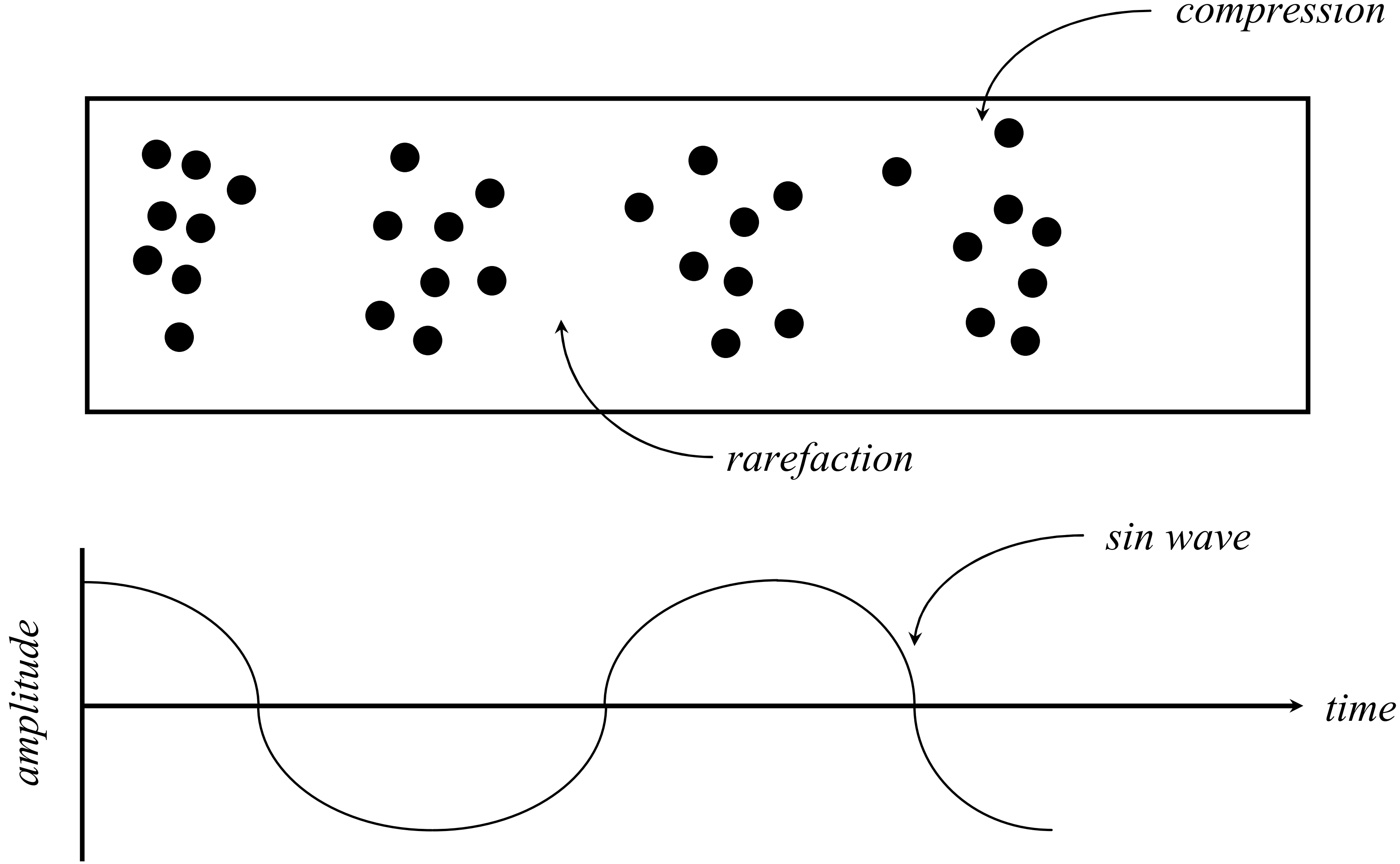
Audio Fundamentals

Audio Fundamentals

- Acoustics is the study of sound
  - *Generation, transmission, and reception of sound waves*
  - *Sound wave - energy causes disturbance in a medium*
- Example is striking a drum
  - *Head of drum vibrates => disturbs air molecules close to head*
  - *Regions of molecules with pressure above and below equilibrium*
  - *Sound transmitted by molecules bumping into each other*

Sound Waves



Sending/Receiving

- Receiver

  A microphone placed in sound field moves according to pressures exerted on it

  Transducer transforms energy to a different form (e.g., electrical energy)

- Sending

  A speaker transforms electrical energy to sound waves

Signal Fundamentals

- Pressure changes can be periodic or aperiodic
- Periodic vibrations

  cycle - *time for compression/rarefaction*
  cycles/second - *frequency measured in hertz (Hz)*
  period - *time for cycle to occur (1/frequency)*
- Frequency ranges

  barametric pression is 10^{-6} Hz
  cosmic rays are 10^{22} Hz
  human perception [0, 20kHz]

Wave Lengths

- Wave length is distance sound travels in one cycle
  - *20 Hz is 56 feet*
  - *20 kHz is 0.7 inch*
- Bandwidth is frequency range
- Transducers cannot linearly produce human perceived bandwidth
  - *Frequency range is limited to [20 Hz, 20 kHz]*
  - *Frequency response is not flat*

Measures of Sound

- Sound level is a logarithmic scale

  SPL = 10 log (pressure/reference) decibels (dB)

  where reference is $2*10^{-4}$ dyne/cm^2

  0 dB SPL - essentially no sound heard

  35 dB SPL - quiet home

  70 dB SPL - noisy street

  120 dB SPL - discomfort

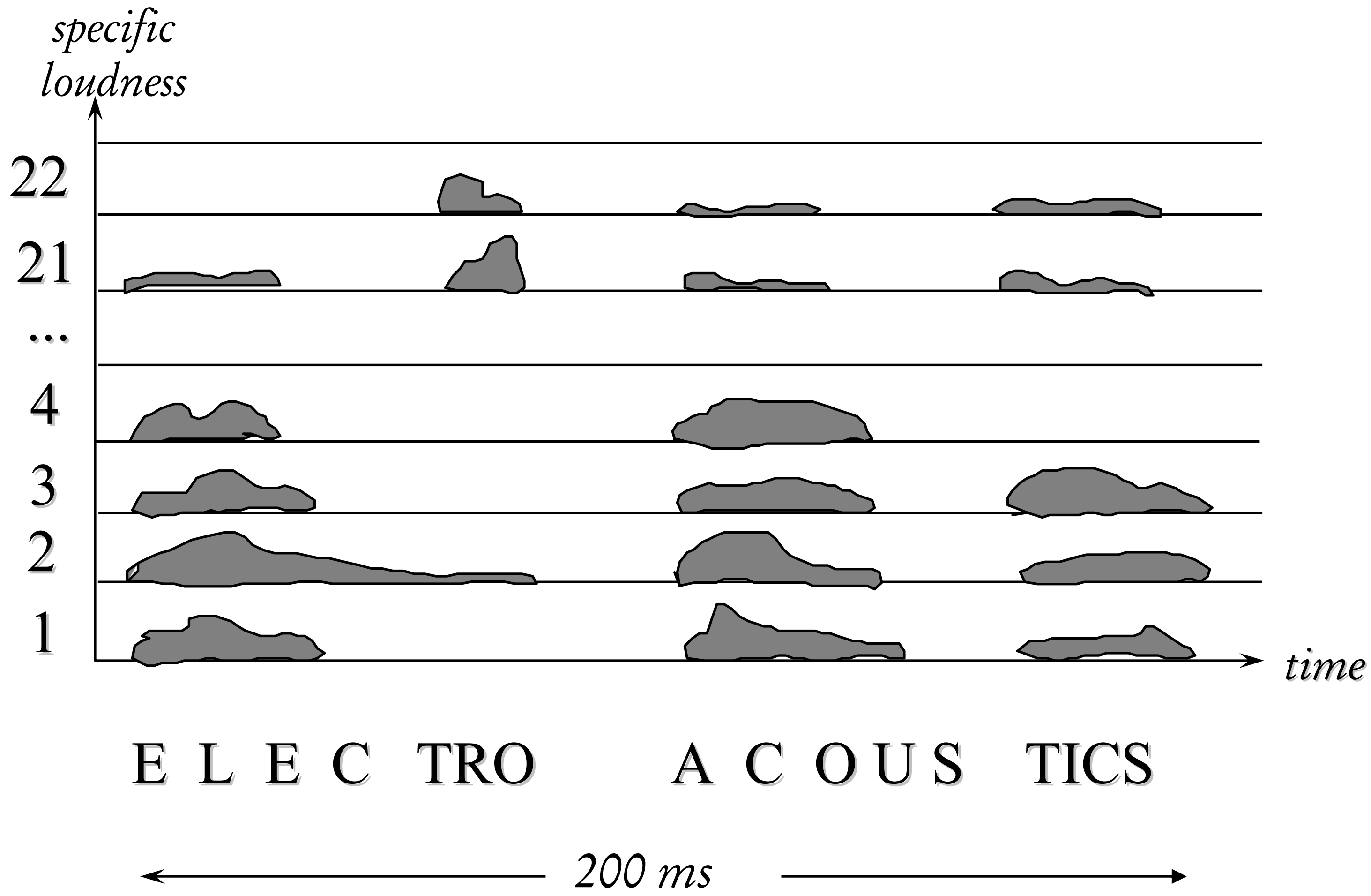
Sound Phenomena

- Sound is typically a combination of waves
  - *Sin wave is fundamental frequency*
  - *Other waves added to it to create richer sounds*
  - *Musical instruments typically have fundamental frequency plus overtones at integer multiples of the fundamental frequency*
- Waveforms out of phase cause interference
- Other phenomena
  - *Sound reflects off walls if small wave length*
  - *Sound bends around walls if large wave lengths*
  - *Sound changes direction due to temperature shifts*

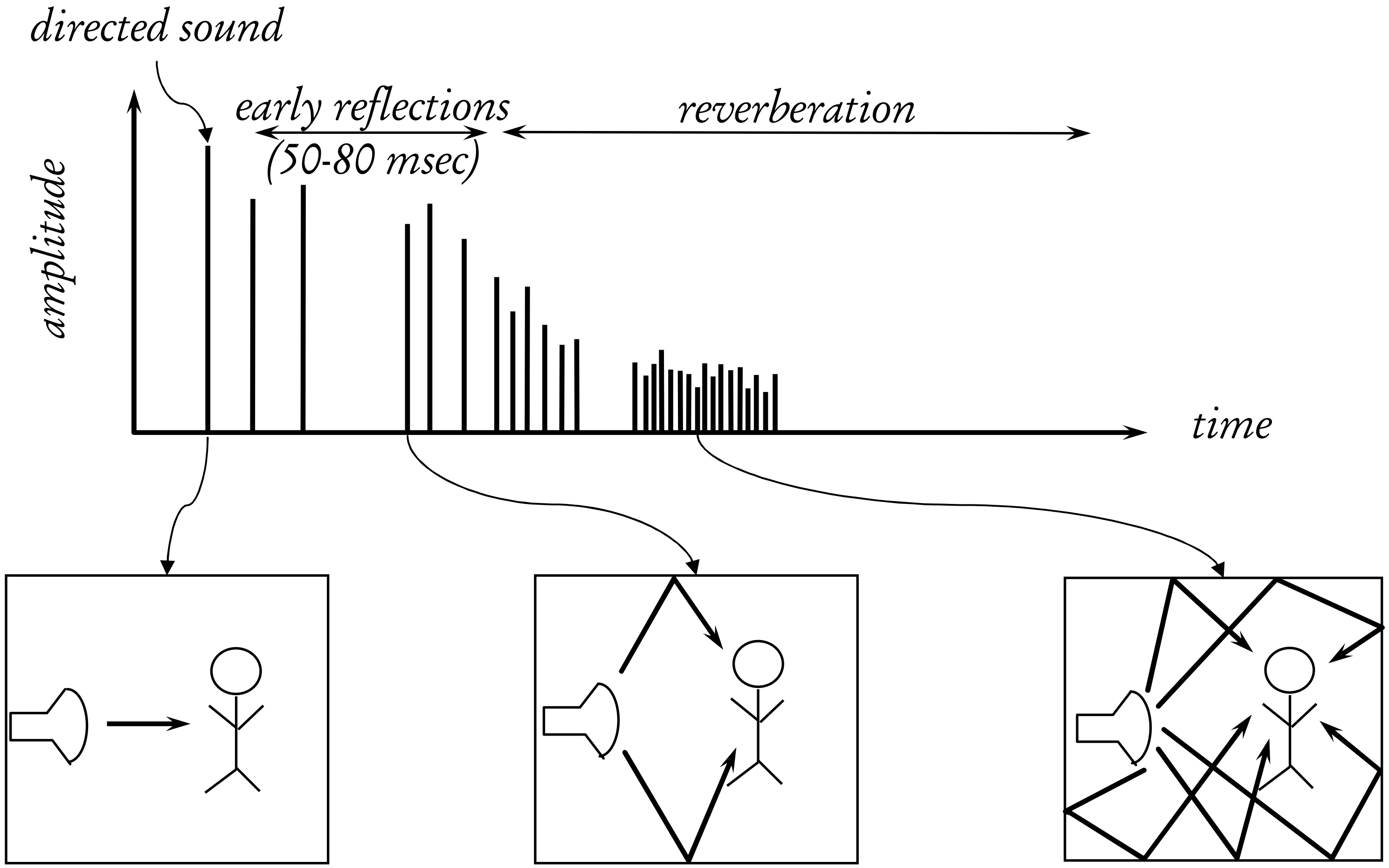
Human Perception

- Speech is a complex waveform
  - *Vowels and bass sounds are low frequencies*
  - *Consonants are high frequencies*
- Humans most sensitive to low frequencies
  - *Most important region is 2 kHz to 4 kHz*
- Hearing dependent on room and environment
- Sounds masked by overlapping sounds

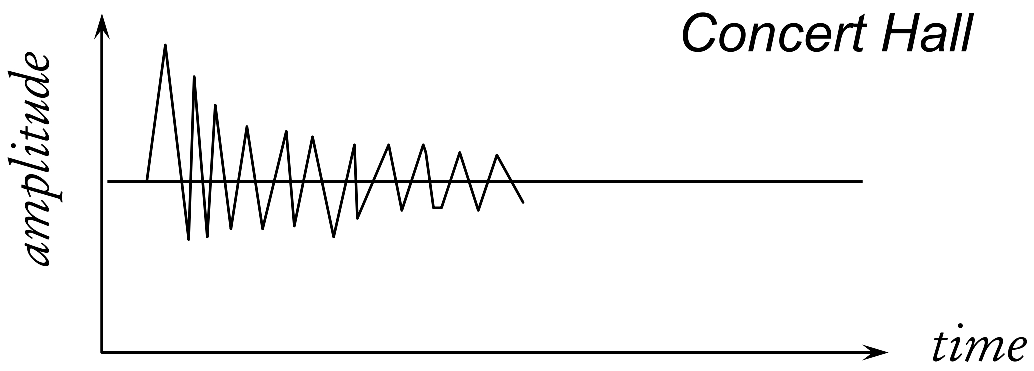
Critical Bands

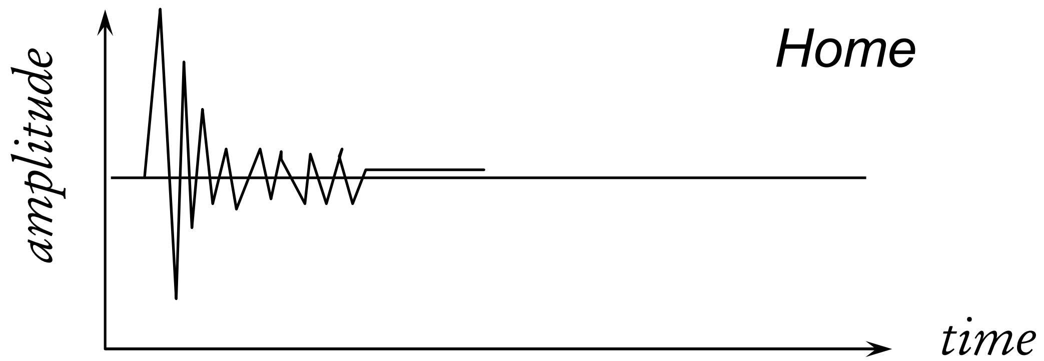


Sound Fields

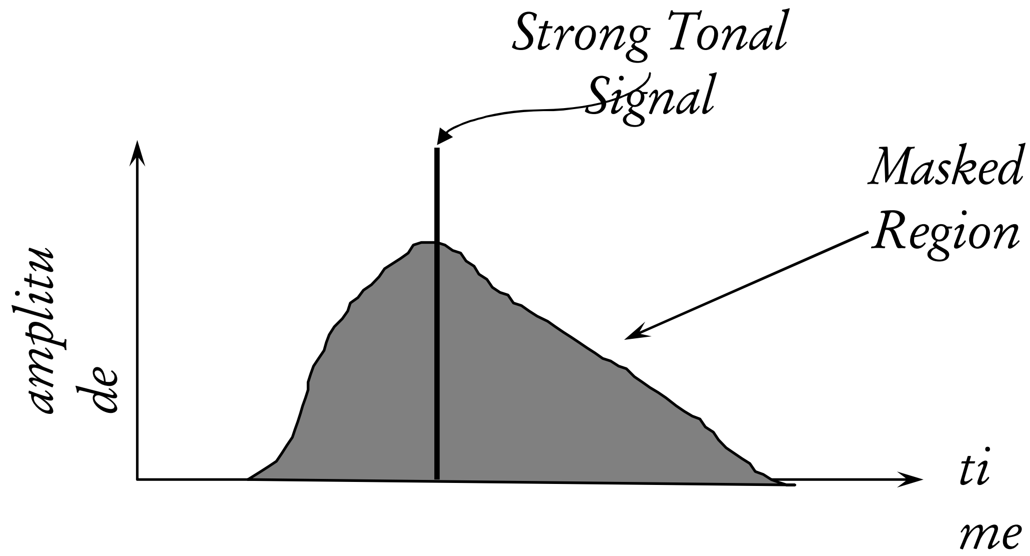


Impulse Response

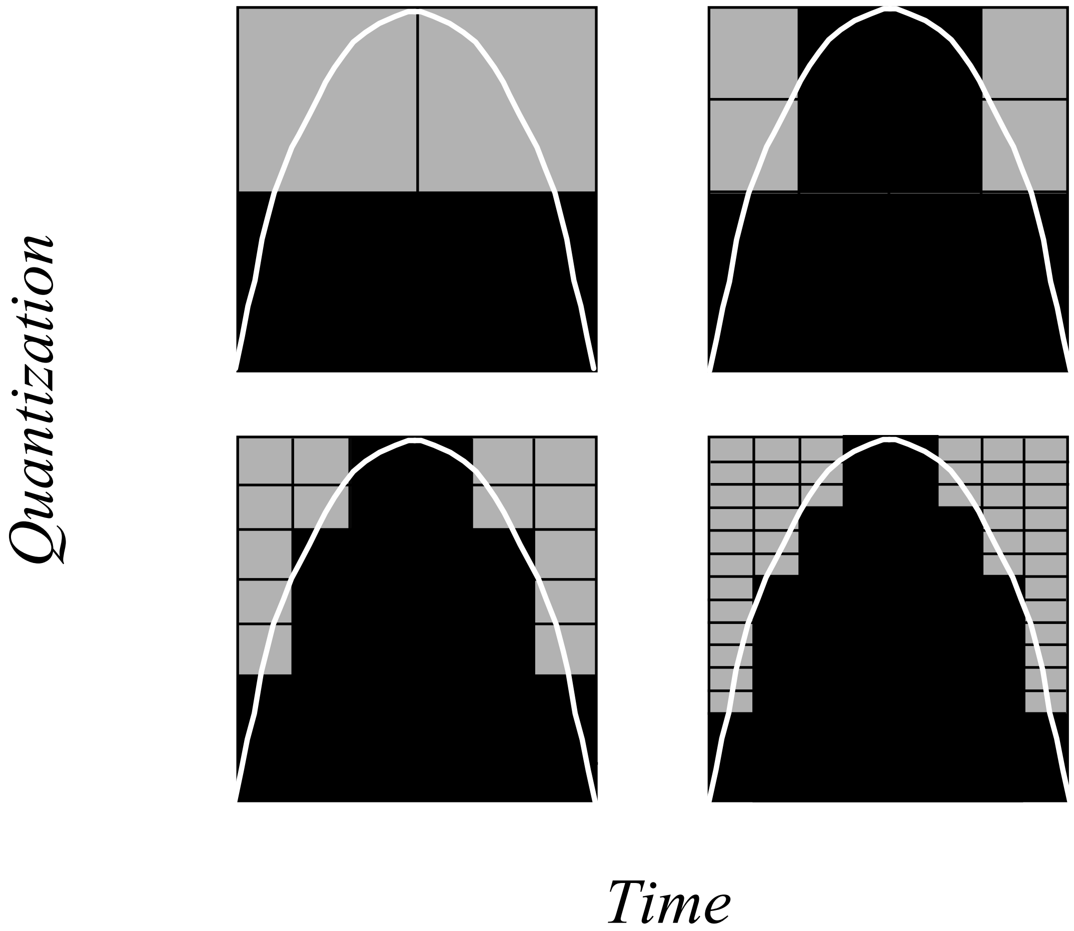




Audio Noise Masking



Audio Sampling



Audio Representations

Optimal sampling frequency is twice the highest frequency to be sampled (Nyquist Theorem)

Format	Sampling Rate	Bandwidth	Frequency Band
Telephony	8 kHz	3.2 kHz	200-3400 Hz
Teleconferencing	16 kHz	7 kHz	50-7000 Hz
Compact Disk	44.1 kHz	20 kHz	20-20,000 Hz
Digital Audio Tape	48 kHz	20 kHz	20-20,000 Hz

Jargons/Standards

- Emerging standard formats

  8 kHz 8-bit U-LAW mono
  22 kHz 8-bit unsigned linear mono and stereo
  44 kHz 16-bit signed mono and stereo
  48 kHz 16-bit signed mono and stereo

- Actual standards

  G.711 - A-LAW/U-LAW encodings (8 bits/sample)
  G.721 - ADPCM (32 kbs, 4 bits/sample)
  G.723 - ADPCM (24 kbs and 40 kbs, 8 bits/sample)
  G.728 - CELP (16 kbs)
  GSM 06.10 - 8 kHz, 13 kbs (used in Europe)
  LPC (FIPS-1015) - Linear Predictive Coding (2.4kbs)
  CELP (FIPS-1016) - Code excitied LPC (4.8kbs, 4bits/sample)
  G.729 - CS-ACELP (8kbs)
  MPEG1/MPEG2, AC3 - (16-384kbs) mono, stereo, and 5+1 channels

Audio Packets and Data Rates

- Telephone uses 8 kHz sampling

  ATM uses 48 byte packets ➔ 6 msecs per packet

  RTP uses 160 byte packets ➔ 20 msecs per packet

- Need many other data rates

  ≤ 30 kbs ➔ audio over 28.8 kbs modems

  32 kbs ➔ good stereo audio is possible

  56 kbs or 64 kbs ➔ conventional telephones

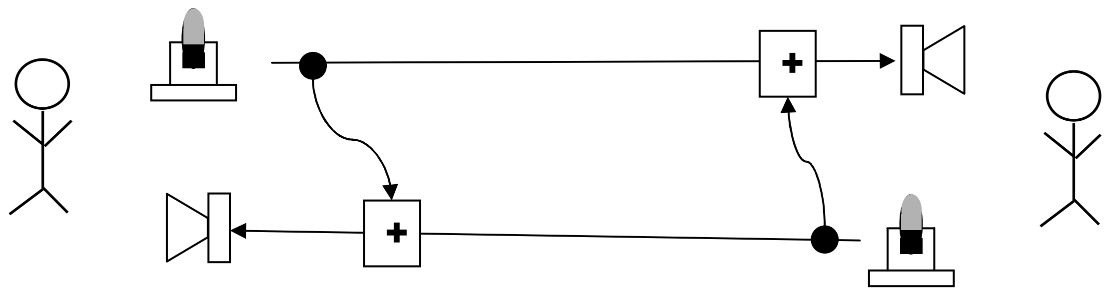
  128 kbs ➔ MPEG1 audio

  256 - 384 kbs ➔ higher quality MPEG/AC3 audio

Discussion

- Higher quality
  - *Filter input*
  - *More bits per sample (i.e. 10, 12, 16, etc.)*
  - *More channels (e.g. stereo, quadraphonic, etc.)*
- Digital processing
  - *Reshape impulse response to simulate a different room*
  - *Move perceived location from which sound comes*
  - *Locate speaker in 3D space using microphone arrays*
  - *Cover missing samples*
  - *Mix multiple signals (i.e. conference)*
  - *Echo cancellation*

Interactive Time Constraints



- Maximum time to hear own voice: *100 msec*
- Maximum round-trip time: *300 msec*

Importance of Sound

- Passive viewing (e.g. film, video, etc.)
  - *Very sensitive to sound breaks*
  - *Visual channel more important (ask film makers!)*
  - *Tolerate occasional frame drops*
- Video conferencing
  - *Sound channel is more important*
  - *Visual channel still conveys information*
  - *Some people report that video teleconference users turn off video*
  - *Need to create 3D space and locate remote participants in it*

Producing High Quality Audio

- Eliminate background noise
  - *Directional microphone gives more control*
  - *Deaden the room in which you are recording*
  - *Some audio systems will cancel wind noise*
- One microphone per speaker
- Keep the sound levels balanced
- Sweeten sound track with interesting sound effects

Audio *-vs-* Video

- Some people argue that sound is easy and video is hard because data rates are lower

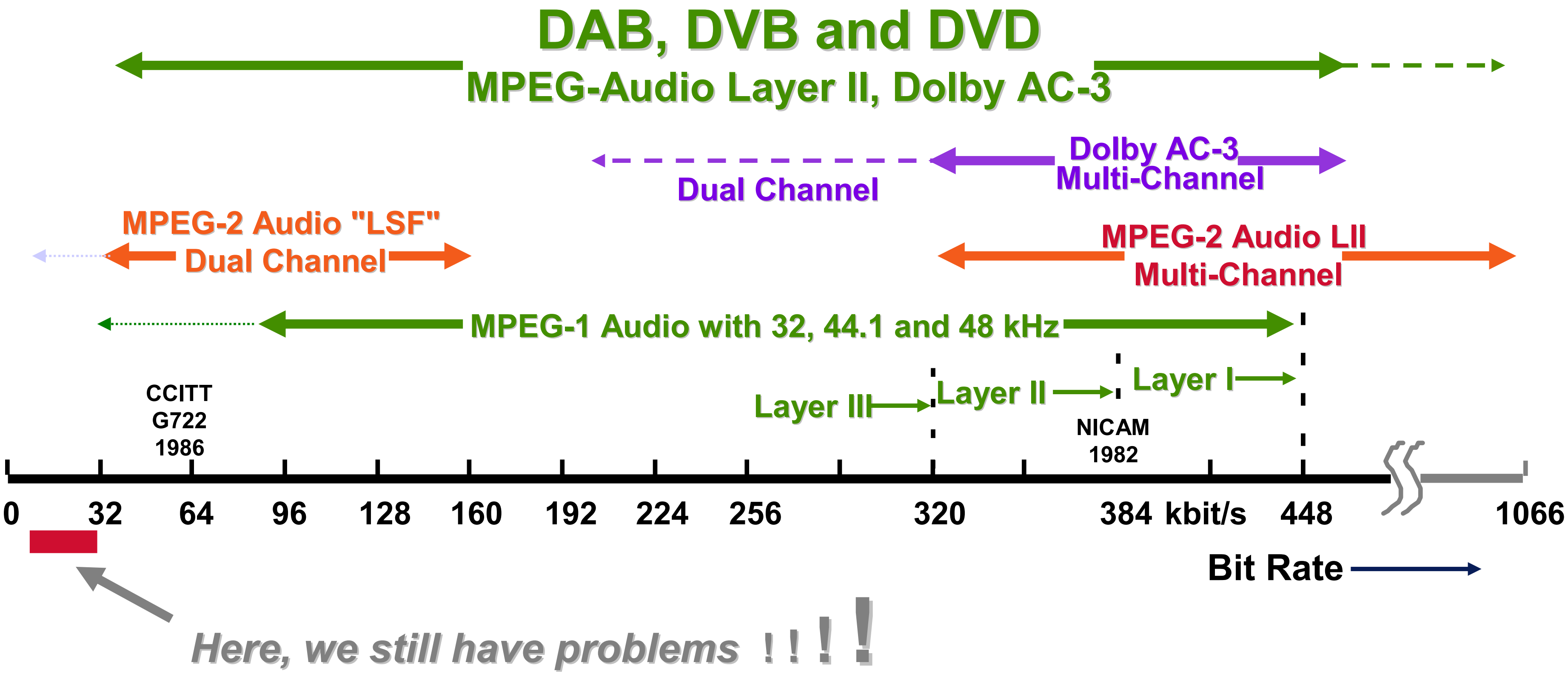
  Not true $\Rightarrow$ audio is every bit as hard as video, just different!

