ENVELOPE ANALYSIS METHODS FOR TONALITY ESTIMATION

Armin Taghipour, Bharadwaj Desikan, and Bernd Edler

International Audio Laboratories Erlangen
A joint institution of Friedrich-Alexander-University Erlangen-Nürnberg (FAU) and
Fraunhofer Institute for Integrated Circuits (IIS), Am Wolfsmantel 33, 91058, Erlangen, Germany
armin.taghipour@audiolabs-erlangen.de

ABSTRACT

In perceptual audio coders, the audio signal masks the quantization noise. The masking effectiveness depends on the degree of tonality/noisiness of the signal. Hence, in psychoacoustic models (PM) of perceptual coders, the level of the estimated masking thresholds can be adjusted by tonality estimation methods. This paper introduces three envelope analysis methods for tonality estimation: optimized amplitude modulation ratio (AM-R), auditory image correlation, and temporal envelope rate. The methods were implemented in a filter bank-based PM. In a subjective quality test, they were compared to each other and to another existing method, partial spectral flatness measure (PSFM). The PSFM and the AM-R were rated significantly higher than the other methods.

Index Terms— Perceptual Model, Psychoacoustic Model, Perceptual Audio Coding, Tonality Estimation

1. INTRODUCTION

Perceptual audio coders exploit the masking properties of the auditory system to reduce the data rate of the input signal. Commonly, the input is split into frames and decomposed into a number of subbands, such that it is divided into time-frequency segments. A psychoacoustic model (PM) estimates the masking threshold (MT) evoked by the input and controls the bit allocation for the time-frequency segments.

The effectiveness of the masking depends on the tonality/noisiness of the signal. This phenomenon is called "asymmetry of masking." Hellman [1] showed that a noise masks a tone more effectively than vice versa. Hall [2] extended these results to narrowband noise probes and maskers with irregular temporal structure and differing bandwidths. For probe bandwidths exceeding that of the masker, the situation that is predominantly relevant for perceptual coders, thresholds decreased with increasing probe bandwidth and with decreasing masker bandwidth. Hall suggested the results might be explained in terms of the temporal structure of the stimuli, specifically the extent to which the modulation pattern of the signal resembled that of the masker. Verhey [3] performed a similar study and found similar results.

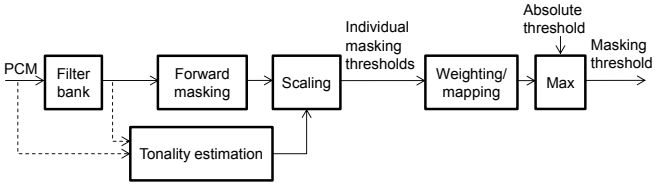

Fig. 1. The psychoacoustic model [7].

In perceptual audio coders, the audio content may include complex tones as well as noise-like sounds. Gockel et al. [4] measured the asymmetry of masking between harmonic complex tones and wideband noise. The irregular noise was a stronger masker. These studies indicate that tone-like sounds with no or slow envelope fluctuations and/or very regular envelope fluctuations are less effective maskers of a noise (with equal or greater bandwidth) than noise-like sounds with strong, irregular and more rapid envelope fluctuations [5].

The results of recent psychoacoustic studies on distinction between noise and tone indicate that envelope analysis methods for the tonality estimation should consider varying temporal resolution when analyzing segments with low, middle, and high center frequencies [6]: longer analysis buffers should be used for lower frequencies. Additionally, a recent study showed that the masked thresholds increased with increasing masker bandwidth and were lowest for medium center frequencies [5]. This should also be taken into account in the development of tonality estimation methods.

2. PSYCHOACOUSTIC MODEL

Taghipour et al. [7–9] introduced a PM which is shown in Figure 1. A filter bank decomposes the signal into its spectral components [8–10]. The filters were designed based on the Bark scale [11], similar to the subbands in most conventional codecs. The center frequencies of neighboring filters are spaced $\frac{1}{4}$ Bark apart from each other. The bandpass filters, which take into account the spreading in simultaneous masking, were designed from reversed masking curves [8,11]. For a sampling rate of 48 kHz and with a Bark-based design, the filter bank consisted of a total of 104 complex, minimum-phase, infinite impulse response (IIR) filters [9, 10].

