The role of the cochlea in Human speech recognition

Where is the speech information lurking?

Jont Allen

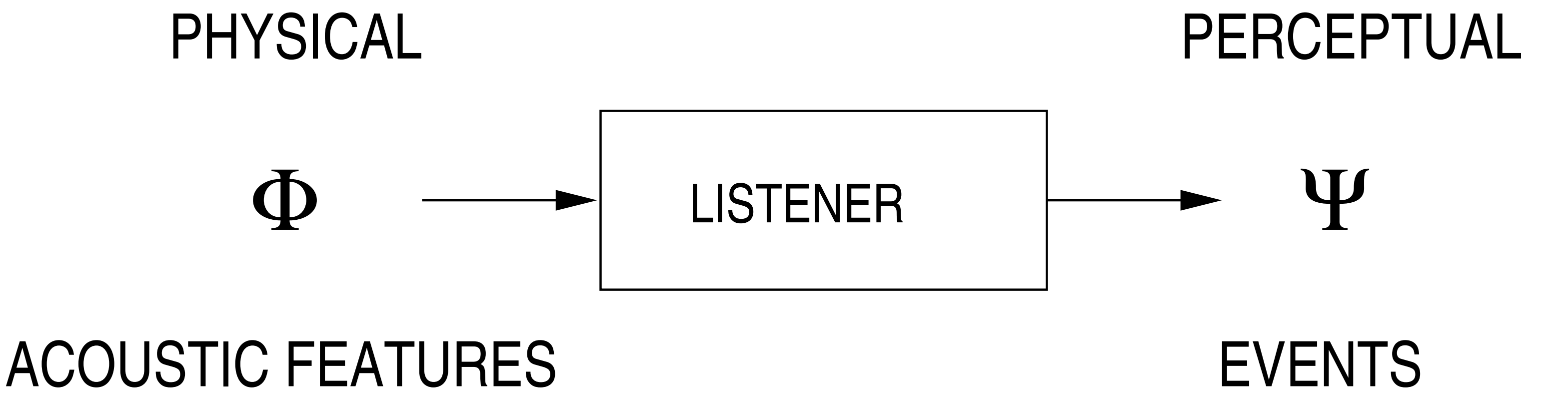
Sandeep Phatak

Marion Regnier

Feipeng Li

Univ. of IL, Beckman Inst., Urbana IL

Objective



- Develop rigorous procedures for analyzing and modifying speech in noise, to:
  - identify perceptual features, denoted events
  - Develop a theory of human speech recognition (HSR) based on two basic measures:
    1. AI-Gram (speech audibility measure)
    2. Confusion matrix (speech discrimination measure)
- Show that across-frequency timing cues are events

Human listeners as a Shannon Channel

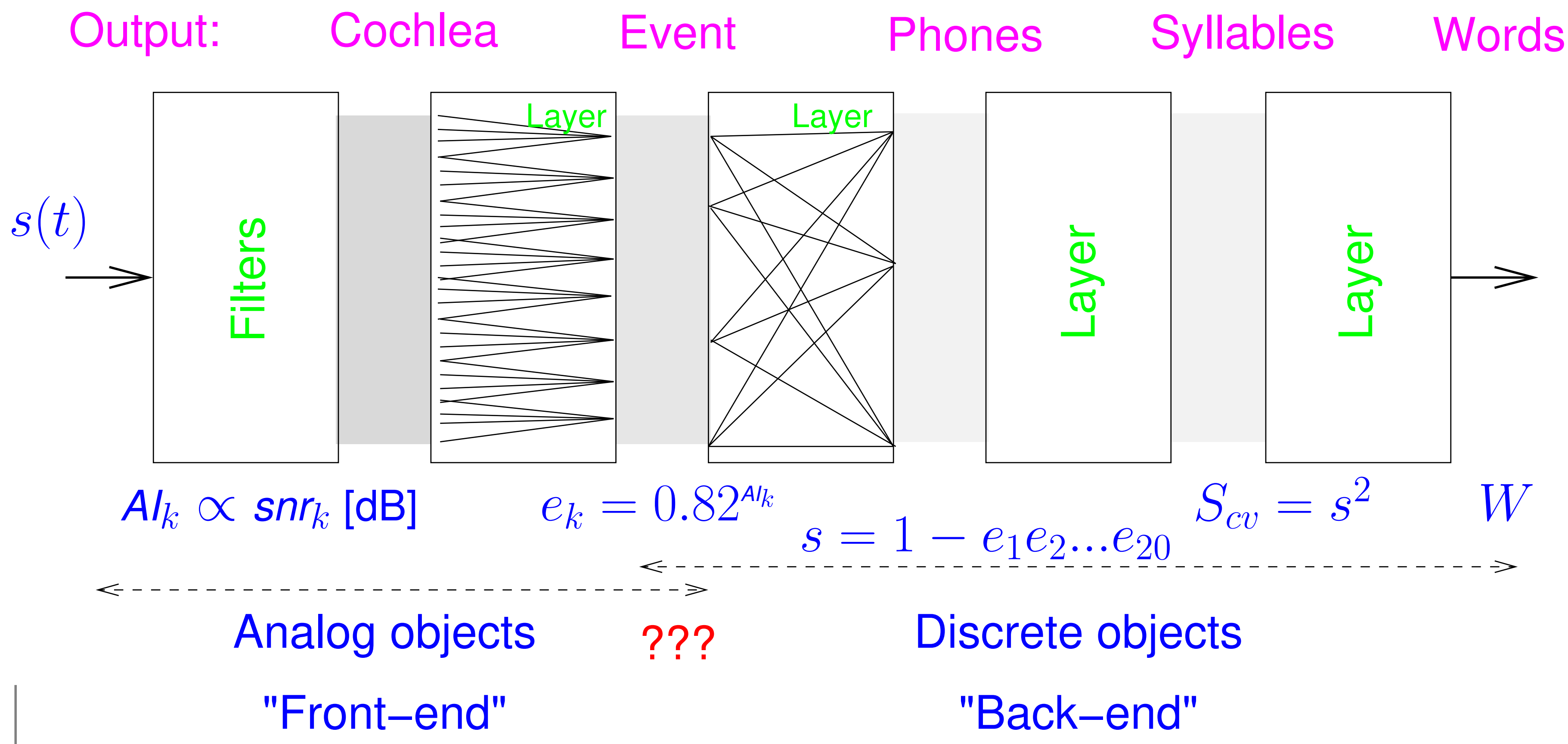
- My approach is inspired by information theory using a classic 3-pronged approach: Simplify, simplify, simplify
  1. The Channel capacity theorem gives the maximum information rate as:

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

  2. The basic idea is to use a Maximum entropy (MaxEnt) speech source, and reduce the maximum information rate for by increasing the noise.
- Take full advantage of *Articulation Index* predictions of the average phone score $s = P_c(AI)$

Model of human speech recognition (HSR)

- The research goal is to identify *elemental HSR events*
  - An event is defined as a *perceptual feature*
  - Event errors are measured by band errors e_k



Articulation Matrices and *elemental events*

- Miller-Nicely's 1955 articulation matrix $\mathcal{A}$ 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

