University of California
Berkeley

College of Engineering
Department of Electrical Engineering
and Computer Sciences

Professors : N.Morgan / B.Gold

EE225D

Human Speech Recognition

Lecture 19
Human Speech Recognition

• Experiments dating from 1918 dealing with noise and reduced BW

• Statistics of CVC perception

• Comparisons between humans and today’s best systems

• Brief peak at human-inspired ASR research
Assessing Recognition Accuracy

• Articulation

• Intelligibility
Fletcher Experiments

• CVC, VC, CV nonsense syllables
• 74% of syllables used
• Tests over different SNR, bands
<table>
<thead>
<tr>
<th>No.</th>
<th>The first group is</th>
<th>I will now say</th>
<th>Write down</th>
<th>As the fourth write</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>Listen carefully to</td>
<td>The last group is</td>
<td>n̄, ŉ, ŏ, ŏ́, ŏ́̀</td>
<td>p̆́, y̆́, y̆́́, k̆́, ṭ̆́</td>
</tr>
<tr>
<td>20</td>
<td>Please try to hear</td>
<td>nğ́, n̆́, ŋ̆́, ŋ̆́́, ŋ̆́̀</td>
<td>t̆́, t̆́́, d̆́, d̆́́, t̆́̀</td>
<td>d̆́, d̆́́, k̆́, ṭ̆́</td>
</tr>
<tr>
<td>19</td>
<td>I am about to say</td>
<td></td>
<td>t̆́, t̆́́, s̆́, s̆́́, t̆́̀</td>
<td>c̆́, c̆́́, t̆́, t̆́́</td>
</tr>
<tr>
<td>18</td>
<td>You should observe</td>
<td></td>
<td>ch̆́, ch̆́́, ch̆́̀, ch̆́́́</td>
<td>s̆́, c̆́, t̆́, t̆́́</td>
</tr>
<tr>
<td>17</td>
<td>Number 16 is</td>
<td></td>
<td>s̆́, s̆́́, s̆́́́, s̆́́́́</td>
<td>s̆́, s̆́́, s̆́́́, s̆́́́́</td>
</tr>
<tr>
<td>16</td>
<td>Write clearly</td>
<td></td>
<td>s̆́, s̆́́, s̆́́́, s̆́́́́</td>
<td>s̆́, s̆́́, s̆́́́, s̆́́́́</td>
</tr>
</tbody>
</table>

Syllable Test Record
Observer W.H.S.
Condition tested: 1000-7500 Hz
Test No. 10
Call No. 59-37

Date 3-16-28
Practice tests
Speech was low- and high-pass filtered with very sharp filters having a cutoff frequency defined by the abscissa. Two things were measured for each filtered speech sound, the RMS level and the articulation. The speech energy for the two filter outputs is shown by the dashed lines and the articulations are shown by the solid lines. The curve labeled “Articulation H” is the same as $s_H$, and the curve labeled “Articulation L” is the same as $s_L$. Note how the energy curves cross at the 50% point, as they should for two sharp filters. Note how the articulation curves do not cross at 50% but at 60%. Also, the frequency of the crossover is very different for energy and articulation. The equal energy point is at 450 Hz, while the equal articulation point is at 1550 Hz.
Articulation Results

- $S = \nu c^2$

- Error independence between bands
Articulation Index (Al)

\[ (1 - s(a, c)) = (1 - s(a, b))(1 - s(b, c)) \] \hspace{1cm} (1)

\[ \log_{10}(1 - s(a, c)) = \log_{10}(1 - s(a, b)) + \log_{10}(1 - s(b, c)) \] \hspace{1cm} (2)

\[ AI(s) = \frac{\log_{10}(1 - s)}{\log_{10}(1 - s_{max})} \] \hspace{1cm} (3)

\[ AI(s(a, c)) = AI(s(a, b)) + AI(s(b, c)) \hspace{1cm} f_a \leq f_b \leq f_c \] \hspace{1cm} (4)
Underlying Density

\[ AI(s(0, f_c)) = \int_0^{f_c} D(f) df \quad (5) \]

and

\[ D(f) = \frac{\partial}{\partial (f_c)} AI(s(0, f_c)) \quad (6) \]

Finally, for each of K bands,

\[ D_k = \int_{f_k}^{f_{k+1}} D(f) df \quad (7) \]

where limits chosen so all \( D_k \) are equal.
Multi-independent Channel Model

• Fletcher’s articulatory band: 1mm along the basilar membrane (20 between 300 and 8000Hz)

• A single zero error band means zero error overall!!

