Visual speech processing interfaces with areas selective for spoken words in ventral and dorsal speech pathways

Lillian Chang[1], Srikanth Damera[1], Benson W. Stevens[1], Silvio P. Eberhardt[2], Edward T. Auer Jr[2]., Lynne E. Bernstein[2], Maximilian Riesenhuber[1]

[1]Georgetown University Medical Center — Department of Neuroscience, [2]George Washington University – Department of Speech, Language & Hearing Sciences

Introduction

Neural basis of auditory speech processing: a dual stream model[1,2] (Fig 1A)

- Ventral pathway: spoken word recognition (e.g., a so-called auditory word form area (AWFA) with lexical representations for spoken words[3-5], paralleling the visual word form area for reading).
  - Poster 241.23 at SFN this year.
- Dorsal pathway: sensorimotor integration via a bidirectional connection between the posterior superior temporal cortex (pSTC) and the premotor cortex (PMC)[6,7] (e.g., mapping speech sounds to their corresponding motor articulations through an "inverse" model)[8].

Neural basis of visual speech processing (Fig 1B)

- The so-called temporal visual speech area (TVSA) in the ventral pSTC (Fig 1B) responds more to visual speech than non-speech gestures[9,10].
- This region is part of a third, lateral visual pathway posited to be important for social perception[11].

Objective

- Although visual cues (i.e., lipreading) provide important information for articulation and perceiving speech, it is unclear if and how the TVSA interfaces with the dual-stream pathways for speech processing.
- We hypothesized that visual speech interfaces with both streams — the AWFA and PMC showing selectivity for speech over non-speech stimuli, as well as coupling between these areas and the TVSA (Fig. 1C).
- To test our hypotheses, we conducted two localizer scans to identify the AWFA, PMC, and TVSA for each individual subject. We then examined the crossmodal activation profiles and functional connectivity between our three ROIs.

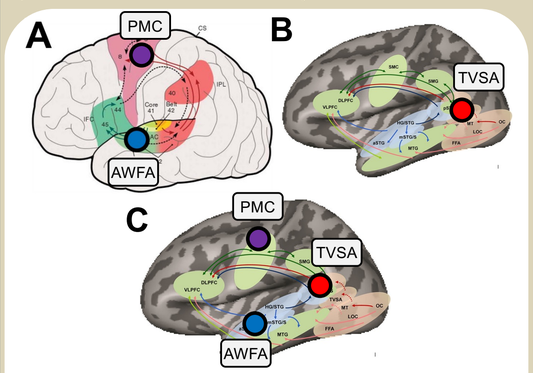

Fig 1A. Model of auditory speech processing from Rauschecker & Scott (2009)[1] with the AWFA[3] (blue) and PMC (purple) included. **B.** Neuroanatomical model of audiovisual speech perception in the left hemisphere. Audiovisual speech is processed in auditory (blue) and visual (pink) areas projecting to amodal (green) regions. Figure adapted from Bernstein & Liebenthal (2014)[9]. TVSA location (red) based on Bernstein et al. (2011)[8]. **C.** Regions-of-interest from the auditory and visual speech processing models as part of our hypotheses.

Experimental Design

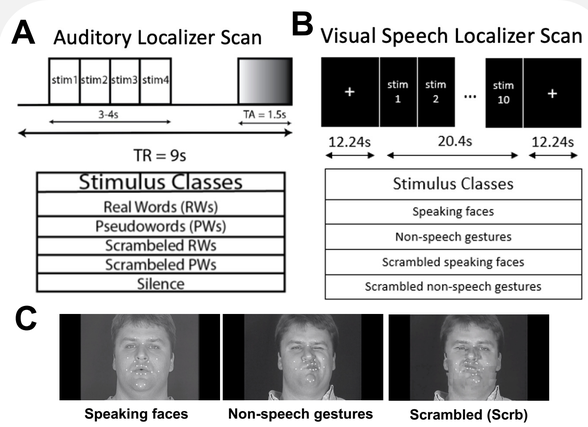

Figure 2: Localizer Experimental Paradigms. A. Slow ER paradigm used in the auditory localizer scan. Figure adapted from Damera et al. (2022)[4]. **B.** Continuous acquisition paradigm used for the visual speech localizer scan. **C.** Example stimuli for the visual speech localizer, first used in Bernstein et al. (2011)[8]. The "scrambled" control stimuli were created by scrambling the temporal order of the video frames from the stimuli in the first two conditions.

