TEMPORAL MASKING FOR BIT-RATE REDUCTION IN AUDIO CODEC BASED ON FREQUENCY DOMAIN LINEAR PREDICTION

Sriram Ganapathy1,2, Petr Motlicek1, Hynek Hermansky1,2, Harinath Garudadri3

1IDIAP Research Institute, Martigny, Switzerland
2École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
3Qualcomm Inc., San Diego, CA, USA
{ganapathy,motlicek,hynek}@idiap.ch, {hgarudad}@qualcomm.com

ABSTRACT

Audio coding based on Frequency Domain Linear Prediction (FDLP) uses auto-regressive model to approximate Hilbert envelopes in frequency sub-bands for relatively long temporal segments. Although the basic technique achieves good quality of the reconstructed signal, there is a need for improving the coding efficiency. In this paper, we present a novel method for the application of temporal masking to reduce the bit-rate in a FDLP based codec. Temporal masking refers to the hearing phenomenon, where the exposure to a sound reduces response to following sounds for a certain period of time (up to 200 ms). In the proposed version of the codec, a first order forward masking model of the human ear is implemented and informal listening experiments using additive white noise are performed to obtain the exact noise masking thresholds. Subsequently, this masking model is employed in encoding the sub-band FDLP carrier signal. Application of the temporal masking in the FDLP codec results in a bit-rate reduction of about 10% without degrading the quality. Performance evaluation is done with Perceptual Evaluation of Audio Quality (PEAQ) scores and with subjective listening tests.

Index Terms— Audio coding, Psychoacoustic modelling, Temporal masking, Frequency Domain Linear Prediction (FDLP)

1. INTRODUCTION

A new audio coding technique based on modeling the spectral dynamics has been proposed in [1], [2]. The input audio signal is first decomposed into frequency sub-bands using a Quadrature Mirror Filter (QMF) bank. Subsequently, each sub-band signal is transformed using Discrete Cosine Transform (DCT) and linear prediction is performed in the DCT domain. This results in an approximation to the Hilbert envelope of that sub-band signal. The technique is referred to as Frequency Domain Linear Prediction (FDLP) [3]. The basic FDLP codec achieves acceptable quality, but at the expense of higher bit rate, necessary for encoding the sub-band FDLP carrier (residual) signal [2].

The auditory masking properties of the human ear provide an efficient solution for quantization of a signal in such a way that the audible distortion is minimized. In particular, temporal masking is a property of the human ear, where the sounds appearing within a temporal interval of about 200 ms after a signal component get masked. A simple first order model of forward masking was proposed in [4].

The long term processing (1000 ms) in the FDLP codec allows for a straightforward exploitation of the temporal masking, while its implementation in more conventional short-term spectra based codecs has been so far quite limited, one notable exemption being the recently proposed wavelet-based codec [5].

In this paper, the linear forward masking model proposed in [4] is applied to the QMF sub-band signal. Since the real conditions differ from the assumptions made in this linear model, listening experiments are conducted to determine the correction factors. This is done by adding white noise of decreasing power till the noise becomes just inaudible. The thresholds obtained for white noise, also indicate the maximum permissible power level for the quantization noise. These masking thresholds are then utilized in quantizing the sub-band FDLP carrier signals.

The paper is organized as follows: In Section 2, a mathematical description of the temporal masking property is given along with the determination of the masking thresholds using the white noise listening experiments. Section 3 explains the application of temporal masking for bit-rate reduction in the FDLP codec. The results of the objective and subjective evaluations are reported in Section 4.

