An Excitation Level Based Psychoacoustic Model for Audio Compression

Ye Wang
Nokia Research Center
Speech and Audio Systems Lab
Tampere, Finland
Tel.: +358 3 272 5609
E-mail: ye.wang@nokia.com

Miikka Vilermo
Nokia Research Center
Speech and Audio Systems Lab
Tampere, Finland
Tel.: +358 3 272 5826
E-mail: miikka.vilermo@nokia.com

ABSTRACT

This paper describes an excitation level based psychoacoustic model to estimate the simultaneous masking threshold for audio coding. The system has the following stages: 1) a windowing function; 2) a time-to-frequency transformation; 3) an excitation level calculation block similar to that in Moore and Glasberg's loudness model; 4) a correction factor for estimating masking threshold; 5) the inclusion of the absolute masking threshold; 6) the output Signal-to-Masking ratio. We have evaluated the performance by integrating the proposed psychoacoustic model into an audio coder similar to MPEG-2 AAC, which contains only the basic coding tools. Our model performs better than or as well as the psychoacoustic model suggested in the MPEG-2 AAC audio coding standard for all the test signals. We can achieve almost transparent quality with bitrate below 64 kbps for most of the critical test signals. Significant improvements have been achieved with speech signals, which are always difficult for transform audio coders.

Keywords

Psychoacoustic model, excitation level, masking threshold, audio compression.

1. INTRODUCTION

Combining psychoacoustic models into audio coders significantly improves the coding efficiency. However, the psychoacoustic models used so far in perceptual coders are based on very simplified assumptions, which may result in much less accurate approximations of masking thresholds. For example, the psychoacoustic models suggested in the audio parts of MPEG-1 and MPEG-2 use a DFT of successive blocks of the audio signal, which gives the associated spectral components of the blocks. For each spectral component an individual masking threshold is generated. The overall masking threshold follows from superposition of the individual thresholds, which is carried out by simply adding up the threshold at the corresponding frequencies [1]. This masking threshold determines the maximum quantization noise energy that can be added to the original signal to keep the noise inaudible. These models are quite approximate, when a complex target (quantization noise) has to be masked by a complex masker comprising multiple spectral components (either speech or musical sounds) [11]. Further bit rate reduction heavily depends on the accurate estimation of the masking threshold both in the time and frequency domains.

To simulate the human ear better, some ear models have been developed [4][5][6][10]. Our model is based on Moore and Glasberg's excitation level calculation. This is quite different from psychoacoustic models commonly used, and it leads to some advantages in masking threshold estimation.

2. MODEL DESCRIPTION

Figure 1 shows the block diagram of our method. The following steps are performed:

A windowing function is first applied to the input audio signal. We apply the same window function as specified in MPEG-2 AAC. Depending on the signal, the model changes the time/frequency resolution by using two different windows: LONG_WINDOW = 2048 and SHORT_WINDOW = 256. We have applied two different transition windows LONG_START_WINDOW and LONG_STOP_WINDOW in case of switching between long and short windows. The transition windows have not been used in the psychoacoustic model suggested in MPEG-AAC. Using the exactly same window switching in both the psychoacoustic model and in the MDCT (Modified Discrete Cosine Transform) helps to reduce some coding artifacts.

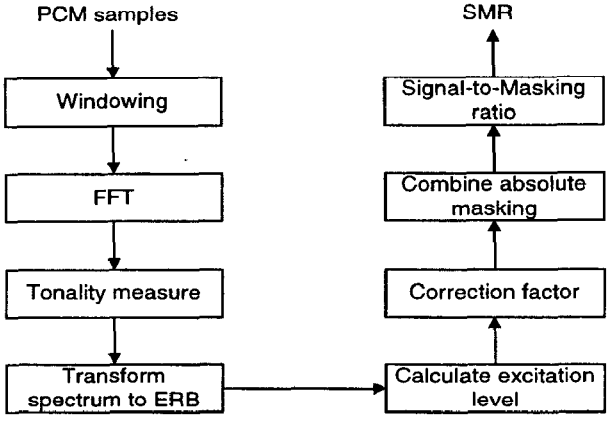

Figure 1. Block diagram of sequence of stages in the model

The reason for window switching is that Moore's loudness model is designed for steady sounds. It can not cope with transient signals well, and at the moment, we solve this problem by introducing window switching.

The FFT has been chosen for the time-to-frequency transformation. The transform block length is 32768 ($=2^{15}$) for practical reasons: Moore and Glasberg's model uses the equivalent rectangular bandwidths (ERBs), which are similar to traditional critical bands at low frequencies (see below). To ensure that each ERB has at least one frequency line, the FFT block length has to be increased by padding with zeros after the actual data, which are 2048 points for the long window and 256 for the short window. This increases the number of frequency lines, while preserving the shape of the spectrum.

Because tonal and non-tonal components have very different masking properties, we introduce the tonality measure as a weighting function of the frequency components. Currently we use unpredictability as a tonality measure similar to the method specified in MPEG-2 AAC. However, our model predicts from both the past and the future two frames. We choose the one with less prediction error for calculating the unpredictability measure. This remarkably improves the coding efficiency for some signals.

