Lecture notes for SSY150: Multimedia and video communications

Compression of Speech and Audio Signals

(Lecture 2)

Irene Y.H. Gu

Chalmers Univ. of Technology, Sweden

March 26, 2020

1

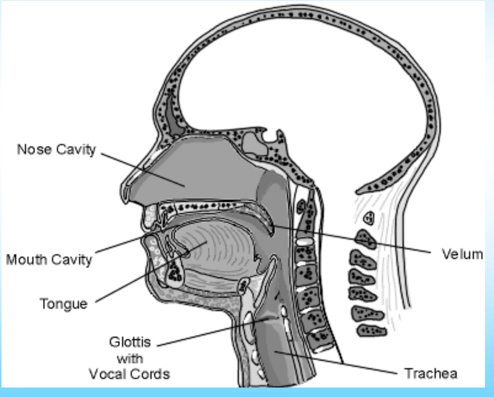
Content

2

1. Speech/Audio production

Human Speech Organ:

Speech: results from the combination of the lung, glottis (vocal cords), and mouth-nose cavity



(*K.Fellbaum,* Brandenburg Tech. Univ. Cottbus, Germany)

Fig. Mouth and nose cavity acting as an articulating tract

(From: http://www.kt.tu-cottbus.de/speech-analysis)

3

2. Basic methods for speech compression

Parametric (model-based)

- LPC analysis /synthesis
- Code-excited LPC (CELP) analysis / synthesis

Non-parametric (non-model based)

- Subband coding
- Transform coding
- Vector quantization

Compression is achieved by:

- using model parameters

- remove small coefficients in freq. bands
- different bit allocation in freq. bands

4

Parametric methods
(model-based methods)

5

a) LPC model for speech analysis/synthesis

LPC – Linear Predictive Coding: is related to a AR model/all pole model

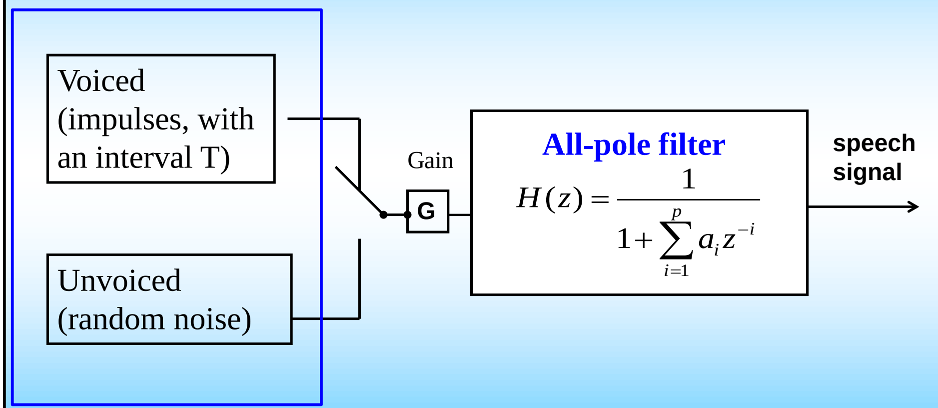
$$s(n) = \frac{G}{1 + \sum_{i=1}^{p} a_i z^{-i}} w(n)$$

For each short time (e.g. 10-20ms) of speech (approx. stationary!), only a few parameters are required for synthesizing the speech:

- p LPC-coefficients (how many p's ? What is the principle to choose #p ?),
- Gain G,
- voiced/unvoiced indicator
- pitch period T (for voiced speech, to generate impulse sequence with T interval)

6

LPC Speech Synthesis

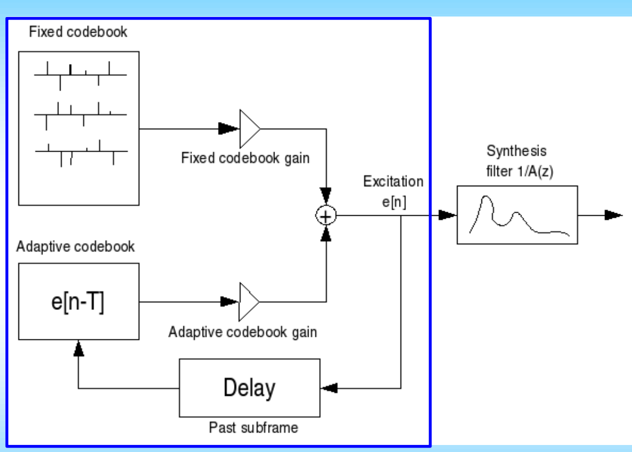


7

b) Code-Excited LP (CELP) for speech analysis/synthesis

Modeling the noise

Modeling the pitch



(figure from 'wikipedia, the free encyclopedia')

