EXPLOITING EXCESS MASKING FOR AUDIO COMPRESSION

YE WANG AND MIIKKA VILERMO

Nokia Research Center
Speech and Audio Systems Lab.
Tampere, Finland

ye.wang@research.nokia.com
miikka.vilermo@research.nokia.com

In order to improve audio coding performance, excess masking has been employed for the compression of complex audio signals. A new algorithm is developed to classify and pre-process maskers. A psychoacoustic model is used to estimate simultaneous masking threshold. This masking threshold is used for quantizing audio signal coefficients in the frequency domain. Preliminary test results show improved coding efficiency.

INTRODUCTION

Auditory masking plays a major role in audio coding, because all coding algorithms engender a certain level of undesirable low-level quantization noise that occurs simultaneously with the desired coded signal. All perceptual audio encoders have a psychoacoustic model, which calculates the masking threshold to determine the maximum allowable noise injection level without audible distortion. These models simulate masking effects from psychoacoustic studies. There is a major challenge however: Only simple stimuli such as sinusoids and bands of noise have been used in most psychoacoustical studies. In audio coding we are dealing with real life audio signals. That is, a multi-component complex masker (coded audio signal) must mask the spectrally complex target (quantization noise).

In our previous paper [1], we have applied an excitation-pattern model to estimate the simultaneous masking threshold for audio coding. This model performs fairly well for narrow-band-noise masking, but may overestimate the masking produced by tonal components [2][3]. We have introduced a weighting function, which includes the tonality measure to solve this problem [1].

On the other hand, the excitation-pattern model seems to underestimate the combined masking effects of multiple-component maskers [4][5]. More specifically, it underestimates the combined effects of two maskers both when the masker frequency components fall within the maskee auditory-filter bandwidth, and when they fall outside this bandwidth [5]. We hereby present some initial work that we have done to exploit the excess masking of two-tone maskers within the equivalent rectangular bandwidths (ERBs) [3] for audio compression.

Excess masking has been discussed in many publications since 1960's. In essence, the masking produced by the combination of simple maskers (sinusoids or bands of noise) is not a simple summation of the masking produced by the individual maskers. Several studies [6][7][8] have shown that the combined masking effect of two equally-effective simultaneous maskers is 3 to 15 dB greater than the masking predicted by the linear addition of masker energies. This "additional" amount of masking is defined as excess masking. Excess masking exists not only in frequency domain but also in time domain [9]. But the time domain excess masking will not be covered in this paper.

1. MODEL DESCRIPTION

The model includes the following stages: 1) time-to-frequency domain transformation, which is a FFT in our case; 2) masker classification and pre-processing, in which maskers are classified by their types and spectral structure; 3) masking threshold estimation, including excess masking as well as the absolute masking threshold as employed in the MPEG-2 AAC standard; 4) SMR (Signal-to-Masking Ratio) calculation, as the output of the model, used to control the quantizer in the audio encoder.

Because this is a modification of the model described in [1], the basic structure is essentially the same. Difference happens in stage 2 and 3. In stage 2, the algorithm searches for components that are subject to the following criteria: 1) The components must be local maxima; 2) They have to be tonal (predictable) i.e. the

unpredictability measure has to be under a certain threshold; 3) They have to be greater than 10.0 dB. Then the algorithm finds the first component that meets the criteria. Afterwards it searches for other components within one ERB that fulfill the criteria and that differ in amplitude less than 3 dB. If such components are found, then these components, together with the first one, are marked to cause excess masking (refer to the circles on top of some spectral lines in Figures 1, 2, 3). For every marked component, a 6-dB excess masking is introduced by modifying the weighting function described in [1].

The weighting function is introduced to integrate the tonality measure to the excitation-pattern model. 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. For tonal components with excess masking we have lifted the masking threshold by 6 dB. That is, the masking threshold is about 12 dB below the masker excitation level for a tonal masker with excess masking. We have introduced this difference before excitation level calculation. The weighting function is described by

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

if there is no excess masking for this component,

$$Spectrum\_weighted = 10^{(12(1-CW-0.5))/10} Spectrum \quad (2)$$

if excess masking occurs for this component,

where CW is the unpredictability measure. The weighting function requires further optimization. The weighting function differs a bit from [1], because the spectrum is an amplitude spectrum in that case, a power spectrum in this paper.