A critical problem is how to integrate the tonality measure with the masking threshold. From psychoacoustical experiments, the masking threshold is about 18 dB below the masker excitation level for a tonal masker, but about 6 dB below for a narrow band noise masker. We have introduced this difference before excitation level calculation. The weighting function is described by

$$\text{Spectrum\_weighted} = 10^{-((12(1-CW))/20}\text{Spectrum}, \quad (1)$$

where CW is the unpredictability measure. The weighting function requires further optimization.

At moderate sound levels, the ERB width is described by

$$\text{ERB} = 24.7(4.3F+1), \quad (2)$$

where the ERB is in hertz and the center frequency F is in kilohertz. This function is similar to the "traditional" critical bandwidth (CB) function at medium to high frequencies, but gives markedly lower values than the CB function at center frequencies below 500 Hz. [5]

The next step is to transform from the frequency domain to the ERB scale, which is described by

$$\text{Number of ERBs} = 21.4 \log10(4.37F+1), \quad (3)$$

where the frequency is in kilohertz [5].

In our model we have not used the outer and middle ear transfer function, because the final masking threshold for coding must be transformed back to free field sound pressure level. We assume that the forward and backward transfer function of outer and middle ear cancel each other.

The excitation pattern for a given spectrum is calculated being the pattern of outputs from the auditory filters. Each auditory filter is assumed to be quasi-linear at a given level, but to change shape with frequency and with level in a way similar to that described by Moore and Glasberg [8].

It is assumed that the masking pattern should be parallel to the excitation pattern of the masker, but shifted vertically downwards by a small amount [9], we have introduced the *CORRECTION FACTOR* to represent that shift and tried to find out the optimal correction factor experimentally. For all test materials used, 6 dB is a suitable correction factor. We have also modified the correction factor below 500 Hz according to [5]. The influence on bitrate versus audio quality seems to be minimal.

Because Moore's model does not cover the whole audible frequency range up to 20 kHz, we combine the calculated masking threshold with the absolute masking threshold as the global masking threshold. Choosing the higher of the two thresholds approximates this combination. Finally we output the Signal-to-Masking ratio (SMR) for each scalefactor band.

3. EXPERIMENTAL RESULTS AND DISCUSSION

The psychoacoustic model is built into a codec similar to MPEG-2 AAC. The test materials are provided by MPEG and commonly used in audio coding evaluation. These include English and German speech spoken by male and female, female singing in English without instrumental accompaniment, harpsichord, castanets, pitchpipe, bagpipe, glockenspiel, plucked strings, trumpet concerto, symphony orchestra and contemporary pop music. Our model performs better than the MPEG-2 AAC psychoacoustic model for all signals. To achieve the same audio quality, we can save 10-20% bits.

Moore and Glasberg's loudness model is intended for stationary signals, but we have used it for real audio signals, which sometimes have strong transients. The transients should be tackled by using e.g., window switch, more accurate detection of transients, better exploiting temporal masking, short window grouping etc. Preliminary tests show that an additional 10% reduction in bit rate can be achieved through combining simultaneous masking and forward masking. It should be noted that we have not used any prediction for our test. Backward adaptive prediction would improve coding efficiency for some signals, such as the pitchpipe and bagpipe. However, it does not help very much for other test signals.

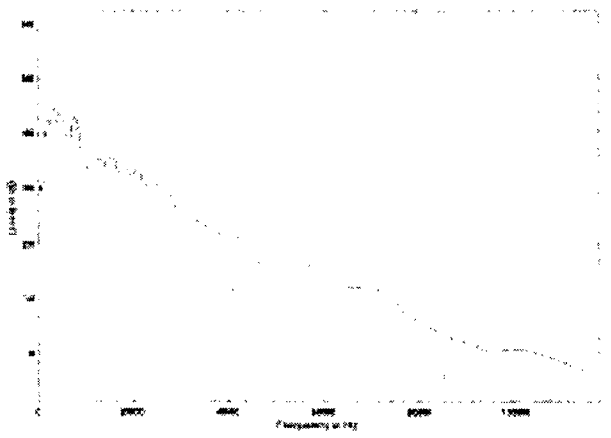

Figure 2 Masking threshold with (solid) and without (dotted) the tonality measure calculation for a noise-like signal

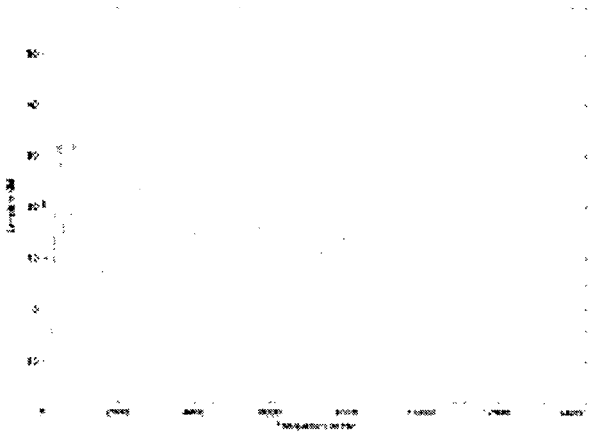

Figure 3 Masking threshold with (solid) and without (dotted) the tonality measure calculation for a signal with significant sinusoidal components

