Cortical Auditory Attention Decoding for Music Listening: Performance and Differences with Speech Listening

Simon, Adèle Maryse Danièle

DOI (link to publication from Publisher):
10.54337/aau548864023

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

CORTICAL AUDITORY ATTENTION DECODING FOR MUSIC LISTENING: PERFORMANCE AND DIFFERENCES WITH SPEECH LISTENING

BY
ADÈLE SIMON

DISSERTATION SUBMITTED 2023

AALBORG UNIVERSITY
DENMARK

Cortical Auditory Attention Decoding for Music Listening: Performance and Differences with Speech Listening

Ph.D. Dissertation

Adèle Simon

Dissertation submitted June 23, 2023

Dissertation submitted: June 23, 2023

PhD supervisor: Prof. Jan Østergaard
Aalborg University

Industrial PhD Supervisor: Prof. Søren Bech
Bang & Olufsen and Aalborg University

Co-PhD Supervisor: Dr. Gérard Loquet
University of Melbourne

Co-Industrial PhD Supervisor: Dr. Pablo Martinez Nuevo
Bang & Olufsen

PhD committee: Associate Professor Daniel Overholt (chairman)
Aalborg University, Denmark

Professor Edmund Lalor
University of Rochester, USA

Principal Scientist Hamish Innes-Brown
Eriksholm Research Center, Oticon, Denmark

PhD Series: Technical Faculty of IT and Design, Aalborg University

Department: Department of Electronic Systems

ISSN (online): 2446-1628
ISBN (online): 978-87-7573-682-9

Published by:
Aalborg University Press
Kroghstræde 3
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Stibo Complete, 2023

Curriculum Vitae

Adèle Simon

Adèle Simon obtained a Bachelor's degree in Audiovisual Sciences from the University of Valenciennes in 2013. Continuing her educational journey, she received an M.Sc. degree in Acoustic and Signal Processing applied to Music from IRCAM in 2015. Additionally, she pursued further studies and achieved an M.A. in Music Psychology and Neurosciences from the University of Jyväskylä in 2018. Adèle arrived at Bang & Olufsen in 2018 as a research engineer and has continued working there as an industrial PhD student since February 2020. Her research interests include cognitive neuroscience, audio perception, and audio signal processing.

Curriculum Vitae

Abstract

Music listening, especially on headphones, is often accompanied by intrusive sounds from the environment, speech from people, or background noise. While it generally impacts the quality of the listening experience by distracting from the headphone media, the listener might also want to listen to the external sound. Therefore, it would be beneficial to determine which sound the listener wants to pay attention to, as it becomes possible to use signal processing to enhance the attended sound and reduce the interfering sounds.

Based on linear modelling, cortical auditory attention decoding (AAD) algorithms are a promising path to determine the sound of interest in complex sound scenes. This approach is based on the principle that sounds temporal variations (e.g., amplitude envelope) are traceable from an electroencephalogram (EEG). In addition, the tracking of attended sounds is increased compared to sounds outside of the focus of attention. AAD have been developed and tested mostly on classical cocktail party scenarios with two (or more) streams of speech. Listening to other types of sound, such as music, is also quite common for humans. However, only a few studies tested AAD techniques on music listening in the presence of other sounds. In the present thesis, linear models based on the reconstruction of the audio envelope are used, and the feasibility, performance, and specificity of AAD in situations where the listener is focused on music listening are compared to conditions where they listen to speech signals.

The current project revolved around two experiments. In a first study, the performances of AAD are compared between situations where the listener focused either on speech or music, to test if speech-AAD algorithms can be applied to music. Based on the results of this first study, a second study has been conducted to understand better the differences observed between speech and music AAD. In parallel, the data from the first study have been further explored to investigate the potential differences in underlying mechanisms during speech and music listening. Overall, the work reported in this thesis contributes to 1) assess and understand how the AAD can be used for music listening and how and why performance differs from speech-listening situations, and 2) explore the cortical mechanisms involved in attentive speech and

music listening through linear modelling approaches. The thesis comprises two parts. In the first part, the extended summary, the motivation and background for the work are presented. The second part contains the research conducted during the project in the form of four research papers.

Resumé

Når man lytter til musik, især over hovedtelefoner, forstyrres man ofte af udefrakommende lyde, tale fra personer i nærheden eller generel baggrundsstøj. Generelt set vil disse lyde fra omgivelserne distrahere lytteren fra hovedtelefonlyden, hvilket reducerer kvaliteten af lytteoplevelsen. Der findes dog også situationer hvor lytteren ønsker at høre den omgivende lyd. Det er derfor interessant at finde ud af hvilke lyde lytteren ønsker at være opmærksom på, hvorefter det med signal processering er muligt at fremhæve den ønskede lyd frem for den forstyrrende lyd.

Algoritmer til kortikal auditorisk-opmærksomhed-dekodning (AAD), der er baserede på lineære modeller, er en lovende metode til at bestemme, hvilke lyder der er interessante i en kompleks lydscene. Denne fremgangsmåde er baseret på princippet om at temporale variationer (fx amplitudens indhyllingskurve) er identificerbare ud fra et elektroencefalogram (EEG), og at lyde lytteren er opmærksom på, vil være mere tydelige i forhold til lyde lytteren ikke er opmærksom på. AAD er blevet udviklet og primært afprøvet i det klassiske cocktail party scenarie hvor to (eller flere) personer taler samtidig. At lytte til andet end tale, fx musik, er meget almindeligt. På trods af det eksisterer der kun enkelte studier hvor AAD-metoden er anvendt i forbindelse med lytning til musik mens lytteren er omgivet af andre lyde. I denne afhandling anvendes lineære modeller baseret på rekonstruktion af audio indhyllingskurven. Anvendeligheden, ydelsen og karakteristika for AAD er undersøgt ved at sammenligne situationer hvor en lytter fokuserer på afspillet musik med situationer hvor lytteren fokuserer på tale.

Dette projekt omhandler to eksperimenter. I det første eksperiment blev det undersøgt hvorvidt AAD kan anvendes i forbindelse med lytning til musik. Dette blev gjort ved at sammenligne ydelsen af AAD i situationer hvor lytteren skulle fokusere på enten tale eller musik. Baseret på resultaterne af det første studie blev et yderligere eksperiment udført for at opnå bedre indsigt i forskellene mellem AAD anvendt ved lytning til henholdsvis tale og musik. Sideløbende blev den indsamlede data, fra det første eksperiment, nærstuderet for at kortlægge de potentielle forskelle i de underliggende mekanismer der træder i kraft når der lyttes til enten tale eller musik. Arbejdet i denne

afhandling bidrager til 1) at vurdere og forstå hvordan AAD kan anvendes når der lyttes til musik, samt hvordan og hvorfor ydelsen er forskellig fra lytning til tale; og 2) at undersøge, ved hjælp af lineær modellering, de kortikale mekanismer der er involverede når der aktivt lyttes til tale og musik. Afhandlingen består af to dele. I den første del, det udvidede resume, præsenteres motivationen og baggrunden for det udførte arbejde. Den anden del beskriver den udførte forskning i form af fire videnskabelige artikler.

Contents

List of Publications

This thesis is based on the following publications:

[A] **Simon, A.**, Loquet, G. Østergaard, & J., Bech, S., "Cortical Auditory Attention Decoding During Music And Speech Listening", Submitted to *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, In review

[B] **Simon, A.**, Østergaard, J., Bech, S., & Loquet, G., "Optimal time lags for linear cortical auditory attention detection: differences between speech and music listening," *19th International Symposium on Hearing (ISH2022)*, Lyon, France, 2022.

[C] **Simon, A.**, Bech, S., Loquet, G., & Østergaard, J., " Electrodes selection for cortical auditory attention decoding with EEG during speech and music listening," *25th International Conference on Information Fusion (FUSION)*, Linköping, Sweden, 2022

[D] **Simon, A.**, Bech, S., Loquet, G., & Østergaard, J., "Cortical linear encoding and decoding of sounds: Similarities and differences between naturalistic speech and music listening," Submitted to *European Journal of Neuroscience*, In review

Abbreviations

AAD: Auditory Attention Decoder

AC: Auditory Cortex

ANC: Active Noise Cancelling

ASA: Auditory Scene Analysis

BCI: Brain-Computer Interface

EEG: Electroencephalogram

ERP: Event-Related Potential

GFP: Global Field Power

LOO: Leave-One-Out

LUFS: Loudness Unit Full Scale

PAD: Personal Audio Device

RQ: Research Question

TRF: Temporal Response Function

Preface

This thesis was submitted to the Doctoral School of IT and Design at Aalborg University in partial fulfilment of the requirements for the degree of Doctor of Philosophy (PhD). The work was carried out over the period of February 1st 2020, to June 23rd 2023. The work was carried out at Aalborg University and Bang & Olufsen a/s, who sponsored the work. The research leading to these results has received funding from the Innovation Fund Denmark Program under grant agreement No.9065-00270B.

I'd like to express my profound gratitude to the people who have played a significant role in the completion of this work. First and foremost, I would like to extend my most profound appreciation to my supervisors, Søren Bech, Jan Østergaard and Gérard Loquet, for their support, encouragement, constructive feedback, thoughtful discussions and patience throughout this project. And a special thanks also goes to Søren Bech for trusting and supporting me even before this PhD started and for allowing me to explore my ideas through this project.

I am thankful to the colleagues I had over these three years. For making this journey an enriching and rewarding experience- I am glad and grateful to had the chance to be surrounded by such brilliant and kind people.

Thanks to my colleagues at B&O, and specifically the research team. Thanks to Martin, Sarvesh, Sven, Pablo, Niels, Pia, Jon, Line, Kate and Klaus and the intern I've met through the years for your kindness and the stimulating intellectual atmosphere of the research team. I also thank all my AAU colleagues for the supportive and collaborative environment fostered by the group. It was a pleasure to work there.

Thanks to the people of the InHear project, both supervision teams and my fellow PhD students, for the great collaboration.

I am grateful to my study participants, who trusted me with their brains. Without their contributions, this research would not have been possible.

I sincerely thank my family and friend for their support through the years.

Finally, an exceptional thanks to Gaël for his unwavering love and understanding throughout this thesis journey (and before). And thank you also for your contributions to this thesis, the design of the figure and the (numerous and tedious) proofreading. I'm lucky to have you by my side.

While it is impossible to individually name every person who has contributed to this work, I extend my appreciation to all those who have played a part in making this thesis a reality, no matter how big or small.

Adèle Simon
Aalborg University and Bang & Olufsen, June 23, 2023

Part I

Extended Summary

Auditory Attention Decoding for Everyday Music Listening

1 Introduction

With the development of Portable Audio Devices (PAD), such as headphones or earbuds, one can now listen to their own music or podcast whenever and wherever they want: when walking outside, on public transportation, at work, or at home. However, in those environments, many sound sources coexist and may disturb the listening experience of the PAD user. Some of the sounds are not desirable for the listener (e.g. noise pollution [104]), but others may sometime be useful (e.g. alarm, social interaction). The user can use some strategies to mitigate the distraction induced by external sounds, such as raising the volume or turning active noise cancellation (ANC). However, if the masking of the external sound is too important, a lack of environmental awareness can also be a problem. That is where a technology that detects auditory attention can be of interest.

1.1 Motivation of the project

Let us consider the case of Lisa, as pictured in Figure 1. Lisa is waiting for her train in a crowded station. She puts her favourite playlist through her headphones to pass the time. It also serves to isolate her from the background noise of the environment. But Lisa does not want to be completely isolated from the sound around her, as it might feel unnatural and reduce environmental awareness [108]. She needs to hear the voice announcement from the train station in case there is information about her train or if another person is trying to interact with her. Then, to be sure that she will not miss any essential information, she has to stop her music or remove her headphones whenever she hears some voice that could be relevant to her. As the station is busy, she has to interrupt her listening often and frequently, which badly impacts the music listening experience and can lead to fatigue because of the environmental

noise.

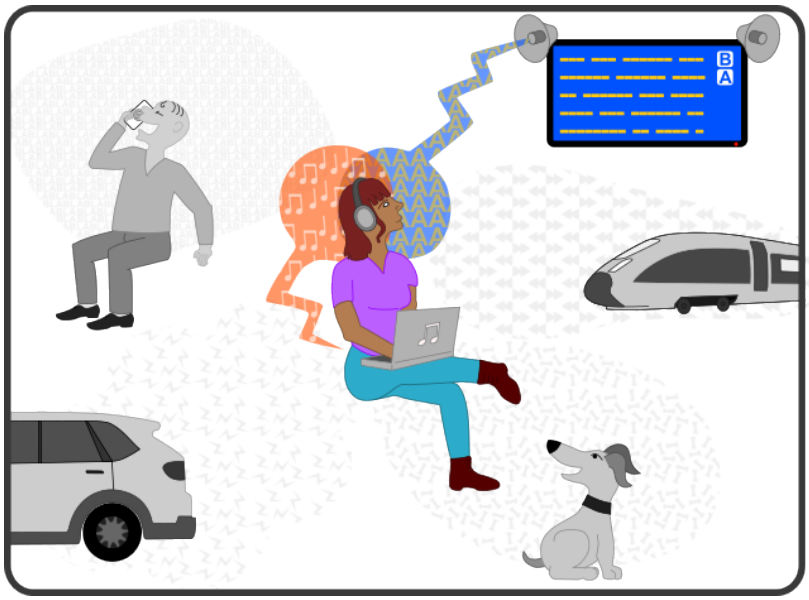

Fig. 1: Illustration of a typical situation where the attention of the listener can be directed either at on the headphone media (music) or at the external world (voice announcement). In such a situation, AAD could help to know what she wants to attend and adapt the sound accordingly.

Similar situations happen in many other settings: when walking in the street, in domestic situations, in an open-plan office, or in cafés. PAD wearers have to compromise between the amount of external noise they want (for convenience, safety, or social reason) and the acceptable level of distraction or annoyance during the listening. One solution could be to implement the PAD with some automatic voice detector that would adapt the content of the headphones if a speech source is detected in the vicinity of the user [20, 107, 214]. It can increase environmental awareness while suppressing some undesirable background noise (e.g., engine noise in a train or traffic in the street). However, as seen in the aforementioned example or as often mentioned by PAD users [108], not all speech signals are desirable for the listener. Another solution that has been proposed is to equip PAD with sensors that would make them sensitive to non-verbal conversational cues from people around the user: when someone approaches the user [167] or gazes toward the user [152], transparency mode (i.e. using the PAD internal microphone to let the external sound pass through the PAD as if the user was not wearing it) would be automatically activated to ease verbal interactions. Nevertheless, the user might not want to be interrupted by this person that wishes to interact with

them (any woman who experienced catcalling in public would easily see the limitation of such a system [89]). In addition, the sound that a listener wants to attend to is not always speech: listeners may want to direct their attention to music or other environmental sounds. Therefore, determining which sounds are of interest to the listener and which sound is a distracter, solely from the audio signal, is arduous. The annoyance or interest in one sound will also depend on the situation or the listener: for instance, some might consider the bird singing through their windows early in the morning annoying, as it would wake them up, while others would enjoy being able to listen to it (see [231, Chapter 2] for more details on the factor influencing the perception of different sounds as wanted or pleasant).

To determine the desired sound a listener wishes to focus on, one approach is to examine the neural activity occurring in their brain while they are engaged in listening. Attention has been shown to modulate the cortical signal related to auditory processes [2, 187] (see section 2.3). In the last decade, the question of auditory attention decoding (AAD) based on Electroencephalography (EEG) has gained interest, especially in the field of hearing aid devices, where knowing what a listener is trying to attend to can help amplify relevant sound and reduce interfering noises.

Based on linear models, one approach has been developed to detect to which source a listener is paying attention in a cocktail-party scenario. This technique relies on the fact that the brain tracks the temporal variations of the audio signal and that the tracking is enhanced for attended sounds compared to ignored sounds [80, 159, 176]. Contrary to more traditional neuroscience processing techniques, this technique allows decoding attention during continuous listening without needing repetitions or averaging of the cortical recording. It makes this approach a suitable candidate for AAD to implement into PAD. This idea is not new and has been suggested and explored by several research teams [23, 101]. However, it is mostly considered for hearing-aids scenarios and thus focused on decoding speech signals [23, 101]. A speech AAD implemented in hearing aids would help in many difficult situations for the hearing-impaired patient [99, Section 1.1].

Nevertheless, it does not consider other types of listening, such as music listening. With an average of 20,1 hours of music listening per week [126], music listening is an everyday activity for humans that could also benefit from sound enhancement driven by AAD. However, AAD for music listening has been only sparsely investigated. The motivation of the present thesis is to close this gap by evaluating an AAD approach that has been used for speech-AAD on music listening and to investigate the differences between speech and music listening in the presence of distracting sounds.

1.2 Context of the work: InHear Project

The present PhD project took place in a larger project, the InHear project[1] (Intelligent Hearables with Environment-Aware Rendering) This project was a collaboration between Bang & Olufsen, Aalborg University, and The University of Surrey (UK). The basis of the project relies on the idea that, with modern PADs, we could use information from various sensors to adapt and enhance the user's listening experience depending on the listening context (context of listening, environment and noise around the users, or attentional state).

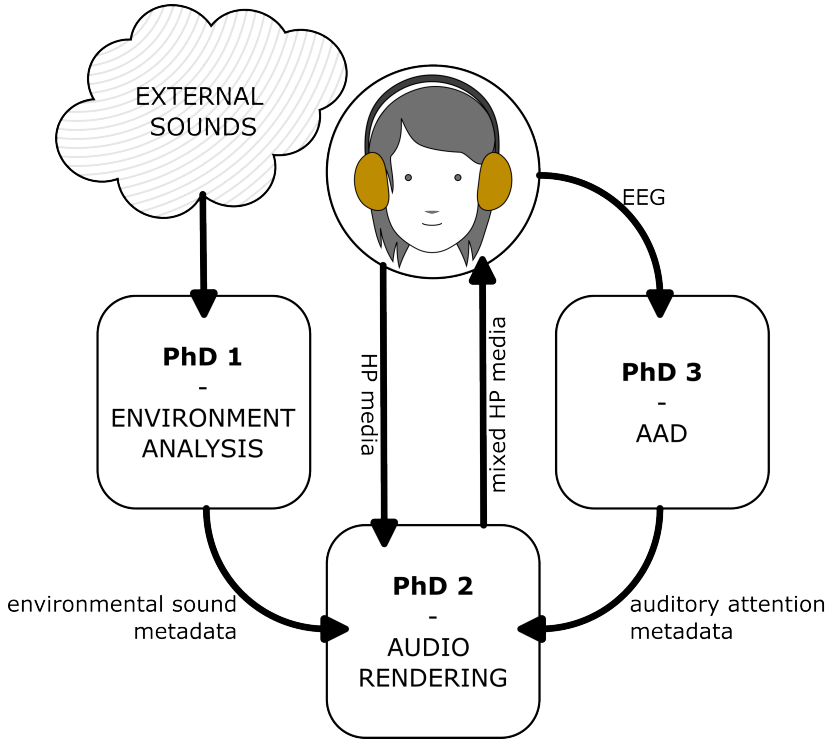

Fig. 2: Concept of the "Intelligent Hearables" and structure of the InHear project

The project consists of three PhD projects, as shown in Figure 2 focused on:

- Environment capture and analysis: detection of auditory events around the listener (e.g. type of event, direction of arrival). It would be used to adapt the rendering to the environment.
- Audio rendering: modify rendering methods (equalising, spatialisation

[1] https://iosr.uk/projects/InHEAR/

or something else of the PAD media to facilitate the attentional need of the user and enhance their listening experience)

- Cortical auditory attention decoding: Project presented in this thesis.

Taken together, the three projects aim to develop a system that can seamlessly adapt to the context of listening and to the attentional needs of the listener. In the present work, we assume that thanks to this collaboration, we could have access to information about the sound scene, both external sound and media played by a PAD (classification of the sound present, onset and offset of each auditory stream, their direction of arrival). It is also assumed that strategies can be implemented (e.g., setting of the PAD, equalising of the sound, change in the spatial placement of the audio object, change in the target-to-interference ratio) [193, 195] that could nudge the sound scene to adapt to the attentional needs of the listener.

1.3 Research Questions

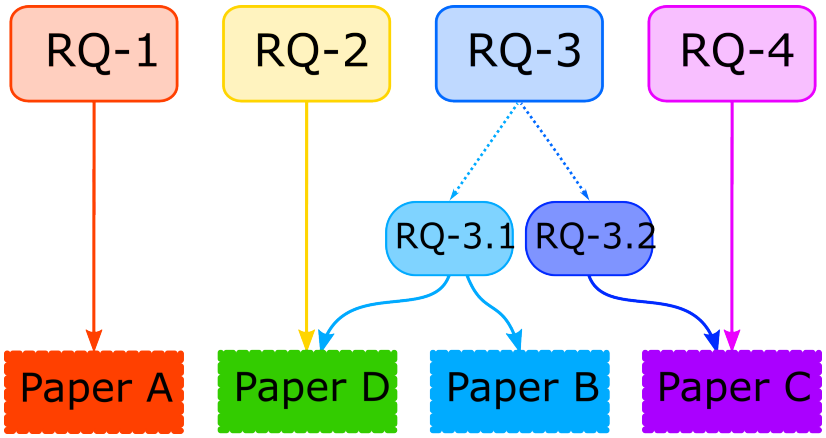

Fig. 3: Relationship between the research questions and the research papers presented in the thesis.

The primary goal of this work is to explore the feasibility, performance, and limits of cortical auditory attention for listening to music in the presence of other sounds. Early in the project, it was decided to use linear modeling, especially backward models [2] for AAD. This technique was proven successful

[2]Backward model, decoding model, or stimulus reconstruction approach are equivalent and used interchangeably throughout this thesis. Similarly, forward model, encoding models or EEG prediction approach are used interchangeably.

on speech signals but rarely investigated for music signals. The first main question has then been formulated as follows:

- RQ1: Can linear backward modeling be used to decode auditory attention during music listening in the presence of other sounds?

The RQ1 has been addressed in Paper A. The results of Paper A, which showed that the reconstruction accuracy obtained for music is lower than for speech, lead to another research question.

- RQ2: What causes the difference in reconstruction accuracies between music and speech listening obtained with linear models?

This question is the focus of Paper D. Also following the results of Paper A, a third research question was explored. It is more open and explored via different standing points in several papers.

- RQ3: How does cortical tracking differ between speech and music listening?
  - RQ3-1: How does the temporal aspect of cortical tracking differ between speech and music? (Paper B & D)
  - RQ3-2: How does the spatial aspect of cortical tracking differ between speech and music? (Paper C)

In addition to addressing RQ3, Paper C also explores another research question relevant to the future development of AAD.

- RQ4: How a reduction of electrodes impacts the performance of AAD?

1.4 Scope, limitations and assumptions

AAD based on linear modeling is promising for implementing Brain Computer Interface (BCI) in audio products [23, 101]. However, little is known about their performance during music listening in the presence of other sounds. This project aimed to contribute to filling this gap in research. At the start of the project, the theoretical bases of cortical AAD through linear modeling were laid, and several studies demonstrated good performance for auditory attention decoding for speech listening. However, only a few papers investigated this approach for music listening and only did so with the attention directed toward a specific musical instrument within polyphonic music. Based on listening habits in daily life, which often include music, the first question was to see how the AAD that is usually used for speech performs on music. To do so, and due to the industrial collaboration of the project, we employ a top-down approach. Naturalistic stimuli were used to test if significant differences arose from realistic situations to investigate the putative factors that could explain

the differences. The goal was to get as close as possible to realistic listening conditions and incorporate listening behaviours as in real life. The first experiment, presented in Paper A, was based on this approach and compared the speech and music signals usually listened to through PAD (audiobooks and existing music).

Based on the first study's results, a second experiment was designed with the same philosophy in mind: using naturalistic stimuli, we aimed to untangle the factors that could explain the differences observed between speech and music listening. Throughout this thesis, in addition to AAD performance, we also aimed to exploit the output of the linear models, both backward and forward, to explore differences between music and speech listening to gain information on the underlying cortical mechanisms that could be used to inspire future study.

Although the project started with the idea of a versatile AAD that can be implemented in PAD and used in real-world settings, the focus of this dissertation is to address essential initial questions related to the development of this technology, rather than comprehensively covering all aspects of its development. The goal of the present work is not to develop an AAD but to investigate the behaviours and limitations of linear AAD during music listening. Therefore, questions related to hardware and software development of a PAD-embedded AAD are outside the scope of this thesis, and it is assumed that such implementation would be feasible (see section 2.6). In addition, some experimental design choices have been made to isolate the factors under investigation (i.e. type of sound, either speech or music) and limit interaction that could have limited the interpretation of the results.

Soundscapes

Simple sound environments were used for the two experiments conducted during this project: two concurrent sound streams for the first experiment and only one sound stream for the second. This choice was made to simplify the audio sound scene presented to the listeners and avoid as many interferences as possible that would increase the number of factors influencing the results. Such simplified sound is not representative of natural soundscapes, and future work should aim to expand the current results to more complex sound scenes. In the context of speech listening, AAD has been shown to be robust in the presence of more sound stream [98, 199] or the presence of noise [8, 67, 98]. While further research would be needed to confirm that this robustness to noise or more complex sound scenes also occurs for music listening, it is assumed that the present findings would also apply to more complex sound scenes.

Access to sound sources

In the presence of noise, access to clean audio sources may become a challenge. In the scenario of the PAD-embedded AAD, access to clean PAD media is not a problem. However, access to a clean signal of an external distracter is technically challenging. PAD's microphones would capture a mixture of external sound, with both distracter streams and irrelevant noise. While it would most likely decrease the attention detection performance [12], some previous work suggest that AAD could still be useable without access to clean sources [13, 14]. In the present work, it was assumed that clean signals for all sounds would be available.

Multimodal interaction

In the present work, only the auditory modality was under investigation to reduce interfering factors that may have hindered questions under investigation. However, humans have other senses besides hearing, and cross/multimodal interactions are expected to impact the neural tracking of auditory input. Previous studies have shown that visual stimuli can affect the neural tracking of auditory input [59, 174] and that incorporating visual information into the AAD model may improve its performance [59]. Nonetheless, in the initial PAD use case, the sound of interest (the PAD media) is primarily acousmatic, thus not related to any visual information. In that case, it is also relevant to investigate the auditory modality in isolation.

1.5 Structure of the Thesis

The present dissertation is structured as a collection of papers written during the PhD project. It is divided into two parts. The first part consists of an introduction to the topic of the PhD work, background information and an overview of the contributions made during this project. It is followed, in the second part, by the four papers addressing the research questions defined in section 1.3. The first part comprises general background (Section 2) on habits of listening with PADs; auditory scene analysis and auditory attention; linear model for exploration of neural data, and more specifically for AAD; differences between speech and music in the human brain; and finally, on existing technology that could allow for a mobile AAD. It is followed by a summary of the research findings and the contributions (Section 3). Section 4 presents the conclusion and consideration for future work.

2 Background

2.1 Everyday music listening in the presence of noise

The present work aims to explore the feasibility of a versatile Auditory Attention Decoder embedded in a PAD, such as headphones or earbuds, used to listen to music, podcast or audiobooks. The following section will focus on music listening, but it can also be applied to podcasts, which are widely used on PADs [86, 146], or any other auditory content that is played on PADs (radio or audiobook). To contextualise the use and the benefit of such AAD technology for PADs, it is necessary to better understand how headphones are used and how the attentional state and needs of the listener may vary in different situations.

Over the past century, technological progress led to fast development and democratisation of personal and mobile listening devices [10, 88, 236]. With the invention of the Walkman in the late 1970s, followed by the Discman, listeners can enjoy their own music everywhere, even on the move [43, 237]. The development of MP3 players, followed by streaming services, allowed listeners to enjoy a tremendous variety of music anytime, anywhere. These technological evolutions profoundly change how people listen to music [127, 145, 161, 203]. Nowadays, personal music listening is ubiquitous [108, 138]. PADs are used in various situations: at home [108, 173, 196]; outside, in urban environments [108, 116, 173, 196]; in public places (café, shops) [208, 209]; in public transport [108, 130, 196]; at the workplace [106, 108, 130]; or while driving [173]. In all these situations, music listening is often accompanied by some external noise that can influence the listening experience.

Furthermore, personal music listening can serve different purposes: entertainment, emotion regulation, motivation, to help to pass the time, creating their own "auditory bubble" [43, Page 344], to create a suitable atmosphere or as a soundscape for their life [97, 173] (For more details, see [43, 106, 116, 194]. Each purpose or situation influences the choice of music and the way to listen to it [105]. Each listening episode requires different levels of attention on the PADs media and can lead to various needs and strategies to mitigate the effect of the external sounds [95, Page 16]. Background listening, used to mitigate boredom during housework, may not require the listener's full attention. On the contrary, during active listening, accompanied by singing along in their head, the user's attention is focused on the music stream.

However, neither the context of listening nor the purpose of listening can alone predict the attentional needs of the user. Some users might enjoy the background music played in a café while others may find it highly annoying and would prefer to block it to enjoy their own music [208]. When commuting or walking in public spaces, some users welcome potential distractions [116], while others use PADs to create their own "auditory bubble" [116]. In the

latter case, distractions from external noise are considered an annoyance, and interruptions from the listening would be seen as unpleasant [42]. The PADs media or technical options such as ANC are used as an active strategy to reduce distractions from the external auditory world and enhance the listening experience [133, 195, 196].

While in some situations, the auditory bubble strategy is purposefully used to block all distractions (trying to relax on a busy train or doing focused work in an open-plan office). In other cases, the bubble needs to allow some permeability. It could be for safety reasons, as a distraction from listening to PADs can be a safety hazard [116] if some sounds remain unnoticed, for example, while driving [177], riding a bike [108] or walking in busy streets [147, 238]. In other situations, the user may want to be reached by some information (e.g., voice announcements on trains or airports), but all the information might not be relevant. It leads to a trade-off between the need to access the external world and the flow of listening, as reported by one of the participants from [108, Page 63]:

"If you are sitting at the airport, listening to content, then every time you hear the 'DingDong', in theory, I still have one hour until my flight, but sometimes they'll relocate the gate, maybe they do this and that, and then you constantly take off the headphones."

The user in this specific example could benefit from an AAD system that could adapt the sound of their headphones based on their auditory attention. First, the 'DingDong' captures their attention and can lead to a reduction of the level of the headphone media to allow them to hear the announcement. If the announcement is of interest, and the user's attention remains on the external voice, the headphone level would stay low. Otherwise, if they switch their attention back to the music, the headphone will automatically increase the volume to the original level, thus removing the need for the user to take their headphones on and off.

In order to modify the sound scene to increase the quality of the listening experience and reduce distraction from external sound, it is crucial to understand how the human hearing system and brain process sound. Research on auditory scene analysis and auditory attention can give insight into what is happening when one listens to a sound in a noisy soundscape and how to take those mechanisms into account in the design of an AAD.

2.2 Auditory Scene Analysis

One sound is rarely alone in a sound scene. In most situations, a source of sound is mixed with several other sounds: environmental sounds, human-made noise, speeches, and music. When walking outside, for instance, a person may hear engine noise from passing vehicles, voices from a passerby, chirps from birds, and a plethora of different sounds. In such situations,

humans (as well as other animals [91]) can segregate all the auditory input into different perceptual units, allowing them to attend to one sound while ignoring the others. This ability is often called "The cocktail party effect". This term, coined by Colin Cherry in 1953 [50], is based on the human capacity to listen to one specific person in a crowded and noisy room while ignoring other people talking [11]. While the task is intuitive and easy for a human listener, it is complex from an acoustic point of view to separate one stream of speech from a mixture containing several sound sources [11, 35, 87].

In 1990, Bregman summarised work on perceptual sound segregation and perceptual grouping of sound in his book on Auditory Scene Analysis (ASA) [35]. Auditory scene analysis encompasses perceiving individual sound streams in an acoustic mixture that reaches the ears and how to segregate and process them.

Fig. 4: When a sound scene is composed of several sources (here, a dog and a cat), what arrives at the ear is a mixture of the sound signals. Through the process of auditory scene analysis, each sound is perceived as an individual stream

The basis of auditory scene analysis relies on the idea that the auditory system separates the incoming mixture of sound into auditory streams. Sometimes also called auditory objects, an auditory stream is defined as the perceptual construct of the sound [31, Pages 2-3], *"...part of the spectrogram that goes together "* [35, Page 9] or *"perceived as a whole coherent entity"* ([198], as cited by [52]). The formation of auditory streams, also called auditory streaming, is *"the human ability to analyse an auditory scene i.e. to attribute portions of the incoming sound sequences to distinct sound-generating entities"* [18].

Several factors play a role in the segregation of the sound scene and the grouping of the auditory streams, and have considerably been studied through listening, behavioural and brain-imaging experiments (See for reviews on the

neuronal mechanisms of ASA [3, 74, 201, 248]). Some of the grouping mechanisms are primitive and depend mainly on the acoustical feature of the audio input. Such primitive grouping is believed to be automatic and has been observed in infants [154, 241] or animals [22, 90]. Segregation and grouping of auditory streams can be influenced by, for instance: simultaneous onsets or offsets [66, 205]; continuity of the signal [27, 30]; common amplitude modulation [65, 205]; spatial localisation of the sound [27, 48, 158, 165, 212], through binaural cues [26, 49, 50]; timbre [128, 207, 239]; fundamental frequency [165].

From the above list of factors, we can recognise several Gestalt principles [135] that have been extensively studied for grouping visual objects and apply to auditory streaming [35, 128, 136]. Temporal variations of those features over time also influence auditory scene analysis, and the coherence of the different features guides segregation [200]. For more extensive reviews of the Gestalt principles involved in auditory streaming, see [72, 74] or [35, Chapters 2-3].

The segregation and integration of auditory streams are not solely steered by the acoustical properties of the sound scene. Context influences the integration of auditory streams [219, 240]. Prior experience with some sound features can facilitate the integration of one part of the spectrogram into one auditory stream [35, Chapter 4]. Such processes are not innate as primitive segregation and take advantage of schema-based learning [244]. Knowledge about the content, some features, or parts of the auditory scene can influence how it is segregated [28, 157]. For speech, familiarity with a voice [129] or word understanding can ease the segregation and integration into separate streams [156].

In most cases, an auditory stream corresponds to a physical sound source (e.g. one auditory speech stream usually originates from one physical person talking) [48]. However, it is not always so straightforward, especially when listening to music. Music is created by humans, musicians or composers that shape sounds to influence their perception [124]. The aforementioned auditory streaming principles still apply [74, 155]) but can also be used to create musical streams that do not necessarily correspond to every single instrument: harmonic relations can be used to link instruments into a single stream, or a single instrument can be used differently over time to create distinct streams [35, Chapter 5], [74, 156, 190] (See Figure 5). It has been shown that, when listening to polyphonic music organised through melodic or harmonic rules or western music composing, listeners do not attend to one melodic line only but instead integrate the whole polyphonic musical object [19]. As for any other audio input, familiarity [28, 83] and prior knowledge of music also influence auditory scene analysis. However, as seen in section 2.1, music listening is not always the primary task of the listener. Thus, the listener's attention might not be focused on the music. In this case, a whole musical piece may be perceived as one common auditory stream by integrating together its distinct sub-stream [29], [35, Pages 460-461].

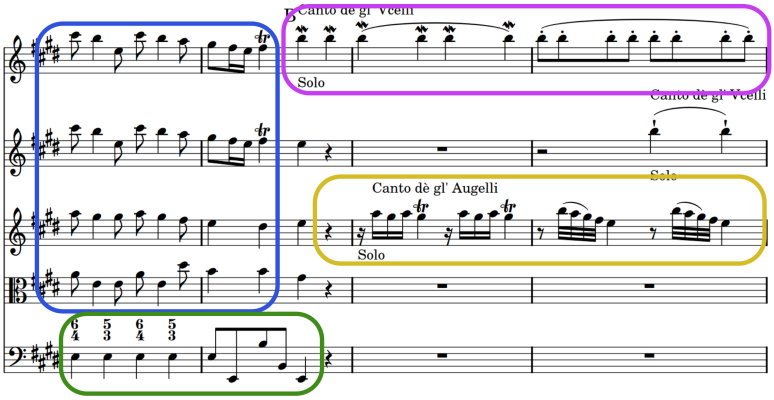

Fig. 5: Considering the first movement of the Spring in Vivaldi's Four Seasons, all instruments are first likely to be split into two auditory streams, both contending several instruments (Main melody and accompaniment, blue and green circles). From the third bar, due to the change of melody and use of vibrato, the first and third violin may later be perceived as individual streams (circles pink and yellow).

2.3 Auditory Attention

Another cognitive process that highly influences auditory scene analysis is attention. It is commonly accepted that attention plays a vital role in our ability to perceive the world around us. However, due to its multifaceted aspects, the term attention is challenging to summarise with a concise definition [55]. This section will focus specifically on selective auditory attention, which can be described as "the mental ability to resist distracter stimuli and select relevant information from the surrounding acoustic events, as illustrated in the "cocktail party effect" [102]. When attending to one sound and ignoring the other, the sound of interest will be further processed by the brain compared to the other [156]. In a cocktail party scenario, one will understand and remember the topic of the conversation they are attending. At the same time, they will not be able to recall the topic of the discussion that was not attended, even though their ear heard it. The attention can be driven either by exogenous factors, such as salient sounds that would capture a listener's attention or from endogenous factors, such as top-down attention [41, 64, 132].

When the attention is directed to one stream of sound, the brain further processes this steam: the content is processed and can lead to cognitive or behavioural responses. This processing is limited for the sound outside of the focus of attention. To what extent the brain processes the ignored streams is not yet completely understood. Some studies showed that the content of an unattended speech is often unnoticed [51, 166], and that listeners fails to

notice changes in a speech sound when the attention is directed to another sound (if the gender of the speaker is changed [51], or replaced by backward speech [211] or unexpected speech message [63]). This effect has also been observed during music listening [137]. Based on this idea that unattended sound is not processed, an early model of attention was developed, suggesting that attention filters out irrelevant sounds [84].

However, this effect has been challenged by studies that showed that unattended sounds are not completely unprocessed [58, 197, 229]. It has been suggested that attention is not acting as a blocking mechanism that only lets the attended signal through for further processing, but more as an attenuation filter [229, 230]. One famous example of this processing of the unattended stream is that people tend to notice if they hear their own name in an unattended auditory stream [166, 243]. Several factors may modulate to what extent ignored streams are processed, either due to cognitive factors [57], acoustic factors [16], or factors related to perceptual load (see [168] for review).

In addition, it poses the question of the relationship between auditory attention and auditory stream segregation. The early model [37] relied on the idea that attention acts as a filter separating attended and unattended sounds and that this attentional filter happens before auditory streaming. Based on this model, only the attended audio would then be segregated and grouped into streams. However, this model has been challenged and evolved [84], and nowadays, it is admitted that auditory streaming happens either before or outside of attention [3, 240]. Some auditory scene analysis is performed before the attentional filter, and the attention occurs on auditory streams or objects [153]. (See Figure 6 - a))

Nevertheless, the two processes are not entirely independent [216]. Attention is sometimes necessary to refine stream segregation and integration for ambiguous stimuli [3, 204]. Changes in the focus of attention may also "reset" the auditory scene analysis process to provide different sets of streams that are more relevant to the listening goal of the listener [62, 210, 218, 226]. Such effect can happen, for example, for music listening: when polyphonic music is not the focus of attention, all instruments tend to be grouped into one single auditory stream, but if the listener starts paying attention to one specific melody line, the relevant instrument would be treated as a single stream. (See Figure 6 - b))

The attention modulates the brain signal, with stronger responses related to attended auditory streams than non-attended sounds [30, 132, 216]. This property of the auditory system can be used to decode auditory attention from neuroimaging techniques.

