The role of the cochlea in Human speech recognition

Where is the speech information lurking?

Jont Allen
Sandeep Phatak
Marion Regnier
Feipeng Li

Univ. of IL, Beckman Inst., Urbana IL
Objective

Develop rigorous procedures for analyzing and modifying speech in noise, to:

- identify perceptual features, denoted events
- Develop a theory of human speech recognition (HSR) based on two basic measures:
  1. AI-Gram (speech audibility measure)
  2. Confusion matrix (speech discrimination measure)
- Show that across-frequency timing cues are events
Human listeners as a Shannon Channel

My approach is inspired by information theory using a classic 3-pronged approach: Simplify, simplify, simplify.

1. The Channel capacity theorem gives the maximum information rate as:

\[
C \equiv \int \log_2 \left(1 + snr^2(f)\right) df
\]  

(1)

2. The basic idea is to use a Maximum entropy (MaxEnt) speech source, and reduce the maximum information rate for by increasing the noise.

Take full advantage of Articulation Index predictions of the average phone score \( s = P_c(AI) \)
The research goal is to identify *elemental HSR events*. An event is defined as a *perceptual feature*. Event errors are measured by band errors $e_k$.

Model of human speech recognition (HSR)

Output: Cochlea Event Phones Syllables Words

$s(t)$ Filters Layer Layer Layer Layer

$A_{I_k} \propto snr_k$ [dB] $e_k = 0.82^{A_{I_k}}$ $s = 1 - e_1 e_2 \ldots e_{20}$ $S_{cv} = s^2$ $W$

Analog objects "Front–end"

??? Discrete objects "Back–end"

$W$

Articulation Matrices and **elemental events**

- **Miller-Nicely’s 1955 articulation matrix** $A$ measured at $[-18, -12, -6 \text{ shown, } 0, 6, 12]$ dB SNR

**Table III. Confusion matrix for $S/N = -6$ dB and frequency response of 200–6500 cps.**

<table>
<thead>
<tr>
<th></th>
<th>p</th>
<th>l</th>
<th>k</th>
<th>f</th>
<th>θ</th>
<th>s</th>
<th>ʃ</th>
<th>b</th>
<th>d</th>
<th>g</th>
<th>v</th>
<th>ữu</th>
<th>z</th>
<th>ʒ</th>
<th>m</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>80</td>
<td>43</td>
<td>64</td>
<td>17</td>
<td>14</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>l</td>
<td>71</td>
<td>84</td>
<td>55</td>
<td>5</td>
<td>9</td>
<td>3</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>k</td>
<td>66</td>
<td>76</td>
<td>107</td>
<td>12</td>
<td>8</td>
<td>9</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>f</td>
<td>18</td>
<td>12</td>
<td>9</td>
<td>175</td>
<td>48</td>
<td>11</td>
<td>1</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>θ</td>
<td>19</td>
<td>17</td>
<td>16</td>
<td>104</td>
<td>64</td>
<td>32</td>
<td>7</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>s</td>
<td>8</td>
<td>5</td>
<td>4</td>
<td>23</td>
<td>39</td>
<td>107</td>
<td>45</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>ʃ</td>
<td>1</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>29</td>
<td>195</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

**Confusion groups imply underlying elemental events**
Case of /pa/, /ta/, /ka/ with /ta/ spoken

- Phone groups imply sub-phonemic units (i.e., events)

- How many events, and of what form?
- Plot of $A_{i,j}(snr)$ for row $i=2$
- Solid red curve is total error $e_2 \equiv 1 - A_{2,2} = \sum_{j\neq2} A_{2,j}$
The case of /ma/ vs. /na/

This 2-group of sounds is closed since
\[ \mathcal{A}_{/ma/\rightarrow/ma/}(SNR) + \mathcal{A}_{/ma/\rightarrow/na/}(SNR) \approx 1 \]

- There can be only 1 event
- Solid red curve is the total error:
  \[ e_i \equiv 1 - \mathcal{A}_{i,i} = \sum_{j \neq i} \mathcal{A}_{i,j}(SNR) \]
Fletcher’s Lopass/Hipass result

