 14. In the first condition (left panel) the masked thresholds of a 1-kHz and a 2.1-kHz Gaussian noise of 50 Hz bandwidth was generated for each single masker and their combination. The 1-kHz masker was replaced by a 1.8-kHz masker in the second condition (right panel), so that both masker center frequencies had a distance of only 1 Bark. The results are plotted in Fig. 14 for a frequency range where the masking effectivity of the single maskers is comparable and where the maximum increase in masking is expected for the combined masker. The maximum masking increment ΔL_{T} in the simulation for the first condition is 22.6 dB and for the second 13.3 dB. This result shows that the considerable masking excess compared to a linear intensity summation can be demonstrated by the model and that higher masking increments are generated for maskers with a larger frequency distance which is consistent with psychoacoustical results.

In a second experiment, the additivity of masking was simulated according to a psychoacoustical test reported in [13]. This test demonstrates that the additivity of masking also depends on the temporal envelope of the masker which consisted of two narrow-band noises at 1.4 kHz and 1.6 kHz, each with a bandwidth of 50 Hz, and a pure-tone test signal at 2 kHz. In the first condition the two noise maskers were uncorrelated while in the second, the noises had identical temporal envelopes. The masker level for each single masker was adjusted in the psychoacoustical experiment and in the simulation, so that equal masked thresholds are obtained at the test signal frequency before the maskers were combined. In one example of the psychoacoustical experiment [13], a masked threshold of 35 dB SPL at 2 kHz was achieved for a masker level of 78 dB for the 1.4-kHz noise or 70 dB for the 1.6-kHz noise (the same level for correlated and uncorrelated noise). The combined masker produced a masked threshold of 49 dB for the uncorrelated noises and 39 dB for the correlated noises which is equal to a masking increment of 14 dB and 4 dB respectively. In the simulation, a masked threshold of 40 dB SPL results from an 1.4-kHz noise at 74.5 dB, an uncorrelated 1.6-kHz noise at 70 dB, or a correlated 1.6-kHz noise at 67.5 dB SPL. The threshold for the combined uncorrelated maskers amounts to 49.7 dB and for the combined correlated maskers to 41.9 dB SPL. The calculated masking increment is 9.7 dB for the correlated and 1.9 dB for the uncorrelated masker from the simulation.

The latter masking simulation confirms the suggestion in [13], that the test signal is detected in temporal minima of the masker envelope. Since these minima are identical for the combined correlated masker compared to each masker component, the masked threshold increase is on the order of intensity summation (3 dB per intensity doubling). The minima of the uncorrelated masker envelope are reduced, because the envelope minima of a single masker can only be preserved, if the superimposed second masker has the envelope minima at the same temporal position.

The simulation results for the additivity of masking are important for the generation of masked thresholds of arbitrary audio signals which usually consist of a superposition of a variety of different simple masker signals. The highly nonlinear additivity of masked thresholds observed for single maskers according to [12] is a masking effect which is also present in the simulation results shown in Fig. 14 indicating that the ear model is suitable for masked threshold generation of complex audio signals.

3.3 Postmasking

Apart from the test signal delay after masker offset, the model considers explicitly the parameter-dependency of postmasking from masker level and masker duration. Level-dependent postmasking patterns are shown in Fig. 15. The model parameters were adjusted so that the level-dependent decay slopes which are approximately linear on a logarithmic time scale are obtained. The test signal used for the psychoacoustical test and the simulation was a very brief click consisting of a 20-μs Gaussian pulse.

Postmasking patterns for a 2-kHz, 5-ms tone burst are shown in Fig. 16. In this case the masker is a uniform-exciting noise with a duration of 200 ms or 5 ms. For test signal delays of more than 10 ms the threshold differences observed for the two masker durations are on the same order from the model. For longer delays the threshold in quiet is reached which is adjusted lower in the model than measured, as previously shown in Fig. 9. At small delay times interactions of the probabilistic masker with the test tone may cause threshold variations. These variations could be lowered by averaging the thresholds of simulations with independent noise masker signals. However, the simulated threshold was obtained from only a single masking noise.

3.4 AM and FM Detection Threshold

The detection thresholds of amplitude and frequency modulation can be explained by combined spectral and temporal masking. The thresholds are shown in Fig. 17 for an amplitude and a frequency modulated 1-kHz tone at different modulation frequencies. The magnitude spectrum of the AM signal closely matches the FM signal spectrum at small modulation degrees (AM) or indices (FM) less than 30% respectively. The spectrum consists of three significant components: one spectral line at the carrier frequency and spectral side lines at the sum and difference of carrier and modulation frequency. The difference between both signals is due to the phase which is 90° shifted for the side lines of the FM compared to the AM spectrum. The modulation index for FM is calculated as $\eta = \Delta f / f_{\mathrm{mod}}$ with the maximum frequency deviation Δf of the FM signal with respect to the carrier frequency and the modulation frequency f_{mod}. The degree of modulation m is the maximum relative amplitude increment in AM with respect to the unmodulated signal.

