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

Jont B. Allen
Retired:
AT&T Labs Research
Florham Park NJ, 07973

jba@auditorymodels.org

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

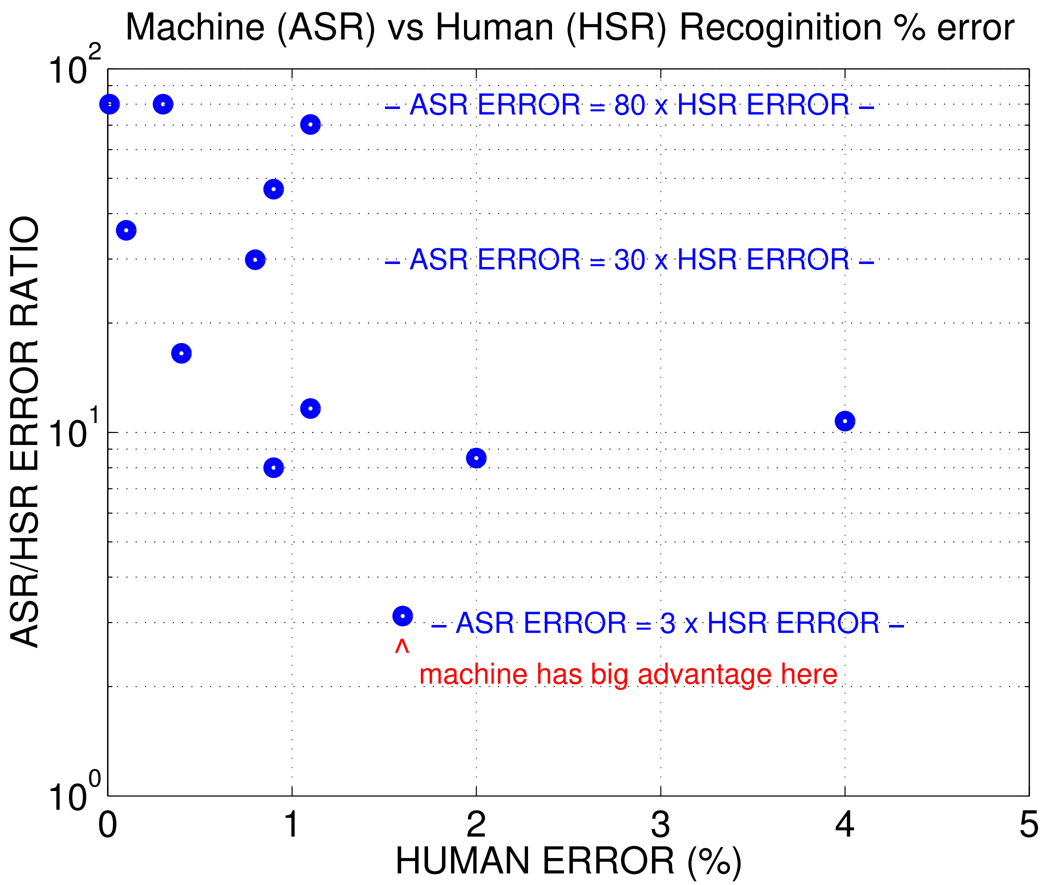
HSR VS. ASR ERROR

- Table summarizing the results of Lippmann 1997, sorted on the ratio of machine to human error.

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

MOTIVATION

- Lippmann (1997) compared human (HSR) and machine (ASR)



- Machine recognition error 3-80 times human rate
- Modern Speech Recognizers are not robust to noise

ABBREVIATIONS

ASR	**Automatic Speech Recognition**
HSR	**Human Speech Recognition**
SNR	**Signal to Noise Ratio**
AI	**Articulation Index**
IT	**Information Theory**
CV	**consonant-vowel** (ex. “pa, zee”)
CVC	**consonant-vowel-consonant** (ex. “poz, hub”)
VOT	**Voice onset time**

JONT's DEFINITIONS

phone	A speech sound e.g., consonant, vowel, nonsense word
word	A meaningful phone or phone cluster
phoneme	The smallest phone conveying a distinction in meaning
allophones	All the phone variants for a given phoneme
recognition	Probability measure p_n of correct phoneme identification
intelligibility	Recognition of words (i.e., meaningful speech)
articulation	Recognition of "nonsense words"
robustness	Relative recognition with filtering and noise
event	A binary subunit of articulation [e.g., Voicing: /ba/ vs. /pa/]
trial	A single presentation of a set of events
context	A phone sequence constraint (e.g., words have context)
information	$I_n = \log_2(1/p_n),\ n = 1, \cdots, N$
entropy	Average information: $\mathcal{H} = \sum_{n=1}^{N} p_n I_n$
conditional entropy	A measure of context: high entropy $\Longrightarrow$ low context

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
  Details were proprietary within AT&T
  - Detailed review of Fletcher's AI theory: Allen IEEE 1994
- French and Steinberg 1947 following work during WWII
- Shannon 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
- Miller 1962 Grammer $\equiv$ 4 dB of SNR
- Boothroyd JASA 1968; Boothroyd & Nittrouer JASA 1988
- Bronkhorst et al. JASA 1993
- Van Petten *et al.* 1994

WHAT I WANT TO SHOW:

- 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 does not seem to interact with semantic context effects
    - ∗ HSR: robust articulation; excellent context models
    - ∗ ASR: bad articulation; weak context models
  - It is critical to control for language context effects
- Comments:
  - ASR is a top-down strategy, largely driven by low-entropy models
  - For continuity, results will presented in cronological order

