Music & Engineering: Digital Encoding and Compression

Tim Hoerning

Fall 2008

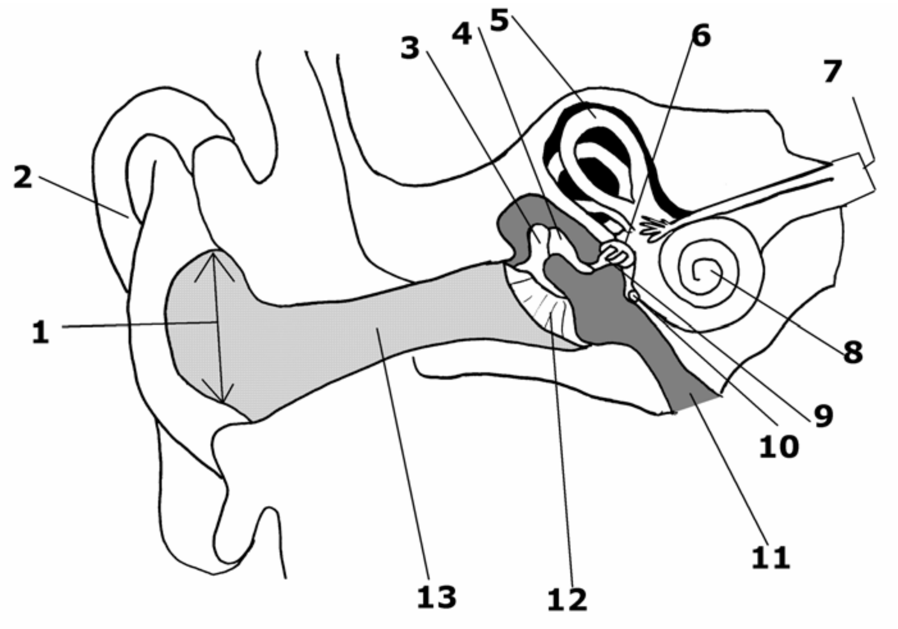
(last modified 10/29/08)

Overview

- The Human Ear
- Psycho-Acoustics
  - Masking
  - Critical Bands
- Digital Standard Overview
  - CD
  - ADPCM
  - MPEG (history, layers, etc)
- Compression
  - Lossless (are then any methods?)
  - Lossy
    - Perceptual Encoding
      - MPEG Audio Layers
    - Source Encoding
      - Vocoder

The Human Ear

Ear Diagram



1. External Auditory Opening
2. External Ear (Pinna)
3. Maleus
4. Incus
5. Semicircular Canals
6. Oval Window (under Stapes)
7. Vestibulocochlear Nerve
8. Cochlea
9. Stapes
10. Round Window
11. Eustachian Tube
12. Tympanic Membrane
13. External Auditory Meatus

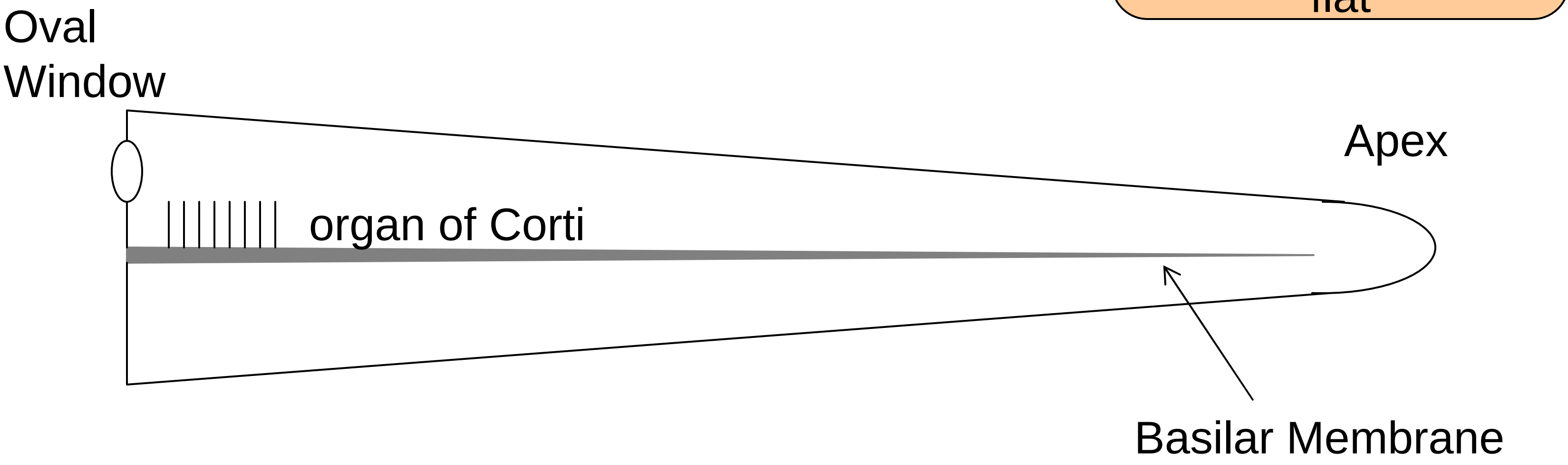
http://www.mansfield.ohio-state.edu/~jbradley/EarA.html

Ear Parts

- Outer Ear – Pinnae & Auditory Canal
  - Affects higher frequency sounds – assists localization
  - Canal is a 1/4 wavelength resonator in the 3-5kHz range. (one end open, other end closed in *tympanic membrane* or *ear drum*)
- Middle Ear – Impedance matching
  - Tympanic membrane is connected to 3 bones *malleus*, *incus* and *stapes* or *hammer, anvil* and *stirrup* (smallest bones in the body)
  - Best match is ~ 1kHz – above which the middle ear is a low pass filter
  - *Eustachian tube* – equalizes pressure between inside and outside
- Inner Ear
  - Cochlea – Transforms mechanical into electrical signals.
  - Auditory Nerve – The “output” from the ear

Cochlea

This shows the Cochlea unwrapped and laid flat



- The Basilar membrane extends the length of the Cochlea (~35mm)
- Pressure changes are introduced from the oval window connected to the middle ear
- The rigidity varies with length, causing each area to have a different resonant response (tuned to a range of frequencies)
  - High frequencies are detected near the base and low frequencies near the apex
- A Traveling wave is generated from the pressure inputs. The resonances of the Basil Membrane cause a Fourier like perception of frequency
- The Basil Membrane is covered by the organ of Corti
  - 30,000 hair cells attached to the auditory nerve
  - Bending of the inner-hairs hairs exacts the Fourier analysis
  - Firing responds to changes in detection (see figure 14.9 of [GOL00]

Fletcher Munson

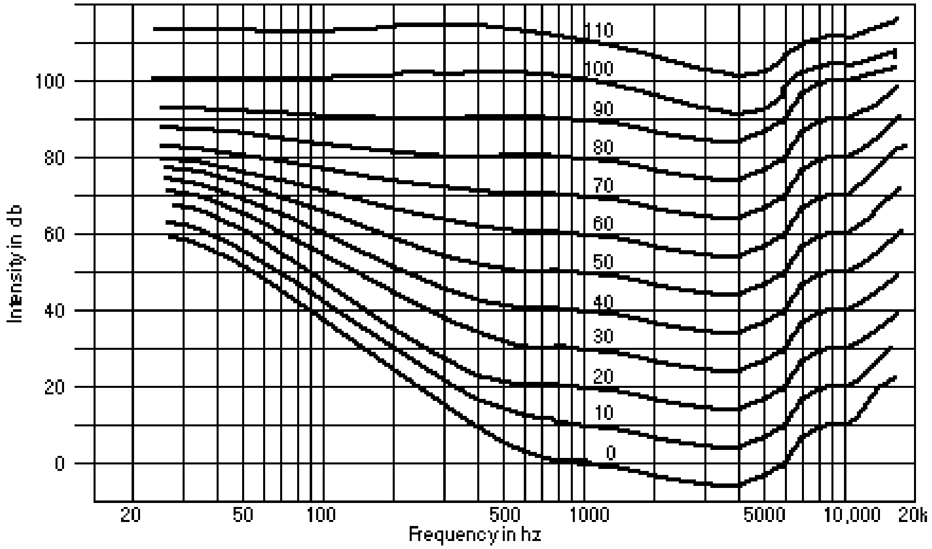


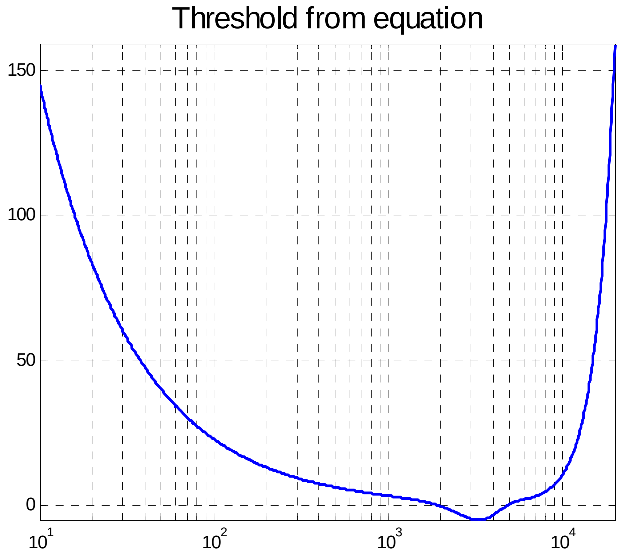
Diagram taken from http://www.webervst.com/fm.htm

- The Fletcher Munson show the results of experiments where the subjects were supposed to indicate when two tones of differing frequencies sounded “equally” as loud.
- All curves together show the frequency response of the human ear as pressure level varies.
- The lowest line indicates the threshold of hearing below this line sounds are not detectable by the human ear.

Threshold Model

- As reported in [Cav02] and [Ter82], The threshold of hearing can be modeled with the following equation
- This equation was derived from experimental data
- This allows analytical modeling of the audibility threshold for use in perceptual encoding
- Are there equations for the rest of the Fletcher Munson Curves?

$$T(f)=3.64\left(\frac{f}{1000}\right)^{-0.8}-6.5\exp\left[-0.6\left(\frac{f}{1000}-3.3\right)^{2}\right]+10^{-3}\left(\frac{f}{1000}\right)^{4}$$


Threshold from equation

Hearing Damage

- Healthy Hearing is nominally 20Hz to 20kHz
  - Generally lose high end as we grow older
  - Most of the frequencies “required” for human speech are under 4kHz
- Repeated exposure to loud sounds damages the Organs of Corti
  - Since the high frequency hairs are closest to the ear drum, they take the most abuse
    - Low frequencies can hurt the high frequency hairs
  - Long term exposure of sounds > 90dBA will damage hearing
    - OSHA guidelines -> long term = 8 hours per day for >= 10 years
    - Acceptable time halves for every 5 dB increase
      - 95dBA = 4 hours
      - 100dBA = 2 hours
    - Examples
      - Driving with the windows down = 90-95 dB
      - Rock Concerts 100 – 130dB
    - Damage usually starts around 3-4kHz and spreads up
  - Tinnitus – ringing in the ear
    - Can be temporary, recurring or permanent

Psycho Acoustics

Testing vs. Specs / Modeling

- Need Testing to determine quality
  - Amount of distortion introduced
    - “toll quality”
- Newer computers can reflect accumulated knowledge of what sounds “good”
- Some aspects can be modeled and used by a computer to simulate an ear (i.e. exploit deficiencies to compress)
- Modeling Methods
  - Physiology – Understand the physical aspects of how the human ear works
  - Psychoacoustics - Model different behaviors with simple functions

Psycho-Acoustics

- Psycho-Acoustics
  - Critical Bands
    - Relationship to Frequency
- Masking
    - Noise Masking
      - Noise of certain bandwidths can mask tones
      - Tones can mask noise in other bands
    - Tone Masking
      - One tone can mask another

Critical Bands

- Critical bands were discovered by Fletcher in the 1940s
  - Measured audibility of sinusoid with narrowband noise (centered at same frequency)
  - Increased noise bandwidth while keeping power density constant
  - Audibility threshold increased with bandwidth up to a point – after that the effect was greatly diminished
- Leads to model of the Ear as a series of Band Pass filters - *Auditory Filters*
  - Regions of Basilar Membrane are associated with a characteristic frequency
  - Response of membrane is output of Auditory filter centered at characteristic frequency.
- One belief is that the ear integrates sound within a critical band
  - When two sounds are within a critical band, the louder sound dominates.
  - When two sounds are within a critical band they sound less loud than when two tones are in different critical bands.

