Aalto University
School of Electrical
Engineering

Communication acoustics
Ch 16: Speech technologies

Ville Pulkki and Matti Karjalainen

Department of Signal Processing and Acoustics
Aalto University, Finland

October 4, 2022

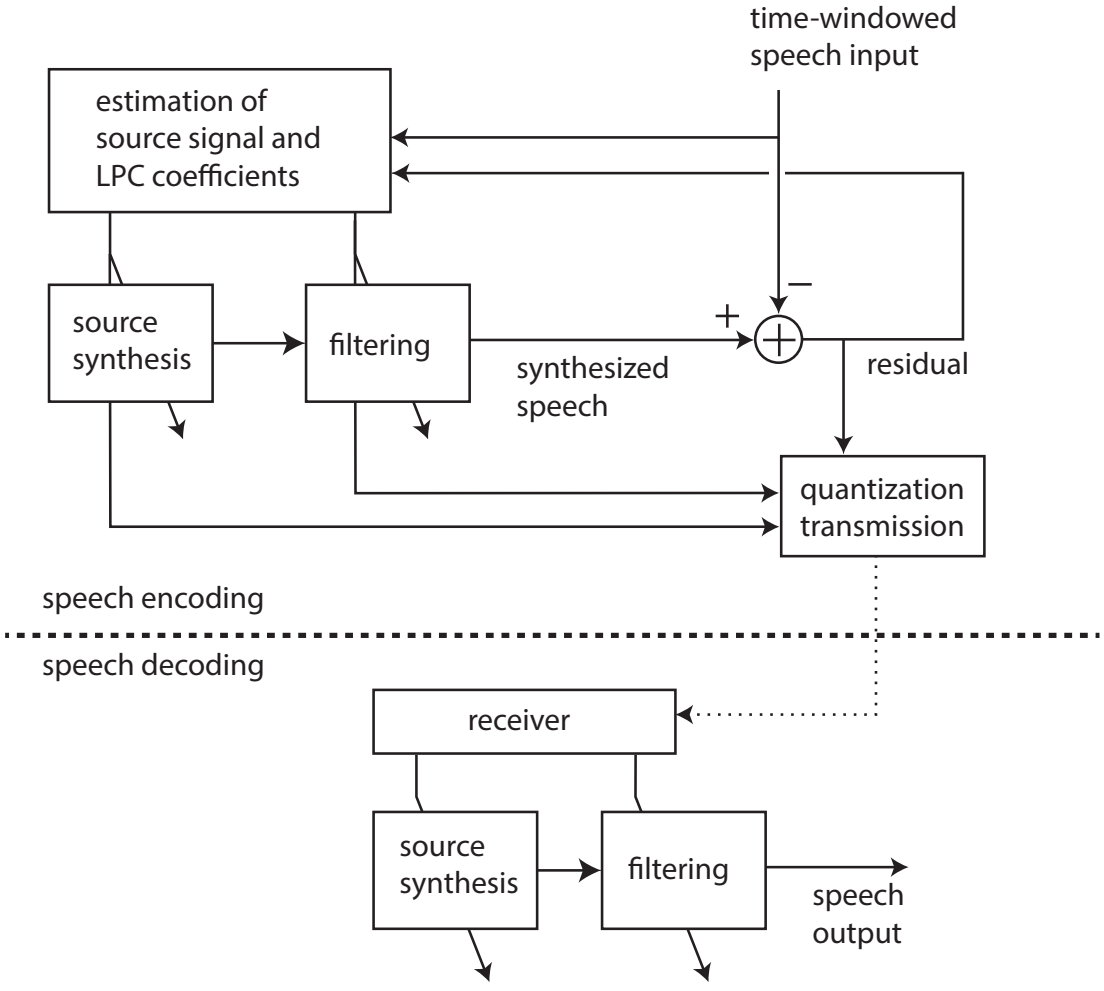
Speech technologies

- Speech coding
- Speech synthesis
- Speech recognition
- [Speech enhancement]

Speech coding

- To transmit or store and replay speech signals using minimal information capacity (number of bits) and with the best possible sound quality
- Early telephone techniques,
  - Electric signal over network
  - Technical challenges in transmission
  - Telephone band from 300 Hz to 3400 Hz, best compromise in quality
- Digital delivery G.711 (70's) with 64 kbit/s rate, logarithmic quantization
- Source-filter models for GSM mobile phones in 80's, e.g., 13 kbit/s
- Multi-rate codecs and wide-band codecs available (2010's)

