

# **Perceptual Audio Coding**

## **An Overview**

### **of Basic Principles and Standards**

**Marina Bosi**

**Stanford University**



**IEEE Signal Processing Society**

**Santa Clara Valley Chapter**

**March 5, 2015**

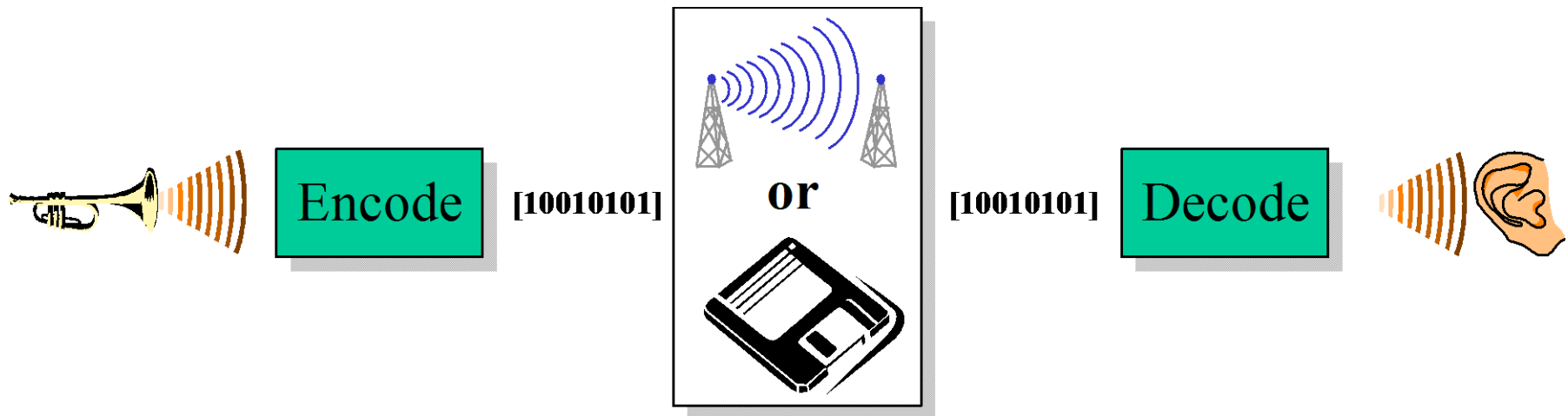
# What Will We Be Talking About?

- **General Ideas Behind Perceptual Audio Coding**
- **What to Listen For?**
- **Design Choices in State-of-the-Art Coders**
- **What will the future bring?**

# **General Ideas Behind Audio Coding**

# Audio Coding

- In general, an audio coder (or codec) is an apparatus whose input is an audio signal and whose output is an audio signal which is perceptually identical (or at least very close) to the (somewhat delayed) input signal**



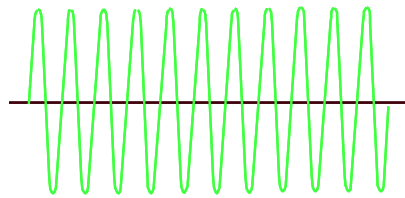
# Audio Coding Chain

- **The beginning of the coding chain is the source of the sound**
  - Source modeling is important in order to optimize the audio signal representation
  - Unfortunately, the exact nature of the source is not known a priori, we only know the statistical distribution of audio signals
- **The last stage is the human ear**
  - Modeling of the ear and the processing of the acoustical stimuli is important to minimize the irrelevant data contained in the signal representation
  - Heuristic models based on the outer/middle/inner ear behavior are widely used in perceptual audio coding

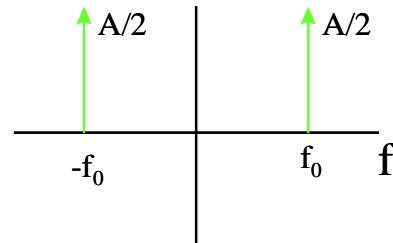
# Redundancy

“Redundant *adj* 1. Exceeding what is necessary or normal. 2. Characterized by or containing an excess: *specif* more words than necessary....” [Websters Dictionary]

- In audio coding redundant means that the same information can be represented with fewer bits
- For example, consider a sine wave signal:
  - Redundant: sample the waveform 44,100 times per second and describe each sample with 16 bits
  - Concise: Describe the amplitude, frequency, phase, and duration



$x(t)$



$X(f)$




- Notice that the concise representation of the sine wave is basically equivalent to the information in its Fourier Transform.
- Since music and many other audio signals are very tonal, most coders work in the frequency domain to reduce redundancy

# Irrelevancy



•“Irrelevant *adj* 1. Not having significant and demonstrable bearing on the matter at hand.” [Websters Dictionary]

- In audio coding irrelevant data means that you can't hear any difference in the audio signal if those data are omitted
- Main causes of irrelevancy:
  - Hearing Threshold
  - Masking
- Hearing Threshold
  - We can't hear sounds below a certain frequency-dependant level
- Masking
  - Loud sounds can prevent us from hearing softer sounds nearby in time or frequency
- Exploiting irrelevancy
  - Don't code signal components you can't hear
  - Only quantize audible signal components with enough bits to keep quantization noise below the level it can be heard

# Demo: 13 dB Miracle

- The “13 dB miracle” paradox (Johnston and Brandenburg ‘91), where the original signal  was injected with noise that was either
  - a) shaped according to psychoacoustic masking models 
  - b) white 

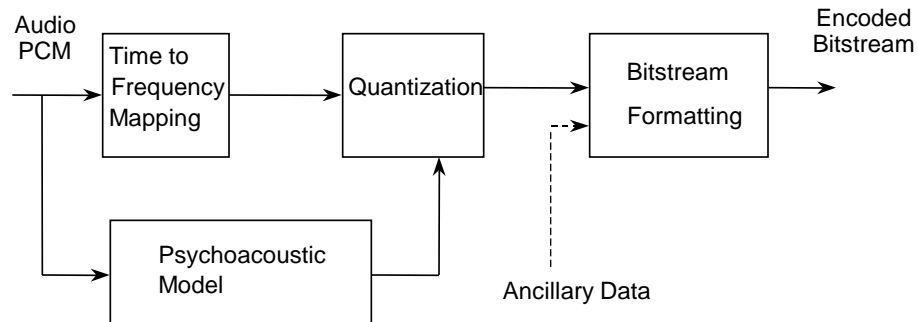
**shows that two systems with identical SNR = 13 dB have very different perceived audio quality**

- In case a) the quantization noise is shaped so that it is contained below masked thresholds 
- In case b) the quantization noise is shaped so that it is uniformly distributed in frequency (in general above masked thresholds) 

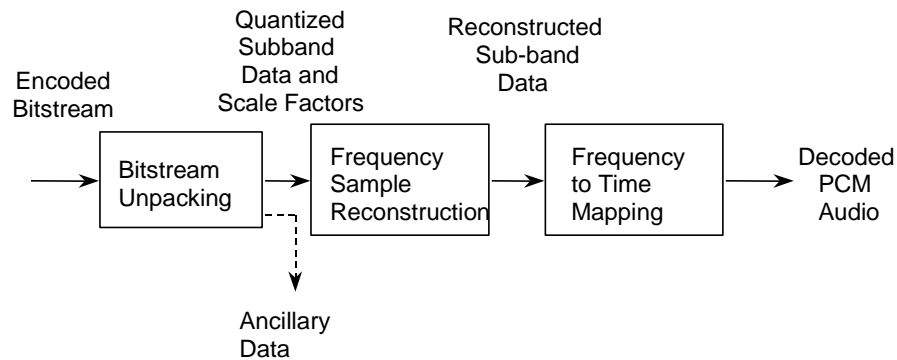


# Building Blocks for a Perceptual Audio Coder

Encode:



Decode:



# The Main Building Blocks

- **Time to Frequency Mapping**
  - Subdivides the signal content into its frequency components
  - Frequency representation of the signal is a good framework for exploiting redundancies and irrelevancies in the signal
  - Sets up the stage for the basic coding efficiency (or coding gain) of the coding system
    - More frequency lines imply higher coding gains
    - Longer blocks may cause temporal artifacts (pre-echo)
- **Psychoacoustic Models**
  - Based on the signal frequency components and heuristic models determines an “irrelevant” noise threshold for each block of the input signal
- **Quantization and coding**
  - Spectral shape of quantization noise is designed to be below the noise threshold to avoid audibility of coding artifacts

