# 01076566 Multimedia Systems

## Chapter 9: Media Compression: Audio

Pakorn Watanachaturaporn
*pakorn.wa@KMITL.ac.th*

Bachelor Program in Computer Engineering (B.Eng.)
Faculty of Engineering

King Mongkut's Institute of Technology Ladkrabang

# Outline

KMITL
พระจอมเกล้าลาดกระบัง
pakorn.wa@kmitl.ac.th

- The need for audio compression
- Audio compression theory
- Audio as a waveform
- Audio compression using psychoacoustics
- Model-based audio compression
- Audio compression using event lists
- Audio coding standards

# The need for audio compression

KMITL
พระจอมเกล้าลาดกระบัง
pakorn.wa@kmitl.ac.th

| Audio data | Speech (mono) | CD (stereo) | FM radio (stereo) | 5.1 surround sound |
|---|---|---|---|---|
| Sample size | 8 bits | 16 bits | 16 bits | 16 bits |
| Sample rate | 8 KHz | 44.1 KHz | 22 KHz | 44.1 KHz |
| One second of uncompressed size (B for bytes) | 8 KB | 88.2 KB (stereo) | 44 KB (stereo) | 530 KB (6 channels) |
| One second of transmission bandwidth (b for bits) | 64 Kbps | 1.4 Mbps (stereo) | 704 Kbps (stereo) | 4.23 Mbps (6 channels) |
| Transmission times for one second of data (56 Kb modem) | 1.14 seconds | 25 seconds | 12.5 seconds | 76 seconds |
| Transmission times for one second of data (780 Kb DSL) | 0.08 seconds | 1.8 seconds | 0.9 seconds | 5.4 seconds |

*Figure 9-1 Examples showing storage space, transmission bandwidth, and transmission time required for uncompressed audio formats*

# Audio compression theory

KMITL
พระจอมเกล้าลาดกระบัง
pakorn.wa@kmitl.ac.th

- Digital audio is represented as a one-dimensional set of PCM samples
  - images have spatial 2D representations
  - video has 3D spatiotemporal representations
- The human auditory system (HAS) in the ear is **different** from the human visual system (HVS)
- HAS is more sensitive to noise and quality degradation when compared with HVS
- Therefore, audio-compression techniques need to be careful with any loss and distortion that might be imposed during compression

pakorn.wa@kmitl.ac.th

- Digital audio is represented in the form of **channels**.
- The number of channels describes whether the audio signal is **mono**, **stereo**, or even **surround sound**.
- Each channel consists of a sequence of samples and can be described by a sampling rate and quantization bits.
- When this digital representation is compressed, a number of trade-offs need to be addressed.
  - The perceived quality after compression and decompression needs to be maximized
  - While at the same time the amount of information needed to represent the audio signal needs to be minimized.

KMITL
พระจอมเกล้าลาดกระบัง
pakorn.wa@kmitl.ac.th

- Considerate properties
  - ↑ **Fidelity** : how perceptually different the output of the codec is when compare with the original sound.
    - E.g., telephone quality, HiFi
  - ↓ **Data rate or bit rate**
    - Throughput
    - Storage
  - ↓ **Complexity**
    - Hardware costs in the encoder and decoder
    - Software costs in the encoder and decoder

pakorn.wa@kmitl.ac.th

- Audio-compression techniques can be categorized into different groups
  - Assume that the signal is a simple digital waveform.
    - The compression techniques aim to reduce only the statistical redundancy.
  - Other techniques borrow more sophisticated theories from psychoacoustics
    - remove redundancy in digital audio depending on what the HAS **can** or **cannot** perceive.
  - Another class of techniques focuses purely on low-band speech, and aims to parameterize human voice models.
  - The area of **structured audio** views the audio signal as composed of a list of events, as in an orchestra, the signal can be compressed by semantically describing these events and their interaction.

pakorn.wa@kmitl.ac.th

- Common to all techniques is the understanding of the way sound amplitudes are measured
  - Use to validate the perception of frequencies, and the control of thresholds in the compression process
- Sound amplitude is measured in decibels – which is a relative measuring system

KMITL
พระจอมเกล้าลาดกระบัง
pakorn.wa@kmitl.ac.th

- The decibel scale is a way of comparing two sound intensity levels (watt per square centimeter)
- The decibel scale is defined as $10\log\left(\frac{I}{I_0}\right)$ where $I$ is the intensity of the measured sound signal and $I_0$ is an arbitrary reference intensity
- For audio purposes,
  - $I_0$ is taken to be the intensity of sound that humans can barely hear
  - $I_0$ has been experimentally measured to be $10^{-16}$ watts/sq.cm.

| Type of sound | Decibel level |
|---|---|
| Threshold of hearing (just silent) | 0 dB |
| Average whisper | 20 dB |
| Conversation | 40 dB |
| Traffic on a busy street | 70–90 dB |
| Threshold of pain (going deaf) | 120–130 dB |
| Jet aircraft at takeoff | 140 dB |

*Figure 9-2 Audio decibel measurements for some commonly occurring sounds*

# Audio as a waveform

KMITL
พระจอมเกล้าลาดกระบัง
pakorn.wa@kmitl.ac.th

- This class of algorithms makes no special assumption about the audio signal, and performs all the analysis in the time domain.
- Compression here can only be achieved by different quantization strategies, followed by the removal of statistical redundancies using entropy coding.
- Fail to provide good compression ratios when compared with recent standards that make use of psychoacoustics.

pakorn.wa@kmitl.ac.th

- DPCM and Entropy Coding
  - The range of differences in the amplitude between successive audio samples is less than the range of the actual sample amplitudes.
  - As a result, coding the differences between the successive signal samples reduces the entropy (variation in amplitudes) when compared with the original signal samples.
  - The smaller differences $d(n)$ can then be coded with fewer bits.
  - However, reducing the number of bits introduces quantization errors.
  - Compression can be achieved if these errors are tolerable for human perception.

pakorn.wa@kmitl.ac.th

- DPCM is a **prediction-based** method
  - The next sample is predicted as the current sample
  - The difference between the predicted and actual sample is coded
  - The closer the predicted value is to the actual one, the smaller the range of variation in the difference errors.

pakorn.wa@kmitl.ac.th

- Predicting the next sample value based only on one previous sample may be simple but not as effective as involving a number of immediately preceding samples in the prediction process.
- In such cases, the preceding samples are weighted by coefficients that suggest how much a preceding sample influences the current sample value.
- These are called **predictor coefficients**.

KMITL
พระจอมเกล้าลาดกระบัง
pakorn.wa@kmitl.ac.th

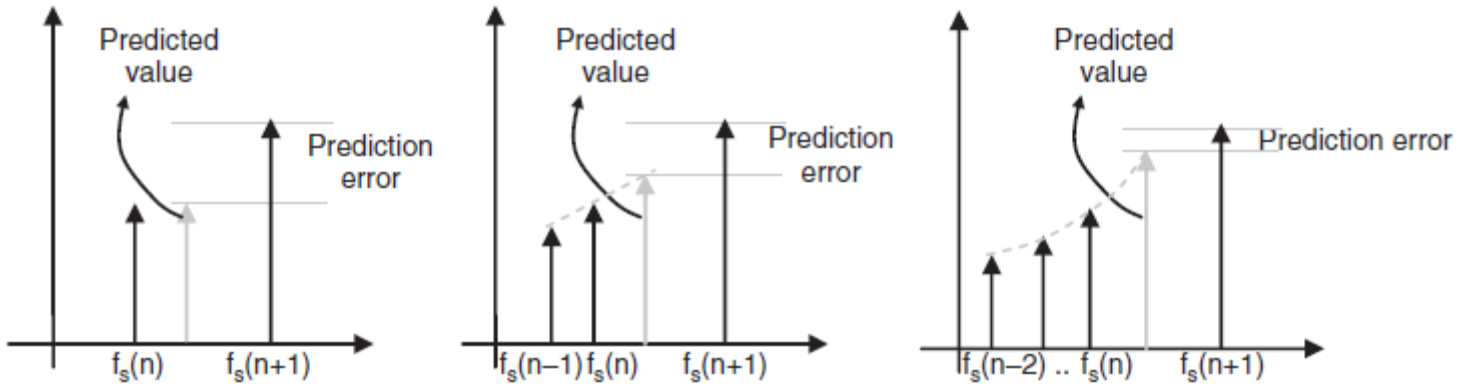



*Figure 9-3 Prediction of sample values in DPCM techniques. The next sample to be coded is predicted based only on preceding sample (left), as a linear combination of two preceding samples (middle), and as a weighted combination of three previous samples (right). The prediction error shown as the difference between the horizontal lines in each case successively decreases as more samples are used.*

pakorn.wa@kmitl.ac.th

- Delta Modulation
  - Similar to DPCM, except that the differences $d(n)$ are coded with one bit—0 or 1.
  - Given a current state,
    - 0 signifies that the next sample has increased by a constant delta amount
    - 1 indicates that the next sample has decreased by a constant delta amount
  - Delta amount is fixed and can be decided depending on the input signal's sampling rate

pakorn.wa@kmitl.ac.th

- ADPCM : Adaptive differential pulse code modulation
  - Similar to DPCM, except that it adaptively modifies the number of bits used to quantize the differences

