



HAL
open science

Neuro-steered music source separation

Giorgia Cantisani

► **To cite this version:**

Giorgia Cantisani. Neuro-steered music source separation. Signal and Image Processing. Institut Polytechnique de Paris, 2021. English. NNT : 2021IPPAT038 . tel-03511225

HAL Id: tel-03511225

<https://theses.hal.science/tel-03511225v1>

Submitted on 4 Jan 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



INSTITUT
POLYTECHNIQUE
DE PARIS

NNT : 2021IPPAT038

Thèse de doctorat



Neuro-steered Music Source Separation

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom Paris

École doctorale n°626 École doctorale de l'Institut Polytechnique de Paris (EDIPP)
Spécialité de doctorat : Signal, Images, Automatique et robotique

Thèse présentée et soutenue à Palaiseau, le 13 décembre 2021, par

GIORGIA CANTISANI

Composition du Jury :

Isabelle Bloch Professor, Télécom Paris, France	Présidente
Alexandre Gramfort Senior Research Scientist, Inria, France	Rapporteur
Shihab A. Shamma Professor, University of Maryland & ENS, France	Rapporteur
Elaine Chew Senior CNRS Researcher, IRCAM, France	Examinatrice
Blair Kaneshiro Adjunct Professor, Stanford University, California	Examinatrice
Slim Essid Professor, Télécom Paris, France	Co-directeur de thèse
Gaël Richard Professor, Télécom Paris, France	Directeur de thèse
Alexey Ozerov Senior Research Scientist, Ava, France	Invité

IMPRINT

Neuro-steered Music Source Separation

Copyright © 2020 by Giorgia CANTISANI.

All rights reserved. Compiled at home, printed in France.

COLOPHON

This thesis was typeset using \LaTeX and the memoir documentclass. It is based on Aaron Turon's thesis *Understanding and expressing scalable concurrency*¹ (as re-implemented by Friedrich Wiemer² and Diego Di Carlo³), itself a mixture of classicthesis⁴ by André Miede and tufte-latex⁵, based on Edward Tufte's *Beautiful Evidence*.

Graphics and plots are made with Matplotlib⁶, Seaborn⁷, and statannot.⁸ Drawings and schemes are made with Excalidraw⁹ and draw.io.¹⁰ Icons are downloaded from the www.flaticon.com/ The bibliography was processed by Biblatex.

The body text is set to 10/14pt (long primer) on a 26pc measure. The margin text is set to 8/9pt (brevier) on a 12pc measure. Linux Libertine acts as both the text and display typeface.

¹<https://people.mpi-sws.org/~turon/turon-thesis.pdf>

²<https://github.com/pfasante/phd-thesis>

³<https://github.com/Chutlhu/PhD-manuscript>

⁴<https://bitbucket.org/amiede/classicthesis/>

⁵<https://github.com/Tufte-LaTeX/tufte-latex>

⁶<https://matplotlib.org/>

⁷<https://seaborn.pydata.org>

⁸<https://github.com/webermarcolivier/statannot>

⁹<https://github.com/excalidraw>

¹⁰<https://www.diagrams.net/>

Abstract

Music source separation is the task of isolating individual instruments that are mixed in a musical piece. This task is particularly challenging, as state-of-the-art models can hardly generalise to unseen test data. Nevertheless, additional information about individual sources can be used to better adapt a generic model to the observed mixture signal. Examples of such information are the music score, the lyrics, visual cues, or the user’s feedback. Beyond metadata and manual annotations, our body’s reaction to auditory stimuli manifests itself through many observable physiological phenomena (e.g. heartbeat variability, body movements, brain activity). Among those, we focused on the neural response and the concept of selective auditory attention, which allows humans to process concurrent sounds and isolate the ones of interest. The attended source’s neural encoding appears to be substantially stronger than the others, allowing to decode which sound source a person is “focusing on”. This task is known as auditory attention decoding (AAD) and has been studied mostly for speech perception in noisy or multi-speaker settings.

In this thesis, we explored how the neural activity reflects information about the attended instrument and how we can use it to inform a source separation system and adapt it to the corresponding stimulus. We were particularly interested in electroencephalographic signals (EEG), which allow for non-invasive neural activity acquisition with high temporal resolution. First, we studied the problem of EEG-based AAD of a target instrument in polyphonic music, showing that the EEG tracks musically relevant features which are highly correlated with the time-frequency representation of the attended source and only weakly correlated with the unattended one. Second, we leveraged this “contrast” to inform an unsupervised source separation model based on a novel non-negative matrix factorisation (NMF) variant, named contrastive-NMF (C-NMF) and automatically separate the attended source. We conducted an extensive evaluation of the proposed system on the MAD-EEG dataset which was specifically assembled for this study, obtaining encouraging results, especially in difficult cases where non-informed models struggle.

Unsupervised NMF represents a powerful approach in such applications with no or limited training data as when neural recording is involved. Indeed, the available music-related EEG datasets are still costly and time-consuming to acquire, precluding the possibility of tackling the problem with fully supervised deep learning approaches. In the last part of the thesis, we explored alternative learning strategies to alleviate this problem. Specifically, we investigated if it is possible to inform a source separation model based on deep learning using the time activations of the sources manually provided by the user or derived from his/her EEG response available at test time. This approach can be referred to as one-shot adaptation, as it acts on the target song instance only. Even if immature, the results are encouraging and point at promising research directions.

Keywords:

Music source separation, Auditory attention decoding, Electroencephalography, Multi-modal processing; Matrix factorisation, Deep learning, One-shot Domain Adaptation.

Résumé en français

La séparation de sources musicales vise à isoler les instruments individuels qui sont mélangés dans un enregistrement de musique. Cette tâche est particulièrement complexe, car même les modèles les plus performants restent peu efficaces sur des données nouvelles ou très différentes de des données utilisées pour l'apprentissage. Néanmoins, des informations supplémentaires sur les sources individuelles peuvent être utilisées pour mieux adapter un modèle de séparation de sources générique au signal observé. Des exemples de telles informations sont : la partition de la musique, les paroles des chansons, les vidéos de performance musicale ou le feedback de l'utilisateur. Au-delà de ces métadonnées et annotations manuelles, la réaction de notre corps aux stimuli auditifs se manifeste par de nombreux phénomènes physiologiques observables (par exemple, la variabilité du rythme cardiaque, les mouvements du corps, l'activité neuronale). Parmi ceux-ci, nous nous sommes concentrés sur la réponse neuronale et le concept d'attention auditive sélective, qui permet aux humains de traiter des sons simultanés et d'isoler ceux qui les intéressent. Le codage neuronal de la source à laquelle on porte son attention semble être sensiblement plus fort que celui des autres sources, ce qui permet de décoder la source sonore sur laquelle une personne se " concentre ". Cette tâche est connue sous le nom de décodage de l'attention auditive (AAD) et a été étudiée principalement pour ce qui concerne la perception des sources vocales dans des environnements bruyants ou à plusieurs voix.

Dans cette thèse, nous avons investigué comment l'activité neuronale reflète des informations sur l'instrument de musique auquel l'auditeur porte son attention et comment nous pouvons l'utiliser pour informer un système de séparation de sources et l'adapter au stimulus correspondant. Nous nous sommes concentrés sur les signaux électroencéphalographiques (EEG), qui permettent une acquisition non invasive de l'activité neuronale avec une haute résolution temporelle. Tout d'abord, nous avons étudié le problème du décodage par l'EEG de l'attention auditive d'un instrument spécifique dans une pièce musicale polyphonique, en montrant que l'EEG suit les caractéristiques musicales pertinentes qui sont fortement corrélées avec la représentation temps-fréquence de la source à laquelle on porte son attention et seulement faiblement corrélées avec les autres. Ensuite, nous avons exploité ce "contraste" pour informer un modèle de séparation de sources non supervisé basé sur une nouvelle variante de factorisation en matrices positives (*NMF* : *non-negative matrix factorization*), appelée *contrastive-NMF* (*C-NMF*) et séparer automatiquement la source à laquelle on porte son attention. Nous avons effectué une évaluation approfondie du système proposé sur le jeu de données MAD-EEG qui a été spécifiquement collecté pour cette étude. Nous avons analysé l'impact de multiples aspects des stimuli musicaux, tels que le nombre et le type d'instruments dans le mélange, le rendu spatial et le genre musical, obtenant des résultats encourageants, en particulier dans les cas difficiles où les modèles non informés sont défailants.

Mots-clés :

Séparation des sources musicales, Décodage de l'attention auditive, Electroencéphalographie, Traitement multimodal, Factorisation matricielle, Apprentissage profond.

La NMF non supervisée représente une approche efficace dans de telles applications ne disposant pas ou peu de données d'apprentissage, comme c'est le cas dans des scénarios nécessitant des enregistrements EEG. En effet, les jeux de données EEG liés à la musique disponibles sont encore coûteux et longs à acquérir, ce qui exclut la possibilité d'aborder le problème par des approches d'apprentissage profond entièrement supervisées. Ainsi, dans la dernière partie de la thèse, nous avons exploré des stratégies d'apprentissage alternatives. Plus précisément, nous avons étudié la possibilité d'informer un modèle de séparation de sources basé sur l'apprentissage profond en utilisant les activations temporelles de sources fournies manuellement par l'utilisateur ou dérivées de sa réponse EEG disponible au moment du test. Cette approche peut être considérée comme étant "à adaptation unitaire" (*one-shot*), car l'adaptation agit uniquement sur une instance de chanson. Bien que préliminaires, les résultats obtenus sont encourageants et indiquent des directions de recherche prometteuses.

Acknowledgements

I would like to thank all the people that believed in me and were co-protagonists of this adventure. Your technical and emotional support was fundamental, and I would like to take this occasion to express my gratitude to:

- Slim and Gaël for advising me and being a source of inspiration for the researcher I want to be. Despite my sabotage attempts and impostor syndrome, you supported me with patience and perseverance.
- Alexey, for welcoming me to your lab in Rennes during my internship and sharing your expertise and valuable feedback with me.
- Blair, for mentoring me throughout all these years. You have been a source of inspiration but, most of all, a true friend.
- Giovanni, for the enthusiastic and stimulating discussions that hopefully will continue after this thesis work.
- Isabelle, Alexandre, Shihab, Alain, Elaine, Blair. It was an honour and challenge to have you on my thesis committee, and I would like to thank you for the time you dedicated to reading these pages.
- Colleagues from Télécom Paris and the ADASP group. You made me feel welcome, and I consider you more as friends than colleagues.
- Ondřej, Kilian, Karim and Javier for the memorable time travelling around the world and living this beautiful experience together.
- Marie Skłodowska-Curie Actions and the MIP-Frontiers network for making my PhD possible.¹ Special thanks to Alvaro, who supported us students technically and emotionally as only a sensitive person can.
- Carmen, Laurence, Delphine, Isabelle and Françoise for supporting me with the scary French language and administrative duties with kindness.
- My friends in France, Italy and around the world. Living far away from home has been hard, but thanks to your friendship, kindness, and help, I never felt lonely.
- Cecilia, Camille and Basile, for always making me feel at home and part of their wonderful family.
- My family & *animals* for their love and support, even if I risked to lose my manuscript because of Mina walking on my Glenny.
- My love Diego, who takes care of me with great courage and lightheartedness even in extreme situations such as the thesis writing.
- Giovanni De Poli, for making me and Diego curious about music and computer science and secretly being the start of all this story. It's been almost 10 years, and you still follow us with great affection.

¹This work was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 765 068 (MIP-Frontiers).

Last but not least, I would like to give credit to all the open-source projects for scientific computing that made possible or, at least, greatly facilitated the work behind this thesis. Python libraries for scientific computing, data analysis and visualisations that I would like to mention are: NumPy [Harris et al. 2020], SciPy [Virtanen et al. 2020], Scikit-learn [Pedregosa et al. 2011], PyTorch [Paszke et al. 2019], PyTorch-Lightning [Falcon et al. 2019], Matplotlib [Hunter 2007], Seaborn [Waskom 2021], Pandas [McKinney 2010] and statsannot [Weber 2020]. A special mention for open-source projects specific for audio and EEG processing which were fundamental for carrying out this thesis: Librosa [McFee et al. 2015], museval [Stöter et al. 2018], Asteroid [Pariante et al. 2020], MNE-Python [Gramfort et al. 2013], Braindecode [Schirrneister et al. 2017] and pymtrf [Steinkamp 2020].

Contents

ABSTRACT	v
RÉSUMÉ EN FRANÇAIS	vii
ACKNOWLEDGEMENTS	xi
CONTENTS	xvi
GLOSSARY	xix
NOTATIONS	xxiii
I INTRODUCTION	1
1 INTRODUCTION	3
1.1 Motivation and objective	3
1.2 Background	5
1.3 Contributions and thesis outline	11
1.4 List of publications	14
1.5 Vademecum	15
II DECODING OF AUDITORY ATTENTION TO MUSIC	17
2 mad-eeeg: AN EEG DATASET FOR DECODING AUDITORY ATTENTION TO A TARGET INSTRUMENT IN POLYPHONIC MUSIC	19
2.1 Introduction	19
2.2 Related works	20
2.3 Dataset creation	20
2.4 Conclusions	27
3 maad: EEG-BASED DECODING OF AUDITORY ATTENTION TO A TARGET INSTRUMENT IN POLYPHONIC MUSIC	29
3.1 Introduction	29
3.2 Related works	30
3.3 Methods	31
3.4 Experiments	34
3.5 Conclusions	39
III NEURO-STEERED MUSIC SOURCE SEPARATION	41
4 c-nmf: NEURO-STEERED MUSIC SOURCE SEPARATION WITH EEG-BASED AUDITORY ATTENTION DECODING AND CONTRASTIVE-NMF	43
4.1 Introduction	43
4.2 Related works	44
4.3 Methods	46
4.4 Experiments	52
4.5 Conclusions	58

5	ugosa: USER-GUIDED ONE-SHOT DEEP MODEL ADAPTATION FOR MUSIC SOURCE SEPARATION	61
5.1	Introduction	61
5.2	Related works	63
5.3	Methods	64
5.4	Experiments	66
5.5	Conclusions	77
	IV EPILOGUE	79
6	CONCLUSIONS	81
6.1	Summary of contributions	81
6.2	Future perspectives	82
	V APPENDICES	89
	STATISTICAL TESTING	91
	Looking at differences	91
	Comparing distributions of scores	91
	Comparing classification performance to chance level	92
	DETAILED DERIVATION OF THE MULTIPLICATIVE UPDATE RULES FOR THE CONTRASTIVE-NMF	95
	Update rule for \mathbf{W}	96
	Update rule for \mathbf{H}	97
	SCIENCE DISSEMINATION: THE MIP-FRONTIERS VIDEO COMMUNICATION PROJECT	101
	Science dissemination	101
	Making-of	103
	Special Thanks	104
	BIBLIOGRAPHY	107

Glossary

AAD	Auditory Attention Decoding	9
AE	Amplitude Envelope	31
AuSS	Audio Source Separation	5
BCI	Brain-Computer Interface	3
C-NMF	Contrastive-NMF	43
DL	Deep Learning	5
DNN	Deep Neural Network	10
ECoG	Electrocorticography	7
ECG	Electrocardiography	26
EEG	Electroencephalography	7
EMG	Electromyography	26
EOG	Electrooculography	26
ERP	Event-related Potential	8
ESU	External Sync Unit	26
fMRI	functional Magnetic Resonance Imaging	7
HCI	Human-Computer Interface	3
ICA	Independent Component Analysis	27
iSTFT	Inverse Short Time Fourier Transform	48
LFP	Local Field Potentials	7
MAG	Magnitude Spectrogram	31
MEG	Magnetoencephalography	7
MEL	Mel Spectrogram	31
MFCC	Mel-frequency cepstral coefficient	52
MIP	Music Information Processing	102
MIR	Music Information Research	3
MMSE	Minimum Mean Squared Error	30
MSS	Music Source Separation	4
MU	Multiplicative Update	49
MWF	Multichannel Wiener Filter	30
NMF	Nonnegative Matrix Factorization	6
NTF	Nonnegative Tensor Factorization	6
PCA	Principal Component Analysis	46
PCC	Pearson correlation coefficient	31
SAR	Signal to Artifacts Ratio	66
SDR	Signal to Distortion Ratio	53

SDRi	Signal to Distortion Ratio Improvement	53
SIR	Signal-to-Interference Ratio	66
SNR	Signal-to-Noise-Ratio	7
SSL	Self Supervised Learning	84
STFT	Short Time Fourier Transform	46
TF	Time-Frequency	6
TL	Transfer Learning	83
WF	Wiener Filter	48

Notations

GENERAL

x	scalars
\mathbf{x}	vector
\mathbf{X}	matrix
x_i	i -th entry of \mathbf{x}
$\hat{\mathbf{x}}$	estimated value of \mathbf{x}
\mathbf{X}^T	Transpose of matrix \mathbf{X}
$\ \mathbf{x}\ _p$	ℓ_p norm of a vector \mathbf{x} .
$\ \mathbf{x}\ _F^2$	squared Frobenius norm of a vector \mathbf{x} .
\mathbf{I}	Identity matrix
\mathbb{R}	set of real numbers
\mathbb{R}_+	set of real nonnegative numbers

INDEXING

t	discrete time index in $\{1, \dots, T\}$
n	frame index in $\{1, \dots, N\}$
f	discrete frequency index in $\{1, \dots, M\}$
i	EEG channel index in $\{1, \dots, C\}$
j	audio source index in $\{1, \dots, J\}$
k	feature index in $\{1, \dots, K\}$
τ	time lag index in $\{1, \dots, L\}$

SIGNALS

$x(t)$	time domain mixture signal
$\mathbf{X}(f, n)$	magnitude spectrogram of $x(t)$
$\tilde{\mathbf{X}}(f, n)$	complex spectrogram of $x(t)$
$s_j(t)$	time domain signal of source j
$s_a(t)$	time domain signal of the attended source
$s_u(t)$	time domain signal of the unattended source
$r_i(t)$	time domain signal of the i -th EEG channel

Part I

INTRODUCTION

1 INTRODUCTION

1.1	Motivation and objective	3
1.2	Background	5
1.3	Contributions and thesis outline	11
1.3.1	Chapters summary	11
1.4	List of publications	14
1.5	Vademecum	15

1

Introduction

1.1 MOTIVATION AND OBJECTIVE

Over the past decades, the availability of services and tools for music creation, recording, production and distribution has increased exponentially. These have become accessible to a broader public thanks to many factors, such as the increasing accessibility of music technologies and the connectivity to the Internet, profoundly changing the music landscape and culture where home music production is no more an exception.

These changes have undoubtedly affected the music demography and culture [Walzer 2017], leading to a more democratic music landscape where more musicians can create and share their music with people from all around the world without intermediaries. Nevertheless, while these services and tools have become affordable and reliable, the technical skills and expertise required for using them may still represent an entry barrier for most users. Even professionals and experts are slowed down in their workflow by the complexity of some interfaces, which are not flexible and often limit the user creativity by sets of hardly-interpretable functionalities and parameters. There is an unavoidable learning curve that the user has to face to learn and adapt to an interface that is not a neutral intermediary between the user and the desired output. These intrinsic limitations can be overcome only by radically rethinking the way we interact with machines and by fully considering the user from the very beginning in the design of music technologies whose interfaces should be effortless and friction-free.

In parallel, the incredible growth of Human-Computer Interfaces (HCIs) led to a new way to interact with technology so that the interfaces are maximally simplified and adaptive to the user. Among those, Brain-Computer Interfaces (BCIs) are paving the way for a direct communication between humans and their devices by directly decoding the user's brain activity [Wolpaw and Wolpaw 2012]. The applications of BCIs are countless and span nowadays from clinical to home entertainment such as neurogaming and VR/AR [Kawala-Sterniuk et al. 2021]. In Music Information Research (MIR), the research field behind many music technologies, BCIs are still far from complementing the classical interfaces and being considered an integral part of the various applications except for music making and performance, which represent an exciting avant-garde mostly of musicians.¹ BCIs could help narrowing the intention gap which is a common experience of the user when dealing with complex

DESIGNER	WHAT THEY ARE RESPONSIBLE FOR
UI	ELEMENTS OF THE INTERFACE THAT THE USER ENCOUNTERS
UX	THE USER'S EXPERIENCE OF USING THE INTERFACE TO ACHIEVE GOALS
UZ	THE PSYCHOLOGICAL ROOTS OF THE USER'S MOTIVATION FOR SEEKING OUT THE INTERACTION
U α	THE USER'S SELF-ACTUALIZATION
U Ω	THE ARC OF THE USER'S LIFE
U ∞	LIFE'S EXPERIENCE OF TIME
U \bullet	THE ARC OF THE MORAL UNIVERSE

FIGURE 1.1: U^U: user1's skills that are used to get user2's experience allowing user1 to make a model for user2's interface. Image courtesy of xkcd, number 2141.

¹This avant-garde dates from 1965 Alvin Lucier's piece *Music for Solo Performer* and has evolved during the past decades with the more disparate interpretation of BCI for music making and performance. The reader can refer to [Williams and Miranda 2018] by Williams and Miranda for a nice review of brain-computer music interfaces.

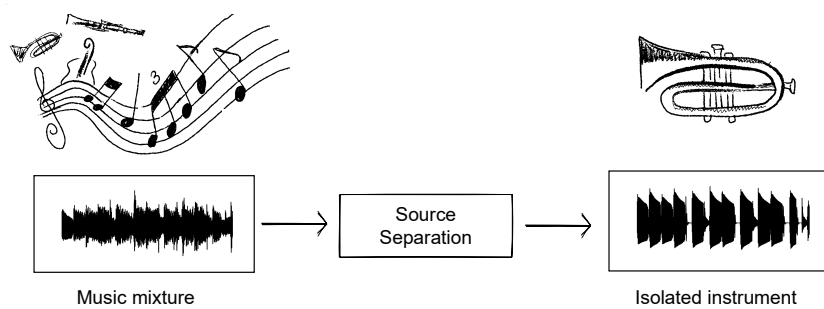


FIGURE 1.2: Music source separation process.

concepts. For instance, in a music recommendation system, the mental image of the desired music is often much clearer than the queries that the user needs to type in the interface to retrieve it. In music production, an audio effect can be described better by our mental idea of it than by the tuning of a set of hardly-interpretable parameters. **BCIs** may also significantly speed up and help the workflow of professional sound engineers and musicians, potentially uncover new understanding about the underlying creative process, or even discover new techniques for approaching it. In other cases, instead, the mental guidance can just replace the classical interfaces such as a mouse or a keyboard. However, the **BCI**, or, more generally, the **HCI**, is not only beneficial for the user but also for the underlying **MIR** algorithms that can leverage significant human expertise and knowledge to improve their performances.

In this thesis work, we make a first attempt of addressing the challenge of integrating **BCI** and music technologies on the specific **MIR** application of Music Source Separation (**MSS**), which is the task of isolating individual sound sources that are mixed in the audio recording of a musical piece (see Figure 1.2). A **MSS** system can be either directly exploited by the end-user (*e.g.*, a musician or a sound engineer) or be an intermediate step that significantly helps other downstream tasks such as automatic music transcription, instrument classification, score following, lyrics alignment, and many more others. This problem has been investigated for decades in the **MIR** community, but never considering **BCI** as a possible way to guide and inform **MSS** systems. This type of guidance can give the user an improved listening experience and boost many **MIR** downstream tasks making them interactive. The potential applications could target both the general audience and expert users such as sound engineers, video designers, and musicians.² Specifically, we explored how to perform a multimodal **MSS** exploiting previously not considered modalities, for instance the user's selective auditory attention to a source characterized in terms of his/her neural activity. Among the signals that can characterize the brain response, we consider the Electroencephalography, which is privileged when monitoring the brain activity for **BCI** because it allows for non-invasive acquisition with high temporal resolution and a reasonable cost.

The rest of the Chapter will introduce the reader to both music source separation and selective auditory attention. Finally, our contributions will be listed and the structure of the thesis will be exposed.

²People could thus enhance the instrument of interest during a concert by only “focusing” on it. Musicians could better study during live performances: imagine a student attending a concert who can enhance different instruments by switching their attention. Sound engineers could improve their workflow through intelligent neuro-steered headphones while remastering songs or soundtracks or video designers editing a video/movie.

1.2 BACKGROUND

- **MUSIC SOURCE SEPARATION** aims to isolate individual sources, such as singing voice, guitar, drums, cello, etc., mixed in an audio recording of a musical piece. More precisely, such individual voices can be referred to as *stems*, *i.e.*, recordings of individual instruments that are arranged together and mastered into the final audio mix.

Considering the case of single-channel recordings, one can assume that the mixture signal $x(t)$ at sample t is a linear mixture³ of J sources $s_j(t)$ such as:

$$x(t) = \sum_{j=1}^J s_j(t). \quad (1.1)$$

Given only $x(t)$, the goal of a general Audio Source Separation (**AuSS**) system is to recover one or more sources $s_j(t)$, where $j \in \{1, \dots, J\}$.

Nowadays, most state-of-the-art **MSS** systems are based on supervised Deep Learning (**DL**) systems [Stoller et al. 2018b; Défossez et al. 2019; Stöter et al. 2019; Hennequin et al. 2020], where an extensive collection of mixtures and corresponding isolated sources are needed during a training phase. Despite the release of dedicated datasets for this task [Rafii et al. 2017; Bittner et al. 2014], it is still hard for those models to generalize to unseen test data with significant timbral variation compared to training, and high-quality **MSS** remains an open problem for most instruments and music genres.

To mitigate this issue, one can inform the separation process with any prior knowledge one may have about the sources and the mixing process along with the audio signal. In this case, the approach is referred to as *informed MSS* and was often shown to enhance the separation result, especially for complex music mixtures, if compared to purely data-driven methods [Liutkus et al. 2013]. When the additional information comes from another modality than the audio itself, one can refer to it as *multimodal MSS* and this is the case depicted in [Figure 1.3](#). Examples of such side information include the score [Ewert et al. 2014; Ewert and Sandler 2017], the pitch contour [Virtanen et al. 2008], the lyrics [Schulze-Forster et al. 2019], the motion of the sound sources and spatial cues [Parekh et al. 2017].

³Usually, the final audio recording is not a linear sum of its stems due to the mixing and mastering steps, which includes multiple non-linear transformations and audio effects. This signal model still holds if we consider that the non-linear effects are applied to the individual stems, which are later summed to obtain the mixture.

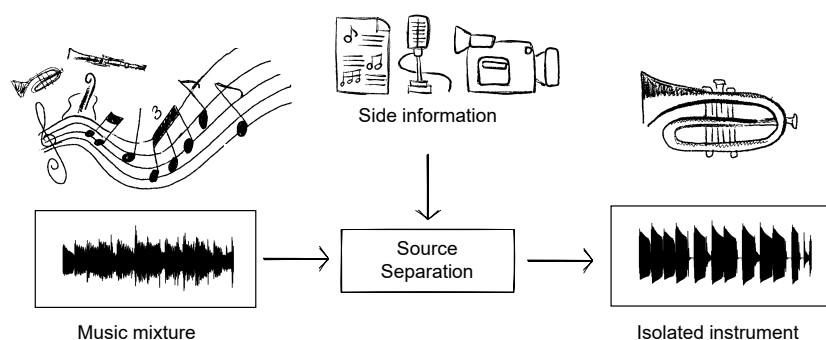


FIGURE 1.3: Informed source enhancement process. The aim is to separate one target source from the mixture exploiting any prior information we may have about the source.

One of the most underrated and powerful additional modalities is the user feedback which may leverage significant human expertise.

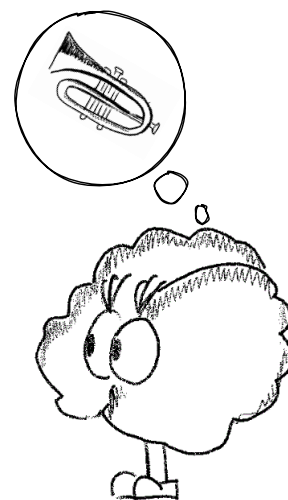
Particularly prolific was the use of time annotations provided by the user to learn an AuSS systems based on Nonnegative Matrix Factorization (NMF) or Nonnegative Tensor Factorization (NTF) [Bui et al. 2016; Laurberg et al. 2008; Ozerov et al. 2011; Duong et al. 2014a]. Some of them rely on dedicated graphical user interfaces, while others are interactive, where the user can iteratively improve and correct the separation [Bryan and Mysore 2013; Duong et al. 2014b]. Time annotations were also extended to more general Time-Frequency (TF) annotations [Lefevre et al. 2012; Lefèvre et al. 2014; Jeong and Lee 2015; Rafii et al. 2015] but those require much more expertise and effort from the user (and a more complicated user interface).

In DL-based systems, time activations have already been used in multi-task learning paradigms where the AuSS and the instrument activity detection tasks are jointly optimized [Stoller et al. 2018a; Hung and Lerch 2020]. Often, the time activations are relaxed to weak class labels, indicating a given instrument in a specific time interval, and are used as an input conditioning for the separation system [Swaminathan and Lerch 2019; Slizovskaia et al. 2019; Seetharaman et al. 2019; Karamatli et al. 2019].

There are also some interesting works where the user can hum [Smaragdis and Mysore 2009], sing or play [FitzGerald 2012] the source he/she wants to enhance as an example to the separation system. In the work from El Badawy et al., the user may listen to an audio mixture and type some keywords (e.g., “dog barking”, “wind”) describing the sound sources to be separated [El Badawy et al. 2014]. These keywords are then used as text queries to search for audio examples from the internet to guide the separation process. The user can also provide the fundamental frequency or manually correct it [Durrieu and Thiran 2012; Nakano et al. 2020] or associate each instrument to a microphone in a multi-channel recording [Di Carlo et al. 2017].

Beyond manual annotations, our body’s reaction to auditory stimuli manifests itself through many observable physiological phenomena. Reaction to music can be seen in the heartbeat variability [Chew et al. 2019; Chew 2021], in the body movements [Müller 2007], as well as in the neural activity [Sturm 2016] to mention a few. Such kind of information would help the separation process and make it also *interactive*, allowing for a number of futuristic applications where the human-machine interaction is simplified and natural. Among those physiological responses to music stimuli, we are interested in the neural response, focusing on the concept of *selective auditory attention*.

- ▶ **SELECTIVE AUDITORY ATTENTION** refers to a multitude of behavioural and cognitive mechanisms that allow humans to process concurrent sounds in a complex auditory scene to isolate the ones of interest [Kaya and Elhilali 2017]. The resulting perceptual effect is known as the “*Cocktail party effect*” and was first described by Cherry in [Cherry 1953] in relation to the perception of speech sources in noisy or multi-speaker settings. In practice, we can follow a single conversation while filtering out competing speakers, other sounds and noise. Therefore, one can define attention as “*the set of processes that allow the cognitive system to select the relevant information in a given context*” [Turatto



2006]. The need for attention comes from the fact that the cognitive system and the substrate on which it is based, *i.e.*, the brain, is not able to analyze all incoming information at the same level of detail [Turatto 2006]. In other words, it is not possible to be aware of everything at the same time.

Attention can be triggered via *bottom-up* mechanisms or by *top-down* factors [Kaya and Elhilali 2017]. In the first case, our attention is involuntarily attracted by sounds like a phone ringing, an alarm, a baby crying, which significantly differ from the ones of the background/neighbourhood, making them salient in that context [Koch and Tsuchiya 2007]. What first makes a stimulus salient is its sudden appearance (onset), but there are other aspects which are mostly studied for visual stimuli such as chromatic or shape characteristics [Jonides and Irwin 1981]. In the second case, our brain is voluntarily paying attention to a circumscribed region in space (focal attention), a particular feature (feature-based attention) or an object (object-based attention) in order to achieve a given task [Koch and Tsuchiya 2007]. In this case, the condition is that the subject knows the characteristic that defines the target [Bravo and Nakayama 1992]. It is the case of our brain focusing on a single conversation at a cocktail party or a particular musical instrument during a concert.

Since Cherry first published his work, the interest in auditory attention has increased substantially, producing a vast literature about it.⁴ As many other cognitive processes, auditory attention had been studied in earlier times through psychoacoustic experiments. The turning point happened with the advent of finer techniques for measuring the neural activity which can give a deeper insight into how, where and when selective attention manifests in our brain.

► MEASUREMENTS OF NEURAL ACTIVITY represent a crucial choice for designing experiments and BCIs based on selective attention. The most common techniques to measure neural activity as a response of a given stimulus are:

- Electroencephalography (EEG);
- Magnetoencephalography (MEG);
- Electrocorticography (ECoG);
- Local Field Potentials (LFP);
- functional Magnetic Resonance Imaging (fMRI).

Each technique has advantages and disadvantages in terms of invasiveness degree, spatial and temporal resolution. EEG and MEG [Cohen 1968], are non-invasive techniques with a high temporal resolution (order of milliseconds), making them appealing for tracking dynamic changes in the brain. They respectively measure the electric and magnetic fields on the scalp which comes with a low spatial resolution (order of centimeters) and low Signal-to-Noise-Ratio (SNR). To improve the SNR and spatial resolution, the electrodes can be implanted on the cortical surface, below the skull, using ECoG [Jasper and Penfield 1949], or directly inside the brain to record the LFP [Einevoll et al. 2013]. As those techniques are highly invasive, micro-electrodes are only implanted to monitor epileptic patients making this data rare and private.

ATTENTION AND CONSCIOUSNESS are different processes with different functions. Selective attention filters what can be elaborated in depth, so that it has access to the subject's consciousness. Attention is therefore a complex process of information selection that takes place in similar ways in the different sensory modalities: visual, tactile, auditory, etc. Consciousness, instead, summarize all information which are previously selected in order to perform deeper elaboration like decision making, language, rational thought, and so on [Koch and Tsuchiya 2007].

⁴The reader can refer to the paper review of Kaya and Elhilali [Kaya and Elhilali 2017] for a nice and concise review.

Moreover, the limited brain coverage gives information only regarding a restricted view of the auditory processing hierarchy [O’sullivan et al. 2014].

fMRI [Ogawa et al. 1990] represents a complementary technique. It has a limited temporal resolution (order of seconds) associated with a much higher spatial resolution (order of millimeters) which allows for precisely localizing the anatomical areas involved in a certain cognitive process. While this technique is largely used in cognitive research, its temporal resolution does not enable the dynamic tracking of attention which occurs within the milliseconds range, but only its allocation in space [Wang et al. 2017].

As we have seen, the choice of one of those techniques is strictly connected with the goal one wants to achieve. In our specific case, we need a high enough temporal resolution to dynamically track the auditory attention to a sound source. Secondly, we need to exclude invasive techniques which are not appropriate for a use in real-life scenarios as a part of **BCIs**. Thus the most natural choice for us is to consider **EEG** signals, which represent the most portable method by which the neural activity can be recorded and from now on we will restrict our focus to them.

- ▶ **THE NEURAL RESPONSE TO MUSIC** characterized in terms of **EEG** signals can be analyzed and understood using diverse approaches. As mentioned in the previous paragraph, **EEG** signals exhibit a very low **SNR** and therefore, it is hard to study a single phenomenon of interest. A typical work-flow is to repeat the stimulus several times and then average the **EEG** responses in order to keep only the stimuli-relevant information and attenuate noise. This approach relies on specifically selected or designed short stimuli, which are only appropriate to study specific attributes of music or the reaction to isolated sounds. Short stimuli generate a well-defined response in the **EEG** signals, called Event-related Potentials (**ERPs**).

ERPs exhibit a characteristic morphology: peaks are observed at a specific time-latency in the average **EEG** responses, which in the literature are referred to, for instance, as N100, P300, etc. **ERPs** are actually generated either from short stimuli or stimuli with high contrast with the background. The last characteristics can be re-created in experimental settings through the so-called *oddball paradigm*, where the subject is stimulated with a rare *deviant event* occurring among more frequent *standard events* [Treder et al. 2014]. In practice, less expected musical events produce stronger neural responses. This process is usually associated with bottom-up mechanisms and it is then difficult to distinguish effects due to the perceptual novelty of the stimulus from the ones due to the stimulus significance.

ERP are typically considered to study attention to particular musical structures such as note onsets, rhythm and pitch patterns or, at least unattended musical deviants among standard and attended events [Treder et al. 2014]. Some studies aim at understanding how the brain processes basic structural components of music such as pitch [Hyde et al. 2008; Kumar et al. 2011; Nan and Friederici 2013; Plack et al. 2014], timbre [Deike et al. 2004; Goydke et al. 2004; Caclin et al. 2007] as well as sensory dissonance, high-level melodic characteristics (e.g., melodic contour, key, mode, scale) and music-syntactic congruity [Koelsch et al. 2013; Sturm 2016]. The most studied components in the perception of

music appear to be note onsets, beats, rhythm and meter [Thaut 2005; Cirelli et al. 2014; Sturm 2016; Stober et al. 2016]. Music presents strong timing mechanisms that have been recognized to engage human behavior and brain function in multiple ways [Thaut 2005]. In particular, low-level structural elements of music, such as note onsets, can be considered distinct auditory events which allow the perception of more complex entities such as beat, rhythm, and meter [Sturm 2016].

Even though ERPs can give insights into how these musical attributes individually relate to neural processes, a different approach is needed to study the continuous brain response to a complex *naturalistic* stimulus such as a musical piece in its entirety. A few attempts have been made to track the dynamics of naturalistic music stimuli in the EEG signal. Cong et al. found evidence that the sound time-domain envelope is consistently reflected in the EEG [Cong et al. 2012]. Ofner and Stober reconstructed the spectrograms of both perceived and imagined music from the EEG [Ofner and Stober 2018], while Sturm et al. did the same for the note onset sequence [Sturm et al. 2015a; Sturm et al. 2015b]. Kaneshiro et al. investigated how musical engagement is reflected in the EEG-response and to what extent this is related to the temporal organization of acoustical events, their novelty and repetition [Kaneshiro et al. 2016b; Kaneshiro 2016; Kaneshiro et al. 2020; Kaneshiro et al. 2021b].

Di Liberto et al. showed that the cortical tracking of the music envelope is significantly modulated by cognitive factors such as attention and expectation, which strongly depend on the listener’s musical culture and expertise during both the listening [Di Liberto et al. 2020b; Di Liberto et al. 2020a] and imagery tasks [Marion et al. 2021; Di Liberto et al. 2021].⁵ Nevertheless, these works focused on stimulus reconstruction and not on decoding the attended instrument. Two attempts have been made for classifying the attended [Schaefer et al. 2013] or imagined [Marion et al. 2021] music and the attended instrument [Treder et al. 2014] but both of them focused on the elicited ERPs.

- ▶ **AUDITORY ATTENTION DECODING** Selective attention has been observed to modulate the neural activity in several different ways. Effects can be seen as an enhancement of neural activity [Hillyard et al. 1973; Woldorff and Hillyard 1991; Woldorff et al. 1993; Jäncke et al. 1999], connectivity [Lipschutz et al. 2002; Tóth et al. 2019] and synchronization [Doesburg et al. 2012] or as a more robust encoding of the attended source compared to the unattended ones [Mesgarani et al. 2009]. The latter makes it possible to decode the auditory attention, *i.e.*, determining which sound source a person is “focusing on”, by just observing at the listener’s brain response. This task is known as Auditory Attention Decoding (AAD), and typical applications are intelligent hearing aids where a neuro-steered enhancement of the attended speaker is desired [Han et al. 2019; Das et al. 2020b; Aroudi and Doclo 2020].

Previous AAD studies based on continuous MEG [Ding and Simon 2012; Akram et al. 2014; Brodbeck et al. 2018], ECoG [Mesgarani et al. 2009; Pasley et al. 2012; Mesgarani and Chang 2012] and EEG [O’sullivan et al. 2014; Crosse et al. 2016] signals have shown that the neural activity tracks dynamic changes in the audio stimulus and can be successfully used to decode selective attention in a complex auditory scene.

⁵Many studies had already shown that violations of music expectations, for instance, out-of-key notes embedded in chords [Koelsch et al. 2000], unlikely chords [Koelsch et al. 2007], elicit consistent ERPs. However, to elicit ERPs one requires substantial violations, which the listener can consider as a musician’s mistakes, while even the valid sequential events in a given musical culture do not have the same probability of occurring [Pearce 2005; Pearce and Wiggins 2012; Temperley 2008; Rohrmeier and Cross 2008; Temperley and Clercq 2013]. Thus, the associated expectation can vary accordingly in the full expectation strength range, and this could only be studied using continuous naturalistic musical stimuli [Di Liberto et al. 2020a].

In a number of works [Mesgarani et al. 2009; Pasley et al. 2012; Mesgarani and Chang 2012; O’sullivan et al. 2014; Crosse et al. 2016], a feature representation of the stimulus is reconstructed from the multi-channel neural recordings through a multi-channel Wiener-filter which is learned by solving a linear regression problem [Crosse et al. 2016]. Mesgarani and Chang were among the first to show that such reconstructed feature representations (in this case spectrograms) were highly correlated with the salient time-frequency features of the attended speaker’s voice, and were only weakly correlated with the unattended speaker ones [Mesgarani and Chang 2012].

These works all focused on reconstructing a specific category of stimuli, *i.e.*, speech. Much less developed is AAD research applied to other types of naturalistic stimuli such as music. In the latter case, one can recast the problem as one of decoding the attention to the “voice” of a particular musical instrument playing in an ensemble.⁶ However, this transposition is not straightforward as, unlike in the cocktail party problem where there is one source of interest to separate from unrelated background noise or speakers, music consists of multiple voices playing together in a coordinated way. Thus the sources are generally highly correlated, making the decoding problem even more difficult [Treder et al. 2014; An et al. 2014].

⁶Zuk et al. also showed that EEG responses are stronger to speech and music than to other natural sounds [Zuk et al. 2020], but, among the two, speech evokes larger responses and unique effects at low frequencies, leading to better reconstructions of the speech envelope than for music [Zuk et al. 2021].