Critical Bands Cont

- Critical bands were measured in by Scharf [Car00]
  - Identified 25 Non-Overlapping Critical Bands
  - Bandwidths are larger at lower frequencies
  - Bandwidths are constant below 500Hz
  - Majority of bands are below 5kHz
  - The bandwidth of the critical bands matches the following equation

$$BW(f) = 25 + 75(1 + 1400 f^{2})^{0.69}$$

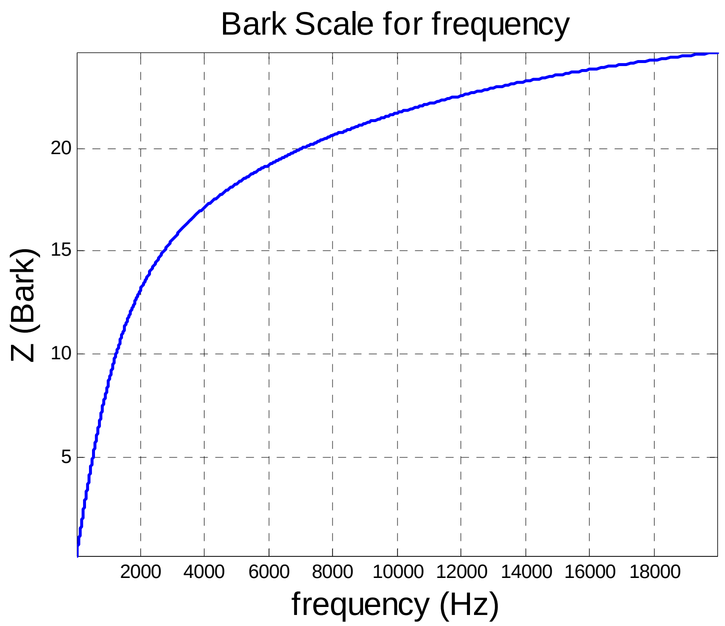
Critical Band #	Low Freq (Hz)	Center Freq (Hz)	High Freq (Hz)	Bandwidth (Hz)	Q (f/BW)
1	0	50	100	100	0.50
2	100	150	200	100	1.50
3	200	250	300	100	2.50
4	300	350	400	100	3.50
5	400	450	510	110	4.09
6	510	570	630	120	4.75
7	630	700	770	140	5.00
8	770	840	920	150	5.60
9	920	1000	1080	160	6.25
10	1080	1170	1270	190	6.16
11	1270	1370	1480	210	6.52
12	1480	1600	1720	240	6.67
13	1720	1850	2000	280	6.61
14	2000	2150	2320	320	6.72
15	2320	2500	2700	380	6.58
16	2700	2900	3150	450	6.44
17	3150	3400	3700	550	6.18
18	3700	4000	4400	700	5.71
19	4400	4800	5300	900	5.33
20	5300	5800	6400	1100	5.27
21	6400	7000	7700	1300	5.38
22	7700	8500	9500	1800	4.72
23	9500	10500	12000	2500	4.20
24	12000	13500	15500	3500	3.86
25	15500	19500			

Frequencies are from [Cav00]

Measuring Critical Bands

- Because Hz is not a good measure of perceived frequency, another scale was created to describe frequency in the perceptual domain: the Bark Scale

$$Z = 13\arctan\left(\frac{76}{100000}f\right) + 3.5\arctan\left(\frac{f}{7500}\right)^2$$



Masking

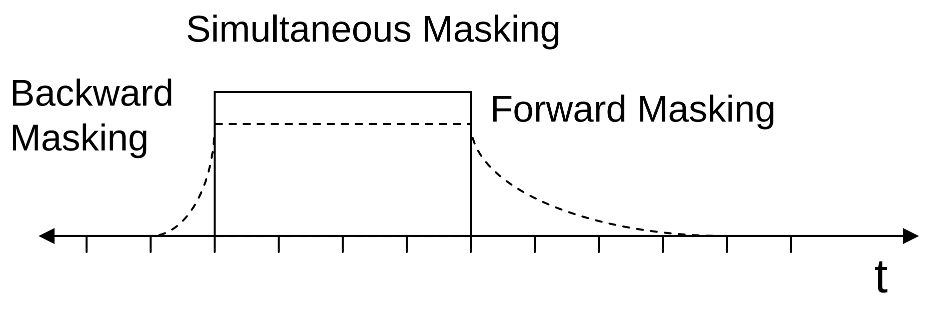
- Auditory Artifact/Feature
  - Detection of one sound “masked” (i.e. suppressed) by existence of another
  - Strong signal masks a weaker signal
    - You can hear a whisper in a quiet room, but not at a party
    - Affected by critical bands, etc
- Strong Tone is called “masker” and the weaker is the “maskee” or “target” signal
- Effects exists in time (temporal masking) and frequency domains (simultaneous masking)

Simultaneous Masking

- Two hypotheses for origins of masking [Cov00]
  - “Swamping” Adding a signal 20dB below a masking signal within a critical band adds only a small amount of signal at the detector (0.04dB in this case)
  - “Suppression” masker suppresses neural activity on the basilar membrane

Temporal Masking

- Masker and target have a temporal offset
- Masker can have effect on signals previous to onset



Dotted line is the masking threshold for a 200ms burst of tone (each tick is 50 ms)

Backward Masking

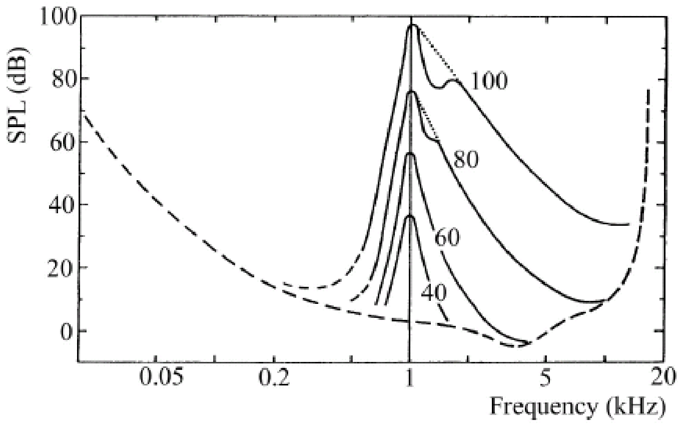
- Two hypotheses for origins of backward temporal masking [Cov00] were suggested by Moore
  - Intense signals are processed more rapidly. Processing takes over “slower” processing from previous weaker signal.
  - Results from the temporal resolution of the ear
- Only begins 20ms prior to the masker onset (~ an order of magnitude less than forward masking)

Forward Masking

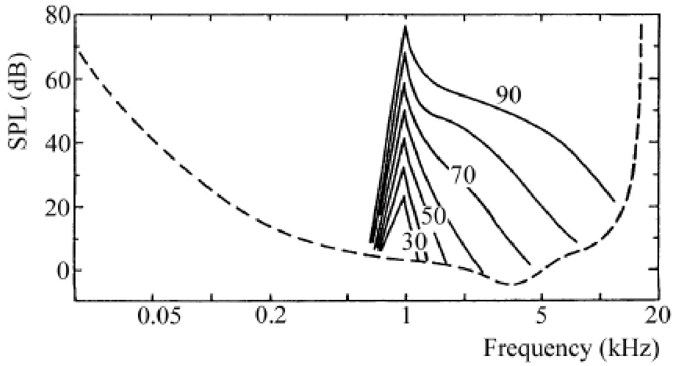
- Three factors for forward temporal masking [Cov00] were suggested by Moore
  - The Basilar membrane continues to “ring” after a loud sound. New sounds don’t create as much deflection as the end of the last sound
  - Fatigue in the auditory nerve
  - Neural activity persists at higher processing levels
- Effects last up to 200ms after masker ends
  - Masking length is largely independent of masker level (decay is always in the 100-200 ms range)
  - Amount of masking increases proportionally to duration of masker and inversely (logarithmically) proportionally to the delay between masker offset and target onset.

Masking Patterns

- Two types signals are of interest: tones and noise. This leads to four types of masking experiments
  - Noise masking noise
  - Noise masking tone
  - Tone masking noise
  - Tone masking tone
- Most experiments use a tone as the target, but audio compressions is more interested in noise as the target
- Results are varied and the debate still goes on regarding the results when the noise is the target (see [Moo98])
- All of this is of questionable relevance to speech and audio where maskers and targets move more rapidly than in masking experiments



(a) Masking patterns produced by a Bark-wide noise masker with a sinusoidal target.



(b) Masking patterns produced by a 1 kHz sinusoidal masker with a noise target.

Additively of Masking

- How do masking patterns add with multiple maskers?
  - Proposal by Theile that the most conservative estimate is to assume the maximum masking applied at any frequency
  - Green measured between 9 and 13dB additional masking when two equal maskers (i.e. they should produce the same amount of masking) are applied simultaneously
  - Lufti measured 10dB to 17dB additional masking when using multiple types of equal power maskers (two sinusoids, two noise maskers and one sinusoid and one narrowband noise masker)
    - Lufti's measurements can be fitted using the following equation

$$F\left(M_{AB\ldots K}\right) = F\left(M_A\right) + F\left(M_B\right) + \ldots + F\left(M_K\right)$$

$$F\left(M_A\right) = \left(10^{\frac{M_A}{10}}\right)^p - \left(10^{\frac{Q_T}{10}}\right)^p$$

Where $M_{A\ldots K}$ are the masking thresholds, Q_T is the threshold in quiet for a given frequency and *p* is a number determined by fitting to experimental data (usually around 0.33)

Quantization Noise

- Remember from the DSP class about Quantization Noise
  - Quantization can be modeled as white noise
  - Quantization noise can be considered independent of the signal when
    - There are sufficient quantization levels
    - The signal is sufficiently “complicated” (i.e. speech, music, etc – not a step or sine wave)
  - The SNR from quantization can be roughly determined from the following equation (where *B* is the number of bits):

$$SNR \approx 6B - 1.25\,\mathrm{dB}$$

– From this equation we can see that 16 bit CD quality has an SNR in the 95dB range.

Digital Encoding

Digital Standard Overview

- Digital Standard Overview
  - Non-Compressed
    - CD
    - SPDIF
  - Compressed
    - Waveform Encoded
      - Sample Domain Compression – μLaw, ADPCM
    - Perceptually Encoded (source agnostic)
      - MPEG
      - AAC
    - Source Encoded (based on specific source models)
      - Vocoders

Compact Disc

- Designed to hold over one hour of music (~74 minutes)
- Created by Phillips and Sony – debuted in Japan in '82 and USA in '83
- Physical Aspects
  - Disc is 120mm in diameter
  - Minimum useable diameter (radius?) is 24.8mm
  - Maximum useable diameter (radius?) is 58mm
  - Track Pitch is in the range of 1.5 – 1.7 microns
  - Linear velocity can vary from 1.2 to 1.4 m/s
- By adjusting these values it is possible to alter the play time of a CD

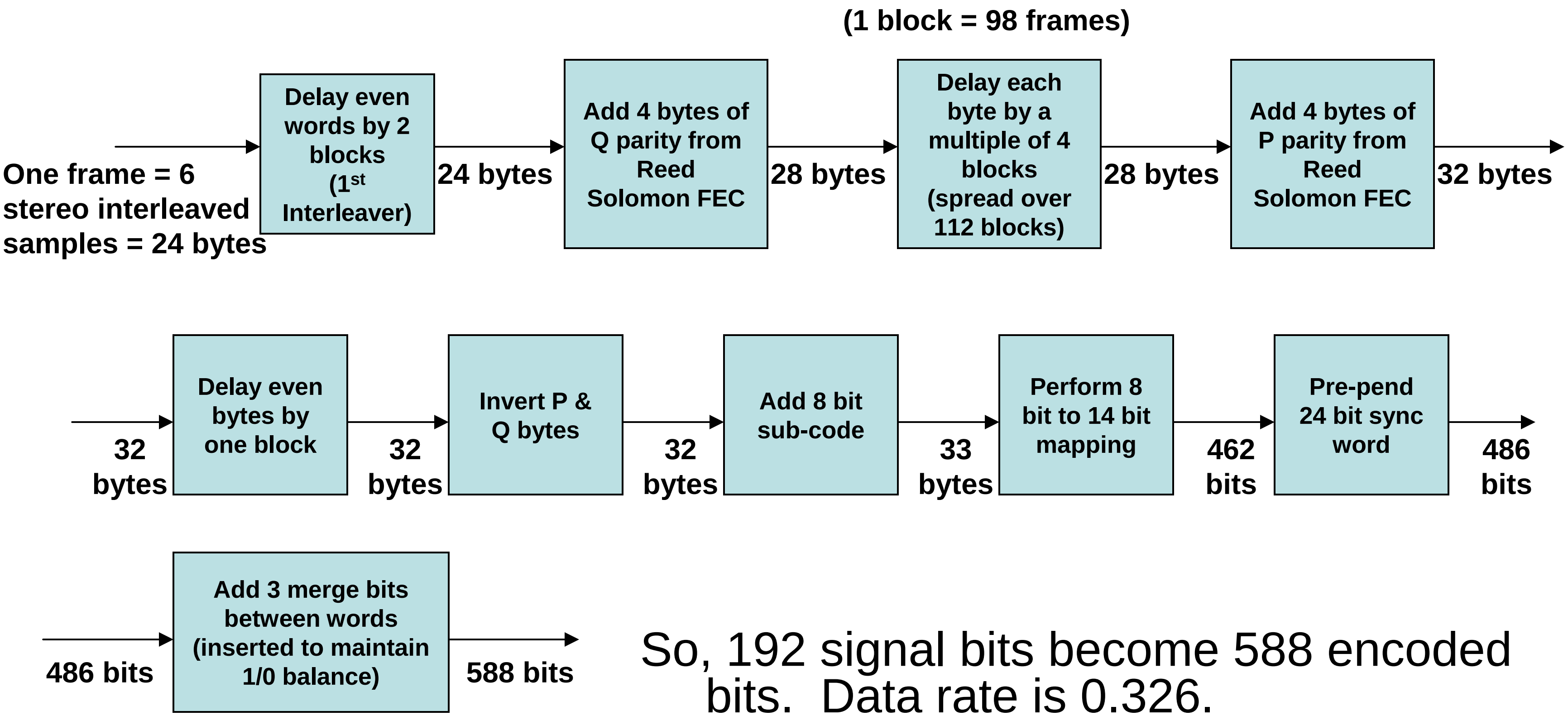
	Median Range	Min Velocity	Min Track Pitch	Min Both
Linear velocity (m/s)	1.3	1.2	1.3	1.2
Track pitch (microns)	1.6	1.6	1.5	1.5
Playing Time (mins)	69.2	75	73.8	80

http://www.disctronics.co.uk/technology/cdaudio/cdaud_intro.htm

CD encoding

- All sounds are sampled as PCM
- The raw sampling rate is 44.100kHz and all samples are 16 bit linear
- Standard CD encodes 2 independent channels (typically stereo)
- Data rate 44100*16*2 = 1.41 Mbps (good for early eighties distribution on physical media, way too high for download)

CD Error Correction



(1 block = 98 frames)

One frame = 6 stereo interleaved samples = 24 bytes

Delay even words by 2 blocks (1st Interleaver)

24 bytes

Add 4 bytes of Q parity from Reed Solomon FEC

28 bytes

Delay each byte by a multiple of 4 blocks (spread over 112 blocks)

28 bytes

Add 4 bytes of P parity from Reed Solomon FEC

32 bytes

32 bytes

Delay even bytes by one block

32 bytes

Invert P & Q bytes

32 bytes

Add 8 bit sub-code

33 bytes

Perform 8 bit to 14 bit mapping

462 bits

Pre-pend 24 bit sync word

486 bits

486 bits

Add 3 merge bits between words (inserted to maintain 1/0 balance)

588 bits

So, 192 signal bits become 588 encoded bits. Data rate is 0.326.

