ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Multimedia Technology

Section # 8: Audio Coding

Instructor: George Xylomenos

Department: Informatics

Contents

- Channel coding for voice
- Source coding for voice
- Perceptual coding
- MPEG-1 audio coding
- MPEG-2 audio coding
- MPEG-4 audio coding

Channel coding for voice

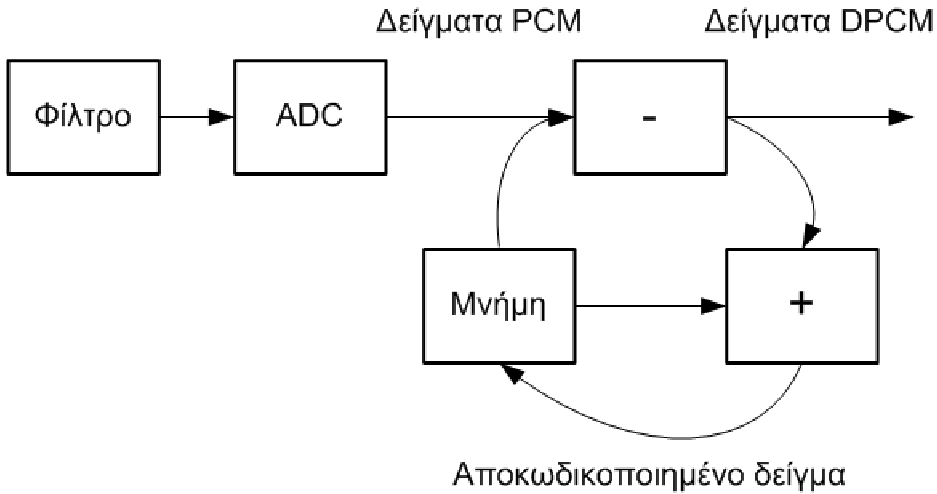
Class: Multimedia Technology, **Section # 8:** Audio Coding

Instructor: George Xylomenos, **Department:** Informatics

G.711

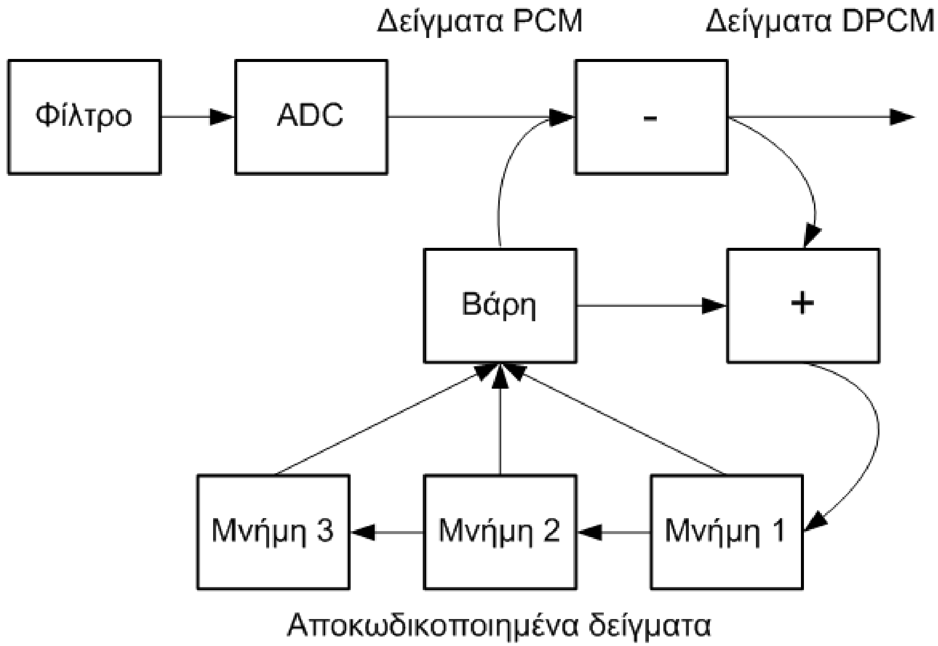
- The ITU G.711 standard
  - Used in POTS (Plain Old Telephony Service)
  - Filters frequencies in the 300-3400 Hz band
  - Logarithmic sampling (“compression”)
  - 8 KHz sampling rate (2x4 KHz, with guard bands)
  - 64 Kbps final bitrate
- Serves as the base for many other standards
  - The G series channel coders

