SPEECH IS THE PROBLEM, HEARING IS THE SOLUTION

CHANNEL CAPACITY AND THE "FRONT END"

Jont B. Allen
ECE Univ. of IL
Beckman Inst., Urbana IL

October 7, 2003

jba@auditorymodels.org

MY VIEW

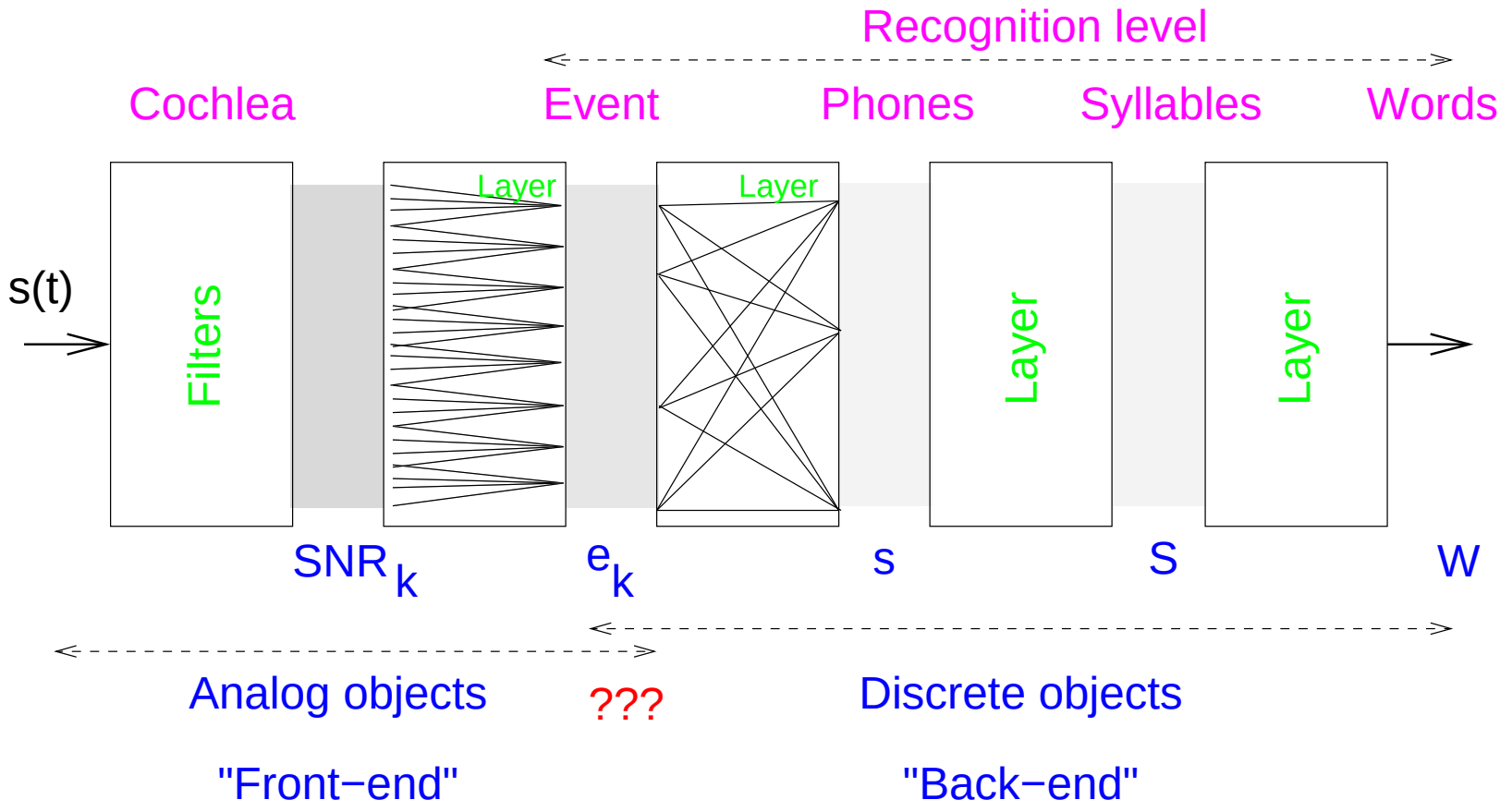
- "We have met the enemy, and he is us." *–Pogo (Walt Kelly)*
  - Why aren't we benchmarking against human performance?
- Biological systems are the ultimate information processors
  And, we need to learn from them

WHAT I WANT TO SHOW:

- Humans have an intrinsic robustness to noise and filtering
  - Robustness is not due to semantic context effects
- HSR is a bottom–up, divide and conquer strategy
  - We recognize speech based on a hierarchy of context layers
  - As in vision, entropy decreases as we integrate context

HOW WE RECOGNIZE SPEECH?

