74.419 Artificial Intelligence 2004
Speech & Natural Language Processing

• Natural Language Processing
  • written text as input
  • sentences (well-formed)

• Speech Recognition
  • acoustic signal as input
  • conversion into written words

• Spoken Language Understanding
  • analysis of spoken language (transcribed speech)
Figure 1.4 Basic system architecture of a spoken language understanding system.
Speech & Natural Language Processing

Areas in Natural Language Processing
- Morphology
- Grammar & Parsing (syntactic analysis)
- Semantics
- Pragamatics
- Discourse / Dialogue
- Spoken Language Understanding

Areas in Speech Recognition
- Signal Processing
- Phonetics
- Word Recognition
Speech Production & Reception

Sound and Hearing

• change in **air pressure** ≡ **sound wave**
• reception through **inner ear membrane** / **microphone**
• break-up into **frequency components**: receptors in **cochlea** / mathematical frequency analysis (e.g. Fast-Fourier Transform FFT) → **Frequency Spectrum**
• perception/recognition of **phonemes** and subsequently **words** (e.g. **Neural Networks**, **Hidden-Markov Models**)
Figure 2.10 The structure of the peripheral auditory system with the outer, middle, and inner ear.
Speech Recognition Phases

Speech Recognition

• acoustic signal as input
• signal analysis - spectrogram
• feature extraction
• phoneme recognition
• word recognition
• conversion into written words
Speech Signal

Speech Signal
composed of different (sinus) waves with different frequencies and amplitudes
• Frequency - waves/second \(\cong\) like pitch
• Amplitude - height of wave \(\cong\) like loudness

+ noise  (not sinus wave)

Speech Signal
composite signal comprising different frequency components
Waveform (fig. 7.20)

Amplitude/Pressure

"She just had a baby."
Waveform for Vowel ae (fig. 7.21)
Speech Signal Analysis

Analog-Digital Conversion of Acoustic Signal

Sampling in Time Frames ("windows")

- Frequency = 0-crossings per time frame
  → e.g. 2 crossings/second is 1 Hz (1 wave)
  → e.g. 10kHz needs sampling rate 20kHz

- Measure amplitudes of signal in time frame
  → digitized wave form

- Separate different frequency components
  → FFT (Fast Fourier Transform)
  → spectrogram

- Other frequency based representations
  → LPC (linear predictive coding),
  → Cepstrum
Waveform and Spectrogram (figs. 7.20, 7.23)
Waveform and LPC Spectrum for Vowel ae
(figs. 7.21, 7.22)

Amplitude/Pressure

Energy

Formants

Time

Frequency
Speech Signal Characteristics

From Signal Representation derive, e.g.

- **formants** - dark stripes in spectrum
  strong frequency components; characterize particular vowels; gender of speaker

- **pitch** – fundamental frequency
  baseline for higher frequency harmonics like formants; gender characteristic

- **change in frequency distribution**
  characteristic for e.g. plosives (form of articulation)
Figure 2.4 A schematic diagram of the human speech production apparatus.
Video of glottis and speech signal in lingWAVES (from http://www.lingcom.de)
Table 2.9 Manner of articulation of English consonants.