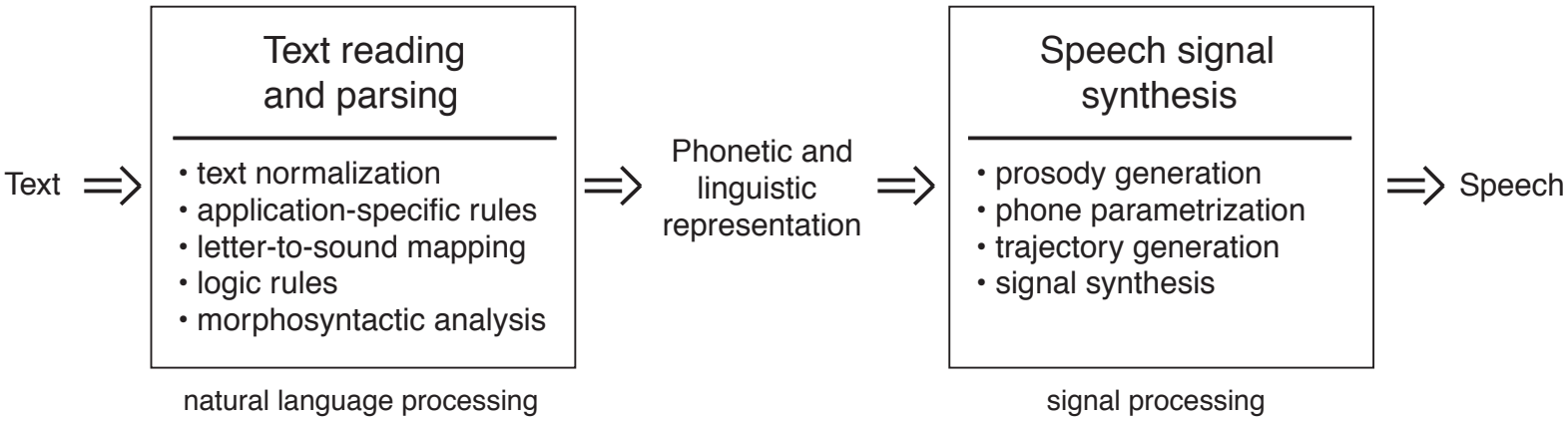
Speech coding with source-filter model



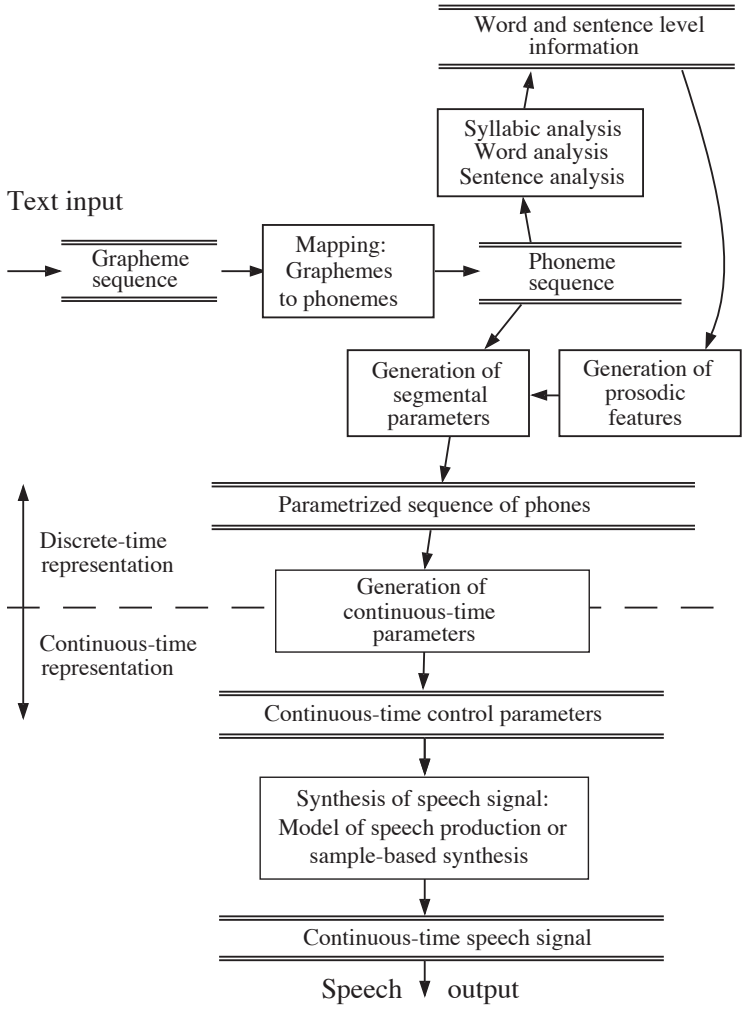
Text-to-speech synthesis

- Late 1800's, speech sounds with mechanical devices
- Digital speech synthesizers from 50's-70's
- Knowledge-based synthesis
- Unit-selection synthesis
- Statistical parametric synthesis

Knowledge-based speech synthesis



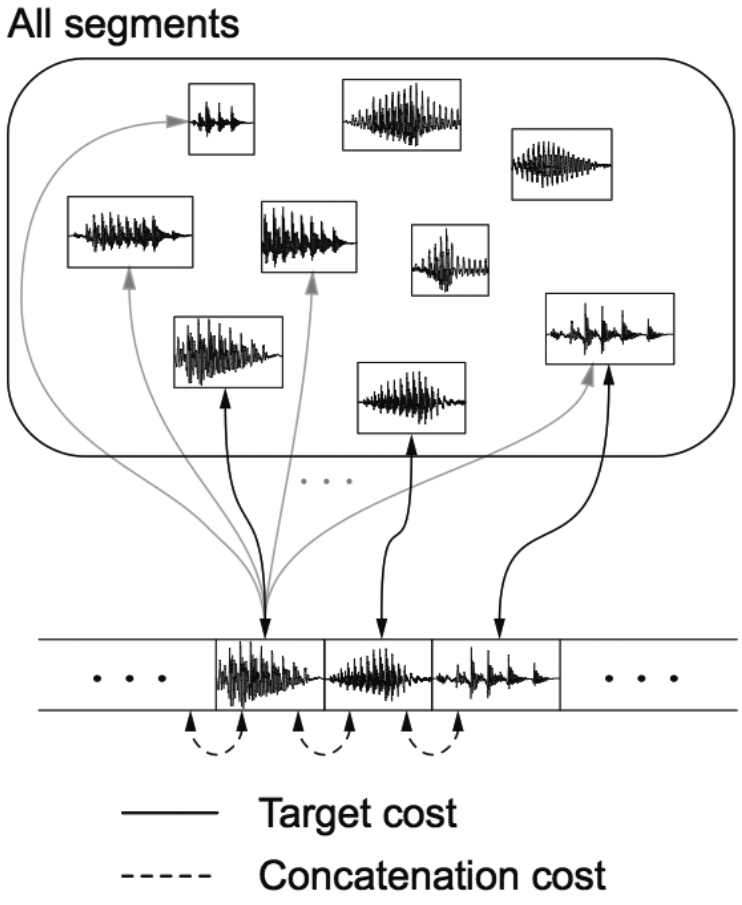
Knowledge-based speech synthesis



Unit-selection synthesis

- Tens of hours of recorded speech database
- Segmentation: individual phones, diphones, half-phones, syllables, morphemes, words, phrases, and sentences
- The text to be synthesized is given, and a set of required units is formed
- The desired target utterance is created by determining the best chain of candidate units from the database,

