FROM LORD RAYLEIGH TO SHANNON: HOW DO WE DECODE SPEECH?

Jont B. Allen
ECE Univ. of IL
Beckman Inst.
Urbana IL
jba@auditorymodels.org

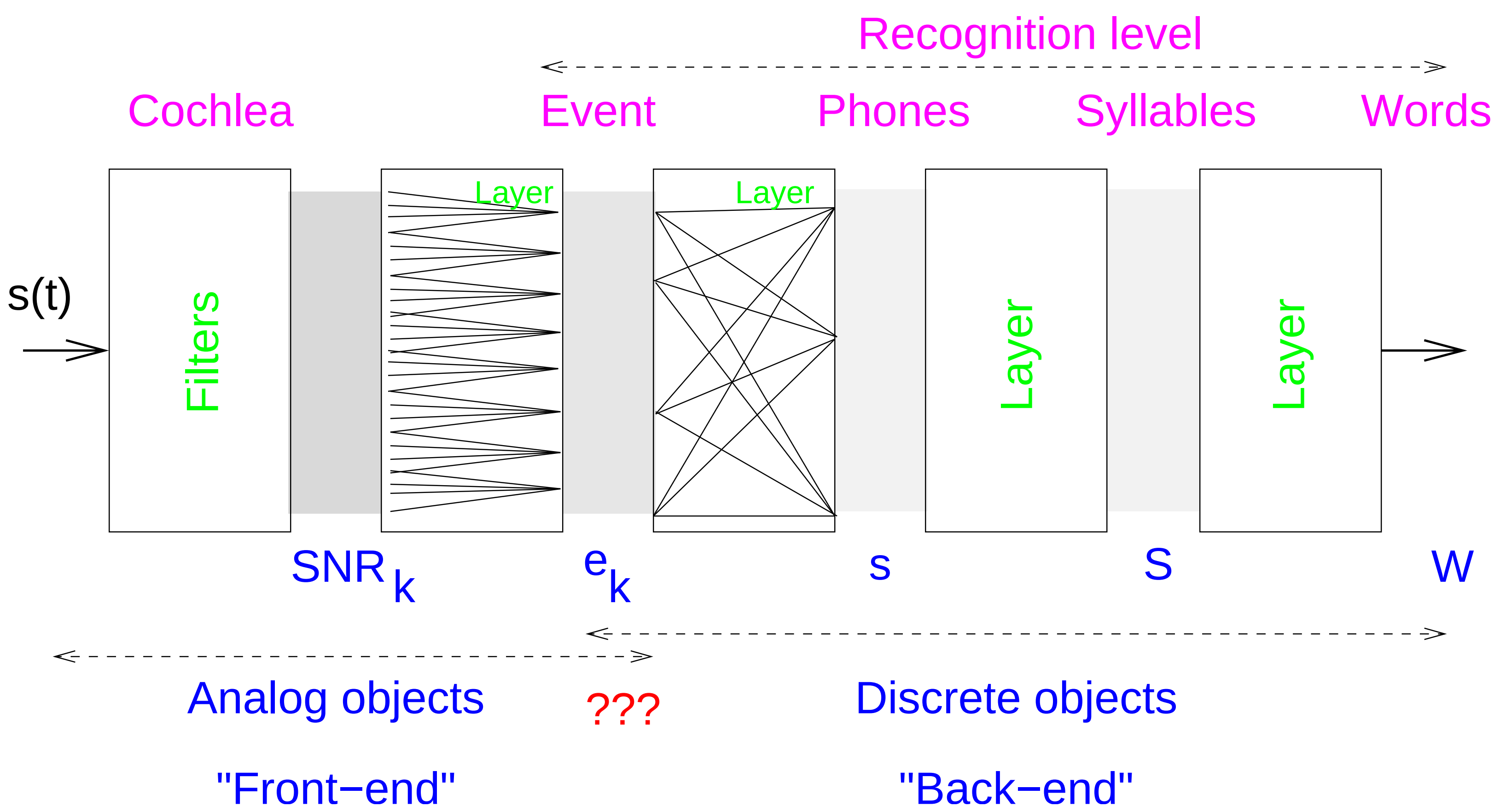
http://auditorymodels.org/CUNY/
http://auditorymodels.org/jba/PAPERS/ICASSP/

WHAT I WANT TO SHOW:

- Biological systems are the ultimate information processors
- 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
- Humans have an intrinsic robustness to noise and filtering
  - Robustness is not due to semantic context effects

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

DEFINITIONS

$\surd$	phone	A consonant (C) or vowel (V) sound
	word	A meaningful phone or phone sequence (i.e., cat $\equiv$ CVC)
	phoneme	The replaceable set of phones which leave a word meaning invariant
	recognition	Probability measure P_c of correct phoneme identification
$\surd$	articulation	Recognition of “nonsense words”
$\surd$	intelligibility	Recognition of words (i.e., meaningful speech)
	robustness	Relative recognition with filtering and noise
	confusion matrix	Table of identification frequencies $N_{sr} \equiv N_{r\|s}$
	articulation matrix	A *confusion matrix* composed of nonsense sounds
$\surd$	articulation event	A discrete subunit of articulation [e.g., Voicing: /ba/ vs. /pa/]
	trial	A single presentation of a set of events
	state	A values of a set of events at some instant of time
	state machine	A machine (program) that transforms from one state to another
	noiseless state machine	A deterministic state machine
	context	Coordinated combinations of events within a trial
	message	Specific information transmitted by a trial
	p_n	Probability of event n, of N possible events
	information density	$I_n = \log_2(1/p_n), \qquad n = 1, \cdots, N$
	entropy	Average information: $H = \sum_{n=1}^{N} p_n I_n$

KEY HSR STUDIES

- The first articulation experiments date from Lord Rayleigh's 1908 and George Campbell 1910 phoneme identification experiments
- A basic probabilistic approach was developed by Stewart & Fletcher 1921
  - Detailed review of Fletcher's AI theory: Allen IEEE 1994
- French and Steinberg 1947 WWII studies
- Shannon's Information theory 1948+
- G.A. Miller, Heise and Lichten 1951; G.A. Miller & Nicely 1955
- *Language and communication* G.A. Miller, 1951 McGraw Hill
  Miller first introduces IT to language modeling, following Shannon
- Boothroyd JASA 1968; Boothroyd & Nittrouer JASA 1988
- Bronkhorst et al. JASA 1993, 2002
- Van Petten *et al.* 1994
- Detailed review chapter Allen 2003

MOTIVATION

- Results of Lippmann 1997, sorted by Error Ratio

Corpus	Size in Words	Conditions	% Error Machine	% Error Human	Error Ratio
Alphabetic	26	20-talkers 8-listeners	5.0 isolated	1.6 continuous	3
Resource	1000	null grammar	17	2	8
WSJ-NAB	5000	quiet (trained)	7.2	0.9	8
Switchboard	14,000	spontaneous (tel. BW)	43	4	11
WSJ-NAB	5000	10 dB (trained)	12.8	1.1	12
WSJ-NAB	65,000	close mic	6.6	0.4	16
WSJ-NAB	65,000	omni mic	23.9	0.8	30
Resource	1000	word–pair grammar	3.6	0.1	36
WSJ-NAB	5000	quiet (not trained)	42	0.9	47
WSJ-NAB	5000	22 dB (not trained)	77.4	0.9	86
word spotting	20	judgment errors	24	0.3	80
TI-digit	10	connected	0.72	0.009	80

DEMO ScanMail examples /Audio/ScanMailExample

TYPICAL ARTICULATION TEST RECORD

- Fletcher’s method of nonsense phone error analysis

ARTICULATION TEST RECORD

March 1928

DATE 3-16-28 SYLLABLE ARTICULATION 51.5%

TITLE OF TEST PRACTICE TESTS CONDITION TESTED 1500~ LOW PASS FILTER

1500 Hz lowpass filtering

NO.		OBSERVED	CALLED	OBSERVED	CALLED	OBSERVED	CALLED
1	THE FIRST GROUP IS	ma'v	na'v	po'z	po'th	kōb	✓ kōb
2	CAN YOU HEAR	pōch	✓ pōch	nēz	nēzh	sheth	siz
3	I WILL NOW SAY	seng	✓ seng	jo'ch	✓ jo'ch	fūch	✓ fūch
4	AS THE FOURTH WRITE	chūd	✓ chūd	tha'm	✓ tha'm	thōl	✓ thōl
5	WRITE DOWN	run	✓ run	hab	✓ hab	poth	✓ poth

DATA

$$S \equiv P_c(syllable) = 0.515$$
$$v \equiv P_c(vowels) = 0.909$$
$$c \equiv P_c(consonants) = 0.74$$

MODELS

$$\hat{S} = cvc = 0.498$$ (CVC syllable model)
$$s \equiv P_c(phone) = (v + 2c)/3 = 0.796$$
$$s^3 = 0.505$$ (3 phone syllable model)

THE METHOD

- The data bases they used were formed from
  - statistically balanced
  - nonsense
  - CVC, CV and VC syllable lists
    where C represents a consonant and V a vowel
- The syllable lists were spoken, and the listeners recorded what they heard
- Probabilities-correct c and v for the sound-units were computed
- The average $\{$C,V$\}$ speech-unit articulation probability s was computed from the composition of $\{$C,V$\}$ units in the data base

  (i.e. $s = (2c + v)/3$ for CVC's, $s = (c + v)/2$ for CV's)
  - Measure s looks like a sufficient statistic

WHAT THEY FOUND

- Nonsense phones are recognized as independent units:
  - The probability of correct recognition for the average phoneme s accurately predicts the nonsense syllable score S_{cvc}, where

$$\begin{aligned} S_{cvc} &= c^2 v \\ &= s^3 \end{aligned}$$

  $*$This is a necessary but insufficient condition for *independence*

- These statistical models are highly accurate
- !!! Remember: This only applies to “nonsense words” !!!

QUESTIONS?

