APPLICATION OF A PHYSIOLOGICAL EAR MODEL TO IRRELEVANCE REDUCTION IN AUDIO CODING

FRANK BAUMGARTE

Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung
Universität Hannover, Hannover, Germany
baumgart@tnt.uni-hannover.de

A previously published physiological ear model is applied as perceptual model to an audio coder complying with the ISO/MPEG-2 AAC standard. The achieved subjective sound quality is compared to results from an optimized psychoacoustical model. Significant deviations of the generated masked thresholds from the physiological ear model and the psychoacoustical model are evaluated with respect to psychoacoustical measurements.

INTRODUCTION

High-quality audio coding for target bit rates of 64 kbit/s per channel and below requires a sophisticated perceptual model for the reduction of irrelevance. In this application both, irrelevance and redundancy reduction provide a significant contribution to the overall coding gain.

The primary task of the perceptual model is the prediction of the masked threshold for introduced quantization noise. Subband or transform coding schemes use a time-to-frequency mapping with subsequent quantization and coding of the spectral components. These schemes currently offer the best audio quality at a given bit rate for high quality applications. Quantization noise originates from amplitude quantization of spectral component samples and typically consists of narrow-band noise with a bandwidth determined by the decoder signal synthesis employing the inverse time-to-frequency mapping. In case of a very coarse quantization consecutive reconstructed samples of a spectral component can be equal to zero so that these components are removed from the original signal. These considerations illustrate that irrelevance reduction is achieved on one hand by permitting quantization noise up to a level where it remains just inaudible and on the other hand by omitting signal components which are inaudible.

Perceptual models currently applied widely ignore the highly nonlinear properties of the human auditory system which influence masking. These models predict significantly deviating masked thresholds compared to psychoacoustical measurements especially in situations where the masked threshold is mainly determined by nonlinear effects [1]. Masking from complex audio signals is assumed to depend to a great extend on the nonlinear sound processing. A previously published ear model [1][2][3] overcomes these limitations by rebuilding the sound processing of the auditory system based on physiology. The model was verified using results from psychoacoustical measurements of masked thresholds including masking effects mainly determined by the nonlinear sound processing of the auditory system. For the model verification masked thresholds for simple test signals like a pure tone or narrow-band noise were considered.

In order to control the quantization of an audio coder, an appropriate procedure for adjusting the step sizes depending on the ear model output is necessary. In contrast to the model verification with pre-defined masker and test signals the introduced quantization distortions are not known a priori. It is only possible to estimate the power spectral density without actually carrying out the quantization and reconstruction. For the physiological ear model it is necessary to have the temporal waveform of the audio signal and the reconstructed (decoded) signal available. The audio signal is associated with the masker and the reconstructed signal is associated with the masker plus superimposed test signal in the terminology of psychoacoustical masking experiments. The model predicts whether the distortion is audible or not and provides a distance measure between the distortion perception measure and an internal threshold value.

In order to utilize the ear model for audio coding a quantizer step size adjustment procedure controlled by the physiological ear model is integrated into an ISO/MPEG-2 AAC encoder and evaluated. The subjective sound quality is assessed using a variable bit rate in order to shape the quantization noise such that it is just inaudible according to the ear model prediction. This approach avoids the influence of the bit allocation algorithm on the introduced noise level which is necessary to achieve a fixed bit rate.

This paper is focussed on the quantizer step size adjustment procedure and the achieved sound quality obtained from the physiological ear model. Section 1 contains a brief review of the physiological ear model structure. The step size adjustment procedure is presented in Section 2. First results from the ISO/MPEG-2 AAC implementation are reported in Section 3. Conclusions are drawn in the last Section.

1 PHYSIOLOGICAL EAR MODEL

The physiological modeling approach can only be realized for the processing stages of the auditory system with known physiological properties. While the physiology of the peripheral ear up to the auditory nerve is widely explored, there is less knowledge available about the physiology of the neural processing stages in the central ear. Therefore, the ear model is composed of a physiological model for the peripheral ear and complemented by a psychoacoustically based model for the neural processing in the central ear. Both model parts fit in the conception of signal detection theory which provides a framework for an analytical description of the detectability of a signal in noise.

1.1 Model Conception

A general model structure for the prediction of masked thresholds can be derived from the conception of signal detection theory [5]. A simple example will be given here in order to illustrate some basic ideas behind the theory and their consequences for this application.

The conception assumes that a virtual observer which only has access to a signal with superimposed distortions has to decide whether the signal is present in the observed signal or not. The signal corresponds to the test signal used in psychoacoustics and will be referred to as test signal in the following. A simple signal model assumes two sources of distortions as shown in Figure 1. The test signal itself is distorted by additive external noise. This distorted test signal is preprocessed by a system which can be associated with the peripheral ear. The observer has access to the internal signal representation given by the preprocessed signal with superimposed internal noise. The observer is assumed to be realized in the central ear, where the neural signals are to be evaluated. An optimal observer shows the best performance in terms of the detection error probability. It performs a measurement on the observed distorted test signal and compares the measurement result with a threshold value as shown in Figure 2. The test signal is detected by the observer when the measurement exceeds the threshold value.

Signal detection theory can be applied to psychoacoustical masked threshold measurements by assuming that the external noise in Figure 1 is introduced by a masker signal which is added to the test signal. In general the masker is not restricted to consist of a stochastic noise signal but it may as well consist of an arbitrary deterministic signal like a pure tone. The internal noise is associated with noise produced by the signal transduction and additional distortions which are present in the inner ear.

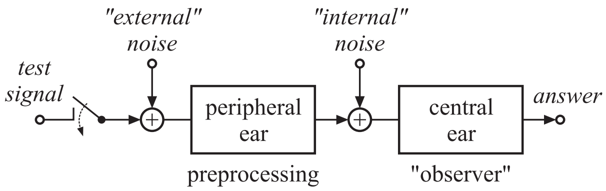

Figure 1: Test signal detection model. The detectability of a test signal at the input is limited by superimposed noise.