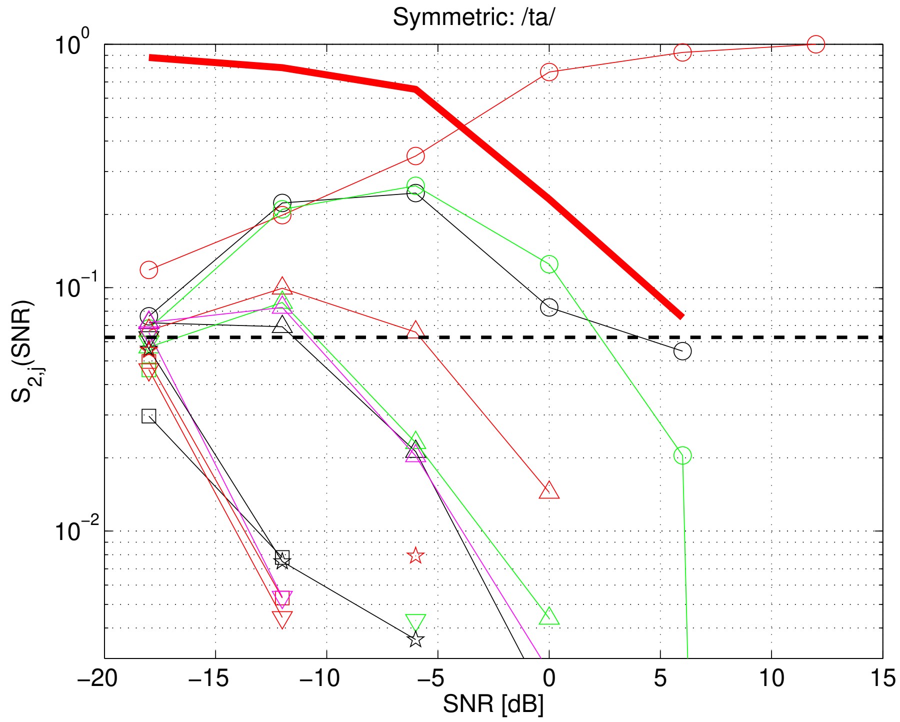
UNVOICED VOICED NASAL

RESPONSE

- Confusion groups imply underlying elemental events

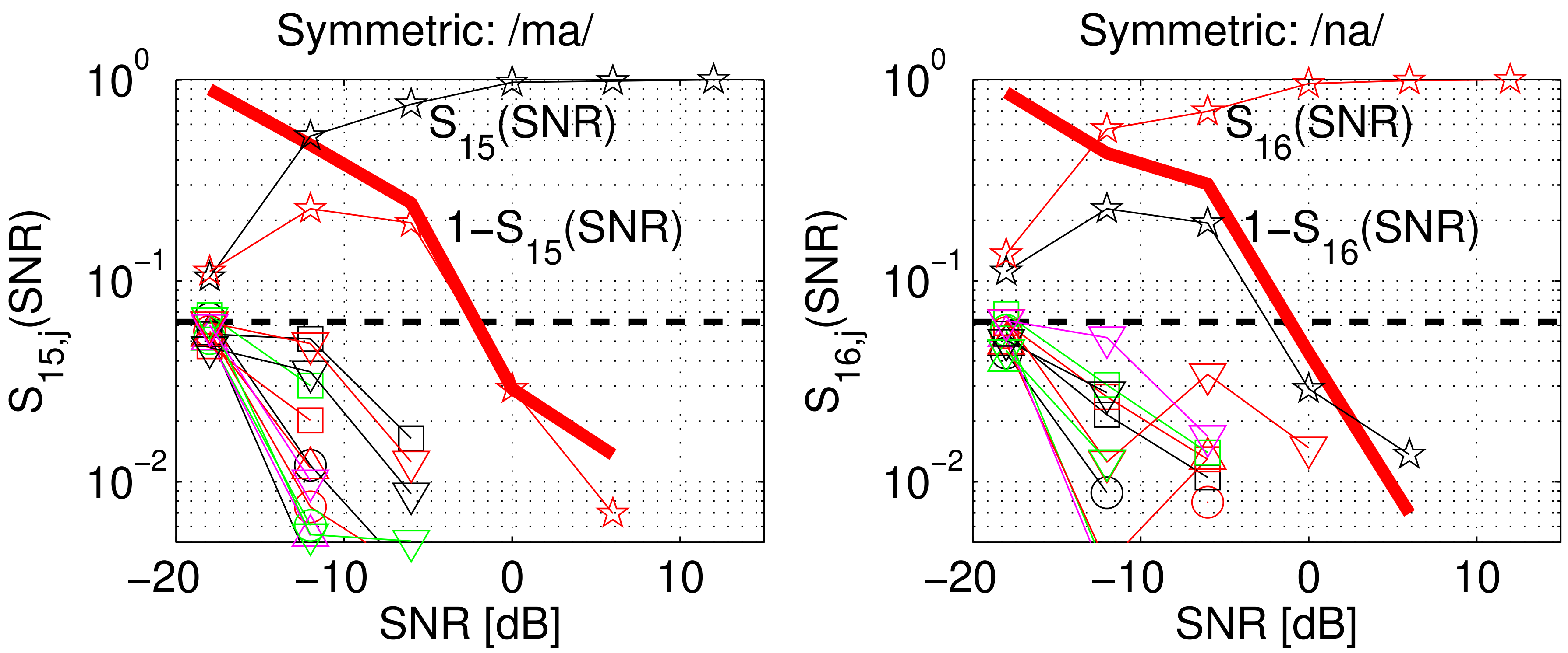
Case of /pa/, /ta/, /ka/ with /ta/ spoken

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



- How many events, and of what form?
- Plot of $\mathcal{A}_{i,j}(snr)$ for row i =2
- Solid red curve is total error $e_2 \equiv 1 - \mathcal{A}_{2,2} = \sum_{j\neq 2} \mathcal{A}_{2,j}$

The case of /ma/ vs. /na/

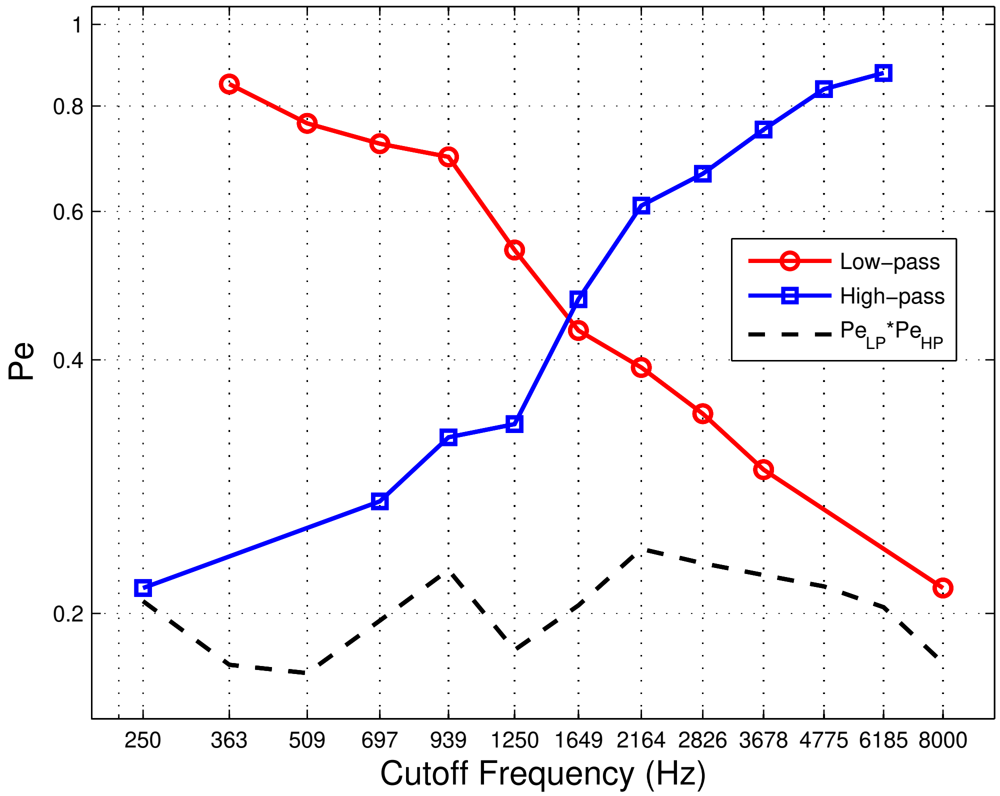


- This 2-group of sounds is closed since $\mathcal{A}_{/ma/,/ma/}(\mathit{SNR}) + \mathcal{A}_{/ma/,/na/}(\mathit{SNR}) \approx 1$
  - There can be only 1 event
  - Solid red curve is the total error: $e_i \equiv 1 - \mathcal{A}_{i,i} = \sum_{j \neq i} \mathcal{A}_{i,j}(\mathit{SNR})$