- Computer Scientists will learn about audio and video just as we learned about printing with the introduction of desktop publishing

Audio

- Some techniques for audio compression:
  - ADPCM
  - LPC
  - CELP

Digital Audio for Transmission and Storage
Target Bit Rates for MPEG Audio and Dolby AC-3



Possible candidates to solve these problems:

- MPEG-2 AAC and MPEG-4 Audio
- Internet Radio Audio Package Manufacturers

History of MPEG-Audio

- MPEG-1 Two-Channel coding standard (Nov. 1992)
- MPEG-2 Extension towards Lower-Sampling-Frequency (LSF) (1994)
- MPEG-2 Backwards compatible multi-channel coding (1994)
- MPEG-2 Higher Quality multi-channel standard (MPEG-2 AAC) (1997)
- MPEG-4 Audio Coding and Added Functionalities (1999, 2000)

Audio

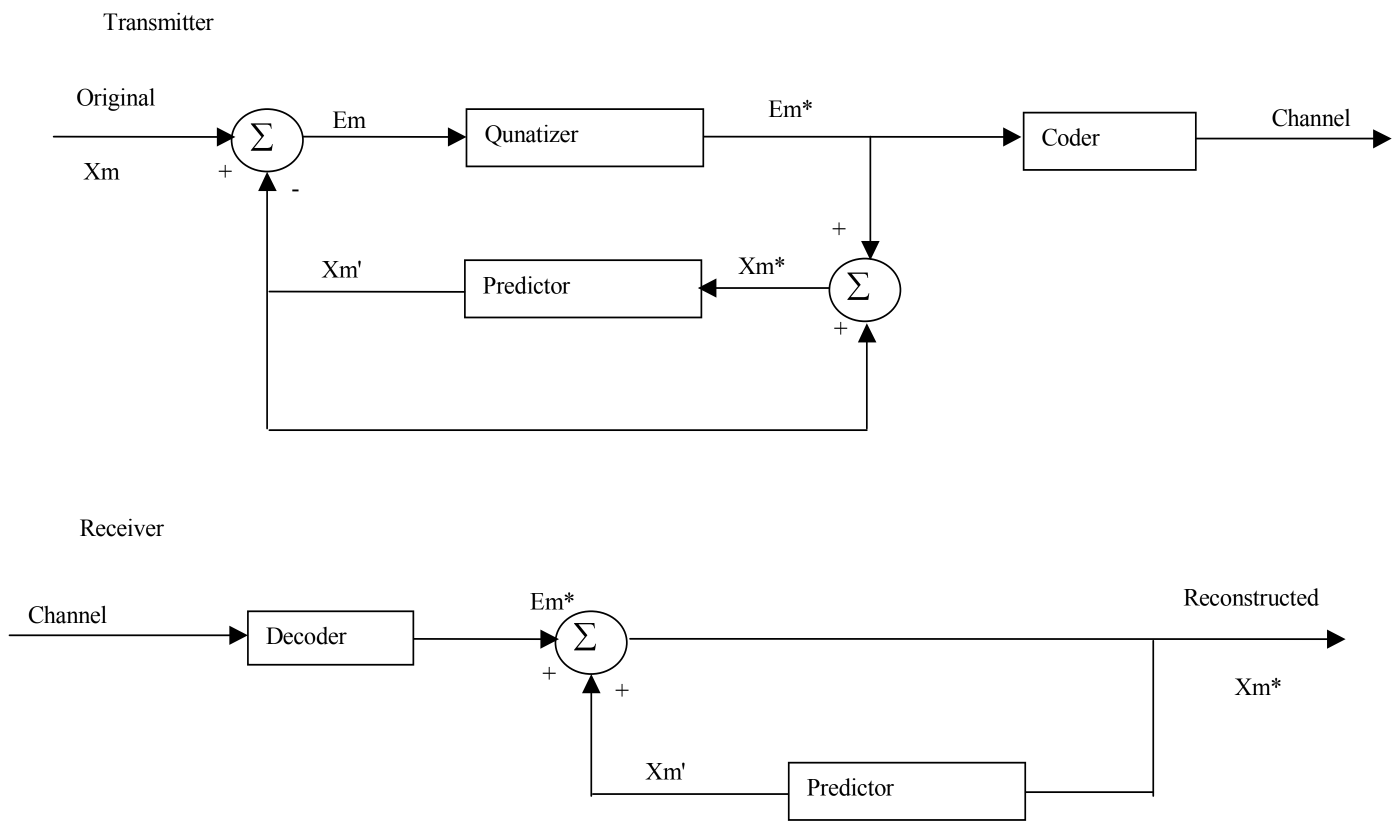
- ADPCM -- Adaptive Differential Pulse Code Modulation

  ADPCM allows for the compression of PCM encoded input whose power varies with time.

  Feedback of a reconstructed version of the input signal is subtracted from the actual input signal, which is quantised to give a 4 bits output value.

  This compression gives a 32 kbit/s output rate.

Audio


Transmitter
Original
Xm
+
-
Em
Qunatizer
Em*
Coder
Channel
+
Xm'
Predictor
Xm*
+
Receiver
Channel
Decoder
Em*
+
+
Reconstructed
Xm*
Xm'
Predictor

Audio

- LPC -- Linear Predictive Coding

  The encoder fits speech to a simple, analytic model of the vocal tract. Only the parameters describing the best-fit model is transmitted to the decoder.

  An LPC decoder uses those parameters to generate synthetic speech that is usually very similar to the original.

  LPC is used to compress audio at 16 Kbit/s and below.

Audio -- CELP

- CELP -- Code Excited Linear Predictor

  CELP does the same LPC modeling but then computers the errors between the original speech and the synthetic model and transmits both model parameters and a very compressed representation of the errors.

  The result of CELP is a much higher quality speech at low data rate.

Digital Audio Recapture

- Digital audio parameters
  - Sampling rate
  - Number of bits per sample
  - Number of channels (1 for mono, 2 for stereo, etc.)

- Sampling rate
  - Unit -- Hz or sample per second
  - Sampling rate is measured per channel
    - For stereo sound, if the sampling rate is 8KHz, that means 16K samples will be obtained per second

Sampling Rate & Applications

Sampling Rate	Applications
8KHz	Telephony standard
11.025KHz	Web applications
22KHz	Mac sampling rate
32 KHz	Digital radio
44.1 KHz	CD quality audio
48 KHz	DAT (Digital Audio Tape)

- Higher sampling rate → better quality → larger file

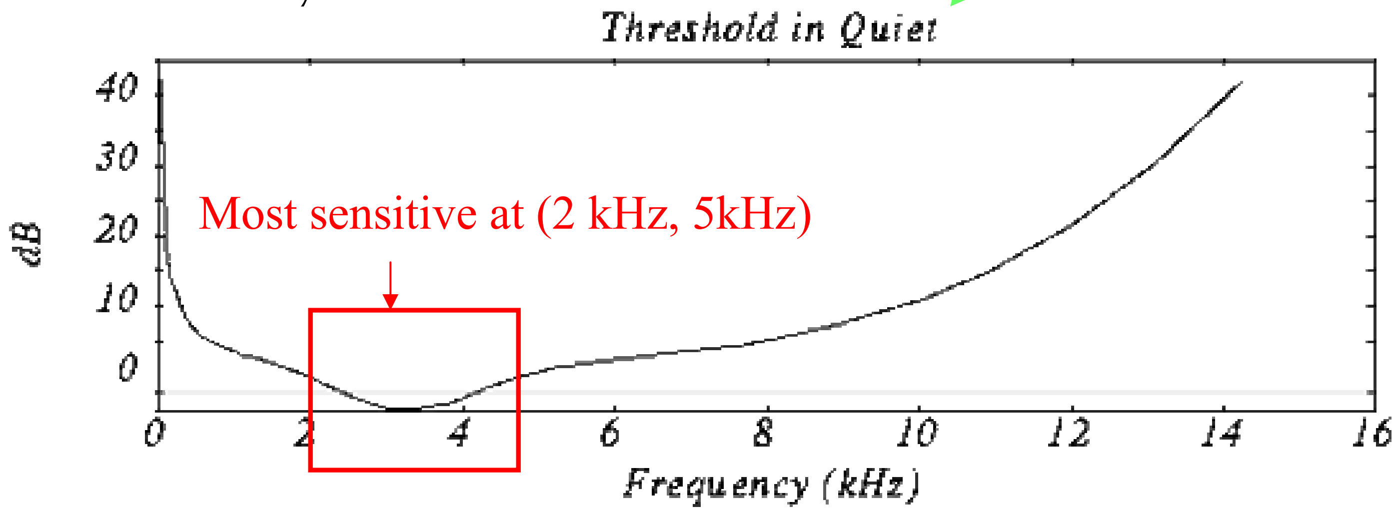
Speech Compression

- Speech compression technologies
  - Silence suppression – detect the “silence”, only code the “laud” part of the speech (currently a technique combined with other methods to increase the compression ratio)
  - Differential PCM – a simple method
  - Utilize the speech model
    - Linear Predictive Coding (LPC): fits signal to speech model and transmits parameters of model
    - Code Excited Linear Predictor (CELP): Same principle as LPC, but instead of transmitting parameters, transmit error terms in codebook
  - Quality of compresses audio
    - LPC -- Computer talking alike
    - CELP – Better quality, audio conference

Audio compression

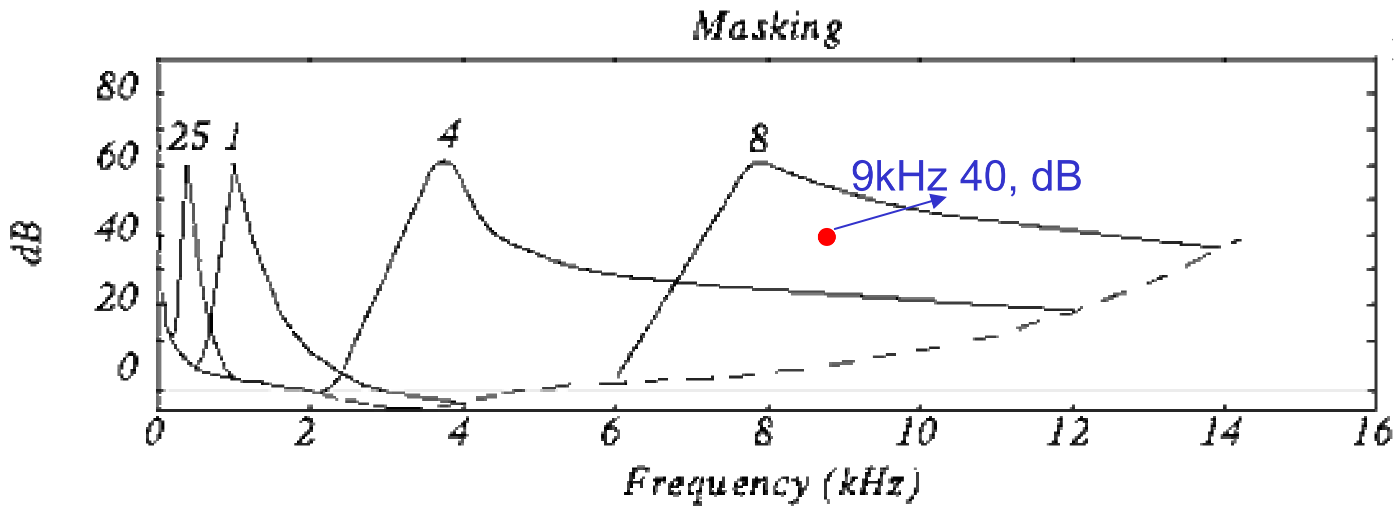
- Audio vs. Speech
  - Higher quality requirement for audio
  - Wider frequency range of audio
- Psychoacoustics model
  - Sensitivity of human ears

Result of ear sensitivity test*



Principle of Audio Compression (1)

- Psychoacoustics model (cont.)

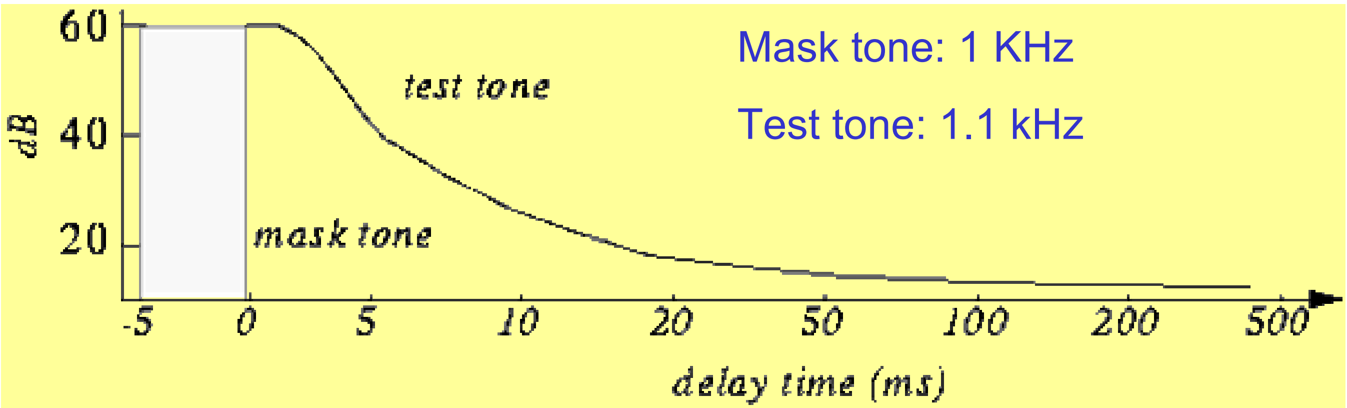


Frequency masking at different tones (60 dB) *

Thinking: if there is a 8 kHz signal at 60 dB, can we hear another 9 kHz signal at 40 dB?

Principle of Audio Compression (2)

- Psychoacoustics model (cont.)
  - Critical bandwidth – the range of frequencies that are affected by the masking tone beyond a certain degree
    - Critical bandwidth increases with the frequency of masking tone
    - For masking frequency less than 500 Hz, critical bandwidth is around 100 Hz; for frequency greater than 500 Hz, the critical bandwidth increase linearly in a multiple of 100 Hz
  - Temporal masking -- If we hear a loud sound, then it stops, it takes a little while until we can hear a soft tone nearby



Principle of Audio Compression (3)

- Audio compression – Perceptual coding
  - Take advantage of psychoacoustics model
  - Distinguish between the signal of different sensitivity to human ears
    - Signal of high sensitivity – more bits allocated for coding
    - Signal of low sensitivity – less bits allocated for coding
  - Exploit the frequency masking
    - Don't encode the masked signal (range of masking is 1 critical band)
  - Exploit the temporal masking
    - Don't encode the masked signal
- Audio coding standard – MPEG audio codec
  - Have three layers, same compression principle

MPEG Audio Codec (1)

- Basic facts about MPEG audio coding
  - Perceptual coding
  - Support 3 sampling rates
    - 32 kHz – Broadcast communication
    - 44.1 kHz – CD quality audio
    - 48 KHz – Professional sound equipment
  - Supports one or two audio channels in one of the following four modes:
    - Monophonic -- single audio channel
    - Dual-monophonic -- two independent channels (similar to stereo)
    - Stereo -- for stereo channels that share bits
    - Joint-stereo -- takes advantage of the correlations between stereo channels

MPEG Audio Codec (2)

- Procedure of MPEG audio coding

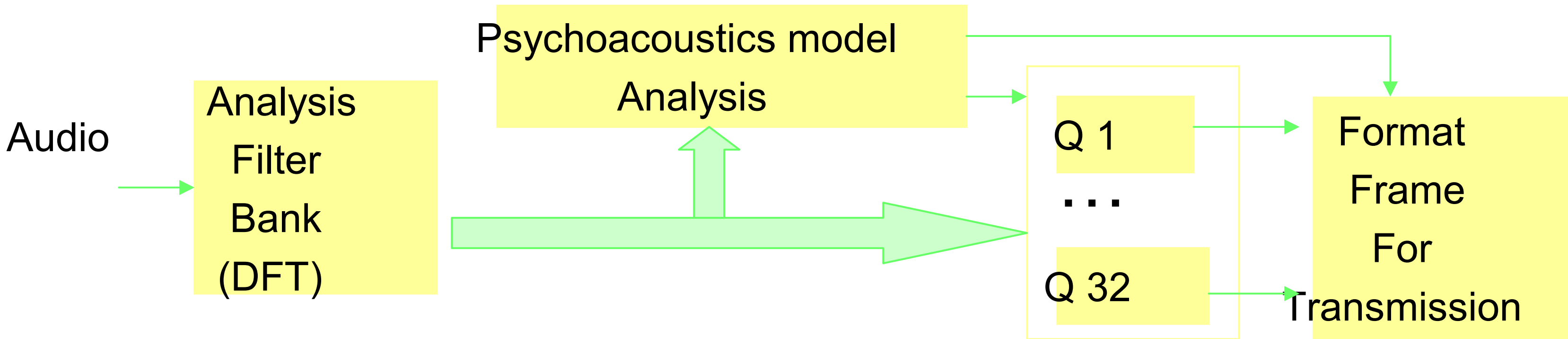
  Apply DFT (Discrete Fourier Transform) → decomposes the audio into frequency subbands that approximate the 32 critical bands (sub-band filtering)

  Use psychoacoustics model in bit allocation

    If the amplitude of signal in band is below the masking threshold, don't encode

    Otherwise, allocate bits based on the sensitivity of the signal

  Multiplex the output of the 32 bands into one bitstream



MPEG Audio Codec (3)

- MPEG audio frame format

  Audio data is divided into frames, each frame contains 384 samples

  After subband filtering, each frame (384) is decomposed into 32 bands, each band has 12 samples

  The bitstream format of the output MPEG audio is:

Header	SBS format	12x32 subband samples (SBS)	Ancillary Data

  The minimum encoding delay is determined by the frame size and the number of frames accumulated for encoding

MPEG Audio Codec (4)

- MPEG layers

  MPEG defines 3 layers for audio. Basic compression model is same, but codec complexity increases with each layer

  The popular MP3 is MPEG audio codec layer 3

  Layer 1:

    DCT type filter apply to one frame

    Equal frequency spread per band

    Frequency masking only

  Layer 2:

    Use three frames in filter (previous, current, next)

    Both frequency and temporal masking.

  Layer 3:

    Better critical band filter is used (non-equal frequencies)

    Better psychoacoustics model

    Takes into account stereo redundancy, and uses Huffman coder.

Perceptual Coding
of
Audio Signals – A Tutorial

What is Coding for?

- Coding, in the sense used here, is the process of reducing the bit rate of a digital signal.
- The coder input is a digital signal.
- The coder output is a smaller (lower rate) digital signal.
- The decoder reverses the process and provides (an approximation to) the original digital signal.

Historical Coder "Divisions":

- Lossless Coders
- vs.
- Lossy Coders

- Or

- Numerical Coders
- vs.
- Source Coders

Lossless Coding:

- Lossless Coding commonly refers to coding methods that are completely reversible, i.e. coders wherein the original signal can be reconstructed bit for bit.

Lossy Coding:

- Lossy coding commonly refers to coders that create an approximate reproduction of their input signal. The nature of the loss depends entirely on the kind of lossy coding used.

Source Coding:

- Source Coding can be either lossless or lossy.
- In most cases, source coders are deliberately lossy coders, *however*, this is not a restriction on the method of source coding. Source coders of a non-lossy nature have been proposed for some purposes.

Source Coding:

- Removes redundancies through estimating a model of the source generation mechanism. This model may be explicit, as in an LPC speech model, or mathematical in nature, such as the ”transform gain” that occurs when a transform or filterbank diagonalizes the signal.

Source Coding:

- Typically, the source coder users the source model to increase the SNR or reduce an other error metric of the signal by the appropriate use of signal models and mathematical redundancies.

Typical Source Coding Methods:

- LPC analysis (including dpcm and its derivatives and enhancements)
- Multipulse Analysis by Synthesis
- Sub-band Coding
- Transform Coding
- Vector Quantization

This list is not exhaustive

Well Known Source Coding Algorithms:

- Delta Modulation
- DCPM
- ADPCM
- G721
- G728
- LDCELP
- LPC-10E

Numerical Coding:

- Numerical coding is a almost always a lossless type of coding. Numerical coding, in its typical usage, means a coding method that uses abstract numerical methods to remove redundancies from the coded data.
- New Lossy Numerical coders can provide fine-grain bit rate scalability.

Common Numerical Coding Techniques:

- Huffman Coding
- Arithmetic Coding
- Ziv-Lempel (LZW) Coding
- This list is not exhaustive

Numerical Coding (cont.):

- Typically, numerical coders use “entropy coding” based methods to reduce the actual bit rate of the signal.
- Source coders most often use signal models to reduce the signal redundancy, and produce lossy coding systems.
- Both methods work by considering the source behavior.
- Both methods attempt to reduce the *Redundancy* of the original signal.

Perceptual Coding:

- Perceptual coding uses a model of the destination, i.e. the human being who will be using the data, rather than a model of the signal source.

- Perceptual coding attempts to remove parts of the signal that the human cannot perceive.

Perceptual Coding (cont.):

- Is a ***lossy*** coding method.

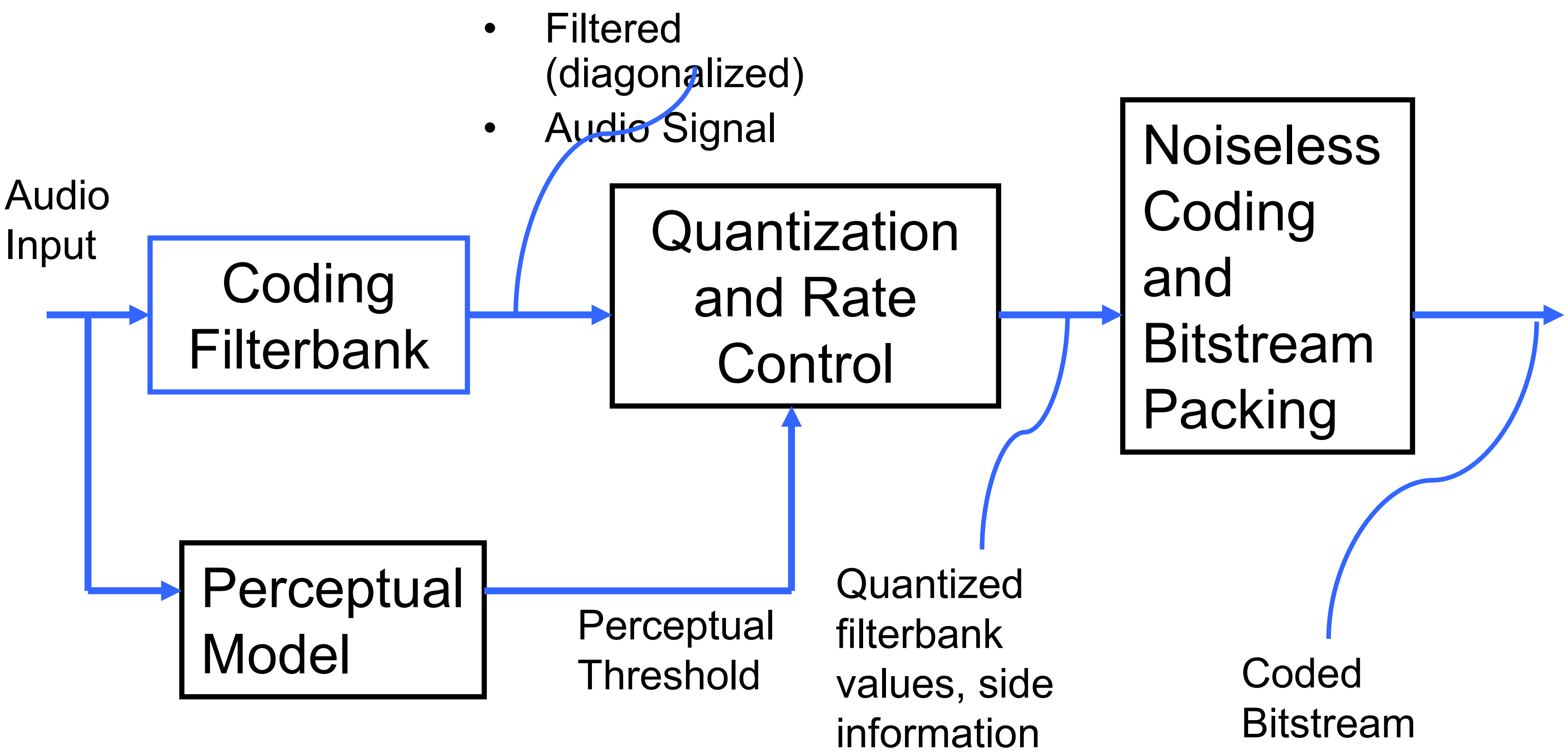
- The imperceptible information removed by the perceptual coder is called the
- ***irrelevancy***
- of the signal.
- In practice, most perceptual coders attempt to remove both ***irrelevancy*** and ***redundancy*** in order to make a coder that provides the lowest bit rate possible for a give audible quality.

Perceptual Coding (cont.):

- Perceptual coders will, in general, have a *lower* SNR than a source coder, and a higher perceived quality than a source coder of equivalent bit rate.

Perceptual Audio Coder

Block Diagram



Auditory Masking Phenomena:

- The “Perceptual Model”

What is Auditory Masking:

- The Human Auditory System (HAS) has a limited detection ability when a stronger signal occurs near (in frequency and time) to a weaker signal. In many situations, the weaker signal is imperceptible even under ideal listening conditions.

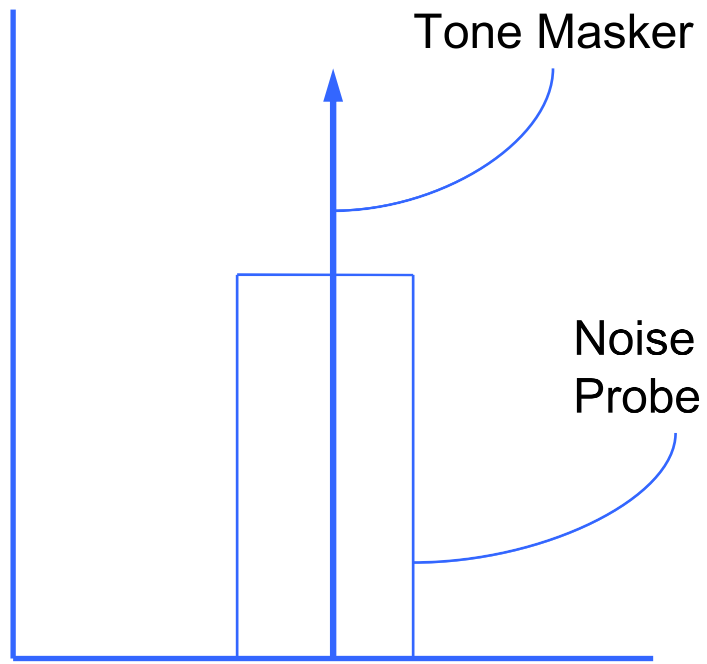
- Auditory Masking Phenomena (cont.)

First Observation of Masking:

- If we compare:

- Tone Masker
- to
- Tone Masker plus noise

- The energy of the 1-bark wide probe is 15.0 dB below the energy of the tone masker.



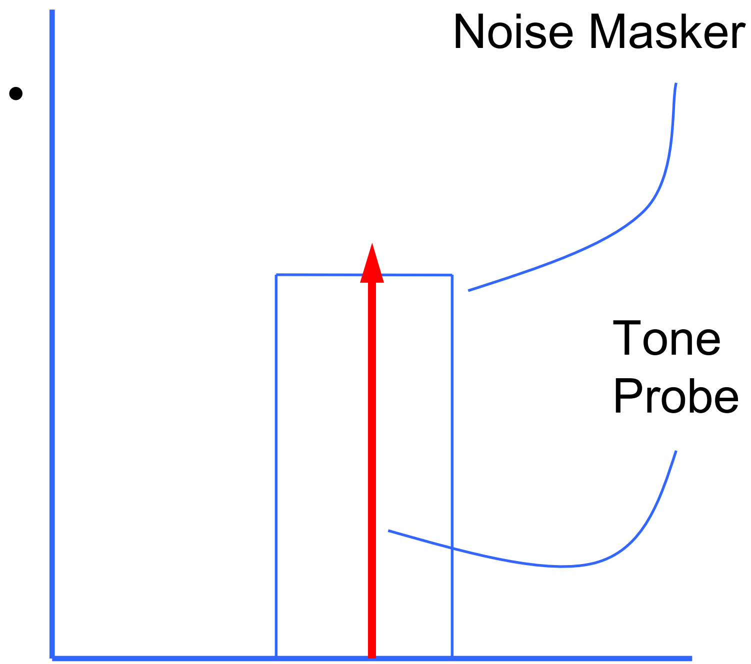
<u>THE NOISE IS AUDIBLE</u>

The Noise is *NOT* Masked!

- In this example, a masker to probe ratio of approximately 25 dB will result in complete masking of the probe.

2nd Demonstration of Masking:

- If we compare:
- Noise Masker
- to
- Noise Masker plus tone probe
- The energy of the 1-bark wide masker is 15 dB above the tone probe.



<u>The Tone is NOT Audible</u>

Auditory Masking Phenomena (cont.)

The Tone is *COMPLETELY* Masked

- In this case, a masker to probe ratio of approximately 5.5 dB will result in complete masking of the tone.

Auditory Masking Phenomena (cont.):

- There is an asymmetry in the masking ability of a tone and narrow-band noise, when that noise is within one critical band.
- This asymmetry is related to the short-term stability of the signal in a given critical bandwidth.

Critical Bandwidth?

- What's this about a ***critical bandwidth***?

- A critical bandwidth dates back to the experiments of Harvey Fletcher. The term critical bandwidth was coined later. Other people may refer to the "ERB" or equivelent rectangular bandwidth. They are all manifestations of the same thing.

- What is that?

A *critical band* or *critical bandwidth*

- is a range of frequencies over which the masking SNR remains more or less constant.

- For example, in the demonstration, any noise signal within +- .5 critical band of the tone will produce nearly the same masking behavior as any other, as long as their energies are the same.

Auditory Masking Phenomena (cont.)

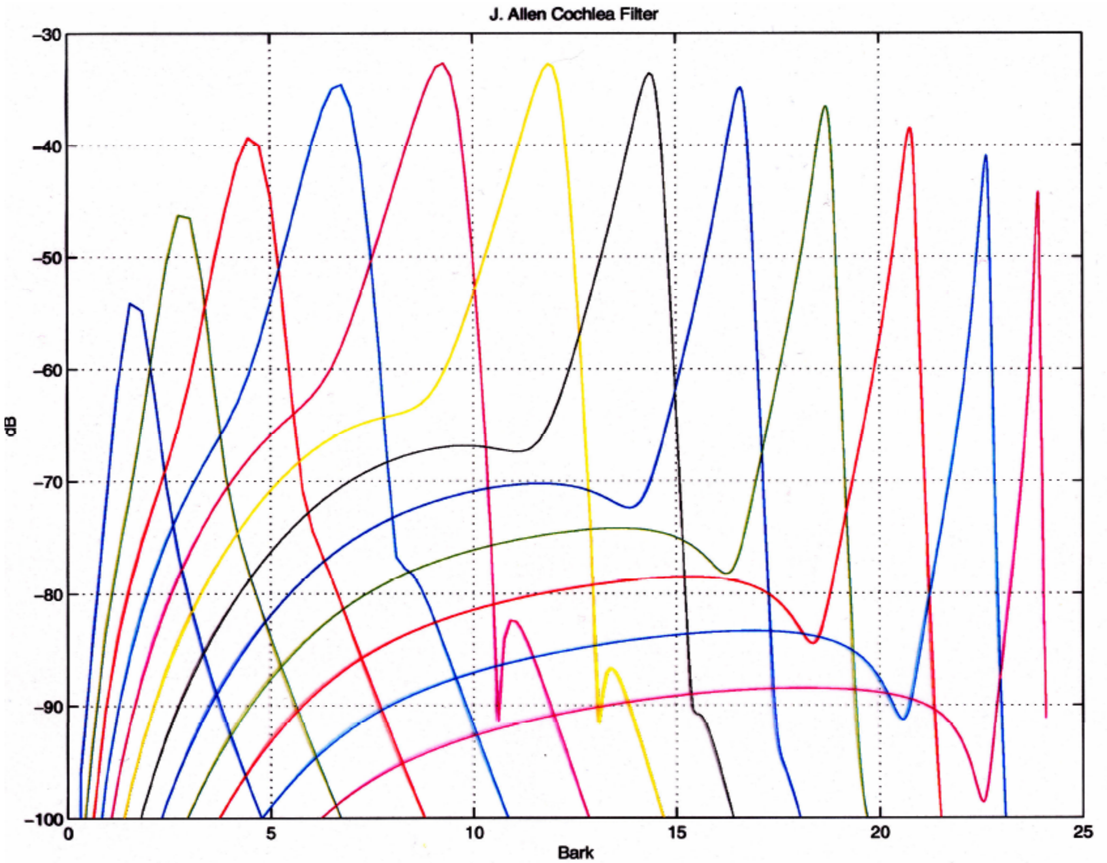
Auditory Filterbank:

- The mechanical mechanism in the human cochlea constitute a mechanical filterbank. The shape of the filter at any one position on the cochlea is called the ***cochlear filter*** for that point on the cochlea. A ***critical band*** is very close to the passband bandwidth of that filter.