- Hierarchical “bottom up” analysis
- Accurate statistical models of performance at each stage



- Entropy drops (i.e., context is integrated) in stages

MODEL OF BAND EVENT ERRORS

- When the SNR is varied they found that the event-error is

$$e_k = e_{min}^{SNR_k/K}$$

where SNR_k is the signal to noise ratio in dB, divided by 30, such that

$$0 \leq SNR_k \leq 1$$

-

$$SNR_k \equiv \begin{cases} 0 & 20\log_{10}(snr_k) < 0 \\ 20\log_{10}(snr_k)/30 & 0 < 20\log_{10}(snr_k) < 30 \\ 1 & 30 < 20\log_{10}(snr_k). \end{cases}$$

- Total error:

$$e = e_1 e_2 \cdots e_K = e_{min}^{(SNR_1 + SNR_2 \cdots SNR_K)/K}$$

- The speech SNR (not the energy) determines the event errors e_k and thus the phoneme articulation $s = 1 - e_1 e_2 \cdots e_K$

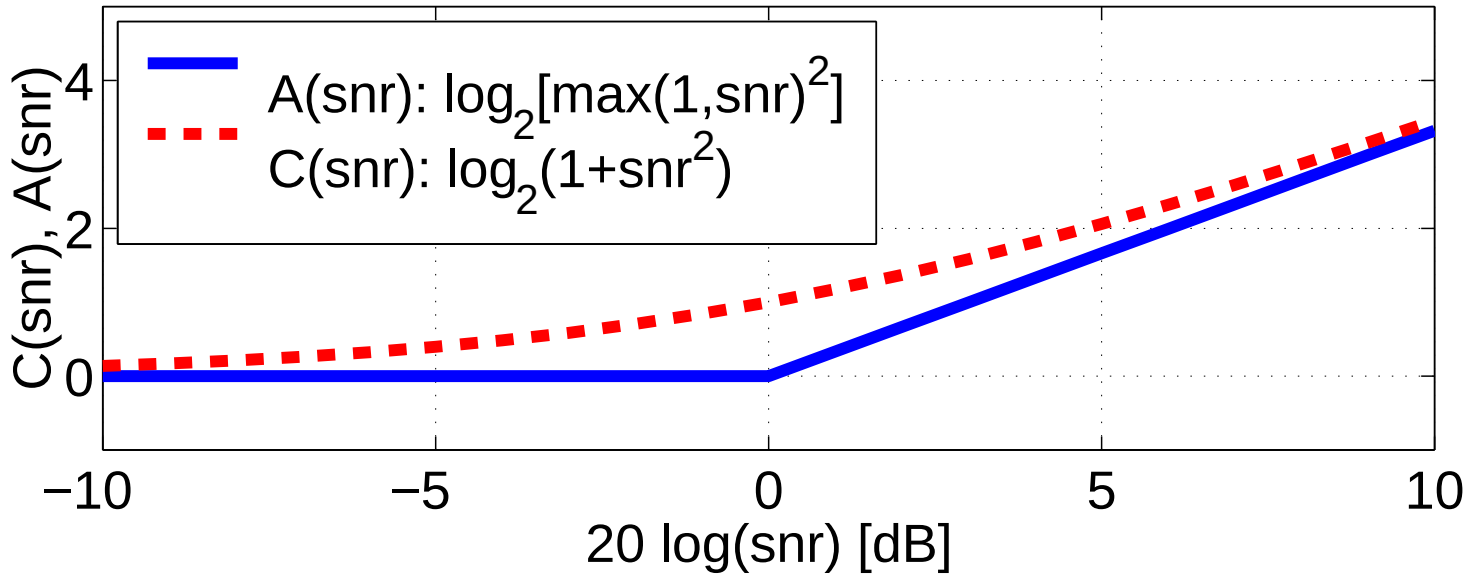
AI AS A CHANNEL CAPACITY

- Since $\Sigma_k(\log snr_k) = \log(\Pi_k snr_k)$

$$\mathcal{A} \equiv \frac{1}{K}\sum_k SNR_k \propto \log\left(\prod_k snr_k\right)^{1/K} \quad (1)$$

