Any questions?

- Practicalities?
- Any open issues from yesterday?

Lecture 5: *audiovisual prosody*

Bosker, H. R., & Peeters, D. (2021). Beat gestures influence which speech sounds you hear. *Proceedings of the Royal Society B*, 288(1943), 1–9. doi:10.1098/rspb.2020.2419.

Hans Rutger Bosker

Speech Perception in Audiovisual Communication [SPEAC] lab

Donders Institute, Radboud University, Nijmegen, The Netherlands

https://hrbosker.github.io

hansrutger.bosker@donders.ru.nl

Listening with your eyes!

- We can ‘see’ speech from articulatory movements alone
    - Lip-reading / speech-reading
- We use visual articulatory signals to inform speech perception
    - Audiovisual enhancement of speech-in-noise
    - Classic McGurk effect

SPEAC

McGurk & MacDonald, 1976; Sumby & Pollack, 1954

Listening with your eyes!

- Classic McGurk effect

McGurk & MacDonald, 1976

Listening with your eyes!

- Classic McGurk effect

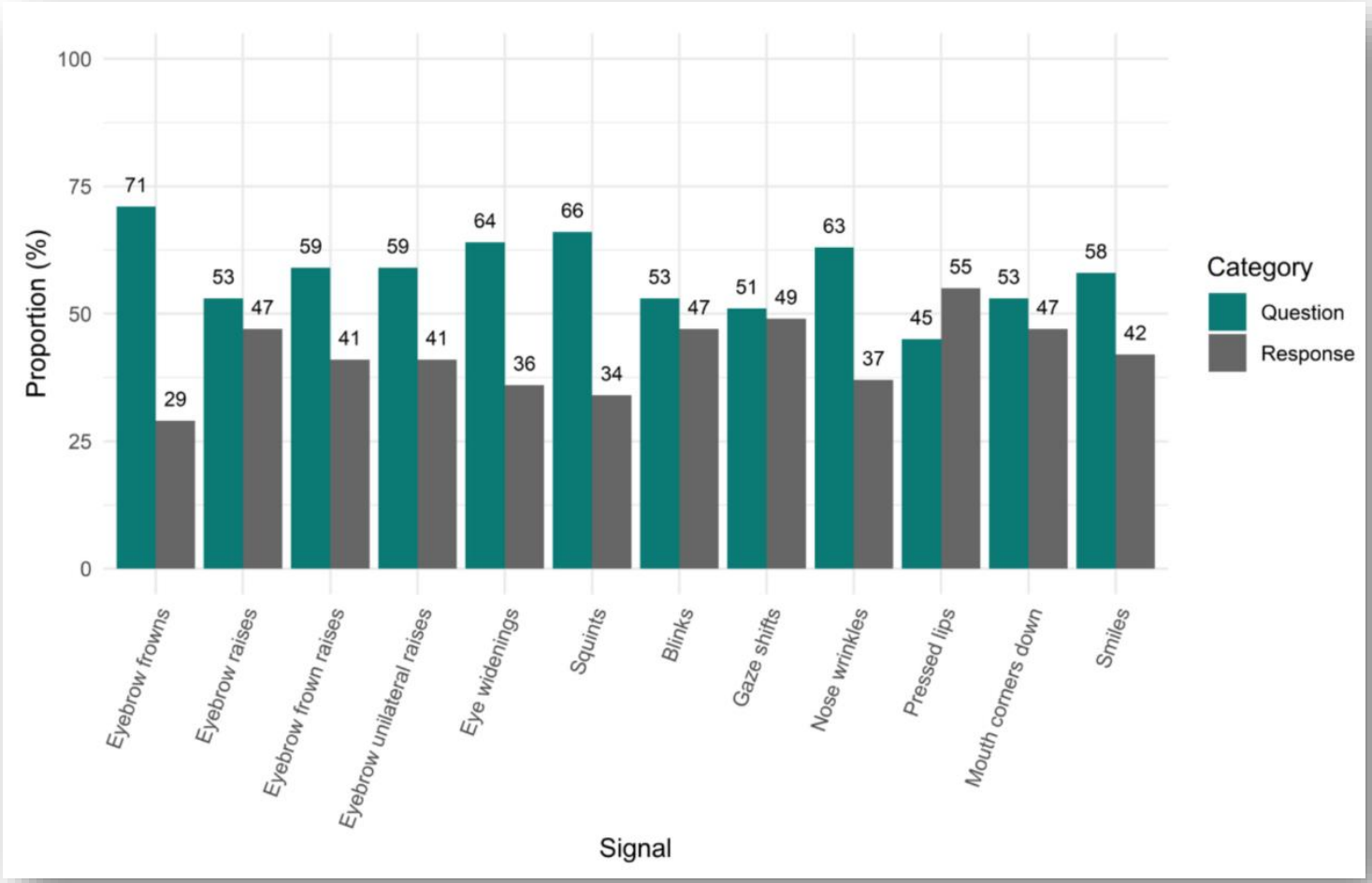
McGurk & MacDonald, 1976

Listening with your eyes!

- Classic McGurk effect

Can we *see* prosody?

McGurk & MacDonald, 1976



Nota et al., 2021

Can we *lip-read* prosody?

- Speech rate estimates are equally accurate in V-only vs. A-only (Green, 1987)
- ??? intensity ???
- Phonological contrasts that primarily rely on *f*0 are much harder to lip-read:
  - Lexical tone
    - Video-only accuracy is at chance (Burnham et al., 2001)
    - …but V-only training can help (Chen & Massaro, 2008)
  - Lexical stress
    - Video-only accuracy above ch
    - …and Dutch (less segmental r

Do we use this info?

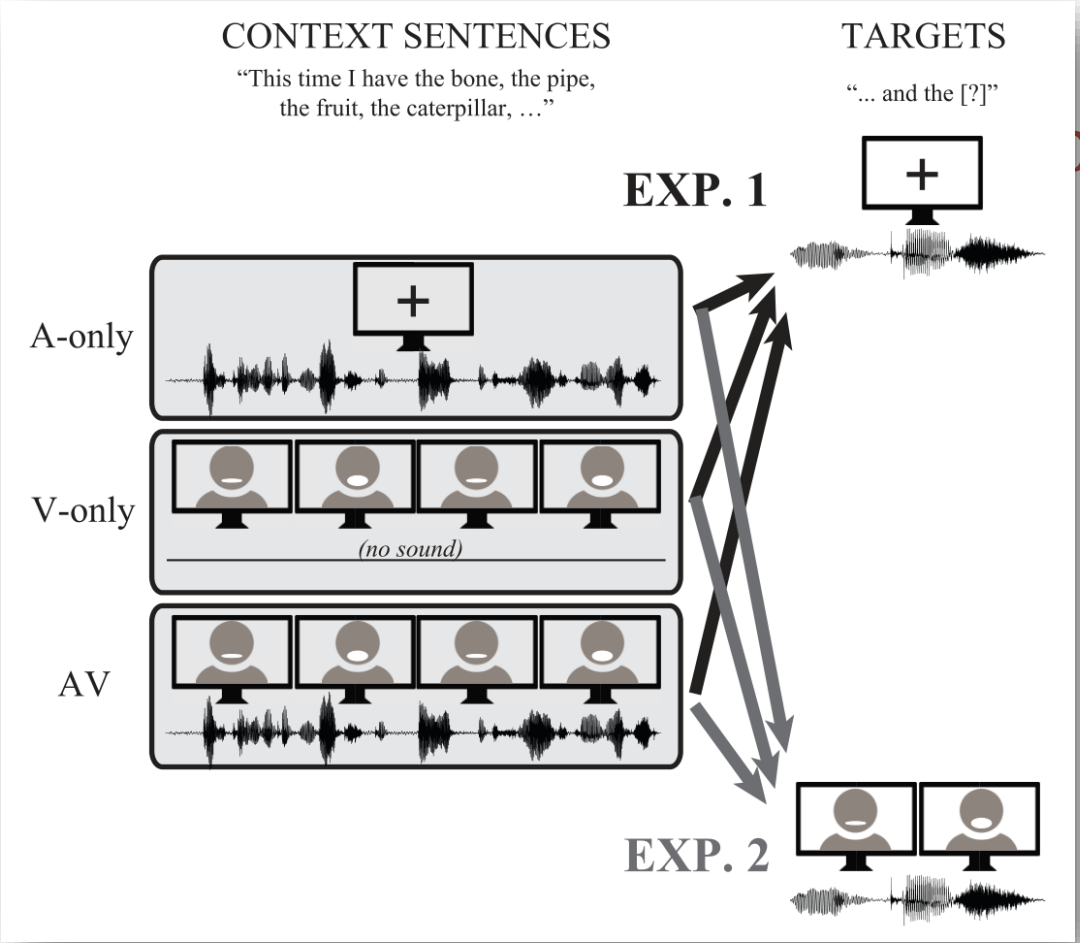
Do we *use* visual articulatory cues to prosody in audiovisual speech perception?

- Speech rate
  - A video of a fast /pi/ + /bi-pi/ auditory continuum: +/p/ responses

    Green & Miller, 1985
  - A video of a fast /ti/ + /bi-pi/ auditory continuum: +/p/ responses

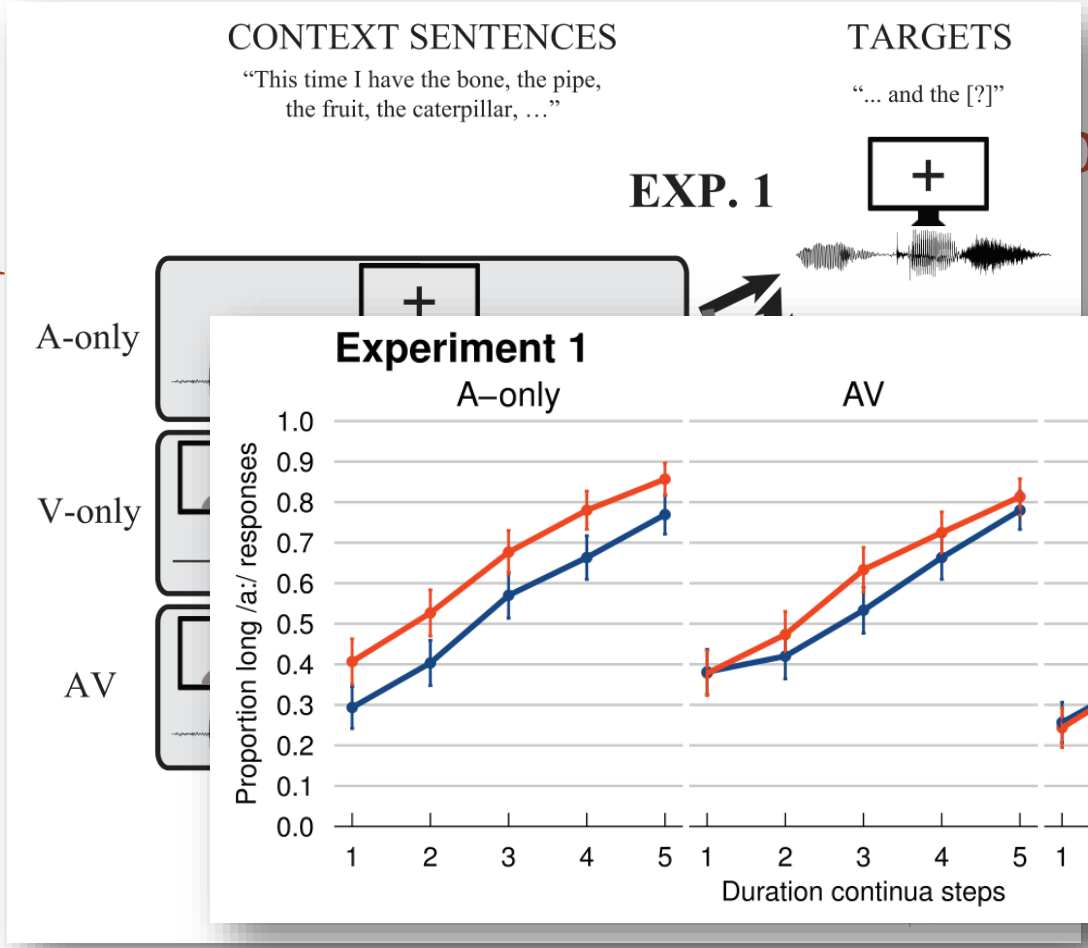
    Brancazio & Miller, 2005
  - Rate normalization by visual speech rate?

