Practice in Data Compression
Lectures 13: Audio data coding

Boris Kudryashov

UT (Estonia),
ITMO University (St.-Petersburg, Russia)
The goal is to provide CD-quality audio over telecommunications networks. Almost all CD audio coders are based on the so-called psychoacoustic model (PAM) of the human auditory system. PAM allows to neglect the parts of the signal that the human cannot perceive and calculate the amount of quantization noise that is inaudible. Audio coders work in the range from 20Hz to 20 kHz. The human auditory system has detection capability with a dynamic range of over 120 dB from very quiet to very loud sounds. The absolute threshold of hearing characterize the amount of energy in a pure tone such that it can be detected by listener in noiseless environment.
Louder sounds mask or hide weaker ones. PAM takes into account this masking property that occurs whenever the presence of strong audio signal makes a temporal or spectral neighborhood of weaker audio signals imperceptible. The noise detection threshold is a modified version of the absolute threshold with its shape determined by the input signal at any given time.
Observations:

- The cochlea can be viewed as a bank of highly overlapping bandpass filters. The filters passbands are of non-uniform bandwidth, and the bandwidths increase with increasing frequency.

- Dependence of resolution on frequency for human auditory system can be expressed in terms of critical-bandwidths. A critical band is a range of frequencies over which the masking SNR remains more or less constant. Critical bandwidths are less than 100 Hz for the lowest audible frequencies and more than 4 kHz at the highest.

- Noise and tone have different masking properties.
• If $B$ is the critical-band number then:

Tone masking noise $= E_N = E_T - (14.5 + B)$, dB
Noise masking tone $= E_T = E_N - K$, dB

where $E_N$ and $E_T$ are noise and tone energy, respectively, $K$ is constant in the range of 3-6 dB.

• Speech and audio signals are neither pure tones nor pure noise but rather mixture of both.

• Typically, each frame contains a collection of both masker types. Individual masking thresholds are combined to form a global masking threshold.

• **Inter-band masking** means that masker centered within one critical band has effect on thresholds in other critical subbands. This effect is known as spread of masking and is modeled as a spreading function.
Cochlea Filter Modeling → Threshold Estimation → Absolute Threshold
Masking thresholds for subbands
Signal-to-mask ratio

Tonality estimate

Short Term Signal Spectrum → Tonality Calculation → Estimated Psychoacoustic Threshold
• Using short-term spectrum we classify frequency components of each subband as tone or noise
• For tone-like and noise-like components individual masking thresholds are computed
• The individual masking thresholds and $T_q$ are combined to form a global threshold and minimum over subband components is computed
• Signal-to-mask ratio for subband is computed
• Mask-to-noise ratio for subband is computed
• Quantization step for each subband is computed
Perceptual audio coder and decoder

PCMs audio input

Coding filter bank (computing transform coefficients)

Bit allocation, Quantization and coding

Bitstream formatting

Psychoacoustic model

encoded bitstream

Bitstream unpacking

Transform coefficient reconstruction

Inverse transform

Decoded PCM audio

Encoded bitstream

Boris Kudryashov, St. Petersburg-Tartu, 2018, Page 10 of 22
In audio standards the audio stream passes through a filter bank that divides the input into multiple subbands of frequency. (subband coding). This transform is time-overlapped.

The input audio stream simultaneously passes through a PAM module that determines the ratio of the signal energy to the masking threshold for each subband.

The quantization and coding block uses PAM data to determine bit allocation and quantize transform coefficients. Usually scalar uniform quantization of transform coefficients is used. Quantized coefficients are coded by the Huffman code.
ISO Motion Pictures Experts Group (ISO-MPEG/audio) is one part of three part compression standard that includes video and systems.

- **The original MPEG** (sometimes it is referred as MPEG-1) was created for mono sound systems and had three layers, each providing greater compression ratio.
- **MPEG-2** was created to provide stereo and multichannel audio capability.
- MPEG advanced audio coder (MPEG-AAC) has the same quality as MPEG-2 but at half the bit rate.
- **MPEG-4** audio is a complete toolbox to do everything from low bit rate speech coding to high quality audio coding or music synthesis. Contains HE AAC and HE AAC v2 modes.
• **Layer I**, the simplest provides bit rates above 128 kb/s per channel (compression ratios 3-4).

• **Layer II**, has intermediate complexity and provides bit rates around 128 kb/s per channel (compression ratios 5-6). Main applications: storage video on CD-ROM and transmitting of audio information over ISDN channels.

