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

Outline I

1. Intro + Objectives 3 mins $\Sigma 3$
   - Research objectives 5 mins $\Sigma 8$

Outline I

1. Intro + Objectives 3 mins $\Sigma 3$
   - Research objectives 5 mins $\Sigma 8$
2. Historical overview 20 mins $\Sigma 28$
   - AG Bell 1860, Rayleigh 1910, Fletcher 2021, Shannon 1948
   - Speech-feature studies 1950-1990; $>$1991

Outline I

1. Intro + Objectives 3 mins $\Sigma 3$
   - Research objectives 5 mins $\Sigma 8$
2. Historical overview 20 mins $\Sigma 28$
   - AG Bell 1860, Rayleigh 1910, Fletcher 2021, Shannon 1948
   - Speech-feature studies 1950-1990; $>$1991
3. Phone Recognition Models 21 mins $\Sigma 49$
   - Channel capacity and the Articulation Index
   - Speech Psychophysics; AIgram/3DDS (cues); Primes and Morphs;
   - Classification models (e.g., DFs)

1. Intro + Objectives 3 mins $\Sigma 3$
   - Research objectives 5 mins $\Sigma 8$
2. Historical overview 20 mins $\Sigma 28$
   - AG Bell 1860, Rayleigh 1910, Fletcher 2021, Shannon 1948
   - Speech-feature studies 1950-1990; $>$1991
3. Phone Recognition Models 21 mins $\Sigma 49$
   - Channel capacity and the Articulation Index
   - Speech Psychophysics; AIgram/3DDS (cues); Primes and Morphs;
   - Classification models (e.g., DFs)
4. Cochlear Mechanics 16 mins $\Sigma 65$
   - CBands, NL, Masking, Role re Speech perception; HI ears

Outline I

1. Intro + Objectives 3 mins $\Sigma 3$
   - Research objectives 5 mins $\Sigma 8$
2. Historical overview 20 mins $\Sigma 28$
   - AG Bell 1860, Rayleigh 1910, Fletcher 2021, Shannon 1948
   - Speech-feature studies 1950-1990; $>$1991
3. Phone Recognition Models 21 mins $\Sigma 49$
   - Channel capacity and the Articulation Index
   - Speech Psychophysics; AIgram/3DDS (cues); Primes and Morphs;
   - Classification models (e.g., DFs)
4. Cochlear Mechanics 16 mins $\Sigma 65$
   - CBands, NL, Masking, Role re Speech perception; HI ears
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) $F0$ 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) $F0$ 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) $F0$ 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 ≈50 *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) $F0$ 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 ≈50 *hearing impaired* ears 2009
   - Characterize hearing impaired (HI) CV confusions 2010

1. Objectives of the UIUC HSR Group

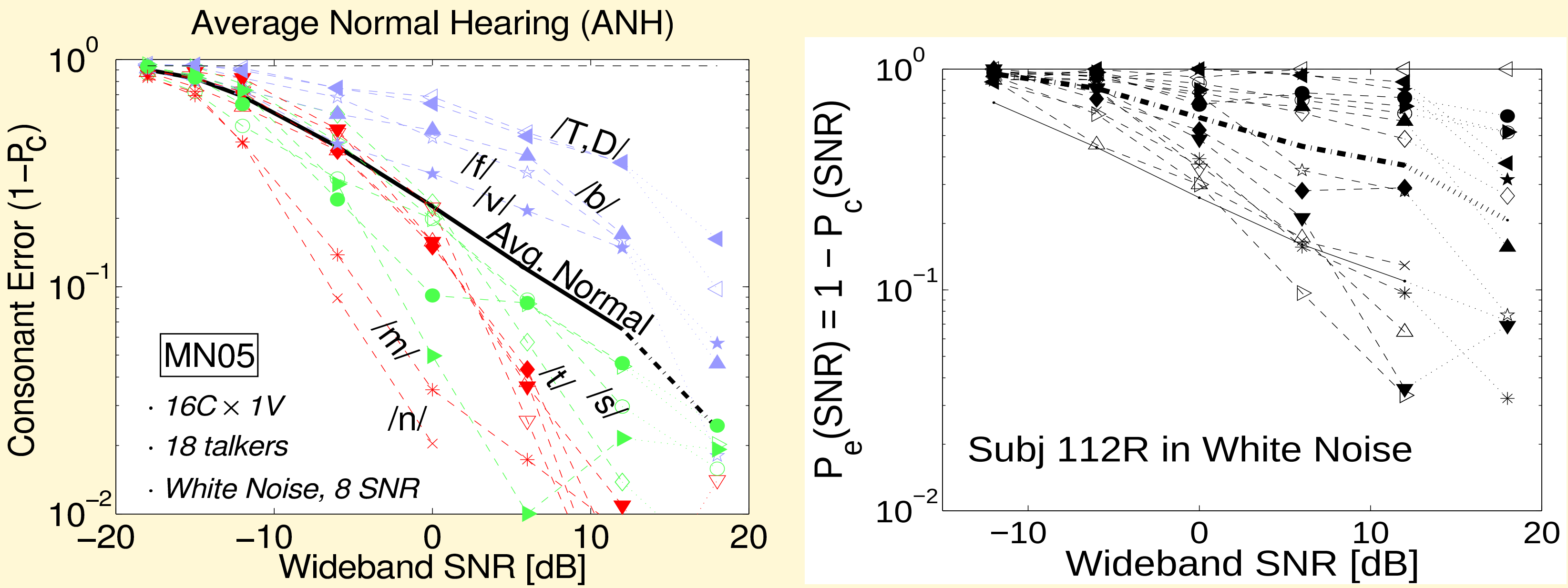
1. Repeat classic experiments on human speech CV sounds 2005
2. Identify acoustic cues in CV tokens 2007
   - *Findings:* a) Onset burst, b) Frequency edge, c) Duration, d) $F0$ modulation, e) Voicing 2007-11
   - Consonant recognition is binary (Threshold @ SNR_{90}) 2012
   - Full analysis of the Articulation Index (AI) 2012
3. Measure CV confusions in ≈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 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 ≈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
   - Hypothesis: HI Consonant discrimination in noise is due to:
     - ⇒ Poor acoustic time/freq edge detection?
     - ⇒ Auditory plasticity?
     - ⇒ Cochlear Dead regions?

- 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 (**?**)

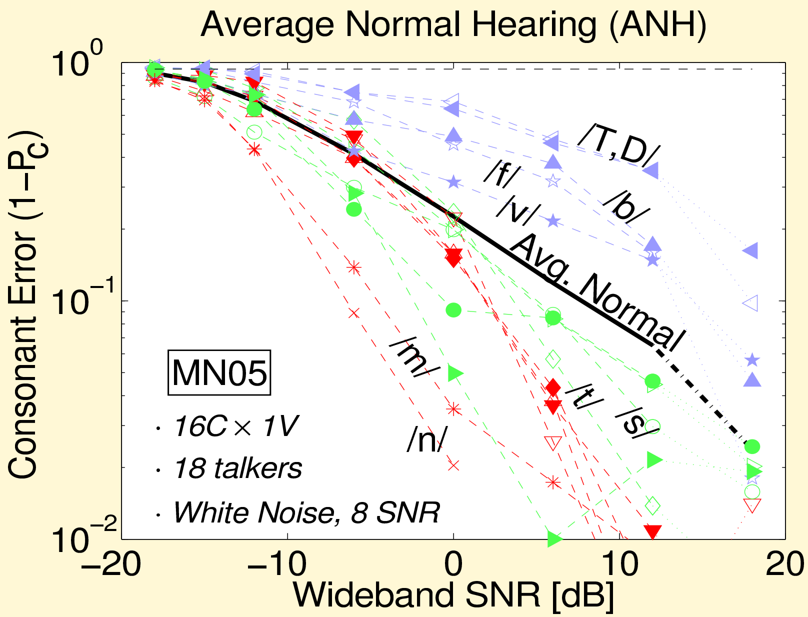


(c) Phone-error patterns for normal ears

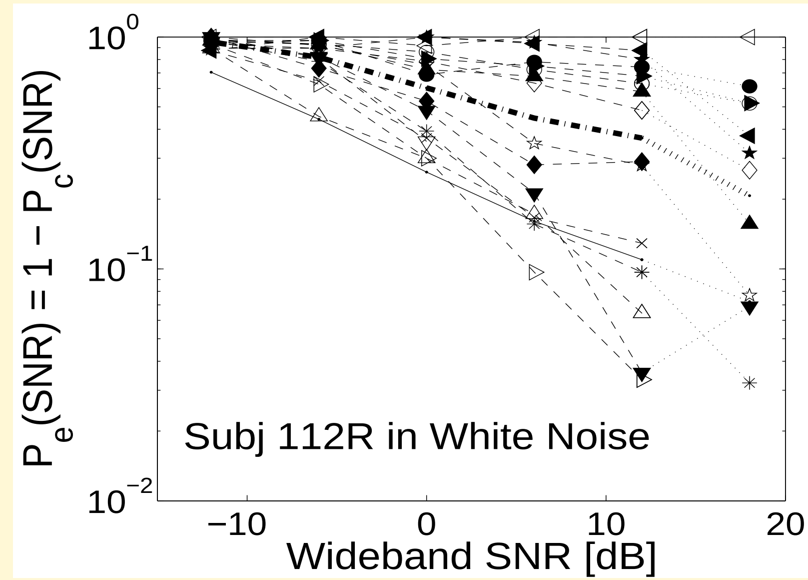
(d) Phone-error patterns for HI subject 112R

- Normal Hearing have zero error $\geq$ -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

Classic Speech Studies

- Lord Rayleigh's 1908 and George Campbell's 1910
  - Based on AG Bell's speech studies 1860

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 $P_c(SNR)$

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 $P_c(\mathit{SNR})$
- Shannon The theory of Information 1948+
  - G.A. Miller, Heise and Lichten *Role of Entropy* 1951
  - G.A. Miller & Nicely CM $P_{h|s}(\mathit{SNR})$ 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
  - AT&T Bell Labs: Theory of HSR 1994-2003
  - 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
  - AT&T Bell Labs: Theory of HSR 1994-2003
  - UIUC: AI theory 2006-2012
  - UIUC: HI Confusion matrices 2007-2013
- ASR
  - CMU
  - IBM
  - BBN
  - Bell Labs
  - MIT
  - Johns Hopkins
  - . . .

Recent Speech Studies 2000-2013

- Three Recent Literature Reviews:
  1. Wright 2004 "A review of perceptual cues and cue robustness"
  2. Allen 2005 *"Articulation & Intelligibility"* Morgan-Claypool
  3. 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"
  6. Alwan 2011 "Perception of place of articulation . . . "
  7. Jørgensen-Dau 2011; 3 dB change; Modulation references
  8. Das-Hansen 2012 "Speech Enhancement $\bar{c}$ Phone Classes"
  9. Singh-Allen 2012 "Stop consonant features & AI"

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

3. McMurray-Jongman 2011

1. Goal 1: “What acoustic cues support human-like phone recognition?”

3. McMurray-Jongman 2011

1. Goal 1: "What acoustic cues support human-like phone recognition?"
2. "Listeners are not at ceiling for naturally produced unambiguous tokens"

3. McMurray-Jongman 2011

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"

3. McMurray-Jongman 2011

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”

3. McMurray-Jongman 2011

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."

3. McMurray-Jongman 2011

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"

3. McMurray-Jongman 2011

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"

3. McMurray-Jongman 2011

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"

3. McMurray-Jongman 2011

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.”

3. McMurray-Jongman 2011

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"

3. McMurray-Jongman 2011

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"

3. McMurray-Jongman 2011

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"

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”
6. Alwan 2011 “Perception of place of articulation . . . ”
7. Das-Hansen 2012 “Speech Enhancement c̄ Phone Classes”
8. Jørgensen-Dau 2011; Modulation references; 3 dB change
9. Singh-Allen 2012 “Stop consonant features & AI”

1 Jongman “Acoustic characteristics of fricatives” 2000

- 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)

2 Smits "Temporal distribution ... in VCVs" 2000

- 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 /ɑCu/ 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

3 Hazan-Simpson ”. . . cue-enhancement” 1998,2000

- 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 $P_c(\mathit{SNR} \pm 6\text{dB})$

4 Jiang “perception of voicing in plosives” 2006

- 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).

5 McMurray-Jongman "speech categorization" 2011

1. Analysis summary (a must-read):

- "Information" ≡ acoustic features; "categorization" ≡ perception
- The *naïve 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."

6 Alwan “Perception of place of articulation . . . ” 2011

- 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;

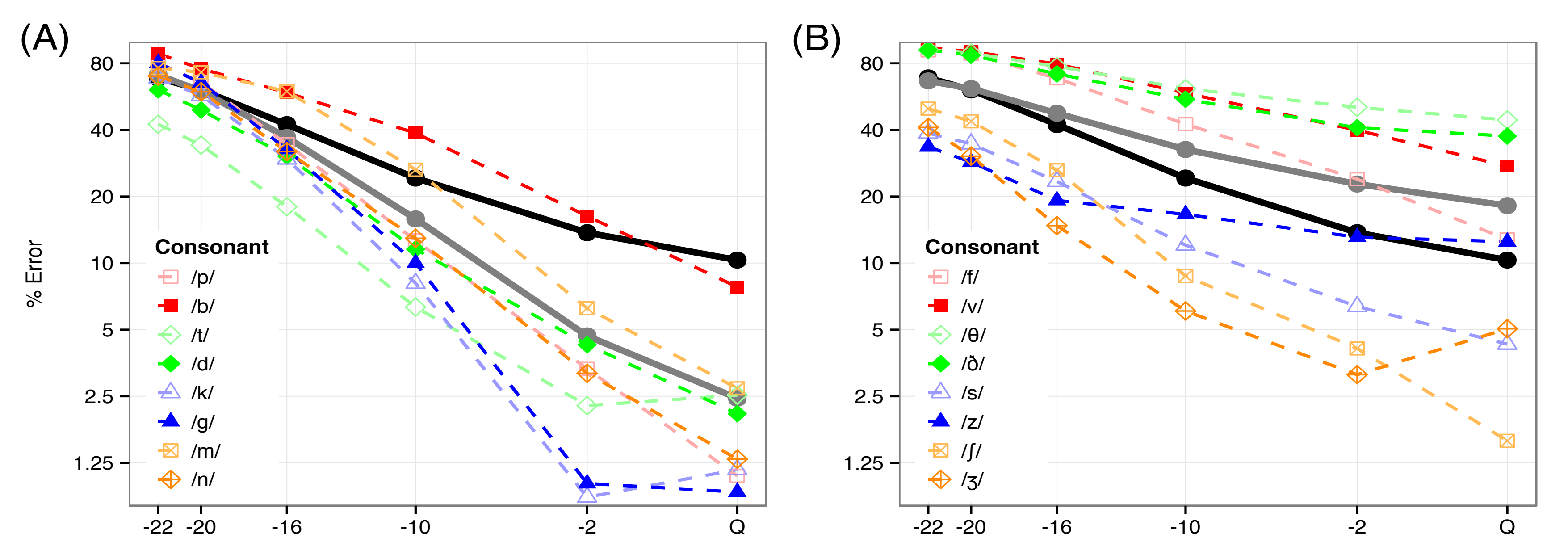
- Very strange: log(p/1-p) where p is 0 or 1. This seems a serious error.
- Fig 2: ΔF2 correlated to burst for /a/, thus in agreement with Allen et al.
- Fig 2: Not so for /i,u,/
- Makes the case that each of the 24 CVs has one set of support features @80%
- Correlations are quite low 0.2–0.68 with 25% 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 c̄ Phone Classes