- and from Shannon

$$C = \int_{-\infty}^{\infty} \log_2[1 + snr^2(f)]df, \quad (2)$$

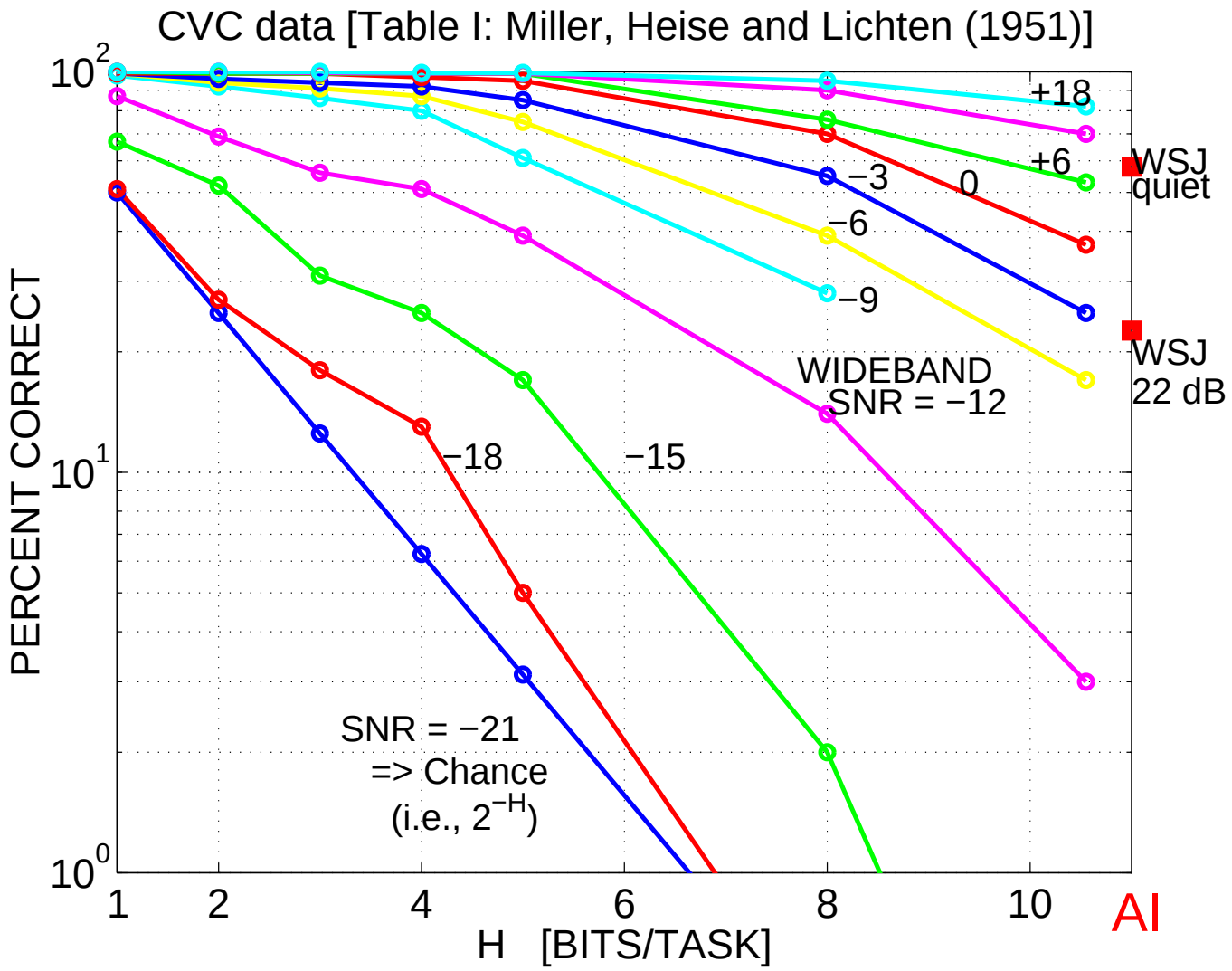


AI IS A CHANNEL CAPACITY

- In conclusion:
  - The channel capacity is the maximum information rate that can asymptotically be sent over a channel without error
  - The AI is basically a channel capacity

SPEECH ENTROPY VS. THE WIDEBAND SNR

- $P_c(\mathcal{H}, SNR)$ Miller, Heise and Lichten 1951
- Many of the results of MHL51 expand on the AI model



CONFUSION MATRIX PARTITIONING

- Miller & Nicely 1955 Confusion Matrix (Table III)
  - MN55 established a natural phone hierarchical clustering:

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

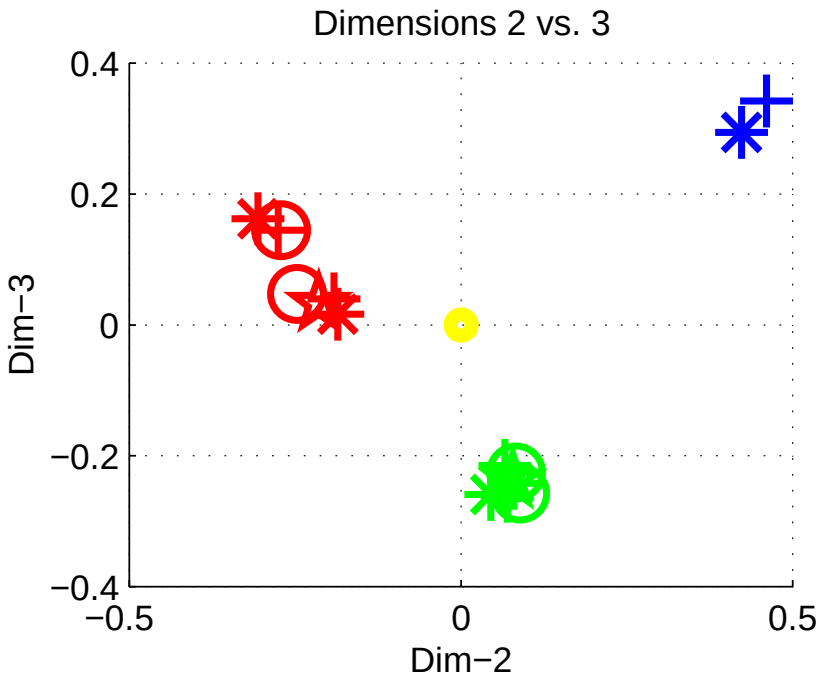
UNVOICED VOICED NASALS

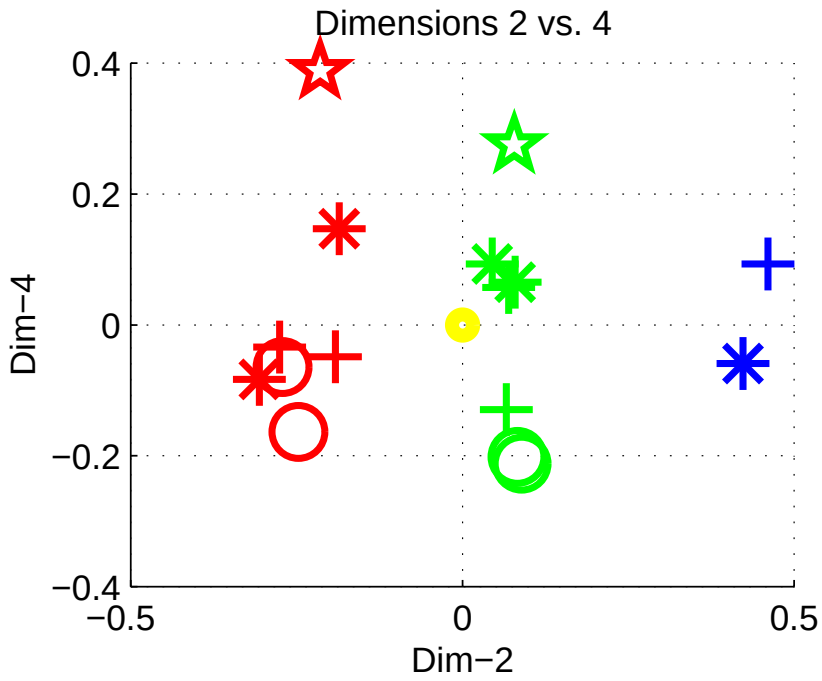
RESPONSE

"This breakdown of the confusion matrix into five smaller matrices . . . is equivalent to . . . five communication channels . . . ." –*Miller & Nicely 1955*

