An investigation of non-uniform bandwidths auditory filterbank in audio coding

Andrew Lin, Stevan Berber, Waleed Abdulla

Department of Electrical and Computer Engineering
University of Auckland, New Zealand
alin051@ec.auckland.ac.nz, s.berber@auckland.ac.nz, w.abdulla@auckland.ac.nz

Abstract

This paper presents an investigation on the use of non-linear auditory filterbank in wideband audio coding. The perceptually based parameterization of the audio signal using gammatone filterbank is examined and discussed. Conventional gammatone filters requires high order FIR filters in the synthesis stage which introduces long delay and large computation cost. Here, a simple and efficient synthesis technique is investigated and embedded into a perceptual audio coder. A limitation of this system is that the outputs of the filterbank can not be maximally decimated due to the non-uniform bandwidths of the filterbank, yet the coder achieves near transparent quality at a bit rate of 120 kbps.

1. Introduction

Much of the popular audio coders nowadays are perceptual audio coders (PAC), which perform lossy data compression by minimizing the perceived distortion (M. Bosi, 2003; Marina Bosi & Goldberg, 2003; Sinha, Johnston, Dorward, & Quackenbush, 1998). Models of auditory system are employed to determine the masking threshold, which defines the boundary of audible and inaudible parts of the signal. Compression is achieved by reducing the number of bits allocated to the audio signal such that quantization noise is below the masking threshold, hence inaudible.

The coding gain is very much depended upon the quality of the psychoacoustic model and the parameterization of the signal (Marina Bosi & Goldberg, 2003). A well-defined psychoacoustic model that closely mimics the auditory system can save bits from signal redundancies, while efficient parameterization of the signal can reduce the signal irrelevancies. The major drawback in the current MPEG-1 audio coders is that it fails to parameterize the signal suitable for the auditory system. In MPEG-1, uniform (equal bandwidth) spectral decomposition is performed to account for the band-pass filtering effect of the cochlea (Baumgarte, 2002). However, the actual auditory system performs non-uniform spectral decomposition (Fletcher, 1940; Moore & Glasberg, 1983). Uniform bandwidth filterbanks were initially used because of their low computational cost.

This paper describes the use of auditory-domain filterbanks in wideband audio coding. Gammatone filterbank (GTFB) has been selected because of its inherent auditory-based properties (de Boer, 1975; Patterson et al., 1992; Slaney, 1993). GTFB have been extensively utilized in computational auditory models of the cochlea (Lopez-Poveda & Meddis, 2001; Patterson et al., 1992) and speech processing (Abdulla, 2002; Ambikairajah & Gunawan, 2004; Ambikairajah, Lin, & Holmes, 2004), but it has not been well investigated in audio and speech coding. Kubin (Kubin & Bastiaan Kleijn, 1999) first started the use of gammatone filterbank in speech coding based on the model of auditory neuron firing behavior. Ambikairajah and Lin (Ambikairajah, Epps, & Lin, 2001) further enhanced the coding in (Kubin & Bastiaan Kleijn, 1999) by introducing the temporal and simultaneous masking properties. Both implementations though, have only studied the coding of speech and audio with 8 kHz signal bandwidth. In this paper, wideband 44.1 kHz sampled audio signals are considered. Furthermore, we investigate a simple and efficient synthesis approach presented in (Hohmann, 2002) instead of the usual time-reversed FIR synthesis technique found in (Ambikairajah et al., 2001; Kubin & Bastiaan Kleijn, 1999).

This paper is organised as follows: section 2 starts with an overview of the conventional analysis/synthesis GTFB, followed by a discussion on the issues observed when they are used in wideband audio coding. Then, a more efficient implementation of the GTFB from (Hohmann, 2002) is described. Section 3 presents the
employment of such GTFB in a perceptual audio coder and the limitation of this coder is discussed. Section 4 concludes this paper.

