Extraction and Representation of Features, Spring 2011

Lecture 4: Speech and Audio: Basics and Resources

Zheng-Hua Tan

Multimedia Information and Signal Processing
Department of Electronic Systems
Aalborg University, Denmark
zt@es.aau.dk

Outline

- Introduction
- Speech basics
- Sound basics
- Resources

Computer as dream of human being

- HAL talks, listens, reads lips and solves problems
  - Nature and effortless for huamn
  - Hard for computer
  - Dream of AI scientists and human
  - True in *2001: A Space Odyssey*

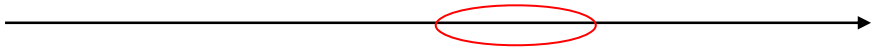
(*After 2001: A Space Odyssey, 1968*)

Computer as a reality: state-of-the-art

- Demos
  - Microsoft
  - Nuance
  - Text to speech (TTS)
    - Festival TTS @ CSTR Edinburg University
    - Next generation TTS @ AT&T

Information in Speech

- Speech coding data rates



200k 100k 64k 32k 16k 12k 9k 4.8k 2k 1k 500 100 60

ADPCM, DPCM, PCM LPC, CELP, MELP, Vocoders

Waveform coding Parametric (source) coding

Human can understand text:

10 char/sec x 6 bits/ASCII char = 60 bits/sec

Is content in speech more than 60 bits/sec?

Information in Speech – cont.

- Examples

"That's one **small step for man**; one **giant leap for mankind.**"

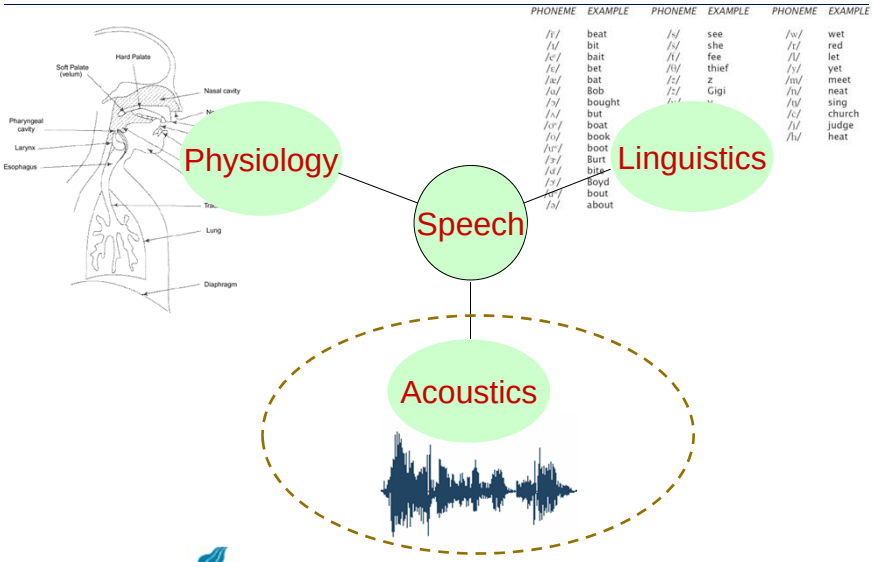
-- Neil Armstrong, *Apollo 11 Moon Landing Speech*

"**I have a dream** that my four little children will one day live in a nation where they will not be judged by the color of their skin but by the content of their character. I have a dream today!"