SVD REPRESENTATION OF THE PERCEPTUAL SPACE

- 4^{dim} SVD perceptual representation of the confusion matrix



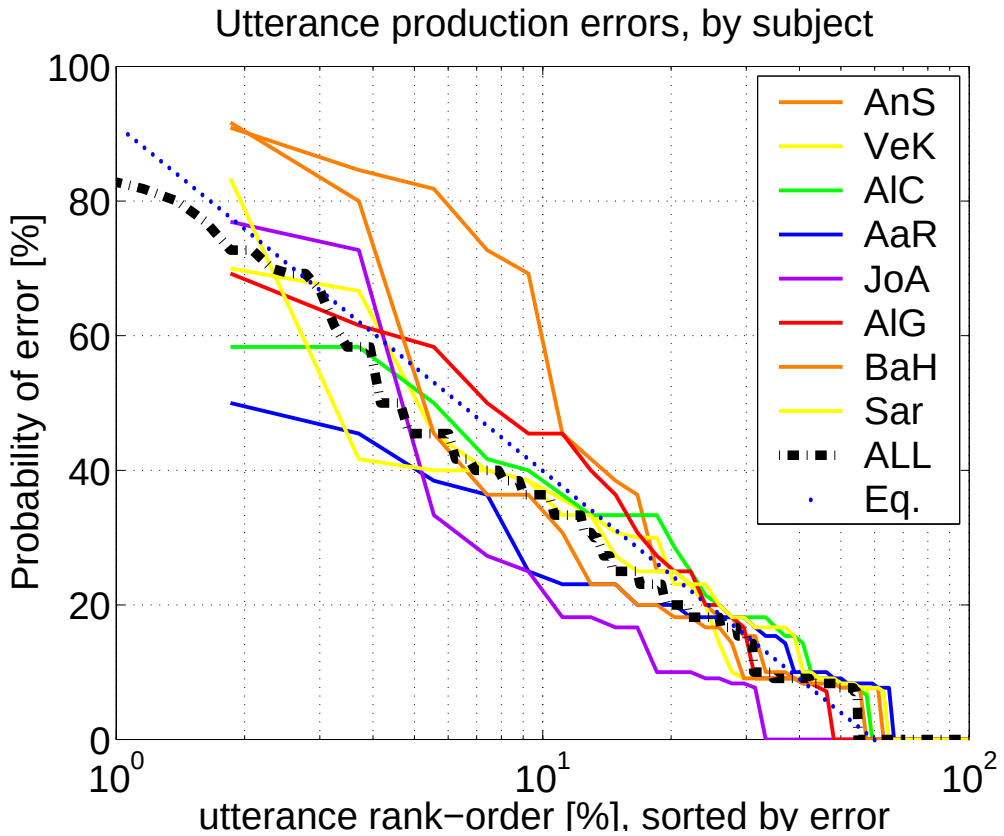


DEMO

DISTRIBUTION OF UTTERENCE ERRORS

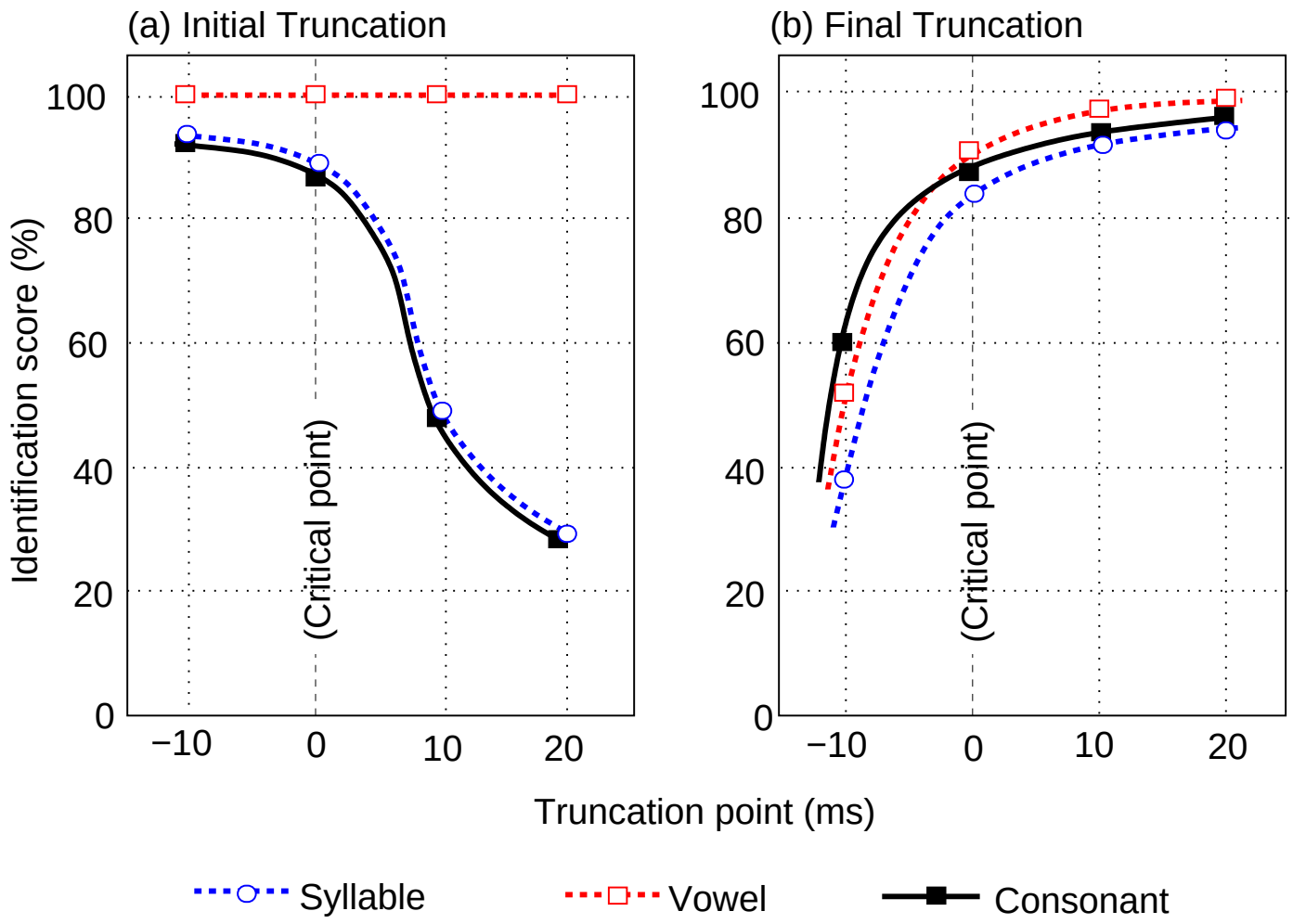
- What determines $s_{max} = 1 - e_{min}$?
- Utterance *talker mispronunciations*, as defined by 32 listeners
- Errors are distributed like Zipf's Law [$\cdots N/N_T \approx 0.6e^{-4.48P_e}$]

  35% of the utterances have no error
  33% have $>$ 10% error, 10% $>$ 35% error, 5% $>$ 50% error



TEMPORAL RESOLUTION OF PHONE RECOGNTION

- Phones are recognized in on a 10 ms time scale (Furui 1986)



