Cochlear nonlinearities and phoneme recognition

Finding the features in individual consonants.

Jont Allen

Feipeng Li

Univ. of IL, Beckman Inst., Urbana IL

Notable quotes

We need to know more about human speech processing and and natural speech variation
–Sadaoki Furui (ASRU 2009)

This is so true!
–Jont Allen

Question your assumptions:

- Elephant in the room:
  Human CV speech is *not* variable.
- CV speech is *not* redundant.
- Why we don't know anything about the topic?
  $ spent on basic speech research $\rightarrow$0,

Outline of talk

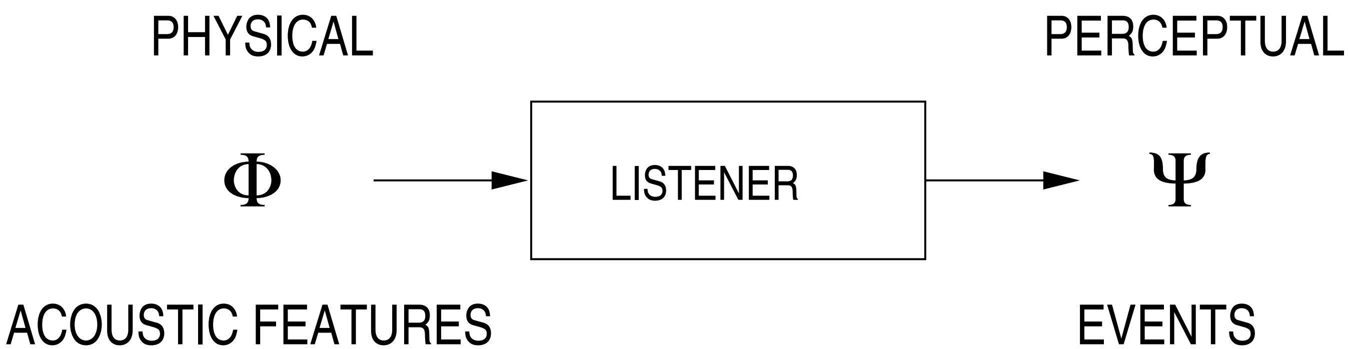
1. Intro + Objectives (5 mins)
   - The research goal is to
     - Identify the *elemental HSR events* in
     - Example consonants
2. Historical overview (5 mins $\Sigma 10$)
   - Rayleigh (1910) to Shannon (1948)
3. Methods (15 mins $\Sigma 25$)
   - -Information Theory; -Signal processing
   - -Psychophysics; -Articulation Index;
4. Results (30 mins $\Sigma 55$)
   - Confusions; Primes and Morphs;
   - Speech Modifications; Conflicting cues
5. Summary + Conclusions (5 mins $\Sigma 60$)

I – Introduction (5 mins)

- Statement of the problem:
  - A fundamental understand the Human Speech code
- Short-term Goal:
  - Identify the key features in individual CV utterances
    - -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, ε, ı/
- Applications:
  - Reduce variability in ASR at frontend
  - Hearing Aids, Cochlear Implants
  - Smart Telcom products
  - TTS (Text to speech)
  - Intelligibility modifications (Robustness problem)
    - Speech enhancement in noise

Objective

- To develop rigorous procedures for analyzing and modifying speech in noise
- To identify perceptual features, denoted events



- Based on two basic measures:
  - AI-Gram (speech audibility measure)
  - Confusion matrix (CV discrimination measure)
- We will show that onset and durational timing cues form the consonant events

II – Historical HSR Studies (5 mins)

- Lord Rayleigh's 1908 and George Campbell 1910
  - First electronic articulation experiments
- Harvey Fletcher's 1921 Articulation Index AI
  - Accurate predictions of nonsense syllable scores
  - French and Steinberg 1947 first publish AI
- Shannon The thoeory of Information 1948+
  - G.A. Miller, Heise and Lichten Entropy $\mathcal{H}$ 1951
  - G.A. Miller & Nicely CM $P_{h|s}(SNR)$ 1955
- Context:
  - G.A. Miller 1951 *Language and communication*
  - G.A. Miller 1962 5-word Grammer $\equiv$ 4 dB of SNR
  - Boothroyd JASA 1968; Boothroyd & Nittrouer 1988
  - Bronkhorst et al. JASA 1993

Speech feature research

- 1910-1980: Bell Labs
- 1940-1960: Haskins Lab
- 1960-1990: MIT
- 1980-2010: ASR at AT&T, IBM, BBN, University research

Cochlear research

- 1910-1950: Bell Labs
- 1960-2010: MIT + Harvard HSTB
- 1980-2010: NIH funded University research

Speech feature research

- 1910-1980: Bell Labs
- 1940-1960: Haskins Lab Synthetic speech
- 1960-1990: MIT Consonant features unknown
- 1980-2010: ASR at AT&T, IBM, BBN, University research Not designed to be robustness to noise

Cochlear research

- 1910-1950: Bell Labs
- 1960-2010: MIT + Harvard HSTB
- 1980-2010: NIH funded University research

III – Methods (15 mins)

- Information Theory IT $\equiv$ Articulation index AI
  - Confusion matrix CM scores: $P_{h|s}(\mathit{SNR})$
  - AI to model mean phone errors $\sum_h P_{h|s}(\mathit{SNR})$
- Psychophysics
  - Real consonant-vowel CV speech
  - Several types of additive noise
  - Large number of trials
    - $>$20 talkers and $>$20 listeners
- Signal processing
  - AI-gram (crude cochlear model)
  - Frequency, time, intensity truncation 3^d-DS
  - Short-Time Fourier Transform STFT modifications

The CM and the *elemental–event*

- 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	*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

