FROM LORD RAYLEIGH TO SHANNON: HOW DO WE DECODE SPEECH?

Jont B. Allen
ECE Univ. of IL
Beckman Inst.
Urbana IL
jba@auditorymodels.org

http://auditorymodels.org/CUNY/
http://auditorymodels.org/jba/PAPERS/ICASSP/
WHAT I WANT TO SHOW:

- Biological systems are the ultimate information processors
- **HSR** is a bottom–up, divide and conquer strategy
  - We recognize speech based on a hierarchy of context layers
  - As in vision, entropy decreases as we integrate context
- Humans have an intrinsic robustness to noise and filtering
  - Robustness is not due to semantic context effects
HOW WE RECOGNIZE SPEECH?

- Hierarchical “bottom up” analysis
- Accurate statistical models of performance at each stage

- Entropy drops (i.e., context is integrated) in stages
## DEFINITIONS

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>phone</strong></td>
<td>A consonant (C) or vowel (V) sound</td>
</tr>
<tr>
<td><strong>word</strong></td>
<td>A <em>meaningful</em> phone or phone sequence (i.e., cat ≡ CVC)</td>
</tr>
<tr>
<td><strong>phoneme</strong></td>
<td>The replaceable set of phones which leave a word meaning invariant</td>
</tr>
<tr>
<td><strong>recognition</strong></td>
<td>Probability measure $P_c$ of <em>correct</em> phoneme identification</td>
</tr>
<tr>
<td><strong>articulation</strong></td>
<td>Recognition of “nonsense words”</td>
</tr>
<tr>
<td><strong>intelligibility</strong></td>
<td>Recognition of words (i.e., <em>meaningful</em> speech)</td>
</tr>
<tr>
<td><strong>robustness</strong></td>
<td>Relative recognition with filtering and noise</td>
</tr>
<tr>
<td><strong>confusion matrix</strong></td>
<td>Table of identification frequencies $N_{sr} \equiv N_{r</td>
</tr>
<tr>
<td><strong>articulation matrix</strong></td>
<td>A <em>confusion matrix</em> composed of nonsense sounds</td>
</tr>
<tr>
<td><strong>articulation event</strong></td>
<td>A <em>discrete subunit</em> of articulation [e.g., Voicing: /ba/ vs. /pa/]</td>
</tr>
<tr>
<td><strong>trial</strong></td>
<td>A single presentation of a set of events</td>
</tr>
<tr>
<td><strong>state</strong></td>
<td>A values of a set of events at some instant of time</td>
</tr>
<tr>
<td><strong>state machine</strong></td>
<td>A machine (program) that transforms from one state to another</td>
</tr>
<tr>
<td><strong>noiseless state machine</strong></td>
<td>A <em>deterministic</em> state machine</td>
</tr>
<tr>
<td><strong>context</strong></td>
<td>Coordinated combinations of events within a trial</td>
</tr>
<tr>
<td><strong>message</strong></td>
<td>Specific information transmitted by a trial</td>
</tr>
<tr>
<td>$p_n$</td>
<td>Probability of event $n$, of $N$ possible events $I_n = \log_2(1/p_n)$, $n = 1, \ldots, N$</td>
</tr>
<tr>
<td><strong>information density</strong></td>
<td>$H = \sum_{n=1}^{N} p_n I_n$</td>
</tr>
<tr>
<td><strong>entropy</strong></td>
<td>Average information: $H = \sum_{n=1}^{N} p_n I_n$</td>
</tr>
</tbody>
</table>
KEY HSR STUDIES

