Application of a spectrally integrating auditory filterbank model to audio coding

Steven van de Par, Armin Kohlrausch, Philips Research Laboratories, Prof. Holstlaan 4, 5656 AA Eindhoven, E-mail:Steven.van.de.Par@philips.com

Introduction

Lossy audio compression depend on auditory masking. Such methods encode an audio signal in a very efficient, but inaccurate manner, thereby introducing considerable amounts of distortion within the audio signal. This is allowed as long as this distortion is masked by the audio signal itself. To control this process, models of auditory masking have been used, cf. [1, 2]. These models are essentially based on the assumption that the detectability of distortions is governed by the auditory filter that receives the largest distortion-to-masker ratio. Improvements have been made to this assumption by making a distinction between tonal and noisy maskers to exploit the knowledge that noises are better maskers than tonal components [1].

In this study we want to investigate the implications of a number of studies that have shown that the detectability of distortion is mediated not only by the best auditory filter, but by a process where information is integrated over a range of auditory filters. Through this integration, detectability is improved compared to the detectability that would be obtained based on the best auditory filter only (i.e. the auditory filter with the largest distortion-to-noise ratio). Buus et al. [3] compared masked thresholds for individual tones and for a complex of 18 tones, separated by approximately one critical bandwidth, that were masked by a uniformly masking noise. Results showed that the thresholds for the tone-complex were about 6 dB lower than for the individual tones expressed in dB-SPL per component. Such a result is in line with the multi-band energy detector model [4] which assumes that decision variables associated with the individual auditory filters are added and which assumes that the internal noises within each auditory filter are indepedent. A later study [5] confirmed the finding that across-frequency integration occurs for both monaural and binaural masking experiments.

Of course these findings have implications for audio coding algorithms because they imply that spectrally separated distortion components that were inaudible individually in the presence of the masker can become audible when presented simultaneously. In this paper we will discuss the implications of spectral integration for perceptual audio coding in more detail. It will be shown that this leads to different solutions for the spectral shape of the distortion signal than by assuming that the best filter determines the detectability of the distortion. To conclude, a listening test was conducted to investigate whether this approach leads to a better quality for the same bit rate.

Theoretical implications of spectral integration

Most perceptual audio compression algorithms achieve their bit-rate reduction by a signal transformation which decomposes the signal into a representation where each time-frequency interval of the input signal is represented by a single number. When these numbers are represented by a limited number of bits a reduction in bitrate can be achieved, but inevitably it will also lead to quantization noise. When the quantization noise level is low enough, it will be masked by the rest of the signal. Usually, a group of intervals that correspond to a single time interval and a range of consecutive frequency intervals are quantized with the same accuracy to avoid excessive amounts of overhead information needed by the decoder for resynthesizing the signal. These groups are termed scale-factor bands and their width is assumed here to be proportional to the auditory filter bandwidth.

The classical approach to determine quantization-noise levels is by using a masking model. The quantization noise in each band is adjusted in such a way that each band is below the threshold of detectability. It is assumed then that there is no additivity of the detectability which is justified under the assumption that each scale-factor band affects only the output of one auditory filter, and under the assumption that the detectability is determined only by the auditory filter that has the best detectability. In line with these assumptions the goal would be to adjust the sensitivity index (d') of each band to make it equal or smaller than unity [4].

Several studies have shown, however, that targets separated by at least one critical bandwidth and each being masked individually, can be audible when presented at the same time. In the study by Buus et al. [3], it was shown that the complex of 18 tones had a masked threshold (dB/component) that was 6 dB lower than that for the individual tones of the complex. This indicates that spectral integration of information contained in separate auditory filters can be integrated by the human auditory system. Buus et al. [3] proposed an Euclidian addition rule for individual tonal d's to obtain the resulting d' for the tone complex:

$$d'^2_{\mathrm{tot}} = \sum_{n=1}^{18} d'^2_n, \qquad (1)$$

where d'_n is the sensitivity index associated with the $n-$th tone of the complex and d'_{tot} the sensitivity index of the tone complex. This would account for a 6.3 dB improvement in threshold for the 18-tone complex.

The above mentioned spectral integration has implications for the quantization noise that can be allowed in audio coders. The additivity of d's leads to an improved detectability of quantization noise which implies that a more conservative level of the masking threshold is required. This, however, is not the only implication.

Let us first consider the situation that d's for each individual scale-factor band are adjusted to the same level. In other words, the quantization noise is equally detectable within each frequency interval of the audio signal. Now consider that scale-factor bands are broader at high frequencies. This implies that more data are needed to represent the information within this scale-factor band. It can reasonably be assumed that when the d' is slightly increased in this scale-factor band, (i.e., the quantization-noise level is increased) a relatively large amount of bits is saved. When these bits are then used to reduce d's at low frequencies, a larger reduction in these d's is expected.

