Using Salient Envelope Features for Audio Coding

Joachim Thiemann[1] and Peter Kabal[1]

[1]*McGill University, Montréal, Québec, Canada*

Correspondence should be addressed to Joachim Thiemann-Author
(Joachim.Thiemann@Mail.McGill.CA)

ABSTRACT

In this paper, we present a perceptual audio coding method that encodes the audio using perceptually salient envelope features. These features are found by passing the audio through a set of gammatone filters, and then computing the Hilbert envelopes of the responses. Relevant points of these envelopes are isolated and transmitted to the decoder. The decoder reconstructs the audio in an iterative manner from these relevant envelope points. Initial experiments suggest that even without sophisticated entropy coding a moderate bitrate reduction is possible while retaining good quality.

1. INTRODUCTION

Perceptual coding of audio is a concept that has been around since the 1980s. Every audio codec in widespread use today uses auditory modelling in some way to reduce the bitrate by not allocating bits to information that is inaudible to most people.

In general, codecs use fast block-transforms to perform a spectral analysis of the audio, and then code the coefficients of these transforms. Perceptual encoding applies rules derived from experimental observations to allocate bits in such a manner so most quantization noise is expected to be inaudible.

In related research, models of the human auditory system have been created by examining the physiology of the ear. Combined with experimental observations, these models have been successful at predicting human perception when evaluating codecs or sound enhancement algorithms.

This leads to the idea of perceptual coding using model inversion. By definition, the auditory model variables represent the audible part of an audio signal. Therefore, suitable quantisation of these variables should be a reasonable method of coding the audio signal. This would minimise the distortion as measured by the model. In this paper, we refer to the quantized model variables as Salient Features (SF).

The codec presented here ensures that the reconstructed audio is consistent with the model variables encoded in the bitstream; the key of the decoder is that it corrects the reconstructed audio to match the SF. While this increases overall delay and computational complexity (in particular at the decoder) an efficient encoding of the audio signal should be possible.

2. CODEC DESCRIPTION

The encoder presented here evolved from the research done by Feldbauer, Kubin, and Kleijn [1], where the underlying auditory model is based on the Dau model [2]. In that model, the movement of the Basilar Membrane (BM) is modelled by analysis of the audio signal with a set of gammatone filters. In contrast to Feldbauer's work, where a neural pulse model is used, in our work the SF are obtained from the Hilbert envelope of the gammatone filter responses. Previous work [3] has shown that this information is sufficient to reconstruct a perceptually similar audio signal.

The reconstruction of the audio signal is achieved

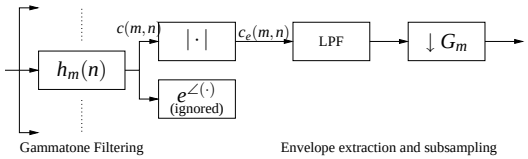

Fig. 1: Analysis of input signal

through a synthesis-by-analysis scheme where the initial synthesized audio signal is reanalyzed, corrected and synthesized in a loop until the analysis variables match the encoded parameters.

2.1. Auditory Hilbert envelopes

We denote the input signal to be coded $x(n)$, and assume that it is band-limited to the range of the filterbank described below. The input signal is passed through a parallel set of M complex FIR filters. The impulse responses of these filters are a sampled gammatone filter response with complex excitation

$$h_m(n) = t^{(l-1)} e^{-2\pi b(m) n/f_s} e^{-2\pi j f_c(m) n/f_s}, \quad m = 1,\ldots,M. \quad (1)$$

The parameters l and $b(m)$ are taken from Patterson [4]. The resonant frequencies of the filters $f_c(m)$ are chosen such that over the bandwidth of $x(n)$ there are two filters per critical band (CB).

The outputs of this filterbank are a set of complex modulated signals, which can be separated into envelopes $c_e(m,n)$ and complex carriers with unit magnitude $c_\phi(m,n)$,

$$\begin{aligned} c(m,n) &= \sum_{l=0}^{L-1} x(n-l) h_m(l), \\ c_e(m,n) &= \|c(m,n)\|, \\ c_\phi(m,n) &= c(m,n)/c_e(m,n), \\ &\qquad \forall m = 1,\ldots,M. \end{aligned} \quad (2)$$

Note that the original signal $x(n)$ can not be recreated exactly unless the carrier is known, since without the carrier, the absolute phase is unknown. However, if some conditions are met, a signal can be found that is perceptually equivalent given only the envelope information [3].