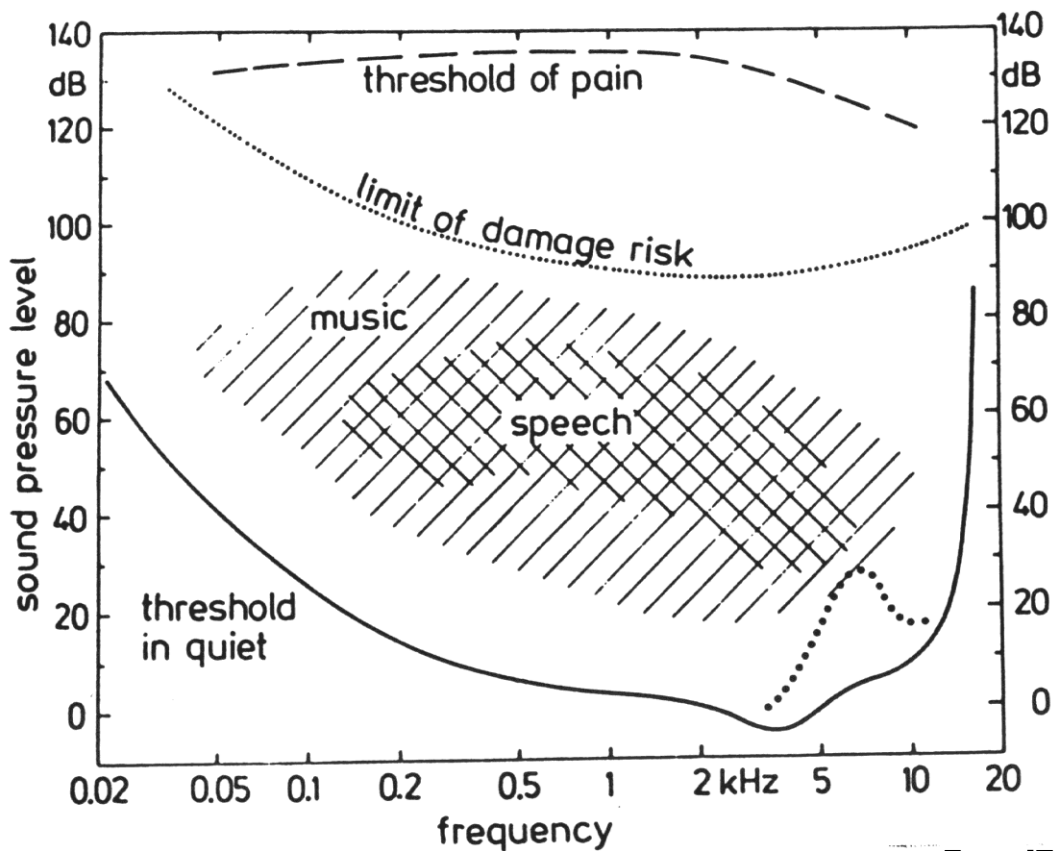
# Time-to-Frequency Mapping

- **PQMF**
  - MPEG Layers I and II: 32-band, 511 PQMF
- **DCT**
  - OCF: 512 frequency lines; 1024 impulse response
- **MDCT/MDST**
  - AC-2A: 256/64 frequency lines; 512/128 impulse response
- **MDCT**
  - AC-3: 256/128 frequency lines; 512/256 impulse response
  - MPEG AAC, PAC: 1024/128 frequency lines; 2048/256 impulse response
- **Hybrid**
  - (PQMF + MDCT) MPEG Layer III : 576/192 frequency lines; 1664/896 impulse response
  - (QMF + MDCT) ATRAC : 512/64 frequency lines; 1072/304 impulse response
  - (MDCT + DCT) E-AC-3 : 1536/128 frequency lines; 3072/256 impulse response
- **Wavelets (EPAC)**
  - Tree structure with higher frequency resolution at low frequencies and higher temporal resolution at high frequencies, utilized during transients only
- **Int MDCT (Lossless Coding)**
  - MPEG-4 SLS : same as AAC with 4x over sampling also enabled

# Masking Phenomena

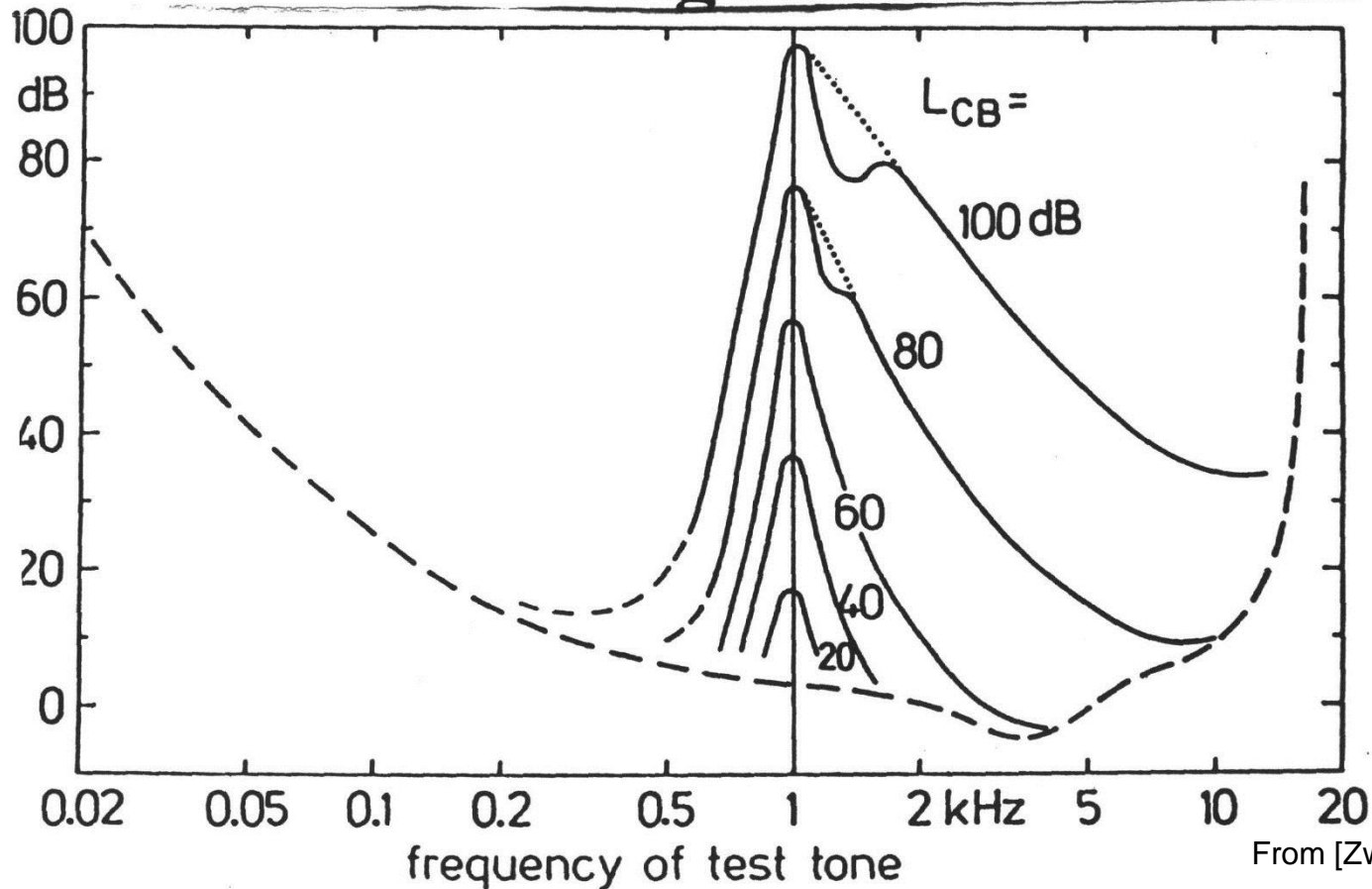
- **In perceptual audio coding signal irrelevancies are exploited**
  - Only quantize audible signal components with enough bits to keep quantization noise below the level that can be heard
- **Main causes of irrelevancy:**
  - **Hearing Threshold**
    - We can't hear sounds below a certain frequency-dependent level
  - **Masking**
    - Loud sounds can prevent us from hearing softer sounds nearby in time or frequency
- **Masking depends on the spectral composition of both masker and maskee, on their temporal characteristics, and intensity**
  - “Asymmetry of Masking” [Hellman 72], i.e. noise-like signals are “better masker” than tone-like signals
  - The shape of the masking curve depends on the masker's level
- **Masking most pronounced at the center frequency of the masker and it is “flat” within a critical band**

# Hearing Threshold



From [Zwicker & Fastl 90]

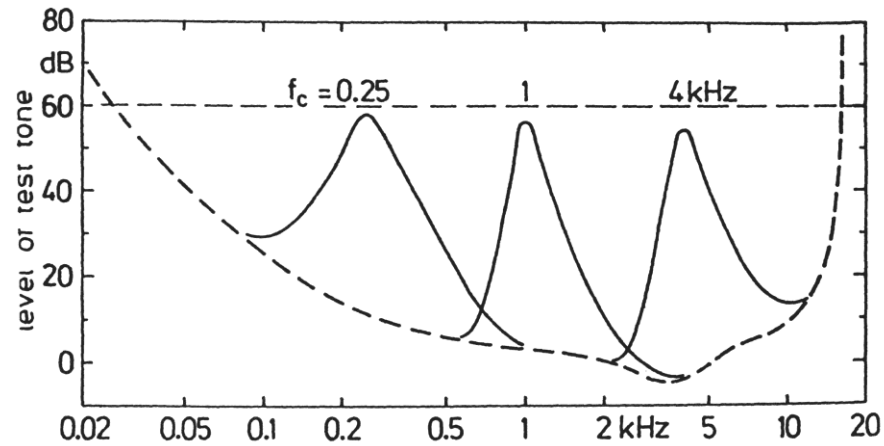
# Masking Thresholds at Different Masking Levels