Channel coding (1 of 5)



- DPCM coding
  - Encode differences instead of samples
    - Send an approximation of the difference
  - Reference: the previous approximation
    - Note that decoding loop at the encoder

Channel coding (2 of 5)



- DPCM with linear prediction
  - Linear combination of previous values
  - Better prediction with fewer bits

Channel coding (3 of 5)

- Adaptive DPCM
  - Multiple values used for prediction
  - Allows changing the quantization step
- G.721: G.711 quality at 32 kbps (DECT)
  - Uses 8 previous values for prediction
  - G.723: similar for 24 and 40 kbps
- G.726: extends/unifies G.721 and G.723
  - Supports 16, 24, 32 and 40 Kbps

Channel coding (4 of 5)

- How to vary the quantization step?
  - The basic step is multiplied by a factor μ
  - We monitor the differences
  - Closer to 0: more detail
    - Reduce factor μ
  - Further from 0: less detail
    - Increase factor μ

Channel coding (5 of 5)

- G.722: 64 kbps but for 7 KHz (HD Voice)
  - Splits voice in two frequency bands
  - Uses ADPCM on each band
  - 0-3,5 KHz: assigned 48 kbps
    - Quality similar to POTS (G.711)
  - 3,5-7 KHz: assigned 16 kbps
    - Adds higher frequencies
    - More natural results

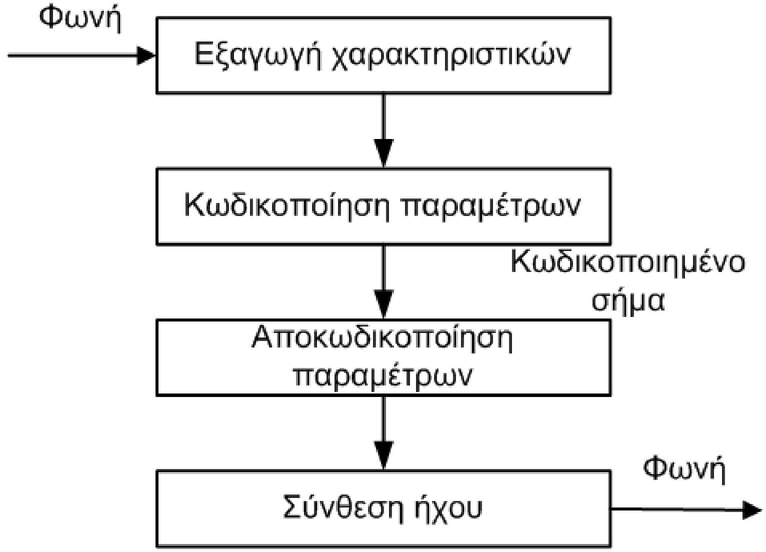
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Source coding for voice

Class: Multimedia Technology, **Section # 8:** Audio Coding

Instructor: George Xylomenos, **Department:** Informatics

Source coding (1 of 8)



- Voice coders (vocoders)
  - Based on a model of the human voice
  - Extracts and transmits voice characteristics
    - Characteristics: parameters of the voice model

Source coding (2 of 8)

- Phonemes: basic sounds of a language
- Voiced sounds
  - Those produced by the vocal chords
  - Vowels and some consonants (e.g., a, b)
  - Their waveforms are periodic
- Unvoiced/Voiceless sounds
  - All the other consonants (e.g., p)
  - Their waveforms look like noise

Source coding (3 of 8)

- Formants
  - Frequencies with peak energy
  - Each phoneme has specific formants (2 or 3)
    - Modulated by chords, mouth, tongue
  - Can be detected via filters
    - We analyze a “frame” of samples (a fixed number)
    - We detect the peak frequencies (formants)
    - We determine the underlying phoneme

Source coding (4 of 8)