ERB

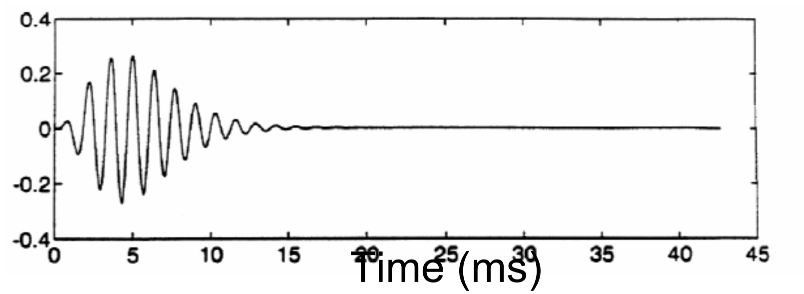
- A newer take on the bandwidth of auditory filters is the “Equivalent Rectangular Bandwidth”. It results in filters slightly narrower at low frequencies, and substantially narrower at mid and high frequencies.
- The “ERB scale”is not yet agreed upon.

J. Allen Cochlea Filters



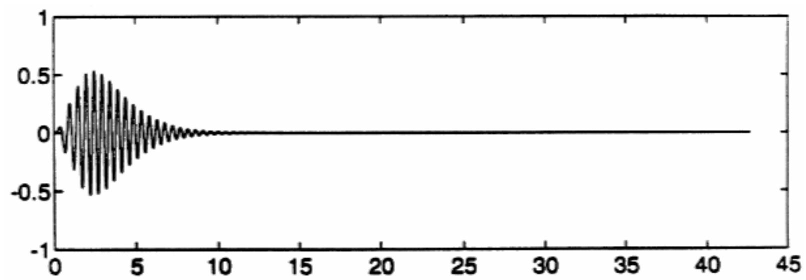
Two Example Cochlear Filters: Time-Domain Response

Impulse response, cochlear filter centered at 750 Hz



Time (ms)

Impulse response, cochlear filter centered at 2050 Hz



Time (ms)

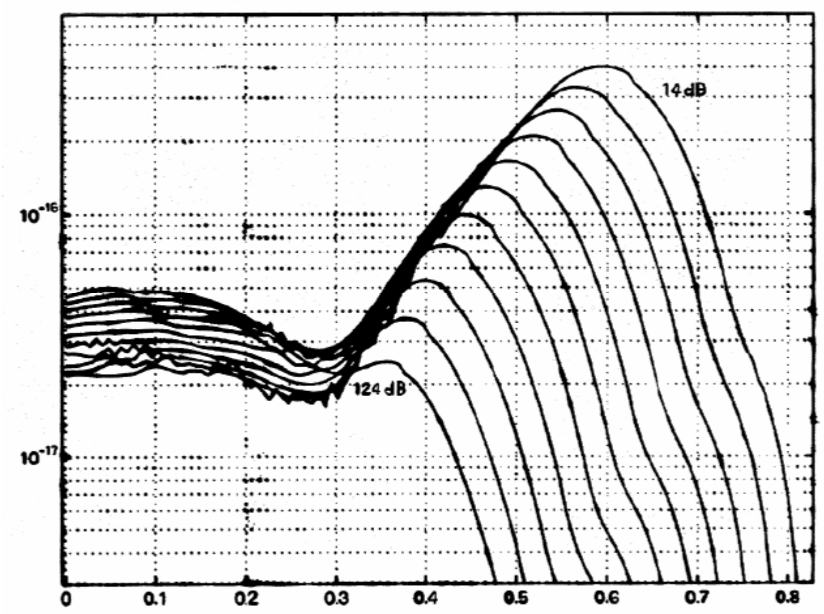
Auditory Masking Phenomena (cont.)

The Cochlear Filterbank

- At this time, it seems very likely that the cochlear filterbank consists of two part, a lowpass filter and a highpass filter, and that one filter is tuned via the action of outer hair cells.

- This tuning changes the overlap of the two filters and provides both the compression mechanism and the behavior of the upward spread of masking.

Neural Tuning for 5kHz tonal Stimulus
(14 – 124 dB SPL)



Distance from Stapes (cm)

Auditory Masking Phenomena (cont.)

The *Bark Scale*

- The bark scale is a standardized scale of frequency, where each “Bark” (named after Barkhausen) constitutes one critical bandwidth, as defined in Scharf’s work. This scale can be described as approximately equal-bandwidth up to 700Hx and approximately 1/3 octave above that point.

Auditory Masking Phenomena (cont.)

- A convenient and reasonably accurate approximation for conversion of frequency in Hz to Bark frequency is:

- B= 13.0 ARCTAN($\frac{0.76f}{1000}$)
- +
- 3.5 ARCTAN(($\frac{f}{7500}$)^2)

Auditory Masking Phenomena (cont.)

- The Bark scale is often used as a frequency scale over which masking phenomenon and the shape of cochlear filters are invariant. While this is not strictly true, this represents a good first approximation.

ERB's Again

- The ERB scale appears to provide a more invariant scale for auditory modelling. With the Bark scale, tone-masking-noise performance varies with frequency.

- With a good ERB scale, tone-masking-noise performance is fixed at about 25-30dB.

The Practical Effects of the Cochlear Filterbank in Perceptual Audio Coding:

- Describes spreading of masking energy in the frequency domain
- Explains the cause of pre-echo and the varying time dependencies in the auditory process
- Offers a time/frequency scale over which the time waveform and envelope of the audio signal can be examined in the cochlear domain.

The Spread of Masking in Frequency:

- The spread of masking in frequency is currently thought to be due to the contribution of different frequencies to the signal at a given point on the basilar membrane, corresponding to one cochlear filter.

Time vs. Frequency Response of Cochlear Filters

- The time extent, or bandwidth, of cochlear filters varies by at least a factor of 10:1 if not more.

- As a result, the audio coding filterbank must be able to accommodate changes in resolution of at least 10:1.

Time Considerations in Masking:

- Simultaneous Masking
- *Forward Masking* – Masking of a signal by a masker that precedes the masked (probe) signal
- *Backward Masking* – Masking of a probe by a masker that comes after the probe

Forward Masking:

- Forward masking of a probe by a masker exists both within the length of the impulse response of the cochlear filter, and beyond that range due to integration time constants in the neural parts of the auditory system.
- The length of this masking is >20ms, and is sometimes stated to be as long as several hundred milliseconds. In practice, the decay for post masker masking has two parts, a short *hangover* part and then a longer *decaying* part. Different coders take advantage of this in different ways.

Limits to Forward Masking

- Signals that have a highly coherent envelope across frequency may create low-energy times when coding noise can be unmasked, even when forward masking may be expected to work.

- For such signals, Temporal Noise Shaping, (TNS) was developed.

Backward Masking:

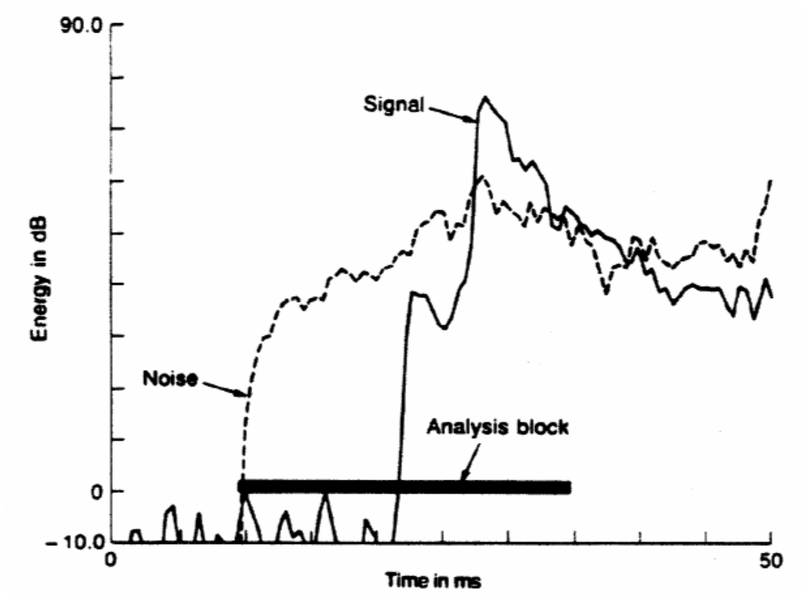
- Backward masking appears to be due to the length of the impulse response of the cochlear filter. At high frequencies, backward masking is less than 1ms for a trained subject who is sensitive to monaural time-domain masking effects. Subjects vary significantly in their ability to detect backwardly masked probes.

Effects of the Time Response of the Cochlear Filter on the Coder Filterbank:

- The short duration of backward masking is directly opposed to the desire to make the filterbank long in order to extract the signal redundancy. In successful low-rate audio coders, a switched filterbank is a necessity.

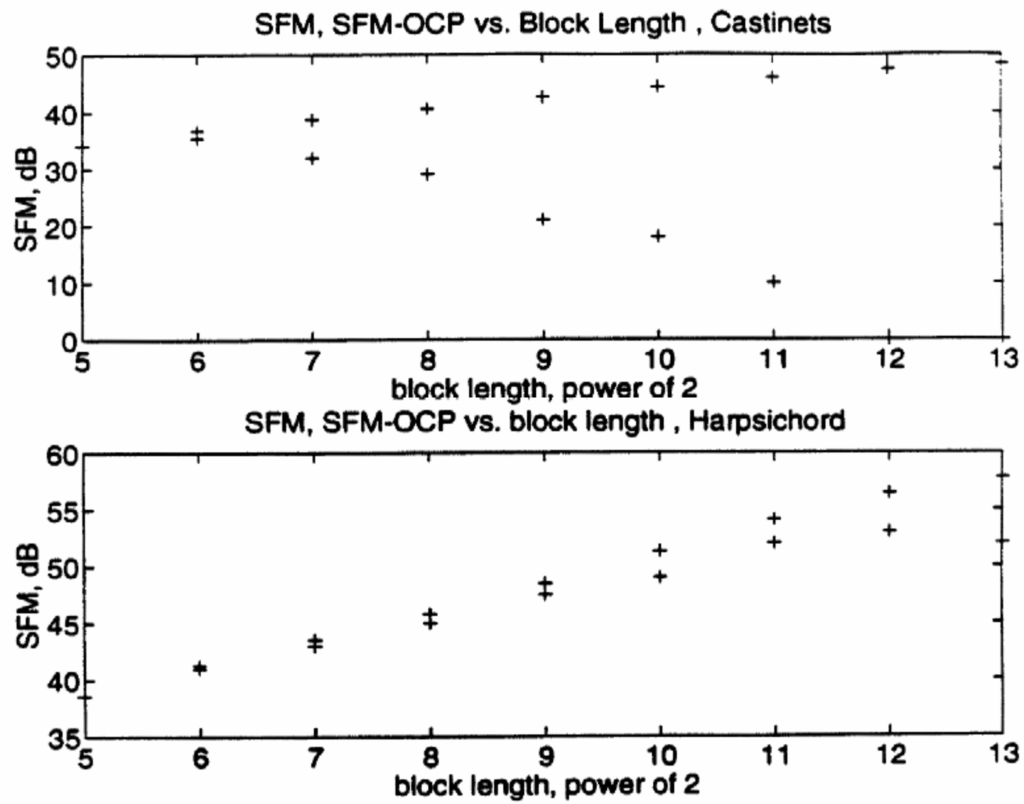
The Spread of Masking in Time:

- An example of how a filterbank can create a pre-echo.

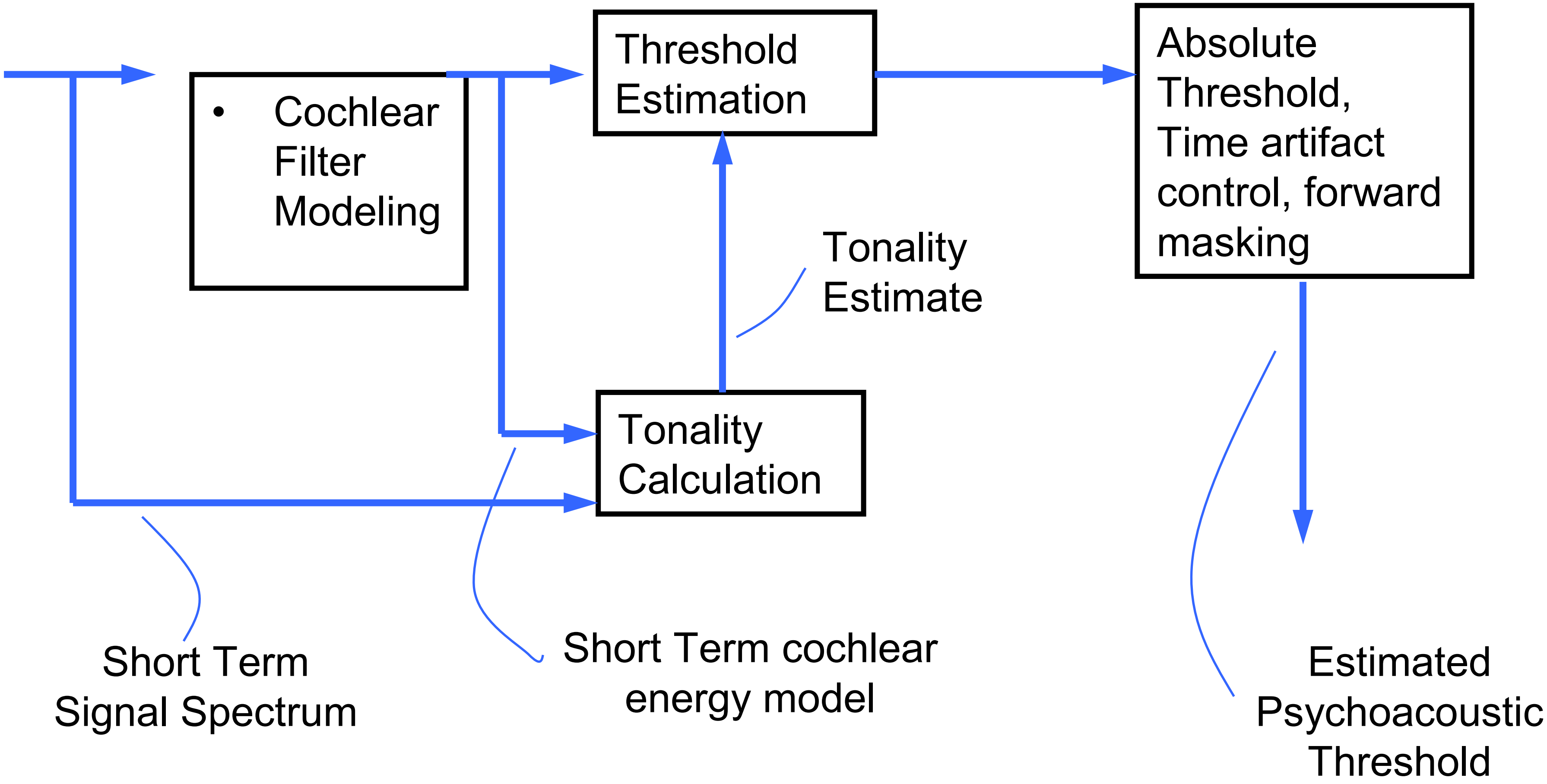


Auditory Masking Phenomena (cont.)

An Example of the Tradeoff of Time-Domain Masking Issues vs. Signal Redundancy for Two Signals:



A Typical Psychoacoustic Model:



Issues in Filterbank Design vs. Psychoacoustic Requirements

- There are two sets of requirements for filterbank design in perceptual audio coders:

- ***They conflict.***

- Remember: **ft >= 1 :** The better the frequency resolution, the worse the time resolution.

Requirement 1:

- Good Frequency Resolution

- Good frequency resolution is necessary to two reasons:

1) Diagonalization of the signal (source coding gain)
2) And
3) 2) Sufficient frequency resolution to control low-frequency masking artifacts. (The auditory filters are quite narrow at low frequencies, and require good control of noise by the filterbank.)

The Problem with Good Frequency Resolution:

- Bad time resolution

Requirement 2:

- Good Time Resolution

- Good time resolution is necessary for the control of time-related artifacts such as pre-echo and post-echo.

Problems with Good Time Resolution

- Not enough signal diagonalization, i.e. not enough redundancy removal.
- Not enough frequency control to do efficient coding at low frequencies.

Rule # 2

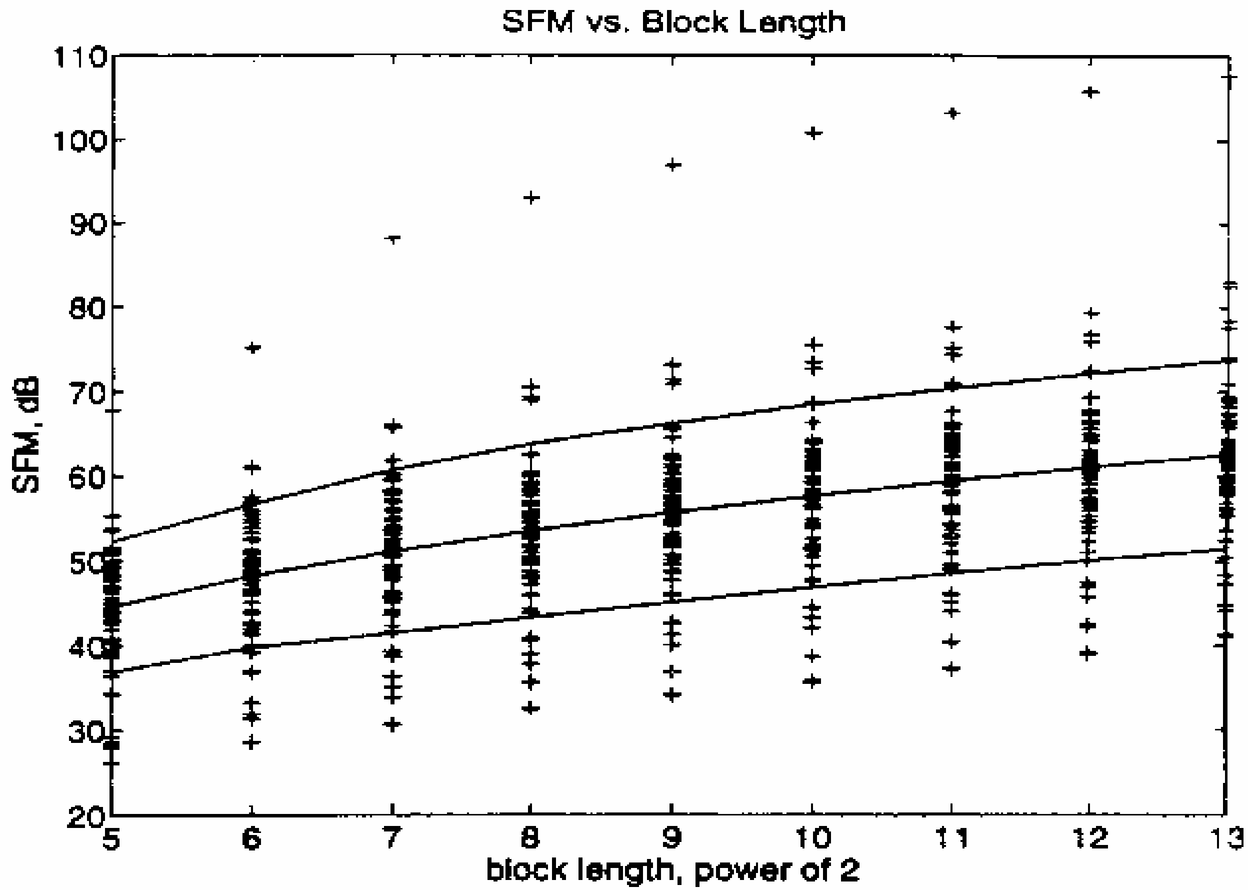
- The filterbank in an audio coder must have both good time resolution ***AND*** good frequency resolution in order to do an efficient job of audio coding.

Rule #2a

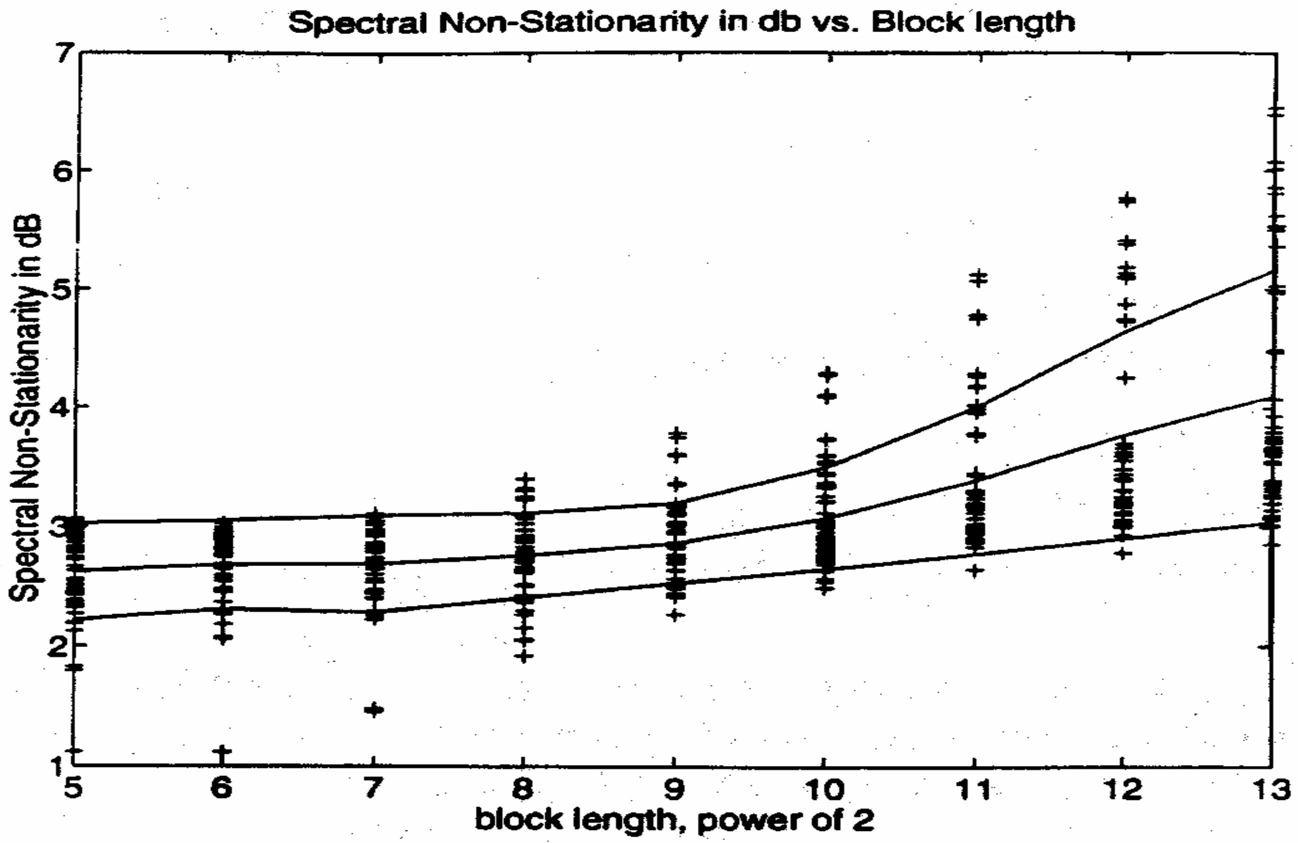
- An efficient audio coder must use a time-varying filterbank that responds to both the signal statistics ***AND*** the perceptual requirements.

Some signal statistics relevant to audio coder filterbank design

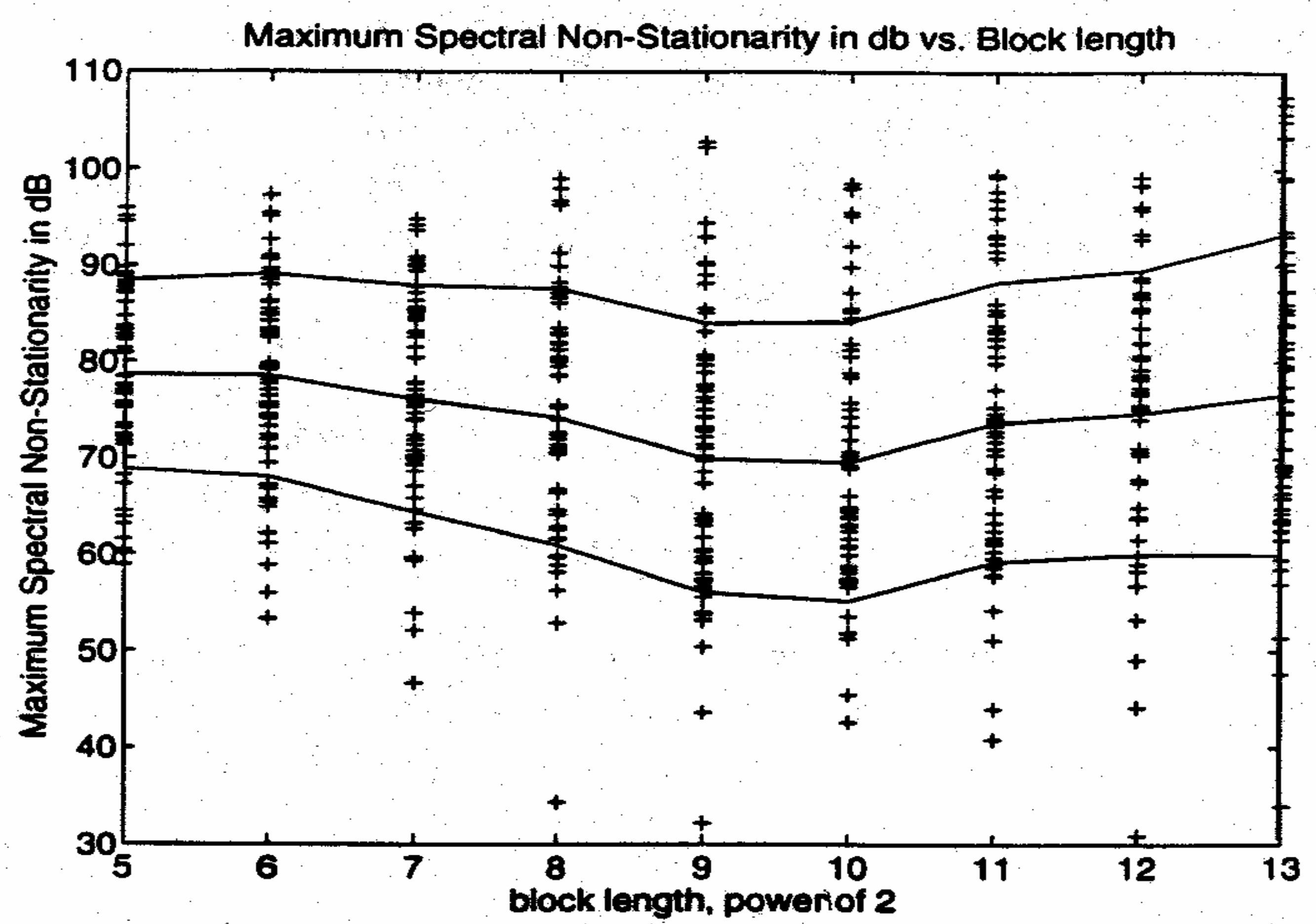
Spectral Flatness Measure as a function of block length



Mean Nonstationarity in Spectrum as a function of block length



Maximum Spectral Nonstationarity



Conclusions about Filterbanks

1) A length of about 1024 frequency bins is best for most, if not all, stationary signals.

2) A length of 64-128 frequency bins is appropriate for non-stationary signals.

Quantization and Rate Control:

- The purpose of the quantization and rate control parts of a perceptual coder is to implement the psychoacoustic threshold to the extent possible while maintaining the required bit rate.

- There are many approaches to the quantization and rate control problem. All of them have the same common goals of:

1) Enforcing the required rate
2) Implementing the psychoacoustic threshold
3) Adding noise in less offensive places when there are not enough bits

Quantization and Rate Control Goals:

- Everyone's approach to quantization and rate control is ***different***. In practice, one chooses the quantization and rate control parts that interact well with one's perceptual model, bitstream format, and filterbank length(s).

The Use of Noiseless Coding in Perceptual Audio Coders:

- There are several characteristics of the quantized values that are obtained from the quantization and rate control part of an efficient perceptual coder.

These Characteristics are:

1) The values around zero are the most common values
2) The quantizer bins are not equally likely
3) In order to prevent the need for sending bit allocation information,
4) *and*
5) in order to prevent the loss of efficiency due to the fact that quantizers do not in general have a number of bins equal to a power of two,
6) some self-termination kind of quantizer value transmission is necessary.

Huffman Codes:

1) Are the best-known technique for taking advantage of non-uniform distributions of single tokens.
2) Are self-terminating by definition.

Huffman Codes are NOT good at:

1) Providing efficient compression when there are very few tokens in a codebook.
2) Providing efficient compression when there is a relationship between successive tokens.

- Arithmetic and LZW coding are good at dealing with symbols that have a highly non-uniform conditional symbol appearance, and with symbols that have a wide probability distribution
- ***but***

1) They require either extra computation, integer specific programming, or extra RAM in the decoder
2) They require a longer training sequence, or a stored codebook corresponding to such a sequence or
3) They have a worse bound on compression efficiency and
4) They create difficulties with error recovery and/or signal break in because of their history dependence, ***therefore***
5) the more sophisticated noiseless coding algorithms are not well fitted to the audio coding problem.

The Efficient Solution:

- Multi-symbol Huffman codes, i.e. the use of Huffman codes where more than one symbol is included in one Huffman codeword.

- Such codebooks eliminate the problems inherent with “too small” codebooks, take a limited advantage of inter-symbol correlation, and do not introduce the problems of history or training time.

An Example Codebook Structure:

The MPEG-AAC Codebook Structure

Codebook Number	Largest Absolute Value	Codebook Dimension	Signed or Unsigned
0	0	*	*
1	1	4	s
2	1	4	s
3	2	4	u
4	2	4	u
5	3	2	s
6	3	2	s
7	7	2	u
8	7	2	u
9	12	2	u
10	12	2	u
11	16 (esc)	2	u

The Problem of Stereo Coding:

- There are several new issues introduced when the issue of stereophonic reproduction is introduced:

1) The problem of Binarual Masking Level Depression
2) The problem of image distortion or elimination

What is Binaural Masking Level Depression (BLMD)?:

- At lower frequencies, <3000 Hz, the HAS is able to take the phase of interaural signals into account. This can lead to the case where, for instance, a noise image and a tone image can be in different places. This can reduce the masking threshold by up to 20dB in extreme cases.

Stereo Coding (cont.):

- BMLD can create a situation whereby a singal that was “the same as the original” in a monophonic setting sounds substantially distorted in a stereophonic setting.

- Two good, efficient monophonic coders do ***NOT*** make one good efficient stereo coder.

Stereo Coding (cont.):

- In addition to BLMD issues, a signal with a distorted high-frequency envelope may sound “transparent” in the monophonic case, but will ***NOT*** in general provide the same imaging effects in the stereophonic case.