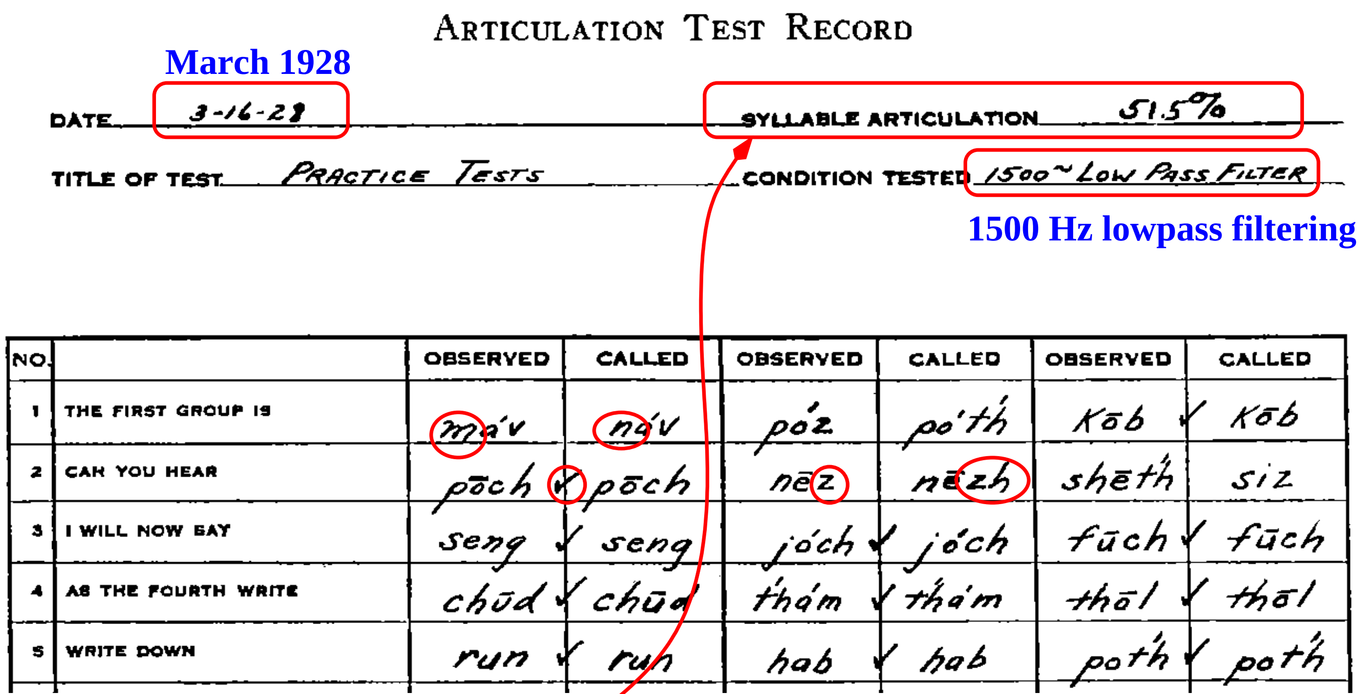
FLETCHER'S ARTICULATION EXPERIMENT

- Play nonsense syllables (CV, VC & CVC) to maximize sound entropy
  - Max(Entropy) $\Leftrightarrow$ Min(context effect)
- Hold the speech corpus constant for each experiment
  - constant source entropy
- Average over many (e.g., 10x10) talker-listener pairs
- Vary the phone articulation by:

  changing the SNR
  LP-HP filtering the speech

TYPICAL ARTICULATION TEST RECORD

- Basic method of phone (nonsense syllables) 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	shēth	siz
3	I WILL NOW SAY	seng	✓ seng	jóch	✓ jóch	fūch	✓ fūch
4	AS THE FOURTH WRITE	chūd	✓ chūd	thám	✓ thám	thōl	✓ thōl
5	WRITE DOWN	run	✓ run	hab	✓ hab	póth	✓ póth

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)

WHAT THEY FOUND

- 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” !!!

QUESTION

- What do these models imply about coarticulation?

THE NEXT STEP

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

SPECIFIC DEFINITIONS

SYMBOL	DEFINITION
α	gain applied to the speech
$c(\alpha) \equiv P_c(\text{consonant}\vert\alpha)$	consonant articulation
$v(\alpha) \equiv P_c(\text{vowel}\vert\alpha)$	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
$S(\alpha)$	nonsense syllable (CVC) articulation
$W(\alpha)$	word intelligibility
$I(\alpha)$	sentence intelligibility

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

$$\mathcal{A}(s_L) + \mathcal{A}(s_H) = \mathcal{A}(s)$$

  - 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:

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

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

  - This relationship took years to discover from the empirical curves

- 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

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

where

$$e_i \equiv 1 - s_i$$

–This formula forms the basis of articulation index theory

–It was never *formally* tested

–Why $K = 20$ bands?

