Audio Engineering Society

Convention Paper10277

Presented at the AES 160th Convention
2026 May 28–30, Copenhagen, Denmark

*This paper was peer-reviewed as a complete manuscript for presentation at this convention. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.*

# Perceptual Model Considering Comodulation Masking Release by Postmasking Adaptation

Fabian M. Schaller[1] and Bernd Edler[1,2]

[1] *Fraunhofer IIS, Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany*
[2] *International Audio Laboratories Erlangen, Germany**

Correspondence should be addressed to Fabian Schaller (fabian.schaller@iis.fraunhofer.de)

**ABSTRACT**

This work presents a perceptual model based on a complex IIR filterbank. The filterbank with a frequency resolution of 4 bands per Bark consists of 104 filters whose slopes are designed to take spectral masking effects into account. The filter outputs are used to obtain masking thresholds with the following post processing. To obtain resonable masking thresholds from the spreading outputs, a post masking stage is required. Here, we propose a comodulation dependent adaptation of the postmasking decay to model Comodulation Masking Release (CMR) effects. This approach explicitely considers the dip-listening effect known from literature. The final masking thresholds are obtained by weighting the postmasking outputs by a tonality dependent gain, controlled using spectral flatness estimation. A listening test compares the proposed method to an already known approach using direct CMR based modification of the masking threshold gains.

## 1 Introduction

In perceptual audio coding, very compact representations maintaining high audio quality can be achieved by exploiting masking effects in the human auditory system [1]. Specifically, the masking of quantization errors by the audio signal itself allows relatively coarse quantization, which leads to a reduction of the bit rate for the storage or transmission of the quantized data. As the masking effect varies over time and frequency, depending on the corresponding characteristics of the audio signal, time and frequency dependent quantizer control must be possible. For that, the audio signal must be converted to a suitable time-frequency representation, which is often obtained by applying a Modified Discrete Cosine Transform (MDCT). In parallel, the audio signal needs to be analyzed by a perceptual model, which estimates the strength of the masking effects, so that quantization can be applied appropriately [2].

Under some circumstances, maskers with equal low frequency modulation even can mutually reduce their masking effects [3, 4]. This phenomenon is called *comodulation masking release* (CMR). Since our proposed perceptual model is based on CMR, we describe in the following the

*A joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS, 91058 Erlangen, Germany

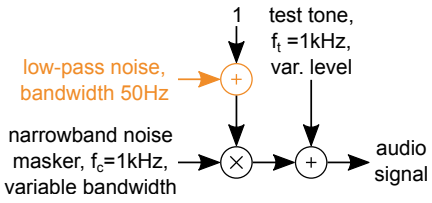



**Fig. 1:** Setup of experiment investigating CMR

fundamental experiment and effect.

CMR was shown in an experiment with the setup illustrated in Figure 1. Two types of maskers were used in this experiment. The first type was a stationary band limited noise centered around 1 kHz with the bandwidth as a control parameter. The second was noise with the same bandwidth, but multiplied by a randomly, relatively slowly changing amplitude modulation function (indicated in orange).

The test signal was a 1 kHz tone adjusted in its level, so that it just started becoming audible, as in usual masking experiments.

The result shown in Figure 2 for the non-modulated masker (black +) looks as expected: the masking threshold increases with bandwidth until the critical bandwidth is reached and then remains constant.

The result for the modulated masker (orange ■) however looks surprising. At first, the masking threshold increases with bandwidth, but then decreases again. This means that, although the total masker level increases, the masking threshold decreases. In other words, the noise components added by the masker bandwidth increase rather help for the audibility of the test tone, instead of preventing it.

