Neuro-steered music source separation with EEG-based auditory attention decoding and Contrastive-NMF

Giorgia Cantisani, Slim Essid, Gaël Richard

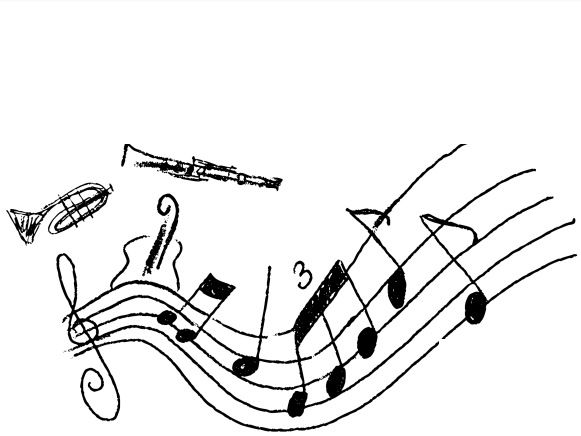
MIPFrontiers

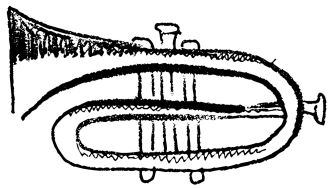
INSTITUT POLYTECHNIQUE DE PARIS

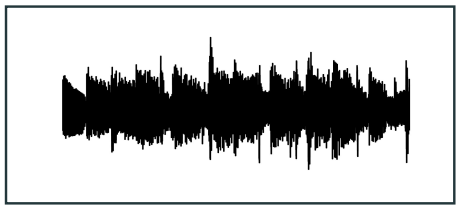
This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No.765068.