- See http://www.ee.washington.edu/conselec/CE/kuhn/cdmulti/95x7/iec908.htm for the nitty gritty details

SubCode Channels

- 8 Channels labeled P – W
- Each SubCode byte contains one bit of each channel.
- There are 96 bits per subchannel (2 bits are used for sync)
- Channels have special purposes
  - P indicates the start and end of each channel (P=1 is the start of the track P=0 during the track)
  - Q indicates the time codes, Table of Contents (during the lead in), track #, index and ISRC (unique identifier)
  - R—W are reserved for Text encoding

Tracks

- May have 1 – 99 tracks
- Tracks must be at least 4 seconds long and typically have 2 seconds of silence between them
- Each Track can have 2 – 99 indices within them (index 0 is the lead-in, index 1 is the main track, others are used less often)
- The “table of contents” is transmitted during the lead-in area of every track. This gives the start time for all tracks from 0—99.
  - Time code is specified in Minutes, Seconds and Frames

SPDIF

- Sony / Philips Digital Interface
- Typically uses an RCA connector
- Designed for digital connections, not digital storage
- Specs
  - One way per cable
    - 16 or 24 bit
    - 44,100kHz , 48kHz or 32kHz sampling
    - Modulated as Bi-Phase Mark Code (1s are represented by a phase transition within the bit period, 0s are represented by a lack of transition within the bit period. There is always a transition on bit boundaries

Simple Compression

Mu Law

- Telephony Network building block is the DS0 = 64kbps channel (really 56kbps given robbed bit signaling)
- 24 DS0s in a DS1 (plus one framing bit) = 1.544 Mbps

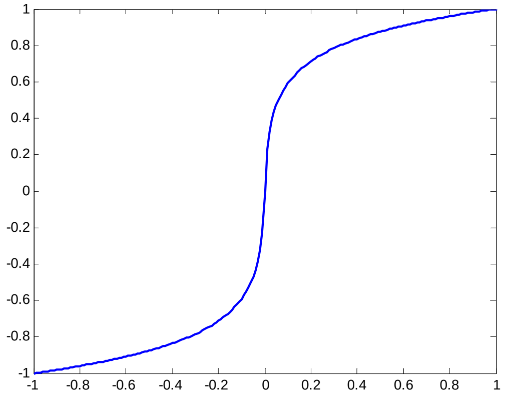
□ μ-law compression is defined in CCITT G.711

- Uses a logarithmic companding equation to let 8bit data cover the dynamic range of 14 bit data
- Lower amplitude data values are preserved better than high amplitude ones

Mu Law

$$y = \left\{ \begin{array}{l} 255 - \dfrac{127}{\ln(1+\mu)} \times \ln\left(1+\mu|x|\right) \text{for } x \geq 0 \\ 127 - \dfrac{127}{\ln(1+\mu)} \times \ln\left(1+\mu|x|\right) \text{for } x < 0 \end{array} \right\}$$

From Pan



$$y = 127 + sign(x) \times \frac{127}{\ln(1+\mu)} \times \ln\left(1+\mu|x|\right)$$

Reworked

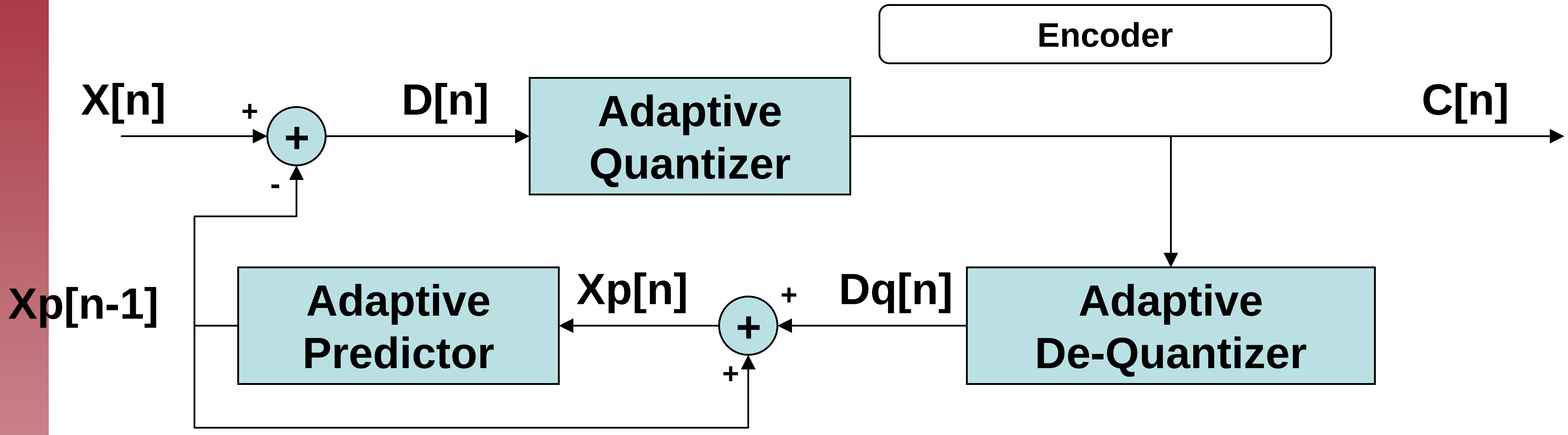
- y is the output, x is the input (normalized to +/- 1 and μ is 255

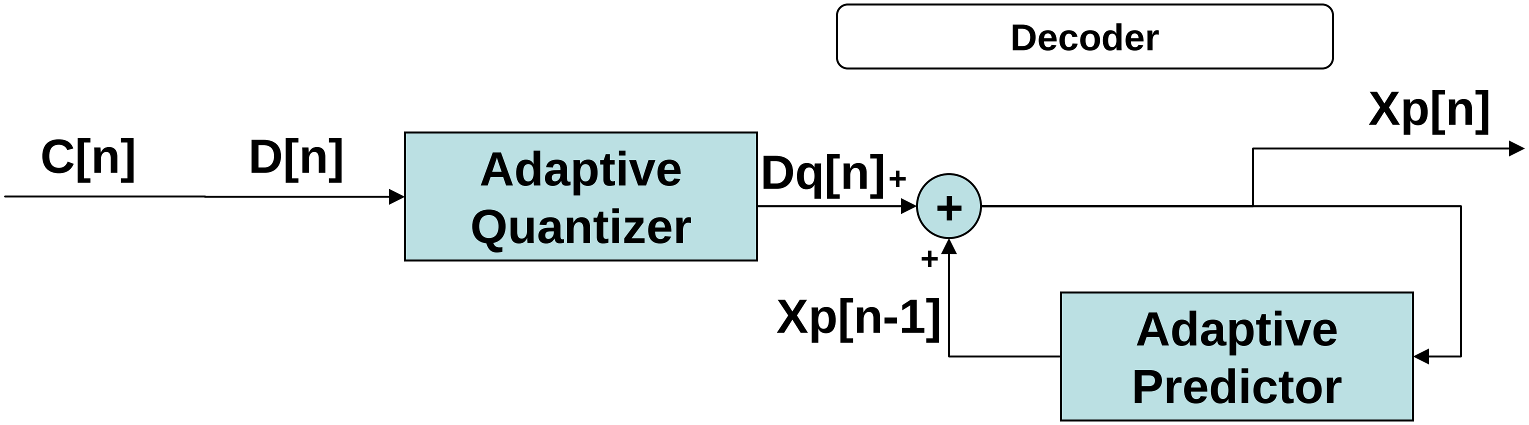
Adaptive Differential Pulse Code Modulation

- Designed to exploit sample redundancy – i.e. samples close together in time are often similar
- Samples aren't represented individually as PCM. The Difference is encoded
  - Actually encodes the difference between the actual sample and the predicted sample
  - Adaptive systems typically required "side information" to determine the range of the samples or for error recovery.
- Outputs are in "number of quantizer levels"
  - The level steps are adaptively resized during the encoding.
  - Requantizer (converts delta modulated symbols to PCM) multiplies quantized sample by step size (and possibly re-centers)
- Adaptation occurs by changing size of quantizer or predictor (or both)

From [Pan 93]

ADPCM encoder and decoder

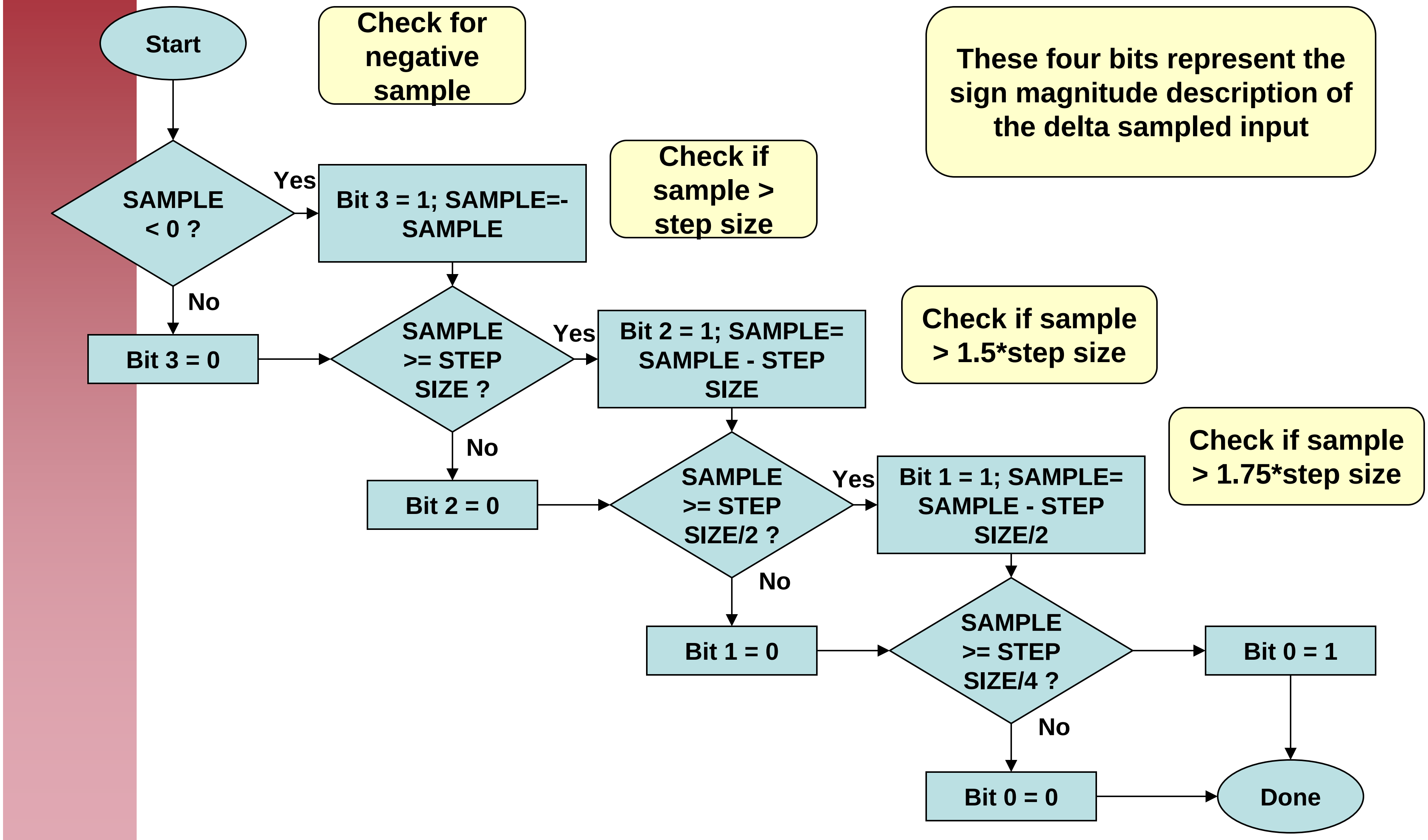




Interactive Multimedia Association ADPCM

- 4 to 1 compression in number of bits
- Alternatives to G.721 (2 to 1 – 32kbps output) and G.723 (8 to 3 – 24kbps)
- An attempt at a de-facto standard for PCs
  - Decoder had to be software only
    - 44.1kHz on a 20MHZ 386 computer
    - IMA algorithm was low complexity enough to allow for a real-time encoding in the same platform
  - Decoder was relatively simple to meet this
    - Predictor is merely a time delay. Prediction is the decoded previous sample. It is not adaptive. Therefore no side information is needed
    - Quantizer outputs four bits to indicate the magnitude of the number of quantizer levels for the input sample

IMA ADPCM Quantizaion

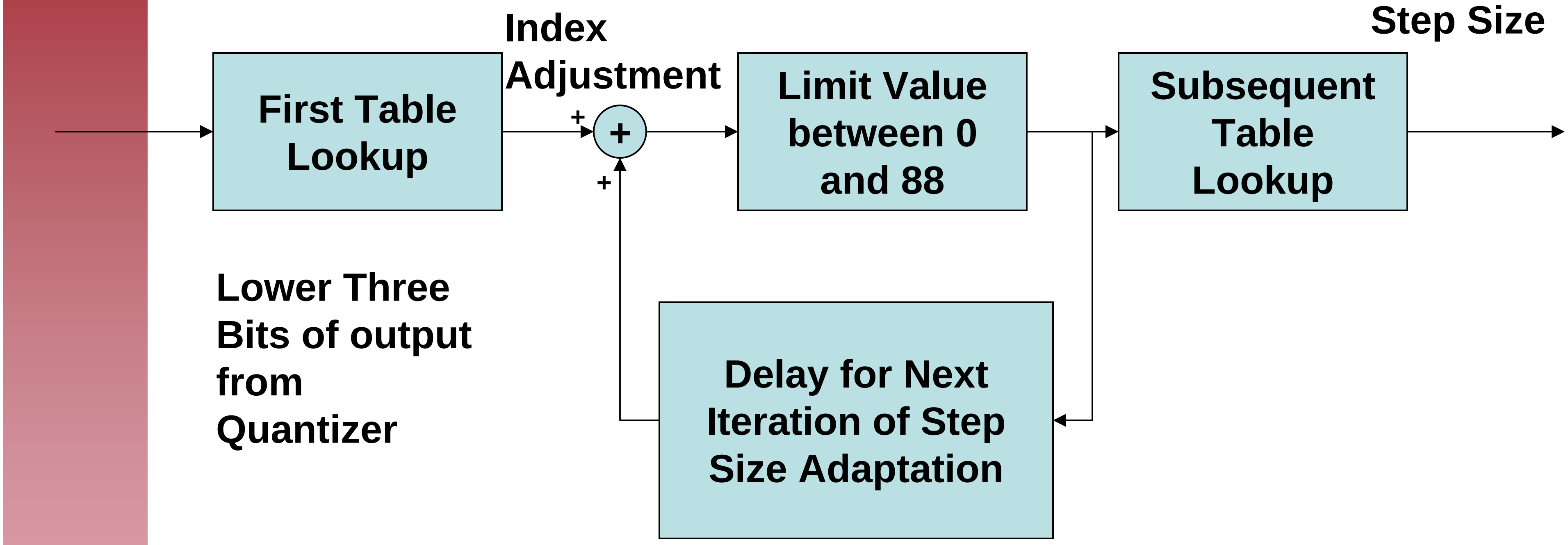

Start
Check for negative sample
These four bits represent the sign magnitude description of the delta sampled input
SAMPLE < 0 ?
Yes
Bit 3 = 1; SAMPLE=-SAMPLE
Check if sample > step size
No
Bit 3 = 0
SAMPLE >= STEP SIZE ?
Yes
Bit 2 = 1; SAMPLE= SAMPLE - STEP SIZE
Check if sample > 1.5*step size
No
Bit 2 = 0
SAMPLE >= STEP SIZE/2 ?
Yes
Bit 1 = 1; SAMPLE= SAMPLE - STEP SIZE/2
Check if sample > 1.75*step size
No
Bit 1 = 0
SAMPLE >= STEP SIZE/4 ?
Bit 0 = 1
No
Bit 0 = 0
Done

