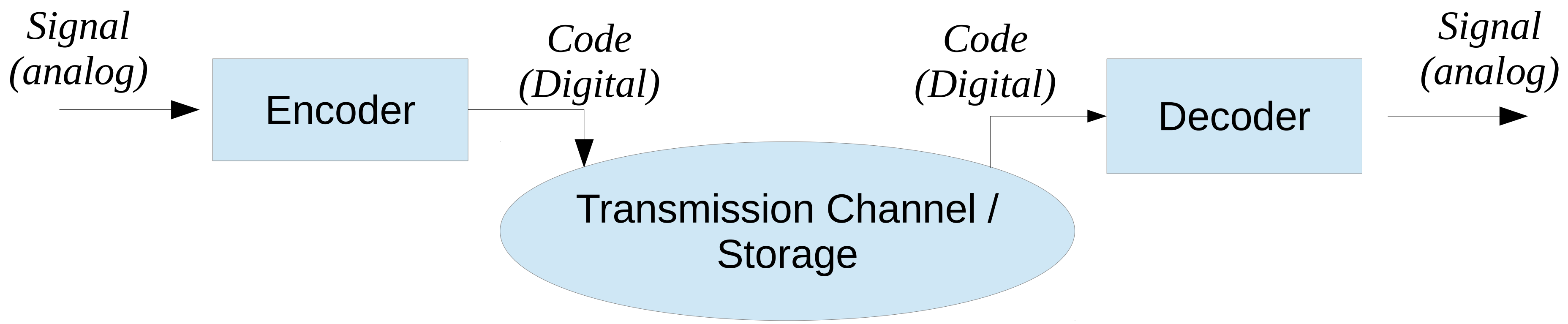
ELL 788
Computational Perception & Cognition

July – November 2015

Module 11

Audio Engineering: Perceptual coding

Coding and decoding



Coding

- For digitizing (recording / transmission)
- Signal to be “satisfactorily” reproduced
- Compact (Compression)
- Time-efficient processing

- *Aims at reproducing an **audio event***
  - *NOT reproducing the original signal*

