Models and Modes of Audiovisual integration

Andersen, Tobias

Publication date:
2015

Document Version
Peer reviewed version

Models and Modes of Audiovisual integration

Tobias Andersen
Technical University of Denmark

Cognitive Systems

DTU Compute
Department of Applied Mathematics and Computer Science

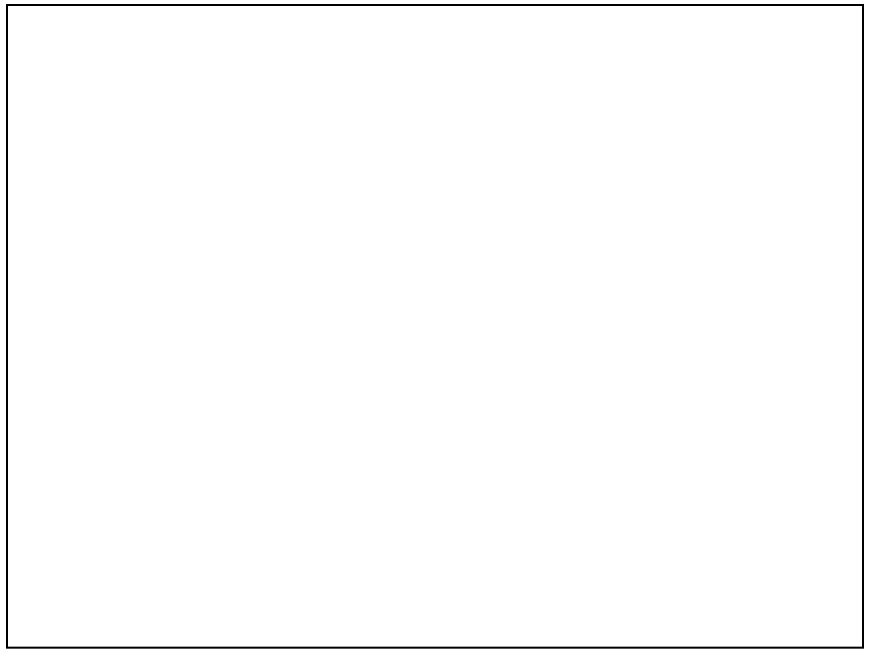
Outline

- Categorical audiovisual perception
  - What's so special?
    - Categorical, non-linear changes
      - The McGurk effect
      - Flashes and beeps

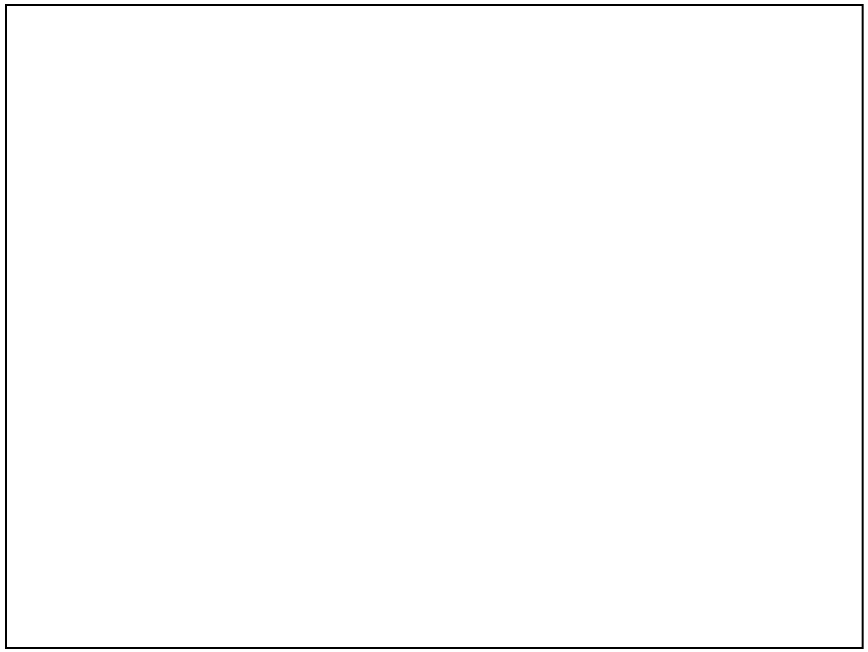
McGurk

McGurk and MacDonald, Nature, 1976

Illusory flashes and beeps



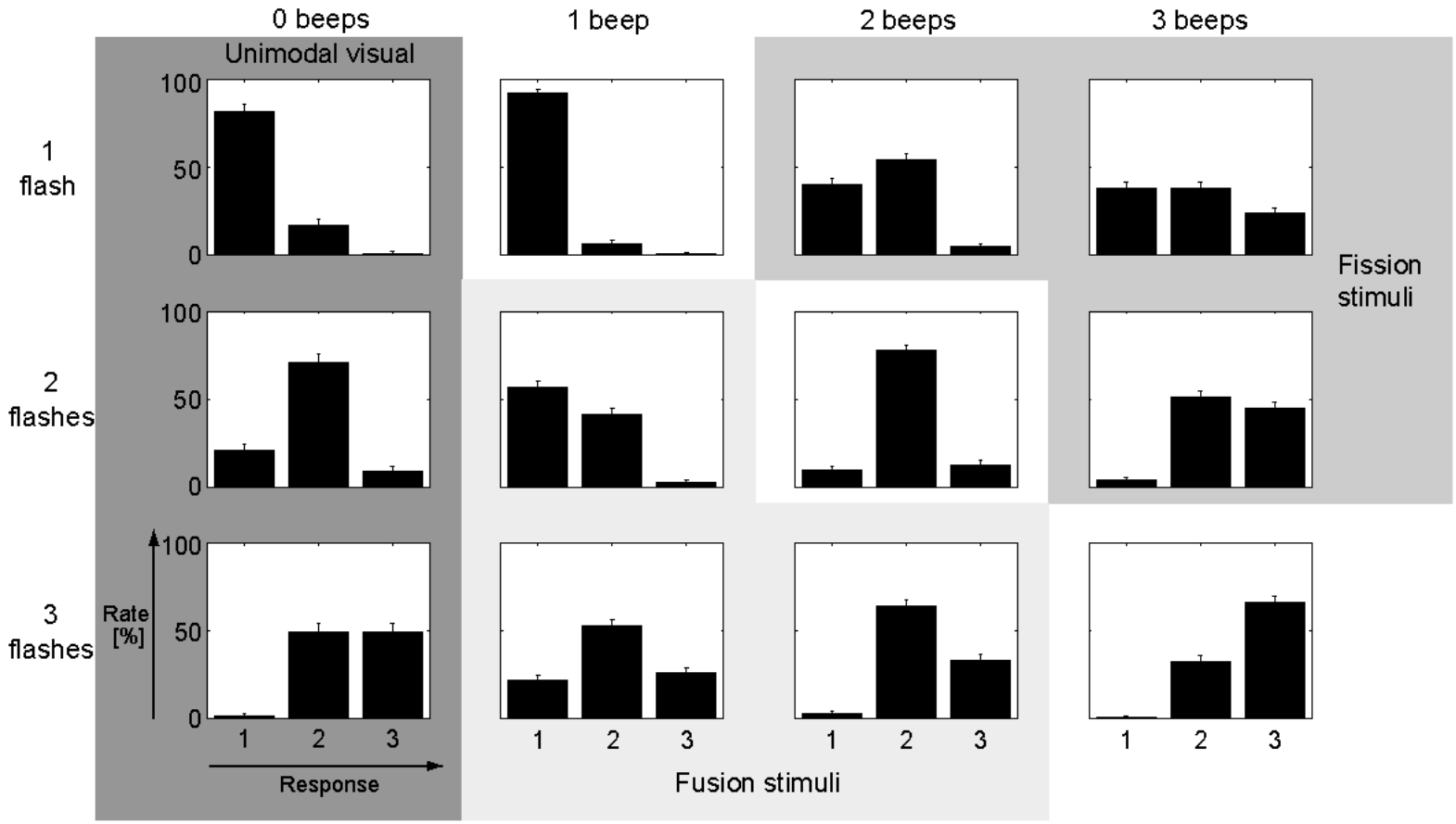
2



1

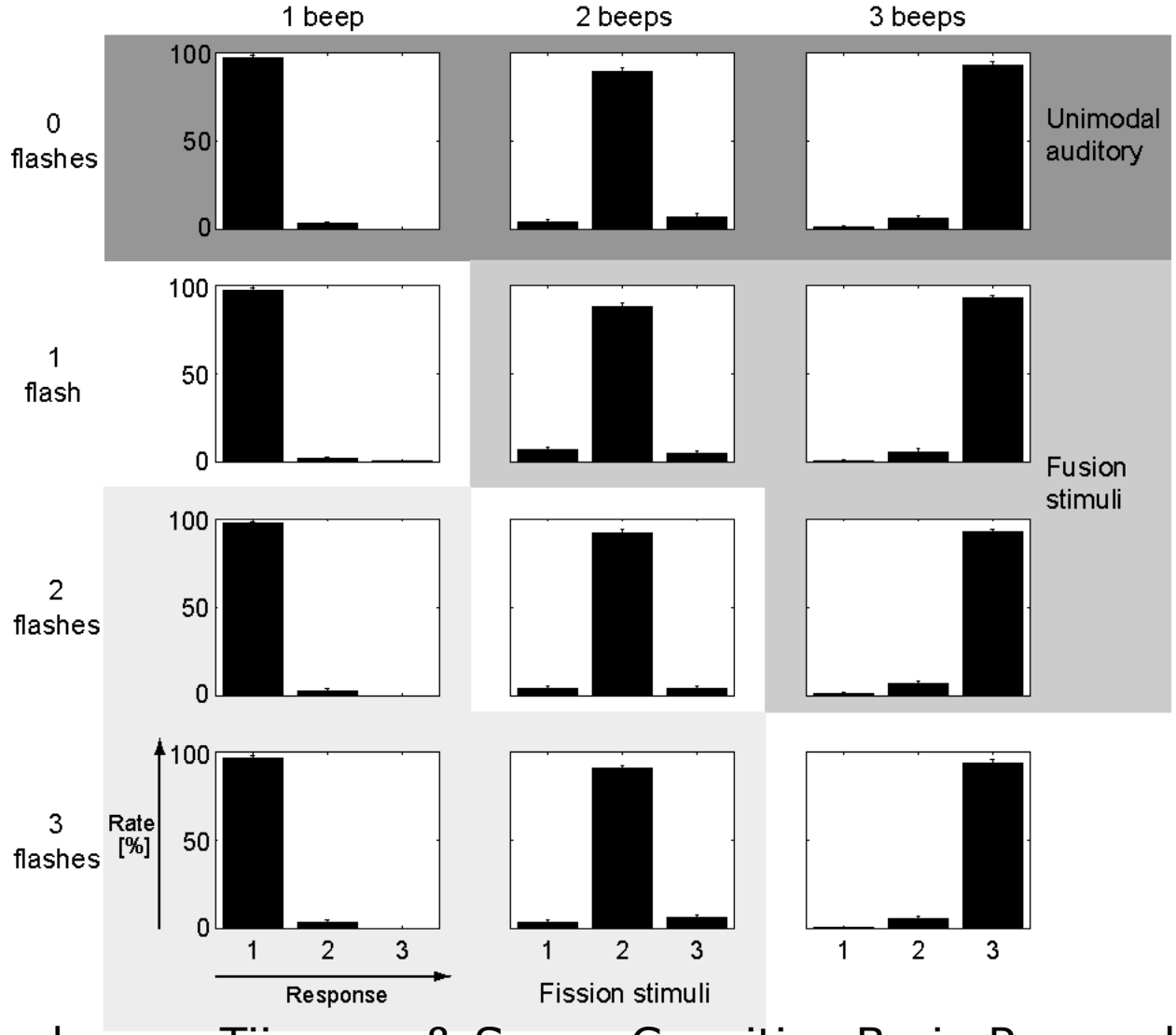
Shams, Kamitani & Shimojo, Nature, 2000

Illusory flashes and beeps



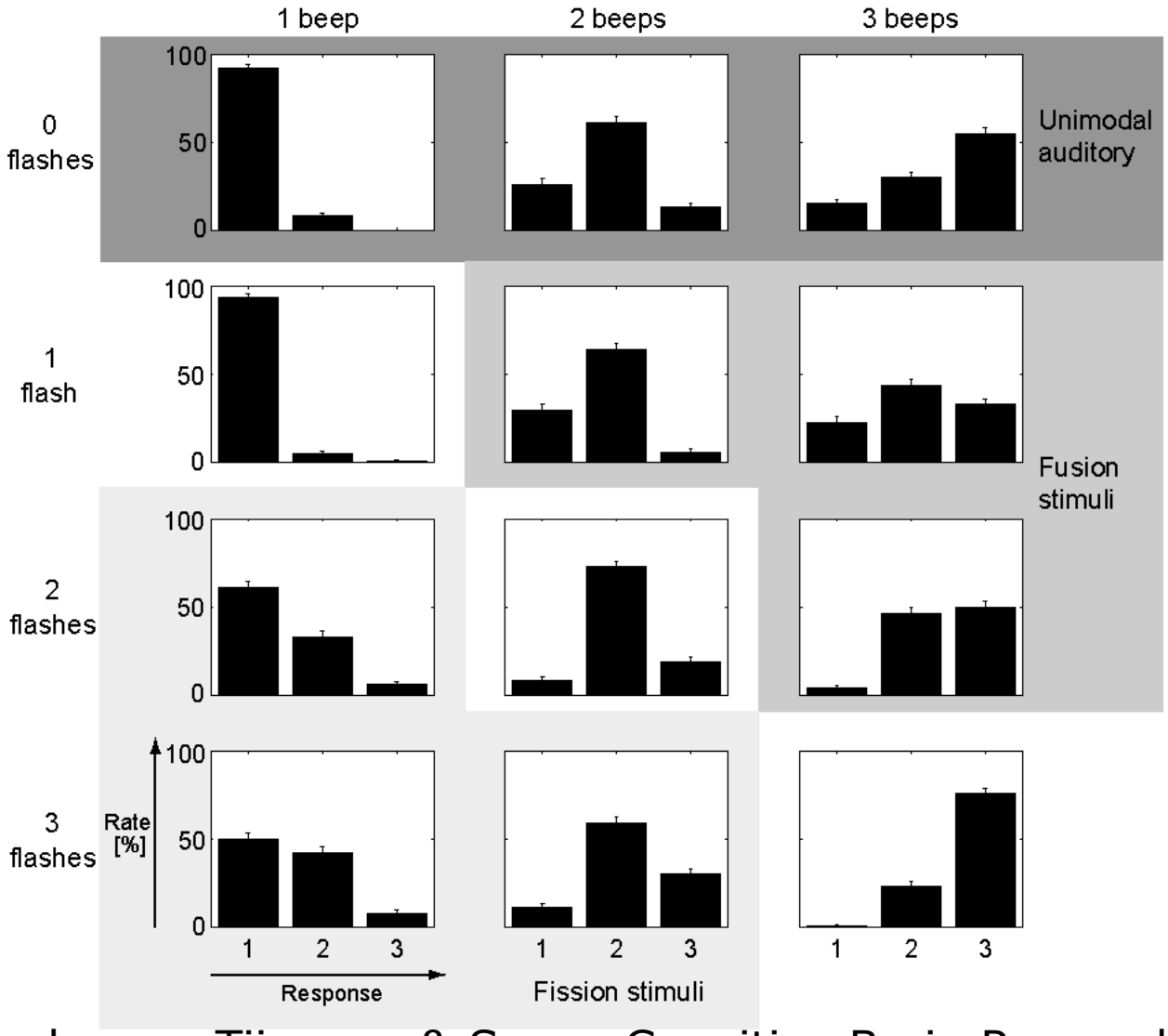
Andersen, Tiippana & Sams, Cognitive Brain Research, 2004

Illusory flashes and beeps



Andersen, Tiippana & Sams, Cognitive Brain Research, 2004

Illusory flashes and beeps

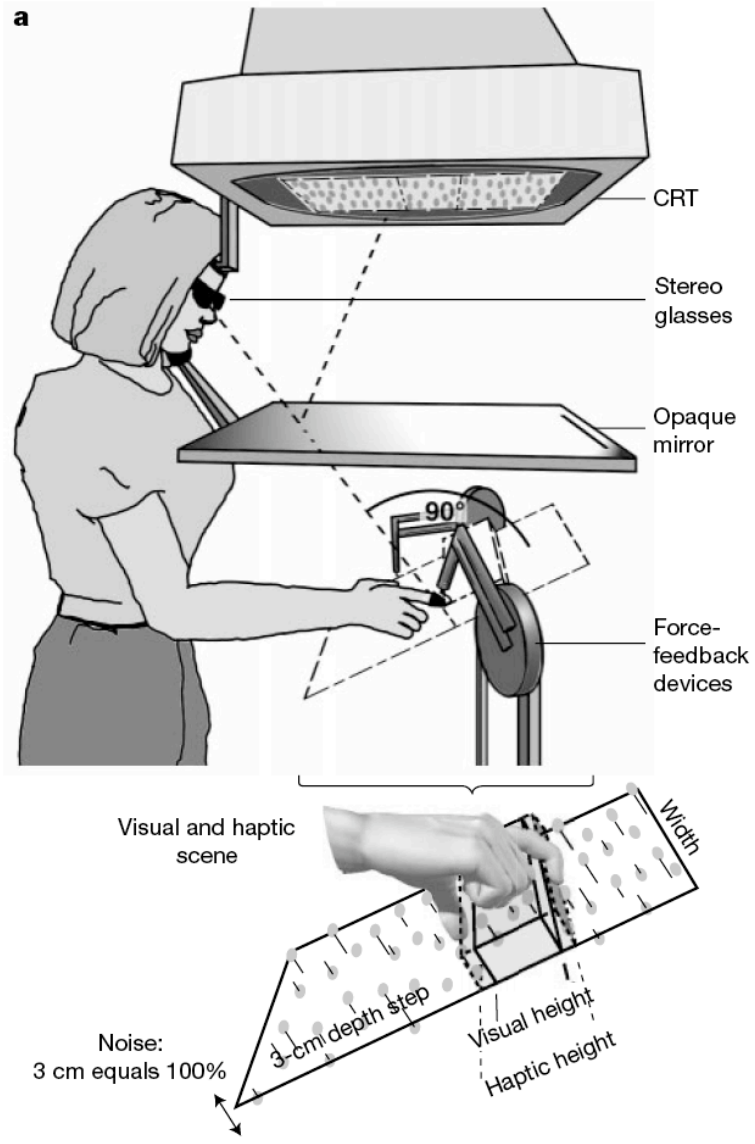


Andersen, Tiippana & Sams, Cognitive Brain Research, 2004

Illusory flashes and beeps

- Governing principles
  - Information reliability
    - The strength of cross-modal influence depended on sound level
  - Modality appropriateness
    - The sound had to be at threshold to be influenced
    - The flashes was influenced also well above threshold
  - Directed attention
    - Possible to count either flashes or beeps

Maximum Likelihood Estimation (MLE)



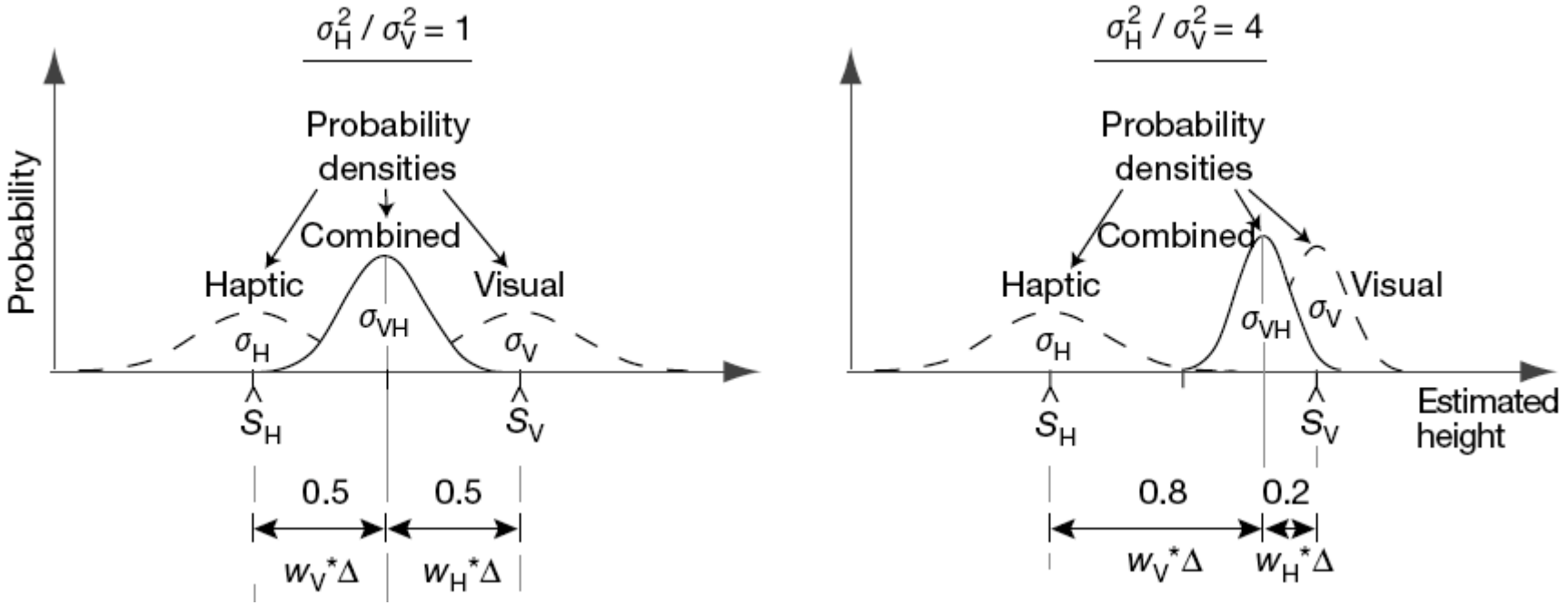
- Height can be estimated from
  - sight
  - proprioception
- Independent stimuli can be created with
  - Force feedback device
  - mirrored stereo display

