

2. Introduction to Speech Processing

The Speech processing stack

Speech Applications: Coding, synthesis, recognition, understanding, speaker verification, language translation, speed-up/slow-down

Speech Measurements: energy, zero crossings, autocorrelations

Speech properties: speech-silence, voiced-unvoiced, pitch, formants

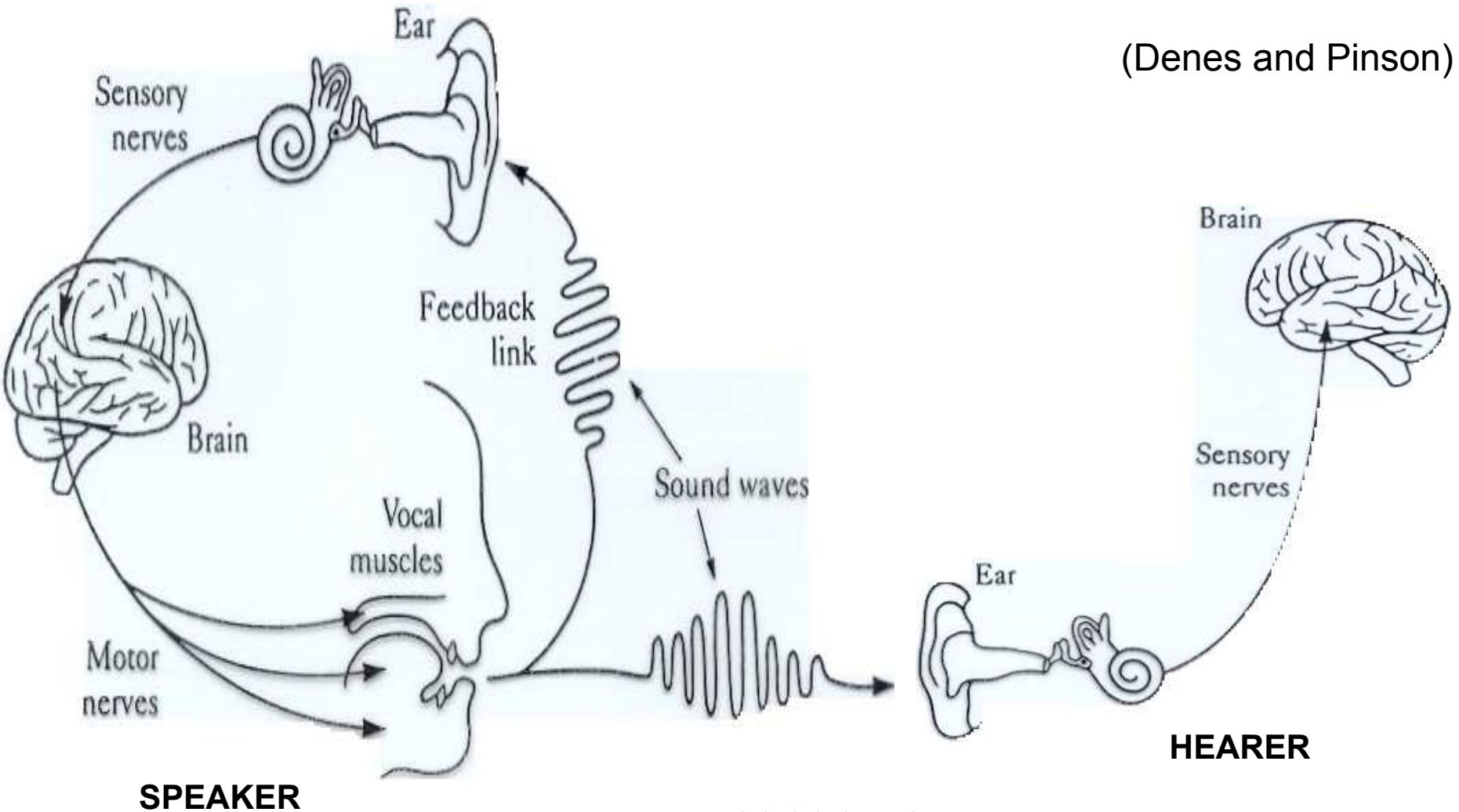
Speech representations: temporal, spectral, homomorphic, LPC