Background

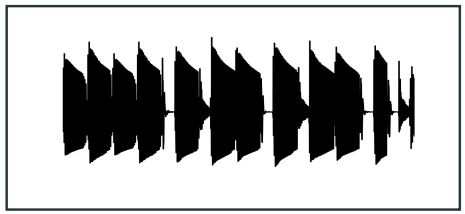
Blind Audio Source Separation







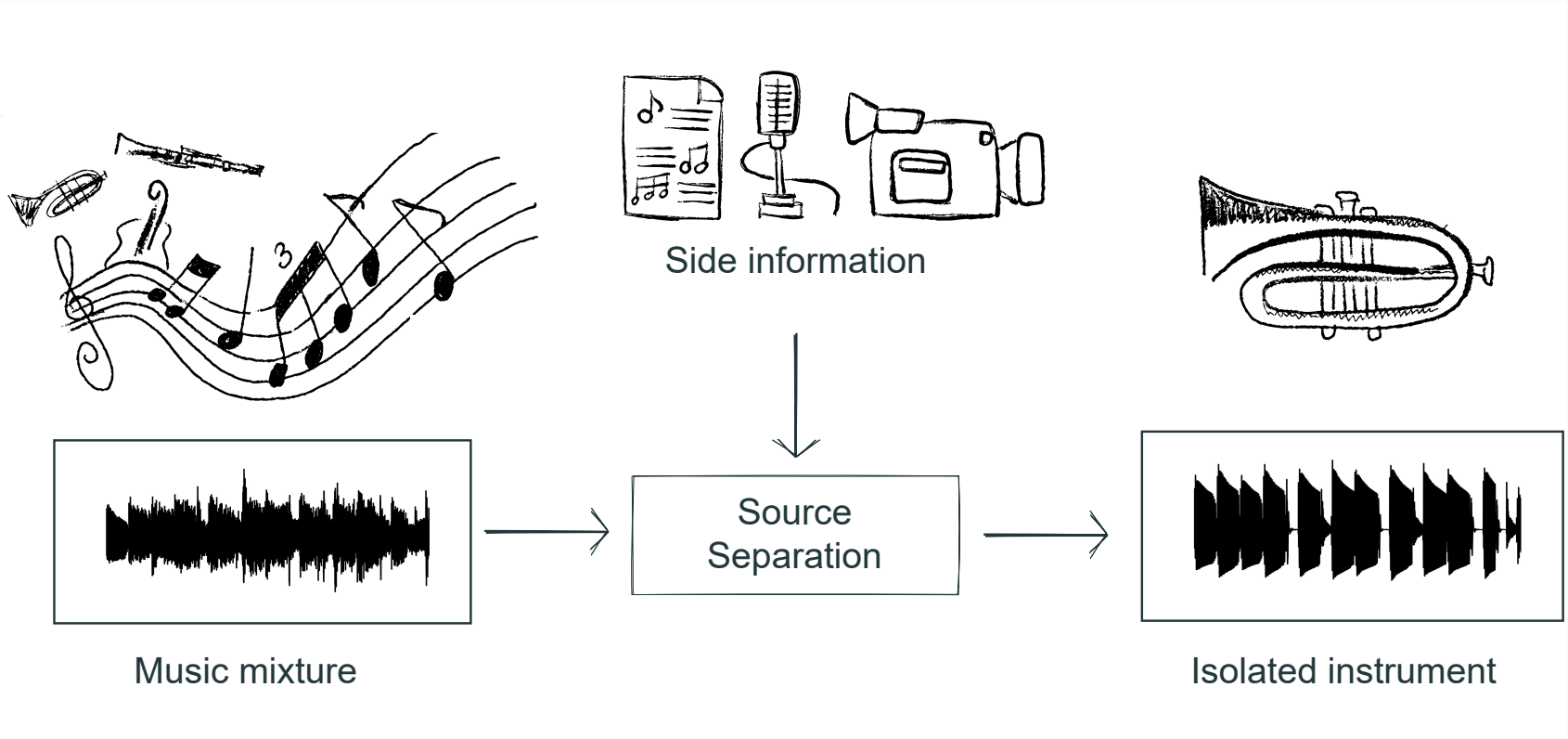
Source Separation



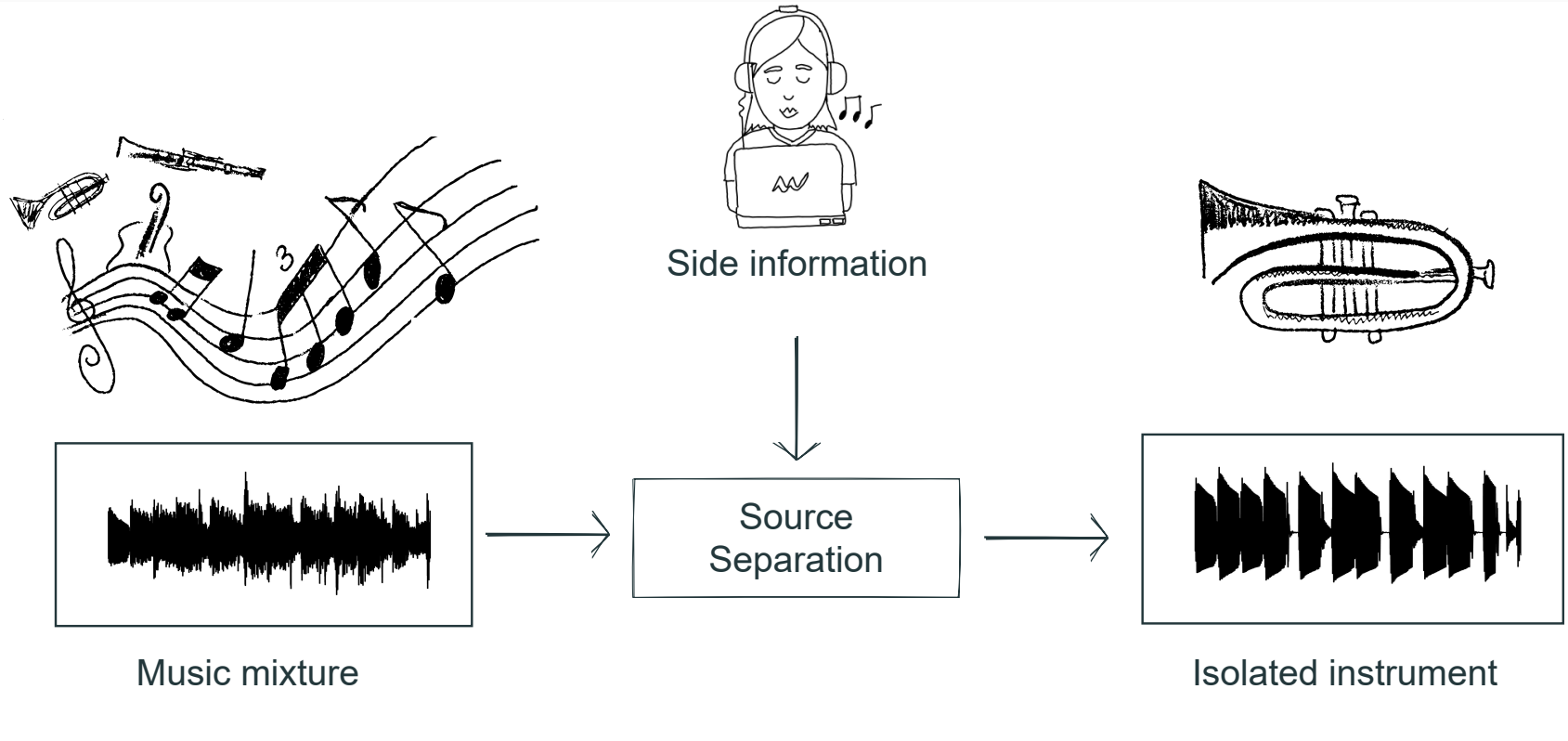
Music mixture

Isolated instrument

Informed Audio Source Separation



User-driven Audio Source Separation



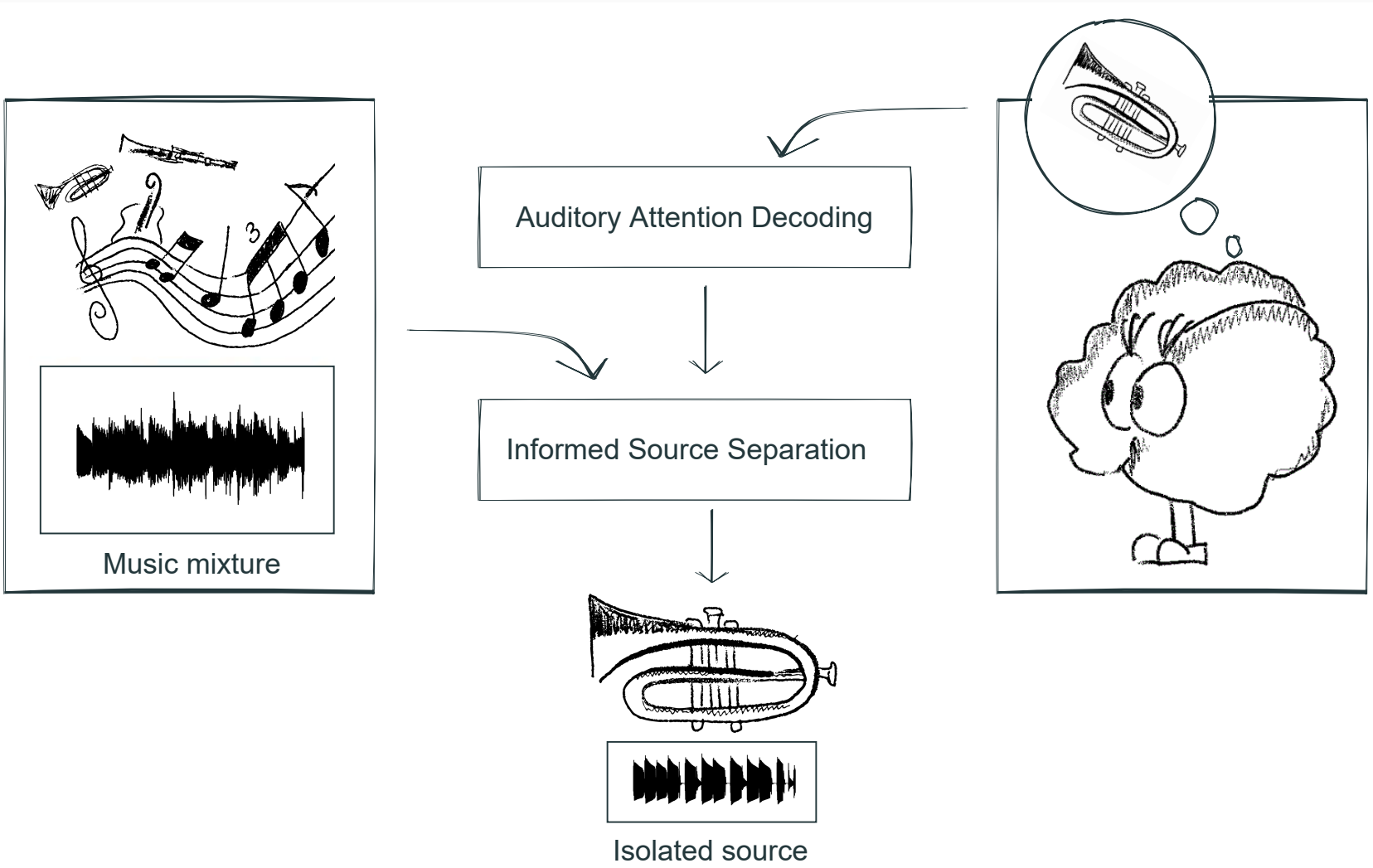