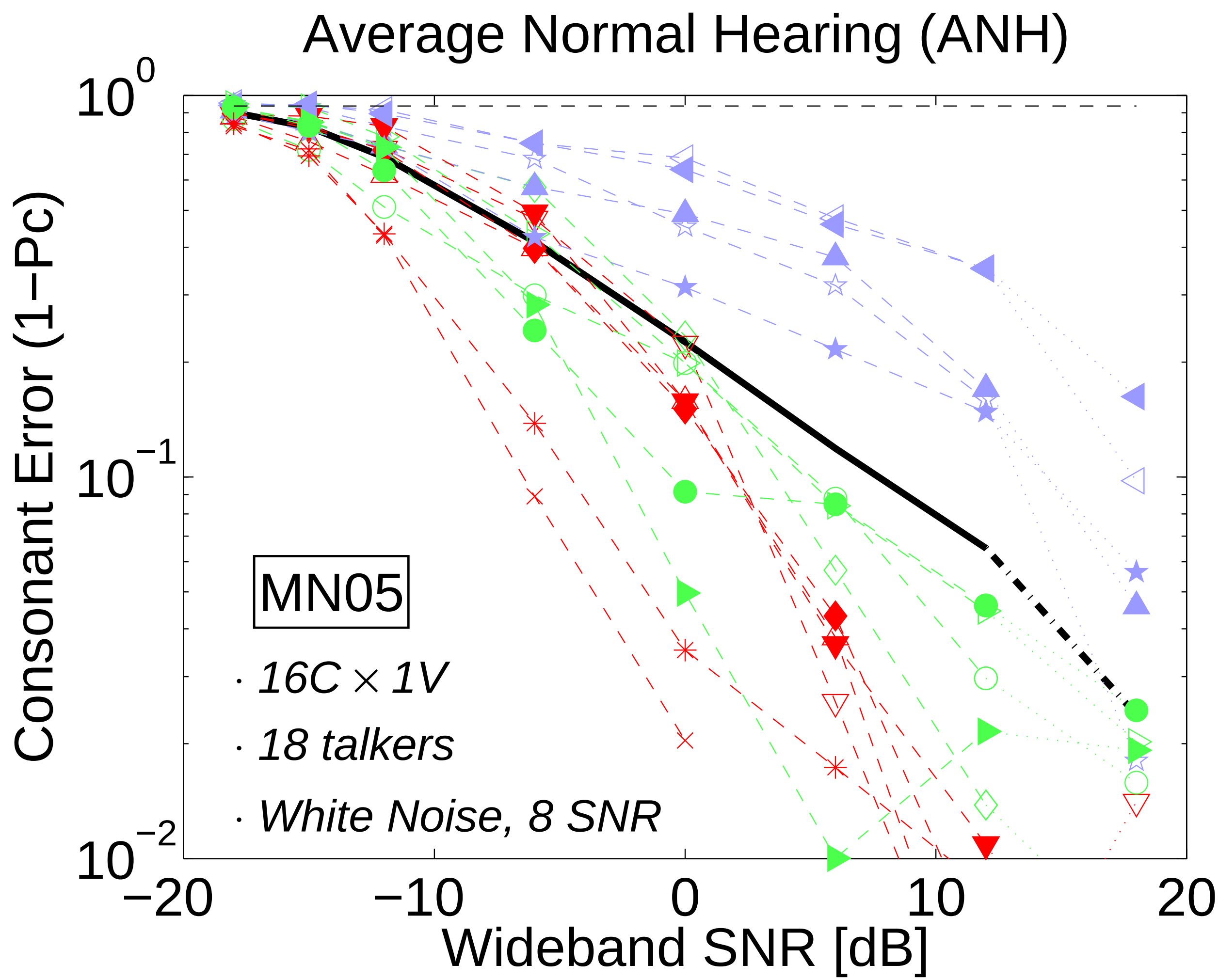
UNVOICED VOICED NASAL

RESPONSE

- Confusion groups $\equiv$ *inhomogeneous elemental-events*

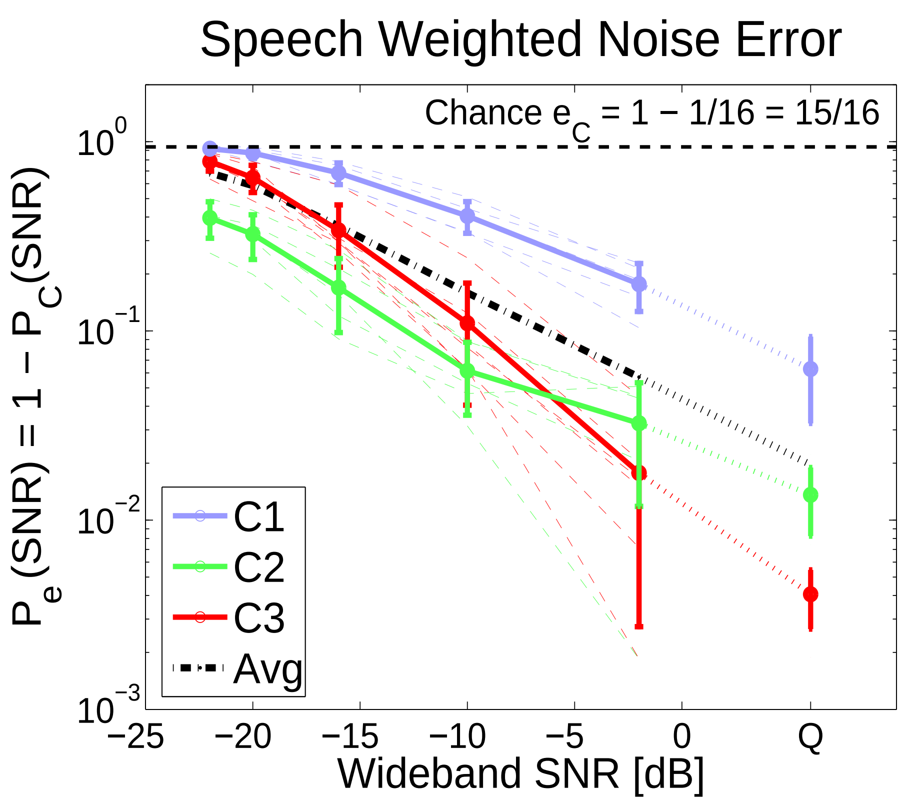
Average phone scores vs. SNR

- Consonant chance performance is -20 dB-SNR in white noise   Phatak Allen, 2007

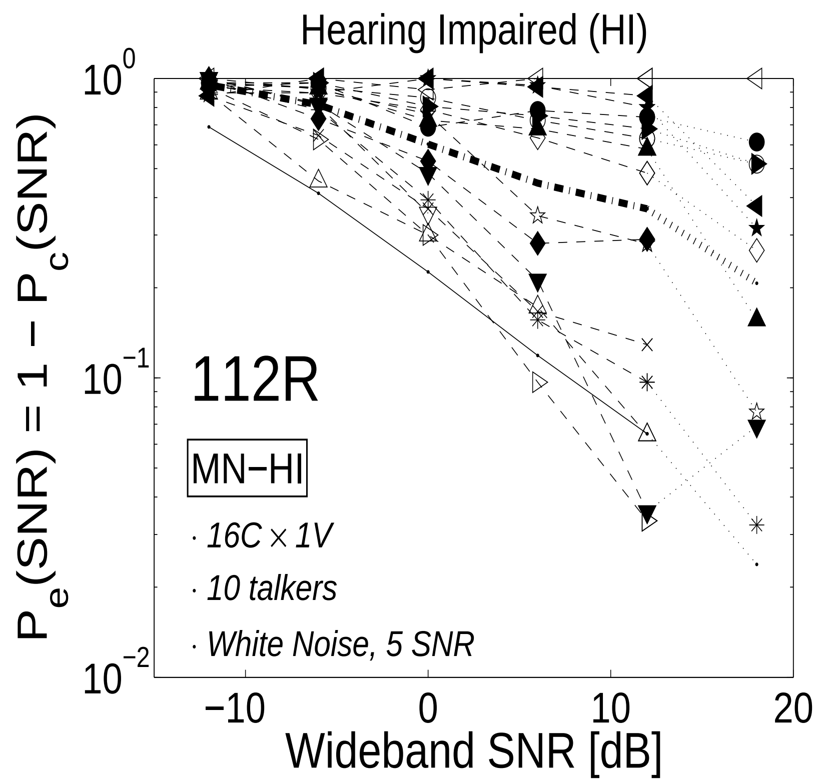


Consonant Variability

- Avg. Contant error $P_{h|s}(SNR)$ strongly hetrogeneous!
- NH listeners above chance at $<$ –25 dB SNR in SWN
  - HI $P_e(SNR) >> $ ANH $P_e(SNR)$



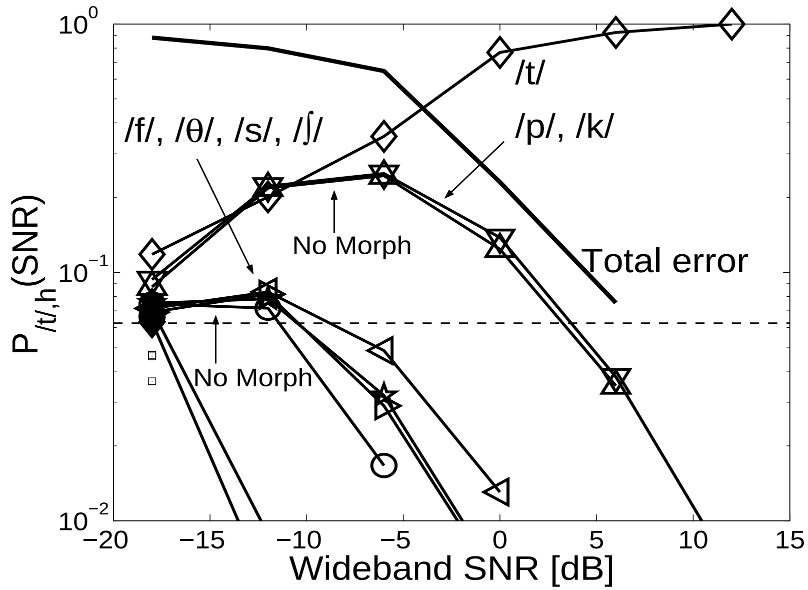
(a) Consonant errors ANH



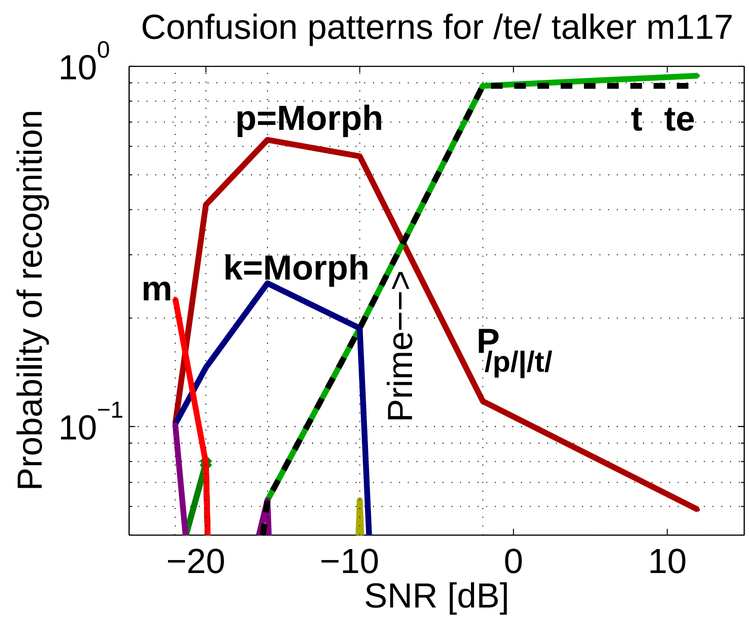
(b) Consonant errors HI

Row of CM $P_{h|/t/}$

- Utterance phone scores are hetrogeneous!



(c) Average over all /t/s.

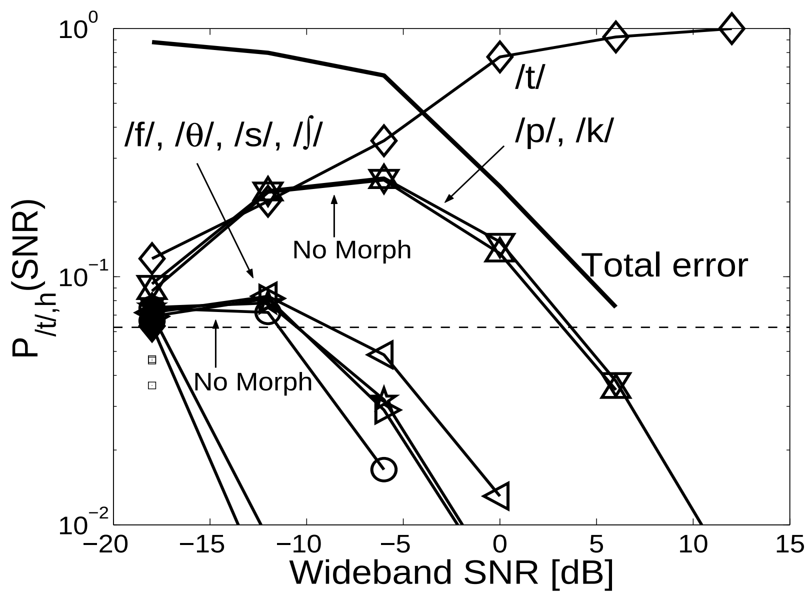


(d) Talker m117 /te/ $P_{h|/ta/}(SNR)$

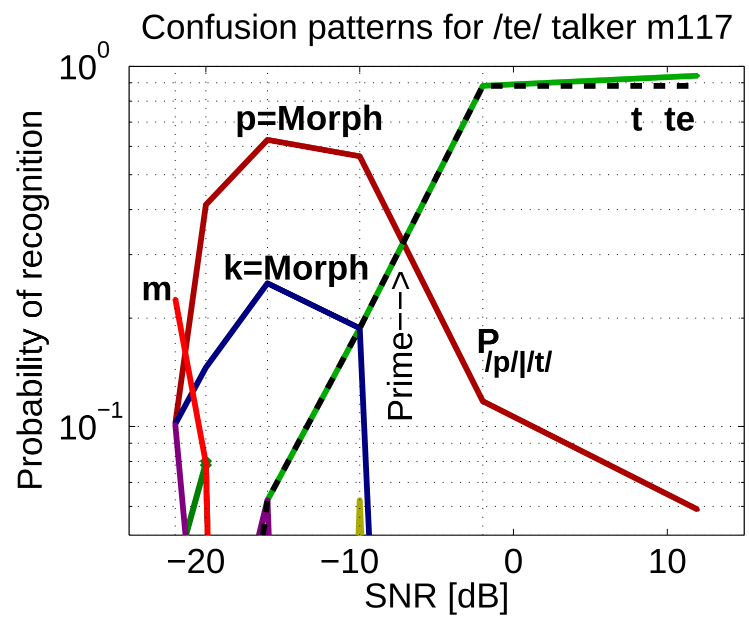
- Phone groups are due to shared sub-phonemic units
  - CV Morphs
  - Morphing sentences

Row of CM $P_{h|/t/}$

- Utterance phone scores are hetrogeneous!