The simulated detection thresholds show the pronounced difference between AM and FM detection for low modulation frequencies as observed in psychoacoustical tests. At higher modulation frequencies, the FM detection threshold approaches the threshold for AM which starts to decrease for increasing modulation frequencies, because the modulation perception vanishes and the side lines become audible as separate tones. This point is reached in the psychoacoustical experiment for a modulation frequency of 75 Hz indicating the half of the critical bandwidth of the auditory filter at 1 kHz.

The different thresholds of FM and AM at low modulation frequencies occur in the simulation because the modulated amplitude of the AM signal leads to excitation patterns with a level dependent slope steepness as shown in [4]. The steepness is increased for lower levels so that the excitation amplitude variation increases at the far end of the excitation slope. Therefore, the amplitude modulation is detected at a location, where a signal with significantly higher frequency than the carrier would have the main excitation. The level-dependent slopes cannot be used for an improved detection of FM due to the constant signal envelope which causes the threshold difference of AM and FM.

4 CONCLUSIONS

A physiological ear model was evaluated by comparing psychoacoustical measurements with simulation results for masking effects relevant to audio coding such as simultaneous masking, postmasking, and modulation detection. It turns out that the simulated masked thresholds are in agreement with psychoacoustical measurements for a wide range of audible frequencies and test conditions, which extends the first results for medium frequencies and pure-tone test signals in an earlier publication [4].

In subband audio coding the quantization noise should be masked by the audio signal. The masked signal in this configuration can be represented as a narrow-band-noise test signal in masking experiments, which are of basic interest in audio coding. Since such masked threshold data are rarely available from the literature, a psychoacoustical test was carried out for this type of test signal and a pure-tone masker. In simultaneous masking with narrow-band-noise or pure-tone test signals it should be taken into account that the results can be significantly influenced by audible distortion products of the inner ear which are prominent for test signal frequencies higher than the masker frequency. Distortion products with frequencies lower than the masker frequency are often masked less than the test signal and thus have a reduced masked threshold.

The results from psychoacoustical measurements of simultaneous masking with noise and pure-tone test signals were satisfactory reproduced by the ear model. The detection of distortion products, which is called ”off-frequency listening” in masking experiments, is represented by the model. However, the masked threshold reduction due to off-frequency listening is overestimated by the model. In case of subband coding, this overestimation of distortion-product detection can be assumed as a minor effect, because distortion products generated by the nonlinear interaction of the audio signal with quantization noise are likely to be masked by the distortion products of the audio signal itself.

Other known masking experiments for combined maskers show increased thresholds compared to the masker components. In most cases the threshold increase is significantly higher than predicted from an intensity summation of the thresholds of each single masker component. This excess masking is also present in the modelling results for two masker configurations. From psychoacoustical tests it is known that excess masking vanishes if the two masker components of a combined masker have the same temporal envelope. This property, which is an exception from the nonlinear additivity of masking, is replicated by the ear model.

Simulation results for postmasking show that the dependency on masker level and duration as well as the dependency on test signal duration are considered by the ear model and comparable to psychoacoustical measurements. The combined effect of simultaneous and postmasking was explored by simulating modulation thresholds. At low modulation frequencies the ear is more sensitive to amplitude modulation than frequency modulation. This property and a far reaching quantitative agreement is achieved by the simulation of the detection thresholds.

The simulation results suggest that the physiological ear model is applicable to arbitrary audio signals. This follows from the validity of the generated masked thresholds for signal levels and frequencies covering the hearing area, for combined maskers, for temporal masking, and modulation detection. Therefore, the ear model meets the requirements for the intended application in audio coding.

The ear model, as described above, currently only allows the masked threshold generation for a single pre-defined test signal, usually with narrow bandwidth. This procedure results in one masked threshold value, i.e. the level at which the test signal is just audible. Therefore, the model needs to be extended to generate a time- and frequency-dependent masked threshold as required for quantizer control in perceptual audio coding.

5 ACKNOWLEDGMENT

This work was supported by the ”Deutsche Forschungsgemeinschaft” (DFG) as part of a joined project titled ”Source coding for the transmission of stereo sound signals at 128 kbit/s”.

6 REFERENCES

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

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

[3] *ISO/IEC JTC1/SC29/WG11 Doc. M2951*; ”Report on the MPEG-2 AAC Stereo Verification tests”. ISO/IEC, 1998.

[4] Baumgarte, F.; ”A Physiological Ear Model for Auditory Masking Applicable to Perceptual Coding”. *Presented at the 103rd AES Convention*, Preprint no. 4511, New York, Sep. 1997.

[5] Baumgarte, F.; ”A Physiological Ear Model for Specific Loudness and Masking”. *Proc. Workshop on Applications of Sig. Proc. to Audio and Acoustics*, New Paltz, Oct. 1997.

