Consonant perception

Sources of perceptual variability and modeling approaches

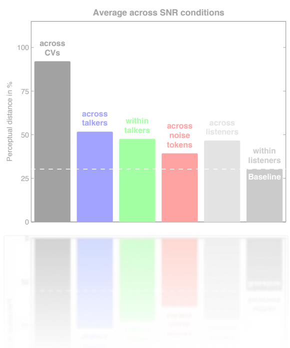
Johannes Zaar and Torsten Dau

Centre for Applied Hearing Research, Technical University of Denmark

7th Speech in Noise Workshop

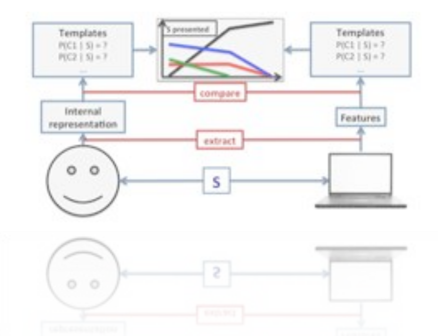
9 January 2015, Copenhagen

Outline



Part I. Sources of perceptual variability in consonant perception

- Experimental data analysis
- Based on perceptual distance

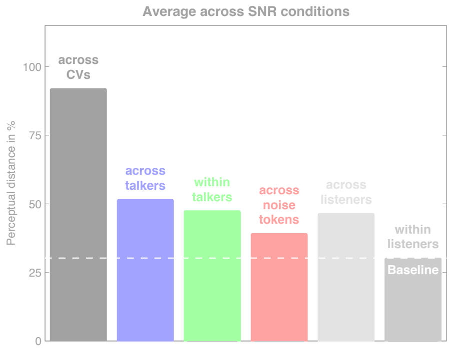


Part II. Modeling consonant perception

- Audibility and modulation front ends
- Template-matching back end
- Evaluation of model predictions

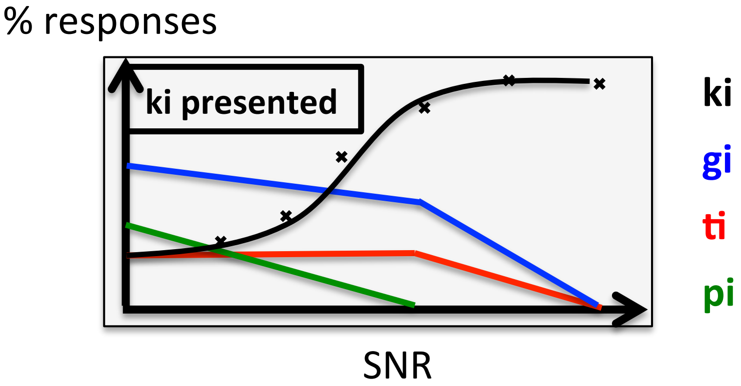
Part I.

Sources of perceptual variability in consonant perception



Consonant perception measurement

- Non-sense short-term stimuli
- Consonant-vowel combinations (CVs) like /ki/ in noise
- Percentage of correct responses
- Percentage of confusions
- Considered per consonant individually



- ➢ No effects of lexicon, context, or syntax
- ➢ Detailed measurement of low-level speech perception

Experimental approach

Large variability observed in responses of NH listeners due to...?

- ***Source-induced* variability**
  1. *Speech-induced* variability
     (across talkers / within talkers)
  2. *Noise-induced* variability

- ***Investigated based on***

  Different speech tokens for same CV
  (different talkers / same talker)

  Same speech token, different frozen noise tokens

- ***Receiver-related* variability**
  1. *Across-listener* variability
  2. *Within-listener* variability
     (internal noise)

- ***Investigated based on***

  Physically identical stimuli
  (different listeners)

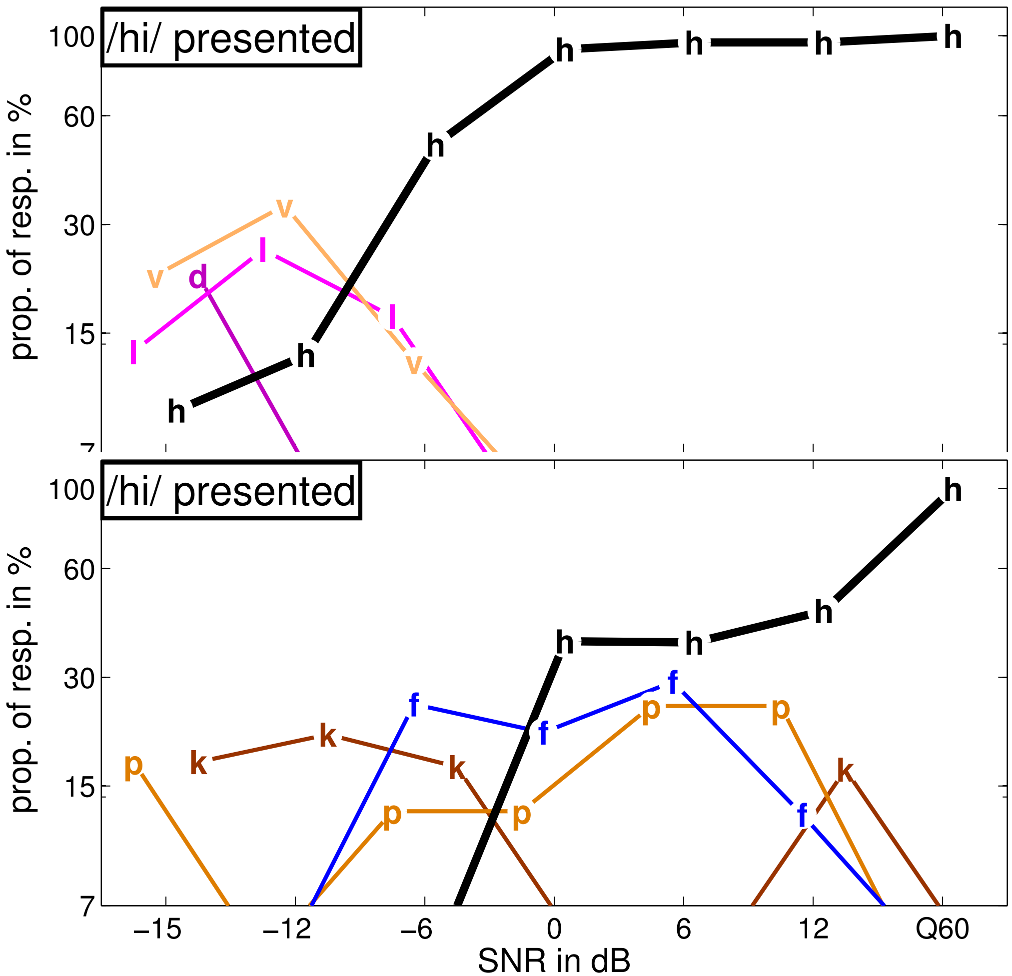
  Physically identical stimuli
  (given listener, test versus retest)