For the outcomes of this experiment and other studies some possible explanations are given, such as suppression of *envelope locking* [5] and that flanking components guide to the optimum time for detection (*dip listening*) [6, 7]. The effect of CMR has been investigated for varying masker bandwidth, center frequency, modulator bandwidth, modulator type, and signal duration and it was found that the reduction of masking can be up to 18 dB [8]. An approach for considering CMR in

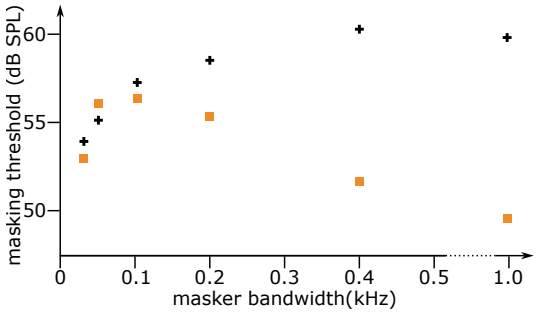



**Fig. 2:** Result of CMR experiment: audibility thresholds of a 1 kHz test tone in dependency of noise masker bandwidth without amplitude modulation (+) and with low frequency random modulation (■). (Reproduced from [3])

a psychoacoustic model for the direct control of an audio encoder was presented in [9]. Disch et al. presented a model for the perceptual comparison of differently encoded audio data increasing the sensitivity in time-frequency regions with high comodulation [10]. Here, we propose a comodulation dependent adaptation of the postmasking decay to model Comodulation Masking Release (CMR) effects. This approach explicitely considers the dip-listening effect.

## 2 Methods

### 2.1 Estimation of Comodulation in Filter Bank based Perceptual Models

Assuming a perceptual model using $K$ filters $k$ with $0 \leq k < K$ for approximating the spectral decomposition in the human auditory system.

In the approach of Sinha and Ferreira [9], cross-correlations are estimated by inner products of normalized envelopes in the corresponding bands, which are obtained as output magnitudes of a complex valued filter bank. From those, an "effective bandwidth" for each band is obtained by averaging over cross-correlations with multiple (e.g. 5) lower and upper neighbors. The amount of masking threshold reduction is then derived from this effective bandwidth and the ratio between the minimum and the maximum of the envelope in this band within a certain time interval (e.g. $20 \ldots 30ms$).

This ratio is introduced to take into account the dip listening effect. The overall complexity of these operations, however is relatively high and thus they are applied with a relative low temporal resolution.

In our perceptual model using a complex IIR filterbank, see Section 2.2, comodulation is estimated in the following way. Since magnitude responses of directly neighboring bands significantly overlap, their output magnitudes show strong covariances even with strongly band limited modulated maskers. Therefore, an appropriate distance $d$ between band indices can be selected. The covariance of bands $k$ with $k \geq d$ and $l = k - d$ at sample position $n$ can then be estimated using short time averages $\overline{e_k(n)}$ and correlations $\overline{e_k(n)e_l(n)}$ from the envelopes $e_k(n)$ and $e_l(n)$.

As averaging method first order IIR lowpass filters are used. For example, the averages are calculated as

$$a_i(n) = \overline{e_i(n)} = \rho_i a_i(n-1) + (1-\rho)e_i(n), \quad (1)$$

where the decay factor $\rho_i$ with $0 < \rho_i < 1$ determines the time constant and depends on the band index $i$. The averaging in the correlation calculations are obtained correspondingly. This method has the advantage of low complexity, which even allows sample-wise application.

The following more complex operations are performed on a reduced sampling rate, here for every 32nd sample. To provide output values at every sample instance, the calculated values are just kept until the next update. This behavior is denoted by the use of $m = 32\lfloor n/32 \rfloor$. First, the covariance of bands $k$ and $l$ is estimated using

$$C_{kl}(m) = \overline{e_k(m)e_l(m)} - a_k(m)\,a_l(m) \quad (2)$$

with $a_k(m) = \overline{e_k(m)}$ and $a_l(m) = \overline{e_l(m)}$.

