AUDIO COMPRESSION

TOPICS TO BE COVERED

Introduction

- **Characteristics of Sound and Audio Signals**
- **Applications**
- **Waveform Sampling and Compression**

Types of Sound Compression techniques

- **Sound is a Waveform – generic schemes**
- **Sound is Perceived – psychoacoustics**
- **Sound is Produced – sound sources**
- **Sound is Performed – structured audio**

Sound Synthesis

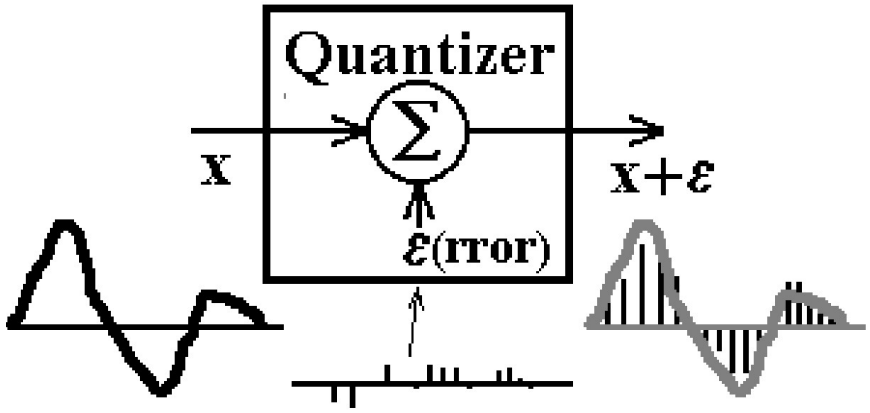
Audio standards

- **ITU - G.711, G.722, G.727, G.723, G.728**
- **ISO – MPEG1, MPEG2, MPEG4**

INTRODUCTION

Physics Introduction – Sound is a Waveform

Recording instruments convert it to an electrical waveform signal, which is then sampled and quantized to get a digital signal



Quantization introduces error! – Listen to 16, 12, 8, 4 bit music and see the difference.

A COMPARISON TO THE VISUAL DOMAIN

Sound is a 1D signal with amplitude at time *t* – which leads us to believe that it should be simple to compress as compared to a 2D image signal and 3D video signals

Is this true?

- **Compression ratios attained for video and images are far greater than those attained for audio**
- **Consider human perception factors – human auditory system is more sensitive to quality degradation than the visual system. As a result humans are more prone to compressed audio errors than compressed image and video errors**

APPLICATIONS

Telephone-quality speech

- **Sampling rate = 8KHz**
- **Bit rate is = 128 Kbps**

CDs (stereo channels)

- **Sampling rate = 44.1KHz**
- **Bit rate is 2x16x44100 = 1.4 Mbps!**
- **CD Storage = 10.5 Megabytes / minute**
- **CD can hold on 70 minutes of audio**

Surround Sound Systems with 5 channels

NEED FOR COMPRESSION

We need to take advantage of redundancy/correlation in the signal by statistically studying the signal – but just that is not enough!

The amount of redundancy that can be removed all through out is very little and hence all the coding methods for audio generally give a lower compression ratio than images or video

Apart from Statistical Study, more compression can be achieved based on

- **Study of how sound is perceived**
- **Study of how it is produced**

TYPES OF AUDIO COMPRESSION TECHNIQUES

Audio Compression techniques can be broadly classified into different categories depending on how sound is "understood"

Sound is a Waveform

- **Use Statistical Distribution / etc.**
- **Not a good idea in general by itself**

Sound is Perceived - Perception-Based

- **Psycho acoustically motivated**
- **Need to understand the human auditory system**

Sound is Produced - Production-Based

- **Physics/Source model motivated**

Music (Sound) is Performed/Published/Represented

- **Event-Based Compression**

SOUND AS A WAVEFORM

Uses variants of PCM techniques. PCM techniques produce a high data rates but can be reduced by exploiting statistical redundancy (information theory)

Differential Pulse Code Modulation (DPCM)

- **Get differences in PCM signals**
- **Entropy code differences**

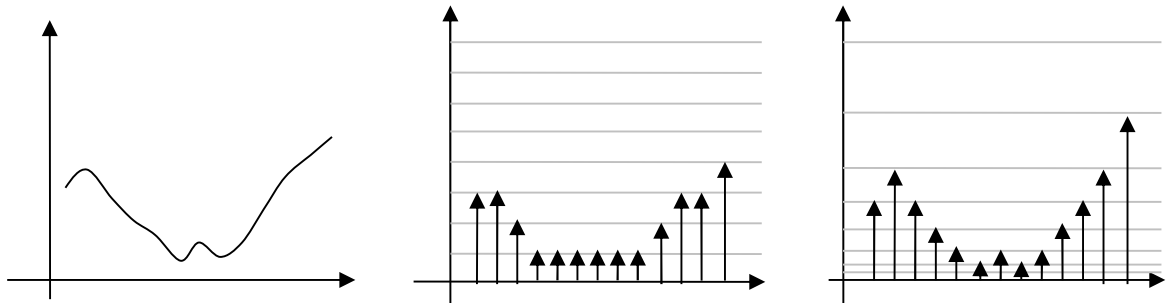
Delta Modulation

- **Like DPCM but only encodes differences using a single bit suggesting a delta increase or a delta decrease**
- **Good for signals that don't change rapidly**

Adaptive Differential Pulse Code Modulation

- Sophisticated version of DPCM.
- Codes the differences between the quantized audio signals using only a small number of specific bits which adaptively vary by signal
- Normally operates in one of two – high frequency mode or low frequency mode (why?)