Experimental results – *Speech-induced variability*

Across talkers: /hi/

Talker A

Talker B

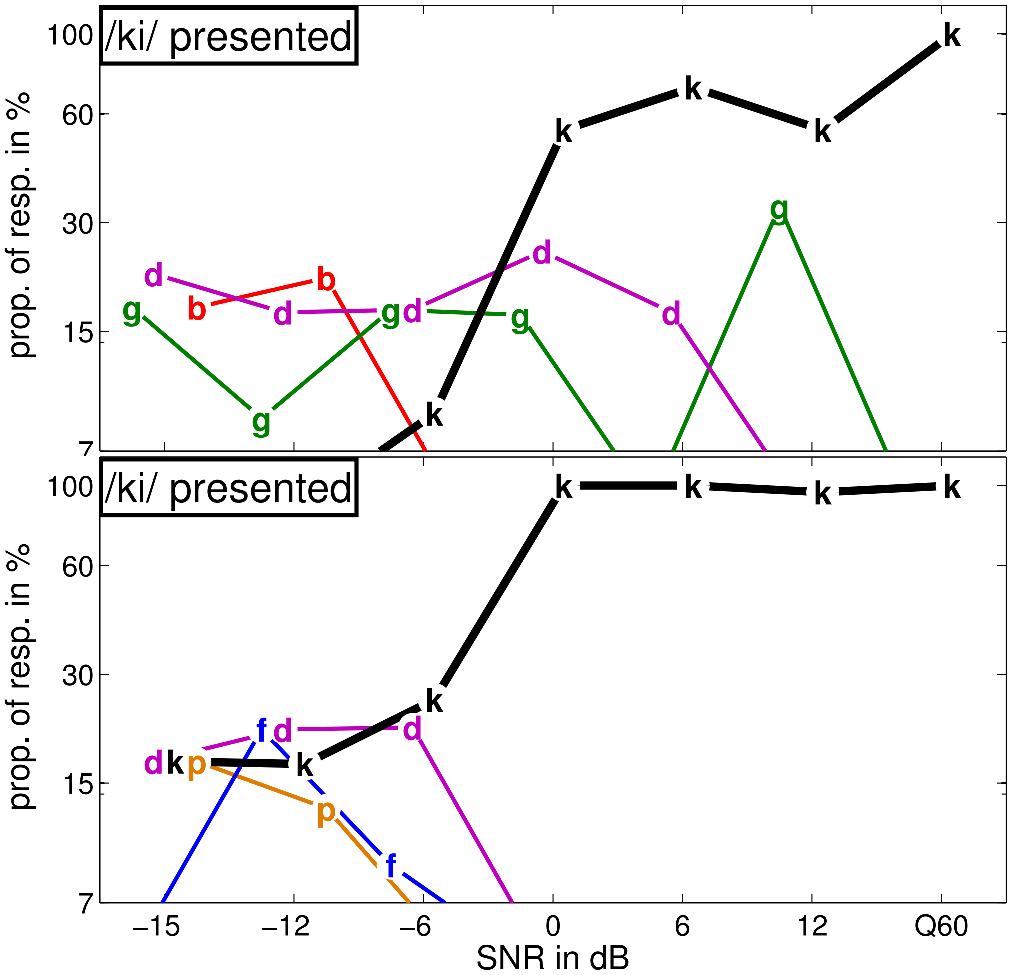


Experimental results – *Speech-induced variability*

Within talkers: /ki/

Talker A, Recording 1

Talker A, Recording 2

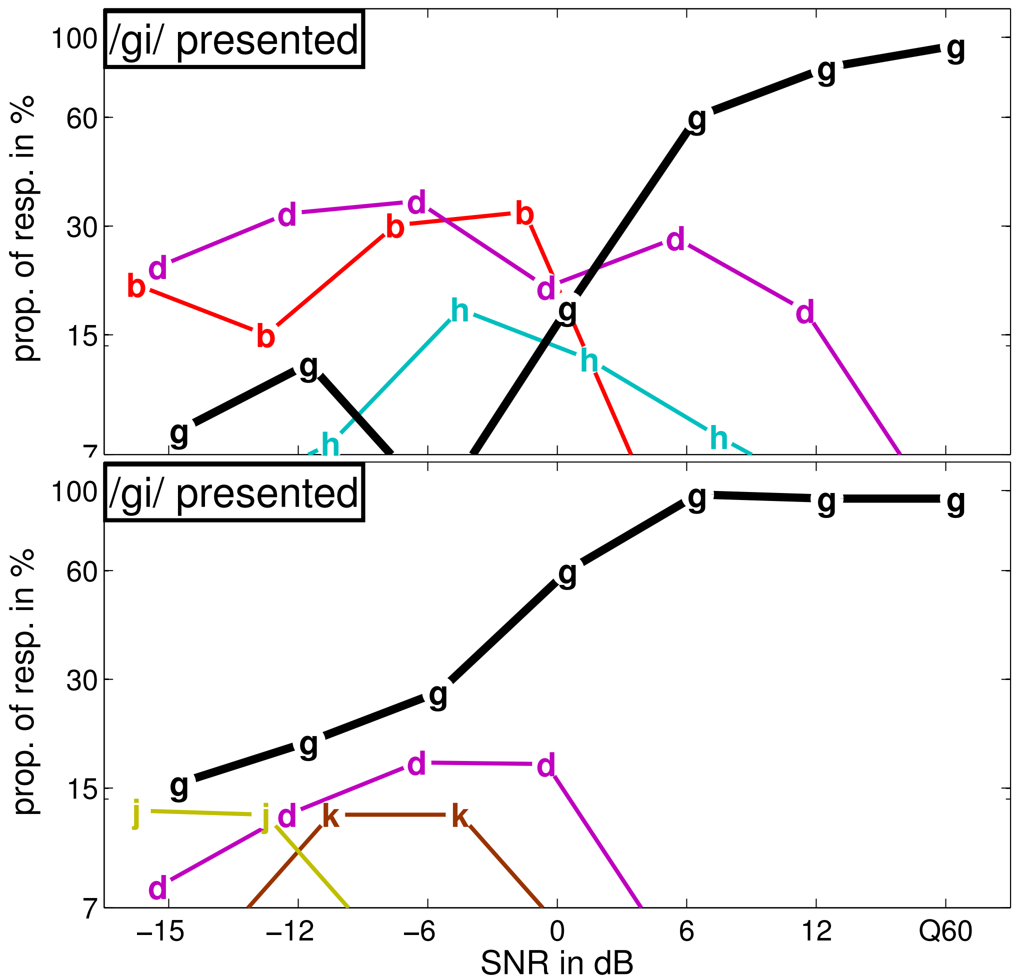


Experimental results – *Noise-induced variability*

Speech token /gi/ mixed with…

Frozen noise A

Frozen noise B

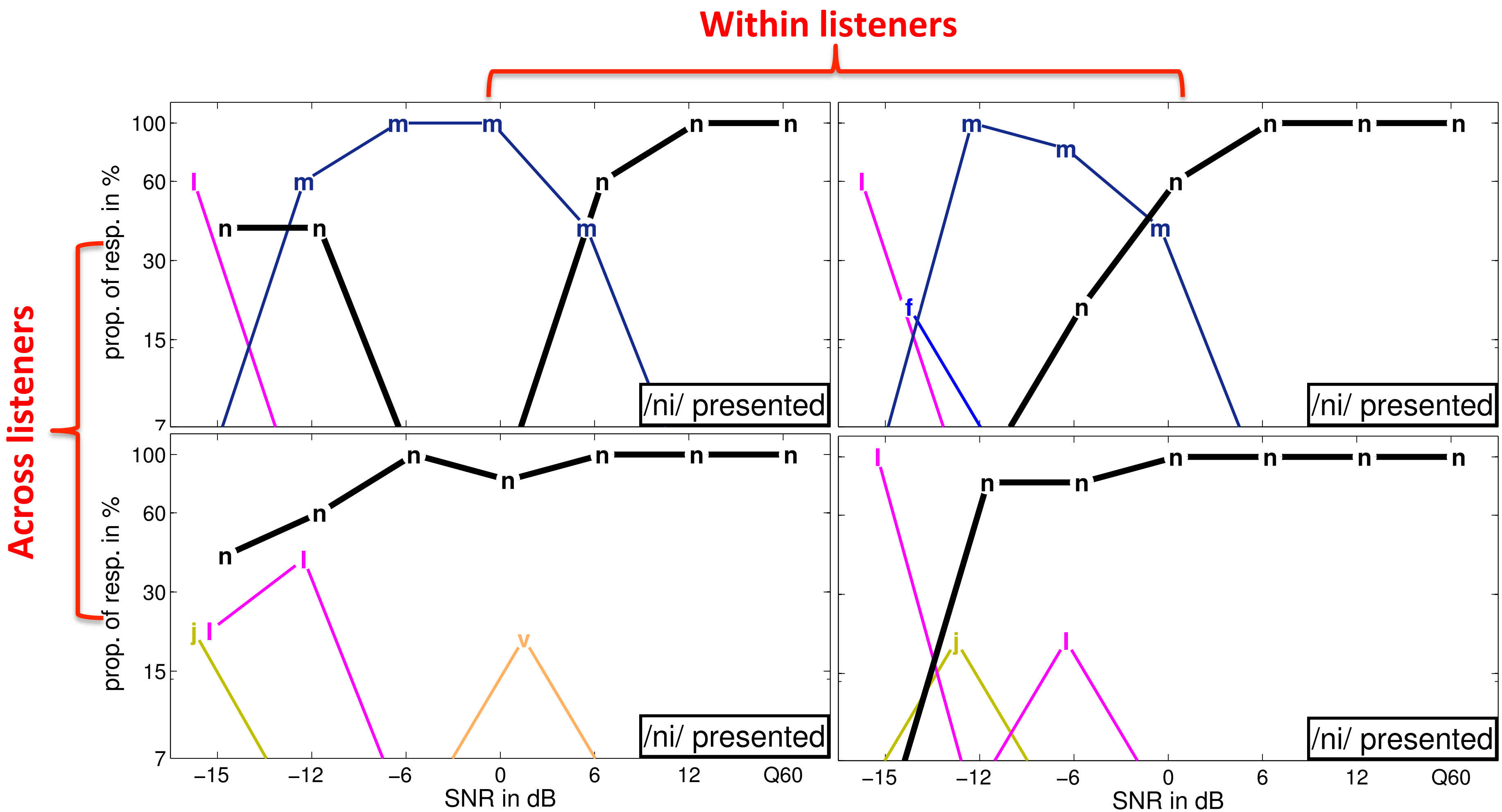


CAHR Centre for Applied Hearing Research

SpeechiNoise

Experimental results – *Receiver-related variability*

one specific token of /ni/ + frozen noise



Analysis – *Perceptual distance calculation*

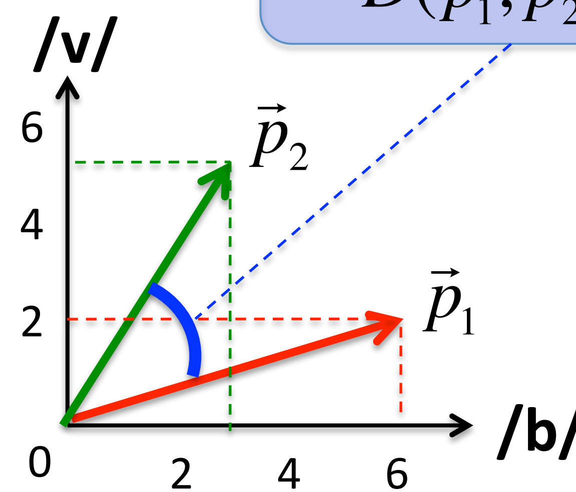
