Sub-lexical processing of audiovisual speech constrains lexical competition.

Aaron Nidiffer (aaron.nidiffer@rochester.edu)

Department of Neuroscience, University of Rochester
Del Monte Institute for Neuroscience, University of Rochester
Rochester, NY, 14627, USA

Edmund C Lalor (elalor@ur.rochester.edu)

Department of Biomedical Engineering, University of Rochester
Department of Neuroscience, University of Rochester
Del Monte Institute for Neuroscience, University of Rochester
Rochester, NY, 14627, USA

Abstract:

As we listen to natural connected speech, we effortlessly transform speech acoustics into linguistic units. As that transformation begins, so does a lexical inference process that updates as each phoneme is uttered. In noisy environments, this process can become disrupted by poor inference at the phoneme level, leading to increased lexical competition and reduced word comprehension. Seeing a speaker's face can restore comprehension, in part by constraining the competition to words consistent with auditory and visual speech. There is evidence that vision can constrain inference at the lexical level, but it is unknown whether those effects can be attributed to sub-lexical interactions or whether constraint happens only after auditory and visual lexical processes are complete. In this study we fit and evaluate EEG encoding models of lexical competition that vary depending on acoustic and visual uncertainty, and the constraint imposed by their set intersection. We use linear modeling of electroencephalography and find evidence that audiovisual lexical processing is affected by visual constraint as a word unfolds.

Keywords: audiovisual speech; EEG; natural language processing; lexical selection

Introduction

Human speech perception is a remarkable feat. Seemingly effortlessly, we process on-going streams of complex spectro-temporal acoustic patterns and rapidly classify them into phonemic categories (Di Liberto et al., 2015; Marslen-Wilson & Warren, 1994). As we infer which word is being spoken, the brain integrates incoming phonemes in real time, removing competing lexical representations that are inconsistent with the current sequence of phonemes (Brodbeck et al., 2018).

In a noisy environment, our perception of acoustic speech suffers, leading to phoneme confusion and word recognition errors. We benefit from seeing the face of the person speaking (Ross et al., 2007), in part by utilizing a noise-robust, complementary visual speech representation (e.g., visemes; Campbell, 2008) to constrain the inference of the phonemes and word being spoken (Campbell, 2008; Peelle & Sommers, 2015). One consequence of the complementarity of auditory and visual speech cues is that they produce unique lexical activations, with behavior driven by the number of lexical competitors of the target word (Luce & Pisoni, 1998; Mattys et al., 2002). For example, "moon" and "noon" are auditory, but not visual, lexical competitors, and vice-versa for "bid" and "pit." During audiovisual presentations, word comprehension depends on the intersection of those activations, and the removal of lexical competitors inconsistent with either auditory or visual input (Tye-Murray et al., 2007).

The visual system generates its own linguistic representation that can support rudimentary lexical detection (Nidiffer et al., 2023) and enhance phonetic representations in the auditory system (O'Sullivan et al., 2021). However, it's unclear whether each incoming viseme can constrain the on-going lexical competition or whether competition is resolved separately for auditory and visual systems before being integrated at the lexical level. In this study, we explore the possibility that visual sub-lexical processes provide early resolution of lexical competition, as the word unfolds.

Methods

EEG Recordings. We reanalyzed preprocessed data from two publicly available datasets (Crosse et al., 2015, 2016) containing EEG responses from two groups of 21 individuals who watched videos of continuous and natural auditory-only (A), visual-only (V), and audiovisual (AV) speech in quiet (group 1) and in noise (-9 dB SNR; group 2).

Stimulus Features. Using the SUBTLEX corpus (Brysbaert & New, 2009) and the Carnegie Mellon Pronouncing Dictionary, we quantified a measure, cohort entropy, reflecting the evolution of lexical competition during the utterance of a word. Briefly, it is the Shannon entropy of the cohort of words consistent with input up to the *ith* phoneme of that word:

$$H_i - \sum_{word}^{cohort_i} p_{word} \log p_{word}$$

where $\boldsymbol{p_{word}}$ is the relative probability of the current word in cohort *i*.

