Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Investigating the neurophysiology of auditory attention in multi-talker scenarios: from sustained attention to attention switching

Sara Carta

Academic supervisors:
Dr. Giovanni Di Liberto, Dr. Alejandro Lopez Valdes
Industry supervisors:
Dr. Emina Aličković, Dr. Johannes Zaar

A dissertation submitted to the
Trinity College Dublin, the University of Dublin
In fulfilment of the requirements for the degree of
Doctor of Philosophy
Academic Year 2026

School of Computer Science and Statistics
Trinity College Dublin, the University of Dublin

Declaration

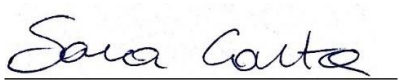
I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I confirm that the work has been completed in full compliance with relevant university policies, including the Academic Integrity policy, the Trinity Policy on Good Research Practice, and the guidance on AI and GenAI in Teaching, Learning, Assessment and Research.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

Signed,



Sara Carta

September 30th, 2025

Summary

Despite the complexity of our auditory world and the limitations of listeners' cognitive and perceptual resources, the human ability to use and comprehend speech is remarkably effective. Its robustness makes it possible to sustain attention on a person of interest among competing speakers, to redirect the focus of attention at will, and to rapidly reorient attention on unexpected auditory events. As such, speech processing abilities are tightly intertwined with the ability to maintain and shift attention.

The neural cortical mechanisms underpinning attention to speech have been explored with invasive and non-invasive technologies, establishing the existence of a selective enhanced representation of the speaker of interest (target) in cortical signals, along with a weaker encoding of interfering speakers outside the focus of attention (maskers). However, several questions remain unanswered, especially regarding the limits of the attentional filter on different levels of speech information (from acoustics to abstract meaning), and the ability to flexibly reorient the focus of attention on multiple speakers alternatively. Furthermore, there is even less clarity on how those cortical mechanisms may operate across ages and hearing abilities.

In this thesis, I investigate the neural basis of attention to speech across different populations and attention scenarios. Electroencephalography (EEG) recordings of the listeners' brain signals were related to various speech features through methods based on lagged linear regression, yielding measures of neural tracking of competing speakers in multi-talker scenes characterised by different attentional demands.

Firstly, this work offers novel insights into the neural encoding of competing talkers in listeners with hearing impairment, focussing on the level of phonological information. The study presented in Chapter 3 indicates that neural tracking of speech phonological information is increased for the target relative to the masker speaker, although phoneme onsets of the masker are more neurally represented than those of the target. This mechanism may represent a neural basis for the increased distractibility in multi-talker scenes in listeners with hearing impairment.

The studies presented in Chapter 4 and Chapter 5, on the other hand, investigate the neural underpinnings of flexible attention reorienting between two competing speakers, with a multi-talker scenario requiring the listener to perform frequent switches of attention. Through this dynamic attention paradigm, the experiment in Chapter 4 offers a novel framework to interpret the attention switch in terms of the disengagement and engagement of attention

between competing speakers, exploring their temporal dynamics and establishing a transient simultaneous tracking of the two streams while a switch of attention unfolds. That study also finds a robust reduction in brain activity within the alpha frequency band (8-12 Hz) after the switching cue, and that the trough of the alpha decrement follows the neural attention switch. Furthermore, we investigate the impact of previous context on the neural tracking of the current speech block, comparing four alternative models on their ability to explain the EEG data, and revealing that a model that resets the context at every switching cue has the highest fit to the neural signals.

Finally, in Chapter 5, we extend our attention switching paradigm by including frequent changes of topic within the same stream after a switch of attention, challenging the reliance on previous context, and we test in on two cohorts of younger and older listeners. The results suggest that neural tracking of speech is highly impacted by attention switching cues, decreasing over the course of each trial, in particular for older listeners. Neural responses to the speech acoustics are also impacted by age, with differences in the latency and amplitude of N1 and P2 components. Furthermore, the attention switching cue introduces a P300-like detection response, which also displays delayed temporal dynamics in older listeners. Similarly to the study in Chapter 4, also here an alpha activity decrement follows the attention switching cue, with similar dynamics for younger and older listeners.

Acknowledgements

This work would have not been possible without the support of many people, whose contribution I would like to acknowledge:

Giovanni, thank you for being an incredible mentor. For teaching me the importance to think deeply about an idea and ask the right questions, to pause and ponder, when necessary, and to move forward, when it is time. Your trust in me, passion for science, and your ability to always find the bright side in any circumstance have truly helped me throughout this journey. Thanks for all the fun chats, the coaster-flipping games, the Wicklow hikes, the lab lunches and, last but not least, thank you for teaching me to always – always – increase the font size.

Alejandro, thank you for your amazing guidance over these four years. For teaching me the value of experimental and technical rigour, and for patiently explaining the details of audio setups and sound calibrations to a non-engineer like me. No dataset would have been collected without your *ghostbusting* support through temperamental soundcards, unpredictable loudspeakers and mysterious electrical interferences. Thanks for the fun moments, and for always lightening things up with the right joke at the right time.

Emina and Johannes, thank you for your precious advice and support over the years. For the productive conversations on research ideas, hearing-aid technologies, potential of our research to translate into real-world applications. Thank you for always welcoming me during my visits in Denmark and helping me through the audio setups in the Pitch Lab, introducing me to the amazing Eriksholm Research Center crowd.

A special thanks goes to the people at Eriksholm, for being so welcoming during my time there, and for enriching my scientific perspective with interesting conversations and interactions.

Giorgio, thank you for involving me in your study on foreign-directed speech, it was a pleasure working and learning from you. Thank you for the time spent together in Dublin, the hikes and the pints.

Bonnie, thank you for offering me the opportunity to apply my methods to the study on infants, it has been a fruitful collaboration.

To all the members of the Di Liberto Lab, thank you for always listening to my waffles, for the useful discussions and explanations on things I did not understand, for the help with data collection. Most of what I learned was learned with you. I will miss the fun conversations, cheeky pints and lunch breaks at the Business School canteen.

To the *OG team*: Emily, thank you for being a friend and always supporting me at any stage and mental state of this PhD. For all our endless conversations on life, research, future, what not to put on pizza. Thank you for making this journey easier. Amir, thank you for the encouragement and help throughout the years, for welcoming me among your gym bros and for the football games. And thanks for making me understand that I cannot play pool.

To Richard and the members of the Reilly and Lopez Valdes lab, thank you for your interest in my work, and for giving me the opportunity to present my research on several occasions.

Dai, thank you for all the coffee breaks together, plant exchanges, and for the emotional support.

To the *Italiani a Dublino* – Tecla, Vincenzo, Gabo, Antonio – thank you for always lifting my spirits through gourmet dinners, pub nights, and card games. And for volunteering your brains as well.

Nicolò, thank you for being the first friend I had in Dublin and for introducing me to the Irish ways. For the many inspiring discussions and dinners together, and for your constant encouragement.

To my friends in Italy, thank you for always supporting and reassuring me, despite the distance and the business of adult life. Thankfully, some things do not change.

Ai miei genitori e ai miei fratelli – Claudia, Riccardo, Matteo – grazie per il vostro affetto e sostegno durante questi anni, per la vostra capacità di sdrammatizzare i problemi e per la piena fiducia nelle mie abilità.

To my husband Emanuele, thank you for firmly believing in me even when I do not, for your willingness to listen and for your genuine curiosity towards what I do. For the interesting conversations, the challenging questions, the unfailing support, for marrying me in the middle of this PhD. Thank you for loving me and for being loved.

This research was kindly supported and co-funded by Taighde Éireann – Research Ireland and the William Demant Fonden.

SARA CARTA

Trinity College Dublin
September 2025

Publications and Presentations arising from this thesis

Journal Publications

(Study 1) Carta, S., Aličković, E., Zaar, J., Valdés, A. L., & Di Liberto, G. M. (2024). Cortical encoding of phonetic onsets of both attended and ignored speech in hearing impaired individuals. *PloS one*, *19*(11), e0308554. https://doi.org/10.1371/journal.pone.0308554

(Study 2) Carta, S., Aličković, E., Zaar, J., Valdés, A. L., & Di Liberto, G. M. Simultaneous cortical tracking of competing speech streams during attention switching, *under review*

(Study 3) Carta, S., Aličković, E., Zaar, J., Valdés, A. L., & Di Liberto, G. M. Investigating the neurophysiology of attention switching in the ageing brain, *in preparation*

Additional publications

Carta, S., Mangiacotti, A. M. A., Valdes, A. L., Reilly, R. B., Franco, F., & Di Liberto, G. M. (2023). The impact of temporal synchronisation imprecision on TRF analyses. *Journal of neuroscience methods*, *385*, 109765. https://doi.org/10.1016/j.jneumeth.2022.109765

Piazza, G., **Carta, S.**, Ip, E. Y. J., Pérez-Navarro, J., Kalashnikova, M., Martin, C. D., & Di Liberto, G. M. (2025). Are you talking to me? How the choice of speech register impacts listeners' hierarchical encoding of speech. *Imaging neuroscience (Cambridge, Mass.)*, *3*, imag_a_00539. https://doi.org/10.1162/imag_a_00539

Carta, S., Ahmed F., Lau B.K., Di Liberto, G. M. Investigating speech perception in multi-talker scenarios in infants, *in preparation*

Poster and Oral Presentations

March 2025 Oral presentation, Eriksholm Attention Symposium, Snekkersten, Denmark. "Attention switching at the cocktail party. Insights from older and younger listeners"

Dec 2024 Poster presentation, TCBE (Trinity Centre for Biomedical Engineering) Research Day, Dublin. "Attention Switching at the Cocktail Party"

March 2024 Oral presentation, Association for Research in Otolaryngology (ARO) Midwinter Meeting, Orlando, USA. "Investigating the Hierarchical Neural Mechanisms Underlying Auditory Attention Switching"

Jan 2024 Oral presentation, Eriksholm Attention Symposium, Snekkersten, Denmark. "EEG metrics of sustained and dynamic attention in realistic listening scenarios"

March 2023 Invited Talk to Laboratoire des Systèmes Perceptifs, École Normale Supérieure, Paris, France. “Cortical Encoding of Phonetic Features of Attended and Ignored speech in listeners with hearing-impairment”.

Feb 2023 Oral presentation, Association for Research in Otolaryngology (ARO) Midwinter Meeting, Orlando, USA.
“Cortical Encoding of Phonetic Features of Attended and Ignored speech in listeners with hearing-impairment”.

Dec 2022 Oral presentation, TCBE (Trinity Centre for Biomedical Engineering) Research Day, Dublin.
“Cortical Encoding of Phonetic Features of Both Attended and Ignored speech in Hearing-Impaired Individuals”

Oct 2022 Poster presentation, Neuroscience Ireland Young Investigator Symposium, Dublin. “Investigating the impact of hearing-aid processing strategies on hierarchical speech processing”

June 2022 Oral presentation, Eriksholm Attention Symposium, Snekkersten, Denmark.
“Pay attention! Phonetic feature encoding in hearing-impaired listeners in Cocktail Party scenarios”

June 2022 Poster presentation, International Conference for Cognitive Hearing Science and Communication (CHSCOM), Linkoping, Sweden.
“Investigating the impact of hearing aid processing strategies on hierarchical speech processing”

Table of contents

Chapter 1. Introduction 1

1.1 Background 4

1.2 Research goals and Collaborations 6

Chapter 2. Perceiving and attending to speech 7

2.1 From the auditory periphery to the linguistic brain 7

2.1.1 Peripheral auditory system 7

2.1.2 Central auditory system 8

2.1.3 The language network 9

2.2 From sound to meaning 11

2.2.1 A cortical hierarchy of linguistic representations 12

2.2.2 Predicting speech 16

2.3 Neural tracking of speech 18

2.3.1 Measuring neural signals 19

2.3.2 Electroencephalography 20

2.3.3 From Event-Related to Continuous speech paradigms 20

2.3.4 Analytical tools to quantify neural tracking 22

2.3.5 The Temporal Response Function (TRF) framework 23

2.4 The neurophysiology of auditory attention 26

2.5 Cortical processing of speech across ageing and hearing loss 30

Chapter 3. Study 1: Investigating phonological processing of competing speakers in hearing impaired listeners 36

3.1 Introduction 36

3.2 Methods 37

3.2.1 Ethics Statement 37

3.2.2 Data Availability 37
3.2.3 Participants 38
3.2.4 Experimental design 39
3.2.5 Stimulus feature extraction 41
3.2.6 EEG data preprocessing 42
3.2.7 TRF computation 42
3.2.8 Speech feature models 43
3.2.9 Phonetic distance maps 44
3.2.10 Statistical analysis 45
3.3 Results 45
3.3.1 Robust phonetic-feature TRF for target but not masker speech in listeners with HI 45
3.3.2 Cortical encoding of both target and masker phoneme onsets in listeners with HI 50
3.4 Discussion 52
3.4.1 Phonological cortical encoding of target and masker speech in listeners with HI. 53
3.4.2 Phoneme onsets are encoded for both target and masker streams 54
3.5 Summary 56
Chapter 4. Study 2: Simultaneous cortical tracking of competing speech streams during attention switching 58
4.1 Introduction 58
4.2 Methods 60
4.2.1 Participants and experimental procedure 60
4.2.2 Stimuli 62
4.2.3 EEG data preprocessing 62
4.2.4 Speech features 62
4.2.5 TRF and analysis procedure 63

4.2.6 Event-related spectral perturbation (ERSP) 65
4.2.7 Statistical analysis 66
4.3 Results 66
4.3.1 Behavioural performance 66
4.3.2 Decoding of selective auditory attention in a dynamic switching scenario 66
4.3.3 Neural tracking of competing speech streams in a dynamic switching scenario reflects the listener's focus of attention and is related to changes in alpha ERSP 67
4.3.4 Dissecting the temporal dynamics of attentional disengagement and engagement during attention switching 70
4.3.5 Determining how lexical predictions are built during attention switching 71
4.4 Discussion 74
4.5 Summary 79
Chapter 5. Study 3: Investigating the neurophysiology of attention switching in the ageing brain .. 80
5.1 Introduction 80
5.2 Methods 81
5.2.1 Participants 81
5.2.2 EEG experimental procedure 82
5.2.3 Determining the SNR 83
5.2.4 Cognitive test battery 84
5.2.5 Experimental stimuli 85
5.2.6 Speech features 85
5.2.7 EEG data preparation 86
5.2.8 TRF analysis 86
5.2.9 Event-Related Spectral Perturbation (ERSP) 88
5.2.10 Event Related Potential (ERP) time-locked to the switch 88
5.3 Results 89

5.3.1 Behavioural performance 89

5.3.2 Neural encoding of speech features in younger and older listeners during an attention switching task 89

5.3.3 Frequent attention switching disrupts sustained speech tracking, particularly in older listeners 91

5.3.4 Attention switching elicits a P300-like response 95

5.3.5 Neural tracking around the switch in younger and older listeners 96

5.4 Summary 103

Chapter 6. General Discussion 105

6.1 Neural tracking beyond the envelope to reveal speech processing differences 106

6.2 A novel approach to study dynamic attention 108

6.3 Perspectives on ageing and attention reorienting 111

6.4 Ongoing and Future Work 113

Appendix 115

Bibliography 118

Chapter 1. Introduction

Language is a hallmark of human nature. The ability to translate one's thoughts and feelings into sophisticated verbal or signed motor plans has provided humans with an evolutionary advantage compared to other species (Hauser et al., 2002). This adaptive advantage lies not only in the ease and speed of communication enabled by language, but also in the capacity to exchange and accumulate knowledge across generations. As such, language has effectively shaped cultural transmission of complex abstract information, supported the development of social structures and, ultimately, scaffolded the emergence of human civilisations (Fedorenko, Piantadosi, et al., 2024).

If the primary purpose of language is to communicate, then the neural underpinnings of speech and language abilities should be studied primarily in settings that reflect real-world scenarios and the challenges they entail. In fact, despite our seemingly effortless and instinctive ability to use and understand speech (Pinker, 1994), our daily auditory environment is complex, filled with background noise, overlapping speakers, echoes, varied voices and accents, ambiguities, and interruptions. Even when considering simpler perceptual scenarios, it is not straightforward to explain how abstract meaning is successfully reconstructed by the listener from acoustic signals. In the quest to understand the neural basis of naturalistic speech perception, one key factor to consider is the speed at which language unfolds, which makes it necessary to measure brain activity with invasive and non-invasive technologies capable of high temporal resolution, such as electro-, magneto-encephalography (EEG and MEG) and intracranial EEG (iEEG). Their millisecond resolution allows to monitor neural processing dynamics almost in real-time as speech unfolds, offering the opportunity to study the impact of fast-changing speech properties. In addition to the temporal aspect of speech perception, the attempt to identify brain network(s) anatomically and functionally specialised in the processing of language has also propelled neurophysiology research with technologies capable of high spatial resolution, such as functional Magnetic Resonance Imaging (fMRI) and functional Near-Infrared Spectroscopy (fNIRS).

What makes the neural processing of language particularly difficult to investigate is the fact that speech is much more than an acoustic signal. To understand even the most basic sentence, one needs to recognise phonemes - the smallest sound units of a language - that, once

combined, form words. Single words have meanings per se, but those meanings are highly dependent on the meanings of nearby words, on the order in which they appear in the sentence, and on the logical structure (syntax) that embeds them. The word *bar* in "*She was admitted to the bar"* means something completely different depending on whether the person wanted to drink a beer or to practice law. As such, extracting meaning from a sentence requires the access to stored phonetic and lexical knowledge of the language, learned combinatorial rules, and sensitivity to the context. The brain needs to analyse these nested and interdependent levels of information with parallel and interactive neural mechanisms that are difficult to disentangle, temporally and spatially (de Heer et al., 2017; Peelle, 2012).

To complicate the picture even further, the brain is performing all these operations in real-time, with stringent time constraints to reach successful understanding and communication. To optimise for these constraints, comprehension does not start at the end of the sentence, but it unfolds incrementally, integrating phonemes or words as soon as they become available, and even anticipating them with top-down predictive mechanisms, based on the statistical knowledge accumulated through extensive experience of the language (Heilbron et al., 2022). This predictive coding theoretical framework (Friston & Kiebel, 2009; Hesselmann et al., 2010), originated in the domain of visual sensory experience (Alink et al., 2010; Rao & Ballard, 1999), is now applied to many sensory and cognitive domains, including language processing. In this view, the brain is constantly attempting to predict upcoming linguistic information, and it is constantly updating its internal model based on the bottom-up input it receives. When the perceived stimulus does not match our expectations, e.g., when we hear a surprising word in a sentence, a prediction-error signal is generated and propagated back through the processing hierarchy, reshaping and informing our expectations based on the newly available evidence (Kutas & Federmeier, 2011; Kutas & Hillyard, 1983). Not only prediction errors, but also linguistic predictions per se impact what we perceive, for instance allowing us to reconstruct in our auditory cortex, and actually hear, entire missing words from a noisy input (Leonard et al., 2016; Sohoglu & Davis, 2016).

Despite this complexity, the complementary strengths of various neuroimaging technologies have shed light on some of the neural mechanisms responsible for speech processing, which is proposed to depend on the parallel and distributed temporal integration of information (Hamilton et al., 2021) at multiple hierarchical levels, from the sound acoustics to higher-level properties such as word predictability and syntactic structure (Broderick et al., 2018; Caucheteux et al., 2023; Di Liberto et al., 2015; Di Liberto et al., 2021; Heilbron et al., 2022; Weissbart & Martin, 2024). The speech processing hierarchy also corresponds to a

functional neural hierarchy involving progressively higher cortical areas for more complex attributes of the speech input (Fedorenko et al., 2010; Hickok & Poeppel, 2007) within specialised language networks (Hagoort, 2019; Regev et al., 2024).

As mentioned above, speech perception in the real-world often occurs in noisy scenarios with overlapping speech streams, requiring listeners to direct their attention according to behavioural goals, e.g., to enhance a specific speech stream while ignoring others (Power et al., 2012), to rapidly monitor the entire auditory scene (Kaufman & Zion Golumbic, 2023) or to switch attention to a previously unattended stimulus (Getzmann et al., 2020). Regardless of the specific objective, attention to speech has been found to engage a broad brain network including auditory association areas and frontoparietal regions (Agmon et al., 2021), which is differentially modulated depending on the attentional demands. Neural markers to various levels of linguistic information are impacted by the listener's focus of attention, reflecting the target (attended) speaker more than the masker (unattended) competing speakers (Mesgarani & Chang, 2012; Zion Golumbic et al., 2013). For example, an attentional bias emerges for the neural tracking of the amplitude envelope of natural speech (Ding & Simon, 2012a; Power et al., 2012; Puvvada & Simon, 2017; Simon, 2015), and has been used to robustly classify the attended speaker from brain responses (de Cheveigné et al., 2018; O'Sullivan, Power, et al., 2015), with potential use for Brain-Computer Interface (BCI) applications (Dau et al., 2018; Hjortkjær et al., 2025). Recent work has also found attentional selectivity in the neural processing of phonetic information in multi-talker scenes, indicating that cortical responses encode phonemes more strongly for the target than for the masker speech (Raghavan et al., 2023; Teoh et al., 2022).

However, much work on attention to speech in multi-talker contexts has focussed on exploring the neural underpinnings of sustained attention to one speaker, studying mostly young participants with normal hearing. It is still unclear how attention modulates neural responses to various speech levels in listeners with hearing-impairment, or in the ageing brain. Moreover, in real-world scenarios, we often disengage from one particular talker and turn our focus to a previously ignored speech stream, rather than consistently sustaining attention to one speaker, but the study of how attention modulates neural responses to speech in the process of switching is still in its infancy.

In this thesis, I set out to investigate the cortical mechanisms of sustained attention and attention switching, offering new insights on the contribution of phonological information in hearing-impaired listeners, and on the dynamic reorienting of attention to speech in younger and older populations. This work is the outcome of fruitful academic collaborations with

industry partners, and of extensive EEG data collection, resulting in two full datasets with a novel attention switching paradigm.

1.1 Background

Extracting meaning from sound waveforms is a computationally hard task. Despite the intricacies of the phenomenon, we learn to use speech at a surprisingly fast pace, with infants widely surpassing the most advanced artificial language models (Rowland et al., 2025), and we effectively understand speech across a variety of communication scenarios, overcoming background noise and sound distortions (Fuglsang et al., 2017). Using language feels easy, but even with the most tightly controlled experimental setups and optimal selection of participants, our field has yet to reach a full neurophysiological account for how neural mechanisms lead to speech comprehension. Why then should we attempt to study speech perception in more naturalistic scenarios, and across diverse populations?

First, because of the very nature of language and communication, which is, etymologically, *to share*, *to put in common*. Language is used in the real world, in imperfect interactions, and refers to things of the human experience. Our neural mechanisms have evolved to accommodate these functions and if we turn speech into a hyper-controlled laboratory phenomenon, we will miss the full extent of its neural underpinnings. As an analogy, imagine a scenario where an alien encounters a boat for the first time. Unsure about what that object could be, the alien dissects it into parts and analyses them one by one. Eventually, it manages to start the engine and believes it has finally solved the mystery. However, despite all the efforts, the alien never realises that the boat could float on water. In a similar vein, much was learned in auditory neuroscience thanks to carefully controlled experimental settings isolating particular attributes of speech stimuli (Kutas & Federmeier, 2011; Kutas & Hillyard, 1980). However, now that advances in experimental paradigms, technological equipment and engineering methodologies allow it, we can finally take the opportunity to study the neural encoding of language in its *natural habitat*.

Secondly, to fully describe a system, one needs to observe it and evaluate it across a range of scenarios. As such, including populations different from younger participants with normal hearing is important to gather a more complete picture of the speech perception dynamics. Pioneers of speech neurophysiology, such as Pierre Paul Broca or Carl Wernicke, achieved foundational discoveries through lesion studies (Ahmad et al., 2024; Dronkers et al., 2007; Lorch, 2008), and more can be understood about the general cortical mechanisms of

speech perception if we also consider how those processes change across ages and hearing abilities.

Thirdly, there is an argument for the usefulness of research on ageing or hearing-impaired populations towards clinical applications. With ageing, hearing thresholds increase naturally, particularly at higher frequencies, due to degradation of the peripheral and central auditory system (Jin et al., 2024), resulting in a declining sensitivity to sound and often in age-related hearing loss. A growing body of literature indicates a strong correlation between hearing deficits and cognitive decline (Livingston et al., 2020; Loughrey et al., 2018; Thomson et al., 2017), but it is unclear if the two phenomena share a common neural basis, or if there is any particular direction of causality. In this context, investigating speech perception across the life span can help to disentangle physiological from pathological hearing decline and identify potential neural markers of degraded speech perception (Peelle & Wingfield, 2016). Furthermore, elucidating the different cortical processing strategies employed by older individuals or listeners with hearing impairment can effectively aid in novel clinical interventions and in the design of a new generation of neurally-steered hearing aids (Dau et al., 2018). For example, hearing aids that could detect the focus of the listener's attention almost in real time (Hjortkjær et al., 2025) and enhance the voice of the attended speaker from an auditory mixture.

Studying these populations presents significant challenges. Due to the specific age or hearing-level criteria, there are increased difficulties in participants' recruitment, which I experienced first-hand while conducting my latest study with a cohort of participants above fifty years of age. Other issues relate to the homogeneity of the test sample, due to the increasing co-occurrence of various health conditions, that increase inter-subject variability. When considering participants with hearing impairment, the variability in neural responses is even higher (Billings et al., 2015), due to the concerted effects of ageing and adaptation to degraded speech input (Herrmann, 2025; Peelle & Wingfield, 2016). Investigating speech perception in experimental settings closer to naturalistic scenarios also introduces its own set of complexities, with less control over participants' behaviour and attention (Makov et al., 2023), and reduced ability to disentangle the impact of highly overlapping speech features, which also needs to be carefully annotated on the stimulus (Agmon et al., 2025).

1.2 Research goals and Collaborations

In this thesis, I aim to elucidate the cortical basis of attention to speech in multi-talker scenarios, across younger and older populations with hearing impairment and with normal hearing abilities, using EEG recordings. Firstly, I investigate the contribution of phonological information to the attentional selectivity in older listeners with hearing impairment. In this work, I test the hypothesis that phonological information could be encoded simultaneously for target and masker streams, while sustaining attention on one particular speaker. Secondly, I study the cortical processing of speech in less explored attention switching scenarios, delving into the neural basis of attention reorienting in younger and older participants with normal hearing. In particular, this work aims to elucidate the temporal dynamics of disengaging from one speaker and engaging to a newly attended speaker, and on the lexical processes underlying context switching. Exploring these main complementary paths, the present research work converges on a common scientific intent to clarify how various speech levels interact to guide attentional selection, how they are engaged and disengaged in contexts of flexible attention reorienting, and how the neural mechanisms underpinning these phenomena vary across ages and hearing levels.

The studies presented in this thesis are the result of a tight collaboration with Dr. Johannes Zaar and Dr. Emina Aličković, from the Eriksholm Research Centre (ERH), part of Oticon, which specialises in audiological and cognitive hearing research. The EEG dataset involving older listeners with hearing impairment (presented in Study 1 of this thesis) was collected at the ERH and shared as part of the ongoing collaboration with Trinity College. The cooperation with ERH was equally important for the design of the attention switching experiments, which I also had the opportunity to pilot at ERH laboratories, during my research stay in Denmark, on two older participants with hearing loss.

Furthermore, the interaction with Dr. Giorgio Piazza (Basque Center on Cognition, Brain and Language, now at University of Padova) provided an opportunity to investigate the cortical dynamics of phonological encoding of speech across various speech registers (foreign-directed or native-directed speech) in native and non-native English speakers. Finally, the collaboration with Dr. Bonnie Lau (University of Washington), allowed for the investigation of the attentional mechanisms in infants exposed to multiple speech streams. While the work arising from those collaborations is not presented in this thesis, it is an integral part of the broader contribution of my PhD.

Chapter 2. Perceiving and attending to speech

This chapter summarises the literature relevant to the work presented in this thesis and is divided into five sections. The first section describes the journey of the speech waveform from the auditory periphery to auditory cortical areas and speech-specialised networks. The second section presents how abstract meaning is extracted from lower-level speech features, and how predictive mechanisms impact speech perception. The following section describes the phenomenon of neural tracking of speech, with a focus on recording techniques, analytical tools, and experimental paradigms. The fourth section presents an overview of the current knowledge on the neural mechanisms of auditory attention, with a particular focus on selective attention in multi-talker scenarios. The fifth section focusses on the neural basis of speech perception in populations of older adults and adults with hearing impairment, with a brief detour on the concept of listening effort.

2.1 From the auditory periphery to the linguistic brain

Spoken language first reaches the listener as a sound-pressure waveform processed through the auditory periphery, as many other auditory stimuli. After being transformed into electrical impulses at the level of the auditory nerve, speech is then processed by a specialised language network within the auditory cortex, where it becomes meaning.

2.1.1 Peripheral auditory system

Speech, similar to other sounds, is an acoustic waveform initially detected by the peripheral auditory system (**Figure 2.1 A**). The sound is first captured through the *outer ear*, channelled from the ear's *pinna* into the *auditory canal*. It then reaches the *tympanic membrane* and causes it to vibrate. The vibrations of the tympanic membrane propagate through the air-filled cavity of the *middle ear,* where they are transmitted to three auditory ossicles, the *malleus*, *incus*, and *stapes*. The movement of this bone structure causes a second membrane, the *oval window*, to vibrate, propagating into the *inner ear*.

Here, vibrations reach the fluid-filled cochlea, characterised by the presence of *stereocilia*, i.e. hair-cells, that are located on the inner surface of the cochlea, on the *basilar membrane*. They float in the cochlea's fluid and are extremely sensitive to any oscillation of the basilar membrane, which causes them to deflect. The mechanical stimulation induces the opening of

ion channels on the stereocilia's surface, which triggers a cascaded exchange of positively charged ions and eventually leads to the depolarisation of the receptor.

If a threshold depolarisation is reached, the stereocilia will release a transmitter to an afferent fibre of the *auditory* (or *cochlear*) *nerve*, initiating an electrical neuronal signal. The sensitivity of stereocilia to particular sound frequencies, depending on their position along the basilar membrane, creates a *tonotopic map* that is responsible for the first instance of frequency tuning in the auditory system. Damage, loss, or fusion of stereocilia is a cause of hearing loss particularly in older individuals (Wu & Liberman, 2022). Greater damage is typically observed in the stereocilia located on the outer, more exposed regions of the basilar membrane, which are tonotopically tuned to higher frequencies. This pattern helps explain the age-related elevation in hearing thresholds, particularly for high-frequency sounds.

2.1.2 Central auditory system

Once the mechanical stimulation of stereocilia has been converted into an electrical impulse through the auditory nerve, sound processing is carried out by the central auditory system, which includes subcortical and cortical processing stages.