Each band equals 1mm along the basilar membrane

–It *was* observed to hold over a hundreds of transmission systems, giving a solid indirect confirmation

- 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

MODEL OF BAND EVENT ERRORS

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

$$e_k = e_{min}^{\mathrm{SNR}_k / K}$$

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

$$0 \leq SNR_k \leq 1.$$

- Total error:

$$e = e_1 e_2 \cdots e_K = e_{min}^{(\mathrm{SNR}_1 + \mathrm{SNR}_2 \cdots \mathrm{SNR}_K)/K}$$

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

THE RECOGNITION CHAIN

- The cochlear critical bandwidth defines the SNR_k
- The *event-error* model: $e_k \propto e_{\min}^{\mathrm{SNR}_k}$ (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* Fletcher; Boothroyd; see 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 $\Pi_i \, P^{(i)}_{\mathrm{error}}$ versus product $\Pi_i \, P^{(i)}_{\mathrm{correct}}$?
  - Parallel processing $\Rightarrow P_e = \Pi_k e_k$
    e.g., $e = e_L e_H$ and the McGurk example
  - Serial processing $\Rightarrow P_c = \Pi_k s_k$
    e.g., $S = s^3$
- HSR seems to be a problem in combinatorics,
  of elementary events.

CONCLUSIONS ABOUT FLETCHER'S AI (HSR)

- Context effects are strong and confound the study of recognition
- To study HSR, we must control for language context
  - Maximizing entropy factors HSR models (e.g., $S = s^3$)
- We recognize speech based on a cascade of layers
  - Entropy decreases along this cascade
- The phone $s \equiv P_c$ is derived from independent event error probabilities

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

- Elementary events seem to account for Fletcher's "independent-band articulation" channels
- Each event error probability depends on the band SNR$_k$, not the energy

$$e_k \propto e_{min}^{\mathrm{SNR}_k}$$

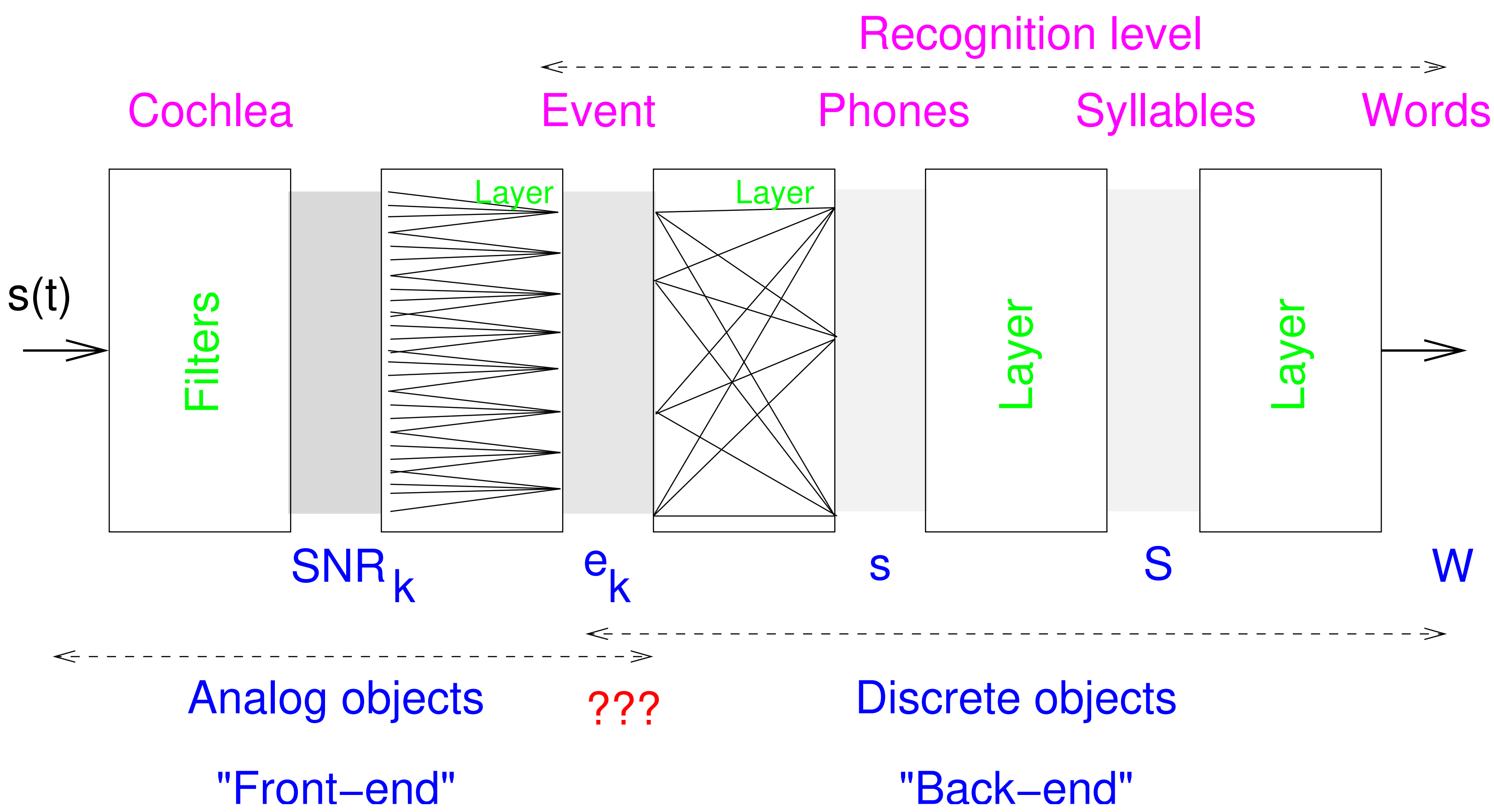
SUMMARY OF FLETCHER'S RESULTS

- 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

HOW WE RECOGNIZE SPEECH?

