Audio and Speech
Digital sound

amplifier  anti-aliasing  filter  codec

1mV  G.7xx  packet-ization
Digital audio

- sample each audio channel and quantize \(\leftrightarrow\) pulse-code modulation (PCM)
- Nyquist bound: need to sample at twice (+ \(\varepsilon\)) the maximum signal frequency
- analog telephony: 300 Hz – 3400 Hz \(\leftrightarrow\) 8 kHz sampling \(\rightarrow\) 8 bits/sample, 64 kb/s
- FM radio: 15 kHz
- audio CD: 44,100 Hz sampling, 16 bits/sample (based on video equipment used for early recordings)
- more bits \(\leftrightarrow\) more dynamic range, lower distortion
- audio highly redundant \(\leftrightarrow\) compression
- almost all codecs fixed rate
# Audio coding

<table>
<thead>
<tr>
<th>application</th>
<th>frequency</th>
<th>sampling</th>
<th>AD/DA bits</th>
<th>application</th>
</tr>
</thead>
<tbody>
<tr>
<td>telephone</td>
<td>300-3400 Hz</td>
<td>8 kHz</td>
<td>12–13</td>
<td>PSTN</td>
</tr>
<tr>
<td>wide band</td>
<td>50-7000 Hz</td>
<td>16 kHz</td>
<td>14–15</td>
<td>conferencing</td>
</tr>
<tr>
<td>high-quality</td>
<td>30-15000 Hz</td>
<td>32 kHz</td>
<td>16</td>
<td>FM, TV</td>
</tr>
<tr>
<td></td>
<td>20-20000 Hz</td>
<td>44.1 kHz</td>
<td>16</td>
<td>CD</td>
</tr>
<tr>
<td></td>
<td>10-22000 Hz</td>
<td>48 kHz</td>
<td>≤ 24</td>
<td>pro-audio</td>
</tr>
</tbody>
</table>
Digital audio: sampling

distortion: signal-to-(quantization) noise ratio
Digital audio: compression

Alternatives for compression:

- companding: non-linear quantization \(\mu\)-law (G.711)
- waveform: exploit statistical correlation between samples
- model: model voice, extract parameters (e.g., pitch)
- subband: split signal into bands (e.g., 32) and code individually MPEG audio coding

Newer codings: make use of \textit{masking properties} of human ear
Judging a codec

- bitrate
- quality
- delay: algorithmic delay, processing
- robustness to loss
- complexity: MIPS, floating vs. fixed point, encode vs. decode
- tandem performance
- can the codec be *embedded*?
- non-speech performance: music, voiceband data, fax, tones, ...
Quality metrics

- speech vs. music
- communications vs. toll quality
- mean opinion score (MOS) and degradation MOS

<table>
<thead>
<tr>
<th>score</th>
<th>MOS</th>
<th>DMOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>excellent</td>
<td>inaudible</td>
</tr>
<tr>
<td>4</td>
<td>good, toll quality</td>
<td>audible, but not annoying</td>
</tr>
<tr>
<td>3</td>
<td>fair</td>
<td>slightly annoying</td>
</tr>
<tr>
<td>2</td>
<td>poor</td>
<td>annoying</td>
</tr>
<tr>
<td>1</td>
<td>bad</td>
<td>very annoying</td>
</tr>
</tbody>
</table>

- diagnostic rhyme test (DRT) for low-rate codecs (96 pairs like “dune” vs. “tune”)
  - 90% = toll quality

August 13, 2001
Companding: $\mu$-law for G.711 (“PCMU”)
Silence detection (VAD)

- avoid transmitting silence during sentence pauses and/or other person talking
- detect silence based on energy, sound
- hangover – unvoiced segments at end of words
- conferencing!
- comfort noise – white noise, shaped noise with periodic updates
- transmit update (4 byte) when things change
Audio silence detection

- needed in conferences to avoid drowning in fan noise
- also reduces data rate
- in use in transoceanic telephony since 1950’s (TASI: time-assigned speech interpolation)
- use energy estimate ($\mu$-law already close) or spectral properties (difficult)
- difficulty: background noise, levels vary
- vary noise threshold: threshold = running average + hysteresis
- if above threshold, increase running average by one for each block
- if below threshold, update running average
- speech has soft (unvoiced) beginnings and endings $\Rightarrow$ hang-over, pre-talkspurt burst
Speech codecs

- waveform codecs exploit sample correlation: 24-32 kb/s
- linear predictive (vocoder) on frames of 10–30 ms (stationary): remove correlation $\rightarrow$ error is white noise
- vector quantization
- hybrid, analysis-by-synthesis
- entropy coding: frequent values have shorter codes
- runlength coding
## Digital audio: compression