(e) Average over all /t/s.



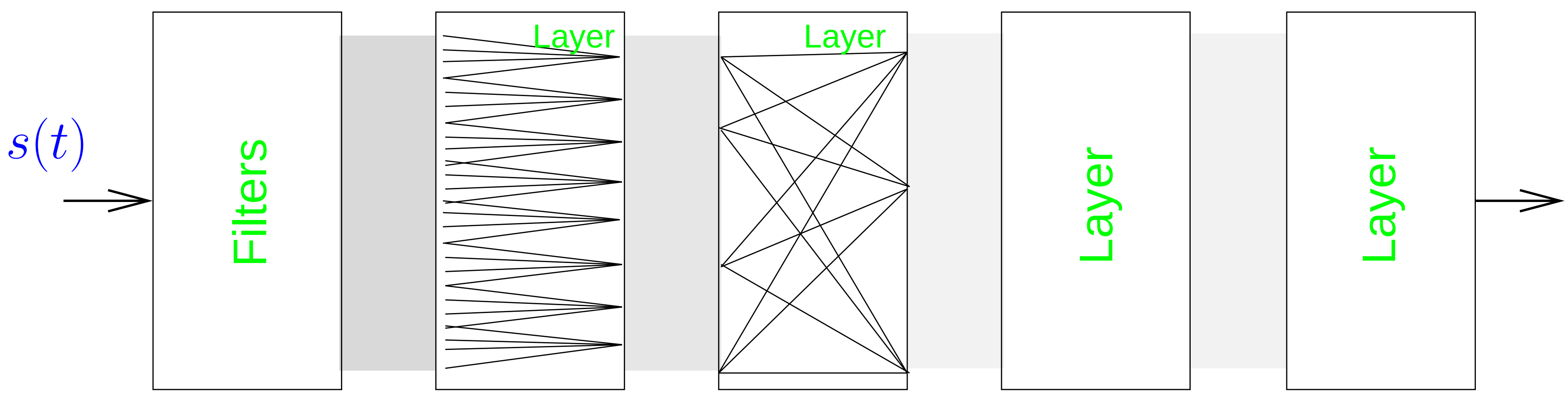
(f) Talker m117 /te/ $P_{h|/ta/}(SNR)$

- Phone groups are due to shared sub-phonemic units
  - CV Morphs DEMO
  - Morphing sentences DEMO

Model of human speech recognition HSR

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

Output:	Cochlea	Event	Phones	Syllables	Words	



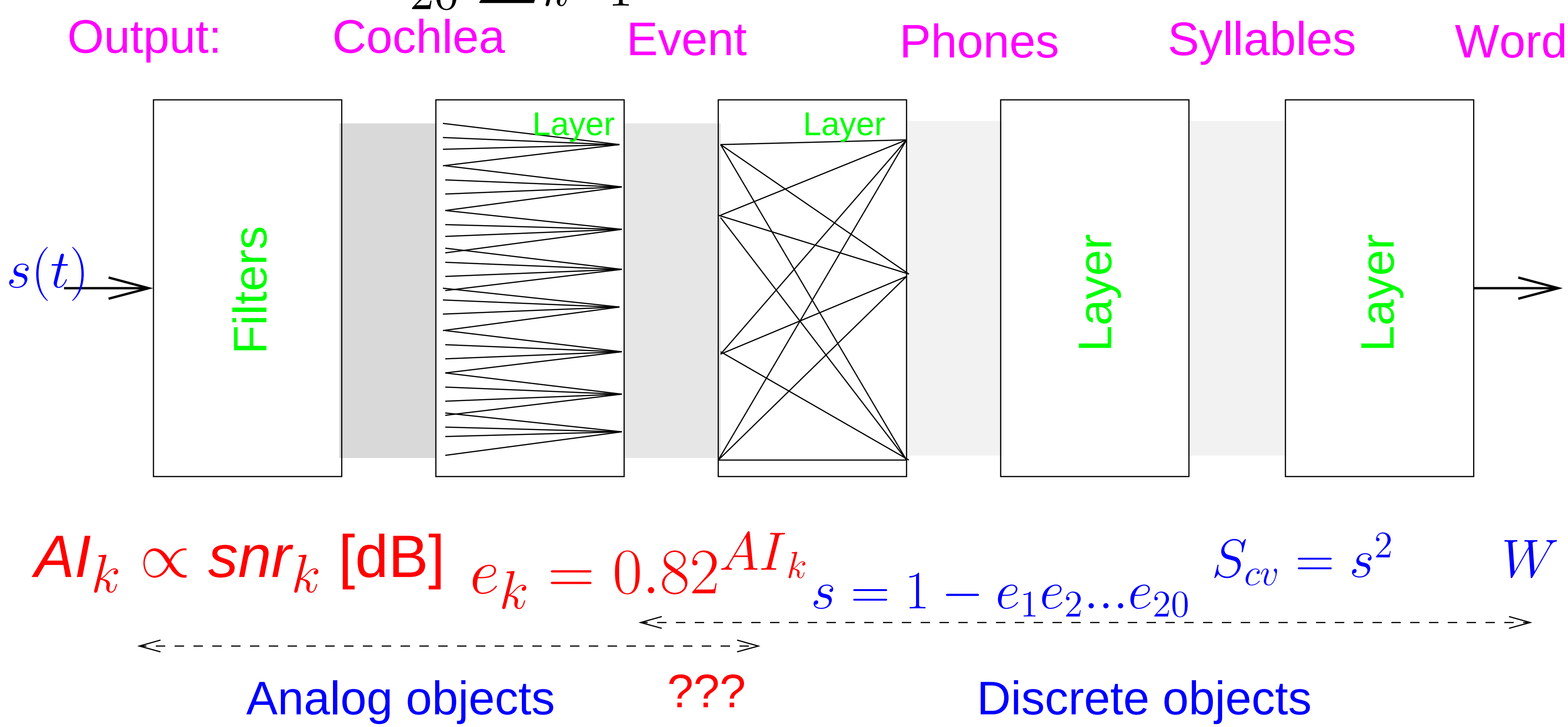
Measure:	AI_k	e_k	s	$S = s^3$	W
Formula:	$\propto$ *snr*$_k$ dB	$= 0.82^{AI_k}$	$= 1 - e_1 e_2 \ldots e_{20}$		

Analog objects ??? Discrete objects

Φ "Front-end" Ψ "Back-end"

Definition and use of the *AI*

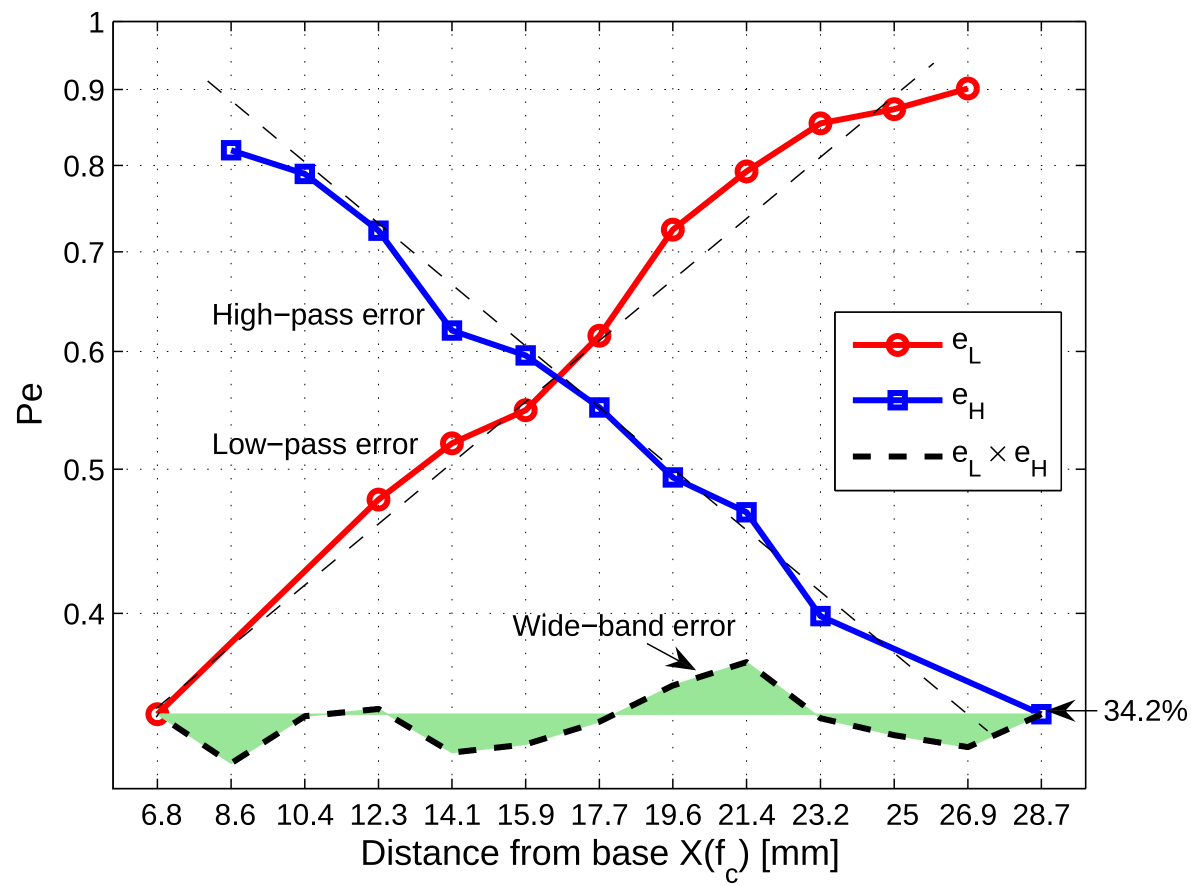
- The average error is: $P_e(\mathsf{SNR}) \equiv \prod_k e_k = 0.02^{\mathsf{AI}}$
  - $e_k = 0.822^{\mathsf{AI}_k(\mathsf{snr}_k)}$ cochlear k^{th} band-error
  - $AI_k = \log_{10}(1 + 4\mathsf{snr}_k^2)^{1/3}$ band channel-capacity
  - $\mathsf{AI} \equiv \overline{\mathsf{AI}_k} = \frac{1}{20}\sum_{k=1}^{20} AI_k,$