Logarithmic Quantization



Different type of waveform based coding schemes

- A-law (Europe, ISDN 8KHz, 13 bits mapped to 8 log bits
- μ-law (America,Japan – maps 14 bits to 8 log bits)

SOUND IS PERCEIVED

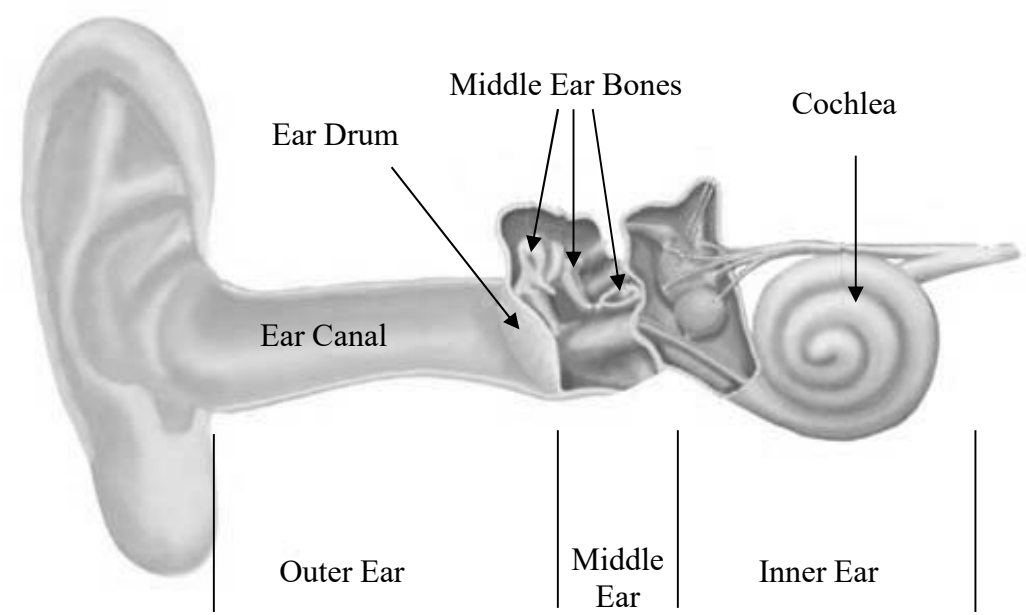
Compression attained by variations of PCM coding techniques alone are not sufficient to attain data rates for modern applications (CD, Surround Sound etc)

Perception of Sound additionally can help in compression by studying

- **What frequencies we hear**
- **When do we hear them**
- **When do not hear them**

This branch of study – Psychoacoustics – deals with sound perception science. "Auditory Masking" is a perceptual weakness of the ear, which can be used to exploit compression without compromising quality

STRUCTURE OF HUMAN EAR



LIMITS OF HUMAN HEARING

The human auditory system, although very sensitive to quality, has a few limitations, which can be analyzed by considering

- **Time Domain Considerations**
- **Frequency Domain (Spectral) Consideration**
- **Masking or hiding – which can happen in the Amplitude, Time and Frequency Domains**

Time Domain

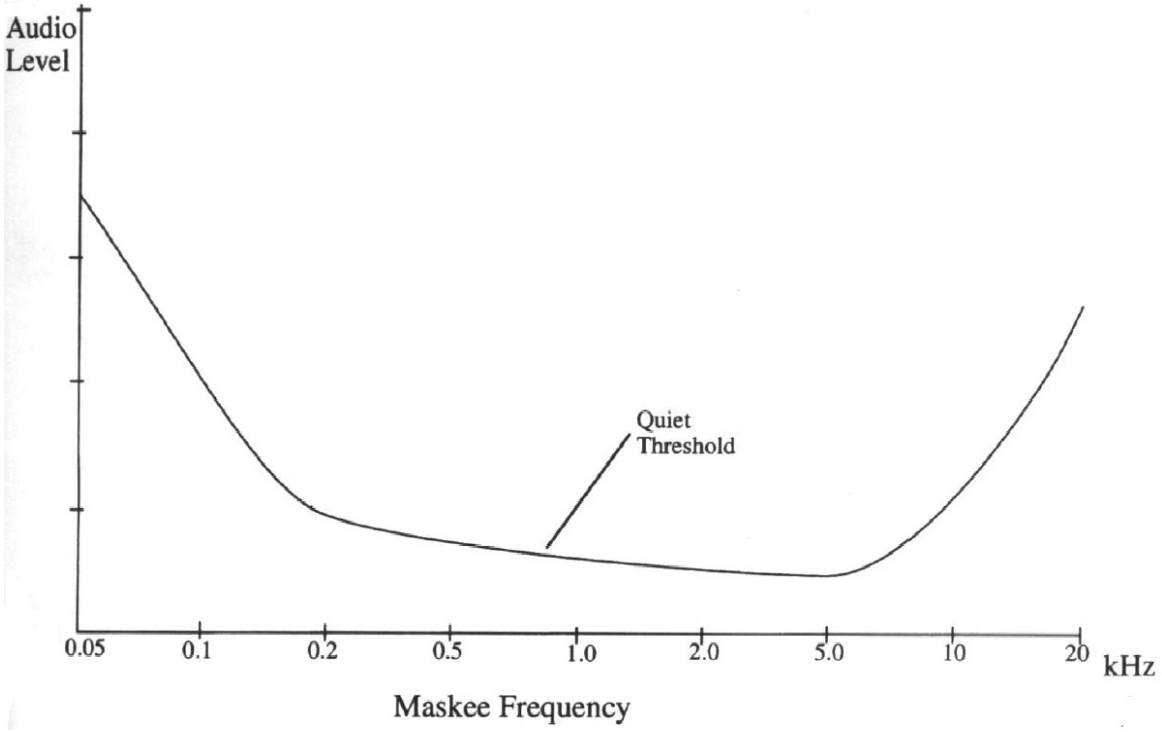
Events longer than 0.03 seconds are resolvable in time. Shorter events are perceived as *features in frequency*