- ▶ **NEURO-STEERED MUSIC SOURCE SEPARATION** The main limitation of most AAD approaches is their use of the separate “clean” audio sources. In fact, the feature representations extracted from the isolated sources are correlated with the ones predicted with the neural data to determine the attended source [Mesgarani and Chang 2012; O’sullivan et al. 2014]. However, the isolated sources are not available in realistic scenarios (*e.g.*, hearing aids) where only the mixture of the sound scene recorded by their microphones is available and an AuSS step, where single audio sources are extracted from their mixture, is needed. This limitation is strongly intertwined with a specular aspect of AuSS, whose process can be informed by prior knowledge one has about the sources. Few works have been proposed in the last years that relate speech source separation with AAD, but most of the time, the two tasks are tackled independently. Either the separated sources are used as clean sources to decode attention, or the EEG is used to decode which source needs to be enhanced. This has been implemented in multi-channel scenarios using beamforming [Van Eyndhoven et al. 2017; Aroudi et al. 2018; Aroudi and Doclo 2019; Aroudi and Doclo 2020] and in single-channel scenarios using Deep Neural Networks (DNNs) [O’Sullivan et al. 2017; Das et al. 2017; Han et al. 2019; Das et al. 2020b].

Only a few works proposed in parallel to our own use directly the neural activity of the listener to inform a source separation model, but they all focused on speech sources [Pu et al. 2019; Ceolini et al. 2020]. In [Pu et al. 2019], the authors propose an adaptive beamformer that reduces noise and interference but, at the same time, maximises the Pearson correlation between the envelope of its output and the decoded EEG. In [Ceolini et al. 2020], instead, a speech separation DNN is informed with the decoded attended speech envelope. Nevertheless, none of these works considers music audio signals.

1.3 CONTRIBUTIONS AND THESIS OUTLINE

Within this PhD project, we have investigated how **MSS** methods can be guided interactively by the user's brain activity. We focused on the concept of selective auditory attention, which allows humans to process concurrent sounds and isolate the ones of interest. Our work investigates how to leverage this phenomenon to guide a **MSS** system and automatically separate the attended source. Such a formulation would also allow reformulating the **AAD** problem without the need for the ground truth sources.

Among the signals that can characterize the brain response, we consider the **EEG**, which allow for non-invasive neural activity acquisition with high temporal resolution, making it an ideal candidate for developing **BCIs**.

The proposed approach is summarized in **Figure 1.4** and can be divided into two main tasks, which can be tackled jointly:

- *Decoding auditory attention to a target instrument in polyphonic music mixtures;*
- *Neuro-steered source separation of the target instrument from a polyphonic music mixture.*

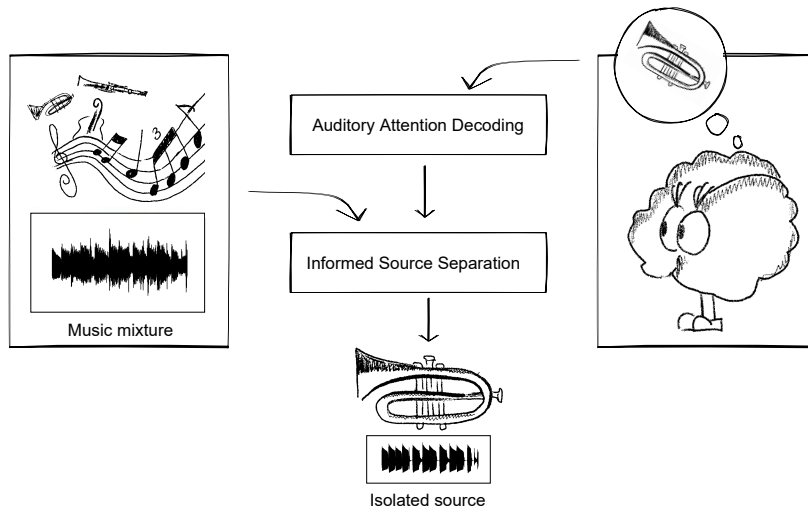


FIGURE 1.4: Schematics of the proposed framework: the source separation algorithm is guided by the user's selective auditory attention to that instrument, which is tracked in his/her neural response to music.

1.3.1 Chapters summary

We briefly describe here the contents of each Chapter, emphasizing the contributions and listing the associated publications. While **Part I** introduces the motivation and objective behind this work, the rest of the thesis is divided in two main parts:

► **PART II AUDITORY ATTENTION DECODING**

This part focuses on the first task, *i.e.*, *decoding auditory attention to a target instrument in polyphonic music mixtures*.

Chapter 2 This Chapter describes in detail the music-related EEG dataset we have assembled for the thesis, namely MAD-EEG, which allows for studying the problems of single-trial EEG-based Auditory Attention Decoding and EEG-guided Music Source Separation. It is crucial for the reader to have a clear understanding of the recording protocol and how the stimuli were built to easily follow the rest of the thesis, and how the proposed algorithms are applied to this data. MAD-EEG, represents our first main contribution and is available to the research community as a free resource. The dataset was acquired by my colleague Gabriel Trégoat during his internship at Télécom Paris and finalised by me, leading to the following conference publication:

- Cantisani, Giorgia, Gabriel Trégoat, Slim ESSID, and Gaël Richard (2019b). “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)*, pp. 51–55

Chapter 3 This Chapter describes the second main contribution of the thesis, relating to the problem of *decoding the auditory attention to a target instrument in polyphonic music* which was extensively investigated on the MAD-EEG dataset. The primary outcome of this study is that the EEG tracks musically relevant features highly correlated with the attended source and weakly correlated with the unattended one making it possible to decode the auditory attention towards a specific instrument in the mixture. This study is particularly important within the thesis, as the proposed neuro-steered Music Source Separation approaches are built upon the results of this Chapter. This work led to the following conference publication:

- Cantisani, Giorgia, Slim ESSID, and Gaël Richard (2019a). “EEG-Based Decoding of Auditory Attention to a Target Instrument in Polyphonic Music”. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*

► **PART III** NEURO-STEERED SOURCE SEPARATION

This part focuses on the second task, *i.e.*, *neuro-steered source separation of the target instrument from a polyphonic music mixture*.

Chapter 4 This Chapter introduces the central contribution of the thesis, a *neuro-steered music source separation* framework built upon the results of the previous Chapter and conducts an extensive evaluation of the proposed system on the MAD-EEG dataset. Specifically, we leverage the fact that the attended instrument’s neural encoding is substantially stronger than the one of the unattended sources left in the mixture to inform a source separation model based on a new variant of NMF named Contrastive-NMF and automatically separate the attended source. This unsupervised NMF variant is particularly advantageous as it allows us to incorporate additional information in a principled optimisation fashion and does not need training data, which is particularly difficult to acquire

for applications involving EEG recording. This work led to the following conference publication and preprint:

- Cantisani, Giorgia, Slim Essid, and Gaël Richard (2021b). “Neuro-steered music source separation with EEG-based auditory attention decoding and contrastive-NMF”. in: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*
- Cantisani, Giorgia, Slim Essid, and Gaël Richard (2021a). “EEG-based Decoding of Auditory Attention to a Target Instrument for Neuro-steered Music Source Separation”. In: *journal in preparation*

Chapter 5 The scarcity of music-related EEG data precludes the possibility of tackling the problem of neuro-steered music source separation with fully supervised deep learning approaches. In this chapter, we explored alternative learning strategies to alleviate this problem. Specifically, we propose to adapt a state-of-the-art music source separation model to a specific mixture using the time activations of the sources provided manually by the user or derived from his/her neural activity which are available only at test time. This paradigm can be referred to as *one-shot* adaptation, as it acts on the target song instance only. A large part of the material presented in the chapter is the result of a work conducted during my internship at InterDigital R&D France under the supervision of Alexey Ozerov and led to the following conference publication:

- Cantisani, Giorgia, Alexey Ozerov, Slim Essid, and Gaël Richard (2021c). “User-guided one-shot deep model adaptation for music source separation”. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*

- ▶ **FINALLY**, the dissertation concludes with **Chapter 6**, which discusses the principal findings of the current investigation, together with discussion on future perspectives and research directions.
- ▶ **IN THE APPENDICES**, the reader will find a chapter about a science dissemination project I have coordinated which led to the release of a short video explaining in simple terms what Music Information Research (**MIR**) is all about. This part is not strictly related to the research topic of the thesis but the more general problem of science communication and dissemination.

1.4 LIST OF PUBLICATIONS

- Cantisani, Giorgia, Gabriel Trégoat, Slim Essid, and Gaël Richard (2019b). “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)*, pp. 51–55
- Cantisani, Giorgia, Slim Essid, and Gaël Richard (2019a). “EEG-Based Decoding of Auditory Attention to a Target Instrument in Polyphonic Music”. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*
- Cantisani, Giorgia, Slim Essid, and Gaël Richard (2021b). “Neuro-steered music source separation with EEG-based auditory attention decoding and contrastive-NMF”. in: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*
- Cantisani, Giorgia, Slim Essid, and Gaël Richard (2021a). “EEG-based Decoding of Auditory Attention to a Target Instrument for Neuro-steered Music Source Separation”. In: *journal in preparation*
- Cantisani, Giorgia, Alexey Ozerov, Slim Essid, and Gaël Richard (2021c). “User-guided one-shot deep model adaptation for music source separation”. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*

1.5 VADEMECUM

The reader will have already noticed that a large margin is left free on each manuscript page. We will use it to insert additional insights, notes and figures to complete each subject. This graphic template is inspired by the work of Tufte and Graves-Morris [Tufte and Graves-Morris 1983]⁷ and exhibits some peculiarities:

- at most three levels of sub-headings: section, subsection, and Tufte's *new-thought* [Tufte and Graves-Morris 1983] and ► to capture attention;
- reference sidenotes on the margin are used as footnotes, providing additional insights;
- *italic* sidenotes and figures without proper reference numbers on the margin are meant to provide optional information and can be read in a second time;
- orange is used for clickable internal reference, such as for sections § 1.2 and acronyms AAD;
- grey and ↗ is used for clickable external link, such as my website↗.

⁷The colophon of the thesis reports more information on the template.

Part II

DECODING OF AUDITORY ATTENTION TO MUSIC

2	mad-eeg: AN EEG DATASET FOR DECODING AUDITORY ATTENTION TO A TARGET INSTRUMENT IN POLYPHONIC MUSIC	
2.1	Introduction	19
2.2	Related works	20
2.3	Dataset creation	20
2.3.1	Participants	21
2.3.2	Stimuli	22
2.3.3	Recording protocol	23
2.3.4	Data Acquisition and Preprocessing	26
2.4	Conclusions	27
3	maad: EEG-BASED DECODING OF AUDITORY ATTENTION TO A TARGET INSTRUMENT IN POLYPHONIC MUSIC	
3.1	Introduction	29
3.2	Related works	30
3.3	Methods	31
3.3.1	Audio Feature Extraction	32
3.3.2	Temporal Response Function	32
3.3.3	Regularization	34
3.4	Experiments	34
3.4.1	Evaluation	34
3.4.2	Experimental Results	34
3.5	Conclusions	39

2

MAD-EEG: an EEG dataset for decoding auditory attention to a target instrument in polyphonic music

- ▶ **SYNOPSIS** This Chapter describes in detail the music-related **EEG** dataset we have assembled for the thesis, namely MAD-EEG, which allows for studying the problems of single-trial **EEG**-based Auditory Attention Decoding and **EEG**-guided Music Source Separation. It is crucial for the reader to have a clear understanding of the recording protocol and how the stimuli were built to easily follow the rest of the thesis, and how the proposed algorithms are applied to this data. MAD-EEG, represents our first main contribution and is available to the research community as a free resource. The dataset was acquired by my colleague Gabriel Trégoat during his internship at Télécom Paris and finalised by me, leading to the following conference publication:

- Cantisani, Giorgia, Gabriel Trégoat, Slim Essid, and Gaël Richard (2019b). “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)*, pp. 51–55

2.1 INTRODUCTION

MAD-EEG is a new, freely available dataset for studying **EEG**-based **AAD** considering the challenging case of subjects attending to a target instrument in polyphonic music. The dataset represents the first music-related **EEG** dataset of its kind, enabling, in particular, studies on single-trial **EEG**-based **AAD** while also opening the path for research on other **EEG**-based music analysis tasks such as neuro-steered **MSS**.

MAD-EEG has so far collected 20-channel **EEG** signals recorded from 8 subjects listening to solo, duo and trio music excerpts and attending to one pre-specified instrument. The stimuli were designed considering variations in the number and type of instruments in the mixture, spatial rendering, music genre and melody, which allow testing the influence of certain factors on the **AAD** and neuro-steered **MSS** performance.

It is worth noting that the setting is entirely different from the ones previously proposed. The experimental protocol usually applied for **AAD** data acquisi-

Keywords: Auditory attention, Polyphonic music, EEG.

Resources:

- 🔗 [Paper](#)
- 🔗 [Open access](#)
- 🔗 [Dataset](#)

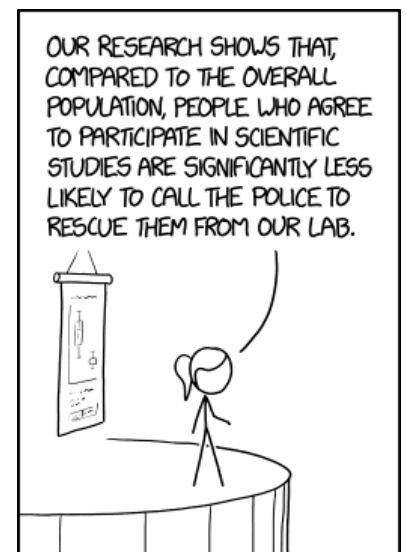


FIGURE 2.1: Our research shows that, compared to the overall population, people who agree to participate in my scientific studies are significantly more likely to ask me to participate in their studies. Image courtesy of xkcd, number 1999.

tion like the ones of [O’sullivan et al. 2014; Crosse et al. 2016; Treder et al. 2014], considers two monaural sources each played to a different ear through headphones. Instead, in our recording sessions, the stimuli were reproduced using speakers, and the audio was rendered in varying spatial configurations.

2.2 RELATED WORKS

A few publicly available music-related EEG datasets exist. Stanford University researchers have assembled a number of such datasets: the Naturalistic Music EEG Dataset-Hindi (NMED-H) [Kaneshiro et al. 2016a], the Naturalistic Music EEG Dataset-Tempo (NMED-T) [Losorelli et al. 2017], the Naturalistic Music EEG Dataset-Rhythm Pilot (NMED-RP) [Appaji and Kaneshiro 2018], the Naturalistic Music EEG Dataset-Elgar (NMED-E) [Kaneshiro et al. 2021a] and the Naturalistic Music EEG Dataset - Minimalism (NMED-M) [Dauer et al. 2021]. Each of these contains EEG and behavioural responses to different kinds of naturalistic music stimuli. Other music-related EEG datasets are the OpenMIIR dataset [Stober et al. 2015] acquired for studying music perception and imagination and the DEAP [Koelstra et al. 2012] and the BCMI [Daly et al. 2020] databases, acquired for studying affective responses to music. Recently, it was released a dataset that allows to compare brain responses to the music of individuals with an intellectual and developmental disorder and typically developing ones [Sareen et al. 2020]. In such datasets, the user focused on the entire stimulus and not on a particular instrument. Thus they are not relevant for the study of the AAD problem. The only publicly available music-related EEG dataset where participants were asked to attend to a target instrument in the music mixture is the music BCI dataset collected by Treder et al. [Treder et al. 2014]. The dataset was explicitly designed for studying ERP-based AAD using a multi-streamed oddball paradigm, where a repetitive musical pattern is interspersed with a randomly occurring deviant pattern that yields clean P300 ERPs.² However, the oddball paradigm’s assumption does not often hold in real-world music compositions as we have seen in § 1.2.

The situation is different when considering datasets for AAD in speech. In this case, several datasets and methods were designed to study this problem using a single-trial approach. Nevertheless, only a few of them are accessible [Fuglsang et al. 2017; Das et al. 2020a]

Taking inspiration from the speech-related EEG datasets, we assembled our EEG dataset from subjects listening to realistic polyphonic music and attending to a particular instrument in the mixture. Our dataset represents the first EEG dataset designed explicitly for studying AAD applied to realistic polyphonic music using single-trial techniques.

2.3 DATASET CREATION

Surface EEG signals were recorded from 8 subjects while listening to polyphonic music stimuli. For each audio stimulus consisting of a mixture containing from two to three instruments, the subjects were asked to attend to a particular instrument.

²P300: is an ERP across the parietal-central area of the skull that occurs around 300 ms after stimuli presentation [Fabiani et al. 1987]. Its wave is larger after the target stimulus and only occurs if the subject actively engages in detecting the targets. Its amplitude varies with the target improbability, while its latency varies with the difficulty of discriminating the target from the standard stimuli.

Each subject listened to a total of 78 stimuli presented in a random order, each one consisting of 4 repetitions of the same roughly 6-second-long music excerpt, leading to a total of approximately 30-32 minutes of 20-channel EEG recordings per subject. Each subject listened to 14 solos, 40 duets and 24 trios, except one subject who only listened to 7 solos, 29 duets and 17 trios.

2.3.1 Participants

Eight volunteers (7 males and one female, all but one right-handed, aged between 23 and 54 years, mean age 28) took part in the study. All of them were healthy and reported no history of hearing impairments or neurological disorders. All participants signed a consent that informed them about the experiment's modalities and purposes. All the data was anonymized.

The study conforms with the Declaration of Helsinki [World Medical Association 2013]. The data were collected preventing all possible health risks for the participants. In particular, they were not exposed to sound pressure levels that can impair their hearing, be painful or lead to other adverse effects; they were taught carefully about how to behave correctly and safely during the experiment and with the EEG acquisition equipment.

The participants were hired within our laboratory, and took part in the experiments as volunteers. In particular, 2 were PhD students, 5 Master students of our lab, and one a sound engineer of the school. They were all non-professional musicians with varying years of musical experience (from 7 to 30 years, mean 13.5), as can be seen in Figure 2.2. However, they all defined themselves as beginners. Five out of them play the guitar, one the bass, one the drums, and one is a multi-instrumentalist playing the drums, guitar and bass. They all practised regularly with their instrument (from 2 to 14 hours per week, mean 6.25). Figure 2.3 presents the number of hours per week that each subject usually spends listening to music (blue) and practices his/her instrument (red). All of them were familiar with the modern instruments in the dataset (drums, guitar, bass and singing voice), while for specific classical instruments (bassoon, French horn and oboe) not all of them were equally confident as can be seen in Figure 2.4. Thus, they were trained to recognize them before the experiment

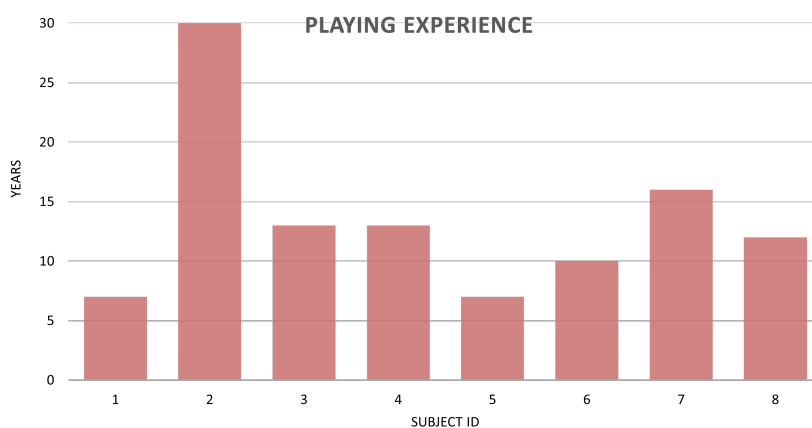


FIGURE 2.2: Years of musical instrument playing experience for each subject.

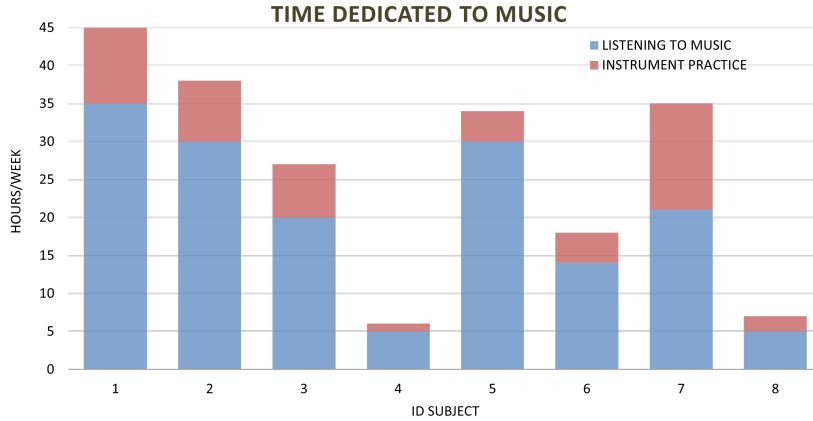


FIGURE 2.3: Hours per week dedicated by each subject respectively to listening to music (blue) and practising their instrument (red).

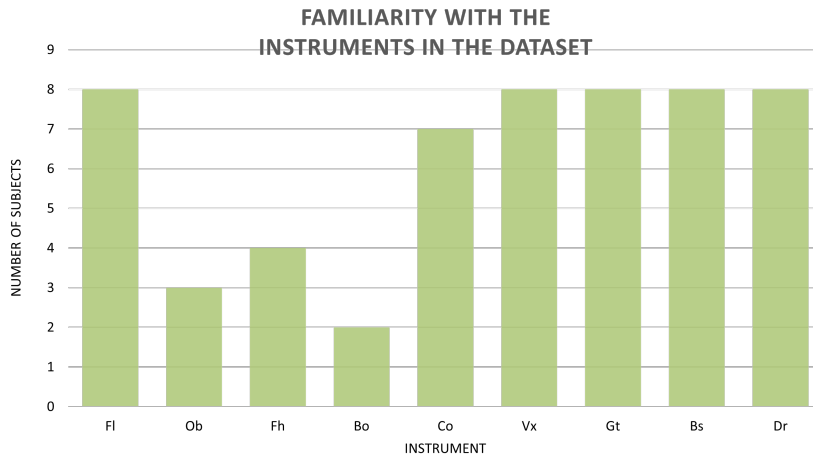


FIGURE 2.4: The number of subjects familiar with each instrument that appears in the dataset. The naming convention for the instruments is: Fl for Flute, Ob for Oboe, Fh for French Horn, Bo for Bassoon, Co for Cello, Vx for Voice, Gt for Guitar, Bs for Bass and Dr for Drum.

using excerpts not used as stimuli.

2.3.2 Stimuli

The stimuli consist of realistic polyphonic music mixtures containing two to three instruments played concurrently in an ensemble. The chosen mixtures reproduce a realistic setting. In particular, real music compositions for which we had access to the isolated instrument tracks were chosen for pop pieces. For Classical music pieces, instead, a selection of excerpts played by single instruments were linearly mixed as follows:

$$x(t) = \sum_{j=1}^J g_j s_j(t), \quad (2.1)$$

where $s_j(t)$ is the mono-channel audio track of the single instrument j , g_j is the corresponding gain, T its number of samples, and J is the number

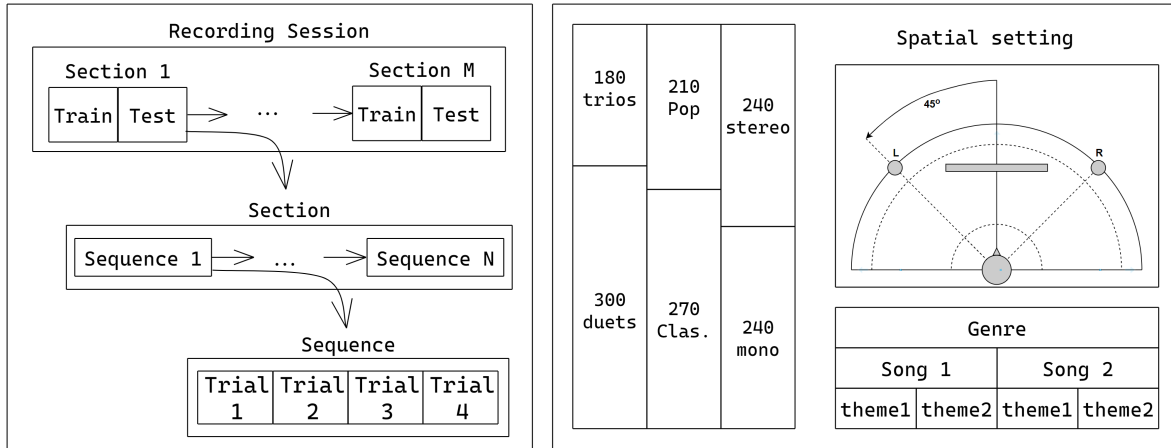


FIGURE 2.5: On the left, an illustration of the recording session for one subject. A recording session is divided into sections. Each section is associated with a given musical piece and consists of a training and a test phase, where a series of stimuli sequences is played. Each stimulus sequence consists of 4 trials where the same stimulus is listened to repetitively. On the right, details about the mixtures and how they are spatially rendered.

of instruments in the mixture. Finally, the sound volume was normalized to avoid bias due to the loudness of the audio.

In order to test the influence of certain factors on the attention decoding performance, we considered different configurations in the choice of the musical stimuli (see Figure 2.5 for a map of the variants):

- Two musical *genres*: pop and Classical music. Pop excerpts were chosen with sharp rhythmical and harmonic patterns to contrast with the Classical music ones, mostly melodic.
- Two musical *pieces* per genre and two *themes* per musical piece. That is, for the same piece, two different excerpts corresponding to different parts of the score.
- Two *ensemble types*: duets and trios.
- Two *spatial rendering configurations*: monophonic and stereo. The loudspeakers were situated $\pm 45^\circ$ along the azimuth direction relative to the listener (see Figure 2.5). The stereo spatial rendering was implemented by merely using conventional stereo panning where one has one instrument mostly on the right and the other one mostly on the left for duets, while for trios the third instrument is in the centre. The target instrument is never in the same position across different sequences.
- *Musical instruments* present in the mixture: different combinations of flute, oboe, French horn, bassoon and cello for Classical pieces, along with singing voice, guitar, bass and drums for pop excerpts.

2.3.3 Recording protocol

Each stimulus duration had to be long enough to allow the study of AAD on a single-trial basis while targeting realistic music excerpts. On the other hand, the experiment's duration had to remain reasonably short to control

the subject's cognitive load and avoid an unsatisfactory concentration level throughout the session. Consequently, we limited the duration of a stimulus to around 6 seconds. Then, during the experiment, each stimulus was heard by the subject four consecutive times, referred to as *trials*, corresponding to around 24 seconds of EEG recordings, which is long enough for studying single-trial methods while still making it possible to consider EEG-signal averaging techniques. Since each subject listened to 78 stimuli, this corresponds to approximately 30-32 minutes of recordings per subject.

For each subject the *recording session* was divided in *sections* as can be seen in Figure 2.5. In each section a series of stimuli *sequences* is played. Each section is actually composed of a *training* and a *test* phase. During the training phase, single instrument tracks of a given piece are played separately as solos, in a random order. Then, during the test phase, all the corresponding duo and trio variants of the same piece are played in a random order, but with a potentially different spatial rendering and considering a different theme of the same musical piece. For each instrument solo of a given piece, between 2 and 6 mixtures where the same instrument is attended to are available, but the theme and spatial rendering may differ. This is meant to allow studies on the generalization ability of an AAD system when the pitch contour varies between training and testing.

A section is presented to the user through a slide-show video showing instructions, displayed as white text on a black background, asking the participant to attend to a particular instrument and visually fix a cross at the centre of the screen. A "beep" precedes each stimulus launch.

ATTENTION SELF-ASSESSMENT Right after each section, the subjects were asked to self-assess the level of attention they paid to each stimulus on a discrete scale ranging from 1 to 5. The level of attention was generally high, except for only a few stimuli (see Figure 2.6) which can be used to evaluate how

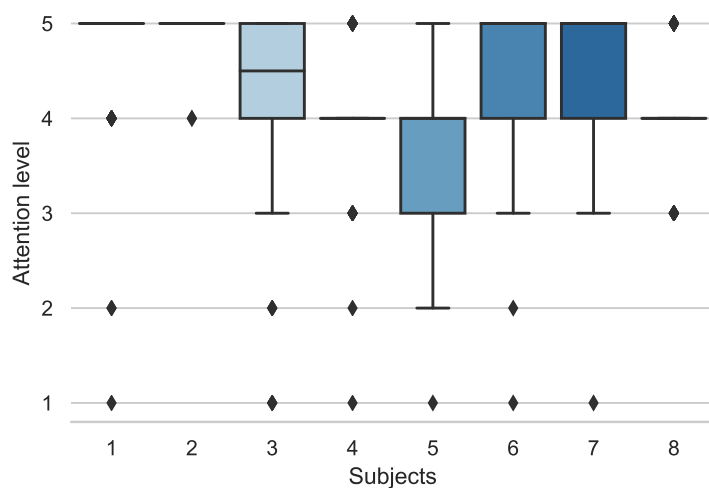


FIGURE 2.6: Statistics of the level of attention reported by each subject at the end of the sections for each stimulus proposed in the section. The scale goes from 1 to 5, where 0 represents no attention and 5 the maximum level.

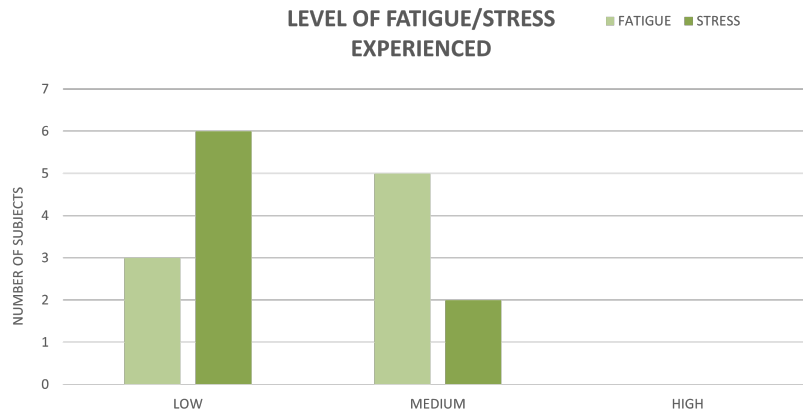


FIGURE 2.7: Number of subjects that experiences respectively a low, medium or high level of fatigue and stress.

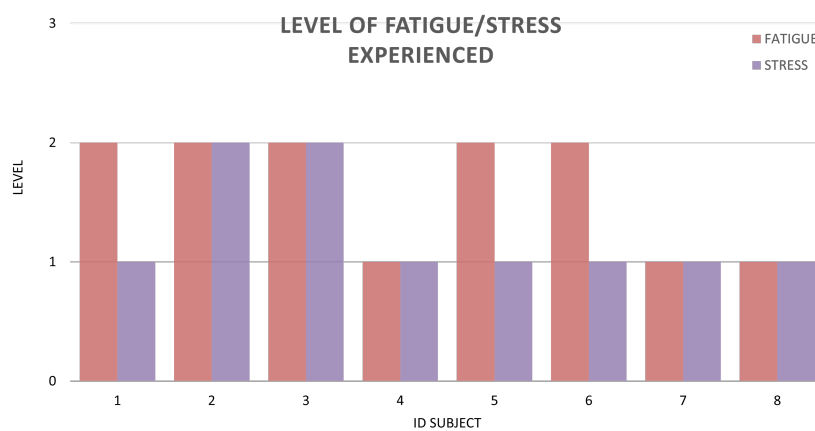


FIGURE 2.8: Fatigue and stress experienced by each subject: low (1), medium (2) or high (3).

the performance of an attention decoding system changes with the attention self-assessed by the subjects. At the end of the session, the participants were also asked to indicate the level of fatigue and stress experienced (low, medium or high) and if they had comments or remarks on the whole process. In general, the fatigue/stress experienced was reasonable as can be seen in [Figure 2.7](#) and [Figure 2.8](#), ensuring that the quality of the collected data is good since the subjects were not overloaded.

EEG SIGNAL ARTIFACTS Blinking, heartbeat, and other involuntary movements significantly modify the EEG recordings while being independent of the stimuli and, therefore, can bias the recorded signals' interpretation. Thus, subjects were instructed to maintain, for the duration of each trial, visual fixation on a cross at the centre of the screen and minimize eye blinking and other motor activities. Moreover, during breaks at the beginning, middle and end of the experiment, a series of instruction videos were used to ask the participants to perform different gestures (shake their cheeks, blink their eyes), each of which has a particular influence on the EEG. This portion of EEG signals are also available within the dataset and can be used by those interested in studying artefact removal techniques, possibly using them on the music-related portions

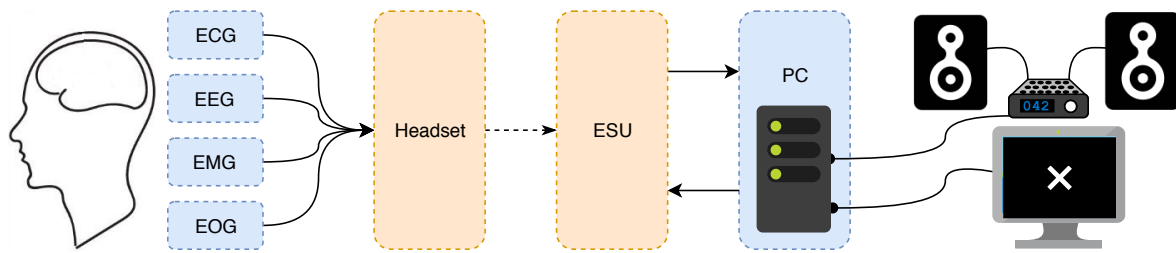


FIGURE 2.9: Block diagram of the acquisition system: surface EEG, EOG, EMG and ECG are acquired through a set of electrodes connected to the headset. An ESU receives the EEG data and timestamps from the headset via Bluetooth and transmits them to the acquisition software. The beginning of the stimulus playback is detected in real-time through a Python script monitoring the playback PC sound card output and then transmitted to the ESU for synchronization.

of this EEG dataset.

2.3.4 Data Acquisition and Preprocessing

A B-Alert X24¹ headset was used to record the surface EEG, Electrooculography (EOG), Electromyography (EMG) and Electrcardiography (ECG) signals of the participants, as well as their head motion acceleration, thanks to an integrated inertial measurement unit, all at a sampling frequency $f_s = 256Hz$. The EEG headset consists of a wireless digital acquisition unit connected to an electrode strip as the one in Figure 2.10. The strip features electrodes F1, F2, F3, F4, Fz, C1, C2, C3, C4, Cz, CPz, P1, P2, P3, P4, Pz, POz, O1, O2 and Oz, placed according to the 10-20 montage system. Active electrodes were referenced to the left mastoid in a unipolar setting.

EOG electrodes were placed above and below each eye diagonally, while ECG ones were placed in the middle and left side of the last rib. EMG electrodes were placed in such a way that they record the activity of the big zygomatic (whose activity can be recorded at a position situated at mid-distance between the top of the ear and the eye of the mouth), and the inferior palpebral orbicularis, which can be contracted simultaneously only involuntarily.

SYNCHRONIZATION A custom software interface automatizes the whole acquisition process and save the necessary information to synchronize the stimuli and the EEG responses. An External Sync Unit (ESU) receives data from the EEG headset via Bluetooth and passes it over to the acquisition software along with timestamps associated with each EEG signal sample as can be seen in Figure 2.9. This ESU can also receive the custom experimenter's auxiliary data and record it along with the EEG data. We use this feature of the ESU to record stimulus playback start times accurately. Thus, the beginning of the stimulus playback is detected in real-time through a Python script monitoring the playback PC sound card output. These playback start-events are then sent through the PC's serial port to the ESU so they can be marked as timestamps for the stimuli. This is done to detect the exact time instant when each stimulus starts within a 10-ms tolerance window. The EEG and the stimuli timestamps are thus saved by the EEG recording software and can be subsequently used offline for synchronization.

¹<https://www.advancedbrainmonitoring.com/xseries/x24/>

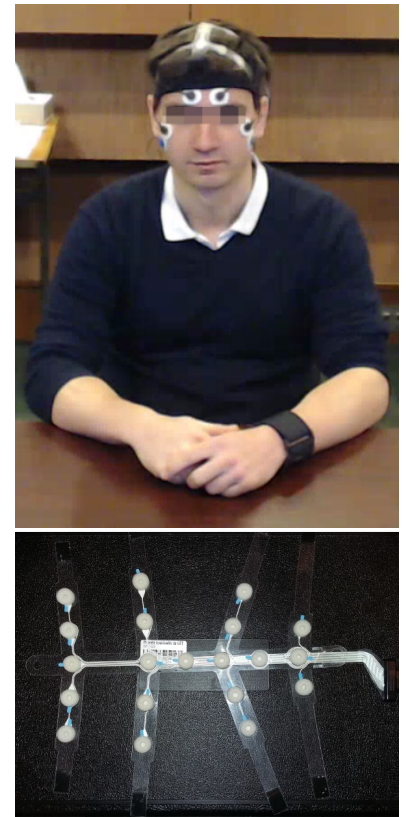


FIGURE 2.10: A wireless B-Alert X24 headset was used to record the EEG data. In the figure, one can see the EEG electrodes strip (the left part corresponds to the frontal electrodes).

AUDIO PLAYBACK Music stimuli were presented using a Hi-Fi audio playback system (JMLab chorus lcr700 speakers and Yamaha DSP-AX2 Natural Sound AV amplifier). The listener was seated at the centre of the room, 2 meters from a full HD TV (165 cm) screen and 2.8 meters from the two speakers. The speakers were positioned $\pm 45^\circ$ along the azimuth direction relative to the listener as depicted in Figure 2.11. The spatial rendering was implemented by merely using conventional stereo panning. This means, that for each instrument in the mixture the left and right channels are obtained as follows:

$$\begin{bmatrix} L \\ R \end{bmatrix} = \begin{bmatrix} \alpha \\ 1 - \alpha \end{bmatrix} s_j(t), \quad (2.2)$$

where $\alpha \in [0, 1]$ and $s_j(t) \in \mathbb{R}^{1 \times T}$ is the mono-channel audio track of the single instrument j . The volume was set to be comfortable and was kept constant across all sessions.

DATA PREPROCESSING Firstly, the EEG data were visually inspected to detect anomalies and keep only valid recording takes (e.g., subject 5 has EEG responses to 53 stimuli instead of 78). Then, the acquired EEG data was synchronized with each stimulus, the 50 Hz power-line interference was removed using a notch filter, and EOG/ECG artefacts were detected and removed using Independent Component Analysis (ICA). The frequencies below 1 Hz were filtered out using a Butterworth zero-phase filter of order 2. Each channel was normalized to ensure zero mean and unit variance. All the data was anonymized.

2.4 CONCLUSIONS

MAD-EEG is a novel, free dataset that enables studies on the problem of EEG-based Auditory Attention Decoding to a target instrument in realistic polyphonic music and EEG-guided Music Source Separation. The numerous variants in the stimuli and the behavioural data allow for investigating how such factors impact on the AAD and neuro-steered MSS performance.

It represents the first dataset of its kind for music stimuli and can be differentiated also from those commonly used for studying AAD for speech stimuli. In fact, the proposed experimental setting differs from the ones previously considered as the stimuli are polyphonic and are played to the subject using speakers instead of headphones.

It is a common experience that acquiring such a type of dataset is time-consuming and expensive. It requires specific equipment and experience but as well a long phase of preparation and experimental design. It takes much time to recruit participants who can participate in the experiment, and each of those is available for a limited amount of time and cannot be overloaded with too long recording sessions. Therefore, the dataset is limited in terms of recording hours and the number of participants but still allows for studying those problems if specific strategies are adopted to avoid overfitting.

MAD-EEG represents the first main contribution of the thesis and is made available to the research community as a free resource.

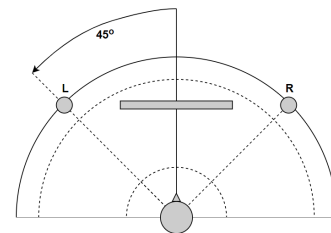


FIGURE 2.11: The speakers were positioned $\pm 45^\circ$ along the azimuth direction relative to the listener.

3

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

- ▶ **SYNOPSIS** This Chapter describes the second main contribution of the thesis, relating to the problem of *decoding the auditory attention to a target instrument in polyphonic music* which was extensively investigated on the MAD-EEG dataset. The primary outcome of this study is that the **EEG** tracks musically relevant features highly correlated with the attended source and weakly correlated with the unattended one making it possible to decode the auditory attention towards a specific instrument in the mixture. This study is particularly important within the thesis, as the proposed neuro-steered Music Source Separation approaches are built upon the results of this Chapter. This work led to the following conference publication:

- Cantisani, Giorgia, Slim Essid, and Gaël Richard (2019a). “EEG-Based Decoding of Auditory Attention to a Target Instrument in Polyphonic Music”. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*

3.1 INTRODUCTION

In this Chapter, we address the problem of **EEG**-based decoding of auditory attention to a target instrument in realistic polyphonic music. To this end, we exploit the so-called *backward model*, which was proven to decode the attention successfully to speech in multi-speaker environments [O’sullivan et al. 2014; Crosse et al. 2016]. To our knowledge, this model was never applied before to musical stimuli for **AAD** and we extensively evaluated it on the MAD-EEG dataset. The task we consider here is quite complex compared to the classical one for speech stimuli which considers two monaural sources each played to a different ear through headphones. Here, the music stimuli are polyphonic, including duets and trios, and the mixtures are reproduced using loudspeakers in varying spatial configurations. We consider the decoding of three different audio representations and investigate the influence on the decoding performance of multiple variants of musical stimuli, such as the number and type of instruments in the mixture, the spatial rendering, the

Keywords: Auditory attention decoding, Polyphonic music, EEG, Stimulus reconstruction model.

