Perceptual audio coding

Nicola Orio

Dipartimento di Ingegneria dell’Informazione

IV Scuola estiva AISV, 8-12 settembre 2008

Introduction

- Current and future visual communications for applications such as
  - broadcasting
  - storage
  - videotelephony
  - video- and audiographic-conferencing
  - news gathering services
  - interactive multimedia (information, training, entertainment)

    services assume a substantial audio component.

- Even text, graphics, fax, still images, email documents, etc. will gain from voice annotation and audio clips.

Motivations

- Main motivations for low bit rate coding:
  - need to minimize transmission costs or provide cost efficient storage
  - demand to transmit over channels of limited capacity such as mobile radio channels
  - need to support variable-rate coding in packet-oriented networks
  - need to share capacity for different services (voice, audio, data, graphics, images) in integrated service network

PCM Audio Data Rate and Data Size

Quality	Sampling Rate (KHz)	Bits per Sample	Data Rate Kbits/s Kbytes/s	Data Size in 1 minute 1 hour
Telephone	8	8 (Mono)	64Kbps 8	480KB 28.8MB
AM Radio	11.025	8 (Mono)	88.2Kbps 11.0	660KB 39.6MB
FM Radio	22.050	16 (Stereo)	705.6Kbps 88.2	5.3MB 317.5MB
CD	44.1	16 (Stereo)	1.41Mbps 176.4	10.6MB 635MB

Conclusion → need advanced coding for compressing sound data

Basic requirements

- High quality of the reconstructed signal with robustness
  - to variations in spectra and levels
  - random and bursty channel bit errors
- Low complexity and power consumption of the codecs
  - more constraints on decoders than on encoders
- Additional network-related requirements:
  - low encoder/decoder delays
  - robust tandeming of codecs, transcodability
  - a graceful degradation of quality with increasing bit error rates (mobile radio and broadcast applications)
- Coded bit streams must allow
  - editing, fading, mixing, and dynamic range compression
- Synchronization between audio and video bitstreams

Wideband audio vs. Speech

- First proposals to reduce wideband audio coding rates have followed those for speech coding
- Speech and audio are still quite different and audio has
  - higher sampling rate
  - better amplitude resolution
  - higher dynamic range
  - larger variations in power density spectra
  - differences in human perception
  - higher listener expectation of quality
  - stereo and multichannel audio signal presentations
- Speech can be coded very efficiently because a speech production model is available
  - nothing similar exists for audio signals.

Evolution of audio coding

- Rapid progress in source coding
  - linear prediction
  - subband coding
  - transform coding
  - vector quantization
  - entropy coding

- Currently good coding quality can obtained with bit rates of
  - 1 bit/sample for speech
  - 2 bits/sample for audio

- Expectations over the next decade
  - 0.5 bit/sample for speech
  - 1 bits/sample for audio

Quality measures – 1

- Digital representations of analog waveforms cause some kind of distortion which can be specified
  - by *subjective* criteria as *mean opinion score* (MOS) as a measure of perceptual similarity
  - by simple objective criteria (i.e. SNR) as a measure of waveform similarity between source and reconstructed signal
  - by objective measures of perceptual similarity which take into account facts about human auditory perception.
- Mean opinion score (MOS)
  - subjects classify the quality of coders on an *N*-point quality scale
  - the final result is an averaged judgment called MOS
  - two five-point adjectival grading scales are in use, one for signal quality, and the other one for signal impairment, and an associated numbering

Quality measures – 2

- MOS advantages
  - different impairment factors can be assessed simultaneously
  - even small impairments can be graded
- MOS disadvantages
  - MOS value vary with time and listener panel to listener panel
  - difficult to duplicate test results at a different test site
  - in the case of audio signals, MOS values depend strongly on the selected test items
- ISO/MPEG tests
  - three signals A,B,C; A is unprocessed source, B and C are the reference and the system under test
  - the selection B/C is double blind
  - subjects have to decide if B or C is the reference and have to grade the remaining one

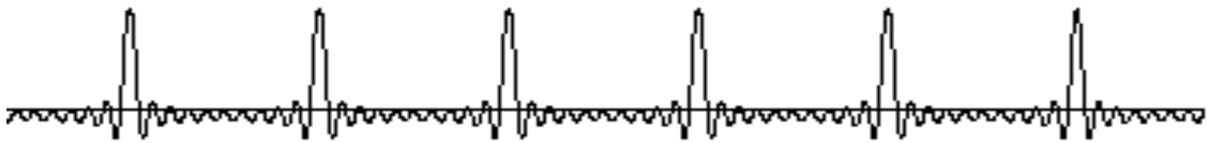
Audio coding

Perception

Perception : Frequency representation

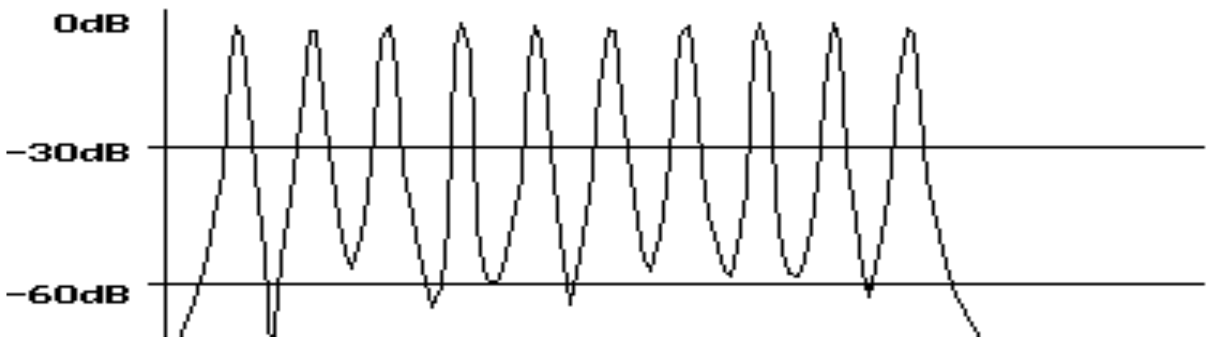
- The inner ear performs short-term analyses where frequency-to-place transformations occur along the basilar membrane
  - two sounds with different waveforms but same frequency components are perceived almost identical

sound 1



sound 2

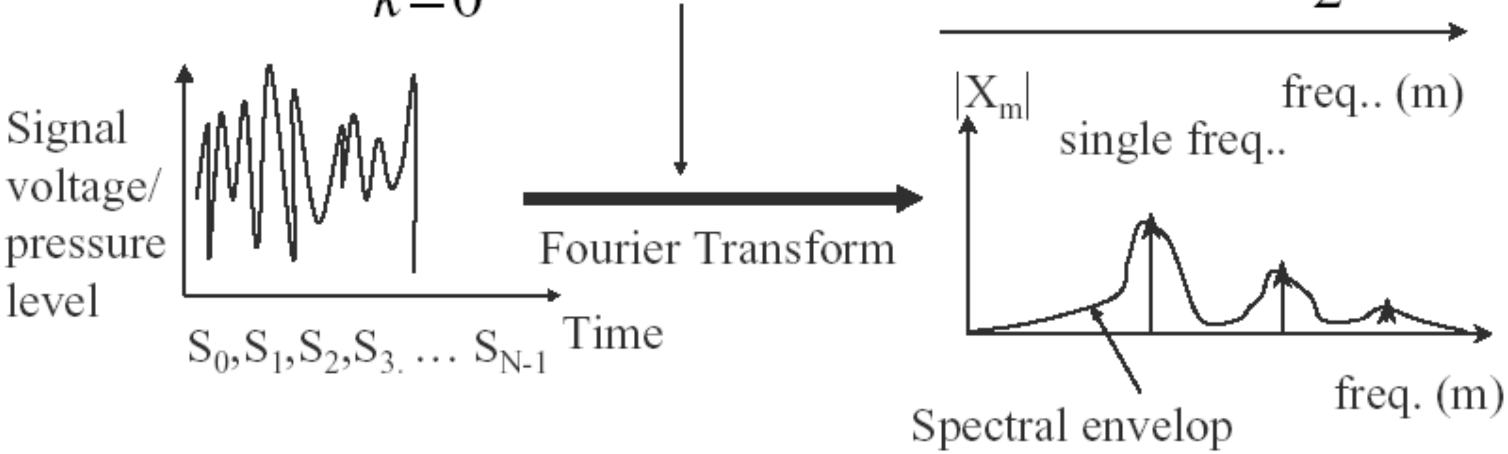
same spectrum



Perception : Fourier analysis

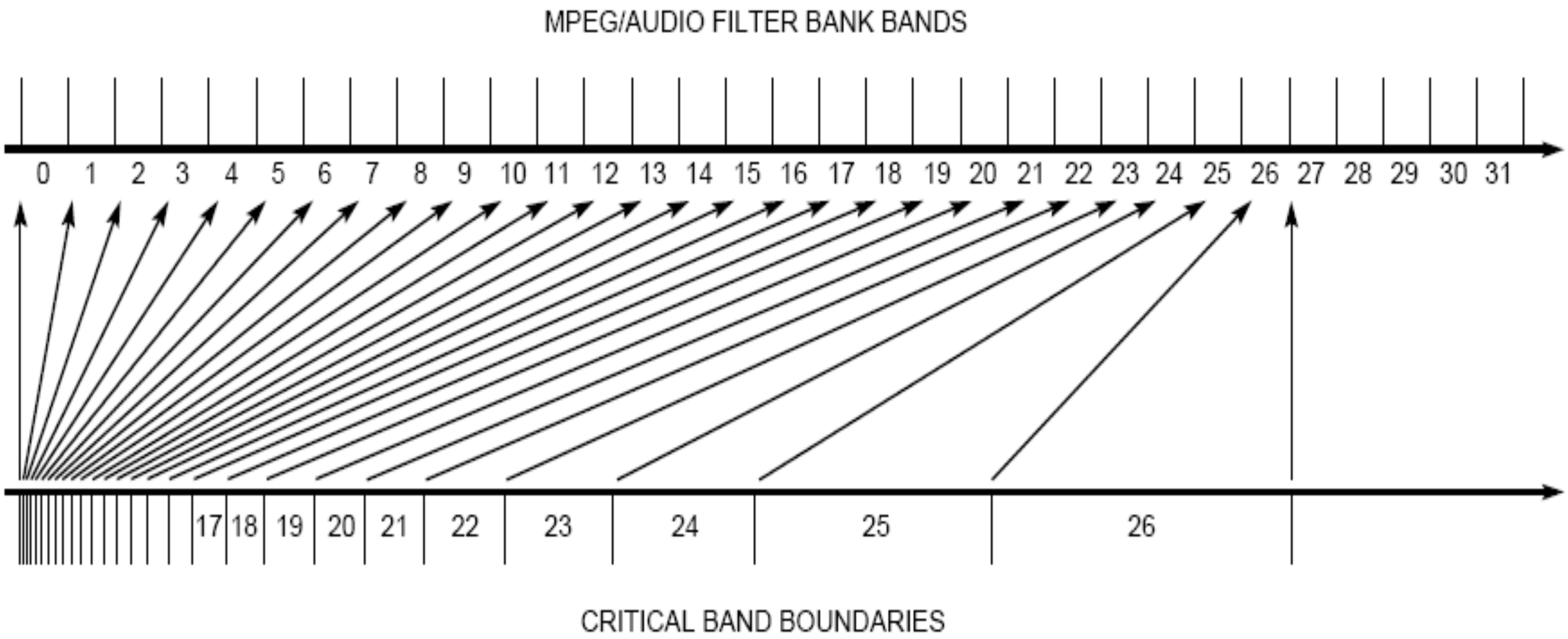
- Fourier analysis is a useful tool for sound processing
  - phase information is not as perceptually important as amplitude

$$X_m = \sum_{k=0}^{N-1} s_k \, e^{-j\frac{2\pi k m}{N}}, \qquad m = 0,1,2,\ldots,\frac{N}{2}$$



Perception : Critical bands – 1

- The power spectra are not presented on a linear frequency scale but on limited frequency bands called *critical bands*.
  - Rough description as a filterbank of bandpass filters with bandwidths that increase with the center frequency.