THE NEXT STEP

- Next they dissected $s \equiv P_{correct}(phone)$ into frequency bands!

SPECIFIC DEFINITIONS

SYMBOL	DEFINITION
α	gain applied to the speech
$c(\alpha) \equiv P_c(\text{consonant}\|\alpha)$	consonant articulation
$v(\alpha) \equiv P_c(\text{vowel}\|\alpha)$	vowel articulation
$s(\alpha) = [2c(\alpha) + v(\alpha)]/3$	average phone articulation for CVC's
$e(\alpha) = 1 - s(\alpha)$	phone articulation error
f_c	high– and low–pass cut–off frequency
$s_L(\alpha, f_c)$	s for low-pass filtered speech
$s_H(\alpha, f_c)$	s for high-pass filtered speech

FLETCHER'S TWO BAND FORMULATION

- Split the speech into low and high bands, having articulations

$$s_L(\alpha, f_c) \text{ and } s_H(\alpha, f_c)$$

- Fletcher proposed a linearizing transformation of the phone articulations

$$(s_L) + \quad (s_H) = \quad (s)$$

  - This is a nonlinear transformation of probabilities
  - There was no guarantee that such a transformation exists
    However, Fletcher's intuition was correct

WHAT THEY FOUND

- For nonsense {C,V} syllables the phone articulation transformation is:

$$(s) = \frac{\log(1-s)}{\log(e_{min})},$$

with $e_{min} = 0.015$ (1.5% error, or 98.5% correct)

  – This relationship must have taken years to discover!

- Solving for $e \equiv 1 - s(\mathcal{A})$:

$$e = e_{min}^{\mathcal{A}(s)} = e_{min}^{\mathcal{A}(s_L)+\mathcal{A}(s_H)} = e_{min}^{\mathcal{A}(s_L)} e_{min}^{\mathcal{A}(s_H)}$$

- In terms of the error probabilities $e = 1 - s$, $e_L = 1 - s_H$ and $e_L = 1 - s_L$:

$$e = e_L e_H.$$

FLETCHER'S TWO BAND EXAMPLE

- If we have 100 spoken sounds, and 10 errors are made while listening to the low band, and 20 errors are made while listening to the high band, then

$$e = 0.1 \times 0.2 = 0.02,$$

namely 2 errors will be made when listening to the full band, so

$$s = 1 - 0.02 = 0.98$$

$$S = s^3 = 0.941$$

- This is an unexpected, simple, and amazing result
  - What does this mean? Why does it turn out this way?

DEMO of the the McGurk effect

THE FLETCHER-STEWART MULTI-CHANNEL MODEL

- Fletcher 1921 generalize the two-band case to $K = 20$ frequency bands

$$
\begin{aligned}
1 - s &= e_1 e_2 \cdots e_k \cdots e_K \;\times\; e_{\text{visual}} \\
&= (1 - s_1)(1 - s_2) \cdots (1 - s_K) \;\times\; (1 - s_{\text{visual}})
\end{aligned}
$$

where

$$e_i \equiv 1 - s_i$$

–This formula forms the basis of articulation index theory

–Why $K = 20$ bands?

Each band equals 1mm along the basilar membrane

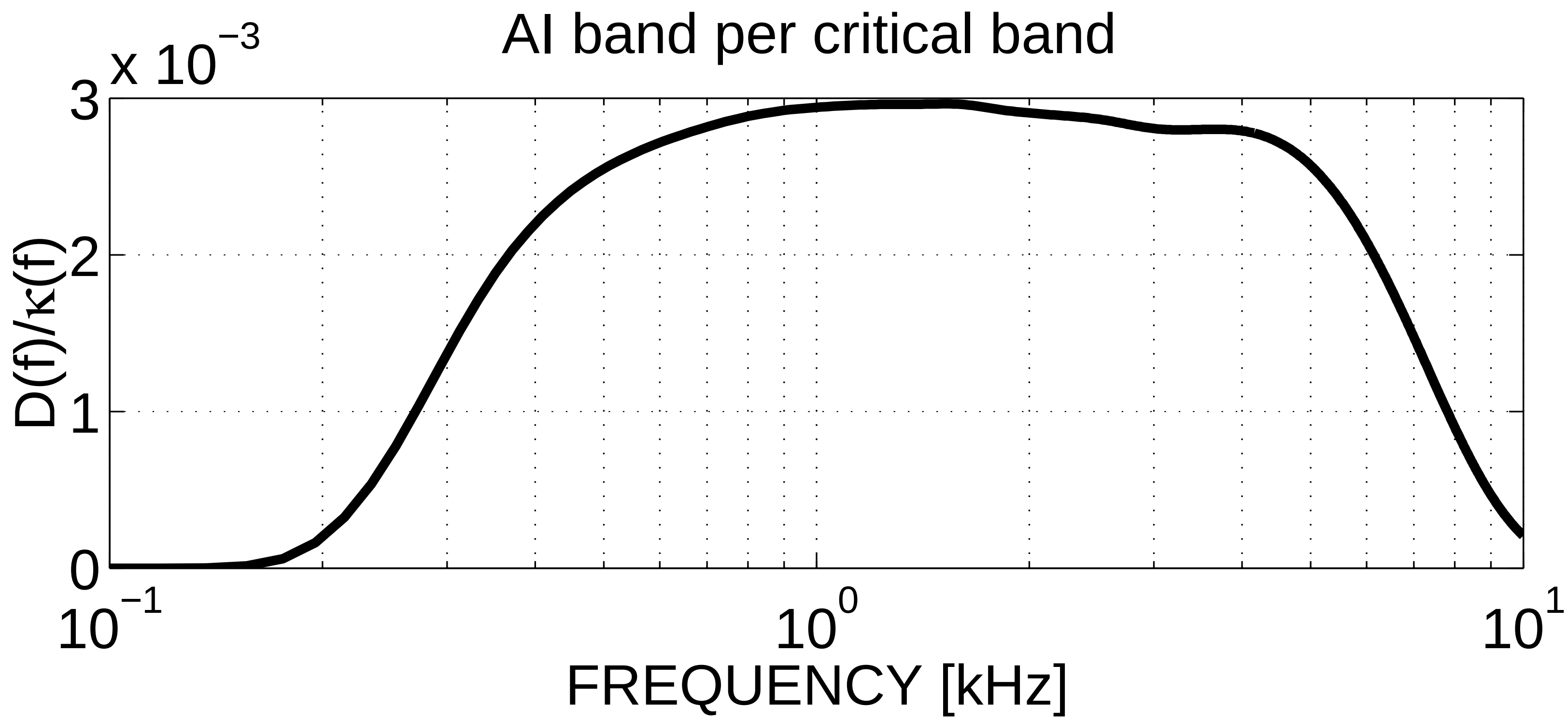
- I have added a visual channel, to account for the McGurk effect (Channel 21)
- Probability of error e_i models events, as in the visual example

DENSITY OF ARTICULATION PER CRITICAL BAND

- This plot is the ratio of $D(f)/\kappa(f)$, where $D(f)$ is the articulation density

$$D(f_c) \equiv \frac{\partial \mathcal{A}_L}{\partial f_c}, \qquad K \text{ AI bands}$$

$\kappa(f)$ = the critical ratio [$\propto$ cochlear filter bandwidth (ERB)]



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

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

Thus

$$0 \le SNR_k \le 1.$$

- Total error:

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

- The speech *SNR in dB* (not the energy) determines the event errors e_k, and thus the phoneme articulation

$$s = 1 - e_1 e_2 \cdots e_K$$

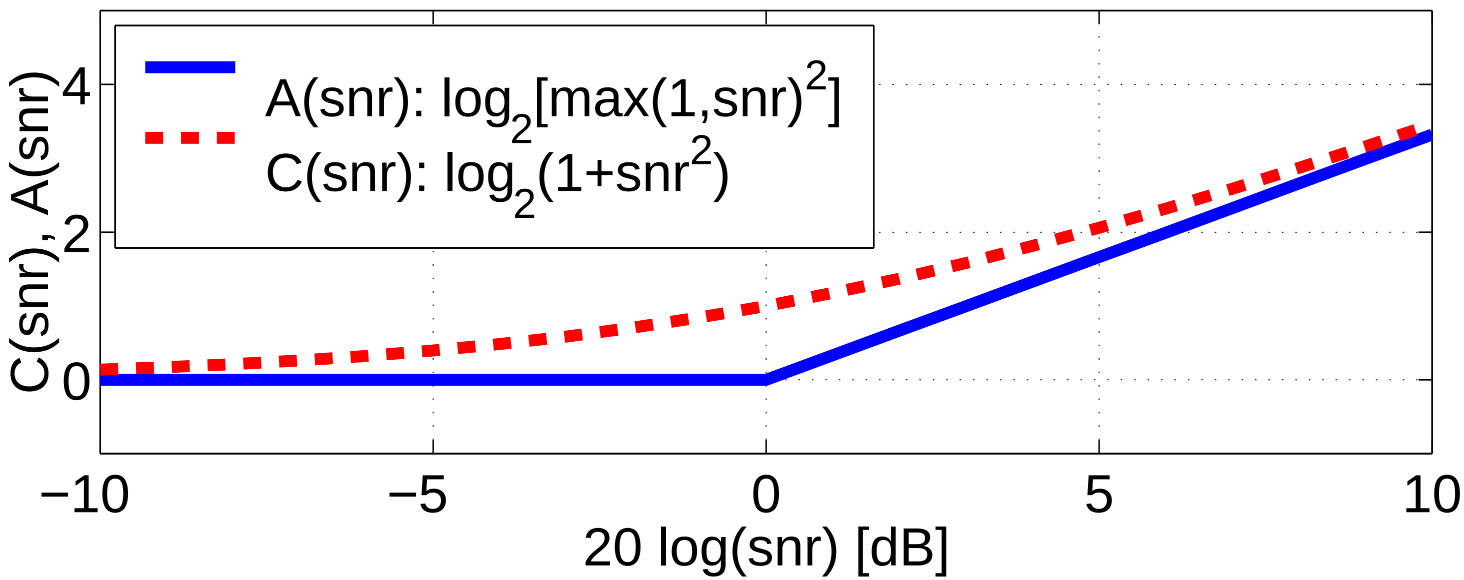
AI AS A CHANNEL CAPACITY

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

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

- and from Shannon (for the Gaussian channel)