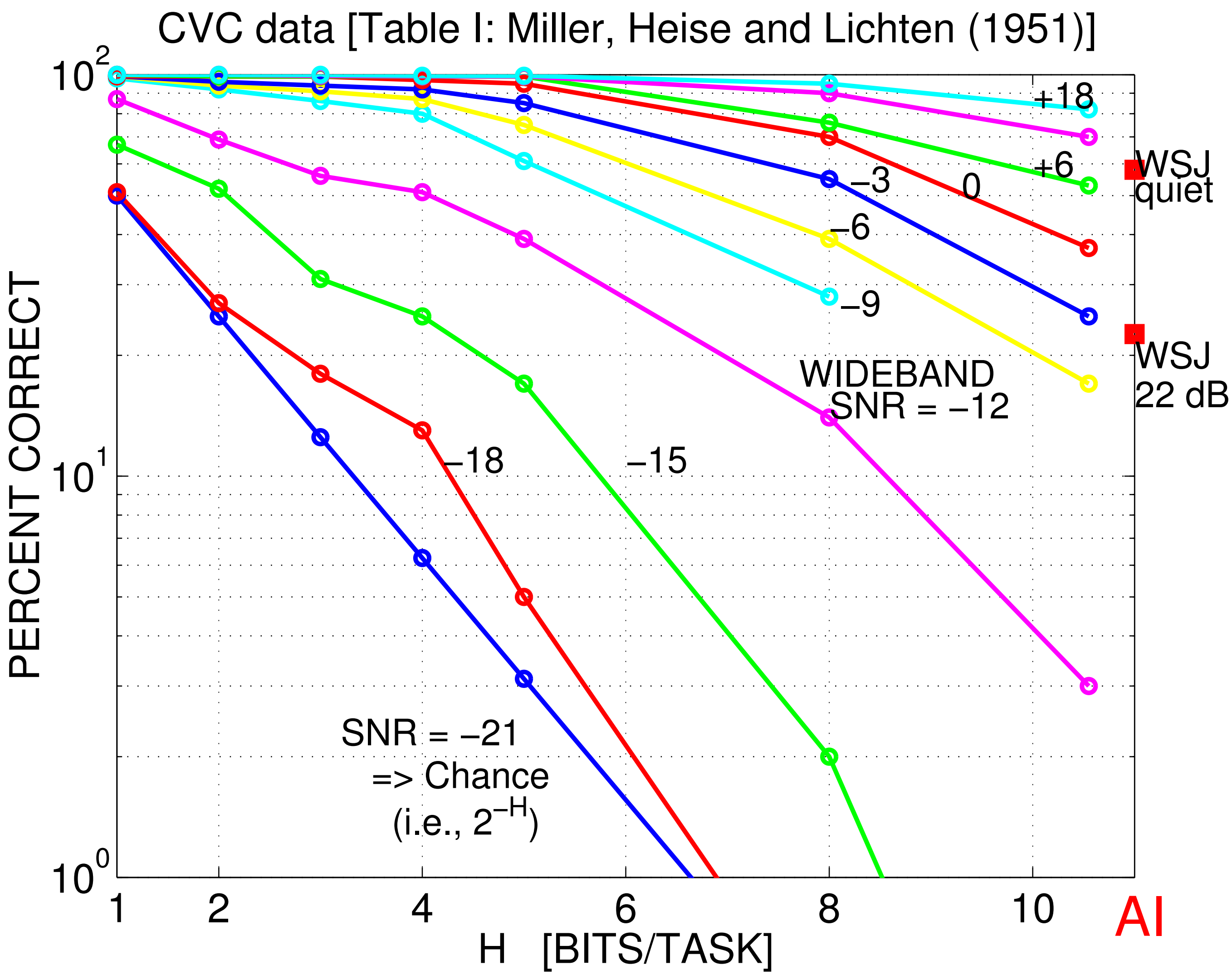
- Hierarchical "bottom up" analysis
- Accurate statistical models of performance at each stage



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

SPEECH ENTROPY VS. THE WIDEBAND SNR

- $P_c(\mathcal{H}, \mathrm{SNR})$ 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

