Speaker-Specific Semantic Priors Enhance Both Expected and Unexpected Speech Across Processing Levels

Fabian Schneider (f.schneider@uke.de
Institute of Systems Neuroscience, University Medical Centre Hamburg-Eppendorf, Martinistr. 52
Hamburg, Hamburg, 20246, Germany

Helen Blank (h.blank@uke.de)
Faculty of Psychology, Ruhr University Bochum, Universitätsstr. 150
Bochum, North-Rhine Westphalia, 44801, Germany

Abstract

Predictive processing is fundamental to speech perception, yet how priors shape neural representations at different hierarchical levels remains debated. Here, we investigate how humans combine expectations about what another person is going to talk about, i.e., speaker-specific semantic priors, with ambiguous sensory inputs. Using a combination of stimulus reconstruction models, representational similarity analysis, and single-trial encoding models of EEG responses, we show two complementary processes of speaker-specific semantic priors: sharpening of low-level acoustic representations, pulling them towards the expected acoustic signal and that prediction errors only at higher levels of the neural hierarchy, signaling semantic surprisal. Critically, speaker-specific priors were not applied when incoming words clearly deviated, indicating flexibility as a function of their relative likelihood. Together, these findings provide evidence for a unified theory of predictive processing in the brain in which priors enhance both expected and unexpected information at different levels of the processing hierarchy.

Keywords: predictive coding; Bayesian brain; semantics

Introduction

High-level priors, such as semantic expectations, are widely believed to generate low-level predictions to facilitate perception (Friston, 2010; de Lange et al., 2018). However, how these priors interact with novel sensory input at a mechanistic level remains unclear (Aitchison & Lengyel, 2017; de Lange et al., 2018). Hierarchical predictive coding (hPC) suggests that the brain emphasizes unexpected input through prediction error signals (Blank & Davis, 2016; Heilbron et al., 2022; Caucheteux et al., 2023; Millidge et al., 2021), while Bayesian theories propose that predictions enhance expected features through sharpening (Kok et al., 2012; Jaramillo & Zador, 2010). These views are often seen as conflicting, but may in fact reflect complementary processes operating at different levels of the neural hierarchy (Press et al., 2020).

To test this hypothesis, we conducted a series of preregistered experiments, where participants ($N_1 = 35$, $N_2 = 35$) were cued with one of six speakers and asked to identify acoustic morphs between two words (e.g., *sea-tea*), of which only one was semantically coherent with the speaker. Speaker-specific feedback reinforced consistent semantic associations (e.g., $\text{speaker}_{\text{food}}$: tea, $\text{speaker}_{\text{nature}}$: sea). Thus, while the acoustic signal remained constant, prior expectations varied by speaker cues. We recorded neural responses via electroencephalography (EEG). An independent, preregistered validation study ($N = 40$) controlled context-free perception of morphs.

In real time, we estimated speaker-specific (i.e., one prior per speaker) and speaker-invariant (i.e., one prior across speakers) semantic priors by fitting multivariate Gaussian distributions over preprocessed semantic embeddings (Pennington et al., 2014; Raunak et al., 2019) via free-energy optimisation (Bogacz, 2017). These priors allowed us to generate acoustic and semantic predictions from the current context and examine their influence on neural responses.

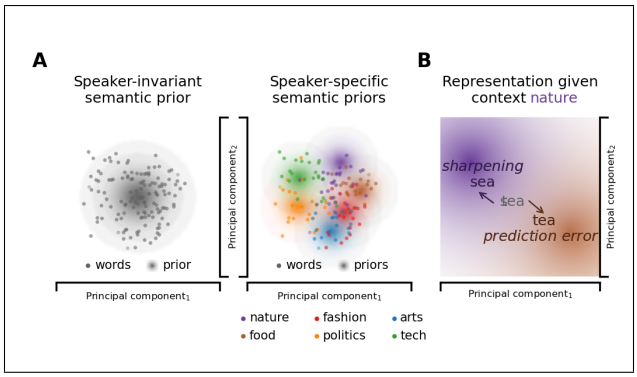

Figure 1: **A** We tested two forms of semantic priors: Speaker-invariant priors (one prior over all speakers) and speaker-specific priors (one prior per speaker). **B** On a mechanistic level, prior expectations may either pull neural representations of incoming words towards expected signals (sharpening) or repel them away (prediction error).