From Ernst and Banks, Nature, 2002

Multisensory integration

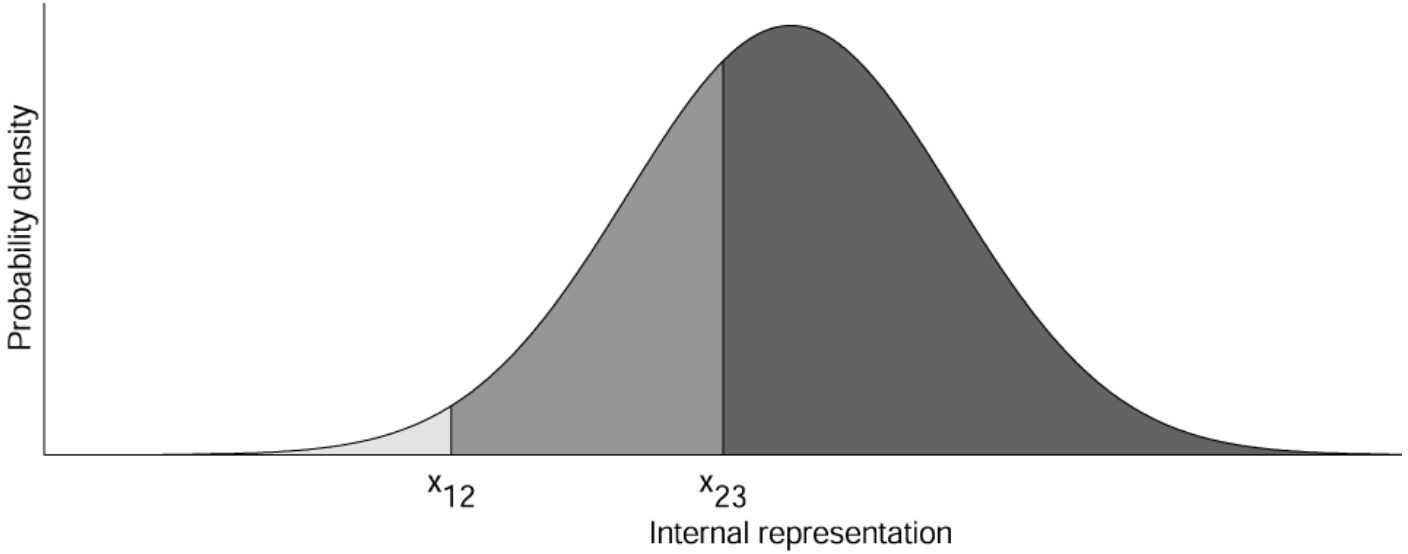
- Maximum likelihood rule nice and simple for Gaussian noise

$$P(S \mid H,V) = \frac{P(S \mid H)P(S \mid V)}{\int_{S'} P(S' \mid H)P(S \mid V)dS'}$$

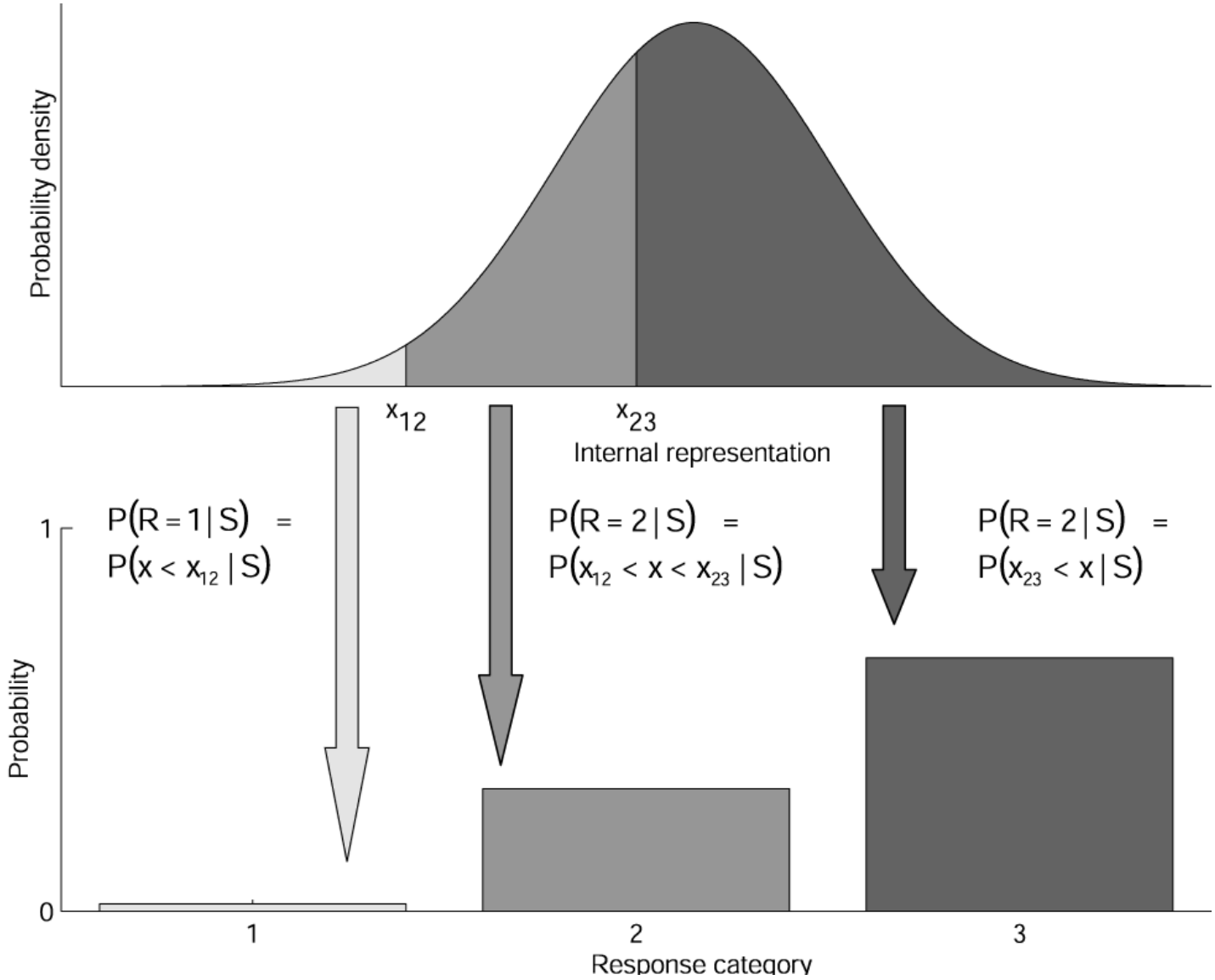


From Ernst and Banks, Nature, 2002

Early MLE - Classification



Early MLE - Classification



Late MLE (a.k.a. FLMP)

$$P(R_i \mid A,V) = \frac{P(R_i \mid A) \times P(R_i \mid V)}{\sum_{j=1}^{N} P(R_j \mid A) \times P(R_j \mid V)}$$

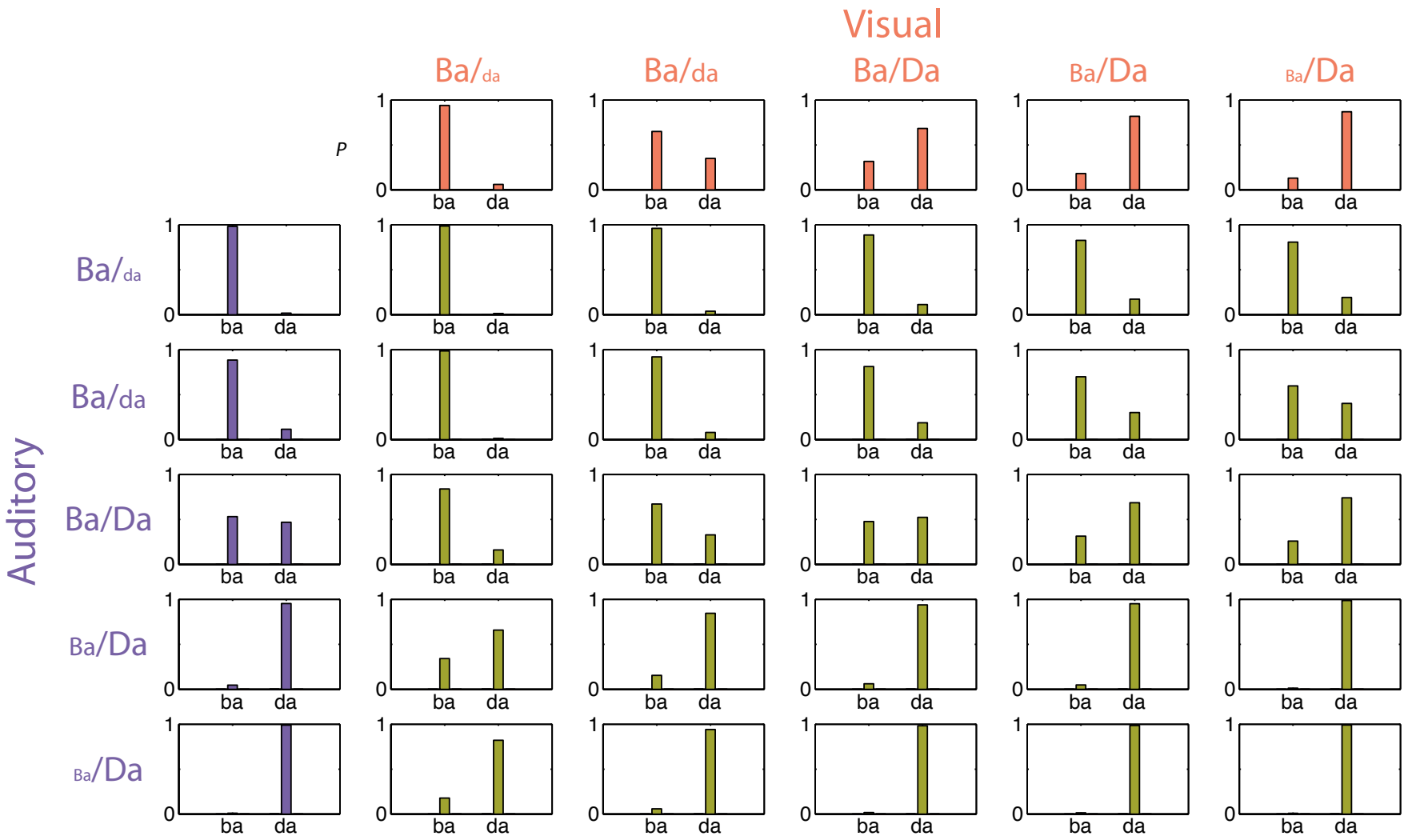
- Late integration (occurs after categorization)
- Only parameters: Unimodal response probabilities
- Generally good fits

Early vs. Late MLE

- Applied to illusory flashes and beeps
  - Early MLI generally has fewer free parameters
  - Early MLI fits our data better
  - Early MLI parameterizes reliability
    - a more parsimonious model
  - Early MLI orders responses / stimuli
    - 1 flash < 2 flashes < 3 flashes

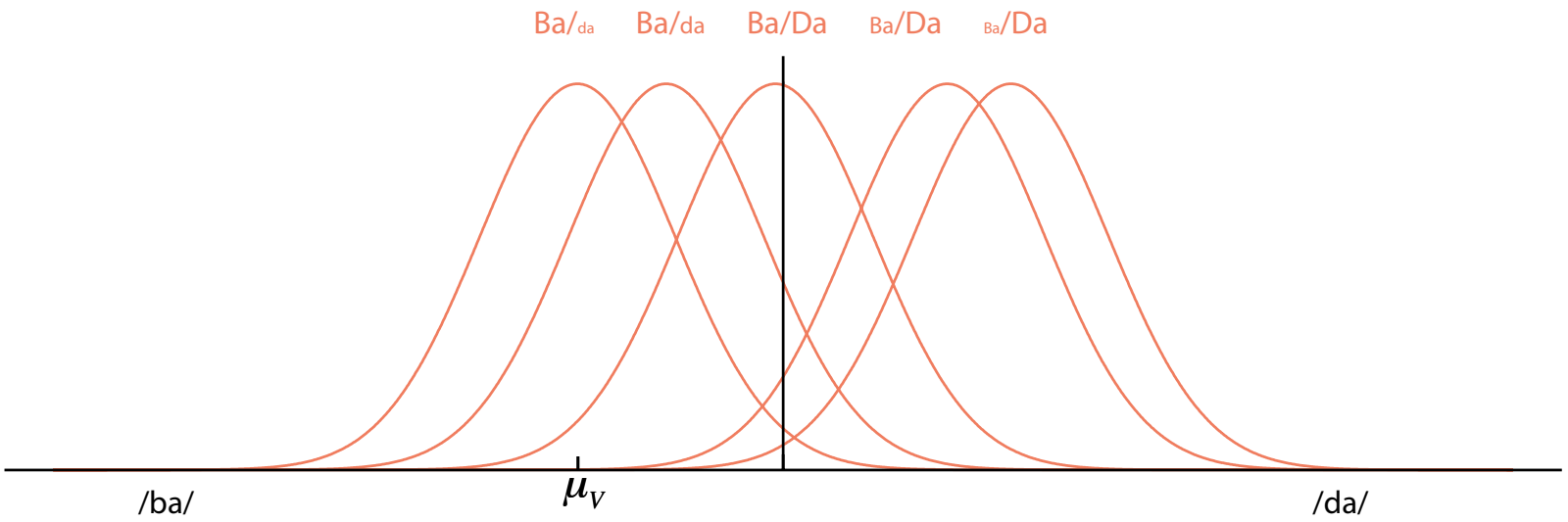
Andersen, Tiippana & Sams, 2005

The UCSC corpus



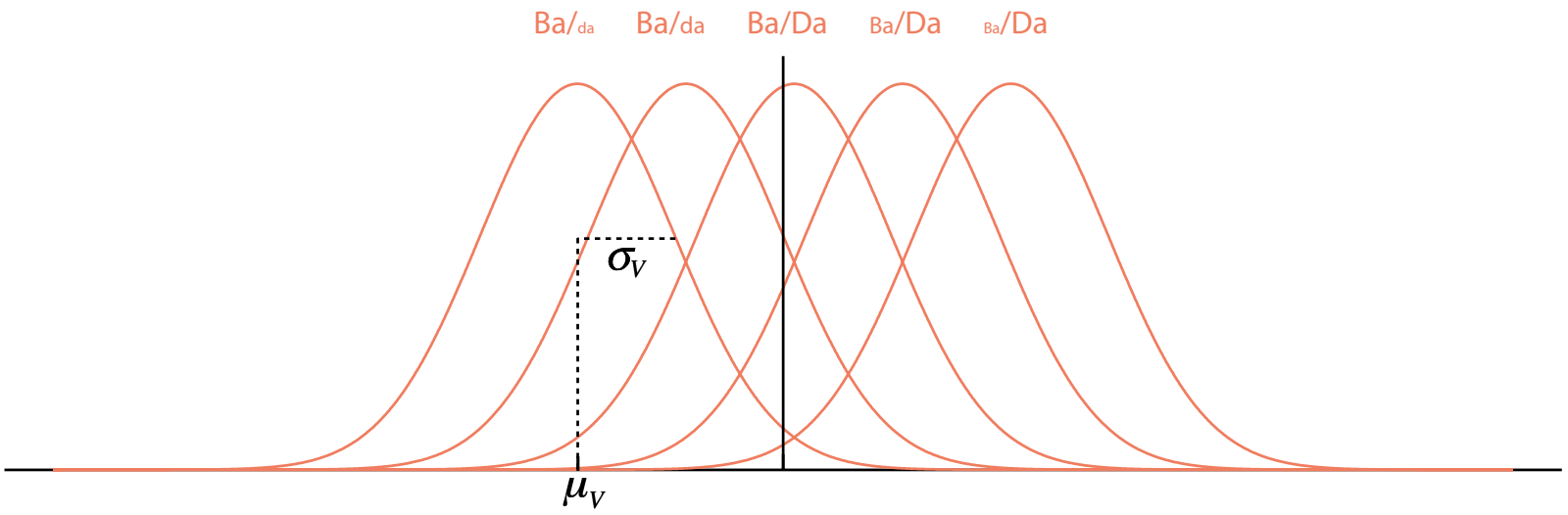
Massaro (1998)

Early MLE applied to the UCSC corpus



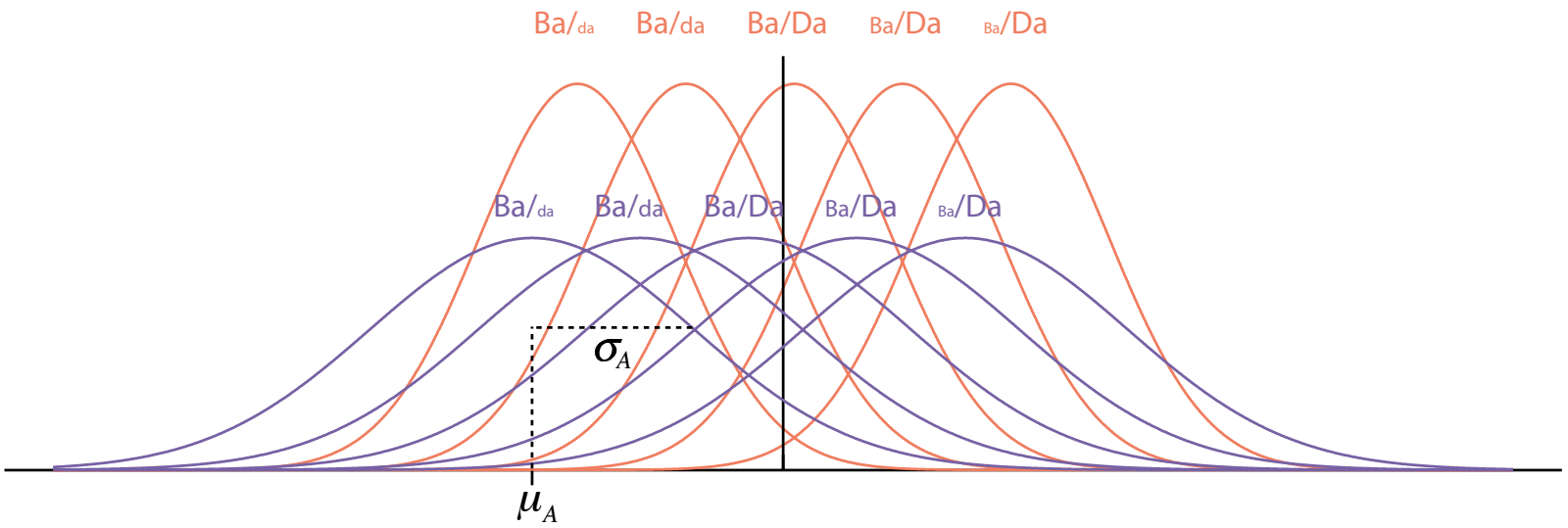
Andersen (forthcoming), JASA

Early MLE applied to the UCSC corpus



Andersen (forthcoming), JASA

Early MLE applied to the UCSC corpus



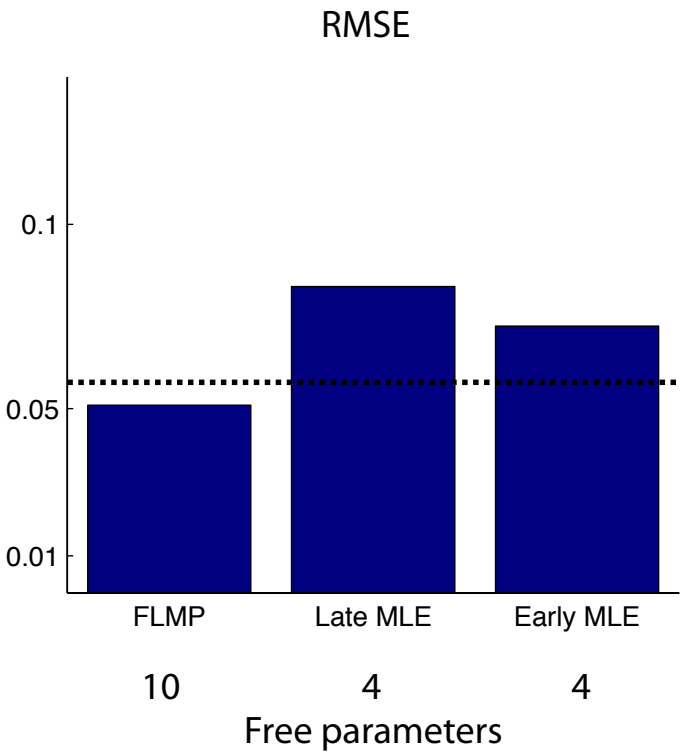
Andersen (forthcoming), JASA

Early MLE applied to the UCSC corpus

- Linear spacing constraint
  - Reflect the experimental design
  - Reduces model complexity (10 -> 4 free parameters)
  - Allows Early MLE