Frequency Domain

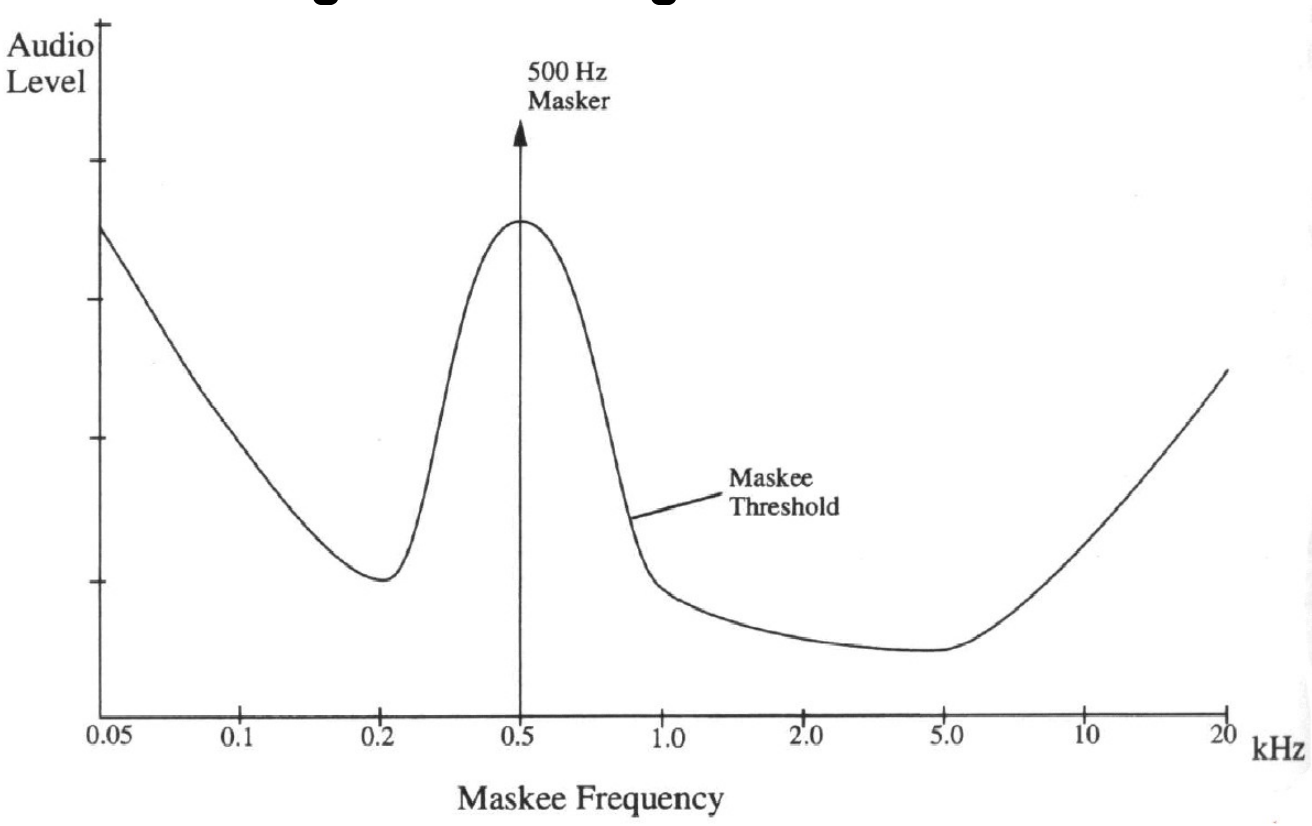
20 Hz. < Human Hearing < 20 KHz.
"*Pitch*" is perception related to frequency. Human Pitch Resolution is about 40 - 4000 Hz.

Masking

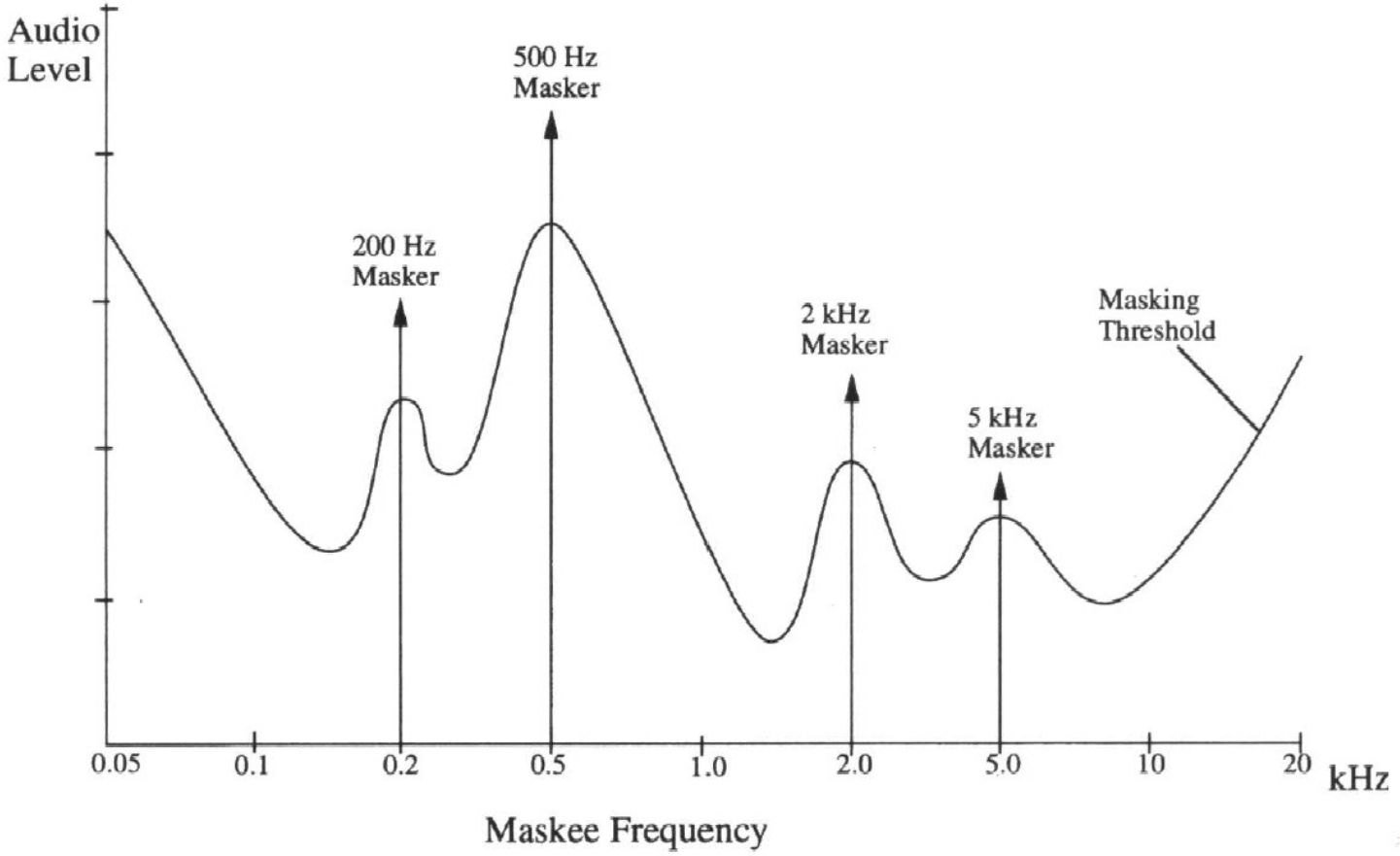
- **Masking as defined by the American Standards Association (ASA) is the amount (or the process) by which the threshold of audibility for one sound is raised by the presence of another (masking) sound.**
- **Masking Threshold Curve - A tone is *audible* only if its power is above the absolute threshold level**



- **If a tone of a certain frequency and amplitude is present, the *audibility threshold curve* is changed. Other tones or noise of *similar frequency*, but of *much lower amplitude*, are not audible – loud stereo sound in a car masks the engine noise.**
- **Masking Effect – Single Masker**



- **Masking Effect – Multiple Masker**



Masking in Amplitude

- **Loud sounds 'mask' soft ones – eg Quantization Noise. Intuitively, a soft sound will not be heard if there is a competing loud sound.**

- This happens because of gain controls within the ear – stapedes reflex, interaction (inhibition) in the cochlea and other mechanisms at higher levels

Masking in Time

- A soft sound just before a louder sound is more likely to be heard than if it is just after.
- In the time range of a few milliseconds
- A soft event following a louder event tends to be grouped perceptually as part of that louder event
- If the soft event precedes the louder event, it might be heard as a separate event.