Fundamentals: acoustics, linguistics, pragmatics, speech perception

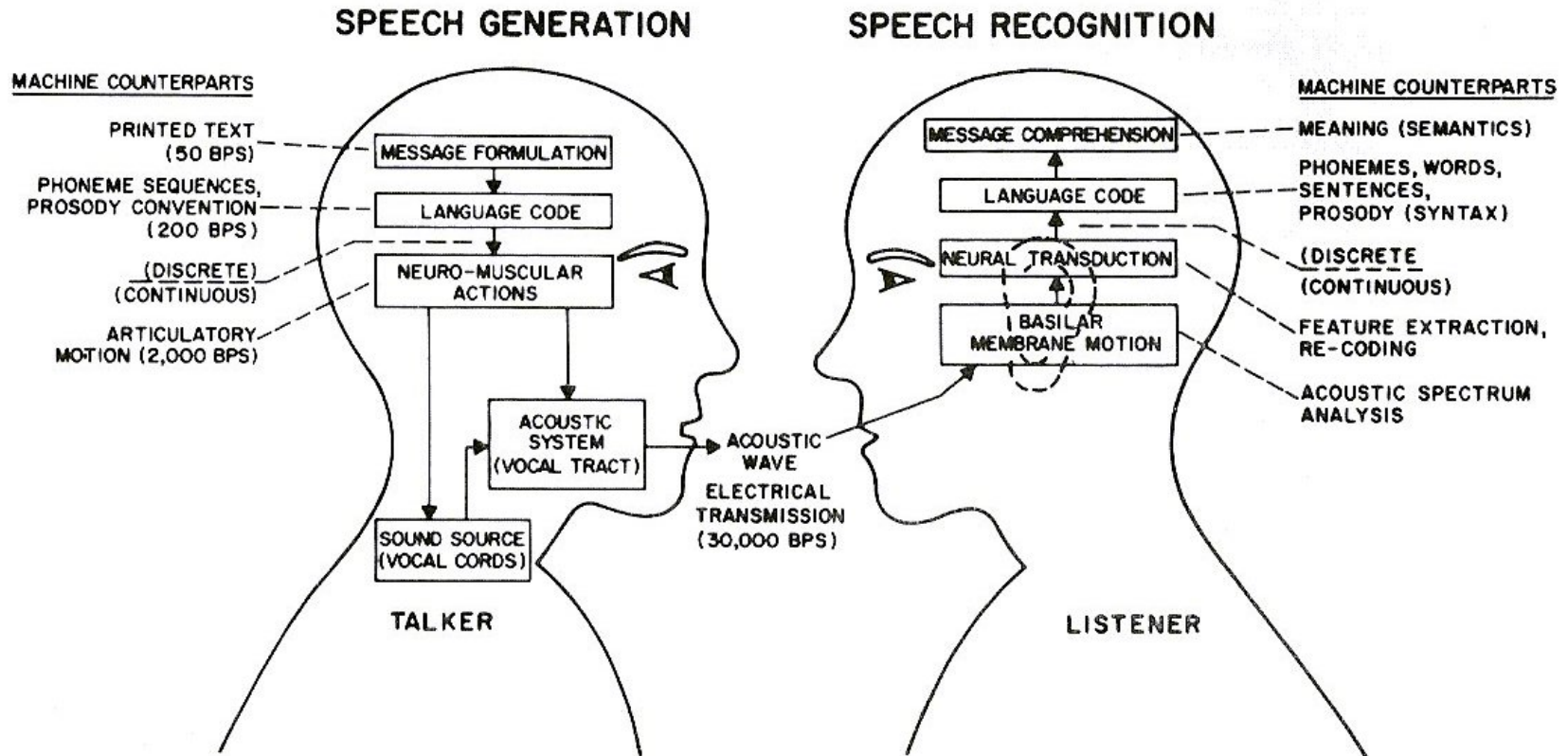
SPEECH GENERATION AND TRANSMISSION

Speech Chain

(Denes and Pinson)

