Tampere University

Audio coding

SGN 14007

Lecture 6

Annamaria Mesaros

Introduction

- Transmission bandwidth increases continuously, but the demand increases more
  → need for compression technology

- Applications of audio coding
  - audio streaming and transmission over the internet
  - audio streaming over Bluetooth
  - mobile music players
  - digital broadcasting
  - soundtracks of digital video (e.g. digital television and DVD)

Requirements for audio coding

- **Compression efficiency**: sound quality vs. bit-rate
  - Typical goal: maximize perceived quality with minimal number of used bits
- **High fidelity**
  - Typical requirement: given sufficiently high bit-rate, no audible difference compared to CD-quality original audio
- **Complexity**
  - Computational complexity is a factor for mobile & embedded devices (power consumption and speed of execution)
  - Storage requirements: factor for dedicated silicon chips
  - Encoder vs. decoder complexity:
    - The encoder is usually much more complex than the decoder
    - Encoding can be done offline in some applications

Requirements for audio coding (cont.)

- **Algorithmic delay**
  - Depending on the application, the delay is or is not an important criterion
  - Very important in two way communication (~ 20 ms)
  - Not important in storage applications
  - Somewhat important in digital TV/radio broadcasting (~ 100 ms)
- **Random access**
  - A certain point in audio signal can be accessed from the coded bitstream
  - Requires that the decoding can start at (almost) any point of the bitstream
- **Error resilience**
  - Susceptibility to single or burst errors in the transmission channel
  - Usually combined with error correction codes, but that costs bits

Source coding vs. perceptual coding

- Usually signals have to be transmitted with a given fidelity, but not necessarily perfectly identical to the original signal
- Compression can be achieved by removing:
  - Redundant information that can be reconstructed at the receiver
  - Irrelevant information that is not important for the listener

- **Source coding**: emphasis on redundancy removal
  - Speech coding: a model of the vocal tract defines the possible signals, parameters of the model are transmitted
    - works poorly in generic audio coding: any kind of signals are possible
  - Spectral band replication (SBR): model the high frequency parts of audio using typically highly correlated lower frequencies. Reconstruct at receiver.

- **Perceptual coding**: emphasis on the removal of perceptually irrelevant information
  - Minimize the audibility of distortions
  - Minimize the number of used bits

Source coding vs. perceptual coding

- Speech and non-speech audio are quite different
  - In the coding context, the word ”audio” usually refers to non-speech audio!

- For audio signals (as compared to speech), typically:
  - Sampling rate is higher
  - Wideband speech codec up to 7kHz vs. CD 22.05 kHz
  - Dynamic range is wider
  - Power spectrum shape varies more
  - High quality is more crucial than in the case of speech signals
  - Stereo and multichannel coding can be considered

- The bitrate required for speech signals is much lower than that required for audio/music

Lossless coding vs. lossy coding

- Lossless or noiseless coding
  - Able to reconstruct perfectly the original samples
  - Compression ratios approximately 2:1
  - Can only utilize redundancy reduction

- Lossy coding
  - Not able to reconstruct perfectly the original samples
  - Compression ratios around 10:1 or 20:1 for perceptual coding
  - Based on perceptual irrelevancy and statistical redundancy removal

Measuring audio quality

- Lossy coding of audio causes inevitable distortion to the original signal

- The amount of distortion can be measured using:
  - **Subjective listening tests**, for example using mean opinion score (MOS): the most reliable way of measuring audio quality
  - **Simple objective criteria** such as signal-to-noise ratio between the original and reconstructed signal (quite non-informative from the perceptual quality viewpoint)
  - **Complex criteria** such as objective perceptual similarity metrics that take into account the known properties of the auditory system (for example the masking phenomenon)

Mean opinion score MOS

- Test subjects rate the encoded audio using N-step scale from 1 to 5
- MOS is defined as the average of the subjects' ratings
- Widely used but has also drawbacks:
  - results vary across time and test subjects
  - results vary depending on the chosen test signals (typical audio material vs. critical test signals)

- Example scale for rating the disturbance of coding artefacts:

Grade (MOS)	Subjective opinion	Quality
5 Excellent	Imperceptible	Transparent
4 Good	Perceptible, but not annoying	Toll
3 Fair	Slightly annoying	Communication
2 Poor	Annoying	Synthetic
1 Bad	Very annoying	Bad

Psychoacoustics and perceptual coding

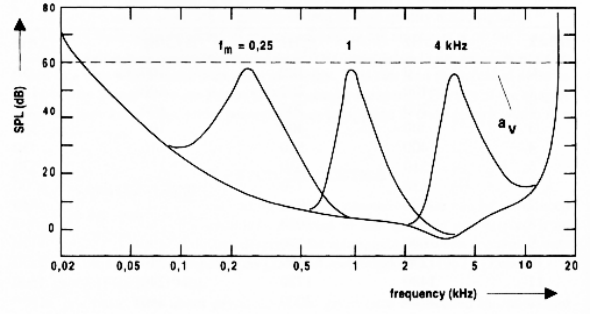
- Main question in perceptual coding: **How much noise (distortion, quantization noise) can be introduced into a signal without it being audible?**

- The answer can be found in psychoacoustics
  - Psychoacoustics studies the relationship between acoustic stimuli and the corresponding auditory sensations

- Most important keyword in audio coding is **”masking”**
  - Masking describes the situation where a weaker but clearly audible signal (maskee) becomes inaudible in the presence of a louder signal (masker)
  - Masking depends both on the spectral composition of the maskee and masker, and their variation over time