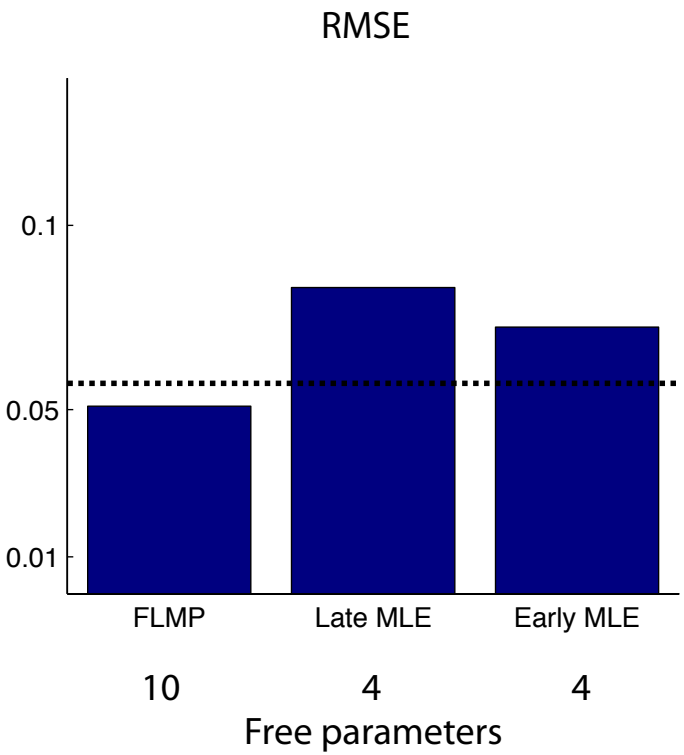
Andersen (forthcoming), JASA

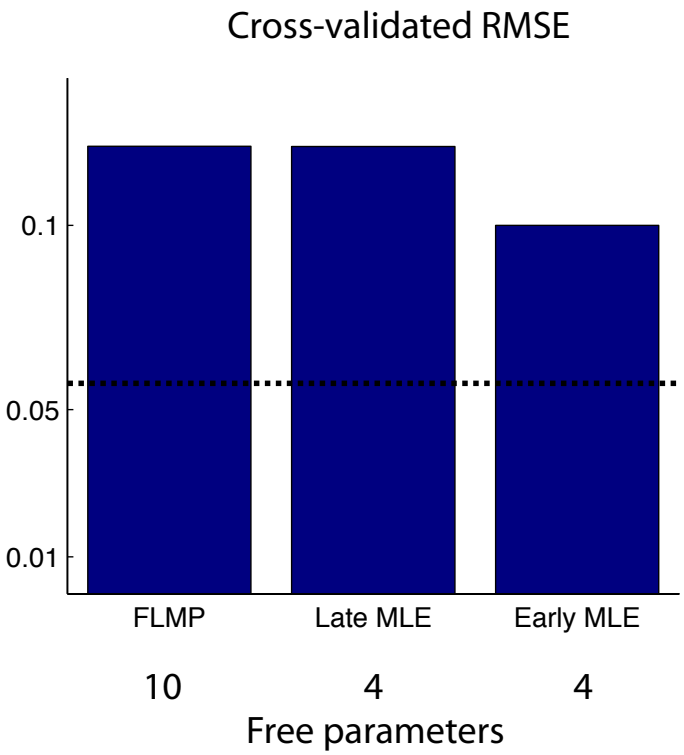
Results



Andersen (forthcoming), JASA

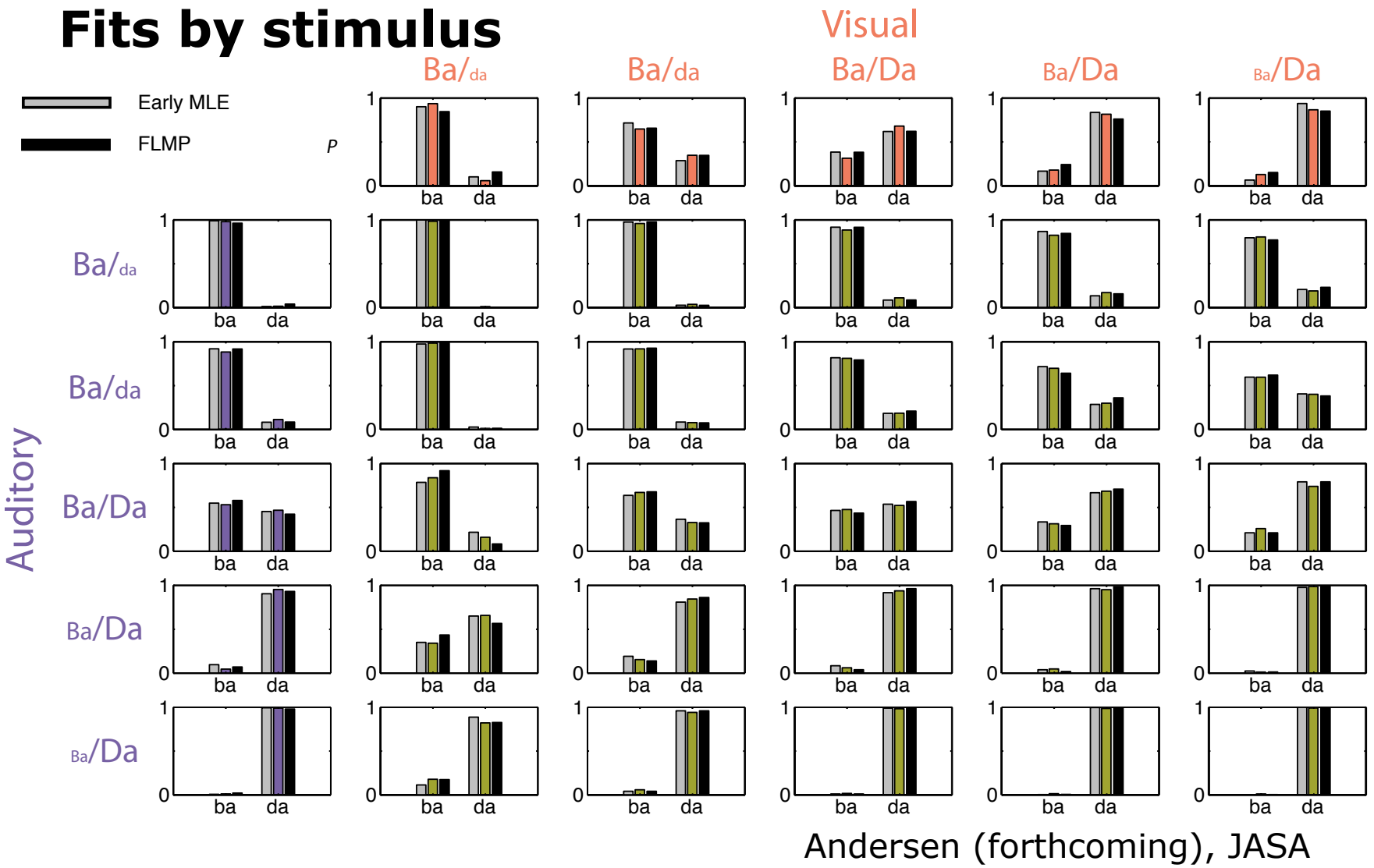
Results





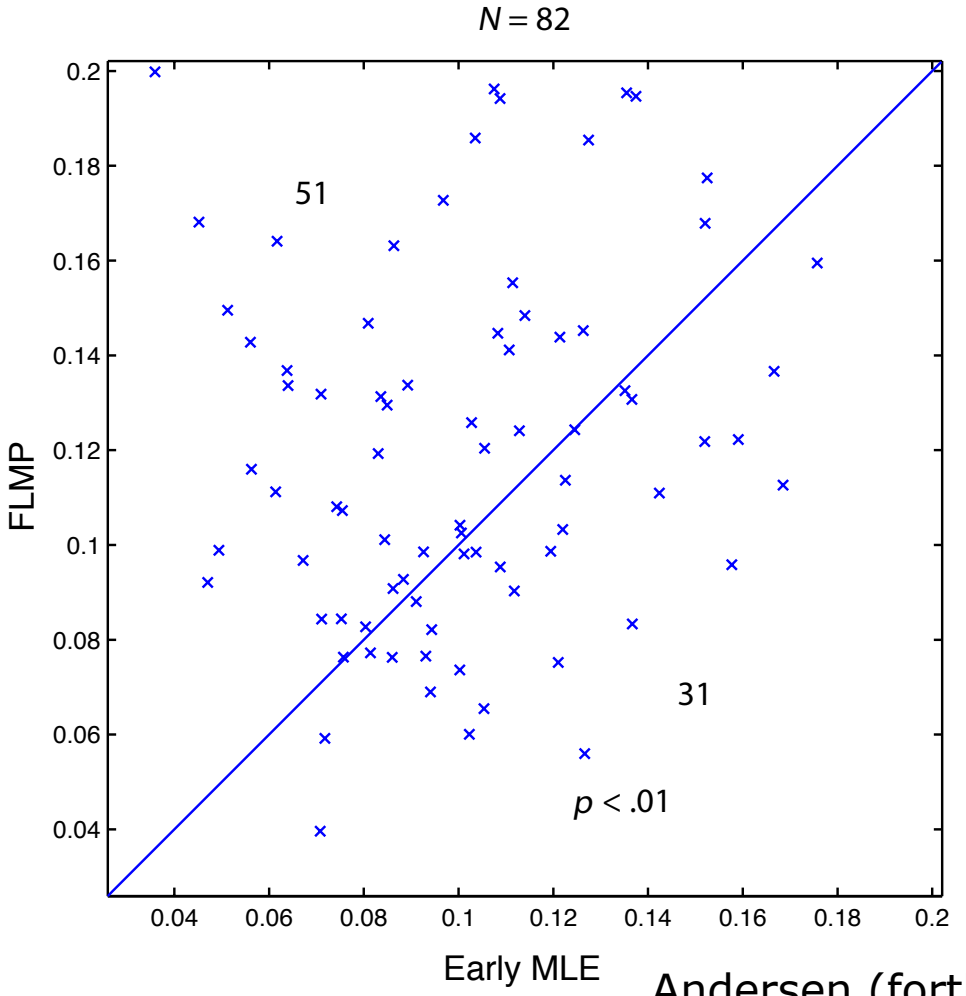
Andersen (forthcoming), JASA

Fits by stimulus



Andersen (forthcoming), JASA

Results by subject



Andersen (forthcoming), JASA

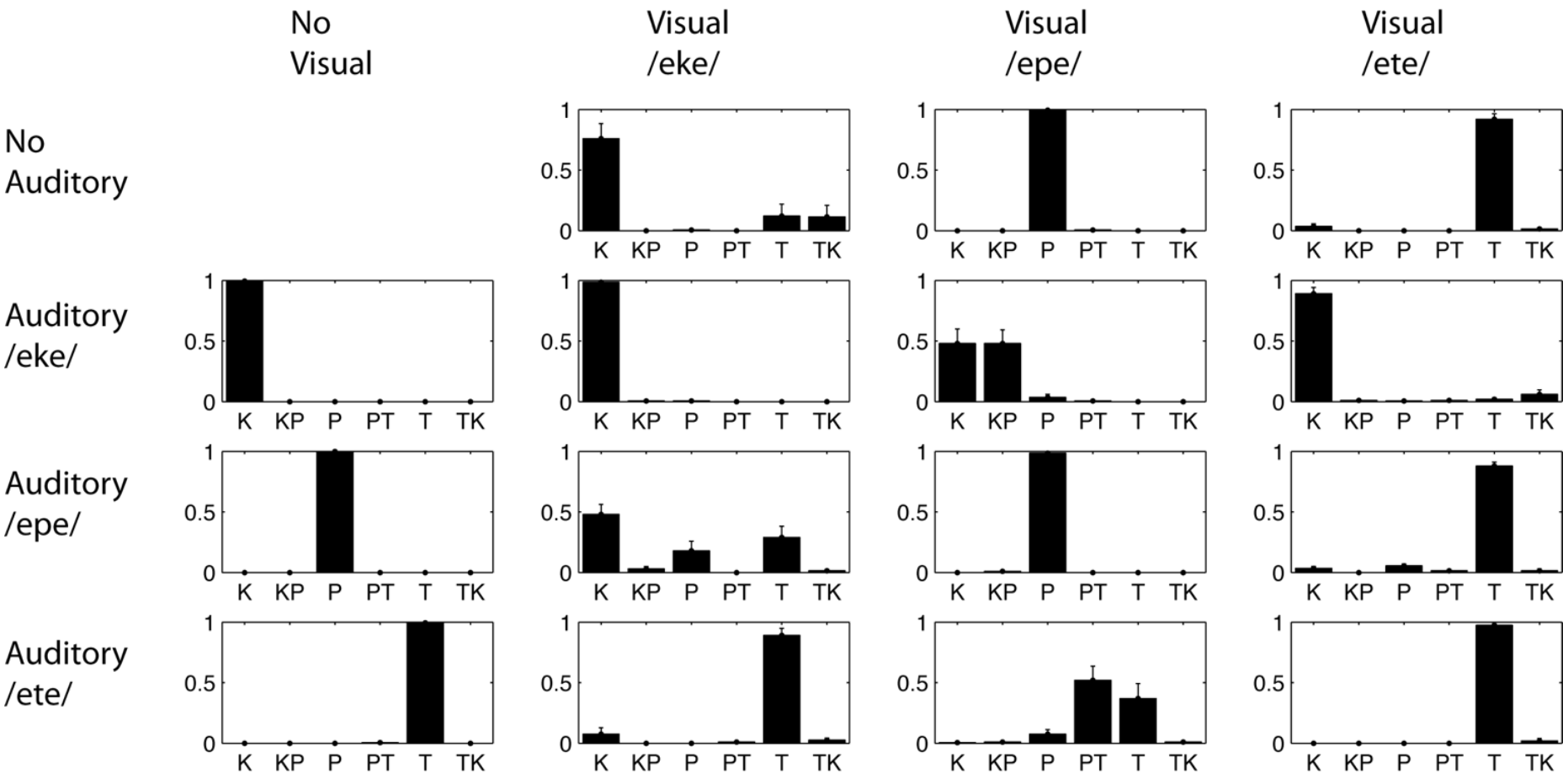
Other models

- Free weight model
  - Separates spacing from variance
  - 1 additional free parameter
  - Better fit – worse prediction

- Equal weight model
  - with a logistic noise distribution it is equivalent to late MLE
  - No improvement in fit / prediction

