EEG-based decoding of auditory attention to a target instrument in polyphonic music

Giorgia Cantisani [1] Slim Essid [1] Gaël Richard [1]

[1]LTCI, Télécom Paris, Institut Polytechnique de Paris, 75013 Paris, France

MIPFrontiers

TELECOM Paris

IP PARIS

Summary

Auditory attention decoding aims at determining which sound source a subject is "focusing on". The goal is to determine the attended instrument based on 24-second long EEG excerpts aligned to corresponding audio stimuli.

Research Questions

- *Are we **tracking attention** or a general music entertainment?*
- *Are we tracking the **target instrument**?*
- *Which is the most suitable **audio descriptor** for such a task?*
- *How much **variants in the stimuli** influence the performances?*

State of the Art

Speech Stimuli

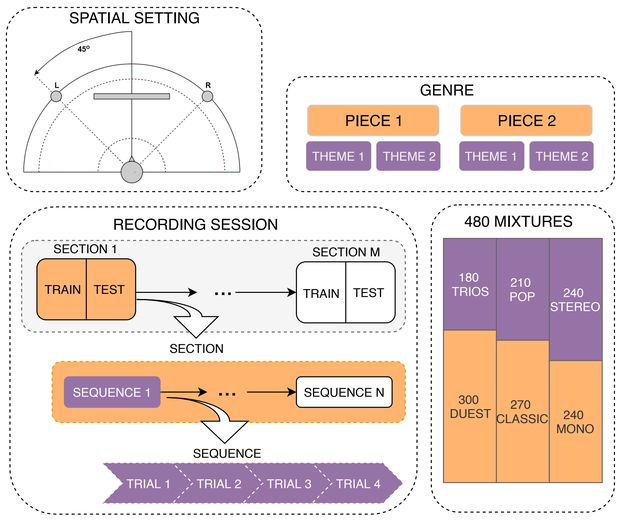
[1, 2] have shown that the EEG tracks dynamic changes in the speech stimulus and can be used to decode selective attention in a multispeaker environment.

Music Stimuli

Attention to speech is mostly semantic while attention to a musical instrument could stem from multiple factors (e.g. timbre, melody, rhythm, harmony, etc).

Data

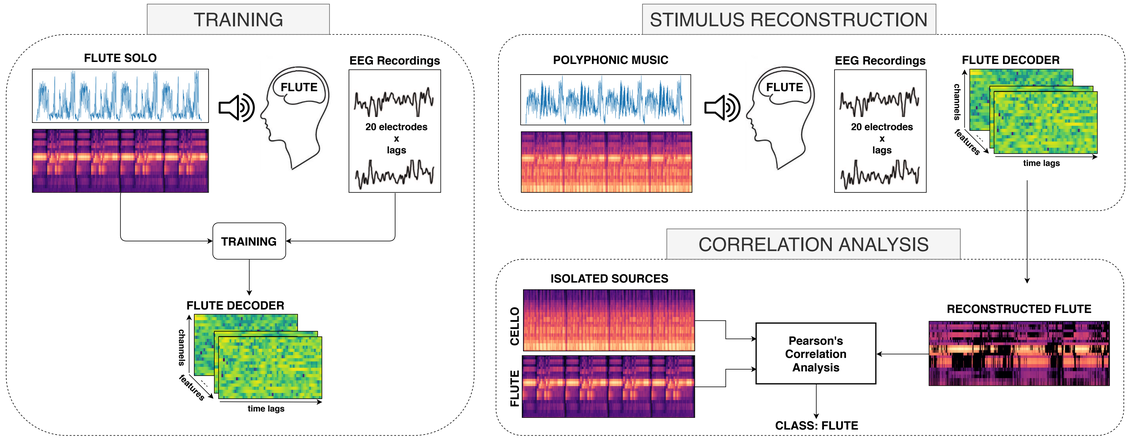
20-channel EEG signals recorded from 8 subjects while they were attending to a particular instrument in realistic polyphonic music using loudspeakers. The attended instrument is previously heard in solo, as part of a *training phase*.



Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 765068

Procedure



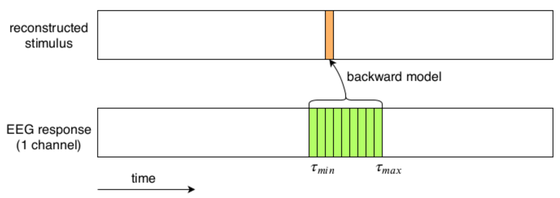
For each of the 8 subjects train on 14 solos, test on 40 duets and 24 trios.

Stimulus Reconstruction

A stimulus representation $\hat{s}$ is estimated from multi-channel neural data r through a model g which behaves like a *multi-channel Wiener filter*:

$$\hat{s}(t,f) = \sum_n \sum_\tau g(\tau, f, n) r(t-\tau, n)$$

The filter is learned by solving a **linear regression** problem: $\min \sum_t \sum_f [s(t,f) - \hat{s}(t,f)]^2$



Stimuli representations

- *Amplitude Envelope (AE)*
- *Magnitude spectrogram (MAG)*
- *Mel spectrogram (MEL)*

Conclusions

Take-home

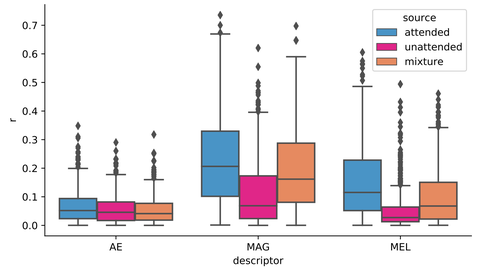
- the model is **tracking attention** and not a general entertainment to the music;
- the neural activity is correlated with **musically relevant features of the attended source**.
- benefits from TF audio representations which highlight amplitude modulations in different **frequency bands**.

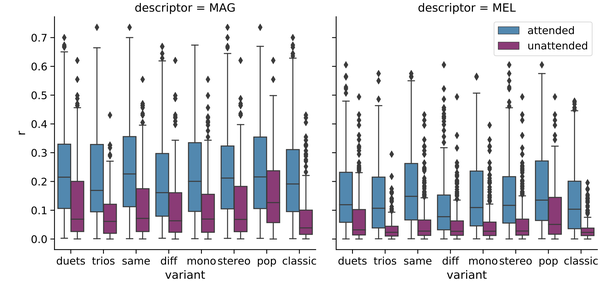
Limitations

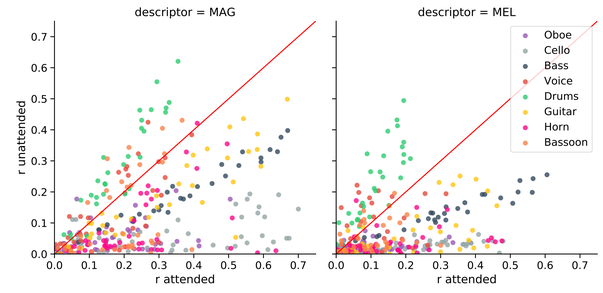
- **limited generalization capability**;
- the model is tracking mostly the **pitch/harmonic contour** of the attended instrument;
- the **more instruments** in the mixture, the **more difficult** is the attention task.

Results

	F1 score (%)								
	all	ensemble		melody/rhythm		rendering		genre	
		duets	trios	same	diff	mono	stereo	pop	classic
AE	51 *	58 *	37 n.s.	48 n.s.	53 *	53 *	48 n.s.	54 *	48 n.s.
MAG	72 **	74 **	66 **	76 **	**65** **	73 **	**72** **	**64** **	79 **
MEL	**73** **	**79** **	**73** **	**79** **	60 **	**74** **	71 **	60 **	**83** **







References

[1] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, p. 233, 2012.

[2] J. A. O'sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional selection in a cocktail party environment can be decoded from single-trial eeg," *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, 2014.