The identification and modification of consonant perceptual cues in natural speech
Part I

Jont Allen
Andrea Trevino
UIUC & Beckman Inst, Urbana IL

August 23, 2013
1. **Intro + Objectives**
   - Research objectives
     - 5 mins Σ8
1. Intro + Objectives 3 mins Σ3
   - Research objectives 5 mins Σ8
2. Historical overview 20 mins Σ28
   - AG Bell 1860, Rayleigh 1910, Fletcher 2021, Shannon 1948
1. **Intro + Objectives**
   - Research objectives
   - 3 mins Σ3

2. **Historical overview**
   - AG Bell 1860, Rayleigh 1910, Fletcher 2021, Shannon 1948
   - 20 mins Σ28

3. **Phone Recognition Models**
   - Channel capacity and the Articulation Index
   - Speech Psychophysics; Algram/3DDS (cues); Primes and Morphs;
   - Classification models (e.g., DFs)
   - 21 mins Σ49
1. **Intro + Objectives**
   - Research objectives

2. **Historical overview**
   - AG Bell 1860, Rayleigh 1910, Fletcher 2021, Shannon 1948

3. **Phone Recognition Models**
   - Channel capacity and the Articulation Index
   - Speech Psychophysics; Algram/3DDS (cues); Primes and Morphs;
   - Classification models (e.g., DFs)

4. **Cochlear Mechanics**
   - CBands, NL, Masking, Role re Speech perception; HI ears
1. **Intro + Objectives**  
   - Research objectives  
   3 mins $\Sigma 3$

2. **Historical overview**  
   - AG Bell 1860, Rayleigh 1910, Fletcher 2021, Shannon 1948  
   20 mins $\Sigma 28$

3. **Phone Recognition Models**  
   - Channel capacity and the Articulation Index  
   - Speech Psychophysics; Algram/3DDS (cues); Primes and Morphs;  
   - Classification models (e.g., DFs)  
   21 mins $\Sigma 49$

4. **Cochlear Mechanics**  
   - CBands, NL, Masking, Role re Speech perception; HI ears  
   16 mins $\Sigma 65$

5. **Summary + Conclusions + Questions**  
   3+3+4 mins $\Sigma 75$
1. Objectives of the UIUC HSR Group

1. Repeat classic experiments on human speech CV sounds 2005
1. Objectives of the UIUC HSR Group

1. Repeat classic experiments on human speech CV sounds 2005
2. Identify acoustic cues in CV tokens 2007
1. Objectives of the UIUC HSR Group

1. Repeat classic experiments on human speech CV sounds 2005
2. Identify acoustic cues in CV tokens 2007

- Findings: a) Onset burst, b) Frequency edge, c) Duration, d) $F_0$ modulation, e) Voicing 2007-11
1. Objectives of the UIUC HSR Group

1. Repeat classic experiments on human speech CV sounds 2005
2. Identify acoustic cues in CV tokens 2007
   - Findings: a) Onset burst, b) Frequency edge, c) Duration, d) $F^0$ modulation, e) Voicing 2007-11
   - Consonant recognition is binary (Threshold @ $SNR_{90}$) 2012
1. Objectives of the UIUC HSR Group

1. Repeat classic experiments on human speech CV sounds 2005
2. Identify acoustic cues in CV tokens 2007
   - Findings: a) Onset burst, b) Frequency edge, c) Duration, d) $F_0$ modulation, e) Voicing 2007-11
   - Consonant recognition is binary (Threshold @ $SNR_{90}$) 2012
   - Full analysis of the Articulation Index (AI) 2012
1. Objectives of the UIUC HSR Group

1. Repeat classic experiments on human speech CV sounds 2005
2. Identify acoustic cues in CV tokens 2007
   - Findings: a) Onset burst, b) Frequency edge, c) Duration, d) $F0$ modulation, e) Voicing 2007-11
   - Consonant recognition is binary (Threshold @ $SNR_{90}$) 2012
   - Full analysis of the Articulation Index (AI) 2012
3. Measure CV confusions in $\approx50$ hearing impaired ears 2009
1. Objectives of the UIUC HSR Group

1. Repeat classic experiments on human speech CV sounds 2005
2. Identify **acoustic cues** in CV tokens 2007
   - *Findings:* a) Onset burst, b) Frequency edge, c) Duration, d) $F_0$ modulation, e) Voicing 2007-11
   - Consonant recognition is binary (Threshold @ $\text{SNR}_{90}$) 2012
   - Full analysis of the **Articulation Index (AI)** 2012
3. Measure CV confusions in $\approx 50$ *hearing impaired* ears 2009
   - Characterize hearing impaired (HI) CV confusions 2010
1. Objectives of the UIUC HSR Group

1. Repeat classic experiments on human speech CV sounds 2005
2. Identify acoustic cues in CV tokens 2007
   - Findings: a) Onset burst, b) Frequency edge, c) Duration, d) $F0$ modulation, e) Voicing 2007-11
   - Consonant recognition is binary (Threshold @ $SNR_{90}$) 2012
   - Full analysis of the Articulation Index (AI) 2012
3. Measure CV confusions in $\approx$50 hearing impaired ears 2009
   - Characterize hearing impaired (HI) CV confusions 2010
   - Explain HI re NH feature extraction deficiencies, based on individual-differences in CV confusions 2012-13
1. Objectives of the UIUC HSR Group

1. Repeat classic experiments on human speech CV sounds 2005
2. Identify \textit{acoustic cues} in CV tokens 2007
   - \textit{Findings:} a) Onset burst, b) Frequency edge, c) Duration, d) $F_0$ modulation, e) Voicing 2007-11
   - Consonant recognition is binary (Threshold @ $SNR_{90}$) 2012
   - Full analysis of the \textit{Articulation Index} (AI) 2012
3. Measure CV confusions in $\approx$50 \textit{hearing impaired} ears 2009
   - Characterize hearing impaired (HI) CV confusions 2010
   - Explain HI re NH feature extraction deficiencies, based on \textit{individual-differences} in CV confusions 2012-13
   - Hypothesis: HI Consonant discrimination in noise is due to:
     \begin{itemize}
     \item $\Rightarrow$ Poor acoustic time/freq edge detection?
     \item $\Rightarrow$ Auditory plasticity?
     \item $\Rightarrow$ Cochlear Dead regions?
     \end{itemize}
Normal Hearing listeners can identify most consonant-vowel (CV) sounds above chance at -18 dB SNR-SWN (?)
Normal Hearing listeners can identify most consonant-vowel (CV) sounds above chance at -18 dB SNR-SWN (SNR) = 1 − P<sub>e</sub>.

Phone-error patterns for normal ears:

Phone-error patterns for HI subject 112R:

Normal Hearing have zero error ≥ -2dB SNR.
Normal Hearing listeners can identify most consonant-vowel (CV) sounds above chance at -18 dB SNR-SWN.