Unit-selection synthesis

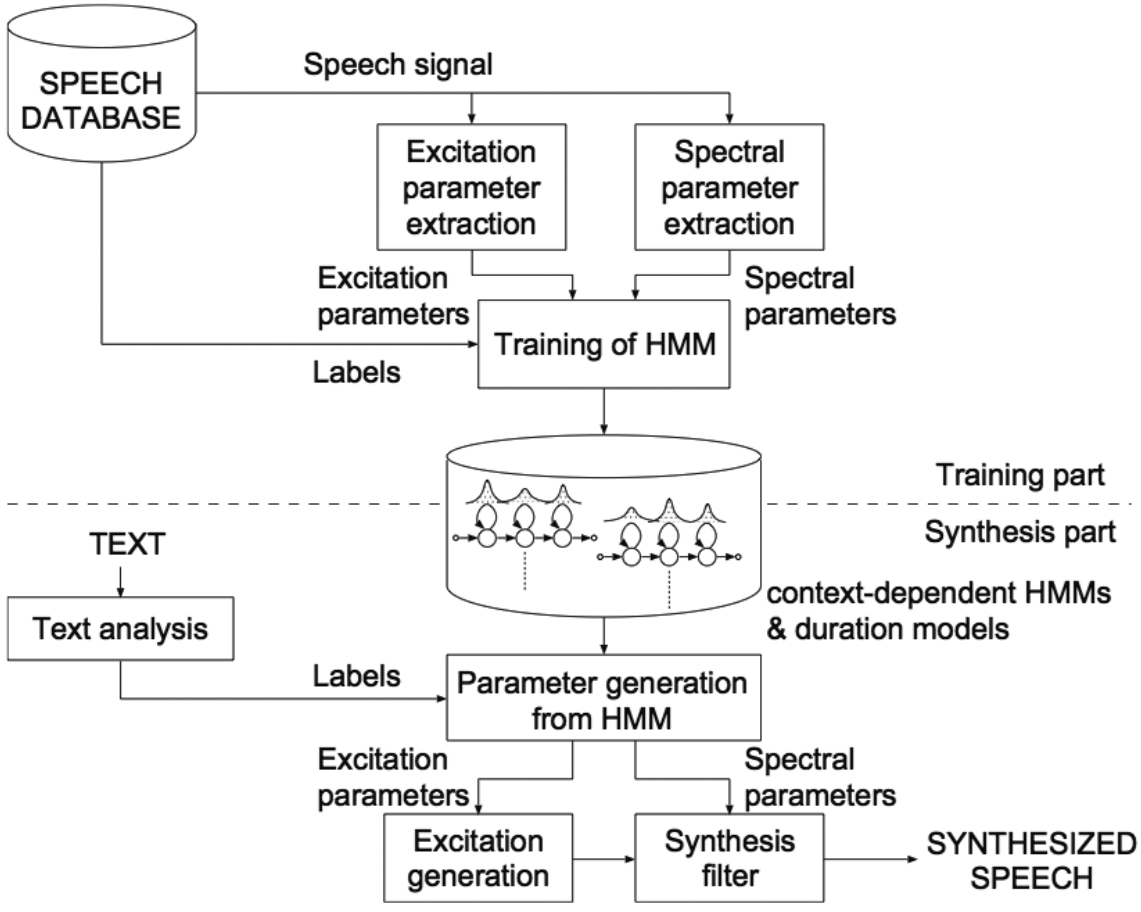


Adapted from (Zen et al., 2009)

Statistical parametric synthesis

- Tens of hours of recorded speech database
- Segmentation: individual phones, diphones, half-phones, syllables, morphemes, words, phrases, and sentences
- Hidden Markov models (HMMs) are trained with database
- Text is given, and HMMs are used to generate source-filter parameters
- Speech is synthesized

Statistical parametric synthesis



Adapted from (Zen et al., 2009)

Speech synthesis demos

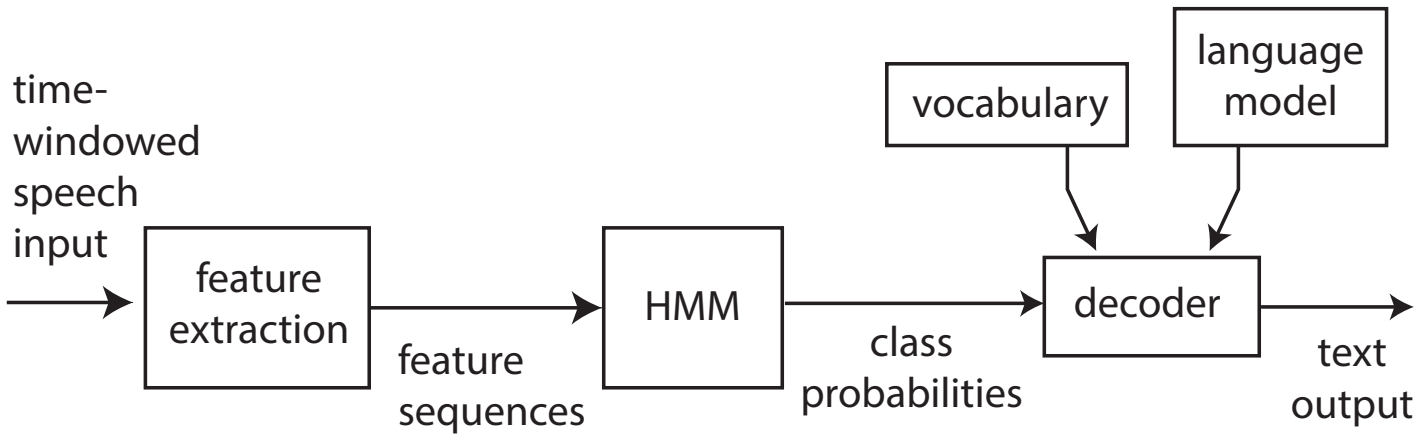
- ▸ Link to Voder (1939) video
- ▸ Link to replica of Voder (2020) video
- Voder sound demo
- Speak&spell toy, 1978
- Synte 2, 1970's (Karjalainen)
- CSTR Festival (2010's)
  http://www.cstr.ed.ac.uk/projects/festival/morevoices.html
  - HMM-based male
  - HMM-based female
  - Unit selection male
  - Unit selection female