Fletcher's Lowpass/Highpass result

- The AI is based on the *band-error product formula*

$$P_e(\mathsf{snr}, \cancel{f_c}) \equiv e_{lp}(\mathsf{snr}, f_c) \times e_{hp}(\mathsf{snr}, f_c)$$

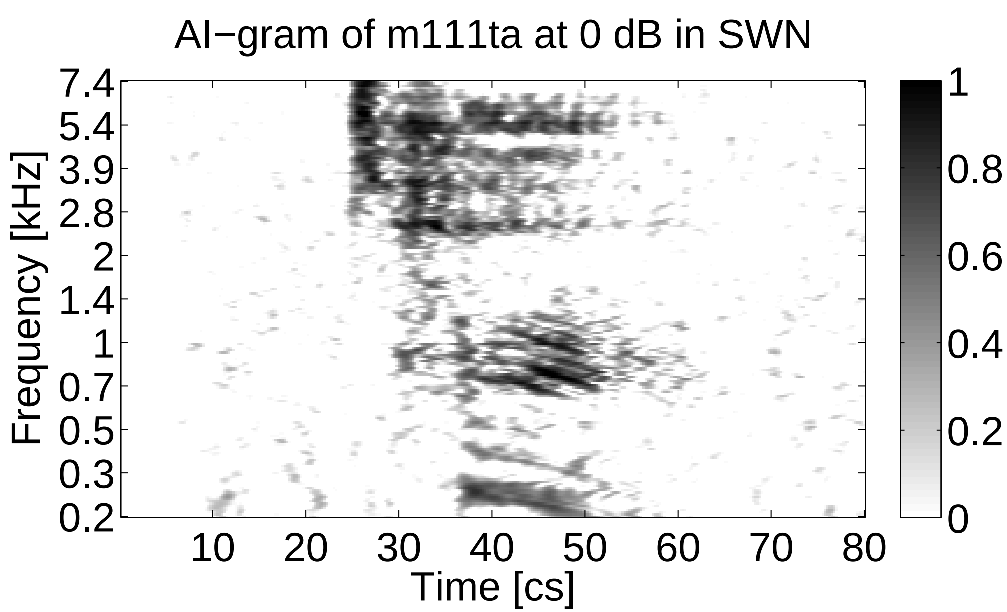


Human listeners as a Shannon Channel

- The Channel capacity theorem gives the maximum information rate as:

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

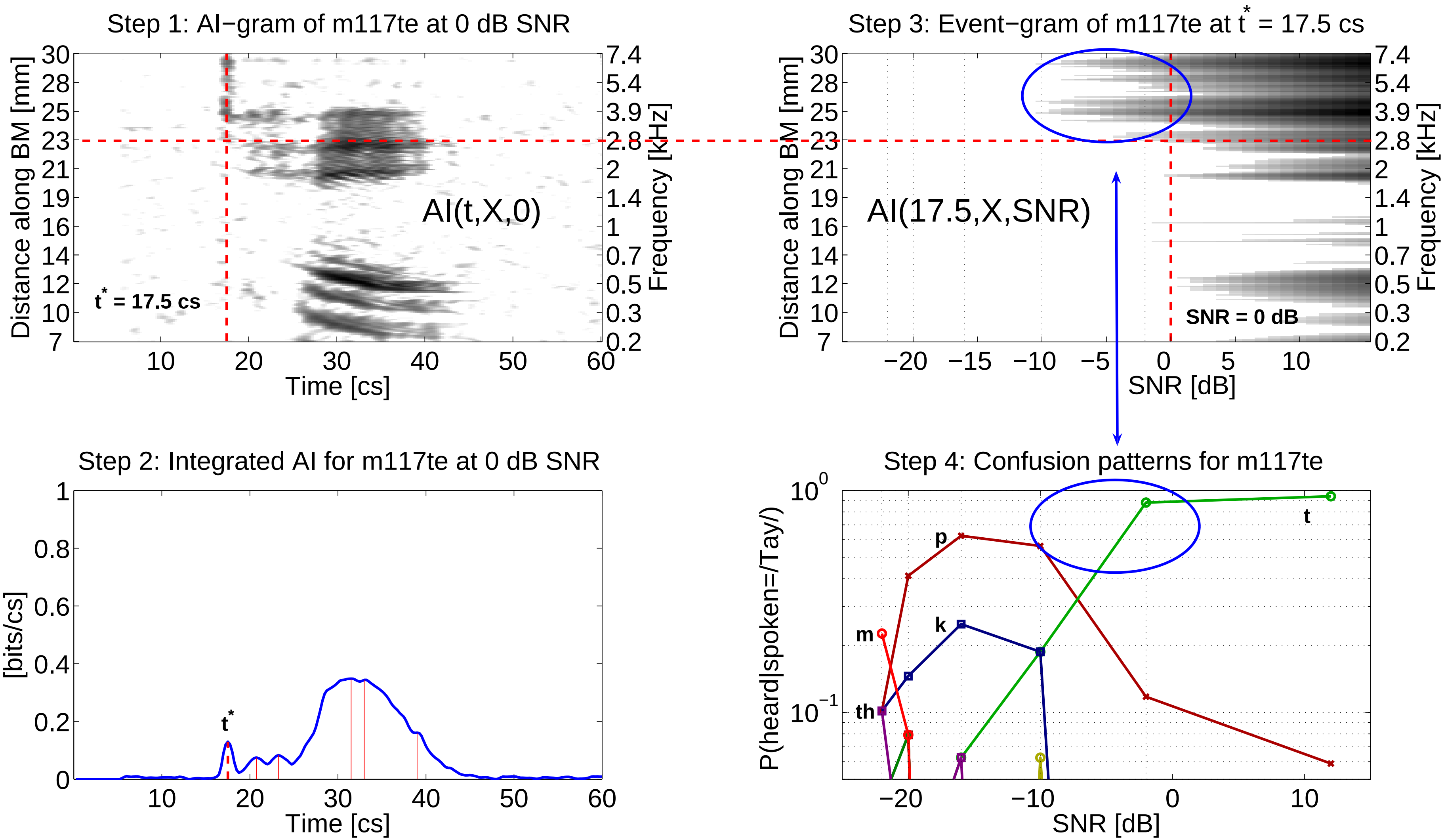
- For a Maximum Entropy (MaxEnt) speech source, the maximum information rate is determined by the SNR
- The AI-gram is a closely related measure:



III–Results (30 mins)

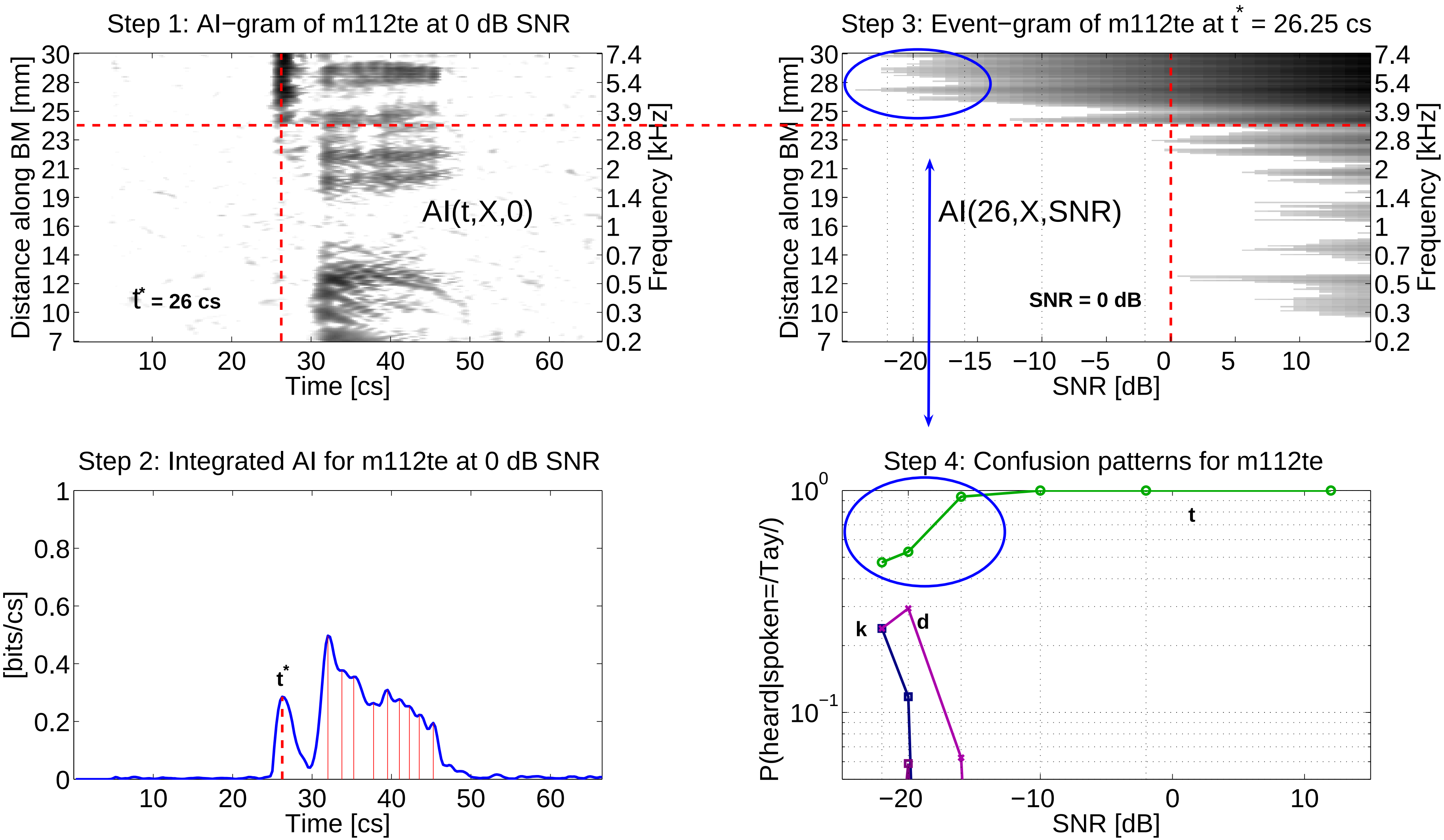
- Examples and Demos of events
  - Plosive CV events
  - Fricative CV events
- Conflicting cues
- DEMOS:
  - Event isolation
  - Consonant morphing
  - Consonant enhancement
  - Conflicting cues within consonants
  - Sentence meaning modification

m117/tɛ/ in speech-weighted noise



- /t/ confusion threshold at $P_c(\mathit{SNR}^* = -2) = 0.9$ correlated to Event-gram

