Properties of distracting speech affect attentional entrainment to audiovisual "cocktail-party" speech

Patrik Wikman (patrik.wikman@helsinki.fi)[1]
[1]Department of Psychology and Logopedics, PO Box 21 FI-00014 University of Helsinki, Finland

Alessandra Bombino (alessandra.bombino@helsinki.fi)[1]
[1]Department of Psychology and Logopedics, PO Box 21 FI-00014 University of Helsinki, Finland

Artturi Ylinen (artturi.ylinen@helsinki.fi)[1]
[1]Department of Psychology and Logopedics, PO Box 21 FI-00014 University of Helsinki, Finland

Ilkka Muukkonen (ilkka.muukkonen@helsinki.fi)[1]
[1]Department of Psychology and Logopedics, PO Box 21 FI-00014 University of Helsinki, Finland

The attentional effects caused by characteristics of unattended speech in cocktail-party settings are poorly understood. We measured EEG (n = 19) and fMRI (n = 20) to naturalistic audiovisual dialogues with concurrent distracting speech, varying in semantic and physical similarity to the attended speaker. We used EEG speech reconstruction analysis to study how the temporal dynamics of selective attention depended on features of the unattended speech, across and within sentences. For the fMRI data we used representational similarity analysis, Procustes analysis, and hierarchical clustering to explore spatiotemporal changes in attentional modulation across sentences. Attentional entrainment to the relevant speech stream was affected by the properties of distracting speech. The fMRI data revealed that the representational structure of attended speech changed distinctly across time in different nodes of the speech processing network. We underscore the value of using multiple measurement techniques, incorporating different spatial and timescales, in attention research.

Keywords: "cocktail-party" speech, attention, distracton, fMRI, EEG , RSA, hyperaligement, hierarchical clustering

Introduction

M/EEG and fMRI studies have shown that neural entrainment to attended speech is enhanced in auditory scenes with concurrent speakers. This entrainment depends on both the physical and linguistic properties of attended speech (Broderick et al., 2019; Puschmann et al., 2024; Wikman et al., 2024). In contrast, how properties of distracting speech affect entrainment is poorly understood (Kaufman & Golumbic, 2023). Early behavioural and neurophysiological experiments suggest that mainly physical features (e.g., pitch) of distracting speech affect attentional selection (Broadbent, 1954). Effects of linguistic properties, e.g., semantic similarity between distracting and attended speech, are small and transient, occurring only in the beginning sentences (see: Näätänen et al., 1992).

We studied how semantic and physical features of distracting speech affect attentional modulation of attended speech using EEG and fMRI. We used audiovisual (AV) video clips of dialogues, each containing seven consecutive sentences (lines), with a concurrent distracting background speaker. Importantly, the distracting speech varied in semantic and physical similarity to the attended speech. We used speech envelope reconstruction (SER) (Crosse et al., 2016) to estimate from the EEG data how the properties of the distracting speech affected neural tracking of attended speech, both within each line and across lines of the dialogue. Representational similarity analysis (RSA) was used on the fMRI data to show how attentional modulation changed as a function of properties of distracting speech across the dialogue.

Methods

The stimuli were 55-65s long AV dialogues. Each dialogue contained 7 lines (~5.4s/line, pause between lines ~3.4s) alternately spoken by a male and female speaker. The distractor speech stream and attended speech stream had the same onset in each line, and the distractor speech stream was 6 dB louder. The distracting speech was semantically equivalent (but used different words, Sem. related) or different (Sem. unrelated) and was spoken by a speaker of the same sex (same fundamental frequency; i.e., physically similar) or opposite sex to the attended speaker. The fMRI experiment included the same attend speech task and conditions as the EEG experiment. However, the fMRI experiment additionally included an ignore speech version of all conditions (ignore speech task). In this task participants counted rotations of a visual cross presented below the speakers' faces and ignored the dialogue and the distracting speech. The experiments comprised 32–36 dialogues. Dialogues were randomly allocated to each condition.