2. Gammatone Analysis/Synthesis Filterbank

Gammatone filters are a special type of filters whose pass-band bandwidth matches closely to the equivalent rectangular bandwidth (ERB) of the our auditory system (Patterson et al., 1992). The derivation of ERB is based on the theory of critical bandwidth except it provides finer measurement for frequencies below 500 Hz. The analytical expression of ERB as a function of its center frequency $f_c$ is defined by:

$$ERB = l + f_c / q$$

where $l = 24.7$ and $q = 9.265$

The ERB is associated with the gammatone filter by the filter’s sampled impulse response:

$$g(n) = A(nT)^{N-1} e^{2\pi q f_c nT} \cos(2\pi f_c nT + \phi)$$

Where $A$ is the gain factor for unity magnitude response, $T$ is the sampling period, $N$ is the order of the filter, $f_c$ is the centre frequency of the filter and $\phi$ is the phase term for complex filters. Patterson (Patterson et al., 1992) has shown that 4th order filter ($N=4$) is a good approximation of the human auditory response. Discrete gammatone filters can be implemented with FIR or IIR scheme. For FIR, the gammatone impulse response in (2) specifies the coefficients of the FIR filter, and for IIR, a 4th order IIR gammatone filter can be implemented by cascading four first-order band-pass filter (Hohmann, 2002; Slaney, 1993). An all-pole approximation IIR can also be used for simplicity because zeros have a small effect near the centre frequency of each filter (Slaney, 1993). The transfer function of an all-pole 4th order gammatone filter is defined by:

$$G(z) = \frac{1}{(1 - \alpha z^{-1})^4}$$

where the coefficients $\alpha$ can be real or complex.

A general block diagram of the analysis/synthesis GTFB is shown in Figure 1, which consists of $K$ gammatone filters connected in parallel and each channel is a gammatone band-pass filter centered at $f_{c1}$ to $f_{cK}$. The synthesis GTFB is usually implemented using the time-reversed impulse response of the analysis filter, as found in (Ambikairajah et al., 2001; Kubin & Bastiaan Kleijn, 1999).

Ideally the center frequencies of each channel should be designed such that there is no gap between each pass-band. A so-called ERB scale is defined in (Hohmann, 2002) to measure the relative distance between each filter. The ERB scale is almost logarithmic and is defined by:

$$ERB_{scale}(f) = q \log(l + f / lq)$$

where $l = 24.7$ and $q = 9.265$

By placing the center frequency of each channel at fixed steps on the ERB scales, the pass-bands of the filters cover the whole spectrum without any gap. A fully covered frequency spectrum ranging from 50 to 22050 Hz requires 40 gammatone filters as shown in Figure 2.

2.1 Gammatone filterbank design issues

The analysis/synthesis filterbank can be implemented in FIR/FIR, IIR/FIR, or IIR/IIR scheme (Lin, Ambikairajah, & Holmes, 2001). The major problem with the use of FIR is that high order filters are usually required to accurately define the gammatone filters centered at low frequencies. For example, the impulse response of a gammatone filter centered at 100 Hz with a sampling rate of 44.1 kHz is shown in Figure 3. To measure the effective length of this impulse response, a
method based on the accumulated energy of the impulse response from (Laakso & Valimaki, 1999) is used. The effective length for 95% accumulated energy of the impulse response in Figure 3 is 1130 samples, hence, the FIR gammatone filters would require at least 1130 taps.

Figure 3: Impulse response of a gammatone filter centered at 100 Hz with a sampling rate of 44.1 kHz.

High order FIR is computationally expensive and the large delay associated with long filters is undesirable, but insufficient length of the filters can result in distortion of the reconstructed signal especially at low frequencies. Figure 4 shows the overall frequency response of a 40 channel FIR/FIR analysis/synthesis filterbank system at low frequencies. The system bandwidth is ranging from 100 to 22050 Hz and the filter length is set to be 50% accumulated energy of the impulse response of the lowest frequency channel.

Figure 4: Overall analysis/synthesis filterbank frequency response at low frequencies. Insufficient filter length causes large ripples at low frequencies.

Subjective listening tests suggest that the filter length should be kept at least 80% accumulated energy of the impulse response of the lowest frequency channel to avoid distortion, this usually requires the FIR filters to have more than 1000 coefficients.

2.2 Gammatone Filterbank with a constant delay

Another analysis/synthesis GTFB system proposed in (Hohmann, 2002) avoids the use of FIR filters in synthesis. Each output of the analysis filterbank is delayed by a calculated amount such that the maximum of their real part align at the same time instant. Signal reconstruction is then performed by a weighted summation of the aligned signals in each channel. The block diagram of such analysis/synthesis filterbank system is shown in Figure 5.

Figure 5: Gammatone analysis/synthesis filterbank with a constant delay.