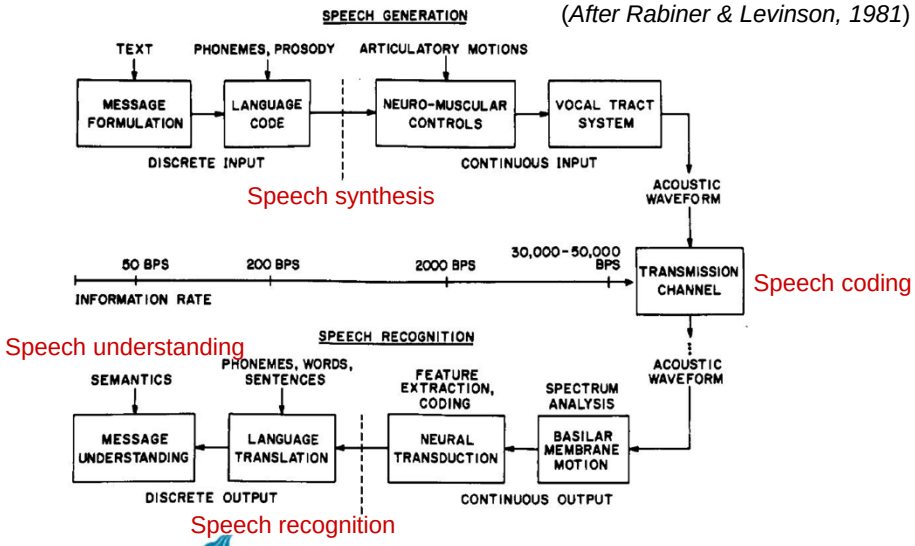
-- Martin Luther King, Jr., I Have a Dream

Speech contains speaker identity, emotion, meaning, text. → speech techniques

Speech is a complex process



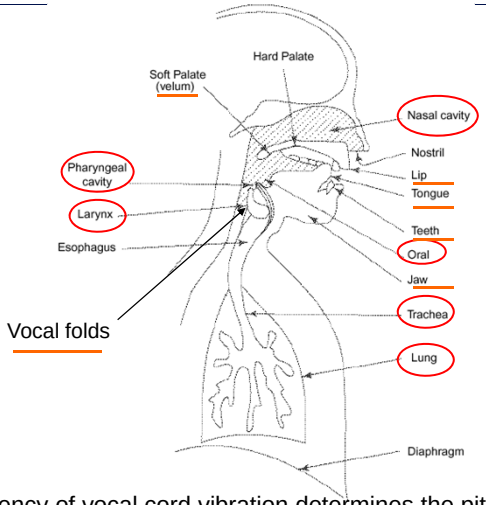
Human speech communication process



Literature

- Textbook:
  - J Deller, J Hansen and J Proakis, Discrete-Time Processing of Speech Signals, IEEE Press, 2000.
- References:
  - Huang, Acero and Hon, Spoken Language Processing, Prentice-Hall, 2001.
  - D. O'Shaughnessy, Speech Communications, IEEE Press, 2000

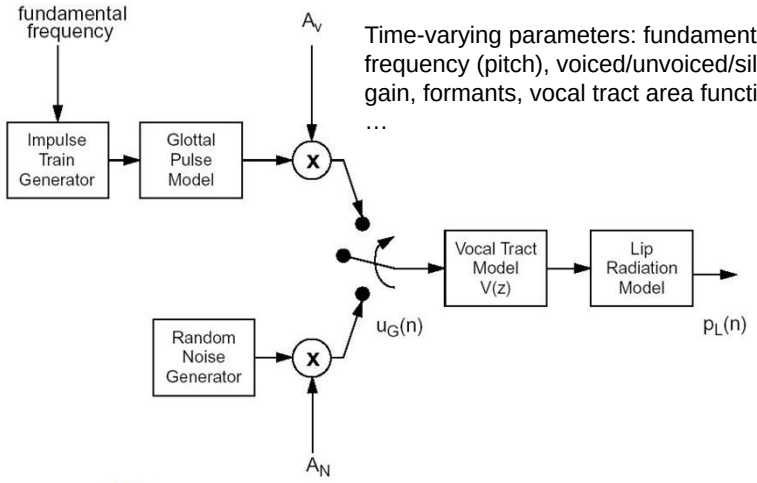
Schematic diagram of speech production



The frequency of vocal cord vibration determines the pitch of the voice (for a male, 50-200Hz; for a female, up to 500Hz).

Model of speech production

- Digital model of speech production



Time-varying parameters: fundamental frequency (pitch), voiced/unvoiced/silence, gain, formants, vocal tract area functions, …

Vocal tract modelling

- Source-filter model

- Type of excitation (source)
  - Voiced: produced by forcing air through the glottis. Vowels are voiced.
  - Unvoiced: generated by forming a constriction at some point along the vocal tract and forcing air through the constriction.