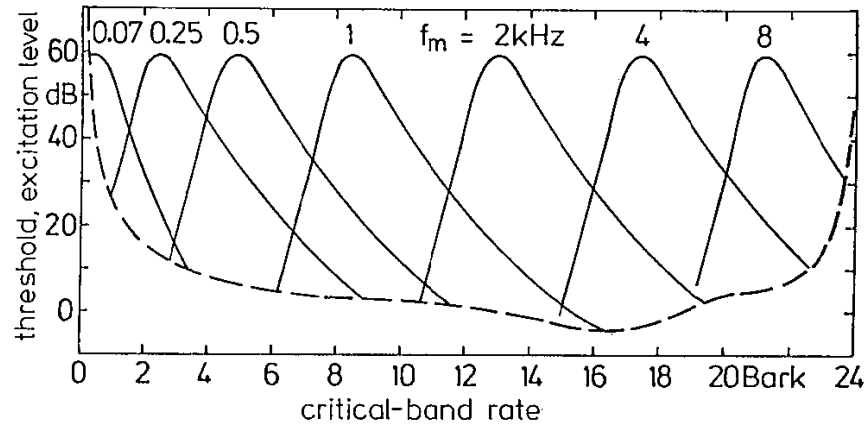
From [Zwicker & Fastl 90]

# The Bark Scale and Masking

Log Frequency Representation



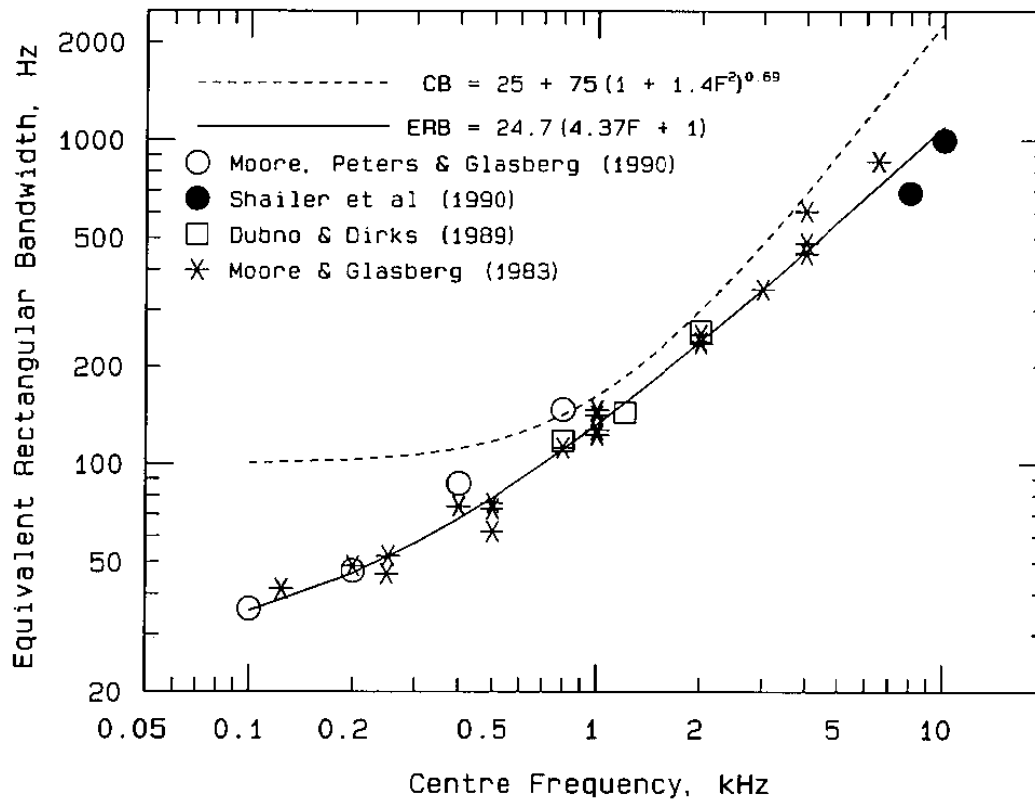
Bark Scale Representation



From [Zwicker & Fastl 90]

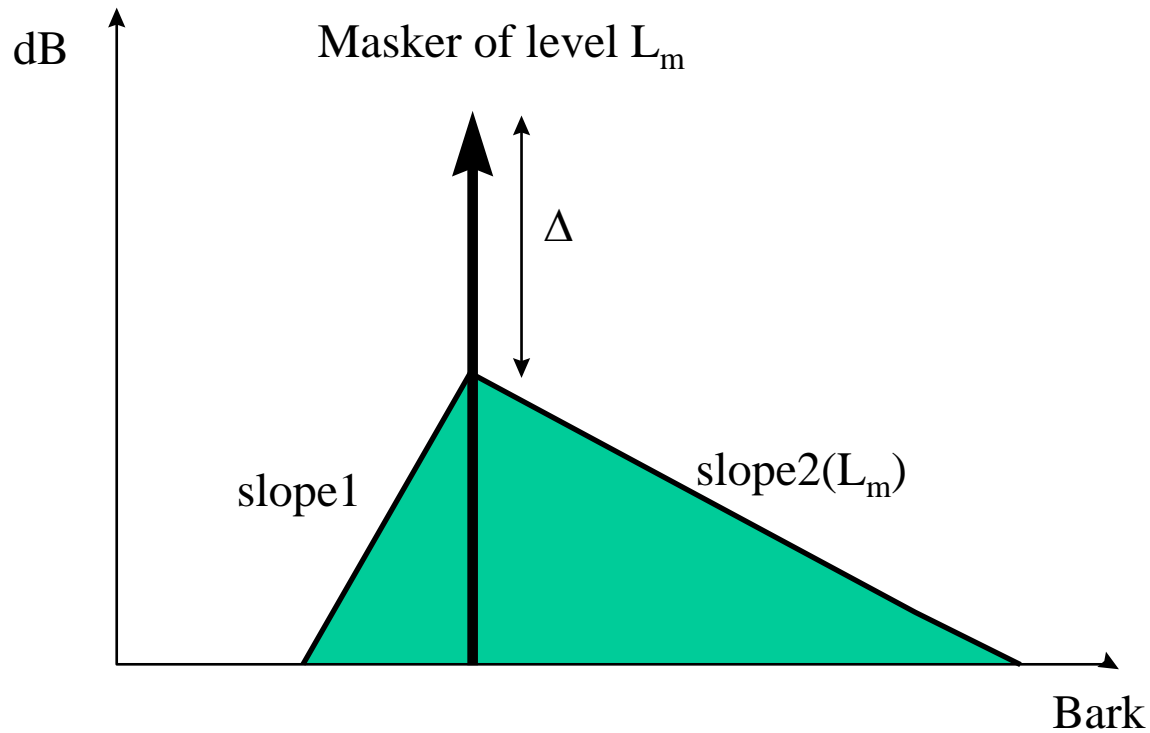
# Critical Bandwidth and ERB

- **Zwicker and Fastl 90:**  $CB/Hz = 25 + 75 \left[ 1 + 1.4 (f_c / \text{kHz})^2 \right]^{0.69}$
- **Moore 96:**  $ERB/Hz = 24.7 (4.37 f_c / \text{kHz} + 1)$