- Linear Predictive Coding (LPC)
  - Analyzes a frame of samples
    - Not necessarily the same length as a phoneme
  - Is the sound voiced or unvoiced?
    - Voiced: simulate with frequency generator and filters
    - Unvoiced: simulate noise generator
  - Use old parameters to predict new ones
    - Liner prediction from previous parameters
    - Calculation of differences to send

Source coding (5 of 8)

- Linear Predictive Coding (LPC)
  - Decoding
    - Calculate linear prediction from past values
    - Add the transmitted difference
  - LPC-10: linear combination of 10 frames
    - Frame: 180 samples at 8 kHz = 22.5 ms
    - Bitrates as low as 2.4 Kbps
    - Recognizable but robotic voice

Source coding (6 of 8)

- Code excited LPC (CELP)
  - Two levels of prediction
    - Short term at the sample level (STP)
    - Long term at the frame level (LTP)
  - Use a library of existing characteristics
    - Each one is a set of encoder parameters
    - Find the best match with the current frame
    - Essentially, this is vector quantization

Source coding (7 of 8)

- Code excited LPC (CELP)
  - Add new characteristics to library
    - Use previous predictions for LTP
- G.723.1: 5.3 or 6.3 kbps
  - Used for teleconference over PSTN
  - 8 kHz sampling at 16 bit, 30 ms frames
    - Broken down into 4 subframes for prediction

Source coding (8 of 8)

- G.728: 16 kbps
  - Used for conferencing over ISDN
  - Lower delay compared to plain CELP
- G.729: 8 kbps
  - Used in cellular telephony
    - G.729a used in GSM (2G)
  - 10 ms frame to reduce delay
  - Protection from parameter loss (due to wireless)

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Perceptual coding

Class: Multimedia Technology, **Section # 8:** Audio Coding

Instructor: George Xylomenos, **Department:** Informatics

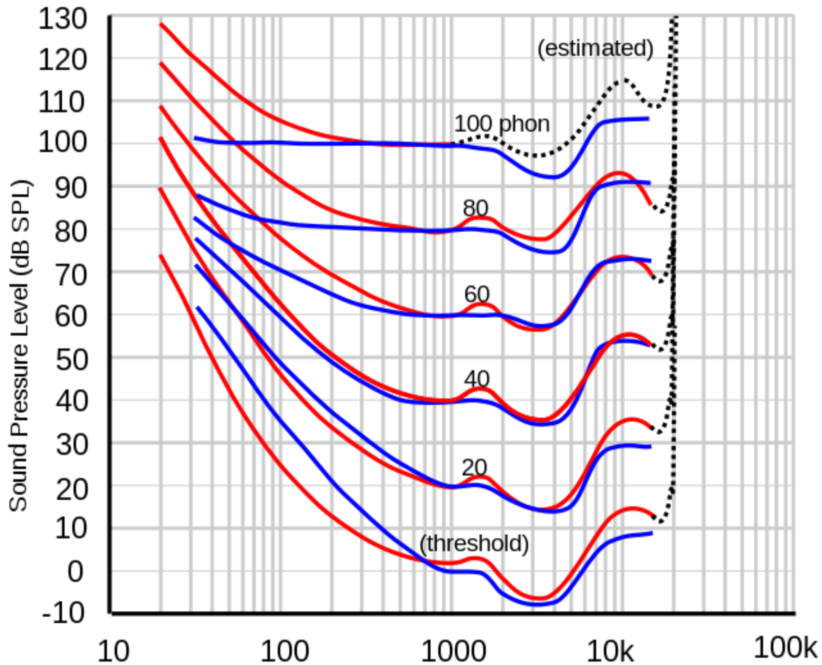
Encoding of random sounds

- Voice is a very specific type of sound
  - Can be modeled in a specific way
  - Specific phonemes and formants, periodicity
- How do we encode any type of sound?
  - For example, music
  - Source coding is infeasible
    - We cannot assume specific sources
  - But we do know how people hear!