GRAMMATICAL CONTEXT

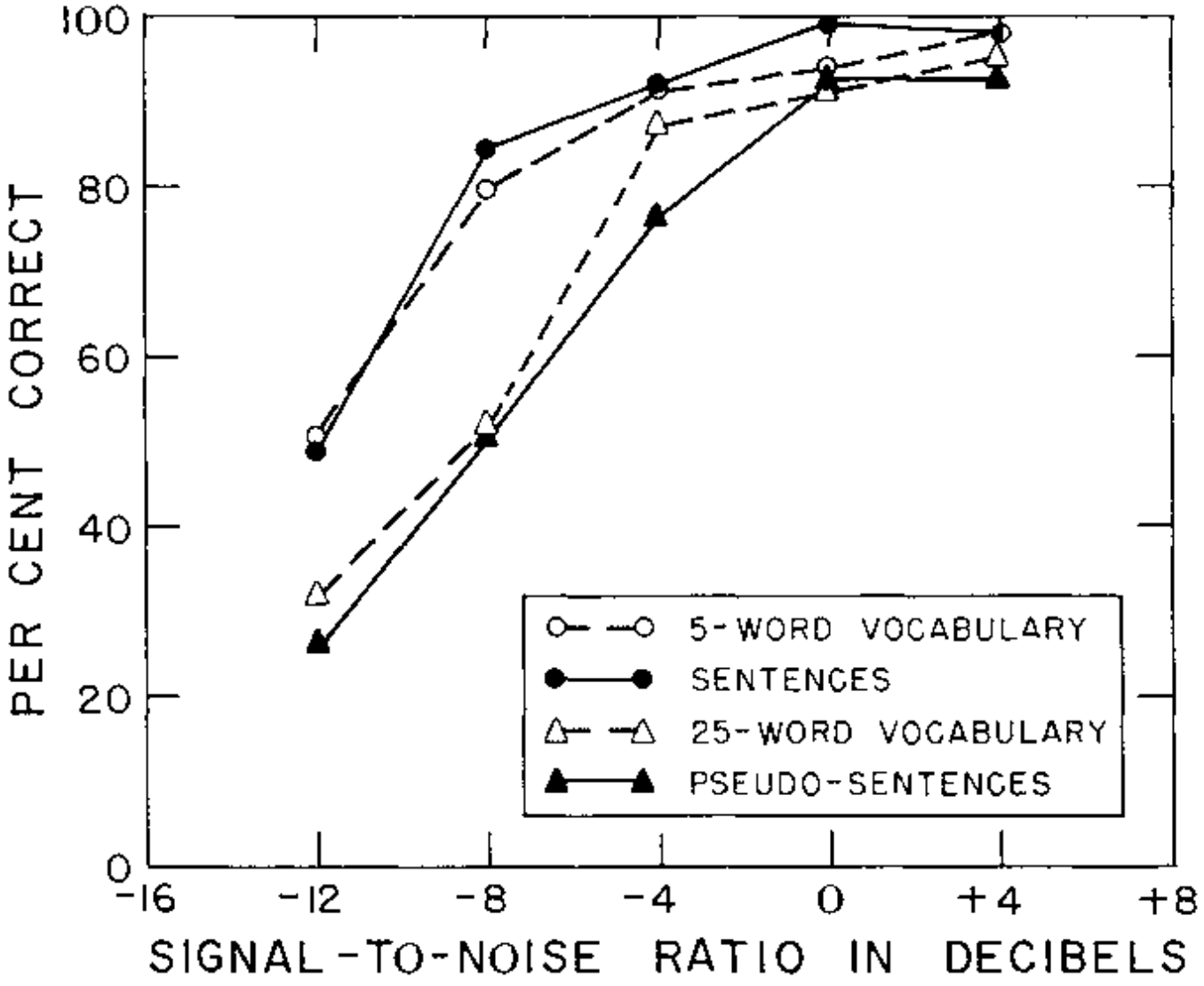
- Five groups of five words that form grammatical sentences:

Don	Brought	His	Black	Bread
He	Has	More	Cheap	Sheep
Red	Left	No	Good	Shoes
Slim	Loves	Some	Wet	Socks
Who	Took	The	Wrong	Things

- Tests:
  - 5 word lists
  - 25 word
  - 25 words with grammatical context
    - Example: He left no black socks
  - 25 words reverse order
    - Example: Socks black no left he.