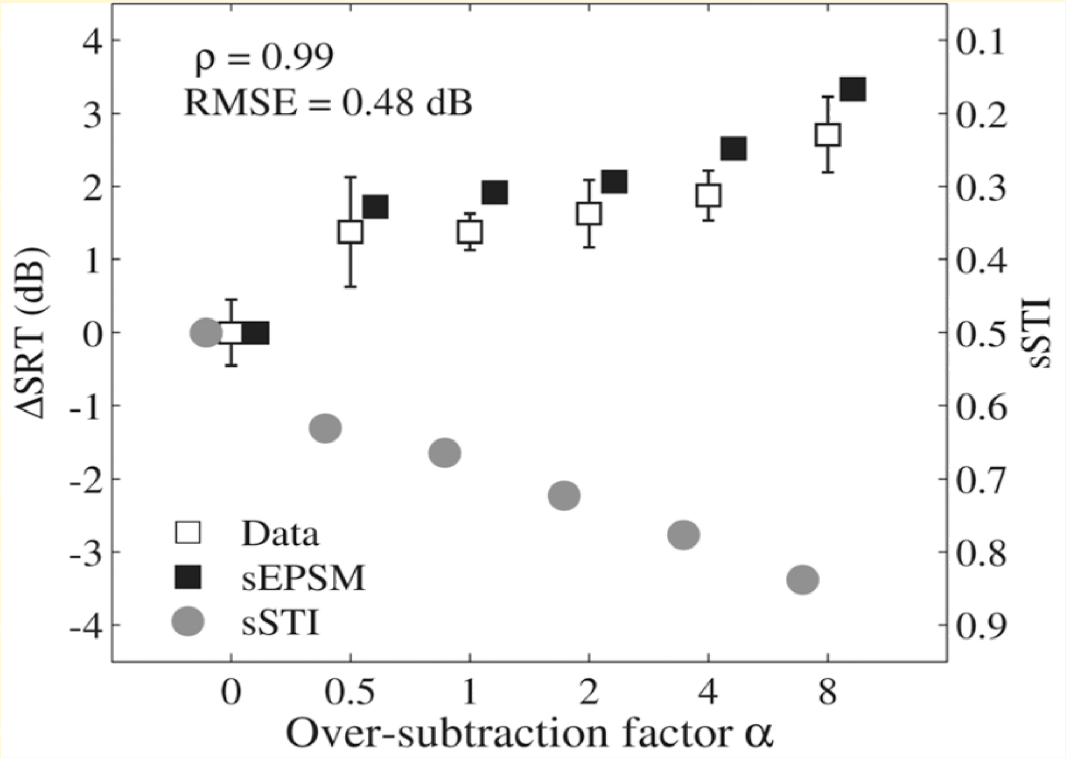
True Class ↓	*VQ Recognized Class* →							
	Vow	Semi	Nas	Aff	Fric	Stop	Clos	Sil
Vow	**70.33**	11.02	5.51	0.27	3.57	5.12	3.31	0.87
Semi	17.84	**46.69**	10.87	0.52	6.78	8.14	6.06	3.10
Nas	13.22	11.21	**42.96**	1.70	8.08	8.52	6.77	7.54
Aff	3.79	1.55	2.59	**56.04**	10.51	9.14	10.52	5.86
Fric	3.59	1.61	5.56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

Phatak-Allen 2007:



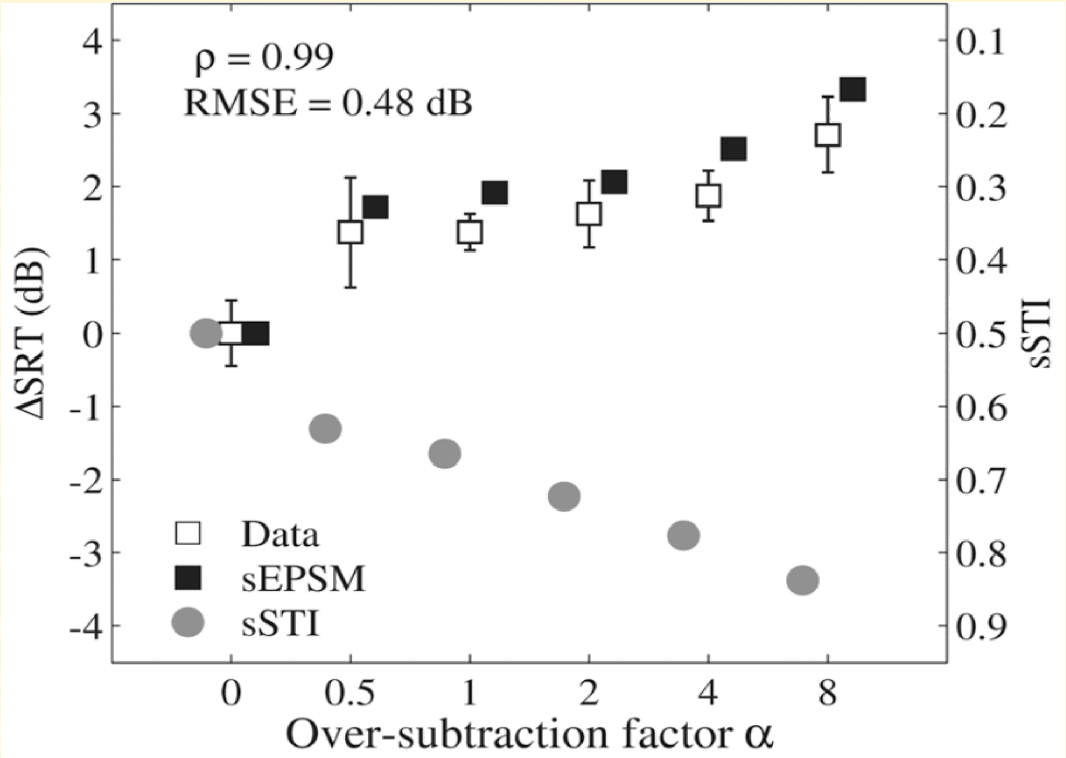
9 Jørgensen-Dau: Modulation filter-bank & STI 2011

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



9 Jørgensen-Dau: Modulation filter-bank & STI 2011

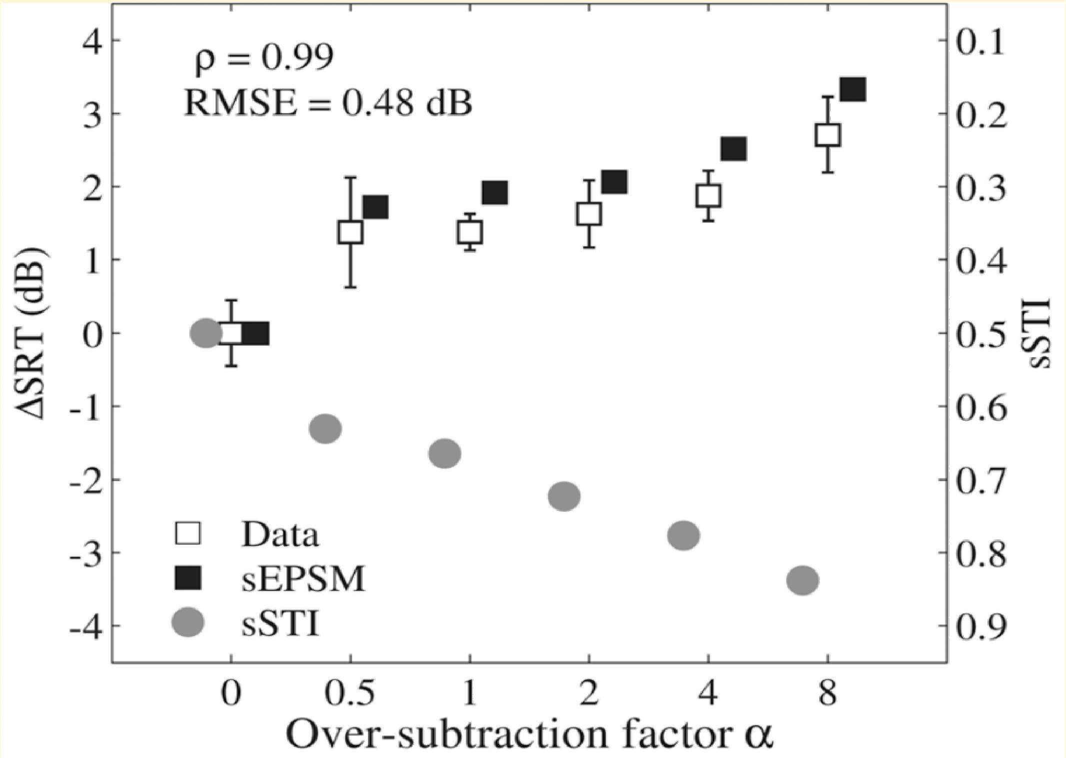
- Based on the utility of the AI(*SNR*) they consider the modulation domain SNR as an important speech metric
- 1.5 dB enhancement



- Would *Forward masking* interfere with their hypothesis?

9 Jørgensen-Dau: Modulation filter-bank & STI 2011

- 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*

Allen et. al Methods 2004-2011

- Speech perception is a difficult unsolved problem, $>$100 years old
  The present methods are not working: McMurray&Jongman

Allen et. al Methods 2004-2011

- Speech perception is a difficult unsolved problem, >100 years old
  The present methods are not working: McMurray&Jongman Why?

Allen et. al Methods 2004-2011

- 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)

Allen et. al Methods 2004-2011

- 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?

Allen et. al Methods 2004-2011

- 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.

Allen et. al Methods 2004-2011

- 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,

Allen et. al Methods 2004-2011

- 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

Allen et. al Methods 2004-2011

- 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

Allen et. al Methods 2004-2011

- 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

3. Allen et. al HSR Experiments 2004-2011

Year	Experiment	Student &Allen	Details	Publication
2004	MN04(MN64)	Phatak	16C+4V SWN	JASA (2007)
2005	MN16R	Phatak, Lovitt	MN55R	JASA (2008)
	HIMCL05	Yoon, Phatak	10 HI ears	JASA (2009)
2006	HINALR05	Yoon *et al.*	10 HI ears	JSLR (2012)
	Verification	Regnier	/ta/ feature	JASA (2008)
	CV06-s/w	Phatak/Regnier	8C+9V SWN/WN	
2007	CV06	Pan	Vowels	
	HL07	Li	Hi/Lo pass	JASA (2009)
2008	TR08	Li	Time-truncation	ASSP (2009)
2009	3DDS	Li	Stops	TASLP (2011)
	3DDS	Li	Stops	JASA (2010)
	Verification	Abhinauv	burst mods	JASA (2012)
	Verification	Cvengros	burst mods	(2012)
	MN64 NZE	Singh	within-C P_e; AI	JASA (2012)
2011	3DDS	Li,Trevino	Fricatives	JASA (2012)
	HINAL11-IV	Han	17 HI ears+NALR	Thesis Ch. 3
2010	HIMCL10-II	Trevino	17 HI ears @MCL	JASA (2013)

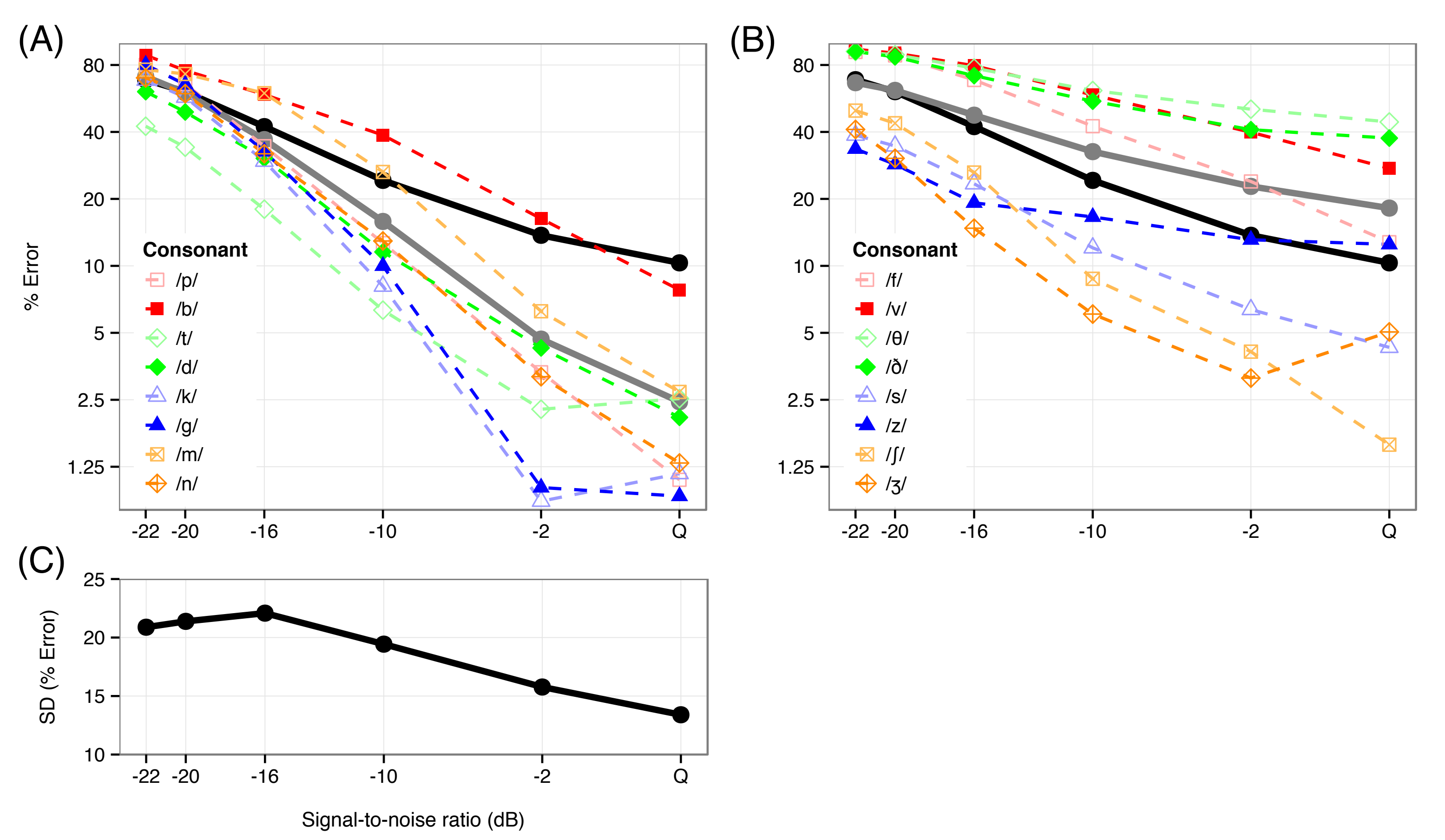
Allen et. al observations: 2004-2011

- Theory should be based on Shannon's Theory of Information

  1. SNR and Entropy (& token!) are key variables:
     AI(SNR) and channel capacity $\mathcal{C}(SNR)$
  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"