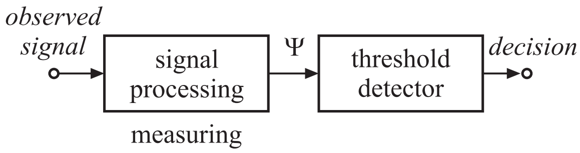

Figure 2: Model of the observer consisting of a measuring unit providing the measure Ψ and a threshold detector which compares Ψ with a threshold value Ψ_S.

The operation of the observer can be illustrated for the simple example of a narrow-band masker as external noise and a test signal with the same center frequency and bandwidth. Internal noise is assumed to have a negligible level in comparison to the external noise. In this case only the energy of the observed signal is changed by the test signal compared to the sole masker signal. The optimal observer measures the signal energy in the observation-time interval. The probability density function of the observed signal energy for the cases with applied test signal $p(\Psi_T)$ and without test signal $p(\Psi_R)$ is outlined in Figure 3. The distributions of the observed energy are assumed as Gaussian of standard deviation σ. The error probability P_e is defined as the sum of the probabilities that a test signal detection occurs without applying a test signal and that no detection occurs when a test signal is present. P_e is minimized by adjusting the energy threshold value Ψ_S so that the energy probability densities with and without test signal are equal at the threshold value.

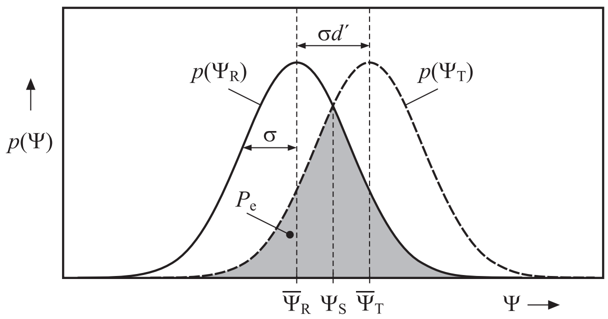

Figure 3: Gaussian probability density functions of the measured energy Ψ of the observed signal. $P(\Psi_{\mathrm{T}})$ is valid when the test signal is present at the input. $P(\Psi_{\mathrm{R}})$ results without test signal. The detection error probability P_{e} is represented by the filled area.

From Figure 3 it is obvious that the error probability increases with the masker energy variance and with decreasing test signal energy, which results in a reduced distance of both distributions. The masker energy variance depends on the type of masker and reaches a minimum for a pure tone masker which results in a lower masked threshold for a pure tone compared to a narrow-band noise. The asymmetry of masking between noise and tone [4] can be explained with this model in terms of observed energy variability. The detection threshold is usually defined as a value of $d' = 1$ which results in a difference of the distribution means $\overline{\Psi}_R$ and $\overline{\Psi}_T$ of σ in this example.

The observer can reach a higher performance, if the internal representation of the test signal waveform is known. In this case the observer can perform a correlation measurement, which allows for a more certain test signal detection and a lower detection error probability. Results from further signal configurations can be found in the literature [5].

The theoretical concept of the observer can be utilized for the design of the central ear model. The signal processing unit and threshold value can be derived by psychoacoustical measurements. The influence of the peripheral preprocessing on masking also considered by this concept.

1.2 Model Structure

An overview of the physiological ear model structure is given in Figure 4. The ear model was already presented in a previous paper [2] and is only briefly reviewed here. The input sound pressure signal is filtered in the first block rebuilding the simplified outer and middle ear (OME) properties. The inner ear model is realized in 251 sections with each section rebuilding the properties of a small slice of the cochlea which contains the sound processing part of the inner ear. The mechanical properties are represented by the hydromechanical model part (HM). The outer hair cells (OHC) which act as amplifiers with saturation are considered in a feedback loop. The maximum amplification is achieved for low-level signals and amounts to approximately 60 dB compared to the passive case. The mechanical to neural transduction is represented by an inner hair cell model (IHC) which consists of a square function and a first order low-pass filter. The output of the inner hair cell models represent the firing rate of the associated auditory nerve fibers, which are input to the neural processing model.

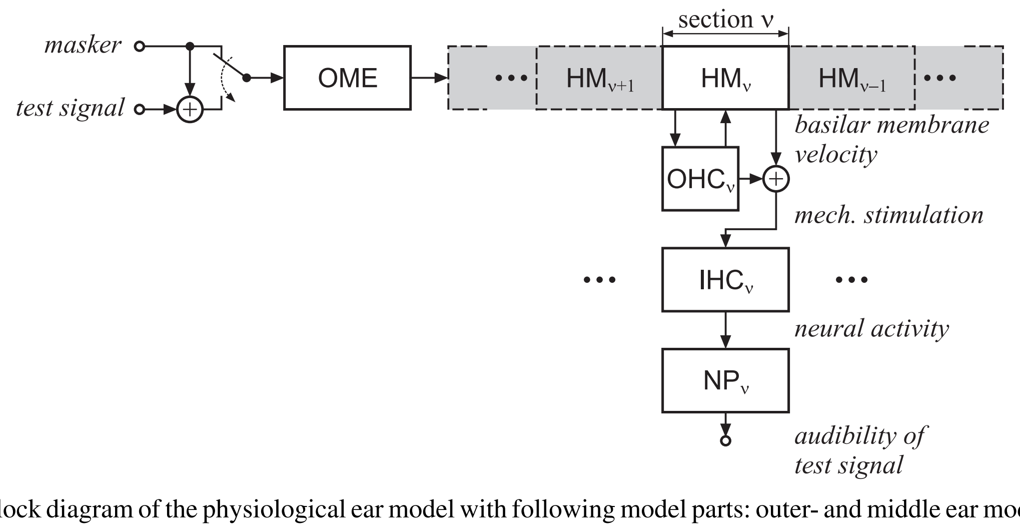

Figure 4: Block diagram of the physiological ear model with following model parts: outer- and middle ear model (OME), cochlear hydromechanics (HM_ν), outer hair cell (OHC_ν), inner hair cell (IHC_ν), and neural processing (NP_ν). Only one section is shown — other sections have identical structure.

