Network & Media Laboratory

Fundamentals of Multimedia

Authored by Ze-Nian Li Mark S. Drew

Lecturer： Lu Dongming

Lecture 12
Audio Compression Techniques

Content

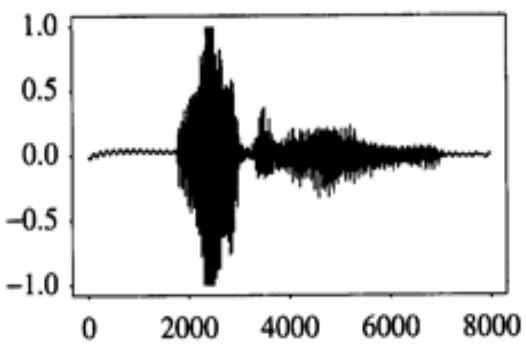
- Basic Audio Compression Techniques
  - ADPCM In Speech Coding
  - G.726 ADPCM
  - VOCODERS
- MPEG Audio Compression
  - PSYCHOACOUSTICS
  - MPEG Audio

1. Basic Audio Compression Techniques

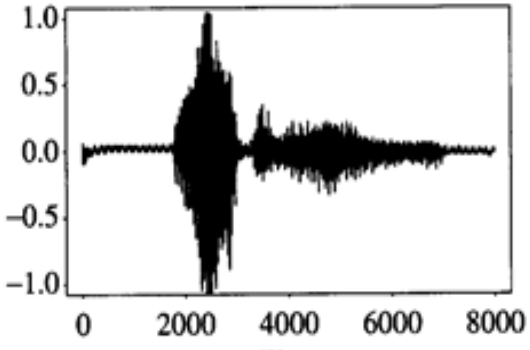
- ADPCM In Speech Coding
- G.726 ADPCM
- VOCODERS
  - Phase Insensitivity
  - Channel Vocoder
  - Formant Vocoder
  - LPC (Linear Predictive Coding)

1.1 ADPCM In Speech Coding

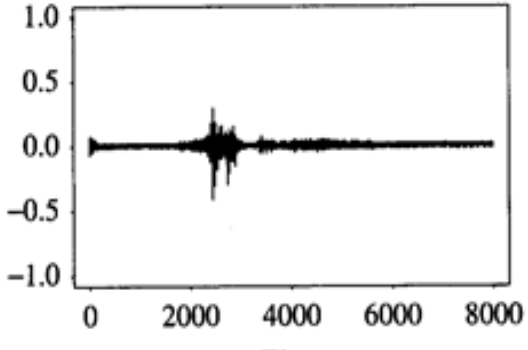
- **ADPCM (Adaptive Differential Pulse Code Modulation)**
  - **The heart of the ITU's speech compression standard G.721, G.723, G.726, G.727**
  - **Bitrate and details of the algorithms of these standards are different**



Audio signal stored as linear PCM 8KHz 16bits per sample



G.721 compressed audio 4bits per sample

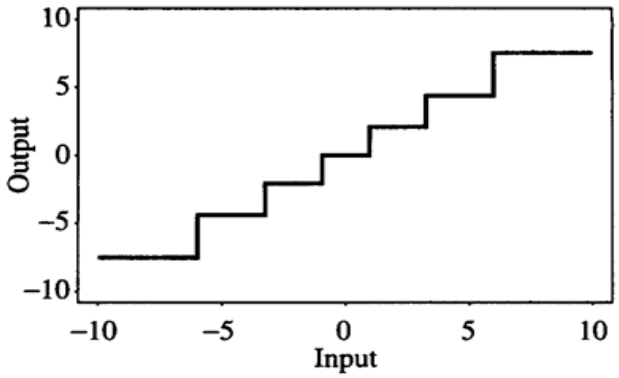


Difference between compressed audio and original audio

1.2 G.726 ADPCM

- Supersedes G.721 and G.723
  - Encode 13bit,14bit PCM samples or 8bit u-law, A-law encode data into 2-,3-,4-,or 5bit codewords
- G.726 defines a factor α
  - The input value is defined as a ratio of a difference with the factor α.

$$e_n = s_n - \hat{s}_n$$
$$f_n = e_n / \alpha$$



G.726 Quantizer

1.2 G.726 ADPCM

- By changing the value of α, the Quantizer can adapt to change in the range of the difference signal
- Principle of **Backward-Adaptive Quantizer** :
  - If too many quantized values far from zero， enlarge the quantize step
  - On the contrary, then reduce the quantize step
  - The algorithm used is called Jayant.

- **Jayant quantizer**
  - Assigns multiplier values M_k to level K, for signal f_n,
  - The quantizer step size Δ is changed according to the quantized value of f_{n-1}

$$\Delta = M_k \Delta$$

1.2 G.726 ADPCM

- **Jayant quantizer**
  - Uses fixed quantizer steps based on the logarithm of the input difference signal,
  - The divisor α is written in terms of logarithm: $\beta=\log_2\alpha$
- **How α is change depends on**
  - If the audio signal is actually speech or voice band data
  - Sample to sample differences of the former fluctuate great, but the latter more smoothing
- **Distinguish between the two situations**
  - α is divided as locked part and unlocked part α_L, α_U
  - The corresponding log formats are β_L, β_U

