Perceptual Audio Codec: Block-Switching with Psychoacoustic Modeling

AJ Ferrick [aferrick@stanford.edu]
Gabriele Carotti-Sha [gcarotti@stanford.edu]

ABSTRACT. We have implemented a digital perceptual audio codec that extends the baseline codec programmed throughout the quarter and which approximates the MPEG-I standard. This extended codec maintains the same compression ratios while being perceptually more accurate than the baseline codec and scores higher on ITU-R standardized tests. The codec implements advanced features such as MDCT-analysis/synthesis with block-size switching, masker decimation and critical band modeling, and a rudimentary bit reservoir.

I. BACKGROUND

Baseline Codec

The baseline coder resembles the structure of MPEG-1 [1] without some deviations, such as the PQMF filter bank. A16-bit stereo PCM audio buffer is fed into two parallel stages: frequency space transformations and the psychoacoustic modeling. Frequency space transforms are facilitated by the Modified Discrete Cosine Transform (MDCT), applied to a block of Sine Window PCM samples. The psychoacoustic modeling applies an FFT to the same block of Hanning Window PCM samples. With these frequency samples, masking curves are created and signal-to-mask ratios (SMRs) in each critical frequency band [2] are computed. The SMRs inform the compressor how to allocate bits into different frequency bands as the frequency data is quantized for storage. Under the block floating point quantization scheme, scale and mantissa bit allocation is determined based on the expected energy perceived in each band, represented by the SMRs. Finally, the data block is encoded and decoded according to the computed bit allocation per band.

This codec scheme presents noticeable impairments, however. Whenever abrupt peaks in energy (e.g. sharp transients or musical attacks) occur in the signal, the spread of quantization noise may precede both the peak and its masker. Perceptually, this effect is known as *pre-echo* since artifacts are introduced leading into each intensity spike. Speech reverberation can also be heard for certain signals (such as the German speaker test file, discussed in the Results section). These impairments are very annoying to listeners, enough to motivate removing them from the decompressed signal.

Extended Codec

The extended codec implements a block-switching scheme, which mitigates the effects of energy spreading in the time domain to increase temporal resolution at the cost of frequency resolution. This means reducing the block size from 1024 samples, to 1024 / 8 = 128 samples. This allows for a much finer analysis of transient components.

The psychoacoustic model implements masker decimation in order to reduce computation complexity, while preserving perceptual quality. The overall masking threshold is computed in order to obtain SMRs that inform the subsequent bit allocation.

The baseline codec accepts 16-bit PCM stereo files. When sampled at 44.1 kHz, the target compression ratio of the baseline codec is about 5.3, with a target data rate of 128 kb per second. The extended codec maintains this compression ratio: compression is not sacrificed for perceptual quality.

II. Implementation

The structure for the codec we implemented is depicted below. For a given input file, each channel is encoded and decoded independently — i.e. a stereo signal is treated as a pair of mono signals. We do not take into account any correlations between channels.

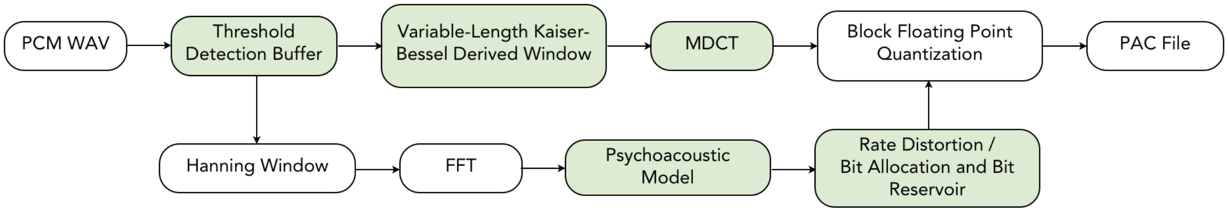
PCM WAV
Threshold Detection Buffer
Variable-Length Kaiser-Bessel Derived Window
MDCT
Block Floating Point Quantization
PAC File
Hanning Window
FFT
Psychoacoustic Model
Rate Distortion / Bit Allocation and Bit Reservoir

Fig 1. Block diagram of the extended codec. Highlighted blocks are additions or extensions.

