Cochlear modeling and its role in human speech recognition

Miller Nicely confusions and the articulation index

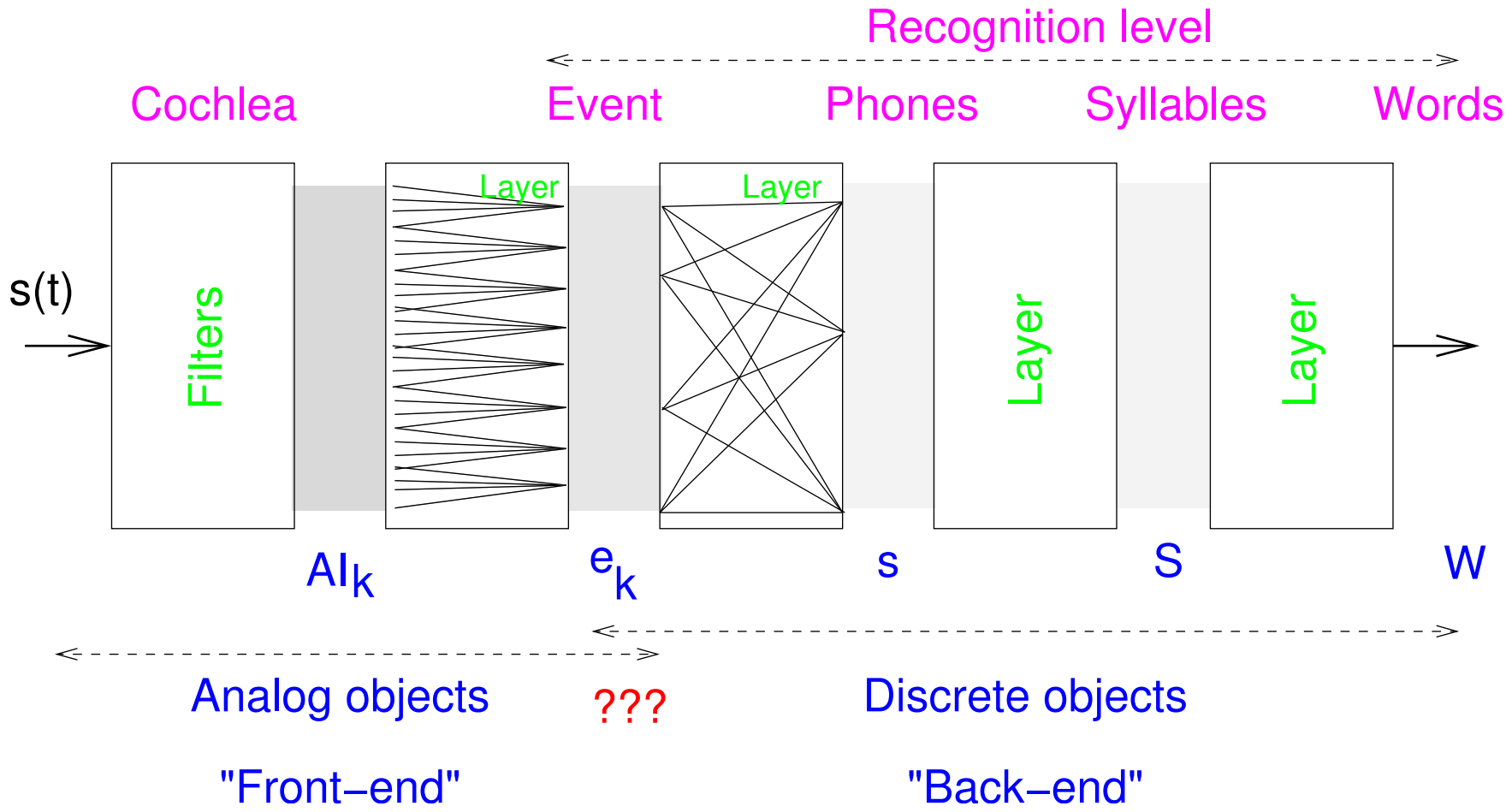
Jont Allen

Univ. of IL,

Beckman Inst., Urbana IL

Model of human speech recognition (HSR

- The research goal is to identify *elemental HSR events*
  - An event is defined as a *perceptual feature*



Modeling MaxEnt HSR

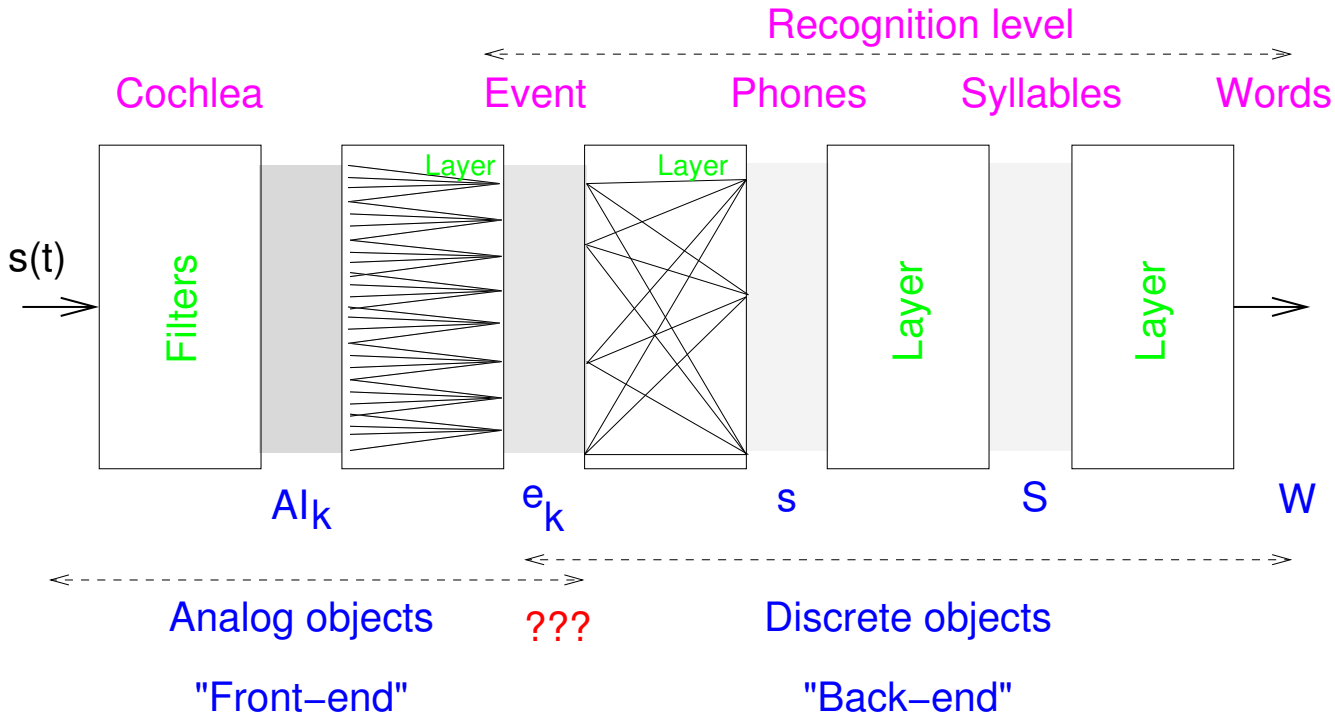
- Definition of MaxEnt (a.k.a. "nonsense") syllables:
  - A fixed set of { C,V } sounds are drawn from the language of interest
  - A uniform distribution for each C and V is required, to minimize syllable context ($\Rightarrow$ MaxEnt)
  - MaxEnt CV or CVC syllable score: $S_{cv} = s^2$, $S_{cvc} = s^3$
  - MaxEnt syllables first described in Fletcher, 1929
- A set of meaningful words is not MaxEnt
  - Modeling non-MaxEnt syllables require the context models of Boothroyd, 1968; Bronkhorst et al., 1993
- Fletcher's 1921 AI *band independence model*:

$$s(AI) \equiv 1 - e = 1 - e_1 e_2 e_3 \ldots e_{20} = 1 - e_{\min}^{AI} \qquad (1)$$

accurately models MaxEnt HSR (Allen, 1994)

Probabilistic measures of recognition

- Recognition measures (MaxEnt ≡ Maximum Entropy):
  - k^{th} band articulation index: $AI_k \propto \log(snr_k)$
  - Band recognition error: $e_k = e_{\min}^{AI_k/K}$ with $K = 20$
  - MaxEnt phone score: $s = 1 - e_1 e_2 \ldots e_K = 1 - e_{\min}^{\overline{AI_k}}$
  - MaxEnt syllable score: $S_{cv} = s^2$, $S_{cvc} = s^3$



What is an *elemental event*?

- 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	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