- The first articulation experiments date from Lord Rayleigh’s 1908 and George Campbell 1910 phoneme identification experiments
- A basic probabilistic approach was developed by Stewart & Fletcher 1921
  – Detailed review of Fletcher’s AI theory: Allen IEEE 1994
- French and Steinberg 1947 WWII studies
- Shannon’s Information theory 1948+
- G.A. Miller, Heise and Lichten 1951; G.A. Miller & Nicely 1955
  – *Language and communication* G.A. Miller, 1951 McGraw Hill
    Miller first introduces IT to language modeling, following Shannon
- Boothroyd JASA 1968; Boothroyd & Nittrouer JASA 1988
- Bronkhorst et al. JASA 1993, 2002
- Van Petten *et al.* 1994
- Detailed review chapter Allen 2003
## MOTIVATION

- Results of Lippmann 1997, sorted by Error Ratio

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Size in Words</th>
<th>Conditions</th>
<th>% Error</th>
<th>Error Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alphabetic</td>
<td>26</td>
<td>20-talkers 8-listeners</td>
<td>5.0</td>
<td>isolated</td>
</tr>
<tr>
<td>Resource</td>
<td>1000</td>
<td>null grammar</td>
<td>17</td>
<td>2</td>
</tr>
<tr>
<td>WSJ-NAB</td>
<td>5000</td>
<td>quiet (trained)</td>
<td>7.2</td>
<td>0.9</td>
</tr>
<tr>
<td>Switchboard</td>
<td>14,000</td>
<td>spontaneous (tel. BW)</td>
<td>43</td>
<td>4</td>
</tr>
<tr>
<td>WSJ-NAB</td>
<td>5000</td>
<td>10 dB (trained)</td>
<td>12.8</td>
<td>1.1</td>
</tr>
<tr>
<td>WSJ-NAB</td>
<td>65,000</td>
<td>close mic</td>
<td>6.6</td>
<td>0.4</td>
</tr>
<tr>
<td>WSJ-NAB</td>
<td>65,000</td>
<td>omni mic</td>
<td>23.9</td>
<td>0.8</td>
</tr>
<tr>
<td>Resource</td>
<td>1000</td>
<td>word–pair grammar</td>
<td>3.6</td>
<td>0.1</td>
</tr>
<tr>
<td>WSJ-NAB</td>
<td>5000</td>
<td>quiet (not trained)</td>
<td>42</td>
<td>0.9</td>
</tr>
<tr>
<td>WSJ-NAB</td>
<td>5000</td>
<td>22 dB (not trained)</td>
<td>77.4</td>
<td>0.9</td>
</tr>
<tr>
<td>word spotting</td>
<td>20</td>
<td>judgment errors</td>
<td>24</td>
<td>0.3</td>
</tr>
<tr>
<td>TI-digit</td>
<td>10</td>
<td>connected</td>
<td>0.72</td>
<td>0.009</td>
</tr>
</tbody>
</table>

DEMO ScanMail examples /Audio/ScanMailExample
**TYPICAL ARTICULATION TEST RECORD**

- Fletcher’s method of nonsense phone error analysis

---

**Articulation Test Record**

**March 1928**

<table>
<thead>
<tr>
<th>DATE</th>
<th>SYLLABLE ARTICULATION</th>
<th>FILTERED</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-16-29</td>
<td>51.5%</td>
<td>Low Pass</td>
</tr>
</tbody>
</table>