Resources:

- 🔗 Paper
- 🔗 Code randomization test

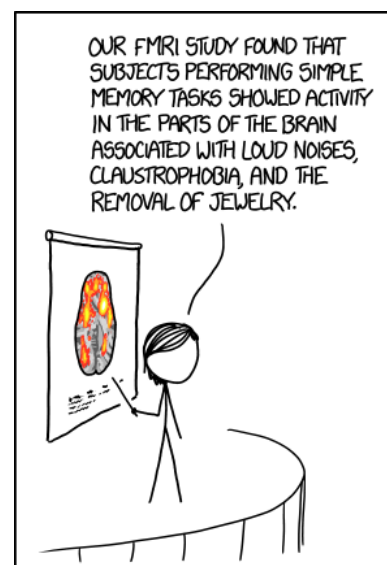


FIGURE 3.1: We found that subjects performing a simple attention task show brain activity associated with sticky hair, deadlines, PhD duties and worthwhile rewards. Image courtesy of xkcd, number 1453.

music genre and the melody/rhythmical pattern that is played. We obtain promising results comparable to those obtained on speech data in previous works and confirm that it is thus possible to correlate the human brain’s activity with musically relevant features of the attended source.

3.2 RELATED WORKS

EEG-based Auditory Attention Decoding aims at determining which sound source a person is “focusing on” by analysing the listener’s brain response. Most of the literature in the field focuses on decoding auditory attention to naturalistic speech in multi-speaker or noisy scenarios from the brain’s electric activity measured on the scalp [O’Sullivan et al. 2014; O’Sullivan et al. 2015]. Indeed, the topic is raising more and more interest thanks to the multitude of promising applications, especially concerning hearing aids and cochlear implants [Van Eyndhoven et al. 2017; Aroudi and Doclo 2020; Han et al. 2019; Das et al. 2020b; Pu et al. 2019; Ceolini et al. 2020].

First studies on **AAD** based on continuous **ECoG** [Mesgarani et al. 2009; Mesgarani and Chang 2012; Pasley et al. 2012] and **EEG** [O’Sullivan et al. 2014; O’Sullivan et al. 2015; Crosse et al. 2016] responses have shown that changes in the audio stimulus can be tracked in the neural activity. They evidenced how the attended source’s neural encoding is substantially stronger than the one of the other sources left in the mixture, allowing for a successful decoding of selective attention to a speaker. Similarly to Treder et al. [Treder et al. 2014], we recast the **AAD** problem in the music domain as one of decoding attention to a specific musical instrument playing in a musical ensemble.

The decoding procedure is usually two-fold [O’Sullivan et al. 2014]: firstly, a feature representation of the attended audio source is reconstructed from the neural response. Secondly, the reconstruction is correlated with the ground truth sources to determine the attended source. The stimulus reconstruction is referred to as the *backward* problem, as one goes from the brain response back to the stimulus. The mapping is usually done using linear models: a Multichannel Wiener Filter (**MWF**) maps the neural activity back to a stimulus feature representation [Lalor et al. 2009; Crosse et al. 2016]. Such a filter is known in the field as *backward model* and is estimated on a training set using a Minimum Mean Squared Error (**MMSE**) criterion [Crosse et al. 2016]. Therefore, assuming the system to be linear and time-invariant, the relation between stimulus and neural response can be described as a convolution where the impulse response is represented by the backward model [Crosse et al. 2016].

The majority of works studying auditory attention represented the speech by its broadband temporal envelope [Lalor and Foxe 2010; Fuglsang et al. 2017; O’Sullivan et al. 2014; O’Sullivan et al. 2015]. Others obtained promising results with speech spectrograms [Mesgarani et al. 2009; Mesgarani and Chang 2012; Pasley et al. 2012], phonemes [Di Liberto et al. 2015], or semantic features [Broderick et al. 2018].

The choice of the speech representation is critical as different features are supposed to map onto different hierarchical levels of brain processing [Di Liberto et al. 2015; Di Liberto et al. 2018]. Many studies suggest that speech

perception results from a hierarchical auditory system that processes attributes of the audio stimulus with an increasing level of complexity: earlier areas of the auditory system respond to low-level spectrotemporal and acoustic dynamics, while later areas to semantic and phonetic features of the stimulus [Okada et al. 2010; Peelle et al. 2010; Chang et al. 2010]. This hierarchical encoding of speech ensures that low-level descriptors of the audio stimulus, such as spectrotemporal and acoustic dynamics and high-level ones such as phonetic or semantic features, are reflected differently in the EEG. Something similar can be said for music, where low-level acoustic features such as the acoustic energy or the fundamental pitch are theorized to be encoded differently than high-level ones such as the musical structure [Di Liberto et al. 2020a].

However, when the aim is to perform **AAD** and not to conduct a neuroscientific study, the focus stays on spectrotemporal/acoustic descriptors which have been proven to be robust for that task. Here we compare multiple acoustic representations of the music stimulus, namely the broadband Amplitude Envelope (**AE**), the Magnitude Spectrogram (**MAG**) and the Mel Spectrogram (**MEL**).

3.3 METHODS

The goal is to determine the attended instrument in a single-trial fashion based on 24-second long **EEG** excerpts aligned to corresponding audio stimuli (of the same length). Our approach can be summarized in two steps and is similar to the one commonly used for decoding the attention to speech [Mesgarani et al. 2009; Pasley et al. 2012; Mesgarani and Chang 2012; O’sullivan et al. 2014; Crosse et al. 2016] and it is shown in **Figure 3.2**. First, an audio representation of the attended instrument is reconstructed from the single-trial **EEG** response of the subject exploiting a decoder previously trained on solos of that instrument. Second, given the isolated instrumental tracks, the attended instrument is recognized as the one that has the highest correlation with the reconstructed stimulus in terms of Pearson correlation coefficient (**PCC**).

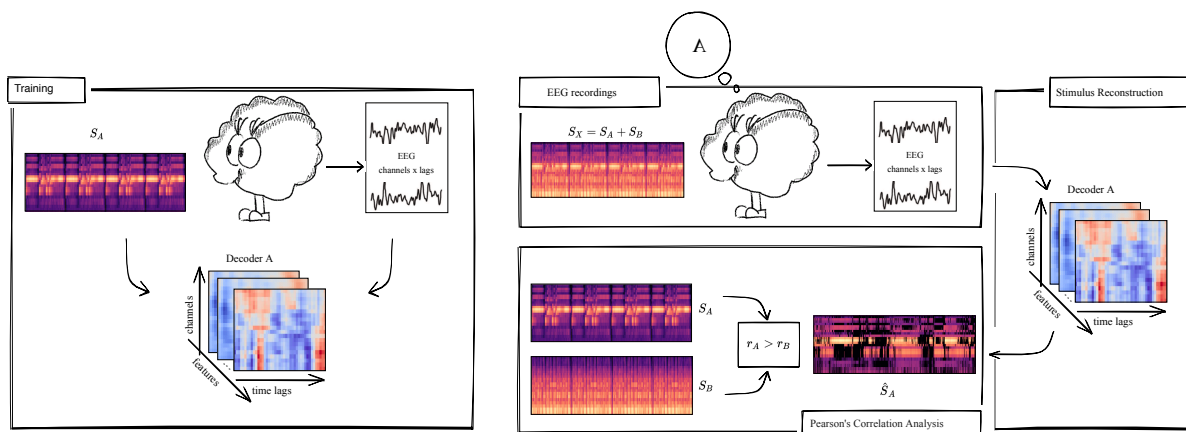


FIGURE 3.2: A subject-specific model is learned for each instrument from its solo and the **EEG** response collected while listening to it. Then, the same model is used to predict a representation of the attended instrument from the **EEG** response to a mixture containing that instrument. The attended instrument is the one that is mostly correlated with the reconstructed stimulus in terms of Pearson's correlation coefficient.

3.3.1 Audio Feature Extraction

Choosing the audio representation is a crucial point of **AAD**, as this choice includes a hypothesis about the neural coding of the stimulus and can significantly impact the reconstruction quality and the decoding performance. We studied three different audio representations, one in the time domain and two in the time-frequency (**TF**) domain: the time domain Amplitude Envelope (**AE**) computed using the Hilbert transform, the Magnitude Spectrogram (**MAG**), and the Mel Spectrogram (**MEL**), a perceptually-scaled representation commonly used for music analysis.

The **AE** is one of the most used audio descriptor for **AAD** with speech stimuli as the **EEG** was shown to track slowly varying changes in the audio stimulus [Golumbic et al. 2013; O’sullivan et al. 2014]. The assumption is that the **EEG** is linearly related to the broadband energy envelope of the stimulus. However, frequency modulations, *i.e.*, envelope fluctuations at specific frequencies, can give a more complete view of the audio signal. In fact, the spectrogram envelope of natural sounds fluctuates across both frequency and time, and this was shown, for instance, to be important for the intelligibility of speech [Pasley et al. 2012]. **TF** audio representations have already shown good performance for speech stimulus-reconstruction tasks [Mesgarani et al. 2009; Pasley et al. 2012; Mesgarani and Chang 2012]. A recent work explored auditory spectrograms modelling the peripheral auditory system [Akbari et al. 2019] as they may better model how the attended source is reflected in the **EEG**.

The same can be said for music, where the modulations’ complexity is much higher than in speech. In practice, the spectrogram can be seen as a time-varying representation of the amplitude envelope at each frequency bin [Pasley et al. 2012]. Thus, we will assume that the neural responses are linearly related to the spectrogram channels, seen as subband temporal envelopes.

3.3.2 Temporal Response Function

A feature representation of the attended source $\hat{\mathbf{S}} \in \mathbb{R}^{K \times N}$ where K is the number of features coefficients and T is the number of time samples, is reconstructed from the **EEG** using the backward model commonly used in the **AAD** framework [Crosse et al. 2016]. This filter can be seen as a *spatio-temporal decoder* which linearly maps the neural activity back to the audio feature representation, as a weighted sum of activity at each electrode in a given temporal context, as follows:

$$\hat{\mathbf{S}} = \mathbf{g}^T \mathbf{R}, \quad (3.1)$$

where

$$\mathbf{g} = [\text{flatten}(\mathbf{g}_1), \dots, \text{flatten}(\mathbf{g}_K)] \in \mathbb{R}^{CL \times K} \quad (3.2)$$

is a matrix composed by the column-wise concatenation of K multi-channel Wiener filters $\mathbf{g}_k \in \mathbb{R}^{C \times L}$ which are reshaped in row-major order into vectors of length CL . C represents the number of **EEG** channels and L the number of time lags, *i.e.*, the temporal context where we assume to see the **EEG** response to the stimulus. The time lags range between τ_{min} and τ_{max} and build the temporal context where we assume to see the **EEG** response

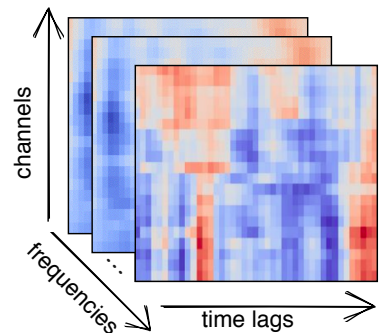


FIGURE 3.3: Visualization of a the spatio-temporal decoder reshaped as a tensor whose shape is given by the number of **EEG** channels C , the number of time lags L and the number of feature coefficients K of the audio representation we need to reconstruct. If the audio representation is the magnitude or Mel spectrogram, then the features will coincide with the frequency bins as displayed in the figure. For the broadband amplitude envelope, $K = 1$ and thus the tensor becomes a matrix.

to the stimulus as shown in Figure 3.4. An example of decoder for the MEL spectrogram is shown in Figure 3.3. Similarly, $\mathbf{R} \in \mathbb{R}^{CL \times N}$ is obtained as the row-wise concatenation of C lagged $L \times T$ time series matrices of the neural response recorded at electrode i . Such matrices are only padded with zeros on the left to ensure causality [O’sullivan et al. 2014].

$$\mathbf{R} = \begin{bmatrix} r_1(1) & r_1(2) & r_1(3) & \dots & \dots & \dots & r_1(N) \\ 0 & r_1(1) & r_1(2) & \dots & \dots & \dots & r_1(N-1) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & \dots & r_1(1) & \dots & r_1(N-L) \\ \hline \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \hline r_C(1) & r_C(2) & r_C(3) & \dots & \dots & \dots & r_C(N) \\ 0 & r_C(1) & r_C(2) & \dots & \dots & \dots & r_C(N-1) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & \dots & r_C(1) & \dots & r_C(N-L) \end{bmatrix} \quad (3.3)$$

In practice, each k -th feature coefficient of $\hat{\mathbf{S}}$ is reconstructed independently from the others using a multi-channel Wiener filter \mathbf{g}_k , which is learned through an MMSE criterion on a training set of solos of the same instrument. Each filter is estimated independently as the normalized reverse correlation:

$$\mathbf{g}_k = \mathbf{C}_{\mathbf{RR}}^{-1} \mathbf{C}_{\mathbf{RS}_k}, \quad (3.4)$$

where

$$\mathbf{C}_{\mathbf{RR}} = \mathbf{R}\mathbf{R}^T \quad (3.5)$$

is the estimated auto-correlation of the EEG data and

$$\mathbf{C}_{\mathbf{RS}_k} = \mathbf{R}\mathbf{S}_k^T \quad (3.6)$$

is the estimated cross-correlation of the stimulus and EEG data across all electrodes and time-lags for the k -th feature coefficient.

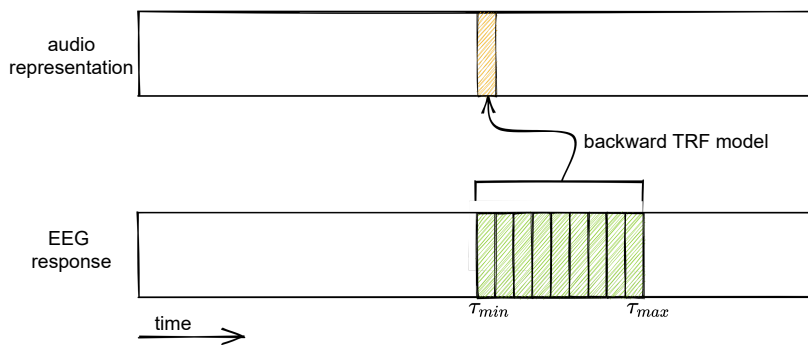


FIGURE 3.4: Schematic of the temporal context used by the backward model to reconstruct one frame of stimulus from the EEG data. For the sake of simplicity, we represented only one EEG channel.

3.3.3 Regularization

Since EEG signals are high-dimensional, autocorrelated, noisy data with high trial-to-trial variability, the estimate of the covariance matrices can be imprecise and subject to overfitting due to the high number of parameters to estimate [Blankertz et al. 2011]. Several methods have been proposed in the literature for regularization and prevent overfitting. Wong et al. nicely compared those techniques on a benchmark dataset for attention decoding to a target speaker [Wong et al. 2018].

We choose to use a shrinkage regularization to constrain the model coefficients by smoothly penalizing extreme eigenvalues [Blankertz et al. 2011]: the diagonal of the autocovariance matrix $\mathbf{C}_{\mathbf{RR}}$ is then weighted as follows:

$$\mathbf{C}'_{\mathbf{RR}} = (1 - \lambda)\mathbf{C}_{\mathbf{RR}} + \lambda\nu\mathbf{I}, \quad (3.7)$$

where \mathbf{I} is the identity matrix, ν is the average eigenvalue trace of $\mathbf{C}_{\mathbf{RR}}$, and $\lambda \in [0, 1]$ is the smoothing parameter.

3.4 EXPERIMENTS

3.4.1 Evaluation

We evaluate the reconstruction capabilities through the Pearson correlation coefficient (PCC) of the reconstructed stimulus representation with the attended instrument $r_{attended}$, the unattended instrument $r_{unattended}$ and the mixture $r_{mixture}$.

Besides the reconstruction capabilities, we also evaluate the decoding performance in terms of accuracy on the AAD task. Their statistical significance was assessed using an adaptation of the computationally-intensive randomization test [Noreen 1989], a non-parametric hypothesis test, comparing to chance, which does not make any assumption on the score distribution [Yeh 2000]. The considered significance levels are 5%, 1%, 0.1% and 0.01%, and the tests were performed over 10^4 iterations. For further explanations and details about the test, please refer to § 6.2.

3.4.2 Experimental Results

All audio representations were time aligned to the EEG responses acquired at 256Hz. Through a grid search over a set of reasonable values for each parameter ($\lambda \in [0.1, 1]$, $\tau_{max} \in [250, 500]$ ms, number of Mel bands $\in [12, 60]$), we found the best value for the shrinkage parameter to be $\lambda = 0.1$, for τ_{max} to be 250ms post stimulus, and for the number of Mel bands to be 24, using the following train/test splits: for each subject train on 14 solos, test on 40 duets and 24 trios.

DECODING PERFORMANCE In Table 3.1 one can see the decoding accuracy with respect to the three audio descriptors and number of instruments in the mixture. All the scores are significantly above the chance level, which is 50% for duets, around 33% for trios, and around 44% for all the test mixtures together. TF representations are clearly beneficial for the decoding indicating that envelope fluctuations at specific frequencies can give a complete view of the music audio signal. The two spectrograms, especially the MEL, also proved

Accuracy (%)	All	Duets	Trios
AE	52 ***	59 **	40*
MAG	75 ****	78 ****	69****
MEL	75 ****	76 ****	74 ****

TABLE 3.1: Decoding accuracy for different subsets of the test set. “****” denotes very high ($p < 0.0001$), “***” high ($p < 0.001$), “**” good ($p < 0.01$), “*” marginal ($p < 0.05$) and “n.s.” no ($p > 0.05$) statistical significance for a non-parametric randomization test.

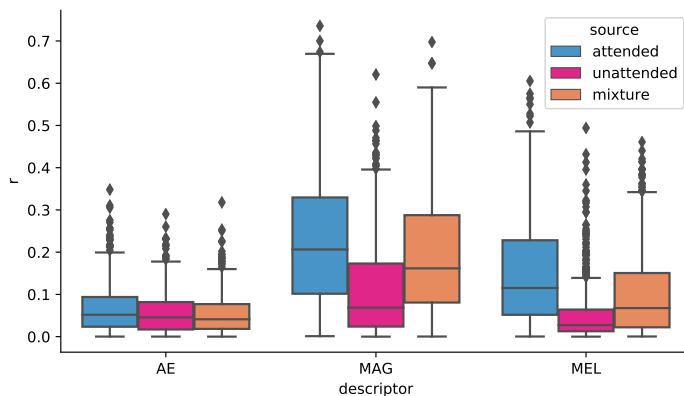


FIGURE 3.5: PCC of the reconstructed stimulus with the attended source (blue), the unattended one (pink) and the mixture (orange) for the three audio descriptors.

to be more robust to the mixture’s number of instruments. Nevertheless, even if the accuracy scores obtained with the AE are drastically below those obtained with the other two descriptors, they are still statistically significant.

CORRELATION ANALYSIS In Figure 3.5 one can see the PCCs of the reconstructed stimulus with the attended source (blue), the unattended one (pink) and the mixture (orange) for the three audio descriptors. In Figure 3.6 one can see the PCC coefficients of the reconstructed stimulus with the attended source contrasted with the one of respectively the mixture and the unattended source (only for duets). The correlation scores are very low, indicating that the reconstructions are highly deteriorated. Nevertheless, the “contrast” between $r_{attended}$ and $r_{unattended}$ is evident, especially for the two TF descriptors, confirming the decoding results of Table 3.1. Thus, the decoding seems to clearly benefit from the use of a finer audio representation, highlighting amplitude modulations in different frequency bands.

The lowest $r_{attended}$ Pearson’s coefficients are those related to the AE (median $r = 0.049$) but are still comparable to those obtained by O’Sullivan et al. in [O’Sullivan et al. 2014] for speech with the same audio descriptor (median $r = 0.054$). However, since the contrast between $r_{attended}$ and $r_{unattended}$ is only marginal, the decoding accuracy is much lower than the one obtained by the same authors. The broadband envelope is probably enough for discriminating between attended and unattended speakers but is not enough when dealing with music. Music present complex modulations both in time and frequency, for which the energy envelope is not enough representative.

Here the model is likely to account for effects more related to the whole mixtures than individual instruments, causing $r_{attended}$ and $r_{mixture}$ to be similar.

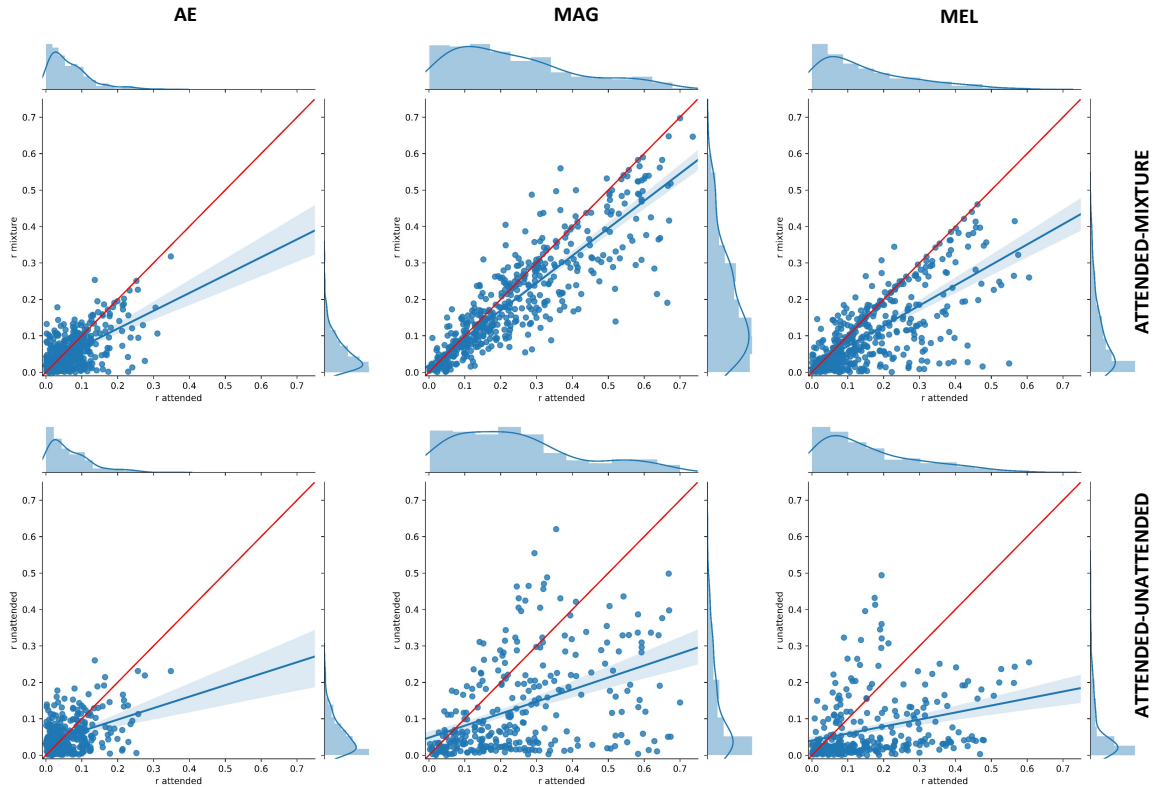


FIGURE 3.6: In the first row, $r_{attended}$ is plotted against $r_{mixture}$ while in the second row, $r_{attended}$ is plotted against $r_{unattended}$ (only duets) for all the audio descriptors. Data points below the red line $r_{attended} = r_{unattended}$ are classified correctly. Data points in the bottom-right corner are classified correctly with a large margin.

Moreover, we have to consider that in our case the stimuli were polyphonic and rendered through loudspeakers while in [O’sullivan et al. 2014] two concurrent speech stimuli were presented as monaural sources using headphones playing a different source to each ear. In general, when both $r_{attended}$ and $r_{unattended}$ are low and similar, the quality of the reconstructed stimulus is highly deteriorated making it hard to correctly decode the attended instrument. Even so, we can state that the $r_{attended}$ and $r_{unattended}$ distributions are statistically different ($p = 0.0042$ using a Wilcoxon test).

Also in the case of the linear spectrogram, the obtained correlations are comparable in terms of magnitude order to the ones obtained previously by [Mesgarani and Chang 2012] for speech in a different setting. From the same plot, we can observe that the correlations obtained with the **MAG** spectrogram are marginally higher than the ones obtained with the **MEL** one (median $r = 0.215$ for **MAG**, median $r = 0.119$ for **MEL**). However, the “contrast” between $r_{attended}$ and $r_{unattended}$ is higher for **MEL**, which is reflected in the decoding accuracy. The **MEL** spectrogram is a perceptually scaled and compact version of the linear spectrogram (**MAG**). A non-linear transformation of the frequency scale based on the perception of pitches (Mel scale) is applied to the linear spectrogram so that two pairs of frequencies that are equidistant in the Mel scale are perceived as being equidistant by humans. We observed that a lower number of features K , or **MEL** bands, is beneficial for the performance during the experiments. In particular, we tested values $\in [12, 60]$, and the results we show are relative to 24 Mel bands. Probably, the **MAG** representa-

	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 **

TABLE 3.2: F1 scores for different subsets of the test set: *all* for all the test mixtures, *duets* and *trios* for those containing respectively 2 or 3 instruments, *sm* and *dm* for those which exhibit respectively the same or a different melody/rhythmical pattern as the solo used to train the model, *mono* and *stereo* for those rendered respectively in mono or stereo. “***” denotes high ($p < 0.001$), “**” good ($p < 0.01$), and “n.s.” no ($p > 0.05$) statistical significance of the results.

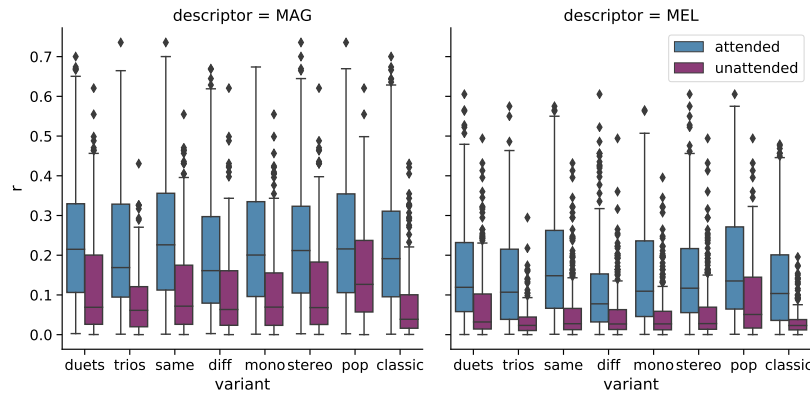


FIGURE 3.7: $r_{attended}$ and $r_{unattended}$ scores for all the stimuli variants. Only MAG and MEL descriptors are considered. $r_{attended}$ and $r_{unattended}$ distributions are significantly different for all the variants ($p < 0.001$, non-parametric Wilcoxon test).

tion has a too high number of features K , as it corresponds to the number of frequency bins (in our experiments 512), which might be too complex for the AAD task. Also for the TF descriptors, many misclassifications happen when the reconstructed stimulus quality is low, *i.e.*, when both $r_{attended}$ and $r_{unattended}$ coefficients are very low ($r < 0.2$) and close (see Figure 3.8). Here the model is accounting for effects which are probably more related to the whole mixtures than individual instruments. When $r_{attended}$ is high, usually the corresponding $r_{unattended}$ is low, meaning that the model is discriminating the two instruments.

NUMBER OF INSTRUMENTS As expected, *the number of instruments* in the mixtures seems to affect the performances, which are better for duets than trios as can be seen in Table 3.2. Some previous works on AAD applied to speech [Fuglsang et al. 2017; Das et al. 2018] showed that the attention task is more challenging for the listener with an increasing number of sources and noise levels. In practice, high noise levels can impact the listener’s ability to segregate the source of interest leading to poor decoding quality. We can assimilate a multi-instrumental musical piece to a particularly complex multi-speaker environment. The more instruments we have, the more difficult is the attention task. In music, this problem can also be related to how much the attended instrument is in the foreground, *i.e.*, to its predominance.

Nevertheless, also the results for trios are still statistically better than chance (considering that the chance levels are 50% for duets, 33% for trios). The MEL descriptor is particularly robust to this variant, both for the F1 scores and

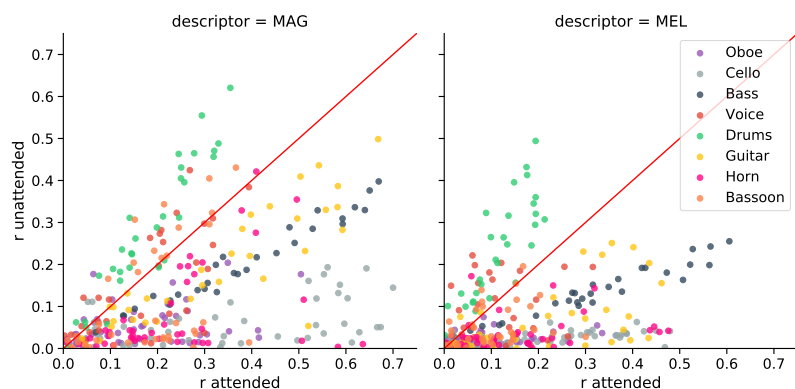


FIGURE 3.8: $r_{attended}$ is plotted against $r_{unattended}$ for each duet in the test set (only **MAG** and **MEL** descriptors). Data points below the red line $r_{attended} = r_{unattended}$ are classified correctly. Data points in the bottom-right corner are classified correctly with a large margin. The instruments are marked with different colors.

$r_{attended}$ (duets median $r = 0.12$, trios median $r = 0.11$ vs the one of the **MAG** which is $r = 0.22$, trios median $r = 0.17$).

SPATIAL RENDERING The stimuli were played to the subjects with two possible spatial renderings: one where both instruments are in the centre denoted as *mono* modality, and one where the instruments are spatialized, denoted as *stereo*. Intuitively, the stereo setting should help the subject in focusing on the target instrument as it makes it easier to localize it, leading to a better reconstruction of its features and finally giving a better decoding performance. However, it seems that the spatial rendering does not significantly affect neither the decoding performance nor the correlation values, with the differences not being statistically significant ($p > 0.05$, non-parametric Wilcoxon test). More data and experiments are needed to verify this hypothesis.

MUSIC GENRE *The genre*, instead, is highly influencing the performances. Both the **TF** descriptors behave much better for the Classical music mixtures compared to Pop ones as can be seen in [Table 3.2](#) and [Figure 3.7](#). This probably happens because the nature of the Pop excerpts used as stimuli is mostly repetitive musical patterns, which are essentially rhythmical. In our dataset, this is particularly true in mixtures with the drums and the bass, which usually have to guide the rhythm. The Classical mixtures used are inherently different: they exhibit long melodic lines which can be translated in well-defined varying pitch contours. Thus, the very good performances on the classical pieces can be explained by the fact that our model is tracking well the pitch/harmonic contour of the attended instrument. Usually, when one attends to an instrument one focuses on following the melody line or rhythm played.

That is why we tested if our models are invariant to *the melody/rhythmical pattern* that is played. In fact, the performance clearly changes when we test the models on different musical pieces from those which were used for training, and is better when the melody/rhythmical pattern remains the same. It is worth clarifying that even in this case, though the same solo excerpt is used during training and testing, during the latter, that solo excerpt is played as part of a mixture (duet or trio) and the **EEG** response is obviously completely different from that of the training with the solo-only stimulus. This performance degradation observed when the pitch contours vary between

training and testing is coherent with the explanation we gave before for the difference of performance among the genres. However, this also means that the generalization ability of the considered models is limited. Even if the models are not invariant to the changing pitch contour, the performance still remains significantly better than chance for the two TF representations. In this case, the linear spectrogram seems to be more robust than the Mel one. The lower performance on the Pop excerpts can be explained also by the fact that the drums are always misclassified as bass. Our tentative explanation is that when the subject is listening to the drums and the bass, the brain's activity is mostly tracking the rhythm. More experiments using recordings with clearer distinction between melody and rhythm will be needed to confirm these initial findings.

3.5 CONCLUSIONS

In this Chapter, we investigated for the first time the problem of AAD to a target instrument in polyphonic music based on the continuous EEG response. This study is critical within the thesis, as the proposed neuro-steered Music Source Separation approaches are built upon the results of this Chapter.

We conducted an extensive evaluation on the MAD-EEG dataset analysing the influence on the performance of multiple variants of musical stimuli, such as the number and type of instruments in the mixture, the spatial rendering, the music genre and the melody/rhythmical pattern that is played. We considered three different acoustic representations: the amplitude envelope, the magnitude and the MEL spectrograms.

Stimulus reconstruction based on a simple linear regression model yields promising results for decoding the attended instrument. Through experimental evaluation, we have shown that the EEG tracks musically relevant features which are highly correlated with the TF representation of the attended source and only weakly correlated with the unattended one making it possible to decode the auditory attention towards a specific instrument in the mixture. This contrast is particularly significant when using TF audio representations, highlighting amplitude modulations in different frequency bands. Among the two TF representations, the more compact and perceptually scaled representation given by the MEL spectrograms appears to be more robust to highlight the contrast. We have shown that we are tracking attention since these features are related to the attended source and not the mixture as a whole. However, it seems that the models are mostly tracking the instrument's pitch contour, which reduces its generalisation capabilities.

The main limitation is that this approach employs the separate "clean" sources of each instrument present in the mixture (to correlate their feature representation to the one predicted with the EEG data). This condition is never met in realistic music listening scenarios where only the mixtures are available. Moreover, the linear model is not tracking all the non-linearity of the EEG signals.

Part III

NEURO-STEERED MUSIC SOURCE SEPARATION

4	c-nmf: NEURO-STEERED MUSIC SOURCE SEPARATION WITH EEG-BASED AUDITORY ATTENTION DECODING AND CONTRASTIVE-NMF	
4.1	Introduction	43
4.2	Related works	44
4.3	Methods	46
4.3.1	NMF-based audio source separation	46
4.3.2	A novel NMF variant: Contrastive-NMF (C-NMF)	48
4.4	Experiments	52
4.4.1	Evaluation	53
4.4.2	Experimental results	53
4.5	Conclusions	58
5	ugosa: USER-GUIDED ONE-SHOT DEEP MODEL ADAPTATION FOR MUSIC SOURCE SEPARATION	
5.1	Introduction	61
5.2	Related works	63
5.3	Methods	64
5.3.1	Proposed adaptation loss	64
5.3.2	Model	65
5.4	Experiments	66
5.4.1	Experiment with manually annotated activations	67
5.4.2	Experiment with EEG-derived activations	71
5.5	Conclusions	77

4

C-NMF: Neuro-steered music source separation with EEG-based auditory attention decoding and contrastive-NMF

- **SYNOPSIS** This Chapter introduces the central contribution of the thesis, a *neuro-steered music source separation* framework built upon the results of the previous Chapter and conducts an extensive evaluation of the proposed system on the MAD-EEG dataset. Specifically, we leverage the fact that the attended instrument’s neural encoding is substantially stronger than the one of the unattended sources left in the mixture to inform a source separation model based on a new variant of **NMF** named Contrastive-NMF and automatically separate the attended source. This unsupervised **NMF** variant is particularly advantageous as it allows us to incorporate additional information in a principled optimisation fashion and does not need training data, which is particularly difficult to acquire for applications involving **EEG** recording. This work led to the following conference publication and preprint:

- Cantisani, Giorgia, Slim Essid, and Gaël Richard (2021b). “Neuro-steered music source separation with EEG-based auditory attention decoding and contrastive-NMF”. in: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*
- Cantisani, Giorgia, Slim Essid, and Gaël Richard (2021a). “EEG-based Decoding of Auditory Attention to a Target Instrument for Neuro-steered Music Source Separation”. In: *journal in preparation*

4.1 INTRODUCTION

In this Chapter, we propose an unsupervised Nonnegative Matrix Factorization (**NMF**) variant, named Contrastive-NMF (**C-NMF**), that separates a target instrument, guided by the user’s selective auditory attention to that instrument, which is tracked in his/her **EEG** response to music. Specifically, we exploited the “contrast” among the sources that can be extracted from the neural response using a decoding model.

From the experiments presented in **Chapter 3**, we know that the reconstruction of the audio modulations we can get from the **EEG** is more correlated with

Keywords: Audio source separation, Polyphonic music, EEG, Matrix factorisation, Multimodal processing.

Resources:

- 🔗 [Paper](#)
- 🔗 [Code](#)
- 🔗 [Demo](#)



FIGURE 4.1: **WHo** doesn’t? Image courtesy of xkcd, number 2506.

those of the attended instrument than with those of the unattended one. We observed that this reconstruction is highly deteriorated but still “good enough” to discriminate between the attended and unattended sources. These two facts can be naturally exploited in an informed NMF-based sound source separation system, where the sources are decomposed into spectral patterns and corresponding activations.

Our proposal is then to reconstruct the attended source’s activations from the EEG using the backward model introduced in Chapter 3. Indeed, the NMF activations can be seen as modulations across time of specific spectral patterns found by the factorisation. Thus, they will represent a rough approximation of the TF representations used in the experiments in Chapter 3.

One advantage over other source separation models is that NMF allows to incorporate additional information about the sources directly in its optimisation cost without requiring a data-intensive training phase. The additional information at our disposal is represented by the attended source’s temporal activations for a given set of spectral patterns representing that source reconstructed from the EEG. Since those reconstructed activations are significantly deteriorated, it is hard to use them directly. Nevertheless, these reconstructions are good enough to discriminate the attended instrument from the unattended one. In the proposed C-NMF, this “contrast” is used to guide the separation. The factorisation and the decoding are learnt jointly, allowing for adapting both models to the specific test mixture and leading to encouraging results.

The main advantage of the C-NMF formulation is that it allows us to reformulate the AAD problem without access to the ground truth sources, paving the way for real-life applications. The attended instrument is the one that is automatically separated by the separation system thanks to the contrast.

We conduct an extensive evaluation of the proposed system on the MAD-EEG dataset analysing the impact of multiple aspects of the musical stimuli, such as the number and type of instruments in the mixture, the spatial rendering and the music genre, obtaining encouraging results, especially in difficult cases where non-informed models struggle.

4.2 RELATED WORKS

The AAD task is naturally related to audio source separation. As previously explained in Chapter 3, the decoding paradigm requires access to the ground truth sources, to correlate them to the neural data. However, this situation is never met in realistic scenarios such as hearing aids and cochlear implants, where only the mixture of the sound scene recorded by their microphones is available. In such scenarios, an additional *audio source separation* step is needed to extract the reference sources needed for the decoding. 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.

Most of the studies that relate speech source enhancement and AAD have worked in this direction. Many of them focused on the multi-channel audio scenario using beamforming [Aroudi et al. 2018; Aroudi and Doclo 2019; Aroudi and Doclo 2020] and multi-channel Wiener filtering [Van Eyndhoven

et al. 2017; Das et al. 2017; Das et al. 2020b] as hearing aids can be equipped with a microphone array. Both techniques estimate spatial filters that return the target speech when applied to the mixture while suppressing the background noise and interfering sources. These approaches use spatial information such as the directions of arrival and the target activity to compute the second-order statistics of the noise and interferers. One of the main limitations lies in estimating the spatial location and the voice activity, which may be difficult in challenging scenarios (e.g., overlapping speakers in space, time or frequency, high reverberation, moving speakers).

Other works focus on the single-channel scenario using DL-based approaches. O’Sullivan et al. were the first along this line [O’Sullivan et al. 2017]. However, their model requires prior training on the target speakers, which is a substantial limitation in real scenarios. The problem is tackled by Han et al. with a speaker-independent source separation system able to generalize to unseen speakers [Han et al. 2019]. Such a system relies on a deep attractor network, which projects the mixture’s time–frequency representation in a high-dimensional space where the speakers are separable [Chen et al. 2017; Luo et al. 2018]. The main difference with the deep clustering (DC) approach [Luo et al. 2017], is that the DNN is trained end-to-end to estimate a mask, while in DC a post-clustering step on the embedding is required, giving an advantage in terms of separation performance. Ceolini et al., instead, informed a speech separation neural network with the decoded attended speech envelope, leading to the extraction of the attended source [Ceolini et al. 2020]. However, the training of the source separation model and that of the AAD model are still decoupled, due to the lack of large datasets collected for AAD.

In general, performing the source separation and AAD steps independently is sub-optimal. In their work Pu et al. propose a unified model for joint AAD and binaural beamforming [Pu et al. 2019]. An adaptive beamformer is learned thanks to an objective which minimizes noise and interference but, at the same time, controls the target speaker distortion and maximizes the Pearson correlation coefficient (PCC) between the envelope of the beamformer output and the decoded EEG. In a later work [Pu et al. 2020], the same authors showed that their algorithm is robust to attention switching, which can be tracked in real-time thanks to the joint approach.

In this work, we pursued the joint approach and propose to adapt an NMF-based source separation model to a specific mixture using a weak signal decoded from the EEG using an AAD model. The AAD model is not fixed and is as well updated during the optimization. Our work differs from those by Pu et al. [Pu et al. 2019] as our aim is not to maximize the PCC between the envelope of the beamformer output and the decoded EEG. Since the decoded output can be significantly deteriorated (see Chapter 3), we leverage instead the fact that the attended instrument’s neural encoding is substantially stronger than the one of the unattended sources left in the mixture. This “contrast” is maximized when solving our separation model estimation problem.

