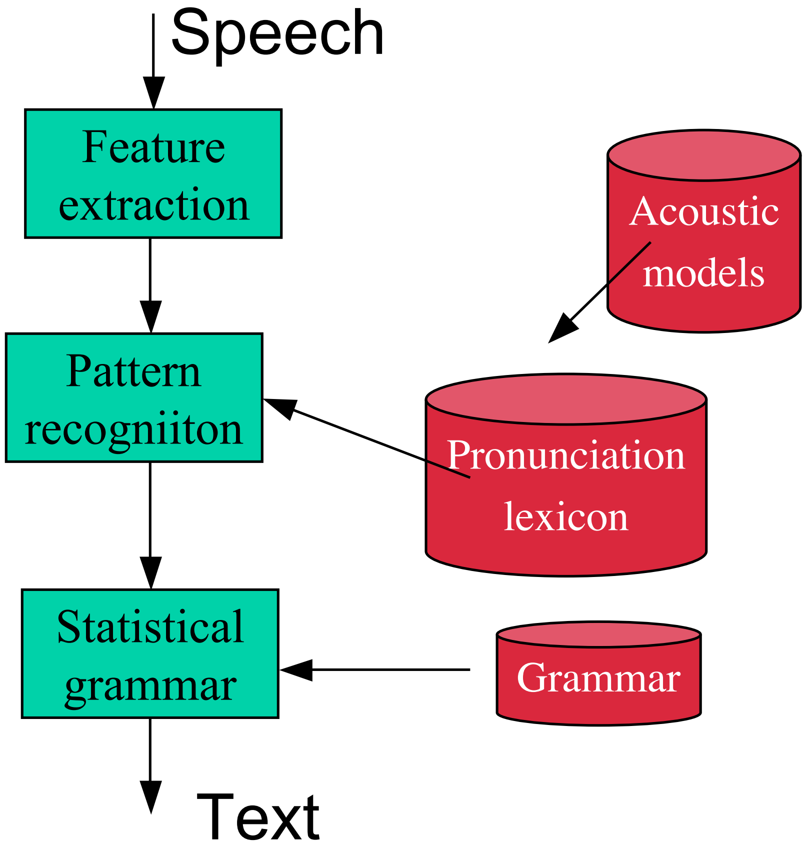
Speech recognition and speaker verification

Speech recognition

- Speech-to-Text
- International status: Good performance in controlled environments
- Problems:
  - Noise (background, line)
  - Speaker variation
  - Pronunciation variation, accents, dialects
  - Sentence patterns and ways of expression
- Need for robust speech recognition



Speech recognition

- Complexity (and performance) depends on:
  - Speech mode
    - Isolated utterances - continuous speech
  - Speaker mode
    - Speaker trained - speaker independent - speaker adaptive
  - Vocabulary (size and content)
  - Naturalness
    - Read speech/dictation
    - Spontaneous, natural speech
  - Noise environment

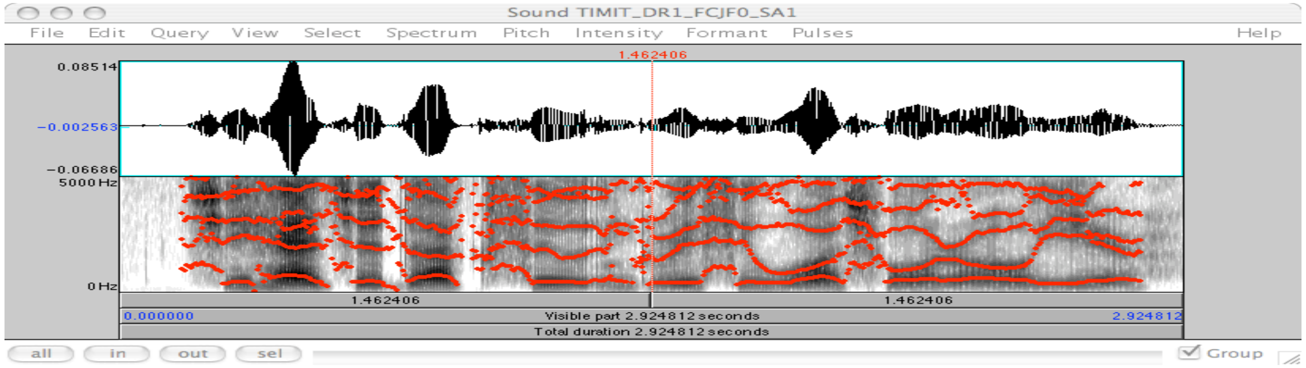
Intra-speaker variability

- Speaking rate and timing variability
- Speaking style
  - Read (careful) vs. spontaneous (casual)
  - Formal vs informal
  - Emontional state influences speech (neutral, happy, angry, afraid ...)
  - Environment influences speech - Lombard effect
- Co-articulation
  - Phonetic context influences pronunciation

Inter-speaker variability

- Differences in physiology
  - E.g. vocal tract length
- Voice quality differences
  - Age, creakiness, nasality
- Accent/dialect variations
- Sociolinguistic variations
- Individual speaking characteristics