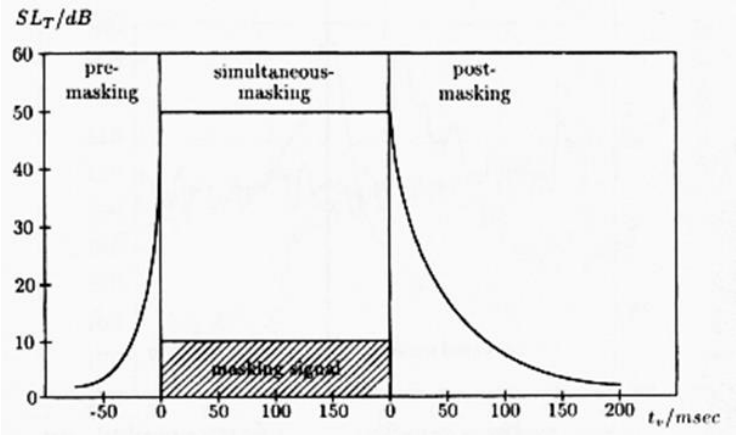
Masking in frequency domain

- Frequency analysis in the auditory system:
  - Subdivision of the frequency axis into critical bands
  - Frequency components within a same critical band mask each other easily
  - Bark and ERB scales: frequency scales that are derived by mapping frequencies to critical band numbers
- Narrowband noise masks a tone (sinusoidal) easier than a tone masks noise
- Masked threshold refers to the raised threshold of audibility caused by the masker
  - Sounds with a level below the masked threshold are inaudible
  - Masked threshold in quiet = threshold of hearing in quiet
  - Additivity of masking: joint masked threshold is approximately (but slightly more than) sum of the components



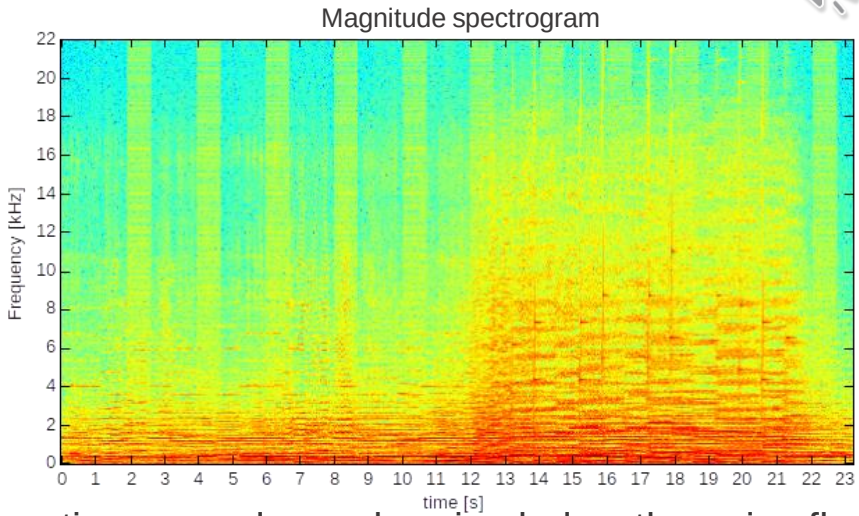
Masking in time domain

- Forward masking (post-masking)
  - masking effect extends to times after the masker is switched off
- Backwards masking (pre-masking)
  - masking extends to times before the masker is been switched on
- Forward/backward masking does not extend far in time
- Simultaneous masking is more important



Masking example

- Samples form http://ethanwiner.com/audibility.html
  - Noise is 15 to 25 dB below the music
  - Noise is 45 to 55 dB below the music



- Interesting examples on hearing below the noise floor

Thinking break

Perceptual audio coding

Overview of perceptual audio coding

- Basic idea is to **hide quantization noise below the signal-dependent threshold of hearing (masked threshold)**
  - Quantization can be performed for each frequency band
  - Different amount of quantization at different frequencies

- Modeling the masking effect
  - Most important masking effects are described in the frequency domain
  - On the other hand, effects of masking extend only up to about 15ms distance in time

- Consequence:
  - Perceptual audio coding is best done in time-frequency domain
  - Common basic structure of perceptual coders

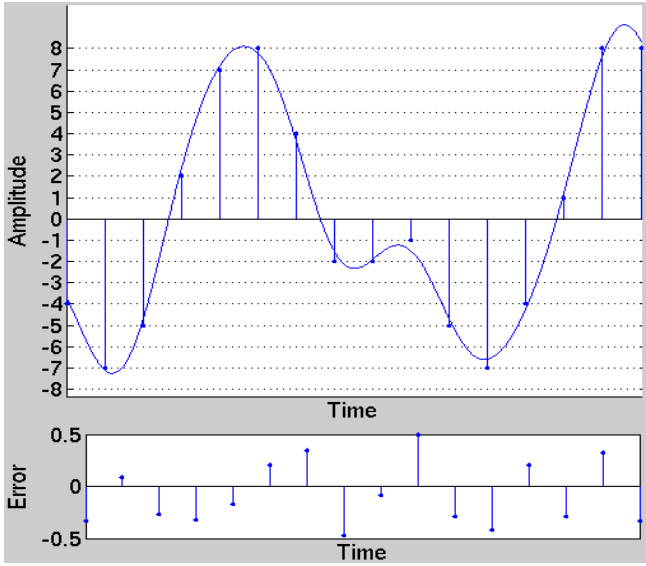
Quantization

- Recall: analog to a digital conversion requires sampling (at specific rate) and quantization (with finite number of bits)
- Quantized signal can be written as $x_q(n) = x(n) + e(n)$,
  - x(n) is the original (continuous valued) signal,
  - $x_q(n)$ is the discrete valued signal
  - Step-size $\Delta = 2x_{max}/(2^w)$
  - e(n) is the quantization error (Fig.:lower panel); white noise (same level across frequencies)
- Signal-to-noise ratio:

$$SNR = 10log_{10}\left(\frac{\sigma_x^2}{\sigma_e^2}\right)$$

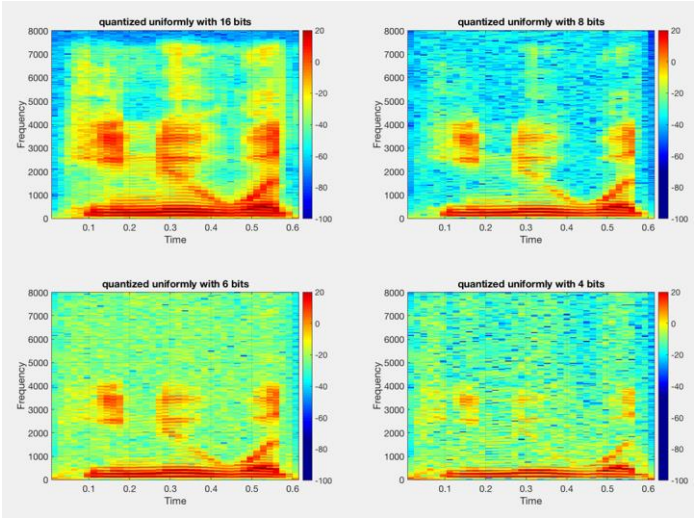
σ_x^2 and σ_e^2 are signal and noise powers ($E[x(n)^2]$ and $E[e(n)^2]$)

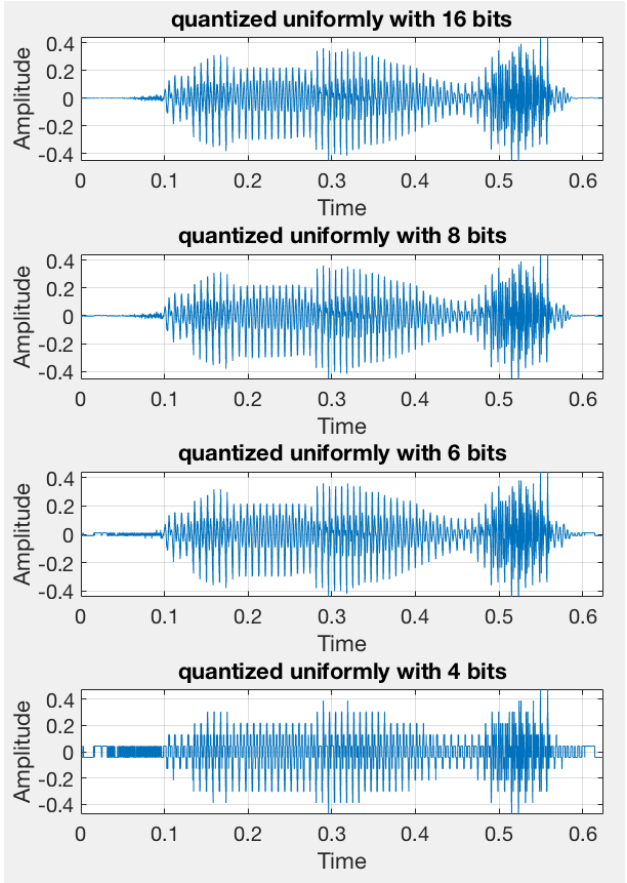
- Quantization with more bits (w) increases the SNR:
  - Approximately SNR = 6.02w
  - Quantiz. error level decreases when bits are added
  - 16 bits results in approximately 96 dB SNR



Quantization example

- Example: w=4,6,8,16
- Magnitude spectrogram below:
  - Quantization noise affects all frequencies.



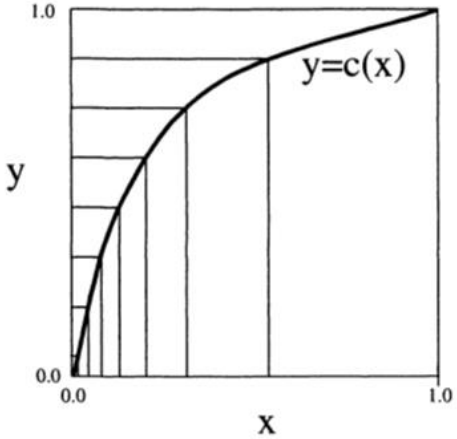


e.g. with w=4 bits, there are 2^4 (=16) possible amplitude values to represent the signal

Quantization

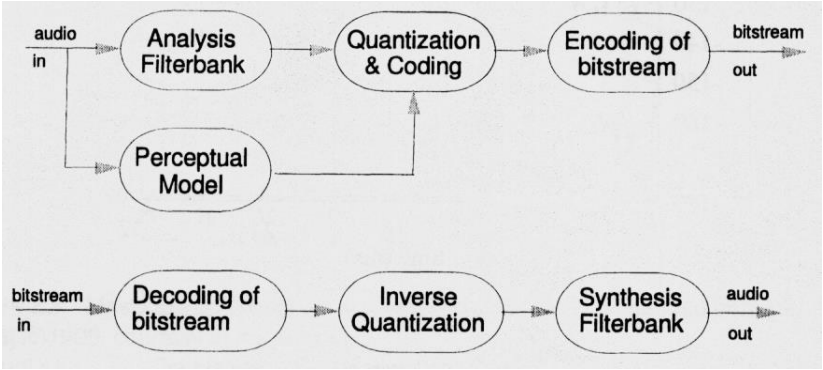
- Uniform quantization
  - Fixed quantization step size: $\Delta = 2x_{max}/2^{w}$