The comodulation strength should be independent of the total strength of the signal. This can be achieved by using the correlation coefficient:

$$C'_{kl}(m) = \frac{C_{kl}(m)}{\sqrt{\overline{e_k^2(m)}\,\overline{e_l^2(m)}}} \quad (3)$$

Another normalization, however, was found to be more useful, since for bands $k$ with $k \geq d$ and $k < K - d$, a combination of correlations with a lower neighbor $l = k - d$ and an upper neighbor $u = k + d$ can be used to give a better indication of comodulations. In this case, the following combination of $C_{kl}$ and $C_{ku}$ and the corresponding averages was found to give good results:

$$s_k(m) = \frac{G\left(\sqrt{C_{kl}(m)} + \sqrt{C_{ku}(m)}\right)}{2a_k(m) + a_l(m) + a_u(m)} \quad (4)$$

where $G$ is a normalization factor to obtain averages similar to the cases, in which only one neighbor with the desired distance $d$ can be used. In those cases, where either no lower or no upper neighbor is available, another normalized covariance with the upper or lower neighbor can be used, i.e.:

$$s_k(m) = \frac{2\sqrt{C_{kj}(m)}}{a_k(m) + a_j(m)} \quad (5)$$

with

$$j = \begin{cases} u & \text{if } k < d \\ l & \text{if } k \geq K - d \end{cases}$$

This way estimates are obtained by averaging over exponentially weighted data giving the strongest weight to the latest samples. This leads to a fast adaptation to changes in the signal characteristics, so that usually no look-ahead is required. Hence, the estimation of comodulation adds no additional latency to the psychoacoustic model.

The comodulation strength $s_k(m)$ is finally converted to a comodulation coefficient $c_k(m)$ within an interval $[0, 1]$ by mapping and clipping:

$$c'_k(m) = \frac{s_k(m) - M_{min}}{M_{max} - M_{min}} \quad (6)$$

and

$$c_k(m) = \begin{cases} 0 & \text{if } c'_k(m) < 0 \\ 1 & \text{if } c'_k(m) > 1 \\ c'_k(m) & \text{else} \end{cases} \quad (7)$$

with clipping parameters $M_{min}$ and $M_{max}$.

In our experiments, $d \approx 1$ Bark, $G = 2.8$, $M_{min} = 0.3$ and $M_{max} = 1$ are used. The decay factor $\rho_i$ is calculated to resemble a decay time of 10 ms for the band with the center frequency of approx. 2.3 kHz.

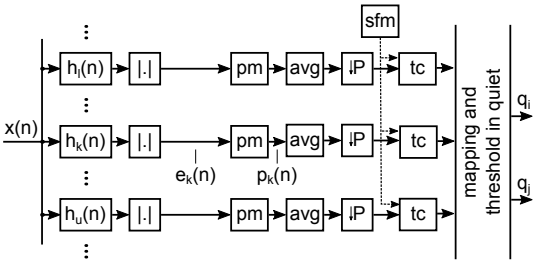



**Fig. 3:** Diagram of the simple perceptual model used in this work. (pm = postmasking control, tc = tonality control)

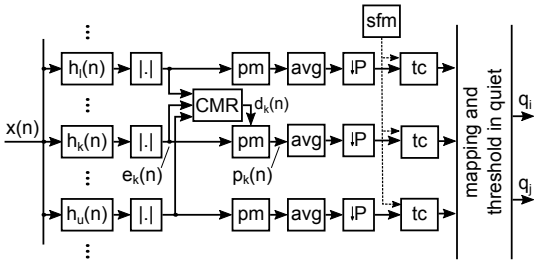



**Fig. 4:** Diagram of the perceptual model using a comodulation masking release dependent postmasking decay. (pm = postmasking control, tc = tonality control, CMR = comodulation masking release, $d_k$ = decay factor)

The decay time is reduced for bands with higher bandwidth (in our case larger $i$) and enlarged for bands with lower bandwidth (in our case smaller $i$).

Since the computationally complex operations are performed at a reduced sampling rate, we consider the complexity moderate compared to the other transforms in a psychoacoustic model and transform coders.

### 2.2 Perceptual model using a complex IIR filterbank

The new approach of controlling the postmasking decay considering the comodulation is implemented in a simple perceptual model. A diagram of the simple model is demonstrated in Figure 3. This perceptual model uses a complex valued IIR filterbank with $K = 104$ third-order minimum phase bandpass filters and a frequency resolution of 4 bands per Bark, resulting in a resonably fine resolution for modelling the auditory system. The filter slopes are designed to take spectral masking effects into account. Details of the filterbank can be found in ref. [11].