(e) Phone-error patterns for normal ears

(f) Phone-error patterns for HI subject 112R

- Normal Hearing have zero error $\geq$ -2dB SNR
- Hearing Impaired (HI) listeners have high error for a few tokens
2. Historical Overview

1. Intro + Objectives 3 mins Σ3
   - Research objectives 5 mins Σ8

2. Historical overview 20 mins Σ28
   - AG Bell 1860, Rayleigh 1910, Fletcher 2021, Shannon 1948

3. Phone Recognition Models 8 mins Σ36
   - Channel capacity and the Articulation Index
   - Speech Psychophysics; Algram/3DDS (cues); Primes and Morphs;
   - Classification models (e.g., DFs)

4. Cochlear Mechanics 15 mins Σ51
   - CBands, NL, Masking, Role re Speech perception; HI ears

5. Summary + Conclusions + Questions 3+3+4 mins Σ76
Lord Rayleigh’s 1908 and George Campbell’s 1910

- Based on AG Bell’s speech studies 1860
■ Lord Rayleigh’s 1908 and George Campbell’s 1910
  ◆ Based on AG Bell’s speech studies 1860
■ Harvey Fletcher’s Articulation Index Al 1921
  ◆ Al first publish: French and Steinberg 1947
  ■ The Al accurately predicts average CV scores $P_c(SNR)$
Lord Rayleigh’s 1908 and George Campbell’s 1910
- Based on AG Bell’s speech studies 1860

Harvey Fletcher’s Articulation Index AI 1921
- AI first publish: French and Steinberg 1947
  - The AI accurately predicts average CV scores $P_c(SNR)$

Shannon  The theory of Information 1948+
- G.A. Miller, Heise and Lichten  *Role of Entropy* 1951
- G.A. Miller & Nicely CM $P_{h|s}(SNR)$ 1955
Classic Speech Studies

- Lord Rayleigh’s 1908 and George Campbell’s 1910
  - Based on AG Bell’s speech studies 1860
- Harvey Fletcher’s Articulation Index AI 1921
  - AI first publish: French and Steinberg 1947
    - The AI accurately predicts average CV scores $P_c(SNR)$
- Shannon The theory of Information 1948+
  - G.A. Miller, Heise and Lichten Role of Entropy 1951
  - G.A. Miller & Nicely CM $P_{h|s}(SNR)$ 1955
- Context effects:
  - G.A. Miller 1951 Language and communication
  - G.A. Miller 1962 5-word Grammar $\equiv$ 4 dB of SNR
  - Boothroyd JASA 1968; Boothroyd & Nittrouer 1988
  - Bronkhorst et al. JASA 1993
Consonant Feature Studies 1950-1990

- Bell Labs 1914-1997
  - Fletcher, Steinberg, French; Shannon; Flanagan; Allen
Consonant Feature Studies 1950-1990

- Bell Labs 1914-1997
  - Fletcher, Steinberg, French; Shannon; Flanagan; Allen
- Haskins Labs 1950-1980
  - Cooper, Liberman, et. al.
Consonant Feature Studies 1950-1990

- **Bell Labs 1914-1997**
  - Fletcher, Steinberg, French; Shannon; Flanagan; Allen

- **Haskins Labs 1950-1980**
  - Cooper, Liberman, et. al.

- **MIT 1970-1990**
  - Stevens+Blumstein; +Alwan, et. al.; +...
Consonant Feature Studies 1950-1990

- Bell Labs 1914-1997
  - Fletcher, Steinberg, French; Shannon; Flanagan; Allen
- Haskins Labs 1950-1980
  - Cooper, Liberman, et. al.
- MIT 1970-1990
  - Stevens+Blumstein; +Alwan, et. al.; +...
- IU 1970-1990
  - Pisoni et. al.; Kewley-Port & Luce 84
- AT&T Labs 1998-2003
  - Allen
Consonant Feature Studies 1950-1990

- Bell Labs 1914-1997
  - Fletcher, Steinberg, French; Shannon; Flanagan; Allen
- Haskins Labs 1950-1980
  - Cooper, Liberman, et. al.
- MIT 1970-1990
  - Stevens+Blumstein; +Alwan, et. al.; +...
- IU 1970-1990
  - Pisoni et. al.; Kewley-Port & Luce 84
- AT&T Labs 1998-2003
  - Allen
- UIUC 2004-2011
  - Allen et. al.: Confusion matrices on NH, HI
Speech Recognition Studies 1990-2013

- **HSR**
  - MIT: Stevens+; Braida+Grant+Rankovic+Alwan+ ...
  - UCLA: Alwan 2000-2013
  - UIUC: AI theory 2006-2012
  - UIUC: HI Confusion matrices 2007-2013
Speech Recognition Studies 1990-2013

- HSR
  - MIT: Stevens+; Braida+Grant+Rankovic+Alwan+
  - UCLA: Alwan 2000-2013
  - UIUC: AI theory 2006-2012
  - UIUC: HI Confusion matrices 2007-2013

- ASR
  - CMU
  - IBM
  - BBN
  - Bell Labs
  - MIT
  - Johns Hopkins
  - ...
Three Recent Literature Reviews:

1. Wright 2004 “A review of perceptual cues and cue robustness”
2. Allen 2005 “Articulation & Intelligibility” Morgan-Claypool

Ten Detailed Studies:

1. Jongman 2000 “Acoustic characteristics of fricatives”
2. Smits 2000 “Temporal distribution . . . in VCVs”
3. Hazan-Simpson 2000 ”cue-enhancement . . . of nonsense words”
4. Jiang 2006 “perception of voicing in plosives”
5. McMurray-Jongman 2011 “information for speech categorization”
7. Jørgensen-Dau 2011; 3 dB change; Modulation references
8. Das-Hansen 2012 “Speech Enhancement ċ Phone Classes”
1. Wright 2004

1. Detailed summary of literature of perceptual cues
   - Bursts, Nasal, VOT, ...
   - Excellent discusses of the Auditory Nerve response (Boosts)

2. Conclusions:
   - Disparity of results (Conclusions weak & unclear)
   - Theories based on very little data
     - most arguments seem dogmatic: neither empirical nor theoretical
   - Lack of theoretical constructs
   - Acoustic cues vary with context (co-articulation)
   - F2 Transitions dominate place perception
   - Burst is a weak cue (susceptible to a low SNR)
     - Fricative noise more robust to noise
   - Extended discussion on robustness and gestures (cue overlap)

Summary: Nice summary of the many misguided attempts at finding speech cues
Review makes it clear there is little agreement in the literature