- Non-uniform quantization
  - Step size Δ varies with input amplitude x
  - y=c(x), c() denotes companding and $c()^{-1}$ is the inverse operation
  - y is then uniformly quantized
  - To obtain the linear-domain quantized sample: $x'=c^{-1}(y)$
  - Different types of operations for c()



Perceptual audio coding: block diagram

- **Filter bank**
  - Decomposes an input signal into subbands or spectral components (time-frequency domain)
  - Determine the basic structure of a coder
- **Perceptual model** (aka psychoacoustic model)
  - Analyzes the input signal
  - Computes signal-dependent masked threshold based on psychoacoustics
  - The algorithmic core of an audio coder
- **Quantization and coding**
  - Spectral components are quantized and encoded
  - Goal is to keep quantization noise below the masked threshold
  - Implement the actual data reduction
- Frame packing
  - Bitstream formatter assembles the bitstream, which typically consists of the coded data and some side information

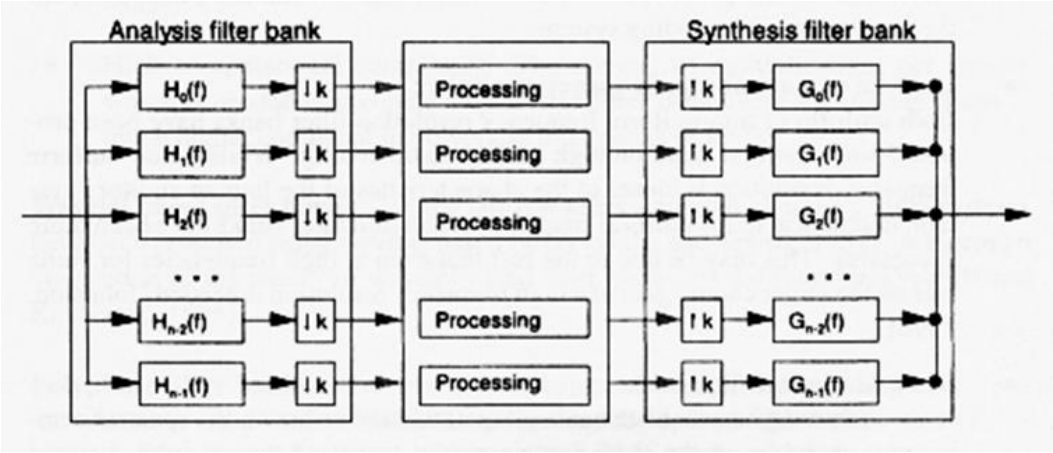


Filterbanks

Filterbanks

- The filter bank determines the basic structure of a code

- Example below: block diagram of a static n-channel analysis/synthesis filterbank [Herre95]
  - Downsampling by factor k at each channel
  - Bandwidths are identical uniform frequency resolution
  - Critical sampling if k=n



Filterbank parameters

- **Frequency resolution:**
  - Low resolution filter banks (e.g. 32 subbands), often called subband coders: quantization module usually works on blocks in time direction
  - High frequency resolution filter banks (e.g. 512 subbands), often called transform coders: quantization module usually works by combining adjacent frequency lines (recent coders)
  - Mathematically, all transforms used in audio coding systems can be seen as filter banks (distinction makes no sense theoretically)

- **Perfect reconstruction** filter banks
  - Enable lossless reconstruction of the input signal in an analysis/synthesis system, if quantization is not used
  - Simplifies the design of the other parts of a coding system
  - Usually either perfect or near perfect reconstruction filter banks are used

- **Prototype window** (windowing of the time frame)
  - Especially at low bit rates, characteristics of the analysis/synthesis prototype window are a key performance factor

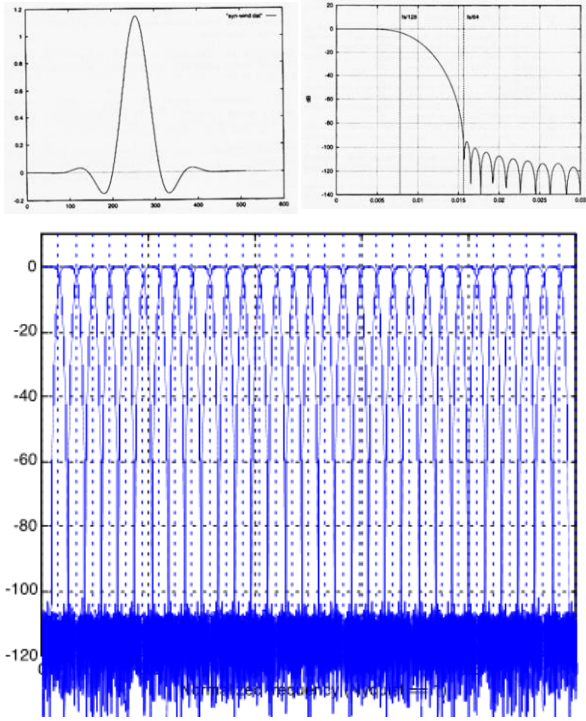
Filterbank parameters

- Uniform or non-uniform **frequency resolution**
  - Non-uniform frequency resolution is closer to the characteristics of the human auditory system
  - In practice, uniform resolution filter banks have been more successful (simplifies the coder design)