4.3 METHODS

The goal is to separate a target instrument from a given music mixture. Along with the audio signal, we have access to the EEG recorded while the subject was listening to the given mixture and attending to the target instrument. From this signal we can reconstruct the attended source’s activations from the EEG using a backward model. Those reconstructed activations are significantly deteriorated but create a “contrast” that allows to discriminate the attended instrument from the unattended one.

In the proposed C-NMF, this “contrast” is used to guide the separation. The factorisation and the decoding are learnt jointly. The target instrument’s activations are reconstructed from the multi-channel EEG at first using a pre-trained backward model. Then they are used to guide the mixture’s factorisation and cluster the components into the respective sources. At the same time, the decoding model is updated every certain number of NMF iterations to adapt to the observed signal. A good initialisation of the decoder can be learned from a small training set of solos and corresponding EEG recordings from the same subject.

4.3.1 NMF-based audio source separation

The proposed Contrastive-NMF (C-NMF) is a novel variant of *Nonnegative Matrix Factorization*, a technique for data decomposition which has been very popular in many audio inverse problems such as source separation, enhancement or transcription as it is able to unmix superimposed spectral components [Févotte et al. 2018]. Among other factorization techniques (e.g., Principal Component Analysis (PCA), Independent Component Analysis (ICA)), NMF distinguishes itself through its nonnegativity constraints which lead to a part-based representation of the data that is *interpretable* [Lee and Seung 1999]. In Figure 4.2 one can see an example of NMF decomposition of a single source audio signal.

In the case of single-channel audio source separation, one can assume that an audio signal $x(t)$ at time sample t is given by the linear mixture of J sources $s_j(t)$:

$$x(t) = \sum_{j=1}^J s_j(t). \quad (4.1)$$

Observing the mixture $x(t)$, a source separation system aims to recover one or more sources $s_j(t)$ of interest. Such a mixture can be represented in matrix form through its magnitude spectrogram $\mathbf{X} \in \mathbb{R}_+^{M \times N}$, where M represents the number of frequency bins and N the number of Short Time Fourier Transform (STFT) frames.

\mathbf{X} can be factorized into two unknown matrices \mathbf{W} and \mathbf{H} such that $\mathbf{X} \approx \mathbf{WH}$, where the columns of $\mathbf{W} \in \mathbb{R}_+^{M \times K}$ are interpreted as non-negative audio spectral patterns, expected to correspond to different sources and the rows of $\mathbf{H} \in \mathbb{R}_+^{K \times N}$ as their temporal activations. Usually, one refers to \mathbf{W} as the *dictionary* and to \mathbf{H} as the *activation matrix*. When K , namely the rank of the factorization, is much smaller than M , \mathbf{WH} represents a *low-rank*

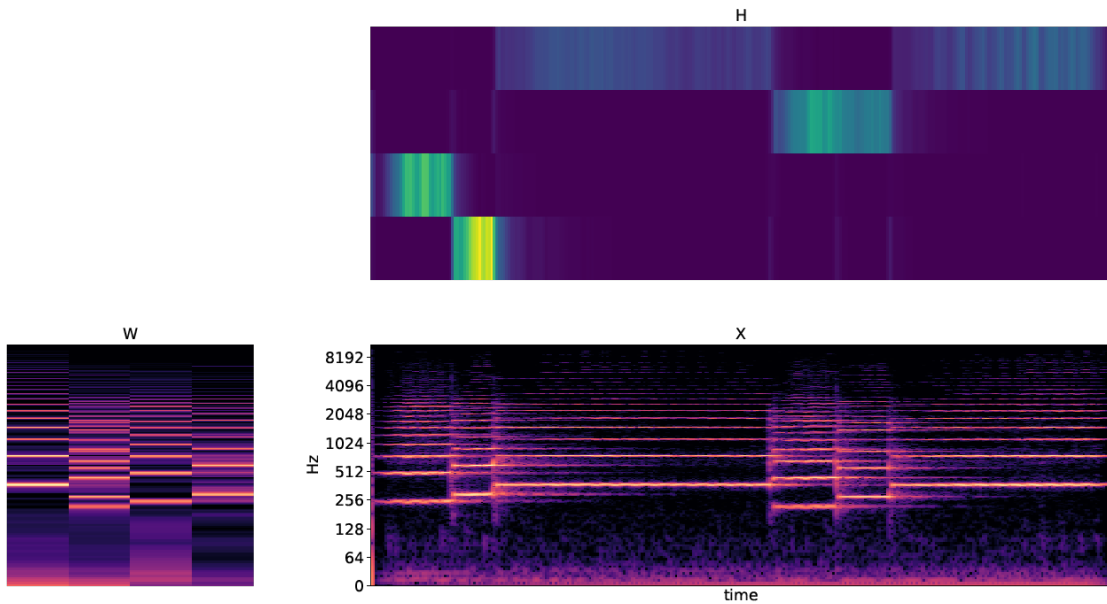


FIGURE 4.2: Example of NMF decomposition of the magnitude spectrogram of a single-source audio recording. The time-frequency matrix \mathbf{X} is approximated as a product of two non-negative matrices \mathbf{W} and \mathbf{H} having a much smaller rank. \mathbf{W} can be seen as a dictionary of spectral components representing elementary sound units (notes, chords, percussive sounds) and \mathbf{H} as their activations.

approximation of the data matrix \mathbf{X} [Févotte et al. 2018]. The factorisation can be achieved by minimizing a cost function as the following:

$$\begin{cases} C(\mathbf{W}, \mathbf{H}) = \underbrace{D(\mathbf{X}|\mathbf{WH})}_{\text{audio factorization}} + \underbrace{\mu\|\mathbf{H}\|_1 + \beta\|\mathbf{W}\|_1}_{\text{sparsity}} \\ \mathbf{W}, \mathbf{H} \geq 0. \end{cases} \quad (4.2)$$

Usually, for the mixture reconstruction β -divergences are used, which have been very popular for audio inverse problems. It is also common to impose a sparsity constrain on both \mathbf{W} and \mathbf{H} using an ℓ_1 regularization controlled by the hyperparameters μ and β , respectively, to improve the source modelling. In fact, music is often given by a repetition of a few audio patterns, thus we can easily assume that the activations are sparse [Vincent et al. 2018]. The same can be assumed for the spectral patterns as there is only a low probability that two given sources are highly activated in the same set of frequency bins [Yilmaz and Rickard 2004].

At this point, the separation problem reduces to the assignment of each NMF component to the corresponding source j . Then, the complex-valued spectrogram \mathbf{S}_j of each source can be estimated by Wiener filtering as [Févotte et al. 2018]:

$$\tilde{\mathbf{S}}_j = \frac{\mathbf{W}_j \mathbf{H}_j}{\mathbf{WH}} \otimes \tilde{\mathbf{X}}, \quad (4.3)$$

where the element-wise division $(\mathbf{W}_j \mathbf{H}_j)/(\mathbf{WH})$ is the soft mask associated to source j and $\tilde{\mathbf{X}}$ is the complex spectrogram of the mixture. \otimes denotes an element-wise multiplication.

Through an Inverse Short Time Fourier Transform (iSTFT) one can recover the corresponding audio signal in the time domain. For a schematic of the NMF-based separation pipeline, the reader can refer to Figure 4.3.

What we have described so far is the so-called *unsupervised NMF*, i.e., a blind signal decomposition where both the dictionary and the activations are estimated from the mixture [Févotte et al. 2018]. However, in real music compositions a source plays several notes with different pitches and it might be hard to represent it with a single component. Moreover, two sources may be represented by similar components as they might overlap and be highly correlated. Therefore, the component assignment might be hard and requires specific classification or clustering techniques to group together components associated to the same source. In such a complex situation, the factorization needs to be “guided” by incorporating prior information about the sources to return a meaningful representation [Vincent et al. 2014].

Starting from the unsupervised formulation, one can incorporate prior knowledge directly in the optimisation cost, e.g., through hard or soft constraints, specific regularizers, pretrained dictionaries, or forcing the elements of \mathbf{W} and/or \mathbf{H} to follow a given distribution [Vincent et al. 2018]. For example, in the case of music, it is possible to impose properties like harmonicity of the spectral patterns or smoothness and sparsity to the activations [Vincent et al. 2018]. Particularly interesting is the multimodal scenario, where one has access to multiple views of the same phenomenon (e.g., video, motion capture data, score) which are synchronized with the audio. Seichepine et al. [Seichepine et al. 2014], for instance, propose to impose the equality (hard constraint) or the similarity (soft constraint) of the source activations in the two modalities. This is not applicable in our case as the time activations we can reconstruct from the EEG are very deteriorated, making it hard to use them directly. Nevertheless, these reconstructions are “good enough” to discriminate the attended instrument from the unattended one, leading to a “contrast” that can guide the separation.

4.3.2 A novel NMF variant: Contrastive-NMF (C-NMF)

The general idea of discriminating sources according to some criterion for NMF-based audio source separation was already explored in the past but most of the proposals refer to fully supervised or semi-supervised scenarios, where the basis functions are learned in a training phase. Weninger et al. and Kitamura et al. propose to learn basis matrices that are as much discriminative as possible to have unique spectral templates for each source [Weninger et al. 2014; Kitamura et al. 2016]. Grais and Erdogan propose to minimize the cross-coherence between dictionaries belonging to different sources [Grais and Erdogan 2013], while Chung et al. to learn a factorization so that each basis is classified into one source [Chung et al. 2016].

Kumar et al., in a different application setting, propose a max-margin framework, where the projections are learned to maximize an Support Vector Machine (SVM) classifier’s discriminative ability [Kumar et al. 2012].

Within this work, instead, the projections are learned by an unsupervised NMF to maximize the discrimination ability of a decoding model. Specifically, the proposed cost aims at decomposing the audio spectrogram while maximizing

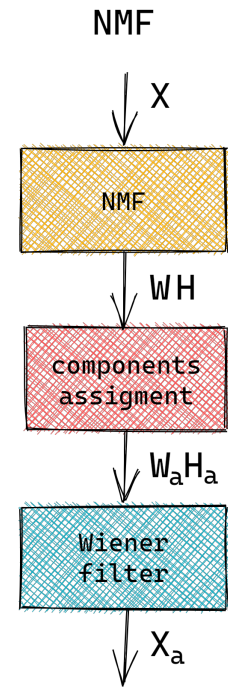


FIGURE 4.3: NMF-based source separation pipeline: first the magnitude or power spectrogram is decomposed in meaningful spectral components and corresponding time activations. At this point, the separation problem reduces to the assignment of each component to the corresponding source which can then be recovered through a Wiener Filter (WF) where the NMF representation is used as source variance model.

the similarity of the EEG-derived activations with the audio-derived ones for the target source and minimizing it for the interference sources. Thanks to this formulation, the components resulting from the decomposition should already be clustered into the target and interference sources.

Let us analyze the novel cost function. Considering a mixture $x(t)$ given by the linear mixing of the attended source $s_a(t)$ and some interferers $s_u(t)$, let $\mathbf{W}_a \in \mathbb{R}_+^{M \times K_a}$ be a sub-dictionary of \mathbf{W} containing a set of basis vectors representing source $s_a(t)$ and $\mathbf{H}_a \in \mathbb{R}_+^{K_a \times N}$ be their activations. \mathbf{H}_a can be roughly approximated by $\mathbf{S}_a \in \mathbb{R}_+^{K_a \times N}$ reconstructed from the time-lagged EEG response \mathbf{R} ,¹ the assumption being that it is likely to be more correlated with the NMF-derived activations of the attended source \mathbf{H}_a than with the ones of the interferers \mathbf{H}_u . This contrast can be integrated in the unsupervised NMF cost function as follows:

$$\left\{ \begin{array}{l} C(\mathbf{W}, \mathbf{H}) = \underbrace{D_{KL}(\mathbf{X}|\mathbf{W}\mathbf{H})}_{\text{audio factorization}} + \underbrace{\mu\|\mathbf{H}\|_1 + \beta\|\mathbf{W}\|_1}_{\text{sparsity}} + \\ \quad - \underbrace{\delta(\|\mathbf{H}_a\mathbf{S}_a^T\|_F^2 - \|\mathbf{H}_u\mathbf{S}_a^T\|_F^2)}_{\text{contrast}} \\ \mathbf{W}, \mathbf{H}, \mathbf{S}_a \geq 0 \\ \|\mathbf{h}_{k\cdot}\|_2 = 1, \|\mathbf{s}_{k\cdot}\|_2 = 1. \end{array} \right. \quad (4.4)$$

where $D_{KL}(\cdot|\cdot)$ is the Kullback-Leibler divergence, μ and β are regularization parameters and δ is a parameter weighting the contrast term. $\mathbf{h}_{k\cdot}$ and $\mathbf{s}_{k\cdot}$ represent the rows of \mathbf{H} and \mathbf{S}_a respectively and are normalized to have unit ℓ_2 norm in order to minimize the effect of a scale mismatch between the modalities.

We derived the update rules for \mathbf{H} and \mathbf{W} using the Multiplicative Updates (MUs) heuristic, which is based on gradient descent [Févotte and Idier 2011]. A solution is searched by moving in the direction opposite to the gradient's: \mathbf{W} and \mathbf{H} are updated alternately according to a scheme called *block-coordinate descent*: each variable is updated assuming the other to be constant.

The learning rate is adaptively chosen so as to have multiplicative updates, which cannot generate negative elements when starting from positive values [Lee and Seung 2001]. The same algorithm can be derived using the heuristic proposed Févotte and Idier, which consists in computing the gradient of the cost $\nabla C(\theta)$, splitting it into its negative and positive parts, *i.e.*, writing $\nabla C(\theta) = \nabla_{\theta^+} C(\theta) - \nabla_{\theta^-} C(\theta)$, and building the rules as follows [Févotte and Idier 2011]:

$$\theta \leftarrow \theta \otimes \frac{\nabla_{\theta^-} C(\theta)}{\nabla_{\theta^+} C(\theta)} \quad (4.5)$$

With $\theta = \{\mathbf{W}, \mathbf{H}\}$, the update rules can be computed as:

$$\mathbf{W} \leftarrow \mathbf{W} \otimes \frac{\nabla_{\mathbf{W}^-} C(\mathbf{W}, \mathbf{H})}{\nabla_{\mathbf{W}^+} C(\mathbf{W}, \mathbf{H})} \quad (4.6)$$

$$\mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\nabla_{\mathbf{H}^-} C(\mathbf{W}, \mathbf{H})}{\nabla_{\mathbf{H}^+} C(\mathbf{W}, \mathbf{H})} \quad (4.7)$$

The cost function in Eq. (4.4) is completely separable, therefore, one can compute the gradient for the Kullback-Leibler divergence, the sparsity con-

¹There is no constraint on the backward model such that the reconstructed activations \mathbf{S}_a are non-negative because the assumption behind a linear regression model is that the output variable follows a Gaussian distribution. To ensure that the reconstruction is non-negative, a generalized regression model should be used, where the output variable is constrained to follow an inverse-Gaussian or Gamma distribution. In our case, we observed that the negative values in the reconstructions had small amplitude and were similar to noise, thus we set them directly to zero.

straints and the contrast term separately. The derivation of the update rule for \mathbf{W} is trivial because it does not involve the contrast term:²

$$\mathbf{W} \leftarrow \mathbf{W} \otimes \frac{(\Lambda^{-1} \otimes \mathbf{X})\mathbf{H}^T}{\mathbf{1}\mathbf{H}^T + \beta} \quad (4.8)$$

where \otimes , divisions and exponents denote element-wise operations, $\mathbf{1}$ is a matrix of ones whose size is given by context and $\Lambda = \mathbf{W}\mathbf{H}$. The update rule for \mathbf{H} requires more attention. In fact, in the contrast term we have the two matrices \mathbf{H}_a and \mathbf{H}_u which are respectively the activations of the attended and interference sources. Thus, the gradient with respect to \mathbf{H} of the contrast term, will be equal to the gradient computed with respect to \mathbf{H}_a for the first K_a rows and equal to the gradient computed with respect to \mathbf{H}_u for the remaining rows:

$$\nabla_{\mathbf{H}}(-\delta(\|\mathbf{H}_a\mathbf{S}_a^T\|_F^2 - \|\mathbf{H}_u\mathbf{S}_a^T\|_F^2)) = \begin{cases} -2\delta\mathbf{H}_a\mathbf{S}_a^T\mathbf{S}_a, & \text{if } 1 < k < K_a \\ +2\delta\mathbf{H}_u\mathbf{S}_a^T\mathbf{S}_a, & \text{if } K_a + 1 < k < K \end{cases} \quad (4.9)$$

leading to the following update rule:

$$\mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\mathbf{W}^T(\mathbf{X} \otimes \Lambda^{-1}) + \delta\mathbf{P}^-}{\mathbf{W}^T\mathbf{1} + \mu + \delta\mathbf{P}^+} \quad (4.10)$$

where \mathbf{P}^- , $\mathbf{P}^+ \in \mathbb{R}_+^{K \times N}$ are auxiliary matrices defined as:

$$\mathbf{P}^- = \begin{cases} \mathbf{H}_a\mathbf{S}_a^T\mathbf{S}_a, & \text{if } 1 < k < K_a \\ 0, & \text{if } K_a + 1 < k < K \end{cases} \quad (4.11)$$

$$\mathbf{P}^+ = \begin{cases} 0, & \text{if } 1 < k < K_a \\ \mathbf{H}_u\mathbf{S}_a^T\mathbf{S}_a, & \text{if } K_a + 1 < k < K. \end{cases} \quad (4.12)$$

The derived update rules for \mathbf{H} and \mathbf{W} are given in lines (10) and (12) of [Algorithm 1](#) respectively. This pseudo-code provides all the details of the algorithm, including also the update of the decodi model \mathbf{g} . In fact, the factorization and the decoding are learnt jointly to improve the source modelling for both the source separation and [AAD](#) tasks.

Specifically, the target instrument's activations \mathbf{S}_a are first reconstructed from the time-lagged [EEG](#) data matrix \mathbf{R} using a pre-trained backward model \mathbf{g} . Then those activations are used to guide the mixture's factorization and cluster the components into the respective sources obtaining two submatrices \mathbf{W}_a and \mathbf{H}_a associated with the attended source. At the same time, the decoding model \mathbf{R} is updated every certain number of [NMF](#) iterations to adapt to the observed signal using \mathbf{W}_a as a new feature extractor. After convergence, the dictionary and the activations related to the attended source are used to obtain the Wiener filter mask. The complete pipeline is depicted in [Figure 4.4](#).

²For the detailed derivation, please refer to [§ 6.2](#) in the appendices.

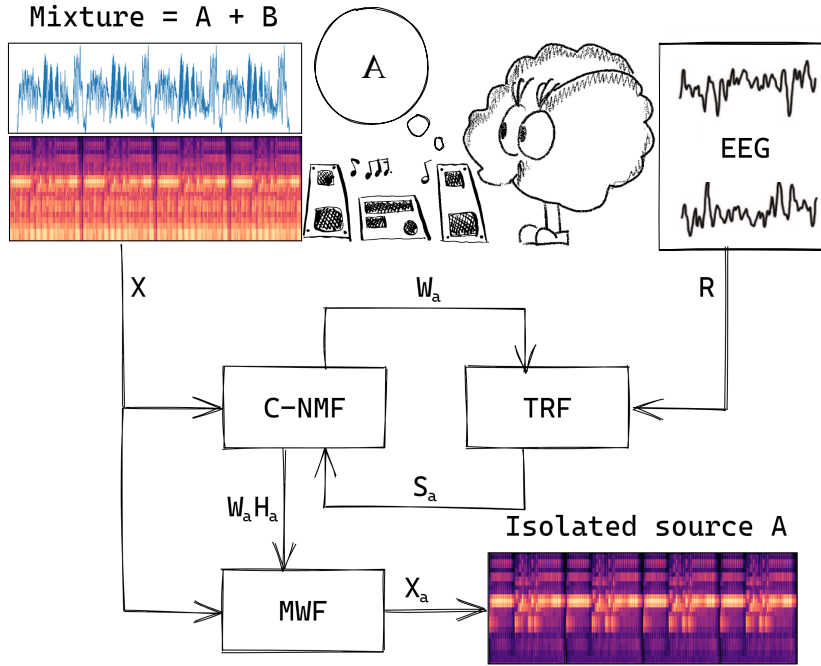


FIGURE 4.4: Proposed scheme: the target instrument’s activations are reconstructed from the listener’s multi-channel EEG using a pre-trained backward model. They are then used to guide the mixture’s factorisation and cluster the components into the respective sources (C-NMF). At the same time, the decoding model is updated every certain number of C-NMF iterations to adapt to the observed signal. After convergence, the dictionary and the activations related to the attended source are used to obtain the WF soft-mask.

Algorithm 1: Contrastive NMF pseudo-code

input : $\mathbf{X}, \mathbf{R}, \mu \geq 0, \beta \geq 0, \delta \geq 0, \gamma \in [0, 1]$
output : $\mathbf{W}_a, \mathbf{H}_a$

- 1 $\mathbf{W}, \mathbf{H}, \mathbf{g}$ initialization
- 2 $\mathbf{H} \leftarrow \text{diag}(\|\mathbf{h}_{1:}\|^{-1}, \dots, \|\mathbf{h}_{K:}\|^{-1})\mathbf{H}$ ▷ normalization
- 3 $\mathbf{W} \leftarrow \mathbf{W} \text{diag}(\|\mathbf{h}_{1:}\|, \dots, \|\mathbf{h}_{K:}\|)$ ▷ re-scaling
- 4 $\Lambda = \mathbf{W}\mathbf{H}$
- 5 **repeat**
- 6 $\mathbf{S}_a \leftarrow \mathbf{g}^T \mathbf{R}$
- 7 $\mathbf{S}_a \leftarrow \text{diag}(\|\mathbf{s}_{1:}\|^{-1}, \dots, \|\mathbf{s}_{K:}\|^{-1})\mathbf{S}_a$
- 8 **repeat**
- 9 $\mathbf{P} \leftarrow [-\mathbf{H}_a \mathbf{S}_a^T \mathbf{S}_a, \mathbf{H}_u \mathbf{S}_a^T \mathbf{S}_a]^T$
- 10 $\mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\mathbf{W}^T (\mathbf{X} \otimes \Lambda^{-1}) + \delta \mathbf{P}^-}{\mathbf{W}^T \mathbf{1} + \mu + \delta \mathbf{P}^+}$
- 11 $\mathbf{H} \leftarrow \text{diag}(\|\mathbf{h}_{1:}\|^{-1}, \dots, \|\mathbf{h}_{K:}\|^{-1})\mathbf{H}$
- 12 $\mathbf{W} \leftarrow \mathbf{W} \text{diag}(\|\mathbf{h}_{1:}\|, \dots, \|\mathbf{h}_{K:}\|)$
- 13 $\Lambda = \mathbf{W}\mathbf{H}$
- 14 $\mathbf{W} \leftarrow \mathbf{W} \otimes \frac{(\Lambda^{-1} \otimes \mathbf{X}) \mathbf{H}^T}{\mathbf{1} \mathbf{H}^T + \beta}$
- 15 $\Lambda = \mathbf{W}\mathbf{H}$
- 16 **until convergence;**
- 17 update \mathbf{g}
- 18 **until convergence;**
- 19 **return** $\mathbf{W}_a, \mathbf{H}_a$

4.4 EXPERIMENTS

The experiments are designed to evaluate if the EEG information helps the separation process. However, to verify that the improvement is due to the EEG and not to the cost function’s discriminative capacity, it was not enough to have the blind NMF as the only baseline. Therefore, we built a second baseline which consists of the C-NMF to which meaningless side information is given. The meaningless side information consists of random activations sampled from a Gaussian distribution. To summarise, we tested three models:

1. Blind NMF (NMF);
2. Contrastive NMF + Random side activations (C-NMF-r);
3. Contrastive NMF + EEG-derived activations (C-NMF-e).

As the models are entirely unsupervised, the factorised components need to be assigned to each source before applying the multi-channel Wiener filter. In the two baselines, the components are clustered according to their Mel-frequency cepstral coefficients (MFCCs) similarity. The different pipelines are depicted in Figure 4.5.

In the case of the C-NMF-e, the EEG information automatically identifies and gathers the target instrument components. Thanks to this we can reformulate the AAD problem exposed in Chapter 3, where we had access to the ground truth sources, differently. This time, the instrument which is predicted as being the attended one is the one that is automatically separated by the proposed source separation system. Specifically for our formulation, the attended instrument is the one represented by the \mathbf{W}_a dictionary and \mathbf{H}_a activations.

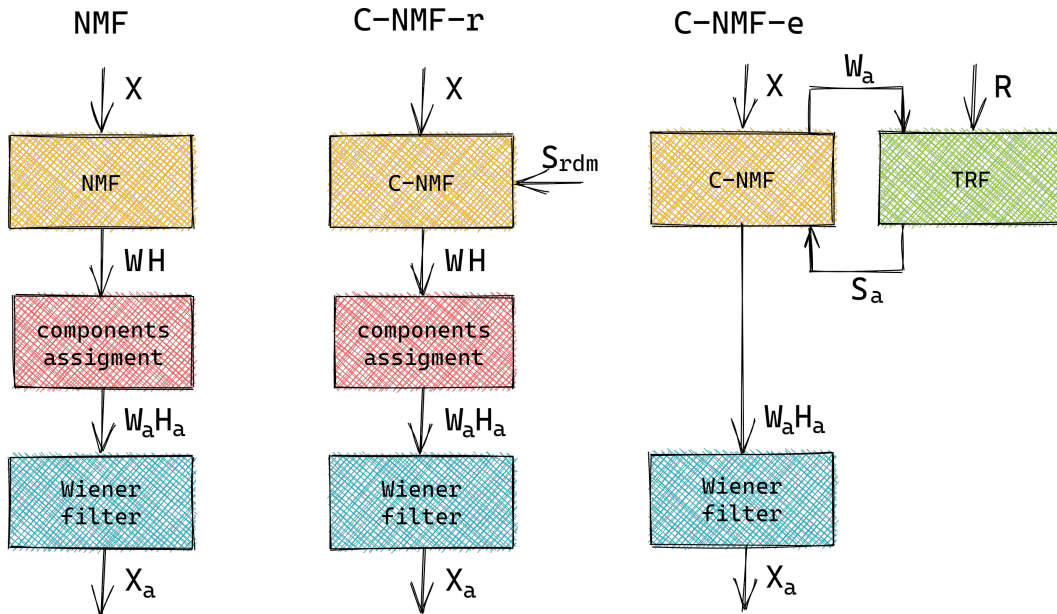


FIGURE 4.5: The proposed EEG-informed source separation algorithm (C-NMF-e) and the two baselines: the blind NMF (NMF) and the randomly-informed C-NMF.

For each method, NMF is run for 400 iterations while the backward model is updated every 100 iterations of the C-NMF-e. For each method, the initialization of \mathbf{W} and \mathbf{H} is obtained by applying a blind NMF to the mixture for 200 iterations. For a given mixture, the initialization of the three models is the same to guarantee a fair comparison. As a reconstruction cost, we chose the Kullback-Leibler divergence. We learned a good initialization of the backward model from a training set of solos (different from the ones used in the test mixtures) and corresponding EEG recordings for each subject and instrument. The Ridge parameter is set to be $\gamma = 0.1$ and the considered temporal context is $[0, 250]$ ms post-stimulus as done in the experiments of Chapter 3.

4.4.1 Evaluation

The models are evaluated using a standard metric in music source separation, *i.e.*, the Signal to Distortion Ratio (SDR) expressed in decibel (dB) and computed using BSSEval v4 [Vincent et al. 2006; Stöter et al. 2018]³. Sometimes, we will refer to the Signal to Distortion Ratio Improvement (SDRi) over the input, as some mixtures may be more or less difficult than others. The metric is computed over the whole length of each music excerpt (around 24 seconds). In the tables below are reported median values. To assert the statistical significance of our model’s improvement over the baselines, we opted for a non-parametric Wilcoxon test on the metrics’ linear values. The considered significance levels are 5%, 1%, 0.1% and 0.01%.

³<https://github.com/sigsep/bseval>

Beside the separation quality, we also evaluate the decoding performance in terms of accuracy on the AAD task as done in Chapter 3. However, here the AAD problem is formulated differently: the instrument that is automatically separated by the proposed source separation system, *i.e.*, the one represented by the \mathbf{W}_a and \mathbf{H}_a , is predicted as being the attended one. The statistical significance was assessed using an adaptation of the computationally-intensive randomization test [Noreen 1989] already introduced in Chapter 3. The considered significance levels are 5%, 1%, 0.1% and 0.01%, and the tests were performed over 10^4 iterations. For details about the test, please refer to § 6.2. It is worth noting that given the user-driven nature of the EEG-driven separation system, the performance, both in terms of separation quality and decoding performance, not only depends on the algorithm but also on the subject’s ability to properly attend to the target instrument. Similarly, the decoding performance now depends not only on the decoding model and the subject’s ability and attention as in Chapter 3 but also on the separation model and the difficulty of the mixture to be separated.

4.4.2 Experimental results

SEPARATION QUALITY In Table 4.1, one can see the median SDR values for different methods, instruments and, in the last two rows, different spatial renderings. As far as spatial rendering is concerned, it is important to keep in mind that the audio signal processed by the source separation system is always mono (*i.e.*, the task is single-channel audio source separation). The “mono” and “stereo” results relate to the way the stimuli were played to the subjects (which we will refer to as spatial rendering) which differently affects

SDR [dB]	Pop								Classical									
	Guitar		Vocals		Drums		Bass		Oboe		Flute		Horn		Cello		Bassoon	
	Duo	Trio	Duo	Trio	Duo	Trio	Duo	Trio	Duo	Trio	Duo	Trio	Duo	Trio	Duo	Trio	Duo	Trio
NMF	3.4	1.9	2.3	5.4	-2.0	7.8	0.6	-12.5	4.4	5.3	6.3	3.7	5.9	5.3	5.5	6.3	4.7	-2.9
C-NMF-r	1.0	2.8	3.2	5.6	0.4	0.9	0.4	-14.9	3.9	-1.7	1.2	1.6	3.7	2.2	7.3	6.6	4.6	1.8
C-NMF-e	4.4	3.4	3.8	5.1	5.6	2.0	5.2	3.9	5.4	1.4	3.0	1.7	2.1	1.6	4.5	3.6	3.6	3.7
Mono	3.4	3.5	3.6	5.2	5.8	1.7	5.2	3.7	5.5	4.8	2.9	2.1	2.3	1.6	4.9	2.9	3.6	3.7
Stereo	4.5	3.4	4.0	3.2	5.4	2.5	9.0	4.0	4.9	-3.9	3.0	1.4	2.0	2.3	4.5	4.1	4.5	3.9

TABLE 4.1: SDR separation results for different models, ensemble types and instruments. The metrics are shown in dB and all values are medians over the corresponding subset of the test set. In the last two rows, the SDR results of the proposed method C-NMF-e are split for stereo and mono listening tests.

their EEG response. For a deeper insight, in Figure 4.6 one can see the same results visualized with boxplots.

Looking at Table 4.1, it is immediate to see that the contrast derived from the EEG can improve the separation quality for all the pop instruments, especially when separated from duets. Particularly significant is the improvement over the blind baseline (NMF) for the drums (more than 7 dB).

It is also clear that the proposed model needs to be fed with meaningful side information and that the activations reconstructed with the backward model are indeed meaningful. In fact, the same model informed with the random side information (C-NMF-r) performs significantly worse than the one fed with the EEG-derived contrast (drums and bass $p < 0.0001$, guitar $p < 0.01$, singing voice $p < 0.05$, Wilcoxon test). In general, the C-NMF-r model introduces lots of artefacts, even without removing the interferers. Moreover, the random side information can even fool the factorization leading to a degradation of the performance w.r.t. the blind NMF. Only in some rare cases (e.g., vocals, drums, and cello), even with the random information, the proposed approach “guides” the separation indirectly by imposing that the \mathbf{H}_a and \mathbf{H}_u activations are different, leading to a little improvement over the blind NMF.

The situation is different for Classical music instruments, where the improvement over the baselines is statistically significant only for the oboe’s separation from duets and the bassoon’s separation from trios. However, this is not in contrast with the results obtained in Chapter 3 where the decoding performances were better for Classical music instruments than pop ones because here we observe also the effect of the separation system and of the difficulty of separating certain mixtures which are dominant factors. The blind NMF is already obtaining a good separation (see NMF results for Classical music instruments in Table 4.1), as the Classical music mixtures of the MAD-EEG dataset can be too easy to separate (e.g., high/low pass filter), and the EEG information helps especially in difficult cases, where the baselines suffer from the task’s complexity. An explicative example of such cases is represented, for instance, by the separation of the drums where the proposed method is significantly better than the baselines. For easy mixtures, instead, it is hard to see the beneficial effects of the additional information w.r.t. the baseline. In any case, the baseline’s results are overall not significantly better than the one of the proposed model ($p > 0.05$, Wilcoxon test).

We remark that the results in Table 4.1 were obtained with $K = 16$, $\mu = \beta =$

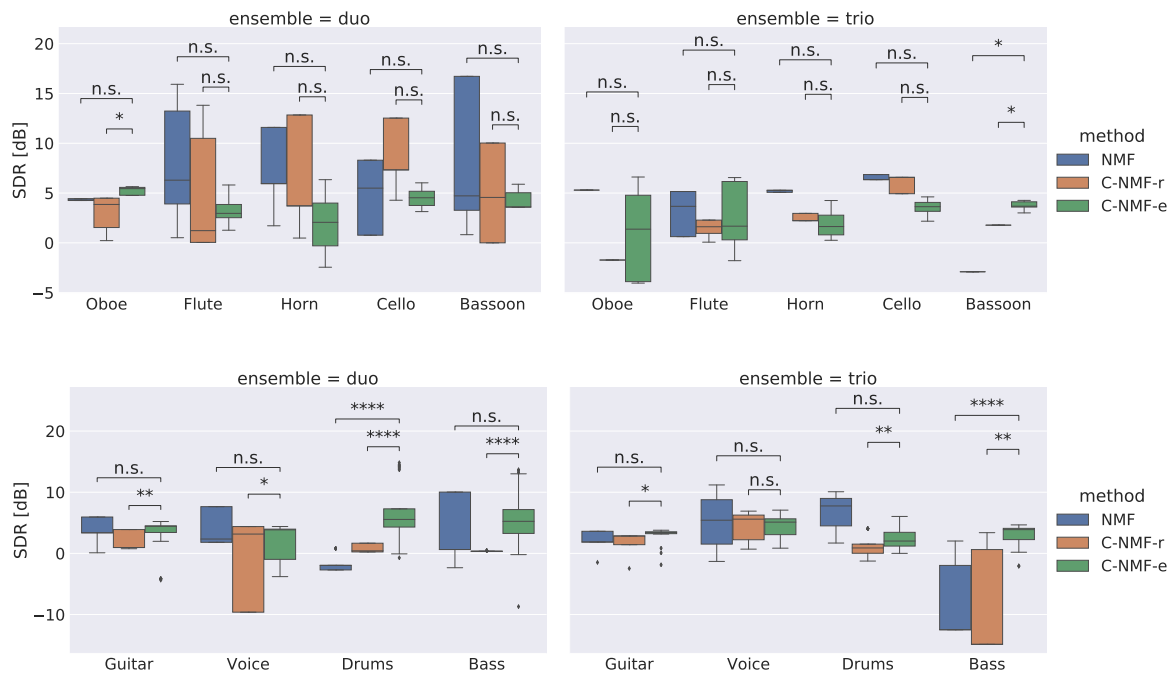


FIGURE 4.6: Signal to Distortion Ratio (SDR) expressed in dB for different instruments in the dataset. In different colors are underlined different methods. To assert the statistical significance of the proposed model with respect to the baselines we opted for a non-parametric Wilcoxon test. The hypothesis test was performed on the linear SDR. "****" denotes high ($p < 0.001$), "****" good ($p < 0.01$), "***" marginal ($p < 0.05$) and "n.s." no ($p > 0.05$) statistical significance.

10, and $\delta = 10^4$, set of values which was found to give good overall results. However, we observed that specific instruments and mixtures would need a specific hyperparameter tuning to maximize the performance. To give an example, by only reducing the value of μ from 10 to 1 when separating the oboe from trios, one can improve the SDR by more than 4 dB. This data-dependent behaviour of NMF scheme's hyperparameters was previously observed [Parekh et al. 2017] and can be mitigated by allowing a user of the system to adjust the hyperparameter values typically through a knob/slider.

SPATIAL RENDERING The stimuli were played to the subjects with two possible spatial renderings: one where both instruments are in the centre denoted as *mono* modality, and one where the instruments are spatialized, denoted as *stereo*. The last two rows of Table 4.1 show the results for these two different cases for all the instruments in the dataset. The results are differentiated w.r.t. the number of instruments in the mixture, and all values are medians over the test set. Intuitively, the stereo setting should help the subject in focusing on the target instrument as it makes it easier to localize it, leading to a better reconstruction of its activations and finally giving a better separation. However, as in Chapter 3, we did not observe statistically significant differences between the two conditions except for a few pop instruments when listened to in duets (guitar $p < 0.01$, singing voice $p < 0.001$, drums and bass $p < 0.05$, Wilcoxon test). In all the other cases, we cannot make any statement ($p > 0.05$, Wilcoxon test).

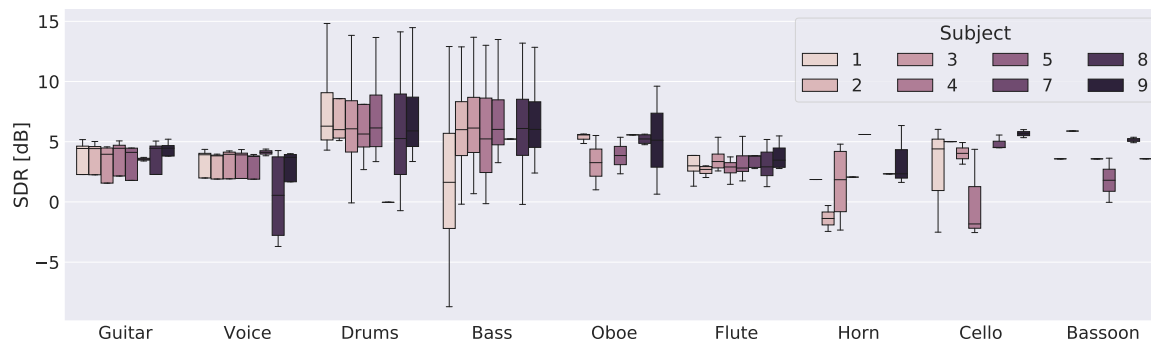


FIGURE 4.7: Inter and intra-subject variability in duets: the SDR results are expressed in dB and different nuances of pink indicate different subjects.

INTER AND INTRA-SUBJECT VARIABILITY Part of the high variance in the SDR performances is because different mixtures in the dataset can be more or less difficult for the separation system. However, most of the variance comes from the very high *inter* and *intra-subject* variability. The attention task may be more or less difficult for different subjects (inter-subject variability), which may depend on factors such as musical training and attention capacity [Di Liberto et al. 2020b]. Simultaneously, one single subject may perform differently throughout the experiment (intra-subject variability), maybe due to stress and fatigue that affect the attention level. These effects are evident in Figure 4.7, where the SDR results for duets are differentiated according to the participants involved in the experiment and the target instrument. Looking at Figure 4.7, one can realise that for a given instrument different subjects may behave very differently while for other ones they behave similarly. Moreover, for single instruments, subject’s performance may span a wide SDR range. For example, regarding Classical instruments, one can observe that the intra-subject variability is generally lower while sometimes there is a clear inter-subject variability. This may be due to the subjects’ unfamiliarity with some instruments like the French horn and the bassoon (see Figure 2.4 in Chapter 2). Another factor is that some instruments can be more difficult than others to follow. For instance, instruments like the bass and the drums, which usually guide the rhythm and tempo, are notably more difficult to track, especially for non-professional musicians and this is reflected in the very high inter and intra-subject variability.

ATTENTION DECODING PERFORMANCES Even if the SDR improvement is not systematic for all the instruments, the main advantage of the C-NMF-e model is that it gives an automatic clustering of the components and automatically enhances the attended source. Therefore, the instrument that is automatically separated by the proposed source separation system, *i.e.*, the one represented by the \mathbf{W}_a and \mathbf{H}_a , is predicted as being the attended one. It is an asset w.r.t. the baselines, which need an additional step to cluster the components and cannot automatically identify the target source.

In Figure 4.8, we report the AAD accuracy values for different instruments and ensemble types. The blue and the red lines represent the chance level for the

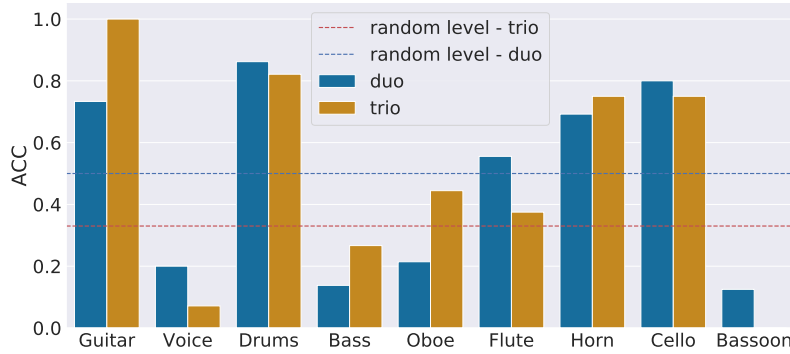


FIGURE 4.8: Decoding accuracy for different instruments and ensemble types compared with the chance level for duets and trios respectively.

duets and the trios. The accuracy is satisfactory and statistically above chance for four instruments: guitar (duo $p < 0.01$, trio $p < 0.0001$, randomization test), drums (duo and trio $p < 0.0001$, randomization test), French horn (trio $p < 0.05$, randomization test), and cello (duo $p < 0.05$, trio $p < 0.01$, randomization test). For some other instruments (singing voice, bass, bassoon, and oboe), the accuracy is much below chance indicating that the contrastive term is always forcing them not to be represented by \mathbf{W}_a and \mathbf{H}_a . The reason for this behaviour lies in a non-customized tuning of the δ parameter. We observed, for instance, that $\delta = 10^4$ causes a drop of the performances for the singing voice and the bassoon, which instead were much above chance with $\delta < 10^4$. As we said previously, this can be easily solved by a customized fine-tuning of the hyperparameters by the user. In the following section we will further analyze the effect of the hyperparameters on the system performance.

