Decoding attention in polyphonic music listening

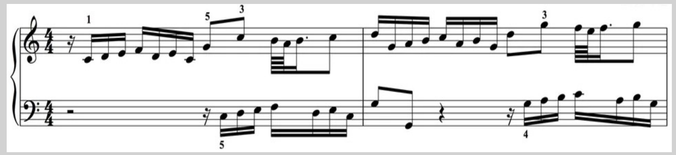
Lucas Klein, Aedan Rourke, Hany Tawfik, Dan Bosnyak & Laurel J. Trainor
Dept. of Psychology, Neuroscience & Behaviour, McMaster University

BACKGROUND

- Humans can distinguish and separate sounds from different sources within acoustically rich environments (known as the **cocktail party problem**).
- **Neural responses** encode spectrotemporal characteristics of sound envelopes (Aiken & Picton, 2008; Ding & Simon, 2012), and to a higher degree for **attended sounds** than unattended sounds.
- Single-trial EEG can be used to **decode** the target of auditory selective attention to continuous speech (Mesgarani & Chang, 2012; O'Sullivan et al., 2015), even while walking (Straetmans et al., 2021).
- Unlike competing talkers, **musicians coordinate** to compose separate music parts that fit together.
- **Music listening involves integrating sounds** to reveal musical elements, such as harmony and rhythm.

HIGH

LOW



Study Aims:

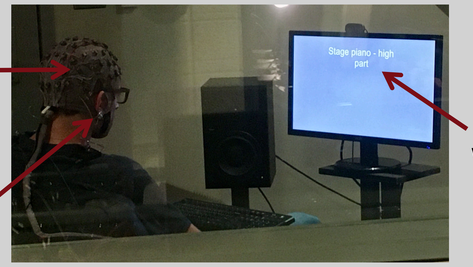
1. *Can the target of auditory attention to polyphonic music be decoded from single-trial EEG?*
2. *Does timbre combination affect the tendency to integrate rather than separate different parts?*

METHODS

- **Stimuli:** twelve 25-s clips of Bach's two-part Inventions (three clips from each of four different Inventions). For each Invention, the high and low parts (right hand and left hand, respectively) were played with a different timbre combination.
- **Procedure:** We collected 64-channel EEG at 256 Hz sampling rate while participants listened to each clip three times: while *(1)* attending to the high part and ignoring the low, *(2)* vice versa, and *(3)* attending to both.

EEG collected at 256 Hz, and synchronized to audio

Audio presented through in-ear headphones



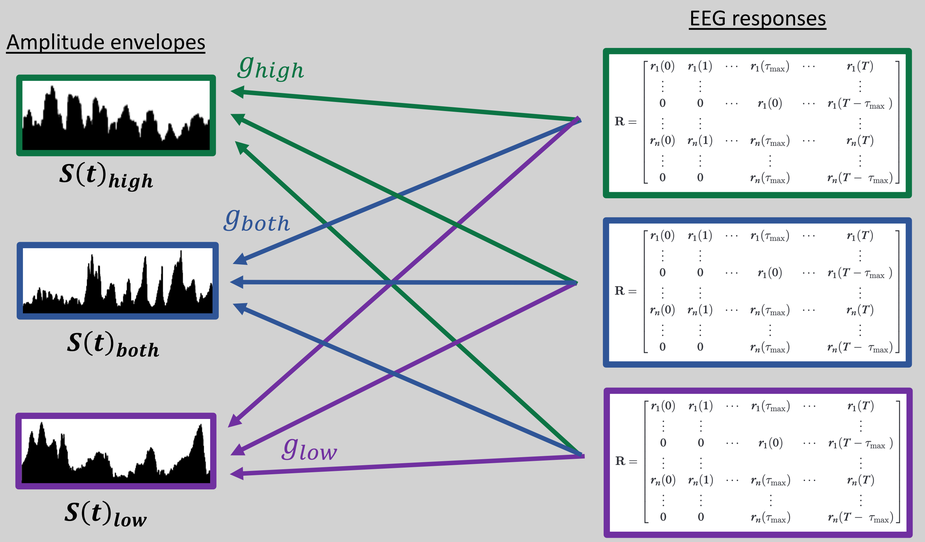
Instructions indicated which part to attend to