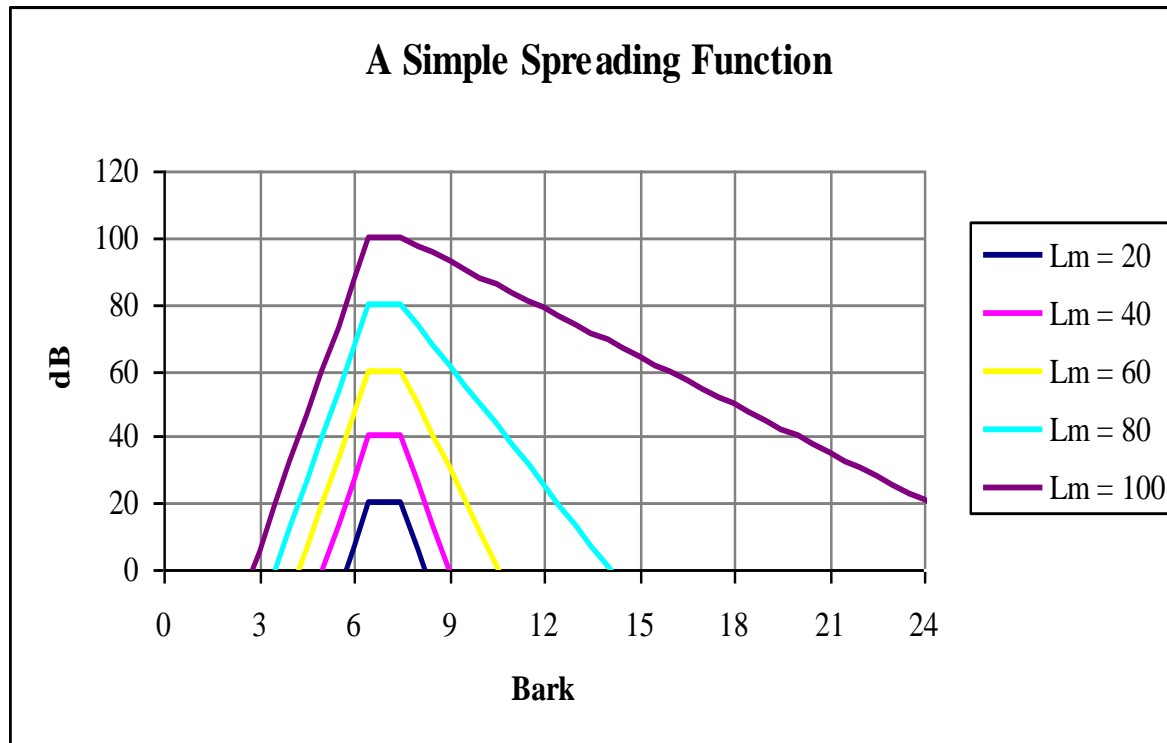




# Simple Masking Model

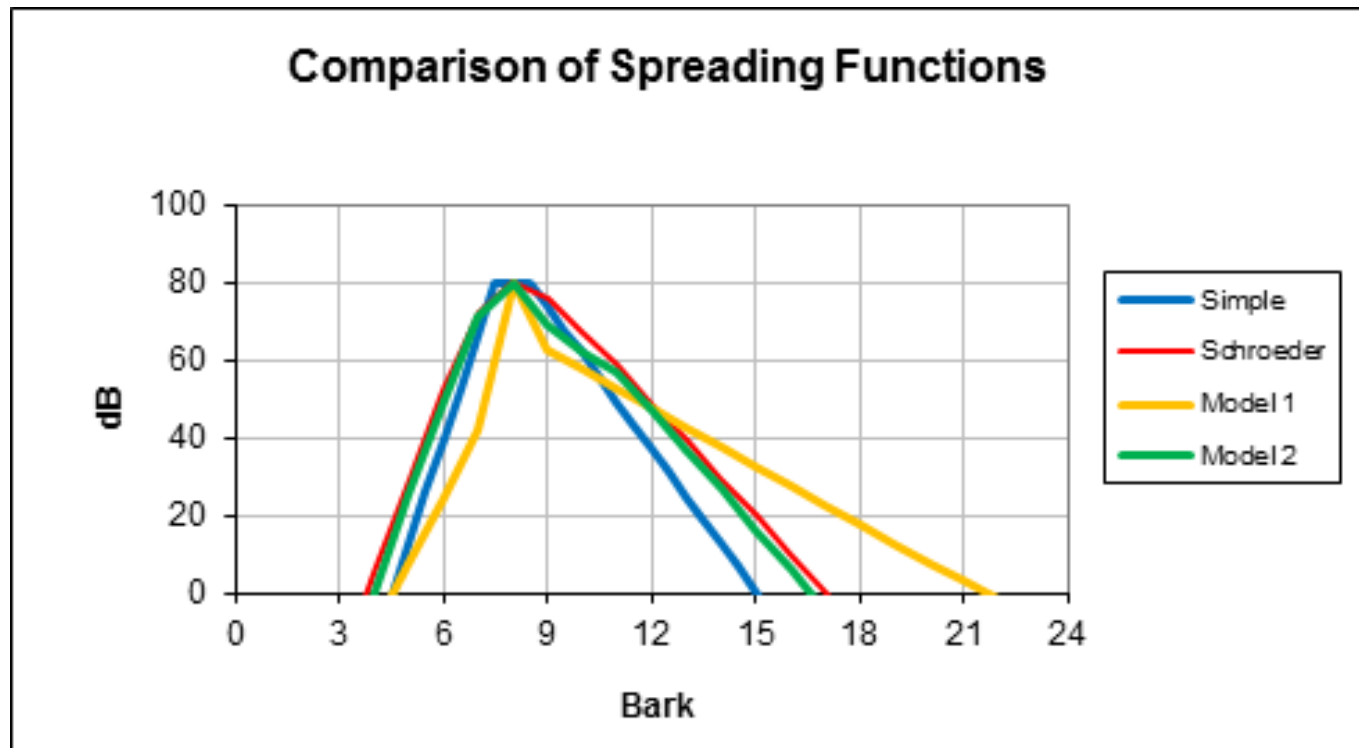


# A simple spreading function for computing masking curves



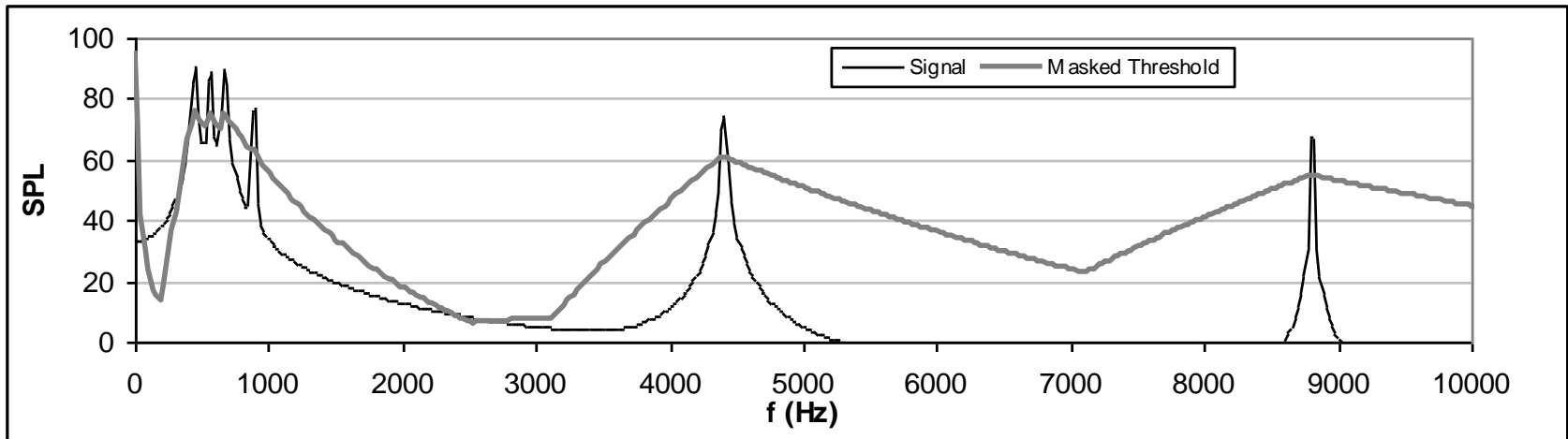
$$10\log_{10}(F(dz)) = \begin{cases} 0 & |dz| \leq 0.5 \\ (-27 + 0.37 \text{MAX}\{L_M - 40, 0\} \theta(dz)) (|dz| - 0.5) & \textit{elsewhere} \end{cases}$$

# Spreading Functions



# Masked Threshold

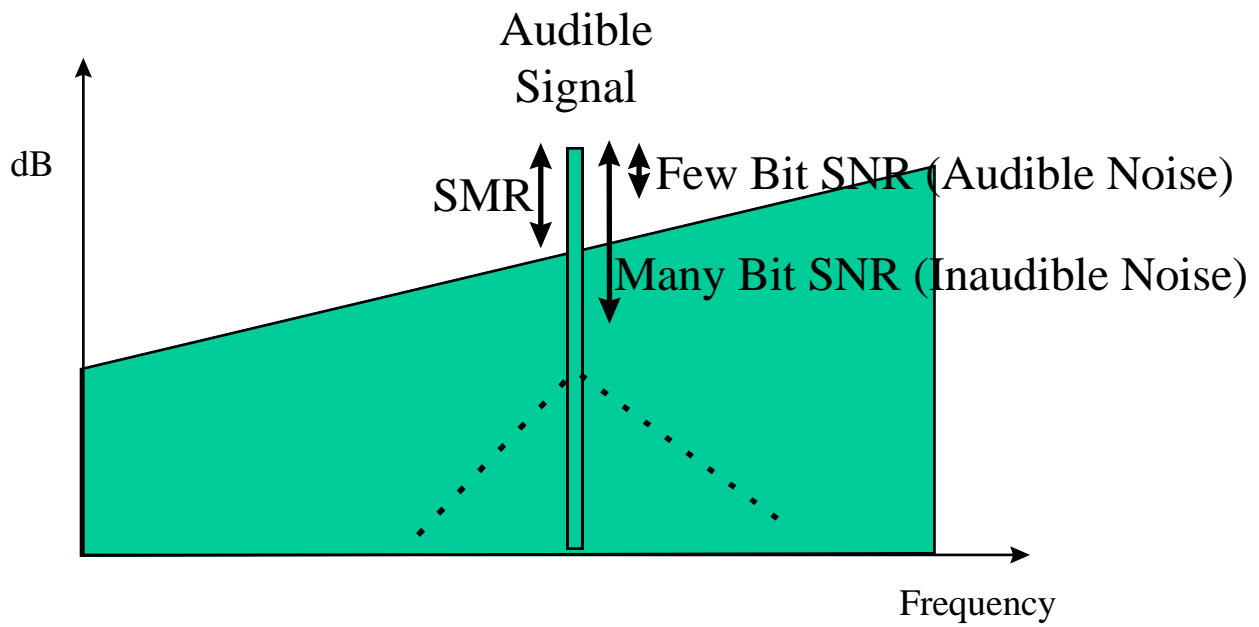
- **The hearing threshold can be combined with the effects of masking from the signal to create the Masked Threshold**
- **The Masked Threshold represents the level below which noise added to the signal should be inaudible**



# Quantization

- Quantization is the representation of a continuous signal amplitude (time or frequency sample) with a finite number of bits
- Quantization is a lossy process and is the main source of signal degradation in a digital audio coder
- Uniform quantization: each additional bit buys you about 6 dB more of signal to noise
- Scalar quantization mostly utilized in perceptual audio coding
  - Floating point quantization (linearized A-law), block floating point quantization, non-uniform quantization followed by noiseless coding
- If you know where the Masked Threshold is, you know how many bits are needed to get quantization noise
- Psychoacoustics-based bit allocation is the secret to Perceptual Audio Coders!