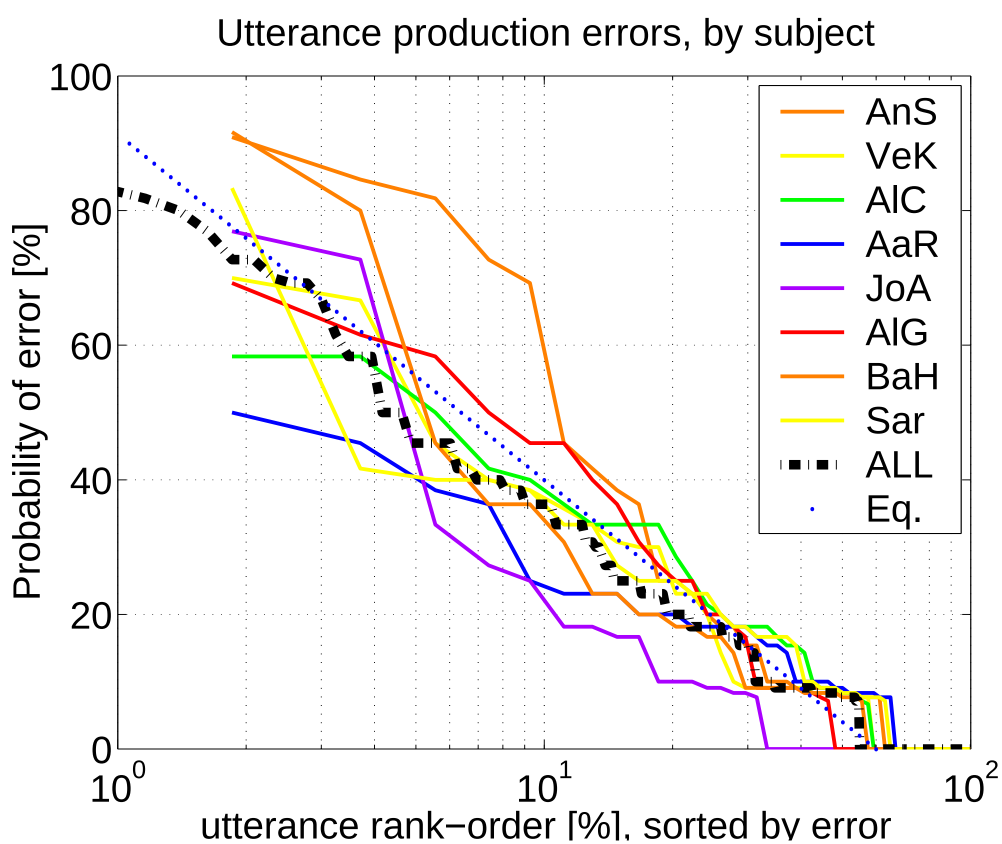
$$C = \int_{-\infty}^{\infty} \log_2[1 + snr^2(f)]df, \qquad (2)$$



TALKER PRODUCTION 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

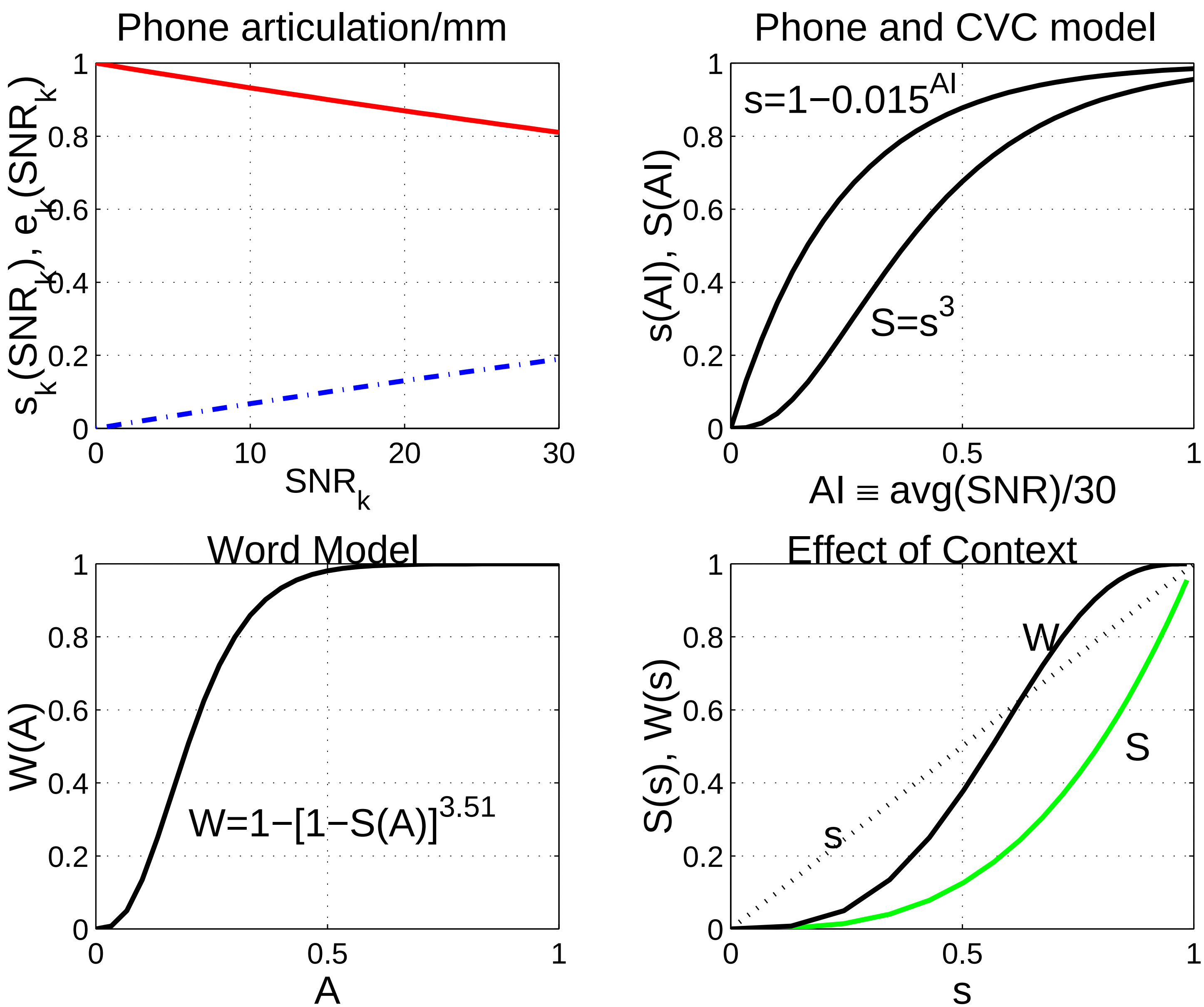


SOURCES OF ERROR

- Talker production errors
  - Production errors are defined by token utterances error over listeners
  - Once poor utterances are identified, they may be selectively removed
    - This method allows us to control the gross error rate
  - With this method we can obtain a 100% score in the clear
    - The price for this is a reduced $N_{utterances}$
- Listener errors (after selectively removing production errors)
  - Listener bias may be determined from individual confusion matrices
  - This bias can be a function of the production error threshold
    - The main effect is on L_2 listeners

EXAMPLE CALCULATIONS

Wide-band channel vs. *SNR*



THE RECOGNITION CHAIN

- The cochlear critical bandwidth defines the *SNR*$_k$
- The *event-error* model: $e_k \propto e_{\min}^{\mathit{SNR}_k}$ (*SNR* in dB units)
- The *average-phone articulation* model:

$$s = 1 - e_1 e_2 \cdots e_k \cdots e_K$$

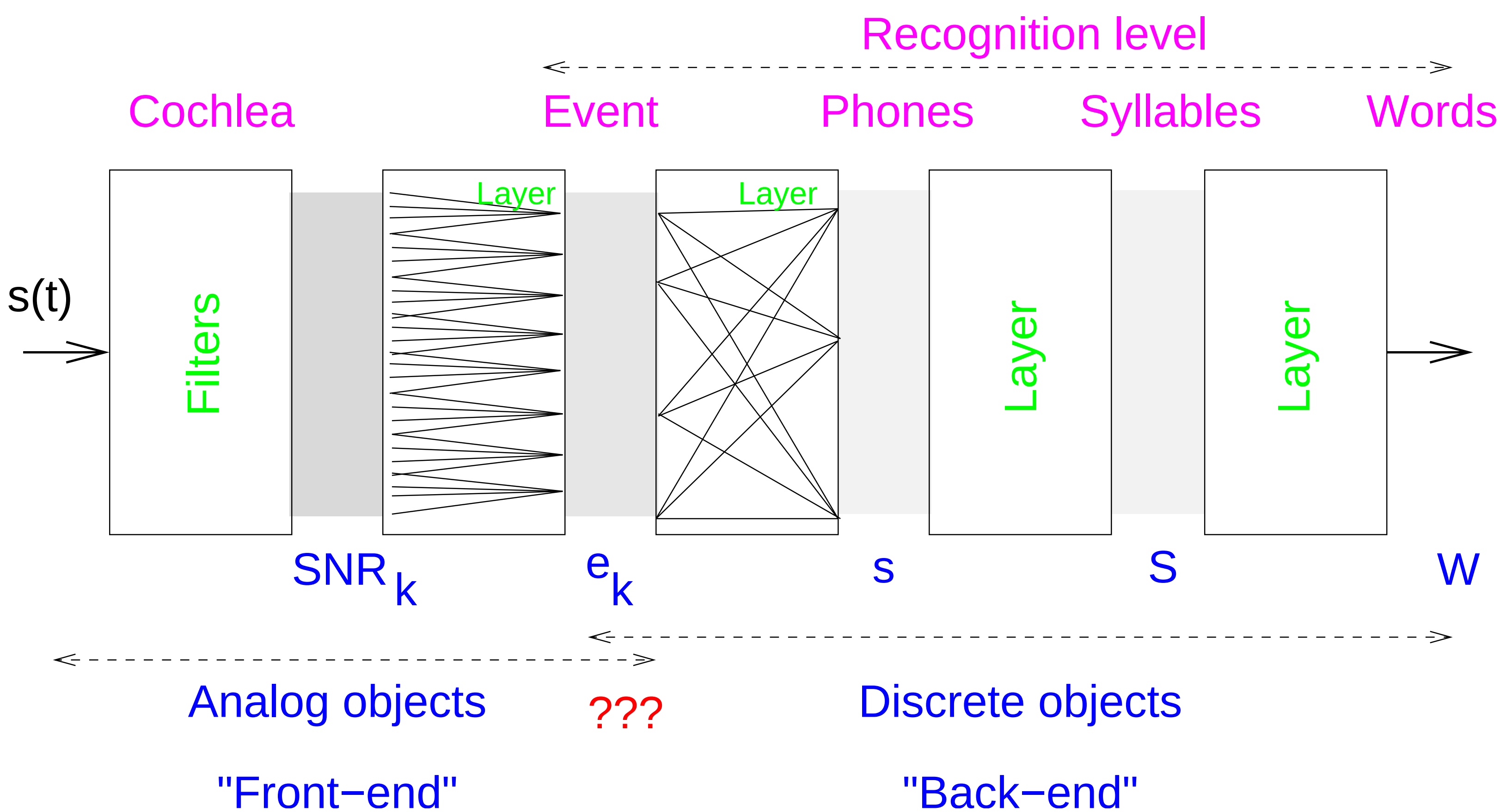
- The nonsense CVC *syllable articulation* model: $S = s^3$
- Heuristic degree of freedom *context models* Boothroyd (see discussion Allen 1994)
  - Word: $W = 1 - (1 - S)^j$
  - Sentence: $I = 1 - (1 - W)^k$
  - Sentence with context: $C = 1 - (1 - I)^l$
- Layers of context:
  - j depends on the ratio of words to pseudo-words in the corpus,
  - k depends on the number of salient words in a sentences,
  - l depends on the word salience and topic context.

COMPOSITION LAWS

- Rules regarding $\Pi_i \; P^{(i)}_{\rm error}$ versus product $\Pi_i \; P^{(i)}_{\rm correct}$?
  - Parallel processing: $P_e = \Pi_k e_k$
    - $\ast$ Errors in many bands have no effect
    - $\ast$ One band with small error (i.e., $e_k = 0$) dominates e.g., $e = e_{_L} e_{_H}$; $e = e_1 e_2 \cdots e_K$; the McGurk example
  - Serial processing: $P_c = \Pi_k s_k$
    - $\ast$ All items of a string must be correct for success e.g., $S_{cvc} = cvc \approx s^3$; $S_{cv} \approx s^2$
- HSR seems to be a problem in combinatorics, of elementary pre-phonic events.

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

SUMMARY OF MODEL RESULTS

- Hierarchical probability relations:

  band *SNR* $\rightarrow$

  band errors (events) $\rightarrow$

  phoneme errors $\rightarrow$

  syllable errors $\rightarrow$

  nonsense word errors $\rightarrow$

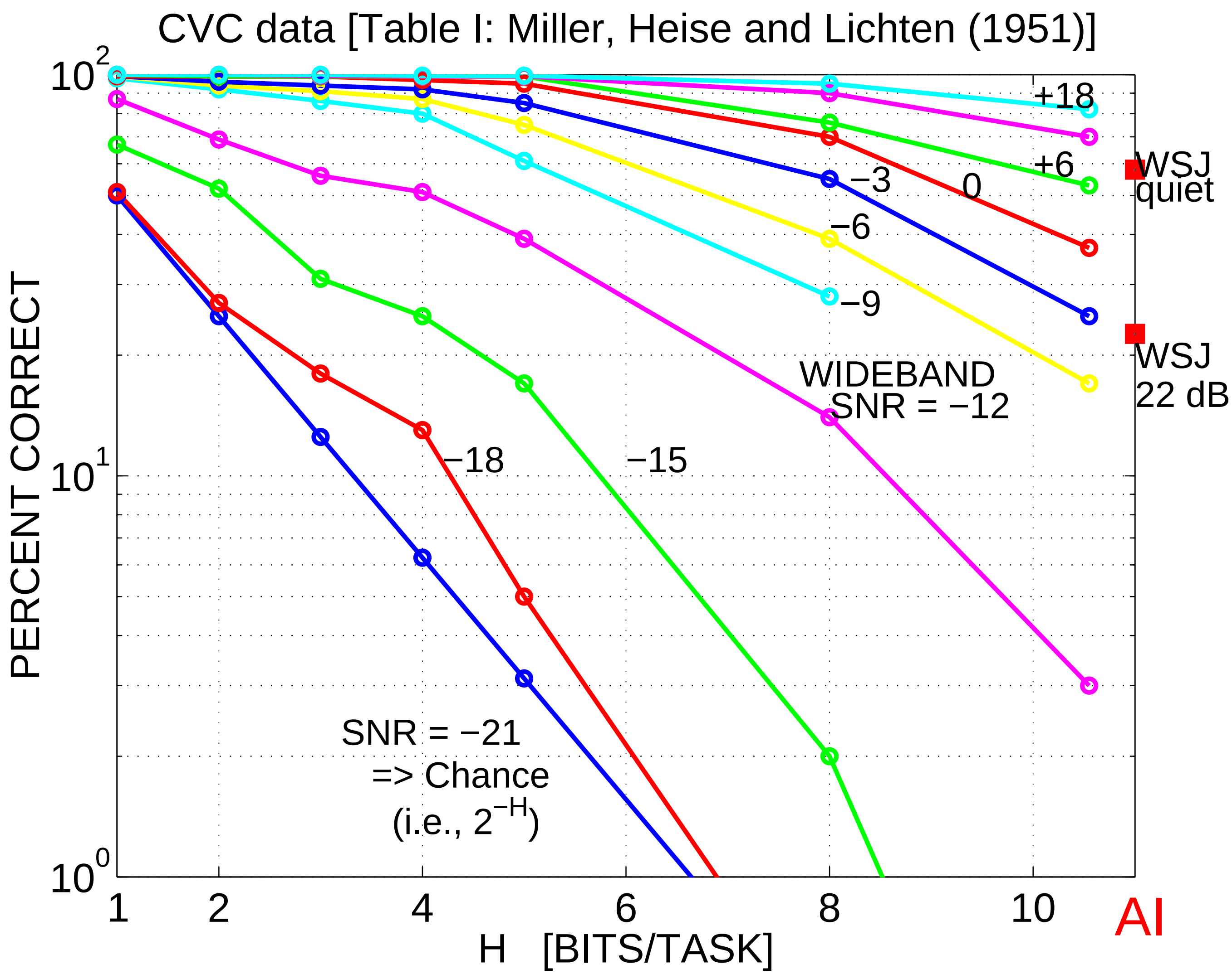
  true word errors, etc.

- The HSR error is established well before language is accessed!

  HSR error depends only on the *SNR* in bands

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



GRAMMATICAL CONTEXT

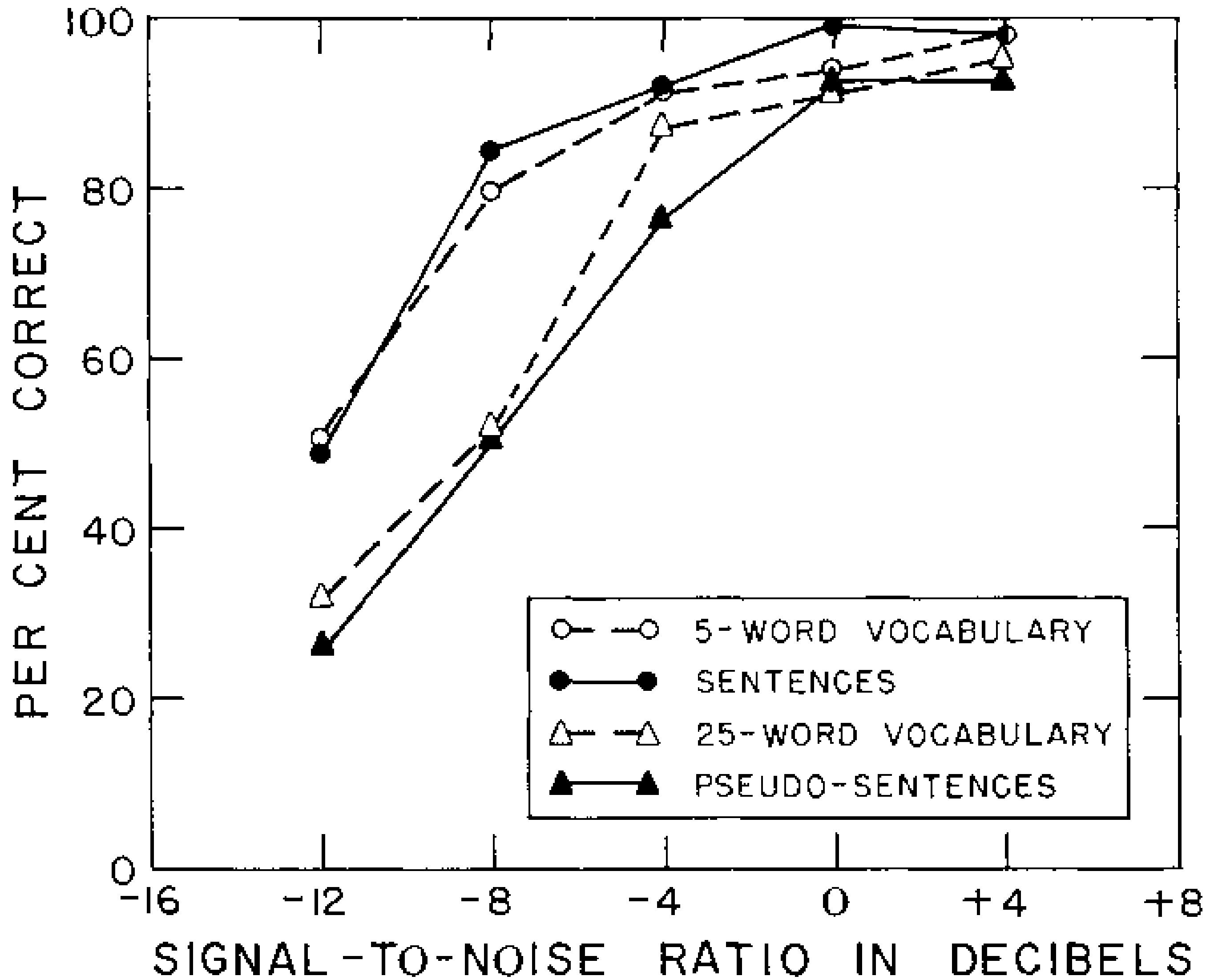
- Five groups of five words that form grammatical sentences:

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



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*

MILLER'S BINARY FEATURES

- Miller & Nicely derived binary consonant features [i.e., events]

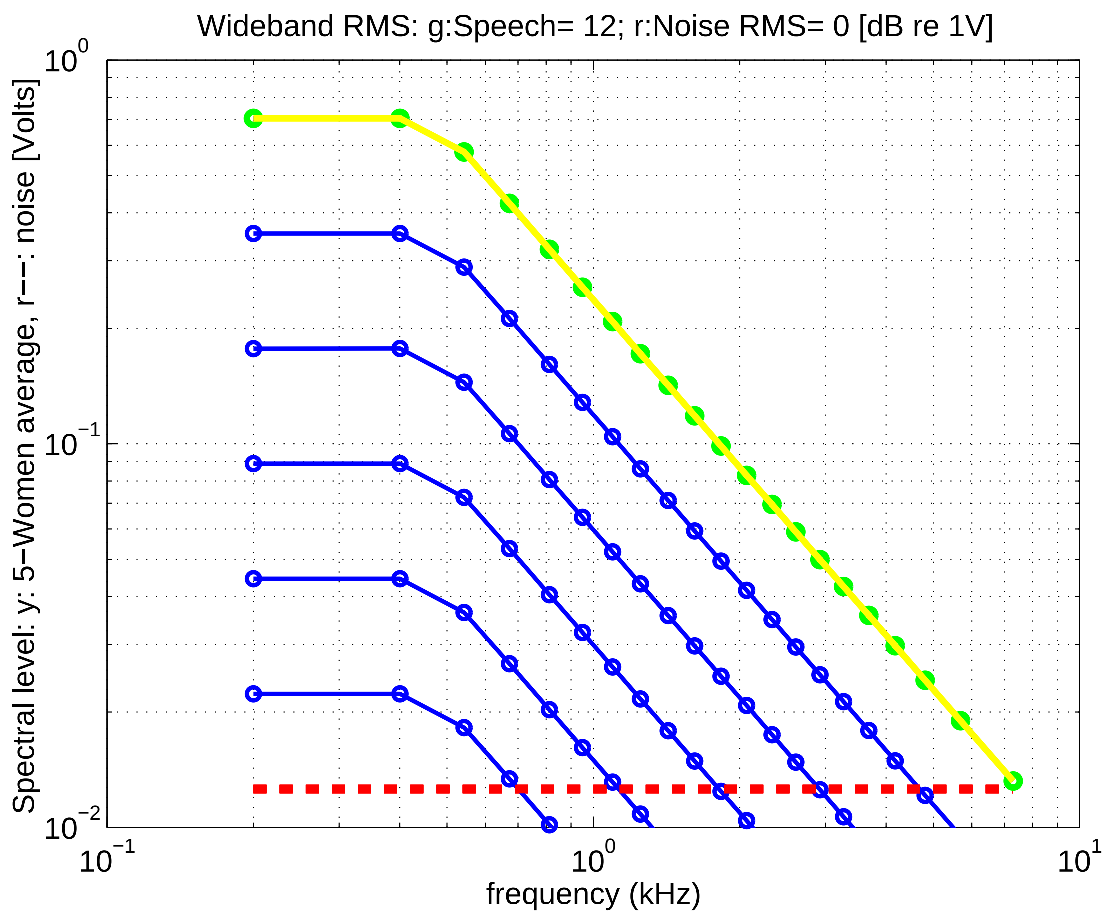
TABLE XIX. Classification of consonants used to analyze confusions.

Consonant	Voicing	Nasality	Affrication	Duration	Place
p	0	0	0	0	0
t	0	0	0	0	1
k	0	0	0	0	2
f	0	0	1	0	0
θ	0	0	1	0	1
s	0	0	1	1	1
ʃ	0	0	1	1	2
b	1	0	0	0	0
d	1	0	0	0	1
g	1	0	0	0	2
v	1	0	1	0	0
ð	1	0	1	0	1
z	1	0	1	1	1
ʒ	1	0	1	1	2
m	1	1	0	0	0
n	1	1	0	0	1

" . . . the impressive thing to us was that . . . the [binary] features were perceived almost independently of one another." *–Miller & Nicely 1955*