Masking in Frequency

- Masking in Frequency – Loud 'neighbor' frequency masks soft spectral components. Low sounds mask higher ones more than high masking low.

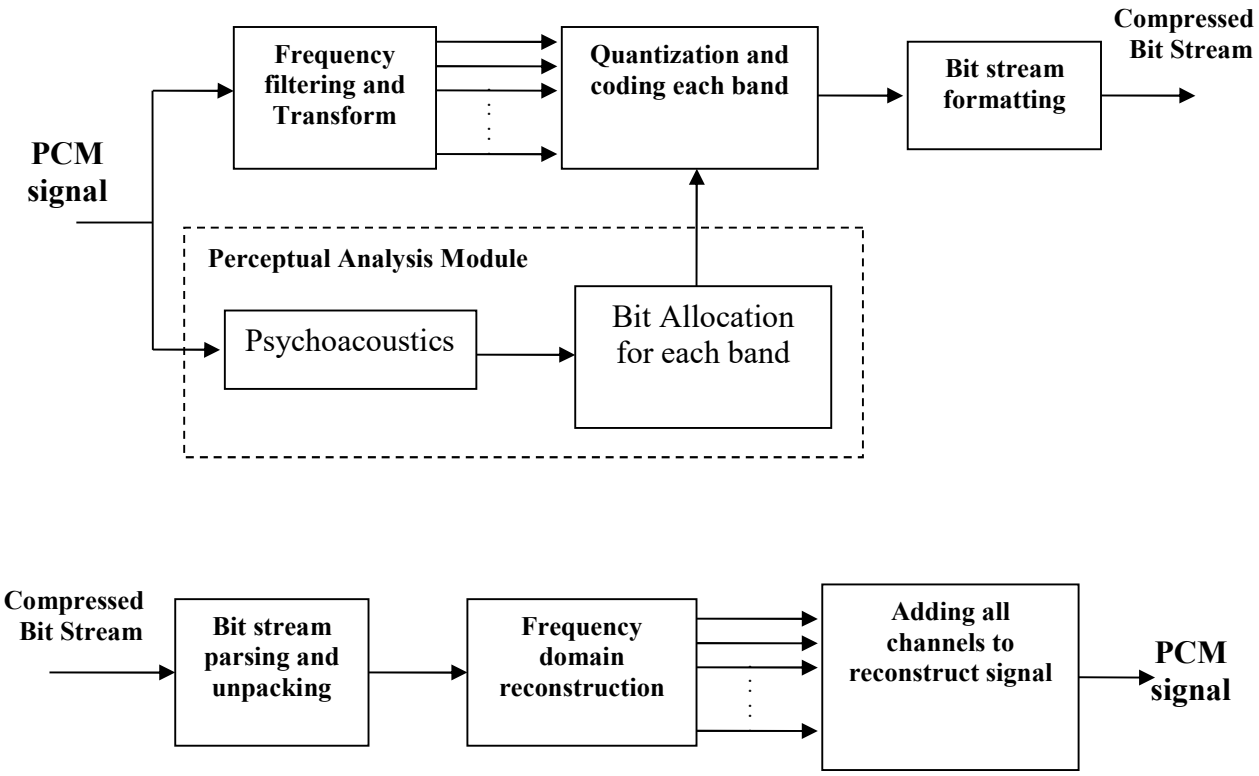
PERCEPTUAL CODING

Perceptual coding **tries to minimize the** ***perceptual distortion*** **in a transform coding scheme**

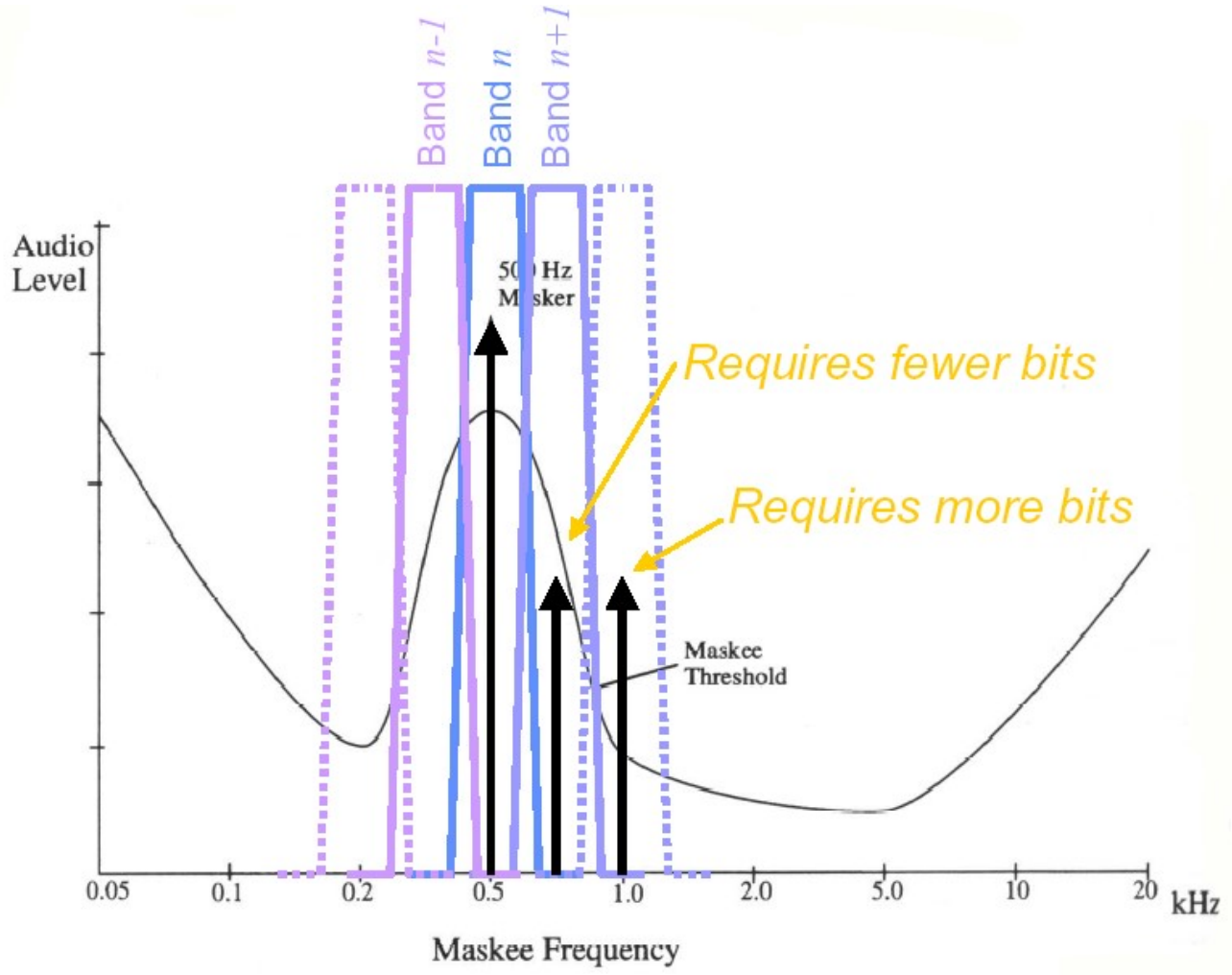
Basic concept: allocate more bits (more quantization levels, less error) to those channels that are most audible, fewer bits (more error) to those channels that are the least audible