KMITL
พระจอมเกล้าลาดกระบัง
pakorn.wa@kmitl.ac.th

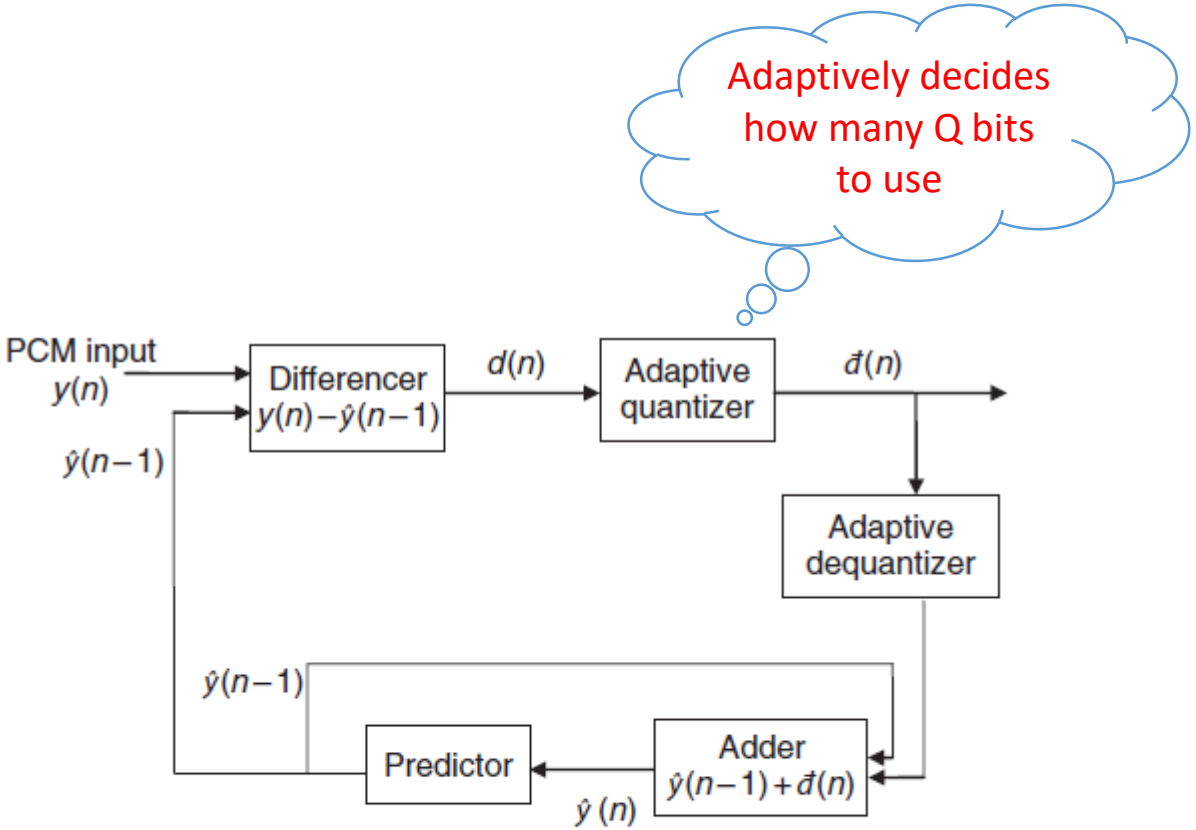



*Figure 9-4 Adaptive DPCM. The block diagram is similar to DPCM with an adaptive quantizer inserted after the differencer. This block adaptively decides how many quantization bits to use to quantize d(n) to đ(n).*

pakorn.wa@kmitl.ac.th

- ADPCM (Cont'd)
  - Have an additional complexity imposed by the need to insert control bits whenever the quantization bits change.
  - Frequent insertion of control bits can easily increase the entropy and the bit rate.
  - Most ADPCM use two modes for adaptive quantization
    - One for low frequency
    - The other for high frequency

pakorn.wa@kmitl.ac.th

- Logarithmic Quantization Scales
  - **A**-law (Europe)
  - $\mu$ law (N.America & Japan)
- Nonuniform quantization intervals are used in human speech compression standards.
- Audio speech signals tend to have a large dynamic range, up to 60 decibels,
  - Requiring a large number of quantization levels, typically 4096 levels, which require 13 bits.
  - The human ear, however, is more sensitive to lower amplitude intensities than to larger amplitude values.

- Better compression results are obtained by minimizing quantization errors at lower amplitudes than at higher amplitudes.
- To do this, the sampled signal at 13 bits is companded (logarithmically mapped) to 256 nonuniform 8-bit quantization levels.
- Two particular logarithmic quantization techniques for PCM audio are defined by the ITU (International Telecommunication Union) and are in worldwide use.
  - The American $\mu$ law standard, which compands 14 bits to 8 bits
  - The European A-law standard, which compands 13 bits to 8 bits.

pakorn.wa@kmitl.ac.th

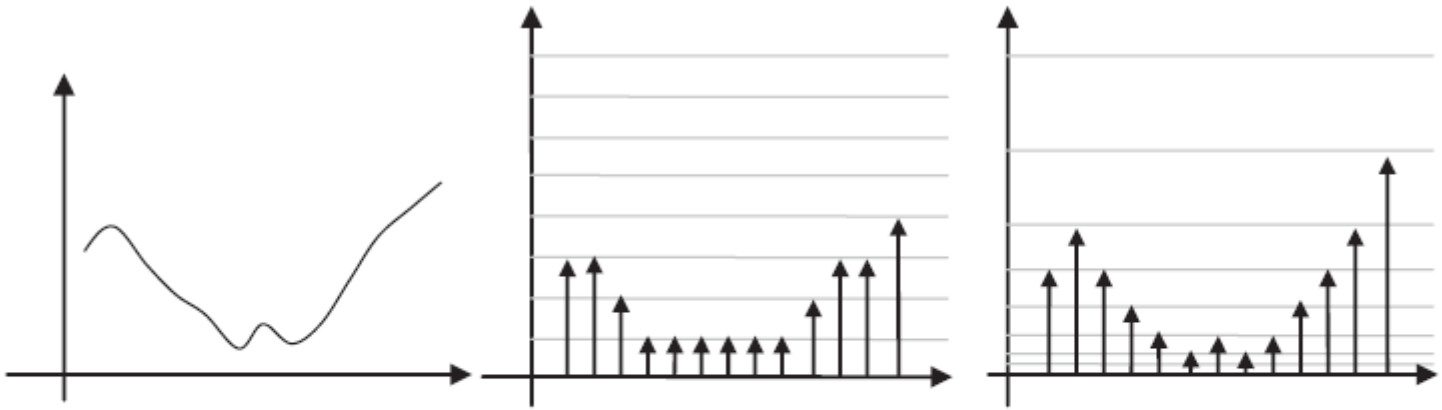
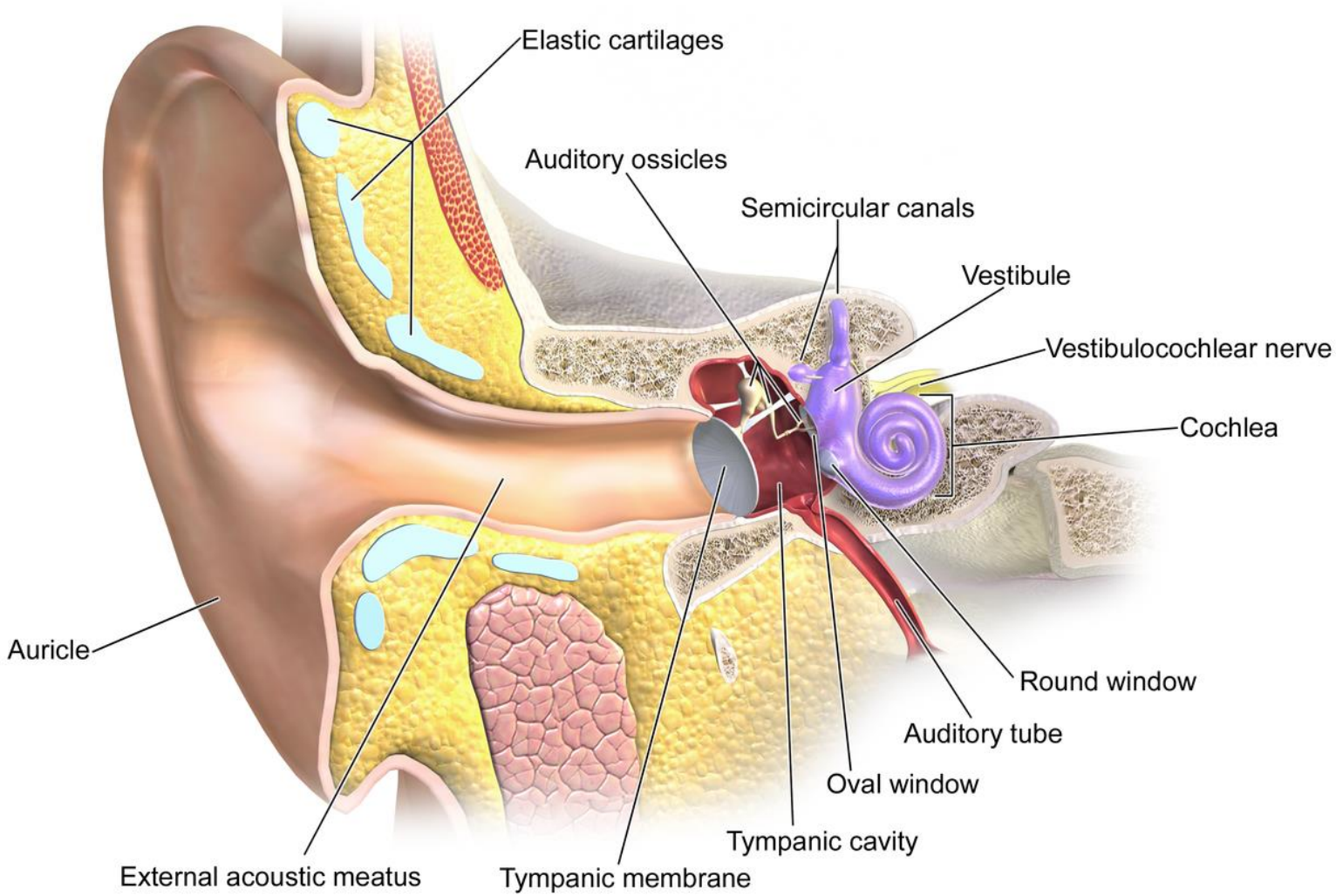