# Bit/Noise Allocation Using Masked Threshold



# Joint Stereo Coding

- **Goal: To exploit spatial redundancies and irrelevancies between audio channel**
- **M/S Stereo: Coding of Mid/Side (sum/difference) signal rather than L and R signals**
  - works best for near-monophonic signals, avoids “stereo unmasking”  
( Binaural Masking Level Differences, BMLD)
- **Intensity Stereo: Human auditory system evaluates signal envelope rather than its waveform at high frequencies (> 4 kHz)**
  - send 1 common set of spectral values for L & R, plus separate sets of scale factors
  - significant bitrate saving, affects stereo image !

**What Are we Listening For?**



# What Are we Listening For?

(from AES CD-ROM)

Original

Coded

- **Castanet (pre-echo)**
- **Quartet (aliasing)**
- **German male speaker (speech reverberation)**
- **Applause (stereo imaging)**



# **Audio Coders Design Choices**

# Brief History of MPEG Standards

- **The Moving Picture Expert Group (MPEG) was established in '88 in the framework of the joint ISO (the International Organization for Standardization) and IEC (International Electrotechnical Commission) Technical Committee, JTC 1, on information technology with the mandate to develop standards for coded representation of moving pictures, associated audio, and their combination**
- **Later it became working group 11 (WG 11) of JTC1/SC29**
- **Standardization phases:**
  - MPEG-1, coding up to 1.5 Mb/s (ISO/IEC 11172)
  - MPEG-2, coding of moving pictures and audio (ISO/IEC 13818)
  - **MPEG-3, coding up to 40 Mb/s**
  - MPEG-4, coding of audio-visual objects (ISO/IEC 14496)
  - MPEG-7, multimedia content description interface (ISO/IEC 15938)
  - MPEG-21, multimedia framework (ISO/IEC 21000)

# MPEG Standards (cont.)

- **23000 (MPEG-A) Multimedia Application Formats**
  - Part 1 Purpose for Multimedia Application formats, Part 2 Music Player Application Format, Part 3 Photo Player Application Format
- **23001 (MPEG-B) MPEG Systems Technologies**
  - Part 1 Binary MPEG format for XML
- **23002 (MPEG-C) MPEG Video Technologies**
  - Part 1 Accuracy specification for implementation of integer-output IDCT
  - Part 2 Fixed point implementation of DCT/IDCT
  - Part 3 Auxiliary Video Data Representation
- **23003 (MPEG-D) MPEG Audio Technologies**
  - Part 1 MPEG Surround
  - Part 2 SAOC
  - Part 3 USAC
  - Part 4 DRC
- **23008 (MPEG-H) MPEG Multimedia Middleware**
  - Part 1 Media Transport
  - Part 2 HEVC
  - Part 3 3D Audio

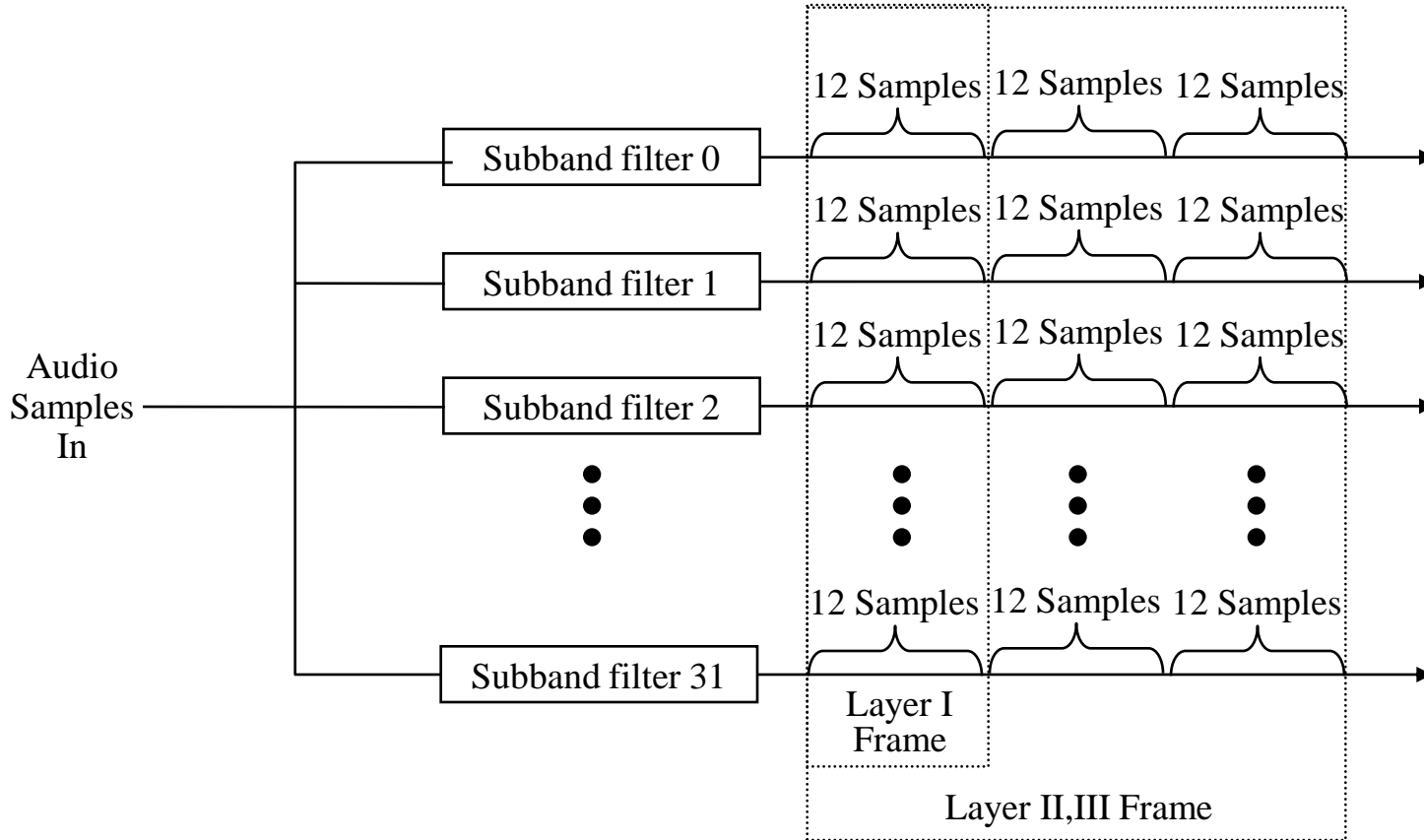
# MPEG-1 Audio

- **Coding of 16-bit mono and stereo signals**
- **Sampling frequencies: 48, 44.1, 32 kHz**
- **Data rates from 32 kbit/s to 448 kbit/s**
- **Three "Layers":**
  - Layer I: lowest complexity
  - Layer II: increased complexity and quality
  - Layer III: highest complexity and quality at low bit-rates
- **Target bitrates 384 kbits/s, 256 kbit/s,  $\leq 192$  kbit/s for Layers I, II, III respectively**

# MPEG-1 Building Blocks

- **Basic building blocks include:**
  - Time to frequency mapping filter bank: it subdivides the input signal onto 32 sub-bands (Layers I and II) or 576/192 sub-bands (Layer III)
  - Psychoacoustic model: the input signal is analyzed and the ratio of the signal energy to masking energy is computed for each sub-band
  - Bit or noise allocation: the signal to mask ratios determine the proportion of the total bits per block available to be allocated to each sub-band
  - Bit stream formatting: the representation of the quantized sub-band samples is formatted with side information and ancillary data
- **The decoder interprets the bitstream, restores the quantized sub-band values, and reconstructs the audio signal from its frequency representation**