- Static or adaptive **filter bank**
  - Quantization error spreads in time over the entire synthesis window
  - **Pre-echo** can be avoided if filter bank is not static but switches between different time-/frequency resolutions
  - Example: adaptive window switching where the system switches to a shorter window in transient-like moments of change

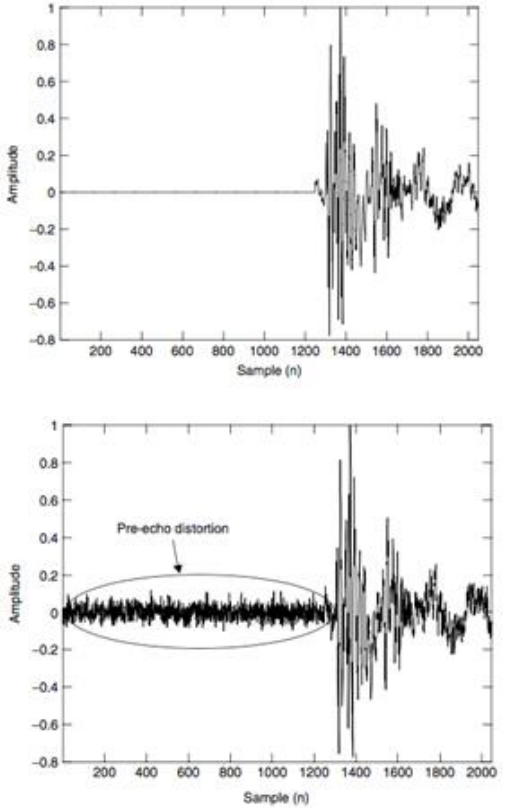
Filterbanks in use

- MPEG-1 Audio prototype LP filter
  - A low pass prototype filter is modulated with cosine function; we obtain for each freq band a bandpass filter
  - Prototype filter design is flexible
  - MPEG-1 audio: 511-tap prototype filter, very steep response
  - Reasonable trade-off between time behaviour and frequency resolution
- Transforms
  - In practice, modern encoders use modified discrete cosine transform (MDCT) as primary signal to be encoded
  - Window takes the part of a prototype filter, the transform modulates the filtered signal into the baseband
  - MPEG-2 AAC, HE-AAC, OPUS
  - MDCT from prev.lecture:
    - Window function is constructed in such a way that it satisfies the perfect reconstruction condition:
    - Critical sampling and overlapping frames



Pre-echo

- In general large number of subbands is beneficial to exploit masking phenomenon accurately
  - This results in long analysis windows (dual nature of time/frequency representation)
- **Pre-echo:** coder-generated audible artifacts (distortions) spread in time to **precede** the signal itself
  - The inverse transforms used in transform coders cause temporal spreading of **quantization noise** within the single frame due to time- frequency uncertainty
- Figure: Original signal and pre-echo distortion (below), visible before main peak.
- Shorter time window could be used to avoid pre-echo

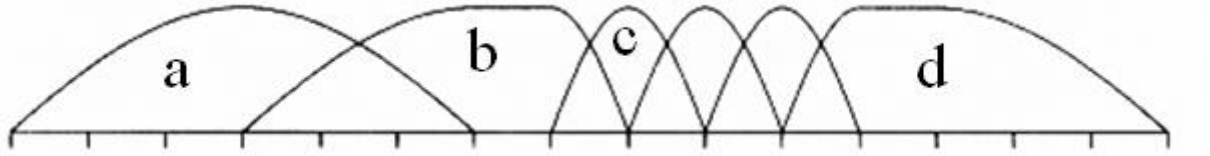


[Spanias et.al]

Adaptive filterbanks

- In the basic configuration, time-frequency decomposition is static
- Adaptive window switching is used e.g. in MPEG-1 Layer 3 (mp3)

- Figure: example sequence
  - a) long window: normal window type used for stationary signals
  - b) start window: ensures time domain alias cancellation for the part which overlaps with the short window
  - c) short window: same shape as a), but 1/3 of the length time resolution is enhanced to 4 ms (192 vs. 592 frequency lines)
  - d) stop window: same task as that of the start window
- Short windows used around transients for better time resolutions



Break

Perceptual models

Perceptual model role

- The perceptual model constitutes the algorithmic core of a coding system
- Most coding standards only define the data format
  - Allows changes and improvements to the perceptual model after the standard is fixed
  - e.g. ”mp3” format was standardized 1992 but became popular much later and is still widely used

- Main task: **deliver accurate estimates of the allowed noise**

- Additional tasks include
  - control of adaptive window switching (if used)
  - control of bit reservoir (if used)
  - control of joint stereo coding tools

