AUDIO AND VOICE COMPRESSION

N. C. State University

CSC557 ♦ Multimedia Computing and Networking

Fall 2001

Lectures # 07&08

AUDIO AND VOICE COMPRESSION

N. C. State University

CSC557 ♦ Multimedia Computing and Networking

Fall 2001

Lectures # 07&08

Types of Audio Compression

- First: general (MPEG-I) compression
- Second: speech compression

Some Non-Speech Audio Compression Standards

Standard	Frequency Range	Compression Method	Sampling Rate	Precision	Bit Rate	Quality
IMA-ADPCM	200-20000Hz	ADPCM	8-44.1 KHz	4 bits	32-350 Kb/s	Telephone …CD
Audio CD	20-20000 Hz	Linear PCM	44.1 KHz	16 bits	1400 Kb/s (stereo)	CD!
MPEG-1	20-20000 Hz	Sub-band coding	32-48 KHz	2-15 bits	128-384 Kb/s	Near-CD

MPEG-1 Audio Compression

- MPEG = Motion Picture Expert Group
  - MPEG-I, Layer III = "MP3"
- Layer I = near CD stereo quality, 384 Kb/s, 32, 44.1 or 48KHz sampling
- Layer II = near CD stereo quality, 256 Kb/s
- Layer III = less than CD stereo quality, 128 Kb/s
- All methods are "lossy"

"Psycho-Acoustic Model"

- MPEG-I audio compression heavily exploits the properties of human hearing
- Property #1: the threshold of hearing is frequency dependent
  - Established an "Absolute Threshold", or "Quiet Threshold"
- Property #2: masking of frequencies by other frequencies
- Properties #1 + #2 → "Masking Threshold"
  - The minimum level of audibility in the presence of masking "noise"
  - Varies over time, as the noise varies
- Used to determine maximum allowable quantization noise at each frequency to minimize noise perceptibility

Frequency vs. Loudness Perception

Source: Haskell 1997, *Digital Video, An Introduction to MPEG-2*

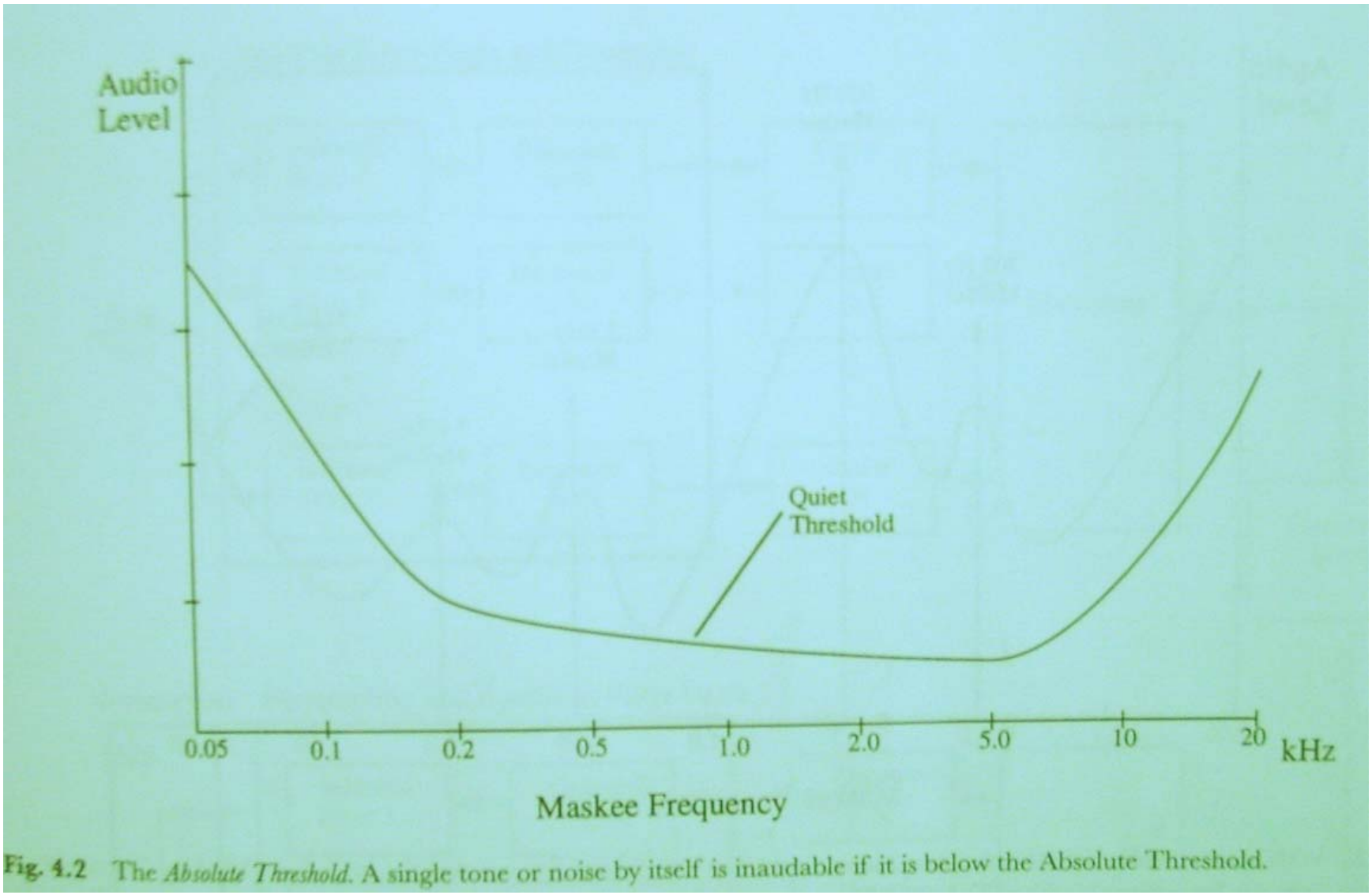


Fig. 4.2 The *Absolute Threshold*. A single tone or noise by itself is inaudable if it is below the Absolute Threshold.

