The Role of Cue Enhancement and Frequency Fine-tuning in Hearing Impaired Phone Recognition

Ali Abavisani

May 14, 2019

Outline

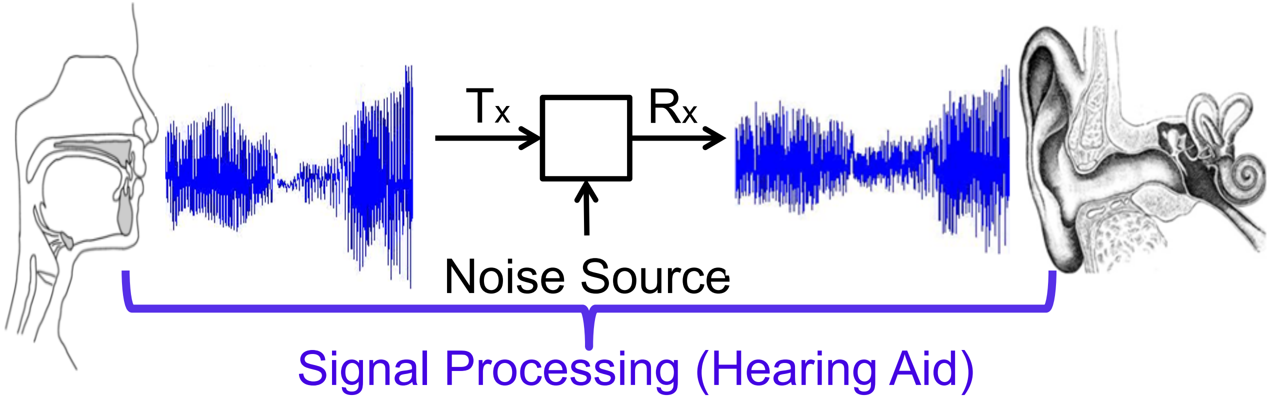
- Problem statement and motivation
- Background studies related to current research
- Proposed experiments to investigate the problem
- Preliminary results

Problem Statement and Motivation

- Hearing impairment profile in the US [Worlds Health Org.]
  - 38 million (12.2%) Americans have significant hearing loss
  - 3 out of 1000 (0.3%) of new born babies in the US are born with hearing loss
  - 1 out of 3 people over the age 65 are living with hearing loss in the US

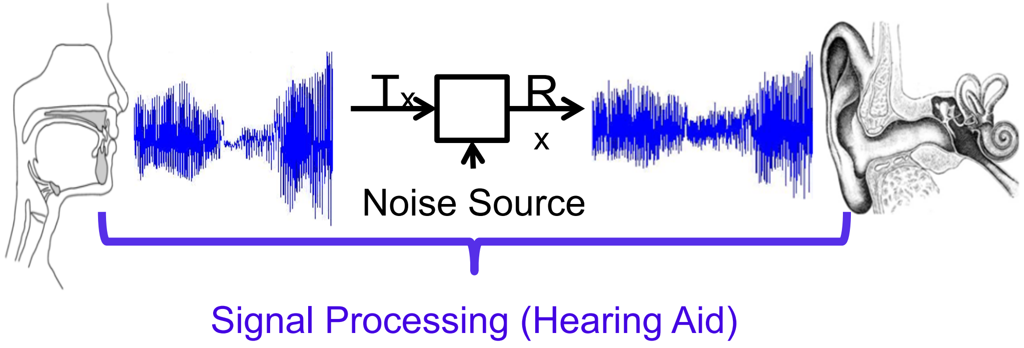
- Hearing Aids (HA)
  - Compensate for hearing loss based on pure-tone thresholds (PTT)
    - ✓ Makes speech signal audible

- Persistency of problem for HI listeners
  - Users of hearing aids have difficulty in speech recognition specially at noisy environments where the background noise is similar to speech
  - This can be related to the focus on audibility of speech through applying frequency dependent amplification, as opposed to a speech-based test



Problem Statement and Motivation

- NH speech perception
  - Speech cues can be determined by LP/HP filtering, time truncation
  - We can assign a perceptual measure as noise threshold level to each token by testing them at various SNRs
  - NH listeners respond to cue enhancement in the presence of noise
- HI speech perception
  - HI confusion patterns are similar to NH
  - PTT-based audibility amplification is not always helpful
    - ✓ more complex approach is needed
  - Noise threshold plays an important role in HI phone recognition



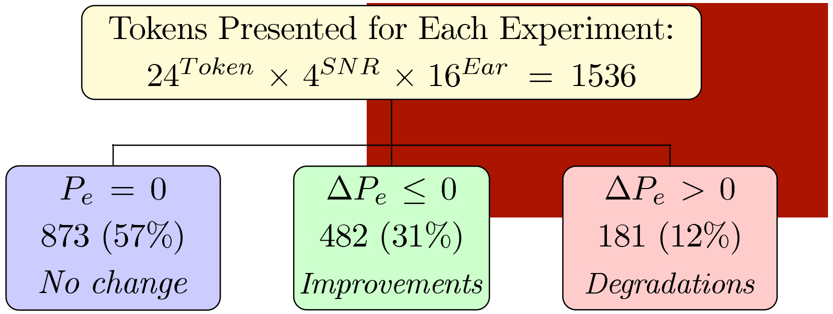
- Motivation
  - Assist HA amplification strategy
    - ✓ Identify problematic consonants
    - ✓ Investigate correct strategy for speech enhancement
  - Identify the appropriate amplification amount for target phones

HI Speech Perception background

- Hearing impairment
  - Hearing Loss (HL) above 20 [dB] in 0.25-8 [kHz]
  - Ears can have mild (< 40 dB), moderate (< 70 dB), severe (< 90 dB), and profound HL (above 90 dB)
- Speech tests for HI
  - Around 58% of words in spoken English consists of consonants [Mines et. al. 1978]
  - Accuracy of consonant recognition is highly correlated with SNR for HI ears [Plomp 1986, Kreul et. al. 1969]
  - Non-sense speech syllables such as Consonant-Vowel (CV) is one way to examine consonant recognition in speech based tests [Kreul et. al. 1969, Boothroyd 1995]
- HI phone recognition
  - A lot of complexity
    - ✓ Same CV sound has different confusion patterns [Trevion & Allen, 2013]
    - ✓ Same HA gain can help recognize some CVs, but reduce recognition for other CVs [Abavisani & Allen 2017]
    - ✓ Phone recognition is idiosyncratic for HI ears [Abavisani & Allen, 2017]

HI Speech Perception background

- Prior experiments showed that a few sounds were erroful for each HI ear with or without frequency dependent insertion gain
- HA insertion gain improved phone recognition accuracy for HI ears in most cases not all

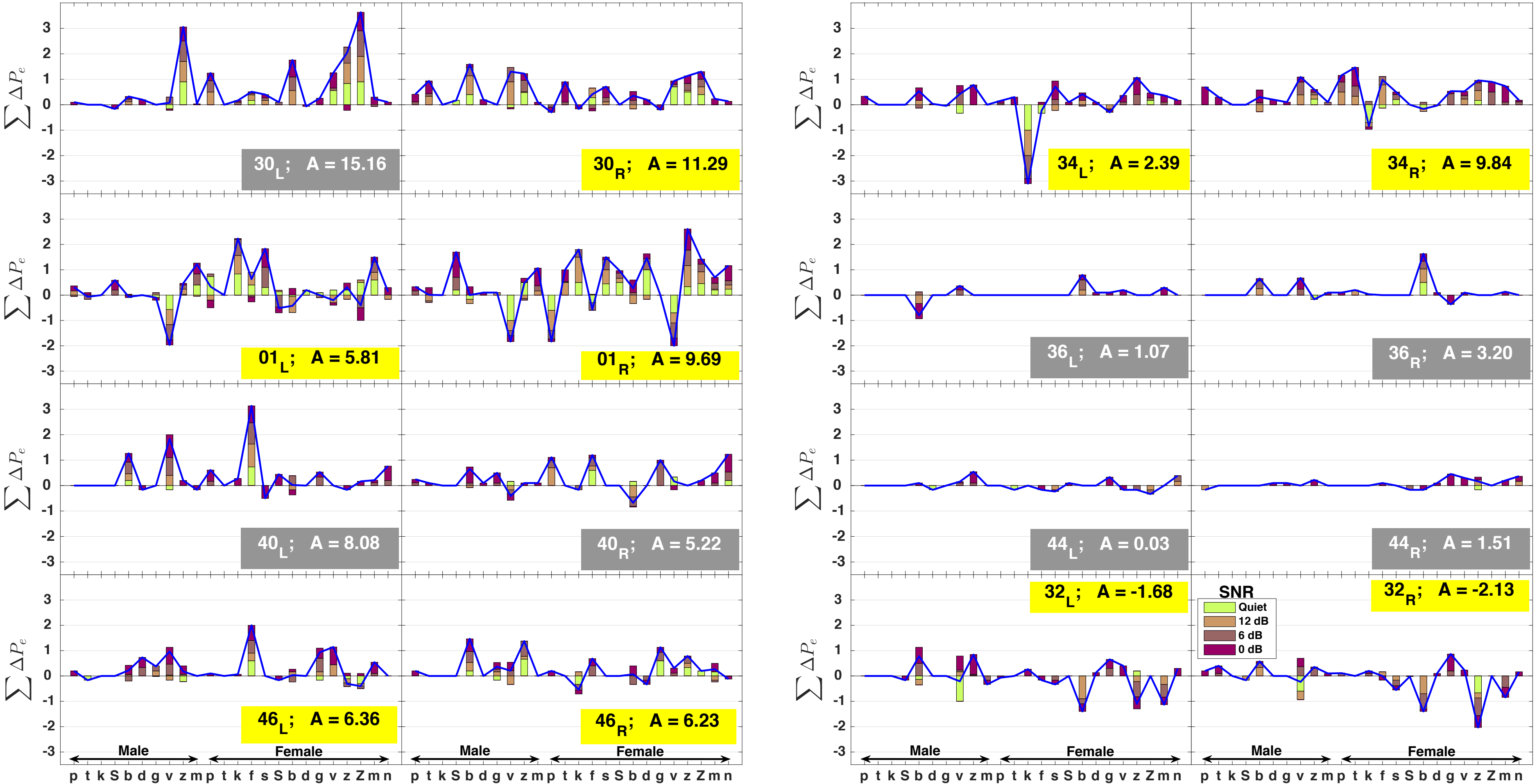


Tokens Presented for Each Experiment: $24^{Token} \times 4^{SNR} \times 16^{Ear} = 1536$

$P_e = 0$
873 (57%)
No change

$\Delta P_e \leq 0$
482 (31%)
Improvements

$\Delta P_e > 0$
181 (12%)
Degradations



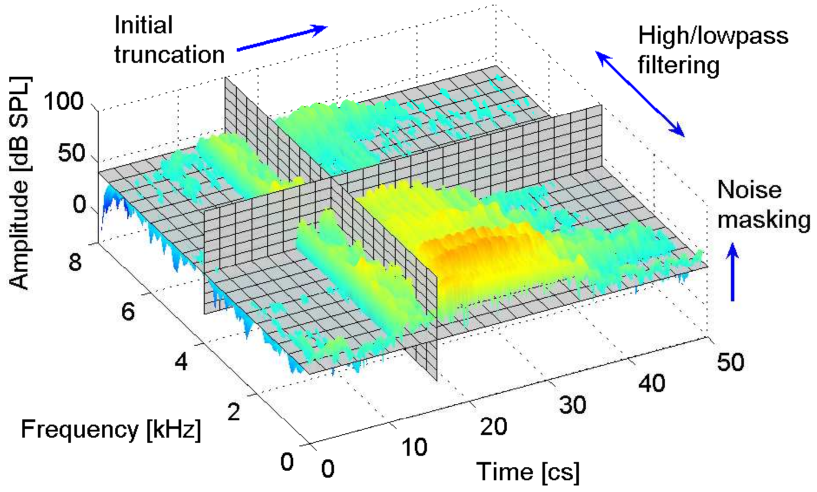
Reference:

Evaluating hearing aid amplification using idiosyncratic consonant errors

[Abavisani and Allen, 2017]

NH Speech Perception Background

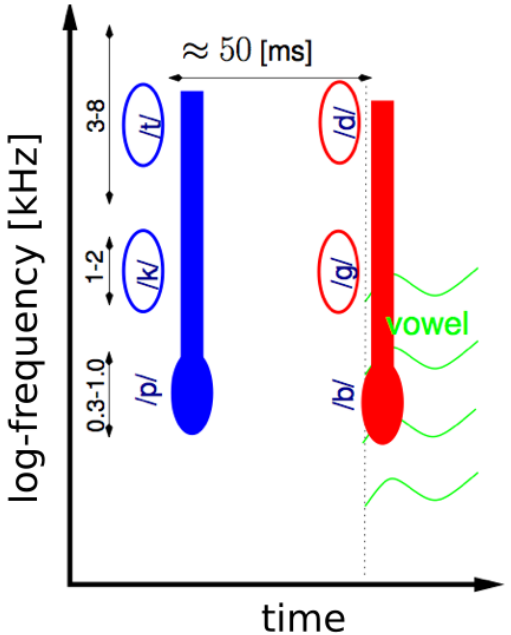
- AI-gram
  - Time frequency speech feature that includes SNR in human critical bands
  - It is an image corresponding to audible speech features in the masking noise
  - Used to identify primary cue region in speech tokens

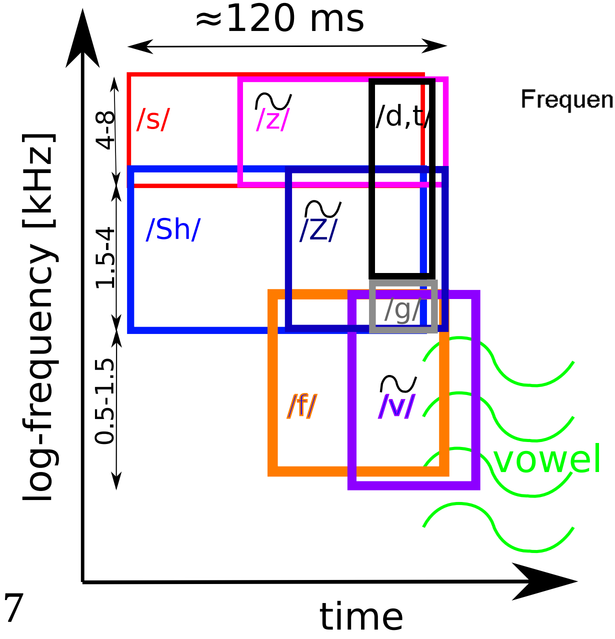


- 3D Deep Search to identify perceptual cues in tokens
  - Low/High pass filtering, Time truncation, SNR adjustment

- Perceptual cues

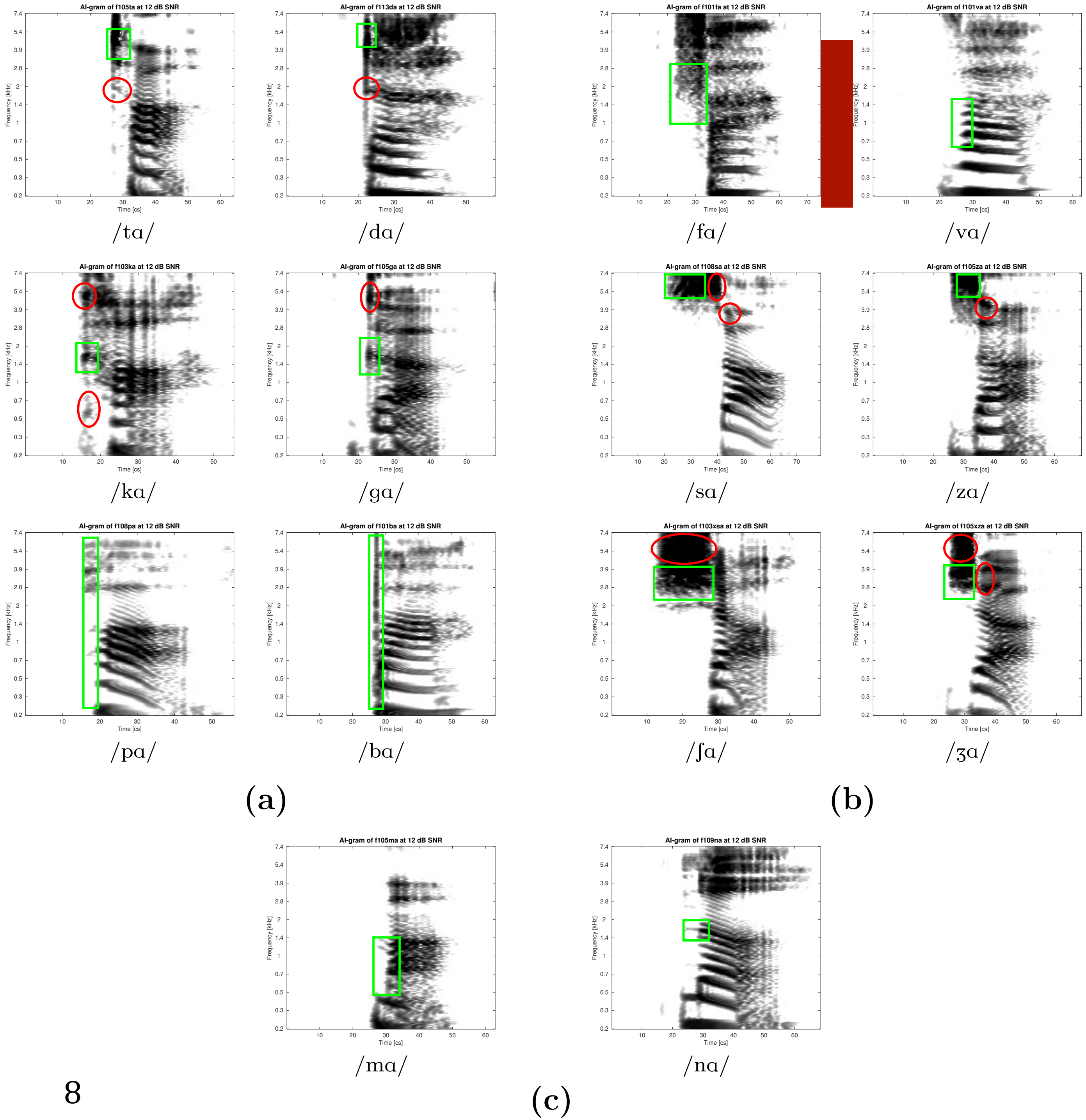
Plosives: /t, k, p, d, g, b/+/ɑ/





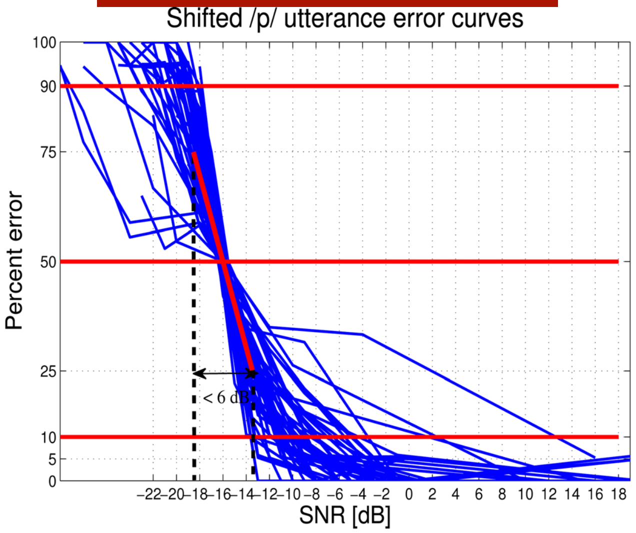
Examples of perceptual cues

- Primary cue region (green)
- Conflicting cue region (red)



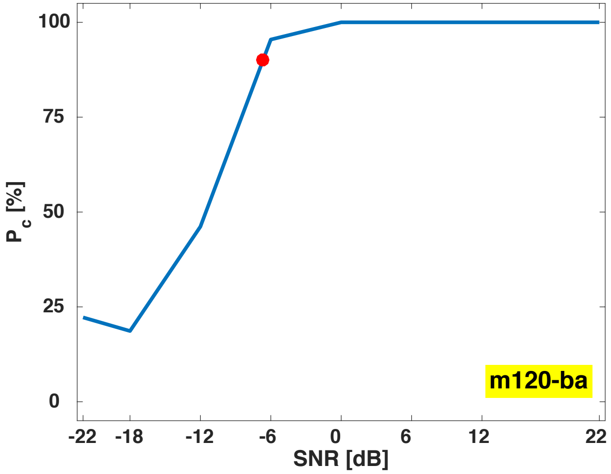
Experiments to determine SNR_{90}

- SNR_{90}
  - ➢ SNR in which NH listeners on average can recognize the sound at least 90% correct
  - ➢ Is a useful summary of the perceptual response of NH ears to a particular token
  - ➢ SNR_{90}, SNR_{50}, and SNR_{10} predict one another with low error for almost all tokens
  - ➢ If we shift P_e [%] curves to align their SNR_{50}, we observe that within a range of a few dB (i.e., +/-6 [dB]), the score drops around 50%
  - ➢ Enforce consistency by removing outliers (tokens whose SNR_{50} and SNR_{90} are not consist)



[Singh & Allen, 2012]

- Present the CV tokens to +30 NH listeners in a random fashion
  - ➢ Start at high SNR (SNR > 20 dB)
  - ➢ Two down, one up procedure
    - ✓ If subject recognizes the CV correctly, play the CV at two SNR levels down
    - ✓ If subject have error in the CV, play the CV at one SNR level up
  - ➢ Continue until reaching three cycles within a same loop
  - ➢ Plot the average score versus SNR, the SNR in which the plot passes 90% from the right for first time, is the SNR_{90}
- The SNR_{90} of CV is the average SNR_{90} thresholds across all NH subjects



SNR_{90}:
A Perceptual Measure for Understanding Speech in Noise

- Experiment I: Try to improve intelligibility for HI listeners by improving SNR_{90}
  - Experiment: improve SNR_{90} by choosing a different talker
  - Changing the talker may change the score, depending to the SNR_{90} of CV [Toscano & Allen, 2014]
  - NH listener should recognize the CV correctly at any SNR at least 6 [dB] above the SNR_{90} [Singh & Allen, 2012]

- Experiment 2: Change to a token with different vowel, but with the same SNR_{90}
  - Changes the formant transitions [Ohman 1966, Delattre et. al., 1966, Sussman et. al., 1991]
  - Changes the center frequency of burst spectrum [Winitz et. al., 1972]
  - Changes the acoustic specrotemporal context of relevant cues [Lisker 1975]
  - Changes the lexical context related to the CV [Ganong, 1980]

