Multisensory integration of speech in social context

Magdalena Matyjek
Universitat Pompeu Fabra, Spain

Sotaro Kita
University of Warwick, UK

Salvador Soto Faraco
Universitat Pompeu Fabra, Spain

BACKGROUND

- **Speech in social contexts** involves integration of auditory and visual information across interrelated levels of representation ranging from purely spatio-temporal correlations to semantics.
- **In the process of multisensory integration (MSI)**, visual speech enhances auditory speech perception, especially in adverse listening conditions [1, 2, 3].
- Although the expression of audio-visual **speech 'in the wild'** involves multiple levels of information processing (including phonology, prosody, syntax, semantics, and pragmatics), these are seldom represented together in laboratory studies, which typically use isolated syllables or, at most, out of context.
- Yet, multilevel contextual cues help create expectations that trickle down to early processing stages of speech perception [4].

AIM & HYPOTHESIS

- We investigated gestural and visual-speech enhancement of auditory speech perception using stimuli embedded in a coherent discursive and social context, and therefore more ecologically valid.
- We hypothesise that the bimodal (audio-visual; AV) in comparison to unimodal (only A or only V) speech would elicit multisensory integration effects observed in behaviour (accuracy and reaction times) and in neuronal responses (alpha suppression).

METHODS

Sample

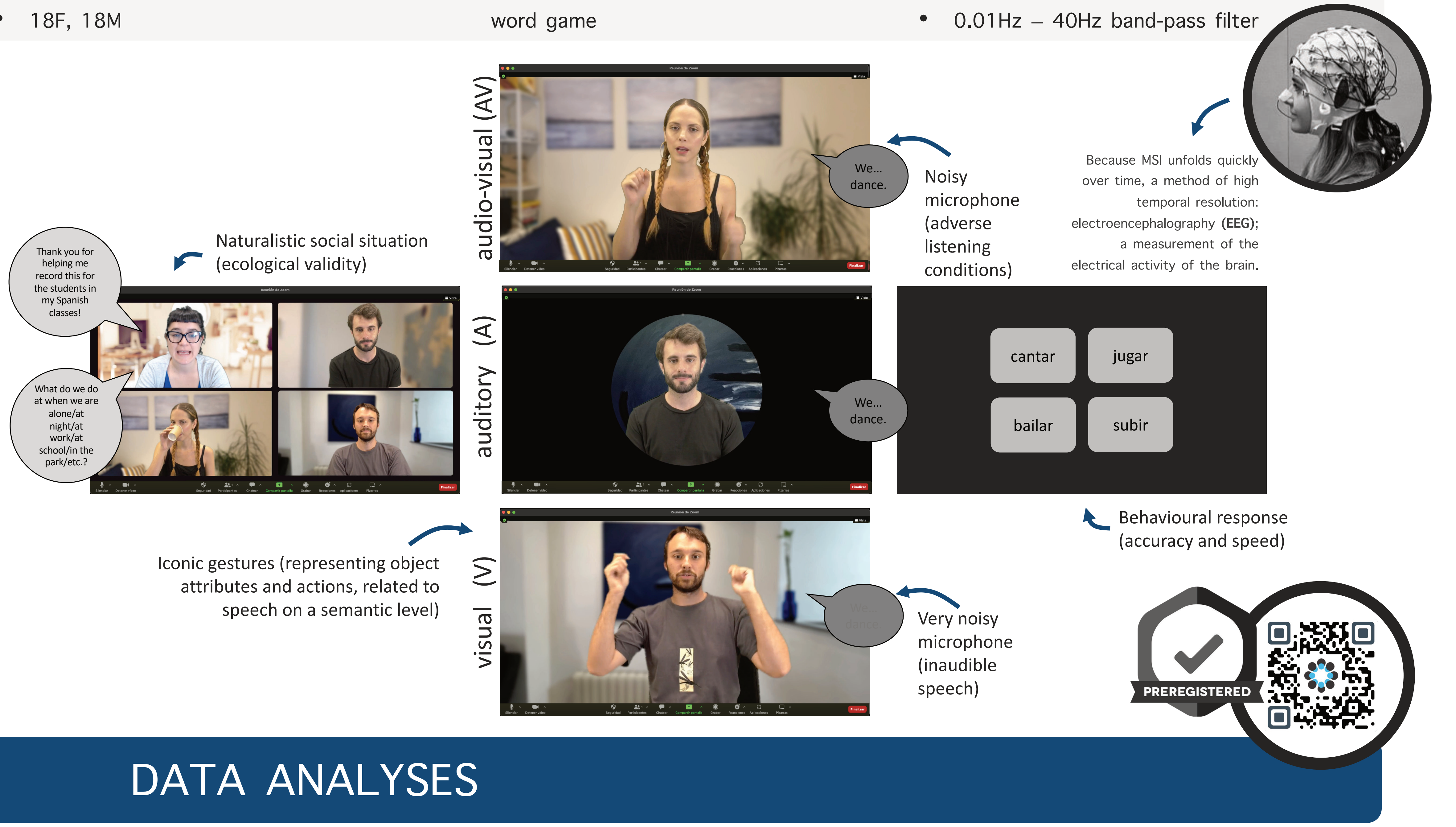
- N = 36, typically developing adults
- M age = 25 (6.6), min/max: 19/49
- 18F, 18M