1.2 G.726 ADPCM

- The logarithm value is written as a sum of two pieces,

$$\beta = A\beta_U + (1-A)\beta_L$$

- The "unlocked" part adapts via the equation

$$\alpha_U \leftarrow M_k \alpha_U$$

$$\beta_U \leftarrow \log_2 M_k + \beta_U$$

- The "locked part" via

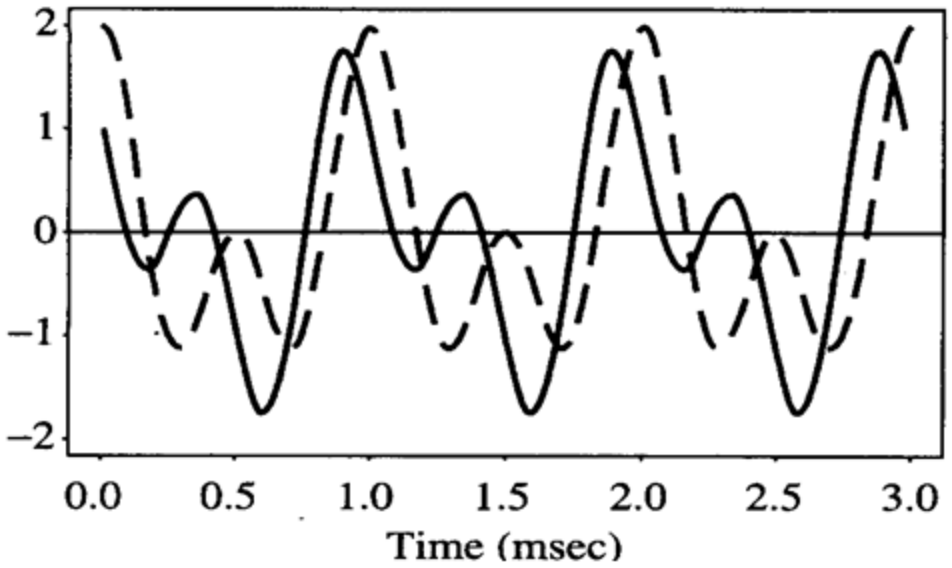
$$\beta_L \leftarrow (1-B)\beta_L + B\beta_U$$

$$B\,is\,a\,small\,number,\quad say 2^{-6}$$

- G.726 predictor is **complicated**:
  - Uses a linear combination of 6 quantized differences and two reconstructed signal values from the previous six signal values f_n

1.3 VOCODERS

- **Phase Insensitivity** (“phase” : A shift in the time argument)
  - Focus on energy spectrum, we are not sensitive to the exact waveform



The solid line shows the superposition of two cosines, with a phase shift. The dashed line shows the same with no phase shift. The wave is very different, yet the sound is the same, perceptually.

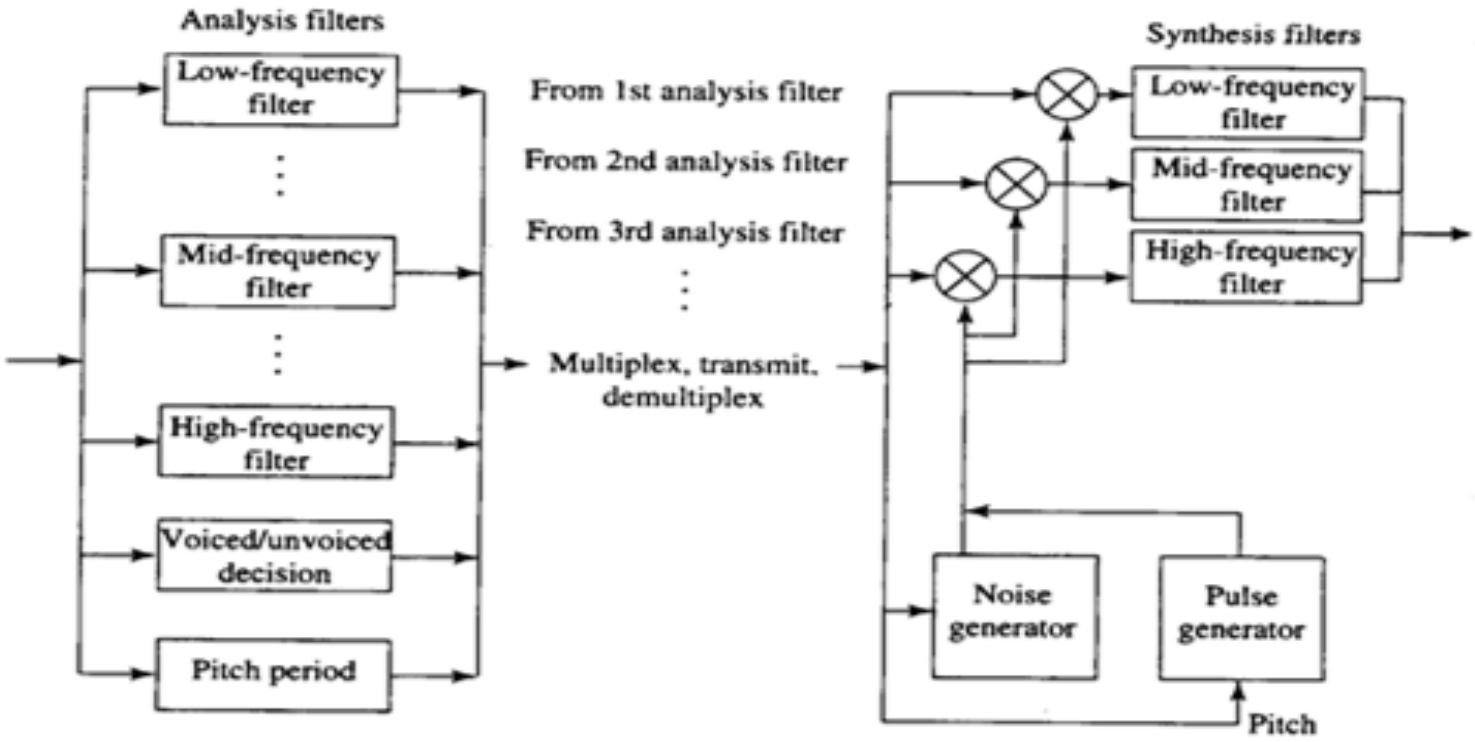
1.3 VOCODERS : Channel Vocoder

- Sub-band filtering
  - A bank of band-pass filters to the analog signal
  - Carrying out the frequency decomposition
  - Sub-band coding can achieve better compression
- First applies a filter bank to separate out the different frequency components, then the sub-band coding fulfilled
- Analyzes the signal to determine the pitch and excitation of the speech

1.3 VOCODERS : Channel Vocoder

- Speech excitation concerned with whether a sound is voiced or unvoiced
  - A sound is unvoiced if its signal looks like noise, otherwise it's voiced
  - Vowels is voiced, sounds like s, f are unvoiced, some consonants like b, d, g are voiced, because at the tail vowel lasting a few milliseconds
- Voiced sounds: a periodic pulse generator recreates voices sounds
- Unvoiced sounds: a pseudo-noise generator is applied
  - All values are scaled by the energy estimates given by the band-pass filter set

