Title: How does the cochlea decode a CV speech sounds with zero error?

Two recent publications [Singh and Allen (2012); Toscano and Allen, (2014)] argue that consonant decoding is a binary process, with zero error above a token-dependent critical SNR threshold, typically below -2 dB SNR. From an information theoretic point of view, this is a "game changer," because it means that human consonant perception is operating below the Shannon channel capacity theoretical bound. We shall review these arguments, and based on our present understanding of cochlear signal processing, explain how this decoding strategy functions. The emphasis is on how the hearing impaired ear fails to perform this task. Speech cues are not "in the gaps," as is commonly assumed. An important question is the nature of the limits of the hearing impaired ear. Existing literature will be reviewed.

Refs:

–Riya Singh and Jont Allen (2012). *The influence of stop consonants perceptual features on the Articulation Index model,* J. Acoust. Soc. Am., v.131, 3051-3068

–Toscano, Joseph and Allen, Jont B (2014). *Across and within consonant errors for isolated syllables in noise,* Journal of Speech, Language, and Hearing Research, doi:10.1044/2014_JSLHR-H-13-0244

Cochlear nonlinearities and phoneme recognition

Jont Allen
UIUC & Beckman Inst, Urbana IL

June 14, 2015

Outline

- Intro + Objectives + Applications 3 mins
- Historical Overview 4 mins $\Sigma 7$
  - <1929 pre Telephone-age
  - 1930-1944 (Telephone-age + WWII)
  - 1945-1985 (Information-theoretic age)
  - >1985 (Computer-Renaissance)
- Methods 8 mins $\Sigma 15$
  - Theory (Information Theory; Signal processing)
  - Data collection (Psychophysics; Consonant confusions)
  - Analysis (Articulation Index; Confusion Patterns: $P_{h|s}(SNR)$)
- Results 21 mins $\Sigma 36$
  - Confusions; Primes and Morphs;
  - Examples of Speech Modifications; Conflicting cues
  - Binary nature of consonant recognition
  - How the AI works
- Cochlear speech processing 12 mins $\Sigma 48$
  - Neural coding of Consonants
- Summary + Conclusions 3 mins $\Sigma 51$

Outline

- Intro + Objectives + Applications 3 mins
- Historical Overview 4 mins Σ7
  - <1929 pre Telephone-age
  - 1930-1944 (Telephone-age + WWII)
  - 1945-1985 (Information-theoretic age)
  - >1985 (Computer-Renaissance)
- Methods 8 mins Σ15
  - Theory (Information Theory; Signal processing)
  - Data collection (Psychophysics; Consonant confusions)
  - Analysis (Articulation Index; Confusion Patterns: $P_{h|s}(SNR)$)
- Results 21 mins Σ36
  - Confusions; Primes and Morphs;
  - Examples of Speech Modifications; Conflicting cues
  - Binary nature of consonant recognition
  - How the AI works
- Cochlear speech processing 12 mins Σ48
  - Neural coding of Consonants
- Summary + Conclusions 3 mins Σ51

Outline

- Intro + Objectives + Applications 3 mins
- Historical Overview 4 mins $\Sigma 7$
  - <1929 pre Telephone-age
  - 1930-1944 (Telephone-age + WWII)
  - 1945-1985 (Information-theoretic age)
  - >1985 (Computer-Renaissance)
- Methods 8 mins $\Sigma 15$
  - Theory (Information Theory; Signal processing)
  - Data collection (Psychophysics; Consonant confusions)
  - Analysis (Articulation Index; Confusion Patterns: $P_{h|s}(SNR)$)
- Results 21 mins $\Sigma 36$
  - Confusions; Primes and Morphs;
  - Examples of Speech Modifications; Conflicting cues
  - Binary nature of consonant recognition
  - How the AI works
- Cochlear speech processing 12 mins $\Sigma 48$
  - Neural coding of Consonants
- Summary + Conclusions 3 mins $\Sigma 51$

Outline

- Intro + Objectives + Applications 3 mins
- Historical Overview 4 mins $\Sigma 7$
  - <1929 pre Telephone-age
  - 1930-1944 (Telephone-age + WWII)
  - 1945-1985 (Information-theoretic age)
  - >1985 (Computer-Renaissance)
- Methods 8 mins $\Sigma 15$
  - Theory (Information Theory; Signal processing)
  - Data collection (Psychophysics; Consonant confusions)
  - Analysis (Articulation Index; Confusion Patterns: $P_{h|s}(SNR)$)
- Results 21 mins $\Sigma 36$
  - Confusions; Primes and Morphs;
  - Examples of Speech Modifications; Conflicting cues
  - Binary nature of consonant recognition
  - How the AI works
- Cochlear speech processing 12 mins $\Sigma 48$
  - Neural coding of Consonants
- Summary + Conclusions 3 mins $\Sigma 51$

Outline

- Intro + Objectives + Applications 3 mins
- Historical Overview 4 mins Σ7
  - <1929 pre Telephone-age
  - 1930-1944 (Telephone-age + WWII)
  - 1945-1985 (Information-theoretic age)
  - >1985 (Computer-Renaissance)
- Methods 8 mins Σ15
  - Theory (Information Theory; Signal processing)
  - Data collection (Psychophysics; Consonant confusions)
  - Analysis (Articulation Index; Confusion Patterns: $P_{h|s}(SNR)$)
- Results 21 mins Σ36
  - Confusions; Primes and Morphs;
  - Examples of Speech Modifications; Conflicting cues
  - Binary nature of consonant recognition
  - How the AI works
- Cochlear speech processing 12 mins Σ48
  - Neural coding of Consonants
- Summary + Conclusions 3 mins Σ51

Outline

- Intro + Objectives + Applications 3 mins
- Historical Overview 4 mins $\Sigma 7$
  - <1929 pre Telephone-age
  - 1930-1944 (Telephone-age + WWII)
  - 1945-1985 (Information-theoretic age)
  - >1985 (Computer-Renaissance)
- Methods 8 mins $\Sigma 15$
  - Theory (Information Theory; Signal processing)
  - Data collection (Psychophysics; Consonant confusions)
  - Analysis (Articulation Index; Confusion Patterns: $P_{h|s}(SNR)$)
- Results 21 mins $\Sigma 36$
  - Confusions; Primes and Morphs;
  - Examples of Speech Modifications; Conflicting cues
  - Binary nature of consonant recognition
  - How the AI works
- Cochlear speech processing 12 mins $\Sigma 48$
  - Neural coding of Consonants
- Summary + Conclusions 3 mins $\Sigma 51$

I – Introduction (3 mins)

- Statement of the problem:
  - A fundamental understanding of the Human Speech code
  - Identify the cues in individual CV utterances
    - -Plosives (e.g., /p, t, k/ and /b, d, g/)
    - -Fricatives (e.g., /θ, ʃ, ʧ, s, h, f/ and /z, ʒ, v, ð/)
    - -With vowels /o, ɛ, ı/
- Applications:
  - Reduce variability in ASR at front-end
  - Hearing Aids, Cochlear Implants
  - Smart Telcom products
  - TTS (Text to speech)
  - Intelligibility modifications (Robustness problem)
    - Speech enhancement in noise

I – Introduction (3 mins)

- Statement of the problem:
  - A fundamental understanding of the Human Speech code
  - Identify the cues in individual CV utterances
    - -Plosives (e.g., /p, t, k/ and /b, d, g/)
    - -Fricatives (e.g., /θ, ʃ, ʧ, s, h, f/ and /z, ʒ, v, ð/)
    - -With vowels /o, ε, ı/
- Applications:
  - Reduce variability in ASR at front-end
  - Hearing Aids, Cochlear Implants
  - Smart Telcom products
  - TTS (Text to speech)
  - Intelligibility modifications (Robustness problem)
    - Speech enhancement in noise

Objective

- Rigorous procedures for analyzing and modifying speech in noise
- Objective: Identify perceptual features, i.e., speech cues

- Methods: Three metrics:
  - AI-Gram (speech audibility measure)
  - Confusion matrix $P_{h|s}$ (CV discrimination measure)
  - Confusion patterns ($P_{h|s}(SNR)$)
- Results: ONSETS, MODULATIONS and DURATION define the cues

Objective

- Rigorous procedures for analyzing and modifying speech in noise
- Objective: Identify perceptual features, i.e., speech cues

- Methods: Three metrics:
  - AI-Gram (speech audibility measure)
  - Confusion matrix $P_{h|s}$ (CV discrimination measure)
  - Confusion patterns ($P_{h|s}(SNR)$)
- Results: ONSETS, MODULATIONS and DURATION define the cues

Objective

- Rigorous procedures for analyzing and modifying speech in noise
- Objective: Identify perceptual features, i.e., speech cues

- Methods: Three metrics:
  - AI-Gram (speech audibility measure)
  - Confusion matrix $P_{h|s}$ (CV discrimination measure)
  - Confusion patterns ($P_{h|s}(SNR)$)
- Results: ONSETS, MODULATIONS and DURATION define the cues

Objective

- Rigorous procedures for analyzing and modifying speech in noise
- Objective: Identify perceptual features, i.e., speech cues