The *subcortical auditory system* consists of a series of nuclei within the brainstem and midbrain, each contributing uniquely to auditory signal processing before information reaches the cortex (**Figure 2.1 B**). First, the *cochlear nucleus* (CN), in the medulla, receives input from the auditory nerve fibres, and contains a variety of cell-types (Pressnitzer et al., 2008) preserving frequency information, extracting sound timing and intensity, and detecting onsets (Oline et al., 2016; Rhode & Smith, 1986). Axons from the CN project to the *Superior Olivary Complex* (SOC), in the pons, which integrates input from both ears, computing interaural time differences (van der Heijden et al., 2013). Through the *Lateral Lemniscus*, auditory information passes from the brainstem to the *Inferior Colliculus* (IC) in the midbrain, which displays rich connections with almost every other auditory structure, and it is fundamental for multi-sensory integration with motor areas (Winer & Schreiner, 2005). The next step in the subcortical pathway is the *Medial Geniculate Nucleus* (MGN) of the thalamus, which finally projects to the primary auditory cortex. For a more detailed general overview of the subcortical pathways, please see (Felix et al., 2018).

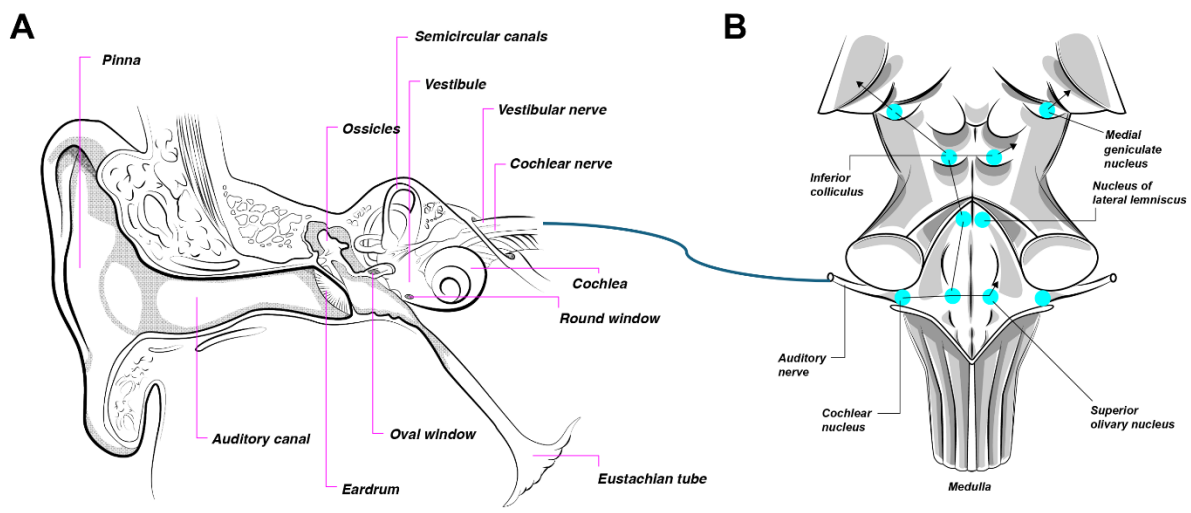


Figure 2.1: Peripheral and subcortical auditory system. (A) Overview of the structure of the outer and inner ear. **(B)** Structures involved in sound processing at the level of the brainstem and midbrain. Relevant sound processing nuclei are highlighted. (Adapted from the MUsic Technology Online Repository team, available at https://mutor-2.github.io/team.html).

The studies presented in this thesis focus on the neural underpinnings of sustained and dynamic attention to speech in the *auditory cortex*. As such, I first describe the anatomical organisation of cortical areas involved in sound processing, before addressing how these areas are organised into a speech-specialised network in the human brain.

The auditory cortex is located bilaterally in the *superior temporal gyrus* (STG) of the temporal lobe, developing primarily within and around the *Transverse temporal gyrus* (TTG), and extends into surrounding cortical areas. It is functionally and anatomically organised into the primary *core* (*Heschl's gyrus* , HG), buried within the lateral sulcus, which receives input from the thalamus, and non-primary *belt* and *parabelt* regions around it, forming a hierarchical processing system (Chevillet et al., 2011; Rauschecker & Afsahi, 2023; Sweet et al., 2005).

Belt and parabelt areas surround HG developing anteriorly, on the *Planum polare* (PP), and posteriorly, on the *Planum temporale* (PT). The auditory cortex maintains the tonotopic organisation of subcortical areas, but additionally processes sound in a highly context-dependent manner, integrating auditory input with sensory and motor signals, modulating its activity based on experience, and adapting dynamically to cognitive demands such as attention (King et al., 2018).

2.1.3 The language network

Beyond the domain-general auditory cortical areas, speech sounds engage a wider ensemble of regions, which together constitute the *language network*. This network - showing a left-

hemisphere bias (Zatorre, 2022), despite being bilaterally organised (Rogalsky et al., 2022) - is specialised for the perception and production of speech stimuli, and characterised by tight connections with perceptual and motor cortical areas, along with areas involved in abstract thinking and reasoning (Fedorenko, Ivanova, et al., 2024). However, others claim that language is coarsely represented in the brain (Drijvers et al., 2025), as part of a more general tendency in the field of neuroscience to rethink localisationist approaches (Noble et al., 2024).

Despite acknowledging the importance of investigating the multidimensional and holistic nature of neural processing dynamics, this thesis adopts the perspective of a highly interconnected network specialised for language (**Figure 2.2**), spanning *anterior* and *posterior temporal lobes*, *middle frontal gyrus* and *inferior frontal gyrus* (Cai et al., 2025; Fedorenko, Piantadosi, et al., 2024; Goldstein et al., 2025; Hickok & Poeppel, 2007).

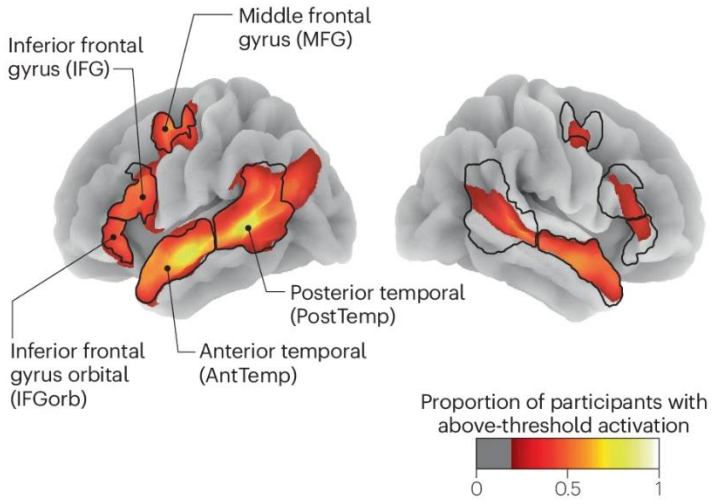


Figure 2.2: Cortical language network. Defined by a contrast of linguistically meaningful vs. non-meaningful information. This is an fMRI probabilistic activation map based on 806 individual participants. (Fedorenko, Ivanova, et al., 2024)

Within this network, experimental and theoretical work in the past twenty years has suggested that increasingly more complex speech properties are processed by increasingly higher-order cortical areas, from primary to non-primary regions (Chevillet et al., 2011; Hickok & Poeppel, 2007; Karunathilake et al., 2025; Poeppel et al., 2012). As such, primary auditory areas - especially HG, with its tonotopic organisation - have been related particularly to the processing of the envelope and spectral features of speech (Kubanek et al., 2013), whereas secondary areas are more specialised in the processing of higher speech properties such as phonemes and words (DeWitt & Rauschecker, 2012; Mesgarani et al., 2014).

Specific areas though react differently depending on the type of experimental stimuli used, e.g. it is different to probe phonological cortical representations with isolated syllables in controlled experimental designs (Liebenthal et al., 2005), or with continuous narrative speech, where the phonological dimension has to be disentangled from neural responses to continuous and complex speech streams (Di Liberto et al., 2015). Thus, when attempting to reconstruct a functional hierarchy of speech perception, it should be first made clear what kind of speech stimuli are being used (Peelle, 2012), as the responses observed in the language network are modulated depending on the surrounding stimulus context (King et al., 2018).

For instance, a concurrent activation of regions of the motor cortex has been reported while processing language, lending support to the motor theory of speech perception (Liberman, 1957). According to this classical framework, motor areas are fundamental for processing language, or at least they aid in speech perception. Many of the studies supporting this view, however, were conducted by asking participants to discriminate between syllables or single sounds (Pulvermüller et al., 2006), and the engagement of the motor system might be a consequence of the particular discrimination task demands, rather than of speech perception per se (Rogalsky et al., 2022).

2.2 From sound to meaning

Assuming that the goal of speech perception research is to study speech in its neural journey from sound to meaning, rather than simple sound contrasts, then experimental stimuli should be the most ecologically-valid as possible. In fact, as mentioned above, the cortical processes involved in language comprehension are not the linear sum of the mechanisms responsible for processing individual linguistic components (such as acoustics, phonology, and lexical items in isolation). Rather, because language is symbolic in nature, the brain must perform a series of nonlinear operations that reflect the inherently nested and nonlinear structure of language, and its arbitrary mapping between sound and meaning. As such, to fully characterise the neural dynamics underlying speech perception, the system should be considered in its entirety, despite the complications that this entails.

A useful approach to tackle this problem is to consider it in its threefold aspects of *representations*, *operations* and *information flow* (Gwilliams et al., 2025). Through a series of linear and nonlinear *operations* (like frequency mapping, filtering, and categorical perception), the brain generates *representations* in response to various features of the speech input, such as the meaning of a single word, or the spectral profile of a sentence. Neural representations along

with the operations generating them display temporal and spatial variability, which determines the *information flow* of the system. Within this highly flexible system, representations are subject to a range of transformations as they are processed by different neuronal populations, moving within and across brain areas through feedforward and feedback connections.

2.2.1 A cortical hierarchy of linguistic representations

Since neural operations and information flow are defined with respect to neural representations, it is necessary, as a first step, to specify those representations. Neural representations in turn are defined with respect to a particular feature or set of features of the speech stimulus, and a crucial part of speech perception research lies in the identification of features that are biologically plausible (Agmon et al., 2025), that is, meaningfully encoded in brain activity patterns. As one moves up the speech processing hierarchy, the features become increasingly abstract and less directly tied to the physical properties of the speech signal.

Lower-level features

Lower-level features, like the broadband amplitude envelope or the frequency spectrum, can be more easily extracted from the speech signal and readily related to brain activity. The brain's sensitivity to these acoustic features is not surprising, since it is present also for complex non-speech stimuli (Singh & Theunissen, 2003; Smith et al., 2002), but there is a particular bias towards speech sounds (Belin et al., 2000; Luo & Poeppel, 2007; Prinsloo & Lalor, 2022), even in primary auditory areas (Khalighinejad et al., 2021). Similarly, in animal studies, heightened neural activity has been observed for conspecifics' vocalisations as opposed to other sounds (Grace et al., 2003), pointing to a potential behavioural relevance explanation for this selectivity. Encoding of the speech envelope in brain responses has been extensively investigated, revealing that neural activity robustly represents temporal fluctuations of speech sounds (L. Decruy et al., 2020; Giraud & Poeppel, 2012; Kubanek et al., 2013), with mixed findings on the relationship between neural envelope encoding and speech intelligibility, since envelope tracking is also observed for unintelligible or time-reversed speech (Di Liberto et al., 2015; Howard & Poeppel, 2010; Steinschneider et al., 2013). However, envelope encoding also reflects higher level cognitive processes, as is shown by its modulation based on attention, which leads to stronger representation of envelopes of attended speech streams compared to ignored ones (Zion Golumbic et al., 2013).

A fundamental ability of the human speech perception system is its robustness to acoustic variability, allowing syllables or phonemes spoken in different voices and accents to

be consistently mapped onto the same underlying representation. In other words, a *"p"* sound is always recognised as a *"p"*, regardless of who pronounces it, adjacent sounds, and background noise. This implies a certain degree of categorical perception, or at least an acoustic invariance acting as a sort of perceptual magnet (Di Liberto et al., 2015; Guenther & Gjaja, 1996), which develops gradually, with increasing sensitivity to the native language, over the first year of life (Di Liberto et al., 2023; Menn et al., 2023). These phonological representations are encoded in the STG (Leonard et al., 2024; Liebenthal et al., 2005; Mesgarani et al., 2014; Yulia Oganian et al., 2023), where they coexist, but at distinct clustered cortical depths, with cortical representations relating to other speech features, such as sentence onsets and offsets, prosodic stress patterns and sound intensity (Hamilton et al., 2018; Leonard et al., 2024). Since different phonemes are distinguished based on their unique spectrotemporal profiles, there is naturally considerable overlap between acoustic and phonological representations. This has sparked a debate, leading some to reasonably question the distinct contribution of phonological information to neural responses to speech (Daube et al., 2019), while others found specific representations of phonological units (Di Liberto et al., 2015), especially when dissociating the two representations through points of acoustic-phonetic divergence (Mai et al., 2024). Cortical representations of the phonetic content of speech appear to be encoded earlier for more predictable phonemes, and also to persist in the neural responses to be integrated with upcoming information, especially in contexts of higher ambiguity (Gwilliams et al., 2022). Overall, phonological information is encoded in brain responses, with representations that are highly redundant and parallel with respect to acoustic features.

Higher-level features

While acoustic and phonological features are still tightly linked to the physical properties of the speech signal, more symbolic features must be extracted to ultimately arrive at meaning. This higher-level feature space comes into play as soon as phonological information needs to be combined into separate *lexical items* (i.e., words). In fact, unlike written text, where each word is neatly separated with a space, spoken language is a continuous stream of sounds, with co-articulated phonemes that cannot be easily assigned to distinct lexical items.

As such, the brain is confronted with the complex task of retrieving word boundaries from a raw continuous acoustic signal. To solve this problem, the brain can rely on complementary sources of information, such as *prosody* - the rhythm of the speech signal derived from pitch variation, sound durations and pauses - and *syntax* - the rules of how single words can be compositionally combined to yield meaningful sentences. Models of syntax based

on a hierarchical and compositional account of syntactic structure were found to be better descriptors of neural activity compared to accounts which only relied on the linear sequence of words, recruiting a network which includes bilateral temporal, parietal and prefrontal cortices (Brennan & Hale, 2019; Caucheteux et al., 2021). Beyond the compositional structure, also prosodic cues often determine the syntactic structure of a sentence and its correct interpretation. For example, "*I am hungry I am going to eat grandma*" shifts towards a more reassuring interpretation once a pause is inserted between "*eat*" and "*grandma*", and it becomes clear that the verb *to eat* is used in an intransitive sense. Prosodic and syntactic information are encoded in brain responses and their cortical representations can be disentangled (Chalas et al., 2024; Degano et al., 2024; Hamilton et al., 2021; Teoh et al., 2019).

While syntactic and prosodic information are useful cues, a prerequisite for lexical segmentation is to access lexical items from a *mental lexicon* (Aitchison, 2003), a stored vocabulary of learned words in a language, with their associated meanings. There are different views on what type of information is stored in the mental lexicon, how modular this information is, and how it is organised (Baayen, 2007; Stella et al., 2024). However, most models agree on the existence of a network-like structure within the mental lexicon, with nodes and relationship between them. Words with similar meanings will be localised closer together within the network and will have stronger connections. As such, word meanings can be described as vectorised representations in a multi-dimensional space, i.e. *word embeddings*, where meaning is defined relative to the other nodes of the network (Mitchell & Lapata, 2010). Yet the dimensions along which such a semantic network should be constructed are far from obvious, due to the complex relationship between word semantics and the conceptual system more in general (Liuzzi et al., 2020): should words be organised taxonomically, based on whether the concepts they refer to share specific features? For example, relevant dimensions in such a multi-dimensional space would be defined explicitly as *"has eyes"*, *"is animate"*, *"is pink"* and other similar properties with a binary realisation. Or should those organising dimensions be defined in a more data-driven way, based on distributional properties of word use within the language, extracted from large text corpora?

In the work presented here, as in the vast majority of recent research on the neural representations of lexical semantics, I will rely on the distributional approach, which allows to model the predictability structure of language based on the words' occurrence in large collections of text (Pennington et al., 2014). This approach is advantageous, as neurophysiological research over the past forty years has shown that the brain is sensitive to the predictability of speech, responding to an unexpected word with a characteristic deflection

in activity occurring reliably around 400 ms after its onset, generating the so called *N400 effect* (Kutas & Federmeier, 2011; Kutas & Hillyard, 1980). Importantly, recent research employing naturalistic speech listening paradigms highlighted that neural responses react not only to particularly unexpected words, but continuously track the information content of speech (Brodbeck et al., 2022; Broderick et al., 2018; Heilbron et al., 2022). As such, particularly in the context of encoding/decoding models (which will be explained in detail below), cortical representations of semantic features can be described with respect to features such as word surprisal, dissimilarity and entropy (Shannon, 1948), relative to their preceding words.

The first paradigm shift in the distributional representation of lexical semantics came from models like Word2Vec and Glove (Mikolov et al., 2013; Pennington et al., 2014), which were the first to enable efficient training on very large datasets, allowing the representation of fine-grained lexical meaning within reasonable computational demands. These models represented word meanings as numerical embeddings learned from large text corpora, capturing complex relationships among words and global statistical patterns. Speech neurophysiology research naturally leveraged these models to extract numerical word meanings and relate them to neural activity, showing that distributional lexical models mirrored how language is represented in the brain (Broderick et al., 2018; Gillis et al., 2021; Zhang et al., 2020). While those models represented a fundamental step to study the neural basis of lexical semantics, they had the limitation of encoding word meanings as static embeddings across various lexical contexts. As a result, these models did not distinguish between semantically ambiguous words ("*bat*" as animal vs. tool) and mapped a particular word used in very different contexts to the same embedding. For example, they did not account for the different meaning of the verb *"to die"* in *"I would <u>die</u> for a coffee right now"* and *"I would <u>die</u> for you"*.

The recent revolution of deep Large Language Models (LLMs) has opened new possibilities to model a word's meaning in context, with embeddings that are highly sensitive to both the syntactic and semantic context of a word, and thus better capture the compositionality of meaning (Gwilliams et al., 2025). Word predictability features extracted from these models, or their contextualised word embeddings, have been successfully used to identify neural representations of lexical semantics across multiple neuroimaging modalities (Caucheteux et al., 2022; Caucheteux & King, 2022; Chalehchaleh et al., 2025; Goldstein et al., 2022). The layered deep structure of LLMs has also been related to brain activity patterns, revealing a correspondence between layers of the models and processing stages and cortical areas in the brain (Caucheteux et al., 2022; Caucheteux & King, 2022; Kell et al., 2018;

Mischler et al., 2024; Schrimpf et al., 2021; Tikochinski et al., 2025). However, most LLMs, such as GPT, are generative models (Radford & Narasimhan, 2018), built and optimised for next-token prediction, and not explicitly trained to extract any particular feature of speech. This leads to two main issues: First, LLMs might employ very different strategies compared to the human brain in order to satisfy their training objective (Moro et al., 2023). Second, as LLMs become more and more specialised in their predictive task, they might increasingly diverge from human capabilities. For example, while increasing the length of lexical context used to predict the next token might be advantageous for the model, it does not improve the model's ability to describe neural data, probably due to limitations in human memory while processing speech (Tikochinski et al., 2025). Despite these caveats, LLMs are proving to be useful tools to extract complex and contextualised representations of meaning.

2.2.2 Predicting speech

The brain's tracking of predictability is not a phenomenon confined to lexical semantics, but it is part of a more general *predictive processing* (PP) strategy. PP, initially conceptualised in the domain of vision research (Rao & Ballard, 1999) and soon applied to many sensory modalities, including audition, has become a popular theory for how the brain represents and reacts to the environment.

The main idea of PP is that the brain uses prior knowledge to generate expectations about incoming sensory input. It then compares these expectations to the actual input, and any mismatch results in a prediction error. The influence of this error on updating internal models depends on how reliable or precise the sensory input is considered to be. In the classic formulation of PP (Friston, 2005, 2009; Friston & Kiebel, 2009; Keller & Mrsic-Flogel, 2018), higher-level cortical areas send predictions down the processing hierarchy, while prediction errors flow upward to update internal models based on sensory information.

Predicted inputs suppress neural activity, while unexpected ones enhance it. In almost all formulations of PP, predictions and prediction errors are assumed to be generated by different populations of neurons located at different cortical layers (Bastos et al., 2012; Clark, 2013), but there are contrasting views on this, with some suggesting that errors are computed at the single-neuron level (Mikulasch et al., 2023) and others recently advocating for a less hierarchical approach, proposing a *predictive routing* alternative (Gabhart et al., 2025).

In the auditory domain, even for simple non-speech sounds, there is evidence that neural responses are shaped by anticipatory expectations and that these expectations are hierarchically organised (Heilbron & Chait, 2018; Walsh et al., 2020), but it is more difficult to conclusively

describe the neuronal dynamics of predictions and prediction error as being implemented by separate units. In earlier work, the brain's ability to track the predictability structure of auditory stimuli was mostly investigated with simple paradigms where sounds displayed various predictability patterns, often with omissions or deviations from the usual sequence (Näätänen et al., 1978). Clever designs were used to disentangle habituation effects from actual processing of the pattern's statistics (Wacongne et al., 2012), and to test for the hierarchical nature of the predictions (Lecaignard et al., 2015), exploring longer-range statistics.

As introduced above, the neural processing of statistical regularities in the context of language has also been extensively studied, initially with controlled paradigms relying on semantic or syntactic violations (Coulson et al., 1998; Kutas & Hillyard, 1980), and more recently with studies probing these mechanisms with continuous and naturalistic speech stimuli (Brennan & Hale, 2019; Brodbeck et al., 2022; Broderick et al., 2018; Caucheteux et al., 2023). Given the compositional and hierarchical structure of language, predictions operate not only within single linguistic levels but interact in complex ways across the language hierarchy and temporal timescales (Heilbron et al., 2022; Schmitt et al., 2021), with higher-level predictions informing lower-level ones. Furthermore, linguistic predictions are specific to the language network and can be mapped to distinct cortical spatial locations for different speech levels (e.g., phonemic vs. lexical predictions) (Heilbron et al., 2022; Shain et al., 2020). This framework even opened the possibility to decode word predictions as they occur from brain activity patterns, before the word is perceived, (Goldstein et al., 2022; Leonard et al., 2016), offering insights into the online processing dynamics of speech. However, some are recently casting doubts on these pre-stimulus effects, attributing them to structural dependencies contained in the speech stimulus itself, and inviting more caution in their interpretation (Schönmann et al., 2025).

The findings that speech regressors based on predictability allow to accurately describe neural responses, together with the mapping of LLMs layers to cortical processing stages (Caucheteux et al., 2023; Caucheteux & King, 2022; Gwilliams et al., 2023) have contributed to the idea that linguistic processing is heavily based on predictive mechanisms. While this is a plausible hypothesis, the correlation between features extracted from a model performing a lexical predictive task and neural activity in response to the same speech stimuli does not represent direct evidence that the brain is indeed performing exactly the same task as the LLM, i.e. predicting the next word (Antonello & Huth, 2024). In general, the rise of LLMs and the notion of predictive speech processing should coexist with efforts to develop explanatory

theories of the neurophysiological mechanisms underpinning speech processing (Slaats & Martin, 2025).

Our current understanding of these neurophysiological mechanisms speaks in favour of a system that is hierarchically organised, with neuronal populations across cortical areas sensitive to different temporal scales (Donhauser & Baillet, 2020; Norman-Haignere et al., 2022; Regev et al., 2024), showing the integration of speech information occurring at multiple timescales. However, the hierarchical nature of cortical *representations* and *operations* does not implicate a strict serial order. Instead, the *information flow* of the system is highly parallel, with linear and nonlinear operations occurring simultaneously, features being represented with a distributed cortical code (Gwilliams et al., 2022; Gwilliams et al., 2023; Hamilton et al., 2021; Leonard et al., 2024), and complex interactions of bottom-up and top-down connections across all speech levels (Heilbron et al., 2022; Weissbart & Martin, 2024).

2.3 Neural tracking of speech

When investigating the cortical representations of speech, it is necessary to perform two complementary tasks: on one hand, brain activity patterns relating to those features should be recorded; on the other hand, relevant features should be extracted from the speech signal. Ultimately, the relationship between measured brain activity and speech features, referred to as *neural tracking*, needs to be quantified. The cortical mechanisms underlying the neural tracking phenomenon, along with the methodologies to measure it, are diverse, but there is agreement on its general definition as the alignment of neural signals to properties of the speech sounds.

Since the foundational work of Luo and Poeppel (Luo & Poeppel, 2007), who established the possibility to discriminate naturalistic spoken sentences simply based on the phase of neural theta band (4-8 Hz) activity, an extensive body of research has investigated the neural tracking of speech, with heated debates around its cortical mechanisms. An intense debate regards whether neural tracking stems from endogenous oscillatory activity aligning to stimulus properties, or actively sampling speech stimuli at particular rates corresponding to the frequency of relevant (abstract or physical) speech units (Doelling et al., 2014; Giraud & Poeppel, 2012; Meyer et al., 2020; Zoefel & Kösem, 2024), or instead it is the result of transient stimulus-evoked activity (Y. Oganian et al., 2023). Regardless of the diverse views on its cortical origins, which are beyond the scope of this thesis, neural tracking has proven to be a powerful conceptual tool for studying speech perception.

2.3.1 Measuring neural signals

As was mentioned in the previous section, an essential step for quantifying neural tracking is to record neural activity patterns, which are then related to features of the experimental stimulus. Participants' brain activity is typically measured using non-invasive technologies that vary in their temporal and spatial resolution, depending on the physiological neural mechanisms that they target. Technologies such as functional Magnetic Resonance Imaging (fMRI) and functional Near-Infrared Spectroscopy (fNIRS) leverage magnetic and optical properties, respectively, of oxygenated and de-oxygenated hemoglobin (Irani et al., 2007; Ogawa et al., 1990), to infer changes in blood flow associated with increased neural activity in particular brain areas. Given their reliance on the blood oxygen level dependent (BOLD) signal, they have high spatial resolution, but slow temporal resolution, since changes in blood flow occur in the order of 1 to 2 seconds (Logothetis, 2008). On the other end of the temporal resolution spectrum, there are technologies such as Electroencephalography (EEG) and Magnetoencephalography (MEG) (Berger, 1929; Brookes et al., 2022; Cohen, 1972), that measure electric and magnetic fields produced by the concerted activation of neuronal populations. As such, they can record at very high rates, but it is more difficult to reconstruct the sources of the recorded signals (Wendel et al., 2009). These different non-invasive neuroimaging technologies can also be combined to leverage their complementary strengths, obtaining measurements that allow for high temporal and high spatial resolution simultaneously (Ru et al., 2022).

In rare cases, it is also possible to record electrical neural activity invasively, with electrodes implanted below the scalp as part of surgical procedures to identify the focus of epileptic activity on patients with a diagnosis of intractable epilepsy. These implants, in the form of grids on the surface of the cortex – electrocorticography (ECoG) (Mesgarani & Chang, 2012; Penfield & Jasper, 1954) - or arrays of depth electrodes - stereotactic EEG (sEEG) (Herff et al., 2020) - offer a very high temporal resolution, Signal-to-Noise ratio (SNR), and an extremely precise local spatial resolution. Recent advances such as the Neuropixel probe (Jun et al., 2017) even allow to record from sites at multiple cortical depths with a single-neuron precision (Leonard et al., 2024). However, these invasive methods pose some serious limitations: they cover very limited cortical areas, hindering the investigation of global activity patterns; they are hardly accessible due to their clinical nature; and they are recordings from patients, with the possibility that the observed neural signals might be impacted by functional and structural changes due to epilepsy (Parvizi & Kastner, 2018).

2.3.2 Electroencephalography

In speech processing research, EEG is a widely used technique due to its cost-effectiveness, portability, and ability to capture the rapid cortical dynamics of speech processing. A light cap, sixty-four (or less) electrodes non-invasively applied on the scalp and some conductive gel, along with proper signal amplification, are enough to reliably decode a listener's attention between competing speech streams (Lauteslager et al., 2014), compare the level of speech comprehension across first- and second-language speakers (Piazza et al., 2025), and observe the neural encoding of nursery rhymes in infants (Attaheri et al., 2022), among many other results. In all of the studies reported in this thesis, EEG was the recording technique of choice.

Through surface electrodes, EEG is able to capture minute electrical fields, in the order of 20-40 μV, resulting from the synchronised post-synaptic potentials of populations of cortical neurons. The possibility to measure macroscopic electrical activity from surface electrodes is offered by the passive conductivity of the tissues of the brain, skull, and scalp, which makes it possible for population activity to be recorded externally with precise temporal resolution, in the order of milliseconds. On the other hand, these biological layers also generate smearing of the signal, causing the low spatial resolution of this recording technique. Furthermore, the SNR of the recorded brain signal is strongly affected by concurrent biological activity, such as blinking, ocular movements (EOG), cardiac activity (ECG) and muscular activity (EMG), which are referred to as *artifacts* or *noise*. The occurrence of these events needs to be carefully controlled during the experimental session, by instructing participants to maintain their gaze steady and avoid movement. In the offline data pre-processing stages, the EEG signal is further cleaned from artifacts by using appropriate filtering, identification of electrodes with low SNR, and signal separation algorithms.

2.3.3 From Event-Related to Continuous speech paradigms

As described in Section 2.2.1, speech unfolds very rapidly and consists of nested levels of information of increasing complexity that interact and overlap with each other. As a result, EEG and MEG, with their millisecond resolution, have proven to be valuable technologies to measure the corresponding fast cortical dynamics. However, even if these recording methods can match the speed of the speech stimulus, the problem remains of disentangling the distinct contributions of the various levels of linguistic information from the listener's neural responses. One approach to solve this challenge is to select a particular level of linguistic information, e.g. the syllable or the word level, and present isolated stimuli from those categories, such that the

resulting cortical activity can be interpreted without confounds from other linguistic levels. Each of these isolated stimuli is defined as an *event*, and neural activity is measured time-locking it to the start of the stimulus, resulting in the so-called Event-Related framework. However, the SNR of neural signals is low compared to brain activity unrelated to the task and to external sources of noise, like electrical powerline noise or movements of the participant. As such, a reliable neural response can be obtained only after averaging over many repetitions (typically in the order of hundreds) of the same event.

Event-Related Potentials (*ERPs*) and *Event-Related Fields* (*ERFs*) measured in response to speech and non-speech sounds have enabled the identification of standard auditory *components*, i.e. stereotypical peaks and troughs of neural activity at predictable temporal latencies, such as the N1-P2 complex (Hillyard & Picton, 1978; Näätänen & Picton, 1987), which can be used as a useful clinical tool for audiological assessment of paediatric or non-verbal populations (Hyde, 2009). Crucially, these components have been shown to reflect different stages of the cortical processing of sounds (Picton, 2013) and also attentional demands (Lewald & Getzmann, 2015). Additionally, ERPs and ERFs have been observed in response to the brain's automatic detection of deviations in regular auditory patterns, giving rise to the Mismatch Negativity (MMN) response (Winkler & Czigler, 1998). The detection of linguistic violations also elicits characteristic event-related responses - namely, the N400 and P600 - occurring approximately 400 ms and 600 ms after stimulus onset, and associated with violations of semantic and syntactic expectations, respectively (Coulson et al., 1998; Kutas & Hillyard, 1980, 1983).