IMA ADPCM

- Adaptation is only in the quantizer block Adjusts based on:
  - Current step size
  - Output of previous input
- Table on the right shows step size adjust
  - If the Sample is < STEP SIZE, The step size will be reduced (by one)
  - If the Sample is >STEP SIZE, the step size will be increased (by a value determined by the other bits)

Three Bits Quantized Mag	Index Adjustment
000	-1
001	-1
010	-1
011	-1
100	2
101	4
110	6
111	8

IMA ADPCM Step Size Adaptation


Index Adjustment
Step Size
First Table Lookup
+
+
+
Limit Value between 0 and 88
Subsequent Table Lookup
Lower Three Bits of output from Quantizer
Delay for Next Iteration of Step Size Adaptation

- Given a starting index, it is always possible to generate the next index given the incoming data stream

Step Size Lookup Table

Index	Step Size	Index	Step Size	Index	Step Size	Index	Step Size
0	7	22	60	44	494	66	4,026
1	8	23	66	45	544	67	4,428
2	9	24	73	46	598	68	4,871
3	10	25	80	47	658	69	5,358
4	11	26	88	48	724	70	5,894
5	12	27	97	49	796	71	6,484
6	13	28	107	50	876	72	7,132
7	14	29	118	51	963	73	7,845
8	16	30	130	52	1,060	74	8,630
9	17	31	143	53	1,166	75	9,493
10	19	32	157	54	1,282	76	10,442
11	21	33	173	55	1,411	77	11,487
12	23	34	190	56	1,552	78	12,635
13	25	35	209	57	1,707	79	13,899
14	28	36	230	58	1,878	80	15,289
15	31	37	253	59	2,066	81	16,818
16	34	38	279	60	2,272	82	18,500
17	37	39	307	61	2,499	83	20,350
18	41	40	337	62	2,749	84	22,358
19	45	41	371	63	3,024	85	24,623
20	50	42	408	64	3,327	86	27,086
21	55	43	449	65	3,660	87	29,794
						88	32,767

Notice that each entry is approximately 1.1 times the previous entry sample

The Impact of Errors

- Generally Errors for this encoder don’t have a disastrous effect.
  - Not always true for systems involving prediction (i.e. decision direction)

$$Xp[n] = Xp[n-1] + STEP\_SIZE[n] \times C'[n]$$

- Where C’[n] is “one-half plus a suitable numeric conversion” of C[n]
- Ignoring limiting (below 0 & above 88), The step size is a function of the previous step size

$$STEP\_SIZE[n] = STEP\_SIZE[n-1] \times F(C[n-1])$$

- Combining these two gives an expression for the sample at time n given the same at time m

$$Xp[n] = Xp[m] + STEP\_SIZE[m] \times \sum_{i=m+1}^{n} \left\{ \prod_{j=m+1}^{i} F(C[j]) \right\} \times C'[i]$$

- Assume an error in Xp[m]. This introduces a “DC” offset into the output signal.
- It also adjusts the STEP_SIZE such that it amplifies or attenuates more than is correct.
- Clipping helps to correct when the STEP_SIZE is too large
- Any sequence of 88 samples with magnitude 3 or less will drive STEP_SIZE back to minimum value – a kind of reset. Even at 8kHz this only 11ms of sound

Psycho Acoustic Compression

MPEG History

- MPEG is short for Motion Pictures Expert Group
- MPEG-1 was the first to come out (started 1988, standardized 1992) – mainly designed for low bit rate video. First international standard for compressing high-fidelity audio.
- MPEG-2 was a flexible standard which introduced a lot of new stuff (5.1 surround signal, multiple resolutions, more sampling rates
- MPEG-3 was going to be HDTV standard, but MPEG-2 was already good enough for that so MPEG-3 was abandoned. (MP3 does not mean MPEG-3!)
- MPEG-4 – Next major standard (standardized in 1998) Focus on new functions over better compression. Extra audio coding, but also includes AAC
- MPEG-7 – a content representation standard. No compression algorithms included.

MPEG Audio

- MPEG-1 (IS 11172-3)
  - Multiple Sampling Rates (32kHz, 44.1kHz, 48kHz)
  - 4 Stereo modes (mono, dual mono, bit-sharing, and joint)
  - Multiple Rates (32kbps to 224kbps – compression factors of 2.7 to 24)
  - Audio Layers
    - Layer 1 – Simplest compression (good for over 128kbps) – DCC uses this layer
    - Layer 2 – Intermediate complexity (good for around 128kbps) – used for Digital Audio Broadcasting (DAB) and Video CD
    - Layer 3 – Most Complex – best quality and capable of 64kbps (suitable for ISDN). Most popular layer
  - Error Correction – optional CRC
  - Ancillary data can be included in the bitstream
- MPEG-2 (IS-13818-3) & (IS-13818-7)
  - Added backwards compatible 5.1 surround to MPEG-1 Audio Layers
  - Added sampling rates of 14kHz, 22.05kHz and 24kHz
  - Added new Advanced Audio Coding (AAC) standard which is not backward compatible with MPEG-1 audio layers.
  - MPEG 2.5 is a proprietary format from the Frauhofer group
- MPEG 4 (IS 14496-3)
  - Adds really low bit rate compression (down to 2kbps)
  - Adds high-quality compression at 64kbps
  - Maintains AAC for most audio coding

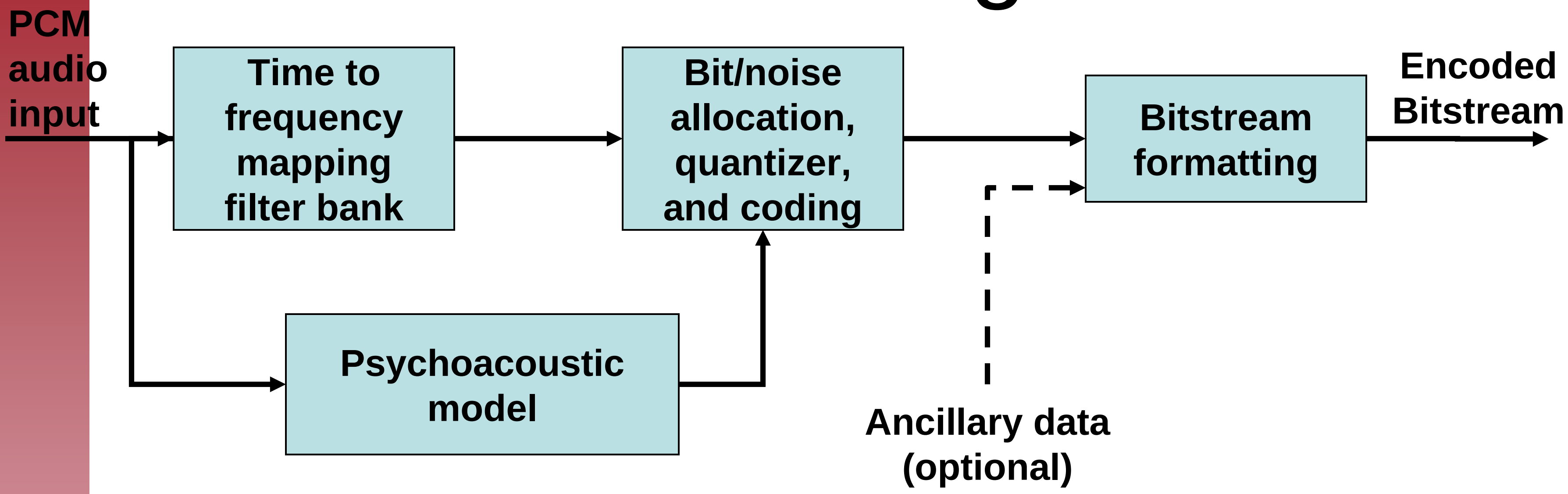
MPEG Quality and Popularity

- The compression is obviously lossy
- However, it is designed to be “perceptually lossless” or “transparent”
- MPEG performed listening tests
  - 6 to 1 compression (16 bit samples, 48kHz sampling to 256kbps)
  - Samples chosen were “difficult”
  - Expert listeners could not distinguish the difference
- MPEG is an open standard. While parts are patented, they are all licensed on “fair and reasonable terms”

Standards and Compliance

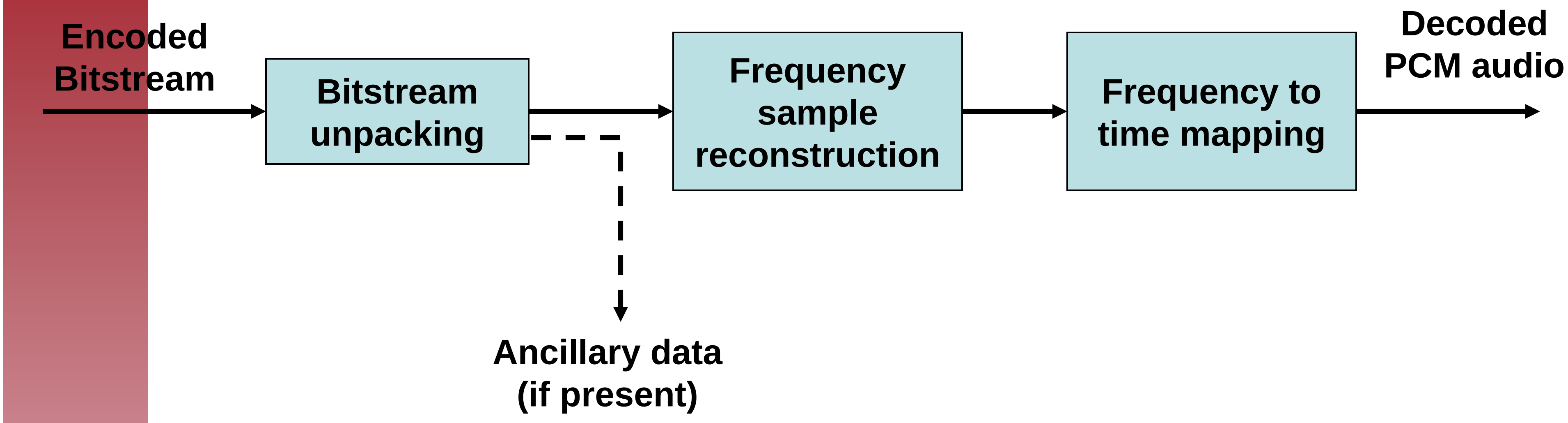
- Standards were designed to “mandate” the minimal amount of normative requirements
- Normative elements
  - Bit Stream formatting
  - Decoder Structure
    - Bit exact compliance is not required
    - Compliance defined by maximum deviation relative to a reference decoder (provided) using floating point precision.
- Informative
  - Encoding techniques are completely determined by the implementer
  - References provided (i.e. psycho-acoustic models), but not perfectly tuned (possibly by design)
  - Because of this, the number of independent implementations are relatively small.

Encoding


PCM audio input
Time to frequency mapping filter bank
Bit/noise allocation, quantizer, and coding
Bitstream formatting
Encoded Bitstream
Psychoacoustic model
Ancillary data (optional)

- Input is applied to filter bank to divide it into multiple sub-bands
  - Sub-bands are then passed to the next block for sub-band coding and quantizer level adjustment
- Input is also passed to psychoacoustic model
  - Determines the ratio of signal energy to masking threshold (for each sub-band)
  - Ratios are used by bit allocation block. Attempts to allocate bits per sub-block to minimize noise
- Quantized signal is passed to last block for bitstream formatting (includes ancillary data if present)

MPEG Decoding

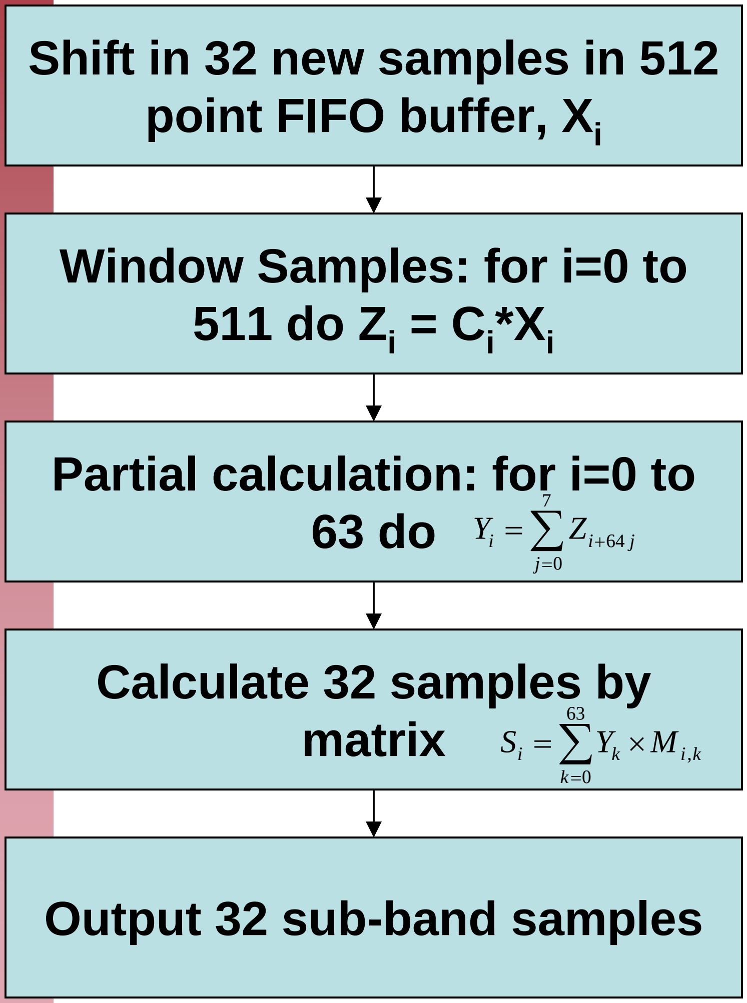


- Performs the reverse of the encoder
  - Decodes bitstream
  - Restores quantized sub-band values (puts them all on the same scale)
  - Reconstructs the PCM audio
- No analysis required, all the difficult work is done at the encoder

Polyphase Filter Bank - Overview

- Common to all Layers of MPEG compression
- Divides audio into 32 equally space bands
  - Good time resolution
  - Reasonable frequency resolution
- 3 Problems
  - Equal Width Sub-bands don’t match the Critical band model of the human ear
    - At lower frequencies, the sub-bands cover multiple critical bands
    - Cannot tune quantization enough to maximize the possible effectiveness in the noise masking
    - Band with least noise-masking dictates quantization for the whole band
  - Filter (and inverse) are not lossless. Error is always introduced, but it is kept unnoticeable by design (< 0.07dB)
  - There is significant overlap between bands. Some samples fall in two bands.