Given a desired overall system delay, the number of samples to be delayed for a particular channel is calculated by the relative position between the impulse response envelope peak and the desired delay position. Figure 6(a) shows the impulse response of channel 20 gammatone filter to be delayed to the 4ms position, that is, sample number 177 with a sampling rate of 44.1kHz shown by the vertical line. Figure 6(b) illustrates the delayed impulse response.

Figure 6: (a) The envelope peak of the impulse response is to be delayed to the desire delay position (vertical line). (b) The delayed impulse response.

For lower frequency channels the position of the envelope peak may be greater than that of the desired delay, as shown in Figure 7(a). In that case, the first local maximum of the real part of that impulse response is delayed to the desire delay position, depicted in Figure 7(b).
The envelope peak of the impulse response at channel 5 is greater than the desired delay of 4 ms. The delayed impulse response with its first local maximum at the desired delay position.

The trade-off between an optimal system delay and the quality of the reconstructed signal needs to be considered. If the system delay is chosen to be too low, severe fluctuation of the magnitude response at lower frequencies will result. A plot of the signal to noise ratio of the reconstructed signal with respect to system delay is given in Figure 8. The SNR remains constant when the system delay is more than 12 ms, because the desired delay position is now greater than the envelope peak position of the lowest frequency channel.

The synthesis technique in (Hohmann, 2002) offers a better solution than the conventional FIR/FIR or IIR/FIR analysis/synthesis filterbank as the large computational cost and delay from the high order FIR filters are avoided. However, the optimal delay factor for the system needs to be considered carefully because it affects the quality of the reconstructed signal. Evaluation from (Hohmann, 2002) suggests a system delay of 4 ms is acceptable in terms of perceptual quality.

3. GTFB in wideband audio coding

The block diagram of the proposed perceptual audio encoder and decoder with the use of GTFB is shown in Figure 9.

The input signal is a CD-quality mono digital signal, sampled at 44.1 kHz with 16 bits per sample. The frame length is chosen to be 512 samples which is similar to that in MPEG-1 audio. Generally, long frame-length provides a better frequency resolution while short frame-length gives a better temporal resolution (Marina Bosi & Goldberg, 2003). The optimal frame length for the proposed coder still needs to be addressed.

3.1 Encoder

The block schematic of the perceptual audio coder with gammatone filterbank. (a) Encoder (b) Decoder

Figure 9: Block schematic of the perceptual audio coder with gammatone filterbank. (a) Encoder (b) decoder
The analysis GTFB consists of K channels with each gammatone filter $G_k(z)$ for $k=1,2,...,K$ centered at frequencies ranging from 50 Hz to 16 kHz. The reason for not having gammatone filters at more than 16 kHz is because the psychoacoustic model reveals that above 16 kHz the masking threshold is generally greater than the signal energy. The position of the center frequency of each channel is set according to the critical bands defined in (Zwicker & Fastl, 1999). There are 24 filters used.

The output from each channel is downsampled to reduce the data rate. The optimal down sampling factor for the proposed system is difficult to define since the bandwidths of the gammatone filter are non-linear and the filterbank itself is not orthogonal (Kubin & Bastiaan Kleijn, 1999). Usually in uniform filterbanks the downsampling factor is constant for every band. The aliasing term of the filter in one subband can be cancelled by another aliasing term of a filter in another subband with careful design of the synthesis filters. However for non-linear filterbanks many theoretical issues for perfect reconstruction still remain unresolved. Vaidyanathan (Akkarakaran & Vaidyanathan, 1999; Hoang & Vaidyanathan, 1989) has proposed some necessary conditions for the decimation of non-uniform filterbank but their method requires synthesis filters which is not used in our system. Therefore, the downsampling factor for our proposed system is chosen such that aliasing does not occur in the highest channel bandwidth. The highest bandwidth is 1700 Hz, so a downsampling factor of 12 is used for all channels.

The psychoacoustic model from MPEG-1 Layer II is used to determine the masking threshold (ISO/IEC Int'l Standard IS 11172-3) but a 2048 point FFT is used instead of a 1024 points FFT for finer frequency resolution. The bit allocation process allocates the minimum number of bits that will be used to code each channel such that the quantization noise level is below the masking threshold.