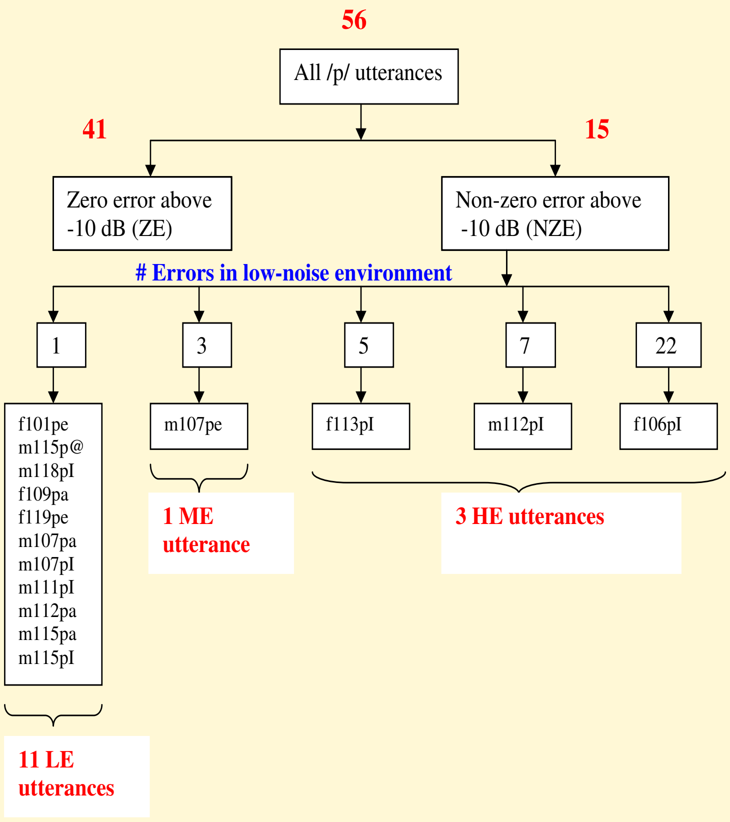
Across-consonant Token error & SD

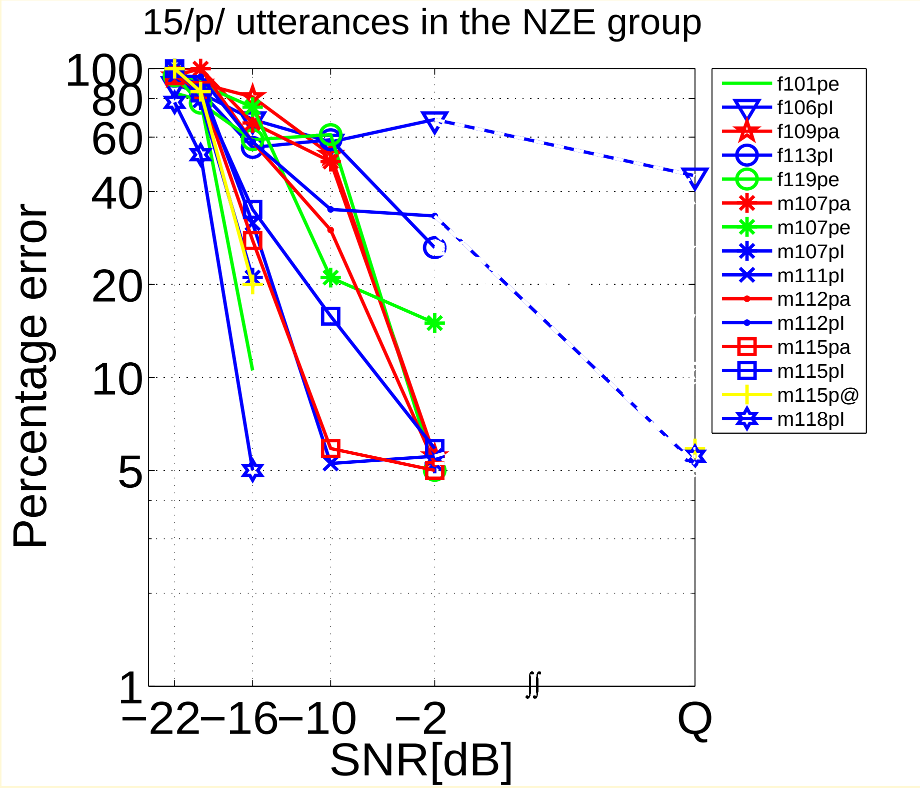


- AI(*SNR*) characterizes the average consonant error ($P_e = e_{\min}^{AI}$)
- AI ignores the huge *across-consonant* Standard Deviation (SD)
- as well as the huge *within-consonant* SD Singh-Allen 2012

Within-consonant Error /p/ Singh-Allen 2012

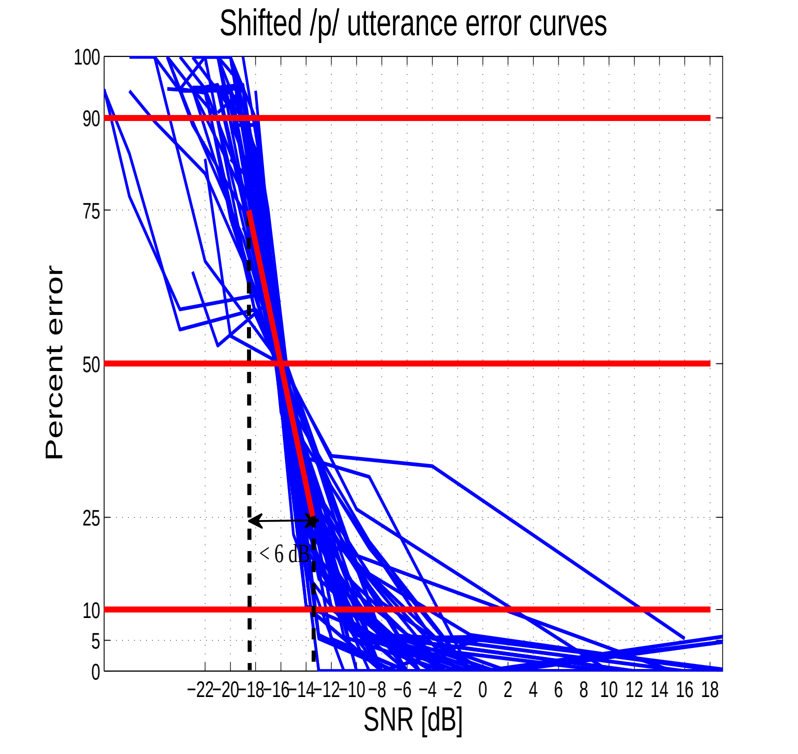
- 56 /p/+/o,e,ɪ/ CV tokens: SNR > -10 dB SNR
- Bimodal error distribution:
  - 41/56: Zero error (ZE); $N_{trials} = 38$, $N_{subj}=25$
  - 15/56: Non-zero error (NZE); 11 ≈ ZE (error: 1/38)



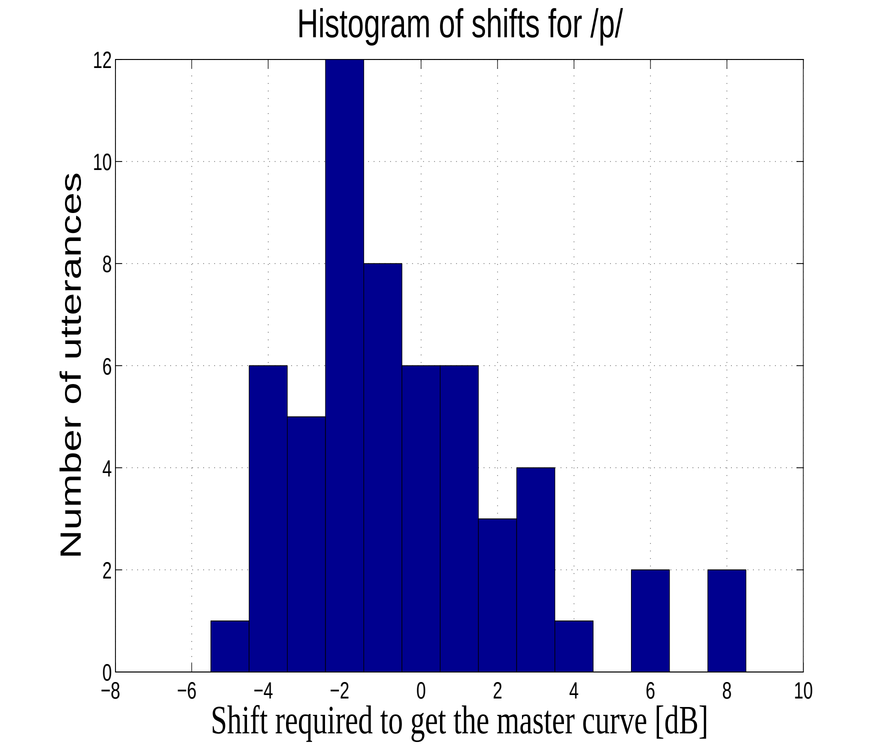


Within-consonant error $P_e(SNR - SNR^*_{50})$ for /p/

- Error vs. SNR shifted to 50% threshold SNR^*_{50} (LEFT)
- Histogram of 50% error thresholds (RIGHT)
  - Sharp transition ⇒ Binary Plosive identification!



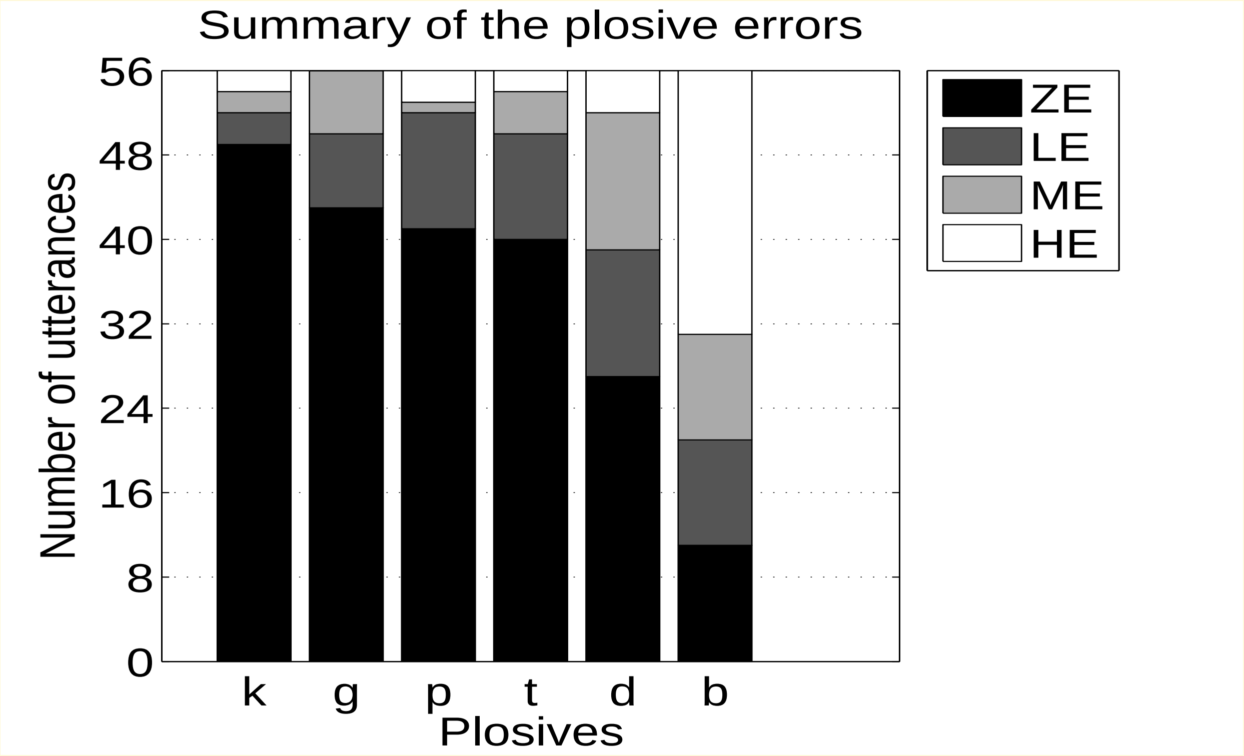
(a) $P_e(SNR - SNR^*_{50})$



(b) Distribution of SNR^*_{50}

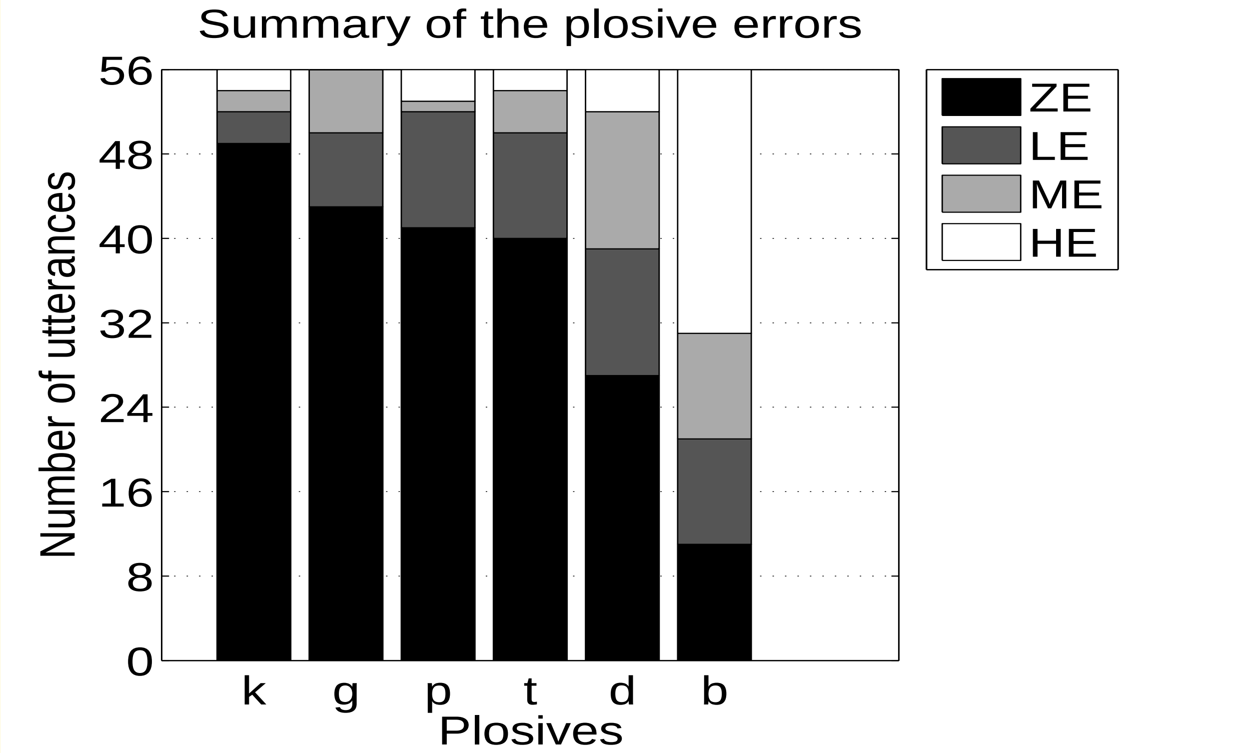
Error summary for Stops Singh-Allen 2012

