Decoding drums, instrumentals, vocals, and mixed sources in music using human brain activity with fMRI

Vincent K.M. Cheung[1], Lana Okuma[2], Kazuhisa Shibata[2], Kosetsu Tsukuda[3], Masataka Goto[3], Shinichi Furuya[1]

[1] Sony Computer Science Laboratories, Tokyo, [2] RIKEN Center for Brain Science, [3] National Institute of Advanced Industrial Science and Technology (AIST), Japan

Motivation

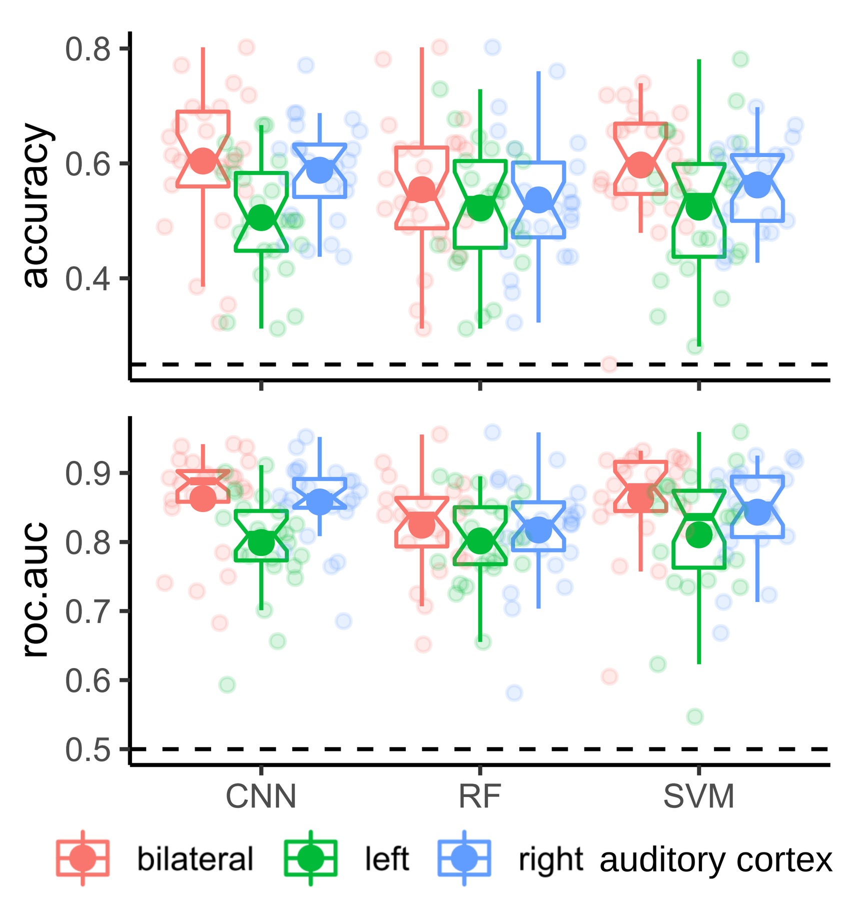
- Decomposing a sound mixture into a linear combination of instrumental sources is a well-established MIR task
- However, current brain decoding models only classify musical instruments from single- or a few notes [1,2], or via attention deployment to a given source [3,4]
- We show that instrument sources in natural music can be decoded from human auditory cortex activity using functional magnetic resonance imaging (fMRI)