Fletcher's Lopass/Hipass result

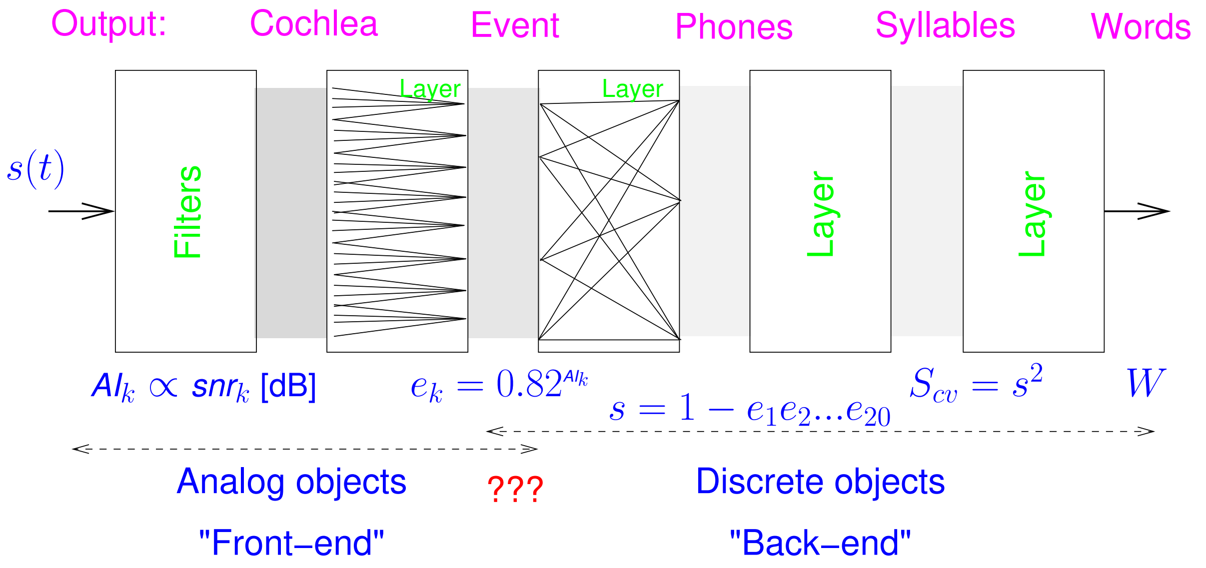
- The AI is based on the *band error product formula*

$$1 - P_c(SNR) \equiv e_{total}(SNR, f_c) = e_{lp}(SNR, f_c) e_{hp}(SNR, f_c) \quad (2)$$



Probabilistic measures of recognition

- k^{th} band articulation index: $AI_k = \frac{10}{30}\log(1 + c^2 snr_k^2)$
  - $c = 2$, $k = 1 \cdots K$ with $K = 20$
- Band (event) error: $e_k = e_{\min}^{AI_k/K} = \sqrt[20]{0.02}^{AI_k} = 0.822^{AI_k}$
- The $AI \equiv \frac{1}{K}\sum_k AI_k$, **?**
- MaxEnt phone score: $s = 1 - e_1 e_2 \ldots e_K = 1 - e_{\min}^{AI}$

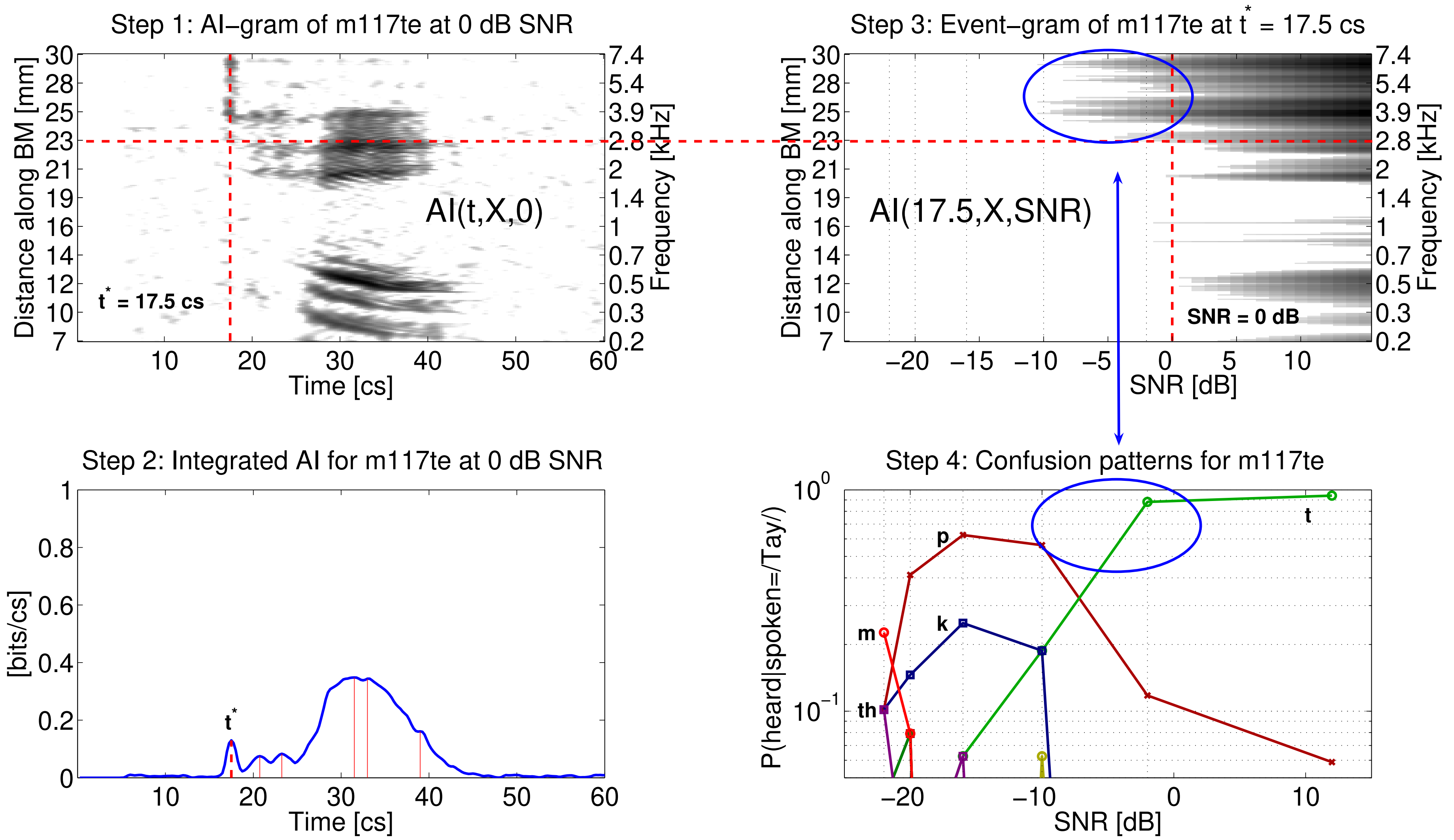


- MaxEnt syllable model score: $S_{cv} = s^2$, $S_{cvc} = s^3$

How can we find events?