Do we in aud rosody



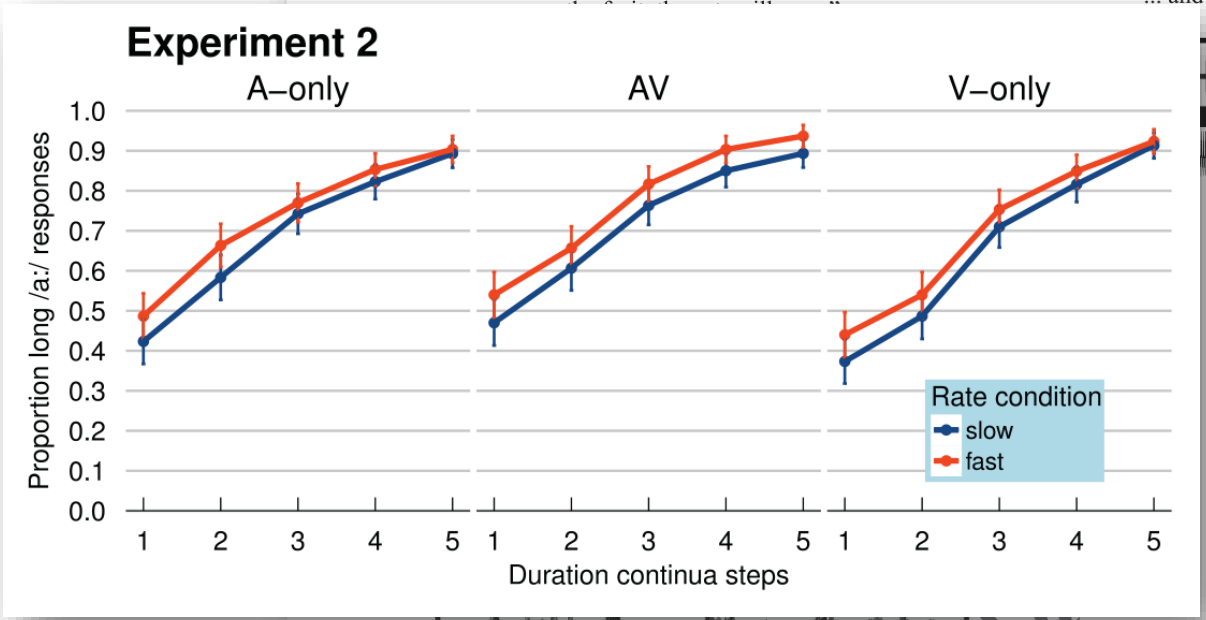
Bosker et al., 2020, *QJEP*

Do we … in aud… prosody



Bosker et al., 2020, *QJEP*

rosody



CONTEXT SENTENCES

"This time I have the bone, the pipe,

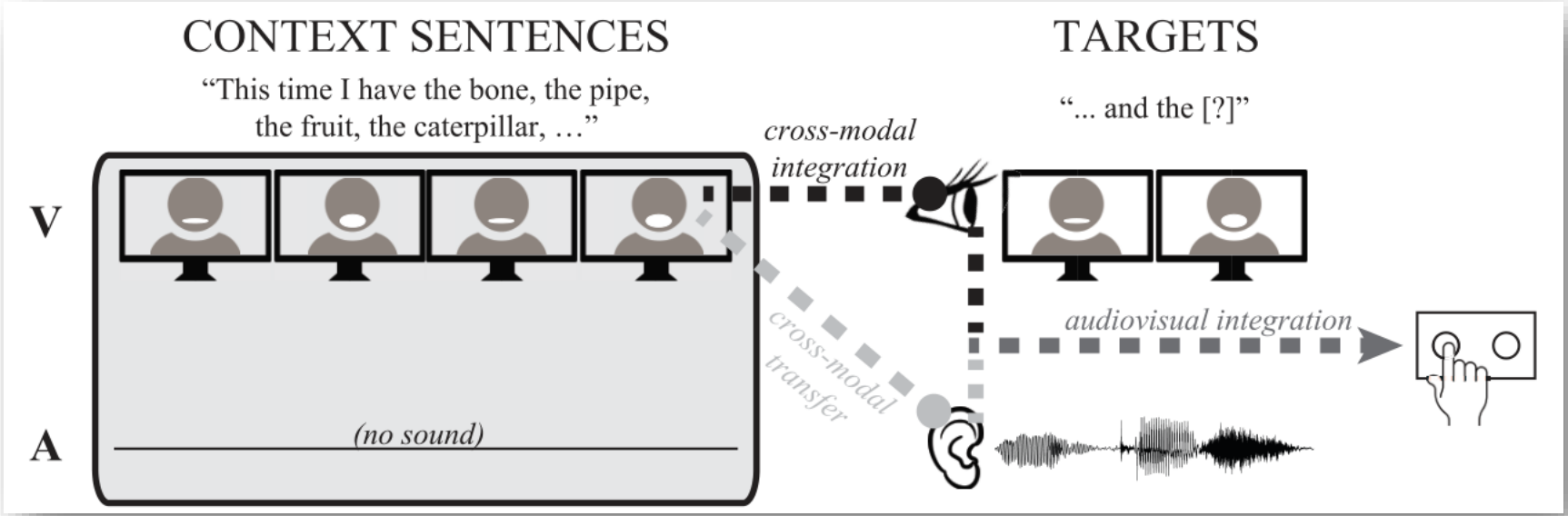
TARGETS

"... and the [?]"

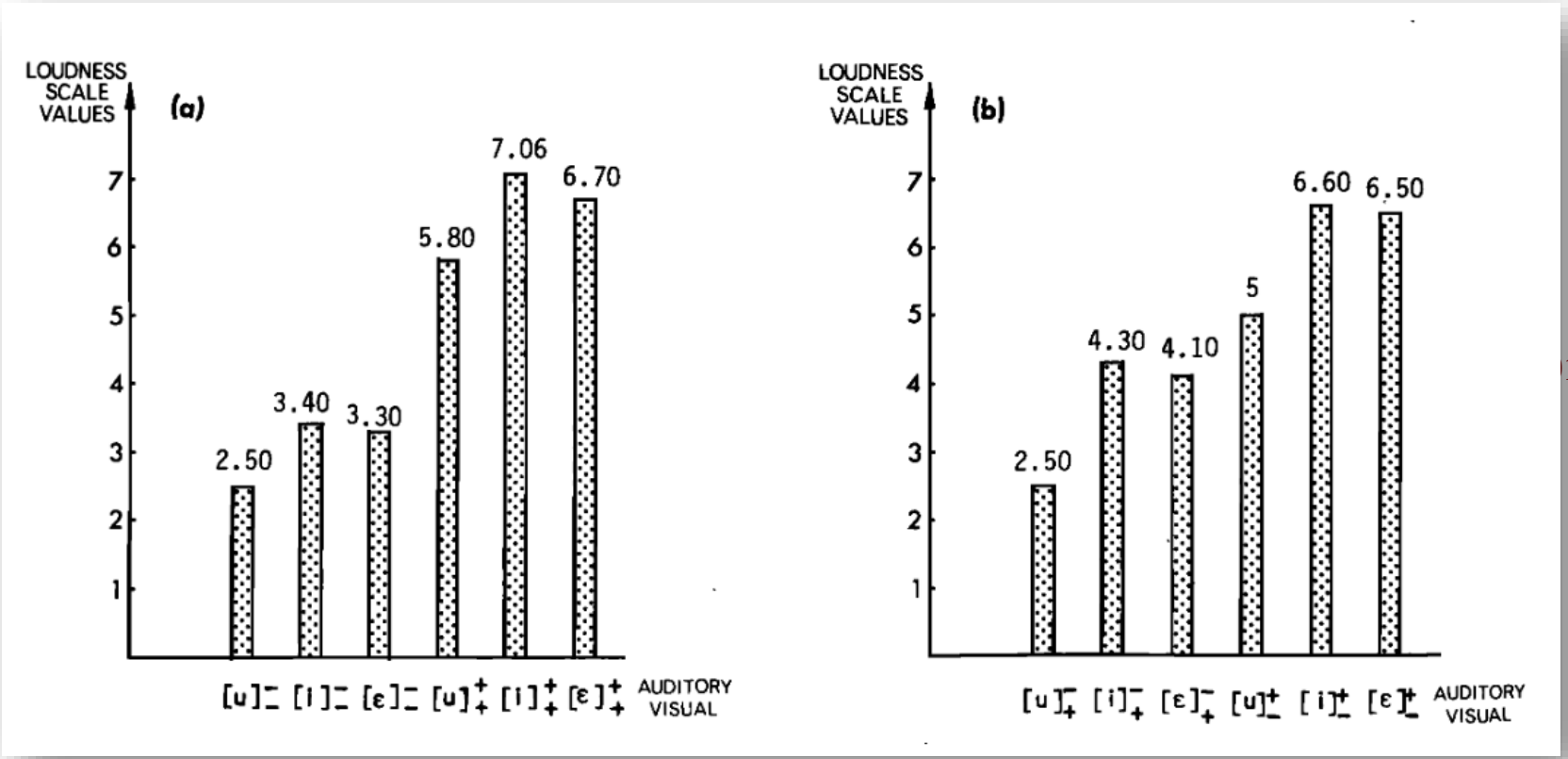
EXP. 2

Bosker et al., 2020, *QJEP*

Do we *use* visual articulatory cues to prosody in audiovisual speech perception?

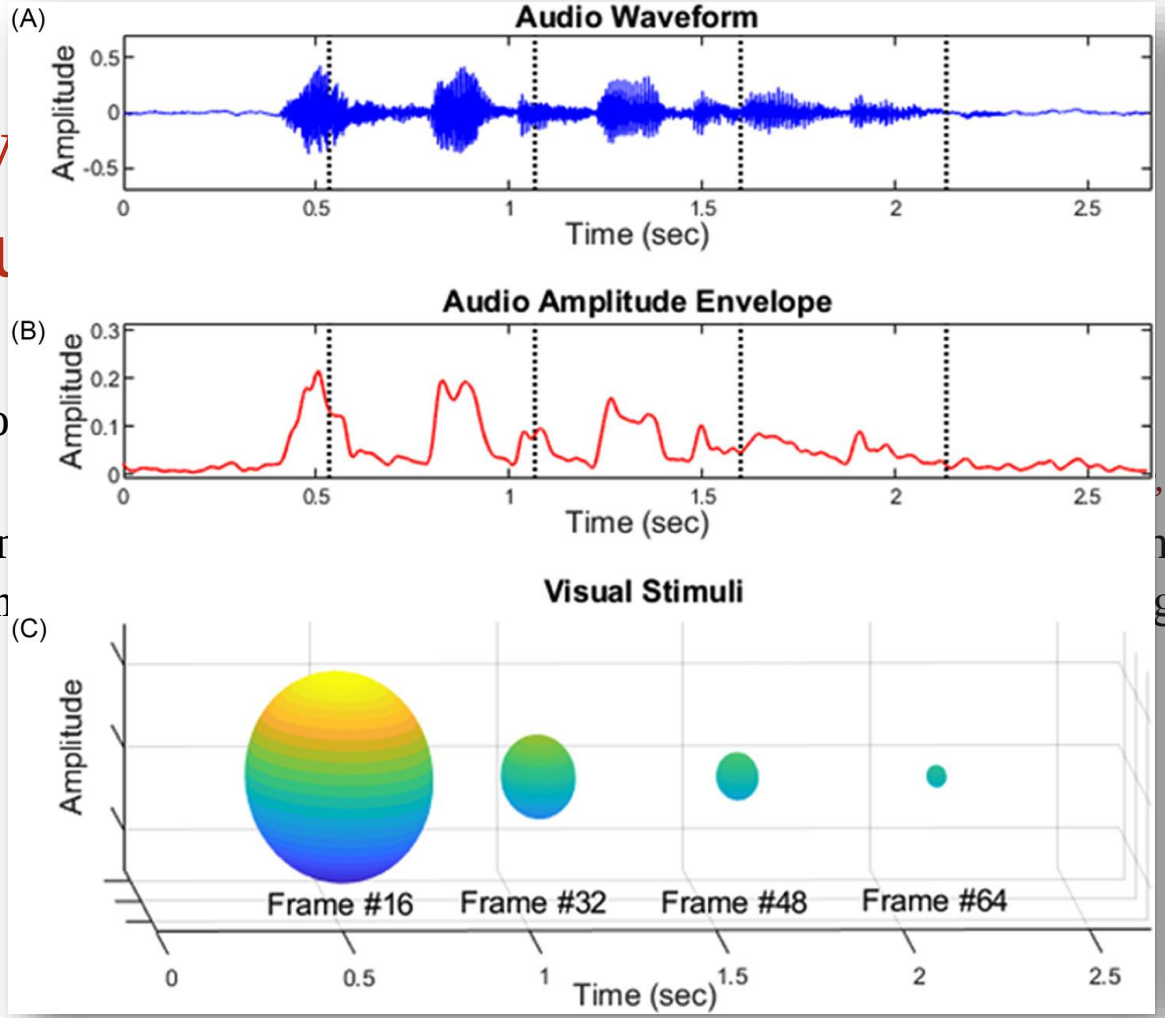


Bosker et al., 2020, *QJEP*


LOUDNESS SCALE VALUES
(a)
7.06
6.70
5.80
3.40
3.30
2.50
AUDITORY
VISUAL
LOUDNESS SCALE VALUES
(b)
6.60
6.50
5
4.30
4.10
2.50
AUDITORY
VISUAL

Do we *use* v
in audiovisu

- Intensity
  - Seeing vo
  - Seeing a r
    of speech



, 1991

nm)

g'?) …

Do we *use* visual articulatory cues to prosody in audiovisual speech perception?

- Intensity
  - Seeing vocal effort boosts perceived loudness

    Glave & Rietveld, 1979; Rosenblum & Fowler, 1991
  - Seeing a modulating circle in sync with the amplitude envelope (rhythm) of speech boosts its intelligibility… (…link to neural 'speech tracking'?)

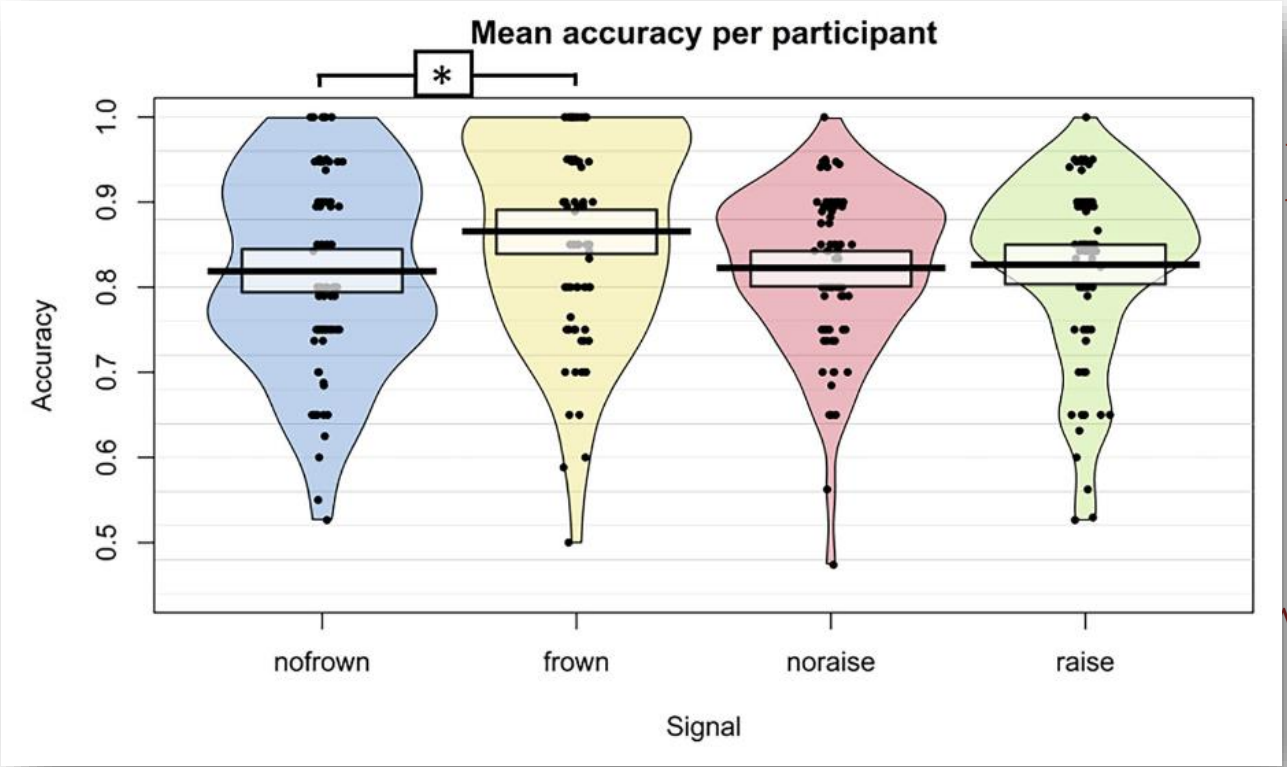
    Yuan et al., 2020
  - …or does it?