[6] Fettweis, A., ”Wave Digital Filters: Theory and Practice”, *Proc. IEEE Trans. ASSP*, Vol. 23, 1975, pp. 516–525.

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

[8] 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.

[9] 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.

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

[11] Schroeder, M. R.; Atal, B. S.; Hall, J. L.; ”Optimizing digital speech coders by exploiting masking properties of the human ear”, *J. Acoust. Soc. Am.*, Vol. 66 (6), pp. 1647–1652, Dec. 1979.

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

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

[14] E. Zwicker, E.; Jaroszewski, A.; ”Inverse frequency dependence of simultaneous tone-on-tone masking patterns at low levels”, *J. Acoust. Soc. Am.*, Vol. 71 (6), pp. 1508–1512, Jun. 1982.

[15] Fastl, H.; Baumgartner, H.; ”Ruhehörschwellen für Sinustöne bzw. Schmalbandrauschen”, (German), *Acustica*, Vol. 34, pp. 111–114, 1975.

7 FIGURES

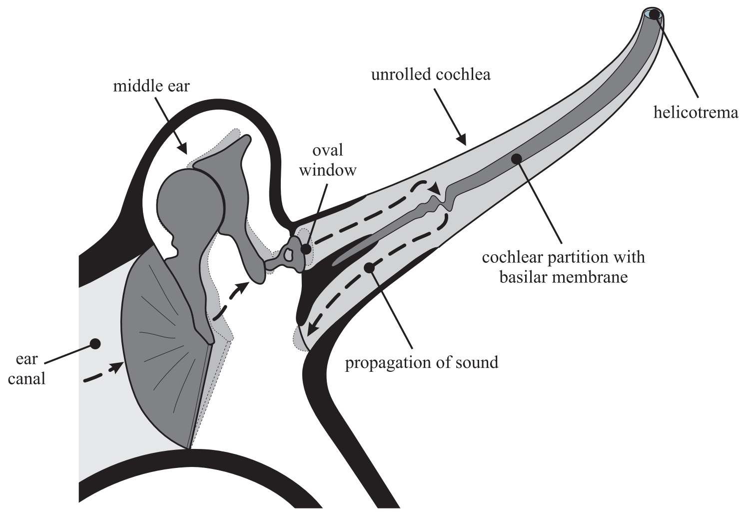

Figure 1. Sound propagation in the peripheral auditory system. Pressure variations in air travel through the ear canal to the eardrum. The eardrum displacement is coupled to the oval window via the middle ear ossicles. The displacement of the oval window results in a corresponding displacement of the round window since the fluid inside the cochlear duct consisting of a stable bone is incompressible. For a given stimulus frequency the fluid displacement in the transverse direction is maximum at a certain resonance location of the cochlear partition.

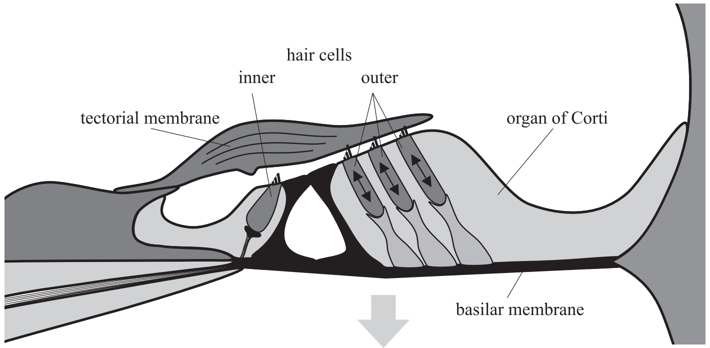

Figure 2. Cross-section of the cochlear partition with organ of Corti, tectorial membrane, and basilar membrane (BM) as main parts. A transverse displacement (gray arrow) of the BM results in a stereociliae (small hairs on top of each hair cell) displacement of the inner hair cells (IHCs) which transduce the mechanical motion signal into electrical action potentials.

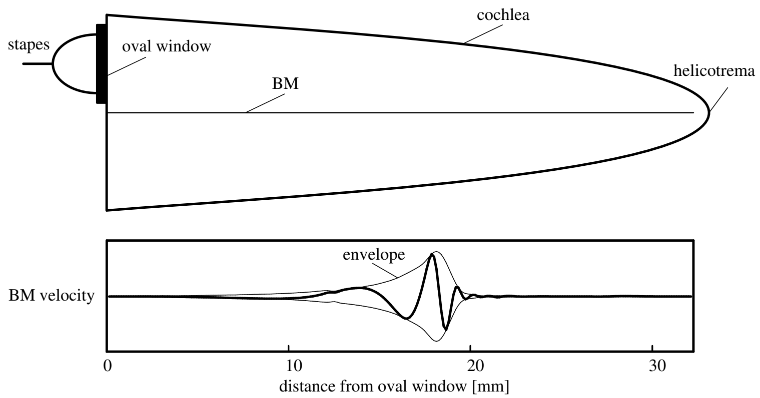

Figure 3. *Upper diagram*: cross-section of the unrolled cochlea with oval window at the entrance from the middle ear, and the basilar membrane (BM) dividing the cochlear duct into two chambers. *Lower diagram*: instantaneous transverse BM velocity at each location from the oval window to the helicotrema for a 1270-Hz tone with the calculated envelope (simulation).

