Lecture notes for SSY150: Multimedia and video communications

Compression of Speech and Audio Signals

(for lectures 2 and 3)

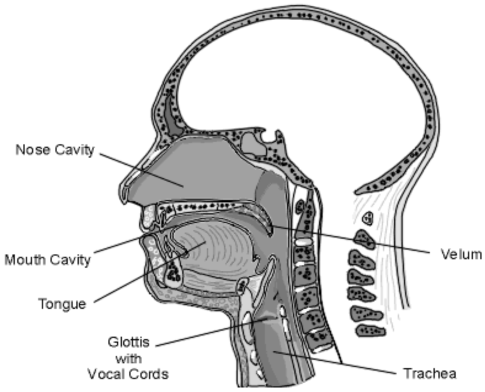
Irene Y.H. Gu, Dept. of Signals and Systems,
Chalmers Univ. of Technology, Sweden
March 19 and 24, 2009

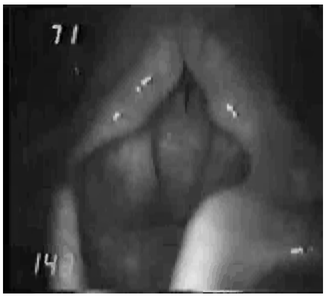
Contents

1. Speech/Audio production

Human Speech Organ:

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





Movement in vocal cord
(*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)

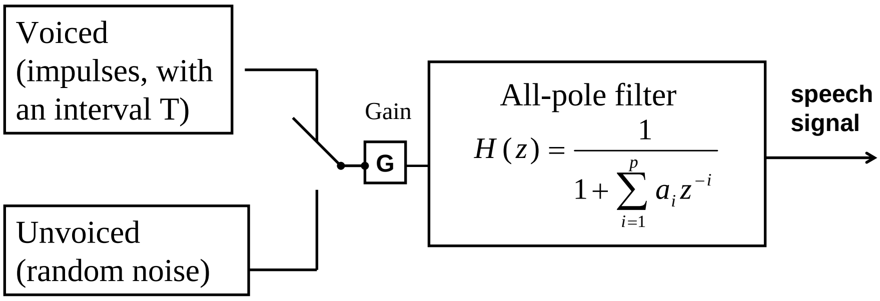
2. Basic methods for speech modeling and encoding

Typical Source Coding Methods

- LPC analysis
- Multi-pulse analysis by synthesis
- Sub-band coding
- Transform coding
- Vector quantization

….

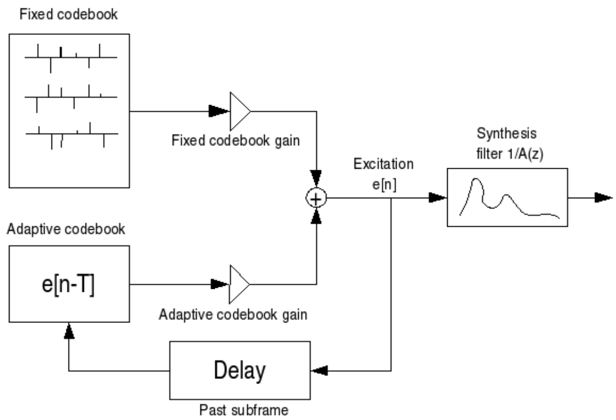
The LPC model for speech



Code-Excited LP (CELP) for speech

Modeling the noise

Modeling the pitch



(figure from ‘wikipedia, the free encyclopedia’)

Source model: a LP filter with an all pole-model
Excitations: an adaptive (pitch) codebook + a fixed stochastic VQ codebook

3. Fundamentals of Audio Compression

(a) Compression is achieved through using models

e.g.1. Speech is generated by a LPC model

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

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

- p LPC-coefficients,
- Gain G,
- period excitations with a pitch interval T, or, white noise sequence

Code-excited LP (CELP): replaces the excitation in the model by a codebook

e.g. 2. music signals described by a damped sinusoidal model in noise

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

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 algorithm.

