Joint Time-frequency Coding of Audio Signals

KARTHIKEYAN UMAPATHY and SRIDHAR KRISHNAN
Dept. of Electrical and Computer Engineering,
Ryerson University,
350, Victoria street, Toronto, ON M5B 2K3, CANADA.
Email: (karthi)(krishnan)@ee.ryerson.ca, url:www.ee.ryerson.ca/~krishnan/ATFT.html

Abstract: Wideband audio signals have a very high data-rate associated with them due to their complex nature and demand for high quality reproduction. Efficient utilization of channel bandwidth and storage space drive the need for low bit-rate, high quality audio coder for multimedia applications. In this paper, we propose a novel low bit rate high quality coder where the joint time-frequency (TF) properties of the non-stationary nature of the audio signals are exploited in creating a compact energy representation of the signal into fewer coefficients. The decomposition coefficients are then quantized to produce a low bit-rate output carrying perceptually relevant information for reproduction of the audio signal. This approach of exploiting the TF properties also has a major advantage of automatic filtering of noise (which are normally non-localised in both time and frequency domains) from the signal. Initial studies with a variety of audio signals has given a good compression ratio in the range of 6 to 25.

Keywords: time-frequency functions, variable quantizers, Gabor functions, audio coding, matching pursuit, perceptual coding.

1. Introduction

Digital audio coding has undergone many developmental phases by different signal processing techniques over the years. Most of the existing audio coding schemes, encode the data either in the time domain or frequency domain alone [1]. Time domain coders exploit the temporal redundancy between audio samples, and use the difference between the samples to represent the correlated waveform, so that the same SNR is maintained at a reduced bit rate. In the frequency domain coders, the audio signal is transformed into frequency domain, and processed further to remove the redundant information. Table 1 lists few well known coders of wideband audio with their associated techniques and maximum compression ratios.

However all these existing techniques suffer from limited time and/or frequency resolution. It is known that audio signals are non-stationary in nature and require a time-frequency approach in processing them for achieving optimal time and frequency resolution. Therefore, the motivation of this paper is to exploit joint time-frequency (TF) correlation to efficiently process and code the audio data into compact energy representations. The approach taken in this paper is signal decomposition using Gabor TF functions as basis functions. The signal is first decomposed by using an adaptive TF decomposition algorithm such as matching pursuit(MP) [3] and the coefficients are processed and filtered for perceptual threshold of hearing to achieve compression. The block diagram of the proposed coder is shown in Figure 1. The perceptual filtering used in the coder is applied in a novel way to the decomposed components of a signal rather than on the signal itself as used by other coders.

2. Materials and Methods

2.1 Adaptive Signal Decomposition

The adaptive signal decomposition algorithm used is the MP algorithm based on a Gabor dictionary [3]. In the MP algorithm any signal \( x(t) \) is decomposed into a linear combination of TF functions \( g(t) \) selected from a overcomplete dictionary of TF functions.

\[
x(t) = \sum_{n=0}^{\infty} a_n g_m(t),
\]  

(1)
Table 1: SC-Subband coding, PM -Psycho acoustic model, TC -Transform coding, CR -Compression ratio

<table>
<thead>
<tr>
<th>Codec</th>
<th>Technique</th>
<th>CR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISO/MPEG1-L3</td>
<td>SC,TC,PM</td>
<td>14</td>
</tr>
<tr>
<td>ISO/MPEG1-L2</td>
<td>SB,TC,PM</td>
<td>8</td>
</tr>
<tr>
<td>MUSICAM</td>
<td>SC,PM</td>
<td>11</td>
</tr>
<tr>
<td>PASC</td>
<td>SC,PM</td>
<td>5.5</td>
</tr>
<tr>
<td>ASPEC</td>
<td>TC,PM</td>
<td>11</td>
</tr>
</tbody>
</table>

where

$$g_{\gamma_n}(t) = \frac{1}{\sqrt{s_n}} g\left(\frac{t-p_n}{s_n}\right) \exp\left\{j(2\pi f_n t + \phi_n)\right\}, \quad (2)$$

and $a_n$ are the expansion coefficients. The scale factor $s_n$ is used to control the width of the window function, and the parameter $p_n$ controls the temporal placement. The parameters $f_n$ and $\phi_n$ are the frequency and phase of the exponential function respectively.

The signal $x(t)$ is projected over a redundant dictionary of TF functions with all possible combinations of scaling, translations and modulations. The dictionary of TF functions can either suitably be modified or selected based on the application in hand. In our technique we are using the Gabor dictionary of Gaussian functions which has the best TF localization properties [6]. At each iteration the best correlated TF function is selected from the dictionary. Each TF function is a Gaussian blob or in other words a Gauss modulated sine waveform with a definite time, frequency, and energy localizations. The remaining signal called the residue is further decomposed in the same way at each iteration subdividing them into TF functions. After $M$ iterations, signal $x(t)$ can be expressed as

$$x(t) = \sum_{n=0}^{M-1} \langle R^n x, g_{\gamma_n}\rangle g_{\gamma_n}(t) + R^M x(t), \quad (3)$$

where the first part of $x(t)$ is the decomposed TF functions till $M$ iterations, and the second part is the residue which will be decomposed in the subsequent iterations. This process is repeated until the energy level drops to 0.5% of the original signal energy or the rate of residue curve falls below a certain percentage (0.0002%).