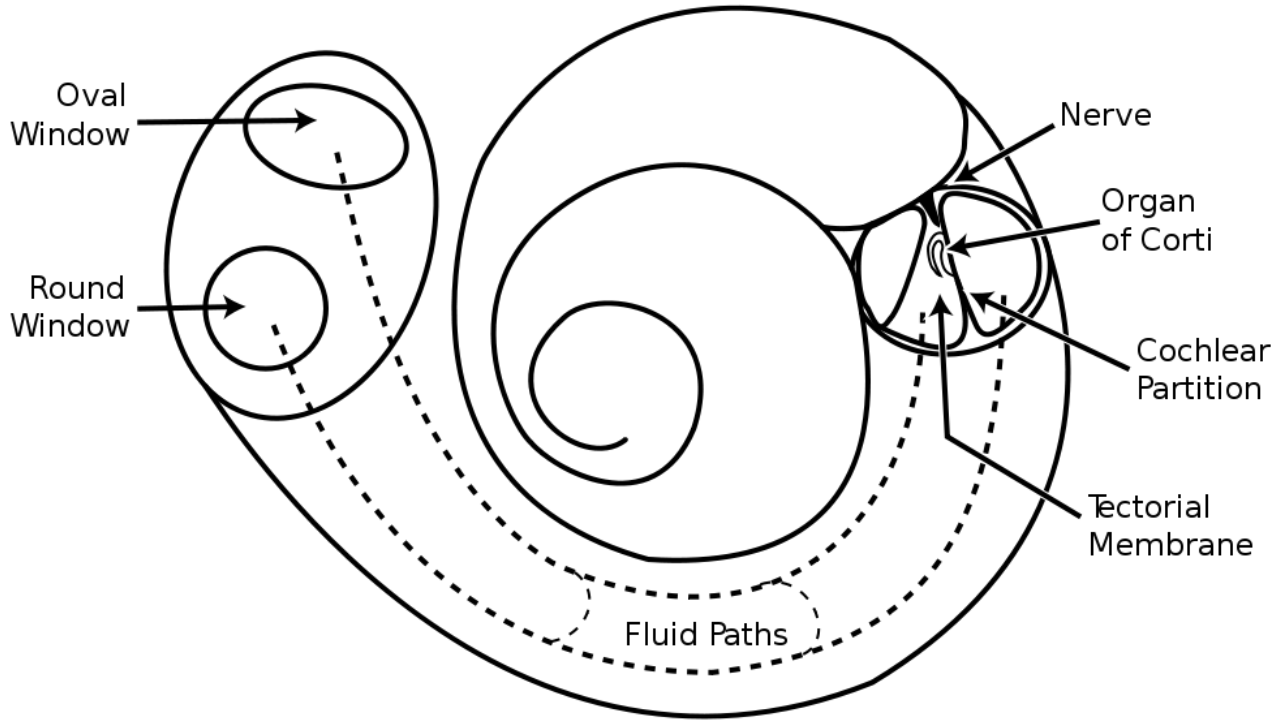
*Figure 9-5 Nonlinear quantization scales. The left image shows the original analog signal. The corresponding digitized signals are shown in the center and right. The center signal is obtained by uniform quantization intervals, whereas the far-right signal is obtained by a logarithmically quantized interval scale. The digitized signal on the right better approximates the original signal.*

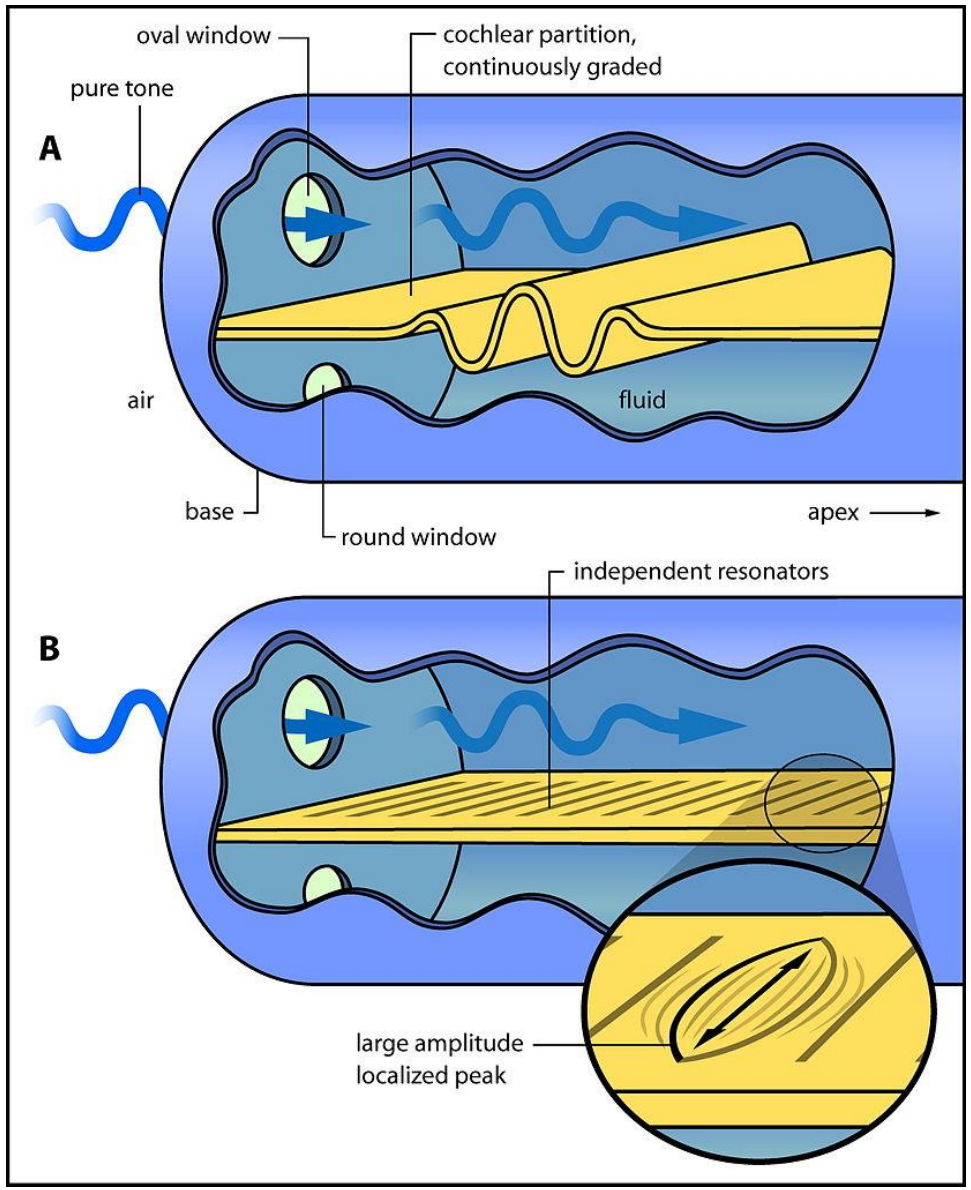
# Audio compression using psychoacoustics

KMITL
พระจอมเกล้าลาดกระบัง
pakorn.wa@kmitl.ac.th

- The ear is a physical structure and, thus, has its own limitations and artifacts.
- The understanding of the perceptual limitations of the ear can be exploited for audio compression by appropriately quantizing and even discarding parts of the signal that are not perceived by the human ear.


Elastic cartilages
Auditory ossicles
Semicircular canals
Vestibule
Vestibulocochlear nerve
Cochlea
Auricle
Round window
Auditory tube
Oval window
Tympanic cavity
External acoustic meatus
Tympanic membrane


Oval
Window
Round
Window
Nerve
Organ
of Corti
Cochlear
Partition
Tectorial
Membrane
Fluid Paths


oval window
cochlear partition,
continuously graded
pure tone
A
air
fluid
base
round window
apex
independent resonators
B
large amplitude
localized peak

pakorn.wa@kmitl.ac.th

- The human auditory system behaves very nonlinearly when it comes to perceiving specific frequencies due to
  - the shape of the cochlea
  - the nonuniform distribution of the sensory cells
  - the overall perceptual mechanism.

pakorn.wa@kmitl.ac.th

- Frequency domain limits
  - The human auditory system is capable of perceiving frequencies between 20 Hz and 20 KHz.
  - The dynamic range of hearing, defined as the ratio of the maximum sound amplitude to the quietest sound humans can hear is around 120 decibels.
  - Prior to compression, appropriate filters can be utilized to ensure that only the "audible frequency content" is input to the encoder.

pakorn.wa@kmitl.ac.th

- Time domain limits
  - A sound signal in the time domain can be decomposed into events that occur at specific times.
  - Two events can be uniquely perceived by the ear depending on how far apart in time they are separated.
  - Events separated by more than 30 milliseconds can be resolved separately.
  - The perception of simultaneous events (less than 30 milliseconds apart) is resolved in the frequency domain.

# Masking or Hiding

KMITL
พระจอมเกล้าลาดกระบัง
pakorn.wa@kmitl.ac.th

- Masking
  - If two sound tones or frequencies are present in the signal, the presence of one might hide the perception of the other, the result being that only one tone is perceived.
  - The audibility level of one frequency is affected by the presence of another neighborhood frequency.