GRAMMATICAL CONTEXT

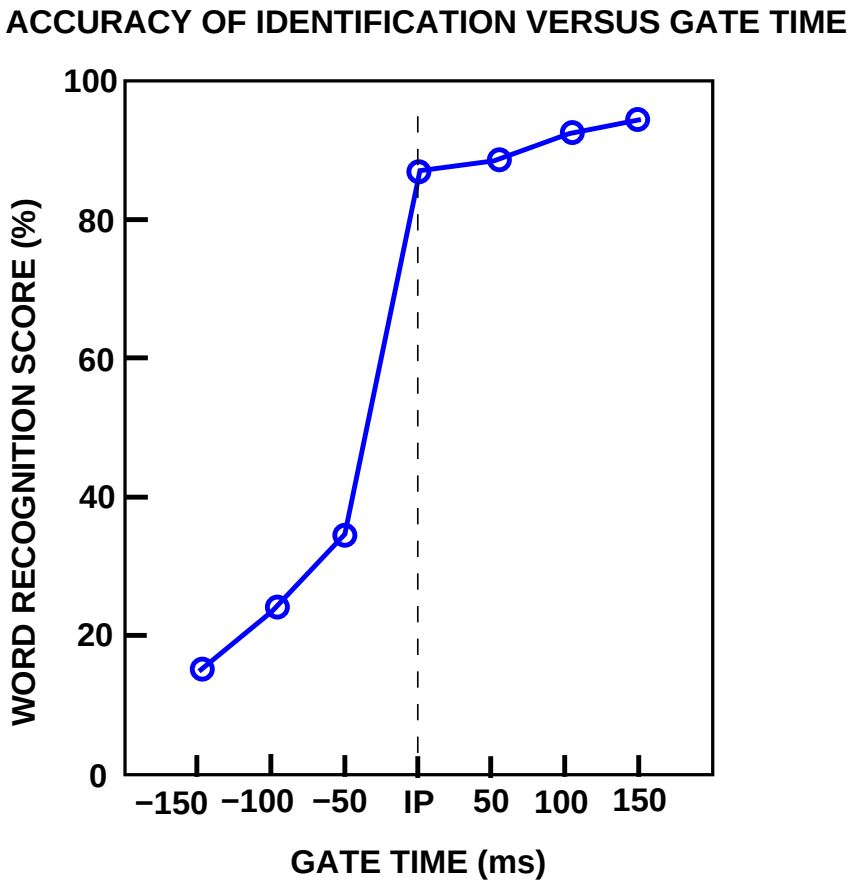
- Results of tests



WORD SEMANTICS: IP DEFINITION

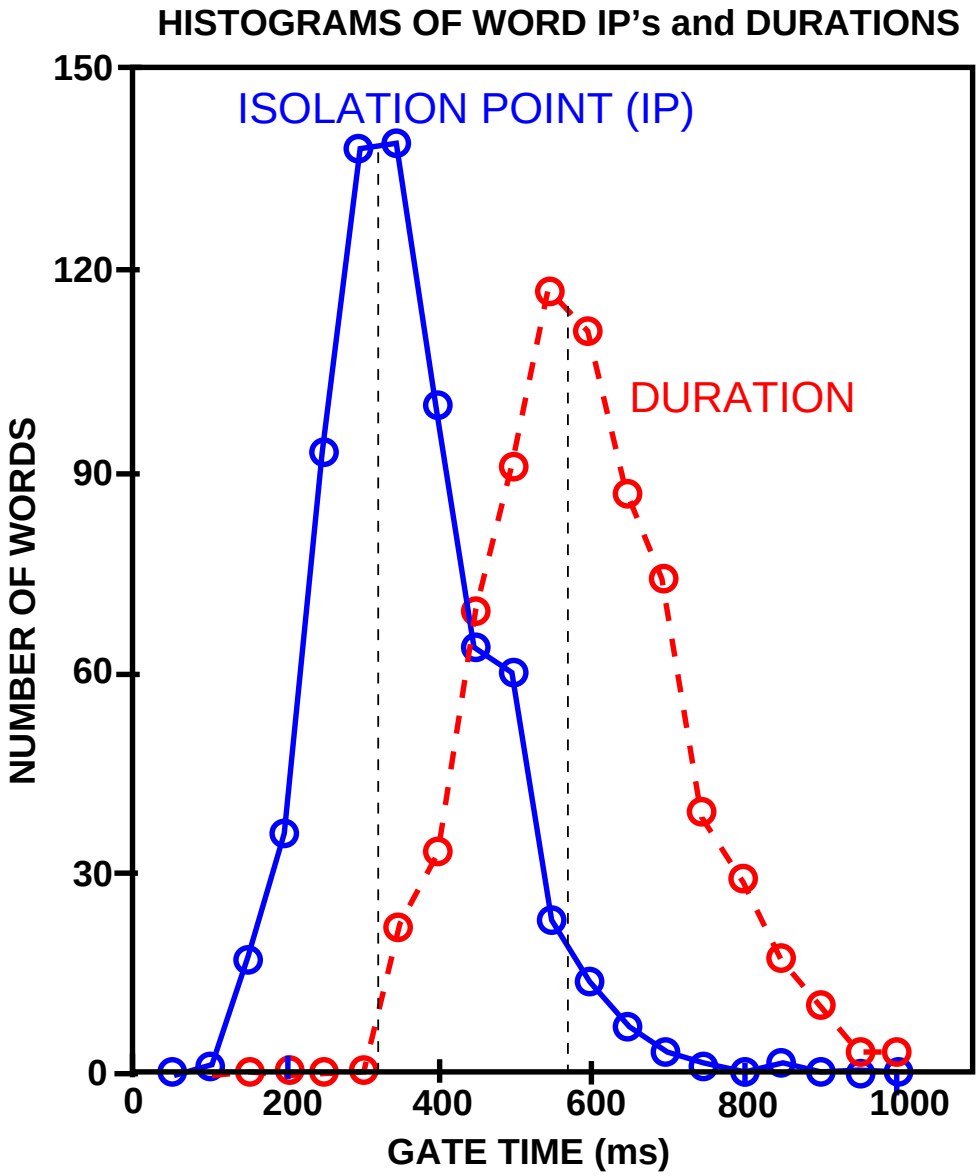
- 704 isolated words were truncated in 50 ms steps Van Petten 1999
- Isolation point is defined as *the time of the discontinuity in recognition*

  Expt. I – Neutral sentences: "The next word is *test-word*."