Auditory Attention

Auditory attention *is the cognitive mechanism that allows humans to focus on a sound source of interest in noisy environment.*

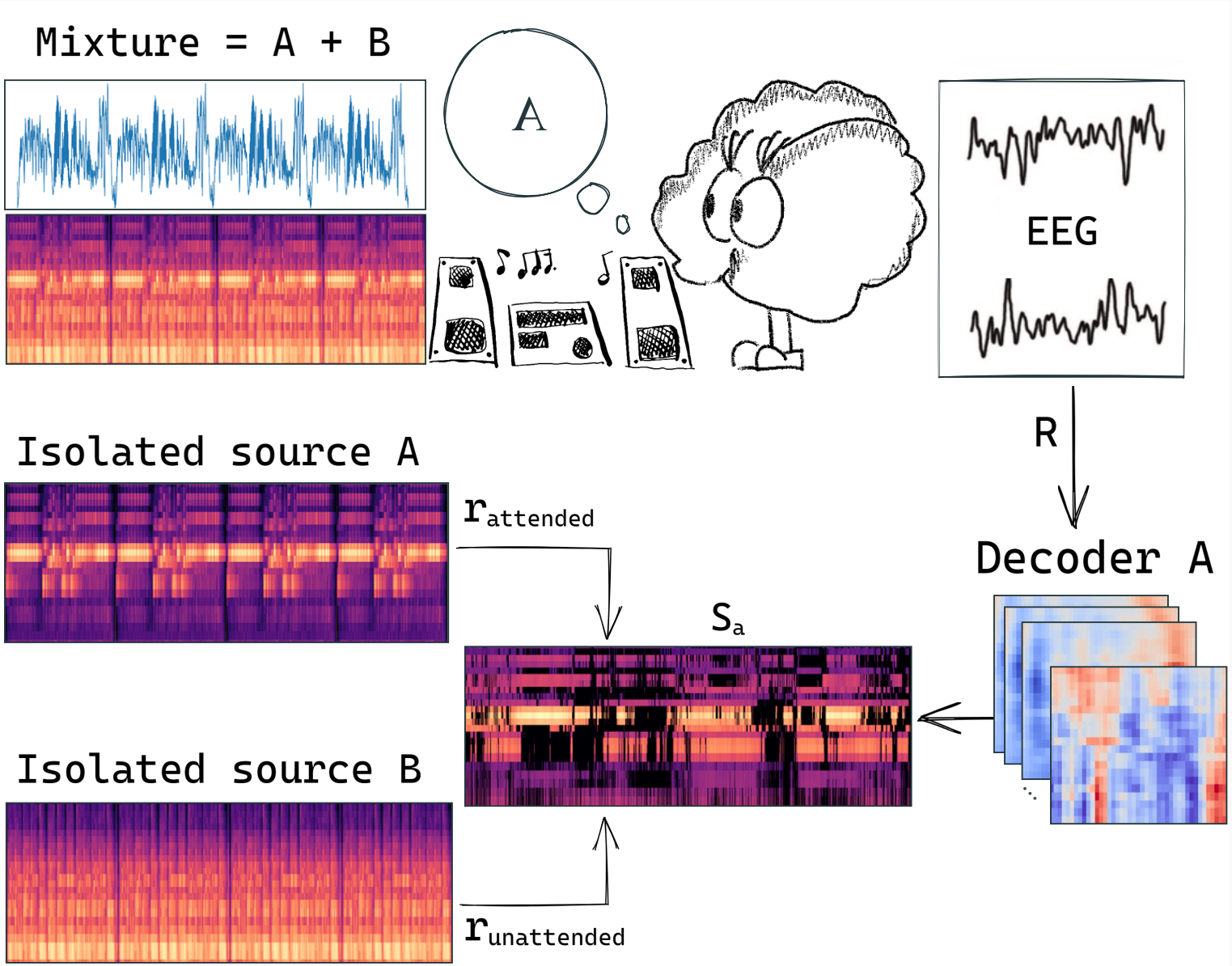
- can be tracked in the **neural activity** (EEG, ECoG, MEG);
- the attended source's neural encoding is **stronger** than the other ones;