KMITL
พระจอมเกล้าลาดกระบัง
pakorn.wa@kmitl.ac.th

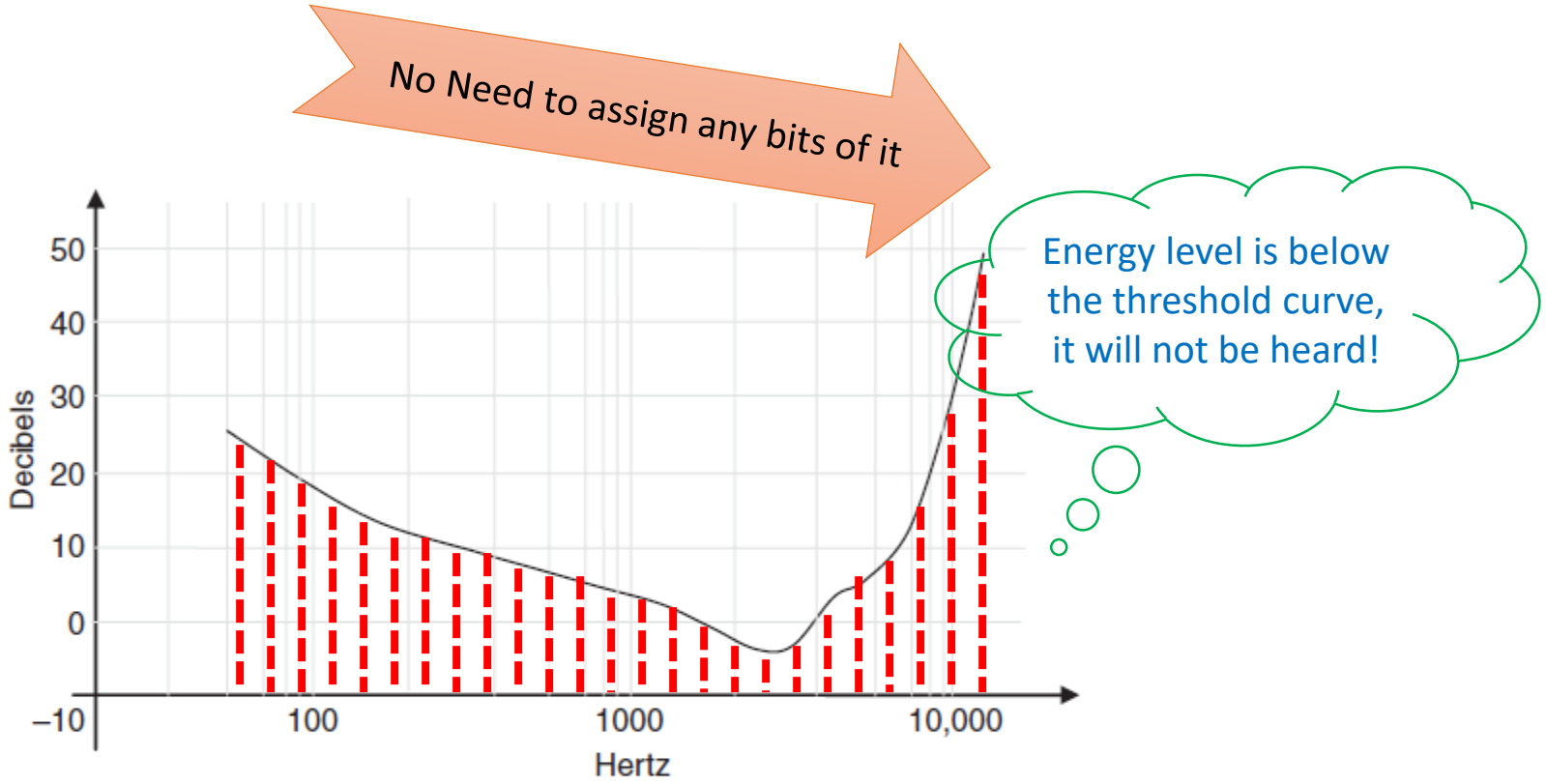
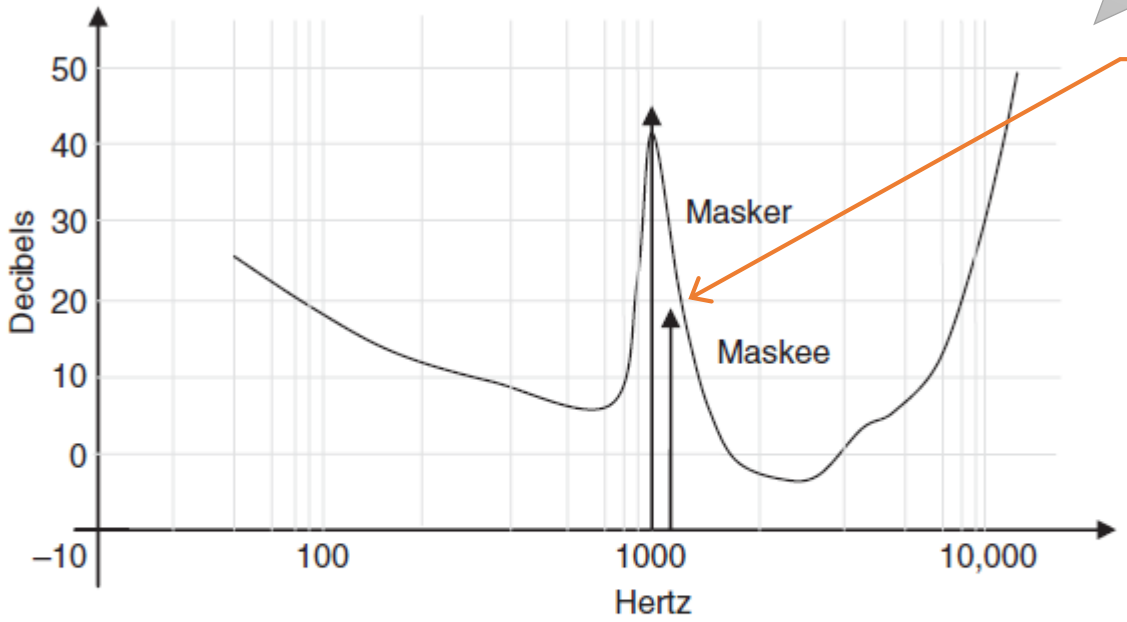



*Figure 9-7 Threshold curve for human hearing. The curve shows a plot for the audible level at all frequencies. The ear's response is not the same for all frequencies and changes nonlinearly. For clarity reasons, the graph plot uses a piecewise logarithmic frequency scale on the x-axis.*

KMITL
พระจอมเกล้าลาดกระบัง
pakorn.wa@kmitl.ac.th

No need to encode it




In the presence of the stronger 1 kHz tone, 1.1 kHz (that normally would be heard) will not be heard

*Figure 9-8 Threshold curve in the presence of a 1 KHz frequency at 30 dB. The curve has changed in relation to the threshold curve in the quiet of Figure 9-7. Also shown is a 1.1 KHz tone at 20 dB. Normally, this tone would have been heard, but now gets masked (under the curve) by the stronger 1 KHz tone.*

KMITL
พระจอมเกล้าลาดกระบัง
pakorn.wa@kmitl.ac.th

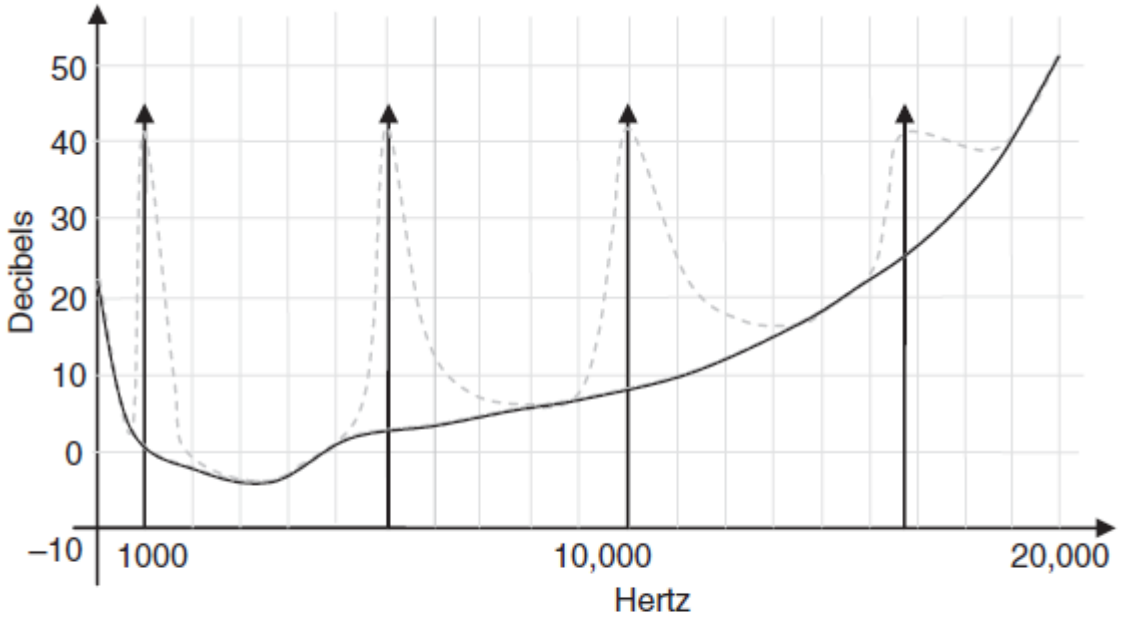



*Figure 9-9 Threshold curves for different individual frequencies. Four frequencies are shown, along with the way each one modifies the threshold curve. At the same amplitude level, the higher the frequency, the greater the area it masks.*

The higher the frequency of the masker, the broader the range of influence it has and more neighborhood frequencies it masks

KMITL
พระจอมเกล้าลาดกระบัง
pakorn.wa@kmitl.ac.th

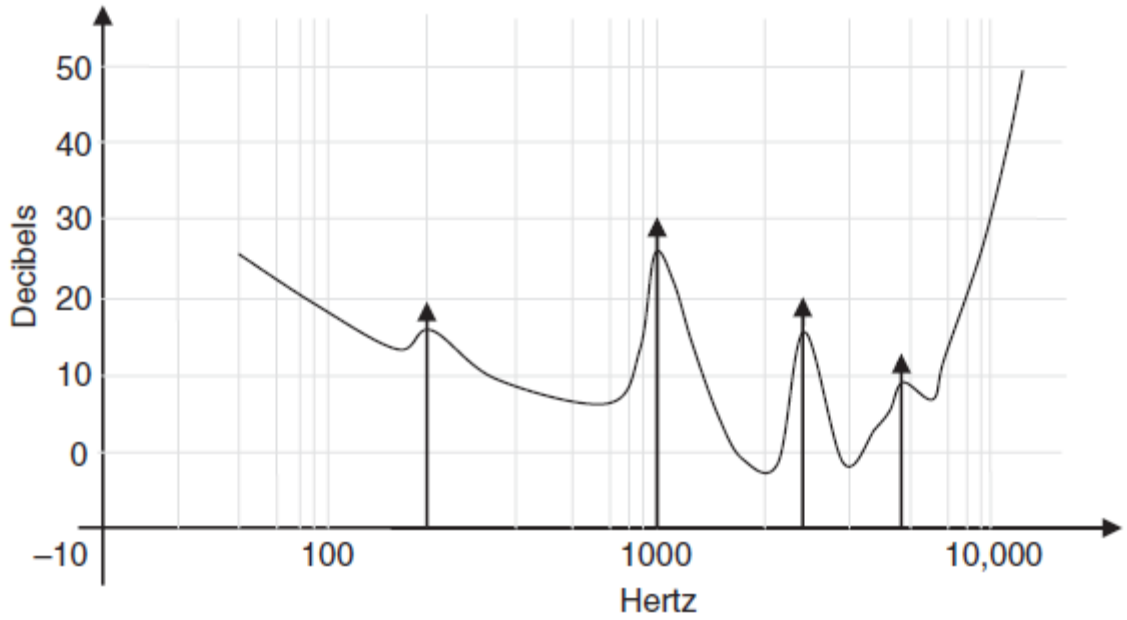



*Figure 9-10 Masking effects with multiple maskers. In the presence of multiple maskers, the threshold curve gets raised, masking a lot more frequencies.*