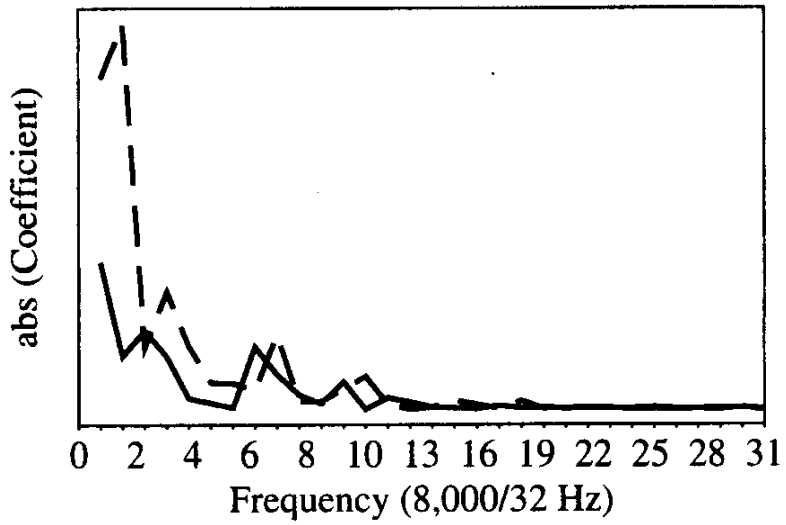
1.3 VOCODERS : Channel Vocoder



Channel vocoder

1.3 VOCODERS: Formant Vocoder

- All frequencies present in speech are not equally important, only certain frequencies contain most of the energy



- The solid line shows frequencies present in the first 40 msec of the speech.
- The dashed line shows similar frequencies are still present on second later.

1.3 VOCODERS : Linear Predictive Coding

- Extract salient features of speech directly
  - From the waveform, rather than transforming the signal to the frequency domain
- Uses a time-varying model
  - A set of parameters modeling the shape and excitation of the vocal tract， not actual signals or differences.
- Denote output as S(n), for input values f(n). The output depends on p previous output sample values.

$$S(n) = \sum_{i=1}^{p} a_i s(n-i) + Gf(n)$$

  - Here, G is known as the gain factor.

1.3 VOCODERS: Linear Predictive Coding

- **Input signal is analyzed in speech frames**
  - **The frame length is usually as 22.5ms, which corresponds to 180 samples for 8kHz sampled digital speech.**
- **LP coefficients can be solved by：**

$$\min E\left\{[s(n)-\sum_{j=1}^{p}a_j s(n-j)]^2\right\}$$

After getting the LP coefficients ,gain G can be calculated as：

$$G=E\left\{[s(n)-\sum_{j=1}^{p}a_j s(n-j)]^2\right\}$$

$$=\phi(0,0)-\sum_{j=1}^{p}a_j\phi(0,j)$$

1.3 VOCODERS: Linear Predictive Coding

- **Voiced and Unvoiced:**
  - **If V(P) is less than some given threshold, the current frame is classified as Unvoiced frame and will be reconstructed by white-noise sequences.**
  - **Otherwise it's judged as voiced frame and reconstructed with periodic wave.**

$$V(i) = \sum_{n=m}^{N-1+m} s(n)s(n-i) \Big/ \left[\sum_{n=m}^{N-1+m} s^2(n) \sum_{n=m}^{N-1+m} s^2(n-i) \right]^{1/2}$$

$i \in [P_{\min}, P_{\max}]$

$[P_{\min}, P_{\max}]$ *is often selected as* [12,140] *for 8kHz speech*

2. MPEG Audio Compression

- **PSYCHOACOUSTICS**
  - Equal-Loudness Relations
  - Frequency Masking
  - Temporal Masking
- **MPEG Audio**
  - MPEG Layers
  - MPEG Audio Strategy
  - MPEG Audio Compression Algorithm
  - MPEG-2 AAC (Advanced Audio Coding)
  - MPEG4 Audio

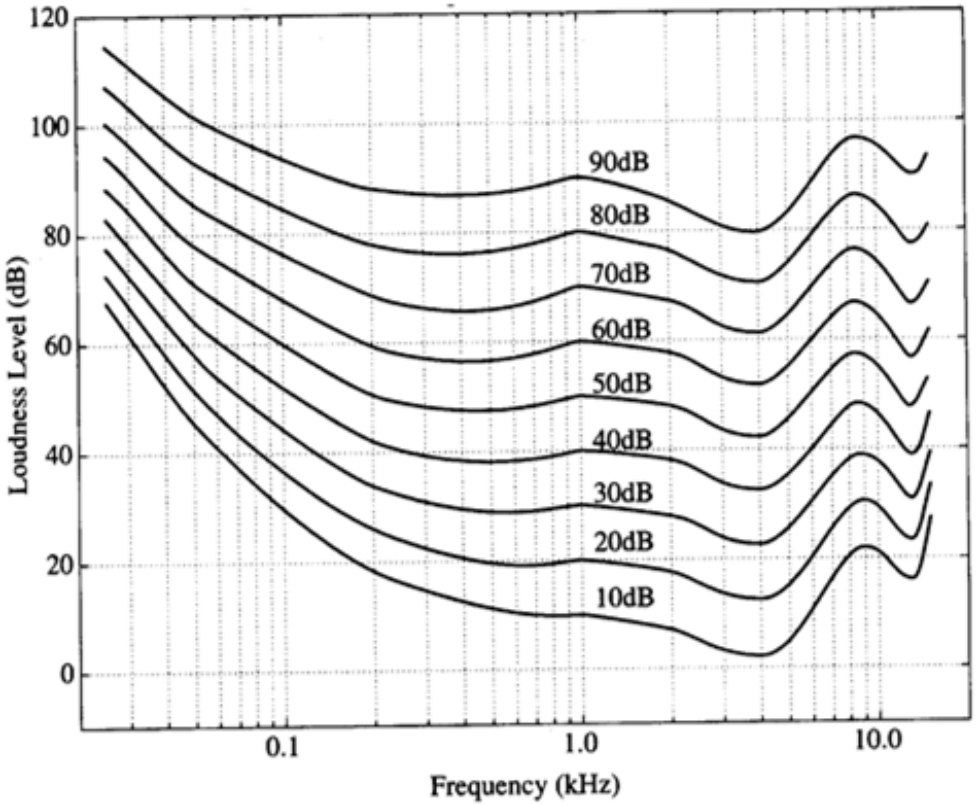
2.1 PSYCHOACOUSTICS

- Equal-Loudness Relations
- Frequency Masking
- Temporal Masking

2.1.1 Equal-Loudness Relations

- The ear has different sensitivities for different audio frequencies, most sensitive to 1kHz~5kHz
- Fletcher-Munson equal-loudness curves
  - Relationship between loudness for a given sound pressure level
- The curves are arranged that the perceived loudness level are same as the loudness level at 1kHz
- As the loudness increase, the curves flatten
  - Perceive most low frequencies better than high ones at high volume level
  - This is why people prefer loud music rather than quiet music.