- Methods: Three metrics:
  - AI-Gram (speech audibility measure)
  - Confusion matrix $P_{h|s}$ (CV discrimination measure)
  - Confusion patterns ($P_{h|s}(SNR)$)
- Results: ONSETS, MODULATIONS and DURATION define the cues

II – Historical HSR Studies (4 mins)

- Lord Rayleigh 1908 and George Campbell 1910
  - First electronic articulation experiments
- Harvey Fletcher's 1921 Articulation Index AI
  - Ψ: Massive data collection, for 30 years
  - Φ: Accurate AI predictions of Average Syllable Scores
    - French and Steinberg 1947 first publish AI
- Shannon The Theory of Information (TI) 1948+
  - Miller's work based on Shannon's TI
  - G.A. Miller, Heise and Lichten Entropy $\mathcal{H}$ 1951
  - G.A. Miller & Nicely CM $P_{h|s}(SNR)$ 1955
- Context studies:
  - Boothroyd JASA 1968; Boothroyd & Nittrouer 1988
  - Bronkhorst et al. JASA 1993

II – Historical HSR Studies (4 mins)

- Lord Rayleigh 1908 and George Campbell 1910
  - First electronic articulation experiments
- Harvey Fletcher's 1921 Articulation Index AI
  - Ψ: Massive data collection, for 30 years
  - Φ: Accurate AI predictions of Average Syllable Scores
    - French and Steinberg 1947 first publish AI
- Shannon The Theory of Information (TI) 1948+
  - Miller's work based on Shannon's TI
  - G.A. Miller, Heise and Lichten Entropy $\mathcal{H}$ 1951
  - G.A. Miller & Nicely CM $P_{h|s}(SNR)$ 1955
- Context studies:
  - Boothroyd JASA 1968; Boothroyd & Nittrouer 1988
  - Bronkhorst et al. JASA 1993

II – Historical HSR Studies (4 mins)

- Lord Rayleigh 1908 and George Campbell 1910
  - First electronic articulation experiments
- Harvey Fletcher's 1921 Articulation Index AI
  - Ψ: Massive data collection, for 30 years
  - Φ: Accurate AI predictions of Average Syllable Scores
    - French and Steinberg 1947 first publish AI
- Shannon The Theory of Information (TI) 1948+
  - Miller's work based on Shannon's TI
  - G.A. Miller, Heise and Lichten Entropy $\mathcal{H}$ 1951
  - G.A. Miller & Nicely CM $P_{h|s}(SNR)$ 1955
- Context studies:
  - Boothroyd JASA 1968; Boothroyd & Nittrouer 1988
  - Bronkhorst et al. JASA 1993

II – Historical HSR Studies (4 mins)

- Lord Rayleigh 1908 and George Campbell 1910
  - First electronic articulation experiments
- Harvey Fletcher's 1921 Articulation Index AI
  - Ψ: Massive data collection, for 30 years
  - Φ: Accurate AI predictions of Average Syllable Scores
    - French and Steinberg 1947 first publish AI
- Shannon The Theory of Information (TI) 1948+
  - Miller's work based on Shannon's TI
  - G.A. Miller, Heise and Lichten Entropy $\mathcal{H}$ 1951
  - G.A. Miller & Nicely CM $P_{h|s}(SNR)$ 1955
- Context studies:
  - Boothroyd JASA 1968; Boothroyd & Nittrouer 1988
  - Bronkhorst et al. JASA 1993

Speech research

- 1910-1960: Bell Labs (Galt, Fletcher, Kelly)
- 1940-1960: Haskins Lab Synthetic speech (Cooper, Liberman)
- 1960-1990: MIT Consonant features unknown (Ken Stevens et al.)
- 1980-2010: ASR at AT&T, IBM, BBN, University research
  Not designed to be robustness to noise
- 2003-2015: UIUC (Allen)

Cochlear research

- 1910-1950: Bell Labs (Wegel+Lane, Fletcher, Munson, Steinberg)
- 1960-2015: MIT+Harvard HSBT
- 1970-2015: NIH funded University research
- 1970-2003 Bell Labs (Allen)

Speech research

- 1910-1960: Bell Labs (Galt, Fletcher, Kelly)
- 1940-1960: Haskins Lab Synthetic speech (Cooper, Liberman)
- 1960-1990: MIT Consonant features unknown (Ken Stevens et al.)
- 1980-2010: ASR at AT&T, IBM, BBN, University research
  Not designed to be robustness to noise
- 2003-2015: UIUC (Allen)

Cochlear research

- 1910-1950: Bell Labs (Wegel+Lane, Fletcher, Munson, Steinberg)
- 1960-2015: MIT+Harvard HSBT
- 1970-2015: NIH funded University research
- 1970-2003 Bell Labs (Allen)

Allen et. al HSR Experiments 2004-2011

Year	Experiment	Student &Allen	Details	Publications
2004	MN04(MN64)	Phatak, Lovitt	MNR	JASA
2005	MN16R	Phatak, Lovitt	MN55R	JASA
2005	HIMCL05	Yoon, Phatak	10 HI ears	JASA
2006	HINALR05	Yoon *et al.*	10 HI ears	JSLR (2011)
2006	Verification	Regnier	/ta/	JASA
2006	CV06-s/w	Phatak/Regnier	8C+9V SWN/WN	
2007	CV06	Pan	CV06	MS Thesis
2007	HL07	Li	Hi/Lo pass	JASA
2008	TR08	Li	Furui86	ASSP
2009	3DDS	Li	plosives	JASA: TLSP
2009	Verification	Cvengros	burst mods	Thesis
2009	Verification	Abhinauv	burst mods	JASA
2009	mn64 NZE	Singh	PA07	JASA
2010	HIMCL10-I,II,III	Trevino, Han	46 HI ears @MCL	JASA/Sem Hear.
2011	3DDS	Li	Fricatives	JASA
2011	HINAL11-IV	Han	17 HI ears w NALR	PhD Thesis (Ch. 3)
2014	CV06	Toscano	30 NH ears	JSLHR

Recent Speech Studies 2000-2013

- Three Recent Literature Reviews:
  - Wright 2004 "A review of perceptual cues and cue robustness"
  - Allen 2005 *"Articulation & Intelligibility"* Morgan-Claypool
  - McMurray-Jongman 2011 "speech categorization"
- Ten Detailed Studies:
  - Jongman 2000 "Acoustic characteristics of fricatives"
  - Smits 2000 "Temporal distribution . . . in VCVs"
  - Hazan-Simpson 2000 "cue-enhancement . . . of nonsense words"
  - Jiang 2006 "perception of voicing in plosives"
  - McMurray-Jongman 2011 "information for speech categorization"
  - Alwan 2011 "Perception of place of articulation . . . "
  - Jørgensen-Dau 2011; 3 dB change; Modulation references
  - Das-Hansen 2012 "Speech Enhancement $\bar{c}$ Phone Classes"
  - Consonant perception is binary with variable thresholds
    - Singh-Allen 2012
    - Toscano-Allen 2013

Recent Speech Studies 2000-2013

- Three Recent Literature Reviews:
  - Wright 2004 "A review of perceptual cues and cue robustness"
  - Allen 2005 *"Articulation & Intelligibility"* Morgan-Claypool
  - McMurray-Jongman 2011 "speech categorization"
- Ten Detailed Studies:
  - Jongman 2000 "Acoustic characteristics of fricatives"
  - Smits 2000 "Temporal distribution . . . in VCVs"
  - Hazan-Simpson 2000 "cue-enhancement . . . of nonsense words"
  - Jiang 2006 "perception of voicing in plosives"
  - McMurray-Jongman 2011 "information for speech categorization"
  - Alwan 2011 "Perception of place of articulation . . . "
  - Jørgensen-Dau 2011; 3 dB change; Modulation references
  - Das-Hansen 2012 "Speech Enhancement $\bar{c}$ Phone Classes"
  - Consonant perception is binary with variable thresholds
    - Singh-Allen 2012
    - Toscano-Allen 2013

Recent Speech Studies 2000-2013

- Three Recent Literature Reviews:
  - Wright 2004 "A review of perceptual cues and cue robustness"
  - Allen 2005 *"Articulation & Intelligibility"* Morgan-Claypool
  - McMurray-Jongman 2011 "speech categorization"
- Ten Detailed Studies:
  - Jongman 2000 "Acoustic characteristics of fricatives"
  - Smits 2000 "Temporal distribution ... in VCVs"
  - Hazan-Simpson 2000 "cue-enhancement ... of nonsense words"
  - Jiang 2006 "perception of voicing in plosives"
  - McMurray-Jongman 2011 "information for speech categorization"
  - Alwan 2011 "Perception of place of articulation ..."
  - Jørgensen-Dau 2011; 3 dB change; Modulation references
  - Das-Hansen 2012 "Speech Enhancement c̄ Phone Classes"
  - Consonant perception is binary with variable thresholds
    - Singh-Allen 2012
    - Toscano-Allen 2013

III – Methods 8 mins

- Psychophysics:
  - Consonant-vowel CV speech recognition $P_{h|s}(SNR)$
  - Several types of additive noise
  - Large number of trials
    - $>$20 talkers and $>$20 listeners
- Modeling:
  - Information Theory IT $\equiv$ Articulation index AI
  - Confusion matrix CM scores: $P_{h|s}(SNR)$
  - AI to model mean phone errors $P_c(SNR|s) = \sum_h P_{h|s}(SNR)$
- Signal processing:
  - AI-gram (crude cochlear model)
  - Frequency, time, intensity truncation 3^d-DS
  - Short-Time Fourier Transform STFT modifications