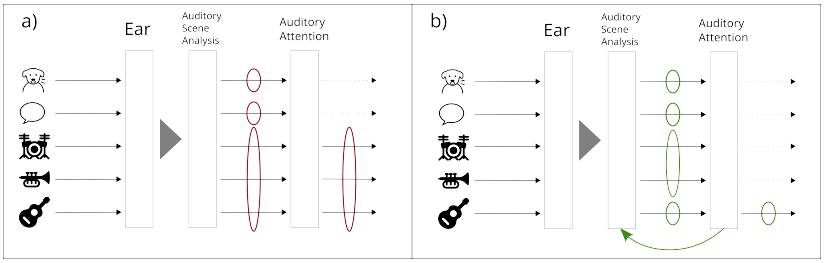

Fig. 6: Auditory scene analysis happen before attentional processes: the dog and the speech are each segregated into individual streams, while all instruments are grouped into one stream (A). Focusing the attention on finer aspects of the sound (here, one of the instruments) reset the ASA process, and induce finer stream segregation, turning the attended instrument into an individual stream (B).

2.4 EEG & Cortical AAD

Electroencephalography, or EEG, is a non-invasive neuroimaging technique that uses electrodes placed on the scalp to measure the brain's electrical activity. The electrodes capture the electric field resulting from multiple post-synaptic potentials [160]. The high temporal resolution offered by EEG is useful for analysing the temporal dynamics of brain processing during auditory tasks.

Electroencephalography has been used in auditory science for a long time. Traditional methods, such as Event-Related Analysis (ERP) [36, 149] help to understand the mechanism of hearing and perception of sound [4, 118, 169] and have been extensively used to study ASA [3] or auditory attention mechanisms [118, 210, 217]. Auditory attention has been shown to modulate the amplitude of auditory ERPs [53]. However, such methods require some stimulus and task design that make the use of ERPs irrelevant for an online AAD that would be used for natural sound sources such as continuous music or speech. First, auditory ERPs are helpful in exploring the cortical response to discrete sound units: by definition, they are time-response elicited by an event, such as the onset of a sound or an oddball stimulus [149]. ERP can be extracted from continuous listening, following some changes in the sound. For naturalistic music listening, large variations in spectral flux or brightness have been shown to elicit an ERP response [188]. However, in realistic listening situations, even if such auditory events are present, it is challenging to precisely identify the event likely to elicit an ERP that could be used for attention detection. Secondly, the amplitude of an ERP is low compared to the amplitude of the ongoing EEG, and the latter also contains cortical activity unrelated to the auditory stimuli under test [150, Page 8]. To extract the relevant information, the stimulus that elicits the ERP is repeated over several trials, and the cortical response of each trial is averaged together to obtain a usable averaged ERP [36, 149]. The

need for repetition and averaging is unsuitable for an online cortical AAD that should detect attention based on a single listening of the stimuli.

Some attempts at BCI based on the influence of attention on ERP have been developed [54, 117, 228], but require to use stimuli that are specifically designed to elicit clear ERP. ERP-based BCI can be useful for specific clinical applications, such as using auditory attention to control a speller BCI for communication with locked-in patients [117]. However, as previously mentioned, the limitations inherent to ERP analysis exclude these techniques from further consideration for the present use case.

Cortical modeling

In the last decade, linear models have gained popularity and demonstrated promising findings for decoding auditory attention from continuous cortical decoding. This modeling approach relies on the idea that some information enters our brain (e.g., sound through our ear and hearing system) and that our brain processes this information and can influence behavioural and cognitive functions. It assumed that the incoming information is represented in some neural activity [71]. Linear models are statistical models that assume a linear relationship between the input and neural representation [119]. They can be either: Encoding, or forward modeling, to predict a cortical response from a stimulus) ; Decoding or backward modeling to estimate a stimulus representation based on the cortical response [111, 139]. Both approaches can also be combined to create bidirectional models [68, 69, 82, 101].

Compared to the more traditional ERP, the modeling approach allows the investigation of continuous stimuli that are more similar to ecologically valid situations [109, 119, 142, 225]. The algorithm used to fit the model may vary, as well as the stimulus features being encoded or decoded (see review in [23, 101]). In the auditory domain, models have been trained to either encode or decode an audio envelope [14, 76, 77, 113, 176, 191, 249] or a multivariate feature such as spectrograms [47, 76, 81, 159]. Higher-lever features have also been used to study cortical response to speech signals, such as phonemes, phonetics, and semantic or linguistic features [40, 75, 76, 78, 103, 221].

Auditory attention decoder

In the context of auditory attention, linear models have been used to investigate how attended and ignored speech signals can be encoded or decoded and what is the influence of the attentional factor. When a listener is presented with two concurrent streams of sound, decoding models have shown that the cortical representations of the sound (or some features of the sound signal) are impacted by attentional factors [235]. Features from both attended and non-attended streams can be decoded. However, the stimulus reconstruction

accuracy (i.e., the correlation between the reconstructed and the original signal) or the EEG prediction accuracy (i.e. the correlation between predicted and original EEG) is higher for the attended stream compared to the non-attended one. This effect of attention has been shown with intracranial-EEG [159] or magnetoencephalography [80]. Similar methods were then tested on EEG recording, leading to similar conclusions [60, 114, 176]. Whereas the reconstruction accuracy was predictably reduced compared to intra-cranial EEG, the difference between the reconstruction accuracy obtained for attended and non-attended remains [38].

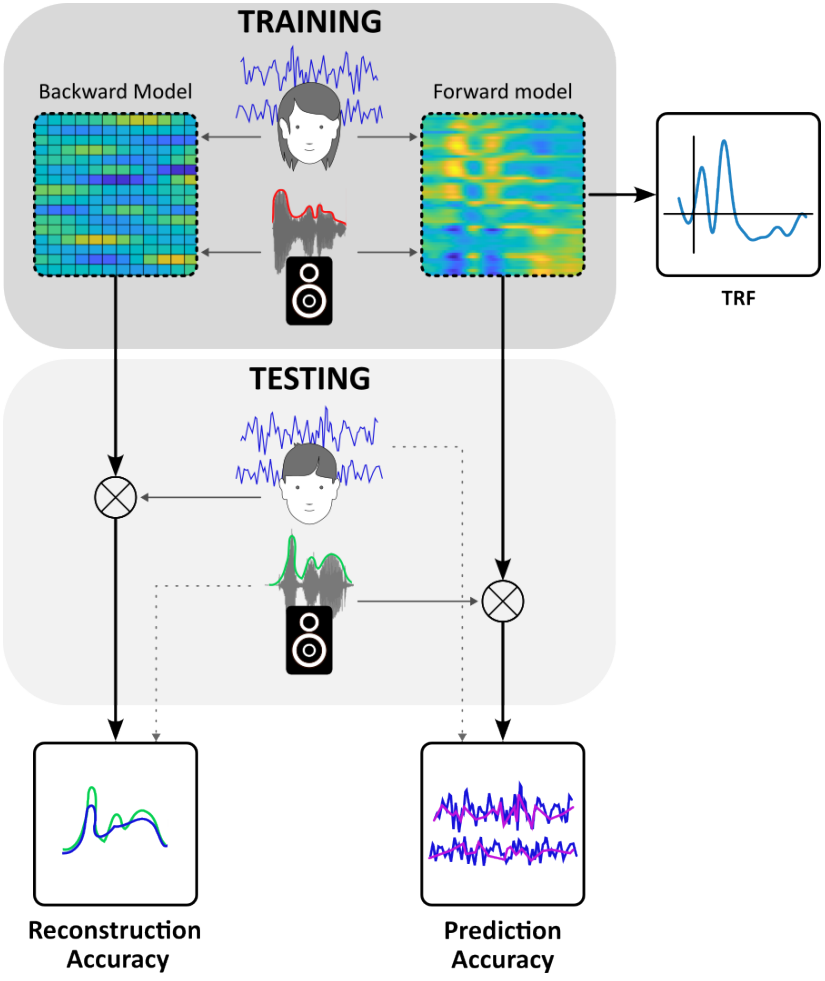

Fig. 7: Schematic of the backward and forward modeling approaches used in the present thesis.

The most common design of cortical AAD is based on a backward linear

modeling approach, that aims to reconstruct features of the audio stimulus from the cortical signal. This approach uses a linear model (W_b), which acts as a multi-channel filter, to create an estimate (S') of an input stimulus (S) from an EEG recording (R) [6, 60, 101, 176].

$$S'(t) = \sum_{n} \sum_{\tau} W_b(\tau, n) R(t + \tau, n) \tag{1}$$

The input stimuli S refer to the sound's envelope played to the participant. $R(t - \tau, n)$ refers to the EEG signal recorded by electrode n at the time $(t - \tau)$ during the stimuli S presentation. The EEG recording is then mapped to the stimulus envelope through the model W_b. Using the time-lagged version of the EEG, via the τ parameter, includes EEG activity relative to auditory or attentional cortical processing at different latencies post-stimulus [61, 176]. To estimate the model W_b, the mean-squared error between the reconstructed stimuli S' and the original stimuli S is minimised:

$$\min \epsilon(t) = \sum_{t} [S(t) - S'(t)]^2 \tag{2}$$

This can be solved by calculating the normalised reverse correlation [213] as follows:

$$W_b = (R^T R)^{-1} R^T S \tag{3}$$

Regularisation can be added to the model to prevent overfitting and increase reconstruction performance (see [101, 119, 242]).

In order to decode auditory attention, the model should be trained on EEG signals and features of attended sounds. The model can then be tested on unseen data of a cortical recording of a listener in a cocktail party scenario. In that case, the reconstructed stimulus can be compared to the sounds that were presented to the listener, and the higher reconstruction accuracy would be sound for the attended sound source [101, 176].

One of the limitations of the backward model is that the models themselves can not provide direct insight: the model weights are not straightforwardly interpretable [111, 139]. The forward encoding approach can also be used to gain supplementary information. The forward modeling follows the same theoretical approach as the backward modeling, except that instead of reconstructing stimuli from the brain signal, it aims to predict the brain signal (R') from the stimulus:

$$R'(t, n) = \sum_{\tau} W_f(\tau, n) S(t - \tau) \tag{4}$$

The model W_f is then estimated by minimising the mean-square error between the original EEG recording R and the reconstructed EEG R' through normalised reverse correlation, comparably to the backward approach:

$$\min \epsilon(t,n) = \sum_{t} [R(t,n) - R'(t,n)]^2 \tag{5}$$

$$W_f = (S^T S)^{-1} S^T R \tag{6}$$

One of the benefits is the interpretability of the model's weights W_f. The model obtained with such an encoding method produces a Temporal Response Function (TRF), which shows similarities to ERP related to the listening task [61, 111, 142, 143].

For linear regression, forward models are outperformed by backward models for stimulus reconstruction or EEG prediction: the correlation between the reconstructed signal and the original one is generally higher when using a backward model compared to the prediction accuracy obtained with a forward [242]. Compared to the forward approach, performance in the AAD task is also better with backward models [6, 242]. The forward and backward approaches can also be combined into a canonical component analysis [69]. This approach is promising for future implementation as it leads to a better success rate in AAD compared to backward or forward models used in isolation [69, 101]. For BCI implementation, linear approaches (both backward and forward and CCA) are still limited by the amount of cortical data required for accurate reconstruction or prediction (20 to 60 seconds), as the decoding performance significantly drops for short segments.

In addition to linear methods, AAD has also been investigated through non-linear methods (see [101, 192] for review). Some of those approaches rely on the stimulus reconstruction method [1, 175, 220, 227], while others perform direct classification of the attended sound [56, 234]. They generally outperform the linear approach [56, 227, 234] or can reduce the length of the EEG signal required for attention decoding [101, 234]. However, there are limitations with non-linear methods: they are more prone to overfitting, especially with a small dataset, which is typical for cortical recordings [101, 234], and they often fail to generalise to new datasets [101].

AAD for music & speech

AAD using linear models on EEG data have been primarily investigated in speech, in cocktail party scenario where both the target sound and distracter sounds are speech signals [23, 114, 121, 176, 221, 242]. In such situations, backward models have been proven successful for auditory attention decoding, even in the presence of noise [7, 8, 12, 14, 67, 98], or in the presence of multiple distracters [98, 114, 199].

The backward modeling has also been applied to music. The technique can successfully reconstruct musical sound envelopes when listened to in isolation [75, 77, 249]. In AAD cases, the majority of the work done for music

listening is based on situations where the listener has to attend to one specific instrument within a polyphonic musical piece [9, 47, 113, 215]. While these studies show that backward linear modeling can successfully decode attention during music listening, they do not allow for direct comparison with speech or predict how a versatile AAD would perform in situations where both speech and music are present. When the present project started, only one study had been conducted with soundscapes composed of music and speech, and even though speech/music comparison was not the main focus of this study, results suggested that there might be a difference to explore [114]. Since then, another study [249] found differences in the reconstruction performance between speech and music, with better reconstruction for speech than for music. Such results might indicate different brain processing, specific to speech or music, which needs to be explored to move toward a versatile AAD.

2.5 The brain on music & speech

When building a cortical AAD for speech and music, one fundamental question is how the brain processes either of those sounds. Backward models trained for AAD take advantage of cortical patterns to reconstruct stimuli. Therefore, the underlying question is whether these cortical patterns are similar for speech and music listening. Differences in these cortical patterns would require different model training to adapt to each type of processing specifically.

The neural mechanisms involved in the auditory processing of complex sounds, such as speech and music, are not fully understood, nor how music and speech processing may differ. Neither music nor speech processing relies on only one unitary brain process but implies several sub-processes [182]. From neuroimaging, neurophysiology and cognitive sciences, it seems that both general mechanisms, which can be involved in both speech music listening, and domain-specific mechanisms coexist. Previous research comparing behavioural, subcortical and cortical study emphasise the influence of transfer effect from one domain to another: musical expertise influence speech processing, and, to a lesser extent, speech processing expertise influence the processing of some musical aspect (See [15, 25] for a review). Such transfer effect suggests that some auditory mechanisms are common in speech and music listening. In addition, multiple fMRI studies have shown some neural overlapping between music and speech listening [134, 184]. Similarities between speech and music have also been found for syntax processing (see for review [178]) or between prosody processing and musical patterns [112, 179].

On the other hand, neurophysiological research on music disorders (amusia) or speech disorders (aphasia) revealed a double dissociation between those two types of disorders. Some aphasia patients present no deficit in music perception [232, 245], while most amusia patients present no speech-related disorders [183, 186]. Comparable dissociation also exists between speech pro-

duction (i.e. speaking) and music production (i.e., singing) [181]. It suggests that some distinct processes coexist for either speech or music. It also has been shown that even when speech and music are interleaved (e.g. in a song with lyrics), semantic and harmonic processing are independent [24, 34]. In the case of amusia, dissociation has also been found between pitch processing processes and rhythm perception [125, 180]. It indicates that different features of the input audio signal, either speech or music, might be processed by separate brain sub-mechanisms, which may work independently [182, 185]. Some specialisations may not be domain-specific (with a strict separation between speech and music neuropathways) but instead rely on low-level features and specialised processing, which may be involved in music and speech listening [247].

The separation between speech and music processing has also been observed, via neuroimaging techniques, through the variations in the lateralisation of the brain activity during music and speech listening [223]. Neuroimaging research reveals a higher contribution of the left Auditory Cortex (AC) for speech sounds and a higher contribution of the right AC for musical sounds [45, 222–224]. However, this lateralisation may not be driven by the existence of speech or music domain-specific mechanisms. Subprocesses are specialised in the processing of specific lower features of the auditory input, with a specialisation on temporal processing in the left AC and a specialisation of spectral processing in the right AC [5, 94, 246, 247]. The link between this feature-specific lateralisation and the apparent domain-specific lateralisation may be partially explained by the difference in the importance of the spectro temporal cues for the perception of speech and music: while speech perception relies mainly on temporal cues [5, 202], spectral cues are more critical for melody processing [5, 247].

When using modeling approaches, it is not straightforward to know what brain mechanisms are considered in the model. It could be explored through the selection of factors used in the design of the model or the feature of the input signal used: for instance, when training models with the envelope of the sound signal, it relies on temporal tracking of the sound by the brain, thus it could be assumed that the model would represent temporal processing mechanism and not subprocesses related to pitch [75]. Through the use of different input features linear model can be used as another tool to investigate auditory processing at various levels: either via acoustic or semantic features [39, 75, 76].

In AAD for speech listening, the feature of interest is generally the audio envelope, which is unsurprising when considering the importance of the temporal cues for speech perception [5, 202]. However, the envelope might not be the best feature for music-AAD. Music listening also relies on temporal cues [112], and it has been shown that linear modeling can also successfully reconstruct musical envelopes (See section 2.4). However, temporal modulations of speech generally differ from temporal modulation of music [79], which

might influence the modeling processes. In addition, it has been shown that including spectral information as an input of linear models can increase reconstruction accuracy for music reconstruction [47]. Due to these differences in the brain processes between speech and music listening, multiple questions remain to explore the performance of an envelope-based AAD during music listening and how it differs from speech listening. This project aims to provide some answers to these questions.

2.6 Toward AAD in audio devices

EEG are non-invasive, small and low-cost, compared to other neuroimaging techniques, making it a suitable candidate for BCI in audio products. However, most of the work done for EEG-AAD, including this thesis, relied on data recorded using a full-scalp EEG system. Such devices are usually bulky, connected to a heavy amplifier, and often gel-based, which is unsuitable for PAD implementation at that stage. The PAD implementation that the users could adopt would require a smaller EEG device, which could be embedded within the PAD, and that would not require using gel.

Several smaller, wireless, dry, and wearable EEG systems designs have been designed in recent years. Some have already been implemented into consumer products for BCI[3], gaming[4] , meditation[5] or sleep [6]. The data quality of such smaller devices is often reduced compared to full-scalp montages [85, 144], but depending on the expected measure, they can provide satisfactory results for BCI application or in-the-wild research (e.g. ERP detection [17, 70, 140, 141], spectral analysis [46, 144, 233]).

In the context of BCI for auditory applications, especially for AAD, the optimal design would embed electrodes directly within the audio product: either in the ear canal for earbuds or around the ear for circumaural headphones. Prototypes have been developed for in-ear EEG [131] or around-ear EEG [32]. They both show good performance for the measure of auditory ERP [33, 73, 92, 120], which is promising for using such small EEG devices in PAD-embedded AAD. With this idea of a neuro-steered audio device in mind, several studies investigated the performances of linear AAD with fewer electrodes and smaller EEG devices through electrode reduction. Overall, they show that the performance of a traditional 64-channel EEG montage can be obtained with a smaller number of electrodes: one study showed that 25 channels are sufficient [163], others concluded that optimal performance could be obtained with 20 electrodes [98, 164], and in the present project, it has been shown that four well-located electrodes match the performance of a 64-channel

[3] https://www.emotiv.com/
[4] https://store.neurosky.com/
[5] https://choosemuse.com/
[6] https://www.advancedbrainmonitoring.com/products/sleep-

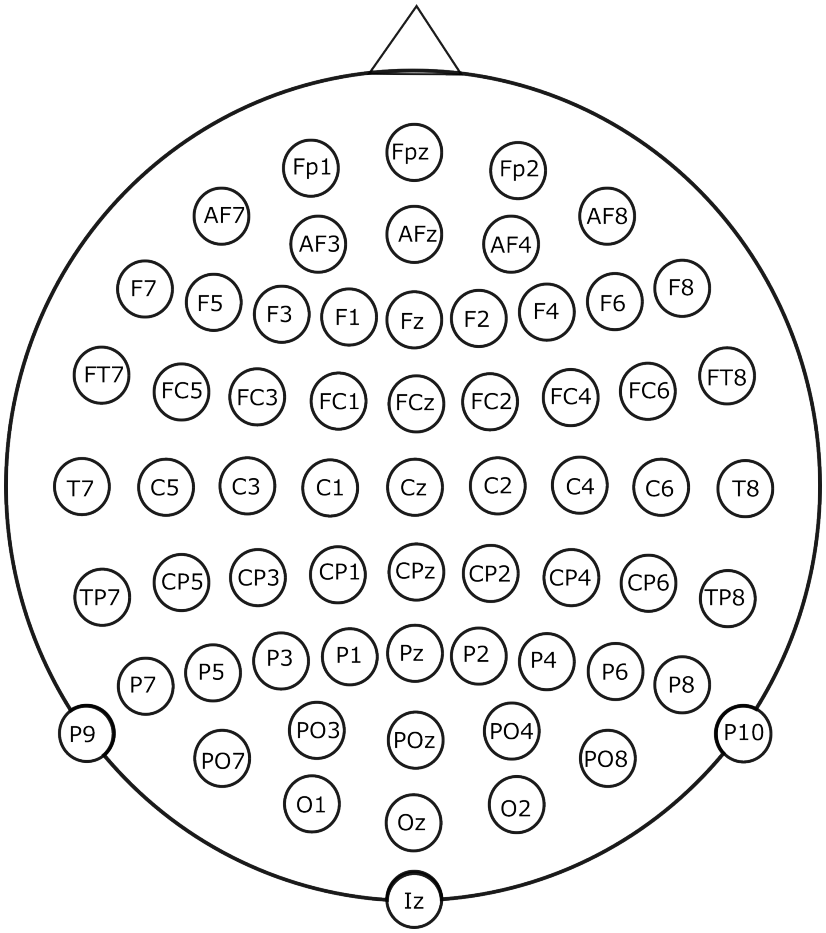

Fig. 8: Map of the 64-channels ull scalp EEG layout used through the present project

EEG montage [206]. Reducing the number of channels can also increase the performance, as it also reduces the presence of irrelevant noise [164].

The optimal placement of the electrodes is primarily located in the temporal regions, around the ear [98, 170–172, 206]. Those regions correspond to the placement of headphones cushion on a listener's head, which would ease the implementation of AAD into headphones. In addition, AAD algorithms have been tested on existing in-ear or around-ear EEG devices. In both cases, the auditory attention decoding is successful, but the decoding performance tends to be reduced compared to full-scalp EEG [93, 122, 162].

3 Contributions

3.1 Paper A: Cortical auditory attention decoding during music and speech listening

Summary

At the time of the experiment design, the question of auditory attention decoding for music listening was not widely investigated. The experimental paradigm to use the stimulus reconstruction approach to musical signals was limited to decoding the attention to one instrument in a polyphonic musical piece [9, 47, 113]. Those studies show that auditory attention decoding through stimulus reconstruction is applicable to musical signals. However, such a setup differs from the initial use-case of the current project, where the goal is to decode the attention paid to a musical stream as a whole (e.g., a song, with the accompaniment and voice included in one single auditory stream) in the presence of distracting sound.

Paper A aimed to test the performance of a linear AAD during music listening in the presence of a distracter (either speech or music) and to compare the results obtained to AAD for speech listening in a similar situation. This paper is based on the research question 1. To that aim, listeners were presented with two audio streams, each coming from a different loudspeaker. They could be two musical pieces, two speech streams or one of each. The two sound sources were separated in space (see Figure 9), and both audios were played at an equal perceived loudness of -23 LUFS (Loudness Unit Full Scale). The idea was to have both target and distracter sounds played at a similar level: no physical factor varies across trials. The only factor influencing each trial is the listener's attention, which focuses on one of the sounds while ignoring the other.

Outcomes

Using linear models based on stimulus reconstruction of the audio envelope, it is possible to reconstruct both the sound where a listener's attention is directed and those that are outside the focus of attention. This effect was achieved for both speech signals and musical signals. However, this study's results showed significant performance variations between speech and music reconstruction, with a reduced reconstruction for musical signals. This effect is consistent for target reconstruction and also for distracter reconstruction. Such results raise multiple questions, and several hypotheses may explain such differences. Exploring these hypotheses became the focus of a second study, presented in PaperD.

Fig. 9: Experimental setup for the experiment described in Papers A, B and D

The differences in stimulus reconstruction are important for a future design of a versatile AAD. As the reconstruction accuracy level obtained for target reconstruction for music is comparable to the ones obtained for decoder reconstruction for speech, it could lead to incorrect classification of the target when music and speech are present in the auditory scene. The results of Paper A also highlight the importance of the training of the decoder. The result, both reconstruction accuracy and decoding success are improved when using a decoder that corresponds to the sound under test, especially for music reconstruction. While this effect is not yet completely understood, it must be considered for future implementation of an AAD. It may require having several decoders available, specialised either for speech or music, depending on the content of the auditory scene. The specialisation of the decoder could be achieved by a relevant selection of the data set used for training and also by taking advantage of the cortical differences in cortical tracking for speech and music, either on the temporal factors (See Paper B) or spatial factors (See Paper C).

3.2 Paper B: Optimal time lags for linear cortical auditory attention detection: differences between speech and music listening

Summary

Traditionally, the stimulus reconstruction approach is performed on a multi-lags model for AAD purposes. In this approach, multiple versions of the EEG data, each delayed by one sample to cover a few hundred milliseconds of lags, are combined for training and testing. Adding lagged versions takes into account the time the brain needs to process the auditory information. While the multi-lagged approach (as used in Paper A) leads to increase reconstruction accuracies [8], training single lags models allows the exploration of temporal variations of the tracking performance over different delays [113, 114, 163, 176]. The motivation for this paper was two-fold:

- Get a better understanding of the temporal aspects of cortical tracking that can inform on underlying mechanisms: differences between speech and music listening and the effect of attention. (RQ 3-1)
- Find time lags that enhance reconstruction for speech and music listening. If differences exist, they could be used to tune specific decoders specialised in either speech or music reconstruction. Tuning the decoder on temporal factor could help to increase the reconstruction performance to counterbalance the variation of performance found in Paper A.

Outcomes

This study confirmed results previously obtained on speech [163, 176] and music listening [113, 114]. Two peaks of increased reconstruction can be found for all stimulus types, which is coherent with the hypothesis of hierarchical processing of the auditory scene (see section 2.3). The first peak is comparable for music and speech and insensitive to attentional factors. It suggests that primary auditory tracking is common for all types of audio stimuli and not affected by top-down attention at early latencies. While attention affects both music and speech tracking at middle/late latencies, the differences in the delay of the second optimal peak suggest that there may be different cortical processing for the attentive tracking of music and speech. Such differences in the timing of cortical tracking might be beneficial to train a music or speech-specific decoder. However, further work is necessary to explore the temporal factor and to evaluate the trade-off of using more focused time lags in an AAD context [8].

3.3 Paper C: Electrodes selection for cortical auditory attention decoding with EEG during speech and music listening

Summary

As the imagined use case of the project is to advance toward the future development of AAD embedded in PAD, reducing the number of electrodes necessary is essential. The results obtained in Paper A are based on a full-scalp, 64-channel EEG cap. It seems unlikely that any user would agree to be equipped with such bulky equipment. Therefore, we should reduce the number of electrodes necessary for an AAD and their placement. Several options for smaller EEG devices are proposed in the academic or industrial worlds (See section 2.6). Some, such as the in-ear EEG [148]) or around-ear EEG [32], could be a good fit for PAD, as the placement of the electrodes could fit the shape of earbuds or headphones. However, it is not clear if the placement of electrodes is optimal for cortical AAD.

This paper implemented a data-driven approach to explore how many electrodes can be sufficient to obtain AAD performance comparable to a full-scalp AAD and where they should be placed. Each type of stimuli (music and speech) was explored individually to highlight potential differences in the optimal electrodes' localisations (RQ 3-2). This study also investigated if the placement of optimal electrodes differs for speech or music listening (RQ 4).

Outcomes

The results obtained through this study are promising for developing AAD for smaller EEG devices, as only four electrodes can perform as well as a full-scalp system. Furthermore, optimal electrodes are distributed mainly in the temporal areas. Considering that the spatial resolution of EEG is poor [44], we could imagine that integrating electrodes into an earbud or cushion of headphones would place the electrodes close enough to the relevant temporal regions to provide satisfying results for a PAD-embedded AAD.

The results also show different lateralisation between speech- and music-optimal electrodes. It adds to existing work on the hypothesis that lateralised processes are involved in cortical audio processing [5, 94, 247]. Furthermore, it could be used to fit a specialised decoder, which would utilise specific electrodes for either speech or music decoding, which could be of interest to counterweight the reduced performance obtained when a listener is listening to music in the presence of speech.

3.4 Paper D: Cortical linear encoding and decoding of sounds: Similarities and differences between naturalistic speech and music listening.

Summary

The results obtained from the study described in Paper A show that the reconstruction accuracy performance differs significantly depending on the type of signal reconstructed: music can be reconstructed to a weaker extent than speech reconstruction. It is consistent with other work that compares linear models for speech and music listening [249]. Nonetheless, as seen in Paper A, it may cause some issues for versatile AAD, as target music's mean reconstruction accuracy is comparable to a speech distracter's reconstruction accuracy. The study presented in Paper D aimed to explore reasons that could explain these differences between speech and music reconstruction ((RQ 2)). By using speech, music, and mixed stimuli, three hypotheses have been tested: one related to the acoustical properties of the speech signal; one related to semantical content processing; and one focusing on the model design, especially the relevance of using the audio envelope as an input for the linear models. The experiment was designed with a top-down approach. Starting from naturalistic stimuli, this experiment aimed to explore the potential influence of those factors (acoustic, semantic or model design) to untangle and highlight their contribution to the results obtained in Paper A.

While the underlying motivation is auditory attention, this second study aims to reduce the impact of the attentional factor and facilitate the auditory task by using a single auditory stream per trial. In the versatile AAD scenario, the attention states might differ depending on the type of target and the combination target/distracter [96].

In addition to decoding modeling, encoding models were also computed. The encoding models are designed to predict cortical responses from the audio input. The weight of such an encoding system, known as a TRF, can be compared to an ERP [60, 111, 143]. Examining the morphology, amplitude and timing of peak can provide useful information on the cortical processing of each input type.

Outcomes

The experiment conducted in this paper led to the rejection of all the tested hypotheses. No significant differences have been found between any stimulus types under test. It suggests that contrary to the results obtained for the previous experiment, linear models for stimulus and cortical reconstruction based on audio envelope performs as well for speech and music stimuli. It is encouraging for the development of AAD, as, in theory, it should work for

both types of stimuli, and the audio envelope is a good predictor of neural entrainment.

Notwithstanding those results, the results obtained in Paper A remain unexplained. Further investigation should look into other hypotheses to understand why the performance is reduced for active listening. The effect of the task should be investigated: both the target type and distracter task interaction as well as the type of listening (active/passive or for information gathering, entertainment or just as background). It could influence attention, listening effort, environmental awareness, or distraction. All those factors are likely to influence the cortical tracking of the different streams of the sound scene.

In addition to the reconstruction results, the TRFs have been compared between types of listening material. The comparison of TRFs obtained revealed differences between speech and music listening. It could indicate a difference in the cortical processes involved in speech and music.

4 Summary & Conclusion

The goal of this project was two-fold:

- Investigate how auditory attention could be decoded from the brain during music listening in noisy environments.
- Explore how auditory attention decoding and cortical temporal tracking differ between music and speech listening.

At the beginning of the project, following a literature review and pilot experiment, the project focused on the use of backward linear models for AAD. This approach, which relies on reconstructing some audio features from neural signals, has succeeded in both speech and music reconstruction and is effective for AAD in cocktail-party scenarios. However, at the start of this project, it has yet to be tested in situations where listeners may want to listen to either speech or music. This approach also has the advantage of enabling the analysis of continuous cortical recordings. It allows using ecologically valid stimuli such as existing music and natural speech. Due to the industrial context of the current project, the ecological validity of the experiment design and the stimuli used was important: to investigate fundamental phenomena and evaluate the feasibility of future implementation of this research.

The first step of this project aimed to fill this research gap. An experiment was designed to test how such a cocktail-party AAD algorithm performed during music listening (RQ 1, Paper A). The results show that the stimulus reconstruction approach can successfully reconstruct speech and music, leading to a positive answer to the first research question. However, differences emerge between the two types of sounds. Music, overall, was reconstructed to

a lower extent than speech, and these differences had a detrimental influence on the performance of the AAD when this attention was directed to music. The results are consistent with previous findings [114, 249]. But it led to a new question, the RQ 2: what is causing such differences between speech and music reconstruction?

Several hypotheses have been formulated to explain such differences, either based on the acoustic differences between speech and music, high-level factors carried by the signal, specificity of the models constructed, or cognitive factors. A second experiment was designed to test some hypotheses (Paper D). Due to the many factors that may influence the results, this experiment was conceived with a top-down approach. The goal was to highlight some factors' putative influence (or lack of influence) to enable further research on the influential factors. Results lead to the rejections of all hypotheses under test and suggest that obtaining similar reconstruction accuracy for both speech and music is possible. While such results were not expected, especially as they seem to contradict the effect observed, they show that for AAD, the use of envelope-based linear AAD can be suitable for musical-cocktail party scenarios.

In parallel to the results on stimulus reconstruction, supplementary analyses were conducted to explore differences between the models for speech and music listening, as stated in RQ 3. On the data from the first listening experiment, spatial (Paper C) and temporal (Paper B) patterns of the performance of music-fitted and speech-fitted models were compared. On the data from the second experiment, forward models were used to compare the TRF for each listening condition (Paper D). All those results conclude that the cortical mechanisms captured by the models differ between speech and music listening. Those results aligned with existing literature on how the brain processes involved in music and speech perception differ (see section 2.5).

Overall, this project significantly contributes to research on music and speech listening. The findings highlight the differences in cortical mechanisms involved in processing these two types of sounds. The employed modeling approach enables the exploration of cortical responses during extended audio stimuli, underscoring the presence of speech/music differences during naturalistic listening.

From an innovation and industrial standpoint, the project's results are promising. The neural tracking and cortical modeling methods successfully reconstructed temporal features, particularly the envelope, for both music and speech above chance levels. The results suggest that the envelope-based cortical model performs similarly well for music than speech. This indicates the potential for further exploration and development of auditory attention decoding technologies that could be incorporated into portable audio devices. In addition, results from Paper C, based on RQ 4, showed that AAD does not require electrodes distributed on the whole scalp of the user and that good performance can be obtained with a smaller number of electrodes, both

for music and speech listening. Furthermore, the optimal placement of the electrodes corresponds to the standard placement of headphones' cushions, which can ease the implementation.

However, several questions remain unanswered. The differences observed between music-AAD and the speech listening scenario, consistent with previous studies [114, 249], still require further explanation. Although some hypotheses were ruled out in this project, the underlying phenomenon warrants continued investigation in future research.

4.1 Future Work

The second experiment (Paper D) showed that differences in reconstruction accuracy are not driven by acoustic or semantic aspects of the input stimuli nor by choice of feature for the model design. However, the speech/music differences observed in Paper A remain unexplained. One of the hypotheses that have not been explored is that the listener's attentional state may vary between speech and music depending on the nature of the distracting sound (speech or music). This effect can be observed in the subjective rating presented in Paper A: e.g. background music is less distracting than speech when attending to a speech signal. One may also speculate that, despite the instruction and participant feedback, the listeners' attention was not always focused on the target signal [123, 151]. As seen in section 2.1, music listening is often used as a background activity and may not require the same attentional level compared to speech listening. In the experiment described in Paper A, the difference in AAD performance might be because listeners were able to direct their attention to it and ignore the distracter when the target was speech. When the target was music, especially with speech distracters, the attention might have been more divided to allow some tracking of the background speech (both perceptually and on the cortical level) [115]. Nevertheless, this hypothesis is mostly putative at that stage. Further research would be needed to explore the attentional state during music and speech listening in such situations. It would also be interesting to look at the different modes of listening (e.g., active, passive, background listening, for information gathering or entertainment) in noisy situations to better understand the mechanisms (perceptual or neural) in place during everyday music listening.

Using linear models can be useful to explore the neural mechanisms during such naturalistic listening, as it can be used with continuous cortical recording. For instance, the present work shows differences between speech and music listening when training models with temporal cues (i.e., envelope). Future work could expand this investigation by exploring the influence of other cues (e.g., spectral, spatial) on the modeling of cortical activity. It could be a step toward better comprehending the cortical mechanisms involved in auditory perception, and untangling when and where auditory processing differs

between speech and music. It could also help to enhance the performance of reconstruction or prediction accuracy of the linear model by, for instance, incorporating multiple audio features as the audio input.

Concerning the idea of a versatile AAD that could be integrated into PAD, future work may focus on improving attention decoding performance. While the results obtained in Paper A demonstrate the feasibility of attention decoding (with success rates of 70 to 80%), it may be insufficient to ensure the user adoption of such technology. To that aim, several tracks could be explored. One could be to enhance the performance of the decoding through different model designs or by combining several modeling approaches: non-linear models or canonical correlation analysis (see section 2.1). The specificities of music and speech modeling could also be used to enhance the performance of an AAD, both in the temporal factor, as seen in Paper B, and spatial, as seen in Paper C. Using either specific time-lags or subsets of electrodes could maximise the performance, and it would be interesting to see combining those spatial and temporal aspects to see how it influences the results. Another way would be to explore other physiological or cortical measures of attention that could be combined with the AAD to increase the decoding robustness. For example, spatial decoding of attention (is the listener attending to a sound coming from the left or the right) could be an interesting addition to AAD, either through eye gaze [189] or alpha power lateralisation [21, 110]. It also has to be noted that all the work done in the present project relied on subject-independent models, where models were trained with a dataset from several participants, as the focus of the project was to investigate general differences between types of auditory input. Training individualised models fitted to a specific listener has been shown to improve decoding accuracy [61, 176]. This aspect should also be considered for potential implementation, with some unsupervised learning that could adapt to individual users [100].