EEG (n = 19, 6 males) preprocessing included band-pass filtering (1-10 Hz), downsampling (64 Hz) and ICA (to remove artefacts). Hilbert transform was used to extract speech amplitude envelopes separately for the attended and distractor speech streams (same sample rate and passband). Thereafter, all 128 EEG channels were used as input in the SER analysis (leave-one-trial-out), with time lags of -200-0ms, common regularization parameter $\lambda=10^4$. SERs were estimated separately for 4 segments of each line of each dialogue.

Preprocessed (fMRIprep) fMRI data (n = 20, 5 males) were projected to the fsaverage surface space (Fischl, 2012). All experimental conditions (7 lines x 8 conditions) were included in the GLM as separate regressors. Representational dissimilarity matrices (RDMs) were constructed from the vertex patterns across all conditions and vertices separately for the 360 ROIs of the HCP parcellation (Glasser et al., 2016), and correlated with different model RDMs (semantic, sex, line-number and their interactions). Because we were interested in how attention modulated neural patterns dependent on our experimental conditions, separate RDMs were always calculated for the attend speech task and ignore speech task, and we report only where the attend speech task caused stronger model correlations than the ignore speech task.

Results

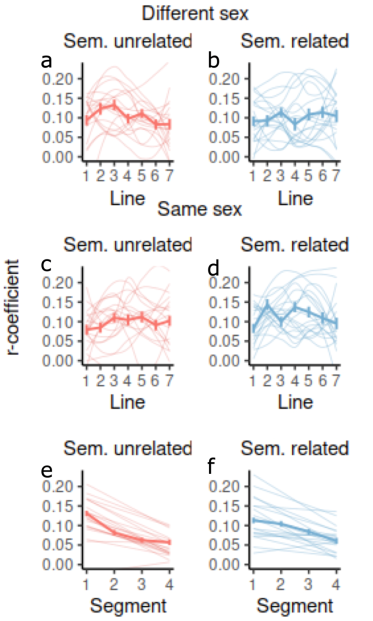

Figure 1. **(**a-d) SER accuracy of attended speech changed between seven lines of the dialogues in a nonlinear fashion. This trend depended on both the semantic similarity and the sex of the distractor. (e-f) SER accuracy linearly decreased as the line progressed. The slope of the linear decrease depended on semantic similarity (± SEM). Transparent lines represent subject trajectories.

As expected, in the EEG data SER accuracy for attended speech changed as a function of line number, semantic relatedness, and sex of the unattended speaker ($F_{6,108} = 2.4$, $p < 0.05$) (Figure 1, a–d). Further, semantic similarity also affected SER accuracy within the line of the attended dialogue (F3,54 = 4.4, $p < 0.01$) (Figure 1 (e-f).

The line-number model showed stronger correlations with fMRI data RDMs during attended speech than during ignored speech across sensory, frontal and sensomotor brain regions (FDR-corrected, $p < .05$, Figure 2). The semantic, sex and all interaction model correlations for the two tasks did not differ in any of the 360 ROIs.

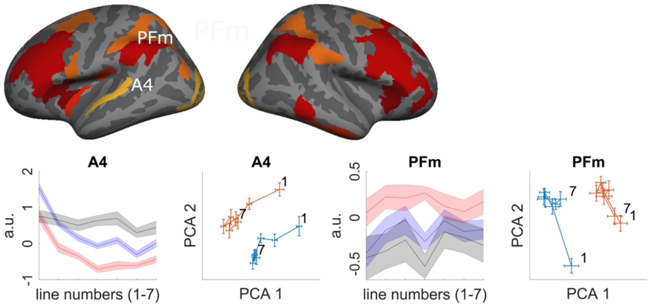

Figure 2: *Upper:* ROIs where correlations with the line model were significantly stronger during the attend speech than the ignore speech task (FDR-corrected $p < .05$). To visualize the neural similarities between the significant ROIs, we aligned the vertex patterns of each ROI-pair using Procrustes analysis, performed hierarchical clustering (Ward's method) based on the Procrustes distances and defined six clusters (different shades between red and yellow). *Lower:* (1st and 3rd image) The average ROI signal (across vertices) during the attend speech task (blue), ignore speech task (red) and their difference (grey) across the lines of the dialogues for two example ROIs (± SEM). (2nd and 4th image) The first two principal components (PC), derived from vertex patterns hyperaligned (Generalized Procrustes) across participants (cross-validated, split-half). The average scores are displayed for the respective line of attended (blue) and ignored speech (red) dialogues (± SEM).