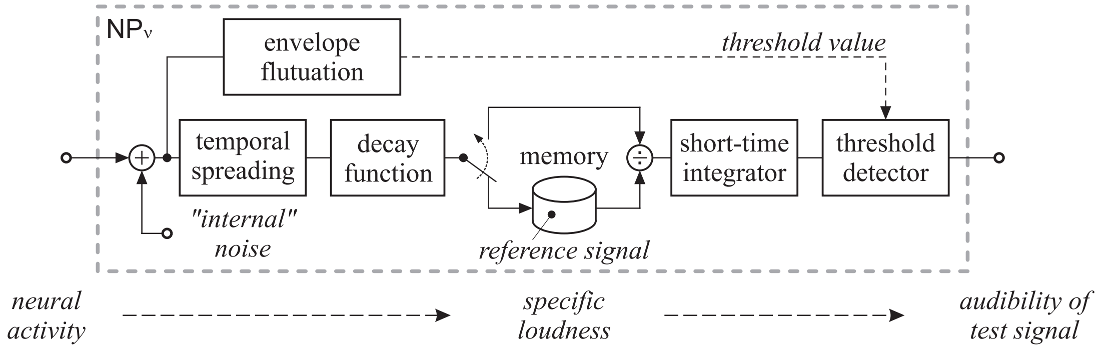

Figure 5: Block diagram of neural processing model in one section. The masked threshold is derived by first generating the specific loudness of the reference signal (masker) which is stored in the memory. In subsequent iteration steps the test signal level is adjusted so that the specific loudness change due to the test signal superposition is at the internal threshold value.

The simplified block diagram of the neural processing model is given in Figure 5. It consists of different modules considering properties of temporal masking effects and a threshold detector. At the input the addition of internal noise determines the audibility of a test signal at low masker levels resulting in the threshold in quiet. The temporal spreading function accounts for the properties of premasking. The following decay function is adapted to the slower postmasking decline after a masker is turned off. The output is assumed to represent specific loudness [6], e.g. the loudness distribution on a spectral Bark scale.

In order to generate the masked threshold for a test signal superimposed to a masker signal the specific loudness of the masker is stored as ‘‘internal’’ reference in a memory. In a first iteration step the test signal is added to the masker and the amount of specific loudness change of masker plus test signal and reference is derived by a short-time integration of its ratio. The output is evaluated by a threshold detector assuming that the change is audible whenever the internal threshold value is exceeded. In this case the test signal level is reduced for the next iteration step and the ratio is calculated again. This procedure is repeated until the specific loudness change is just below the threshold value.

The threshold value is adapted to the envelope fluctuation of the input signal of the neural processing model with internal noise added. This mechanism considers the asymmetry of masking between noise and tone in a way that the low envelope fluctuation of a tone results in a reduced threshold value used by the threshold detector.

The structure of the physiological ear model is consistent with the conception of signal detection theory. The preprocessing of the peripheral ear is realized in the OME, HM, OHC, and IHC models. The band-pass filtering of each model section allows for an independent signal detection in each section of the model. Observers are assumed to be represented by the neural processing model in each section. However, the realized observer has only suboptimal performance since it is for example not able to evaluate a correlation measure in case of a known internal test signal representation. The external noise corresponds to the masker signal which limits the detectability of a test signal and the internal noise determines the absolute threshold. The variability of the observed signal is estimated by the envelope fluctuation measure which accounts for different test signal detectabilities depending on the energy distribution (cf. Figure 3).

1.3 Adjustment of Quantizer Step Sizes

Reduction of irrelevance by a perceptual audio coding scheme requires the adjustment of the quantizer step sizes so that the introduced quantization noise is just below the masked threshold. This criterion can be evaluated with the ear model by a comparison of the original and decoded signal. This comparison results in a distance measure as a function of time for the internal loudness change in relation to the internal threshold value of each ear model section. This distance measure is utilized for an iterative adjustment of the quantizer step size such that the distance between the internal specific loudness change and the threshold value is minimized.

Irrelevance reduction is investigated here based on the ISO/MPEG–2 Advanced Audio Coding (AAC) standard. This standard employs a filterbank to derive a time-to-frequency mapping resulting in a subband signal representation with uniform spectral resolution. Adjacent subbands are grouped into scalefactor bands and use a common step size for the quantization of the subband samples. The bandwidth of the scalefactor bands are related to the critical bands of a Bark scale. It is approximately constant up to a center frequency of 1 kHz and increases at higher center frequencies. Applying the distance measure to the quantizer adjustment requires the consideration of different delays of the filterbank in the encoder and the ear model. Additionally, the model sections must be assigned to corresponding subband quantizers which operate at the same center frequencies. The quantizer step size for the first iteration is derived from the power density spectrum. The permitted quantization noise is calculated by limiting the slope steepness of the scalefactor band energy spectrum and reducing the spectral level by 10 dB as illustrated in Figure 6. The slopes used are 15 dB per scalefactor band in the direction to lower frequencies and –3 dB to higher frequencies. The calculated masked threshold is finally compared to the absolute masked threshold and the larger value in each scalefactor band is taken as initial masked threshold. The initial adjustment is used as starting value for an iterative procedure of evaluating the decoded signal with the ear model and adapting the step sizes accordingly. The final result is taken from the iteration procedure after a fixed number of five iterations.

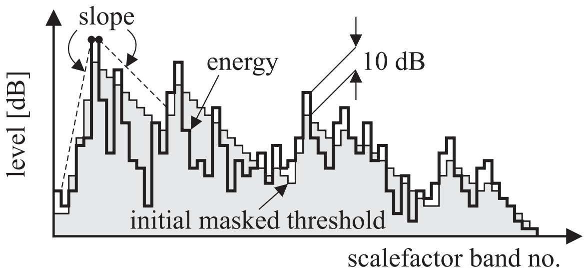

Figure 6: Initial masked threshold calculation for the first iteration step. Shown for an example energy distribution.