Overall, the ERP framework has proven to be a powerful tool in auditory and speech research, enabling the isolation of specific cortical indices of sound and speech processing through carefully controlled experimental designs and linguistic manipulations. However, as introduced in Section 1.1, these experimental designs are far from how speech is perceived in the real world, and do not necessarily probe the cortical dynamics of the speech network as it would operate while listening to continuous naturalistic speech. In other words, what can be measured, and the interpretation attributed to it, depends on what speech is defined to mean (Peelle, 2012). As a result, recent research on the neurophysiological underpinnings of speech processing tends to rely on more naturalistic experimental paradigms, where participants are exposed to continuous speech, e.g. in the form of podcasts or audiobooks. This paradigm shift in experimental design allows researchers to study the neural speech network in conditions that more closely reflect its natural function, but it also introduces challenges in controlling experimental variables of interest. In this context, novel analytical tools (Crosse et al., 2016;

Di Liberto et al., 2024) have been adapted or developed to isolate the contributions of different levels of information and to reliably assess the relationship between brain activity and continuous speech.

2.3.4 Analytical tools to quantify neural tracking

How is neural tracking quantified in the context of continuous speech stimuli? One approach is to quantify it in terms of coherence among participants' neural signals, based on the assumption that relevant features of the stimulus will be consistently represented and evoke similar brain activity patterns. Inter-Subject Correlation (ISC) measures coherence as the correlation between the amplitude of neural activity time courses of different participants undergoing the same experiment (Nastase et al., 2019), e.g. listening to the same story (Madsen & Parra, 2022) or watching the same video (Hasson et al., 2004). Crucially, this metric is not simply driven by the physical properties of the stimulus, but it also reflects higher level cognitive processes, showing sensitivity to the subjective interpretation of an abstract animation (Nguyen et al., 2019) and even allowing to predict TV preferences of a crowd (Dmochowski et al., 2014). Across-subject coherence can also be quantified as the alignment in the phase of neural signals across participants (Puschmann et al., 2021).

Another possibility is to rely on single-participant metrics, aimed at quantifying the relationship between an individual participant's brain activity and the stimulus. As for the across-subject metrics, also in this case analytical tools can focus on measuring either the amplitude or phase of neural signals, with different approaches. Methods focussing on phase estimation quantify the consistency of the phase of neural signals across trials (Inter-Trial Phase Coherence – ItpC) (Koerner & Zhang, 2015) or the consistency of the phases of neural signals and stimulus dynamics (Phase Coherence or Phase-Locking Value – PLV) (Lachaux et al., 1999; Zoefel & VanRullen, 2016). The brain-stimulus relationship can also be quantified in bits, as the amount of information, in an information-theoretic sense, that is shared between them. This Mutual Information (MI) approach avoids making explicit assumptions about this relationship and as such it is capable of accounting for non-linear dynamics (De Clercq et al., 2023).

It is also possible to assess the brain-stimulus relationship by cross-correlating the two signals (Petersen et al., 2017) or performing back-to-back regression (King et al., 2020). One popular approach is to employ (typically linear) encoding or decoding models that either predict neural responses based on the stimulus features (*encoding* or *forward* model) or reconstruct stimulus features based on neural responses (*decoding* or *backward* model). In this

context, neural tracking is quantified as the correlation value between predicted and actual brain activity or speech features (Crosse et al., 2016; Crosse et al., 2021). There is also a bidirectional approach, Canonical Component Analysis (CCA), where both the stimulus and the neural data are projected onto a common space of pairs of maximally correlated canonical components (CCs) (de Cheveigné et al., 2018). Encoding and decoding models can be non-linear, as when they are implemented with neural networks (Thornton et al., 2022). Despite their ability to capture non-linear dynamics, it is more difficult to interpret these models' behaviours in terms of neurophysiological phenomena.

2.3.5 The Temporal Response Function (TRF) framework

Most of the encoding and decoding approaches treat the human brain as a linear and time-invariant system, where speech features are considered the input and brain activity patterns the resulting output. This is a gross oversimplification, especially in the domain of speech processing, because comprehending speech requires prediction, integration of information at multiple timescales, and a certain degree of perceptual invariance. This assumption is useful nonetheless, because it reduces the problem of describing the complex brain-speech relationship to the simpler task of identifying the impulse response of the system. In a linear system, the impulse response is the characteristic response to a transient input (an impulse), and it fully captures the system's temporal dynamics.

In the context of speech processing, the impulse response is a time-locked and feature-specific mapping to neural activity. The most important advantage of the approach lies in its neurophysiological interpretability, both temporally and spatially. The mapping is interpretable in time as it represents the effect that a particular speech feature at time *t* has over subsequent time points, similarly to an ERP. It is also interpretable spatially as it is computed for each electrode separately, and it can thus describe where the response is localised in the brain.

On the other hand, it is still possible to model the brain-speech relationship as a non-linear system, for example with neural networks. This approach is theoretically closer to the reality of how the brain operates, but it presents significant challenges, particularly in terms of model definition and neurophysiological interpretability. While a linear system can be fully described by its impulse response, the identification of the precise non-linear operations involved in the neural dynamics underlying speech processing is not as feasible. As a result, non-linear models like neural networks incur in the *black box* problem: they are less tied to precise neurophysiological operations and thus less interpretable.

In this trade-off of interpretability and adherence to biological principles, linear systems are an approximate yet simple and computationally tractable way to model an intricate problem.

Once the impulse response of a linear system is estimated for a certain input at time *t*, it can be used to compute the system's output for any input (*encoding* or *forward modelling*) or conversely to reconstruct any input from the system's output (*decoding* or *backward modelling*). In this context, the correlation between predicted and actual data represents the metric of evaluation of the neural tracking of one or more speech features. In the framework adopted in this work, the estimated impulse response of the system is referred to as the Temporal Response Function (TRF) (Crosse et al., 2016; Crosse et al., 2021). The forward model can be represented as:

$$r(t,n) = \sum_{\tau} w(\tau,n)s(t-\tau) + \varepsilon(t,n),$$

where $r(t,n)$ is the neural response at time t and channel n, and $s(t-\tau)$ is the stimulus (e.g., speech envelope), at a time lagged by τ. The model is estimated by summing over a range of time lags, because the relationship between any stimulus feature and the resulting brain response is not instantaneous, and the current response at time t is impacted by past stimulus values. The TRF, $w(t,n)$, is essentially a filter consisting of the optimal set of weights that describe the linear transformation of the stimulus at lag τ into the neural response at time t and channel n. There is also a noise component, $\varepsilon(t,n)$, for each time point and channel, representing activity that is not explained by the model.

The TRF weights are estimated by minimising the mean-squared error between the neural response $r(t,n)$and its prediction $\hat{r}(t,n)$:

$$\min\big(\varepsilon(t,n)\big) = \sum_{\tau}[r(\tau,n) - \hat{r}(t,n)]^2.$$

This minimisation problem is resolved by applying reverse correlation, which can be implemented as:

$$w = (S^T S)^{-1} S^T r,$$

where S is the matrix containing time-lagged versions of the stimulus feature s, with dimensions $T \times \tau_{window}$:

$$\begin{bmatrix} s(1-\tau_{min}) & (-\tau_{min}) & \cdots & s(1) & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots & s(1) & \dots & \vdots \\ \vdots & \vdots & \dots & \vdots & \vdots & \dots & 0 \\ \vdots & \vdots & \dots & \vdots & \vdots & \dots & s(1) \\ s(T) & \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ 0 & s(T) & \dots & \vdots & \vdots & \dots & \vdots \\ \vdots & 0 & \cdots & \vdots & \vdots & \dots & \vdots \\ \vdots & \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & s(T) & s(T-1) & \dots & s(T-\tau_{max}) \end{bmatrix}$$

For each time point t of the stimulus feature, $s(t)$, ranging from 1 to T, each row of S represents a time-lagged segment of the stimulus centred around the current $s(t)$, including past and, possibly, future values of the stimulus (positive and negative lags). Each column of S represents a time lag applied to the stimulus, ranging from 1 to $\tau_{window} = [\tau_{max} - \tau_{min}]$. Due to the shifting in time of the stimulus feature, there are instances where the stimulus values are absent, and in that cases zero-padding is applied, in order to maintain causality.

The variable r represents the neural signals, and it is organised in a matrix of shape $T \times N$, with time points as rows and channels as columns. Therefore, the resulting TRF w is a $\tau_{window} \times N$ matrix, where each column reflects the temporal mapping from the stimulus feature s to the neural response at each channel.

The idea behind the formula is to project the neural response matrix r onto a space spanned by the stimulus features in S, to assess how each feature impacts the neural responses. However, stimulus features in S might be highly correlated, especially due to the overlap of time lags in a non-random signal such as speech. To correct for these correlations in the stimulus space, the auto-covariance matrix $(S^T S)$ of the lagged stimulus features needs to be inverted. Matrix inversion can be problematic when the matrix has linearly dependent columns (i.e., it is singular), or the data is noisy, causing the model to capture random noise rather than meaningful structure (i.e. overfitting). To address both of these problems, a small regularisation term, λ, is introduced to stabilise the inversion, adding it to the diagonal of the auto-covariance matrix before inversion. This method, known as ridge regression or Tikhonov regularisation, ensures that the matrix is invertible and less prone to instability:

$$w = (S^T S + \lambda I)^{-1} S^T r$$

The λ term that maximises the correlation between actual and predicted neural responses is selected using cross-validation.

2.4 The neurophysiology of auditory attention

Rather than being simply a bottom-up response to auditory stimuli, neural tracking of speech is modulated by higher-order cognitive processes, including attentional focus. Leveraging such attentional bias in the neural tracking response, it is possible to investigate realistic speech perception scenarios, such as how we selectively focus on a talker of interest while others are speaking at the same time.

The *Cocktail Party* problem (Cherry, 1953) has been studied since the second half of the twentieth century, with pioneering behavioural experiments probing listeners' ability to separate two messages presented simultaneously, either mixed or delivered to right and left ears separately. These very first studies using continuous speech already highlighted phenomena that are considered relevant for the processing of multi-talker scenes even to this day, such as the contribution of statistical and semantic information to the successful attentional allocation, the contrast of voices with different pitch (male vs. female) and different languages (English vs. German), the importance of temporal delays to merge two speech messages into one single percept during dichotic listening. This line of experiments concluded that speech out of the focus of attention could be easily ignored, to the point that listeners failed to detect a change of language in the ignored message or distinguish whether it consisted of normal speech or time-reversed speech, let alone recall its content. This work also introduced the notion of an *attentional filter* gating the processing of the ignored message, which was central to early theories on how selective attention to speech operates. These theories differed in the proposed stage at which they located the attentional bottleneck, resulting in different views on the extent to which ignored speech was processed, with early-filtering, late-filtering or attenuation models (Broadbent, 1965; Deutsch & Deutsch, 1963; Treisman, 1964).

However, between the 1960s and 1990s, research on attention became largely dominated by the field of vision, while auditory attention studies focussed on the separation of simple sounds especially at the level of the auditory periphery, examining the role of directional, spatial, and spectro-temporal cues (French & Steinberg, 1947; Freyman et al., 1999), along with processing advantages resulting from binaural hearing cues (Bronkhorst & Plomp, 1988; Durlach, 1972). This body of work mostly investigated how speech or non-speech sounds interfere with each other in terms of their physical properties, a phenomenon known as *energetic masking* (Arbogast et al., 2002), and how speech can be successfully perceived amid various kinds of unwanted noise. This resulted in the development of behavioural tests to predict a listener's ability to detect speech (ANSI, 1997; French &

Steinberg, 1947; Kryter, 1962; Nilsson et al., 1994), i.e. speech-reception threshold (SRT), which became particularly useful in the context of audiological testing and are still being constantly improved (Zaar et al., 2024). Along with energetic masking, researchers also introduced the broad notion of *informational masking* (Arbogast et al., 2002), defined as an interference resulting not from overlapping physical properties of sounds, e.g. their relative loudness, but from cognitive factors, such as similarity or confusion between the target and the masker signals. For example, attending to speech against a background of white noise is an instance of energetic masking, while when the interference consists of another intelligible speech stream, it is a case of informational masking. Therefore, informational masking refers to a more cognitive source of the interference in sound separation, rather than to the auditory periphery.

Some EEG studies investigated the effects of auditory selective attention on cortical responses using ERP paradigms, showing a modulatory effect due to the attentional focus. However, some of those experiments did not involve speech stimuli, but simple tones of different frequencies in deviant-detection tasks (Hansen & Hillyard, 1980; Näätänen et al., 1978) or the detection of synthetic sounds embedded within competing speech streams (Hink & Hillyard, 1976). Sometimes, no competing sound was included and, instead, ERP components were compared while participants attended to the detection task versus when they focussed on an irrelevant task, like reading a book (Picton & Hillyard, 1974).

Concepts from the visual attention domain, such as the contrast of exogenous and endogenous attention mechanisms (Posner, 1980), were transferred to the auditory attention domain (Mondor & Zatorre, 1995; Spence & Driver, 1994), but again with simple paradigms. From research on vision also came the notion of the integration of different features, forming separate *objects* (Treisman & Gelade, 1980). However, in the auditory domain, the definition of what constitutes an object is complicated by the temporally evolving nature of the stimulus, with contrasting views on whether auditory objects are formed automatically and then selected based on attentional goals (Bregman, 1990), or rather they are formed only through attentional binding mechanisms (Cusack et al., 2004). In realistic speech listening conditions, both of these mechanisms likely play a role, with different contributions depending on the characteristics of the auditory scene (Elhilali et al., 2009; Shamma et al., 2011; Shinn-Cunningham et al., 2015). It is only in the past fifteen years that research has begun to systematically examine the neurophysiological basis of selective attention in naturalistic scenarios of competing speakers (Lee et al., 2014). By using continuous listening paradigms where participants are instructed to sustain attention on one speaker and ignore other competing speakers, it became feasible to

explore the neural tracking of attended (target) and unattended (masker) speech in multi-talker scenes.

This work showed that the focus of a listener's attention could be accurately decoded from neural activity measured invasively with ECoG (Mesgarani & Chang, 2012) and non-invasively from single-trial EEG (O'Sullivan, Power, et al., 2015). A selectively enhanced representation of the target stream coexists with a parallel, though weaker, representation of the masker stream. Indeed, both target and masker are concurrently encoded in cortical signals as separate auditory objects, with a stronger representation of the target stream (Ding & Simon, 2012a; Power et al., 2012; Zion Golumbic et al., 2013), at critical temporal latencies (100 ms in MEG, 200 ms in EEG), and in higher-order cortical areas (Puvvada & Simon, 2017).
These seminal experiments all relied on the neural tracking of the speech envelope, since it is a readily accessible and informative feature of the speech signal. However, as we saw in Chapter 2.2, comprehending speech requires abstracting away from its physical properties to reach a higher level of symbolic meaning, naturally raising the question of how that speech hierarchy unfolds in the case of multi-talker scenes. Intuitively, the target speaker should be processed in full, through all levels of the speech hierarchy, since it is the focus of attention. However, the cortical representation of the target stream in a multi-talker scenario will be different from that of the same stream presented in isolation, due to the interference of the masker. Therefore, to solve the *Cocktail Party* problem, it is necessary to better characterise the extent to which streams out of the focus of attention are processed, and to localise attentional bottlenecks to specific speech processing stages.

Recent studies have started to clarify this aspect by investigating how higher levels of speech information are impacted by the focus of attention. They indicate that an attentional filter on the masker stream emerges quite early in the speech hierarchy, since phonological information is meaningfully encoded for the target but not for the masker stream (Teoh et al., 2022). Consistently with this, robust cortical responses to acoustic features are observed regardless of the focus of attention, while higher-level lexical and semantic properties are significantly represented only for the target speech (Anderson et al., 2024; Christian Brodbeck et al., 2018; Broderick et al., 2018). The extent to which these attentional filters are stable across diverse listening conditions and personal characteristics of the listener, like age, language competence, and hearing ability, is still an object of investigation.

In these experiments, listeners are instructed to direct and keep their attention selectively on one speaker, A, while ignoring a competing speaker, B, resulting in the classic *sustained attention* paradigm. In an ideally simplified interpretation of this scenario, one could

think that speaker A is uniformly and perfectly attended, while B is consistently ignored. If this were to be the case, the attentional gain would be time-invariant and well predictable. However, there are studies indicating that attention is not as static as in that idealised interpretation. For example, it was shown that phonological information of the ignored speaker is robustly represented in cortical responses when it can be glimpsed from the auditory mixture thanks to a momentarily favourable acoustic SNR between the two streams, indicating that rapid instances of attention reorienting do occur even in sustained attention scenarios (Raghavan et al., 2023). In another experiment, it was highlighted that around 70% of the listeners did not exhibit a significant attentional bias towards the target in the envelope tracking response, a finding that could not be explained based on low neural SNR or a lack of compliance to the sustained attention task (Kaufman & Zion Golumbic, 2023). In contexts where the masker speech is semantically salient, its cortical representation was found to be more robust than that of a less semantically salient target speech, suggesting that masker streams can undergo broad linguistic processing sufficient to get the gist of the unattended story (Park & Gross, 2023). In another study, the masker stream consisted of rhythmic syllable sequences ordered to form either coherent or random linguistic content. Despite sustaining attention to another naturalistic speech stream, cortical responses still encoded the phrasal structure of the linguistically coherent masker, which also produced greater interference with the representation of the target stream. This pattern of results did not hold when the structure of the masker was linguistically incoherent, despite similar acoustic and phonetic content, suggesting that listeners were indeed extracting phrasal information of a stream out of the focus of attention (Har-shai Yahav & Zion Golumbic, 2021).

Taken together, this body of literature indicates that sustained attention is a dynamic process, with attentional bottlenecks emerging at various stages of the speech processing hierarchy but differing in their strength and stage of occurrence. Uncovering the exact cognitive operations underlying attentional allocation is a complex task, particularly due to the difficulty in estimating the true internal state of the listener. When instructed to keep attention on speaker A and ignore B, does the listener consistently follow the instructions? The experiments described above suggest that there are situations where despite successfully focussing on the target, attention can be flexibly shifted to the masker or even, to some extent, directed towards both of the streams, when enough cognitive resources are available. Adjudicating between these two scenarios is nonetheless complicated, as mentioned above, by the difficulty in understanding the *ground truth* of the attentional state (Makov et al., 2023).

In real-world listening scenarios, switching and flexibly reorienting attention is as important as sustaining it on one particular sound. Firstly, the attentional system is always open to behaviourally relevant events, that capture attention through their bottom-up salience (Fiedler et al., 2025). For example, the sound of an object suddenly falling, as well as a person's own name (Cherry, 1953), will result in an instinctive and automatic reorienting of auditory attention. Secondly, attention can be reoriented at will between different sound sources, responding to top-down attentional goals. This voluntary redirecting of attention is particularly relevant for understanding speech in complex real-world listening conditions, where we often shift our attention between multiple simultaneous speakers, or we have to keep track of multiple conversations unfolding at the same time. Initial interest in the dynamic nature of reorienting auditory attention was once again motivated by efforts to establish parallels with the visual domain, and test for the existence of a supra-modal attention reorienting network (Huang et al., 2012; Larson & Lee, 2013; Shomstein & Yantis, 2006). These experiments identified the engagement of posterior parietal cortices in cued auditory attention switching and demonstrated the existence of a shared attentional network involved in both auditory and visual attention switching. However, these studies used very simple stimuli, mostly consisting of letters and digits uttered by competing speakers and relied on tightly controlled experimental designs. As such, they did not allow for investigating the online dynamics of the attention switching process as it naturally occurs when disengaging from one speaker and turning attention to another competing speaker. Even the later *Cocktail Party* literature with continuous speech streams, as we have discussed, has mostly focussed on exploring the neural basis of sustained attention, with few exceptions (Haro et al., 2022). The studies presented in Chapters 4 and 5 in this thesis aim to address this gap in knowledge.

2.5 Cortical processing of speech across ageing and hearing loss

The neural mechanisms underlying speech perception and comprehension evolve throughout a person's lifetime. Listening to speech, especially in multi-talker scenarios, becomes more challenging for older adults (Schneider et al., 2002) due to increased hearing difficulties - sometimes not reflected in traditional audiograms, but more related to the processing of fine temporal information of speech (Parthasarathy et al., 2020) - and to broader cognitive differences impacting various domains, such as working memory, inhibitory control, selective attention (Salthouse, 2010).

In terms of neural responses to the speech acoustics, most of the studies on the neural tracking of the envelope reveal an exaggerated envelope tracking response in older compared to younger adults, across both single- and multi-talker scenarios and under various masking conditions (C. Brodbeck et al., 2018; Decruy et al., 2019; Herrmann, 2025; Karunathilake et al., 2023; Presacco et al., 2016a, 2016b; Zan et al., 2020). Higher envelope tracking is often observed in combination with lower performance on cognitive tasks probing attention and inhibitory control (Decruy et al., 2019; Presacco et al., 2016a), hinting at the potential connection of this phenomenon with the cognitive resources available to the listener. Explanations for the enhanced neural responses to the speech acoustics in older individuals have been related to the compensatory recruitment of additional brain regions (C. Brodbeck et al., 2018) or increased activation within one brain region (Peelle et al., 2010), and to a general hyperactivity of the auditory pathway resulting from reduced sensory input, and manifesting as a sort of compensatory gain, particularly in cortical auditory areas (Herrmann & Butler, 2021). The observation that higher envelope tracking in older adults is associated with better (Decruy et al., 2019) or poorer (Goossens et al., 2018) speech comprehension scores across different experiments invites caution on its interpretation: it might reflect heightened attention to the stimulus, increased listening effort, or a suboptimal use of cognitive resources. Furthermore, it should be noted that there also studies reporting no difference in neural tracking of acoustic features between younger and older listeners (Gillis et al., 2023; Tune et al., 2021), motivating further investigations to shed light on these mechanisms.

Beyond the processing of the speech acoustics, age-related differences also emerge in speech comprehension and production metrics and in the neural encoding of linguistic features, suggesting that older adults may be adopting different strategies to maintain effective speech processing abilities. At the behavioural level, older listeners are less accurate in replying to comprehension questions and reporting the final word of a sentence, and they generally display longer reaction times (Murphy et al., 2006; Pichora-Fuller et al., 1995) compared to younger listeners, especially when speech is presented against a noisy background (Presacco et al., 2016a). They also seem to benefit less from release from masking, at least in paradigms where they are required to listen to isolated sentences (Irsik et al., 2022). On the production side, the coherence of the discourse seems to decrease as speakers age, both locally and globally (Hoffman et al., 2018). A stronger tendency to go off-topic when prompted to speak about a particular theme also emerges in older compared to younger people (Mortensen et al., 2006). At the neurophysiological level, cortical measures of lexical processing, such as the N400 effect, have proven to be impacted by the listener's age, both in traditional ERP paradigms

where simple sentences were used, and in more recent experiments involving continuous speech stimuli and realistic listening scenarios. The main differences observed involve the latency and the amplitude of the N400 deflection[1], which generally appears delayed and smaller in amplitude (Joyal et al., 2020; Payne & Federmeier, 2018; Wlotko et al., 2010), when compared to that of younger participants. Some recent studies employing TRF methods offer a less clear picture on the effects of ageing on this neurophysiological measure, due to different experimental paradigms, models (univariate or multivariate) and choice of features to represent higher-order lexical information, such as lexical dissimilarity or surprisal. In these experiments, the N400 response has been reported either not to differ between older and younger adults in a multi-talker listening scenario (Mesik et al., 2021), to be delayed but of similar amplitude (Broderick et al., 2021), or to be delayed and of smaller amplitude (Gillis et al., 2023) while listening to monologues in quiet. Similarly, those studies have yielded mixed results on measures of neural tracking, painting a complex picture. Some studies have observed a decreased tracking response to linguistic features (Gillis et al., 2023), others an increased tracking (Mesik et al., 2021). Another study reported no difference between younger and older participants when using word surprisal as a regressor, while a difference emerged when employing word dissimilarity, with decreased tracking for older listeners (Broderick et al., 2021). This conflicting evidence highlights that the choice of linguistic features is important, along with the necessity to account for the contribution of the acoustics (e.g., Mesik and collaborators did not disentangle their linguistic tracking effects from the tracking of the envelope).

The factors proposed to be at play to explain these neurophysiological and behavioural differences include increased vocabulary size (Hoffman, 2019; Hoffman et al., 2018) and reduced ability to suppress irrelevant information (Dave et al., 2021; Mortensen et al., 2006) in older listeners. Indeed, a number of studies highlight the importance of semantic control, i.e. the ability to inhibit irrelevant semantic information, which seems to be less effective as listeners age, both in speech perception and production (Hoffman, 2019; Hoffman et al., 2018; Wu & Hoffman, 2022). To overcome these difficulties, older listeners seem to rely on the previous linguistic context to a higher extent than younger people (Aydelott et al., 2010; Irsik et al., 2022; Kathleen Pichora-Fuller, 2008; Rogers & Wingfield, 2015; Wang et al., 2018), and

[1] Here I am referring to N400 broadly, including the widely known ERP-N400 and its continuous counterpart TRF-N400.

to process speech based on stored semantic knowledge more than online lexical predictions (Broderick et al., 2021; Federmeier et al., 2002).

An additional layer of complexity in the interpretation of age-related neurophysiological changes in speech processing is that older individuals often display severe forms of hearing decline, resulting in clinically relevant hearing loss. As such, it is difficult to disentangle the effects of healthy ageing from those of hearing impairment. Similarly to their normal-hearing peers, older adults with hearing impairment also display an over-representation of the speech acoustics, with ample evidence for an increased neural tracking of the envelope while listening to monologues or in multi-talker scenarios (Lien Decruy et al., 2020; Fuglsang et al., 2020; Gillis et al., 2022; Goossens et al., 2018; Millman et al., 2017; Schmitt et al., 2022). Yet, there are also studies that have found different patterns of results, reporting no difference in neural tracking measures between older adults with normal hearing or hearing impairment (Presacco et al., 2019) and, in multi-talker scenarios, observing differences only in the target-masker bias, and not in the neural tracking of the target (Petersen et al., 2017). However, as noted by (Schmitt et al., 2022), these results can be reconciled with the rest of the literature if we consider differences in the participants' hearing levels across experiments, which are important due to the non-linear relationship between hearing loss and auditory hyperactivity (Herrmann & Butler, 2021), and the choice of different temporal processing windows over which the differences in neural tracking were calculated.

While the neural tracking of the envelope in listeners with hearing impairment has been extensively studied, less is known regarding the neural encoding of linguistic features in this population. Very few studies indirectly investigated the topic, without specifically addressing a population with hearing impairment, and found decreased neural representations of linguistic features (Bolt & Giroud, 2024; Gillis et al., 2023). However, it is difficult to isolate the effects of decreased hearing abilities from the effects of age and other cognitive factors, in those non-specific designs. Understanding how hearing deficits affect higher-level aspects of speech processing is nevertheless crucial to fully characterise the neural dynamics of listeners with hearing impairment, and this consideration motivated Study 1 of this thesis, presented in Chapter 3.

Listening effort

Comprehending speech, especially in challenging conditions, requires the listener to exert a certain amount of what is commonly referred to as *mental energy*. The demands of a particular

listening situation interact with the person's internal motivation to determine the amount of mental energy to allocate, defined as listening effort (Peelle, 2018; Pichora-Fuller et al., 2016). The demands of a listening situation are both external - e.g. acoustic quality of the speech signal, or its loudness compared to background noise - and internal to the listener - e.g. the person's cognitive profile, or proficiency in the language (**Figure 2.3**).

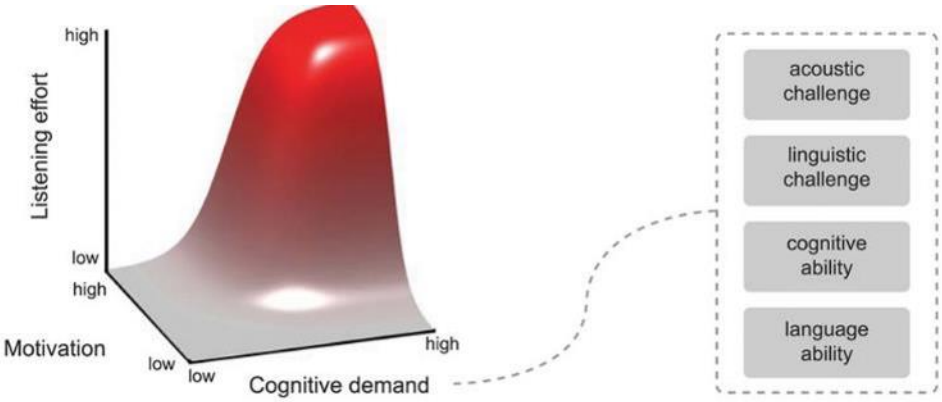


Figure 2.3: Schematic of listening effort. It results from the interaction of motivation and cognitive demands, which might be internal or external to the listener. (Peelle, 2018; Pichora-Fuller et al., 2016)

Behaviourally, decreased memory performance is associated with listening effort, since acoustically degraded speech is typically more difficult to remember. As such, speech repetition or recall tests have been used to quantify listening effort under various conditions (Pichora-Fuller et al., 1995). Performance in dual-tasks has also been used as a metric of listening effort, based on the idea that a system with limited capacity will show signs of effort in the form of reduced performance in an unrelated secondary task (Kahneman, 1973).

Physiologically, the pupil increases its diameter as a consequence of cognitive demands (Kahneman & Beatty, 1966), thus becoming a gold standard in the literature on effortful processing from various domains, including speech listening (Winn et al., 2018). Neurophysiologically, changes in the power of brain activity in the alpha frequency range (8-12 Hz) have been shown to reflect listening effort (Obleser et al., 2012). However, these various metrics of listening effort are not interchangeable, since they are often found to be independent of each other, and they most likely underlie different aspects of the phenomenon (Alhanbali et al., 2019). It also not clear how measurements of listening effort relate to ratings of self-perceived effort, thus making the connection between subjective and objective metrics complicated (Francis & Love, 2020).

Cognitive and hearing decline (as a consequence of healthy ageing or hearing loss) naturally impact the level of effort that the listener has to exert in order to reach successful speech comprehension, thus increasing the cognitive demands of the listening situation (Peelle, 2018). Many studies have reported measures of increased effort in these populations, resulting in listening fatigue (Dimitrijevic et al., 2019; Gagné et al., 2017; Barbara Ohlenforst et al., 2017; Paul et al., 2021; Zink et al., 2025). Despite providing adequate hearing compensation, evidence is still ambiguous regarding the effectiveness of hearing aids in reducing listening effort (Barbara Ohlenforst et al., 2017), with recent studies indicating a beneficial effect, mostly on pupillometry measures (Fiedler et al., 2021; Wendt et al., 2017). For these populations in particular, engaging speech material and content might actively lower the listening effort, since the level of motivation and engagement is tightly linked to the cognitive resources that the listener is willing to recruit, despite challenging listening demands (Herrmann & Johnsrude, 2020; Peelle, 2018).