Principle – perceptual coding

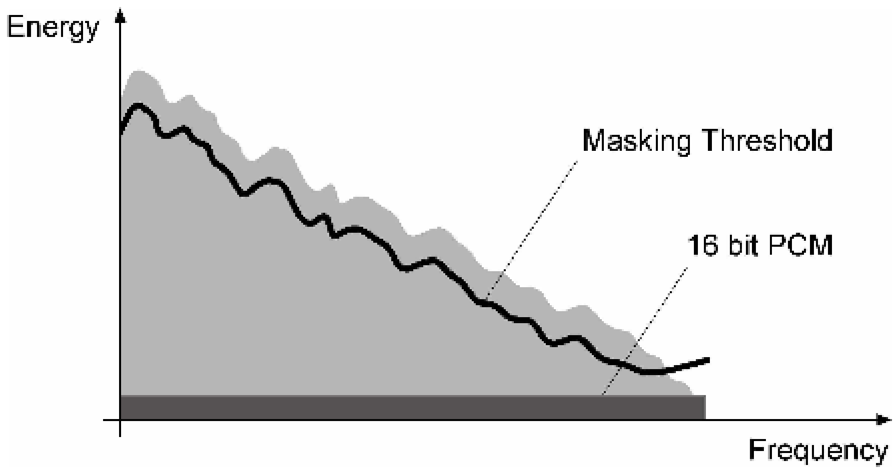


Figure 1: Spectrum and Masking Threshold

Shape permissible quantization noise in frequency domain

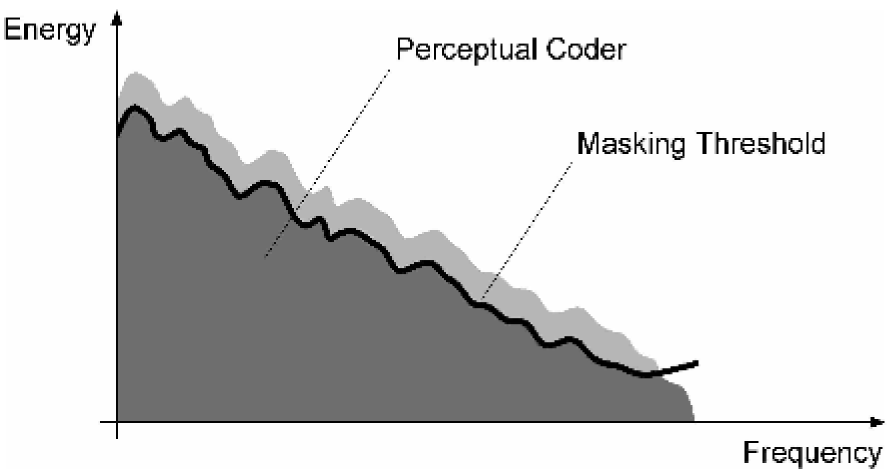


Figure 2: Ideal Perceptual Coding

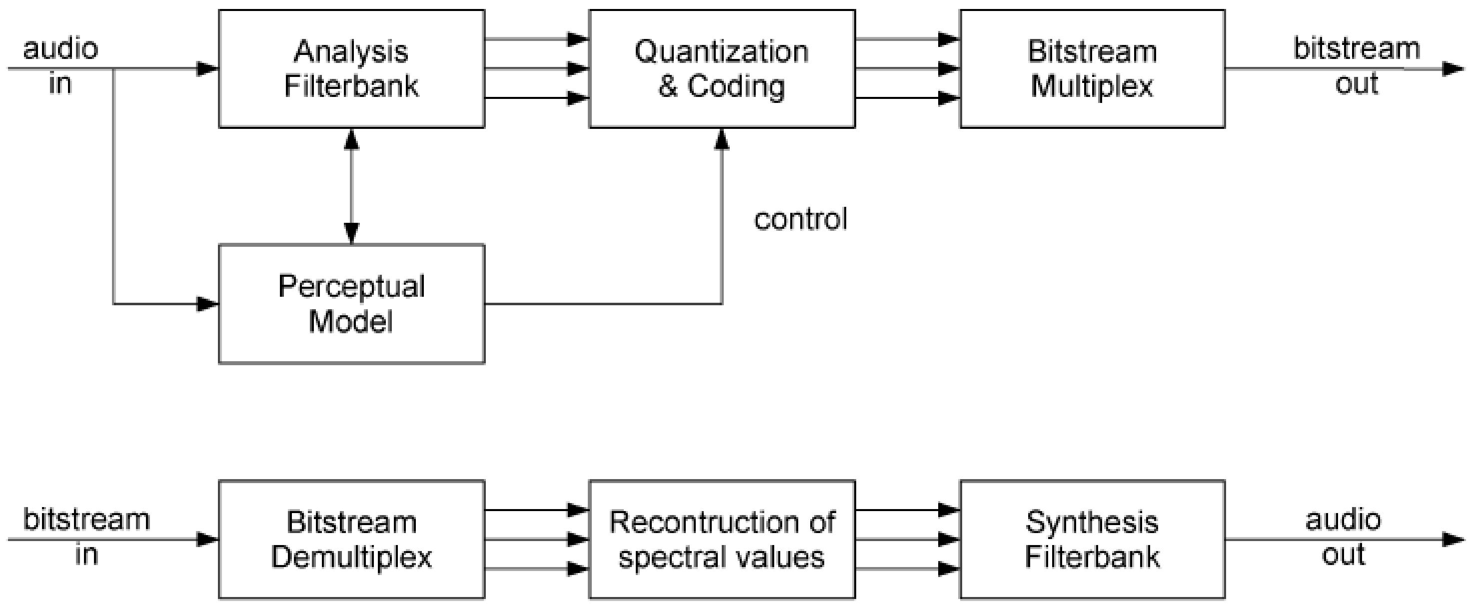
Source: Dietz, et al.