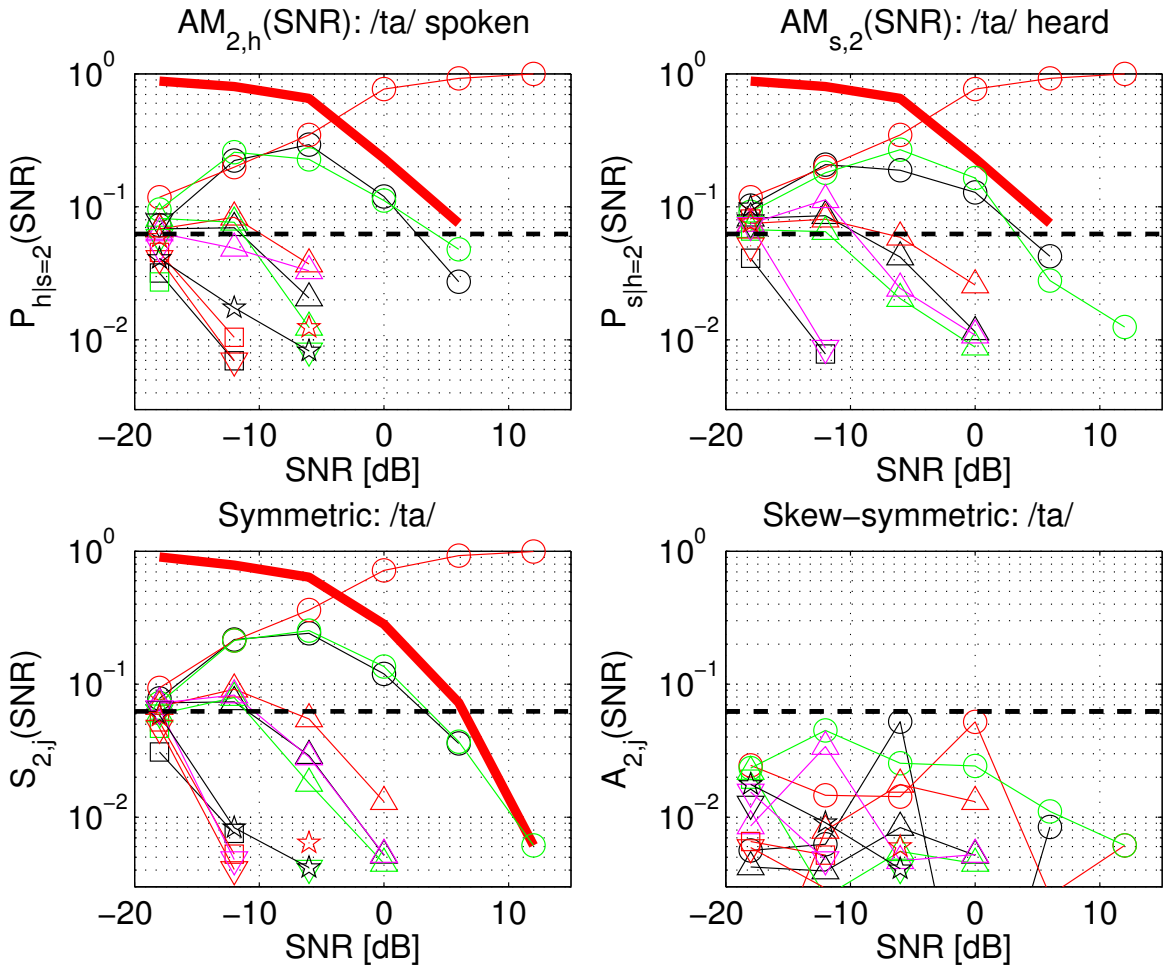
UNVOICED VOICED NASAL

RESPONSE

- Confusion groups formed in $\mathcal{A}(snr)$

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

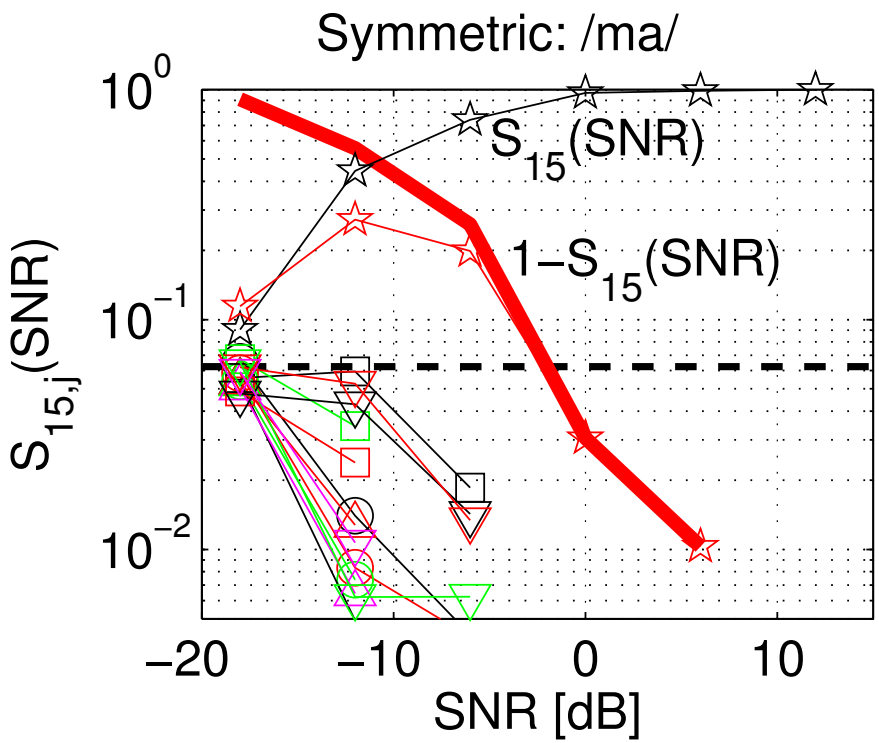
- Plot of $\mathcal{A}_{i,j}(snr)$ for row i =2 and column j =2

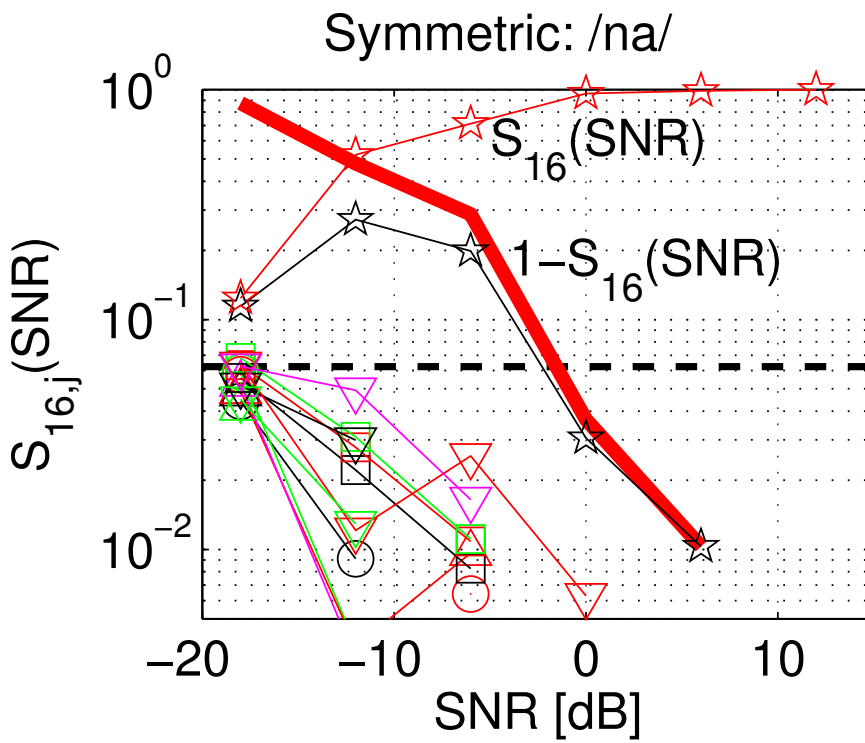


- Solid red curve is total error $e_2 \equiv 1 - \mathcal{A}_{2,2} = \sum_{j \neq 2} \mathcal{A}_{2,j}$

The case of /ma/ vs. /na/

- Plots of $S(\mathit{snr}) \equiv \frac{1}{2}(\mathcal{A} + \mathcal{A}^t)$, /ma/, /na/ spoken
- Solid red curve is total error $e_i \equiv 1 - S_{i,i} = \sum_{j \neq i} S_{i,j}$





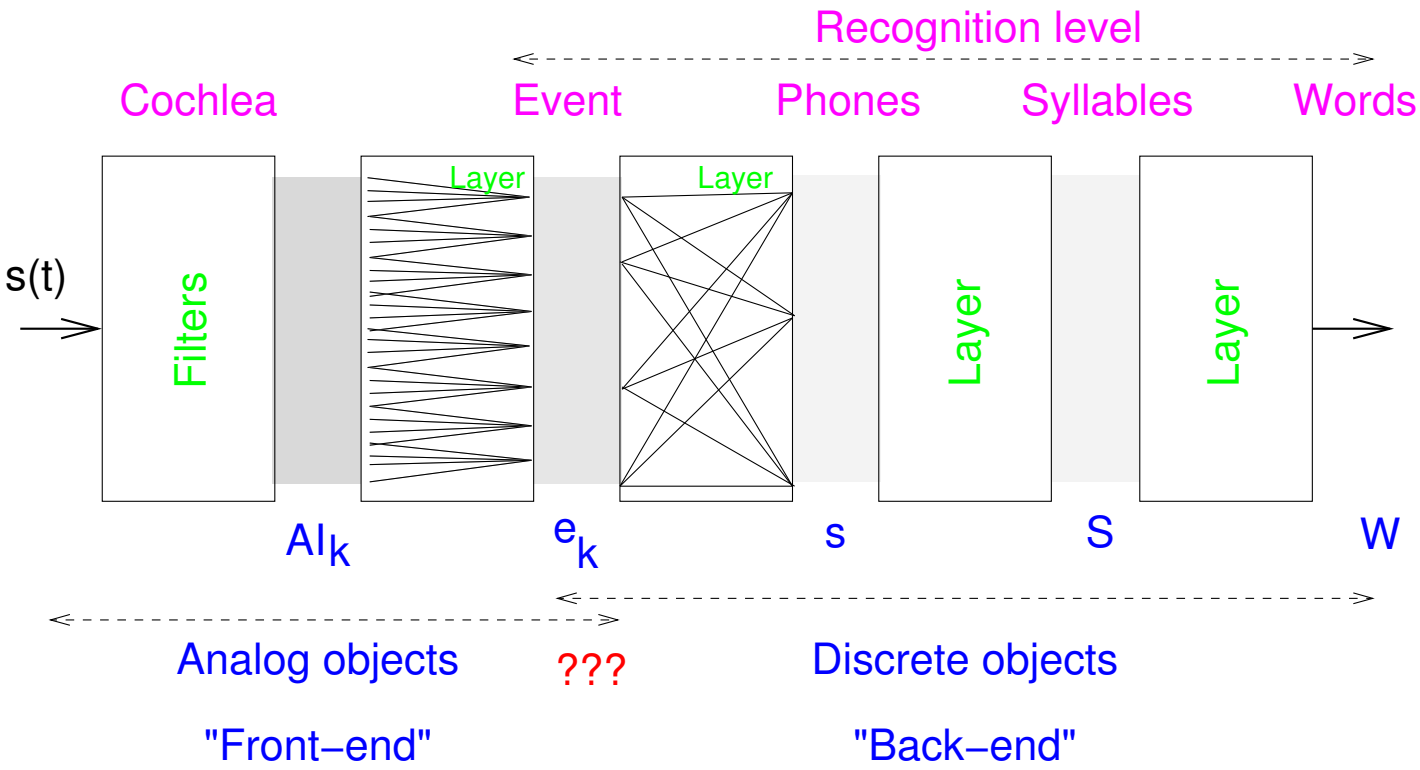
- This 2-group of sounds is closed since
  $S_{/ma/,/ma/}(\mathit{snr}) + S_{/ma/,/na/}(\mathit{snr}) \approx 1$