    Strand et al., 2020

Do we *use* visual articulatory cues to prosody in audiovisual speech perception?

- *f*0
    - Sentence accent
        - Which word in this sentence is the most prominent?
        - Slow down in RTs when V is incongruent with A

Swerts & Krahmer, 2008



prosody

verts & Krahmer, 2008

Nota et al., 2022

Is this driven by articulatory cues?

Do we *use* visual articulatory cues to prosody in audiovisual speech perception?

- *f*0
  - S…
  - …spectral energy in female talkers

Other visual cues to *f*0?

Jongman et al., 2000

  - Same fricative on /ʃ-s/ continuum more /ʃ/-like if female speech

Mann & Repp, 1980

  - Just **seeing** a female face also leads to more /ʃ/-responses

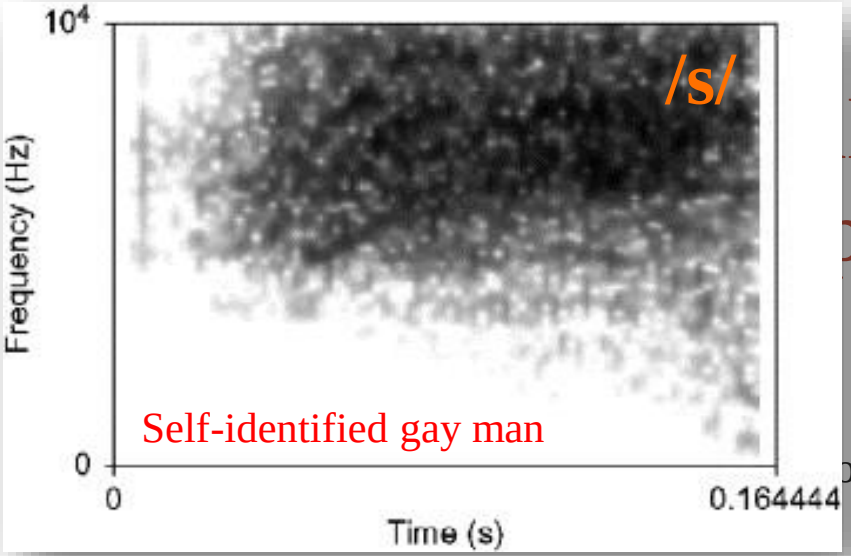
Strand et al., 1996

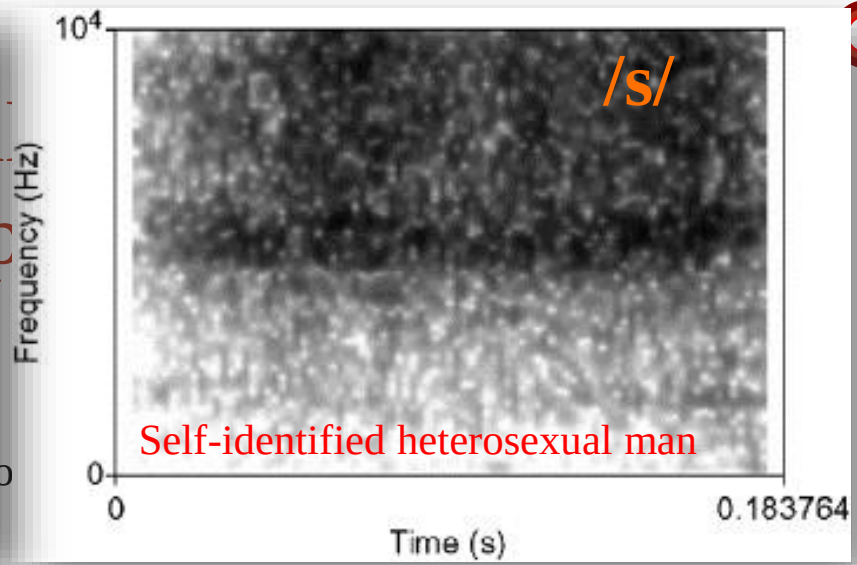
  - Even **imagining** listening to a female talker induces normalization

    ~ prior knowledge?

Johnson et al., 1999

SPEAC





- Common societal impression: higher-frequency spectral energy in gay talkers + lower-frequency spectral energy in lesbian talkers

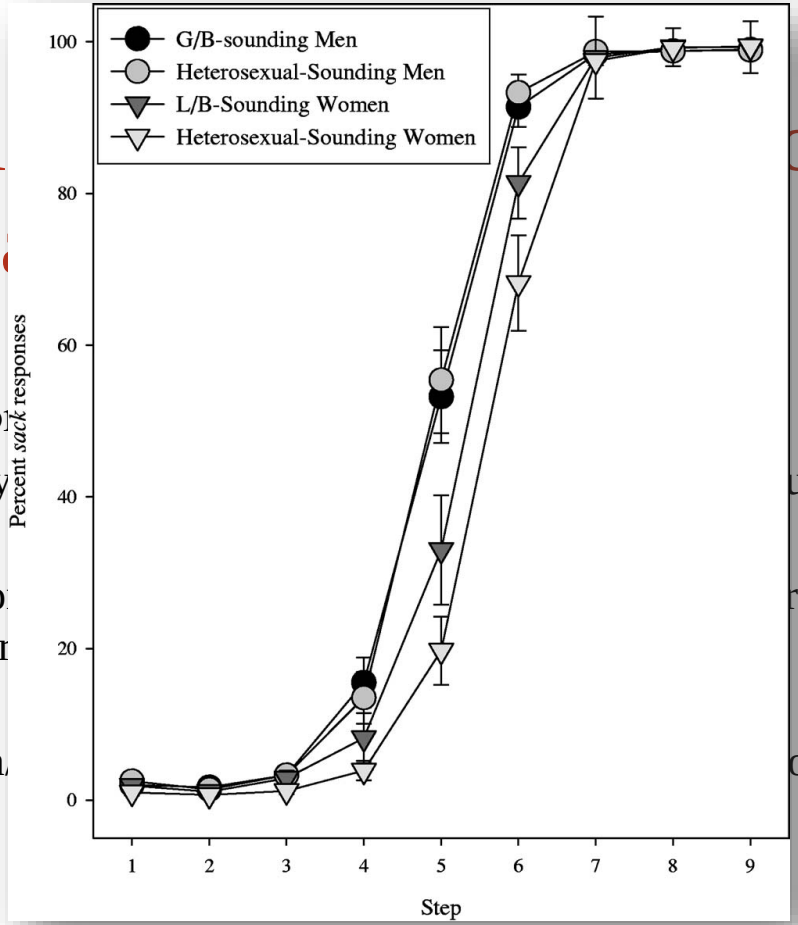
Munson et al., 2006, *JPhon*

- Lesbian/bisexual-sounding talkers elicit more /s/-responses

Munson et al., 2006, *JASA*

Do we *use* vi… …o prosody in audiovisu…

- *f*0
  - Spectral no…
    - Visually… …users
      Winn et al., 2013
    - Commo… …ral energy in gay talkers
      + lower
      Munson et al., 2006, *JPhon*
    - Lesbian… …onses
      Munson et al., 2006, *JASA*



Do we *use* visual articulatory cues to prosody in audiovisual speech perception?

- *f*0
  - Lexical tone
    - AV accuracy = A accuracy in Mandarin

      Burnham et al., 2001
    - A ye1 + V ye2 = A ye1 (incongruent AV stimuli)

      Hannah et al., 2017
    - AV advantage only surfaces when is degraded (noise/vocoded speech)

      Mixdorff et al., 2005; Burnham et al., 2015
    - Perhaps no surprise because hardly any visual cues to lexical tone (V-only)

      Burnham et al., 2001

Do we *use* visual articulatory cues to prosody in audiovisual speech perception?

- *f*0
  - Lexical stress
    - Visual cues to stress are detectable in production…

      Scarborough et al., 2009
    - …and people can ‘lip-read’ stress from V-only videos…

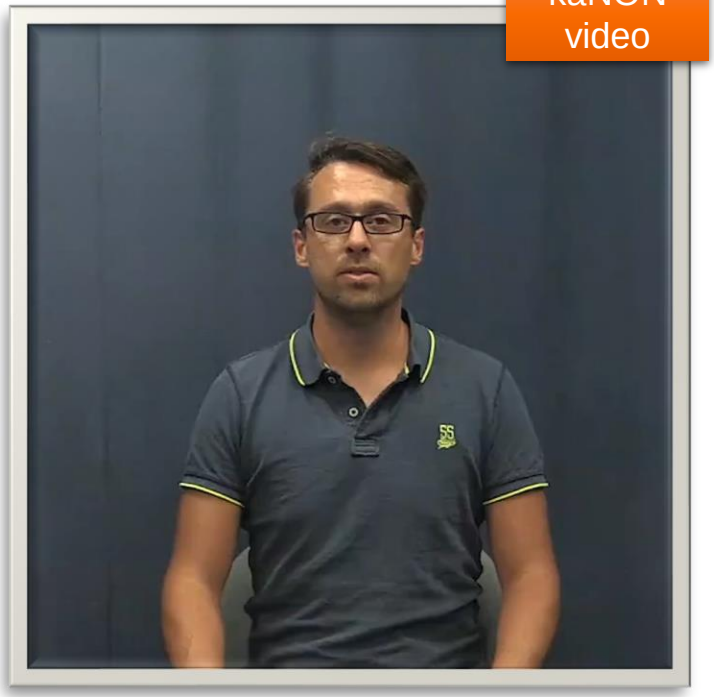
      Scarborough et al., 2009; Jesse & McQueen, 2014
    - …so surely people should use these cues in AV speech perception, right?