Perception : Critical bands – 2

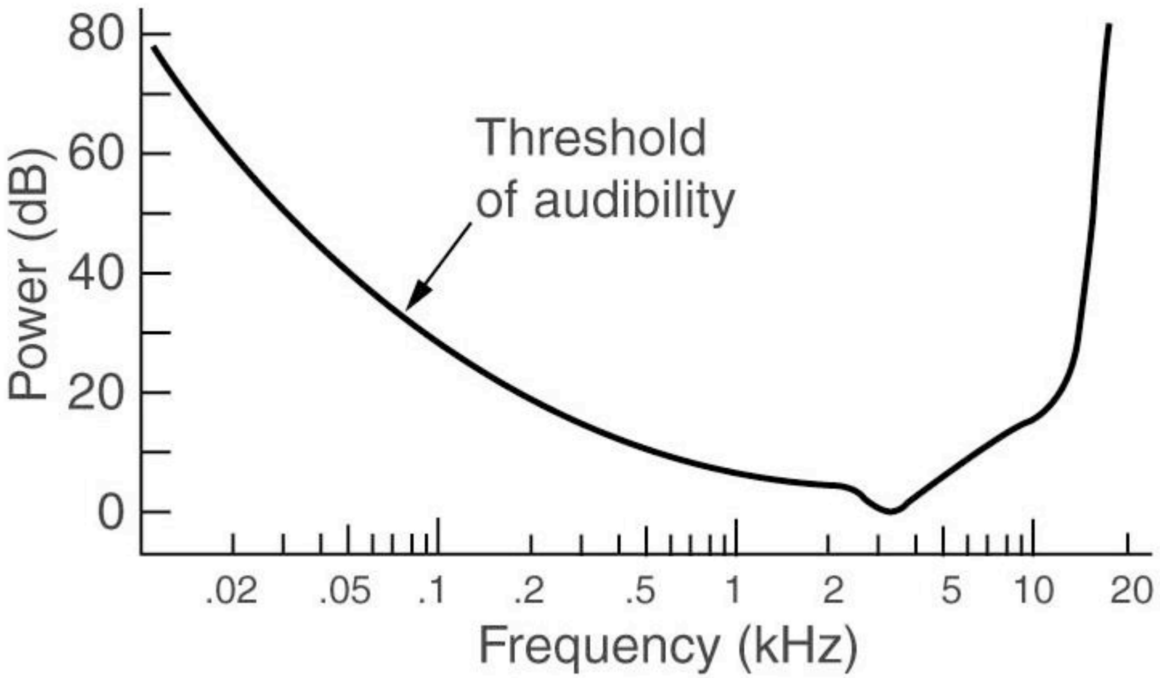
- Strong perceptual interference between frequencies in the same critical bands
  - Explanation: it looks that critical bands are related to sections of the acoustic nerve
- The scale related to critical bands is called **bark scale**

- Bandwidth of about 100 Hz below ~500 Hz
- Bandwith increase of about 20% above ~500 Hz

band	*center*	*bounds*
1	50	-100
2	150	100-200
3	250	200-300
…		
7	700	630-770
…		
11	1370	1270-1480
…		
15	2500	2320-2700
…		
19	4800	4400-5300
20	5800	5300-6400
…		
25	19500	15500-

Perception : Hearing limitations – 1

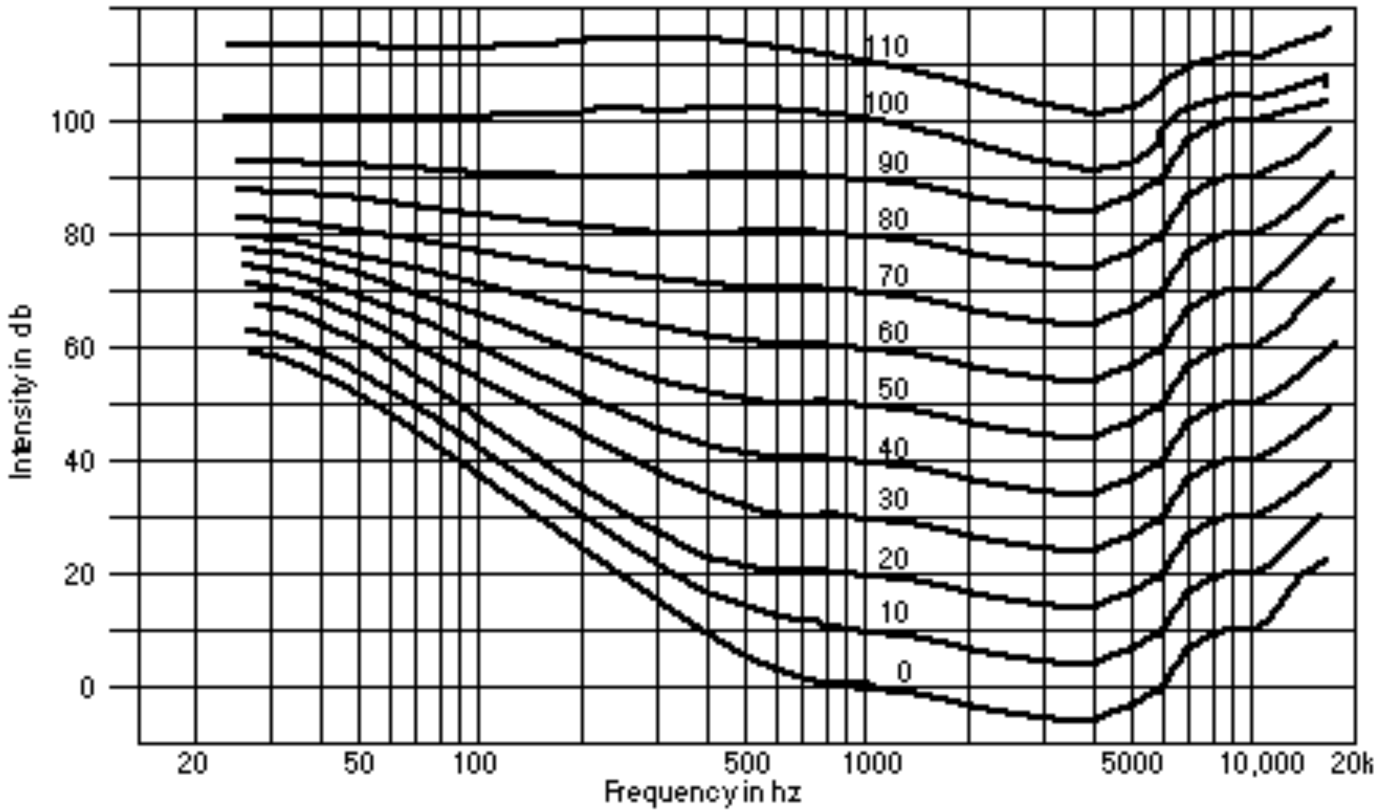
- Frequencies < 16-20 Hz and > 16-20 kHz are not perceived
- Amplitudes below a given threshold are not perceived
  - the threshold depends on the frequency



(a)

Perception : Hearing limitations – 2

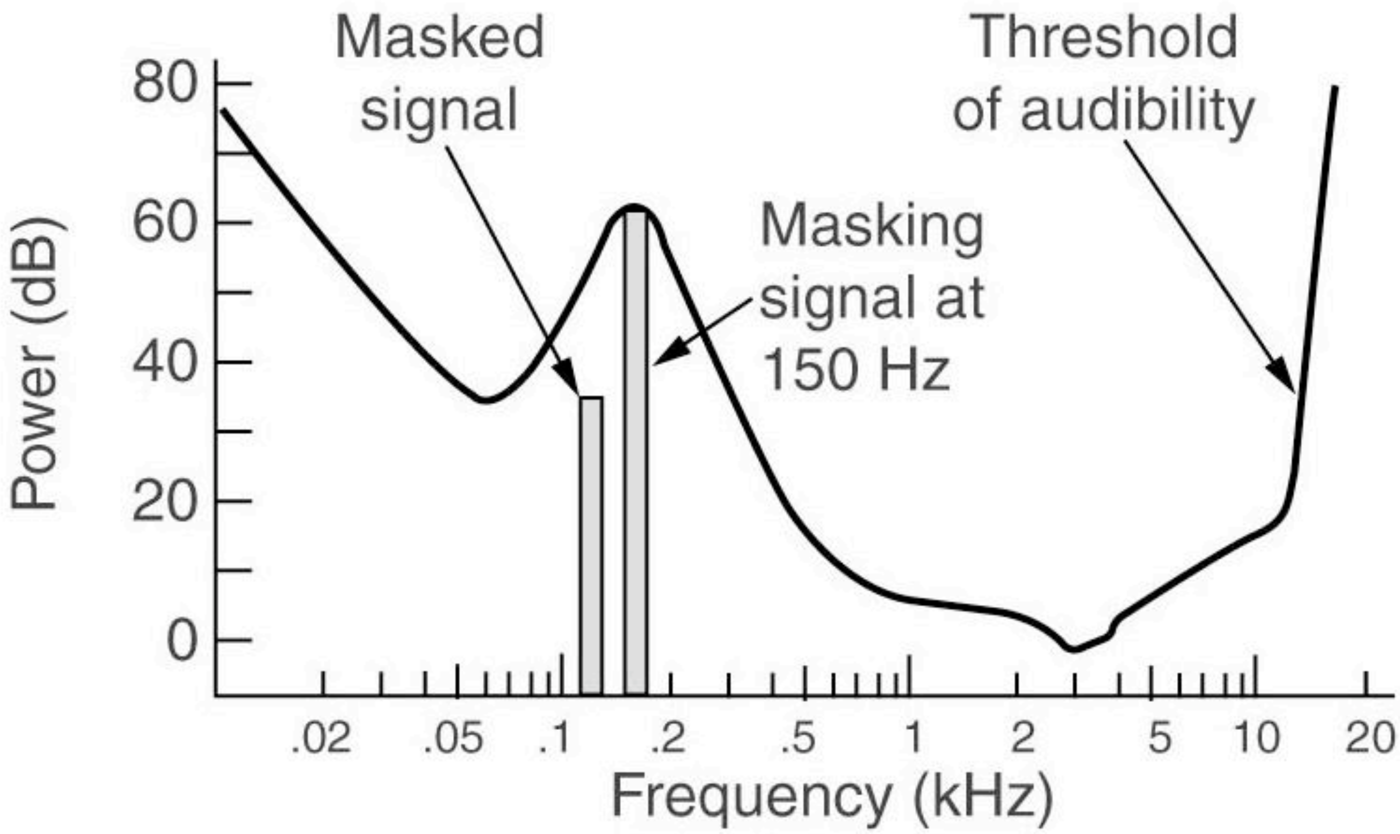
- Perceived intensity depends also on frequency
  - dynamic range (quitest to loudest) is about 100 dB