Definition of the *AI*

- Let AI_k be the k^{th} band cochlear channel capacity

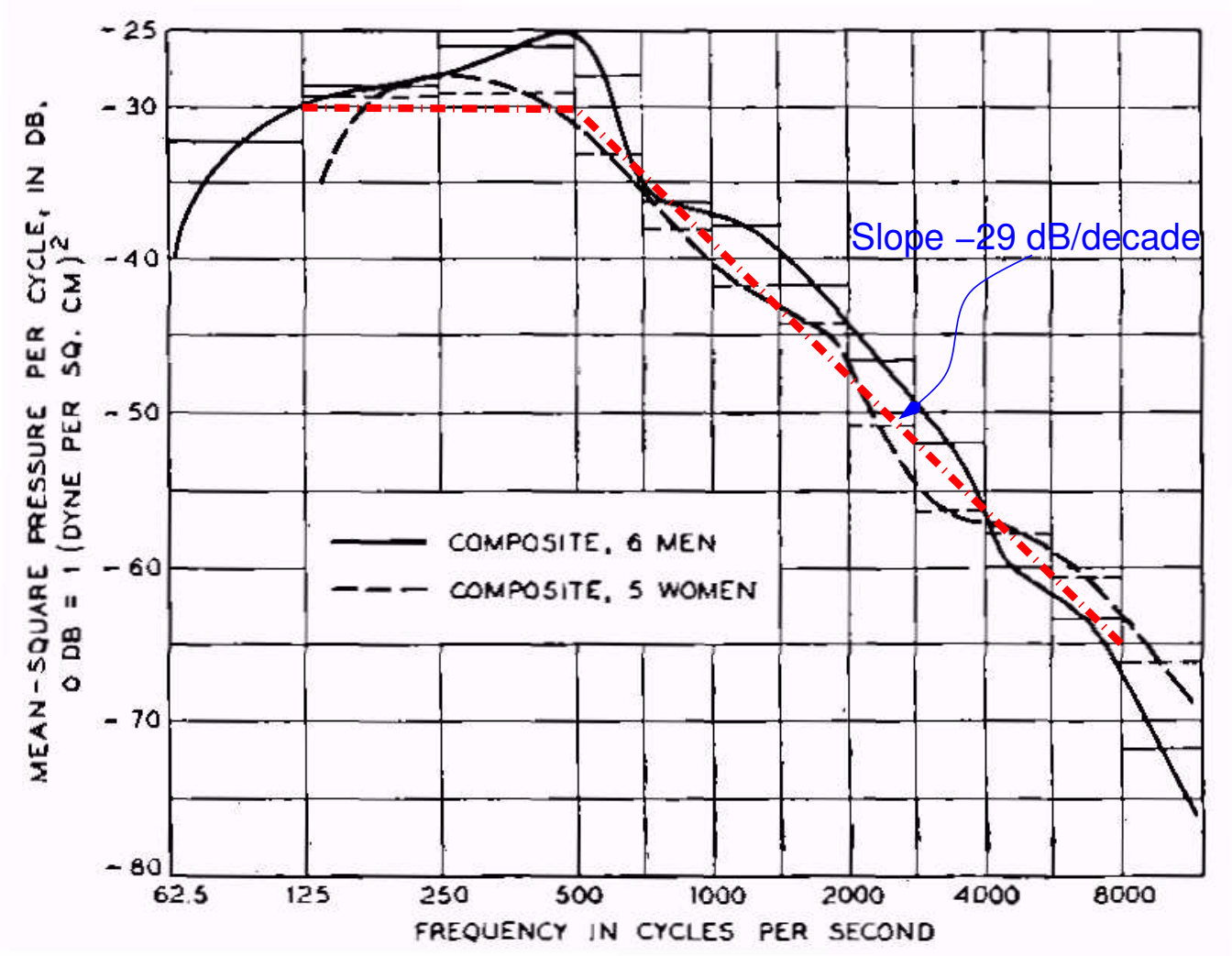
$$AI_k \equiv \frac{10}{30}\log_{10}(1 + c^2 snr_k^2), \text{ with } c = 2, AI_k \leq 1,$$

then $AI \equiv \frac{1}{K}\sum AI_k$ (Allen, 1994; Allen, 2004)



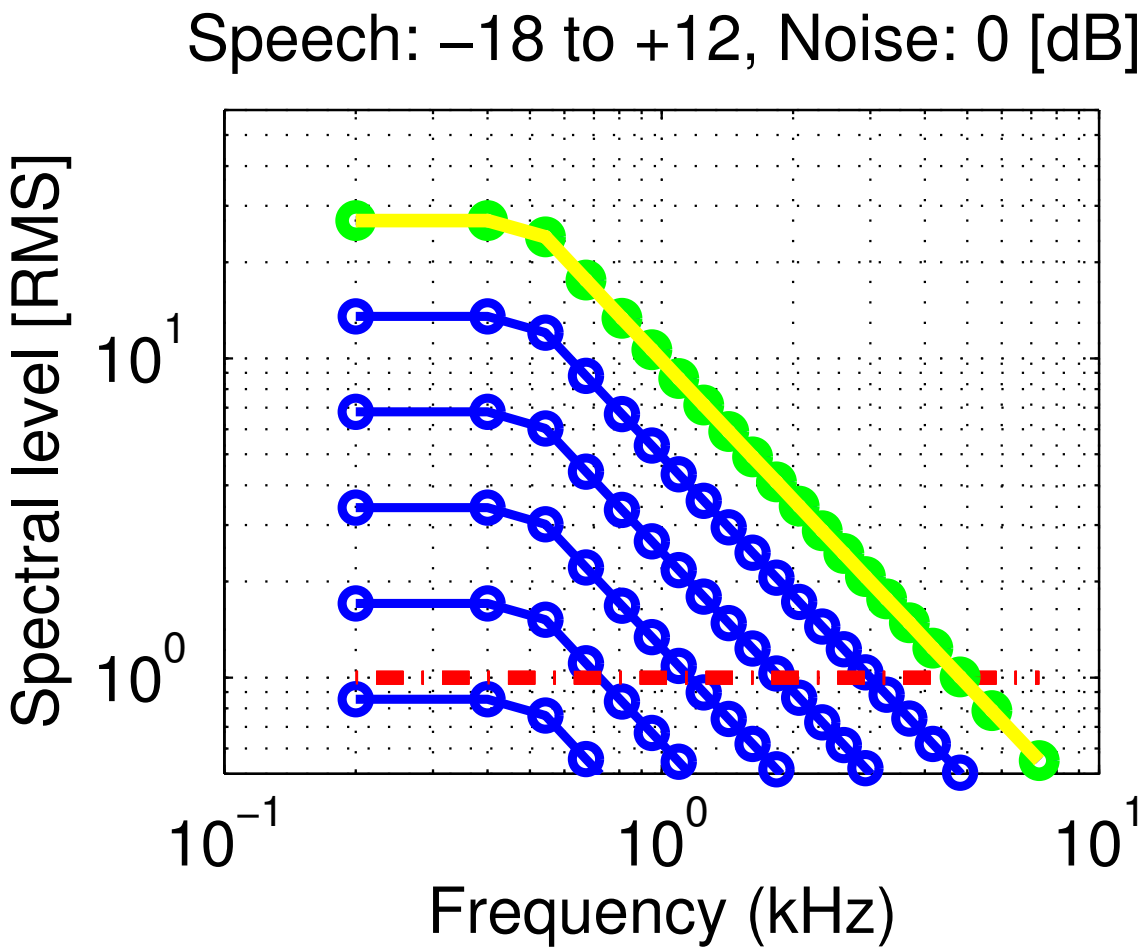
Long-term spectrum of female speech

- Speech spectrum for female speech Dunn and White
  - Dashed red line shows the approximation



Conversion from SNR to AI

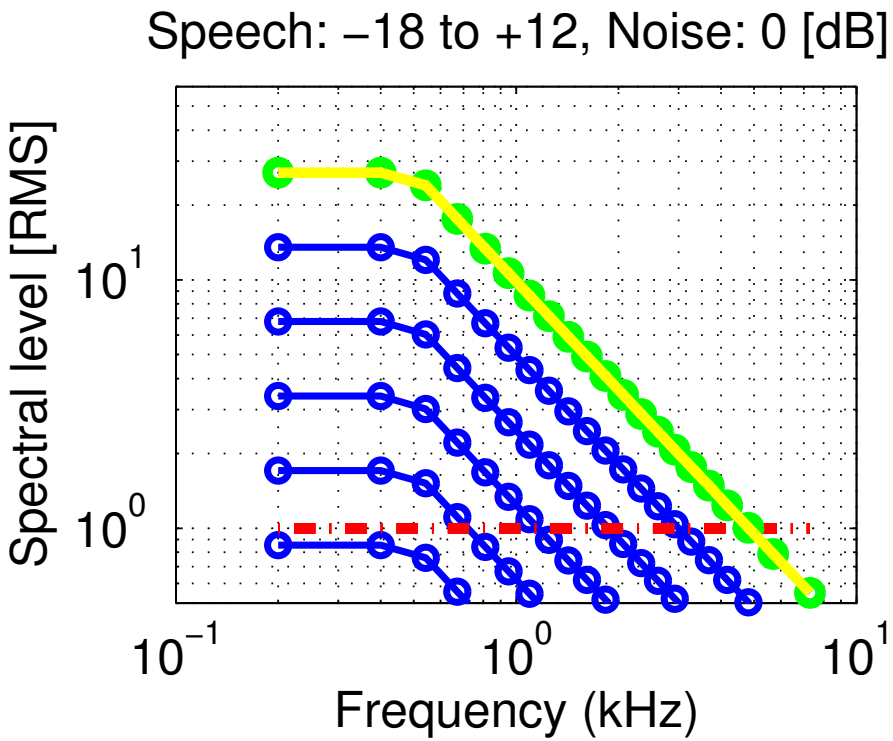
- Spectra for SNRs of [-18, -12, -6, 0, 6, 12] dB