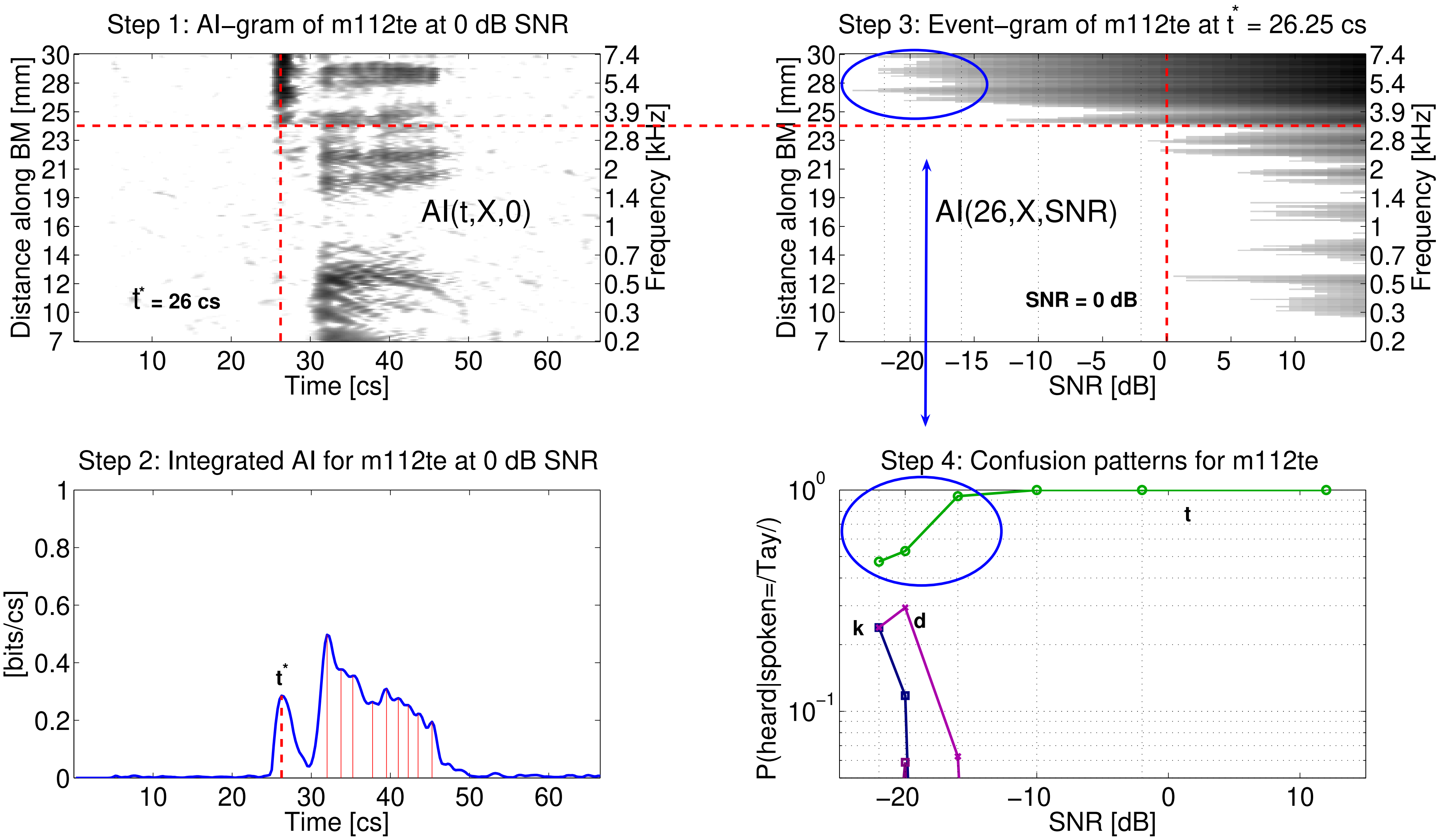
- A *4-Step analysis* relates confusions to an audibility measure (???):
- Modification of speech sounds
  - We developed a tool based on the Short-Time Fourier Transform (STFT) (?) that allows us to selectively:
  - Mask with noise specific time and frequency regions
    - so that this specific part of the speech becomes inaudible
    - selectively amplify specific regions
    - to increase intelligibility
    - We will present audio examples of original and modified sounds