Masking Threshold

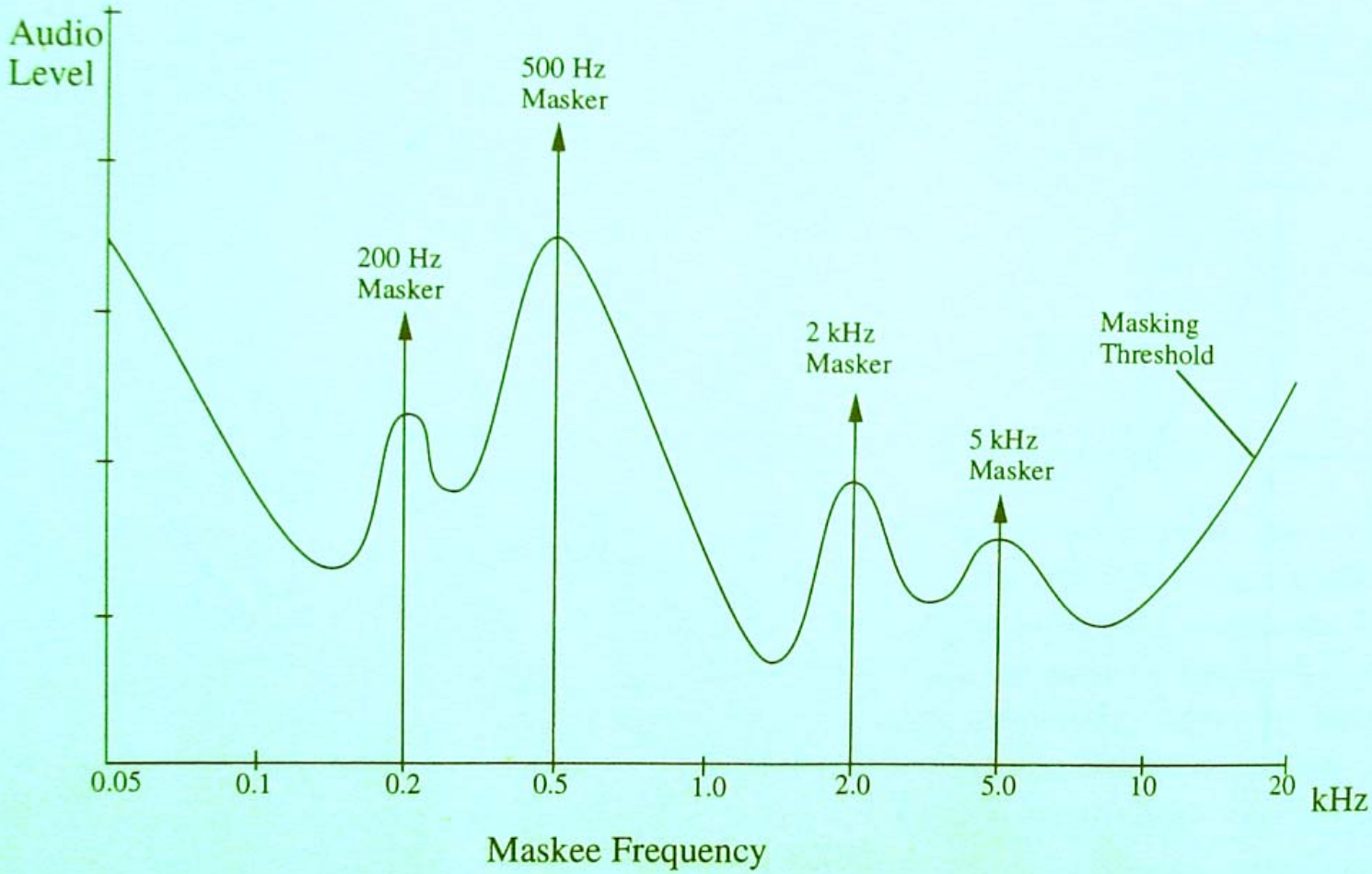


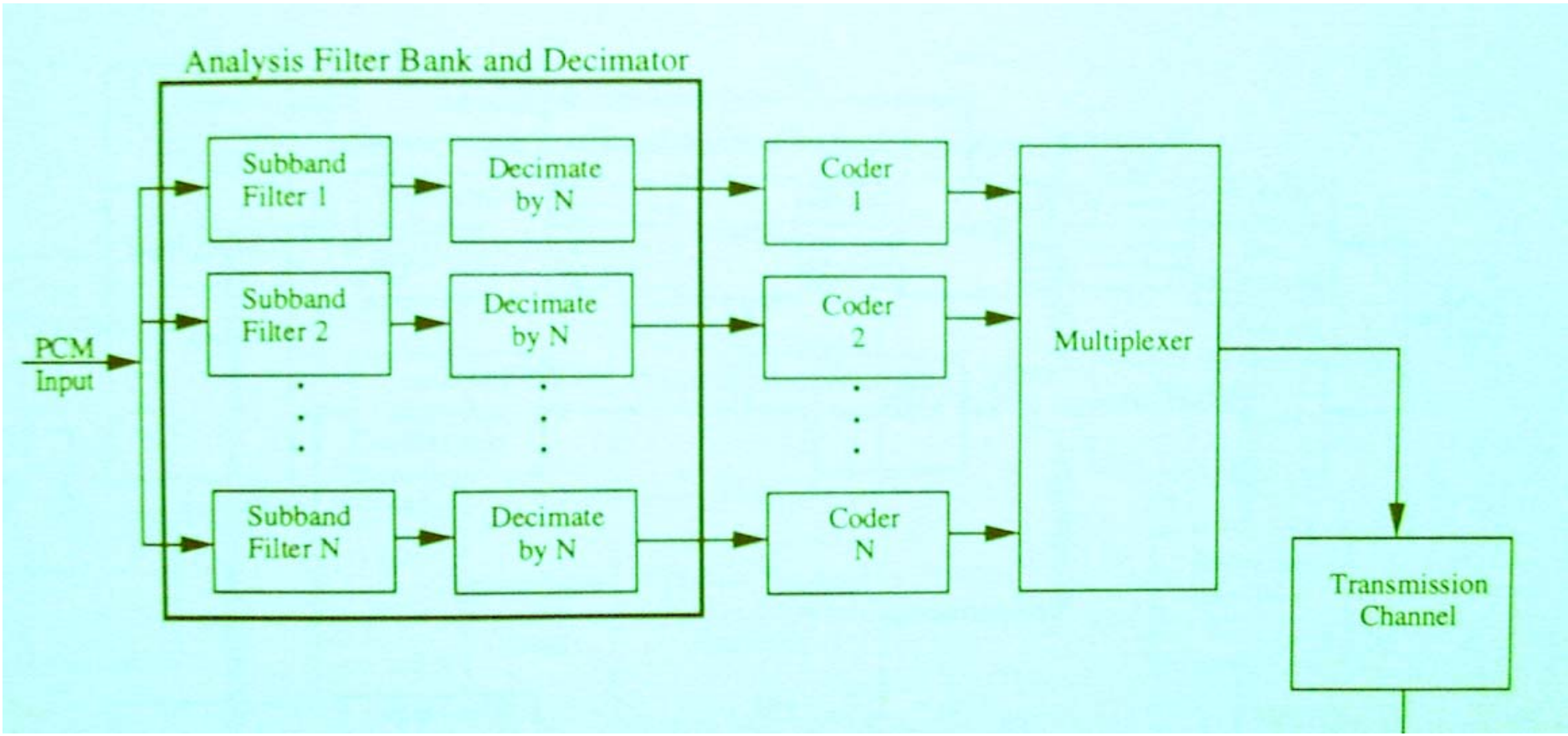
Fig. 4.3 With normal audio signals containing many maskers at a variety of frequencies, the overall net Masking Threshold is calculated from the Frequency Masking, Temporal Masking and Absolute Threshold for all the maskers. The Masking Threshold is a time varying function of frequency that indicates the maximum inaudible noise at each frequency.

Sub-Band Audio Coding

- Incoming signal decomposed into N digital bandpassed signals by a bank of digital bandpass analysis filters
- Each bank independently subsampled ("decimated") so total number of samples from all banks = number of samples in original input
  - Ex.: 4 banks, reduce # of samples from each bank by 3/4, so total # of samples remains the same
- Now compress each bank independently
- Reconstruction (decoding):
  - Increase sampling rate of each bank back to original rate
  - Synthesize signal for each bank
  - Sum together outputs of banks