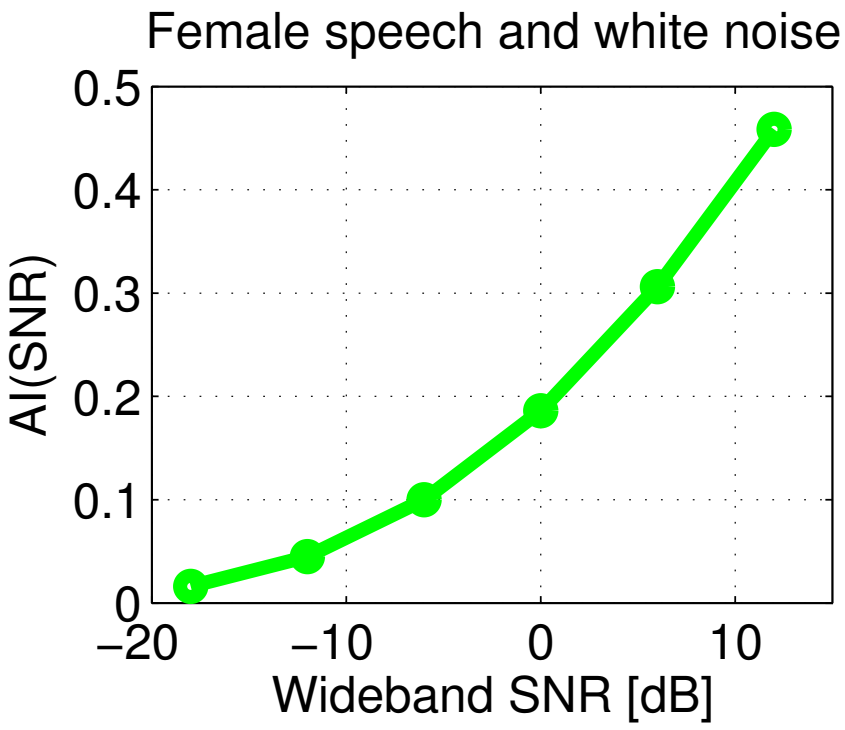


- snr_k is determined for $K = 20$ cochlear critical bands

AI(SNR) for Miller Nicely's data

- AI(SNR) computed from the Dunn and White spectrum



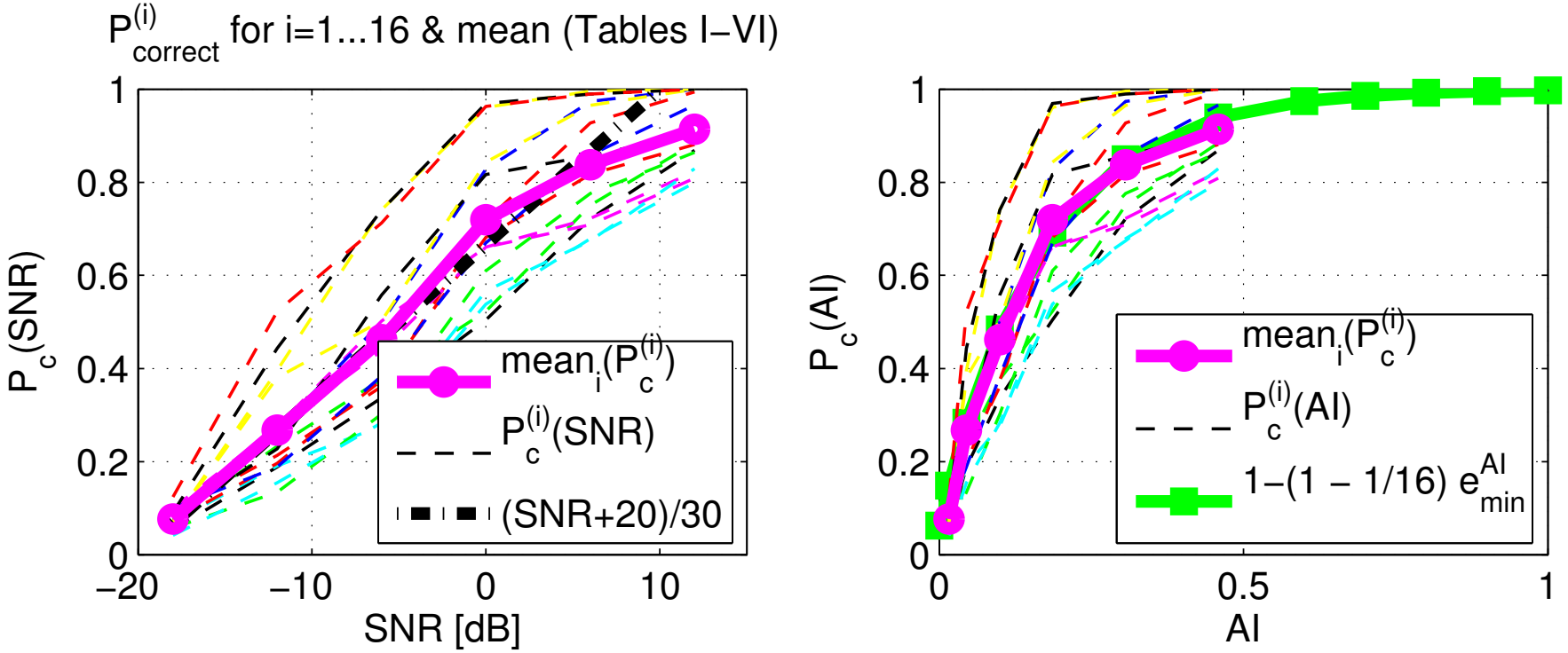


- Conversion from SNR_k to AI [Allen 2005, JASA]:

$$AI \equiv \overline{AI_k} = \frac{1}{60}\sum_{k=1}^{20}\log_{10}(1+c^2 snr_k^2), \quad c=2, AI_k \leq 1$$

MN16 and the AI model

- $P_c^{(i)}(AI)$ for the i^{th} consonant and $P_c(AI) \equiv \frac{1}{16}\sum_i P_c^{(i)}(AI)$
- $P_c(AI) = 1 - (1 - \frac{1}{16})e_{\min}^{AI}$ is the chance–corrected model



- Band independence seems a perfect fit to MN16 !!

Log-error probability

- Band-independence, corrected for chance,

$$P_e(AI, \mathcal{H} = 4) \equiv 1 - P_c(AI) = (1 - 2^{-\mathcal{H}}) e_{\min}^{AI}$$

- This suggests linear log-error plots vs. AI:
  - MN16 CVs $\log(P_e^{(i)}) = \log(1 - \frac{1}{16}) + \log(e_{\min})AI$
- This relation is of the form

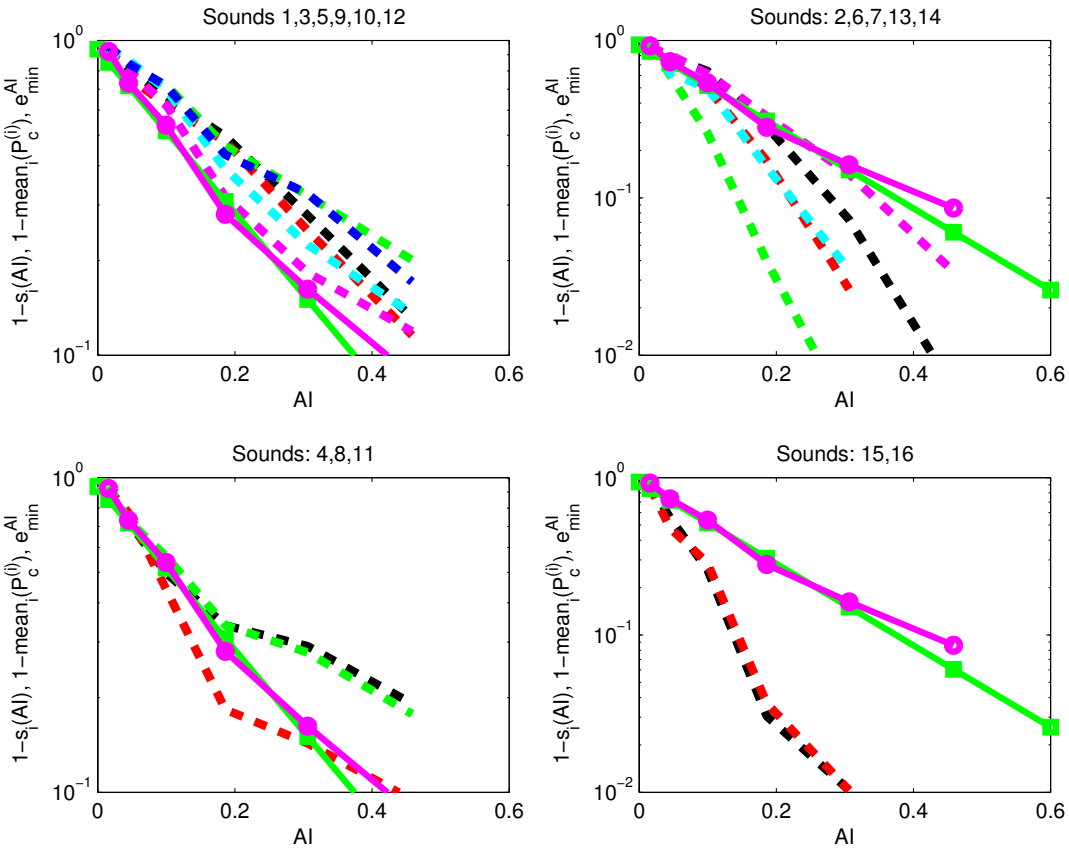
$$Y = a + bx$$

- Plots of $\log(P_e)(AI)$ vs. AI provide a nice test of Fletcher's "band independence" model
- Individual sounds $P_e^{(i)}(AI)$ from MN16 may be tested for band independence.