Basic principle

Angular distance
$$D(\vec{p}_1, \vec{p}_2)$$

responded

presented

	/b/	/v/	
s1	6	2	$\vec{p}_1$
s2	3	5	$\vec{p}_2$



Can be used to compare any pair of responses of arbitrary dimensionality!

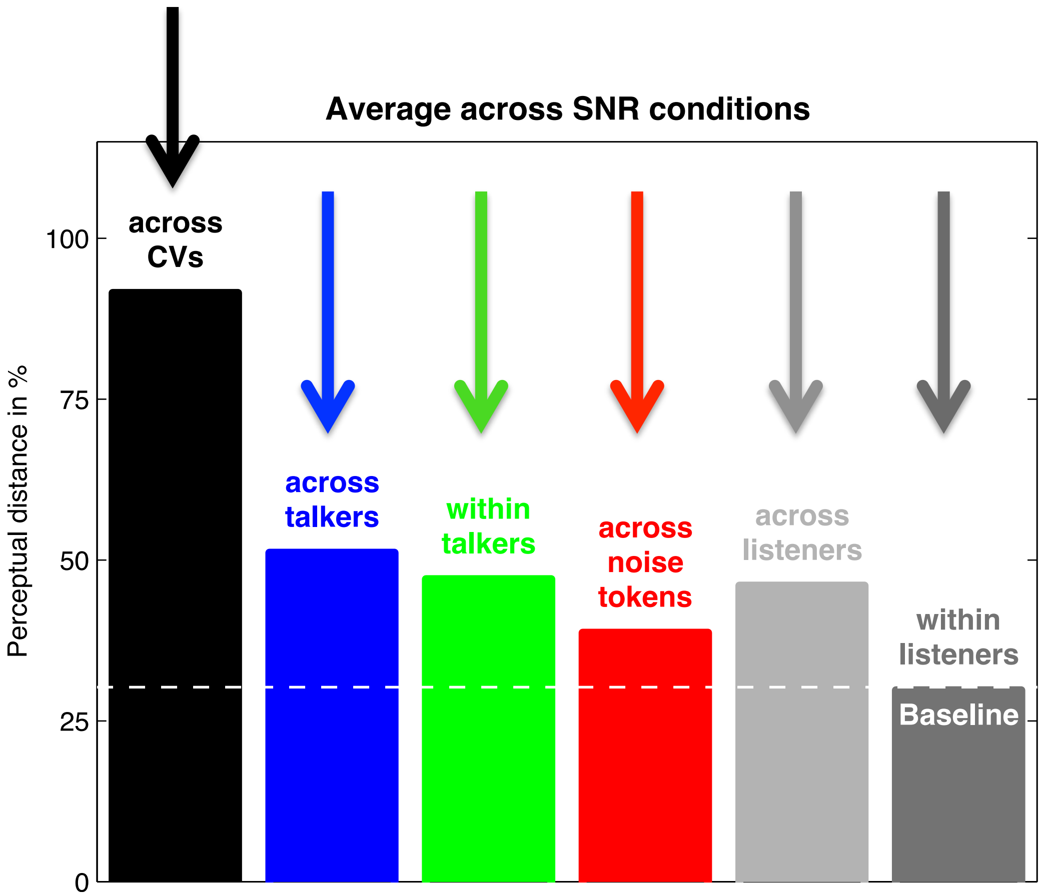
Analysis – *Perceptual distance calculation*

- 15-dimensional vector space (15 consonants as response alternatives)
- Calculated based on individual listener responses:
  - **Across CVs** (reference for maximal distance)
  - ***Source-induced* variability:**
    - **Across talkers**
    - **Within talkers**
    - **Across frozen-noise tokens**

    for stimuli of the same phonetic identity
  - ***Receiver-related* variability:**
    - **Across listeners**