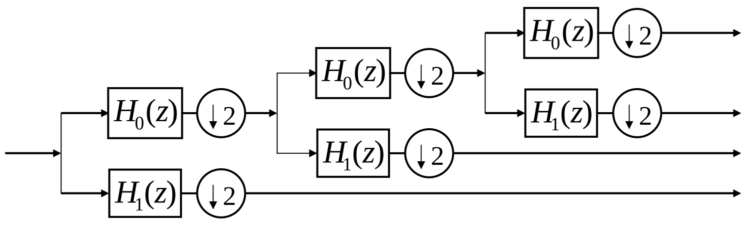
3.Fundamentals of audio compression (cont'd)

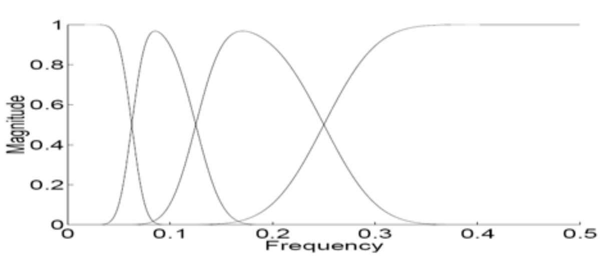
b) compression from non-parametric methods

- Decompose audio signal using subband filters/transformation
  (with different bandwidths)
- Bit allocation
  (set different number of quantization levels for different bands)
- Variable length coding
  (Set the length according the probability of quantizer outputs)

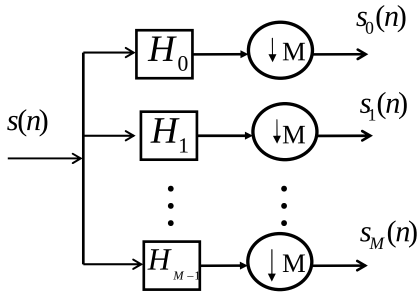
Compression through subband filters, vector/scalar quantization and variable-length source coding

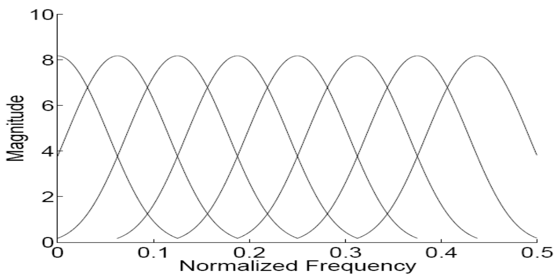
e.g.: Subband filters, with octave bandwidths





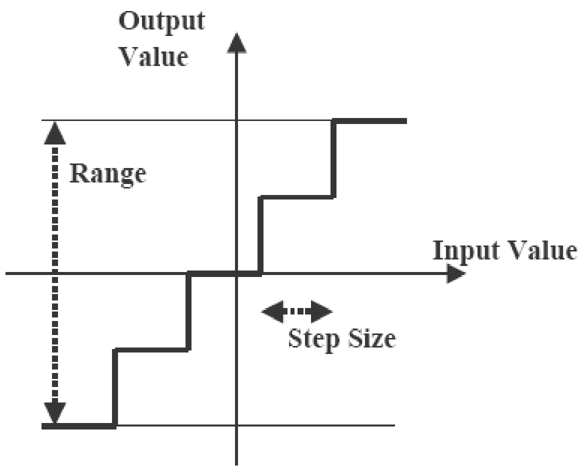
e.g.: Subband filters with an equal bandwidth





From STFT

e.g.: Scalar quantization



Lossless encoding of source symbols

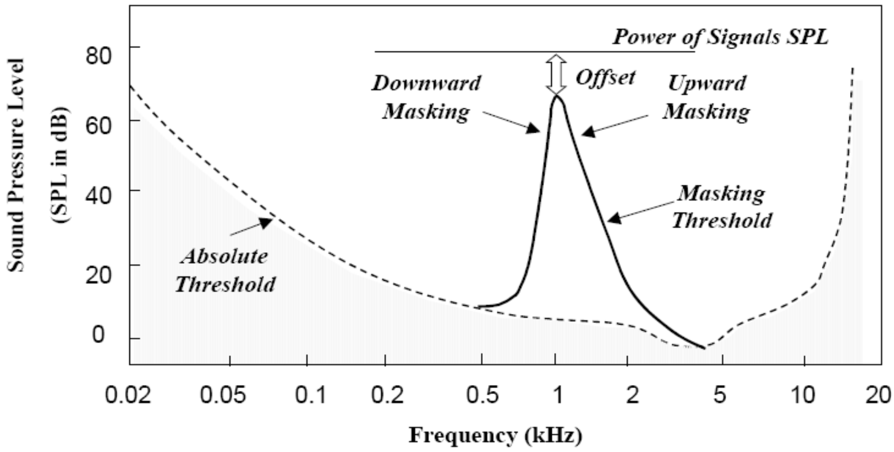
- e.g. Huffman coding; arithmetic coding; Ziv-Lempel (LZW) coding …

- 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.