Perception : Masking – 1

- **Simultaneous masking** is a frequency domain phenomenon where a low-level signal (the maskee) can be made inaudible by a simultaneously occurring stronger signal (the masker)
  - masker and maskee should have close enough frequencies
- The *masking threshold*, in the context of source coding also known as *threshold of just noticeable distortion* (JND), varies with time. It depeds on
  - the sound pressure level (SPL),
  - the frequency of masker,
  - the characteristics of masker and maskee
- Without masker, a signal is inaudible if its sound pressure level is below the *threshold of audibility* anyway
- The distance between the level of the masker and the masking threshold is called *signal-to-mask ratio* (SMR).

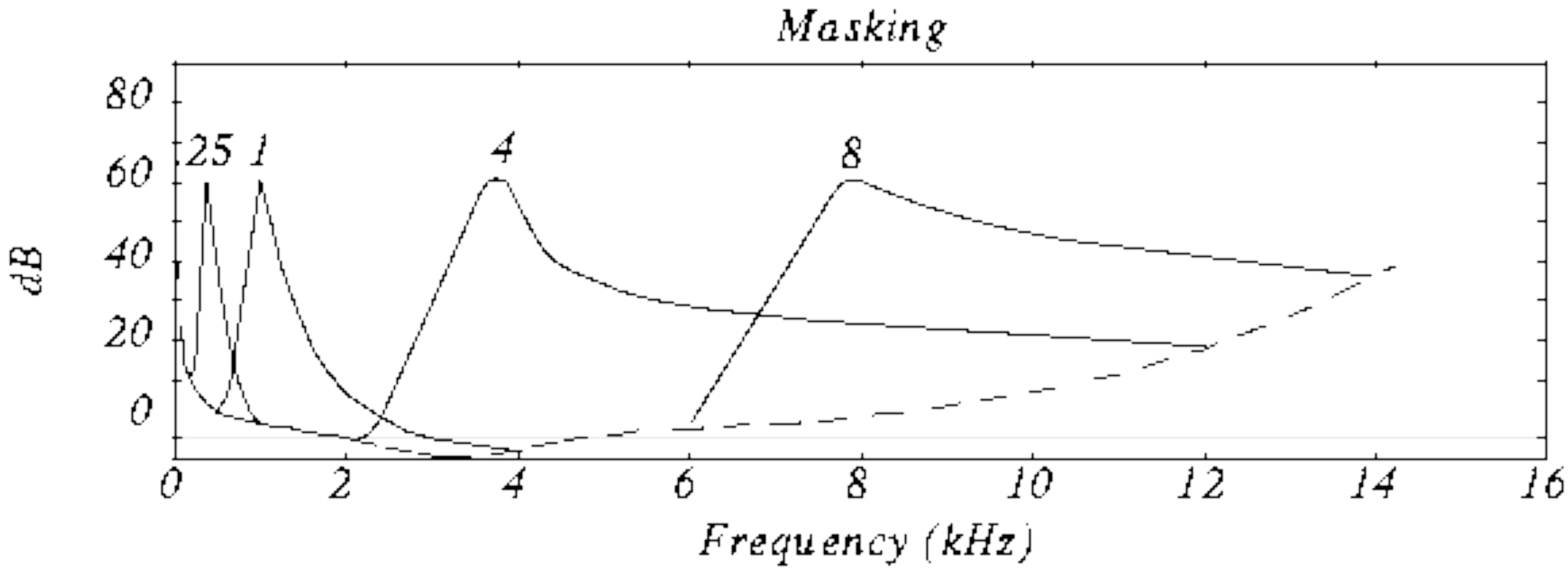
Perception : Masking – 2



(b)

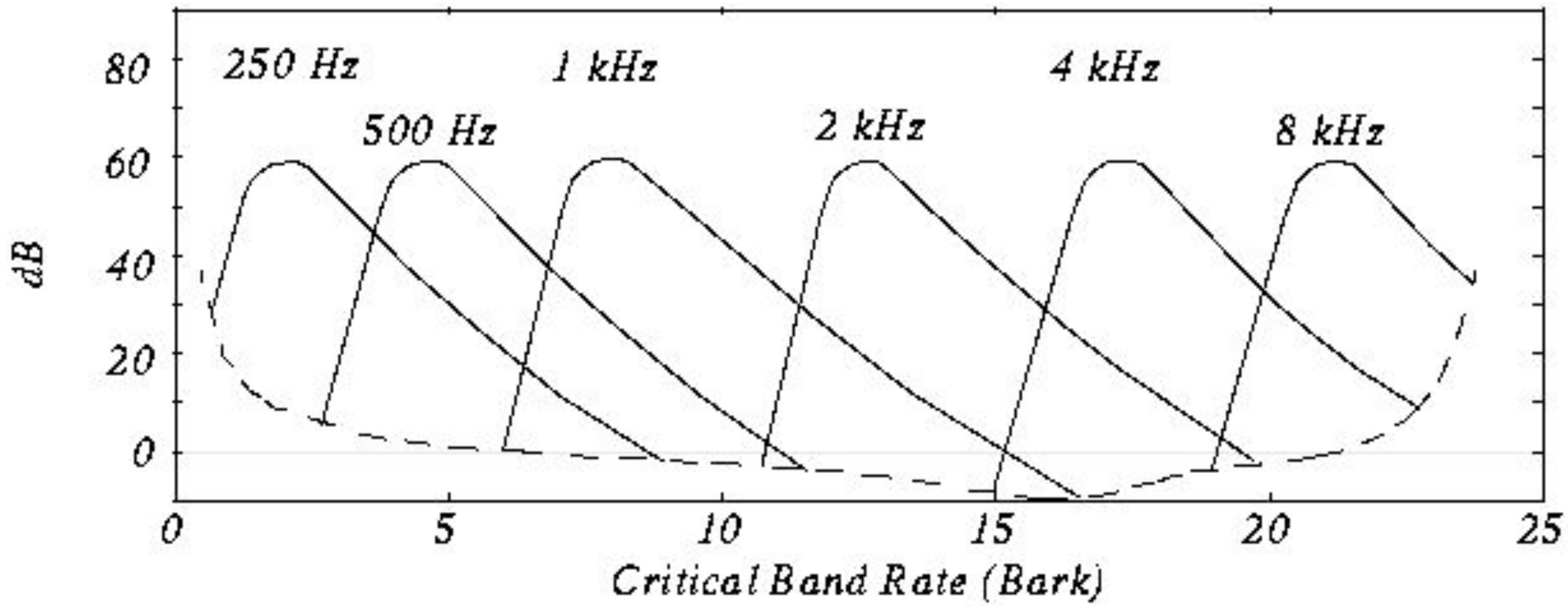
Perception : Masking – 3

- Different masking effects appear when masking and maskee are tones or noise
  - tone masking noise => strong
  - noise masking tone => weak
- Simultaneous masking changes with frequency