MILLER'S BINARY FEATURES

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

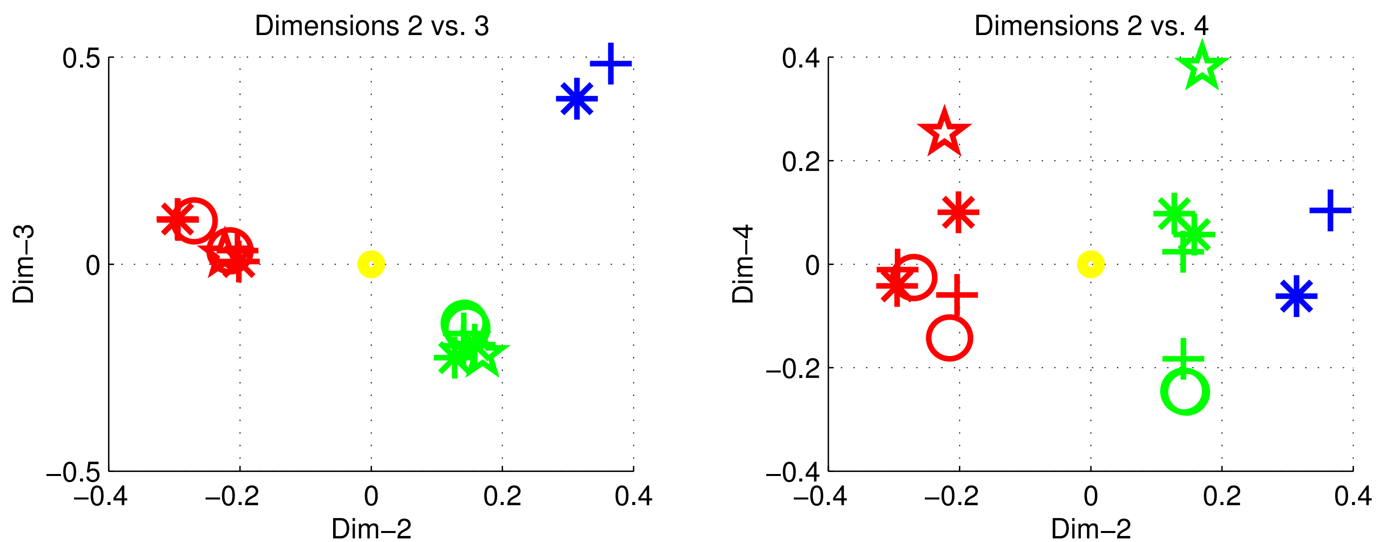
TABLE XIX. Classification of consonants used to analyze confusions.

