Thursday 8/26/99
Note-2
Fundamentals of Spoken Language Processing
CECS401
C. Summary of current system performances

Automatic speech recognition

<table>
<thead>
<tr>
<th>Task</th>
<th>Style</th>
<th>Vocabulary size</th>
<th>Word-error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>connected digit string</td>
<td>spontaneous</td>
<td>10</td>
<td>0.3%</td>
</tr>
<tr>
<td>airline travel information</td>
<td>spontaneous</td>
<td>2,500</td>
<td>2.0%</td>
</tr>
<tr>
<td>wall street journal</td>
<td>read</td>
<td>64,000</td>
<td>8.0%</td>
</tr>
<tr>
<td>radio broadcast</td>
<td>mixed</td>
<td>64,000</td>
<td>27%</td>
</tr>
<tr>
<td>switch board</td>
<td>conversation</td>
<td>10,000</td>
<td>38%</td>
</tr>
<tr>
<td>recorded telephone speech</td>
<td>conversation</td>
<td>-</td>
<td>54%</td>
</tr>
</tbody>
</table>

by Dr. John Makhoul, BBN, 1998.
Look up the web, for example, help://www.tiec.net/users/rwilcox/speech.

### Commercial Products

By Dr. Rich Cox, 1998

<table>
<thead>
<tr>
<th>Language</th>
<th>Quality</th>
<th>Intelligibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japanese, French</td>
<td>High</td>
<td>English</td>
</tr>
<tr>
<td>Scottish</td>
<td>Low</td>
<td>High</td>
</tr>
</tbody>
</table>

### Text-to-Speech Synthesis

<table>
<thead>
<tr>
<th>Codec Type (per sec)</th>
<th>96 - 128 KB</th>
<th>64 KB</th>
<th>16 KB</th>
<th>6.4 KB</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD quality audio</td>
<td>Wideband audio</td>
<td>Wideband speech</td>
<td>Narrowband speech</td>
<td>Real-time speech</td>
</tr>
</tbody>
</table>

Speech & audio compression
"We speak with our ears."

Harvey Fletcher (1953)

Speaking and hearing is an interactive process.

Topic-2 Speech Production
Four grades of audio signal bandwidth:

High frequency components (3,400 – 7,000 Hz) provide greater intelligibility for speech and 이제 difference (e.g., in all vs. f in fil)

Mid frequency components (200 – 3,400 Hz) contribute to the distinction of sounds and presence of speech

Low frequency components (50 – 200 Hz) contribute to the bulkiness and sound of vibration waves in the frequency range of 20 to 20,000 Hz

A sound
Human ear is most sensitive to the frequency band of 500 Hz to 10,000 Hz.

\[ P = 10 \log \left( \frac{I/f^2}{I_0} \right) \]

Intensity level is defined as

\[ I_{air} = 10^{-12} \text{ watts/m}^2 \text{ in air} \]

Reference intensity is defined as

Audible intensity (I) range from 10^-12 to 10 watts/m^2

Sound perception by human ears:
Fig. 11.5. Thresholds and free-field equal-loudness-level contours for pure tones with various frequencies and intensities.
Cross-section of vocal apparatus of an adult

B. Mechanism of Speech Production
**Eating:**

- Food mouth esophagus (the gate under the epiglottis) to the stomach
- Teeth

**Breathing:**

- Air nostrils nasal cavity

**Speaking, breathing, coughing:**

**Functions of vocal apparatus:**

The measures are based on average adult males.

<table>
<thead>
<tr>
<th>nasal tract</th>
<th>oral tract</th>
<th>larynx</th>
<th>lips</th>
<th>glottis</th>
<th>vocal tract</th>
<th>start</th>
<th>end</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 5 cm sg</td>
<td>0 - 20 cm sg</td>
<td>larynx</td>
<td>length</td>
<td>cross-section area</td>
<td>larynx</td>
<td>length</td>
<td>cross-section area</td>
</tr>
</tbody>
</table>
The frequency (F₀) of the vocal cords is called fundamental frequency (F₀).

To produce voiced sounds, the vocal cords vibrate, generating quasiperiodic pulses of air. This excites the vocal tract, producing voiced sounds (e.g., "a") in speaking.

<table>
<thead>
<tr>
<th>200-600</th>
<th>150-400</th>
<th>70-200</th>
</tr>
</thead>
<tbody>
<tr>
<td>children</td>
<td>women</td>
<td>men</td>
</tr>
</tbody>
</table>

Built up against to force the cords apart builds up pressure within the air pressure between the vocal cords apart. Pressure is reduced to draw the two cords together. Vibration of vocal cords produces voiced sounds.
(e.g. p)

Production of plosive sounds (e.g. p)

Release → Produce unvoiced sounds

Lateral tracheal muscles → Push air out of the lungs → Form constriction

Production of voiced sounds (e.g. b, v)

Describe the shape of vocal tract and determine the type of voiced sound.