EFFECT OF HYPERPARAMETERS We first analyze the number of NMF components necessary to describe each instrument testing 4 values ($\{4, 8, 16, 32\}$). We observe that an increasing number of components improves the separation performance as it allows a more accurate description of the sources. As for the impact of the sparsity constraints imposed on \mathbf{H} and \mathbf{W} by μ and β , respectively, which in our experiments are set to be equal, we tested 4 values ($\{0, 0.1, 1, 10\}$), observing that higher μ and β improve the separation quality as it allows a better source modelling.

Lastly, we tested four reasonable values for δ ($\{10^1, 10^2, 10^3, 10^4\}$), which weights the contrastive term in the C-NMF cost function. We observed that increasing values of δ lead to significantly higher SDR for all the tested instruments except for the French horn, for which there is no significant difference ($p > 0.05$, Wilcoxon test). However, one has to be careful not to choose a too high value of δ , which may push to a trivial solution where the activations of the interferers \mathbf{H}_u are set to zero and all the sources in the mixture are represented by the \mathbf{W}_a and \mathbf{H}_a . This effect is reflected in the AAD accuracy reported in Figure 4.9, where the performance drops for $\delta = 10^4$ for the vocals and the bassoon. However, this effect is strictly instrument-dependent as for other instruments like the cello, the decoding accuracy becomes statistically better than chance only with $\delta = 10^4$ ($p < 0.0001$, randomization test).

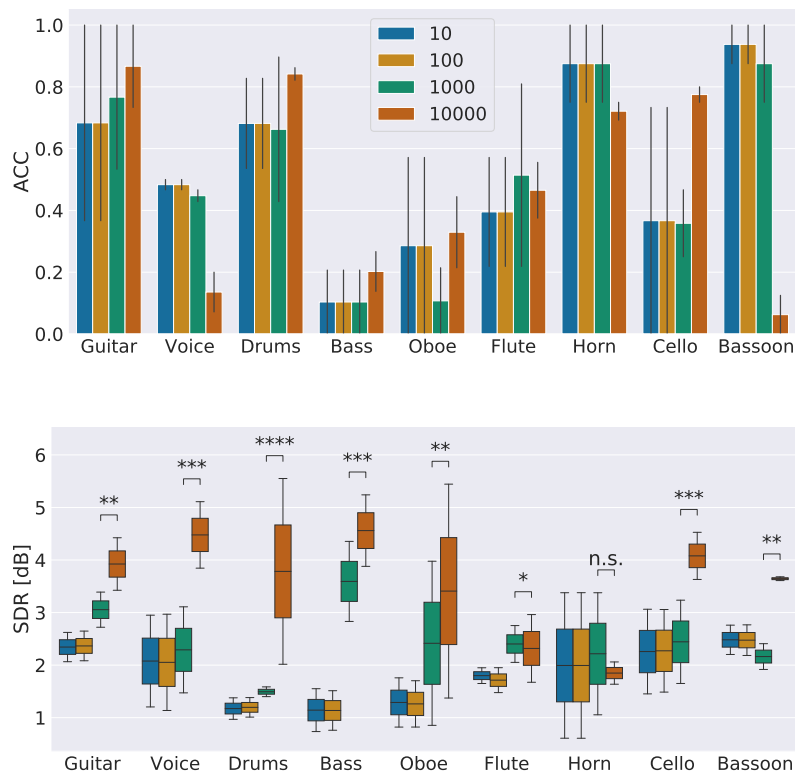


FIGURE 4.9: Decoding accuracy and Signal to Distortion Ratio (SDR) expressed in dB for different instruments and values of the hyperparameter δ that weights the contrastive term. “****” denotes very high ($p < 0.0001$), “***” denotes high ($p < 0.001$), “**” good ($p < 0.01$), “*” marginal ($p < 0.05$) and “n.s.” no ($p > 0.05$) statistical significance for a non-parametric Wilcoxon test on the linear SDR.

4.5 CONCLUSIONS

This Chapter describes a novel *neuro-steered music source separation* framework and conducts an extensive evaluation of the proposed system on the MAD-EEG dataset. The results support the thesis that the EEG can guide and help a source separation system, especially in difficult cases where non-informed models struggle. Our ablation study, where the proposed model is informed with random side information, shows that the C-NMF formulation is not enough by itself but needs to be informed with meaningful side information and that the activations reconstructed with the decoding model may indeed be meaningful. We could reformulate the AAD problem without needing access to the “clean” audio sources, which are absent in real-life scenarios. In fact, thanks to the C-NMF formulation and the EEG-guidance, the NMF components are clustered into the target and interference sources and the attended instrument is the one that is automatically separated by the separation system.

The EEG-driven C-NMF system has the intrinsic limitation of the subject-related variability: if the level of attention of the subject is not sufficient, this will inevitably impact the performance. Another factor that needs to be considered is musical expertise and training, which may help the subject while attending to an instrument.

We believe that this NMF variant is advantageous for neuro-steered music source separation. Indeed the available music-related EEG datasets are still costly and time-expensive to acquire, precluding the possibility to tackle the problem with data-driven approaches. Unsupervised NMF represents a powerful approach in such applications where there is no or a limited amount of training data. Moreover, additional information can be easily incorporated into the model cost function directly at test time. However, even if the C-NMF is unsupervised, we need to keep in mind that we still need pairs of EEG and music data for training the backward model.

Moreover, the proposed algorithm can be generalised and used with temporal activations derived from other modalities than the EEG (*e.g.*, video, score, motion capture data) or from a manual annotation provided by the user (*e.g.*, a sound engineer annotating when the source of interest is active).

5

UGOSA: User-guided one-shot deep model adaptation for music source separation

- ▶ **SYNOPSIS** The scarcity of music-related EEG data precludes the possibility of tackling the problem of neuro-steered music source separation with fully supervised deep learning approaches. In this chapter, we explored alternative learning strategies to alleviate this problem. Specifically, we propose to adapt a state-of-the-art music source separation model to a specific mixture using the time activations of the sources provided manually by the user or derived from his/her neural activity which are available only at test time. This paradigm can be referred to as *one-shot* adaptation, as it acts on the target song instance only. A large part of the material presented in the chapter is the result of a work conducted during my internship at InterDigital R&D France under the supervision of Alexey Ozerov and led to the following conference publication:

- Cantisani, Giorgia, Alexey Ozerov, Slim Essid, and Gaël Richard (2021c). “User-guided one-shot deep model adaptation for music source separation”. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*

5.1 INTRODUCTION

Deep Learning (DL) has profoundly changed the Music Source Separation (MSS) scene of the last years thanks to the appearance of large datasets where the isolated tracks of a set of instruments, usually the most common ones, are available along with the mixture [Rafii et al. 2017; Bittner et al. 2014]. As a consequence, most state-of-the-art MSS systems consist nowadays of Deep Neural Networks (DNNs) trained in a fully supervised fashion [Stöter et al. 2018; Takahashi et al. 2018; Stöter et al. 2019; Luo and Mesgarani 2019; Défossez et al. 2019; Hennequin et al. 2020; Samuel et al. 2020; Takahashi and Mitsufuji 2020; Choi et al. 2021; Li et al. 2021; Sawata et al. 2021]. Those models have proven to be extremely powerful, but only when the training data is enough for learning the enormous amount of parameters they have. However, the availability of a large datasets is not always realized, especially when working on informed MSS where the annotation of the side information is needed and costly to obtain. When the side information is the listener’s

Keywords: Music Source Separation, User-guided, One-shot Domain Adaptation, EEG.

Resources:

- 🔗 Paper
- 🔗 Code
- 🔗 Demo

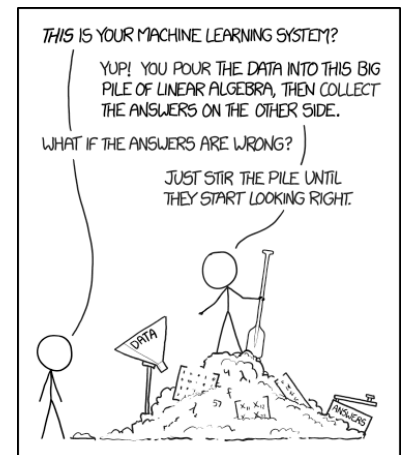


FIGURE 5.1: Is such a pile big enough to hide the corpse of the user? I’m afraid we don’t have enough data. Image courtesy of xkcd, number 1838.

EEG response, this is particularly true as the acquisition process of such data can be very long and expensive as described in Chapter 2.

As seen in Chapter 4, unsupervised techniques like NMF are ideal in such cases as it is easy to incorporate additional information about the sources directly in the optimization cost without requiring a data-intensive training phase. However, NMF-based MSS has its limitations in terms of separation performances and it is desirable to find alternative strategies to inform DL-based systems with side information that is available only at test time.

Usually in DL-based informed MSS, the model is learned using both the side information and the audio material (mixtures) to be separated. One may want, instead, to choose a powerful deep model which was trained in a fully supervised fashion for the MSS task only and adapt it to a specific mixture using the additional information available only at test time. Specifically, we investigated if it is possible to inform a MSS model based on DL using the time activations of the sources provided by the user at test time.

We propose a *User-guided one-shot deep model adaptation for music source separation* (UGOSA), where the time activations of the sources provided by the user are used to fine-tune a pre-trained deep MSS model to the specific test mixture he/she is listening to as in Figure 5.2. The adaptation is made possible thanks to a proposed loss function which aims to minimize the energy of the silent sources while at the same time forcing the perfect reconstruction of the mixture. We underline that the adaptation is *one-shot*, as it acts on the target song instance only and not on a new dataset as most fine-tuning strategies do. The activations of the sources can be manually annotated by the user through an interface or, in a more challenging scenario, be directly derived from his/her neural activity using decoding models like the ones presented in Chapter 3. The EEG-guided variant, namely EGOSA, can be seen as a particularly challenging case of UGOSA, where the interface for annotating the data is replaced by a BCI. This approach also allows us to reformulate the AAD problem of Chapter 3 using the separation model estimates instead of the ground truth sources.

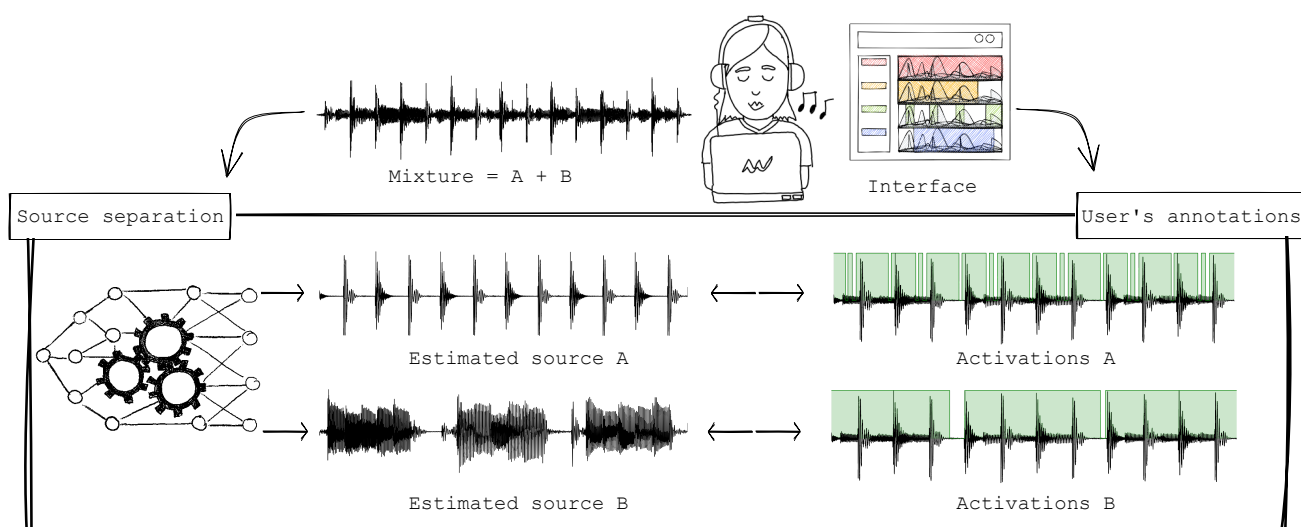


FIGURE 5.2: Time activations of the sources annotated by the user are used to adapt a pre-trained deep MSS model to one specific test mixture.

Even if immature, the results are encouraging and point at promising research directions. While with “ideal” manually annotated activations the preliminary experiments show significant improvements with the adaptation, the EEG-informed case is clearly more challenging and needs further refinements, mainly because the MAD-EEG dataset was not ideal for this study.

5.2 RELATED WORKS

The idea of using time annotations provided by the user to inform a source separation system was already explored in many previous works, mainly based on NMF or NTF [Laurberg et al. 2008; Ozerov et al. 2011; Duong et al. 2014a].¹ Some of them rely on dedicated graphical user interfaces, while others are interactive, where the user can iteratively improve and correct the separation [Bryan and Mysore 2013; Duong et al. 2014b]. Time annotations were also extended to more general TF annotations [Lefevre et al. 2012; Lefèvre et al. 2014; Jeong and Lee 2015; Rafii et al. 2015] but those require much more expertise and effort from the user (and a more complicated user interface). There are also some interesting works where the user can hum [Smaragdis and Mysore 2009], sing or play [FitzGerald 2012] the source he/she wants to enhance as an example to the source separation system. The user can also provide the fundamental frequency or manually correct it [Durrieu and Thiran 2012; Nakano et al. 2020] or associate each instrument to a microphone in a multi-channel recording [Di Carlo et al. 2017].

Only a few works use directly the neural activity of the listener to inform a speech separation model [Pu et al. 2019; Ceolini et al. 2020]. In [Pu et al. 2019], the authors propose an adaptive beamformer that reduces noise and interference but, at the same time, maximizes the Pearson correlation between the envelope of its output and the decoded EEG. In [Ceolini et al. 2020], instead, a speech separation neural network is informed with the decoded attended speech envelope. Ceolini et al. hacked the problem of having a large EEG dataset for training the network using what they called a “noise training scheme”. The model is trained using the ground truth envelopes to which Gaussian noise is increasingly added across the epochs to simulate the deteriorated speech envelopes that will be reconstructed at test time from the EEG [Ceolini et al. 2020].

Within this work, we explore if adaptation is beneficial for DL-based MSS models, as nowadays, most state-of-the-art models are based on a fully data-driven approach without adaptation [Défossez et al. 2019; Luo and Mesgarani 2019; Stöter et al. 2019; Stöter et al. 2018].

Considering the case of user-guided audio source separation based on DL, we observed that the additional information provided by the user is needed in some form also at training time, constraining the architecture and precluding the use of DNNs already pre-trained on other tasks and/or datasets. In the work of Nakano et al., the model was initially trained for both singing voice separation and fundamental frequency estimation and was then adapted using the F0 loss only [Nakano et al. 2020]. In the work of Ceolini et al. the network is built to take as input the amplitude envelope of the sources along with the mixture [Ceolini et al. 2020].

¹The literature review focuses on user-guided methods as it is the primary focus of this investigation. In particular, we considered works where the user is in the loop and actively provides additional information to the source separation system to improve its performance via adapting it to a specific mixture.

In our case, instead, we are interested in a more general framework, where the DNN is trained on the source separation task only, and the activations are used solely for the adaptation. This approach is general since it allows for adapting any DL-based source separation model, using the activations of the target song instance only.

5.3 METHODS

The scope of this work was to investigate if it is possible to adapt a pre-trained DNN for MSS to a particular music piece using the time annotations provided manually by the user, or, in a more challenging scenario, derived from his/her neural activity, no matter which model is used. To this aim, we choose a state-of-the-art MSS model working in the time-domain whose pre-trained weights were made available, and we study fine-tuning strategies using a new loss function we propose which makes use of those time annotations.

5.3.1 Proposed adaptation loss

In supervised training of a MSS model working in the time-domain, the mixture is provided as input; the model outputs the estimated sources which are then compared to the original sources used to create the mixture. The difference between the estimated and the original sources is used to update the model parameters during training. Typically, an ℓ_1 or ℓ_2 loss is adopted, which respectively represents the average absolute error or average mean squared error between waveforms.

In our case, during adaptation, we do not have access to the isolated sources anymore but only to their binary temporal activations. To adapt the weights of the model to the test mixture, we introduce a new loss function based on the binary activations $h_j(t)$ (active: $h_j(t) = 1$ / non-active: $h_j(t) = 0$) of each instrument j at sample t .

We consider \mathcal{X}_t the set of instruments that are present in the mixture x at time frame t . When one instrument is absent, the loss minimizes the ℓ_1 -norm of its estimate while at the same time, it forces the perfect reconstruction of the mixture.² To improve the readability, the time information will be considered in the subscript.

$$L = \frac{1}{T} \sum_{t=1}^T \left[\sum_{j \in \mathcal{X}_t} |\hat{s}_{j,t} - x_t| + \sum_{j \notin \mathcal{X}_t} |\hat{s}_{j,t}| \right] \quad (5.1)$$

Given the binary activations $h_{j,t}$ of each instrument j at time frame t , this formulation can be implemented as follows:

$$L = \frac{1}{T} \sum_{t=1}^T \left[\underbrace{\sum_{j=1}^J (h_{j,t} \cdot \hat{s}_{j,t}) - x_t}_{\text{reconstruction loss}} + \lambda \underbrace{\sum_{j=1}^J |(1 - h_{j,t}) \cdot \hat{s}_{j,t}|}_{\text{activations loss}} \right] \quad (5.2)$$

where the total cost is composed by two terms: the first one concerns the perfect reconstruction of the mixture while the second one the energy

²Often, it is desirable to relax the time activations to weak class labels, indicating a given instrument in a specific time interval. In such a case, it is straightforward to modify the formulation and define \mathcal{X} as the set of instruments that are present in the mixture segment x , obtaining:

$$L = \frac{1}{T} \sum_{t=1}^T \left[\sum_{j \in \mathcal{X}} |\hat{s}_{j,t} - x_t| + \sum_{j \notin \mathcal{X}} |\hat{s}_{j,t}| \right]$$

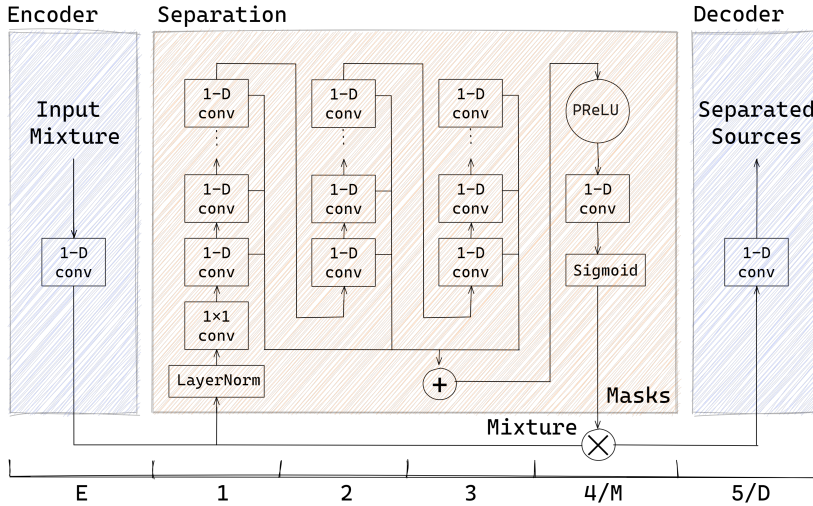


FIGURE 5.3: ConvTasnet architecture.

minimization of the silent sources. If the instrument is active in a given frame t , then $h_{j,t} = 1$ and the energy minimization term is 0.

On the contrary, if $h_{j,t} = 0$, then the energy of $\hat{s}_{j,t}$ is minimized. Only if the instrument is active, it will concur to the mixture reconstruction loss. λ is a hyper-parameter that weights the contribution of the energy minimization term in the total loss.

5.3.2 Model

The source separation model chosen for our experiment is ConvTasnet. This architecture was proposed for single-channel speech separation by Luo and Mesgarani [Luo and Mesgarani 2019] and extended to multi-channel music separation by Défossez et al. [Défossez et al. 2019]. It achieves state-of-the-art results in both tasks, and its implementation and weights are publicly available, reasons why this model was a good candidate for our experiments. Nevertheless, we underline that the proposed approach is general and can be applied to different deep model architectures working in the waveform domain. ConvTasnet is structured as three main blocks:

- The encoder (E in Figure 5.3) transforms a mixture's segments into a non-negative representation in an intermediate feature space;
- this representation is then used to estimate a mask for each source at each time step in the separation subnetwork (1-M in Figure 5.3);
- the isolated waveforms are finally reconstructed transforming the masked encoder features using the decoder (D in Figure 5.3).

Further details about the model can be found in the original paper [Luo and Mesgarani 2019], while for details about its multi-channel implementation for music, please refer to [Défossez et al. 2019].

5.4 EXPERIMENTS

In this work, we considered the implementation of ConvTasnet for multi-channel music separation provided by [Défossez et al. 2019]. The weights of the model pre-trained on the MUSDB18 dataset³ were downloaded from the author’s Github page⁴, where the reader can find further details about the model implementation. The model is built to separate the mixture into four tracks associated with the categories drums, bass, vocals, and others.

To adapt the network to each test mixture we fine-tuned it for 10 epochs on 4-second-long segments extracted from the mixture. The initial learning rate was set to 10^{-5} , batch size to 1 and Ranger was used as the optimizer.⁵ Specifically, Ranger combines RAdam [Liu et al. 2020] and LookAhead [Zhang et al. 2019] optimiser together. Our source code is publicly available.⁶

ADAPTATION STRATEGIES When adapting a DL model for a new task, it is often useless and counterproductive to fine-tune all the network parameters as, for example, the first layers extract some general features which might be useful also for the new task. In our case, the adaptation is not performed over a new task but over a specific instance of the test set. Thus, the task remains the same as the one for which the network was trained. Moreover, the data on which to perform the adaptation is extremely limited (just one mixture), increasing the risk of overfitting. Those factors make the choice of parameters to fine-tune critical and will largely influence the performance.

Let “P” stand for proposed while “B” stand for baseline. “Lx:y” indicates the layers that are fine-tuned (e.g., P-L2:D means that the network is fine-tuned from the second block to the last one using the proposed loss). Please refer to Figure 5.3 for the layer’s names. We consider as the main baseline the original ConvTasnet trained on the MUSDB18 training set (B0). Moreover, for each of the proposed fine-tuning strategies, we obtain a specific baseline B-Lx:y where the model is adapted in an unsupervised manner using the mixture reconstruction loss only and ignoring the activations.

EVALUATION The models are evaluated using standard metrics in MSS, i.e., Signal to Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), and Signal to Artifacts Ratio (SAR) expressed in decibel (dB) and computed using the BSSEval v4 [Vincent et al. 2006] as in Chapter 4. As the SDR is not defined for silent frames, the evaluation is done only where the sources are non-silent.⁷ Each tested configuration is evaluated in terms of the median over all tracks of the median SDR, SIR, and SAR over each track, as done in the SiSEC Mus evaluation campaign [Stöter et al. 2018].

To assert the statistical significance of our model’s improvement over the baselines and to compare different hyper-parameters tuning, we opted for a Wilcoxon test on the linear values of the metrics as in Chapter 4.

Beside the separation quality, we also evaluated the accuracy on the AAD task. The statistical significance was assessed using an adaptation of the randomization test [Noreen 1989] explained in § 6.2.

³The MUSDB18 dataset [Rafi et al. 2017] consists of 150 full-length music stereo tracks of various genres sampled at 44.1 kHz. For each track, it provides a linear mixture along with the isolated tracks for the four categories: drums, bass, vocals, and others. The “others” category contains all other sources in the mix that are not the drums, bass, or vocals.

⁴<https://github.com/facebookresearch/demucs>

⁵<https://github.com/lessw2020/Ranger-Deep-Learning-Optimizer>

⁶<https://github.com/giorgiacantisani/ugosa>

⁷The SDR is not defined on silent frames. Thus, we evaluated the system using BSSEval only on active segments, while when the reference source was silent, we evaluated the predicted energy at silence (PES) proposed by the authors of [Schulze-Forster et al. 2019]. However, by construction, the system is very good at predicting silence, thus the PES metric was not informative and we ended up considering only the SDR for non-silent frames.

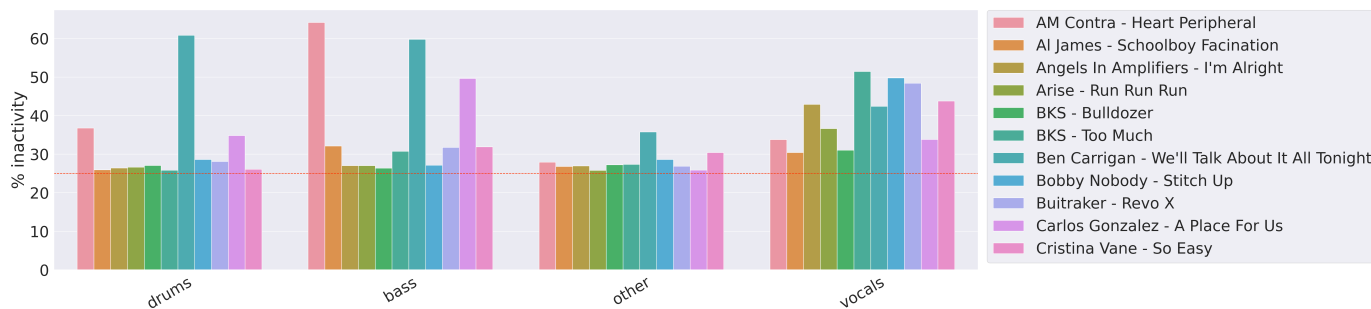


FIGURE 5.4: Amount of silence of each instrument throughout the test songs. The red line represents the 25% of silence we manually set for each source. Without any action, the instruments would be almost always activated, making it hard to evaluate the proposed loss function.

5.4.1 Experiment with manually annotated activations

We first validated the proposed approach in the most straightforward scenario, the one where we assume the user to manually annotate the time activations of each source in the mixture. In order to simulate this situation and work in a controlled setting, the time activations were computed synthetically from the ground truth sources. A first evaluation was performed on the MUSDB18 dataset [Rafii et al. 2017], that is, the same dataset on which the model was pre-trained. In particular, we use the first ten songs of the test set together with the binary temporal activations of each instrument computed in a controlled way to have a clear understanding of how the proposed loss function works and what its weaknesses are.

SYNTHETIC ACTIVATIONS The procedure to obtain the activations is two-fold. First, we manually set to zero each source composing a mixture for one-quarter of the song so as to have at least 25% of silence for each instrument. This step is necessary because otherwise the MUSDB18 test mixtures we wanted to evaluate did not have enough silent parts as can be clearly seen in Figure 5.4. Indeed, what makes the proposed loss different from a simple mixture reconstruction loss is the energy minimisation of the silent sources (second term of Eq. (5.2)). If there are no silences in the mixture, the activations will provide no additional information, and the adaptation would be completely unsupervised.

This procedure belongs to a data preparation step before computing the frame-wise activations. For each test mixture, the procedure is as follows:

1. segment the mixture into four segments of equal length,
2. assign each segment to one source,
3. set each source to zero in the assigned segment.

The source to segment assignment (see step 2. above) is performed randomly to avoid systematic bias. The sources are set to zero in the Short Time Fourier Transform (STFT) domain, so as to have smooth transitions in time between silent and non-silent segments thanks to the STFT windowing.

Then, the time annotations were obtained using the same procedure and hyper-parameters used to annotate the MedleyDB dataset [Bittner et al. 2014], a music dataset which provides the temporal activations of each instrument. The

amplitude envelopes were generated for each source $s_{j,t}$ using a standard *envelope following* technique, consisting of half-wave rectification, compression, smoothing, and down-sampling. The resulting envelope $a_{j,t}$ is then normalized to account for overall signal energy and the total number of sources in the mixture. Finally, the confidence $c_{j,t}$ of the activations $a_{j,t}$ of instrument j at time frame t can be approximated via a logistic function:

$$c_{j,t} = 1 - \frac{1}{1 + e^{\gamma(a_{j,t} - \theta)}}, \quad (5.3)$$

where $\gamma = 20$ controls the slope of the function, and $\theta = 0.15$ controls the threshold of activation. If $c_{j,t} \geq 0.5$, then instrument j is considered active ($h_{j,t} = 1$) at time frame t . Otherwise, if $c_{j,t} < 0.5$, it is considered silent ($h_{j,t} = 0$). No manual corrections were performed on the annotations. An example of the computed envelopes and activations is depicted in Figure 5.5.

HYPER-PARAMETER SENSITIVITY We verified the influence of the hyper-parameter λ on the performances by testing nine different values of λ ranging from 10^{-4} to 10^4 with a logarithmic step. Those results were obtained on the P-L3:M configuration using a window length of 10 seconds. λ expresses the weight of the term that minimizes the energy of the absent sources in the total cost function. In Figure 5.7 (first row) we can see the influence of the hyper-parameter λ on the performances. Only the vocals performances are pretty stable with respect to this parameter with no statistically significant difference in the **SDR**, **SAR** and **SIR** across different values of λ ($p > 0.05$, Wilcoxon test). For the other classes, a higher λ leads to a higher **SIR**, meaning that the suppression of the interferences is more aggressive. This effect is particularly evident for the bass, where one has 5 dB of **SIR** increment, which, however, is not statistically significant ($p > 0.05$, Wilcoxon test). A more aggressive separation is often counterbalanced by a significant deterioration of the **SAR**

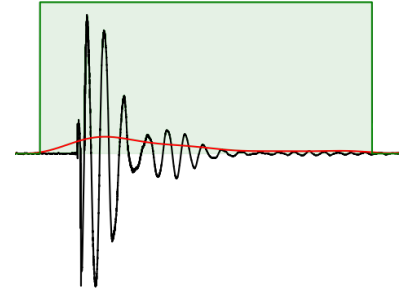


FIGURE 5.5: Detail of the time activations computed synthetically from the ground-truth source of the drums. The audio waveform s_j is represented in black, the amplitude envelope a_j in red and the binary time activations h_j in green.

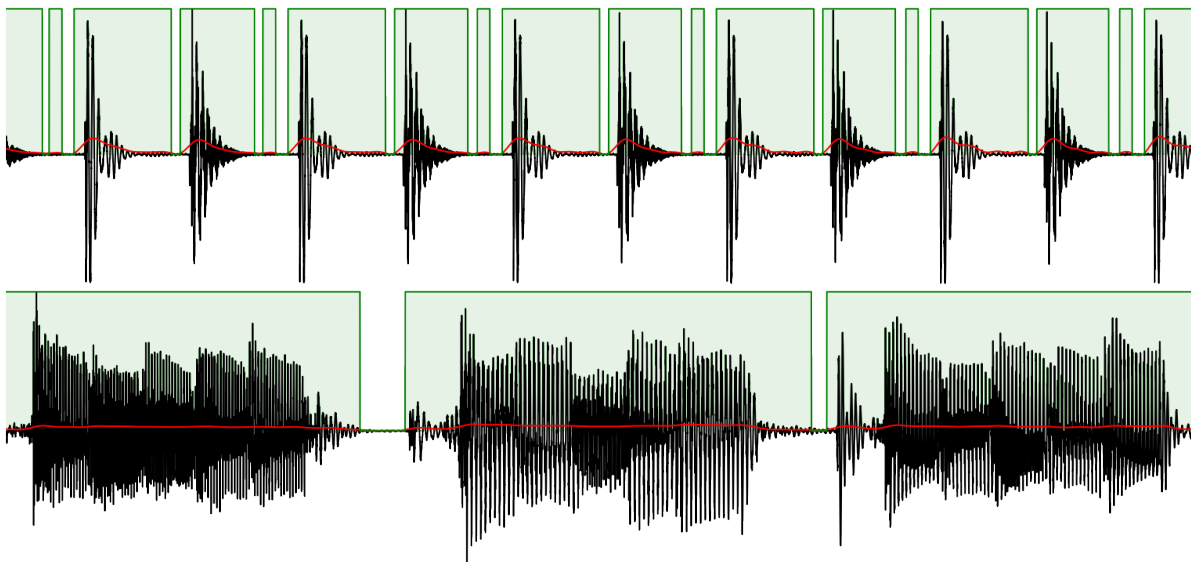


FIGURE 5.6: Example of time activations computed synthetically from the ground-truth sources of the drums (upper plot), and the bass (lower plot) of one song of MAD - EEG. The audio waveform s_j is represented in black, the amplitude envelope a_j in red and the binary time activations h_j in green.

($p < 0.0001$, Wilcoxon test), meaning more artifacts, and of the overall **SDR** (other $p < 0.001$, drums $p < 0.01$ and bass $p < 0.05$, Wilcoxon test).

The performances are not sensitive, instead, to the length of the input segments. The results in the bottom row of [Figure 5.7](#) were obtained on the P-L3:M configuration with $\lambda = 1$ for different lengths of the input segments. We tested five different lengths from 2 to 10 seconds obtaining no statistically significant differences in the **SDR** and **SAR** performances ($p > 0.05$, Wilcoxon test) except for the class “other”, where, with a window below 4 seconds, the **SDR** and the **SAR** marginally decreases ($p < 0.05$, Wilcoxon test). This parameter does not significantly influence the **SIR** ($p > 0.05$, Wilcoxon test) except for the vocals, where it significantly decreases below 4 seconds ($p < 0.01$, Wilcoxon test). For all the instruments except the vocals, a longer context seems to be beneficial to reduce the artifacts however the improvement is not statistically significant ($p < 0.05$, Wilcoxon test) except for the other class where the improvement over the 2 seconds case is significant ($p < 0.0001$ with respects to 4 and 6 s and $p < 0.01$ with respects to 8 and 10 s, Wilcoxon test).

SEPARATION QUALITY In [Figure 5.8](#) one can see the **SDR** expressed in dB for different fine-tuning strategies and instruments in the dataset. Blue bars correspond to models fine-tuned with the proposed loss while orange ones correspond to models fine-tuned using the mixture reconstruction loss only. The red line represents the B0 baseline, *i.e.*, the original ConvTanset trained on the MUSDB18 training set and not adapted at all.

We can see how the **SDR** changes with respect to the block from which we start fine-tuning the network. It is necessary to fine-tune at least from the third block to obtain a significant improvement over the baseline B0. We have to keep in mind that fine-tuning starting from a deeper block corresponds to millions more parameters to fine-tune. If the number of such parameters is high, it requires a proportional amount of training data, which in our case is not possible, as the “adaptation” data comes from only one mixture.

The improvement over the baseline is particularly pronounced for the category “other”, for which the original baseline B0 was struggling the most. As we said before, this category in the MUSDB18 dataset does not represent a specific

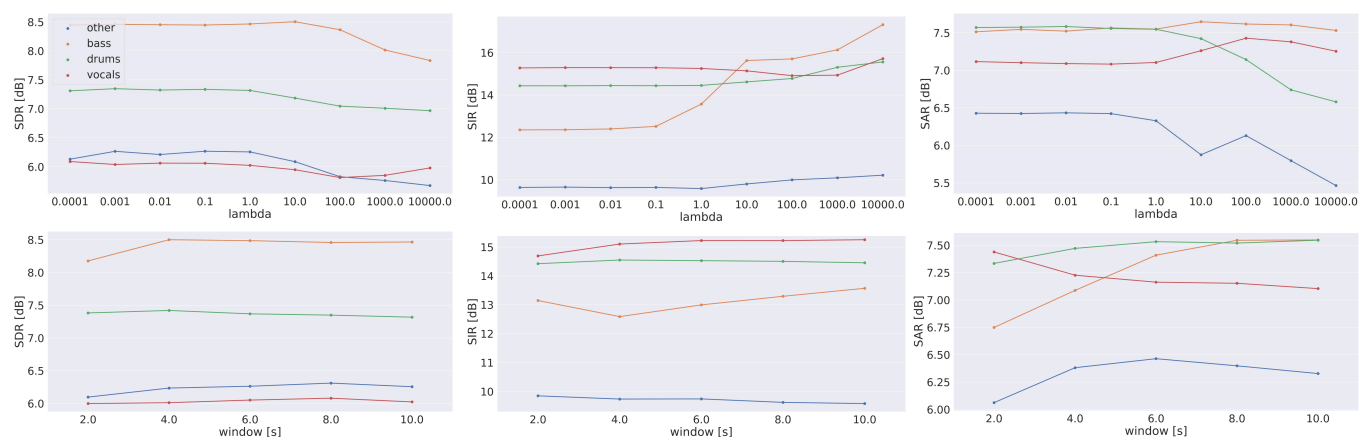


FIGURE 5.7: **SDR**, **SAR** and **SIR** expressed in dB: median over frames, median over tracks for different values of λ and window length.

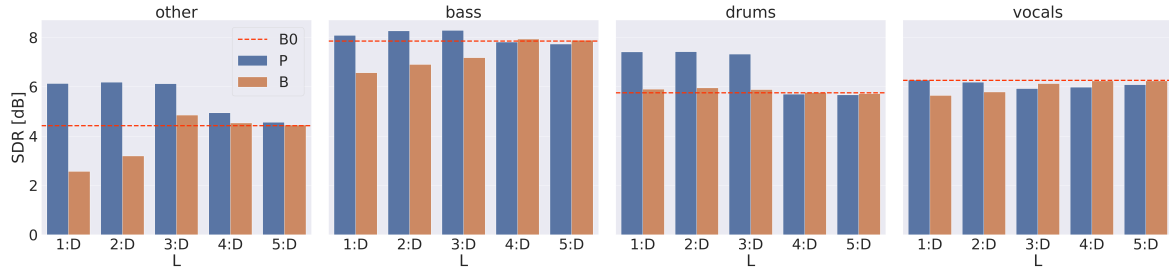


FIGURE 5.8: Median over all tracks of the median SDR (expressed in dB) over each track for different fine-tuning strategies and different instruments in the dataset. Blue bars correspond to models adapted with the proposed loss while Orange ones correspond to models adapted using a reconstruction loss only. The horizontal red line represents the B0 baseline, *i.e.*, the original ConvTanset before adaptation.

		other			bass			drums			vocals		
	#TP	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
P-L1:D	8.2M	6.1	9.3	6.7	8.1	15.3	7.6	7.4	14.6	7.5	6.3	15.9	7.3
P-L2:D	5.6M	6.2	9.5	6.5	8.3	15.3	7.6	7.4	14.5	7.6	6.2	15.7	7.1
P-L3:D	2.9M	6.1	9.5	6.5	8.3	12.3	7.0	7.3	14.2	7.3	5.9	14.3	7.3
P-L4:D	0.4M	4.9	8.9	5.6	7.8	10.4	7.3	5.7	12.7	6.1	6.0	16.5	6.9
P-L5:D	0.01M	4.6	9.1	5.1	7.7	10.9	7.3	5.7	13.7	6.0	6.1	16.8	6.7
B0	-	4.4	10.0	4.5	7.9	11.2	7.4	5.8	15.4	5.9	6.3	18.9	6.7

TABLE 5.1: SDR, SIR, SAR expressed in dB: median over frames, median over tracks for different fine-tuning strategies and different instruments in the dataset. #TP stands for the number of trainable parameters which are fine-tuned during adaptation.

instrument. So, it has much more variability than the other classes which are homogeneous in terms of type of instruments, and the network struggles to find a common representation for those sounds. Adaptation is then particularly useful in this situation, where we need to adapt to a specific instrument which may be different from the ones seen in the training phase. The vocals are the only instrument where we do not improve over the baseline, indicating that probably this class was already well represented in the training data, leaving small room for improvement.

In general, the deeper we fine-tune, the higher the improvement of the proposed model over the corresponding unsupervised baseline, showing that the activations play an active role in the adaptation and that the improvement over B0 cannot be achieved easily in a completely unsupervised fashion.

Looking at Table 5.1, we can have an insight into the evolution of all the metrics. The SDR improvement is mostly due to a SAR improvement, while at the same time, the SIR drops. This means that there are fewer artefacts than before the adaptation, but at the same time, the interferences are not entirely removed. The only instrument which shows a different trend is the bass, for which the SIR and SDR increase and the SAR drops. The bass is the only instrument for which the SIR improves over B0. Separating the bass often corresponds to a low-pass filter and probably fine-tuning allows for better adapting the filter to the register played by the bass in the given piece of music. Motivated by the observation that the decoder has the general function of going back from the feature to the waveform domain, two other fine-tuning strategies

were experimented: one where the decoder weights are frozen during fine-tuning (P-Lx:M) and one where both the decoder and masking blocks are frozen (P-Lx:3). We experimented those variants for all the fine-tuning depths and compared them to the corresponding variants where the network is fine-tuned until the last layer (P-Lx:D). The three variants' performances are not significantly different, indicating that there is no need to fine-tune the decoder or the masking blocks and giving us an insight into the network functionality.

5.4.2 Experiment with EEG-derived activations

Now that we have validated our approach on the manually annotated time activations, we can move to the experiments that use the EEG-derived ones, which are more deteriorated and imprecise.