4. The masking effect of the human auditory system (HAS) and perceptually relevant speech/audio compression

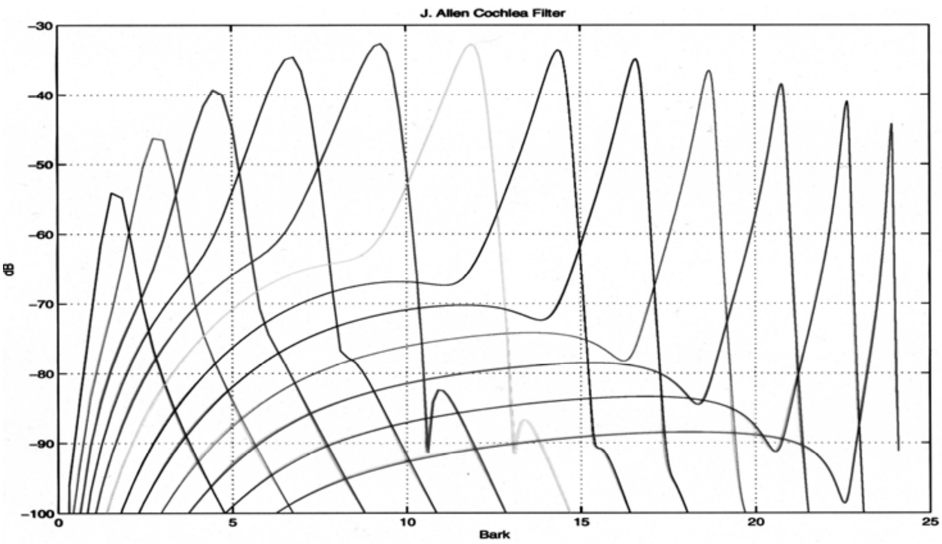
Within a "critical band", a stronger tone masks the remaining weaker sounds (making them inaudible)

Psychoacoustic modeling: The masking effect of the human auditory system



(from: Kari Pihkala)

Cochlear filters for HAS modeling



(From J. Allen: Cochlear Filters)

Critical Band, Bark scale and frequency

- **Critical band**:

  A range of frequencies where 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.

Critical band, Bark scale and frequency (cont'd)

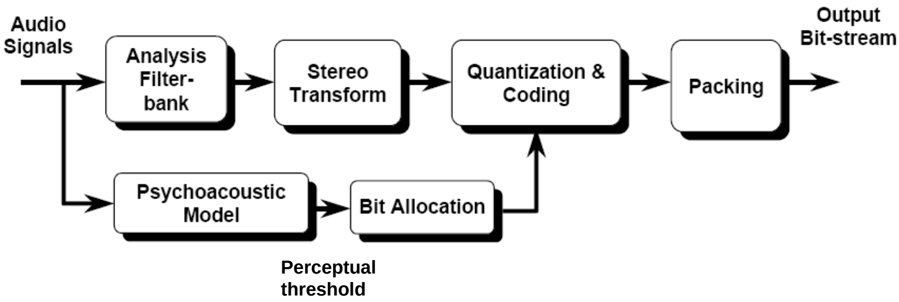
- A frequency scale over which the masking phenomenon and the shape of cochlear filters are approximately invariant.