Prosody on the face

Prosody on the face

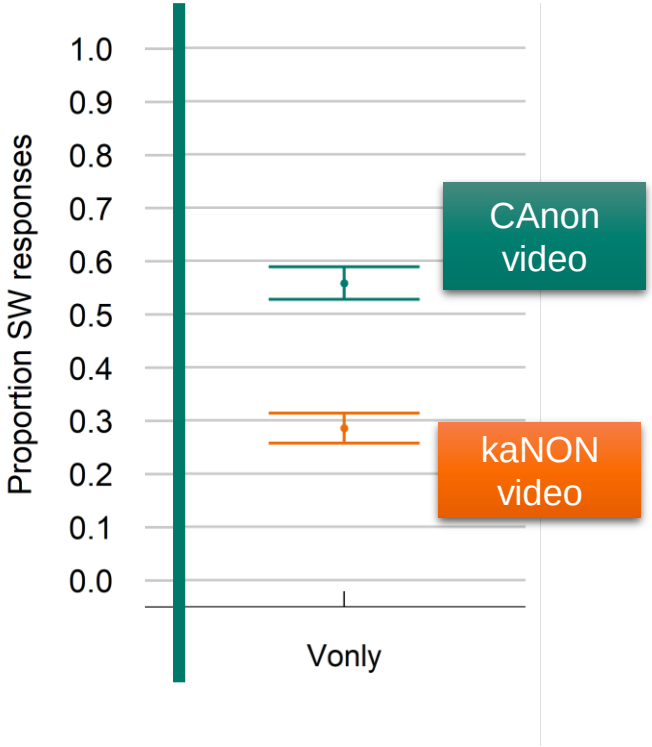
CAnon video

kaNON video



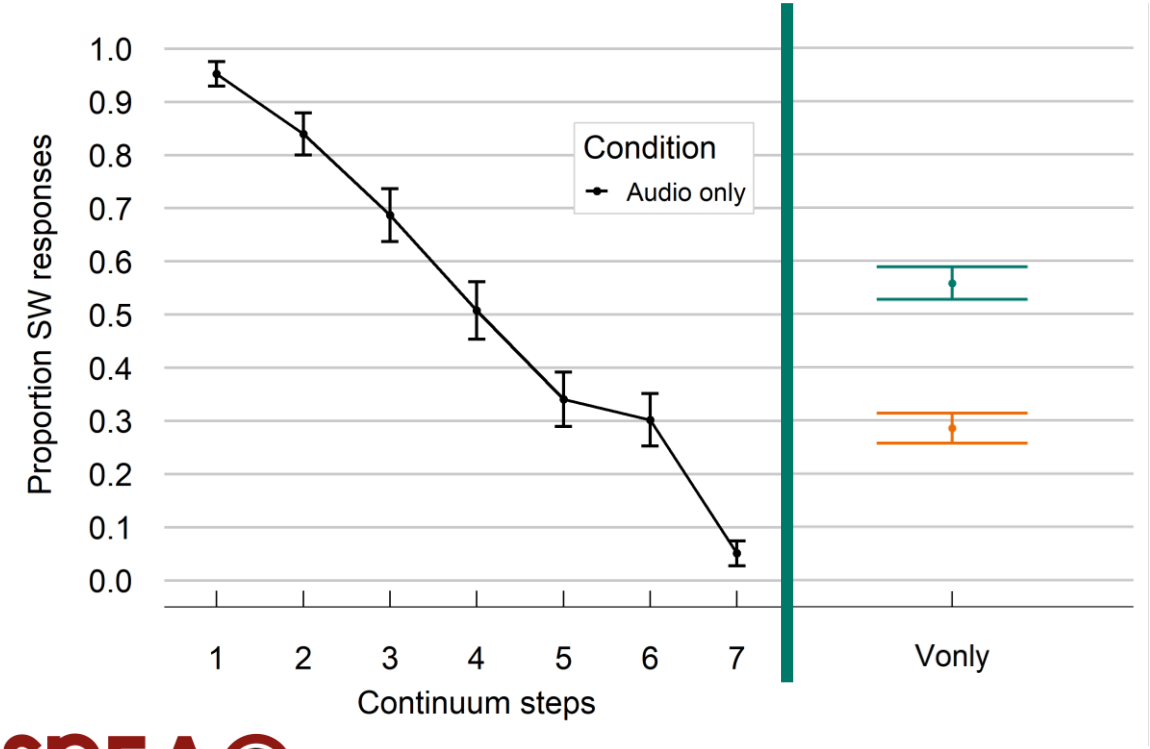
Bujok et al., 2024, *L&S*

Prosody on the face



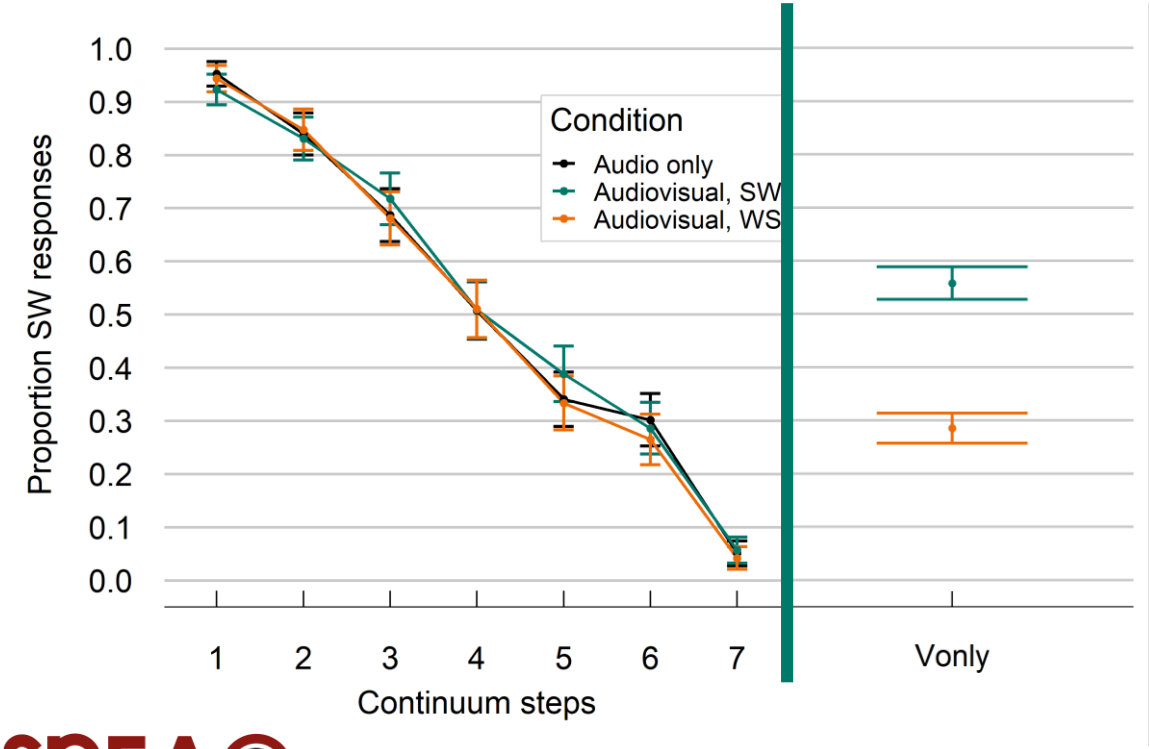
Bujok et al., 2024, *L&S*

Prosody on the face



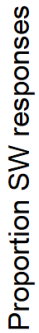
Bujok et al., 2024, *L&S*

Prosody on the face



Bujok et al., 2024, *L&S*

Prosody on the face



Audio step

Mok, Bujok, & Bosker, *in prep.*

Do we *use* visual articulatory cues to prosody in audiovisual speech perception?

- *f*0
  - Lexical stress
    - Visual articulatory cues on the face hardly contribute to AV speech perception
      Bujok et al., 2024, *L&S*
    - This is in stark contrast to how people use visual segmental cues (McGurk)
    - Low reliability of visual prosody cues leads people to down-weight these cues in quiet…
    - …but up-weigh their contribution when listening is hard.

Are there any other visual cues to prosody?

Prosody in the hands

- We move to the rhythm of our speech

"... and when and how and on – what – ba sis..."

Prosody in the hands

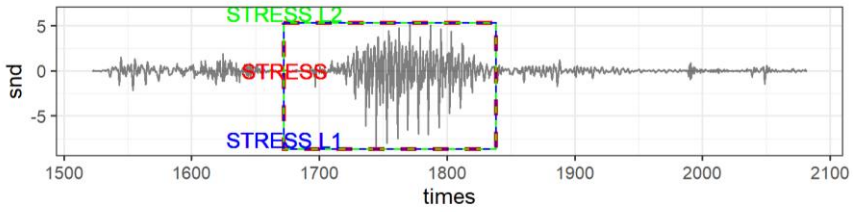
- We move to the rhythm of our speech
- Relatively simple up-and-down hand movements make up over 90% of our everyday gestures

Shattuck-Hufnagel & Ren, 2018

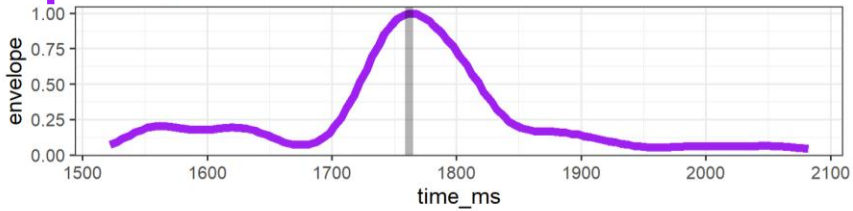
- Closely coupled to speech prosody, typically falling on stressed syllables
- Gesture's apex time-locked to pitch peak in stressed syllable.

Prosody in the hands

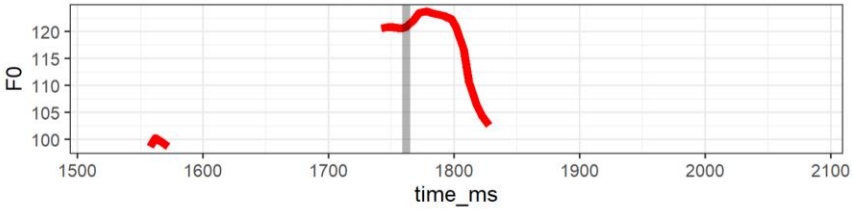
Spanish: “his.TÓ.ri.co”



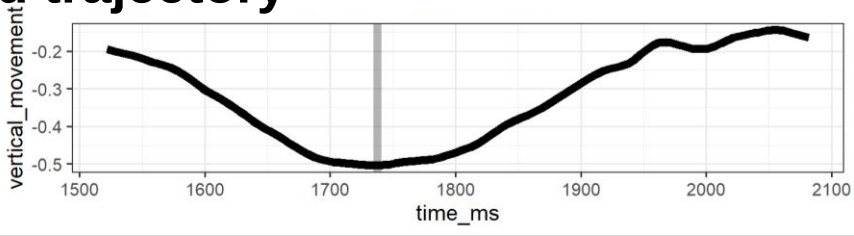
amplitude



f0



hand trajectory



Prosody in the hands

- Why do we beat on the stressed syllable?

1. Universal biomechanics:
   moving your limbs raises your voice (amplitude, *f*0)

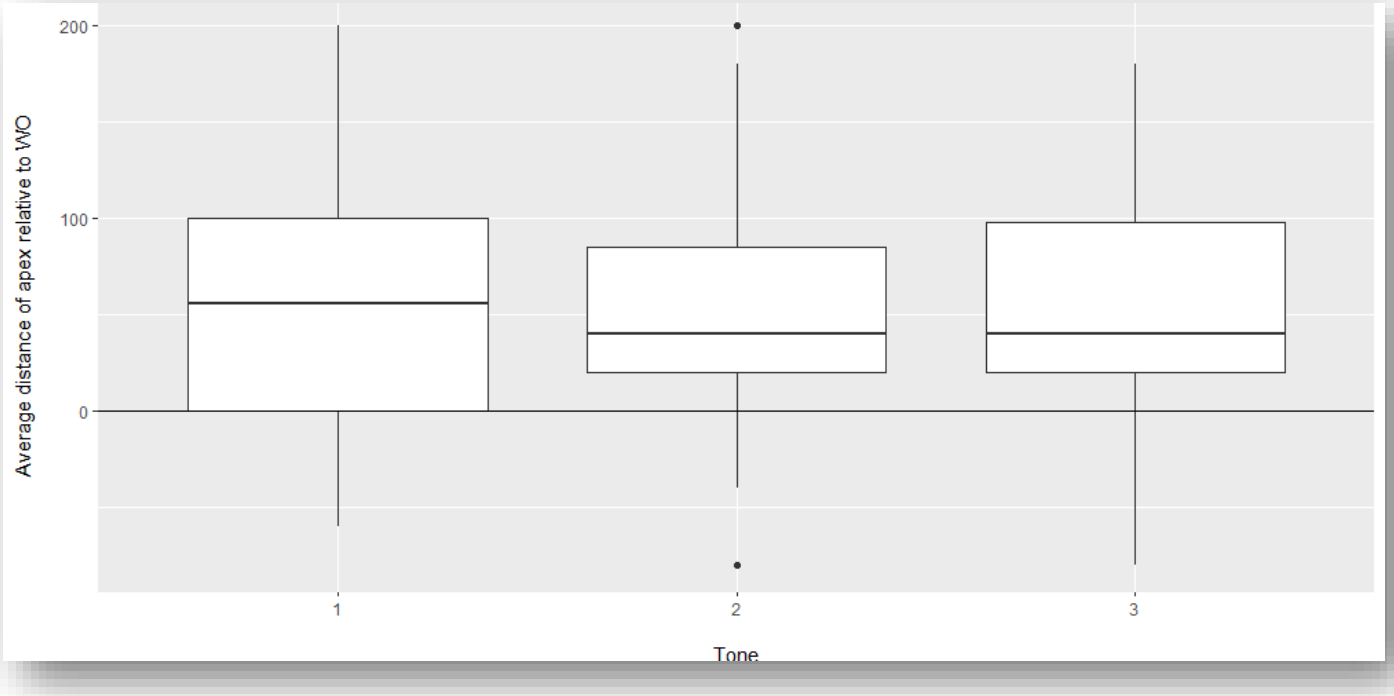
Pouw et al., 2020

2. Language-specific multimodal phonology:
   no evidence for coupling to *f*0 contour in tone languages

Rohrer et al., 2024

SPEAC

Prosody in the hands



Pouw et al., 2020

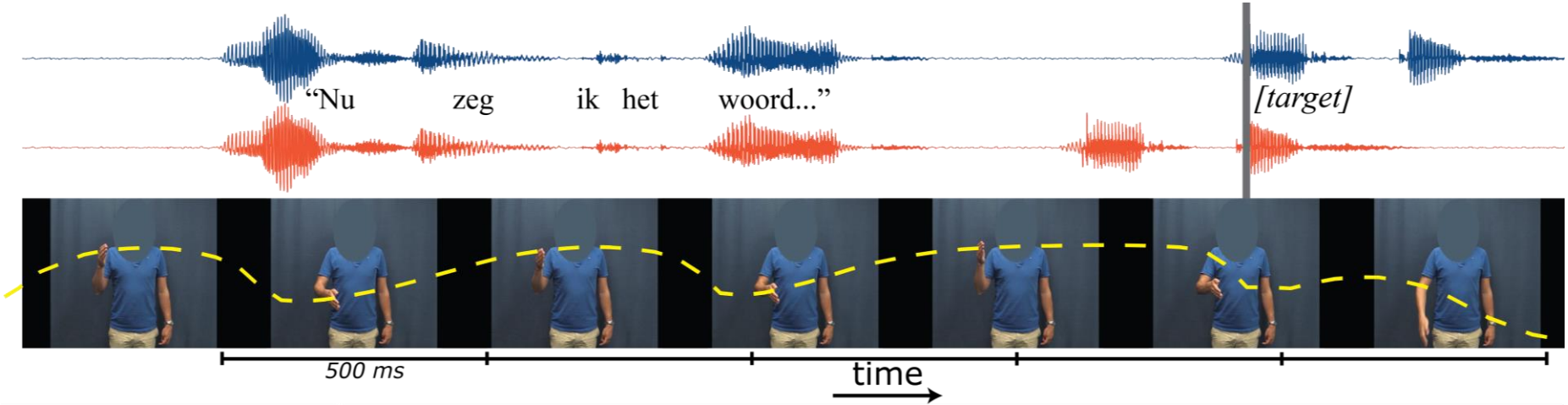
Rohrer et al., 2024

Prosody in the hands

- If talkers tend to gesture on stressed syllables…
- …can listeners use this visual temporal cue to perceive lexical stress?

Prosody in the hands

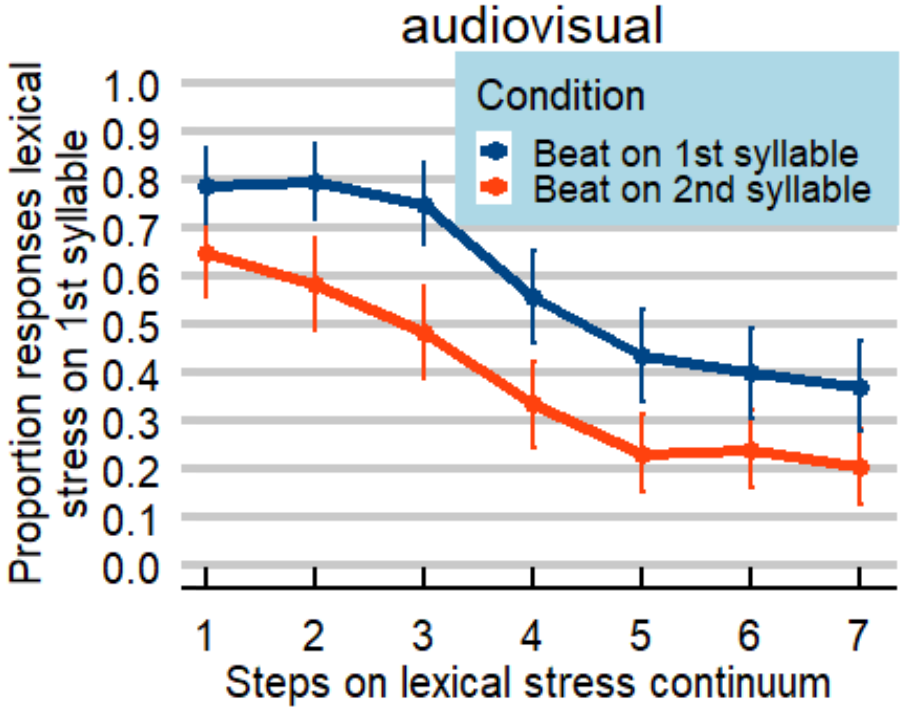
Prosody in the hands



Bosker & Peeters, 2021

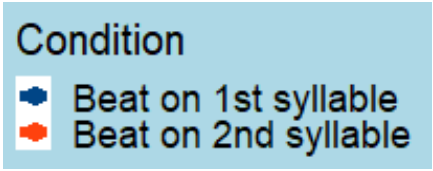
Prosody in the hands

- ‘manual McGurk effect’



Bosker & Peeters, 2021

Prosody in the hands



nonwords *real words* *shadowing* *vowel length*

A beat gesture on the first syllable of /b?x.pɪf/ will bias towards: short /b**ɑ**x.pɪf/ or long /b**aː**x.pɪf/