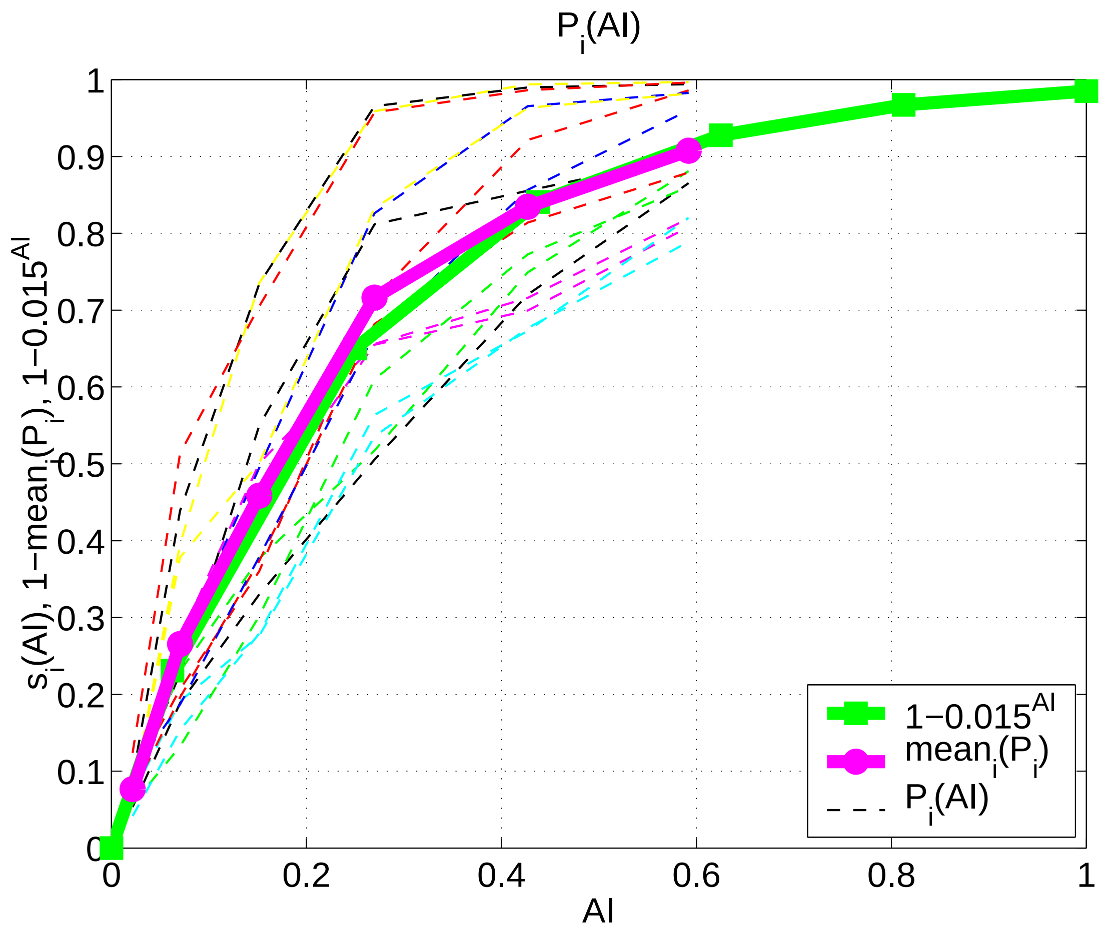
FINDING THE AI FOR MILLER NICELY TALKERS

- Average spectrum for female talkers



TRACE OF MILLER NICELY AND THE AI

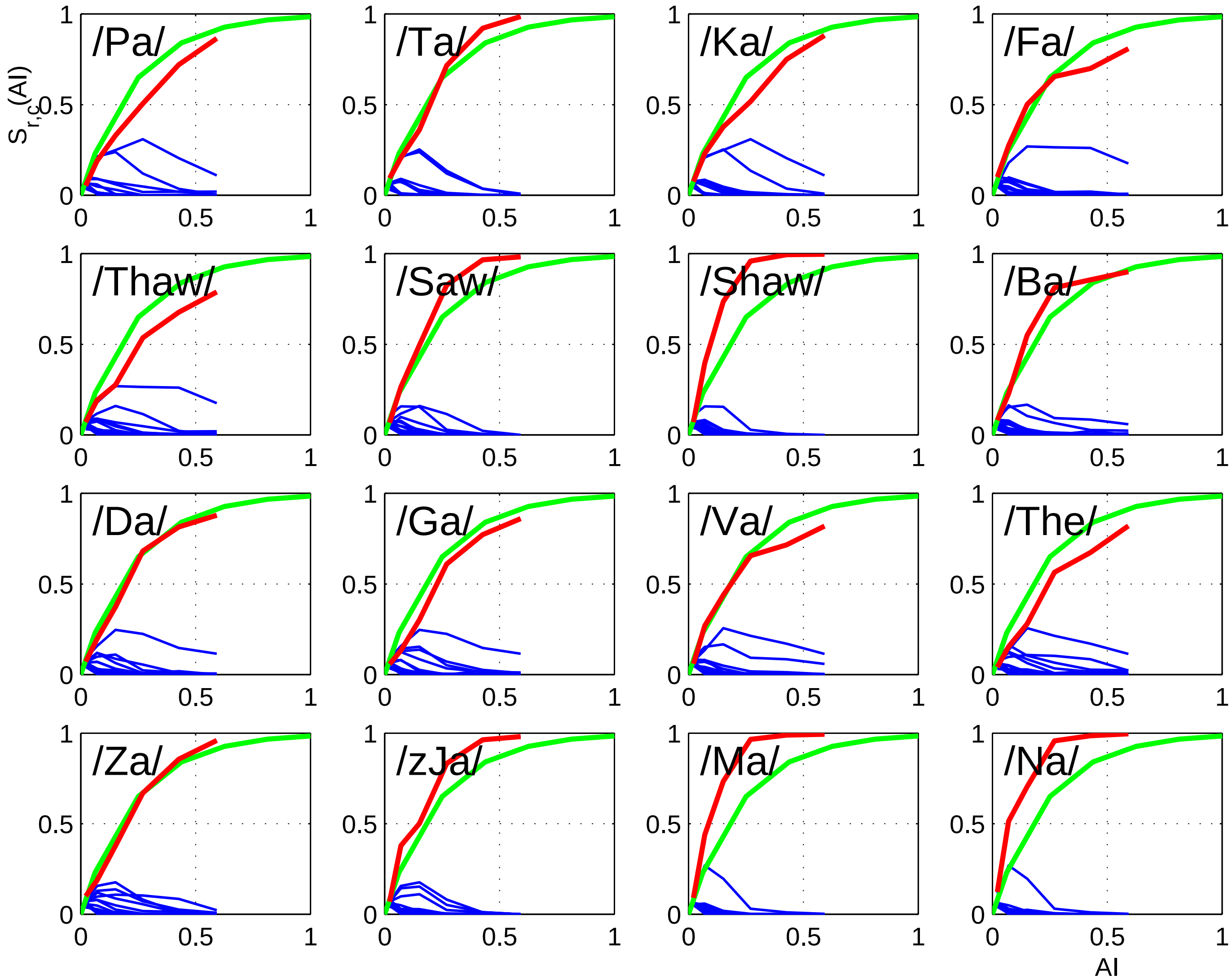
- Next we look at the average PI function vs. AI



SYMMETRIC COMPONENT OF $P_C(SNR)$

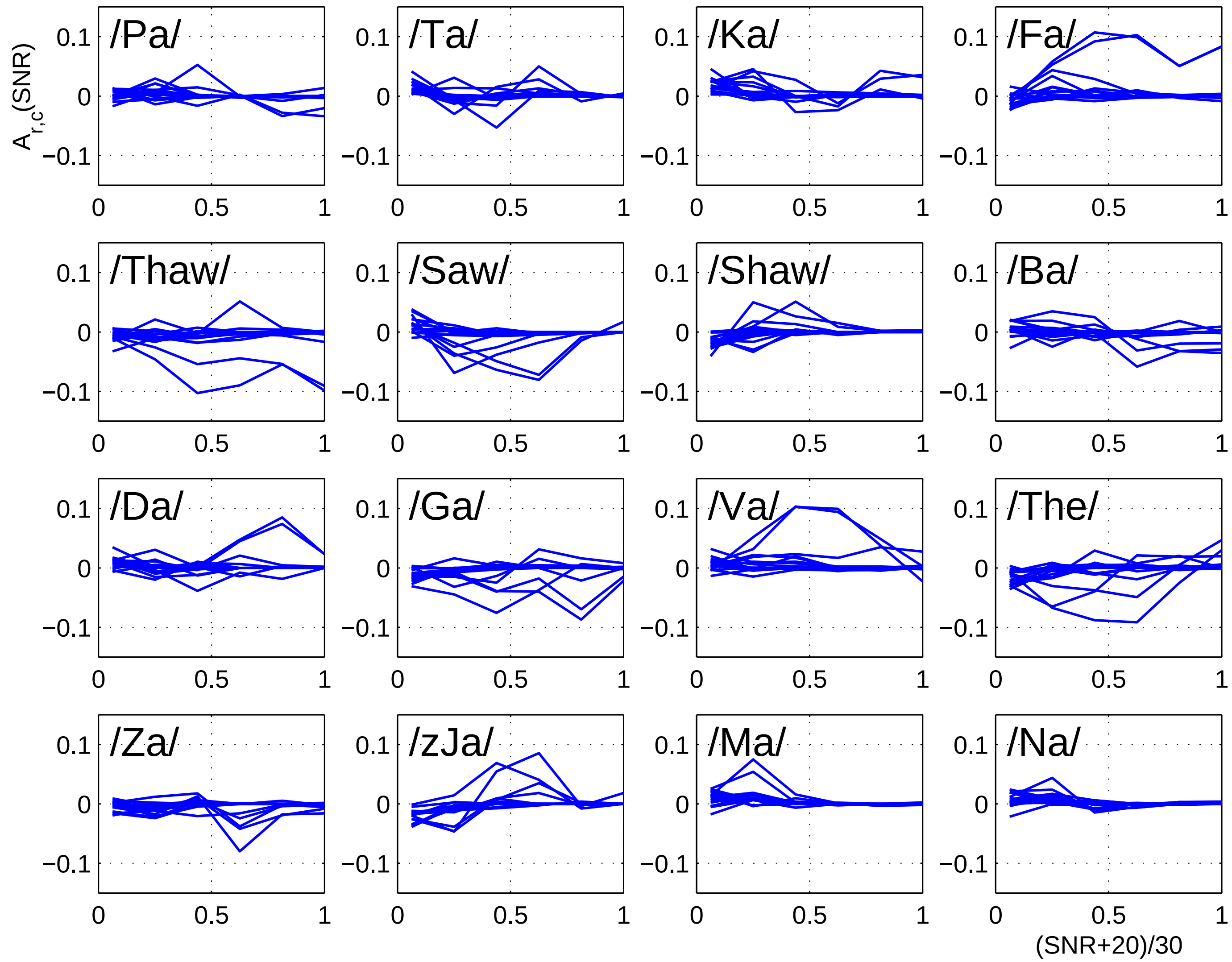
- We stand to learn from linear operations on $P_{ij}(snr)$

  Symmetric: $S_c(snr) \equiv [P_{ij}(snr) + P_{ji}(snr)]/2$



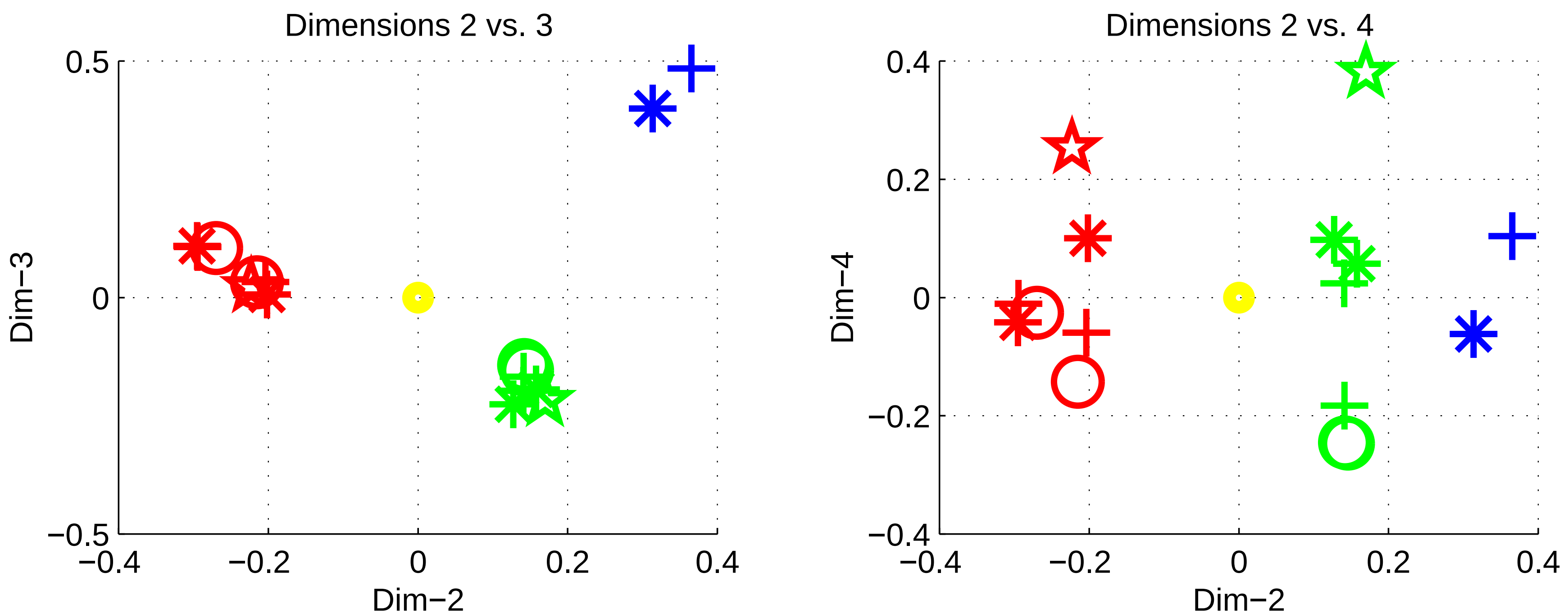
SKEW-SYMMETRIC COMPONENT OF $P_C(SNR)$

- Skew: $A_c(\mathit{snr}) \equiv [P_{ij}(\mathit{snr}) - P_{ji}(\mathit{snr})]/2$



