

Audio and Video

CSC 790

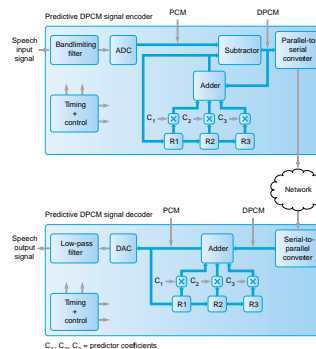
WAKE FOREST
UNIVERSITY

Department of Computer Science

Fall 2009

Adaptive PCM

- Adaptive PCM (ADPCM) varies number bits to encode differences
 - Large steps for high frequencies and small for low frequencies
 - Use previous samples to predict changes in the future
 - ITU-T Recommendation G.721
- For example, 3 predictor coefficients (samples) are used below



Linear Predictive Coding

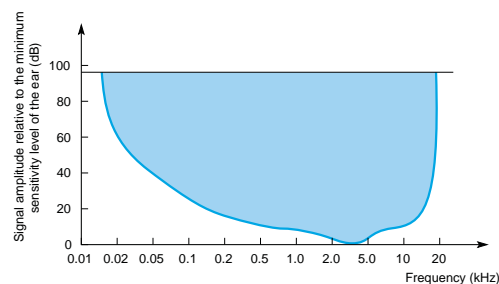
- Previous methods are based on PCM and sending the difference
- Alternative methods analyze the signal for certain features
 - Features are digitized then sound *synthesized* at receiver
 - Methods called **Linear Predictive Coding** (LPC)

What is this method limited to?

- Features that are noticeable to the human ear include
 - Pitch (similar to frequency), duration, and loudness
- Also extract origins of the sound
 - Voiced or unvoiced sounds
- Code-excited LPC (CLEP) has been standardized by ITU-T Recommendations G.728, 729, and 729(A)

Perceptual Encoding

- Use a psychoacoustic model to exploit limitations of the human ear
 - Human ear can only hear frequencies from 15 Hz to 20 kHz



- **Frequency masking** can occur
 - Loss of one frequency due to another
- **Temporal masking** can also occur
 - After loud sound, human ear needs finite amount of time before a quieter sound can be heard