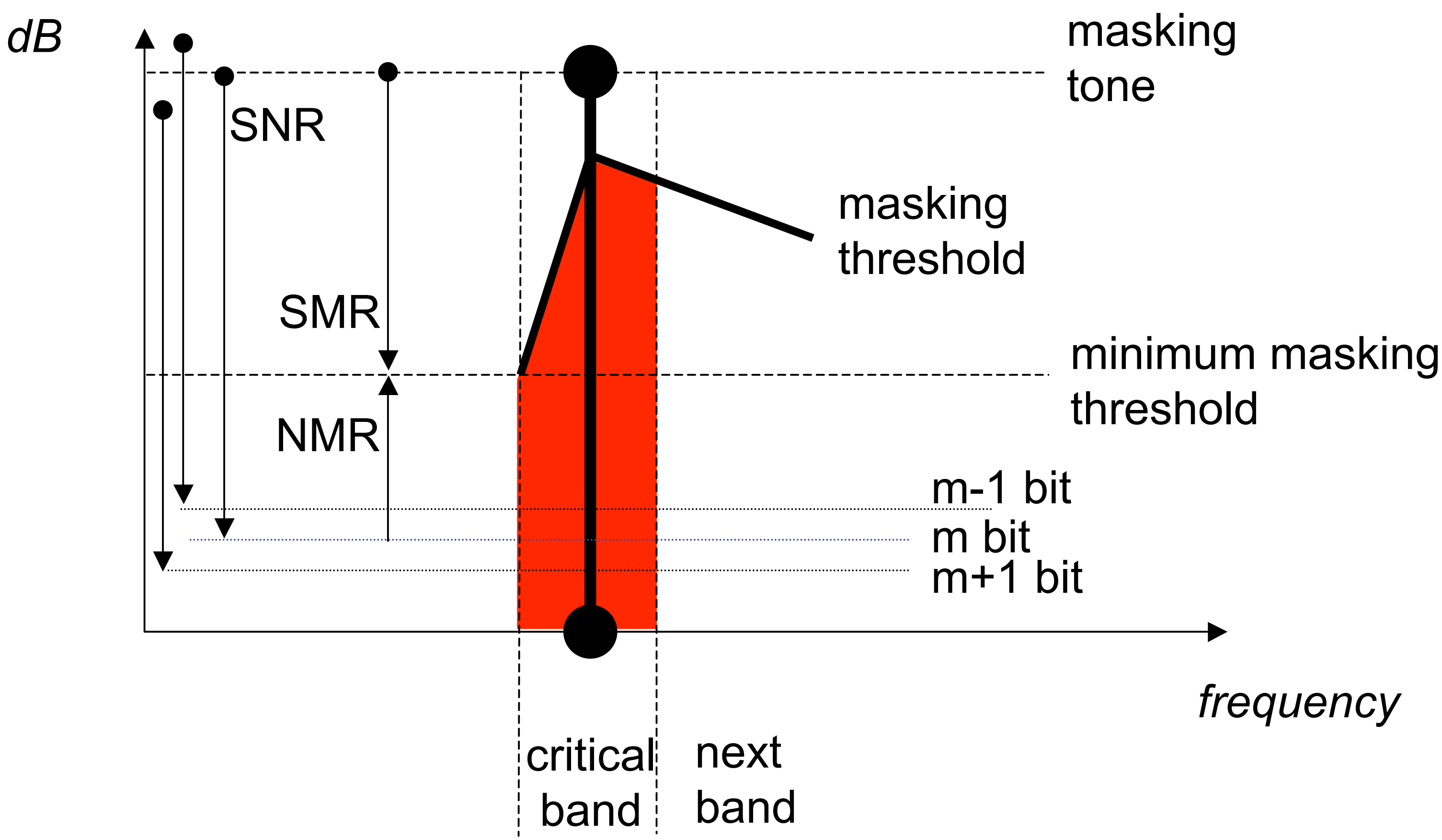
Perception : Masking – 4

- Masking effect is strictly related to the presence of critical bands
  - a weak stimulus is not perceived when a strong one excites the same perceptors
- almost constant in bark scale



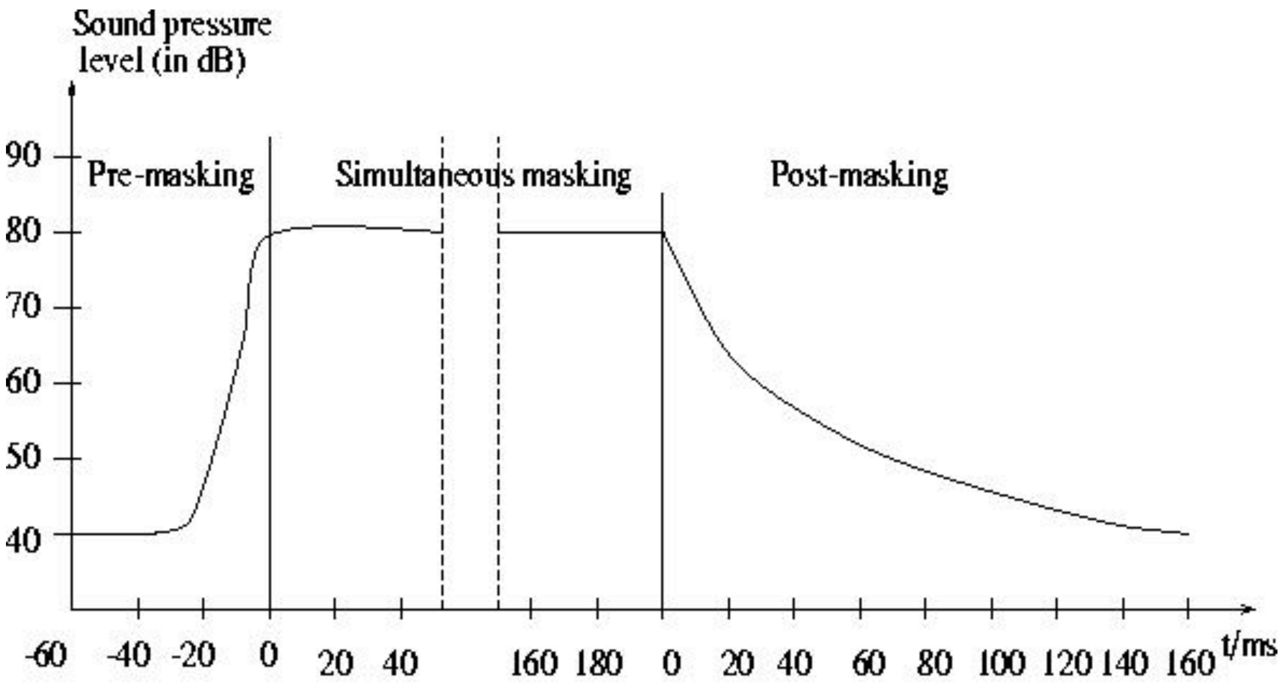
Perception : Masking – 5

- SNR = SMR (Signal to Mask Ratio) + NMR (Noise to Mask Ratio)



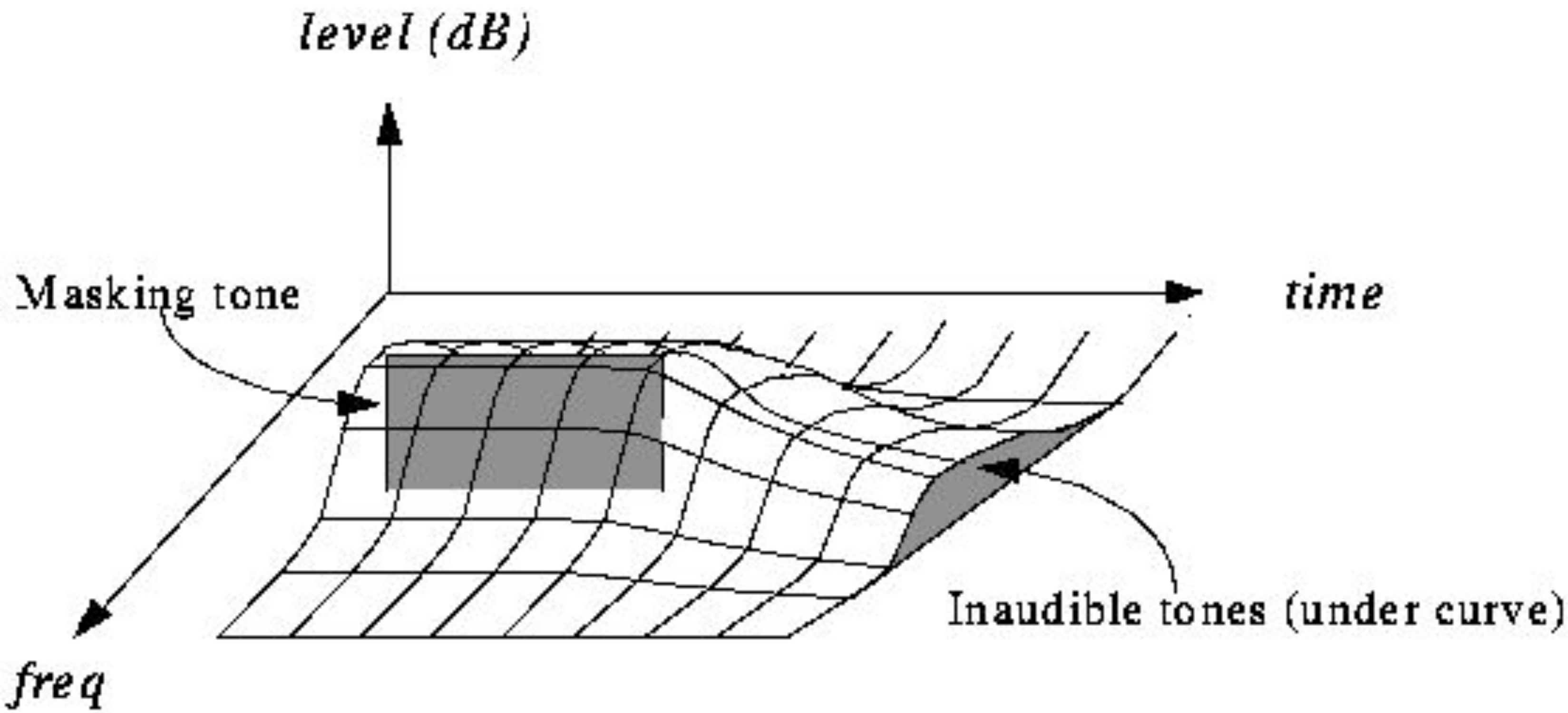
Perception : Temporal masking

- **Temporal masking** may occur when two sounds appear within a small interval of time.
  - a stronger sound may mask the weaker one, even if the maskee precedes the masker (pre- and post masking)



Perception : Global masking

- Simultaneous and temporal masking are combined together
  - effect is on the frequency axis, and
  - has influence on the temporal axis



Perception : Source localization

- Sound perception has some of limitations related to the source localization
  - stereo signals (or more)
- Low frequencies:
  - impossible to localize the audio source, mono is enough
- High frequencies:
  - localization is based on amplitude envelope only

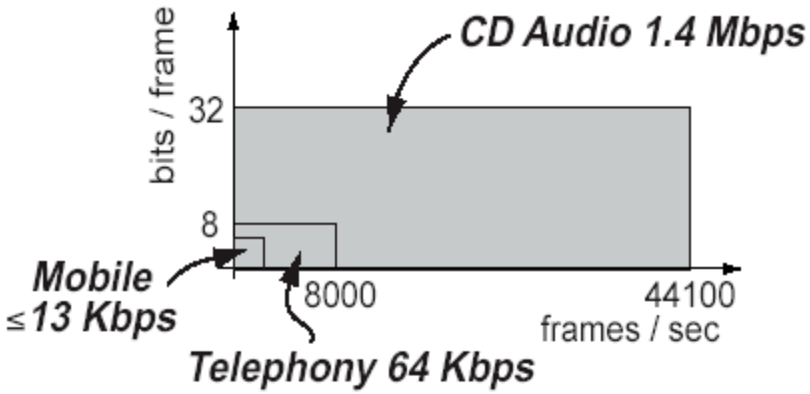
- In general, for stereo signals there may be
  - interchannel dependencies
  - interchannel masking effects
  - stereo-irrelevant components of the multichannel signal