Speaker-specific priors guide perception

Participants reported the word that was in line with the speaker-specific prior (GLMM, $\beta = 1.64 \pm 0.12, z = 13.11, p = 2.71 \times 10^{-39}$), and this influence increased with exposure ($\beta = 0.43 \pm 0.03, z = 12.77, p = 2.41 \times 10^{-37}$). These results show that listeners used speaker-specific priors to interpret ambiguous sensory signals.

Priors sharpen acoustic representations

To examine the content of sensory representations, we trained stimulus reconstruction models (Crosse et al., 2016) on EEG data to reconstruct spectrograms from neural responses to morphs. First, we confirmed reconstruction success (one-sample t-test, $r = 0.06 \pm 0.02, t(34) = 17.06, p = 8.96 \times 10^{-17}$). To test how semantic priors influence the representation of morphs at the acoustic level, we computed sensory representational similarity matrices (RSMs) (Kriegeskorte, 2008) between reconstructions and the corresponding clear words across the different speaker contexts:

$$\mathrm{RSM}(n,t) = \begin{bmatrix} f(/\_\mathrm{i}:/|\mathrm{nature}, /\mathrm{si}:/) & f(/\_\mathrm{i}:/|\mathrm{food}, /\mathrm{si}:/) \\ f(/\_\mathrm{i}:/|\mathrm{nature}, /\mathrm{ti}:/) & f(/\_\mathrm{i}:/|\mathrm{food}, /\mathrm{ti}:/) \end{bmatrix}$$

Here, f denotes cosine similarity, n is the morph index, and t is time. We then constructed hypothesis RSMs based on 1) baseline acoustic similarity, 2) top-k prior-weighted acoustic predictions, and 3) semantic embeddings of priors and corresponding clear words.

Encoding models revealed significant modulation of sensory RSMs by both speaker-invariant and speaker-specific acoustic predictions (invariant-baseline: $t(34) = 22.40$, $p = 1.47 \times 10^{-20}$; specific-baseline: $t(34) = 25.12$, $p = 4.04 \times 10^{-22}$). Specific predictions outperformed invariant ones ($t(34) = 4.18$, $p = 2.11 \times 10^{-3}$), and their contributions were additive (both-specific: $t(34) = 15.03$, $p = 2.35 \times 10^{-15}$; both-invariant: $t(34) = 17.87$, $p = 1.33 \times 10^{-17}$). Semantic RSMs failed to improve encoding (all $t \leq -2.19$), suggesting sharpening at the acoustic level.

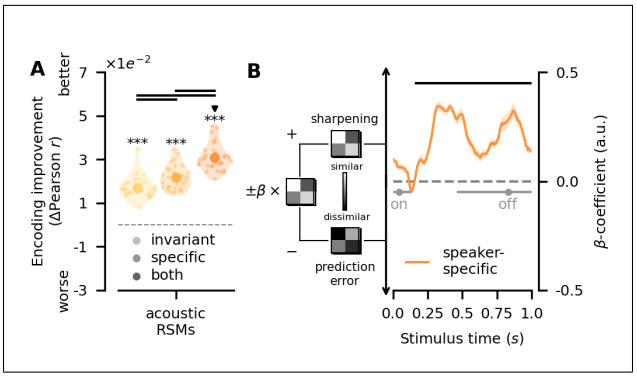

Figure 2: We encoded representational similarity of reconstructed morphs and corresponding clear words. **A** Only acoustic RSMs boosted encoding of sensory RSMs. The effects of speaker-invariant and -specific predictions were additive. **B** Coefficients revealed a positive relationship between speaker-specific acoustic predictions and sensory RSMs, suggesting a sharpening of sensory representations.

Critically, model coefficients revealed a significant cluster of positive values for speaker-specific acoustic predictions ($p \leq 8.00 \times 10^{-4}$) from 165–1000ms post-stimulus, supporting a sharpening mechanism wherein priors pull neural representations toward expected acoustic features.