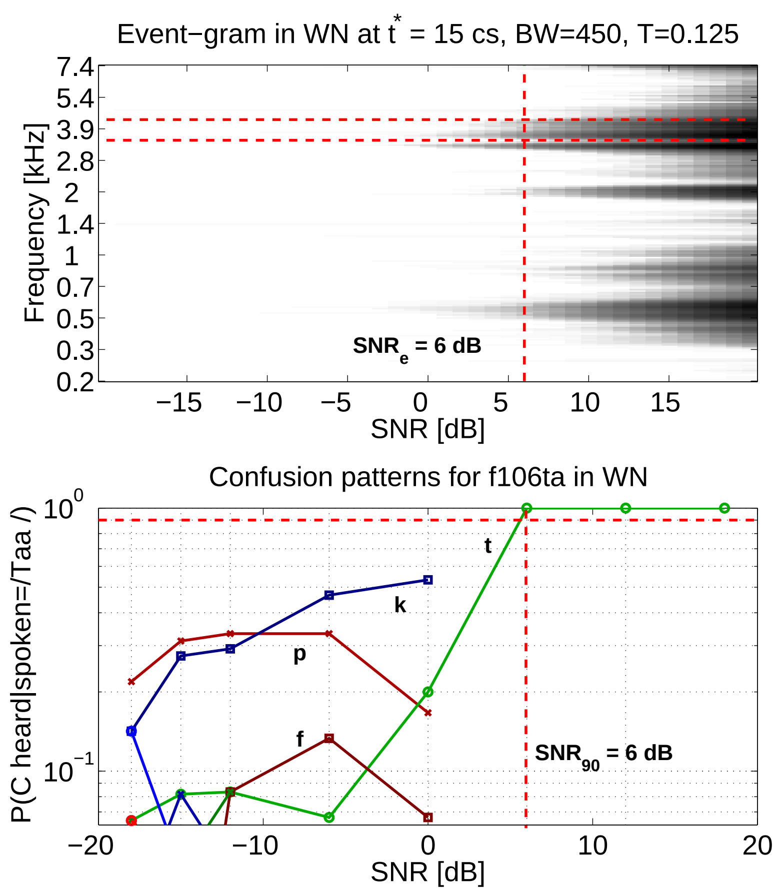
m112/tɛ/ in speech-weighted noise



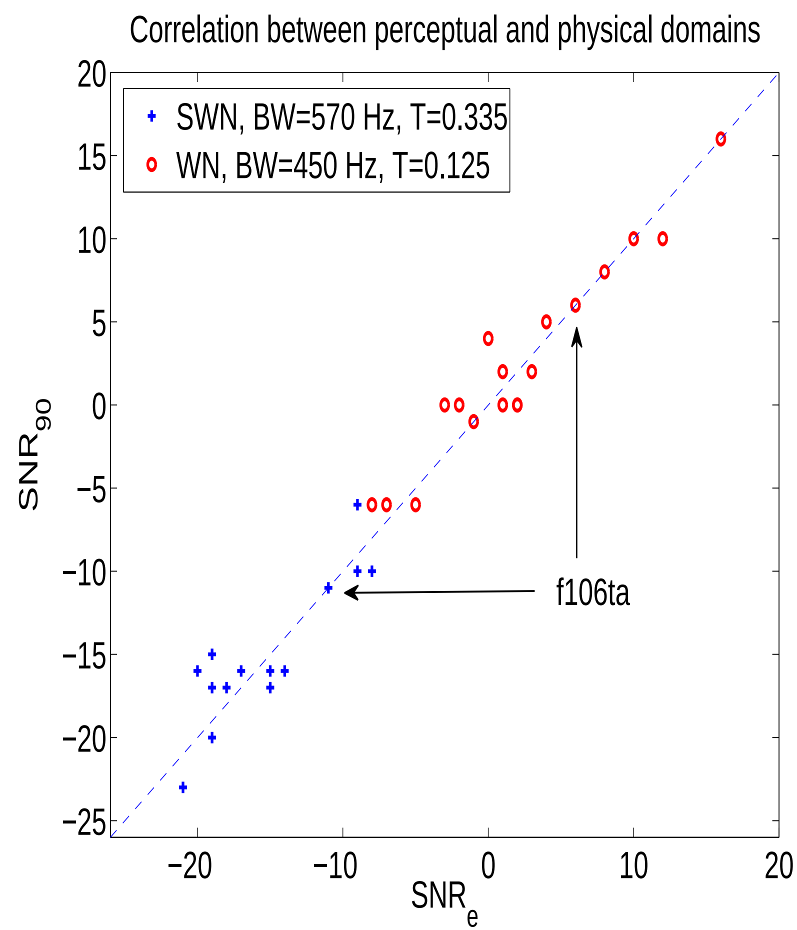
- /t/ confusion threshold at $P_c(\mathit{SNR}^* = -16) = 0.9$ correlated to Event-gram

Correlations of /t/ events

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

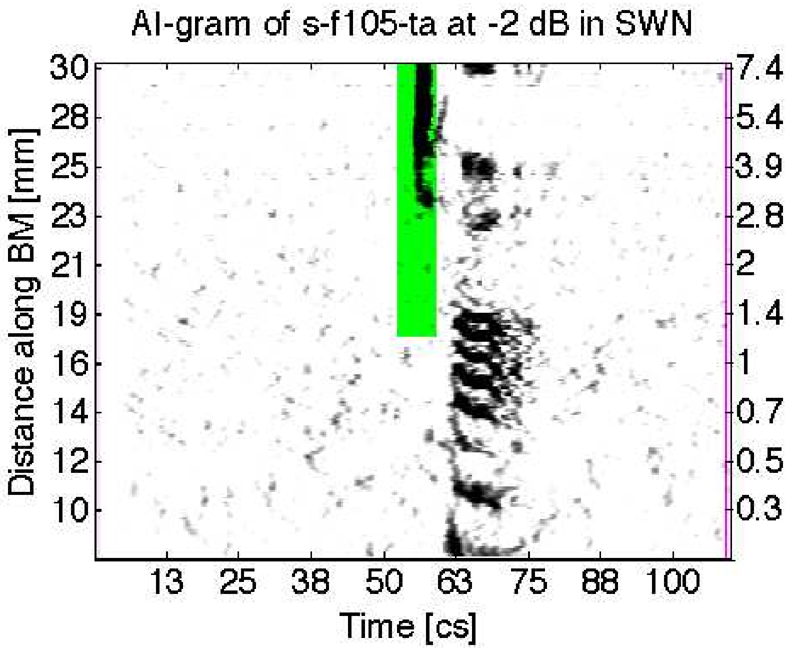


(g)

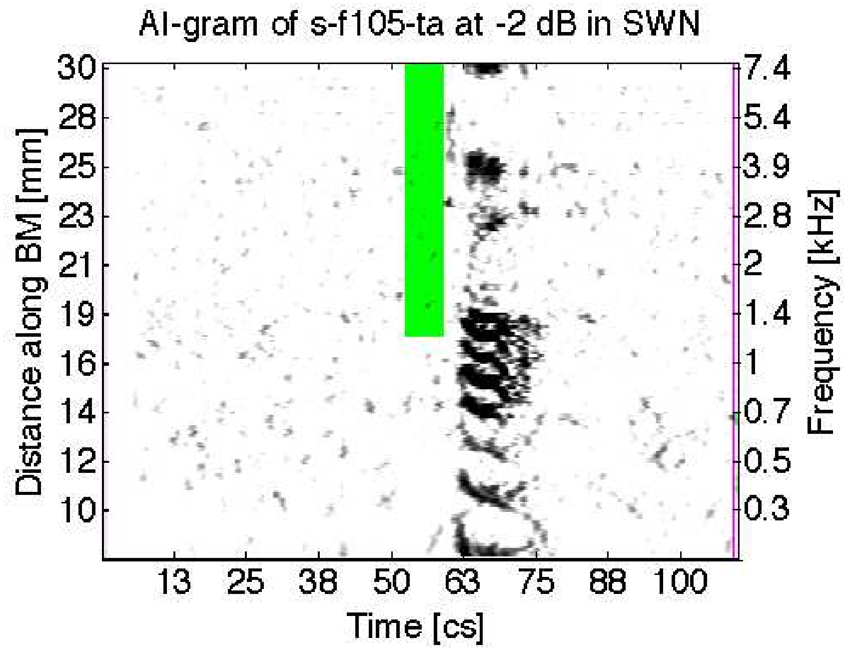


(h)