m117/tɛ/ in speech-weighted noise



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

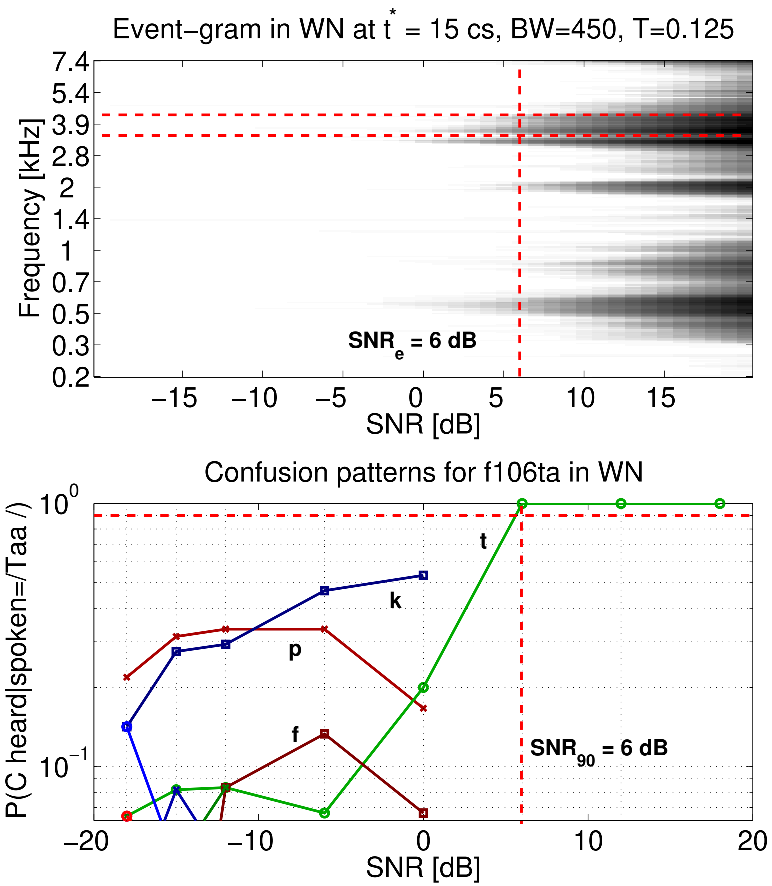
m112/tɛ/ in speech-weighted noise



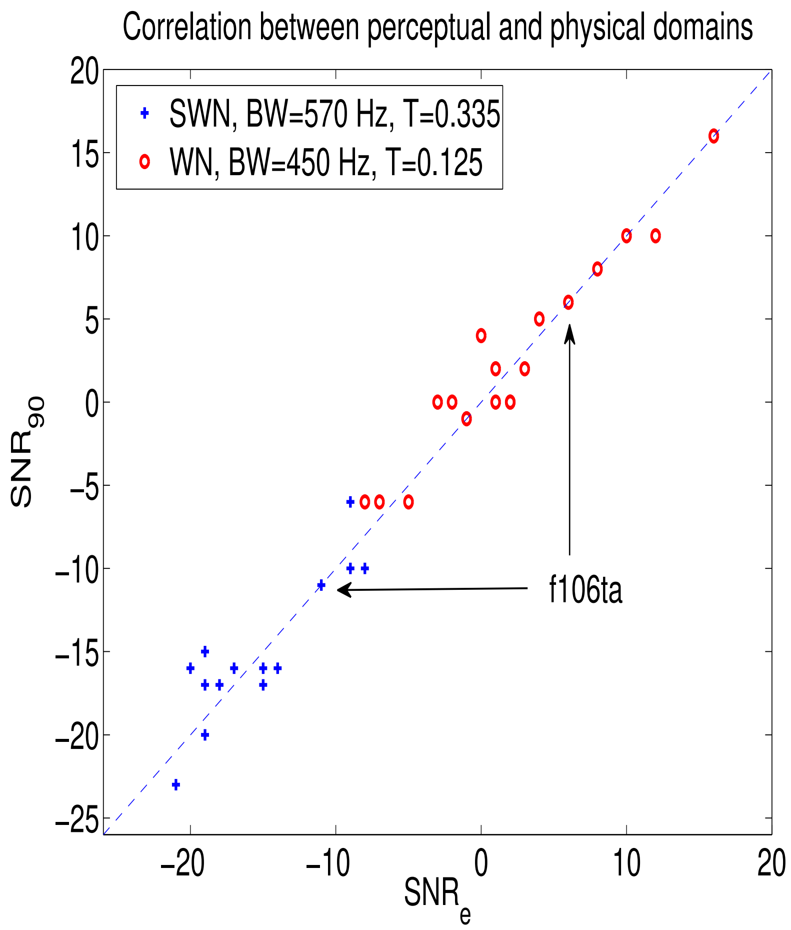
- /t/ confusion threshold at $P_c(\mathit{SNR}^* = -16) = 0.9$ correlated to Event-gram

Correlations of /t/ events

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

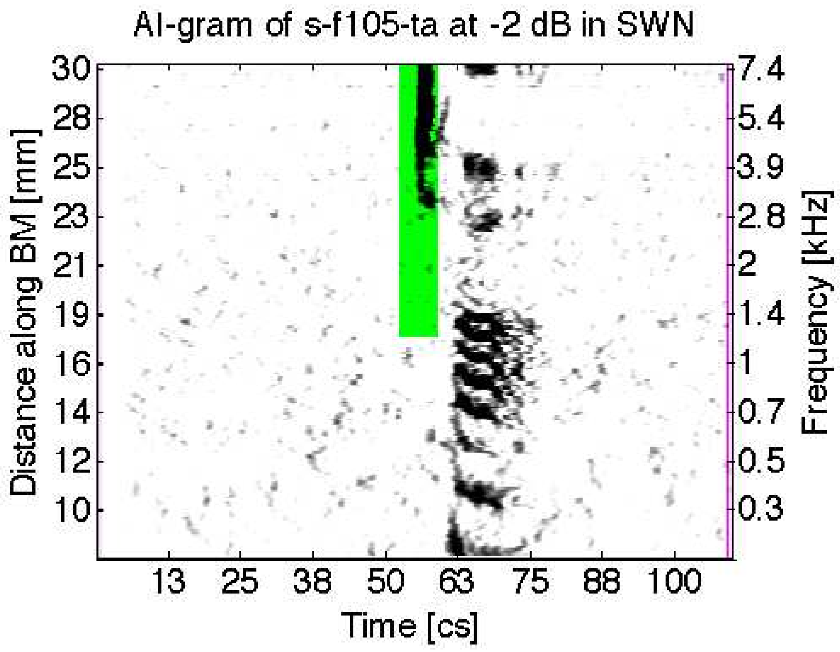


(a)

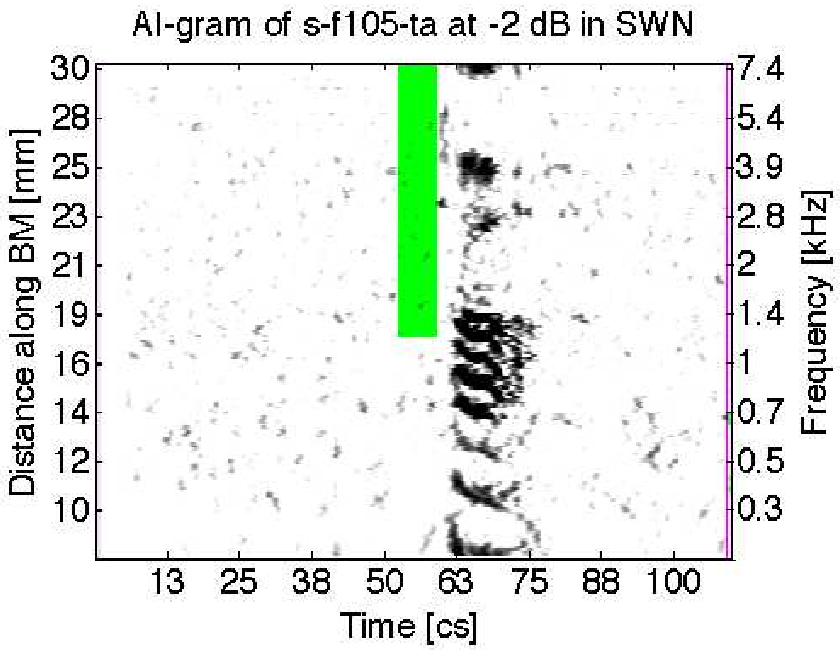


(b)