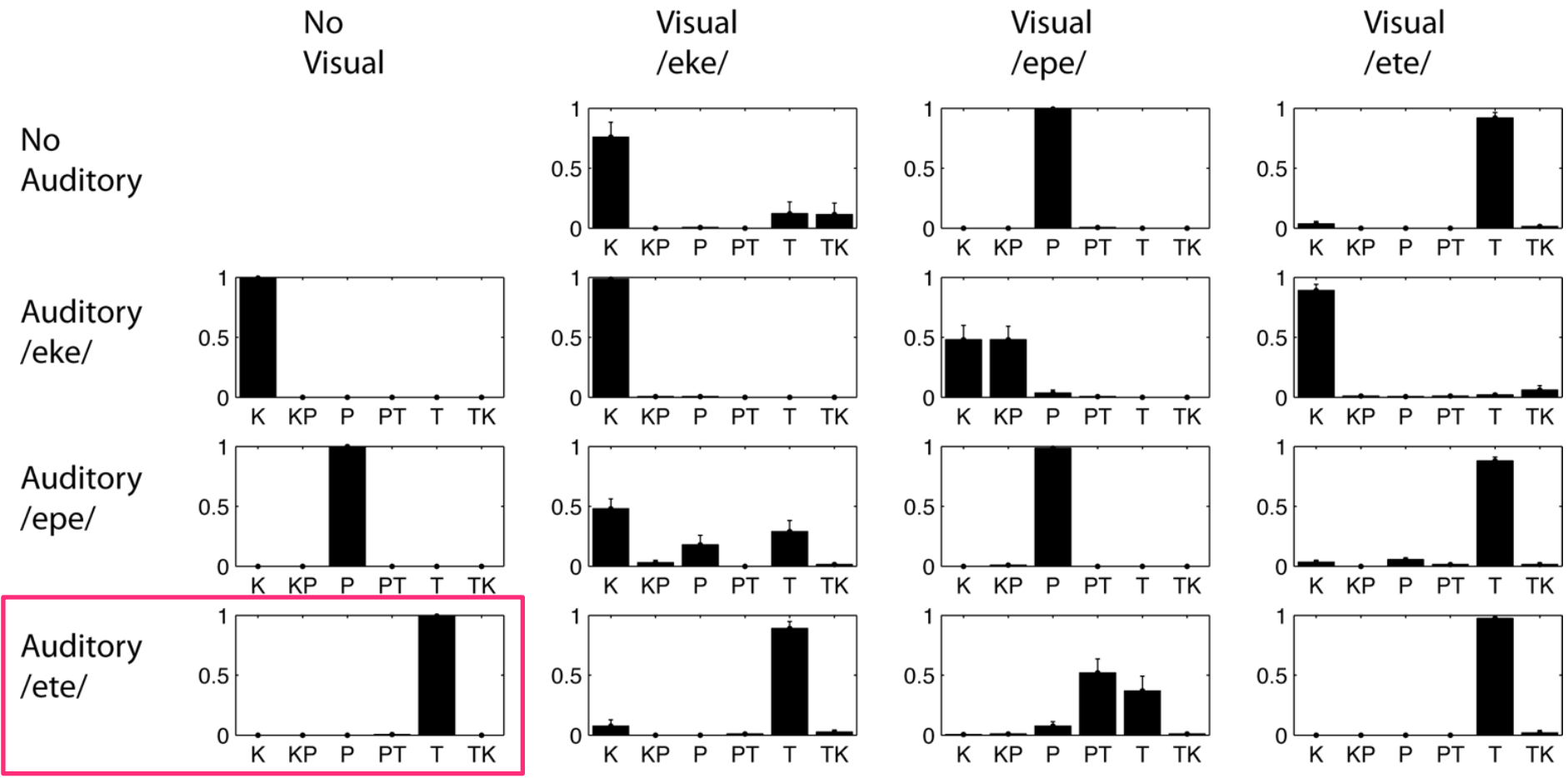
Andersen (forthcoming), JASA

Audiovisual speech perception



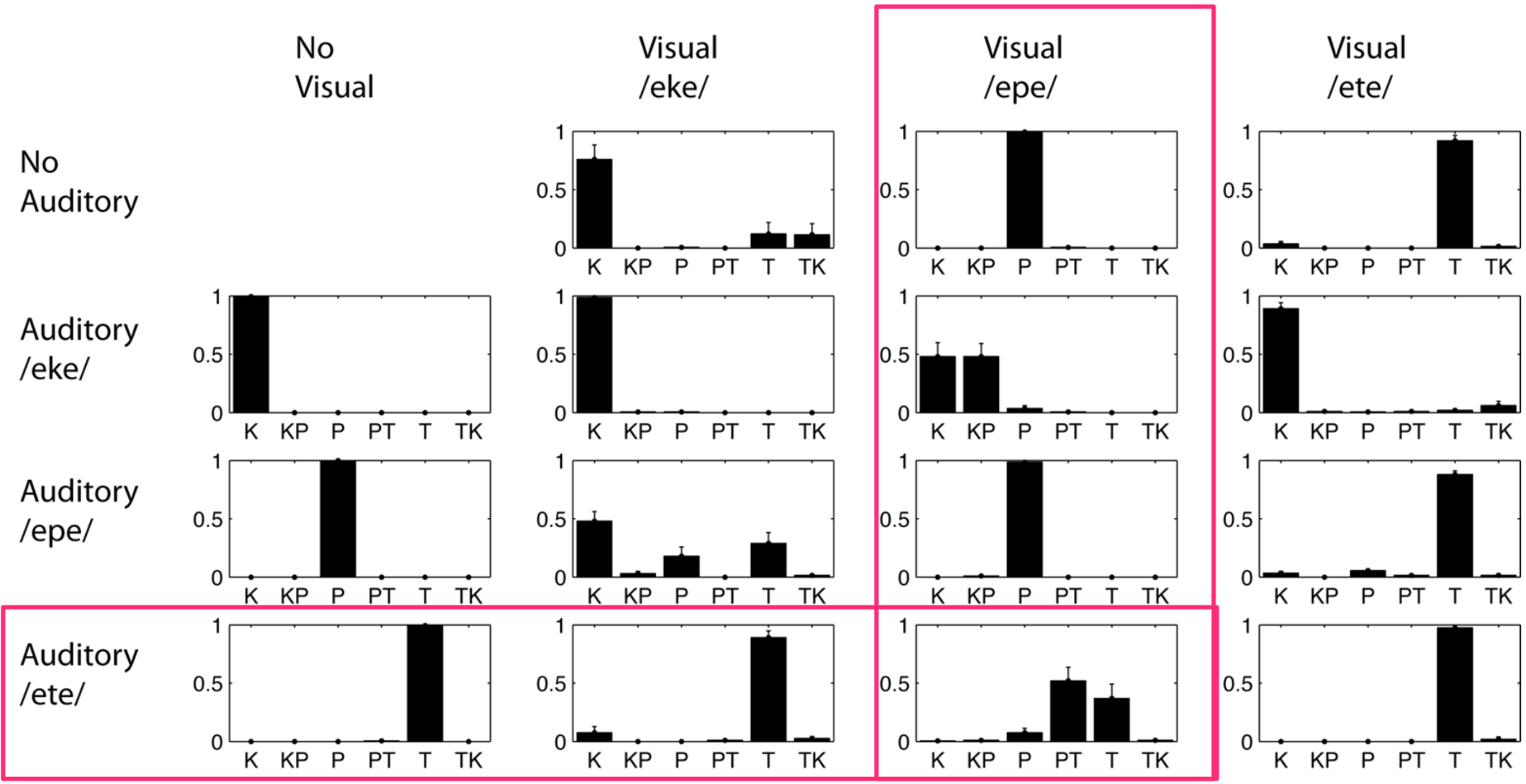
Andersen & Winther, in preparation

Audiovisual speech perception



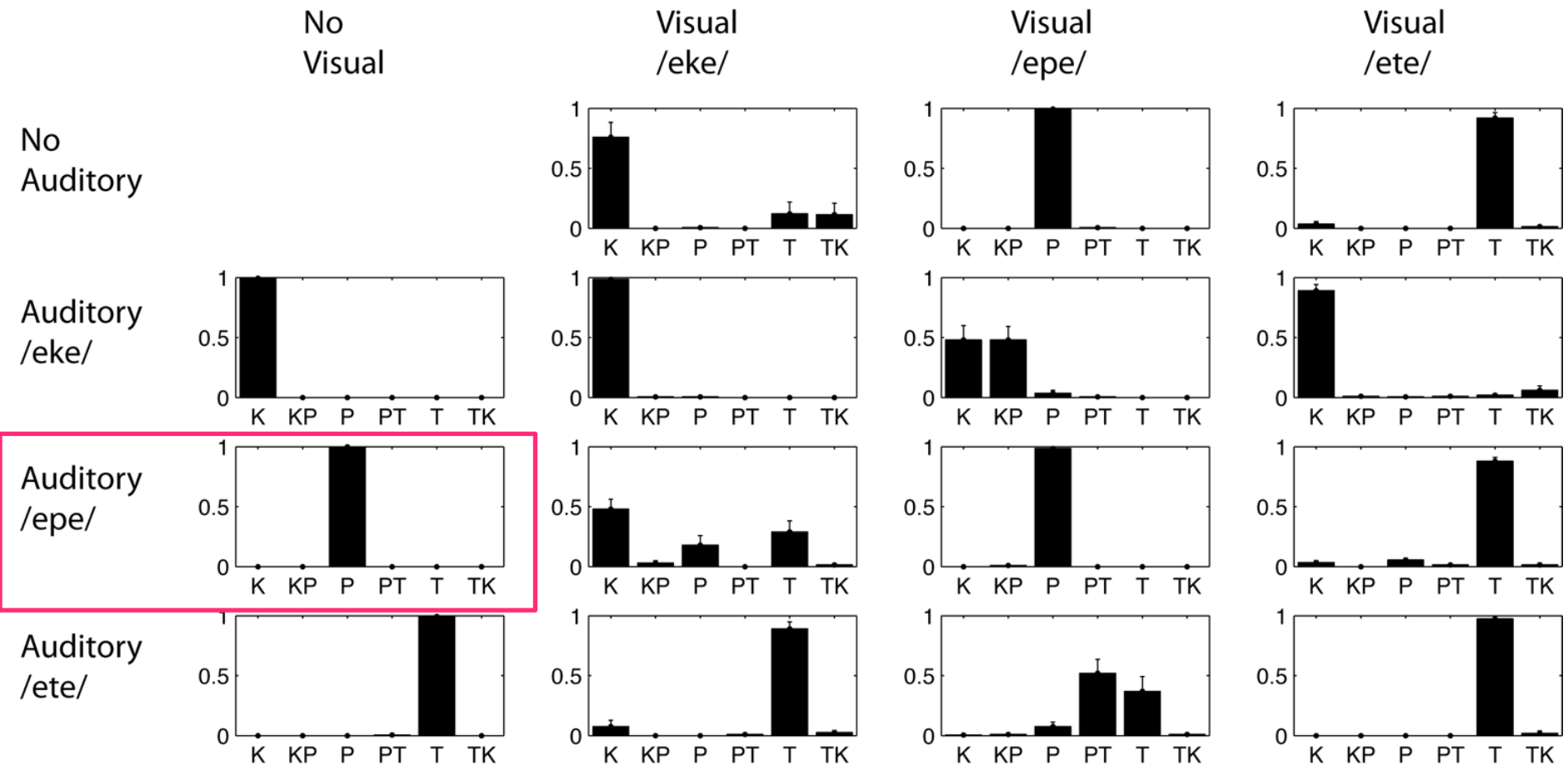
Andersen & Winther, in preparation

Audiovisual speech perception



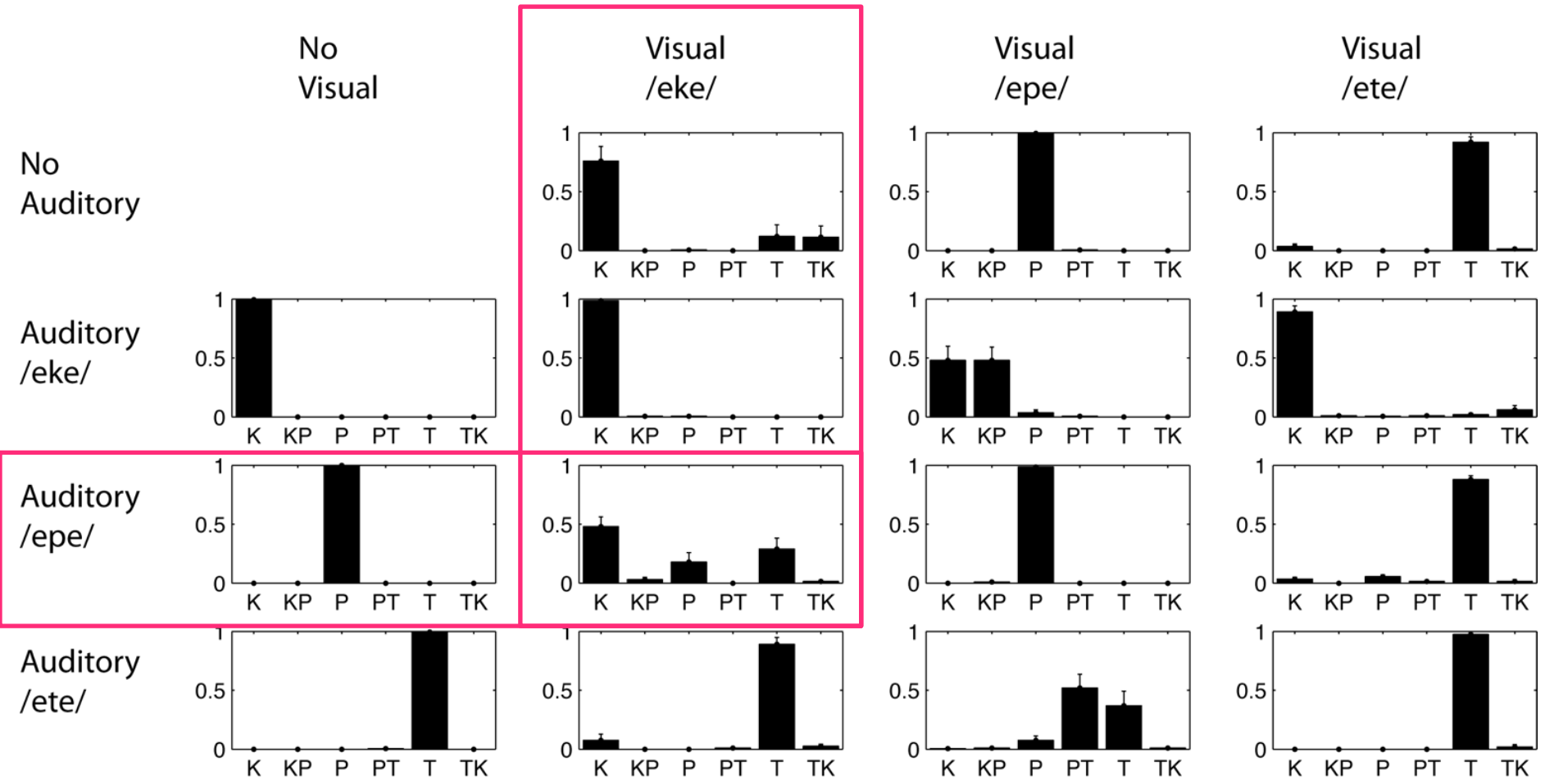
Andersen & Winther, in preparation

Audiovisual speech perception



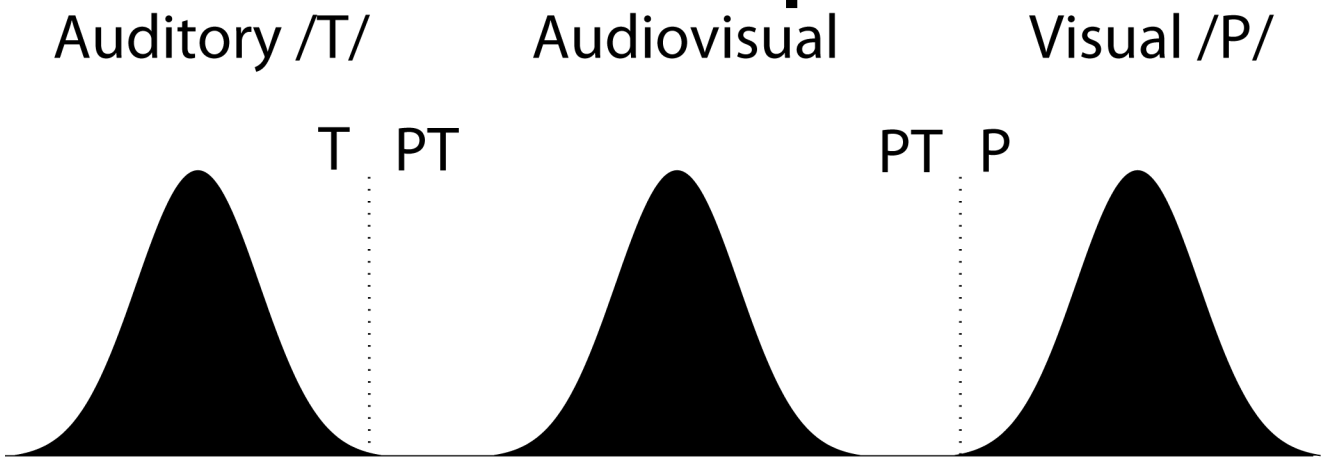
Andersen & Winther, in preparation

Audiovisual speech perception



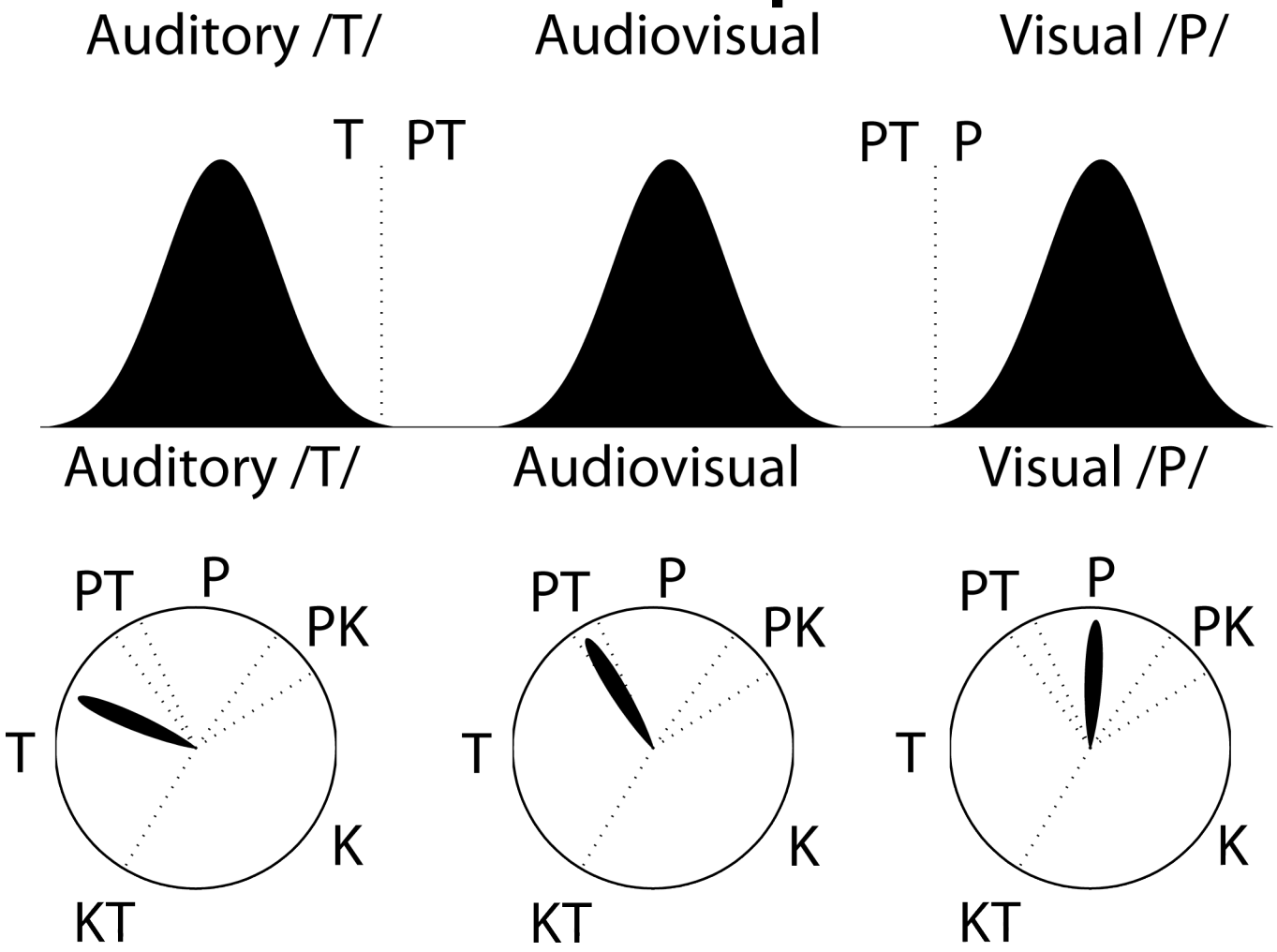
Andersen & Winther, in preparation

The continuous internal representation



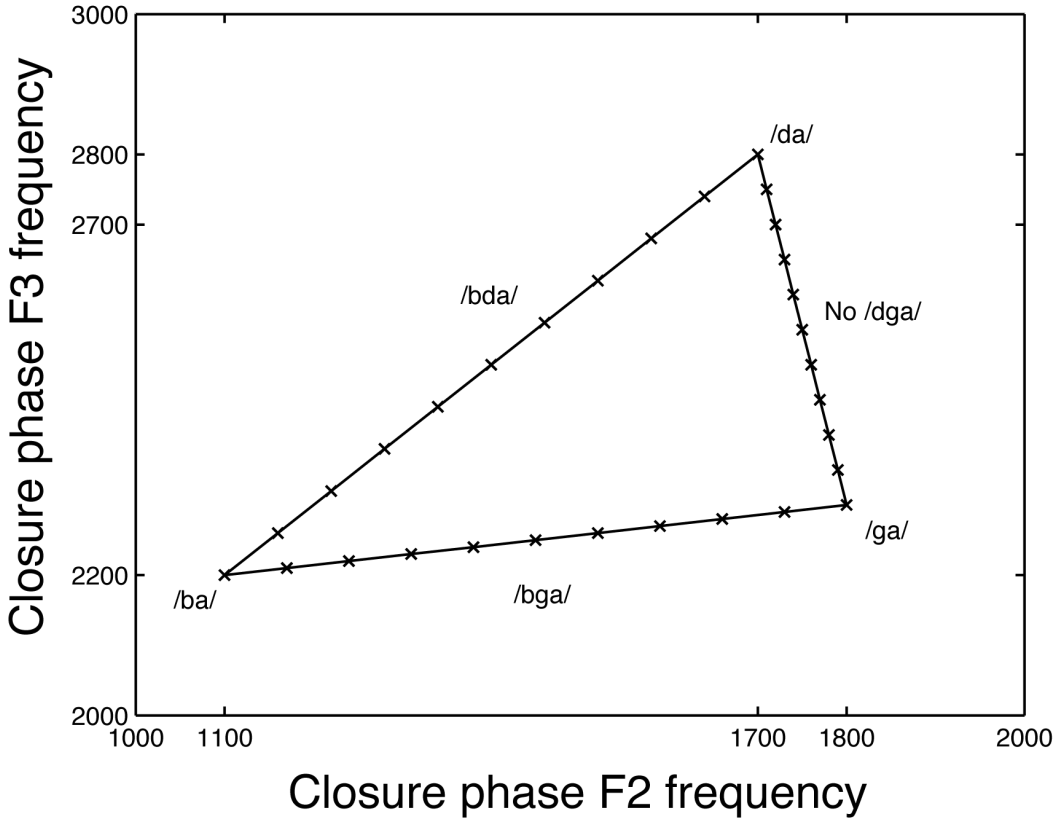
Andersen & Winther, in preparation

The continuous internal representation



Andersen & Winther, in preparation

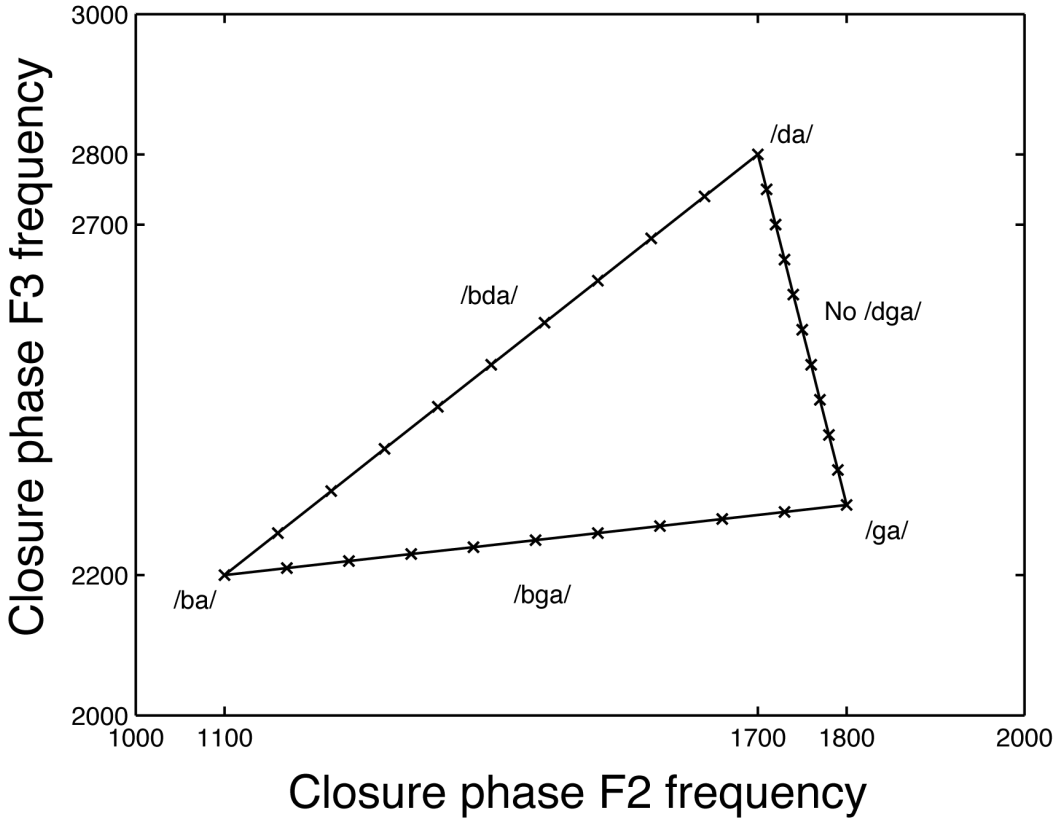
The continuous internal representation



Summerfield, Phonetica, 1979

Andersen & Winther, in preparation

The continuous internal representation

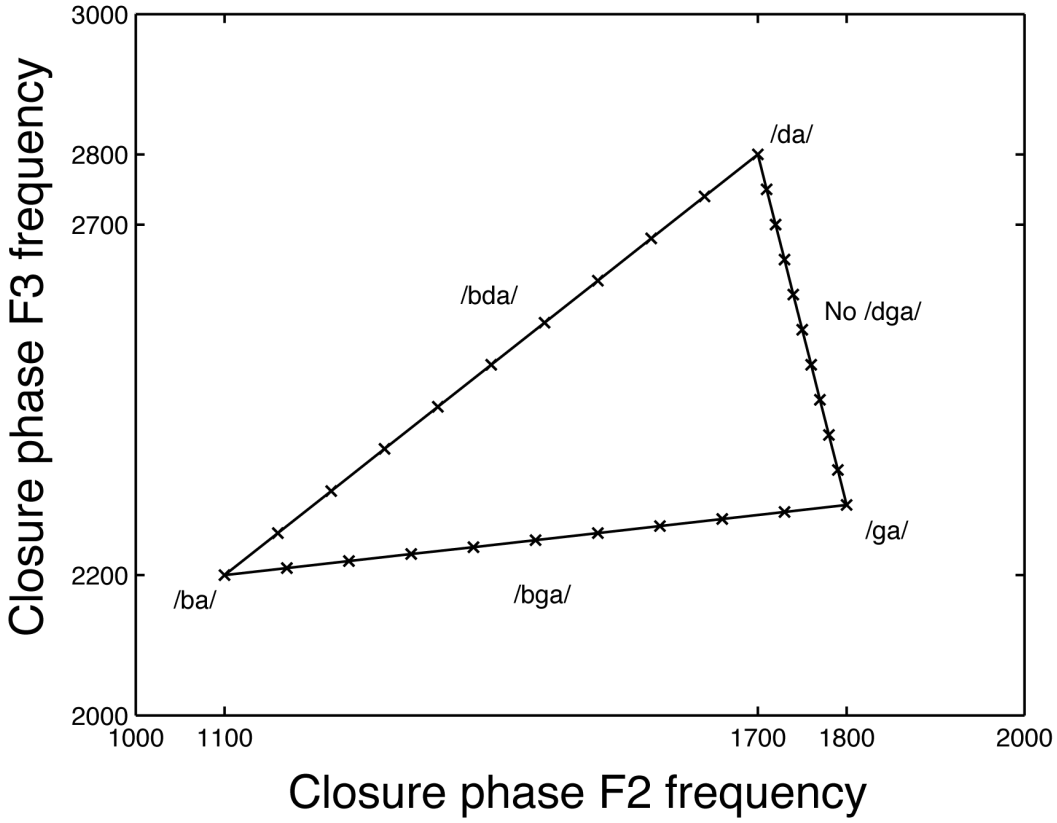


Cyclical model
without integration
35 parameters
RMSE = 0.004

Summerfield, Phonetica, 1979

Andersen & Winther, in preparation

The continuous internal representation



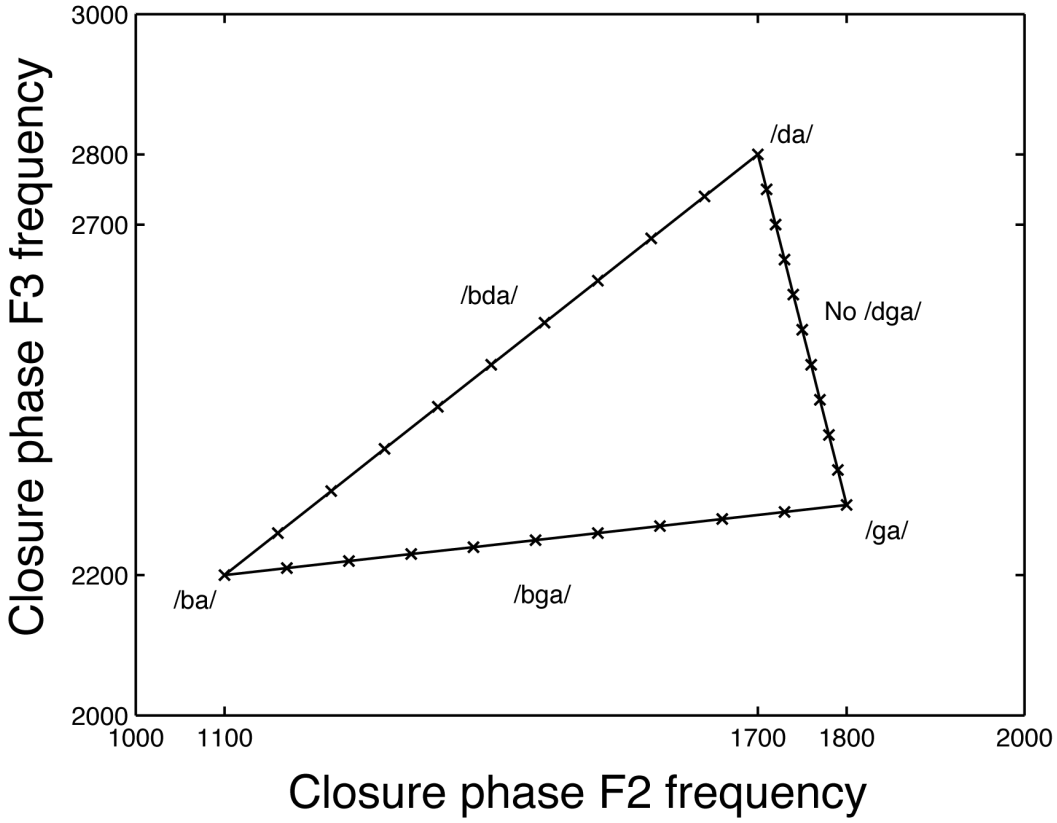
Cyclical model
without integration
35 parameters
RMSE = 0.004

Early MLE
13 parameters
RMSE = 0.02

Summerfield, Phonetica, 1979

Andersen & Winther, in preparation

The continuous internal representation



Cyclical model without integration
- 35 parameters
- RMSE = 0.004

Early MLE
- 13 parameters
- RMSE = 0.02

Late MLE / FLMP
- 30 parameters
- RMSE = 0.01

Summerfield, Phonetica, 1979

Andersen & Winther, in preparation