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

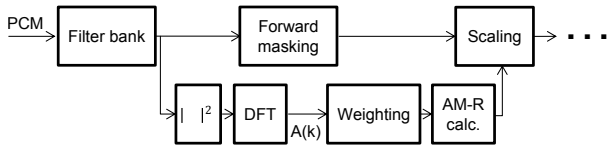

Fig. 2. Block diagram of the optimized AM-R. Only one output channel of the filter bank is shown. Short-time magnitude spectral coefficients of the squared envelope of the filter output are weighted for the calculation.

Forward masking is modeled with a decaying exponential. Based on the degree of tonality of the time-frequency segments, the estimated MTs are scaled; see Section 3. The model can work with short or long coding blocks. First, MTs of short segments are calculated. When long blocks are used, the estimated MTs of short blocks are combined to generate one MT for a long block taking into account the weighting of quantization error by the synthesis window.

Since the spreading in simultaneous masking has already been taken into account in the filter design, no further superposition was necessary. The high-resolution MTs estimated by the PM are mapped onto the spectral resolution of the transform. The final MT is compared to the absolute threshold of hearing. Whenever the levels of the spectral components of the MT are below the absolute threshold, they are replaced by the level of the absolute threshold at those spectral points.

3. TONALITY ESTIMATION METHODS

In the following, four methods are presented that estimate the tonality/noisiness of time-frequency segments of signals.

3.1. Amplitude modulation ratio (AM-R)

Chen et al. [10] introduced a measure referred to as amplitude modulation ratio (AM-R). The basic idea is that the envelope of the output of a filter can indicate tonality; for details, see [10]. The AM-R is optimized in this paper [7].

The bandwidth of a noise can be related to the degree of fluctuations in its envelope [3, 5, 12]. Hence, envelope analysis using the amplitude modulation allows an estimation of the effective bandwidth of the masking signal. As mentioned earlier, in the context of the asymmetry of masking in coding, Hall [2] suggested to analyze the temporal structure of the input, since the threshold decreases with decreasing masker bandwidth. The relation between the modulation frequency and the bandwidth of a masker can be most easily explained by the example of a complex-tone masker: the resulting modulation frequencies are proportional to the frequency difference of the tones, and thus, to the bandwidth of the masker. A weighting function $w(k)$ was added to the formula of the initial AM-R to take this into account.

The optimized AM-R computes the tonality after weighting the modulation frequency components of the modulation spectrogram with a perceptual weighting function $w(k)$:

$$\text{AM-R} = \frac{\sum_{k=1}^{\frac{l}{2}} |w(k) \times A(k)|}{|A(0)|} \quad (1)$$

where $A(k)$ is the amplitude modulation spectrum. The optimized AM-R processing is depicted in Figure 2. Different spectral resolutions were generated by using different DFT lengths in different bands (l = 4096; 2048; 1024; 512; 256), which led to higher spectral resolution for the calculation of AM-R in low frequency regions

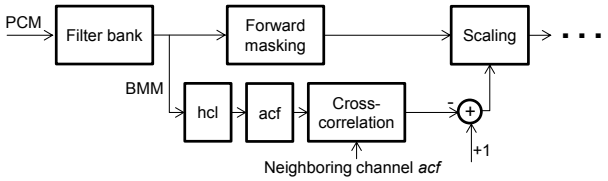

Fig. 3. Block diagram of the AIC. The real-part of the filter output is used as BMM. hcl, acf, and cross-correlation with the neighboring channel follow.

3.2. Auditory image correlation (AIC)

The auditory image correlation (AIC) [7] is a tonality measure based on the so called "auditory image model" (AIM) [13,14]. AIM simulates the auditory processing of everyday sounds in the form of an "auditory image" (AI) that is intended to represent our initial impression of the sound. At higher abstraction level, the AI is constructed in three steps: a filter bank simulates the basilar membrane motion (BMM). Then, a multichannel neural activity pattern (NAP) is computed like that observed in the auditory nerve. Further, a bank of strobed temporal integration (STI) units, one per channel, converts this neural activity pattern into a dynamic AI [14].