Testing the band-independence model

- CV log-error model:

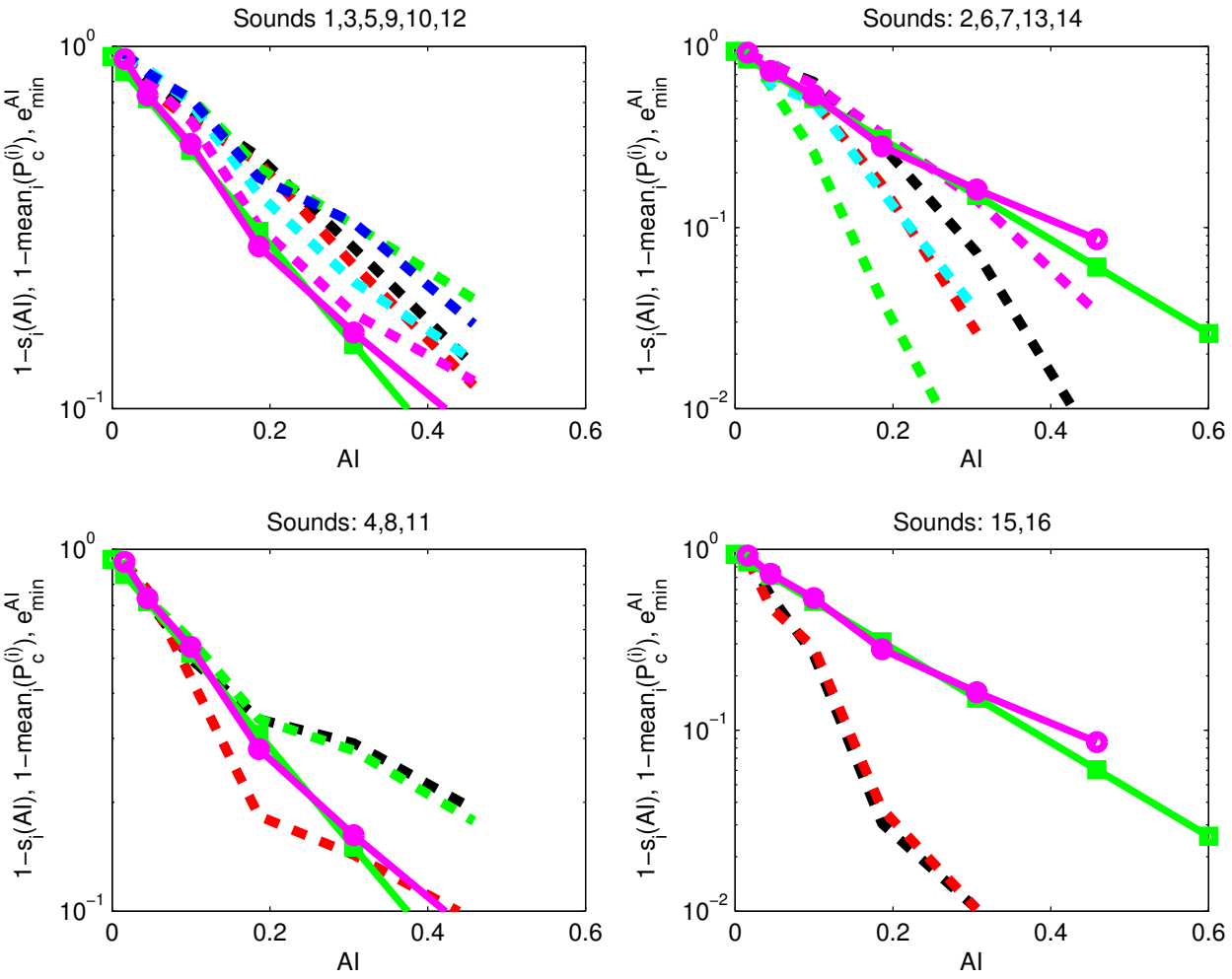
$$\log\left(P_e^{(i)}(AI)\right) = \log\left(1 - 2^{-4}\right) + \log\left(e_{\min}^{(i)}\right) AI$$



- AI Model; MN16 $P_c(AI)$; - - -, - - -: CV sounds

Testing the band-independence model

- Sounds: [1, 2, 3, 5, 8, 9, 10, 12, 13, 14] pass
- Sounds: [4, 8, 11, 15, 16] fail (nonlinear wrt *AI*)

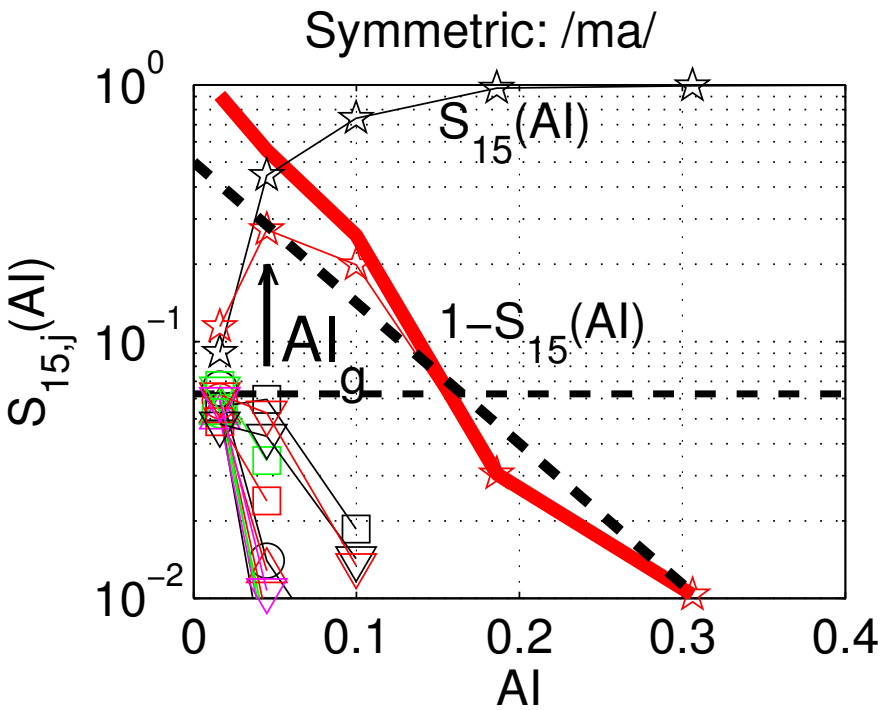


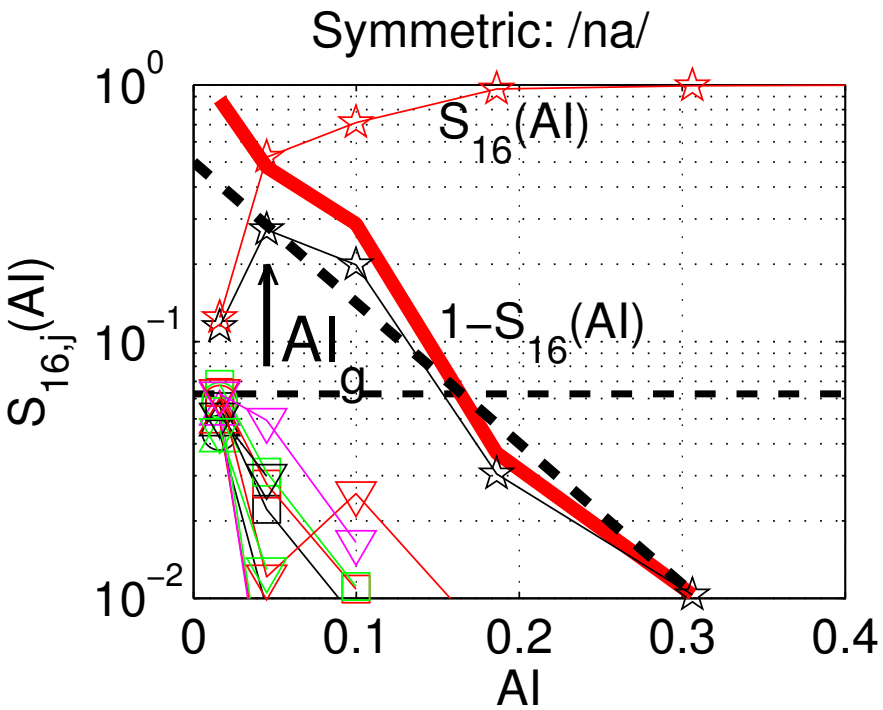
/ma/ and /na/ vs. AI

- The multichannel model is valid for the nasals

$$S_{i,j}(AI) \approx \delta_{i,j} + (-1)^{\delta_{i,j}} \left(1 - 2^{-\mathcal{H}_g}\right) \left(e_{\min}^{(i)}\right)^{AI} \quad \text{for } AI > AI_g$$

  - ${e_{\min}}^{(i)} \equiv 1 - S_{i,i}|_{AI=1}$
  - $\mathcal{H}_g = 1$ [bit] (i.e., a 2-group)

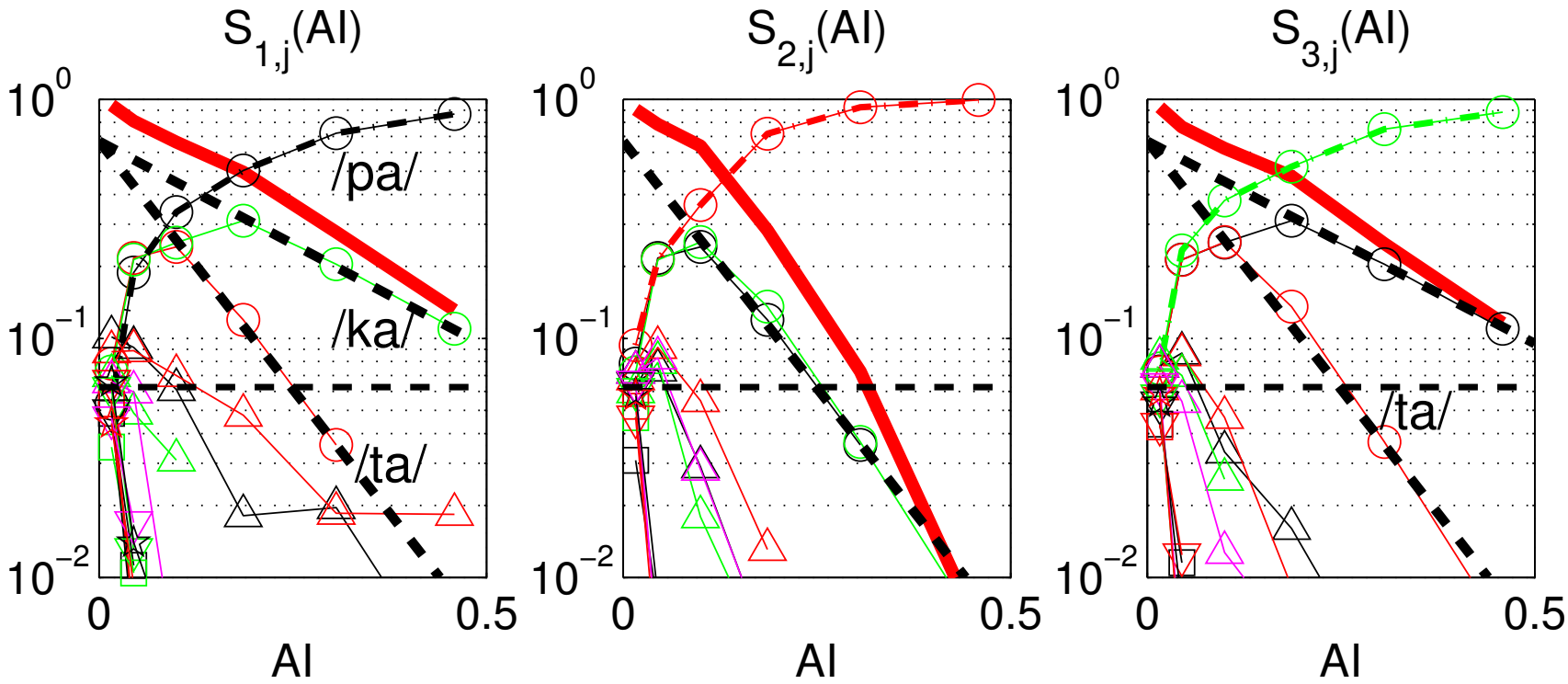




Case of /pa/, /ta/, /ka/ vs. AI

- Fletcher's multichannel model is valid for $i = 1, 2, 3$:

$$S_{i,j}(AI) \approx \delta_{i,j} + (-1)^{\delta_{i,j}} \left(1 - 2^{-\mathcal{H}_g}\right) \left(e_{\min}^{(i,j)}\right)^{AI} \quad \text{for } AI > AI_g$$



- Band independence holds for the [/pa/, /ta/, /ka/] group: $\mathcal{H}_g = \log_2(3)$, $AI_g = 0.15$ and $e_{\min}{}^{(i,j)}$ depends on i, j.