References

[1] H. Akbari, B. Khalighinejad, J. L. Herrero, A. D. Mehta, and N. Mesgarani, "Towards reconstructing intelligible speech from the human auditory cortex," *Scientific Reports*, vol. 9, no. 1, p. 874, jan 2019. [Online]. Available: https://www.nature.com/articles/s41598-018-37359-z

[2] C. Alain, "Breaking the wave: Effects of attention and learning on concurrent sound perception," *Hearing Research*, vol. 229, no. 1-2, pp. 225–236, jul 2007. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0378595507000184

[3] C. Alain and L. J. Bernstein, "Auditory Scene Analysis," *Music Perception*, vol. 33, no. 1, pp. 70–82, sep 2015. [Online]. Available: https://online.ucpress.edu/mp/article/33/1/70/62681/Auditory-Scene-AnalysisTales-from-Cognitive

[4] C. Alain and K. Tremblay, "The Role of Event-Related Brain Potentials in Assessing Central Auditory Processing," *Journal of the American Academy of Audiology*, vol. 18, no. 07, pp. 573–589, jul 2007. [Online]. Available: http://www.thieme-connect.de/DOI/DOI?10.3766/jaaa.18.7.5

[5] P. Albouy, L. Benjamin, B. Morillon, and R. J. Zatorre, "Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody," *Science*, vol. 367, no. 6481, pp. 1043–1047, feb 2020. [Online]. Available: https://www.science.org/doi/10.1126/science.aaz3468

[6] E. Alickovic, T. Lunner, F. Gustafsson, and L. Ljung, "A Tutorial on Auditory Attention Identification Methods," *Frontiers in Neuroscience*, vol. 13, no. March, pp. 1–17, mar 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2019.00153/full

[7] E. Alickovic, T. Lunner, D. Wendt, L. Fiedler, R. Hietkamp, E. H. N. Ng, and C. Graversen, "Neural Representation Enhanced for Speech and Reduced for Background Noise With a Hearing Aid Noise Reduction Scheme During a Selective Attention Task," *Frontiers in Neuroscience*, vol. 14, no. September, pp. 1–16, sep 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2020.00846/full

[8] E. Alickovic, E. H. N. Ng, L. Fiedler, S. Santurette, H. Innes-Brown, and C. Graversen, "Effects of Hearing Aid Noise Reduction on Early and Late Cortical Representations of Competing Talkers in Noise," *Frontiers in Neuroscience*, vol. 15, no. March, mar 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2021.636060/full

[9] W. W. An, B. Shinn-Cunningham, H. Gamper, D. Emmanouilidou, D. Johnston, M. Jalobeanu, E. Cutrell, A. Wilson, K.-J. Chiang, and I. Tashev, "Decoding Music Attention from "EEG Headphones": A User-Friendly Auditory Brain-Computer Interface," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, IEEE. IEEE, jun 2021, pp. 985–989. [Online]. Available: https://ieeexplore.ieee.org/document/9414492/

[10] J. J. Arango, P. A. Barbosa, and E. Astorga, "Headphone Listening Cultures," *Revista Música*, vol. 20, no. 1, pp. 473–492, jul 2020. [Online]. Available: https://www.revistas.usp.br/revistamusica/article/view/172212

[11] B. Arons, "A review of the cocktail party effect," *Journal of the American Voice I/O society*, vol. 12, no. 7, pp. 25–50, 1992. [Online]. Available: http://ocw.mit.edu/terms.

[12] A. Aroudi and S. Doclo, "EEG-based auditory attention decoding: Impact of reverberation, noise and interference reduction," in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, oct 2017, pp. 3042–3047. [Online]. Available: http://ieeexplore.ieee.org/document/8123092/

[13] A. Aroudi, B. Mirkovic, M. De Vos, and S. Doclo, "Auditory attention decoding with EEG recordings using noisy acoustic reference signals," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, mar 2016, pp. 694–698. [Online]. Available: http://ieeexplore.ieee.org/document/7471764/

[14] A. Aroudi, B. Mirkovic, M. De Vos, and S. Doclo, "Impact of Different Acoustic Components on EEG-Based Auditory Attention Decoding in Noisy and Reverberant Conditions," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 4, pp. 652–663, apr 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8662636/

[15] S. S. Asaridou and J. M. McQueen, "Speech and music shape the listening brain: evidence for shared domain-general mechanisms," *Frontiers in Psychology*, vol. 4, no. JUN, pp. 1–14, 2013. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fpsyg.2013.00321/abstract

[16] J. Aydelott, Z. Jamaluddin, and S. Nixon Pearce, "Semantic processing of unattended speech in dichotic listening," *The Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 964–975, aug 2015. [Online]. Available: http://dx.doi.org/10.1121/1.4927410https://pubs.aip.org/asa/jasa/article/138/2/964-975/917177

[17] N. A. Badcock, P. Mousikou, Y. Mahajan, P. de Lissa, J. Thie, and G. McArthur, "Validation of the Emotiv EPOC ® EEG gaming system for measuring research quality auditory ERPs," *PeerJ*, vol. 1, p. e38, feb 2013. [Online]. Available: https://peerj.com/articles/38

[18] D. Barniv and I. Nelken, "Auditory Streaming as an Online Classification Process with Evidence Accumulation," *PLOS ONE*, vol. 10, no. 12, p. e0144788, dec 2015. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0144788

[19] K. C. Barrett, R. Ashley, D. L. Strait, E. Skoe, C. J. Limb, and N. Kraus, "Multi-Voiced Music Bypasses Attentional Limitations in the Brain," *Frontiers in Neuroscience*, vol. 15, no. January, pp. 1–8, jan 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2021.588914/full

[20] S. Basu, B. Clarkson, and A. Pentland, "Smart headphones: enhancing auditory awareness through robust speech detection and source localization," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 5. IEEE, 2001, pp. 3361–3364. [Online]. Available: http://ieeexplore.ieee.org/document/940379/

[21] A. Bednar and E. C. Lalor, "Neural tracking of auditory motion is reflected by delta phase and alpha power of EEG," *NeuroImage*, vol. 181, no. June, pp. 683–691, nov 2018. [Online]. Available: https://doi.org/10.1016/j.neuroimage.2018.07.054https://linkinghub.elsevier.com/retrieve/pii/S1053811918306694

[22] M. A. Bee, "Treefrogs as animal models for research on auditory scene analysis and the cocktail party problem," *International Journal of Psychophysiology*, vol. 95, no. 2, pp. 216–237, feb 2015. [Online]. Available: http://dx.doi.org/10.1016/j.ijpsycho.2014.01.004https://linkinghub.elsevier.com/retrieve/pii/S0167876014000051

[23] J. Belo, M. Clerc, and D. Schön, "EEG-Based Auditory Attention Detection and Its Possible Future Applications for Passive BCI," *Frontiers in Computer Science*, vol. 3, p. 661178, apr 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcomp.2021.661178/full

[24] M. Besson, F. Faïta, I. Peretz, A.-M. Bonnel, and J. Requin, "Singing in the Brain: Independence of Lyrics and Tunes," *Psychological Science*, vol. 9, no. 6, pp. 494–498, nov 1998. [Online]. Available: http://journals.sagepub.com/doi/10.1111/1467-9280.00091

[25] M. Besson, J. Chobert, and C. Marie, "Transfer of Training between Music and Speech: Common Processing, Attention, and Memory," *Frontiers in Psychology*, vol. 2, no. MAY, pp. 1–12, 2011. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fpsyg.2011.00094/abstract

[26] V. Best, F. J. Gallun, S. Carlile, and B. G. Shinn-Cunningham, "Binaural interference and auditory grouping," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 1070–1076, feb 2007. [Online]. Available: https://pubs.aip.org/asa/jasa/article/121/2/1070-1076/920923

[27] V. Best, E. J. Ozmeral, N. Kopčo, and B. G. Shinn-Cunningham, "Object continuity enhances selective auditory attention," *Proceedings of the National Academy of Sciences*, vol. 105, no. 35, pp. 13 174–13 178, sep 2008. [Online]. Available: https://pnas.org/doi/full/10.1073/pnas.0803718105

[28] C. Bey and S. McAdams, "Schema-based processing in auditory scene analysis," *Perception & Psychophysics*, vol. 64, no. 5, pp. 844–854, jul 2002. [Online]. Available: http://link.springer.com/10.3758/BF03194750

[29] E. Bigand, S. McAdams, and S. Forêt, "Divided attention in music," *International Journal of Psychology*, vol. 35, no. 6, pp. 270–278, dec 2000. [Online]. Available: http://doi.wiley.com/10.1080/002075900750047987

[30] J. K. Bizley and Y. E. Cohen, "The what, where and how of auditory-object perception," *Nature Reviews Neuroscience*, vol. 14, no. 10, pp. 693–707, oct 2013. [Online]. Available: https://inatal.org/component/content/article/66-lactancia/lactancia-materna-por-que/394-que-beneficios-le-aporta-a-la-mama-amamantar-a-su-bebe.htmlhttp://www.nature.com/articles/nrn3565

[31] J. Blauert, *Spatial Hearing*. The MIT Press, oct 1996. [Online]. Available: https://direct.mit.edu/books/book/4885/Spatial-HearingThe-Psychophysics-of-Human-Sound

[32] M. G. Bleichner and S. Debener, "Concealed, Unobtrusive Ear-Centered EEG Acquisition: cEEGrids for Transparent EEG," *Frontiers in Human Neuroscience*, vol. 11, no. April, pp. 1–14, apr 2017. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fnhum.2017.00163/full

[33] M. G. Bleichner, B. Mirkovic, and S. Debener, "Identifying auditory attention with ear-EEG: cEEGrid versus high-density cap-EEG comparison," *Journal of Neural Engineering*, vol. 13, no. 6, p. 066004, dec 2016. [Online]. Available: http://dx.doi.org/10.1088/1741-2560/13/6/066004https://iopscience.iop.org/article/10.1088/1741-2560/13/6/066004

[34] A. M. Bonnel, F. Faita, I. Peretz, and M. Besson, "Divided attention between lyrics and tunes of operatic songs: Evidence for independent processing," *Perception & Psychophysics*, vol. 63, no. 7, pp. 1201–1213, oct 2001. [Online]. Available: http://link.springer.com/10.3758/BF03194534

[35] A. S. Bregman, *Auditory Scene Analysis*. The MIT Press, 1990. [Online]. Available: https://direct.mit.edu/books/book/3887/auditory-scene-analysisthe-perceptual-organization

[36] S. L. Bressler and M. Ding, "Event-Related Potentials," in *Wiley Encyclopedia of Biomedical Engineering*. Hoboken, NJ, USA: John Wiley & Sons, Inc., apr 2006, no. 6. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/9780471740360.ebs0455

[37] D. E. Broadbent, *Perception and communication.*, P. Press, Ed. Elmsford: Pergamon Press, 1958. [Online]. Available: http://content.apa.org/books/10037-000

[38] C. Brodbeck, A. Jiao, L. E. Hong, and J. Z. Simon, "Neural speech restoration at the cocktail party: Auditory cortex recovers masked speech of both attended and ignored speakers," *PLOS Biology*, vol. 18, no. 10, p. e3000883, oct 2020. [Online]. Available: http://dx.doi.org/10.1371/journal.pbio.3000883https://dx.plos.org/10.1371/journal.pbio.3000883

[39] M. P. Broderick, A. J. Anderson, G. M. Di Liberto, M. J. Crosse, and E. C. Lalor, "Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech," *Current Biology*, vol. 28, no. 5, pp. 803–809.e3, mar 2018. [Online]. Available: https://doi.org/10.1016/j.cub.2018.01.080https://linkinghub.elsevier.com/retrieve/pii/S0960982218301465

[40] M. P. Broderick, A. J. Anderson, and E. C. Lalor, "Semantic Context Enhances the Early Auditory Encoding of Natural Speech," *The Journal of Neuroscience*, vol. 39, no. 38, pp. 7564–7575, sep 2019. [Online]. Available: https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.0584-19.2019

[41] A. W. Bronkhorst, "The cocktail-party problem revisited: early processing and selection of multi-talker speech," *Attention, Perception, & Psychophysics*, vol. 77, no. 5, pp. 1465–1487, jul 2015. [Online]. Available: http://link.springer.com/10.3758/s13414-015-0882-9

[42] M. Bull, "The Seduction of Sound in Consumer Culture," *Journal of Consumer Culture*, vol. 2, no. 1, pp. 81–101, mar 2002. [Online]. Available: http://journals.sagepub.com/doi/10.1177/146954050200200104

[43] ——, "No dead air! The iPod and the culture of mobile listening," *Leisure studies*, vol. 24, no. 4, pp. 343–355, 2005.

[44] B. Burle, L. Spieser, C. Roger, L. Casini, T. Hasbroucq, and F. Vidal, "Spatial and temporal resolutions of EEG: Is it really black and white? A scalp current density view," *International Journal of Psychophysiology*, vol. 97, no. 3, pp. 210–220, sep 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0167876015001865

[45] D. E. Callan, V. Tsytsarev, T. Hanakawa, A. M. Callan, M. Katsuhara, H. Fukuyama, and R. Turner, "Song and speech: Brain regions involved with perception and covert production," *NeuroImage*, vol. 31, no. 3, pp. 1327–1342, jul 2006. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1053811906000553

[46] C. Cannard, H. Wahbeh, and A. Delorme, "Validating the wearable MUSE headset for EEG spectral analysis and Frontal Alpha Asymmetry,"

in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, dec 2021, pp. 3603–3610. [Online]. Available: https://ieeexplore.ieee.org/document/9669778/

[47] G. Cantisani, S. Essid, and G. Richard, "EEG-Based Decoding of Auditory Attention to a Target Instrument in Polyphonic Music," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, vol. 2019-Octob, no. 765068. IEEE, oct 2019, pp. 80–84. [Online]. Available: https://ieeexplore.ieee.org/document/8937219/

[48] S. Carlile, "Active listening: Speech intelligibility in noisy environments," *Acoustics Australia*, vol. 42, no. 2, pp. 90–96, 2014.

[49] R. P. Carlyon, "How the brain separates sounds," *Trends in Cognitive Sciences*, vol. 8, no. 10, pp. 465–471, oct 2004. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364661304002128

[50] C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[51] ——, *On Human Communication*, M. Press, Ed., Cambridge, MA, USA, 1957.

[52] M. Chion, *Guide To Sound Objects*, Buchet Chastel, Ed., 1983.

[53] I. Choi, S. Rajaram, L. A. Varghese, and B. G. Shinn-Cunningham, "Quantifying attentional modulation of auditory-evoked cortical responses from single-trial electroencephalography," *Frontiers in human neuroscience*, vol. 7, p. 115, 2013.

[54] ——, "Quantifying attentional modulation of auditory-evoked cortical responses from single-trial electroencephalography," *Frontiers in Human Neuroscience*, vol. 7, no. APR 2013, pp. 1–19, 2013. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fnhum.2013.00115/abstract

[55] M. M. Chun, J. D. Golomb, and N. B. Turk-Browne, "A Taxonomy of External and Internal Attention," *Annual Review of Psychology*, vol. 62, no. 1, pp. 73–101, jan 2011. [Online]. Available: https://www.annualreviews.org/doi/10.1146/annurev.psych.093008.100427

[56] G. Ciccarelli, M. Nolan, J. Perricone, P. T. Calamia, S. Haro, J. O'Sullivan, N. Mesgarani, T. F. Quatieri, and C. J. Smalt, "Comparison of Two-Talker Attention Decoding from EEG with Nonlinear Neural Networks and Linear Methods," *Scientific Reports*, vol. 9, no. 1, p. 11538, aug 2019. [Online]. Available: https://www.nature.com/articles/s41598-019-47795-0

[57] A. R. A. Conway, N. Cowan, and M. F. Bunting, "The cocktail party phenomenon revisited: The importance of working memory capacity," *Psychonomic Bulletin & Review*, vol. 8, no. 2, pp. 331–335, jun 2001. [Online]. Available: http://link.springer.com/10.3758/BF03196169

[58] N. Cowan and N. L. Wood, "Constraints on Awareness, Attention, Processing, and Memory: Some Recent Investigations with Ignored Speech," *Consciousness and Cognition*, vol. 6, no. 2-3, pp. 182–203, jun 1997. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1053810097903007

[59] M. J. Crosse, J. S. Butler, and E. C. Lalor, "Congruent Visual Speech Enhances Cortical Entrainment to Continuous Auditory Speech in Noise-Free Conditions," *The Journal of Neuroscience*, vol. 35, no. 42, pp. 14 195–14 204, oct 2015. [Online]. Available: https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.1829-15.2015

[60] M. J. Crosse, G. M. Di Liberto, A. Bednar, and E. C. Lalor, "The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli," *Frontiers in Human Neuroscience*, vol. 10, no. NOV2016, pp. 1–14, nov 2016. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fnhum.2016.00604/full

[61] M. J. Crosse, N. J. Zuk, G. M. Di Liberto, A. R. Nidiffer, S. Molholm, and E. C. Lalor, "Linear Modeling of Neurophysiological Responses to Speech and Other Continuous Stimuli: Methodological Considerations for Applied Research," *Frontiers in Neuroscience*, vol. 15, no. November, nov 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2021.705621/full

[62] R. Cusack, J. Decks, G. Aikman, and R. P. Carlyon, "Effects of Location, Frequency Region, and Time Course of Selective Attention on Auditory Scene Analysis." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 30, no. 4, pp. 643–656, 2004. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/0096-1523.30.4.643

[63] P. Dalton and N. Fraenkel, "Gorillas we have missed: Sustained inattentional deafness for dynamic events," *Cognition*, vol. 124, no. 3, pp. 367–372, sep 2012. [Online]. Available: http://dx.doi.org/10.1016/j.cognition.2012.05.012https://linkinghub.elsevier.com/retrieve/pii/S0010027712001047

[64] P. Dalton and R. W. Hughes, "Auditory attentional capture: implicit and explicit approaches," *Psychological Research*, vol. 78, no. 3, pp. 313–320, may 2014. [Online]. Available: http://link.springer.com/10.1007/s00426-014-0557-5

[65] C. J. Darwin, "Auditory grouping," *Trends in Cognitive Sciences*, vol. 1, no. 9, pp. 327–333, dec 1997. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364661397010978

[66] C. Darwin, "Listening to speech in the presence of other sounds," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 363, no. 1493, pp. 1011–1021, mar 2008. [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rstb.2007.2156

[67] N. Das, A. Bertrand, and T. Francart, "EEG-based auditory attention detection: boundary conditions for background noise and speaker positions," *Journal of Neural Engineering*, vol. 15, no. 6, p. 066017, dec 2018. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2552/aae0a6

[68] A. de Cheveigné, G. M. Di Liberto, D. Arzounian, D. D. Wong, J. Hjortkjær, S. Fuglsang, and L. C. Parra, "Multiway canonical correlation analysis of brain data," *NeuroImage*, vol. 186, no. June 2018, pp. 728–740, feb 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1053811918321049

[69] A. de Cheveigné, D. D. Wong, G. M. Di Liberto, J. Hjortkjær, M. Slaney, and E. Lalor, "Decoding the auditory brain with canonical component analysis,"

NeuroImage, vol. 172, no. January, pp. 206–216, may 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1053811918300338

[70] S. Debener, F. Minow, R. Emkes, K. Gandras, and M. de Vos, "How about taking a low-cost, small, and wireless EEG for a walk?" *Psychophysiology*, vol. 49, no. 11, pp. 1617–1621, nov 2012. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/j.1469-8986.2012.01471.x

[71] R. C. DeCharms and A. Zador, "Neural Representation and the Cortical Code," *Annual Review of Neuroscience*, vol. 23, no. 1, pp. 613–647, mar 2000. [Online]. Available: https://www.annualreviews.org/doi/10.1146/annurev.neuro.23.1.613

[72] S. Denham and I. Winkler, "Auditory Perceptual Organization," in *Encyclopedia of Computational Neuroscience*. New York, NY: Springer New York, 2014, pp. 1–15. [Online]. Available: https://link.springer.com/10.1007/978-1-4614-7320-6{_}100-1

[73] F. Denk, M. Grzybowski, S. M. A. Ernst, B. Kollmeier, S. Debener, and M. G. Bleichner, "Event-Related Potentials Measured From In and Around the Ear Electrodes Integrated in a Live Hearing Device for Monitoring Sound Perception," *Trends in Hearing*, vol. 22, p. 233121651878821, jan 2018. [Online]. Available: http://journals.sagepub.com/doi/10.1177/2331216518788219

[74] D. Deutsch, "Grouping Mechanisms in Music," in *The Psychology of Music*. Elsevier, 2013, pp. 183–248. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/B9780123814609000067

[75] G. M. Di Liberto, G. Marion, and S. A. Shamma, "Accurate Decoding of Imagined and Heard Melodies," *Frontiers in Neuroscience*, vol. 15, no. August, pp. 1–11, aug 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2021.673401/full

[76] G. M. Di Liberto, J. A. O'Sullivan, and E. C. Lalor, "Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing," *Current Biology*, vol. 25, no. 19, pp. 2457–2465, oct 2015. [Online]. Available: http://dx.doi.org/10.1016/j.cub.2015.08.030https://linkinghub.elsevier.com/retrieve/pii/S0960982215010015

[77] G. M. Di Liberto, C. Pelofi, S. Shamma, and A. de Cheveigné, "Musical expertise enhances the cortical tracking of the acoustic envelope during naturalistic music listening," *Acoustical Science and Technology*, vol. 41, no. 1, pp. 361–364, jan 2020. [Online]. Available: https://www.jstage.jst.go.jp/article/ast/41/1/41{_}E19257/{_}article

[78] G. M. Di Liberto, D. Wong, G. A. Melnik, and A. de Cheveigné, "Low-frequency cortical responses to natural speech reflect probabilistic phonotactics," *NeuroImage*, vol. 196, no. October 2018, pp. 237–247, aug 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1053811919303234

[79] N. Ding, A. D. Patel, L. Chen, H. Butler, C. Luo, and D. Poeppel, "Temporal modulations in speech and music," *Neuroscience & Biobehavioral Reviews*, vol. 81, pp. 181–187, oct 2017. [Online]. Available: https://doi.org/10.1016/j.neubiorev.2017.02.011https://linkinghub.elsevier.com/retrieve/pii/S0149763416305668

[80] N. Ding and J. Z. Simon, "Emergence of neural encoding of auditory objects while listening to competing speakers." *Proc. Natl. Acad. Sci. U.S.A.*, vol. 109, no. 29, pp. 11 854–11 859, jul 2012. [Online]. Available: https://pnas.org/doi/full/10.1073/pnas.1205381109

[81] ——, "Neural coding of continuous speech in auditory cortex during monaural and dichotic listening," *Journal of Neurophysiology*, vol. 107, no. 1, pp. 78–89, jan 2012. [Online]. Available: https://www.physiology.org/doi/10.1152/jn.00297.2011

[82] J. P. Dmochowski, J. J. Ki, P. DeGuzman, P. Sajda, and L. C. Parra, "Extracting multidimensional stimulus-response correlations using hybrid encoding-decoding of neural activity," *NeuroImage*, vol. 180, no. May 2017, pp. 134–146, oct 2018. [Online]. Available: https://doi.org/10.1016/j.neuroimage.2017.05.037https://linkinghub.elsevier.com/retrieve/pii/S1053811917304299

[83] W. Dowling, "The perception of interleaved melodies," *Cognitive Psychology*, vol. 5, no. 3, pp. 322–337, nov 1973. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0010028573900406

[84] J. Driver, "A selective review of selective attention research from the past century," *British Journal of Psychology*, vol. 92, no. 1, pp. 53–78, feb 2001. [Online]. Available: http://doi.wiley.com/10.1348/000712601162103

[85] M. Duvinage, T. Castermans, T. Dutoit, M. Petieau, T. Hoellinger, C. D. Saedeleer, K. Seetharaman, and G. Cheron, "A P300-based Quantitative Comparison between the Emotiv Epoc Headset and a Medical EEG Device," in *Biomedical Engineering / 765: Telehealth / 766: Assistive Technologies*. Calgary,AB,Canada: ACTAPRESS, 2012. [Online]. Available: http://www.actapress.com/PaperInfo.aspx?paperId=453213

[86] Edison Research and Triton Digital, "The Infinite Dial 2022," Tech. Rep., 2022. [Online]. Available: https://www.edisonresearch.com/the-infinite-dial-2022/

[87] M. Elhilali, "Modeling the Cocktail Party Problem," in *The Auditory System at the Cocktail Party*, 2017, pp. 111–135. [Online]. Available: http://link.springer.com/10.1007/978-3-319-51662-2{_}5

[88] B. Engdahl and L. Aarhus, "Personal Music Players and Hearing Loss: The HUNT Cohort Study," *Trends in Hearing*, vol. 25, p. 233121652110158, jan 2021. [Online]. Available: http://journals.sagepub.com/doi/10.1177/23312165211015881

[89] O. Farmer and S. Smock Jordan, "Experiences of Women Coping With Catcalling Experiences in New York City: A Pilot Study," *Journal of Feminist Family Therapy*, vol. 29, no. 4, pp. 205–225, oct 2017. [Online]. Available: https://doi.org/10.1080/08952833.2017.1373577https://www.tandfonline.com/doi/full/10.1080/08952833.2017.1373577

[90] R. R. Fay and A. N. Popper, "Evolution of hearing in vertebrates: the inner ears and processing," *Hearing Research*, vol. 149, no. 1-2, pp. 1–10, nov 2000. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0378595500001684

[91] A. S. Feng and R. Ratnam, "Neural Basis of Hearing in Real-World Situations," *Annual Review of Psychology*, vol. 51, no. 1, pp. 699–725, feb 2000. [Online].

Available: https://www.annualreviews.org/doi/10.1146/annurev.psych.51.1.699

[92] L. Fiedler, J. Obleser, T. Lunner, and C. Graversen, "Ear-EEG allows extraction of neural responses in challenging listening scenarios — A future technology for hearing aids?" in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, vol. 2016-Octob. IEEE, aug 2016, pp. 5697–5700. [Online]. Available: http://ieeexplore.ieee.org/document/7592020/

[93] L. Fiedler, M. Wöstmann, C. Graversen, A. Brandmeyer, T. Lunner, and J. Obleser, "Single-channel in-ear-EEG detects the focus of auditory attention to concurrent tone streams and mixed speech," *Journal of Neural Engineering*, vol. 14, no. 3, p. 036020, jun 2017. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2552/aa66dd

[94] A. Flinker, W. K. Doyle, A. D. Mehta, O. Devinsky, and D. Poeppel, "Spectrotemporal modulation provides a unifying framework for auditory cortical asymmetries," *Nature Human Behaviour*, vol. 3, no. 4, pp. 393–405, mar 2019. [Online]. Available: http://dx.doi.org/10.1038/s41562-019-0548-zhttps://www.nature.com/articles/s41562-019-0548-z

[95] J. Francombe, "Perceptual evaluation of audio-on-audio Interference in a personal sound zone system," Ph.D. dissertation, University of Surrey, 2014.

[96] J. Francombe, R. Mason, M. Dewhirst, and S. Bech, "Determining the threshold of acceptability for an interfering audio programme," *132nd Audio Engineering Society Convention*, pp. 333–349, 2012.

[97] C. Fuentes, J. Hagberg, and H. Kjellberg, "Soundtracking: music listening practices in the digital age," *European Journal of Marketing*, vol. 53, no. 3, pp. 483–503, apr 2019. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/EJM-10-2017-0753/full/html

[98] S. A. Fuglsang, T. Dau, and J. Hjortkjær, "Noise-robust cortical tracking of attended speech in real-world acoustic scenes," *NeuroImage*, vol. 156, no. April, pp. 435–444, aug 2017. [Online]. Available: http://dx.doi.org/10.1016/j.neuroimage.2017.04.026https://linkinghub.elsevier.com/retrieve/pii/S105381191730318X

[99] S. Geirnaert, "Signal Processing Algorithms for EEG-based Auditory Attention Decoding," Ph.D. dissertation, KU Leuven, 2022.

[100] S. Geirnaert, T. Francart, and A. Bertrand, "Unsupervised Self-Adaptive Auditory Attention Decoding," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 10, pp. 3955–3966, oct 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9416746/

[101] S. Geirnaert, S. Vandecappelle, E. Alickovic, A. de Cheveigne, E. Lalor, B. T. Meyer, S. Miran, T. Francart, and A. Bertrand, "Electroencephalography-Based Auditory Attention Decoding: Toward Neurosteered Hearing Devices," *IEEE Signal Processing Magazine*, vol. 38, no. 4, pp. 89–102, jul 2021. [Online]. Available: http://arxiv.org/abs/2008.04569https://ieeexplore.ieee.org/document/9467380/

[102] M. H. Giard, "Neurophysiological mechanisms of auditory selective attention in humans," *Frontiers in Bioscience*, vol. 5, no. 1, p. d84, 2000. [Online]. Available: https://imrpress.com/journal/FBL/5/3/10.2741/giard

[103] M. Gillis, J. Vanthornhout, J. Z. Simon, T. Francart, and C. Brodbeck, "Neural Markers of Speech Comprehension: Measuring EEG Tracking of Linguistic Speech Representations, Controlling the Speech Acoustics," *The Journal of Neuroscience*, vol. 41, no. 50, pp. 10 316–10 329, dec 2021. [Online]. Available: https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.0812-21.2021

[104] L. Goines and L. Hagler, "Noise Pollution: A Modern Plague," *Southern Medical Journal*, vol. 100, no. 3, pp. 287–294, mar 2007. [Online]. Available: http://sma.org/southern-medical-journal/article/noise-pollution-a-modern-plague

[105] F. Greb, W. Schlotz, and J. Steffens, "Personal and situational influences on the functions of music listening," *Psychology of Music*, vol. 46, no. 6, pp. 763–794, nov 2018. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0305735617724883

[106] A. B. Haake, "Individual music listening in workplace settings," *Musicae Scientiae*, vol. 15, no. 1, pp. 107–129, mar 2011. [Online]. Available: http://journals.sagepub.com/doi/10.1177/1029864911398065

[107] G. Haas, E. Stemasov, M. Rietzler, and E. Rukzio, "Interactive Auditory Mediated Reality," in *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. New York, NY, USA: ACM, jul 2020, pp. 2035–2050. [Online]. Available: https://dl.acm.org/doi/10.1145/3357236.3395493

[108] G. Haas, E. Stemasov, and E. Rukzio, "Can't You Hear Me?" in *Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia*. New York, NY, USA: ACM, nov 2018, pp. 59–69. [Online]. Available: https://dl.acm.org/doi/10.1145/3282894.3282897

[109] L. S. Hamilton and A. G. Huth, "The revolution will not be controlled: natural stimuli in speech neuroscience," *Language, Cognition and Neuroscience*, vol. 35, no. 5, pp. 573–582, jun 2020. [Online]. Available: https://doi.org/10.1080/23273798.2018.1499946https://www.tandfonline.com/doi/full/10.1080/23273798.2018.1499946

[110] S. Haro, H. M. Rao, T. F. Quatieri, and C. J. Smalt, "EEG alpha and pupil diameter reflect endogenous auditory attention switching and listening effort," *European Journal of Neuroscience*, vol. 55, no. 5, pp. 1262–1277, mar 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/ejn.15616

[111] S. Haufe, F. Meinecke, K. Görgen, S. Dähne, J.-D. Haynes, B. Blankertz, and F. Bießmann, "On the interpretation of weight vectors of linear models in multivariate neuroimaging," *NeuroImage*, vol. 87, pp. 96–110, feb 2014. [Online]. Available: http://dx.doi.org/10.1016/j.neuroimage.2013.10.067https://linkinghub.elsevier.com/retrieve/pii/S1053811913010914

[112] M. Hausen, R. Torppa, V. R. Salmela, M. Vainio, and T. Särkämö, "Music and speech prosody: a common rhythm," *Frontiers in Psychology*, vol. 4, no. SEP, pp. 1–16, 2013. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fpsyg.2013.00566/abstract

[113] L. Hausfeld, N. R. Disbergen, G. Valente, R. J. Zatorre, and E. Formisano, "Modulating Cortical Instrument Representations During Auditory Stream Segregation and Integration With Polyphonic Music," *Frontiers in Neuroscience*, vol. 15, no. September, pp. 1–15, sep 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2021.635937/full

[114] L. Hausfeld, L. Riecke, G. Valente, and E. Formisano, "Cortical tracking of multiple streams outside the focus of attention in naturalistic auditory scenes," *NeuroImage*, vol. 181, no. March, pp. 617–626, nov 2018. [Online]. Available: https://doi.org/10.1016/j.neuroimage.2018.07.052https://linkinghub.elsevier.com/retrieve/pii/S1053811918306670

[115] B. Herrmann and I. S. Johnsrude, "A model of listening engagement (MoLE)," *Hearing Research*, vol. 397, p. 108016, nov 2020. [Online]. Available: https://doi.org/10.1016/j.heares.2020.108016https://linkinghub.elsevier.com/retrieve/pii/S0378595520302872

[116] A. Heye and A. Lamont, "Mobile listening situations in everyday life: The use of MP3 players while travelling," *Musicae Scientiae*, vol. 14, no. 1, pp. 95–120, mar 2010. [Online]. Available: http://journals.sagepub.com/doi/10.1177/102986491001400104

[117] N. J. Hill and B. Schölkopf, "An online brain–computer interface based on shifting attention to concurrent streams of auditory stimuli," *Journal of Neural Engineering*, vol. 9, no. 2, p. 026011, apr 2012. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2560/9/2/026011

[118] S. A. Hillyard and M. Kutas, "Electrophysiology of Cognitive Processing," *Annual Review of Psychology*, vol. 34, no. 1, pp. 33–61, jan 1983. [Online]. Available: https://www.annualreviews.org/doi/10.1146/annurev.ps.34.020183.000341

[119] C. R. Holdgraf, J. W. Rieger, C. Micheli, S. Martin, R. T. Knight, and F. E. Theunissen, "Encoding and Decoding Models in Cognitive Electrophysiology," *Frontiers in Systems Neuroscience*, vol. 11, no. September, sep 2017. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fnsys.2017.00061/full

[120] D. Hölle, J. Meekes, and M. G. Bleichner, "Mobile ear-EEG to study auditory attention in everyday life," *Behavior Research Methods*, vol. 53, no. 5, pp. 2025–2036, oct 2021. [Online]. Available: https://link.springer.com/10.3758/s13428-021-01538-0

[121] B. Holtze, M. Jaeger, S. Debener, K. Adiloğlu, and B. Mirkovic, "Are They Calling My Name? Attention Capture Is Reflected in the Neural Tracking of Attended and Ignored Speech," *Frontiers in Neuroscience*, vol. 15, no. March, pp. 1–15, mar 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2021.643705/full

[122] B. Holtze, M. Rosenkranz, M. Jaeger, S. Debener, and B. Mirkovic, "Ear-EEG Measures of Auditory Attention to Continuous Speech," *Frontiers in Neuroscience*, vol. 16, no. May, pp. 1–14, may 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2022.869426/full

[123] M. P. Huet, C. Micheyl, E. Parizet, and E. Gaudrain, "Behavioral Account of Attended Stream Enhances Neural Tracking," *Frontiers in*

Neuroscience, vol. 15, no. December, pp. 1–13, dec 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2021.674112/full

[124] D. Huron, "Tone and Voice: A Derivation of the Rules of Voice-Leading from Perceptual Principles," *Music Perception*, vol. 19, no. 1, pp. 1–64, sep 2001. [Online]. Available: https://online.ucpress.edu/mp/article/19/1/1/62106/Tone-and-Voice-A-Derivation-of-the-Rules-of-Voice

[125] K. L. Hyde and I. Peretz, "Brains That Are out of Tune but in Time," *Psychological Science*, vol. 15, no. 5, pp. 356–360, may 2004. [Online]. Available: http://journals.sagepub.com/doi/10.1111/j.0956-7976.2004.00683.x

[126] IFPI, "IFPI Engaging with music Report 2022," International Federation of the Phonographic Industry, Tech. Rep., 2022. [Online]. Available: https://www.ifpi.org/wp-content/uploads/2021/10/IFPI-Engaging-with-Music-report.pdf

[127] ——, "IFPI Global Music Report 2022- State of the industry," pp. 76–107, 2022. [Online]. Available: https://www.ifpi.org/wp-content/uploads/2022/04/IFPI{_}Global{_}Music{_}Report{_}2022-State{_}of{_}the{_}Industry.pdf

[128] P. Iverson, "Auditory stream segregation by musical timbre: Effects of static and dynamic acoustic attributes." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 21, no. 4, pp. 751–763, 1995. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/0096-1523.21.4.751

[129] I. S. Johnsrude, A. Mackey, H. Hakyemez, E. Alexander, H. P. Trang, and R. P. Carlyon, "Swinging at a Cocktail Party," *Psychological Science*, vol. 24, no. 10, pp. 1995–2004, oct 2013. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0956797613482467

[130] M. Kamalzadeh, D. Baur, and T. Möller, "A survey on music listening and management behaviours," in *Proceedings of the 13th International Society for Music Information Retrieval Conference*, 2012, pp. 373–378.

[131] S. L. Kappel, M. L. Rank, H. O. Toft, M. Andersen, and P. Kidmose, "Dry-Contact Electrode Ear-EEG," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 1, pp. 150–158, jan 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8357918/

[132] E. M. Kaya and M. Elhilali, "Modelling auditory attention," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 372, no. 1714, p. 20160101, feb 2017. [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rstb.2016.0101

[133] L. Kiss and K. J. Linnell, "Making sense of background music listening habits: An arousal and task-complexity account," *Psychology of Music*, vol. 51, no. 1, pp. 89–106, jan 2023. [Online]. Available: http://journals.sagepub.com/doi/10.1177/03057356221089017

[134] S. Koelsch, T. C. Gunter, D. v. Cramon, S. Zysset, G. Lohmann, and A. D. Friederici, "Bach Speaks: A Cortical "Language-Network" Serves the Processing of Music," *NeuroImage*, vol. 17, no. 2, pp. 956–966, oct 2002. [Online]. Available: http://link.springer.com/10.3758/BF03194534https://linkinghub.elsevier.com/retrieve/pii/S1053811902911547

[135] K. Koffka, *Principles Of Gestalt Psychology*. Routledge, oct 2013. [Online]. Available: https://www.taylorfrancis.com/books/9781136306815

[136] Y.-Y. Kong, A. Mullangi, and N. Ding, "Differential modulation of auditory responses to attended and unattended speech in different listening conditions," *Hearing Research*, vol. 316, no. 5, pp. 73–81, oct 2014. [Online]. Available: http://dx.doi.org/10.1016/j.heares.2014.07.009https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3624763/pdf/nihms412728.pdfhttps://linkinghub.elsevier.com/retrieve/pii/S1364661308000600https://linkinghub.elsevier.com/retrieve/pii/S0378595514001270

[137] S. Koreimann, B. Gula, and O. Vitouch, "Inattentional deafness in music," *Psychological Research*, vol. 78, no. 3, pp. 304–312, may 2014. [Online]. Available: http://link.springer.com/10.1007/s00426-014-0552-x

[138] A. E. Krause, A. C. North, and L. Y. Hewitt, "Music-listening in everyday life: Devices and choice," *Psychology of Music*, vol. 43, no. 2, pp. 155–170, mar 2015. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0305735613496860

[139] N. Kriegeskorte and P. K. Douglas, "Interpreting encoding and decoding models," *Current Opinion in Neurobiology*, vol. 55, no. April, pp. 167–179, apr 2019. [Online]. Available: https://doi.org/10.1016/j.conb.2019.04.002https://linkinghub.elsevier.com/retrieve/pii/S0959438818301004

[140] O. E. Krigolson, M. R. Hammerstrom, W. Abimbola, R. Trska, B. W. Wright, K. G. Hecker, and G. Binsted, "Using Muse: Rapid Mobile Assessment of Brain Performance," *Frontiers in Neuroscience*, vol. 15, no. January, pp. 1–11, jan 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2021.634147/full

[141] O. E. Krigolson, C. C. Williams, A. Norton, C. D. Hassall, and F. L. Colino, "Choosing MUSE: Validation of a Low-Cost, Portable EEG System for ERP Research," *Frontiers in Neuroscience*, vol. 11, no. March, pp. 1–10, mar 2017. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fnins.2017.00109/full

[142] E. C. Lalor and J. J. Foxe, "Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution," *European Journal of Neuroscience*, vol. 31, no. 1, pp. 189–193, jan 2010. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/j.1460-9568.2009.07055.x

[143] E. C. Lalor, A. J. Power, R. B. Reilly, and J. J. Foxe, "Resolving Precise Temporal Processing Properties of the Auditory System Using Continuous Stimuli," *Journal of Neurophysiology*, vol. 102, no. 1, pp. 349–359, jul 2009. [Online]. Available: https://www.physiology.org/doi/10.1152/jn.90896.2008

[144] J. LaRocco, M. D. Le, and D.-G. Paeng, "A Systemic Review of Available Low-Cost EEG Headsets Used for Drowsiness Detection," *Frontiers in Neuroinformatics*, vol. 14, no. October, pp. 1–14, oct 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/fninf.2020.553352/full

[145] S. Lepa and A.-K. Hoklas, "How do people really listen to music today? Conventionalities and major turnovers in German audio repertoires,"

Information, Communication & Society, vol. 18, no. 10, pp. 1253–1268, oct 2015. [Online]. Available: https://doi.org/10.1080/1369118X.2015.1037327https://www.tandfonline.com/doi/full/10.1080/1369118X.2015.1037327

[146] A. Li, A. Wang, Z. Nazari, P. Chandar, and B. Carterette, "Do podcasts and music compete with one another? Understanding users' audio streaming habits," in *Proceedings of The Web Conference 2020*. New York, NY, USA: ACM, apr 2020, pp. 1920–1931. [Online]. Available: https://dl.acm.org/doi/10.1145/3366423.3380260

[147] R. Lichenstein, D. C. Smith, J. L. Ambrose, and L. A. Moody, "Headphone use and pedestrian injury and death in the United States: 2004–2011," *Injury Prevention*, vol. 18, no. 5, pp. 287–290, oct 2012. [Online]. Available: https://injuryprevention.bmj.com/lookup/doi/10.1136/injuryprev-2011-040161

[148] D. Looney, C. Park, P. Kidmose, M. L. Rank, M. Ungstrup, K. Rosenkranz, and D. P. Mandic, "An in-the-ear platform for recording electroencephalogram," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, aug 2011, pp. 6882–6885. [Online]. Available: http://ieeexplore.ieee.org/document/6091733/

[149] S. J. Luck, "Event-related potentials." in *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics.*, American Psychological Association, Ed. Washington: American Psychological Association, 2012, pp. 523–546. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.458.6801{&}rep=rep1{&}type=pdfhttp://content.apa.org/books/13619-028

[150] ——, *An Introduction to the Event-Related Potential Technique*, M. Press, Ed., 2014.