Figure 2 and 3 show the effect of the tonality measure. Figure 2 shows the spectrum of a piece of symphony orchestra signal and its masking thresholds calculated with and without the tonality measure. Figure 3 shows the spectrum of a section of pitchpipe signal and its masking thresholds calculated with and without the tonality measure. The symphony orchestra signal is more noise-like and the difference between the two masking thresholds is rather small. The pitchpipe contains rich sinusoidal harmonics and the difference between the two masking thresholds is more significant.

Figure 4 shows the spectrum of a section of the symphony orchestra signal and its masking thresholds from the MPEG2-AAC model (dotted) and our model (solid). Our model shows a different distribution of the allowed quantisation noise compared to the MPEG2-AAC model.

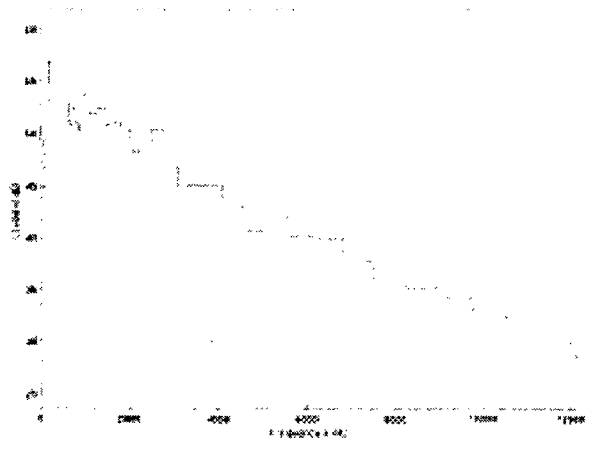

Figure 4 Masking threshold from the MPEG2-AAC model (dotted) and our model (solid)

4. CONCLUSION AND FUTURE WORK

The proposed psychoacoustic model can predict the masking threshold quite well for most test signals. Particularly, the performance with speech signals makes it very promising for a future hybrid speech and audio coders. Based on experimental codes in MATLAB, we have implemented our model in C language with some optimization for real-time applications.

What could be done in the future is:

- To find some other tonality measure which is more reliable than the unpredictability measure;
- In order to tackle transient signals better, the window switch mechanism has to be improved;
- In order to squeeze the bitrate further, short window grouping can be tested.

5. ACKNOWLEDGMENT

The authors wish to thank their superior, Mr. Mauri Väänänen (Nokia Research Center) for supporting this research; Prof. Brian C. J. Moore (University of Cambridge) for his careful reading of the manuscript and critical comments on improving the technical content and presentation of this paper, especially regarding the outer and middle ear transfer function; Dr. Jilei Tian (Nokia Research Center) for very inspiring discussion.

6. REFERENCES

[1] ISO/IEC JTC1/SC29/WG11, "Coding of moving pictures and audio- MPEG-2 Advanced Audio Coding AAC", ISO/IEC 13818-7 International Standard, 1997.

[2] N. S. Jayant, P. Noll, "Digital Coding of Waveforms", Prentice-Hall, Englewood Cliffs, NJ0732, U.S.A, 1984.

[3] E. Zwicker, H. Fastl: "Psychoacoustics, Facts and Models", Springer-Verlag, Berlin Heidelberg, Germany, 1990.

[4] J. G. Beerends, J. A. Stemerdink, "A Perceptual Audio Quality Measure Based on a Psychoacoustic Sound Representation", J. Audio Eng. Soc., Vol. 40, No. 12, 1992.

[5] B. C. J. Moore, B. R. Glasberg, T. Baer, "A Model for the Prediction of Thresholds, Loudness, and Partial Loudness", J. Audio Eng. Soc., Vol. 45, No. 4, 1997.

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

[7] B. C. J. Moore, "An Introduction to the Psychology of Hearing", 4. Edition, Academic Press, 1997.

[8] B. R. Glasberg and B. C. J. Moore, "Derivation of Auditory Filter Shapes from Notched-Noise Data", Hear. Res., Vol. 47, pp.103-138 (1990).

[9] B. C. J. Moore, "Masking in the Human Auditory System", Collected Papers On Digital Audio Bit-Rate Reduction, special publication of AES, 1996.

[10] C. Colomes, M. Lever, J. B. Rault, Y. F. Dehery, "A Perceptual Model Applied to Audio Bit-Rate Reduction", J. Audio Eng. Soc., Vol. 43, No. 4, 1995 April.

[11] B. Espinoza-Varas, S. V. Cherukuri, "Evaluating a model of auditory masking for applications in audio coding", proc. 1995 IEEE ASSP Workshop on Application of Signal Processing to Audio & Acoustics. New Paltz, New York.