A PSYCHOACOUSTIC MODEL WITH PARTIAL SPECTRAL FLATNESS MEASURE FOR TONALITY ESTIMATION

Armin Taghipour[1]*, Maneesh Chandra Jaikumar*[2]*, and Bernd Edler*[1]

[1] International Audio Laboratories Erlangen, Am Wolfsmantel 33, 91058 Erlangen, Germany
[2] Fraunhofer Institute for Integrated Circuits IIS, Am Wolfsmantel 33, 91058 Erlangen, Germany
armin.taghipour@audiolabs-erlangen.de

ABSTRACT

Psychoacoustic studies show that the strength of masking is, among others, dependent on the tonality of the masker: the effect of noise maskers is stronger than that of tone maskers. Recently, a Partial Spectral Flatness Measure (PSFM) was introduced for tonality estimation in a psychoacoustic model for perceptual audio coding. The model consists of an Infinite Impulse Response (IIR) filterbank which considers the spreading effect of individual local maskers in simultaneous masking. An optimized (with respect to audio quality and computational efficiency) PSFM is now compared to a similar psychoacoustic model with prediction based tonality estimation in medium (48 kbit/s) and low (32 kbit/s) bit rate conditions (mono) via subjective quality tests. 15 expert listeners participated in the subjective tests. The results are depicted and discussed. Additionally, we conducted the subjective tests with 15 non-expert consumers whose results are also shown and compared to those of the experts.

Index Terms**—** Perceptual Model, Psychoacoustic Model, Perceptual Audio Coding, Spectral Flatness, Tonality Estimation

1. INTRODUCTION

One of the main goals of audio coding has been to reduce the requirements for storage and transmission of the audio signal data via compression. This is partly done by employing predictive and entropy coding, which reduce redundancies in the signal. However, redundancy reduction alone does not lead to low bit rate audio coding. Hence, in transform-based perceptual audio coding, psychoacoustic models (PM) are used to control the quantizers of spectral components and consequently reduce irrelevancies in the audio signal. A block diagram of such an audio coder is shown in Figure 1. In the optimal case, after compression, the quantization noise should be imperceptible to human listeners at the output of the decoder - i.e. it should lie just below the masking threshold. Audio coders such as mp3 (MPEG-1/2 Audio Layer 3 [1,2]) or AAC (MPEG-2/4 Advanced Audio Coding [3, 4]) use such psychoacoustic models which approximate the masking effects to the best possible extent. However, at medium and low bit rates, the estimated masking threshold has to be violated. Especially for low bit rate coding of speech and complex sounds, we see room for improvement in estimating masking thresholds which would bring the best possible subjective audio quality. Quick switches between tonal and transient-like segments, especially diverging characteristic changes in various spectral parts, are challenging.

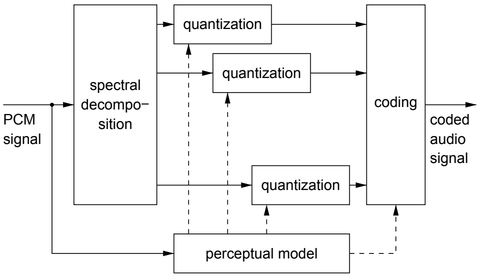


Fig. 1. Basic structure of a transform based audio encoder. A transform decomposes frames of the input audio signal in their spectral components. The psychoacoustic model calculates estimated masking thresholds for the frames with which it controls quantization steps for the individual subbands.

Psychoacoustic studies show an asymmetry of masking strengths of tonal and noise-like maskers. Narrowband noise has a stronger masking effect than a tone of the same energy placed at its center frequency [5–8]. An example of modeling this asymmetry for perceptual audio coding is described in [9, 10]. Furthermore, latest studies show that human listeners can not necessarily distinguish easily between narrowband noise and tone bursts. There is a duration dependent distinction threshold which lies somewhere between 3 and 30 ms for frequencies up to 2.7 kHz. The duration threshold depends on the center frequency and bandwidth of the narrowband noise [11]. The higher the frequency and the wider the bandwidth, the lower (shorter) the threshold. This is in accord with the fact that the temporal resolution of the human auditory system increases with frequency [6]. Unlike the conventional psychoacoustic models, our model takes into account this frequency dependency at the stage of tonality estimation, and consequently by estimating the masking threshold level.