2. TEMPORAL MASKING IN HUMAN AUDITORY SYSTEM

Temporal masking can be explained as a change in the time course of recovery from masking or as a change in growth of masking at each signal delay. The amount of forward masking is determined by the interaction of a number of factors including masker level, the temporal separation of the masker and the signal, frequency of the masker and the signal and duration of the masker and the signal [4]. A simple first order mathematical model, which provides a sufficient approximation for the amount of temporal masking, is given in Equation 1:

\[
M[n] = a(b - \log_{10}(\Delta t))(X[n] - c),
\]

where \(M\) is the temporal mask in dB Sound Pressure Level (SPL), \(X\) is the signal dB SPL level, \(n\) is the sample index, \(\Delta t\) is the time delay in ms, \(a\), \(b\) and \(c\) are the constants. At any sample point, multiple mask estimates arising from the several previous samples are present and the max of it is chosen as the mask in dB SPL at that point. The optimal values of these parameters, as defined in [5], are as follows:
\[ a = k_2 f^2 + k_1 f + k_0, \]  
\[ c = 3.64 f^{0.8} - 6.5 e^{0.6(f-3.3)^2} + 0.001 f^4. \]

2.1. An alternative SPL definition

A short-term SPL definition is needed to estimate the masking threshold at each sample index. For this purpose, the signal is divided into 10 ms overlapping frames with frame shifts of 1 sample. The estimated short-term power in SPL is assigned to the middle sample of the frame:

\[ X[n] = 10 \log_{10} \left[ \sum_{i=n-L/2}^{n+L/2} x^2[i] \right], \]

where \( X \) is the signal in dB SPL, \( x \) denotes the original time domain signal and \( L \) denotes the frame length (10 ms).

2.2. White noise experiments for determining the thresholds from temporal masking

The linear masking model is based on a set of assumptions like sinusoidal nature of the masker and signal, minimum duration of the masker (300 ms), minimum duration of the signal (20 ms) etc [4]. Such assumptions do not hold in a practical case, where we need to determine the masking thresholds for every sample point. Hence, the actual masking thresholds are much below the thresholds from the linear masking model. Therefore, we perform listening experiments using additive white noise to obtain the correction factors on the linear model.

The main goal of these experiments is to determine the maximum imperceptible SPL of white Gaussian noise that can be added to the audio signal. The power level of white noise is chosen as a sum of the linear masking model and a correction factor to be reduced from the linear model. Also, the correction factor is made dependent on the ATH in that frequency band. For different levels of signal SPL, white noise of decreasing powers (in steps of 5 dB SPL) is added until the noise becomes just imperceptible. The temporal masking thresholds obtained from these experiments can be mathematically written down as:

\[ T[n] = L_m[n] - (35 - c), \quad \text{if} \quad L_m[n] \geq (35 - c) \]
\[ = L_m[n] - (25 - c), \quad \text{if} \quad (25 - c) \leq L_m[n] \leq (35 - c) \]
\[ = L_m[n] - (15 - c), \quad \text{if} \quad (15 - c) \leq L_m[n] \leq (25 - c) \]
\[ = c, \quad \text{if} \quad L_m[n] \leq (15 - c) \]

where \( L_m \) is maximum of the temporal mask \( M \) computed from its previous samples. An example of a 1000 ms duration of a sub-band signal in the dB SPL, its linear masking model obtained from Equation 1 and the white masking thresholds obtained using Equation 5 is shown in Fig. 1.

3. APPLICATION OF THE TEMPORAL MASK FOR ENCODING THE SUB-BAND FDLP CARRIERS

The white noise temporal masking thresholds also denote the maximum permissible quantization noise in that sub-band. The starting point is the base-line FDLP codec without temporal masking [2]. The number of bits required for representing the sub-band FDLP carrier is reduced in accordance with the temporal masking thresholds. Since the sub-band signal is the product of its FDLP envelope and carrier, the masking thresholds for the carrier signal are obtained by subtracting the dB SPL of the envelope from that of the sub-band signal.

The first step is to estimate the quantization noise present in the base-line version; if the mean of the quantization noise (in 200 ms sub-band signal) is above the masking threshold, no bit-rate reduction is applied. If the temporal mask mean is above the noise mean, then the amount of bits needed to encode that sub-band carrier signal is reduced in such a way that the noise level becomes similar to the masking threshold.