Block Switching Scheme

Our main point of departure from the baseline coder is the addition of block-switching. Although this feature adds complexity, it is essential in reducing pre-echo effects in sounds that are too sharp to capture in a ~23 ms block (i.e. 1024 samples per block at a sampling rate of 44100 samples per second). We allow the codec to encode transient events in ~2 ms blocks (i.e. 128 samples per block at a sampling rate of 44100 samples per second). Acoustic events with large amounts of high-frequency content (i.e. sudden and rapid changes in the time domain) trigger a block switch. These events include sharp percussive hits or sharp musical attacks with strong overtone series.

Onset Detection

Onset detection is performed through a modified queue data structure. The codec depends on having a queue of data which buffers the next PCM samples it is going to encode. The onset detection operates on this queue, so that it effectively predicts which block sizes will be appropriate for the samples the codec will encounter next. Its effectiveness is displayed with the SQAM castanets in Figures 2 and 3.

The algorithm for detecting onsets is inspired by Bello et al [3]. The onset detection algorithm uses a weighted energy measure which prefers high-frequency content. This energy measure is compared to a threshold value, which varies depending on the current window size — short coding windows have a smaller threshold than long coding windows. If the signal energy exceeds this threshold, an onset is triggered. A state machine paired with the buffer controls how the onset informs which window size is to be used next (shown in Figure 4).

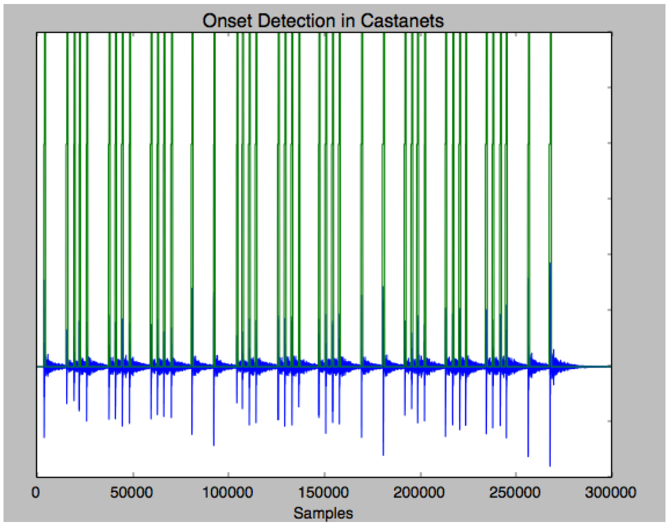

Fig 2. Onset detection on the SQAM Castanets

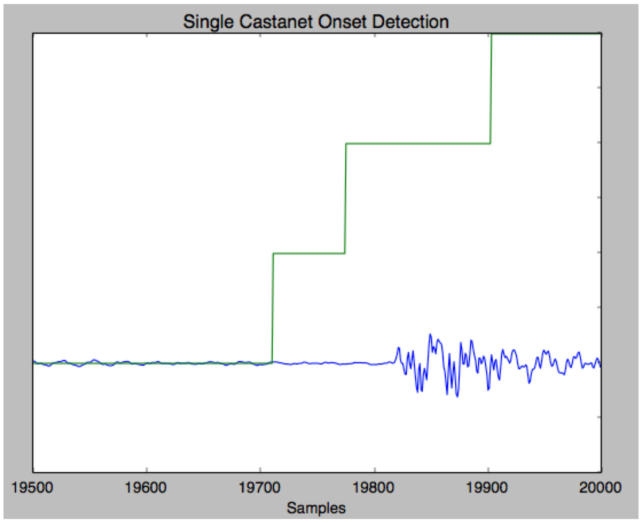

Fig 3. Single onset detection. Second level is the "short" window state.

The codec simply asks this transient-detecting queue for the next block, and the queue returns a list of samples and the current window the codec should use.

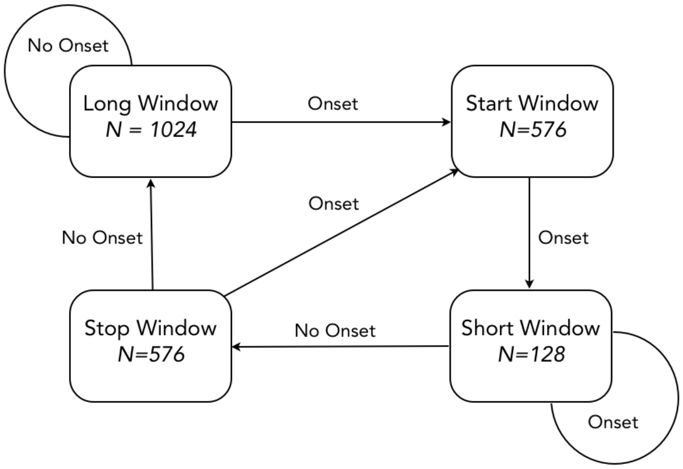

Fig 4. State machine controlling the current window size. It transitions state based on the whether there is an onset in the buffer.