Bosker & Peeters, 2021

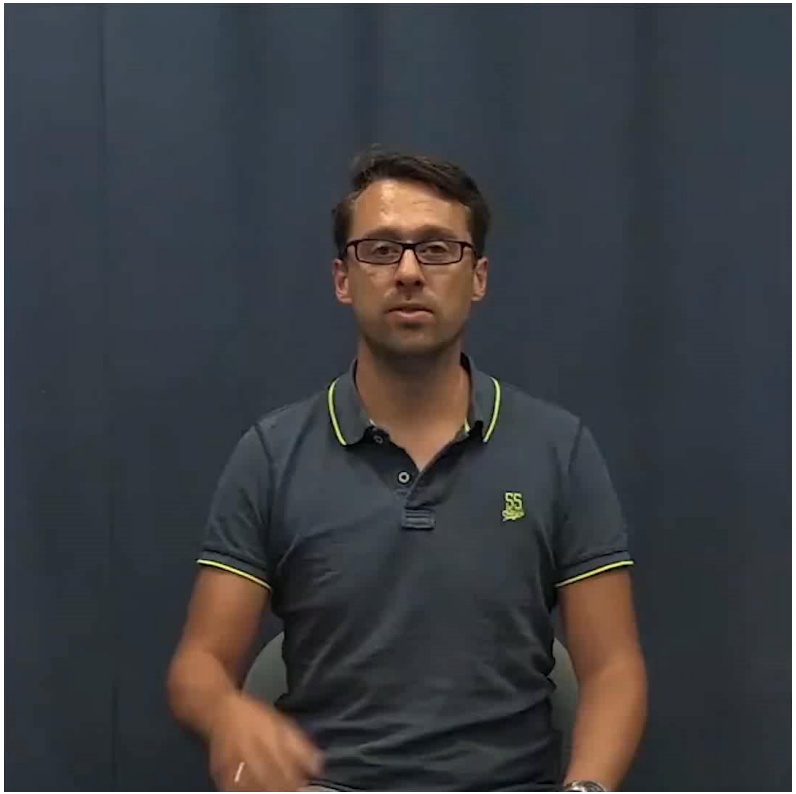
Prosody in the hands

head from no-gesture condition + *gesturing body* = *new fully-crossed stimuli*



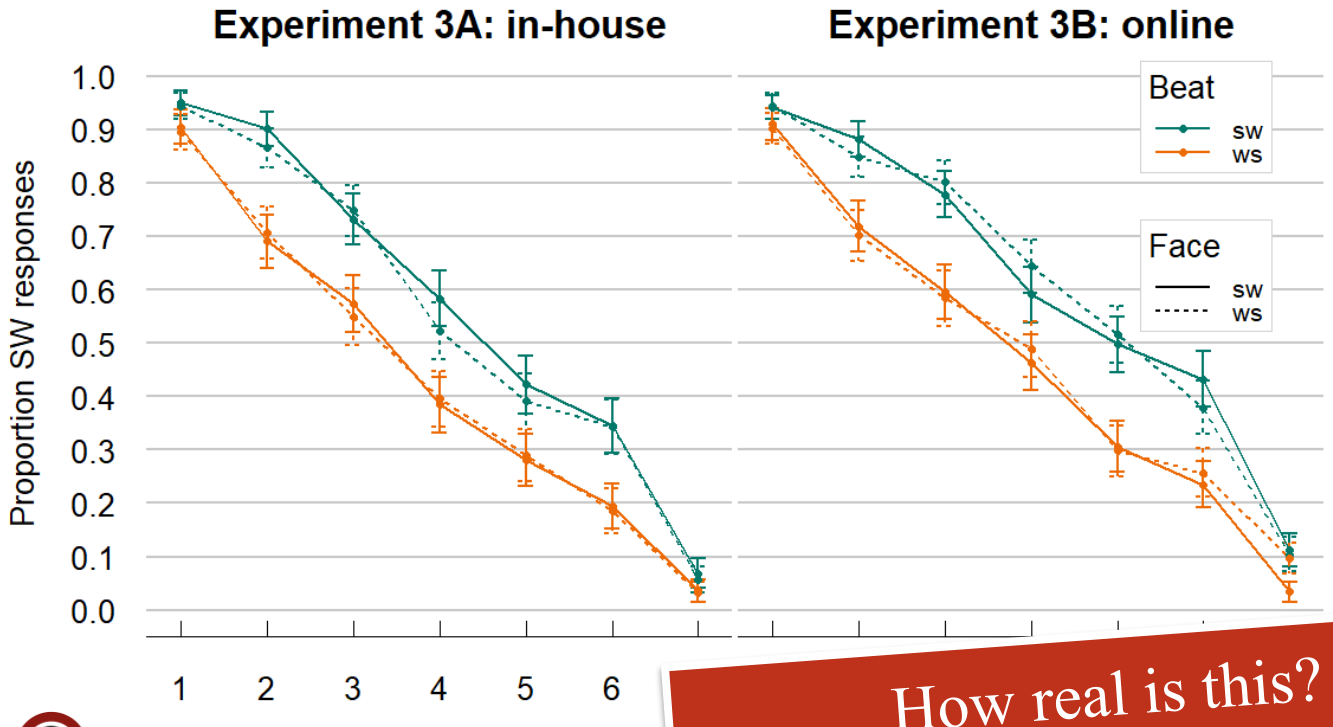
Bujok et al., 2024, *L&S*

Prosody in the hands



Bujok et al., 2024, *L&S*

Prosody in the hands

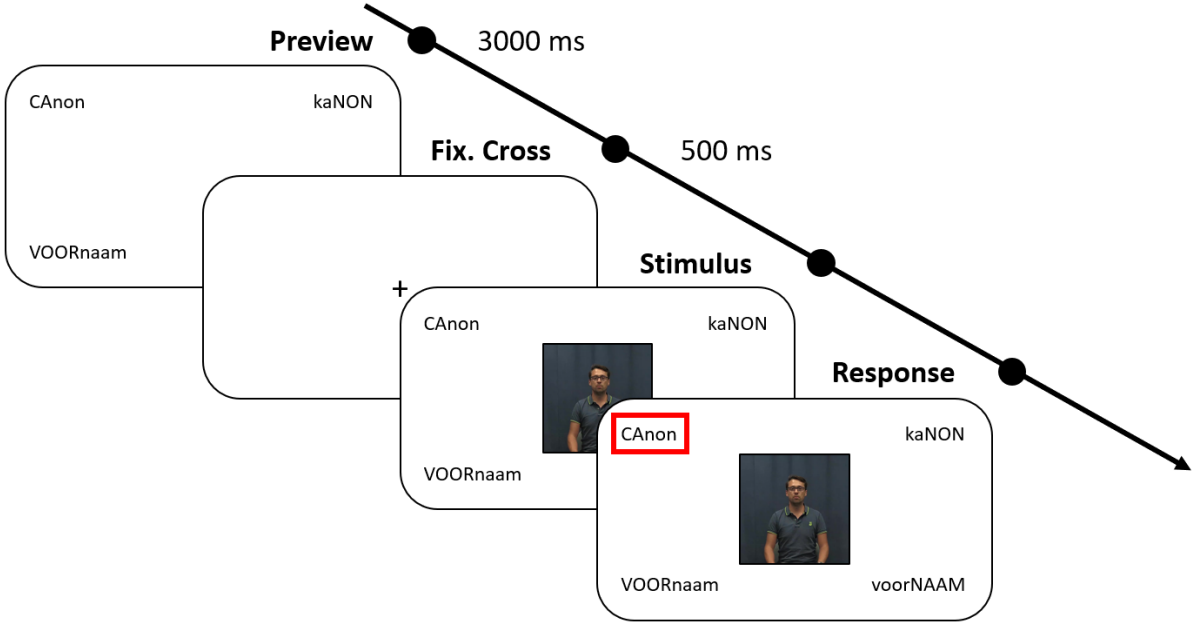


How real is this?

Bujok et al., 2024, *L&S*

Prosody in the hands: immediate impact

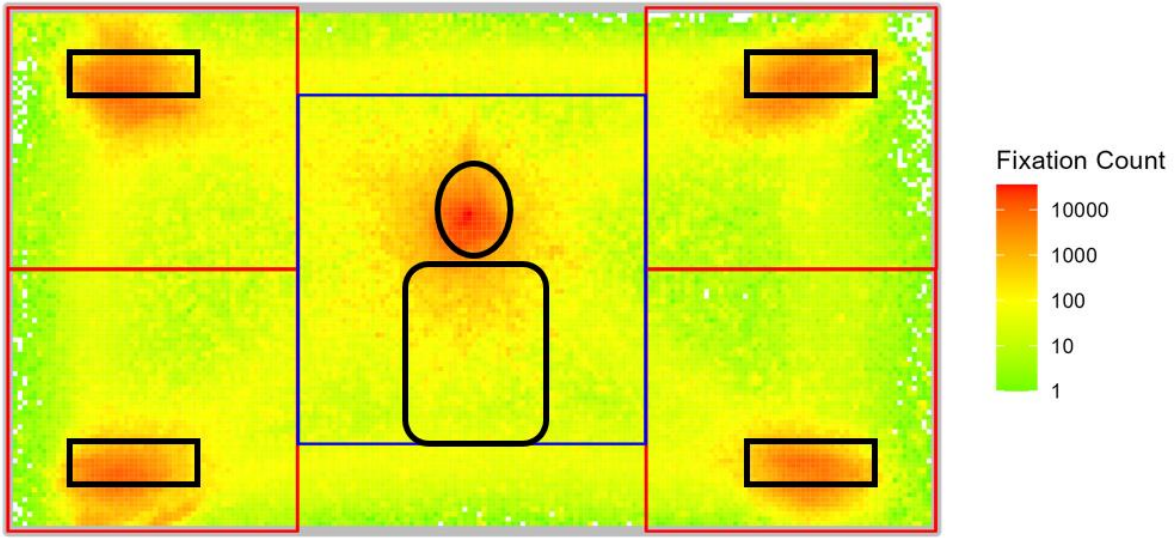
- Eye-tracking: visual world paradigm



Bujok, Meyer & Bosker, *in prep.*

Prosody in the hands: immediate impact

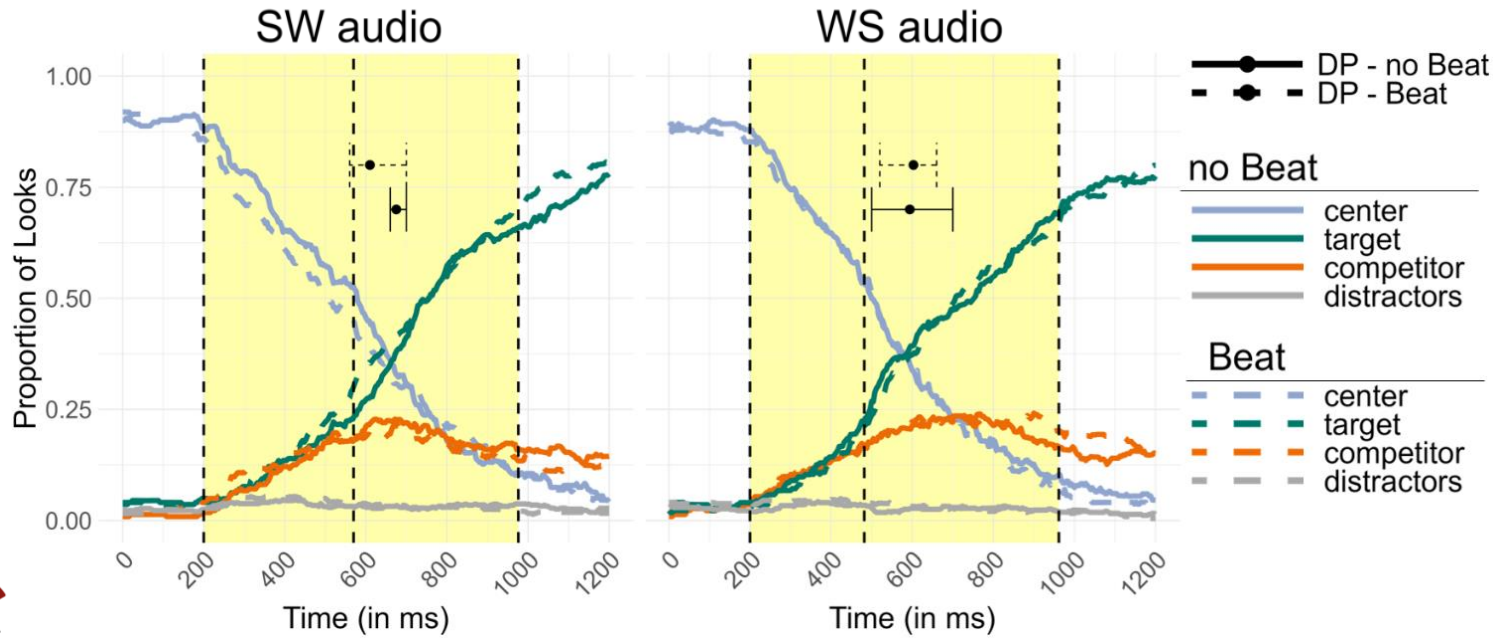
- Hardly any fixations on the gesture



Bujok, Meyer & Bosker, *in prep.*

Prosody in the hands: immediate impact

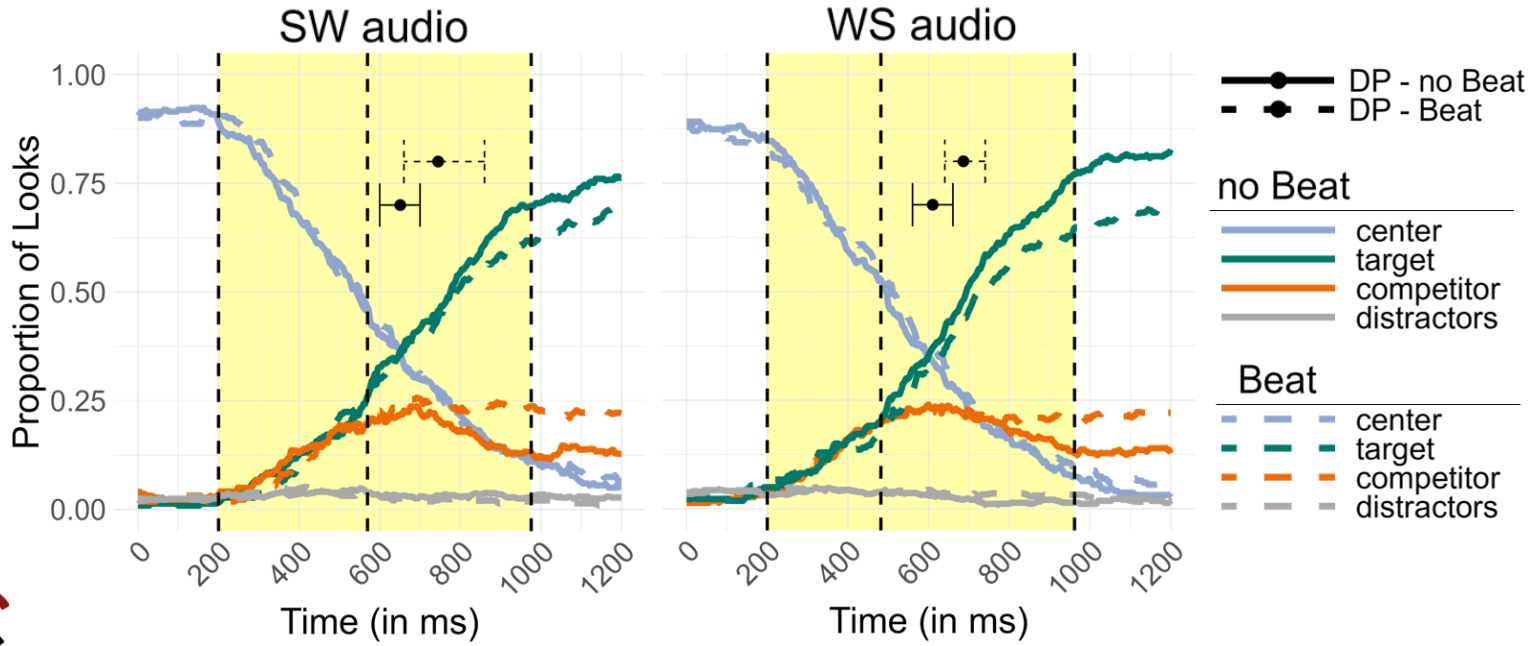
- When audio clear, no (reliable) beneficial effect of stress-congruent beat.