[151] S. Makov, D. Pinto, P. Har-shai Yahav, L. M. Miller, and E. Zion Golumbic, ""Unattended, distracting or irrelevant": Theoretical implications of terminological choices in auditory selective attention research," *Cognition*, vol. 231, no. October 2022, p. 105313, feb 2023. [Online]. Available: https://doi.org/10.1016/j.cognition.2022.105313https://linkinghub.elsevier.com/retrieve/pii/S001002772200302X

[152] A. Mamuji, R. Vertegaal, C. Sohn, D. Cheng, and O. N. Kl, "Attentive Headphones : Augmenting Conversational Attention with a Real World TiVo ®," in *Extended Abstracts of CHI*, Portland, Oregon, USA, 2005, pp. 2223–2226.

[153] G. Marinato and D. Baldauf, "Object-based attention in complex, naturalistic auditory streams," *Scientific Reports*, vol. 9, no. 1, p. 2854, feb 2019. [Online]. Available: http://dx.doi.org/10.1038/s41598-019-39166-6https://www.nature.com/articles/s41598-019-39166-6

[154] S. McAdams and J. Bertoncini, "Organization and discrimination of repeating sound sequences by newborn infants," *The Journal of the Acoustical Society of America*, vol. 102, no. 5, pp. 2945–2953, nov 1997. [Online]. Available: https://pubs.aip.org/asa/jasa/article/102/5/2945-2953/558119

[155] S. McAdams and A. S. Bregman, "Hearing Musical Streams," *Computer Music Journal*, vol. 3, no. 4, pp. 26–43, 1979.

[156] J. H. McDermott, "The cocktail party problem," *Current Biology*, vol. 19, no. 22, pp. R1024–R1027, dec 2009. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0960982209016807

[157] J. H. McDermott, D. Wrobleski, and A. J. Oxenham, "Recovering sound sources from embedded repetition," *Proceedings of the National Academy of Sciences*, vol. 108, no. 3, pp. 1188–1193, jan 2011. [Online]. Available: https://pnas.org/doi/full/10.1073/pnas.1004765108

[158] K. L. McDonald and C. Alain, "Contribution of harmonicity and location to auditory object formation in free field: Evidence from event-related brain potentials," *The Journal of the Acoustical Society of America*, vol. 118, no. 3, pp. 1593–1604, sep 2005. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.2000747https://pubs.aip.org/asa/jasa/article/118/3/1593-1604/992922

[159] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, pp. 233–236, may 2012. [Online]. Available: http://www.nature.com/articles/nature11020

[160] C. M. Michel and M. M. Murray, "Towards the utilization of EEG as a brain imaging tool," *NeuroImage*, vol. 61, no. 2, pp. 371–385, jun 2012. [Online]. Available: http://dx.doi.org/10.1016/j.neuroimage.2011.12.039https://linkinghub.elsevier.com/retrieve/pii/S1053811911014418

[161] M. J. Miquel-Romero and J. D. Montoro-Pons, "Consumption habits, perception and positioning of content-access devices in recorded music," *International Journal of Arts Management*, vol. 19, no. 3, pp. 4–18, 2017.

[162] B. Mirkovic, M. G. Bleichner, M. De Vos, and S. Debener, "Target Speaker Detection with Concealed EEG Around the Ear," *Frontiers in Neuroscience*, vol. 10, no. JUL, pp. 1–11, jul 2016. [Online]. Available: http://journal.frontiersin.org/Article/10.3389/fnins.2016.00349/abstract

[163] B. Mirkovic, S. Debener, M. Jaeger, and M. De Vos, "Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications," *Journal of Neural Engineering*, vol. 12, no. 4, p. 046007, aug 2015. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2560/12/4/046007

[164] J. Montoya-Martínez, J. Vanthornhout, A. Bertrand, and T. Francart, "Effect of number and placement of EEG electrodes on measurement of neural tracking of speech," *PLOS ONE*, vol. 16, no. 2, p. e0246769, feb 2021. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0246769

[165] B. C. Moore, H. Gockel, and H. Cockel, "Factors Influencing Sequential Stream Segregation," *Acta Acustica United with Acustica*, vol. 88, no. 3, pp. 320–333, 2002.

[166] N. Moray, "Attention in Dichotic Listening: Affective Cues and the Influence of Instructions," *Quarterly Journal of Experimental Psychology*, vol. 11, no. 1, pp. 56–60, feb 1959. [Online]. Available: http://journals.sagepub.com/doi/10.1080/17470215908416289

[167] F. Mueller and M. Karau, "Transparent hearing," in *CHI '02 Extended Abstracts on Human Factors in Computing Systems*. New York, NY, USA: ACM, apr 2002, pp. 730–731. [Online]. Available: https://dl.acm.org/doi/10.1145/506443.506569

[168] S. Murphy, C. Spence, and P. Dalton, "Auditory perceptual load: A review," *Hearing Research*, vol. 352, pp. 40–48, sep 2017. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0378595516304373

[169] R. Näätänen, P. Paavilainen, T. Rinne, and K. Alho, "The mismatch negativity (MMN) in basic research of central auditory processing: A review," *Clinical Neurophysiology*, vol. 118, no. 12, pp. 2544–2590, dec 2007. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1388245707001939

[170] A. M. Narayanan and A. Bertrand, "The effect of miniaturization and galvanic separation of EEG sensor devices in an auditory attention detection task," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, vol. 2018-July. IEEE, jul 2018, pp. 77–80. [Online]. Available: https://ieeexplore.ieee.org/document/8512212/

[171] ——, "Analysis of miniaturization effects and channel selection strategies for eeg sensor networks with application to auditory attention detection," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 1, pp. 234–244, 2019.

[172] A. M. Narayanan, R. Zink, and A. Bertrand, "EEG miniaturization limits for stimulus decoding with EEG sensor networks," *Journal of Neural Engineering*, vol. 18, no. 5, p. 056042, oct 2021. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2552/ac2629

[173] A. C. North, D. J. Hargreaves, and J. J. Hargreaves, "Uses of Music in Everyday Life," *Music Perception*, vol. 22, no. 1, pp. 41–77, sep 2004. [Online]. Available: https://online.ucpress.edu/mp/article/22/1/41/62190/Uses-of-Music-in-Everyday-Life

[174] A. E. O'Sullivan, C. Y. Lim, and E. C. Lalor, "Look at me when I'm talking to you: Selective attention at a multisensory cocktail party can be decoded using stimulus reconstruction and alpha power modulations," *European Journal of Neuroscience*, vol. 50, no. 8, pp. 3282–3295, oct 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/ejn.14425

[175] J. O'Sullivan, Z. Chen, S. A. Sheth, G. McKhann, A. D. Mehta, and N. Mesgarani, "Neural decoding of attentional selection in multi-speaker environments without access to separated sources," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, jul 2017, pp. 1644–1647. [Online]. Available: https://ieeexplore.ieee.org/document/8037155/

[176] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG," *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, 01 2015. [Online]. Available: https://doi.org/10.1093/cercor/bht355

[177] D. Pandey, S. Saroshe, S. Dixit, and Y. Sabde, "Estimation of prevalence of headphone usage during driving and awareness about their health hazards among medical undergraduates," *International Journal of Community Medicine and Public Health*, vol. 2, no. 2, p. 167, 2015. [Online]. Available: http://ijcmph.com/index.php/ijcmph/article/view/943

[178] A. D. Patel, "Language, music, syntax and the brain," *Nature Neuroscience*, vol. 6, no. 7, pp. 674–681, jul 2003. [Online]. Available: http://www.nature.com/articles/nn1082

[179] A. D. Patel, I. Peretz, M. Tramo, and R. Labreque, "Processing Prosodic and Musical Patterns: A Neuropsychological Investigation," *Brain and Language*, vol. 61, no. 1, pp. 123–144, jan 1998. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0093934X97918629

[180] I. Peretz, "Brain Specialization for Music: New Evidence from Congenital Amusia," in *The Cognitive Neuroscience of Music*. Oxford University Press, jul 2003, pp. 192–203. [Online]. Available: https://academic.oup.com/book/26285/chapter/194530817

[181] ——, "Music, language, and modularity in action," in *Language and Music as Cognitive Systems*. Oxford University Press, oct 2011, pp. 254–268. [Online]. Available: https://academic.oup.com/book/4665/chapter/146841860

[182] I. Peretz and M. Coltheart, "Modularity of music processing," *Nature Neuroscience*, vol. 6, no. 7, pp. 688–691, jul 2003. [Online]. Available: http://www.nature.com/articles/nn1083

[183] I. Peretz, R. Kolinsky, M. Tramo, R. Labrecque, C. Hublet, G. Demeurisse, and S. Belleville, "Functional dissociations following bilateral lesions of auditory cortex," *Brain*, vol. 117, no. 6, pp. 1283–1301, 1994. [Online]. Available: https://academic.oup.com/brain/article-lookup/doi/10.1093/brain/117.6.1283

[184] I. Peretz, D. Vuvan, M.-É. Lagrois, and J. L. Armony, "Neural overlap in processing music and speech," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 370, no. 1664, p. 20140090, mar 2015. [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rstb.2014.0090

[185] I. Peretz and R. J. Zatorre, "Brain Organization for Music Processing," *Annual Review of Psychology*, vol. 56, no. 1, pp. 89–114, feb 2005. [Online]. Available: https://www.annualreviews.org/doi/10.1146/annurev.psych.56.091103.070225

[186] M. Piccirilli, "Modularity of music: evidence from a case of pure amusia," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 69, no. 4, pp. 541–545, oct 2000. [Online]. Available: https://jnnp.bmj.com/lookup/doi/10.1136/jnnp.69.4.541

[187] T. Picton and S. Hillyard, "Human auditory evoked potentials. II: Effects of attention," *Electroencephalography and Clinical Neurophysiology*, vol. 36, no. C, pp. 191–200, jan 1974. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0013469474901564

[188] H. Poikonen, V. Alluri, E. Brattico, O. Lartillot, M. Tervaniemi, and M. Huotilainen, "Event-related brain responses while listening to entire pieces of music," *Neuroscience*, vol. 312, pp. 58–73, 2016.

[189] U. Pomper and M. Chait, "The impact of visual gaze direction on auditory object tracking," *Scientific Reports*, vol. 7, no. 1, p. 4640, jul 2017. [Online]. Available: https://www.nature.com/articles/s41598-017-04475-1

[190] D. Pressnitzer, "Auditory scene analysis: the sweet music of ambiguity," *Frontiers in Human Neuroscience*, vol. 5, no. December, pp. 1–11, 2011. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fnhum.2011.00158/abstract

[191] K. D. Prinsloo and E. C. Lalor, "General auditory and speech-specific contributions to cortical envelope tracking revealed using auditory chimeras," *The Journal of Neuroscience*, vol. 42, no. 41, pp. JN–RM–2735–20, aug 2022. [Online]. Available: https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.2735-20.2022

[192] C. Puffay, B. Accou, L. Bollens, M. J. Monesi, J. Vanthornhout, H. Van hamme, and T. Francart, "Relating EEG to continuous speech using deep neural networks: a review," *arXiv*, feb 2023. [Online]. Available: http://arxiv.org/abs/2302.01736

[193] J. Rämö, S. Bech, and S. H. Jensen, "Validating a real-time perceptual model predicting distraction caused by audio-on-audio interference," *The Journal of the Acoustical Society of America*, vol. 144, no. 1, pp. 153–163, jul 2018. [Online]. Available: http://dx.doi.org/10.1121/1.5045321https://pubs.aip.org/asa/jasa/article/144/1/153-163/855420

[194] W. M. Randall and N. S. Rickard, "Reasons for personal music listening: A mobile experience sampling study of emotional outcomes," *Psychology of Music*, vol. 45, no. 4, pp. 479–495, jul 2017. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0305735616666939

[195] M. Rane, P. Coleman, R. Mason, and S. Bech, "Quantifying headphone listening experience in virtual sound environments using distraction," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2022, no. 1, p. 30, dec 2022. [Online]. Available: https://doi.org/10.1186/s13636-022-00262-7https://asmp-eurasipjournals.springeropen.com/articles/10.1186/s13636-022-00262-7

[196] ——, "Survey of User Perspectives on Headphone Technology," *152nd Audio Engineering Society Convention*, pp. 106–115, 2022.

[197] M. Rivenez, C. J. Darwin, and A. Guillaume, "Processing unattended speech," *The Journal of the Acoustical Society of America*, vol. 119, no. 6, pp. 4027–4040, jun 2006. [Online]. Available: https://pubs.aip.org/asa/jasa/article/119/6/4027-4040/542375

[198] P. Schaeffer, *Traité des objets musicaux*, E. Seuil, Ed., Paris, 1966.

[199] P. J. Schäfer, F. I. Corona-Strauss, R. Hannemann, S. A. Hillyard, and D. J. Strauss, "Testing the Limits of the Stimulus Reconstruction Approach: Auditory Attention Decoding in a Four-Speaker Free Field Environment," *Trends in Hearing*, vol. 22, p. 233121651881660, jan 2018. [Online]. Available: http://journals.sagepub.com/doi/10.1177/2331216518816600

[200] S. A. Shamma, M. Elhilali, and C. Micheyl, "Temporal coherence and attention in auditory scene analysis," *Trends in Neurosciences*, vol. 34, no. 3, pp. 114–123, mar 2011. [Online]. Available: http://dx.doi.org/10.1016/j.tins.2010.11.002https://linkinghub.elsevier.com/retrieve/pii/S0166223610001670

[201] S. A. Shamma and C. Micheyl, "Behind the scenes of auditory perception," *Current Opinion in Neurobiology*, vol. 20, no. 3, pp. 361–366, jun 2010. [Online]. Available: http://dx.doi.org/10.1016/j.conb.2010.03.009https://linkinghub.elsevier.com/retrieve/pii/S0959438810000474

[202] R. V. Shannon, F.-g. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech Recognition with Primarily Temporal Cues," *Science*, vol. 270, no. 5234, pp.

303–304, oct 1995. [Online]. Available: https://www.science.org/doi/10.1126/science.270.5234.303

[203] L. V. Sharakhina, N. V. Mikhailov, K. D. Selyankina, and A. S. Semkina, "Digital Technologies in Development of Modern Music Industry," in *2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*. IEEE, jan 2020, pp. 71–76. [Online]. Available: https://ieeexplore.ieee.org/document/9039328/

[204] B. G. Shinn-Cunningham, "Object-based auditory and visual attention," *Trends in Cognitive Sciences*, vol. 12, no. 5, pp. 182–186, may 2008. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364661308000600

[205] B. G. Shinn-Cunningham and V. Best, "Selective Attention in Normal and Impaired Hearing," *Trends in Amplification*, vol. 12, no. 4, pp. 283–299, dec 2008. [Online]. Available: http://journals.sagepub.com/doi/10.1177/1084713808325306

[206] A. Simon, S. Bech, G. Loquet, and J. Ostergaard, "Electrodes selection for cortical auditory attention decoding with EEG during speech and music listening," in *2022 25th International Conference on Information Fusion (FUSION)*, IEEE. IEEE, jul 2022, pp. 01–06. [Online]. Available: https://ieeexplore.ieee.org/document/9841365/

[207] P. G. Singh and A. S. Bregman, "The influence of different timbre attributes on the perceptual segregation of complex-tone sequences," *The Journal of the Acoustical Society of America*, vol. 102, no. 4, pp. 1943–1952, oct 1997. [Online]. Available: https://pubs.aip.org/asa/jasa/article/102/4/1943-1952/562467

[208] J. Sloboda, "Everyday Uses of Music Listening: A Preliminary Study," in *Exploring the Musical MindCognition, emotion, ability, function*. Oxford University Press, dec 2004, vol. 354, no. February, pp. 318–331. [Online]. Available: https://academic.oup.com/book/27084/chapter/196429437

[209] J. Sloboda, S. O'Neill, and A. Ivaldi, "Functions of Music in Everyday Life: An Exploratory Study Using the Experience Sampling Method," *Musicae Scientiae*, vol. 5, no. 1, pp. 9–32, mar 2001. [Online]. Available: http://journals.sagepub.com/doi/10.1177/102986490100500102

[210] J. S. Snyder, C. Alain, and T. W. Picton, "Effects of Attention on Neuroelectric Correlates of Auditory Stream Segregation," *Journal of Cognitive Neuroscience*, vol. 18, no. 1, pp. 1–13, jan 2006. [Online]. Available: https://direct.mit.edu/jocn/article/18/1/1/4095/Effects-of-Attention-on-Neuroelectric-Correlates

[211] C. Spence and V. Santangelo, *Auditory attention*. Oxford University Press, jan 2010, no. March. [Online]. Available: https://academic.oup.com/edited-volume/28094/chapter/212182476

[212] W. Spieth, J. F. Curtis, and J. C. Webster, "Responding to One of Two Simultaneous Messages," *The Journal of the Acoustical Society of America*, vol. 26, no. 3, pp. 391–396, may 1954. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.1907347https://pubs.aip.org/asa/jasa/article/26/3/391-396/744529

[213] G. B. Stanley, F. F. Li, and Y. Dan, "Reconstruction of Natural Scenes from Ensemble Responses in the Lateral Geniculate Nucleus," *The Journal of*

Neuroscience, vol. 19, no. 18, pp. 8036–8042, sep 1999. [Online]. Available: https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.19-18-08036.1999

[214] E. Stemasov, G. Haas, M. Rietzler, and E. Rukzio, "Augmenting Human Hearing Through Interactive Auditory Mediated Reality," in *Adjunct Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. New York, NY, USA: ACM, oct 2018, pp. 69–71. [Online]. Available: https://dl.acm.org/doi/10.1145/3266037.3266104

[215] I. Sturm, M. Treder, D. Miklody, H. Purwins, S. Dähne, B. Blankertz, and G. Curio, "Extracting the neural representation of tone onsets for separate voices of ensemble music using multivariate EEG analysis." *Psychomusicology: Music, Mind, and Brain*, vol. 25, no. 4, pp. 366–379, dec 2015. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/pmu0000104

[216] E. Sussman, "Auditory Scene Analysis: An Attention Perspective," *Journal of Speech, Language, and Hearing Research*, vol. 60, no. 10, pp. 2989–3000, oct 2017. [Online]. Available: http://pubs.asha.org/doi/10.1044/2017{_}JSLHR-H-17-0041

[217] E. Sussman, W. Ritter, and H. G. Vaughan, "Attention affects the organization of auditory input associated with the mismatch negativity system," *Brain Research*, vol. 789, no. 1, pp. 130–138, apr 1998. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0006899397014431

[218] E. Sussman, I. Winkler, M. Huotilainen, W. Ritter, and R. Näätänen, "Top-down effects can modify the initially stimulus-driven auditory organization," *Cognitive Brain Research*, vol. 13, no. 3, pp. 393–405, may 2002. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0926641001001318

[219] E. S. Sussman, "Integration and segregation in auditory scene analysis," *The Journal of the Acoustical Society of America*, vol. 117, no. 3, pp. 1285–1298, mar 2005. [Online]. Available: https://pubs.aip.org/asa/jasa/article/117/3/1285-1298/543894

[220] T. Taillez, B. Kollmeier, and B. T. Meyer, "Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech," *European Journal of Neuroscience*, vol. 51, no. 5, pp. 1234–1241, mar 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/ejn.13790

[221] E. S. Teoh, F. Ahmed, and E. C. Lalor, "Attention Differentially Affects Acoustic and Phonetic Feature Encoding in a Multispeaker Environment," *The Journal of Neuroscience*, vol. 42, no. 4, pp. 682–691, jan 2022. [Online]. Available: https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.1455-20.2021

[222] M. Tervaniemi, A. Kujala, K. Alho, J. Virtanen, R. Ilmoniemi, and R. Näätänen, "Functional Specialization of the Human Auditory Cortex in Processing Phonetic and Musical Sounds: A Magnetoencephalographic (MEG) Study," *NeuroImage*, vol. 9, no. 3, pp. 330–336, mar 1999. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1053811999904056

[223] M. Tervaniemi and K. Hugdahl, "Lateralization of auditory-cortex functions," *Brain Research Reviews*, vol. 43, no. 3, pp. 231–246, dec 2003. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0165017303002236

[224] M. Tervaniemi, S. V. Medvedev, K. Alho, S. V. Pakhomov, M. S. Roudas, T. L. van Zuijen, and R. Näätänen, "Lateralized automatic auditory processing of phonetic versus musical information: A PET study," *Human Brain Mapping*, vol. 10, no. 2, pp. 74–79, jun 2000. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-0193(200006)10:2{%}3C74::AID-HBM30{%}3E3.0.CO;2-2

[225] F. E. Theunissen and J. E. Elie, "Neural processing of natural sounds," *Nature Reviews Neuroscience*, vol. 15, no. 6, pp. 355–366, jun 2014. [Online]. Available: http://www.nature.com/articles/nrn3731

[226] S. K. Thompson, R. P. Carlyon, and R. Cusack, "An objective measurement of the build-up of auditory streaming and of its modulation by attention." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 37, no. 4, pp. 1253–1262, 2011. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/a0021925

[227] M. Thornton, D. Mandic, and T. Reichenbach, "Robust decoding of the speech envelope from EEG recordings through deep neural networks," *Journal of Neural Engineering*, vol. 19, no. 4, p. 046007, aug 2022. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2552/ac7976

[228] M. S. Treder, H. Purwins, D. Miklody, I. Sturm, and B. Blankertz, "Decoding auditory attention to instruments in polyphonic music using single-trial EEG classification," *Journal of Neural Engineering*, vol. 11, no. 2, p. 026009, apr 2014. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2560/11/2/026009

[229] A. Treisman, "Contextual Cues in Selective Listening," *Quarterly Journal of Experimental Psychology*, vol. 12, no. 4, pp. 242–248, 1960.

[230] ——, "Monitoring and storage of irrelevant messages in selective attention," *Journal of Verbal Learning and Verbal Behavior*, vol. 3, no. 6, pp. 449–459, dec 1964. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0022537164800153

[231] B. Truax, *Acoustic communication*, 1984.

[232] C. Tzortzis, M.-C. Goldblum, M. Dang, F. Forette, and F. Boller, "Absence of Amusia and Preserved Naming of Musical Instruments in an Aphasic Composer," *Cortex*, vol. 36, no. 2, pp. 227–242, jan 2000. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0010945208705264

[233] B. Van Hal, S. Rhodes, B. Dunne, and R. Bossemeyer, "Low-cost EEG-based sleep detection," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, aug 2014, pp. 4571–4574. [Online]. Available: http://ieeexplore.ieee.org/document/6944641/

[234] S. Vandecappelle, L. Deckers, N. Das, A. H. Ansari, A. Bertrand, and T. Francart, "EEG-based detection of the locus of auditory attention with convolutional neural networks," *eLife*, vol. 10, pp. 1–17, apr 2021. [Online]. Available: https://elifesciences.org/articles/56481

[235] J. Vanthornhout, L. Decruy, and T. Francart, "Effect of Task and Attention on Neural Tracking of Speech," *Frontiers in Neuroscience*,

vol. 13, no. September, pp. 1–11, sep 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2019.00977/full

[236] H. Weber, *Sound Souvenirs : Audio Technologies, Memory and Cultural Practices*, K. Bijsterveld and J. Van Dijck, Eds. Amsterdam: Amsterdam University Press, 2009. [Online]. Available: http://www.oapen.org/record/340032

[237] P. Webster, "Historical Perspectives on Technology and Music," *Music Educators Journal*, vol. 89, no. 1, pp. 38–43, sep 2002. [Online]. Available: http://journals.sagepub.com/doi/10.2307/3399883

[238] H. L. Wells, L. A. McClure, B. E. Porter, and D. C. Schwebel, "Distracted Pedestrian Behavior on two Urban College Campuses," *Journal of Community Health*, vol. 43, no. 1, pp. 96–102, feb 2018. [Online]. Available: http://link.springer.com/10.1007/s10900-017-0392-x

[239] D. L. Wessel, "Timbre Space as a Musical Control Structure," *Computer Music Journal*, vol. 3, no. 2, p. 45, jun 1979. [Online]. Available: https://www.jstor.org/stable/3680283?origin=crossref

[240] I. Winkler, "Preattentive auditory context effects," *Cognitive, Affective, & Behavioral Neuroscience*, vol. 3, no. 1, pp. 57–77, 2003.

[241] I. Winkler, E. Kushnerenko, J. Horváth, R. Čeponienė, V. Fellman, M. Huotilainen, R. Näätänen, and E. Sussman, "Newborn infants can organize the auditory world," *Proceedings of the National Academy of Sciences*, vol. 100, no. 20, pp. 11 812–11 815, sep 2003. [Online]. Available: https://pnas.org/doi/full/10.1073/pnas.2031891100

[242] D. D. E. Wong, S. A. Fuglsang, J. Hjortkjær, E. Ceolini, M. Slaney, and A. de Cheveigné, "A Comparison of Regularization Methods in Forward and Backward Models for Auditory Attention Decoding," *Frontiers in Neuroscience*, vol. 12, no. AUG, pp. 1–16, aug 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2018.00531/full

[243] N. Wood and N. Cowan, "The cocktail party phenomenon revisited: How frequent are attention shifts to one's name in an irrelevant auditory channel?" *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 21, no. 1, pp. 255–260, 1995. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/0278-7393.21.1.255

[244] K. J. P. Woods and J. H. McDermott, "Schema learning for the cocktail party problem," *Proceedings of the National Academy of Sciences*, vol. 115, no. 14, pp. E3313–E3322, apr 2018. [Online]. Available: https://pnas.org/doi/full/10.1073/pnas.1801614115

[245] A. Yamadori, Y. Osumi, S. Masuhara, and M. Okubo, "Preservation of singing in Broca's aphasia." *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 40, no. 3, pp. 221–224, mar 1977. [Online]. Available: https://jnnp.bmj.com/lookup/doi/10.1136/jnnp.40.3.221

[246] R. J. Zatorre, "Hemispheric asymmetries for music and speech: Spectrotemporal modulations and top-down influences," *Frontiers in Neuroscience*, vol. 16, no. December, pp. 1–7, dec 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2022.1075511/full

[247] R. J. Zatorre, P. Belin, and V. B. Penhune, "Structure and function of auditory cortex: music and speech," *Trends in Cognitive Sciences*, vol. 6, no. 1, pp. 37–46, jan 2002. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364661300018167

[248] E. M. Zion Golumbic, N. Ding, S. Bickel, P. Lakatos, C. A. Schevon, G. M. McKhann, R. R. Goodman, R. Emerson, A. D. Mehta, J. Z. Simon, D. Poeppel, and C. E. Schroeder, "Mechanisms Underlying Selective Neuronal Tracking of Attended Speech at a "Cocktail Party"," *Neuron*, vol. 77, no. 5, pp. 980–991, mar 2013. [Online]. Available: http://dx.doi.org/10.1016/j.neuron.2012.12.037https://linkinghub.elsevier.com/retrieve/pii/S0896627313000457

[249] N. J. Zuk, J. W. Murphy, R. B. Reilly, and E. C. Lalor, "Envelope reconstruction of speech and music highlights stronger tracking of speech at low frequencies," *PLOS Computational Biology*, vol. 17, no. 9, p. e1009358, sep 2021. [Online]. Available: http://dx.doi.org/10.1371/journal.pcbi.1009358https://dx.plos.org/10.1371/journal.pcbi.1009358

Part II

Papers

Paper A

Cortical Auditory Attention Decoding During Music And Speech Listening

Adèle Simon, Gérard Loquet, Jan Østergaard and Søren Bech

The paper has been submitted to the
IEEE TRANSACTIONS ON NEURAL SYSTEMS AND REHABILITATION ENGINEERING.

The layout has been revised.

Abstract

It has been demonstrated that from cortical recordings, it is possible to detect which speaker a person is attending in a cocktail party scenario. The stimulus reconstruction approach, based on linear regression, has been shown to be useable to reconstruct an approximation of the envelopes of the sounds attended to and not attended to by a listener from the electroencephalogram data (EEG). Comparing the reconstructed envelopes with the envelopes of the stimuli, a higher correlation between the envelopes of the attended sound is observed. Most of the studies focused on speech listening and only a few studies investigated the performances and the mechanisms of auditory attention decoding during music listening. In the present study Auditory Attention Aetection (AAD) techniques, that have been proven successful for speech listening, were applied to a situation where the listener is actively listening to music, concomitant with a distracting sound. Results show that AAD can be successful for both speech and music listening, while showing differences in the reconstruction accuracy. The results of this study also highlighted the importance of the training data used in the construction of the model. This study is a first attempt to decode auditory attention from EEG data in situations where music and speech are present. The results of this study indicate that linear regression can also be used for AAD when listening to music if the model is trained for musical signals.

1 Introduction

In complex sound scenes, human beings have the ability to segregate sound streams and to focus their attention on one of the multiple sounds present [1]. A considerable amount of literature has been published on this ability, often called the cocktail party effect [2, 3]. These studies particularly focus on situations where multiple speech signals are presented to a listener. However, less research has explored this effect where music is also present.

The last two decades have seen a growing trend toward auditory attention decoding (AAD) from neuroimaging, as a way to understand the underlying mechanisms and as a potential application for future brain-computer interface (BCI) or neuro-steered hearing assisting devices [4, 5]. Auditory attention has been shown to induce neural responses [6–8], for example by modulating some neural frequency bands [9], or by reshaping neural events [10], both during speech listening [11] or music listening [12–14]. These variations in event-related potentials (ERP) can be used to decode auditory attention [12, 14, 15], but present several limitations. One of these limitations is the requirement of the specific onset of the auditory signal, while another challenge is the noisy nature of the neural signal. Those limitations rise the need for several repetitions of the task to extract useful ERP for hearing assistive BCI.

AAD has also been explored by looking at the mechanisms of neural entrainment. Some acoustic features of the audio heard, such as the temporal envelope of the audio signal heard by a listener, are tracked by the brain [16]. It led to new methods to analyse cortical responses due to continuous audio stimuli based on linear (or non-linear) models that estimate: either the neural response from the audio signal (encoding) [17]; or the audio signal from the cortical response (decoding) [18–20]. The decoding process, also known as the backward method or stimulus reconstruction method, has been demonstrated to be sensitive to auditory attention: when the listener focuses on one source of sound in a complex auditory environment, the cortical tracking of that attended sound is increased compared to the tracking of the unattended sounds [17, 19–21]. Several studies have demonstrated this influence of attentional factors, from magnetoencephalography [22], intracranial EEG [23] and intensively from scalp EEG [5, 20, 24–28]. Most of these studies are based on speech listening scenarios, where listeners have to solve cocktail party effects with often two competing speech streams.

The stimulus reconstruction approach has recently been successfully applied to reconstruct musical signals [29–32] but to a lesser extent than speech signals [32, 33]. AAD has also been applied to music, where the goal was to decode attention directed towards individual instruments in a multi-instrumental musical piece [4, 34, 35]. This new research focus is relevant as music is often present in natural sound scenes, either as a distracter or as a target of attention. Therefore, performance and potential specificities of the stimulus reconstruction method in a so-called "musical cocktail party scenario" could be explored in a context where multiple sounds that are speech or music are present and compete for the listeners' attention.

The present study investigated the performance of auditory attention based on a linear stimulus reconstruction method in a musical cocktail party scenario. The primary goal of this study is to test the performance of an AAD based on previously used methodology in a listening situation that includes music. In the experiment, participants listened to a target sound which was either speech or music, in the presence of a competing distracter sound, which was also either speech or music. During the listening task, the participants' cortical responses were continuously measured with high-density EEG and used to train a linear model that was then used to reconstruct the target stimuli and decode attention. The experimental strategy was designed to test the hypotheses that the temporal envelope of the target signal can be reconstructed regardless of whether it is speech or music, and with an accuracy above chance level; this decoding approach can be successfully used to decode attention in a musical cocktail party scenario.

2 Methods

2.1 Participants

Thirty-five participants (14 females), aged between 21 and 33 year-old (mean=26,29) took part in the experiment. No participants reported a known history of neurological disorder or hearing loss. Apart of the three participants who were native English speakers, all the others had working experience or followed education in English. The participants were compensated for their participation and written informed consent was obtained from all the participants. After recording, two participants were excluded from the data due to poor data quality in the raw data, and thus not used for further analysis (due to the large number of artefacts that contaminate every trial).

2.2 Procedure

Each participant undertook 32 trials of one minute each. In each trial they were presented with two different sound streams coming from separate loudspeakers. The loudspeakers were separated in space in front of the listener (+/- 30° azimuth). For each trial, participants were asked to pay attention to one of the sounds (the target), while ignoring the other sound (the distracter). The target, as well as the distracter, could be either speech or music (see figure A.1).

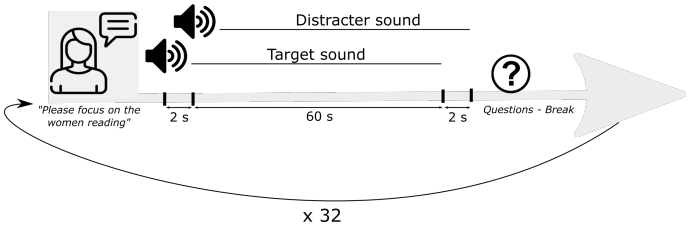

Fig. A.1: A trial started with a visual cue that indicates which sound is the target. Right after, the two sound streams start with a 2s offset, to help the participant to focus on the target. Participants listen to the two concurring sounds for 60s. At the end of each trial, participants have to answer two questions.

Before starting the task, the participants did a training trial consisting of a trial similar to the real one with stimuli that were not later reused in the study. They had the opportunity to repeat this training as many times as they wanted and to ask questions about the task before starting. At the end of each trial the participants answered two questions related to their attention level and the quality of their listening experience. Both questions were rated on a

continuous scale with endpoint labels offset 1.5 cm after the start and before the end of a 15 cm long scale.

- *"How difficult was it to focus on the target stimuli?"*- Endpoint labels: *Easy* and *Difficult*
- *"How would you describe your listening experience?"*- Endpoint labels *Bad* and *Excellent*

The subjects could take a break between trials. The participants were instructed to keep their gaze fixed on a cross in the middle of the screen for the entire duration of each trial and asked to minimize body movements and blinking.

2.3 Stimuli

Four categories of stimuli were used, divided into two types (music and speech), with each type separated into two genres.

- Piano Music: 8 excerpts of mono instrumental pieces played on a piano
- Electronic music: 8 excerpts of polyphonic pieces of instrumental electronic music
- Speech female: 8 excerpts of an audiobook read by a woman in English
- Speech male: 8 excerpts of an audiobook read by a man in English

Each excerpt was one minute long, and the participants actively listened to the target throughout the whole minute. Participants listened to the same type of target for a full block (e.g. first a block of 8 trials of Piano Music, then a block of 8 trials of Speech female). The order of the block was randomized across participants.

In the same trial, the target and the distracter could both be music, speech, or one of each type. Each excerpt was used only once as a target. Distracters were so that a balanced number of trials across conditions was obtained. For each trial, the distracter was randomly drawn from the pool of the relevant genre. In the case where both the target and distracter were music in a trial, the two excerpts could not belong to the same musical genre (e.g., target = piano music & distracter = piano music).

2.4 EEG data acquisition and pre-processing

The experiment was carried out in a single session for each participant. Continuous EEG data were recorded at 512 Hz using a 64-channel g.HIamp-Research system (g.tec Medical Engineering GmbH, Austria). The electrodes

were placed on the scalp according to the 10-20 international system. The impedance of each electrode was maintained at lower than 5kOhms.

After data collection, pre-processing of the data was carried out using EEGLAB v2021.1 [36]. The EEG data were referenced to the average of all scalp electrodes. The EEG channels contaminated by noise were visually inspected and interpolated from neighbouring electrodes. Independent Component Analysis (ICA) was run from EEGLAB and the automated detection plugin [37] allowed to remove the artefacts related to eye blinks or eye movements. The EEG data were bandpass filtered between 1 and 8 Hz and downsampled to a sampling rate of 64 Hz. The choice of cutoff frequency was based on previous studies on cortical stimulus reconstruction done on speech signal [20]. The influence of cutoff frequency was also tested on the present dataset: the results obtained support the importance of the 1-8 Hz frequency range for both speech reconstruction and music reconstruction (See Supplementary Material). The trials where the artefacts were too significant were discarded (e.g., movements). The discarded data correspond to 7,68% of the total data.

For the signals, amplitude envelopes from both target and distracter were extracted using a Hilbert transform and then downsampled to the same sampling rate of 64 Hz. Examples of the envelopes for speech and music can be seen in Figure A.2. The shape of envelopes for speech and music differ due to the nature of the signal: for speech signal, due to the pause between words, the envelope tends to drop to zero and show a greater depth of modulation, compared to music envelopes.