# MPEG Layers Frames

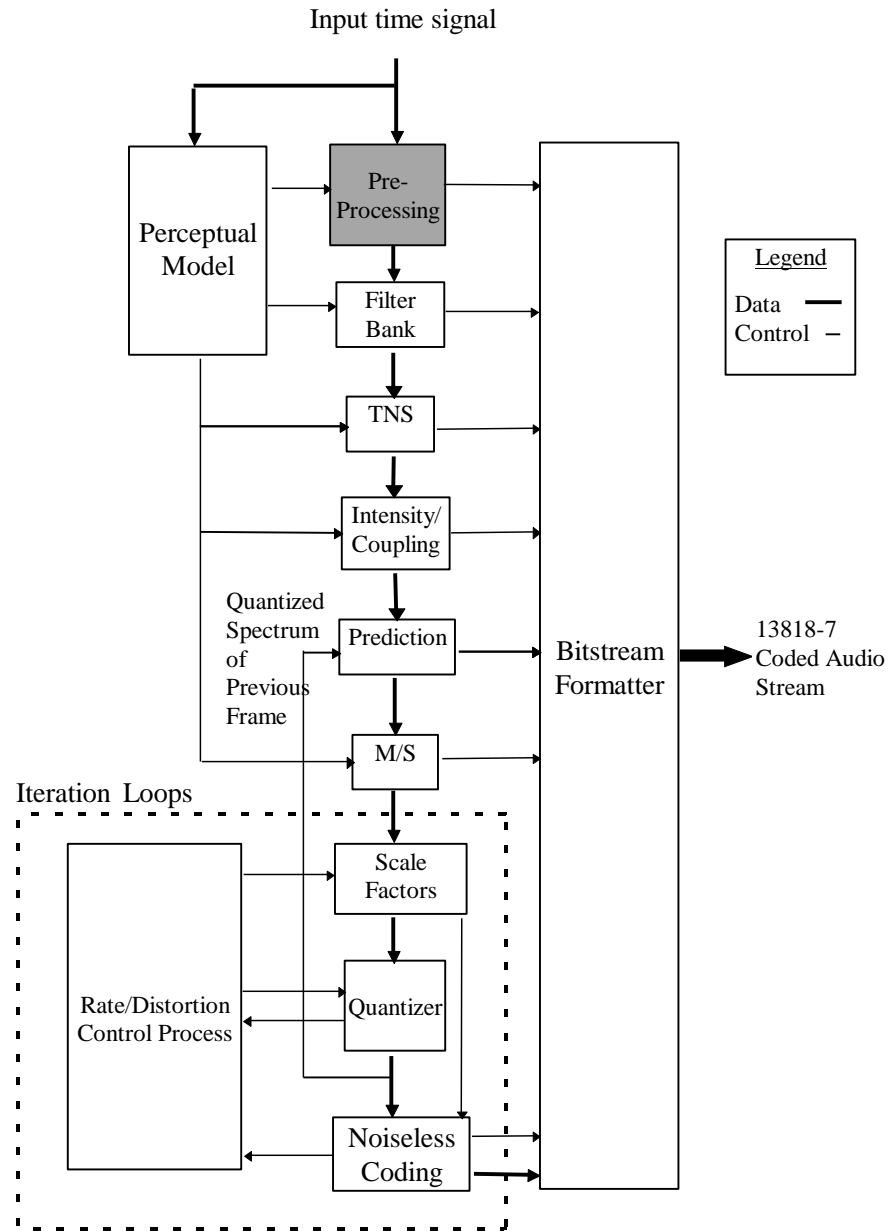


# MPEG-2 Audio

- **Same basic technology as in MPEG-1; in addition:**
  - Extension to multichannel signals (BC)
    - Up to 5.1 main audio channels plus up to 7 multilingual channels
    - Increased input precision (24 bits)
    - Data rates up to 1.1 Mb/s
  - Low Sampling Frequencies (LSF): 24, 22.05, 16 kHz
    - Increased frequency resolution/increased coding efficiency
    - Reduced bandwidth (max 12 kHz)
    - Data rates down to 8-128 kb/s per channel
- **MP3 or “MPEG 2.5” based on MPEG-2 LSF but also supports 12, 11.025, 8 kHz sampling frequencies**
- **MPEG-2 Advanced Audio Coding (AAC, formerly NBC)**



# AAC Encoder Configuration



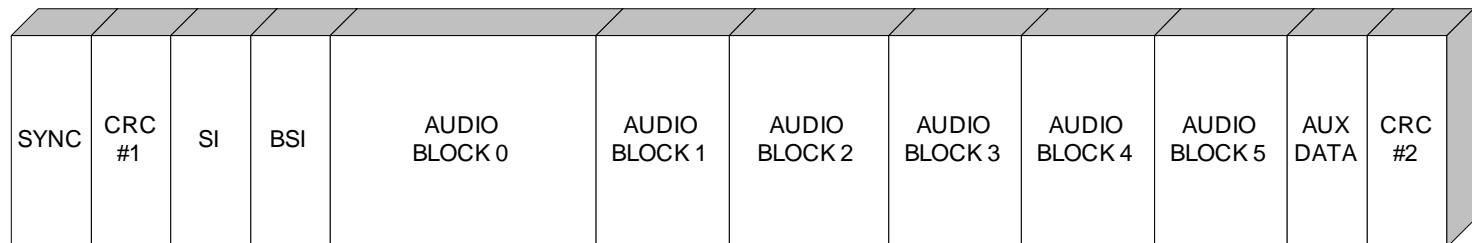
# Enhanced Coding Efficiency

- **Higher frequency resolution (MPEG-1/2 Layers I, II and III use 32 / 576 frequency lines vs. AAC's 1024)**
- **KBD window and improved block switching**
  - Window length 2048 / 256 (Layer III 1152 / 384); impulse response: 5.3 ms at 48 kHz (Layer III 18.6 ms)
- **Temporal noise shaping (TNS)**
  - Prediction in frequency domain to achieve noise shaping control in time domain
- **Prediction**
  - Backward adaptive line by line
- **Improved joint stereo coding**
  - Generalized intensity coding
  - Band-wise Mid/Side coding
- **Additional compression for the side information**
- **More flexible Huffman coding**
- **Sampling rates range from 8 kHz up to 96 kHz, up to 48 channels, data rates up to 576 kb/s/ch**

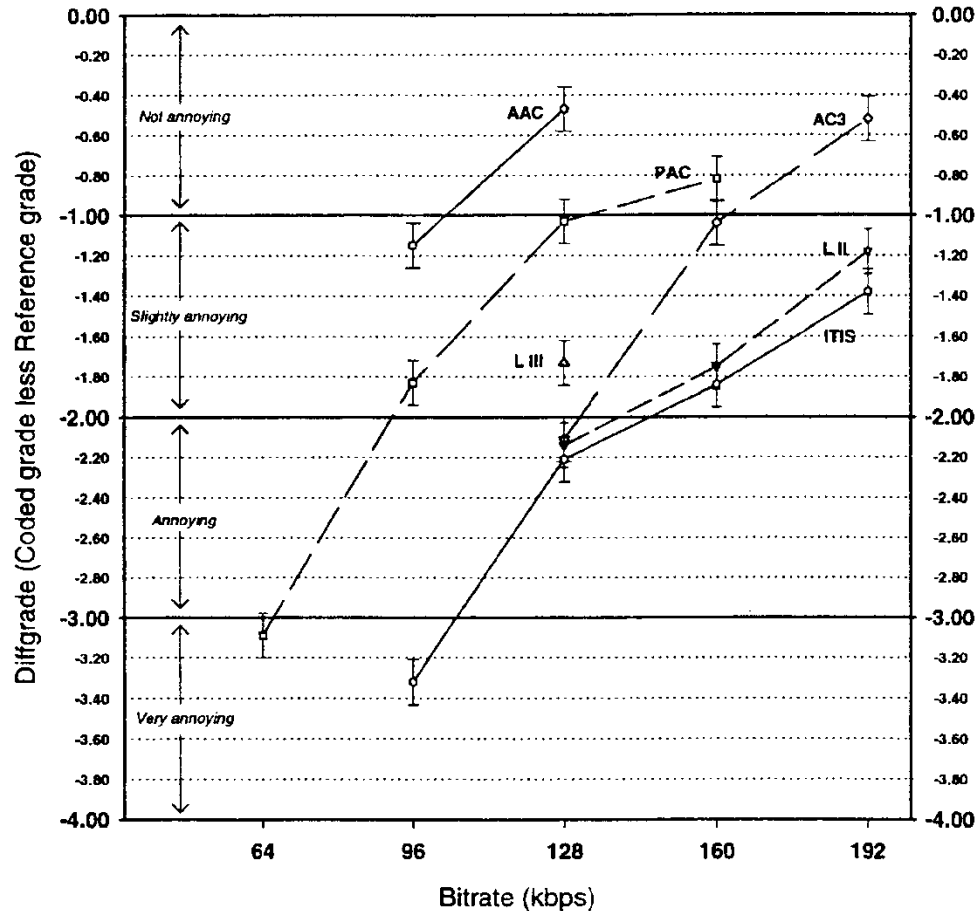
# Dolby AC-3 Audio Configurations

- **AC-3 design conceived to satisfy multichannel applications (sound for movies)**
- **Sampling rates:**  
48 kHz, 44.1 kHz, 32 kHz
- **Data rates ranging from 32 kb/s to 640 kb/s (6:1 to 24:1 compression ratios)**
- **Channel configurations:**  
3/2, 2/2, 3/1, 2/1, 3/0, 2/0, 1/0, 1+1
- **256/128 Frequency Lines MDCT, KBD window**
- **Backwards/Forwards Adaptive Psychoacoustics Model /Bit Allocation**
- **AC-3 Bitstream**

← 1536 PCM samples →



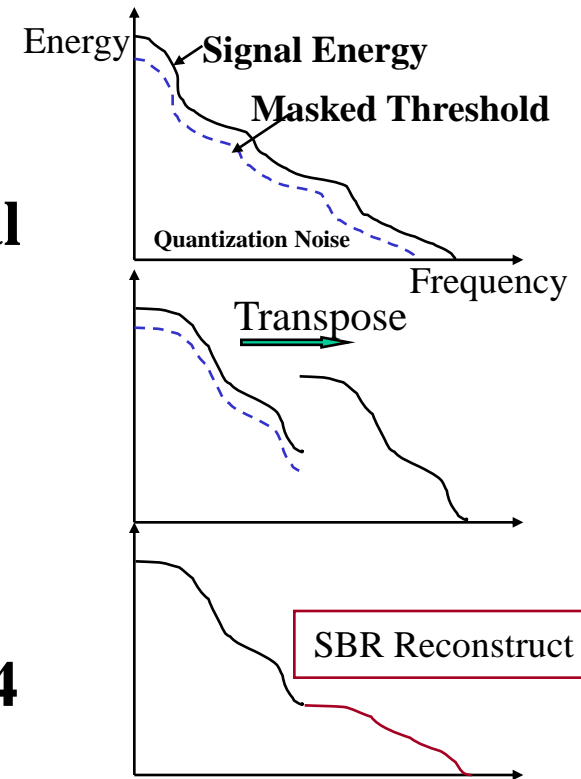
# Comparison of AC-3, MPEG-2 AAC, MPEG LII and LIII