Bujok, Meyer & Bosker, *in prep.*

Prosody in the hands: immediate impact

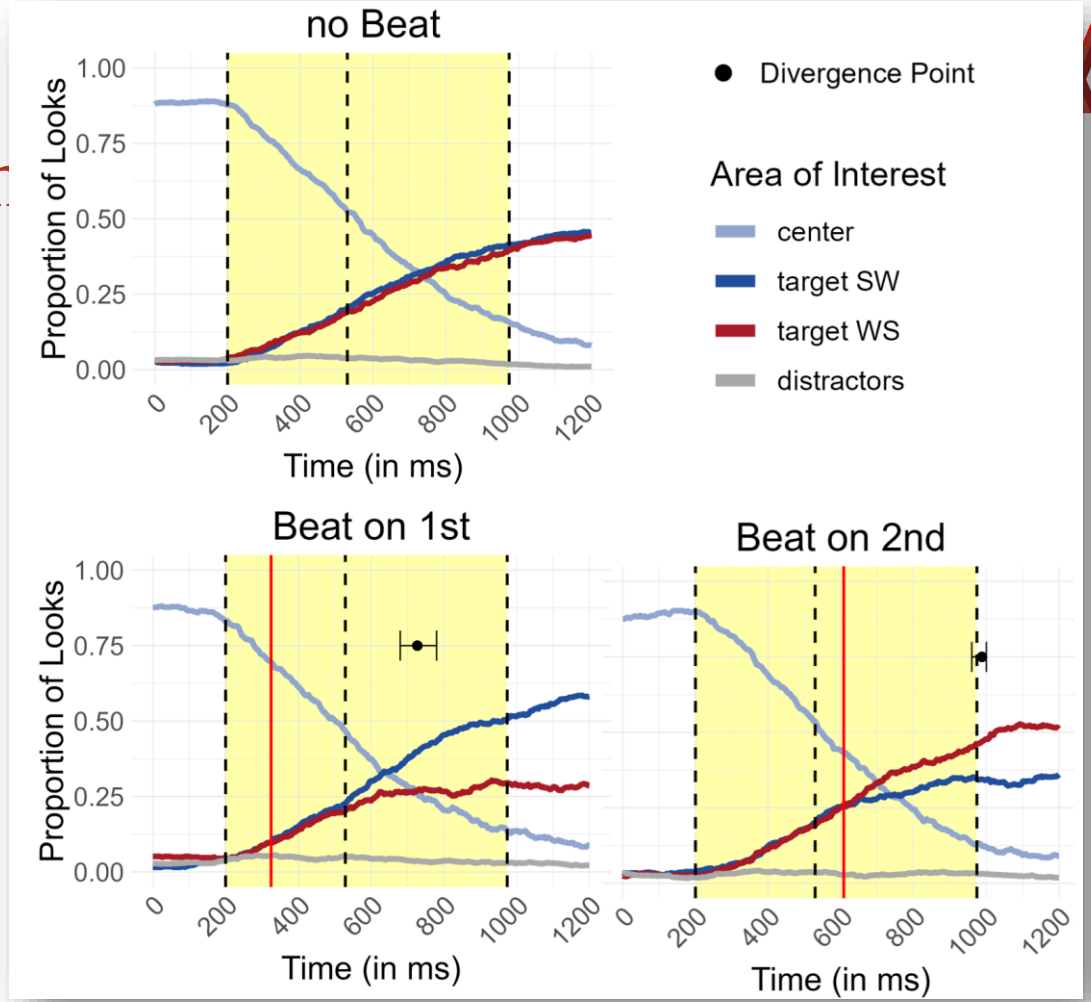
- When audio clear, reliable interfering effect of stress-**incongruent** beat.



Bujok, Meyer & Bosker, *in prep.*

Prosody in the har

- When audio ambiguous, beat-on-1 biases to SW, beat-on-2 biases to WS.
- Divergence point lies before word offset.
- Manual McGurk effect is an early perceptual bias, not only a decision-making effect.



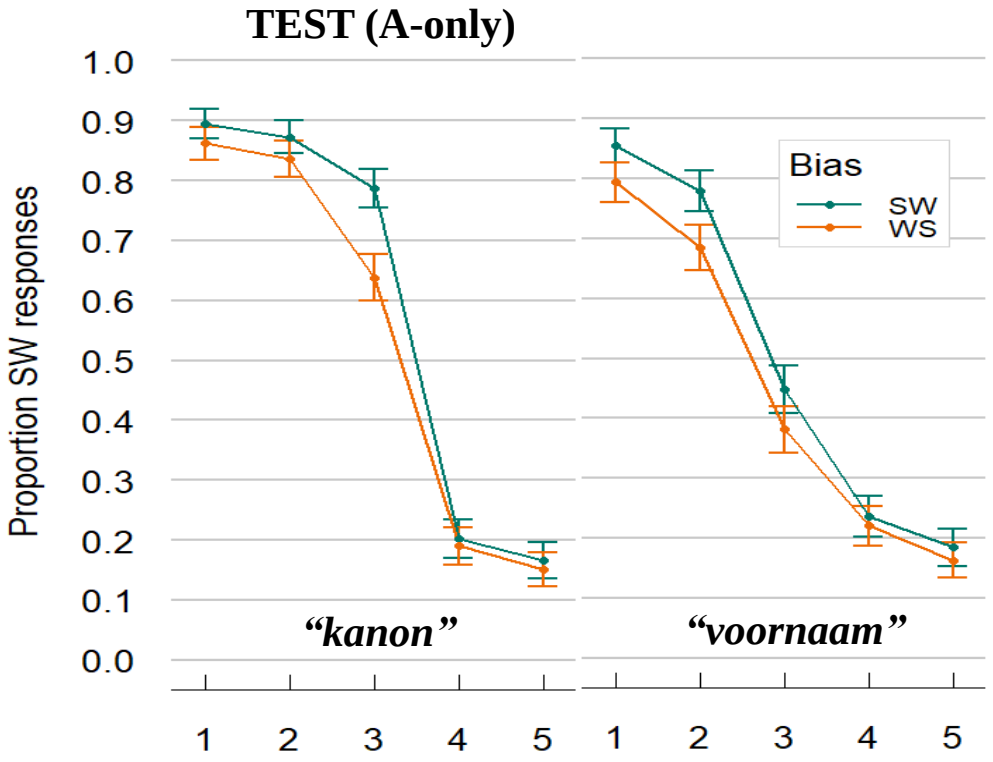
Bujok, Meyer & Bosker, *in prep.*

Prosody in the hands: lasting impact

- Audiovisual recalibration of lexical stress

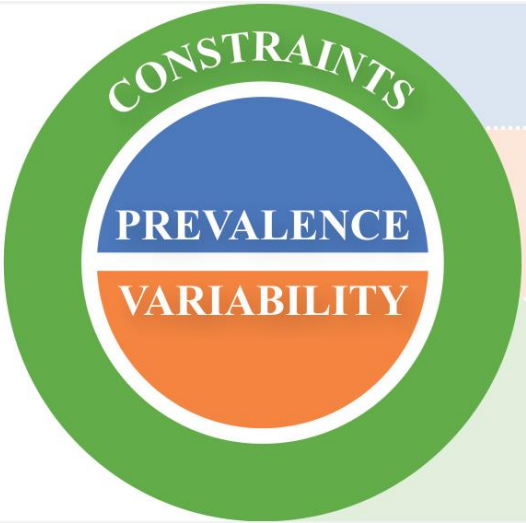
EXPOSURE (AV)

- Group 1: /ka.nɔn/?
  + beat on syllable 1
- Group 2: /ka.nɔn/?
  + beat on syllable 2



Bujok, Peeters, Meyer & Bosker, *subm.*, osf.io/9sua6

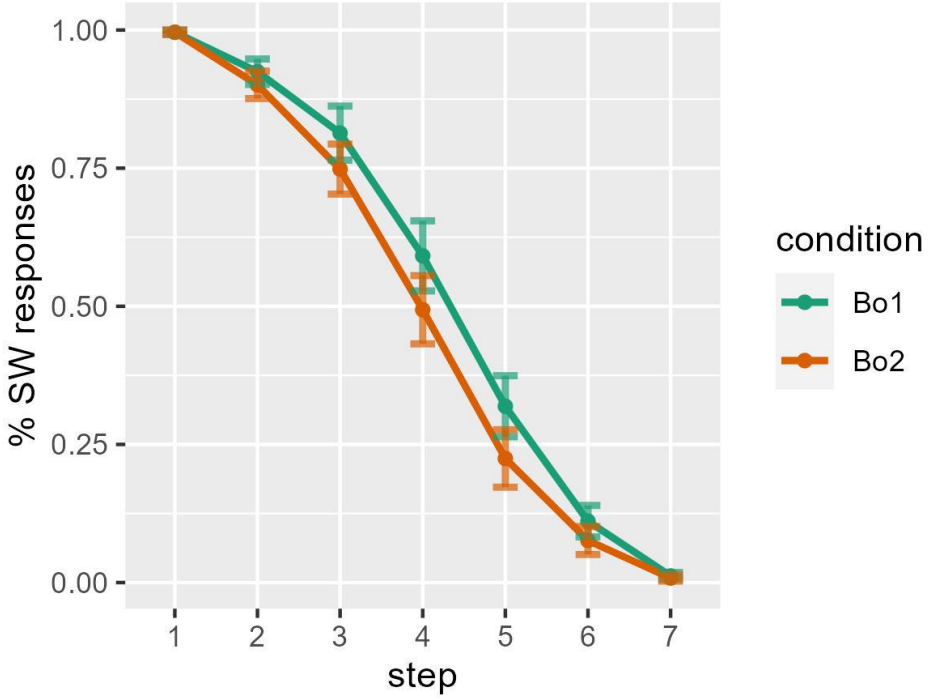
ERC StG HearingHands



WP1 (PD)	Gesture-speech synchrony in production & perception in free-stress, fixed stress, and tonal languages
WP2 (PD)	Neurobiology of gesture-speech synchrony integration and individual differences in the neurotypical and autistic population
WP3 (PHD)	Communicative, situational, and cognitive constraints on the effects of gesture-speech synchrony on spoken word recognition in eye-tracking and VR

ONGOING:

“Efecto McGurk manual”

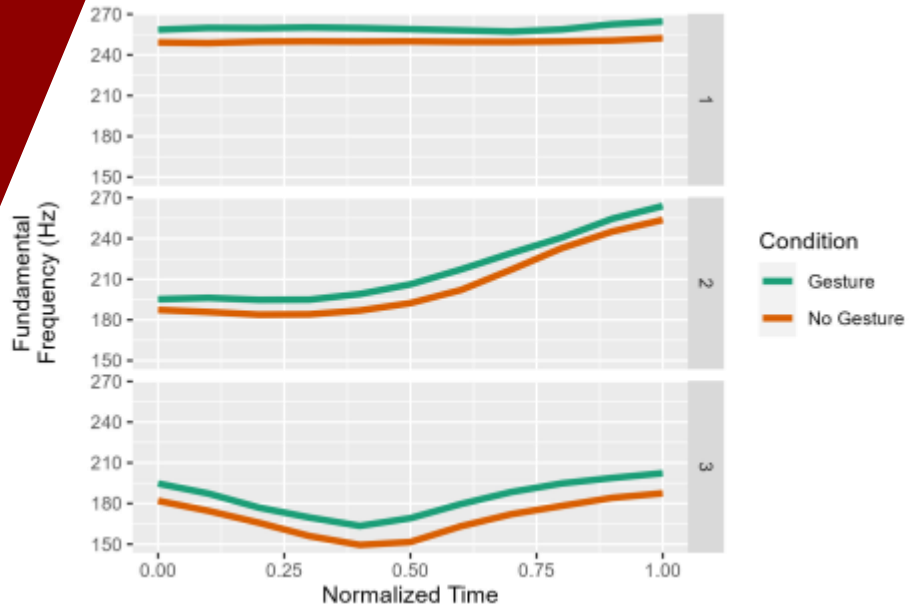


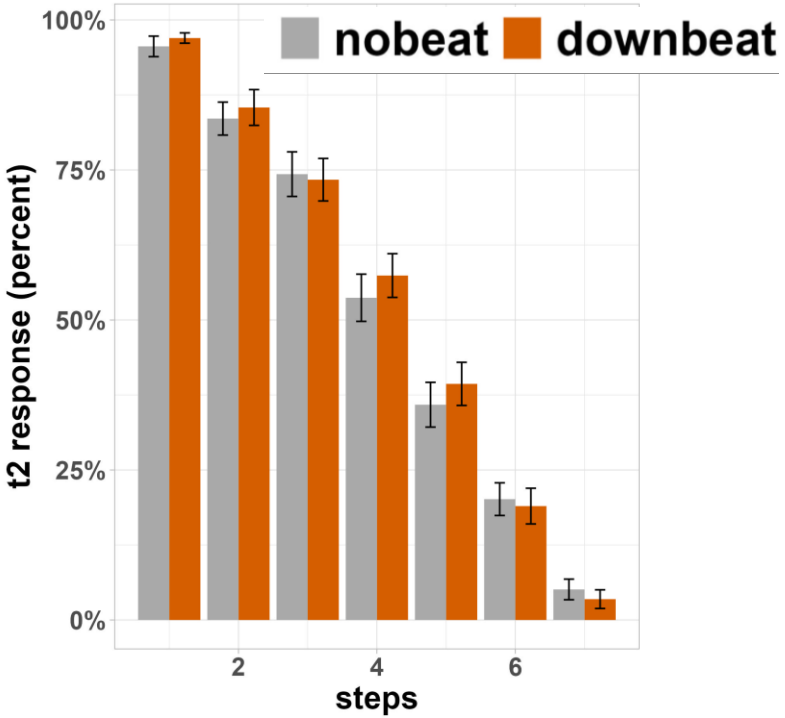
canto *cantó*

Rohrer, Bujok, Van Maastricht, & Bosker, *in prep.*

ONGOING:

Gesture-speech perception in Mandarin





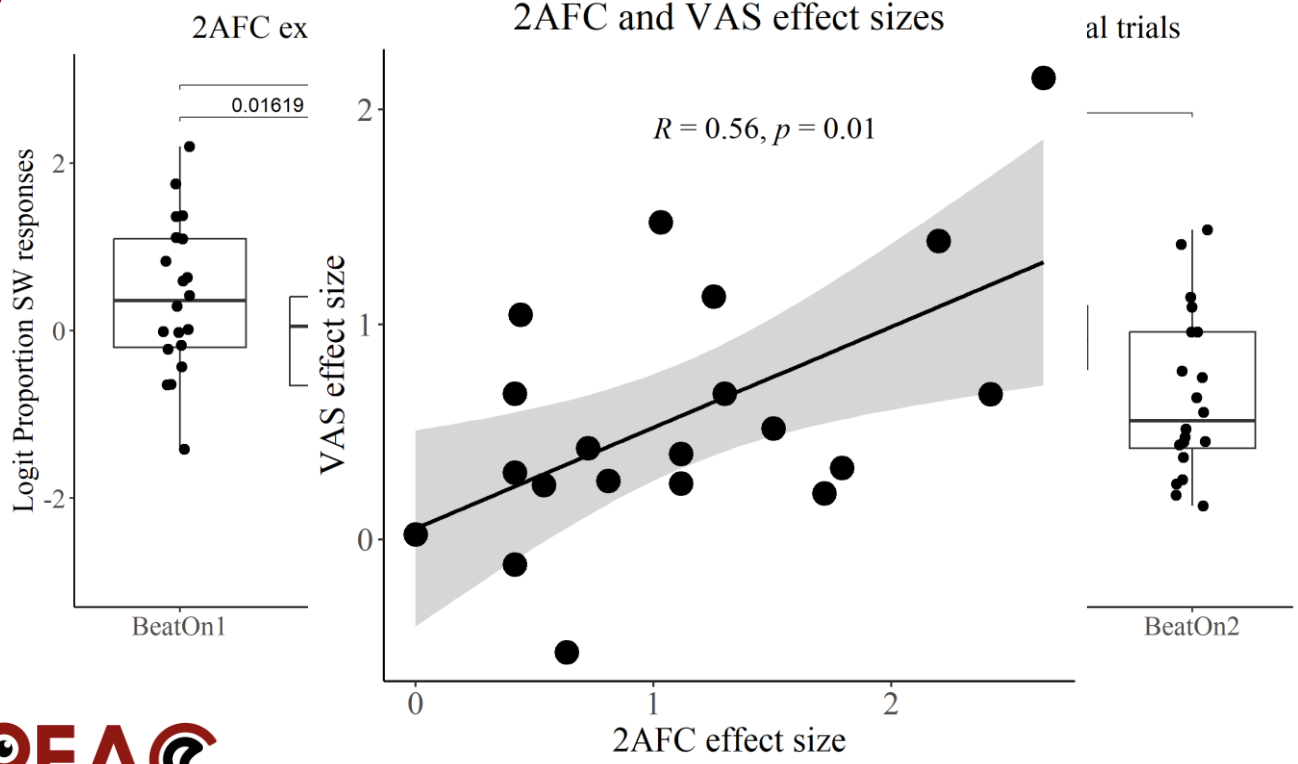
Hong, Rohrer, & Bosker, *in prep.*

SPEAC

Radboud University Radboudumc

ONGOING:

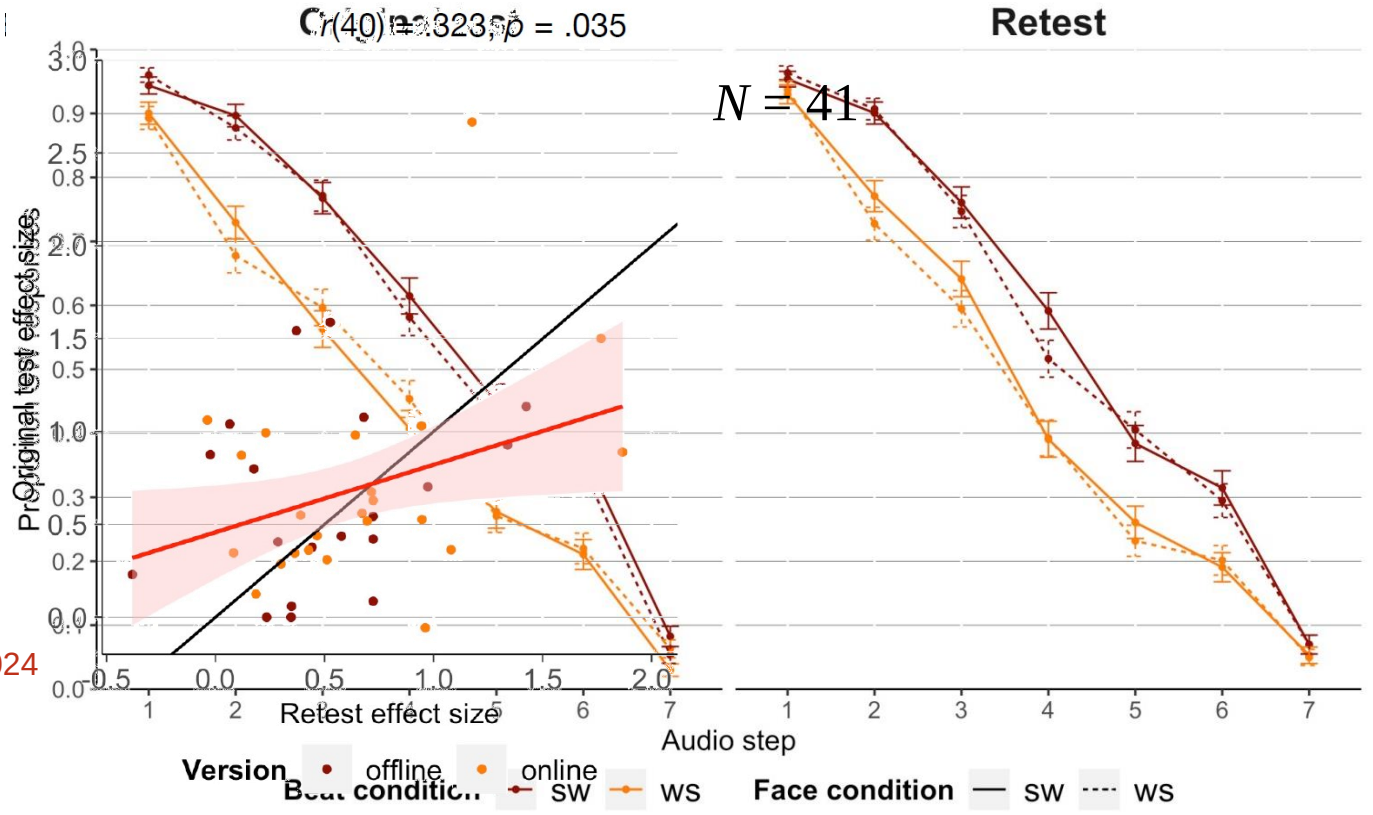
Minitest for inclusion in test batteries



Maran & Bosker, 2024

ONGOING:

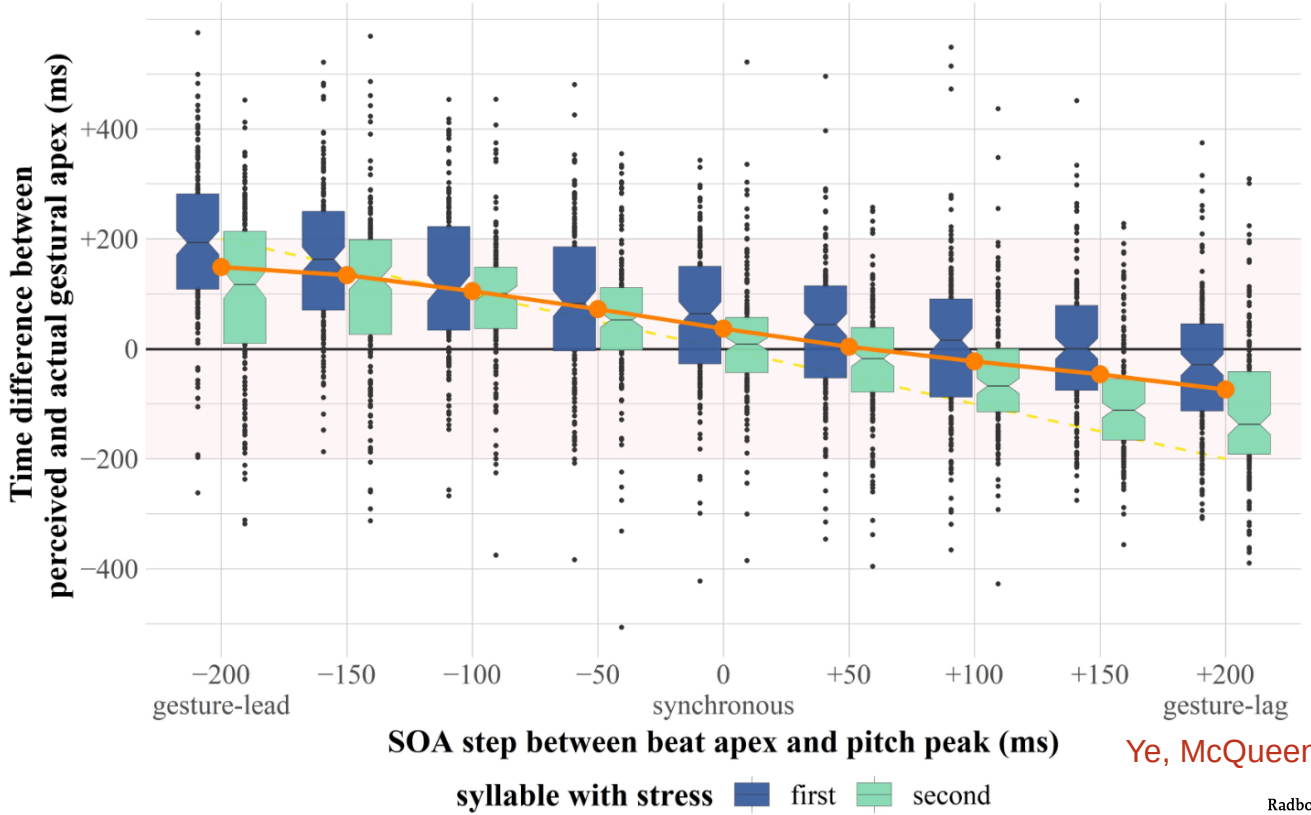
Test-retest reliability after >1.5 years



Cos, Bujok, & Bosker, 2024

ONGOING:

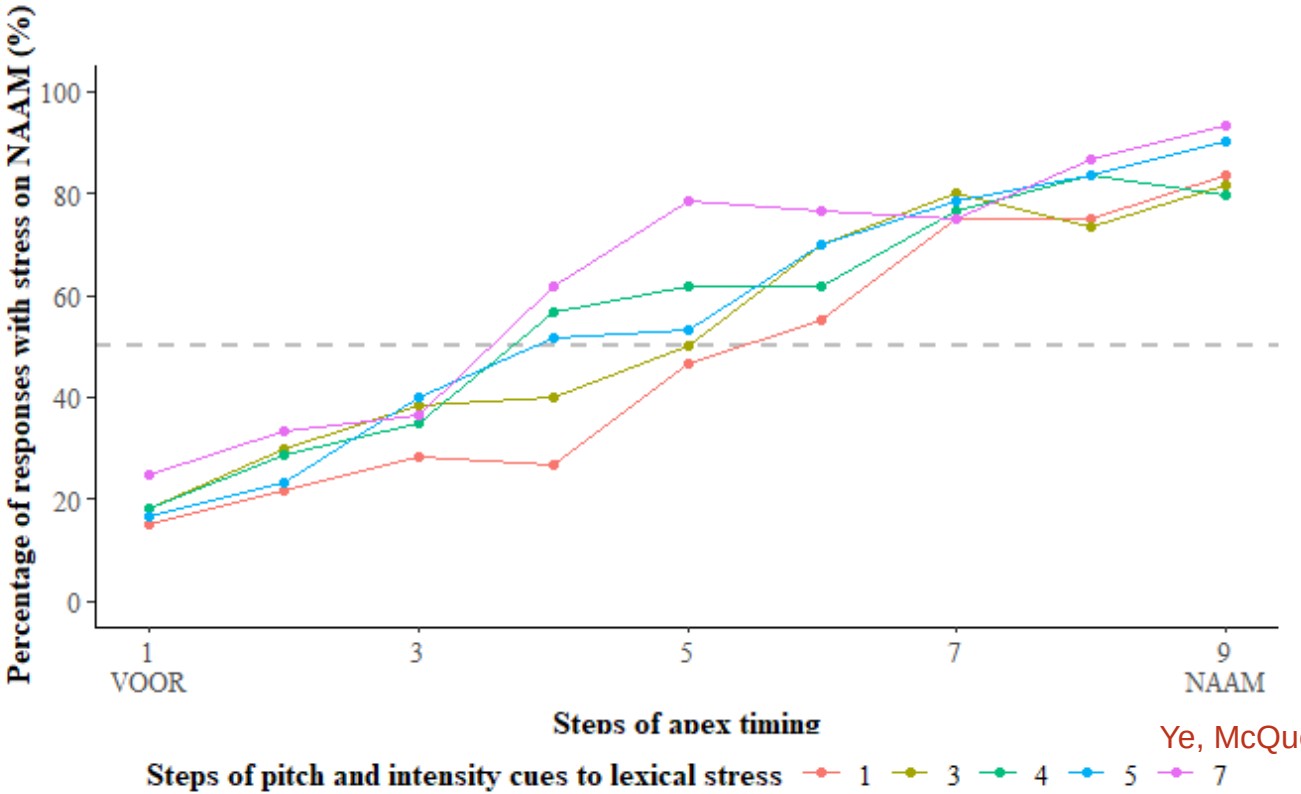
stress attracts perceived gestural timing



Ye, McQueen, & Bosker, *in prep.*

ONGOING:

stress attracts perceived gestural timing



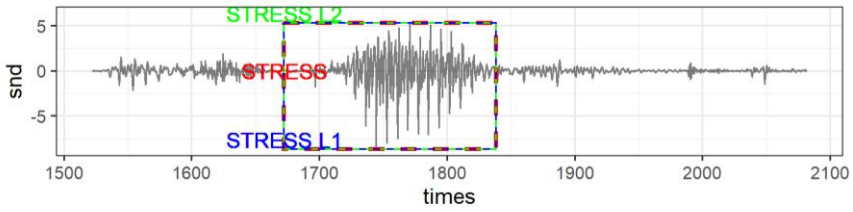
Ye, McQueen, & Bosker, *in prep.*

ONGOING:

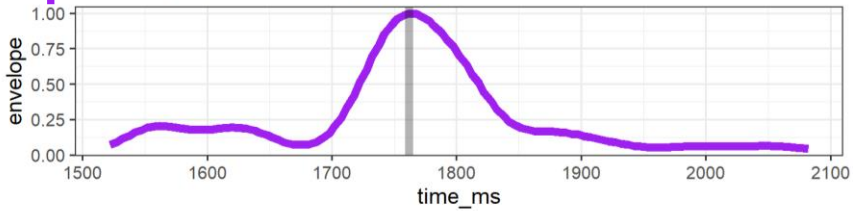
gesture-speech synchrony in L2

Spanish: “his.TÓ.ri.co”

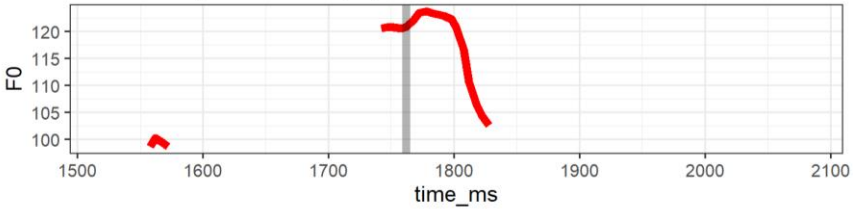
- How do L2 learners acquire L2 prosody in the auditory and gestural systems?