# Perceptual Encode

KMITL
พระจอมเกล้าลาดกระบัง
pakorn.wa@kmitl.ac.th

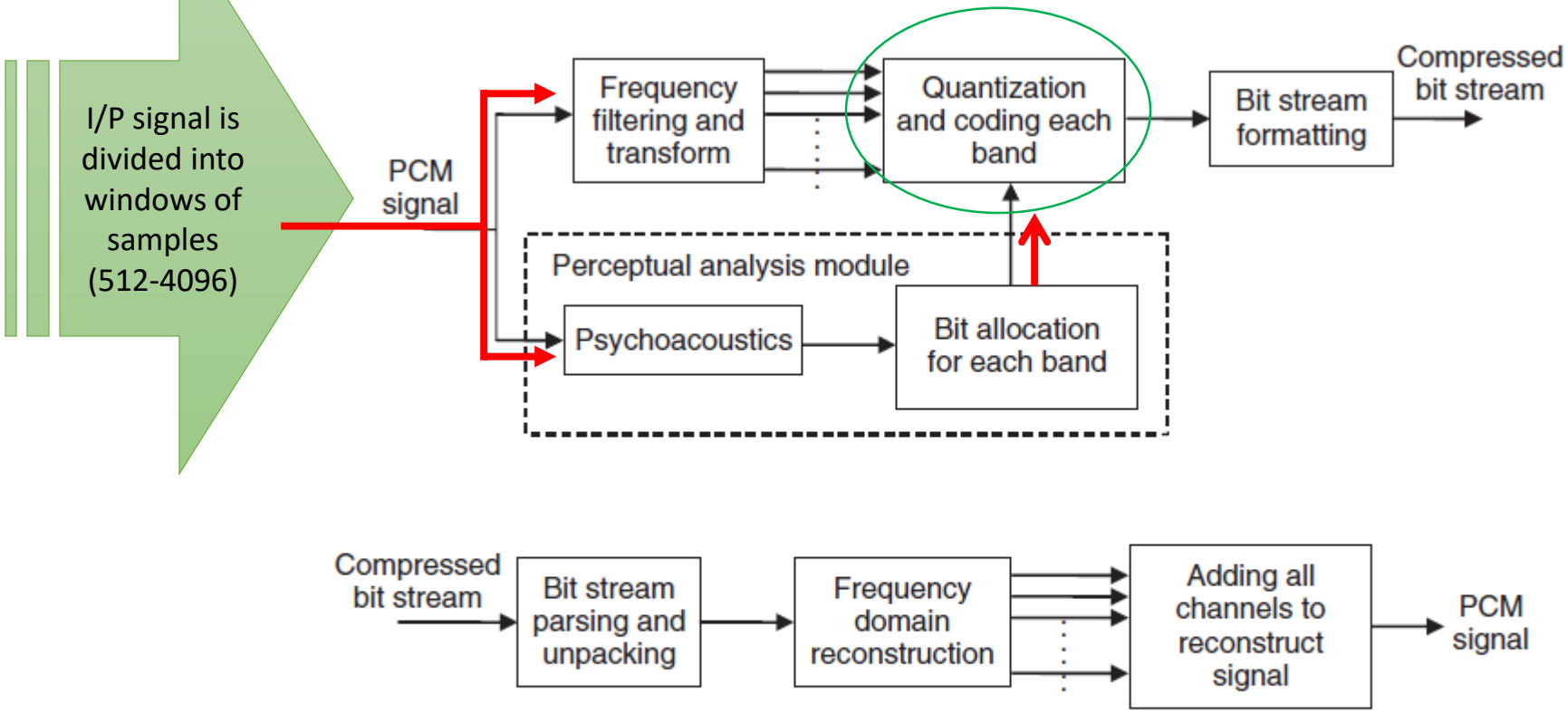



*Figure 9-11 Generic perceptual encoder (above) and decoder (below). The encoder is more complex than the decoder, and constantly analyzes incoming frequencies relative to an auditory perceptual model. Frequencies that are masked are assigned no bits or fewer bits.*

KMITL
พระจอมเกล้าลาดกระบัง
pakorn.wa@kmitl.ac.th

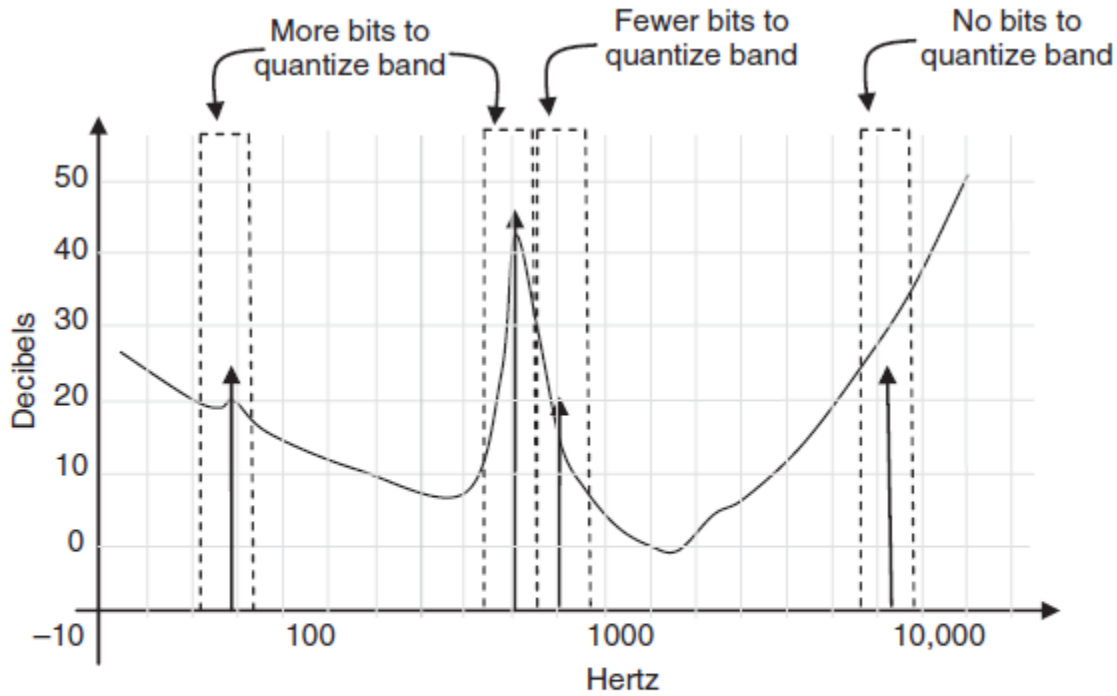



*Figure 9-12 Bit distribution among different frequency bands. The frequencies above the psychoacoustically modeled threshold curve need more bits compared with bands at the threshold boundary or those below the boundary.*

# Model-based Audio Compression

KMITL
พระจอมเกล้าลาดกระบัง
pakorn.wa@kmitl.ac.th

- Assumption
  - The sound source is well known and, hence, can be parameterized or modeled
  - The encoders – aka source codes – attempt to extract parameters of the model from the input signal that has to be coded.
  - The encoders communicate these parameters to the decoders.
    - E.g., synthesized sound – the pitch, amplitude levels, periodicity, and other intonations in the sound

pakorn.wa@kmitl.ac.th

- LPC : **Linear predictive coding**
  - An encoder: an analyzer
    - Take the speech signal
    - Break it into blocks or audio frames
    - Analyze each audio frame to determine
      - Voice (from the larynx)
      - Unvoice (silent)
  - A decoder: a synthesizer
    - Synthesizing the buzz from voiced segments and using random noise for the unvoiced segments

pakorn@live.kmitl.ac.th

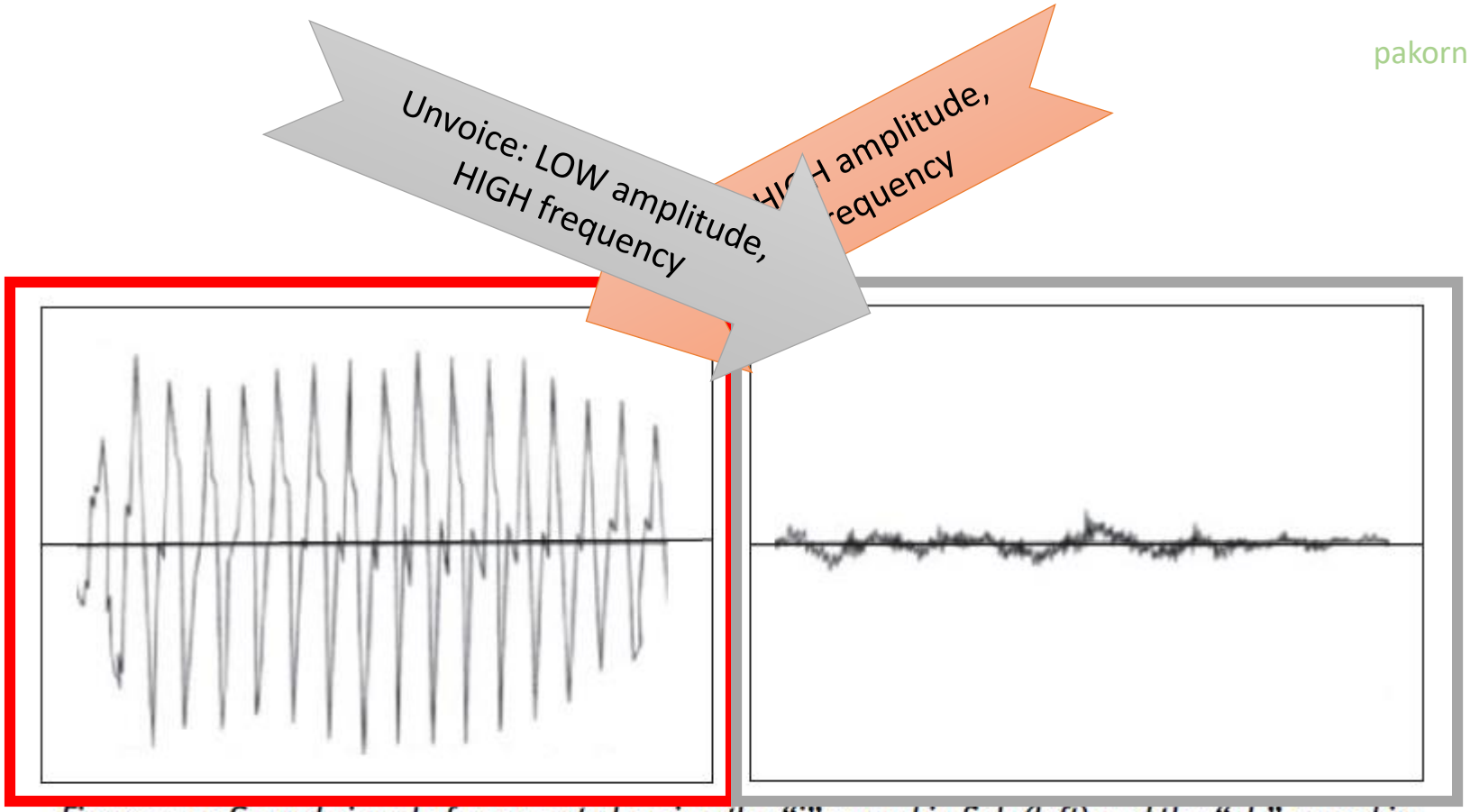



*Figure 9-13 Sound signal of segment showing the "i" sound in fish (left) and the "sh" sound in fish (right). The left sound is "voiced" from the larynx showing high amplitude and low frequency. The right sound is "unvoiced" characterized by higher frequency and low amplitude.*