Postmasking effects decay more slowly than the impulse responses of the IIR filters, which is modelled in an additional stage. An envelope detector (peak detector) with an appropriate decay factor $d_k$ is used:

$$p_k(n) = \max\left(e_k(n), d_k p_k(n-1)\right) \tag{8}$$

The temporal resolution of the postmasking output $p_k(n)$ can be adjusted to the resolution required for proper quantizer control in the audio coder. This is achieved by averaging followed by downsampling.

To consider tonality, the averaged and downsampled postmasking outputs are weighted by a tonality dependent gain, controlled using spectral flatness estimation, see tonality control (tc) in Figure 3. The spectral flatness estimation is based on multiple FFTs with different time/frequency resolutions calculated from the input audio signal [12, 13].

The output is then fed to a final stage for mapping to an appropriate level and the required frequency resolution for quantizer control, and incorporating the influence of the threshold in quiet. The whole system delivers the quantizer step sizes to be used in an audio encoder, see Section 2.3.

The new approach in this work is the extension of the model by introducing a time-varying postmasking decay factor $d_k(n)$, which is controlled by the comodulation coefficient $c_k(m)$. A diagram of this version is presented in Figure 4. The value of the decay factor $d_k(n)$ is updated at the same reduced sampling rate as $c_k(m)$, so that the corresponding time constant lies between $T_{min}$ and $T_{max}$:

$$\tau_k(m) = T_{max} - c_k(m) \cdot (T_{max} - T_{min}) \tag{9}$$

$$d_k(n) = 1 - \frac{1000}{\tau_k(m) \cdot f_s} \tag{10}$$

with $f_s$ being the sample rate of the IIR filterbank and $T_{min}$ and $T_{max}$ the minimum and maximum postmasking decay time constant.

This is in contrast to the already known method where the thresholds after postmasking are directly reduced by a gain factor dependent on a comodulation coefficient [9].

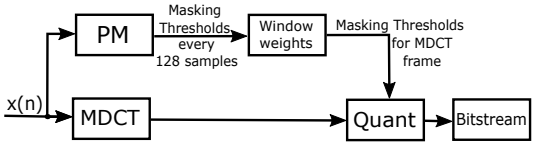



**Fig. 5:** Diagram of the encoding process. (PM = perceptual model, Quant = quantizer)

### 2.3 Perceptual models in the encoder framework

A basic transform coder using standardized transforms (as e.g. in AAC) was choosen for the experiments. The transform coder uses MDCT long blocks with sine windows, a window length of 2048 audio samples and a hop size of 1024 audio samples. In general, the proposed method can be implemented in any encoder framework using a psychoacousic model considering postmasking decay.

Three different perceptual models are used in the encoder framework and the following listening test:

- **CMR no**: perceptual model using always a postmasking decay of 2.4 ms.
- **CMR pm**: perceptual model using a varying postmasking decay between $T_{min} = 2.0$ ms and $T_{max} = 10.0$ ms based on the comodulation coefficient $c_k(n)$.
- **CMR gain**: perceptual model reducing the masking threshold based on the comodulation coefficient $c_k(n)$, similar to [9].

The threshold in quiet is set to be the same for all models to receive a comparable transmitted frequency bandwidth. To adjust the bitrate, the received thresholds after the postmasking and tonality correction can be scaled up or down, before taking the maximum with the threshold in quiet. The masking thresholds are calculated every 128 samples. To receive the final masking thresholds for a full audio frame, a weighting with an inverse sine window is performed [11]. This weighting shapes the quantization error considering the window function. Afterwards, the maxima for each frequency range are taken as masking thresholds for the quantizer.

| item | CMR no | CMR pm | CMR gain |
|---|---|---|---|
| applause | 49.4 | 50.0 | 50.6 |
| castanets | 57.2 | 58.0 | 57.8 |
| es02 | 50.6 | 44.8 | 52.0 |
| fatboy | 70.0 | 77.7 | 78.3 |
| hallelujah | 37.6 | 37.8 | 34.8 |
| harpsi | 42.0 | 43.1 | 40.3 |
| pitchpipe | 38.1 | 39.9 | 37.4 |
| rock | 39.5 | 39.3 | 37.0 |
| vega | 53.0 | 50.3 | 51.8 |

**Table 1:** Bitrates in kbps for individual items for the different psychoacoustic models used in the listening test.

The quantizer step sizes calculated by the perceptual model are used in the transform coder. A simple bitstream format (similar to AAC) was used to transmit the bitstream and for bitrate measurements. A diagram of the encoding process is shown in Figure 5.