Masking of /tɑ/ timing cue



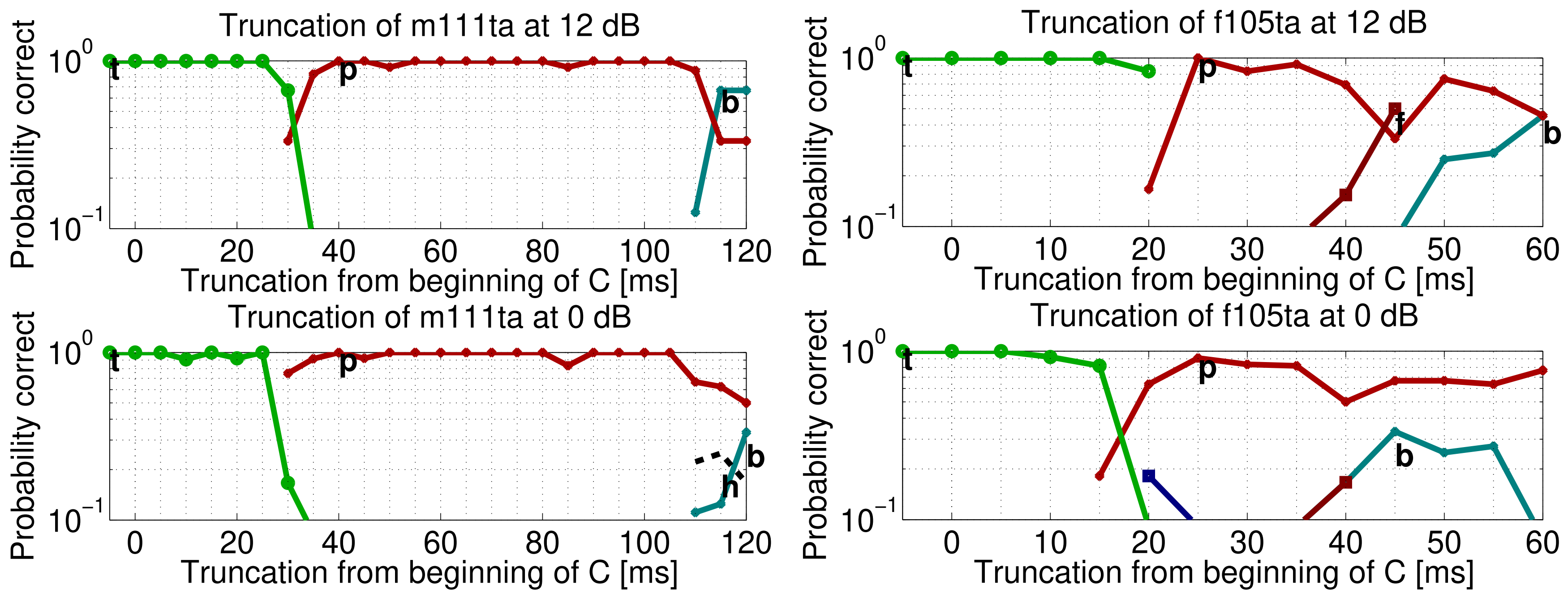
(c) Original /tɑ/



(d) Modified /tɑ/

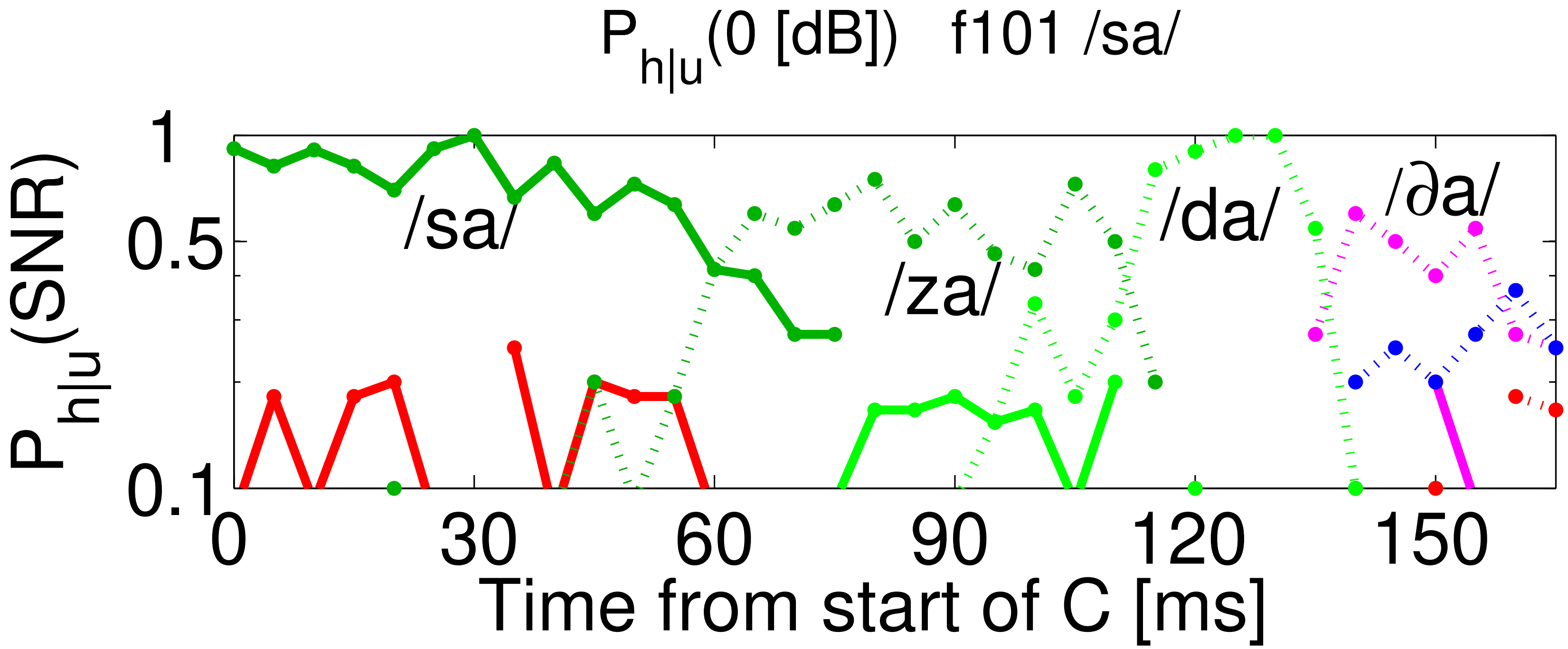
- When the /t/ burst is masked by noise, the perception morphs to /p/
- DEMO 4

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 previous studies **?**, and our more extensive results

Truncation of f101 /sa/

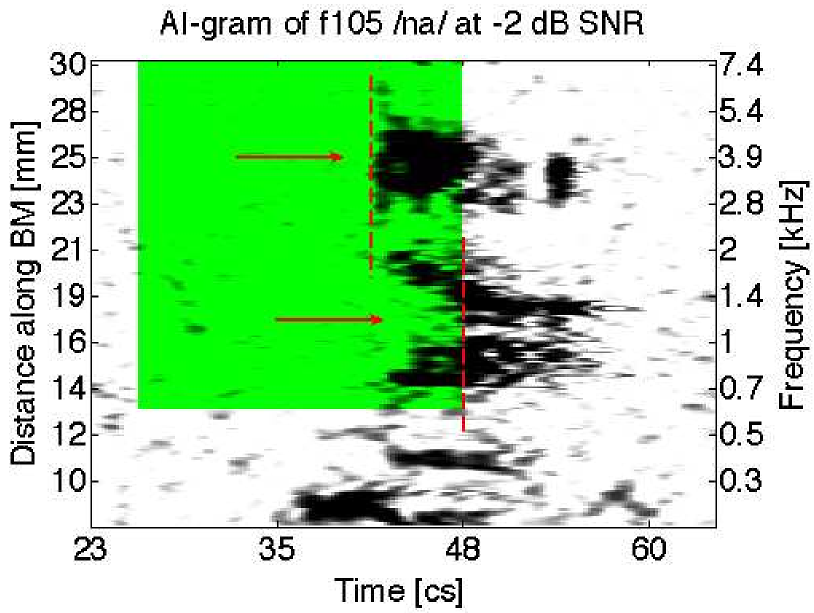


- This represents the normal hearing responses to a truncated /sɑ/, from the start of the consonant
- Morphing from /sɑ/ to /zɑ/ to /dɑ/ to /ðɑ/
- Duration seems to be a fricatives event