**Title of Test:** Practice Tests

**Condition Tested:** 1500 Hz lowpass filtering

---

<table>
<thead>
<tr>
<th>NO</th>
<th>OBSERVED</th>
<th>CALLED</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>mān</td>
<td>máv</td>
</tr>
<tr>
<td></td>
<td>pāz</td>
<td>pōth</td>
</tr>
<tr>
<td></td>
<td>Kāb ✓</td>
<td>Kāb ✓</td>
</tr>
<tr>
<td>2</td>
<td>pōch ✓</td>
<td>pōch ✓</td>
</tr>
<tr>
<td></td>
<td>nez ✓</td>
<td>nezh</td>
</tr>
<tr>
<td></td>
<td>shēth ✓</td>
<td>sīz ✓</td>
</tr>
<tr>
<td>3</td>
<td>seng ✓</td>
<td>seng ✓</td>
</tr>
<tr>
<td></td>
<td>jōch ✓</td>
<td>jōch ✓</td>
</tr>
<tr>
<td></td>
<td>fūch ✓</td>
<td>fūch ✓</td>
</tr>
<tr>
<td>4</td>
<td>chūd ✓</td>
<td>chūd ✓</td>
</tr>
<tr>
<td></td>
<td>thām ✓</td>
<td>thām ✓</td>
</tr>
<tr>
<td></td>
<td>thāl ✓</td>
<td>thāl ✓</td>
</tr>
<tr>
<td>5</td>
<td>run ✓</td>
<td>run</td>
</tr>
<tr>
<td></td>
<td>hab ✓</td>
<td>hab</td>
</tr>
<tr>
<td></td>
<td>poth ✓</td>
<td>poth</td>
</tr>
</tbody>
</table>

**Data**

\[
S = P_c(syllable) = 0.515 \\
v = P_c(vowels) = 0.909 \\
c = P_c(consonants) = 0.74
\]

**Models**

\[
\hat{S} = cvc = 0.498 \quad \text{(CVC syllable model)} \\
s = P_c(phone) = \left(\frac{v + 2c}{3}\right) = 0.796 \\
s^3 = 0.505 \quad \text{(3 phone syllable model)}
\]
THE METHOD

- The data bases they used were formed from
  - statistically balanced
  - nonsense
  - CVC, CV and VC syllable lists
    where C represents a consonant and V a vowel
- The syllable lists were spoken, and the listeners recorded what they heard
- Probabilities-correct $c$ and $v$ for the sound-units were computed
- The average $\{C,V\}$ speech-unit articulation probability $s$ was computed from the composition of $\{C,V\}$ units in the data base
  (i.e. $s = (2c + v)/3$ for CVC’s, $s = (c + v)/2$ for CV’s)
  - Measure $s$ looks like a sufficient statistic
WHAT THEY FOUND

- Nonsense phones are recognized as independent units:
  - The probability of correct recognition for the average phoneme $s$ accurately predicts the nonsense syllable score $S_{cvc}$, where
    \[ S_{cvc} = c^2 v = s^3 \]
    *This is a necessary but insufficient condition for independence*

- These statistical models are highly accurate

- !!! Remember: This only applies to “nonsense words” !!!

QUESTIONS?
THE NEXT STEP

• Next they dissected \( s \equiv P_{correct}(\text{phone}) \) into frequency bands!

SPECIFIC DEFINITIONS

<table>
<thead>
<tr>
<th>SYMBOL</th>
<th>DEFINITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha )</td>
<td>gain applied to the speech</td>
</tr>
<tr>
<td>( c(\alpha) \equiv P_c(\text{consonant}</td>
<td>\alpha) )</td>
</tr>
<tr>
<td>( v(\alpha) \equiv P_c(\text{vowel}</td>
<td>\alpha) )</td>
</tr>
<tr>
<td>( s(\alpha) = \frac{2c(\alpha) + v(\alpha)}{3} )</td>
<td>average phone articulation for CVC’s</td>
</tr>
<tr>
<td>( e(\alpha) = 1 - s(\alpha) )</td>
<td>phone articulation error</td>
</tr>
<tr>
<td>( f_c )</td>
<td>high- and low-pass cut-off frequency</td>
</tr>
<tr>
<td>( s_L(\alpha, f_c) )</td>
<td>( s ) for low-pass filtered speech</td>
</tr>
<tr>
<td>( s_H(\alpha, f_c) )</td>
<td>( s ) for high-pass filtered speech</td>
</tr>
</tbody>
</table>
FLETCHER’S TWO BAND FORMULATION

• Split the speech into low and high bands, having articulations
  \[ s_L(\alpha, f_c) \text{ and } s_H(\alpha, f_c) \]