The AIC uses intra- and inter-channel correlations of the multi-channel activity patterns to analyze envelope fluctuations. As shown in Figure 3, AIC uses the filters of the PM instead of the gammatone filters in the original AIM [13, 14]; the real parts of the filter outputs are taken as the BMM. The conversion to the NAP is done using "half-wave rectification," logarithmic "compression," and "lowpass" filtering (hcl). Half-wave rectification simulates the unipolar response of the hair cells. The progressive loss of phase locking with increasing frequency was implemented using a second order leaky integrator with a cutoff of 1200 Hz [15].

Since AIC is designed only to estimate tonality, it was sufficient to model STI with an autocorrelation function (acf): strong periodicity leads to maxima in the output of the acf at positions corresponding to the period [7, 13]. However, in this setup, dirac-shaped acf of white noise signals are smeared due to the shaped (filtered) white noise, which leads to errors in the estimation. To overcome this, an inter-channel measure was added using the cross-correlation coefficient of the

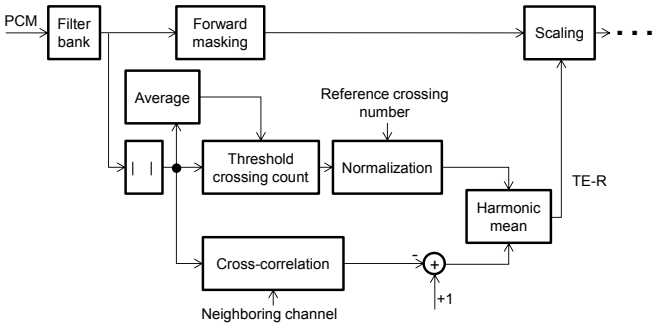

Fig. 4. Block diagram of the TE-R. The measure is calculated from the harmonic mean of the TETC and the TECC.

individual channels' autocorrelation of the neighboring channel [16], as shown in Figure 3. Further, the measure was normalized with respect to the white noise to have a tonality mapping similar to the AM-R (0 for tonality, 1 for noisiness). The AIC is defined as:

$$\text{AIC} = 1 - \frac{\sum_{k=1}^{N} acf_{c,k} \times acf_{c-1,k}}{\sqrt{\sigma_c^2 \times \sigma_{c-1}^2}} \tag{2}$$

where $acf_{c,k}$ is the acf for channel c and sample index k and σ_c^2 is the variance of the acf of channel c.

3.3. Temporal envelope rate (TE-R)

Temporal envelope rate (TE-R) is introduced as another approach to analyze the envelope fluctuations with lower computational complexity [7]. As depicted in Figure 4, TE-R is based on an intra- and an inter-channel correlation analysis. In contrast to AIC, TE-R avoids the computation of entire acfs with very long buffer lengths.

An intra-channel temporal envelope measure was designed based on the idea of using threshold crossings to determine noisiness on the temporal envelope: for every channel, the short-time average of the magnitude filter output (temporal envelope) is taken as a threshold value. Fluctuations around this threshold are counted. The number of crossings is then normalized to a reference crossing value, which is obtained by averaging crossing numbers for white noise input. An inter-channel temporal envelope measure analyzes the correlations between different channels of the temporal envelopes of the filter outputs. The number of channels for the estimation of the inter-channel measure is generally a design parameter. The immediate lower neighboring channel was considered for the current implementation. As shown in Figure 4 the harmonic mean of the intra- and inter-channel measures generates the TE-R.

Similar to the AM-R and the AIC, the TE-R gives a tonality/noisiness value which lies between 0 and 1, indicating strong tonality or strong noisiness, respectively.

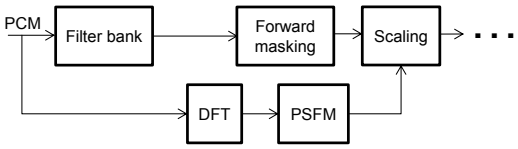

Fig. 5. Block diagram of the PSFM.