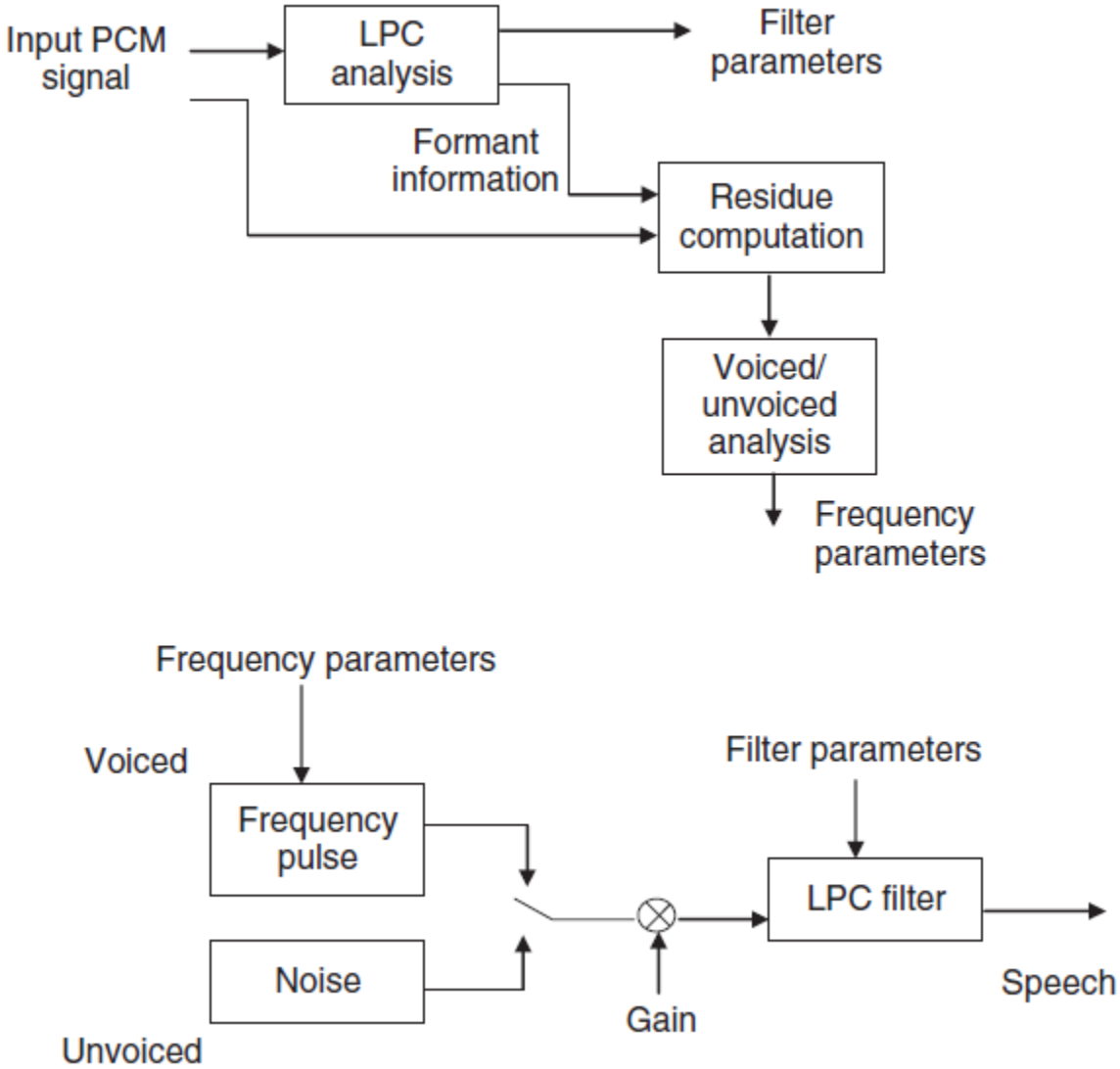



*Figure 9-14 Model-based voice compression using Linear Predictive Coding (LPC). The encoder (above) computes the filter parameters that model the voice tract using LPC. The base residue is analyzed for voiced and unvoiced segments resulting in frequency parameters for voiced segments. The decoder (below) uses the frequency parameters to synthesize voiced and unvoiced speech segments, which are then inverse filtered to produce speech.*

# Audio compression using event lists

KMITL
พระจอมเกล้าลาดกระบัง
pakorn.wa@kmitl.ac.th

- ***Describes*** the sound data using parameters that denote the music-making process as in a musical orchestra
  - The **sequence of notes** (frequencies) played, the **timing** between the notes, the **mixing effects**, the **timbre**, and so on.
  - Timbre describes the envelopes that encapsulate differences of the same sound played on different instruments.
- As such, these well-structured representations do not have any sound samples, but rather are **made up of semantic information** that can be interpreted by an external model to synthesize sounds
- E.g., MIDI (Musical Instrument Digital Interface), MPEG-4's SAOL (Structured Audio Orchestra Language),
  SASL (Structured Audio Score Language)

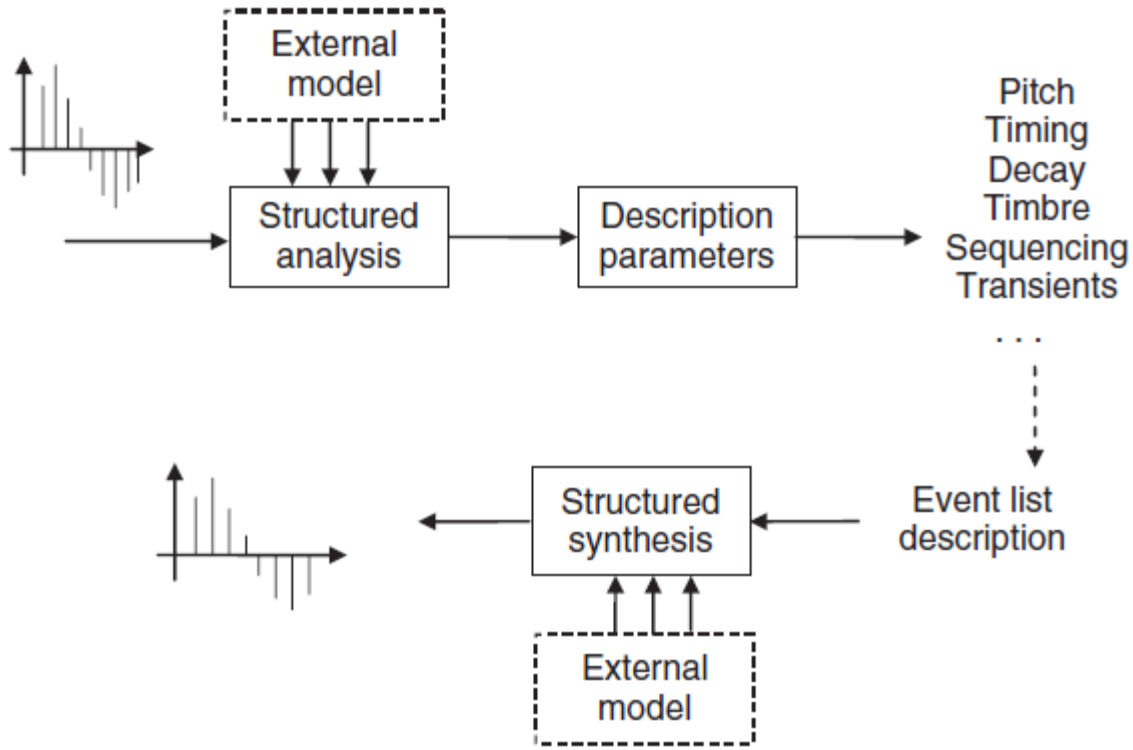



*Figure 9-15 Event list encoder (above) and decoder (below). The encoder analyzes the sound to extract semantic information to create a list of events, each represented by musical parameters, such as pitch, time of decay, tempo, and so on. Any structured audio decoder (MIDI) takes this description and synthesizes sound based on instrument models (such as a piano, guitar, and so on).*

# Structured representations and synthesis methodology

KMITL
พระจอมเกล้าลาดกระบัง
pakorn.wa@kmitl.ac.th

- Additive synthesis
  - Using the basic sinusoidal fundamental frequencies
  - Time-varying amplitudes and frequencies are superposed additively
- Subtractive synthesis
  - The output of a periodic oscillator that uses fundamental sinusoids (or others) is passed through filters and amplifiers.
- Frequency modulation
  - A host frequency gets modulated by a modulating frequency
- Physical modeling
  - Emulate the actual physical structure of the source and the sound