2.1.1 Equal-Loudness Relations



Equal Loudness Response Curves for the Human Ear

2.1.2 Frequency Masking

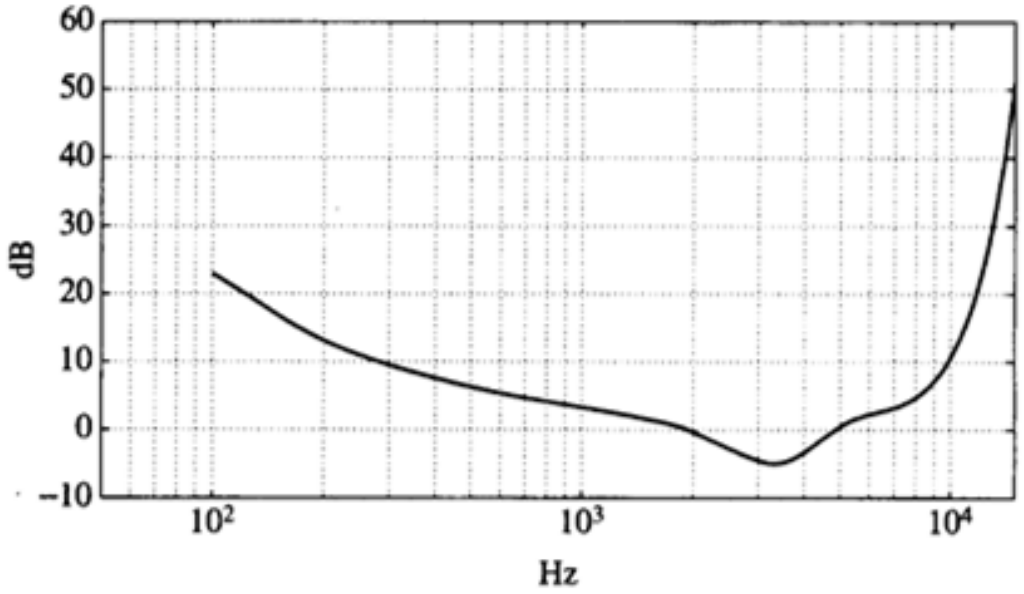
- A lower tone can effectively mask a high tone
- Higher tones can mask lower tones
  - But not effectively as they mask higher-frequency ones
- The greater the power in the masking tone, the wider its influence
- If two tones are widely separated in frequency, little masking occurs

2.1.2 Frequency Masking

- **Threshold of hearing**

  An approximate formula:

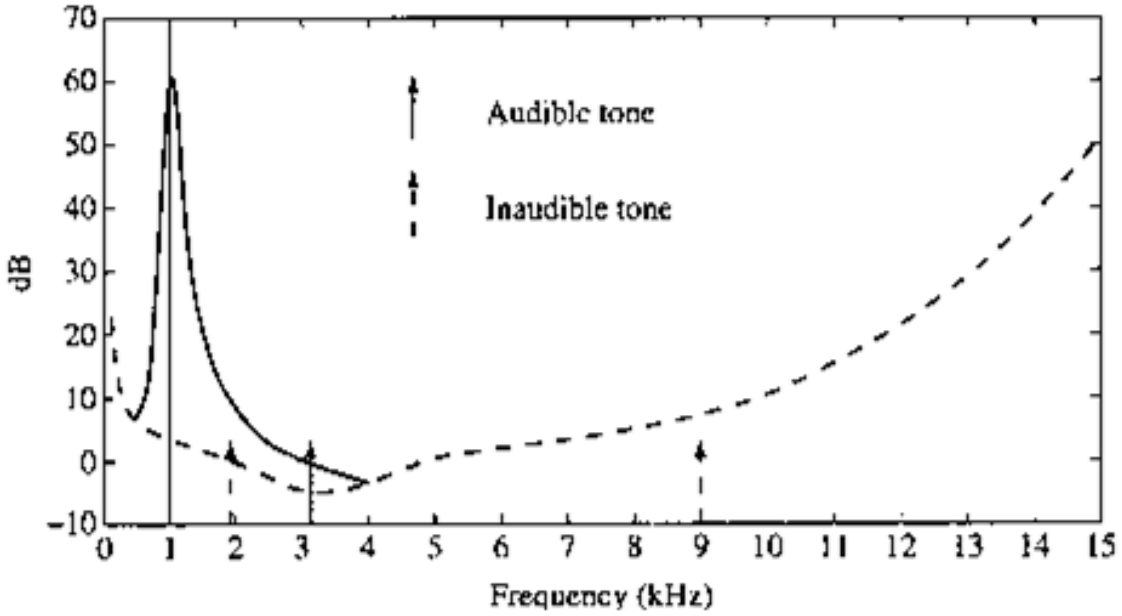
$$Threshold(f) = 3.64(f/1000)^{-0.8} - 6.5e^{-0.6(f/1000-3.3)^2} + 10^{-3}(f/1000)^4$$



Threshold of humman hearing, for pure tones.