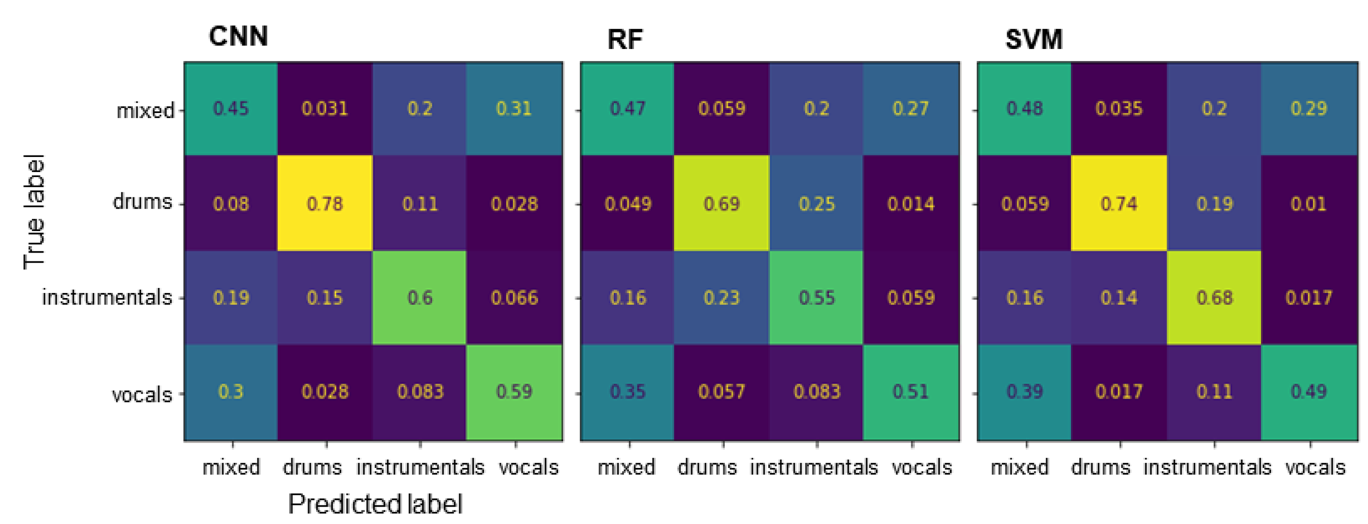
Experiment

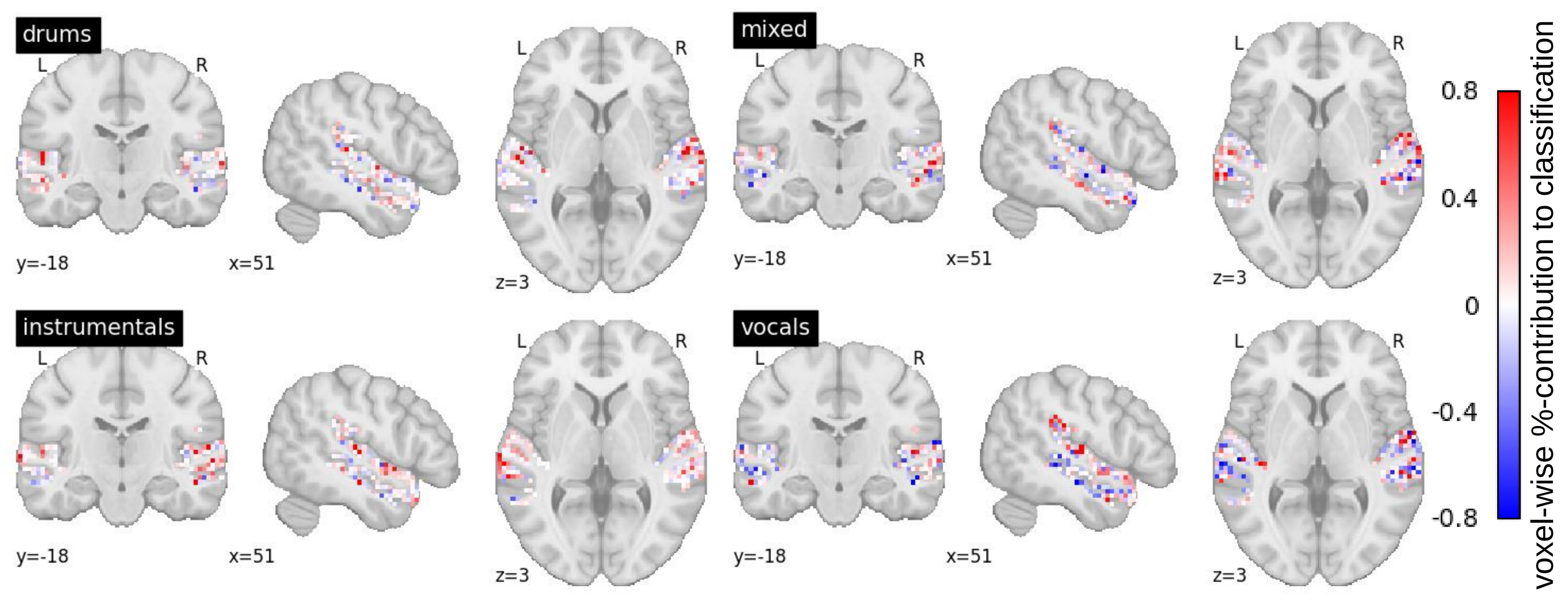
- 96 loudness-normalised stimuli were derived from the first 15s of the chorus in 24 unreleased pop/rock songs separated into four sources using Demucs v4 [5]:
  - Drums
  - Vocals
  - Instrumentals (= bass + others)
  - Mixed (= drums + vocals + instrumentals)
- Brain activity from 24 healthy adults was recorded using 3T MRI scanner during stimulus presentation