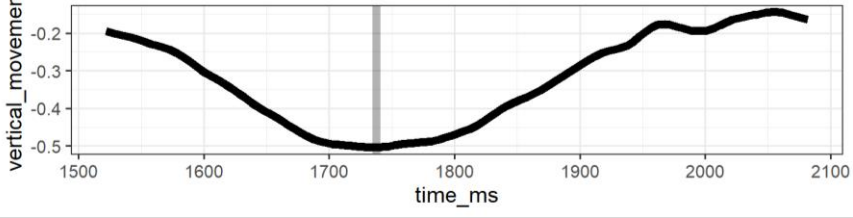
amplitude



***f*0**



hand trajectory

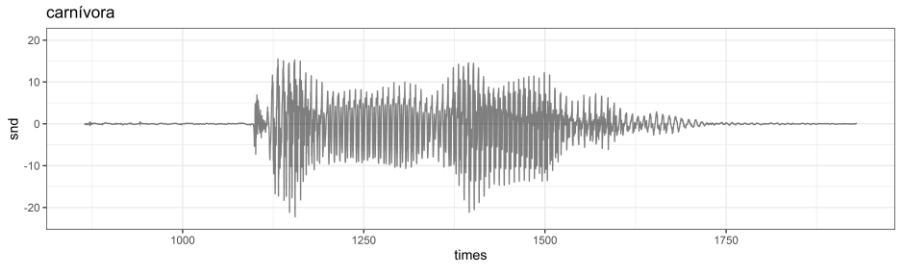


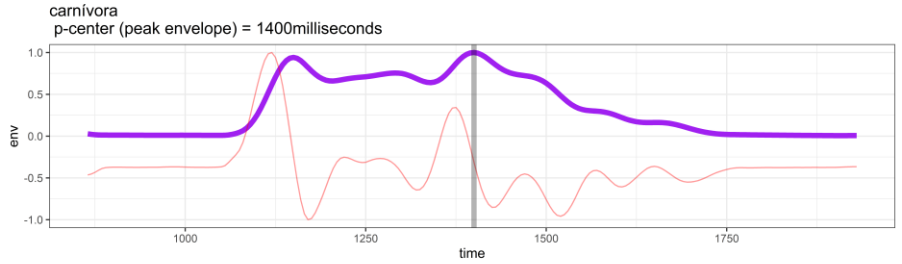
Example data

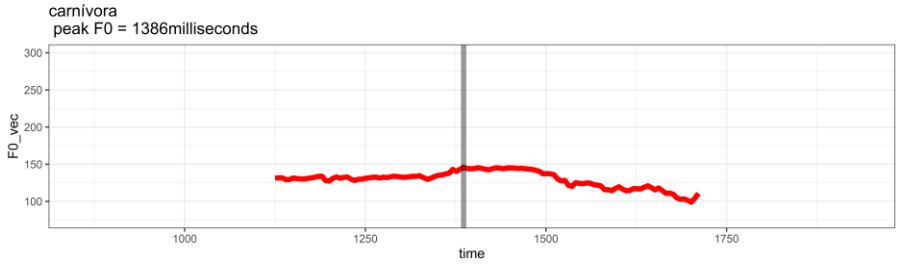
stress-mismatch

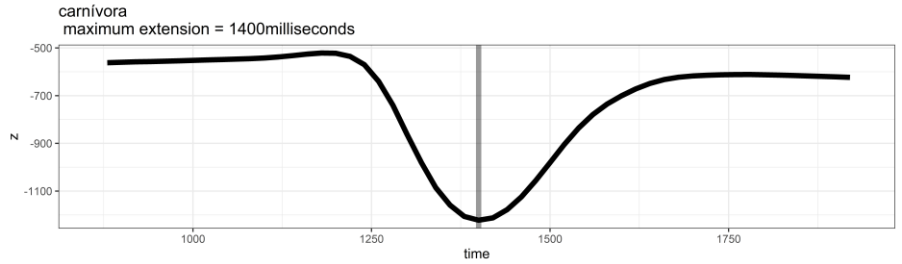
Dutch: *car.ni.VOOR*

Spanish: *car.NÍ.vo.ra*









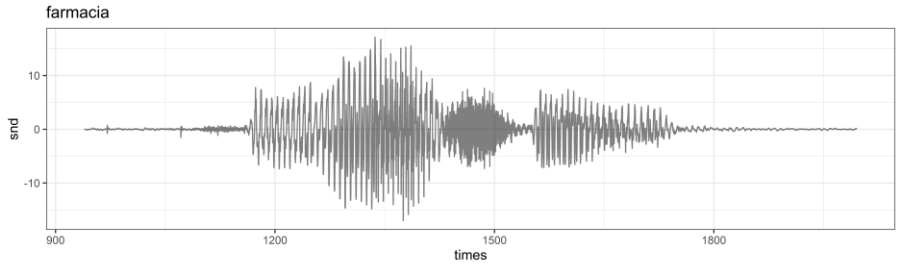
TEAM SCIENCE:
Bosker, Hoetjes, Pouw, Van Maastricht, *subm.*
https://osf.io/w2ezs,

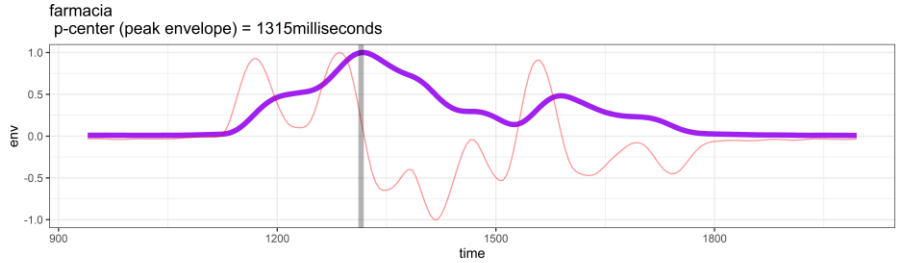
Example data

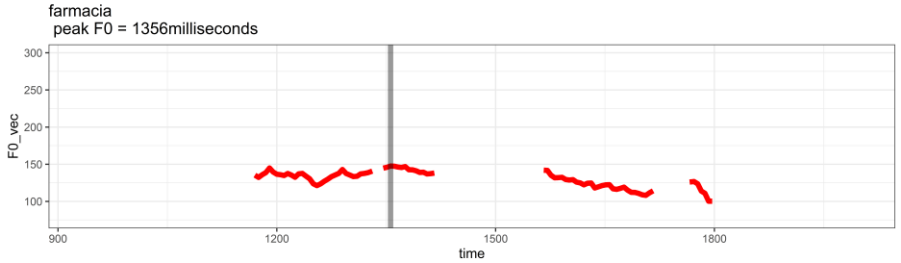
stress-mismatch

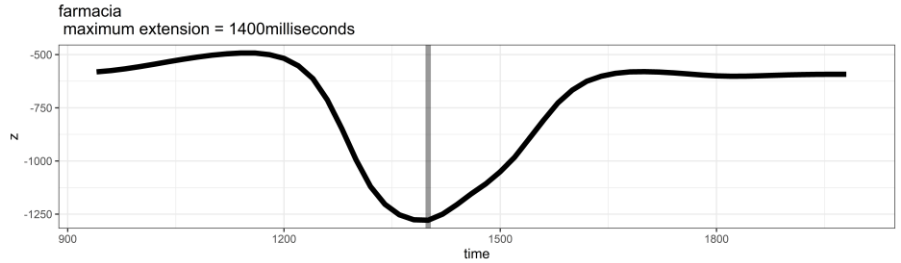
Dutch: *far.ma.CIE*

Spanish: *far.MA.cia*









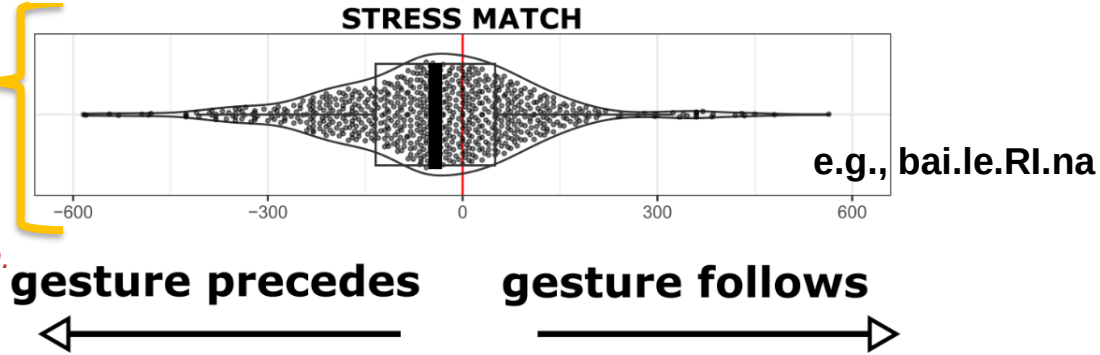
TEAM SCIENCE:
Bosker, Hoetjes, Pouw, Van Maastricht, *subm.*
https://osf.io/w2ezs,

Results: gesture-speech synchrony

L1 attraction

stress-mismatching cognates,
but correct L2 stress placement

stress-matching cognates
with correct L2 stress placement



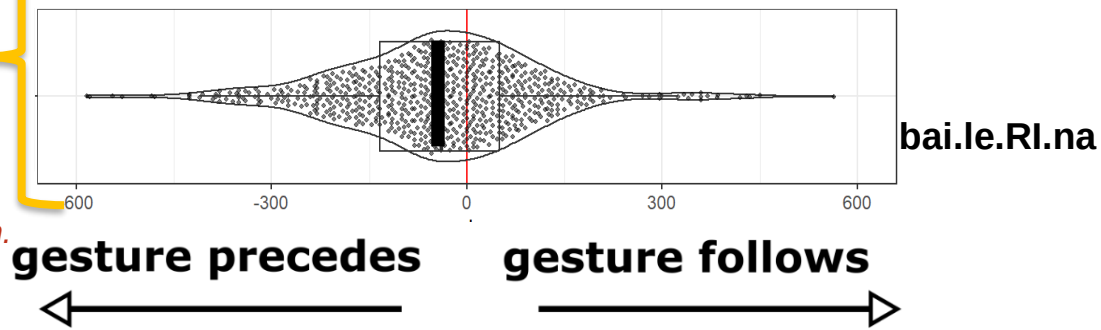
TEAM SCIENCE:
Bosker, Hoetjes, Pouw, Van Maastricht, *subm.*
https://osf.io/w2ezs,

Results: gesture-speech synchrony

L2 attraction

stress-mismatching cognates,
but ***incorrect*** L2 stress placement
(i.e., on L1 target syllable)

stress-matching cognates
with correct L2 stress placement



TEAM SCIENCE:
Bosker, Hoetjes, Pouw, Van Maastricht, *subm.*
https://osf.io/w2ezs,

Prosody in the hands: L2

- When L2 stress placement was **correct**, the hands showed **L1 attraction**
  - **‘kinematic accent’**

- When L2 stress placement was **incorrect**, erroneously stressing the L1 target, the hands showed **L2 attraction**

TEAM SCIENCE:
Bosker, Hoetjes, Pouw, Van Maastricht, *subm.*
https://osf.io/w2ezs,

Prosody in the hands: L2

- Snapshot of an unstable, developing, multimodal prosody system

- Overall strong evidence for multimodal gesture-speech coupling
- Yet separate unimodal L1/L2 attraction is reliably detectable
- Unique timing regimes for gestures vs. spoken prosody?
- Does a ‘kinematic accent’ of an L2 talker influence lexical stress perception in L1 listeners?

TEAM SCIENCE:
Bosker, Hoetjes, Pouw, Van Maastricht, *subm.*
https://osf.io/w2ezs,

Wrap-up of today

- Prosody is visible in articulatory movements, facial cues and expressions, and in the body.
- Multisensory cue weighting in audiovisual prosody perception, depending on the listening conditions
- Simple up-and-down hand gestures can influence what you hear.

Wrap-up of the course

- Prosody in Speech Perception
- Course aim: to reveal the central role that prosody plays in low-level speech perception and spoken word recognition.
- Course objectives:
    - to be familiar with key concepts in the area of speech prosody and speech perception
    - to be familiar with recent advances and paradigms in the speech perception literature
    - to understand how prosody influences the perception of vowels, consonants, and words
    - to understand the different processing mechanisms that underlie these influences
    - to understand the open issues and debates in the field of speech perception

Wrap-up of the course

- Normalization, neural speech tracking, prediction, talker-specific learning, audiovisual integration…
  - …are tightly interconnected.
  - …are not exclusive
  - …showcase the ‘cognitive toolkit’ listeners have at their disposal

Prosody matters!

WANNA JOIN US?

SPEAC

Hans Rutger Bosker

Speech Perception in Audiovisual Communication [SPEAC] lab

Donders Institute, Radboud University, Nijmegen, The Netherlands

https://hrbosker.github.io

hansrutger.bosker@donders.ru.nl