3.4. Partial Spectral Flatness Measure (PSFM)

Johnston [17, 18] applied a spectral flatness measure (SFM) to perceptual audio coding. Based on the idea of perceptual entropy, the model deployed SFM as a distinction measure between tone-like and noise-like maskers [17,18]. The model used a short-time power spectrum with fixed analysis frame length for Bark-wide bands. Similarly, Taghipour et al. [9] introduced the tonality measure "partial spectral flatness measure" (PSFM) as the ratio of the geometric and the arithmetic means of short-time squared magnitude spectrum, $|S_{st}(k)|^2$, of the input signal. Short-time spectra of different spectral resolutions are generated by discrete Fourier transforms (DFT) of different lengths of 4096, 2048, or 1024 for low, middle and high frequencies, respectively [9]. The PSFM is calculated corresponding to the individual band-pass filters as:

$$\text{PSFM} = \frac{\sqrt[N2-N1+1]{\prod_{k=N1}^{N2} |S_{st}(k)|^2}}{\frac{1}{N2-N1+1}\sum_{k=N1}^{N2} |S_{st}(k)|^2} \tag{3}$$

where $0 \leq \text{PSFM} \leq 1$. $N1$ and $N2$ were chosen in a way that for each filter output the range extended to the double of its efficient bandwidth [9]. The block diagram of the PSFM is shown in Figure 5. Only three FFTs[1] are calculated for one set of tonality values. In this paper, the PSFM is used as a reference for the comparison.

4. THE EXPERIMENTAL CODING SETUP

The tonality estimation methods were implemented in the PM. For the subjective quality test, an MDCT[2]-based experimental coding setup was chosen. A fixed input frame length of 1024 samples was used, which is equivalent to an MDCT window length of 2048 samples. Different variants of the PM were applied for quantizer control, each of which included one of the tonality estimation methods. Although an entropy coding was not applied, entropy rates were estimated. The different versions were controlled to have a desired average data rate, estimated for a large set of standard test audio signals with varying characteristics. This was done by applying a scaling factor to the estimated MTs. An equal average entropy rate of 48 kbit/s was chosen.

[1] Fast Fourier transform
[2] Modified discrete cosine transform

5. SUBJECTIVE QUALITY TEST

The four tonality estimation methods were compared to each other by means of MUSHRA (multiple stimuli with hidden reference and anchor) test [19]. Six items were chosen, which possess various characteristics (all mono with 48 kHz sampling frequency). Table 1 shows a list of used items with the corresponded estimated entropy in each condition. The subjects performed the MUSHRA test using a Graphical User Interface (GUI) developed by Fraunhofer-IIS [9, 10].

For each test item, subjects rated the quality of each condition in comparison to the reference/original. The conditions were: a hidden reference, a lowpass filtered anchor (f_c = 3.5 kHz), and four coded versions with AIC, TE-R, AM-R, and PSFM. Subjects were asked to rate the hidden reference at 100, as far as they could detect this. The order of appearance of the audio files (items) and the order of the codecs (conditions) were randomized, as described in [19].

16 normal-hearing, well-trained subjects participated in the test. However, inspection of the results showed that two subjects should be discarded. For several items, they had not detected the hidden reference and/or the lower anchor. Additionally, their ratings for the coded conditions showed an extremely low variability. Hence, the final statistical analysis was based on the results of 14 subjects: 13 male and 1 female, aged between 21 and 39 (mean 30, median 30).

6. RESULTS

The results of the MUSHRA test are depicted in Figure 6. For each item and condition, the mean rating across subjects and the corresponding 95% confidence interval are shown. The scores for the hidden reference and the low anchor are represented by orange and magenta, respectively. Red, green, blue, and black represent AIC, TE-R, the AM-R and the PSFM, respectively. To analyze the results, a two-way repeated-measures ANOVA was carried out with factors "item" and "condition." A significant effect of item was found $[F(5, 65) = 10.0, p < 0.001]$. The effect of condition was also significant $[F(5, 65) = 148.6, p < 0.001]$. There was a significant interaction between item and condition $[F(25, 325) = 8.4, p < 0.001]$.