III – Methods 8 mins

- Psychophysics:
  - Consonant-vowel CV speech recognition $P_{h|s}(SNR)$
  - Several types of additive noise
  - Large number of trials
    - $>$20 talkers and $>$20 listeners
- Modeling:
  - Information Theory IT $\equiv$ Articulation index AI
  - Confusion matrix CM scores: $P_{h|s}(SNR)$
  - AI to model mean phone errors $P_c(SNR|s) = \sum_h P_{h|s}(SNR)$
- Signal processing:
  - AI-gram (crude cochlear model)
  - Frequency, time, intensity truncation 3^d-DS
  - Short-Time Fourier Transform STFT modifications

III – Methods 8 mins

- Psychophysics:
  - Consonant-vowel CV speech recognition $P_{h|s}(SNR)$
  - Several types of additive noise
  - Large number of trials
    - $>$20 talkers and $>$20 listeners
- Modeling:
  - Information Theory IT $\equiv$ Articulation index AI
  - Confusion matrix CM scores: $P_{h|s}(SNR)$
  - AI to model mean phone errors $P_c(SNR|s) = \sum_h P_{h|s}(SNR)$
- Signal processing:
  - AI-gram (crude cochlear model)
  - Frequency, time, intensity truncation 3^d-DS
  - Short-Time Fourier Transform STFT modifications

The CM $P_{h|s}(SNR)$

- Miller-Nicely's 1955 articulation matrix $P_{h|s}(SNR)$, 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 ≡ *inhomogeneous cues*

The CM $P_{h|s}(SNR)$

- Miller-Nicely's 1955 articulation matrix $P_{h|s}(SNR)$, 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

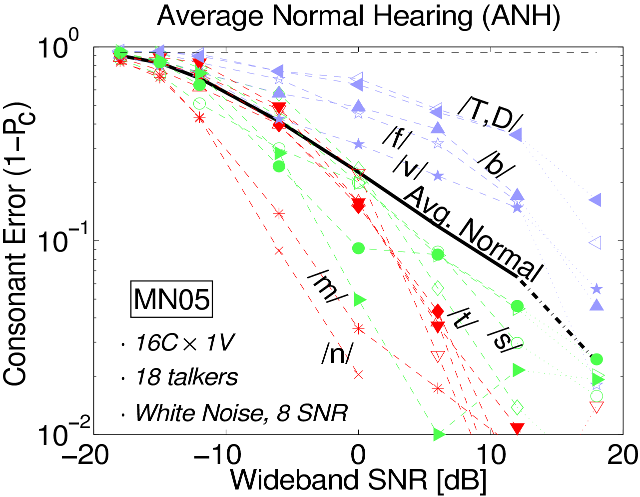
UNVOICED — VOICED — NASAL

RESPONSE

- Confusion groups $\equiv$ *inhomogeneous cues*

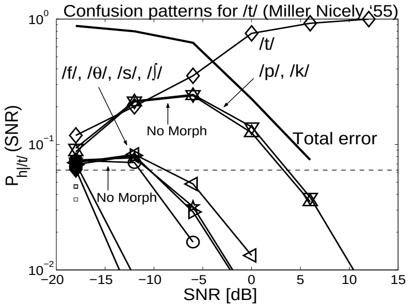
Average phone scores vs. SNR: $P_{h|s}(SNR)$

- Consonant chance performance is -20 dB-SNR in white noise Phatak Allen, 2007

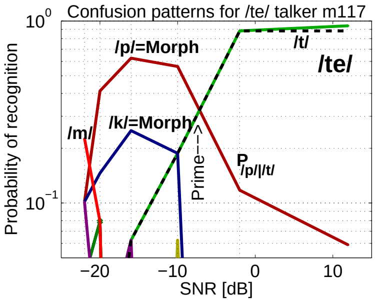


Row of CM $P_{h|/t/}$

- Utterance phone scores are heterogeneous!



(a) Average over all /t/s.

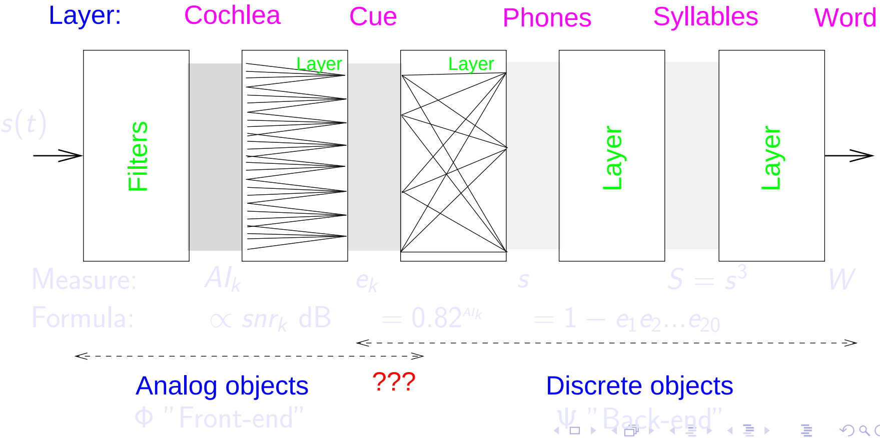


(b) Talker m117 /te/ $P_{h|/ta/}(SNR)$

- Phone groups are due to shared sub-phonemic units
  - CV Morphs

AI Model of *human speech recognition* HSR

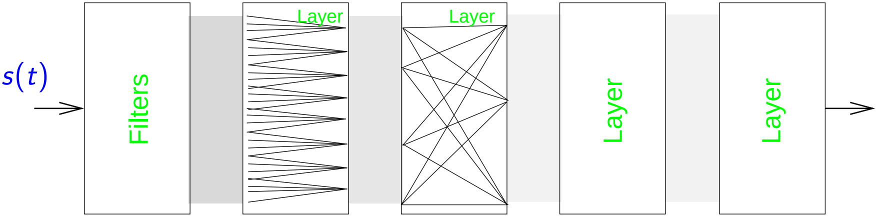
- Research Goal:
  - Identify *elemental speech cues*
  - A cue is defined as a *perceptual feature*
  - Cue errors are measured by band errors e_k



AI Model of *human speech recognition* HSR

- Research Goal:
  - Identify *elemental speech cues*
  - A cue is defined as a *perceptual feature*
  - Cue errors are measured by band errors e_k

Layer:	Cochlea	Cue	Phones	Syllables	Word



Measure:	AI_k	e_k	s	$S = s^3$	W
Formula:	$\propto snr_k$ dB	$= 0.82^{AI_k}$	$= 1 - e_1 e_2 \ldots e_{20}$		

<------------------------------>   <--->

Analog objects
Φ "Front-end"

???

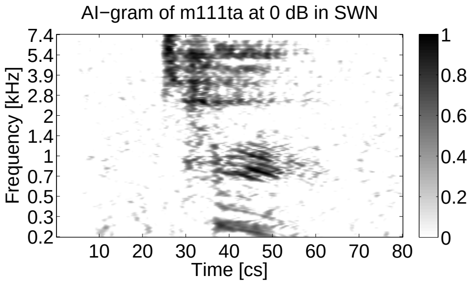
Discrete objects
Ψ "Back-end"

Model of human listeners as a Shannon Channel

- Channel capacity theorem specifies the maximum information rate

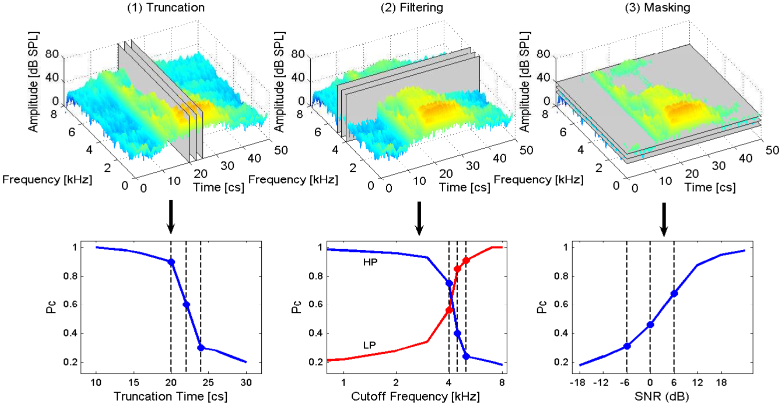
$$\mathcal{C} \equiv \int \log_2 \left(1 + SNR^2(f)\right) df \qquad (1)$$

- For a Maximum Entropy (MaxEnt) speech source, the maximum information rate is determined by the SNR
- The AI-gram is a related measure:



Methods: 3^d Deep Search (3DDS)

- 3^d Deep-Search via truncation:
  - SNR truncation (i.e., masking)
  - Frequency truncation (High/Low-pass filtering)
  - Time truncation (Furui 1986)



III–Results 21 mins

- Discussion of AI
  - Across consonant error
  - Within consonant error
- Examples and Demos of events
  - Plosive CV events
  - Fricative CV events
- Conflicting cues
- DEMOS:
  - Event isolation
  - Consonant morphing
  - Consonant enhancement
  - Conflicting cues within consonants
  - Sentence meaning modification