BMLD

- Both the low-frequency BLMD and the high-frequency envelope effects behave quite similarly in terms of stereo image impairment or noise unmasking, when we consider signal envelope at high frequencies or waveforms themselves at low frequencies. The effect is not as strong between 500Hz and 2 kHz.

Stereo Coding (cont):

- In order to control the imaging problems in stereo signals, several methods must be used:

1) A psychoacoustic model that takes account of BMLD and envelope issues must be included.
2) BMLD is best calculated and handled in the M/S paradigm
3) M/S, while very good for some signals, creates either a false noise image or a substantial overcoding requirement for other signals.

M/S Coding

- M/S coding is mid/side, or mono/stereo coding, M and S are defined as:
- M=L+R
- S=L-R

- The normalization of ½ is usually done on the encoding side. L in this example is the left channel, R the right.

Stereo Coding (cont.):

- A good stereo coder uses both M/S and L/R coding methods, as appropriate.
- The MPEG-AAC algorithm uses a method whereby the selection of M/S vs. L/R coding is made for each of 49 frequency band in each coding block. Protected thresholds for M, S, L, and R are calculated, and each M/S vs. L/R decision is made by calculating the bit cost of both methods, and choosing the one providing the lowest bit rate.

Stereo Coding (cont.):

- An M/S coder provides a great deal of redundancy elimination when signals with strong central images are present, or when signals with a strong “surround” component are present.

Stereo Coding (cont.):

- Finally, an M/S coder provides better signal recovery for signals that have “matrixed” information present, by preserving the M and S channels preferentially to the L and R channels when one of M or S has the predominant energy.

What’s This About “Intensity Stereo” or the MPEG-1 Layer 1,2 “Joint Stereo Mode”?

- Intensity stereo is a method whereby the relative intensities of the L and R channels are used to provide high-frequency imaging information. Usually, one signal (L+R, typically) is sent, with two gains, one for L and one for R.

“Intensity Stereo” (cont.):

- “Intensity Stereo” Methods do not guarantee the preservation of the Envelope of the Signal for High Frequencies.

- For “lower quality” coding, intensity stereo is a useful alternative to M/S stereo coding,
- ***and***
- For situations where intensity stereo DOES preserve the high-frequency signal envelope, it is useful for high quality audio coding. Such situations are not as common as one might prefer.

- Think of intensity stereo as the coder equivalent of a “pan-pot”.

Temporal Noise Shaping

- Temporal Noise Shaping (TNS) can help with preserving the envelope in the case of intensity stereo coding,
- HOWEVER
- the control of TNS and intensity stereo is not yet well understood.

Stereo Coding (cont.):

- Finally, a stereo coder must consider the joint efficiency issues when "block switching" on account of a signal attack. If the attack is present in only one channel, the pre-echo must be prevented, while at the same time maintaining efficient coding for the non-attach channel.

- This is a tough problem.

What About Intensity Stereo in the MPEG-AAC Standard?

- In the AAC standard, intensity stereo can be activated by using one of the “spare” codebooks. The ability to use M/S or intensity stereo coding as needed, in each coding block, allows for extremely efficient coding of both acoustic and pan-pot stereo signals.

So, what about rate control and all that good stuff?

- Because the rates of the M, S, L, and R components vary radically from instant to instant, the only reasonable way to do the rate control and quantization issues is to do an "overall" rate control, hence my unwillingness to say "this is 48 kb/s/channel" as opposed to "this is a 96 kb/s stereo-coded signal."

- The more information that one can put under the rate control mechanism at one instant, the better the coder can cross-allocate information in a perceptually necessary sense, hence the same is true for multi-channel audio signals, or even sets of independent audio signals.

Multichannel Audio Issues:

- The issue of multichannel audio is a natural extension of the stereophonic coding methods, in that symmetric pairs must be coded with the same stereophonic imaging concerns, and in that joint allocation across all channels is entirely desirable.

Multichannel (cont.):

- There are some problems and techniques unique to the multichannel environment:

1) Inter-channel prediction.
2) Pre-echo in the multichannel coder
3) Time delay to “rear” channels

Inter-channel Prediction:

- It is thought that under some circumstances, the use of inter-channel prediction may reduce the bit rate.

- To the present, this has not been realized in a published coder due to the delay issues in rear channels and the memory required to realize such inter-channel predictors.

Pre-echo in the Multi-channel Setting:

- Due to the delay in signals between channel pairs, it is necessary to provide independent block switching for each channel pair, at least, in order to eliminate situations where enormous over-coding requirements occur due to the need to suppress pre-echos.

Time Delay to the "Rear" Channels:

- In multi channel audio, there is often a long time delay to the rear channels. While the problems this introduces have, in a sense, been addressed in the prediction and pre-echo comments, this delay in fact makes "joint" processing of more than channel pairs difficult on many if not all levels.

- On the other hand, as this decorrelates the bit-rate demand for front and rear channels, it raises the gain available when all channels are jointly processed in the quantization and rate-control sense.

Multichannel:

- On the issue of “Backward Compatibility”, or the ability to either send or derive a stereo mixdown from the multichannel coded signal:
- (turn on echoplex)
- ***The use of “Matrix” or “L’,R’” matrixing inside the coder is a***
- ***BAD IDEA!***

Multichannel:

- When the 2-channel mixdown is required, it is better to send it as a separate signal pair, rather than as either a pre- or post-matrixed signal, and allowing the appropriate extra bit rate for the extra two channels.
- The same is true for the Monophonic mixdown channel.
- ***Why?***

Multichannel:

- There are several reasons:

1) It is better, from the artist’s and producer’s point of view, to have separate, and deliberately mixed, 1, 2, and multichannel mixdowns.
2) In the process of assuring the quality of either the L or L’ channel, whichever is derived, the peak bit rate will be the same or more than that which is required to simply send that channel outright. Further more, in this case, additional decoder complexity and memory will be required.

Using Perceptual Audio Coding:

- Perceptual audio coding is intended for final delivery applications. It is not advisable for principle recording of signals, or in cases where the signal will be processed heavily ***AFTER***
- the coding is applied.

Using Perceptual Audio Coding:

- Perceptual audio coding is applicable where the signal will NOT be reprocessed, equalized, or otherwise modified before the final delivery to the consumer.

The “Tandeming” or “Multiple Encoding” Problem:

- There is a one-word solution to the problem of using multiple encodings.

- ***<u>DON’T</u>***

Multiple Encoding (cont.):

- If you are in a situation where you must do multiple encodings:

1) Avoid it to the extent possible and
2) Use a high bit rate for all but the final delivery bitstream.

Finally:

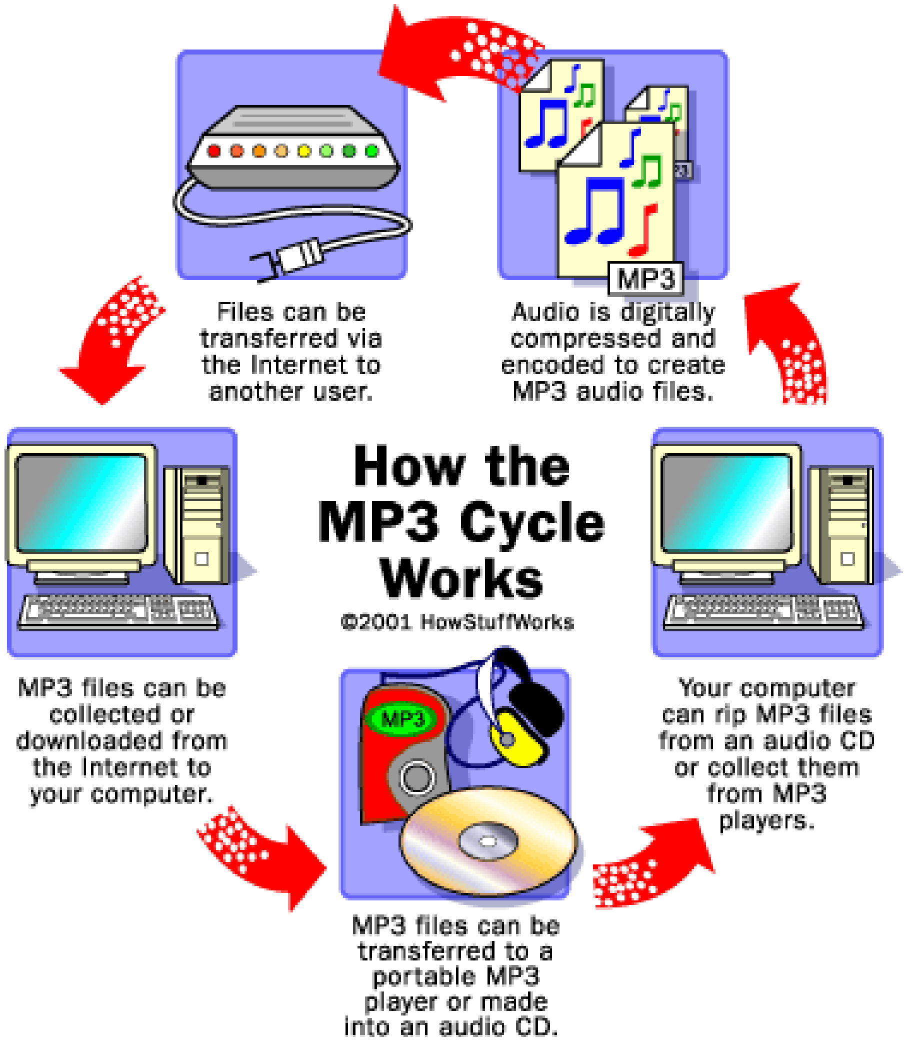
- Perceptual coding of audio is a very powerful technique for the ***final*** delivery of audio signals in situations where the delivery bit rate is limited, and/or when the storage space is small.

Psychoacoustics and the MP3

Evolution and Implications of MP3s

- Storage
- Internet
- Lawsuits

MP3 Cycle


How the MP3 Cycle Works
©2001 HowStuffWorks
Audio is digitally compressed and encoded to create MP3 audio files.
MP3
Files can be transferred via the Internet to another user.
MP3 files can be collected or downloaded from the Internet to your computer.
MP3
MP3 files can be transferred to a portable MP3 player or made into an audio CD.
Your computer can rip MP3 files from an audio CD or collect them from MP3 players.

Brief History of MP3

- **MPEG** (Motion Picture Experts Group)
  International standards of audio and video
  Agreed on in 1993

- MPEG 1 Audio Layer 3
  Audio part of MPEG

Conventional Digital Audio

- Analog to Digital (**A-D**)

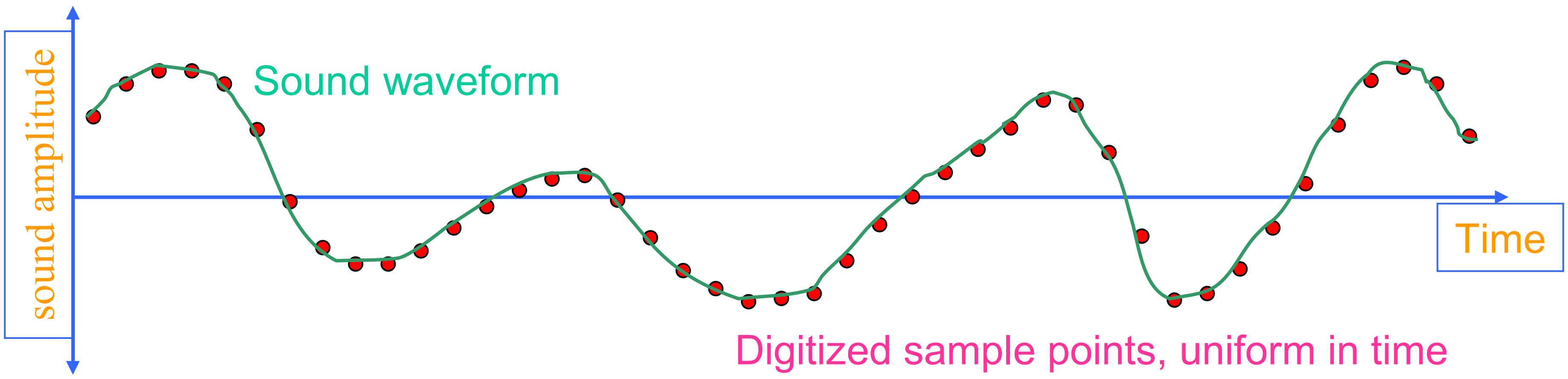
  Encodes sound to digital (binary)

- Digital to Analog (**D-A**)

  Converts into playable audio by reverse process

Conventional Digital Audio

- **PCM (Pulse Code Modulation)**
  - Process of digitizing and retrieval
  - Sampling



Sampling Frequency

- Number of samples of sound per second (function of time)

- 44.1kHz is CD quality standard
  - Shannon-Nyquist Theorem
  - Will explain later
  - Professionals differ

Bit Depth and Amplitude

- 1s and 0s describe data: Bit words

  In this case, sound waves amplitude

- So, the more you have the better the description

- 16-bits per sample is standard CD quality due to dB range (see previous slide)

Bit Depth

How binary works

Base 2: 1's and 0's only

0 → 00000000 (8-bit)

1 → 00000001 (8-bit)

2 → 00000010 (8-bit)

3 → 00000011 (8-bit) (1 + 2)

4 → 00000100 (8-bit)

127 → 01111111 (8-bit) (1 + 2 + 4 + 8 + 16 + 32 + 64)

–127 → 11111111 (8-bit): first bit indicates negative

Math of PCM

- (44,100 Hz) x (16-bit/sec)/(8) = 88,200 b/s
- X2 for stereo = 176,400 b/s
- X60 seconds for minute = 10,584,000 b/m

Around 10MB for minute of recording… that's a lot

56K modem: One song in 2hrs

Why Compress?

- Computer space
- File sharing/transfer
- Server space

Methods

- Halving the sample rate

  Cuts size in half!

  High frequency loss (lacks clarity)

- **Shannon-Nyquist Theorem**

  Based on highest frequencies heard

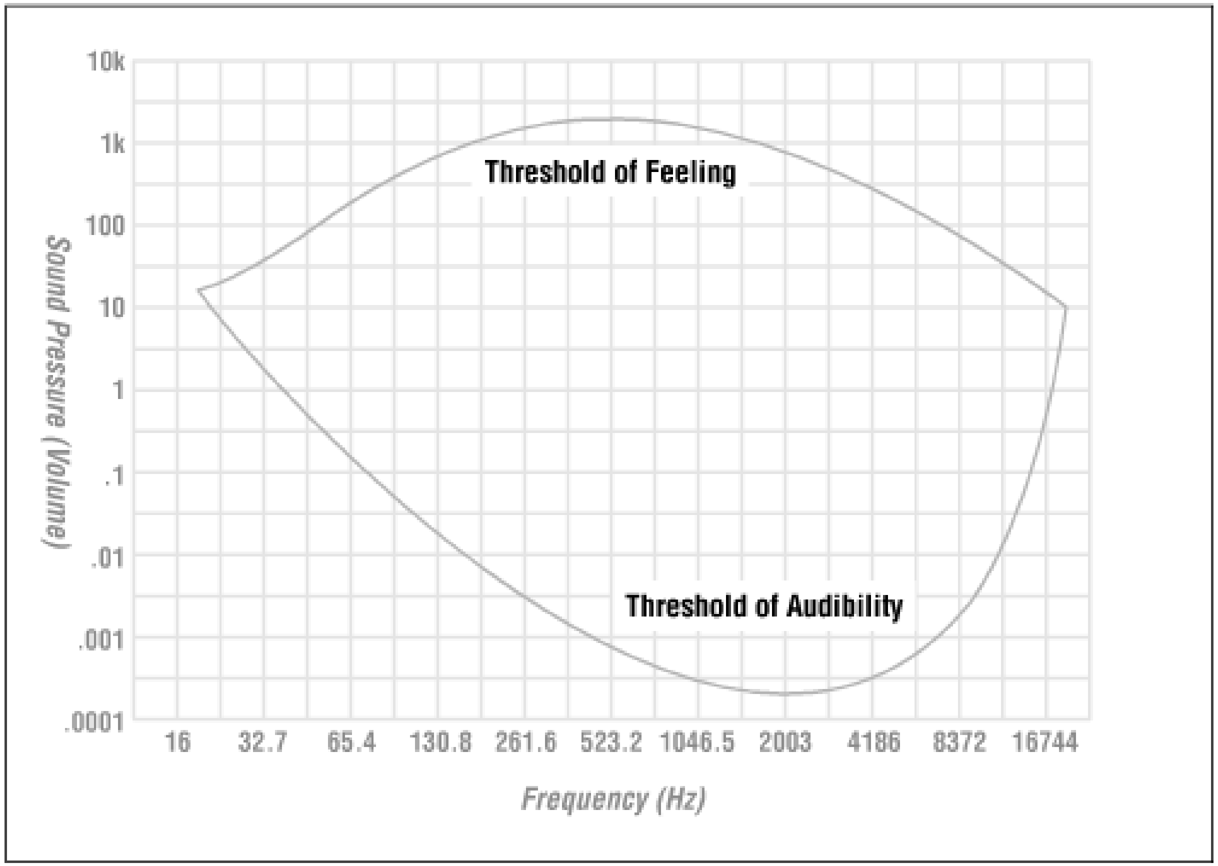
  Must record at 2x highest frequency heard

  44.1 KHz – **Nyquist Limit** is a standard

Hearing Sensitivity

Shannon-Nyquist Theorem

Range of Hearing



Methods

- Reducing bit depth from 16 to 8
  - Cuts the size in half!
  - Reduces quality to much more than half due to base-2 possibility
  - Sound harsh and unnatural anyway
- **Temporal Redundancy Reduction**
  - Loss-less
  - Text, Programs, ZIP
  - Cuts down about 10-15% only

Intro to Selective Compression
Perceptual Coding

- Called **Lossy**

  Because you lose information forever

- Exploits selective perception

  Not exact realization of physical world

  Example: Amplitude to loudness

  **Other things are not perceived at all

Psychoacoustics:
Sound For Listener

- Does away with data that is:
    - Unperceived
    - Redundant

- Is a concept of storing data that is relevant to human ear and networks

- **File size reduced 10 times!**

Psychoacoustics:
Masking

- **Masking** (term borrowed from vision)
  - Tendency to prioritize certain stimuli ahead of others, according to the context in which they occur

- **Loudness and frequency dependent**
  - Simultaneous or near-simultaneous stimuli
  - Masking occurs

Masking

- **Loudness**

  Quiet sound around a relatively loud sound won't be perceived

- **Frequency**

  Low sound around a relatively high sound won't be perceived

Masking

- Can be **forward** or **backward masking** that is dependent on time
  - **Forward** < 200ms
  - **Backward** is much shorter

  Due to cochlear response as well as neural pathway limitations

Encoding an MP3

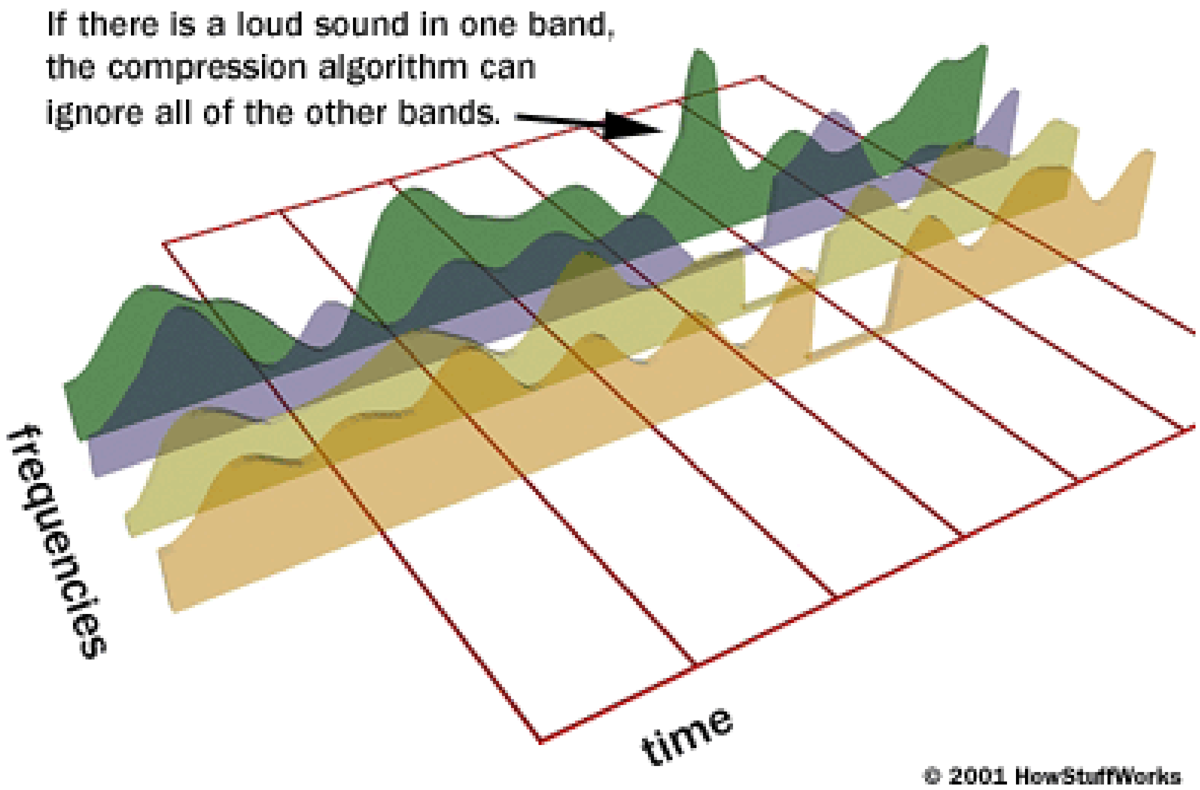
- Still 16-bit and 44 KHz
- Cut into **frames** first
- **Fourier Analysis**

  Divides into 32 sub-bands of frequency spectrum

Changing Bit Rate

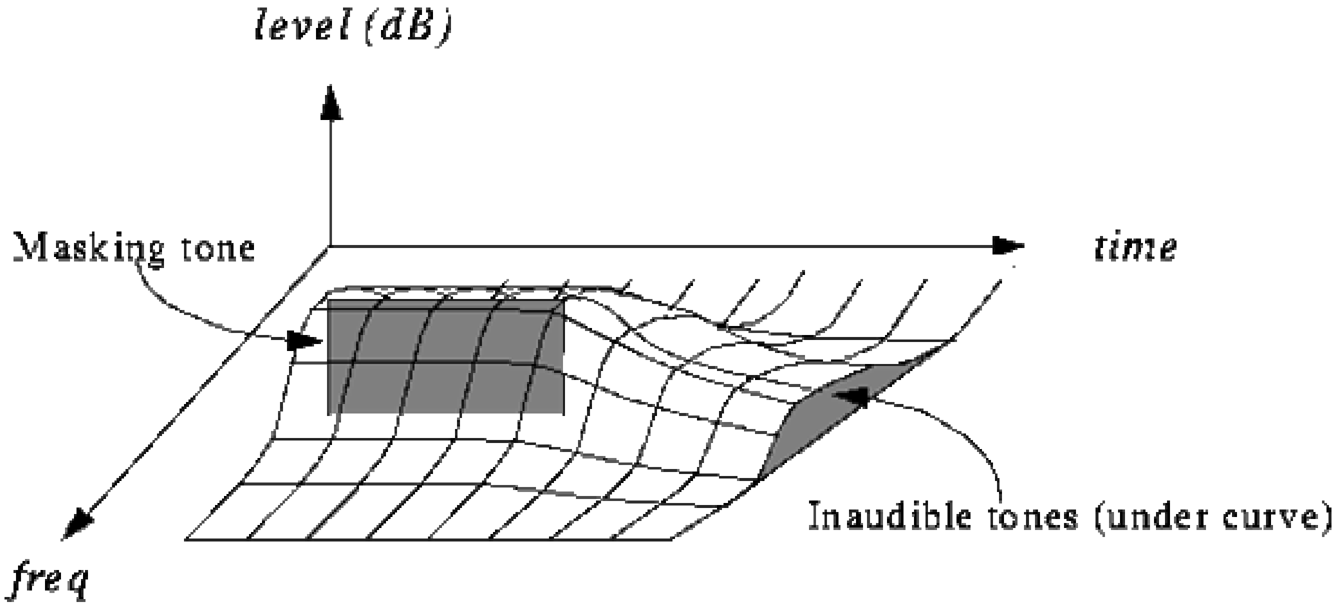
- **Bit depth**
  - Assigns less bits to describe irrelevant (masked) sounds
  - Assigns more bits to describe more relevant (masking) sounds
  - Uses much less bits on average

Bands In Masking



Masking

Frequency and Volume



Critical Bandwidth

- This is the bandwidth around a tone that acts as a masker which cannot be perceived
- **Signal to Noise Ratio** is used to determine the critical bandwidth
- Similar to the **Cochlear Filter** shape
  - Related to response of the cochlea to sounds in time (disputed)

Encoding

- Quality according to bits

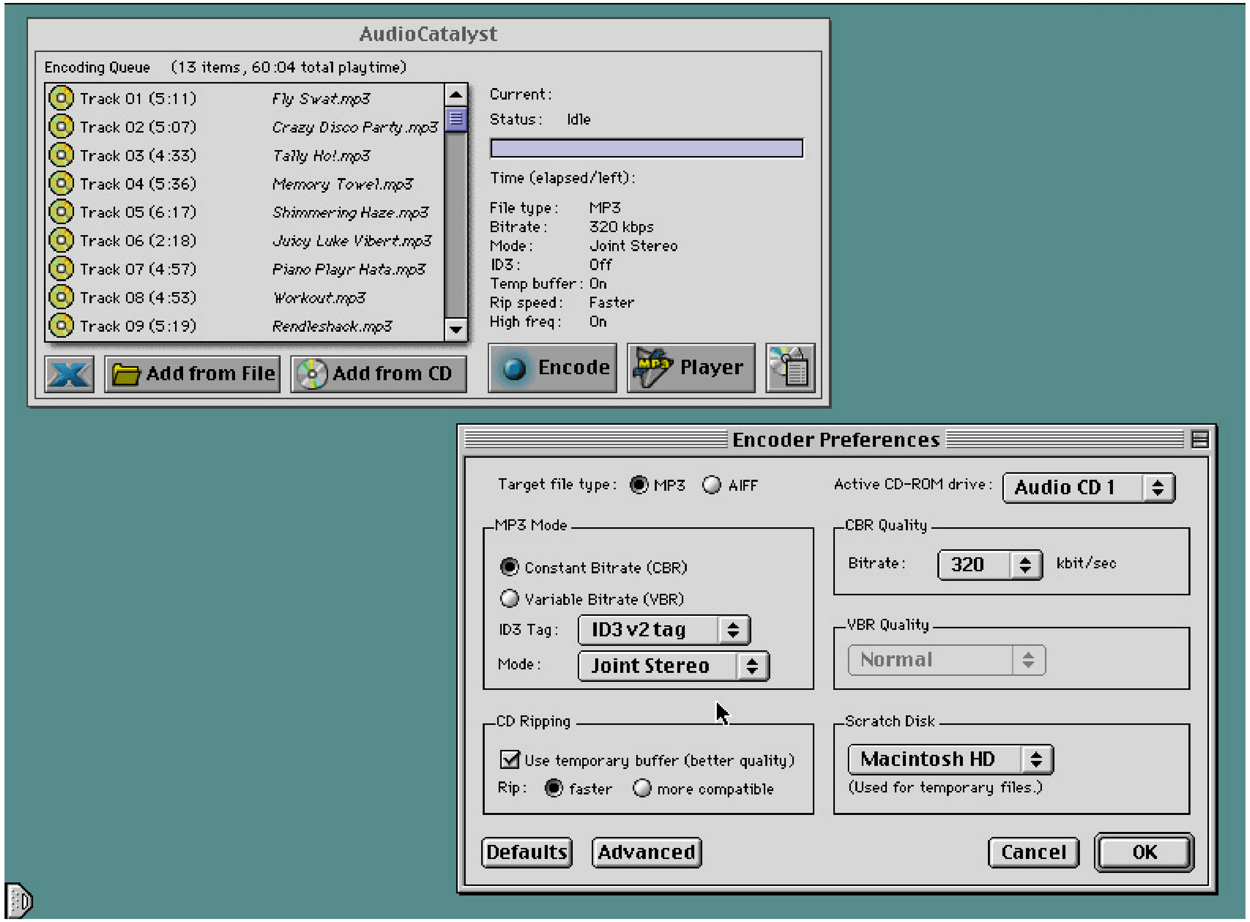
  128+ kilobits per second (kbps) for high quality priority

  But could be as high as 224, 256 and 320kbps

  Chosen before encoding according to needs of the user

  Size, speed, or quality

Computer View



Decoding

- Reverse process
- Simpler
- Decoders are more common than encoders

Improvements

- **VRB** (Variable Bit Rate) Encoding

  Bit rate is altered continuously according to the content complexity of the music

  e.g. Orchestral music

  File size slightly larger

Open Source Competition

MP3
AAC
MP4
WMA

- Offer better than common compression
- Problems with *backward compatibility* so they have to be really good

Future of Digital Audio

- With faster Internet and cheaper and larger memory and storage
  - Why compress?
- Compression will use better psychoacoustic models
  - Elimination of all unperceived components
  - No discoloration of quality
  - Better than CD quality

Summary (Important terms are *italicized* or bold in the presentation)

- Digitizing analog sounds relies on perceptual understandings of the listener
- Digitizing is done through sampling of physical properties
- Shannon-Nyquist Theorem is used as a basic guide
- Frequency of sampling coresponds to time and bits per sample corresponds to amplitude
- Masking works due to cochlear and neural limitations
- Masking allows variable bit sampling after Fourier Analysis
- Perceptual Encoding (yield of psychoacoustical research) offers reduction of file size by 10 times
- Future of digital audio depends on understanding of perceptual systems that could achieve better than CD-quality sound

References

- *Perceptual Coding of Digital Audio*, Painter, T., **Proceedings of Institute of Electrical and Electronics Engineers**, Vol. 88, No. 4, April 2000, http://www.eas.asu.edu/
- *Multiple Description Perceptual Audio Coding with Correlating Transforms*, Arean, R., Kovacevic, J., **IEEE Transactions on Speech and Audio Processing**, Vol. 8, No. 2, March 2000, http://www.rle.mit.edu/
- Digital Audio Compression, Yen Pan, D., Digital Technical Journal, Vol. 5, No. 2, Spring 1993, http://www.iro.umontreal.ca
- A Tutorial on MPEG/Audio Compression, Pan, D., IEEE Multimedia Journal, Summer 1995, http://www.cs.columbia.edu