2.1.2 Frequency Masking

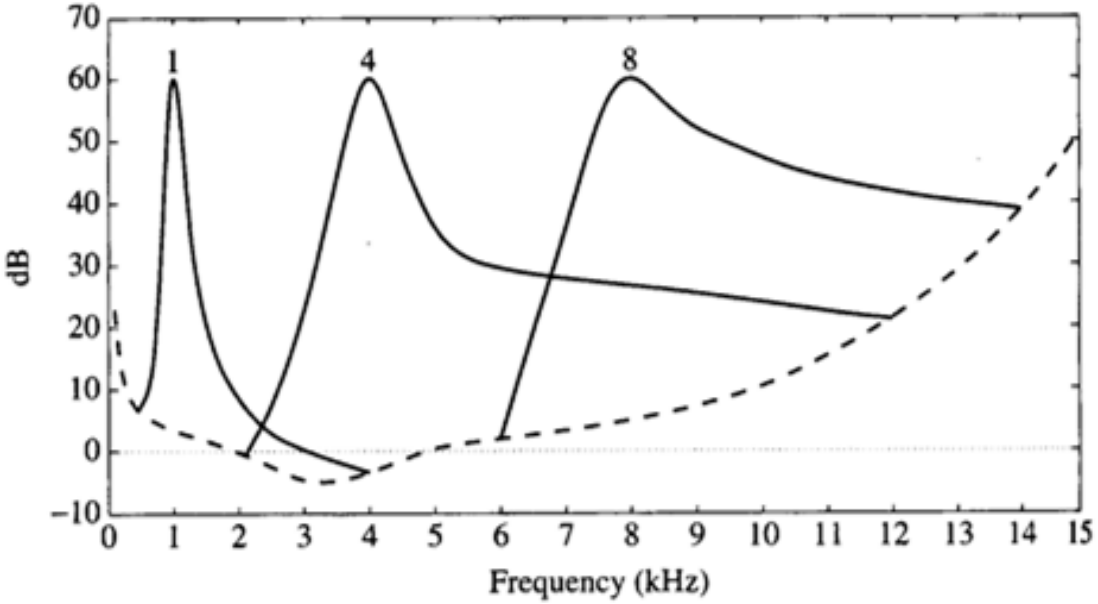
- **Frequency Masking Curves**
  - **Generate a 1kHz masking tone at a fixed sound level of 60dB, test the frequency masking effect:**



Effect on threshold of humman hearing for a 1kHz masking tone

2.1.2 Frequency Masking

- **Frequency Masking Curves**
  - **Different frequencies have different masking effect**



Effect of masking tones at three different frequencies

2.1.2 Frequency Masking : Critical Bands

- **Critical bandwidth**: the smallest frequency difference between two partials that each can still be heard separately
- The low frequency end: critical bandwidth is less than 100Hz
- High frequency: the width can be greater than 4kHz
- Experiments indicate
  - When mask frequency below 500Hz, critical bandwidth is approximately 100Hz
  - For frequencies above 500Hz, critical bandwidth approximately linear with frequency.

2.1.2 Frequency Masking : Critical Bands

- The audio frequency range for hearing be partitioned into 24 critical bands
- The ear operates like a set of band-pass filters
  - Allow a limited range of frequencies through and blocks all others
- A constant volume sound will seem louder if it spans two critical bands
- The ear is not very discrimination within a critical band

2.1.2 Frequency Masking: Bark Unit

- The range of frequencies affected by masking is broader for higher frequencies
- Define a new frequency unit so that the masking curves have about the same width.
  - The new unit defined is "Bark"
- The conversion between a frequency f and its corresponding critical-band number b is as follows:

$$Critical\ band\ number(Bark) = \begin{cases} f/100, & for\ f < 500 \\ 9 + 4\log_2^{(f/1000)}, & for\ f \geq 500 \end{cases}$$

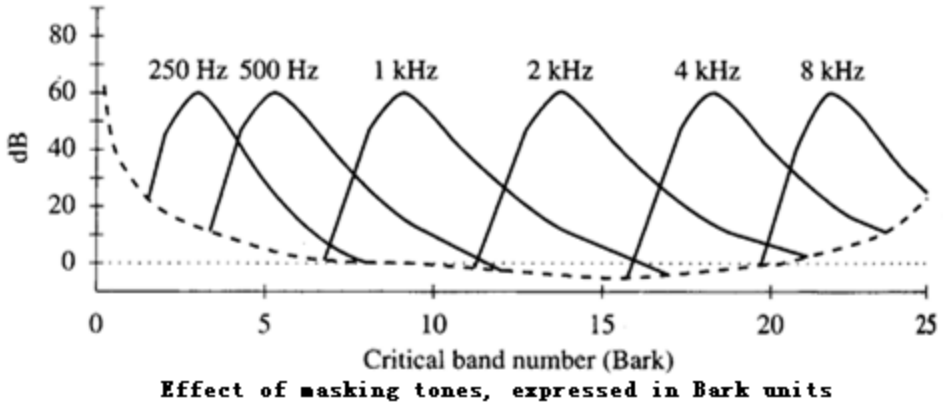
2.1.2 Frequency Masking: Bark Unit

- **The inverse equation from b to f(kHz) is:**

$$f = [(\exp(0.219 \times b) / 352) + 0.1] \times b - 0.032 \times \exp[-0.15 \times (b-5)^2]$$

- **Given a center frequency f, critical bandwidth (df) can be approximated by:**