Subjects: 20 right-handed English speakers (15 females). Ages 18-26.

Auditory word localizer (Fig. 2A)

- Task: Participants responded with a button press if any of the 4 stimuli in a short block was a repeat.

Visual speech localizer (Fig. 2B)

- Task: 1-back. Subjects responded with a button press if any of the stimuli was an immediate repeat.

Acquisition Parameters:

- 3T Siemens TIM Trio scanner with 12-channel head coil (N=17).
- 3T Siemens Prisma scanner with 64-channel head coil (N=3).
- Auditory word localizer: TR=9s; TA=1.5s.
- Visual speech localizer: TR=2.04s.

Defining Individual-subject ROIs

Regions-of-interest (ROIs) were defined through contrast masking at the first-level analyses based on methods from the literature (SPM12).

Contrasts

- AWFA & PMC: RW versus Silence (p<0.001, uncorrected) masked by the contrast of RW versus Scrambled Words (p-uncorrected<0.05).
- TVSA: Speech versus Gesture (p<0.001, uncorrected) masked by Speech versus Scramble (p<0.05, uncorrected).

ROI clusters and sizes: Thresholds were adjusted to obtain clusters closest to the literature coordinates that were between 10 and 50 voxels.

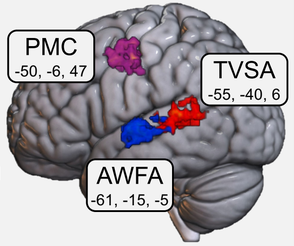

Fig 3. Individual-subject ROIs. AWFA: auditory word form area, PMC: premotor cortex, TVSA: temporal visual speech area. Literature coordinates were used as reference locations when choosing clusters for individual-subject ROIs.

Auditory Regions Selective for Visual Speech

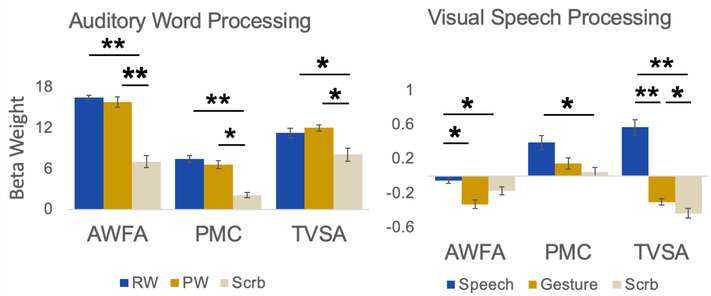

Fig 4. Activation profiles for **A.** auditory word processing and **B.** visual speech processing. One-tailed paired t-tests. p*<0.05, p**<0.0001. Error bars indicate within-subject SEM.

- Significant activation differences in the AWFA during visual speech processing compared to gesture (p=0.0003, one-tailed) and the scrambled control stimuli (p=0.0357, one-tailed).
- The PMC shows greater selectivity for visual speech over the control (p=0.0099, one-tailed).

Functional Connectivity Between Auditory and Visual Speech Regions

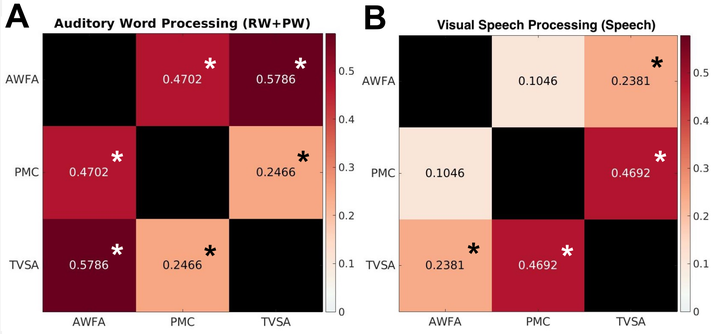

Fig 5. Correlation coefficient values representing ROI-to-ROI stimulus driven functional connectivity **A.** auditory word processing and **B.** visual speech processing. Connection threshold: p-uncorrected<0.05, two-sided. Cluster threshold: ROI-level p-FDR <0.05. Asterisks indicate significant connectivity at the thresholds.