Human hearing (1 of 2)

- How loud is a sound?
  - Sound pressure level (SPL): measured in dB
    - Relative to the threshold of human hearing
    - This is an objective measure
  - Human perception depends on frequency
  - phon: perception of intensity by humans
    - 1 phon = loudness of 1 dB SPL at 1kHz
    - This is a subjective measure

Human hearing (2 of 2)



Equal-loudness contours (red) (from ISO 226:2003 revision)
Fletcher–Munson curves shown (blue) for comparison

- Fletcher-Munson curves
  - Human ears are most sensitive at 2-5 KHz

Psychoacoustic model (1 of 4)

- Perceptual coding
  - Exploits the psychoacoustic model
  - Looks for less important frequencies
  - Encodes them with fewer bits (or not at all)
- Two basic techniques
  - Frequency masking
  - Temporal masking

Psychoacoustic model (2 of 4)

- Frequency masking
  - Loud signals reduce the dynamic range
    - Increase threshold of hearing in nearby frequencies
    - The masking effect is frequency dependent
    - Most evident at higher frequencies
  - Division of the spectrum in critical bands
    - The ear does not distinguish frequencies in a band
    - The width of these bands grows with frequency

Psychoacoustic model (3 of 4)

- Temporal masking
  - Loud signals affect hearing for some time
    - Increase threshold of hearing in nearby frequencies
    - Effect is reduced with time
    - The masking effect is frequency dependent
  - Two dimensional masking curve
    - Temporal masking
    - Frequency masking

Psychoacoustic model (4 of 4)

- Exploiting masking
  - Start with a frame (consecutive samples)
  - Break down the signal by frequency range
  - In each range locate the louder signals
  - Calculate the masking effects
    - Each range has different characteristics
  - Isolate the less audible signals
    - Encode them with fewer bits

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

MPEG-1 audio coding

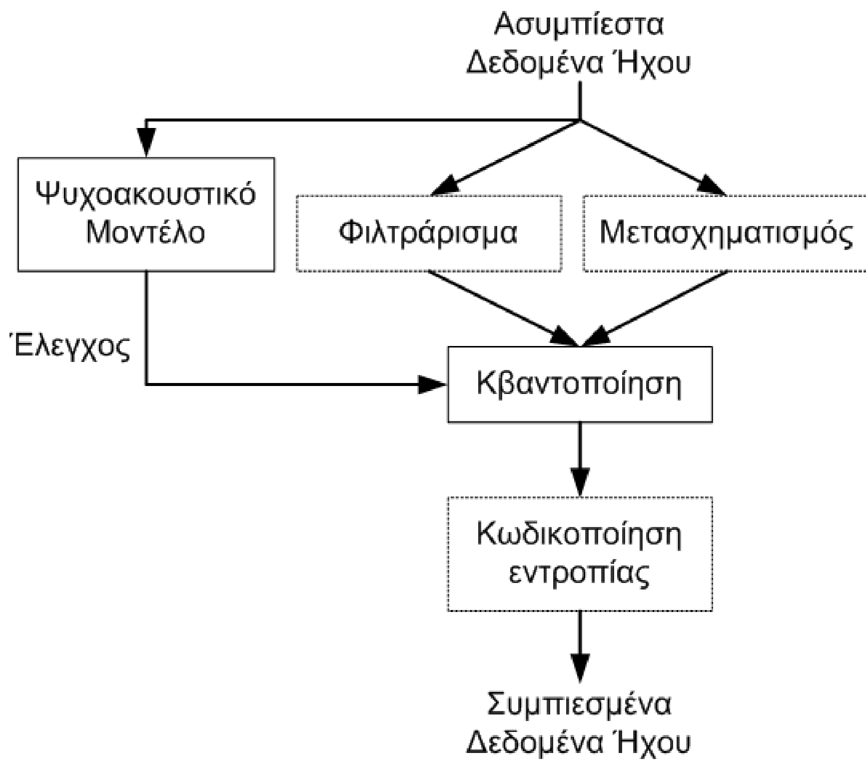
Class: Multimedia Technology, **Section # 8:** Audio Coding

Instructor: George Xylomenos, **Department:** Informatics

MPEG-1 Audio (1 of 7)

- MPEG Audio Layer 1, 2, 3
  - Standardized as part of MPEG-1
  - Originally used in Video-CD (predates DVD)
- Three levels, backward compatibility
  - Level 3 is the most popular (MP3)
  - Higher complexity and latency
- Signal digitization in MPEG-1 Audio
  - 48, 44,1 or 32 KHz, 16 bit samples per channel