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

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

A typical perceptual audio encoder

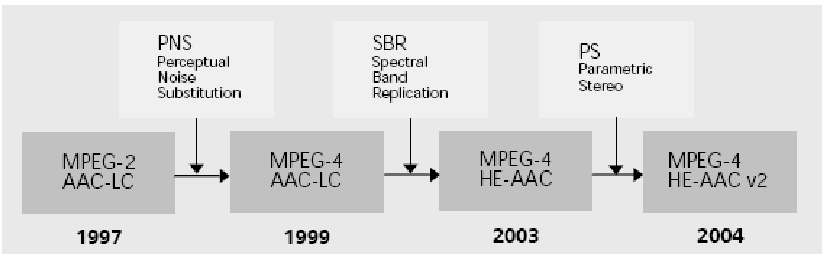
(single scale, multi-channel, e.g. MPEG-1 layer I & II)



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

5. Audio Coding Standards: Brief

Progress in AAC (Advanced Audio Coding) standards



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

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)

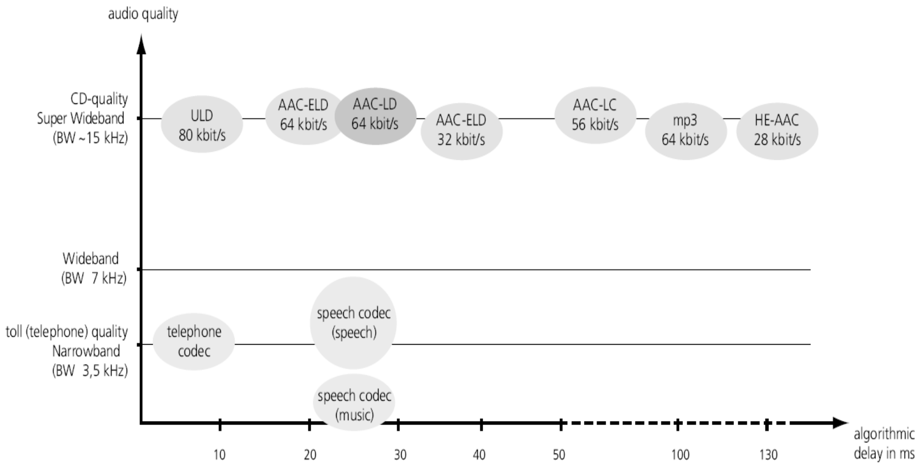
MPEG-2 (part 7) and MPEG-4 (part 3):
use Advanced Audio Coding (AAC) schemes

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.

Codecs: audio quality vs. time delay



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

Comparisons: Audio quality and time delay (mono bit rate)

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

Parametric/non-parametric methods in audio coding standards

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

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

MDCT and Hybrid filterbank

- Modified Discrete Cosine Transform (MDCT)

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

- Hybrid filterbank (or, subband MDCT)

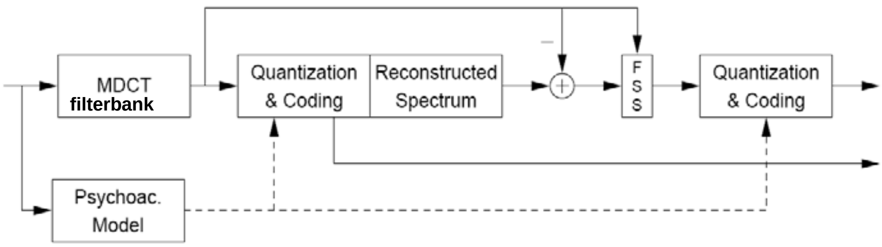
4 PQF (polyphase QF) subbands followed by a MDCT

- 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

Main features in MPEG-4 audio coding

- Two basic algorithms
  - HVXC (Harmonic Vector eXcitation Coding)
  - CELP (Code Excited Linear Prediction)
- Multi bit-rates
  - 1.5 ~ 24 kbps
- Narrow-band and wide-band - CELP
- Lowest bit-rate as an international standard coding - HVXC
  - 2.0 kbps (fixed) ave 1.5 kbps (var)
- New Functionalities
  - Speed / Pitch change - HVXC
  - Bit-rate scalability - HVXC, CELP
  - Bandwidth scalability - CELP

MPEG-4 (bitrate) scalable AAC coding



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

FSS: frequency selective switch

5. About Lab. exercise-1

Tasks:

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, the excitations to the filter are replaced by using 15 prominent residuals 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, or the ACF method.

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

Some useful websites:

Software:http://www.utdallas.edu/~loizou/speech/colea.htm (software)

http://www.iis.fraunhofer.de/bf/amm/

http://neural.cs.nthu.edu.tw/jang