Block Sizes and KBD Windows

Both the baseline and extended codec perform reconstruction with the MDCT using an overlap-and-add procedure. However, because of the different block sizes, two intermediate *transition windows* must be used. These transition windows allow the codec to transition from overlap-and-add with long windows to overlap-and-add with short windows. The length of a transition window is computed as L = (N_{long} + N_{short}) / 2. The extended codec uses 1024 samples for long windows, 128 samples for short windows, and therefore 576 samples for transition windows (illustrated in Figure 5). For a block size of N samples, the MDCT provides N/2 spectral samples. This also means the MDCT kernel's phase offset must be adjusted so that the 50% overlap-and-add will match up the middle of the *right-hand window*, and not simple the middle of the sample. Practically, this means the phase offset term contains either an (N_{long} + 1) / 2 term or an (N_{short} + 1) / 2 term.

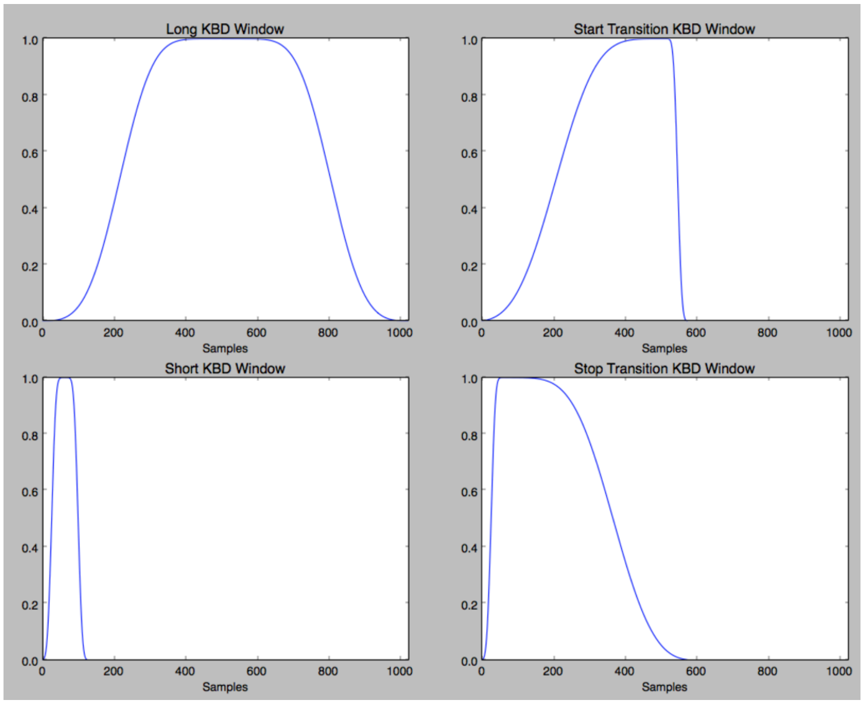

Fig 5. KBD windows for various block sizes.

Before the MDCT takes place, each block of time-domain data is windowed using a Kaiser-Bessel Derived (KBD) window. The KBD window satisfies the perfect reconstruction condition of the MDCT and allows for a parameter, alpha, to control frequency resolution versus side-band rolloff. This is essential to the block switching scheme. Adjacent blocks are overlapped and added with an N/2 offset from each other. Once a transient is detected and the change of state is triggered, a start window must be applied to prevent aliasing from an abrupt change in temporal resolution. This transition window is constructed as the concatenation of the left half of a long window together with the right half of a short window. As a result, subsequent blocks can be overlapped an added while maintaining a shorter size until the state machine reverts back to long blocks. Likewise, a transition window from short to long (a stop window) can be easily constructed as symmetrical with respect to the start window.

In the extended codec, an alpha parameter of 4 is used for the KBD window segments corresponding to long half-blocks, and an alpha parameter of 6 is used for segments corresponding to short half-blocks.