Prediction errors at higher levels

To test for additional information theoretic prediction errors, we used single-trial encoding of broadband EEG (Crosse et al., 2016) with surprisal computed from wav2vec2.0 activations (Baevski et al., 2020). These representations were projected into a 5D subspace, with transformer layer 12 selected independently via back-to-back decoding (King et al., 2020).

Only speaker-specific semantic surprisal improved model performance (semantic-baseline: $t(34) = 5.33$, $p = 1.90 \times 10^{-5}$). Temporal knock-out analyses revealed a significant effect from 150–630ms ($p \leq 1.80 \times 10^{-3}$), suggesting prediction errors occurred at higher levels such as phonemes or semantics, but not at the acoustic level.

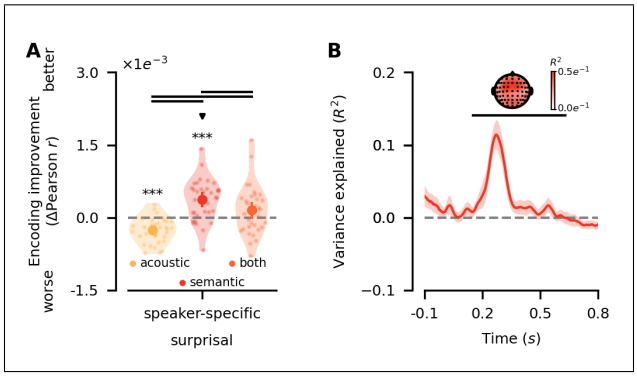

Figure 3: **A** Single-trial encoding revealed only speaker-specific semantic surprisal to improve model fit, indicating that prediction errors emerged at higher levels of processing. **B** Knock-out analysis of coefficients revealed significant variance explained by speaker-specific semantic surprisal around $150\text{-}630ms$, in line with higher level processing.

Flexible deployment of priors

In a final task, participants identified degraded unmorphed words that were highly congruent or incongruent with a given speaker's prior, without feedback.

Participants were overall slower to respond to incongruent trials (LMM, $\beta = 0.12 \pm 0.02$, $t = 6.86$, $p = 7.85 \times 10^{-12}$). In congruent trials, response speed scaled only with speaker-specific prior probability (EMTs, $\beta = -0.11 \pm 0.01, t = -9.19$, $p = 1.12 \times 10^{-19}$); this was not observed in incongruent trials ($\beta = -0.01 \pm 0.01, t = -0.85, p = 0.39$).

Encoding mirrored this pattern: Only speaker-specific surprisal improved encoding in congruent trials ($t(34) = 8.39$, $p = 3.39 \times 10^{-8}$), while only invariant surprisal mattered in incongruent ones ($t(34) = 3.03$, $p = 1.88 \times 10^{-2}$). This double dissociation of specificity and congruency demonstrates that listeners deploy and discard priors dynamically as a function of contextual plausibility.

Conclusion

Listeners use speaker-specific semantic priors during speech perception and apply priors flexibly, depending on contextual congruency. Critically, these priors reveal two complementary underlying processes: they sharpen low-level acoustic representations and generate prediction errors at higher levels. Together, our findings reconcile sharpening and prediction error computations within a unified Bayesian framework that departs from more traditional theories of hierarchical predictive coding (Press et al., 2020).

Acknowledgments

This work was funded by the Emmy Noether program of the Deutsche Forschungsgemeinschaft (German Research Foundation; Grant No DFG BL 1736/1-1).