Sub-Band Encoding

Source: Haskell 1997, *Digital Video, An Introduction to MPEG-2*



Sub-Band Decoding

Source: Haskell 1997, *Digital Video, An Introduction to MPEG-2*

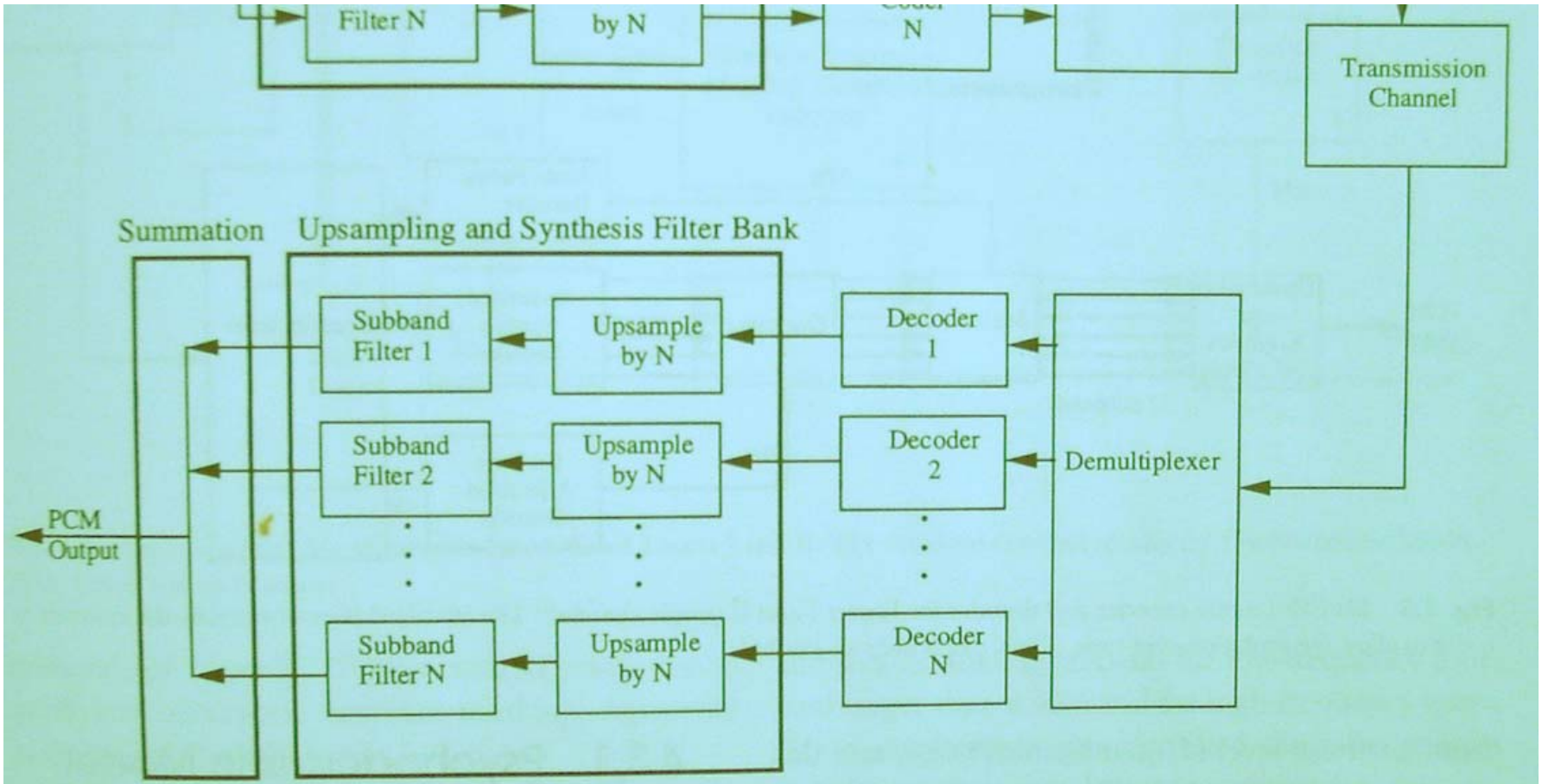


Fig. 4.4 Subband Audio Coder (single channel). The incoming signal is fed to a parallel bank of N digital bandpass *Analysis* filters to produce N digital bandpass signals. These are then decimated by a factor of N, coded and transmitted. Each coded bandpass signal thus consists of one sample for every N input PCM samples. At the decoder the bandpass signals are upsampled by a factor of N by the insertion of zeros, then fed to a bank of digital bandpass *Synthesis* filters and finally summed to produce the

Steps in MPEG-I Audio Encoding

I. Sample input at 16 bits/sample

II. Decompose input into frequency "sub-bands" by bandpass filtering

- 32 sub-bands, filter order = 511

III. Subsample each sub-band (decimation)

- throw out 31 of every 32 samples for each sub-band

IV. A "Block" = 12 consecutive samples in each (decimated) sub-band

- this is the unit of compression
- 32 sub-bands * 12 samples = 384 samples to compress

Encoding (cont.)

V. Scale so the largest value in each sub-band is slightly less than 1
- From a table of 63 scaling factor values
- remember (record) the scaling factor for each sub-band

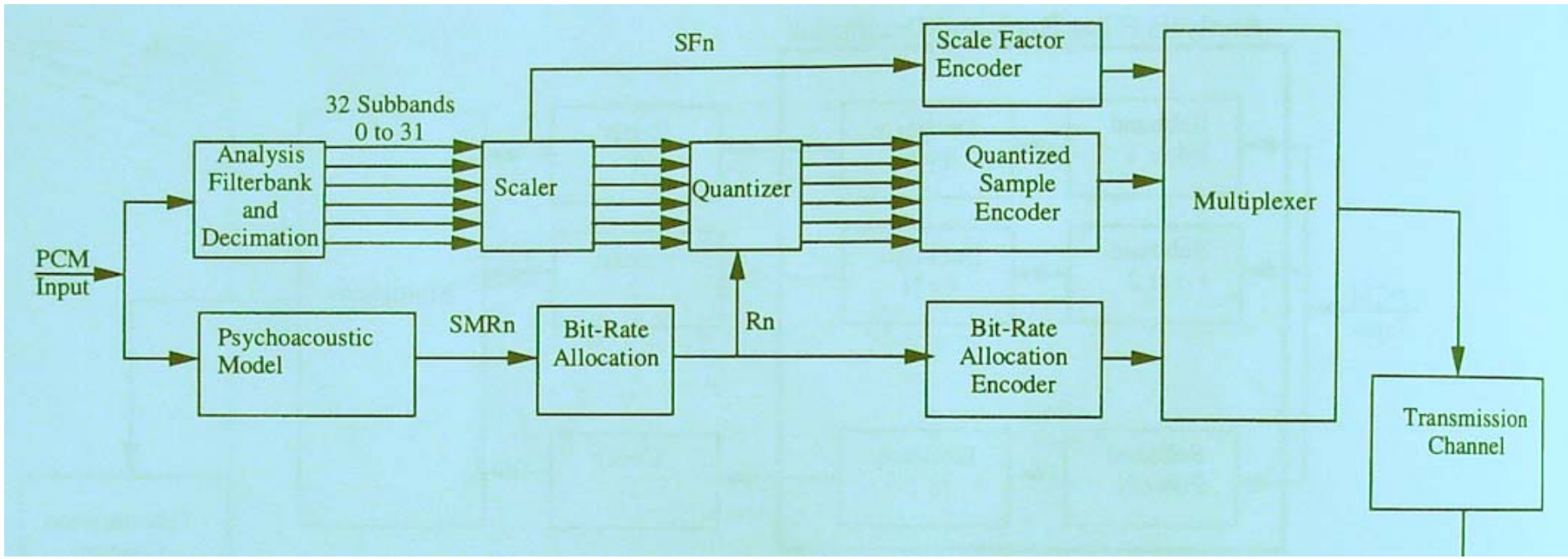
VI. Determine maximum allowable quantization noise in each sub-band
- "signal masking ratio" (SMR) computed by psycho-acoustic model (more to come…)
- allocate bits/sample for each sub-band so that ratio of quantization noise to SMR is roughly the same for each sub-band

VII. Quantize samples in a sub-band according to this bit allocation
- using uniform "mid-step" quantizer

VIII. Multiplex sub-band blocks into a single output stream

MPEG-I Encoding Block Diagram

Source: Haskell 1997, *Digital Video, An Introduction to MPEG-2*



Psycho-acoustic Model

i. Transform the original input (not band-pass filtered) to the frequency domain
   - using 512-point DFT or 1024-point DFT
   - divide the result into sub-bands

ii. Compute "sound level" within each sub-band
   i. Maximum amplitude of any frequency within sub-band

iii. Separate out the major tonal frequencies (peaks); the remainder = a "lumped noise" source
   - lumped noise source assigned a representative frequency
   - noise source + major frequencies = the "maskers"

iv. Discard frequencies that are inaudible due to absolute threshold, or due to masking by a louder "nearby" frequency

Psycho-acoustic Model (cont'd)

iv. Compute the masking threshold from absolute threshold + effects of maskers

v. Compute "minimum masking threshold" for each sub-band
   - "the" threshold = minimum for any frequency in that sub-band

vi. Compute signal-to-minimum-masking threshold ratio (SMR) for each sub-band
   i. SMR = sound level – minimum masking threshold