From [Soulodre et. al. 98]

# Spectral Band Replication (SBR)

- **Only the low part of the signal spectrum is waveform coded**
- **The high frequency components of the signal are reconstructed from the low frequency components of the signal through a small amount of side information**
- **Compression efficiency can be significantly improved by using SBR (mp3PRO, MPEG-4 HE AAC, HE AAC v2)**
- **Similar principles applied in E-AC-3**



# E-AC-3

- **Similar to AC-3 with enhancements to increase coding efficiency including:**
- **Data rates ranging from 32 kb/s to 6.144 Mb/s**
- **1-8 Programs carried via time multiplexing**
- **Hybrid MDCT/DCT with 1536 frequency lines**
- **SBR**
- **Transient pre-noise signal substitution**

# Parametric Stereo/Multichannel Coding

- **Exploit correlations between stereo/multichannel signals**
- **Parametric stereo coding**
  - The stereo signal is coded as a monaural signal plus a small amount of stereo parameters
  - Inter-channel Intensity Difference (IID), describing the intensity difference between the channels
  - Inter-channel Cross-Correlation (ICC), describing the cross correlation or coherence between the channels
  - Inter-channel Phase Difference (IPD), describing the phase difference between the channels
  - The Inter-channel Time Difference (ITD) can be considered as an alternative to IPD.
- **Similar matrix basis changes can be extended to multichannel coding, see MPEG Surround**