References

Aitchison, L., & Lengyel, M. (2017, October). With or without you: predictive coding and Bayesian inference in the brain. *Current Opinion in Neurobiology*, *46*, 219–227. Retrieved 2022-10-11, from https://linkinghub.elsevier.com/retrieve/pii/S09594388173 doi: 10.1016/j.conb.2017.08.010

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations.* arXiv. Retrieved 2025-01-22, from https://arxiv.org/abs/2006.11477 doi: 10.48550/ARXIV.2006.11477

Blank, H., & Davis, M. H. (2016, November). Prediction Errors but Not Sharpened Signals Simulate Multivoxel fMRI Patterns during Speech Perception. *PLOS Biology*, *14*(11), e1002577. Retrieved 2023-06-28, from https://dx.plos.org/10.1371/journal.pbio.1002577 doi: 10.1371/journal.pbio.1002577

Bogacz, R. (2017, February). A tutorial on the free-energy framework for modelling perception and learning. *Journal of Mathematical Psychology*, *76*, 198–211. Retrieved 2023-03-28, from https://linkinghub.elsevier.com/retrieve/pii/S0022249615 doi: 10.1016/j.jmp.2015.11.003

Caucheteux, C., Gramfort, A., & King, J.-R. (2023, March). Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, *7*(3), 430–441. Retrieved 2025-01-16, from https://www.nature.com/articles/s41562-022-01516-2 doi: 10.1038/s41562-022-01516-2

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016, November). The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. *Frontiers in Human Neuroscience*, *10*. Retrieved 2024-01-19, from http://journal.frontiersin.org/article/10.3389/fnhum.2016 doi: 10.3389/fnhum.2016.00604

de Lange, F. P., Heilbron, M., & Kok, P. (2018, September). How Do Expectations Shape Perception? *Trends in Cognitive Sciences*, *22*(9), 764–779. Retrieved 2022-10-11, from https://linkinghub.elsevier.com/retrieve/pii/S13646613183 doi: 10.1016/j.tics.2018.06.002

Friston, K. (2010, February). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. Retrieved 2024-01-23, from https://www.nature.com/articles/nrn2787 doi: 10.1038/nrn2787

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & De Lange, F. P. (2022, August). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, *119*(32), e2201968119. Retrieved 2023-06-12, from https://pnas.org/doi/full/10.1073/pnas.2201968119 doi: 10.1073/pnas.2201968119

Jaramillo, S., & Zador, A. (2010, October). Auditory cortex

mediates the perceptual effects of acoustic temporal expectation. *Nature Precedings*. Retrieved 2025-02-05, from https://www.nature.com/articles/npre.2010.5139.1 doi: 10.1038/npre.2010.5139.1

King, J.-R., Charton, F., Lopez-Paz, D., & Oquab, M. (2020, October). Back-to-back regression: Disentangling the influence of correlated factors from multivariate observations. *NeuroImage*, *220*, 117028. Retrieved 2025-01-22, from https://linkinghub.elsevier.com/retrieve/pii/S1053811920305140 doi: 10.1016/j.neuroimage.2020.117028

Kok, P., Jehee, J., & de Lange, F. (2012, July). Less Is More: Expectation Sharpens Representations in the Primary Visual Cortex. *Neuron*, *75*(2), 265–270. Retrieved 2024-12-18, from https://linkinghub.elsevier.com/retrieve/pii/S0896627312004382 doi: 10.1016/j.neuron.2012.04.034

Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*. Retrieved 2025-01-22, from http://journal.frontiersin.org/article/10.3389/neuro.06.004.2008/abstract doi: 10.3389/neuro.06.004.2008

Millidge, B., Seth, A., & Buckley, C. L. (2021). Predictive Coding: a Theoretical and Experimental Review. Retrieved 2023-11-13, from https://arxiv.org/abs/2107.12979 doi: 10.48550/ARXIV.2107.12979

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics. Retrieved 2023-02-10, from http://aclweb.org/anthology/D14-1162 doi: 10.3115/v1/D14-1162

Press, C., Kok, P., & Yon, D. (2020, January). The Perceptual Prediction Paradox. *Trends in Cognitive Sciences*, *24*(1), 13–24. Retrieved 2023-11-15, from https://linkinghub.elsevier.com/retrieve/pii/S136466131930261X doi: 10.1016/j.tics.2019.11.003

Raunak, V., Gupta, V., & Metze, F. (2019). Effective Dimensionality Reduction for Word Embeddings. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)* (pp. 235–243). Florence, Italy: Association for Computational Linguistics. Retrieved 2023-02-10, from https://www.aclweb.org/anthology/W19-4328 doi: 10.18653/v1/W19-4328