MPEG-1 Audio (2 of 7)



- Compression based on psychoacoustic model
  - Controls coding rate based on model

MPEG-1 Audio (3 of 7)

- Basic coding: Layer 1
  - Starts with 384 audio samples
  - Uses a filter bank to break down the signal
    - 32 equal width frequency bands
      - Some overlap between bands
    - They do not grow with frequency
      - Unlike critical bands
    - Considers 12 samples per band

MPEG-1 Audio (4 of 7)

- Psychoacoustic model
  - Locates loudest signal in each band
  - Estimates how important each band/sample is
    - 1024-point Fourier transform
  - Assigns bits based on importance
- Quantization
  - Linear quantization of the coefficients
  - Scaling factor used to control quantization
    - Goal: produce a fixed output bit rate

MPEG-1 Audio (5 of 7)

- Intermediate coding: Layer 2
  - Uses three frames in each repetition (1152 samples)
    - Increases latency, but has better accuracy
    - Also exploits temporal masking
- Advanced coding: Layer 3
  - Unequal frequency bands (like critical bands)
  - Modified Discrete Cosine Transform (MDCT)
    - Better at masking than Fourier transform
  - Non-linear quantization

MPEG-1 Audio (6 of 7)

- Entropy coding in the final stage
  - MP1/2: simple PCM
  - MP3: Each pair of coefficients is Huffman coded
    - Huffman tree selected based on input
- Dual adaptive loop (MP3)
  - Internal: based on entropy coding
    - Modifies quantization step to achieve bit rate
  - External: based on noise per band
    - Modifies quantization factors per band

MPEG-1 Audio (7 of 7)

- Final coding
  - Level 1 and 2: constant bit rate
  - Level 3: optional variable bit bit
    - Can change in every audio frame
- Bit rate: at least 32 Kbps
  - Level 1: Up to 448 Kbps
  - Level 2: Up to 384 Kbps
  - Level 3: Up to 320 Kbps
    - The quality of each rate depends on the level

Stereo audio (1 of 3)

- Stereo audio
  - Two audio channels for more realistic sound
    - Microphones / speakers at different locations
  - People perceive stereo in two ways
    - Differences in timing
    - Differences in loudness
- MPEG-1 stereo coding
  - Independent or joint (exploits commonality)

Stereo audio (2 of 3)

- Intensity joint stereo coding
  - Lower frequencies
    - We mostly perceive timing differences
  - Higher frequencies
    - We mostly perceive loudness differences
  - Mixes left and right channels
    - Adds information for the intensity per channel

Stereo audio (3 of 3)

- Mid-side joint stereo coding
  - The central channel is the sum
    - M=L+R
  - The side channel is their difference
    - S=L-R
    - Can be encoded with fewer bits
  - Transform to original channels
    - L=(M+S)/2, R=(M-S)/2

MPEG-1 bit stream (1 of 2)

- MPEG-1 bit stream
  - MP3 files may have a header
    - Depends on file format, not the standard
  - The bit stream is divided into frames
    - 24 ms of audio at 48 KHz
  - Each frame has a header
    - Allows decoding from the beginning of the frame
  - Timing word: check for periodic appearance
    - May also appear inside the data

MPEG-1 bit stream (2 of 2)

- Frame header
  - Bit rate: can be changed per frame
  - Sampling rate: can be changed per frame
  - Level: 1, 2, 3 or variants
  - Coding type
    - Stereo, joint stereo, etc
  - Protection bits: rarely used

ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

MPEG-2 audio coding

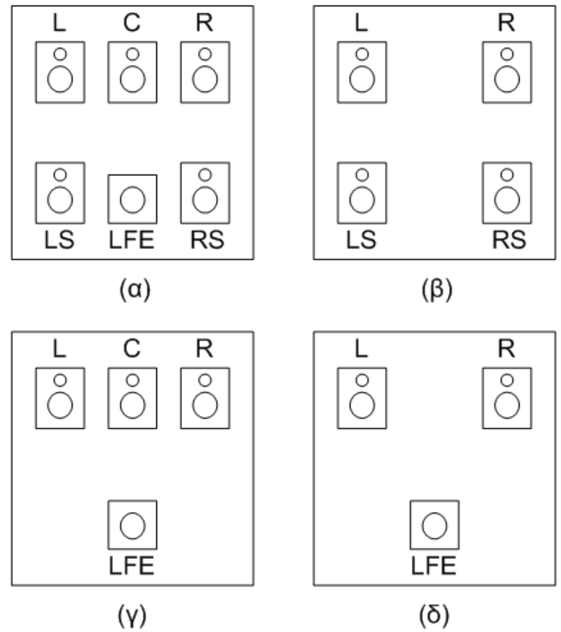
Class: Multimedia Technology, **Section # 8:** Audio Coding

Instructor: George Xylomenos, **Department:** Informatics

MPEG-2 Additions (1 of 2)

- MPEG-2: used in DVDs
  - Allows multichannel sound
  - Up to five full range channels
    - Central, front L/R, peripheral L/R
  - Low Frequency Enhancement (LFE) for 15-120 Hz
  - Different combinations are possible
  - Multiple audio tracks (dubbing, commentaries)
  - Offers 5.1 cinematic sound

MPEG-2 Additions (2 of 2)



- Other MPEG-2 extensions
  - Extended bitrates: 8-96 kHz
  - Works well with 64 Kbps per channel

Limitations of MPEG-1 (1 of 2)

- Distortion in digital audio
  - Very different from distortion in analog audio
- Loss of quality
  - In any frequency band
    - Unlike analog harmonic distortion
  - Can change in each frame
  - Some frequency ranges can disappear
    - Some coefficients are drop to achieve bit rate

Limitations of MPEG-1 (2 of 2)

- Pre-echo
  - Sudden change within a frame
  - Causes distortion in the entire frame
    - Due to lack of sufficient bits
- Double speak
  - Different sound and coding periods
  - Speech is periodic
  - Can be distorted by coding

MPEG-2 AAC (1 of 4)

- MPEG-2 Advanced Audio Coding (AAC)
  - New coded for MPEG-2
  - More efficient than MPEG-1
    - Reduces bit rate by 30% for the same quality
  - Not backwards compatible
    - Same basic idea, but too many modifications
  - Main audio codec for MPEG-4

MPEG-2 AAC (2 of 4)

- Coding improvements
  - 256 or 2048 samples per frame
    - Uses only MDCT, not filter banks
    - 256 samples: lower latency
    - 2048 samples: lower bit rate
  - Splits coefficients into 49 bands
    - Similar to critical bands
  - Coefficient prediction within each bank

MPEG-2 AAC (3 of 4)

- Coding improvements
  - Improved joint stereo coding
  - Huffman coding of four coefficients at a time
- Quality improvements
  - Reduced pre-echo
    - Due to smaller frames
  - Temporal Noise Shaping (TNS)
    - Prevents double speak phenomenon

MPEG-2 AAC (4 of 4)

- AAC bit stream: two options
  - Audio Data Interchange Format (ADIF)
    - All information in a single header
    - Decoding must start at beginning of file
  - Audio Data Transport Stream (ADTS)
    - Per frame headers
    - Similar to MPEG-1
    - Also allows variable length frames
    - Level 4 in the header (MP4!)

ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

MPEG-4 audio coding

Class: Multimedia Technology, **Section # 8:** Audio Coding

Instructor: George Xylomenos, **Department:** Informatics

MPEG-4 audio tools (1 of 2)

- Multiple audio codecs supported
  - 2-6 Kbps: LPC
  - 6-24 Kbps: CELP
  - 24-64 Kbps: AAC
- Text to speech (TTS)
  - 200 bps to 1,2 Kbps bit rates
  - Simple text or text with timing information

MPEG-4 audio tools (2 of 2)

- Score-based audio synthesis
  - 2-3 Kbps bitrate
  - An orchestra consisting of instruments
  - Instructions to each instrument
  - Sound bank plus sound effects
- MPEG-2 AAC expansions
  - HD AAC: lossless compression
  - HE AAC: even lower bit rate

End of Section # 8

Class: Multimedia Technology, **Section # 8:** Audio Coding

Instructor: George Xylomenos, **Department:** Informatics