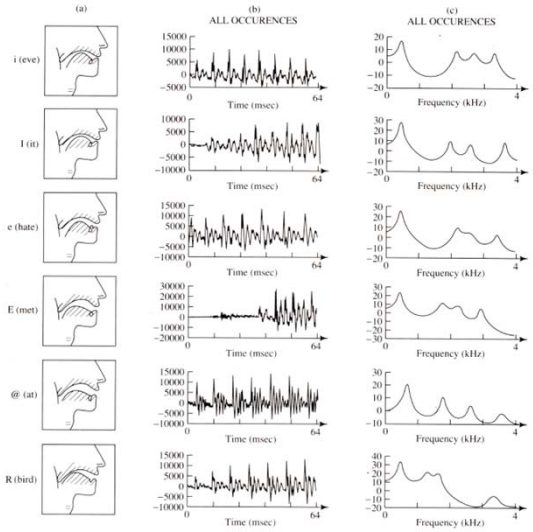
Role of the vocal tract

- Vowels: produced by exciting a fixed vocal tract with quasi-periodic pulsed of air caused by vibration of the vocal cords
- Consonants: a significant restriction and thus weaker in amplitude and noisy-like
- Formants: resonances determined by the shape of vocal tract, which form the overall spectrum and the properties of the filter

The speech signal

- Speech is a sequence of highly changing sounds
- When producing sounds, the vocal cords and the various articulators slowly change over time

Vowel production: examples



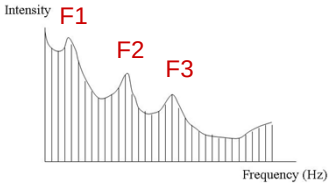
(*After Joseph Picone*)

- Fixed vocal tract shape
- Voiced
- Cross-sectional area

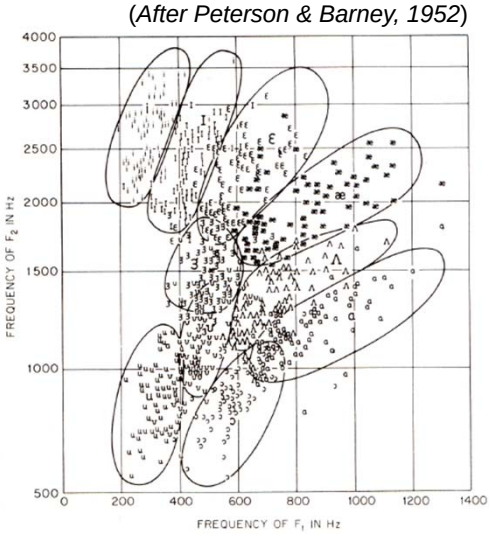
  → F_i

The vowel space

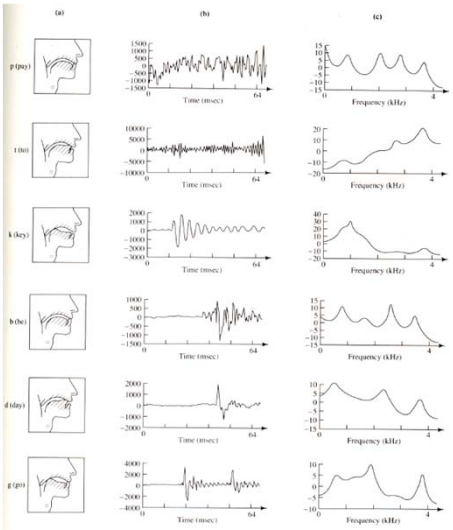
By the locations of the first and second formant frequencies:



(*After Peterson & Barney, 1952*)