• Fletcher proposed a linearizing transformation of the phone articulations
  \[ \mathcal{A}(s_L) + \mathcal{A}(s_H) = \mathcal{A}(s) \]
  – This is a nonlinear transformation of probabilities
  – There was no guarantee that such a transformation exists
    However, Fletcher’s intuition was correct
WHAT THEY FOUND

- For nonsense \{C,V\} syllables the phone articulation transformation is:

\[
\mathcal{A}(s) = \frac{\log(1 - s)}{\log(e_{\text{min}})},
\]

with \(e_{\text{min}} = 0.015\) (1.5% error, or 98.5% correct)

- This relationship must have taken years to discover!

- Solving for \(e \equiv 1 - s(\mathcal{A})\):

\[
e = e_{\text{min}} = e_{\text{min}} + \mathcal{A}(s_L) + \mathcal{A}(s_H) = e_{\text{min}} \cdot \mathcal{A}(s_L) \cdot \mathcal{A}(s_H)
\]

- In terms of the error probabilities \(e = 1 - s\), \(e_L = 1 - s_H\) and \(e_L = 1 - s_L\):

\[
e = e_L e_H.
\]
FLETCHER’S TWO BAND EXAMPLE

• If we have 100 spoken sounds, and 10 errors are made while listening to the low band, and 20 errors are made while listening to the high band, then

\[ e = 0.1 \times 0.2 = 0.02, \]

namely 2 errors will be made when listening to the full band, so

\[ s = 1 - 0.02 = 0.98 \]

\[ S = s^3 = 0.941 \]

• This is an unexpected, simple, and amazing result

  — What does this mean? Why does it turn out this way?

  DEMO of the the McGurk effect
THE FLETCHER-STEWARD MULTI-CHANNEL MODEL

- Fletcher 1921 generalize the two-band case to $K = 20$ frequency bands

$$1 - s = e_1 e_2 \cdots e_k \cdots e_K \times e_{\text{visual}}$$

$$= (1 - s_1)(1 - s_2) \cdots (1 - s_K) \times (1 - s_{\text{visual}})$$

where

$$e_i \equiv 1 - s_i$$

- This formula forms the basis of articulation index theory
- Why $K = 20$ bands?
  Each band equals 1mm along the basilar membrane

- I have added a visual channel, to account for the McGurk effect (Channel 21)

- Probability of error $e_i$ models events, as in the visual example
DENSITY OF ARTICULATION PER CRITICAL BAND

This plot is the ratio of $D(f)/\kappa(f)$, where $D(f)$ is the articulation density

$$D(f_c) \equiv \frac{\partial A_L}{\partial f_c}, \ K \text{ AI bands}$$

$\kappa(f) = \text{the critical ratio} \ [\propto \text{cochlear filter bandwidth (ERB)}]$
MODEL OF BAND EVENT ERRORS

- When the SNR is varied they found that the event-error is

\[ e_k = e_{min}^{\frac{SNR_k}{K}} \]

where \( SNR_k \) is the signal to noise ratio in dB, divided by 30, such that

\[ SNR_k \equiv \begin{cases} 
0 & \text{if } 20 \log_{10}(snr) < 0 \\
20 \log_{10}(snr)/30 & \text{if } 0 < 20 \log_{10}(snr) < 30 \\
1 & \text{if } 30 < 20 \log_{10}(snr). 
\end{cases} \]

Thus

\[ 0 \leq SNR_k \leq 1. \]

- Total error:

\[ e = e_1 e_2 \cdots e_K = e_{min}^{(SNR_1 + SNR_2 \cdots SNR_K)/K} \]

- The speech SNR in dB (not the energy) determines the event errors \( e_k \), and thus the phoneme articulation

\[ s = 1 - e_1 e_2 \cdots e_K \]
**AI AS A CHANNEL CAPACITY**

- Since $\sum_k (\log snr_k) = \log(\prod_k snr_k)$

  \[
  A \equiv \frac{1}{K} \sum_k SNR_k \propto \log \left( \prod_k snr_k \right)^{1/K} \tag{1}
  \]