2. “Listeners are not at ceiling for naturally produced unambiguous tokens”
2. “Listeners are not at ceiling for naturally produced unambiguous tokens”
3. “Recognition depends on multi-dimensional continuous acoustic cues”
2. “Listeners are not at ceiling for naturally produced unambiguous tokens”
3. “Recognition depends on multi-dimensional continuous acoustic cues”
4. “The nature of the perceptual dimensions may matter”
2. “Listeners are not at ceiling for naturally produced unambiguous tokens”
3. “Recognition depends on multi-dimensional continuous acoustic cues”
4. “The nature of the perceptual dimensions may matter”
5. “It’s widely ... accepted that perception compensates for variance.”
2. “Listeners are not at ceiling for naturally produced unambiguous tokens”
3. “Recognition depends on multi-dimensional continuous acoustic cues”
4. “The nature of the perceptual dimensions may matter”
5. “It’s widely ... accepted that perception compensates for variance.”
6. “The interpretation of a cue may depend on the category of others”
2. “Listeners are not at ceiling for naturally produced unambiguous tokens”
3. “Recognition depends on multi-dimensional continuous acoustic cues”
4. “The nature of the perceptual dimensions may matter”
5. “It’s widely ... accepted that perception compensates for variance.”
6. “The interpretation of a cue may depend on the category of others”
7. “Speech perception is a map from continuous acoustic cues to categories”
2. “Listeners are not at ceiling for naturally produced unambiguous tokens”
3. “Recognition depends on multi-dimensional continuous acoustic cues”
4. “The nature of the perceptual dimensions may matter”
5. “It’s widely ... accepted that perception compensates for variance.”
6. “The interpretation of a cue may depend on the category of others”
7. “Speech perception is a map from continuous acoustic cues to categories”
8. “Most speech cues are context-dependent and there are few invariants”
   “there is little question that this is a fundamental issue”

2. “Listeners are not at ceiling for naturally produced unambiguous tokens”
3. “Recognition depends on multi-dimensional continuous acoustic cues”
4. “The nature of the perceptual dimensions may matter”
5. “It’s widely ... accepted that perception compensates for variance.”
6. “The interpretation of a cue may depend on the category of others”
7. “Speech perception is a map from continuous acoustic cues to categories”
8. “Most speech cues are context-dependent and there are few invariants”
   “there is little question that this is a fundamental issue”
9. “Fricatives are signaled by a large number of cues.”
2. “Listeners are not at ceiling for naturally produced unambiguous tokens”
3. “Recognition depends on multi-dimensional continuous acoustic cues”
4. “The nature of the perceptual dimensions may matter”
5. “It’s widely ... accepted that perception compensates for variance.”
6. “The interpretation of a cue may depend on the category of others”
7. “Speech perception is a map from continuous acoustic cues to categories”
8. “Most speech cues are context-dependent and there are few invariants”
   “there is little question that this is a fundamental issue”
9. “Fricatives are signaled by a large number of cues.”
10. “Normalization required to account for large talker variability”
1. **Goal 1:** “What acoustic cues support human-like phone recognition?”
2. “Listeners are not at ceiling for naturally produced unambiguous tokens”
3. “Recognition depends on multi-dimensional continuous acoustic cues”
4. “The nature of the perceptual dimensions may matter”
5. “It’s widely ... accepted that perception compensates for variance.”
6. “The interpretation of a cue may depend on the category of others”
7. “Speech perception is a map from continuous acoustic cues to categories”
8. “Most speech cues are context-dependent and there are few invariants”
   “there is little question that this is a fundamental issue”
9. “Fricatives are signaled by a large number of cues.”
10. “Normalization required to account for large talker variability”
11. Using only a few cues “oversimplifies issues & exaggerates problems”
12. “Speech categorization fundamentally requires massive cue-integration”
2. “Listeners are not at ceiling for naturally produced unambiguous tokens”
3. “Recognition depends on multi-dimensional continuous acoustic cues”
4. “The nature of the perceptual dimensions may matter”
5. “It’s widely ... accepted that perception compensates for variance.”
6. “The interpretation of a cue may depend on the category of others”
7. “Speech perception is a map from continuous acoustic cues to categories”
8. “Most speech cues are context-dependent and there are few invariants”
   “there is little question that this is a fundamental issue”
9. “Fricatives are signaled by a large number of cues.”
10. “Normalization required to account for large talker variability”
11. Using only a few cues “oversimplifies issues & exaggerates problems”
12. “Speech categorization fundamentally requires massive cue-integration”

Summary: Main Goal of study: Resolve significant literature uncertainty
Strong conjectures based on uncertain speech perception literature
“Recognition & normalization deeply intertwined”
Recent Consonant Studies 2000-2013

Two Recent Literature Reviews:
- Wright 2004 “A review of perceptual cues and cue robustness”
- McMurray-Jongman 2011 “information for speech categorization”

Ten Detailed Studies:
1. Jongman 2000 “Acoustic characteristics of fricatives”
2. Smits 2000 “Temporal distribution . . . in VCVs”
3. Hazan-Simpson 2000 ”cue-enhancement . . . of nonsense words”
4. Jiang 2006 “perception of voicing in plosives”
5. McMurray-Jongman 2011 “information for speech categorization”
7. Das-Hansen 2012 “Speech Enhancement ċ Phone Classes”
8. Jørgensen-Dau 2011; Modulation references; 3 dB change
Q: How is place coded for /f, v, θ, ð, s, z, ʃ,ʒ/?

Method: Combinations of 5 static and 2 dynamic measures

Pros:
- Large study: 20 talkers
- High specificity & sensitivity (not for /f, v/ & /θ, ð/)?

Cons:
- Not systematic (trial and error search with many possibilities)
  - No gold standard error control (i.e., human responses)
  - 4 spectral moments (unlikely auditory system to measure these)
  - 4 measures ignore temporal variations
- Claims to solve the fricative phone recognition problem
- Few quantitative conclusions (mostly negative)
Quest for acoustic cues near closure and release in CVC

- Temporal gating of closure & release
- Multi-dimensional scaling (MDS) analysis (4D)
- Transmitted information (with no added noise)

Stimuli: 51 /ąCu/ tokens; 2 talkers (1M, 1F); 17 C, 3 V

Analysis: Response set averaged: Initial+Final Fric, Nasal, Stop
MDS to describe “major confusion patterns”

Results: Distinctive Feature (DF) main variable
Variables: Speaker, vowel context, stress, DF all significant

Conclusions: Results highlight the problem of a rigorous CM analysis
Only a few conclusions
The enhancement of the burst portion of the consonant increases the consonant’s robustness.