Data collection and analysis are ongoing

STIMULUS RECONSTRUCTION

Stimulus reconstruction uses a linear regression model that predicts a stimulus representation (e.g., amplitude envelope) from EEG. A multivariate temporal response function (mTRF) ($\boldsymbol{g}$) that maps electrode responses ($\boldsymbol{R}$) to an audio envelope ($\boldsymbol{S}$) (Crosse et al., 2016) is constructed by solving: $\boldsymbol{g} = \left(\boldsymbol{RR^T}\right)^{-1}\boldsymbol{RS^T}$

Reconstruction models ($\boldsymbol{g_{high}}$ and $\boldsymbol{g_{low}}$) are trained (using leave-one-out cross-validation) on EEG responses using the *high* and *low* isolated envelopes, respectively, and reconstructions (estimates) are calculated by convolving $\boldsymbol{g}$ with $\boldsymbol{R}$, in both cases.



$$\hat{S}(t) = \sum_{n}\sum_{\tau}[\ g(\tau,n) * R(t-\tau,n)\ ]$$

Estimated stimulus

Reconstruction model (g) consisting of a set of weights for each electrode and each time lag

Response of electrode n at time $t = 1 \ldots T$ for time lags $\tau = 1 \ldots \tau_{max}$

DECODING

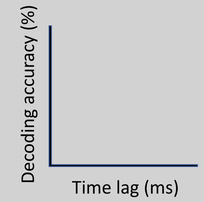
To identify the target to attention, calculate Pearson's r between each reconstruction ($\hat{\boldsymbol{S}}$) and each (isolated or combined) envelope ($\boldsymbol{S}$):

- $r\left(\hat{\boldsymbol{S}}(\boldsymbol{t})_{\boldsymbol{high}}, \boldsymbol{S}(\boldsymbol{t})_{\boldsymbol{high}}\right)$
- $r\left(\hat{\boldsymbol{S}}(\boldsymbol{t})_{\boldsymbol{low}}, \boldsymbol{S}(\boldsymbol{t})_{\boldsymbol{low}}\right)$
- $r\left(\hat{\boldsymbol{S}}(\boldsymbol{t})_{\boldsymbol{both}}, \boldsymbol{S}(\boldsymbol{t})_{\boldsymbol{both}}\right)$

max → **target of attention**

Decoding accuracy = percentage of trials for which the maximum Pearson correlation corresponds to the correct condition

Decoding time-course = decoding accuracy across time lags



HYPOTHESES

Given that the auditory cortex represents spectrotemporal features of attended sounds to a higher degree than unattended, we expect that:

1. Reconstructions of the **attended part** will be more highly correlated with those sounds' amplitude envelopes.
2. Reconstructions of the **combined audio** will have the highest correlations across conditions, but will be maximal for similar timbre combinations
3. Decoding **time courses** will reveal peak decoding accuracies at lags of approximately 200-250 ms (following O'Sullivan et al. (2015))

REFERENCES

Aiken, S. J., & Picton, T. W. (2008). Human Cortical Responses to the Speech Envelope. *Ear and Hearing*, *29*(2), 139–157. https://doi.org/10.1097/aud.0b013e31816453dc

Crosse, M. J., Liberto, G. M. D., 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

Ding, N., & Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, *109*(29), 11854–11859. https://doi.org/10.1073/pnas.1205381109

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*(7397), 233–236. https://doi.org/10.1038/nature11020

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A., & Lalor, E. C. (2015). Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. *Cerebral Cortex*, *25*(7), 1697–1706. https://doi.org/10.1093/cercor/bht355

Straetmans, L., Holtze, B., Debener, S., Jaeger, M., & Mirkovic, B. (2021). Neural tracking to go: auditory attention decoding and saliency detection with mobile EEG. *Journal of Neural Engineering*, *18*(6), 066054. https://doi.org/10.1088/1741-2552/ac42b5

ACKNOWLEDGEMENTS

NSERC-CREATE: Complex Dynamics of Brain and Behaviour

NSERC CRSNG

SSHRC CRSH
Social Sciences and Humanities Research Council of Canada
Conseil de recherches en sciences humaines du Canada