[Mesgarani et al., 2009, Mesgarani and Chang, 2012, O'sullivan et al., 2014]

Neuro-steered Audio Source Separation



EEG-based Auditory Attention Decoding



Neuro-steered Speech Enhancement

Typically the separation and the decoding tasks are tackled **sequentially**:

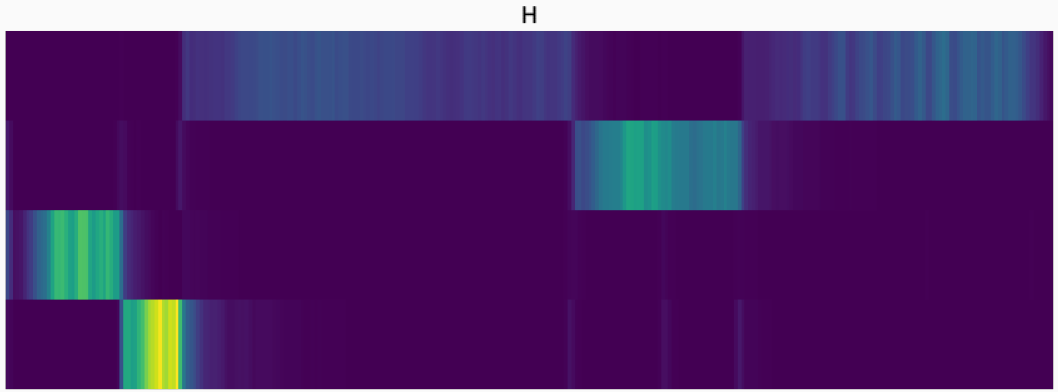
- a separation system provides the reference sources for the decoding;
- and the decoding system selects the source which needs to be enhanced.

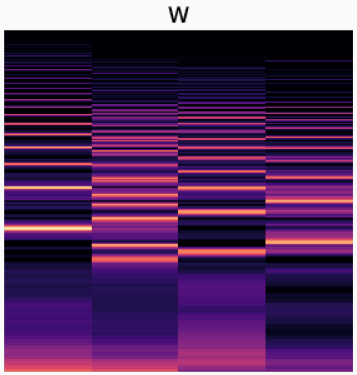
[Aroudi and Doclo, 2020, Van Eyndhoven et al., 2017, Das et al., 2020, O'Sullivan et al., 2017, Han et al., 2019, Ceolini et al., 2020]

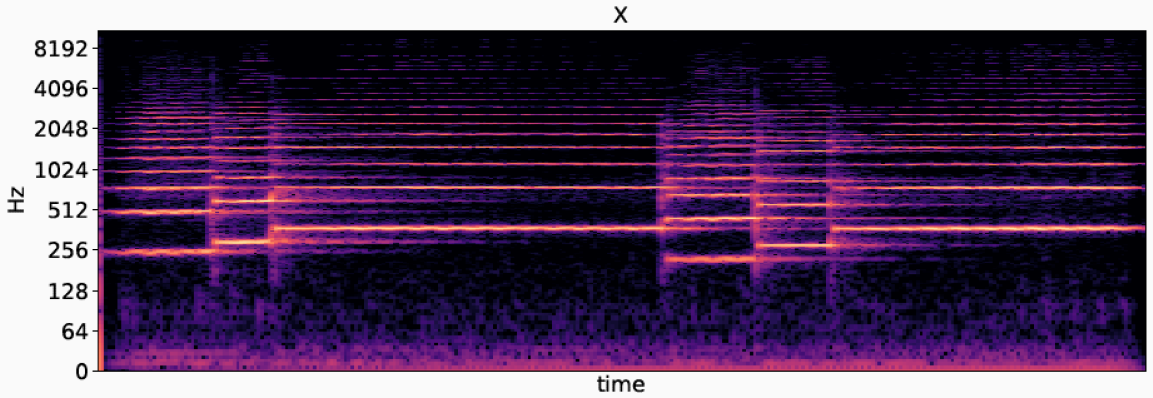
Framework

NMF for Audio Source Separation

$$\begin{cases} C(W,H) = \underbrace{D(X|WH)}_{\text{audio factorization}} + \underbrace{\mu\|H\|_1 + \beta\|W\|_1}_{\text{sparsity}} \\ W,H \geq 0. \end{cases}$$