The codec presented here encodes the envelopes $c_e(m,n)$ in a practical way. As described in Eq. (2), $c_e(m,n)$ is a M-fold increase in data, but highly redundant. This redundancy is due to the large amount of overlap between filters in frequency domain, as well as oversampling of the envelope. Since the envelopes are in an auditory domain, we can now reduce redundancy by isolating perceptually salient features.

2.2. The Feldbauer sparse pulse algorithm

In the codec proposed by Feldbauer *et al* in [1], the output of a gammatone filterbank is used to extract pulses modelling auditory nerve behaviour. In order to achieve efficient coding of these auditory pulses, Feldbauer *et al* described an algorithm that classified these pulses as relevant or not, and only encodes the relevant ones [5].

A relevant auditory pulse acts as a masker to other pulses in its vicinity (in time and frequency). To determine if another pulse (the "probe") is being masked, the masking pulse is passed through the synthesis stage and then reanalyzed, giving a unit excitation pattern; the probe is then compared to the excitation pattern (scaled by an "impact factor"). This synthesis-analysis ("transmultiplexing") is a *spreading* of an auditory stimulus and models perceptual masking in both frequency and time domain.

Given this spread excitation pattern, a probe with energy below this scaled excitation pattern is removed. The algorithm begins with the pulse with largest energy and then proceeds to the next largest remaining pulse.

2.3. Relevant envelope sample points

The algorithm we use works on the same principle, but in the absence of auditory pulses considers a subsampling of the envelopes $c_e(m,n)$. Each channel m of $c_e(m,n)$ is subsampled at a rate G_m based on data by Ghitza [6] and the resulting envelope samples $c(m,G_m k)$ are treated in a similar manner as Feldbauer's auditory pulses.

Let $P = \{p_k\}_{k=1,K}$ be the set of all envelope sample points with energy above the absolute threshold of hearing. Each element is a triplet $\{p_{E,k}, p_{m,k}, p_{n,k}\}$ denoting the point's energy, channel and time position respectively. The set is sorted in descending order of energy. Let $R = \{r_l\}_{l=1,L}$ be the set of relevant points, initially empty.

Let $E(p_k, m, n)$ be a function giving the envelope value in channel m at time index n due to an isolated auditory pulse with energy $p_{E,k}$ in channel $p_{m,k}$ at time $p_{n,k}$. Now the following algorithm is repeated until the set P is empty:

1. Move p_1 from P into R.
2. Evaluate each remaining point $p_k, k = 2, \ldots, K$ in P: if $p_{E,k} < \gamma \max(\{E(r_l, p_{m,k}, p_{n,k}\}_{l=1,L})$, remove p_k from P.

The impact factor γ is used to control the bitrate to quality tradeoff. A value of $\gamma > 1$ can be used for aggressive bitrate reduction.

It is important to note a key difference between envelope sample points and auditory pulses. Envelope sample points are at known temporal locations given by multiples of G_m, thus even "discarded points", without being transmitted explicitly to the decoder carry information: the envelope at a point $c(m, G_m k)$ for which no sample point was transmitted does not exceed the excitation pattern from the surrounding sample points. This information can be used in the reconstruction process.

The relevant sampled points of the envelopes are the SF and can be quantised and passed through an entropy coder. Ideally, a vector quantiser should be designed that takes into account the nonlinear properties of loudness perception and the distortion of the envelopes due to the subsampling and relevancy selection. For the purpose of initial evaluation, the envelope magnitudes were quantised on a dB scale with a scalar quantiser.

2.4. Iterative Reconstruction

Given the SF as a set of "relevant" envelope samples, the decoder can construct a copy of $c_e(m,n)$, where some sections are fixed by a transmitted sample point, and other sections have an inferred maximum value. From this set of envelopes, the decoder attempts to reconstruct a signal $\hat{x}(n)$. Below, this partially fixed estimate is denoted $c'_e(m,n)$.

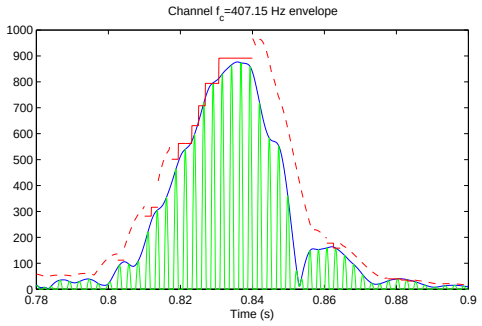

Fig. 2: Fixed quantised envelope regions and "masked" maxima

Figure 2 shows an example of the original envelope $c_e(m,n)$ (dark solid line) with the underlying carrier (real part half-wave, bounded by the envelope). The stepped solid line is the quantised and subsampled *fixed* envelope to be reconstructed, along with the *inferred* maximum that the envelope is allowed to attain during reconstruction (dashed). The discontinuity between the fixed to the inferred sections is primarily due to the impact factor setting, and the fact that the inferred level may be due to data from adjacent channels.