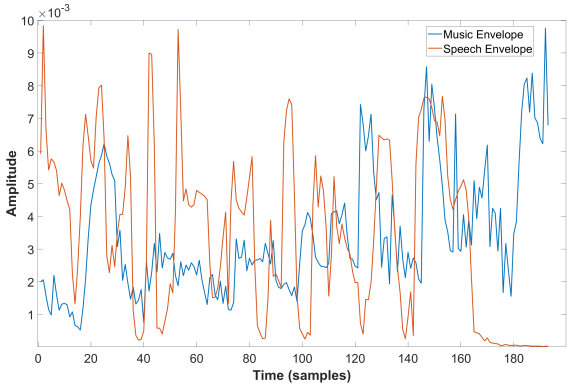

Fig. A.2: Example of 3 seconds of an envelope for speech and music.

2.5 Stimulus reconstruction and attention decoding

The decoding of auditory attention from the EEG signal was done with a conventional stimulus reconstruction method [17, 19, 20]. With this method, the EEG signal is used to reconstruct an estimate of the input stimulus through a linear reconstruction multi delay model. This model maps the cortical activity measured with the EEG to the stimulus envelope, as follows:

$$s'(t) = \sum_{n} \sum_{\tau} g(\tau, n) R(t - \tau, n) \tag{A.1}$$

where s' is the reconstructed envelope, $R(t - \tau, n)$ is the EEG response at time $(t - \tau)$ for electrode n, and g is the linear model, which is a function of electrode n and time lags τ. The time lags τ cover the interval from 0 ms to 500 ms post-stimulus, in order to take into account time lags that have been shown to influence AAD for both speech [20] and Music [32]. The model g can be estimated by minimizing the mean squared error between the original and the reconstructed envelopes, which can be solved analytically using ridge regularization methods [21]:

$$g = (R^T R + I\lambda)^{-1} R^T S \tag{A.2}$$

where I is the identity matrix, and λ is the regularization parameter used to prevent overfitting [19, 21]. The hyperparameter λ was estimated through a cross-validation approach, as described in [17]. This test was run for each separated subset of data (target = speech female, target = speech male, target = music electronic, and target = music piano), in order to assure that the regularization factor is optimized for each stimulus type. For those 4 categories, the optimal regularization factor that produced the highest reconstruction accuracy was similar at $\lambda = 10^5$.

The reconstruction accuracy is measured by calculating the Pearson's r, the correlation coefficient, between the original target envelope and the reconstructed one (r_{target}). The correlation between the reconstructed envelope and the envelope extracted from the distracter was also calculated ($r_{distracter}$). The correlation is calculated with an entire trial, corresponding to 60 seconds of the reconstructed envelope and 60 seconds of the original envelope.

For each reconstruction, the attention decoding was evaluated by comparing the two correlation coefficients. A trial was considered successfully decoded if the reconstructed envelope had a greater correlation with the target envelope compared to the correlation with the distracter envelope (i.e., $r_{target} > r_{distracter}$). For the present study, the stimulus reconstruction approach was done using a custom-made analysis script, on Matlab R2021a.

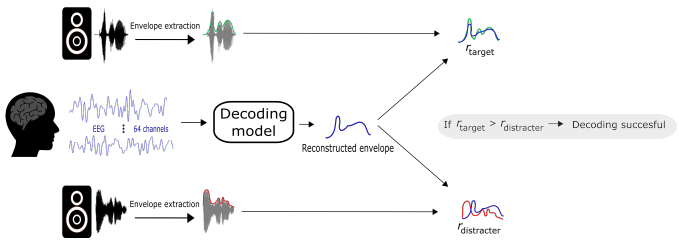

Fig. A.3: Schematic summary of auditory attention decoder

2.6 Model training

For each trial, a leave-one-out cross-validation method was used to train the models. Each trial was decoded using a model obtained by averaging the parameter of the models trained on all other trials. Through this experiment, several types of models were used, all created with different sets of training data:

- Trained on all: All trials, both when the target is speech and music, minus the one under test, are used to calculate the models.
- Trained on the same type and the same genre: Congruent model, where all trials where the target is of the same type and the same genre as the one under test (e.g., Piano music), minus the one under test, are used to calculate the models.
- Trained on the same type: All trials where the target is of the same type as the one under test (either music or speech), minus the one under test, are used to calculate the models.
- Trained on opposite type: All trials where the target is not of the same type as the one under test are used to calculate the models.

3 Results

Two measures were used to assess the performance of the auditory attention detection. The first one is the success rate of the attention detection, which corresponds to the percentage of trials that were successfully decoded. To that aim, the correlation between the reconstructed envelope is compared to either the target's envelope (r_{target}) or the distracter's envelope ($r_{distracter}$). A trial is successfully decoded when the correlation of the reconstructed envelope with the target is greater than with the distracter ($r_{target} > r_{distracter}$). This success

rate can indicate if the model allows decoding auditory attention better than chance. Chance level is calculated by taking the mean and the confidence interval of a binomial distribution with a success chance of 50%, corresponding to a random binary decision.

Following that, the reconstruction accuracy was also investigated, which corresponds to Pearson's correlation coefficients between the reconstructed envelope and the target envelope. The goal is to investigate if the linear model can reconstruct the target envelope better than chance, and then explore potential differences between the reconstruction of musical envelopes compared to the reconstruction of speech envelopes.

To establish the chance level all conditions were compared with a "random reconstruction accuracy". The random reconstruction accuracy was calculated with a reconstructed envelope and an unrelated original envelope: e.g. the envelope of the target used for trial 1 of Subject 1, where the target was piano music, was correlated with the envelope reconstructed from trial 14 from Subject 6, where the target was female speech. The pairing between the original and reconstructed envelopes was randomized. Following that, permutation tests were used to compare the reconstruction accuracy for each condition to the random reconstruction accuracy, with 10 000 permutations. For each condition, sample sizes used for the calculation of actual accuracy and random accuracy were equal.

3.1 Congruent Model

The models were first calculated with training data picked only from trials where the target was similar to the trial under test, i.e. of the same type and same genre. By using condition-specific reconstruction filters, the assumption is that the models were not influenced by other listening conditions and would be fitted to each specific trial. The highest decoding success rate is obtained when the target of attention is speech, either when the distracter is music (Success rate = 81.43%) or when the distracter is speech (Sucess rate = 73.52%). When the listener is actively listening to music, the success rate of the AAD is a bit lower, at 70.76% when the distracter is also music, or 67.19% when the distracter is speech.

Success Rate

As shown in Figure A.4, when using a congruently trained model, auditory attention can be successfully decoded, above chance level (chance level interval = 43.89% to 56.11%), for all listening conditions.

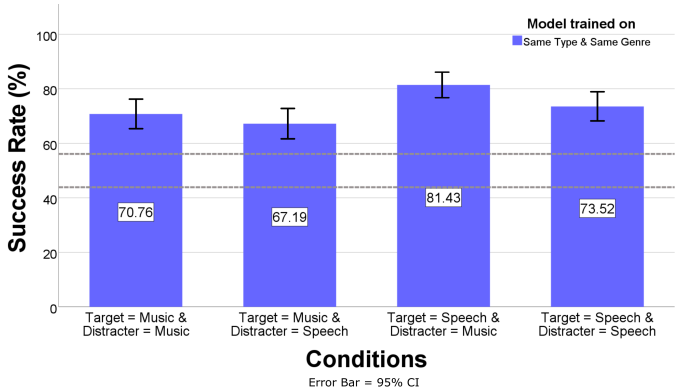

Fig. A.4: Success rates across conditions obtained with models trained on the same type and same genre as the target - Chance level is indicated by dashed gray lines

Reconstruction accuracy

For all conditions, reconstruction accuracy was significantly higher than the random reconstruction accuracy ($p < 0.0001$), suggesting that for all listening conditions, stimulus reconstructions are feasible above chance level.

To explore differences between reconstruction accuracy across conditions, Anova based on general linear model was performed to explore the main effect of fixed (listening conditions). Participant was included as a random factor.

When comparing across conditions, the ANOVA shows a significant effect of listening conditions ($F(3, 998) = 25.980, p < 0.001$), with significant differences, calculated by posthoc comparisons with Bonferroni corrections, between both of the "target music" conditions and the "target speech" conditions ($p < 0.001$). Significant differences were also found between the two "target speech" conditions ($p = 0.004$). These results suggest that the stimulus reconstruction approach can better reconstruct the target stimulus when the listener is listening to speech compared to situations where the listener is listening to music, especially when the distracter is a musical sound (see figure A.6-A).

3.2 Other training conditions

In a second analysis, the training of the model was also considered to evaluate the success rate and the reconstruction accuracy when using models trained on various types of sounds. The goal of this analysis is to explore the influence of the training data on the performance of the model.

Success Rate

Figure A.5 shows the success rate of auditory attention decoding when using models that are differently trained. For the "target speech" conditions, the success rate appears to be unaffected by the training data use, apart from the "trained on opposite type" condition. In that condition, the auditory attention model performs around the chance level, suggesting that in such a case, the auditory attention cannot be successfully detected. For the other conditions when the target of attention is music, the training of the model influences the performances.

For the "target music & distracter music" condition, only the most congruent condition (trained on the same type and same genre) allows for successful decoding, while all the other training conditions perform around the chance level. For the "target music & distracter speech" condition, while both congruent training conditions perform above chance level (trained on the same type and same genre and trained on the same type), for the two other conditions the success rate is considerably low. The low success rate, below chance level, suggests that in these cases, the model tends to reconstruct the distracter better than the target. The poor success rate observed, in this condition suggests that a general decoder, trained on both speech and music tends to be biased toward speech reconstruction.

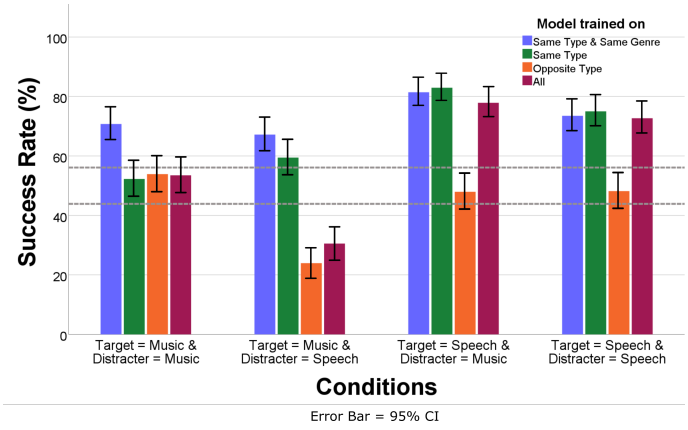

Fig. A.5: Success rates across conditions obtained with differently trained models - Chance level is indicated by dashed gray lines.

Reconstruction accuracy

As for the congruent models, reconstruction accuracy was significantly higher than the random reconstruction accuracy ($p < 0.0001$), for all conditions, and

for both target reconstruction accuracy and distracter reconstruction accuracy. It suggests that for all listening conditions, stimulus reconstructions are feasible above chance level, either with the target or with the distracter.

A three-way ANOVA based on general linear model was calculated with both listening condition and model training as fixed factors and participants as a random factor. Main effects and interactions between fixed factors (Listening Condition x Model Training) were explored, and residuals were checked to be normally distributed. Results showed significant effects of both Listening conditions ($F(3, 4088) = 118.293, p < 0.001$) and Training conditions ($F(3, 4088) = 96.642, p < 0.001$). A significant interaction between the two factors was also found ($F(9, 4088) = 7.065, p < 0.001$). (See figure A.6-A for results) These results suggest that when the target of attention is speech, the reconstruction is less precise when using a model trained on different types of signals (here trained only on music). For all other training conditions, reconstruction performances are equivalent. When music is the target of attention, the training of the model influences the results. Training on an incongruent model, with only trials from a different type, decrease the reconstruction performance, for speech. Contrary to the "target speech" conditions, here using a model trained on the same musical genre increase the reconstruction performance compared to a generic musical model ("trained on the same type"), or a generic model ("trained on all").

Figure A.6-B shows the correlations obtained when calculating the correlation between the reconstructed envelope and the distracter envelope. First, it can be observed that the results obtained are above chance level, for all conditions. That suggests that it is also possible to reconstruct sounds that were heard, but not actively focused on. However, for most situations, this reconstruction is lower, as the correlation with the original envelope is smaller, indicating that the distracter sound can also be reconstructed, but to a lesser extent than the target.

The situation is different when the target is music and the distracter is speech. In that case, when the model is trained only on musical signal ("trained on the same type" or "trained on the same type and same genre"), the reconstruction accuracy for the distracter is lower than the reconstruction accuracy with the target, and follow a similar trend compared to the other listening conditions. When the model is trained on speech signal ("trained on opposite signal" or "trained on all"), the reconstruction accuracy of the distracter is greater. It is still lower than the target reconstruction accuracies obtained with speech as a target, but on average greater than the target reconstruction accuracy obtained with music as a target. This suggests that when trained with speech signals, the model may be biased towards speech reconstruction, which can also explain the poor success rate obtained for the conditions where music is a target, speech is a distracter, and the model is trained on speech.

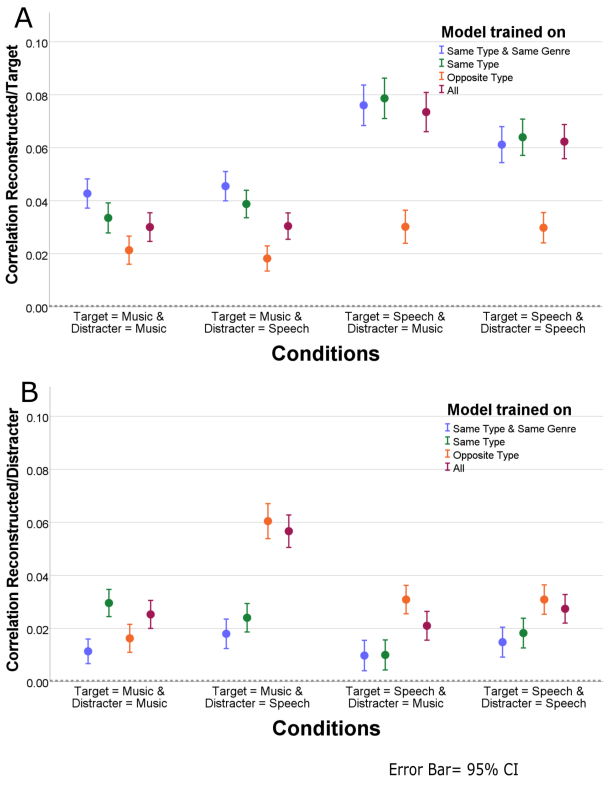

Fig. A.6: Reconstruction accuracies across conditions obtained with differently trained models, between A- reconstructed and target stimuli ; B- reconstructed and distracter stimuli - Chance level is indicated by dashed gray lines.

3.3 Effect of size of training set

In the aforementioned analysis, the size of the training set differed: as the general model used all available data, the training dataset is larger than for the model trained on only a subset of data (e.g. congruent models trained on one specific type and genre). This approach was chosen to optimize training by using as much data as possible. However, this difference in the size of the training sets may influence the conclusions of the current study. To control this factor, the success rates for AAD were recalculated with models trained on smaller training sets, to ensure that all training subsets were of equal size (i.e. the models were trained on 252 trials of one minute). For each training condition, the training subsets were randomly selected from the available training data. During the selection, the type of data was controlled to ensure a bal-

anced distribution of conditions in the smaller training set (e.g. for the subset selection for a model trained on both speech and music, the number of trials where Target = Music is equal to the number of trials where Target = Speech). An exact McNemar's test was used to test if the success rates obtained with equally-sized training sets differ from the success rates from unequally-sized training sets. The test determined that there were no significant differences between the two conditions (p=.275).

3.4 Subjective ratings

Two-way mixed model ANOVAs were conducted to examine the effect of condition and participants (included as a random factor), as well as the interactions, on the subjective ratings of the participants, attention and quality of listening experience (QoLE), residuals were checked to be normally distributed.

For attention, significant effects were found for the condition factor $F(3, 998) = 20.081, p < 0.001$. For QoLE, significant effects were found for the condition factor $F(3, 998) = 48.618, p < 0.001$. This suggests that while there are differences between the conditions in term of difficulty to focus on the target stimuli and quality of listening experience, it also varies across individuals. Results can be seen in figure A.7.

In order to explore a potential link between cortical reconstruction accuracy and subjective ratings, Pearson's correlations have been run between the reconstruction accuracy values and the subjective rating. For the attention ratings, a small but significant correlation has been found with the reconstruction accuracy ($r(5168) = 0.052, p < 0.001$). Similarly, the ratings of QoLE ratings are slightly correlated with the cortical reconstruction accuracy ($r(5168) = 0.042, p = 0.003$).

4 Discussion

This study attempts to decode auditory attention from continuous cortical responses, measured with EEG, in a musical cocktail party scenario. Participants were presented with two streams of sounds simultaneously, which could be either speech or music, and asked to focus on one of the sounds (target) while ignoring the other (distracter). A linear regression method that maps the cortical data to the audio signal was used to reconstruct the input stimuli and the reconstruction was used to decode the attention of the listener.

For each trial, the attention decoding was done by comparing the reconstructed envelope with both the target envelope and the distracter envelope. As the stimulus reconstruction approach has been shown to be sensitive to selective auditory attention, it was hypothesised that the reconstructed envelope should correlate better with the envelope of the target stimulus compared to

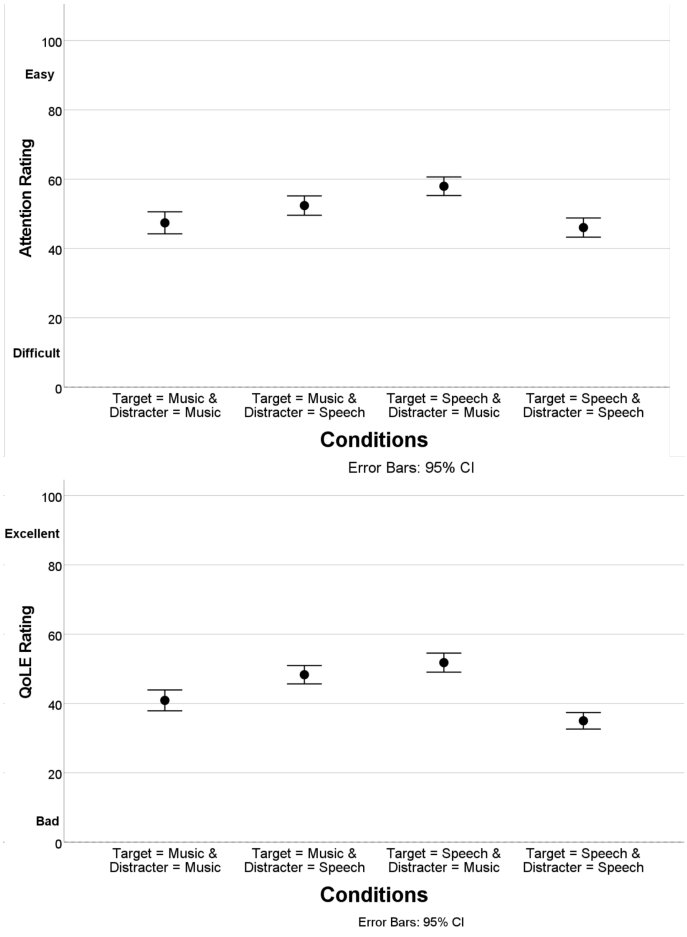

Fig. A.7: Subjective ratings across conditions, mean and 95% CI across participants.

the correlation with the distracter stimulus, irrespective of the type of target of distracter (music or speech).

When tested on a congruently trained model (i.e., trained and tested on trials with the target of the same type and genre), auditory attention can be successfully decoded for all listening conditions, above chance level. The type of data used for the training of the model also impacted the success rate of the decoding. When the target of attention was speech, all training conditions lead to successful decoding, with an equivalent success rate, except for one condition. If the model was trained only on musical trials and tested on speech, the success rate dropped around the chance level. In this condition the decoding was unsuccessful.

When the target of attention was music, results also vary across conditions depending on the type of distracter. If the target was music, and the distracter was also music, only the congruent training (same type & same genre), leads to successful decoding. All the other training conditions perform around chance level. However, when the target was music and the distracter was speech, both congruently trained models (model trained on the same type & same genre, as well as model same type) perform above chance level. However, when the training set also includes speech data, the success rate dropped considerably. For the "trained on all" and "trained on opposite type" (i.e., here trained on speech), the success rate was below chance level (23.95 – 30.53%), which suggests that the model reconstructs better the distracter than the target. This finding was unexpected and suggests that, in these conditions, the model is more influenced by aspects specific to speech than by aspects related to auditory attention.

For all listening conditions, the values of reconstruction accuracies obtained are better than chance. It suggests that, even when multiple sounds are present in a sound scene, the linear regression approach can be used to reconstruct the stimulus that a listener is attentive to, both when this stimulus is speech and music. These results, especially the order of magnitude of the reconstruction accuracy obtained through this study, are consistent with previous work on auditory attention decoding using linear regression during a cocktail party scenario with only speech [20, 26, 28], or previous work using this reconstruction approach during music listening [29, 34, 35]. Reconstruction accuracies are overall greater when the target of attention was speech compared to the trials where the target of attention is music. This difference in tracking accuracies matched those observed by [33]. In addition, the results showed that the training of the model can influence the reconstruction accuracy. An incongruently trained model (e.g., a model trained on music and tested on speech), significantly reduces the reconstruction accuracies. When the target of attention is music, the reconstruction accuracy seems to be more sensitive to the training of the model: The reconstruction accuracy is significantly higher when decoded with a model trained only on the same musical genre, compared to the other models (i.e., trained only on music trials, but including also other musical genres, or trained on both speech and music). Nevertheless, it is coherent with the results obtained for the success rate of attention decoding. This difference might be greater than the difference due to the auditory attention, leading to a better reconstruction of speech than music, even if the attention was directed to speech.

The difficulty of the task and the difficulty of attending to the target stimulus might influence the decoding performance, as the attention of the listener might not be perfectly on the target during a challenging trial. However, the correlation between subjective ratings of attention and reconstruction accuracy has been found to be small. This difficulty in attending to the target might

explain the small decrease in performance for speech listening in presence of speech compared to speech listening in presence of music. It is, however, different for the situation where the target of attention is music and the distracter speech. While being rated easier than the speech on speech or music on music situation, the decoding performances are worse. More research would be needed to explore the relationship between cortical reconstruction accuracy, the difficulty of the task and listening effort. The fact that participants were not native English speakers may also influence the neural response (and thus the reconstruction accuracy), or modulate how they attend to the speech signal [38].

The choice of the bandpass filter (1 to 8 Hz) might also have influenced the reconstruction accuracy, as it has been shown that the cortical tracking of sound differs between speech and music listening at lower frequencies [33]. This study aimed to test an AAD method, that has previously proven successful on speech listening, on a music listening task. However, that means that the method and the parameters have been optimized for speech, and may be suboptimal for music. Further research should be undertaken to investigate how the performances are influenced by some of the signal or model parameters such as filtering of the EEG data, and choice of audio features used in the AAD. In order to increase reconstruction and attention decoding performances for music listening, further work could also explore stimulus reconstruction based on other audio features than the envelope, that may be better tracked by the brain during music listening such as mel spectrogram [34] or notes onset timing [29]. For music listening, it also has been suggested that spectral modulation plays a greater role than temporal modulations [39]. The open questions that arose after this study are: to what extent the parameters used for the models (such as the audio features) maximize the reconstruction for speech compared to music; and if there are other parameters that can be more suited for music reconstruction.

The differences between speech and music conditions, both in reconstruction accuracy and success rate could also be due to separate cortical processes for speech or music listening, involving different parts of the brain [40]. In the context of stimulus reconstruction, when trying to maximize reconstruction accuracy, differences were found between speech and music listening for the optimal latencies [41] and the selection of the electrode [42]. These results suggest temporal and topographic variations that could indicate the presence of differences in cortical processes activated during speech or music listening. Additional studies would be needed to investigate further the temporal and topographic variations during cortical music and speech processing.

Furthermore, the reconstruction accuracy results might be due to an enhanced cortical tracking for speech compared to music, which would be in line with the recent findings by [33]. Differences could be related to some brain processing specific to speech, or they could be due to some acoustical

aspects specific to speech signals. Another potential explanation, as suggested by [33], is that these differences may be linked to some speech-specific features, such as phoneme processing [43] or phonotactic probabilities [44] or semantics aspects [45] that could increase the cortical tracking of speech and thus lead to greater reconstruction accuracy. Aspects related to the signal itself might influence the differences obtained for signal reconstruction. The shape of the envelopes differs between speech and music. For speech, due to the pause between words, the envelopes go down to zero, followed by sharp jumps to high amplitude a the beginning of a new word. For the music envelope, the envelopes are more "flat", resulting in a lower dynamics range (see Figure A.2). Further studies, which take these variables into account, would be needed to get a better understanding of cortical auditory tracking, and variations between speech and music, in order to untangle how different aspects of speech influence the reconstruction accuracy.

Independent of the underlying factors influencing the reconstruction accuracy, the present results should be considered when designing a versatile AAD, especially in cases where both speech and music could be the target of attention. Due to the difference in reconstruction accuracy observed between speech and music reconstruction, the current decision criteria (i.e., comparing the reconstruction accuracy for both target and distracter) may not be suitable as it does not take into account these differences. Other approaches, for instance, using thresholds to classify between target or distracter, might be more appropriate (e.g., the reconstruction accuracy should be above a music threshold to be considered as a target of attention if there is music in the sound scene). Thresholds could also be used to correct the bias toward speech found in the general model (i.e., trained on both speech and music) to ensure that the apriori probability for a trial to be classified as speech or music is equal. While this approach would be interesting to develop AAD, it is outside the scope of the present study to determine such thresholds: more data and more diverse situations would be needed to explore and determine relevant thresholds. In addition, for AAD implementation, using threshold would require knowledge about the sound scene when using the AAD, such as an acoustic scene classification to inform about the presence of speech or music in the sound scene to decode.

Overall, this study shows that auditory attention decoding is feasible for musical cocktail party scenarios, both during active speech listening and active music listening. However, for music listening the decoding model needs to be fitted to the target stimulus. This is a limitation for the potential applications of such technologies, as it would be necessary to know the type of stimuli present in the sound scene to apply the right model. Future work could further explore the differences between speech listening and music listening for auditory attention decoding, to gain knowledge about the underlying mechanisms for both music listening and speech listening and attempt to improve

the performance in AAD during music listening.

References

[1] E. Sussman, "Auditory Scene Analysis: An Attention Perspective," *Journal of Speech, Language, and Hearing Research*, vol. 60, no. 10, pp. 2989–3000, oct 2017. [Online]. Available: http://pubs.asha.org/doi/10.1044/2017{_}JSLHR-H-17-0041

[2] C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[3] J. H. McDermott, "The cocktail party problem," *Current Biology*, vol. 19, no. 22, pp. R1024–R1027, dec 2009. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0960982209016807

[4] W. W. An, B. Shinn-Cunningham, H. Gamper, D. Emmanouilidou, D. Johnston, M. Jalobeanu, E. Cutrell, A. Wilson, K.-J. Chiang, and I. Tashev, "Decoding Music Attention from "EEG Headphones": A User-Friendly Auditory Brain-Computer Interface," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, IEEE. IEEE, jun 2021, pp. 985–989. [Online]. Available: https://ieeexplore.ieee.org/document/9414492/

[5] J. Belo, M. Clerc, and D. Schön, "EEG-Based Auditory Attention Detection and Its Possible Future Applications for Passive BCI," *Frontiers in Computer Science*, vol. 3, p. 661178, apr 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcomp.2021.661178/full

[6] C. L. Baldwin, "Auditory cognition and human performance: Research and applications." *CRC Press*, 2012.

[7] J. C. Hansen and S. A. Hillyard, "Endogeneous brain potentials associated with selective auditory attention." *Electroencephalogr. Clin. Neurophysiol.*, vol. 49, no. 3-4, pp. 277–290, 1980.

[8] S. Crottaz-Herbette and V. Menon, "Where and when the anterior cingulate cortex modulates attentional response: combined fMRI and ERP evidence." *J. Cogn. Neurosci.*, vol. 18, no. 5, pp. 766–780, 2006.

[9] J. R. Kerlin, A. J. Shahin, and L. M. Miller, "Attentional gain control of ongoing cortical speech representations in a "cocktail party"." *J. Neurosci.*, vol. 30, no. 2, pp. 620–628, 2010.

[10] T. Picton and S. Hillyard, "Human auditory evoked potentials. II: Effects of attention," *Electroencephalography and Clinical Neurophysiology*, vol. 36, no. C, pp. 191–200, jan 1974. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0013469474901564

[11] D. J. Lee, H. Jung, and P. Loui, "Attention modulates electrophysiological responses to simultaneous music and language syntax processing." *Brain Sci.*, vol. 9, no. 11, p. 305, 2019.

[12] I. Choi, S. Rajaram, L. A. Varghese, and B. G. Shinn-Cunningham, "Quantifying attentional modulation of auditory-evoked cortical responses from single-trial electroencephalography," *Frontiers in Human Neuroscience*, vol. 7, no. APR 2013, pp. 1–19, 2013. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fnhum.2013.00115/abstract

[13] N. J. Hill and B. Schölkopf, "An online brain–computer interface based on shifting attention to concurrent streams of auditory stimuli," *Journal of Neural Engineering*, vol. 9, no. 2, p. 026011, apr 2012. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2560/9/2/026011

[14] M. S. Treder, H. Purwins, D. Miklody, I. Sturm, and B. Blankertz, "Decoding auditory attention to instruments in polyphonic music using single-trial EEG classification," *Journal of Neural Engineering*, vol. 11, no. 2, p. 026009, apr 2014. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2560/11/2/026009

[15] P. Loui, T. Grent, D. Torpey, M. Woldorff *et al.*, "Effects of attention on the neural processing of harmonic syntax in western music." *Brain Cogn.*, vol. 25, no. 3, pp. 678–687, 2005.

[16] N. Ding and J. Z. Simon, "Cortical entrainment to continuous speech: functional roles and interpretations." *Front. Hum. Neurosci.*, vol. 8, p. 311, 2014.

[17] M. J. Crosse, G. M. Di Liberto, A. Bednar, and E. C. Lalor, "The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli," *Frontiers in Human Neuroscience*, vol. 10, no. NOV2016, pp. 1–14, nov 2016. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fnhum.2016.00604/full

[18] B. N. Pasley, S. V. David, N. Mesgarani, A. Flinker, S. A. Shamma, N. E. Crone, R. T. Knight, and E. F. Chang, "Reconstructing speech from human auditory cortex." *PLoS Biol.*, vol. 10, no. 1, 2012.

[19] E. Alickovic, T. Lunner, F. Gustafsson, and L. Ljung, "A Tutorial on Auditory Attention Identification Methods," *Frontiers in Neuroscience*,

vol. 13, no. March, pp. 1–17, mar 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2019.00153/full

[20] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG," *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, 01 2015. [Online]. Available: https://doi.org/10.1093/cercor/bht355

[21] D. D. E. Wong, S. A. Fuglsang, J. Hjortkjær, E. Ceolini, M. Slaney, and A. de Cheveigné, "A Comparison of Regularization Methods in Forward and Backward Models for Auditory Attention Decoding," *Frontiers in Neuroscience*, vol. 12, no. AUG, pp. 1–16, aug 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2018.00531/full

[22] N. Ding and J. Z. Simon, "Emergence of neural encoding of auditory objects while listening to competing speakers." *Proc. Natl. Acad. Sci. U.S.A.*, vol. 109, no. 29, pp. 11 854–11 859, jul 2012. [Online]. Available: https://pnas.org/doi/full/10.1073/pnas.1205381109

[23] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, pp. 233–236, may 2012. [Online]. Available: http://www.nature.com/articles/nature11020

[24] W. Biesmans, N. Das, T. Francart, and A. Bertrand, "Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario." *IEEE Trans. Neural Syst.*, vol. 25, no. 5, pp. 402–412, 2016.

[25] A. Aroudi, B. Mirkovic, M. De Vos, and S. Doclo, "Impact of Different Acoustic Components on EEG-Based Auditory Attention Decoding in Noisy and Reverberant Conditions," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 4, pp. 652–663, apr 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8662636/

[26] S. A. Fuglsang, T. Dau, and J. Hjortkjær, "Noise-robust cortical tracking of attended speech in real-world acoustic scenes," *NeuroImage*, vol. 156, no. April, pp. 435–444, aug 2017. [Online]. Available: http://dx.doi.org/10.1016/j.neuroimage.2017.04.026https://linkinghub.elsevier.com/retrieve/pii/S105381191730318X

[27] B. Mirkovic, S. Debener, M. Jaeger, and M. De Vos, "Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications," *Journal of Neural Engineering*, vol. 12, no. 4, p. 046007, aug 2015. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2560/12/4/046007

[28] B. Mirkovic, M. G. Bleichner, M. De Vos, and S. Debener, "Target Speaker Detection with Concealed EEG Around the Ear," *Frontiers in Neuroscience*, vol. 10, no. JUL, pp. 1–11, jul 2016. [Online]. Available: http://journal.frontiersin.org/Article/10.3389/fnins.2016.00349/abstract

[29] G. Marion, G. M. Di Liberto, and S. A. Shamma, "The music of silence: Part I: Responses to musical imagery encode melodic expectations and acoustics." *J. Neurosci.*, vol. 41, no. 35, pp. 7435–7448, 2021.

[30] G. M. Di Liberto, C. Pelofi, S. Shamma, and A. de Cheveigné, "Musical expertise enhances the cortical tracking of the acoustic envelope during naturalistic music listening," *Acoustical Science and Technology*, vol. 41, no. 1, pp. 361–364, jan 2020. [Online]. Available: https://www.jstage.jst.go.jp/article/ast/41/1/41{_}E19257/{_}article

[31] G. M. Di Liberto, G. Marion, and S. A. Shamma, "Accurate Decoding of Imagined and Heard Melodies," *Frontiers in Neuroscience*, vol. 15, no. August, pp. 1–11, aug 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2021.673401/full

[32] L. Hausfeld, L. Riecke, G. Valente, and E. Formisano, "Cortical tracking of multiple streams outside the focus of attention in naturalistic auditory scenes," *NeuroImage*, vol. 181, no. March, pp. 617–626, nov 2018. [Online]. Available: https://doi.org/10.1016/j.neuroimage.2018.07.052https://linkinghub.elsevier.com/retrieve/pii/S1053811918306670

[33] N. J. Zuk, J. W. Murphy, R. B. Reilly, and E. C. Lalor, "Envelope reconstruction of speech and music highlights stronger tracking of speech at low frequencies," *PLOS Computational Biology*, vol. 17, no. 9, p. e1009358, sep 2021. [Online]. Available: http://dx.doi.org/10.1371/journal.pcbi.1009358https://dx.plos.org/10.1371/journal.pcbi.1009358

[34] G. Cantisani, S. Essid, and G. Richard, "EEG-Based Decoding of Auditory Attention to a Target Instrument in Polyphonic Music," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, vol. 2019-Octob, no. 765068. IEEE, oct 2019, pp. 80–84. [Online]. Available: https://ieeexplore.ieee.org/document/8937219/

[35] L. Hausfeld, N. R. Disbergen, G. Valente, R. J. Zatorre, and E. Formisano, "Modulating Cortical Instrument Representations During Auditory Stream Segregation and Integration With Polyphonic Music," *Frontiers in Neuroscience*, vol. 15, no. September, pp. 1–15, sep 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2021.635937/full

[36] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis." *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, 2004.

[37] L. Pion-Tonachini, K. Kreutz-Delgado, and S. Makeig, "Iclabel: An automated electroencephalographic independent component classifier, dataset, and website." *NeuroImage*, vol. 198, pp. 181–197, 2019.

[38] R. Reetzke, G. N. Gnanateja, and B. Chandrasekaran, "Neural tracking of the speech envelope is differentially modulated by attention and language experience," *Brain and Language*, vol. 213, p. 104891, 2021.

[39] I. Wollman, P. Arias, J.-J. Aucouturier, and B. Morillon, "Neural entrainment to music is sensitive to melodic spectral complexity." *J. Neurophysiol.*, vol. 123, no. 3, pp. 1063–1071, 2020.

[40] P. Albouy, L. Benjamin, B. Morillon, and R. J. Zatorre, "Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody," *Science*, vol. 367, no. 6481, pp. 1043–1047, feb 2020. [Online]. Available: https://www.science.org/doi/10.1126/science.aaz3468

[41] A. Simon, S. Bech, J. Østergaard, and G. Loquet, "Optimal latencies for linear cortical auditory attention detection: differences between speech and music listening." in *Proceedings of International Symposium on Hearing 2022.*, 2022.

[42] A. Simon, S. Bech, G. Loquet, and J. Ostergaard, "Electrodes selection for cortical auditory attention decoding with EEG during speech and music listening," in *2022 25th International Conference on Information Fusion (FUSION)*, IEEE. IEEE, jul 2022, pp. 01–06. [Online]. Available: https://ieeexplore.ieee.org/document/9841365/

[43] G. M. Di Liberto, J. A. O'Sullivan, and E. C. Lalor, "Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing," *Current Biology*, vol. 25, no. 19, pp. 2457–2465, oct 2015. [Online]. Available: http://dx.doi.org/10.1016/j.cub.2015.08.030https://linkinghub.elsevier.com/retrieve/pii/S0960982215010015

[44] G. M. Di Liberto, D. Wong, G. A. Melnik, and A. de Cheveigné, "Low-frequency cortical responses to natural speech reflect probabilistic phonotactics," *NeuroImage*, vol. 196, no. October 2018, pp. 237–247, aug 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1053811919303234

[45] M. P. Broderick, N. J. Zuk, A. J. Anderson, and E. C. Lalor, "More than words: Neurophysiological correlates of semantic dissimilarity depend on comprehension of the speech narrative." *bioRxiv*, 2020.

Paper B

Optimal time lags for linear cortical auditory attention detection: differences between speech and music listening

Adèle Simon, Jan Østergaard, Søren Bech, and Gérard Loquet

The paper has been published in the
Proceedings of the *19th International Symposium on Hearing* 2022.

The layout has been revised.

Abstract

In recent decades, there has been a lot of interest in detecting auditory attention from brain signals. Cortical recordings have been demonstrated to be useful in determining which speaker a person is listening to a mixed variety of sounds (the cocktail party effect). Linear regression, often called the stimulus reconstruction method, shows that the envelope of the sounds heard can be reconstructed from continuous electroencephalogram recordings (EEG). The target sound, to which the listener is paying attention, can be reconstructed to a greater extent compared to other sounds present in the sound scene, which can allow attention decoding. Reconstruction can be obtained with EEG signals that are delayed compared to the audio signal, to consider the time for neural processing. It can be used to identify latencies where the reconstruction is optimal, which reflects a cortical process specific to the type of audio heard. However, most of these studies used only speech signals and did not investigate other types of auditory stimuli, such as music.