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



Error summary for Stops Singh-Allen 2012

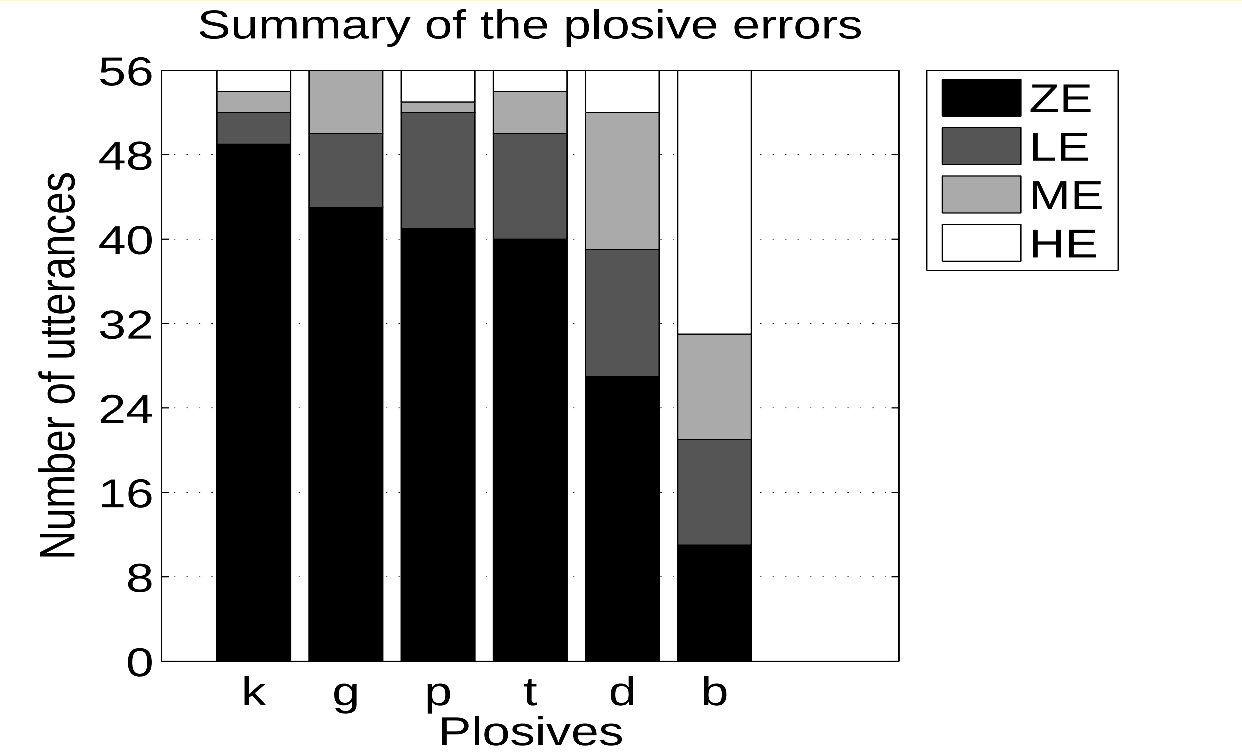
- Most stops have zero error (ZE+LE) above -10 dB SNR



- Bimodal error distribution for $\geq$ -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



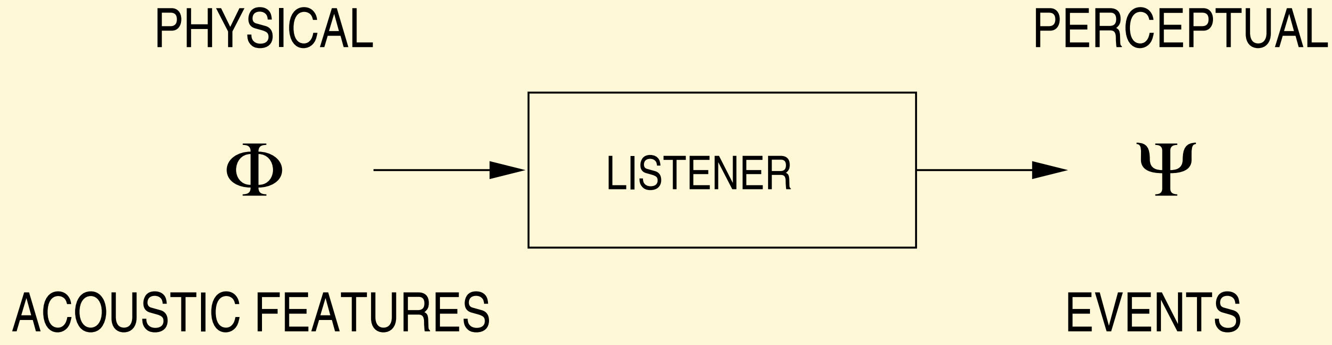
- Bimodal error distribution for $\geq$ -2 dB SNR
- While speech is highly variable, NH listeners are not
- The AI is an average measure
  - Huge 'across– & 'within–consonant' SD (85% of the variance)
  - SNR depends only on binary threshold distributions

3. Phone Recognition Models

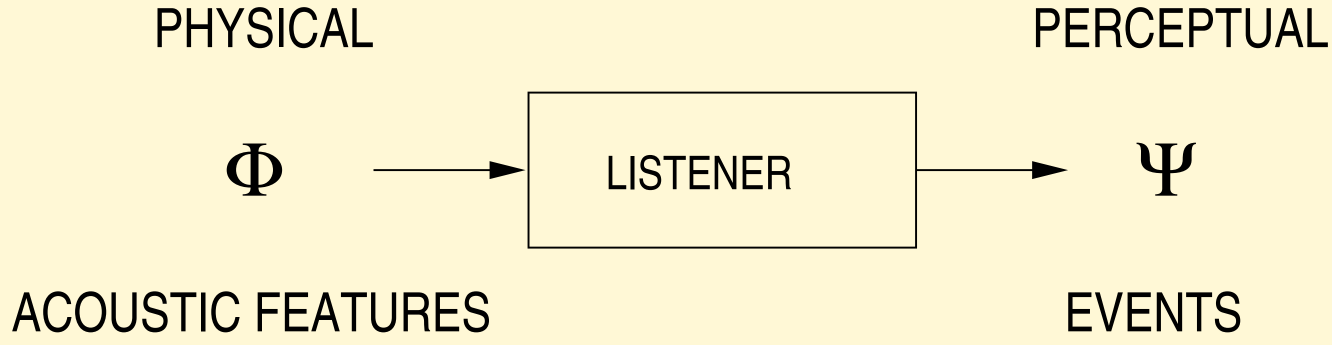
The Role of Models

- We need rigorous procedures for analyzing speech elements

- We need rigorous procedures for analyzing speech elements
  - Basic model of acoustic vs. perceptual cue identification



- We need rigorous procedures for analyzing speech elements
  - Basic model of acoustic vs. perceptual cue identification

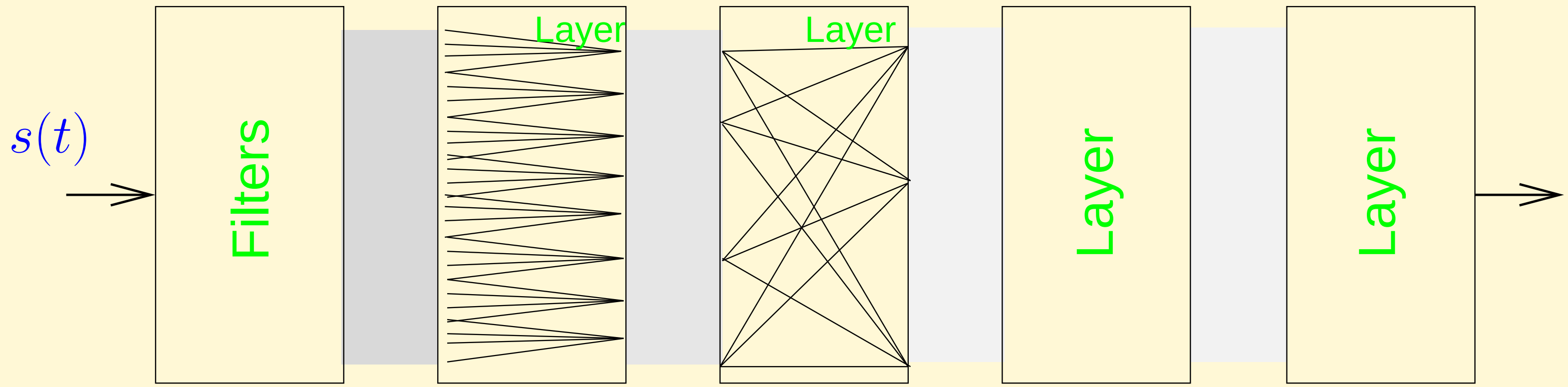


- We define two basic measures:
  - Physical Input: AI-Gram
  - Perceptual Output: Confusion matrix

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



Measure: AI_k e_k s $S = s^3$ W

Formula: $\propto snr_k$ dB $= 0.82^{AI_k}$ $= 1 - e_1 e_2 \ldots e_{20}$

Analog objects ??? Discrete objects

Φ "Front-end" Ψ "Back-end"

Human listeners as a Shannon Channel

- The Channel capacity theorem gives the zero error SNR bound:

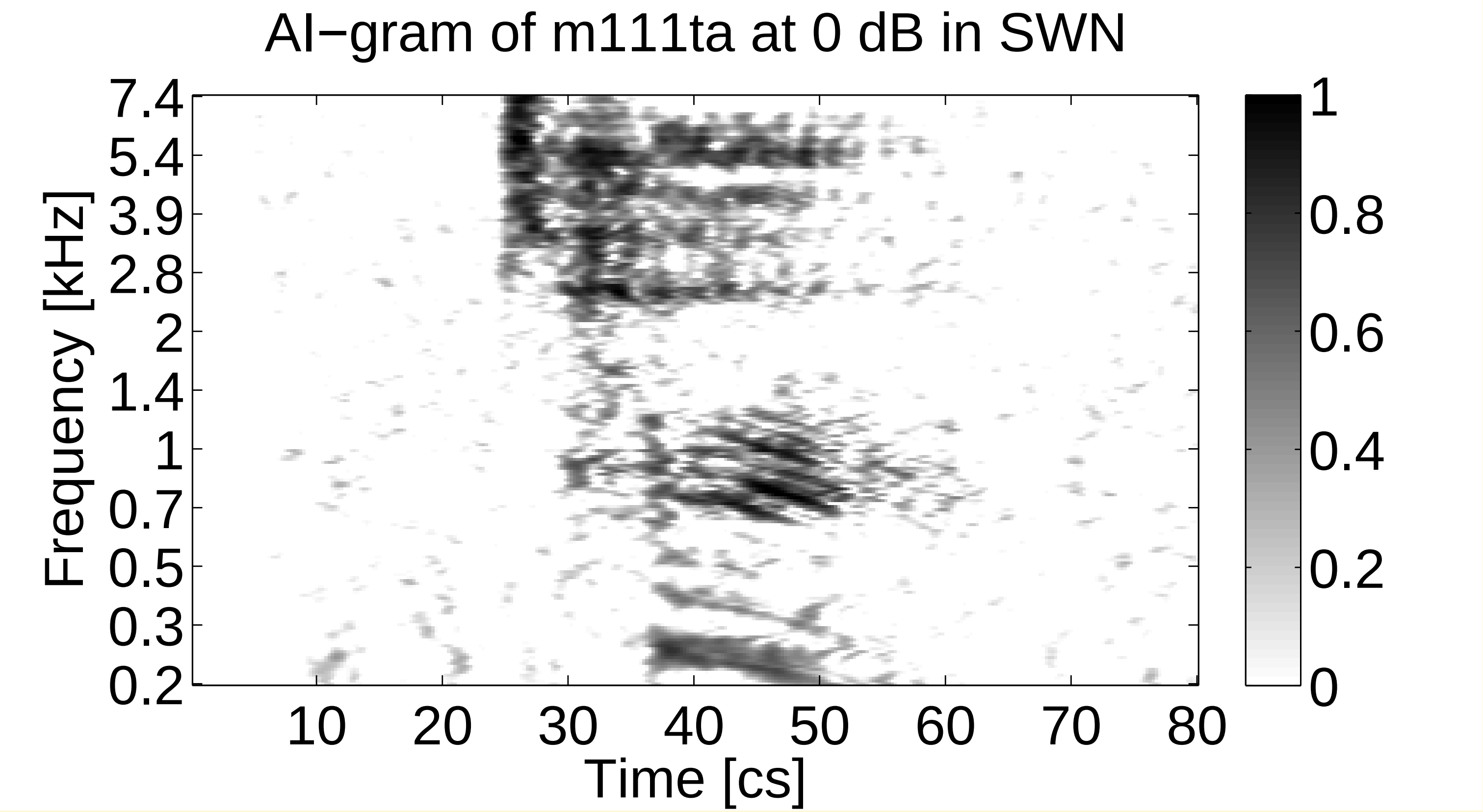
$$\mathcal{C}(SNR) \equiv \int \log_2 \left(1 + snr^2(f)\right) df \approx AI(SNR) \quad (1)$$

Human listeners as a Shannon Channel

- The Channel capacity theorem gives the zero error SNR bound:

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

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



Human listeners as a Shannon Channel

- The Channel capacity theorem gives the zero error SNR bound:

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

  - 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

Human listeners as a Shannon Channel

- The Channel capacity theorem gives the zero error SNR bound:

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

  - 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?