    - **Within listeners** (baseline for minimal distance)

    for physically identical stimuli

Analysis – *Perceptual distance*



Implications for experimental design

I. “*Global*“ experiment:

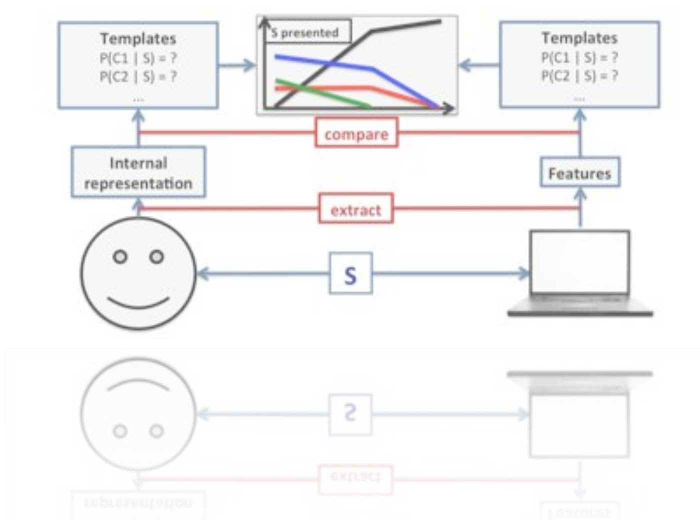
- Avoid bias due to individual speech tokens, noise tokens, and listeners
- Present many speech tokens per consonant in random noise to many listeners
- Average across speech tokens and listeners

II. “*Detailed*“ experiment (investigating consonant cues):

- Evaluate responses for each speech token separately
- Use unique combinations of speech tokens and noise tokens (across SNRs)
- Evaluate responses for each listener individually

Part II.