The EGOSA approach is depicted in Figure 5.9: the amplitude envelopes of each source are reconstructed from the multichannel EEG using a decoding model like the ones presented in Chapter 3 and then binarised to obtain the binary temporal activations necessary for the adaptation. This approach also allows us to reformulate the AAD problem exposed in Chapter 3 using the separation model estimates instead of the ground truth sources.

We performed this second evaluation on the MAD-EEG dataset, which was extensively presented in Chapter 2. We considered only pop mixtures, as, by construction, the network separates the four classes of instruments “bass”, “drums”, “vocals” and “others”. Note that the category “other” in the case of the MAD-EEG dataset coincides with a specific instrument, and precisely the guitar.

EEG-DERIVED ACTIVATIONS of each instrument in the mixture are obtained in two steps. First, the amplitude envelope of each source is reconstructed from the multichannel EEG exploiting the stimulus reconstruction approach explained in Chapter 3. In particular, we estimated subject-specific reconstruction filters for each instrument in the mixture by training a backward model on EEG response of solos with their amplitude envelopes (computed as in

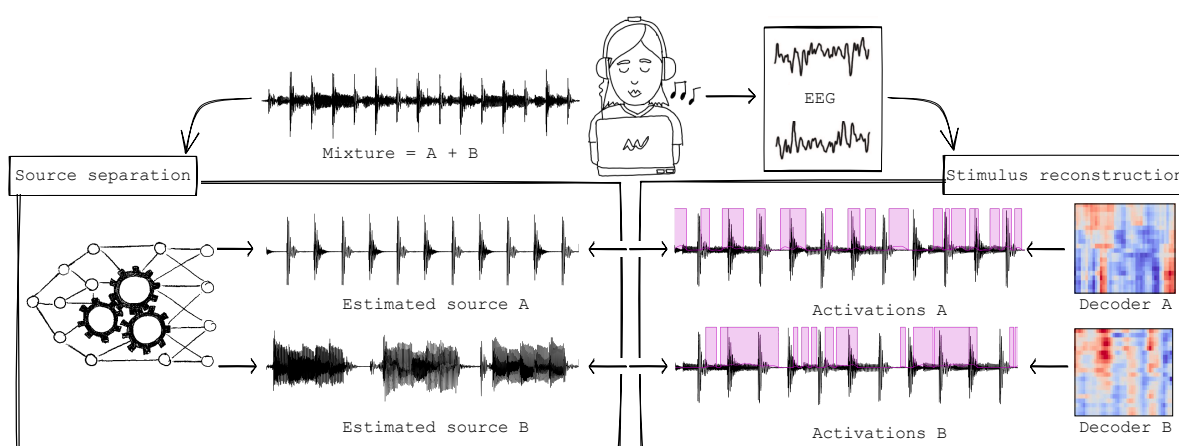


FIGURE 5.9: EGOSA: the time activations of the sources derived from the user’s neural activity are used to fine-tune a pre-trained deep source separation model to the specific test mixture he/she is listening to. Specifically, we first reconstruct the amplitude envelope of each source from the multichannel EEG using a decoding model. Secondly, those amplitude envelopes are binarised according to a threshold to obtain the binary temporal activations.

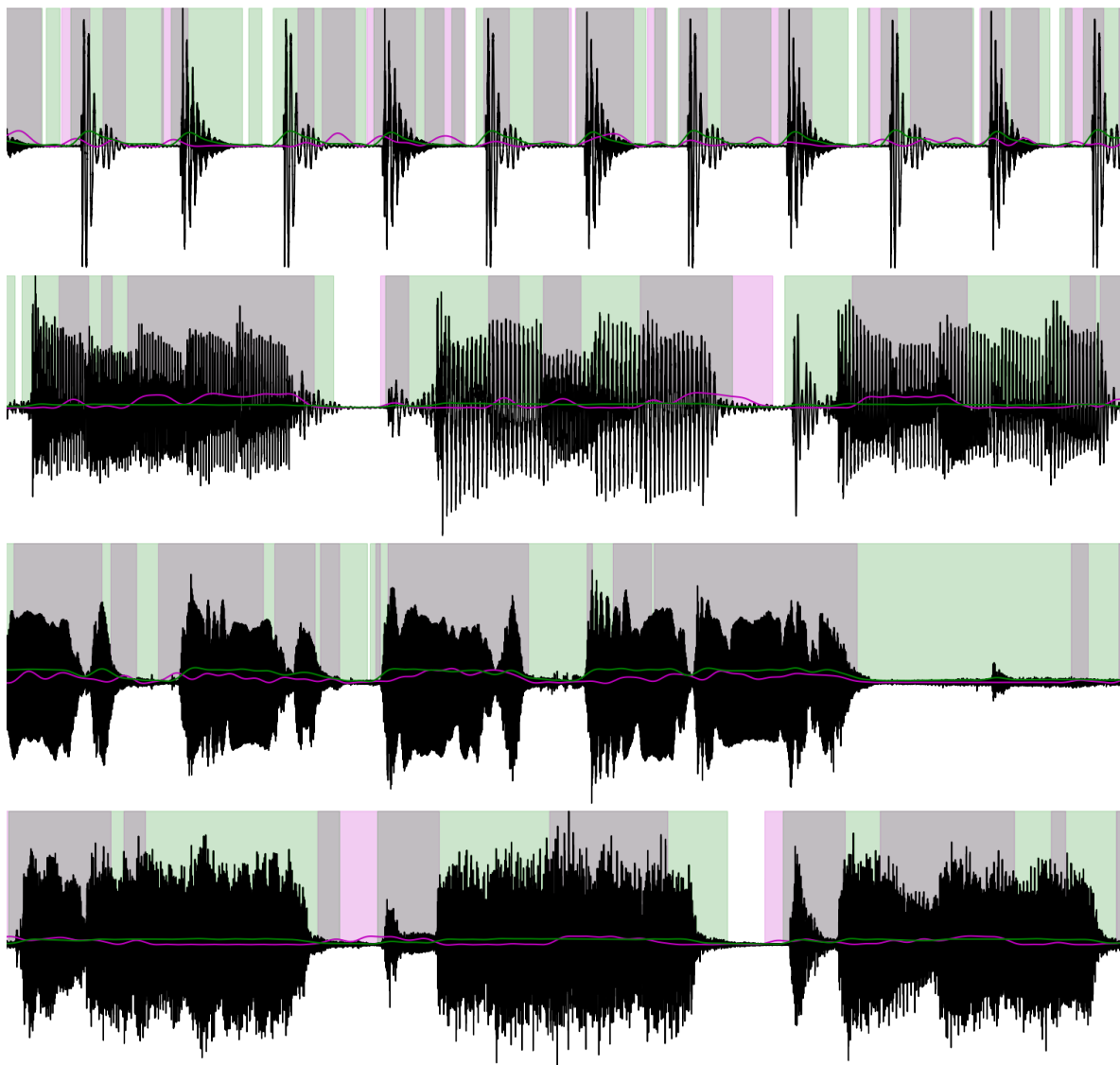


FIGURE 5.10: Example of envelopes and binary activations computed synthetically from the ground-truth sources (in green) or derived from the EEG (in this case, of subject 2) using a decoding model (pink) for (from the upper to the lower plot) the drums, bass, vocals and guitar of one song of MAD-EEG. The audio waveform s_j is represented in black, the amplitude envelopes a_j with a continuous curve and the binary time activations h_j with a colored region.

§ 5.4.1) as targets for the regression. The Ridge parameter is set to be $\gamma = 0.1$ and the considered temporal context is $[0, 250]$ ms post-stimulus as done in the experiments of Chapter 3 and Chapter 4. Secondly, the reconstructed amplitude envelopes are binarized following the same process used in § 5.4.1. An example of how those activations look like and relate to the ones computed synthetically from the ground-truth sources can be seen in Figure 5.10. One can immediately see that the EEG-derived activations are quite imprecise with respect to the synthetic ones. However, the main problem seems to be that the sources are always activated, leaving small room for action to the proposed loss function. In fact, in the MAD-EEG dataset, the sources in the mixtures are almost always activated. As we have seen in the experiments in § 5.4.1, we need enough silent portions in the mixture to benefit from the adaptation as we proposed it. Therefore, we do not expect a clear improvement over the

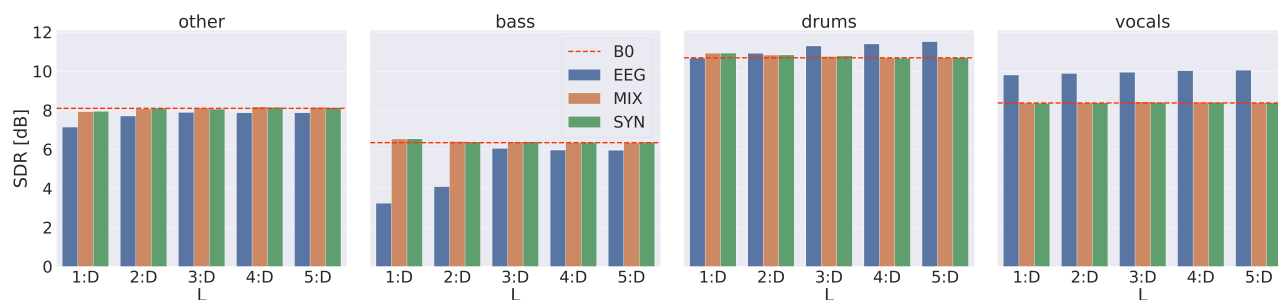


FIGURE 5.11: Median over all tracks of the median SDR (expressed in dB) over each track for different fine-tuning strategies and different instruments in the dataset. Blue bars correspond to models adapted with the proposed loss using the EEG-derived activations, while Orange ones correspond to models fine-tuned using the mixture reconstruction loss only. Green bars correspond to models fine-tuned with the proposed loss using the synthetic activations. The horizontal red line represents the B0 baseline, *i.e.*, the original ConvTasnet before adaptation.

non-adapted model because the data itself is not ideal for this study. Nevertheless, it is still interesting to perform the experiments and analyze the results to better understand the problematic.

SEPARATION QUALITY In Figure 5.11 one can see the SDR expressed in dB for different fine-tuning strategies and instruments in the dataset. On the x-axis, one can see how the SDR changes with respect to the block from which we start fine-tuning the network. The red line represents the B0 baseline, *i.e.*, the original ConvTasnet trained on the MUSDB18 training set and not adapted at all. Blue bars correspond to models fine-tuned with the proposed loss function using the EEG-derived activations. In contrast, orange ones correspond to models fine-tuned using the mixture reconstruction loss only. As an additional control, we fine-tuned the model with the proposed loss function using the synthetic activations computed from the ground-truth sources as described in § 5.4.1 (green bars). This additional experiment aims at distinguishing two different effects on the performance, the ones related to the audio data and the ones related to the EEG. In Table 5.2, the reader can have an insight into the evolution of all the metrics for the baseline B0 and the model adapted using the EEG-derived activations for different fine-tuning depths.

The first observation is that the baseline B0 performs very well on the MAD-EEG dataset, achieving excellent performances for all the instrument classes. The reader can compare the results of B0 presented in Table 5.2 with those of Table 5.1 and immediately see that the MAD-EEG mixtures seem much easier to separate for the ConvTasnet than those of the MUSDB19 dataset, especially for the class “other” and the drums, where the SDR almost doubles. While those results are excellent from the more general point of view of source separation, this also indicates that the MAD-EEG dataset is probably too easy for a state-of-the-art model as ConvTasnet to verify if the EEG information can help the separation. The better the original model’s performance and the easier the mixtures to separate, the harder it will be to see an improvement.

A second observation can be made by observing the performances of the model adapted with the mixture reconstruction loss only (orange bars) and the one adapted with the “ideal” synthetic activations computed from the ground-truth sources (green bars) in Figure 5.11. One can see that there is little

	#TP	other			bass			drums			vocals		
		SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
P-L1:D	8.2M	7.1	12.9	9.4	3.2	15.6	1.1	10.7	18.3	13.8	9.8	13.7	12.6
P-L2:D	5.6M	7.7	13.0	9.6	4.1	16.8	2.2	10.9	18.7	13.9	9.9	13.7	12.7
P-L3:D	2.9M	7.9	13.1	9.8	6.0	15.6	5.8	11.3	19.8	14.0	9.9	13.8	12.7
P-L4:D	0.4M	7.9	13.9	10.0	6.0	15.2	5.8	11.4	20.4	14.2	10.0	14.1	12.5
P-L5:D	0.01M	7.9	14.3	10.0	6.0	15.3	5.8	11.5	20.5	14.2	10.1	14.2	12.5
B0	-	8.1	14.4	9.9	6.3	15.3	7.0	10.7	21.4	12.3	8.4	12.5	12.5

TABLE 5.2: SDR, SIR, SAR expressed in dB: median over frames, median over tracks for different fine-tuning strategies and different instruments in the dataset. #TP stands for the number of trainable parameters which are fine-tuned during adaptation.

or no difference among the two cases indicating that the sources are almost always activated and confirming our worries about the MAD-EEG dataset. When the sources are always active, the proposed adaptation loss reduces to the only mixture reconstruction term, leading to an unsupervised adaptation where no activations are needed. In such a case, it is simply not possible to assert the influence of the EEG information.

Nevertheless, we can look at the results obtained using the EEG-derived activations as we can still get some interesting insights for future works. In this case, it is not necessary to fine-tune in-depth as for the previous experiment on the MUSDB18 data. On the contrary, this deteriorates the results, in some cases proportionally to the fine-tuning depth (see the guitar and the bass in Figure 5.11). This probably happens for two main reasons. Firstly, fine-tuning starting from a deeper block corresponds to millions of more parameters to fine-tune. If the number of such parameters is high, this requires a proportional amount of training data, which in our case is not possible, as the “adaptation” data comes from only one mixture. In the MAD-EEG dataset, the mixtures are only about 24-second long, while the MUSDB18 mixtures are full-length tracks lasting some minutes. Secondly, the fact that the model was already performing very well, combined with the fact that the activations are highly deteriorated and imprecise, leads to a degradation of the performances w.r.t. B0. The more parameters are fine-tuned, the more one can degrade the performances.

Two emblematic cases of this situation are the guitar and the bass. In the first case, we know that the guitar is almost always active throughout most of the mixtures of MAD-EEG and the big difference between green and blue bars indicates that the EEG-derived activations are highly imprecise (e.g., the case of the lower plot of Figure 5.10). In the second case, instead, we know that the bass is not always active. Therefore, the main problem lies in the EEG-derived activations. In Table 5.1 we can see that the SDR degradation for the bass is primarily due to a degradation of the SAR, which measures the artefacts as a consequence of more aggressive removal of the interferer. The explanation is that the EEG-derived activations are zeros where the bass is instead active (see, for instance, the second row of Figure 5.10). This error forces the separation model to output zeros where the source was instead active, removing the other sources better and increasing the artefacts.

On the contrary, the SDR improvement of the proposed approach over B0 is

SDR [dB]	All			other			bass			drums			vocals		
	all	duets	trios	all	duets	trios	all	duets	trios	all	duets	trios	all	duets	trios
attended	8.2	7.4	9.7	7.7	7.2	8.2	6.1	3.2	6.1	11.1	10.1	11.9	10.6	7.8	12.0
unattended	8.5	7.2	8.8	7.9	7.5	8.1	5.9	3.2	5.9	11.4	10.1	11.9	10.1	8.1	10.5

TABLE 5.3: SDR for the best configuration (P-L5:D) differentiated to whether the separated instrument was also the attended one.

clear for the vocals and the drums, which are the less activated sound source in the dataset. The vocals present an SDR improvement which is mainly due to better removal of the interferers, which does not increase the artefacts. For the drums, instead, the separation is simply less aggressive therefore reducing the artefacts. However, we must underline that those improvements over B0 are not statistically significant ($p > 0.05$, Wilcoxon test).

In Figure 5.10, one can see that the activations computed from the ground truth source of these two sources indicate that they are always activated. This happens because, in MAD-EEG, the ground truth sources present some cross-talk between microphones (the instruments were not recorded separately). Therefore, even if the source is silent, it is possible to track the energy of the residual sources. The EEG-derived activations, instead, even if imprecisely, indicate much more silence.

AUDITORY ATTENTION DECODING In Table 5.3 we untangled the SDR results for one of the best configurations (P-L5:D) according to whether the separated instrument was also the attended one. Intuitively, we should get a higher SDR if the separated instrument is also the attended one as we should get a better reconstruction of its activations from the EEG. However, the difference between these attended and unattended is never statistically significant ($p > 0.05$, Mann-Whitney test). As we have seen previously, it seems that the dominant factor in the results is the fact that the sources are always activated and that the mixtures are very easy to separate, precluding the possibility to evaluate the influence of the EEG information.

In the left plot of Figure 5.12, the reader can see the PCC computed between the AE reconstructed from the EEG and the ones derived from the estimates of the separation system. The difference between the distribution of PCCs computed with the estimates of the attended source and the ones with the unattended sources is only significant for the bass. This fact should not come as a surprise because, as we have seen in Chapter 3, the AE is a poor descriptor of the music signal, and to have a stronger contrast, we would need to use a TF representation as target.

IMPROVING AAD WITH TF AUDIO DESCRIPTORS Therefore, we tested a second audio descriptor, the MEL spectrogram, which was proven to be the more robust for the AAD task for music (see Chapter 3). The adaptation procedure does not change because the MEL is used only as a target for the decoding. For the adaptation, the binary activations are derived from the MEL spectrogram as follows: we consider MEL bands as narrow-band amplitude envelopes at specific frequencies. The broadband amplitude envelope can then be reconstructed as the average narrow-band amplitude envelopes across the frequency

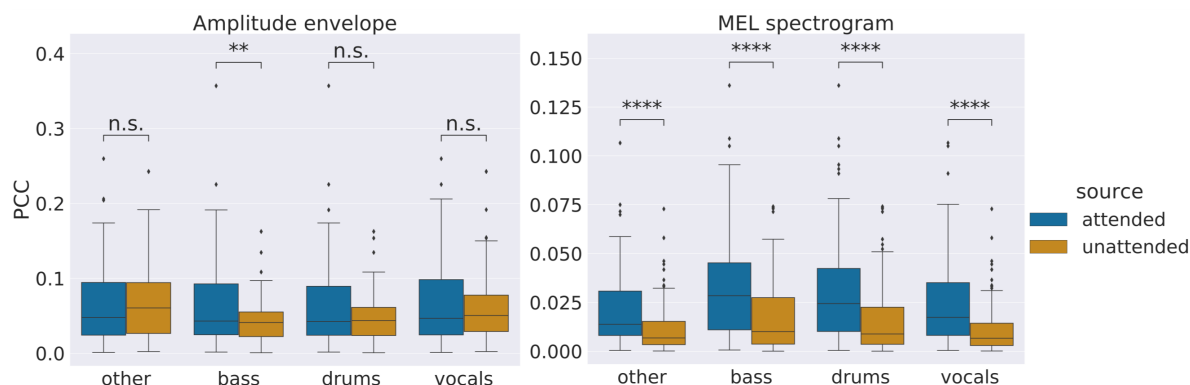


FIGURE 5.12: PCC of the attended and unattended sources for different instruments and audio descriptors. “****” denotes very high ($p < 0.0001$), “****” high ($p < 0.001$), “***” good ($p < 0.01$), “**” marginal ($p < 0.05$) and “n.s.” no ($p > 0.05$) statistical significance of the difference among the two conditions for a non-parametric Wilcoxon test.

bands and binarized following the same process used in § 5.4.1.

In the right plot of Figure 5.12, the reader can see the PCC computed between the MEL spectrograms reconstructed from the EEG and the ones derived from the estimates of the separation system. For the MEL the contrast between the attended and the unattended source is much more remarkable than for the AE, similarly to what previously verified Chapter 3.

This fact is also reflected in the decoding performances in Table 5.4. Note that here the AAD problem is tackled without access to the ground truth sources as in Chapter 3 but using the separation system estimates. Specifically, we computed the PCC between the audio representation reconstructed from the EEG with the ones computed from the separation model estimates, and the attended instrument is recognised as the one that has the highest correlation. The chance level is 50% for duets, around 33% for trios, and approximately 44% for all the test mixtures together.

The MEL descriptor indeed achieves better performances than the AE for all the instruments except for the bass, for which they perform similarly. Indeed the bass was the only instrument for which the difference among the two distributions of PCCs was statistically significant also for the AE descriptor. It is also more robust for trios, as already verified in Chapter 3.

Except for a slight drop in the performances, the results are consistent with the ones obtained in Chapter 3 where we were using the ground truth sources. This fact indicates that the estimates of the source separation system are good enough to replace the ground truth sources.

Even so, the separation performances do not improve significantly with the new audio descriptor. Note that the reconstructed MEL is not used directly in the adaptation, but it is used to derive the energy envelope of the signal and then its binary activations. Therefore, the proposed loss function was not adapted to deal with TF representations, and we reserve it for future works.

Accuracy(%)	All	Duets	Trios	other	bass	drums	vocals
AE	46 *	55 n.s.	38 n.s.	36 n.s.	68 ****	44 n.s.	41 n.s.
MEL	66 ****	66 ****	67 ****	75 ****	65 ****	77 ****	51 *

TABLE 5.4: Decoding accuracy for the best configuration (P-L5:D) and the [AE](#) and [MEL](#) audio descriptors. “****” denotes very high ($p < 0.0001$), “***” high ($p < 0.001$), “**” good ($p < 0.01$), “*” marginal ($p < 0.05$) and “n.s.” no ($p > 0.05$) statistical significance compared to chance level for a non-parametric randomization test.

5.5 CONCLUSIONS

In this chapter, we proposed a *User-guided one-shot deep model adaptation for music source separation*, where the temporal segmentation annotated by the user is used to adapt a pre-trained deep source separation model to one specific test mixture. The adaptation is possible thanks to a newly proposed loss function that aims to minimize the energy of the silent instruments while at the same time forcing the perfect reconstruction of the mixture. We emphasise that the proposed approach is general and can be applied to other types of audio sources (speech, natural sounds) or different deep model architectures. We experimented with two variants: one where the user manually annotates the activations of the sources and a more challenging one where the activations are reconstructed from the neural activity.

The results show that for improving the separation quality, we need at least a weak guiding signal (time activations) in a semi-supervised setting and that an utterly unsupervised adaptation is not enough (mixture reconstruction loss only). The results obtained with “ideal” manually annotated activations in the experiments on the MUSDB19 dataset are promising. They show that a state-of-the-art [MSS](#) model like ConvTasnet may be significantly improved via adaptation with a few epochs to the specific test mixture, especially in complex cases. However, the improvement is not systematic when it comes to using the [EEG](#)-derived time-activations on our data, mainly because MAD-EEG was not ideal for this study. Firstly, in this dataset, the sources tend to be constantly activated, making it hard to see an influence of an adaptation based on time activations. Secondly, the mixtures to be separated are too easy for a state-of-the-art model such as ConvTasnet, making it hard to see the influence of the [EEG](#)-derived information.

Even if the separation quality does not improve systematically, thanks to the proposed approach, it is possible to reformulate the [AAD](#) problem exposed in [Chapter 3](#) using the separation model estimates instead of the ground truth sources. The results obtained with the Mel spectrogram as audio descriptor for the decoding are satisfactory, with only a marginal drop in the performances, if compared with the ones obtained in [Chapter 3](#) where we were using the ground truth sources. This fact indicates that the source separation system estimates are good enough to replace the ground truth sources

Part IV

EPILOGUE

6

Conclusions

- ▶ WITHIN THIS WORK, we explored how to inform and guide a Music Source Separation system exploiting previously not considered modalities such as the user's selective auditory attention to a source characterized in terms of his/her neural activity. Specifically, we investigated two main problems which are intrinsically intertwined with each other:

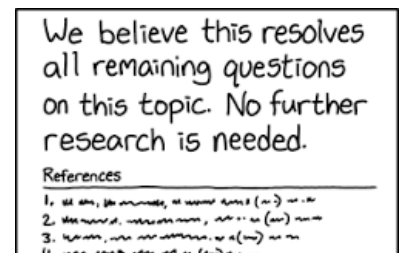
- A. *EEG-based decoding of auditory attention to a target instrument in polyphonic music mixtures;*
- B. *Neuro-steered source separation of the target instrument from a polyphonic music mixture.*

In this chapter, we will summarize the principal findings of the current investigation, and present a discussion on future perspectives and research directions. We hope the results and ideas investigated in this dissertation will stimulate and encourage novel works in this fascinating research direction.

6.1 SUMMARY OF CONTRIBUTIONS

After introducing the motivation and objective behind this work, in [Part I](#), the contributions of this thesis were presented in [Part II](#) and [Part III](#), elaborating on the two problems above. The pursuit of these goals led into the following contributions and outcomes:

- ▶ **PART II AUDITORY ATTENTION DECODING**
 - **MAD-EEG** We assembled a music-related **EEG** dataset which allows for studying the problems of single-trial **EEG**-based **AAD** and **EEG**-guided **MSS** for realistic polyphonic music. It represents the first dataset of its kind for music stimuli and can also be differentiated from those commonly used for studying **AAD** for speech stimuli. The proposed experimental setting differs from the ones previously considered as the stimuli are polyphonic and are played to the subject using speakers instead of headphones. **MAD-EEG** represents our first main contribution and is available to the research community as a free resource.



JUST ONCE, I WANT TO SEE A RESEARCH PAPER WITH THE GUTS TO END THIS WAY.

FIGURE 6.1: Three years ago, I once dreamed of writing it in my PhD dissertation. Image courtesy of xkcd, number 2268.

- MAAD We investigated for the first time the problem of AAD to a target instrument in polyphonic music based on the continuous EEG response. To this end, we exploited the so-called backward model, which was proven to successfully decode the attention to speech in multi-speaker environments. To our knowledge, this model was never applied before to musical stimuli for AAD, and we extensively evaluated it on MAD-EEG. The primary outcome of this study is that the EEG tracks musically relevant features highly correlated with the attended source and weakly correlated with the unattended one making it possible to decode the auditory attention towards a specific instrument in the mixture.

► PART III NEURO-STEERED SOURCE SEPARATION

- C-NMF We proposed a *neuro-steered* MSS framework where we leverage the fact that the attended instrument’s neural encoding is substantially stronger than the one of the unattended sources left in the mixture to inform a source separation model based on NMF and automatically separate the attended source. Thanks to the C-NMF formulation, we could reformulate the AAD problem differently, without needing access to the “clean” audio sources, which are absent in real-life scenarios. We extensively evaluated the proposed system on MAD-EEG, obtaining encouraging results, especially in difficult cases where non-informed models struggle.
- UGOSA We investigated whether it is possible to inform a MSS model based on DL using the time activations of the sources manually annotated by the user or derived from his/her EEG response available at test time. Indeed, the scarcity of music-related EEG data precludes the possibility of using fully supervised DL approaches, which, however, represent the state-of-the-art in MSS. This approach can be referred to as *one-shot*, as the adaptation acts on the target song instance only. Thanks to the proposed approach, we could reformulate the AAD problem using the separation model estimates instead of the ground truth sources. Even if immature, the results are encouraging and point at promising research directions.

Taken together, we hope that these contributions make a step forward towards the direction of integrating BCI and MSS. Nevertheless, much remains to be done, and many research questions arise from the conducted investigation.

6.2 FUTURE PERSPECTIVES

In the previous section, we have summarized the main findings and contributions of the thesis. Nevertheless, we have only scratched the surface of many problems related to EEG-based AAD and neuro-steered MSS. Besides, there are many limitations and much room for improvement in the methods proposed here. This section elaborates on short and long-term research directions that arise as natural follow-ups to the topics discussed so far.

- ▶ **LACK OF DATA** The lack of freely available music-related EEG datasets has been a strong hindering factor for the research in this field. It is a common experience that acquiring such a type of dataset is time-consuming and expensive. It requires specific equipment and experience and a long phase of experimental design, preparation and participants recruitment. Participants are available for a limited time and cannot be overloaded with too long recording sessions. Therefore, those datasets are often limited in terms of recording hours and the number of participants. These factors represent a significant obstacle to the research in the field. With MAD-EEG we hope to help researchers pursuing research in the field, especially for those working in MIR who usually do not have the equipment and expertise of a cognitive or neuroscience oriented laboratory. We are aware of the intrinsic limitations of the dataset, and in future works, we plan to extend the dataset in terms of the number of EEG recordings and stimuli variants and behavioural data. Although, the dataset will still have a size that does not allow studying DL models unless specific learning strategies are adopted. In such low-labelled data regimes, it is necessary to focus on alternative strategies to exploit the large amount of unlabelled data that is often available from close and similar domains.

- ▶ **SUBJECT-INDEPENDENT MODELS** Throughout the thesis, the decoder was always trained in a *subject-specific* fashion, which means that only data recorded from the same subject under test are used for training the decoder. Subject-specific models represent standard practice in BCI because the EEG temporal and spatial characteristics vary significantly between subjects. However, this can be a substantial limitation, especially in the optic of real-life applications, as this approach requires a time consuming and inconvenient calibration phase. A *subject-independent* model would allow avoiding such a calibration phase and also incredibly enlarge the amount of training data of the already existing and available datasets. Indeed, a subject-independent model still requires labelled data, but this can also come from subjects different from the one under test. This fact would allow pre-training such models and make them much more practical for realistic applications. However, subject-independent BCIs have generally shown poor performances in the literature if compared to subject-specific ones [Ghane et al. 2021] and this applies also to AAD models [O’sullivan et al. 2014] mainly due to the high inter-subject variability in the EEG data. Therefore, specific research needs to be conducted in this direction. Transfer Learning (TL) techniques to adapt a pre-trained decoder to unseen subjects as proposed by Geirnaert et al. might be the way to go [Geirnaert et al. 2021a].

- ▶ **ADAPTIVE DECODING MODELS** Another under-considered aspect in AAD research is that the decoding models are not adaptive to the new test data leading to suboptimal results. Adaptation could help in all situations with changing environmental conditions, audio sources, and brain activity which is non-stationary. The EEG temporal and spatial characteristics vary significantly between subjects (different scalps, electrodes placement/impedances) but as well among the data of the same subject (electrode displacements, change of electrode-skin contact impedance, different recording sessions). There have been a few efforts in this direction [Akram et al. 2017; Miran et al. 2018;

Aroudi et al. 2020; Geirnaert et al. 2021a]. Akram et al. employ state-space models to compute a dynamic estimate of the decoder over time [Akram et al. 2017] while Miran et al. extended this work by making it able to operate near real-time [Miran et al. 2018]. In both cases, the decoder is estimated for each new data segment in an unsupervised fashion and then applied again to that same data segment to determine the attended source. Hence, the model is adaptive with respect to the new incoming data. Geirnaert et al. instead proposed to adapt subject-specific models to new subjects in an unsupervised manner. The autocorrelation matrix is updated on the new batch of data, and the subject-specific model is thus updated. After prediction, one can also compute the cross-correlation matrix and re-update the decoder iteratively. All those approaches are defined as unsupervised because they do not require the attended/unattended labels but still require the isolated sources to update the cross-correlation.¹

- ▶ **DEEP LEARNING FOR BRAIN SIGNALS** The scarcity of music-related EEG data involving attention precludes the possibility of tackling the problem of neuro-steered MSS with fully supervised DL approaches. As explained in Chapter 4, unsupervised techniques such as NMF are ideal in such cases as it is easy to incorporate additional information about the sources directly in the optimization cost without requiring a data-intensive training phase. However, MSS systems based on NMF have their limitation in terms of separation performances if compared to DL approaches. It is desirable then to use DL models and possibly find alternative learning strategies to alleviate the problem of the lack of music-related EEG data involving attention. In Chapter 5 we proposed a one-shot adaptation of a pre-trained DNN for MSS to a specific mixture using the user's EEG response, which is available only at test time. It was a first, straightforward attempt to work around the lack of significant training data. However, it has its limitations, and there are many other directions we would like to explore in future works. For instance, only the MSS is DL-based, while the EEG decoding part still relies on a linear regression model. Generally speaking, it is not yet clear if DL has significant advantages compared to traditional approaches for a variety of different BCI and monitoring applications [Roy et al. 2019] and the same can be said for AAD [Geirnaert et al. 2021b]. Certainly, such data-hungry models struggle in low-labelled data regimes, which are the standard when working with EEG [Roy et al. 2019]. We indeed lack data related to attention, but there exist many other small music-related EEG datasets where the subjects were not attending any particular source in the mixture. It would then make sense to aggregate all these heterogeneous auxiliary datasets to scale the training data and use the attention-related dataset only in a second phase as a target dataset thanks to some TL techniques.
- ▶ **SELF-SUPERVISION** One possible strategy is to use Self Supervised Learning (SSL) to learn from the auxiliary datasets a feature representation for both the audio and neural data optimal for joint tasks. The idea is then to use those representations to solve multi-modal problems such as AAD and neuro-steered MSS on the target dataset. SSL allows learning representations from unlabeled data by exploiting the intrinsic structure of the data in a pretext task

¹In Chapter 3, the decoder adaptation happens differently: every certain number of NMF iterations, the dictionary associated with the attended source is used as an updated feature extractor for the decoder training data (solos of the attended instrument). The autocorrelation data is not updated, and only the cross-correlation matrix changes. It would be now interesting to explore another variant, where the solos of the attended instrument are used only to obtain a good initialization of the decoding model, which is then updated in an unsupervised fashion using the test data only. The adaptation, in this case, can act on both the autocorrelation and cross-correlation matrices as proposed by Geirnaert et al. [Geirnaert et al. 2021a], obtaining an adaptation that acts on both the neural (non-stationarity, different subject) and audio data (different environmental conditions, different source type).

[Jing and Tian 2020; Banville et al. 2021a]. These representations can then be used in a *downstream task* for which there are limited or no annotated data. Specifically, the unsupervised problem is then reformulated as a supervised one by automatically generating the annotations from the data with the condition that the pretext and downstream tasks must be sufficiently related. Despite its potential, only a few works have used SSL to improve EEG-related tasks over standard approaches [Yuan et al. 2017; Banville et al. 2021a; Kostas et al. 2021], but never in multimodal scenarios.

In our case, a good pretext task can be the one of *relative positioning*, i.e., determining whether a pair of representations, one for the audio and one for the EEG are synchronized as proposed by [Banville et al. 2021b].² One can artificially generate millions of *positive* and *negative* pairs of EEG and audio data, which are respectively located within a local positive window or outside a long-range negative window and train the system in a *contrastive* fashion. One can also consider using triplets, where the anchor is represented by an EEG segment, and the positive and negative examples are audio segments respectively located within the positive window and outside the negative one. Here, the non-stationarity of the EEG data is not a drawback but a necessary assumption. Moreover, the smoothness assumption (neighbouring representations have the same label) is fair because the EEG tracks slow-varying features of the audio, such as the amplitude envelope.

However, in the experiments of Banville et al., the time windows employed are pretty long (order of minutes), while in our case, we are interested in features on a smaller time scale (order of seconds). Assuming that representations within the positive window are similarly labelled, it might not be easy to expand to time scales closer to that of one of our trials. Additionally, in our case, mining negative pairs hard enough for learning might be tricky. The negative examples need to be difficult enough to prevent the network from learning other data features we do not want it to learn (e.g., subject, song, trial).³ Adversarial training can be a helpful coupling strategy to make the representations learned with self-supervision independent and robust of those features of the data we are not interested in. We are currently conducting experiments in this direction.

- ▶ **INCONSISTENCIES ACROSS DATASETS** Another significant challenge that immediately emerges is the fact that the auxiliary datasets are often heterogeneous and inconsistent. It means, for instance, that they were collected using different protocols and headsets, resulting in varying channel ordering, numbers, and often different signal references. It is also common to have noisy or even missing channels. Consequently, available music-related EEG datasets are heterogeneous, and all of them are very small. Scaling EEG training data seems, therefore, only possible by aggregating them according to some strategy. Moreover, transferring trained systems across datasets, for instance, from the auxiliary to the target one, exhibits the same difficulty. These factors require specific strategies for training a system with heterogeneous datasets and enable TL across the auxiliary and the target datasets. Wu et al. propose to use the common subset of channels shared between headsets [Wu et al. 2016]. Other more elaborate approaches propose to use attention mechanisms that

²A similar approach, called *match/mismatch paradigm* based on [Cheveigné et al. 2018] was proposed by Cheveigné et al. to compare performance of different linear stimulus-response models [Cheveigné et al. 2021]. The paradigm was investigated by Accou et al. in the non-linear case [Accou et al. 2021]. However, in both cases, the match/mismatch paradigm is not used as pre-text task for pre-training a DL-based model for AAD (downstream task).

³We do not have, in fact, long recordings as the ones for studying sleep staging [Banville et al. 2021a], but shorter recordings of subjects listening to different musical pieces. For instance, if the negative pair is chosen on a different audio stimulus, the network might learn to classify the song. If we choose a negative window for another subject or recording session, the network might learn to classify different trials or subjects. If we choose the negative pair within the same subject and same audio stimulus, we have to be careful with the repetitions that often occur in music, and that can elicit a similar EEG response.

recombine the input channels into a fixed number of virtual channels [Nasiri and Clifford 2020; Guillot and Thorey 2021; Saeed et al. 2021], which, however, are not easily transferable when going from the pretext to the downstream task. In contrast, the dynamic spatial filtering proposed by Banville et al. allows for re-using the same filter learned in the pretext task as it is in the downstream task, allowing the transferability among the two tasks [Banville et al. 2021b]. The latter approaches are based on differentiable preprocessing. It means that the network will learn the best recombination of channels according to their predictive power. In our current investigation, we are interpreting this challenge as a data augmentation problem, where different electrodes configurations represent different and *augmented* views of the same data. To preserve the spatial consistency across training examples, one can select a subset of electrodes as fixed centroids and, for each training example, sample an electrode in its neighbourhood. This way, one can potentially generate multiple and augmented versions of the same training example simulating electrodes displacements and different head shapes, making the system robust to such variations.

Part V

APPENDICES

Statistical testing

- **SYNOPSIS** Throughout the thesis, we evaluate the statistical significance of the presented results using hypothesis testing. In this chapter, the reader can find further explanations and details about each hypothesis test used in the manuscript.

LOOKING AT DIFFERENCES

The choice of the statistical test depends first of all on the research question we want to answer. In our case, we want to find differences: is there a significant difference among the **PCCs** in the two conditions attended and unattended instrument? Are the **SDR** scores of the proposed method better than those of the baseline? Are we decoding attention better than a random classifier?

If we look at differences, one must understand if the samples are independent (unpaired) or related (paired). When we compare the **SDR** of the baseline with the **SDR** of the proposed method, the samples are paired: one music mixture is separated by both the methods obtaining two **SDRs** which are related. The samples would be unpaired if we tested the baseline and the proposed method on a different set of mixtures, but this is not the case when comparing source separation systems. The same consideration can be made when evaluating the decoding performances: since each mixture contains a different combination of instruments, the probability of randomly choosing one instrument as the attended one will vary between mixtures.

The second aspect to consider when choosing a statistical test is the distribution of the data. Parametric tests such as the *t*-test assume that the data are randomly sampled from a population whose distribution of scores is characterized by a fixed number of parameters, *e.g.*, a normal distribution parametrized by the mean and standard deviation. If we want the outcome of a statistical test to be valid, assumptions on the data distribution must be met. Otherwise, non-parametric or randomization-based tests should be used. In the case of a small sample size, it might be pretty hard to assess whether an assumption is met or not, and it is better to opt for non-parametric or randomization-based tests which do not make any assumption on the distribution of the data.

COMPARING DISTRIBUTIONS OF SCORES

In our evaluation, to compare two distributions of scores (*e.g.*, **PCCs**, linear **SDR**) we opted for two non-parametric tests:

- Wilcoxon signed rank test in the case of paired samples [Conover 1999];
- Mann & Whitney U-test in the case of unpaired samples [Mann and Whitney 1947].

Keywords: Hypothesis testing, non-parametric tests, randomization test.

Resources:

🔗 [Code randomization test](#)

P-VALUE	INTERPRETATION
0.001	HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04	SIGNIFICANT
0.049	
0.050	OH CRAP, REDO CALCULATIONS.
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
0.08	
0.09	
0.099	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS
≥0.1	

FIGURE 1: If all else fails, use “significant at a $p > 0.05$ level” and hope no one notices. Image courtesy of xkcd, number 1478.

COMPARING CLASSIFICATION PERFORMANCE TO CHANCE LEVEL

Instead, in the cases where we needed to evaluate the decoding performances, the statistical significance of classification results (*e.g.*, accuracy, F1 score) compared to chance level was assessed using an adaptation of the computationally-intensive randomization test [Noreen 1989], a non-parametric hypothesis test, which does not make any assumption on the score distribution and can be used also for complex non-linear measures such as F1 score [Yeh 2000]. In our specific case, the test is built by implementing the following procedure: first, we considered a random classifier that, given a test mixture, chooses the attended instrument randomly among the instruments in the given mixture. Then, the performances were computed over the random predictions on the complete test set. This procedure was repeated 10000 times, which resulted in a distribution of the performances. This empirical distribution was then approximated with a theoretical distribution which could be a normal or a *t*-distribution (the one that fits better). Then we evaluated how likely our model's actual performances were to be produced by this artificial distribution of performances obtaining the P-value. Our implementation of the hypothesis test can be found at.⁴

⁴<https://github.com/giorgiacantisani/randomization-test>

Detailed Derivation of the Multiplicative Update Rules for the Contrastive-NMF

- **SYNOPSIS** This Chapter provides the detailed derivation of the Multiplicative Update (MU) rules for the Contrastive-NMF (C-NMF) which was presented in Chapter 4. The material reported here is extracted from the supplementary material accompanying the work in [Cantisani et al. 2021b].