Human listeners as a Shannon Channel

- The Channel capacity theorem gives the zero error SNR bound:

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

  - 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:

Human listeners as a Shannon Channel

- The Channel capacity theorem gives the zero error SNR bound:

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

  - 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

Human listeners as a Shannon Channel

- The Channel capacity theorem gives the zero error SNR bound:

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

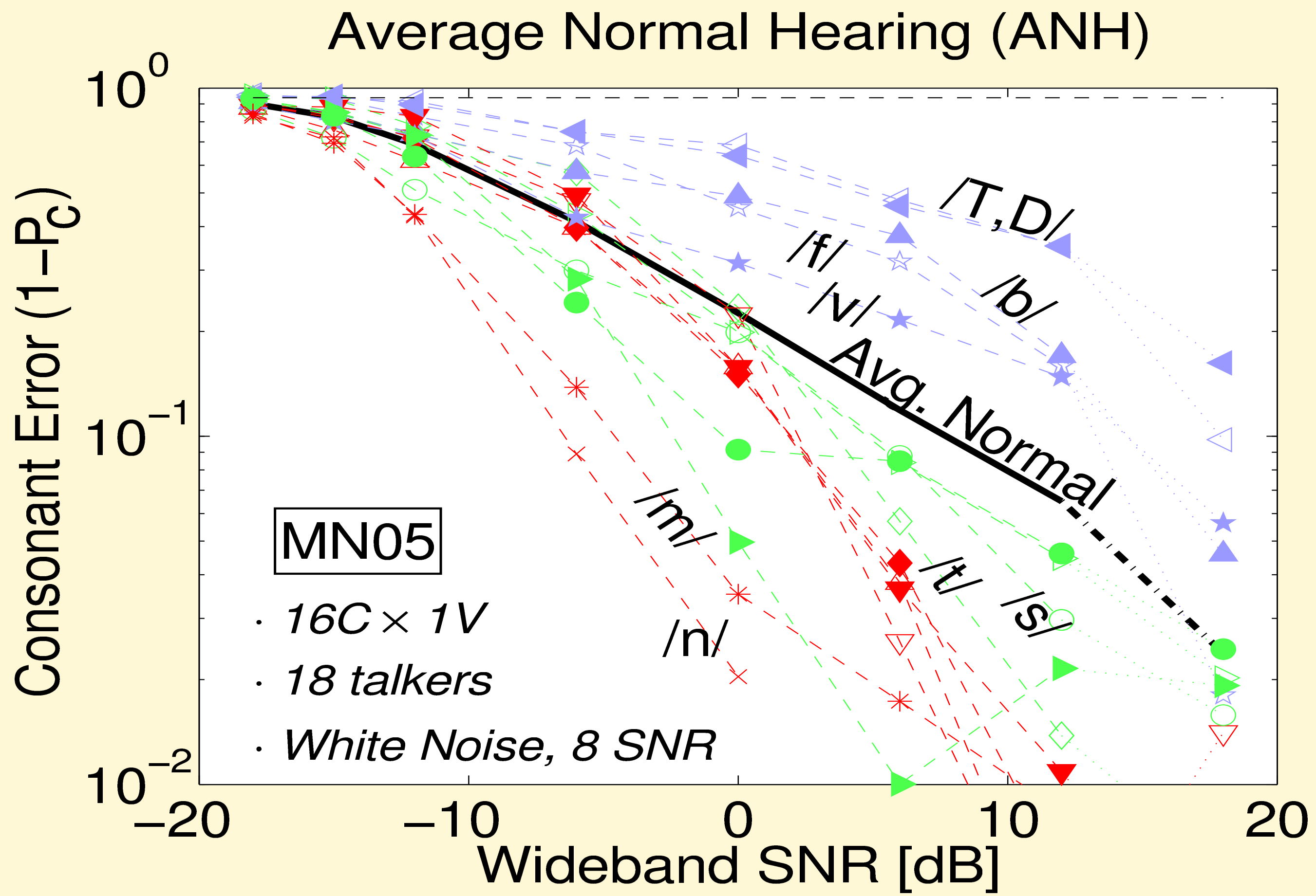
  - For a Maximum Entropy (MaxEnt) speech source, the maximum information rate is determined by the AI(SNR)
  - The AI-gram is a closely related measure

- 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

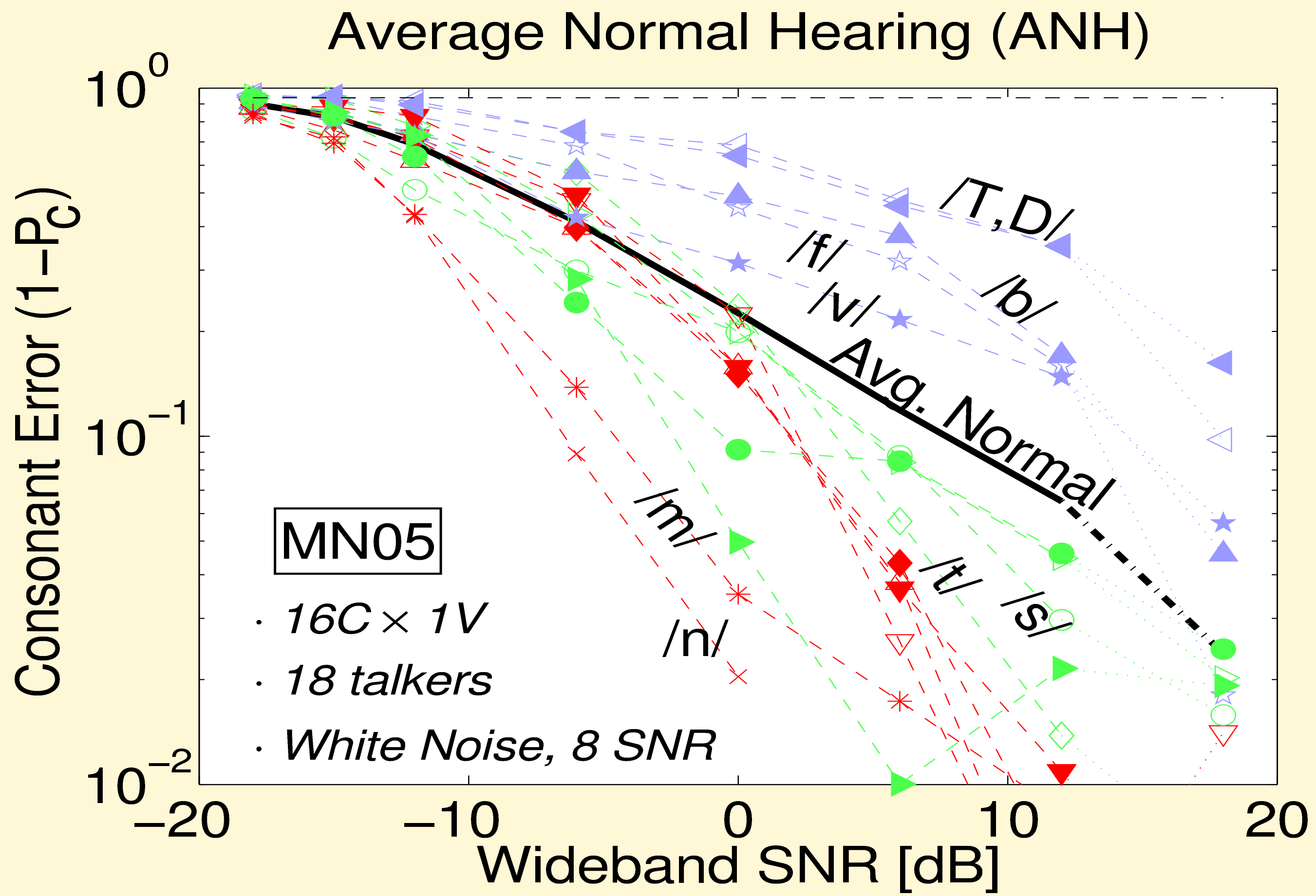
- The AI predicts $P_e(SNR)$, but with a huge SD ($\sigma_{AI}(SNR)$)

- 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 ($\sigma_{AI}(SNR)$)



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

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

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

STIMULUS \ RESPONSE	p	t	k	f	θ	s	ʃ	b	d	g	v	ð	z	ʒ	m	n
p	80	43	64	17	14	6	2	1	1		1	1			2	
t	71	84	55	5	9	3	8	1				1	2		2	3
k	66	76	107	12	8	9	4					1			1	
f	18	12	9	175	48	11	1	7	2	1	2	2				
θ	19	17	16	104	64	32	7	5	4	5	6	4	5			
s	8	5	4	23	39	107	45	4	2	3	1	1	3	2		1
ʃ	1	6	3	4	6	29	195		3							1
b	1			5	4	4		136	10	9	47	16	6	1	5	4
d							8	5	80	45	11	20	20	26	1	
g					2			3	63	66	3	19	37	56		3
v				2		2		48	5	5	145	45	12		4	
ð					6			31	6	17	86	58	21	5	6	4
z					1	1	1	7	20	27	16	28	94	44		1
ʒ								1	26	18	3	8	45	129		2
m	1							4			4	1	3		177	46
n					4			1	5	2		7	1	6	47	163

STIMULUS

UNVOICED VOICED NASAL

RESPONSE

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

STIMULUS \ RESPONSE	*p*	*t*	*k*	*f*	*θ*	*s*	*ʃ*	*b*	*d*	*g*	*v*	*ð*	*z*	*ʒ*	*m*	*n*
p	80	43	64	17	14	6	2	1	1		1	1			2	
t	71	84	55	5	9	3	8	1				1	2		2	3
k	66	76	107	12	8	9	4					1			1	
f	18	12	9	175	48	11	1	7	2	1	2	2				
θ	19	17	16	104	64	32	7	5	4	5	6	4	5			
s	8	5	4	23	39	107	45	4	2	3	1	1	3	2		1
ʃ	1	6	3	4	6	29	195		3							1
b	1			5	4	4		136	10	9	47	16	6	1	5	4
d							8	5	80	45	11	20	20	26	1	
g					2			3	63	66	3	19	37	56		3
v				2		2		48	5	5	145	45	12		4	
ð					6			31	6	17	86	58	21	5	6	4
z					1	1	1	7	20	27	16	28	94	44		1
ʒ								1	26	18	3	8	45	129		2
m	1							4			4	1	3		177	46
n					4			1	5	2		7	1	6	47	163

STIMULUS

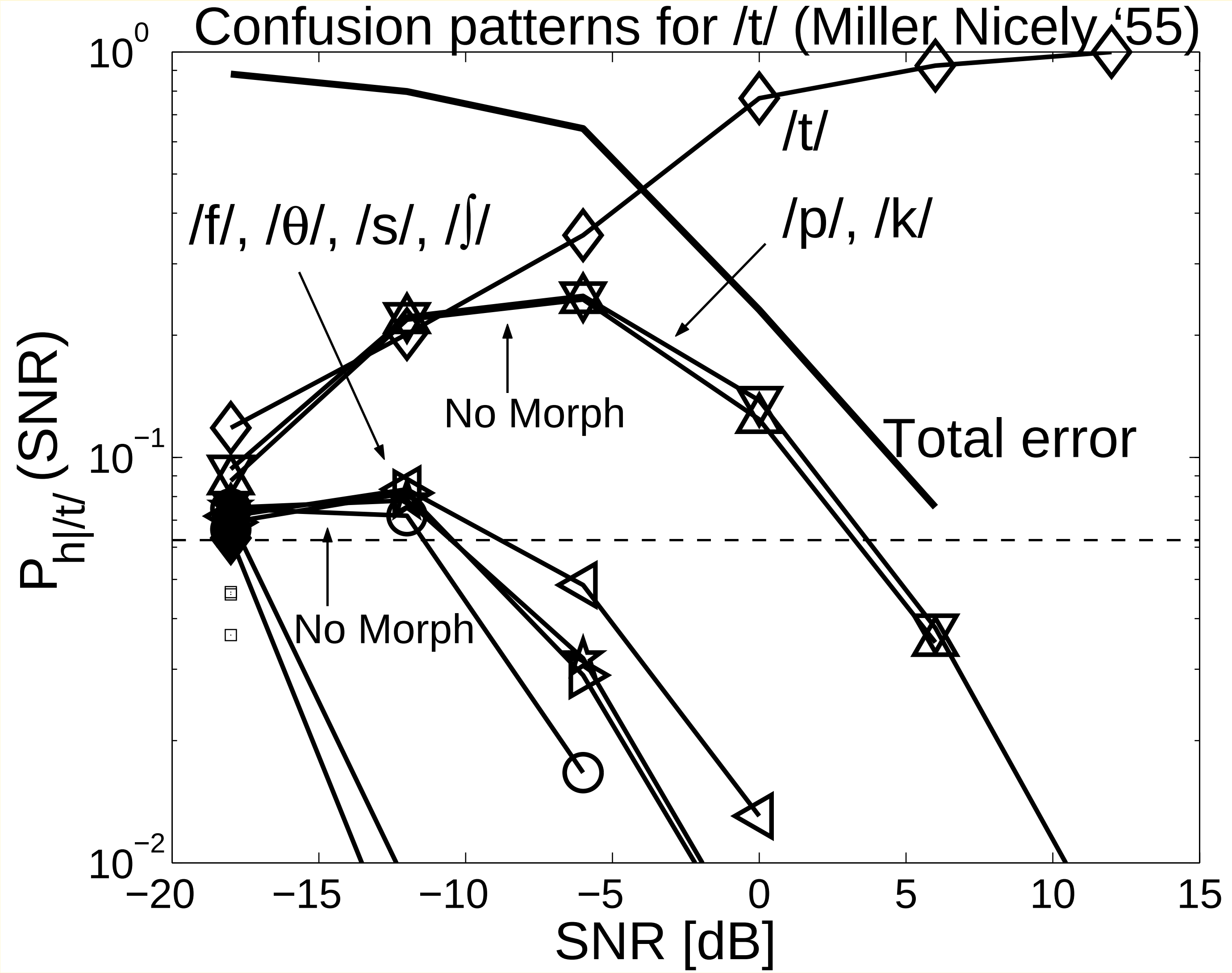
UNVOICED VOICED NASAL

RESPONSE

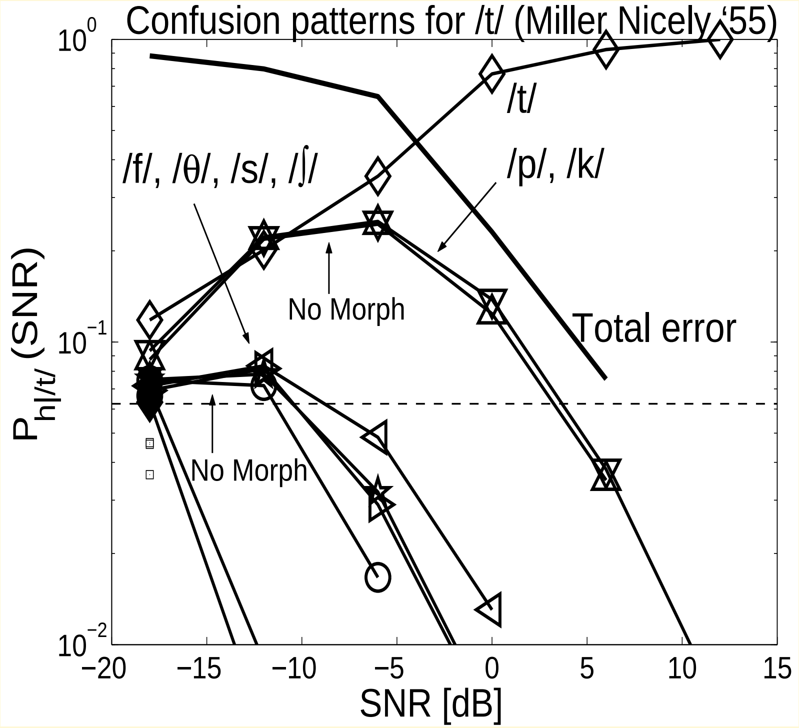
- Confusion groups $\equiv$ *inhomogeneous confusions*

Row of confusion matrix (CM) $P_{h|/t/}$

- This *confusion pattern* characterizes the /t/ row vs SNR

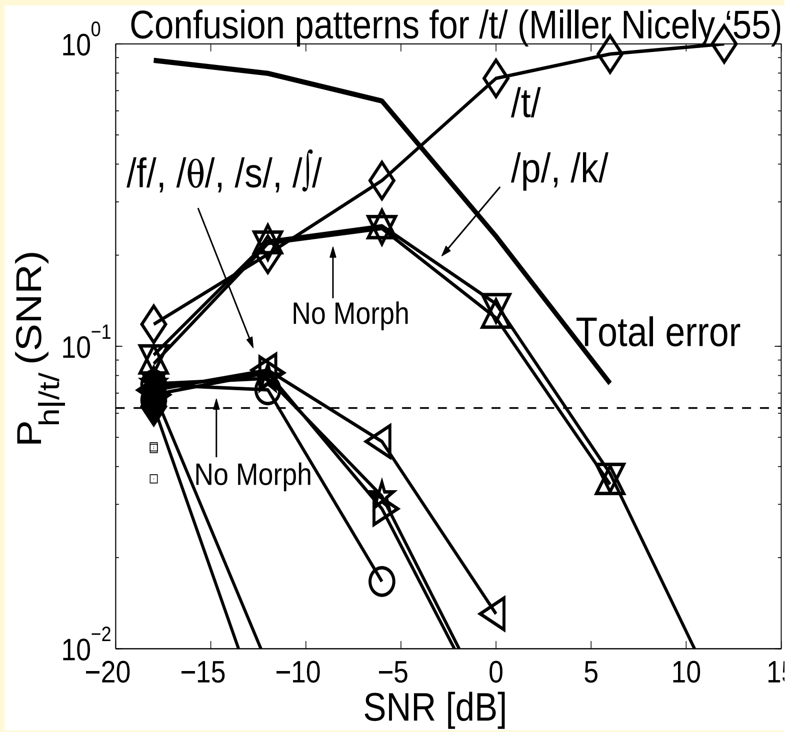


- The SIN_t of averaging within-consonants (i.e., tokens):
  - Token confusions are strongly heterogeneous!
  - Averaging obscures per-token confusions