Vocal tract: LPC (AR model)
Excitations: an adaptive (pitch) codebook + a fixed stochastic VQ codebook

8

c) Damped sinusoids in white noise for music compression

$$s(n) = \sum_{i=1}^{K} a_i e^{-\beta_i n} \cos(\omega_i n + \phi_i) + v(n)$$

For each 10-20ms of audio signal, only a few parameters are required for re-synthesizing the signal:

- damping factors β_i
- amplitudes a_i
- frequency ω_i
- initial phase ϕ_i

$i = 1, \cdots K$

These parameters can be estimated by the ESPRIT / MUSIC algorithms.

9

Non-parametric methods
(non-model based methods)

10

General principles of non-parametric methods

Decompose audio/speech signal by:

- subband filters (filterbank), with
  - equal bandwidths
  - octave bandwidths
  - transformation

Achieve compression through:

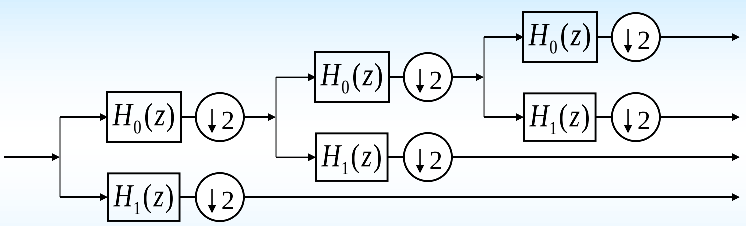
- set the bandwidth consistent to human auditory system
- bit allocation in different bands
  (set different number of quantization levels for different bands)
- variable length coding
  (set the length according the probability of quantizer outputs)

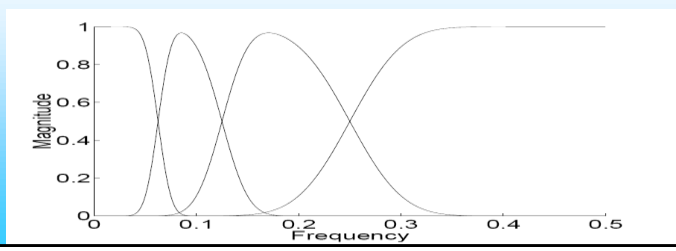
MP3 and MP4 belong to this category!

1) Subband Filters

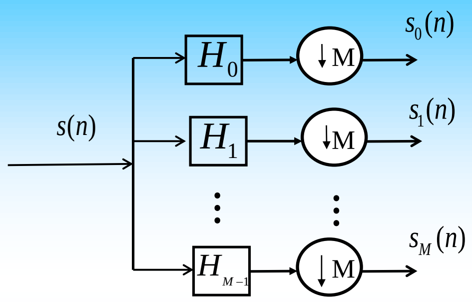
- Decompose signal into frequency bands, followed by down-sampling

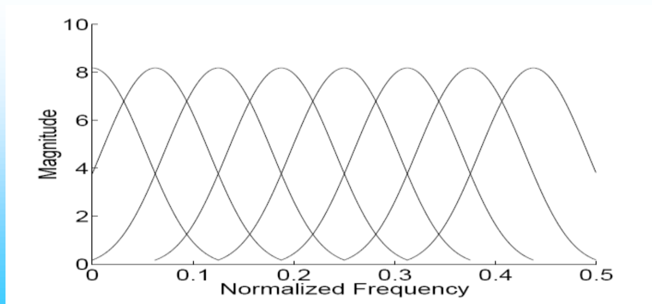
Example: Subband filters with octave bandwidths





Example: subband filters with equal bandwidths





2) DCT and Modified DCT

Conventional DCT:

Forward 1D DCT

$$f(k) = \frac{w_k}{\sqrt{N}} \sum_{n=0}^{N-1} s(n) \cos\frac{(2n+1)\pi k}{2N}, \quad k = 0, \cdots N-1, \; w_k = \begin{cases} 1 & k = 0 \\ \sqrt{2} & k > 0 \end{cases}$$

Set A as transform matrix: $\mathbf{A} = \left[c_k(n)\right]$ $\quad c_k(\mathrm{n}) = \begin{cases} \frac{1}{\sqrt{N}} & k = 0 \\ \sqrt{\frac{2}{N}} \cos\frac{(2n+1)\pi k}{2N} & 0 < k \le N-1 \end{cases}$

→ **DCT (in the vector and matrix form)** $\quad$ **f=As**

Inverse 1D DCT: $\mathbf{s} = \mathbf{A}^{-1}\mathbf{f} = \mathbf{A}^T\mathbf{f}$

↑

(Since DCT is real and orthonomal $\Rightarrow \mathbf{A}^T = \mathbf{A}^{-1}, \mathbf{A}^{\mathbf{T}} = \mathbf{A}^*$

Modified DCT (MDCT)

For each block of data (length of 2N),

$$X_k = \sum_{n=0}^{2N-1} x_n \cos\left[\frac{\pi}{N}\left(n + \frac{1}{2} + \frac{N}{2}\right)\left(k + \frac{1}{2}\right)\right]$$

Data: signal itself, or outputs of a subband filter.

MDCT: a special case of subband filters
(filter kernel length = data block size)

For compression:
remove small value DCT coefficients (set to 0 values)

15

3. Quantization and encoding

16

Quantization

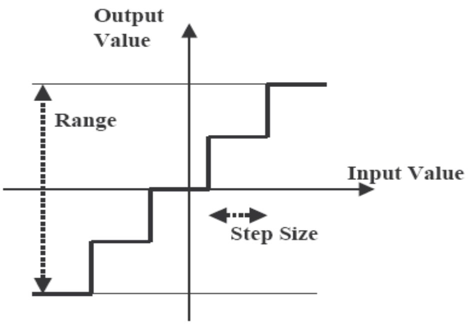
aim: continuous value magnitude → discrete values

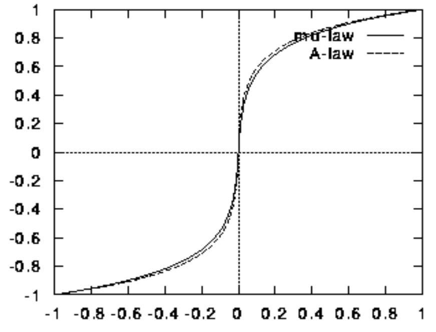
type: scalar /vector quantizer

step size: uniform, logarithm, power …

large step size: high compression,

but low quality (high quantization error)





17

Source symbols encoding (lossless)

There are many,

e.g. Huffman coding;

arithmetic coding;

Ziv-Lempel (LZW) coding

…

- Huffman coding: is an entropy-based lossless coding method. Takes advantage of non-uniform distributions of symbols, where different code lengths are given according to the probabilities of symbols.

18

4. Psychoacoustic model, and parameters for HAS (Human Auditory System)

19

Why Psychoacoustic model ?

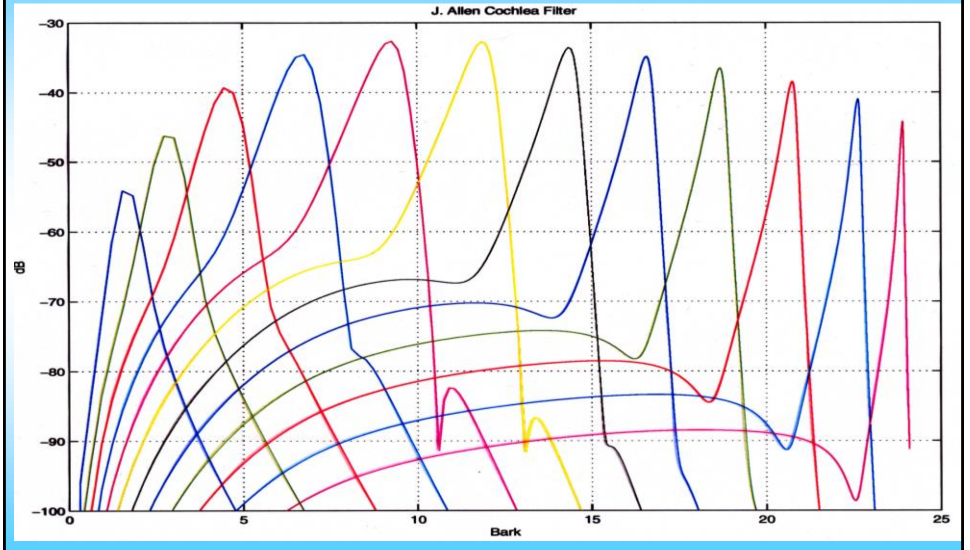
- ”Cheat” human ears: lossy compression, but perceptually lossless
- No bits attribute to sound that is non-audible
- No extra bits to sound components than human ears needed.