WORD SEMANTICS: IP VS. DURATION

- Isolation point vs. word durations (real words, no sentence context)

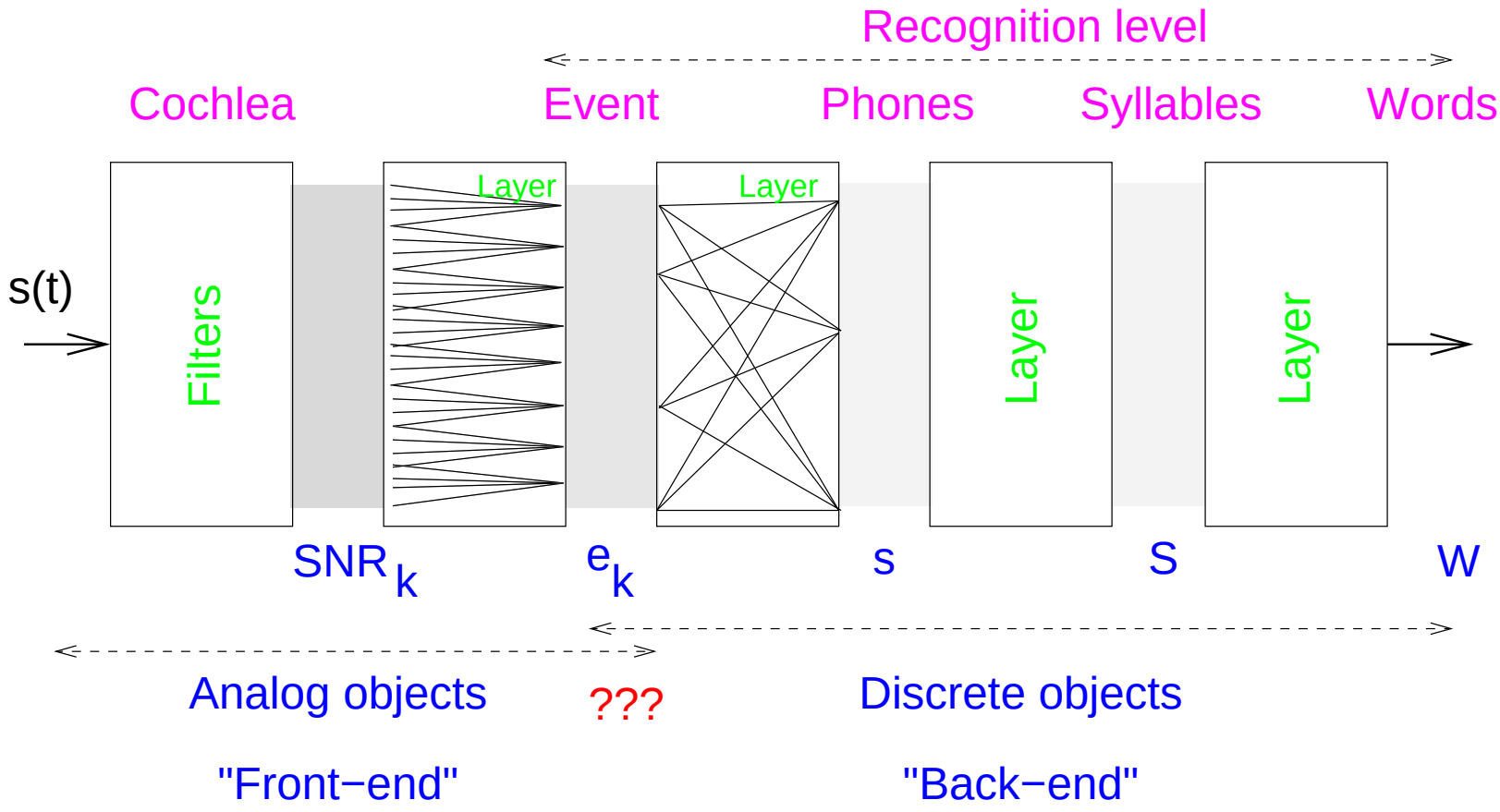


FROM CONTINUOUS TO DISCRETE

- Φ-domain signals
  - Speech signal
  - Cochlear filter outputs
  - Neural rate
  - Voltage in cochlear nucleus cells
- Ψ-domain objects
  - Words
  - Syllables
  - Phonemes
  - Events [Miller's features]

HOW WE RECOGNIZE SPEECH?

- Hierarchical “bottom up” analysis
- Accurate statistical models of performance at each stage



- Entropy drops (i.e., context is integrated) in stages