III–Results 21 mins

- Discussion of AI
  - Across consonant error
  - Within consonant error
- Examples and Demos of events
  - Plosive CV events
  - Fricative CV events
- Conflicting cues
- DEMOS:
  - Event isolation
  - Consonant morphing
  - Consonant enhancement
  - Conflicting cues within consonants
  - Sentence meaning modification

III–Results 21 mins

- Discussion of AI
  - Across consonant error
  - Within consonant error
- Examples and Demos of events
  - Plosive CV events
  - Fricative CV events
- Conflicting cues
- DEMOS:
  - Event isolation
  - Consonant morphing
  - Consonant enhancement
  - Conflicting cues within consonants
  - Sentence meaning modification

III–Results 21 mins

- Discussion of AI
  - Across consonant error
  - Within consonant error
- Examples and Demos of events
  - Plosive CV events
  - Fricative CV events
- Conflicting cues
- DEMOS:
  - Event isolation
  - Consonant morphing
  - Consonant enhancement
  - Conflicting cues within consonants
  - Sentence meaning modification

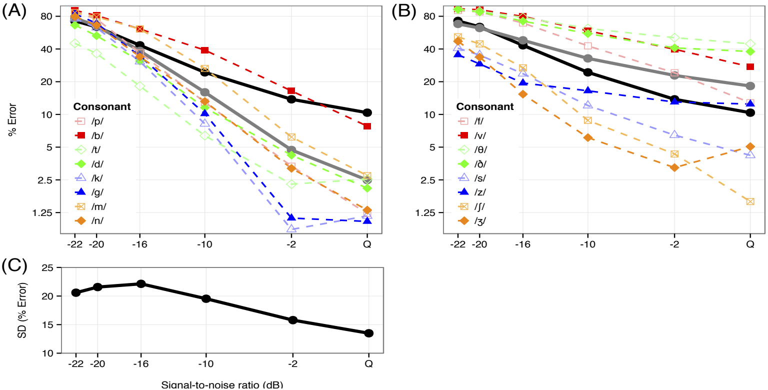
III–Results 21 mins

- Discussion of AI
  - Across consonant error
  - Within consonant error
- Examples and Demos of events
  - Plosive CV events
  - Fricative CV events
- Conflicting cues
- DEMOS:
  - Event isolation
  - Consonant morphing
  - Consonant enhancement
  - Conflicting cues within consonants
  - Sentence meaning modification

III–Results 21 mins

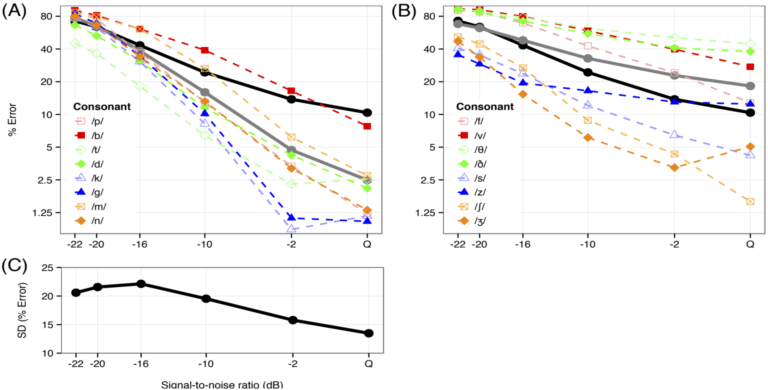
- Discussion of AI
  - Across consonant error
  - Within consonant error
- Examples and Demos of events
  - Plosive CV events
  - Fricative CV events
- Conflicting cues
- DEMOS:
  - Event isolation
  - Consonant morphing
  - Consonant enhancement
  - Conflicting cues within consonants
  - Sentence meaning modification

Results 1: The Across-consonant variance is Huge



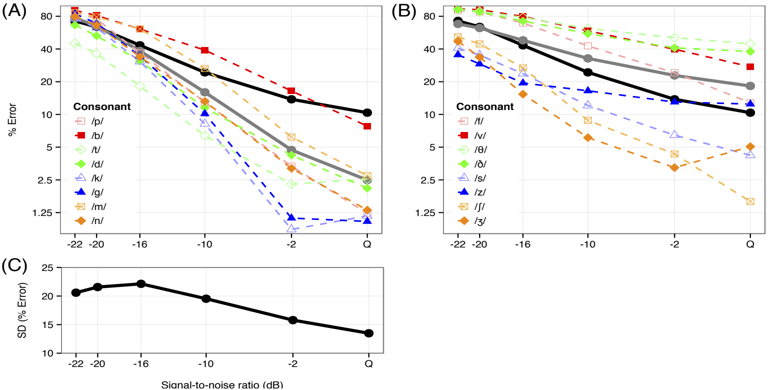
- AI(SNR) characterizes the average consonant error ($P_e = e_{chance} e_{\min}^{AI}$)
- $AI \approx SNR$ assuming SWN
- Log-error is linear in AI: $\log P_e = \log e_{chance} + AI \cdot \log e_{\min} = \beta_0 + \beta_1 AI$
- Note the huge *across-consonant* Standard Deviation

Results 1: The Across-consonant variance is Huge



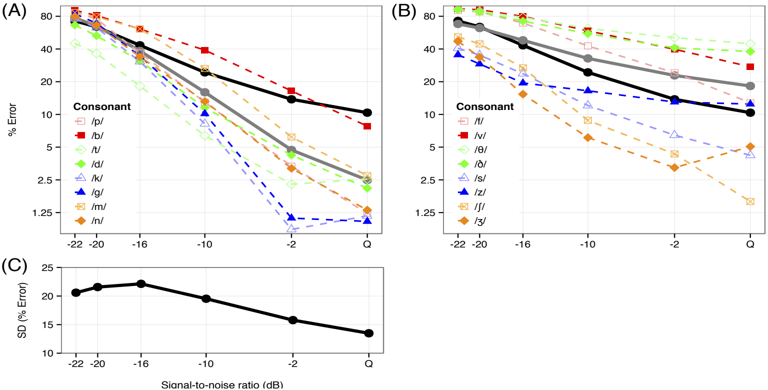
- AI(SNR) characterizes the average consonant error ($P_e = e_{chance} e_{\min}^{AI}$)
- $AI \approx SNR$ assuming SWN
- Log-error is linear in AI: $\log P_e = \log e_{chance} + AI \cdot \log e_{\min} = \beta_0 + \beta_1 AI$
- Note the huge *across-consonant* Standard Deviation

Results 1: The Across-consonant variance is Huge



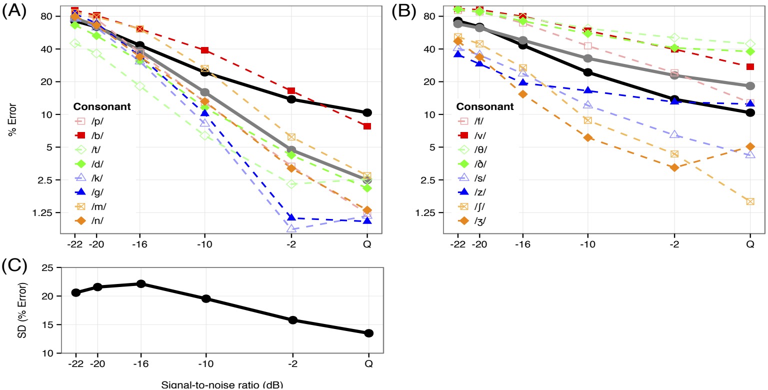
- AI(SNR) characterizes the average consonant error ($P_e = e_{chance} e_{\min}^{AI}$)
- $AI \approx SNR$ assuming SWN
- Log-error is linear in AI: $\log P_e = \log e_{chance} + AI \cdot \log e_{\min} = \beta_0 + \beta_1 AI$
- Note the huge *across-consonant* Standard Deviation

Results 1: The Across-consonant variance is Huge



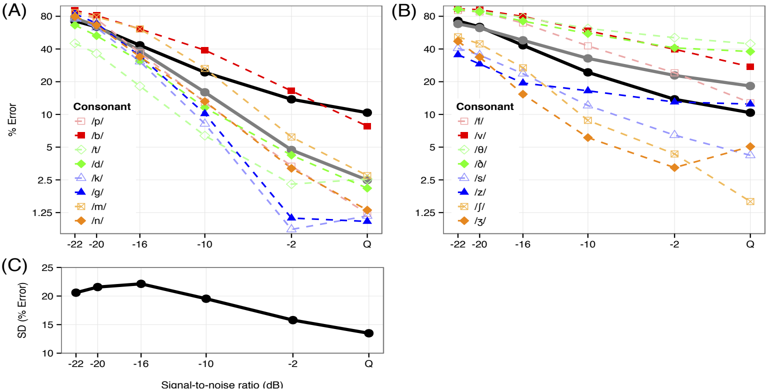
- AI(SNR) characterizes the average consonant error ($P_e = e_{chance} e_{\min}^{AI}$)
- $AI \approx SNR$ assuming SWN
- Log-error is linear in AI: $\log P_e = \log e_{chance} + AI \cdot \log e_{\min} = \beta_0 + \beta_1 AI$
- Note the huge *across-consonant* Standard Deviation