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

	AIC	TE-R	AM-R	PSFM
Female vocal - es01	58.40	59.39	61.68	57.06
Eng. m. speech - es04	61.44	56.25	54.36	49.97
Harpsichord - si01	29.87	30.40	34.31	38.13
Castanet - si02	72.17	90.73	79.86	73.32
S. orchestra - sc02	39.29	47.82	49.28	45.39
Pitch-pipe	26.87	24.33	29.24	31.04
Average	48.01	51.49	51.46	49.15

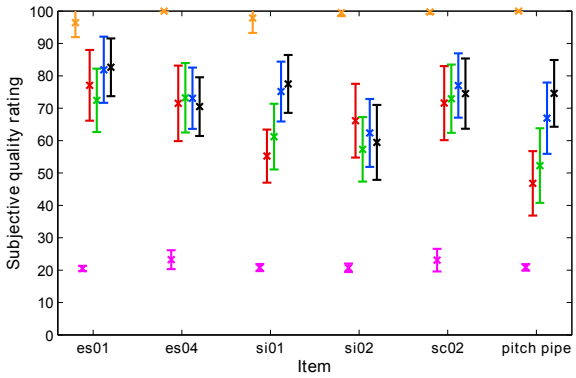

Fig. 6. Results: mean ratings and confidence intervals (95%) across 14 subjects [7]. The abscissa shows the items. Results for the hidden reference are shown by orange, the lower anchor by magenta, the AIC by red, the TE-R by green, the AM-R by blue, and the PSFM by black. For each item, the subjective quality ratings for different conditions are shown. The ordinate shows a quality scale between 0 (very bad) and 100 (excellent).

It can be observed in Figure 6 that the hidden reference was rated significantly higher than all the other conditions; similarly, the low anchor scored significantly lowest (all pairs $p < 0.001$). Since the focus was to compare the four coded versions to each other, an additional two-way repeated-measures ANOVA was carried out with the same factors, but only on data for the four coded conditions. A significant effect of item was found $[F(5, 65) = 10.0, p < 0.001]$. A significant effect of condition was found $[F(3, 39) = 31.1, p < 0.001]$. There was a significant interaction between item and condition $[F(15, 195) = 8.6, p < 0.001]$.

Post hoc pairwise comparisons (Fisher's least significant difference, LSD) between the four coded conditions showed that AM-R and the PSFM were rated significantly higher than AIC and TE-R (all pairs $p < 0.001$). There was no significant difference between AIC and TE-R ($p = 0.820$). Similarly, AM-R and PSFM were not rated significantly different ($p = 0.666$).

Due to the interaction between item and condition, it seemed reasonable to analyze possible differences between the conditions, item by item. A separate one-way repeated-measures ANOVA was conducted for each item, with 4 coded conditions.

For Femal vocal, there was a significant effect of condition $[F(3, 39) = 4.0, p < 0.05]$. Post hoc comparisons showed that PSFM was rated significantly higher than AIC ($p < 0.05$) and TE-R ($p < 0.01$), even though this method showed the lowest estimated entropy for this item among the 4 conditions; see Table 1. No further significant differences were found. For English male speech $[F(3, 39) = 0.5, p > 0.5]$ and Symphony orchestra $[F(3, 39) = 2.0, p > 0.1]$, no significant effect of condition was found. For Harpsi-

chord, there was a significant effect of condition $[F(3, 39) = 28.9, p < 0.001]$. Post hoc comparisons showed that both the AM-R and the PSFM were rated significantly higher than both AIC and TE-R (all pairs $p < 0.01$). TE-R was rated significantly higher than AIC ($p < 0.05$). No significant difference was found between the two best conditions. For Castanets, there was a significant effect of condition $[F(3, 39) = 3.1, p < 0.05]$ when sphericity was assumed. However, use of the Greenhouse-Geisser correction led to no significance ($p = 0.072$). For Pitch-pipe, there was a significant effect of condition $[F(3, 39) = 25.6, p < 0.001]$. LSD tests showed that the PSFM was rated significantly higher than AIC, TE-R, and AM-R (all pairs $p < 0.01$). It can be observed in Table 1 that PSFM allocates more (estimated) bits to this item than other methods. AM-R was rated higher than AIC and TE-R (all pairs $p < 0.01$). No significant difference was found between the AIC and TE-R.