Needs to continuously analyze the signal to determine the current ***audibility threshold*** **curve using a perceptual model**

PERCEPTUAL CODING



PERCEPTUAL CODING – EXAMPLE



SOUND IS PERCEIVED – REVISITED

The auditory system does not hear everything. The perception of sound is limited by the properties discussed above.

There is room to cut *without us knowing about it*! - by exploiting perceptual redundancy.

To summarize -

- **Bandwidth is limited – discard using filters**
- **Time resolution is limited – we can't hear over sampled signals**
- **Masking in all domains - psychoacoustics is used to discard perceptually irrelevant information. Generally requires the use of a perceptual model.**

SOUND IS PRODUCED

This is based on the assumption that a "perfect" model could provide the perfect compression. In other words, an analysis of frequencies (and variations) produced by the sound sources, yield properties of the signal it produces. A model of this sound production source is then built and the model parameters are adjusted according to sound it produces.

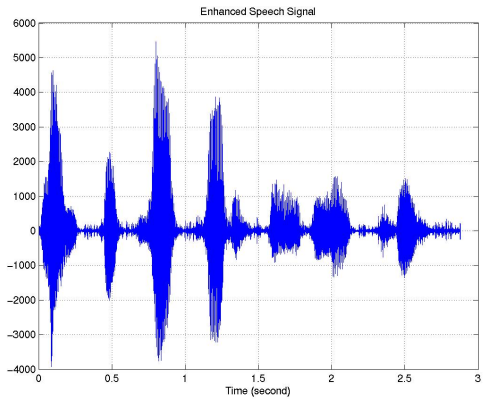
For example, if the sound is human speech then a well parameterized vocal model can yield high quality compression

Advantage - great at compression and maybe quality

Drawbacks - signal sources must be assumed, known apriori, or identified. Complex when a sound scene has one or more widely different sources.

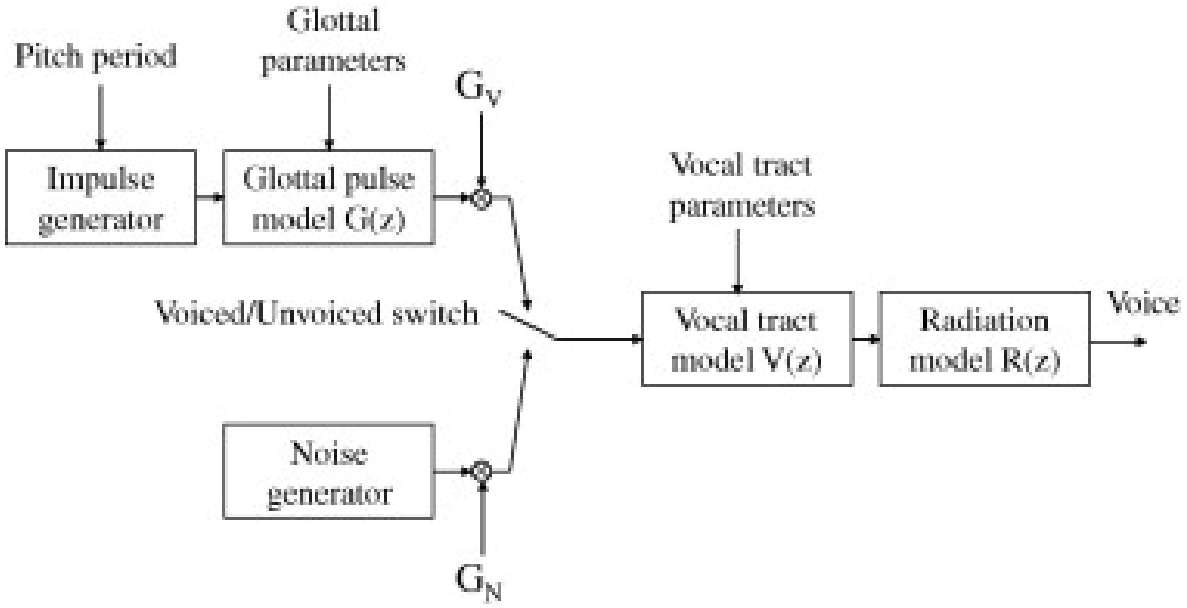
LINEAR PREDICTIVE CODING (LPC)

- **Stationary vs Non Stationary**



- **Human Speech is very highly non stationary**
- **Dynamically changes over time, change is quick – need to approximate to stationary via frame blocking**
- **Each window may be categorized as "voiced" (vowels, consonants) or "unvoiced"**

LPC DECODER / SYNTHESIZER



SOUND IS PERFORMED OR PUBLISHED

This sound is also known as *event based audio* or *structured audio*

Description format that is made up of *semantic information* **about the sounds it represents, and that makes use of** *high-level* **(algorithmic) models**

***Event-list representation*: sequence of control parameters that, taken alone, do not define the quality of a sound but instead specify the ordering and characteristics of parts of a sound with regards to some external model**

EVENT-LIST REPRESENTATION

Event-list representations are appropriate to soundtracks, piano, percussive instruments. Not good for violin, speech and singing