In Section 2, a filterbank based PM is described. Masking thresholds are calculated with respect to the frequency dependency of masking for the individual band-pass filters [12]. More specifically, an optimized Partial Spectral Flatness Measure (PSFM) estimates tonality. By means of subjective tests (described in Section 3) this model is compared to a similar model which uses prediction based tonality estimation. In Section 4, the results are presented and discussed in detail.

2. PSYCHOACOUSTIC MODEL (PM)

Transform based codecs process input signals framewise. For example, in MPEG Advanced Audio Coding (AAC) [3, 4] there is a distinction between usage of long frames of 2048 samples for stationary parts, and short frames of 256 samples for transient parts of signals. In doing so, either high frequency resolution, or high temporal resolution is achieved, respectively. As long as the signal segments are strongly harmonic or strongly transient, this distinction is easily achieved. However, for more complex sounds, an individual segment of the signal could have different characteristics along the spectral axis. Therefore, it is desirable to have a PM which analyzes the signal in a way that the resulting masking threshold can be used for short frames as well as long frames.

Recently, we introduced a PM [12,13]. It consists of complex IIR[1] band-pass filters which take into account the spreading in simultaneous masking [5]. Their center frequencies are spaced at intervals of Bark/4 (Bark scale: [5, 14]). The structure of the PM, the filters' frequency responses and the calculation of the masking thresholds are detailed in [12, 13]. The model considers temporal and spectral characteristics of the signal: for a frame of 1024 samples, it first calculates masking thresholds for 8 short frames of 128 samples (which leads to high temporal resolution), and further, a masking threshold for the long frame by combining the individual thresholds of the short blocks within the current long block [13].

2.1. Partial Spectral Flatness Measure

Spectral Flatness Measure (SFM) has broadly been used as a measure of tonality. In [15], SFM is described for the continuous-frequency case. J. D. Johnston applied discrete SFM to perceptual audio coding [9, 10, 16]. His model deployed SFM as a distinction measure between tone and noise maskers while calculating masking thresholds [9, 10]. The model used short-time power spectrum with fixed analysis frame length for Bark-wide bands.

In [12], we introduced the tonality measure PSFM as the ratio of geometric and arithmetic mean of short-time squared magnitude spectrum, $|S_{st}(k)|^2$. In that model, the magnitude spectrum of each IIR-filter output was individually analyzed by a Discrete Fourier Transform (DFT), and PSFM was calculated for a range of coefficients around its center frequency:

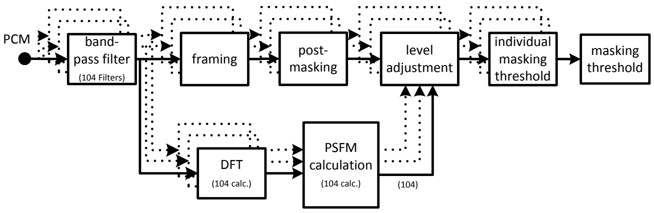


Fig. 2. Block diagram of the initial PSFM and calculation of masking thresholds in the PM [12]. 104 DFTs are applied to the outputs of the filters. Spectral resolution of PSFM calculation varies with the DFT length; it is higher for low frequencies. All in all, four different DFT lengths were used. From the individual masking thresholds of 104 bands, a global masking threshold is calculated for a short frame.

$$PSFM = \frac{\sqrt[N]{\prod_{k=N1}^{N2} |S_{st}(k)|^2}}{\frac{1}{N}\sum_{k=N1}^{N2} |S_{st}(k)|^2}, \tag{1}$$

where $0 \leq PSFM \leq 1$, and $N = N2 - N1 + 1$. The limits $N1$ and $N2$ were chosen in a way that for each filter output the range extended to the "double of its efficient bandwidth" (more details in [12]).