- <u>ILLUSTRATIONS:</u>

  http://www.sparta.lu.se/~bjorn/whitney/compress.htm

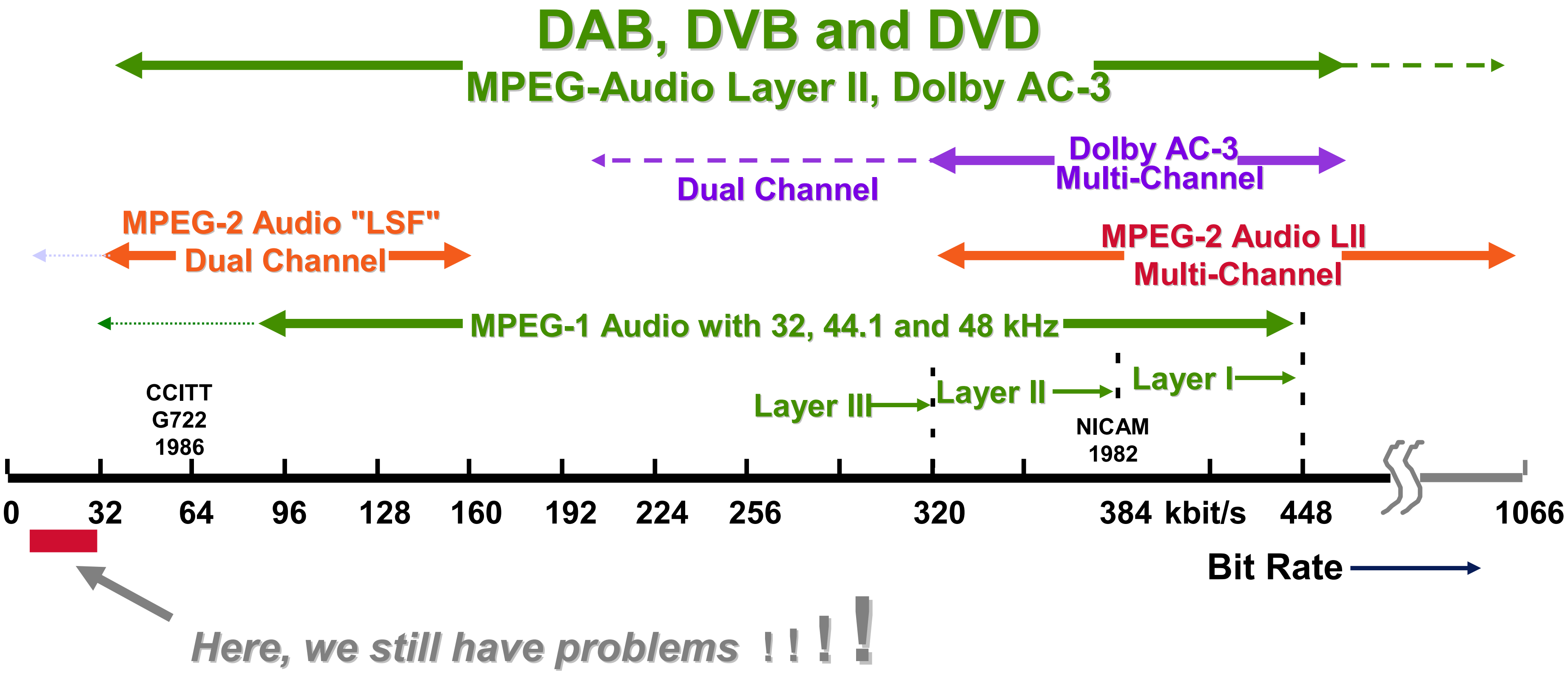
  http://www.howstuffworks.com/

  http://www.mp3-converter.com/

MPEG-4 Audio Coding

- Audio Coding for Transmission and Storage
- MPEG-4 Audio Toolbox: Generic Audio Coding
  - *Natural Audio Coding*
  - *Parametric Audio Coding*
  - *Structured Audio Orchestra Language*
- MPEG-4 Audio Toolbox: Speech Coding
  - *CELP Coding*
  - *HVXC Parametric Speech Coding*
  - *TTS Text-To-Speech*
- MPEG-4 Audio Codecs and typical Bit-rates
- Audio Demonstration

Digital Audio for Transmission and Storage
Target Bit Rates for MPEG Audio and Dolby AC-3



Possible candidates to solve these problems:

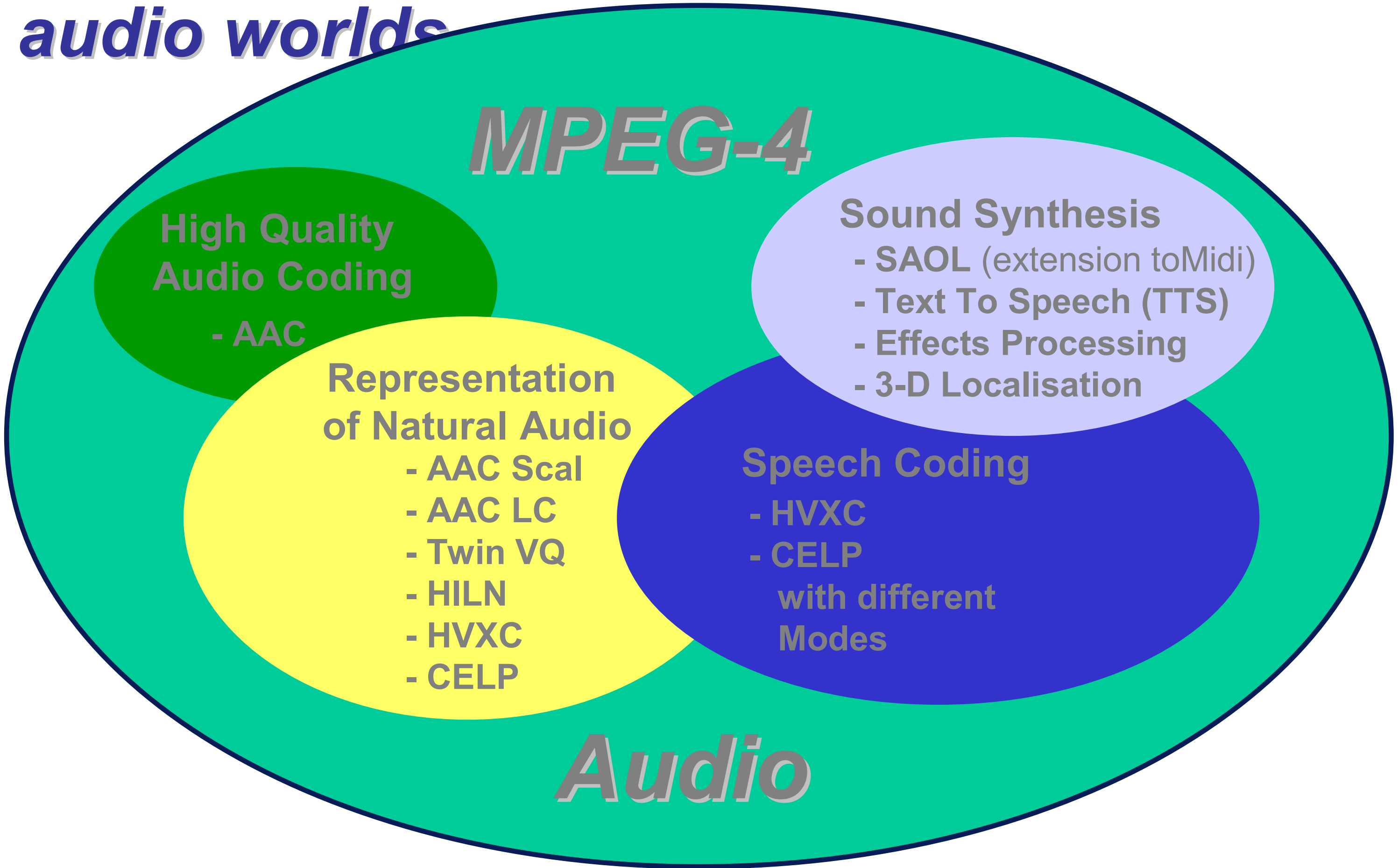
- MPEG-2 AAC and MPEG-4 Audio
- Internet Radio Audio Package Manufacturers

History of MPEG-Audio

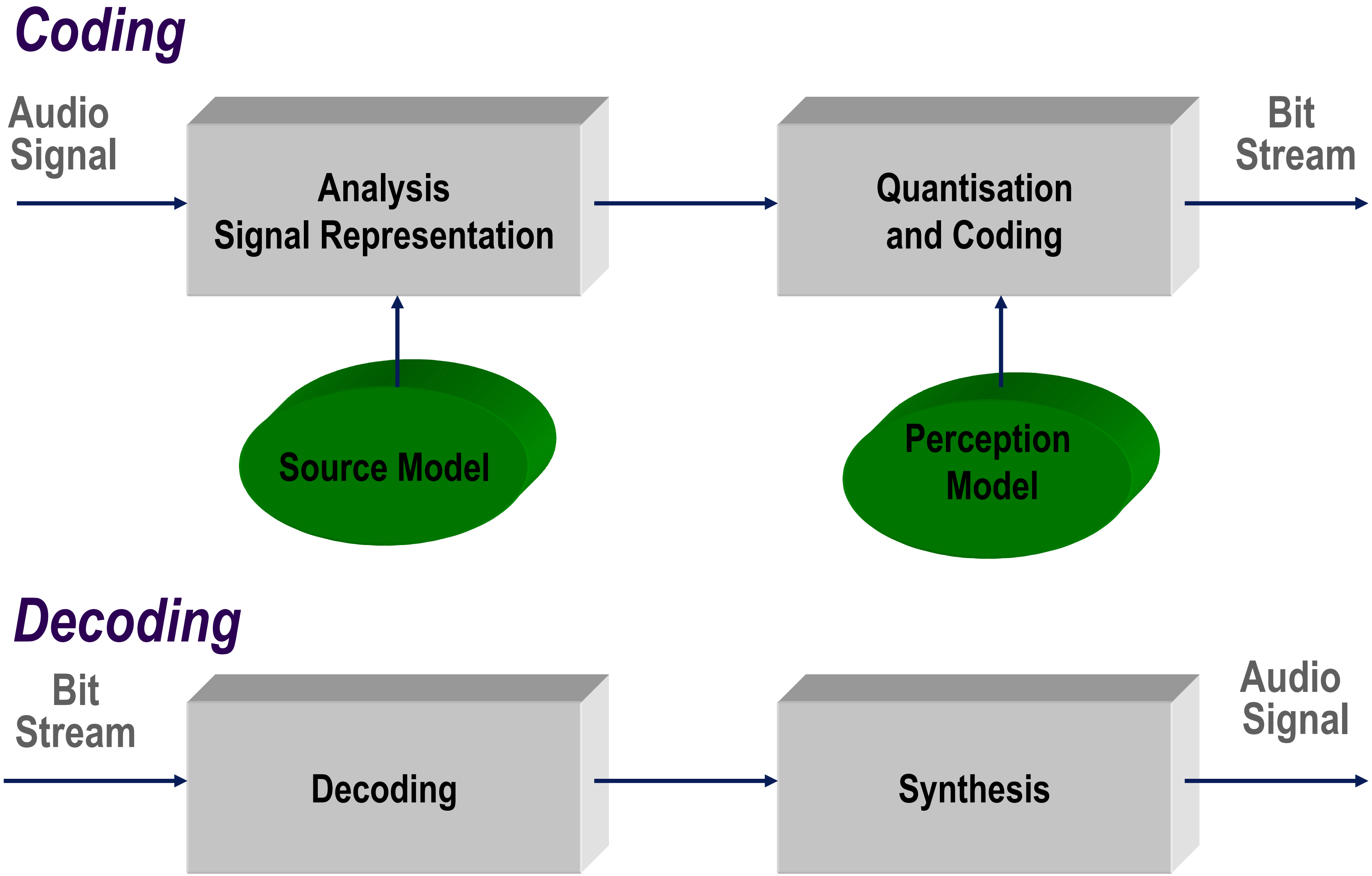
- MPEG-1 Two-Channel coding standard (Nov. 1992)
- MPEG-2 Extension towards Lower-Sampling-Frequency (LSF) (1994)
- MPEG-2 Backwards compatible multi-channel coding (1994)
- MPEG-2 Higher Quality multi-channel standard (MPEG-2 AAC) (1997)
- MPEG-4 Audio Coding and Added Functionalities (1999, 2000)

MPEG-4 Audio: The Audio Toolbox

- ***MPEG-4 Audio integrates the different audio worlds***



MPEG-4 Audio:
Conceptual diagram of basic audio coding



MPEG-4 Audio: Natural and Unnatural (Structured) Audio

- High Quality Audio Tools:
  - AAC: „**A**dvanced **A**udio **C**oding"
  - Twin VQ: „**T**ransform domain **w**eighted **in**terleaved **V**ector **Q**uantisation"

- Low Bit-rate Audio Tools:
  - CELP „**C**ode **E**xcited **L**inear **P**redictive coding"
  - Parametric, i.e. coding based on parametric representation of audio signal
    - **HVXC: „Harmonic Vector eXcitation Coding"**
    - **HILN: „Harmonic and Individual Line and Noise"**

- SNHC Audio Tools
  - **S**ynthethic/**N**atural **H**ybrid **C**oding
    - **SAOL „Structured Audio Orchestra Language"**
    - **TTS „Text-To-Speech"**

MPEG-4 Audio: Functionalities

- Bit-rate compression

  About 2 kbit/s/ch to 64 kbit/s/ch
- Scaleability

  Bit-rate

  Steps of 16 kbit/s/ch with AAC Scalable

  Complexity

  Delay
- Pitch change in encoded domain
- Speed change in encoded domain
- Text-To-Speech (TTS)
- Structured Audio (SA)

Let's Listen!

Let's Listen!

MPEG-4 Audio Version 1: Tools

- Coding based on T/F (Time to Frequency) mapping
  - *Advanced Audio Coding (AAC)*
    - Long term prediction
    - Bit Sliced Arithmetic Coding
    - Perceptual Noise Shaping
  - *Twin VQ*
- Code Excited Linear Prediction
  - *With NB and WB CELP*
- Parametric Coding
- Structured Audio Orchestra Language (SAOL)
- Structured Audio Score Language (SASL)

MPEG-4 Audio Version 2:
New Tools (Part 1)

- Error Resilience
  - Error robustness,
    - Huffman codeword reordering (HCR) in ACC bitstream
    - Reversible Variable Length Coding (RVLC)
    - Virtual Code-Books (VCB11) to extend the sectioning info
  - Error protection
    - Unequal Error Protection (UEP)
    - Forward Error Correction (FEC)
    - Cyclic Redundancy Check (CRC)
- Low-Delay Audio Coding for AAC
- Small Step Scalability
  - Steps of 1 kbit/s/ch with BSAC „Bit-Sliced Arithmetic Coding“ in combination with AAC
- Parametric Audio Coding
  - Harmonic and Individual Line plus Noise (HILN)

MPEG-4 Audio Version 2:
New Tools (Part 2)

- Environmental Spatialisation
  - Physical approach, based on description of the acoustical properties of the environment
  - Perceptual approach, based on high level perceptual description of audio scenes
  - Version 2 Advanced Audio BIFS (Binary Format for Scene description)
- CELP Silence Compression
- MPEG-4 File Format: MP4
  - Independent of any particular delivery mechanism
  - Streamable format, rather than a streaming format
  - Based on the QuickTime format from Apple Computer Inc.
- Backchannel Specification
  - Allows for user-controlled streaming

MPEG-4 Audio Version 2:
Profiles and Levels

- Low Delay Audio Profile

  for speech and generic audio coding

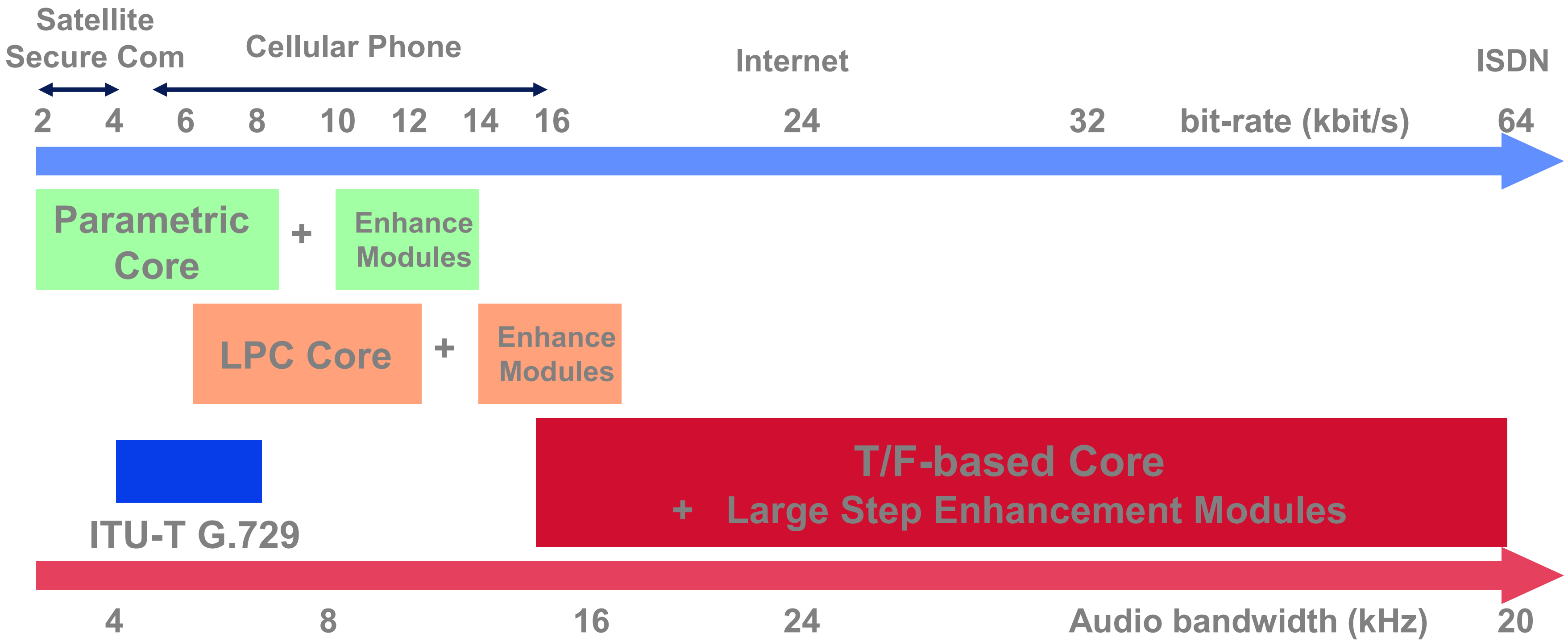
- Scalable Internet Audio Profile

  extends scalable profile of version1 by small step scalability and Parametric Audio Coding

- MainPlus Audio Profile

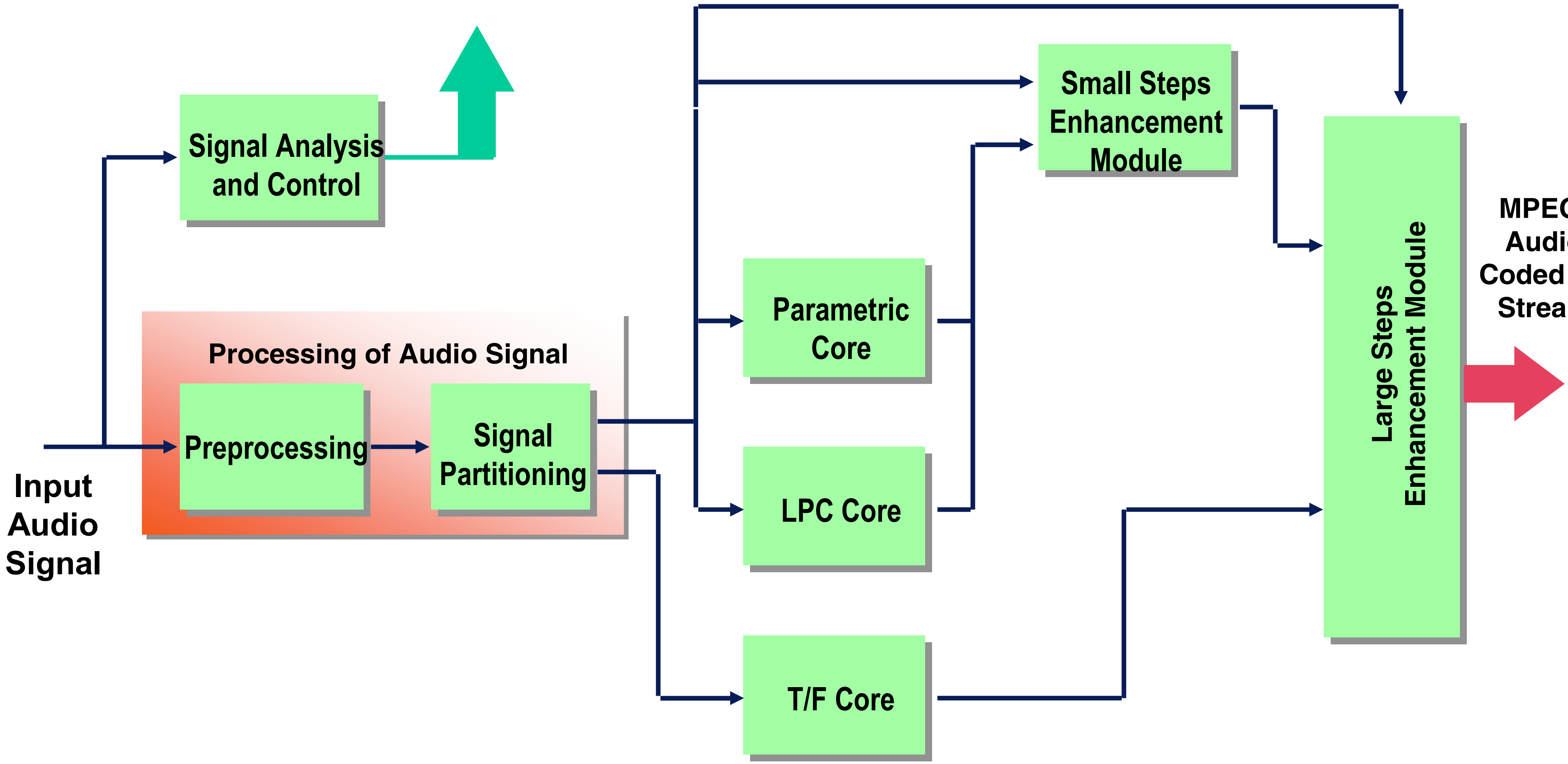
  provides superset of all audio profiles

MPEG-4 Audio:
Different Codecs and their typical bit-rate range

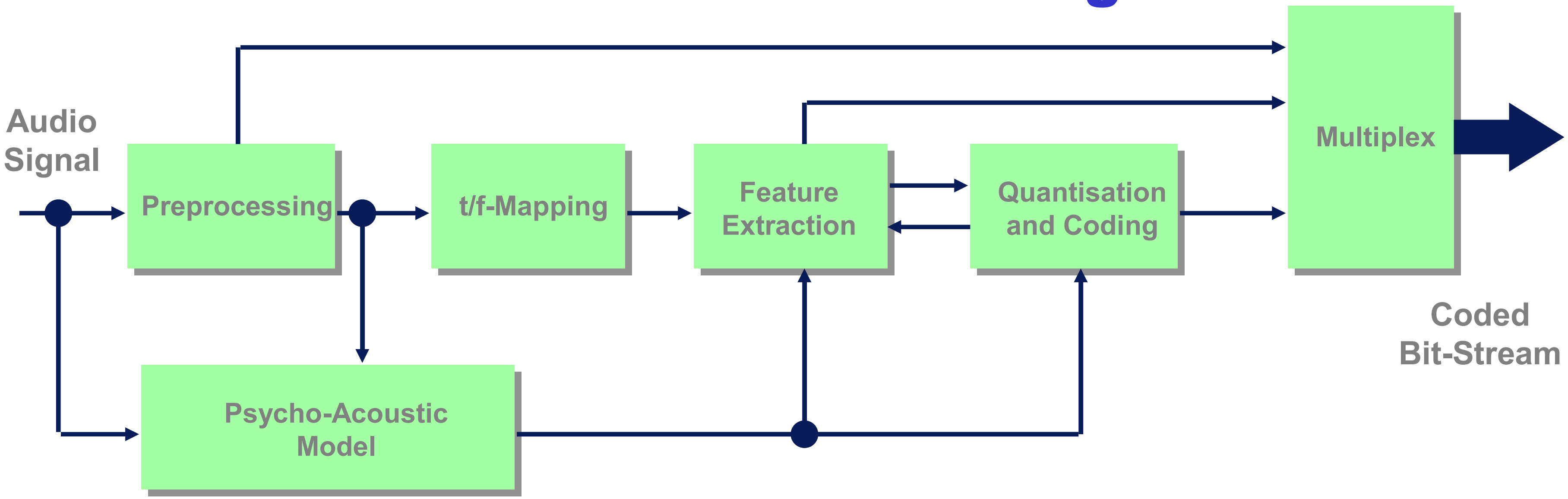


- *Class A:* **based on Time/Frequency mapping (T/F), i.e. transform coding**
- *Class B:* **Linear Predicting Coding (LPC) based analysis/synthesis codec**
- *Class C:* **Codecs, based on parametric description of audio signal**

Conceptual Diagram: MPEG-4 Audio VM framework scalable audio encoder



Conceptual Diagram of a class A Core Codec: t/f-based audio coding



- t/f-mapping based on transform codecs
- Generic Coding System
- Performs best at bit-rates ≥ 24 kbit/s per channel

MPEG-2 Advanced Audio Coding: Profiles

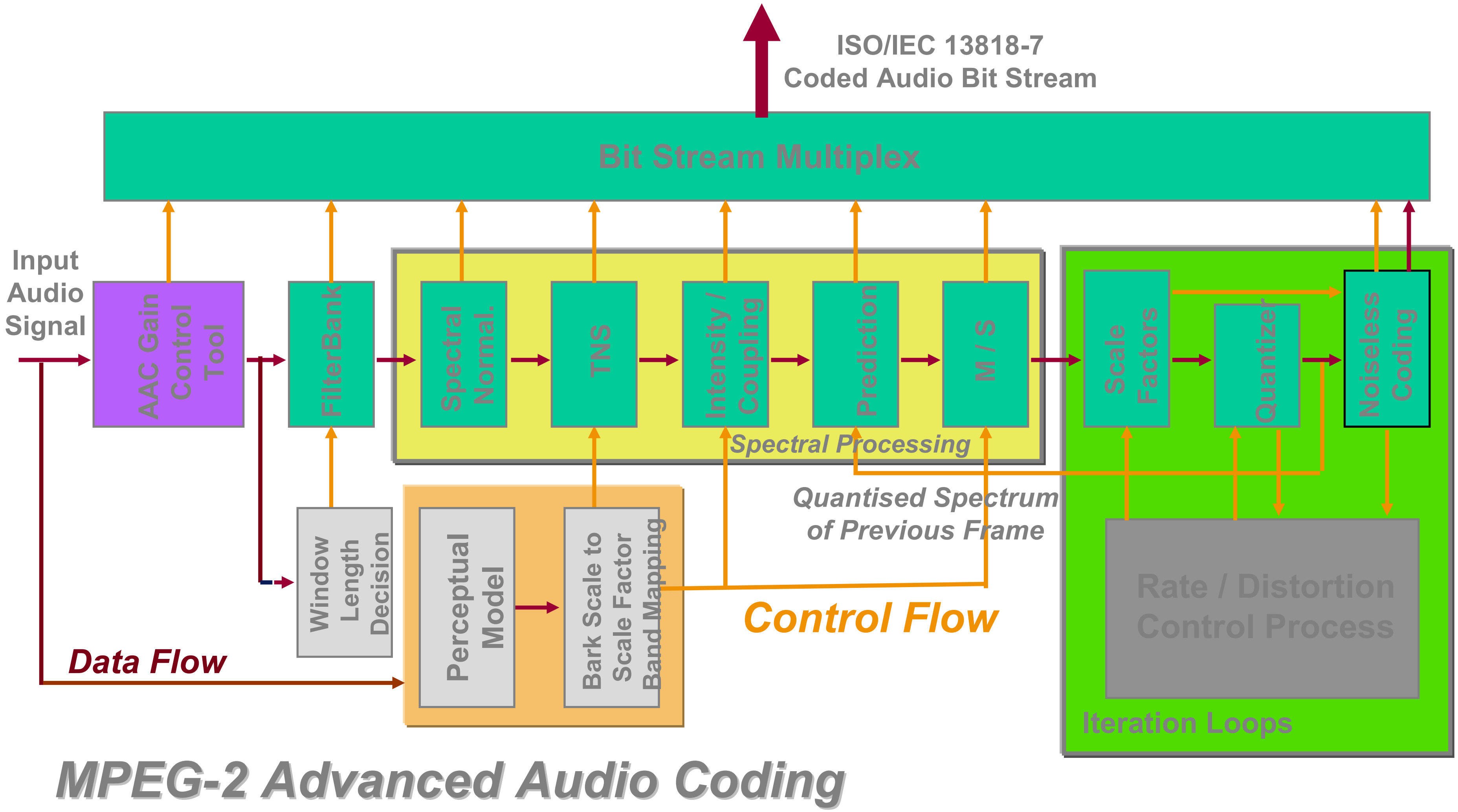
- ISO/IEC 13818-7 (2nd Phase of MPEG-2 Audio)
  MPEG-2 NBC (Non Backwards compatible Coding) finalized in April 1997
- Renamed to AAC (Advanced Audio Coding)

- MPEG-2 AAC Profiles
  - Main Profile
  - Low Complexity (LC) Profile
  - Sample Rate Scaleable (SRS) Profile
    - of particular interest to Internet Radio and Digital AM, SW systems