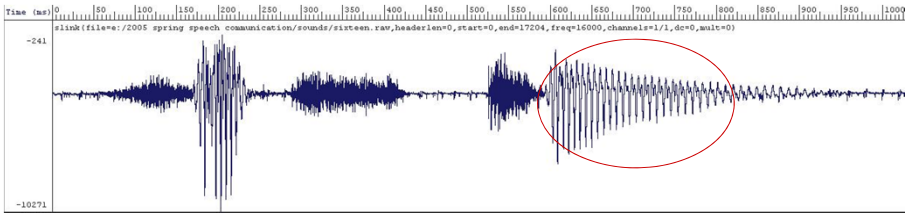
Consonant production: examples



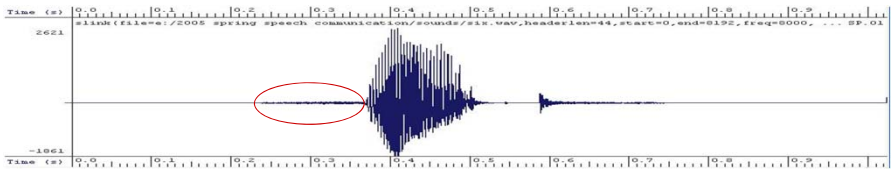
(*After Joseph Picone*)

Speech sounds and waveforms

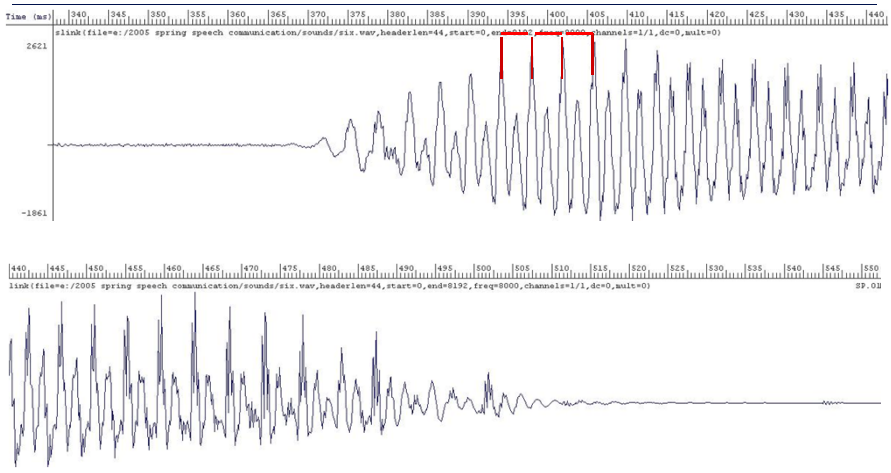
sixteen /s/ /i/ /k/ /s/ /t/ /ee/ /n/



six periodicity, intensity, duration, boundary, etc



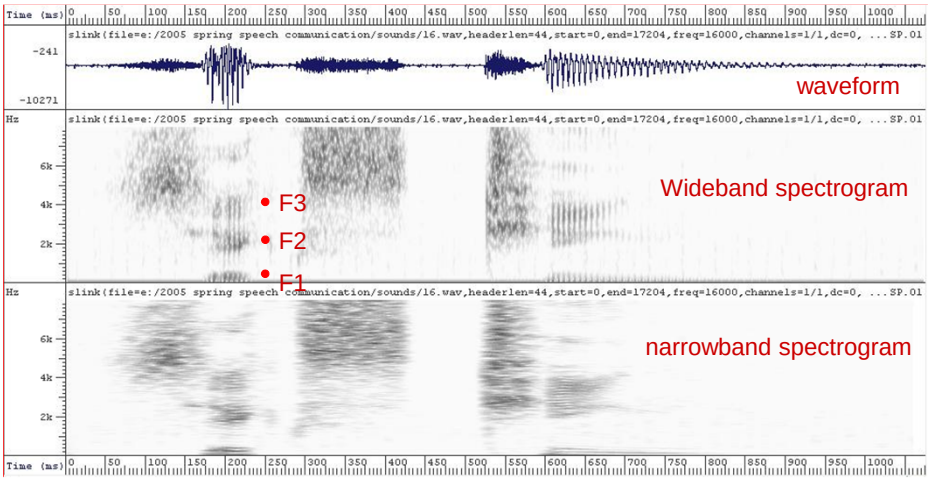
Observing pitch from waveforms



Spectrogram

- Spectrogram
  - two-dimensional waveform (amplitude/time) is converted into a three-dimensional pattern (amplitude/frequency/time)
  - Wideband spectrogram: analyzed on 15ms sections of waveform with a step of 1ms
    - voiced regions with vertical striations due to the periodicity of the time waveform (each vertical line represents a pulse of vocal folds) while unvoiced regions are solid/random, or 'snowy'
  - Narrowband spectrogram: on 50ms
    - pitch for voiced intervals in horizontal lines