Paradigm

- Participants watched a recording of a videoconference with 4 actors playing a word game

EEG recording

- 61 active electrodes (Brain Products)
- 500 Hz, reference: average
- 0.01Hz – 40Hz band-pass filter


audio-visual (AV)
auditory (A)
visual (V)
Thank you for helping me record this for the students in my Spanish classes!
What do we do at when we are alone/at night/at work/at school/in the park/etc.?
We... dance.
We... dance.
We... dance.
cantar
jugar
bailar
subir
PREREGISTERED

Naturalistic social situation (ecological validity)

Noisy microphone (adverse listening conditions)

Because MSI unfolds quickly over time, a method of high temporal resolution: electroencephalography **(EEG)**; a measurement of the electrical activity of the brain.

Behavioural response (accuracy and speed)

Iconic gestures (representing object attributes and actions, related to speech on a semantic level)

Very noisy microphone (inaudible speech)

DATA ANALYSES

EEG data

Alpha suppression

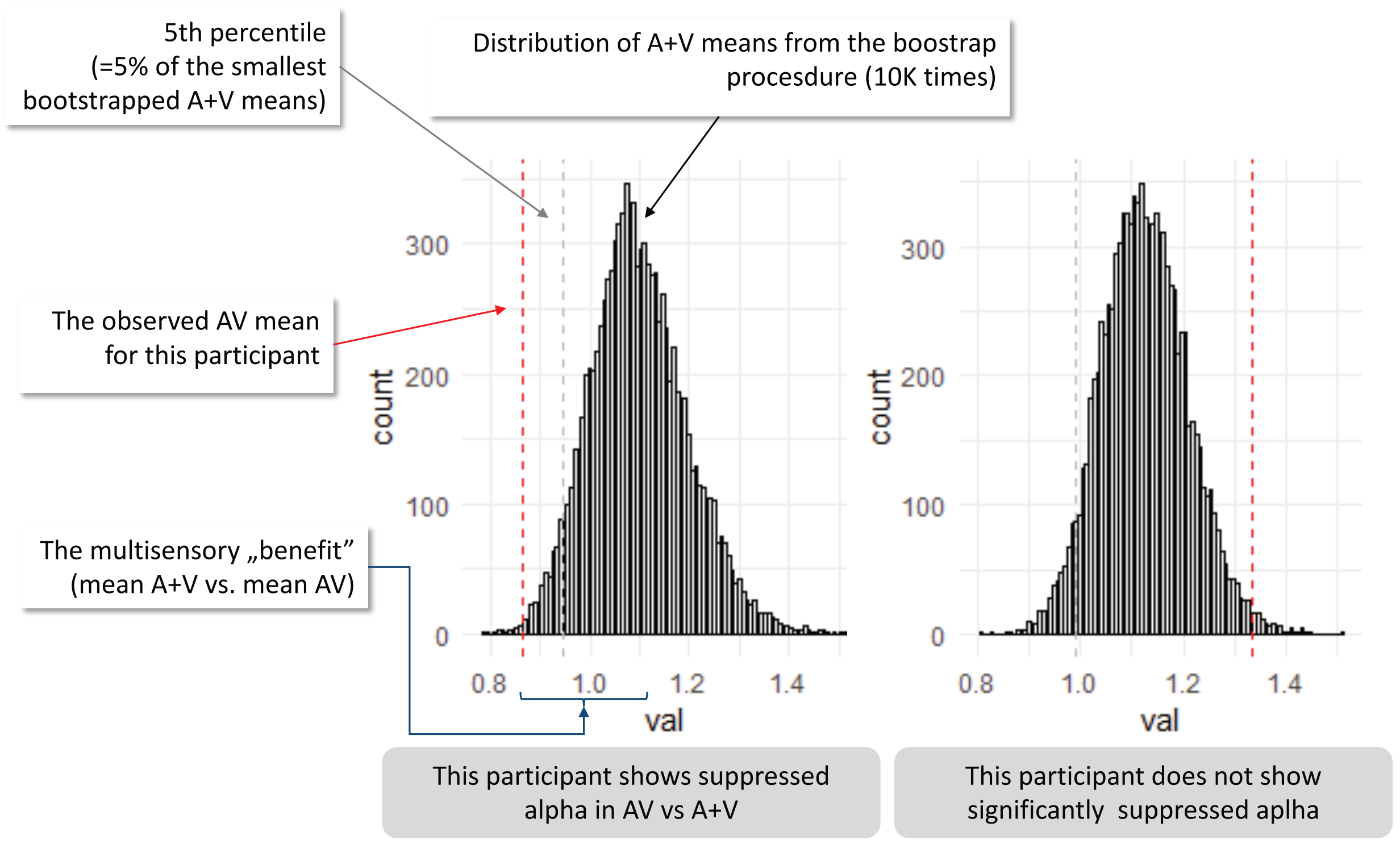
We expected to observe stronger parieto-occipital alpha suppression in AV trials (linked to integratory processes [5, 6, 7]) in the first second after stimulus onset.

MSI estimation using oscillatory power

To investigate MSI, we need to compare the AV condition against the sum of A and V. However, linear operations are not appropriate for oscillatory power computed with wavelet transformations. Instead, we used the following analysis [based on 8]

Analysis steps:

1. Linear summation of all A and V combinations in the time domain
2. Power calculations for A+V and AV (wavelets)
3. Bootstrapping A+V
4. Comparing the observed AV mean against the distribution of means of A+V
5. Group statistics:
   a. Chi-squared test
   b. T-test with z-scores



5th percentile (=5% of the smallest bootstrapped A+V means)

Distribution of A+V means from the boostrap procesdure (10K times)

The observed AV mean for this participant

The multisensory „benefit" (mean A+V vs. mean AV)

This participant shows suppressed alpha in AV vs A+V

This participant does not show significantly suppressed aplha

Behavioural data

Planned: Accuracy

We built a generalised linear mixed model (GLMM) with binary dependent variable for single-trial correct and incorrect responses:

```
glmer(formula = correct ~ cond + (1|ID) + (1 +
cond|verb), family = binomial(link = "logit")
```

Exploratory: Reaction times (RTs)