Psychoacoustic Model

Noise and Tonal Maskers

After determining block size and block samples, the psychoacoustic model analyzes the block spectrum in greater detail via the Fast Fourier Transform. It is assumed that both tonal and noise components of a signal produce the perceptual effect of masking; that is, for a given sound pressure level (SPL), energy present at a given frequency will mask the perception of nearby signal components.

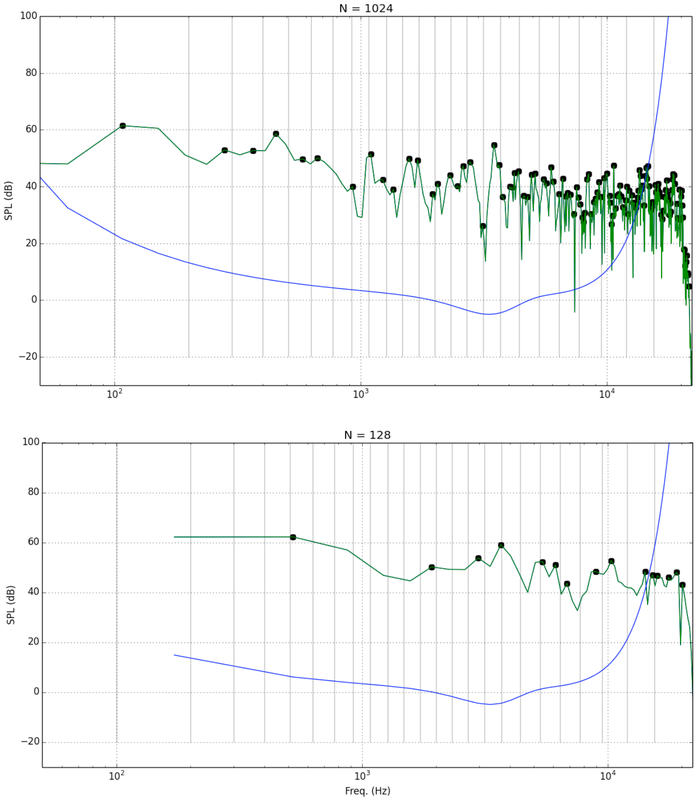

Fig 6. Peak detection for long and short block sizes.

The implemented model is similar to the baseline coder, which in turn is similar to MPEG-I Psychoacoustic Model 1 [1]. A fast peak detection algorithm is applied to determine likely tonal components in the block. These peaks are used to construct a series of tonal maskers. Each masker is admitted to the final masking threshold curve only if the power of its tonal peak exceeds the human threshold of hearing (see Figure 6). The *residual spectrum* is then computed by subtracting these peaks from the original spectrum. Noise maskers are created for each critical band by summing over the energy in the residual spectrum. The maximum value of each masker corresponds to the difference between the SPL value and 15 bark for tonal maskers, 5.5 bark for noise. This means that noise maskers produce a higher threshold.

Masker Decimation

Along with a less permissive peak detection algorithm, the new codec incorporates masker decimation in order to save both computation time and space. Maskers whose center SPL values are close by 0.5 Barks are resolved to just the predominant one. Thus, clusters of potential maskers generated by corresponding peaks are avoided.

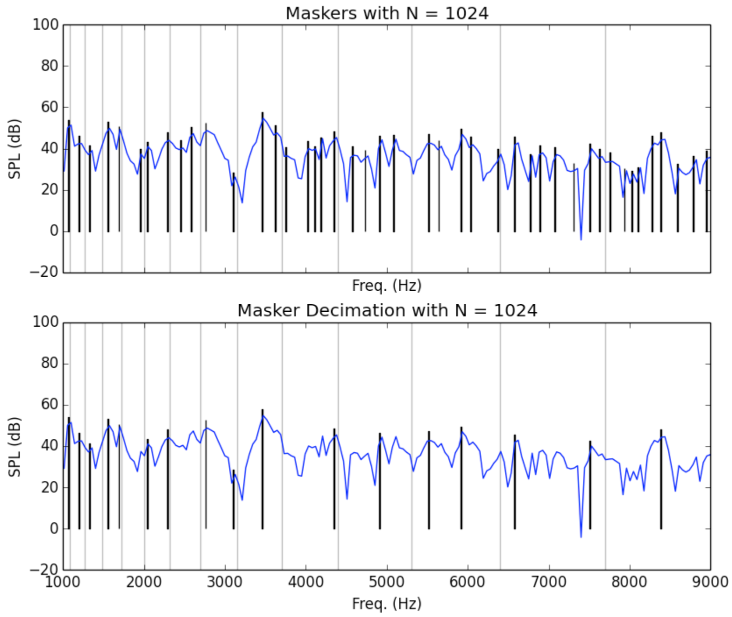

Fig 7. Masker decimation for harpsichord sound sample under a long block of 1024 samples.