7. CONCLUSION

Three envelope analysis methods were introduced for tonality estimation. They were implemented in a filter bank-based PM and were compared to each other and to an existing method. Overall, the AM-R and the PSFM scored significantly higher than AIC and TE-R, but were comparable to each other. The analysis of the results of the MUSHRA tests, presented here and in [9, 10], indicate that both the PSFM and the AM-R are appropriate, valid methods for tonality estimation.

The AIC and the TE-R might be improved by further optimizations [7]. The varying analysis lengths for different channels in the PSFM and the AM-R should also be incorporated for the AIC and the TE-R. Additionally, since the computational complexity of the TE-R is much lower compared to the AIC, it is recommended for future improvements.

REFERENCES

[1] R. P. Hellman, "Asymmetry of masking between noise and tone," *Perc. Psychoph.*, vol. **11**, pp. 241–246, 1972.

[2] J. L. Hall, "Asymmetry of masking revisited: Generalization of masker and probe bandwidth," *J. Acoust. Soc. Am.*, vol. **101**, pp. 1023–1033, 1997.

[3] J. Verhey, "Modeling the influence of inherent envelope fluctuations in simultaneous masking experiments," *J. Acoust. Soc. Am.*, vol. **111**, pp. 1018–1025, 2002.

[4] H. Gockel, B. C. J. Moore, and R. D. Patterson, "Asymmetry of masking between complex tones and noise: the role of temporal structure and peripheral compression," *J. Acoust. Soc. Am.*, vol. **111**, pp. 2759–2770, 2002.

[5] A. Taghipour, B. C. J. Moore, and B. Edler, "Masked threshold for noise bands masked by narrower bands of noise: effects of masker bandwidth and center frequency," *J. Acoust. Soc. Am.*, (under review).

[6] A. Taghipour, B. C. J. Moore, and B. Edler, "Durations required to distinguish noise and tone: effects of noise bandwidth and frequency," *J. Acoust. Soc. Am.*, (under review).

[7] A. Taghipour, "Psychoacoustics of detection of tonality and asymmetry of masking: implementation of tonality estimation methods in a psychoacoustic model for perceptual audio coding," Doctoral Thesis. Friedrich-Alexander-University of Erlangen-Nürnberg, Germany, 2016.

[8] A. Taghipour, N. Knölke, B. Edler, and J. Ostermann, "Combination of different perceptual models with different audio transform coding schemes: implementation and evaluation," *129th AES Convention, San Francisco, USA*, 2010, paper number 8283.

[9] A. Taghipour, M. C. Jaikumar, and B. Edler, "A psychoacoustic model with partial spectral flatness measure for tonality estimation," *22nd Eur. Signal Process. Conf. (EUSIPCO), Lisbon, Portugal*, 2014, pp. 646-650.

[10] H. Chen, A. Taghipour, and B. Edler, "Comparison of two tonality estimation methods used in a psychoacoustic model," *4th IEEE Int. Conf. Audio Lang. Image Process. (ICALIP), Shanghai, China*, 2014, pp. 706-710.

[11] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*, Springer, Berlin, Germany, 3rd edition, 2007.

[12] C. E. Bos and E. de Boer, "Masking and discrimination," *J. Acoust. Soc. Am.*, vol. **39**, pp. 708–715, 1966.

[13] R. D. Patterson, "Auditory images: how complex sounds are represented in the auditory system," *J. Acoust. Soc. Jpn.*, vol. **21**, pp. 183–190, 2000.

[14] R. D. Patterson and J. Holdsworth, "A functional model of neural activity patterns and auditory images," *Adv. speech hear. lang. process.*, vol. **3**, pp. 547–563, 1996.

[15] I. T. Bleeck, S. and R. D. Patterson, "Aim-mat: the auditory image model in matlab," *Acta Acust. united Acust.*, vol. **90**, pp. 781–787, 2004.

[16] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neur. Net.*, vol. **15**, pp. 1135–1150, 2004.

[17] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Sel. Areas Commun.*, vol. **6**, pp. 314–323, 1988.

[18] J. D. Johnston, "Estimation of perceptual entropy using noise masking criteria," *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 1988, pp. 2524-2527.

[19] "*Method for the subjective assessment of intermediate quality level of coding systems, Recommendation ITU-R BS.1534-1*," International Telecommunication Union, Geneva, Switzerland, 2003.