Currently in each iteration step the complete signal is processed first by the encoder and decoder and afterwards evaluated by the ear model. For a frame-by-frame iteration this procedure has to be modified to complete all iterations in one frame before proceeding to the next. This method is not restricted to off-line applications and reduces the necessary memory capacity. Due to the maximum signal delay of the ear model of approximately 10 ms one additional frame must be available in the encoding process. Thus the coding/decoding delay is increased by one frame. Additional delay can be caused by the temporal SMR smoothing if SMR values are used from following frames. The current implementation uses two frames in either temporal direction.

The quantizer step size adjustment is controlled via the individual SMR values in each scalefactor band. The SMR values are temporally smoothed in order to avoid large temporal SMR changes. The iterative adjustment itself is controlled by the obtained distance between the distortion perception measure and internal threshold value. Since more than one ear model section is assigned to one scalefactor band, the distance is defined as the maximum distance obtained from all assigned sections during the belonging frame. This distance is the input argument of a manually optimized nonlinear adjustment function which provides the SMR modification factor for the next iteration step in order to minimize the distance. The amount of SMR modification depends additionally on the presence of a threshold excess. If the internal threshold is exceeded the SMR is increased by a larger factor compared to the SMR reduction factor used in case of a distortion below the internal threshold. This ensures that audible distortions are rapidly reduced and oscillations in the SMR adjustment in consecutive are prevented.

Since the nonlinearity of the ear model creates distortion products at frequencies different from the input signal [1] a threshold excess in a specific ear model section can originate from quantization distortions different from the center frequency of that section. The most dominant distortion from the ear model is the distortion product at the cubic difference frequency $f_3 = 2f_A - f_B$ created by two superimposed sinusoidals with the frequencies f_A and f_B ($f_B > f_A$). The simple iteration method described above fails if distortion products cause an internal threshold excess since the distortion product will not be reduced by increasing the SMR at the subband frequency of the cubic difference frequency. From the iteration results of different audio material it is observed that the distortion products have only little effect on the SMR adjustment and, if present, they most likely occur at very low frequencies. This observation confirms an earlier assumption [1] that due to the high spectral resolution of the decoder filterbank the frequencies of the quantization noise is always close to a frequency component of the audio signal. Therefore, the cubic distortion frequency will be very low and most likely be masked or below absolute threshold. Distortion products are not considered in the SMR adjustment procedure since their detection causes considerable computational costs while only a small performance gain is expected. However, distortion products may cause an unbounded continuous SMR growth during the iterative step size adjustment procedure which is prevented by limiting the SMR to a maximum value of 23 dB. This SMR limit is increased by 1 dB per scalefactor band for the band number 8 down to 1.

The quantizer step sizes must be iteratively adjusted since it is not possible to derive an inverse ear model without significant model simplifications. An inverse ear model would allow to calculate the test signal level at the masked threshold directly from the internal threshold value. Due to the nonlinearity of the model an inversion can only be derived for a linearized model at the signal-dependent operating point and for one center frequency. Even if the linearization is possible, a significant number of linearized models have to be created for a sufficient number of frequencies. Such an inverse ear model thus provides no reduction in computational complexity compared to the iterative approach using the nonlinear model.

It is necessary for the iterative procedure to generate a variable bit rate so that no additional bit allocation algorithm influences the noise shaping which is applied to force a fixed bit rate encoding. The variable rate encoding permits that the introduced quantization distortions are close to the threshold level predicted by the perceptual model. The masked threshold can be verified by subjective assessment of the decoded variable rate bitstream.

In applications where a fixed bit rate is appropriate audible distortions are expected if the amount of available bits is insufficient to keep the quantization noise below the masked threshold. In these situations the noise level is usually increased by a constant level offset to the masked threshold until the number of bits allows to achieve that noise level. With the ear model the noise level above masked threshold can be shaped in way that audible distortions lead to a constant excess of the internal threshold value. This constant internal threshold excess is assumed to result in a noise energy distribution which can be less audible than a distribution according to a constant masked threshold offset.

2 RESULTS

First results from the ear model quantization control are obtained from an implementation into an ISO/MPEG-2 Advanced Audio Coding (AAC) [7] compliant encoder. The reference encoder utilizes an optimized psychoacoustical model which was already in use in former listening tests during the MPEG-standardization process [8][9] [10].

Compared to other coding schemes AAC currently achieves the highest subjective audio quality at bit rates in the range of 64 kbit/s per channel which enable a close-to-CD quality.

AAC uses a spectral decomposition of the input signal into critically sampled subband signals. The application of two alternative uniform spectral resolutions provided by the filterbank with either 1024 or 128 subbands allows a signal-adaptive decomposition which provides the choice between the standard high spectral resolution and an increased temporal resolution in conjunction with the reduced spectral resolution. The temporal resolution follows from the block size of the filterbank input samples which amounts to 2048 sample intervals and 256 sample intervals for the high and low spectral resolution respectively. Adjacent subbands are grouped into 49 scalefactor bands for the high spectral resolution and 14 scalefactor bands for the low spectral resolution.

In case of variable rate coding the quantizer step size is derived from the signal-to-mask ratio (SMR) in each scalefactor band. This ratio determines the maximum permitted quantization noise level in relation to the energy level such that the noise level does not exceed the masked threshold. The SMR level is approximately proportional to the number of bits necessary to encode a subband sample.

The results presented here were derived from the reference AAC encoder and the modified version with the psychoacoustical model replaced by the physiological ear model. Only one channel (mono) signals are used since the ear model does not take into account any binaural masking effects. The encoding options were chosen to include intra-channel prediction but no temporal noise shaping (TNS) since the TNS option of this encoder implementation results for some test sequences in a quality reduction. The bandwidth of the encoded signal was limited to 15.5 kHz.

