Perceptual Wavelet Packet Audio Coder

Teddy Surya Gunawan, Eliathamby Ambikairajah, Julien Epps

School of Electrical Engineering and Telecommunications
The University of New South Wales, Australia
tsgunawan@ee.unsw.edu.au;ambi@ee.unsw.edu.au

National Information Communications Technology
Australia (NICTA)
 julien.epps@nicta.com.au

Abstract

Traditional wavelet packet audio compression algorithms do not utilize the temporal masking properties of the human auditory system, relying instead on simultaneous masking models. This paper presents the design and implementation of a perceptual wavelet audio coder by incorporating temporal and simultaneous masking models. The efficiency of the encoder was assessed based upon the number of bits required to code wavelet packet coefficients in each critical band, while retaining perceptual transparency. Subjective listening tests conforming to ITU-R BS.1116 revealed the bit rate is reduced by more than 17% compared to using a coder that only employs a simultaneous masking model.

1. Introduction

The discrete wavelet transform and wavelet packets are efficient multi-resolution decomposition tools in source coding and have been widely used in still image and video compression. Previous research work [1-3] has demonstrated the performance of wavelets for low bit rate coding of audio signals. Wavelet packet decomposition provides a closer approximation to the critical bands and normally acts as a front-end to the psychoacoustic model. The incorporation of auditory masking effects in the quantization process of spectrally transformed signals has led to important gains in terms of bit rate while preserving perceptually relevant signal components [3, 4].

Simultaneous masking models have been used in audio coding for bit rate reduction [2, 4]. Most recently, temporal masking has been incorporated in speech and audio coding [5, 6]. We have also demonstrated that the incorporation of temporal masking, along with simultaneous masking, into an audio coding algorithm, results in further bit rate reduction compared to using simultaneous masking alone, while retaining perceptual quality [7]. In this paper, we present an efficient implementation of a temporal masking model for a wavelet packet based audio coder. The temporal masking model, combined with the simultaneous masking model, is then used to calculate the combined masking threshold to determine the bit allocation.

To evaluate the performance of our perceptual wavelet packet audio coder, we carried out subjective tests according to the ITU-R BS.1116 using various audio materials sampled at 44.1 kHz. We trained the listener to become familiar with coding artifacts, such as pre-echo, aliasing, birdies, and speech reverberation as described in AES CD-ROM [8]. The ABC/Hidden Reference Software Audio Comparison Tool was then used to evaluate our coder performance for various audio materials, in which the listener grades the audio material.

2. Wavelet Packet Audio Coding

A block diagram of the 28-band wavelet packet (WP) decomposition used in this work is shown in Fig. 1. Eight-level decomposition was employed and the bands were grouped together in such a way as to approximate 28 critical bands. The wavelet packet decomposition was designed using 16-tap filters based on the Daubechies wavelet, providing an acceptable compromise between the sub-band separation and increased computational load [2]. Moreover, wavelet packet decomposition provides a convenient signal representation for calculating both masking thresholds in the time-frequency domain.

![Block Diagram of the Encoder](image-url)

**Figure 1:** Block Diagram of the Encoder

In a tree structured sub-band decomposition, the amount of coder delay accumulated by each sub-band depends on the order of the tree. Without compensation for this delay the reconstruction of the sub-band signal components will lead to an incoherent output signal. The total amount of delay compensation necessary for sub-band is calculated as,

\[ \Delta d(k) = (N - 1) \left( 2^{s_{\text{max}} - s(k)} - 1 \right) \]  

where \( N \) is the filter order and \( s_{\text{max}} \) is the filter order and is the maximum decomposition stage.
3. Auditory Masking Models

Principles of psychoacoustic masking play a key role in enabling coding of audio signals with minimum audible degradation. A temporal masking model previously developed by us [6] was modified and combined with a simultaneous masking model [2] in order to calculate the combined masking thresholds in the time-frequency domain.

3.1. Sound Pressure Level (SPL) Calculation

The SPL calculation is necessary as it is a standard metric that quantifies the intensity of an acoustical stimulus. Auditory psychophysical phenomena addressed in this paper are treated in terms of SPL. Therefore, it is necessary to relate the PCM input signal with SPL. The common method to normalize the PCM input signal is as follows [3, 4]:

\[ P_{SPL} = PN + 10 \log_{10} \left| P_{z_l} \right|^2 \]  

where \( PN \) is the power normalization term and fixed at 90 dB, \( \left| P_{z_l} \right|^2 \) is the critical band power for \( z_l = 1 \ldots 28 \).