Speech recognition

- Motivation
  - Human-computer interface
  - Cars, handheld devices, TV sets, automated telephone service
  - Future telephone service: speech recognition - instantaneous translation - speech synthesis
- Applications already available

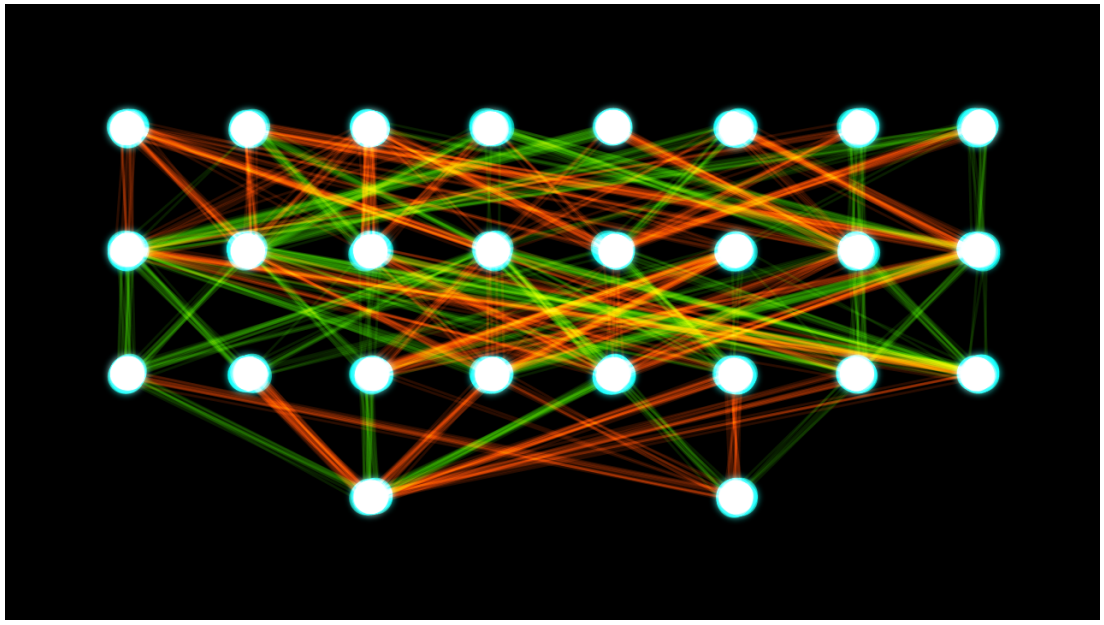
Speech recognition with HMMs

- Similar to speech synthesis, large speech database, segmentation, training etc.
- Speech input is windowed, features computed (e.g., mel-cepstra), HMM recognition
- Language model, vocabulary needed



Speech recognition with neural networks

- Deep learning, neural networks
- The gains of the net are trained with bigg amount of data
- State-of-the art results



Speech recognition

Performance depends heavily on the definition of the task.

- Simplest case
  - one known speaker uttering temporally separated words from a small, known vocabulary
- Harder cases
  - vocabulary is made larger or not restricted at all
  - words are not separated by silences
  - multi-language speech is allowed
  - content of speech is not limited to any specific topi
  - multiple speakers
  - the level of background noise is high.

References

These slides follow corresponding chapter in: Pulkki, V. and Karjalainen, M. Communication Acoustics: An Introduction to Speech, Audio and Psychoacoustics. John Wiley & Sons, 2015, *where also a more complete list of references can be found.*