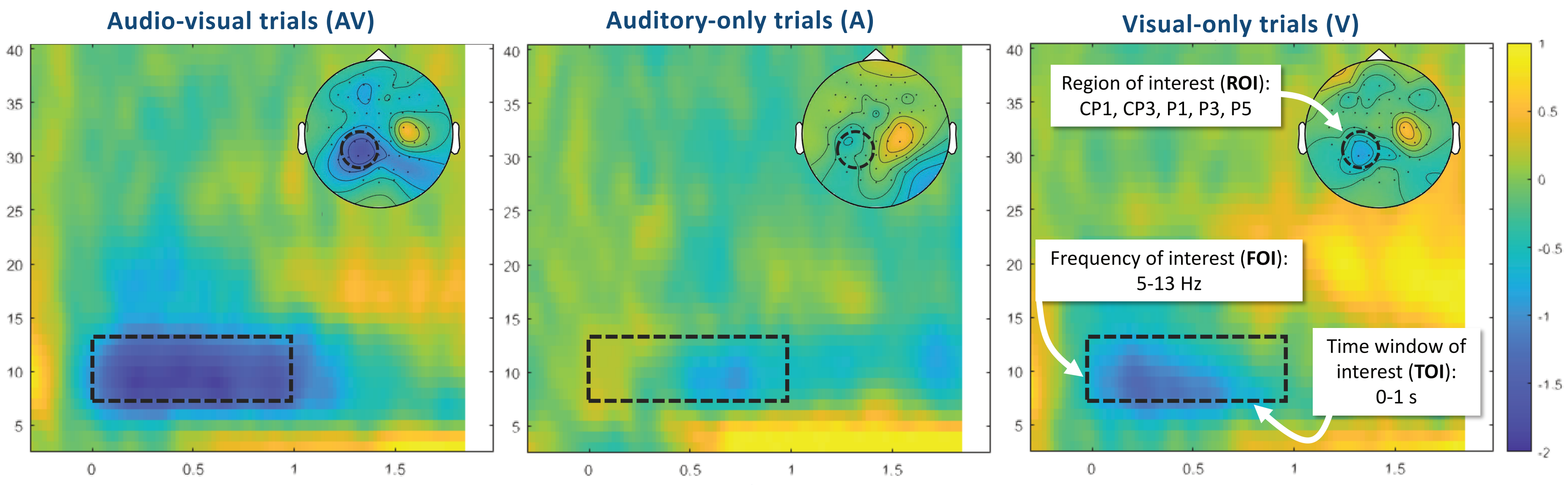
To eliminate motor preparation from the early EEG signal, we included a 200-ms buffer between the stimulus and response [9]. Hence, RTs may not be sensitive to MSI. We built an exploratory linear mixed model:

```
glmer(RT ~ cond + (1|ID) + (1|verb),
family=Gamma(link="identity"))
```

RESULTS

EEG data

Time-frequency representations (TFR): Hanning window of 0.5s, 5 to 30 Hz, in frequency steps of 1 Hz. As preregistered, the electrodes for the alpha suppression analysis were chosen based on the TFRs [8].