Conclusions I

- The AI predicts above chance performance near -20 dB
- HSR performance saturates near 0 dB SNR
- No overlap in ASR vs. HSR
  - ASR chance performance near 0 dB
  - ASR performance saturates near +20 dB SNR

Conclusions II

- Fletcher's AI theory is based on band independence

$$e(\mathit{snr}) = e_{\min}^{\frac{1}{K}\mathit{AI}_1} e_{\min}^{\frac{1}{K}\mathit{AI}_2} e_{\min}^{\frac{1}{K}\mathit{AI}_3} \ldots e_{\min}^{\frac{1}{K}\mathit{AI}_K} = e_{\min}^{\mathit{AI}} \qquad (2)$$

- Recognition of MaxEnt phones satisfy 2
- MaxEnt close set tests scores may be predicted by

$$P_c(\mathit{AI}, \mathcal{H}) = 1 - (1 - 2^{-\mathcal{H}}) {e_{\min}}^{\mathit{AI}}$$

- The first sign of $>$ chance performance is grouping
- The sound groups depend on the noise spectrum

Conclusions III

- The average over the 16 Miller Nicely consonant confusions $P_c^{(i)}(\mathit{snr})$ is accurately predicted by the Articulation Index:

$$\overline{P_c}(\mathit{snr}) \equiv \frac{1}{16}\sum_{i=1}^{16} P_c^{(i)}(\mathit{snr}) = 1 - \left(1 - \frac{1}{16}\right) e_{\min}^{AI}$$

- $P_c^{(i)}$ is the probability of correct identification of i^{th} CV
- $AI = \frac{10}{30}\sum_k \log_{10}(1 + 4\mathit{snr}_k^2)$ where k is the AI band index
- Chance error is given by $(1 - 1/16)$
- $e_{\min} \approx 0.005 \equiv \left.\frac{1-P_c(AI)}{(1-1/16)}\right|_{AI=1}$

Conclusions IV

- The *AI* model, corrected for 1/16 chance guessing, predicts the average Miller Nicely P_c quite well
- There are no free parameters in this model of $\overline{P_c}(snr)$
- For the MN experiment, the *AI* $\leq$ 0.5
- The *AI*(*snr*) is not linear in *snr*, due to the shape of the *snr(f)*
- The individual curves remain to be analyzed and modeled (if possible)
  - How can we explain the variance across sounds?

Conclusions V

- The MN data is very close to symmetric
- There are a few exceptions, but even for these, the asymmetric part is very small

Conclusions VI

- The off-diagonal PI(AI) functions linearly decompose error for i^{th} sound $P_c(i, AI)$, since $\sum_j P_{ij} = 1 \Rightarrow$

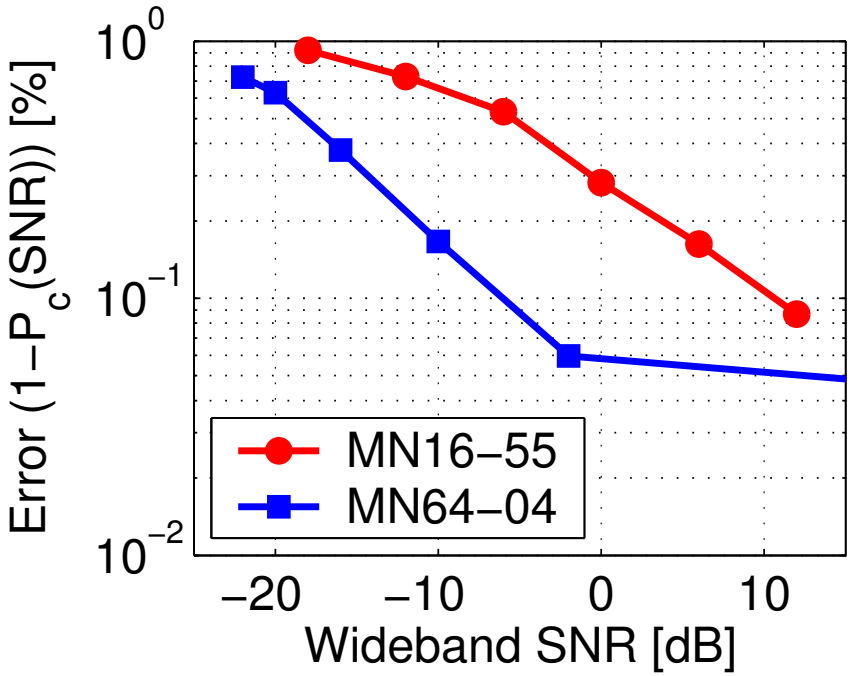
$$P_e(i, AI) \equiv 1 - P_{i,i}(AI) = \sum_{j \neq i} P(j|i, AI)$$

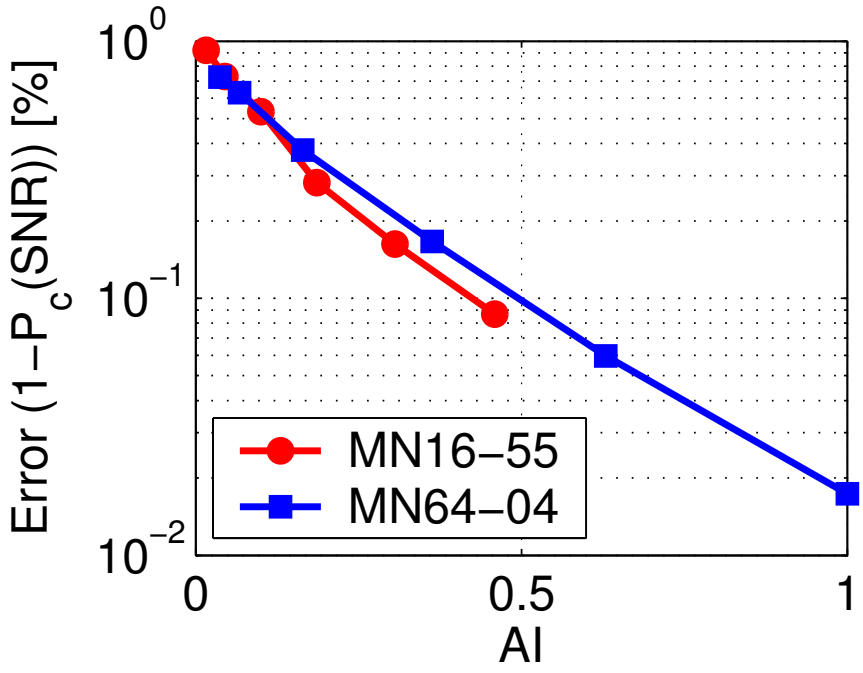
- In the previous figure the red curve is identically the sum of all the blue curves
- The green curve is the model

$$P_c(AI, \mathcal{H} = 4) = \left(1 - \frac{1}{16}\right) e_{\min}^{AI}$$

MN16 vs. MN64

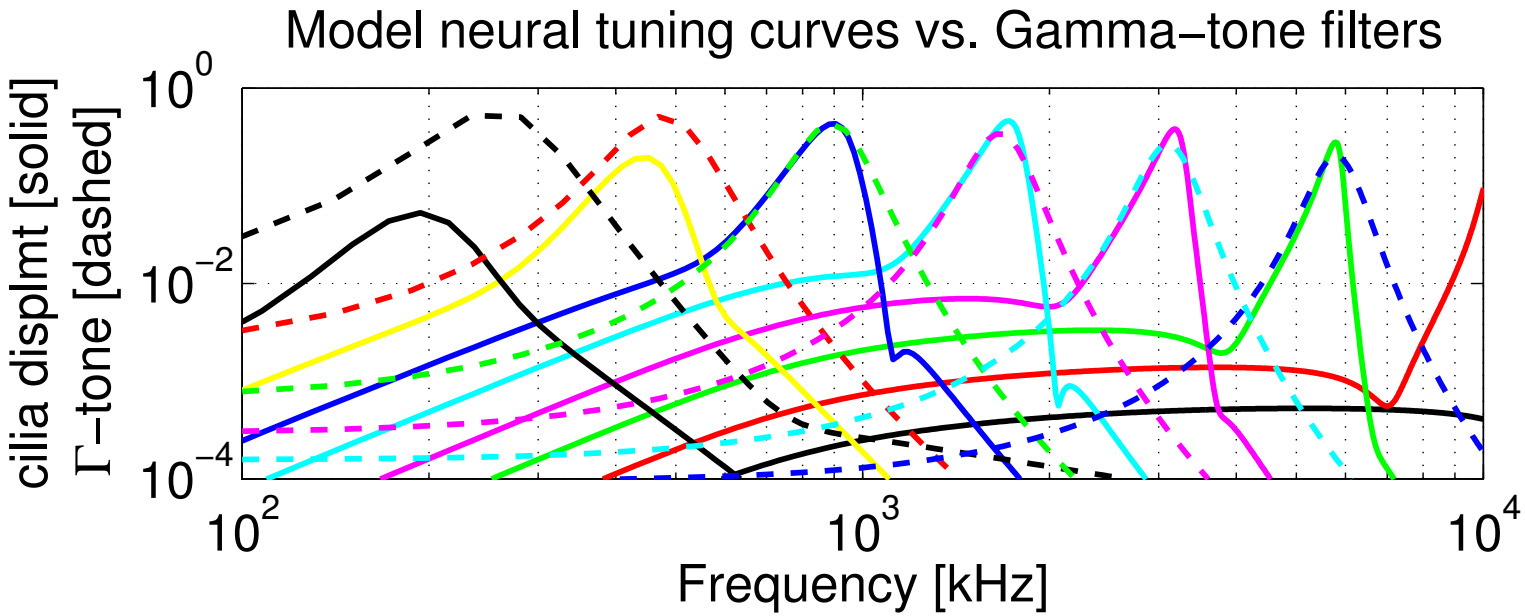
- Results of the MN16 and the MN64 CV experiments, plotted as a function of both the SNR and the AI.