3.2. Simultaneous Masking Model

Since the masking level is a function of signal power, the psychoacoustic model first computes the signal power within each sub-band by selecting the maximum power of the wavelet coefficients in that particular sub-band. The masking threshold for each band is calculated using the MPEG psychoacoustic model 1 outlined in [2, 4].

3.3. Temporal Masking Model

In this paper, we consider only forward masking model. Backward masking is not considered here, as their effects in coding applications are somewhat limited. Recently, Najafzadeh et al. [5] incorporated the following forward masking model into the MPEG1 Layer II coder, in which they achieve significant coding gain,

\[ M_f(t,z_l) = a_0 \log_{10} \Delta t(L_m(t,z_l) - c) \]  

where \( M_f(t,z_l) \) is the amount of forward masking (dB) of the current frame in the ith critical band, \( \Delta t \) is the time difference between the masker and the maskee, \( L_m(t,i) \) is the masker level (dB) of the current frame in the ith critical band, and \( a, b, c \) are parameters that determine the amount of masking. In [5], the parameter \( a, b, c \) are set to 0.2, 2.3, and 20, respectively.

Assuming that forward masking has a duration of 200 milliseconds, \( b \) may be chosen as \( \log_{10}(200) \) [5], Fig. 2 shows the effect of varying the parameter of \( a, c, L_m, \) and \( \Delta t \) to the amount of masking \( M_f(t,z_l) \).

Motivated by Fig. 2, the amount of masking can be controlled easily by varying one parameter and keeping the other parameters constant. From equation (3) and Fig. 2, the parameter \( a \) was chosen to be varied and the parameter \( b \) and \( c \) were set to 2.3 and 20, respectively, as specified in [5]. By exploiting the property that in the human ear high frequencies are barely perceptible and the fact that the high frequencies of the WP coder have more coefficients than the low frequencies, we expect a good compromise between the bit rate reduction and audio quality by setting the amount of masking to be monotonically increasing towards the high frequency.

![Figure 2](image_url): The amount of forward masking, \( M_f \), with varying \( a, c, L_m, \) and \( \Delta t \).

After several psychoacoustic experiments with eight audio materials, we found that the following expression provides a good compromise between quality and bit rate reduction,

\[ a(z_l) = 0.5 + 0.0071 \times z_l \]  

The temporal information in the wavelet packet based audio coder is obtained by temporal distances between frames,

\[ \Delta t = \frac{FrameSize}{F_s} \times 10^{-3} \text{ ms} \]  

Since the longest duration of forward masking is 200 ms, then forward masking is calculated over \( N_f \) successive frames as follows:

\[ N_f = \left[ \frac{200}{\Delta t} \right] \]  

The amount of temporal masking level \( TM \) for each sub-band is then chosen as follows:

\[ TM(t,z_l) = \max\{M_f(t,z_l), \cdots, M_f(t-k\Delta t,z_l)\} \]  

where \( k = 1,2,\cdots,N_f \) and \( \Delta t \) is the time difference between frames.

3.4. Combined Masking Threshold

We use the power law method for combining the thresholds as follows [4]:

\[ MT(t,z_l) = \left( TM(t,z_l)^p + SM(t,z_l)^p \right)^{1/p} \quad p \geq 1 \]  

where \( MT(t,z_l) \) is the total masking threshold, \( TM(t,z_l) \) is temporal masking threshold, and \( SM(t,z_l) \) is the simultaneous masking threshold. The parameter \( p \) defines the way in which the masking thresholds are combined. Setting
\( p=1 \) corresponds to intensity addition, while taking larger values of \( p \) corresponds to using the highest masking threshold.

In order to evaluate which value could give a good audio quality, we calculated the total bit rate for various values of \( p \). The total bit rate is the number of bits required to code the wavelet packet coefficients plus the required side information that is sent to decoder. The result of the total bit rate versus the parameter \( p \) is shown in Fig. 3.