The outcome of this model was compared to a similar model with predictor based tonality estimation [13] by means of subjective tests. Although the models showed distinct abilities in compression of different types of audio items, there was no significant difference between these models.

As depicted in Figure 2, in this initial model, the "shaped" (filtered) outputs of the preceding filter stage were the inputs of the PSFM [12]. This led to two shortcomings: first, since PSFM was not calculated from the original input signal, even for white noise input, the measure has never achieved the maximum flatness. Second, the calculation had a high computational complexity whereby 104 Fast Fourier Transforms (FFT) had to be considered.

In order to overcome these problems, a further optimized model for PSFM is presented here. The block diagram of the PM (including the optimized PSFM) is shown in Figure 3. For the input signal, short-time spectra of different spectral resolutions are generated by DFTs of different lengths (4096, 2048, or 1024 for low, middle and high frequencies, respectively). PSFM is calculated corresponding to the individual band-pass filters (over double of their efficient bandwidth). In doing so, only 3 FFTs were needed.

A transform audio coding scheme with Modified Discrete Cosine Transform was chosen. We used fixed frame length of 1024 samples for coding. The PMs were applied to the coding structure. Although entropy coding was not applied, entropy rates were estimated. By applying a scaling factor to the calculated masking thresholds, the codecs could be brought to desired average data rates, estimated for a large set of stan-

[1] Infinite Impluse Response

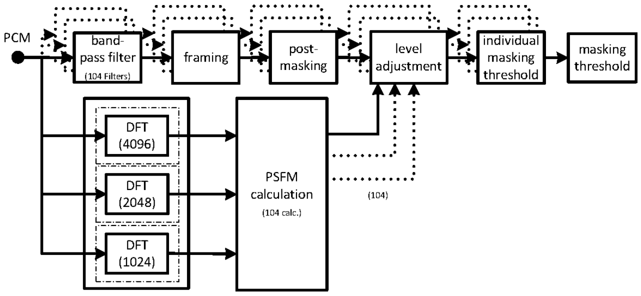


Fig. 3. PM with the optimized PSFM: varying DFT lengths are shown for analysis of the input signal at low, middle and high frequencies (4096, 2048 and 1024, respectively). From the individual masking thresholds of 104 bands, a global masking threshold is calculated for a short frame.

dard test audio signals. For subjective tests, the codecs under test were designed to have equal average entropy rates of 48 kbit/s (medium quality) and 32 kbit/s (low quality).

3. SUBJECTIVE TESTS

By means of MUSHRA (Multiple Stimuli with Hidden Reference and Anchor) [17] test, the optimized PSFM was compared to the prediction based tonality estimation. Initially, we conducted a pilot MUSHRA test with 3 audio items, in which 5 listeners participated. The results of this test showed a tendency for higher subjective quality ratings of PSFM for all 3 items. The average entropy rate was 48 kbit/s. Based on these results, we decided to test the items at different average entropy rates. The hypothesis was that, presumably, in a lower quality range the differences would be more distinct.

For the final MUSHRA test a set of 9 items (all mono with 48 kHz sampling frequency) with various characteristics were used in two item-groups:

1. speech and vocal
   - es01 - Female vocal, “Tom’s Diner”
   - es02 - German male speech
   - es04 - English male speech
   - es05 - German female speech
2. music
   - si01 - Harpsichord
   - si02 - Castanet
   - sc02 - Symphony orchestra
   - sc03 - News intro
   - pipe-short10 - Pitch-pipe

These two item-groups were presented to the subjects in two separate test sessions, otherwise the test would have taken more than 50 minutes, including the training phase. Subjects took between 15 and 25 minutes for each test session.