Perceptual audio coding

Fundamentals

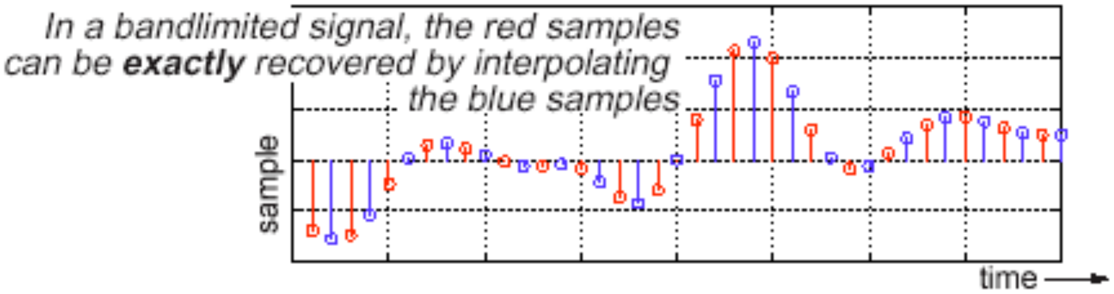
Compression & Quantization

- **How big is audio data? What is the bitrate?**
  - *Fs* frames/second (e.g. 8000 or 44100)
    x*C* samples/frame (e.g. 1 or 2 channels)
    x*B* bits/sample (e.g. 8 or 16)
  - ➔ $Fs \cdot C \cdot B$ bits/second (e.g. 64 Kbps or 1.4 Mbps)
- **How to reduce?**
  - lower sampling rate ➔ less bandwidth (muffled)
  - lower channel count ➔ no stereo image
  - lower sample size ➔ quantization noise
- **Or: use data compression**



Data compression: Redundancy vs. Irrelevance

- **Two main principles in compression:**
  - remove redundant information
  - remove irrelevant information
- **Redundant info is implicit in remainder**
  - e.g. signal bandlimited to 20kHz, but sample at 80kHz
  - ➔ can recover every other sample by interpolation:
- **Irrelevant info is unique but unnecessary**
  - e.g. recording a microphone signal at 80 kHz sampling rate



Irrelevant data in audio coding

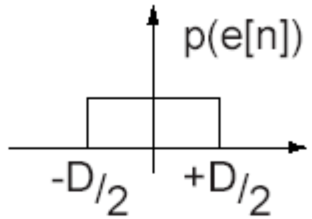
- **For coding of audio signals, irrelevant means perceptually insignificant**
  - an empirical property
- **Compact Disc standard is adequate:**
  - 44 kHz sampling for 20 kHz bandwidth
  - 16 bit linear samples for ~ 96 dB peak SNR
- **Reflect limits of human sensitivity:**
  - 20 kHz bandwidth, 100 dB intensity
  - sinusoid phase, detail of noise structure
  - dynamic properties - hard to characterize
- **Problem: separating salient & irrelevant**

Quantization

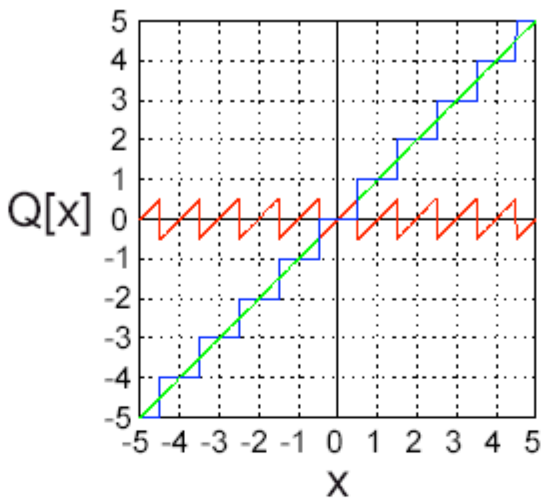
- Represent waveform with discrete levels
- Equivalent to adding error e[n]:

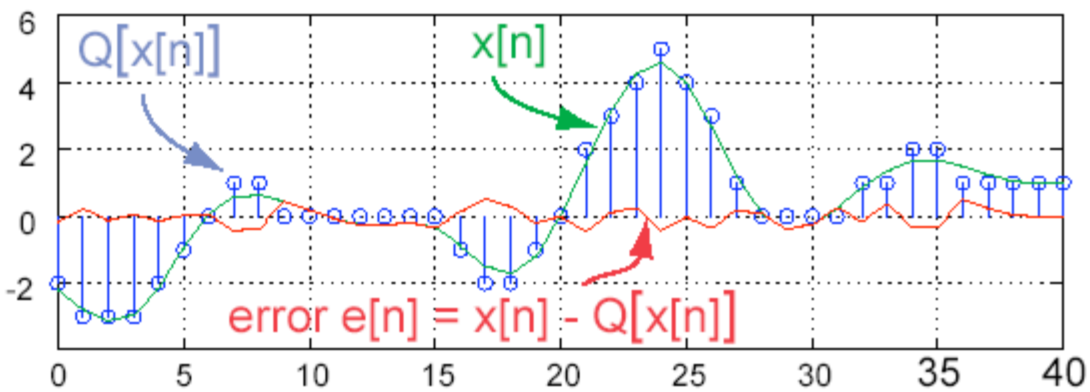
$$x[n] = Q[x[n]] + e[n]$$

- e[n] ~ uncorrelated, uniform white noise



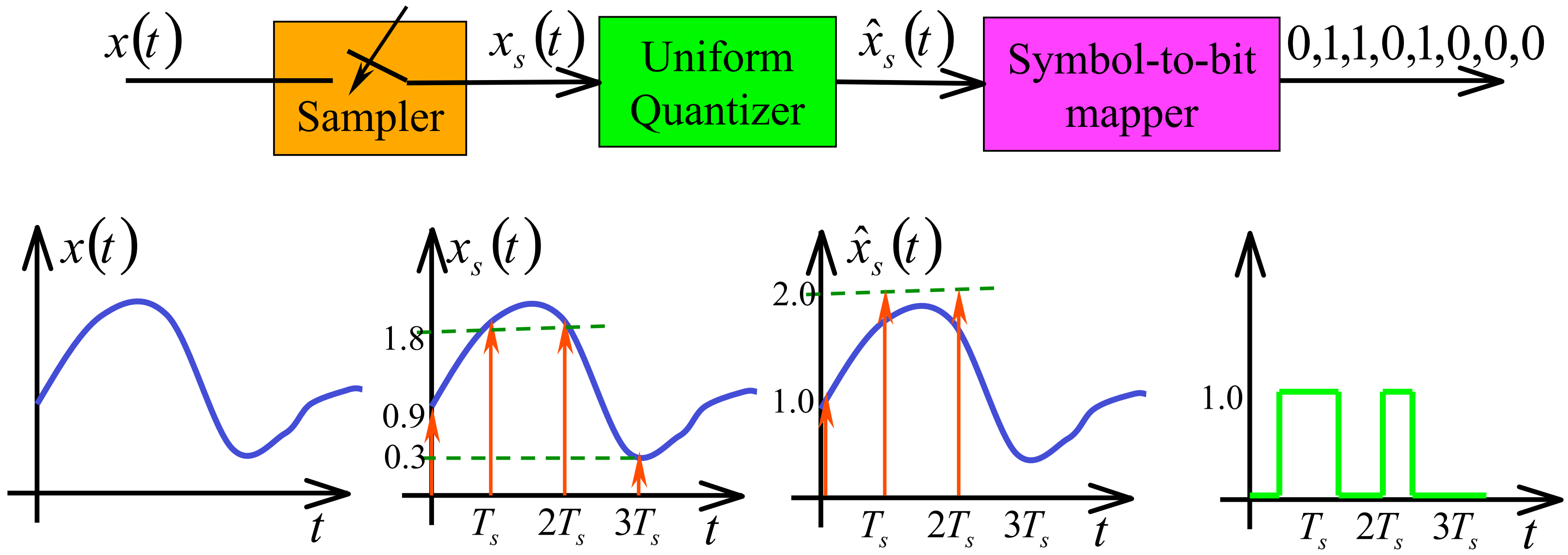
$$\rightarrow \text{variance } \sigma_e^2 = \frac{D^2}{12}$$





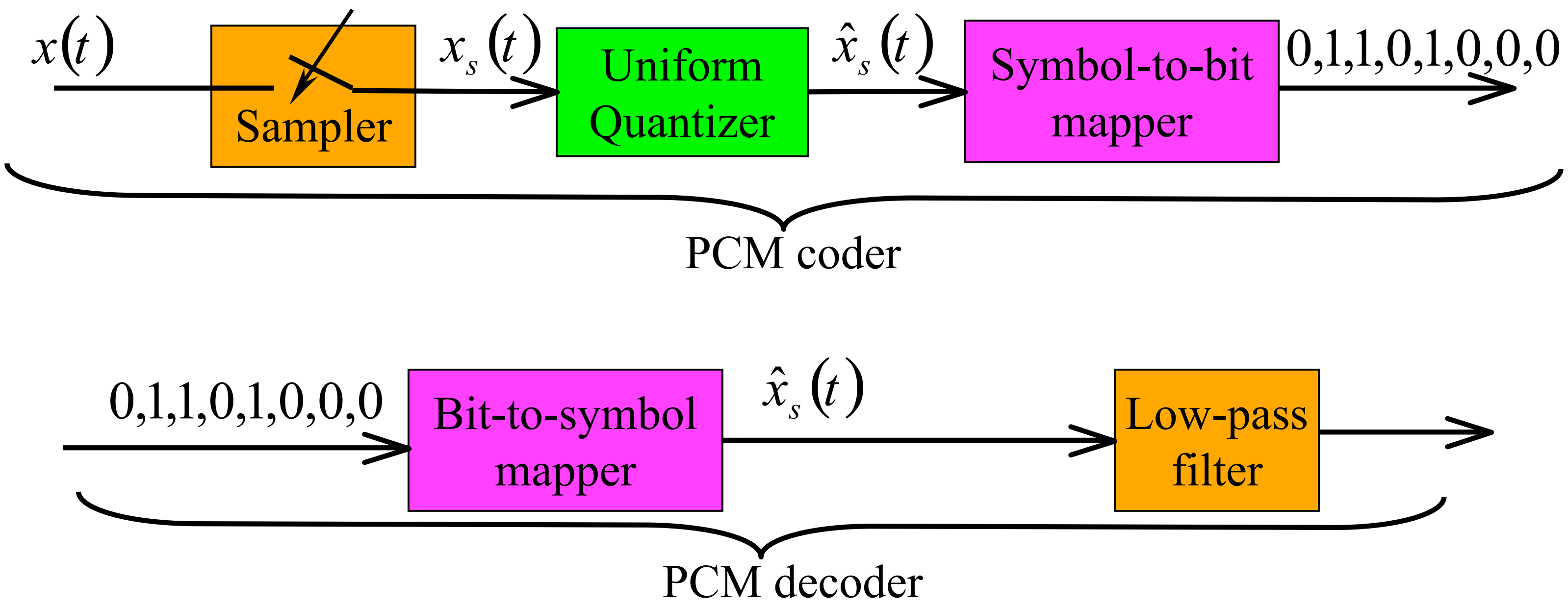
Linear Pulse Code Modulator (PCM)