Approach

- Use the psycho-accoustic model
  - Cochlear time-frequency resolution
    - *Discard frequencies that we cannot hear*
  - Auditory masking (Simultaneous and Temporal)
    - *Allow noise to the extent that cannot be perceived*

- Stereo-coding issue – just two channels do not help
  - Not optimal
  - Localization problem (Random uncorrelated noise)
- Sophisticated methods vs. Commercial systems
  - Commodity encoder / decoder

Time-frequency based audio coding



Source: Brandenburg 2013

- <u>Analysis filter-banks</u>: decompose the input signal into a time sequence of vectors with as many elements as spectral components
- <u>Perceptual model</u>: actual (time- and frequency-dependent) masking threshold is computed using rules derived from psychoacoustics
- <u>Quantization and coding</u>: goal is to keep the resulting error signal (usually referred to as quantization noise) below the masking threshold
- <u>Bit stream multiplex</u>: quantized and coded spectral coefficients and some side information + metadata

Sub-band coding

- Modified DCT (MDCT) used to realize polyphase filter-banks
  - DCT is like DFT; uses only cosine functions
  - MDCT operates on overlapping blocks – avoids artefacts arising out of block boundaries
  - High frequency resolution – poor time resolution
- A set of time samples $\rightarrow$ a vector of spectral energy components
  - These spectral components are orthogonal -- Avoids redundancy
- Retain only that human ear will recognize
  - Filter-banks cover range of human auditory system
  - Apply masking in close frequencies

Key for perceptual compression

Quantization and coding

- The energy co-efficients at different frequencies need to be quantized and coded
- Quantization error introduces noise
  - Should be below hearing threshold – *varies with frequency*
- Each quantized level is represented with a symbol
  - Entropy coding: Allocate less bits for frequently used symbols
    - *e.g. Huffman coding*
- Arithmetic coding, or range coding
  - Represents entire signal segment with one number
  - More optimal than Entropy coding

Fixed rate and variable rate coding

- Bit-rate depends on
  - Filter banks
  - Quantization
  - Coding
- Entropy coding leads to variable rate coding
- Fixed rate transmission is often desirable
  - VBR coding can still be used if delay is allowed
    - Typically in one-way transmission, e.g. music system, broadcasting, etc.
    - *NOT for telephony*
  - Buffer control is important

Perceptual audio codecs: MP3

- Sampling frequencies: 32, 44.1 and 48 kHz Sampling frequencies
  - Frame-length 24 ms @ 48 kHz sampling frequency → 1152 time samples
- Auditory masking to remove redundant signal
  - MDCT for coding
  - 2 filter-bank blocks with 576 sub-bands each
  - Quantization based on power law (loudness)
- Huffman coding: a set of predefined tables
- Bit-rate: 32 – 320 kbps, or VBR
- High encoding complexity – but low decoding complexity
- Bit-reservoir and back frames
  - Bits preserved during coding of silent frames can be used to encode audio frames with more audio content
- 2-channel stereo

Quality

- Trade-off between bit-rate and the sound quality

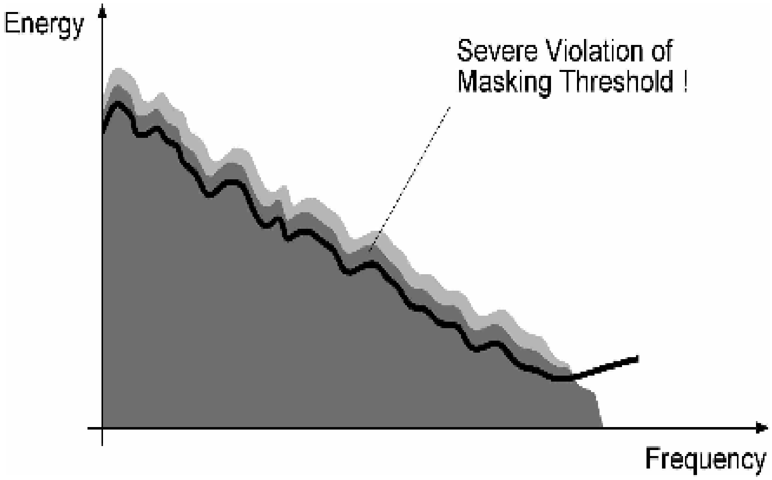


Figure 4: Waveform coding beyond its limits

- Depends on nature of audio being compressed
- Random sound is difficult to compress
  - Compression artifacts