Masking of /tɑ/ timing cue



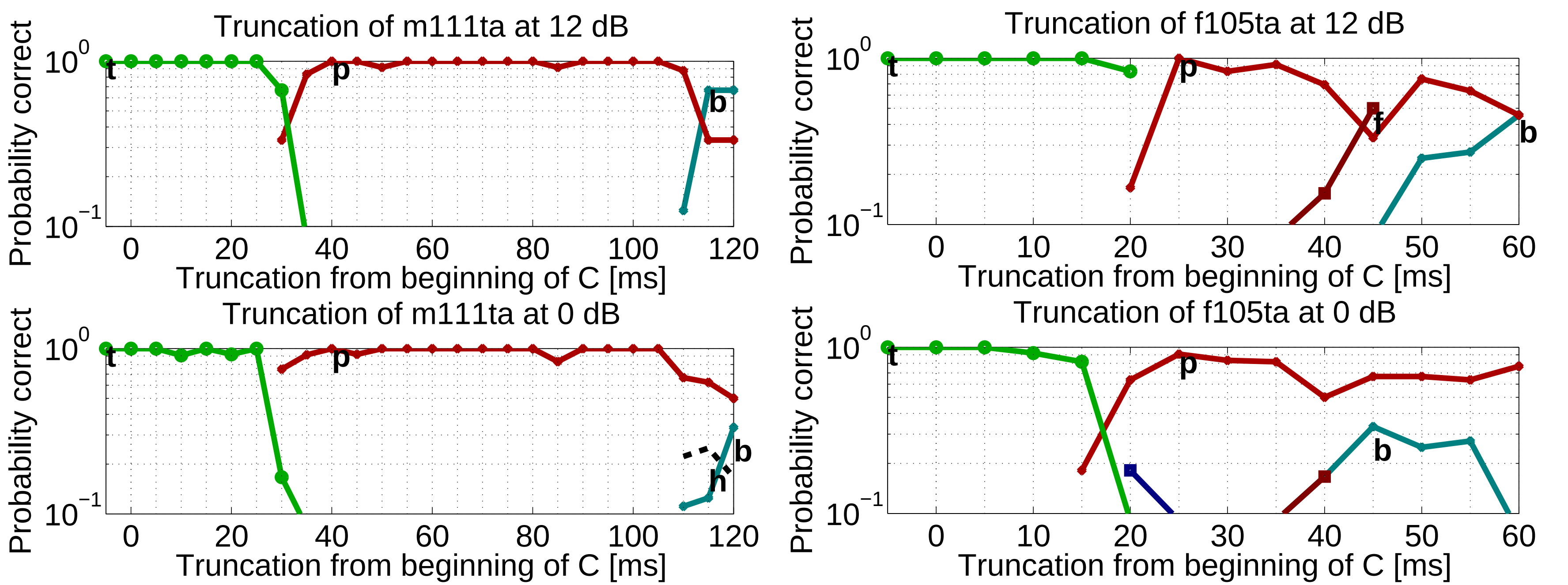
(i) Original /tɑ/



(j) Modified /tɑ/

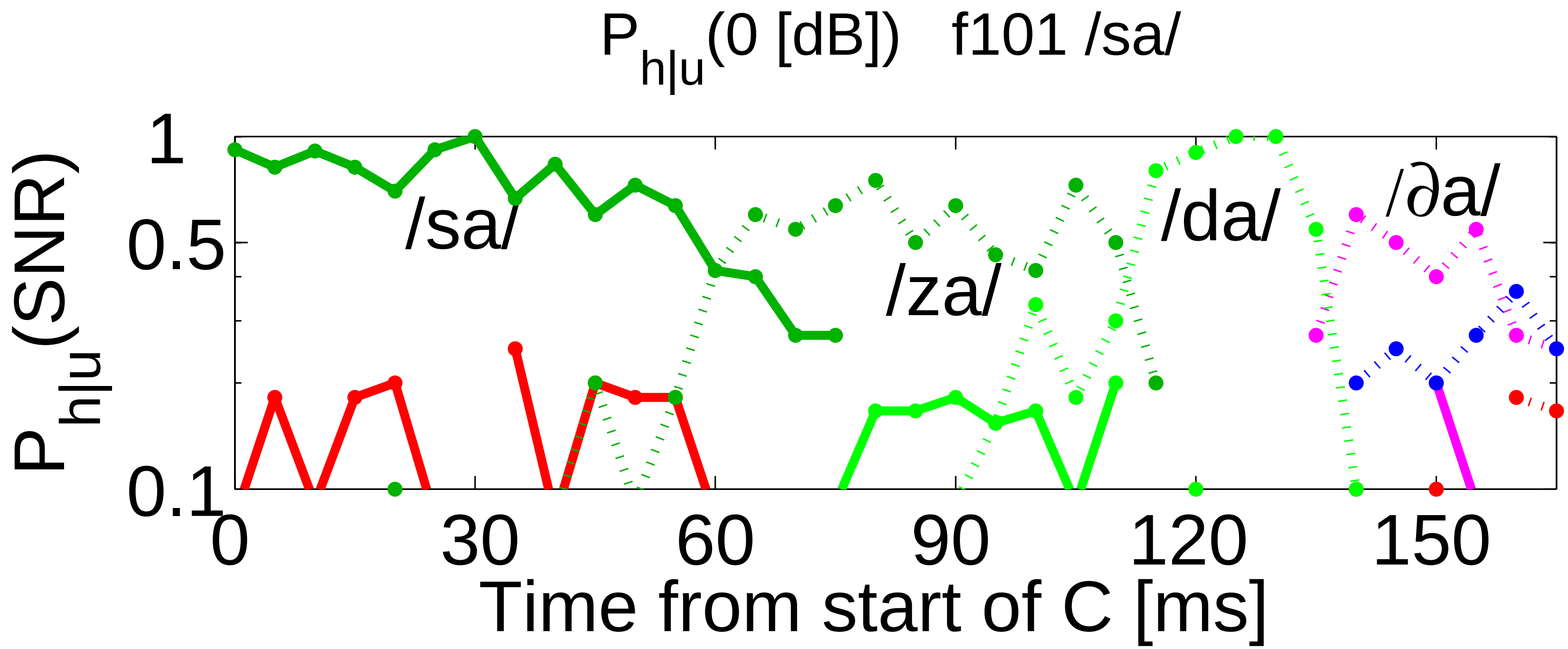
- When the /t/ burst is masked by noise, the perception morphs to /p/
- DEMO 4

Truncation of /tɑ/



- This represents the normal hearing responses to a truncated /tɑ/, from the start of the consonant
- Morphing from /tɑ/ to /pɑ/ to /bɑ/ at 0 and 12 dB SNR
- Similar to Furui 1986, and our extensive results

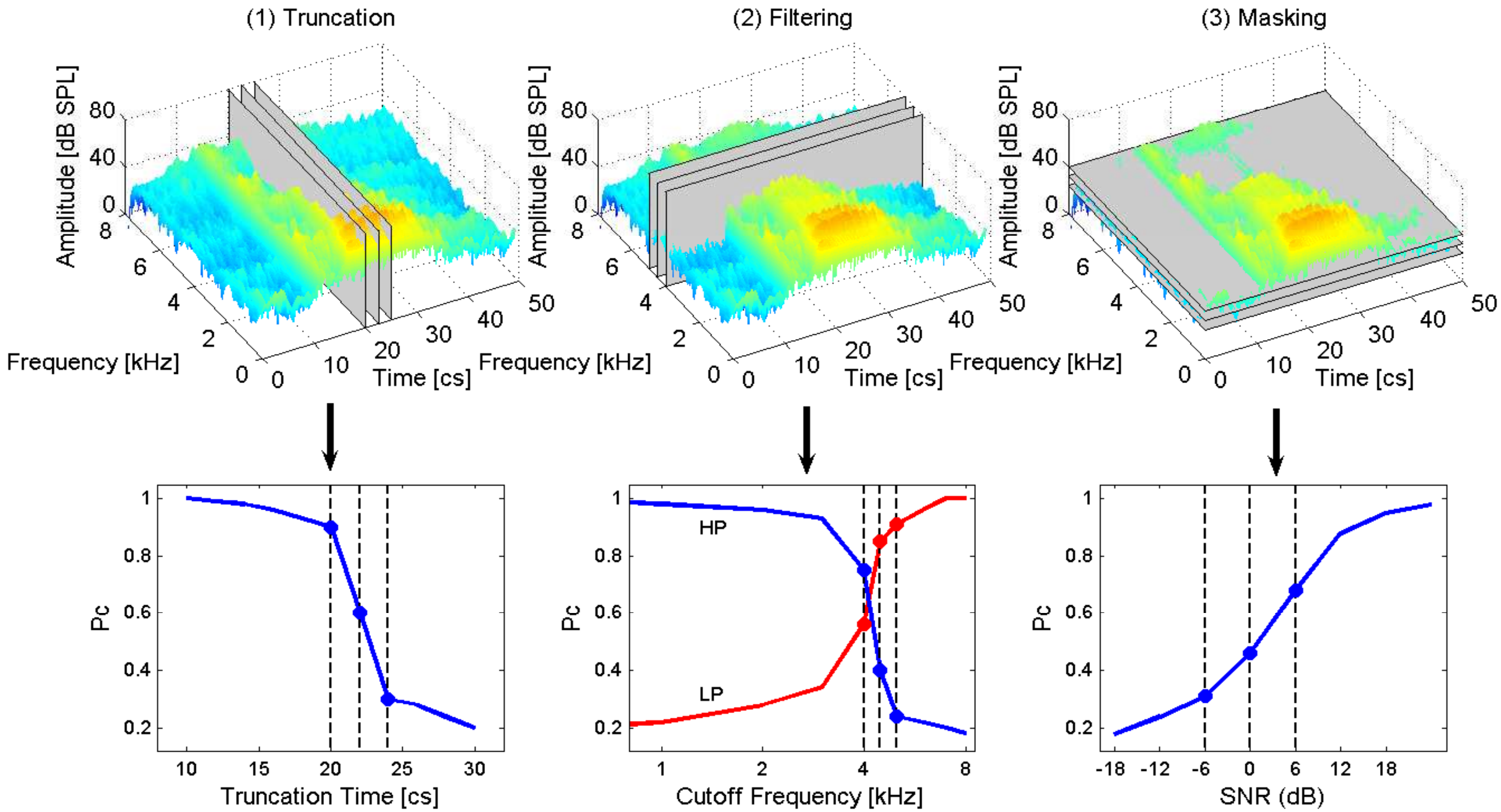
Truncation of f101 /sa/



- This represents the normal hearing responses to a truncated /sɑ/, from the start of the consonant
- Morphing from /sɑ/ to /zɑ/ to /dɑ/ to /ðɑ/
- Duration seems to be a fricatives event