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*

SVD REPRESENTATION OF THE PERCEPTUAL SPACE

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

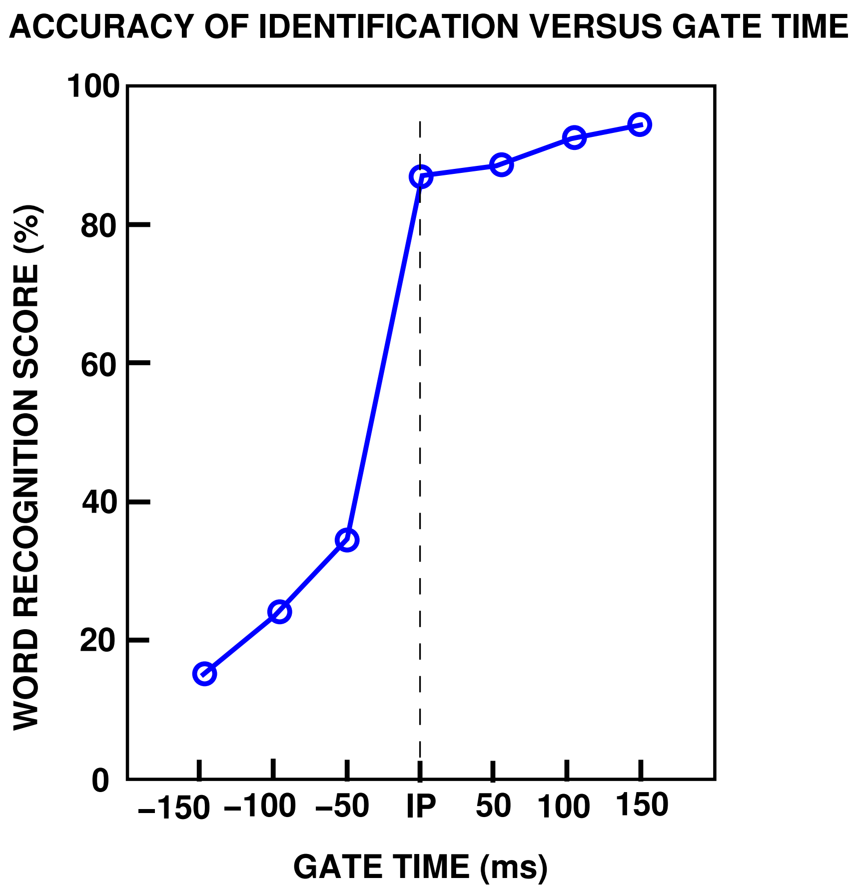


DEMO

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*

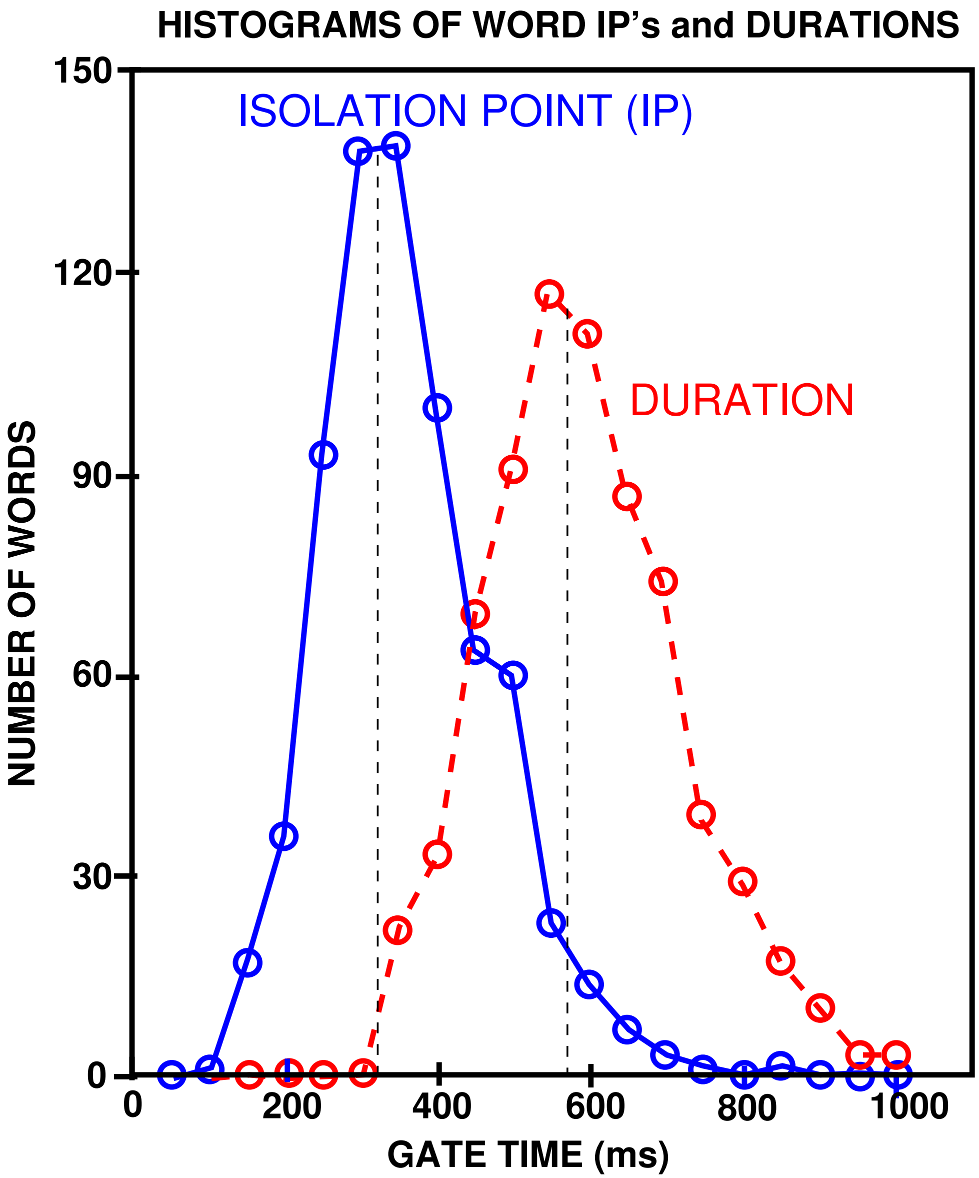
  Expt. I – Neutral sentences: “The next word is *test-word*.”