Modeling consonant perception



CAHR Centre for Applied Hearing Research

SpeechiNoise

Motivation

Macroscopic speech intelligibility models

- Prediction of *average recognition (SRTs)*

Audibility (classical)

Analysis of speech-to-noise energy in spectral bands

Articulation **I**ndex – **AI**

Speech **I**ntelligibility **I**ndex – **SII**

Modulation masking (more recent)

Depth and rapidity of the amplitude fluctuations in the noisy speech envelope

Speech **T**ransmission **I**ndex – **STI**

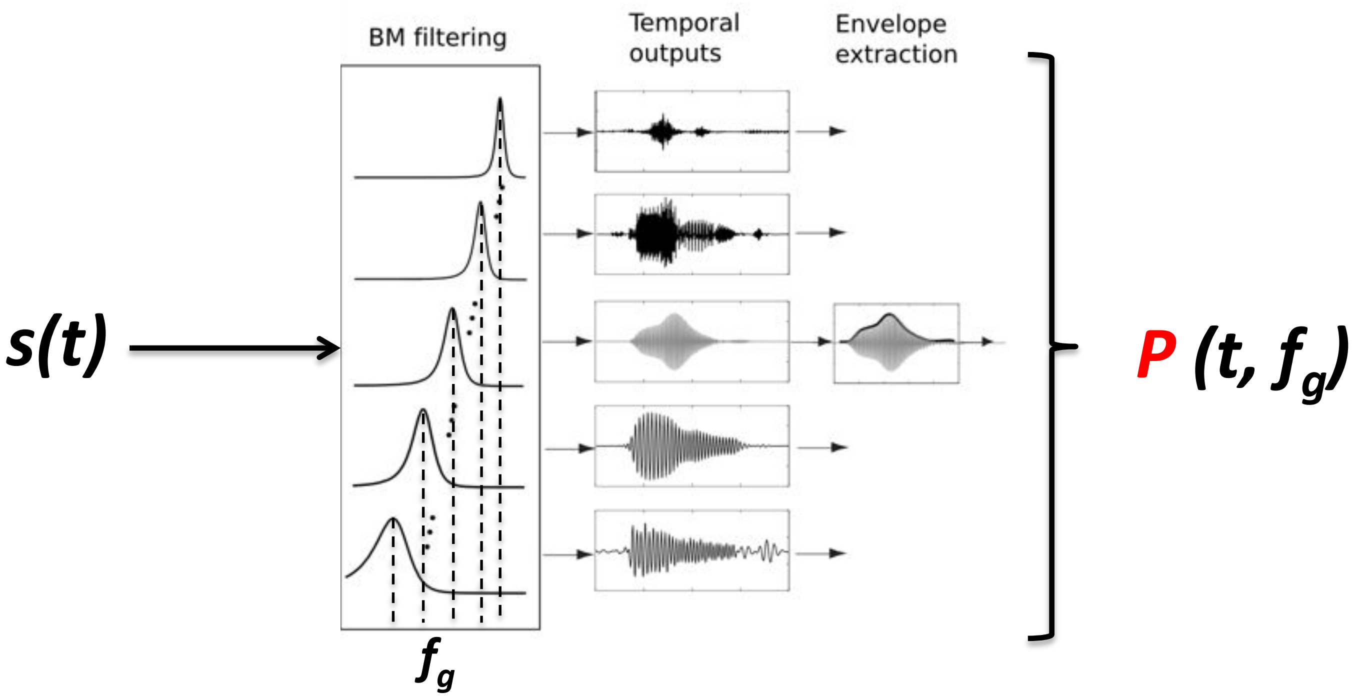
speech-based **E**nvelope **P**ower **S**pectrum **M**odel – **sEPSM**

Microscopic consonant perception modeling

- Prediction of *consonant-specific recognition and confusions*

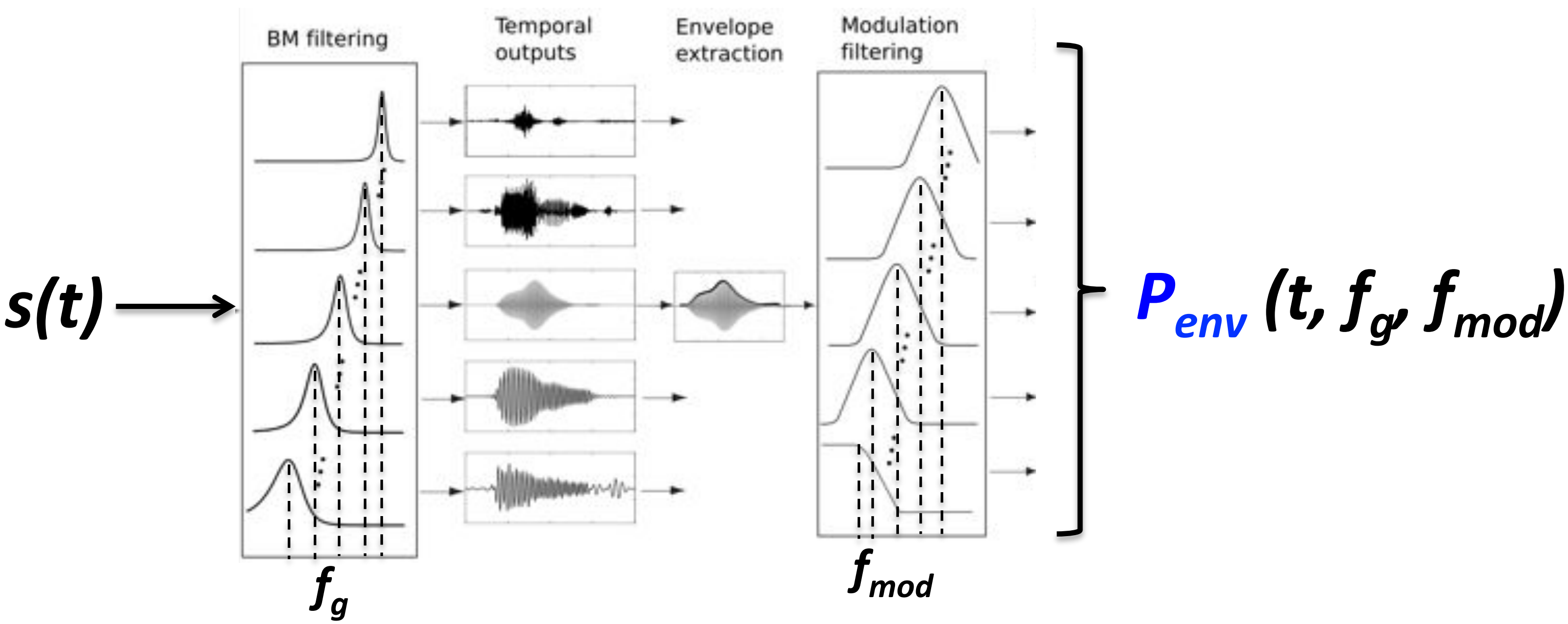
Which macroscopic modeling concept is more suitable for consonant perception modeling – audibility or modulation masking?