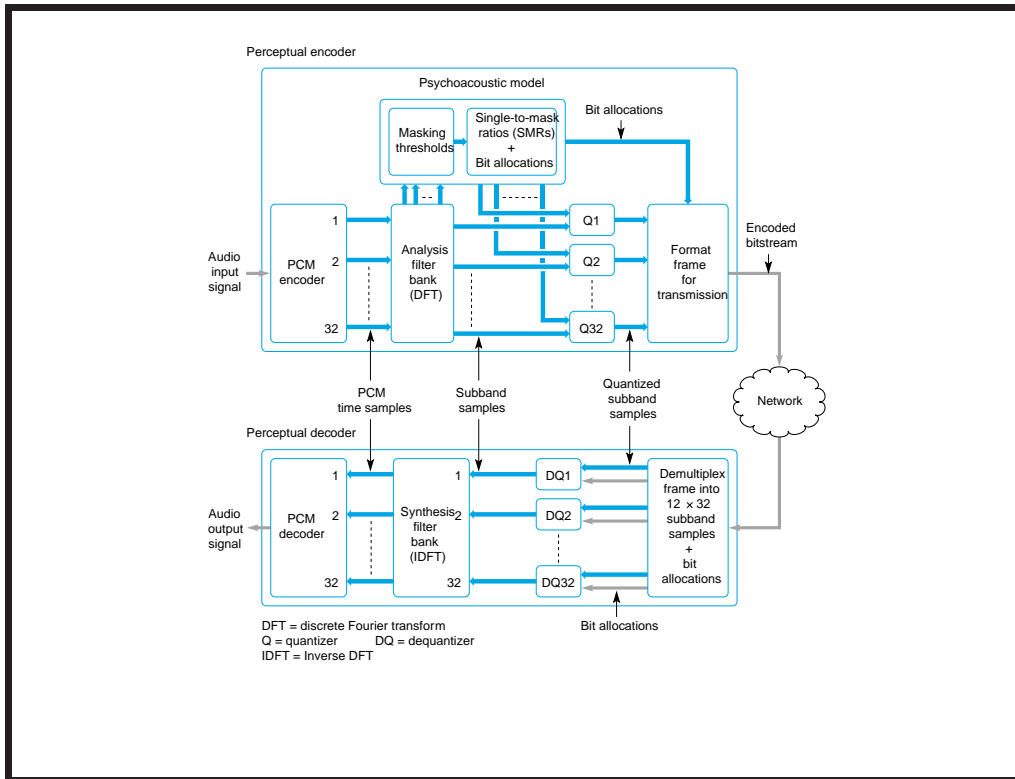
MPEG Audio

- Audio encoder/decoder associated with MPEG video
 - First applies filters to audio, to determine frequencies
 - In parallel, psychoacoustic model applied the quantized
- MPEG audio has three downward-compatible *layers* of compression
 - Layer 1 has high quality and requires a high bit rate
 - Layer 2 is more complex, proposed for digital audio broadcast
 - Layer 3 (mp3) was designed for audio transmission over ISDN
 - *Main difference is in the psychoacoustic model used*

Layer	Application	Bit Rate (kbps)	Quality	Input to Output Delay
1	Digital recording	32 - 448	HiFi at 192 kbps	20 msec
2	Digital broadcasting	32 - 192	CD at 192 kbps	40 msec
3	CD quality transmission	64	CD quality at 64 kbps	60 msec

MPEG Audio Compression

- Audio input is sampled a regular intervals
 - Quantized using PCM (*quality depends on application*)
- Bandwidth available is divided into 32 **frequency subbands**
 - Each frequency subband has equal width
- Filter divides input into the different subbands
 - Input \Rightarrow 32 PCM samples, output \Rightarrow 32 frequency coefficients
 - Frequency information used by psychoacoustic model
- Psychoacoustic model used to determine quantization values



E. W. Fulp

Fall 2009

Audio Compression Standards

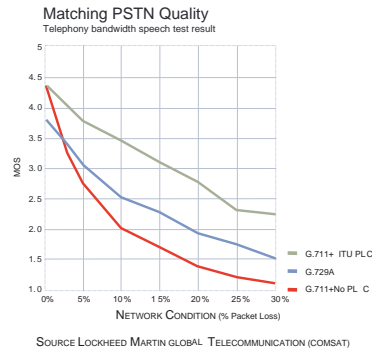
- Below standards are from the ITU-T, except for LPC-10
 - These standards are for speech encoding

Standard	Compression	Bit Rate (kbps)	Speech Quality	Application
G.711	PCM & companding	64	Good	PSTN/ISDN telephony
G.721	ADPCM	32 (16)	Good (fair)	Telephony
G.722	ADPCM & subbanding	64, 56/48	Excellent (Good)	Audio conference
G.726	ADPCM & subbanding	40/32 (24/16)	Good (Fair)	Telephony
LPC-10	Linear Pred. Code	2.4/1.6	Poor	Military telephony
G.728	CLEP	16	Good	Low Delay Telephony
G.729	CLEP	8	Good	Cellular Telephony
G.729(A)	CLEP	8	Good	Simultaneous Telephony & FAX
G.723.1	CLEP	6.3 (5.3)	Good (fair)	VoIP

E. W. Fulp

Fall 2009

Speech Quality



- Limited information comparing the *quality* of speech encoders
 - If used over the Internet, loss and delay are important

Is it possible that VoIP could be better than POTS?

Silence Detection

- Simple method to reduce network usage is **not** to transmit silence
This helps in to ways, what are they?

- Question is *how do we detect silence?*

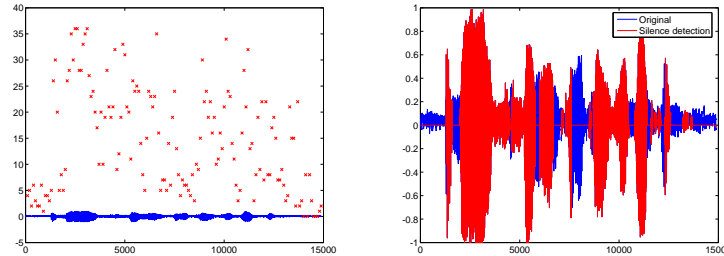
- Magnitude based method, measure difference in signal amplitude
 - Simple technique, compare each sample to threshold
 - Some difficulty with *background noise*
 - Can also compare successive samples

How do you distinguish between speech and silence?

- Another method is zero-crossing rate of the signal
 - Number of times the signal cross zero divided by number of samples (speech typically has higher zero-crossing rate)

Example of Silence Detection

- *What is the appropriate window and number of crossings?*
 - Example below, window = 100 samples, threshold = 10



- Problems at the beginning and end of speech, as well as lack of background noise

How does this help again? transmission or encoding?