- Categorical perception

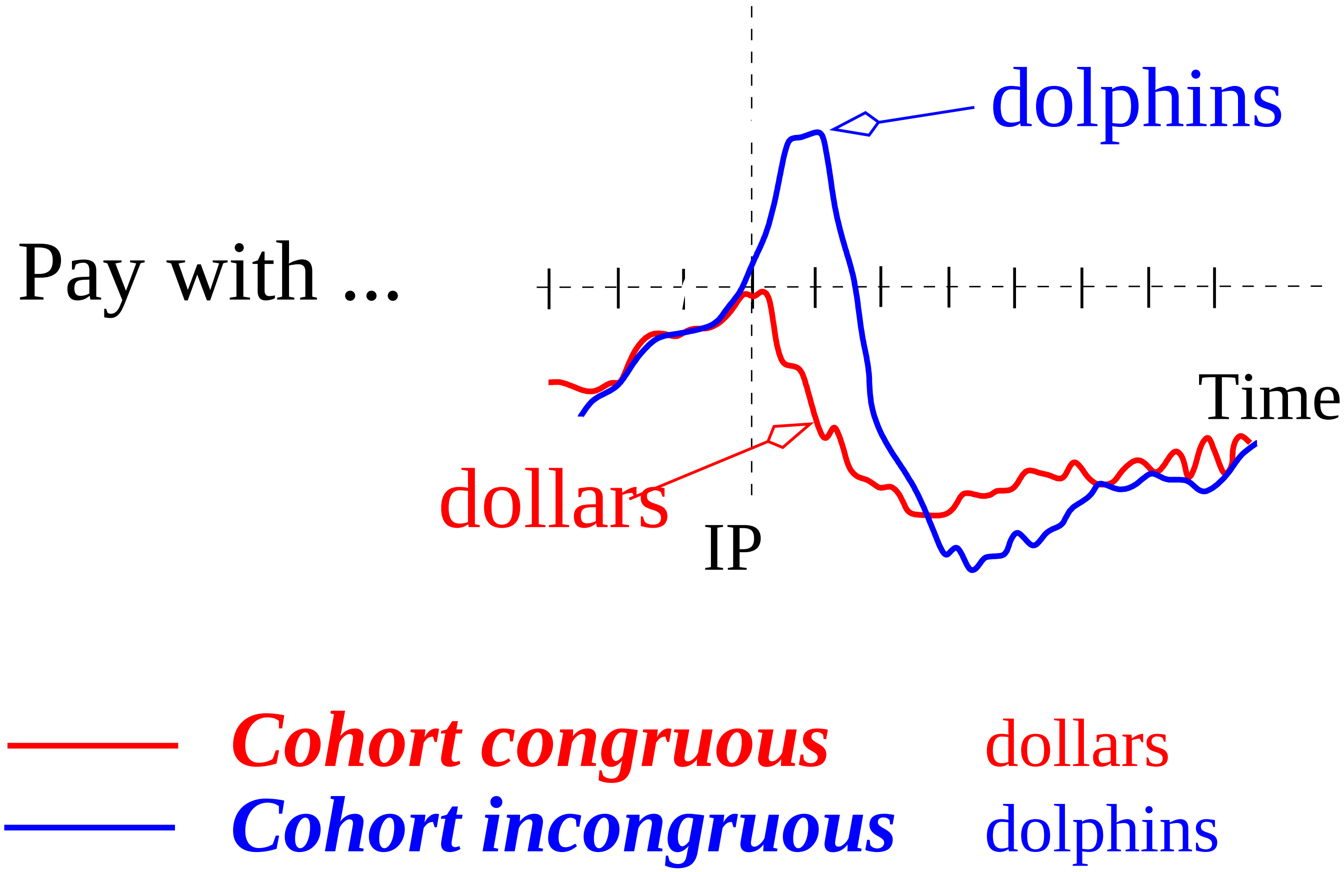
WORD SEMANTICS: IP VS. DURATION

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



ERP MEASURE OF CONTEXT RE IP

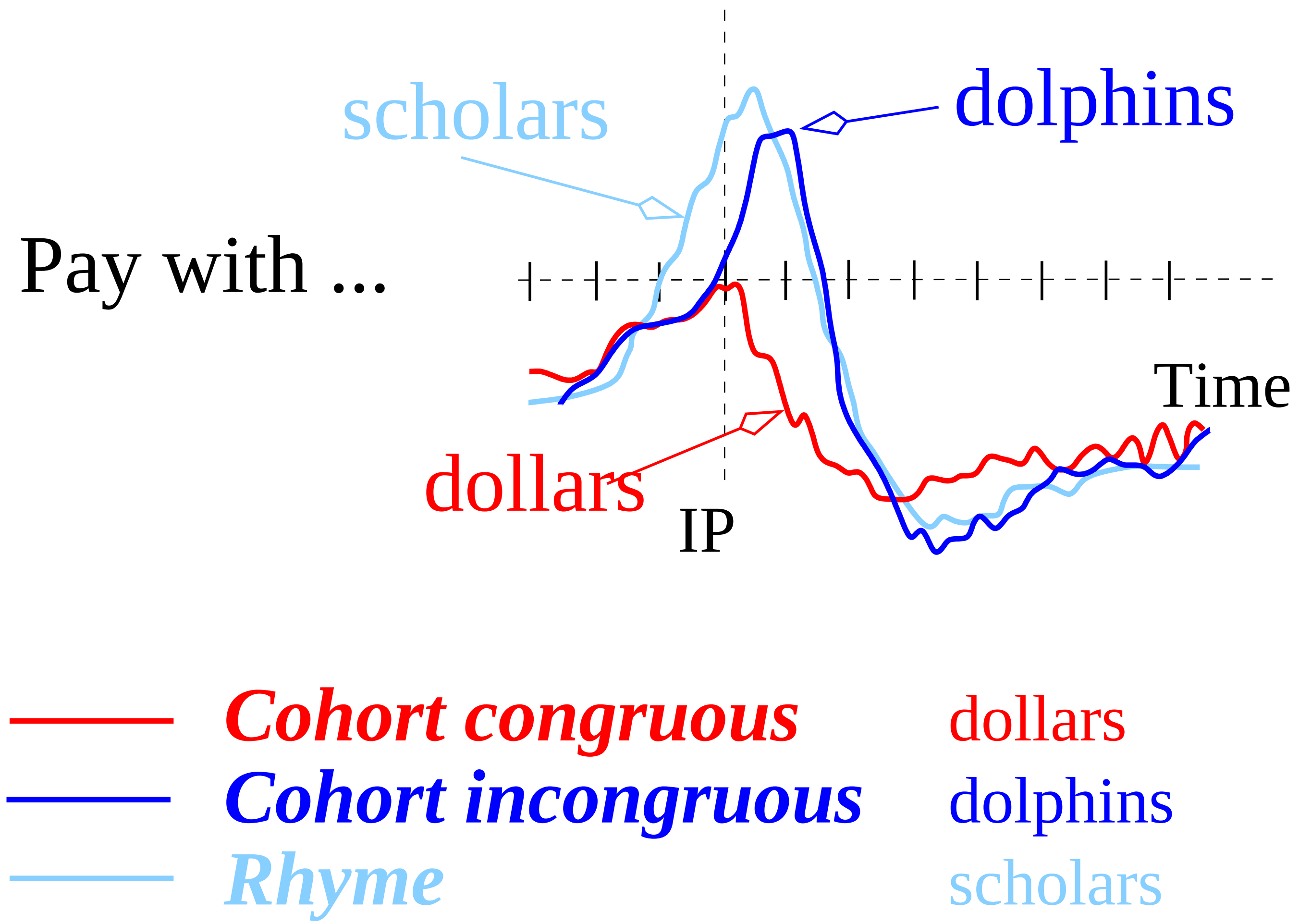
- Expt. II – Event related scalp potential (ERP) re IP, from Exp. I
  Sentence semantics effects



- dollars vs. dolphins: Word context, as measured by the IP, is independent of the sentence context!

ERP MEASURE OF CONTEXT RE IP

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



- Rhyme word scholars is recognized as being out of context before it is even recognized (at its IP)!