MPEG Audio Filter band computation



Shift in 32 new samples in 512 point FIFO buffer, X_i

Window Samples: for i=0 to 511 do $Z_i = C_i * X_i$

Partial calculation: for i=0 to 63 do $Y_i = \sum_{j=0}^{7} Z_{i+64j}$

Calculate 32 samples by matrix $S_i = \sum_{k=0}^{63} Y_k \times M_{i,k}$

Output 32 sub-band samples

This particular implementation is very efficient

The combined equation is as follows:

$$s_t[i] = \sum_{k=0}^{63} \sum_{j=0}^{7} M[i][k] \times \left(C[k+64j] \times x[k+64j]\right)$$

$$M[i][k] = \cos\left[\frac{(2\times i+1)\times(k-16)\times\pi}{64}\right]$$

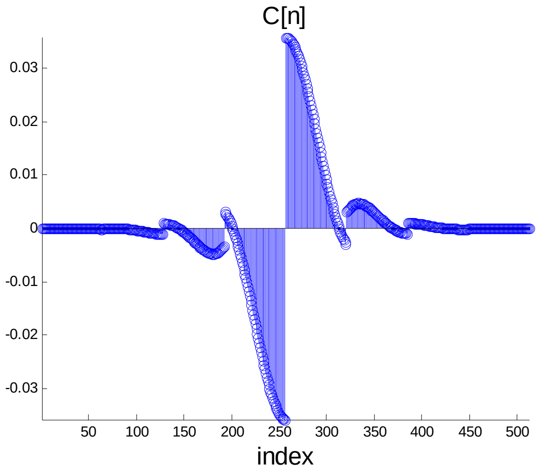
Reproduced from [Pan95]

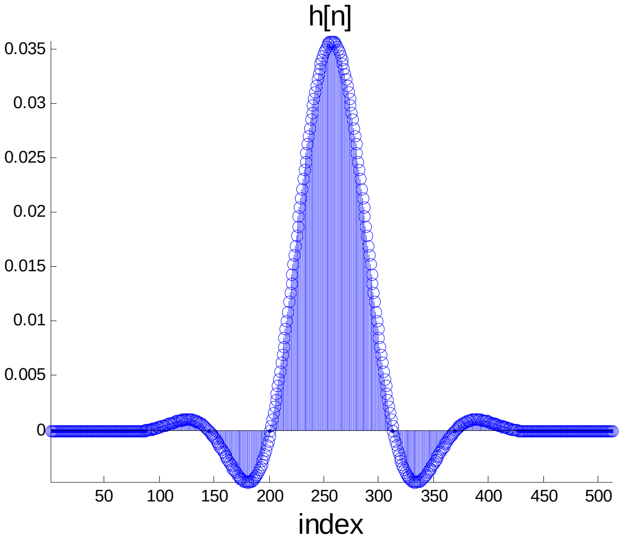
Alternate View of the filter

- This can also be rewritten as the following for easier analysis

$$s_t[i] = \sum_{n=0}^{511} x[t-n] \times H_i[n]$$

$$H_i[n] = h[n] \times \cos\left[\frac{(2\times i+1)\times(n-16)\times\pi}{64}\right]$$

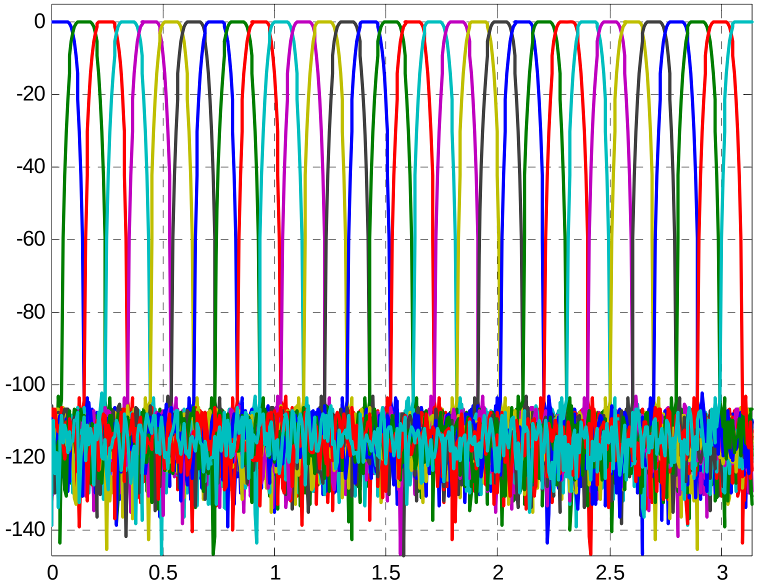




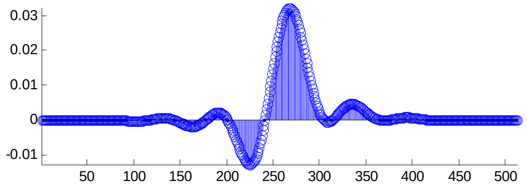
Polyphase Filter Bank - Response

- The equation on the previous slide shows that the filters are h[n] times a cos term which shifts the frequency
- This the reason for the term “polyphase” filters. Centers are at π/(64T) where T is 1/fs and each filter has BW π/(32T)
- Notice that the filters aren’t “brickwall” filters. There is a significant amount of overlap. The decimation by 32 causes aliasing, but the overlap and phase shifts cancel the aliasing at the decoder
- The overlap means some frequencies fall in two sub-bands

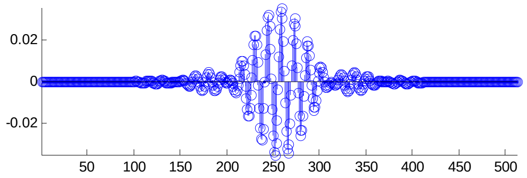
32 Sub-band filter response



First Bandpass filter response



Fifth Bandpass filter response



Psychoacoustics and Noise

- MPEG standards exploit the psychoacoustic results discussed earlier in the presentation
- Basic idea is the divide the audio spectrum into frequency sub-bands that are intended to represent critical bands (for a evaluation of this method see Figure 7 and accompanying text from [Pan93])
- These sub-bands are quantized with just the right levels to make the noise in the sub-band imperceptible

MPEG Psycho Acoustic Models

- Designed to model the masking phenomenon of the human ear
- Analyzes signal and computes masking thresholds as a function of frequency (using frequency, sub-band and loudness)
- Output of psycho-acoustic model is the control input to the quantizer block. This determines the quantization level (and therefore the added noise) for each sub-band
- Two models are available from the standard – both work for any of the layers
  - Model 1 is less complex and therefore less accurate
  - Model 2 is more complex and includes special modifications to work with Layer 3
- Lots of flexibility in the model. It all depends on the target bit rate
  - Higher bit rates require less compression and therefore less psycho acoustic analysis

MPEG psycho-acoustic overview

- Time align audio data
- Convert audio to frequency domain representation
- Analyze frequency components with respect to Critical Bands
- Separate Tonal vs. non-tonal components
- Apply a spreading function
- Set a lower bound for the threshold values
- Find the masking value for each sub-band
- Calculate the signal to mask ratio

Time-alignment of Audio

- One psycho-acoustic evaluation per frame (Frame size = 384 samples for layer 1 frame, 1152 samples for Layers 2 & 3)
- Must align the data sent through the filter bank (with it's delay) and offset to center data in the middle of psychoacoustic window (Layer one window is 512 samples)
- Example – Layer 1
  - Delay through filter is 256 samples
  - Need to add 64 samples to center 384 sample frame within 512 sample analysis window.
  - Combined this means a 320 sample delay to align the psychoacoustic model with the filtered data

Convert Audio to frequency domain

- Must be a separate conversion from filter bank – finer resolution required for analysis.
- Both models us a fourier transform with Hann weighting (reduces edge effects of transform window)
- Window sizes vary by model and layer
  - Model 1
    - 512 sample analysis window for Layer 1 (> 384 frame samples)
    - 1024 sample window for Layers 2 & 3 (<1152 frame samples)
  - Model 2
    - 1024 sample window for all layers
      - Layer 1 centers 384 samples in center of window
      - Layers 2 and 3 do two analysis windows
        - » First one centers the first 576 samples of the 1152 sample frame
        - » Second one centers the next 576 samples of the 1152 sample frame
      - Signal to mask ratios are combined using the higher of the two ratios – selects the lower of the two thresholds
- Each frequency is a “perceptual quanta” – each frequency considered separately

Identify tonal vs. non-tonal

- Both models need to determine tonal vs. non-tonal because they have different masking effects
- Model 1
  - Identifies local peaks in the power spectrum
  - Nontonal noise is summed into one nontonal component per critical band with the frequency value being the geometrical mean of the critical band
- Model 2
  - Doesn’t actually separate tonal vs. non-tonal
  - Creates a tonality index for each frequency (a measure of the tonality of each frequency)
  - Masking values are an interpolation of tone and noise masking thresholds based on the index.
  - Tonality is based on predictability – previous two analysis windows are used to predict the current sample (via linear extrapolation). Higher tonality indices indicate a high degree of predictability and more tonality (and less noise)

Spreading and combining

- Given that maskers impact their complete critical band, the thresholds are determined by a masking function which is empirically derived (Model 1) or a spread function (Model 2)
- Determine the lower bound for all spectral components – use the threshold of hearing.
- Determine the masking threshold for each sub-band.
  - Must combine the masking ability of all maskers in a sub-band
    - Model 1
      - selects the minimum masking threshold in the band
      - This works well in lower sub-bands where multiple critical bands are within one filter
      - Less accurate for high frequencies where critical bands span sub-bands
      - In-accuracies arise because model 1 treats all non-tonal energy as single component with a center frequency
    - Model 2
      - selects the minimum but only when sub-band is wider than critical band
      - Uses average of thresholds when sub-band is narrow relative to critical band
      - Has same accuracy for high and low bands
- Calculate the Signal to Mask ratio
  - Ratio between signal energy and minimum masking threshold for the sub-band.
  - This is the data passed to the quantizer block

Example from Pan95

- The signal is presented in Figure 7
  - Strong 11,250Hz tone
  - Low Pass noise
- Consider model 2
  - Figure 8a shows audio in perceptual domain
    - 63 1/3 critical band partitions
    - Spreading function is applied to perceptual data
    - About the data
      - The low frequency noise is stretched out
      - The high frequency signal appears closer to the upper end of the plot
  - Figure 8b shows tonality index.
    - Notice how the spreading function makes the tonality indices higher in the region around the tone
  - Figure 9 shows the spreading functions. They all appear similar

Example from Pan95 continued

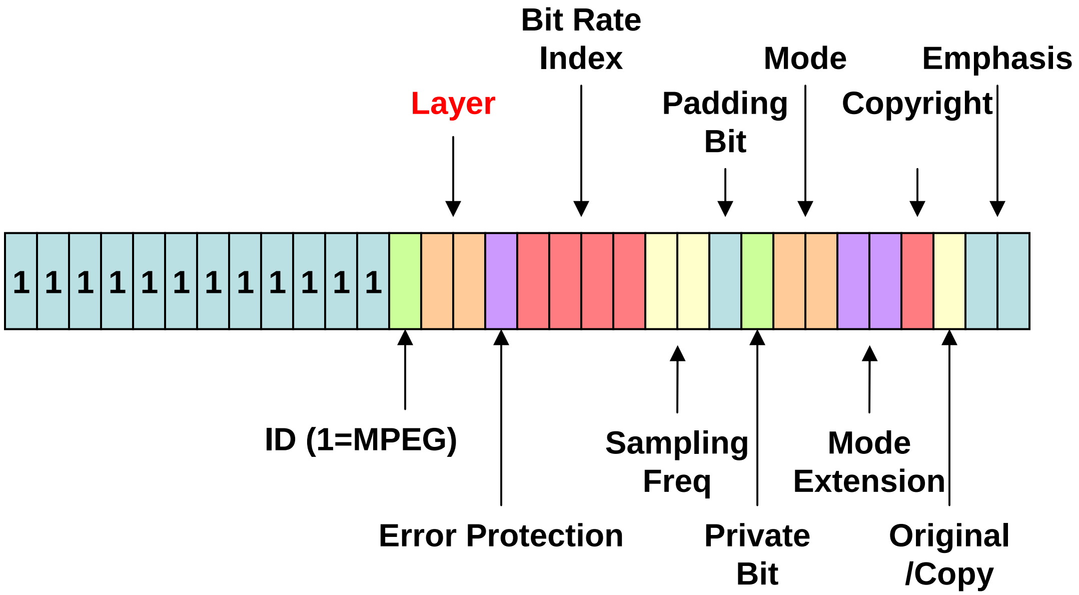
- Consider model 2 continued
  - Figure 10a shows plot of masking effects
    - One line shows masking based on spread samples
    - Other line combines this with the threshold of hearing.
      - This has a significant impact on masking effect at very high frequencies
      - Notice the masking within the critical band near the sinusoid
    - Notice that this is displayed in the frequency domain (not perceptual)
  - Figure 10b shows Signal to Mask ratios (x axis is filter indices)
  - Figure 10c shows the resultant frequency response after encoding the signal
    - This example used 768 to 64 kbps compression (12x)
    - Not all quantization noise was masked

Example from Pan95 continued

- Consider model 1
  - – Figure 11a shows the tonal versus non-talk detection
  - – Figure 11b shows results of decimation
    - Removes components below the quiet threshold
    - Removes weaker tonal signals within 1/2 critical band of strong components
    - These results are used to set a masking threshold for the 32 sub-bands
  - – Figure 11c shows masking thresholds
    - Minimum global masking threshold is used
  - – Figure 12a shows signal to mask ratios and 12b shows frequency response of encoded signal