Human vs. Γ-tone filters

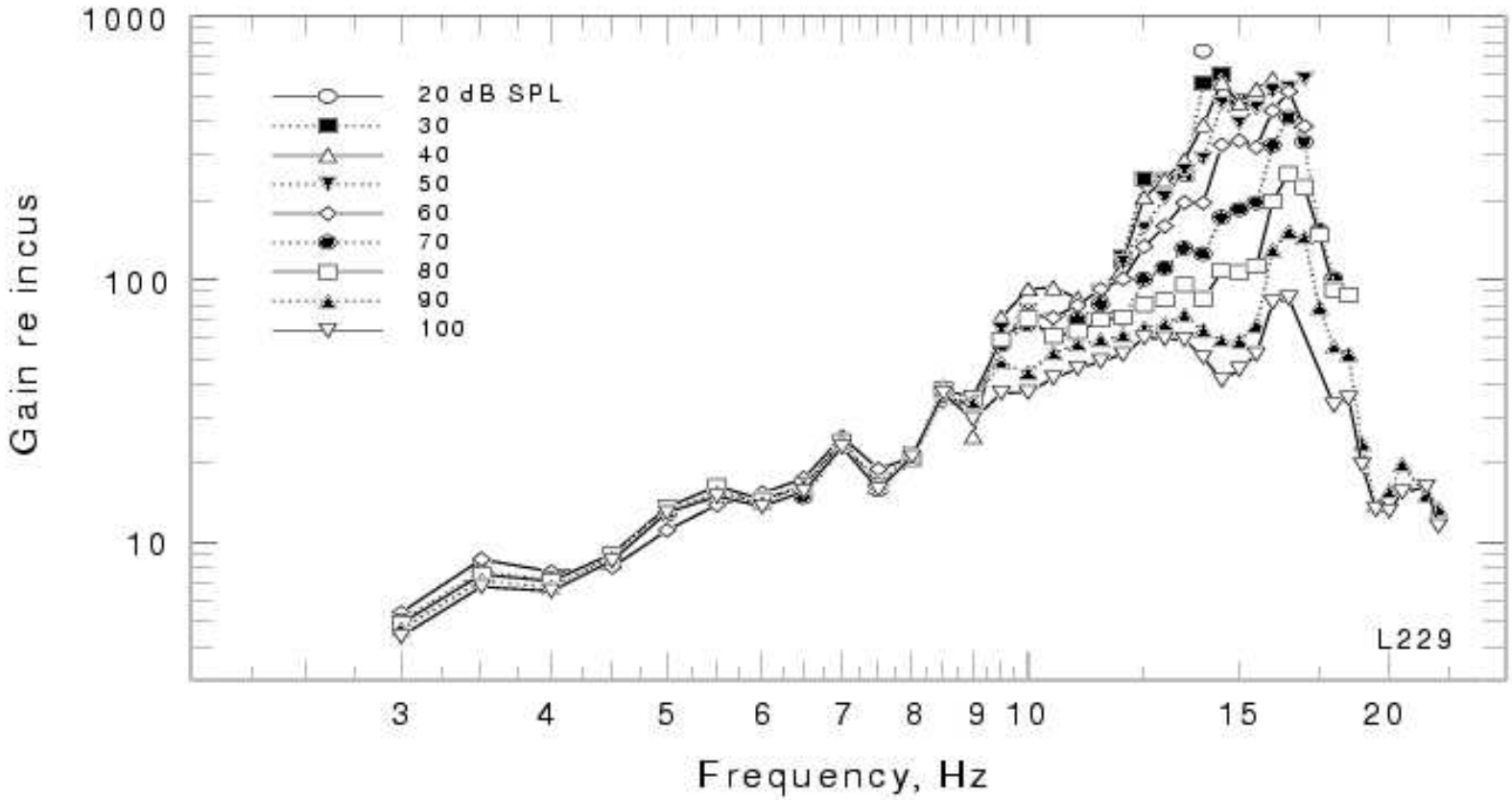
- $h(t, f_0) = t^3 e^{-2.22\,\pi\,\mathrm{ERB}(f_0)\,t}\cos(2\pi f_0 t)$ (Γ-tone filter)
- $\mathrm{ERB}(f_0) = 24.7(4.37 f_0/1000 + 1)$



- High frequency slope problem
- Low-frequency tails are wrong
- Wrong bandwidth at higher frequencies
- Missing middle ear high-pass response

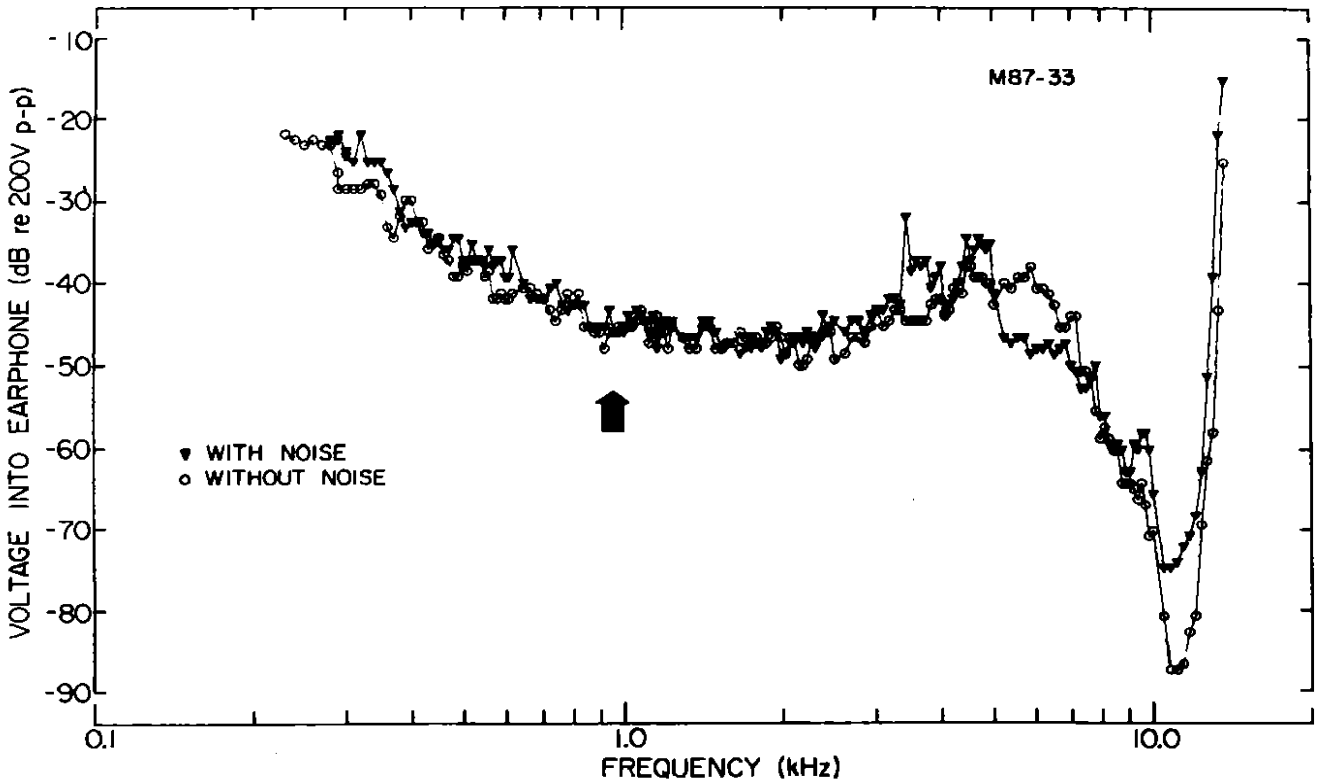
Narayan et al. data

- Narayan et al. 2000 (Gerbil)