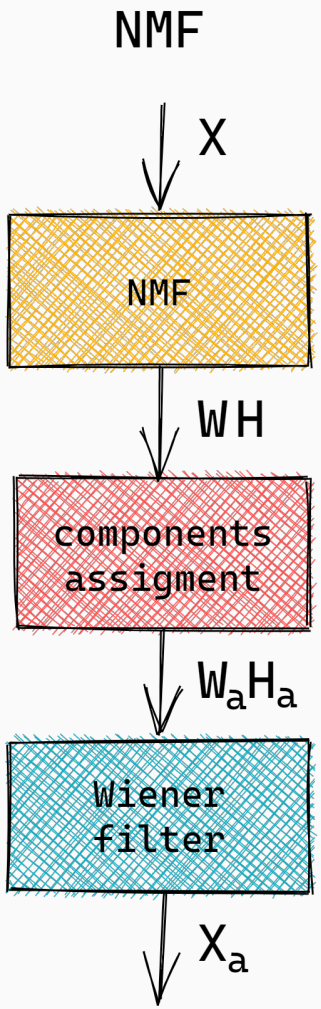




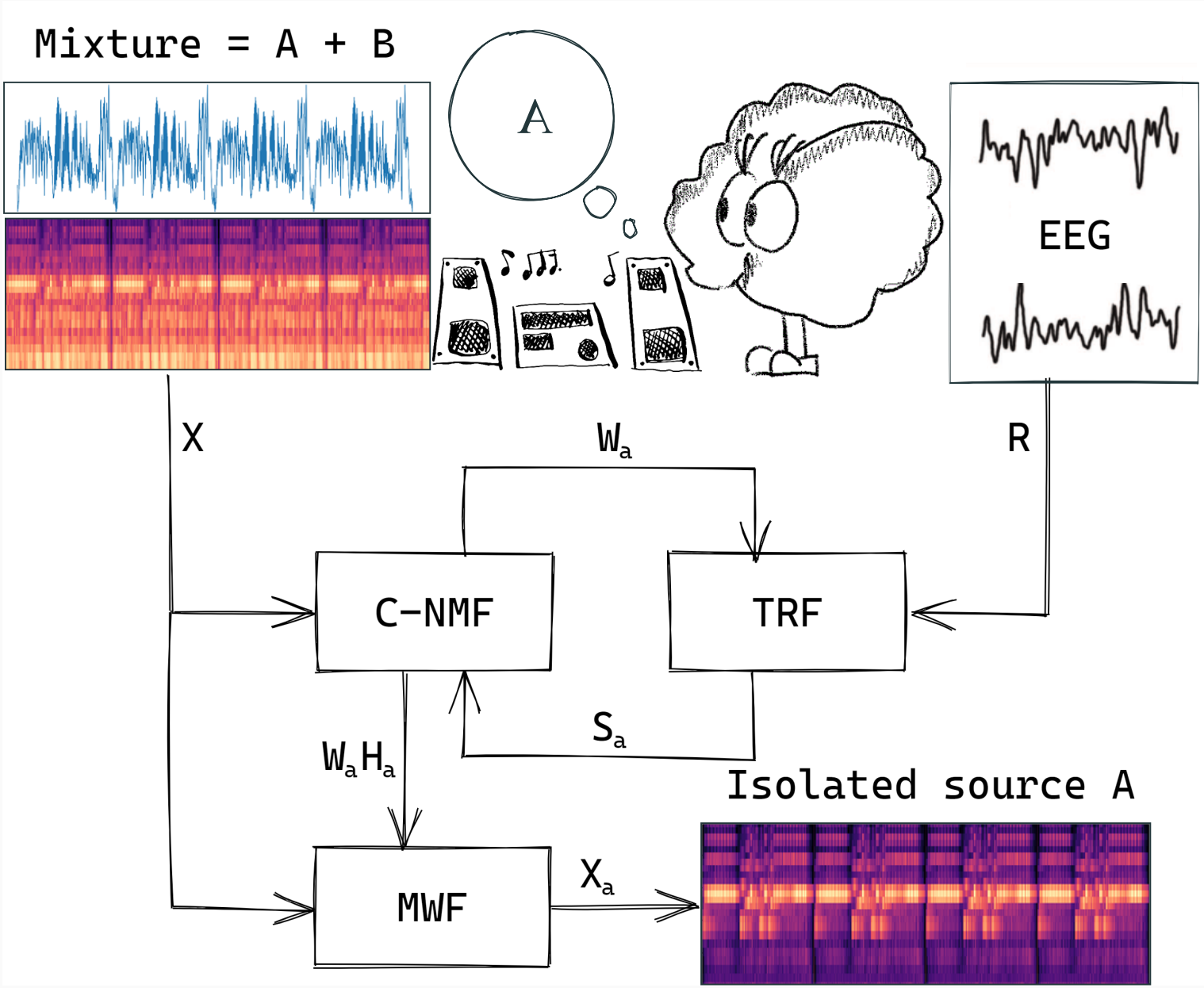
NMF for Audio Source Separation

$$\begin{cases} C(\mathrm{W},\mathrm{H}) = \underbrace{D(\mathrm{X}|\mathrm{WH})}_{\text{audio factorization}} + \underbrace{\mu\|\mathrm{H}\|_1 + \beta\|\mathrm{W}\|_1}_{\text{sparsity}} \\ \mathrm{W},\mathrm{H} \geq 0. \end{cases}$$

$$\mathrm{X}_a = \frac{\mathrm{W}_a\mathrm{H}_a}{\mathrm{WH}} \otimes \tilde{\mathrm{X}}.$$