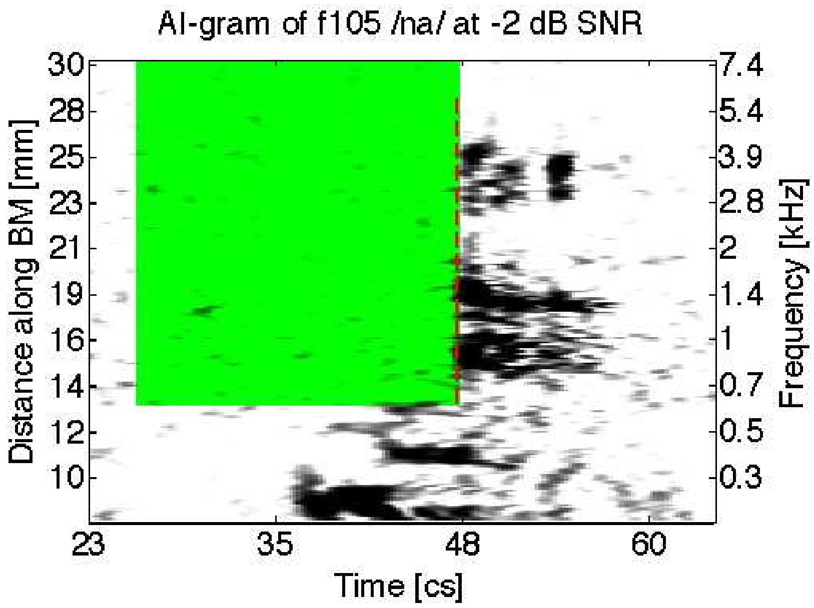
/mɑ/- /nɑ/ discrimination

- /nɑ/ recognition from /mɑ/ relies on a ≈ 50 ms delay formed from the F_1 and F_2 collision
- When we edit the speech so that the onset is simultaneous above 0.6 kHz, the /nɑ/ is robustly and naturally heard as /mɑ/
- METHODS: 9 listeners evaluated these sounds in open response random trial experiment.