- We would like to control these effects by controlling over the SNR_{90}

SNR_{90}:
A Perceptual Measure for Understanding Speech in Noise

- Changing the token changes a lot of details of the waveform
- All tokens are pre-evaluated by SNR_{90}
- For NH listener, if CV_1 and CV_2 have similar SNR_{90}
  - ➢ primary consonant cue is about the same level in both CVs
- If HI have different P_e for these two CVs
  - ➢ must be caused by something other than the level of primary cue
    - ✓ Co-articulatory cues [Lisker 1975, Ohman 1966]
    - ✓ Spectrotemporal context [Stevens 1987]
    - ✓ Lexical neighborhood density [Ganong 1980]
- By controlling over SNR_{90}, we rule out the primary cue level as cause of perceptual deficiency

Usage of SNR_{90} in Experiment I: Talker Change

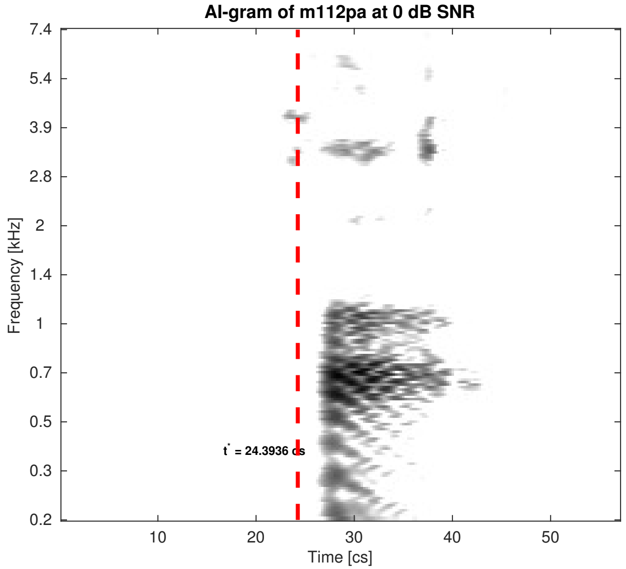
- For NH listeners, if we amplify the primary cue of the erroful CV to the levels ~ 6 [dB] above CV's SNR_{90}, the error should drop to ~ 0 [Kapoor & Allen, 2012]
  - ➢ Also, if we replace the CV by the same CV but with different talker with more clear voice (more salient CV), that has SNR_{90} well above previous CV, the error will drop to ~ 0 [Toscano & Allen, 2014]

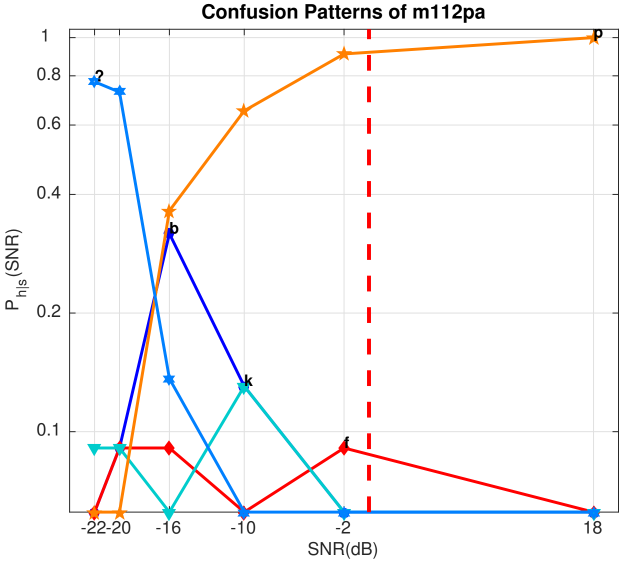
- We would like to investigate this fact on HI listeners (experiment I)
  - ➢ Hypothesis: In HI phone recognition, if we replace the CV by the same CV but with different talker with more clear voice (more salient CV), that has SNR_{90} well above previous CV, the error should drop
    - ✓ Replace CV_1 by CV_2 (same consonant and vowel) where $\mathrm{SNR90}_2 \geq \mathrm{SNR90}_1 + 6$ [dB]
    - ✓ This will constitute a change in the intensity of the primary cue region

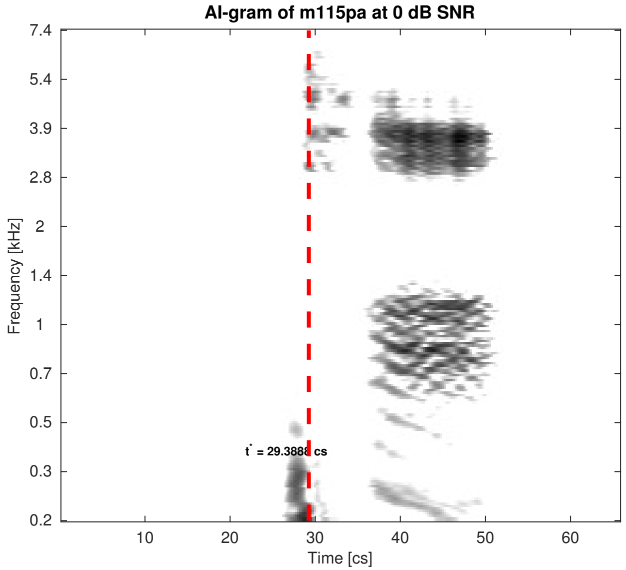
- Check the impact of this change on error, entropy, confusion pattern of the HI CV recognition

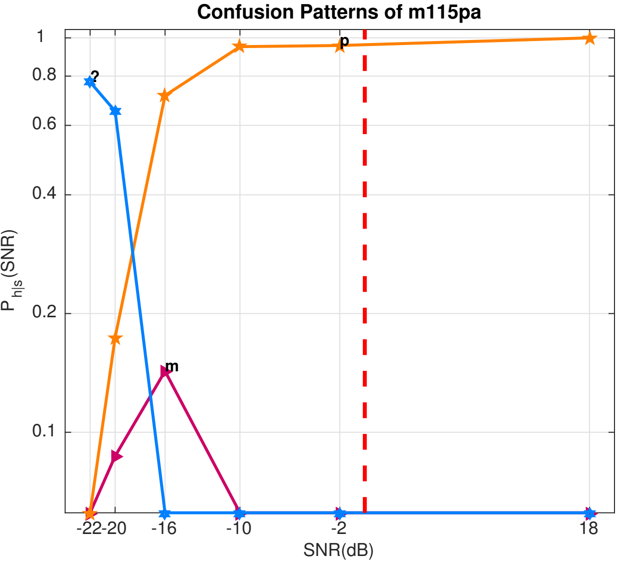
Example of cue change in Experiment I

- Replace /pa/ with more salient /pa/


AI-gram of m112pa at 0 dB SNR


Confusion Patterns of m112pa


AI-gram of m115pa at 0 dB SNR


Confusion Patterns of m115pa

Usage of SNR_{90} in Experiment II: Vowel Change

- NH CV recognition is affected by changing the vowel as a result of:
  - Formant transitions [Ohman 1966, Delattre et. al., 1966, Sussman et. al., 1991]
  - Displace of center frequency of burst spectrum [Winitz et. al., 1972]
  - Acoustic specrotemporal context variations of relevant cues [Lisker 1975]
  - Changes the lexical context related to the CV [Ganong, 1980]

- We would like to investigate whether these effect play role in HI phone recognition??
  - For this matter, we replace CV_1 by CV_2 with same consonant but with different vowel
  - CV_1 and CV_2 should have similar SNR_{90} ($|\Delta\mathrm{SNR}_{90}| \leq 3$ dB)
  - This will constitute a change in the spectrotemporal features of the consonant

- Check the impact of this change on error, entropy, confusion pattern of the HI CV recognition

Example of cue change in Experiment II

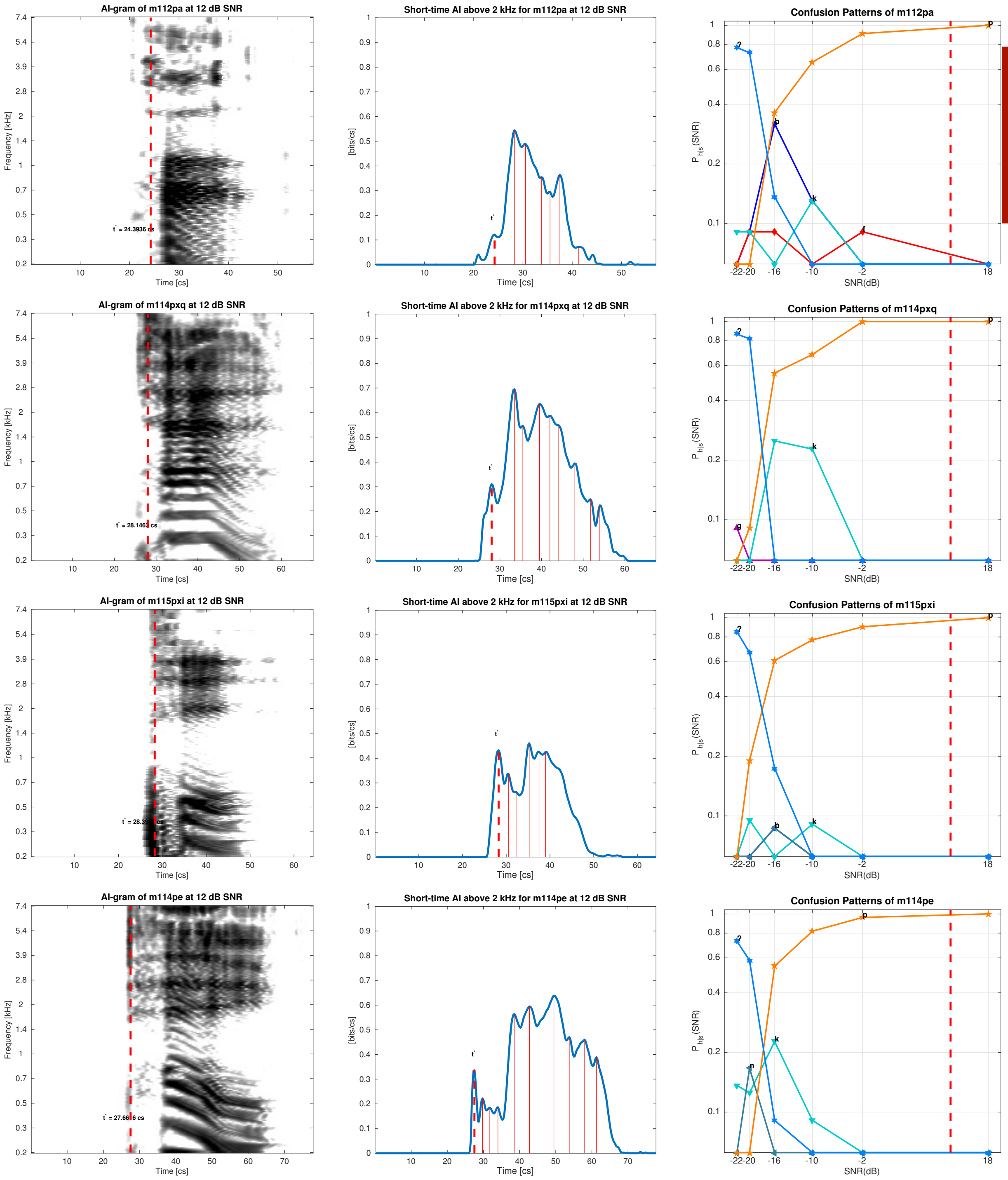
- Replace /pa/ with /p/+vowel with similar SNR$_{90}$