Proposed model



Contrastive-NMF (C-NMF)

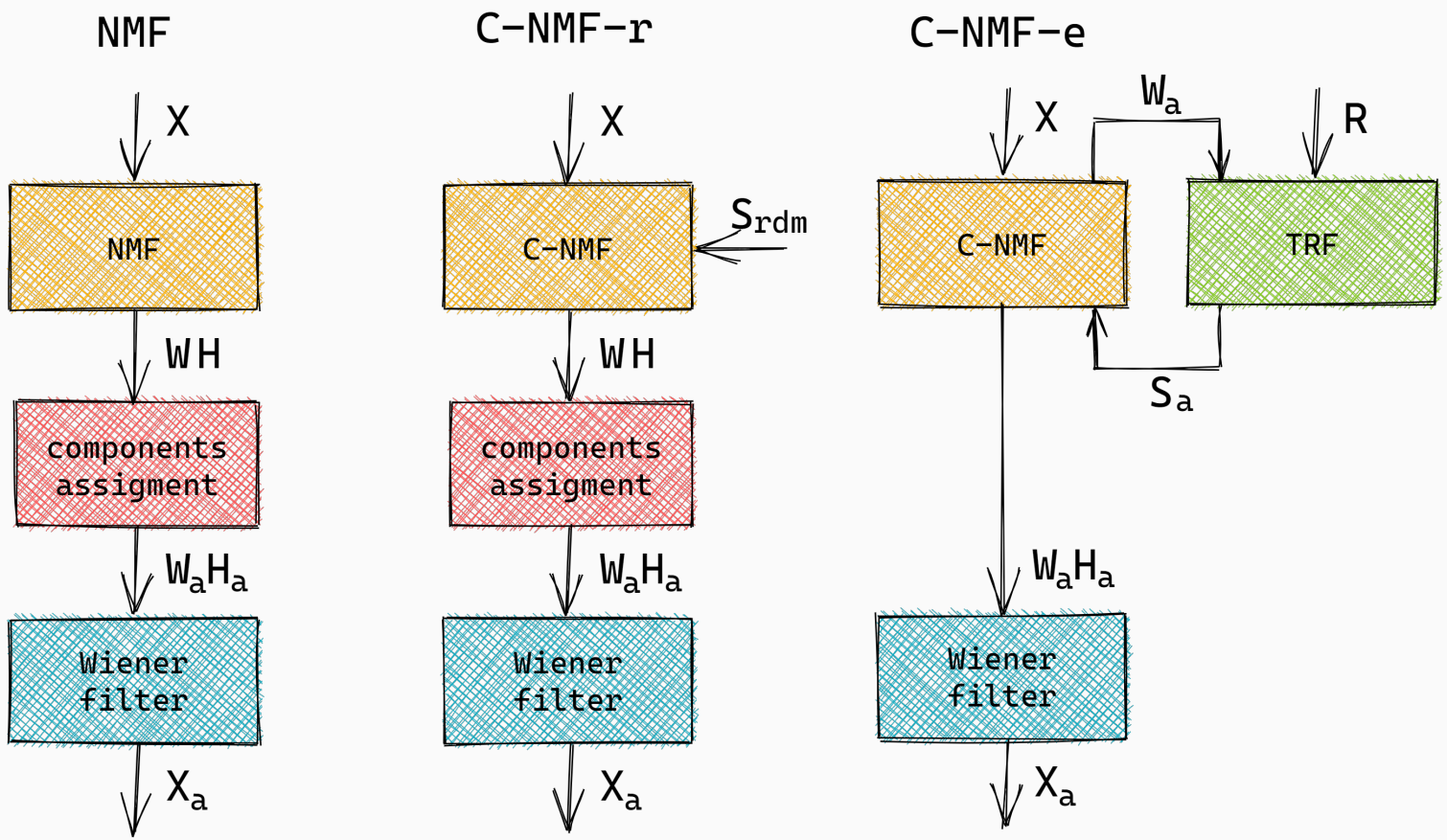
$$\begin{cases} C(\mathrm{W},\mathrm{H}) = \underbrace{D(\mathrm{X}|\mathrm{WH})}_{\text{audio factorization}} + \underbrace{\mu\|\mathrm{H}\|_1 + \beta\|\mathrm{W}\|_1}_{\text{sparsity}} - \underbrace{\delta(\|\mathrm{H}_a\mathrm{S}_a^T\|_F^2 - \|\mathrm{H}_u\mathrm{S}_a^T\|_F^2)}_{\text{contrast}} \\ \mathrm{W},\mathrm{H},\mathrm{S}_a \geq 0 \\ \|\mathbf{h}_{k:}\|_2 = 1, \|\mathbf{s}_{k:}\|_2 = 1. \end{cases}$$

The proposed cost aims at:

- **decomposing** the mixture spectrogram;
- **maximising** the similarity of S_a with H_a;
- **minimising** the similarity of S_a with H_u.