**Figure 3**: Total bitrate versus of the parameter \( p \).

After listening to the audio output, we found that the optimum \( p \) that gives a reasonable bit rate and transparent audio quality is when \( p=2 \). Hence, the parameter \( p \) is set accordingly.

4. Performance Evaluation

In this section, the performance of the WP coder was evaluated using the subjective listening tests complied with ITU-R BS.1116. Moreover, the compression rate using perceptual WP coder was investigated.

4.1. Audio Materials

The range of audio materials for the subjective testing of an audio coder should be chosen carefully so that it will provide the most demanding possible conditions to the coder. For this purpose, a wide range of audio materials has been selected with total of 8 audio files sampled at 44.1 kHz as shown in Fig. 4.

**Figure 4**: Audio materials used in the subjective listening tests, (a) Mariah Carey, (b) Eric Clapton, (c) Susan Vega, (d) Tracy Chapman, (e) Hani Anggraini, (f) Castanets, (g) Jazz, and (h) Male Speech.

4.2. Subjective Testing

In order to evaluate the quality of the compressed audio materials, we performed subjective tests conforming to ITU-R BS.1116. Six subjects took part in the listening tests. First, the subjects were trained to become familiar with the coding artifacts commonly found in audio coding [8], such as pre-echo, aliasing, birdies, and speech reverberation.

The next step was to conduct listening tests following ITU-R BS.1116 recommendation which is well known as “double-blind triple-stimulus with hidden reference”. The basic principle of that particular test method is briefly described as follows: the listener can select between three sources (“A”, “B” and “C”). The known Reference Signal is always available as source “A”. The Hidden Reference Signal and the Signal Under Test are simultaneously available but are “randomly” assigned to “B” and “C”, depending on the trial.

The listener is asked to assess the impairments on “B” compared to “A”, and “C” compared to “A”, according to the continuous five grade impairment scale. One of the sources, “B” or “C”, should be indiscernible from source “A”; the other one may reveal impairments. Any perceived differences between the reference and the other source must be interpreted as impairment. The grading scale shall be treated as continuous with “anchors” derived from the ITU-R five-grade impairment scale given in Recommendation ITU-R BS.562. The grading scale was continuous from 1 (very annoying) to 5 (no difference between the reference and the coded file).

During the listening test, each subject was free to take as much time as required on any trial, switching freely among the three stimuli as often as desired. The ABC/HR software was used for that purpose. Furthermore, the high quality headphone, i.e. Plantronics DSP-500, was utilized.

There are a total of 16-compressed audio materials, 8 using simultaneous masking only (SM) and 8 using combined masking (SM+TM). Table 1 shows the subjective difference grade (SDG) result for 6 subjects while Fig. 5 shows the average subjective scores. Note that, SDG was calculated as follows:

\[
SDG = \text{Grade}_{\text{Signal Under Test}} - \text{Grade}_{\text{Reference Signal}} \tag{10}
\]
Using the SDG, it is possible to know when the subject misclassified the reference signal as the coded signal. In that case, the SDG will have a positive value (highlighted at Table 1) instead of a negative value. As stated in [4], a misclassification of 50% shows that the transparent quality has been achieved. Except for Castanets and Male Speech, which have sharp attack, (see Figure 4), we can achieve a transparent audio quality. The prominent artifact of both signals is pre-echo distortion. Further research is required to control the distortion, such as bit reservoir, window switching, gain modification, and temporal noise shaping [3].

![Figure 5: Average Subjective Scores](image)

4.3. Bitrate reduction

![Table 2: Required average bitrate (in kbit/s)](image)

Eight audio materials were encoded and decoded using the WP coder with SM only and with SM+TM producing total of 16 outputs. The listening tests as described in the previous section showed that the high quality of the compressed sounds could be maintained while the bit rate was reduced by more than 17% on average if we include the proposed temporal masking method. Note that the bit rates shown in Table 2 have included the side information.

5. Conclusions

We have presented the design and implementation of a perceptual wavelet audio coder that incorporates temporal and simultaneous masking models. The performance of our coder was evaluated using eight audio files. Subjective listening tests conforming to ITU-R BS.1116 standard revealed that transparent quality is achievable, while the bit rate is reduced by more than 17% on average compared with using a coder that only employs a simultaneous masking model. Further investigation is required to control the pre-echo distortion found in some audio recordings.

6. References