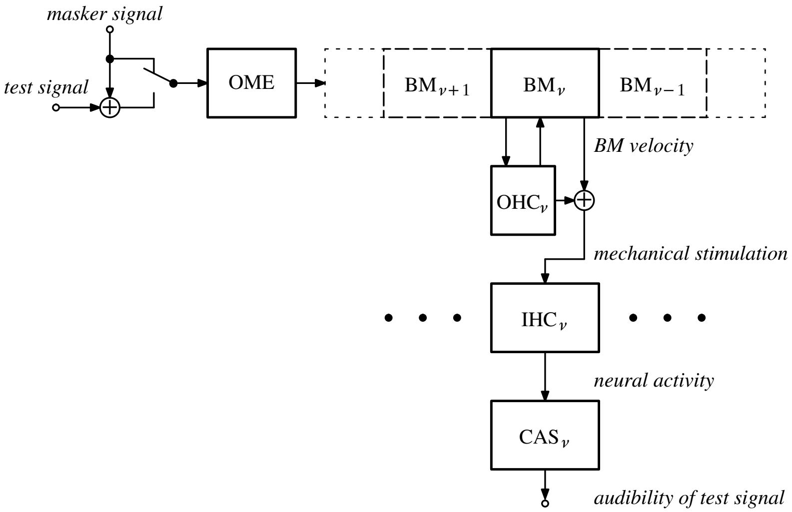

Figure 4. Block diagram of the physiological ear model. The peripheral auditory system is rebuild by the models of the outer and middle ear (OME), the basilar membrane (BM), the outer hair cell (OHC) and the inner hair cell (IHC) which is connected to the central auditory system (CAS) model. The BM model is realized in 251 sections. Each section is connected to an OHC model in a feedback loop and an IHC model. The feedback signal is amplified by the OHC model and added with the BM velocity signal resulting in the IHC stimulation. The audibility of a test signal is derived by first calculating the internal representation inside the CAS model of the masker signal alone and afterwards assessing the amount of change of the internal representation when the test signal is superimposed to the masker.

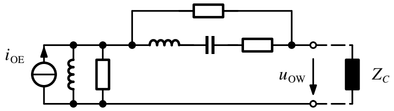

Figure 5. Electrical network model of outer and middle ear (OME). The current i_{OE} corresponds to the sound pressure at the outer ear entrance, the voltage u_{OW} corresponds to the velocity of the oval window of the cochlea and Z_{C} represents the input impedance of the cochlear model.

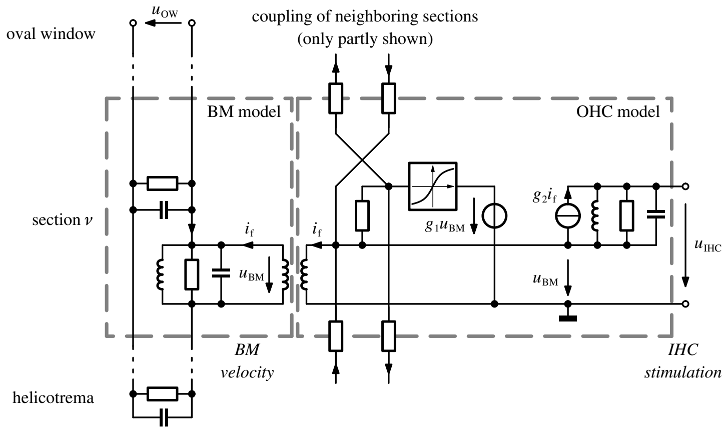

Figure 6. One section of the electrical network model of the cochlea without inner hair cell (IHC) model. The voltage u_{OW} from the outer and middle ear (OME) model is input at the oval window. The cascade of model sections is completed by the RC network rebuilding the helicotrema. The basilar membrane (BM) model in each section rebuilds the local hydromechanics of the cochlea with the voltage u_{BM} representing the local transverse BM velocity. The outer hair cell (OHC) model acts partly inside a feedback loop resulting in a high u_{BM} gain at low levels. The second amplification stage outside the feedback loop contributes additionally the amplified feedback signal to the output voltage u_{IHC}, the IHC stimulation signal.