The coding results are compared at the same mean bit rate calculated from all test sequences. The bit rate obtained from the modified encoder under ear model control is used as reference. The reference encoder is adjusted to the same mean bit rate by applying a constant level offset of –5.05 dB to the masked threshold generated by the psychoacoustical model.

2.1 Subjective Quality

Seven audio signals showing the critical artefacts after coding at a fixed bit rate of 64 and 56 kbit/s were chosen from a larger set. In line with observations from earlier listening tests carried out during the MPEG standardization process [11] male speech signals turned out to result in clearly audible distortions. Other selected items are female vocals, castanets and harpsichord. Each test item has a duration of approximately 10 seconds. A listening test was performed using the triple stimulus / double blind / hidden reference methodology based on ITU-R Recommendation BS.1116 [12]. In each trial the listener is presented with three signals starting with the original. The remaining two consist of the original again and the decoded signal in arbitrary order. The quality of the latter two signals is graded in comparison to the original using the ITU-R 5-point impairment scale. Possible gradings for introduced distortions range from 1 for ‘‘very annoying’’ to 5 which means ‘‘imperceptible’’ on a continuous scale. The test results are usually presented as mean difference gradings and 95% confidence intervals from all listeners. A difference grading is defined as the difference of the gradings for the hidden decoded signal and the hidden reference. Figure 7 shows the results from 7 listeners for each sequence in the test and the mean results over all sequences. The mean bit rate measured for each sequence and encoder is shown in Figure 8.

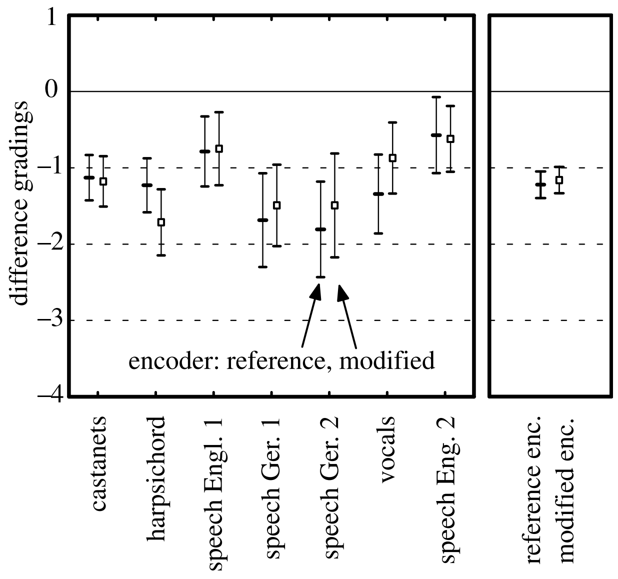

Figure 7: Difference gradings and 95% confidence intervals of 7 subjects for the selected set of 7 test sequences. *Left*: Values averaged over all subjects. *Right*: Averaged values over all subjects and sequences for each encoder.

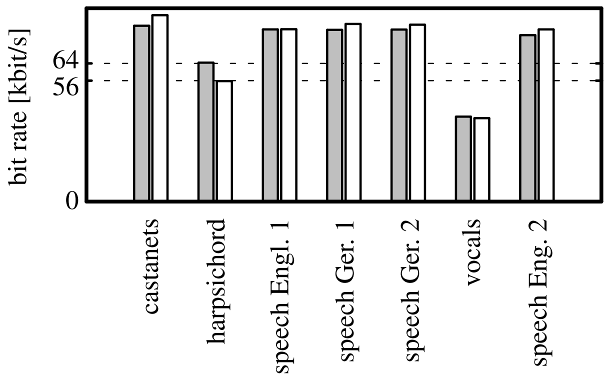

Figure 8: Mean bit rate of each test sequence from the reference encoder (*grey*) and the modified encoder (*white*).

While the overall mean quality grading shows no significant deviation from the reference encoder and the modified encoder with the psychoacoustical model replaced by the ear model, there are some implications from the different gradings of the individual audio signals. The largest quality differences of both coders are observed for the signals “German speech”, “harpsichord”, and “vocals”. The large confidence intervals are caused mainly by different absolute gradings of the subjects so that for example “German speech 2” from the modified encoder was never graded worse than from the reference encoder. The grading of the harpsichord recording from the modified encoder shows the largest deviation towards lower quality gradings in comparison to the reference encoder due to audible artefacts occurring at the lowest tone played on that instrument. It should be noted that the lower sound quality is partly caused by a reduced bit rate from the modified encoder as outlined in Figure 8. “German speech” shows higher quality from the modified encoder indicating that the fast changes in signal statistics inherent in speech is adequately resolved by the ear model. General speech signals often contain variations of the fundamental frequency in voiced parts which can be interpreted as a frequency modulation of the harmonics. This modulation may cause an increased masked threshold from the reference encoder in comparison to an unmodulated signal. Frequency-modulated signals with a sufficiently fast changing frequency are not classified as a purely tonal signal since the tonality measure is based on a prediction technique. A tonal signal results in a significantly higher SMR than a noise-like signal. In order to verify this assumption two synthetic frequency-modulated signals were explored.

The first signal consists of an sinusoidally-frequency-modulated pure tone with superimposed pink noise. The FM-signal frequency varies between 600 and 1900 Hz with a modulation frequency of 8 Hz as outlined in the spectrogram in Figure 9a.

The FM signal was processed by the reference and the modified encoder using variable rate and applying the same masked threshold offset to the reference encoder as in the quality assessment described above. The decoded-signal spectrograms are shown in Figure 9b from the reference encoder and in Figure 9c from the modified encoder. Both decoded signals show differences compared to the original at the FM-signal slopes as well as areas where parts of the pink noise are not encoded. Both decoded signals differ in the amount of distortions at the signal slopes which are reduced in case of the modified encoder. The audibility of these distortions was evaluated by a subjective test in order to assess the obtained signal quality. The mean difference gradings of the decoded signals from 5 subjects were 0.56 grades higher for the modified encoder compared to the reference encoder. This result suggests that the reduced distortions from the modified model also lead to an improved subjective quality. The bitstream of the reference encoder comprises a total number of 84760 bits compared to the modified encoder with 84928 bits.