$$df = 25 + 75 \times [1 + 1.4(f^2)]^{0.69}$$

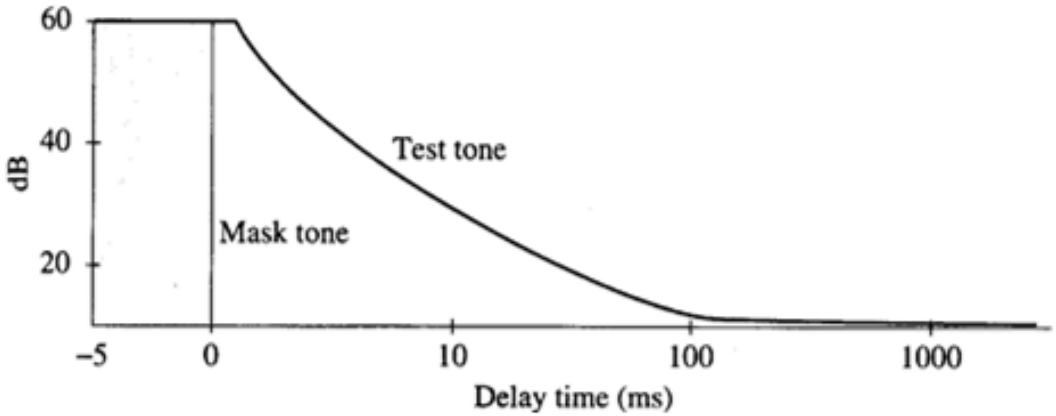


Effect of masking tones, expressed in Bark units

2.1.3 Temporal Masking

- **After dance it takes quite a while for our hearing to return to normal**

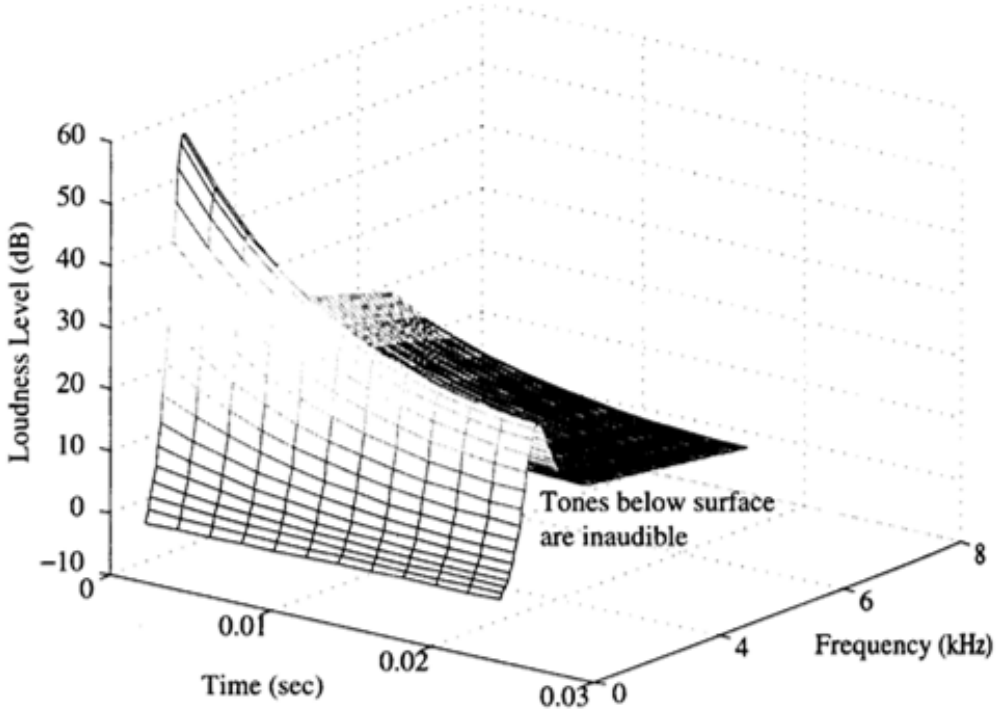
Play a masking tone at 1kHz with volume level 60dB, the smallest amount of time we can hear 1.1kHz,40dB tone again after the masking is turn off.



The louder the test tone, the shorter the amount of time required before the test tone is audible once the masking tone is removed.

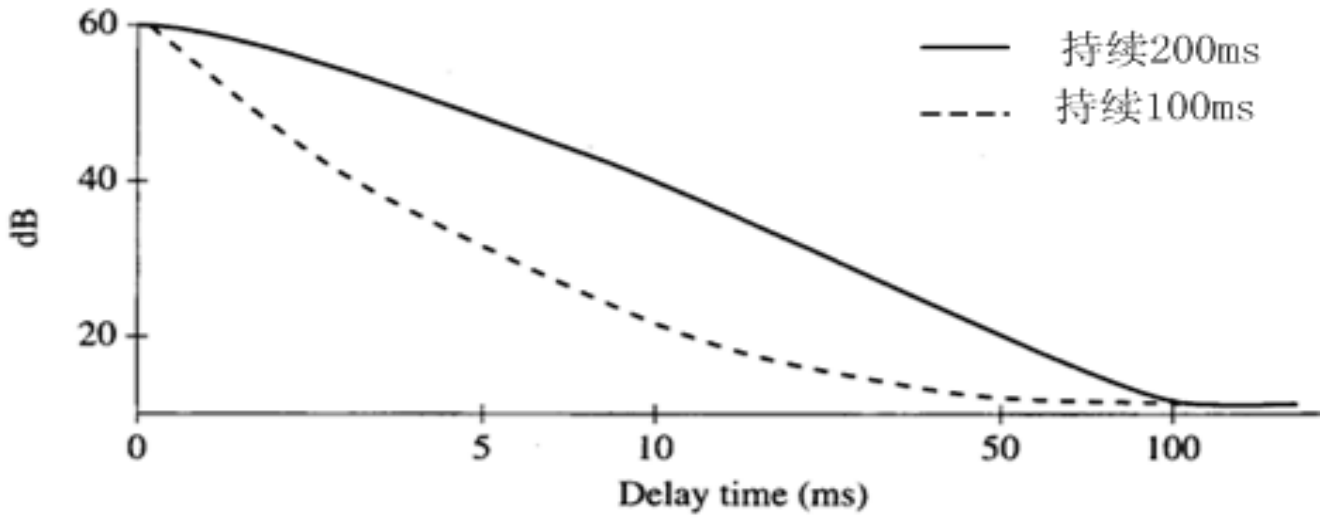
2.1.3 Temporal Masking

- **Effect of temporal masking depends on both time and closeness in frequency**



2.1.3 Temporal Masking

- **Effect of temporal masking also depends on the length of time the masking tone is applied.**



掩蔽效应与掩蔽声音(masking tone)的持续时间的关系

2.2 MPEG Audio

- MPEG Layers
- MPEG Audio Strategy
- MPEG Audio Compression Algorithm
- MPEG-2 AAC (Advanced Audio Coding)
- MPEG4 Audio

2.2.1 MPEG Layers: MP3

- 3 stands for 3 layers，MP stands for MPEG-1
  - Released in 1992, published as ISO/IEC 11172-3 in 1993
- All layers include the same file header information
  - Layer 1 quality can be very good
  - Layer2 has more complexity for use in digital audio broadcasting
  - Layer3 is most complex and was originally aimed at audio transmission over ISDN lines
- Most of the complexity is at the encoder rather than at the decoder side
  - This accounts for the popularity of MP3 players

2.2.2 MPEG Audio Strategy

- MPEG encoder employs a bank of filters
  - Layer1 and layer2 use a quadrature-mirror filter bank
  - layer3 adds a DCT
- Psychoacoustic model: frequency masking to estimate the noticeable level
  - Balance the masking behavior and the available number of bits
- More sophisticated model: account the actual width of critical bands
  - For simplicity, the model adopts a uniform width for all frequency analysis filters
  - Layer1 uses only frequency masking
  - Layer2 uses some temporal masking
  - Layer3 uses a more sophisticated subband analysis, with nonuniform subband width