Model description – Audibility front end (AI, SII)



22 gammatone filters, logarithmically spaced between 63 Hz and 8 kHz

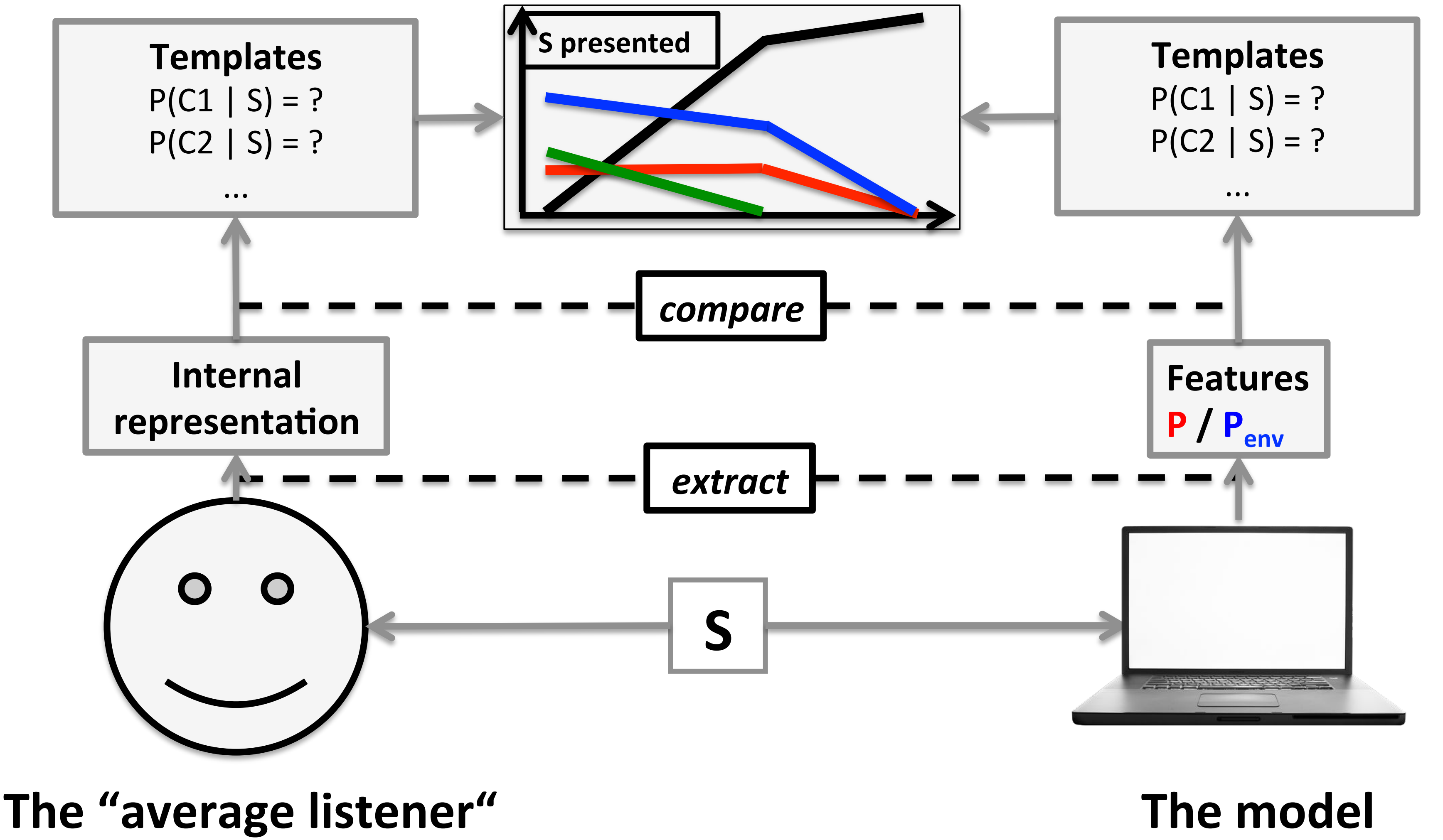
Model description – Modulation front end (STI, sEPSM)