Perceptual coding

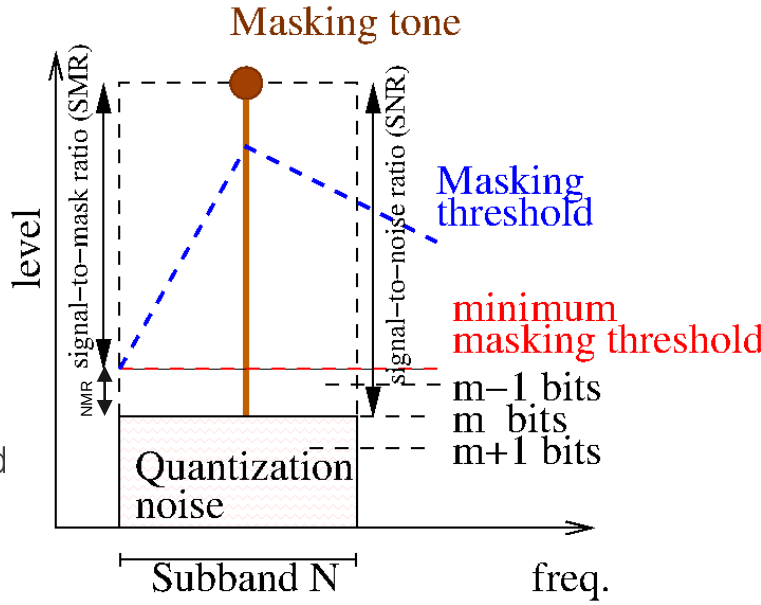
- Strong and prominent components of the signal being encoded hide the noise and distortion resulting from coding.

- Masking thresholds must be inferred from analyzed data.
  - The masking effect depends on frequency and level (and time)
  - Generally, masking is asymmetric and spreads more upwards towards higher frequencies.
  - Spread increases with level.
  - Tonal masker are different from noise maskers. Noise raises the threshold more than tonal signals with equal power.
  - Audio signals contain multiple maskers.

- Individual masking curves are combined by summing masking levels and combined with the threshold of quiet.
  - Result: Global masking threshold curve over the frequency range (per analysis frame).

Bit allocation in perceptual models

- Given the **masked threshold curve**, we can maximize quantization noise within each sub-band while keeping noise inaudible (noise shaping)
- Uniform quantization noise within subband decreases 6.02 dB by adding 1 bit
- Figure: the quantization affects the SNR of sub-band.

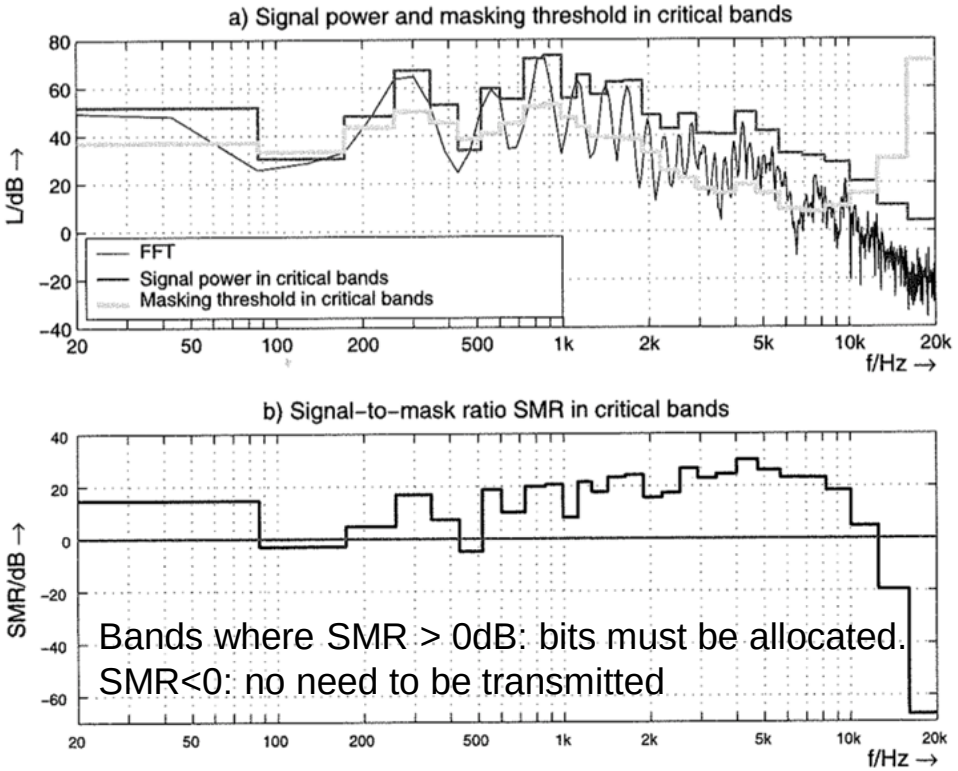


- **Signal-to-mask ratio SMR** is the band's margin to mask quantization noise.
  - SMR = Signal level – min(subband masking threshold

- **Noise-to-Mask ratio** (NMR) = SNR – SMR [dB]
- Noise is inaudible (perceptual transparency) if NMR>0

SMR defines number of bits to use to mask quantization ~SMR/6.02

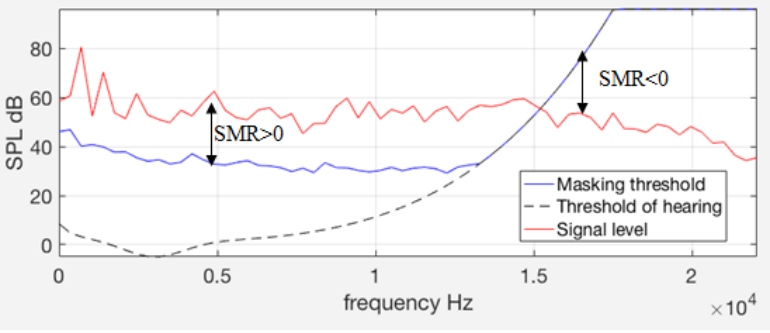
Masking threshold and SMR in critical bands