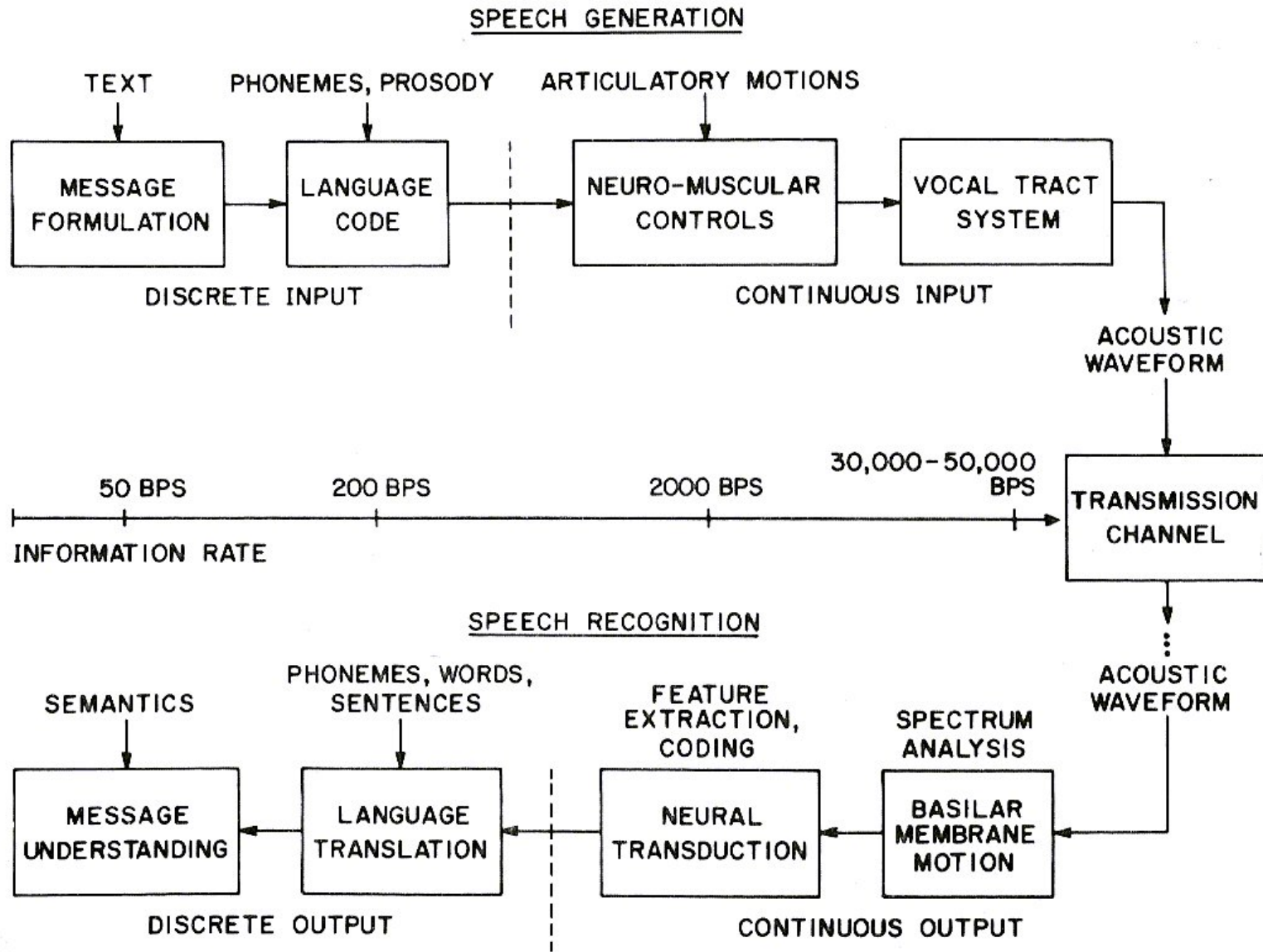


Speech Production/Perception



(after Flanagan)