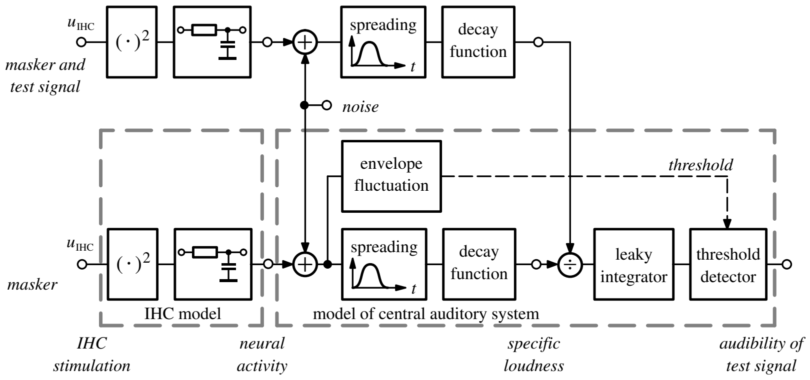

Figure 7. Block diagram of the inner hair cell (IHC) model and simplified block diagram of the central auditory system (CAS) model. The IHC stimulation signal u_{IHC} of the preprocessed masker is applied to the upper signal path, the preprocessed masker with superimposed test signal is applied to the lower path. Corresponding blocks in the upper and lower path are identical and the IHC and CAS models in all sections are identical.

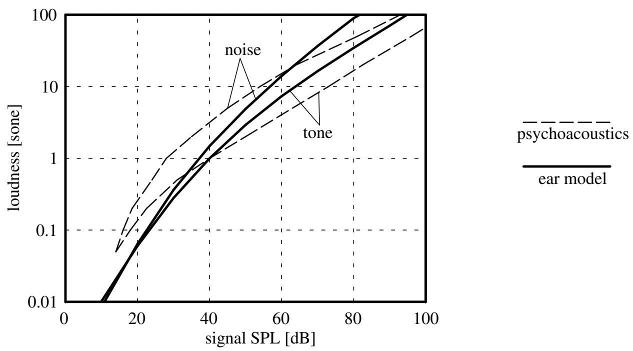

Figure 8. Total loudness of a pure tone and uniform-exciting noise at different sound pressure levels (SPLs). (Psychoacoustical data adapted from [10])

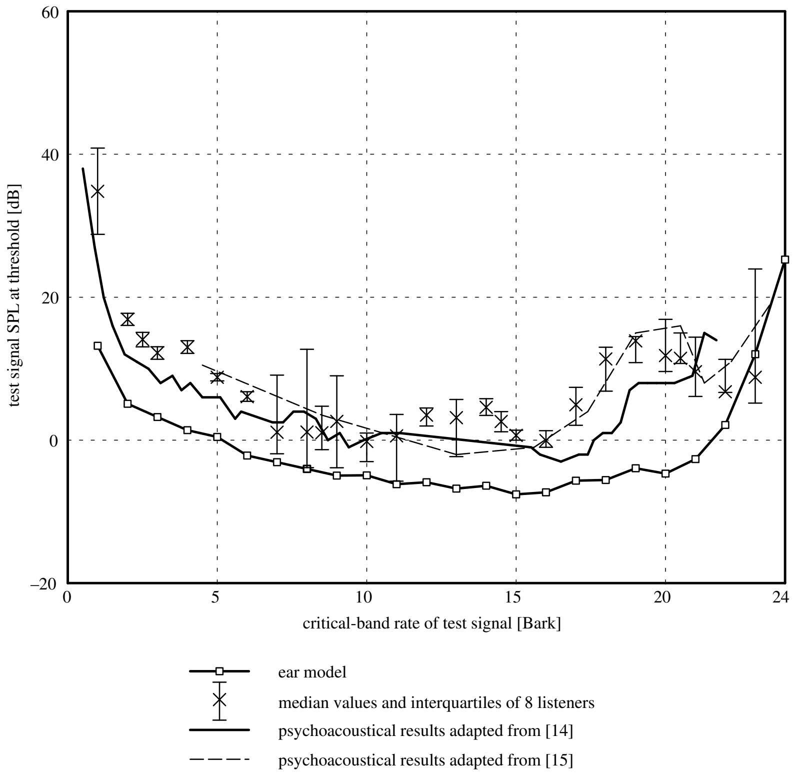

Figure 9. Masked threshold for pure tones in quiet. Simulation results compared to psychoacoustical measurements from the literature and a performed psychoacoustical test with 8 participants.