Bands where SMR > 0dB: bits must be allocated.
SMR<0: no need to be transmitted

[Zölzer]

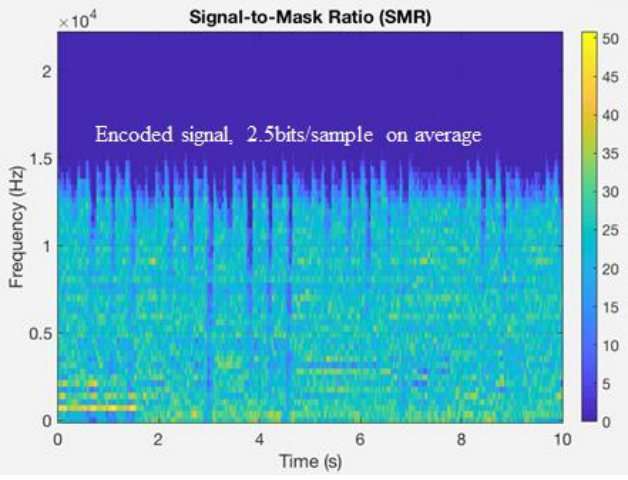
Example: encoded vs direct quantization

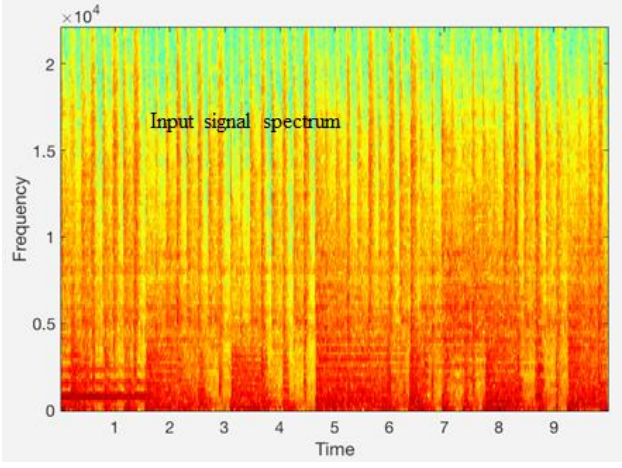


Original

Encoded

Uniform quantiz.





Quantization and coding

- Quantization and coding implement the actual data-reduction task in an encoder
- Remember that quantization is an essential part of analog-to-digital conversion
  - Analog sample values (signal levels) are converted to (binary) numbers

- In **coding**, digital signal values are further quantized to represent the data more compactly (and more coarsely)
  - In perceptual audio coding, quantization is performed in the time-frequency domain

- The quantized values are stored and/or transmitted either directly or as entropy coded words.
  - Entropy is a lower bound on the average number of bits needed to represent a set of symbols.

- **In audio coding SMR is the basis for bit allocation process**

Quantization and coding

- Perceptual models attempt to estimate a time-dependent signal-to-mask ratio (SMR) for each subband

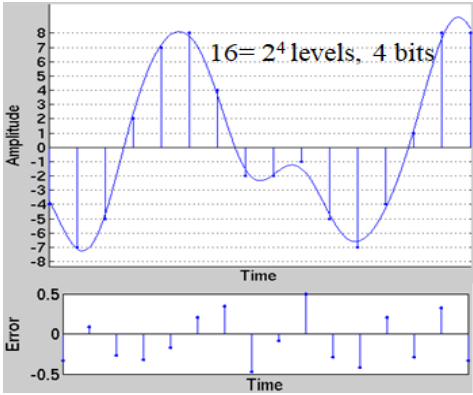


Figure: illustration of uniform quantization error in time-domain

- In perceptual audio coding, quantization is performed in time-frequency domain.
- ”block companding” using a *scalefactor* that is i) common for all bands or ii) band specific
- The same quantization step is used for all samples in the same block
- Uniform quantization results in noise with flat spectrum

Quantization and coding

Design options:

- **Quantization**: uniform or non-uniform quantization (MPEG-1 and MPEG-2 audio use power-law quantization)
- **Coding**: quantized spectral components are transmitted directly, or as entropy coded words (e.g. Huffman coding)

Quantization and coding control structures (two in wide use):

- **Bit allocation** (direct structure): a bit allocation algorithm driven (either by data statistics or) by a perceptual model decides number of bits per each frequency band. Bit allocation is done before the quantization.
- **Noise allocation** (indirect structure): data is quantized according to a perceptual model. The number of bits used for each component can be counted only after the process is completed. (since the lossless encoding (Huffman) encodes the quantized values with variable bit-rate)

Joint stereo coding

- Exploit redundancy of stereo signals and the irrelevancy of certain stereo features
  - Joint encoding of audio channels should lead to lower bitrate required to encode the channels separately.