/pa/

/pae/

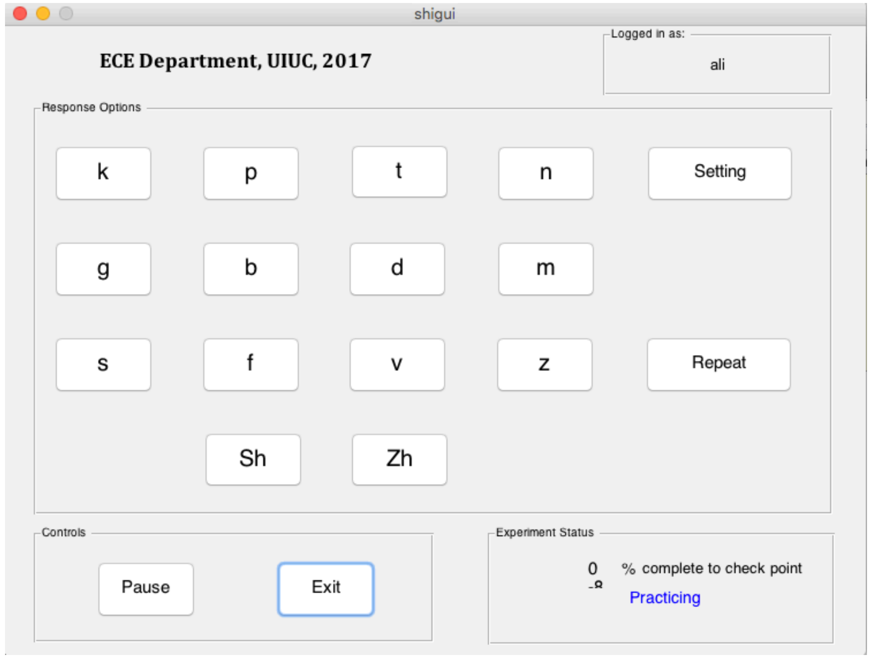
/pI/

/pɛ/



Designed Software for Adaptive Testing

- Subjects: HI subjects, age < 64, with mild to moderate hearing loss
- SNR = 0, 6, 12 dB and Quiet
- Speech material: Male+Female /p, t, k, f, s, S, b, d, g, v, z, Z, m, n/+/a, æ, I, ε/, presented at the Most Comfortable Level (MCL)
- Experiment I: Change Talker (Change intensity of primary cue)
  - ➢ Screening in List 1: Start with less salient CV at SNR = 0 dB
    - ✓ If CV had error, copy to List 2
  - ➢ Evaluation in List 2: present CV two times at SNR = 0 dB and one time at SNR = 6 dB
    - ✓ If two errors occurred out of three presentations, copy CV to List 3
    - ✓ Copy same CV with new more salient talker to List 2 (|ΔSNR90| > 6 dB)
    - ✓ Copy confusing sounds associated with this CV to List 2
  - ➢ Test in List 3: Present same CV 8 times at each SNR (total 32 presentations), record the response
- Experiment II: Change Vowel (shift frequency of primary cue)
  - ➢ Screening in List 1: Start with less salient C+/a/ at SNR = 0 dB (screening)
    - ✓ If CV had error, copy to List 2
  - ➢ Evaluation in List 2: present CV two times at SNR = 0 dB and one time at SNR = 6 dB
    - ✓ If two errors occurred out of three presentations, copy CV to List 3
    - ✓ Copy same consonant with 3 new vowels /æ, I, ε/ to List 2 (|ΔSNR90| < 3 dB)
    - ✓ Copy confusing sounds associated with these CVs to List 2
  - ➢ Test in List 3: Present same CV 8 times at each SNR (total 32 presentations), record the response



shigui

ECE Department, UIUC, 2017

Logged in as: ali

Experiment Status: 0 % complete to check point

Practicing

Designed Software for Adaptive Testing

- Confusing sounds pattern to induce more error
  - Derived from previous phone recognition experiments
  - Each consonant has up to 3 confusing consonants
  - Uniform transition probability for outgoing paths

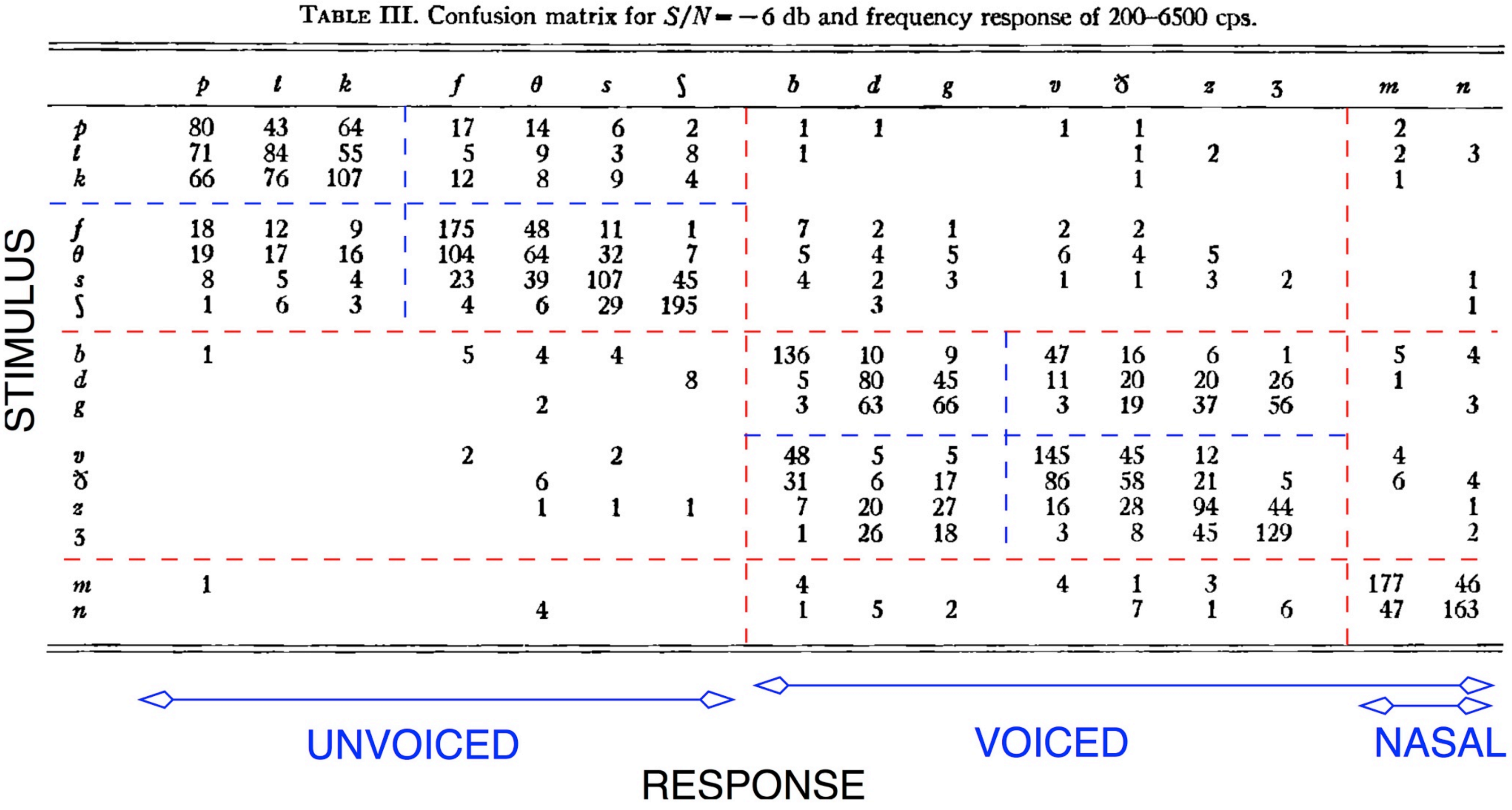
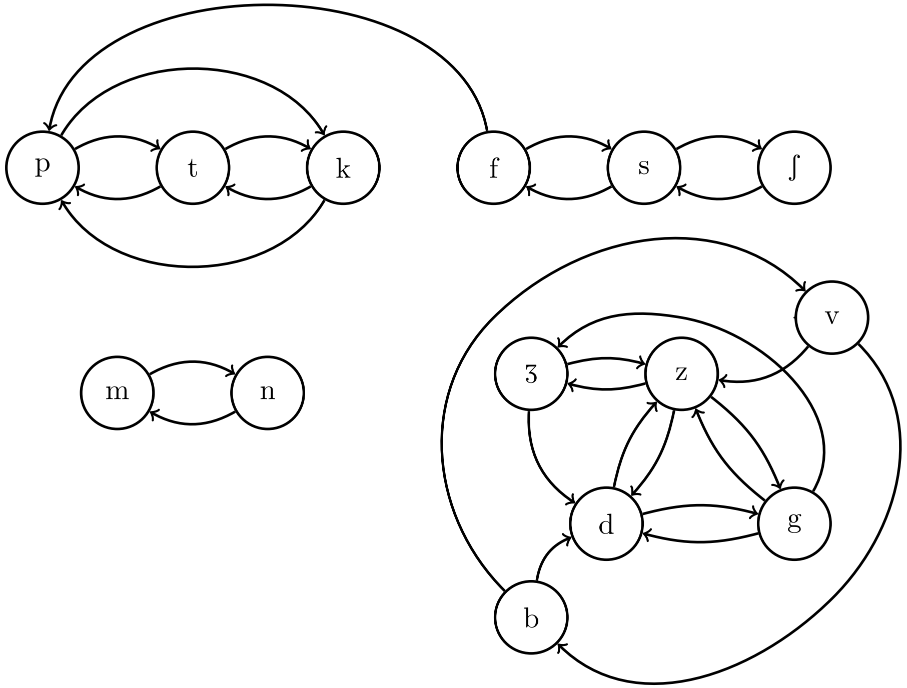


TABLE III. Confusion matrix for $S/N = -6$ db and frequency response of 200–6500 cps.