Wide- and narrow-band spectrograms



Outline

- Introduction
- Speech basics
- Sound basics
- Resources

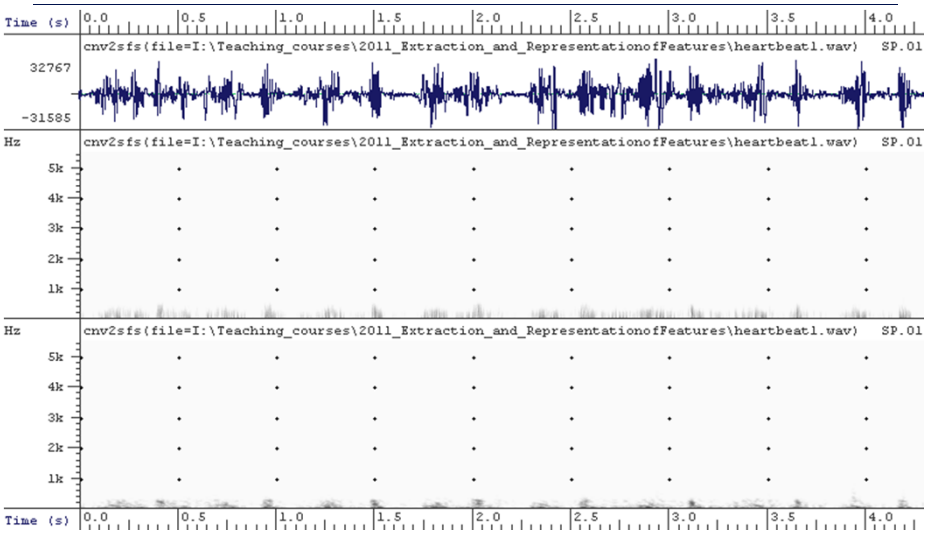
Sound basics

- Audio (sound) wave
  - one-dimensional acoustic pressure wave
  - causes vibration in the eardrum or in a microphone
- Frequency range of human ear
  - 20 – 20.000 Hz (20 KHz)
  - perception nearly logarithmic, relation of amplitudes A and B is expressed as $dB = 20 \log_{10} (A/B)$

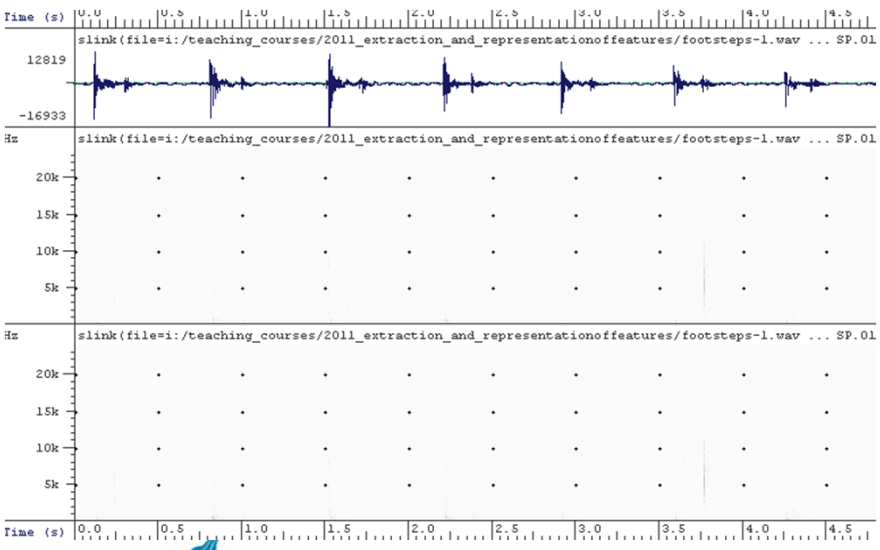
Sounds

(From Sound Jay)