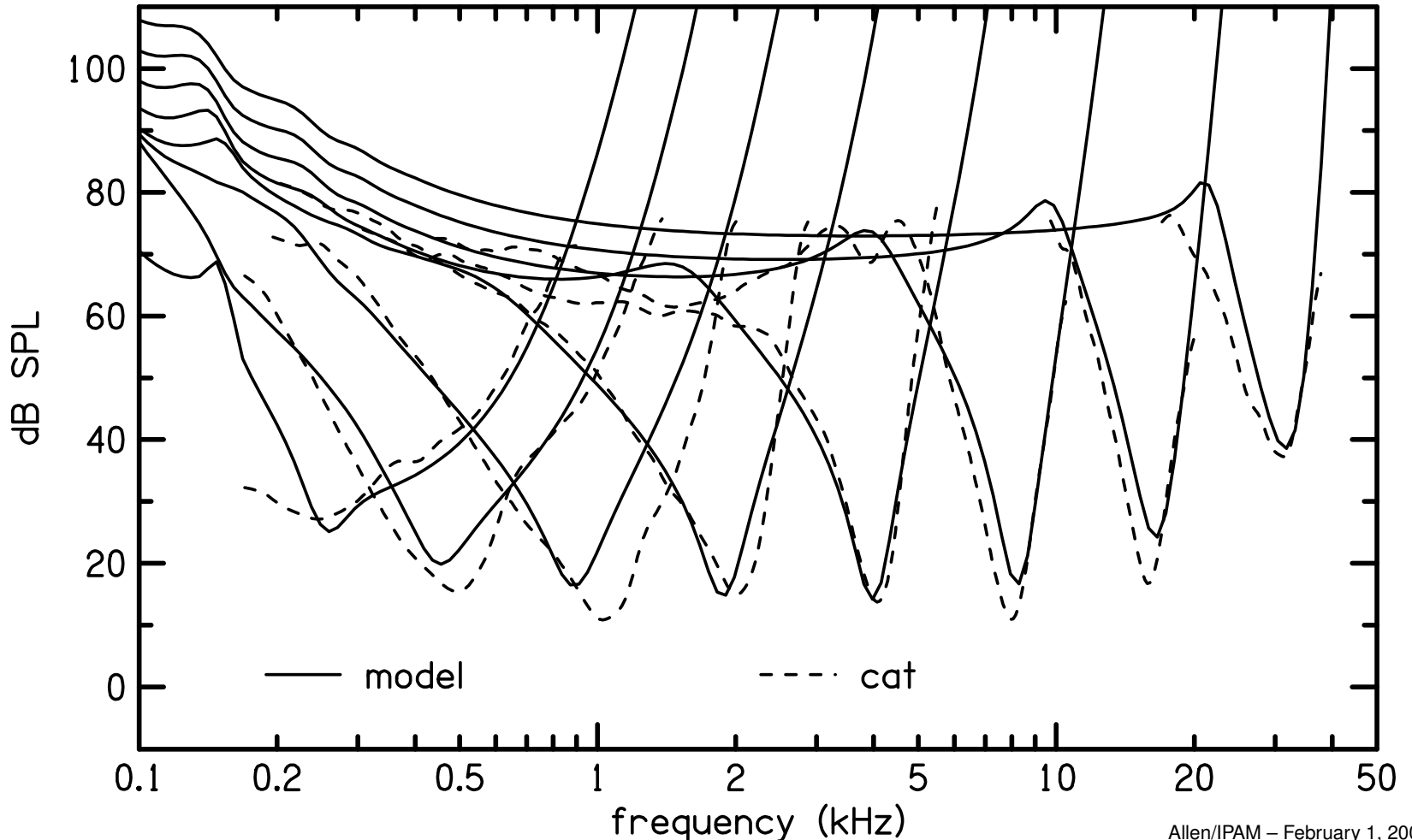
Kiang and Moxon 1979 cochlear USM

- Nonlinear upward spread of masking



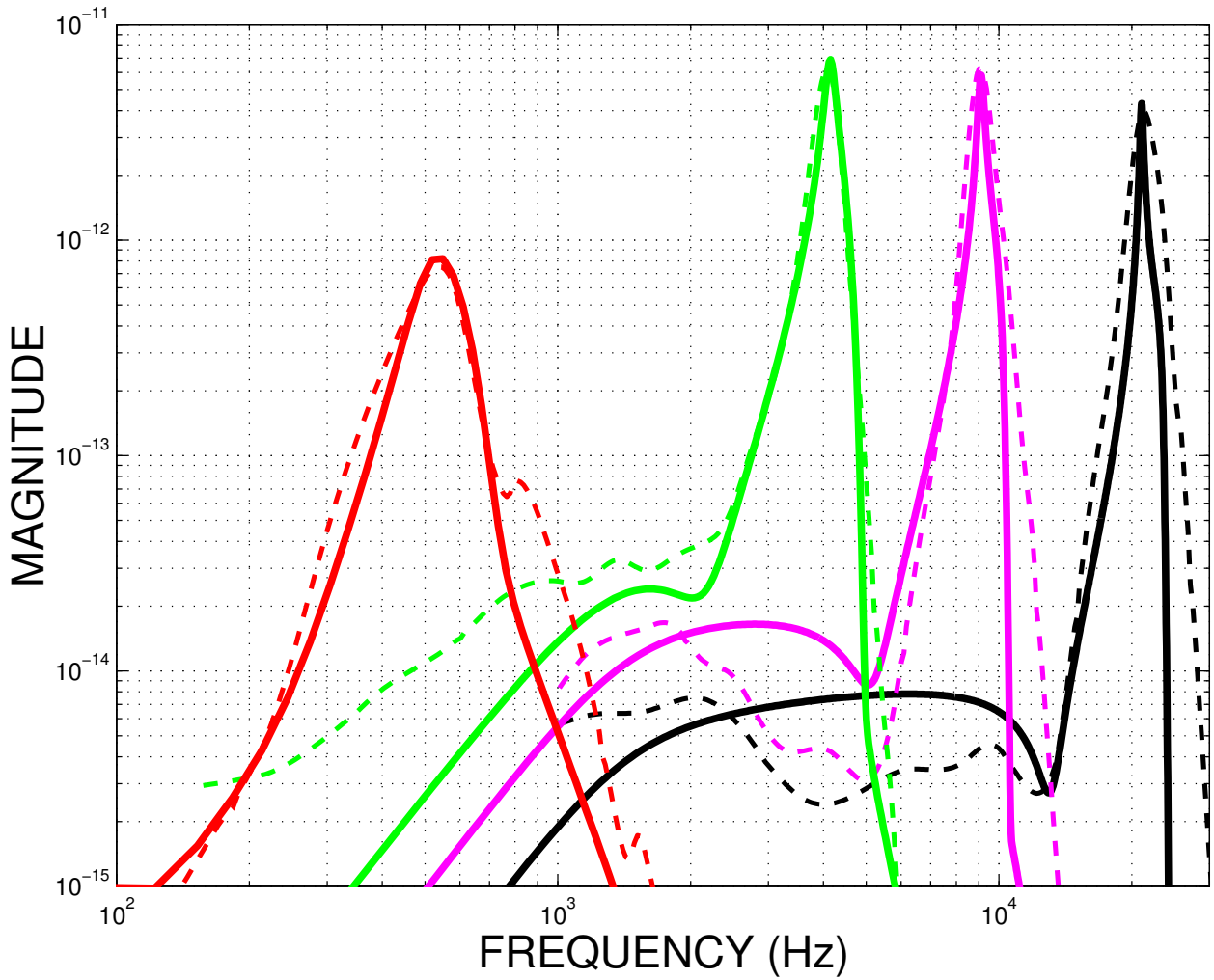
Neely model for Cat 1986

- Active model of the cochlea and cilia, based on the resonant TM model.
- Cat data from Liberman and Delgutte.

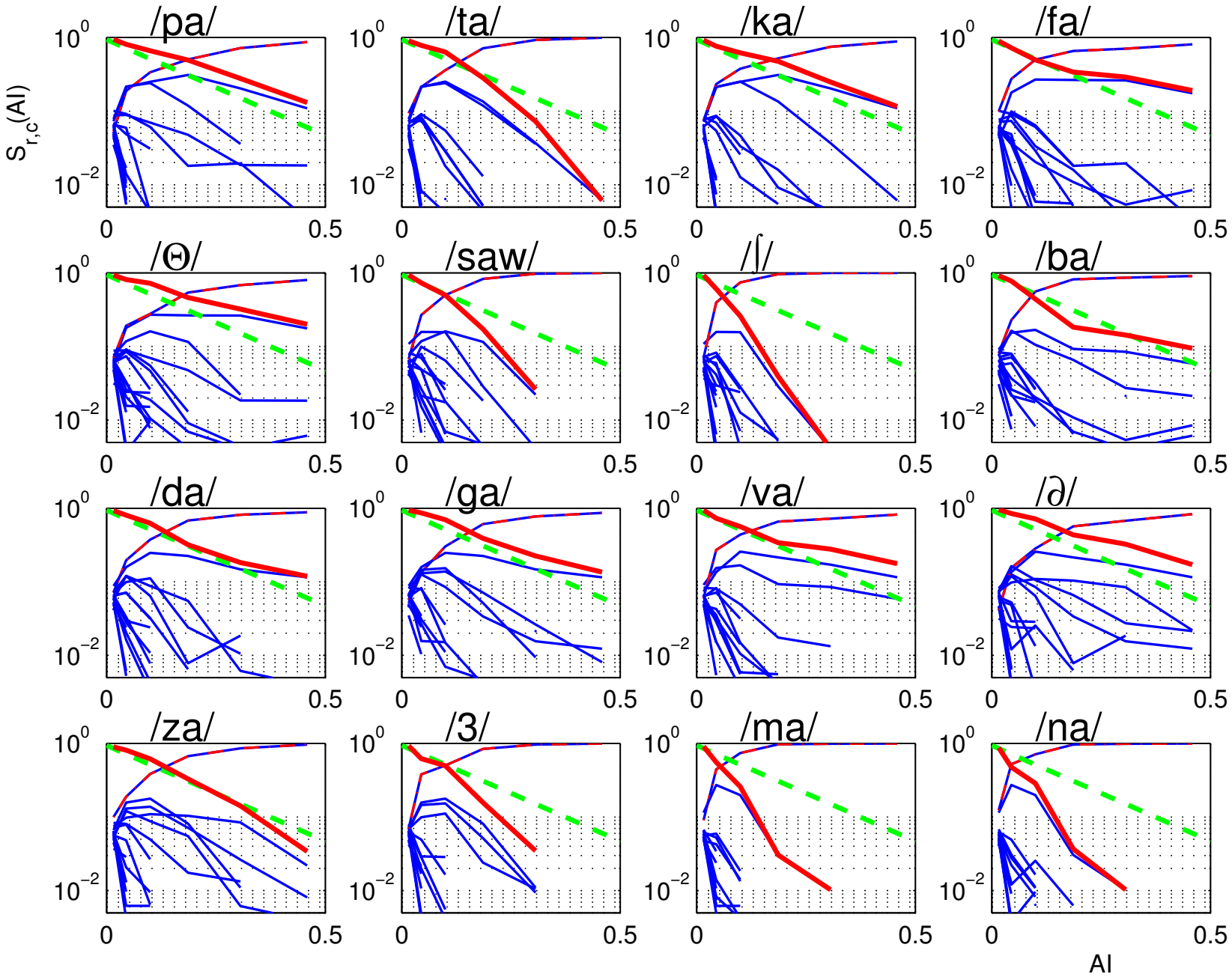


Allen model 1980

- Solid: Model; dashed Cat FTC (Liberman and Delgutte)



Symmetric form $S = (\mathcal{A} + \mathcal{A}^t)/2$



References

Allen, J. B. (1994). How do humans process and recognize speech? *IEEE Transactions on speech and audio*, 2(4):567–577.

Allen, J. B. (2004). The articulation index is a Shannon channel capacity. In Pressnitzer, D., de Cheveigné, A., McAdams, S., and Collet, L., editors, *Auditory signal processing: physiology, psychoacoustics, and models*, chapter Speech, pages 314–320. Springer Verlag, New York, NY.

Boothroyd, A. (1968). Statistical theory of the speech discrimination score. *J. Acoust. Soc. Am.*, 43(2):362–367.

Bronkhorst, A. W., Bosman, A. J., and Smoorenburg, G. F. (1993). A model for context effects in speech recognition. *J. Acoust. Soc. Am.*, 93(1):499–509.

Fletcher, H. (1929). *Speech and Hearing*. D. Van Nostrand Company, Inc., New York.

Miller, G. A. and Nicely, P. E. (1955). An analysis of perceptual confsions among some english consonants. *J. Acoust. Soc. Am.*, 27(2):338–352.

title

- Latest from Sandeep