***Sequencer*s: allow the specification and modification of event sequences**

MIDI

MIDI (Musical Instrument Digital Interface) is a system specification consisting of both hardware and software components that define interconnectivity and a communication protocol for electronic synthesizers, sequencers, rhythm machines, personal computers and other musical instruments

***Interconnectivity* defines standard cabling scheme, connectors and input/output circuitry**

***Communication protocol* defines standard multibyte *messages* to control the instrument's voice, send responses and status**

MIDI COMMUNICATION PROTOCOL

The MIDI communication protocol uses multibyte messages of two kinds: *channel messages* and *system messages. Channel message*s address one of the 16 possible *channels*

***Voice Message*s: used to control the voice of the instrument**

- **Switch notes on/off**
- **Send key pressed messages**
- **Send control messages to control effects like vibrato, sustain and tremolo**
- **Pitch-wheel messages are used to change the pitch of all notes**
- ***Channel key pressure* provides a measure of force for the keys related to a specific channel (instrument)**

MIDI FILES

MIDI messages are received and processed by a MIDI sequencer asynchronously (in real time)

- **When the synthesizer receives a "note on" message it plays the note**
- **When it receives the corresponding "note off" it turns it off**

If MIDI data is stored as a data file, and/or edited using a sequencer, the tone form of "time stamping" for the MIDI message is required and is specified by the Standard MIDI file specifications.

SOUND REPRESENTATION AND SYNTHESIS

Sampling –

Individual instrument sounds (notes) are digitally recorded and stored in memory in the instrument. When the instrument is played, the note recording are reproduced and mixed to produce the output sound

Takes a lot of memory! To reduce storage:

- **Transpose the pitch of a sample during playback**
- **Quasi-periodic sounds can be "looped" after the *attack transient* has died**

Used for creating sound effects for film *(Foley)*

SOUND REPRESENTATION AND SYNTHESIS (2)

Additive and subtractive synthesis –

- **Synthesize sound from the superposition of sinusoidal components** *(additive*) **or from the filtering of an harmonically rich source sound** *(subtractive*)
- **Very compact but with "analog synthesizer" feel**

Frequency modulation synthesis –

- **Can synthesize a variety of sounds such as brass-like and woodwind-like, percussive sounds, bowed strings and piano tones**
- **No straightforward method available to determine a FM synthesis algorithm from an analysis of a desired sound**

AUDIO CODING: MAIN STANDARDS

MPEG **(Moving Picture Expert Group) family**

- **MPEG1 - Layer 1, Layer 2, Layer 3 (MP-3)**
- **MPEG2 - Back-compatible with MPEG1, AAC (non-back-compatible)**
- **MPEG4 – CELP and AAC**

Dolby **AC3**

ITU Speech Coding Standards

- **ITU G.711**
- **ITU G.722**
- **ITU G.726, G.727**
- **ITU G.729, G.723**
- **ITU G.728**

MPEG-1 AUDIO CODER

Layered Audio Compression Scheme, each being backward compatible

Layer1

- **Transparent at 384 Kbps**
- **Subband coding with 32 channels (12 samples/band)**
- **Coefficient normalized (extracts Scale Factor)**
- **For each block, chooses among 15 quantizers for perceptual quantization**
- **No entropy coding after transform coding**
- **Decoder is much simpler than the encoder**

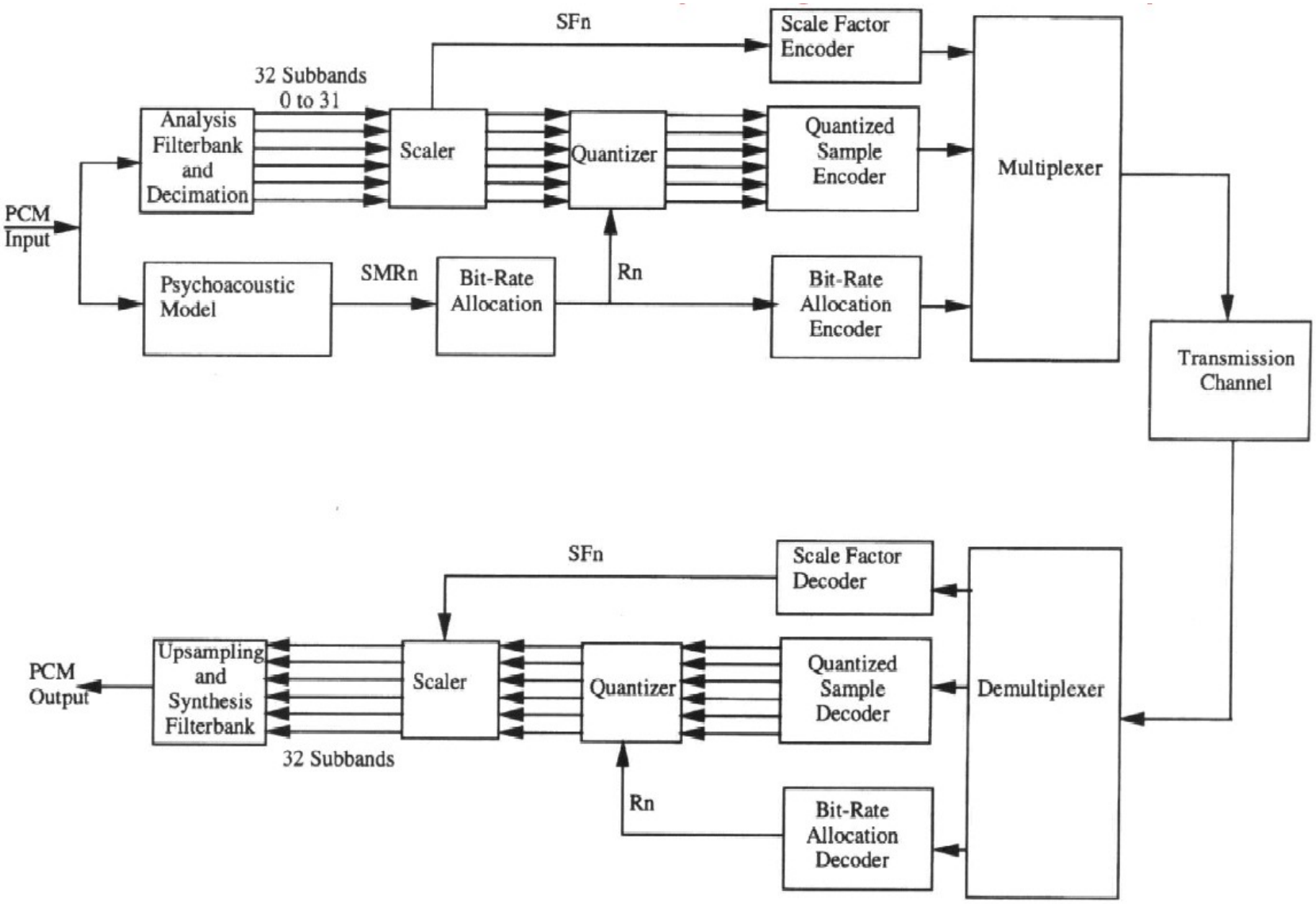
Layer2

- **Transparent at 296 Kbps**
- **Improved perceptual model**
- **Finer resolution quantizers**