The second synthetic signal used for the evaluation of FM-signal components consists of an artificial vowel. The signal was generated using an impulse train with varying impulse rate as excitation signal for a vocal-tract filter. The filter resonances are visible from the spectrogram in Figure 10 as three maxima of the spectral envelope with

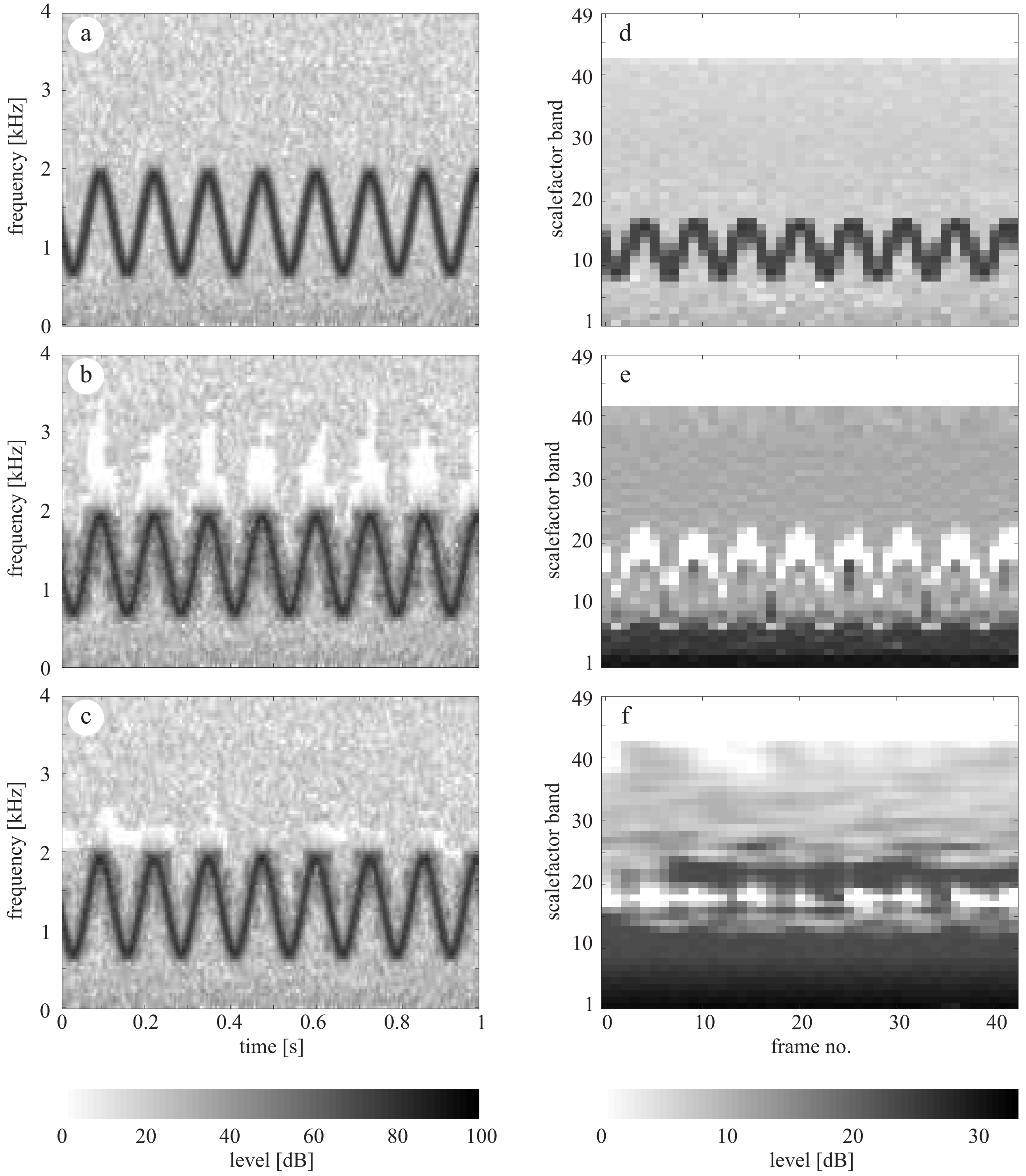

Figure 9: Sinusoidally-frequency modulated pure tone with superimposed pink noise. *Left column*: Spectrograms of original (a) and decoded audio signals from reference encoder (b) and modified encoder (c). *Right column*: Energy of original signal (d) and signal-to-mask ratios used in the reference (e) and modified encoder with physiological ear model (f). The greyscale-to-level assignment below the right column is valid for the graphs (e) and (f).

constant frequency. The impulse sequence creates an harmonic line spectrum with a fundamental frequency corresponding to the impulse repetition rate.

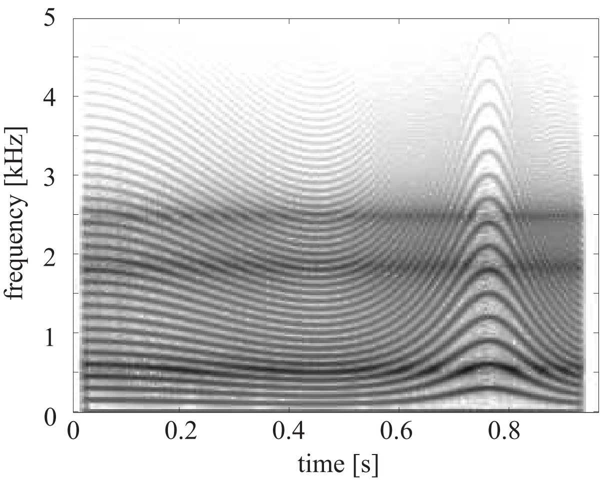

Figure 10: Spectrogram of a synthetic vowel.