<table>
<thead>
<tr>
<th>Coding</th>
<th>kb/s</th>
<th>MOS</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>LPC-10</td>
<td>2.4</td>
<td>2.3</td>
<td>robotic, secure telephone</td>
</tr>
<tr>
<td>G.723.1</td>
<td>5.3/6.3</td>
<td>3.8</td>
<td>videotelephony (room for video)</td>
</tr>
<tr>
<td>GSM HR</td>
<td>5.6</td>
<td>3.5</td>
<td>GSM 2.5G networks</td>
</tr>
<tr>
<td>IS 641</td>
<td>7.4</td>
<td>4.0</td>
<td>TDMA (N. America) mobile (new)</td>
</tr>
<tr>
<td>IS 54/136</td>
<td>7.95</td>
<td>3.5</td>
<td>TDMA (N. America) mobile (old)</td>
</tr>
<tr>
<td>G.729</td>
<td>8.0</td>
<td>4.0</td>
<td>mobile telephony</td>
</tr>
<tr>
<td>GSM EFR</td>
<td>12.2</td>
<td>4.0</td>
<td>GSM 2.5G</td>
</tr>
<tr>
<td>GSM</td>
<td>13.0</td>
<td>3.5</td>
<td>European mobile phone</td>
</tr>
<tr>
<td>G.728</td>
<td>16.0</td>
<td>4.0</td>
<td>low-delay</td>
</tr>
<tr>
<td>G.726</td>
<td>16-40</td>
<td>4.0</td>
<td>low-complexity (ADPCM)</td>
</tr>
<tr>
<td>DVI</td>
<td>32.0</td>
<td>4.1</td>
<td>low-complexity (ADPCM)</td>
</tr>
<tr>
<td>G.722</td>
<td>64.0</td>
<td>4.0</td>
<td>toll-quality (Intel, Microsoft)</td>
</tr>
<tr>
<td>G.711</td>
<td>64.0</td>
<td>4.5</td>
<td>7 kHz codec (subband)</td>
</tr>
<tr>
<td>MPEG L3</td>
<td>56-128.0</td>
<td>N/A</td>
<td>CD stereo</td>
</tr>
<tr>
<td>16 bit/44.1 kHz</td>
<td>1411</td>
<td></td>
<td>compact disc</td>
</tr>
</tbody>
</table>
Distortion measures

- SNR *not* a good measure of perceptual quality
- segmental SNR: time-averaged blocks (say, 16 ms)
- frequency weighting
- subjective measures:
  - A-B preference
  - subjective SNR: comparison with additive noise
  - MOS (mean opinion score of 1-5), DRT, DAM, …
MOS vs. packet loss

G.711 Bernoulli (10ms)
G.711 Bursty (10ms)
G.729 Bursty ($p_a=30\%$, 20ms)
Objective speech quality measurements

- approximate human perception of noise and other distortions
- distortion due to encoding and packet loss (gaps, interpolation of decoder)
- examples: PSQM (P.861), PESQ (P.862), MNB, EMBSD – compare reference signal to distorted signal
- either generate MOS scores or distance metrics
- much cheaper than subjective tests
- only for telephone-quality audio so far
Objectice quality measures

**PSQM:** perceptual distance; can’t handle delay offset

**PESQ:** MOS scores; automatically detects and compensates for time-varying delay offsets between reference and degraded signal

- time-frequency mapping (FFT)
- frequency warping from Hertz scale to critical band domain (Bark spectrum)
- calculate noise disturbance as the difference of compressed loudness (Sone) intensity in each band between the two signals, with threshold masking
- asymmetry modeling (addition of an unrelated frequency component is worse than omission of a component of the reference signal)
Objective vs. Subjective MOS

Objective MOS tools don’t always handle loss impairments correctly:
Audio traffic models

talkspurt: constant bit rate: one packet every 20...100 ms ➤ mean: 1.67 s

silence period: usually none (maybe transmit background noise value) ➤ 1.34 s

➤ for telephone conversation, both roughly exponentially distributed

• double talk for “hand-off”

• may vary between conversations... ➤ only in aggregate
Multiplexing traffic

In a diff-serv buffer, with \( R = 0.5 = \text{reserved/peak} \):

**Effect of \( N \) (multiplexing factor) and \( R \) (token rate) on \( p_o \)**

- \( N = 5 \)
- \( N = 30 \)
- \( N = 100 \)

G.729B: about 42-43% silence
References


See also http://www.cs.columbia.edu/~hgs/audio