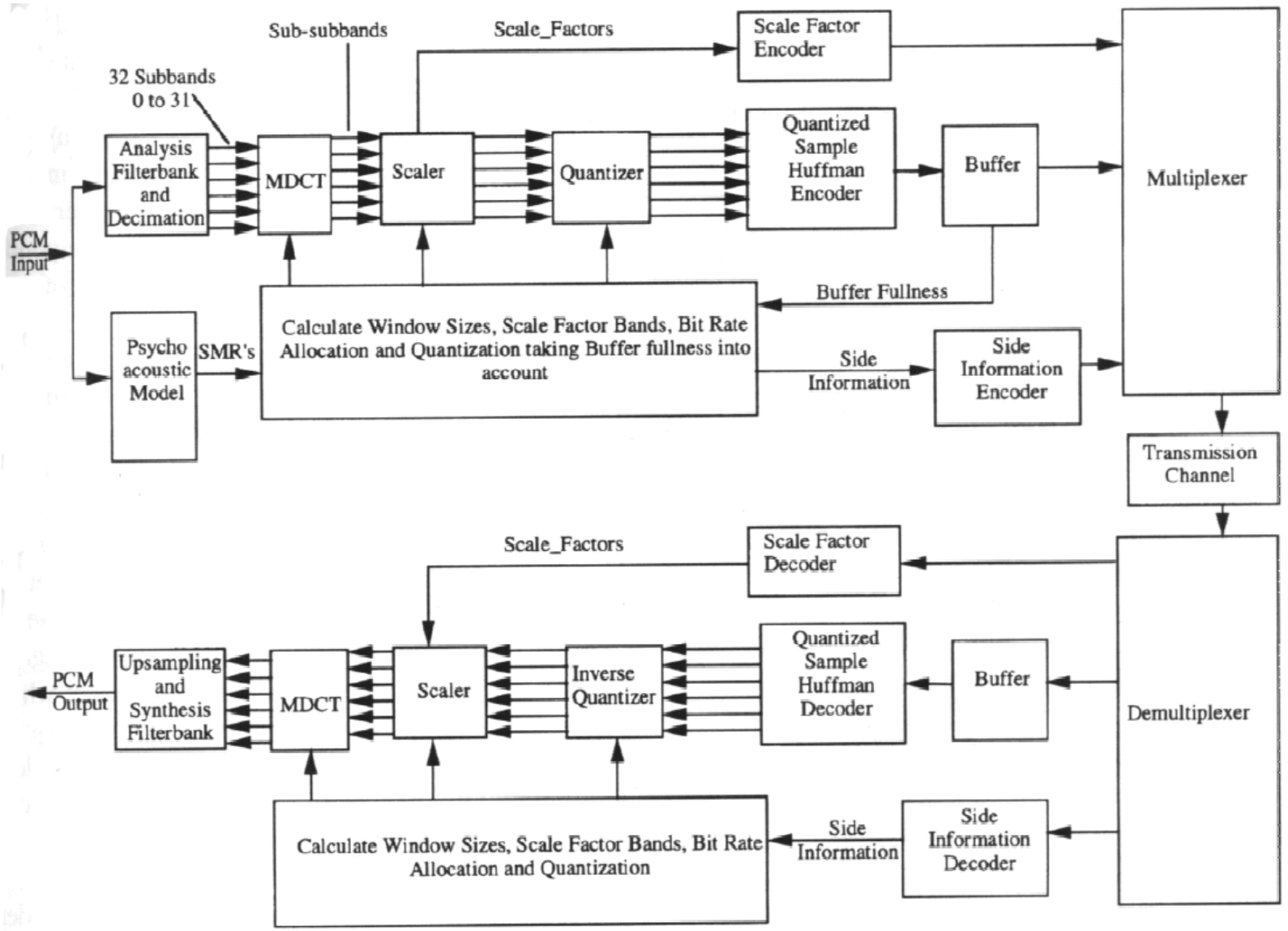
Layer 3

- Transparent at 96 Kb/s per channel
- Applies a variable-size modified DCT on the samples of each subband channel
- Uses non-uniform quantizers
- Has entropy coder (Huffman) - requires buffering!
- Much mode complex than Layer 1 and 2

MPEG-1 LAYERS 1 AND 2 AUDIO CODEC



MPEG-1 LAYER 3 (MP3) AUDIO CODEC



MPEG-2 AUDIO CODEC

Designed with a goal to provide theater-style surround-sound capabilities and backward compatibility. Has various modes of surround sound operation:

- **Mono-aural**
- **Stereo**
- **Three channel (left, right and center)**
- **Four channel (left, right, center, rear surround)**
- **Five channel (four channel + center) at 640 kbps**

Non-backward compatible (*AAC*):

- **At 320 Kb/s judged to be equivalent to MPEG-2 at 640 Kb/s for five-channels surround-sound**
- **Can operate with any number of channels (between 1 and 48) and output bit rate (from 8 Kb/s per channel to 182 Kb/s per channel)**
- **Sampling rates between 8Khz and 96 KHz per ch**

DOLBY AC-3

Used in movie theaters as part of the *Dolby digital film* system.