Chapter 3. Study 1: Investigating phonological processing of competing speakers in hearing impaired listeners

3.1 Introduction

In multi-talker scenarios, the human brain faces the challenge of focussing on a speech stream of interest (target), while ignoring competing speech streams (maskers). This task is achieved almost effortlessly in many real-world scenarios, despite being a complex signal-separation challenge. However, masker sounds cannot be entirely ignored, as doing so would prevent us from switching our focus of attention when needed. Furthermore, attention on a stream is not monolithic: depending on the listening demands of the situation and the cognitive resources available, attention shifts to competing streams (Agmon et al., 2021; Kaufman & Zion Golumbic, 2023; Raghavan et al., 2023). Consequently, masker streams are also represented in neural responses, though at distinct cortical locations and time latencies (Ding & Simon, 2012a; Power et al., 2012; Puvvada & Simon, 2017). Yet, it remains unclear to what extent linguistic information from masker streams is processed, especially in listeners with hearing impairments (HI), who often experience increased listening effort and distractibility despite adequate hearing compensation (B. Ohlenforst et al., 2017).

In normal-hearing (NH) listeners, previous findings suggest that attention enhances neural encoding of the target stream across progressively higher levels of speech processing, from acoustic features like the sound envelope to phonemic and lexical features (Broderick et al., 2018; Di Liberto et al., 2021; O'Sullivan, Power, et al., 2015; O'Sullivan et al., 2019). The cortical representation of the masker speech, instead, seems limited to acoustic features, with a less reliable encoding of higher-level linguistic features, including phonological information (Brodbeck et al., 2020; Teoh et al., 2022).

However, the extent to which speech features of masker streams are processed is still unclear (Park & Gross, 2023; Raghavan et al., 2023), and how attentional selection operates in individuals with HI is even less well understood. Although some studies suggest similar attentional modulation of envelope tracking in both NH and HI individuals (Alickovic et al., 2020; Alickovic et al., 2021; Petersen et al., 2017; Seifi Ala et al., 2020), others report over-representation of target speech envelope (Fuglsang et al., 2020) and potential deficits in top-

down suppression mechanisms under challenging condition (Fiedler et al., 2019). As such, there is the possibility that HI listeners may also process masker speech differently, possibly encoding more strongly certain phonological information, which could potentially contribute to increased distractibility and effort in complex auditory environments.

This chapter investigates the degree to which individuals with HI encode phonological information for both target and masker speech, as revealed by EEG neural responses while listening to continuous speech in a multi-talker scenario. In doing so, we provide evidence that phonological encoding is stronger for target than for masker speech, reflecting attentional focus, consistently with results in the NH population. Furthermore, we sought to determine the degree of acoustic invariance of the phonological representation for both target and masker streams and found that an overall acoustic-phonetic representation was present regardless of attentional focus, and was robustly encoded even for the masker speech, potentially contributing to the elevated listening effort experienced by this population. The findings described in this chapter were presented at several international conferences and were published as: "Cortical encoding of phonetic onsets of both attended and ignored speech in hearing impaired individuals", *PLOS One*, November 2024.

3.2 Methods

3.2.1 Ethics Statement

Participants gave their written informed consent to participate in this study. The study was approved by the ethics committee for the capital region of Denmark (journal number H-1-2011-033). Data collection was carried out between October 1st 2019 and January 23rd 2020.

3.2.2 Data Availability

The consent given by participants at the outset of this study did not explicitly detail sharing of the data in any format. This limitation is in keeping with EU General Data Protection Regulation and is imposed by the Research Ethics Committees of the Capital Region of Denmark. These ethical restrictions prevent Eriksholm Research Centre from fully anonymising and sharing the dataset. To inquire on the possibility to access the data, please contact Claus Nielsen, Eriksholm research operations manager, at clni@eriksholm.com.

3.2.3 Participants

Thirty-four hearing-impaired individuals (10 females and 24 males, aged between 21 and 84 years, mean age = 64.2 years, standard deviation = 13.6 years) participated in this study. All participants were native Danish speakers recruited from the Eriksholm Research Centre Hearing Aid User Research Participant Database. This databased is managed by certified research audiologists and abides by the best accepted practices in diagnosis and management of hearing loss. All participants in this study cohort had previously been fitted with hearing aids and had at least four months usage experience. The cohort's hearing profile spanned from mild to moderately severe (bilateral average hearing threshold of 47.5 dBHL based on a 4-frequency threshold average at 0.5kHz, 1kHz, 2kHz and 4kHz) symmetrical (<10 dBHL in three adjacent frequencies) sensorineural hearing impairment. Hearing aids were fitted binaurally with hearing aids available in the market (Oticon Opn S 1^{TM} mini-Receiver-in-the-ear). The Voice Align Compression (VAC+) was used for amplification, offering a dynamic-range compression scheme with higher compression at lower input levels. This compression models, which is partly based on loudness-perception data by (Buus & Florentine, 2002), provides the hearing-aid user with a perception of the speech signal as close as possible to naturalistic conditions, amplifying softer sounds more than louder sounds. This is achieved by varying the amount of compression depending on the input's loudness, with an algorithm featuring low-compression kneepoints, i.e. input levels at which the gain for the amplification is reduced, between 20 and 50 dB SPL, depending on the frequency region of the sound and the user's hearing profile. The VAC+ scheme improves the listener's ability to clearly perceive distinct phonemes, and it particularly benefits the perception of certain phonetic categories characterised by high frequency and low intensity, such as fricatives and affricates (for example: "*th*", "*sh*", "*f*" sounds), which are greatly compromised in most forms of hearing loss.

Participants reported no history of neurological or psychiatric disorders and had normal or corrected-to-normal vision. (Alickovic et al., 2021). The wide age range of this experiment is justified by the presence of two participants below 50 years of age. Excluding those younger subjects (a 21-year-old and a 45-year-old), the age range becomes a more typical 54-84 years and, when re-running the analysis after excluding them from the dataset, the statistical results did not change. Detailed information regarding bilateral air conduction hearing thresholds, age and gender of the participants can be found in **Table 3.1,** ***Appendix***.

3.2.4 Experimental design

The experiment consisted of a simulation of a free-field cocktail-party scenario, with speech stimuli simultaneously delivered to the participants from an array of six loudspeakers placed around them (at ±30°, ±112.5° and ±157.5° azimuth relative to the participants' location). (**Figure 3.1 A**). Participants were instructed to selectively listen to a particular speech stream (target), corresponding to a specific talker and spatial location in the foreground, while ignoring all other talkers (masker and background babble noise). Target and masker speech stimuli were always played from the two frontal loudspeakers (*S1* and *S2*), while the four loudspeakers in the background (*B1-B4*) simultaneously presented a 4-talker babble noise each, creating a 16-talker background noise babble. Both target and masker streams were presented at 73 dB sound pressure level (SPL), while each of the noise babbles was delivered at a level of 64 dB SPL, resulting in a +3dB SNR of the frontal audio streams relative to the background.

Audio stimuli were presented at a sampling rate of 44100 Hz, delivered through a sound card (RME Hammerfall DSP multiface II, Audio AG, Haimhausen, Germany) and played through six loudspeakers Genelec 8040A (Genelec Oy, Iisalmi, Finland). EEG data was simultaneously acquired on a BioSemi ActiveTwo system with a sampling rate of 1024 Hz, from sixty-four electrodes, mounted on a standard cap according to the International 10/20 system (Klem et al., 1999). An active and a passive electrode, CMS and DRL respectively, served as reference to all electrodes, and two electrodes were placed on the mastoids.
This study presents a novel re-analysis of a previously existing dataset from Aličković et al. (2020, 2021). Note that only two out of the four experimental blocks were re-analysed as the rest of the data was not relevant to the goals of this study. As such, the results presented here were derived based on two blocks of twenty trials each, with a listening time of eleven minutes per block.

Despite having an identical selective-attention task, the two blocks differed in the perceived loudness of the background noise, as a consequence of turning on or off a noise-reduction (NR) scheme in the participants' hearing aids. In the NR ON block (Klem et al., 1999), the NR algorithm used a beamformer with a minimum-variance distortionless response to attenuate the babble noise coming from the background (Kjems & Jensen, 2012). In the NR OFF block, the NR scheme was deactivated and, as such, participants were completely exposed to the background noise. The intensity of the background noise is an important factor when examining the neural encoding of phonological information in HI listeners, especially in a context of selective attention. First, hearing loss introduces acoustic and cognitive challenges

to the demands of the listening situation, making HI listeners susceptible to listening fatigue, especially when the SNR of the target speech is lower (Fiedler et al., 2021). Secondly, the presence of background noise likely interacts with the VAC+ amplification scheme used in the hearing aids, increasing the overall loudness and potentially impacting the compression gains. As such, the presence of a louder background noise may directly impact the phonemes' perceptual discriminability by reducing the effectiveness of the VAC+ scheme. Given these considerations, the NR OFF – NR ON comparison was analysed here to assess the impact of background noise on the neural encoding of target and masker speech during the task of selective attention, with the expectation that the suppression of the noise babble in the NR ON block would have improved the separability of cortical responses to target and masker streams in the foreground.

Each 20-trial block was composed of four sub-blocks of five randomized consecutive trials for each of "left male (LM)," "right male (RM)," "left female (LF)," and "right female (RF)." Before each sub-block of 5 trials, a visual cue was presented on the screen located in front of the participants, indicating the sex and location of the frontal speaker they needed to attend to.

Participants underwent a familiarisation phase, with a brief training of 4 trials, one for each of the above combinations of target speakers' sex and location of presentation. Each trial started with the presentation of the background babble noise and, after 5 seconds, the frontal target and masker speakers were added to the auditory scene (**Figure 3.1 B**). Following the end of the trial, participants were required to answer a two-alternative-forced-choice question on the target-speech content to check for their sustained attention to the task. After each block, participants could take a self-paced break before continuing with the experiment.

Individual speech stimuli consisted of short Danish newsclip monologues of neutral content (33s each). To control for differences between male and female target talkers, stimuli were all root-mean-squared (RMS) normalized to the same overall intensity. All silences were shortened to 200ms, and the long-term average spectrum of the babble noise was matched to the overall spectrum of male and female foreground talkers to avoid inconsistencies in masking.

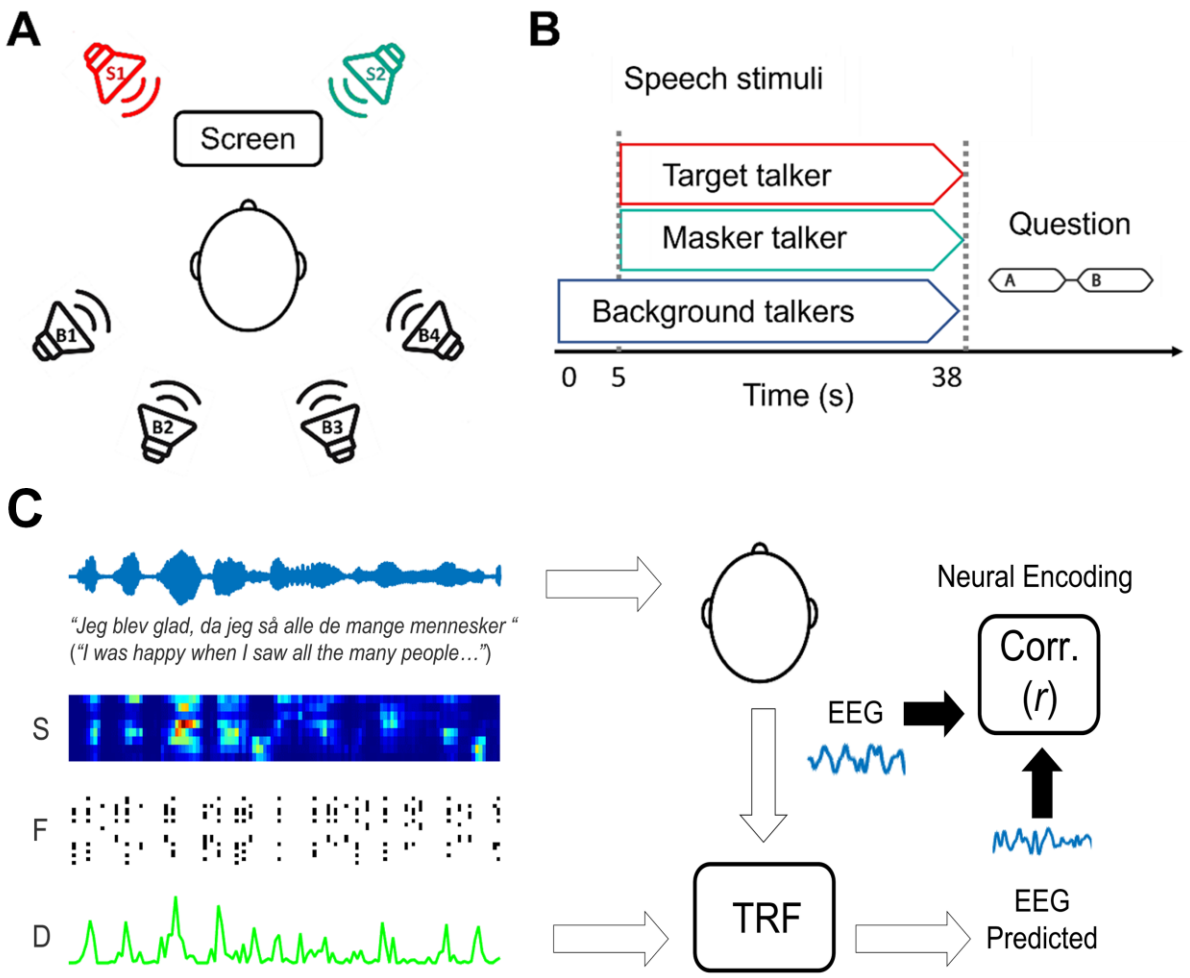


Figure 3.1: Experimental design and analysis framework (A) Location of loudspeakers (for this particular example, S1: target speech, S2: masker speech, B1-B4: babble noise) relative to the participant. **(B)** Timeline of each experimental trial. Start of the noise babble at $t_0 = 0$, followed by target and masker at $t_1 = 5$s. **(C)** Neural signals were recorded with EEG as participants listened to natural speech monologues. A lagged linear regression analysis was carried out to estimate the Temporal Response Function (TRF) describing the relationship between speech phonetic features (S - Spectrogram, F -Phonetic Features, D - Half-wave rectified Spectrogram Derivative) and low-frequency EEG (1-8 Hz).

3.2.5 Stimulus feature extraction

Here, we sought to determine how specific speech properties in this competing-talker attention scenario contribute to the neural-tracking response of target and masker speakers. Therefore, the foreground speech streams were modelled according to a set of features representing their low-level acoustic and phonetic characteristics. For the former set of features, we first extracted the stimuli's spectrograms by applying a filter which partitioned the sound waveforms into eight logarithmically spaced frequency bands, from 250 Hz to 8 kHz (Hamming window, with a length of 18ms, a shift of 1ms; FFT with 1024 samples), according to Greenwood's equation

(Greenwood, 1961). We also extracted, for each frequency band, its half-wave rectified spectrogram derivative. Together, these acoustic features are referred to as '**S**.'

A second set of features was derived to capture categorical phonetic features, describing the sounds typical of the Danish language in terms of articulatory features, e.g., voicing, manner of articulation, and place of articulation. This feature-set is referred to as '**F**.' First, speech stimuli were automatically transcribed, and the accuracy of the transcription was verified by a native Danish speaker. The transcripts were then processed by NordFA, a forced phonetic aligner for Nordic languages (Rosenfelder et al., 2014; Young & McGarrah, 2021), which identified timestamps corresponding to the start and end of each word into its constituent phonemes. The quality of the forced alignment was manually verified using Praat software on about 10% of the speech material, and with custom-made MATLAB scripts to assess that phoneme onsets corresponded to increases in the envelope, as expected in case of satisfactory phoneme alignment. Each phoneme was then represented as a binary vector with eighteen dimensions, each corresponding to one articulatory feature. Specifically, the following eighteen features were considered: syllabic, long, consonantal, sonorant, continuant, delayed-release, approximant, nasal, labial, round, coronal, anterior, dorsal, high, low, front, back, tense (**Figure 3.1 C**).

3.2.6 EEG data preprocessing

Neural data were analysed by using MATLAB software (MathWorks). Analysis code was customised based on publicly available resources (please see the CNSP initiative website: https://cnspworkshop.net). Minimal EEG processing was carried out (Di Liberto et al., 2024). EEG data were bandpass filtered between 1 and 8 Hz (Di Liberto et al., 2021; Zion Golumbic et al., 2013) using a zero-phase shift 4^{th} order Butterworth filter and then downsampled from 1024 to 64 Hz. Noisy EEG channels with a variance three times greater than that of the surrounding sites were replaced by means of a spherical spline interpolation, using a library in the EEGLAB Software (Delorme & Makeig, 2004). Subsequently, we re-referenced EEG signals to the average of the two mastoid channels. Single-subject EEG was also standardised by its overall standard deviation, preserving the ratio across electrodes.

3.2.7 TRF computation

We employed a system identification technique, known as the TRF framework, to estimate the relationship between neural signals and speech features, which can be conventionally approximated as a linear mapping. The TRF represents the optimal set of weights, obtained by

employing regularised linear regression (Crosse et al., 2016), describing the transformation from the stimulus features to the corresponding brain responses. To avoid overfitting, leave-one-out cross-validation was applied across trials, by exploring a parameter space spanning from 10e-6 to 10e4, in search for the optimal lambda, which was selected as the regularisation value yielding the highest Pearson's correlation between predicted and observed EEG data (forward model). Please note that in the case of multivariate speech representations consisting of both categorical and continuous variables (e.g., the FS model with spectrogram, spectrogram derivative and phonetic features) a banded regression was applied, in which the lambda search allowed for a different lambda selection for each set of features, aiming at the optimal combination of lambda values.

Considering that the relationship between any stimulus feature and the resulting EEG response is not instantaneous, TRFs are always estimated by taking into account multiple time-lags. In this case, we considered a time-lag window spanning the interval from -100 ms to 350 ms. Since the stimulus-EEG relationship is estimated for each electrode separately and defined by examining the stimulus effect on the EEG at multiple latencies, TRF weights are interpretable both spatially, through topographies representing the scalp locations where this relationship is stronger or weaker, and temporally, by examining how the stimulus' impact on the EEG signal evolves over time.

3.2.8 Speech feature models

The TRF framework was used to determine the relationship between low-frequency EEG signals and the phonetic information in speech. For both NR-ON and NR-OFF conditions, single-subject TRF models were fit for target and masker stimuli separately.

First, forward TRF models were fit considering the phonetic features (F feature-set), yielding the F model, and a control F*sh* model, with shuffled phonetic features (shuffled F feature-set). Comparing the predictive performance of the F and the F*sh* models, allows us to isolate neural variance that can be explained by the encoding of phonological information.

Next, we performed other TRF analyses to pinpoint the specific aspects contributing to the neural encoding of phonological information, particularly focussing on the acoustically-invariant encoding of phonetic categories, and on the neural encoding of phoneme boundaries, i.e., phoneme onsets. To investigate the cortical representation of phonetic categories, TRFs were fit by considering low-level acoustic features only (S feature-set), and a combination of F and S (FS feature-set combination).

By subtracting the EEG prediction correlations for the FS and S models we derived a metric called PhCat (also called FS-S in the literature), which quantifies EEG variance explained by phonetic features but not by the acoustic features, reflecting acoustically invariant responses. To analyse the neural encoding of phoneme onsets, which would reflect a representation of phoneme boundaries rather than their specific categorical identity, the TRF model fit was re-evaluated after a random shuffling of the phonetic features identity in the FS model, producing the F*sh*S model. Here, phonemes were shuffled while retaining the temporal information, by leaving phonemes' timestamps intact. The difference between the F*sh*S model's prediction correlations and those obtained from the S acoustics-model will represent the unique contribution of phoneme onsets. This prediction correlation gain, (F*sh*S-S), is called the PhOnset metric, because it quantifies the prediction gain due to phoneme onsets while accounting for the acoustics.

3.2.9 Phonetic distance maps

To determine how attentional selection alters the sensitivity to phonetic features, EEG phoneme-distance maps (PDM; Di Liberto et al., 2015; Di Liberto et al., 2021) were compared for the target and masker TRF weights of the PhCat metric (FS-S). The rationale is to project the TRF weights onto a multi-dimensional space where the Euclidean distance between phoneme pairs corresponds to the difference in their EEG responses. In doing so, these maps allow us to determine the sensitivity of the EEG signals to specific phonetic features. PDMs were obtained as follows: First, phoneme weights were calculated as a linear combination of the weights for the corresponding phonetic features. Then, a multi-dimensional scaling analysis (MDS) was carried out to reduce the dimensionality of the TRF weights, considering all time-lags and electrodes simultaneously, while preserving the standardized Euclidean distances between the EEG signals corresponding to different phonemes. The masker PDM was then mapped onto the reference target PDM, for each NR, by isomorphic scaling via Procrustes analysis (MATLAB function *procrustes*).

Four sets of phonetic categories were selected arbitrarily to reflect the major phonetic articulatory groups, ensuring that phonemes belonging to each set would create perceptually relevant phonetic contrasts defining the phonetic features: manner of articulation, place of articulation (lips), place of articulation (tongue), and voicing. For example, when considering the place of articulation (tongue) phonetic set, it is possible to explore the relative similarity of TRF weights in response to the phonetic features: high, low, back and front. If these phonetic contrasts, related to articulatory tongue movements, are perceptually relevant and represented

at the neural level, then the TRF weights in response to each of these phonetic categories should have consistently similar responses, while they would be consistently different across categories. Please note that the symbols used to represent phonetic features are an adaptation of the IPA symbols for Nordic languages. The '?' symbol indicates the Danish phonation feature "Stød", which is a type of glottal stop (Fischer-Jørgensen, 1989).

3.2.10 Statistical analysis

All statistical analyses directly comparing the groups were performed using repeated measures three-way ANOVAs. One-sample *t*-tests were used for post-hoc tests. Correction for multiple comparisons was applied where necessary via the Holm correction. In that case, the adjusted *p*-value was reported. Descriptive statistics for the neurophysiology results are reported as a combination of mean and standard error (SE). All analyses were performed using the JASP software (Love et al., 2019).

3.3 Results

Behavioural scores indicated that participants were able to successfully perform the task, displaying a 73% correct performance on 2-choice questions regarding the target speech for the NR OFF condition, and an 84% correct performance in the NR ON condition, with a significant impact of NR scheme on behavioural accuracy (Alickovic et al., 2021).

3.3.1 Robust phonetic-feature TRF for target but not masker speech in listeners with HI

Forward TRF models were fit to characterise the mapping between phonological information and low-frequency (1-8 Hz) EEG signals. Separate TRF models were fit for target and masker speech streams, first with a phonetic-feature model F and then with a control model, F*sh*, where the phonetic categories were shuffled, while preserving their timing. A repeated-measures three-way ANOVA was conducted to evaluate the impact of NR (OFF and ON), model (F and F_{sh}) and attention (target and masker) on the EEG prediction correlations. This analysis indicated a main effect of attention ($F(1,33) = 24.50$, $p = 2.14e\text{-}05$, $\eta p^2 = 0.43$), with larger EEG prediction correlations for the target compared to the masker speech ($r_{target} > r_{masker}$, post-hoc *t*-test: $t(33) = 4.95$, $p = 2.14e\text{-}05$, Cohen's $d = 0.68$); a main effect of model ($F(1,33) = 31.62$; $p = 2.93e\text{-}06$, $\eta p^2 = 0.49$), with the F model showing greater EEG prediction correlations than its control model, F*sh*, derived by re-running the TRF computation after shuffling the

corresponding phoneme categories ($r_F > r_{Fsh}$; post-hoc *t*-test: $t(33) = 5.62$, $p = 2.93e\text{-}06$, Cohen's $d = 0.54$); and a significant Attention*Model interaction ($F(1,33) = 18.82$, $p = 1.28e\text{-}04$, $\eta p^2 = 0.36$). The interaction effect revealed that an increase in EEG prediction correlation in the F model compared to its shuffled version F*sh* (rF > rF*sh*) emerged for the target stimulus (post-hoc *t*-test: $t(33) = 7.10$, $p = 6.98e\text{-}09$, Cohen's $d = 0.87$) whereas this effect did not emerge for the masker speech (post-hoc *t*-test: $t(33) = 1.77$, $p = 0.16$, Cohen's $d = 0.21$). Finally, we did not find a main effect of NR ($F(1,33) = 0.17$; $p = 0.68$, $\eta p^2 = 0.005$).

First, we report plots of the EEG prediction correlations (averaged across all EEG channels) and single-channel TRF weights (**Figure 3.2 B**) for the NR OFF condition. Statistical post-hoc tests showed that the F model yields significantly higher prediction scores than its shuffled F*sh* version for the target (post-hoc *t*-test: $t(33) = 5.11$, $p = 2.96e\text{-}05$, Cohen's $d = 0.85$), but not for the masker speech (post-hoc *t*-test: $t(33) = 0.65$, $p = 1$, Cohen's $d = 0.15$). Furthermore, an attentional effect emerged, since EEG prediction correlation values were significantly greater for the target compared to the masker, for both the F model (post-hoc *t*-test: $t(33) = 4.67$, $p = 1.96e\text{-}04$, Cohen's $d = 0.1$) and the F*sh* model (post-hoc *t*-test: $t(33) = 1.79$, $p = 1$, Cohen's $d = 0.38$); (**Figure 3.2 A)**.

Similarly, for the NR ON condition, the F model's EEG prediction correlations were significantly greater than those of the control F*sh* model in the case of the target stream (post-hoc *t*-test: $t(33) = 5.41$, $p = 9.14e\text{-}06$, Cohen's $d = 0.9$), while this result did not emerge for the masker (post-hoc *t*-test: $t(33) = 1.23$, $p = 1$, Cohen's $d = 0.2$); (**Figure 3.3 A)**. As for the NR OFF condition, also in the NR ON the accuracy of the EEG prediction was overall greater for the target compared to the masker speech, but for the F model only (post-hoc *t*-test: $t(33) = 4.8$, $p = 1.31e\text{-}04$, Cohen's $d = 1$), while in this portion of the data there was no difference between target and masker in the F*sh* model (post-hoc *t*-test: $t(33) = 1.54$, $p = 1$, Cohen's $d = 0.33$). TRF weights for channel FCz in the NR ON condition are visualised in **Figure 3.3 B.**

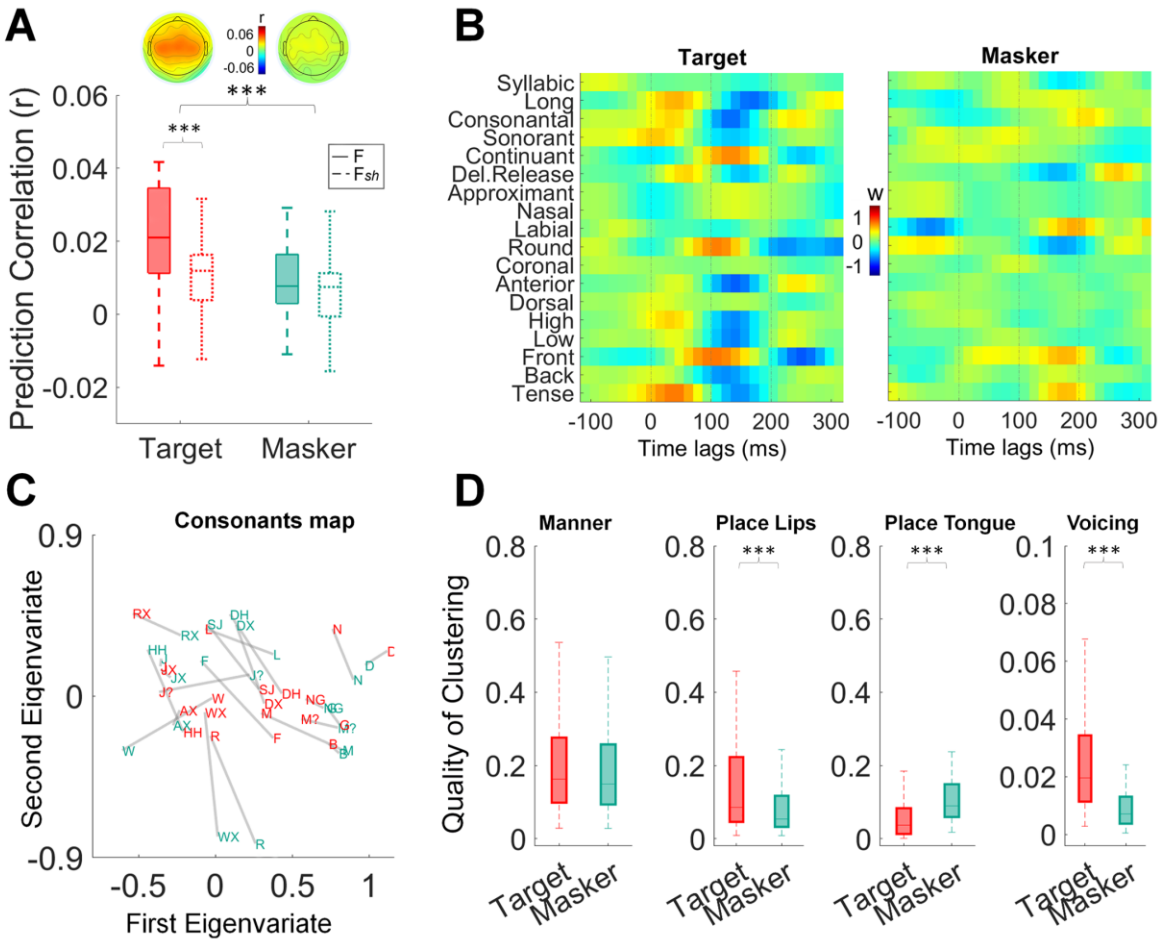


Figure 3.2: Stronger cortical encoding of target acoustic-phonetic information compared to masker information in listeners with hearing impairment in the NR OFF condition. (A) EEG prediction correlations (Pearson's r) for phonetic features models F and shuffled control F_{sh} of the target and masker speech. Boxplots show the median and inter-quartile range (IQR) of the distributions. Scalp topographies represent the distribution of prediction correlations across all channels (average across participants) for the F model. **(B)** TRF weights at channel FCz for the eighteen phonetic features for the F model. **(C)** Phoneme distance maps (PDMs) for target (red) and masker (green) speech. **(D)** EEG sensitivity to groups of phonetic features, i.e., quality of clustering of the EEG responses around relevant phonetic contrasts, for target and masker speech.