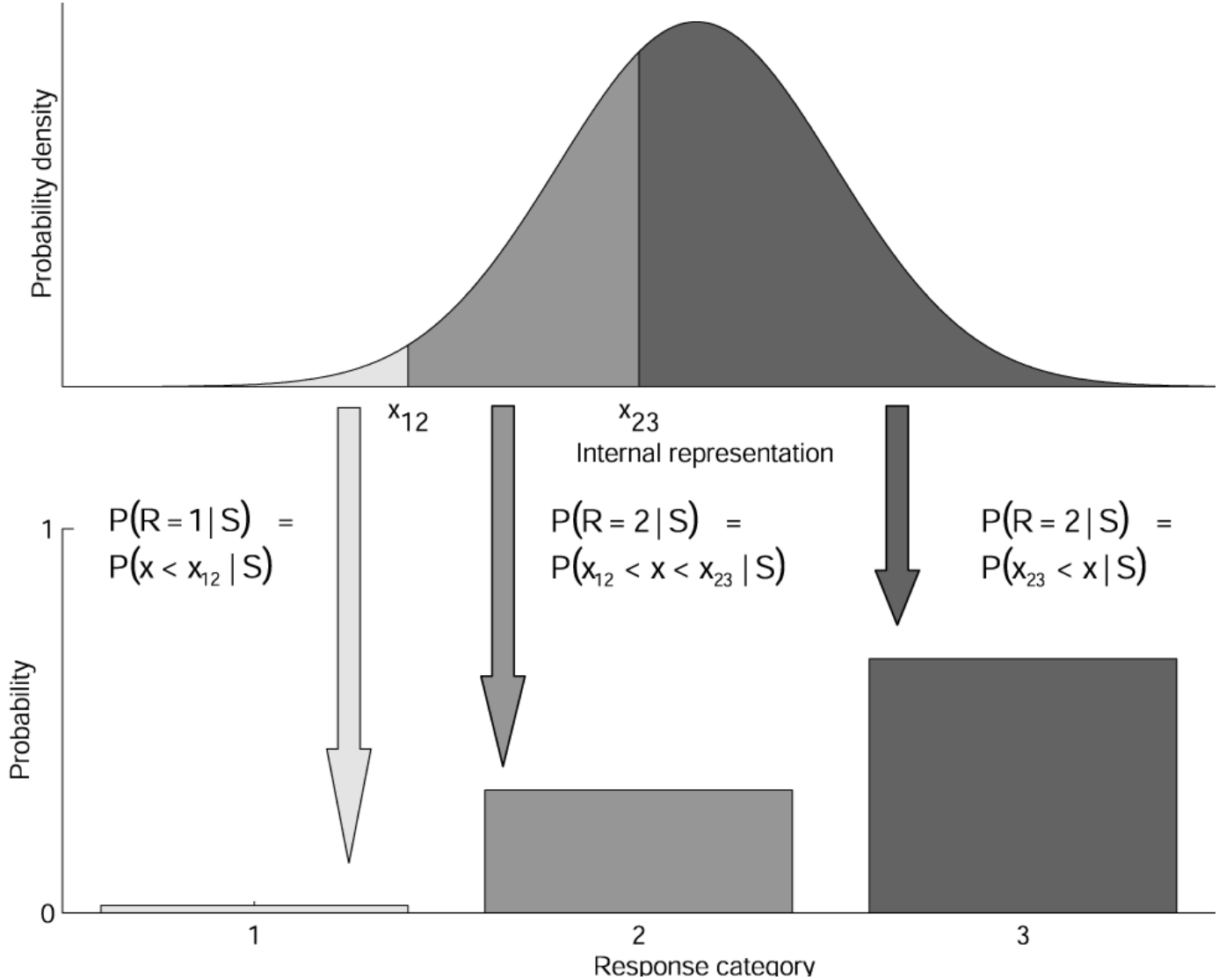
Cross-validation

- Leave one-out cross-validation
  - Late MLE / FLMP: poor results
  - Early MLE: Less poor results (but still poor)

- Why?
  - Non-linearity (not just number of free parameters)
  - Model fits very sensitive to small changes in parameter values
  - Assumes that the internal representation is unrealistically precise

Andersen & Winther, in preparation

Early MLE - Classification



Andersen & Winther, in preparation

regularization

- Don’t like something about your model?
- Optimize it away!

- I don’t like
  - Too high internal precision
    - Unrealistic
    - Makes models too flexible
    - Kills predictive power

- So, I add a penalizing term to the error when fitting

Andersen & Winther, in preparation

regularization

- **Early MLE - Continuous representation**
- The critical parameter is the width, σ, of the distributions
- Apply a Gaussian prior on 1/σ centered at zero (flat distribution)
- Penalizes for high precision

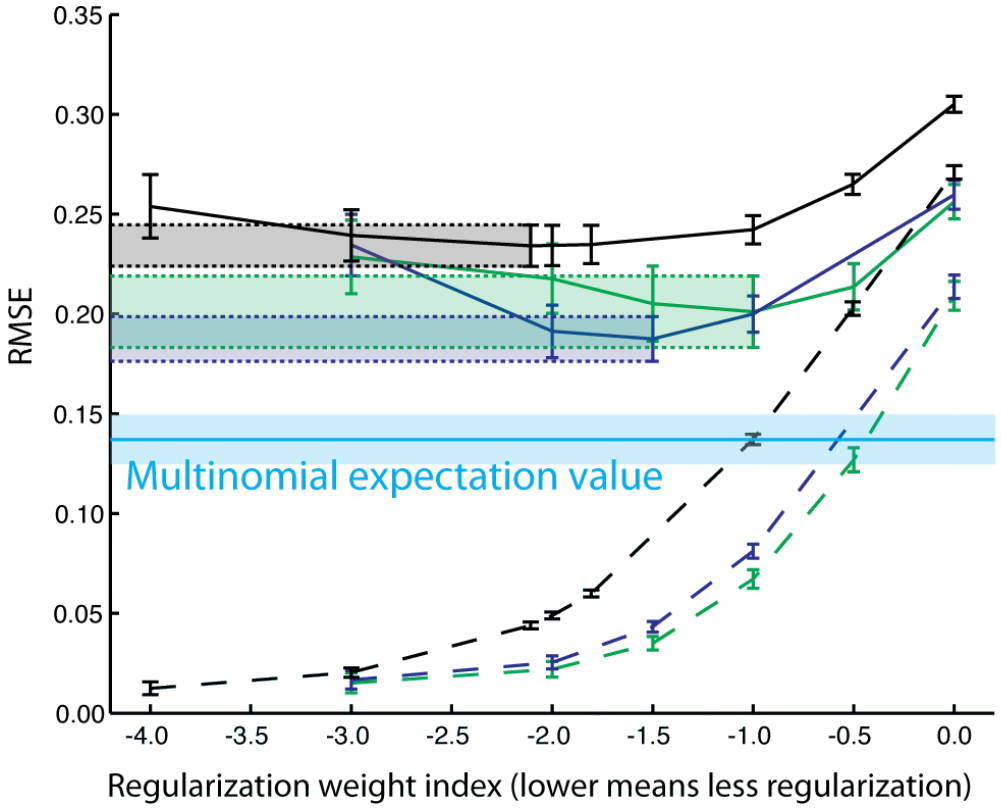
- **Late MLE / FLMP**
- The parameters are the unimodal response probabilities
  - Apply a uniform symmetric Dirichlet prior $P\left(P\left(R_r\right)\right) = \frac{1}{B\left(\alpha\right)} \prod_{r=1}^{N} P\left(R_r\right)^{\alpha-1}$
  - Penalize for negative log prior w/o the normalization term, $B\left(\alpha\right)$

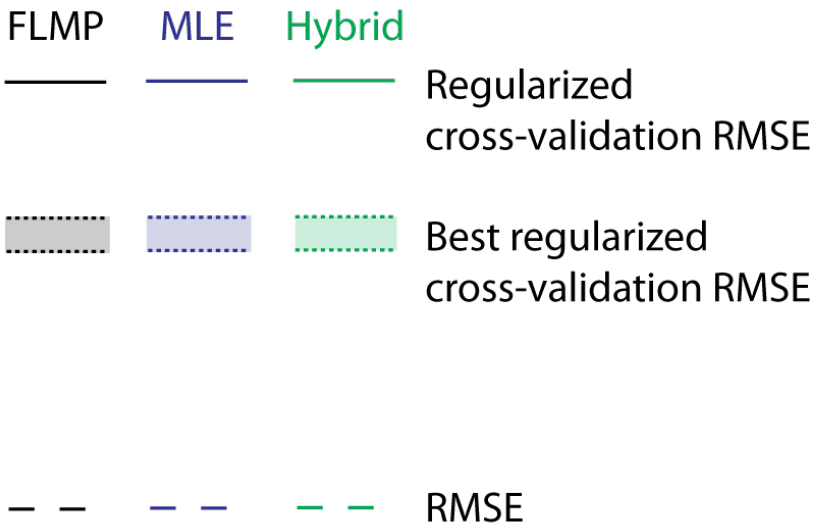
$$-\log\left(P\left(P\left(R_r\right)\right)\right) = -\left(\alpha-1\right)\sum_{r=1}^{N}\log\left(P\left(R_r\right)\right)$$

  - When the concentration parameter, $\alpha = 1$, the distribution is flat
    - Regularization penalizes peaked distributions
    - Peaked distributions are unstable