Main ”features” in HAS:

- Insensitive to phase changes
- Frequency resolution: differ in different frequencies
  HAS: ~ cochlear filters
  imply: different bit allocation to different frequency bands
- Masking effect: within a “critical band”, a stronger tone masks the remaining weaker sounds (making them inaudible)
- Critical band
- Bark scale, Bark frequency

20

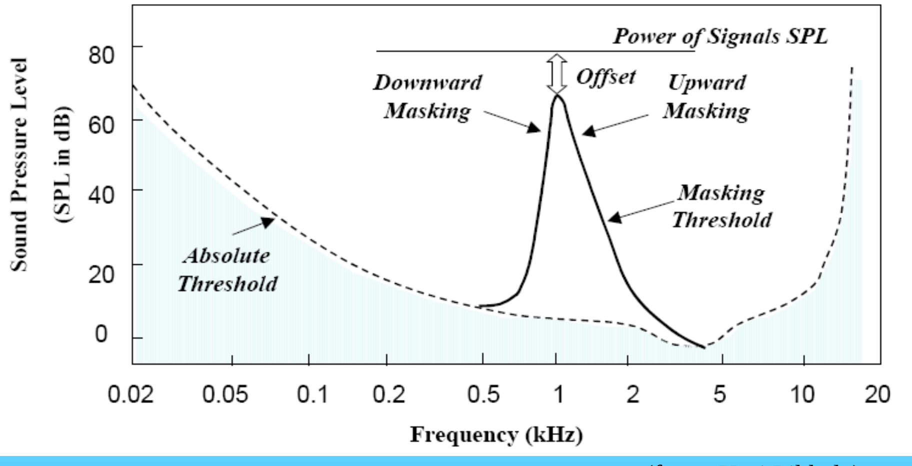
Cochlear filters: HAS modeling



(From J. Allen: Cochlear Filters)

21

Psychoacoustic modeling: Masking effect of HAS



(from: Kari Pihkala)

22

Critical band, Bark scale vs. frequency

- **Critical band**:

  A range of frequencies within which the masking SNR remains a constant.

- **Bark scale:**

  A standardized scale of frequency, where each “Bark” constitutes one critical bandwidth.

  Is approximately equal-bandwidth up to 700Hz, and 1/3 octave above 700Hz.

  A frequency scale, under which the masking phenomenon and shape of cochlear filters are approximately invariant.

23

- **Bark frequency:** can be converted from the usual frequency *f* (in Hz)

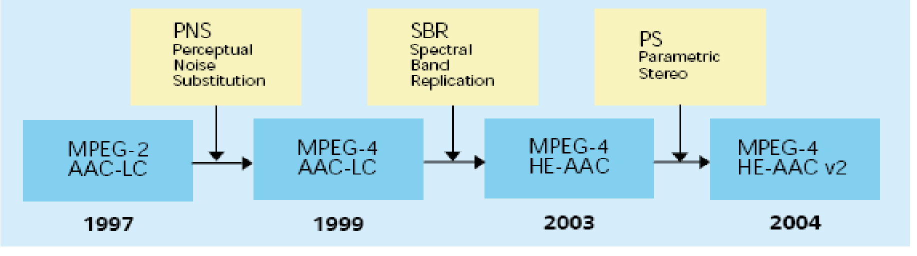
$$B_f = 13\tan^{-1}\left(\frac{0.76f}{1000}\right) + 3.5\tan^{-1}\left(\left(\frac{f}{7500}\right)^2\right)$$