The subjects performed the MUSHRA test using a Graphical User Interface (GUI) developed by Fraunhofer-IIS. The GUI is depicted in Figure 4. For each audio test item, subjects rated the quality of each condition in comparison to the reference/original. The conditions for each item were:

- a hidden reference
- the coded version with predictor based tonality estimation with average entropy rate of 48 kbit/s
- the coded version with PSFM tonality estimation with average entropy rate of 48 kbit/s
- the coded version with predictor based tonality estimation with average entropy rate of 32 kbit/s
- the coded version with PSFM tonality estimation with average entropy rate of 32 kbit/s
- a low pass filtered anchor (f_c = 3.5 kHz)

Subjects were asked to rate the hidden reference at 100 and the low anchor at 20, as far as they could detect these. In the MUSHRA test, the order of appearance of the audio files (items) and the order of the codecs (conditions) were randomized, as described in detail in [17].

Table 1 shows the estimated entropies for the 9 items for both cases (48 and 32 kbit/s), and for white noise (not included in the subjective test; used only as a controlling item for compression ability throughout the implementation).

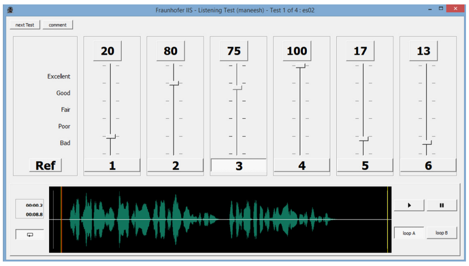
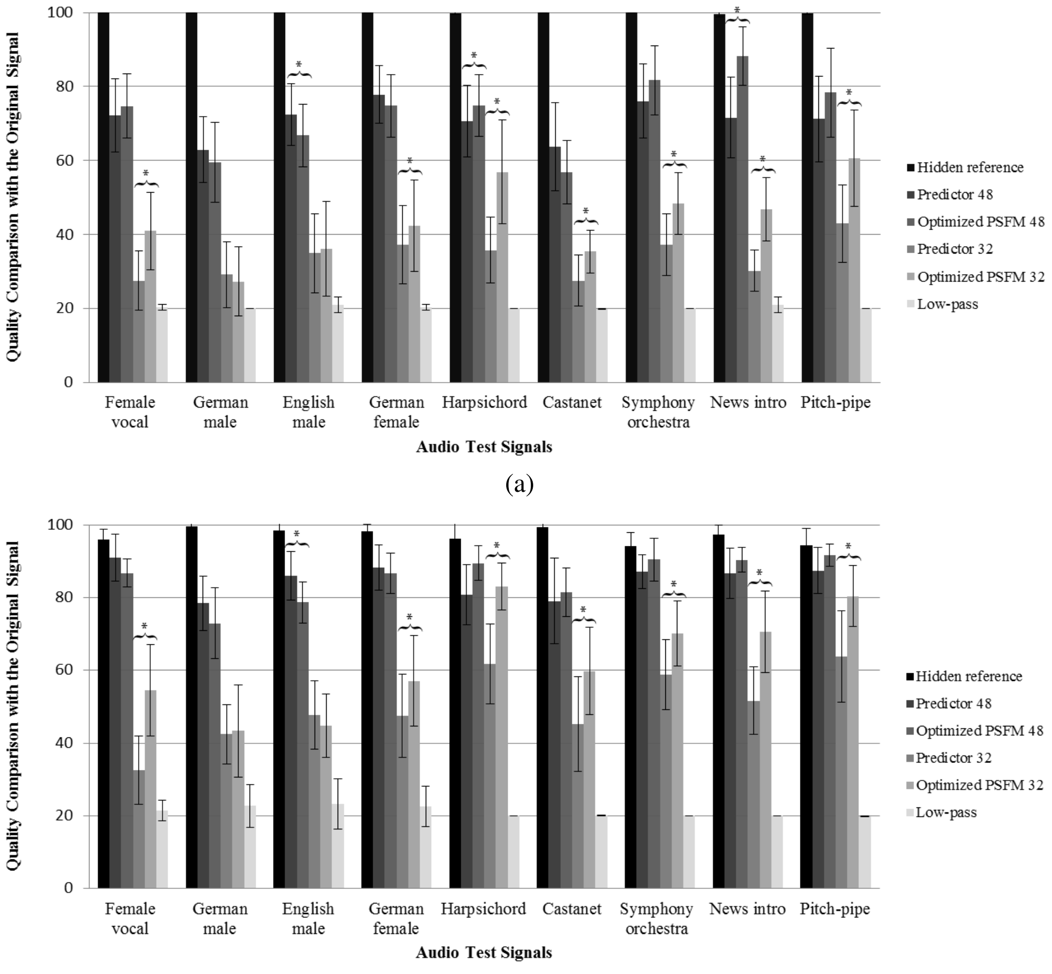