Region of interest (**ROI**): CP1, CP3, P1, P3, P5

Frequency of interest (**FOI**): 5-13 Hz

Time window of interest (**TOI**): 0-1 s

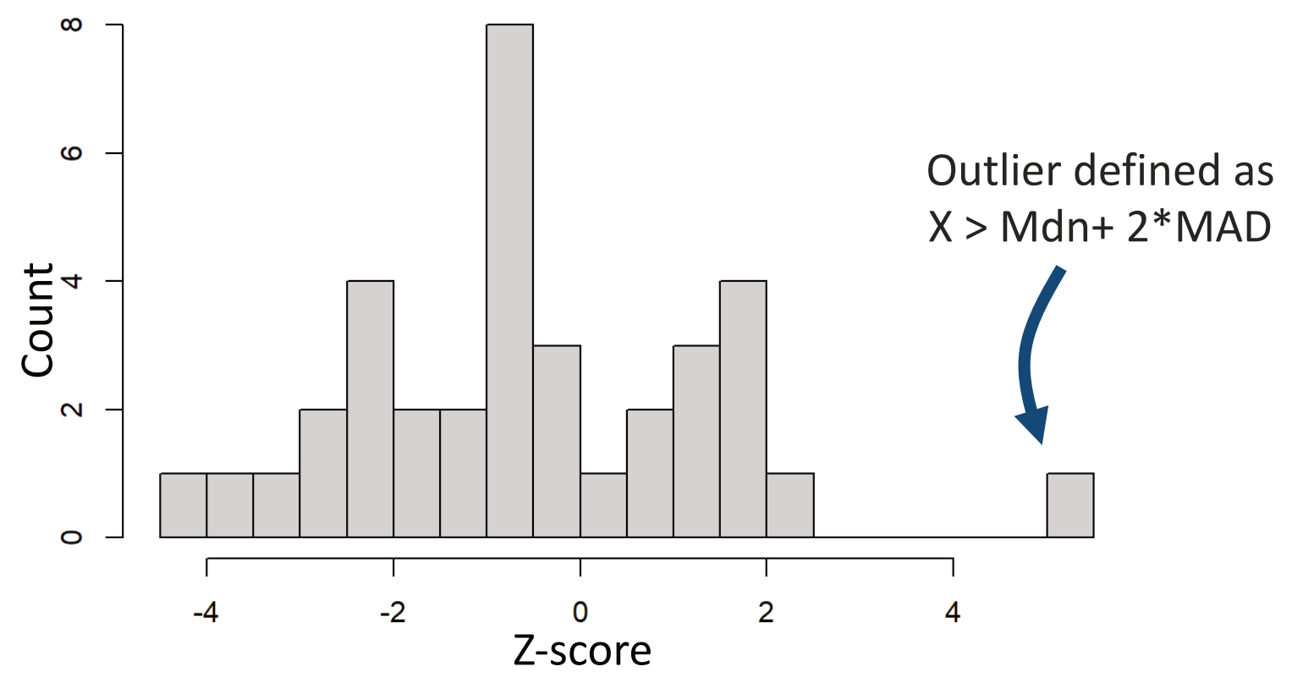
Group analyses

Observed vs. random distribution [8]

	Alpha suppression	No sig. alpha suppression
Observed data	11	25
Expected random	1.8	34.2
	$X^2(1) = 8.04$, $p = .005$	

Significantly more participants show alpha suppression than expected in a random distribution.