24

5. Speech/Audio Coding Standards: examples

Audio Coding Standards: brief Introduction

Progress in AAC (Advanced Audio Coding) standards



(From: **http://www.iis.fraunhofer.de/bf/amm/)**

LC: low complicity; HE: high efficiency

HE-AAC v2 (aacPlus v2):
is also part of the 3GPP standard for the delivery of audio content to 3G devices

Varieties in AAC codecs

Advanced Audio Coding's multiple codecs:

- Low Complexity AAC (LC-AAC)
- High-Efficiency AAC (HE-AAC)
- Scalable Sample Rate AAC (AAC-SSR)
- Bit Sliced Arithmetic Coding (BSAC)
- Long Term Predictor (LTP)
- Low Delay AAC (LD-AAC)

27

MPEG/ISO audio standards

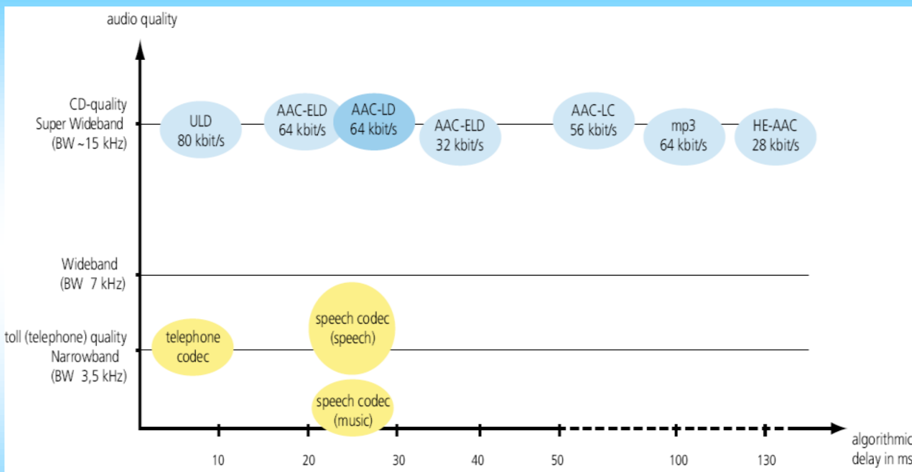
Standards	Audio sampling rate (kHz)	Compressed bit-rate (kbits/sec)	Channels	Standard Approved
MPEG-1 Layer I	32, 44.1, 48	32 – 448	1-2 channels	1992
MPEG-1 Layer II	32, 44.1, 48	32 – 384	1-2 channels	1992
MPEG-1 Layer III	32, 44.1, 48	32 – 320	1-2 channels	1993
MPEG-2 Layer I	32, 44.1, 48	32 – 448 for two BC channels	1-5.1 channels	1994
	16, 22.05, 24	32 – 256 for two BC channels		
MPEG-2 Layer II	32, 44.1, 48	32 – 384 for two BC channels	1-5.1 channels	1994
	16, 22.05, 24	8 – 160 for two BC channels		
MPEG-2 Layer III	32, 44.1, 48	32 – 384 for two BC channels	1-5.1 channels	1994
	16, 22.05, 24	8 – 160 for two BC channels		
MPEG-2 AAC	8, 11.025, 12, 16, 22.05, 24, 32, 44.1, 48, 64, 88.2, 96	Indicated by a 23-bit unsigned integer	1-48 channels	1997
MPEG-4 T/F coding	8, 11.025, 12, 16, 22.05, 24, 32, 44.1, 48, 64, 88.2, 96	Indicated by a 23-bit unsigned integer	1-48 channels	1999

BC: backward compatibility

Table is from C-M Liu and W-W Chang, '99 in http://www.mp3-tech.org/programmer/docs/AudioCoding.pdf