Speech Processing Diagram



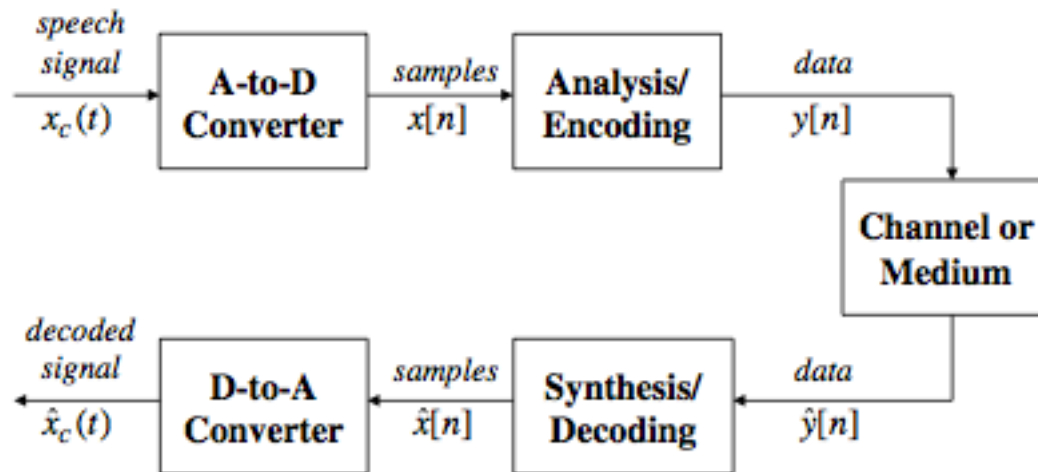
Application of digital speech processing

- Speech coding
- Speech Synthesis (from text to speech)
- Speech recognition
- Speaker/language recognition
- Many others

SPEECH CODING

Speech Coding

- The aim of speech coding is to compress (and then decompress) the speech waveform without any loss of *listenability* or *intelligibility*.
- Various standards exist for speech coding.
- The desired *bit rate* and associated quality of speech is highly application dependent.
 - Low bit rate: these basically have rates of between 75 and 2400 bps (bits-per-second).
 - Medium-to-high bit rate: operate at greater than 2400 bps.



Applications of Speech Coding

- Reduction in bit-rate for transmission/storage
- Speech enhancement (removal of noise)
- Allows the development of applications for
 - Security
 - High definition TV
 - Teleconference
 - Etc.

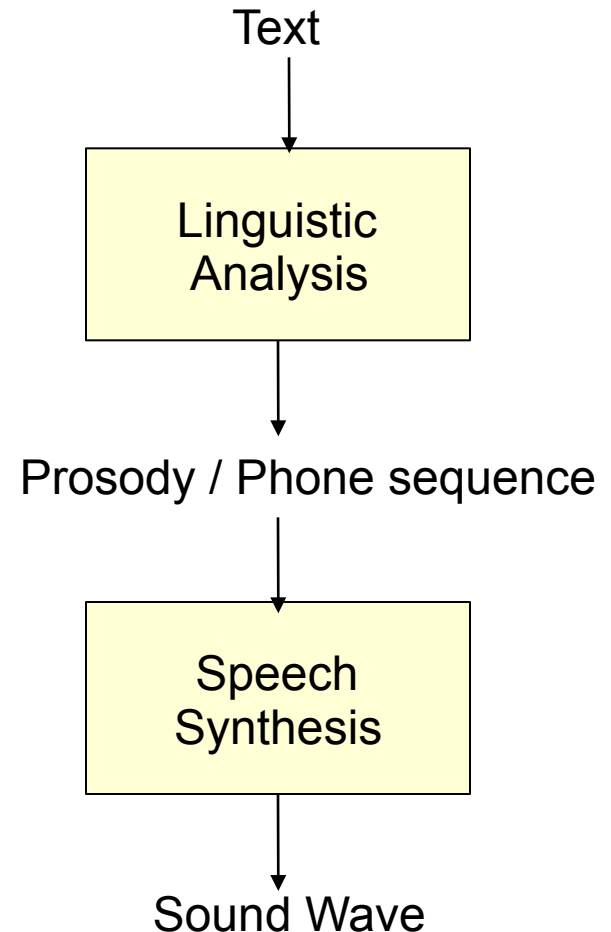
SPEECH SYNTHESIS

Speech Synthesis

- The aim of speech synthesis is to be able to take a word sequence and produce “human-like” speech

- Linguistic analysis stage: maps the input text into a standard form; determines the structure of the input, and finally decides how to pronounce it.