Differences of the decoded signals are not obvious from the spectrograms (not shown). However, the decoded signals have different subjective quality. The mean difference gradings of the decoded signals from 5 subjects were 0.63 grades higher for the modified encoder compared to the reference encoder. The number of encoded bits is virtually identical. Clearly audible distortions from the reference encoder are present in intervals with significant frequency modulation which confirms the presumption of an improper tonality measure in these signal parts.

2.2 Signal-to-Masked Ratio

A deeper insight into the different results obtained from both encoders can be illustrated comparing the different SMRs. The SMR determines the quantization step size in each scalefactor band and thus the number of bits necessary to encode the subband samples. The SMR provides information about the different masked thresholds of the encoders and different shaping of the quantization noise.

A short signal excerpt from the sequence ‘‘German male speech 1’’ is utilized here to illustrate the results. The scalefactor-band energies of the excerpt are shown in Figure 11a. The SMR from both encoders are given in Figure 11b and 11c.

Compared to the modified encoder the reference encoder shows a smoother shape and only a little correlation between SMR and energy spectrum. For the frame indicated in the Figures by a vertical line the SMR is depicted in Figure 12. The bars in this graph are horizontally sized according to a linear frequency scale so that the total area of all bars is approximately proportional to the number of bits necessary to encode the subband samples of that frame.

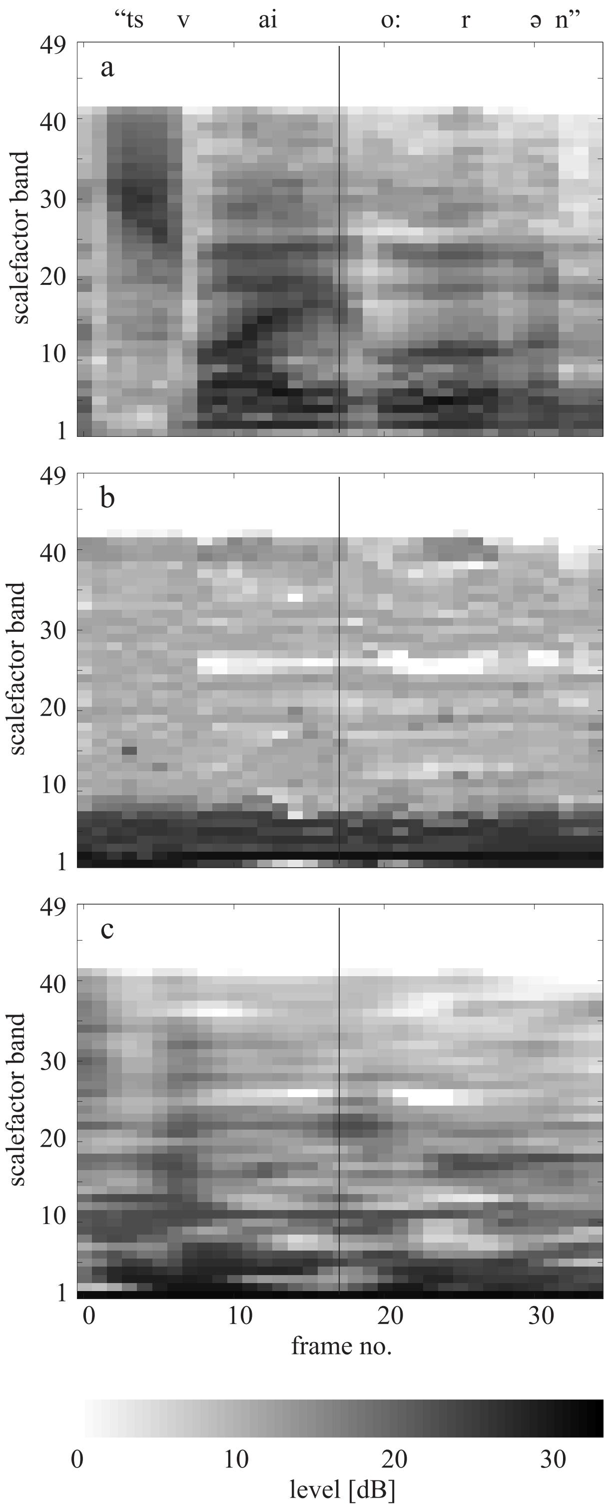

Figure 11: Excerpt from ‘‘German male speech’’ containing the spoken words ‘‘zwei Ohren’’. Scalefactor-band energies (a) and signal-to-mask ratios from the reference encoder (b) and modified encoder (c). The greyscale-to-level assignment is valid for the graphs (b) and (c).

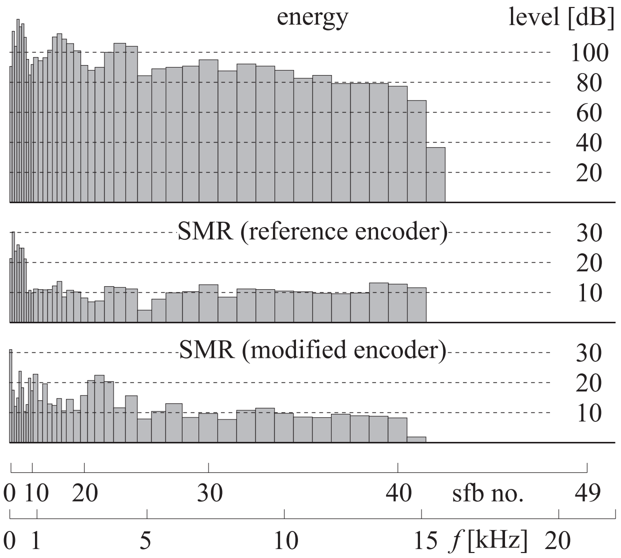

Figure 12: Signal energy and signal-to-mask ratio (SMR) shown for the indicated frame in Figure 11. Each bar represents one scalefactor band (sfb). The width of which is proportional to the bandwidth in Hertz.