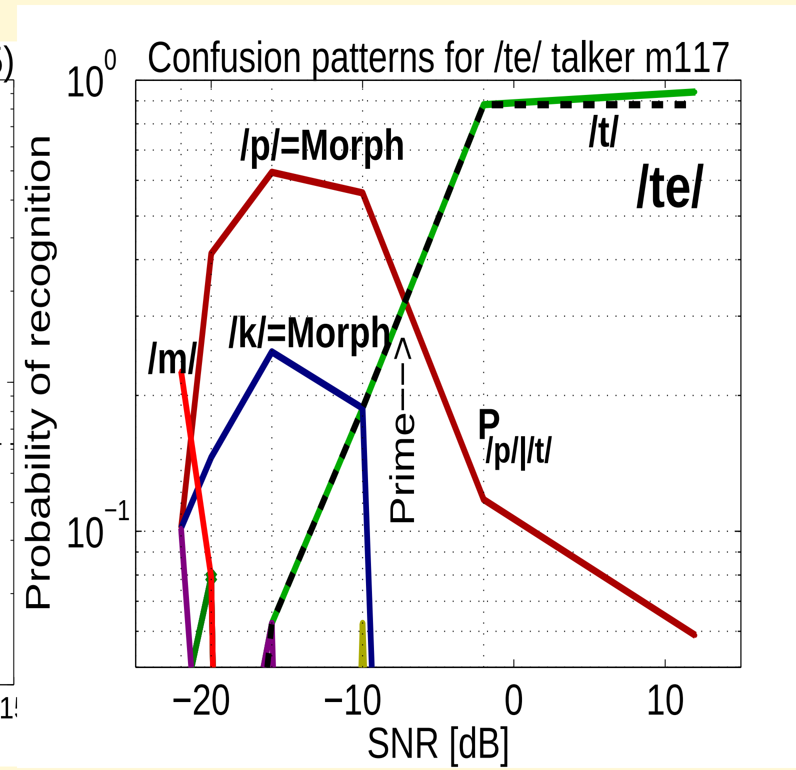


(a) Average over all /t/s.

- The SIN$_t$ of averaging within-consonants (i.e., tokens):
  - Token confusions are strongly heterogeneous!
  - Averaging obscures per-token confusions



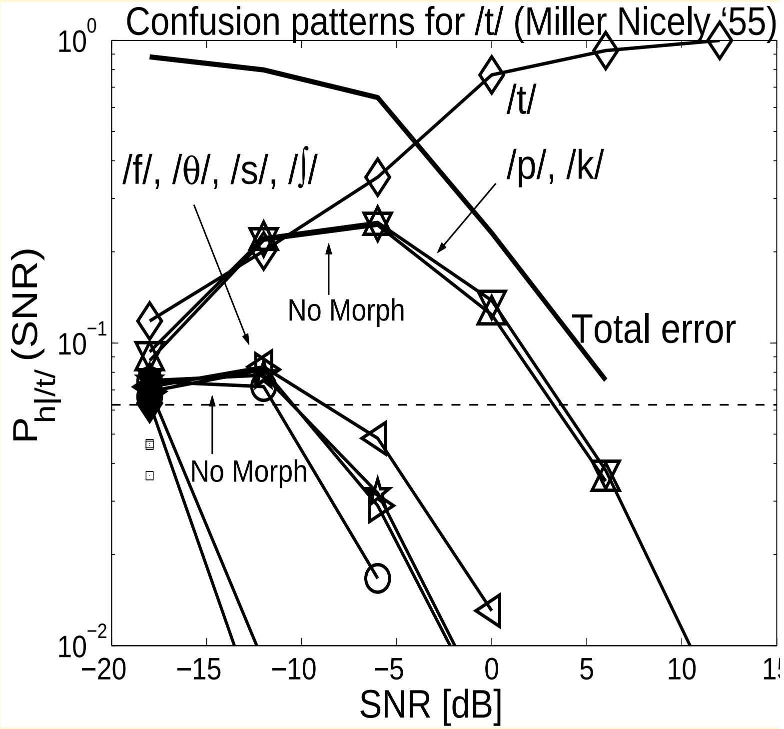
(a) Average over all /t/s.



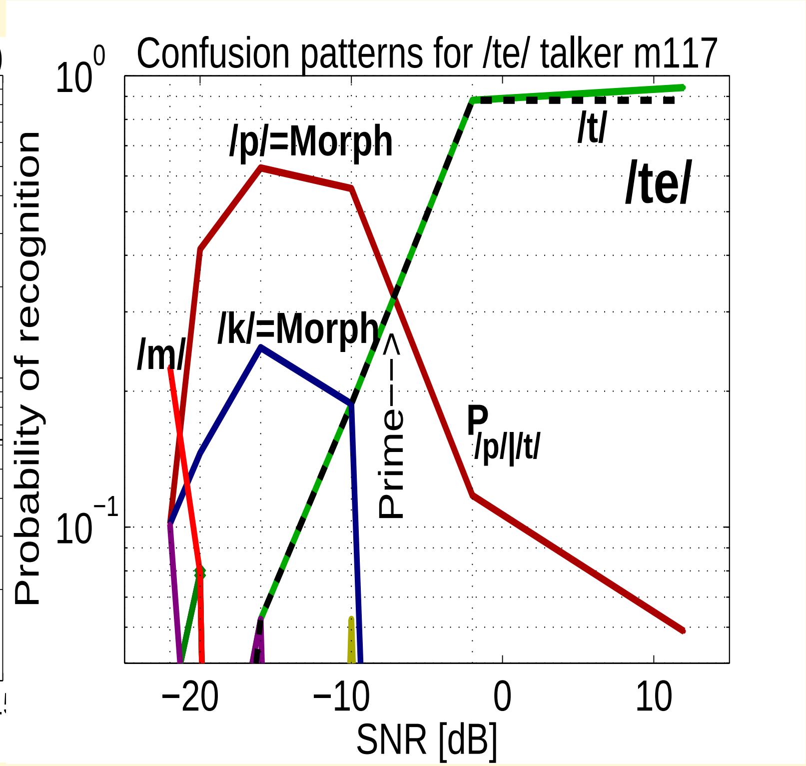
(b) Talker m117 /te/ $P_{h|/ta/}(SNR)$

Row of confusion matrix (CM) $P_{h|/t/}$

- The 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/ $P_{h|/ta/}(SNR)$

- 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, ʧ, s, h, f/ and /z, ʒ, v, ð/)
  - -With vowels /o, e, ı/
    - $\approx$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, ʧ, 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: 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, ʧ, s, h, f/ and /z, ʒ, v, ð/)
  - -With vowels /o, e, ı/
    - ≈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: Cochlear models & signal processing
  - AIgram Régnier & Allen 2008; Li & Allen 2009,10,11

Methods: 3^d Deep Search (3DDS)

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

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

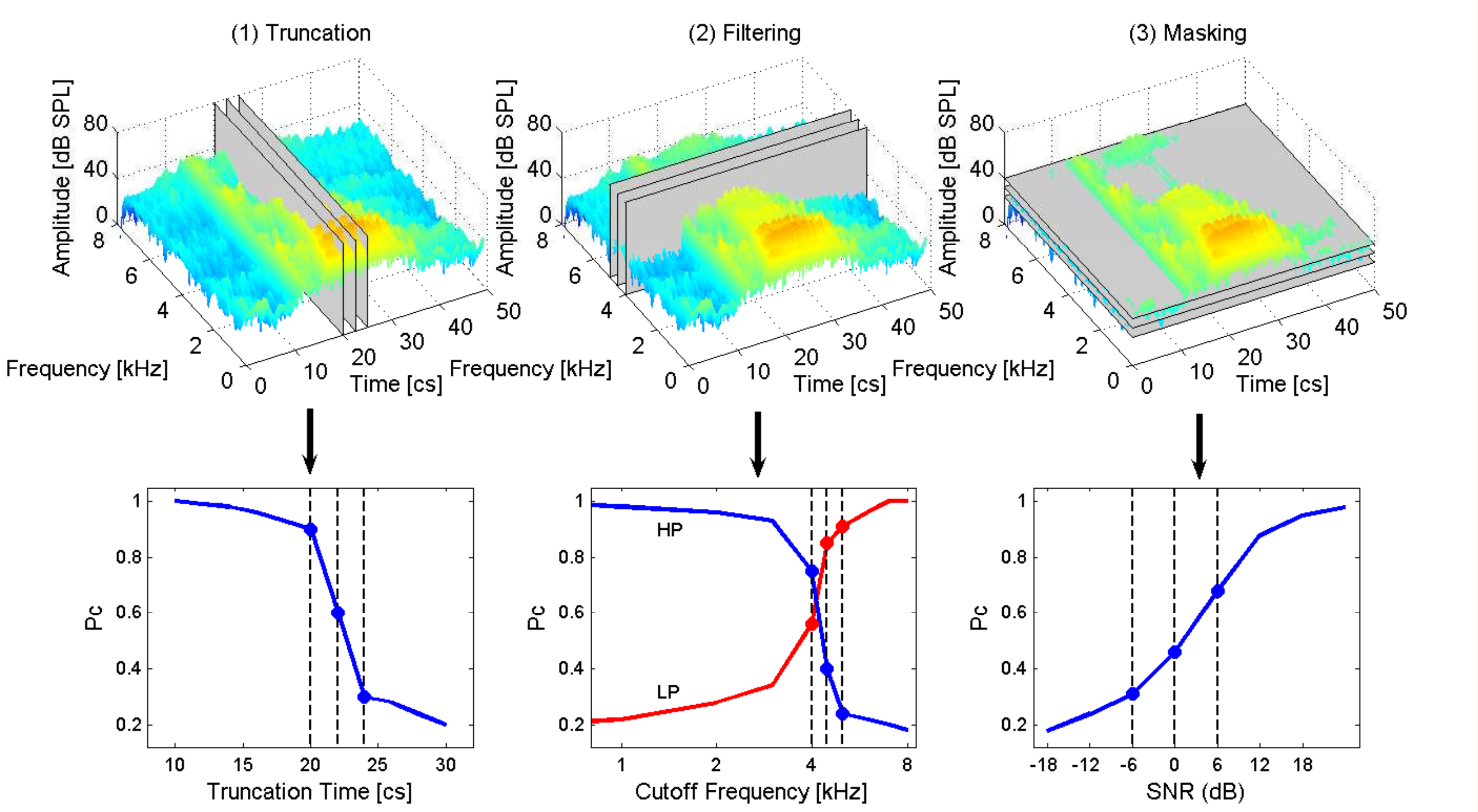
- 3^d Deep-Search (3^d-DS) via truncation (triangulate):
  - Time truncation Furui 1986
  - Frequency truncation (High/Low-pass filtering)

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)

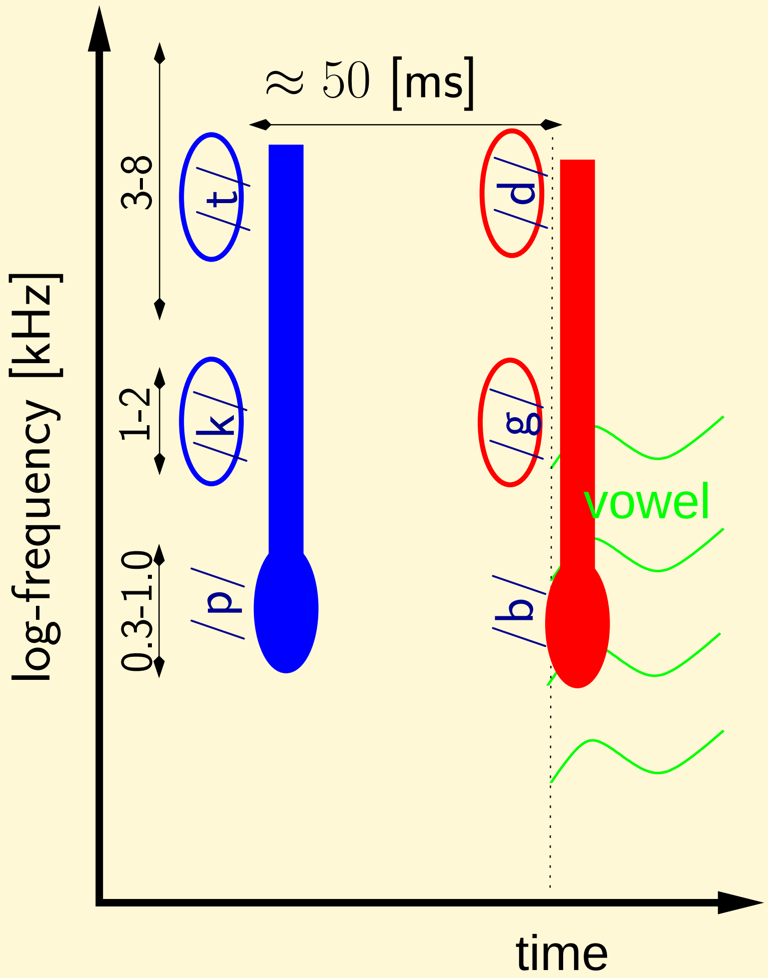
Methods: 3^d Deep Search (3DDS)

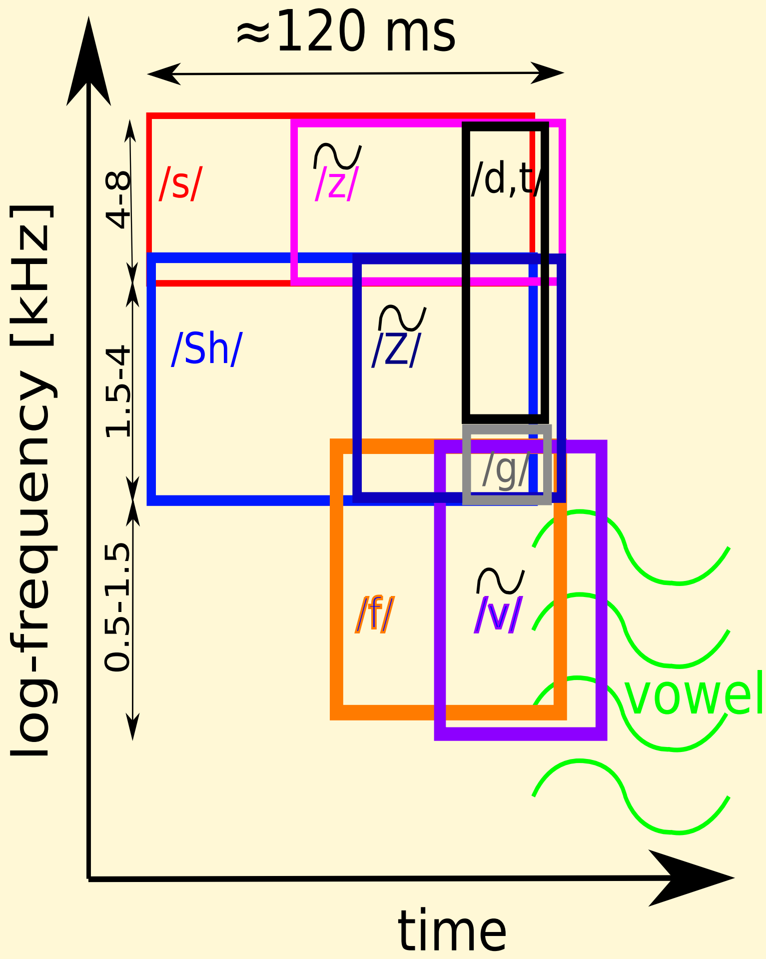
- 3^d Deep-Search (3^d-DS) via truncation (triangulate):
  - Time truncation Furui 1986
  - Frequency truncation (High/Low-pass filtering)
  - Intensity truncation (i.e., masking)



- Time-frequency structure of plosives and fricatives

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





Auditory & Cochlear Modeling 1920-2000

- 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, . . .)