SVD REPRESENTATION OF THE PERCEPTUAL SPACE

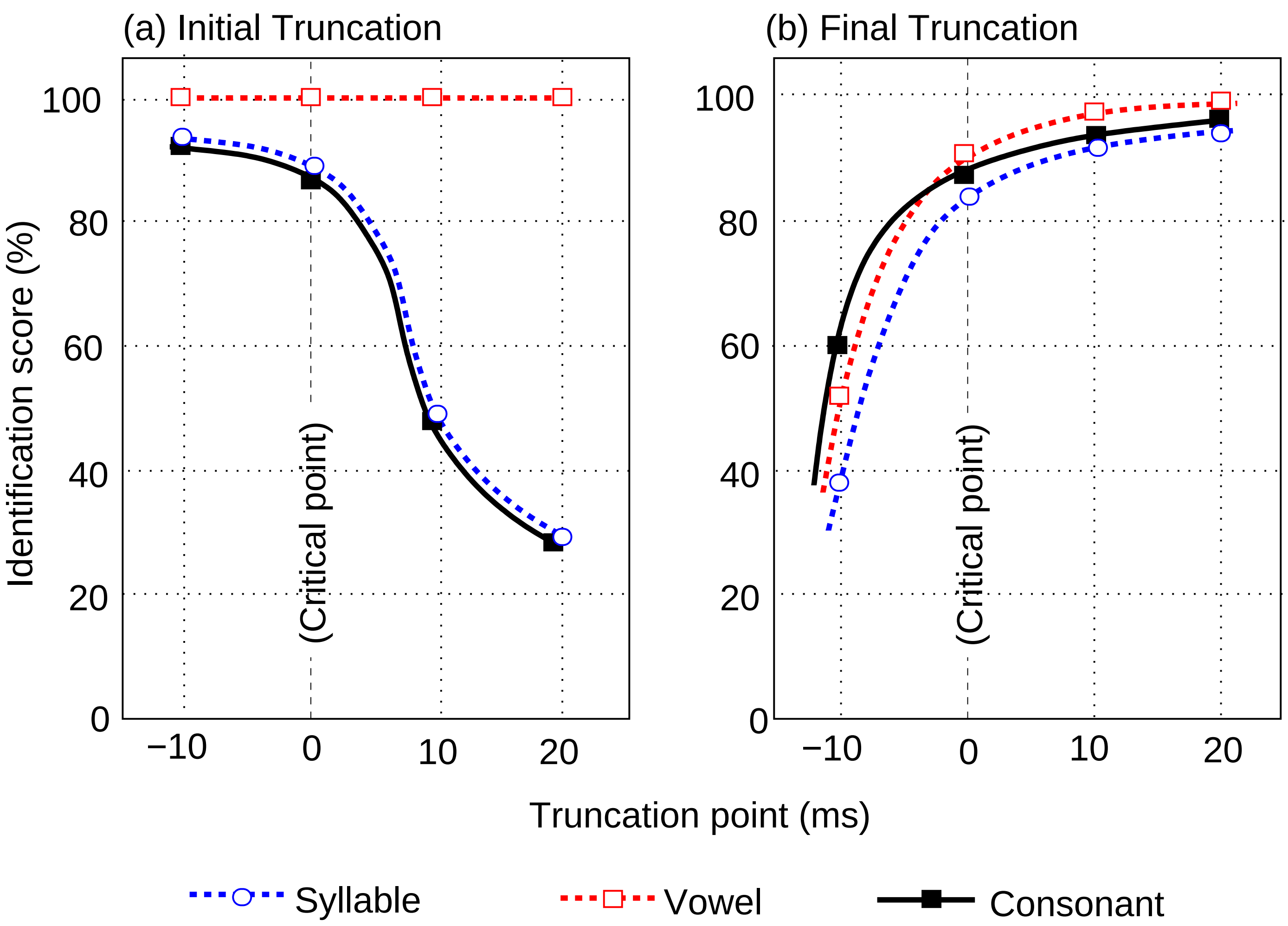
- 4^{dim} SVD perceptual representation of the confusion matrix



DEMO

TEMPORAL RESOLUTION OF PHONE RECOGNTION

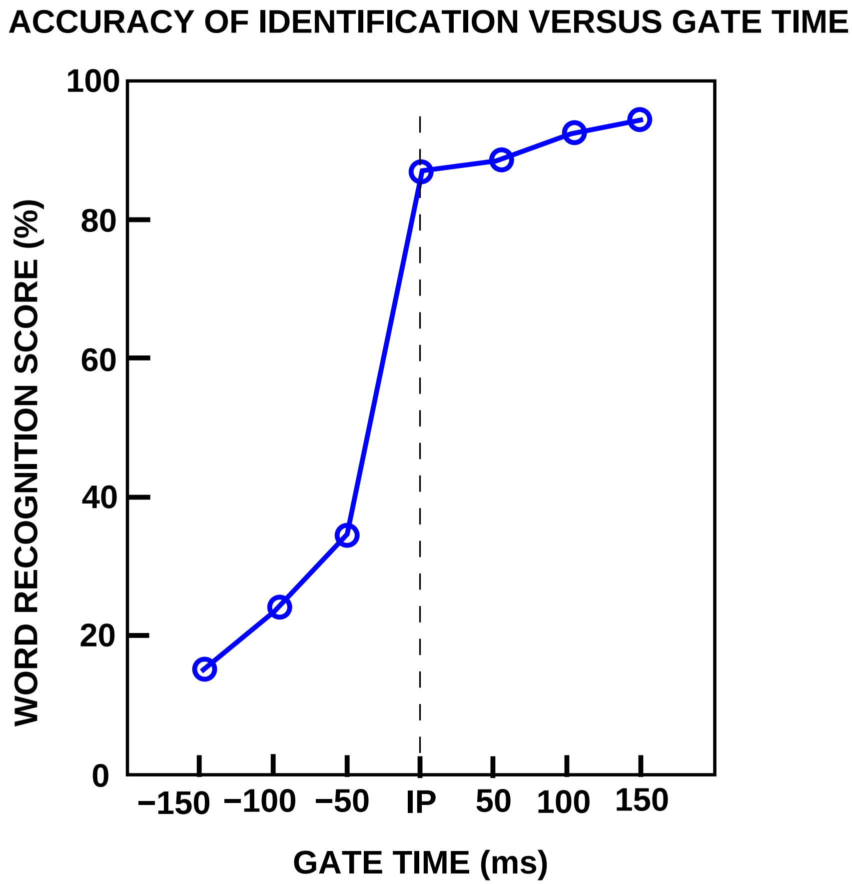
- Phones are recognized in on a 10 ms time scale (Furui 1986)



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*

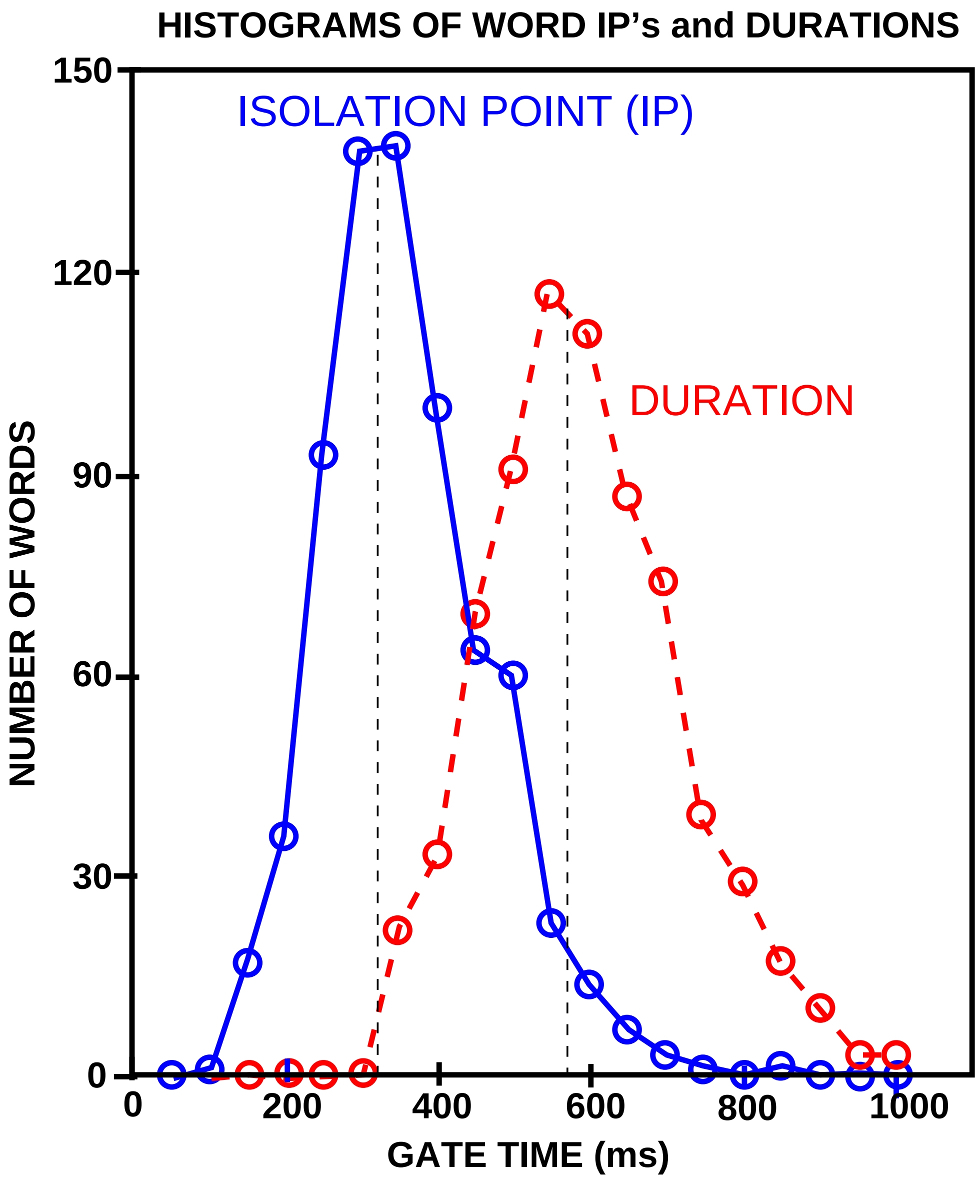
  Expt. I – Neutral sentences: “The next word is *test-word*.”