Magnitude of the effect is about 1-1.5 SD (1<d'<2)

Similar to Kapoor-Allen 2012 which shifted $P_c(SNR \pm 6dB)$
Alwan says “Jiang conducted voicing discrim exps of natural CV syllables by 4 talkers, in variable amounts of white noise. Onset of F1 is critical to perceiving voicing (not VOT).
1. Analysis summary (a must-read):

- "Information" ≡ acoustic features; "categorization" ≡ perception
- The naive invariance hypothesis: "Are a small number unnormalized cues sufficient for classification?"
- This has not yet been attempted with more powerful logistic regression (appeal to the power of statistics)
- "We did not find any cues that were even modestly invariant for place of articulation in non-sibilants"
- "this cue-set was made solely by statistical reliability (rather than via a theory of production)"
- "The cue-integration hypothesis suggests that if sufficient cues are encoded in detail, their combination is sufficient to overcome single cue variability."
- "normalization required to achieve listener-like performance (Cues are talker-dependent)."
- "Any scaled up system, without normalization, would still need to identify vowels and talkers.

i.e., Listeners naturally compensate for tokens."
Define acoustic cues between labial vs alveolar for plosives and fricatives

Methods: 24 CVs (8 C, 3 V); 4 talkers; White noise (SNR=-15:5:20 dB)
Measures: 17 spectral measures (e.g., F1,2,3, Burst, ...); Manner-dependent Threshold $SNR_{79}^*$

Results: Linear Logit analysis;
- Very strange: $\log(p/(1-p))$ where p is 0 or 1. This seems a serious error.
- Fig 2: $\Delta F2$ correlated to burst for /a/, thus in agreement with Allen et al.
- Fig 2: Not so for /i,u,/
- Makes the case that each of the 24 CVs has one set of support features @80%
- Correlations are quite low 0.2–0.68 with 25% mean error (not impressive)
- “Formants more noise-robust than other spectral measures” (-15 dB = chance); voiceless fricatives lower thresholds than plosives (agreeing with MN55?)
- The present study showed that fricatives had lower threshold SNRs? than plosives and that voiceless fricatives were slightly more robust than the voiced ones.
- within- and across-talker variations were not examined. Within- and across-talker variations is an interesting future topic.

Conclusion: Formants are highlighted as the main feature
### 8 Das-Hansen Speech Enhancement ċ Phone Classes

<table>
<thead>
<tr>
<th>True Class</th>
<th>Vow</th>
<th>Semi</th>
<th>Nas</th>
<th>Aff</th>
<th>Fric</th>
<th>Stop</th>
<th>Clos</th>
<th>Sil</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vow</td>
<td>70.33</td>
<td>11.02</td>
<td>5.51</td>
<td>0.27</td>
<td>3.57</td>
<td>5.12</td>
<td>3.31</td>
<td>0.87</td>
</tr>
<tr>
<td>Semi</td>
<td>17.84</td>
<td>46.69</td>
<td>10.87</td>
<td>0.52</td>
<td>6.78</td>
<td>8.14</td>
<td>6.06</td>
<td>3.10</td>
</tr>
<tr>
<td>Nas</td>
<td>13.22</td>
<td>11.21</td>
<td>42.96</td>
<td>1.70</td>
<td>8.08</td>
<td>8.52</td>
<td>6.77</td>
<td>7.54</td>
</tr>
<tr>
<td>Aff</td>
<td>3.79</td>
<td>1.55</td>
<td>2.59</td>
<td>56.04</td>
<td>10.51</td>
<td>9.14</td>
<td>10.52</td>
<td>5.86</td>
</tr>
<tr>
<td>Fric</td>
<td>3.59</td>
<td>1.61</td>
<td>5.56</td>
<td>4.83</td>
<td>52.08</td>
<td>11.04</td>
<td>13.89</td>
<td>7.40</td>
</tr>
<tr>
<td>Stop</td>
<td>3.63</td>
<td>4.31</td>
<td>10.43</td>
<td>2.51</td>
<td>15.30</td>
<td>41.45</td>
<td>17.31</td>
<td>5.06</td>
</tr>
<tr>
<td>Clos</td>
<td>4.29</td>
<td>3.14</td>
<td>3.38</td>
<td>2.41</td>
<td>20.25</td>
<td>10.91</td>
<td>39.72</td>
<td>15.90</td>
</tr>
<tr>
<td>Sil</td>
<td>1.06</td>
<td>1.73</td>
<td>2.87</td>
<td>2.79</td>
<td>13.33</td>
<td>7.41</td>
<td>17.17</td>
<td>53.64</td>
</tr>
</tbody>
</table>

**Phatak-Allen 2007:**

![Graph A](image)

![Graph B](image)
Based on the utility of the AI($SNR$) they consider the modulation domain SNR as an important speech metric.

- 1.5 dB enhancement
Based on the utility of the AI(SNR) they consider the modulation domain SNR as an important speech metric.

1.5 dB enhancement

Would Forward masking interfere with their hypothesis?
Based on the utility of the AI(SNR) they consider the modulation domain SNR as an important speech metric.

1.5 dB enhancement

Would Forward masking interfere with their hypothesis?

The AI has a very large unaccounted variance. Singh-Allen, 2012
Speech perception is a difficult unsolved problem, >100 years old
The present methods are not working: McMurray & Jongman
Speech perception is a difficult unsolved problem, >100 years old. The present methods are not working: McMurray & Jongman. Why?
Speech perception is a difficult unsolved problem, >100 years old
The present methods are not working: McMurray & Jongman
Why?
Bad assumptions? (e.g., Guessing wrong cues?)
Dysfunctional methods? (e.g., Use of synthetic speech)
Speech perception is a difficult unsolved problem, >100 years old. The present methods are not working: McMurray & Jongman

Why?
- Bad assumptions? (e.g., Guessing wrong cues?)
- Dysfunctional methods? (e.g., Use of synthetic speech)

How can we do this differently? Is there a better way?
Speech perception is a difficult unsolved problem, >100 years old. The present methods are not working: McMurray & Jongman Why?

- Bad assumptions? (e.g., Guessing wrong cues?)
- Dysfunctional methods? (e.g., Use of synthetic speech)

How can we do this differently? Is there a better way? I think so.
Speech perception is a difficult unsolved problem, >100 years old. The present methods are not working: McMurray & Jongman. Why?

Bad assumptions? (e.g., Guessing wrong cues?)
Dysfunctional methods? (e.g., Use of synthetic speech)

How can we do this differently? Is there a better way? I think so.
1. Remove ‘irrelevant’ variables (e.g., context, visual)
2. Don’t try to ‘guess’ the answer
3. Use ‘real’ speech,
Speech perception is a difficult unsolved problem, >100 years old.
The present methods are not working: McMurray & Jongman. Why?