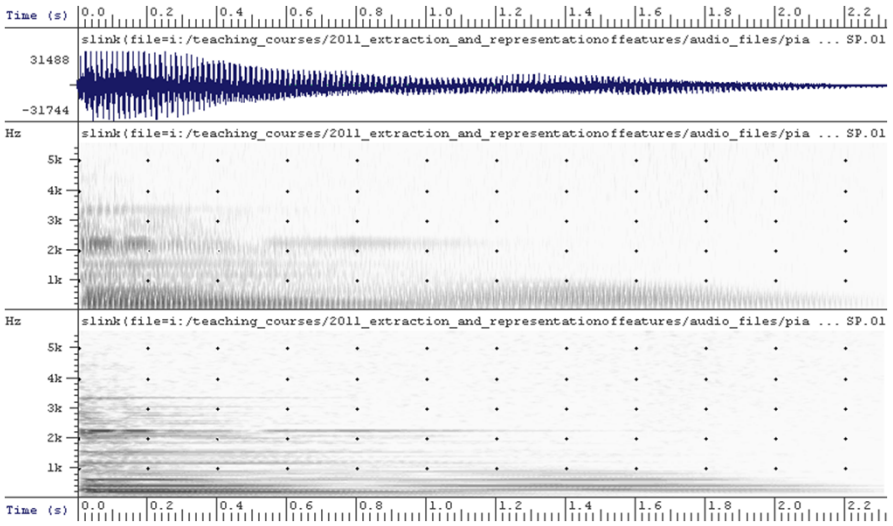
Heartbeat



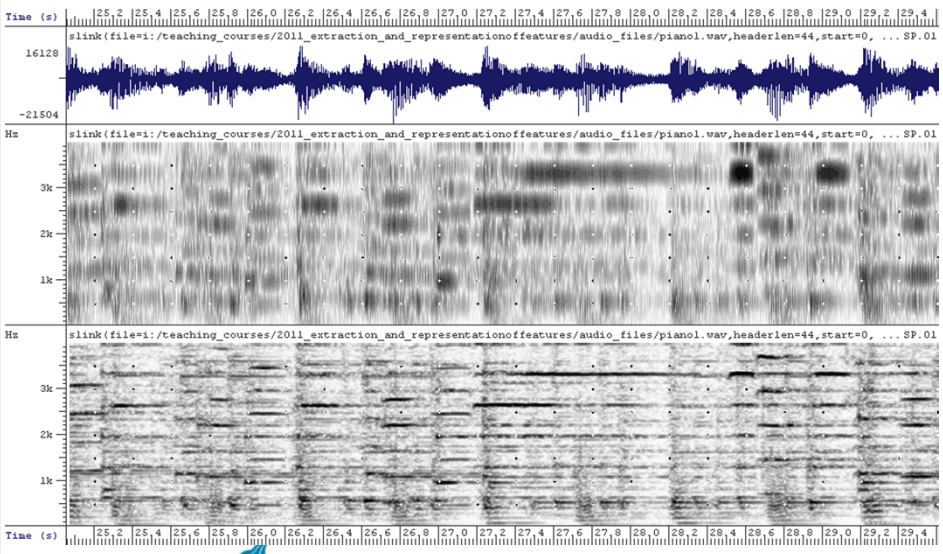
Footstep



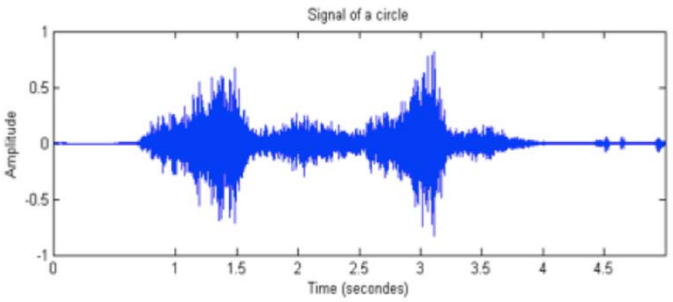
Piano note



Piano



Scratch sound



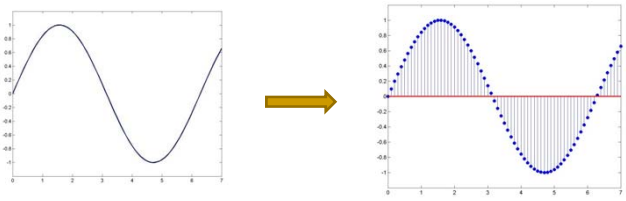
B. Lemoine, J. Nicolaou, A. Osmar, A. Palsky, C. Petitimbert, 2011.