2.2.3 MPEG Audio Compression: Basic Algorithm

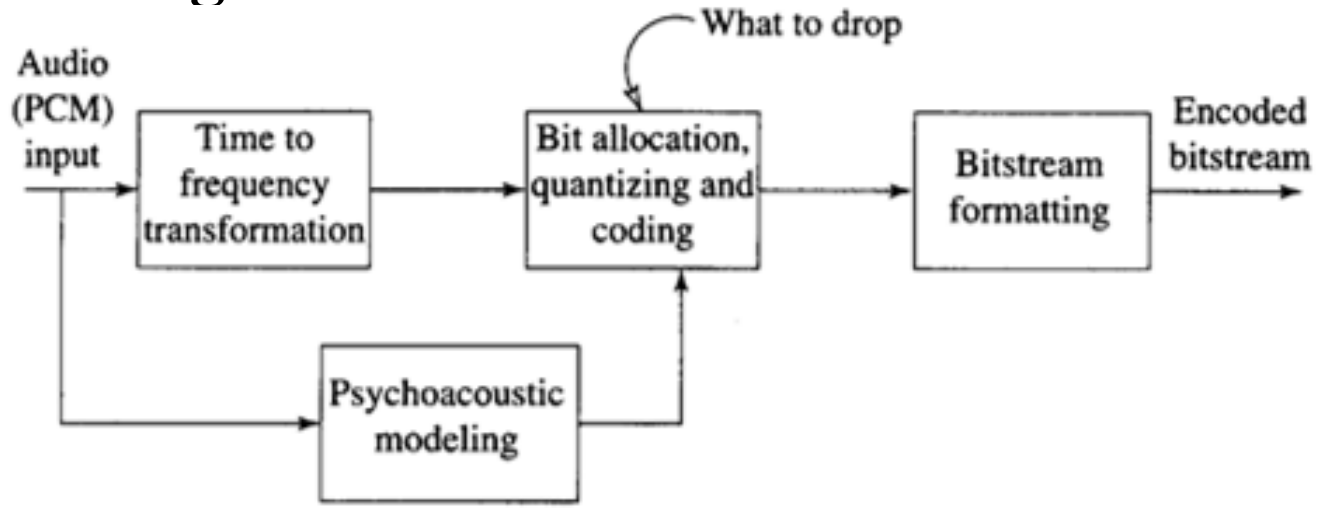
- Dividing the input into 32 frequency subbands
  - In the layer1 encoder, a set of 12 groups of 32 PCM values are called a segment, each subband examine 12 set of values at once
- Consider the 32*12 segment as a 32*12 matrix, scaling factor of the subband is the maximum of the 12 samples.
  - The maximum is passed to the bit-allocation block along with SBS (subband samples)
  - The key point of the bit-allocation block is to apportion the bits to minimize the audibility of quantization noise.
- Psychoacoustic model is fairly complex
  - In layer1, to decide whether each frequency band is tone or noise.
  - From this decision and the scaling factor, a masking threshold is calculated and compared with the threshold of hearing

2.2.3 MPEG Audio Compression: Basic Algorithm

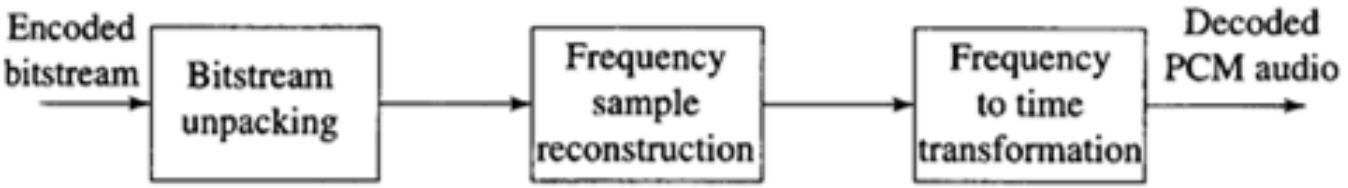
- The model's output consists of a set of signal-to-mask ratios (SMRs), flag frequency components with amplitude below the masking level
- SMR gives the amplitude resolution needed and controls the bit allocation
  - Scaling factors are used to set quantization levels so that the quantization error falls below the masking level
- The scaling factor is quantized using 6bits, then 12values in each subband are quantized using 4bits
  - Quantized scaling factor and 12codewords are grouped into Subband-Sample format.
- On the decoder side, the values are de-quantized.
  - Magnitudes of the 32 samples are reestablished, psychoacoustic model is not needed in the decoder.