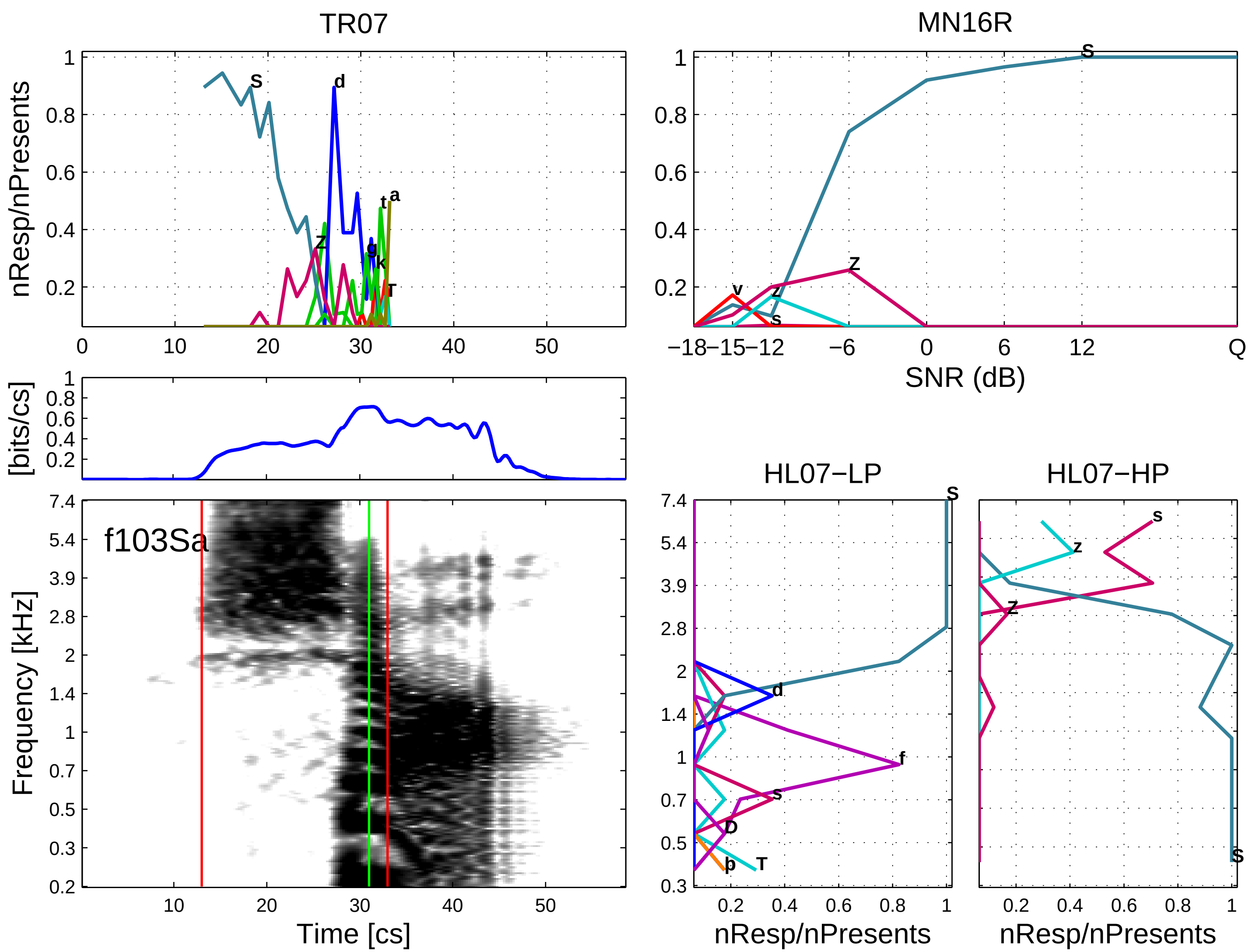
Methods: 3^d Deep Search (3^d-DS)