Advance Audio Coding (AAC) Family

- Improved audio quality for coding at low bit rates
  - < 48 kb/s per channel
- Spectral Band Replication (SBR) for bandwidth extension
  - Low frequency components of signal are transmitted
  - High frequency components are reconstructed
  - Results in major savings in bit-rate
- 5.1 channel stereo

Sprectral Band Reconstruction (SBR)

- Based on the principles
  - Generally, there is a large dependency between the lower and higher frequency parts of an audio signal.
  - The psychoacoustic part of the human brain tends to analyse higher frequencies with less accuracy
- Encoder codes only the low and mid frequencies
  - The high frequency part is can be efficiently reconstructed from the low frequency part
    - Some low bit rate SBR control data is needed

Hi-frequency reconstruction using SBR

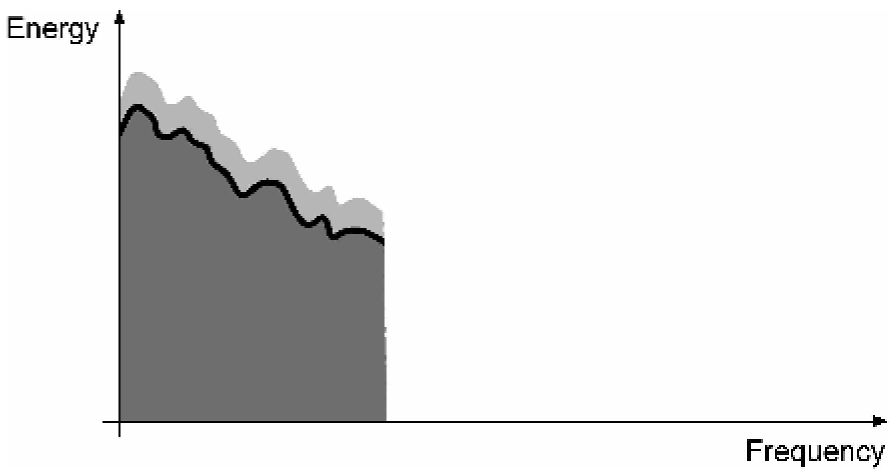


Figure 5: Limiting the audio bandwidth

Spectral envelop and some other control information about the high-frequency components are extracted while coding

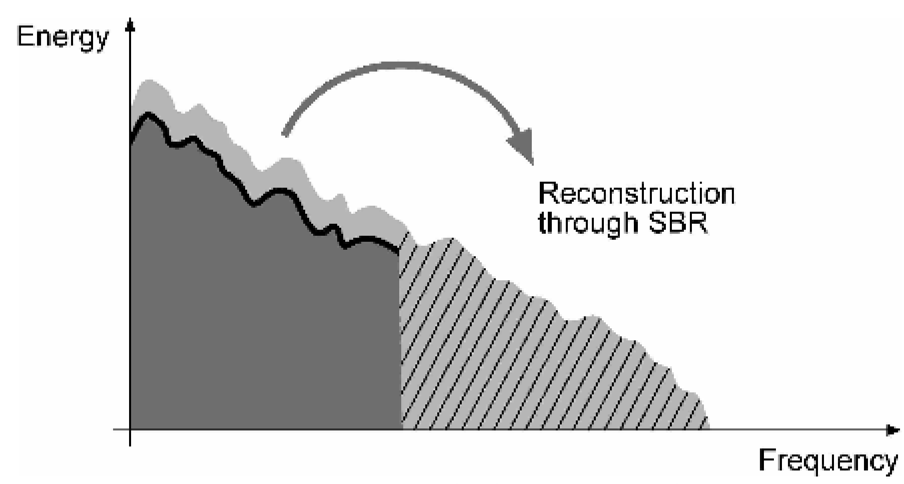
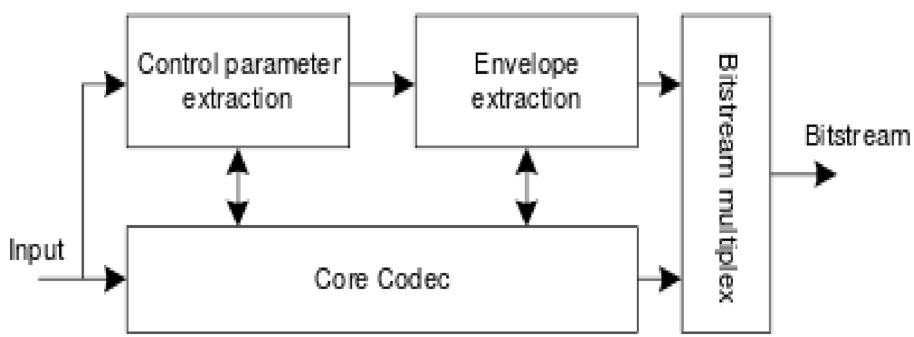