Andersen & Winther, in preparation

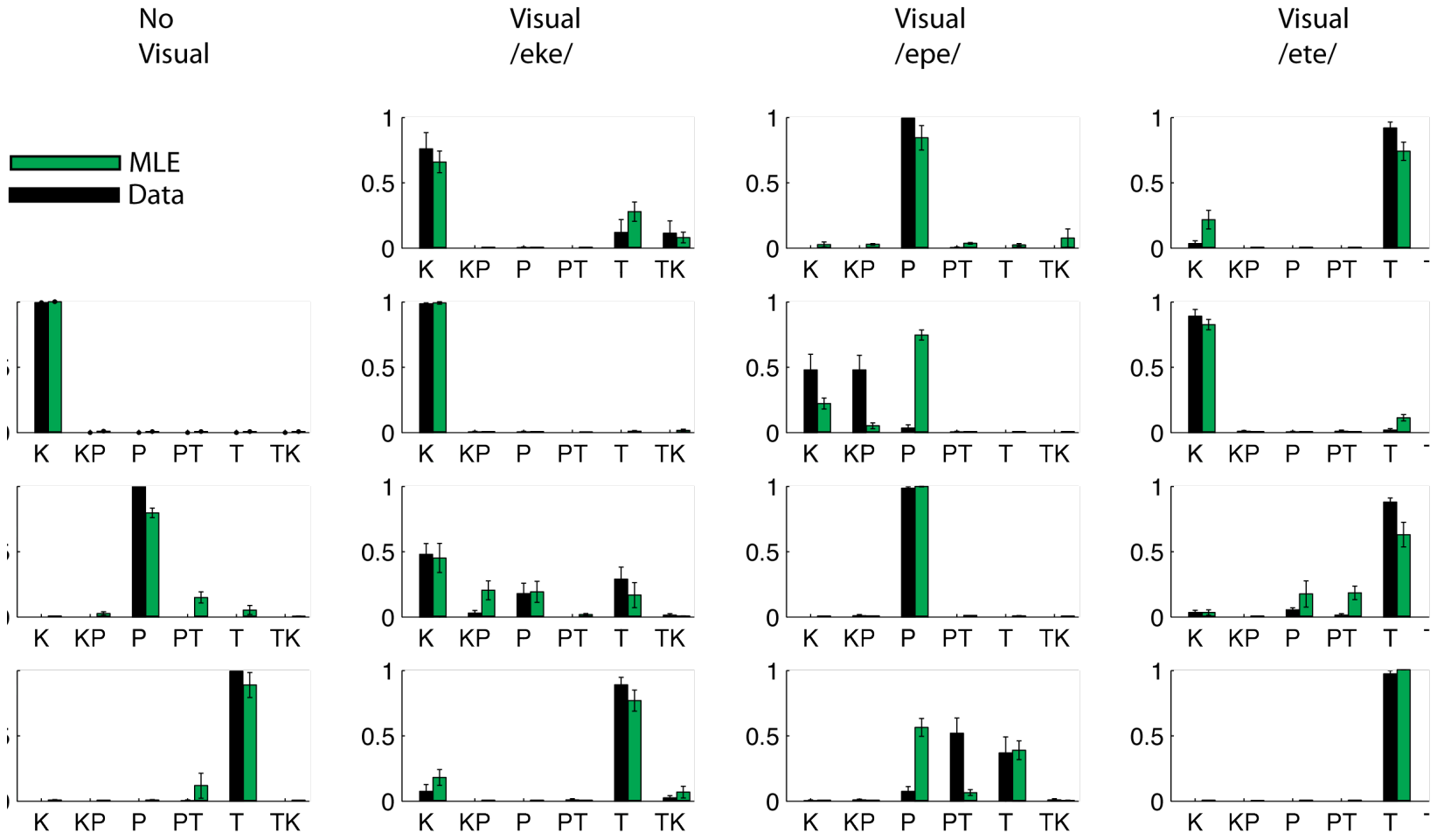
regularization





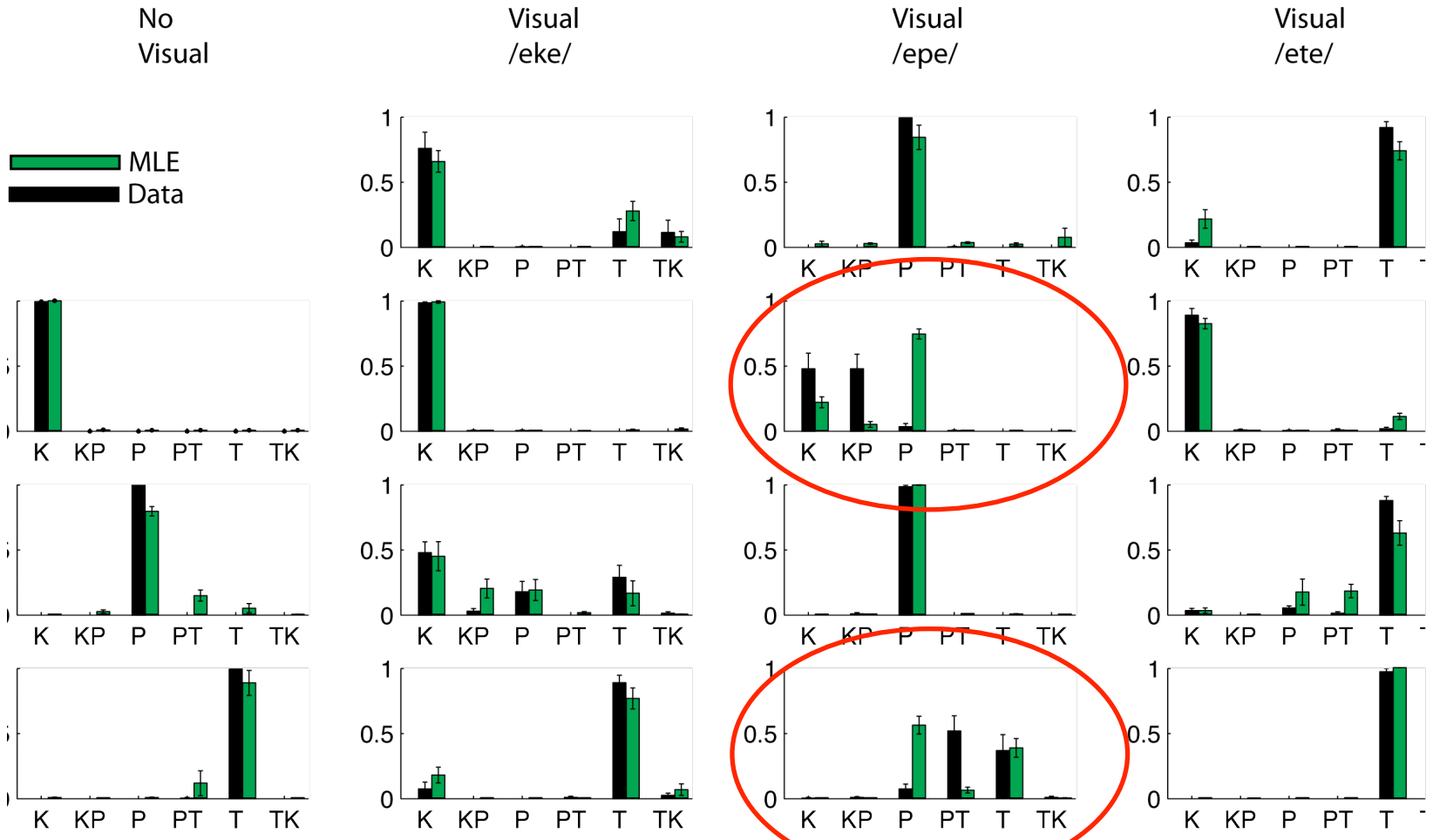
Andersen & Winther, in preparation

How good is Early MLE with regularization?



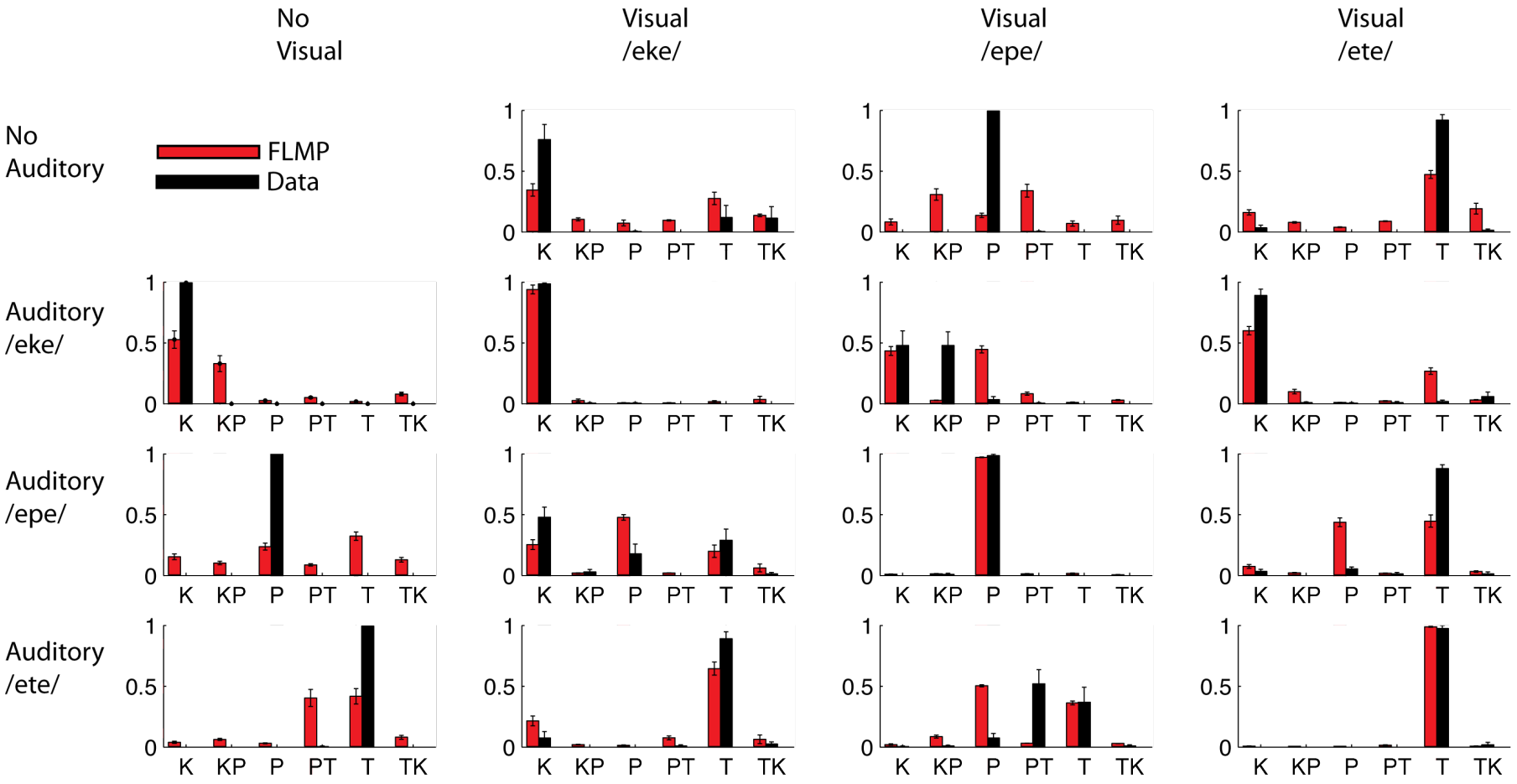
Andersen & Winther, in preparation

How good is Early MLE with regularization?



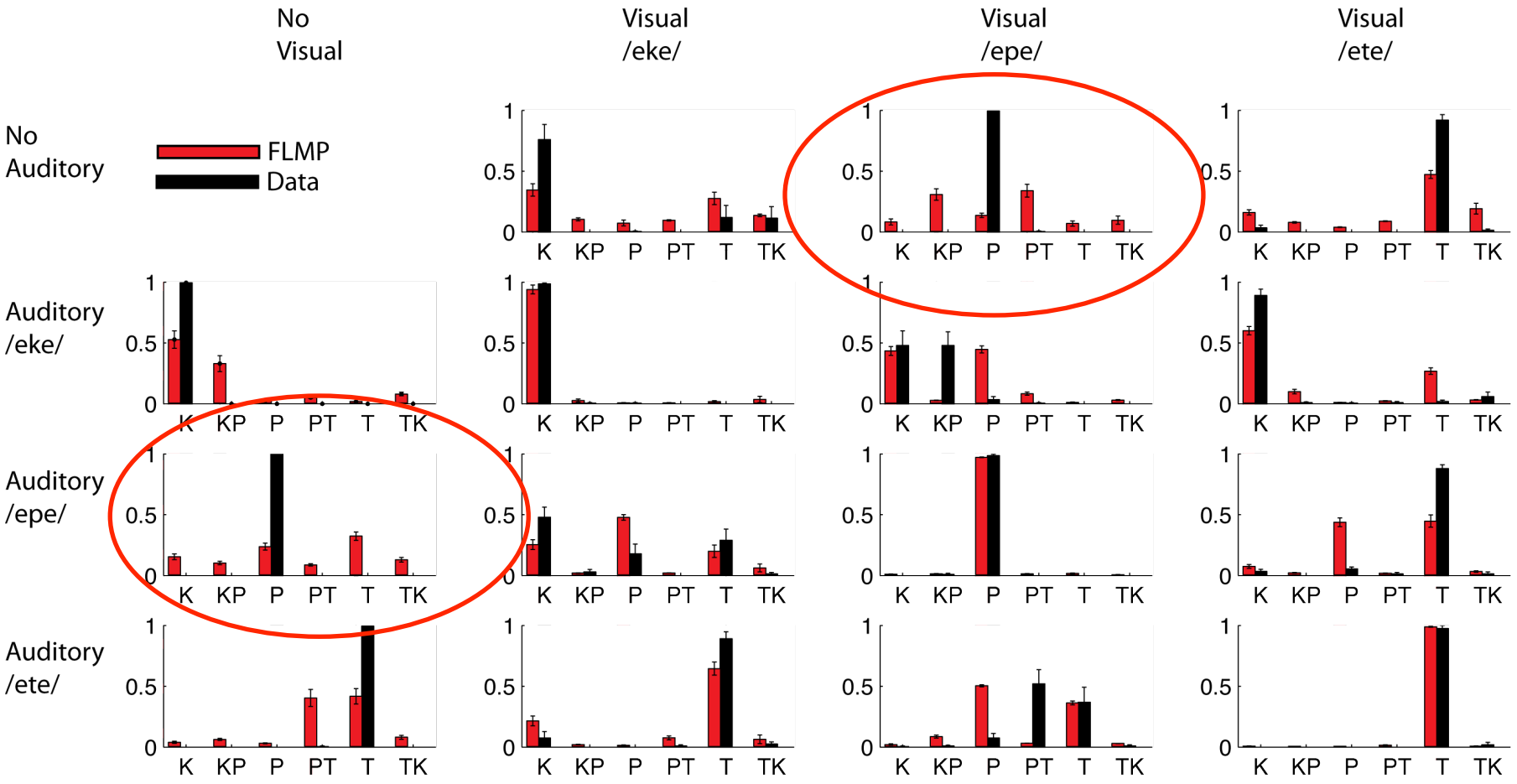
Andersen & Winther, in preparation

How good is Late MLE / FLMP with regularization?



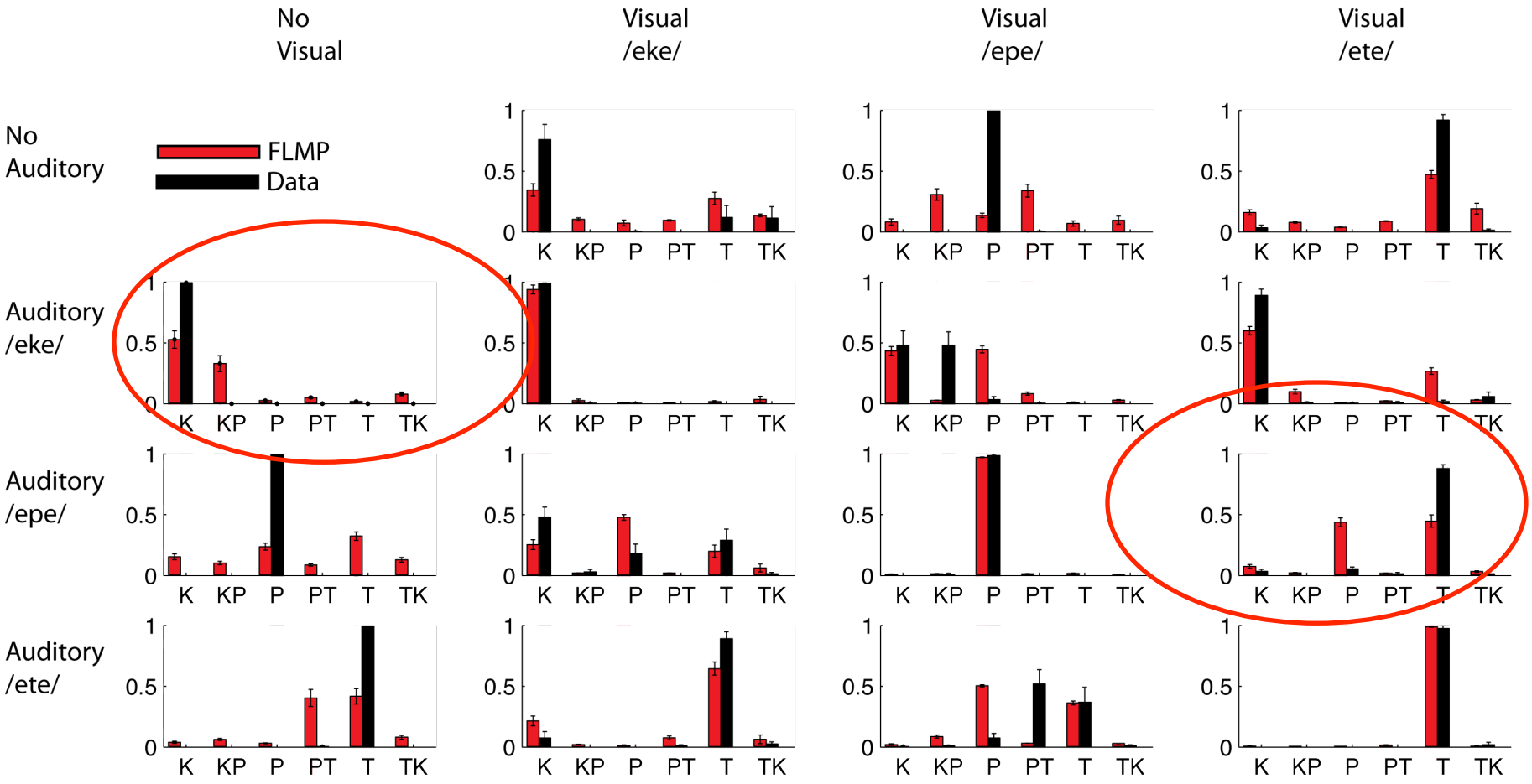
Andersen & Winther, in preparation

How good is Late MLE / FLMP with regularization?



Andersen & Winther, in preparation

How good is Late MLE / FLMP with regularization?



Andersen & Winther, in preparation

Conclusion

- Leave-one (stimulus/condition) out cross-validation is a great way to test models
  - Gives a good estimate of the right kind of generalization error

- Models of audiovisual speech perception benefits from
  - An underlying continuous parameter
  - regularization

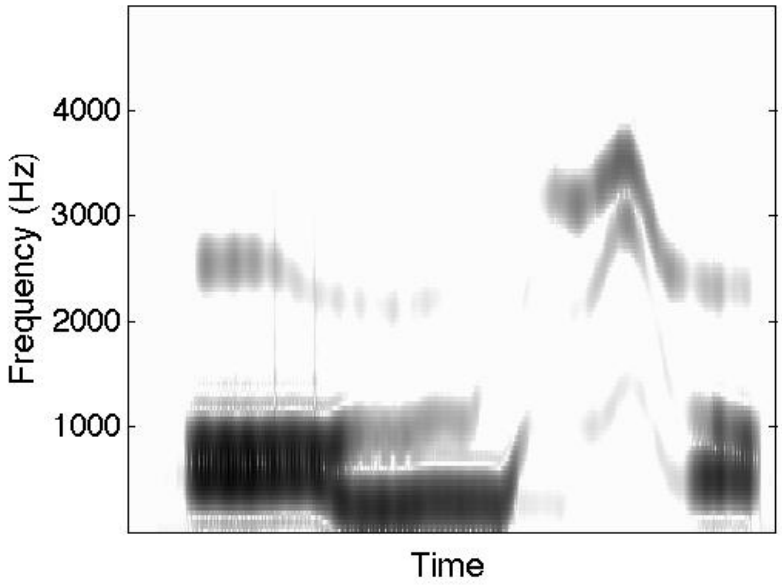
- The computational mechanism of integration is still unknown
  - Current results favor Early MLE
  - The Hybrid model performs almost as well
  - Weighted models make more sense

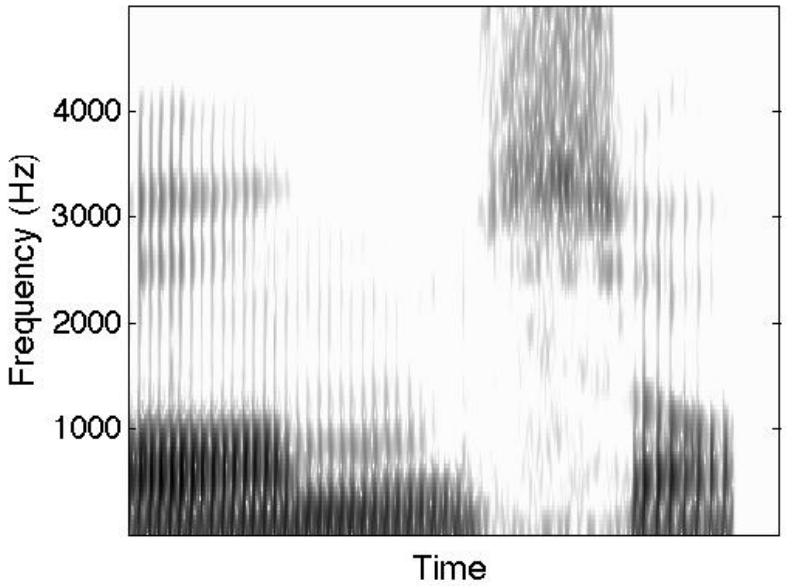
Andersen & Winther, in preparation

Modes of perception

Modes of perception

Sine-wave speech





Sine Wave Speech

- Created by placing time-varying sine wave tones at the three lowest formants of the speech signal
- Naïve observers do not recognize sine wave speech as speech
- Informed observers can understand the phonetic content