- A simple source coder contains sampler, quantizer, and a mapper



Linear PCM Decoding

- A process to undo the effects of speech coding
- Locate at the receiver
- Linear PCM



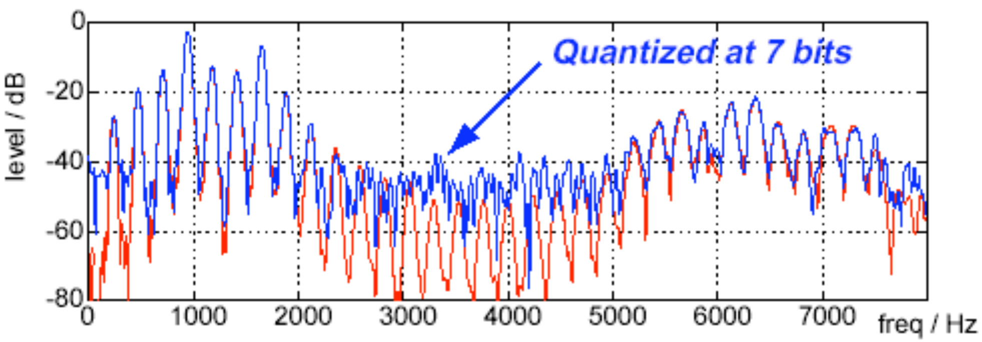
Quantization noise (Q-noise)

- **Uncorrelated noise has flat spectrum**
- **With a *B* bit word and a quantization step *D***
    - max signal range (x) = $-(2^{B-1})\cdot D$ .. $(2^{B-1}-1)\cdot D$
    - quantization noise (e) = $-D/2$ .. $D/2$

→ **Best signal-to-noise ratio (power)**

$$SNR = E[x^2]/E[e^2]$$
$$= (2^B)^2$$

.. or, in dB, $20 \cdot \log_{10} 2 \cdot B \approx 6 \cdot B$ **dB**

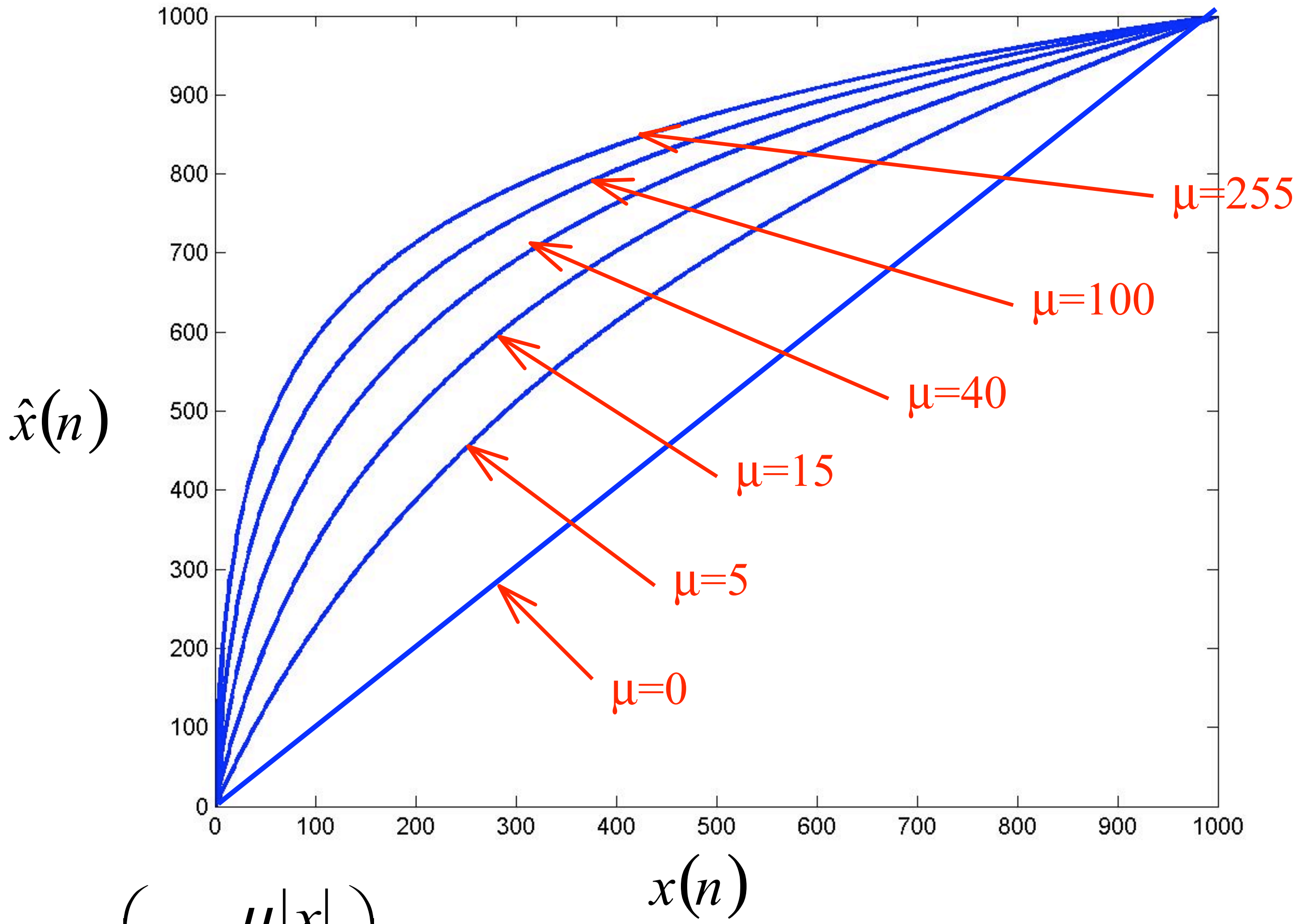


Non-Uniform Quantization

- Used to transmit in ISDN
  - 8 kHz, 8 bits: 64kbps
- Logarithmic quantization
  - dynamic range as 13/14 linear bits
  - 16 bits are compressed with a non linear technique in 8 bits samples

$$y = \begin{cases} 128 + \dfrac{127}{\ln(1+\mu)} \bullet \ln\left(1+\mu|x|\right), & x \geq 0 \\ 127 - \dfrac{127}{\ln(1+\mu)} \bullet \ln\left(1+\mu|x|\right), & x \leq 0 \end{cases}$$

μ-law Compression – 1

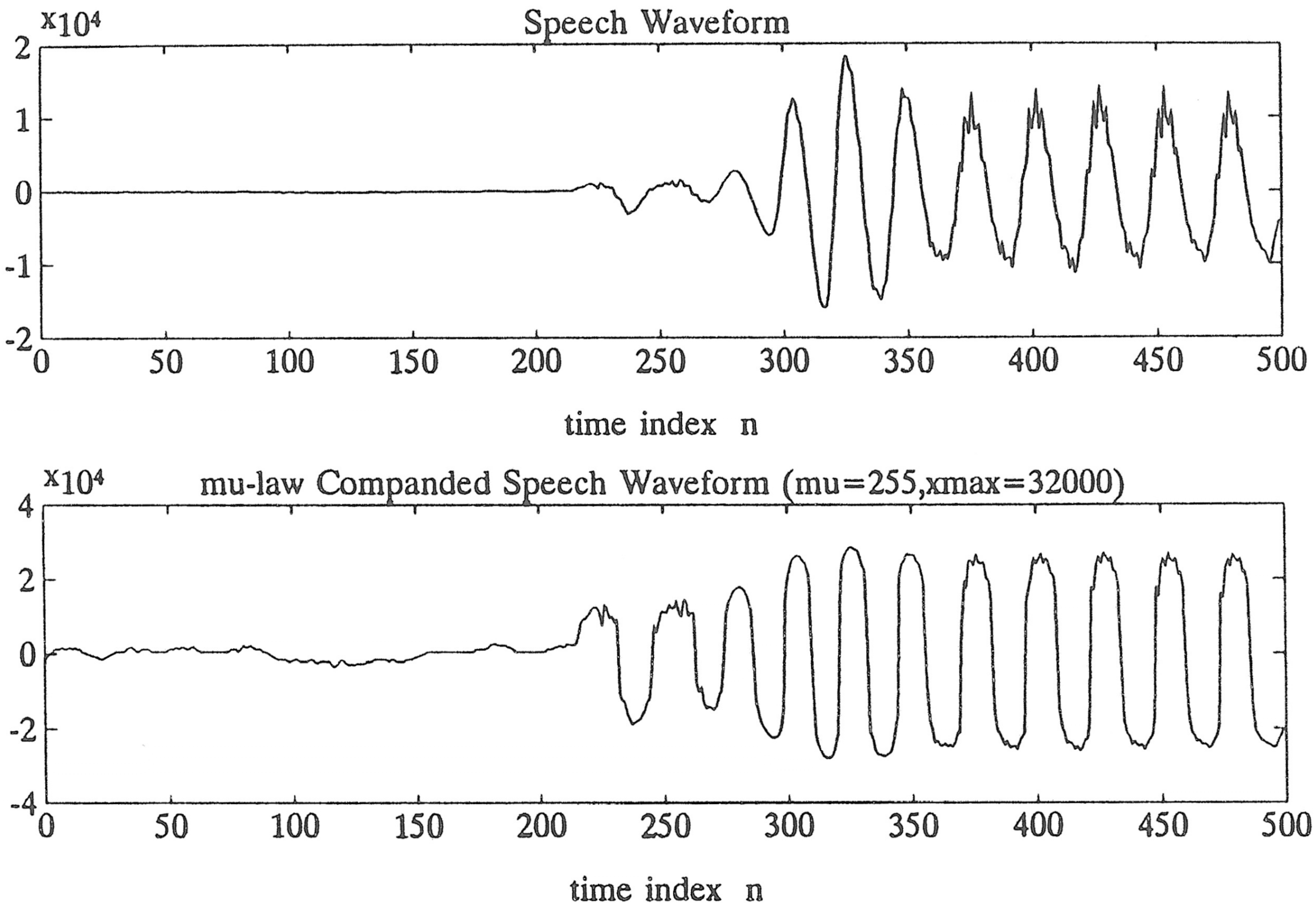


$$G(x) = G_{\max} \frac{\log_e\left(1 + \frac{\mu|x|}{x_{\max}}\right)}{\log_e(1+\mu)} \operatorname{sgn}(x)$$

typically, $\mu = 255$

$$\operatorname{sgn}(x) = \begin{cases} 1, \text{ if } x \geq 0 \\ -1, \text{ if } x < 0 \end{cases}$$

μ-law Compression – 2



Perceptual audio coding

Techniques

Principles in low bit rate coding – 1

- Digital coding at high bit rates is dominantly waveform-preserving, i.e., the amplitude-versus-time waveform of the decoded signal approximates that of the input signal.
  - the basic error criterion of codec design is the difference signal between input and output waveform
- At lower bit rates, facts about the production and perception of audio signals have to be included in coder design.
  - the error criterion has to be in favor of an output signal that is useful to human receiver rather than favoring an output signal that follows and preserves the input waveform
- Basically, an efficient source coding algorithm will
  - remove redundant components of source signal by exploiting correlations between its samples
  - remove components which are irrelevant to the ear.