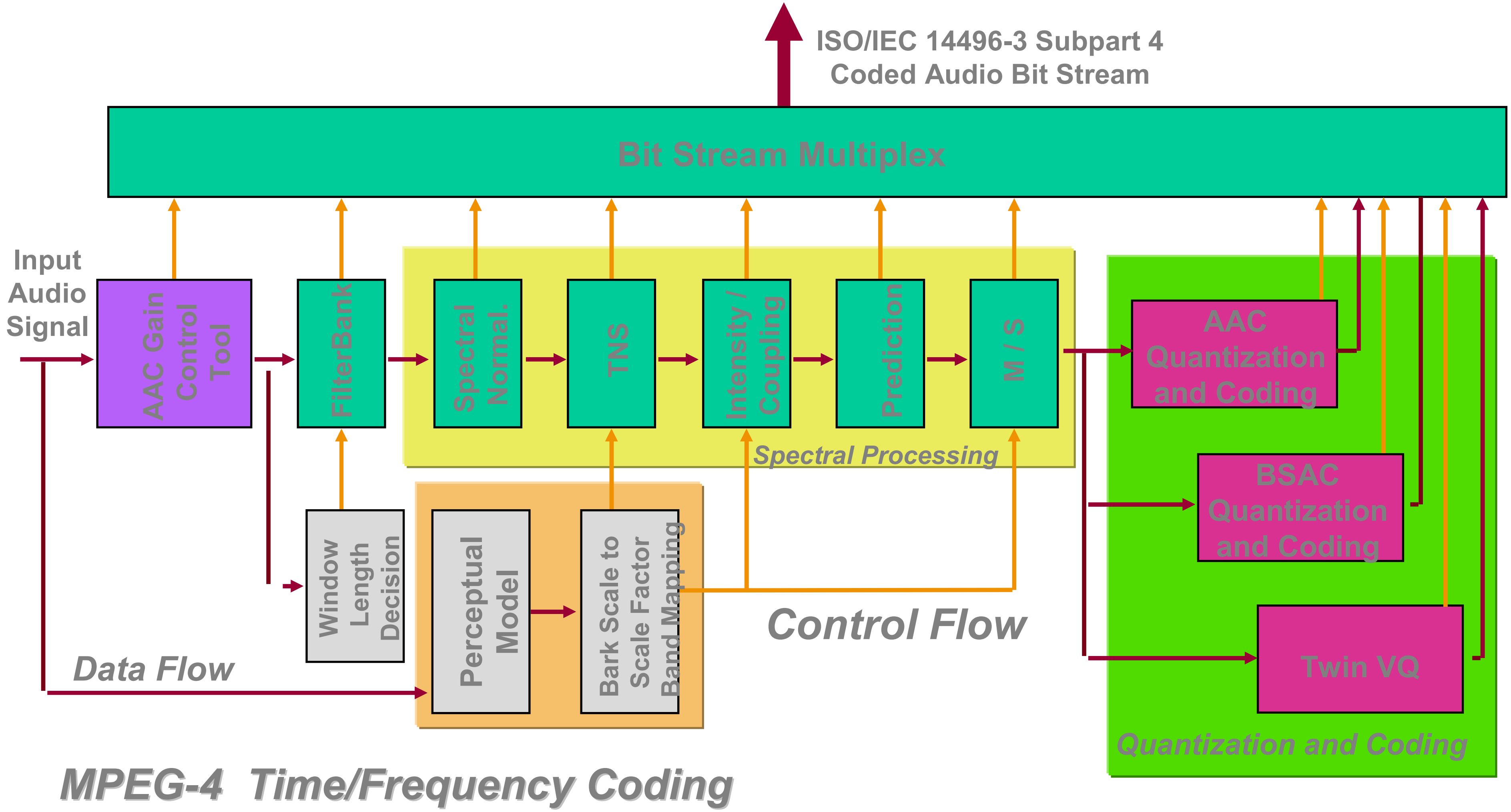
MPEG-2 Advanced Audio Coding: Tools

- MPEG-2 AAC Tools
  - Preprocessing Module: Gain Control, optional
    - mainly used for SRS Profile only
  - Filter bank to split signal into subsampled spectral components with frequency resolution 23 Hz or time resolution 5.3 ms
  - Computing of masking thresholds (similar to MPEG-1 Audio)
  - Temporal Noise Shaping (TNS) to control fine structure of quantisation noise within filter bank window (optional, i.e. simplified version used for LC Profile)
  - Intensity Stereo and coupling channel, optional
  - Perceptual Noise Substitution (PNS), optional
  - Time-domain Prediction of subsampled spectral components (optional, not used for LC Profile)
  - M/S decision, optional

MPEG-4 Audio Coding:
Conceptual diagram of MPEG-2/4 AAC



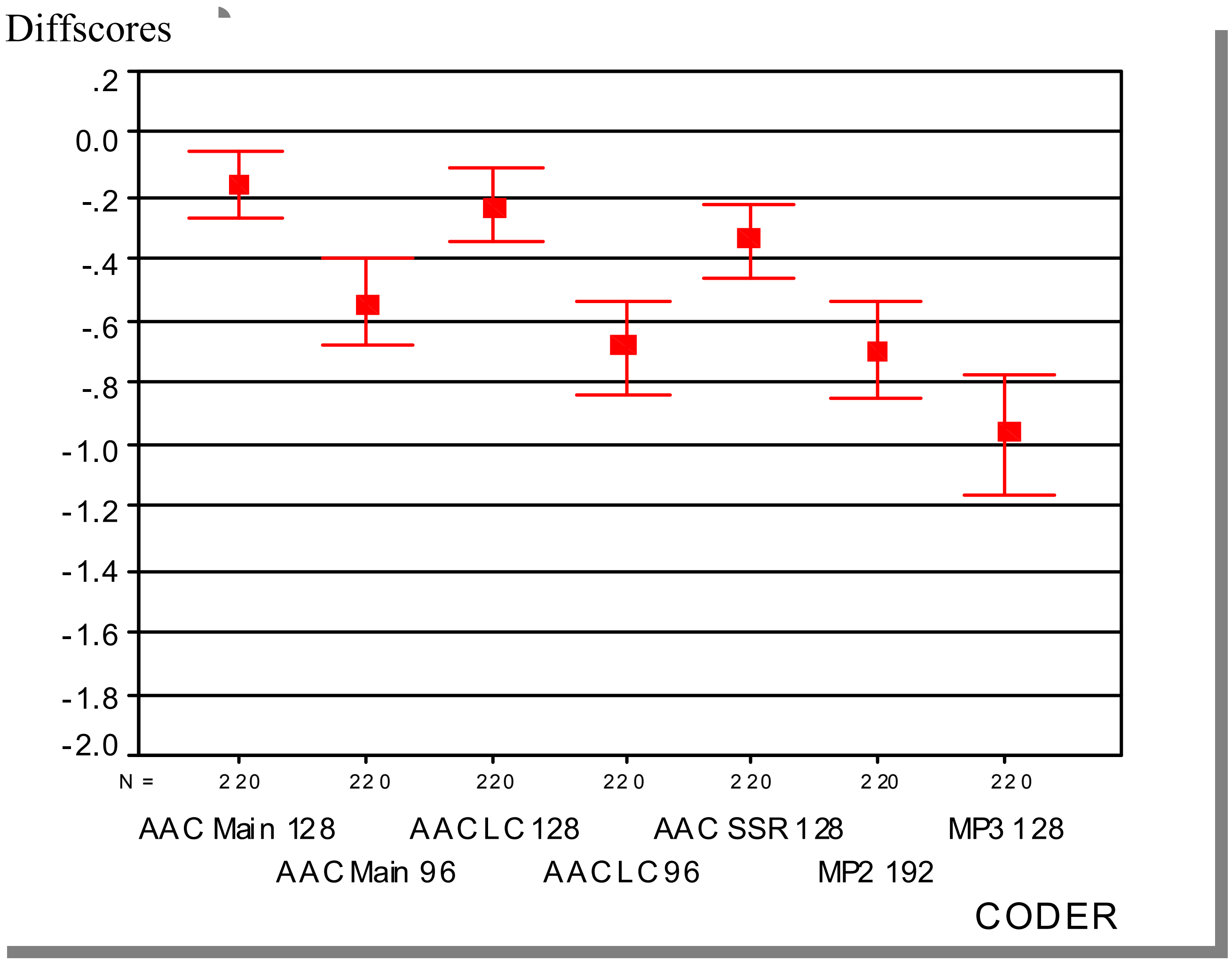
MPEG-4 Audio Coding:
Conceptual diagram of the AAC-based Encoders



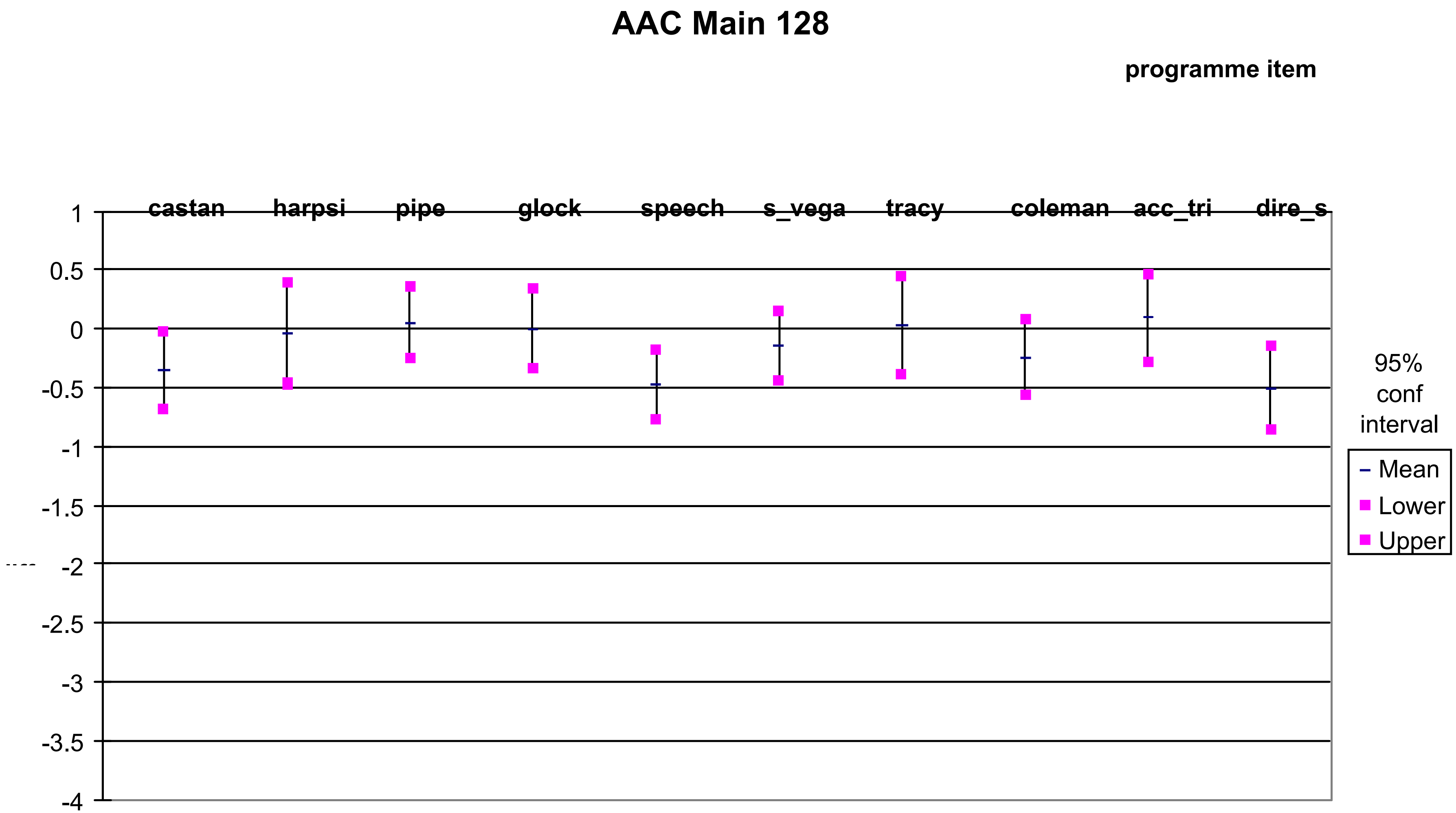
MPEG-4 General Audio Coding:
AAC Low Delay

- Delay mainly caused by:
  - Frame length
  - Analysis and synthesis filter
  - Switching window (2048 versus 256 samples) needs "look-ahead" time in encoder
  - Bit reservoir to equalise Variable Bit-Rate (VBR) demands
- Minimum theoretic delay :
  110 ms plus 210 ms for bit reservoir
  (at 24 kHz Sampling Frequency and bit-rate of 24 kbit/s)
- AAC Low Delay:
  - Frame length: only 512 samples
  - No look-ahead time
  - No Bit reservoir
  - Loss in coding gain: about 20%

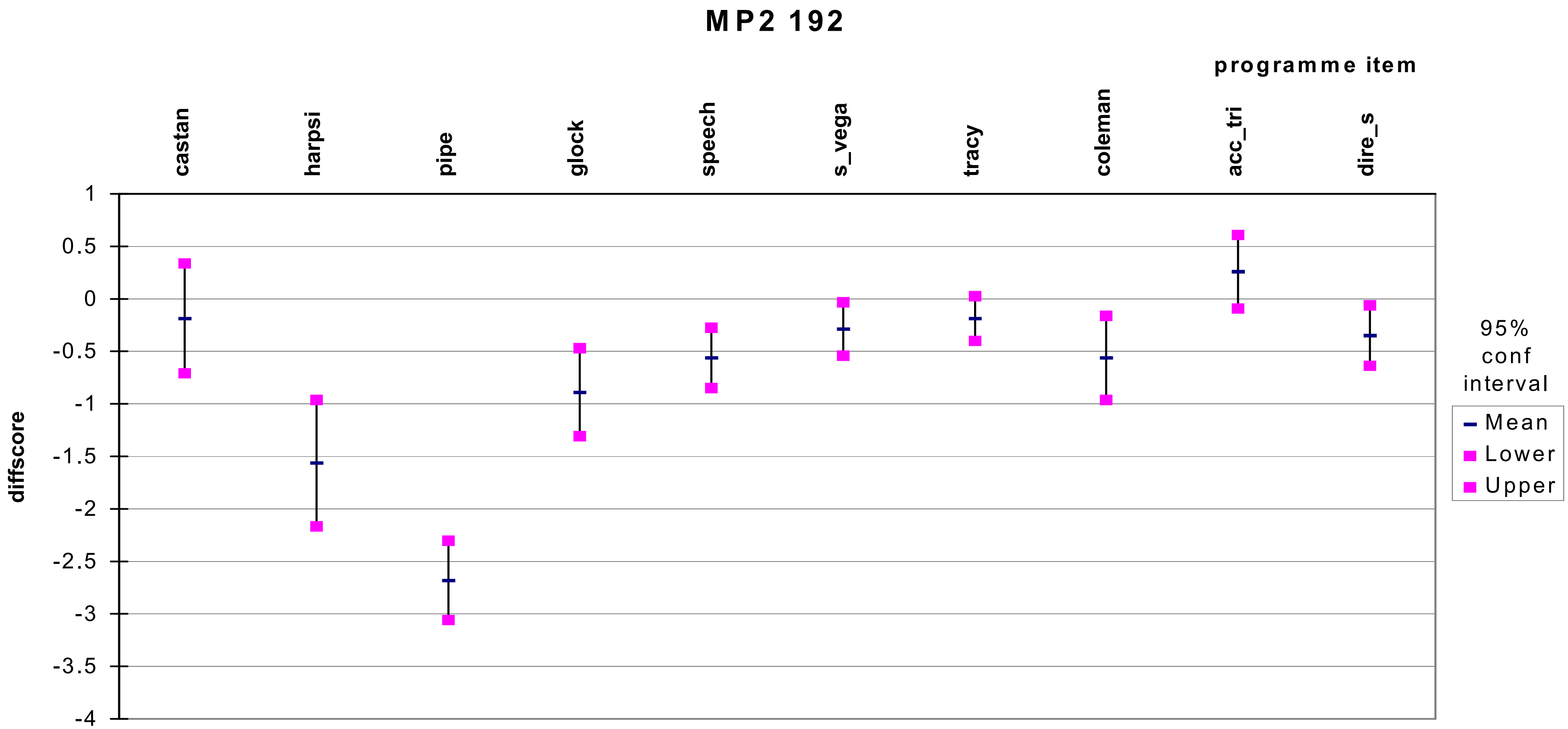
Comparison with MPEG-1 codecs : Overall results (average across programme items)



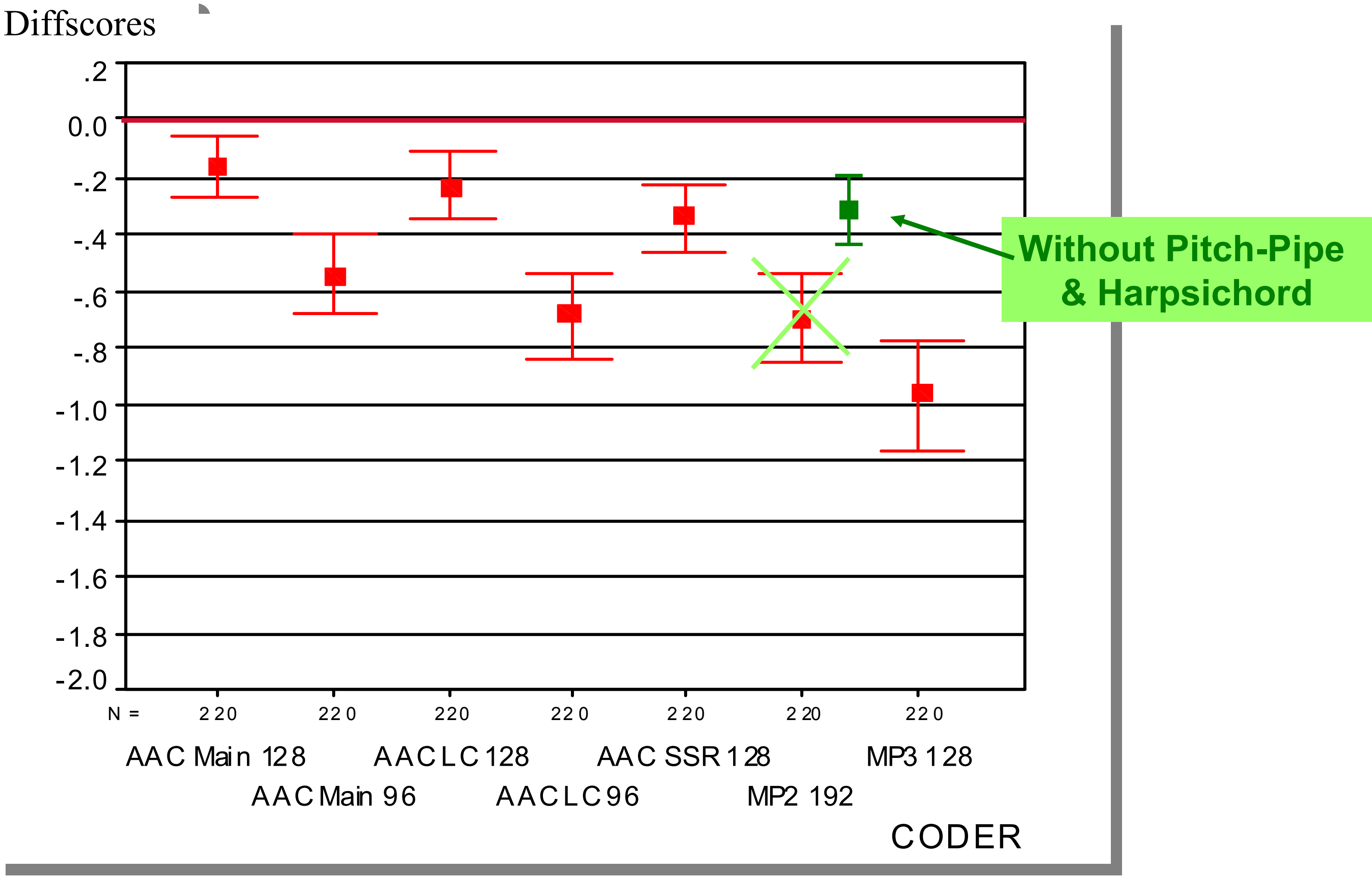
MPEG-2 AAC Verification Tests: Results for AAC Main Profile at 128 kbps



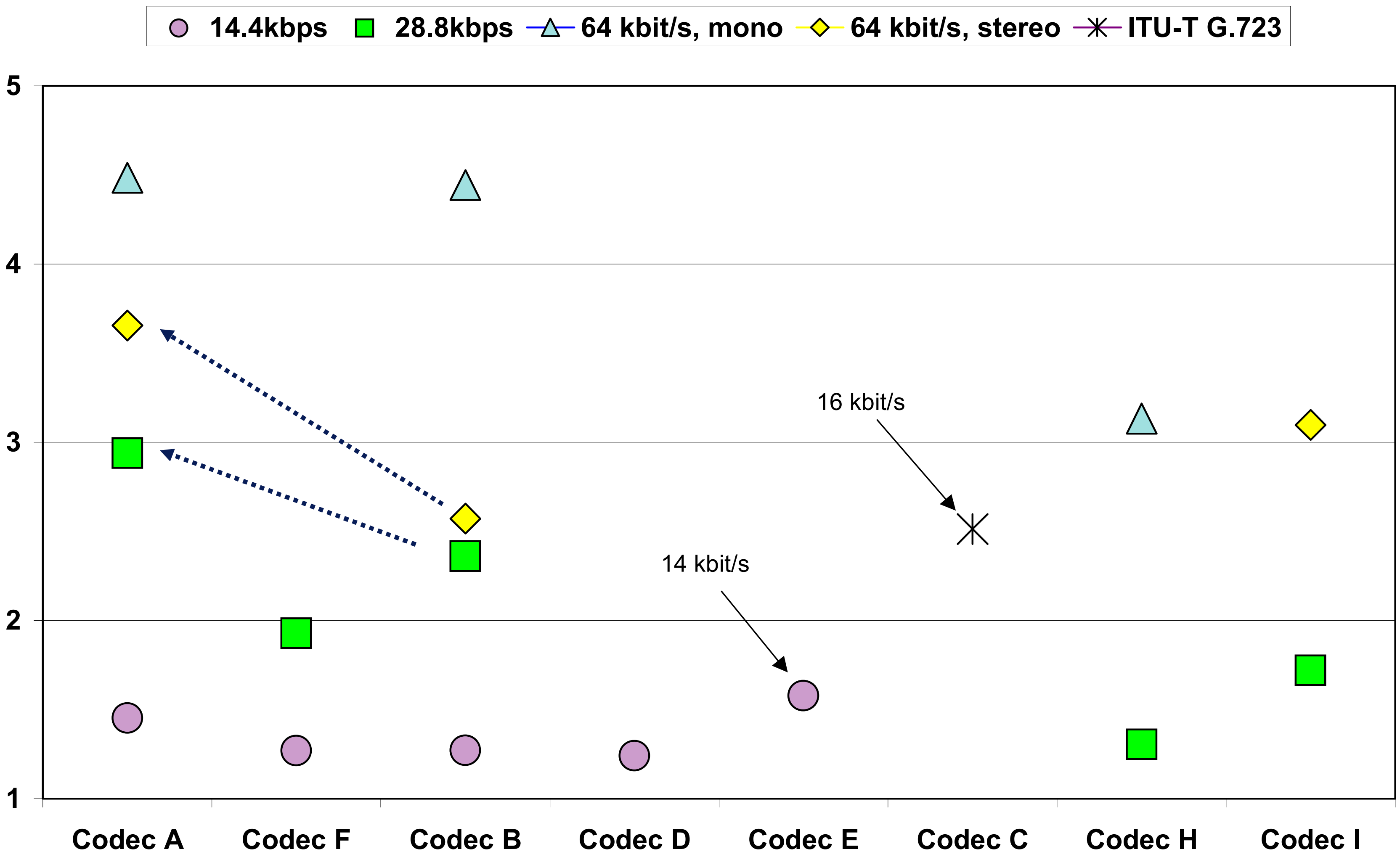
MPEG-2 AAC Verification Tests: Results for MPEG-1 Layer II at 192 kbps



Comparison with MPEG-1 codecs : Overall results (average across programme items)



Audio@Internet-Tests 1997:
Mean Values Audio Quality at different Modem Speed

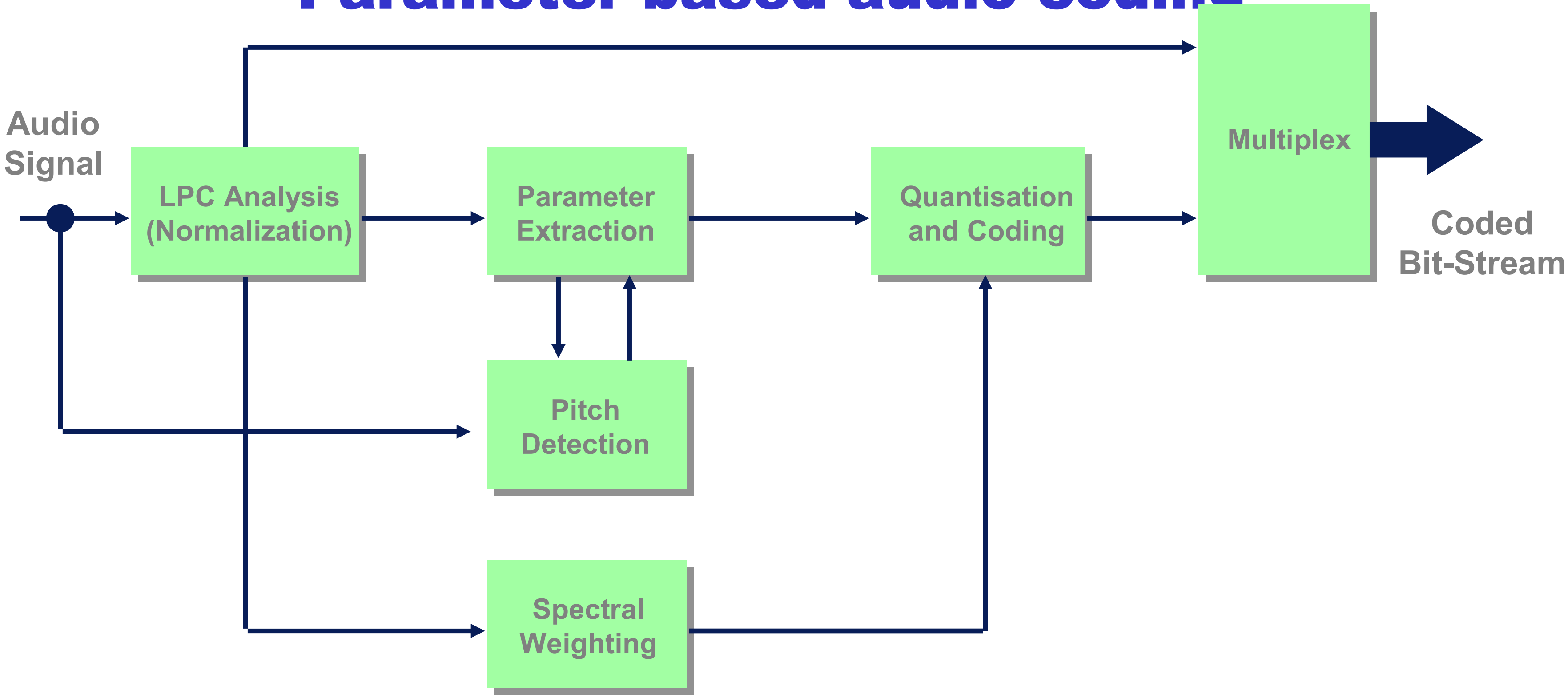


EBU B/AIM Internet Radio Tests '99

- Microsoft Windows Media 4
  - New Version, available since end of August 99
- MPEG-4 AAC
  - FhG-IIS
- MPEG-2.5 Layer III, or MP3
  - Opticom
- Quicktime 4 Music-Codec 2
  - Qdesign, new Version, Sept. 99 - was not yet commerciallyavailable
- RealAudio 5.0
- RealAudio G2
- MPEG-4 TwinVQ
  - Yamaha's SoundVQ

Report at EBU available

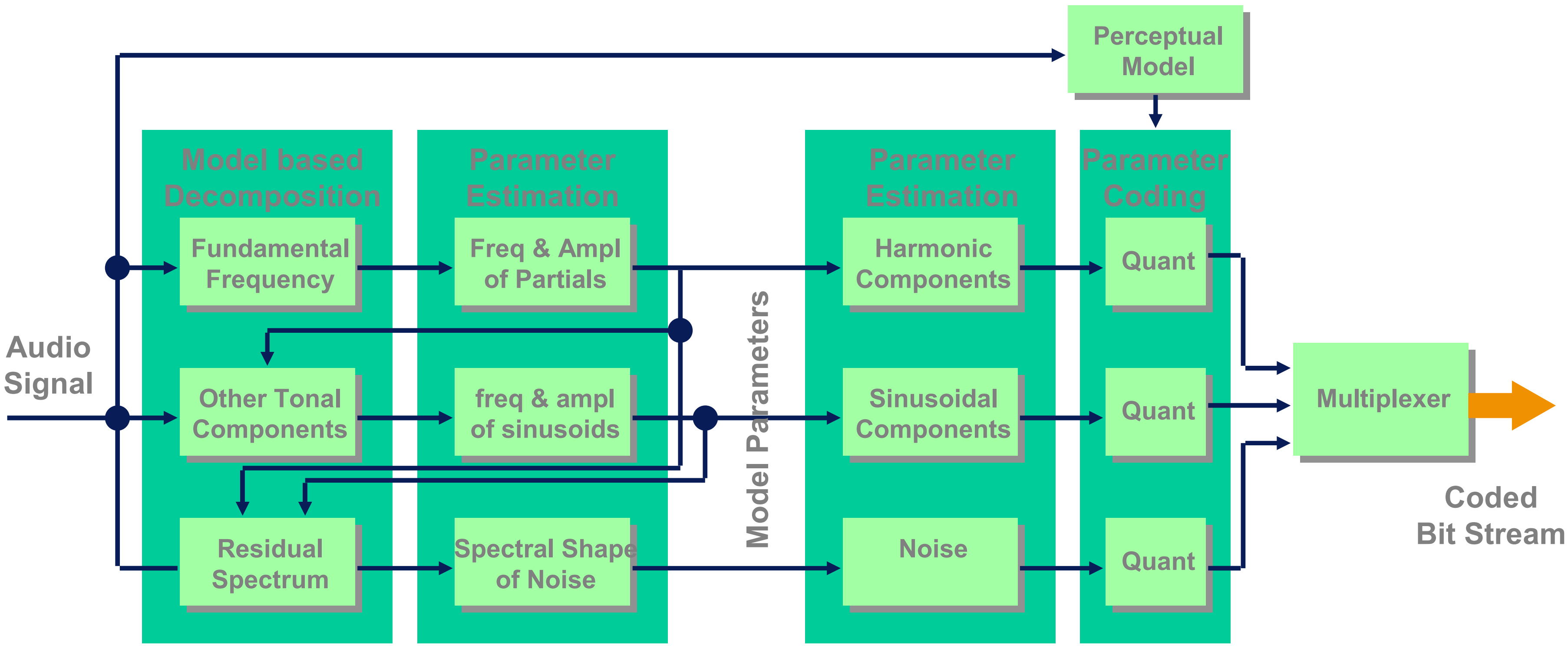
Conceptual Diagram of a class C Core Codec: Parameter-based audio coding



- **Based on parametric description of the audio signal**
- **Objects: Harmonic Tone, individual sinusoid, noise**