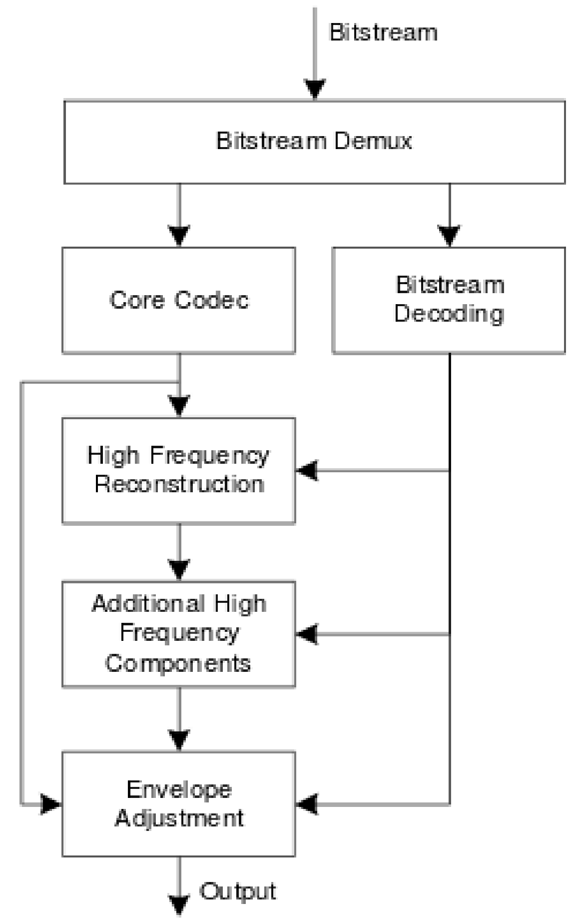
Figure 7: Spectrum after high frequency adjustment

Source: Dietz, et al.

SBR encoder and decoder
(Basic block diagrams)



Encoder



Decoder

Delays

Codec	Typical Delay (HW or DSP)	Typical Application
MP3	140 ms	Music Player
AAC	210 ms	Music Player, Broadcasting
HE-AAC	360 ms	Mobile Music, Satellite Radio

Not suitable for 2-way / multi-party applications
- Telephony, A/V conferencing, Telepresence
- **Requirement: max delay ~20 ms**

Codecs with low delay: AAC-LD

- Uses shorter transform size
  - 480 or 512 subband MDC (vs. 1024 for AAC)
- Avoid look-ahead
  - Bit reservoir is avoided / restricted to only a small number of bits (around 100)
- Less compression
  - Can be used over normal telephone lines / ISDN
  - 32 / 64 kbps or higher
- The achieved coding quality scales up with bitrate
  - The audio quality of AAC LD is slightly better compared to MP3 at the same bitrate.

Spatial audio-coding
(Channel-based spatial coding: stereo / 5.1)

- Encoding each channel separately is suboptimal
  - Also, quality is poor because of uncorrelated noise
- The multi-channel signals are coded as a “downmix”
  - All signal components (disregarding spatial aspects) present in all channels
  - parameters describing ‘‘perceptually relevant differences’’ (in terms of spatial hearing) between the original audio channels
    - Very low bit-rate (order of 2 lower)
- In the decoder, the channels are synthesized from the downmix
  - Close to the original audio channels

References

- Brandenburg, et al. Perceptual Coding of High-Quality Digital Audio. Proc. IEEE, Sept 2013
- Wang and Vilermo. Modified discrete cosine transform ...
- Dietz, et al.
  Spectral Band Replication, a novel approach in audio coding
- Lutzky, et.al. A guideline to audio codec delay