STIMULUS

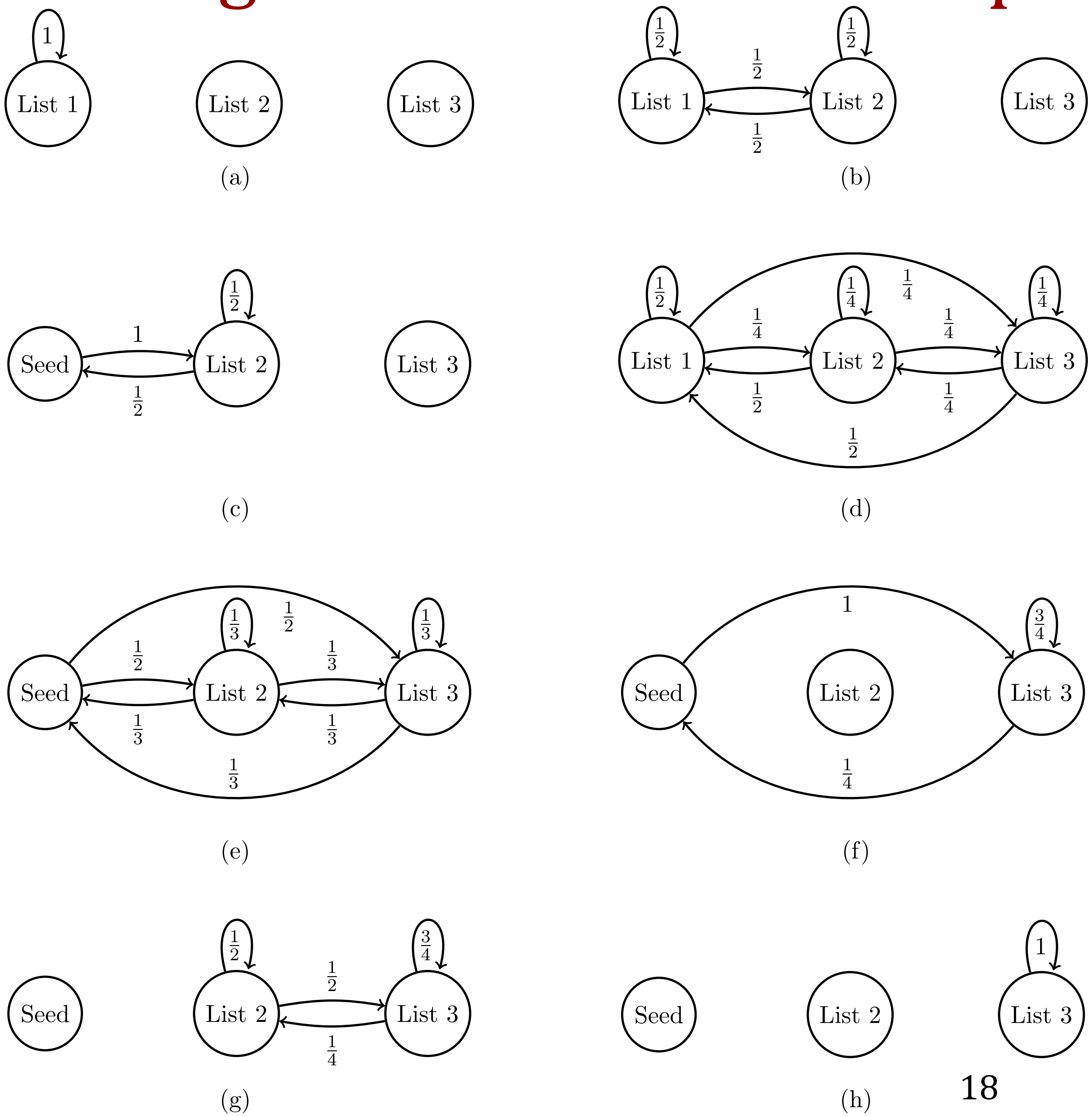
	p	t	k	f	θ	s	ʃ	b	d	g	v	ð	z	ʒ	m	n
p	80	43	64	17	14	6	2	1	1		1	1			2	
t	71	84	55	5	9	3	8	1				1	2		2	3
k	66	76	107	12	8	9	4					1			1	
f	18	12	9	175	48	11	1	7	2	1	2	2				
θ	19	17	16	104	64	32	7	5	4	5	6	4	5			
s	8	5	4	23	39	107	45	4	2	3	1	1	3	2		1
ʃ	1	6	3	4	6	29	195		3							1
b	1			5	4	4		136	10	9	47	16	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

RESPONSE



[Miller & Nicely 1955]

Designed Software for Adaptive Testing



(a) (b)

(c) (d)

(e) (f)

(g) (h)

- Transition probabilities between lists
  - ➢ To increase randomness, we use consonants from different confusion groups as seeds
  - ➢ When there is enough diversity of consonants (9+ different consonants), we use CVs within lists as seeds

Confusion Matrix data Analysis

- Form confusion matrix out of recorded response from List 3
- Convert confusion matrix to probability matrix
  - Divide each element by the row sum
- Probability of error for each token

$$P_e(CV_i, \mathrm{SNR}) = 1 - P_{ii} = \sum_{j \neq i} P\{heard\, CV_j \mid spoken\, CV_i\}$$

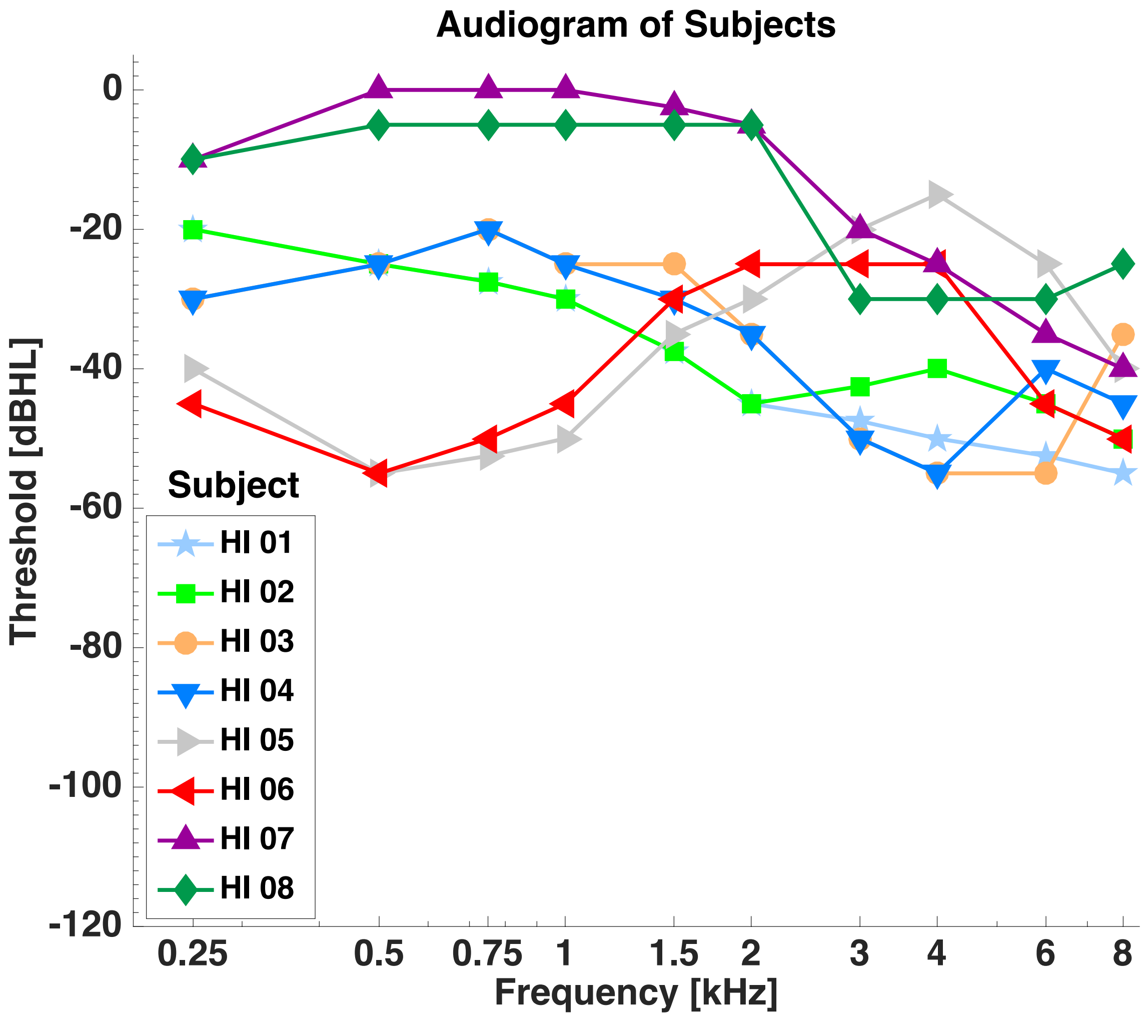
- Entropy of each token

$$\mathcal{H}(CV_i, \mathrm{SNR}) = -\sum_{j=1}^{14} P_{ij} log(P_{ij})$$

- ***Improvement***: error (entropy) in 2nd condition (after change) is smaller than 1st condition
- ***Degradation***: error (entropy) in 2nd condition (after change) is larger than 1st condition

Preliminary Results

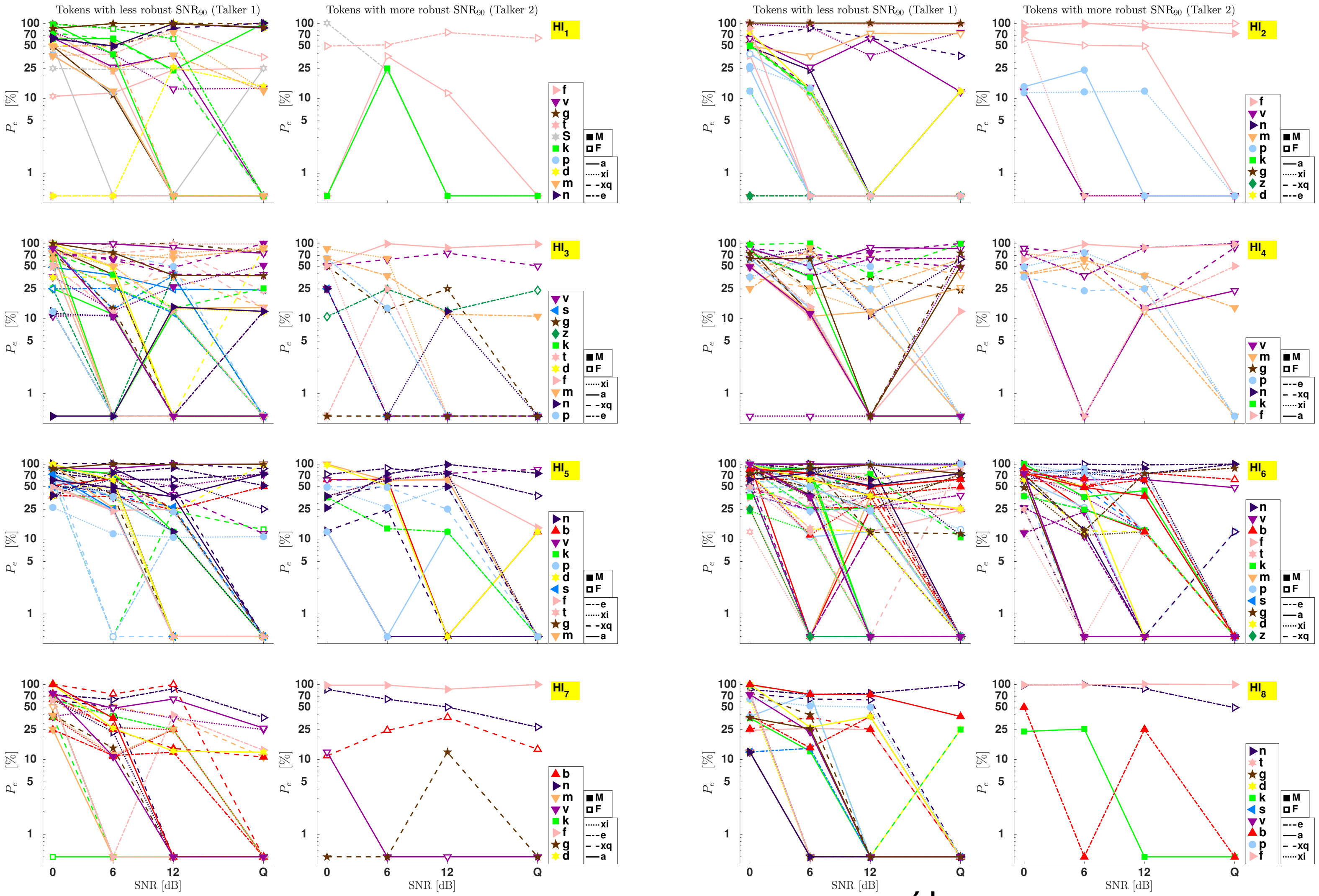
- Pure tone thresholds of 4 HI listeners