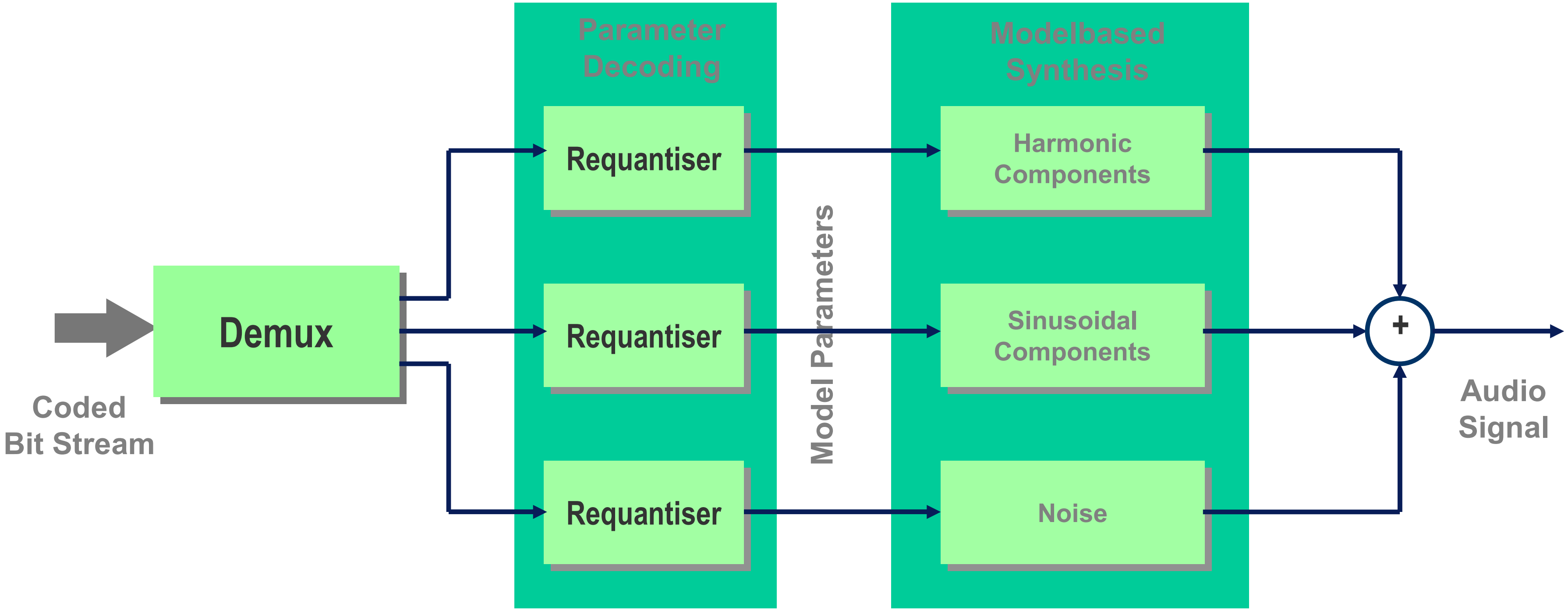
"Harmonic and Individual Lines plus Noise" (HILN) parametric coder

Conceptual Diagram of a class C Core Codec: Parametric-based audio coding

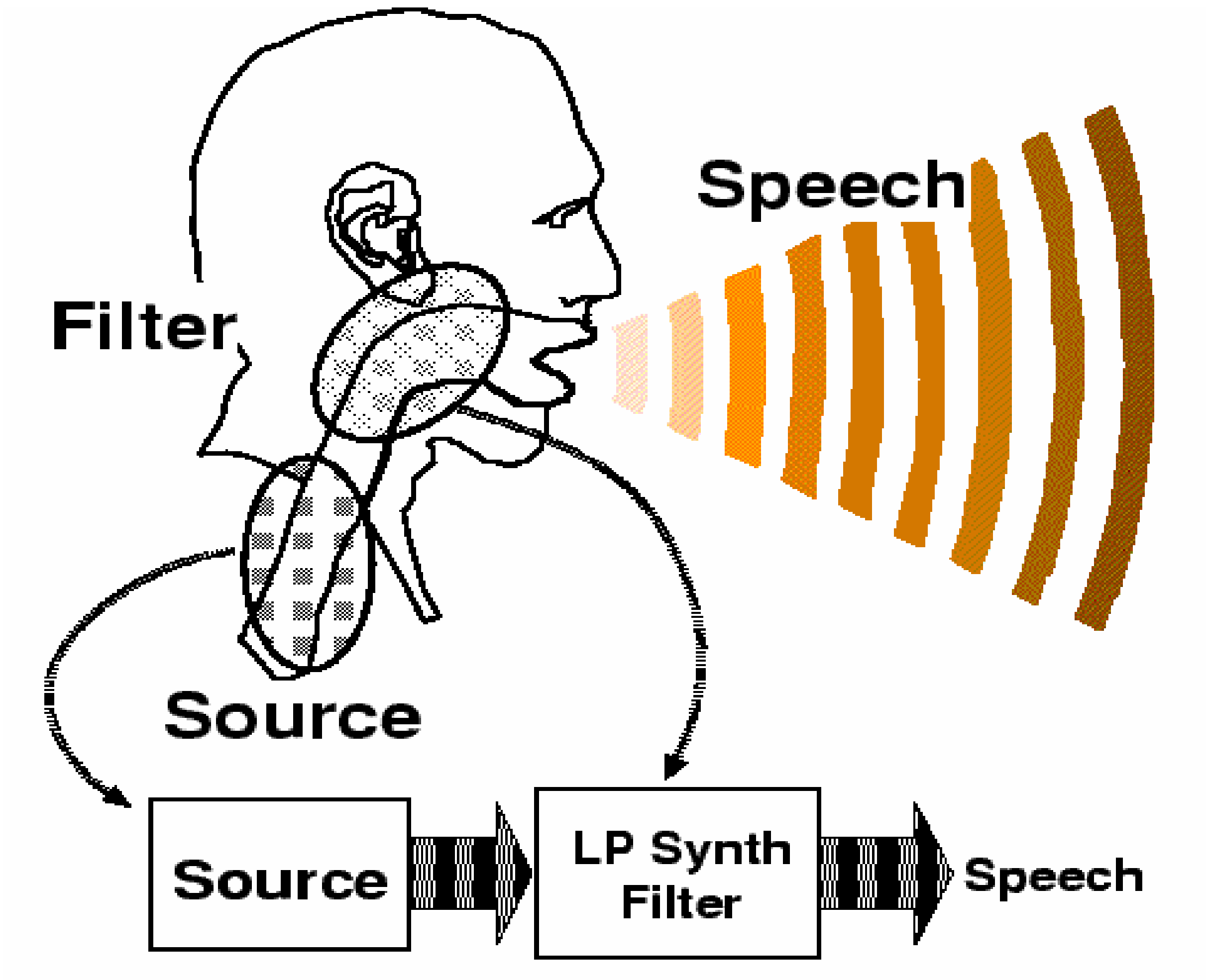


- Based on parametric description of the audio signal
- Objects: Harmonic Tone, individual sinusoid, noise
  - "Harmonic and Individual Lines plus Noise" (HILN) parametric coder

Conceptual Diagram of a class C Core Codec: Parametric-based audio decoding



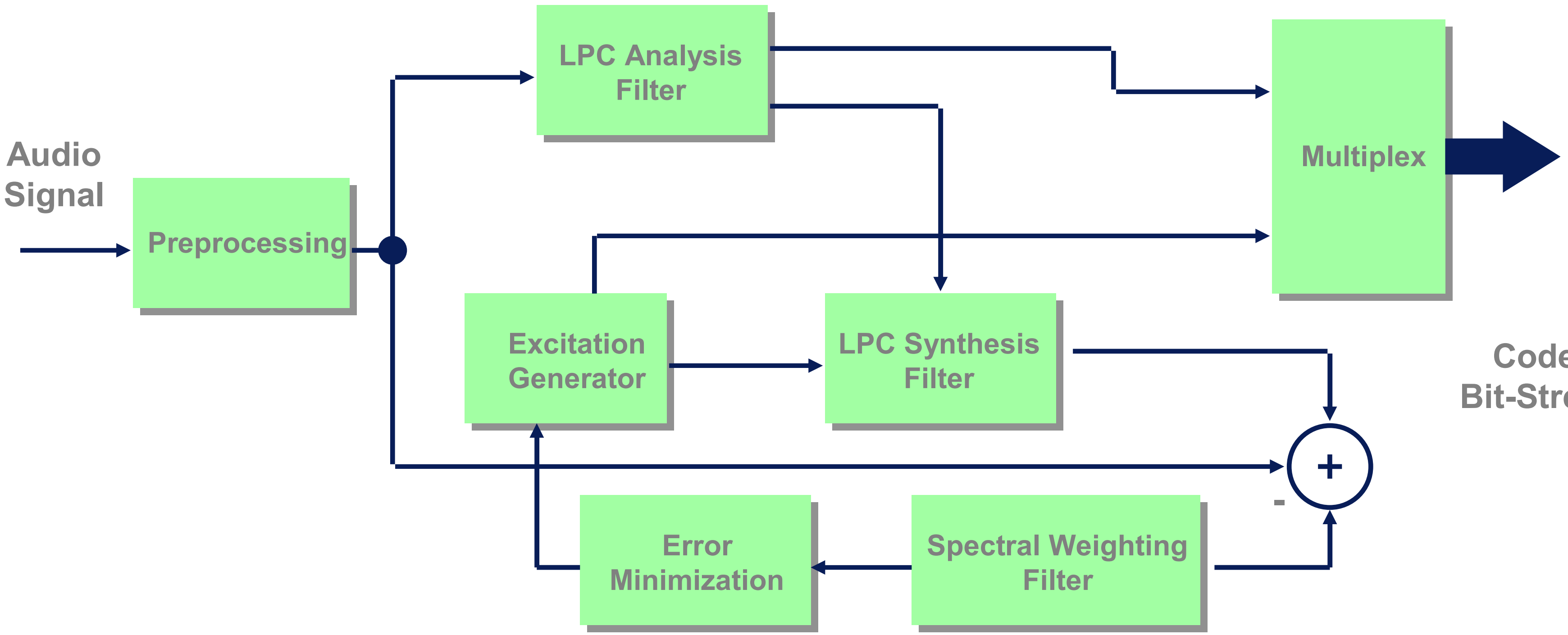
Basics of Speech Coding: Principles of Linear Prediction Coding (LPC)



LP Synth Filter: Linear Prediction Synthesis Filter

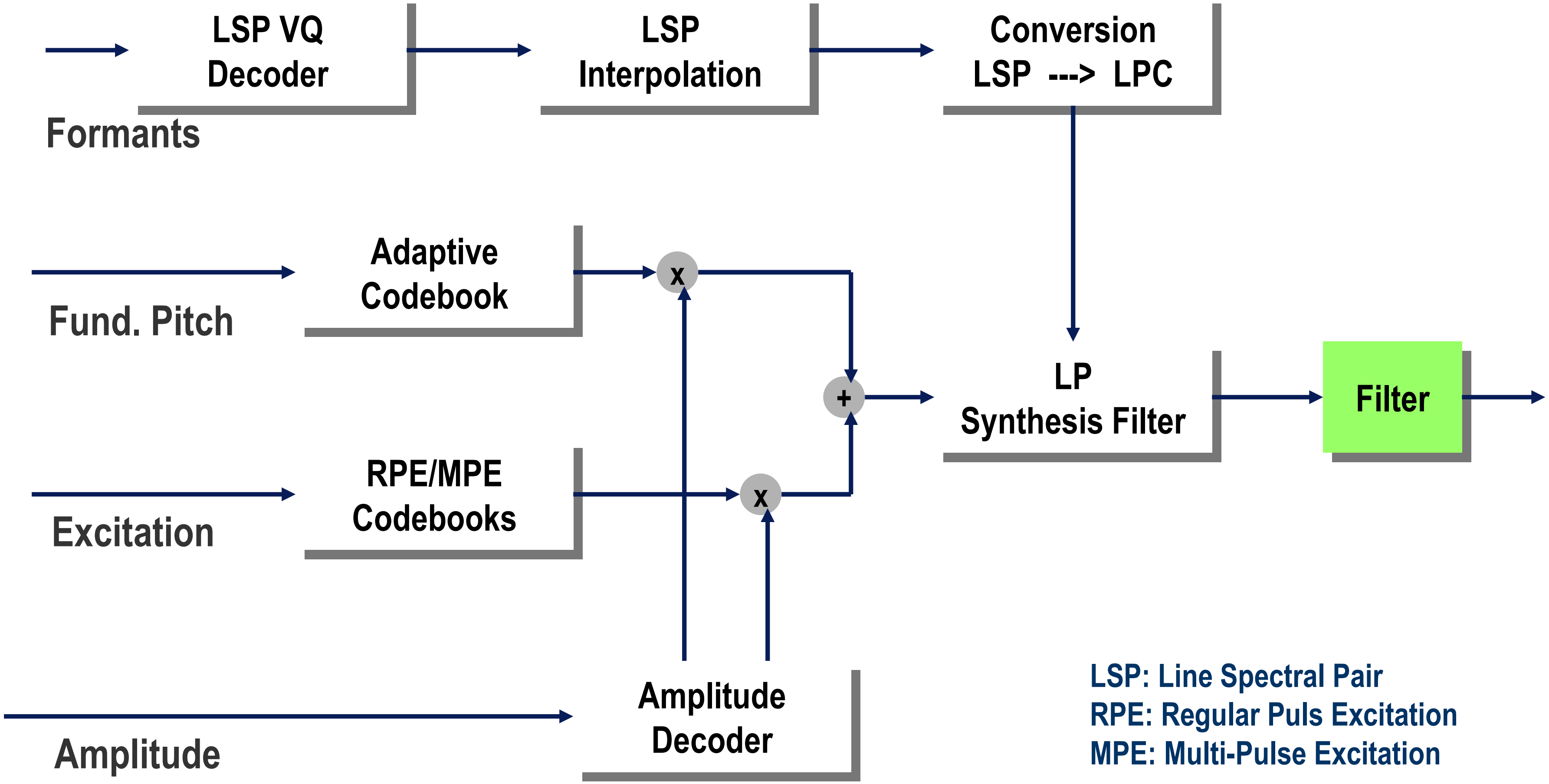
Source: www.cselt.it/leonardo/icjfiles/Mpeg-4_si

Conceptual Diagram of a class B Core Codec: LPC-based audio coding



- **Based on synthesis by analysis**
- **Performs best for speech between 6 kbit/s and 16 kbit/s per channel**
- **Multiple bit-rates**
- **Bit-rate and bandwidth scaleability**

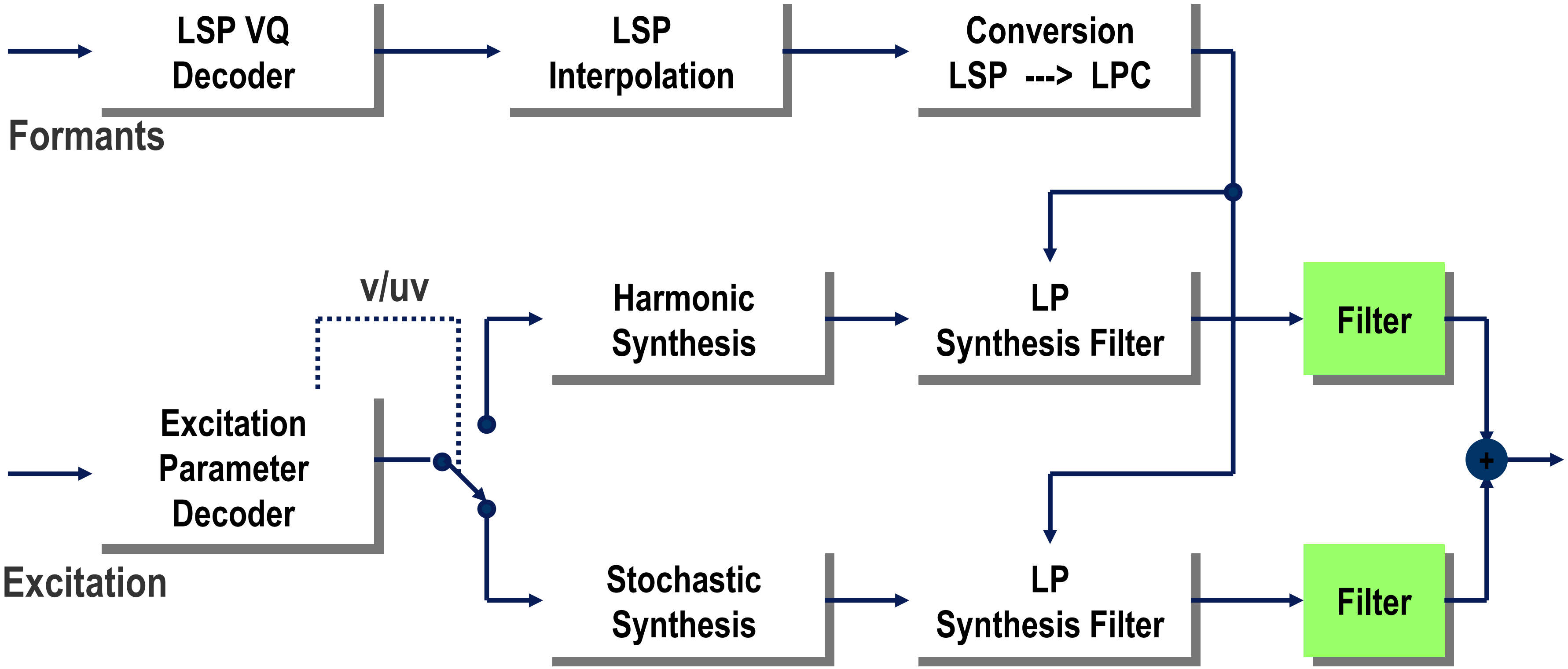
MPEG-4 Natural Speech Coding: Conceptual Diagram of MPEG-4 CELP Decoder



LSP: Line Spectral Pair
RPE: Regular Puls Excitation
MPE: Multi-Pulse Excitation

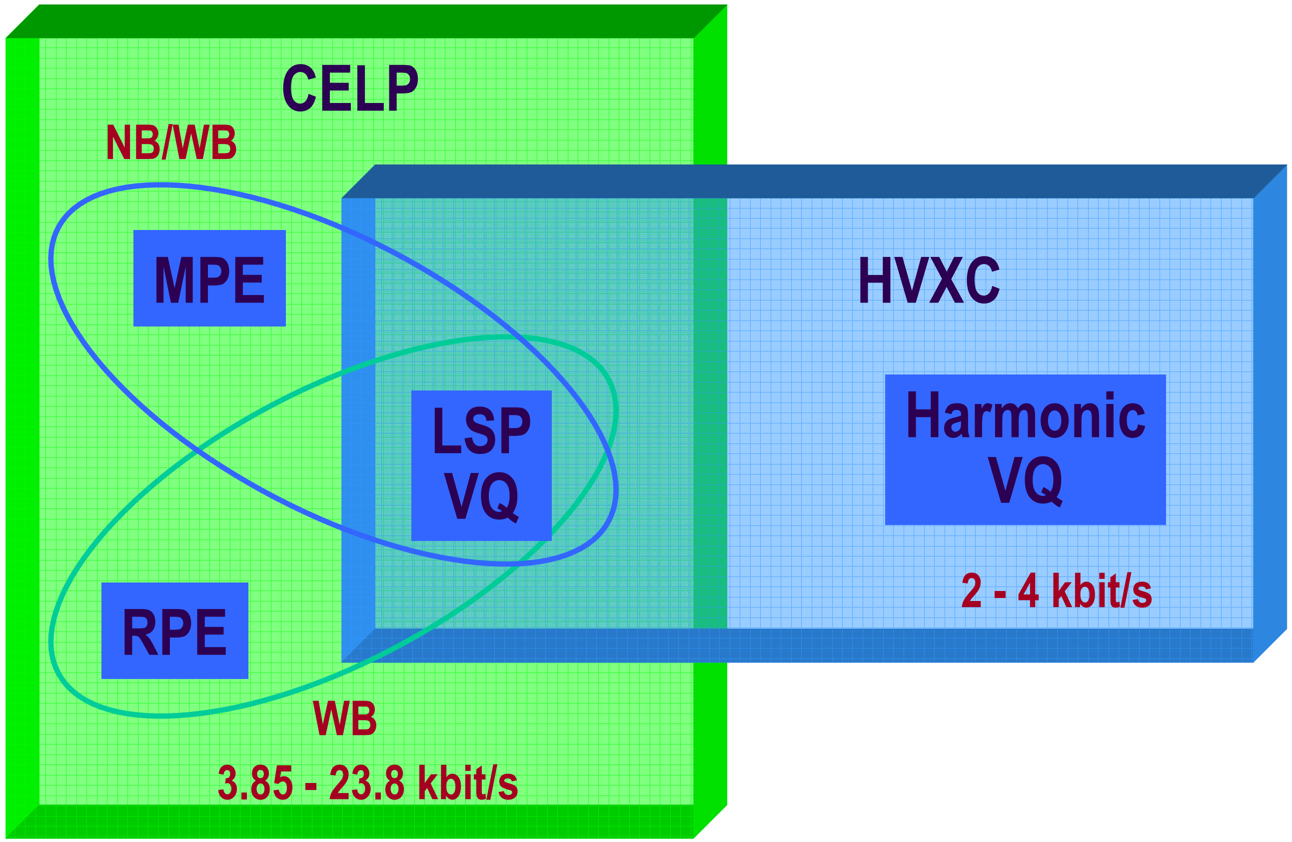
Source: B. Edler, Speech Coding in MPEG-4

MPEG-4 Natural Speech Coding: Conceptual Diagram of MPEG-4 HVXC Decoder



Source: B. Edler, Speech Coding in MPEG-4

MPEG-4 Natural Speech Coding Tools used in CELP and HVXC



HVXC: „Harmonic Vector eXcitation Coding“

Source: www.cselt.it/leonardo/icjfiles/Mpeg-4_si

MPEG-4 Natural Speech Coding:

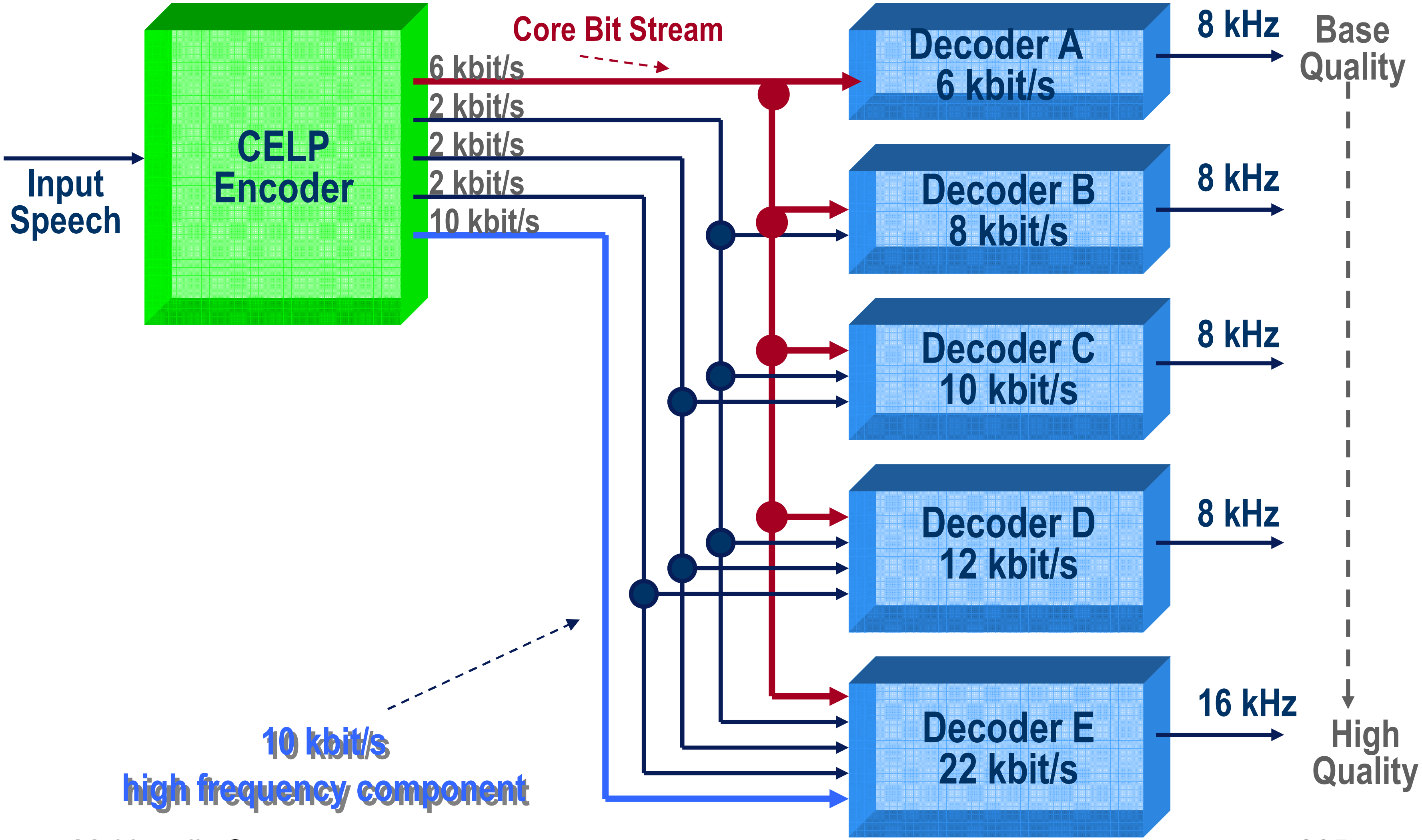
HVXC

Sampling Frequency	8 kHz
Bandwidth	300 ..... 3400 Hz
Bitrate	2 kbit/s and 4 kbit/s
Frame Size	20 ms
Delay	33,5 ..... 56 ms
Features	Multi Bit-rate coding, Bit-rate scalability

CELP

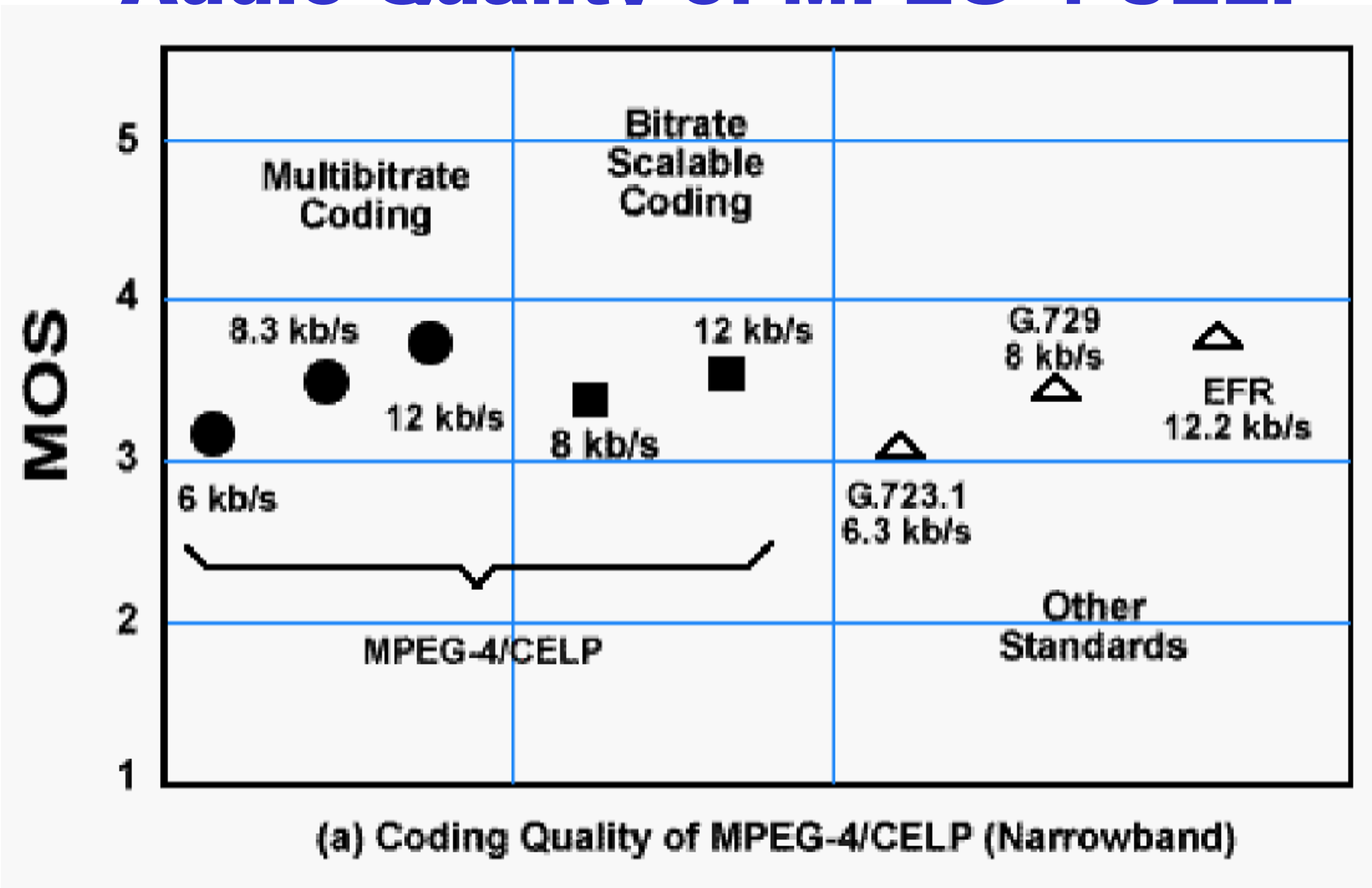
Sampling Frequency	8 kHz	16 kHz
Bandwidth	300 ..... 3400 Hz	50 ..... 7000 Hz
Bit-rate	3,85 ..... 12,2 kbit/s 28 Bit-rates	10,9 ..... 23,8 kbit/s 30 Bit-rates
Frame Size	10 ..... 40 ms	10 ..... 20 ms
Delay	10 ..... 45 ms	15 ..... 26,75 ms
Features	Multi Bit-rate Coding Bit-rate Scalability Bandwidth Scalability	