The decoder begins by modulating the envelope with a unit magnitude (complex) carrier signal centered on the channel frequency. These initial estimates of the channel signals $\hat{c}^{(0)}(m,n)$ are then passed through the synthesis filterbank to form the initial signal estimate, $\hat{x}^{(0)}(n)$,

$$\hat{x}^{(k)}(n) = \sum_{m=1}^{M} \sum_{l=0}^{L} c^{(k)}(m, n-l) g_m(l), \qquad (3)$$

where g_m are scaled time-reverse complex conjugates of h_m. The filterbanks are designed such that when passing $c(m,n)$ directly from the analysis to the synthesis filterbank, $x \approx \hat{x}$ within the design bandwidth.

Given this initial estimate of the input signal, the iterative reconstruction is performed by successive iterations of analysis and synthesis. Starting with $\hat{x}^{(0)}(n)$, the kth estimate of the audio signal is analysed to obtain the channel signal $\hat{c}^{(k+1)}(m,n)$ and its carrier estimate $\hat{c}_\phi^{(k+1)}(m,n)$.

We can now apply the envelope correction. Denoting the corrected channel signal $c^{(k+1)}(m,n)$, in the *fixed* regions of c' we set

$$c^{(k+1)}(m,n) = c'_e(m,n)\hat{c}_\phi^{(k+1)}(m,n), \qquad (4)$$

whereas in the *inferred* region, the channel signal estimate is only modified if its envelope exceeds the inferred maximum,

$$c^{(k+1)}(m,n) = \begin{cases} \hat{c}^{(k+1)}(m,n) & \|\hat{c}^{(k+1)}(m,n)\| \le c'_e(m,n) \\ c'_e(m,n)\hat{c}_\phi^{(k+1)}(m,n) & \text{otherwise.} \end{cases} \qquad (5)$$

Using Eq. (3), the next estimate $\hat{x}^{(k+1)}$ is obtained until a fixed limit of iterations is reached.

In the present implementation, data is processed in frames. The iteration reconstructs the estimate using multiple frames at a time, such that for every new frame received, the oldest frame is considered as fully reconstructed and sent to the output.

3. IMPLEMENTATION AND SIMULATION RESULTS

To evaluate performance, the above coding scheme was implemented to process wideband (16 kHz sampling rate) monophonic speech and audio samples. The analysis and synthesis filterbanks consisted of 62 filters with center frequencies from 40.08 Hz to 6930 Hz. The filter with lowest center frequency had an impulse response of 2067 samples, the filter with highest center frequency 78 samples.

The implementation was based on short blocks of only 60 samples. Due to the length of the filters and the auditory pulse envelopes a large lookahead is required. The overall coding delay was 165 ms.

The subsampling of the envelopes was $G_m = 30$ for the lowest 41 channels, down to $G_m = 10$ for the high frequency channels. Overall, each block of 60 samples results in 170 envelope sample points before relevance selection. The decoder iterates over a buffer of three frames, with 20 iterations for every new frame. Thus, the effective number of iterations is 60.

Seven audio files were encoded, taken from the SQAM database [7]. We selected four speech segments and three audio segments: Test file SQAM35 is a glockenspiel, SQAM48 a vocal quartet, SQAM49 to SQAM54 are speech samples, and SQAM60 is a part of a piano concerto. The files were downsampled to 16 kHz and reduced to a single channel.

3.1. Rate of information

Sample	Total Frames	Nonzero Frames	SF per NZ Frame
SQAM35	2360	2322	18.78
SQAM48	2965	2870	40.07
SQAM49	1824	1770	35.93
SQAM50	2041	1983	35.74
SQAM51	1833	1797	35.72
SQAM54	1993	1853	34.23
SQAM60	3513	3318	31.85

Table 1: Average relevant points, $\gamma = 1.0$

Table 1 shows the length of each test file in frames (3.75 ms) and the number of frames that are nonzero. Since the coding algorithm takes the absolute threshold of hearing into account, quiet passages are encoded as "empty" frames. These empty frames were excluded when calculating the averages. The fourth column shows the average number of SF per frame when the impact factor was set to $\gamma = 1.0$. SQAM35 is interesting due to its low number of SF per frame; the sample is a sequence of single tones and thus spectrally sparse.