2. EXPERIMENTAL RESULTS

For preliminary test purposes, we have used a pitchpipe signal, which contains rich sinusoidal harmonics. Figure 1 shows its amplitude spectrum. Then we have produced a second signal from the previous one by shifting all frequency components upward one semitone. By mixing the above two signals together, we produce a signal, which has equal amplitude component pairs that are close in frequency (see Figure 2). In addition, we have created a major triad in root position with a similar approach using the same pitchpipe signal. These kinds of signals are supposed to produce quite obvious excess masking.

We have evaluated the performance by integrating the modified excitation-pattern model into an MPEG-2 AAC type audio encoder, which contains only the basic coding tools. We first code these mixed signals with the original masking curve calculated with the excitation-pattern model and then with the modified one (exploiting excess masking). Without degrading the subjective audio quality, the average bitrate can be reduced by 5% for both pitchpipe and bagpipe signals, 10% for both two-pitchpipe-mixed signal and the major triad in root position of pitchpipe signal. For many other audio signals such as speech, harpsichord, castanets, glockenspiel, plucked strings, trumpet concerto, symphony orchestra and contemporary pop music, this model seems to have little effect in bitrate and causes no audible degradation in sound quality. Tests were performed informally by the authors and two young colleagues in the same lab.

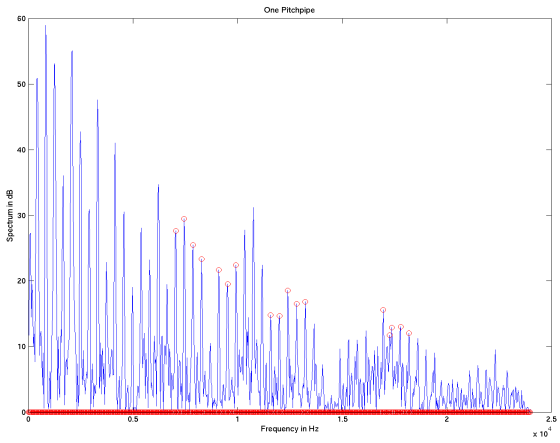

Figure 1. Spectrum of a piece of pitchpipe signal. Components that cause excess masking are marked with circles on top of them.

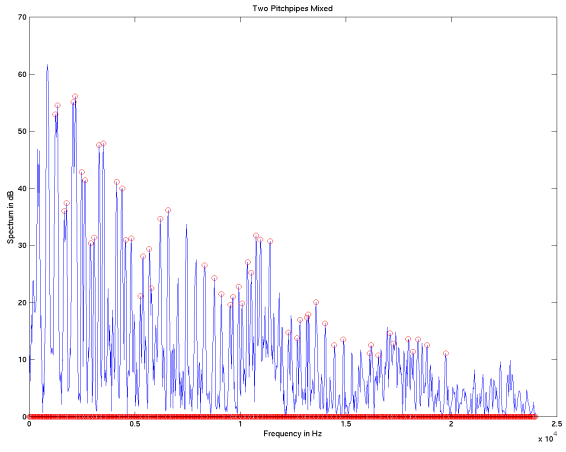

Figure 2. Spectrum of a piece of two-pitchpipe-mixed signal. Components that cause excess masking are marked with circles on top of them.