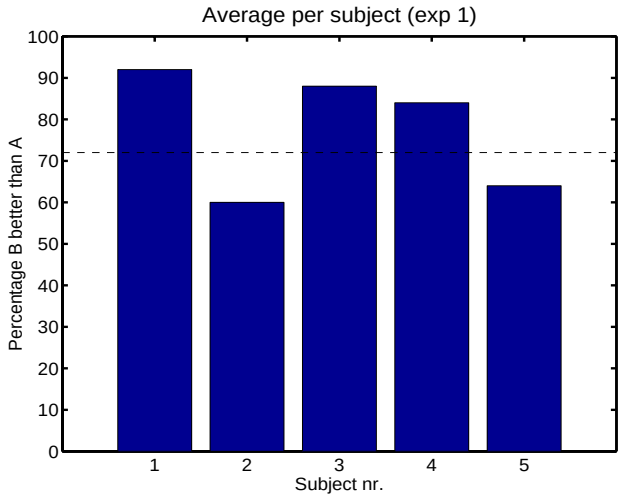

Figure 1: Percentages of preferences for optimized d's over uniformly distributed d's for five subjects. The results are averages over the five excerpts. The dashed line shows the 5% significance level.

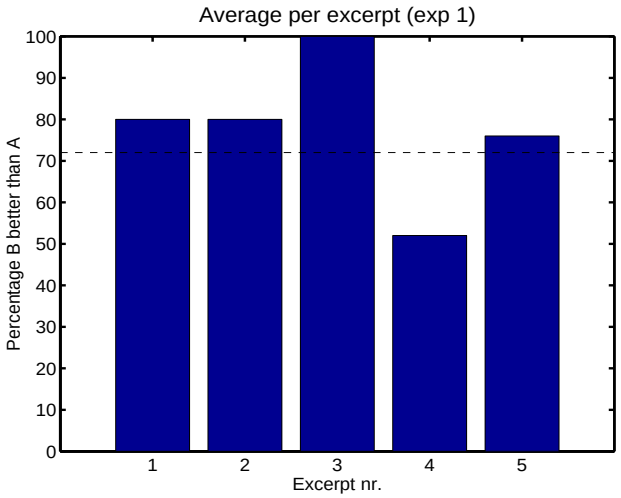

Figure 2: Same as Fig. 1, except now that averages over subjects are shown for all five excerpts.

The result will be that, using the same amount of bits, the total d' is reduced implying a more efficient coding of audio.

Thus, one can conclude that besides the masking properties of the input signal, also the differential bit-costs that are associated with the separate scale-factor bands need to be taken into account to find the most efficient levels for the quantization noise. It is important that the masking is modelled in a way that incorporates spectral integration because only such a model allows a trading of d's as is proposed here. Such a model is proposed by Van de Par et al. It measures distortion-to-masker ratios at the outputs of the filters of an auditory filterbank and integrates these distortion to masker ratios to yield a measure of the perceived distortion [6].

Experiments

To investigate whether the newly proposed quantization strategy indeed leads to an improvement in quality, a short listening test was conducted. A simple transform coder was used for this purpose together with the perceptual distortion measure proposed in Van de Par et al. [6].

In a listening test a 3-Interval 2-Alternative Forced choice procedure was used. In the first interval the original excerpt was presented, in the second and the third interval the method with uniform d's or optimized d's was presented in a random order. Subjects had to indicate their preference for either the second or the third interval. Five different excerpts were used, each of which was presented in five repeated trials.

The average preference results (Fig. 1) show that 3 out of 5 subjects prefer the newly proposed method that optimized d' considering bit costs (as indicated by the bars that exceed the statistical significance level depicted by the dashed line). It also shows that the two remaining subjects had no preference, in other words the optimization of d's did not lead to a reduced quality as would be expected based on a model that does not include spectral integration.

Similarly, the average results per excerpt (Fig. 2) show that only for one excerpt there was no significant preference for either one of the methods while for the other excerpts the method that optimized d's was perfered.

Conclusion

A new method is proposed to determine quantization levels for perceptual audio coding. This method takes into account the masking properties of the input signal, which serves as the masker, but also takes into account the bit-costs that are associated with achieving the quantization levels as such. This allowed for a trading of quantization noise detectability between bands that are expensive to encode and bands that are not so expensive to encode. This trading is possible because spectral integration occurs in auditory masking. In this way a more efficient way of encoding audio signals can be obtained that would not be possible on the basis of masking alone. This theoretical prediction was supported by a listening test.

1. ISO/MPEG Committee. *Coding of moving pictures and associated audio for digital storage media at up to about 1.5Mbit/s - part 3: Audio*, 1993. ISO/IEC 11172-3.
2. T. Painter and A. Spanias. Perceptual coding of digital audio. *Proceedings of the IEEE*, 88:451–513, 2000.
3. S. Buus E. Schorer M. Florentine and E. Zwicker. Decision rules in detection of simple and complex tones. *J. Acoust. Soc. Am.*, 80:1646–1657, 1986.
4. D. M. Green and J. A. Swets. *Signal detection theory and psychophysics.* John Wiley & Sons, New York, London, Sydney, 1966.
5. A. Langhans and A. Kohlrausch. Spectral integration of broadband signals in diotic and dichotic masking experiments. *J. Acoust. Soc. Am.*, 91:317–326, 1992.
6. S. van de Par A. Kohlrausch G. Charestan and R. Heusdens. A new psychoacoustical masking model for audio coding applications. In *IEEE int. Conf. Acoust., Speech and Signal Process.*, Orlando, USA, 2002. In press.