Z-scores against 0

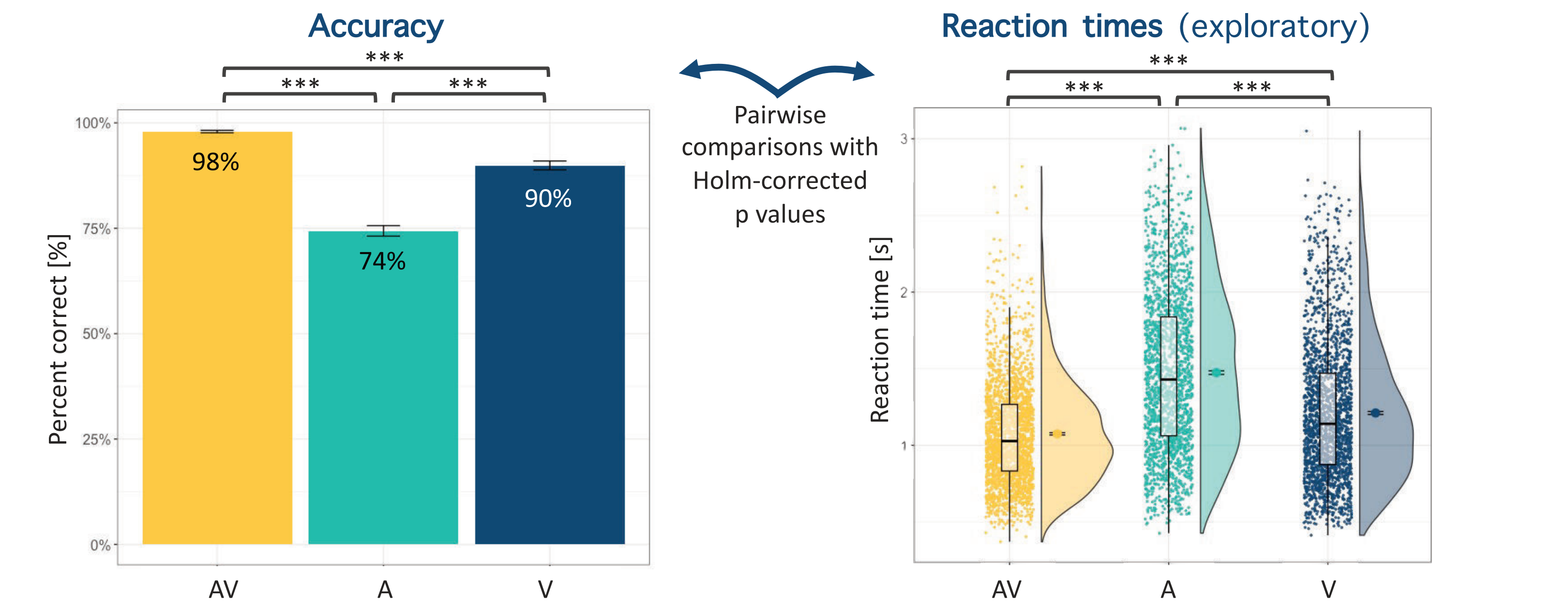


Outlier defined as X > Mdn+ 2*MAD

Without the outlier: $t(34) = -2.23$, $p = .016$
With the outlier: $t(35) = -1.5$, $p = .071$

The sample showed statistically stronger suppression in AV than A+V trials.

Behavioural data



Pairwise comparisons with Holm-corrected p values

Effect of condition: $X^2(2) = 105.55$, $p < .001$

Effect of condition: $X^2(2) = 902.76$, $p < .001$

DISCUSSION

Behavioural benefit from visual speech and iconic gestures

Visual speech and gesture facilitate processing of noisy auditory speech (accuracy and speed).

Neuronal correlates of MSI

Audio-visual speech is linked to alpha suppression (an MSI correlate), which cannot be accounted for by simple additive processing of the partial, audio and visual, information.

Take-home message:

Audio-visual speech accompanied by iconic gestures is readily integrated to the benefit of speech perception in dynamic, complex, and naturalistic social situations.

Bibliography

[1] Bremner, A. J., Lewkowicz, D. J., & Spence, C. (2012). The multisensory approach to development. In Multisensory development (pp. 1–26). Oxford University Press.

[2] Schepers, I., et al. (2013). Noise alters beta-band activity in superior temporal cortex during audiovisual speech processing. NeuroImage, 70, 101–112.

[3] Stekelenburg, J. J., & Vroomen, J. (2007). Neural Correlates of Multisensory Integration of Ecologically Valid Audiovisual Events. Journal of Cognitive Neuroscience, 19(12), 1964–1973.