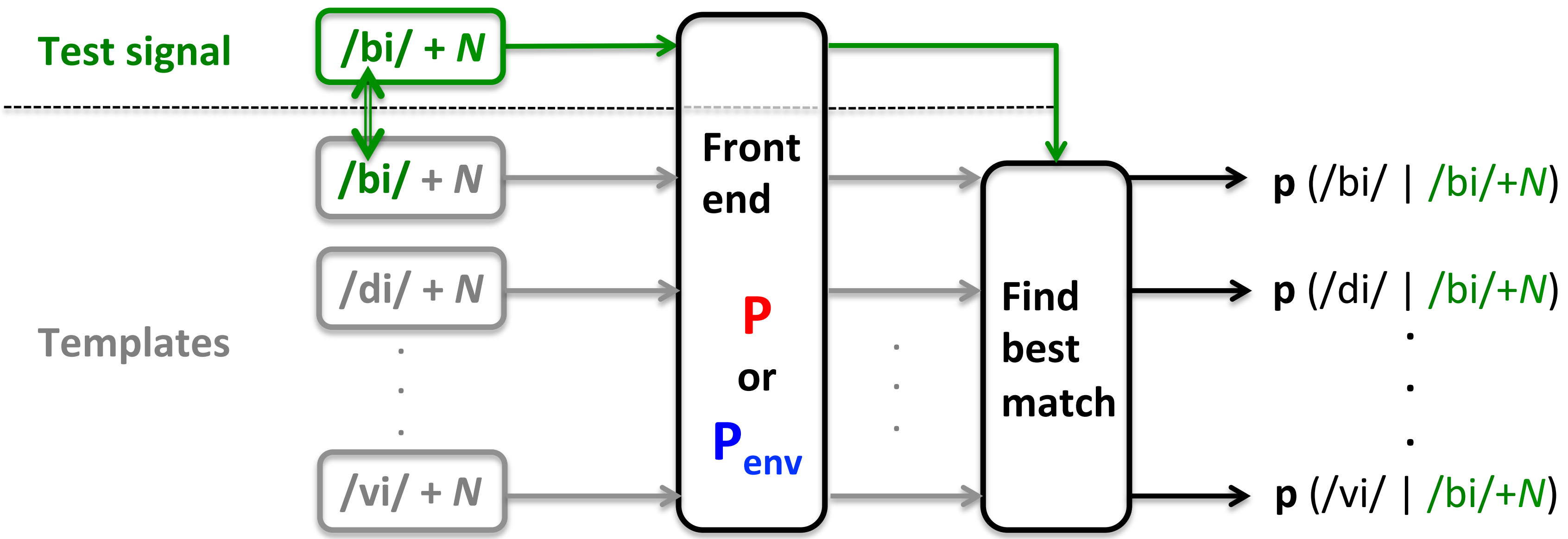
22 gammatone filters, logarithmically spaced between 63 Hz and 8 kHz

9 modulation filters, logarithmically spaced between 1 Hz and 256 Hz

Model description – Basic approach for back end



Model description – Overall modeling scheme



Calculated 10 times (noise always newly generated)

<u>Model knows:</u>

- Clean test speech token
- Speaker
- Noise type (white noise)
- Signal-to-noise ratio

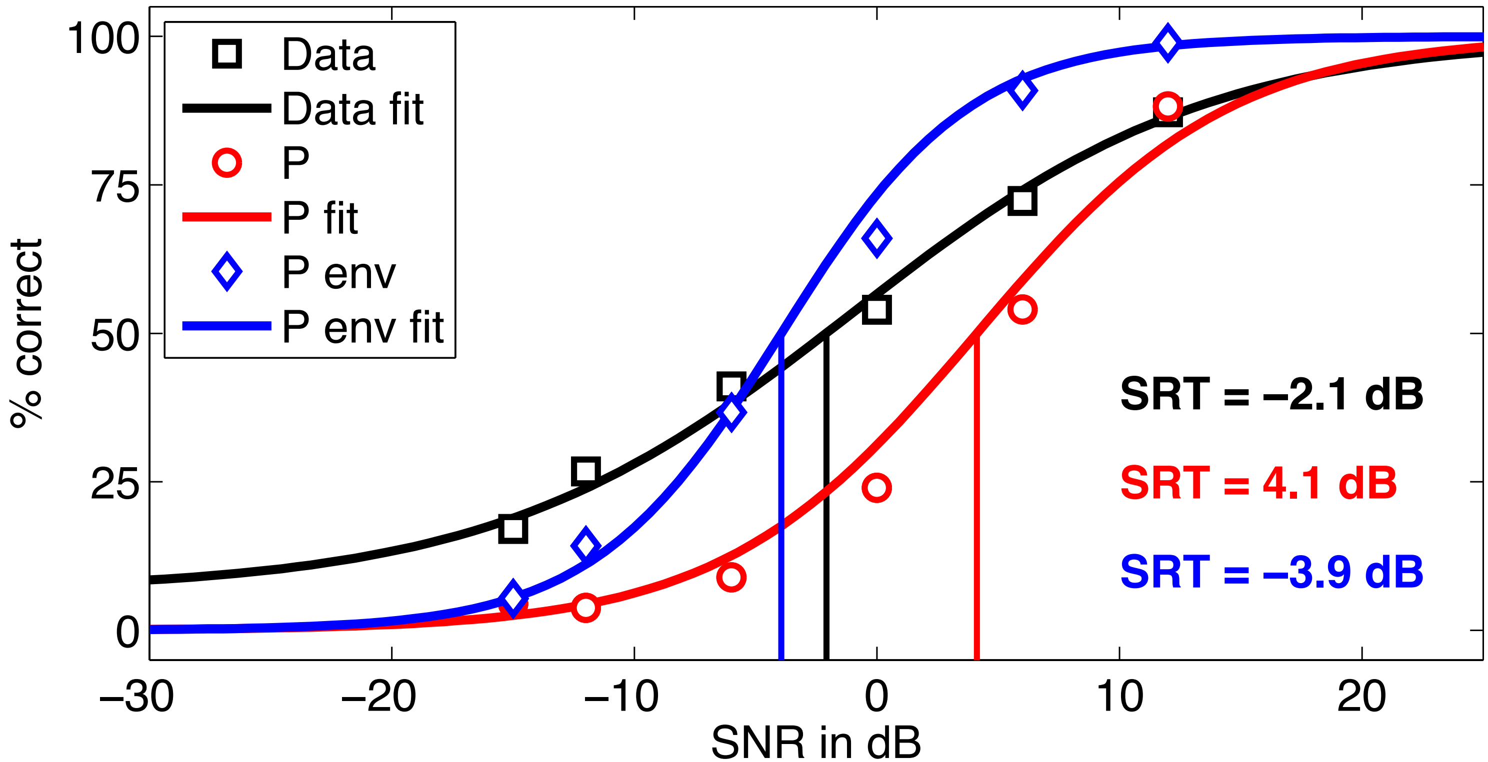
<u>Model doesn‘t know:</u>

- Noise waveform
- Articulatory variability within CVs (uses only one template for each CV)