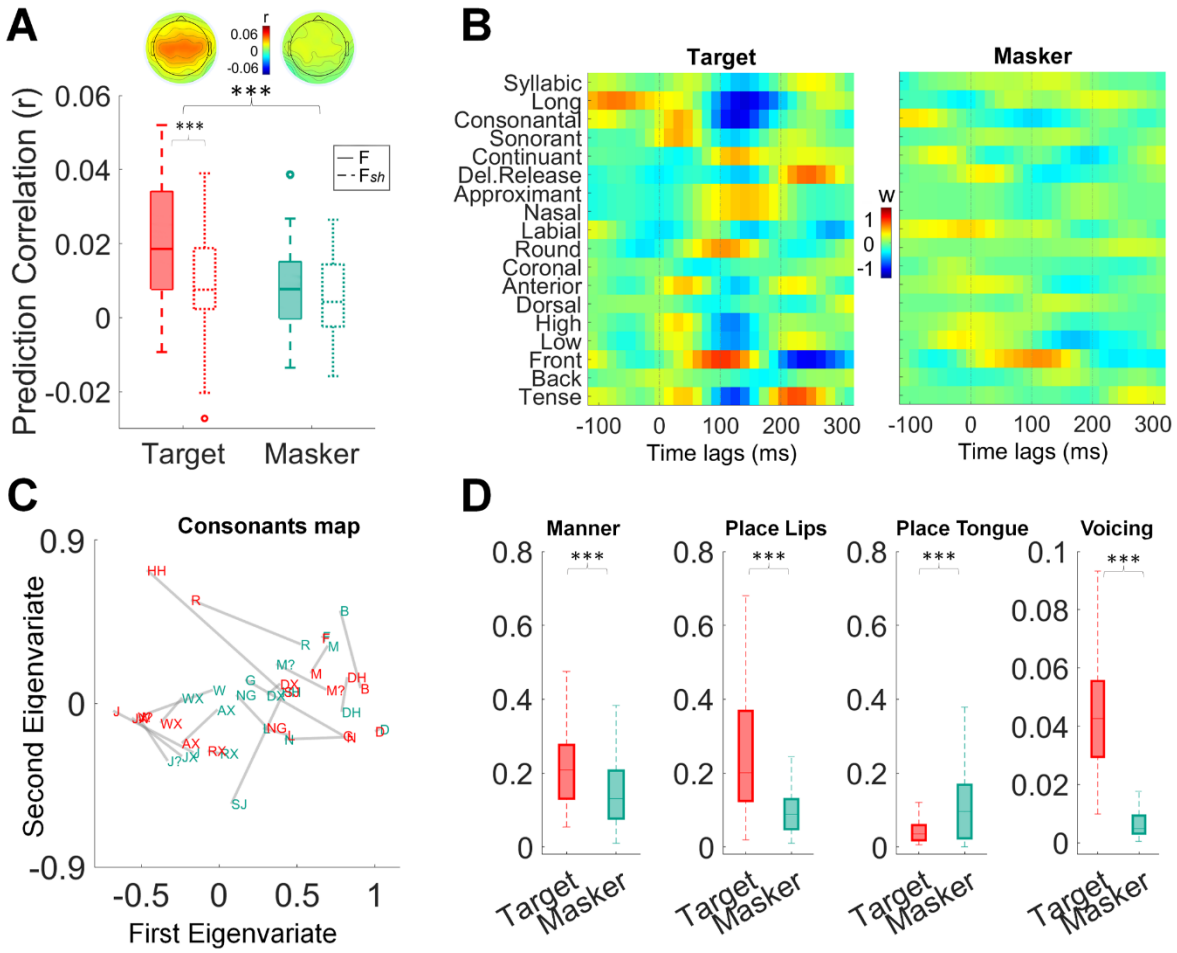


Figure 3.3: Target acoustic-phonetic information is more strongly encoded than the masker's information in listeners with hearing impairment when background noise is removed in the NR ON condition. (A) EEG prediction correlations (Pearson's r) for phonetic features models F and shuffled control F_{sh} of the target and masker speech. Boxplots represent the median and inter-quartile range (IQR) of the distributions. Scalp topographies represent the distribution of prediction correlations across all channels (average across participants) for the F model. Error bars indicate the SEM across participants. **(B)** TRF weights at channel FCz for the eighteen phonetic features for the F model. **(C)** Phoneme distance maps (PDMs) for target (red) and masker (green) speech. **(D)** EEG sensitivity to groups of phonetic features, i.e., quality of clustering of the EEG responses around relevant phonetic contrasts, for target and masker speech.

The results above indicate a robust relationship between EEG signals and phonological information of the target speech, while no such relationship is observed for the masker speech. Subsequent analyses were conducted to clarify whether attention had a more pronounced influence on the encoding of specific groups of phonetic features. To do so, PDMs reflecting the encoding of the target and masker speech were derived from the TRF weights of the target and masker F models separately (**Figure 3.2 C**; see **Methods; Figure 3.3 C**, for NR ON). Calinski-Harabasz metrics were derived to evaluate the quality of clustering for both the target and masker speech. These metrics measure the degree of separation between the relevant

phonetic feature clusters within the EEG-based PDMs. Larger values indicate that the organisation of PDMs aligns well with a specific group of features.

Figure 3.2 D summarises the results of the PDM analysis for the NR OFF condition indicating that, overall, the neural encoding of different phonetic feature sets is more consistently clustered in the target compared to the masker speech, while the impact of NR emerged as significant for some phonetic features sets, but not for others. These results emerged from a two-way repeated measures ANOVA – NR (ON and OFF) and attention (target and masker) - which was conducted four times, one for each phonetic feature set. Results of the PDM analysis for the NR ON condition are reported in **Figure 3.3 D.**

For manner of articulation, the analysis revealed a main effect of attention ($F(1,99) = 15.90$; $p = 1.29e\text{-}04$, $\eta p^2 = 0.14$), with a general effect of target > masker, (post-hoc *t*-test: $t(33) = 3.99$, $p = 1.29e\text{-}04$, Cohen's $d = 0.36$), while there was no effect of NR ($F(1,99) = 2.79$; $p = 0.09$, $\eta p^2 = 0.03$). In the NR OFF condition, there was no statistical difference between target and masker (post-hoc *t*-test: $t(33) = 1.74$, $p = 0.25$, Cohen's $d = 0.23$). In the NR ON condition, instead, the clustering of *manner*-related phonetic features was more consistent for the target speech compared to the masker (post-hoc *t*-test: $t(33) = 3.8$, $p = 9.61e\text{-}04$, Cohen's $d = 0.5$).

For place of articulation (lips), a significant effect of attention emerged ($F(1,99) = 59.55$; $p = 9.65e\text{-}12$, $\eta p^2 = 0.38$), with target > masker, (post-hoc *t*-test: $t(33) = 7.72$, $p = 9.65e\text{-}12$, Cohen's $d = 0.78$). A main effect of NR was also found ($F(1,99) = 12.47$; $p = 6.31e\text{-}4$, $\eta p^2 = 0.11$), with NR ON > OFF (post-hoc *t*-test: $t(33) = 3.53$, $p = 6.31e\text{-}4$, Cohen's $d = 0.37$), together with a statistically significant interaction of attention and NR ($F(1,99) = 5.62$; $p = 0.02$, $\eta p^2 = 0.05$). For the NR OFF condition, the clustering for the target speech was greater than that for the masker (post-hoc *t*-test: $t(33) = 4.10$, $p = 1.82e\text{-}4$, Cohen's $d = 0.56$). The same pattern of results emerged in the NR ON condition (post-hoc *t*-test: $t(33) = 7.3$, $p = 3.57e\text{-}11$, Cohen's $d = 1$).

Regarding the place of articulation (tongue) features, the ANOVA revealed once again a significant main effect of attention ($F(1,99) = 45.86$; $p = 9.15e\text{-}10$, $\eta p^2 = 0.32$), this time with a more effective clustering for the masker than for the target, (post-hoc *t*-test: $t(33) = 6.77$, $p = 9.15e\text{-}10$, Cohen's $d = 0.64$). For NR OFF specifically, it can be observed that masker > target (post-hoc *t*-test: $t(33) = 3.50$, $p = 0.002$, Cohen's $d = 0.50$). Similarly, in the case of the NR ON condition, the masker speech showed a more consistent clustering than the target (post-hoc *t*-test: $t(33) = 5.4$, $p = 1.19e\text{-}06$, Cohen's $d = 0.79$).

In the case of voicing, the 2-way ANOVA revealed a significant effect of attention ($F(1,99) = 221.96$; $p = 4.96e\text{-}27$, $\eta p^2 = 0.70$), with target > masker (post-hoc *t*-test: $t(33) =$

14.90, p = 4.96e-27, Cohen's d = 1.46), a significant effect of NR ($F(1,99)$ = 14.35; p = 2.61e-4, ηp^2 = 0.13), with NR ON yielding a better clustering than NR OFF (post-hoc t-test: $t(33)$ = 3.80, p = 2.61e-4, Cohen's d = 0.39), and a significant attention*NR interaction ($F(1,99)$ = 48.23; p = 4.03e-10, ηp^2 = 0.33). In the NR OFF condition, phonetic features related to voicing were more consistently clustered for the target compared to the masker speech (post-hoc t-test: $t(33)$ = 6.54, p = 1.03e-9, Cohen's d = 0.86). Similarly, the NR ON condition showed a significantly greater clustering for the target compared to the masker stream (post-hoc t-test: $t(33)$ = 15.76, p = 7.54e-36, Cohen's d = 2.07).

3.3.2 Cortical encoding of both target and masker phoneme onsets in listeners with HI

The TRF F model captured phonological information, potentially reflecting both an acoustically invariant cortical encoding of phonetic features as well as responses to other correlated information, such as the acoustic spectrogram (Teoh et al., 2022). To isolate EEG signals that can be explained by phonetic feature categories but not by speech acoustics, TRF models were fit based on an FS set of features comprising the S model (spectrogram and the half-wave rectified spectrogram derivative) and the F model (phonetic-feature categories). The gain in EEG prediction correlations due solely to phonetic features was then assessed by subtracting the FS and the S models, yielding the PhCat metric.

The PhCat metric (FS-S gain) may reflect the EEG encoding of the timing of phonetic feature categories, as well as their identity (Di Liberto et al., 2015; Di Liberto et al., 2021; Liberto et al., 2022). To assess if either or both these properties were encoded in the EEG signals, we also compared the FS-S predictive performance with the results of the same analysis after shuffling the phoneme identities in the transcript, which resulted in the PhOnset metric (F*sh*S-S gain). This PhOnset metric allowed us to isolate the contribution of phoneme onsets, while the PhCat metric represented the contribution of phonetic feature identity.

We compared the EEG prediction correlation gain obtained from the PhCat metric (FS-S) and the one obtained from the PhOnset metric (F*sh*S-S) using a repeated-measures three-way ANOVA with factors: NR (OFF and ON), model gain (FS-S and F*sh*S-S) and attention (target and masker). This analysis was done to dissociate the potential contribution of phoneme onsets from that of phonetic feature identity. Our analysis revealed no significant difference between the EEG prediction increase of the two model gains (PhCat and PhOnset metrics), suggesting that the contribution of phoneme onsets and phonetic feature identity is overall comparable (no

main effect of model gains: $F(1,33) = 1.98$, $p = 0.17$, $\eta p^2 = 0.06$). Furthermore, we observed a significantly greater EEG prediction increase for the masker speech, compared to the target, across the two model gains (significant main effect of attention: $F(1,33) = 11.26$, $p = 0.002$, $\eta p^2 = 0.25$. $r_{masker} > r_{target}$, post-hoc *t*-test: $t(33) = 3.36$, $p = 0.002$, Cohen's $d = 0.42$). The ANOVA results did not indicate a significant main effect of NR on the EEG prediction correlations ($F(1,33) = 0.09$, $p = 0.7$, $\eta p^2 = 0.003$). An interaction between attention and model gain did reach statistical significance ($F(1,33) = 4.26$, $p = 0.047$, $\eta p^2 = 0.11$), showing that the target speech PhOnset metric yielded significantly lower EEG prediction gains compared to both the masker PhCat metric (post-hoc *t*-test: $t(33) = -3.63$, $p = 0.003$, Cohen's $d = -0.5$) and the masker PhOnset metric (post-hoc *t*-test: $t(33) = -3.94$, $p = 0.001$, Cohen's $d = -0.56$).

When considering the NR OFF condition, no difference was found between the PhCat and PhOnset metrics, neither for the target speech (post-hoc *t*-test: $t(33) = 1.10$, $p = 1$, Cohen's $d = 0.15$) nor for the masker speech (post-hoc *t*-test: $t(33) = 1.05$, $p = 1$, Cohen's $d = 0.14$). The statistical analysis showed a greater predictive gain of phoneme onsets for the masker speech, compared to the target (post-hoc *t*-test: $t(33) = 3.24$, $p = 0.046$, Cohen's $d = 0.72$) (**Figures 3.4 A and 3.4 B)**. For the NR ON condition, we found the same pattern of results, with no significant difference between the PhCat and PhOnset metrics, neither for the target (post-hoc *t*-test: $t(33) = 2.38$, $p = 0.43$, Cohen's $d = 0.32$) nor for the masker stream (post-hoc *t*-test: $t(33) = 1.86$, $p = 1$, Cohen's $d = 0.41$) (**Figures 3.5 A and 3.5 B)** .

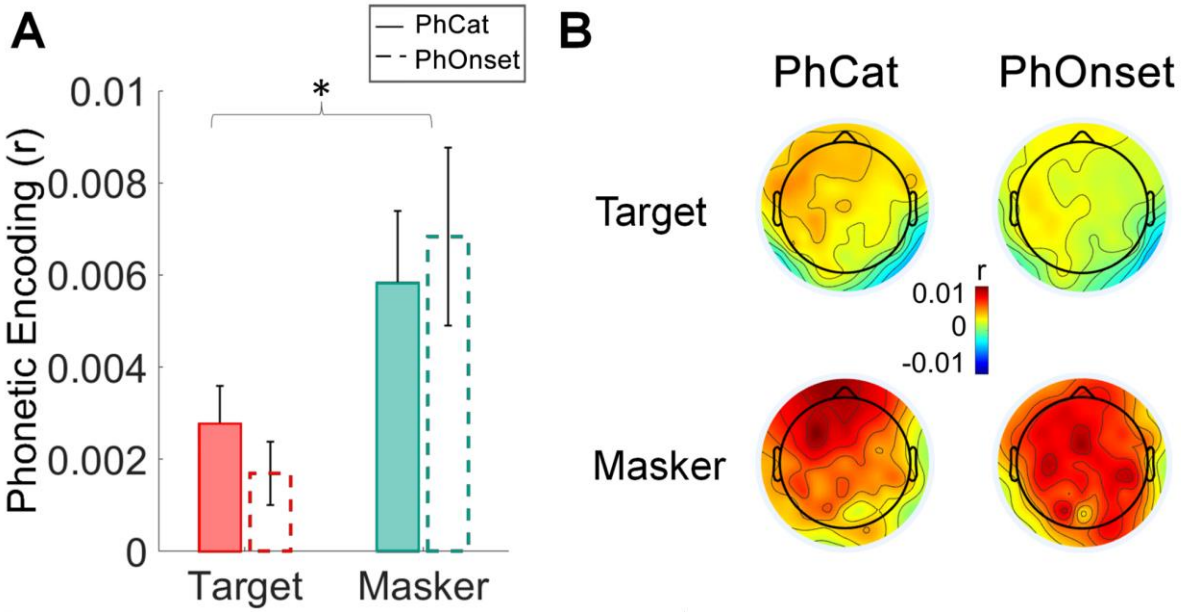


Figure 3.4: Representation of phonemic onsets of the ignored speech sounds in the low-frequency EEG of participants with HI in the NR OFF condition. (A) EEG prediction gains obtained from the PhCat (FS-S) and PhOnset (F_{sh}S-S) metric, for the target and masker speech. Bars represent the increase in prediction correlations (r) averaged across all participants and electrodes. Error bars represent the

SEM across participants. **(B)** Topographical distribution of the average EEG prediction correlation increases from the baseline model S, across all electrode locations.

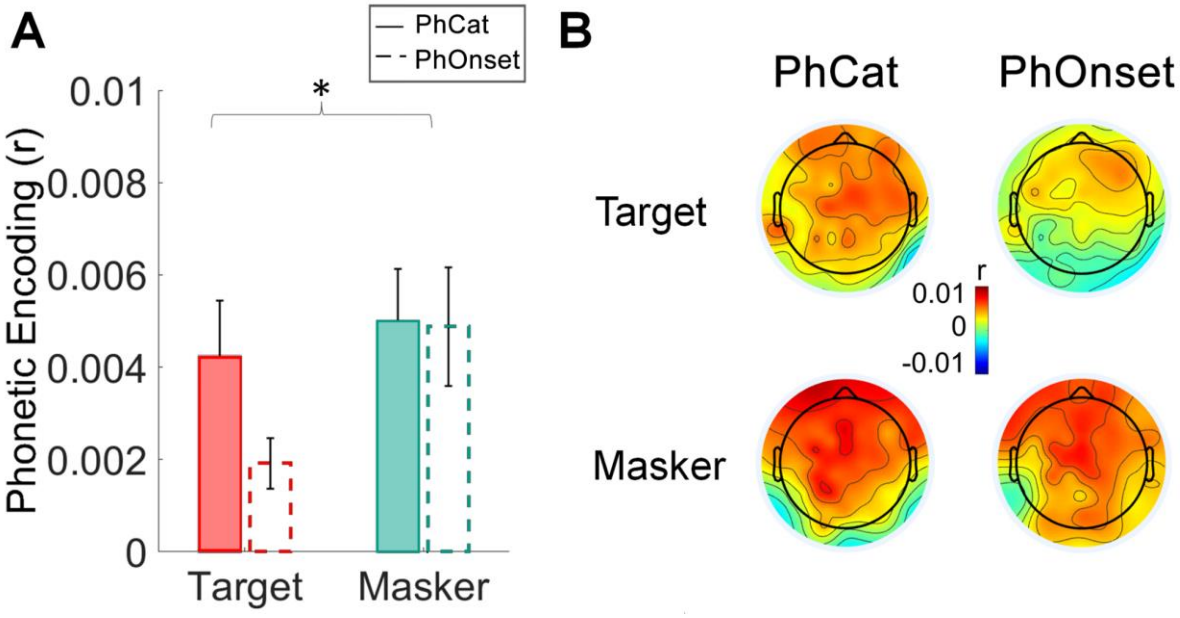


Figure 3.5: Representation of phonemic onsets of the ignored speech sounds in the low-frequency EEG of participants with hearing impairment when background babble noise is suppressed in the NR ON condition. (A) EEG prediction gains obtained from the PhCat (FS-S) and PhOnset (F_{sh}S-S) metric, for the target and masker speech. Bars represent the increase in prediction correlations (r) averaged across all participants and electrodes. Error bars represent the SEM across participants. **(B)** Topographical distribution of the average EEG prediction correlation increases from the baseline model S, across all electrode locations.

3.4 Discussion

Individuals with hearing impairment have difficulties in sustaining attention to the speaker of interest in multi-talker scenarios (Lunner et al., 2020). In this Chapter, we presented results on how attentional selection modulates the cortical encoding of phonological information in participants with HI, testing whether hearing impairment leads to a consistent cortical representation of the masker speech, resulting in high distractibility. Our results show two main patterns: First, we found that cortical responses in participants with HI relate to phonological information more strongly for the target than the masker streams, which is similar to the pattern previously measured in participants without HI (Teoh et al., 2022). Second, when isolating the different contributors to phonological information encoding, we found a significant cortical encoding of phoneme onsets of both target and masker speech, with a stronger representation of the masker.

3.4.1 Phonological cortical encoding of target and masker speech in listeners with HI

Our findings impact the current understanding of how selective attention strategies unfold in the brain of listeners with HI. The finding of a stronger relationship between cortical responses and phonological information of the Target speech suggests that individuals with HI perform attentional filtering at the level of the phonetic encoding, segregating target and masker streams and processing phonetic information of target and masker in a different manner. However, we could not isolate the acoustically invariant component of the phonetic responses here, meaning that the effect of selective attention on the phonetic feature EEG metric likely reflects a combination of acoustically variant and invariant responses to phonetic features. While possible explanations could relate to sample size and data quality, it should be noted that this is the first study measuring phonetic feature TRFs with the Danish language, while previous studies isolating acoustically invariant phonetic-feature TRFs used the English language (and Flemish), (Gillis et al., 2023). Future work could further investigate this question by considering participants with and without HI across different languages.

Our finding that cortex of individuals with HI encodes the phoneme onsets of both target and masker speech confirms the initial hypothesis of a highly interfering representation of phoneme-level information of the masker, which could contribute to explaining the high distractibility reported in the literature. Interestingly, the cortical encoding of phoneme onsets was stronger for the masker than the target. This result suggests that the onset encoding itself changes with the specific phoneme category for the target speech, reducing the consistency across different occurrences when those categories are shuffled in F*sh*. Conversely, the stronger phoneme onset encoding for the masker suggests higher consistency across phoneme categories, as it emerged from the analysis in **Figure 3.2 A**. Overall, these results indicate a significant encoding of phonetic distinctions for the target but not the masker, as well as a strong representation of the masker phoneme onsets, which could underlie participants' high distractibility in the presence of HI. In the context of selective attention, the cortical tracking of the masker phoneme onsets could serve two important functions: keeping a fundamental structural representation of the to-be-ignored stimulus for an effective re-orienting of attention on the masker stimulus at later stages, and building a structural representation of the masker speech, which is useful for suppressing it.

3.4.2 Phoneme onsets are encoded for both target and masker streams

It is worth noting that the robust cortical encoding of phoneme onsets in both target and masker speech streams - despite successful behavioural performance and neural attention bias at the level of overall phonological encoding - might also reflect fluctuations in participants' attention during the course of the trial. A recent electrocorticography study on NH listeners (Raghavan et al., 2023) found that the masker's phonetic features are neurally encoded when the target is silent enough to allow to glimpse the Masker out of the auditory mixture, suggesting that sustained attention on the Target is not necessarily continuous and stable. The fluctuation of attentional spotlight between target and masker might be even more pronounced in participants with HI, particularly in the presence of changing levels of target-masker SNR.

It is interesting to note that the EEG-measured phoneme-level encoding was not influenced by the hearing aid noise-reduction scheme, despite increased comprehension. The NR scheme reduced the sound level of the 16-talker babble noise in the background, without affecting the relative sound level between the target and masker speech streams originating in front of the speaker. As such, while NR ON made the task easier, increasing the overall comprehension (Alickovic et al., 2020; Alickovic et al., 2021), the listener's brain still had to segregate target and masker speech streams. Based on this result, we can speculate that the impact of attentional modulation measured at the phonetic level does not reflect overall speech comprehension, but a more general cortical strategy for selective attention in participants with HI. Instead, measuring overall speech comprehension might be effective by isolating higher order processes, such as lexical predictions (Gillis et al., 2021).

Previous work highlighted the key role of temporal information and phoneme onsets in the segregation of auditory streams in multi-talker scenarios. Temporal coherence was proposed as a key criterion for grouping acoustic information into separate streams (Elhilali et al., 2009; O'Sullivan, Shamma, et al., 2015). Previous work also found acoustic onsets of naturalistic speech to be faithfully represented in the auditory cortex (Daube et al., 2019; Hamilton et al., 2018), guiding the segregation of target and masker streams (Brodbeck et al., 2020). Acoustic-onset enhancement of the target speech signal has proven to benefit sentence recognition in cochlear implant users, even in the presence of competing speech (Koning & Wouters, 2012; Stilp & Kluender, 2010).

Furthermore, the neural encoding of acoustic onsets of the masker speech streams was suggested to serve as a template structure of the masker stream, contributing to its suppression (Brodbeck et al., 2020). Intriguingly, the late encoding of acoustic onsets of the masker in

fronto-parietal regions was also shown to be important for attentional selection, with a stronger encoding emerging when the masker stream was presented at a higher sound level than the target stream (Fiedler et al., 2019). This effect could not be studied with our experimental paradigm, which only modulated the SNR by reducing the sound level of the speech babble in the background (NR OFF vs. NR ON), not the relative sound-level of target and masker streams. Further work exploring different relative sound-levels could inform us on whether the late encoding of acoustic onsets measured in fronto-parietal areas is purely acoustic or related to phonological processing.

One of the main limitations of the present study is the lack of an age-matched control group, which would have allowed us to disentangle the effects of age and hearing loss. As such, while the current results inform us on the neural mechanisms of selective attention to phonological information in a population of HI listeners, they cannot be directly compared to previous studies investigating phonological encoding in younger individuals with normal hearing. To do so, it will be necessary to conduct tailored experiments probing attentional mechanisms across populations with normal hearing and hearing loss, while controlling for the potential confound of age. Future investigations could also examine the relationship between age and the degree of hearing loss with the effects measured in the present study, further explaining what the strong representation of the masker's phonetic onsets reflects exactly.

In conclusion, this study presented novel insights into the impact of selective attention on the neural encoding of speech in listeners with hearing loss. Our results complement previous work on attentional selection by measuring phonological neural encoding in a naturalistic multi-talker scenario. The finding that target and masker streams are segregated in the human cortex is in line with models of object-based auditory attention (Shinn-Cunningham, 2008; Shinn-Cunningham & Best, 2008) and complements previous evidence on the encoding of masker speech streams (Brodbeck et al., 2020), informing us for the first time on the neural underpinnings of this process at the phoneme level in participants with hearing impairment.

While these results shed new light onto the neural encoding of phonological information in a population of listeners with HI, they were obtained with a sustained attention paradigm, which requires participants to continuously maintain their attention on the target and ignore the masker. An equally important but less studied aspect of attention to speech is the ability to switch attention between competing speakers, which could be strongly affected by hearing deficits. The first steps towards studying the flexible reorienting of attention consist in establishing a tailored experimental paradigm, defining metrics that are sensitive to

neurophysiological changes, and an appropriate analysis framework. In the next Chapter, I describe a study focussing on the neural dynamics of attention switching.

3.5 Summary

In summary, this chapter examined how selective sustained attention affects the cortical encoding of phonological information in individuals with hearing impairment (HI) during multi-talker listening. The results reveal two patterns. First, participants with HI exhibited stronger cortical tracking of phonological features for the target speech stream compared to the masker, suggesting that attentional filtering at the phonetic level remains at least partially preserved in this population. Second, a robust encoding of phoneme onsets was observed for both target and masker speech, with the masker showing an even stronger onset representation. This finding supports the hypothesis that individuals with HI may maintain a persistent cortical representation of to-be-ignored speech, potentially contributing to the increased distractibility and listening effort reported in real-world scenarios.

Moreover, this phoneme-level encoding was not significantly influenced by a noise-reduction scheme, despite improved speech comprehension under NR-on conditions. This suggests that the neural tracking of phonetic onsets reflects broader attentional mechanisms rather than comprehension alone. The observed encoding of masker onsets may serve dual roles: providing a structural representation that allows for potential re-orienting of attention if needed, and aiding in the suppression of irrelevant input. Additionally, attention may fluctuate dynamically during the task, especially in challenging signal-to-noise conditions, possibly explaining the persistence of masker encoding even while actively sustaining attention on the target stream.

Chapter 4. Study 2: Simultaneous cortical tracking of competing speech streams during attention switching

4.1 Introduction

To understand speech in multi-talker environments, listeners must isolate a target speaker from competing streams (Power et al., 2012; Puvvada & Simon, 2017; Zion Golumbic et al., 2013). This selective attention process has been widely studied using cocktail-party paradigms (Bronkhorst, 2015; Cherry, 1953), showing how the brain segregates and transforms target speech into meaning, while suppressing masker speech to varying degrees (Har-shai Yahav & Zion Golumbic, 2021; Kaufman & Zion Golumbic, 2023; Rivenez et al., 2006; Teoh et al., 2022). Techniques such as EEG (O'Sullivan, Power, et al., 2015; Power et al., 2012) , iEEG (Mesgarani & Chang, 2012), MEG (Puvvada & Simon, 2017; Simon, 2015), and fMRI (Evans et al., 2016; Hausfeld et al., 2024), have pinpointed auditory cortical loci of segregation and revealed differences in processing target versus masker speech (Ding & Simon, 2012b; Park & Gross, 2023; Power et al., 2012; Teoh et al., 2022). However, most work has examined sustained attention, leaving the neural basis of attention switching largely unknown.

Dynamic switching paradigms are widely used in cognitive control research to study flexibility and stability (Egner, 2023), typically requiring participants to switch tasks based on changing instructions (Luo et al., 2025; Ravizza & Carter, 2008; Wager et al., 2004), with associated switch-costs manifesting as differences in reaction time or accuracy (Rogers & Monsell, 1995). Adapted versions of these paradigms have also been applied to speech in multi-talker settings (Lawo et al., 2014; Meister et al., 2020; Oberem et al., 2017), treating target selection as a task-switching problem, e.g., by spatial location or voice identity (Lawo et al., 2014), but not directly addressing dynamic attention reallocation between continuous speech streams in naturalistic scenarios. More recently, speech neurophysiology studies have begun to include attention switches, either to refine EEG/MEG decoding methods (Akram et al., 2016; Miran et al., 2018) or to examine sustained attention to moving auditory objects (Teoh & Lalor, 2019). However, to the best of our knowledge, only one previous study has specifically focussed on the neurophysiology of attention switching in multi-talker scenarios, relating the neural encoding of speech during attentional re-orienting with EEG alpha activity and pupil

dilation dynamics (Haro et al., 2022). Those findings proved that the neurophysiology of attention switching can be studied non-invasively. Building on that work, our study sheds light on the exact neural dynamics supporting the steering of attention between two competing speech streams, disengaging from the previous target stream while engaging to the new one.

Here, we measure the neural encoding of speech with high temporal resolution as listeners steer their attention from one speaker to another. We test whether engagement with a new speech stream begins before disengagement from the previous target is complete, resulting in a brief period of simultaneous tracking of both streams. Such an asymmetry in the disengagement-engagement processes, even if transient, could support the ability to explore alternative auditory streams while maintaining attention to a given stream (Agmon et al., 2021). The neural encoding of speech was measured from normal-hearing adult participants using EEG during an immersive multi-talker listening task. Participants were exposed to two competing streams from TED talks, presented via two front-facing loudspeakers, with additional speech babble noise from the background (**Figure 4.1 A**). An on-screen arrow cued participants to attend to one of the two speech streams and to shift their attention rapidly whenever the arrow changed direction, approximately every 10–30 seconds (**Figure 4.1 B**). Neural tracking of target and masker speech was quantified using the Temporal Response Function (TRF), describing the linear relationship between each speech stream and the neural responses. First, we confirmed that the attended stream could be reliably decoded from the EEG, consistent with the extensive literature on sustained attention (de Cheveigné et al., 2018; Mesgarani & Chang, 2012; O'Sullivan, Power, et al., 2015). This confirms that the EEG responses in this experiment reflects differential encoding of target versus masker speech (**Figure 4.1 C**).

We next addressed two fundamental questions. First, we asked whether the processes of engaging with a new speech stream and disengaging from a previous one unfold symmetrically (**Figure 4.2 and 4.3**). To test this, we measured the neural tracking of the two competing speech streams over time, by fitting temporally detailed encoding TRF models. This allowed us to characterise the average encoding dynamics surrounding attention switches, comparing disengagement and engagement processes. The second objective was to understand how our brains update and use lexical context when switching attention (**Figure 4.4**). We formulated four hypotheses and built quantitative predictions for each using a state-of-the-art large language model (LLM). This resulted in four TRF regressors reflecting lexical semantic features with a different sensitivity to previous context and switch occurrence. Encoding TRF models were fit for the four hypotheses, revealing the most physiologically plausible lexical

processing mechanism. This study provides substantial new insights into the temporal unfolding and contextual mechanisms guiding attention switching, encompassing both low and high levels of speech abstraction.