[4] Brunellière, A., Sánchez-García, C., Ikumi, N., & Soto-Faraco, S. (2013). Visual information constrains early and late stages of spoken-word recognition in sentence context. International Journal of Psychophysiology, 89(1), 136–147.

[5] Drijvers, L., Özyürek, A., & Jensen, O. (2018a). Hearing and seeing meaning in noise: Alpha, beta, and gamma oscillations predict gestural enhancement of degraded speech comprehension. Human Brain Mapping, 39(5), 2075–2087.

[6] Drijvers, L., Özyürek, A., & Jensen, O. (2018b). Alpha and Beta Oscillations Index Semantic Congruency between Speech and Gestures in Clear and Degraded Speech. Journal of Cognitive Neuroscience, 30(8), 1086–1097.

[7] Drijvers, L., van der Plas, M., Özyürek, A., & Jensen, O. (2019). Native and non-native listeners show similar yet distinct oscillatory dynamics when using gestures to access speech in noise. NeuroImage, 194, 55–67.

[8] Senkowski, D., Gomez-Ramirez, M., Lakatos, P., Wylie, G. R., Molholm, S., Schroeder, C. E., & Foxe, J. J. (2007). Multisensory processing and oscillatory activity: Analyzing non-linear electrophysiological measures in humans and simians. Experimental Brain Research, 177(2), 184–195.

[9] Besle, J., Fort, A., & Giard, M.-H. (2004). Interest and validity of the additive model in electrophysiological studies of multisensory interactions. Cognitive Processing, 5(3), 189–192.

This programme has received co-funding from the European Union's Horizon 2020 research and innovation Maria Sklodowska-Curie Actions, under grant agreement number 945380.