28

Comparison: quality, bit rate, time delay

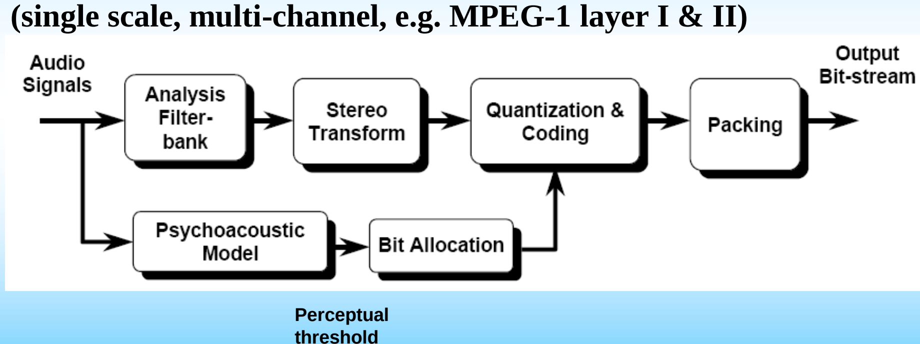


(From: http://www.iis.fraunhofer.de/bf/amm/)

Speech codecs

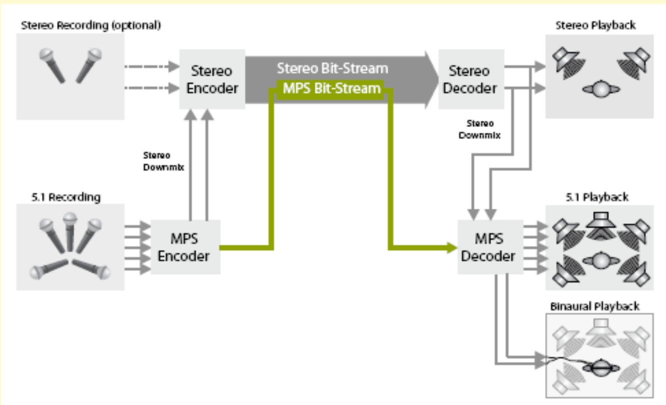
A typical perceptual audio encoder

(single scale, multi-channel, e.g. MPEG-1 layer I & II)



e.g. Use MDCT (modified DCT) as analysis filterbank
(containing 32-band polyphase quadrature filters (PQFs)).

Stereo / 5.1 channel audio:



How MPEG Surround (MPS) works

31

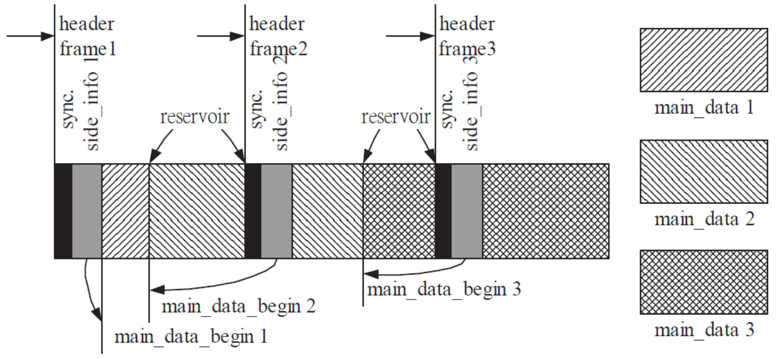
Bit Stream

Header:

bit rate, sample frequency, …

Side information:

block type, Huffman tables, subband gains, …



32

Speech/audio coding standards:

Non-parametric audio coding (subband filters):
in MPEG-1, MPEG-2 standards

Parametric audio coding: (CELP-based)
in MPEG-4 standards

33

MPEG-2 (part 7) and MPEG-4 (part 3):

use Advanced Audio Coding (AAC) schemes
AAC is a standardized, lossy compression and encoding scheme for digital audio.
AAC has a better quality than *MP3* at the same bite-rate, particularly under 192 kb/s.

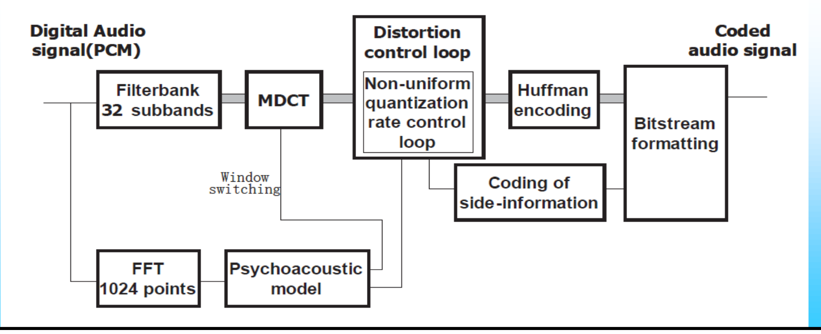
MPEG layer 3 (or, MP3):
is the most popular audio coding standard for digital music in the computer and the Internet. MP3 is a part of the MPEG-1 and the MPEG-2 standards.