Bad assumptions? (e.g., Guessing wrong cues?)
Dysfunctional methods? (e.g., Use of synthetic speech)

How can we do this differently? Is there a better way? I think so.
1. Remove ‘irrelevant’ variables (e.g., context, visual)
2. Don’t try to ‘guess’ the answer
3. Use ‘real’ speech, with natural variability
4. Take advantage of this natural variability
Speech perception is a difficult unsolved problem, >100 years old.
The present methods are not working: McMurray & Jongman

Why?
Bad assumptions? (e.g., Guessing wrong cues?)
Dysfunctional methods? (e.g., Use of synthetic speech)

How can we do this differently? Is there a better way? I think so.
1. Remove ‘irrelevant’ variables (e.g., context, visual)
2. Don’t try to ‘guess’ the answer
3. Use ‘real’ speech, with natural variability
4. Take advantage of this natural variability
5. Rigorous theoretical (i.e., Communication-theory) analysis
6. Use a large N to avoid complex significance arguments
   Detailed Experimental results with Many talker & listeners
Speech perception is a difficult unsolved problem, >100 years old. The present methods are not working: McMurray & Jongman Why?

- Bad assumptions? (e.g., Guessing wrong cues?)
- Dysfunctional methods? (e.g., Use of synthetic speech)

How can we do this differently? Is there a better way? I think so.

1. Remove ‘irrelevant’ variables (e.g., context, visual)
2. Don’t try to ‘guess’ the answer
3. Use ‘real’ speech, with natural variability
4. Take advantage of this natural variability
5. Rigorous theoretical (i.e., Communication-theory) analysis
6. Use a large N to avoid complex significance arguments

Detailed Experimental results with Many talker & listeners

Summary: Rigorous experimental methods & simple analysis $P_{h|s}(SNR)$, based on communication and information theory

<table>
<thead>
<tr>
<th>Year</th>
<th>Experiment</th>
<th>Student &amp;Allen</th>
<th>Details</th>
<th>Publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>MN04(MN64)</td>
<td>Phatak</td>
<td>16C+4V SWN</td>
<td>JASA (2007)</td>
</tr>
<tr>
<td></td>
<td>HIMCL05</td>
<td>Phatak</td>
<td></td>
<td>JASA (2009)</td>
</tr>
<tr>
<td>2006</td>
<td>HINATALR05</td>
<td>Yoon et al. Regnier</td>
<td>10 HI ears /ta/</td>
<td>JSLR (2012)</td>
</tr>
<tr>
<td>2007</td>
<td>CV06</td>
<td>Pan</td>
<td>Vowels</td>
<td>JASA (2009)</td>
</tr>
<tr>
<td></td>
<td>HL07</td>
<td>Li</td>
<td>Hi/Lo pass</td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td>TR08</td>
<td>Li</td>
<td>Time-truncation</td>
<td>ASSP (2009)</td>
</tr>
<tr>
<td>2009</td>
<td>3DDS Verification</td>
<td>Li Abhinauv Cvangros</td>
<td>Stops burst mods</td>
<td>TASLP (2011)</td>
</tr>
<tr>
<td></td>
<td>MN64 NZE</td>
<td>Singh</td>
<td>within-C $P_e; AI$</td>
<td>JASA (2010)</td>
</tr>
<tr>
<td></td>
<td>Verification</td>
<td></td>
<td></td>
<td>JASA (2012)</td>
</tr>
<tr>
<td>2011</td>
<td>3DDS</td>
<td>Li, Trevino</td>
<td>Fricatives</td>
<td>JASA (2012)</td>
</tr>
<tr>
<td>2010</td>
<td>HIMCL10-II</td>
<td>Trevino</td>
<td>17 HI ears @MCL</td>
<td>JASA (2013)</td>
</tr>
<tr>
<td></td>
<td>HINAL11-IV</td>
<td>Han</td>
<td>17 HI ears+NALR</td>
<td>Thesis Ch. 3</td>
</tr>
</tbody>
</table>

22 / 54
Theory should be based on Shannon’s Theory of Information

1. SNR and Entropy (& token!) are key variables: \( AI(SNR) \) and channel capacity \( C(SNR) \)
2. Token Phone error is binary wrt SNR
3. Tokens have a large threshold SD
   - Never Averaging across tokens!
   - Do not use DF (depends on averages)
4. Entropy is the ideal measure of confusions
5. Very few studies consider Entropy vs. SNR
   - NO: Fletcher 1914-1950
   - YES: Miller Nicely 1955
6. The \( AI(SNR) \) has a huge “across & within” consonant SD

Summary: Information Theory: “the systematic way to proceed”
AI($SNR$) characterizes the average consonant error ($P_e = e_{\text{AI min}}$)

AI ignores the huge across-consonant Standard Deviation (SD)

as well as the huge within-consonant SD Singh-Allen 2012
Within-consonant Error /p/  Singh-Allen 2012

- 56 /p/+/o,e,i/ CV tokens: SNR > -10 dB SNR
- Bimodal error distribution:
  - 41/56: Zero error (ZE); \(N_{\text{trials}} = 38\), \(N_{\text{subj}}=25\)
  - 15/56: Non-zero error (NZE); \(11 \approx \text{ZE} \) (error: 1/38)

![Graph showing zero and non-zero error categories for SNR above -10 dB]

- 11 LE utterances
- 1 ME utterance
- 3 HE utterances

- 15 /p/ utterances in the NZE group
Within-consonant error $P_e(SNR - SNR_{50}^*)$ for /p/

- Error vs. $SNR$ shifted to 50% threshold $SNR_{50}^*$ (LEFT)
- Histogram of 50% error thresholds (RIGHT)

- Sharp transition $\Rightarrow$ Binary Plosive identification!

(a) $P_e(SNR - SNR_{50}^*)$

(b) Distribution of $SNR_{50}^*$
Most stops have zero error (ZE+LE) above -10 dB SNR

Summary of the plosive errors

- ZE
- LE
- ME
- HE

Number of utterances

Plosives: k, g, p, t, d, b
Most stops have zero error (ZE+LE) above -10 dB SNR

Bimodal error distribution for ≥ -2 dB SNR

While speech is highly variable, NH listeners are not
Error summary for Stops Singh-Allen 2012

- Most stops have zero error (ZE+LE) above -10 dB SNR

![Summary of the plosive errors]

- Bimodal error distribution for $\geq -2$ dB SNR
- While speech is highly variable, NH listeners are not
- The AI is an average measure
  - Huge ‘across– & ‘within–consonant’ SD (85% of the variance)
  - SNR depends only on binary threshold distributions
3. Phone Recognition Models

1. Intro + Objectives 3 mins Σ3
   - Research objectives 5 mins Σ8

2. Historical overview 20 mins Σ28
   - AG Bell 1860, Rayleigh 1910, Fletcher 2021, Shannon 1948

3. Phone Recognition Models 8 mins Σ36
   - Channel capacity and the Articulation Index
   - Speech Psychophysics; Algram/3DDS (cues); Primes and Morphs;
   - Classification models (e.g., DFs)

4. Cochlear Mechanics 15 mins Σ51
   - CBands, NL, Masking, Role re Speech perception; HI ears

5. Summary + Conclusions + Questions 3+3+4 mins Σ76
We need rigorous procedures for analyzing speech elements
We need rigorous procedures for analyzing speech elements