The positions of the articulators (jaw, tongue, velum, lips, teeth)
optional surrounding consonants
at least one syllable (each syllable must contain a vowel with
The smallest possible unit of words, every word must contain:

Syllable:
the phone unit | in bishop and king
Phone produced in the context of neighboring phones, for example:

Allphone:
    Acoustics realizations of phonemes

Phoneme:
Language down in a systematic and unambiguous way
A set of abstract symbolic units that can be used for writing:

C. Acoustic-phonetics
<table>
<thead>
<tr>
<th>phoneme</th>
<th>ARPABET</th>
<th>examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>/a/</td>
<td>/a/</td>
</tr>
<tr>
<td>aɪ</td>
<td>/aɪ/</td>
<td>/aɪ/</td>
</tr>
<tr>
<td>aʊ</td>
<td>/aʊ/</td>
<td>/aʊ/</td>
</tr>
<tr>
<td>ɑ</td>
<td>/ɑ/</td>
<td>/ɑ/</td>
</tr>
<tr>
<td>aɾ</td>
<td>/aɾ/</td>
<td>/aɾ/</td>
</tr>
<tr>
<td>ɑɪ</td>
<td>/ɑɪ/</td>
<td>/ɑɪ/</td>
</tr>
<tr>
<td>ɑʊ</td>
<td>/ɑʊ/</td>
<td>/ɑʊ/</td>
</tr>
<tr>
<td>ɒ</td>
<td>/ɒ/</td>
<td>/ɒ/</td>
</tr>
<tr>
<td>ɑʊ̯</td>
<td>/ɑʊ̯/</td>
<td>/ɑʊ̯/</td>
</tr>
<tr>
<td>æ</td>
<td>/æ/</td>
<td>/æ/</td>
</tr>
<tr>
<td>æɪ</td>
<td>/æɪ/</td>
<td>/æɪ/</td>
</tr>
<tr>
<td>æʊ</td>
<td>/æʊ/</td>
<td>/æʊ/</td>
</tr>
<tr>
<td>ɛɪ</td>
<td>/ɛɪ/</td>
<td>/ɛɪ/</td>
</tr>
<tr>
<td>ɛʊ</td>
<td>/ɛʊ/</td>
<td>/ɛʊ/</td>
</tr>
<tr>
<td>ɛ̄</td>
<td>/ɛ̄/</td>
<td>/ɛ̄/</td>
</tr>
<tr>
<td>ɛ̄ɪ̯</td>
<td>/ɛ̄ɪ̯/</td>
<td>/ɛ̄ɪ̯/</td>
</tr>
<tr>
<td>ɛ̄ʊ̯</td>
<td>/ɛ̄ʊ̯/</td>
<td>/ɛ̄ʊ̯/</td>
</tr>
<tr>
<td>ɛ̄ʊ̯̄</td>
<td>/ɛ̄ʊ̯̄/</td>
<td>/ɛ̄ʊ̯̄/</td>
</tr>
<tr>
<td>ɛ̄ɪ̯̄</td>
<td>/ɛ̄ɪ̯̄/</td>
<td>/ɛ̄ɪ̯̄/</td>
</tr>
<tr>
<td>ɛ̄ʊ̯̄̄</td>
<td>/ɛ̄ʊ̯̄̄/</td>
<td>/ɛ̄ʊ̯̄̄/</td>
</tr>
</tbody>
</table>

TABLE 2.1: A condensed list of phonetic symbols for American English.
Figure 2.13
Articulatory configurations for typical vowel sounds (after Flanagan [3]).
Radiation is through nostrils.
Oral tract is closed by lips or tongue, nasal tract is open.
Excited by vocal cords vibration.
Nasal consonants and open.
Articulated in a similar way as vowel, but vocal tract is less steady.
Semi-vowels
Movement from one vowel to another within a single syllable.
Diphthongs
Similar to unvoiced fricatives but vocal cords are vibrating.

Voiced Fricatives

Radiation is through mouth

Vocal tract is constricted by tongue or teeth

Excited by steady air flow

Unvoiced Fricatives
These represent vocal cord operation (adapted from Potter, Kope, and Green).
Voiceless and unvoiced stops (pl护士)

Release pressure suddenly.

Constriction is formed by lips or tongue.

Press period vocal cords vibrate for voiced stops.

Form a constriction in the oral tract and build up pressure.

(1) (2) (3)
slacken the constriction to produce turbulent air
form a constriction in the oral tract and build up pressure
articulate
vocal tract is constricted at glottis
excited by steady air flow
whisper