34

MP3 coding

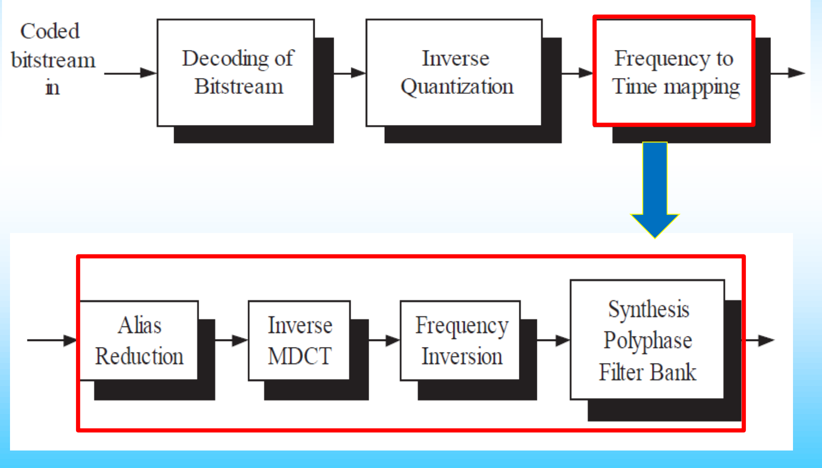
- Analysis filterbank /Synthesis filterbank
- Midified DCT (MDCT) /Inverse MDCT
- Quantizer/dequantizer
- Huffman encoder/decoder
- Psycho-acoustic model (using masking effect, critical band, …)
- Bitstream

Block diagram of MP3 encoder:



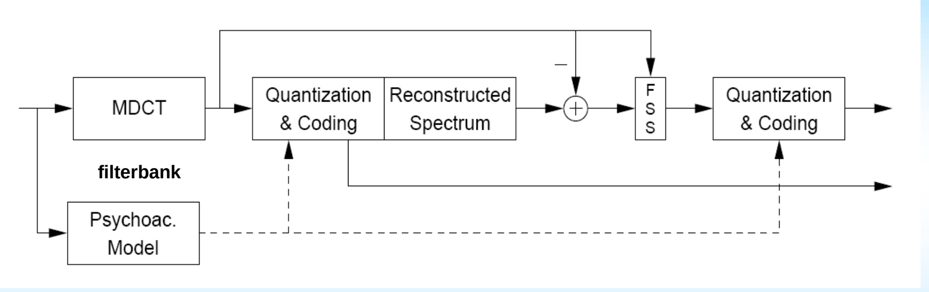
35

Block diagram of MP3 decoder:



36

AAC coding in MPEG-2 / MPEG-4



FSS: frequency selective switch

37

MDCT and hybrid filterbank

- Modified Discrete Cosine Transform (MDCT)

- Hybrid filterbank (or, subband MDCT)

  4 PQF (polyphase quadrature filter) subbands
  followed by a MDCT

  - fs: [8KHz, 96kHz]:
  - narrow/wideband: 10-40ms frame/10-20ms frame
    ➔ high/low frequency resolution
  - Channels: MPEG-4 up to 48 channels;
    MPEG-1: up to 2 channels;
    MPEG-2: up to 5.1 channels

38

6. Objective quality measures for synthetic speech/audio signals

39

Quality measures for synthetic audio/speech

+ Human ears are insensitive to phase changes !
  => Criteria based on speech waveform distortion is NOT suitable

+ Compute spectral distortions **in the frequency-domain**
  (e.g. magnitude spectral distortions between original and synthetic ones)

+ Or, compute distortions **in the Bark frequency-domain**:
  Bark / Modified Bark Spectral distortion (BSD/MBSD)

$$MBSD = \frac{1}{N}\sum_{j=1}^{N}\left[\sum_{i=1}^{K} M(i)\left|L_x^{(j)}(i) - L_y^{(j)}(i)\right|^{n}\right]$$