- Synthesis stage: converts the symbolic representation of what to say into an actual speech waveform.



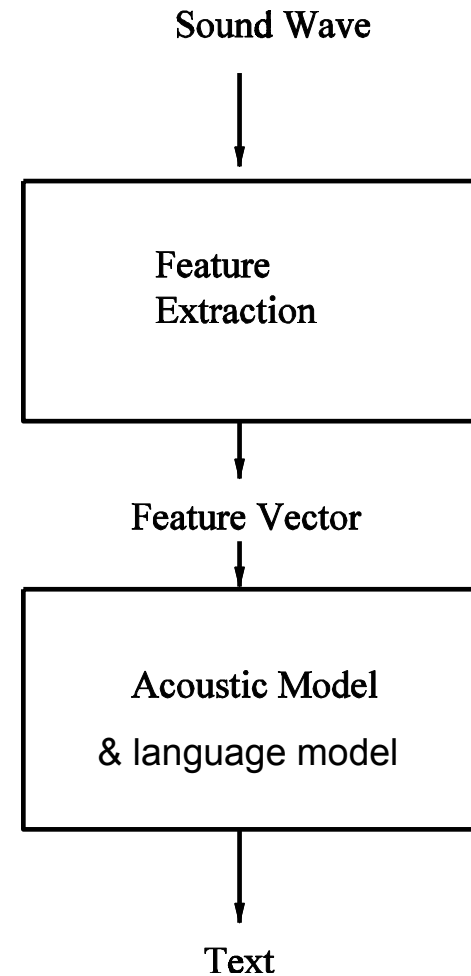
Applications of Speech Synthesis/Text-to-Speech (TTS)

- Games
- Telephone-based Information
 - directions, air travel, banking, etc.
 - Accessing variable information
 - Machine-human interfaces
- Eyes-free (in car)
- Reading/speaking for disabled
 - Reading of texts/books
 - Email access
- Education (Reading tutors)
- Alarm systems
-

AUTOMATIC SPEECH RECOGNITION (ASR)

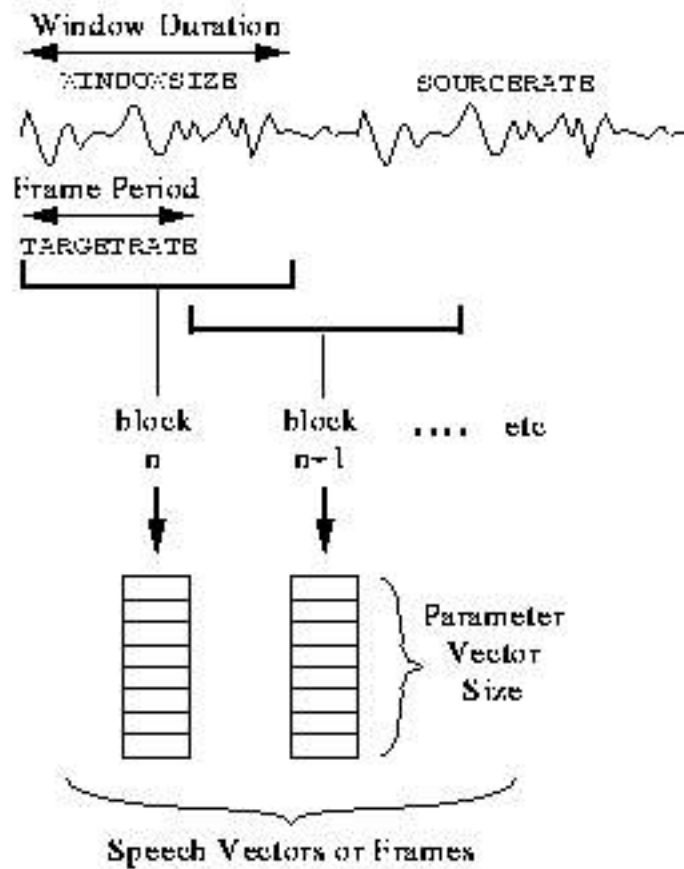
Automatic Speech Recognition

Automatic Speech Recognition (ASR) is the process of converting an unknown speech waveform into the corresponding orthographic transcription.