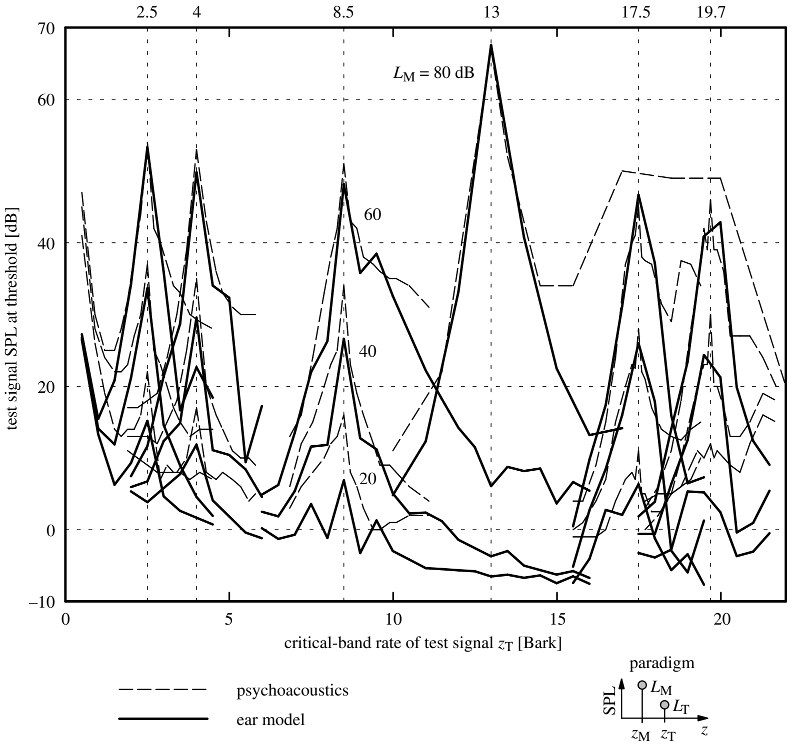

Figure 10. Masked thresholds for a pure tone masked by a pure tone. Modelling results compared to psychoacoustical measurements (adapted from [10],[14]). The critical-band rates of the pure-tone masker used in the different tests corresponding to the masking patterns from left to right as indicated above the graph were 2.5 (250 Hz), 4 (400 Hz), 8.5 (1 kHz), 13 (2 kHz), 17.5 (4 kHz), and 19.7 Bark (6 kHz). The masker levels were 20, 40, and 60 dB SPL except for the 13 Bark masker which was adjusted to 80 dB SPL.

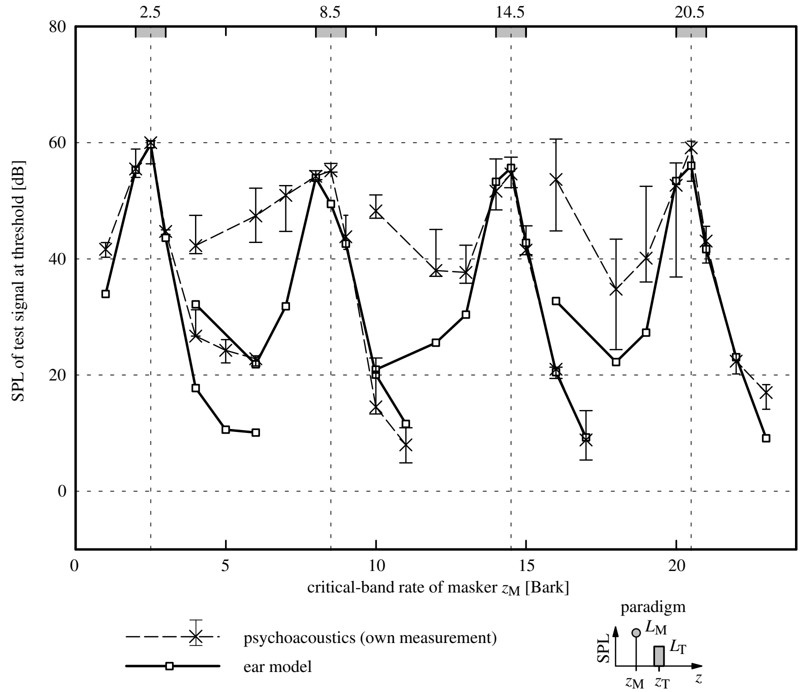

Figure 11. Simulation results together with medians and interquartile ranges of masked thresholds for critical-band noise masked by a pure tone from 8 subjects. The test signal was fixed in frequency (2.5, 8.5, 14.5, or 20.5 Bark) for each masking pattern with a bandwidth of 1 Bark as indicated at the top of the graph. The test signal SPL L_T at threshold is plotted for a given pure tone masker with an SPL of L_M = 80 dB and a critical-band rate z_M. Off-frequency listening was possible in the psychoacoustical test and also simulated by the model.

