A Full Bandwidth Audio Codec with Low Complexity and Very Low Delay

Jean-Marc Valin, Octasic Inc.
Timothy B. Terriberry, Xiph.Org Foundation
Gregory Maxwell, Juniper Networks Inc.

EUSIPCO 2009
Introduction

- Motivations for very low delay
  - Delay-sensitive applications (e.g. live network music)
  - Reduces perception of acoustic echo

- Codec characteristics
  - Speech and music at 48 kHz
  - 5.3 ms frame size (256 samples), 2.7 ms look-ahead
  - 48-128 kb/s per channel (adaptive)
  - Support for frames sizes of 64 – 512 samples
Overview

- Constrained-Energy Lapped Transform (CELT)
- Basic principles
  - MDCT spectrum divided into critical bands
  - Band energy explicitly coded, constrained at decoder
  - Spectral “details” coded with spherical codebook
  - Bit allocation based on shared information
Encoder Block Diagram

Audio \rightarrow \text{Window} \rightarrow \text{MDCT} \rightarrow \frac{z}{x} \rightarrow \text{PVQ} \rightarrow \text{Coarse energy} \rightarrow + \rightarrow \text{Fine energy} \rightarrow \text{Range coder} \rightarrow \text{Bit-stream}

- Band energy
- Desired bit-rate
- Bit allocation

 Quantizers
Transform, Bands

- **Modified Discrete Cosine Transform (MDCT)**
  - Low-overlap window
  - Divided into critical bands (except low frequencies)

- **Implications of short frame size**
  - Poor frequency resolution and leakage
  - High cost of “side information”
Energy Quantization

- Energy computed for each critical band

Coarse-fine strategy

- Coarse energy quantization
  - Scalar quantization with 6 dB fixed resolution
  - Prediction in time (previous frame) and frequency
  - Range-coded with Laplacian probability model

- Fine energy quantization
  - Variable resolution (based on bit allocation)
  - Not entropy-coded

- Any error in the energy quantization is not compensated in the later quantization stages
PVQ Codebook

- Quantizing $N$-dimensional vectors of unit norm
  - $N-1$ degrees of freedom (hyper-sphere)

- Pyramid Vector Quantizer [Fischer, 1986]
  - Algebraic codebook (no table stored)
  - Combinations of $K$ signed “pulses”
  - Set of vectors $y$ such that $\| y \|_{L1} = K$
  - Mapped onto the hyper-sphere: $x = y / \| y \|_{L2}$

- Fast search and indexing algorithms
- Index is range-coded (flat probability)
Perceptual Improvements

- Pre-echo control
  - Multiple smaller MDCTs, interleaved spectra
  - Energy computed as if a single MDCT

- “Birdie” avoidance
  - Adding an “offset” to PVQ quantization
  - Based on lower part of the spectrum
  - Gain = $N / (N + 6K)$
Bit Allocation

- Fundamentally a CBR codec (VBR supported)
- Synchronized allocator in encoder and decoder
  - Allocates fine energy bits and PVQ bits
  - Depends only on shared information
    - Number of compressed bytes
    - Number of bits used so far by the range coder
  - Near-constant bits per band in time
    - Models within-band masking with near-constant SMR
    - Does not model inter-band masking, tone vs noise
  - Implicit psycho-acoustic model (not coded)
Allocation Example (64 kb/s)
# Evaluation

- **MUSHRA listening tests (10 listeners)**
  - CELT version 0.5.0 (proposed)
  - FhG ULD: warped LPC, pre-filtering
  - G.722.1C: MDCT, scalar quantization, uniform bands

<table>
<thead>
<tr>
<th>Codec</th>
<th>Sample rate kHz</th>
<th>Bitrate kbit/s</th>
<th>Frame size sample (ms)</th>
<th>Look-ahead sample (ms)</th>
<th>Total delay sample (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed (64)</td>
<td>48</td>
<td>64</td>
<td>256 (5.3)</td>
<td>128 (2.7)</td>
<td>384 (8)</td>
</tr>
<tr>
<td>Proposed (96)</td>
<td>48</td>
<td>96</td>
<td>128 (2.7)</td>
<td>64 (1.3)</td>
<td>192 (4)</td>
</tr>
<tr>
<td>ULD</td>
<td>48</td>
<td>96</td>
<td>128 (2.7)</td>
<td>128 (2.7)</td>
<td>256 (5.3)</td>
</tr>
<tr>
<td>G.722.1C</td>
<td>32</td>
<td>48</td>
<td>640 (20)</td>
<td>640 (20)</td>
<td>1280 (40)</td>
</tr>
</tbody>
</table>
Complexity and RAM

- Complexity (encoder+decoder average)
  - 17 WMOPS in fixed-point
  - 27 MHz on Intel Core2 (unoptimised floating-point C)

- State data (per channel)
  - Encoder: 0.5 kB
  - Decoder: 0.5 kB (+ 4 kB for PLC)

- Scratch space
  - Encoder+decoder: ~7 kB
• Conclusion

  • Low-delay coded, explicit energy constraint
  • Work in progress
    • Pitch prediction
    • Stereo coupling
  • Submitted to IETF as Internet codec proposal
  • Resources
    • Source code: http://www.celt-codec.org
    • Mailing list: celt-dev@xiph.org
Questions?

Ask me for audio samples after the session
Other Frame Sizes

Overhead is about 42 bits/frame