Auditory & Cochlear Modeling 1920-2000

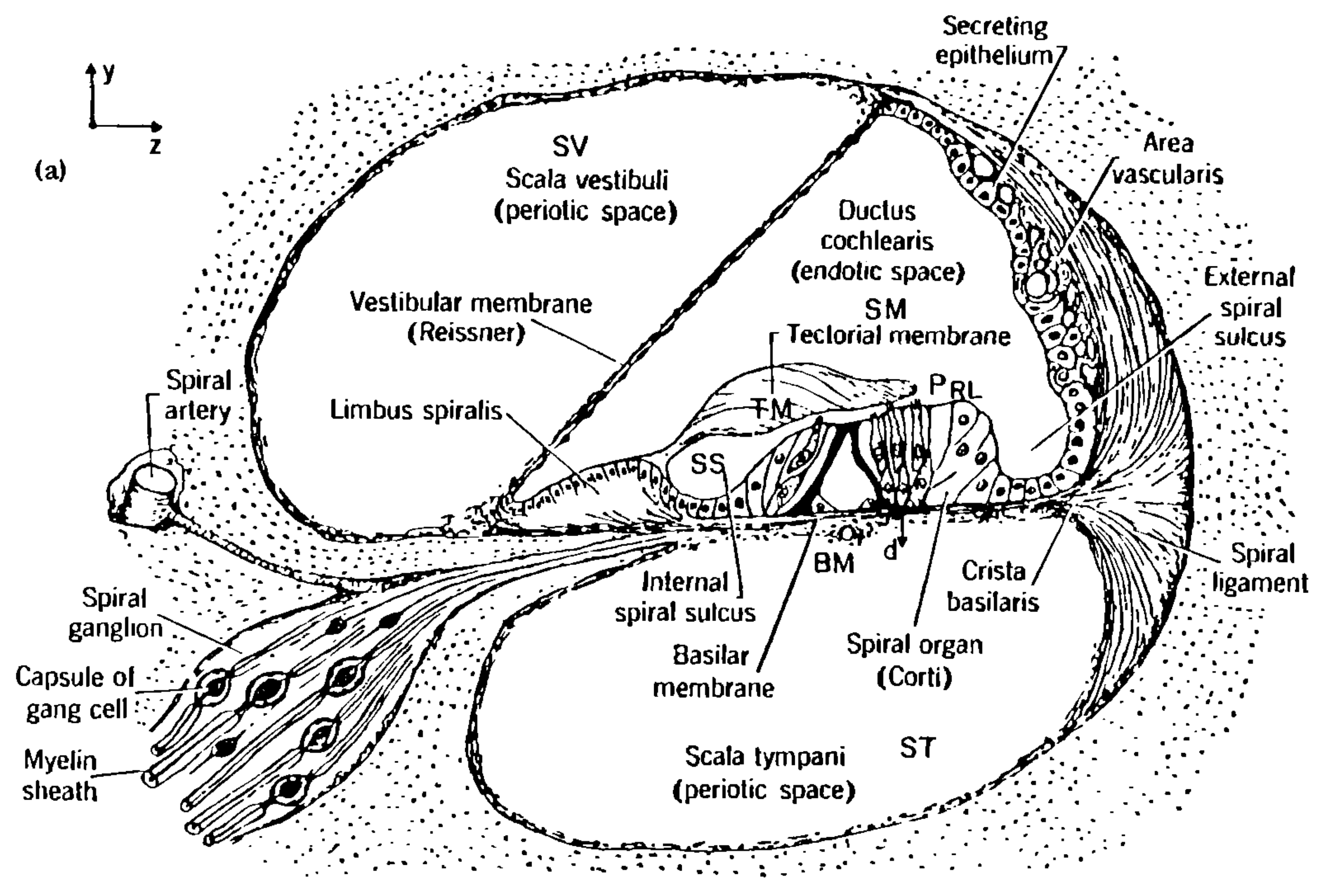
- 1910-1980: Bell Labs (long history)
  - Fletcher 1914; Wegel & Lane 1924; Flanagan; Hall; Allen
- 1960-2010: MIT + Harvard HSBT
  - Eaton Peabody (Kiang, Siebert, Liberman, Guinan, Shera, . . .)
- Netherlands, England
  - deBoer, Duifhuis, Evans, . . .
- Australia (B. Johnstone, . . .)

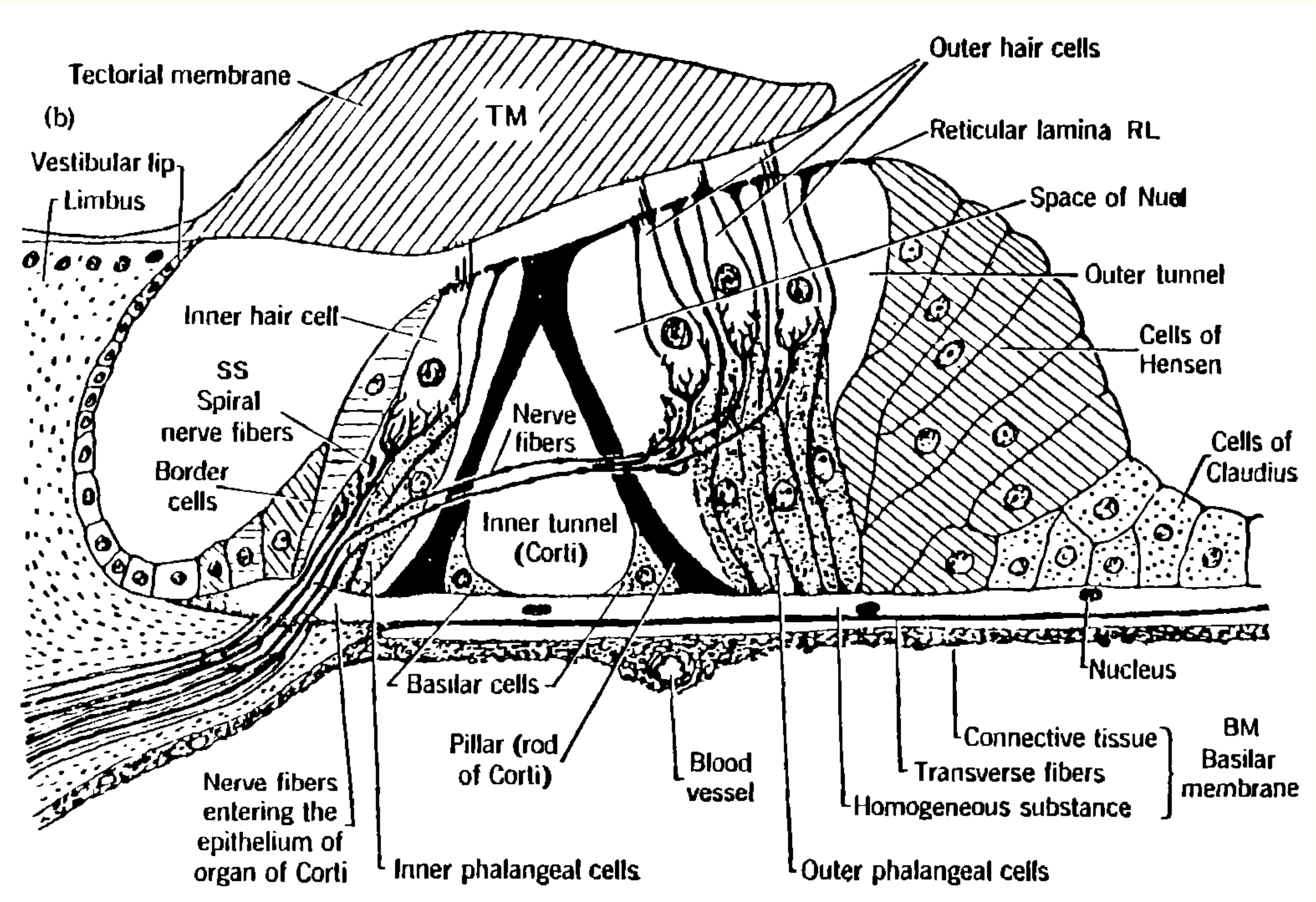
Auditory & Cochlear Modeling 1920-2000

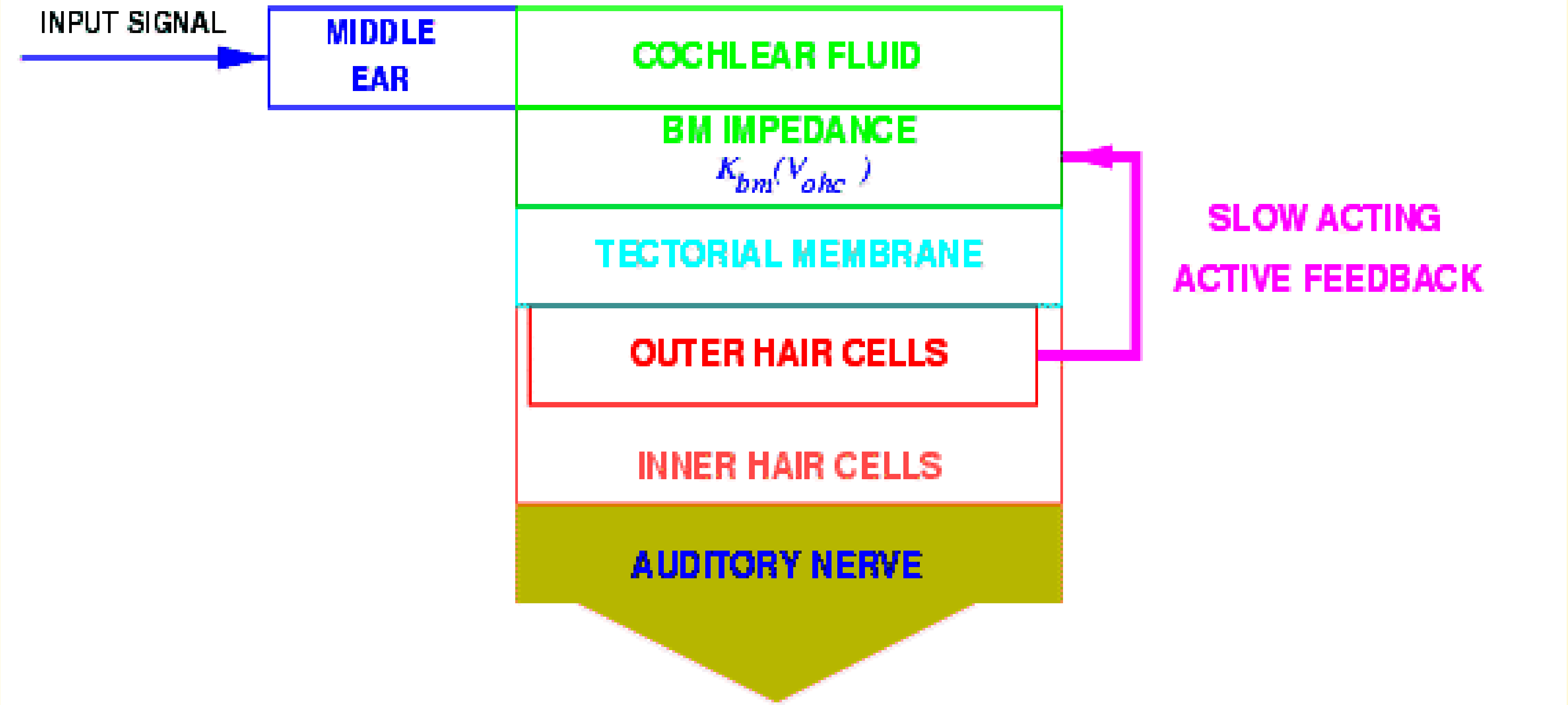
- 1910-1980: Bell Labs (long history)
  - Fletcher 1914; Wegel & Lane 1924; Flanagan; Hall; Allen
- 1960-2010: MIT + Harvard HSBT
  - Eaton Peabody (Kiang, Siebert, Liberman, Guinan, Shera, . . .)
- Netherlands, England
  - deBoer, Duifhuis, Evans, . . .
- Australia (B. Johnstone, . . .)
- 1980-2011: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