Masking Threshold, Again

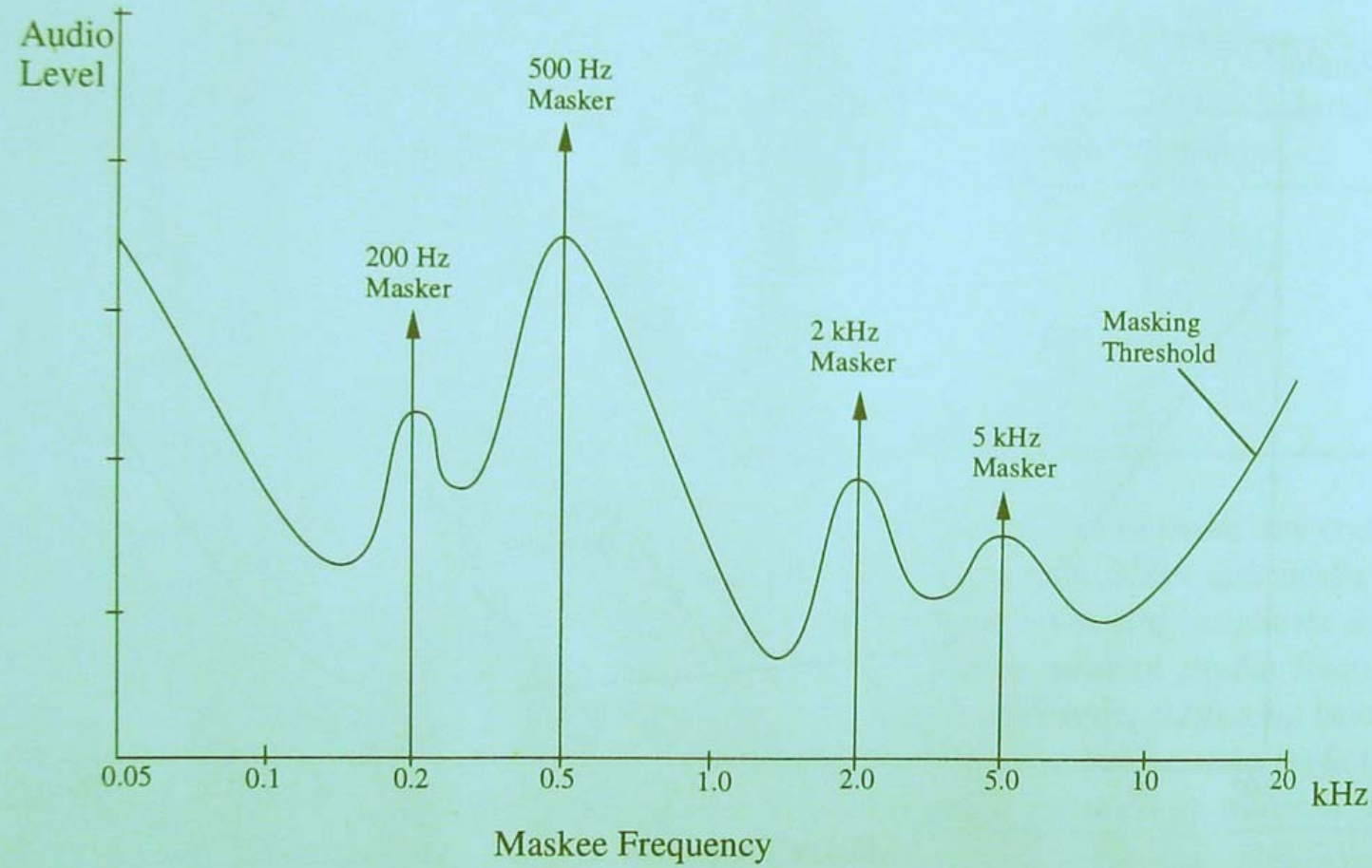


Fig. 4.3 With normal audio signals containing many maskers at a variety of frequencies, the overall net Masking Threshold is calculated from the Frequency Masking, Temporal Masking and Absolute Threshold for all the maskers. The Masking Threshold is a time varying function of frequency that indicates the maximum inaudible noise at each frequency.

Psycho-Acoustic Model (Computing The Threshold)

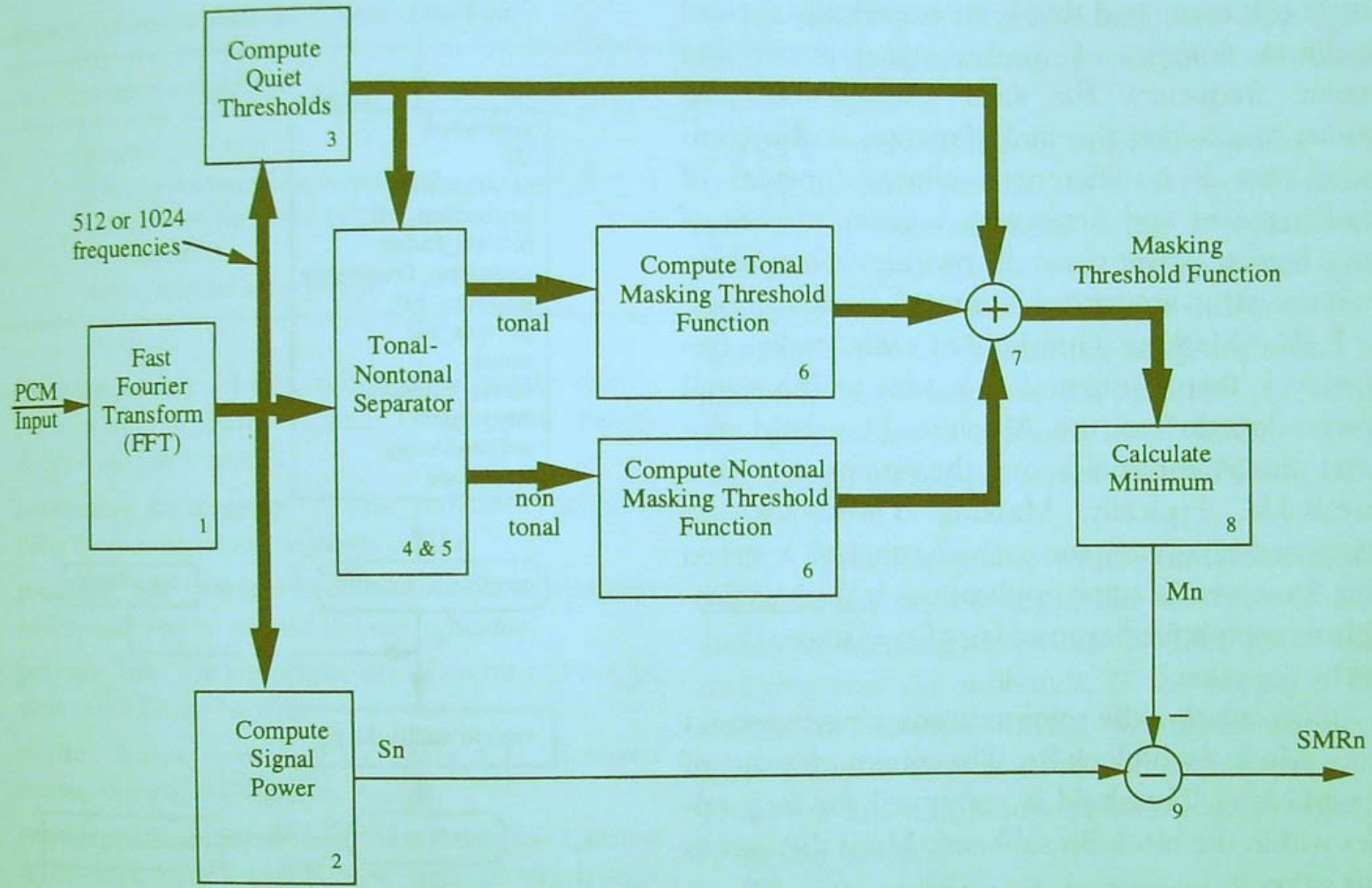


Fig. 4.6 Psychoacoustic Model suitable for MPEG-1 Layers I and II. The standard does not specify the Psychoacoustic Model. Thus, this is only an example.

MPEG-I Audio Decoding Steps

I. Demultiplex

II. Unquantize

III. Unscale (multiply by scale factor)

IV. Unblock

V. Upsampling (zero insertion for removed samples)

VI. Band pass synthesis filter each sub-band, and sum the outputs

- Decoding is the only part of processing that is standardized!
  - encoding can be changed as new methods are created