In the present study, we applied this stimulus reconstruction method to decode auditory attention in a cocktail party scenario that includes both speech and music. Participants were presented with a target sound (either speech or music) and a distracter sound (either speech or music) while continuously recording their cortical response during the listening with a 64-channels EEG system. From these recordings, we reconstructed the envelope of the stimuli, both target and distracter, by using linear ridge regression decoding models at individual time lags. Results showed different time lags for maximal reconstruction accuracies between music and speech listening, suggesting separate underlying cortical processes. Results also suggest that an attentional aspect can influence the reconstruction accuracy for middle/late time-lags.

1 Introduction

The world is composed of complex auditory scenes, where several sources of sounds coexist simultaneously, such as noise, speech, or music, and a listener can actively attend to one of the auditory streams. For instance, when sitting in a cafe, with people talking and background music, one can choose to focus on a conversation or to follow the music [1]. Separating and tracking individual sound streams from a complex sound scene is possible thanks to selective auditory attention.

An effect of selective attention is reflected in the cortical signal of the listener. Several studies recorded continuous neural response of listener presented with two or more sounds, with electroencephalogram or magnetoencephalogram: results showed that the cortical response track the attended sound stream better than an ignored sound stream [2–5]. Using this effect and an approach called stimulus reconstruction method, it has been shown that auditory attention can be decoded from continuous neural recording [4]. This approach

uses linear filters, computed using least-squares optimization, to reconstruct the sound heard by the listener from the cortical recording [6]. This stimulus reconstruction method have be shown to be sensitive to auditory attention for dichotic speech listening [4, 7, 8], and also during music listening [9–11].

When attempting to decode auditory attention with the stimulus reconstruction approach, most studies use multi-lag models to take into account cortical processing time [12]. In such a multi-lags model, the model is trained and evaluated on a combination of EEG recording at different time lags (e.g., 0 to 500ms), relative to the stimulus. While using the multi-lags model can enhance the model prediction and the performance of an auditory attention decoder, it does not allow for investigating the reconstruction performance of individual time lags, which can give information on the temporal neural processing of the sound signal [13]. A single-lags model can be used, to gain insight into the reconstruction accuracy at each time lag. This can help to compare neural processes for different types of signals, or conditions [14, 15]. This method can also be used to explore the effect of the attentional state of the listener on their cortical response: training such models to either reconstruct the target stimulus or the reconstructed stimulus can give information on the effect of attention [4]. Investigating the individual time-lags to find an optimal value that enhances stimulus reconstruction can help to gain insight into the cortical processes involved in speech and music listening. It can also provide useful information to design an auditory attention decoder, which could be fitted to either music or speech listening and enhance the performance of such an auditory attention decoder.

In the present study, we used stimulus reconstruction methods with single lags models to explore differences between cortical processing of music and cortical processing of speech. Subsequently, we compare target-trained and distracter-trained models, for both speech and music listening, to identify time lags affected by auditory attention.

2 Methods

2.1 Participants

For this study, 35 participants (14 female) were recruited, aged between 21- and 33-year-old (mean = 26,29). Participants did not report any hearing disorders or neurological disorders among the participants. Three of the participants were native English speakers, and all the others were fluent, with education or work experience in English. Written informed consent was obtained and participants were compensated for their participation in the study. Due to poor data quality, EEG recordings from two participants were excluded after recording.

2.2 Procedure and Stimuli

For each trial of one minute, the participant was exposed to two separate sound streams originating from separate loudspeakers, placed in front of her/him (+/- 30° azimuth). The direction of arrival of the target sound (left or right loudspeaker) was randomly selected for each trial The participant was instructed to pay attention to one of the sounds (target) while ignoring the other sound (distracter) the target may be either speech or music. During listening, the subject was instructed to keep their eyes fixed on a crosshair and to minimize blinks and movements. There were four categories of stimuli employed, split into two types (music and speech), with each type further subdivided into two genres.

- Piano Music: 8 excerpts of mono instrumental pieces played on a piano
- Electronic music: 8 excerpts of polyphonic pieces of instrumental electronic music
- Speech female: 8 excerpts of an audiobook read by a woman in English
- Speech male: 8 excerpts of an audiobook read by a man in English

In the same trial, the target and the distractor could have been both music, both speech, or one of each type. Each excerpt was used as a target just once. Distracters were selected to obtain a balanced number of trials across conditions (Music/Speech, Music/Music, Speech/Speech, Speech/Music). Participants completed 32 one-minute trials. For each participant the experiment was conducted in a single session.

2.3 Data collection and pre-processing

A 64-channel g.HIamp-Research system was used to record continuous EEG data at 512 Hz (g.tec Medical engineering GmbH, Austria). The electrodes were placed on the scalp in accordance with the international 10-20 system. The impedance of each electrode was kept below 5kOhms.

After data collection, pre-processing of the data was carried out using EEGLAB v2021.1 [16]. The EEG data were referenced to the average of all scalp electrodes. The noise-contaminated EEG channels were visually evaluated and interpolated from neighbouring electrodes. Independent Component Analysis (ICA) was performed in EEGLAB, and the automatic identification plugin allowed the artefacts associated with eye blinks or eye movements to be removed [17]. The envelopes of the sound signal, both target and distracter, were extracted using a Hilbert transform. Both EEG data and audio envelopes were finally bandpass filtered between 1 and 8Hz and downsampled to a 64Hz sampling rate.

2.4 Stimulus reconstruction

We used a classic stimulus reconstruction approach to decode auditory attention from the EEG data [4, 6, 13]. The EEG data is utilized to reconstruct an estimation of the input stimuli using a linear reconstruction l. This model relates EEG-measured brain activity to the stimulus envelope as follows:

$$s'(t) = \sum_{n} \sum_{\tau} g(1, n) R(t, n) \tag{B.1}$$

where s' is the reconstructed envelope, $R(t, n)$ is the EEG response at time t for electrode n, and g is the linear model, which is a function of electrode n.

The model g can be estimated by minimizing the mean squared error between the original and the reconstructed envelopes, which can be solved analytically using ridge regularization methods [18]:

$$g = (R^T R + I\lambda)^{-1} R^T S \tag{B.2}$$

where I is the identity matrix, S is the stimulus envelope, and λ is the regularization parameter used to prevent overfitting [6, 18]. The regularization factor was set to 10^5. This value was chosen by calculating several models with different values of this regularization parameter. The value that produced the highest reconstruction accuracy (measured by the Pearson's correlation coefficient between the original and the reconstructed envelope) was used for the analysis.

We calculated the Pearson's r, or correlation coefficient, between the original target envelope and the reconstructed one (r_{target}) to assess reconstruction accuracy. To obtain r_{target}, a "Target model" was trained by using EEG signals and the original envelope of the target. To assess the processes that are encountered by the distracter stimulus, a "Distractor model" was also trained, with EEG signals and the envelope of the distractor. The obtained reconstructed distractor is then compared to the original distracter to obtain the reconstruction accuracy of the distracter, $r_{distracter}$.

2.5 Single-lag model

To explore variation across time lags, several models have been trained on each individual time lag. The models were trained using the original envelope of the sound stimulus and the corresponding EEG data from the specific time lags. For example, to compute a model g_{30} for a time lag of approximately 30 ms, which corresponds to a time lag of 2 samples at the sampling rate of 64Hz (see figure B.1), we used original envelope S and time-lags EEG R as follows:

$$S = \begin{bmatrix} s(0) \\ s(1) \\ s(2) \\ \vdots \\ s(t) \end{bmatrix} \quad \text{and} \quad R = \begin{bmatrix} r_1(2) & \cdots & r_64(2) \\ r_1(3) & \cdots & r_64(3) \\ \vdots & \ddots & \vdots \\ r_1(T) & \cdots & r_64(T) \\ r_1(T+1) & \cdots & r_64(T+1) \\ r_1(T+2) & \cdots & r_64(T+2) \end{bmatrix}$$

We computed models to covert times lags ranging from 0 ms to 500 ms, at a sample rate of 64Hz. That corresponds to thirty-three individual single-lag models, separated by an interval of 15,625 ms.

All models were trained in a leave-one-out approach, which means that each trial was tested on a model created by averaging the parameters of the models trained on every other trial.

Four categories of single lags models were trained:

- Models optimized for music as a Target, where only trials where the target of attention was music are used for training and testing, and by using the Target envelope for training
- Models optimized for speech as a Target, where only trials where the target of attention was speech are used for training and testing, and by using the Target envelope for training
- Models optimized for music as a Distracter, where only trials where the distracter was music are used for training and testing, and by using the distracter envelope for training
- Models optimized for speech as a Distracter, where only trials where the distracter was speech are used for training and testing, and by using the distracter envelope for training

3 Results

3.1 Differences between speech and music listening

Figure B.2 shows the reconstruction accuracy across different time lags for both trials where the target of attention was music stimulus (music listening) and trials where the target of attention was speech stimulus. Shaded areas correspond to the 95% confidence interval for the reconstruction accuracies. The reconstruction accuracies for speech were obtained similarly, but by testing all speech-target trials on models trained on speech-target trials. In figure B.2, there is a clear difference between reconstruction accuracies for speech and music. For all time lags, the reconstruction accuracy for speech is significantly

higher than the reconstruction accuracies for music (permutation test speech vs music, $p<0.05$ for each time lag).

The second thing that stands out in figure B.2 is the pattern in variation of the reconstruction accuracies. For both speech and music, we can observe two peaks of increased reconstruction accuracy across the different time lags: a first peak at an early time lag and a second, larger peak at a later time lag. The first peak is located at a time lag comprised of between 30 and 50 ms, and the timing of this first peak is similar for both speech and music. However, the timing of the second peak varies between speech and music: ≈170 ms for speech and ≈265 ms for music. This difference in time lags between speech and music could indicate time process differences for speech and music sounds. Maximized reconstructions for speech corroborates results preciously obtained in other studies: [4] describe a two peaks pattern, with increased reconstruction accuracy and increased decoding accuracy for the interval of 170-250 ms; [14] show an increase of reconstruction accuracy when using late EEG response ; [8] found increased decoding accuracy for times lags between 130 to 220ms; [19]'s results showed an increase of cross-correlation between the envelope of the sound signal and M/EEG data at 80 ms, followed by a second peak of increased correlation. For music, the current results can be compared with results obtained by [15], where an early peak was also observed

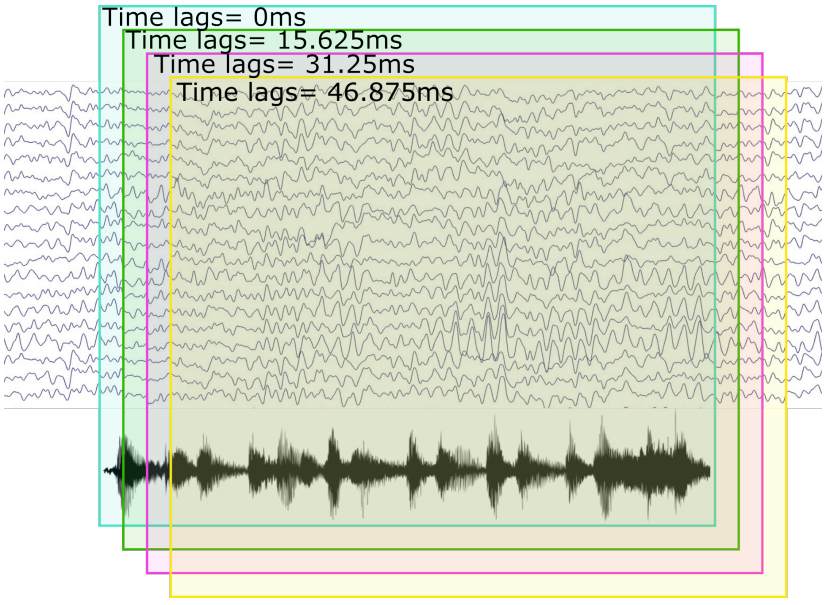

Fig. B.1: Schematic of the EEG data selection used for each single lag model. Each model is trained based on the 1 minute of audio data, and 1 minute of EEG data delayed compared to the audio data.

at -10 to 30 ms. [15] also observed a second peak of increased reconstruction accuracies at late latencies. However, they found this peak happening for time lags comprise between 460 and 500 ms, which is later than what we observe in the current study.

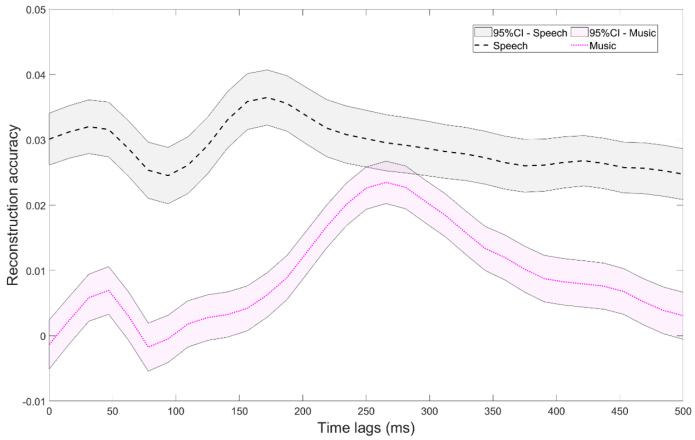

Fig. B.2: Reconstruction Accuracy across all time lags for trials where the target audio is Speech and trials where the target audio is Music.

3.2 Effect of attention

To explore if the attentional processes influence the temporal pattern of reconstruction accuracy, we compared reconstruction accuracy obtained with models trained to reconstruct the target of attention and models trained to reconstruct the distracter stimulus.

To that mean, separate models have been trained to either reconstruct envelopes of the target sound or to reconstruct envelopes of the distracter sound. Target models are trained in a leave-one-out approach, by using trials where the target is of the same time as the trial under test (either music or speech). The distractor models are trained with a similar approach, but by using trials where the distracter was the same type as the distracter of the trial under test. Comparing the reconstructions accuracy of these two models can give information about the effect of attention on cortical auditory processes: increased reconstruction accuracy at a given time lag observable for the target model but not for the distracter model can indicate an attentional effect.

3.3 Music listening scenario

Figure B.3 shows the reconstruction accuracies for both target models and distracter models for music listening. The two peak patterns can be seen for both models, with both peaks happening at similar time lags for both the Target model and the distracter model. Permutation tests, based on 100 000 permutations, were run to compare the reconstruction accuracy between models for each time lag. Significant differences were found for time lags between 260 and 285 ms, and also 320 and 380ms, where the Target model results in higher reconstruction accuracies than the distracter model. The maximum of reconstruction accuracies with a difference between the performance of the Target model compared to the distracter model, which suggests that a music-specific process around 265ms time lags may be affected by attentional processes.

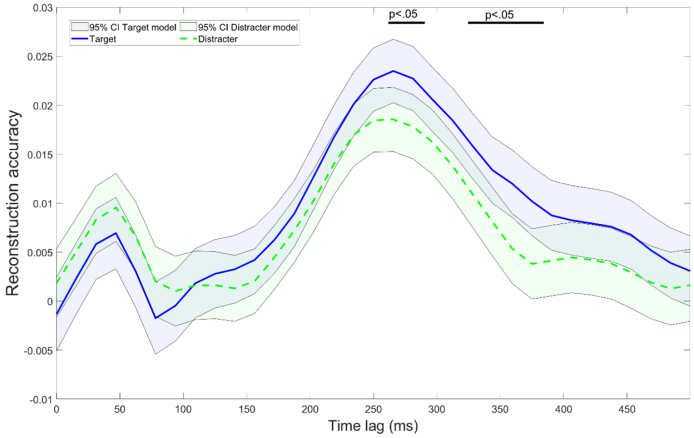

Fig. B.3: Reconstruction accuracy across all time-lags, obtained with Target model, with music as a target and Decoder model,with music as a distracter

3.4 Speech listening scenario

Figure B.4 shows the reconstruction accuracies for both target and distracter models for speech listening. Overall, reconstructions accuracies were obtained with the distracter models compared to the target model, for all time lags. Permutation tests, based on 100 000 permutations were run to compare the target model versus the distracter model, and indicate significant differences ($p<.05$) for all time lags, except between 355 and 410 ms. Despite the difference between the target and distracter models, the variation of the reconstruction accuracies across shows different trends for the models. For both models, the first peak of increased reconstruction accuracies can be observed between 30 and 50 ms.

However, while a second peak with maximal reconstruction accuracies arises at around 170 ms for the target model, there is no such increase for the distracter model. For the distracter model, an increase in reconstruction accuracy is observable at late time lags (350 to 450 ms) Taken together, it suggests that during speech listening, attentional processes affect the reconstruction accuracies level. The absence of a peak of maximal reconstruction when using the decoder model could indicate an increased attentional effect around a time lag of 170ms, as suggested by [4].

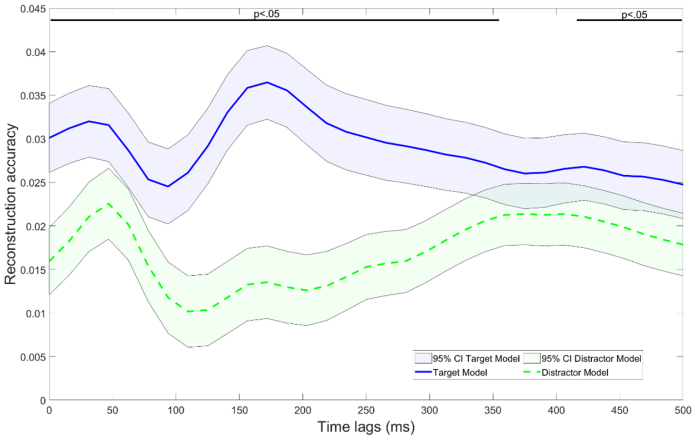

Fig. B.4: Reconstruction accuracy across all time-lags, obtained with Target model, with speech as a target and Decoder model, with speech as a distracter.

4 Discussion

In this study, we used linear regression to reconstruct sound heard from EEG data. By using single-lag models we explore the effect of time lags applied to EEG data to reconstruction accuracy. We compare models trained on speech and on music to highlight temporal differences in the cortical process for speech listening and music listening.

Overall, the reconstruction accuracy is higher for speech listening compared to music listening, for all time lags. This result was expected as performance differences for reconstruction accuracy have previously been observed between speech and music [20, 21].

For both Speech and music listening, a two-peak pattern can be observed: an early first peak of increased reconstruction accuracies for time lags around 30 to 50 ms, and a later second peak of maximal reconstruction accuracy,

where the timing differs between speech and music. This two-peak pattern has previously been observed for Speech listening [4] or Music listening [15].

Results suggest that the maximized reconstruction accuracies are obtained using different time lags for speech listening and music listening. For Speech, optimized reconstruction is obtained by applying a time lag of approximately 170 ms to the EEG data, relative to the audio signal. This is coherent with previous findings on optimal time lag for speech reconstruction or auditory attention decoding [4, 7, 8].

For music listening, the peak of optimized reconstruction is obtained for a time lag of approximately 265 ms. This optimal timing differs from previous results, where maximized reconstruction accuracies were found for music either at short time lags [11], or longer time lags [11, 15].

The early peak could suggest an early auditory process, which is coincident with both speech and music listening. The timing differences at middle/late timelags between speech and music listening suggest different cortical processes in place for speech listening and music listening.

A second question explored in the present study was to investigate the effect of selective auditory attention on reconstruction accuracies. To that mean, decoder models were trained to reconstruct the distracter stimulus, which was ignored by the participant. Comparing the reconstruction accuracies across time lags between the outputs of the target models and the distracter model can provide insight about the effect of attention.

For music, small but significant differences in reconstruction accuracies were found for time lags between 260 and 285 ms, and also 320 and 380ms. This difference is aligned with the peak of maximized reconstruction accuracies, which suggests that the process that creates this peak of maximal reconstruction might be affected by attention. On the other hand, the first peak of maximized reconstruction is similar for both target and distracter models. Taking together, these findings could suggest two separate cortical processes of music, an early one, not affected by selective auditory attention, and a late process, influenced by attentional processes.

For speech, the findings should be interpreted with more caution as a significant difference is observed for the Target and distracter model, across all time lags. These differences might be influenced by the difficulty of the task: when attending to a target sound is more challenging, an increased effort may be necessary to ignore the distracter, and the cortical tracking of the distracter might be reduced. Trials used for the speech decoder model correspond to trials where the listener had to attend to either speech signal or music signal in presence of distracting speech. However, attending to music in presence of speech is not a common task for human beings, compared for example to listening to speech in presence of music. This task might have been more challenging for the participants, which could influence the results observed in figure C.3.

Despite this offset between models, the shape of the peaks is worth considering. For both target and distracter models, an early peak of increased reconstruction accuracy is present for time lags of 30 to 50 ms. This suggests that the underlining cortical process is activated for stimuli that are inside or outside the focus of attention of the listener. For the target model, a second peak is observed at middle time lags (≈170 ms), while for the distracter model this peak is almost inexistent. It could indicate that the underlining cortical process that results in this peak is activated only for attended speech. It corroborates the idea developed in [4], which suggested an important attentional effect between 170 and 250 ms.

5 Conclusion

The present study explored the temporal aspect of cortical auditory processes and the effect of auditory attention by using a stimulus reconstruction approach. The results highlight two phases of auditory processes: an early process, which is concomitant for both speech and music listening and a second process, happening later for music listening than for speech listening. While the first process does not seem to be affected by attentional processes, the second might be enhanced by sounds that are actively attended by the listener. More research should be conducted to replicate, confirm, and elaborate on the current findings. Further analysis, such as using a forward model [6], could provide additional information on the cortical processes in places during listening in complex auditory sound scenes and explore more the similarities and differences between speech and music listening. Overall, this study highlights differences in optimal time-lags for cortical stimulus reconstruction between speech listening and music listening. This suggests temporal differences in cortical processing of speech and music. These differences could be used to fine-tune an auditory attention decoder that would be specifically tuned for music or speech.

References

[1] C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] N. Ding and J. Z. Simon, "Emergence of neural encoding of auditory objects while listening to competing speakers." *Proc. Natl. Acad. Sci. U.S.A.*, vol. 109, no. 29, pp. 11 854–11 859, jul 2012. [Online]. Available: https://pnas.org/doi/full/10.1073/pnas.1205381109

[3] ——, "Neural coding of continuous speech in auditory cortex during monaural and dichotic listening," *Journal of Neurophysiology*, vol. 107, no. 1, pp. 78–89, jan 2012. [Online]. Available: https://www.physiology.org/doi/10.1152/jn.00297.2011

[4] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG," *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, 01 2015. [Online]. Available: https://doi.org/10.1093/cercor/bht355

[5] P. J. Schäfer, F. I. Corona-Strauss, R. Hannemann, S. A. Hillyard, and D. J. Strauss, "Testing the Limits of the Stimulus Reconstruction Approach: Auditory Attention Decoding in a Four-Speaker Free Field Environment," *Trends in Hearing*, vol. 22, p. 233121651881660, jan 2018. [Online]. Available: http://journals.sagepub.com/doi/10.1177/2331216518816600

[6] E. Alickovic, T. Lunner, F. Gustafsson, and L. Ljung, "A Tutorial on Auditory Attention Identification Methods," *Frontiers in Neuroscience*, vol. 13, no. March, pp. 1–17, mar 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2019.00153/full

[7] S. A. Fuglsang, T. Dau, and J. Hjortkjær, "Noise-robust cortical tracking of attended speech in real-world acoustic scenes," *NeuroImage*, vol. 156, no. April, pp. 435–444, aug 2017. [Online]. Available: http://dx.doi.org/10.1016/j.neuroimage.2017.04.026https://linkinghub.elsevier.com/retrieve/pii/S105381191730318X

[8] B. Mirkovic, S. Debener, M. Jaeger, and M. De Vos, "Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications," *Journal of Neural Engineering*, vol. 12, no. 4, p. 046007, aug 2015. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2560/12/4/046007

[9] W. W. An, B. Shinn-Cunningham, H. Gamper, D. Emmanouilidou, D. Johnston, M. Jalobeanu, E. Cutrell, A. Wilson, K.-J. Chiang, and I. Tashev, "Decoding Music Attention from "EEG Headphones": A User-Friendly Auditory Brain-Computer Interface," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, IEEE. IEEE, jun 2021, pp. 985–989. [Online]. Available: https://ieeexplore.ieee.org/document/9414492/

[10] G. Cantisani, S. Essid, and G. Richard, "EEG-Based Decoding of Auditory Attention to a Target Instrument in Polyphonic Music," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*

(WASPAA), vol. 2019-Octob, no. 765068. IEEE, oct 2019, pp. 80–84. [Online]. Available: https://ieeexplore.ieee.org/document/8937219/

[11] L. Hausfeld, N. R. Disbergen, G. Valente, R. J. Zatorre, and E. Formisano, "Modulating Cortical Instrument Representations During Auditory Stream Segregation and Integration With Polyphonic Music," *Frontiers in Neuroscience*, vol. 15, no. September, pp. 1–15, sep 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2021.635937/full

[12] G. M. Di Liberto, J. A. O'Sullivan, and E. C. Lalor, "Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing," *Current Biology*, vol. 25, no. 19, pp. 2457–2465, oct 2015. [Online]. Available: http://dx.doi.org/10.1016/j.cub.2015.08.030https://linkinghub.elsevier.com/retrieve/pii/S0960982215010015

[13] M. J. Crosse, N. J. Zuk, G. M. Di Liberto, A. R. Nidiffer, S. Molholm, and E. C. Lalor, "Linear Modeling of Neurophysiological Responses to Speech and Other Continuous Stimuli: Methodological Considerations for Applied Research," *Frontiers in Neuroscience*, vol. 15, no. November, nov 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2021.705621/full

[14] E. Alickovic, E. H. N. Ng, L. Fiedler, S. Santurette, H. Innes-Brown, and C. Graversen, "Effects of Hearing Aid Noise Reduction on Early and Late Cortical Representations of Competing Talkers in Noise," *Frontiers in Neuroscience*, vol. 15, no. March, mar 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2021.636060/full

[15] L. Hausfeld, L. Riecke, G. Valente, and E. Formisano, "Cortical tracking of multiple streams outside the focus of attention in naturalistic auditory scenes," *NeuroImage*, vol. 181, no. March, pp. 617–626, nov 2018. [Online]. Available: https://doi.org/10.1016/j.neuroimage.2018.07.052https://linkinghub.elsevier.com/retrieve/pii/S1053811918306670

[16] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis." *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, 2004.

[17] L. Pion-Tonachini, K. Kreutz-Delgado, and S. Makeig, "Iclabel: An automated electroencephalographic independent component classifier, dataset, and website." *NeuroImage*, vol. 198, pp. 181–197, 2019.

[18] D. D. E. Wong, S. A. Fuglsang, J. Hjortkjær, E. Ceolini, M. Slaney, and A. de Cheveigné, "A Comparison of Regularization Methods in Forward and Backward Models for Auditory Attention Decoding," *Frontiers in*

Neuroscience, vol. 12, no. AUG, pp. 1–16, aug 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2018.00531/full

[19] M. Wöstmann, L. Fiedler, and J. Obleser, "Tracking the signal, cracking the code: Speech and speech comprehension in non-invasive human electrophysiology," *Language, Cognition and Neuroscience*, vol. 32, no. 7, pp. 855–869, 2017.

[20] A. Simon, G. Loquet, J. Østergaard, and S. Bech, "Cortical Auditory Attention Decoding During Music And Speech Listening," *(In review)*, 2023.

[21] N. J. Zuk, J. W. Murphy, R. B. Reilly, and E. C. Lalor, "Envelope reconstruction of speech and music highlights stronger tracking of speech at low frequencies," *PLOS Computational Biology*, vol. 17, no. 9, p. e1009358, sep 2021. [Online]. Available: http://dx.doi.org/10.1371/journal.pcbi.1009358https://dx.plos.org/10.1371/journal.pcbi.1009358

Paper C

Electrodes selection for cortical auditory attention decoding with EEG during speech and music listening

Adèle Simon, Søren Bech, Gérard Loquet and Jan Østergaard

The paper has been published in the
Proceedings of the *25th International Conference on Information Fusion (FUSION)*, Linköping, Sweden, 2022.

The layout has been revised.

Abstract

In complex sound scenes, where multiple sounds are present around a listener, selective attention to one auditory stream is hypothesized to synchronize low-frequency brain activity with the envelope of the attended streams. Recent research has employed stimulus reconstruction from neural data to decode to which auditory stream a listener is paying attention. This could be used to create an auditory attention decoder (AAD), that could be embedded in smart headphones or hearing aids, that would adapt the sound processing based on the attention of the user.

However, most of these studies use full scalp electroencephalogram, which is not suitable for implementations in audio devices. To that aim, a smaller EEG device, with fewer electrodes could be used.

In the present study, we explore the performance of an AAD based on a smaller number of electrodes during speech and music listening. Participants were presented with two sounds simultaneously, and where asked to attend to one while ignoring the other, and their cortical response was continuously recorded during the listening. Using a greedy approach based on reconstruction accuracy, a subset of EEG electrodes that are optimized for linear stimulus reconstruction were selected. The goal of this study is to explore the performance of a linear AAD when reducing the number of electrodes.

Results suggest that four well-selected electrodes can be sufficient for a miniaturized AAD as it performs as well as a 64-channels setup. The channels selected vary depending on the type of sound attended, suggesting that different electrodes placement should be used to decode attention during music listening and speech listening.

1 Introduction

In complex auditory environments, where multiple sounds are present, human beings have the ability to discriminate these auditory sources and to focus their attention on only one of them. This is useful for following a conversation at a cocktail party or enjoying our own music in a crowded and noisy environment. However, this task can be demanding and difficult in some situations, especially for people with hearing impairment. While is it possible to enhance one source of sounds to make it easier to follow, for example by raising the level of the sound of interest, this requires knowing what the listener wants to listen to.

In recent years, a large number of studies focused on this problem of auditory attention detection, to determine to what sound a listener is paying attention to. Brain recording has shown promising results: It has been demonstrated that auditory attention affects the cortical tracking of the attended stream of sound, compared to ignored sounds [1–3].

This has been used to implement an Auditory Attention Decoder (AAD), using the stimulus reconstruction approach, which consists in reconstructing the envelope of a heard sound from neural data [2, 4]. When a listener is in presence of multiple sound streams, it is possible to reconstruct the sound envelope of the sounds heard by the listener. If the listener's attention is focused on one stream of sound, it can be reconstructed more accurately than other ignored streams of sound. Such a system relies on an electroencephalogram (EEG) to measure the cortical tracking of the sound.

Initially, most of these studies used full scalp EEG, with a large number of electrodes (32, 64 or more). Such equipment is heavy, bulky, and not suited for applications such as neuro-steered hearing aids or other audio devices. Several studies aimed to tackle this issue by investigating how and to what extent the number of electrodes needed for an AAD can be reduced [5–10]. Overall, these studies suggest that, if placed in specific locations, the number of electrodes can be reduced.

An AAD based on stimulus reconstruction have been proven successful for cocktail party scenario [2, 7, 8], where only speech is present, and also during music listening situation, where a listener focuses on one instrument in a multi-instrumental piece [5, 11, 12]. However mixed cases, where both music and speech are presented have rarely been investigated, even though it corresponds to a frequently occurring situation for neuro-steered (e.g., listening to music in a noisy environment, following a conversation in a restaurant with background music).

In the present study, we address this mixed situation. Equipped with a 64-channels EEG, participants in the study listened to a target sound, which was either speech or music, while also being subjected to a distracter sound, which was likewise speech or music. From the recording, a greedy algorithm was used to select EEG channels that maximize stimulus reconstruction, for either speech listening or music listening. The main goal is to determine to what extent the number of electrodes can be reduced without decreasing the performance of the AAD. The second aim of this study is to investigate if the optimal electrode placement differs between speech-listening conditions and music-listening conditions.

2 Methods

Thirty-five English-speaking participants (14 females) were recruited for this study. Participants' age ranged between 21- and 33-year-old (mean = 26,29). None of the participants reported neurological disease or hearing loss. The participants were compensated for their time, and all of them signed a written informed consent form.

2.1 Experiment & Stimuli

During each experiment, the participants were presented to listen to two separate streams of sounds from two different spatial locations, from loudspeakers located at +/- 30-degree azimuth. They were asked to focus on only one of them (target) while ignoring the other (distracter). Stimuli used were excerpts of music and speech, of one minute each:

- Piano Music: 8 excerpts of mono instrumental pieces of classical played on a piano
- Electronic music: 8 excerpts of polyphonic pieces of instrumental electronic music
- Speech female: 8 excerpts of an audiobook read by a woman in English
- Speech male: 8 excerpts of an audiobook read by a man in English

Stimuli were selected to obtain a balanced number of the following conditions: Target = Music & Distracter = Music; Target = Music & Distracter = Speech; Target = Speech & Distracter = Music; Target = Speech & Distracter = Speech. Each excerpt was used once as a target, resulting in thirty-two trials per participant. Distracters were selected randomly for the same pool of stimuli, to obtain the balance number of trials per condition.

2.2 EEG data acquisition and preprocessing

Each participant participated in a single session where their cortical response was continuously recorded at 512 Hz using 64 channels g. HIAMP-Research system (g.tec medical engineering GmbH, Austria). The electrodes were installed following the 10-20 system. The impedance of each electrode was kept below 5 kOhms. After data collection, the EEG data were pre-processed using the EEG lab toolbox [13].

EEG data were referenced to the average of all scalp electrodes. Noise-contaminated EEG channels were visually assessed and interpolated from nearby electrodes. To eliminate components linked to eye blinks, an Independent Component Analysis was performed. The EEG data were downsampled to a sampling rate of 100Hz after being bandpass filtered between 1 and 8Hz. The trials in which the artefacts were too important (e.g., due to movements) were discarded. Data from two participants were excluded because of artefacts. The data that was eliminated accounts for 7.68 percent of the overall data.

The audio signals' amplitude envelopes were extracted using a Hilbert transform and then downsampled to the same sampling rate of 100 Hz.

2.3 Stimulus reconstruction from neural signal & attention decoding

It has been shown that auditory attention decoding from the neural signal can be performed through the stimulus reconstruction method [2, 4, 14]. For this approach, the EEG data is used to reconstruct an approximation of the input stimuli, through a linear reconstruction multi delay model.

$$s'(t) = \sum_{n} \sum_{\tau} g(\tau, n) R(t - \tau, n) \tag{C.1}$$

where s' is the reconstructed envelope, $R(t - \tau, n)$ is the EEG response at the time$(t - \tau)$ for electrode n, and g is the linear model, which is a function of electrode n and time lags τ. For the present experiment, time delays range from 0 to 500 milliseconds following the stimulus.

The mean squared error between the original and reconstructed envelopes, which can be computed analytically using ridge regularization methods, may be used to estimate the model $\hat{g}$. [15]:

$$\hat{g} = (R^T R + I\lambda)^{-1} R^T S \tag{C.2}$$

where S is the envelope of the target sound, I is the identity matrix, λ is the regularization parameter used to prevent overfitting [4, 15]. Here, R is a matrix that contains contains all the delay EEG responses at time-lags τ from 0 to 500 ms post stimulus, in a similar approach that the one described in O'Sullivan et al [2].

In the presented analysis, the value of λ was chosen by calculating several models with different values of this regularization parameter, and set to 10^5, based on the value that maximized the performance of the AAD.

We measured the reconstruction accuracy by calculating the Pearson's correlation coefficient between the original target envelope and the reconstructed one (r_{target}). The correlation between the reconstructed envelope and the envelope extracted from the distracter was also calculated ($r_{distracter}$).

The attention decoding for each reconstruction was assessed by comparing the two correlation coefficients. If the reconstructed envelope showed a higher correlation with the target envelope than the distracter envelope (i.e., $r_{target} > r_{distracter}$), the trial was regarded as successfully decoded.

2.4 Models training

The models were trained using a leave-one-out cross-validation procedure for each trial. Models were not individualized by participants, thus all data from all participants were used for training.

Each trial was decoded with a model created by averaging the parameters of all other trials' models. The training data were chosen to match the trial

under test: the model was only trained on trials in which the target stimulus was from the same genre as the trial under test. For instance, when testing on a trial where the participant was actively listening to piano music, all other trials where the target was also piano music were used to train the model.

2.5 Greedy selection of electrodes based on reconstruction accuracy

In order to reduce the number of electrodes, the goal is to select the channel that contributes the most to the auditory attention decoder. As it is based on stimulus reconstruction of the input stimulus, one approach is to select EEG channels that show the greatest reconstruction accuracy. Here, we select channels one by one, using a greedy algorithm, by using the reconstruction accuracy as a criterion.

The greedy selection was done with data from all 33 participants polled together.

At first, all channels were treated independently, by calculating a model g_i for each channel, by using only the data from the electrode i. The model was used to compute the reconstructed signal s_i, which was compared the input signal s, to obtain the reconstructed accuracy r_i. The channel i that gives the highest reconstruction accuracy r_i was selected and used for the following round of channel selection.

$ElecRemain = [1:64]$
$ElecSelect = []$

while $length(ElecSelect) <= MaxElecs$ **do**
  for $i = 1 : length(ElecRemain)$ **do**
    $Subset = concat(ElecSelect, ElecRemain(i))$
    $g_i = (R_{subset}^T R_{subset} + \lambda I)^{-1} R_{subset} S$
    $S_i' = g_i * R_{subset}$
    $r_i = corr(S, S_i')$
  end for
  % Find electrodes that gives highest correlation
  $j = argmax(r_i)$
  $ElecSelect = concat(ElecSelect, ElecRemain(j))$
  $ElecRemain(j) = []$
end while

It results in an ordered list of the 32 channels that optimized the reconstruction of the input stimulus the most. In the present experiment, two lists of channels were selected:

- A list of electrodes was calculated exclusively on trials where the partici-

pant was focusing on music. The electrodes selected are then optimized for music reconstruction.

- A list of electrodes calculated exclusively on trials where the participant was focusing on speech, with electrodes that maximize speech reconstruction.

3 Results

3.1 Greedy selection based on reconstruction accuracy

An AAD, as described in section 2.3, was tested on data from several subsets of electrodes, selected through the greedy algorithm described in section 2.5. The success rate is calculated for each subset of electrodes and compared to the baseline, which is the success rate obtained with the 64 electrodes dataset. The greedy algorithm was applied on speech and music trials separately, to extract a list of electrodes that optimize music reconstruction and the second list of electrodes that optimize speech reconstruction. The list of electrodes that optimize Music reconstruction (by order of selection) is shown below:

- TP8 – C4 – P9 – C6 – TP7 – C5 – Fp1 – F1 – T7 – C1 – Cz – F8 – CP2 – F2 – P7 – CP5 – F4 – FC2 – FT7 – POz – Pz – P1 – P4 – CP3 – P6 – Oz - O1 – FCz – F5 – FC4 – T8 - Afz

The list of electrodes that optimize Speech reconstruction (by order of selection) is shown below:

- - P9 – FT7 – C6 – C3 – P1 - C5 – TP7 – T8 – AFz - FC5 - F6 – Fp1 – FC3 – Fz – C1 – C2 – F8 – FC6 – P5 – AF3 – P7 – FCz – CP3 – TP8 – P4 – O2 – F5 – F4 – CP1 – AF4 – FT8 – C4

We can observe that some channels are selected in both lists, however, the order varies. P9 and C6, or even C5 or TP7 were selected early in both lists, which could be interpreted as channels useful for generic auditory attention decoding. On the contrary, C4, which is selected early in the music condition, but late for the speech condition may be more specific, and more useful for the reconstruction of the musical signal. For the following analysis, subsets of electrodes will be used based on those two lists, by selecting the N first electrodes of each list. For testing, subsets of electrodes were selected from these lists, following the order of selection: for example, for speech, the subset of two electrodes is composed of P9 and FT7, and the four-electrode subset is composed of P9, FT7, C6 and C3.