<table>
<thead>
<tr>
<th>Consonant Labels</th>
<th>Consonant Examples</th>
<th>Voiced?</th>
<th>Manner</th>
</tr>
</thead>
<tbody>
<tr>
<td>b</td>
<td>big, able, tab</td>
<td>+</td>
<td>plosive</td>
</tr>
<tr>
<td>p</td>
<td>put, open, tap</td>
<td>-</td>
<td>plosive</td>
</tr>
<tr>
<td>d</td>
<td>dig. idea, wad</td>
<td>+</td>
<td>plosive</td>
</tr>
<tr>
<td>t</td>
<td>talk, sat</td>
<td>-</td>
<td>plosive</td>
</tr>
<tr>
<td>g</td>
<td>gut, angle, tag</td>
<td>+</td>
<td>plosive</td>
</tr>
<tr>
<td>k</td>
<td>cut, oaken, take</td>
<td>-</td>
<td>plosive</td>
</tr>
<tr>
<td>v</td>
<td>vat, over, have</td>
<td>+</td>
<td>fricative</td>
</tr>
<tr>
<td>j</td>
<td>fork, after, if</td>
<td>-</td>
<td>fricative</td>
</tr>
<tr>
<td>z</td>
<td>zap, lazy, haze</td>
<td>+</td>
<td>fricative</td>
</tr>
<tr>
<td>s</td>
<td>sit, cass, toss</td>
<td>-</td>
<td>fricative</td>
</tr>
<tr>
<td>dh</td>
<td>then, father, scythe</td>
<td>+</td>
<td>fricative</td>
</tr>
<tr>
<td>th</td>
<td>thin, nothing, truth</td>
<td>-</td>
<td>fricative</td>
</tr>
<tr>
<td>zh</td>
<td>genre, azure, beige</td>
<td>+</td>
<td>fricative</td>
</tr>
<tr>
<td>sh</td>
<td>she, cushion, wash</td>
<td>-</td>
<td>fricative</td>
</tr>
<tr>
<td>jh</td>
<td>joy, agile, edge</td>
<td>+</td>
<td>affricate</td>
</tr>
<tr>
<td>ch</td>
<td>chin, archer, march</td>
<td>-</td>
<td>affricate</td>
</tr>
<tr>
<td>l</td>
<td>lid, elbow, sail</td>
<td>+</td>
<td>lateral</td>
</tr>
<tr>
<td>r</td>
<td>red, part, far</td>
<td>+</td>
<td>retroflex</td>
</tr>
<tr>
<td>y</td>
<td>yacht, onion, yard</td>
<td>+</td>
<td>glide</td>
</tr>
<tr>
<td>w</td>
<td>with, away</td>
<td>+</td>
<td>glide</td>
</tr>
<tr>
<td>hh</td>
<td>help, ahead, hotel</td>
<td>+</td>
<td>fricative</td>
</tr>
<tr>
<td>m</td>
<td>mat, amid, aim</td>
<td>+</td>
<td>nasal</td>
</tr>
<tr>
<td>n</td>
<td>no, end, pan</td>
<td>+</td>
<td>nasal</td>
</tr>
<tr>
<td>ng</td>
<td>sing, anger, drink</td>
<td>+</td>
<td>nasal</td>
</tr>
</tbody>
</table>

Figure 2.21 The major places of consonant articulation with respect to the human mouth.
Table 2.7 Phonological (abstract) feature decomposition of basic English vowels.

<table>
<thead>
<tr>
<th>Vowel</th>
<th>high</th>
<th>low</th>
<th>front</th>
<th>back</th>
<th>round</th>
<th>tense</th>
</tr>
</thead>
<tbody>
<tr>
<td>iy</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>ih</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ae</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>aa</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>ah</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ao</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>ax</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>eh</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ow</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>uh</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>uw</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>

Table 2.8 Consonant manner of articulation.

<table>
<thead>
<tr>
<th>Manner</th>
<th>Sample Phon</th>
<th>Example Words</th>
<th>Mechanism</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plosive</td>
<td>/p/</td>
<td>tat, tap</td>
<td>Closure in oral cavity</td>
</tr>
<tr>
<td>Nasal</td>
<td>/m/</td>
<td>team, meet</td>
<td>Closure of nasal cavity</td>
</tr>
<tr>
<td>Fricative</td>
<td>/s/</td>
<td>sick, kiss</td>
<td>Turbulent airstream noise</td>
</tr>
<tr>
<td>Retroflex liquid</td>
<td>/ɾ/</td>
<td>rat, tar</td>
<td>Vowel-like, tongue high and curled back</td>
</tr>
<tr>
<td>Lateral liquid</td>
<td>/ɹ/</td>
<td>lean, kneel</td>
<td>Vowel-like, tongue central, side airstream</td>
</tr>
<tr>
<td>Glide</td>
<td>/ʃ/, /w/</td>
<td>yes, well</td>
<td>Vowel-like</td>
</tr>
</tbody>
</table>
Figure 2.23 Spectrogram: *bet*, *debt*, and *get* (separated by vertical lines). Note different relative spreads of *F1* and *F2* following the initial stop consonants in each word.
Phoneme Recognition