Results 1: The Across-consonant variance is Huge



- AI(SNR) characterizes the average consonant error ($P_e = e_{chance} e_{\min}^{AI}$)
- $AI \approx SNR$ assuming SWN
- Log-error is linear in AI: $\log P_e = \log e_{chance} + AI \cdot \log e_{\min} = \beta_0 + \beta_1 AI$
- Note the huge *across-consonant* Standard Deviation

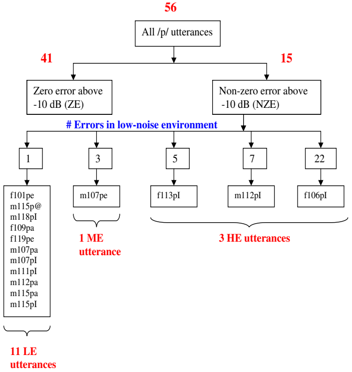
Results 1: The Across-consonant variance is Huge

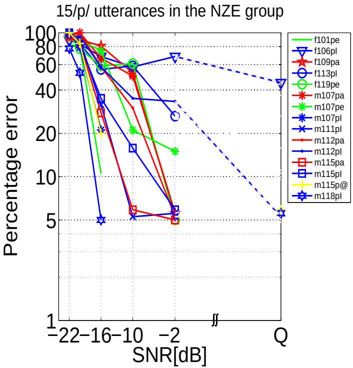


- AI(SNR) characterizes the average consonant error ($P_e = e_{chance} e_{\min}^{AI}$)
- $AI \approx SNR$ assuming SWN
- Log-error is linear in AI: $\log P_e = \log e_{chance} + AI \cdot \log e_{\min} = \beta_0 + \beta_1 AI$
- Note the huge *across-consonant* Standard Deviation

Within-consonant Error /p/ Singh-Allen 2012

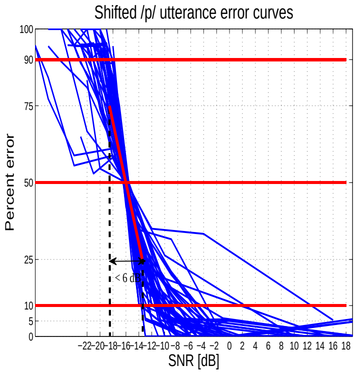
- 56 /p/+/o,e,ɪ/ CV tokens: SNR > -10 dB SNR
- Bimodal error distribution:
  - 41/56: Zero error (ZE); $N_{trials} = 38$, $N_{subj}=25$
  - 15/56: Non-zero error (NZE); 11 ≈ ZE (error: 1/38)



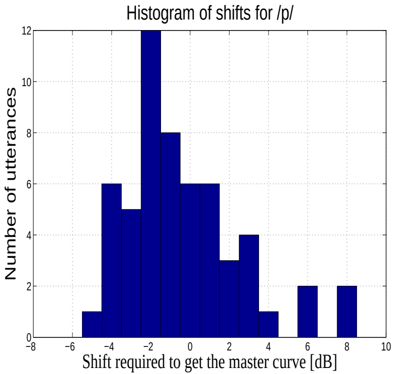


Within-consonant error $P_e(SNR - SNR^*_{50})$ for /p/

- Error vs. SNR shifted to 50% threshold SNR^*_{50} (LEFT)
- Histogram of 50% error thresholds (RIGHT)

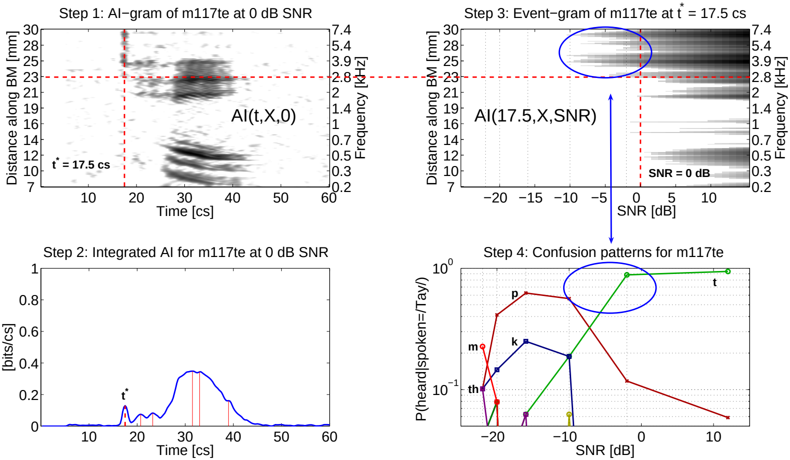


(a) $P_e(SNR - SNR^*_{50})$



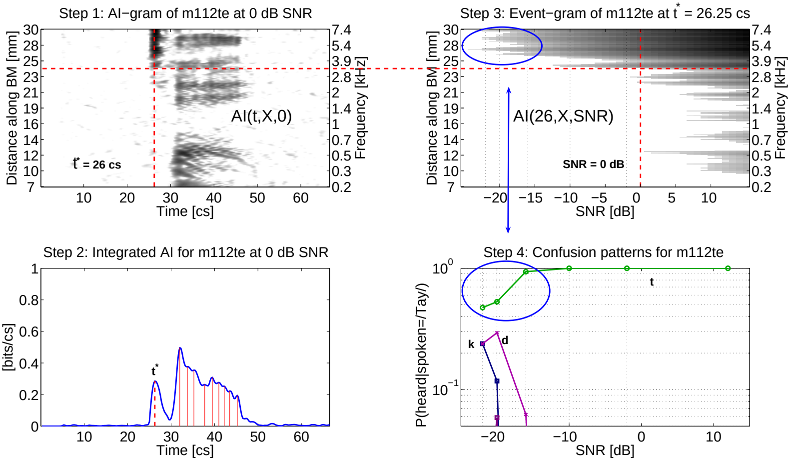
(b) Distribution of SNR^*_{50}

3DDS: m117/tε/ $SNR_{50} = -2$ [dB] (SWN)



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

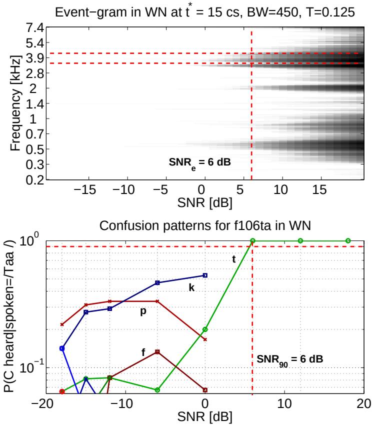
3DDS: m112/tε/ $SNR_{50} = -16$ [dB] (SWN)

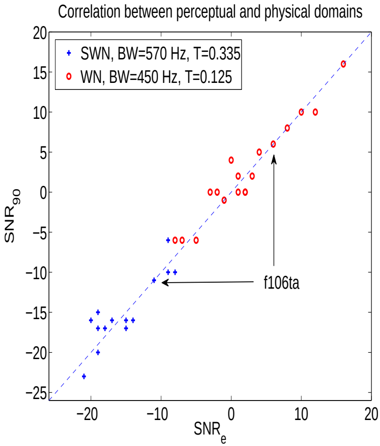


- /t/ confusion threshold at $P_c(SNR^* = -16) = 0.9$ correlated to Event-gram

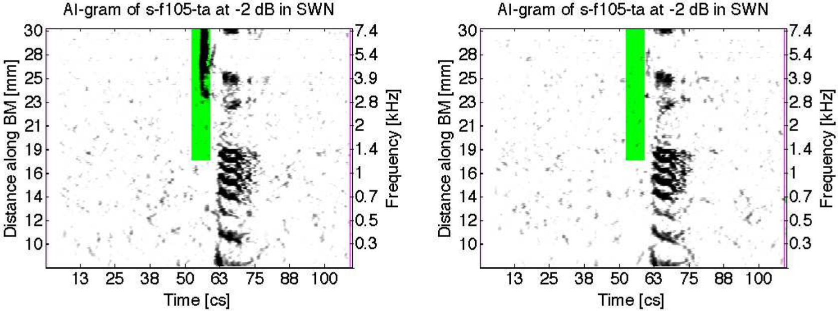
Correlations of all the /t/ events Regnier-Allen (2008)