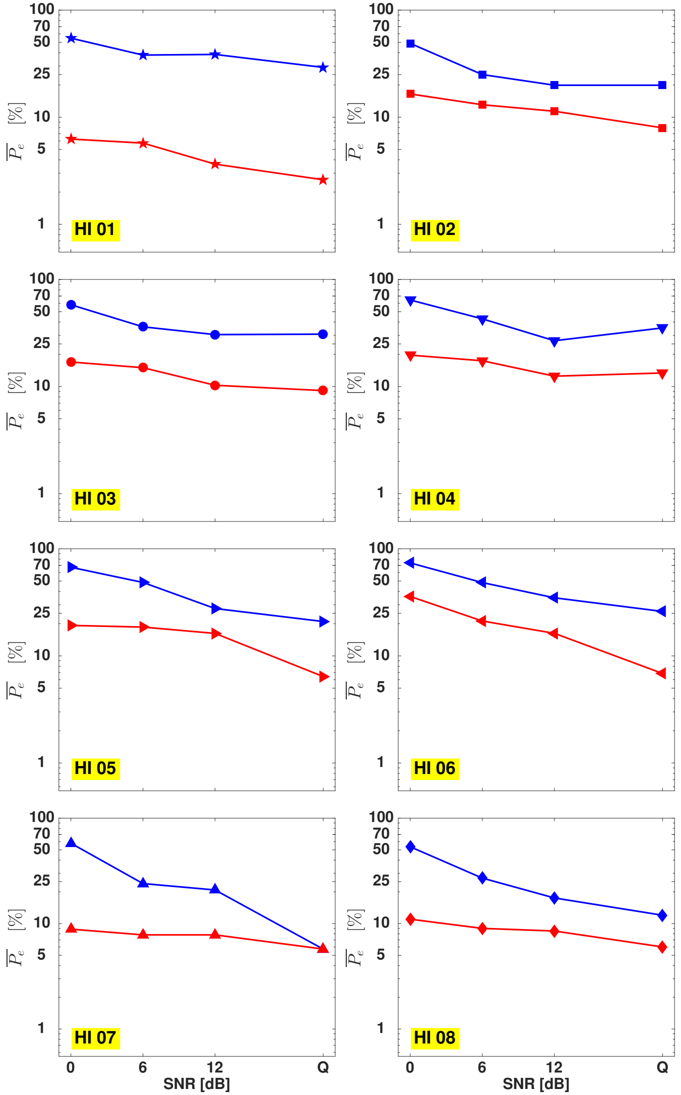
Preliminary Results: Experiment I

- Experiment I: change the talker (intensity of primary cue)
  - ➢ Improving SNR_{90} caused HI listeners to have fewer errors

Individual token errors

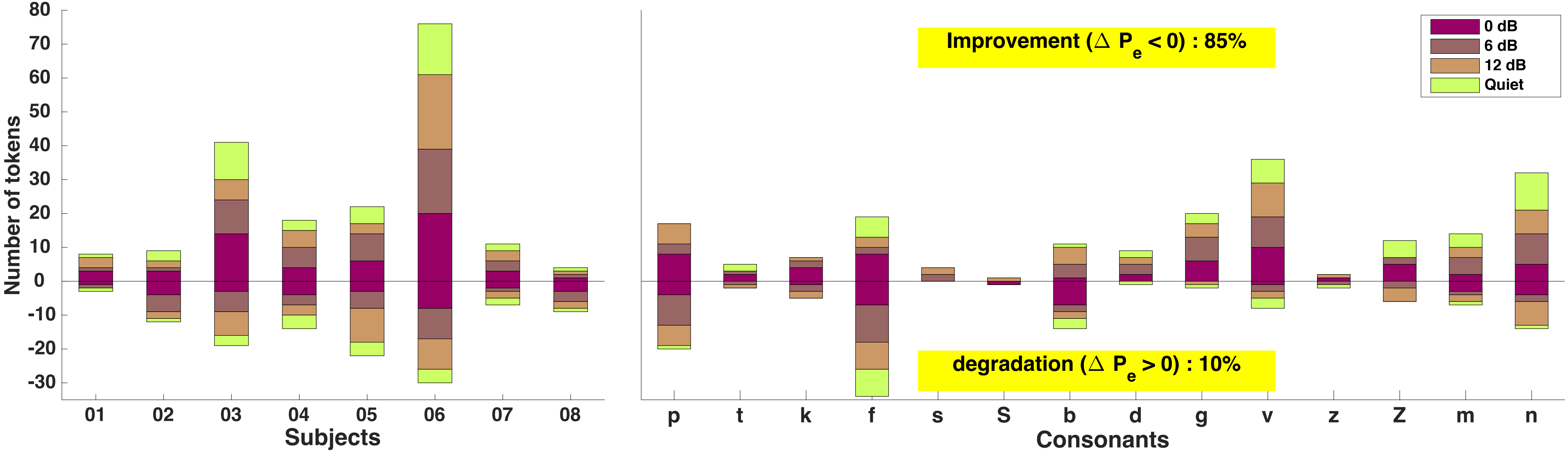


Average error



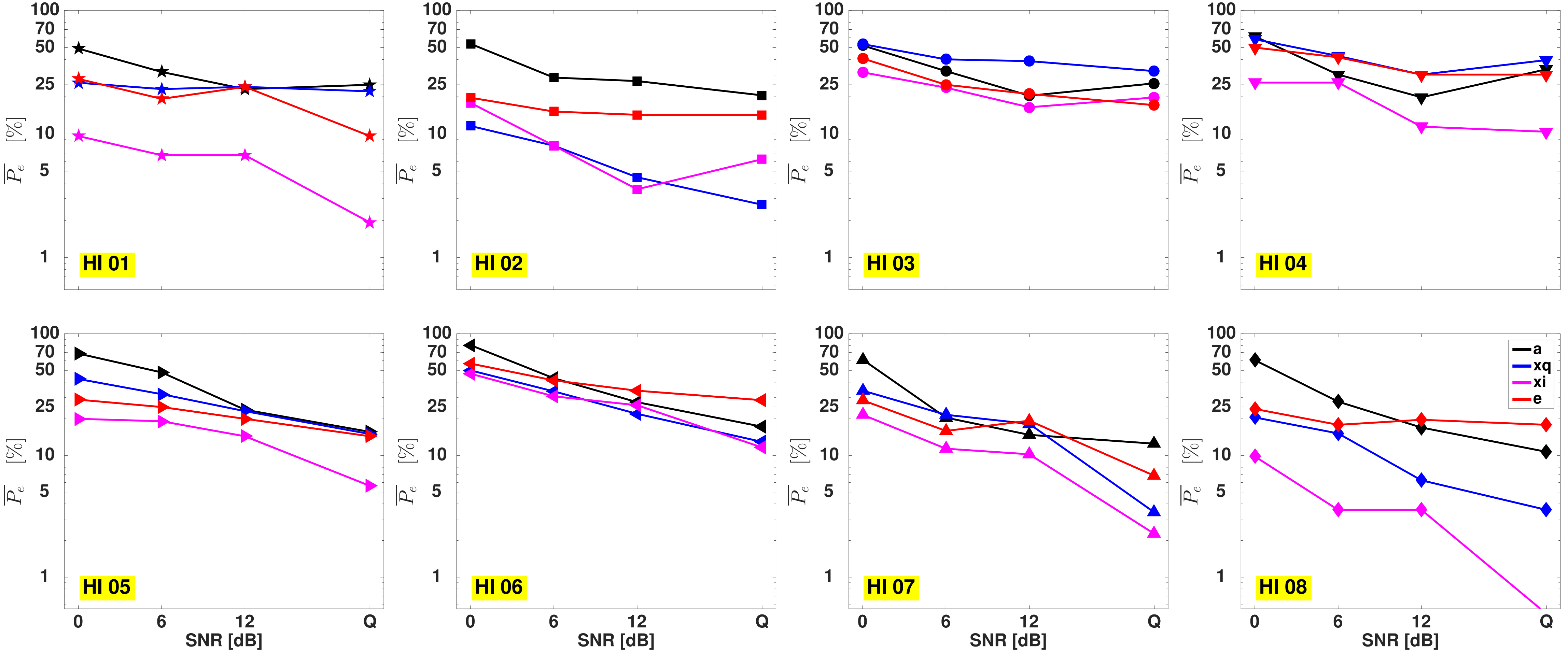
Preliminary Results: Experiment I

- Improvement vs degradation in error for talker change



Preliminary Results: Experiment II

- Experiment II: change the vowel (manipulate frequency of primary cue)
  - Average error for various vowels:



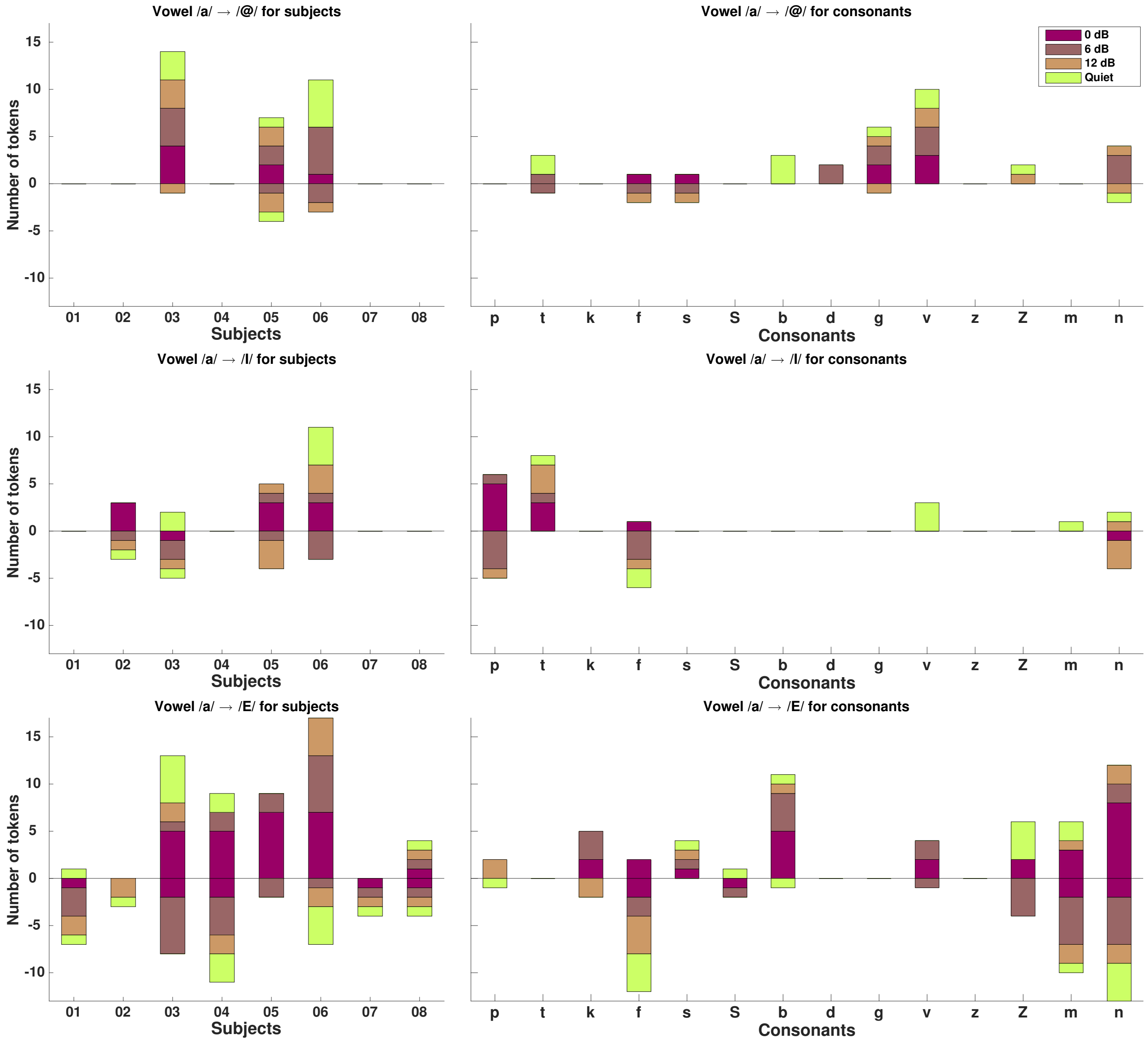
Preliminary Results: Experiment II

- Summary of vowel change improvement vs degradations for different vowels

Changed vowel	Improvement [%]	Degradation [%]
/a/	75	14
/ae/	71	16
/I/	63	24
/ε/	72	18