Recognition Process based on
• features extracted from spectral analysis
• phonological rules
• statistical properties of language/ pronunciation

Recognition Methods
• Hidden Markov Models
• Neural Networks
• Pattern Classification in general
Pronunciation Networks / Word Models as Probabilistic FAs (fig 5.12)

Word model with dialect variation:

Word model with coarticulation and dialect variation:
Pronunciation Network for 'about'
(fig 5.13)
Word Recognition with Probabilistic FA / Markov Chain (fig 5.14)
The **Viterbi Algorithm** finds an optimal sequence of states in continuous Speech Recognition, given an observation sequence of phones and a probabilistic (weighted) FA (state graph). The algorithm returns the path through the automaton which has **maximum probability** and accepts the observation sequence.

$a[s,s']$ is the transition probability (in the phonetic word model) from current state $s$ to next state $s'$, and $b[s',o_t]$ is the observation likelihood of $s'$ given $o_t$. $b[s',o_t]$ is 1 if the observation symbol matches the state, and 0 otherwise.
**Viterbi-Algorithm (fig 5.19)**

```plaintext
function VITERBI(observations of len T, state-graph) returns best-path
num-states ← NUM-OF-STATES(state-graph)
Create a path probability matrix viterbi[num-states+2,T+2]
viterbi[0,0] ← 1.0
for each time step t from 0 to T do
    for each state s from 0 to num-states do
        for each transition s' from s in state-graph
            new-score ← viterbi[s,t] * a[s,s'] * b[s',(o_t)]
            if ((viterbi[s',t+1] = 0) || (new-score > viterbi[s',t+1]))
                then viterbi[s',t+1] ← new-score
                back-pointer[s',t+1] ← s
Backtrace from highest probability state in the final column of viterbi[] and return path
```
**Viterbi-Algorithm Explanation** (cf. Jurafsky Ch.5)

The Viterbi Algorithm sets up a probability matrix, with one column for each time index $t$ and one row for each state in the state graph. Each column has a cell for each state $q_i$ in the single combined automaton for the competing words (in the recognition process).

The algorithm first creates $N+2$ state columns. The first column is an initial pseudo-observation, the second corresponds to the first observation-phone, the third to the second observation and so on. The final column represents again a pseudo-observation. In the first column, the probability of the Start-state is initially set to 1.0; the other probabilities are 0. Then we move to the next state.

For every state in column 0, we compute the probability of moving into each state in column 1. The value $viterbi[t, j]$ is computed by taking the maximum over the extensions of all the paths that lead to the current cell. An extension of a path at state $i$ at time $t-1$ is computed by multiplying the three factors:

- **the previous path probability** from the previous cell $forward[t-1, i]$
- **the transition probability** $a_{i,j}$ from previous state $i$ to current state $j$
- **the observation likelihood** $b_{jt}$ that current state $j$ matches observation symbol $t$.

$b_{jt}$ is 1 if the observation symbol matches the state; 0 otherwise.
Speech Recognition

Acoustic / sound wave

Frequency Spectrum

Features (Phonemes; Context)

Phonemes

Phoneme Sequences / Words

Word Sequence / Sentence

Filtering, Sampling
Spectral Analysis; FFT

Signal Processing / Analysis

Phoneme Recognition:
HMM, Neural Networks

Grammar or Statistics

Grammar or Statistics for likely word sequences
Speech Recognizer Architecture (fig. 7.2)
Speech Processing - Important Types and Characteristics

single word vs. continuous speech
unlimited vs. large vs. small vocabulary
speaker-dependent vs. speaker-independent training
Speech Recognition vs. Speaker Identification
Additional References


Figures taken from:


lingWAVES (from http://www.lingcom.de