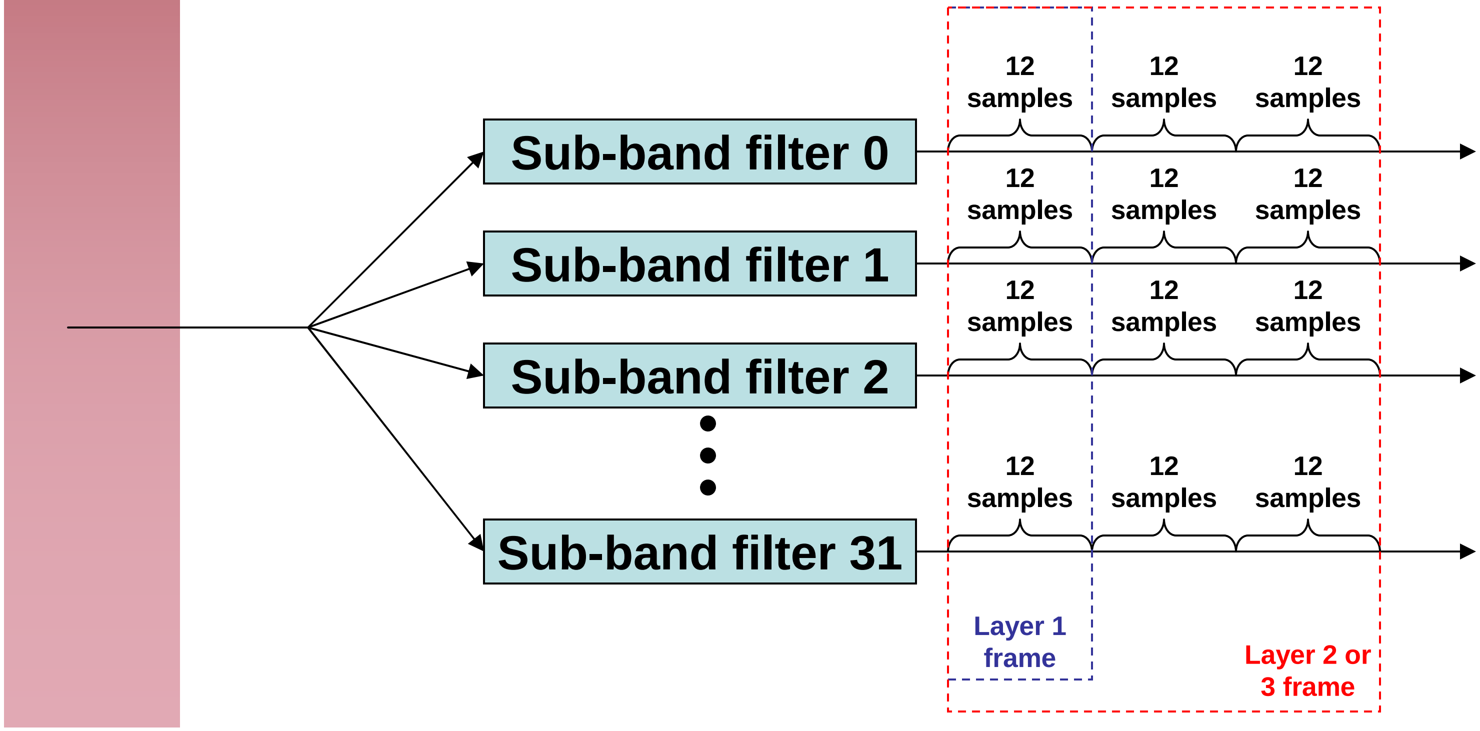
Frame Headers

- All three layers include frame headers for synchronization



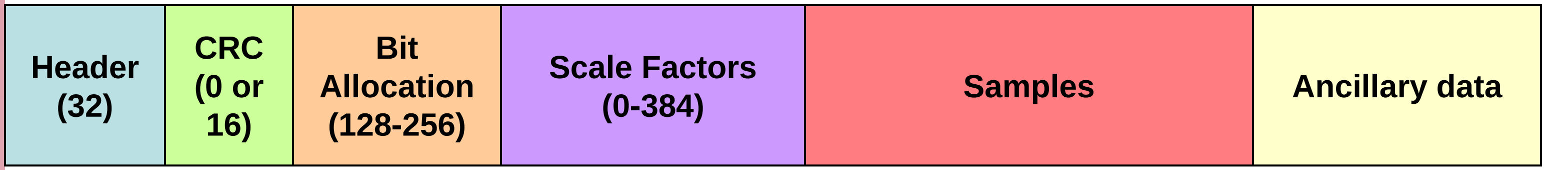
Frame Grouping

- Layer 1 encodes 12 samples from each of the 32 sub-bands
  - Each frame is 384 samples
- Layer 2 encodes 3 times as many bits
  - Each frame is 1152 samples



Layer 1 framing

- In addition to data, the Layer 1 frame contains
  - Header
  - Optional error detection
  - Possibly ancillary data
- Bit Allocation is used to inform the decoder of the number of samples (0—15 bits per subband)
  - Each group of 12 samples gets a bit allocation
- Scale factor is used to scale samples to the full range of the quantizer.
  - Scale factors only assigned if bit allocation > 0
  - Decoder uses scale factor to re-adjust level before combining
  - Scale factors chosen to have 120dB range

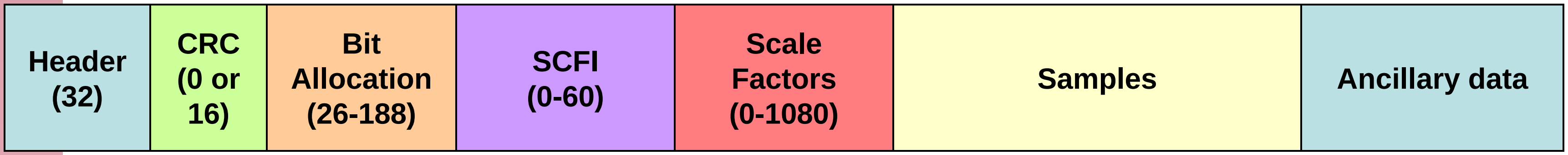


Header (32)	**CRC (0 or 16)**	**Bit Allocation (128-256)**	**Scale Factors (0-384)**	**Samples**	**Ancillary data**

Numbers in parenthesis represent the number of bits used

Layer 2 framing

- Enhancement of Layer 1
- Groups three times as many samples (1152 vs. 384)
- Restrictions on bit allocations for middle and high sub-bands
- Bit allocation, scale factors and quantized samples are coded differently to save space.
- Three groups of 12 samples are coded together
  - The set of three groups is allocated one bit allocation
  - The set can be allocated up to three scale factors (when necessary to avoid distortion)
    - Scale factors are used of in 2 cases
      - Values of the scale factors for individual groups are similar
      - When the added noise falls below the expected threshold
  - SCFSI (scale factor selection information) field is used to relay scale factor sharing patterns to the decoder.
- One more compression for the case of 3,5 or 9 levels of sub-band quantization

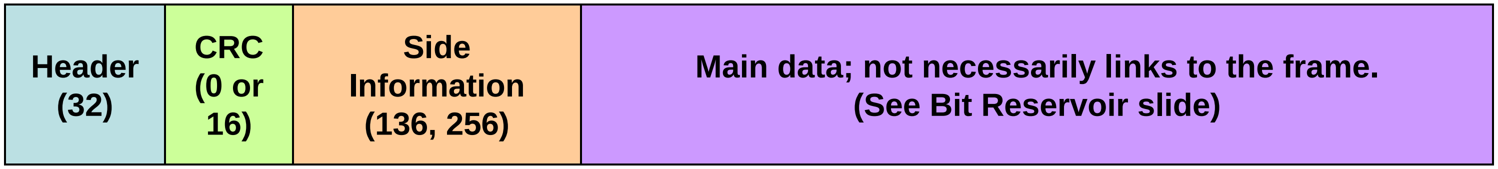


Header (32)	CRC (0 or 16)	Bit Allocation (26-188)	SCFI (0-60)	Scale Factors (0-1080)	Samples	Ancillary data

Numbers in parenthesis represent the number of bits used

Layer 3 framing

- Most complex layer – derived from ASPEC (audio spectral perceptual entropy coding) and OCF (optimal coding the frequency domain)
- Includes compensation for filter bank issues by using a modified DCT (discrete cosine transform) – a lossless transform
  - Subdivides spectrum further
  - Aliasing from poly-phase filter can be cancelled. Decoder must actually re-introduce this
  - Inverse DCT is performed at the decoder
- Supports data rates of 8 kbps to 320kbps
  - Data rate can switch from frame to frame (not supported by all decoders)



Header (32)

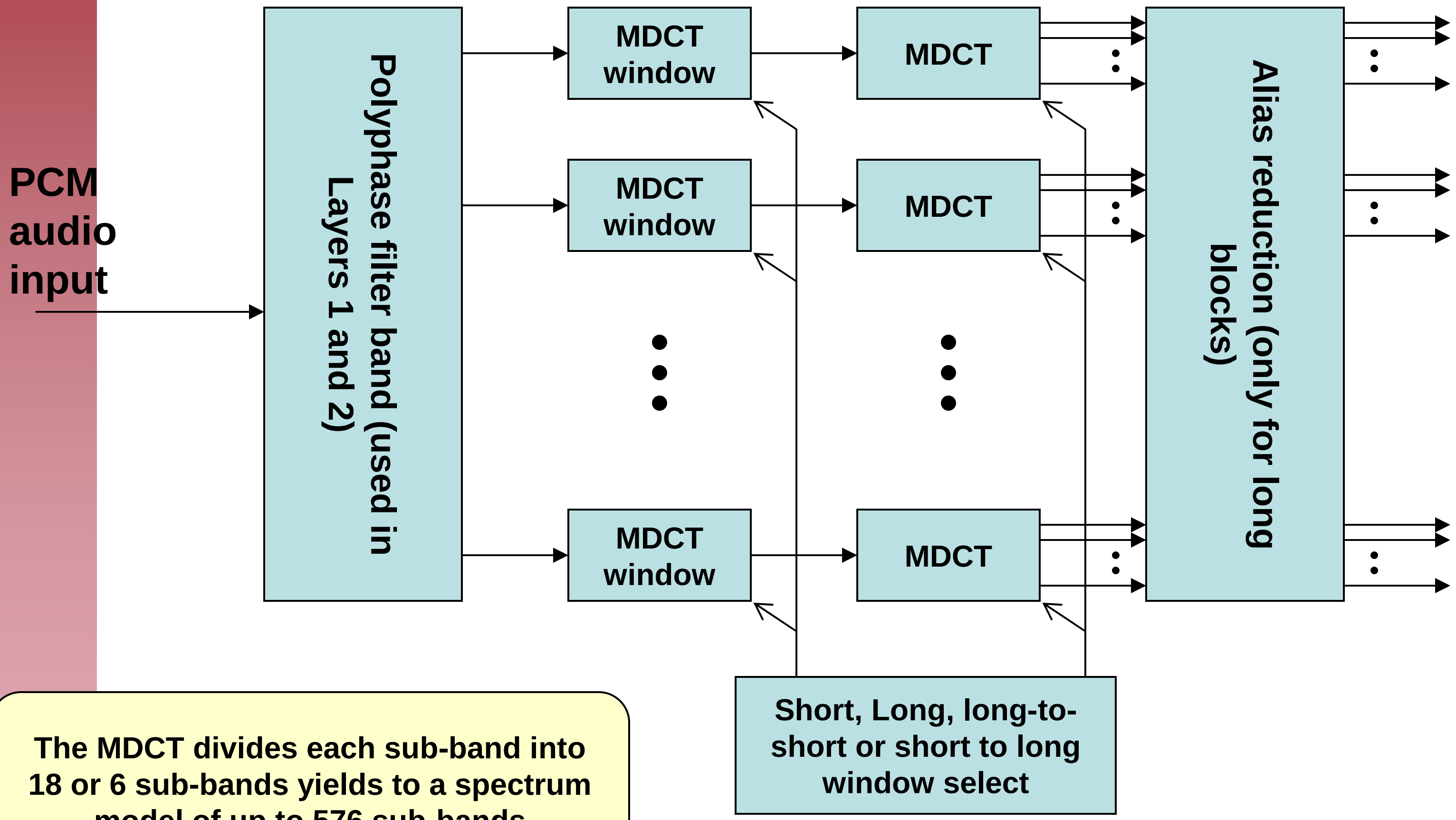
CRC (0 or 16)

Side Information (136, 256)

Main data; not necessarily links to the frame. (See Bit Reservoir slide)

Numbers in parenthesis represent the number of bits used

Block Diagram of Layer 3 enhanced sub-band coding


PCM audio input
Polyphase filter band (used in Layers 1 and 2)
MDCT window
MDCT window
MDCT window
MDCT
MDCT
MDCT
Alias reduction (only for long blocks)
Short, Long, long-to-short or short to long window select

The MDCT divides each sub-band into 18 or 6 sub-bands yields to a spectrum model of up to 576 sub-bands