MPEG Audio Decoding

Source: Haskell 1997, *Digital Video, An Introduction to MPEG-2*

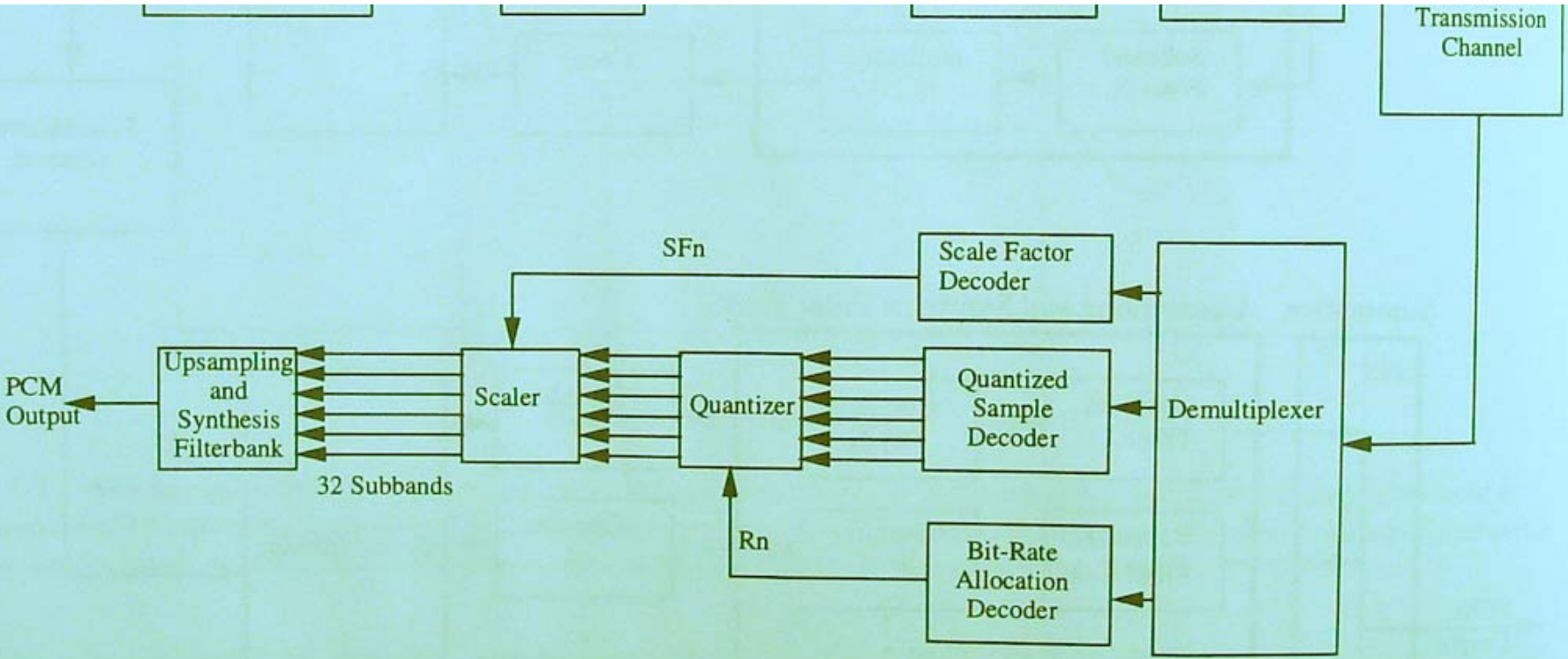


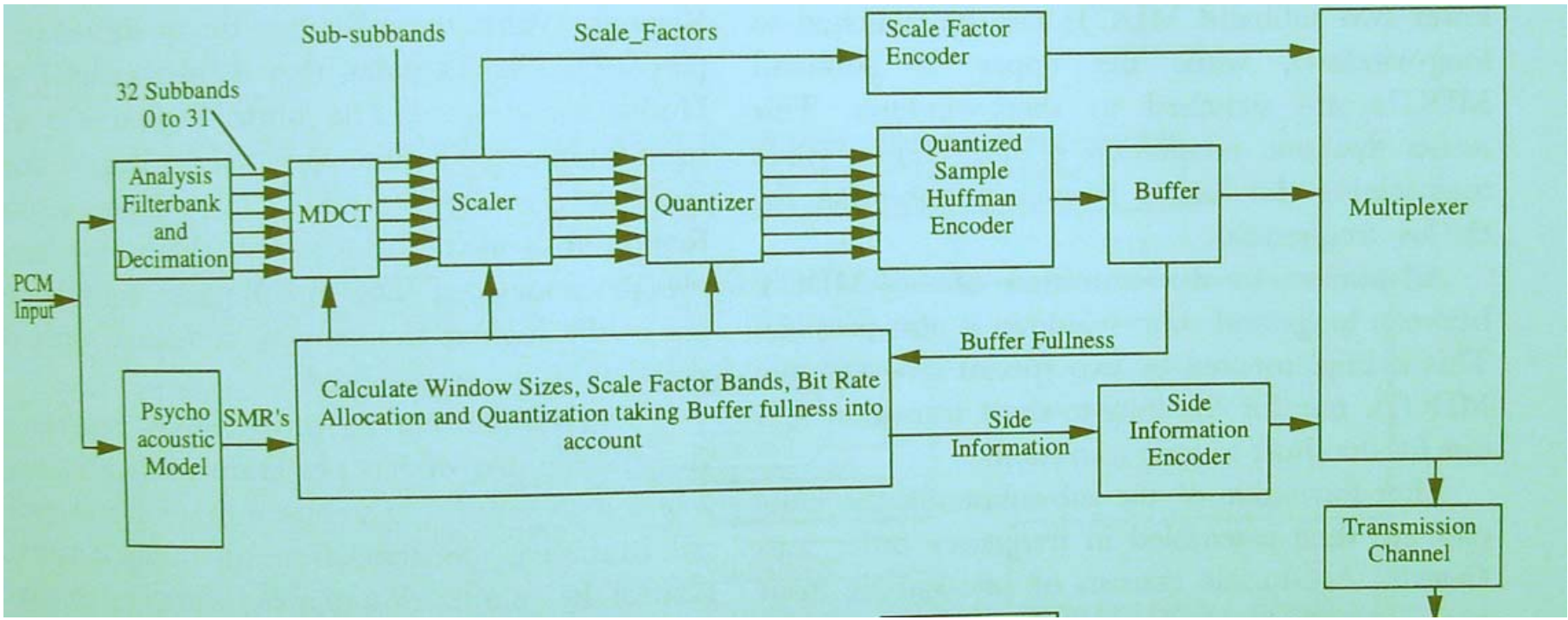
Fig. 4.5 MPEG-1 audio encoder and decoder for Layers I and II (single channel). The standard does not specify the encoder in order to allow for evolution over time. Thus, this is only an example.

MPEG-I Layer III Encode/Decode

- Layer II improvements over Layer I
  - exploit similarities of scaling factors for consecutive blocks
  - reduce the precision (# of bits, or quantization levels) for the high frequency sub-bands
- Layer III improvements over Layer II
  - further frequency resolution into "sub-sub-bands"
  - considerably more complex method; for us, "beyond the scope…"

Layer III Encode

Source: Haskell 1997, *Digital Video, An Introduction to MPEG-2*



Layer III Decode

Source: Haskell 1997, *Digital Video, An Introduction to MPEG-2*

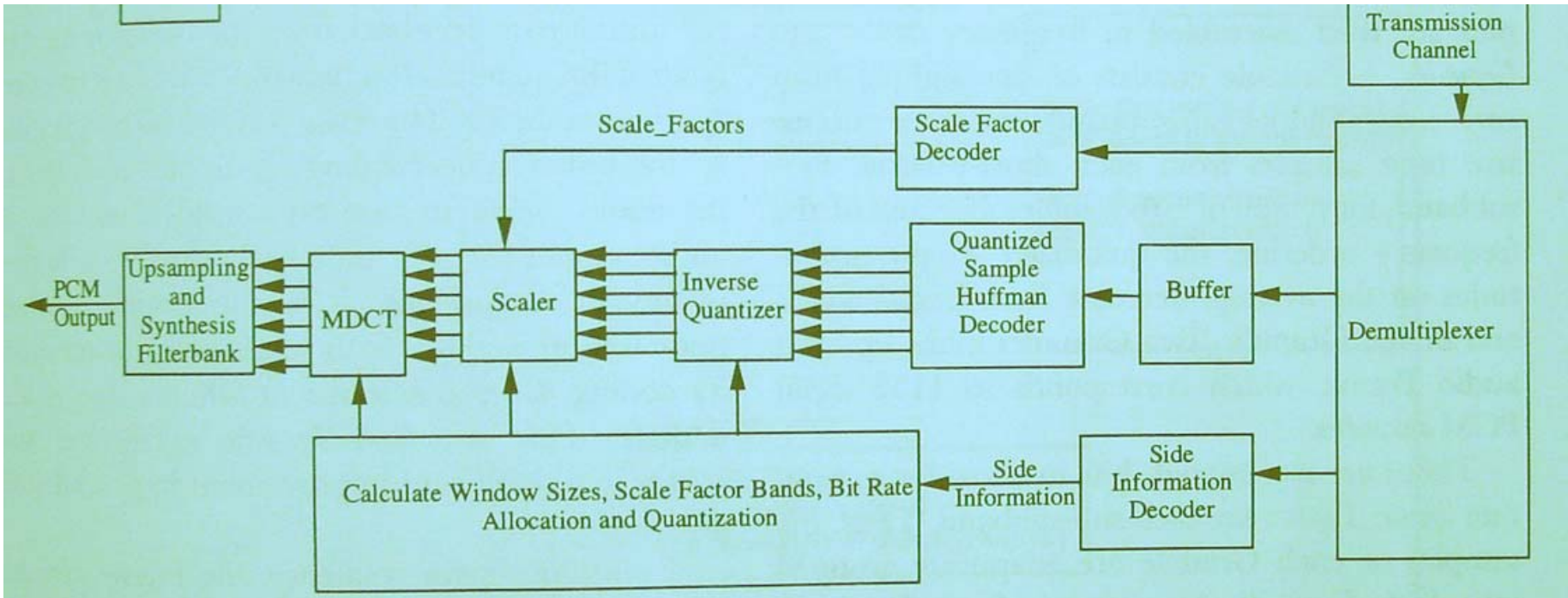


Fig. 4.10 MPEG-1 audio encoder and decoder for Layer III (single channel). The standard does not specify the encoder in order to allow for evolution over time. Thus, this is only an example.