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



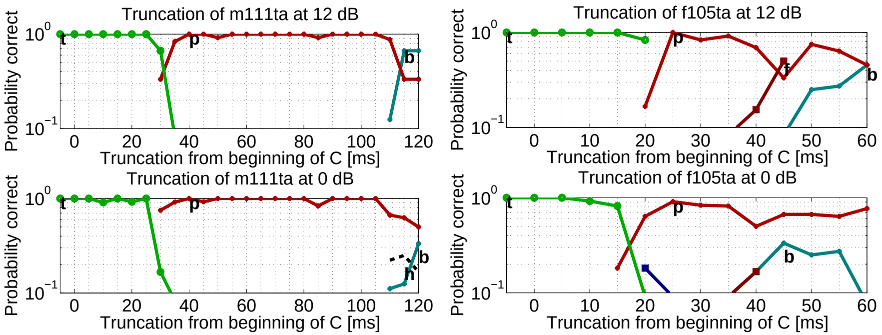


Masking of /tɑ/ timing cue



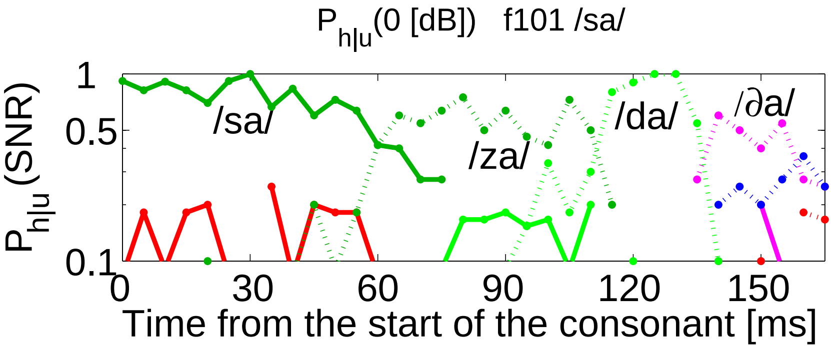
- When the /t/ burst is masked by noise, the perception morphs to /p/

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 Furui 1986, and results of Allen et. al

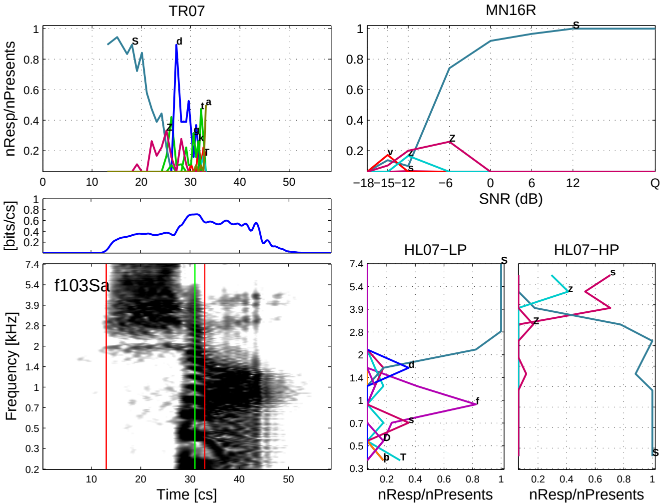
Truncation of f101 /sa/ (fricatives)



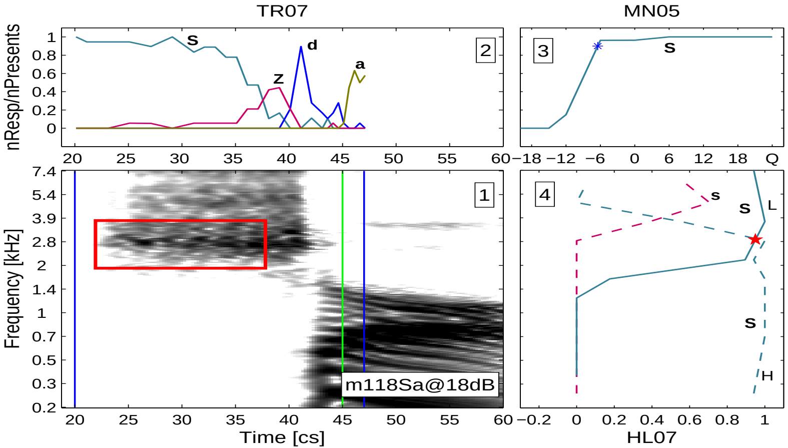
- This represents the normal hearing responses to a truncated /sɑ/, from the start of the consonant
- Morphing from /sɑ/ to /zɑ/ to /dɑ/ to /ðɑ/
- Duration is an important fricatives cue ▸ Sa to Da

3DDS Method /ʃa/

- Truncation in Time, Intensity and Frequency

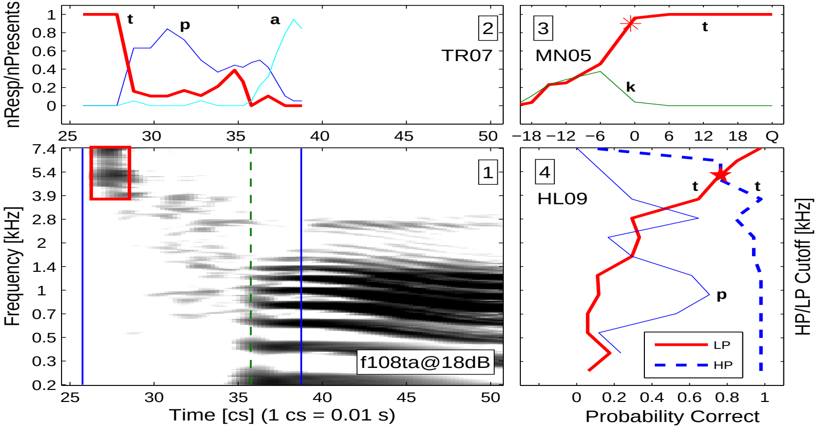


3DDS Method /ʃa/



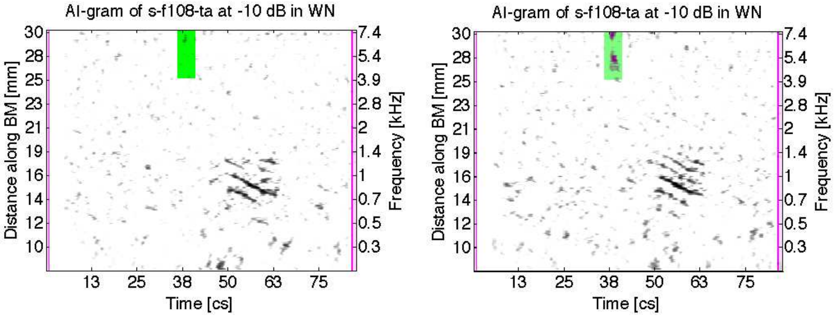
- Truncation in Intensity, time and frequency

3DDS Method /ta/



- Truncation in Intensity, time and frequency

Enhancement of ta event

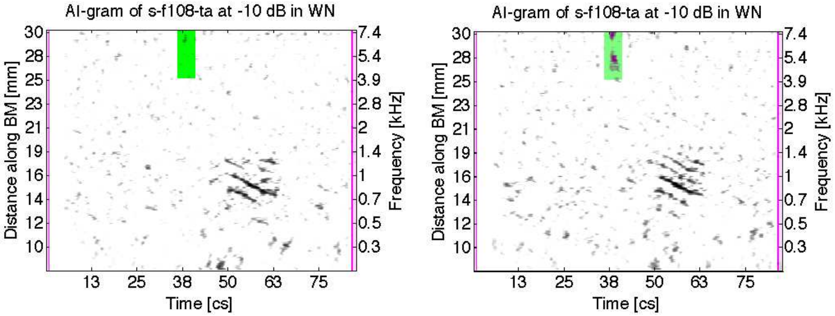


(c) Original /tɑ/

(d) Modified /tɑ/

- METHODS: The /t/ burst is enhanced (14 dB) on the quiet sound, then noise is added
- DEMO /ta/2/ka/

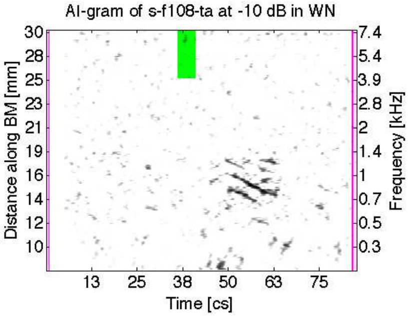
Enhancement of ta event



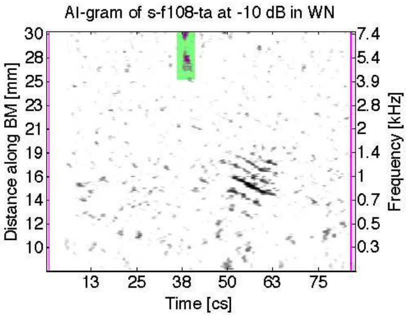
(e) Original /tɑ/ (f) Modified /tɑ/

- METHODS: The /t/ burst is enhanced (14 dB) on the quiet sound, then noise is added
- DEMO ▸ /ta/2/ka/

Enhancement of ta event



(g) Original /tɑ/



(h) Modified /tɑ/

- METHODS: The /t/ burst is enhanced (14 dB) on the quiet sound, then noise is added
- DEMO /ta/2/ka/

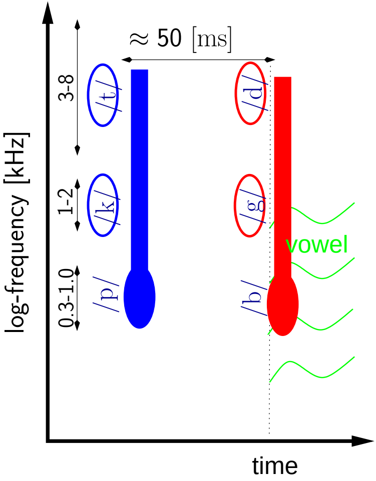
Demos by Andrea Trevino (2013)