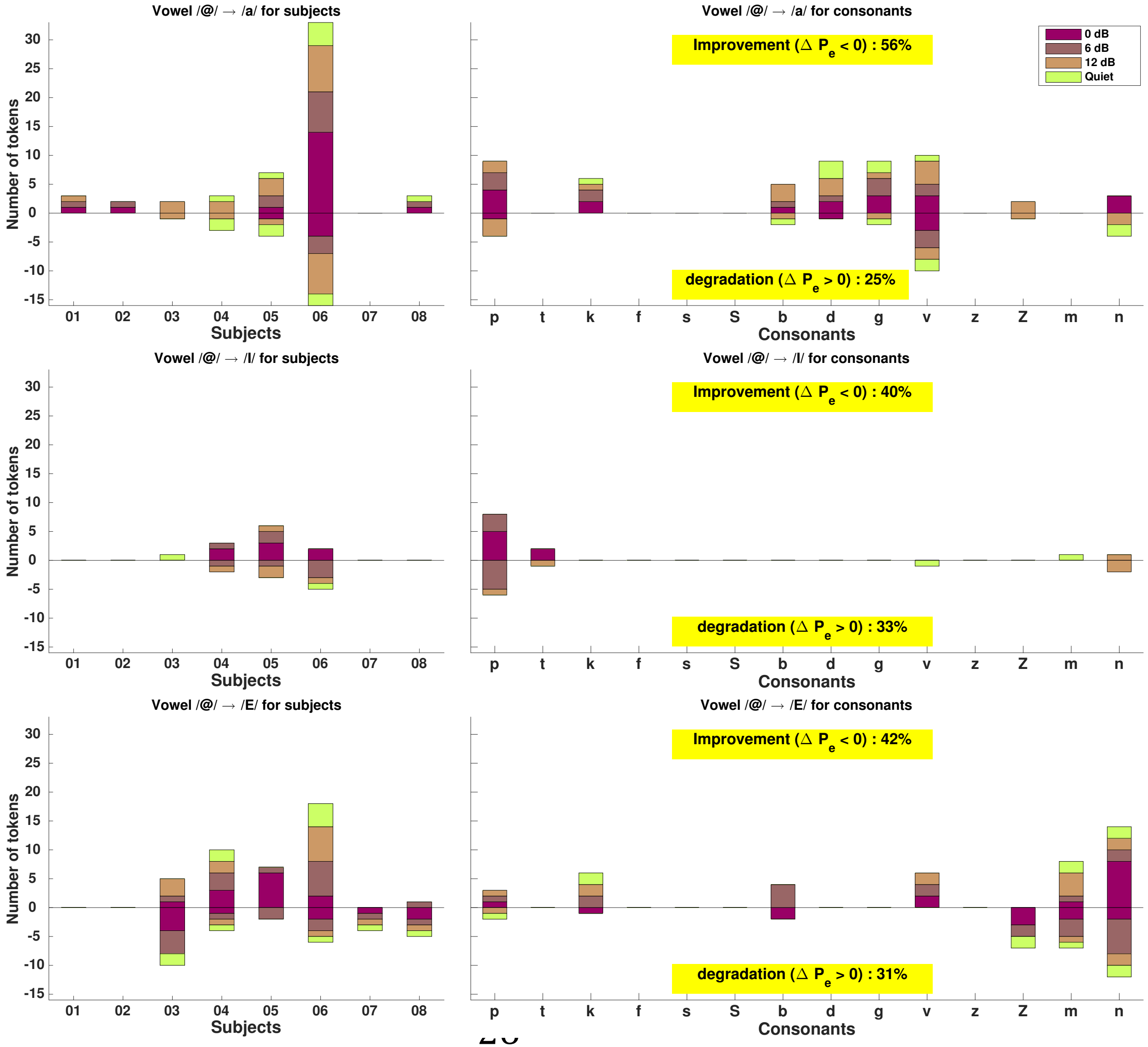
Preliminary Results: Experiment II

- Improvement vs degradation in error for vowel change
  - Vowel /a/ changes
    - To /ae/:
    - To /I/:
    - To /ε/:



Preliminary Results: Experiment II

- Improvement vs degradation in error for vowel change
  - Vowel /ae/ changes
    - To /a/:
    - To /I/:
    - To /ε/:



Vowel /@/ → /a/ for subjects

Vowel /@/ → /a/ for consonants

Improvement (Δ P_e < 0) : 56%

degradation (Δ P_e > 0) : 25%

Vowel /@/ → /I/ for subjects

Vowel /@/ → /I/ for consonants

Improvement (Δ P_e < 0) : 40%

degradation (Δ P_e > 0) : 33%

Vowel /@/ → /E/ for subjects

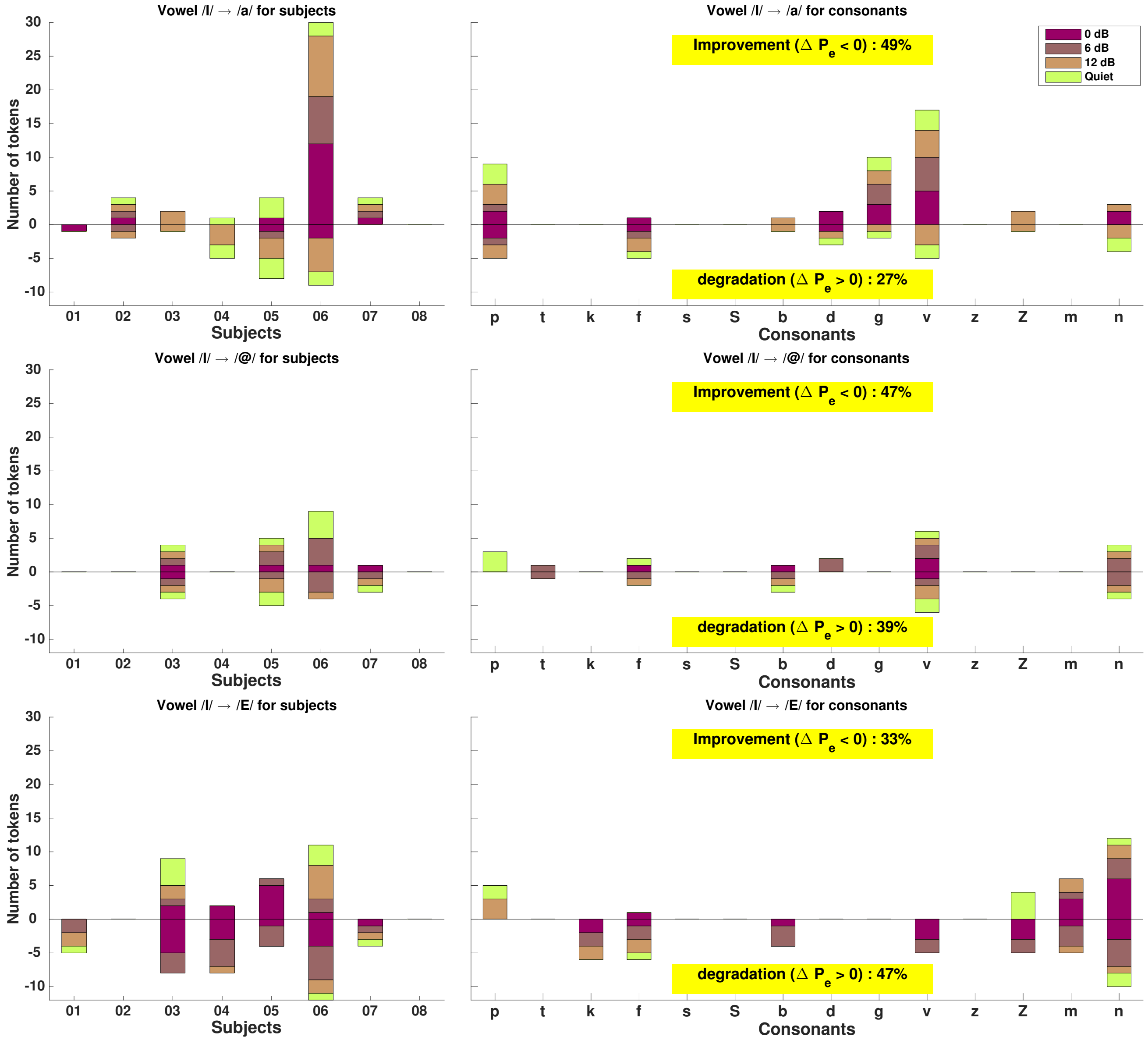
Vowel /@/ → /E/ for consonants

Improvement (Δ P_e < 0) : 42%

degradation (Δ P_e > 0) : 31%

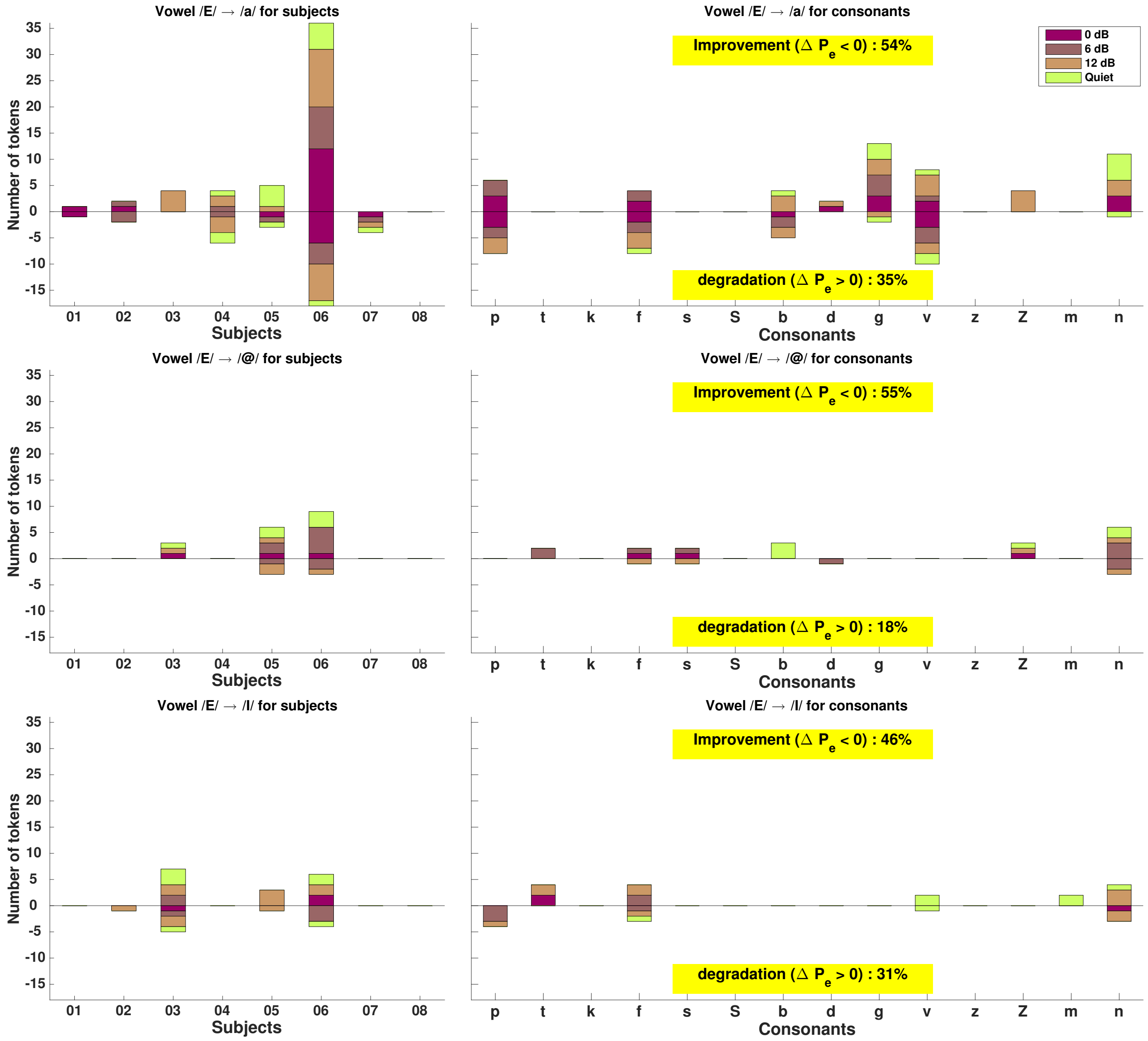
Preliminary Results: Experiment II

- Improvement vs degradation in error for vowel change
  - Vowel /I/ changes
    - To /a/:
    - To /ae/:
    - To /ɛ/:

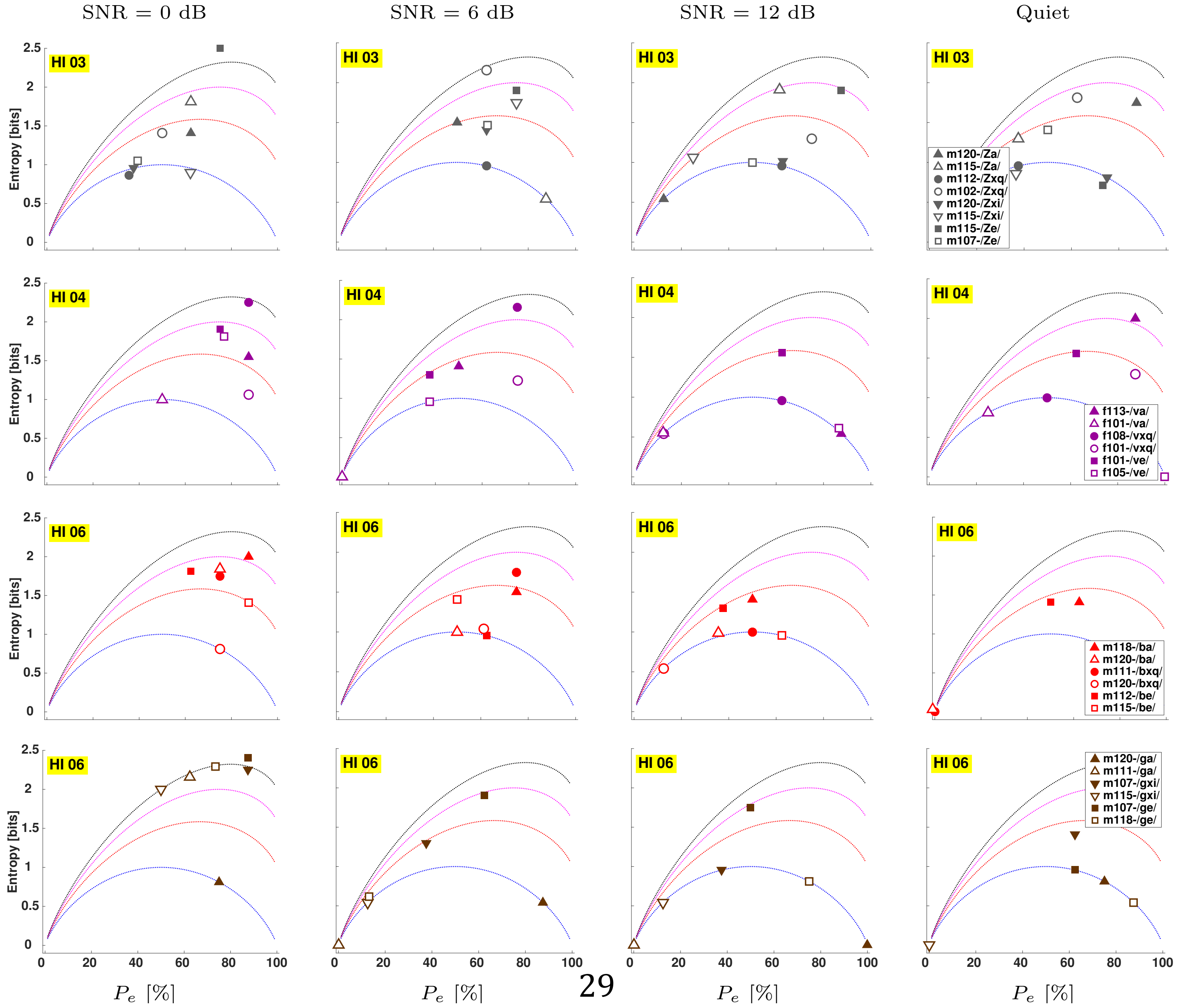


Preliminary Results: Experiment II

- Improvement vs degradation in error for vowel change
  - ➢ Vowel /ε/ changes
    - To /a/:
    - To /ae/:
    - To /I/:



Examples of entropy vs P_e curves



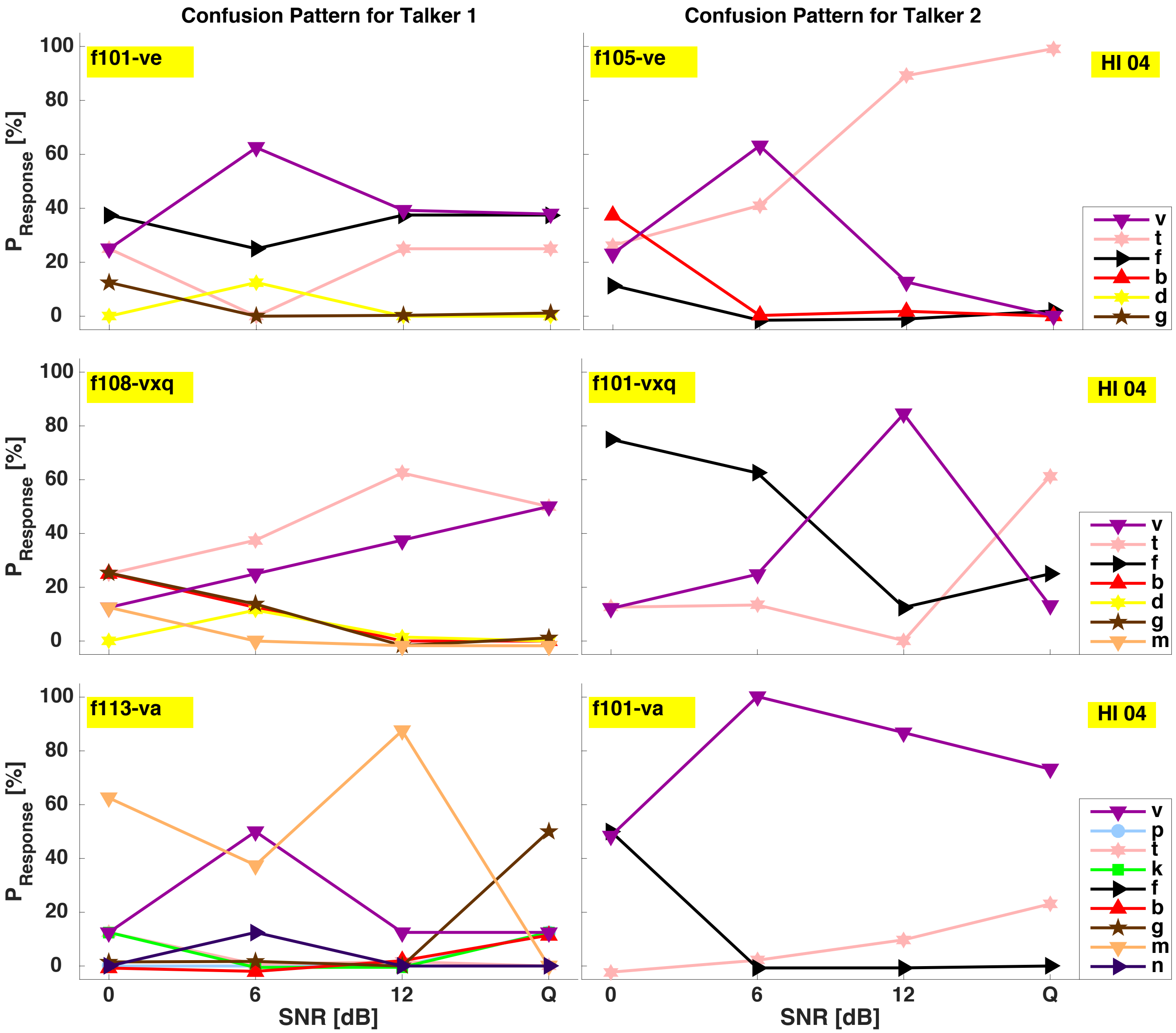
HI Consonant Recognition Predication by SNR_{90}

- HI confusion pattern are similar to NH [Trevino & Allen, 2013]

- SNR_{90}: a perceptual measure of hearing speech on noise, derived from NH data

- SNR_{90} can predict error for HI speech perception
  - ➢ Tokens presented in noise levels well above SNR_{90}, should be recognized by NH and HI
  - ➢ This is not always the case for HI
    - ✓ Higher noise can mask conflicting cues
      - ✓ Reducing noise in these cases may increase the error
    - ✓ Some HI ears do not respond to talker change as expected
      - ✓ Should investigate the conflicting cues
    - ✓ If vowel change (with similar SNR_{90}) increases the error for HI ear
      - ✓ Should investigate the particular changes on formant transitions, spectrotemporal context of CV, etc

Examples of complicated confusion patterns



Conclusions

- Results of this speech based test helps better understand
  - HI phone recognition strategy comparing to NH
    - ✓ The role of replacing talker with more salient talker (variation of intensity of primary cue)
    - ✓ The role of changing the vowel (variation of frequency of primary cue)
  - Categorize HI listeners based on their response (improvement vs degradation) in terms of error and entropy
  - Categorize consonants in terms of positive/negative responding to their acoustic spectrotemporal shift

- Average probability of error is not the best metric to understand HI phone recognition
  - Should look into individual sounds associate the error with confusion pattern

- Experiment on NH listeners verified SNR_{90} labels for test tokens

- Training a model to automatically estimate SNR_{90} perceptual measure for CV sounds helps to estimate the appropriate amplification amount needed for speech perception enhancement
  - Needs data augmentation since current SNR_{90} labeled data is limited
    - ✓ Extreme cases of augmented data should be evaluated by NH experiments to verify their SNR_{90}
  - Explore various models to compare the accuracy in estimation