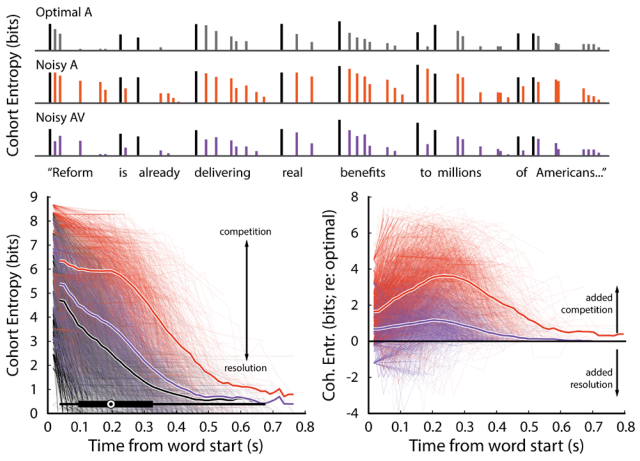

Figure 1: Cohort Entropy models calculated for auditory and audiovisual speech.

To simulate differing listening conditions and visual perceptual sensitivity, we computed this measure using several versions of the phonemic transcription that differed in their assumptions about phoneme confusion (Figure 1, top): 1) Optimal phonemic inference (after Brodbeck et al., 2018); 2) errors based on acoustic similarity defined by one differing phonetic feature (e.g., /m/ and /n/ differ only by place of articulation; Bailey & Hahn, 2005); 3) errors based on visual similarity, i.e., 12 phoneme equivalence classes as defined by Auer and Bernstein (1997; not shown), and 4) the intersection of cohorts (2) and (3), representing visual constraint on auditory lexical processing. The lower panels in Figure 1 show cohort entropy resolved over each word's utterance (left; box and whiskers indicate word durations) and the competition added by noise (right). The framework predicts that noisy acoustics results in increased competition that is largely resolved by visual speech.

Encoding Models

We used the mTRFtoolbox (Crosse, Di Liberto, Bednar, et al., 2016) to fit cross-validated encoding models that predicted EEG responses from the abovementioned stimulus features and evaluate their performance against ground-truth EEG. To assess non-additivity in AV EEG, we also attempted to predict AV responses with an A-only encoder, a V-only encoder, and an A+V encoder (Crosse et al., 2015).

Results and Discussion

For each speech condition, we evaluated a phoneme onset model and the respective cohort entropy model on their ability to predict new EEG. Here, the onset model acts as a baseline condition to account for EEG that responds just because any phoneme has been uttered. Figure 2a shows EEG prediction accuracy across the scalp for the phoneme onset model during auditory-only and audiovisual speech in noise. These data reveal topographies that are consistent with primary and non-primary auditory generators, with additional contributions coming from occipital scalp during audiovisual speech. We evaluated a model consisting of both phoneme onsets and cohort entropy which revealed additional predictive value of cohort entropy in a cluster on the right temporal scalp (Figure 2b, significant electrodes marked, ps<0.05), a signature of sub-lexical visual constraint imposed on lexical processing.

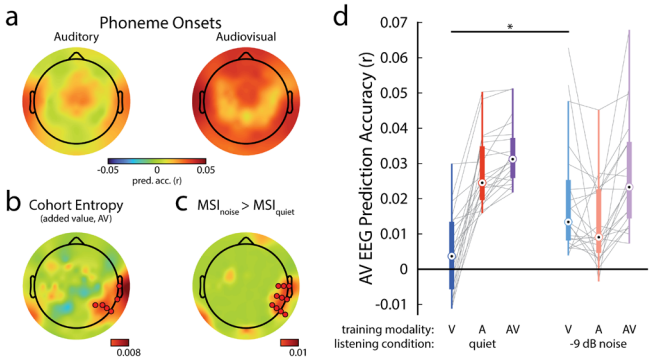

Figure 2: Model evaluation

We then investigated the extent of non-linear multisensory interactions which are expected to occur based on the complementary nature of phonemic and

visemic representations and the hypothesized intersection that imposes the lexical constraint. To do so, using the cohort entropy features, we fit an audiovisual encoder model and constructed an additive multisensory (A + V) model, and compared their abilities to predict left-out AV EEG. Our hypothetical framework predicts a larger potential for visual constraint (and thus more multisensory non-linearities) when acoustic phoneme inference is poor, so we also contrasted these values between noisy and quiet conditions. We found a cluster of electrodes (Figure 2c; ps<0.05) with larger multisensory effects during noisy speech that largely overlapped with the region where cohort entropy adds value to EEG predictions.