Audio File Formats

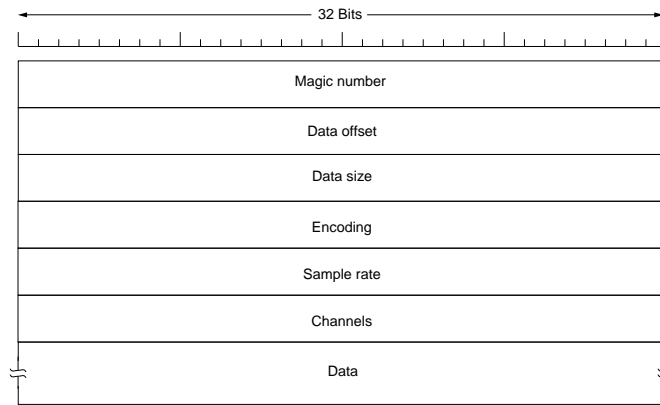
- A container format for storing audio data
 - Includes header (info about codec) and data (audio samples)

Is there a difference between file format and codec?

- There are three major categories of audio file formats
 - Common formats such as .wav aiff, and .au
 - Formats with lossless compressions such as .flac and .wma
 - Formats with lossy compressions such as .mp3 and .aac
- *Most formats support only one codec, however a format can support multiple codecs...*

AU

- Simple audio file format introduced by Sun Systems
 - Support different codecs but μ -law used most often
- File consists of 6 32-bit word header, optional info, then data



- Header contents include the following
 - Magic number is 0x2e736e64
 - Data offset is the offset to the data in bytes
 - What is the minimum number?*
 - Data size is number data bytes, if unknown then 0xffffffff
 - Encoding indicates the type of codec used for the data

Value	Format
1	8-bit G.711 μ -law
2	8-bit linear PCM
27	8-bit G.711 A-law

- Sample rate is number of samples per second
- Channels are the number of interleaved channels

Video Compression Principles

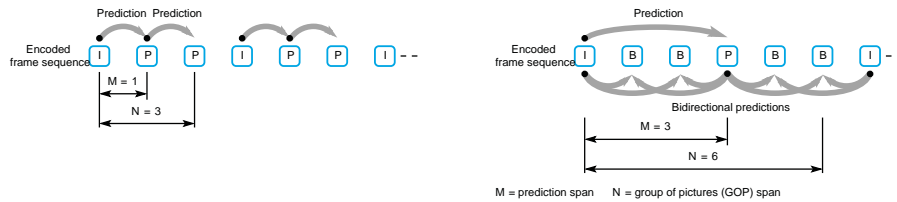
- Video is simply a sequence of digitized pictures (images or frames)
 - Therefore, we could just use JPEG for each picture
 - This is motion JPEG (MJPEG) compression is 10:1 to 20:1
What kind of redundancy is removed with MJPEG?
 - *Not enough compression for the current packet networks*
- Can also remove the redundancy that may occur between frames
 - Some parts of the image may not change from frame to frame
 - Only send information about what has changed between frames
 - *Most video systems use spatial and temporal compression*

Redundancy

- JPEG compression is **spacial redundancy**
 - Compression may not be significant enough for video
- **Temporal redundancy** is between successive frames
 - Video can be considered a *stack* of images
 - Given 15 fps, little changes between successive frames
Does this mean a pixel by pixel difference between frames?

Group of Frames

- There are two basic video frame types
 - Frames encoded independently, called intracoded or **I-frames**
 - Frames that are predictive, **P-frames** or **B-frames**



- Number of frames between successive I-frames is a **group of pictures (GOP)**, denoted as N , can be 3 to 12
 - For example, GOP order could be IBBPBBPBB

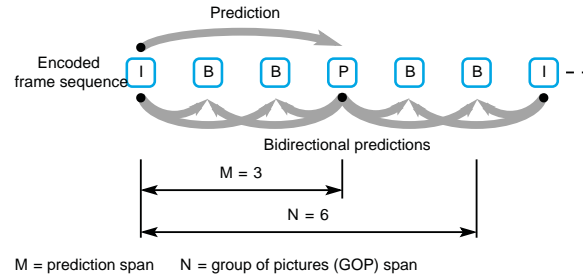
Frame Types

- I-frames are encoded without reference to any other frames
 - Each frame is treated as separate, encoded using JPEG
 - Must appear at regular intervals
- P-frame encoded relative to preceding I or preceding P-frame
 - Encoded with **motion estimation** and **compensation**
 - Transmit the difference between preceding frame
- B-frame encoded using past and future frames
 - Differences between frames is transmitted
 - Have the highest compression ratio

B-frames do not propagate errors, why?