Fig. 4. The GUI for the MUSHRA test: e.g. for the item “German male speech”, 6 different conditions were compared to the reference. The conditions (hidden reference, low anchor, Predictor48, PSFM48, Predictor32 and PSFM32) were randomly placed from 1 to 6.

Items	Predict. (48)	PSFM (48)	Predict. (32)	PSFM (32)
Female vocal	50.30	51.58	31.58	32.16
Ger. m. speech	57.05	47.73	36.39	29.76
Eng. m. speech	51.86	44.31	33.16	28.12
Ger. f. speech	61.93	53.88	42.66	36.53
Harpsichord	45.86	53.12	30.86	37.32
Castanet	73.12	69.08	51.61	48.84
S. orchestra	42.97	42.55	28.17	27.76
News intro	47.68	51.05	31.58	33.36
Pitch-pipe	50.97	48.69	38.00	35.51
Noise	56.50	66.10	34.65	45.39

Table 1. Estimated entropy rates in kbit/s for different items.



(a)

(b)

Fig. 5. Results of MUSHRA tests for (a) 15 expert and (b) 15 non-expert subjects. Abscissa shows different items (and their conditions). Subjective quality ratings of different conditions of the items are depicted. Ordinate shows a scale between 0 and 100 (spanning a quality range from bad to excellent). Mean values and confidence intervals (95%) are shown.

Prior to the test, subjects were asked to take a similar MUSHRA training test with 2 audio items and 4 conditions. All participants were asked whether they consider themselves to have normal hearing, whether they have ever experienced hearing loss and whether they are healthy at the moment of study (no flu, etc.). Furthermore, audiometry tests were conducted with the subjects to confirm their hearing threshold. For the actual MUSHRA test, subjects were chosen who had had experience in MUSHRA tests within the last year, and who are considered as expert listeners according to [17].

Since not all consumers are experts in audio coding, we also conducted the same series of MUSHRA tests with non-expert listeners. The results of 4 of the non-expert participants are not considered in the following statistics (based on their responses to hidden reference and anchor). They either did not understand the task sufficiently, or had severe difficulty in distinguishing the hidden reference and anchor from other conditions. All 30 subjects, whose results are listed in the statistics, fulfilled the health criterion, had normal hearing and responded acceptably for the hidden reference and anchor, according to the criteria of [17]. Average age of the experts and non-experts was 32 and 23, respectively.

4. RESULTS

MUSHRA test results of the 15 expert subjects are depicted in Figure 5a. The 9 audio items are listed along the abscissa including their 6 conditions. The graph shows average subjective quality ratings for the different conditions of the audio items over the ordinate. For each item the average quality rating and the corresponding 95% confidence interval are shown. We analyzed the results and compared those of the two "48 kbit/s"-coded versions to each other and those of the two "32 kbit/s"-coded versions to each other, using two-tailed paired t-tests. Significant differences ($p < 0.05$) are shown with " * " in Figure 5a. For most of the items in medium bit rate cases, there are no significant differences between the models. However, in low bit rate cases, for 7 items, the model with PSFM was rated significantly higher than the predictor based model.