Part II. Speech Compression

Bit Rates of Some Speech Compression Techniques

Standard	Frequency Range (Hz)	Compression Method	Sampling Rate (KHz)	Precision	Bit Rate (Kb/s	MOS
IMA-ADPCM	200-20000	ADPCM	8-44.1	4 bits	32-350	Telephone …CD
G.711	200-3200	Mu-Law PCM	8	8 bits	64	4.4
G.721	200-3200	ADPCM	8	4 bits	32	4.1
GSM	200-3200	CELP	8	variable	13	3.6
G.728	200-3200	LPC + VQ	8	2 bits	16	4.0
G.729	200-3200	Modified CELP	8	variable	8	4.2
G.722	50-7000	DPCM	16	4 bits	64	AM Radio
G.723	200-3200	CELP	8	variable	5.3 and 6.3	4.0

Criteria for Speech Coding

1. Bit rate, or amount of compression
   - phonemic bit rate of speech: approx. 50 bits/sec (bps)
   - cognitive content bit rate of speech: approx. 400 bps
   - how close can we get to this?
2. Intelligibility
3. "Naturalness", quality
4. Processing effort
5. Complexity of implementation
6. Delay (maximum time between receiving a sample and outputting the encoded or compressed value)
7. Robustness (to bit errors)

Speech Compression Categories

- Quantization adaptation
  - Logarithmic PCM, again
- Waveform coding with fixed prediction
  - DPCM, ADPCM (adaptive quantizer), and delta modulation
- Linear predictive coding (LPC)
- Waveform coding with adaptive prediction
- Analysis by synthesis LPC
  - CELP

Logarithmic PCM Again: μ-Law Coding

- Definition
  - X = input voice (sampled) signal
    - normalized to the range $-1 \leq x \leq 1$
  - Y = digitized (quantized) output signal
  - $y = S_{max} * [\ln(1+(\mu * |x|))] / \ln(1+\mu) * sign(x)$
    - S_{max} = maximum possible digital value
    - sign(x) = 1 if x is non-negative, else = -1
    - μ= 255 usually for telephony
  - actual mu-law encoding is a piece-wise approximation to this function
- Also called "companding" (compression+expansion)

Examples (mu-law)

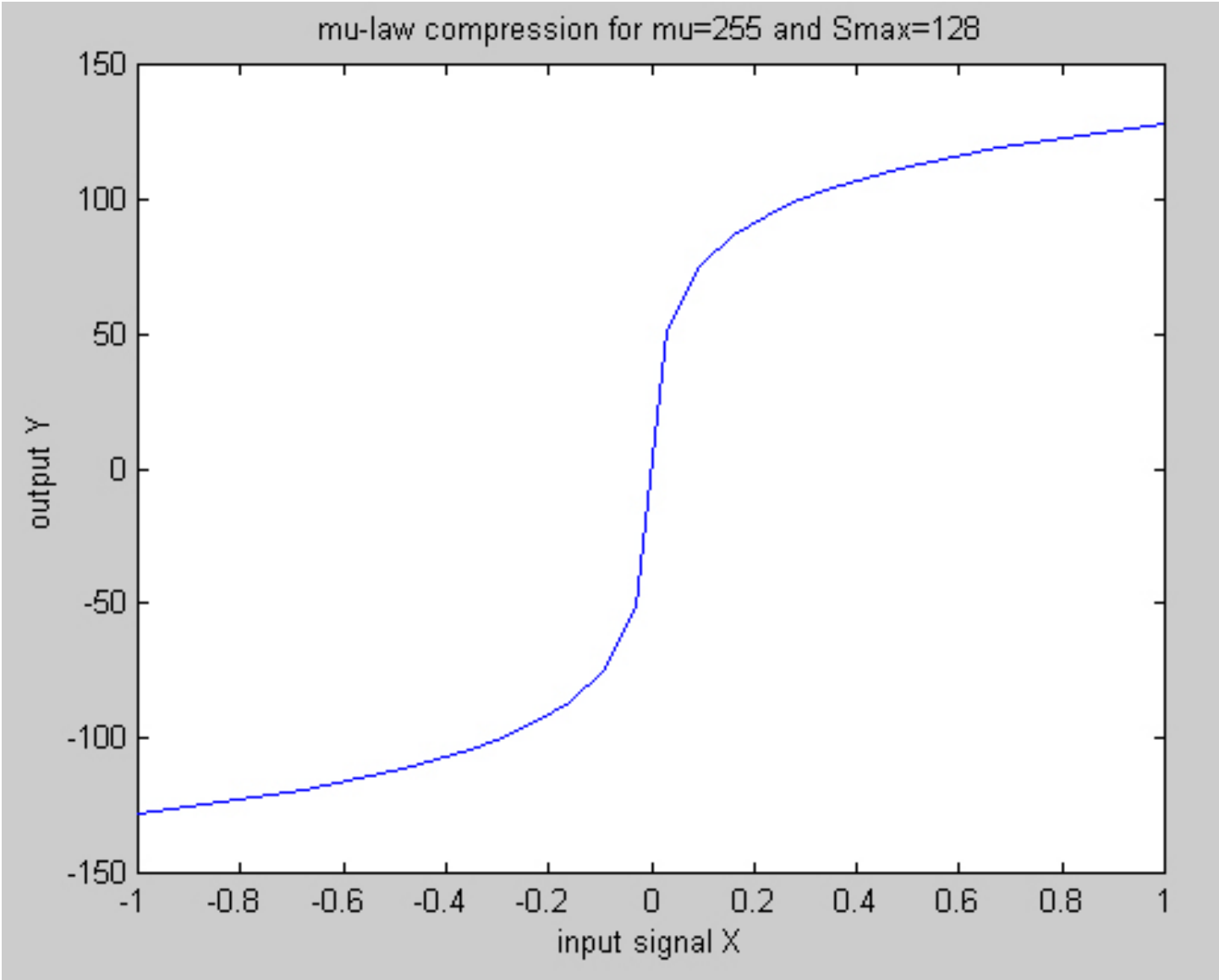
x = .05, y = 128 * [ln(1+(255*.05))] / ln(1+255)*1 = 61

x = .25, y = 128 * [ln(1+(255*.25))] / ln(1+255) * 1 = 96

x = .8, y = 128 * [ln(1+(255*.8))] / ln(1+255) * 1 = 123

x = -.4, y = 128 * [ln(1+(255*.4))] / ln(1+255)* -1 = -107

mu-law



Step Size, or Quantizer, Adaptation

- Adapting the quantization scheme gives better results than a fixed quantization scale
- Generally: set step size proportional to the variance of a neighborhood of values around the current sample
  - higher variance → larger step sizes
  - lower variance → smaller step sizes

Quantizer Adaptation Example

- $\sigma[n] = \sqrt{(1/M * \Sigma_{m=n}^{n+M-1} x^2[m])}$

  = estimated standard deviation of the next m samples starting at sample n

- $\Delta[n] = \Delta_0 * \sigma[n] / 2^{B-1}$

  = step size to use at the n^{th} sample for the next m samples

- Notation
  - x[n] = n^{th} sample of X
  - M = size of the "neighborhood"
  - B = # of bits available for quantization
  - Δ_0 = constant scale factor between 0 and 1

Example (Quantization Adaptation)

- M = 3
- X[I] = 60, x[I+1] = 85, x[I+2] = 120
- B = 8
- $\Delta_0 = 0.5$
- $\sigma[n] = \sqrt{}\ (1/3 * (60^2 + 85^2 + 120^2)) = 91.7$
- $\Delta[n] = 0.5 * 91.7 / 2^7 = .36$ / step
- 85-60 = 25 = 70 steps
- 120-85 = 35 = 98 steps

Waveform Coding with *Fixed* Predictor

- Fixed predictor, fixed quantizer → DPCM, DM
  - Predictor order (number of coefficients) typically 4 or 5
- Fixed predictor, *adaptive* quantizer → ADPCM, ADM

Speech Production

1. Lungs provide power
2. Vocal chords and glottis provide vibrations
   - "voiced sounds" -- generated by vocal chords/glottis
3. Vocal tract (throat, mouth, nose, lips) modifies the result
   - unvoiced ("fricative") sounds -- generated by a constriction of the airflow

- Components of the voice tract change at a "relatively" slow time scale

A Simple Model

1. Power component: what's the "gain", i.e., how much power do the lungs produce?
2. Excitation component: is the sound voiced, or unvoiced?
   - if voiced, at what frequency are the vibrations?
   - if unvoiced, what does the turbulence look like?
     - "noise" or pseudo-random signal close enough?
3. Filter component: what is the effect of the rest of the vocal tract?
   - deriving the coefficients of filter?