- and from Shannon (for the Gaussian channel)

  \[
  C = \int_{-\infty}^{\infty} \log_2 [1 + snr^2(f)] df, \tag{2}
  \]

![Graph showing the relationship between C(snr) and A(snr) vs. 20 log(snr) [dB].]
TALKER PRODUCTION ERRORS

- What determines $s_{\text{max}} = 1 - e_{\text{min}}$?
- Utterance talker mispronunciations, as defined by 32 listeners
- Errors are distributed like Zipf’s Law $N/N_T \approx 0.6e^{-4.48P_e}$

- 35% of the utterances have no error
- 33% have $> 10\%$ error, 10% $> 35\%$ error, 5% $> 50\%$ error
SOURCES OF ERROR

- **Talker production errors**
  - Production errors are defined by *token utterances error* over listeners
  - Once poor utterances are identified, they may be selectively removed
    - *This method allows us to control the gross error rate*
  - With this method we can obtain a 100% score in the clear
    - *The price for this is a reduced $N_{utterances}$*
- **Listener errors** (after selectively removing production errors)
  - Listener bias may be determined from individual confusion matrices
  - This bias can be a function of the production error threshold
    - *The main effect is on $L_2$ listeners*
EXAMPLE CALCULATIONS

Wide-band channel vs. SNR

Phone articulation/mm

\[ s_k(SNR_k), e_k(SNR_k) \]

\[ SNR_k \]

\[ s_k, e_k \]

\[ 0 \leq SNR_k \leq 30 \]

\[ 0 \leq s_k, e_k \leq 1 \]

Phone and CVC model

\[ s = 1 - 0.015^A_{in} \]

\[ S = s^3 \]

\[ A_{in} = \text{avg}(SNR)/30 \]

Word Model

\[ W(A) = 1 - [1 - S(A)]^{3.51} \]

\[ S(s), W(s) \]

Effect of Context

\[ S(s), W(s) \]
THE RECOGNITION CHAIN

- The cochlear critical bandwidth defines the $SNR_k$
- The event-error model: $e_k \propto \frac{SNR_k}{e_{min}}$ (SNR in dB units)
- The average-phone articulation model:
  $$s = 1 - e_1 e_2 \cdots e_k \cdots e_K$$
- The nonsense CVC syllable articulation model: $S = s^3$
- Heuristic degree of freedom context models Boothroyd (see discussion Allen 1994)
  - Word: $W = 1 - (1 - S)^j$
  - Sentence: $I = 1 - (1 - W)^k$
  - Sentence with context: $C = 1 - (1 - I)^l$
- Layers of context:
  - $j$ depends on the ratio of words to pseudo-words in the corpus,
  - $k$ depends on the number of salient words in a sentence,
  - $l$ depends on the word salience and topic context.
COMPOSITION LAWS

- Rules regarding $\prod_i P_{\text{error}}^{(i)}$ versus product $\prod_i P_{\text{correct}}^{(i)}$?

  - Parallel processing: $P_e = \prod_k e_k$
    * Errors in many bands have no effect
    * One band with small error (i.e., $e_k = 0$) dominates
      e.g., $e = e_L e_H$, $e = e_1 e_2 \cdots e_K$; the McGurk example

  - Serial processing: $P_c = \prod_k s_k$
    * All items of a string must be correct for success
      e.g., $S_{cvc} = cvc \approx s^3$; $S_{cv} \approx s^2$

- HSR seems to be a problem in combinatorics, of elementary pre-phonic events.
HOW WE RECOGNIZE SPEECH?