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]

CATEGORICAL PERCEPTION

- Meaningful words are recognized before they end
- Word context (i.e., the IP) seems independent of sentence context

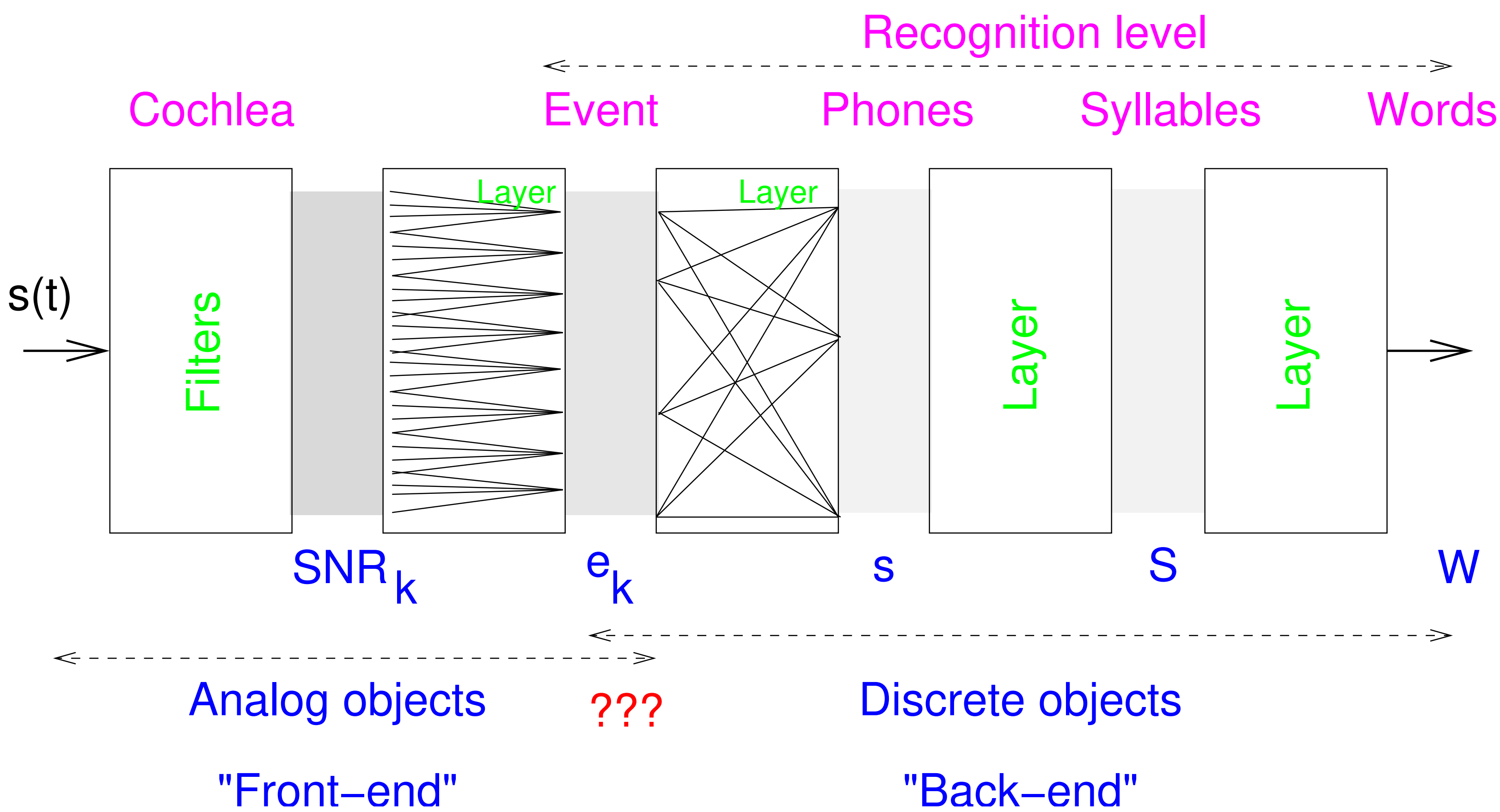
SUMMARY

- Miller & Nicely found 5 independent channels, described by discrete events [Miller's features]
- $\text{SNR}_k \Rightarrow$ events $\Rightarrow$ phones $\Rightarrow$ phonemes $\Rightarrow$ syllables $\Rightarrow$ words $\Rightarrow \ldots$
  - SNR determines discrete event errors
  - Discrete event errors label phone errors
  - Phone errors determine syllable errors
  - Syllable errors determine word errors
  - The HSR word error is established well before language is accessed!
  - HSR error depends only on the SNR in bands
- Language model performance is independent of noise robustness!
  - Cochlear filtering is important to robustness
  - Performance established at the event level
  - Strong parallels to visual processing
- ASR and HSR are fundamentally different
- To study HSR, entropy must be controlled
- Studies need to report raw phone/word errors
- Speech psychophysics is an important tool for studying HSR

FUTURE GOALS

- We need psychophysics to gain insight of how events are extracted
  - What are the physical parameters supporting each event?
- Any increase in insight will lead to invention of new signal processing methods for robust machine speech recognition

THE RECOGNITION CHAIN



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

jba@auditorymodels.org

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

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

Fletcher, H. and Galt, R. (1950). "Perception of speech and its relation to telephony," *J. Acoust. Soc. Am.* **22**:89–151.

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

Lippmann, R. (1997). "Speech perception by humans and machines," *Speech communication* **22**:1–15.

Miller, G. (1962). "Decision units in the perception of speech," *IRE Transactions on Information Theory* pages 81–83.

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

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

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

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

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