CaHR Centre for Applied Hearing Research

Speech in Noise

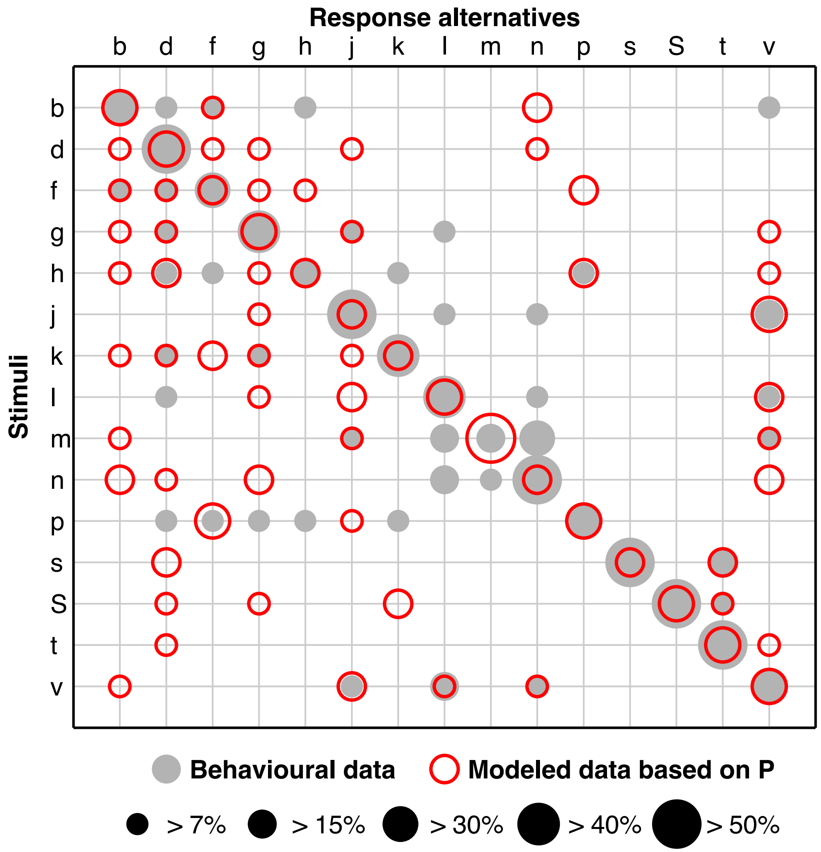
Modeling results – *Average consonant recognition*



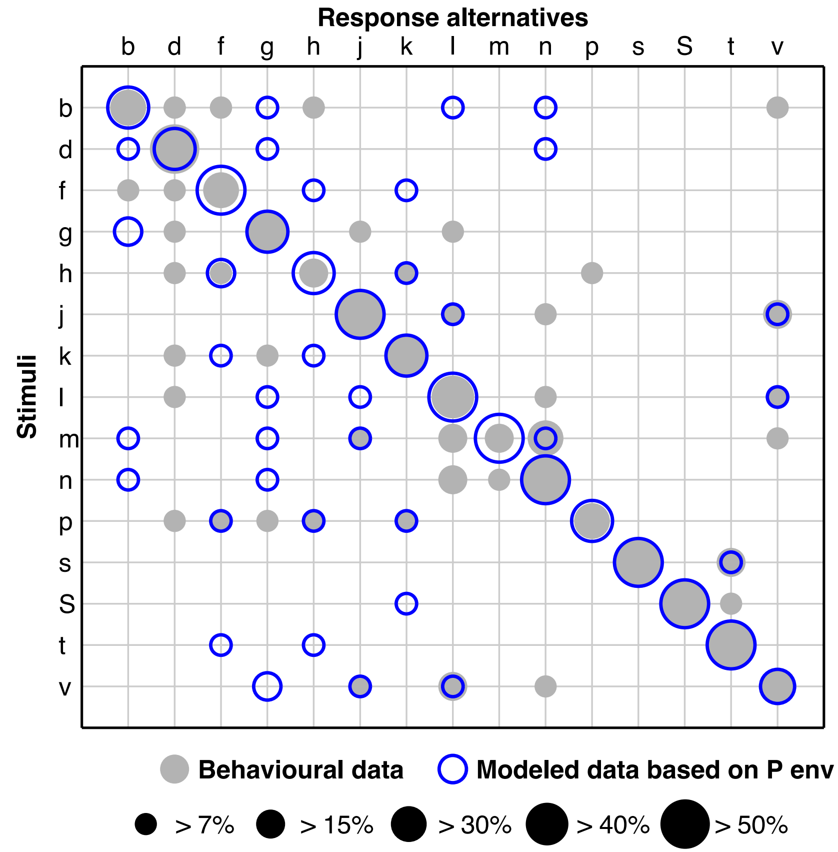
- **P:** far less sensitive than listeners
- **P_{env}:** slightly more sensitive than listeners
- Modeled psychometric function slopes too steep (both front ends)

Model predictions – *Recognition and confusions*

P - based predictions



P_{env} - based predictions



- **P:** underestimation of recognition / confusions only partly captured
- **P_{env}:** good prediction of recognition / confusions only partly captured

Part II. Summary

- Modulation front end seems to capture relevant features for consonant perception better than audibility front end
- Well-predicted using modulation front end:
  - ✓ Grand average SRT
  - ✓ Consonant-specific recognition
- Room for improvement - modulation front end:
  - ❖ Slopes of predicted psychometric functions too steep
  - ❖ Confusion predictions unsatisfactory

Future work

- Comparison of perceptual distance and auditory-feature distance
  - For spectro-temporal representation
  - For modulation-domain representation
- Inclusion of articulatory variability in modeling framework
- Inclusion of internal-noise term and language-specific bias term in modeling framework

Acknowledgements

Torsten Dau

Søren Jørgensen

Hearing Systems Group @ DTU

Thank you for your attention!