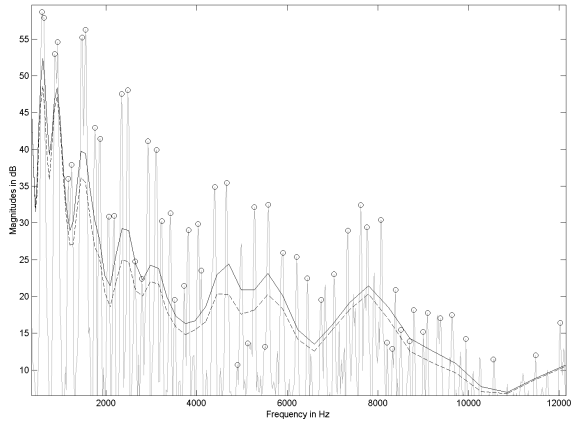

Figure 3. Spectrum of a piece of two-pitchpipe-mixed signal (dotted line), circles indicate components that cause excess masking, masking threshold without excess masking (dashed line), with excess masking (solid line). The masking thresholds are lifted parallel upwards for better visibility.

3. DISCUSSION

This work is only an initial work that utilises excess masking for audio coding applications. We have used only excess masking produced by pairs of sinusoids within one ERB. The algorithm that identifies components, which produce excess masking, is not optimised. It helps to reduce bitrate only for a few special audio signals such as pitchpipe, bagpipe and the mixed signals described earlier.

In addition to sinusoidal pairs, maskers can be two nearby narrow bands of noise, sinusoid combined with a narrow band of noise, etc. Excess masking of 8 dB was found for all masker configurations [7]. Even a pair of maskers outside the maskee auditory-filter bandwidth also produces some excess masking [5]. In addition, excess masking has been found in the time domain as well. That is, if the maskers are close enough in the time domain, the combined masking effect in the arithmetic center of the pair of maskers is not a linear combination of forward and backward masking [9]. In principle, all these excess masking phenomena can be utilised in audio coding. It is however very difficult to find a computationally efficient way to combine all excess masking into a practical audio encoder. It is worthwhile to point out that the maskees in almost all psychoacoustical studies [6][7][8] were sinusoids. To what extend these results can be utilised in audio coding is still an open question, since the maskee of an audio encoder is always the quantization noise, not sinusoids. Essentially what we are looking for is the optimal shaping of quantization noise according to the auditory masking.

In most of the publications, excess masking was measured at one particular point, most commonly in the middle of the pairs of maskers. How about the overall shape of excess masking (excess masking pattern) in the nearby frequency region or time span between the forward and backward maskers? This kind of overall shape would be much more useful in practical applications such as audio coding.

So far we have modified the weighting function to cope with the masking of both tonal components [1] and pairs of sinusoids. We have modified the amplitudes of these components before excitation-pattern model calculation. It is not obvious if this is the optimal way to solve these problems, since the excitation-pattern model is level dependent. In the case of reducing the amplitude of a tonal component, the corresponding auditory filter shape has been changed as well. More research is needed to answer these questions.

4. CONCLUSIONS

This preliminary test result proves that excess masking of sinusoidal pairs within one ERB can be exploited to compress at least some subclasses of audio signals more efficiently, especially for low bitrate applications. However, it is a challenging task to find technically feasible algorithms to include all excess masking into audio coding algorithms.

REFERENCES

[1] Wang Y., Vilermo M., “An Excitation Level Based Psychoacoustic Model for Audio Compression”, submitted to 1999 IEEE ASSP Workshop on Application of Signal Processing to Audio & Acoustics. New Paltz, New York.

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

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

[4] Van der Heijden M., Kohlrausch A., (1994) “Using an Excitation-pattern Model to Predict Auditory Masking”, Hearing Research 80, 38-52.

[5] Espinoza-Varas B., Cherukuri S. V., (1995)

"Evaluating a model of auditory masking for applications in audio coding", IEEE ASSP Workshop on Application of Signal Processing to Audio & Acoustics. New Paltz, New York.

[6] Green D. M. (1967). “Additivity of Masking”, J. Acoust. Soc. Am., 41, 1517-1525.

[7] Lutfi, R. A. (1983). “Additivity of simultaneous masking”, J. Acoust. Soc. Am., 73, 262-267

[8] Humes, L. E. and Jesteadt, W. (1989). “Models of the additivity of masking”, J. Acoust. Soc. Am., 85, 1285-1294.

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