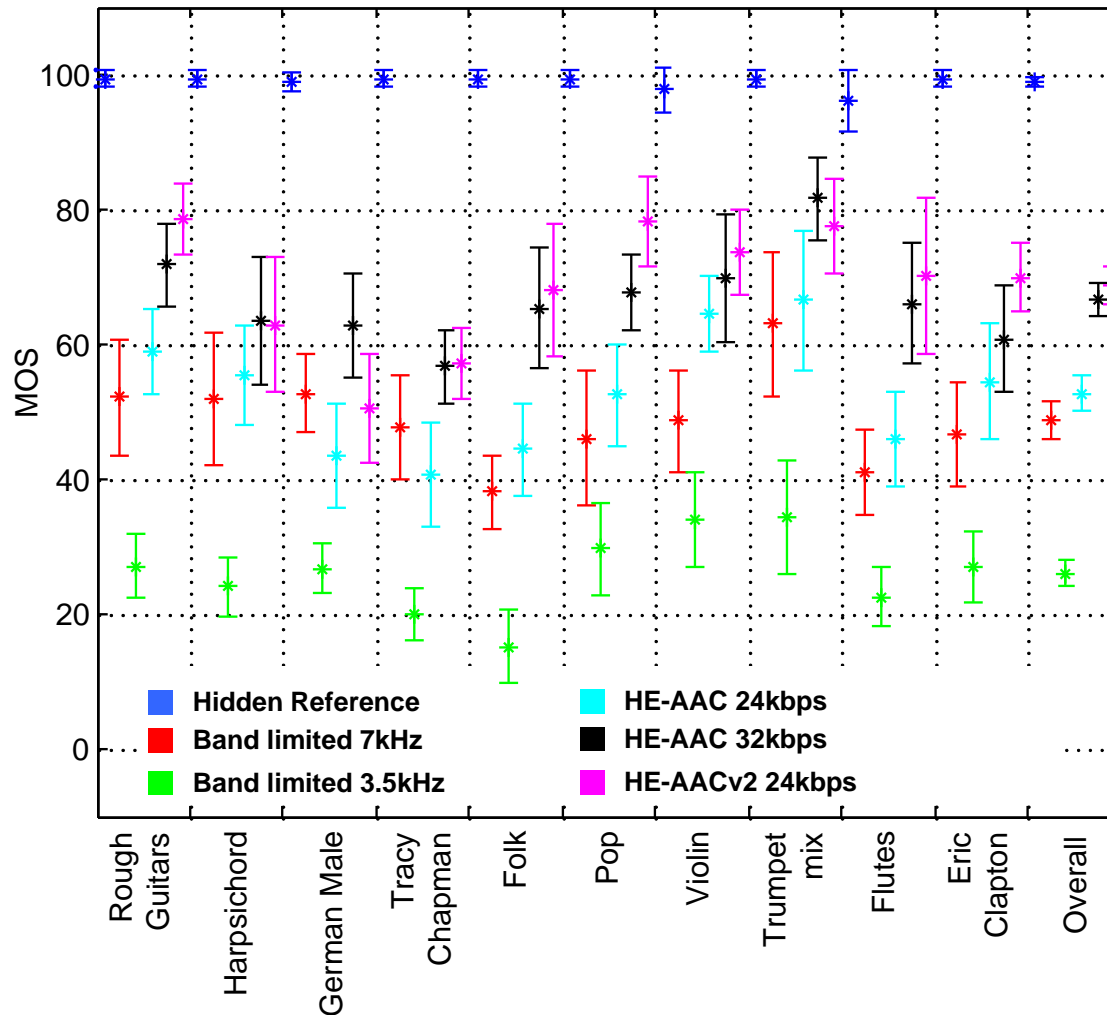
# **Sound Example: HE AAC v2**

**(courtesy of Coding Technologies)**

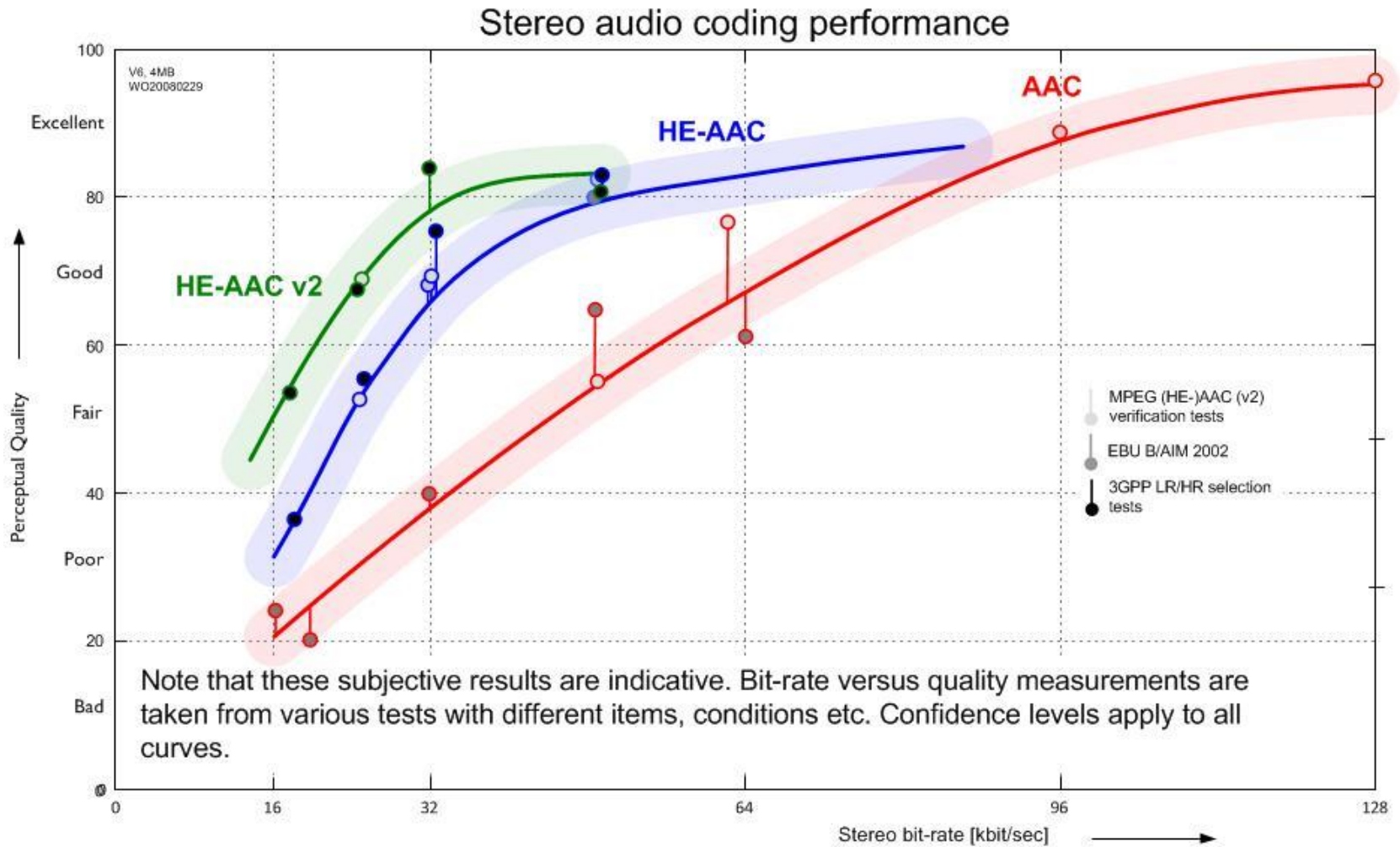


# Verification Test of MPEG-4 HE AAC v2

ISO/IEC JTC 1/SC 29/WG 11N7137, 2005



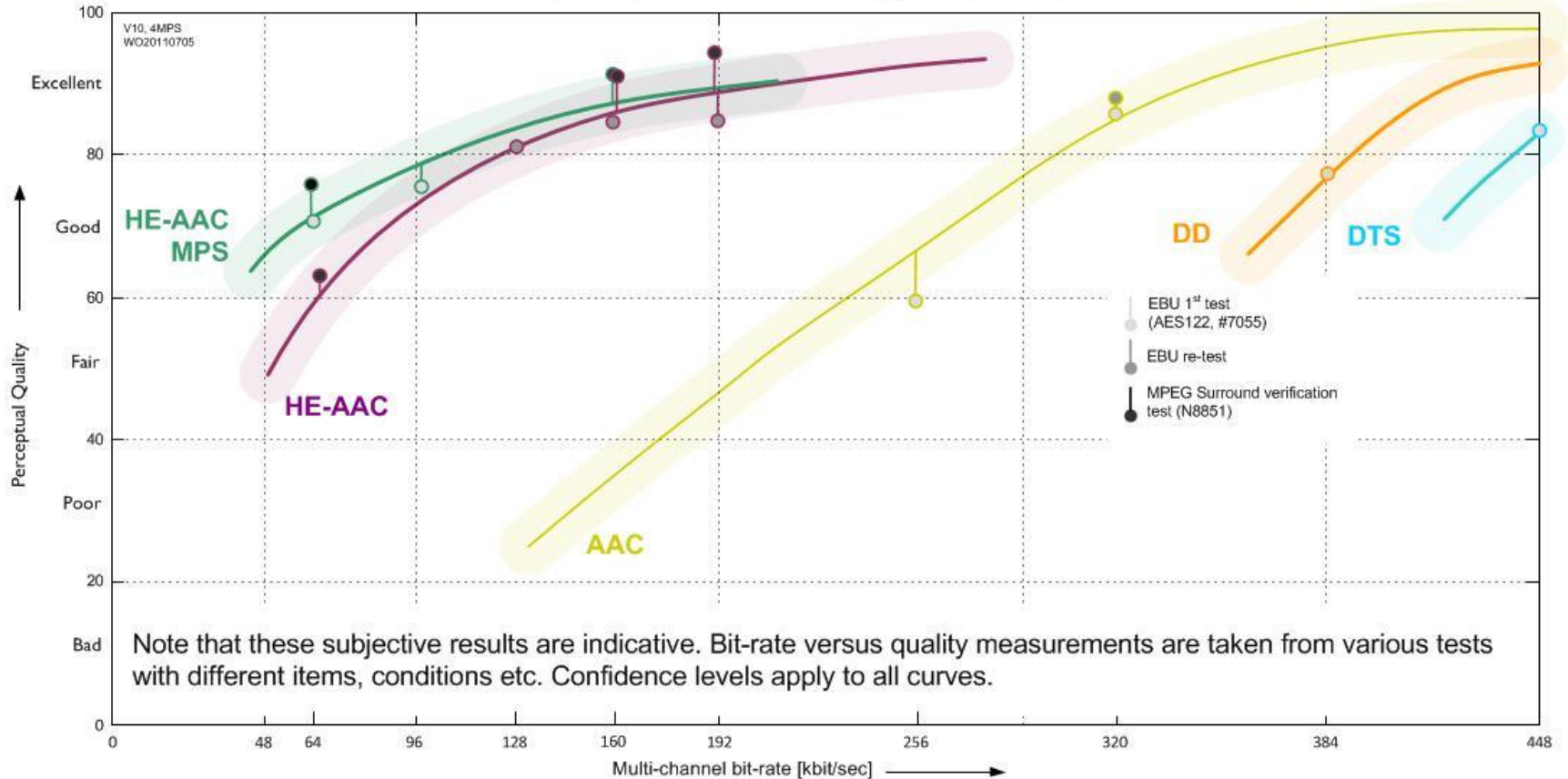
# AAC Family Performance



Thanks to Werner Oomen, Philips

# Multichannel Audio Coding Performance

## 5.1 channel audio coding performance



Thanks to Werner Oomen, Philips

**What will the future bring?**

# Evolution of technology

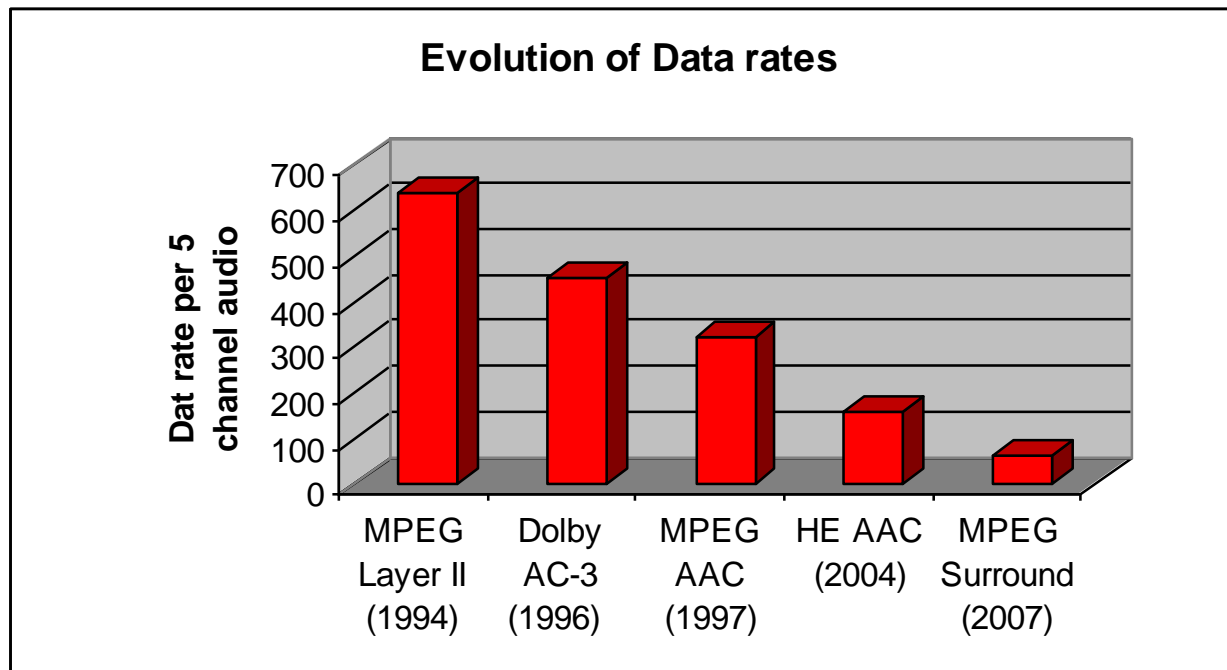
- **1980s**      **ADPCM**
- **1985s**      **PQMF/Transform Coding**
- **1990s**      **MDCT, High Quality Perceptual Coding**
- **1995s**      **Parametric Coding, Scalability, Source Separation**
- **2000s**      **Spectral Band Replication (SBR)**
- **2004**        **Parametric Stereo (PS)**
- **2007**        **Parametric Surround**
- **2012**        **Unified Speech and Audio Coding**
- **2015**        **3D Audio**
- **Future**       **Parametric Techniques in Conjunction with Scene Analysis (Semantic) Techniques ?**

# Evolution of Data Rates for Good Sound Quality for Stereo Signals

- **1992**            **256 kb/s**                            **MPEG Layer II**
- **1993**            **192 kb/s**                            **MPEG Layer III**
- **1994**            **128-192 kb/s**                        **MPEG MP3**
- **1995**            **384-448 kb/s per 5.1 signal**        **AC-3**
- **1997**            **96-128 kb/s**                        **MPEG-2 AAC**
- **2000**            **64-96 kb/s**                        **MPEG-4 AAC**
- **2001**            **48-64 kb/s**                        **AAC+ (HE AAC)**
- **2004**            **24-48 kb/s**                        **AAC+ PS**
- **2007**            **64 kb/s per 5.1 signal**            **MPEG Surround**
- **2012**            **8-32 kb/s**                        **USAC**

# Can We Continue to Achieve Better Compression?

- “Very High” Quality: AAC at 320 kb/s
- “Good Quality”: Increased performance over the last few years
  - See graph below



# To Learn More:

- **M. Bosi and R. E. Goldberg, “ Introduction to Digital Audio Coding and Standards ”, Kluwer Academic Publishers 2003**
- **“Collected Papers on Digital Audio Bit-Rate Reduction” Neil Gilchrist and Christer Grewin, Editors, Audio Engineering Society 1996**
- **E. Zwicker and H. Fastl, “Psychoacoustics”, Springer-Verlag 1990**
- **Proceedings of the AES 17th International Conference on “High-Quality Audio Coding”, K. Brandenburg and M. Bosi Co-chairs, Florence September 1999**
- **AES CD-ROM On Perceptual Audio Coders 2001: "Perceptual Audio Coders: What to Listen For", AES 2001**
- **<http://mpeg.chiariglione.org/standards>**