MPEG-4 Natural Speech Coding Tools
Scalable Bit-rate Coding



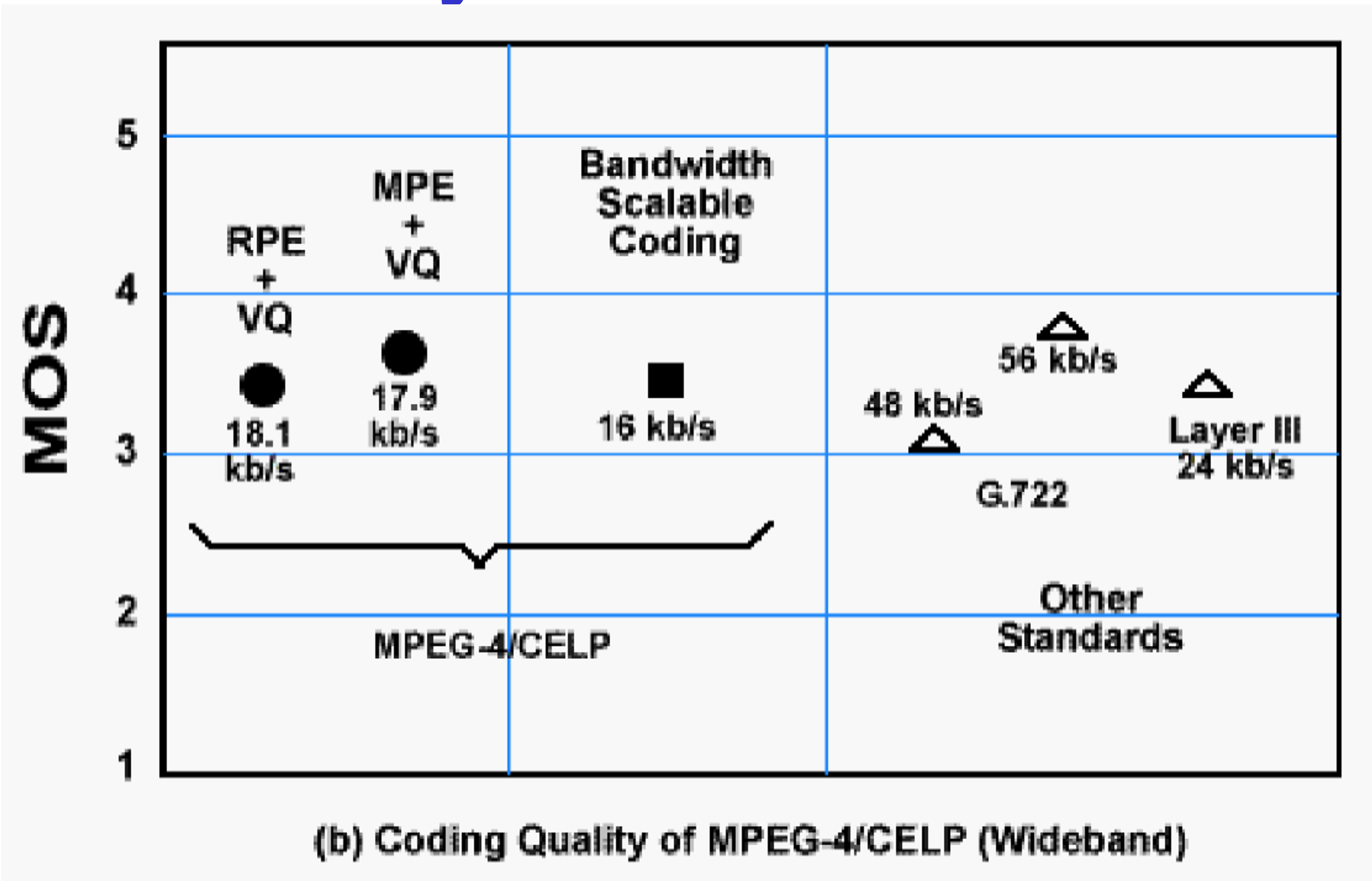
MPEG-4 Natural Speech Coding Tools:

Audio Quality of MPEG-4 CELP



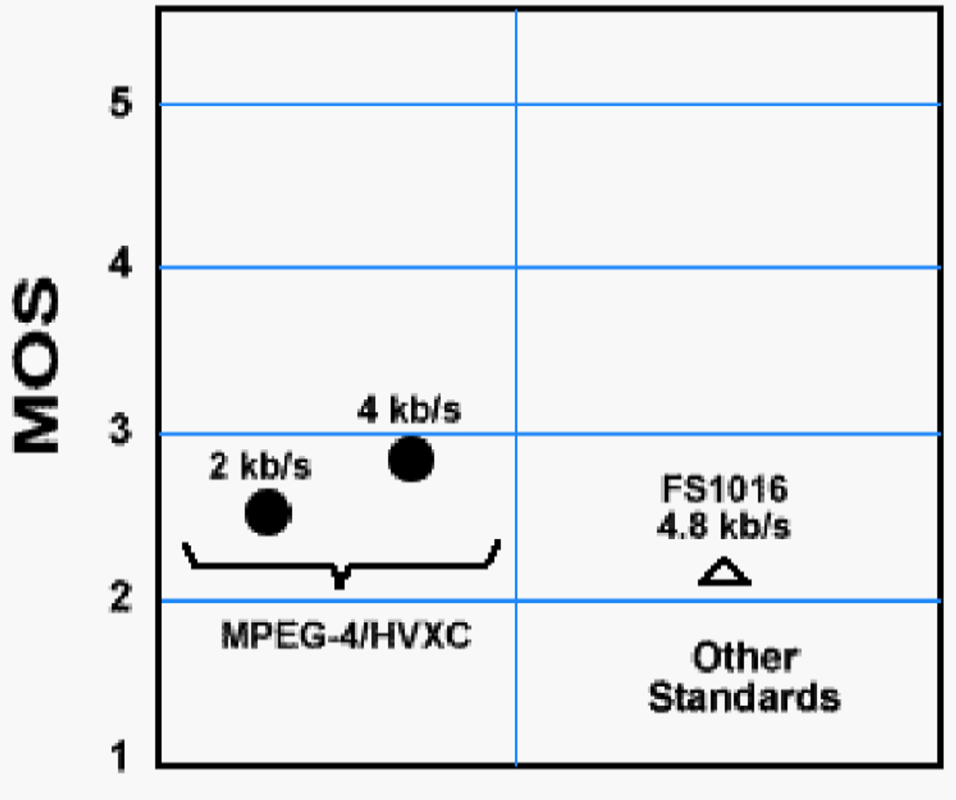
Source: www.cselt.it/leonardo/icjfiles/Mpeg-4_si

MPEG-4 Natural Speech Coding Tools: Audio Quality of MPEG-4 CELP Wideband



Source: www.cselt.it/leonardo/icjfiles/Mpeg-4_si

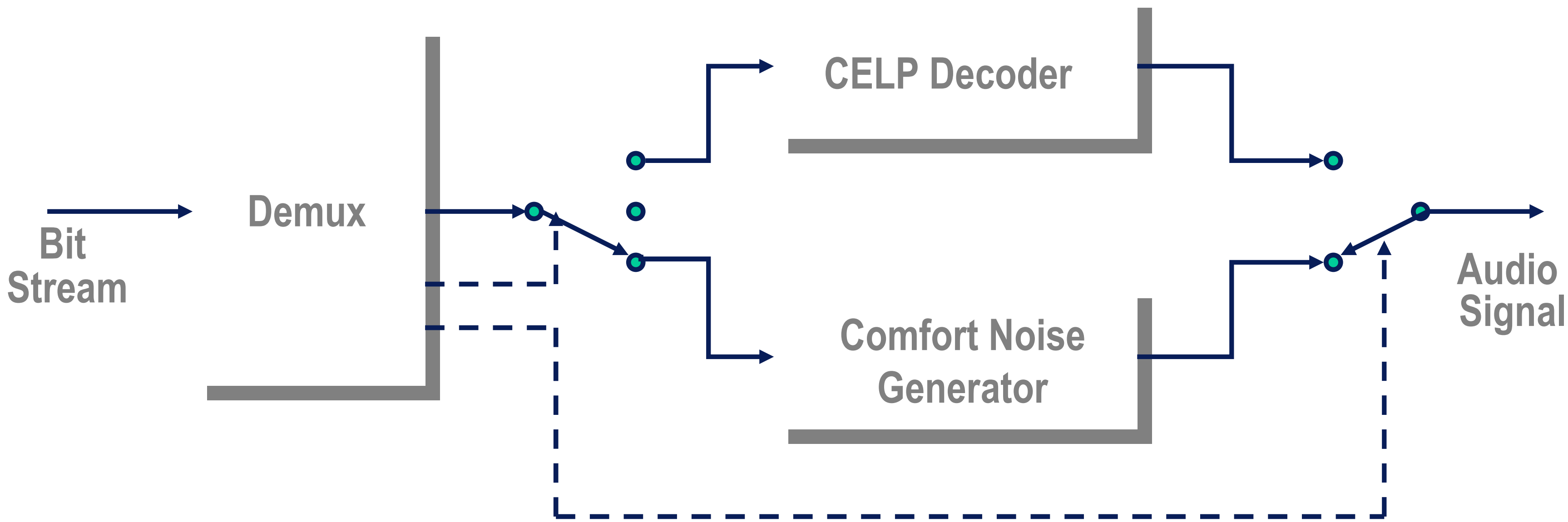
MPEG-4 Natural Speech Coding Tools: Audio Quality of MPEG-4 HVXC



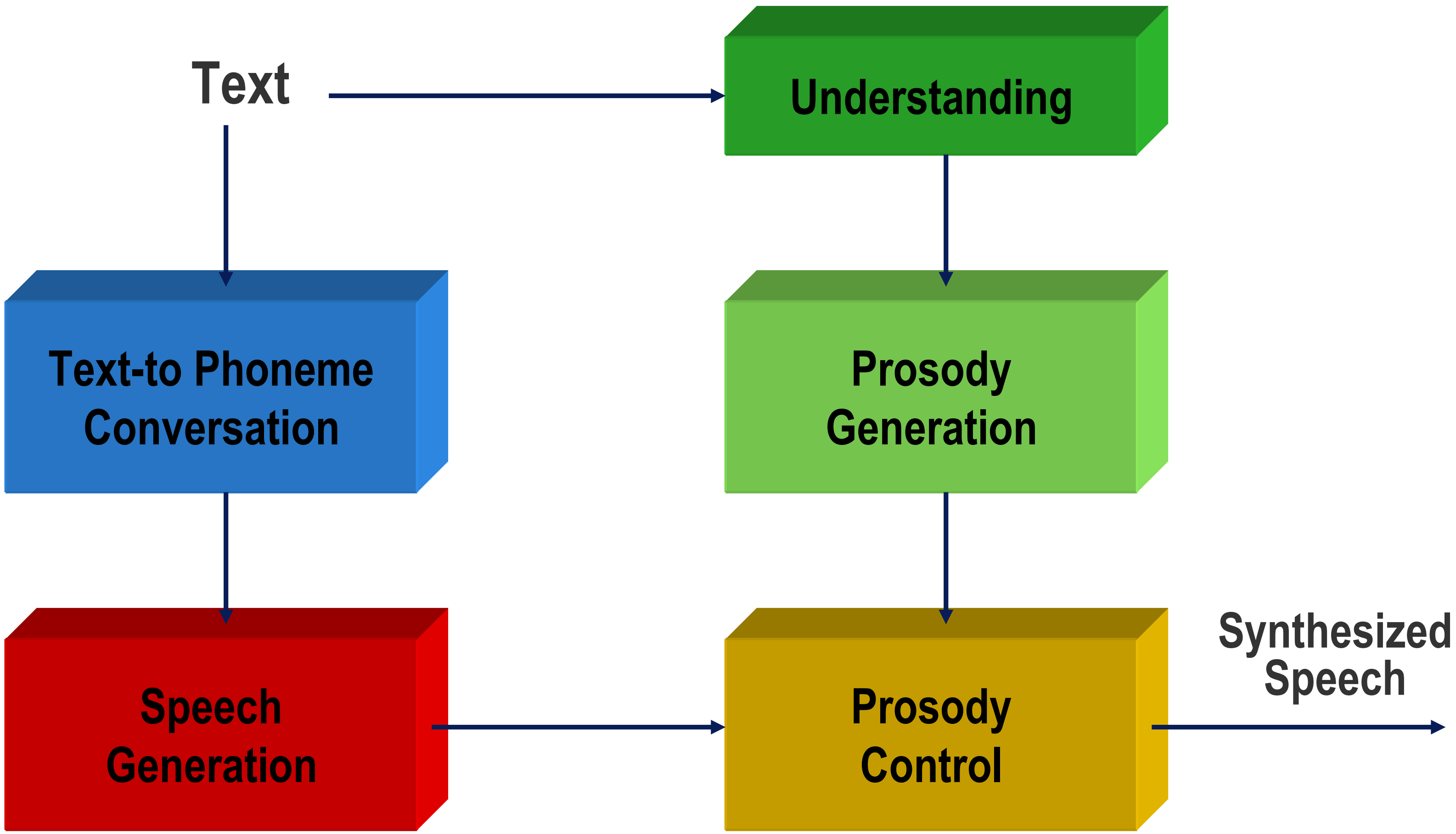
(c) Coding Quality of MPEG-4/HVXC

Source: www.cselt.it/leonardo/icjfiles/Mpeg-4_si

MPEG-4 Audio CELP decoder with Silence Compression Tool

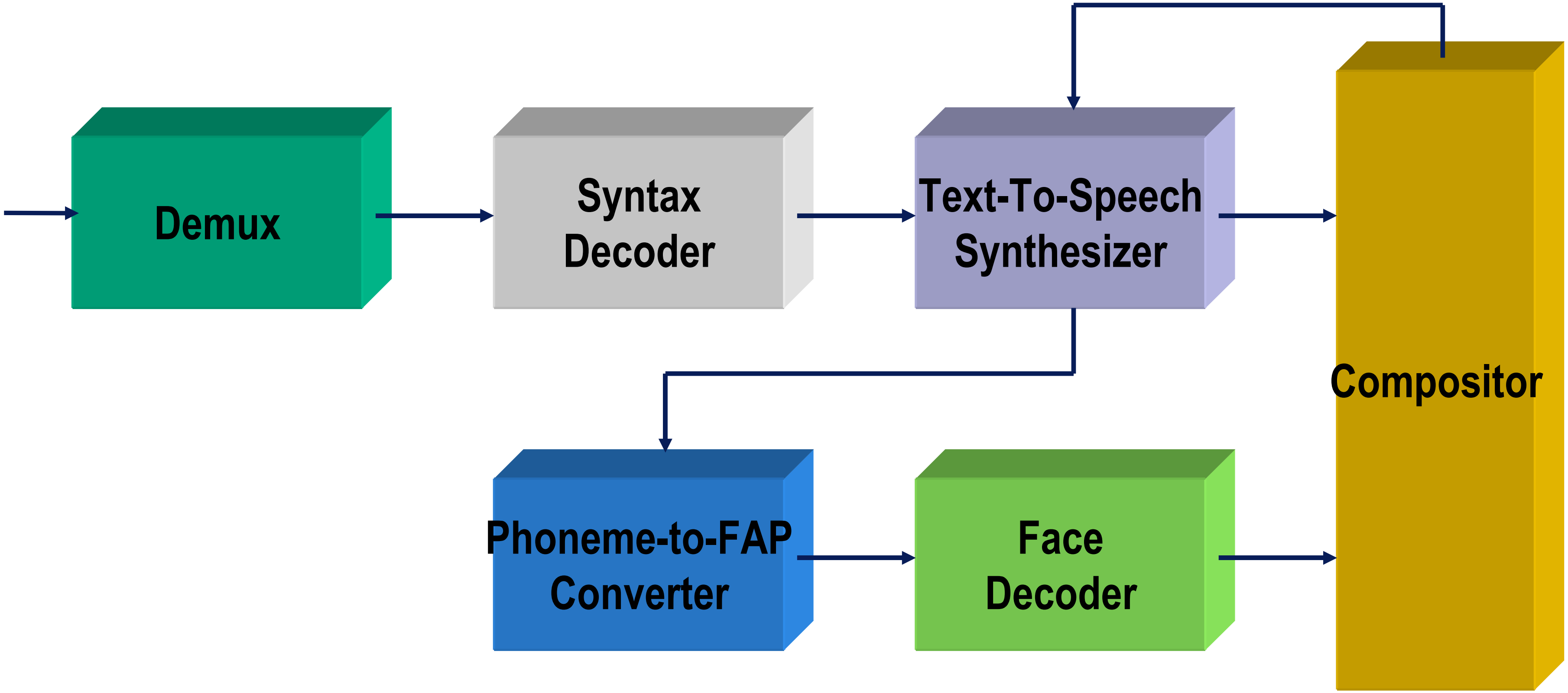


MPEG-4 Text-To-Speech Encoding: Conceptual Diagram



Source: www.cselt.it/leonardo/icjfiles/Mpeg-4_si

MPEG-4 Text-To-Speech Decoding: Conceptual Diagram



Source: www.cselt.it/leonardo/icjfiles/Mpeg-4_si

MPEG-4 SAOL: Structured Audio Orchestra Language

- **SAOL allows for transmission and decoding of**
  - Synthetic sound effects
  - Music
- **High-Quality audio can be created at extremely low bandwidth**
  - Synthetic music possible with 0.01 kbit/s
  - Subtle coding of expressive performance using multiple instruments possible with 2 ... 3 kbit/s
- **MPEG-4 standardizes a method for describing synthesis methods**
  - a particular set of synthesis method is not standardized
- **Any current or future sound-synthesis method may be described in SAOL**

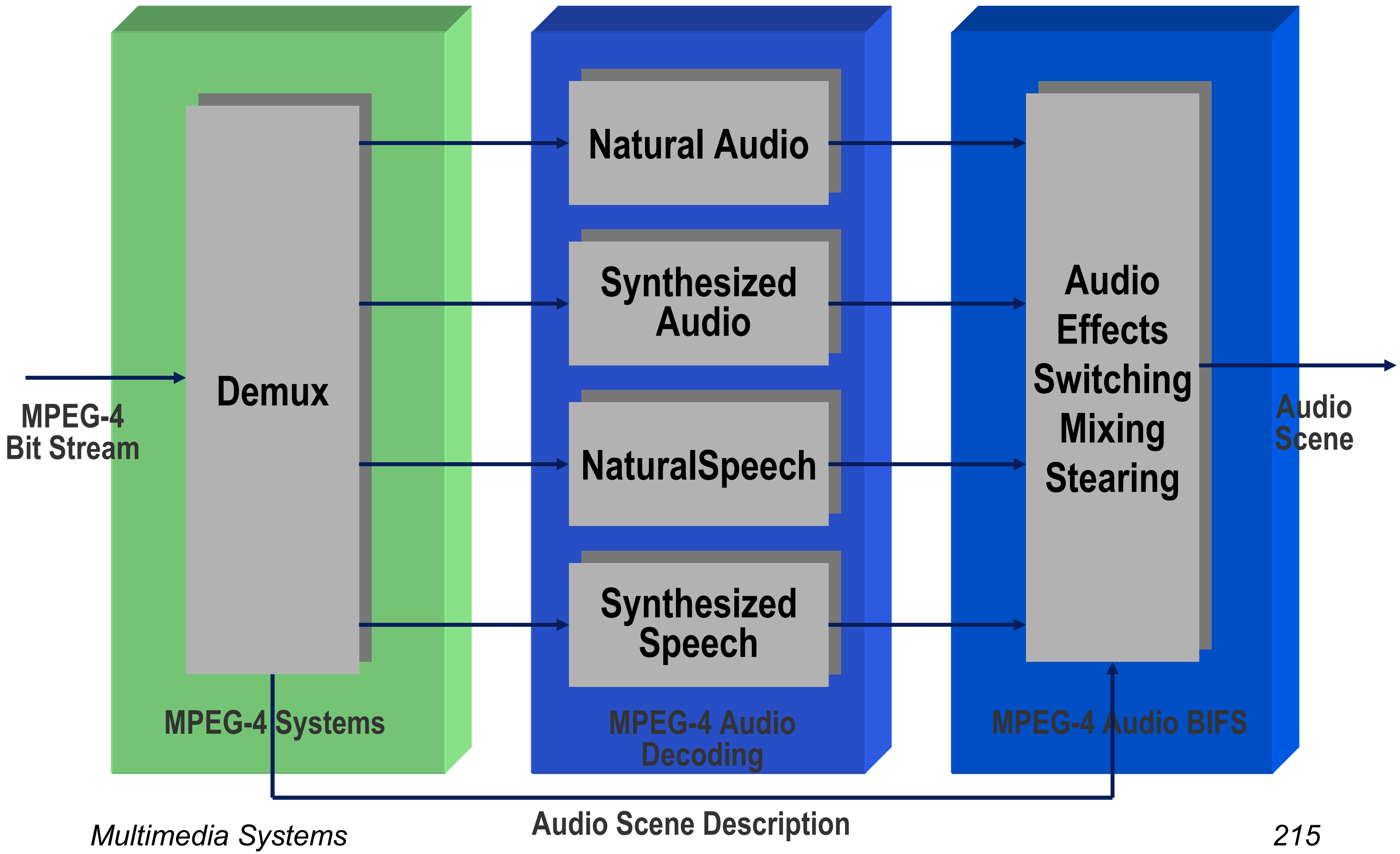
MPEG-4 SAOL: Five major elements to the Structured Audio Toolset

- SAOL is a digital signal processing language which allows
  - for description of arbitrary synthesis and control algorithms
- SASL (Structured Audio Score Language) is a score and control language and describes
  - the manner in which sound-generation algorithms are used to produce sound
- SASBF (Structured Audio Sample Bank Format) allows
  - for transmission of banks of audio samples
  - for description of simple processing algorithms (wavetable synthesis)
- Scheduler description
  - is supervisory run-time element for the SA decoding process
  - maps structural sound control to real-time events
- Normative reference to MIDI (structural control) standards
  - can be used in conjunction with or instead SASL

MPEG-4 Audio:
Coding systems and their typical data rates

Type of Coding Technique	Bitrate in kbit/s
• SAOL (Structured Audio Orchestra Language)	0.01....10
• CELP Narrow Band Speech codec	6
• G.723.1 Narrow Band Speech codec	6
• Wide Band CELP	18
• AAC (Advanced Audio Coding)	18....24
• AAC TwinVQ	24 (8+16)
• AAC CELP	24 (6+18)
• HILN (Harmonic ind. Line and Noise codec)	< 32
• BSAC (Bit Sliced Arithmetic Coding)	
Dynamic scalable AAC	*16...64*
• TTS (Text-To-Speech conversation system)	

MPEG-4 Audio:
Conceptual Diagram of complete Audio Decoder



MPEG-4 Audio:
The AudioBIFS modes

Node Name	*Functionality*
AudioSorce	Connect decoder to scene graph
Sound	Connect audio subgraph to visual scene
AudioMix	Mix multiple channels of sound together
AudioSwitch	Select a subset of a set of channels of sound
AudioDelay	Delay a set of audio channels
AudioFX	Perform audio effects-processing
AudioBuffer	Buffer sound for interactive playback
Listening Point	Control position of virtual listener
TermCap	Query resources of terminal

MPEG-4 Audio: Range of applications

- **Audio on demand on the Web at very low bit-rates**
- **Digital Radio Broadcasting in narrow band channels**
- **Video services on the Web**
- **Interactive multimedia on mobiles (point-to point or point-to-multipoint, e.g. with UMTS standard)**

  MPEG-4 video and audio standards have embedded error resilience
- **Digital multimedia broadcasting**
- **Electronic Program Guides (EPG)**

  MPEG-4 2D composition profiles
- **Virtual Reality experiences on the Web**

  MPEG-4 high compression, partial streams
- **Interactive local multimedia**

  Complex virtual world on a DVD-ROM with last minute updates via the web or a broadcast channel

SAMBITS:
Systems for Advanced Multimedia Broadcast and IT Services

- Main Objectives

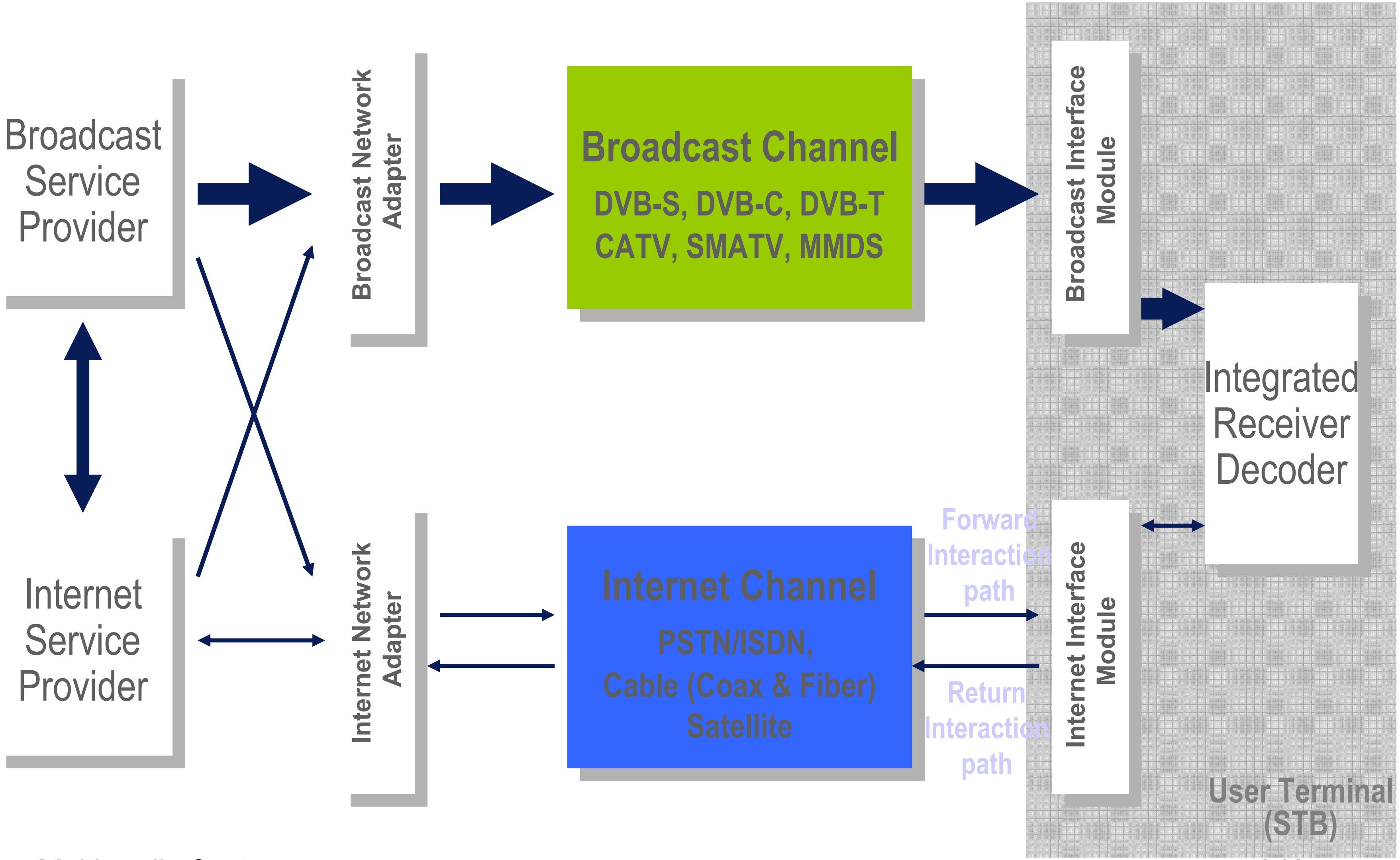
  bring MPEG-4 and MPEG-7 for broadcast technology together with related Internet services

  provide multimedia services to a terminal that can display any type of general interest integrated broadcast/internet services with local and remote interactivity (combination of DVB and Internet infrastructure)

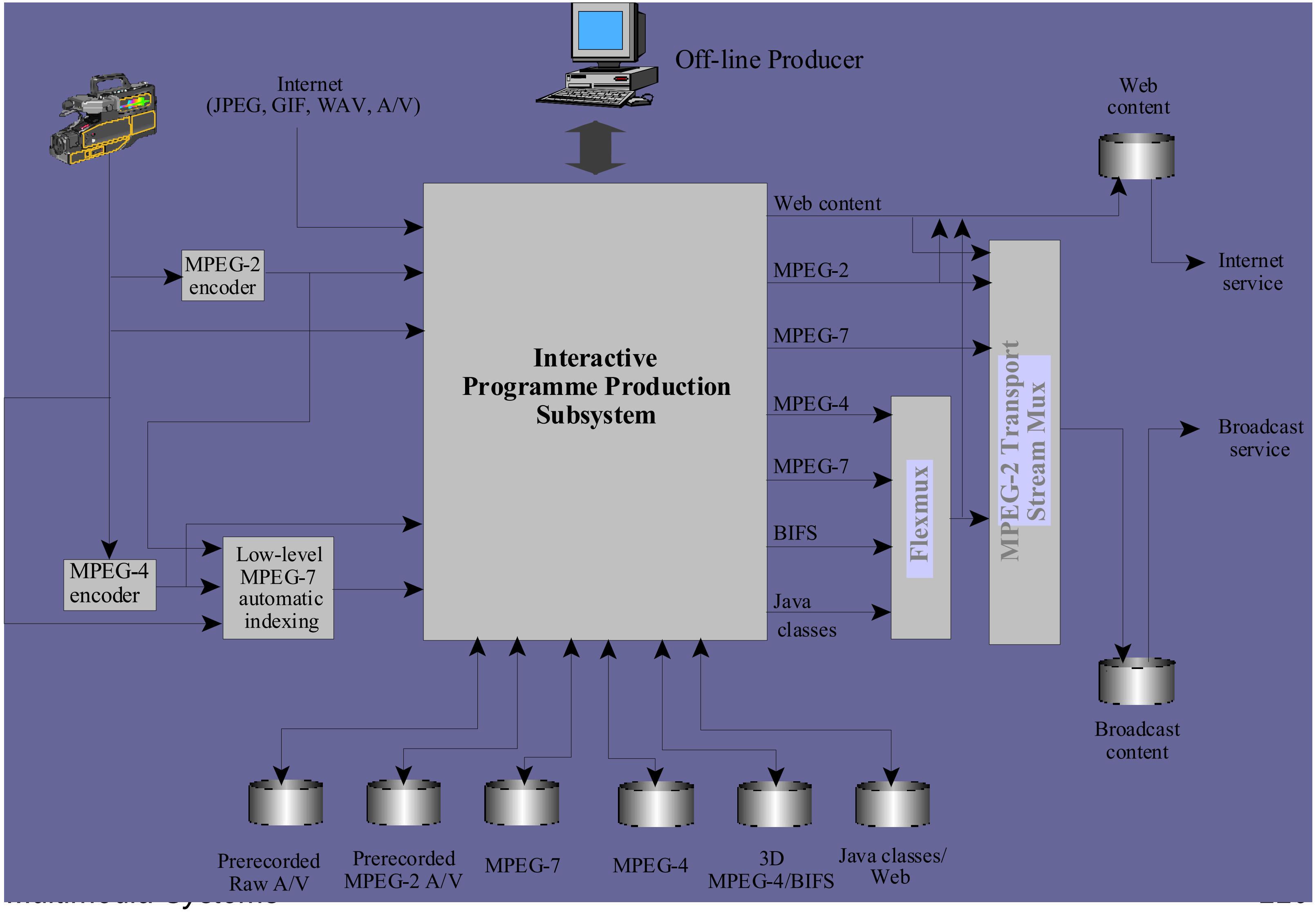
  Improving quality of Web and broadcast experience

  Services consisting of high quality video enhanced by multimedia elements and interactive personalised information retrieval through the Internet

SAMBITS:
The Reference Chain



SAMBITS:


Off-line Producer
Internet
(JPEG, GIF, WAV, A/V)
Web
content
MPEG-2
encoder
Interactive
Programme Production
Subsystem
Web content
MPEG-2
MPEG-7
MPEG-4
MPEG-7
BIFS
Java
classes
Flexmux
MPEG-2 Transport
Stream Mux
Internet
service
Broadcast
service
MPEG-4
encoder
Low-level
MPEG-7
automatic
indexing
Broadcast
content
Prerecorded
Raw A/V
Prerecorded
MPEG-2 A/V
MPEG-7
MPEG-4
3D
MPEG-4/BIFS
Java classes/
Web

MPEG-4 Audio: Audio Demonstration: Demo contains Speech (Soccer Match) and Music (Pop)

- **MPEG-4 parametric based coder:**
  Harmonic and Individual Lines plus Noise (HILN)

  Original - 8 kbit/s - 14 kbit/s

- **Proprietary algorithm (QDesign):**
  Proposed for Audio on the Internet

  Original - 8 kbit/s - 12 kbit/s - 24 kbit/s - 56 kbit/s

- **MPEG-4 t/f based coder:**
  Advanced Audio Coding (AAC)

  Original - 8 kbit/s - 24 kbit/s - 56 kbit/s - Original

Let's Listen!