Experiments

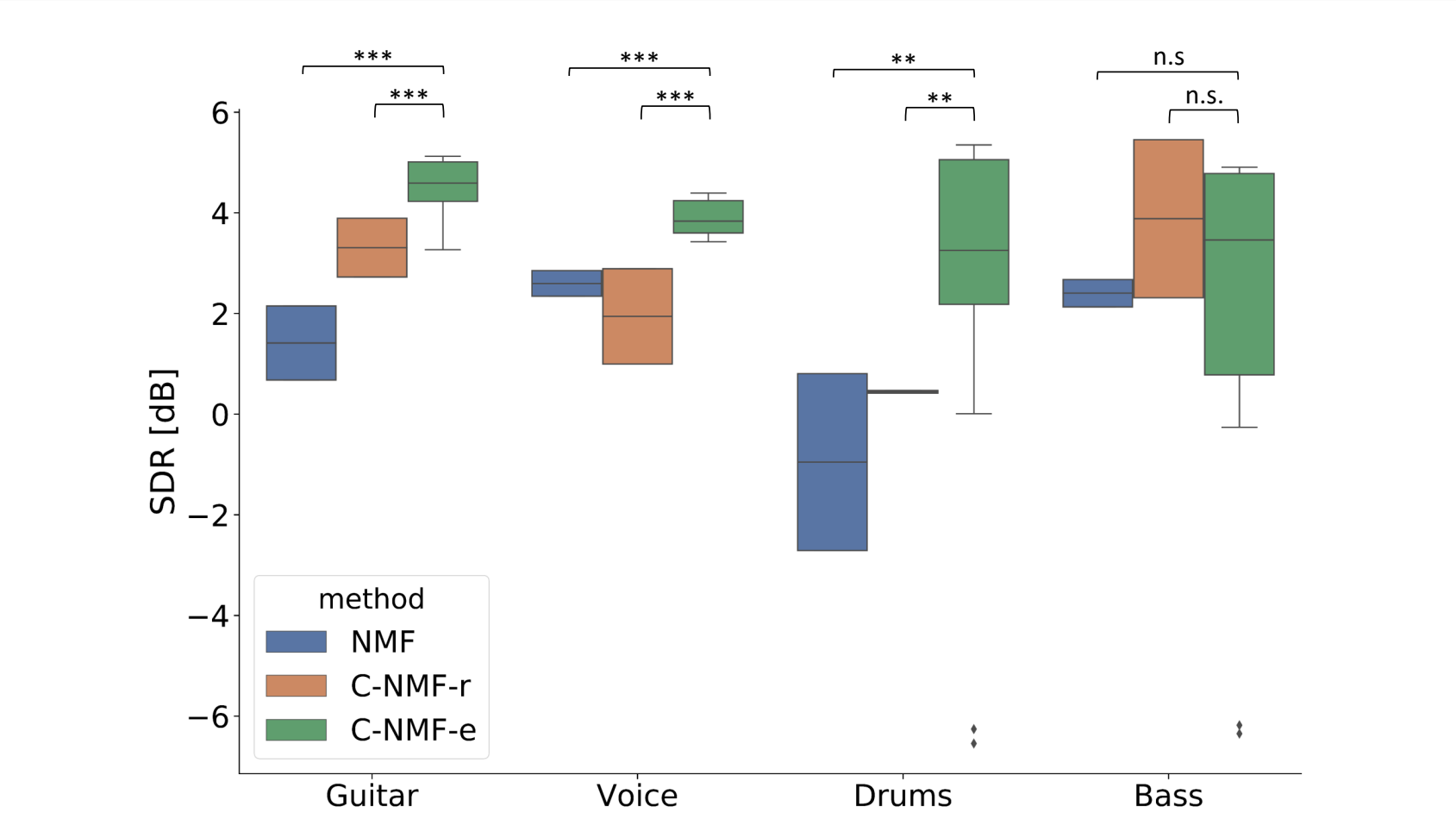
Experiments



Data: 20-channel EEG signals recorded from 8 subjects while they were attending to a particular instrument in polyphonic music [Cantisani et al., 2019].