- Basic model of acoustic vs. perceptual cue identification

![Diagram](image-url)
We need rigorous procedures for analyzing speech elements

- Basic model of **acoustic vs. perceptual cue** identification

<table>
<thead>
<tr>
<th>PHYSICAL</th>
<th>PERCEPTUAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACOUSTIC FEATURES</td>
<td>EVENTS</td>
</tr>
</tbody>
</table>

  $\Phi$ \rightarrow \text{LISTENER} \rightarrow $\Psi$

We define two basic measures:

- Physical Input: AI-Gram
- Perceptual Output: Confusion matrix
Model of Human Speech Recognition HSR

- Research Goal: Identify *elemental HSR cues*
  - An event is defined as a *perceptual feature*
  - Event errors are measured by band errors $e_k$

Output: Cochlea Event Phones Syllables Words

\[ s(t) \]

Measure: $AI_k$
Formula: $\propto snr_k$ dB

\[ e_k = 0.82^{\Delta k} = 1 - e_1e_2...e_{20} \]

Analog objects $\Phi$ "Front-end"
Discrete objects $\Psi$ "Back-end"
The Channel capacity theorem gives the zero error SNR bound:

$$C(SNR) \equiv \int \log_2 \left(1 + snr^2(f)\right) df \approx AI(SNR) \quad (1)$$
The Channel capacity theorem gives the zero error SNR bound:

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

For a Maximum Entropy (MaxEnt) speech source, the maximum information rate is determined by the AI(SNR)
The Channel capacity theorem gives the zero error SNR bound:

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

For a Maximum Entropy (MaxEnt) speech source, the maximum information rate is determined by the AI(SNR)

The Al-gram is a closely related measure
The Channel capacity theorem gives the zero error SNR bound:

\[ C(SNR) \equiv \int \log_2 (1 + snr^2(f)) \, df \approx AI(SNR) \quad (1) \]

- For a Maximum Entropy (MaxEnt) speech source, the maximum information rate is determined by the AI(SNR)
- The AI-gram is a closely related measure

Is the human operating below the channel capacity?
The Channel capacity theorem gives the zero error SNR bound:

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

- For a Maximum Entropy (MaxEnt) speech source, the maximum information rate is determined by the AI(SNR)
- The AI-gram is a closely related measure

Is the human operating below the channel capacity?
- Probably YES:
The Channel capacity theorem gives the zero error SNR bound:

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

- For a Maximum Entropy (MaxEnt) speech source, the maximum information rate is determined by the AI(SNR)
- The AI-gram is a closely related measure

Is the human operating below the channel capacity?
- Probably YES:
- Fletcher’s AI is similar to Shannon’s channel-capacity measure
The Channel capacity theorem gives the zero error SNR bound:

$$C(SNR) \equiv \int \log_2 \left(1 + snr^2(f)\right) df \approx AI(SNR) \quad (1)$$

- For a Maximum Entropy (MaxEnt) speech source, the maximum information rate is determined by the AI(SNR)
- The AI-gram is a closely related measure

Is the human operating below the channel capacity?

- Probably YES:
  - Fletcher’s AI is similar to Shannon’s channel-capacity measure
  - The Phone error is zero above $-10$ dB SNR (Eq. 1)

Singh & Allen 2012
3. Results for Normal Hearing (NH) ears

- The AI predicts $P_e(SNR)$, but with a huge SD ($\sigma_{AI}(SNR)$)
3. Results for Normal Hearing (NH) ears

- The AI predicts $P_e(SNR)$, but with a huge SD ($\sigma_{AI}(SNR)$).

![Graph showing consonant error vs. wideband SNR for Normal Hearing (NH) ears.](image)
3. Results for Normal Hearing (NH) ears

- The AI predicts $P_e(SNR)$, but with a huge SD ($\sigma_{AI}(SNR)$)

- Averaging obscures large across-consonant errors $\sigma_{AI}(SNR)$
- The $SIN_c$ of averaging: across-consonant error
Methods: The count (confusion) matrix

Miller-Nicely’s 1955 articulation matrix $P_{h|s}(SNR)$, measured at [-18, -12, -6 shown, 0, 6, 12] dB SNR

<table>
<thead>
<tr>
<th>STIMULUS</th>
<th>UNVOICED</th>
<th>RESPONSE</th>
<th>VOICED</th>
<th>NASAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>80</td>
<td>43</td>
<td>64</td>
<td>1</td>
</tr>
<tr>
<td>t</td>
<td>71</td>
<td>84</td>
<td>55</td>
<td>1</td>
</tr>
<tr>
<td>k</td>
<td>66</td>
<td>76</td>
<td>107</td>
<td>1</td>
</tr>
<tr>
<td>f</td>
<td>18</td>
<td>12</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>θ</td>
<td>19</td>
<td>17</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>s</td>
<td>8</td>
<td>5</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>s̄</td>
<td>1</td>
<td>6</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>b</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>d</td>
<td>2</td>
<td>3</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>g</td>
<td>136</td>
<td>10</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>v</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>ð</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>z</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>m</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>n</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

Table III. Confusion matrix for $S/N = -6$ db and frequency response of 200–6500 cps.
Methods: The count (confusion) matrix

Miller-Nicely’s 1955 articulation matrix \( P_{h|s}(SNR) \), measured at [-18, -12, -6 shown, 0, 6, 12] dB SNR

<table>
<thead>
<tr>
<th></th>
<th>p</th>
<th>t</th>
<th>k</th>
<th>f</th>
<th>θ</th>
<th>s</th>
<th>§</th>
<th>b</th>
<th>d</th>
<th>g</th>
<th>v</th>
<th>δ</th>
<th>z</th>
<th>3</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></td>
<td>2</td>
</tr>
<tr>
<td>t</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></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></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></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</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></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></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></td>
</tr>
<tr>
<td>§</td>
<td>1</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>29</td>
<td>195</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>b</td>
<td></td>
<td>5</td>
<td>4</td>
<td>4</td>
<td></td>
<td>136</td>
<td>10</td>
<td>9</td>
<td>47</td>
<td>16</td>
<td>6</td>
<td>1</td>
<td></td>
<td>5</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td></td>
<td>2</td>
<td></td>
<td>63</td>
<td>66</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>11</td>
<td>20</td>
<td>20</td>
<td>26</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>g</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>19</td>
<td>37</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>v</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>δ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>z</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>m</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>6</td>
<td>177</td>
<td>46</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TABLE III. Confusion matrix for S/N = -6 db and frequency response of 200–6500 cps.

Confusion groups \( \equiv inhomogeneous confusions \)
This *confusion pattern* characterizes the /t/ row vs SNR.