Selected for the USA Digital TV (DTV) and DVD

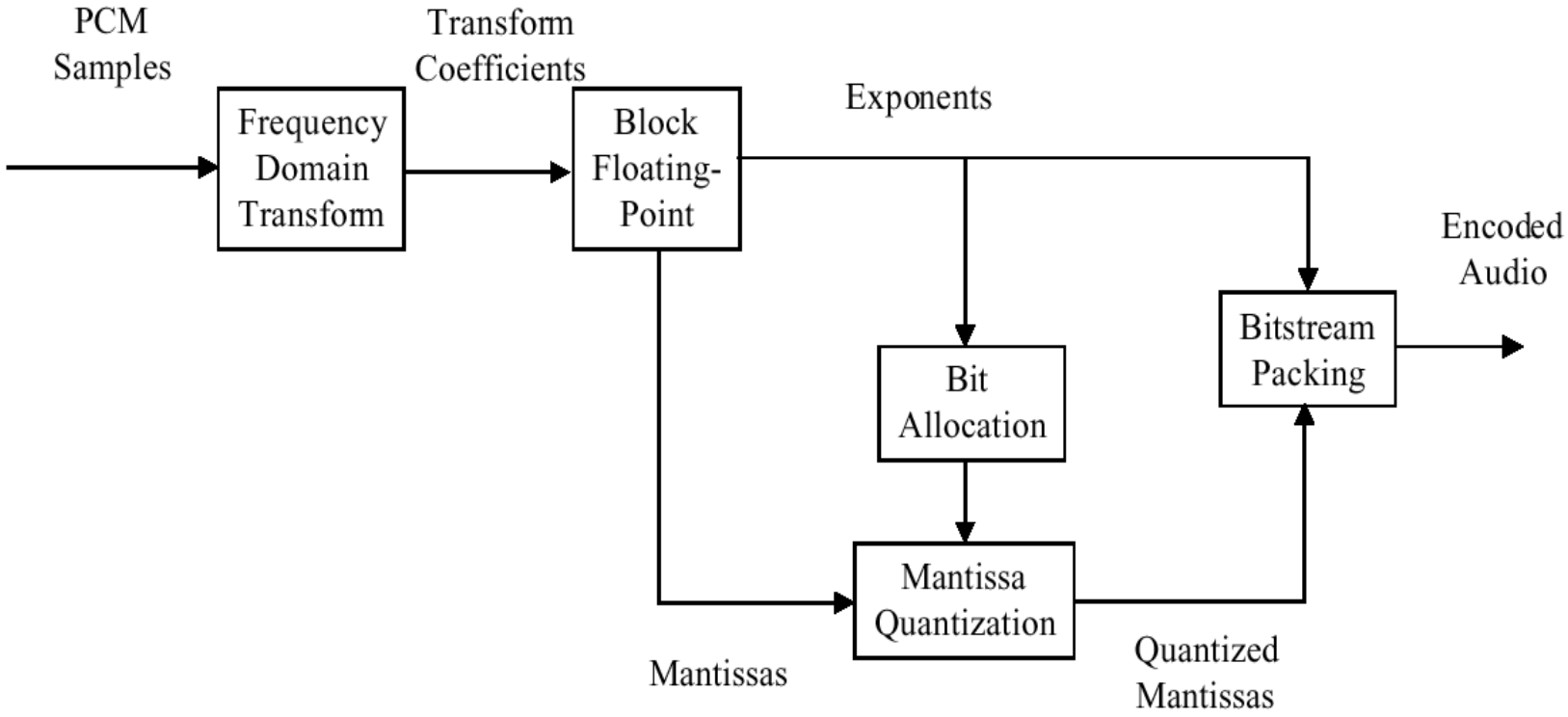
Bit-rate: 320 Kb/s for 5.1 stereo

Uses 512-point Modified DCT (can be switched to 256-point)

Floating-point conversion into exponent-mantissa pairs (mantissas quantized with variable number of bits)

Does not transmit bit allocation but perceptual model parameters

DOLBY AC-3 ENCODER



ITU SPEECH COMPRESSION STANDARDS

Standard	Bit rate	Frame size/ Look-ahead	Complexity
G.711 PCM	64 Kb/s	0 / 0 ms	0 MIPS
G.726, G.727	16,24,32,40 Kb/s	0.125 / 0 ms	2 MIPS
G.722	48,56,64 Kb/s	0.125 / 1.5 ms	5 MIPS
G.728	16 Kb/s	0.625 / 0 ms	30 MIPS
G.729	8 Kb/s	10 / 5 ms	20 MIPS
G.723	5.3 & 6.4 Kb/s	30/7.5 ms	16 MIPS

ITU G.711

Designed for telephone bandwidth speech signal (3KHz)

Does direct sample-by-sample non-uniform quantization (PCM). Provides the lowest delay possible (1 sample) and the lowest complexity. Employs u-law and A-law encoding schemes.

High-rate and no recovery mechanism, used as the default coder for ISDN video telephony

ITU G.722

Designed to transmit 7-Khz bandwidth voice or music

Divides signal in two bands (high-pass and low-pass), which are then encoded with different modalities

But G.722 is preferred over G.711 PCM because of increased bandwidth for teleconference-type applications. Music quality is not perfectly transparent.

ITU G.726, G.727

Has ADPCM (Adaptive Differential PCM) codecs for telephone bandwidth speech. Can operate using 2, 3, 4 or 5 bits per sample

ITU G.729, G.723

Model-based coders: use special models of production (synthesis) of speech

- Linear synthesis
- Analysis by synthesis: the optimal "input noise" is computed and coded into a multipulse excitation
- LPC parameters coding and Pitch prediction

Have provision for dealing with frame erasure and packet-loss concealment (good on the Internet)

G.723 is part of the standard H.324 standard for communication over POTS with a modem

ITU G.728

Hybrid between the lower bit-rate model-based coders (G.723 and G.729) and ADPCM coders

Low-delay but fairly high complexity

Considered equivalent in performance to 32 Kb/s G.726 and G.727

Suggested speech coder for low-bit rate (64-128 Kb/s) ISDN video telephony

Remarkably robust to random bit errors

QUESTION

Both the visual image/video encoders and the psychoacoustic audio encoders work by converting the input spatial or time domain samples to the frequency domain and quantize the frequency coefficients.

- **How is the conversion to the frequency domain different for visual encoders compared to audio encoders? Why is this difference made for audio encoders?**

QUESTION

Both the visual image/video encoders and the psychoacoustic audio encoders work by converting the input spatial or time domain samples to the frequency domain and quantize the frequency coefficients.

- **How is the conversion to the frequency domain different for visual encoders compared to audio encoders? Why is this difference made for audio encoders?**
- **How does the quantization of frequency coefficients in the psychoacoustic encoders differ from that used in the visual media types?**

QUESTION

Both the visual image/video encoders and the psychoacoustic audio encoders work by converting the input spatial or time domain samples to the frequency domain and quantize the frequency coefficients.

- **How is the conversion to the frequency domain different for visual encoders compared to audio encoders? Why is this difference made for audio encoders?**
- **How does the quantization of frequency coefficients in the psychoacoustic encoders differ from that used in the visual media types?**
- **Why is it necessary to transmit the bit rate allocation in the bit stream for audio encoders? Does this have to be done in the beginning, towards the end or often – Explain! How do the visual encoders convey the bit rate allocation?**