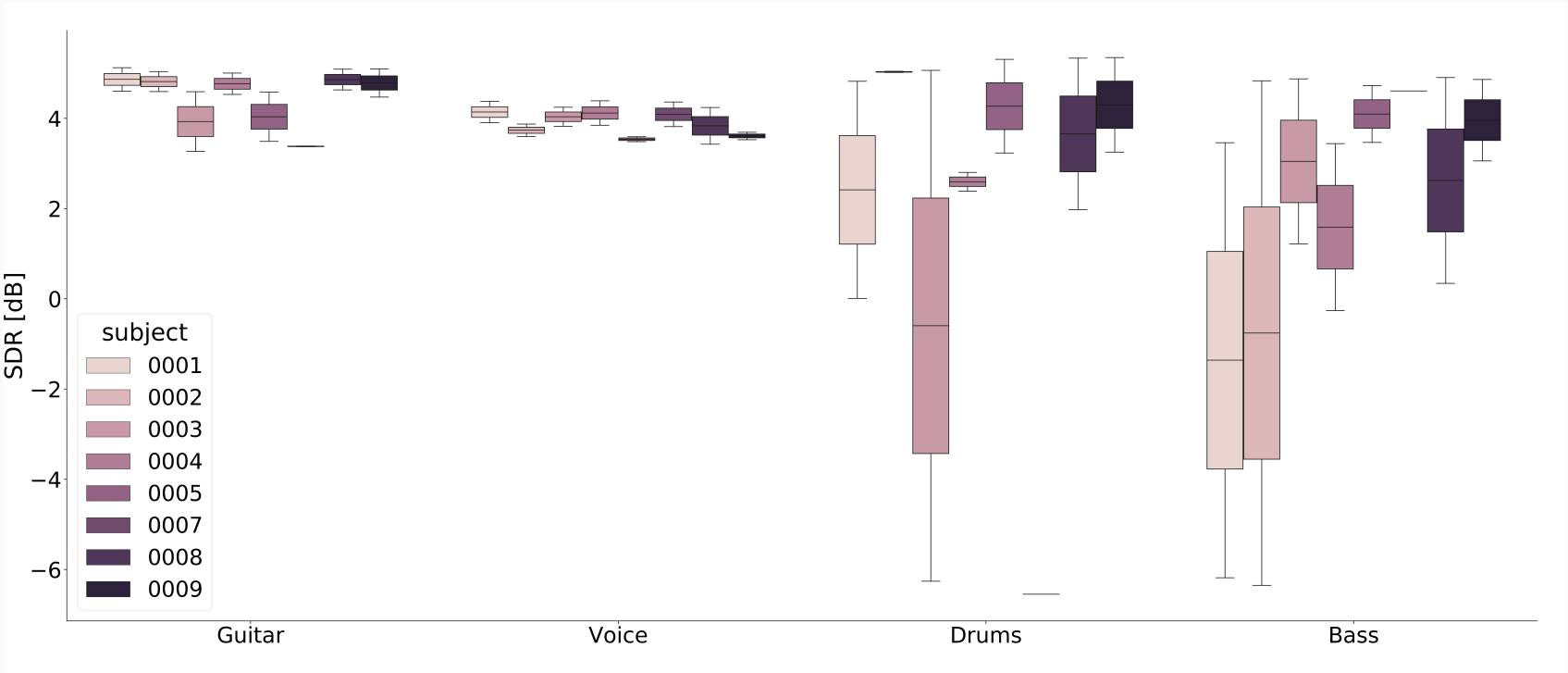
Evaluation: Signal-to-Distortion Ratio (SDR) expressed in dB.

Separation quality



C-NMF performs significantly better than both the baselines except for the bass

Inter-subject variability



- *high inter and intra-subject variability*
- *dependency on the level of attention and musical expertise*

Spatial rendering

SDR [dB]	mono	stereo
Guitar	4.6	**5.0**
Vocals	3.8	**4.3**
Drums	3.2	**5.0**
Bass	**4.0**	0.3

- *the stereo setting helps to localize the target instrument*
- *localizing the instruments helps the attention task*

Conclusions

We propose a new model for **neuro-steered music source separation**:

- **improves the separation**, especially in difficult cases;
- **automatically separates the attended source**;
- auditory attention decoding **without the reference sources**.

Resources

Thank you for the attention!

GitHub Code

Demo

Update rule

References i

Aroudi, A. and Doclo, S. (2020).
Cognitive-driven binaural beamforming using EEG-based auditory attention decoding.
IEEE/ACM Trans. on Audio, Speech and Language Processing (TASLP), 28:862–875.

Cantisani, G., Trégoat, G., Essid, S., and Richard, G. (2019).
MAD-EEG: an EEG dataset for decoding auditory attention to a target instrument in polyphonic music.
In *Proc. Workshop on Speech, Music and Mind (SMM19)*, pages 51–55.

References ii


Ceolini, E., Hjortkjær, J., Wong, D. D., O'Sullivan, J., Raghavan, V. S., Herrero, J., Mehta, A. D., Liu, S.-C., and Mesgarani, N. (2020).
Brain-informed speech separation (BISS) for enhancement of target speaker in multitalker speech perception.
NeuroImage.


Das, N., Zegers, J., Francart, T., Bertrand, A., et al. (2020).
EEG-informed speaker extraction from noisy recordings in neuro-steered hearing aids: linear versus deep learning methods.
BioRxiv.

References iii


Han, C., O'Sullivan, J., Luo, Y., Herrero, J., Mehta, A. D., and Mesgarani, N. (2019).
Speaker-independent auditory attention decoding without access to clean speech sources.
Science advances, 5(5):eaav6134.


Mesgarani, N. and Chang, E. F. (2012).
Selective cortical representation of attended speaker in multi-talker speech perception.
Nature, 485(7397):233.


Mesgarani, N., David, S. V., Fritz, J. B., and Shamma, S. A. (2009).
Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex.
Journal of neurophysiology.

References iv

O'Sullivan, J., Chen, Z., Sheth, S. A., McKhann, G., Mehta, A. D., and Mesgarani, N. (2017).
Neural decoding of attentional selection in multi-speaker environments without access to separated sources.
In *39th Ann. Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC)*.

O'sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A., and Lalor, E. C. (2014).
Attentional selection in a cocktail party environment can be decoded from single-trial EEG.
Cerebral Cortex, 25(7):1697–1706.

References v


Van Eyndhoven, S., Francart, T., and Bertrand, A. (2017).
EEG-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses.
IEEE Trans. Biomed. Engineering, 64(5):1045–1056.