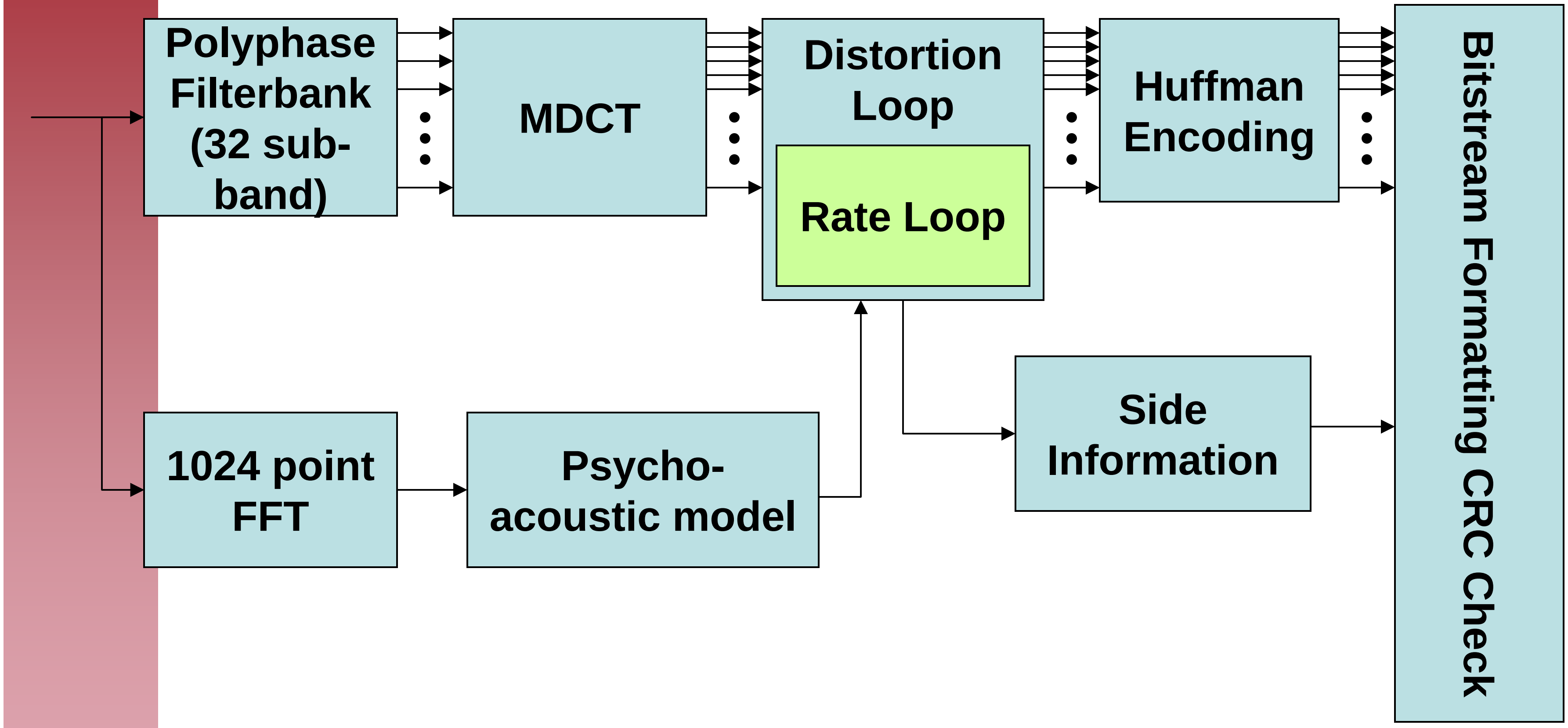
Overlapping DCT windows

- See Figure 17 from [Pan95]
- Two MDCT block lengths (long=18, short=6)
  - 50% overlap between transform windows
  - Window size is 36 or 12 accordingly
    - Long window is intended for stationary audio signals
    - Short window is for transients
  - # of MDCT samples is fixed
    - Long block is 3 times the length of a short block
  - Each frame can encoded in one of three ways
    - All short blocks
    - All Long blocks
    - Mixed Mode – long blocks for lowest 2 bands, short for the rest
  - Special long-to-short and short-to-long transition blocks exist to change the time scale (36 samples long)
  - Remember basic trade-offs:
    - Longer windows have better frequency resolution / worse time resolution
    - Windows are 12 to 36 times the normal (layer 1 or 2) time window. Time resolution is affected accordingly
    - The result of this is usually pre-echo. Masking threshold for sounds early in the block are less than expected based on samples later in the block.
      - Modifications to the psychoacoustic model pre-echo and compensate.
      - Bits can be borrowed from the Bit Reservoir to increase quantization levels
      - Encoder can use shorter time window

Other Layer 3 enhancements

- Alias Reduction
  - Removes artifacts arising from overlapping bands in spectrum
- Non-uniform quantization
  - Raises input to the 3/4 power. This provides a more consistent level to the quantizer (like mu-law)
  - Decoder raises received signals to the 4/3 power.
- Scale-factor bands
  - Not locked to filter sub-bands
  - Cover several MDCT coefficients
  - Have approximate critical band widths
  - Attempts to “color” quantization noise to better match the masking threshold
  - Adjusted as part of the noise-allocation process
- Entropy coding / Huffman codes
- Bit reservoir

Complete MP3 Diagram



Huffman coding

- Lossless Huffman coding added to encode quantized samples
- After coding there are 576 samples (32 sub-bands * 18 MDCT coefficients per sub-band). These are encoded in a specific order
  - Encoding with increasing frequency except for short MDCT block mode
  - Short MDCT blocks are coded by frequency and then by window (there are 3 windows per frequency)
  - Advantage is large values fall at lower frequencies and long runs of low values (zero or non-zero) are in the higher frequencies

Huffman coding continued

- Coefficients are separated into three regions
  - Each region is encoded with a different set of Huffman tables
  - Start at the highest frequency
    - Detect the continuous run of all-zero values. This is one region.
    - This region is not coded as the size can be determined from the size of the other regions.
    - Must contain an even number of zeros to be compatible with other regions
  - Second region is called “count1”
    - Continuous run of values made of only -1,0 or1.
    - Must be four values at a time
  - Third region is for “big values”
    - Covers the remaining values
    - Huffman tables code the values in pairs
    - Are is sub-divided into 3 sub-regions with separate tables
      - Helps error propagation

Bit Reservoir

- Designed to better varying requirements for bits.
- Frames are 1152 samples. But encoded frames aren’t fixed length
  - When less bits than the frame size are required, the extra bits are added to a “reservoir”
  - When more bits are required the encoder borrows from the reservoir
- Bit-stream includes 9 bits pointer “main_data_begin” in the side information
  - Points to the start of the audio data for that frame
  - Offsets for header and side information are not included since they are fixed size.
  - Maximum variation is 29bytes
  - Bit resevoir is also limited by the maximum code buffer of 7680 bits (see [Pan95] for example)

Bit Allocation

- Bit allocation determines the number of code bits used for each sub-band
  - Determined by the psycho-acoustic model
  - Layer 1 & 2 start with Mask to Noise ratio
    - MNR=SNR-SMR (all in dB)
    - Tables specify number of bits that map to specific SNRs
  - Search for lowest mask-to-noise ratio – code bits are allocated to this band first
  - Bands are assigned bits in a loop
    - As the allocation changes, the encoder determines the new estimated SNR and recomputes the MNR
    - This is iterative and repeats until no more bits can be added

Layer 3 Bit Allocation

- Layer 3 uses noise allocation
  - Encoder is iterative
    - It varies the quantizers
    - Quantizes the spectral values
    - Computes the size of the data after Huffman compression
    - Calculates the new noise level
    - Check for excessive distortion
      - If there is too much, increase the values in the scale factor bands (effectively reducing the quantizer step size)
  - Process terminates after the following conditions
    - All Scale-factor bands have lower levels than allowed
    - Next iteration would cause a band to exceed maximum scaling
    - Next iteration would cause all bands to be amplified
    - A time limit is exceeded (in real-time encoders usually)
- Bit rate determination
  - Bit rate is determined by the user and not by the encoder (for all encoders)
  - Bit rate switching should be supported on a frame by frame basis for MPEG audio Layer 3
  - Bit reservoir also helps to allow the frames to vary in rate

Inner and Outer Loops

Outer (Noise Control) Loop

- Apply Scale Factors to each scale factor band
- Adjust scale factors such that the quantization noise is below the mask for that band
  - This increase the number of quantization steps (and a higher bit rate
  - Inner loop needs to be re-run
- Run until the noise is below the threshold for all scale factor bands (critical bands)

Inner (Rate) Loop

- Encode data with Huffman coder
- If code exceeds available bits
  - Adjust global gain – yields larger quantization step & small quantized values
- Else
  - Exit loop

Inner loop will always converge. Combination will not always. Noise Control can force quantization steps smaller and Rate can force the larger. Stopping conditions exist to prevent an infinite loop

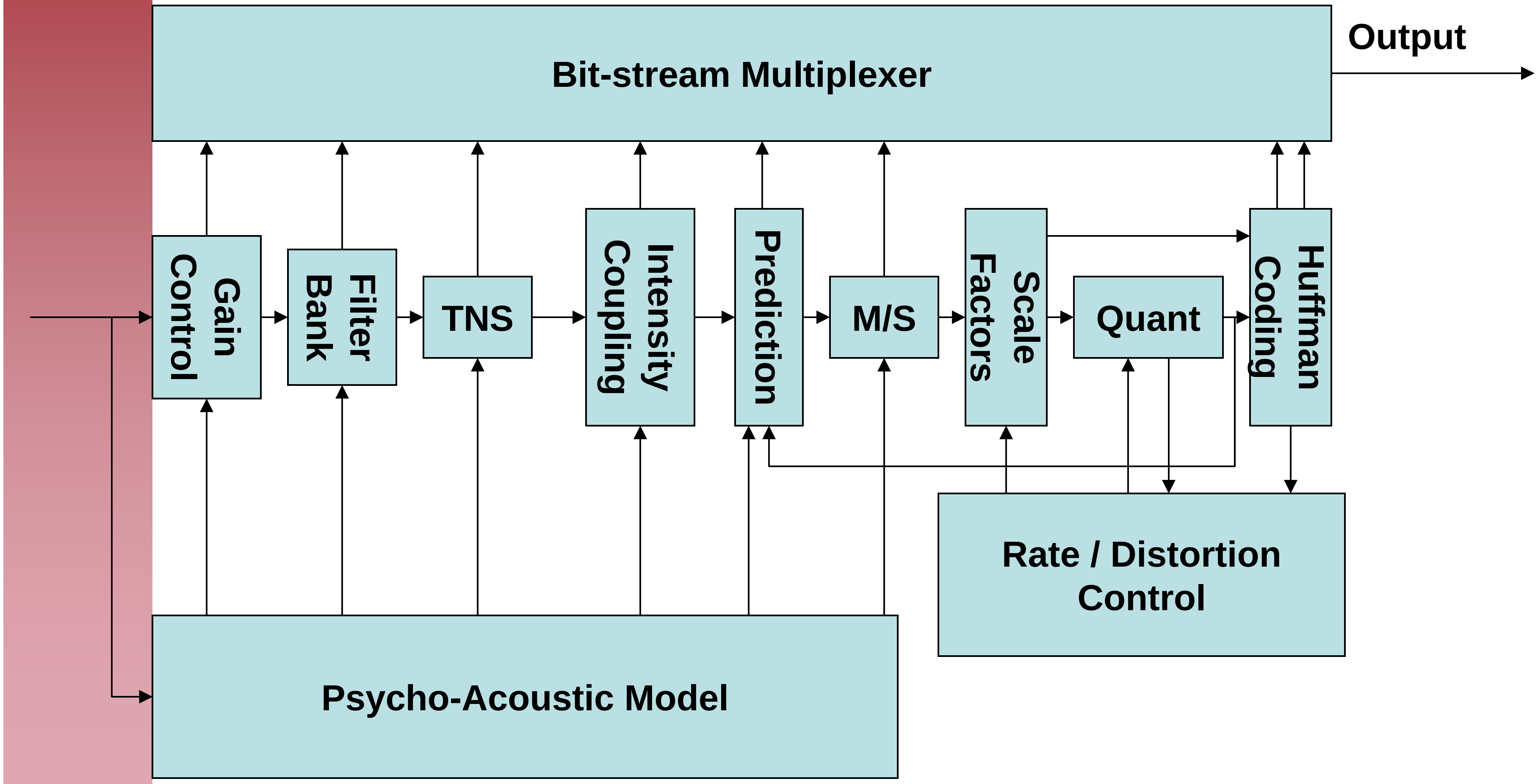
Stereo redundancy

- There two types of stereo coding (dual-mono doesn’t count)
  - Intensity Stereo coding (all layers)
  - Middle-side coding (layer 3)
- Exploits further psycho-acoustics
  - Above 2kHz and within a critical band human perception of stereo is based on temporal envelope
- Intensity Mode encodes some sub-bands with a single summed signal
  - Single signal is passed
  - Individual Scale factors are used for right and left
  - Spectral shape is the same for both sides, but amplitude varies
- Similar to FM
  - Sum of signals is middle
  - Difference of signals is side
  - Sum and Difference only used in certain sub-bands
  - Tuned thresholds compress the side signal more than the middle (or separate left and right)

The Next Generation - AAC

- Advanced Audio Coding (AAC) was introduced as a non-backward compatible audio compression algorithm in MPEG-2.
  - The structure is completely different yet uses many of the key ingredients from MPEG-1 audio layers
- Expected to be the successor of MPEG-1 Layer 3
- Includes more detailed Digital Rights Management
  - MPEG-1 audio layer only use a couple of bits of copyright.
  - AAC combines with watermarking from Secure Digital Music Initiative (SDMI)
- MPEG 2.5 is a proprietary extension of MPEG-2 from the Fraunhofer Institute designed to work at even lower bit rates

AAC Block Diagram



Comparison of AAC to MP3

- At a high level, both contain many similar elements
  - “Better” resolution filter bank (compared to MPEG Layer 1 and MPEG Layer 2)
  - Non-uniform quantization
  - Huffman Coding
  - Iterative Loops
- But there are substantial improvements in the details of each of this blocks…

AAC vs. MP3 – Coding efficiency

- Even Better coding efficiency
  - Number of frequency bins is 1024 over 576 with MP3
- Optional Prediction algorithm
  - Backward prediction yields better coding efficiency with pure tones (and tone-like signals)
- Better Stereo Coding
  - Middle/Side coding is more flexible
- Improved Huffman Coding
  - Coded by quadruples of frequency lines (as compared to doubles)
  - Code tables and code partition assignments are more flexible.

AAC vs. MP3 – Enhanced quality

- Enhanced Block Switching
  - AAC uses a single MDCT filter bank
  - Two windows are specified for the MDCT. Sizes are changed to prevent Pre-Echo
    - Long = 2048 samples
    - Short = 256 samples
- Temporal Noise Shaping
  - Performed in the Time domain using frequency domain prediction
  - Used to model forward masking phenomenon in the ear. Spreads masking in time (where previously they were only spread in frequency)
- Results
  - Same quality as MP3 with 70% of the bit rate

AAC transport formats

- More flexible frame structure than MPEG-1 Layers 1—3
  - MPEG-1 use fixed size frames with a header block for sync and coding parameters
    - Bit rate is adaptable on a frame by frame basis for MP3
    - Each frame is independent and the decoder can start anywhere within the stream
- AAC has two standards for transport of audio
  - Audio Data Interchange Format (ADIF)
    - All decoder required data in a header before the audio
    - Does not allow decoder to start in the middle of the stream
  - Audio Data Transport Stream
    - Very similar to MPEG 1/2 header format
    - Signaled as “Layer 4”
    - Frame length is variable
      - Data is bounded by occurrences of the syncword
    - Allows decoding mid-stream

Some Modifications

- As described in [Dim05], it is required that some modifications are made to the standard encoder to arrive at a more musically pleasing algorithm.
  - Reference software is made to simple to be understandable
  - More research is required to make improvements
  - Reference software co-written by several members from different organizations. Why would they want a freely available “good” implementation? They probably wouldn’t …

Dimkovic’s Modifications

- Psycho Acoustic improvements
  - Baseline model uses algorithm similar to Model 2 from MPEG-1
  - Input Filter Bank modifications
    - Use MDCT instead of FFT for the input to psychoacoustic model
    - Can use Complex MCDCT to get complex data for psychoacoustic model and use real coefficients for the encoding chain.
  - Threshold of Hearing modifications
    - Model uses possibly inaccurate values that are determined for a given loudness.
    - Their model uses a formula which models the threshold for multiple loudness levels
    - Reduced High Frequency artifacts
  - Hard coded psychoacoustic parameters modification
    - Keeps the average distortion the same across frequency bands.