- Hierarchical “bottom up” analysis
- Accurate statistical models of performance at each stage

- Entropy drops (i.e., context is integrated) in stages
SUMMARY OF MODEL RESULTS

- Hierarchical probability relations:
  
  band $SNR \rightarrow$
  
  band errors (events) $\rightarrow$
  
  phoneme errors $\rightarrow$
  
  syllable errors $\rightarrow$
  
  nonsense word errors $\rightarrow$
  
  true word errors, etc.

- The HSR error is established well before language is accessed!
  
  HSR error depends only on the $SNR$ in bands
SPEECH ENTROPY VS. THE WIDEBAND SNR

- $P_c(\mathcal{H}, SNR)$ Miller, Heise and Lichten 1951
- Many of the results of MHL51 expand on the AI model

CVC data [Table I: Miller, Heise and Lichten (1951)]

SNR = −21 $\Rightarrow$ Chance (i.e., $2^{-H}$)
**GRAMMATICAL CONTEXT**

- Five groups of five words that form grammatical sentences:

<table>
<thead>
<tr>
<th>Don</th>
<th>Brought</th>
<th>His</th>
<th>Black</th>
<th>Bread</th>
</tr>
</thead>
<tbody>
<tr>
<td>He</td>
<td>Has</td>
<td>More</td>
<td>Cheap</td>
<td>Sheep</td>
</tr>
<tr>
<td>Red</td>
<td>Left</td>
<td>No</td>
<td>Good</td>
<td>Shoes</td>
</tr>
<tr>
<td>Slim</td>
<td>Loves</td>
<td>Some</td>
<td>Wet</td>
<td>Socks</td>
</tr>
<tr>
<td>Who</td>
<td>Took</td>
<td>The</td>
<td>Wrong</td>
<td>Things</td>
</tr>
</tbody>
</table>

- Tests:
  - 5 word lists
  - 25 word
  - 25 words with grammatical context
    - Example: *He left no black socks*
  - 25 words reverse order
    - Example: *Socks black no left he.*
GRAMMATICAL CONTEXT

- Results of tests
CONFUSION MATRIX PARTITIONING

- **Miller & Nicely 1955** Confusion Matrix (Table III)

  - **MN55** established a natural phone hierarchical clustering:

  "This breakdown of the confusion matrix into five smaller matrices . . . is equivalent to . . . five communication channels . . ." – Miller & Nicely 1955

<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>ʒ</th>
<th>z</th>
<th>s</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>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</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>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</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>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></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>ʒ</th>
<th>z</th>
<th>s</th>
<th>m</th>
<th>n</th>
</tr>
</thead>
<tbody>
<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>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>
</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>
</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>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>b</th>
<th>d</th>
<th>g</th>
<th>v</th>
<th>ʒ</th>
<th>z</th>
<th>s</th>
<th>m</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>b</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>8</td>
<td>136</td>
<td>10</td>
<td>9</td>
<td>47</td>
<td>16</td>
</tr>
<tr>
<td>d</td>
<td>2</td>
<td>3</td>
<td>63</td>
<td>66</td>
<td>3</td>
<td>19</td>
<td>37</td>
<td>56</td>
<td>3</td>
</tr>
<tr>
<td>g</td>
<td>2</td>
<td>2</td>
<td>48</td>
<td>5</td>
<td>5</td>
<td>145</td>
<td>45</td>
<td>12</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>v</th>
<th>ʒ</th>
<th>z</th>
<th>m</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>v</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ʒ</td>
<td>7</td>
<td>3</td>
<td>8</td>
<td>45</td>
<td>129</td>
</tr>
<tr>
<td>z</td>
<td>1</td>
<td>26</td>
<td>18</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>m</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>177</td>
<td>46</td>
</tr>
<tr>
<td>n</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>7</td>
<td>1</td>
</tr>
</tbody>
</table>

**Table III.** Confusion matrix for S/N = -6 db and frequency response of 200-6500 cps.
MILLER’S BINARY FEATURES

- Miller & Nicely derived binary consonant features [i.e., events]

Table XIX. Classification of consonants used to analyze confusions.