The signal level of each band is normalized between -1 and +1, and a uniform midtread quantizer is used to take into account of the zero amplitude. 6 bits are used to code the normalization factor of each channel and 4 bits are used to code the number of bits allocated to each channel, similar to that in MPEG-1 audio.

3.2 Decoder

The decoder decodes and dequantizes the signal of each channel, then upsamples the signal back to the input sampling frequency. Next, image extraction filters that are identical to the analysis GTFB are used to recover the frequency content of each channel. The extracted signal in each channel is then delayed by the calculated amount such that they all align at the desired delay position. A gain factor is applied to all signals before summation to make the overall system gain unity.

3.3 Subjective listening tests

The result average bit rate of the system is about 1400 bits per frame, which rounds to 120 kbps. Two types of listening test were conducted. First, the coded audio signal was compared to the original CD-quality audio signal. A grade between 0 and 5 was used to measure the quality, with 0 being a disturbing degradation of the quality and 5 being unnoticeable quality change. Second, a ‘double blind’ test was conducted to evaluate the quality of the proposed coder with a benchmark coder. MPEG-1 Layer III coded at 96 kbps was chosen to be the benchmark coder. Listeners were asked to select the better quality coder without the knowledge of which coder it was. 1 and 0 were given for the preferred and less-preferred coder respectively, 0.5 was given for ‘not sure’. 10 items were used in the test, which were different genres of music as well as solo instruments and vocals. 5 non-professional listeners were involved in this test.

Table 1 summarises the results of the subjective listening test comparing the original and the coded audio signals. It demonstrates that the proposed coder provides adequate quality. However, some noticeable differences in quality were found with percussive instruments such as the glockenspiel and castanet. Table 2 summarises the results from the ‘double blind’ test. The proposed coder demonstrates compatible quality with the MPEG-1 Layer III coded at 96 kbps for most type of music.

<table>
<thead>
<tr>
<th>Test Item</th>
<th>Average Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jazz</td>
<td>3.5</td>
</tr>
<tr>
<td>Pop</td>
<td>3.33</td>
</tr>
<tr>
<td>Rock</td>
<td>4.3</td>
</tr>
<tr>
<td>Classical</td>
<td>3.87</td>
</tr>
<tr>
<td>Soprano</td>
<td>4</td>
</tr>
<tr>
<td>Bass</td>
<td>2.5</td>
</tr>
<tr>
<td>Violin</td>
<td>3.73</td>
</tr>
<tr>
<td>Trumpet</td>
<td>3.3</td>
</tr>
<tr>
<td>Glockenspiel</td>
<td>2</td>
</tr>
<tr>
<td>Castanet</td>
<td>2.3</td>
</tr>
</tbody>
</table>

Table 1: Quality grading of the proposed coder.
Table 2: Average probability of the proposed coder preferred over MPEG-1 Layer III @ 96 kbps

<table>
<thead>
<tr>
<th>Test Item</th>
<th>Average probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jazz</td>
<td>0.5</td>
</tr>
<tr>
<td>Pop</td>
<td>0.5</td>
</tr>
<tr>
<td>Rock</td>
<td>0.55</td>
</tr>
<tr>
<td>Classical</td>
<td>0.4</td>
</tr>
<tr>
<td>Soprano</td>
<td>0.6</td>
</tr>
<tr>
<td>Bass</td>
<td>0.35</td>
</tr>
<tr>
<td>Violin</td>
<td>0.5</td>
</tr>
<tr>
<td>Trumpet</td>
<td>0.5</td>
</tr>
<tr>
<td>Glockenspiel</td>
<td>0.2</td>
</tr>
<tr>
<td>Castanet</td>
<td>0.2</td>
</tr>
</tbody>
</table>

4. Conclusions

The bit rate of the proposed coder is higher compared to the MPEG-1 Layer III audio coder, which also achieves near transparent quality at a bit rate of 96 kbps. The high bit rate is mostly due to the limitation in decimation, as the outputs from the filterbank can not be maximally downsampled. Nonetheless, a computationally efficient synthesis technique that avoids the use of long length FIR filters is more suitable for wideband audio signals with high sampling frequencies, and a low system delay of 4 ms is attractive for application of real-time encoding/decoding.

5. Acknowledgements

The author would like to thank sincerely to Associate Professor E. Ambikairajah as well as Dr. Volker Hohmann for their provision of the technical papers and simulation codes.

6. References