It is also assumed that for low frequencies a coder's frequency bands tend to be wide compared to the ear's critical bands, whereas they are narrow compared to the critical bands at higher frequencies. To account for this, SMRs are computed by taking the minimum masking threshold for bands corresponding to low frequencies and the mean threshold value for higher bands. This to ensure that the most sensitive critical band is represented for low frequencies. The overall threshold can be seen in Figure 8.

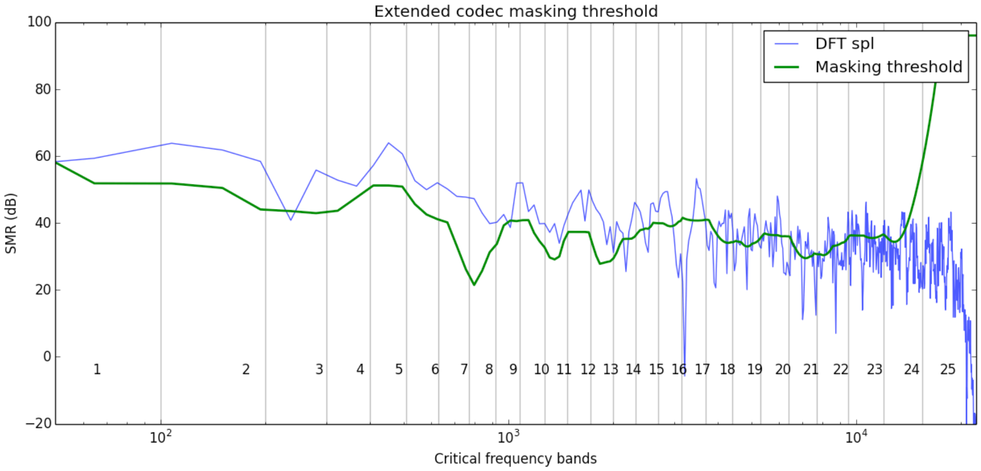

Fig 8. Masking threshold of harpsichord sound sample (long block).

Bit Allocation and Reservoir

Bit Allocation and Rate Distortion

The computed SMRs for each band are then passed to the bit allocation module in the encoder to determine how the available bits will be allocated for this block. The total bit allocation is governed by target data rate, yielding a rate distortion and thus lossy compression. Allocation was achieved through water-filling, by which bands corresponding to the highest SMRs are allocated bits to first. At each iteration, the SMR value in the previously allocated band is decreased by 6.02 dB and the process repeats until the bit budget is exhausted or the maximum number of bits is allocated to a given band (16 bits). The resulting allocation scheme is then applied to the data block processed via MDCT. The final encoding step is the non-uniform midtread quantization of each data block.

Bit Reservoir

One issue encountered was a serious loss of bit resolution for short blocks. After subtracting out the overhead for storing block metadata, the available bits available to short blocks was very small and resulted in very poor allocation — very few frequency samples could be encoded, and those that were suffered from poor resolution. This phenomenon was exemplified most drastically by the SQAM Castanets sample, where rapid successive attacks caused a "thrashing" effect of hearing very little spectral content followed by nearly-full spectral content.

To circumvent this problem, we extended the codec with a naive bit reservoir. The procedure is simple: keep a running tally of any leftover bits not allocated in the long and transition windows. Whenever a short window is encountered, the bit allocation routine may draw bits from this reservoir so that it has a decent resolution. Our coded file format allows for arbitrary data resolution per block, so there was no need to implement pointers within the block header to previous blocks.

The final result of this allocation scheme at our target data rate was that the decompressed signal is effectively lowpass-filtered at around 15500 Hz, which corresponds to the higher critical bands in the SMR vector. Although the difference is perceptible for sound files with strong overtones in these upper spectral bands (e.g. a harpsichord or glockenspiel), this allows for a greater availability of bits at bands containing fundamentals. Since this was not enough to reduce quantization noise, we adopted a simplified bit reservoir technique to increase bit resolution for short windows

III. Results

Effects of the Coder

Transient detection as implemented has proven to be highly accurate for the SQAM Castanets sample, and in general for signals containing sharp and isolated energy peaks. Trickier samples, such as the SQAM Male German Speaker sample, present a challenge due to the absence of overly tonal components, as well as to the presence of essential noise-like sounds such as fricatives, etc. The extended coder was also successful at removing imperceivable spectral content, as shown in Figure 10. However, this rate distortion presented problems for samples high prevalent overtones in high frequencies: there were not enough bits to reconstruct some of these samples at the bit rate while minimizing the perceived impairments.