The cost function of the Contrastive-NMF is formulated as:

$$\begin{cases} C(\mathbf{W}, \mathbf{H}) = \underbrace{D_{KL}(\mathbf{X}|\mathbf{W}\mathbf{H})}_{\text{audio factorization}} + \underbrace{\mu\|\mathbf{H}\|_1 + \beta\|\mathbf{W}\|_1}_{\text{sparsity}} - \underbrace{\delta(\|\mathbf{H}_a\mathbf{S}_a^T\|_F^2 - \|\mathbf{H}_u\mathbf{S}_a^T\|_F^2)}_{\text{contrast}} \\ \mathbf{W}, \mathbf{H}, \mathbf{S}_a \geq 0 \\ \|\mathbf{h}_{k\cdot}\|_2 = 1, \|\mathbf{s}_{k\cdot}\|_2 = 1. \end{cases} \quad (6.1)$$

where $\mathbf{X} \in \mathbb{R}_+^{M \times N}$ is the magnitude spectrogram of the mixture, the columns of $\mathbf{W} \in \mathbb{R}_+^{M \times K}$ are interpreted as non-negative audio spectral patterns expected to correspond to different sources and the rows of $\mathbf{H} \in \mathbb{R}_+^{K \times N}$ as their activations. M represents the number of frequency bins, N the number STFT frames and K the number of spectral patterns.

Let us consider a mixture $x(t)$ given by the linear mixing of the attended source $s_a(t)$ and some interferers $s_u(t)$. Let $\mathbf{W}_a \in \mathbb{R}_+^{M \times K_a}$ be a sub-dictionary of \mathbf{W} containing a set of basis vectors representing source $s_a(t)$ and $\mathbf{H}_a \in \mathbb{R}_+^{K_a \times N}$ be their activations. Let $\mathbf{H}_u \in \mathbb{R}_+^{(K-K_a) \times N}$ be the activations of the interference sources. \mathbf{H}_a can be roughly approximated by \mathbf{S}_a reconstructed from the time-lagged EEG response \mathbf{R} , the assumption being that it is likely to be more correlated with the NMF-derived activations of the attended source \mathbf{H}_a than with the ones of the interferers \mathbf{H}_u .

The rows of \mathbf{H} and \mathbf{S}_a ($\mathbf{h}_{k\cdot}$ and $\mathbf{s}_{k\cdot}$ respectively) are normalized in order to minimize the effect of a scale mismatch between the modalities.

Multiplicative Update Rules

To derive the MU rules, one can compute the gradient of the cost function $\nabla C(\theta)$, split it into its negative and positive parts, *i.e.*,

$$\nabla C(\theta) = \nabla_{\theta^+} C(\theta) - \nabla_{\theta^-} C(\theta), \quad (6.2)$$

and build the rules as follows [Lee and Seung 2001; Févotte and Idier 2011]:

$$\theta \leftarrow \theta \otimes \frac{\nabla_{\theta^-} C(\theta)}{\nabla_{\theta^+} C(\theta)}. \quad (6.3)$$

Since the variables are $\theta = \{\mathbf{W}, \mathbf{H}\}$, the MU rules will be:

$$\mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\nabla_{\mathbf{H}^-} C(\mathbf{W}, \mathbf{H})}{\nabla_{\mathbf{H}^+} C(\mathbf{W}, \mathbf{H})}, \quad (6.4)$$

$$\mathbf{W} \leftarrow \mathbf{W} \otimes \frac{\nabla_{\mathbf{W}^-} C(\mathbf{W}, \mathbf{H})}{\nabla_{\mathbf{W}^+} C(\mathbf{W}, \mathbf{H})}. \quad (6.5)$$

Keywords: Contrastive-NMF, Nonnegative matrix factorisation, Multiplicative updates.

Resources:

- Paper
- Code
- Demo

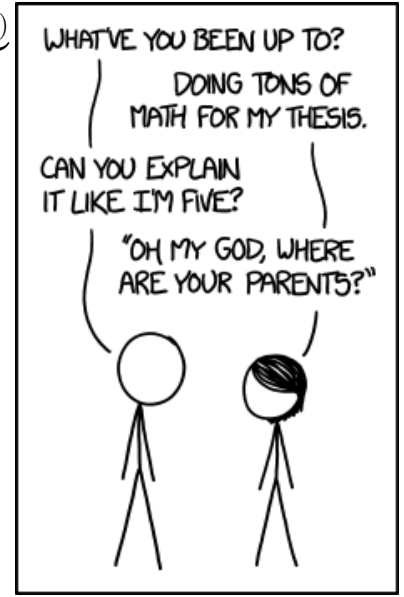


FIGURE 2: “I am Tom, the solution is Jerry: before I catch Jerry, tons of heavy tools fell on me, and it hurt”. Image courtesy of xkcd, number 1364.

UPDATE RULE FOR \mathbf{W}

Since the cost function is completely separable, we can compute the gradient for the KL divergence and for the sparsity constraint separately.

KL Divergence

$$\begin{aligned}
\frac{\partial D_{KL}(\mathbf{X}|\mathbf{WH})}{\partial w_{ij}} &= \frac{\partial}{\partial w_{ij}} \sum_{m=1}^M \sum_{n=1}^N (x_{mn} \log \frac{x_{mn}}{\mathbf{WH}|_{mn}} - x_{mn} + \mathbf{WH}|_{mn}) = \\
&= \sum_{m=1}^M \sum_{n=1}^N \frac{\partial}{\partial w_{ij}} (x_{mn} \log \frac{x_{mn}}{\mathbf{WH}|_{mn}}) + \sum_{m=1}^M \sum_{n=1}^N \frac{\partial}{\partial w_{ij}} (\mathbf{WH}|_{mn}) = \\
&= \sum_{m=1}^M \sum_{n=1}^N \frac{\partial}{\partial w_{ij}} x_{mn} (\log x_{mn} - \log \mathbf{WH}|_{mn}) + \sum_{n=1}^N h_{jn} = \\
&= \sum_{m=1}^M \sum_{n=1}^N x_{mn} \frac{\partial}{\partial w_{ij}} (-\log \mathbf{WH}|_{mn}) + \sum_{n=1}^N h_{jn} = \\
&= \sum_{m=1}^M \sum_{n=1}^N \frac{-x_{mn}}{\mathbf{WH}|_{mn}} \frac{\partial}{\partial w_{ij}} (\mathbf{WH}|_{mn}) + \sum_{n=1}^N h_{jn} = \\
&= \sum_{n=1}^N \frac{-x_{in}}{\mathbf{WH}|_{in}} h_{jn} + \sum_{n=1}^N h_{jn} = \\
&= [-(\Lambda^{-1} \otimes \mathbf{X})\mathbf{H}^T + \mathbf{1}\mathbf{H}^T]_{ij}
\end{aligned} \tag{6.6}$$

- $D_{KL}(p, q) = p \log \frac{p}{q} - p + q$
- $\mathbf{WH}|_{mn} = \sum_k w_{mk} h_{kn}$
- $\Lambda = \mathbf{WH}$
- matrix product derivative:

$$\frac{\partial}{\partial w_{ij}} \mathbf{WH}|_{mn} = \begin{cases} h_{jn} & \text{if } m = i \\ 0 & \text{if } m \neq i \end{cases}$$

Sparsity

$$\frac{\partial \beta \|\mathbf{W}\|_1}{\partial w_{ij}} = \frac{\partial}{\partial w_{ij}} \beta \sum_{m=1}^M \sum_{k=1}^K w_{mk} = \beta \frac{\partial}{\partial w_{ij}} w_{ij} = \beta \tag{6.7}$$

Update rule

$$\boxed{\mathbf{W} \leftarrow \mathbf{W} \otimes \frac{\nabla_{\mathbf{W}} C(\mathbf{W}, \mathbf{H})}{\nabla_{\mathbf{W}} C(\mathbf{W}, \mathbf{H}) + \beta} = \mathbf{W} \otimes \frac{(\Lambda^{-1} \otimes \mathbf{X})\mathbf{H}^T}{\mathbf{1}\mathbf{H}^T + \beta}} \tag{6.8}$$

where \otimes , divisions and exponents denote element-wise operations, $\mathbf{1}$ is a matrix of ones whose size is given by context and $\Lambda = \mathbf{WH}$.

UPDATE RULE FOR \mathbf{H}

As for \mathbf{W} , we can compute the gradient for the KL divergence, the sparsity constraint and for the margin term separately.

KL divergence

$$\begin{aligned}
\frac{\partial D_{KL}(\mathbf{X}|\mathbf{WH})}{\partial h_{ij}} &= \frac{\partial}{\partial h_{ij}} \sum_{m=1}^M \sum_{n=1}^N (x_{mn} \log \frac{x_{mn}}{\mathbf{WH}|_{mn}} - x_{mn} + \mathbf{WH}|_{mn}) = \\
&= \sum_{m=1}^M \sum_{n=1}^N \frac{\partial}{\partial h_{ij}} (x_{mn} \log \frac{x_{mn}}{\mathbf{WH}|_{mn}}) + \sum_{m=1}^M \sum_{n=1}^N \frac{\partial}{\partial h_{ij}} (\mathbf{WH}|_{mn}) = \\
&= \sum_{m=1}^M \sum_{n=1}^N \frac{\partial}{\partial h_{ij}} x_{mn} (\log x_{mn} - \log \mathbf{WH}|_{mn}) + \sum_{m=1}^M w_{mi} = \\
&= \sum_{m=1}^M \sum_{n=1}^N x_{mn} \frac{\partial}{\partial h_{ij}} (-\log \mathbf{WH}|_{mn}) + \sum_{m=1}^M w_{mi} = \\
&= \sum_{m=1}^M \sum_{n=1}^N \frac{-x_{mn}}{\mathbf{WH}|_{mn}} \frac{\partial}{\partial h_{ij}} (\mathbf{WH}|_{mn}) + \sum_{m=1}^M w_{mi} = \\
&= \sum_{m=1}^M \frac{-x_{mj}}{\mathbf{WH}|_{mj}} w_{mi} + \sum_{m=1}^M w_{mi} = \\
&= [-\mathbf{W}^T (\mathbf{X} \otimes \Lambda^{-1}) + \mathbf{W}^T \mathbf{1}]_{ij}
\end{aligned} \tag{6.9}$$

- $D_{KL}(p, q) = p \log \frac{p}{q} - p + q$
- $\mathbf{WH}|_{mn} = \sum_k w_{mk} h_{kn}$
- $\Lambda = \mathbf{WH}$
- matrix product derivative:

$$\frac{\partial}{\partial h_{ij}} \mathbf{WH}|_{mn} = \begin{cases} w_{mi} & \text{if } n = j \\ 0 & \text{if } n \neq j \end{cases}$$

Sparsity constrain

$$\frac{\partial \mu \|\mathbf{H}\|_1}{\partial h_{ij}} = \frac{\partial}{\partial h_{ij}} \mu \sum_{k=1}^K \sum_{n=1}^N h_{kn} = \mu \frac{\partial}{\partial h_{ij}} h_{ij} = \mu \tag{6.10}$$

Contrast term

Recall that the Frobenius norm can be rewritten as:

$$\|\mathbf{X}\|_F = \sqrt{\sum_{i=1}^M \sum_{j=1}^N x_{ij}^2} = \sqrt{\text{Tr}(\mathbf{X}^T \mathbf{X})} \tag{6.11}$$

Since $\mathbf{H}_a \mathbf{S}_a^T$ and $\mathbf{H}_u \mathbf{S}_a^T$ are square matrices, we have:

$$\|\mathbf{H}_a \mathbf{S}_a^T\|_F^2 = \text{Tr}[(\mathbf{H}_a \mathbf{S}_a^T)^T (\mathbf{H}_a \mathbf{S}_a^T)] = \text{Tr}[\mathbf{S}_a \mathbf{H}_a^T \mathbf{H}_a \mathbf{S}_a^T] \tag{6.12}$$

$$\|\mathbf{H}_u \mathbf{S}_a^T\|_F^2 = \text{Tr}[(\mathbf{H}_u \mathbf{S}_a^T)^T (\mathbf{H}_u \mathbf{S}_a^T)] = \text{Tr}[\mathbf{S}_a \mathbf{H}_u^T \mathbf{H}_u \mathbf{S}_a^T] \tag{6.13}$$

The gradient with respect to \mathbf{H} , will be equal to the gradient computed with respect to \mathbf{H}_a for the first K_a rows of \mathbf{H} and equal to the gradient computed with respect to \mathbf{H}_u for the remaining rows:

$$\nabla_{\mathbf{H}} (-\delta(\|\mathbf{H}_a \mathbf{S}_a^T\|_F^2 - \|\mathbf{H}_u \mathbf{S}_a^T\|_F^2)) = \begin{cases} -\delta \nabla_{\mathbf{H}_a} (\|\mathbf{H}_a \mathbf{S}_a^T\|_F^2 - \|\mathbf{H}_u \mathbf{S}_a^T\|_F^2), & \text{if } 1 < k < K_a \\ -\delta \nabla_{\mathbf{H}_u} (\|\mathbf{H}_a \mathbf{S}_a^T\|_F^2 - \|\mathbf{H}_u \mathbf{S}_a^T\|_F^2), & \text{if } K_a + 1 < k < K \end{cases} \tag{6.14}$$

$$\begin{aligned}
\nabla_{\mathbf{H}_a}(\|\mathbf{H}_a \mathbf{S}_a^T\|_F^2 - \|\mathbf{H}_u \mathbf{S}_a^T\|_F^2) &= \nabla_{\mathbf{H}_a} \|\mathbf{H}_a \mathbf{S}_a^T\|_F^2 = \nabla_{\mathbf{H}_a} \text{Tr}[\mathbf{S}_a \mathbf{H}_a^T \mathbf{H}_a \mathbf{S}_a^T] = & \bullet \text{Tr}(\mathbf{ABC}) = \text{Tr}(\mathbf{BCA}) = \\
&= \mathbf{H}_a (\mathbf{S}_a^T \mathbf{S}_a) + \mathbf{H}_a (\mathbf{S}_a^T \mathbf{S}_a)^T = & \text{Tr}(\mathbf{CAB}) \\
&= 2\mathbf{H}_a \mathbf{S}_a^T \mathbf{S}_a & \bullet \nabla_{\mathbf{X}} \text{Tr}(\mathbf{XAX}^T) = \mathbf{X}(\mathbf{A}^T + \mathbf{A}) \\
& & \bullet (\mathbf{X}^T \mathbf{Y})^T = \mathbf{Y}^T \mathbf{X}
\end{aligned} \tag{6.15}$$

$$\begin{aligned}
\nabla_{\mathbf{H}_u}(\|\mathbf{H}_a \mathbf{S}_a^T\|_F^2 - \|\mathbf{H}_u \mathbf{S}_a^T\|_F^2) &= -\nabla_{\mathbf{H}_u} \|\mathbf{H}_u \mathbf{S}_a^T\|_F^2 = -\nabla_{\mathbf{H}_u} \text{Tr}[\mathbf{S}_a \mathbf{H}_u^T \mathbf{H}_u \mathbf{S}_a^T] = \\
&= -(\mathbf{H}_u (\mathbf{S}_a^T \mathbf{S}_a) + \mathbf{H}_u (\mathbf{S}_a^T \mathbf{S}_a)^T) = \\
&= -2\mathbf{H}_u \mathbf{S}_a^T \mathbf{S}_a
\end{aligned} \tag{6.16}$$

Thus, we have:

$$\nabla_{\mathbf{H}}(-\delta(\|\mathbf{H}_a \mathbf{S}_a^T\|_F^2 - \|\mathbf{H}_u \mathbf{S}_a^T\|_F^2)) = \begin{cases} -2\delta \mathbf{H}_a \mathbf{S}_a^T \mathbf{S}_a, & \text{if } 1 < k < K_a \\ +2\delta \mathbf{H}_u \mathbf{S}_a^T \mathbf{S}_a, & \text{if } K_a + 1 < k < K \end{cases} \tag{6.17}$$

Update Rule

$$\boxed{\mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\nabla_{\mathbf{H}^-} C(\mathbf{W}, \mathbf{H})}{\nabla_{\mathbf{H}^+} C(\mathbf{W}, \mathbf{H})} = \mathbf{H} \otimes \frac{\mathbf{W}^T (\mathbf{X} \otimes \Lambda^{-1}) + \delta \mathbf{P}^-}{\mathbf{W}^T \mathbf{1} + \mu + \delta \mathbf{P}^+}} \tag{6.18}$$

where \otimes , divisions and exponents denote element-wise operations, $\mathbf{1}$ is a matrix of ones whose size is given by context and $\Lambda = \mathbf{W}\mathbf{H}$. \mathbf{P}^- , $\mathbf{P}^+ \in \mathbb{R}_+^{K \times N}$ are auxiliary matrices defined as:

$$\mathbf{P}^- = \begin{cases} \mathbf{H}_a \mathbf{S}_a^T \mathbf{S}_a, & \text{if } 1 < k < K_a \\ 0, & \text{if } K_a + 1 < k < K \end{cases} \tag{6.19}$$

$$\mathbf{P}^+ = \begin{cases} 0, & \text{if } 1 < k < K_a \\ \mathbf{H}_u \mathbf{S}_a^T \mathbf{S}_a, & \text{if } K_a + 1 < k < K \end{cases} \tag{6.20}$$

Science dissemination: the MIP-frontiers video communication project

- **SYNOPSIS** This Chapter is about the MIP-frontiers science dissemination project I have coordinated, which led to the release of a short video explaining in simple terms what Music Information Research (MIR) is all about. This part is not strictly related to the research topic of the thesis but the more general problem of science communication and dissemination.

SCIENCE DISSEMINATION

Sharing your research can be very challenging. Sometimes you may need to target a broader audience than simply the colleagues in your particular research field. Colleagues in other communities or disciplines are already less likely to read about your work. When it comes to sharing your research with the general public, things become even more difficult.

There are several reasons why we all should aim to disseminate our research beyond our universities and scientific communities. For instance, it might be essential to explain your research to a general audience because you are doing it thanks to some public funding. In such a case, it is a social duty to inform the citizens about your findings and make your research comprehensible. It's a virtuous circle that produces culture and participation, and in return, can pay for new investments in research.

Another reason is to attract the next generation towards science and your specific research field. This is an aspect that is often underrated because it hasn't an immediate economic and/or social recognition return, but that is critical in the long term. Undergraduate students can orient their education choices and be our future colleagues and enlarge our research community. It's vital then to let them know that your research exists and might be interesting for them. This would also benefit and increase diversity in the community and

Keywords: Science dissemination, Videos about science, Music Information Research.

Resources:

- 🔗 Video
- 🔗 Music




VIDEO ORIENTATION	PROS	CONS
HORIZONTAL 	<ul style="list-style-type: none"> LOOKS NORMAL TO OLD PEOPLE. FORMAT USED BY A CENTURY OF CINEMA 	<ul style="list-style-type: none"> HUMANS ARE TALLER THAN THEY ARE WIDE I'M NOT TURNING MY PHONE SIDELAYS
VERTICAL 	<ul style="list-style-type: none"> HOW MOST NORMAL PEOPLE SHOOT AND WATCH VIDEO NOW SO WE MAY AS WELL ACCEPT IT 	<ul style="list-style-type: none"> HUMAN WORLD IS MOSTLY A HORIZONTAL PLANE.
DIAGONAL 	<ul style="list-style-type: none"> BOLD AND DYNAMIC EQUALLY ANNOYING TO ALL VIEWERS GOOD COMPROMISE 	<ul style="list-style-type: none"> NONE

FIGURE 3: Curious phenomenon: when you are in charge, it comes the time when equally annoying solutions looks easier and funnier. Image courtesy of xkcd, number 2119.

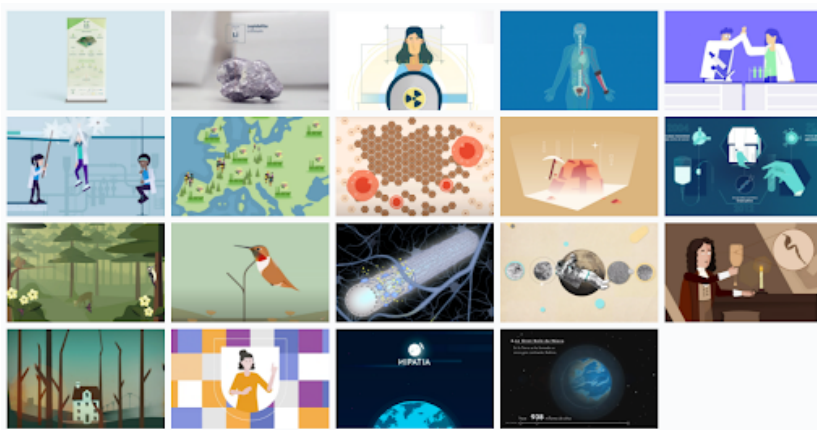


FIGURE 4: Examples of scientific dissemination projects. Image courtesy of Scienseed

reach all those students for whom computer science is not among the options because of societal, demographic, or socioeconomic factors.

In this context, it is still tough for scientists to involve the uninitiated on very specific topics that seem to have almost no connection with their everyday lives. However, many different techniques, tools, and languages have been studied and gradually refined over time. With the increasing amount of information available online, it is becoming more and more important to be concise and attract the audience's attention from the very beginning.

- ▶ **VIDEOS ABOUT SCIENCE** have become more and more popular over the last decade as they are a low-barrier medium to communicate ideas efficiently and effectively. Short videos from 3 to 5 minutes are ideal because they are long enough to explain a concept and sufficiently short for viewers to decide if they are interested. We all have learned about the advantages and disadvantages of this medium during the last year of the pandemic. The format of the conferences has changed, and video abstracts are now a standard. However, video abstracts are intended for peers and not for a broader audience. When disseminating science, complex concepts should be made accessible for the largest audience possible. In such a case, motion graphics and animated storytelling can be a possible solution. Through the process of abstraction in an animated representation, we can effectively simplify the concept we want to transmit. The style, colour palette, transitions, aesthetic and functional choices can all concur to convey the main message.

- ▶ **THE ABSTRACTION PROCESS** is not easy. It takes time, many iterations over the script and many drafts before coming up with something good. You have to learn to work with visual designers who do not know anything about your research. We experienced this when working on the MIP-frontiers video communication project, meant to attract young researchers in our research field. It's very hard to simplify and abstract things you work on every day. It feels like sacrificing many details which are essential to you for the sake of simplicity. Because of that, you have to always keep in mind who's your target audience. In the specific case of this video, there was an additional problem: we needed to cover the most possible areas in Music Information

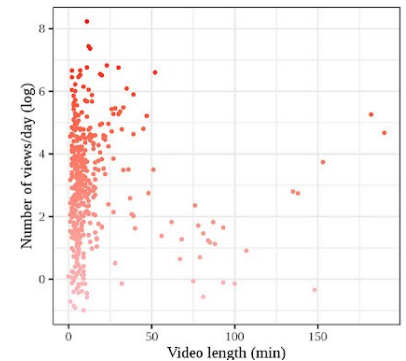


FIGURE 5: Number of views per day (log) x video length (min). Plot courtesy of Velho et al. [Velho et al. 2020].



FIGURE 6: Extract from the MIP-frontiers dissemination video

Processing (MIP), which was quite hard. The trick we found was to trace back the history of a song that an imaginary inhabitant of the future is listening to. We managed to derive a circular story following the song from composition to recording and from distribution to the user experience. Therefore, the music is the backbone of the video, and its choice was crucial.

MAKING-OF

When preparing a motion graphic, you need to provide to the visual designers a *script* (description of the scenes), the *voiceover* (text that an actor needs to read and which describes the scene), and the background *music*. With those three elements, the visual designers built an animation on which you can then give feedback and adapt the voiceover and the music again. This process is reiterated repeatedly until convergence and everyone is happy with the result. In our case, an additional difficulty was that the music wasn't just some "background" music. It was, on the contrary, the absolute protagonist that mainly contributes to conveying the main message. The music evolves throughout the video and changes according to the MIR application we wanted to illustrate. All of this needs a not negligible effort of *synchronization and composition*. In Figure 7, you can see an extract of the script I prepared with the musicians. Regarding the voiceover, we quickly realized how few words can fit a 3-minutes-long video. More importantly, we learned how hard it can be to summarize the vast diversity of research in our community. Moreover, there are synchronization constraints that impose a fixed number of words to express

Scene	Timeline	Voiceover	Description of the action	Sounds / Music
1. Intro: (Mr. Listen)				
Intro	00:00 - 00:03		Spaceship floating in the space and as a subtitle. Zoom in the spaceship. "Somewhere in space in 2080 ..."	1 loop = 4 bar = 12 sec Bass only and fading out / away EFX as the camera zooms out
	00:03 - 00:12	Can you imagine how listening to music might be in the future?	A guy sitting on a space chair in his ultra-modern living room inside the spaceship is listening to a piece of music. Maybe a cat sleeps on his legs (e.g. lofi girl Ahsoka/ Cyberpunk)?	
Transition (Zoom out-Zoom-in)		And what about the process of creating it?	Zoom out of the spaceship and zoom in another galaxy-earth-continent-country-city- dirty rehearsal room where a band is creating a song	
2. Creation: (musicians, MIR avatar)				
Jamming	00:12 - 00:36	Nowadays music results from a creative process that starts with an original idea and culminates in releasing a song. The truth is: creating music can be very hard. Luckily, science can support musicians in such a process.	The band is jamming in the rehearsal room (e.g. garage): a song is sketched, but a lot of errors and noises occur (e.g. wrong notes, wrong keys). Band composition (6 musicians): <ul style="list-style-type: none"> • Guitar • Bass • Drums • Piano • Keyboard • Trumpet Transition (shift to right) on one of the musicians listens to the recordings of that rehearsal and has an idea. He pushes <i>MIRacle</i> "play" (➤) <i>button</i> to summon the MIR	The voiceover starts just before the drums. Drums kick in as in the real song. All instruments attempting to play theme as in a jam 2 loop (8 bar) = 24 sec

FIGURE 7: Draft of the initial script of the MIP-frontiers animation



FIGURE 8: Extract from the MIP-frontiers dissemination video

complex concepts. In the end, we reached a compromise trying to represent as extensively as possible some **MIP** applications.

Once the voiceover, the animation and the music are done, it is not trivial to create the final video anyway. In fact, in addition to a temporal synchronization of events, automation on the volume of the various instruments and the voice are necessary. This operation is always necessary for video production, and the role of a sound engineer is essential for an optimal result. Especially in this work, where music and its evolving parts are the protagonists, this professional figure had a particularly central role in glueing all the components.

SPECIAL THANKS

We really thank **Mandela** (music), **Scienseed** (animation) and **Alberto Di Carlo** (sound engineer) for their great work!

Mandela is an Italian instrumental jazz band from Vicenza. The sound of the band is characterized by a fusion of jazz idioms, rock, world music, psychedelic, and funk. Over the years, the band has performed in several festivals and venues and released 3 full-length albums. These recordings are all available on the major streaming service. Their last release was presented at the festival Rimusicazioni (Bolzano, Italy) and consists of an original soundtrack for “*Grass: A Nation’s Battle for Life*” – one of the earliest documentaries ever produced (1925). For this video, the track *Simple* from the album *Mandela s.t.* was used. The song was remixed and remastered by **Alberto Di Carlo**.

Scienseed is a multifunctional agency for the dissemination of scientific findings. Its founding goal is to promote public engagement in science through all available tools in the Era of IT. We are specialized in the translation of scientific data into different accessible products and activities, aimed at either the scientific community (peers) or the general public (society). We provide support to academic laboratories, research institutes, universities and private institutions to raise public awareness and increase the repercussion of their contribution to science.



Bibliography

- Accou, Bernd, Mohammad Jalilpour Monesi, Jair Montoya, Tom Francart, et al. (2021). “Modeling the relationship between acoustic stimulus and EEG with a dilated convolutional neural network”. In: *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 1175–1179 (cit. on p. 85).
- Akbari, Hassan, Bahar Khalighinejad, Jose L Herrero, Ashesh D Mehta, and Nima Mesgarani (2019). “Towards reconstructing intelligible speech from the human auditory cortex”. In: *Scientific reports* 9.1, p. 874 (cit. on p. 32).
- Akram, Sahar, Jonathan Z Simon, and Behtash Babadi (2017). “Dynamic estimation of the auditory temporal response function from MEG in competing-speaker environments”. In: *IEEE Trans. on Biomedical Engineering* 64.8, pp. 1896–1905 (cit. on pp. 83, 84).
- Akram, Sahar, Jonathan Z Simon, Shihab A Shamma, and Behtash Babadi (2014). “A state-space model for decoding auditory attentional modulation from MEG in a competing-speaker environment”. In: pp. 460–468 (cit. on p. 9).
- An, Xingwei, Johannes Höhne, Dong Ming, and Benjamin Blankertz (2014). “Exploring combinations of auditory and visual stimuli for gaze-independent brain-computer interfaces”. In: *PloS one* 9.10, e111070 (cit. on p. 10).
- Appaji, Jay and Blair Kaneshiro (2018). *Neural tracking of simple and complex rhythms: pilot study and dataset* (cit. on p. 20).
- Aroudi, Ali, Tobias De Taillez, and Simon Doclo (2020). “Improving auditory attention decoding performance of linear and non-linear methods using state-space model”. In: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (cit. on p. 84).
- Aroudi, Ali and Simon Doclo (2019). “Cognitive-driven binaural LCMV beamformer using EEG-based Auditory Attention Decoding”. In: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (cit. on pp. 10, 44).
- (2020). “Cognitive-driven binaural beamforming using EEG-based auditory attention decoding”. In: *IEEE/ACM Trans. on Audio, Speech and Language Processing (TASLP)* 28, pp. 862–875 (cit. on pp. 9, 10, 30, 44).
- Aroudi, Ali, Daniel Marquardt, and Simon Daclo (2018). “EEG-based auditory attention decoding using steerable binaural superdirective beamformer”. In: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (cit. on pp. 10, 44).
- Banville, Hubert, Omar Chehab, Aapo Hyvärinen, Denis-Alexander Engemann, and Alexandre Gramfort (2021a). “Uncovering the structure of clinical EEG signals with self-supervised learning”. In: *Journal of Neural Engineering* 18.4, p. 046020 (cit. on p. 85).
- Banville, Hubert, Sean UN Wood, Chris Aimone, Denis-Alexander Engemann, and Alexandre Gramfort (2021b). “Robust learning from corrupted EEG with dynamic spatial filtering”. In: *arXiv preprint arXiv:2105.12916* (cit. on pp. 85, 86).
- Bittner, Rachel M, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Pablo Bello (2014). “Medleydb: A multitrack dataset for annotation-intensive mir research.” In: *Int. Society for Music Information Retrieval Conf. (ISMIR)* (cit. on pp. 5, 61, 67).
- Blankertz, Benjamin, Steven Lemm, Matthias Treder, Stefan Haufe, and Klaus-Robert Müller (2011). “Single-trial analysis and classification of ERP components - a tutorial”. In: *NeuroImage* 56.2, pp. 814–825 (cit. on p. 34).
- Bravo, Mary J and Ken Nakayama (1992). “The role of attention in different visual-search tasks”. In: *Perception & psychophysics* 51.5, pp. 465–472 (cit. on p. 7).
- Brodbeck, Christian, Alessandro Presacco, and Jonathan Z Simon (2018). “Neural source dynamics of brain responses to continuous stimuli: speech processing from acoustics to comprehension”. In: *NeuroImage* 172, pp. 162–174 (cit. on p. 9).

- Broderick, Michael P, Andrew J Anderson, Giovanni M Di Liberto, Michael J Crosse, and Edmund C Lalor (2018). “Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech”. In: *Current Biology* 28.5, pp. 803–809 (cit. on p. 30).
- Bryan, Nicholas and Gautham Mysore (2013). “An efficient posterior regularized latent variable model for interactive sound source separation”. In: *Int. Conf. on Machine Learning (ICML)* (cit. on pp. 6, 63).
- Bui, Manh-Quan, Viet-Hang Duong, Shih-Pang Tseng, Zhao-Ze Hong, Bo-Chang Chen, Zhi-Wei Zhong, and Jia-Ching Wang (2016). “NMF/NTF-based methods applied for user-guided audio source separation: An overview”. In: *IEEE Int. Conf. on Orange Technologies (ICOT)* (cit. on p. 6).
- Caclin, Anne, Marie-Helene Giard, Bennett K Smith, and Stephen McAdams (2007). “Interactive processing of timbre dimensions: A Garner interference study”. In: *Brain research* 1138, pp. 159–170 (cit. on p. 8).
- Cantisani, Giorgia, Slim Essid, and Gaël Richard (2019a). “EEG-Based Decoding of Auditory Attention to a Target Instrument in Polyphonic Music”. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (cit. on pp. 12, 14, 29).
- Cantisani, Giorgia, Slim Essid, and Gaël Richard (2021a). “EEG-based Decoding of Auditory Attention to a Target Instrument for Neuro-steered Music Source Separation”. In: *journal in preparation* (cit. on pp. 13, 14, 43).
- (2021b). “Neuro-steered music source separation with EEG-based auditory attention decoding and contrastive-NMF”. In: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (cit. on pp. 13, 14, 43, 95).
- Cantisani, Giorgia, Alexey Ozerov, Slim Essid, and Gaël Richard (2021c). “User-guided one-shot deep model adaptation for music source separation”. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (cit. on pp. 13, 14, 61).
- Cantisani, Giorgia, Gabriel Trégoat, Slim Essid, and Gaël Richard (2019b). “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)*, pp. 51–55 (cit. on pp. 12, 14, 19).
- Ceolini, Enea, Jens Hjortkjær, Daniel DE Wong, James O’Sullivan, Vinay S Raghavan, Jose Herrero, Ashesh D Mehta, Shih-Chii Liu, and Nima Mesgarani (2020). “Brain-informed speech separation (BISS) for enhancement of target speaker in multitalker speech perception”. In: *NeuroImage* (cit. on pp. 10, 30, 45, 63).
- Chang, Edward F, Jochem W Rieger, Keith Johnson, Mitchel S Berger, Nicholas M Barbaro, and Robert T Knight (2010). “Categorical speech representation in human superior temporal gyrus”. In: *Nature neuroscience* 13.11, pp. 1428–1432 (cit. on p. 31).
- Chen, Zhuo, Yi Luo, and Nima Mesgarani (2017). “Deep attractor network for single-microphone speaker separation”. In: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (cit. on p. 45).
- Cherry, E Colin (1953). “Some experiments on the recognition of speech, with one and with two ears”. In: *The Journal of the acoustical society of America* 25.5, pp. 975–979 (cit. on p. 6).
- Cheveigné, Alain de, Malcolm Slaney, Søren A Fuglsang, and Jens Hjortkjær (2021). “Auditory stimulus-response modeling with a match-mismatch task”. In: *Journal of Neural Engineering* 18.4, p. 046040 (cit. on p. 85).
- Cheveigné, Alain de, Daniel DE Wong, Giovanni M Di Liberto, Jens Hjortkjær, Malcolm Slaney, and Edmund Lalor (2018). “Decoding the auditory brain with canonical component analysis”. In: *NeuroImage* 172, pp. 206–216 (cit. on p. 85).
- Chew, Elaine (2021). “On making music with heartbeats”. In: *Handbook of Artificial Intelligence for Music*. Springer, pp. 237–261 (cit. on p. 6).
- Chew, Elaine, Peter Taggart, and Pier Lambiase (2019). “Cardiac Response to Live Music Performance: Computing Techniques for Feature Extraction and Analysis”. In: *IEEE Computing in Cardiology (CinC)* (cit. on p. 6).
- Choi, Woosung, Minseok Kim, Jaehwa Chung, and Soonyoung Jung (2021). “LaSAFT: Latent Source Attentive Frequency Transformation for Conditioned Source Separation”. In: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (cit. on p. 61).
- Chung, Hanwook, Eric Plourde, and Benoit Champagne (2016). “Discriminative training of NMF model based on class probabilities for speech enhancement”. In: *IEEE Signal Processing Letters* 23.4 (cit. on p. 48).

- Cirelli, Laura K, Dan Bosnyak, Fiona C Manning, Christina Spinelli, Céline Marie, Takako Fujioka, Ayda Ghahremani, and Laurel J Trainor (2014). “Beat-induced fluctuations in auditory cortical beta-band activity: using EEG to measure age-related changes”. In: *Frontiers in psychology* 5, p. 742 (cit. on p. 9).
- Cohen, David (1968). “Magnetoencephalography: evidence of magnetic fields produced by alpha-rhythm currents”. In: *Science* 161.3843, pp. 784–786 (cit. on p. 7).
- Cong, Fengyu, Anh Huy Phan, Qibin Zhao, Asoke K Nandi, Vinoo Alluri, Petri Toiviainen, Hanna Poikonen, Minna Huotilainen, Andrzej Cichocki, and Tapani Ristaniemi (2012). “Analysis of ongoing EEG elicited by natural music stimuli using nonnegative tensor factorization”. In: *20th European Signal Processing Conf. (EUSIPCO)* (cit. on p. 9).
- Conover, William Jay (1999). *Practical nonparametric statistics*. Vol. 350. John Wiley & Sons (cit. on p. 91).
- Crosse, Michael J, Giovanni M Di Liberto, Adam Bednar, and Edmund C Lalor (2016). “The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli”. In: *Frontiers in human neuroscience* 10, p. 604 (cit. on pp. 9, 10, 20, 29–32).
- Daly, Ian, Nicoletta Nicolaou, Duncan Williams, Faustina Hwang, Alexis Kirke, Eduardo Miranda, and Slawomir J Nasuto (2020). “Neural and physiological data from participants listening to affective music”. In: *Scientific Data* 7.1, pp. 1–7 (cit. on p. 20).
- Das, Neetha, Alexander Bertrand, and Tom Francart (2018). “EEG-based auditory attention detection: boundary conditions for background noise and speaker positions”. In: *Journal of neural engineering* 15.6, p. 066017 (cit. on p. 37).
- Das, Neetha, Tom Francart, and Alexander Bertrand (2020a). *Auditory Attention Detection Dataset KULeuven*. Version 1.1.0. DOI: 10.5281/zenodo.3997352 (cit. on p. 20).
- Das, Neetha, Simon Van Eyndhoven, Tom Francart, and Alexander Bertrand (2017). “EEG-based attention-driven speech enhancement for noisy speech mixtures using N-fold multi-channel Wiener filters”. In: *25th European Signal Processing Conf. (EUSIPCO)* (cit. on pp. 10, 45).
- Das, Neetha, Jeroen Zegers, Tom Francart, Alexander Bertrand, et al. (2020b). “EEG-informed speaker extraction from noisy recordings in neuro-steered hearing aids: linear versus deep learning methods”. In: *BioRxiv* (cit. on pp. 9, 10, 30, 45).
- Dauer, Tysen, Duc T. Nguyen, Nick Gang, Jacek P. Dmochowski, Jonathan Berger, and Blair Kaneshiro (2021). *Naturalistic Music EEG Dataset - Minimalism (NMED-M)*. URL: <https://exhibits.stanford.edu/data/catalog/kt396gb0630> (cit. on p. 20).
- Défossez, Alexandre, Nicolas Usunier, Léon Bottou, and Francis Bach (2019). “Music source separation in the waveform domain”. In: *arXiv preprint:1911.13254* (cit. on pp. 5, 61, 63, 65, 66).
- Deike, Susann, Birgit Gaschler-Markefski, André Brechmann, and Henning Scheich (2004). “Auditory stream segregation relying on timbre involves left auditory cortex”. In: *Neuroreport* 15.9, pp. 1511–1514 (cit. on p. 8).
- Di Carlo, Diego, Ken Déguernel, and Antoine Liutkus (2017). “Gaussian framework for interference reduction in live recordings”. In: *Audio Engineering Society Conf.: 2017 AES Int. Conf. on Semantic Audio*. Audio Engineering Society (cit. on pp. 6, 63).
- Di Liberto, Giovanni M, Michael J Crosse, and Edmund C Lalor (2018). “Cortical measures of phoneme-level speech encoding correlate with the perceived clarity of natural speech”. In: *Eneuro* 5.2 (cit. on p. 30).
- Di Liberto, Giovanni M, Guilhem Marion, and Shihab A Shamma (2021). “The Music of Silence: Part II: Music Listening Induces Imagery Responses”. In: *Journal of Neuroscience* 41.35, pp. 7449–7460 (cit. on p. 9).
- Di Liberto, Giovanni M, James A O’Sullivan, and Edmund C Lalor (2015). “Low-frequency cortical entrainment to speech reflects phoneme-level processing”. In: *Current Biology* 25.19, pp. 2457–2465 (cit. on p. 30).
- Di Liberto, Giovanni M, Claire Pelofi, Roberta Bianco, Prachi Patel, Ashesh D Mehta, Jose L Herrero, Alain de Cheveigné, Shihab Shamma, and Nima Mesgarani (2020a). “Cortical encoding of melodic expectations in human temporal cortex”. In: *Elife* 9, e51784 (cit. on pp. 9, 31).