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



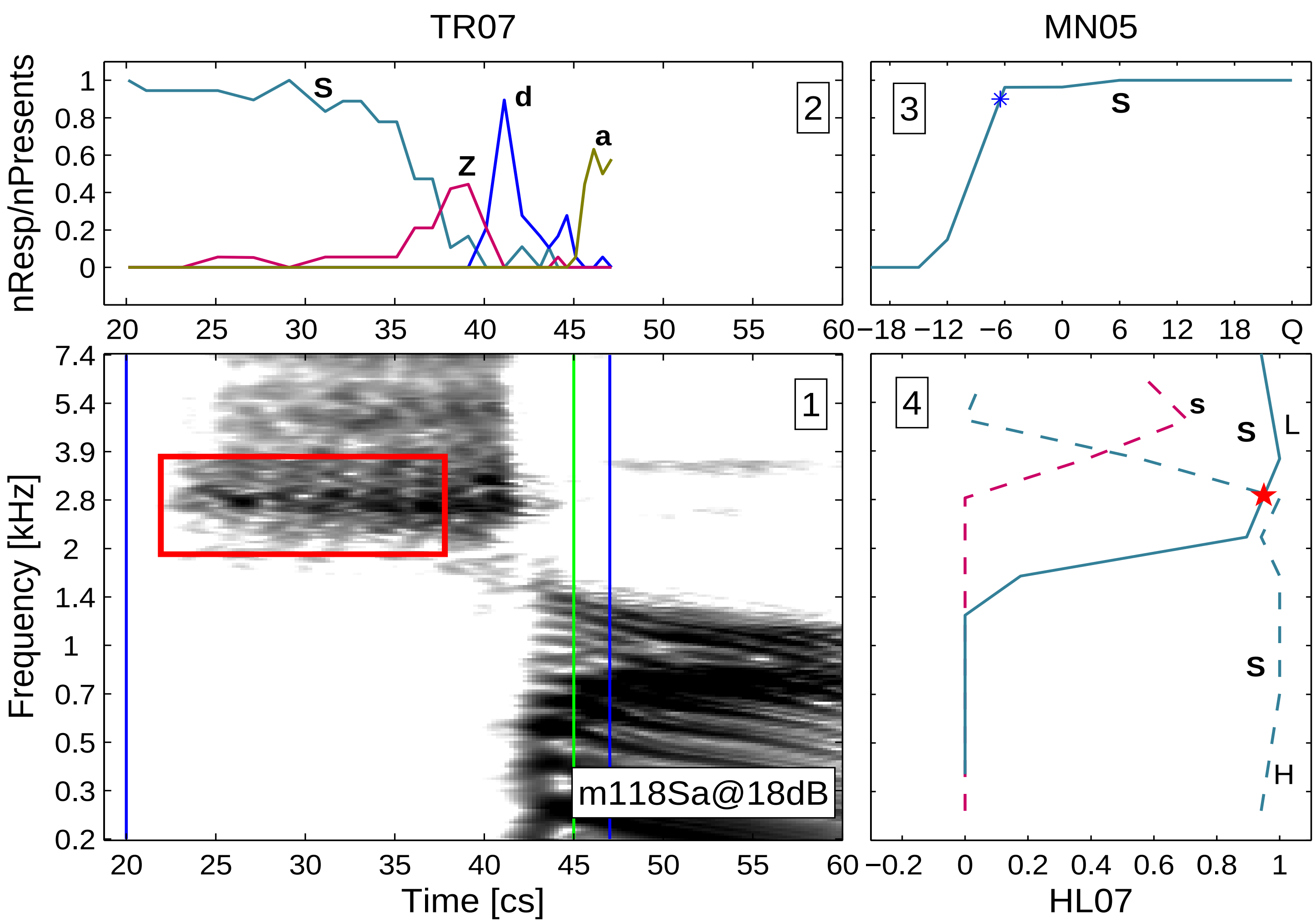
3^d-DS Method /ʃa/

- Truncation in Time, Intensity and Frequency



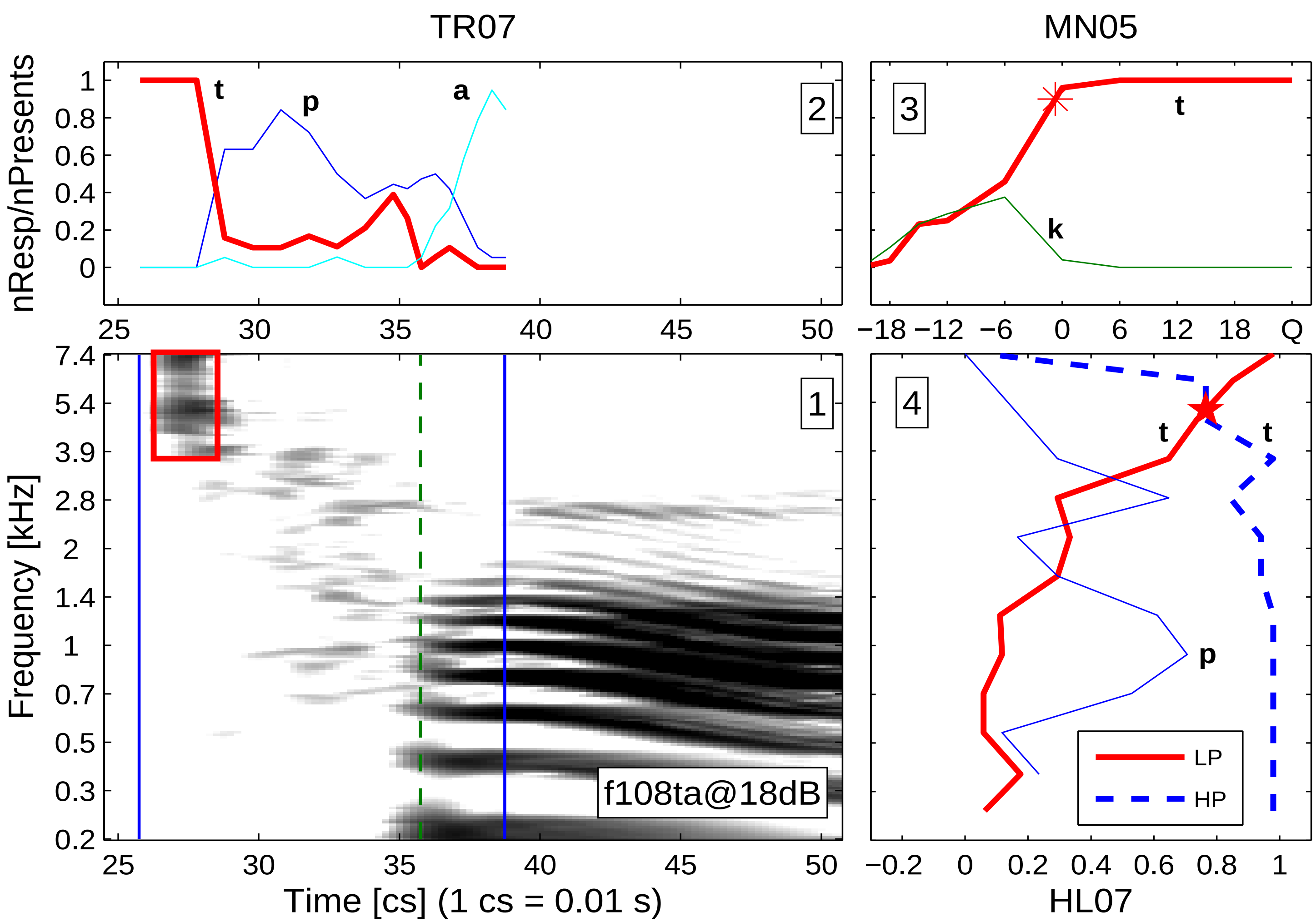
3^d-DS Method /sa/

- Truncation in Intensity, time and frequency

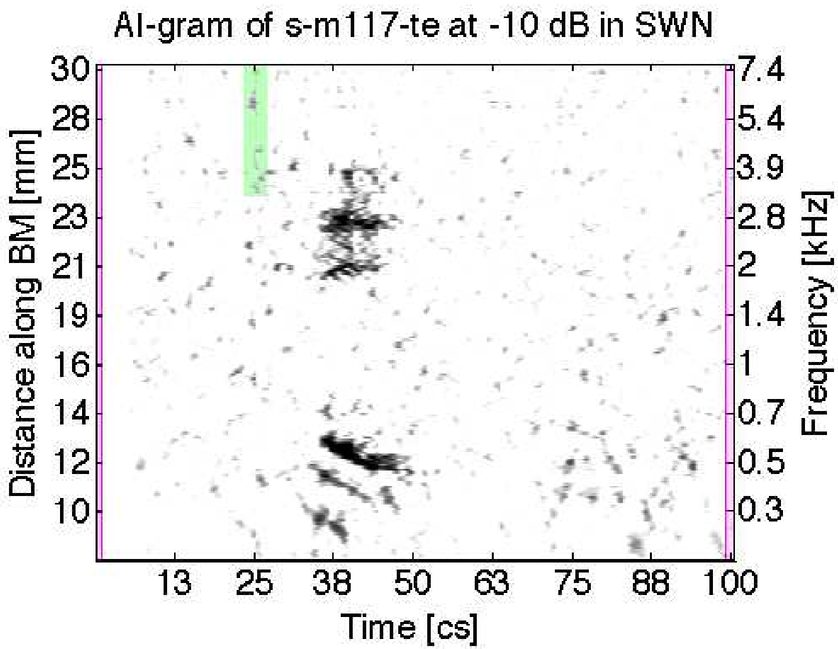


3^d-DS Method /ta/

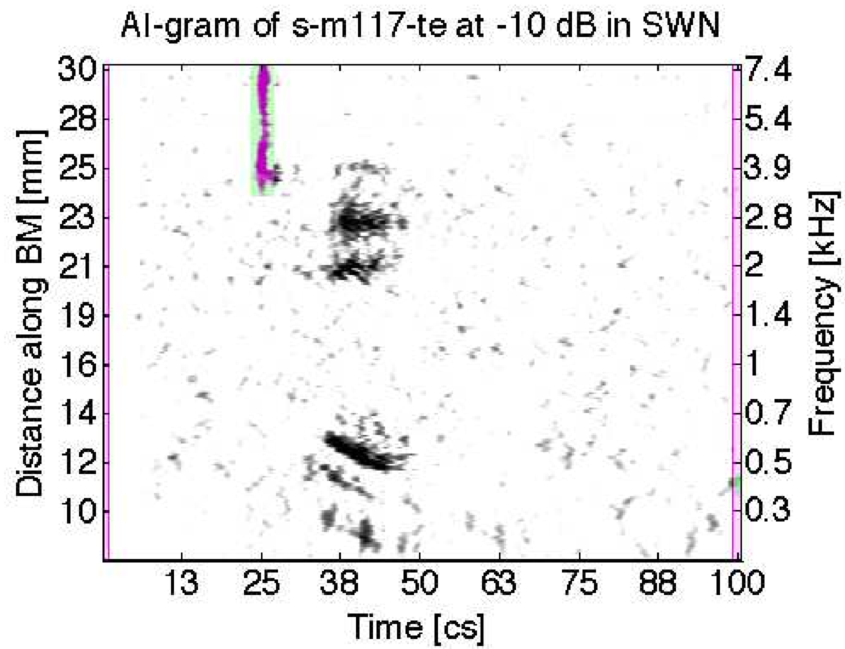
- Truncation in Intensity, time and frequency



Enhancement of /tɛ/ event



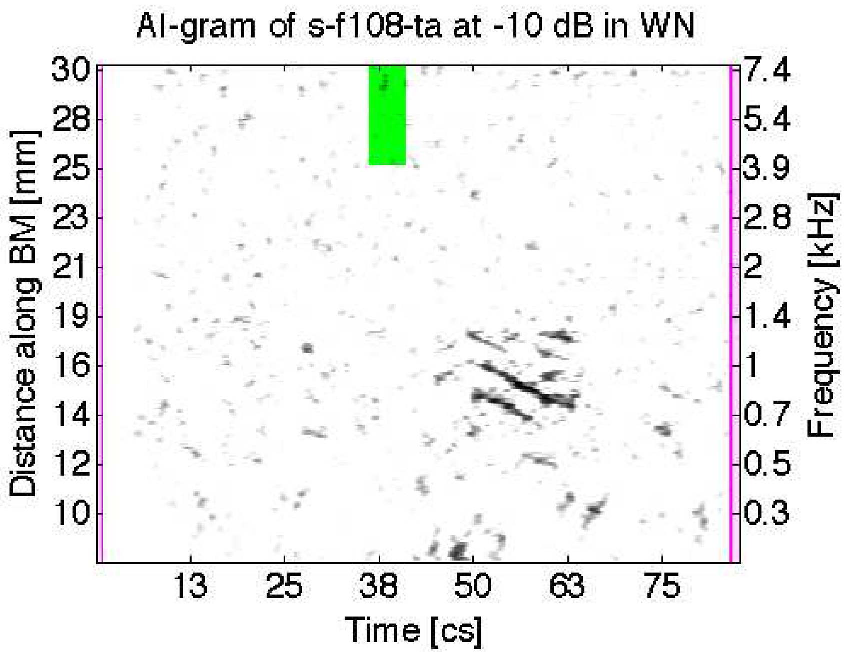
(k) Original /tɛ/



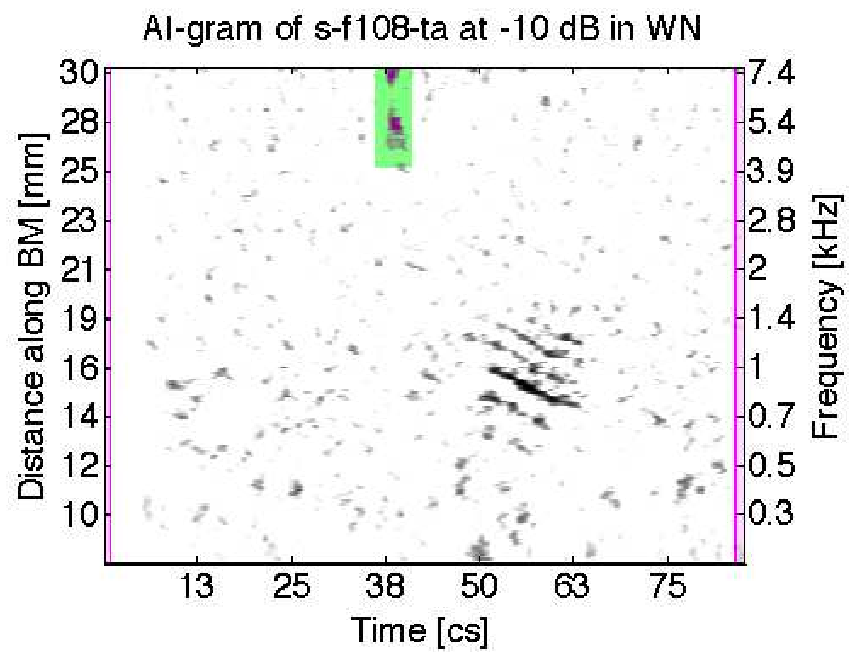
(l) Modified /tɛ/

- The sound is heard as /t/ again, we suppressed the morph (see confusion patterns of slide 4)
- METHODS: The /t/ burst is enhanced (14 dB) on the quiet sound, then noise is added
- DEMO

Enhancement of /tɑ/ event



(m) Original /tɑ/



(n) Modified /tɑ/

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

Conclusion I

We have:

1. isolated events for CV: Plosives /p, t, k/ and /b, d, g/ and Fricatives /θ, ʃ, ʧ, s, h, f/ and /z, ʒ, v, ð/) + Vowels /o, ɛ, ɪ/
   - for many individual talkers
   - via new tools (AI-gram, Event-gram and 3^d-DS)

Conclusion I

We have:

1. isolated events for CV: Plosives /p, t, k/ and /b, d, g/ and Fricatives /θ, ʃ, ʧ, s, h, f/ and /z, ʒ, v, ð/) + Vowels /o, ɛ, ɪ/
   - for many individual talkers
   - via new tools (AI-gram, Event-gram and 3^d-DS)
2. shown that normal listeners use
   - *across-frequency timing coincidences*
   - duration and bandwidth

   to discriminate consonants in noise

Conclusion I

We have:

1 isolated events for CV: Plosives /p, t, k/ and /b, d, g/ and Fricatives /θ, ʃ, ʧ, s, h, f/ and /z, ʒ, v, ð/) + Vowels /o, ε, ı/

- for many individual talkers
- via new tools (AI-gram, Event-gram and 3^d-DS)

2 shown that normal listeners use

- *across-frequency timing coincidences*
- duration and bandwidth

to discriminate consonants in noise

3 developed tools to

- Morphed speech sounds
- Decrease or increase intelligibility. Ex: /tɑ/, /tε/

Conclusion II

We have shown:

1 the existance of conflicting cues

- Thus MaxEnt consonants are NOT redundent

Conclusion II

We have shown:

1 the existance of conflicting cues
   - Thus MaxEnt consonants are NOT redundent

2 that the event threshold is abrupt (i.e., 6 dB)

Conclusion II

We have shown:

1 the existance of conflicting cues
   - Thus MaxEnt consonants are NOT redundent

2 that the event threshold is abrupt (i.e., 6 dB)

3 proven the AI band-product formula (yet again)

Conclusion II

We have shown:

1 the existance of conflicting cues
- Thus MaxEnt consonants are NOT redundent

2 that the event threshold is abrupt (i.e., 6 dB)

3 proven the AI band-product formula (yet again)

4 why the AI works
- Due to the frequency and SNR event distribution

Conclusion II

We have shown:

1 the existance of conflicting cues
   - Thus MaxEnt consonants are NOT redundent

2 that the event threshold is abrupt (i.e., 6 dB)

3 proven the AI band-product formula (yet again)

4 why the AI works
   - Due to the frequency and SNR event distribution

5 the role of forward and upward masking spread

Conclusion III

This could lead to:

1 Improved automatic speech recognition front-ends

Conclusion III

This could lead to:

1 Improved automatic speech recognition front-ends

2 The design of new hearing aids

Question your basic assumptions

Thanks for your attention
http://hear.ai.uiuc.edu