Sine Wave Speech - Stimuli

From Tuomainen, Andersen, Tiippana and Sams, Cognition, 2005

Sine Wave Speech - Paradigm

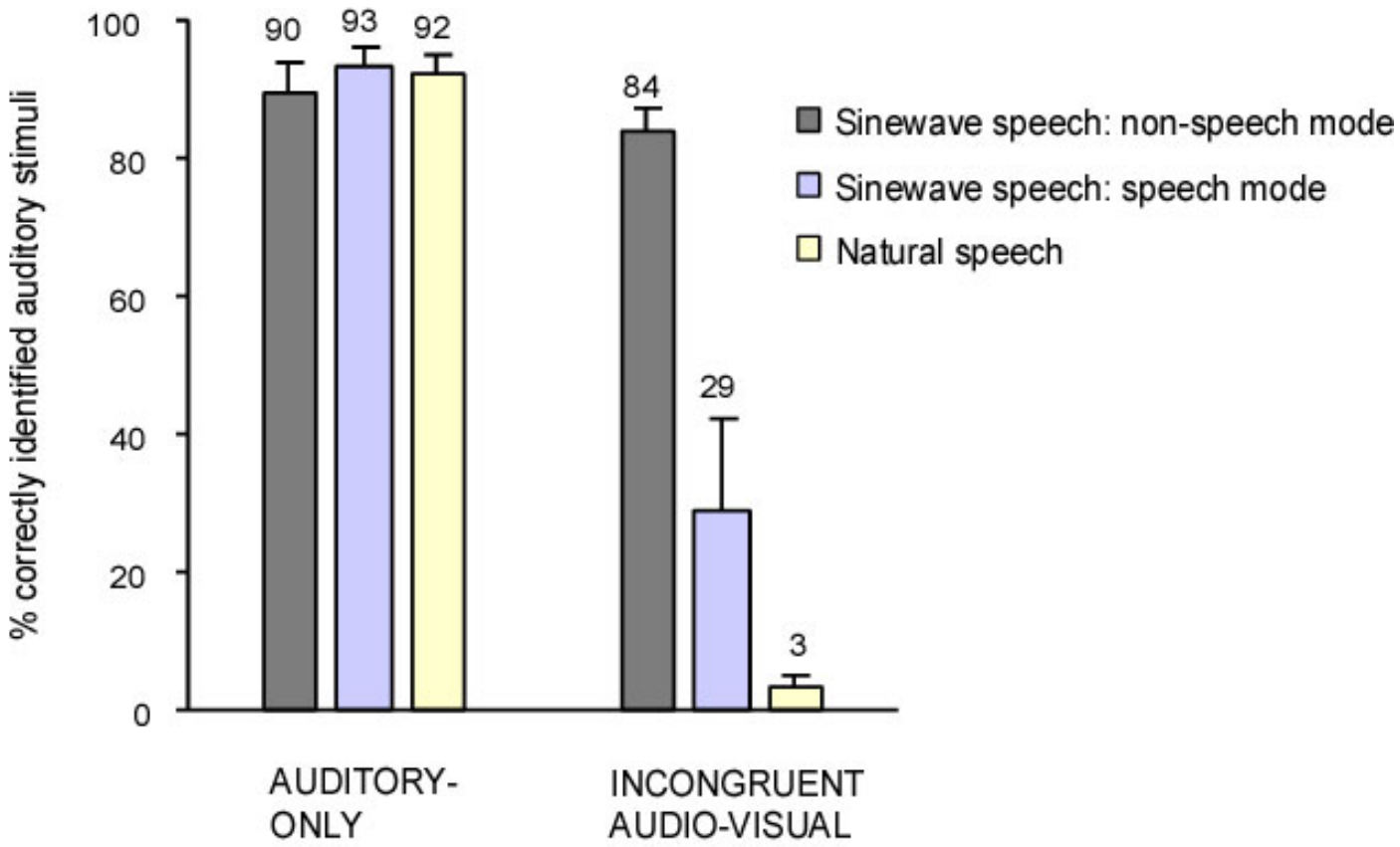
1. Training in non-speech mode (SWS)
2. Testing in non-speech mode (SWS)

3. Testing natural speech

4. Training in speech mode (SWS)
5. Testing in speech mode (SWS)

From Tuomainen, Andersen, Tiippana and Sams, Cognition, 2005

Sine Wave Speech - Results



From Tuomainen, Andersen, Tiippana and Sams, Cognition, 2005

Sine Wave Speech - Conclusion

- Strong audiovisual integration of sine wave speech and the talking face
- But! Only when observers are in speech mode
- Demonstrates strong top-down influence on audiovisual integration of speech

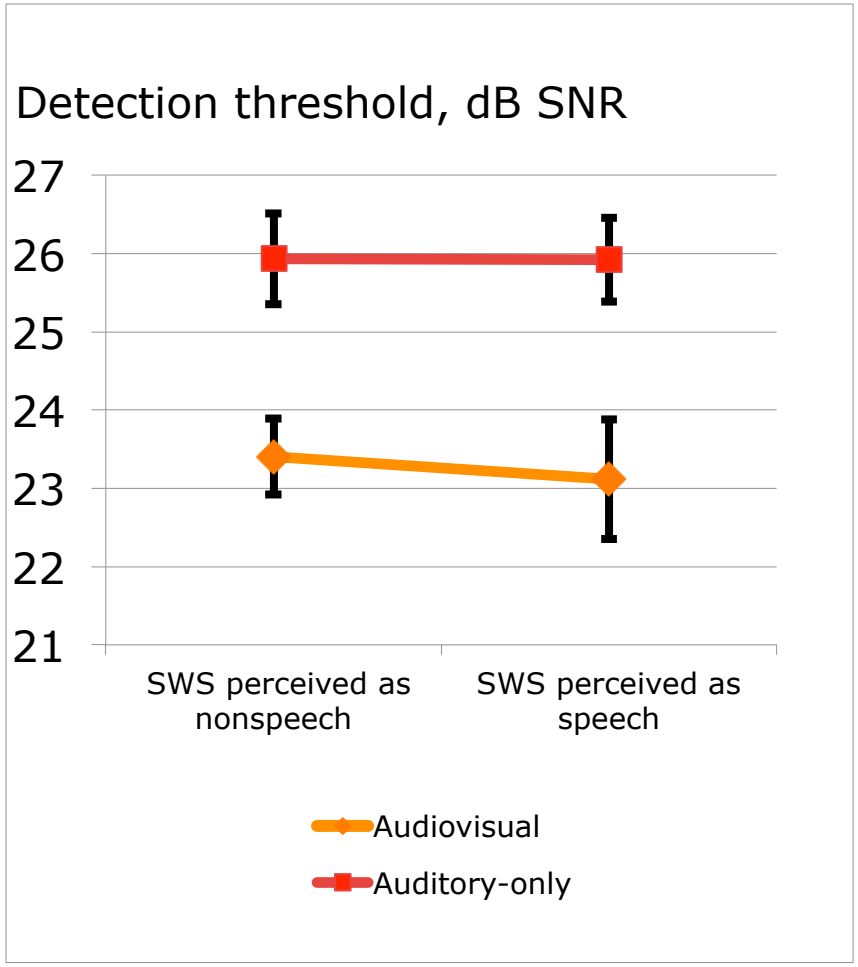
From Tuomainen, Andersen, Tiippana and Sams, Cognition, 2005

Audiovisual detection advantage

- The AV detection advantage (Grant & Seitz, JASA, 2000)
  - Acoustic speech detection threshold lowered by congruent visual speech
  - AV gain sizes reported between 1.6 and 2.7 dB, depending on method
  - Not just a response bias
    - 2 AFC w/ adaptive staircase procedure – visual information identical in the 2 alternatives
  - Is it speech specific?

From Eskelund, Tuomainen & Andersen, Exp. Br. Res., 2011

AV detection - results

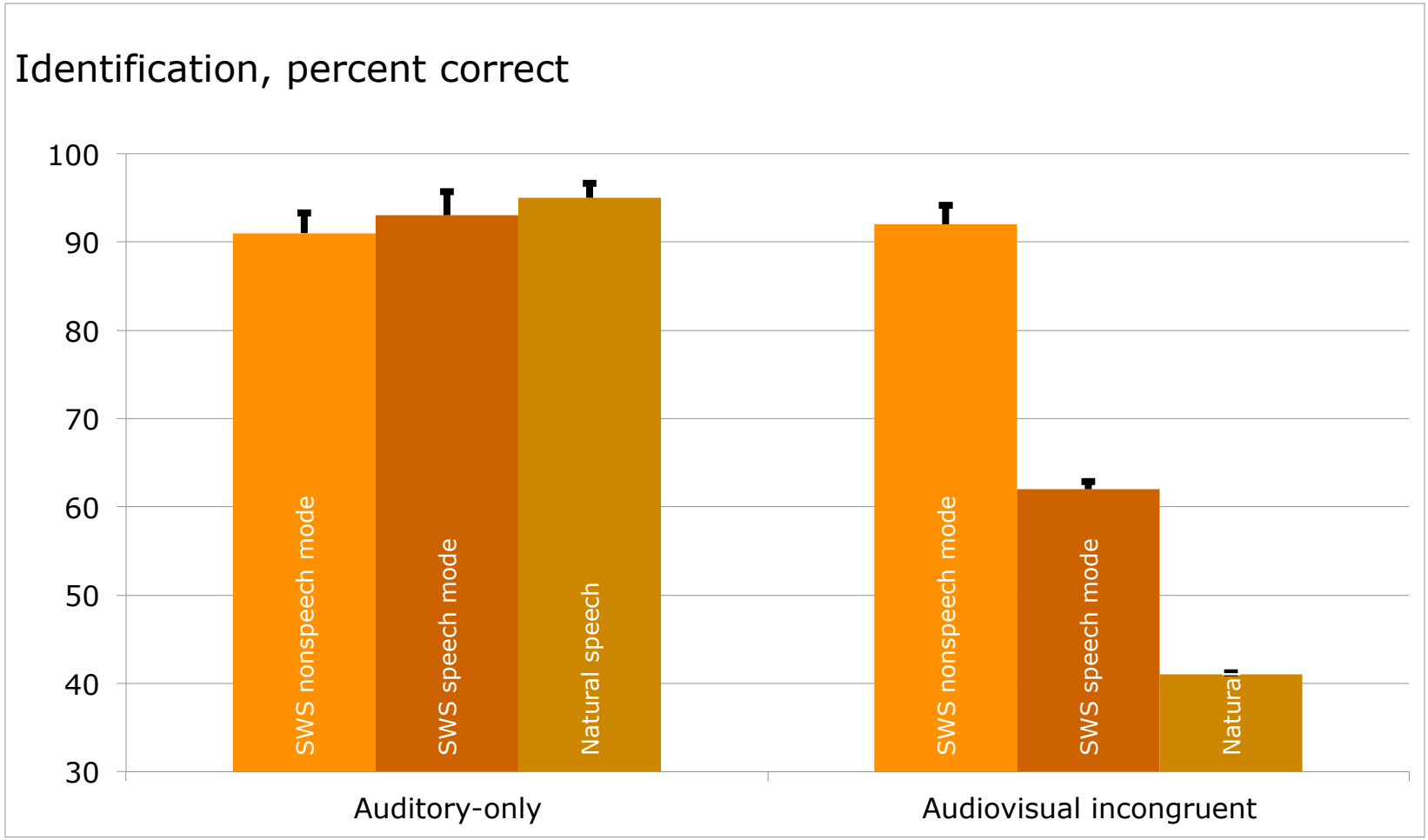


The AV detection advantage occurs also for SWS

No difference in AV detection advantage between nonspeech and speech conditions

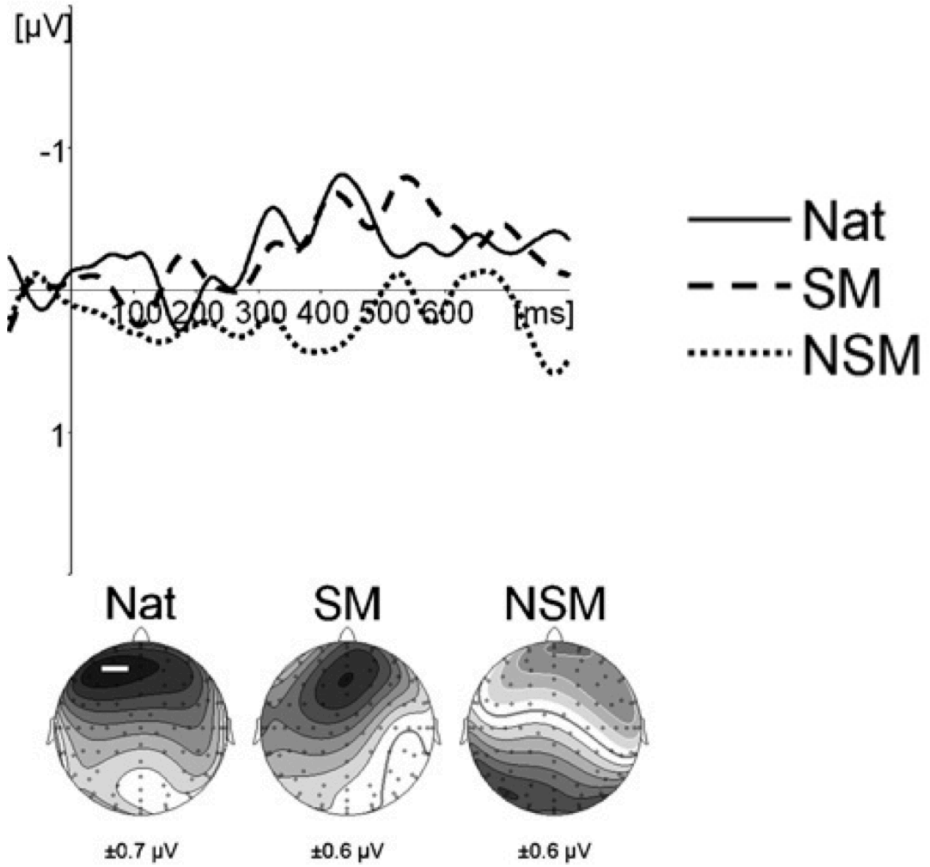
From Eskelund, Tuomainen & Andersen, Exp. Br. Res., 2011

Identification - results



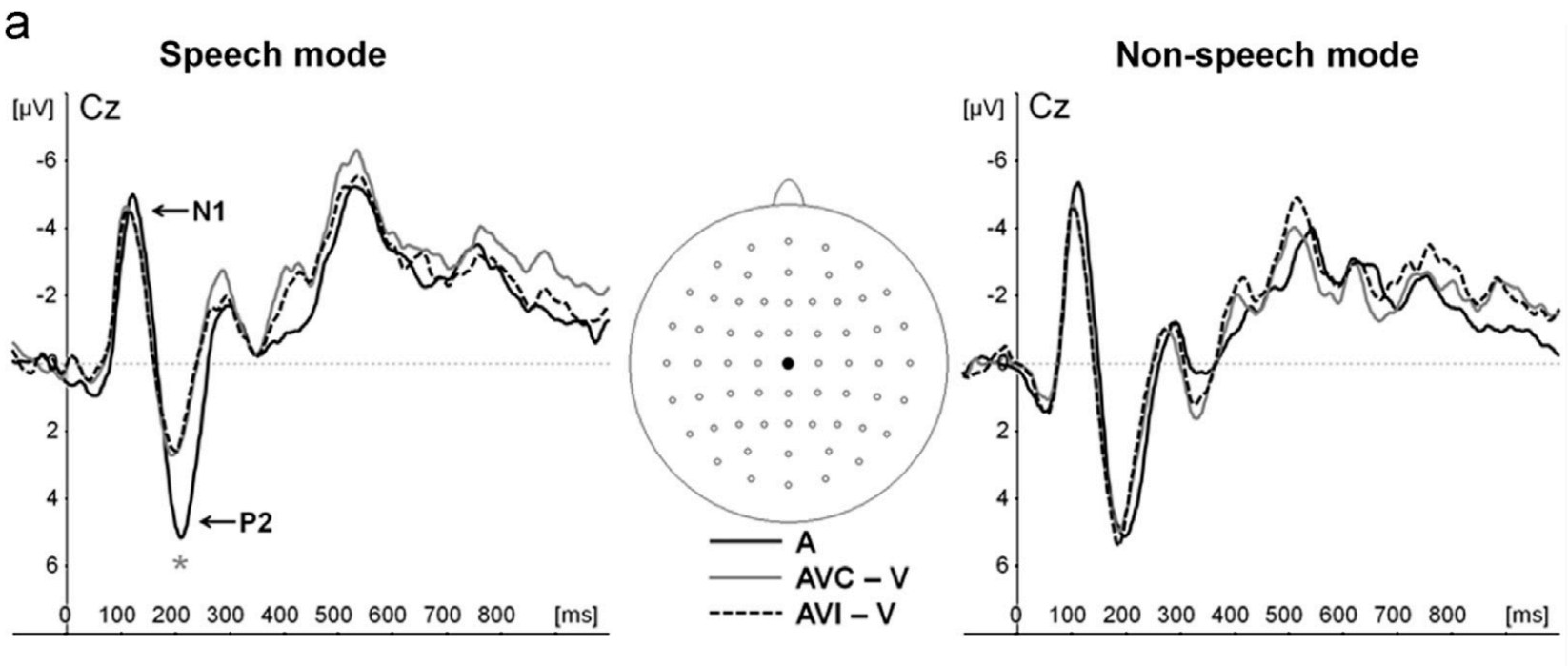
From Eskelund, Tuomainen & Andersen, Exp. Br. Res., 2011

EEG – mismatch negativity MMN



Stekelenburg & Vroomen (2012), Neuropsychologia

EEG – N1 and P2

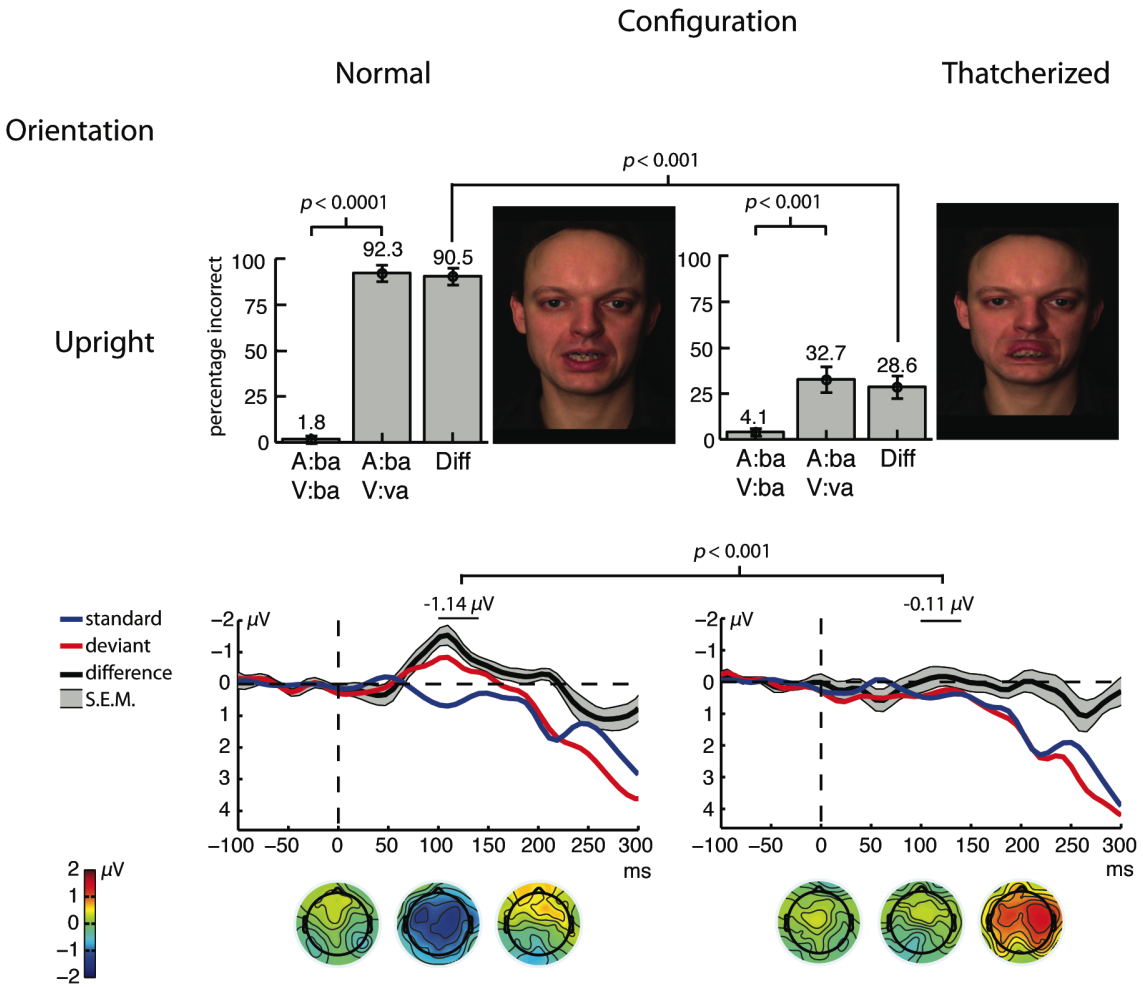


Baart, Stekelenburg & Vroomen (2014), Neuropsychologia

Margaret Thatcher

Margaret Thatcher

The McThatcher MMN



Eskelund, MacDonald & Andersen (2015), Neuropsychologia

Modes of perception

- Phonetic audiovisual integration varies for very similar stimuli
  - Sine-wave speech
  - McThatcher effect

- Audiovisual integration is a multi-stage process
  - Speech mode in the McGurk illusion and the detection advantage

- Phonetic audiovisual integration is reflected in the MMN and the P2
  - But not the N1

Thanks for listening

Any ???

Audiovisual SDT

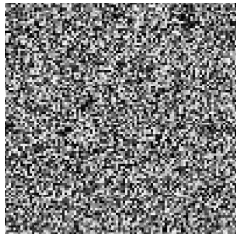
Audiovisual SDT

- Audiovisual integration in signal detection
  - Sound can enhance visual sensitivity
  - Frasinetti et al., 2003
- Integration of magnitude in weak signals
  - Cat chasing mouse in the dusk
  - Involves the Superior Colliculus
  - Direcst attention to the location of a change
    - Stein et al.
- Loudness increase perceived brightness
  - Stein, London, Wilkinson, Price, 1996.