Principles in low bit rate coding – 2

- The dependence of auditory perception on frequency and the perceptual tolerance of errors can directly influence encoder designs
  - *noise-shaping techniques* can shift coding noise to frequency bands where that noise is not of perceptual importance
  - the noise shifting must be dynamically adapted to the actual short-term spectrum in accordance with the signal-to-mask ratio
- the encoding process is controlled by the signal-to-mask ratio versus frequency curve from which the needed amplitude resolution in each critical band is derived
  - the bit allocation and rate in each critical band can be computed
- Given the bitrate for a complete masking distortion
  - the coding scheme will be perceptually transparent
  - the decoded signal is subjectively indistinguishable from a reference.

Principles in low bit rate coding – 3

- if the necessary bit rate for a completely masking of distortions is not possible,
  - the global masking threshold serves as a weighting function for spectral error
  - the resulting error spectrum will have the shape of the global masking threshold
- we cannot go to limits of masking or just noticeable distortion because
  - postprocessing may (e.g. filtering in equalizers) demask the noise,
  - our current knowledge about auditory masking is very limited => safety margin needed

Frequency domain coders

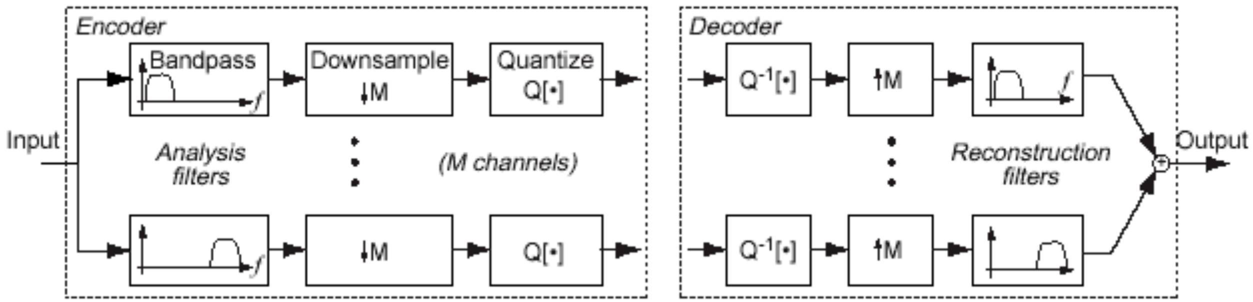
- The short-term spectral characteristics of the signal and the masking properties of the ear are exploited to reduce bitrate
  - Direct method for noise-shaping and suppression of frequency components that not need to be transmitted.
  - Source spectrum is split into frequency bands
  - Each frequency component is quantized separately
    => quantization noise associated with a particular band is contained within that band.
- The number of bits used to encode frequency components varies.
  - component being subjectively more important are quantized more finely, i.e. more bit allocated
- A dynamic bit allocation controlled by the spectral short-term envelope of the source signal is needed.
  - information transmitted to the decoder as side information

Subband coding – 1

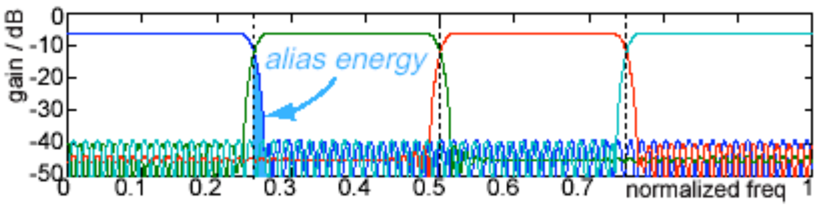
- The source signal is fed into an analysis filter bank consisting of *M* bandpass filters which are contiguous in frequency so that the set of subband signals can be recombined additively to produce the original signal or a close version thereof.
- Each filter output is critically decimated (i.e. sampled at twice the nominal bandwitdth) by a factor equal to *M*.
  => an aggregate number of subband samples that equals that in the source signal
- Each decimated filter output is quantized separately.
- In the receiver, the sampling rate of each subband is increased to that of the source signal by filling in the appropriate number of zeros samples.
- Interpolated subband signals appear at the bandpass outputs of the synthesis filter bank.

Subband coding – 2

- Quantize separately in different bands
  - quantization noise stay within band; gets masked



- Critical sampling: 1/M of spectrum per band
  - aliasing inevitable
  - Quadrature Mirror Filters: cancel with alias of adjacent bands



Polyphase Filter Bank – 1

- Characteristics
  - Lossy (even without quantization)
  - Fairly simple with reasonable resolution

- What It Does:
  - Divides input signal into equal width sub-bands.
  - Sub-bands overlap a lot, introduces error for analyzing.

- MP3 Specific
  - Input signal size is 32 samples which produces 32 sub-bands.
  - Vital part of Layers 1,2,3
  - Examples of subband coding
    - ISO/MPEG Audio Coding, Layers I and II

Polyphase Filter Bank – 2

$$s_t[i] = \sum_{k=0}^{63} \sum_{j=0}^{7} M[i][k] * (C[k+64j] * x[k+64j])$$

where:

i is the subband index and ranges from 0 to 31,

$s_t[i]$ is the filter output sample for subband i at time t, where t is an integer multiple of 32 audio sample intervals,

$C[n]$ is one of 512 coefficients of the analysis window defined in the standard,

$x[n]$ is an audio input sample read from a 512 sample buffer, and

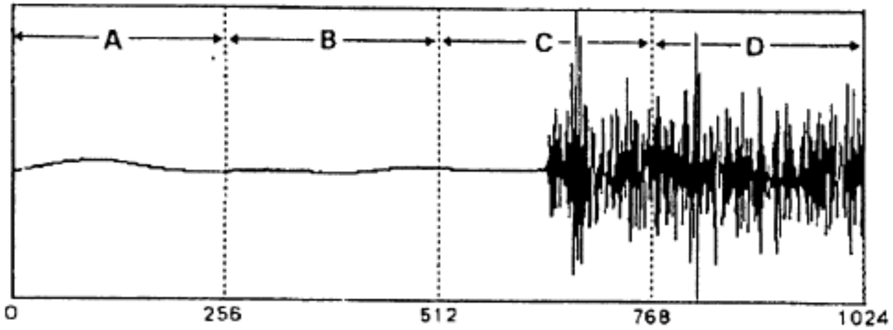
$M[i][k] = \cos\left[\frac{(2*i+1)*(k-16)*\pi}{64}\right]$ are the analysis matrix coefficients.

Pre-echoes – 1

- Crucial part in frequency domain coding of audio signals is the appearance of *pre-echoes*.
  - for example, a silent period is followed be a percussive sound, such as from castanets or triangles, within the same coding block
    => comparably large instantaneous quantization errors
    => pre-echoes can become distinctively audible, especially at low at low bit rates with comparably high error contributions
  - pre-echoes can be masked by the time domain effect of pre-masking if the time spread is of short length (in the order of a few milliseconds)
    => pre-echoes can be avoided by using blocks of short lengths. However, a larger percentage of the total bit rate is required for the transmission of side information if the blocks are too short.
  - a solution to this problem is to switch between block sizes of different lengths

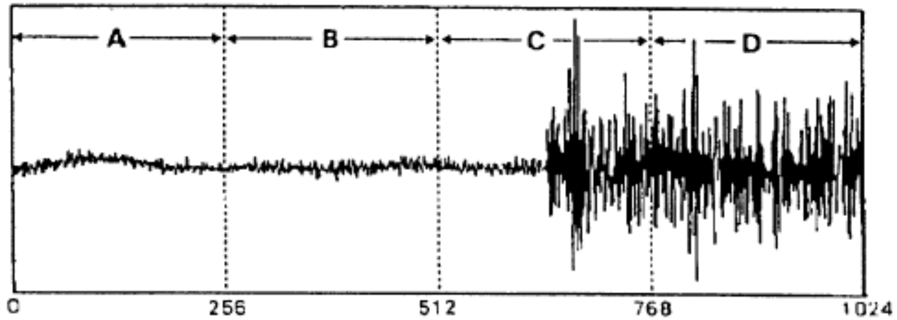
Pre-echoes – 2

original signal



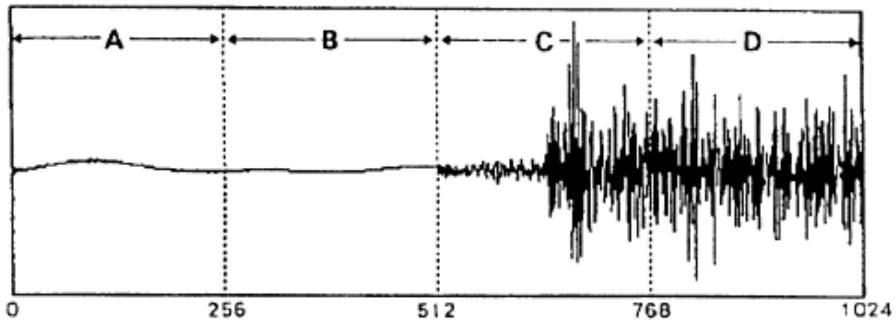
(a)

pre-echo (audible)



(b)

pre-echo (non audible)



(c)

Transform coding

- A block of input samples is linearly transformed via a discrete transform into a set of transform coefficients.
- These coefficients are then quantized and transmitted in digital form to the decoder.
- In the decoder an inverse transform maps the signal back into the time domain.
- Typical transforms are the discrete Fourier Transform (DFT) or the discrete cosine transform (DCT), calculated via FFT, and modified versions thereof.
- Discrete transforms can be viewed as filter banks. The finite lengths of its bandpass impulse responses may be so-called block boundary effects.
  => State-of-the-art coders employ a *modified DCT* (MDCT) with overlapping analysis blocks which can essentially eliminate these effects.