• Robustness to a range of problems
AI and Noise

- Saturating SNR at 0 and 30 dB,

\[ D_k = \frac{1}{K} \frac{SNR_k}{30} \]  \hspace{1cm} (8)

and

\[ AI = \sum_{k=1}^{K} D_k \]  \hspace{1cm} (9)
Fig. 6: Hypothetical cascade of recognition layers, starting with the cochlea. The articulation measures shown at the bottom are defined in Table II. The words along the top describe the physical correlate of the measure. No feedback is assumed between layers in this oversimplified model of HSR. The first layer, the cochlea, determines the signal-to-noise ratio in about 2800 overlapping critical band channels. The next layer extracts features (i.e., partial recognition) from the speech in a local manner, as indicated by the network wiring. The output of this layer is measured in terms of the K=20 or so feature errors ek. Next, the features are mapped onto the M=20 or so phones. This process necessarily integrates across the entire tonotopic axis. Then syllables and words are formed.
Figure 17.1: Block diagram of sound representation in the auditory system.
Issues to be considered:

- Robustness to frequency localized degradation
- Relating synchrony constraint between frequency bands
- Use of different recognition strategies in different bands
- Robustness to input amplitudes in each sub-band
- Degree of syllable, phone, state merging level: word, syllable, phone, state

Multi-band approach potentialities:

- Multi-band model
- Conventional ASR

Feature Extraction

Classifier

Final result

Feedback loop
Figure 1: Block diagram of recognizer using missing-feature adaptation.
Questions about Articulation Index

• Based on phones - the right unit for fluent speech?
• Lost correlation between distant bands?
• Lippmann experiments, disjoint bands
HSR vs ASR

Quantitative Comparisons

• Lippmann compilation

• Range of tasks
Figure 18.1: Six speech recognition corpora.
Figure 18.2: Characteristics of six talker-independent recognition corpora

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Description</th>
<th>Number of Talkers</th>
<th>Vocabulary Size</th>
<th>Number of Utterances</th>
<th>Total Duration</th>
<th>Recognition Perplexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>TI Digits</td>
<td>Read Digits</td>
<td>326</td>
<td>10</td>
<td>25,102</td>
<td>4 hrs</td>
<td>11</td>
</tr>
<tr>
<td>Alphabet Letters</td>
<td>Read Alphabet Letters</td>
<td>150</td>
<td>26</td>
<td>7,800</td>
<td>1 hr</td>
<td>26</td>
</tr>
<tr>
<td>Resource Management</td>
<td>Read Sentences</td>
<td>109</td>
<td>1,000</td>
<td>4,000</td>
<td>4 hrs.</td>
<td>60-1,000</td>
</tr>
<tr>
<td>Wall Street Journal</td>
<td>Read Sentences</td>
<td>84 - 284</td>
<td>5,000 - 20,000</td>
<td>7,200 - 37,200</td>
<td>12 hrs - 62 hrs</td>
<td>45-160</td>
</tr>
<tr>
<td>Credit-Card Continuous Speech Recognition</td>
<td>Spontaneous Telephone Conversations</td>
<td>70</td>
<td>2,000</td>
<td>35 Conversations, 1,600 Segments</td>
<td>2 hrs</td>
<td>100</td>
</tr>
<tr>
<td>Credit-Card Wordspotting</td>
<td>Spontaneous Telephone Conversations</td>
<td>70</td>
<td>20 Keywords</td>
<td>2,000 Keyword Occurrences</td>
<td>2 hrs</td>
<td>—</td>
</tr>
</tbody>
</table>
Figure 18.2: Five comparisons between human and ASR Devices.
<table>
<thead>
<tr>
<th>System</th>
<th>10dB SNR</th>
<th>16dB SNR</th>
<th>Quiet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline HMM ASR</td>
<td>77.4%</td>
<td>42.2%</td>
<td>7.2%</td>
</tr>
<tr>
<td>ASR with noise comp</td>
<td>12.8%</td>
<td>10.0%</td>
<td>-</td>
</tr>
<tr>
<td>Human Listner</td>
<td>1.1%</td>
<td>1.0%</td>
<td>0.9%</td>
</tr>
</tbody>
</table>

Table 18.1: Word error rate for 5000 word Wall Street journal task using additive automotive noise
**HSR vs ASR**

Qualitative Comparisons

- Signal processing
- Subword recognition
- Temporal integration
- Higher levels
HSR vs ASR: Signal Processing

• Many maps versus one
• Sampled in frequency and time
  vs sampled in time (10ms)
• Some aspects of hearing already in ASR
HSR vs ASR: Subword Recognition

• Knowing what is important

• Combining it optimally
HSR vs ASR: Temporal Integration

• Using or ignoring duration
• Compensating for rapid speech
• Incorporating multiple time scales
HSR vs ASR: Higher Levels

• Syntax
• Semantics
• Pragmatics
• Getting the gist
• Dialog to learn more
Conclusions

• Under good conditions, human recognition much better

• Humans need to pay attention

• Some human approaches going into ASR

• Much more to do