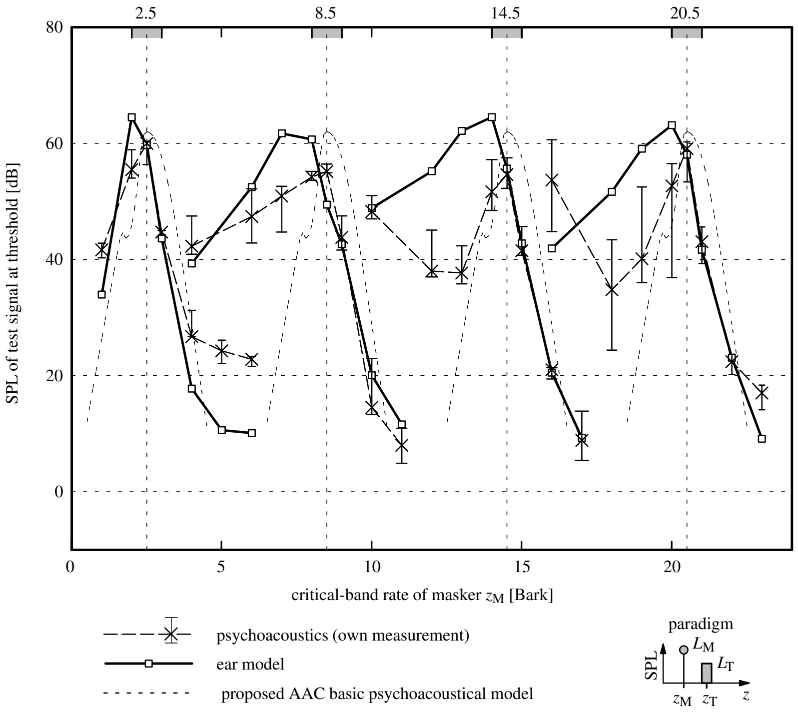

Figure 12. Caption see Fig. 11. In this graph, the masked thresholds from the model were obtained without off-frequency listening. Additionally, calculated masked thresholds from the proposed basic psychoacoustical model of the ISO/MPEG-2 AAC are shown for comparison.

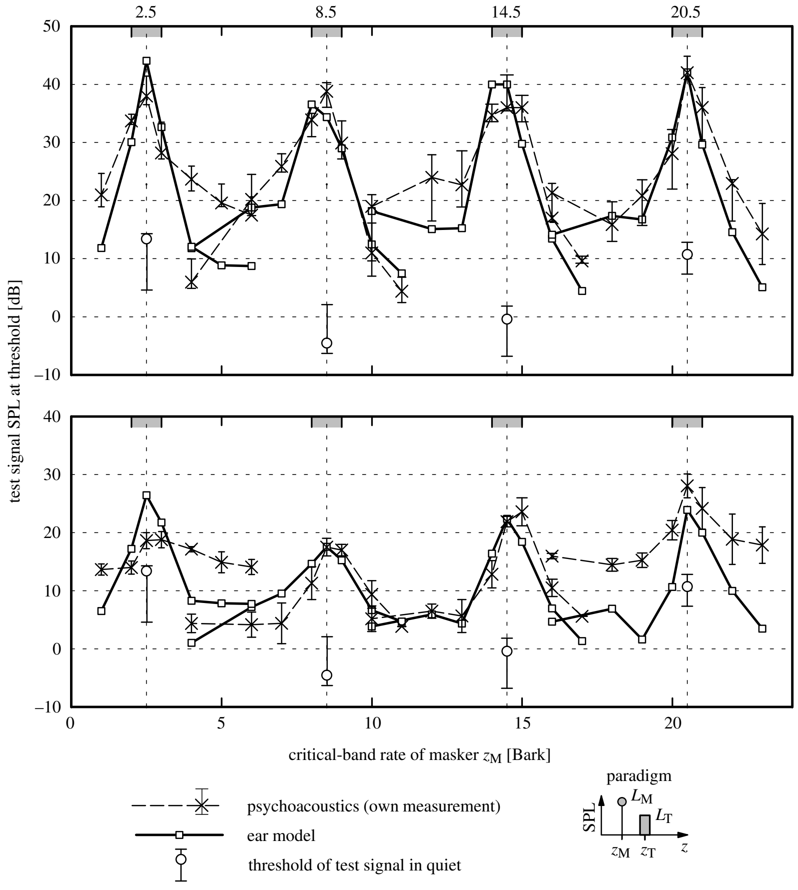

Figure 13. Caption see Fig. 11. *Upper diagram*: Masked thresholds of 3 subjects for a masker level L_M = 60 dB SPL, *lower diagram*: Masked thresholds of 2 subjects for a masker level L_M = 40 dB SPL. The threshold in quiet of 8 subjects for each test signal is plotted as median and interquartile range in both panels.