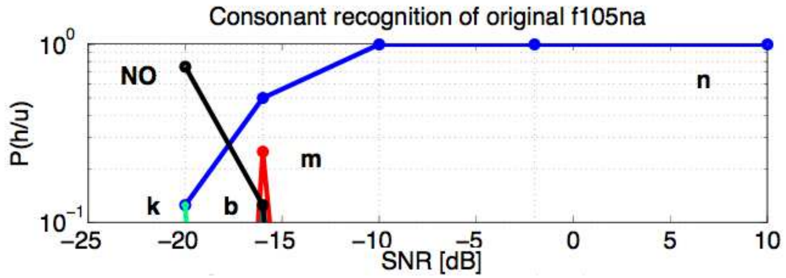
Deletion of /nɑ/ timing cue

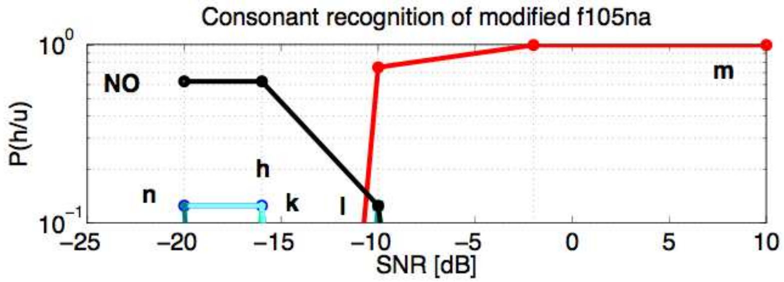


(e) Original /nɑ/



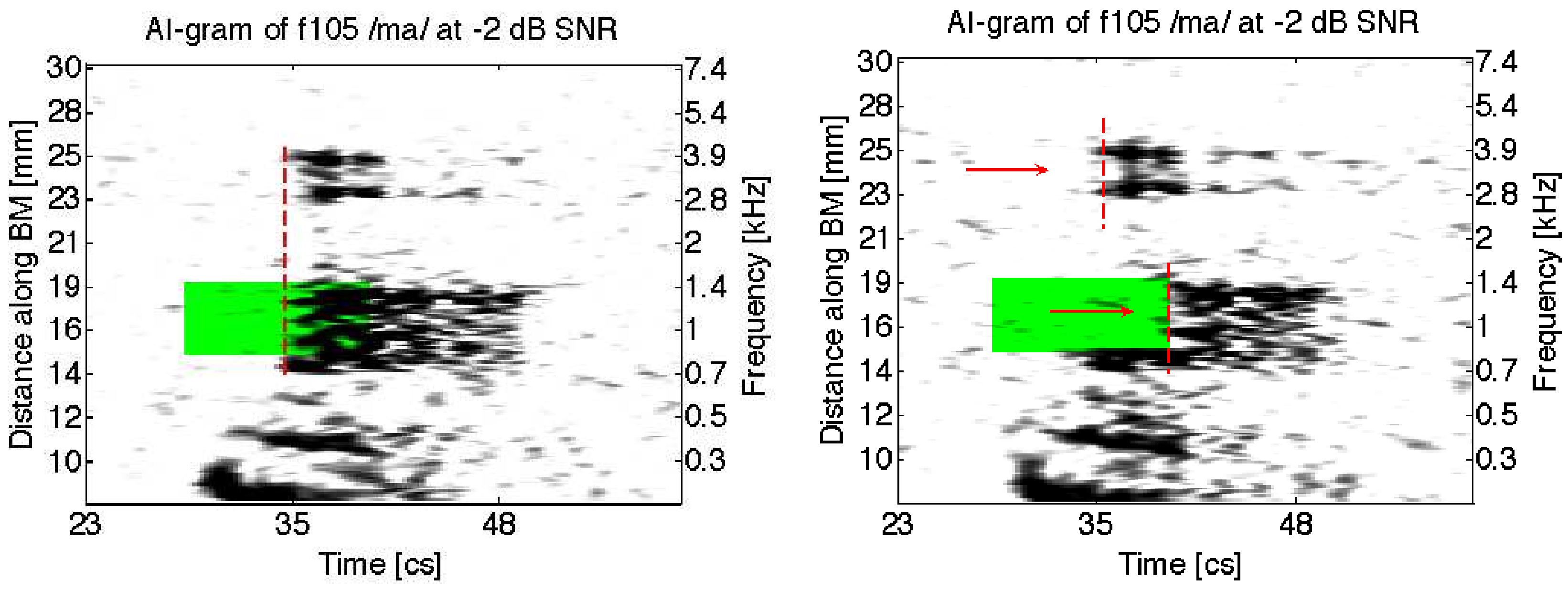
(f) Modified /nɑ/



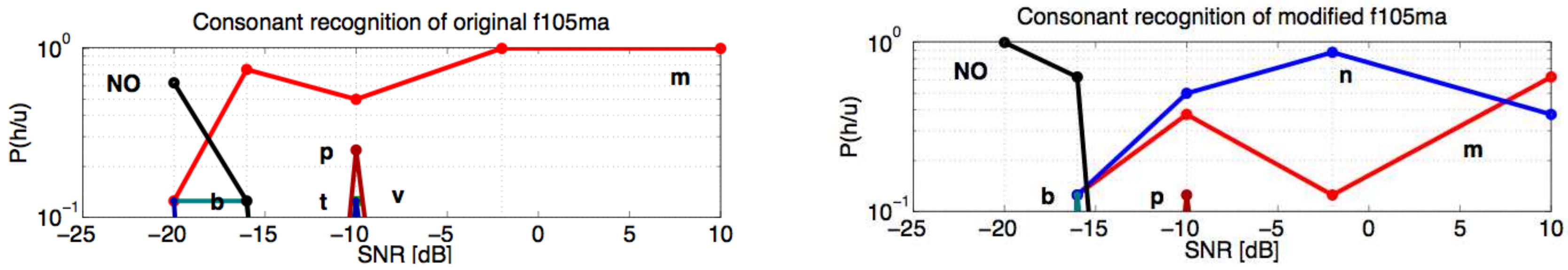


- DEMO

Creation of /nɑ/ timing cue

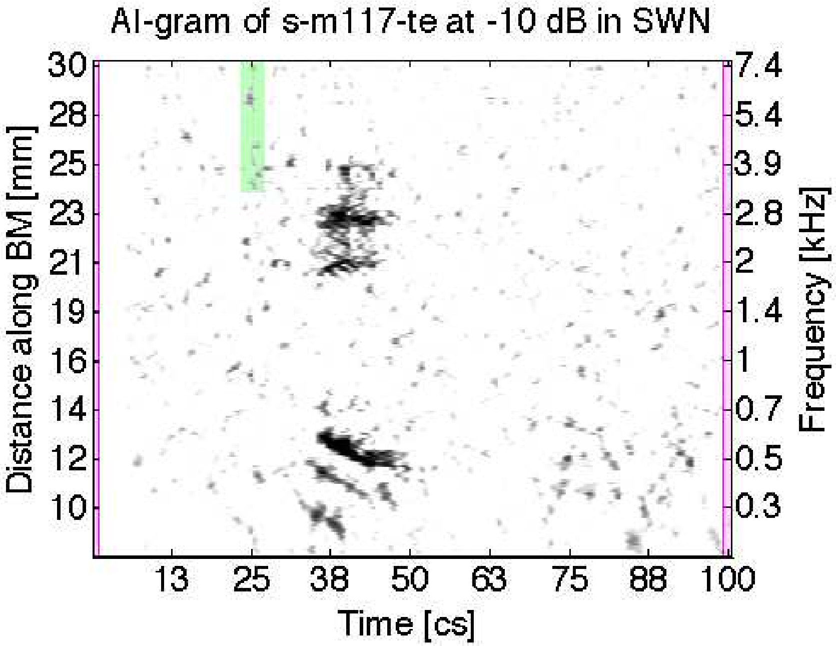


(g) Original /mɑ/ (h) Modified /mɑ/

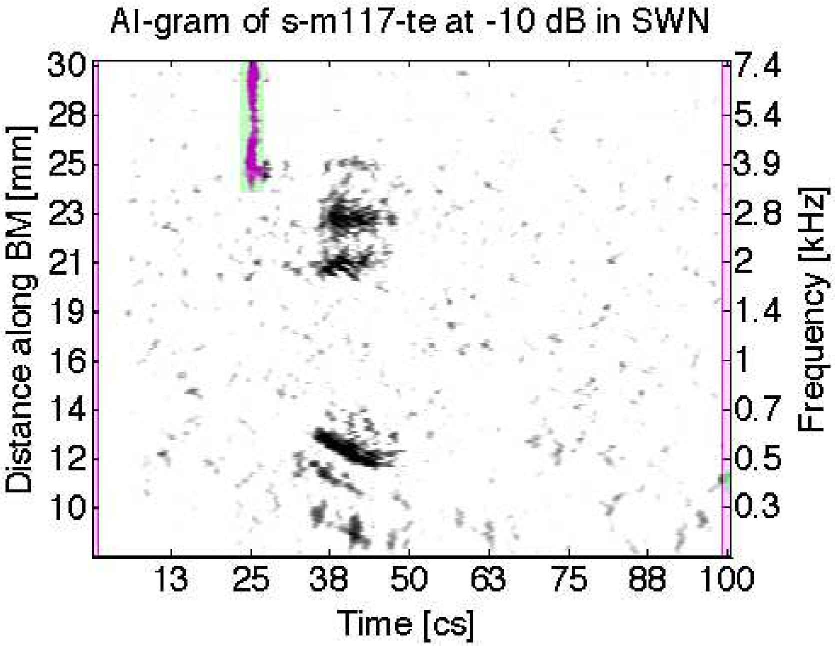


- DEMO

Enhancement of /tɛ/ event



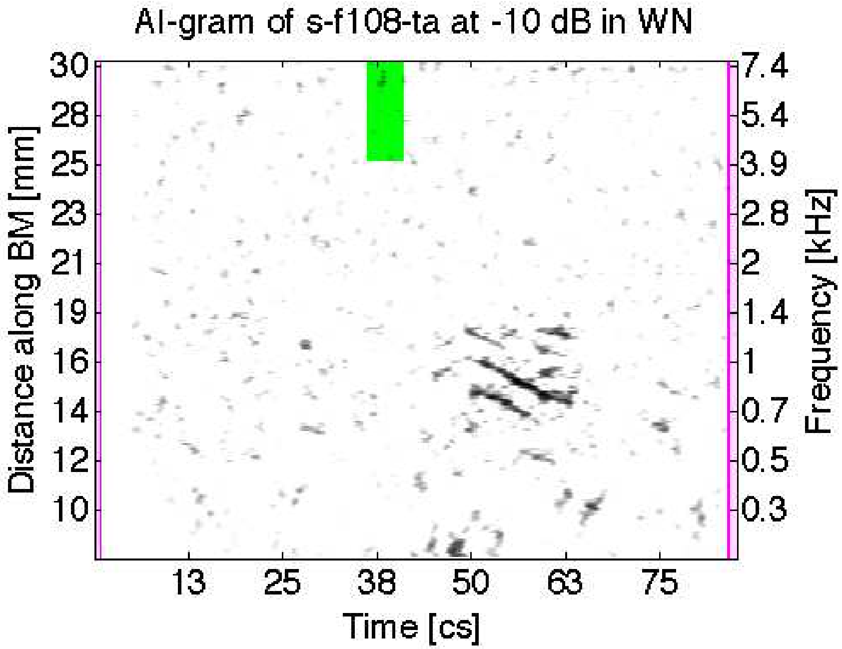
(i) Original /tɛ/



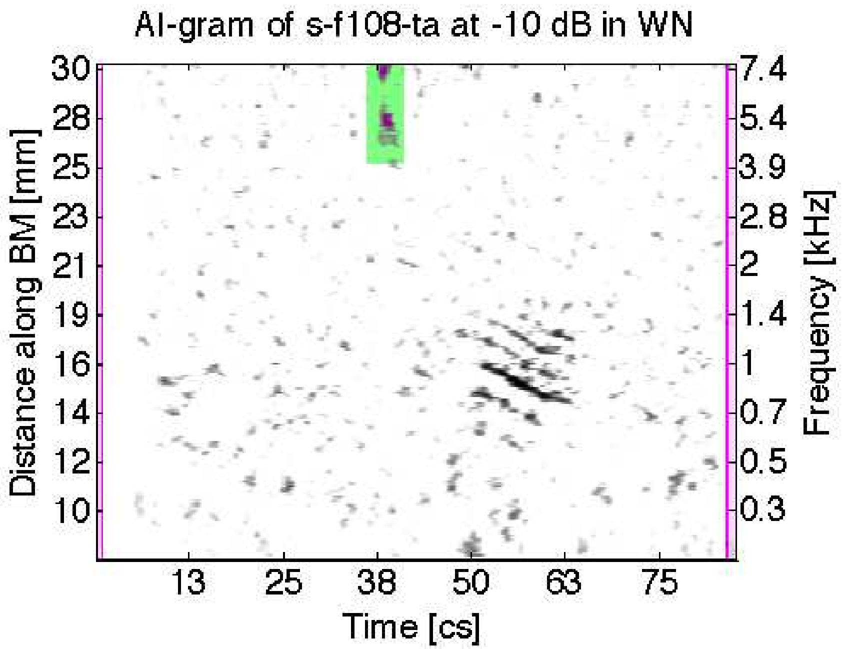
(j) Modified /tɛ/

- 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



(k) Original /tɑ/



(l) 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

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