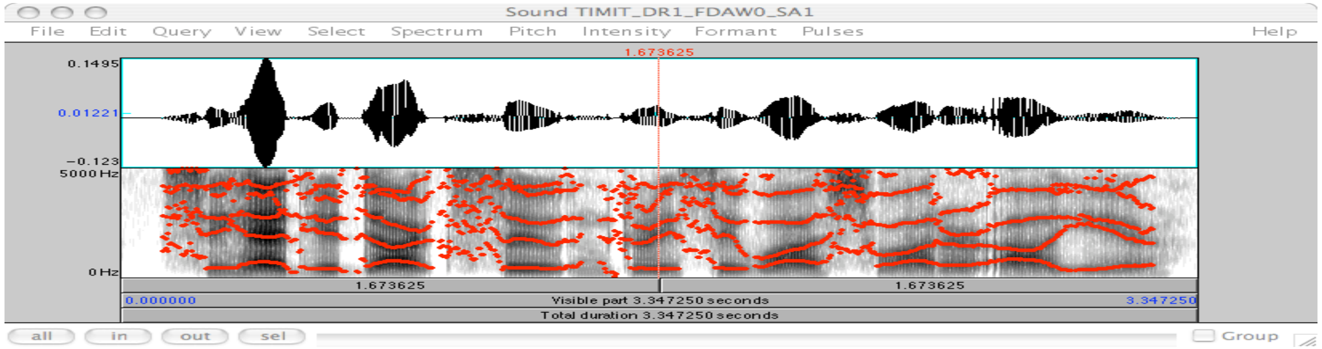
Inter-speaker variability



Sound TIMIT_DR1_FCJF0_SA1

Visible part 2.924812 seconds

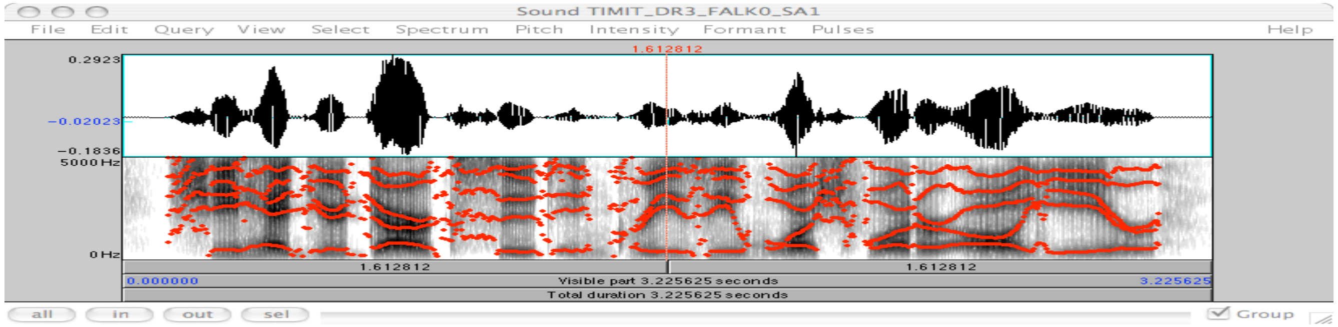
Total duration 2.924812 seconds



Sound TIMIT_DR1_FDAW0_SA1

Visible part 3.347250 seconds

Total duration 3.347250 seconds



Sound TIMIT_DR3_FALK0_SA1

Visible part 3.225625 seconds

Total duration 3.225625 seconds

Environmental influence

- Background noise
  - Traffic, office equipment, factory noise, doors and bells
- Transmission noise and channel distortion in telecommunications
- Room reverberation
- Microphone characteristics

Some important ASR types

- Dictation
  - Transcription of speech
  - Continuous speech, large vocabulary
  - Can be speaker trained
  - All recognition errors are in principle equally important
- Command and control
  - Short commands (one word or short sentence)
  - Limited vocabulary
  - Translation of spoken utterance to an action
- Speech understanding, dialogue systems
  - Literal transcription unimportant, capturing relevant *meaning* paramount
  - Key words/phrases contain the relevant information
  - Semantic processing, NLP

Different types require different design criteria!

Speech recognition performance

Correct: I constantly make severe new errors
Recognized: I count to make several _ errors

- Error types:
  - Substitutions (S)
  - Deletions (D)
  - Insertions (I)

- Percent correct = 100*(N-D-S)/N
  - Where N is the number of words in the (correct) sentence
- Percent accuracy = 100*(N-D-S-I)/N
- Word error rate = 100*(D+S+I)/N

Speech recognition - performance

Task	Type	Vocabulary	WER
Connected digits	Read	10	<0.3%
Air traffic information	Spontaneous	2.500	2%
Wall Street Journal	Read	64.000	7%
Radio news	Mixed	64.000	30%
”Call home”	Conversational	10.000	50%

Source: IEEE Spectrum, Jan. 1997

Outline

- Feature extraction
- Template matching and dynamic programming
- Hidden Markov Models for speech recognition
- Adaptation
- Speaker verification