Dimkovic's Modifications (continued)

- Psycho Acoustic improvements (continued)
  - Tonality Detection modifications
    - Difference of masker type can yield 20dB difference in masking ability
    - Used "intra-frame" tonality measurements
      - Spectral Flatness Measure
      - Peak Detection of tones
      - Special Improved Human Speech Encoding model especially for human speech

Dimkovic's Modifications (continued)

- Psycho Acoustic improvements (continued)
  - – Spreading function modifications
    - ISO model assumes constant loudness
    - Masking ability will change with loudness

$$S_l = 27 dB / Bark$$

$$S_u = \left[22 + \min\left(\frac{230}{f}, 10 \right) - 0.2 \cdot L \right]$$

S_l is the lower part (frequencies below the masker), S_u is the upper part, f is the mid-frequency of the masker and L is the masker loudness

Dimkovic's Modifications (continued)

- Psycho Acoustic improvements (continued)
  - Addition of Temporal Masking
    - Only includes forward (post) masking
    - Required modifications to the psychoacoustic prediction model
  - Block Switching Decision modifications
    - Standard uses "Perceptual Entropy" to detect a change. Change in PE forces a change to short block mode
      - This method has high probability of false detection
    - Modified method uses less than one block
      - Signal is first high-pass filtered (low pass signals can use long block)
      - Perceptual Energies are measured in each sub-block and linear prediction is done in the time-domain
      - Two thresholds are used to enact a switch to the short block size
        - PE threshold
        - Linear prediction threshold

Dimkovic's Modifications (continued)

- Psycho Acoustic improvements (continued)
  - Middle/Side stereo coding
    - Detects differences between Left and Right and switches to M/S encoding only when there are no significant differences
  - Bit Allocation / Quantization
    - ISO model recommends global gain be increased when rate distortion loop cannot converge
      - Doesn't always work, especially at low rates
        - Some frequency bands are more distorted than others
    - Modified model uses analysis by synthesis to estimate bit allocation before loops
      - Based on NMR and Signal energy
      - Allows best quantization even if perceptual goals are not fully met.
  - Bit Reservoir
    - A better management module is introduced to combat problem of reservoir being drained to often.
  - Variable Bit Rate
    - New structure to allow constant quality variable rate (as apposed to the standard constant rate variable quality)

Vocoders

- Source Encoding not perceptual encoding
- Tries to model the human vocal tract using a combination of source signals and filters
- Simpler to create than perceptual encoders
  - Longer history
  - Originally executed in the analog domain
    - See “Duck Call” web-page

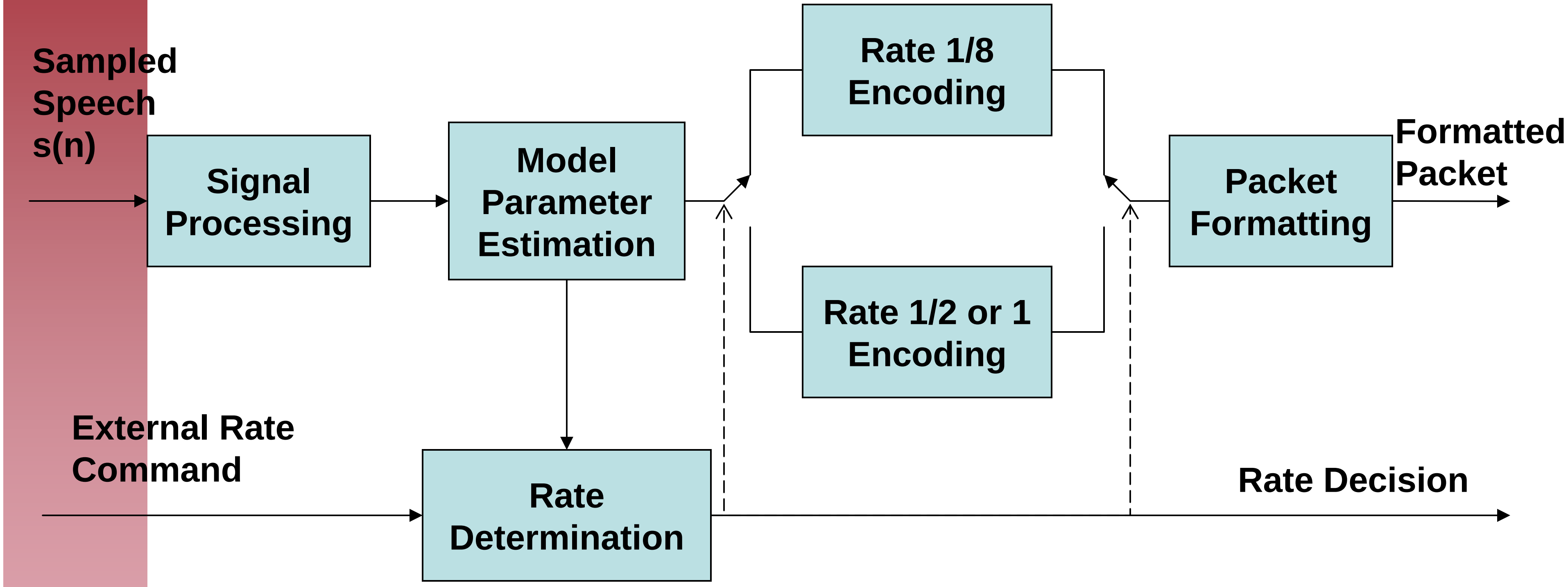
Vocoders by Standard and Generation

- CDMA
  - QCELP – 8k
    - Generally referred to as the worst.
    - Was not used for a significant amount of time in the US
  - QCELP – 13k
    - Similar to 8k, but used more bits when quantizing parameters
  - EVRC – 8k
    - Introduced as a new standard that was tolerable at 8kbps
- GSM
  - Full Rate (FR) Vocoder
  - Half Rate (HR) Vocoder
  - Enhanced Full Rate (EFR) Vocoder
  - Adaptive Multi Rate (AMR) Vocoder

EVRC Overview

- Based on RCELP
  - Modified for variable rate operation
  - Increased robustness for CDMA environment
  - Does not try to match original residual – tries to match pitch contour
  - By using the “perceptual” model, bits are freed up to create a larger code-book
- 3 out of the 4 CDMA packet types are used for vocoding
  - Rate 1 (171 bits/packet)
  - Rate 1/2 (80 bits/packet)
  - Rate 1/8 (16 bits/packet)

EVRC Encoder



EVRC Block functions

- First block performs pre-processing
  - High-pass filtering
  - Adaptive noise suppression filtering
- Model Parameter Estimation
  - Determines Linear Prediction Coefficients (LPC)
  - Converts LPCs into Line Spectral Pairs (LSP) and calculates optimal pitch delay τ
- Rate Determination
  - Applies Voice Activity Detection (VAD) and additional logic to determine packet type
- Rate 1/8 Encoding
  - Background noise
  - No speech periodicity, just encode noise source and energy contour
- Rate 1 or 1/2 Encoding
  - Use RCELP algorithm to match time-warped version of speech residual

LPC and LSPs

- Calculate the formant filter parameters from pre-processed speech.
- Generate the correlation function of a windowed section of the pre-processed speech.
  - First 11 are used for LPC analysis
  - All 17 terms are used for Rate Detection
- Generate the LPCs from the correlation function
- Convert to LSPs
  - LSPs can be encoded with less precision that LPCs
  - Determine Pitch
- Determine Pitch delay and long term prediction gain

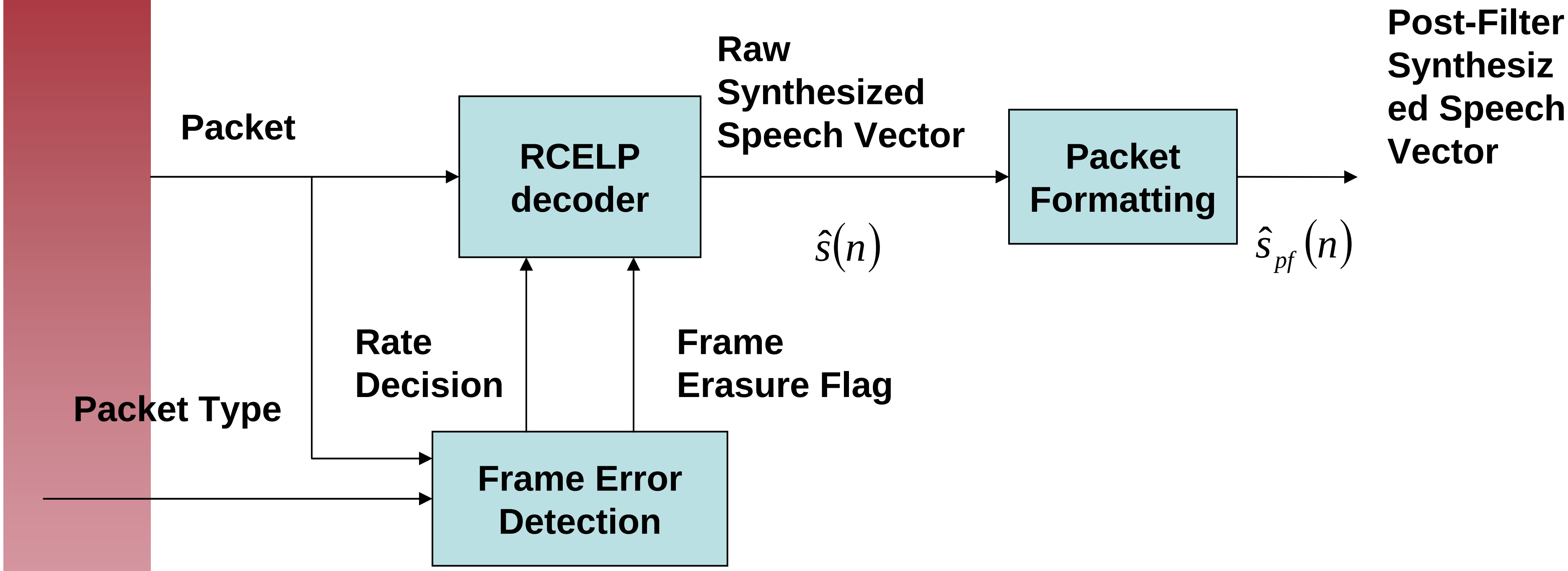
Data Rate Determination

- Used to select between 1,1/2 and 1/8
  - Active Speech is 1 or 1/2
  - Background noise is 1/8
- Band energy is used to determine how voiced a signal is
- Rate Variation
  - Rates can increase as fast as possible
  - Additional delays are introduced when the rate is decreasing.

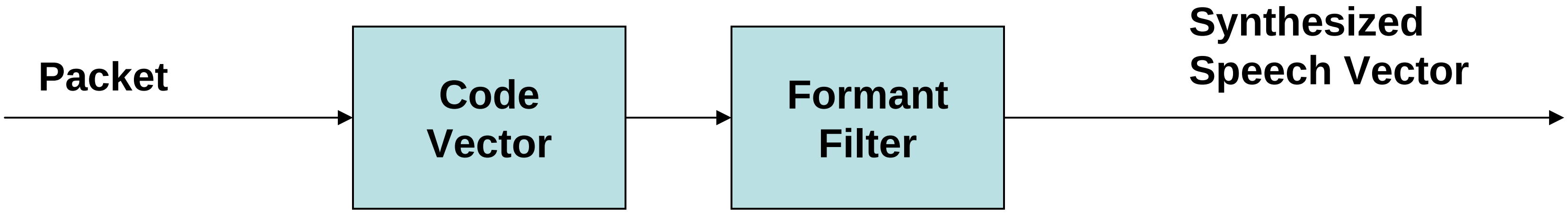
LSP Quantization

- Multiple code books exist for input specification
  - 4 codebooks for full rate
  - 3 codebooks for 1/2 rate
  - 2 codebooks for 1/8 rate
- Signal is used to select which entry in the code book to use
  - Algebraic Code book search
    - Minimizes the mean-squared error between input weighted speech and weight synthesis speech.

EVRC Block Diagram



RCELP Basic description



References

[Bra99] Brandenburg K. “MP3 and AAC Explained”, AES 17th International Conference of High Quality Audio Coding

[Bra00] Brandenburg K. & Popp H. “An introduction to MPEG Layer-3”, EBU Technical Review, June 2000

[Cav02] Cave, C. “Perceptual Modeling for Low-Rate Audio Coding”, Masters Thesis, 2002

[Cav02] Cave, C. “Perceptual Modeling for Low-Rate Audio Coding”, Masters Thesis, 2002

[Chu03] Church, S. “On Beer and Audio Coding”

[Cox05] Cox, T. http://www.acoustics.salford.ac.uk/research/arc/cox/sound_quality/Frequency%20spectrum.htm

[Dim05] Dimkovic, I. “Improved ISO AAC Coder”

[Gol00] Gold, B. & Morgan, N. Speech and Audio Signal Processing, Wiley, 2000

[Moo98] Moore, B. and Alcantara J., “Masking patterns for sinusoidal and narrow-band noise maskers” *J. Acoustic. Soc. Am.,* Vol 104 Aug 1998

[Nav05] Nave, R. http://hyperphysics.phy-astr.gsu.edu/hbase/sound/loud.html

[Pan93] Pan, D. “Digital Audio Compression”, Digital Technical Journal Vol. 5 No. 2, 1993

[Pan95] Pan, D. “A Tutorial on MPEG/Audio Compression”, IEEE Multimedia, 1995

[Smi05] Smith, J. http://ccrma.stanford.edu/~jos/sitemap.html

[Ter82] Terhardt, E. “Algorithm of extraction of pitch and pitch salience from complex tonal signals”, *J. Acousti. Soc. Am.* vol. 71 Mar. 1982

Online Resources

- http://www.mp3-tech.org/programmer/docs/
  - Includes Audio Layer Standard!
- http://www.snopes.com/music/media/cdlength.htm
  - Urban Legend regarding designed length of the CD
- http://www.disctronics.co.uk/technology/cdaudio/cdaud_intro.htm
  - Good article on the tech aspects of CDs
- http://www.ee.washington.edu/conselec/CE/kuhn/cdmulti/95x7/iec908.htm
  - Good description of Error Correction in CDs
- http://www.cdrfaq.org/faq02.html#S2-17
  - FAQ on CD-Rs
- http://www.epanorama.net/documents/audio/spdif.html
  - S/PDIF details
- http://www.exploratorium.edu/exhibits/vocal_vowels/vocal_vowels.html
  - Duck Call with vocal tract
- http://www.3gpp2.org/Public_html/specs/C.S0014-0_v1.0_revised.pdf
  - EVRC Standard