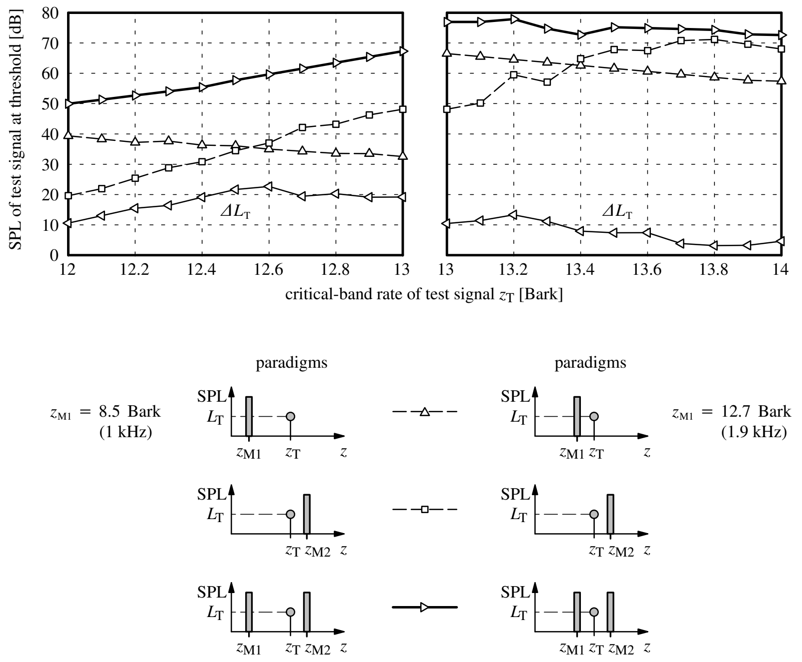

Figure 14. Simulated masked thresholds for a 2-kHz (z_T = 13 Bark) test tone masked by 50-Hz-wide Gaussian noises. In the first paradigm one noise masker with a frequency below the test signal was present while in the second the masker frequency was higher than the test signal frequency. Both maskers were used simultaneously in the third paradigm. The center frequency of the first masker was different for the results shown in the left and right panel as indicated. The second masker was always at 2.1 kHz (z_{M1} = 13.3 Bark) and both masker SPLs were 80 dB. The increase of the masked threshold in the third paradigm compared with the first and second is plotted as lowest curve denoted by ΔL_T. It is calculated as the masked threshold difference of the third paradigm and the maximum of the first and second paradigm.

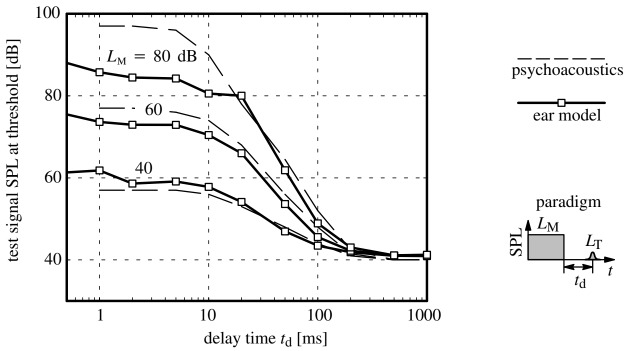

Figure 15. Postmasking patterns for a white-noise masker at different levels and a 20-µs Gaussian test pulse with peak level L_T (Psychoacoustical data adapted from [10]).

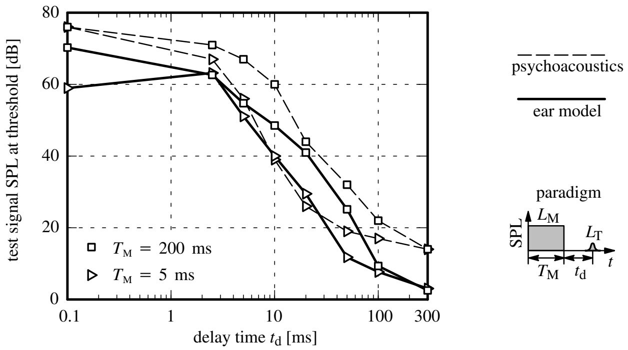

Figure 16. Postmasking patterns for uniform-masking noise with 80 dB SPL and a duration of T_M. The test signal was a 2-kHz test tone burst of 5 ms duration (Psychoacoustical data adapted from [10]).

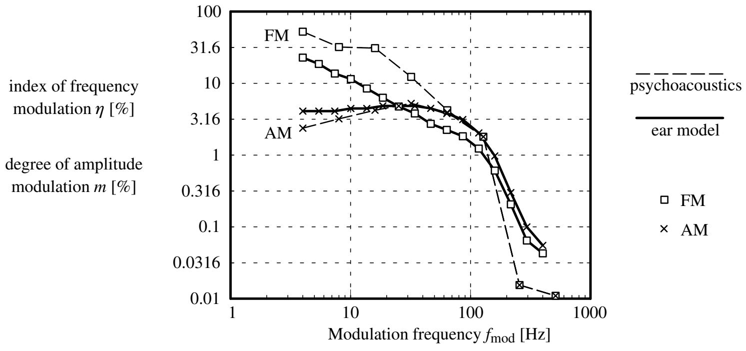

Figure 17. Detection thresholds for sinusoidal frequency modulation (FM) and amplitude modulation (AM) of a 1-kHz tone at 80 dB SPL (Psychoacoustical data adapted from [10]).