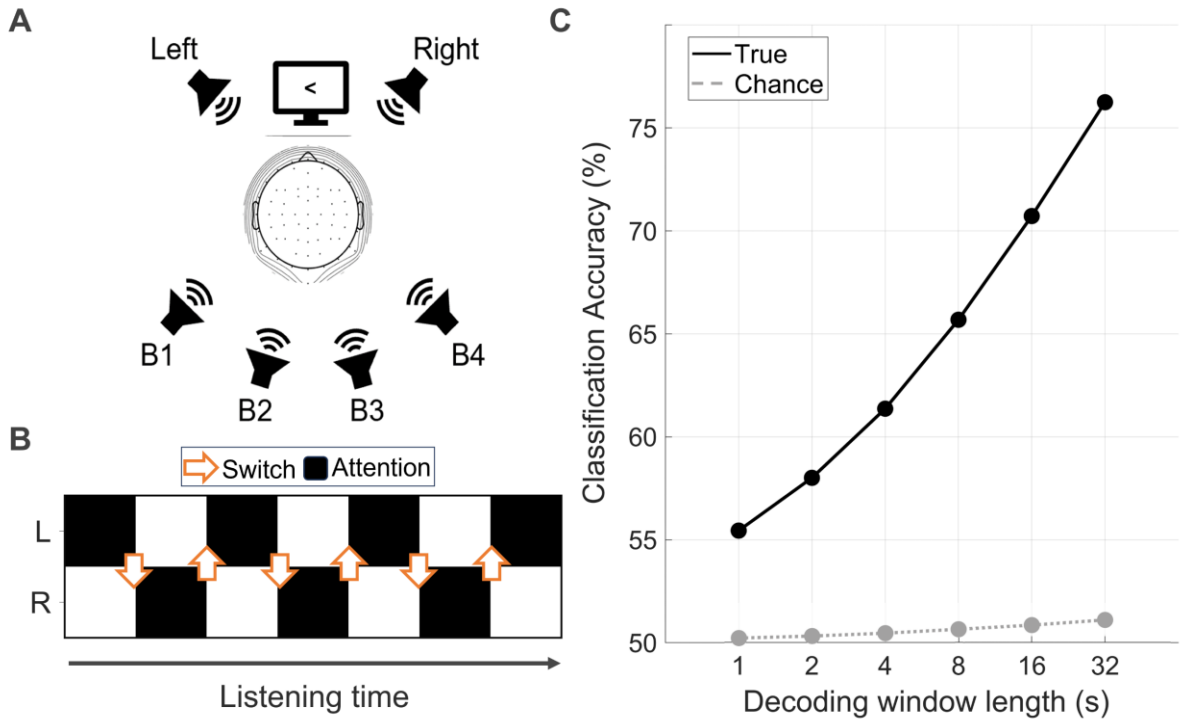


Figure 4.1: Experiment overview and validation. (A) Participants were presented with speech from two loudspeakers placed in front of them with 60° of separation (30° left and 30° right), and with concurrent 16-talker background noise (B1-B4). In each trial, the screen presented an arrow pointing to the target speech stream. Participants were instructed to switch attention as soon as the visual cue changes direction. **(B)** Schematic diagram of one experimental trial. The black area represents blocks of attention either to the left (L) or right (R) front streams. The red arrows indicate the instants where the attention cue switches side (six times per trial). Note that block duration was randomised and always between 15 and 30 seconds, with trials lasting 3 minutes. **(C)** EEG data validation was carried out by running an attention decoding analysis. Progressively longer decoding windows were considered (larger windows use more data, typically leading to more accurate decoding scores). Binary classification scores are reported arbitrating between the target and masker streams. The dashed line indicates the 95th percentile of a random distribution calculated by randomising the classification labels. Statistically significant attention decoding classification scores were measured for all the decoding windows considered, with numerical results comparable with previous studies on selective attention (Choudhari et al., 2024; de Cheveigné et al., 2018; O'Sullivan et al., 2017).

4.2 Methods

4.2.1 Participants and experimental procedure

For the present study, which was approved by the Ethics Committee of the School of Psychology at Trinity College Dublin, we recruited 24 young native English speakers (between

18 and 39 years of age), who gave their written informed consent. Participants had normal hearing, as per a screening pure tone audiogram from 0.25 Hz to 8 kHz and reported no history of neurological or psychiatric disorders and had normal or corrected-to-normal vision.

The experiment simulated a multi-talker scenario (**Figure 4.1 A**) with a circular array (1.50m radius) of six loudspeakers surrounding the listener (at horizontal angles of ±30°, ±112.5°, and ±157.5° relative to the participant). Participants were instructed to dynamically switch their attention between left and right speech streams in the foreground, following a visual cue (left- or right-pointing arrow) indicating the to-be-attended side, which was displayed at the centre of a screen placed in front of them. While switching their attention between the foreground streams, they were also asked to ignore a 16-talker noise played from the four loudspeakers in the background (B1-B4, each of them delivering a 4-talker babble). Frontal streams were presented at 60dB sound pressure level (SPL) each, while each of the noise babbles was delivered at a level of 54dB SPL, resulting in a 3dB SNR of the foreground relative to the background.

Participants were presented with twenty trials (lasting 180 s each) and had to perform 6 attention switches per trial, occurring at semi-random intervals (**Figure 4.1 B**). For this reason, blocks of sustained attention to one particular speech stream varied considerably in duration, spanning from 10s to 30s. For each trial, a different male and female speech stream was played from the left and right loudspeakers in the foreground, counterbalancing throughout the experiment for side of presentation and start of the attention block (i.e., the experiment consisted of five sub-blocks with the following trial sequence: Male Left – Attention Start: Left; Female Left – Attention Start: Left; Male Left – Attention Start: Right; Female Left – Attention Start: Right).

Each trial started with the visual cue pointing towards the to-be-attended side and background noise only, followed by the two foreground speech streams starting simultaneously after 5s. At the end of the trial, participants answered three multiple-choice questions about: the content of the speech streams, their preference based on personal interest (left or right stream), and the perceived difficulty of the attention switching task. The experiment flow was self-paced and, to minimise fatigue, it included three mandatory breaks, each lasting no less than five minutes, every fifth trial.

Speech streams were presented at a sampling rate of 44.1 kHz, delivered through a Roland Octa-Capture 10x10 sound card (24-bit/192 kHz), and played through six PreSonus Eris 4.5BT loudspeakers. Participants' EEG activity was recorded using a BioSemi ActiveTwo system at a sampling rate of 512 Hz, from sixty-four electrodes positioned on a standard cap

following the International 10/20 system. An active (CMS) and a passive (DRL) electrode were used as reference for all electrodes, and two additional electrodes were placed on the mastoids for offline referencing. For 21 of our 24 participants, we additionally recorded electro-oculography (EOG) and electro-myography (EMG). Two electrodes were placed on the left and right temples to capture horizontal eye movements, and two electrodes were positioned above and below the left eye to record vertical eye movements and blinks. To capture EMG activity related to head rotation, an electrode was placed on the left deltoid muscle. Please note that activity from these external electrodes has not been analysed as part of this study.

4.2.2 Stimuli

The foreground speech stimuli included forty TED Talks covering a range of topics, with 20 female and 20 male presenters, each speaking in a variety of English accents. All speech streams were root-mean-squared (RMS) normalised to reduce differences between male and female voices. Each of the 4-talker background babble signals was obtained by summing the audio signals of four separate TED talks. The long-term average spectrum of the babble noise was then adjusted to align with the overall spectrum of both male and female foreground speakers, to prevent inconsistencies in masking.

4.2.3 EEG data preprocessing

Neural data were analysed with custom scripts in MATLAB software (MathWorks), based on publicly available scripts and resources shared as part of the CNSP initiative (Cognition and Natural Sensory Processing; https://cnspworkshop.net). Neural signals were first band-pass filtered between 0.5 Hz and 8 Hz, using a zero-phase shift Butterworth filters of order 4, and then downsampled from 512 Hz to 64 Hz. Spherical spline interpolation was applied to replace channels that were three standard deviations away from the mean. EEG was then re-referenced to the average of the two mastoid channels.

4.2.4 Speech features

The current study aimed to characterise the neural tracking of a dynamic multi-talker scenario by measuring the relationship between EEG data and various features of the foreground speech stimuli, related to their acoustic and lexical properties. To model the speech acoustics, the audios' broadband amplitude envelopes were extracted by taking the absolute of the Hilbert transform. In order to model the lexical properties of speech, the transcribed stimuli and their audios were first automatically aligned using the WebMAUS Basic aligner (Kisler et al., 2017;

Schiel, 1999, 2015), which identified timestamps corresponding to the start and end of each word. The resulting automatic alignment was saved in the TextGrid format and adjusted manually, when necessary, using the Praat software (Boersma, 2007). The time stamps were then used to build binary word onset vectors in MATLAB. While binary word onset vectors represent information related to word segmentation, they can also be modulated according to each word's surprisal or entropy value to represent higher-order semantic information. Word surprisal is a measure of how unexpected a word is given its preceding linguistic context, and it can be computed using Large Language Models (LLMs), as the negative logarithm of the probability of that word given the previous context. Word entropy, on the other hand, measures how uncertain or unpredictable the next word is. Here, we used a pretrained open-source LLM, Mistral-7B-v0.1 (Jiang, 2023) to extract word probabilities, and then computed lexical surprisal and lexical entropy values for each word.

When considering the dynamic nature of the attention switching task, the definition of what context to include for the current word prediction becomes a non-trivial problem. From the machine perspective, given a word *w* belonging to, e.g., the left stream, the LLM would predict it more easily when provided with all the available linguistic context from the left stream. However, from the neural/behavioural perspective, since participants were flexibly re-orienting their attention between left and right streams, the optimal context would be impacted by the attention switch and, potentially, store information of previously attended blocks. To compare these context-accumulation alternatives, we represented context according to four alternative representations. A machine-ideal model, Oracle, considers as context all words preceding the current word in one particular stream, whether they were attended or unattended. As such, this model is *unaware* of the switch of attention. Among more neurally-plausible and switch-aware context accumulation models, we constructed the Attention model, which incorporates as context any previously attended block from both left and right speech streams, and the Speaker Specific model, which displays a speech stream bias, whereby only previously attended blocks of the same stream are included as context for lexical prediction of the current block. Similarly to Attention and Speaker Specific, the fourth model, Reset, is switch-aware, but not context-aware. In fact, it does not keep track of any previous speech block, neither attended nor unattended, and instead resets the context following each attention switch cue.

4.2.5 TRF and analysis procedure

In this study, we quantified neural tracking of speech features as the ability to predict unseen EEG data (encoding) or reconstruct unseen stimulus features (decoding), with a linear model

based on lagged linear regression, the Temporal Response Function (TRF). A TRF model was fit for each participant separately, using a leave-one-trial-out cross-validation procedure to avoid overfitting, and considering a range of time lags from -100ms to 600ms in the case of the encoding models, and between 0ms and 600ms for the decoding models. The TRF model calculation applied Tikhonov regularisation, with the regularisation parameter (λ) tuned via an exhaustive search across a logarithmic scale from 10^{-6} to 10^{4}. This search was conducted on the training fold during each iteration of the cross-validation process, and the optimal λ was selected as the parameter that yielded the highest correlation (Pearson's *r*) between true and predicted/reconstructed data.

We first constructed an Attended speech model from the combination of attended segments of left and right speech streams. We then fitted TRF models by coupling the EEG and the Attended stimulus features. With a decoding approach, we reconstructed the envelope of each trial based on the Attended TRF model and then correlated it with the true envelopes of left and right streams for the corresponding trial. Since the left or right true attention label was known throughout the trial, classification accuracy could be computed as the percentage of left or right segments correctly classified as attended (**Figure 4.1 C**). We also computed the null distribution of the classification accuracy by randomly shuffling the left and right labels for 100 iterations. We employed a range of decoding window lengths (1 s,2 s,4 s,8 s,16 s,32 s) to assess the impact of correlation window size on the model's ability to correctly classify the attended stimulus.

To observe the evolution of neural tracking of left and right streams over time with a neurophysiologically interpretable approach, we calculated an encoding TRF model where true and predicted EEG signals were correlated using a sliding window approach (MATLAB's *movcorr* function) with a step size of 0.125s. The encoding TRFs were fit with a multivariate set of attended stimulus features including envelope (Env), word onset (WO) and word surprisal (WS) or word entropy (WE), to assess the joint impact of both acoustic and semantic features on the temporal dynamics of the attention switch.

The re-orienting of attention around the switch can be analysed in terms of disengaging stream (attended > unattended) and engaging (unattended > attended) stream. To compare the onset and offset points of these processes, we first selected participants depending on their attentional bias over the course of the switch. Based on the reference 4-s encoding window, we determined the single-subject classification accuracy in a 15s window before and after the attention switching cue. We then found those participants displaying a lower-than-chance (50%) classification performance either in the pre- or post-switch window and excluded them

from further analyses on the switching temporal dynamics. Start and end points of disengagement and engagement processes were identified for each subject by fitting a piecewise linear regression model to the EEG prediction correlation values for the disengaging and the engaging streams, around the attention switch cue. For each of the EEG prediction correlation streams, the model incorporates the average pre-switch and post-switch values as constants and tries to estimate the breakpoints at which these flat segments connect, i.e. the onset and offset points of the process. An initial guess for the breakpoints was provided based on reasonable assumptions (e.g., that the start of the disengagement/engagement could not occur before the switching cue), and the fitting was done with MATLAB's *lsqcurvefit* function, which minimises the difference between the model and the observed data.

When computing encoding TRFs comparing different kinds of context-accumulation models, we adopted an approach optimised for highlighting the effects of lexical information processing (Chalehchaleh et al., 2025). With this method, the prediction correlation between true and predicted EEG data is computed within a window covering the range from 300ms to 500ms after word onset, allowing to effectively enhance the impact of semantic encoding.

4.2.6 Event-related spectral perturbation (ERSP)

Previous studies have related decreased parietal alpha-band (8-12 Hz) EEG power to increased listening effort and have shown that more listening effort is required to switch attention than to sustain attention on a particular speaker (Haro et al., 2022). Since we were interested in assessing the time course of alpha power and how it would compare to the temporal dynamics of the TRF prediction correlations, we computed event-related spectral perturbation (ERSP) in the alpha frequency band as a proxy of effort around the switch. For this analysis, EEG data was filtered between 0.5 Hz and 20 Hz and, since neural activity in the alpha frequency band is sensitive to ocular artifacts, Independent Component Analysis (ICA) was performed using standard EEGLAB functions (Delorme & Makeig, 2004). For each participant, we identified independent components reflecting ocular artifacts (such as blinks and horizontal eye movements) based on their temporal characteristics and spatial topographies, and we removed them from the data.

For each participant, trial, and attention switch, ERSP was computed as the relative change in power spectral density (PSD) compared to a baseline. We first defined a segment (±15 s) around the switching cue and, to estimate ERSP with the same temporal resolution used for the neural tracking analysis, we chose a sliding window of 4 s, with a step size of 0.125 s. This analysis was also replicated across a range of sliding window lengths (1s, 2s, 8s), in line

with the neural tracking analysis. Along the switching segment, PSD was calculated at each forward step within the sliding window via Welch's method (MATLAB's *pwelch* function) and summed across the alpha band frequency bins (8-12 Hz). The PSD computed at each step was compared to the reference PSD value of a baseline, consisting of the first sliding window in the pre-switch segment, producing a series of ERSP values (in dB) that reflects the relative change in alpha power during the switch.

4.2.7 Statistical analysis

Statistical analyses were performed using repeated measures ANOVAs with two or three factors, with *F*-values reported as $F(df_{time}, df_{error})$. When the assumption of sphericity wasn't met, as revealed by Mauchly's test of sphericity, a correction was applied with the Greenhouse-Geisser's method. When a significant main effect was found, post-hoc tests were run, and *p*-values were corrected for multiple comparisons with the Holm method. To compare the significance of a time series against zero, a Wilcoxon signed rank test was used for each time point, and the false discovery rate (FDR) was controlled with the Benjamini & Hochberg procedure. We report FDR-adjusted *p*-values. Descriptive statistics for the neurophysiology results are reported as a combination of mean and standard error (SE).

4.3 Results

4.3.1 Behavioural performance

Following each trial, participants were presented with a four-alternative forced-choice question about the content of one of the speech streams to confirm task engagement. Behavioural performance revealed that participants were able to successfully reply to comprehension questions, with an average accuracy of 86.3% (SEM 2.6%). Due to technical issues, behavioural data for one of the twenty-four participants was not available, therefore behavioural performance was computed based on the data from the remaining twenty-three participants.

4.3.2 Decoding of selective auditory attention in a dynamic switching scenario

Participants' attention was decoded with a backward TRF analysis, describing the relationship between the EEG signals and the envelope of the target speech. For each left-out trial, the

speech envelope reconstructed from the target decoding model was correlated with the envelopes of both the left and right speech streams. Attention was classified by determining which speech stream's envelope showed a higher correlation with the reconstructed envelope. Since this was a dynamic attention-switching scenario, the attended speech could alternatively correspond to the left or the right stream. Classification was considered correct when the reconstructed envelope correlated more strongly with the target speech envelope than with the masker envelope. Classification accuracy was then computed as the proportion of instances where this criterion was met. To establish chance performance, left and right labels were randomly shuffled 100 times for each decoding window. As shown in **Figure 4.1 C**, the longer decoding windows led to higher classification performances. However, even with a 1-second window, classification accuracy was significantly above chance level, and all decoding windows yielded classification rates significantly above the 95th percentile of its chance distribution (paired two-tailed t- test, FDR-corrected for multiple comparisons for windows of 1s, 2s, 4s, 8s, 16s, 32s, respectively: $p =$ 0.47e-9; 0.53e-9; 0.27e-9; 0.24e-9; 0.24e-9; 0.24e-9). These findings align with previous work decoding sustained attention or employing match-vs-mismatch classification metrics (Alickovic et al., 2019; de Cheveigné et al., 2018; O'Sullivan, Power, et al., 2015), and confirm that a classification based on the envelope reconstruction can reliably track selective attention even during attention switches.

4.3.3 Neural tracking of competing speech streams in a dynamic switching scenario reflects the listener's focus of attention and is related to changes in alpha ERSP

A multivariate TRF analysis was carried out to characterise the neural tracking of two competing speech streams in a setting where participants were instructed to dynamically switch their attention between the two streams. Single-subject TRFs were trained on the target stream and tested on both speakers (i.e., Spk1 and SpK2) using a multivariate speech representation that included Envelope, Word Onset and Word Surprisal features (for more details see Methods). EEG prediction correlations were computed using a sliding window to correlate true and predicted EEG signals over time, with a leave-one-out cross validation procedure, and were averaged across all EEG channels. In order to analyse robust switching dynamics, we selected 19 participants displaying a reliable attentional bias over the course of the switch, based on an above-chance classification accuracy criterion (>50%) over the course of the switch. Aligning with our expectation (**Figure 4.2 A**), EEG prediction correlations around the switching cue

reflected tracking of Spk1 and Spk2 streams consistent with the attention instructions, such that Spk1 was significantly more tracked than Spk2 before the switch, while the reverse pattern was observed after the switch (paired two-tailed t-test of Spk1-Spk2 difference against zero, FDR-corrected for multiple comparisons, $p < 0.005$).

As the attention switch unfolds, also the grand-mean ERSP in the alpha frequency band displayed a statistically significant change compared to baseline (one-sample t-test against zero, FDR-corrected for multiple comparisons), revealing a pattern of occipito-parietal negativity in the scalp topographies (**Figure 4.2 B**). This is consistent with our expectation of an impact of attentional reorientation on the EEG alpha band, which has already been shown to reflect attention switching behaviour in competing speech listening scenarios (Haro et al., 2022).

EEG prediction correlations for Spk1 and Spk2 converged, before significantly separating again once the switching process was concluded and, presumably, the attention was fully reallocated. Here, we refer to the time point when EEG prediction overlaps between Spk1 and Spk2 as the encoding switch point. Given the observed statistically significant drop in alpha ERSP, we asked how the temporal dynamics of this drop compared to those of the EEG prediction correlations. To address this, for each participant, we identified the time of the alpha ERSP minimum, and the encoding switch point, based on an encoding window of 4 s (**Figure 4.2 C**). The choice of this particular encoding window for our main analysis is justified based on the classification accuracy results (**Figure 4.1 C**), since it is a good compromise between temporal resolution and classification performance. However, the same pattern of results holds when considering multiple encoding windows simultaneously (**Supplementary Figure 4.2**, ***Appendix***). A paired t-test comparing the temporal dynamics of the alpha ERSP and the EEG prediction correlations showed that the minimum of the alpha ERSP drop significantly follows the encoding switch point ($t(18) = 4.07$, $p = 7.23\text{e-}4$, Cohen's $d = 0.93$).

We then evaluated multiple encoding window lengths, assessing the effect of Metric (encoding switch vs. ERSP minimum) and Window (1s, 2s, 4s, 8s) on the timing of the encoding switch and the minimum of the alpha ERSP with a 2-way repeated measures ANOVA. The analysis revealed that the temporal dynamics of both encoding switch and alpha ERSP minimum became longer as the encoding window length increased ($F(1,18) = 42.98$, $p = 1.48\text{e-}8$, $\eta_p^2 = 0.7$; a Greenhouse-Geisser's correction was applied due to sphericity violation), which is unsurprising given the methodological constraints we discussed (see **Methods**). More interestingly for our question, a statistically significant effect of Metric emerged ($F(1,18) = 20.78$, $p = 2.44\text{e-}4$, $\eta_p^2 = 0.54$), with the alpha ERSP minimum occurring

significantly later than the encoding switch across a range of encoding windows (Holm-corrected post-hoc *t*-test: $t(18) = 4.56$, $p = 2.44e\text{-}4$, Cohen's $d = 0.98$).

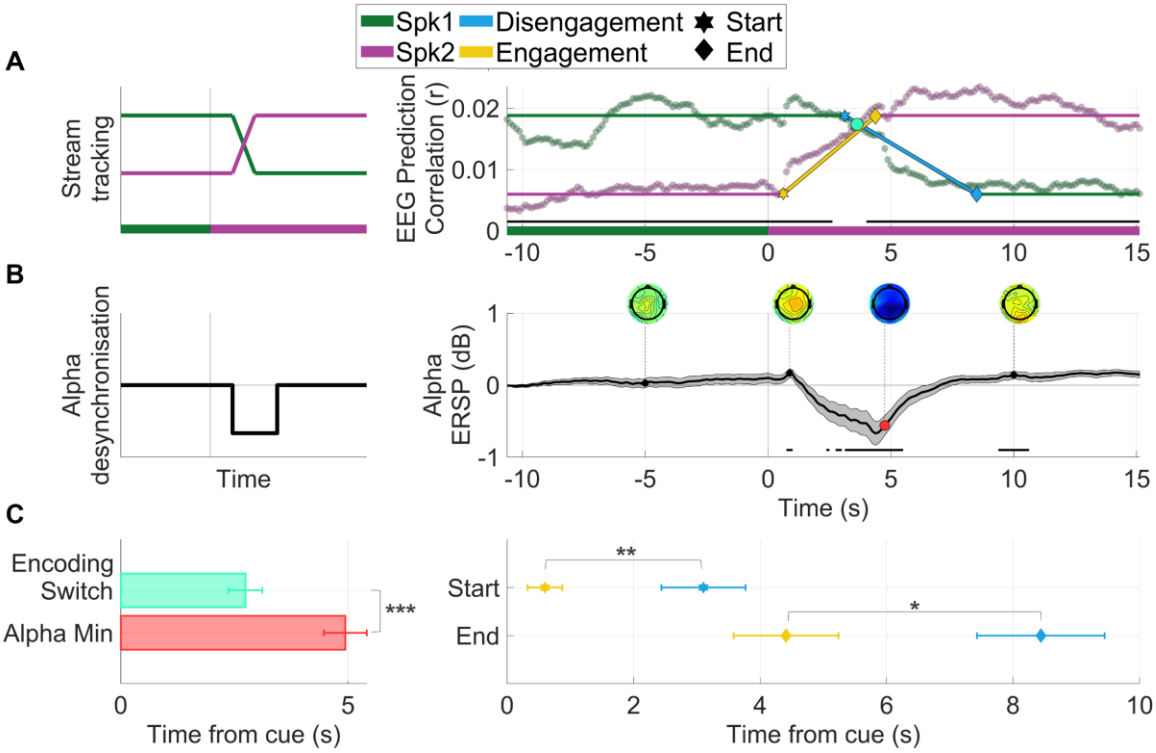


Figure 4.2: The attention-switching cue prompts a robust disengagement from Speaker 1 and engagement to Speaker 2, and it is followed by a significant decrease in the EEG alpha ERSP. Disengagement has longer temporal dynamics compared to engagement. (A) Left: Speech tracking encoding for an attention switch from Speaker 1 and 2. The trajectory in the panel represents our null hypothesis, where the disengagement and engagement processes progress in a symmetric manner after the switch-cue (vertical grey line). **Right:** Results for the neural tracking of Speaker 1 and Speaker 2 across the switching cue. EEG prediction correlations (average across all channels) obtained from a 4-s sliding-window TRF model including Envelope (Env), Word Onset (WO) and Word Surprisal (WS) features. Coloured horizontal bars at the bottom of the plot indicate the attention instruction around the attention switching cue. The turquoise dot indicates the encoding switch of EEG prediction correlations based on Spk1- and Spk2- speech features. The piecewise linear model fit for disengagement and engagement is overlayed on the EEG prediction correlation values (average of all channels and participants)**.** Hexagram shapes indicate the start of the disengagement (blue) and engagement (yellow) processes, while diamonds represent the end of the transitions. **(B) Left:** Diagram of expected results for alpha-band ERSP (event-related spectral perturbation) across the switching cue. **Right:** ERSP of the alpha band (8-12 Hz) around the switching cue (average of all channels), computed with a 4-s sliding-window, as above**.** Scalp topographies at selected time points reveal a pattern of posterior negativity, which drops significantly following the instruction to switch (thick black lines indicate a statistically significant change compared to pre-switch baseline). The red dot represents the average of ERSP minima across participants. The shaded area represents the standard error of the mean (SEM) across participants. **(C) Left:** Comparison of encoding switch of EEG prediction correlations (turquoise bar) and alpha ERSP minimum (red bar) for a 4-s sliding-window. The alpha ERSP reaches its minimum significantly after the Spk1-Spk2 encoding switch point. **Right:** Comparison of temporal dynamics for start and end points of disengagement and engagement processes. Stars indicate significant statistical effects (paired sample t-tests; *$p \leq 0.05$; **$p \leq 0.01$; **$p \leq 0.001$).

4.3.4 Dissecting the temporal dynamics of attentional disengagement and engagement during attention switching

The attention switching cue prompts the listener to reallocate their attention from the previously attended speaker, Spk1, to the newly attended speaker, Spk2. While this re-routing of attention appears to be a single, unified process, it is possible to distinguish two separate operations that are necessary for it to happen: disengagement, which we define as the decrease in neural tracking for the previously attended speech stream, and engagement, which we define as the increase in neural tracking for the previously unattended speech stream. Our goal was to clarify the temporal dynamics of these two operations to understand whether they occur fully in parallel, serially, or with a certain degree of overlap.

As above, we first selected participants displaying a reliable attentional bias over the course of the switch (see **Methods**). For this selection of 19 participants, we fitted a piecewise linear regression on single-subject EEG prediction correlations, and found the optimal breakpoints, corresponding to the start and end time points of disengagement and engagement (**Figure 4.2 C**). As above, we chose to focus on an example window of 4s and later replicated our results on a range of encoding window lengths. Disengagement and engagement processes were compared separately based on their start times and end times, revealing consistently earlier temporal dynamics for the engagement compared to the disengagement. Engagement to the newly attended speaker started significantly earlier than the disengagement from the previously attended speaker (paired-sample *t*-test: $t(18) = 3.06$, $p = 0.007$, Cohen's $d = 0.43$), and finished significantly earlier (paired-sample *t*-test: $t(18) = 2.56$, $p = 0.02$, Cohen's $d = 0.42$).

We then extended our analysis to a range of sliding window lengths and compared start times and end times for disengagement and engagement processes, including Window (1s,2s,4s,8s,16s) and Process (disengagement vs. engagement) as main factors in a repeated measures ANOVA. Regarding the start points (**Figure 4.3 A**), our analyses revealed an expected statistically significant effect of Window ($F(1,18) = 8.37$, $p = 0.003$, $\eta_p^2 = 0.32$; the assumption of sphericity was not met; hence, a Greenhouse-Geisser's correction was applied), with longer temporal dynamics corresponding to longer encoding window lengths. More importantly, we also observed a significant main effect of Process ($F(1,18) = 7.22$, $p = 0.015$, $\eta_p^2 = 0.27$), with engagement to the newly attended stream starting significantly earlier than the disengagement to the previously attended stream (Holm-corrected post-hoc *t*-test: $t(18) = 2.69$, $p = 0.015$, Cohen's $d = 0.64$). The same statistical analysis was repeated separately on the end time points of disengagement and engagement processes (**Figure 4.3 B**), revealing once again

a main effect of Window, whereby longer encoding windows yield longer temporal transitions ($F(1,18) = 36.7$, $p = 6.8e\text{-}8$, $\eta_p^2 = 0.32$; the assumption of sphericity was not met; hence, a Greenhouse-Geisser's correction was applied). A significant main effect of Process also emerged ($F(1,18) = 6.36$, $p = 0.021$, $\eta_p^2 = 0.11$), revealing that the process of engagement to the newly attended speaker, not only starts, but also ends significantly earlier than the disengagement (Holm-corrected post-hoc t-test: $t(18) = 2.52$, $p = 0.021$, Cohen's $d = 0.75$).

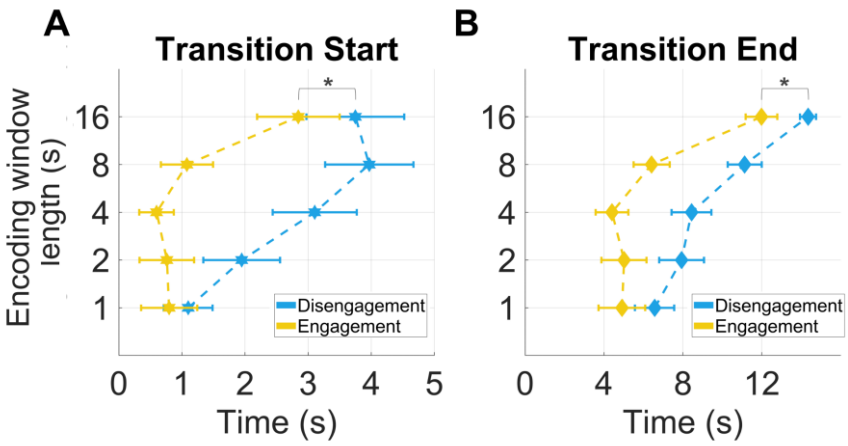


Figure 4.3: Comparing the start and end transition points for the disengagement and engagement processes after the attentions switching cue. The process of engaging to a new speaker begins and ends significantly earlier than disengaging from the previously attended speaker. (A-B) Start and end points of the transition for the disengagement (blue) and engagement (yellow) processes over five TRF sliding window lengths. Error bars represent SEM across participants. Stars indicate significant effects of process type (two-way repeated measures ANOVA; *p ≤ 0.05; **p ≤ 0.01; ***p ≤ 0.001).