Modified Discrete Cosine Transform

- Characteristics
  - lossless, when there is no quantization
  - complex with good resolution
  - optimized for audio
- What it does
  - divides output of PQF into 18 subbands per input subband
    - 32 x 18 = 576 subbands
  - attempt to correct some error from PQF's subband overlapping
- MDCT transform splits each subband sequence further in frequency content to achieve a high frequency resolution.
  - It is specific for MP3 (MPEG1 – Layer III)

Discrete Cosine Transform

- DCT

$$C_k(n) = h(2M-1-n)\sqrt{\frac{2}{M}}\cos[\frac{\pi}{M}(k+k_0)(n+\frac{M+1}{2})]$$

  - Ck = impulse response
  - h = low pass prototype filter
  - M = num. of band
  - k = the order of the filter

- h(n) is restricted to a rectangular window of length M

$$h_{dct}(n) = \begin{cases} 1 \ \text{ for } \frac{\mathrm{M}}{2} \le n \le \frac{3M}{2} \\ 0 \ \text{ for otherwise} \end{cases}$$

Modified Discrete Cosine Transform

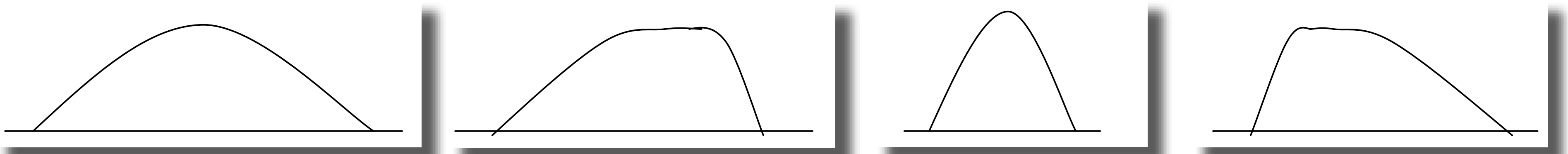
- MDCT
  - h(n) has a longer length

$$h_{MDCT}(n) = \sin[\frac{\pi}{2M}(n+0.5)] \ \text{for n} = 0..2\text{M - 1}$$

- Reduces the spectral leakage between channels by overlapping 50%

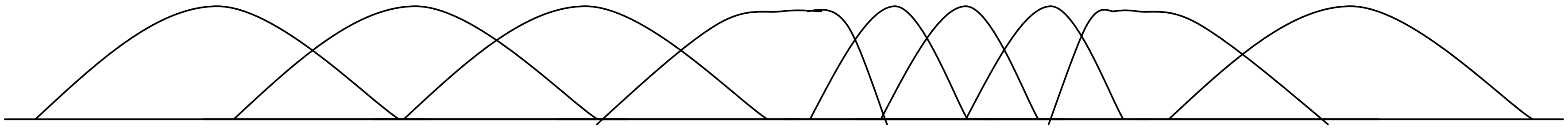
- **Hybrid (Subband/Transform) Coding**
  - Combinations of discrete transform and filter bank implementations
  - Different frequency resolutions can be provided at different frequencies in flexible way by using a cascade of a filterbank (with its short delays) and a linear MDCT transform (*hybrid filterbank*).
    - MDCT transform splits each subband sequence further in frequency content to achieve a high frequency resolution.

Windowing and overlapping



Long window Start window Short window Stop window

- MDCT analyzes data in blocks of length 18 or 6 samples.
  - This is 26 and 12 samples taking into account 50% overlapping.
- Overlapping reduces inconsistencies and misalignment between compressed sections.
- Supports mixed usage of small and large blocks
  - Provides better frequency resolution where it is needed
  - Mixes require transition blocks.



Perceptual audio coding

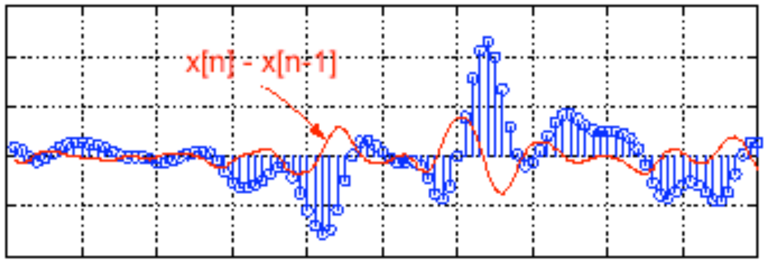
Additional techniques

Joint stereo

- The correlation between left and right channels is exploited to further reduce overall bitrate
  - For low frequencies: impossible to identify source position
    - mono is enough
  - For high frequencies: the identification of source position is based on amplitude envelope
    - mono spectrum
    - two modulations for amplitudes

Redundant information

- **Redundancy removal is lossless**
- **Signal correlation implies redundant information**
  - e.g. if $x[n] = x[n-1] + v[n]$
    $x[n]$ has a greater amplitude range → more bits than $v[n]$
  - sending $v[n] = x[n] - x[n-1]$ can reduce amplitude, hence bitrate



  - 'white noise' sequence has no redundancy

- **Problem: separating unique & redundant**

Optimal coding

- **Shannon information:**
  An unlikely occurrence is more 'informative'

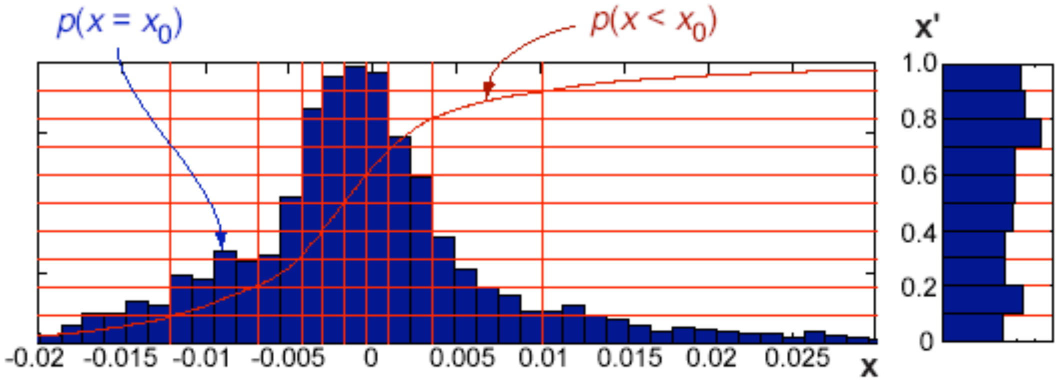
$p(A) = 0.5 \quad p(B) = 0.5$	$p(A) = 0.9 \quad p(B) = 0.1$
ABBBBAAABBABBABBABB	AAAAABBAAAAAABAAAAB
A, **B** equiprobable	**A** is expected; **B** is 'big news'

  - **Information in bits** $I = -\log_2(probability)$
    - clearly works when all possibilities equiprobable

- **Opt. bitrate ➔ token length = entropy** H = E[I]
  - i.e. equal-length tokens are equally likely
- **How to achieve this?**
  - transform signal to have uniform pdf
  - nonuniform quantization for equiprobable tokens
  - variable-length tokens ➔ Huffman coding

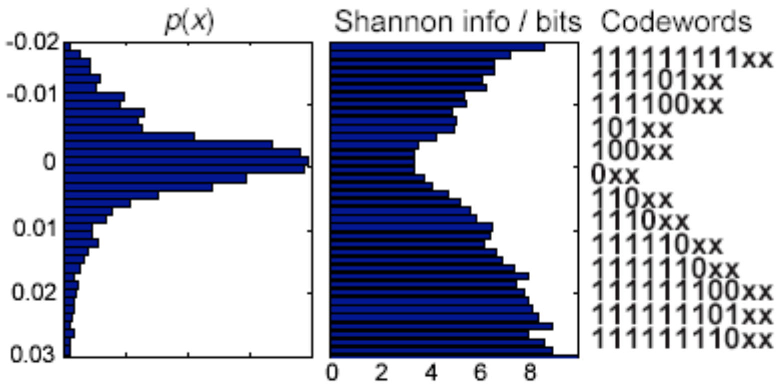
Quantization for optimum bitrate

- **Quantization should reflect pdf of signal:**



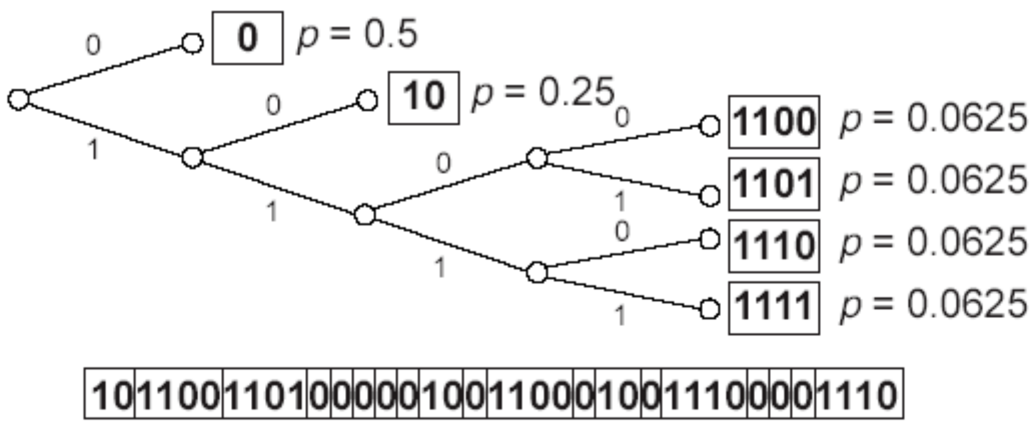
  - cumulative pdf $p(x < x_0)$ maps to uniform x'
  - or: nonuniform quantization bins

- **Or, codeword length per Shannon** $-\log_2(p(x))$**:**



Huffman coding

- Variable-length bit sequence tokens
  - → can code unequally probable events
- Tree-structure for unambiguous decoding:



- Can build tables to approximate arbitrary distributions
- Eliminates irrelevance .. within limits
- Problem: very probable events ➔ short tokens

Bit reservoir

- **Problem**:
  - A frame with little audio interest may require few bits to encode
    - with a constant bitrate these bits may be unused
  - A frame with substantial audio interest may require more bits to encode
    - with a constant bitrate the audio quality may decrease
- **Solution**:
  - Allow frames to give to or take from a reservoir
  - Each frame that save spaces, allows subsequent frames to store bits if needed
    - typical situation of a silence (few bits) followed by an attack (many bits)