Figure 3 shows the effect of varying the impact factor on the average number of relevant points per frame, as well as the maximum. These values are

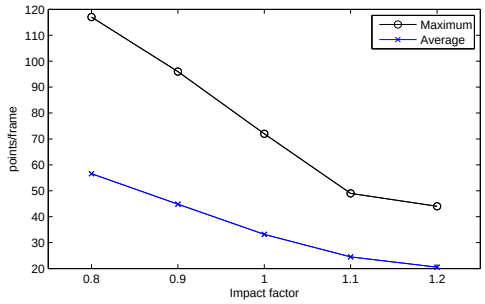

Fig. 3: Effect of impact factor on number of relevant points

for all test files combined. The impact of γ on the rate of SF to be encoded is apparent.

3.2. Quality evaluation

Informal testing of the resulting quality was performed to find reasonable limits of coding and reconstruction parameters before a quantiser design can be considered. In addition to the impact factor γ, the number of iterations during the reconstruction at the decoder was found to be an important factor.

The codec performs reasonably well on speech signals, with few audible distortions for $\gamma < 1$. Distortions were much more apparent with tonal signals, especially SQAM35. The problem of reconstructing tonal signals was already shown in [3], and while simply increasing the number of iterations at the decoder is one solution, the computational requirement becomes a problem.

4. DISCUSSION AND CONCLUSION

We present in this paper a perceptual codec that uses salient features extracted from auditory Hilbert envelopes. The auditory envelopes are subsampled and the resulting envelope samples are evaluated for relevance. The relevant envelope samples may be quantised and transmitted to the decoder.

The decoder reconstructs auditory envelopes based on the transmitted samples, where parts of the envelope are at fixed levels, whereas other parts may vary within bounds inferred from the fixed parts. An iterative algorithm is used to find an audio signal from the reconstructed auditory envelopes.

The codec was tested with wideband speech and audio samples, and showed that a moderate reduction in bitrate should be possible with little impact on perceived quality. A single parameter is used to control quality versus bitrate reduction.

As currently implemented, the codec has very high computational complexity at the decoder, but has a relatively simple encoder. This is opposite to common codecs, and would be prohibitive for portable applications.

However, we show that well-chosen parts of auditory envelopes are sufficient to encode audio signals without perceptual distortions. With processing power even in small devices increasing rapidly, the computational complexity becomes less important. In the short term, the concepts presented here may be useful to improving existing codecs.

There are many areas of further research related to this codec. It is probably possible to further reduce the number of SF to be transmitted, or to increase the quality of the audio by refining the SF selection algorithm. Also, a quantiser needs to be designed that will keep the envelope distortion to a minimum. Finally, it should be possible to improve the decoder such that complexity is comparable to the encoder.

In conclusion, the codec presented here is a novel method to code audio using auditory Hilbert envelopes. This provides a reference for our continued research in auditory envelopes and how they relate to perception. We hope that this will lead to improved perceptual coding methods.

5. REFERENCES

[1] C. Feldbauer, G. Kubin, and W. B. Kleijn, “Anthropomorphic coding of speech and audio: A model inversion approach,” *EURASIP J. Applied Signal Processing*, pp. 1334–1349, Sep 2005.

[2] T. Dau, D. Puschel, and A. Kohlrausch, "A quantitative model of the 'effective' signal processing in the auditory system. I. Model structure," *J. Acoust. Soc. Am*, Jan 1996.

[3] J. Thiemann and P. Kabal, "Reconstructing Audio Signals from Modified Non-Coherent Hilbert Envelopes," in *Proc. Interspeech 2007*, (Antwerpen, Belgium), pp. 534–537, Aug. 2007.

[4] R. D. Patterson and J. Holdsworth, "A functional model of neural activity patterns and auditory images," in *Advances in Speech, Hearing and Language Processing*, vol. 3, pp. 547–558, JAI Press, 1996.

[5] C. Feldbauer and G. Kubin, "How sparse can we make the auditory representation of speech?," in *Proc. Interspeech 2004*, (Jeju Island, Korea), pp. 1997–2000, Oct. 2004.

[6] O. Ghitza, "On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception," *J. Acoust. Soc. Am.*, vol. 110, pp. 1628–1640, Sep. 2001.

[7] EBU, "Sound Quality Assessment Material, Recordings for subjective tests - Users' Handbook for the EBU-SQAM Compact Disc," Tech. Rep. 3253, European Broadcasting Union, 1988.