# Advantage of structured audio

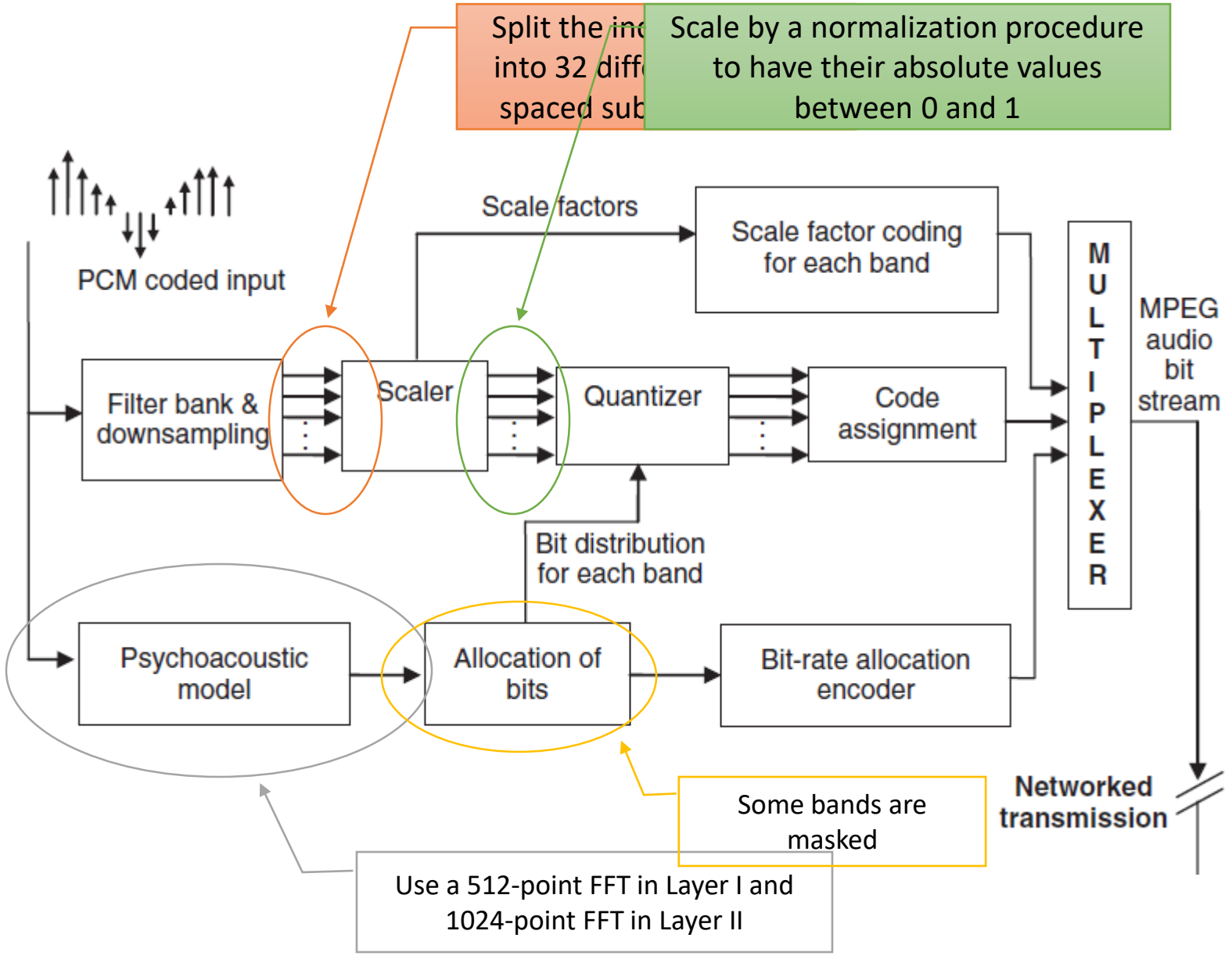
KMITL
พระจอมเกล้าลาดกระบัง
pakorn.wa@kmitl.ac.th

- Low-bandwidth transmission
- Musical synthesis and creation
- Parametric sound generation
- Interactive music applications
- Audio content-based retrieval

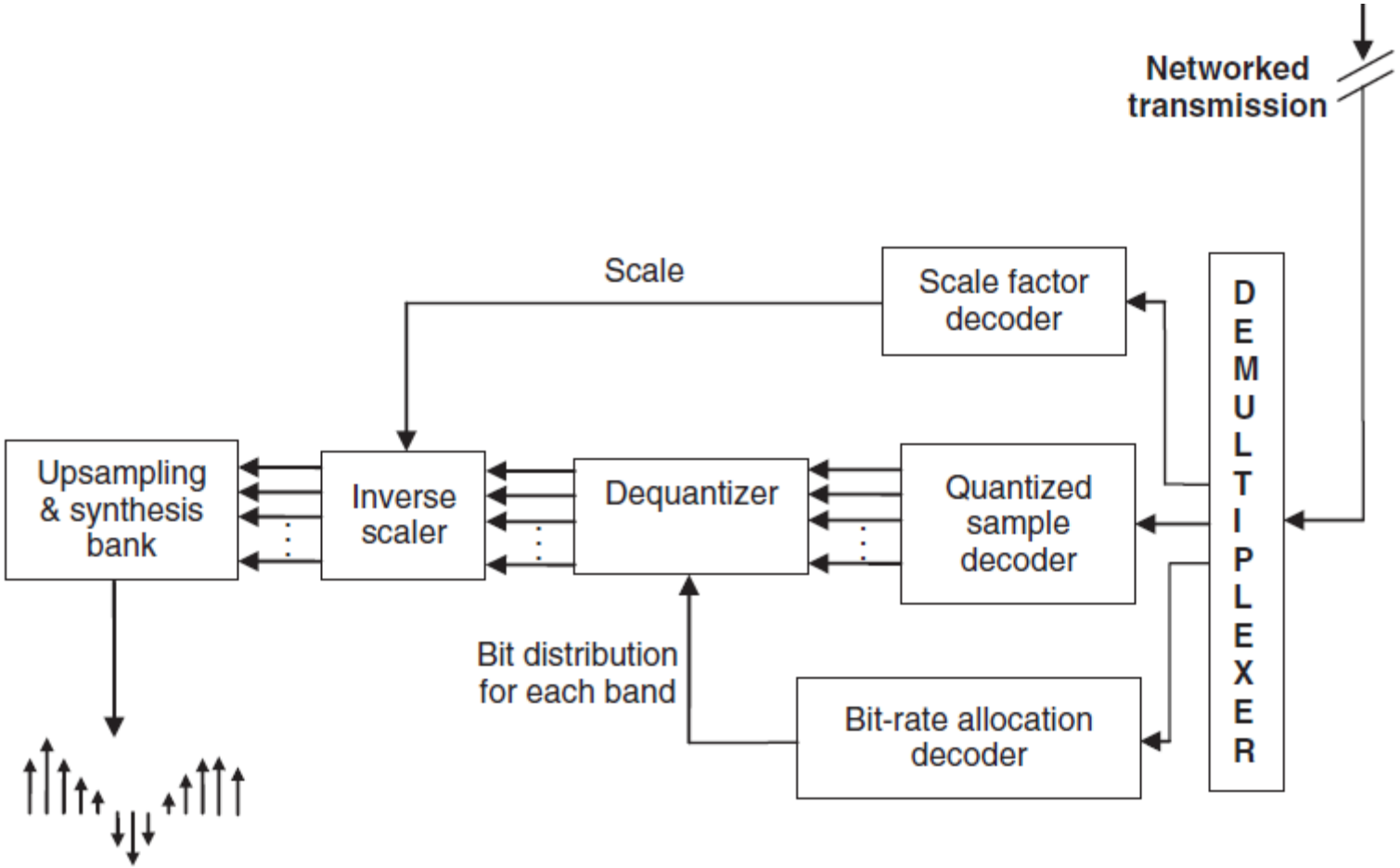
# Audio coding standards

KMITL
พระจอมเกล้าลาดกระบัง
pakorn.wa@kmitl.ac.th

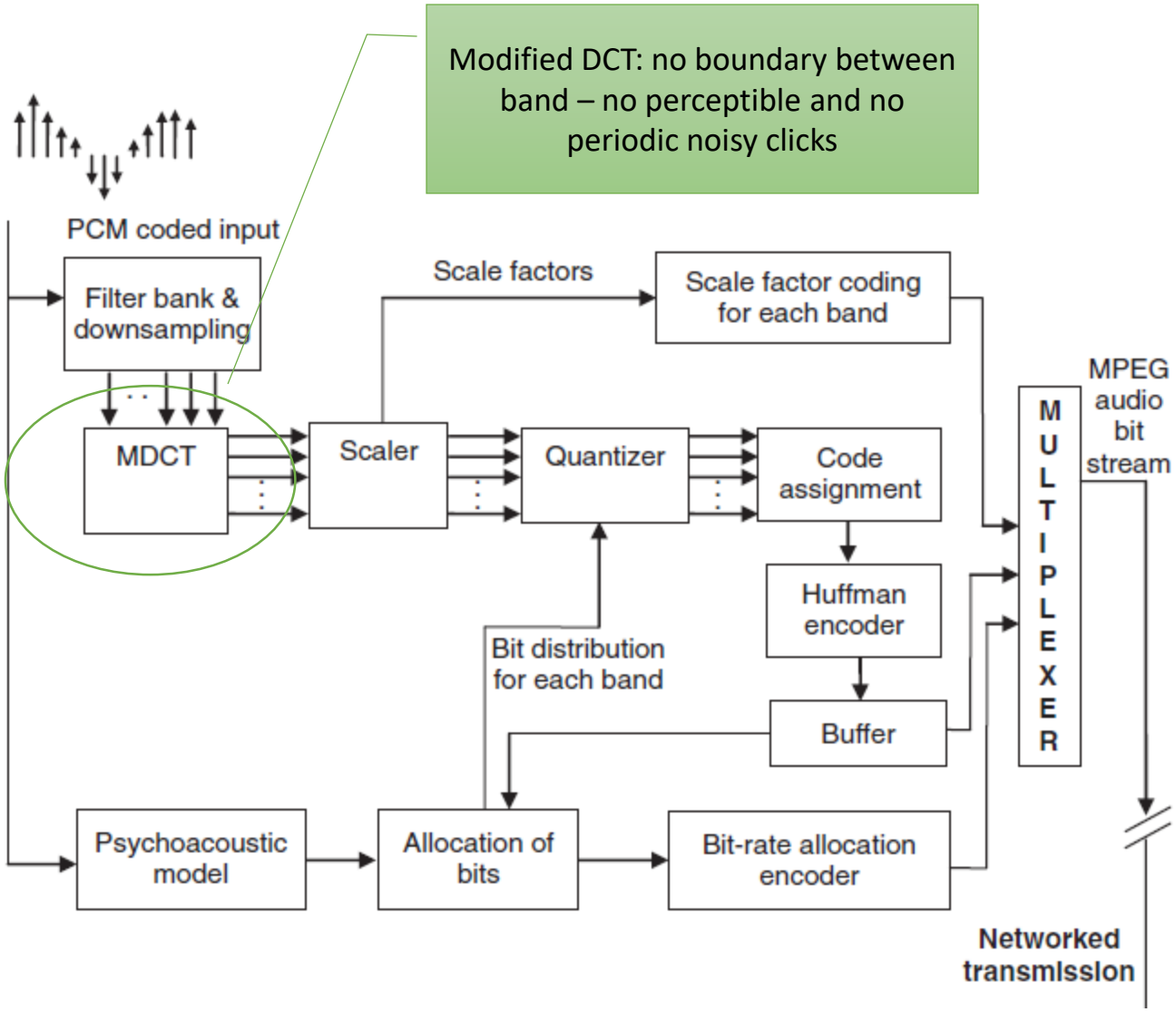
- MPEG-1
  - MPEG-1 Layer I
  - MPEG-1 Layer II
  - MPEG-1 Layer III (MP3)