4.3.5 Determining how lexical predictions are built during attention switching

Reorienting attention to a different speech stream implies a change of context and, consequently, different semantic priors for lexical predictions. We thus hypothesised that incorporating this change of context into the structure of our semantic regressor in a multivariate encoding TRF model would increase EEG prediction correlations, as it would better reflect the dynamically updating neural tracking of the competing speech streams. We compared four alternative models representing how context could be incrementally accumulated for performing lexical predictions at one particular attention block (e.g., B7, in **Figure 4.4 A**). A naïve Oracle model, which uses all available context of previous blocks from the current stream, whether attended or unattended, to predict words from the current block,

served as our baseline, since it was essentially a switch-unaware contextual representation. Speaker-Specific and Attention models were instead switch-aware models, as they only considered previously attended blocks as part of the context for lexical predictions. Speaker-Specific assumed a higher degree of stream segregation, since its context only consisted of previously attended blocks from the same speech stream, while Attention included any previously attended block from both streams. The Reset model instead ignored all previously attended blocks from any of the streams and computed context only over the course of the current block of attention, as if the priors for lexical predictions were reset at each attention switch (**Figure 4.4 A**).

As lexical entropy is a proxy of uncertainty for next-word prediction, its values should be impacted by a switching cue, which determines an abrupt change of context. **Figure 4.4 B** shows the change of average lexical entropy values in words preceding and following the switch cue, which vary depending on the context models. It can be observed that the Reset model peaks with the highest uncertainty and slowly decays over the course of the next words, while the Attention and Speaker Specific models have overall similar lexical entropy dynamics and more stable values. Consistently with its switch-unaware nature, the Oracle model instead displays entropy values that are largely unchanged despite the switch.

Lexical surprisal and lexical entropy were used as semantic information regressors for each context model and separately included in a multivariate stimulus representation to fit single-subject encoding TRFs (Envelope-Word Onset-Word Surprisal and Envelope-Word Onset-Word Entropy). Resulting TRF weights and EEG prediction correlations were then compared across context models, with the hypothesis that switch-aware and context-rich representations (e.g., Speaker-Specific or Attention) would best describe neural activity in attention-switching scenarios.

Before comparing the context models, we first tested whether each of them yielded a significant encoding accuracy gain compared to the baseline model only consisting of acoustic features (Envelope and Word Onset). When using entropy as a regressor for semantics, all models, with the exception of Oracle, showed a statistically significant gain, suggesting a robust tracking of semantic information in addition to the stimulus acoustics (paired t-tests: Oracle vs. Acoustics: $p = 0.29$; Spk.Spec. vs. Acoustics: $p = 0.03$; Attention vs. Acoustics: $p = 0.03$; Reset vs. Acoustics: $p = 0.003$). Employing word surprisal as semantic regressor yielded similar results, with all the models showing a robust encoding of semantic information, apart from Oracle (paired t-tests: Oracle vs. Acoustics: $p = 0.16$; Spk.Spec. vs. Acoustics: $p = 0.04$; Attention vs. Acoustics: $p = 0.03$; Reset vs. Acoustics: $p = 0.02$). The non-significant gain of

the Oracle model compared to the acoustic model was expected, since Oracle was designed as a control *switch-unaware* model.

In contrast to our expectation, the Reset context model was shown to yield higher EEG prediction correlation values when entropy was used as a regressor for semantics (**Figure 4.4 C**). A repeated measures ANOVA revealed a statistically significant effect of the main factor, Context Model ($F(1,23) = 7.12$, $p = 0.001$, $\eta_p^2 = 0.27$; with Greenhouse-Geisser's correction). In the Holm-corrected post-hoc tests, the Reset model was shown to yield significantly higher encoding accuracies than Oracle ($t(23) = 4.55$, $p = 1.32e\text{-}4$, Cohen's $d = 0.14$), and Speaker Specific ($t(23) = 2.96$, $p = 0.021$, Cohen's $d = 0.09$). Reset was also marginally significant compared to the Attention model ($t(23) = 2.52$, $p = 0.056$, Cohen's $d = 0.08$). We then assessed the difference of TRF weights for the entropy feature across the four context models (**Figure 4.4 D**), averaging the weights' amplitude within a window broadly centred around the TRF-N400 latency (350ms-550ms). A repeated measure ANOVA was run on the weights' amplitude values, revealing a main effect of Context Model ($F(1,23) = 14.56$, $p = 1.9e\text{-}7$, $\eta_p^2 = 0.39$; a Greenhouse-Geisser's correction was applied). Post-hoc tests (Holm-corrected) showed that weights for the Reset model had lower TRF-N400 amplitude compared to Oracle ($t(23) = -5.93$, $p = 6.41e\text{-}7$, Cohen's $d = 0.56$), Attention ($t(23) = -4.94$, $p = 2.1e\text{-}5$, Cohen's $d = 0.47$), and Speaker Specific ($t(23) = -5.11$, $p = 1.4e\text{-}5$, Cohen's $d = 0.48$). It is worth noting that the entropy TRF weights for the four semantic models all displayed significant pre-onset activity, as highlighted by a one-sample *t*-test against zero (Oracle: $t(23) = 2.87$, $p = 8.7e\text{-}3$, Cohen's $d = 0.57$; Attention: $t(23) = 3.44$, $p = 5.5e\text{-}3$, Cohen's $d = 0.68$; Speaker-specific: $t(23) = 3.06$, $p = 5.5e\text{-}3$, Cohen's $d = 0.6$; Reset: : $t(23) = 3.93$, $p = 6.5e\text{-}4$, Cohen's $d = 0.78$).

When fitting a multivariate TRF including lexical surprisal as a semantic regressor we were not able to observe a statistically significant difference between the four context models for the EEG prediction correlations, even if the Reset model yielded overall higher accuracy values. No difference was observed also when comparing the TRF-N400 amplitude of the models' TRF weights.

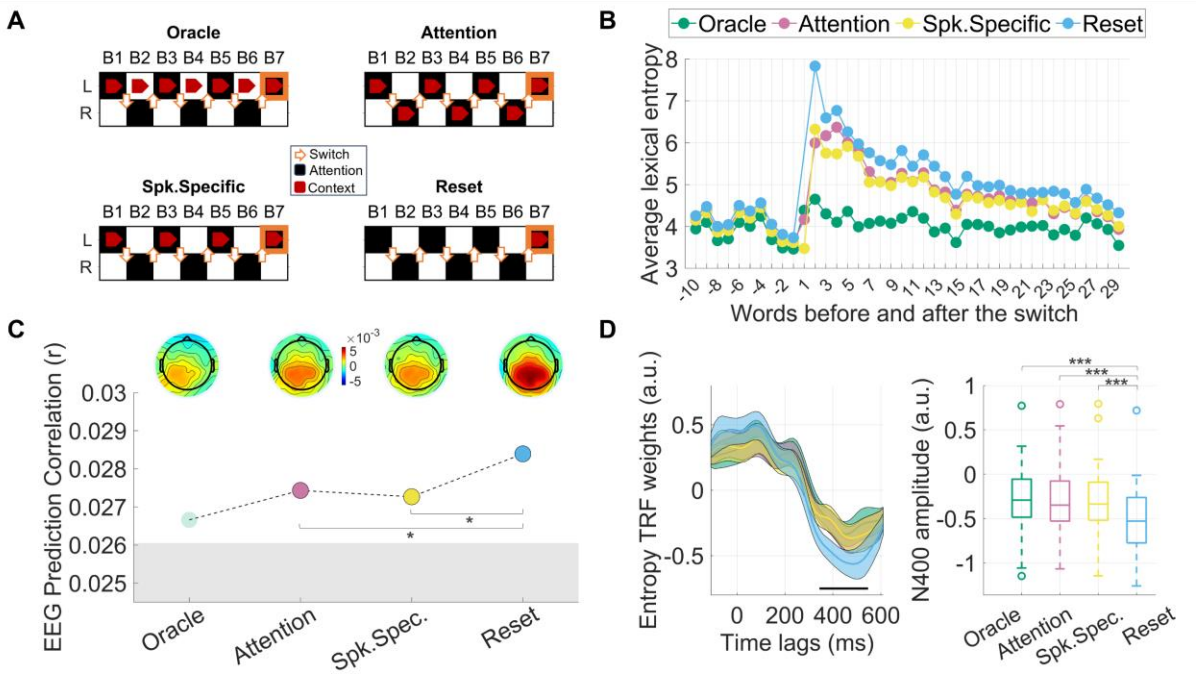


Figure 4.4: Investigating lexical prediction mechanisms during attention switching. (A) Layout of the four context models. Blocks coloured in black illustrate sustained attention either to the Left or Right stream, while orange arrows indicate attention switching cues. The thick red arrow indicates the context used to guide word predictions for the current block (B7, highlighted in orange). **(B)** Average lexical entropy at words preceding and following the attention switch cue. Note that no value for entropy is displayed in the Reset model for the first word after the switch, due to the context being fully reinstated. **(C)** EEG prediction correlations for the four multivariate TRF models, only differing in their entropy feature. Coloured dots indicate the average across all electrodes and participants. The grey area at the bottom represents the average encoding accuracy of a multivariate TRF without any semantic information (Envelope + Word Onset). Stars represent statistically significantly greater EEG prediction correlations for the Reset model compared to the other models (Significance levels: *p < 0.05, **p < 0.01, ***p < 0.001). Topographical patterns illustrate the gain due to semantic information (compared to the Envelope + Word Onset TRF) for the four models. **(D)** (Left) TRF weights for the entropy feature at time-lags between -100ms and 600ms relative to stimulus onset. Transparent shaded areas represent the standard error of the mean (SEM) across participants. The horizontal black line indicates the time window employed to compute the average TRF-N400 amplitude. (Right) Boxplots representing the distribution of the TRF-N400 amplitude across participants for the four context models. The central line within each box represents the median, while the edges of the box indicate the interquartile range (IQR). Whiskers extend to the most extreme data points within 1.5 times the IQR from the quartiles. Outliers are plotted as individual points beyond the whiskers. Stars indicate statistically significant comparisons (Significance levels: *p < 0.05, **p < 0.01, ***p < 0.001).

4.4 Discussion

Speech communication in multi-talker environments requires a skilful combination of sustained attention and rapid attention switching abilities (Agmon et al., 2021; Bronkhorst, 2015). While the neurophysiology of sustained speech attention has been widely studied (Alickovic et al., 2021; Gillis et al., 2021; O'Sullivan, Power, et al., 2015; Power et al., 2012),

less is known about the neural mechanisms of attention switching. Here, we fill this gap with a tailored EEG experiment examining the neurophysiology of attention switching across different levels of speech abstraction. In doing so, we 1) demonstrated an experimental paradigm that can successfully probe both sustained attention and attention switching mechanisms; 2) successfully dissected disengagement and engagement processes with a high temporal resolution, identifying substantial asymmetries in their temporal unfolding and a transient simultaneous encoding of two speech streams; and 3) proposed a neurophysiologically plausible explanation of how our brains update and use lexical context when switching attention.

The findings in this study have several implications for our understanding of speech attention switching mechanisms. The asymmetry measured between disengagement and engagement processes highlights the importance of studying the two processes separately. That distinction was often not considered in previous studies on sustained attention, which often focussed on measures of attention bias or classification (de Cheveigné et al., 2018; Mesgarani & Chang, 2012; O'Sullivan et al., 2017; Raghavan et al., 2024). The effectiveness of such decoding metrics has been a driving force for research on brain-computer interfaces such as cognitively-controlled hearing devices (Geirnaert et al., 2021; Hjortkjær et al., 2025; Katthi et al., 2020; Wilroth et al., 2025). Our finding highlights that encoding metrics enable a sufficient level of detail for disentangling how the encoding of different streams evolves over time. Here, we measured an asymmetry between disengagement and engagement processes during attention switching in a very specific scenario. Indeed, it will be important to determine how that relationship is modulated by factors such as cognitive load, ageing, cognitive abilities, hearing difficulties, interest in the speech content, frequency of attention switches in a trial, among many others. Of course, future work should also scrutinise how the specific nature of the task might impact that phenomenon.

Sustained attention tasks, where participants focus on a target speech while ignoring the masker (Fuglsang et al., 2017; Zion Golumbic et al., 2013), involve a quite particular scenario where listeners have no incentive to monitor unattended streams. In real-life situations, however, listeners may have reasons to explore alternative speech streams, for example due to a lack of interest in the current speaker. Our experimental paradigm more closely mirrors this scenario. Although the instructed nature of the task makes it less realistic, the paradigm incentivises monitoring the masker and being ready to rapidly switch attention, which contrasts with sustained attention tasks. While the asymmetry observed may be specific to this experimental paradigm, the result indicates that our brains can engage with a new target

even before starting the disengagement from the previous one. This *de facto* produces a transient simultaneous cortical tracking of two speech streams, with both measurements comparable to sustained attention in strength, aligning with auditory scene monitoring mechanisms (Bregman, 1990; Shinn-Cunningham, 2008). Intuitively, maintaining a transient parallel representation of multiple speech sources during attention switching is an efficient neural processing strategy. It allows the flexibility to switch back to the previous stream, if necessary, without fully committing to the newly attended speech immediately. This phenomenon supports previous claims that our brains can process speech maskers beyond the acoustic level, encoding linguistic properties to some extent (Har-shai Yahav & Zion Golumbic, 2021; Kaufman & Zion Golumbic, 2023; Park & Gross, 2023).

Unattended speech streams are also represented in human cortical activity, with evidence for lower encoding strengths or longer time latencies than the target stream (Ding & Simon, 2012a; Hausfeld et al., 2018; Power et al., 2012), and with a gradient of attentional bias from primary to nonprimary auditory cortex (Hausfeld et al., 2024; O'Sullivan et al., 2019), whereby the unattended stream encoding tends to be substantially reduced or not measurable in higher-order cortical areas (Anderson et al., 2024; Broderick et al., 2018). Prior research has shown that not only the speech envelope, but also other key features of the unattended speech, such as acoustic onsets, are neurally represented (Brodbeck et al., 2020; Orf et al., 2023) and, when not readily available due to speech masking, they are even restored at later temporal scales (Fiedler et al., 2019). One interpretation is that encoding a template structure of the unattended speech might be a useful strategy to suppress it (Brodbeck et al., 2020) . Other research found attentional fluctuations corresponding to the changes in Target-Masker relative sound energy, with evidence that our brains may encode some phonetic information of unattended streams (Carta et al., 2024; Raghavan et al., 2023). The encoding of such unattended stream information may be one of the factors facilitating the rapid engagement during attention switching.

Using alpha ERSP as indicators of listening effort, this study related perceptual demands with attention switching dynamics. A large reduction in EEG alpha-band power was measured consistently about 4.5 seconds after the attention-switching cue. The ERSP trajectory suggests that a strong listening effort persists throughout the attention switch, with a substantial reduction close to the end of the switch (**Figure 2A**). The late ERSP trough, marking a turning point in listening effort, occurred significantly later than the moment of the attention switch itself. This finding extends prior research on the neural correlates of selective attention and listening effort. Variations in alpha ERSP have been associated not only with auditory attention

effort (Dimitrijevic et al., 2019; Fiedler et al., 2021; Paul et al., 2021; Seifi Ala et al., 2020; Shahsavari Baboukani et al., 2022), but also with the active suppression of irrelevant information in accordance with behavioral goals (Foxe & Snyder, 2011; Herrmann et al., 2023; Wöstmann et al., 2017). Notably, cognitive load and the inhibition of irrelevant stimuli appear to be even more strongly influenced by attention reorienting than by maintaining attention on a single speaker (Haro et al., 2022; McCloy et al., 2017), highlighting the increased cognitive demands of switching attention. While alpha power has commonly been linked to subjective listening fatigue (Dimitrijevic et al., 2019; Paul et al., 2021) or the SNR between attended and unattended streams (Fiedler et al., 2021), it is less frequently associated with neural markers of speech tracking (L. Decruy et al., 2020; Tune et al., 2021). In this study, we identified a link between the temporal dynamics of neural speech tracking and listening effort, suggesting a potentially valuable metric for future research into attention-switching challenges.

This study examined how linguistic context is dynamically updated during attention switching by comparing four context-accumulation models that varied in their sensitivity to attention switches and access to prior context: an Oracle model (context-rich but switch-unaware), Speaker Specific and Attention models (both context-rich and switch-aware but differing in stream selectivity), and a Reset model (aware of the switch but limited to the current block's context). For each model, a multivariate temporal response function was fit using a semantic regressor aligned with the respective context strategy. Unexpectedly, the Reset model best predicted EEG data, outperforming models that retained past context and challenging the assumption that prior semantic information aids comprehension during attention switches. This finding may reflect a cognitive mechanism, suggesting listeners reset context and recalibrate their lexical predictions dynamically when switching attention to a new stream. Another potential explanation highlights the predictive nature of LLMs like Mistral, which are optimised for next-word prediction, without a requirement of being neurophysiologically plausible. As such, despite consistent reports of the growing similarity between more modern LLMs with neurophysiological activity (Caucheteux et al., 2023; Goldstein et al., 2022; Mischler et al., 2024),the higher neurophysiological plausibility of the Reset model tells us that prior context is not availed of by our brains in the way implemented by the competing models, Attention and Spk-Specific. Nonetheless, we don’t exclude that other strategies making use of the prior context might be in place. One possibility is that a dynamic attention switching task prompts a different use of that context, for example by capturing its essence in a more abstract and less specific way, as the gist of the story (Park & Gross, 2023). As such, there could be value in exploring different strategies for context representation, for example by employing

Large Concept Models (LCM team, 2024), which are trained and optimised for sentence prediction. This latter possibility is also supported by recent work on the accumulation of linguistic context in LLMs and the human brain (Tikochinski et al., 2025) while listening to continuous monologues, showing that LLMs with a limited context window (32 tokens) and with access to a coarse summary of the previous context predict neural activity better than LLMs with a higher token-memory.

These fine-grained differences between context-accumulation models only emerged when employing lexical entropy as a regressor of semantic information, while they were not reflected in lexical surprisal. In the particular scenario of this experiment, a potential reason for the observed difference may lie in the neurophysiological relevance of the two features. Lexical surprisal represents the amount of information that the current word adds relative to the preceding context. As such, it can be described as a *backward-looking* feature, because its computation depends on prior context. On the other hand, lexical entropy is more *forward-looking*, as it represents the amount of uncertainty over all possible upcoming words, given the current word. The current experimental paradigm, where attention switching cues determined a frequent change of relevant semantic context, may have caused participants to be less reliant on the linguistic information they had accumulated, and to focus instead on the linguistic uncertainty ahead, making lexical entropy a better suited semantic regressor for the EEG data.

Furthermore, when considering the time course of TRF weights for lexical entropy (**Figure 4.4 D**), the results indicate statistically significant pre-onset activity across all four context-accumulation models. One possible explanation is an anticipatory activation of lexical representations before a word is received. Beyond classical ERP paradigms (DeLong et al., 2005; Dikker & Pylkkänen, 2013), similar pre-activation accounts have also been supported by recent studies relying on naturalistic listening paradigms and modelling speech information content with LLMs (Goldstein et al., 2022). Future analyses could focus on how the level of pre-activation may vary depending on the uncertainty of the prediction or the current level of surprisal.

In summary, this study showed that the process of attention switching in a realistic multi-talker scenario can be investigated in terms of its engagement and disengagement components, with a transient parallel representation of the two streams. We highlight the importance of relating metrics of neural tracking of speech with metrics of listening effort and demonstrate that the listening effort starts decreasing following successful disambiguation of the two streams during attentional re-allocation. Finally, we introduce an approach for modelling lexical context of dynamic attention scenarios, showing the sensitivity of

transformer-based language models to subtle differences in context accumulation strategies. These findings have implications for future investigations into the cortical mechanisms of attention re-orienting and can be employed to highlight differences across diverse populations in terms of age and hearing levels. In the next Chapter, I present a study where I begin to investigate these aspects comparing a younger and an older population of participant, using an experimental paradigm similar to the one adopted in the current study.

4.5 Summary

Successful speech communication in multi-talker scenarios requires a skilful combination of sustained attention and rapid attention switching. While the neurophysiology literature offers detailed insights into the neural underpinnings of sustained attention, there remains considerable uncertainty on how attention switching takes place. In this study, using EEG recordings from normal-hearing adults in an immersive multi-talker environment, we measured the neural encoding of two competing speech streams amid background babble. Participants were cued to switch attention between streams every 10–30 seconds. Neural tracking was assessed via TRFs, confirming reliable decoding of attentional focus. Our results indicate asymmetric disengagement and engagement processes during attention switches, where the neural tracking of the new target stream emerges before disengaging from the previous target, revealing a transient simultaneous encoding of two speech streams. That transition was closely mirrored by a reduction in EEG alpha power, informing on the cognitive effort during different phases of the attention switch. We then isolated cortical activity reflecting lexical prediction mechanisms to determine how lexical context is updated after an attention switch, comparing four numerical hypotheses that were constructed using LLMs. Our findings elucidate both the temporal and contextual mechanisms underlying auditory attention shifts, pointing to the possibility that listeners carry out a reset in lexical context after switching attention. By focusing on dynamic attentional reallocation, this study offers insights into the brain's capacity for flexible speech processing in complex listening environments.

Chapter 5. Study 3: Investigating the neurophysiology of attention switching in the ageing brain

5.1 Introduction

A complex interplay of changes in speech audibility and cognitive abilities is thought to underlie age-related differences in speech processing. On the one hand, a cascade of neurophysiological changes across the auditory pathway determines increased hearing thresholds (Wu & Liberman, 2022) or hidden hearing loss (Parthasarathy et al., 2020), especially in noisy scenarios, resulting in a compensatory reliance on domain-general executive systems and more intense listening effort (Peelle & Wingfield, 2016). On the other hand, executive functions and general cognitive abilities are known to decline with ageing (Salthouse, 2010), especially in the domain of attention, working memory, and inhibitory control, which are important for maintaining focus on a speaker of interest, remembering the details of an unfolding conversation, and suppressing irrelevant information. As such, a decline in general cognitive functions will naturally have an impact on the ability to understand speech, among other behavioural consequences. In the neurophysiology of speech processing, these age-related differences emerge in envelope tracking responses, which appear to be enhanced for older listeners (C. Brodbeck et al., 2018; Decruy et al., 2019), but not consistently across experiments (Gillis et al., 2022; Tune et al., 2021), and in tracking of lexical information, with N400 responses that appear delayed, of smaller amplitude or unchanged, depending on the study (Broderick et al., 2021; Gillis et al., 2023; Mesik et al., 2021). However, age-related differences in the processing of continuous speech have only been studied with monologue listening paradigms, or in multi-talker scenarios with sustained attention paradigms. As a result, the neural mechanisms of attention reorienting in older listeners are still unclear, despite their relevance in realistic listening environments with multiple speakers, where older individuals experience the most intense difficulties (Herrmann, 2025; Herrmann & Johnsrude, 2020).

As described in Chapter 4, the flexible reorienting of attention between two speech streams can be studied non-invasively with EEG, deriving metrics regarding the temporal dynamics of engagement and disengagement from a stream, the listening effort around the switch, and the relevance of lexical context in a dynamic attention scenario. However, that

study has not addressed the question of how the timing and trajectory of these dynamic attentional mechanisms might change across various populations, for example, in the ageing brain.

Here, we employ our novel experimental paradigm and metrics to investigate the process of attention switching across two populations of younger and older listeners with normal hearing (38 subjects) undergoing an EEG experiment and a separate cognitive test battery. Participants were instructed to reorient their attention between two speech streams in the foreground while simultaneously ignoring a background speech babble (16 talkers). In doing so, we compared how switching cues impact sustained attention to speech across the two populations, analyse neural tracking of speech around the switch with temporally fine-grained methods. By further introducing a frequent change of topic within each speech stream, we aim to probe for the reliance on previous lexical context in older and younger participants, with the hypothesis that the lack of relevant semantic context will be more detrimental for older than for younger listeners.

Under the hypothesis that speech processing in ageing is impacted by hearing and cognitive changes, we expect that also switching attention to another speech stream in a multi-talker scenario will reveal age-related differences, with our neural metrics of disengagement-engagement and listening effort showing potentially longer time scales for older individuals, and potentially different trajectories across the stages of tuning out from one stream and tuning in to a novel stream. Furthermore, we investigate neural tracking of sustained attention blocks within our switching paradigm, highlighting the impact of frequent switching instructions on stable attentional allocation.

5.2 Methods

5.2.1 Participants

This study involved 38 native English speakers, divided into two groups of 19 younger (mean age 22.16 years, with a standard deviation of 3.7 years) and 19 older participants (mean age 61.16 years, with a standard deviation of 4.9 years). All participants had no history of neuro-psychiatric disorders and underwent a pure tone audiometry test (from 0.25 Hz to 8 kHz) to confirm their normal hearing levels. The experiment was carried out following a protocol approved by the Ethics Committee of the School of Psychology at Trinity College Dublin.

5.2.2 EEG experimental procedure

Participants were located at the centre of a circular multi-talker scenario consisting of an array of six loudspeakers positioned around the listener at a distance of 1.50 m each (with horizontal angles of ±30°, ±112.5°, and ±157.5°). The experimental task consisted in dynamically shifting attention between foreground speech streams on the left and right, guided by a left- or right-pointing arrow presented at the centre of a screen in front of the participants. While attending to the speech stream indicated by the visual cue, participants were also required to ignore background noise consisting of a 16-talker babble, distributed across four loudspeakers, each playing a 4-talker babble.

The experiment was divided into twenty trials, each lasting 210 s, with eight attention switching cues per trial, occurring semi-randomly. As a result, participants sustained attention on a particular speech stream for a variable amount of time, ranging from 12 s to 30 s.

At the start of each trial, the four background loudspeakers simultaneously played the multi-talker babble, while an arrow on the screen cued the participant by indicating the side of the target stream. After 5 seconds, the left and right foreground speakers were introduced into the auditory scene. After each trial, three multiple-choice questions were presented on the screen, asking participants about their perceived level of difficulty in performing the attention switching task, their preference based on interest between left and right streams, and their understanding about the content of one of the two streams. Over the course of the experiment, care was taken to avoid excessive fatigue by setting three mandatory breaks, every fifth trial, and by allowing participants to decide when to continue with the next trial.

In each trial, a different male and female speech stream was presented from the left and right foreground loudspeakers, with the side of presentation and the starting attention side counterbalanced across the experiment.

The speech streams were sampled at 44.1 kHz and delivered via a Roland Octa-Capture 10×10 sound card (24-bit/192 kHz) to six PreSonus Eris 4.5BT loudspeakers. EEG data was recorded with a BioSemi ActiveTwo system at a sampling rate of 512 Hz from sixty-four electrodes mounted on a standard cap in accordance with the International 10–20 system. A CMS electrode served as the active reference and a DRL electrode as the passive reference for all channels, while two additional electrodes were placed on the mastoids and used for offline re-referencing.

5.2.3 Determining the SNR

The SNR of the foreground speakers relative to the background babble was determined individually for each participant prior to the EEG session, employing a custom-made version of the Hearing in Noise Test (HINT) (Nilsson et al., 1994). The HINT is an adaptive test where listeners are presented with short sentences against a background of noise and are asked to repeat them word by word. For each sentence the participant correctly repeats, the speech level of the next sentence is decreased, making it more difficult to hear. If the participant instead repeats it incorrectly, the task is made easier at the following sentence, by increasing its level. This adaptive procedure continues until the test determines the speech reception threshold (SRT), the level at which the participant can correctly identify a certain percentage of the sentences in noise. In this case, the target SRT was set to be at 80%. The HINT was administered through the same circle of loudspeakers used for the EEG experiment, with the attended stream being alternatively located on the left or right side throughout one single list of sentences.

After training with an example list, participants were presented with two lists, one with the target stream from the left and one from the right side, and their SRTs for the two sides were averaged. The background noise used in this version of the HINT consisted of the same multi-talker babble used in the EEG experiment, while the competing speaker in the foreground was another short sentence from the standard HINT repertoire. The levels of the multi-talker babble and the competing speaker were jointly modulated based on the adaptive procedure to make the task easier or harder. It is important to note, however, that this modulation of the SNR is different from the one adopted during the EEG experiment. In the HINT, as mentioned above, the SNR was computed between the attended speaker in the foreground against any unattended stream, including both the background babble and the unattended competing speaker at the front. In the EEG experiment, on the other hand, both foreground speakers were set at the same level, and the SNR was only modulated for the ensemble of the foreground speakers relative to the background babble noise. As a result, the SRTs obtained from this version of the HINT are probably an optimistic estimate of the listeners' actual speech-in-noise abilities, compared to the increased difficulty of the EEG experiment due to the 0 dB SNR between the two foreground speakers.

5.2.4 Cognitive test battery

In an experimental session separate from the EEG, participants underwent a computer-based cognitive test battery with the aim to assess their fluid intelligence, working memory and processing speed abilities. These cognitive skills appear to be related to speech processing, and to potentially facilitate it under challenging conditions (Lunner, 2003). As the goal would be to better clarify the relationship between these cognitive abilities and speech processing, the current test battery only includes non-verbal tasks, thus minimising potential confounds between the two domains. Please note that results from the cognitive tests have not been employed as part of the analyses presented in this Chapter.

To test for fluid intelligence, a pattern completion task was used, similar to Raven's matrices (Raven & Raven, 2003), but consisting of only 12 puzzles to be solved within 20 minutes. In each puzzle, participants were presented with eight geometric figures, each varying in a range of features, such as colour, orientation and shape, arranged in a 3 by 3 grid, with the bottom-right corner left blank. The goal was to identify the pattern relating those geometric figures and choose the figure that correctly completes the pattern from a separate list of candidate figures.

Working memory was tested with the Corsi block-tapping task (Corsi & Michael, 1972), which probes the ability to replicate a visually-presented sequence. In this test (lasting at most five minutes), some of the squares, initially blank and positioned at different locations on the screen, light up in red one after another, creating a particular spatial sequence. Once the sequence is over, the task is to replicate it by clicking on the blank squares to light them up in the correct order. If performing well, the participant is challenged with longer and more complex sequences, while the test is interrupted earlier after multiple errors.

The Digit Symbol Substitution Test (DSST) (Wechsler, 1940) was used to assess participants' speed of information processing. In this test, lasting two minutes, participants are presented with a coding key where each of nine abstract symbols corresponds to a particular digit from 1 to 9. After familiarising themselves with the key, participants had to fill out a long table by correctly replacing symbols with their corresponding digits. The coding key was always visible on top of the screen, but due to the short time available, the task could be efficiently performed only with a certain degree of learned associations between symbols and digits.