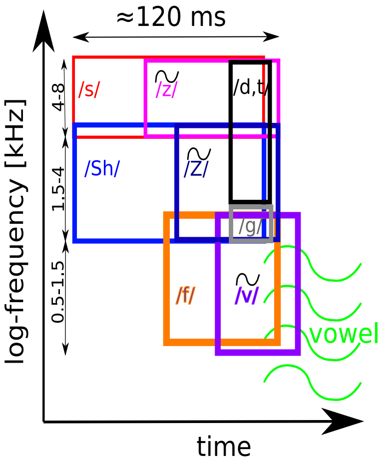
- Demo 1: /tɑ/ remove burst
- Demo 2: ka2ta f103, da2ga f103
- Demo 3: Sa2sa m118
- Demo 4: Sa2da m111
- Demo 5: za voicebar removed, ʃɑ vs ʒɑ same duration

Summary of Consonant structure

- Time-frequency structure of plosives and fricatives

  plosives: /p, t, k, b, d, g/+/a/





Auditory & Cochlear Modeling 1920-2015 12 min

- 1910-1980: Bell Labs (long history)
  - Fletcher 1914; Wegel & Lane 1924; Flanagan; Hall; Allen
- 1960-2010: MIT + Harvard HSBT
  - Eaton Peabody (Kiang, Siebert, Liberman, Guinan, Shera, . . .)
- Netherlands, England
  - deBoer, Duifhuis, Evans, . . .
- Australia (B. Johnstone, . . .)
- 1980-2011: NIH funded University research
  - MIT; Wash U; Boys Town; U. Wisc.; U. Mich.; Nortwestern U.
- The role of cochlear modeling on speech perception is huge!
  - And under appreciated, IMO

Auditory & Cochlear Modeling 1920-2015 12 min

- 1910-1980: Bell Labs (long history)
  - Fletcher 1914; Wegel & Lane 1924; Flanagan; Hall; Allen
- 1960-2010: MIT + Harvard HSBT
  - Eaton Peabody (Kiang, Siebert, Liberman, Guinan, Shera, ...)
- Netherlands, England
  - deBoer, Duifhuis, Evans, ...
- Australia (B. Johnstone, ...)
- 1980-2011: NIH funded University research
  - MIT; Wash U; Boys Town; U. Wisc.; U. Mich.; Nortwestern U.
- The role of cochlear modeling on speech perception is huge!
  - And under appreciated, IMO

Auditory & Cochlear Modeling 1920-2015 12 min

- 1910-1980: Bell Labs (long history)
  - Fletcher 1914; Wegel & Lane 1924; Flanagan; Hall; Allen
- 1960-2010: MIT + Harvard HSBT
  - Eaton Peabody (Kiang, Siebert, Liberman, Guinan, Shera, ...)
- Netherlands, England
  - deBoer, Duifhuis, Evans, ...
- Australia (B. Johnstone, ...)
- 1980-2011: NIH funded University research
  - MIT; Wash U; Boys Town; U. Wisc.; U. Mich.; Nortwestern U.
- The role of cochlear modeling on speech perception is huge!
  - And under appreciated, IMO

Auditory & Cochlear Modeling 1920-2015 12 min

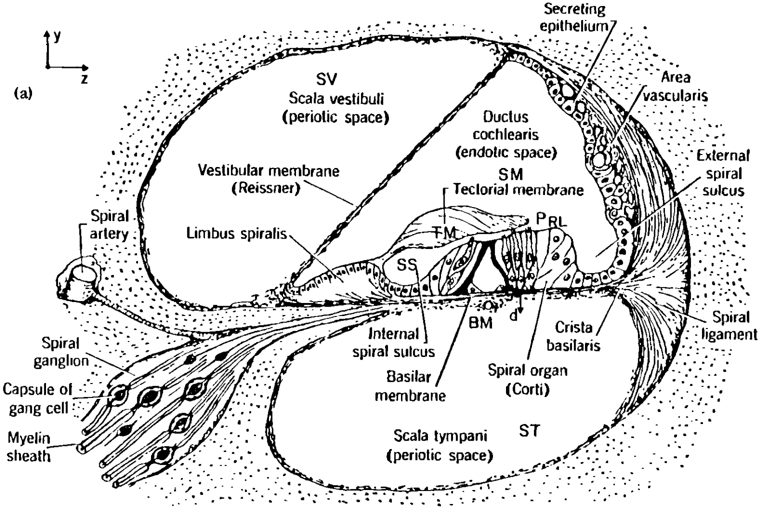
- 1910-1980: Bell Labs (long history)
  - Fletcher 1914; Wegel & Lane 1924; Flanagan; Hall; Allen
- 1960-2010: MIT + Harvard HSBT
  - Eaton Peabody (Kiang, Siebert, Liberman, Guinan, Shera, . . .)
- Netherlands, England
  - deBoer, Duifhuis, Evans, . . .
- Australia (B. Johnstone, . . .)
- 1980-2011: NIH funded University research
  - MIT; Wash U; Boys Town; U. Wisc.; U. Mich.; Nortwestern U.
- The role of cochlear modeling on speech perception is huge!
  - And under appreciated, IMO

Auditory & Cochlear Modeling 1920-2015 12 min

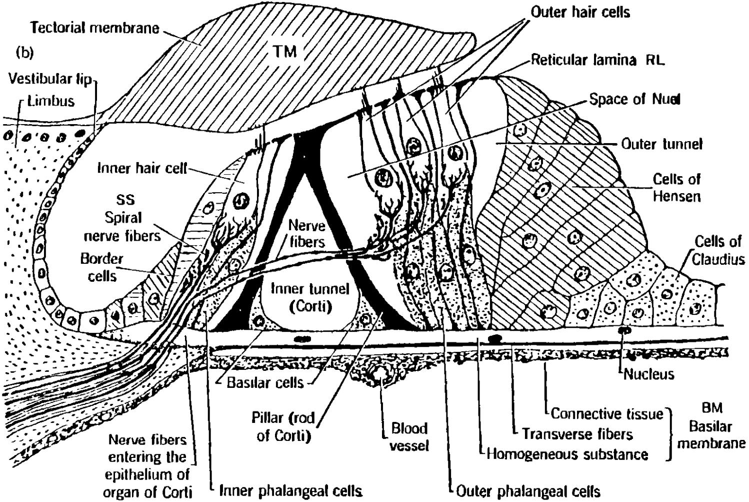
- 1910-1980: Bell Labs (long history)
  - Fletcher 1914; Wegel & Lane 1924; Flanagan; Hall; Allen
- 1960-2010: MIT + Harvard HSBT
  - Eaton Peabody (Kiang, Siebert, Liberman, Guinan, Shera, . . .)
- Netherlands, England
  - deBoer, Duifhuis, Evans, . . .
- Australia (B. Johnstone, . . .)
- 1980-2011: NIH funded University research
  - MIT; Wash U; Boys Town; U. Wisc.; U. Mich.; Nortwestern U.
- The role of cochlear modeling on speech perception is huge!
  - And under appreciated, IMO

The Mammalian Cochlea

The Human Cochlea

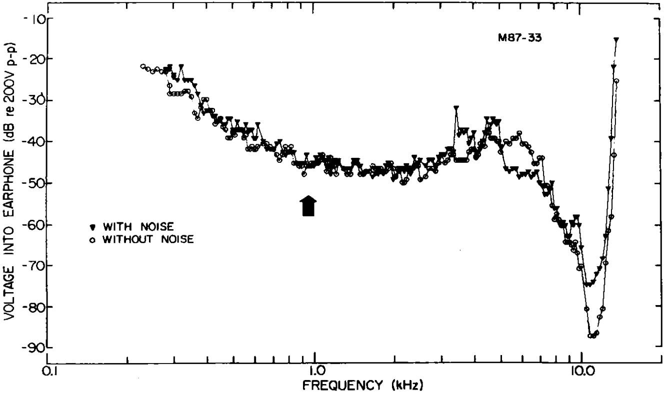


The Cochlear duct



Kiang and Moxon 1979 cochlear USM

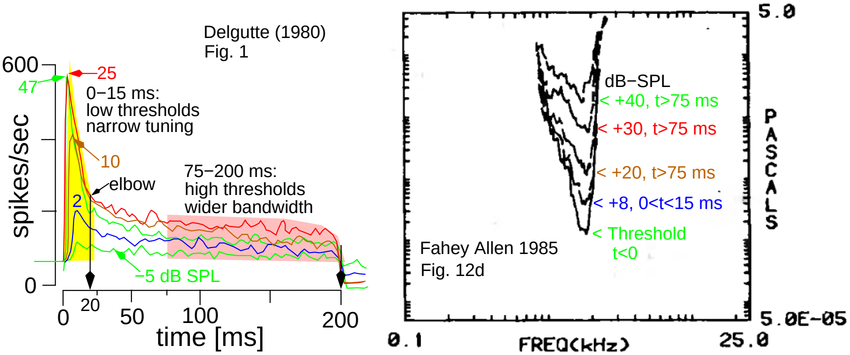
- Nonlinear upward spread of masking



- Sewell, William; Hearing Research v. 14, 305-314 (1984): 1 dB/mv EP threshold sensitivity