MPEG audio architecture for Layer I and Layer II




MPEG audio architecture for Layer I and Layer II




MPEG audio architecture for Layer III

# MPEG-2

KMITL
พระจอมเกล้าลาดกระบัง
pakorn.wa@kmitl.ac.th

- Support a surround sound 5.1 channel input
- Categorized into 2 groups
  - MPEG-2 BC (backward compatible with MPEG-1)
  - AAC (advanced audio coding; aka. NBC [non-backward compatible]) used in DVD, XM radio

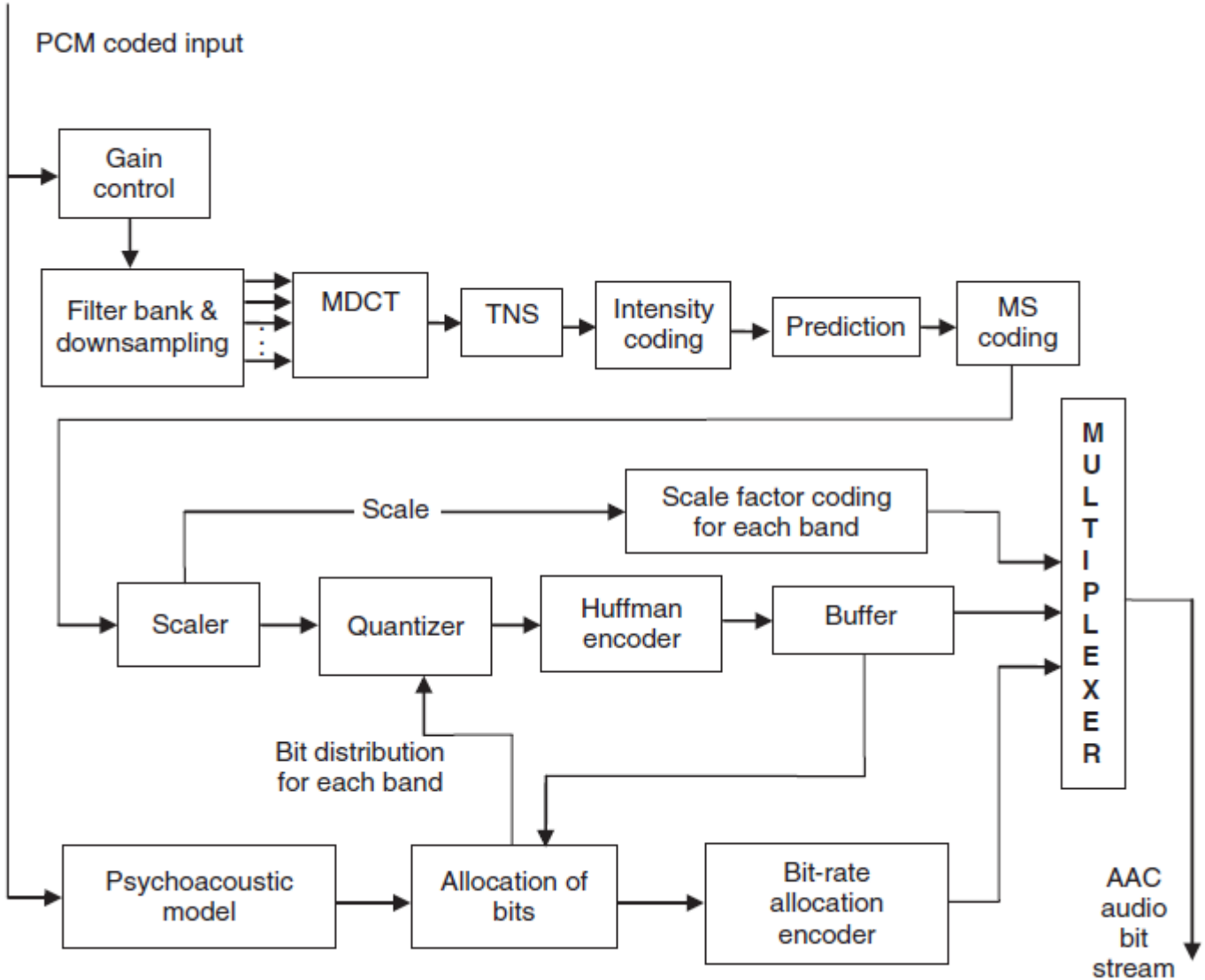
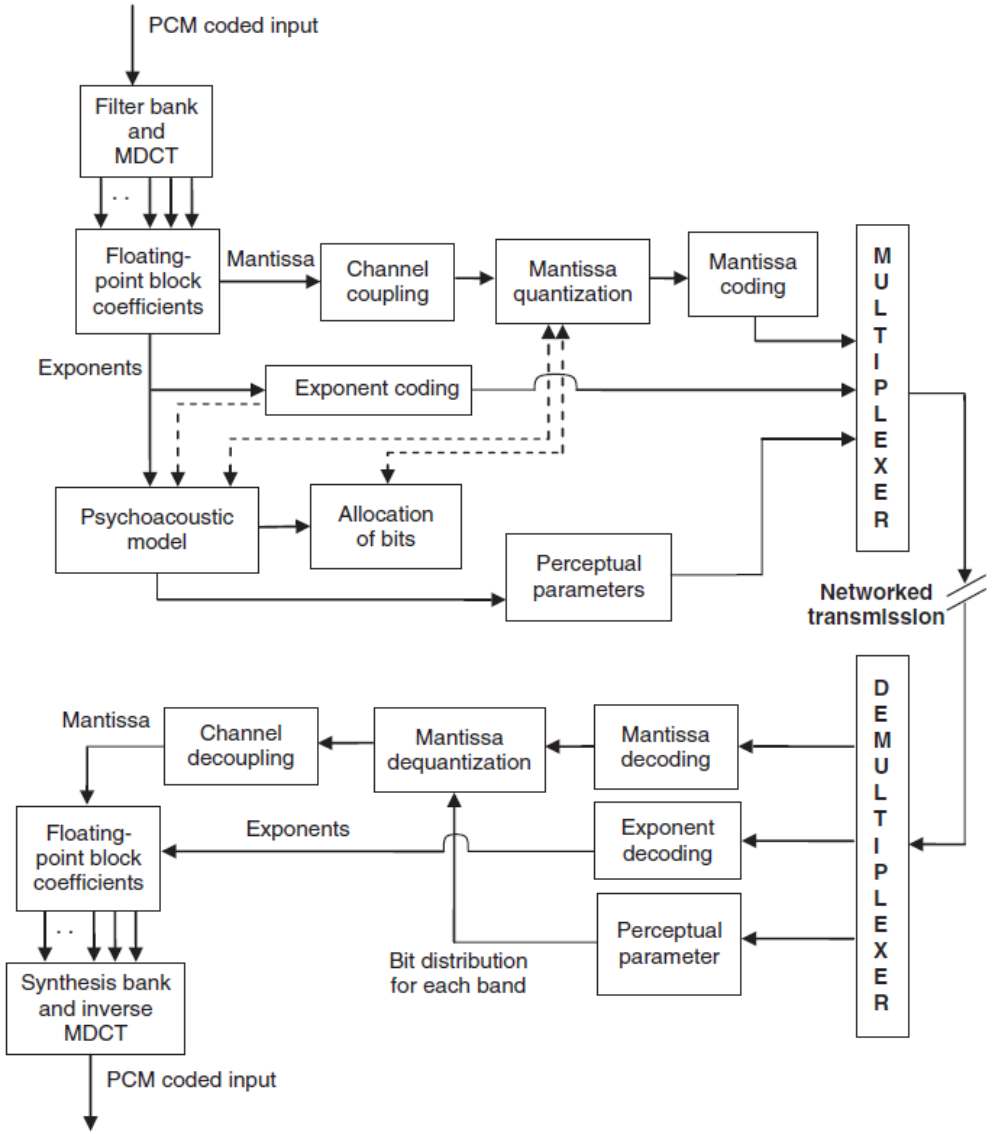



*Figure 9-18 MPEG-2 AAC encoder. Most of the modules, such as the psychoacoustic, quantizer, and Huffman, are similar to the MPEG-1 layers. AAC gains more compression by temporal noise shaping (TNS), stereo intensity coding, prediction, and MS coding modules.*

# Dolby AC-2 and AC-3

KMITL
พระจอมเกล้าลาดกระบัง
pakorn.wa@kmitl.ac.th



# Other standards

KMITL
พระจอมเกล้าลาดกระบัง
pakorn.wa@kmitl.ac.th

- MPEG-4
- ITU G.711 – telecommunication networks
- ITU G.722 – videoconferencing systems
- ITU G.721, ITU G.726, ITU G.727 – telephone bandwidth speech using adaptive differential pulse code modulation techniques
- ITU G.723, ITU G.729 – model-based coders with special models of speech production
- ITU G.728 – hybrid between the higher bit rate ADPCM coders (G.726 and G.727) and the lower bit rate model-based coders (G.723 and G.729)
- MIDI: Musical Instrument Digital Interface

KMITL
พระจอมเกล้าลาดกระบัง
pakorn.wa@kmitl.ac.th

# Q & A