Confusion patterns for /t/ (Miller Nicely ‘55)

Row of confusion matrix (CM) $P_{h|/t/}$
The $\text{SIN}_t$ of averaging within-consonants (i.e., tokens):

- Token confusions are strongly heterogeneous!
- Averaging obscures per-token confusions

(a) Average over all /t/s.
The $\text{SIN}_t$ of averaging within-consonants (i.e., tokens):

- Token confusions are strongly **heterogeneous**!
- Averaging obscures per-token confusions

(a) Average over all /t/s.  
(b) Talker m117 /te/ $P_{h|/t\alpha/}(SNR)$
The $\text{SIN}_t$ of averaging within-consonants (i.e., tokens):

- Token confusions are strongly **heterogeneous**!
- Averaging obscures per-token confusions

(a) Average over all /t/ tokens.

(b) Talker m117 /te/ $P_{h\mid t\alpha\langle SNR\rangle}$

- Never average across tokens!
Methods to Identify Acoustic Features

- Identify the key features in *individual* CV tokens
Methods to Identify Acoustic Features

- Identify the key features in individual CV tokens
  - Plosives (e.g., /p, t, k/ and /b, d, g/)
  - Fricatives (e.g., /θ, ʃ, tʃ, s, h, f/ and /z, ʒ, v, ʒ/)
  - With vowels /o, e, i/

- ≈18 talkers and >20 listeners
- Up to 20 trials per consonant per SNR
Methods to Identify Acoustic Features

- Identify the key features in individual CV tokens
  - Plosives (e.g., /p, t, k/ and /b, d, g/)
  - Fricatives (e.g., /θ, ʃ, tʃ, s, h, f/ and /z, ʒ, v, ʒ/)  
  - With vowels /o, e, i/

- ≈18 talkers and >20 listeners
- Up to 20 trials per consonant per SNR

- Method: $3^d$ Deep-Search (3DDS) via truncations (no guessing):
  - Time truncation Furui 1986
  - Intensity truncation (i.e., masking)
  - Frequency truncation (High/Low-pass filtering)
Methods to Identify Acoustic Features

- Identify the key features in individual CV tokens
  - Plosives (e.g., /p, t, k/ and /b, d, g/)
  - Fricatives (e.g., /θ, ʃ, tʃ, s, h, f/ and /z, ʒ, v, ð/)
  - With vowels /o, e, ɪ/

  - ≈18 talkers and >20 listeners
  - Up to 20 trials per consonant per SNR

- Method: 3rd Deep-Search (3DDS) via truncations (no guessing):
  - Time truncation Furui 1986
  - Intensity truncation (i.e., masking)
  - Frequency truncation (High/Low-pass filtering)

- Methods: Cochlear models & signal processing
  - Algram Régnier & Allen 2008; Li & Allen 2009,10,11
Methods: $3^d$ Deep Search (3DDS)

- $3^d$ Deep-Search ($3^d$-DS) via truncation (triangulate):
Methods: $3^d$ Deep Search (3DDS)

- $3^d$ Deep-Search ($3^d$-DS) via truncation (triangulate):
  - Time truncation Furui 1986
Methods: $3^d$ Deep Search ($3^d$DDS)

- $3^d$ Deep-Search ($3^d$-DS) via truncation (triangulate):
  - Time truncation Furui 1986
  - Frequency truncation (High/Low-pass filtering)
Methods: $3^d$ Deep Search (3DDS)

- $3^d$ Deep-Search ($3^d$-DS) via truncation (triangulate):
  - Time truncation Furui 1986
  - Frequency truncation (High/Low-pass filtering)
  - Intensity truncation (i.e., masking)
Methods: $3^d$ Deep Search ($3^d$DDS)

$3^d$ Deep-Search ($3^d$-DS) via truncation (triangulate):

- Time truncation Furui 1986
- Frequency truncation (High/Low-pass filtering)
- Intensity truncation (i.e., masking)
Summary of Consonant structure

- Time-frequency structure of plosives and fricatives

  Plosives: /p, t, k, b, d, g/+/a/

  Approximate duration: ≈ 50 [ms]
# 4. Cochlear Mechanics

1. **Intro + Objectives**
   - 3 mins Σ3
     - Research objectives 5 mins Σ8

2. **Historical overview**
   - 20 mins Σ28
     - AG Bell 1860, Rayleigh 1910, Fletcher 2021, Shannon 1948

3. **Phone Recognition Models**
   - 8 mins Σ36
     - Channel capacity and the Articulation Index
     - Speech Psychophysics; Algram/3DDS (cues); Primes and Morphs;
     - Classification models (e.g., DFs)

4. **Cochlear Mechanics**
   - 15 mins Σ51
     - CBands, NL, Masking, Role re Speech perception; HI ears

5. **Summary + Conclusions + Questions**
   - 3+3+4 mins Σ76
1910-1980: Bell Labs (long history)

- Fletcher 1914; Wegel & Lane 1924; Flanagan; Hall; Allen
Auditory & Cochlear Modeling 1920-2000

- **1910-1980**: Bell Labs (long history)
  - Fletcher *1914*; Wegel & Lane *1924*; Flanagan; Hall; Allen
- **1960-2010**: MIT + Harvard HSBT
  - Eaton Peabody (Kiang, Siebert, Liberman, Guinan, Shera, …)
1910-1980: Bell Labs (long history)
- Fletcher 1914; Wegel & Lane 1924; Flanagan; Hall; Allen

1960-2010: MIT + Harvard HSBT
- Eaton Peabody (Kiang, Siebert, Liberman, Guinan, Shera, . . .)

Netherlands, England
- deBoer, Duifhuis, Evans, . . .

Australia (B. Johnstone, . . .)
Auditory & Cochlear Modeling 1920-2000

- **1910-1980**: Bell Labs (long history)
  - Fletcher 1914; Wegel & Lane 1924; Flanagan; Hall; Allen
- **1960-2010**: MIT + Harvard HSBT
  - Eaton Peabody (Kiang, Siebert, Liberman, Guinan, Shera, . . .)
- Netherlands, England
  - deBoer, Duifhuis, Evans, . . .
- Australia (B. Johnstone, . . .)
- **1980-2011**: NIH funded University research
  - MIT; Wash U; Boys Town; U. Wisc.; U. Mich.; Northwestern U.
Auditory & Cochlear Modeling 1920-2000

- **1910-1980**: Bell Labs (long history)
  - Fletcher 1914; Wegel & Lane 1924; Flanagan; Hall; Allen

- **1960-2010**: MIT + Harvard HSBT
  - Eaton Peabody (Kiang, Siebert, Liberman, Guinan, Shera, . . .)

- Netherlands, England
  - deBoer, Duifhuis, Evans, . . .

- Australia (B. Johnstone, . . .)

- **1980-2011**: NIH funded University research
  - MIT; Wash U; Boys Town; U. Wisc.; U. Mich.; Nortwestern U.