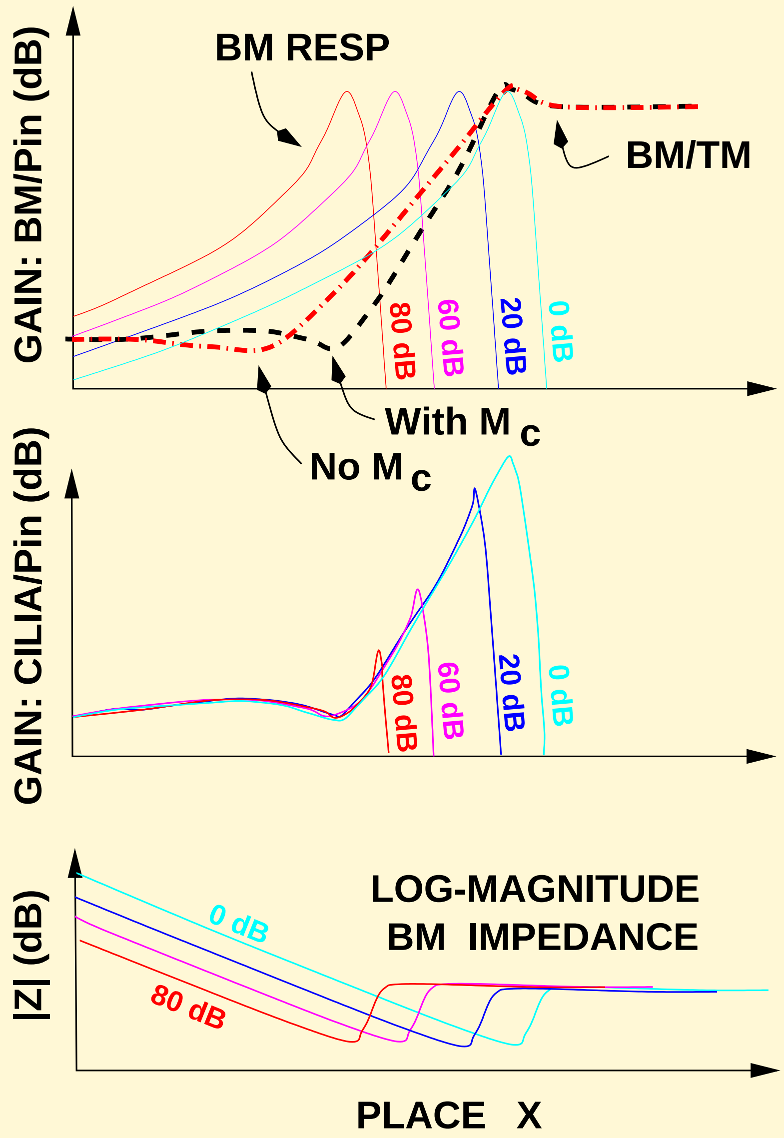
Auditory & Cochlear Modeling 1920-2000

- 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

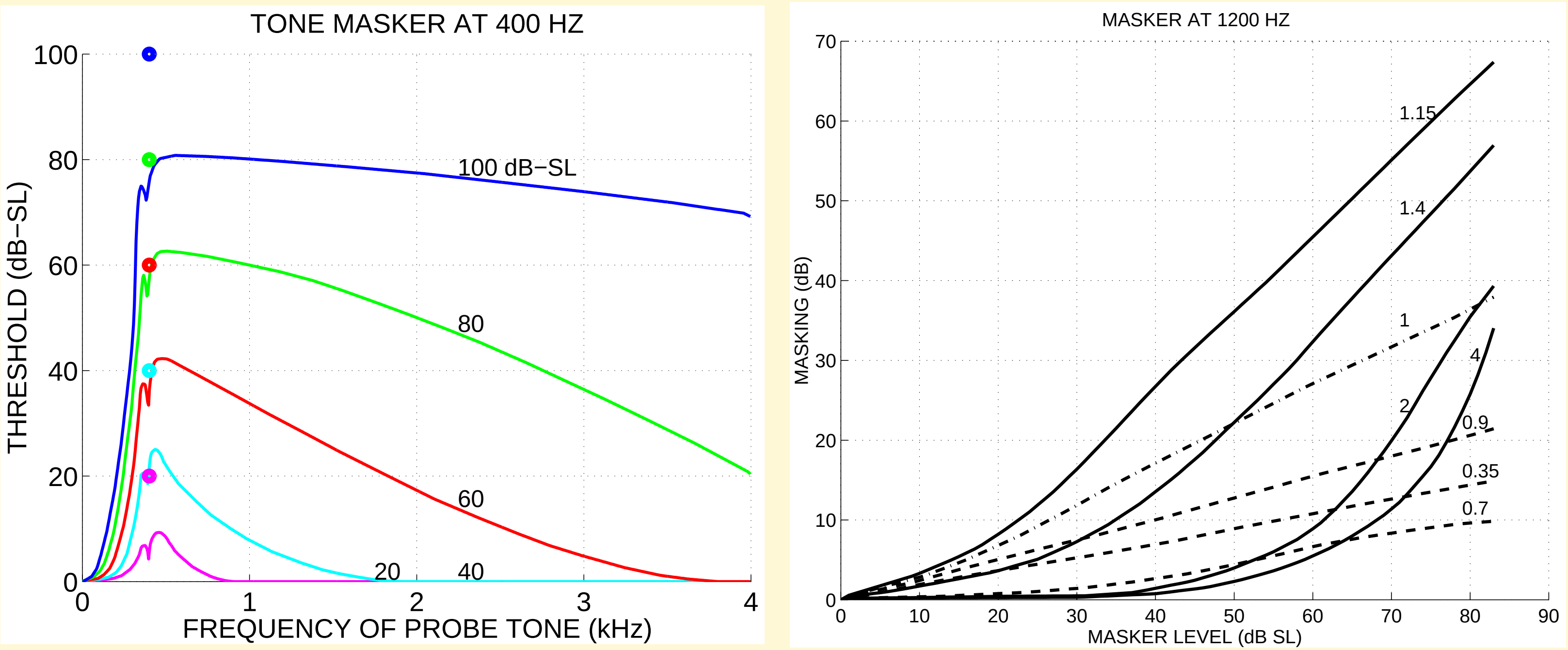

y
z
(a)
SV
Scala vestibuli
(periotic space)
Secreting
epithelium
Area
vascularis
Ductus
cochlearis
(endotic space)
SM
External
spiral
sulcus
Vestibular membrane
(Reissner)
Tectorial membrane
TM
P
RL
Spiral
artery
Limbus spiralis
SS
BM
d
Crista
basilaris
Spiral
ligament
Internal
spiral sulcus
Spiral
ganglion
Capsule of
gang cell
Basilar
membrane
Spiral organ
(Corti)
Myelin
sheath
Scala tympani
(periotic space)
ST


Tectorial membrane
(b)
TM
Vestibular lip
Limbus
Outer hair cells
Reticular lamina RL
Space of Nuel
Outer tunnel
Cells of Hensen
Cells of Claudius
Inner hair cell
SS
Spiral nerve fibers
Border cells
Nerve fibers
Inner tunnel (Corti)
Basilar cells
Pillar (rod of Corti)
Blood vessel
Nucleus
Connective tissue
Transverse fibers
Homogeneous substance
BM Basilar membrane
Nerve fibers entering the epithelium of organ of Corti
Inner phalangeal cells
Outer phalangeal cells


INPUT SIGNAL
MIDDLE EAR
COCHLEAR FLUID
BM IMPEDANCE
$K_{bm}(V_{ohc})$
TECTORIAL MEMBRANE
OUTER HAIR CELLS
INNER HAIR CELLS
AUDITORY NERVE
SLOW ACTING
ACTIVE FEEDBACK

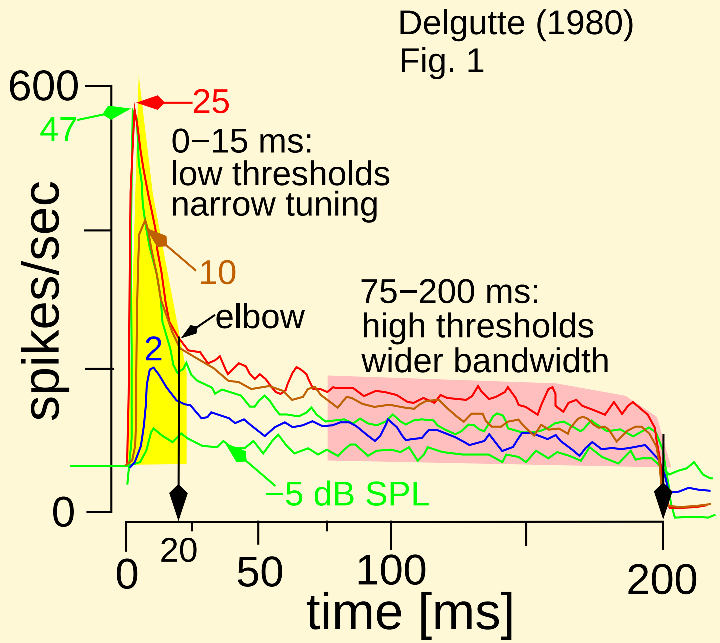

BM RESP
BM/TM
GAIN: BM/Pin (dB)
80 dB
60 dB
20 dB
0 dB
With M c
No M c
GAIN: CILIA/Pin (dB)
80 dB
60 dB
20 dB
0 dB
|Z| (dB)
0 dB
80 dB
LOG-MAGNITUDE
BM IMPEDANCE
PLACE X

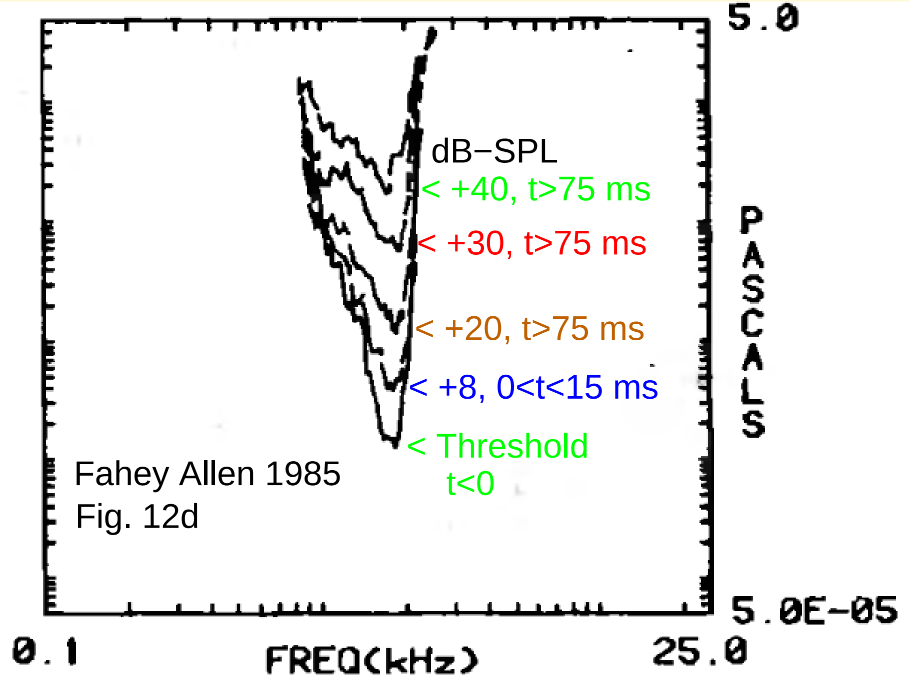
Upward spread of masking



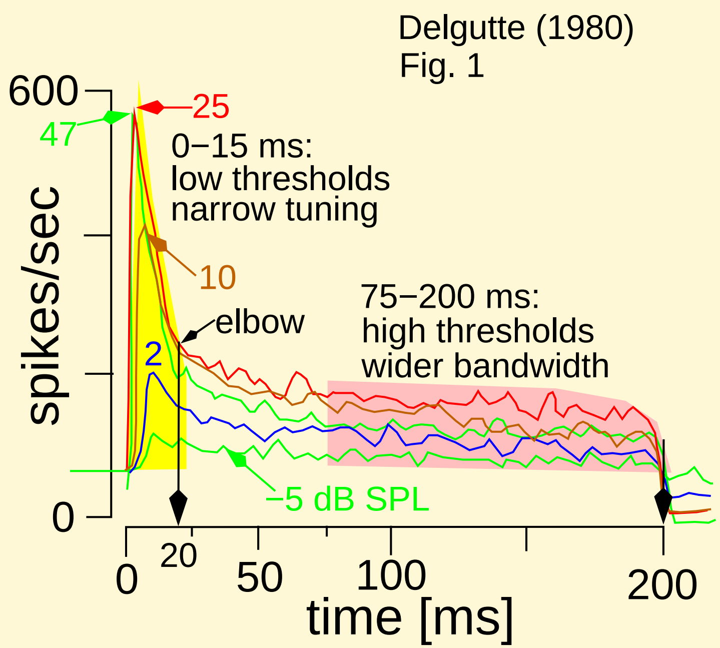
- This effect leads to forward masking
- Forward Masking is a very large effect lasting for up to 200 ms

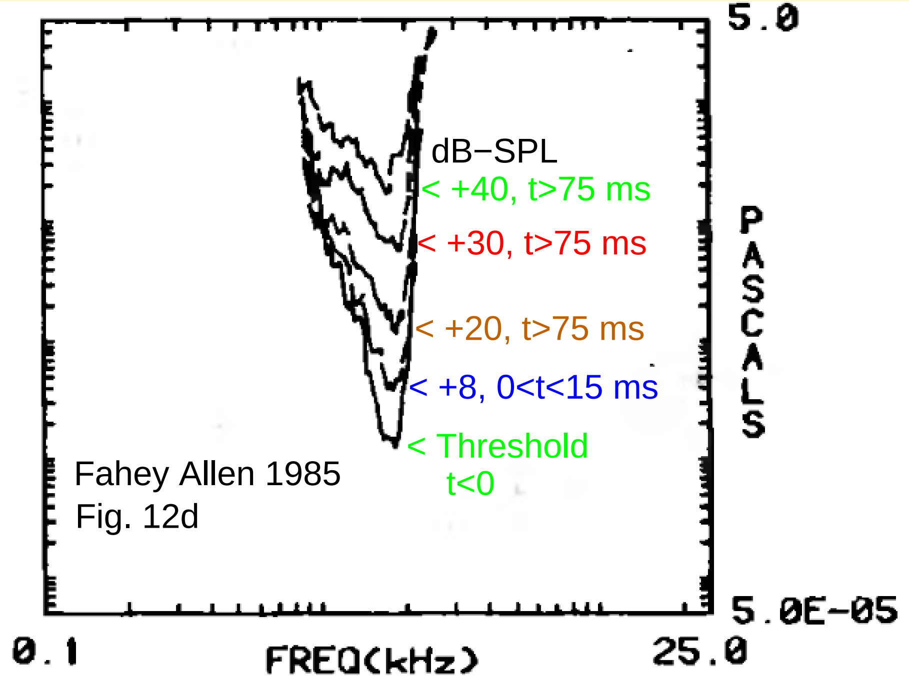
- Onset transients enhance the auditory nerve response, to 2 [cs]





- 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

1. Intro + Objectives 3 mins $\Sigma 3$
   - Research objectives 5 mins $\Sigma 8$
2. Historical overview 20 mins $\Sigma 28$
   - AG Bell 1860, Rayleigh 1910, Fletcher 2021, Shannon 1948
   - Speech-feature studies (1950-1990; $>$1991)
3. Phone Recognition Models 8 mins $\Sigma 36$
   - Channel capacity and the Articulation Index
   - Speech Psychophysics; AIgram/3DDS (cues); Primes and Morphs;
   - Classification models (e.g., DFs)
4. Cochlear Mechanics 15 mins $\Sigma 51$
   - CBands, NL, Masking, Role re Speech perception; HI ears
5. Summary + Conclusions + Questions 3+3+4 mins $\Sigma 76$

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

Conclusion I

- 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, ʧ, 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, ʧ, s, h, f/, /z, ʒ, v, ð/) + vowels /o, e, ɪ/
  3. Data on discriminating consonants in noise, NH listeners use
     - Plosives: *Burst + timing to Voicing*
     - Fricatives: *Low-frequency edge + duration +* F_0 *modulation*

5. STFT to manipulate speech:
   - Morph consonants (e.g., /k/ to /t/ to /p/)
   - Intelligibility: Modify SNR_{90}

Conclusion II

- 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:

Conclusion II

- 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

Conclusion II

- 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
   - $P_e(\text{SNR}) = e_{\min}^{\mathit{SNR}}$ due to SNR^*_{50} distribution

Conclusion II

- 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
   - $P_e(\mathrm{SNR}) = e_{\min}^{\mathit{SNR}}$ due to SNR_{50}^{*} distribution
3. Explored the natural existence of conflicting cues

Conclusion II

- 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
   - $P_e(\text{SNR}) = e_{\min}^{\mathit{SNR}}$ due to SNR^*_{50} distribution
3. Explored the natural existence of conflicting cues
   - This could impact ASR systems

Conclusion III

- Findings re HI ears:

1. HI ears have huge individual differences

Conclusion III

- Findings re HI ears:

1. HI ears have huge individual differences
   - Individual differences dominate HI results

Conclusion III

- 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 $P_e(SNR)$

Conclusion III

- 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 $>$c2016
     - 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 ...?