- Di Liberto, Giovanni M, Claire Pelofi, Shihab Shamma, and Alain de Cheveigné (2020b). “Musical expertise enhances the cortical tracking of the acoustic envelope during naturalistic music listening”. In: *Acoustical Science and Technology* 41.1, pp. 361–364 (cit. on pp. 9, 56).
- Ding, Nai and Jonathan Z Simon (2012). “Emergence of neural encoding of auditory objects while listening to competing speakers”. In: *Proc. Nat. Academy of Sciences* 109.29, pp. 11854–11859 (cit. on p. 9).
- Doesburg, Sam M, Jessica J Green, John J McDonald, and Lawrence M Ward (2012). “Theta modulation of inter-regional gamma synchronization during auditory attention control”. In: *Brain research* 1431, pp. 77–85 (cit. on p. 9).
- Duong, Ngoc QK, Alexey Ozerov, and Louis Chevallier (2014a). “Temporal annotation-based audio source separation using weighted nonnegative matrix factorization”. In: *IEEE Int. Conf. on Consumer Electronics-Berlin (ICCE-Berlin)* (cit. on pp. 6, 63).
- Duong, Ngoc QK, Alexey Ozerov, Louis Chevallier, and Joël Sirot (2014b). “An interactive audio source separation framework based on non-negative matrix factorization”. In: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (cit. on pp. 6, 63).
- Durrieu, Jean-Louis and Jean-Philippe Thiran (2012). “Musical audio source separation based on user-selected F0 track”. In: *Int. Conf. on Latent Variable Analysis and Signal Separation, LVA/ICA*. Springer (cit. on pp. 6, 63).
- Einevoll, Gaute T, Henrik Lindén, Tom Tetzlaff, Szymon Łeski, and Klas H Pettersen (2013). “Local field potentials”. In: *Principles of neural coding* 37 (cit. on p. 7).
- El Badawy, Dalia, Ngoc QK Duong, and Alexey Ozerov (2014). “On-the-fly audio source separation”. In: *IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP)* (cit. on p. 6).
- Ewert, Sebastian, Bryan Pardo, Meinard Müller, and Mark D Plumbley (2014). “Score-informed source separation for musical audio recordings: An overview”. In: *IEEE Signal Processing Magazine* 31.3, pp. 116–124 (cit. on p. 5).
- Ewert, Sebastian and Mark B Sandler (2017). “Structured dropout for weak label and multi-instance learning and its application to score-informed source separation”. In: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (cit. on p. 5).
- Fabiani, Monica, Gabriele Gratton, Demetrios Karis, Emanuel Donchin, et al. (1987). “Definition, identification, and reliability of measurement of the P300 component of the event-related brain potential”. In: *Advances in psychophysiology* 2.S 1, p. 78 (cit. on p. 20).
- Falcon, William et al. (2019). “Pytorch lightning”. In: *GitHub*. Note: <https://github.com/PyTorchLightning/pytorch-lightning> 3, p. 6 (cit. on p. xii).
- Févotte, Cédric and Jérôme Idier (2011). “Algorithms for nonnegative matrix factorization with the β -divergence”. In: *Neural computation* 23.9, pp. 2421–2456 (cit. on pp. 49, 95).
- Févotte, Cédric, Emmanuel Vincent, and Alexey Ozerov (2018). “Single-channel audio source separation with NMF: divergences, constraints and algorithms”. In: *Audio Source Separation*. Springer, pp. 1–24 (cit. on pp. 46–48).
- FitzGerald, Derry (2012). “User assisted separation using tensor factorisations”. In: *20th European Signal Processing Conf. (EUSIPCO)* (cit. on pp. 6, 63).
- Fuglsang, Søren Asp, Torsten Dau, and Jens Hjortkjær (2017). “Noise-robust cortical tracking of attended speech in real-world acoustic scenes”. In: *Neuroimage* 156, pp. 435–444 (cit. on pp. 20, 30, 37).
- Geirnaert, Simon, Tom Francart, and Alexander Bertrand (2021a). “Unsupervised Self-Adaptive Auditory Attention Decoding”. In: *IEEE Journal of Biomedical and Health Informatics* (cit. on pp. 83, 84).
- Geirnaert, Simon, Servaas Vandecappelle, Emina Alickovic, Alain de Cheveigne, Edmund Lalor, Bernd T Meyer, Sina Miran, Tom Francart, and Alexander Bertrand (2021b). “Electroencephalography-based auditory attention decoding: Toward neurosteered hearing devices”. In: *IEEE Signal Processing Magazine* 38.4, pp. 89–102 (cit. on p. 84).

- Ghane, Parisa, Narges Zarnaghinaghsh, and Ulisses Braga-Neto (2021). “Comparison of Classification Algorithms Towards Subject-Specific and Subject-Independent BCI”. In: *IEEE 9th Int. Winter Conf. on Brain-Computer Interface (BCI)* (cit. on p. 83).
- Golumbic, Elana M Zion, Nai Ding, Stephan Bickel, Peter Lakatos, Catherine A Schevon, Guy M McKhann, Robert R Goodman, Ronald Emerson, Ashesh D Mehta, Jonathan Z Simon, et al. (2013). “Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party””. In: *Neuron* 77.5, pp. 980–991 (cit. on p. 32).
- Goydke, Katja N, Eckart Altenmüller, Jörn Möller, and Thomas F Münte (2004). “Changes in emotional tone and instrumental timbre are reflected by the mismatch negativity”. In: *Cognitive Brain Research* 21.3, pp. 351–359 (cit. on p. 8).
- Grais, Emad M and Hakan Erdogan (2013). “Discriminative nonnegative dictionary learning using cross-coherence penalties for single channel source separation.” In: *Interspeech* (cit. on p. 48).
- Gramfort, Alexandre, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti S. Hämäläinen (2013). “MEG and EEG Data Analysis with MNE-Python”. In: *Frontiers in Neuroscience* 7.267, pp. 1–13. DOI: 10.3389/fnins.2013.00267 (cit. on p. xii).
- Guillot, Antoine and Valentin Thorey (2021). “RobustSleepNet: Transfer learning for automated sleep staging at scale”. In: *arXiv preprint arXiv:2101.02452* (cit. on p. 86).
- Han, Cong, James O’Sullivan, Yi Luo, Jose Herrero, Ashesh D Mehta, and Nima Mesgarani (2019). “Speaker-independent auditory attention decoding without access to clean speech sources”. In: *Science advances* 5.5, eaav6134 (cit. on pp. 9, 10, 30, 45).
- Harris, Charles R. et al. (Sept. 2020). “Array programming with NumPy”. In: *Nature* 585.7825, pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: <https://doi.org/10.1038/s41586-020-2649-2> (cit. on p. xii).
- Hennequin, Romain, Anis Khelif, Felix Voituret, and Manuel Moussallam (2020). “Spleeter: a fast and efficient music source separation tool with pre-trained models”. In: *Journal of Open Source Software* 5.50. Deezer Research, p. 2154. DOI: 10.21105/joss.02154. URL: <https://doi.org/10.21105/joss.02154> (cit. on pp. 5, 61).
- Hillyard, Steven A, Robert F Hink, Vincent L Schwent, and Terence W Picton (1973). “Electrical signs of selective attention in the human brain”. In: *Science* 182.4108, pp. 177–180 (cit. on p. 9).
- Hung, Yun-Ning and Alexander Lerch (2020). “Multitask learning for instrument activation aware music source separation”. In: *Int. Society for Music Information Retrieval Conf. (ISMIR)* (cit. on p. 6).
- Hunter, John D (2007). “Matplotlib: A 2D graphics environment”. In: *Computing in science & engineering* 9.03, pp. 90–95 (cit. on p. xii).
- Hyde, Krista L, Isabelle Peretz, and Robert J Zatorre (2008). “Evidence for the role of the right auditory cortex in fine pitch resolution”. In: *Neuropsychologia* 46.2, pp. 632–639 (cit. on p. 8).
- Jäncke, Lutz, Shahram Mirzazade, and Nadim Joni Shah (1999). “Attention modulates activity in the primary and the secondary auditory cortex: a functional magnetic resonance imaging study in human subjects”. In: *Neuroscience letters* 266.2, pp. 125–128 (cit. on p. 9).
- Jasper, Herbert and Wilder Penfield (1949). “Electrocorticograms in man: effect of voluntary movement upon the electrical activity of the precentral gyrus”. In: *Archiv für Psychiatrie und Nervenkrankheiten* 183.1, pp. 163–174 (cit. on p. 7).
- Jeong, Il-Young and Kyogu Lee (2015). “Informed source separation from monaural music with limited binary time-frequency annotation”. In: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (cit. on pp. 6, 63).
- Jing, Longlong and Yingli Tian (2020). “Self-supervised visual feature learning with deep neural networks: A survey”. In: *IEEE Trans. on pattern analysis and machine intelligence* (cit. on p. 85).
- Jonides, John and David E Irwin (1981). “Capturing attention”. In: (cit. on p. 7).

- Kaneshiro, Blair Bohannon (2016). “Toward an objective neurophysiological measure of musical engagement”. PhD thesis. Stanford University (cit. on p. 9).
- Kaneshiro, Blair, Duc T. Nguyen, Jacek P. Dmochowski, Anthony M. Norcia, and Jonathan Berger (2016a). *Naturalistic Music EEG Dataset - Hindi (NMED-H)*. URL: <https://exhibits.stanford.edu/data/catalog/sd922db3535> (cit. on p. 20).
- Kaneshiro, Blair, Duc T Nguyen, Jacek P Dmochowski, Anthony M Norcia, and Jonathan Berger (2016b). “Neurophysiological and behavioral measures of musical engagement”. In: *Proc. 14th Int. Conf. on Music Perception and Cognition* (cit. on p. 9).
- Kaneshiro, Blair, Duc T Nguyen, Anthony M Norcia, Jacek P Dmochowski, and Jonathan Berger (2020). “Natural music evokes correlated EEG responses reflecting temporal structure and beat”. In: *NeuroImage* 214, p. 116559 (cit. on p. 9).
- Kaneshiro, Blair, Duc T. Nguyen, Anthony M. Norcia, Jacek P. Dmochowski, and Jonathan Berger (2021a). *Naturalistic Music EEG Dataset - Elgar (NMED-E)*. URL: <https://exhibits.stanford.edu/data/catalog/pp371jh5722> (cit. on p. 20).
- Kaneshiro, Blair, Duc T Nguyen, Anthony Matthew Norcia, Jacek P Dmochowski, and Jonathan Berger (2021b). “Inter-subject EEG correlation reflects time-varying engagement with natural music”. In: *bioRxiv* (cit. on p. 9).
- Karamatli, Ertuğ, Ali Taylan Cemgil, and Serap Kurbiz (2019). “Audio source separation using variational autoencoders and weak class supervision”. In: *IEEE Signal Process. Lett.* (cit. on p. 6).
- Kawala-Sterniuk, Aleksandra, Natalia Browarska, Amir Al-Bakri, Mariusz Pelc, Jaroslaw Zygarlicki, Michaela Sidikova, Radek Martinek, and Edward Jacek Gorzelanczyk (2021). “Summary of over Fifty Years with Brain-Computer Interfaces—A Review”. In: *Brain Sciences* 11.1, p. 43 (cit. on p. 3).
- Kaya, Emine Merve and Mounya Elhilali (2017). “Modelling auditory attention”. In: *Philosophical Trans. Royal Society B: Biological Sciences* 372.1714, p. 20160101 (cit. on pp. 6, 7).
- Kitamura, Daichi, Nobutaka Ono, Hiroshi Saruwatari, Yu Takahashi, and Kazunobu Kondo (2016). “Discriminative and reconstructive basis training for audio source separation with semi-supervised nonnegative matrix factorization”. In: *IEEE Int. Workshop on Acoustic Signal Enhancement (IWAENC)* (cit. on p. 48).
- Koch, Christof and Naotsugu Tsuchiya (2007). “Attention and consciousness: two distinct brain processes”. In: *Trends in cognitive sciences* 11.1, pp. 16–22 (cit. on p. 7).
- Koelsch, Stefan, Tomas Gunter, Angela D Friederici, and Erich Schröger (2000). “Brain indices of music processing: “non-musicians” are musical”. In: *Journal of cognitive neuroscience* 12.3, pp. 520–541 (cit. on p. 9).
- Koelsch, Stefan, Sebastian Jentschke, Daniela Sammler, and Daniel Mietchen (2007). “Untangling syntactic and sensory processing: An ERP study of music perception”. In: *Psychophysiology* 44.3, pp. 476–490 (cit. on p. 9).
- Koelsch, Stefan, Martin Rohrmeier, Renzo Torrecuso, and Sebastian Jentschke (2013). “Processing of hierarchical syntactic structure in music”. In: *Proceedings of the National Academy of Sciences* 110.38, pp. 15443–15448 (cit. on p. 8).
- Koelstra, Sander, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras (2012). “Deap: A database for emotion analysis; using physiological signals”. In: *IEEE Trans. on Affective Computing* 3.1, pp. 18–31 (cit. on p. 20).
- Kostas, Demetres, Stephane Aroca-Ouellette, and Frank Rudzicz (2021). “BENDR: using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data”. In: *arXiv preprint arXiv:2101.12037* (cit. on p. 85).
- Kumar, BG Vijay, Irene Kotsia, and Ioannis Patras (2012). “Max-margin non-negative matrix factorization”. In: *Image and Vision Computing* 30.4-5 (cit. on p. 48).
- Kumar, Sukhbinder, William Sedley, Kirill V Nourski, Hiroto Kawasaki, Hiroyuki Oya, Roy D Patterson, Matthew A Howard III, Karl J Friston, and Timothy D Griffiths (2011). “Predictive coding and pitch processing in the auditory cortex”. In: *Journal of Cognitive Neuroscience* 23.10, pp. 3084–3094 (cit. on p. 8).
- Lalor, Edmund C and John J Foxe (2010). “Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution”. In: *European journal of neuroscience* 31.1, pp. 189–193 (cit. on p. 30).

- Lalor, Edmund C, Alan J Power, Richard B Reilly, and John J Foxe (2009). “Resolving precise temporal processing properties of the auditory system using continuous stimuli”. In: *Journal of neurophysiology* 102.1, pp. 349–359 (cit. on p. 30).
- Laurberg, Hans, Mikkel N Schmidt, Mads Graesboll Christensen, and Soren Holdt Jensen (2008). “Structured non-negative matrix factorization with sparsity patterns”. In: *IEEE 42nd Asilomar Conf. on Signals, Systems and Computers* (cit. on pp. 6, 63).
- Lee, Daniel D and H Sebastian Seung (1999). “Learning the parts of objects by non-negative matrix factorization”. In: *Nature* 401.6755, p. 788 (cit. on p. 46).
- (2001). “Algorithms for non-negative matrix factorization”. In: *Advances in neural information processing systems*, pp. 556–562 (cit. on pp. 49, 95).
- Lefevre, Augustin, Francis Bach, and Cédric Févotte (2012). “Semi-supervised NMF with time-frequency annotations for single-channel source separation”. In: *Int. Society for Music Information Retrieval Conf. (ISMIR)* (cit. on pp. 6, 63).
- Lefèvre, Augustin, François Glineur, and P-A Absil (2014). “A convex formulation for informed source separation in the single channel setting”. In: *Neurocomputing* (cit. on pp. 6, 63).
- Li, Tingle, Jiawei Chen, Haowen Hou, and Ming Li (2021). “Sams-net: A sliced attention-based neural network for music source separation”. In: *IEEE 12th Int. Symposium on Chinese Spoken Language Processing (ISCSLP)* (cit. on p. 61).
- Lipschutz, Brigitte, Régine Kolinsky, Philippe Damhaut, David Wikler, and Serge Goldman (2002). “Attention-dependent changes of activation and connectivity in dichotic listening”. In: *Neuroimage* 17.2, pp. 643–656 (cit. on p. 9).
- Liu, Liyuan, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han (2020). “On the variance of the adaptive learning rate and beyond”. In: *Int. Conf. on Learning Representations (ICLR)* (cit. on p. 66).
- Liutkus, Antoine, Jean-Louis Durrieu, Laurent Daudet, and Gaël Richard (2013). “An overview of informed audio source separation”. In: *IEEE Int. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)* (cit. on p. 5).
- Losorelli, Steven, Duc T Nguyen, Jacek P Dmochowski, and Blair Kaneshiro (2017). “NMED-T: A tempo-focused dataset of cortical and behavioral responses to naturalistic music”. In: URL: <https://exhibits.stanford.edu/data/catalog/jn859kj8079> (cit. on p. 20).
- Luo, Yi, Zhuo Chen, John R Hershey, Jonathan Le Roux, and Nima Mesgarani (2017). “Deep clustering and conventional networks for music separation: Stronger together”. In: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (cit. on p. 45).
- Luo, Yi, Zhuo Chen, and Nima Mesgarani (2018). “Speaker-independent speech separation with deep attractor network”. In: *IEEE/ACM Trans. on Audio, Speech and Language Processing (TASLP)* 26.4, pp. 787–796 (cit. on p. 45).
- Luo, Yi and Nima Mesgarani (2019). “Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation”. In: *IEEE/ACM Trans. on Audio, Speech and Language Processing (TASLP)* 27.8, pp. 1256–1266 (cit. on pp. 61, 63, 65).
- Mann, Henry B and Donald R Whitney (1947). “On a test of whether one of two random variables is stochastically larger than the other”. In: *The annals of mathematical statistics*, pp. 50–60 (cit. on p. 91).
- Marion, Guilhem, Giovanni M Di Liberto, and Shihab A Shamma (2021). “The Music of Silence: Part I: Responses to Musical Imagery Encode Melodic Expectations and Acoustics”. In: *Journal of Neuroscience* 41.35, pp. 7435–7448 (cit. on p. 9).
- McFee, Brian, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto (2015). “librosa: Audio and music signal analysis in python”. In: *Proc. 14th python in science Conf.* Pp. 18–25 (cit. on p. xii).

- McKinney, Wes (2010). “Data Structures for Statistical Computing in Python”. In: *Proceedings of the 9th Python in Science Conference*. Ed. by Stéfan van der Walt and Jarrod Millman, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a (cit. on p. xii).
- Mesgarani, Nima and Edward F Chang (2012). “Selective cortical representation of attended speaker in multi-talker speech perception”. In: *Nature* 485.7397, p. 233 (cit. on pp. 9, 10, 30–32, 36).
- Mesgarani, Nima, Stephen V David, Jonathan B Fritz, and Shihab A Shamma (2009). “Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex”. In: *Journal of neurophysiology* (cit. on pp. 9, 10, 30–32).
- Miran, Sina, Sahar Akram, Alireza Sheikhattar, Jonathan Z Simon, Tao Zhang, and Behtash Babadi (2018). “Real-time tracking of selective auditory attention from M/EEG: A bayesian filtering approach”. In: *Frontiers in neuroscience* 12, p. 262 (cit. on pp. 83, 84).
- Müller, Meinard (2007). *Information retrieval for music and motion*. Vol. 2. Springer (cit. on p. 6).
- Nakano, Tomoyasu, Yuki Koyama, Masahiro Hamasaki, and Masataka Goto (2020). “Interactive deep singing-voice separation based on human-in-the-loop adaptation”. In: *Proc. 25th Int. Conf. on Intelligent User Interfaces (IUI)* (cit. on pp. 6, 63).
- Nan, Yun and Angela D Friederici (2013). “Differential roles of right temporal cortex and Broca’s area in pitch processing: evidence from music and Mandarin”. In: *Human brain mapping* 34.9, pp. 2045–2054 (cit. on p. 8).
- Nasiri, Samaneh and Gari D Clifford (2020). “Attentive adversarial network for large-scale sleep staging”. In: *Machine Learning for Healthcare Conf. PMLR*, pp. 457–478 (cit. on p. 86).
- Noreen, Eric W (1989). *Computer-intensive methods for testing hypotheses*. Wiley New York (cit. on pp. 34, 53, 66, 92).
- O’Sullivan, James A, Richard B Reilly, and Edmund C Lalor (2015). “Improved decoding of attentional selection in a cocktail party environment with EEG via automatic selection of relevant independent components”. In: *37th Ann. Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBC)* (cit. on p. 30).
- O’Sullivan, James, Zhuo Chen, Sameer A Sheth, Guy McKhann, Ashesh D Mehta, and Nima Mesgarani (2017). “Neural decoding of attentional selection in multi-speaker environments without access to separated sources”. In: *39th Ann. Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBC)* (cit. on pp. 10, 45).
- Ofner, André and Sebastian Stober (2018). “Shared generative representation of auditory concepts and EEG to reconstruct perceived and imagined music”. In: (cit. on p. 9).
- Ogawa, Seiji, Tso-Ming Lee, Alan R Kay, and David W Tank (1990). “Brain magnetic resonance imaging with contrast dependent on blood oxygenation”. In: *Proc. National Academy of Sciences* 87.24, pp. 9868–9872 (cit. on p. 8).
- Okada, Kayoko, Feng Rong, Jon Venezia, William Matchin, I-Hui Hsieh, Kourosh Saberi, John T Serences, and Gregory Hickok (2010). “Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech”. In: *Cerebral Cortex* 20.10, pp. 2486–2495 (cit. on p. 31).
- O’Sullivan, James A, Alan J Power, Nima Mesgarani, Siddharth Rajaram, John J Foxe, Barbara G Shinn-Cunningham, Malcolm Slaney, Shihab A Shamma, and Edmund C Lalor (2014). “Attentional selection in a cocktail party environment can be decoded from single-trial EEG”. In: *Cerebral Cortex* 25.7, pp. 1697–1706 (cit. on pp. 8–10, 20, 29–33, 35, 36, 83).
- Ozerov, Alexey, Cédric Févotte, Raphaël Blouet, and Jean-Louis Durrieu (2011). “Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation”. In: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (cit. on pp. 6, 63).
- Parekh, Sanjeel, Slim Essid, Alexey Ozerov, Ngoc QK Duong, Patrick Pérez, and Gaël Richard (2017). “Guiding audio source separation by video object information”. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (cit. on pp. 5, 55).
- Pariente, Manuel, Samuele Cornell, Joris Cosentino, Sunit Sivasankaran, Efthymios Tzinis, Jens Heitkaemper, Michel Olvera, Fabian-Robert Stöter, Mathieu Hu, Juan M. Martín-Doñas, David Ditter, Ariel Frank, Antoine

- Deleforge, and Emmanuel Vincent (2020). “Asteroid: the PyTorch-based audio source separation toolkit for researchers”. In: *Proc. Interspeech* (cit. on p. xii).
- Pasley, Brian N, Stephen V David, Nima Mesgarani, Adeen Flinker, Shihab A Shamma, Nathan E Crone, Robert T Knight, and Edward F Chang (2012). “Reconstructing speech from human auditory cortex”. In: *PLoS biology* 10.1, e1001251 (cit. on pp. 9, 10, 30–32).
- Paszke, Adam et al. (2019). “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., pp. 8024–8035. URL: <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf> (cit. on p. xii).
- Pearce, Marcus T and Geraint A Wiggins (2012). “Auditory expectation: the information dynamics of music perception and cognition”. In: *Topics in cognitive science* 4.4, pp. 625–652 (cit. on p. 9).
- Pearce, Marcus Thomas (2005). “The construction and evaluation of statistical models of melodic structure in music perception and composition”. PhD thesis. City University London (cit. on p. 9).
- Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. (2011). “Scikit-learn: Machine learning in Python”. In: *the Journal of machine Learning research* 12, pp. 2825–2830 (cit. on p. xii).
- Peelle, Jonathan E, Ingrid Johnsrude, and Matthew H Davis (2010). “Hierarchical processing for speech in human auditory cortex and beyond”. In: *Frontiers in human neuroscience* 4, p. 51 (cit. on p. 31).
- Plack, Christopher J, Daphne Barker, and Deborah A Hall (2014). “Pitch coding and pitch processing in the human brain”. In: *Hearing Research* 307, pp. 53–64 (cit. on p. 8).
- Pu, Wenqiang, Jinjun Xiao, Tao Zhang, and Zhi-Quan Luo (2019). “A joint auditory attention decoding and adaptive binaural beamforming algorithm for hearing devices”. In: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (cit. on pp. 10, 30, 45, 63).
- Pu, Wenqiang, Peng Zan, Jinjun Xiao, Tao Zhang, and Zhi-Quan Luo (2020). “Evaluation of Joint Auditory Attention Decoding and Adaptive Binaural Beamforming Approach for Hearing Devices with Attention Switching”. In: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (cit. on p. 45).
- Rafii, Zafar, Antoine Liutkus, and Bryan Pardo (2015). “A simple user interface system for recovering patterns repeating in time and frequency in mixtures of sounds”. In: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (cit. on pp. 6, 63).
- Rafii, Zafar, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner (2017). *The MUSDB18 corpus for music separation*. DOI: 10.5281/zenodo.1117372 (cit. on pp. 5, 61, 66, 67).
- Rohrmeier, Martin and Ian Cross (2008). “Statistical properties of tonal harmony in Bach’s chorales”. In: *Proc. 10th Int. Conf. on Music Perception and Cognition*. Vol. 6. Hokkaido University Sapporo, Japan, pp. 619–627 (cit. on p. 9).
- Roy, Yannick, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert (2019). “Deep learning-based electroencephalography analysis: a systematic review”. In: *Journal of neural engineering* 16.5, p. 051001 (cit. on p. 84).
- Saeed, Aaqib, David Grangier, Olivier Pietquin, and Neil Zeghidour (2021). “Learning from heterogeneous eeg signals with differentiable channel reordering”. In: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (cit. on p. 86).
- Samuel, David, Aditya Ganeshan, and Jason Naradowsky (2020). “Meta-learning extractors for music source separation”. In: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (cit. on p. 61).
- Sareen, Ekansh, Lakshya Singh, Blessin Varkey, Krishnaveni Achary, and Anubha Gupta (2020). “EEG dataset of individuals with intellectual and developmental disorder and healthy controls under rest and music stimuli”. In: *Data in brief* 30, p. 105488 (cit. on p. 20).

- Sawata, Ryosuke, Stefan Uhlich, Shusuke Takahashi, and Yuki Mitsufuji (2021). “All for One and One for All: Improving Music Separation by Bridging Networks”. In: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (cit. on p. 61).
- Schaefer, Rebecca S, Peter Desain, and Jason Farquhar (2013). “Shared processing of perception and imagery of music in decomposed EEG”. In: *Neuroimage* 70, pp. 317–326 (cit. on p. 9).
- Schirrmester, Robin Tibor, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball (2017). “Deep learning with convolutional neural networks for EEG decoding and visualization”. In: *Human Brain Mapping*. ISSN: 1097-0193. DOI: 10.1002/hbm.23730. URL: <http://dx.doi.org/10.1002/hbm.23730> (cit. on p. xii).
- Schulze-Forster, Kilian, Clément Doire, Gaël Richard, and Roland Badeau (2019). “Weakly informed audio source separation”. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (cit. on pp. 5, 66).
- Seetharaman, Prem, Gordon Wichern, Shrikant Venkataramani, and Jonathan Le Roux (2019). “Class-conditional embeddings for music source separation”. In: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (cit. on p. 6).
- Seichepine, Nicolas, Slim Essid, Cédric Févotte, and Olivier Cappé (2014). “Soft Nonnegative Matrix Co-Factorization.” In: *IEEE Trans. Signal Processing* 62.22, pp. 5940–5949 (cit. on p. 48).
- Slizovskaia, Olga, Leo Kim, Gloria Haro, and Emilia Gomez (2019). “End-to-end sound source separation conditioned on instrument labels”. In: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (cit. on p. 6).
- Smaragdis, Paris and Gautham J Mysore (2009). “Separation by “humming”: User-guided sound extraction from monophonic mixtures”. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (cit. on pp. 6, 63).
- Steinkamp, Simon (2020). “pymtrf”. In: *GitHub*. URL: <https://github.com/SRSteinkamp/pymtrf> (cit. on p. xii).
- Stober, Sebastian, Thomas Prätzlich, and Meinard Müller (2016). “Brain Beats: Tempo Extraction from EEG Data.” In: *Int. Society for Music Information Retrieval Conf. (ISMIR)* (cit. on p. 9).
- Stober, Sebastian, Avital Sternin, Adrian M Owen, and Jessica A Grahn (2015). “Towards Music Imagery Information Retrieval: Introducing the OpenMIIR Dataset of EEG Recordings from Music Perception and Imagination.” In: *Int. Society for Music Information Retrieval Conf. (ISMIR)* (cit. on p. 20).
- Stoller, Daniel, Sebastian Ewert, and Simon Dixon (2018a). “Jointly detecting and separating singing voice: A multi-task approach”. In: *Int. Conf. on Latent Variable Analysis and Signal Separation, LVA/ICA*. Springer (cit. on p. 6).
- (2018b). “Wave-u-net: A multi-scale neural network for end-to-end audio source separation”. In: *Int. Society for Music Information Retrieval Conf. (ISMIR)* (cit. on p. 5).
- Stöter, Fabian-Robert, Antoine Liutkus, and Nobutaka Ito (2018). “The 2018 signal separation evaluation campaign”. In: *Int. Conf. on Latent Variable Analysis and Signal Separation*. Springer (cit. on pp. xii, 53, 61, 63, 66).
- Stöter, Fabian-Robert, Stefan Uhlich, Antoine Liutkus, and Yuki Mitsufuji (2019). “Open-unmix-a reference implementation for music source separation”. In: *Journal of Open Source Software* (cit. on pp. 5, 61, 63).
- Sturm, Irene (2016). “Analyzing the perception of natural music with EEG and ECoG”. PhD thesis. TU Berlin (cit. on pp. 6, 8, 9).
- Sturm, Irene, Sven Dähne, Benjamin Blankertz, and Gabriel Curio (2015a). “Multi-variate EEG analysis as a novel tool to examine brain responses to naturalistic music stimuli”. In: *PLoS one* 10.10, e0141281 (cit. on p. 9).
- Sturm, Irene, Matthias Treder, Daniel Miklody, Hendrik Purwins, Sven Dähne, Benjamin Blankertz, and Gabriel Curio (2015b). “Extracting the neural representation of tone onsets for separate voices of ensemble music using multivariate EEG analysis.” In: *Psychomusicology: Music, Mind, and Brain* 25.4, p. 366 (cit. on p. 9).

- Swaminathan, Rupak Vignesh and Alexander Lerch (2019). “Improving singing voice separation using attribute-aware deep network”. In: *IEEE Int. Workshop on Multilayer Music Representation and Processing (MMRP)* (cit. on p. 6).
- Takahashi, Naoya, Nabarun Goswami, and Yuki Mitsufuji (2018). “Mmdenselstm: An efficient combination of convolutional and recurrent neural networks for audio source separation”. In: *IEEE 16th Int. Workshop on Acoustic Signal Enhancement (IWAENC)* (cit. on p. 61).
- Takahashi, Naoya and Yuki Mitsufuji (2020). “D3net: Densely connected multidilated densenet for music source separation”. In: *arXiv preprint arXiv:2010.01733* (cit. on p. 61).
- Temperley, David (2008). “A probabilistic model of melody perception”. In: *Cognitive Science* 32.2, pp. 418–444 (cit. on p. 9).
- Temperley, David and Trevor de Clercq (2013). “Statistical analysis of harmony and melody in rock music”. In: *Journal of New Music Research* 42.3, pp. 187–204 (cit. on p. 9).
- Thaut, Michael H (2005). “Rhythm, human temporality, and brain function”. In: *Musical communication*, pp. 171–191 (cit. on p. 9).
- Tóth, Brigitta, Dávid Farkas, Gábor Urbán, Orsolya Szalárdy, Gábor Orosz, László Hunyadi, Botond Hajdu, Annamária Kovács, Beáta Tünde Szabó, Lidia B Shestopalova, et al. (2019). “Attention and speech-processing related functional brain networks activated in a multi-speaker environment”. In: *PloS one* 14.2, e0212754 (cit. on p. 9).
- Treder, Matthias S, Hendrik Purwins, Daniel Miklody, Irene Sturm, and Benjamin Blankertz (2014). “Decoding auditory attention to instruments in polyphonic music using single-trial EEG classification”. In: *Journal of neural engineering* 11.2, p. 026009 (cit. on pp. 8–10, 20, 30).
- Tufte, Edward R and Peter R Graves-Morris (1983). *The visual display of quantitative information*. Vol. 2. 9. Graphics press Cheshire, CT (cit. on p. 15).
- Turatto, Massimo (2006). *notes and lessons*. Department of Cognitive Science, University of Trento, Italy (cit. on pp. 6, 7).
- Van Eyndhoven, Simon, Tom Francart, and Alexander Bertrand (2017). “EEG-Informed Attended Speaker Extraction From Recorded Speech Mixtures With Application in Neuro-Steered Hearing Prostheses.” In: *IEEE Trans. Biomed. Engineering* 64.5, pp. 1045–1056 (cit. on pp. 10, 30, 44).
- Velho, Raphaela Martins, Amanda Merian Freitas Mendes, and Caio Lucidius Naberezny Azevedo (2020). “Communicating science with YouTube videos: how nine factors relate to and affect video views”. In: *Frontiers in Communication*, p. 72 (cit. on p. 102).
- Vincent, Emmanuel, Nancy Bertin, Rémi Gribonval, and Frédéric Bimbot (2014). “From blind to guided audio source separation: How models and side information can improve the separation of sound”. In: *IEEE Signal Processing Magazine* 31.3, pp. 107–115 (cit. on p. 48).
- Vincent, Emmanuel, Rémi Gribonval, and Cédric Févotte (2006). “Performance measurement in blind audio source separation”. In: *IEEE/ACM Trans. on Audio, Speech and Language Processing (TASLP)* 14.4 (cit. on pp. 53, 66).
- Vincent, Emmanuel, Tuomas Virtanen, and Sharon Gannot (2018). *Audio source separation and speech enhancement*. John Wiley & Sons (cit. on pp. 47, 48).
- Virtanen, Pauli, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. (2020). “SciPy 1.0: fundamental algorithms for scientific computing in Python”. In: *Nature methods* 17.3, pp. 261–272 (cit. on p. xii).
- Virtanen, Tuomas, Annamaria Mesáros, and Matti Ryynänen (2008). “Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music.” In: *Interspeech* (cit. on p. 5).
- Walzer, Daniel A (2017). “Independent music production: how individuality, technology and creative entrepreneurship influence contemporary music industry practices”. In: *Creative Industries Journal* 10.1, pp. 21–39 (cit. on p. 3).

- Wang, Liting, Xintao Hu, Meng Wang, Jinglei Lv, Junwei Han, Shijie Zhao, Qinglin Dong, Lei Guo, and Tianming Liu (2017). “Decoding dynamic auditory attention during naturalistic experience”. In: *14th IEEE Int. Symposium on Biomedical Imaging (ISBI)* (cit. on p. 8).
- Waskom, Michael L. (2021). “seaborn: statistical data visualization”. In: *Journal of Open Source Software* 6.60, p. 3021. DOI: 10.21105/joss.03021. URL: <https://doi.org/10.21105/joss.03021> (cit. on p. xii).
- Weber, Marc (2020). “statannot”. In: *GitHub*. URL: <https://github.com/webermarcolivier/statannot> (cit. on p. xii).
- Weninger, Felix, Jonathan Le Roux, John R Hershey, and Shinji Watanabe (2014). “Discriminative NMF and its application to single-channel source separation”. In: *15th Ann. Conf. Int. Speech Communication Association* (cit. on p. 48).
- Williams, D and Eduardo R Miranda (2018). “BCI for music making: then, now, and next”. In: *Brain–Computer Interfaces Handbook: Technological and Theoretical Advances*. CRC Press (cit. on p. 3).
- Woldorff, Marty G, Christopher C Gallen, Scott A Hampson, Steven A Hillyard, Christo Pantev, David Sobel, and Floyd E Bloom (1993). “Modulation of early sensory processing in human auditory cortex during auditory selective attention”. In: *Proc. National Academy of Sciences* 90.18, pp. 8722–8726 (cit. on p. 9).
- Woldorff, Marty G and Steven A Hillyard (1991). “Modulation of early auditory processing during selective listening to rapidly presented tones”. In: *Electroencephalography and clinical neurophysiology* 79.3, pp. 170–191 (cit. on p. 9).
- Wolpaw, Jonathan R and E Winter Wolpaw (2012). “Brain-computer interfaces: something new under the sun”. In: *Brain-computer interfaces: principles and practice* 14 (cit. on p. 3).
- Wong, Daniel DE, Søren A Asp Fuglsang, Jens Hjortkjær, Enea Ceolini, Malcolm Slaney, and Alain de Cheveigné (2018). “A Comparison of Temporal Response Function Estimation Methods for Auditory Attention Decoding”. In: *bioRxiv*, p. 281345 (cit. on p. 34).
- World Medical Association (2013). *World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects*. DOI: 10.1001/jama.2013.281053 (cit. on p. 21).
- Wu, Dongrui, Vernon J Lawhern, W David Hairston, and Brent J Lance (2016). “Switching EEG headsets made easy: Reducing offline calibration effort using active weighted adaptation regularization”. In: *IEEE Trans. on Neural Systems and Rehabilitation Engineering* 24.11, pp. 1125–1137 (cit. on p. 85).
- Yeh, Alexander (2000). “More accurate tests for the statistical significance of result differences”. In: *Proc. 18th Conf. on Computational linguistics*. Association for Computational Linguistics (cit. on pp. 34, 92).
- Yilmaz, Ozgur and Scott Rickard (2004). “Blind separation of speech mixtures via time-frequency masking”. In: *IEEE Trans. on signal processing* 52.7 (cit. on p. 47).
- Yuan, Ye, Guangxu Xun, Qiuling Suo, Kebin Jia, and Aidong Zhang (2017). “Wave2vec: Learning deep representations for biosignals”. In: *IEEE Int. Conf. on Data Mining (ICDM)* (cit. on p. 85).
- Zhang, Michael, James Lucas, Jimmy Ba, and Geoffrey E Hinton (2019). “Lookahead optimizer: k steps forward, 1 step back”. In: (cit. on p. 66).
- Zuk, Nathaniel J, Jeremy W Murphy, Richard B Reilly, and Edmund C Lalor (2021). “Envelope reconstruction of speech and music highlights stronger tracking of speech at low frequencies”. In: *PLoS computational biology* 17.9, e1009358 (cit. on p. 10).
- Zuk, Nathaniel J, Emily S Teoh, and Edmund C Lalor (2020). “EEG-based classification of natural sounds reveals specialized responses to speech and music”. In: *NeuroImage* 210, p. 116558 (cit. on p. 10).

Titre : Séparation de sources musicales neuroguidée

Mots clés : Séparation de sources audio, décodage de l'attention auditive, EEG, traitement multimodal

Résumé : Dans cette thèse, nous abordons le défi de l'utilisation d'interfaces cerveau-machine (ICM) sur l'application spécifique de la séparation de sources musicales qui vise à isoler les instruments individuels qui sont mélangés dans un enregistrement de musique. Ce problème a été étudié pendant des décennies, mais sans jamais considérer les ICM comme un moyen possible de guider et d'informer les systèmes de séparation. Plus précisément, nous avons étudié comment l'activité neuronale caractérisée par des signaux électroencéphalographiques (EEG) reflète des informations sur la source à laquelle on porte son attention et comment nous pouvons l'utiliser pour informer un système de séparation de sources.

Tout d'abord, nous avons étudié le problème du décodage par l'EEG de l'attention auditive d'un instrument spécifique dans une pièce musicale polyphonique, en montrant que l'EEG suit les caractéristiques musicales pertinentes qui sont fortement corrélées avec la représentation temps-fréquence de la source à laquelle on porte l'attention et seulement faiblement corrélées avec les autres. Ensuite, nous avons exploité ce "contraste" pour informer un modèle de séparation de sources non supervisé basé sur une nouvelle variante de factorisation en matrices positives (NMF), ap-

pelée *contrastive-NMF (C-NMF)* et séparer automatiquement la source à laquelle on porte l'attention.

La NMF non supervisée est une approche efficace dans de telles applications ne disposant pas ou peu de données d'apprentissage, comme c'est le cas dans des scénarios nécessitant des enregistrements EEG. En effet, les jeux de données EEG liés à la musique disponibles sont coûteux et longs à acquérir, ce qui exclut la possibilité d'aborder le problème par des approches d'apprentissage profond entièrement supervisées. Dans la dernière partie de la thèse, nous avons exploré des stratégies d'apprentissage alternatives. Plus précisément, nous avons étudié la possibilité d'adapter un modèle de séparation de sources de l'état de l'art à un mélange spécifique en utilisant les activations temporelles de sources dérivées de l'activité neuronale de l'utilisateur au moment du test. Cette approche peut être considérée comme étant "à adaptation unitaire" (*one-shot*), car l'adaptation agit uniquement sur une instance de chanson.

Nous avons évalué les approches proposées sur les jeux de données MAD-EEG qui a été spécifiquement assemblé pour cette étude, obtenant des résultats encourageants, en particulier dans les cas difficiles où les modèles non informés sont mis à mal.

Title : Neuro-steered music source separation

Keywords : Music source separation, Auditory attention decoding, EEG, Multimodal processing

Abstract : In this PhD thesis, we address the challenge of integrating Brain-Computer Interfaces (BCI) and music technologies on the specific application of music source separation, which is the task of isolating individual sound sources that are mixed in the audio recording of a musical piece. This problem has been investigated for decades, but never considering BCI as a possible way to guide and inform separation systems. Specifically, we explored how the neural activity characterized by electroencephalographic signals (EEG) reflects information about the attended instrument and how we can use it to inform a source separation system.

First, we studied the problem of EEG-based auditory attention decoding of a target instrument in polyphonic music, showing that the EEG tracks musically relevant features which are highly correlated with the time-frequency representation of the attended source and only weakly correlated with the unattended one. Second, we leveraged this "contrast" to inform an unsupervised source separation model based on a novel non-negative matrix factorisation (NMF) variant, named

contrastive-NMF (C-NMF) and automatically separate the attended source.

Unsupervised NMF represents a powerful approach in such applications with no or limited amounts of training data as when neural recording is involved. Indeed, the available music-related EEG datasets are still costly and time-consuming to acquire, precluding the possibility of tackling the problem with fully supervised deep learning approaches. Thus, in the last part of the thesis, we explored alternative learning strategies to alleviate this problem. Specifically, we propose to adapt a state-of-the-art music source separation model to a specific mixture using the time activations of the sources derived from the user's neural activity. This paradigm can be referred to as *one-shot* adaptation, as it acts on the target song instance only.

We conducted an extensive evaluation of both the proposed system on the MAD-EEG dataset which was specifically assembled for this study obtaining encouraging results, especially in difficult cases where non-informed models struggle.