• **Layer III** is the most complex and provides bit rates around 64 kb/s per channel (compression ratios 10-12). It is used for low-rate compression systems (famous MP3).
Subband decomposition

\[ \hat{x}(n) = g_{M}(n) + \sum_{i=0}^{M-1} g_{i}(n - 2^i) \]

\[ M \text{ is number of subbands, } h_{i}(n) \text{ and } g_{i}(n) \text{ are analysis and synthesis bandpass filters.} \]
LOT-based subband decomposition

\[ N = L \times M \]

\[ w_a \rightarrow \text{MDCT} \rightarrow M \]

\[ w_a \rightarrow \text{MDCT} \rightarrow M \]

\[ w_a \rightarrow \text{MDCT} \rightarrow M \]

\( w_a \) is analysis transform window
LOT-based subband synthesis

\[ N = L \times M \]

\( w_a \) is synthesis transform window
The transform matrix $A$ of size $N \times M$ can be written as

$$A = YT$$

where $T$ of size $M \times M$ is the matrix of DCT-IV transform with entries

$$t_{kn} = \sqrt{\frac{2}{M}} \cos \left( \left( k + \frac{1}{2} \right) \left( n + \frac{1}{2} \right) \frac{\pi}{M} \right)$$

$k, n = 0, 1, \ldots, M - 1$, matrix $Y$ describes the preprocessing step. It has the following form

$$Y = [Y_0 \mid Y_1 \mid -Y_0 \mid -Y_1 \mid Y_0 \mid Y_1 \ldots ]^T$$

where submatrices $Y_0$ and $Y_1$ of size $M \times M$ in turn consist of submatrices of size $M/2 \times M/2$:

$$Y_0 = \begin{pmatrix} 0 & 0 \\ I & -J \end{pmatrix} \quad Y_1 = \begin{pmatrix} -J & -I \\ 0 & 0 \end{pmatrix}$$
where \( I \) and \( J \) are \( M/2 \times M/2 \) identity matrix and contra-identity matrices, respectively. For \( M = 3 \)

\[
I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad J = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}
\]

The analysis and synthesis windows can be chosen to be equal, \( w_a(n) = w_s(n) = w(n) \). The requirement of the perfect reconstruction:

\[
\sum_{l=0}^{L-2s-1} w(n + lM)w(n + lM + 2sM) = \begin{cases} 1, & s = 0 \\ 0, & s \neq 0 \end{cases}
\]

\( n = 0, 1, \ldots, M/2 - 1. \)
Often (e.g. HE-AAC) the MDCT with $N = 2M$ is used, $L = 2$. Perfect reconstruction condition

$$w^2(n) + w^2(n + M) = 1, \quad n = 0, \ldots, M/2 - 1.$$

The example of the perfect-reconstruction window

$$w(n) = \sin \left( \frac{\pi}{2N} \left( n + \frac{1}{2} \right) \right), \quad n = 0, \ldots, N - 1.$$

For windows used in standard implementations, perfect reconstruction requirement is not satisfied. More important requirements are related with transfer functions of band-pass filters.
Sine window

\[ w(n) = \sin \left( \frac{\pi}{2N} \left( n + \frac{1}{2} \right) \right), \quad n = 0, \ldots, N - 1. \]

Hamming window

\[ w(n) = 0.54 - 0.46 \cos \left( \frac{2\pi n}{N} \right); \quad n = 0, \ldots, N - 1. \]

Kaiser window \((l_0(\cdot)\) is the 0-order Bessel function, \(\beta = 5\) \)

\[ w(n) = \frac{l_0 \left( \beta \sqrt{1 \left( \frac{n - N/2}{N/2} \right)^2} \right)}{l_0(\beta)} \quad n = 0, \ldots, N - 1. \]

Vorbis window

\[ w(n) = \sin \left( \frac{\pi}{2} \sin^2 \frac{\pi(i + 1/2)}{n} \right) \quad n = 0, \ldots, N - 1. \]
Several ways to save bits:

- **Intensity stereo** (scaled sum of left and right channels)
- **Middle-side (MS) stereo** (sum and difference of left and right channels)
- **Parametric stereo** (stereo is synthesized from mono signal by using parameters: inter-channel phase difference, inter-channel correlation and so on.)

HE-AAC combines **Spectral Band Replication** (high-frequency part is properly scaled low-frequency part shifted to the high-frequency region) and Parametric stereo