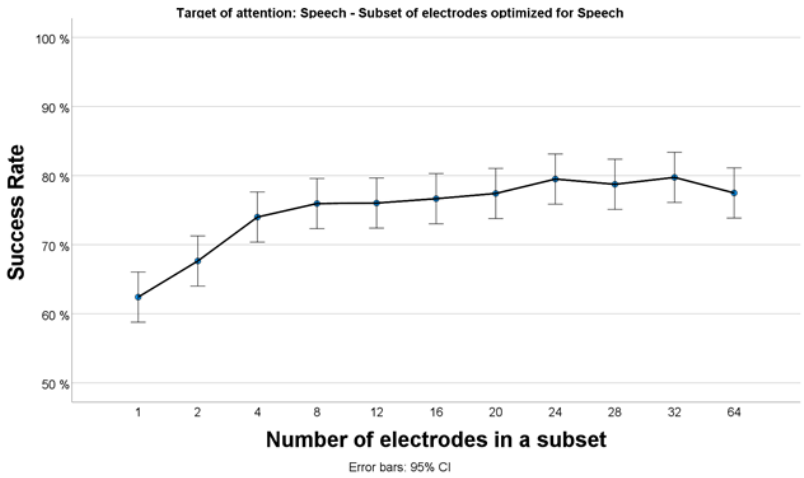

Fig. C.1: Success rate of the electrodes depending on the size of the subset for speech-listening condition.

3.2 Electrodes reduction for speech listening

A first analysis was conducted of the speech data, to compare the performance of each subset of electrodes selected through the greedy algorithm. The goal of this analysis is to find the minimum number of electrodes needed in a well-selected subset to obtain a performance of the AAD that is comparable to a full-scalp 64-channels setup. The analysis was done on the speech data, i.e., only the trial where the Target of attention was speech, and with the subsets of electrodes that were selected via the greedy algorithm to optimize speech reconstruction.

A a full factorial two-way ANOVA was conducted, using the number of electrodes used as a fixed factor and participants as a random factor. There was a significant main effect of the number of electrodes on success rate, $F(10, 346.696) = 17.708, p < 0.001$. Planned contrasts were used to test which subset of electrodes performed significantly worse than the 64-channel baseline. Planned contrasts revealed that having only the subset composed of one electrode ($p < 0.001$) or two electrodes ($p < 0.001$) significantly reduce the success rate of the AAD. (See figure C.1)

This suggests that by using a subset of only four well-selected electrodes, we can obtain performance of the AAD that is comparable to that of an AAD based on full scalp 64-channels EEG. It has to be noted that a significant effect of the participant was also found $F = (32, 320) = 31.58, p < 0.001$, suggesting a variation of the success rate of the AAD across participants. No interactions between the participant's factor and the number of electrodes were observed, which indicates that the placement of the optimal electrodes does not vary

significantly across participants.

3.3 Electrodes reduction for music listening

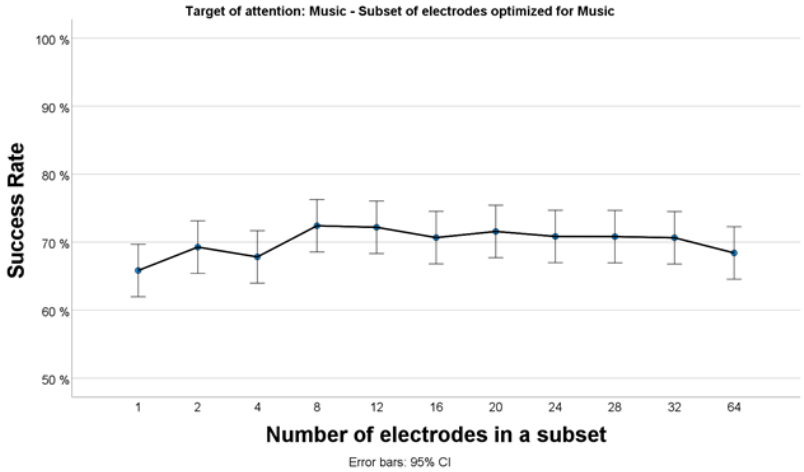

Fig. C.2: Success rate of the electrodes depending on the size of the subset for music-listening condition.

An analysis, similar to the one performed on speech data was done on music data, using the subsets of electrodes selected through the greedy algorithm to optimize music reconstruction.

A two-way ANOVA was conducted that examined the effect of the number of electrodes used on the success rate of an AAD, participants were used as a random factor. There was a significant main effect of the number of electrodes on success rate, $F(10, 326.584) = 2.716, p = 0.003$. Planned contrasts were used to test which subset of electrodes performed significantly worse than the 64-channel baseline. Planned contrasts showed no significant differences in performance between any subset and the 64-channel baseline.

It suggests of any subset of the selected electrodes can perform as well as a full-scalp EEG. However, visual inspection suggests a small increase in AAD performance for subsets composed of either 8 or 12 electrodes. (See figure C.2)

A significant effect of the participant was also found $F = (32, 320) = 37.42, p < 0.001$, suggesting variation of the success rate of the AAD across participants. No interactions between the participant's factor and the number of electrodes were observed.

3.4 Optimization specific to types

The second goal of this experiment is to investigate if some electrodes are specific to speech or music reconstruction. In order to explore this question, we focus on a four-electrodes subset. The above analysis showed that four electrodes well-selected can be sufficient to obtain decoding success comparable to those obtained with a 64-channels setup.

We selected a subset of the four electrodes that optimize Speech reconstruction (Speech subset), and another comprises of the four electrodes that optimize Music reconstruction (Music subset). The two subsets have two electrodes in common (P9 and C6) (see figure C.4). Trials where the target of attention was music was decoded using data from the music-optimized subset (congruent condition) and the speech-optimized (incongruent condition), and respectively for trials where the target of attention was speech.

A two-way ANOVA was conducted to assess the influence of this congruence on the success rate, with target of attention (music or speech) and congruence as factors. There was a significant main effect of the congruence between the target type and the subset used, $F(1, 2064) = 8.326, p = 0.004$. A significant effect of the target type was found $F = (1, 2064) = 12.426, p < 0.001$, suggesting that the performances differ if the target is speech or music. No interactions between the two factors of electrodes were observed. (see figure C.3)

Results are displayed in figure C.3. It can be observed that in both cases, the success rate of the AAD decreases when using a subset of electrodes not optimized for the specific type under test: when using the Speech subset, the success of AAD is reduced when decoding trials where the participant was attentive to music (-7.52% of success rate). A similar pattern is observed for speech listening conditions, but to a lesser extent ($-4,28\%$ of success rate).

4 Discussion

This study used a greedy approach to select, out of a 64-channel EEG recording, which EEG channel maximize stimulus reconstruction, in the context of a linear auditory attention decoder.

The goals were two-fold: explore the performance of AAD based on a limited number of electrodes; investigate differences between music listening and speech listening and explore if some channels maximize the reconstruction of one specific type of sound.

We used a previously published method for auditory attention decoding [2], based on linear stimulus reconstruction of the audio envelope from the neural signal. Several AAD were implemented, based on reduced subsets of electrodes (1,2,4,8,12,16,20.24,28 or 32 electrodes). For electrodes selection,

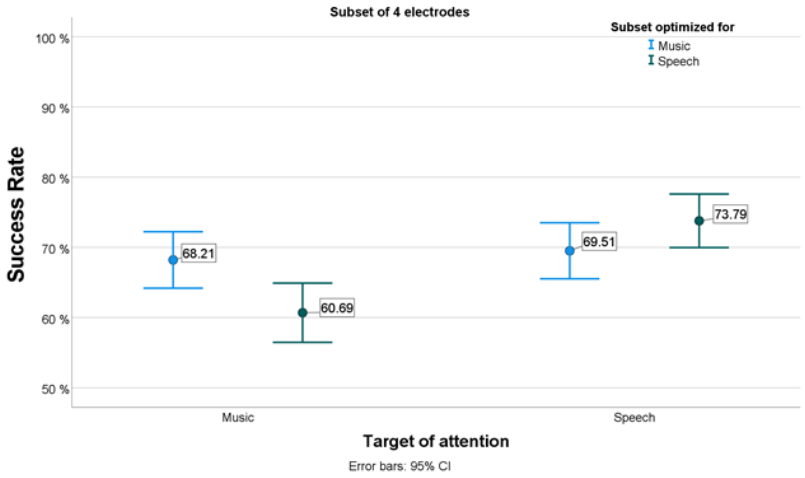

Fig. C.3: Success rate of the AAD, based on listening condition and electrode selection.

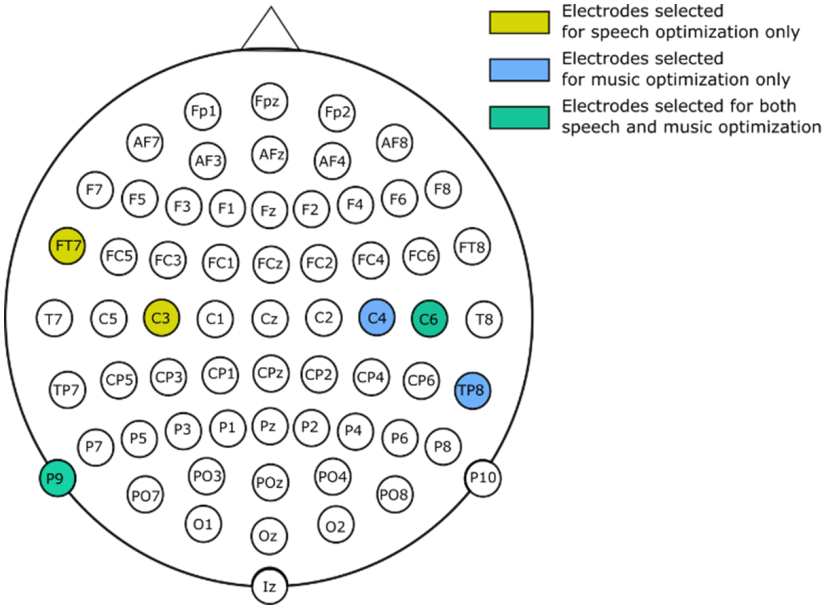

Fig. C.4: Placement of the electrodes selected for speech reconstruction optimization and music reconstruction optimization.

a greedy algorithm was used to select the EEG channels that optimize the reconstruction accuracies of the target sound. The findings suggest that, when well-selected, the number of electrodes can be reduced to four electrodes without significantly reducing the performance of the AAD. During attentive music listening, one or two electrodes may be sufficient. However, in this study, the channel selection was based on the reconstruction accuracy of the target stimulus only. Auditory attention decoding also relies on the reconstruction accuracy of the distracter, assuming that it would be smaller than the targets. To complement the present results, further investigation could explore the influence of channel reduction on the targets and the distracter's reconstruction accuracy.

With respect to the second objective of the study, it was found that the selected channel for speech-listening and music-listening was not similar. In the 4-channels AAD scenario, two electrodes were common to both listening conditions, while the two others differ. We then compare the performance variations induced when using channels selected to optimize for a different type of sound. When listening attentively to music, an AAD using 4-channels selected for speech will be less successful than AAD that is fitted for music reconstruction. A similar pattern arose for attentive listening to speech.

It suggests that while some electrodes contribute to both speech and music reconstruction, other electrodes may maximize either speech or music reconstruction. A possible interpretation of this might be that different cortical regions are involved in speech and music processing. AAD might be using cortical responses from these specific regions, activated during the processing of either speech or music. Nevertheless, further research would be needed to investigate those differences. It has to be noted that the placement of many of the selected electrodes are concentrated around the temporal lobe, which is consistent with the previous findings on EEG-channels reductions for AAD [7, 9, 10].

This study focused on generalised electrodes selection, based on the data from 33 participants, with no individualization by participants. Nonetheless, a significant effect of participants was found during the analysis, future work could explore the individualized selection, as it has been shown to help improve the performance of AAD [2, 8, 10]. However, as no interactions between the electrodes and the subject were found during the analysis, the results suggest that the number of electrodes needed does not depend on the participants.

5 Conclusion

Through a greedy channels selection based on reconstruction accuracy, we showed that an auditory attention decoding is feasible with a well-selected subset of electrodes. Our results suggest that four electrodes could be sufficient

to successfully decode auditory attention, with performance equivalent to a 64-channels setup, for both speech listening and music listening. These results are promising for potential real-life application of AAD. This study also highlights the fact that some electrodes, hence potentially specific brain regions, contribute more to music reconstruction than speech reconstruction. This specificity could be of importance when designing a neuro-steered audio device, where the placement of electrodes should then be done according to the type of audio signal present in the use case.

References

[1] N. Ding and J. Z. Simon, "Emergence of neural encoding of auditory objects while listening to competing speakers." *Proc. Natl. Acad. Sci. U.S.A.*, vol. 109, no. 29, pp. 11 854–11 859, jul 2012. [Online]. Available: https://pnas.org/doi/full/10.1073/pnas.1205381109

[2] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG," *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, 01 2015. [Online]. Available: https://doi.org/10.1093/cercor/bht355

[3] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, pp. 233–236, may 2012. [Online]. Available: http://www.nature.com/articles/nature11020

[4] E. Alickovic, T. Lunner, F. Gustafsson, and L. Ljung, "A Tutorial on Auditory Attention Identification Methods," *Frontiers in Neuroscience*, vol. 13, no. March, pp. 1–17, mar 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2019.00153/full

[5] W. W. An, B. Shinn-Cunningham, H. Gamper, D. Emmanouilidou, D. Johnston, M. Jalobeanu, E. Cutrell, A. Wilson, K.-J. Chiang, and I. Tashev, "Decoding Music Attention from "EEG Headphones": A User-Friendly Auditory Brain-Computer Interface," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, IEEE. IEEE, jun 2021, pp. 985–989. [Online]. Available: https://ieeexplore.ieee.org/document/9414492/

[6] L. Fiedler, M. Wöstmann, C. Graversen, A. Brandmeyer, T. Lunner, and J. Obleser, "Single-channel in-ear-EEG detects the focus of auditory attention to concurrent tone streams and mixed speech," *Journal of Neural*

Engineering, vol. 14, no. 3, p. 036020, jun 2017. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2552/aa66dd

[7] S. A. Fuglsang, T. Dau, and J. Hjortkjær, "Noise-robust cortical tracking of attended speech in real-world acoustic scenes," *NeuroImage*, vol. 156, no. April, pp. 435–444, aug 2017. [Online]. Available: http://dx.doi.org/10.1016/j.neuroimage.2017.04.026https://linkinghub.elsevier.com/retrieve/pii/S105381191730318X

[8] B. Mirkovic, S. Debener, M. Jaeger, and M. De Vos, "Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications," *Journal of Neural Engineering*, vol. 12, no. 4, p. 046007, aug 2015. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2560/12/4/046007

[9] A. M. Narayanan and A. Bertrand, "The effect of miniaturization and galvanic separation of EEG sensor devices in an auditory attention detection task," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, vol. 2018-July. IEEE, jul 2018, pp. 77–80. [Online]. Available: https://ieeexplore.ieee.org/document/8512212/

[10] ——, "Analysis of miniaturization effects and channel selection strategies for eeg sensor networks with application to auditory attention detection," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 1, pp. 234–244, 2019.

[11] G. M. Di Liberto, G. Marion, and S. A. Shamma, "Accurate Decoding of Imagined and Heard Melodies," *Frontiers in Neuroscience*, vol. 15, no. August, pp. 1–11, aug 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2021.673401/full

[12] L. Hausfeld, N. R. Disbergen, G. Valente, R. J. Zatorre, and E. Formisano, "Modulating Cortical Instrument Representations During Auditory Stream Segregation and Integration With Polyphonic Music," *Frontiers in Neuroscience*, vol. 15, no. September, pp. 1–15, sep 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2021.635937/full

[13] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis." *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, 2004.

[14] M. J. Crosse, G. M. Di Liberto, A. Bednar, and E. C. Lalor, "The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli," *Frontiers in Human*

Neuroscience, vol. 10, no. NOV2016, pp. 1–14, nov 2016. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fnhum.2016.00604/full

[15] D. D. E. Wong, S. A. Fuglsang, J. Hjortkjær, E. Ceolini, M. Slaney, and A. de Cheveigné, "A Comparison of Regularization Methods in Forward and Backward Models for Auditory Attention Decoding," *Frontiers in Neuroscience*, vol. 12, no. AUG, pp. 1–16, aug 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2018.00531/full

Paper D

Cortical linear encoding and decoding of sounds: Similarities and differences between naturalistic speech and music listening

Adèle Simon, Søren Bech, Gérard Loquet and Jan Østergaard

The paper has been submitted to the
European Journal of Neuroscience .

The layout has been revised.

Abstract

Linear models are becoming increasingly popular to investigate brain activity in response to continuous and naturalistic stimuli. In the context of auditory perception, these predictive models can be "encoding", when stimulus features are used to reconstruct brain activity, or "decoding", when neural features are used to reconstruct the audio stimuli. These linear models are a central component of some Brain-Computer Interfaces that can be integrated into hearing assistive devices (e.g., hearing aids). Such advanced neurotechnologies have been widely investigated when listening to speech stimuli but rarely when listening to music. Recent attempts at neural tracking of music show that the reconstruction performances are reduced compared to speech decoding. The present study investigates the performance of stimuli reconstruction and EEG prediction (decoding and encoding models) based on the cortical entrainment of temporal variations of the audio stimuli for both music and speech listening. Three hypotheses that may explain differences between speech and music stimuli reconstruction were tested to assess the importance of the speech-specific acoustic and linguistic factors. While the results obtained with encoding models suggest different underlying cortical processing between speech and music listening, no differences were found in terms of reconstruction of the stimuli or the cortical data. As all listening conditions can be successfully reconstructed from both encoding and decoding models based on an audio envelope, the results are promising for future developments of versatile brain-computer interfaces that could be used for listening situations not limited to speech.

1 Introduction

In recent years, an increased interest in linear modelling approaches can be observed in the field of auditory cognitive neuroscience [1–3]. Such approaches attempt to estimate or reconstruct brain activities from an input stimulus (forward or encoding modelling [4] or, conversely, to reconstruct an input stimulus from the corresponding brain recording (backwards or decoding modelling [5]). The present study applies such modelling methods to music and speech listening.

As opposed to traditional methods in auditory neuroscience, which are mostly based on statistical analyses of direct evoked responses in the auditory system, linear modelling approaches provide a data-driven way to study relationships between auditory stimuli and cortical responses. One of the main advantages is that they allow the investigation of continuous stimuli, closer to ecologically valid situations [2, 6].

Such linear modelling techniques can also lead, in addition to research purposes, to concrete applications and Brain-Computer Interfaces. Stimulus reconstruction, through backward modelling, could be used for cortical

auditory attention decoding (AAD) [7–9]. The AAD application could be implemented in hearing aids, hearing assistive devices, or even consumer audio products (headphones or earbuds) [8, 9]. It is based on the fact that the cortical tracking of auditory stimuli to which a listener is actively attending is enhanced compared to distracting sounds that are not the focus of the listener's attention [5, 7]. Knowing to what sound a listener is attending to could help to adapt the sound scene to make listening easier or reduce listening effort: for example, facilitate listening to one external source of the sound through beamforming, or reducing distraction by cancelling surrounding noises if the user wants to focus on the headphone media.

While stimulus reconstruction, especially in the context of AAD, has been widely investigated for speech listening [5, 10–14], the research is not as extended when it concerns other types of signals, especially musical signals. However, music holds an important place in human life [15], and one could imagine extending the use of AAD to more versatile scenarios to make it useable in complex situations where both speech and music could be the target of the listener's attention. Linear modelling has been successful for musical signals [16–22].

However, compared to speech reconstruction while using similar methods, the reconstruction accuracy obtained for music reconstruction tends to be lower [23–25]. In the context of AAD, reduced reconstruction accuracy for music has been observed for both target reconstruction and distracter reconstruction [23, 24]. Such results have implications for future developments of AAD, as reduced reconstruction accuracy for music may bias the attention decoding toward speech in mixed situations.

The motivation for the current study is to move toward a versatile AAD that could decode and adapt to a variety of sounds that one user may encounter in realistic environments. Towards this goal, the present study explores the performances of stimulus reconstruction for different types of auditory stimuli, attempts to replicate existing results, and explores factors that may influence the differences between speech and music stimuli. While the underlying motivation relates to AAD, the attentional factors are not the central point of the present study. In order to investigate how linear modelling performs on speech and music and to reduce biases due to potential variations of attention between speech and music, the current study is based on listening to only one auditory stream at a time.

The present investigation compares both stimulus reconstruction and EEG prediction. In addition, temporal response functions (TRFs) [26] obtained through an envelope-based linear encoding model are compared across conditions to gain insight into the underlying cortical processes at work during speech or music listening. Since the variation between speech and music reconstruction accuracy would influence the performance of a versatile AAD, it is important to understand first better how the reconstruction performance

differs for different types of stimuli.

Inspired by existing results in the literature on speech and music encoding and decoding, five music and speech listening conditions are compared in the present study: instrumental music; music with lyrics understood by the participants; music with lyrics not understood; speech understood; and speech not understood by the participants. Three hypotheses that may explain the difference between speech and music when using encoding and decoding linear models are proposed and tested:

- **H1**: Differences are due to enhanced tracking of speech-specific acoustic processing, such for instance, prosodic features [25, 27, 28]. It also has been shown that temporal modulations of speech differ from temporal modulation of music [29] (as an example, see Supplementary material S-B for the power spectrum of the stimuli used in the present study). The presence of gaps between words, specific to speech, led to differences in shape between the music and speech temporal envelopes [25]. The differences in the acoustical properties could explain the differences in performance:

Higher reconstruction accuracy is expected for speech (both understood and not understood) compared to music (all conditions). Different patterns of TRFs are also expected between speech and music conditions, which could suggest distinct underlying cortical processes involved in naturalistic speech listening or music listening. (Figure D.1- A)

- **H2**: Differences are not due to acoustic factors but to high-level cortical processing of speech, such as content processing or higher linguistic features specific to speech understanding.

While the cortical tracking of the temporal variation of an auditory stimulus (i.e. temporal envelope) is suspected to be mostly driven by general acoustical processing [30], other factors have also been found to influence the cortical envelope tracking, such as intelligibility [30–32], semantic processing [28], clarity, comprehension [33] or language proficiency [34]. In contrast, it has been shown that better speech understanding is not necessarily linked to greater acoustic neural tracking [35]. For the present study, the hypothesis under test is that understanding the speech signal's semantic content increases cortical tracking. Under this hypothesis, different TRF patterns would be expected between the conditions that involve linguistic speech processing (speech understood and music with lyrics understood) compared to the others, and reconstruction accuracy and EEG prediction accuracy would be higher for understood content (Figure D.1-B).

- **H3**: The parameters chosen in the design of the model, especially the audio feature (i.e., temporal envelope), biased the model to favour speech.

For music reconstruction, other audio features may lead to better reconstruction compared to the envelope [18]. Spectral modulation may be more important than temporal modulation for cortical music tracking [1]. For this hypothesis, when using a model based on the audio envelope, stimuli containing speech only should be better reconstructed than with music only. Reconstruction would be maximised for cases where only speech is present and slightly reduced for mixed cases (music with lyrics). In this case, stronger TRFs, which would be comparable to Event-Related Potential (ERP) [36], would be found for speech conditions and weaker TRFs would be seen for music with lyrics conditions. As this hypothesis assumes that the temporal cortical tracking of the envelope would be reduced for the music alone condition, the TRFs obtained for the instrumental music condition would be less defined (Figure D.1-C).

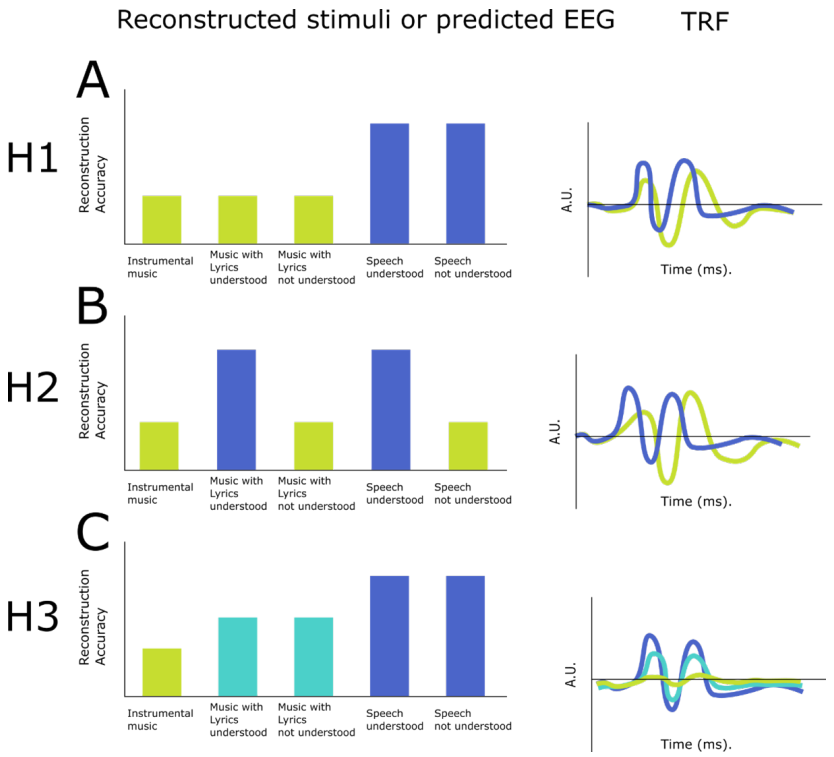

Fig. D.1: Putative patterns of reconstruction accuracy for both encoding and decoding model and encoding Temporal response functions (TRF) for each of the hypotheses tested. A- H1: Cortical tracking is driven mainly by speech-related acoustic factors. B- H2: Cortical tracking is mostly driven by linguistic/content factors. C- H3: Model is biased toward speech-specific features.

In the current study, we aim to test these three hypotheses in participants

listening to sounds while their brain activity is recorded using a 64-channel EEG system. EEG data and the audio signal envelope are used to train both encoding and decoding models, investigate potential differences between speech and music listening, and discuss the potential effects on the creation of versatile AAD. Five different types of sounds were used: instrumental music, music with understandable or not understandable lyrics (foreign language), and continuous speech, understandable or not. Variations of the reconstruction accuracies and/or distinct patterns of TRF are anticipated for each of the aforementioned hypotheses.

2 Methods

2.1 Participants

22 participants (7 males) aged between 20 and 45- years old (mean age: 25.71 ± 5.28) took part in the experiment. All participants were native Danish speakers, without any knowledge of the Finnish language. None of the participants had any known history of hearing loss or a neurological condition. Participants were compensated for their participation. EEG data were visually inspected and if more than a third of the trial was contaminated by large noise artefacts, those participants were excluded from further analysis. This led to the exclusion of 4 participants, which resulted in a final sample of 18 participants (5 males), with ages ranging from 20 to 32 years old (mean age: 24.63 ± 3.15) .

2.2 Stimuli & Procedure

Thirty excerpts of 70 seconds were presented to the participants, divided into five categories:

All musical signals could be considered as a single auditory object, as the organisation of the stimuli (melodic and harmonic relationships between instruments) can lead to a grouped perception of the music as one single and coherent auditory stream [37, 38].

The three musical conditions were created by taking three different versions of the same song: e.g. "How far I'll go", from Disney's movie Moana[1] in both the Danish version, the Finnish version and the instrumental version with the main melody recreated with a synthetic cello[2], to minimise the differences between conditions. For the speech conditions, a language unknown by the participant was chosen to compare natural speeches and maintain natural prosody in both conditions. Finnish was chosen because it is considerably different from Danish [39], which excludes potential content

[1]Moana (2016). Walt Disney Pictures.
[2]SoundKeys: CelloFan. http://soundkeysvst.free.fr/

Stimuli	Description	Speech alone	Music alone	Music with speech	Content understood by the listeners
Cello Music	Excerpts of polyphonic Disney songs with no lyrics. The melody line from the original version was replaced by a similar melody played by a synthetic cello.		X		No
Danish Music	Excerpts of polyphonic Disney songs with lyrics in Danish.			X	Yes
Finnish Music	Excerpts of polyphonic Disney songs with lyrics in Finnish.			X	No
Danish Speech	Excerpts of a Danish audiobook read by a woman.	X			Yes
Finnish Speech	Excerpts of a Finnish audiobook read by a woman.	X			No

recognition/understanding from the participants. For each condition, five different excerpts were used, resulting in 25 trials per participant. The normalised autocorrelation, as well as the power spectrum, of each stimulus can be found in Supplementary Material A and B, respectively. For each trial, excerpts were played on a single loudspeaker located in front of the participant. All stimuli were loudness matched to keep an equal perceived loudness of -23 LUFS (Loudness Unit Full Scale) across trials and correspond to 60 dBA at the listening position. The order of the stimuli was randomised across participants. For each trial, participants were instructed to actively listen and focus on the sound played for the whole 70 seconds. They were instructed to stay still, minimise body movements, and keep their gaze fixed on a cross on a screen placed in from of them. They had breaks in between each trial. At the end of each trial, participants answered two questions about their listening experience. The ratings for both questions were given on a continuous scale with labels placed 1.5 cm before and after the beginning and end of a 15 cm long scale, respectively.

- Q1: How familiar were you with the sound? (“Hvor bekendt er du med denne lyd?”) – From “Not familiar” to “Very familiar” (“Slet ikke bekendt” – “Meget bekendt”)
- Q2: How engaged were you in the listening? (“Hvor nærværende var du i lytteoplevelsen?”) - From “Not engaged ” to “Very engaged” (“Slet ikke nærværende”,”Meget nærværende”)

The choice of a monodic listening task (listeners are presented with only one stream of sound through one loudspeaker) is to minimie the effect of

attention to a different type of signal. Attention and engagement in listening may differ between speech and music listening and may be influenced even more by the distracter sound. While background music leads to minimum distraction during speech listening, the presence of distracting speech makes the listening and attentional task more challenging [24, 40]. Such variations might then be reflected in the cortical tracking of both target and distracter sounds [41]. Here, the listening task was limited to one stimulus at a time, to avoid variations of attention due to a secondary sound stream. In order to monitor potential differences in listening behaviour across types of sound, participants rated their engagement in the listening task for each trial. The engagement question was also used to test if engagement is correlated to cortical reconstruction accuracy.

The familiarity question was used as a screening question to assess if the participants were actively doing the task or just answering randomly to the questions. A high familiarity rating with the Finnish Speech trials would be an exclusion factor, as the probability that one is highly familiar with an audiobook in an unknown language is low and would indicate that the participant did not genuinely answer the question or did actually understand Finnish.

2.3 EEG data acquisition and pre-processing

Data collection was done in one single session for each participant. EEG data were recorded continuously at 2400 Hz with a g.HIamp-Research system [3]. 64 electrodes were placed on the scalp of participants following the international 10-10 system. Every electrode's impedance was kept at or below 5 kOhms.

All the pre-processing of the EEG data was carried out with EEGLAB v2021.1 [42]. Data pre-processing followed the recommendation for such experiment design, as described by [1]. Band-pass FIR filter was applied to filter the data between 0.5Hz and 45Hz, to remove low-frequency noise and 50Hz line noise. EEG data were re-referenced to the average of all scalp electrodes. The noise-contaminated EEG channels were visually examined and interpolated from nearby electrodes. Individual Component Analysis (ICA), with the automated detection plugin of EEGLAB [43], was used to remove artefacts caused by eye blinks and movements. Through visual inspection, the trials in which the artefacts were too large (e.g., movements) were discarded.

Following this initial preprocessing, the EEG data were downsampled to a sampling rate of 128 Hz after being lowpass filtered at 8 Hz (Butterworth filter, order=10). This lowpass cutoff frequency was selected as the cortical tracking of auditory stimuli is linked to cortical activities in the delta and theta band [1, 33, 44, 45] (that also correspond to the beat of the music present in the EEG signal [46, 47]). The highpass cutoff frequency was selected to

[3] g.tec Medical Engineering GmbH, Austria

remove low frequency noise in the EEG signal but to keep frequency where differences in reconstruction accuracy between speech and music have been observed [25]. Finally, the 5 first seconds and the last 5 seconds of the data were removed from each trial to avoid introducing the onset effects of the stimuli into the model, artefacts linked to the filtering, and to keep only data that correspond to ongoing continuous listening [1]. All the preprocessing (EEG and Audio) and the processing related to model training and testing was done using custom code.

2.4 Audio pre-processing

For all audio stimuli, the envelope of the audio signal was extracted, inspired by the approach described in [11]. The signal was divided into 32 sub-bands by a gammatone filter bank, with centre frequencies distributed between 50 and 20000 Hz. For each of these sub-bands, the envelope was extracted by calculating the absolute value of the Hilbert transform. An averaged envelope was obtained by averaging the extracted envelope of each sub-band. Finally, aligned to the preprocessing of the EEG data, the averaged envelope was filtered between 0.5 and 8 Hz (Butterworth filter, order=10), downsampled to a sampling rate of 128 Hz, and the 5 first and last seconds were removed to have both audio and EEG data aligned in time.

2.5 Backward modelling: stimulus reconstruction

The stimulus reconstruction also called backward modelling, aims to reconstruct an estimate S' of input stimuli S from cortical data R through a linear model W_b that behaves like a multi-channel filter [5, 7, 48, 49]. In the present study, the input stimuli S correspond to the envelope of the sound played to the participant, and the cortical data $R(t+\tau, n)$ corresponds to the EEG signal recorded by the electrode n at the time $(t-\tau)$ during the listening of the stimuli S. EEG recording is mapped to the stimulus envelope as follows:

$$S'(t) = \sum_{n} \sum_{\tau} W_b(\tau, n) R(t+\tau, n) \tag{D.1}$$

The estimation of the model W_b can be done with the minimisation of the mean-squared error between the reconstructed stimuli S' and the original stimuli S, which can be solved by the calculation of normalised reverse correlation, using ridge regularisation [50]:

$$W_b = (R^T R + I\lambda)^{-1} R^T S \tag{D.2}$$

Where I is the identity matrix, λ is the ridge regularisation parameter and S the original audio envelope for a single trial at time $t = 1...T$. R contains

several time-delayed versions of the EEG data collected during the trial over 64 electrodes [5, 48].

The time-delayed versions are included in the model to allow the model to capture the relevant time lags that correspond to cortical envelope tracking. In the present study, the stimulus reconstruction used a multi-delay model, which included several delayed versions of the EEG ranging from $\tau_min = 0$ to $\tau_max = 500ms$ for training and testing of the model. This range was chosen to ensure that optimal time lags for both speech and music reconstruction were included in the model [5, 23, 51].

$$R = \begin{bmatrix} r_1(1) & r_1(2) & \dots & r_1(\tau_{max}) & \dots & r_1(T) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & r_1(1) & \dots & r_1(T+\tau_max) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ r_n(1) & r_n(2) & \dots & r_n(\tau_{max}) & \dots & r_n(T) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & r_n(\tau_{max}) & \dots. & r_n(T+\tau_{max}) \end{bmatrix} \tag{D.3}$$

The ridge regularisation parameter λ was used to both enhance the regression and prevent overfitting [1, 7, 50]. In the present case, the value of λ was determined for each training data set through a cross-validation procedure, as described in [48]. The values explored for the parameter λ were logarithmically spaced from $10^(-10)$ to $10^(10)$.

The main information that can be extracted from such linear stimulus reconstruction is the reconstruction accuracy. It is measured by the Pearson correlation between a reconstructed stimulus and the corresponding original stimulus. The model itself, and its weights, cannot directly give interpretable information ([1, 52].

In the present study, condition-specific models were trained: models were trained and then tested on only one type of condition at a time. For example, for the Cello Music condition, models were trained using only trials where cello music was listened to and then tested on different cello music trials. A leave-one-out (LOO) approach was used for training the model by training a model on each trial minus one that is used for testing and averaging the resulting models. Due to the limited amount of data per participant and per condition (maximum 5 minutes for each participant and condition), a generic model design was chosen [1]. For each condition, all trials from all participants were pooled together before the LOO training.

2.6 Forward Modelling: EEG prediction and temporal response functions (TRF)

The forward modelling approach follows the same principle as the backward modelling, but instead of reconstructing the stimuli from the EEG, it attempts to predict the cortical response R' from the stimuli S through a linear model W_f [7, 53].

$$R'(t,n) = \sum_{\tau} W_f(\tau,n) S(t-\tau,n) \tag{D.4}$$

The forward model W_f can be calculated as follow:

$$W_f = (S^T S + I\lambda)^{-1} S^T R \tag{D.5}$$

with I and λ corresponding to the same elements as described in equation D.2. Here R corresponds to the preprocessed EEG recorded from one trial for time $t = 1...T$, and S contains several time-delayed versions of the audio envelope.

$$S = \begin{bmatrix} s(1-\tau min) & s(-\tau min) & \dots & s(1) & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots & s(1) & \dots & \vdots \\ \vdots & \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ \vdots & \vdots & \dots & \vdots & \vdots & \dots & s(1) \\ s(T) & \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ 0 & s(T) & \dots & \vdots & \vdots & \dots & \vdots \\ \vdots & \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & s(T) & S(T-1) & \dots & S(T-\tau max) \end{bmatrix} \tag{D.6}$$

With forward modelling, the prediction accuracy can be extracted from each channel, similarly to stimulus reconstruction. In addition, contrary to the backward modelling, the model itself can provide useful information as the weights describe the effect of the stimulus on the modulation of each EEG channel and are known as TRF. TRF corresponds to the estimated impulse response of the cortical response to the stimuli and can be compared to an event-related potential [1, 36, 48, 52].

The present study used forward models to compare TRFs between stimulus conditions. Forward models have been calculated by including delayed versions of the stimuli S ranging between $\tau_m in = -100$ and $\tau_m ax = 750ms$. For the forward model, the regularisation parameter is optimised through a

cross-validation procedure, similar to the backward model. The values explored for the parameter λ were logarithmically spaced values from $10^{(} - 10)$ to $10^{(}10)$. As for the backward model, a leave-one-out approach was used to train the forward models. The trained models are conditions-dependent (i.e., models are trained and tested only on trials from the same conditions), and subject-independent (i.e. for each condition, trials from all participants are used for the LOO approach).

3 Results

3.1 Engagement and familiarity rating

To assess the differences in rating for engagement between conditions, a two-way general linear mixed model ANOVA was used, with audio conditions as a fixed factor and participants as a random factor [54]. This test shows a significant effect of the audio condition on the rating of engagement (F(4,68.977)=25.806, p<0.001). A smaller effect of participants ($F(17, 68.527) = 6.178, p < 0.001$) and an interaction between audio-condition and participant ($F(68, 340) = 1.764, p = 0.001$) were found. Post-hoc comparisons with Bonferroni correction revealed that the " Finnish Speech " condition was rated significantly lower than all other conditions ($p < 0.001$). The Danish contents, both speech and music, were also rated significantly higher than the Finnish Music condition ($p < 0.001$). (See figure D.2). As explained in the method section, the results of the familiarity question were not further analysed. In addition, the Pearson's correlation between the engagement ratings and the reconstruction accuracy was calculated to test a potential link between the subjective and cortical measures. No significant correlation was found ($r = 0.005, p = 0.786$). A similar analysis investigated a potential link between the engagement rating and the EEG prediction accuracy. No significant correlation was found ($r = -0.042, p = 0.388$).