Analog-to-digital conversion

- Converting an analog signal into a digital signal has 2 sub-processes:
  1. sampling - conversion of a continuous-space/time (audio, video) signal into a discrete-space/time (audio, video) signal.
  2. quantization - converting a continuous-valued (audio, video) signal that has a continuous range (set of values that it can take) of intensities and/or colors into a discrete-valued (audio, video) signal that has a discrete range of intensities and/or colors.

Analog-to-digital converter

- Sampling
  - What the ADC circuit does is to take samples from the analog signal from time to time. Each sample will be converted into a number, based on its voltage level.



Representation of digital audio

- Temporal resolution, i.e. sampling frequency
  - Audible frequency range: ~ 20Hz-20kHz
  - Nyquist-Shannon theorem: 2*20kHz = 40kHz
  - Actual frequency due to production: 44.1 kHz
- Bit depth, i.e. amplitude quantization
  - 16-bit linear PCM (Pulse-code modulation)
    - Digital audio stored in computers: Windows WAV, Apple AIF, Sun AU, Blu-ray (incl. 20-, 24-bit as well)
    - Compact Disc – Digital Audio
  - 16 bit/sample per channel ~ 2^16 = 65,536

Representation of digital audio

- Bit-rate stereo
  - 16*2*44100 ~ 1.4 Mbit/s
- A CD can store up to 74 minutes of music

Total amount of data =

44,100 samples/(channel*second) * 2 bytes/sample *

2 channels * 60 seconds/minute * 74 minutes

= 783,216,000 bytes

- There are CD-Rs that can hold 700 megabytes (734,003,200 bytes) of error corrected data, or 80 minutes of stereo 16 bit 44.1 kHz audio (846,739,200 bytes) (807.51 megabytes) without error correction code.

Coding standards

	Sampling Rate (KHz)	Quantization level (bits)	Bit Rate (Kbps)
Telephone	8	8	64
AM Radio	16	16	256
FM Radio	22.05	16	352.8
CD Stereo	44.1	16	1411.2
DAT	48	16	1536
DVD (Stereo)	192	24	9216

What about mobile communication and VoIP?

Coding standards

- The International Telecommunications Union (ITU)

Standard	Method	Bit rete (kb/s)	MOS	Complexity (MIPS)	Release Time
ITU G.711	Mu/A-law PCM	64	4.3	0.01	1972
ITU G.729	CS-ACELP	8	4.0	20	1996

- The European Telecommunications Standards Institutes (ETSI)

Standard	Method	Bit rete (kb/s)	MOS	Complexity (MIPS)	Release Time
GSM FR	RPE-LTP	13			1987
GSM AMR	ACELP	4.75-12.2			1998

Outline

- Introduction
- Speech basics
- Sound basics
- Resources

Sound and speech databases

- http://www.soundjay.com/
- http://soundjax.com
- http://www.ldc.upenn.edu/

Speech Tool – Audacity

- http://audacity.sourceforge.net
- Audacity is a free audio editor and recorder for many operating systems. Key features:
  - Record live audio.
  - Convert tapes and records into digital recordings or CDs.
  - Edit Ogg Vorbis, MP3, WAV or AIFF sound files.
  - Cut, copy, splice or mix sounds together.
  - Change the speed or pitch of a recording.
  - Analyze audio signal

Speech Tool – speech filing system

- Speech Filing System- Tools for Speech Research
  - It performs standard operations such as recording, replay, waveform editing and labelling, spectrographic and formant analysis and fundamental frequency estimation.
  - http://www.phon.ucl.ac.uk/resource/sfs/
  - Demo

Speech tool – voicebox

- VOICEBOX: Speech Processing Toolbox for MATLAB
- http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
- Demo

Summary

- Introduction
- Speech basics
- Sound basics
- Resources