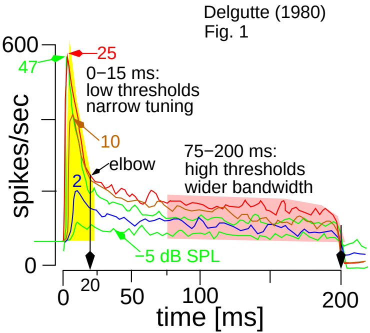
Neural Onset Enhancement   Delgutte 1980

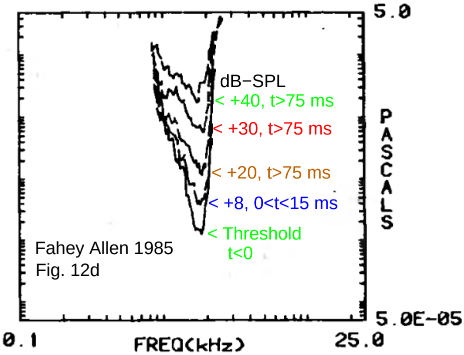
- Onset transients enhance the auditory nerve response, to 2 [cs]



- Forward Masking depresses the response up to 40 dB, to 20 [cs]

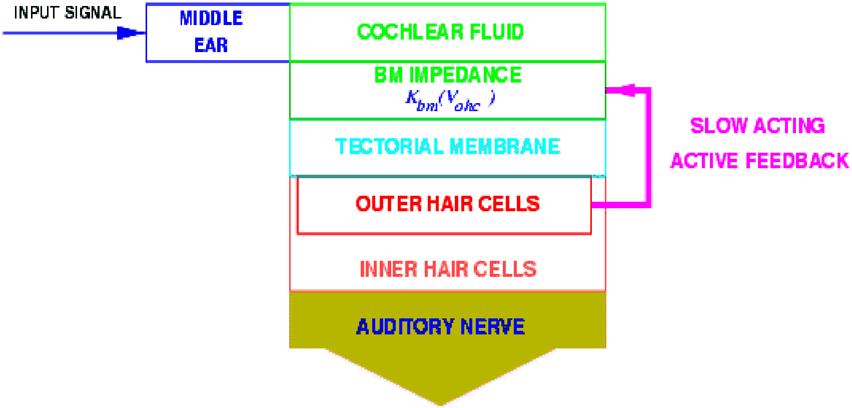
- Onset transients enhance the auditory nerve response, to 2 [cs]

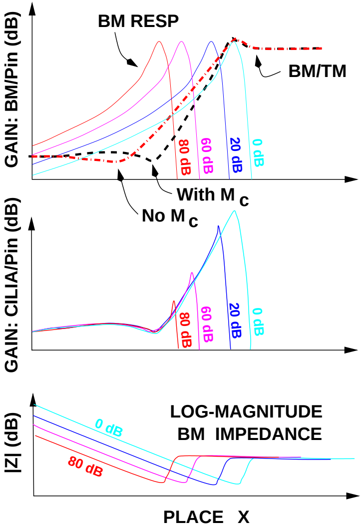




- Forward Masking depresses the response up to 40 dB, to 20 [cs]

The Human Cochlea   Allen 2000




BM RESP
BM/TM
GAIN: BM/Pin (dB)
80 dB
60 dB
20 dB
0 dB
With M c
No M c
GAIN: CILIA/Pin (dB)
80 dB
60 dB
20 dB
0 dB
|Z| (dB)
0 dB
80 dB
LOG-MAGNITUDE
BM IMPEDANCE
PLACE X

Conclusions I

We have:

- Isolated events for CV: Plosives /p, t, k/ and /b, d, g/ and Fricatives /θ, ʃ, ʧ, s, h, f/ and /z, ʒ, v, ð/) + Vowels /o, ɛ, ɪ/
  - for many individual talkers
  - via new tools (AI-gram, Event-gram and 3^d-DS)
- Shown that normal listeners use:
  - *across-frequency timing coincidences*
  - duration, modulation & bandwidth

  to discriminate consonants in noise
- Developed tools to:
  - Morphed speech sounds
  - Decrease or increase intelligibility. Ex: /tɑ/, /tɛ/

Conclusions I

We have:

- Isolated events for CV: Plosives /p, t, k/ and /b, d, g/ and Fricatives /θ, ʃ, ʧ, s, h, f/ and /z, ʒ, v, ð/) + Vowels /o, ɛ, ı/
  - for many individual talkers
  - via new tools (AI-gram, Event-gram and 3^d-DS)
- Shown that normal listeners use:
  - *across-frequency timing coincidences*
  - duration, modulation & bandwidth

  to discriminate consonants in noise
- Developed tools to:
  - Morphed speech sounds
  - Decrease or increase intelligibility. Ex: /tɑ/, /tɛ/

Conclusions I

We have:

- Isolated events for CV: Plosives /p, t, k/ and /b, d, g/ and Fricatives /θ, ʃ, ʧ, s, h, f/ and /z, ʒ, v, ð/) + Vowels /o, ɛ, ɪ/
  - for many individual talkers
  - via new tools (AI-gram, Event-gram and 3^d-DS)
- Shown that normal listeners use:
  - *across-frequency timing coincidences*
  - duration, modulation & bandwidth

  to discriminate consonants in noise
- Developed tools to:
  - Morphed speech sounds
  - Decrease or increase intelligibility. Ex: /tɑ/, /tɛ/

Conclusions I

We have:

- Isolated events for CV: Plosives /p, t, k/ and /b, d, g/ and Fricatives /θ, ʃ, ʧ, s, h, f/ and /z, ʒ, v, ð/) + Vowels /o, ɛ, ɪ/
  - for many individual talkers
  - via new tools (AI-gram, Event-gram and 3^d-DS)
- Shown that normal listeners use:
  - *across-frequency timing coincidences*
  - duration, modulation & bandwidth

  to discriminate consonants in noise
- Developed tools to:
  - Morphed speech sounds
  - Decrease or increase intelligibility. Ex: /tɑ/, /tɛ/

Conclusions I

We have:

- Isolated events for CV: Plosives /p, t, k/ and /b, d, g/ and Fricatives /θ, ʃ, ʧ, s, h, f/ and /z, ʒ, v, ð/) + Vowels /o, ɛ, ɪ/
    - for many individual talkers
    - via new tools (AI-gram, Event-gram and 3^d-DS)
- Shown that normal listeners use:
    - *across-frequency timing coincidences*
    - duration, modulation & bandwidth

  to discriminate consonants in noise
- Developed tools to:
    - Morphed speech sounds
    - Decrease or increase intelligibility. Ex: /tɑ/, /tɛ/

Conclusions II

We have shown:

1. The existence of conflicting cues
    - Thus MaxEnt consonants are NOT redundant
2. that the event threshold is abrupt (i.e., 6 dB)
3. proven the AI band-product formula (yet again)
4. why the AI works
    - Due to the frequency and SNR event distribution
5. the role of forward and upward masking spread

Conclusions II

We have shown:

1. The existence of conflicting cues
    - Thus MaxEnt consonants are NOT redundant
2. that the event threshold is abrupt (i.e., 6 dB)
3. proven the AI band-product formula (yet again)
4. why the AI works
    - Due to the frequency and SNR event distribution
5. the role of forward and upward masking spread

Conclusions II

We have shown:

1 The existence of conflicting cues
   - Thus MaxEnt consonants are NOT redundant

2 that the event threshold is abrupt (i.e., 6 dB)

3 proven the AI band-product formula (yet again)

4 why the AI works
   - Due to the frequency and SNR event distribution

5 the role of forward and upward masking spread

Conclusions II

We have shown:

1. The existence of conflicting cues
   - Thus MaxEnt consonants are NOT redundant
2. that the event threshold is abrupt (i.e., 6 dB)
3. proven the AI band-product formula (yet again)
4. why the AI works
   - Due to the frequency and SNR event distribution
5. the role of forward and upward masking spread

Conclusions II

We have shown:

1. The existence of conflicting cues
   - Thus MaxEnt consonants are NOT redundant
2. that the event threshold is abrupt (i.e., 6 dB)
3. proven the AI band-product formula (yet again)
4. why the AI works
   - Due to the frequency and SNR event distribution
5. the role of forward and upward masking spread

Conclusions III

This could lead to:

1. Improved automatic speech recognition front-ends
2. The design of new hearing aids

Conclusions III

This could lead to:

1. Improved automatic speech recognition front-ends
2. The design of new hearing aids

Topics for discussion

- Theory should be based on Shannon's Theory of Information
  1. SNR and Entropy (& token!) are key variables: AI(SNR) and channel capacity $\mathcal{C}(SNR)$
  2. Token Phone error is binary wrt SNR
  3. Tokens have a large threshold SD
     - Never Averaging across tokens!
     - Do not use DF (depends on averages)
  4. Entropy is the ideal measure of confusions
  5. Very few studies consider Entropy vs. SNR
     - NO: Fletcher 1914-1950
     - YES: Miller Nicely 1955
  6. The AI(SNR) has a huge "across & within" consonant SD
- Summary: Call upon Information Theory to:
  - "We eliminate the suspects one by one. We do not scatter around like puppies."

    –Hercule Poirot

Question your basic assumptions

Thanks for your attention

http://auditorymodels.org

- Status of the cochlear amplifier model: $\cdots$
- Is it time for a paradigm shift?