Simple "Pitch-Excited" Model

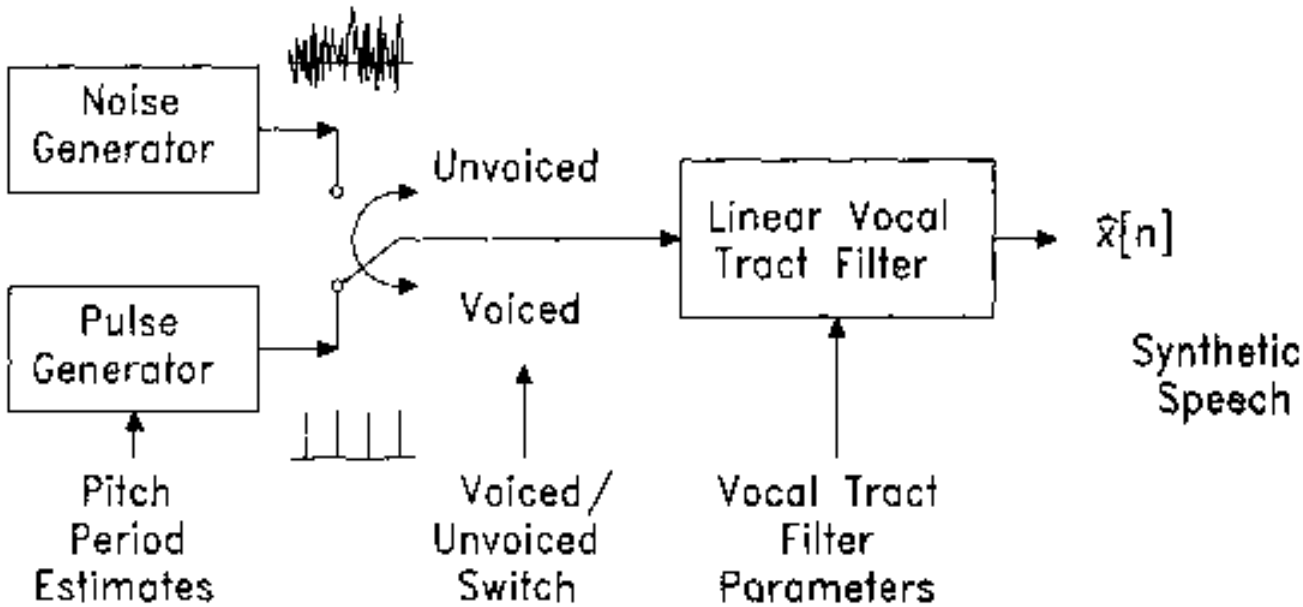


Figure 1.2. Pitch-excited vocoder synthesis model.

Source: [Barnwell 1996] *Speech Coding: A Computer Laboratory Textbook*

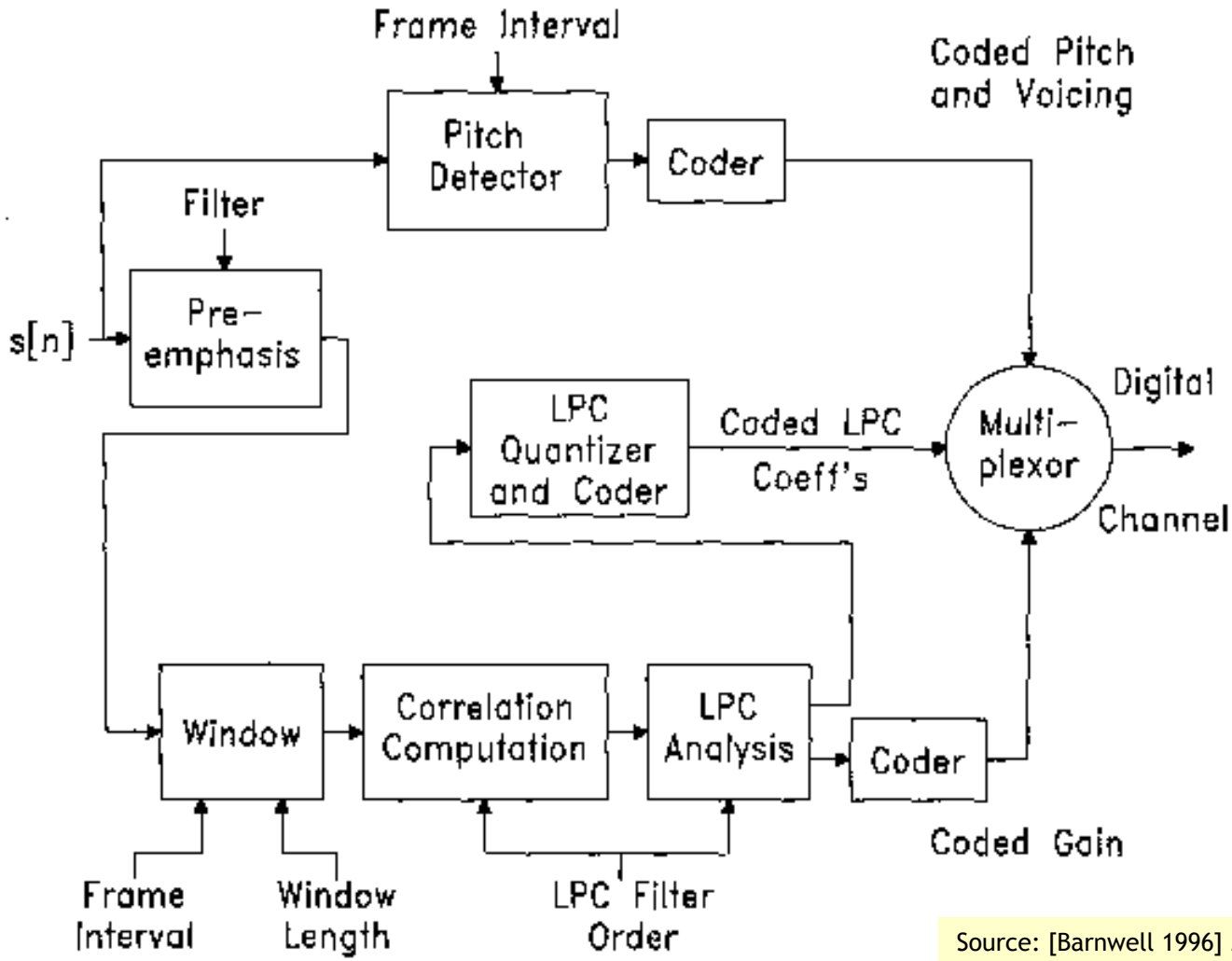
Frames

- Interval of time over which a sound is processed
  - producing one set of parameter values
  - voice signal is roughly stationary over small time intervals
  - this is a good unit of processing for compression purposes
- Typically, 1 frame=10-25 msec.

Pitch-Excited Linear Prediction Coding (LPC)

- Most successful advance in speech compression
  - (we will not discuss details; overview only)
- Two analysis steps:
  1. pitch detection (excitation)
  2. vocal tract analysis
- Frames: 10-25 msec (depending on exact method)

LPC Encoder



Source: [Barnwell 1996] *Speech Coding: A Computer Laboratory Textbook*

Figure 5.1. Block diagram of a pitch-excited LPC transmitter.