Auditory word processing

- Functional connectivity between the TVSA with the PMC (p-FDR=0.0343) and the AWFA (p-FDR=0.00034) during auditory word processing.

Visual speech processing

- There is significant functional connectivity between the TVSA with the PMC (p-FDR=0.001) and the AWFA (p-FDR=0.0153).
- However, no significant connectivity between the AWFA and PMC during visual speech processing (p-FDR=0.247).

Conclusions

- In both the AWFA and the PMC, we found greater activation for speech over gestures and our control stimuli during visual speech processing. There is also significant functional connectivity between the TVSA and AWFA, and the TVSA and PMC.
- Our activation and functional connectivity provides evidence that the TVSA interfaces with both auditory speech processing streams when perceiving visual speech (i.e., lipreading).
- Our findings provide further insight into how visual speech processing couples with the dual-stream auditory speech network.
  - Integration with the dorsal stream could subserve the learning of speech articulation as part of the "inverse" model[8].
  - Integration into the ventral stream is compatible with this stream's role as the speech comprehension pathway, where the AWFA as a lexicon for spoken words could be a substrate for lipreading.

Acknowledgements

- Research supported by the National Science Foundation (BCS-1756313) and Georgetown University.
- Dr. Ian DeWitt for the auditory scrambling code used in the auditory word localizer.

References

1. Rauschecker, J. P., & Scott, S. K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature neuroscience, 12*(6), 718-724.
2. Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature reviews neuroscience, 8*(5), 393-402.
3. DeWitt, I., & Rauschecker, J. P. (2012). Phoneme and word recognition in the auditory ventral stream. *Proceedings of the National Academy of Sciences, 109*(8), E505-E514.
4. Damera, S. R., Mattei, J., Glezer, L., Cox, P., Jiang, X., Rauschecker, J., & Riesenhuber, M. A Lexicon in the Anterior Auditory Ventral Stream: Preliminary evidence from an fMRI-RA study [abstract]. In: Eleventh Annual Meeting of the Society for the Neurobiology of Language; 2019 Aug 20-22; Helsinki, Finland. Abstract E66.
5. Damera, S. R., Chang, L., Nikolov, P., Mattei, J. A., Banerjee, S., Glezer, L. S., Cox, P. H., Jiang, X., Rauschecker, J. P., Riesenhuber, M. (2022). Evidence for a Spoken Word Lexicon in the Auditory Ventral Stream. *bioRxiv*.
6. Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature neuroscience, 7*(7), 701-702.
7. Meister, I. G., Wilson, S. M., Deblieck, C., Wu, A. D., & Iacoboni, M. (2007). The essential role of premotor cortex in speech perception. Current Biology, 17(19), 1692-1696.
8. Chevillet, M. A., Jiang, X., Rauschecker, J. P., & Riesenhuber, M. (2013). Automatic phoneme category selectivity in the dorsal auditory stream. *Journal of Neuroscience*, *33*(12), 5208-5215.
9. Bernstein, L. E., Jiang, J., Pantazis, D., Lu, Z. L., & Joshi, A. (2011). Visual phonetic processing localized using speech and nonspeech face gestures in video and point-light displays. *Human brain mapping*, *32*(10), 1660-1676.
10. Bernstein, L. E., & Liebenthal, E. (2014). Neural pathways for visual speech perception. *Frontiers in neuroscience*, *8*, 386.
11. Pitcher, D., & Ungerleider, L. G. (2021). Evidence for a third visual pathway specialized for social perception. *Trends in Cognitive Sciences*, *25*(2), 100-110.