Figure 5b shows results of the MUSHRA test with non-expert listeners. As expected [18], the results of different coded versions are closer to each other and to the hidden reference. Significant differences are marked with " * ", as above. Here as well, the results of the same 7 items show significant preference of PSFM in low bit rate cases.

There is a preference for the optimized PSFM for non-speech signals, most significantly at low bit rates. For speech signals, despite higher compression (lower entropy rates, Table 1), PSFM achieves comparable quality to the other model.

5. CONCLUSION

An improved tonality estimation method is described which was implemented in a filterbank-based PM. The model can work independent of the block lengths of audio coding, as it operates with a frequency dependent temporal and spectral resolution well adapted to that of the human auditory system. The new enhanced PSFM has computational advantages over the initial model and leads to good subjective qualities. For non-speech signals, the optimized PSFM reaches higher ratings, especially in the low bit rate case where all the differences are significant. This improved version is computationally more efficient, and could be an alternative to the prediction based tonality estimation.

REFERENCES

[1] ISO/IEC JTC/SC29/WG11 MPEG International Standard ISO/IEC 11172, "Coding of moving pictures and associated audio for digital storage media at up to about 1.5 mbit/s," 1993.

[2] G. Stoll and K. Brandenburg, "The iso/mpeg-audio codec: A generic standard for coding of high quality digital audio," in *Audio Engineering Society Convention 92*, Mar 1992.

[3] ISO/IEC JTC1/SC29/WG11 MPEG International Standard ISO/IEC 13818-7, "Generic coding of moving pictures and associated audio: Advanced audio coding," 1997.

[4] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, and M. Dietz, "Iso/iec mpeg-2 advanced audio coding," *J. Audio Eng. Soc*, vol. 45, no. 10, pp. 789–814, 1997.

[5] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*, Springer series in information sciences. Springer, 2007.

[6] B.C.J. Moore, *An Introduction to the Psychology of Hearing*, Academic Press, 2012.

[7] R. P. Hellman, "Asymmetry of masking between noise and tone," *Perception & Psychophysics*, vol. 11, no. 3, pp. 241–246, 1972.

[8] H. Gockel, B.C. Moore, and R.D Patterson, "Asymmetry of masking between complex tones and noise: the role of temporal structure and peripheral compression.," *The Journal of the Acoustical Society of America,*, vol. 111, pp. 2759–70, 2002.

[9] J.D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *Selected Areas in Communications, IEEE Journal on*, vol. 6, no. 2, pp. 314–323, Feb 1988.

[10] J.D. Johnston, "Estimation of perceptual entropy using noise masking criteria," in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, Apr 1988, pp. 2524–2527 vol.5.

[11] A. Taghipour, B. Edler, M. Amirpour, and J. Herre, "Dependency of tonality perception on frequency, bandwidth and duration," *Proceedings of Meetings on Acoustics*, vol. 19, no. 1, pp. –, 2013.

[12] A. Taghipour, M.C. Jaikumar, B. Edler, and H. Stahl, "Partial spectral flatness measure for tonality estimation in a filter bank-based psychoacoustic model for perceptual audio coding," in *AES 134th Convention, Rome, Italy*, 2013, p. 8824.

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

[14] E. Zwicker, G. Flottorp, and S. S. Stevens, "Critical band width in loudness summation," *The Journal of the Acoustical Society of America*, vol. 29, no. 5, pp. 548–557, 1957.

[15] A. Gray and J. Markel, "A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 22, no. 3, pp. 207–217, 1974.

[16] J.D. Johnston and S.S. Kuo, "Audio signal processing based on a perceptual model," Sept. 10 2003, EP Patent App. EP20,030,003,261.

[17] Recommendation ITU-R BS.1534-1, "Method for the subjective assessment of intermediate quality level of coding systems," 2001-2003.

[18] N. Schinkel-Bielefeld, N. Lotze, and F. Nagel, "Audio quality evaluation by experienced and inexperienced listeners," *Proceedings of Meetings on Acoustics*, vol. 19, no. 1, pp. –, 2013.