## 3 Results

Since the encoder framework operates with variable bit rates, the perceptual models were tuned to achieve equal average bit rates for a very long item concatenated from a multitude of test signals. Nevertheless, the bit rates for individual items used in the test can differ, see Table 1. Thus, the test can reveal which of the models better indicates the real time-varying thresholds. The following listening test uses mono items and the average bitrate was set to 48 kbit/s. The test contains items dominated by tonal signals (hallelujah, harpsichord, pitchpipe and rock), as well as items with strong transient components (applause, castanets and fatboy) and pure speech/vocal items (es02 and vega).

As test method, a MUSHRA test is used, standardized in ITU-R BS.1534-3 [14]. The test was performed using the freely available tool webMUSHRA [15]. The listening test contains the above described audio items encoded with the

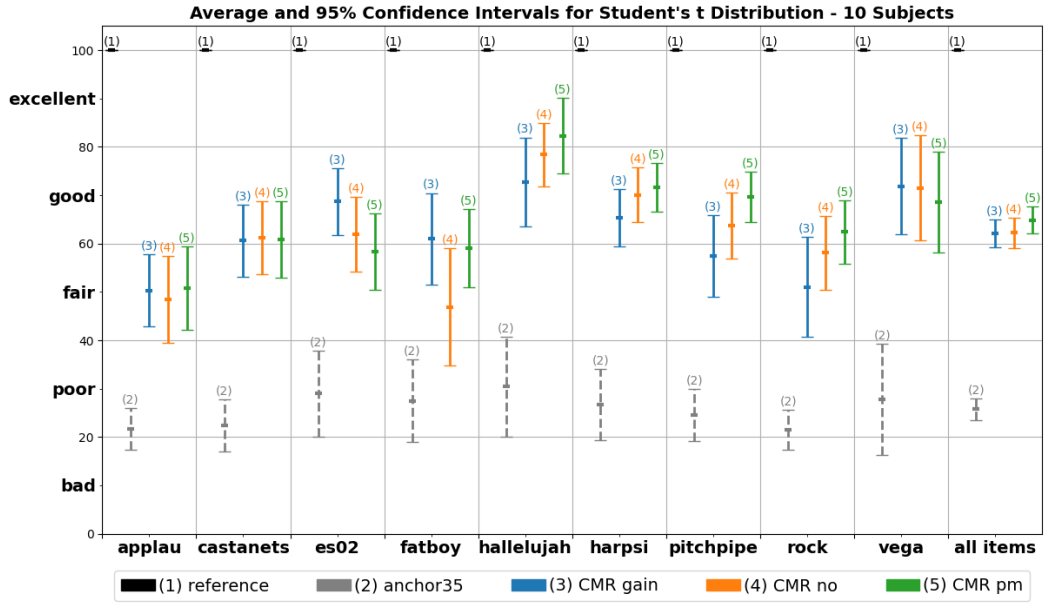



**Fig. 6:** Listening test results. The CMR postmasking model (CMR pm) shows an overall improvement over the CMR gain model (CMR gain) and the basic model without any CMR consideration (CMR no).