Results (using leave-one-subject-out cross-validation)

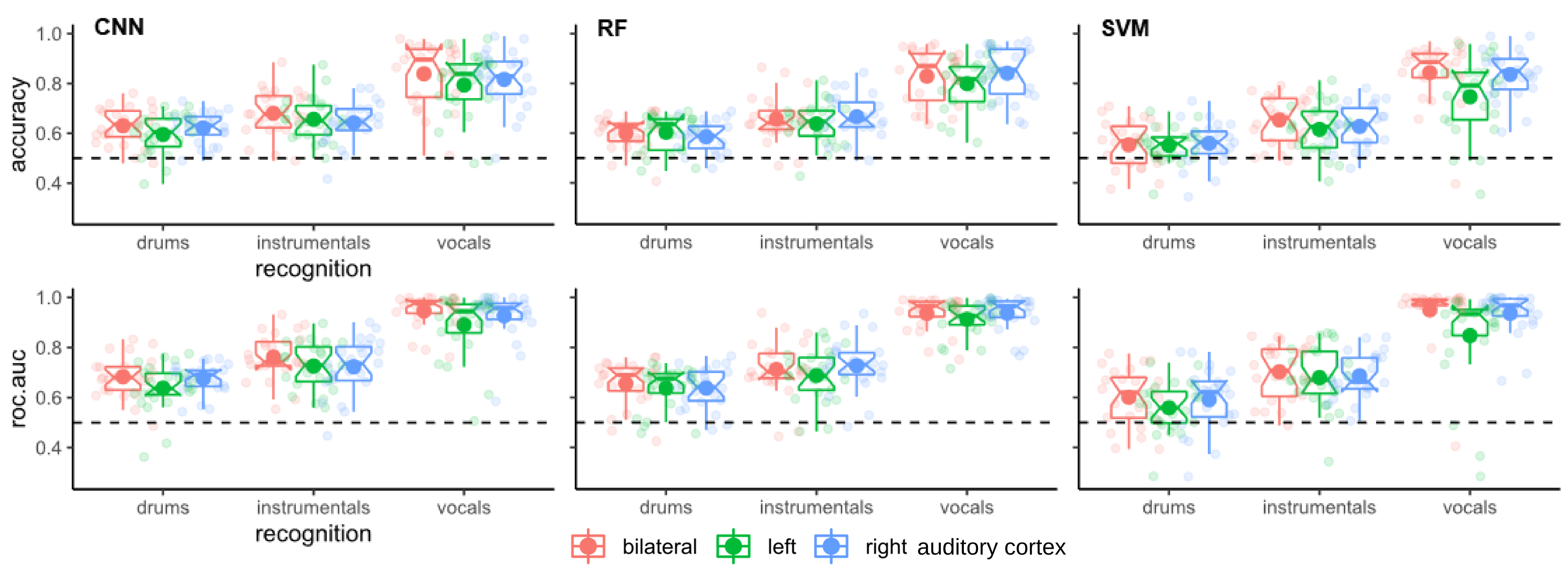
Four-way classification





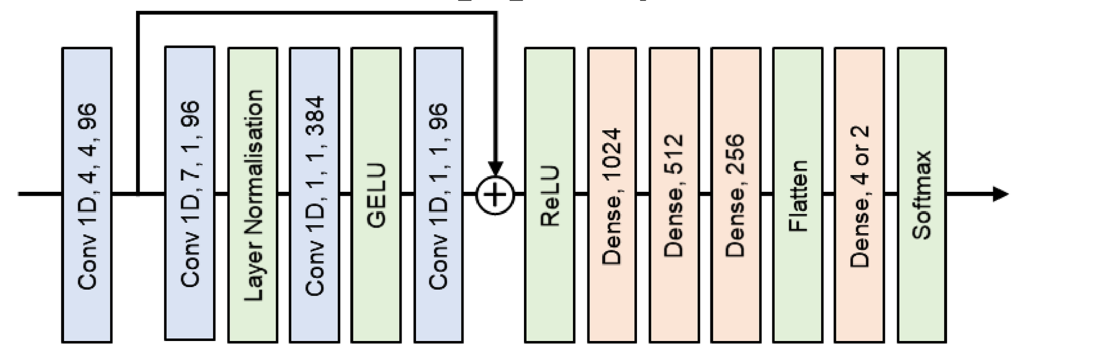


Source recognition



Brain decoders

- ConvNeXt [6]-inspired CNN



- Random forest (RF)
- Support vector machine (SVM)

	CNN		RF		SVM	
	acc	auc	acc	auc	acc	auc
Four-way classification						
l AC	.506	.799	.523	.802	.524	.810
r AC	.588	.858	.536	.817	.563	.843
l+r AC	**.604**	.863	.554	.824	.597	**.863**
l+r PV	.301	.560	.319	.554	.253	.510
l+r SM	.304	.547	.332	.550	.276	.554
Drums recognition						
l AC	.595	.638	.603	.637	.550	.559
r AC	.622	.677	.586	.638	.559	.591
l+r AC	**.630**	**.683**	.599	.655	.553	.601
l+r PV	.507	.505	.528	.533	.526	.531
l+r SM	.517	.544	.530	.545	.490	.500
Instrumentals recognition						
l AC	.656	.726	.638	.688	.615	.679
r AC	.642	.723	.666	.727	.627	.687
l+r AC	**.680**	**.762**	.657	.712	.652	.703
l+r PV	.577	.593	.585	.611	.495	.509
l+r SM	.558	.576	.580	.600	.517	.553
Vocals recognition						
l AC	.794	.891	.799	.913	.746	.847
r AC	.816	.926	.841	.937	.836	.936
l+r AC	.839	.946	.829	.936	**.843**	**.950**
l+r PV	.527	.527	.525	.541	.495	.502
l+r SM	.516	.544	.563	.581	.516	.552

acc = accuracy, auc = ROC AUC; l/r/l+r = left/right/bilateral; AC = auditory, PV = primary visual, SM = somatosensory-motor cortices

Conclusions

- Spatial representations in the human auditory cortex activity provide useful information across classifiers towards decoding different instrument sources
- High performance in recognising vocals suggests enhanced perceptual sensitivity towards vocal information during music listening
- Future work could exploit neural representations as an alternative to subjective tests such as MUSHRA or MOS

References

[1] Paquette, et al., “Cross-classification of musical and vocal emotions in the auditory cortex,” *Annals of the New York Academy of Sciences*, vol. 1423, no. 1, pp. 329–337, 2018.

[2] Ogg, et al., “Separable neural representations of sound sources: Speaker identity and musical timbre,” *NeuroImage*, vol. 191, pp. 116–126, 2019.

[3] Cantisani, et al., “MAD-EEG: an EEG dataset for decoding auditory attention to a target instrument in polyphonic music,” in Proceedings of the Speech, Music and Mind (SMM), Satellite Workshop of Interspeech 2019, 2019.

[4] Cantisani, et al., “Neuro-steered music source separation with eeg-based auditory attention decoding and contrastive-nmf,” in *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, ser. ICASSP 2021, 2021.

[5] Rouard, et al., “Hybrid transformers for music source separation,” in *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, ser. ICASSP 2023, 2023.

[6] Liu, et al., “A convnet for the 2020s,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.