From our experiments it is observed that any signal decomposed to 99.5% of its original energy is sufficient enough for good quality audio reproduction. The MP algorithm breaks the signal based on coherent and non-coherent structures of the signal. The significant energy content of the signal is present in its coherent structures, and the non-coherent noise like structures generally occupy the low energy insignificant band. Since these low energy non-coherent structures do not have a localized TF characteristics, MP algorithm breaks them into finer components trying to find a match for them in the dictionary. This makes the rate of residue decay or the energy decay curve to slow down considerably after certain number of iteration, and it is observed further decomposition does not yield significant information of the signal. Figure 2 and 3 illustrate this effect. From the figures we can observe that for the sample audio signal, MP does not perform a significant decomposition after 6900 TF functions. The number of TF functions needed for every music segment varies based on the amount of coherent and non-coherent structures present in them. This condition limits the number of TF functions needed for any music segment yielding the first stage of compression.

2.2 Perceptual coding

The perceptual coding is based on the human psychoacoustic properties. The non-linear frequency response and the limited frequency resolving
Figure 2: Energy decay curve. au-arbitrary unit

Figure 3: Residue decay curve. au-arbitrary unit

power of our ears enables us to compress the audio signal further by discarding components of the signal that does not satisfy certain conditions. The minimum audible level with respect to frequency location is given by the absolute threshold or threshold in quiet (TIQ) curve [4]. The absolute threshold (TIQ) approach used in the proposed coder is explained in the next section.

2.2.1 Absolute threshold

A novel approach of perceptual filtering is employed in our coder. The perceptual filtering is applied on the MP decomposed TF functions rather than on the signal. This prevents us to use many of the already available standard perceptual curves. To implement the absolute threshold model for our technique two experiments were performed by studying the sound produced by TF functions with different energy levels and frequency spreads.

Experiment 1:

The first experiment was performed on 5 listeners aged between 20 and 30. In the experiments, the listeners were played with individual TF functions with different center frequencies of duration 0.4 sec (Octave 14) through a high quality head phone (Sennheiser), and by varying the energy level. The feedback of the listeners were monitored, and the readings were recorded when the listeners could not make out any perceptual sound. The procedure was repeated till the reproducibility of the results were achieved. Figure 4 shows the average TIQ curve of the 5 listeners.

Figure 4: Threshold in quiet curve (for TF functions)

Experiment 2:

The second experiment was performed to find the relation between the octave and the absolute magnitude of the TF function for a particular energy. In this experiment an individual TF function with a fixed energy was monitored by varying the octaves in steps of 1. Figure 5 demonstrates the relation obtained between octave and the absolute magnitude of a TF function for unit energy. This relation is used to calculate the absolute magnitude of the TF function based on its energy and octave values. The decomposed TF functions are sorted into three groups of center frequencies: below 500 Hz, between 500 Hz to 9 kHz, and above 9 kHz. The TIQ curve is applied on the sorted TF functions and checked for the energy level relation with the frequency and octave. The TF functions that
fell below the TIQ curve were removed.

![Amplitude scaling with octave](image)

Figure 5: Amplitude scaling with octave

2.3 Quantization

The parameters \((a_n, p_n, f_n, s_n, \text{ and } \phi_n)\) from equations 1 and 2) needed to represent each TF function is quantized with varying resolution based on the dynamic range of the parameters. Logarithmic compression were used on the energy \(a_n\) and the frequency \(f_n\) parameters to reduce the number of bits needed. After an extensive study, with 16 signals, 51 bits were arrived at for representing each TF function. The total number of bits needed to represent a music segment would be 51 times the total number of TF functions remaining after perceptual filtering.

3. Results and Conclusion

From the spectrogram displays in Figure 9 and 10 of the original signal Figure 7 and the reconstructed signal Figure 8 respectively of a sample music segment, it can be seen that the perceptionally relevant TF characteristics of the signal are well maintained with an added advantage of denoising [2]. Table 2 gives the compression ratios achieved with 4 different sample signals (10 sec duration each at 44.1 kHz sampling rate) in comparison with wavelet packet coder and the MPEG layer-I coder [5]. The sound version of the sample signals can be heard from the web site url: ee.ryerson.ca/~krishman/ATFT.html.

![Percentage of audience](image)

Figure 6: MOS - Mean opinion score

Both original and compressed version of all the above four music segments were played to an audience of 25 people. The perceptual similarity were rated based on the feedback of the audience. Figure 6 illustrates the mean opinion score (MOS) in which the percentage of audience voted that they could not perceptually identify the difference between the original and compressed signal are graphed. The results suggest that the proposed coder gives very good compression ratio for music signals. Further work involves improving psycho acoustic model by incorporating more perceptual reduction techniques, and testing with wide variety of music signals.

### Table 2: Compression ratios. NT -not tested, WP -Wavelet packet coder, MPEG-LI/I -MPEG Layer-I/psycho acoustic model -I coder, PC -Proposed coder

<table>
<thead>
<tr>
<th>Audio Signal</th>
<th>WP</th>
<th>MPEG-LI/I</th>
<th>PC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enya</td>
<td>9.27</td>
<td>5.49</td>
<td>12</td>
</tr>
<tr>
<td>Harp</td>
<td>10.98</td>
<td>7.10</td>
<td>20</td>
</tr>
<tr>
<td>Piano</td>
<td>14.59</td>
<td>NT</td>
<td>25</td>
</tr>
<tr>
<td>complex</td>
<td>NT</td>
<td>NT</td>
<td>7</td>
</tr>
</tbody>
</table>

Acknowledgements

We gratefully acknowledge MICRONET and NSERC for supporting this project.
Figure 7: Original signal

Figure 8: Reconstructed signal

Figure 9: Spectrogram of the original signal in Figure 7

Figure 10: Spectrogram of the reconstructed signal in Figure 8

References