Since the information regarding the number of quantization bits is to be transmitted to the receiver, the bit-rate reduction is done in a discretized manner. In the proposed version of the codec, the bit-rate reduction is done in 8 different levels (in which the first level corresponds to no bit-rate reduction). The number of bits to be reduced is made dependent on the difference in dB SPL between the quantization noise and the mask threshold. When the difference is higher, bit-rate reduction is severe. Also, level of bit-rate reduction for each sub-band FDLP carrier is sent as side information to the receiver (around 0.5 kbps).

We show two examples of the application of temporal masking to reduce the bits needed for representing the sub-band FDLP carrier signal. For low-energy sub-band signals, the masking threshold, as given by Equation 5, is the ATH. Thus, for these regions of the signal, the number of bits spent on the FDLP carrier signal can be significantly reduced. This is shown in Fig. 2, where we plot a region of the 200 ms sub-band signal of low energy, the temporal mask which is at the ATH for that frequency band, quantization noise in the sub-band with and without temporal masking. The case for signal regions with high energy is shown in Fig. 3.
4. RESULTS

The subjective and objective evaluations of the proposed audio codec are performed using challenging audio signals (sampled at 48 kHz) present in the framework for exploration of speech and audio coding [6]. It is comprised of speech, music, and music over music recordings. The music samples contain a wide variety of challenging audio samples ranging from tonal signals to highly transient signals.

Objective evaluations are performed on 27 audio samples and the results show the bit-rate and quality advantage of using temporal masking in FDLP codec as compared with the base-line version. Subjective listening tests compare the final version of the FDLP codec (utilizing temporal masking) with the state of the art codecs on 8 audio samples from this database.

4.1. Objective Evaluations

The objective measure employed is the Perceptual Evaluation of Audio Quality (PEAQ) distortion measure [7]. In general, the perceptual degradation of the test signal with respect to the reference signal is measured, based on the ITU-R BS.1387 (PEAQ) standard. The output combines a number of model output variables (MOVs) into a single measure, the Objective Difference Grade (ODG) score, which is an impairment scale with meanings shown in Table 1.

Table 1. PEAQ scores and their meanings.

<table>
<thead>
<tr>
<th>ODG Scores</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>imperceptible</td>
</tr>
<tr>
<td>-1</td>
<td>perceptible but not annoying</td>
</tr>
<tr>
<td>-2</td>
<td>slightly annoying</td>
</tr>
<tr>
<td>-3</td>
<td>annoying</td>
</tr>
<tr>
<td>-4</td>
<td>very annoying</td>
</tr>
</tbody>
</table>

Table 2. Average PEAQ scores with and without masking.

<table>
<thead>
<tr>
<th>Base-line codec (73 kbps)</th>
<th>With Temporal Masking (66 kbps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.99</td>
<td>-1.11</td>
</tr>
</tbody>
</table>

4.2. Subjective Evaluations

MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) is a methodology for subjective evaluation of audio quality. It is defined by ITU-R recommendation BS.1534 [8]. We perform the MUSHRA tests on 8 audio samples from the database with 22 listeners. The subjective quality of following 3 codecs are compared:

1. FDLP codec with temporal masking at ~ 66 kbps.
2. LAME MP3 (MPEG 1, layer 3) [9] at 64 kbps.

The results of the MUSHRA tests are shown in Figure 5. It is found that the proposed version of the codec with temporal masking, is competitive with the state of the art codecs at similar bit-rates.
5. CONCLUSIONS

In order to improve the compression efficiency of audio codecs based on spectral dynamics, a method to advantageously use temporal masking phenomenon of the human ear has been proposed. The application of the temporal masking for the FDLP codec has reduced the bit-rates by about 10%. Objective evaluations justify the importance of this novel technique of bit-rate reduction. Also, the version of the codec exploiting temporal masking gives subjective results competitive to the state of the art codecs at similar bit-rates. It is worth noting that this performance is achieved without utilizing standard modules like entropy coding and simultaneous masking. The inclusion of these techniques form part of the future work.

6. REFERENCES