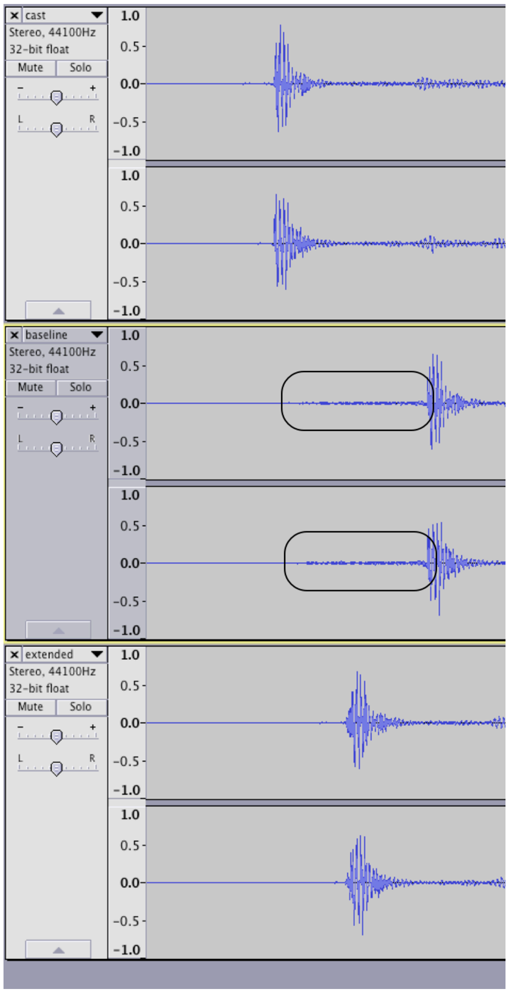

Fig 9. Pre-echo cancellation. The highlighted area shows the pre-echo impairments demonstrated by the baseline coder. The top waveform is the original sample, and the bottom is the reconstruction through the extended coder. No pre-echo is present.

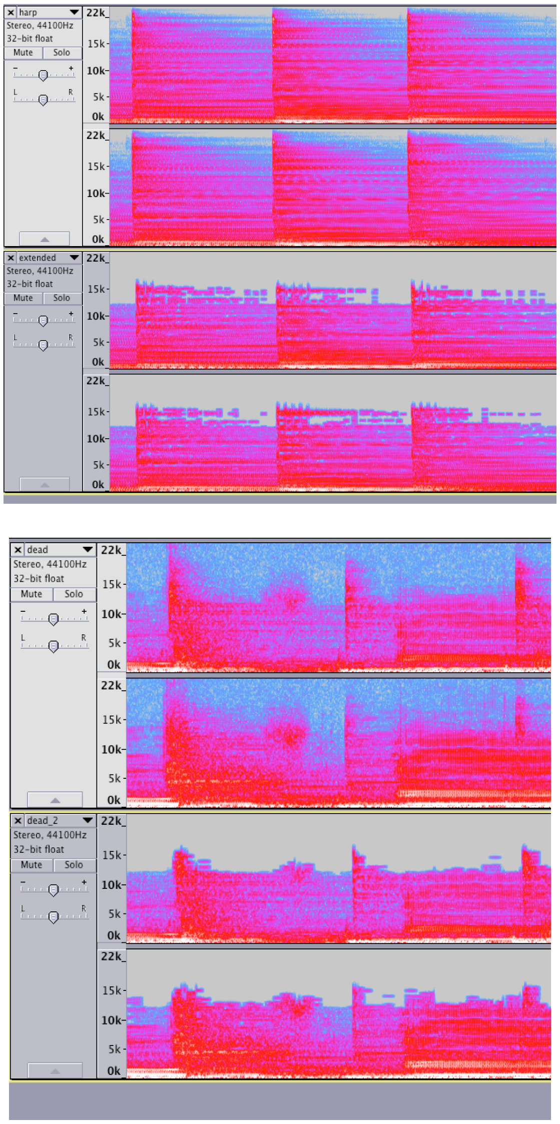

Fig 10. Bit allocations. Note the loss of energy in high frequency bands. Top: SQAM harpsichord. Bottom: "Box of Rain" by The Grateful Dead.