Extraction of feature vectors

Speech signal
Usually every 10ms,
25ms window



Acoustic features
Typically around 39

Current issues in ASR

Steady reduction has been achieved over the last 20 years in many domains. Still more research is needed:

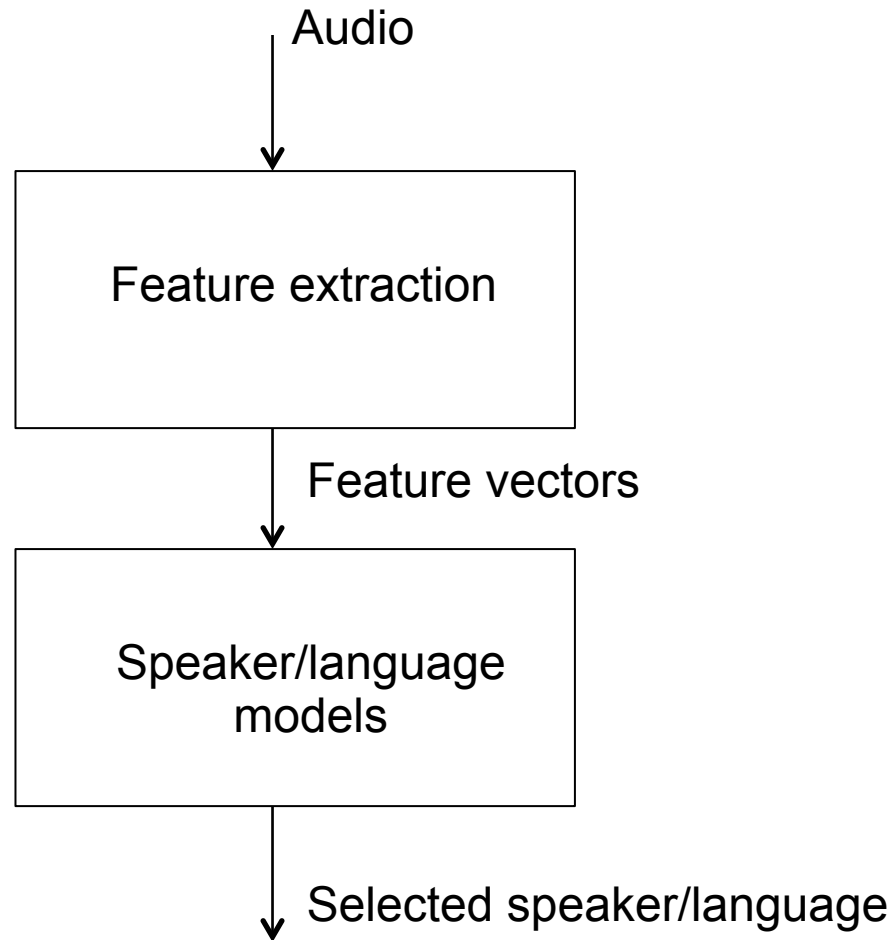
- Increase robustness for new acoustic environments
- Vocabulary increase and topic independence
- Improve OOV (out-of-vocabulary) recognition

Applications of Speech Recognition/ Understanding (ASR/ASU)

- Dictation
- Telephone-based Information
 - directions, air travel, banking, etc
 - Polls, online shopping
 - Call routing
- Hands-free
 - in car, computer, home(domotics), controlling tools
- Second language (accent reduction)
- Audio archive searching
- Help for disabled people

SPEAKER/LANGUAGE RECOGNITION

Speaker/Language identification



Applications of Speaker/Language Recognition

- Language recognition for call routing
- Speaker Recognition:

Speaker verification (binary decision)

- Voice password, telephone assistant

Speaker identification (one of N) (open set/closed set)

- Criminal investigation