Paradigm

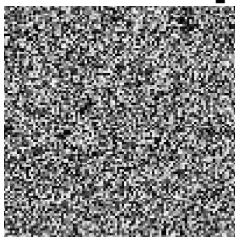
Lumin: ↑

Sound: —



Lumin: ↑

Sound: ↑



Andersen & Mamassian, Vision Research, 08

Perceptual effects

- Sound carries no information
- Bias free paradigm
- Two stimulus attributes may integrate audiovisually:
  - Transients
  - Sustained loudness and brightness

Andersen & Mamassian, Vision Research, 08

Attention

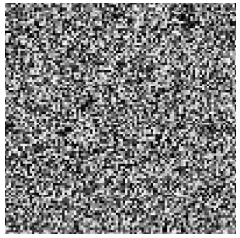
- Directional effects
  - If louder makes brighter, then a luminance *decrease* should be more difficult to detect when the sound becomes louder
- Additional task
  - Identify the luminance change as an increase or decrease
- Attentional effects
  - Exogenous attentional cueing
  - Reduction of temporal uncertainty

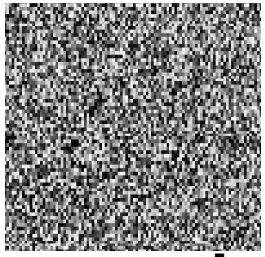
Andersen & Mamassian, Vision Research, 08

Paradigm

Lumin: ↑

Sound: —



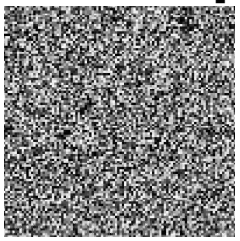


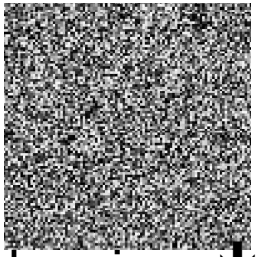
Lumin: ↓

Sound: —

Lumin: ↑

Sound: ↑



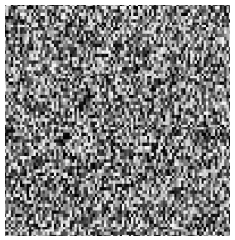


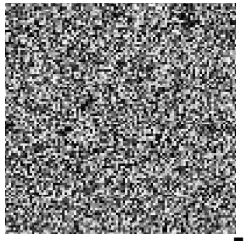
Lumin: ↓

Sound: ↑

Lumin: ↑

Sound: ↓



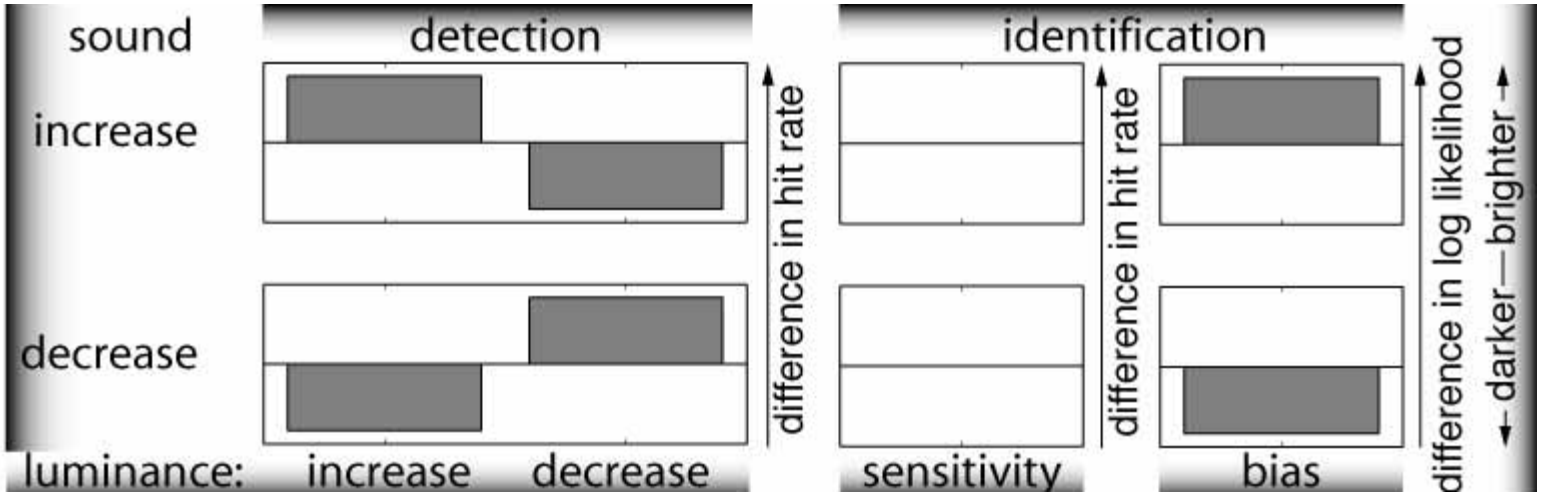


Lumin: ↓

Sound: ↓

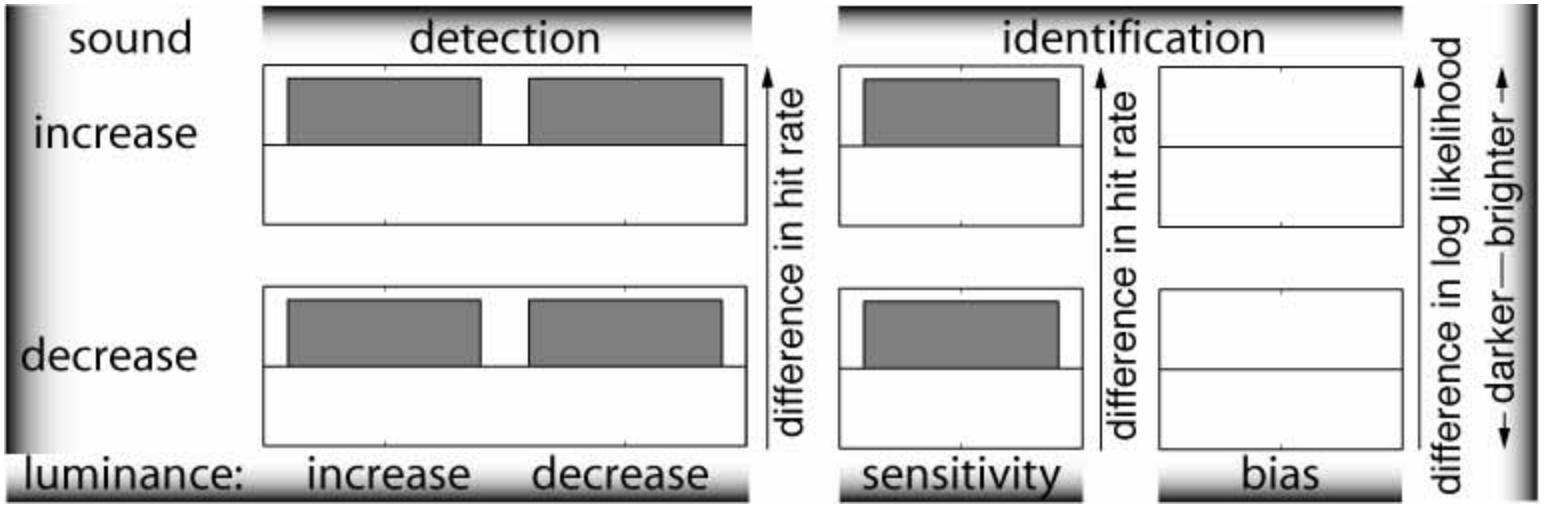
Andersen & Mamassian, Vision Research, 08

Predictions
Loudness/brightness interaction



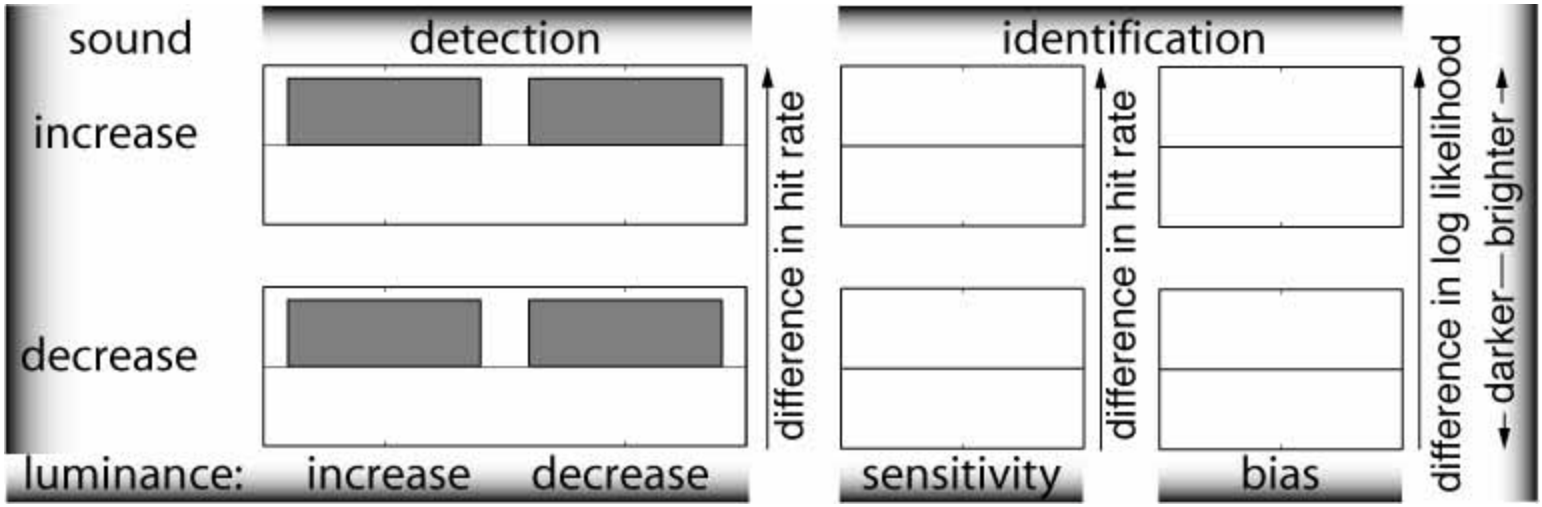
Andersen & Mamassian, Vision Research, 08

Predictions
Attention and Uncertainty



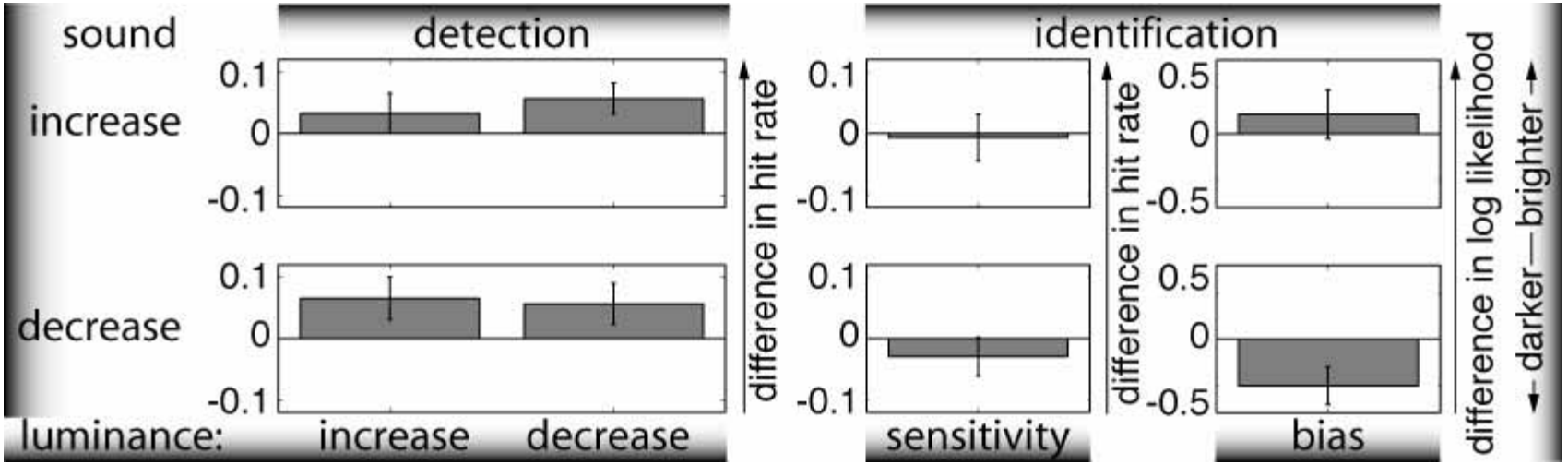
Andersen & Mamassian, Vision Research, 08

Predictions
Transient interactions



Andersen & Mamassian, Vision Research, 08

Results

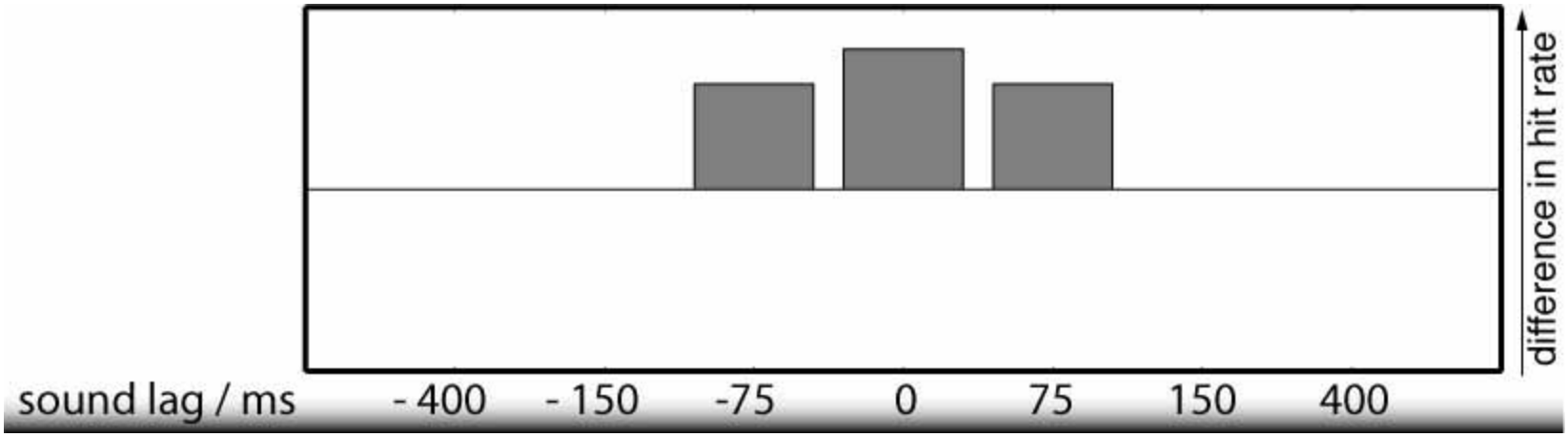


Andersen & Mamassian, Vision Research, 08

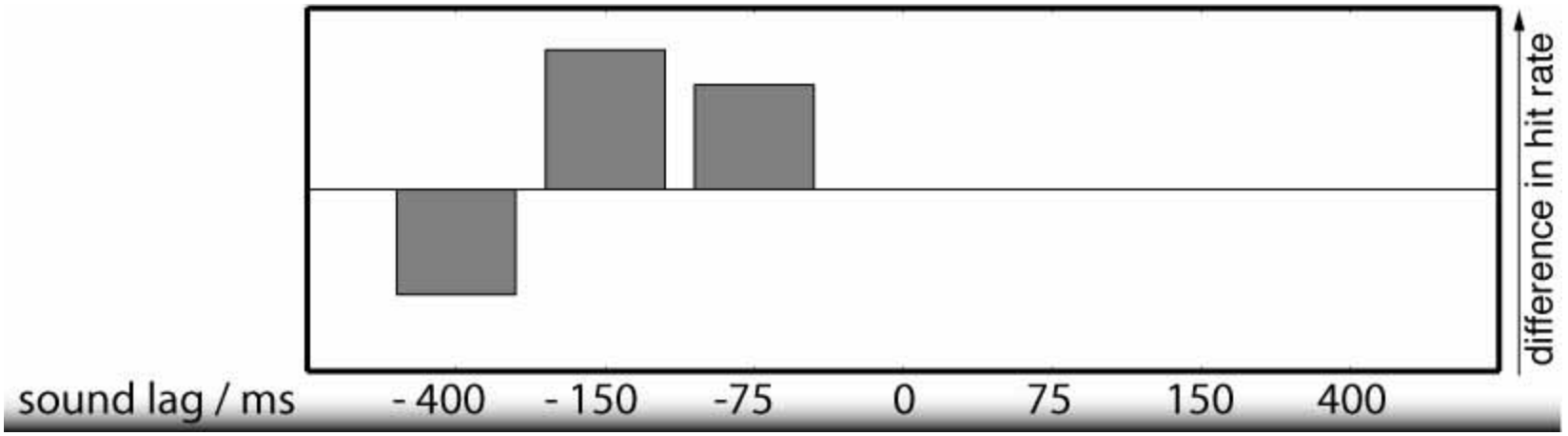
Transient hypothesis

- A true perceptual integration of rapid transients in the intensity of auditory and visual signals
- In excellent agreement with physiological studies of the Superior Colliculus
- These studies predict a temporal window of integration of 100 ms
- This can be tested by varying the audiovisual SOA
  - Should eliminate uncertainty reduction

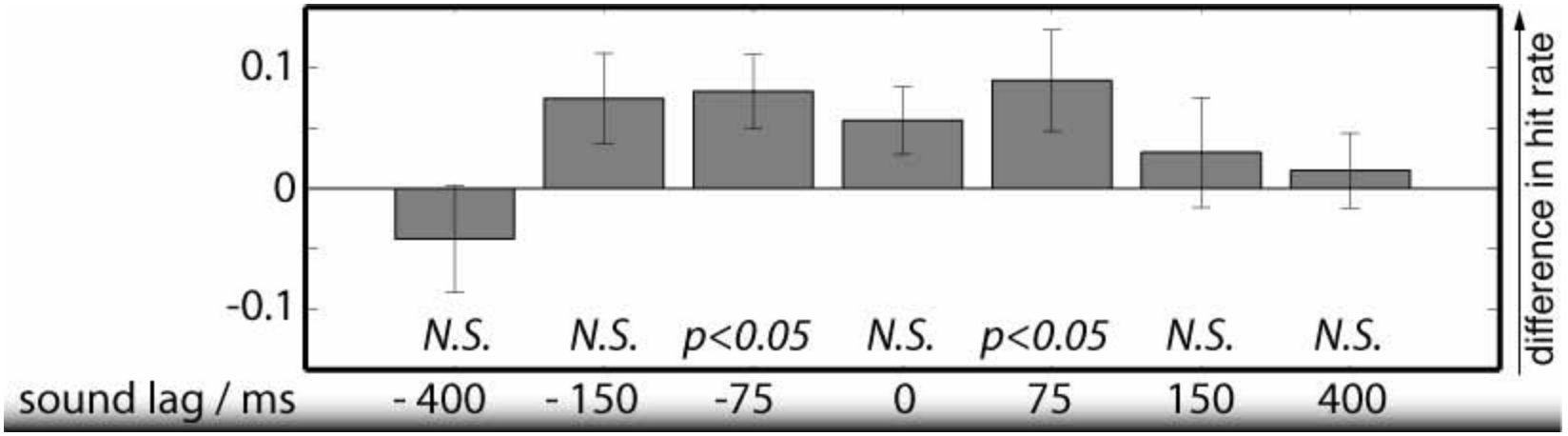
Predictions
Transient interactions



Predictions
Attentional cueing



Results


0.1
0
-0.1
N.S.
N.S.
p<0.05
N.S.
p<0.05
N.S.
N.S.
difference in hit rate
sound lag / ms
- 400
- 150
-75
0
75
150
400

Conclusions

- Sound intensity increase visual sensitivity
  - when lagging with 75 ms but not when lagging 150 ms
    - Cannot be due to exogenous attention
  - When stimulus asynchrony varies randomly
    - Cannot be due to reduction of uncertainty

- In good agreement with response properties of SC neurons

Summary

- Categorical audiovisual perception
  - Special: Strong, non-linear effects
    - Tricky to model!
    - Needs regularization
  - Not so special
    - Information reliability
    - Modality appropriateness
    - Continuous quantitative models apply
      - When adding a response boundary
      - Provides predictive power when regularized
  - McGurk Depends on top-down effects (Speech mode)
  - Multi-dimensional (multi-faceted)

Summary

- Audiovisual integration in signal detection
  - Based on transients
    - Not on intensity
  - Separable from attentional cueing
    - And reduction of temporal uncertainty