Compression Ratios

We used the following audio samples to obtain a measure of compression and for the subsequent ITU-R testing, discussed below.

Sample Name	Coded File Size	Original Size	Compression Ratio
SQAM Castanets	295,129 bytes	1,590,021	5.4 : 1
SQAM Trumpet	570,092 bytes	3,146,652	5.5 : 1
SQAM Harpsichord	545,698 bytes	2,891,112	5.3 : 1
SQAM German Male Speaker	536,857 bytes	2,949,076	5.5 : 1

ITU-R Listening Tests

The goal for the extended codec was to create a more accurate and pleasant reconstruction scheme while maintaining the same compression ratios. To measure the subjective quality of the codec, we performed standardized double-blind, triple-stimulus ITU-R listening tests with six different listeners. These listeners are both untrained and trained (i.e. studied codec

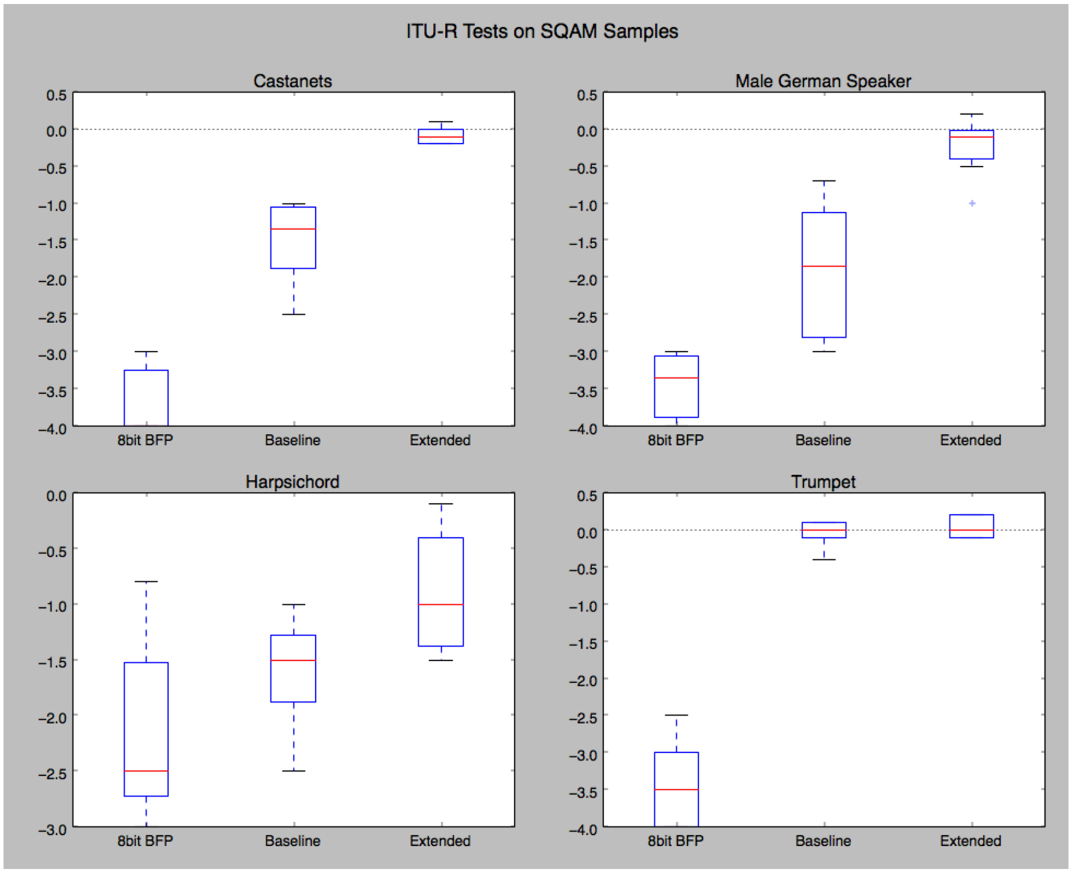

Fig 11. The ITU-R double-blind triple stimulus tests show a marked improvement on every SQAM sample. Notable are the Castanets & German mean and variance improvements.

impairments). The test was administered comparing three different codecs: the baseline codec, the extended codec, and a low-anchoring 8-bit, 1024-block floating point codec. The low anchor has no perceptual bit allocation, and simply served as a way to generate a signal rife with quantization noise, thus providing a low quality codec to compare against the higher-quality codecs.