<table>
<thead>
<tr>
<th>Consonant</th>
<th>Voicing</th>
<th>Nasality</th>
<th>Affrication</th>
<th>Duration</th>
<th>Place</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>t</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>k</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>f</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>θ</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>s</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ſ</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>b</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>g</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>v</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ʰ</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>z</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ż</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>m</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>n</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

"... the impressive thing to us was that ... the [binary] features were perceived almost independently of one another."  — Miller & Nicely 1955
FINDING THE AI FOR MILLER NICELY TALKERS

- Average spectrum for female talkers

Wideband RMS: g:Speech = 12; r:Noise RMS = 0 [dB re 1V]
TRACE OF MILLER NICELY AND THE AI

• Next we look at the average PI function vs. AI
SYMMETRIC COMPONENT OF $P_C(SNR)$

- We stand to learn from linear operations on $P_{ij}(snr)$

Symmetric: $S_c(snr) \equiv [P_{ij}(snr) + P_{ji}(snr)]/2$
SKEW-SYMMETRIC COMPONENT OF $P_{C}(SNR)$

- Skew: $A_c(snr) \equiv [P_{ij}(snr) - P_{ji}(snr)]/2$
**SVD REPRESENTATION OF THE PERCEPTUAL SPACE**

- $4^{dim}$ SVD perceptual representation of the confusion matrix

---

**DEMO**
TEMPORAL RESOLUTION OF PHONE RECOGNITION

- Phones are recognized in on a 10 ms time scale (Furui 1986)

![Graph showing identification score vs truncation point for syllable, vowel, and consonant categories.](image-url)
WORD SEMANTICS: IP DEFINITION

- 704 isolated words were truncated in 50 ms steps Van Petten 1999

- Isolation point is defined as the time of the discontinuity in recognition

Expt. I – Neutral sentences: “The next word is test-word.”

![Graph showing accuracy of identification versus gate time]

- Categorical perception
WORD SEMANTICS: IP VS. DURATION

- Isolation point vs. word durations (real words, no sentence context)

HISTOGRAMS OF WORD IP’s and DURATIONS

GATE TIME (ms)

NUMBER OF WORDS

ISOLATION POINT (IP)

DURATION
ERP MEASURE OF CONTEXT RE IP

- Expt. II – Event related scalp potential (N-400 ERP) re IP, from Exp. I
- Sentence semantics effects

Pay with ...

- **Cohort congruous** dollars
- **Cohort incongruous** dolphins

- Words are recognized on a syllable by syllable basis, **within 50 ms**
- Context is recognized on a syllable by syllable basis, **within 200 ms**
FROM CONTINUOUS TO DISCRETE

Φ

OBSERVER

Ψ

CONTINUOUS

DISCRETE

- **Φ-domain signals**
  - Speech signal
  - Cochlear filter outputs
  - Neural rate
  - Voltage in cochlear nucleus cells

- **Ψ-domain objects**
  - Words
  - Syllables
  - Phonemes
  - Events [Miller’s features]

CATEGORIAL PERCEPTION

- Meaningful words are recognized before they end
- Syllables are recognized within 50 ms
SUMMARY

- Miller & Nicely found 5 independent channels, described by discrete events [Miller's features]
- Speech is recognized in layers:
  \[ SNR_k \Rightarrow \text{events} \Rightarrow \text{phones} \Rightarrow \text{syllables} \Rightarrow \text{words} \Rightarrow \ldots \]
- Language model performance is independent of noise robustness!
- To study HSR, entropy must be controlled
- Speech psychophysics is an important tool for studying HSR
FUTURE GOALS

- Use psychophysics to gain insight into event extraction
- The next break through:
  - More robust ASR
  - An event extracting hearing aid

This talk may be found at:
http://auditorymodels.org/jba/PAPERS/ICASSP/
References