2.2.3 MPEG Audio Compression Algorithm

- **Basic Algorithm**



MPEG Audio encoder



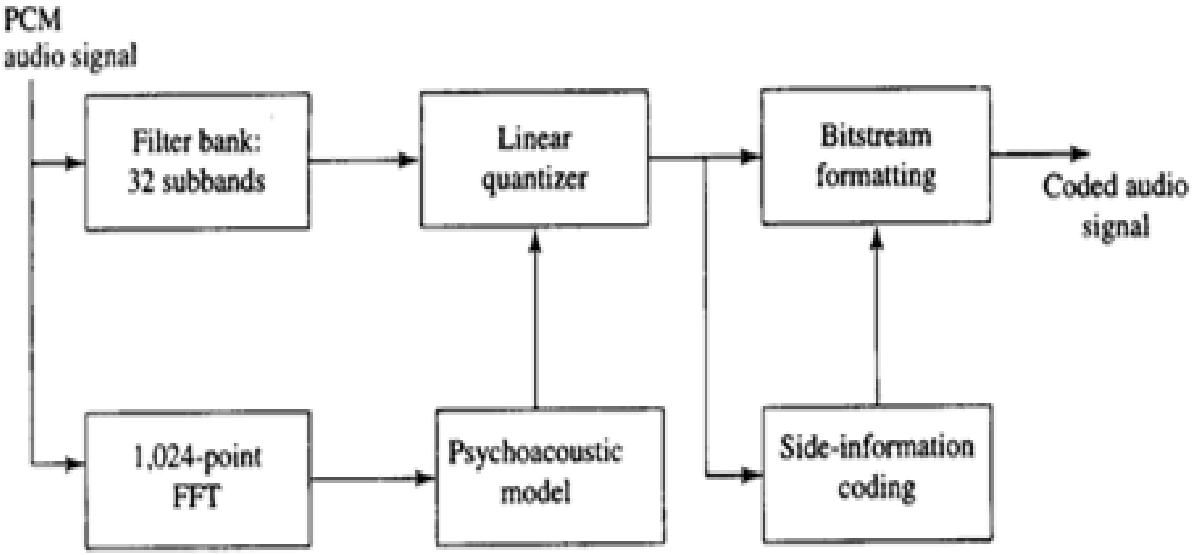
MPEG Audio decoder

2.2.3 MPEG Audio Compression Algorithm: Bit Allocation

- **Bit-allocation is not part of the standard**
  - To ensure that all the quantization noise is below the masking thresholds
- **A lookup table in the MPEG Audio standard provides an estimate of the SNR**
  - Mask-to-Noise Ratio (MNR) is defined as: $MNR_{dB}=SNR_{dB}-SMR_{dB}$
- **For the subband with the lowest MNR**
  - Allocated bits is incremented, and made a new estimation of the SNR
  - The process iterates until no more bits are left to allocate.
- **Mask calculation are performed parallel with subband filtering**
  - Decomposition of frequencies usually calculated with a 1024-point FFT.

2.2.3 MPEG Audio Compression Algorithm

- **Bit Allocation**



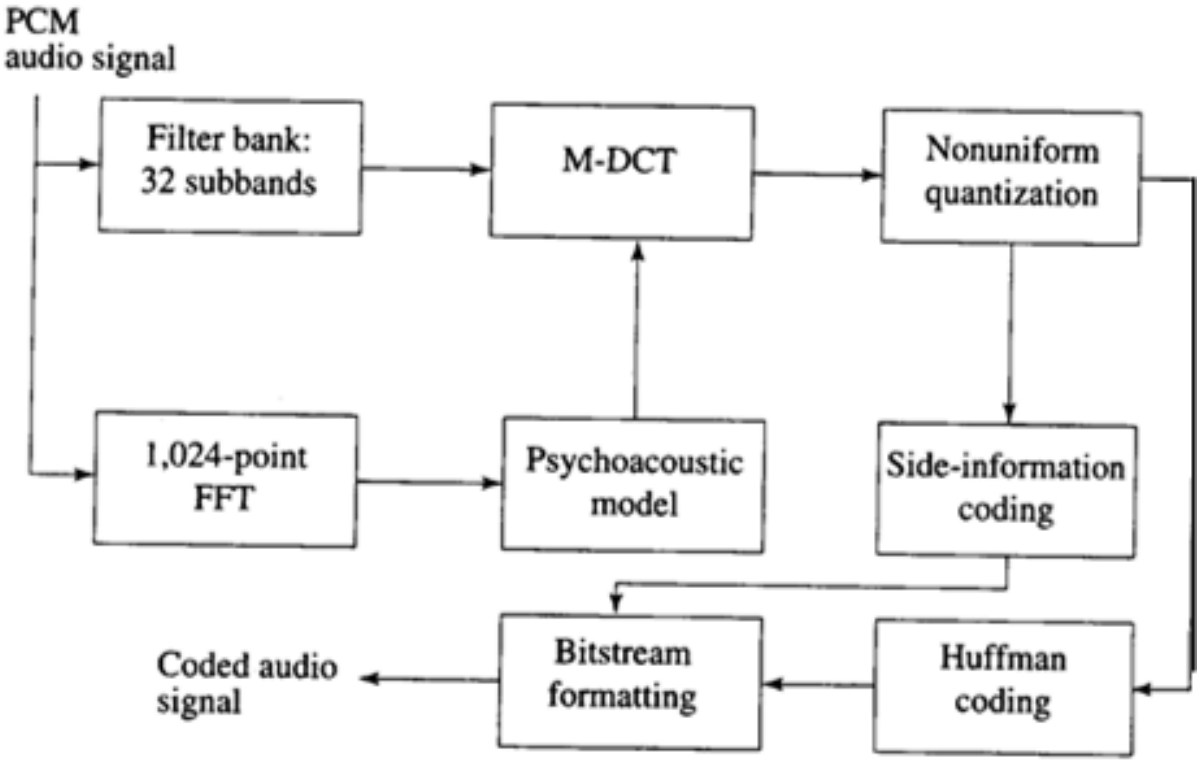
MPEG-1 Audio Layers 1 and 2

2.2.3 MPEG Audio Compression Algorithm

- **Layer 2**
  - **Layer2 improves codec efficiency at the price of an increase in complexity**
  - **Use temporal masking; bit allocation is applied to window lengths of 36 samples instead of 12.**
- **Layer 3**
  - **Better audio quality, at the price of increased complexity.**
  - **MDCT to deal with block edge effects introduced by DCT.**
  - **Carrying out entropy coding and making use of nonuniform quantizers.**

2.2.3 MPEG Audio Compression Algorithm

- **Layer 3**



MPEG-1 Audio Layer 3

2.2.4 MPEG-2 AAC (Advanced Audio Coding)

- Aimed at transparent sound reproduction for theaters, can deliver at 320kbps for 5 channels
- Capable of delivering high-quality stereo sound at bitrates below 128kbps.
- Can support up to 48 channels, sampling rate between 8kHz and 96kHz, bitrates up to 576kbps.
- Three profiles: Main、Low Complexity (LC) and Scalable Sampling Rate (SSR)

2.2.5 MPEG4 Audio

- **MPEG4 Audio integrates several different audio component into one standard:**
  - **Speech compression, perceptually based coders, text-to-speech and MIDI.**
  - **MPEG-4 AAC is similar to the MPEG-2 AAC standards, with some minor changes.**

- **Perceptual Coders**
  - **Perceptual Noise Substitution module, for scale factor bands above 4kHz decide whether they are noiselike or tonelike, just energy is transmitted if noiselike, and the frequency coefficient is set to zero.**
  - **Includes a Bit-Sliced Arithmetic Coding (BSAC) module**
  - **Includes a second perceptual audio coder**

2.2.5 MPEG4 Audio

- **Structured coders**
  - **A low bitrate delivery option:**

    Synthetic/Natural Hybrid Coding (SNHC)
  - **Takes a toolbox approach and allows specification of many such models.**
  - **Structured Audio Orchestra Language (SAOL) allows simple specification of sound synthesis, including special effects such as reverberation.**

The end!

Thanks!