3.2 Backward model: stimulus reconstruction accuracy

Empirical chance level

The chance level for reconstruction was calculated to test if the linear backward model can robustly reconstruct the target signal. To calculate the chance level, the reconstruction accuracy was obtained through chance model, trained after shuffling audio EEG from all participants with irrelevant audio envelope (350 permutations). Chance levels were calculated independently for each condition. Wilcoxon rank sum test, with significance values adjusted through Bonferroni correction, showed a significant difference between the distribution

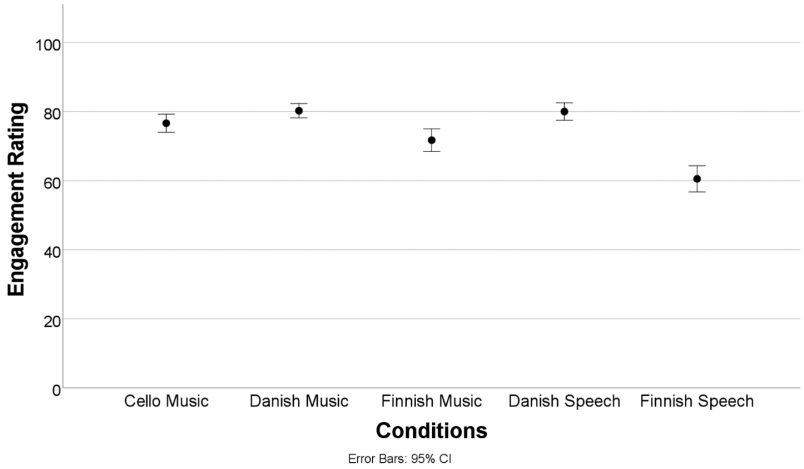

Fig. D.2: Engagement rating across stimuli conditions

of reconstruction accuracy and the distribution of chance level for all stimuli conditions ($p < 0.001$) D.3).

Comparison across conditions

A two-way general linear mixed model ANOVA was used to compare the reconstruction accuracy across audio conditions, with conditions as a fixed factor and participants as a random factor. The results show no significant effect of audio conditions, i.e., the type of sound being listened to by the participants, on the reconstruction accuracy ($F(4, 69.014) = 1.695, p = .161$). A significant effect of participant ($F(17, 68.547) = 1.848, p = 0.039$), and a significant interaction between participants and audio-condition ($F(68, 340) = 1.7, p = 0.001$) were found.

3.3 Forward model: EEG prediction accuracy

Empirical chance level

The chance level for prediction accuracy was computed to determine whether the linear forward model can successfully reconstruct the EEG signal. To calculate the chance level, the prediction accuracy was obtained through chance model, trained after shuffling audio EEG from all participants with irrelevant audio envelope (350 permutations). Chance levels were compared independently for each condition . The distribution of reconstruction accuracy and the distribution of chance level for all stimulation conditions were significantly

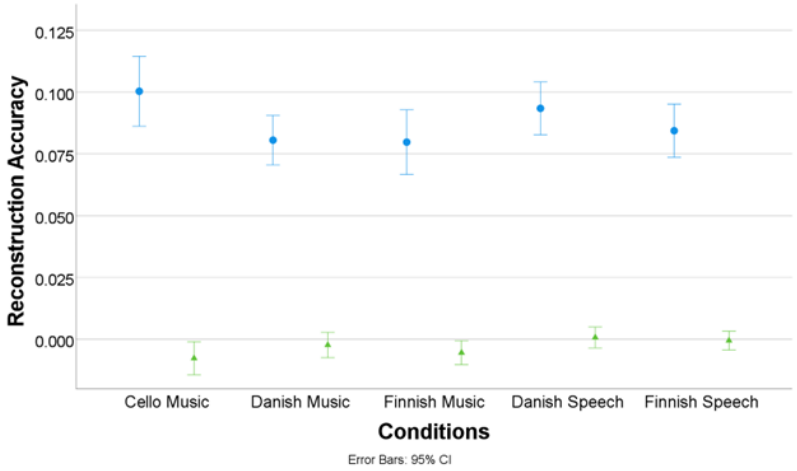

Fig. D.3: Stimulus reconstruction accuracy obtained with a backward model by conditions. Mean and CI are calculated from results from all relevant trials and all participants. Chance level is indicated for each condition by the green triangles.

different, according to the Wilcoxon rank sum test, with significance values corrected by Bonferroni method ($p < 0.001$) (Figure D.4).

Comparison across conditions

To compare the EEG prediction accuracy across conditions, a two-way general linear mixed model ANOVA with participants as a random factor and conditions as a fixed factor was used. The effect of EEG channels was not investigated. Instead, the prediction accuracy was calculated by taking into account prediction values from all 64 EEG channels. The findings demonstrate no significant effect of the stimuli condition on the reconstruction accuracy ($F(4, 69.229) = 1.150, p = 0.340$), and no significant effect of participant ($F(17, 68.684) = 1,529, p = 0.111$). A small but significant interaction between conditions and participants was found ($F(68, 340) = 1.360, p = 0.042$).

3.4 Forward model: Temporal Response Functions

The results obtained for both the stimulus reconstruction accuracy and the EEG prediction accuracy suggest that both forward and backward modelling perform similarly for all audio conditions. The TRFs (i.e., the weights of the forward model) were then examined to investigate if differences could be observed in the model across conditions. As artefacts are common at the beginning and end of the TRFs samples, all TRFs were trimmed by 50 ms at

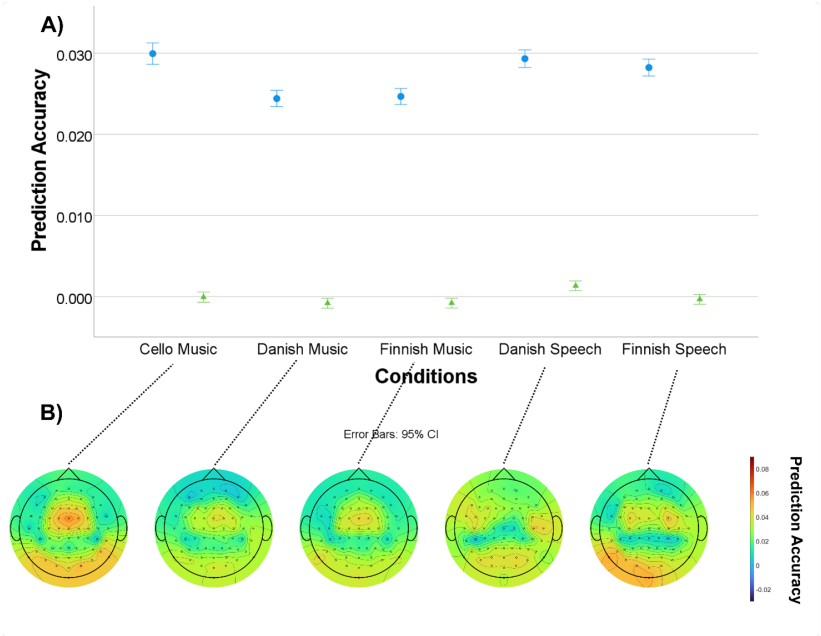

Fig. D.4: A) EEG prediction accuracy obtained with a forward model by stimuli conditions, including all 64 EEG channels. Mean, and CI are calculated from results from all relevant trials and all participants. The chance level is indicated for each condition by the green triangles. – B) Mean prediction accuracy at each electrode, across conditions.

both ends. First, the Global Field Power (GFP) [55] of the 64-channel TRFs was calculated at each time point and examined to determine the timing where TRF components can be observed. All TRF calculated through the LOO procedure in place for calculating the EEG prediction accuracy were used for calculating the GFP. Figure D.5-A) displays the GFP for each condition. From those GFP profiles, several observations can be made. Firstly, a clear component can be observed for both speech and music-with-lyrics conditions, around 200ms. It can also be noted that similarity in the GFP profile can be seen for both speech conditions. For the music-with-lyrics conditions, the profile of the Danish and Finnish Music conditions also show a similar pattern, despite a peak component around 400ms that is only present for the Danish Music condition. However, these two groups differ from each other:

- A TRF component can be observed at $\approx 60ms$ for the Danish Music and Finnish Music conditions and is weaker or absent in the speech GFP profiles.
- For the speech conditions (Danish Speech & Finnish Speech), an early

peak (around 0ms) can be observed. This component is not observable in the other conditions. However, due to the temporal placement of this peak and its amplitude compared to the $\approx 200ms$ component, this early peak should be considered cautiously, as it is likely to correspond to an artefact due to the modelling procedure.

- While the $\approx 200ms$ component is present for speech conditions and music-with-speech conditions, the shape varies between those two groups,

Finally, concerning the instrumental music condition (Cello Music), the profile of the GFP differs from the other conditions, and no specific component can be extracted from the GFP profile, indicating weaker, less robust TRFs. Figure D.5-C) shows the topographic map of the averaged weight of TRFs per condition for each peak identified in the GFP profiles. The topography of all peaks previously identified suggests that the weights are more robust in the frontal-central and occipital regions for all the components observed. A slight divergence from this pattern can be observed for the Danish Speech conditions at the 200ms component, where the maximum weight becomes more lateral and symmetric in the frontal region.

Overall, from the observation of the different TRFs, while the spatial pattern is comparable across conditions, three distinct temporal patterns can be observed: one comprises the speech conditions, one comprised of the music-with-speech conditions and one different for the music instrumental condition. Those differences may indicate distinct underlying cortical mechanisms involved in listening to speech, music with lyrics, or instrumental music.

4 Discussion

In this study, participants performed a listening task where they were instructed to listen to either music or speech actively. The musical conditions were separated into three sub-conditions: instrumental music, music with lyrics in Danish, understood by the participants; music with lyrics in Finnish, not understood by the participants. For speech listening, they listened to excerpts of an audiobook in Danish or an excerpt of an audiobook in Finnish, with comprehension varying accordingly. EEG data were recorded during the listening sessions, and linear models were trained to examine differences between speech and music listening. Based on results from previous research, it was hypothesised that speech would be better reconstructed via the backward decoding model than music [23–25]. Similarly, it was hypothesised that a similar bias toward speech would be found through forward linear modelling by investigating the EEG prediction accuracy and the weights of the Temporal Response Function computed.

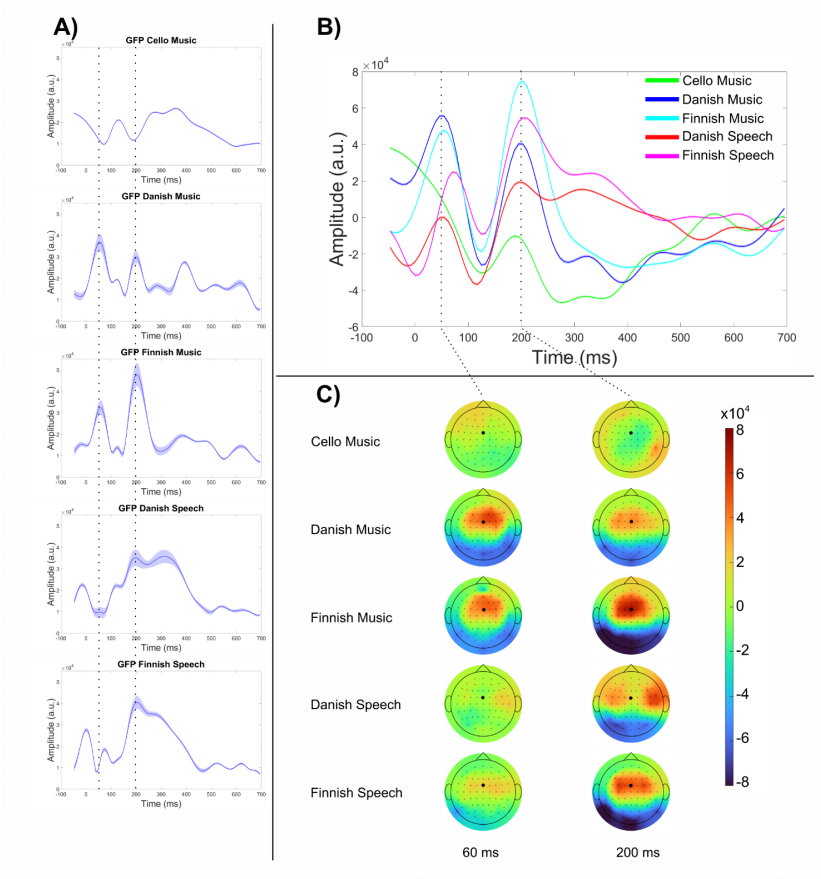

Fig. D.5: A) Mean global field power measured at each time lag averaged across TRFs calculated through a LOO approach, for each condition, shaded areas correspond to standard deviation.- B) Grand average TRFs at FCz, per conditions. Shaded areas correspond to standard deviation. - C) Scalp topographies of the dominant TRF components occurring at ≈ 60 and $\approx 200ms$ for each condition. The black dot represents the position of the electrode FCz, used for illustrating the TRFs.

With the assumption that there are differences in the neural tracking of the temporal envelope between speech and music listening, this study aimed to investigate more precisely the factor contributing to such differences. To that aim, three hypotheses have been formulated and tested.

Two main performance measurements have been used to test the hypothesis. First, the stimulus reconstruction accuracy was calculated using the Pearson correlation between a reconstructed audio envelope through a backward linear model and the original audio envelope. For this measure, no significant

differences have been found between the conditions under test. The second measure presented here is the EEG prediction accuracy, corresponding to the correlation between the EEG signal and predicted EEG through a forward model. Across conditions, the values obtained for EEG prediction accuracy were not significantly different. The lack of differences indicates comparable decoding/encoding performance for music listening and speech listening when using envelope-based linear modelling. These results contradict the experiment's initial assumption, leading to the rejection of all three formulated hypotheses.

The first hypothesis was that a significantly higher neural tracking would be observed for speech conditions than music due to enhanced tracking of speech-specific acoustic processing. As speech and music can be reconstructed or predicted to a comparable extent, this hypothesis must be rejected, despite the acoustic differences that exist between speech and music at the envelope level (See figures D.6 & D.7). The second hypothesis, that semantic processing led to higher neural tracking, is also rejected. Compared to the other conditions, no significant differences have been found between conditions containing understandable linguistic content (Danish Speech and Danish Music, understood by the listeners). It indicates that linguistic content processing does not enhance cortical envelope tracking of audio stimuli. The third hypothesis was related to potential bias in the model design that would enhance reconstruction/prediction accuracies. This hypothesis is rejected as the presence of speech in the original stimulus does not impact the models' performances. It has to be noted that, despite the lack of significant effect of the condition, significant interactions between the condition and participant have been found for those two accuracy measures. It indicates that, at the individual level, the reconstruction/prediction performance may differ between speech and music listening. However, it varies across listeners; no effect is generalised across participants. Due to the current study's design, generic linear models have been trained by including data from several participants in the model's training. It would be of interest to also investigate further the individual differences by training individual models, fitted to only one participant at a time, to explore if the similarities in reconstruction/prediction performance remain and to investigate more how individual differences across listeners may influence the neural tracking of the audio envelope.

Overall, the current results contradict the initial assumption and contrast with previous results, which showed increased reconstruction accuracy for speech compared to music listening [23–25]. The experimental design of this study differs from the one previously used: two of the aforementioned studies were designed to study auditory attention decoding; hence several sounds were concurrently presented to the listener [23, 24]. In that case, the differences in reconstruction may be due to attentional factors that may differ between speech and music listening. In the present study, an affective measurement

(i.e., engagement rating) has been collected in addition to the cortical measures. This measure was added to the experiment design to explore a potential link between listening engagement and the cortical tracking of the temporal variations of the sound. While some differences have been found across conditions for the engagement rating, they were not correlated with the reconstruction or EEG prediction accuracy. It suggests that cortical tracking is not directly related to the level of engagement of the listener. Further research would be needed to explore how attentional or other cognitive factors may influence the neural tracking of different auditory stimuli.

Besides the model's performance, the weights of the model trained in this study were also investigated. As the weights of the backward model are not straightforwardly interpretable [52], only the weight of the forward models has been explored. Observation of the TRFs components indicates differences across conditions, conversely to the reconstruction/prediction performances. Apart from the instrumental music condition, the forward models comprise robust TRFs. In the case of instrumental music, the obtained TRFs are not as robust, with less defined components. However, it does not affect the reconstruction accuracy obtained with this model, which performs comparably to the other conditions. For both speech conditions and both music conditions that contain lyrics, robust TRFs can be observed, and some similarity can be observed: the presence of a component around 200ms and comparable topographic patterns. This topographic pattern, as well as the 200ms component, are consistent with previous studies on envelope-based TRFs on speech [56, 57] and music [19]. However, some differences can also be noticed, with a component present for the music-with-lyrics conditions but not for the speech conditions and shape differences between music and speech at the 200ms component.

Overall, three different patterns of TRFs one for speech, one for music with lyrics and one for instrumental music. While it does not directly fit one of the initial hypotheses presented in this study, such differences could indicate distinct underlying cortical processing for those three conditions. However, surprisingly, such a grouping of the initial condition into three categories is not observed in the reconstruction accuracy results. It suggests that those putative distinct cortical processes of the speech, music or music with speech stimuli may lead to comparable cortical tracking of the temporal variations of the heard sound.

The present results are promising for implementing envelope-based linear models for applications with music or speech, such as Auditory Attention Detection. Comparable reconstruction accuracy across those diverse audio signal types suggests that such a model is not inherently biased toward speech and that a similar model design could be used for speech and music listening. Nevertheless, although the present results suggest that the neural tracking of the temporal envelope is comparable for speech (understood or not), instrumental

music and music with lyrics, further research would be needed to explore the factor that could influence the differences in reconstruction accuracy that have been observed in other studies.

5 Conclusion

This study explores the differences in performance in cortical linear encoding and decoding during naturalistic speech and music listening. The study explored three hypotheses that could explain differences in reconstruction accuracies between speech and music stimuli, and the obtained results led to the rejection of all three hypotheses. Moreover, contrary to previous results, no significant differences in reconstruction accuracy between music and speech listening were found. It can be seen as a promising result for the future development of BCI based on stimuli and/or EEG signal reconstruction, as the present study shows that encoding and decoding linear models based on the envelope can be used for both music, speech and music stimuli that include speech with similar performances. Exploration of the weight of the encoding model (TRFs), however, suggest that listening to those different type of sound might differ in their underlying cortical processes, even though those differences do not explain the variation in reconstruction accuracies. Future research should continue to explore such differences further.

References

[1] M. J. Crosse, N. J. Zuk, G. M. Di Liberto, A. R. Nidiffer, S. Molholm, and E. C. Lalor, "Linear Modeling of Neurophysiological Responses to Speech and Other Continuous Stimuli: Methodological Considerations for Applied Research," *Frontiers in Neuroscience*, vol. 15, no. November, nov 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2021.705621/full

[2] C. R. Holdgraf, J. W. Rieger, C. Micheli, S. Martin, R. T. Knight, and F. E. Theunissen, "Encoding and Decoding Models in Cognitive Electrophysiology," *Frontiers in Systems Neuroscience*, vol. 11, no. September, sep 2017. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fnsys.2017.00061/full

[3] J.-R. King, L. Gwilliams, C. Holdgraf, J. Sassenhagen, A. Barachant, D. Engemann, E. Larson, and A. Gramfort, "Encoding and decoding neuronal dynamics: Methodological framework to uncover the algorithms of cognition," *HAL*, no. ffhal-01848442f, 2018.

[4] G. M. Di Liberto, J. A. O'Sullivan, and E. C. Lalor, "Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing," *Current Biology*, vol. 25, no. 19, pp. 2457–2465, oct 2015. [Online]. Available: http://dx.doi.org/10.1016/j.cub.2015.08.030https://linkinghub.elsevier.com/retrieve/pii/S0960982215010015

[5] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG," *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, 01 2015. [Online]. Available: https://doi.org/10.1093/cercor/bht355

[6] L. S. Hamilton and A. G. Huth, "The revolution will not be controlled: natural stimuli in speech neuroscience," *Language, Cognition and Neuroscience*, vol. 35, no. 5, pp. 573–582, jun 2020. [Online]. Available: https://doi.org/10.1080/23273798.2018.1499946https://www.tandfonline.com/doi/full/10.1080/23273798.2018.1499946

[7] E. Alickovic, T. Lunner, F. Gustafsson, and L. Ljung, "A Tutorial on Auditory Attention Identification Methods," *Frontiers in Neuroscience*, vol. 13, no. March, pp. 1–17, mar 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2019.00153/full

[8] J. Belo, M. Clerc, and D. Schön, "EEG-Based Auditory Attention Detection and Its Possible Future Applications for Passive BCI," *Frontiers in Computer Science*, vol. 3, p. 661178, apr 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcomp.2021.661178/full

[9] S. Geirnaert, S. Vandecappelle, E. Alickovic, A. de Cheveigne, E. Lalor, B. T. Meyer, S. Miran, T. Francart, and A. Bertrand, "Electroencephalography-Based Auditory Attention Decoding: Toward Neurosteered Hearing Devices," *IEEE Signal Processing Magazine*, vol. 38, no. 4, pp. 89–102, jul 2021. [Online]. Available: http://arxiv.org/abs/2008.04569https://ieeexplore.ieee.org/document/9467380/

[10] A. Aroudi, B. Mirkovic, M. De Vos, and S. Doclo, "Impact of Different Acoustic Components on EEG-Based Auditory Attention Decoding in Noisy and Reverberant Conditions," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 4, pp. 652–663, apr 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8662636/

[11] W. Biesmans, N. Das, T. Francart, and A. Bertrand, "Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario." *IEEE Trans. Neural Syst.*, vol. 25, no. 5, pp. 402–412, 2016.

[12] S. A. Fuglsang, T. Dau, and J. Hjortkjær, "Noise-robust cortical tracking of attended speech in real-world acoustic scenes," *NeuroImage*, vol. 156, no. April, pp. 435–444, aug 2017. [Online]. Available: http://dx.doi.org/10.1016/j.neuroimage.2017.04.026https://linkinghub.elsevier.com/retrieve/pii/S105381191730318X

[13] B. Mirkovic, S. Debener, M. Jaeger, and M. De Vos, "Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications," *Journal of Neural Engineering*, vol. 12, no. 4, p. 046007, aug 2015. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2560/12/4/046007

[14] B. Mirkovic, M. G. Bleichner, M. De Vos, and S. Debener, "Target Speaker Detection with Concealed EEG Around the Ear," *Frontiers in Neuroscience*, vol. 10, no. JUL, pp. 1–11, jul 2016. [Online]. Available: http://journal.frontiersin.org/Article/10.3389/fnins.2016.00349/abstract

[15] IFPI, "International federation of the phonographic industry engaging with music report 2021," 2021. [Online]. Available: https://www.ifpi.org/wp-content/uploads/2021/10/IFPI-Engaging-with-Music-report.pdf

[16] W. W. An, B. Shinn-Cunningham, H. Gamper, D. Emmanouilidou, D. Johnston, M. Jalobeanu, E. Cutrell, A. Wilson, K.-J. Chiang, and I. Tashev, "Decoding Music Attention from "EEG Headphones": A User-Friendly Auditory Brain-Computer Interface," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, IEEE. IEEE, jun 2021, pp. 985–989. [Online]. Available: https://ieeexplore.ieee.org/document/9414492/

[17] J. Belo, M. Clerc, and D. Schon, "Attentional inhibition ability predicts neural representation during challenging auditory streaming," *bioRxiv*, 2022.

[18] G. Cantisani, S. Essid, and G. Richard, "EEG-Based Decoding of Auditory Attention to a Target Instrument in Polyphonic Music," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, vol. 2019-Octob, no. 765068. IEEE, oct 2019, pp. 80–84. [Online]. Available: https://ieeexplore.ieee.org/document/8937219/

[19] G. M. Di Liberto, C. Pelofi, S. Shamma, and A. de Cheveigné, "Musical expertise enhances the cortical tracking of the acoustic envelope during naturalistic music listening," *Acoustical Science and Technology*, vol. 41, no. 1, pp. 361–364, jan 2020. [Online]. Available: https://www.jstage.jst.go.jp/article/ast/41/1/41{_}E19257/{_}article

[20] G. M. Di Liberto, G. Marion, and S. A. Shamma, "Accurate Decoding of Imagined and Heard Melodies," *Frontiers in Neuroscience*, vol. 15, no. August, pp. 1–11, aug 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2021.673401/full

[21] O. Etard, R. B. Messaoud, G. Gaugain, and T. Reichenbach, "No evidence of attentional modulation of the neural response to the temporal fine structure of continuous musical pieces," *Journal of Cognitive Neuroscience*, vol. 34, no. 3, pp. 411–424, 2022.

[22] L. Hausfeld, N. R. Disbergen, G. Valente, R. J. Zatorre, and E. Formisano, "Modulating Cortical Instrument Representations During Auditory Stream Segregation and Integration With Polyphonic Music," *Frontiers in Neuroscience*, vol. 15, no. September, pp. 1–15, sep 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2021.635937/full

[23] L. Hausfeld, L. Riecke, G. Valente, and E. Formisano, "Cortical tracking of multiple streams outside the focus of attention in naturalistic auditory scenes," *NeuroImage*, vol. 181, no. March, pp. 617–626, nov 2018. [Online]. Available: https://doi.org/10.1016/j.neuroimage.2018.07.052https://linkinghub.elsevier.com/retrieve/pii/S1053811918306670

[24] A. Simon, G. Loquet, J. Østergaard, and S. Bech, "Cortical Auditory Attention Decoding During Music And Speech Listening," *(In review)*, 2023.

[25] N. J. Zuk, J. W. Murphy, R. B. Reilly, and E. C. Lalor, "Envelope reconstruction of speech and music highlights stronger tracking of speech at low frequencies," *PLOS Computational Biology*, vol. 17, no. 9, p. e1009358, sep 2021. [Online]. Available: http://dx.doi.org/10.1371/journal.pcbi.1009358https://dx.plos.org/10.1371/journal.pcbi.1009358

[26] N. Ding and J. Z. Simon, "Emergence of neural encoding of auditory objects while listening to competing speakers." *Proc. Natl. Acad. Sci. U.S.A.*, vol. 109, no. 29, pp. 11 854–11 859, jul 2012. [Online]. Available: https://pnas.org/doi/full/10.1073/pnas.1205381109

[27] ——, "Adaptive temporal encoding leads to a background-insensitive cortical representation of speech," *Journal of Neuroscience*, vol. 33, no. 13, pp. 5728–5735, 2013.

[28] E. Verschueren, J. Vanthornhout, and T. Francart, "The effect of stimulus choice on an eeg-based objective measure of speech intelligibility," *Ear and hearing*, vol. 41, no. 6, pp. 1586–1597, 2020.

[29] N. Ding, A. D. Patel, L. Chen, H. Butler, C. Luo, and D. Poeppel, "Temporal modulations in speech and music," *Neuroscience & Biobehavioral Reviews*, vol. 81, pp. 181–187, oct 2017. [Online]. Available: https://doi.org/10.1016/j.neubiorev.2017.02.011https://linkinghub.elsevier.com/retrieve/pii/S0149763416305668

[30] K. D. Prinsloo and E. C. Lalor, "General auditory and speech-specific contributions to cortical envelope tracking revealed using auditory chimeras," *The Journal of Neuroscience*, vol. 42, no. 41, pp. JN–RM–2735–20, aug 2022. [Online]. Available: https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.2735-20.2022

[31] I. Iotzov and L. C. Parra, "Eeg can predict speech intelligibility," *Journal of Neural Engineering*, vol. 16, no. 3, p. 036008, 2019.

[32] J. Vanthornhout, L. Decruy, J. Wouters, J. Z. Simon, and T. Francart, "Speech intelligibility predicted from neural entrainment of the speech envelope," *Journal of the Association for Research in Otolaryngology*, vol. 19, no. 2, pp. 181–191, 2018.

[33] O. Etard and T. Reichenbach, "Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise," *Journal of Neuroscience*, vol. 39, no. 29, pp. 5750–5759, 2019.

[34] G. M. Di Liberto, J. Nie, J. Yeaton, B. Khalighinejad, S. A. Shamma, and N. Mesgarani, "Neural representation of linguistic feature hierarchy reflects second-language proficiency," *Neuroimage*, vol. 227, p. 117586, 2021.

[35] E. Verschueren, M. Gillis, L. Decruy, J. Vanthornhout, and T. Francart, "Speech understanding oppositely affects acoustic and linguistic neural tracking in a speech rate manipulation paradigm," *bioRxiv*, 2022.

[36] E. C. Lalor, A. J. Power, R. B. Reilly, and J. J. Foxe, "Resolving Precise Temporal Processing Properties of the Auditory System Using Continuous Stimuli," *Journal of Neurophysiology*, vol. 102, no. 1, pp. 349–359, jul 2009. [Online]. Available: https://www.physiology.org/doi/10.1152/jn.90896.2008

[37] A. H. Gregory, "Listening to polyphonic music," *Psychology of Music*, vol. 18, no. 2, pp. 163–170, 1990.

[38] E. Bigand, S. McAdams, and S. Forêt, "Divided attention in music," *International Journal of Psychology*, vol. 35, no. 6, pp. 270–278, dec 2000. [Online]. Available: http://doi.wiley.com/10.1080/002075900750047987

[39] D. Benedetto, E. Caglioti, and V. Loreto, "Language trees and zipping," *Physical Review Letters*, vol. 88, no. 4, p. 048702, 2002.

[40] Y.-N. Shih, R.-H. Huang, and H.-Y. Chiang, "Background music: Effects on attention performance," *Work*, vol. 42, no. 4, pp. 573–578, 2012.

[41] J. A. Brown and G. M. Bidelman, "Familiarity of background music modulates the cortical tracking of target speech at the "cocktail party"," *Brain Sciences*, vol. 12, no. 10, p. 1320, 2022.

[42] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis." *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, 2004.

[43] L. Pion-Tonachini, K. Kreutz-Delgado, and S. Makeig, "Iclabel: An automated electroencephalographic independent component classifier, dataset, and website." *NeuroImage*, vol. 198, pp. 181–197, 2019.

[44] B. Meltzer, C. S. Reichenbach, C. Braiman, N. D. Schiff, A. Hudspeth, and T. Reichenbach, "The steady-state response of the cerebral cortex to the beat of music reflects both the comprehension of music and attention," *Frontiers in human neuroscience*, vol. 9, p. 436, 2015.

[45] Y. Zhu, C. Zhang, H. Poikonen, P. Toiviainen, M. Huotilainen, K. Mathiak, T. Ristaniemi, and F. Cong, "Exploring frequency-dependent brain networks from ongoing eeg using spatial ica during music listening," *Brain Topography*, vol. 33, no. 3, pp. 289–302, 2020.

[46] F. Cong, V. Alluri, A. K. Nandi, P. Toiviainen, R. Fa, B. Abu-Jamous, L. Gong, B. G. Craenen, H. Poikonen, M. Huotilainen *et al.*, "Linking brain responses to naturalistic music through analysis of ongoing eeg and stimulus features," *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1060–1069, 2013.

[47] S. Nozaradan, I. Peretz, and A. Mouraux, "Selective neuronal entrainment to the beat and meter embedded in a musical rhythm," *Journal of Neuroscience*, vol. 32, no. 49, pp. 17 572–17 581, 2012.

[48] M. J. Crosse, G. M. Di Liberto, A. Bednar, and E. C. Lalor, "The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli," *Frontiers in Human Neuroscience*, vol. 10, no. NOV2016, pp. 1–14, nov 2016. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fnhum.2016.00604/full

[49] P. J. Schäfer, F. I. Corona-Strauss, R. Hannemann, S. A. Hillyard, and D. J. Strauss, "Testing the Limits of the Stimulus Reconstruction Approach: Auditory Attention Decoding in a Four-Speaker Free Field

Environment," *Trends in Hearing*, vol. 22, p. 233121651881660, jan 2018. [Online]. Available: http://journals.sagepub.com/doi/10.1177/2331216518816600

[50] D. D. E. Wong, S. A. Fuglsang, J. Hjortkjær, E. Ceolini, M. Slaney, and A. de Cheveigné, "A Comparison of Regularization Methods in Forward and Backward Models for Auditory Attention Decoding," *Frontiers in Neuroscience*, vol. 12, no. AUG, pp. 1–16, aug 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2018.00531/full

[51] A. Simon, J. Østergaard, S. Bech, and G. Loquet, "Optimal time lags for linear cortical auditory attention detection : differences between speech and music listening," *19th International Symposium on Hearing 19-24 June 2022, Lyon, France*, no. June, pp. 2–5, 2022.

[52] S. Haufe, F. Meinecke, K. Görgen, S. Dähne, J.-D. Haynes, B. Blankertz, and F. Bießmann, "On the interpretation of weight vectors of linear models in multivariate neuroimaging," *NeuroImage*, vol. 87, pp. 96–110, feb 2014. [Online]. Available: http://dx.doi.org/10.1016/j.neuroimage.2013.10.067https://linkinghub.elsevier.com/retrieve/pii/S1053811913010914

[53] N. Ding and J. Z. Simon, "Neural coding of continuous speech in auditory cortex during monaural and dichotic listening," *Journal of Neurophysiology*, vol. 107, no. 1, pp. 78–89, jan 2012. [Online]. Available: https://www.physiology.org/doi/10.1152/jn.00297.2011

[54] T. Næs, P. B. Brockhoff, and O. Tomic, *Statistics for sensory and consumer science*. John Wiley & Sons, 2011.

[55] M. M. Murray, D. Brunet, and C. M. Michel, "Topographic erp analyses: a step-by-step tutorial review," *Brain topography*, vol. 20, no. 4, pp. 249–264, 2008.

[56] M. P. Broderick, N. J. Zuk, A. J. Anderson, and E. C. Lalor, "More than words: Neurophysiological correlates of semantic dissimilarity depend on comprehension of the speech narrative," *European Journal of Neuroscience*, vol. 56, no. 8, pp. 5201–5214, 2022.

[57] E. Verschueren, J. Vanthornhout, and T. Francart, "The effect of stimulus intensity on neural envelope tracking," *Hearing Research*, vol. 403, p. 108175, 2021.

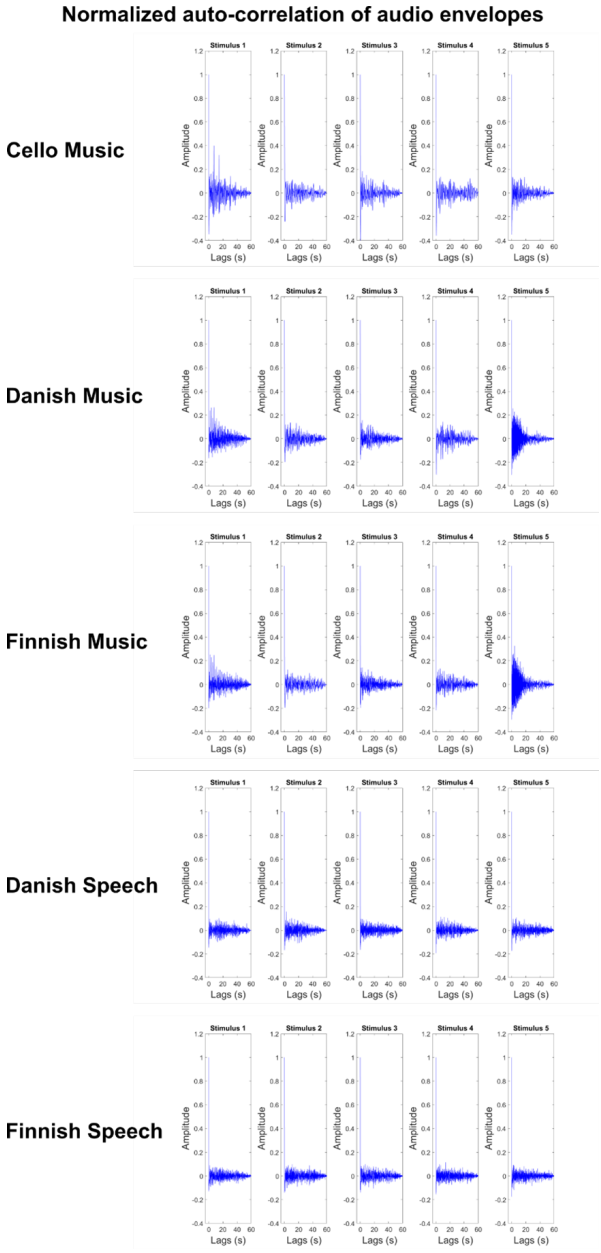

Fig. D.6: Normalised autocorrelation of the envelopes of each sound stimuli used in the experiment.

6 Supplementary Material

6.1 Supplementary Material A:

6.2 Supplementary Material B:

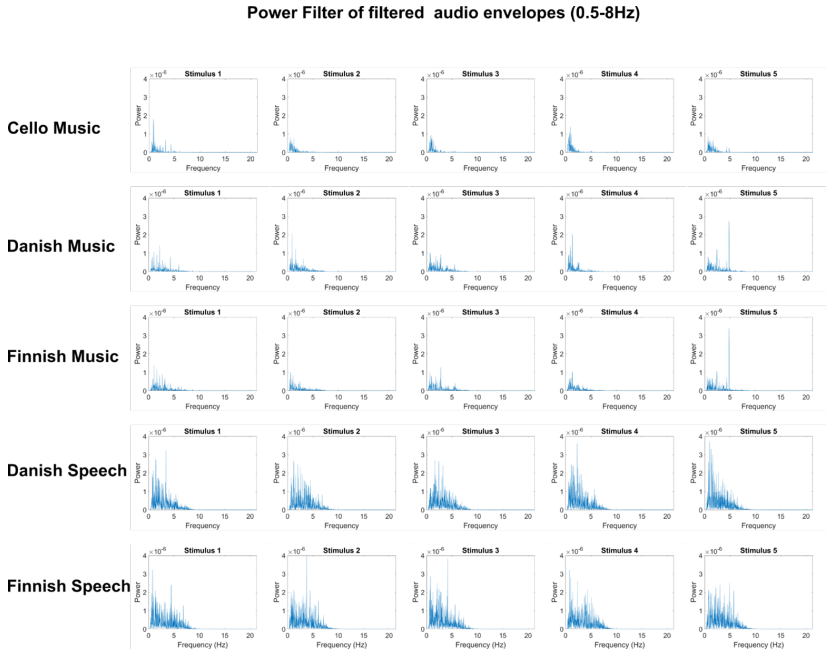

Fig. D.7: Power spectrum of the envelopes of each sound stimuli used in the experiment, after bandpass filtering (0.5-8Hz).

ISSN (online): 2446-1628
ISBN (online): 978-87-7573-682-9

AALBORG UNIVERSITY PRESS