The AI is based on the band error product formula

\[ 1 - P_c(SNR) \equiv e_{total}(SNR, f_c) = e_{lp}(SNR, f_c)e_{hp}(SNR, f_c) \] (2)
Probabilistic measures of recognition

- \( k^{th} \) band articulation index: \( AI_k = \frac{10}{30} \log(1 + c^2 snr_k^2) \)
- \( c = 2, k = 1 \cdots K \) with \( K = 20 \)
- Band (event) error: \( e_k = e_{min} \)
- MaxEnt phone score: \( s = 1 - e_1 e_2 \cdots e_K = 1 - e_{min} \)
- MaxEnt syllable model score: \( S_{cv} = s^2, S_{cvc} = s^3 \)
How can we find events?

- A 4-Step analysis relates confusions to an audibility measure (?) :

- Modification of speech sounds
  - We developed a tool based on the Short-Time Fourier Transform (STFT) (?) that allows us to selectively:
    - Mask with noise specific time and frequency regions so that this specific part of the speech becomes inaudible
    - Selectively amplify specific regions to increase intelligibility
  - We will present audio examples of original and modified sounds
//t/ confusion threshold at $P_c(SNR^* = -2) = 0.9$
correlated to Event-gram
m112/tɛ/ in speech-weighted noise

Step 1: AI-gram of m112te at 0 dB SNR

Step 2: Integrated AI for m112te at 0 dB SNR

Step 3: Event-gram of m112te at $t^* = 26.25$ cs

Step 4: Confusion patterns for m112te

/ʃ/ confusion threshold at $P_c(SNR^* = -16) = 0.9$

correlated to Event-gram
Correlations of /t/ events

- High correlation across all /t/’s in the database

Event−gram in WN at t* = 15 cs, BW=450, T=0.125

Confusion patterns for f106ta in WN

Correlation between perceptual and physical domains

- SWN, BW=570 Hz, T=0.335
- WN, BW=450 Hz, T=0.125
Masking of /tɑ/ timing cue

When the /t/ burst is masked by noise, the perception morphs to /p/

DEMO 4
Truncation of /tɑ/

- This represents the normal hearing responses to a truncated /tɑ/, from the start of the consonant
- Morphing from /tɑ/ to /pɑ/ to /bɑ/ at 0 and 12 dB SNR
- Similar to previous studies?, and our more extensive results
Truncation of f101 /sa/

This represents the normal hearing responses to a truncated /sa/, from the start of the consonant.

Morphing from /sa/ to /za/ to /da/ to /ða/

Duration seems to be a fricatives event.
/mA/- /nA/ discrimination

- /nA/ recognition from /mA/ relies on a \( \approx 50 \) ms delay formed from the \( F_1 \) and \( F_2 \) collision.

- When we edit the speech so that the onset is simultaneous above 0.6 kHz, the /nA/ is robustly and naturally heard as /mA/.

- METHODS: 9 listeners evaluated these sounds in open response random trial experiment.
Deletion of /na/ timing cue

(e) Original /na/

(f) Modified /na/

Consonant recognition of original f105na

Consonant recognition of modified f105na
Creation of /na/ timing cue

(g) Original /ma/

(h) Modified /ma/

Consonant recognition of original f105ma

Consonant recognition of modified f105ma
Enhancement of /tɛ/ event

The sound is heard as /t/ again, we suppressed the morph (see confusion patterns of slide 4)

METHODS: The /t/ burst is enhanced (14 dB) on the quiet sound, then noise is added

DEMO
Enhancement of /ta/ event

- The sound is heard as /t/ again, we increase /t/ recognition
- METHODS: The /t/ burst is enhanced (14 dB) on the quiet sound, then noise is added
- DEMO
Conclusion

- We have shown that normal listeners use *across-frequency timing coincidences* to discriminate consonants in noise.
- We have developed a tool to modify speech sounds.
  - Morph sounds. Ex: /ma/ - /na/
  - Decrease or increase intelligibility. Ex: /ta/, /tɛ/
- This could well lead to the design of new hearing aids.