M(i): Perceptible distortion in i-th critical band

$L_x^j(i)$ **: Bark spectrum of j-th frame of coded speech**

+ Perceptual speech quality measures
  (e.g. ITU-T recommendation P.861)

40

7. About laboratory project-1

41

Lab.-1

Tasks: speech model: analysis, synthesis and compression (dead line: 2020-04-17, 23:55)

1. Record a (stationary) single vowel, and make Matlab programs for LPC analysis and synthesis of stationary (single-tone) speech;
2. Record a (nonstationary) speech sentence, and make Matlab programs for block-based LPC analysis and synthesis of nonstationary speech (using the residual sequence as the excitations);
3. Repeat the task 2, however, excitations to the filter are replaced by using a few prominent residuals (<20) in each block;
4. For the recorded speech sentence, determine whether a speech frame (block) is voiced or unvoiced. For those voiced frames, estimate the pitch periods either from the cepstrum.
5. Objective measures of synthetic speech quality

42

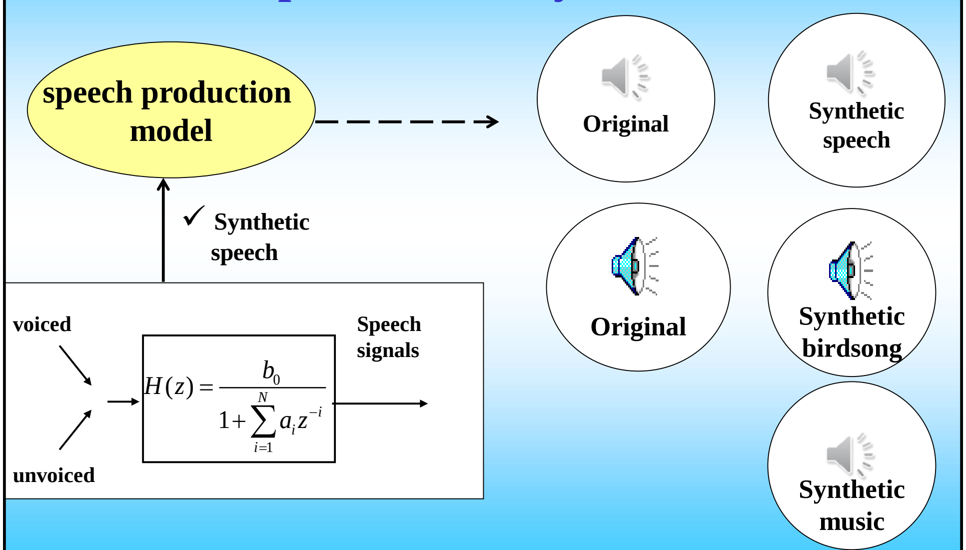
What one shall do in Lab.1:

Speech modeling, analysis, synthesis and compression

- Record and save a sound or speech file to a computer, and then load the speech file in Matlab.
- Make a Matlab program on LPC (Linear predictive coding) analysis and then synthesis of single tone sound (stationary) and a speech sentence (nonstationary). Listen to the resulting sound.
  From this, you can learn how speech compression is achieved:
  a 10ms of speech signal only requires less than 20 parameters to characterize.
- Model the vocal cord excitations of speech by some impulses or white noise, to the LPC model, and listen to the synthetic speech.
- Estimate the pitch period using the cepstrum method.

43

Example of what you may achieve: speech/audio synthesis



44

8. References

[1] Lawrence R. Rabiner, Ronald W. Schafer, Digital Processing of speech signals, Prentice-Hall, Inc., 1978.

[2] John R., Jr. Deller, John H.L. Hansen, John G. Proakis, Discrete-time processing of speech signals, IEEE Press Classic Reissue, 1999.

[3] Alan V. Oppenheim, Ronald W. Schafer, and John R. Buck, Discrete-Time Signal Processing, 2nd edition, Prentice Hall, Inc. 1999.

[4] Wikipedia, the free encyclopedia on CELP: http://en.wikipedia.org/wiki/Code_Excited_Linear_Prediction

45