- Redundancy
  - Contrary to intuition, there is usually not much correlation between the time domain signals of left and right channels
  - But power spectra of the channels are often highly correlated

- Irrelevancy
  - Human ability to localize sound sources decreases towards high frequencies
  - At high frequencies, the spatial perception is mainly based on intensity differences between channels at each frequency

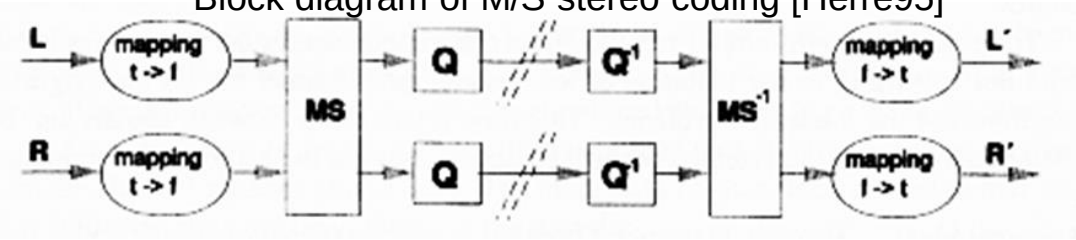
Joint stereo coding: pitfalls

- In some cases, the required bit-rate for stereo coding exceeds that needed for coding two mono channels
- Stereo unmasking effect
  - Certain coding artifacts which are masked in single channel coding can become audible when presented as a stereo signal coded by a dual mono coding system.
  - Binaural masking level difference (esp. at low frequencies)
  - Maximum masking occurs when the direction of the virtual quantization noise source coincides with the direction of the main signal source.
- Precedence effect
  - Sound sources are localized according to the first wavefront
  - Coding techniques may result in a distorted stereo image by introducing timing changes into the first wavefront arrival times.

Mid/Side (M/S) stereo coding

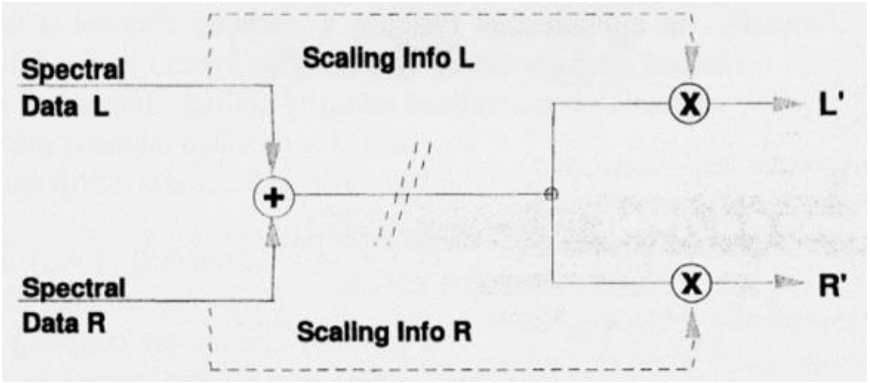
- Normalized sum and difference signals are transmitted instead of left and right channels (Middle and Side)
- Emphasis on **redundancy removal**
- Perfect reconstruction
  - Altering between L+R M+S does not lose information
- Heavily signal dependent bit-rate gain
  - Varies from 50 % (identical left/righ channel signals) to 0 %
- Preserves spatial information (can be applied to all freqs.)

Block diagram of M/S stereo coding [Herre95]



Intensity stereo coding

- For each subband, only the intensity spectrum is retained
  - Directional information is transmitted by encoding independent scalefactor values for left and right channels
- Rather successful at high frequencies (only applied to high freqs.)
  - Main spatial cues are transmitted, some details may be missing
  - Less annoying than other coding errors
- Emphasis on **irrelevancy removal**
  - 50 % data reduction at high frequencies, approx 20 % for the entire signal



Basic principle of intensity stereo coding [Herre95]

Details, not important for exam

Huffmann coding

- Lossless compression applied to quantised coefficients to remove further redundancy

  Imagine you have four symbols “A”,”B”,”C”,”D” to be encoded
  - 2-bits enough to transmit them 00,01,10,11
  - Average bit-rate is 2bits/symbol

  If probabilities of symbols is unequal
  - e.g. If p(“A“)=0.7, p(“B“)=0.15, p(“C“)=0.1,p(“D“)=0.05
  - Huffman coding result in: A=0, B=10, C=110, D=111
  - Average bit-rate 0.7*1+0.15*2+0.1*3+0.05*3=1.45bits/symbol

- Pre-computed tables kept for various codecs
  - Table tells the symbol from the coded representation
- Used in MPEG-1 layer 3 (.mp3) and MPEG-2 AAC

Details, not important for exam

Real coding systems

- MPEG (Moving Pictures Experts Group) standardizes compression techniques for video and audio
- Three low bit-rate audio coding standards have been completed
  - MPEG-1 Audio (layers 1, 2, and 3 (”mp3”))
  - MPEG-2 Audio - backwards compatible coding (multichannel, incl 5.1 configuration, more rates, codng efficiency at low bit- rates)
  - MPEG-2 Advanced Audio Coding (AAC) second generation audio coding scheme (
    - AAC reaches on average the same quality as Layer-3 at about 70 % of the bit-rate.
- MPEG-4 - a **family of coding algorithms** targeted for speech, audio, text-to-speech interfaces, lossless coding, still under development
- Codecs outside MPEG:
  - Ogg Vorbis, Windows Media Audio
  - OPUS: Speech + audio codec, high quality, open source
  - Similar basic principles applied to the discussed coders

Perceptual coding research?

- Filterbanks – still ongoing: wavelet based, low delay, variable filterbanks
- Perceptual models – search for better perceptual models for very low bit-rates and variable coding rate
- Quantization and coding – no major new ideas, but refinements and variations
- Losless and near lossless coding – with very low bit-rates, audio artifacts are accumulating and become audible; no standardization planned, but ongoing work to improve systems

Summary

- Perceptual audio coding exploits psychoacoustics to reduce bit-rate through quantization
  - **Auditory masking** effect plays a key role
- Quantization
  - Using less bits causes quantization noise level increase
  - Is done in separate frequency bands, so that the noise stays below the masking curve
- Redundancy removal: M/S stereo
- Irrelevancy removal: Intensity stereo coding