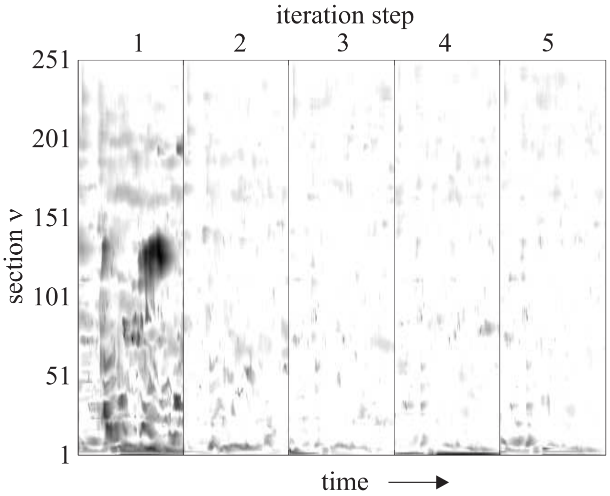

Figure 13: Illustration of the internal threshold excess after several iteration steps. Large values are shown in black, white areas indicate that no audible difference is detected by the ear model. The corresponding input signal is the excerpt from ”German male speech 1” also used in Figure 11. For each iteration step the abscissa represents the input signal duration.

For the sake of completeness the SMR values are also given for the FM signal in the right column of Figure 9. The right top graph (Figure 9d) shows the scalefactor-band energies of the original signal. The SMRs resulting from the reference and modified encoder are illustrated in Figure 9e and 9f respectively. The SMR graphs correspond to the decoded signal spectrograms in the left column. It is apparent that areas with no spectral energy are created in the decoded signals where the corresponding SMR is negative or zero.

The convergence of the iterative quantizer adjustment is illustrated in Figure 13. The results obtained from the decoded signal evaluation by the physiological ear model is shown for five iterations. The associated audio signal in this example is the same excerpt from “German male speech 1” as used in Figure 11. The figure shows a stepwise reduction of threshold excess which is plotted as the distance between the evaluated specific loudness change and the internal threshold value in each section of the model. The first iteration step reflects the distortions detected as audible by the ear model resulting from the initial masked threshold adjustment. These internal threshold exceedings are considerably reduced in the second iteration and successively lowered in further iterations. Since the quantizer step sizes are permitted to grow in case of distortions below the internal threshold value, it is possible that small internal threshold exceedings are created temporally in an iteration.

3 CONCLUSIONS

In this paper results from the application of a physiological ear model to irrelevance reduction in audio coding are reported. The ear model as such was described in earlier publications [1][2][3]. The ear model structure is motivated here considering conceptions of signal detection theory. An iterative quantizer step size adjustment procedure is developed to integrate the ear model into an ISO/MPEG-2-AAC compliant encoder.

Results are given in comparison to a reference AAC encoder utilizing an optimized psychoacoustical model already evaluated in former MPEG listening tests. The subjective quality from the reference encoder and the modified encoder shows the same performance in terms of mean quality over all test items. Slightly different quality is observed from the individual items. The modified encoder with the optimized psychoacoustical model replaced by the ear model shows improved quality for speech but lower quality for one of the single instrument recordings. One reason for the better speech performance is the improper tonality estimation for frequency-modulated signals of the reference encoder. This property is confirmed from additional measurements using synthetic frequency-modulated signals. Another reason is assumed to result from the fast changes of the speech signal statistics which can be more adequately resolved by the ear model.

Comparisons of the signal-to-mask ratio (SMR) obtained from both encoders show significant differences. For the variable rate encoding utilized here the SMR determines the quantizer step sizes so that the quantization-noise level approximates the masked-threshold level. The reference encoder results in a smooth SMR shape over time and frequency while the SMR from the modified encoder has a more signal-dependent shape. From the subjective assessment it can be stated that the perceived differences between the decoded signals are smaller than expected from the different SMRs.

The virtually equal performance of the optimized psychoacoustical model and the physiological ear model in terms of subjective audio quality in first results reported here indicates that the ear model is able to adequately predict masked thresholds for complex audio signals. In this application the ear model still provides room for parameter optimization based on more extensive subjective evaluations than could be realized in the present work.

ACKNOWLEDGEMENTS

The project was supported by the Deutsche Forschungsgemeinschaft (German national research foundation).

REFERENCES

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

[2] Baumgarte F. "A Physiological Ear Model for Auditory Masking Applicable to Perceptual Coding," *103rd AES Convention*, New York, NY, Preprint 4511, 1997.

[3] Baumgarte F. "A Physiological Ear Model for Specific Loudness and Masking," *Proc. IEEE WASPAA*, New Paltz, NY, 1997.

[4] Moore, B. C. J.; Alcántara, J. L.; Dau, T. "Masking patterns for sinusoids and narrow-band noise maskers," *J. Acoust. Soc. Am.*, Vol. 104, No. 2 (1), 1023–1038.

[5] Green, D. M.; Swets, J. A. *Signal detection theory and psychophysics*, Wiley, New York, 1966.

[6] Zwicker E., Fastl H. *Psychoacoustics. Facts and Models*. Springer Verlag, New York, 1999.

[7] ISO/IEC JTC1/SC29/WG11. *Coding of moving pictures and audio – MPEG-2 Advanced Audio Coding*. ISO/IEC 13818-7 international standard, 1997.

[8] ISO/IEC JTC/SC29/WG11/N1279. *NBC Reference Model 3 monophonic subjective tests: overall results*, 1996.

[9] ISO/IEC JTC/SC29/WG11/N1280. *NBC Reference Model 4 stereophonic and multichannel subjective tests: overall results*, 1996.

[10] ISO/IEC JTC/SC29/WG11/N1419. *Report on the formal subjective listening test of MPEG-2 NBC multichannel audio coding*, 1996.

[11] ISO/IEC JTC/SC29/WG11/N2006. *Report on the MPEG-2 AAC stereo verification test*, 1998.

[12] ITU-R. *Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems*. ITU-R Recommendation BS.1116. Geneva, 1994.