- The role of cochlear modeling on speech perception is huge!
  - And underappreciated, IMO
The Human Cochlea
The Human Cochlea

(a)

SV
Scalae vestibuli
(periosteal space)

Vestibular membrane
(Reissner)

Secreting epithelium

Area vascularis

Ductus cochlearis
(endotic space)

SM
Teclorial membrane

External spiral sulcus

Spiral ligament

Spiral artery

Spiral ganglion

Capsule of gang cell

Myelin sheath

Limbus spiralis

Internal spiral sulcus

Basilar membrane

Crista basilaris

Spiral organ (Corti)

Scala tympani
(periosteal space)

ST
The Cochlear duct
This effect leads to forward masking
Forward Masking is a very large effect lasting for up to 200 ms
Onset transients enhance the auditory nerve response, to 2 [cs].
Onset transients **enhance** the auditory nerve response, to 2 [cs]

Forward Masking **depresses** the response up to 40 dB, to 20 [cs]
### 6. Summary + Conclusions + Questions

<table>
<thead>
<tr>
<th>Section</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Intro + Objectives</td>
<td>3 mins (\Sigma 3)</td>
</tr>
<tr>
<td>Research objectives</td>
<td>5 mins (\Sigma 8)</td>
</tr>
<tr>
<td>2. Historical overview</td>
<td>20 mins (\Sigma 28)</td>
</tr>
<tr>
<td>AG Bell 1860, Rayleigh 1910, Fletcher 2021, Shannon 1948</td>
<td></td>
</tr>
<tr>
<td>3. Phone Recognition Models</td>
<td>8 mins (\Sigma 36)</td>
</tr>
<tr>
<td>Channel capacity and the Articulation Index</td>
<td></td>
</tr>
<tr>
<td>Speech Psychophysics; Algram/3DDS (cues); Primes and Morphs;</td>
<td></td>
</tr>
<tr>
<td>Classification models (e.g., DFs)</td>
<td></td>
</tr>
<tr>
<td>4. Cochlear Mechanics</td>
<td>15 mins (\Sigma 51)</td>
</tr>
<tr>
<td>CBands, NL, Masking, Role re Speech perception; HI ears</td>
<td></td>
</tr>
<tr>
<td>5. <strong>Summary + Conclusions + Questions</strong></td>
<td>3+3+4 mins (\Sigma 76)</td>
</tr>
</tbody>
</table>
New methods:

1. *Al-gram* based on centi-second & critical band scales
New methods:

1. **AI-gram** based on centi-second & critical band scales
2. **3DDS** (truncate: time, freq, intensity) to isolated cues: *Plosives* /p, t, k/, /b, d, g/ + *Fricatives* /θ, j, ŋ, s, h, f/, /z, ʒ, v, ð/) + vowels /o, e, ɪ/
New methods:

1. **AI-gram** based on centi-second & critical band scales
2. **3DDS** (truncate: time, freq, intensity) to isolated cues: Plosives /p, t, k/, /b, d, g/ + Fricatives /θ, ʃ, ʒ, s, h, f/, /z, ʒ, v, ɹ/) + vowels /o, e, ɪ/
3. Data on discriminating consonants in noise, NH listeners use
   - Plosives: *Burst + timing to Voicing*
   - Fricatives: *Low-frequency edge + duration + F₀ modulation*

5. **STFT** to manipulate speech:
   - Morph consonants (e.g., /k/ to /t/ to /p/)
   - Intelligibility: Modify SNR₉₀
We have demonstrated:
1. Speech cue detection is binary (6 dB SNR range)
We have demonstrated:

1. Speech cue detection is binary (6 dB SNR range)
2. Explained the AI properties:
We have demonstrated:
1. Speech cue detection is binary (6 dB SNR range)
2. Explained the AI properties:
3. Established the basis of acoustic cues
   - Burst, frequency-edge, timing & SNR\textsubscript{50} distributions
We have demonstrated:

1. Speech cue detection is binary (6 dB SNR range)
2. Explained the AI properties:
3. Established the basis of acoustic cues
   - Burst, frequency-edge, timing & SNR\(_{50}\) distributions
   - \(P_e(SNR) = e^{SNR}_{min}\) due to SNR\(_{50}^*\) distribution
We have demonstrated:

1. Speech cue detection is binary (6 dB SNR range)
2. Explained the AI properties:
3. Established the basis of acoustic cues
   - Burst, frequency-edge, timing & SNR_{50} distributions
   - \( P_e(SNR) = e^{SNR_{min}} \) due to SNR_{50}^* distribution
4. Explored the natural existence of conflicting cues
We have demonstrated:

1. Speech cue detection is binary (6 dB SNR range)
2. Explained the AI properties:
3. Established the basis of acoustic cues
   - Burst, frequency-edge, timing & SNR\textsubscript{50} distributions
   - \( P_e(SNR) = e^{SNR}_{min} \) due to SNR\textsuperscript{*}_{50} distribution

3. Explored the natural existence of conflicting cues
   - This could impact ASR systems
Findings re HI ears:
1. HI ears have huge individual differences
Findings re HI ears:
1. HI ears have huge individual differences
   - Individual differences dominate HI results
Findings re HI ears:

1. HI ears have huge individual differences
   - Individual differences dominate HI results
   - No two ears are the same
Findings re HI ears:

1. HI ears have huge individual differences
   - Individual differences dominate HI results
   - No two ears are the same
   - Low correlations between HL(f) and $P_e(SNR)$
Findings re HI ears:

1. HI ears have huge individual differences
   - Individual differences dominate HI results
   - No two ears are the same
   - Low correlations between HL(f) and $P_e(SNR)$

2. Each ear has a different consonant recognition strategy
Findings re HI ears:

1. HI ears have huge individual differences
   - Individual differences dominate HI results
   - No two ears are the same
   - Low correlations between HL(f) and $P_e(SNR)$

2. Each ear has a different consonant recognition strategy

3. A better understanding of HI acoustic cue detection will lead to:
   - Improved understanding of HSR for NH & HI ears
   - Better signal processing methods
   - Speech-aware hearing aids in 5 years $\geq 2016$
     - Individual fitting based on specific confusions
Question your basic assumptions

Thank you for your attention

http://hear.ai.uiuc.edu/
http://hear.ai.uiuc.edu/wiki/Main/Publications
Discussion: “Helpful” speech-perception categories

- ‘Distinctive features,’ ‘Acoustic cues,’ & ‘Perceptual cues’
- Synthetic speech
  - Assumes cues [F2(t), Modulations, durations, ...]
  - Low Entropy of experimental task?
    - One parameter (e.g., F2) typically varied
    - Human CV speech is an open-set 11 bit task!
    - Context reduces the entropy (Sentences; Key words; Known material)
- Noise (type, amount, analysis method?)
  - “Babble” you can almost understand (e.g., 1-talker)
  - Sine-wave speech
- Magnitude of the result (e.g., <6 dB)
- Suggestions from you ...?