Finally, we sought to examine the contribution of visual speech to the lexical selection process. To that end, we used the abovementioned single-modality (A and V) encoder models and separately tested their ability to predict AV EEG (Figure 2d). This analysis provides insight into the relative contribution from unisensory modalities. First, trivially, the acoustic cohort entropy model performance decreases with noise ($T = 4.3$, $p = 8\times10^{-5}$). We do not interpret this decrease as meaningfully reflecting lexical selection because the difference in listening condition leads to a general decrease of cortical tracking. However, the visual stimulus was identical in both cases, yet the visual model was able to predict more EEG activity when the acoustics were noisy ($T = 3.6$, $p = 9.2\times10^{-4}$). We interpret these findings to reflect, generally, an increased reliance on visual cues when acoustics are degraded and, more specifically, an increased capacity for visual cues to constrain phonemic and lexical inference in those situations.

Acknowledgments

This work was supported by NIH Grant DC016297. We are appreciative to Jin Dou for feedback during the development of this project.

References

Auer, E. T., & Bernstein, L. E. (1997). Speechreading and the structure of the lexicon: Computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness. *The Journal of the Acoustical Society of America*, *102*(6), 3704–3710. https://doi.org/10.1121/1.420402

Bailey, T. M., & Hahn, U. (2005). Phoneme similarity and confusability. *Journal of Memory and Language*, *52*(3), 339–362. https://doi.org/10.1016/J.JML.2004.12.003

Brodbeck, C., Hong, L. E., & Simon, J. Z. (2018). Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech. *Current Biology*, *28*(24), 3976-3983.e5. https://doi.org/10.1016/j.cub.2018.10.042

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977/METRICS

Campbell, R. (2008). The processing of audio-visual speech: empirical and neural bases. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 1001–1010. https://doi.org/10.1098/rstb.2007.2155

Crosse, M. J., Butler, J. S., & Lalor, E. C. (2015). Congruent Visual Speech Enhances Cortical Entrainment to Continuous Auditory Speech in Noise-Free Conditions. *Journal of Neuroscience*, *35*(42), 14195–14204. https://doi.org/10.1523/JNEUROSCI.1829-15.2015

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. *Frontiers in Human Neuroscience*, *10*, 604. https://doi.org/10.3389/fnhum.2016.00604

Crosse, M. J., Di Liberto, G. M., & Lalor, E. C. (2016). Eye Can Hear Clearly Now: Inverse Effectiveness in Natural Audiovisual Speech Processing Relies on Long-Term Crossmodal Temporal Integration. *Journal of Neuroscience*, *36*(38), 9888–9895. https://doi.org/10.1523/JNEUROSCI.1396-16.2016

Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, *25*(19), 2457–2465. https://doi.org/10.1016/j.cub.2015.08.030

Luce, P. A., & Pisoni, D. B. (1998). Recognizing Spoken Words: The Neighborhood Activation Model. *Ear and Hearing*, *19*(1), 1–36. https://doi.org/10.1097/00003446-199802000-00001

Marslen-Wilson, W., & Warren, P. (1994). Levels of perceptual representation and process in lexical access: Words, phonemes, and features. *Psychological Review*, *101*(4), 653–675. https://doi.org/10.1037/0033-295X.101.4.653

Mattys, S. L., Bernstein, L. E., & Auer, E. T. (2002). Stimulus-based lexical distinctiveness as a general word-recognition mechanism. *Perception and Psychophysics*, *64*(4), 667–679. https://doi.org/10.3758/BF03194734/METRICS

Nidiffer, A. R., Cao, C. Z., O'Sullivan, A., & Lalor, E. C. (2023). A representation of abstract linguistic categories in the visual system underlies successful lipreading. *NeuroImage*, *282*, 120391. https://doi.org/10.1016/J.NEUROIMAGE.2023.120391

O'Sullivan, A. E., Crosse, M. J., Liberto, G. M. Di, Cheveigné, A. de, & Lalor, E. C. (2021). Neurophysiological Indices of Audiovisual Speech Processing Reveal a Hierarchy of Multisensory Integration Effects. *Journal of Neuroscience*, *41*(23), 4991–5003. https://doi.org/10.1523/JNEUROSCI.0906-20.2021

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. In *Cortex* (Vol. 68, pp. 169–181). Masson SpA. https://doi.org/10.1016/j.cortex.2015.03.006

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex*, *17*(5), 1147–1153. https://doi.org/10.1093/cercor/bhl024

Tye-Murray, N., Sommers, M., & Spehar, B. (2007). Auditory and Visual Lexical Neighborhoods in Audiovisual Speech Perception: *Trends in Amplification*, *11*(4), 233–241. https://doi.org/10.1177/1084713807307409