- Categorical perception

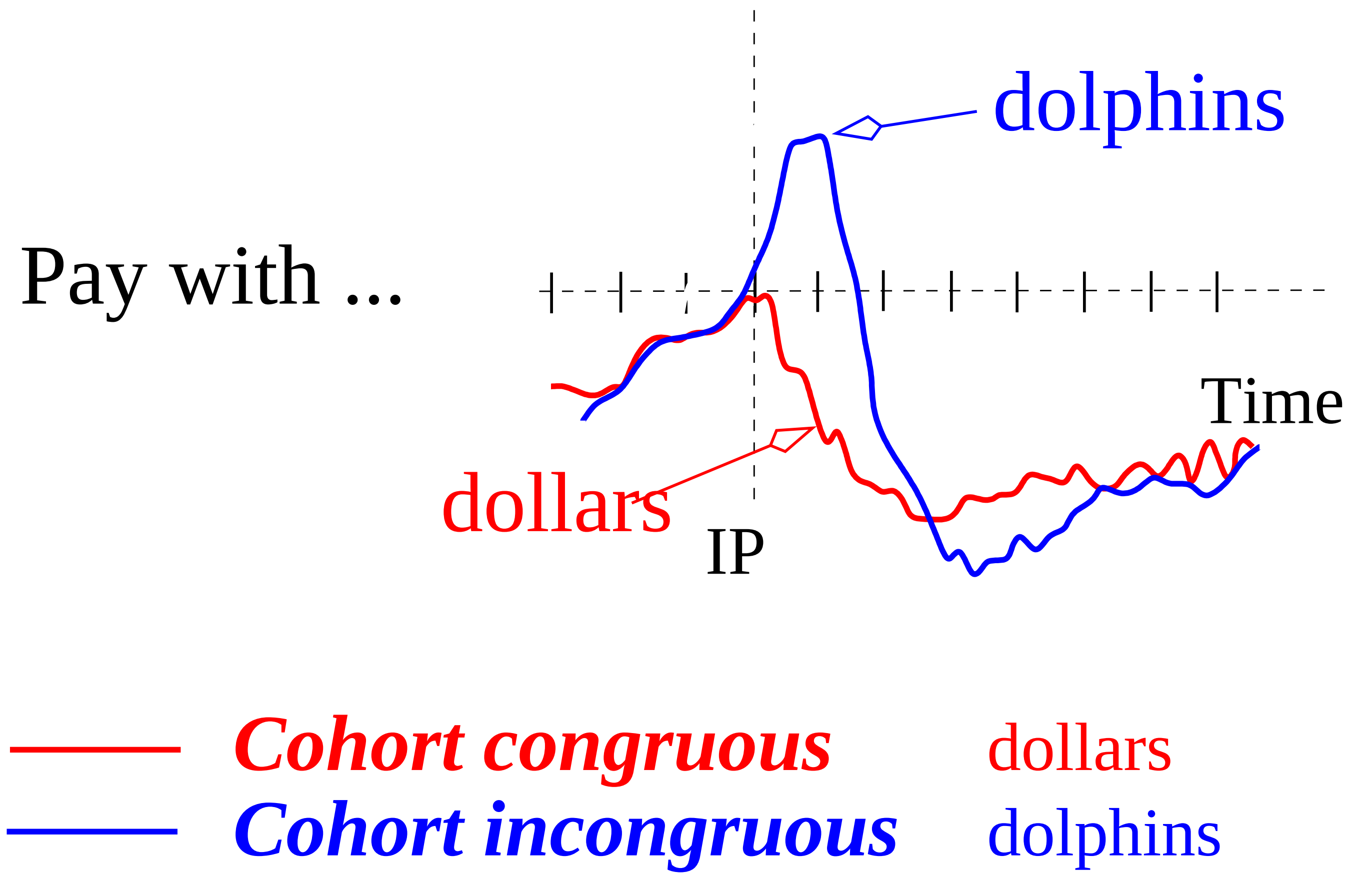
WORD SEMANTICS: IP VS. DURATION

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



ERP MEASURE OF CONTEXT RE IP

- Expt. II – Event related scalp potential (N-400 ERP) re IP, from Exp. I Sentence semantics effects



- Words are recognized on a syllable by syllable basis, within 50 ms
- Context is recognized on a syllable by syllable basis, within 200 ms

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

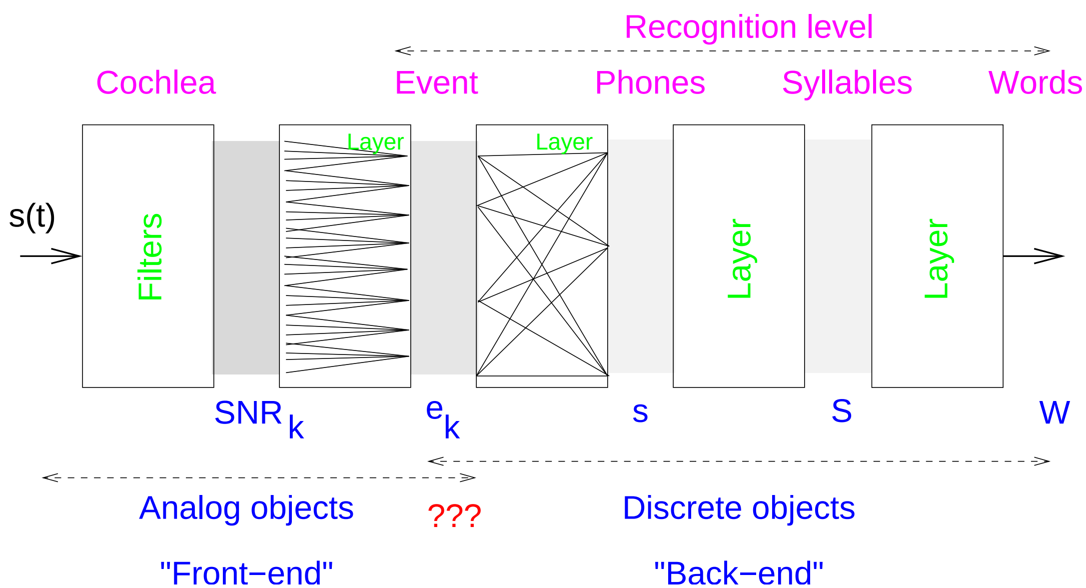
  Events [Miller's features]

CATEGORIAL PERCEPTION

- Meaningful words are recognized before they end
- Syllables are recognized within 50 ms

SUMMARY

- Miller & Nicely found 5 independent channels, described by discrete events [Miller's features]
- Speech is recognized in layers:
  $\mathrm{SNR}_k \Rightarrow$ events $\Rightarrow$ phones $\Rightarrow$ syllables $\Rightarrow$ words $\Rightarrow \ldots$



- Language model performance is independent of noise robustness!
- To study HSR, entropy must be controlled
- Speech psychophysics is an important tool for studying HSR

FUTURE GOALS

- Use psychophysics to gain insight into event extraction
- The next break through:
  - More robust ASR
  - An event extracting hearing aid

This talk may be found at:

http://auditorymodels.org/jba/PAPERS/ICASSP/

References

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

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

Boothroyd, A. and Nittrouer, S. (1988). “Mathematical treatment of context effects in phoneme and word recognition,” *J. Acoust. Soc. Am.* **84**(1):101–114.

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.

Bronkhorst, A. W., Brand, T., and Wagener, K. (2002). “Evaluation of context effects in sentence regonition,” *J. Acoust. Soc. Am.* **111**(6):2874–2886.

Campbell, G. A. (1910). “Telephonic intelligibility,” *Phil. Mag.* **19**(6):152–9.

French, N. R. and Steinberg, J. C. (1947). “Factors governing the intelligibility of speech sounds,” *J. Acoust. Soc. Am.* **19**:90–119.

Furui, S. (1986). “On the role of spectral transition for speech perception,” *Journal of the Acoustical Society of America* **80**(4):1016–1025.

Miller, G. A. (1951). *Languange and communication*. McGraw Hill, New York.

Miller, G. A., Heise, G. A., and Lichten, W. (1951). “The intelligibility of speech as a function of the context of the test material,” *J. Exp. Psychol.* **41**:329–335.

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.

Rayleigh, L. (1908). “Acoustical notes – viii,” *Philosophical Magazine* **16**(6):235–246.

Shannon, C. E. (1948). “The mathematical theory of communication,” *Bell System Tech. Jol.* **27**:379–423 (parts I, II), 623–656 (part III).

Shannon, C. E. (1951). “Prediction and entropy of printed english,” *Bell System Tech. Jol.* **30**:50–64.

Van Petten, C., Coulson, S., Rubin, S., Planten, E., and Parks, M. (1999). “Time course of word identification and semantic integration in spoken language,” *J. of Exp. Psych.: Learning, Memory and Cognition* **25**(2):394–417.

Wang, M. D. and Bilger, R. C. (1973). “Consonant confusions in noise: A study of perceptual features,” *J. Acoust. Soc. Am.* **54**:1248–1266.