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
  - Emotional 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
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

<table>
<thead>
<tr>
<th>Task</th>
<th>Type</th>
<th>Vocabulary</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connected digits</td>
<td>Read</td>
<td>10</td>
<td>&lt;0.3%</td>
</tr>
<tr>
<td>Air traffic information</td>
<td>Spontaneous</td>
<td>2,500</td>
<td>2%</td>
</tr>
<tr>
<td>Wall Street Journal</td>
<td>Read</td>
<td>64,000</td>
<td>7%</td>
</tr>
<tr>
<td>Radio news</td>
<td>Mixed</td>
<td>64,000</td>
<td>30%</td>
</tr>
<tr>
<td>”Call home”</td>
<td>Conversational</td>
<td>10,000</td>
<td>50%</td>
</tr>
</tbody>
</table>

Outline

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