Frame Order

- Given I, P and B frames, the undecoded order is



– For example, IBBPBBPBBPBI

- When transmitted the sequence is typically reordered

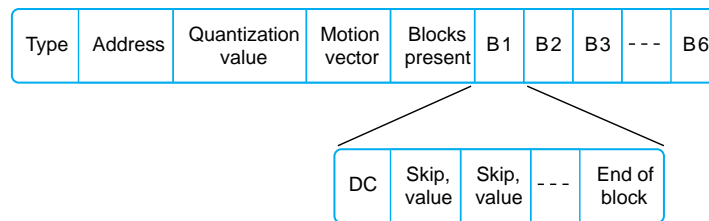
IPBBPBBPIPPBPP

What?

Macroblocks

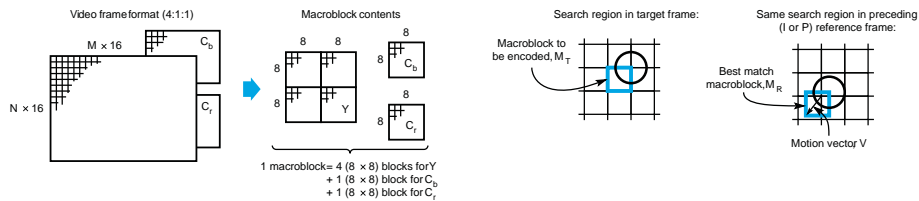
- Frame is divided into $n \times n$ **macroblocks**, general format below

– Macroblock has an address



- Type indicates if the macroblock has been encoded independently (intracoded) or with reference to a macroblock in preceding frame (intercoded)
- Quantization value is the threshold used to quantize all the DCT coefficients

- To encode a P-frame (**target frame**) the macroblocks are compared to the preceding I or P-frame (**reference frame**)
 - If a match is found the address is encoded



- If a close match then address and **motion vector** is encoded
- If no match then macroblock encoded like I-frame

Which macroblock should you search?

H.261

- H.261 is a video compression standard defined by the ITU in 1990
 - Designed for video telephony and conferencing over ISDN
 - Assume bit rate is multiples of 64 kbps, $p \times 64$
- Only I and P-frames are used
 - GOP is three P-frames between I-frames IPPPIPPI
 - Why no B-frames?*
 - Video encoder delay must be < 150 msec
- H.261 also organizes macroblocks into groups

GOB

176 pixels

48 pixels

1 MB = 16 × 16 pixels (luminance)

352 pixels

288 pixels

CIF resolution

176 pixels

144 pixels

QCIF resolution

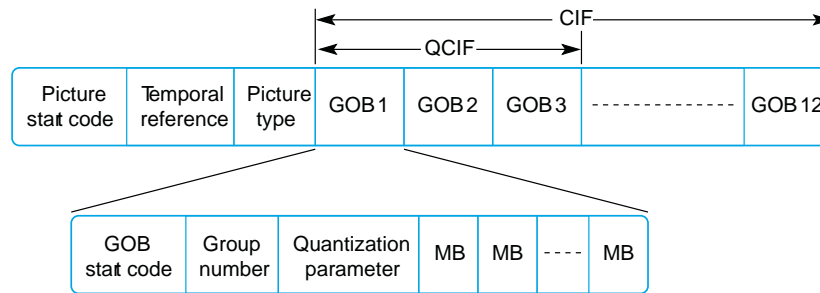
GOB = Group of (macro) blocks

- **Group of Macroblocks** is an 11×3 array of macroblocks
 - Each GOB has a unique *start code*
 - If error occurs the decoder moves to the next GOB start code
 - Start code is also known as the *resynchronization marker*

- If an error, the remaining macroblocks in the GOB are skipped
 - To mask the error, display preceding corresponding GOB
 - Called *error concealment scheme*

Sounds great, but an error may propagate across multiple frames... how?
- Control is at the GOB level, not macroblock
 - If the bit stream is greater than the allowed bit rate, then quantization per GOB is altered (*constant rate not quality*)
 - If quantization is not enough, then GOB are dropped
 - Dropping GOB is not enough, then frames are dropped

H.261 Video Frame



- Start of each new video frame is indicated by *picture start code*
 - Followed by *temporal reference* field time stamp used to synchronize audio with video
 - Picture type indicates I, P, or B-frame
 - GOBs that make the picture follow