Discussion

We show with our EEG data analysis that neural entrainment to attended speech shows a linear decrease within each line of the dialogues (Figure 1, e-f). We suggest that this linear decrease reflects a interaction between attention and speech related prediction errors in the auditory cortex. Importantly, however, we found that decreases in SER were most consistently observed for attended speech. (Wikman et al., 2024). As expected (Näätänen, 1990), semantic relatedness of the distracting speech affected this temporal profile in the beginning of the line. This may indicate that when the two speech streams were semantically similar, attentional allocation to the correct speech stream was delayed. We also show that neural entrainment to attended speech follows a nonlinear temporal profile across the lines of the dialogue (Figure 1, a–d). However, this profile was more labile than expected (Wikman et al., 2024), depending on characteristics of the distractor speech.

Our RSA on the fMRI data revealed that attentional modulation of vertex patterns across sensory, frontal and sensomotor cortical regions changed from line-to-line of the dialogue. This network corresponded broadly to previously reported regions that show univariate attentional changes in a line-to-line fashion using similar AV dialogues (Wikman et al., 2021). However, RSA revealed additional regions in the parietal cortex (PFm, Figure 2) that were not found in our previous study using only the mean signal change instead of neural patterns. Furthermore, as can be seen in Figure 2, attention did not modulate the mean signals between the different lines of dialogue in this study either. Instead, the vertex patterns in the two tasks behave differently across the lines of the dialogue. Thus, we highlight the importance of using both temporally (EEG) and spatially (fMRI) accurate brain research methods in combination with univariate and multivariate methods to gain a comprehensive understanding of spatiotemporal attentional modulation of naturalistic speech.

Acknowledgments

This work was supported by an Academy of Finland (grant number #297848 and #1348353) and the Finnish Cultural Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Broadbent, D. E. (1954). The Role of Auditory Localization in Attention and Memory Span. *Journal of Experimental Psychology*, *47*(3), 191-196. https://doi.org/10.1037/h0054182

Broderick, M. P., Anderson, A. J., & Lalor, E. C. (2019). Semantic Context Enhances the Early Auditory Encoding of Natural Speech. *Journal of Neuroscience*, *39*(38), 7564-7575. https://doi.org/10.1523/Jneurosci.0584-19.2019

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 of Human Neuroscience*, *10*, 604. https://doi.org/10.3389/fnhum.2016.00604

Fischl, B. (2012). FreeSurfer. *Neuroimage*, *62*(2), 774-781.

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., . . . Jenkinson, M. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), 171-178.

Kaufman, M., & Zion-Golumbic, E. (2023). Listening to two speakers: Capacity and tradeoffs in neural speech tracking during Selective and Distributed Attention. *NeuroImage*, *270*, 119984.

Näätänen, R. (1990). The role of attention in auditory information processing as revealed by event-related potentials and other brain measures of cognitive function. *Behavioral and brain sciences*, *13*(2), 201-233.

Näätänen, R., Teder, W., Alho, K., & Lavikainen, J. (1992). Auditory attention and selective input modulation: a topographical ERP study. *Neuroreport*, *3*, 493–496

Puschmann, S., Regev, M., Fakhar, K., Zatorre, R. J., & Thiel, C. M. (2024). Attention-driven modulation of auditory cortex activity during selective listening in a multi-speaker setting. *The Journal of Neuroscience*, e1157232023. https://doi.org/10.1523/jneurosci.1157-23.2023

Wikman, P., Sahari, E., Salmela, V., Leminen, A., Leminen, M., Laine, M., & Alho, K. (2021). Breaking down the cocktail party: Attentional modulation of cerebral audiovisual speech processing. *NeuroImage*, 117365.

Wikman, P., Salmela, V., Sjöblom, E., Leminen, M., Laine, M., & Alho, K. (2024). Attention to audiovisual speech shapes neural processing through feedback-feedforward loops between different nodes of the speech network. *PLoS Biology*, *22*(3), e3002534.