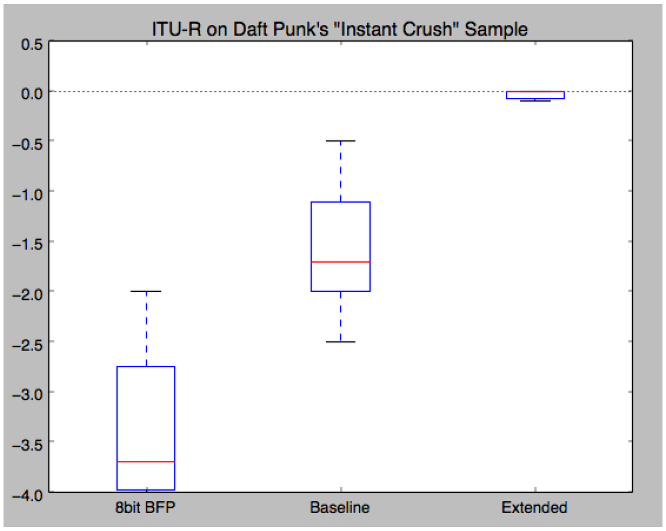

Fig 12. The ITU-R test results for a sample of Daft Punk through the various codecs.

All the samples in the battery demonstrated a marked improvement. In particular, the sample of the castanets improved from a mean score of -1.53 to -0.08 with the extended codec. Similarly, the Male German speaker demonstrated improved quality, raising its mean from -1.9 to -0.25. Additionally, observe the variances in the ITU-R box plot for these two samples: this demonstrates a tighter consensus of quality.

Finally, we ran several popular songs through the codec to observe their results. We performed a similar double-blind triple-stimulus ITU-R test on a 10 second sample of the song “Instant Crush” by Daft Punk to obtain a subjective measure of quality. The results are plotted in Figure 12: While the baseline codec contained noticeable defects, the extended codec scored very highly, with a mean of -0.03.

IV. Conclusions

As illustrated above, the block switching scheme was very effective at highly attenuating, if not altogether removing, the pre-echo impairments suffered by the baseline codec. We were able to achieve this scheme without lowering the compression ratio. The block switching, paired together with the extended psychoacoustic model and rate distortion loop, yielded far better audio quality. This was clearly illustrated in the ITU-R tests, where the listeners reported nearly unanimously that the quality of the extended codec was much higher.

V. Future Work

It would be of interest to compare existing codecs in the market against this extended codec at similar data rates: MPEG-I Layer III, AAC, Dolby AC-3, and Windows Media Audio. Creating this sample bank and administering these tests were outside the scope of this paper, however.

Many improvements can be made to the current codec. The transient detection algorithm suffers in the face of constant high frequency content, yielding many false positives. This detection could be made more accurate by using an adaptive threshold — potentially one that accumulates and rises with high energy content, and over time dissipates back to its original lower threshold (e.g. a "leaky" integrating threshold). Temporal noise shaping algorithms would be useful in providing better control over quantization noise with respect to the masking threshold. This would result in better quality for speech signals as well as aid block switching in a finer representation of transient signals. Additional work could tighten and improve the bit allocation scheme, which currently under-allocates bits for some samples. Further improvements in compression can be achieved by employing a lossless entropy coding scheme, such as Huffman coding. Finally, the codec speed is not at all performant to the point of being able to handle streaming audio.

VI. Thanks

Special thanks to Dr. Marina Bosi for teaching Stanford University's Perceptual Audio Coding course, Tim O'Brien for his advice during the course and the project, and CCRMA for providing facilities and the environment to pursue research in computer audio research.

VII. References

[1] M. Bosi and R. E. Goldberg, “Introduction to Digital Audio Coding and Standards”, Kluwer Academic Publishers, 2003
[2] E. Zwicker and H. Fastl, “Psychoacoustics: Facts and Models”, Springer-Verlag, Berlin Heidelberg 1990
[3] Bello, Juan Pablo et al, "A Tutorial on Onset Detection in Music Signals", IEEE Transactions on Speech and Audio Processing, Vol. 13, No .5, September 2005.