LPC Decoder

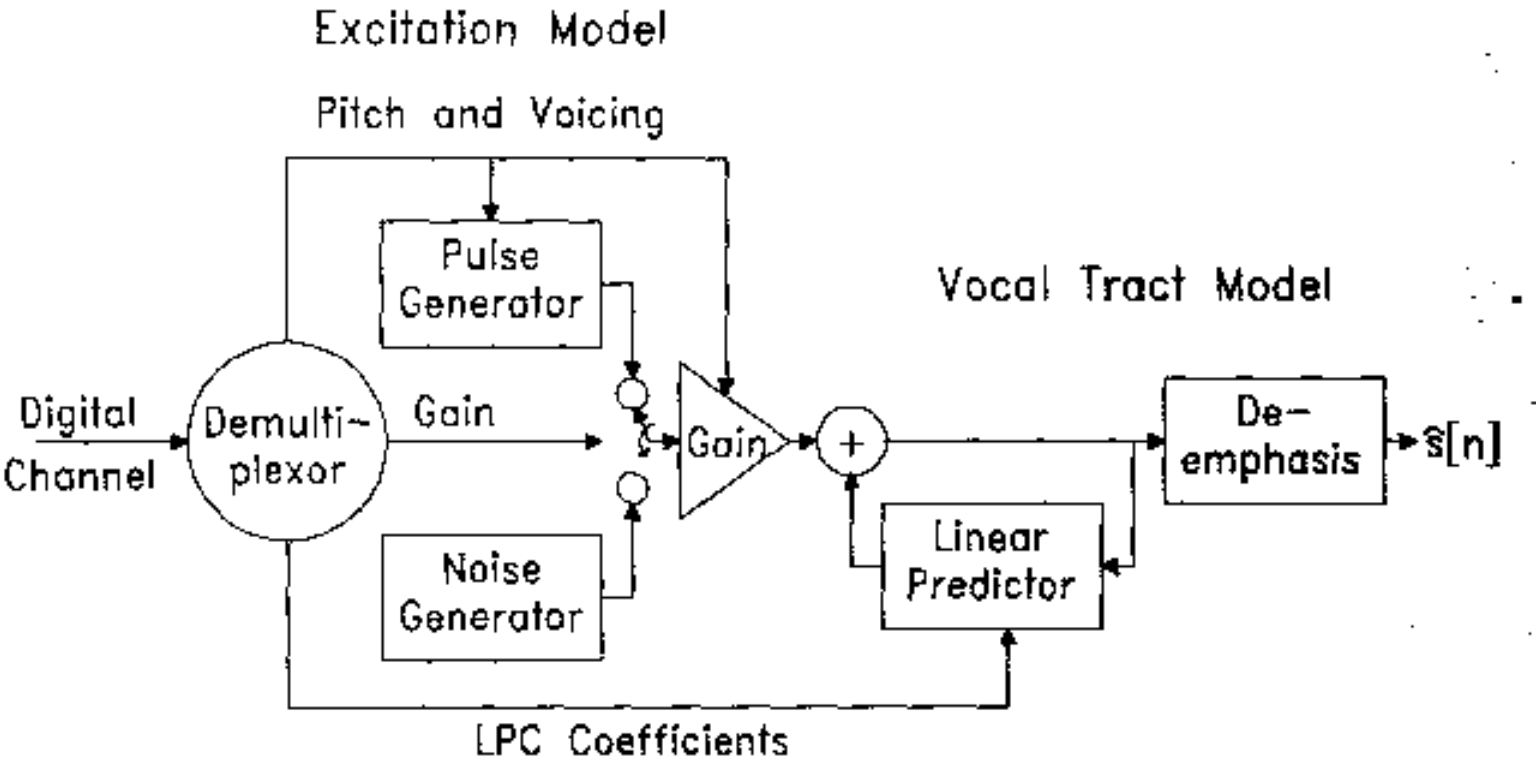


Figure 5.2. Block diagram of a pitch-excited LPC receiver.

Source: [Barnwell 1996] *Speech Coding: A Computer Laboratory Textbook*

Pitch-Excited Linear Prediction Coding (LPC)

- General form: $s[n] = \sum_{i=1}^{P} a_i s[n-i] + Gu[n]$
  - s[n] = output of speech encoder
  - u[n] = excitation signal
    - excitation signal = pulse train (voiced sound) or white noise (unvoiced sound)
  - G = gain (match energy of decoded speech with energy of the input speech signal)

LPC (cont.)

- Pitch detection (for voiced sounds)
  - only looking for the fundamental frequency
- Voiced vs. unvoiced classification
  - counting zero crossings in the time domain
- Gain computation
  - match energy of the filter output on random input with energy of original speech
  - energy = weighted mean-squared sum

LPC (cont.)

- Predictor order (number of coefficients) around 10-15
  - predictor coefficients are adapted to minimize the predicted error
- Quantization
  - linear, coarse quantization OK for gain and pitch period
  - filter coefficients: more precision(8-10 bits) needed, quantization more complex
- LPC assessment
  - intelligible, but not very natural-sounding
  - very low bit-rate

CELP Coding

- Code Excited Linear Prediction
- "Analysis by synthesis"
- Like LPC, but uses a more accurate excitation model
- An excitation generator produces K different sequences
  - try them all!
  - then pick the one that minimizes the energy in the error signal
- The set of possible excitation functions = "the codebook"

CELP (cont.)

- CELP has two components in excitation sequence
  - long-term predictor
  - codebook sequence to use
- For codebook sequence, specify
  - index # in codebook
  - gain to use
- Weighting filter
  - pass noise at high-energy frequencies
  - suppress noise at low-energy frequencies

CELP Diagram

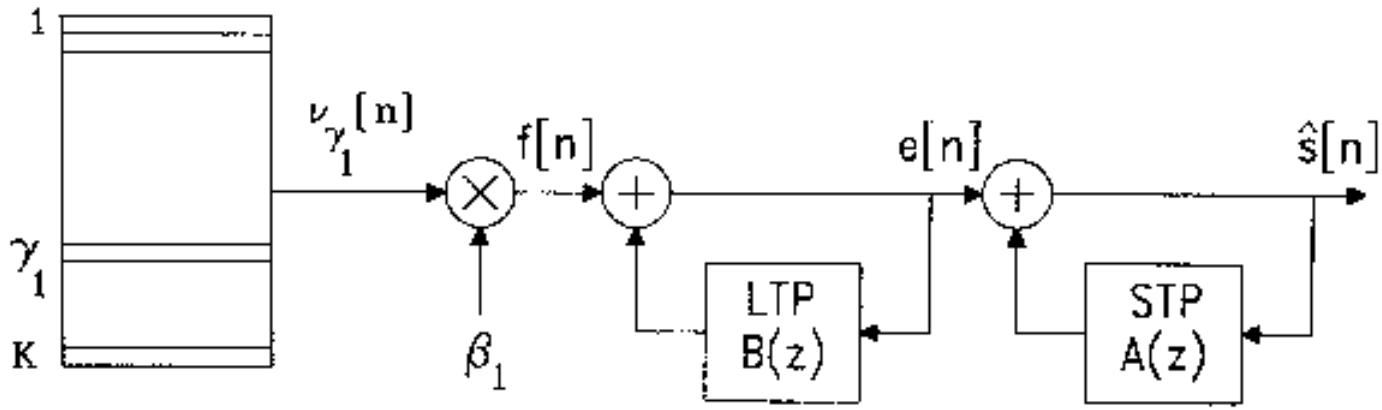


Figure 7.7. Block diagram of a CELP synthesizer.

parameters	name	range	typical values
predictor order	P	1–16	10
LPC window size	L	160-360	240
LPC frame size	I	80-240	120
error-weighting factor	α	0 0.99	0.8
codeword (exc. frame) size	N	20-60	40

Table 7.3. The parameters of the CELP analysis and synthesis.

Source: [Barnwell 1996] *Speech Coding: A Computer Laboratory Textbook*

Example of Speech Compression

- Whole sequence of "wow" files…
  - Reduced sampling rate
  - Reduced # of bits
  - Different compression schemes

Sources of Information

- [Pan] *A Tutorial on MPEG / Audio Compression* (handout)
- [B. Haskell et al], *Digital Video: An Introduction to MPEG-2*
  - Chapter 4 on MPEG Audio Coding, pp. 55-64
- [Barnwell et al 1996] *Speech Coding: A Computer Laboratory Textbook*
  - An overview of speech coding with lots of examples. It is sometimes beyond the scope of our course, but is one of the better treatments.
- [Gibson et al 1998] *Digital Compression for Multimedia*
  - Chapter 5 (Predictive Coding) and Chapter 6 (Linear Predictive Speech Coding Standards) is detailed and dense. Mostly beyond the scope of our course, but a good reference.