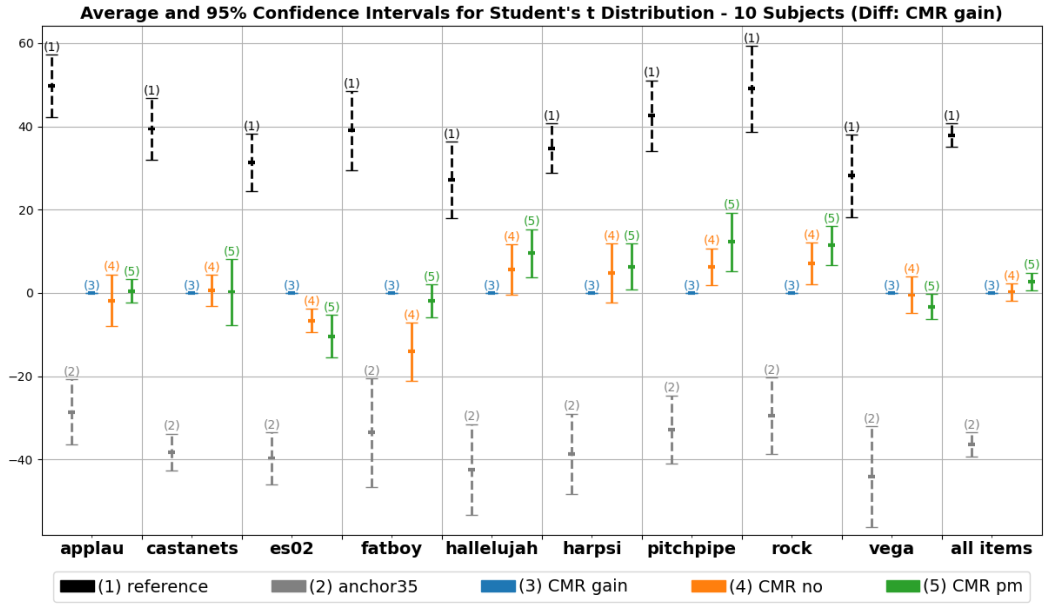



**Fig. 7:** Difference Plot to the CMR gain model. For tonal items a significant improvement in audio quality of the CMR pm model can be observed. Transient items show the same audio quality, pure vocal/speech items show a degradation.

three models (CMR no, CMR pm and CMR gain) as well as a 3.5 kHz anchor and a hidden reference. Ten expert listeners took part in the test using Stax SR-404 headphones on a Stax SRM-006T-II amplifier.

The results are illustrated in Figure 6. The proposed CMR postmasking model shows a very good performance and an overall improvement of 2.5 MUSHRA points over the CMR gain model and the basic model without any CMR consideration, see aggregated scores (all items). For all tonal items, the CMR postmastking model performs the best, although the bitrate of the individual items is, if at all, only marginally higher than the ones of the model not considering CMR, see Table 1.

Significant differences are revealed in the difference plot to the CMR gain model, see Figure 7. Especially tonal items as hallelujah, harpsi, pitchpipe and rock show a significant improvement in the CMR postmasking model over the CMR gain model. For transient items the two models using CMR show almost no difference in audio quality. An exception is the fatboy item, where a significant enhancement of the models using CMR is observed. The pure speech/vocal items, especially es02 which contains german male speech, show a significant degradation in the CMR postmasking model compared to the CMR gain model. One reason is the drop of more than 10% in bitrate for the es02 item with the CMR postmasking model compared to the other models, see Table 1. Comparing the CMR postmasking model to the model not using CMR, the degradation for the pure speech/vocal items in the CMR postmasking model is not significant. Still, this behavior needs further investigation and tuning, to improve the performance of the CMR postmasking model.

Finally, an analysis using an objective measurement of perceived audio quality was performed. An implementation of the advanced PEAQ model, described in ITU-R BS.1387 [16], was used. The Objective Difference Grade (ODG) scores show for the tonal items harpsi, pitchpipe and rock the similar trend as the listening test with expert listeners. For those items the CMR postmastking model outperforms the others. All other items, expecially the transient ones, show only very small differences in the ODG scores. The objective measurement analysis underlines the performance of the CMR postmasking model, especially for tonal items. However, some global trends of the listening test results are not well captured by PEAQ.

## 4 Summary & Outlook

The proposed method by controlling the masking threshold decay via a comodulation strength shows an advantage in the overall bit distribution over the known method by reducing the masking threshold gain. Especially for tonal dominated items an improved audio quality is observed. For speech items further investigations and tunings are needed. A possible improvement could be to include more neighboring bands in the comodulation calculation to improve the stability. Furthermore, the combination of both, the postmasking and gain based CMR model, could be investigated. The experiments in this work were performed using a basic transform coder based on only MDCT long blocks. The masking thresholds in the proposed psychoacoustic models are calculated every 128 samples. To make use of the better time resolution, especially for transient signals, the proposed approach can be extended to additionally use short blocks in the transform coder.

## 5 Acknowledgements

We thank Yaqiong Hou, Paula Schäfer, Max Betz and Paul Meier for their contribution to the software implementation and support in informal listening.

## References

[1] Schroeder, M. R., Atal, B. S., and Hall, J., "Optimizing digital speech coders by exploiting masking properties of the human ear," *The Journal of the Acoustical Society of America*, 66(6), pp. 1647–1652, 1979.

[2] Brandenburg, K., Gerhser, H., Seitzer, D., and Sporer, T., "Transform coding of high quality digital audio at low bit rates-algorithms and implementation," in *IEEE International Conference on Communications, Including Supercomm Technical Sessions*, pp.

932–936 vol.3, 1990, doi:10.1109/ICC.1990.117211.

[3] Moore, B., *An Introduction to the Psychology of Hearing: Fifth Edition (pp 100 - 105)*, Emerald, 2008.

[4] Schnupp, J., Nelken, I., and King, A. J., *Auditory Neuroscience: Making Sense of Sound (pp 228 - 233)*, The MIT Press, 2011.

[5] Las, L., Stern, E. A., and Nelken, I., "Representation of tone in fluctuating maskers in the ascending auditory system," *Journal of Neuroscience*, 25(6), pp. 1503–1513, 2005.

[6] Buus, S., Zhang, L., and Florentine, M., "Stimulus-driven, time-varying weights for comodulation masking release," *The Journal of the Acoustical Society of America*, 99(4), pp. 2288–2297, 1996.

[7] Buus, S., "Release from masking caused by envelope fluctuations," *The Journal of the Acoustical Society of America*, 78(6), pp. 1958–1965, 1985.

[8] Verhey, J. L., Dau, T., and Kollmeier, B., "Within-channel cues in comodulation masking release (CMR): Experiments and model predictions using a modulation-filterbank model," *The Journal of the Acoustical Society of America*, 106(5), pp. 2733–2745, 1999, ISSN 0001-4966, doi:10.1121/1.428101.

[9] Sinha, D. and Ferreira, A. J. S., "A New Broadcast Quality Low Bit Rate Audio Coding Scheme Utilizing Novel Bandwidth Extension Tools," in *Audio Engineering Society Convention 119*, 2005.

[10] Disch, S., van de Par, S., Niedermeier, A., Burdiel Pérez, E., Berasategui Ceberio, A., and Edler, B., "Improved Psychoacoustic Model for Efficient Perceptual Audio Codecs," in *Audio Engineering Society Convention 145*, 2018.

[11] Taghipour, A., Knoelke, N., Edler, B., and Ostermann, J., "Combination of different perceptual models with different audio transform coding schemes: Implementation and evaluation," in *AES 129th Convention, San Francisco, USA*, p. 8283, 2010.

[12] Taghipour, A., Jaikumar, M. C., Edler, B., and Stahl, H., "Partial spectral flatness measure for tonality estimation in a filter bank-based psychoacoustic model for perceptual audio coding," in *Audio Engineering Society (Convention) 2013*, 2013.

[13] Taghipour, A., Jaikumar, M. C., and Edler, B., "A psychoacoustic model with Partial Spectral Flatness Measure for tonality estimation," in *2014 22nd European Signal Processing Conference (EUSIPCO)*, pp. 646–650, 2014.

[14] ITU-R, "Recommendation BS.1534-3: Method for the subjective assessment of indermediate quality levels of coding systems," 10/2015.

[15] Schoeffler, M., Bartoschek, S., Stöter, F.-R., Roess, M., Westphal, S., Edler, B., and Herre, J., "webMUSHRA - A comprehensive framework for web-based listening tests," *Journal of Open Research Software*, 6(1), 2018.

[16] ITU-R, "Recommendation BS.1387: Method for objective measurements of perceived audio quality," 2001.