5.2.5 Experimental stimuli

Throughout the experiment, participants switched their attention between two speech streams uttered by AI generated male and female voices (https://play.ht), with the content consisting of publicly available scripts of real TED talks on diverse topics.

The artificially generated voices were selected based on their perceived naturalness and clarity, and sample sentences for each voice were also presented to a separate group of eight participants (three males, mean age of 34.25 years, standard deviation of 13.65 years) in an online questionnaire, to ensure that they indeed sounded sufficiently natural. For each of the audio samples, participants were asked to rate them on a 5-point Likert scale based on four statements: 1) *The speaker is clear and understandable*; 2) *The speaker is easy to follow*; 3) *The speaker's pacing and intonation is appropriate*.

Once the voices were selected and the audio samples for the experiment were artificially generated, the resulting speech streams were matched in loudness through root-mean-square (RMS) normalisation. Despite the controlled acoustics of these artificially generated stimuli, the procedure is still useful especially to compensate for differences between male and female voices. Each 4-talker background babble was created by combining the audio signals from four different TED talks. The long-term average spectrum of this babble noise was subsequently adjusted to match the overall spectral profile of both the male and female foreground speakers, ensuring consistent masking effects.

Unlike the previous study, in this experiment there was a frequent change of topic in the foreground streams. As such, following a switching cue, the listener reorienting towards a previously attended speaker would sometimes be exposed to a completely different topic. For example, the left speech stream could be initially talking about gardening, but after a few attention switches, the topic could suddenly turn into the Roman empire. This naturally added a further layer of difficulty to the already taxing attention switching task, challenging the reliance on semantic context. Out of eight attention switches per trial, three also introduced a change of topic. In order to avoid confusion at the beginning of the trial, topic changes were set to occur semi-randomly only between the third and eighth switching cues.

5.2.6 Speech features

In this study, we aimed to characterise the neural dynamics of switching attention between two speech streams in a population of younger and one of older listeners. As such, we identified features of the foreground competing speech streams to regress them against the continuous

EEG activity of each participant and assess their neural tracking response. Considering that the focus was on the neural tracking of the streams' acoustics, we identified features relating to the speech sound and to word segmentation.

The broadband speech envelope (Env) of each stream was calculated as the absolute of the stream's Hilbert transform. The envelope derivative (EnvDer) was computed as the pointwise difference between consecutive samples of the envelope and provided an approximation of the envelope's rate of change. To extract information relative to the occurrence of words in the streams, the transcribed stimuli and corresponding audio recordings were first automatically aligned using the WebMAUS Basic aligner (Kisler et al., 2017; Schiel, 1999, 2015), which generated timestamps marking the start and end of each word. The resulting alignments were exported in TextGrid format and, when needed, manually corrected using Praat (Boersma, 2007). These timestamps were then used in MATLAB to construct binary vectors indicating word onsets (WO).

5.2.7 EEG data preparation

Data was analysed offline using the MATLAB software (MathWorks), starting from scripts and resources made publicly available as part of the CNSP initiative (Cognition and Natural Sensory Processing; https://cnspworkshop.net). For the main neural tracking analysis, EEG signals were digitally filtered between 0.5 Hz and 8 Hz in both the forward and backward direction to avoid phase distortion, with a fourth-order Butterworth filter, and then downsampled from 512 Hz to 128 Hz for computational efficiency. EEG channels with variance exceeding the average variance by three standard deviations were considered noisy and replaced by applying spherical spline interpolation. As a last step, EEG signals were re-referenced to the average of the two external channels on the mastoids.

5.2.8 TRF analysis

As in Chapter 4, a TRF analysis was conducted to quantify how well the EEG in an attention switching scenario reflects the encoding of the various speech representations. To do so, a multivariate model - including Env, EnvDer and WO features - was fit to describe the forward mapping from speech features to neural data, and then tested on a left-out trial, by assessing how well the model could predict a segment of data it had not been trained on. The model's performance is quantified as the correlation coefficient (Pearson's r) between true and predicted data, and this procedure is repeated iteratively for every trial (leave-one-out-cross validation). The model was fit for each participant individually, considering a time lag window ranging

from – 100 ms to 600 ms, and applying Tikhonov regularisation with a tuning of the λ parameter from a space spanning logarithmically from 10^{-6} to 10^{4}.

In this study, we aimed to characterise the dynamic process of attention switching across different age groups, and concurrently to clarify how frequent switching instructions impact sustained attention to speech in younger and older listeners. To investigate the neural encoding of the streams during sustained attention, mTRF models were also fit excluding 1 s after each attention switching cue, thus only considering blocks of sustained attention from our switching paradigm.

To compare the neural tracking of the envelope between older and younger listeners, standard negative and positive components, N1 and P2, were identified on the TRF weights based on a Gaussian function fitting: $a(t) = p * exp(-\left(\frac{t-l}{w}\right)^2)$, with t being time, p the deflection - either positive or negative, w representing the width of the deflection and l its latency. Boundaries for the search of these parameters were set for each component separately, with N1: p = [-5 0], l = [70 150], w = [0 50]; and P2: p = [0 5], l = [100 300], w = [0 50]. This procedure and the choice of boundaries were taken from (Borges et al., 2025).

As in Study 2, encoding TRF models were also used to yield a fine-grained temporal representation of neural tracking of the two streams around the switching cue, as the switching process unfolds. First, speech features of attended blocks from the left and right streams were combined to construct an Attended speech model. These attended stimulus features were then related to the EEG activity, resulting in an encoding Attended TRF model. For each trial, EEG signals were predicted using the Attended TRF model in combination with the speech features of the left stream and the right stream, separately. This resulted in two predicted EEG signals, both based on the Attended model, but differing in whether the speech features belonged to the left or the right stream. Those predicted EEG signals were then individually correlated with the true EEG signal for that trial and, to obtain a temporally detailed representation, the correlation was performed over a sliding window of 4 s with a step size of 0.125 s.

The correlation values obtained with the sliding window approach define the temporally evolving neural tracking of left stream and right stream around the switching cue. In Study 2, it was shown that the switching process could be interpreted in terms of disengagement from the currently attended stream and engagement to a new stream, and that for a very short amount of time the two streams were simultaneously represented, before the new stream was fully attended. In this study, the disengagement and engagement processes were compared between younger and older listeners, to consider potential age-related differences in those temporal

dynamics. Furthermore, the introduction of frequent topic changes within each stream makes the attention switching task even more complex, potentially impacting the speed or trajectory of disengagement and engagement. As before, for each participant, piecewise linear regression was applied to EEG prediction correlations around the attention-switching cue to identify the onset and offset of disengagement and engagement. The model assumed constant pre- and post-switch values and estimated the breakpoints connecting them, finding the start and end points of the two processes. Breakpoint initialisation followed logical assumptions (e.g., no onset could be localised before the switching cue), and fitting was performed with MATLAB's *lsqcurvefit* to minimise the error between the model and the data.

5.2.9 Event-Related Spectral Perturbation (ERSP)

Following the procedure used in Study 2, the time course of neural activity power in the alpha band was computed as an event-related spectral perturbation (ERSP). First, EEG data was filtered between 0.5 and 20 Hz, and it was cleaned from eye-movement and blinking artifacts performing Independent Component Analysis (ICA) with standard EEGLAB functions (Delorme & Makeig, 2004). Then, for each participant, trial and attention switch, a switching segment was defined around the switching cue (± 12 s). For consistency with the neural tracking analysis, power spectral density (PSD) was estimated over the switching segment with a sliding window of 4 s and step size of 0.125 s. Within each window, PSD was computed using Welch's method and summed up across frequency bins corresponding to the alpha band (8-12 Hz). The first sliding window before the switch served as the baseline, i.e. the reference level of spectral power. For each successive sliding step, the estimated alpha-band PSD was compared to the PSD of the baseline, yielding an ERSP value (converted in dB), which indicates the relative change of the alpha power over the course of the switch compared to the baseline.

5.2.10 Event Related Potential (ERP) time-locked to the switch

In order to assess overall changes in neural activity due to the attention switching instruction, an ERP analysis was performed by taking a window ranging from -1 s to +3 seconds around the cue. The average activity in the pre-switch window [-1 s to 0 s] was taken as the baseline and subtracted from the activity time course of the whole window. The ERP was computed within-subject, for each trial and electrode. Since neural activity at the switch could be impacted by noise signals, ICA was first performed on the EEG data filtered between 0.5 Hz

and 8 Hz to remove cardiac and ocular activity, as judged based on the components' time course and spatial topography.

5.3 Results

5.3.1 Behavioural performance

At the end of each trial, participants were presented with three multiple-choice questions, focussing on their comprehension of the speech content, their preference between left and right streams, and their evaluation of the difficulty in the switching task for that trial.

The ability to correctly reply to comprehension questions relating to the content of one of the two streams was not significantly impacted by age, with a similar performance between the two groups (Younger: 86% correct, SEM 2.2%; Older: 86.8% correct, SEM 1.7%). The preference between left and right streams also did not differ significantly between age groups, and did not show a systematic bias towards left or right streams for none of the two groups, as assessed running a repeated-measures ANOVA, with Age (Younger vs. Older) as a between-subjects factor and Stream (Left vs. Right) as a within-subjects factor. This analysis revealed no main effect of Stream, ($F(1,36) = 0.25$, $p = 0.62$, $\eta_p^2 = 0.007$), and no main effect of Age group, ($F(1,36) = 1.1e\text{-}28$, $p = 1$, $\eta_p^2 = 2.5e\text{-}16$). The interaction between Stream and Age group was also non-significant, ($F(1,36) = 1$, $p = 0.31$, $\eta_p^2 = 0.028$). Regarding the perceived level of difficulty of the task, older participants reported a higher experienced difficulty (3.1 on average, with a SEM of 0.16) compared to the younger group (2.4 on average, with a SEM of 0.16), on a difficulty scale from 1 (*very easy*) to 5 (*very hard*). However, the comparison between the two groups did not reach statistical significance (two-sample *t*-test: $p = 0.15$).

5.3.2 Neural encoding of speech features in younger and older listeners during an attention switching task

In this study, two groups of younger and older participants were immersed in a multi-talker scenario requiring dynamic switching of attention between two speech streams presented in the foreground, while ignoring a 16-talker noise babble in the background. As a first step, the neural tracking of speech properties relating to acoustics and lexical segmentation was compared between younger and older listeners, with an encoding mTRF model jointly fit on envelope (Env), envelope derivative (EnvDer) and word onset (WO) features. Since the goal

was to evaluate the neural encoding within stable sustained attention blocks, the model was fit with the exclusion of 1 s of data following each attention switching cue.

The resulting TRF weights are plotted in **Figure 5.1**, along with scalp topographies representing the weights' spatial distribution at relevant time points. The main interest was to compare latency and magnitude of the N1 and P2 auditory components between the two age groups. These components were identified by fitting a gaussian function on the single-subject Env TRF weights (**Figure 5.1 A**) and were found to occur at similar latencies across younger and older listeners (**Figure 5.1 B, top**). For the latency of the N1 component, the difference was not statistically significant, but a medium effect size was found, suggesting a potentially meaningful group difference that may not have been detected due to limited sample size (two-sample t-test: $t(36) = 1.71$, $p = 0.09$, Cohen's $d = 0.54$). For the P2 component, no significant difference in latency emerged (two-sample t-test: $t(36) = 0.2$, $p = 0.83$, Cohen's $d = 0.06$).

Younger and older listeners were then compared in terms of the N1-P2 complex, i.e. the difference between the amplitude of the P2 and N1 components (**Figure 5.1 B, bottom**). Here, a statistically significant difference was found, with a larger N1-P2 complex for the younger compared to the older group (two-sample t-test: $t(36) = 2.66$, $p = 0.01$, Cohen's $d = 0.84$). **Figures 5.1 C** and **D** represent TRF weights and scalp topographies for the EnvDer and the WO features, respectively. For the EnvDer TRF, N1 and P2 components were identified as before with a gaussian fit, and they show a pattern broadly consistent with the results on the Env TRF. For the WO feature, the minimum of the N400-like component was found by selecting a window of interest around the N400 latency (between 300 and 500 ms).

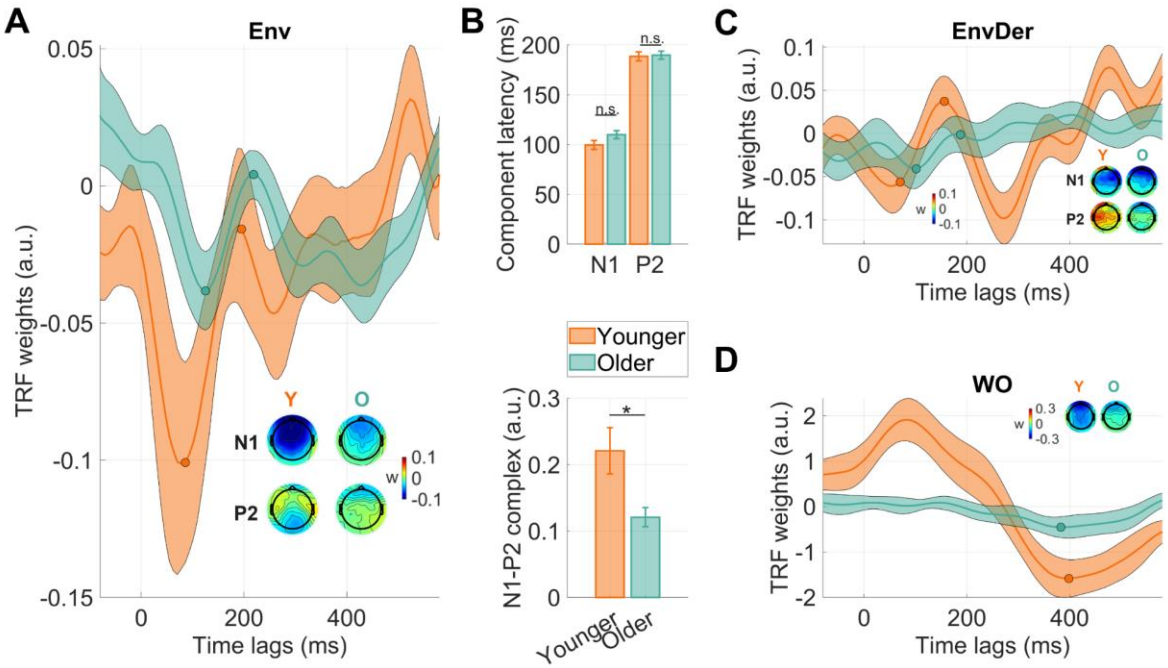


Figure 5.1. Neural encoding of speech features in an attention-switching paradigm across younger and older individuals. A forward mTRF model was fit on sustained attention blocks (excluding 1s after each switching cue), to assess EEG responses within stable attentional demands. **(A)** Average mTRF weights for the speech envelope (Env) in younger (orange) and older (green) listeners, at channel Cz. Shaded areas represent the standard error of the mean (SEM) across participants. Highlighted dots on each line indicate N1 and P2 auditory components, as identified by a gaussian fit. Scalp topographies show the distribution of responses across channels for age-specific N1 and P2 time latencies. **(B) Top:** Average time latency of N1 and P2 responses in the younger and older population. **Bottom:** Average N1-P2 amplitude differences across age-groups. Error bars represent the SEM across participants, and asterisks indicate significant statistical effects (paired sample t-tests; *$p \leq 0.05$; **$p \leq 0.01$; **$p \leq 0.001$; n.s. indicates a non-significant effect). **(C)** Average mTRF weights for the derivative of the speech envelope (EnvDer) at channel Cz. As above, dots on the lines indicate N1 and P2 components. Topographies report the scalp distribution of N1 and P2 components averaged across participants separately for each age-group. **(D)** mTRF weights averaged across participants for the word onset (WO) feature at channel Cz. Highlighted dots indicate the time latency at which the mTRF weights reach their lowest amplitude, corresponding to the minimum of an N400-like response. Topographical maps represent the scalp distribution at the highlighted latency.

5.3.3 Frequent attention switching disrupts sustained speech tracking, particularly in older listeners

Older listeners have often been shown to display an increased envelope tracking response compared to younger listeners, both for single-speaker and multi-talker scenarios (Decruy et al., 2019; Zan et al., 2020). This enhanced speech tracking has been related to a potential compensatory activation of areas responsible for speech processing in the ageing brain.

In the current study, however, a different pattern of results emerged when comparing the neural tracking of target and masker streams across the two age groups. Encoding mTRF

models (with Env, EnvDer and WO features) were fit on the sustained attention blocks of target and masker speech separately, allowing to quantify the neural tracking response to the competing streams as the resulting EEG prediction correlations. Those prediction correlations scores were then compared between target and masker models, with the hypothesis of a significant attentional bias, and between younger and older listeners, with the expectation of an increased speech tracking in the older group (**Figure 5.2 A**). A repeated-measures ANOVA confirmed the first hypothesis, showing a statistically significant target vs. masker bias ($F(1,36) = 13.9$, $p = 6.47\text{e-}4$, $\eta_p^2 = 0.28$) for the within-subjects effect of Stream (Target vs. Masker). However, contrary to the expectations, the between-subjects factor of Age (Younger vs. Older) revealed an overall stronger tracking response for the younger compared to the older listeners ($F(1,36) = 9.7$, $p = 0.004$, $\eta_p^2 = 0.21$). Even more strikingly, when examining the statistically significant interaction Stream*Age ($F(1,36) = 6$, $p = 0.019$, $\eta_p^2 = 0.14$), a significant attentional bias was only found for the younger and not for the older age group (post-hoc comparisons, Younger: $t(36) = 4.38$, $p = 5\text{e-}4$, Cohen's $d = 1$; Older: $t(36) = 0.91$, $p = 0.73$, Cohen's $d = 0.21$). Neural tracking target vs. masker differences are shown at the single-participant level in **Figure 5.2 B**.

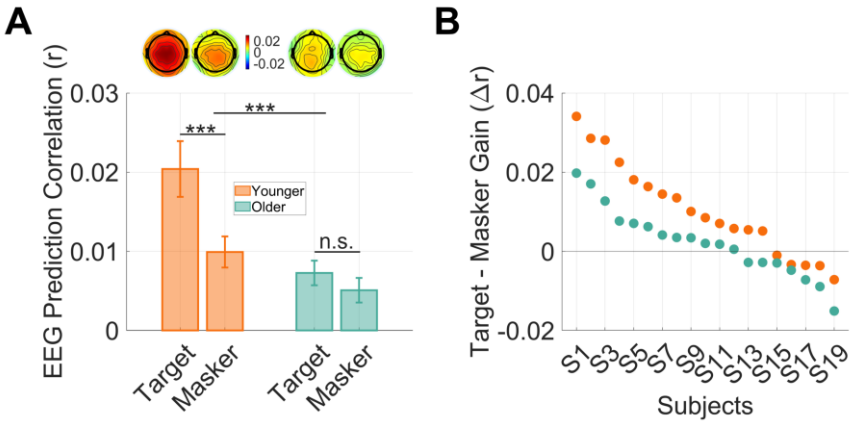


Figure 5.2. Neural tracking of competing speakers and attentional bias are strongly reduced in older listeners. (A) EEG prediction correlations for sustained attention blocks of target and masker streams, from an encoding mTRF model jointly fit on Env, EnvDer and WO features. Bars represent the average across participants and electrodes, error bars indicate the SEM across participants. Scalp topographies report the distribution of prediction correlation values across channels. **(B)** Attentional bias (target – masker EEG prediction correlations) within participant for the younger (orange) and older (green) age groups.

A fundamental difference of the current paradigm compared to previous studies is the frequent reorienting of attention (eight switches per trial) that participants were instructed to perform. As such, the pattern of results described above could potentially be explained by the dynamic attention switching between left and right streams. To clarify this aspect, we started

by fitting target and masker TRF models only selecting the first part (first four sustained attention blocks, B1-B4) and the second part (last four sustained attention blocks, B6-B9) of each trial, with the aim to assess if neural tracking decreases over the course of the trial and if the trajectory of this process differs between younger and older participants, and between target and masker streams. To do so, we fit a repeated-measures ANOVA on the EEG prediction correlation values, examining the impact of the within-subject factors of Stream (Target vs. Masker) and Trial Part (First part vs. Second part), and between-subject factor of Age (Younger vs. Older). This analysis (**Figure 5.3 A**) revealed that the neural tracking of the target stream is significantly enhanced compared to the masker's (Stream: $F(1,36) = 15.73$, $p = 3.32\text{e-}4$, $\eta_p^2 = 0.3$), and that, overall, younger listeners display a significantly increased neural tracking compared to the older group (Age: $F(1,36) = 12.31$, $p = 0.001$, $\eta_p^2 = 0.26$). Importantly, neural tracking significantly decreases in the second part of the trial relative to the first part (Trial Part: $F(1,36) = 6.72$, $p = 0.014$, $\eta_p^2 = 0.16$), with no statistically significant Stream*Trial Part interaction ($F(1,36) = 0.88$, $p = 0.35$, $\eta_p^2 = 0.02$), indicating that the change in the tracking response displays similar trajectories across target and masker streams. Furthermore, the attentional bias (target-masker difference) remained stable over the course of the trial for both age groups, as indicated by the lack of a significant three-way interaction between Stream, Trial Part and Age ($F(1,36) = 0.002$, $p = 0.96$, $\eta_p^2 = 6.65\text{e-}5$).

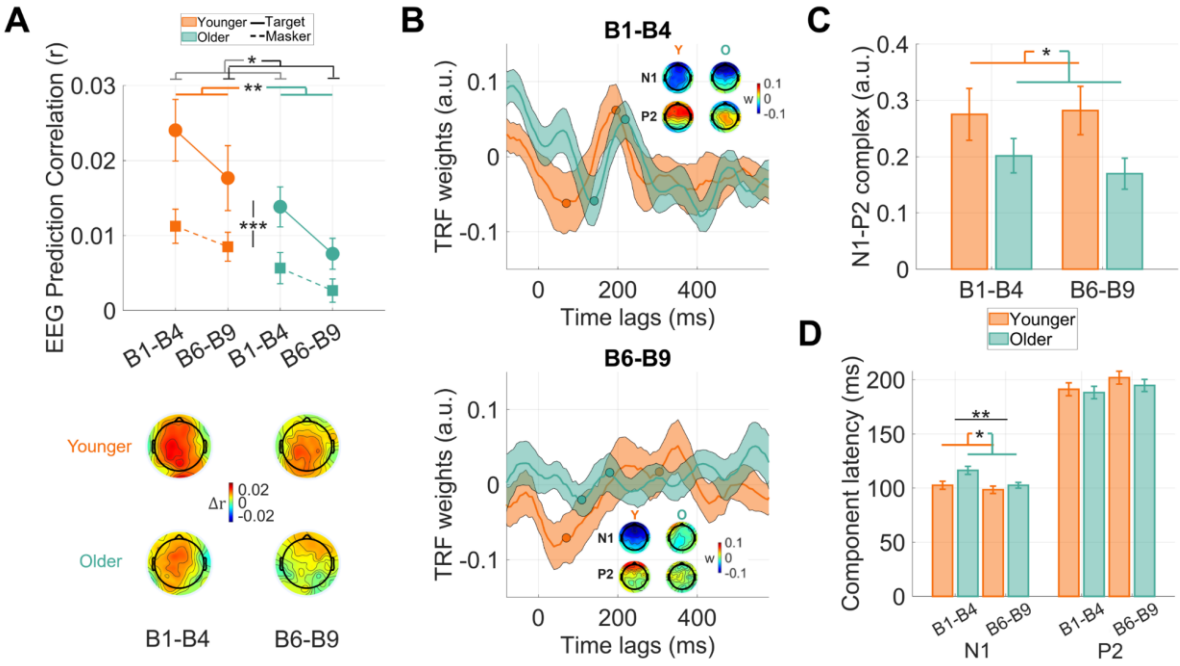


Figure 5.3. Frequent attention switching impacts sustained attention to speech. (A) Top: EEG prediction correlations across the first (B1-B4) and second (B6-B9) halves of each trial for younger (orange) and older (green) listeners. Thick lines and circle marker indicate the target, dotted lines and

square marker indicate the masker. Error bars show the SEM across participants. Asterisks represent statistical significance (mixed-design ANOVA; *$p \leq 0.05$; **$p \leq 0.01$). **Bottom:** Topographical patterns of the EEG prediction correlation gain (target – masker) for older and younger listeners for the first (B1-B4) and second (B6-B9) halves of each trial. **(B)** Speech envelope mTRF weights for younger (orange) and older (green) participants. Lines represent the average across participants for channel Cz, while the shaded area defines the SEM across participants. N1 and P2 components are indicated on each line as a filled dot, with corresponding scalp topographies showing the distribution of the TRF weights over the scalp. Top: first sustained attention blocks (B1-B4); Bottom: last sustained attention blocks (B6-B9). **(C)** Average N1-P2 amplitude difference across B1-B4 and B6-B9, for younger and older groups. Error bars represent the SEM across participants, and asterisks indicate significant statistical effects (paired sample t-tests; *$p \leq 0.05$; **$p \leq 0.01$; **$p \leq 0.001$; n.s. indicates a non-significant effect). **(D)** Time latency of N1 and P2 auditory components across B1-B4 and B6-B9, averaged across participants. Error bars represent SEM across participants.

Beyond the comparison of EEG prediction correlations, we also examined the model's TRF weights to identify potential differences between the first and second halves of the trials across age groups. Specifically, we focussed on the Envelope feature for the target stream (**Figure 5.3 B**) to assess the amplitude and latency of the N1 and P2 TRF components, consistent with the earlier analysis across the full trial length. First, we compared the amplitude of the N1-P2 complex across Trial Parts (within-subject factor, First part vs. Second part) and across age groups (between-subject factor, Younger vs. Older). As in the previous analysis, younger listeners were shown to display an increased amplitude of the N1-P2 complex (**Figure 5.3 C**) compared to older individuals ($F(1,36) = 4.64$, $p = 0.038$, $\eta_p^2 = 0.11$). In contrast, no statistically significant difference emerged depending on the trial part ($F(1,36) = 0.16$, $p = 0.69$, $\eta_p^2 = 0.005$), nor did the effect vary by age group, as shown by the lack of a significant Age*Trial Part interaction ($F(1,36) = 0.39$, $p = 0.53$, $\eta_p^2 = 0.01$). Latencies of the N1 and P2 components were compared separately (**Figure 5.3 D**), as before, for the within-subject Trial Parts factor (First part vs. Second part), and the between-subject of Age (Younger vs. Older) in a repeated-measures ANOVA. The N1 latency was shown to be overall significantly delayed in older listeners ($F(1,36) = 5.3$, $p = 0.027$, $\eta_p^2 = 0.13$), and it was also impacted by the trial part, occurring earlier in the second part of the trials ($F(1,36) = 10.18$, $p = 0.003$, $\eta_p^2 = 0.22$). This effect did not vary depending on the age group, as indicated by the lack of a significant Trial Part*Age interaction ($F(1,36) = 3$, $p = 0.09$, $\eta_p^2 = 0.08$). The P2 latency, on the other hand, did not vary as a function of trial part ($F(1,36) = 2.6$, $p = 0.11$, $\eta_p^2 = 0.07$) or age group ($F(1,36) = 0.67$, $p = 0.42$, $\eta_p^2 = 0.02$), and the effect did not change depending on their interaction ($F(1,36) = 0.16$, $p = 0.69$, $\eta_p^2 = 0.004$).

To complement the main findings on the reduction of neural tracking over the course of the trial, we performed an additional analysis, fitting target-mTRF models (as before, based

on Env, EnvDer and WO features) on single, isolated sustained attention blocks (from B1 to B9). Results from this temporally-detailed analysis (**Supplementary Figure 5.3 A,** ***Appendix***) were tested with a linear mixed model, with Age (Younger vs. Older) and Block (continuous variable, B1-B9) as fixed effects and Subjects modelled as a random effect. A statistically significant effect of Block was found ($F(1,136.94) = 25$, $p = 1.72e\text{-}6$), indicating a reduction in neural tracking over the trial, along with a significant Age*Block interaction ($F(1,136.94) = 6.14$, $p = 0.014$), suggesting that the decrease of neural tracking follows a different trajectory between younger and older listeners. Indeed, a follow-up trend analysis revealed that the slope of the change in neural tracking over attention blocks was not significantly different from zero for the younger age group (β = -0.001, SE = 6.1e-4, z = -1.78, p = 0.075), while it was significantly different from zero for the older age group (β = -0.003, SE = 6.1e-4, z = -5.29, p = 1.24e-7). The difference in slopes between the two age groups was also significant ($\Delta\beta$ = -0.002, SE = 8.58e-4, z = -2.48, p = 0.013). Thus, this block-specific analysis indicates that older individuals show a sharper decay in their neural tracking of the target stream during the trial, compared to the younger group. It is interesting to note that the decrease in neural tracking of the target was only observed within-trial, and not over the whole experiment's duration (**Supplementary Figure 5.3 B,** ***Appendix***). This was found by examining the single-trial EEG prediction correlations from an mTRF model fit on the whole experiment, including all sustained attention blocks from each trial. A linear mixed effect model fit on this data revealed only a significant effect of Age, with younger participants showing higher EEG prediction correlations than older participants ($F(1,36) = 6.19$, $p = 0.018$), but no significant Trial Number effect ($F(1,36) = 0.03$, $p = 0.86$), nor a significant Trial Number*Age interaction ($F(1,36) = 0.98$, $p = 0.33$). This suggests that the neural tracking over the experiment's duration was overall stable across trials, with marked differences only driven by the participants' age.

5.3.4 Attention switching elicits a P300-like response

An ERP time-locked to the attention switching cue was computed for older and younger listeners separately, to characterise the temporal dynamics of neural activity around the switching cue. This analysis revealed a modulation of activity elicited by the attention switching instruction (**Figure 5.4 A**), with one main positive (P1) and two main negative (N1, N2) components. The amplitude of the N1 component was found to significantly differ between the two age groups (**Figure 5.4 B**),, with the older participants displaying a significantly deeper trough (two-sample t-test: $t(37) = 2.1$, $p = 0.045$, Cohen's $d = 0.66$), while no amplitude

difference was found between the N2 (two-sample t-test: $t(36) = 1.49$, $p = 0.15$, Cohen's $d = 0.47$) and P1 components (two-sample t-test: $t(36) = 0.21$, $p = 0.83$, Cohen's $d = 0.07$).

When comparing their latencies (**Figure 5.4 C**), instead, a statistically significant difference only emerged for the N2 component (two-sample t-test: $t(36) = -3.1$, $p = 0.004$, Cohen's $d = -0.98$), occurring at longer latencies for older participants. There was a trending significance for the latency difference between P1 components (two-sample t-test: $t(36) = -1.91$, $p = 0.065$, Cohen's $d = -0.6$), which indeed showed a medium effect size, but did not reach standard significance. No difference was found for the latencies of the N1 component (two-sample t-test: $t(36) = -0.88$, $p = 0.38$, Cohen's $d = -0.27$) between older and younger age groups.

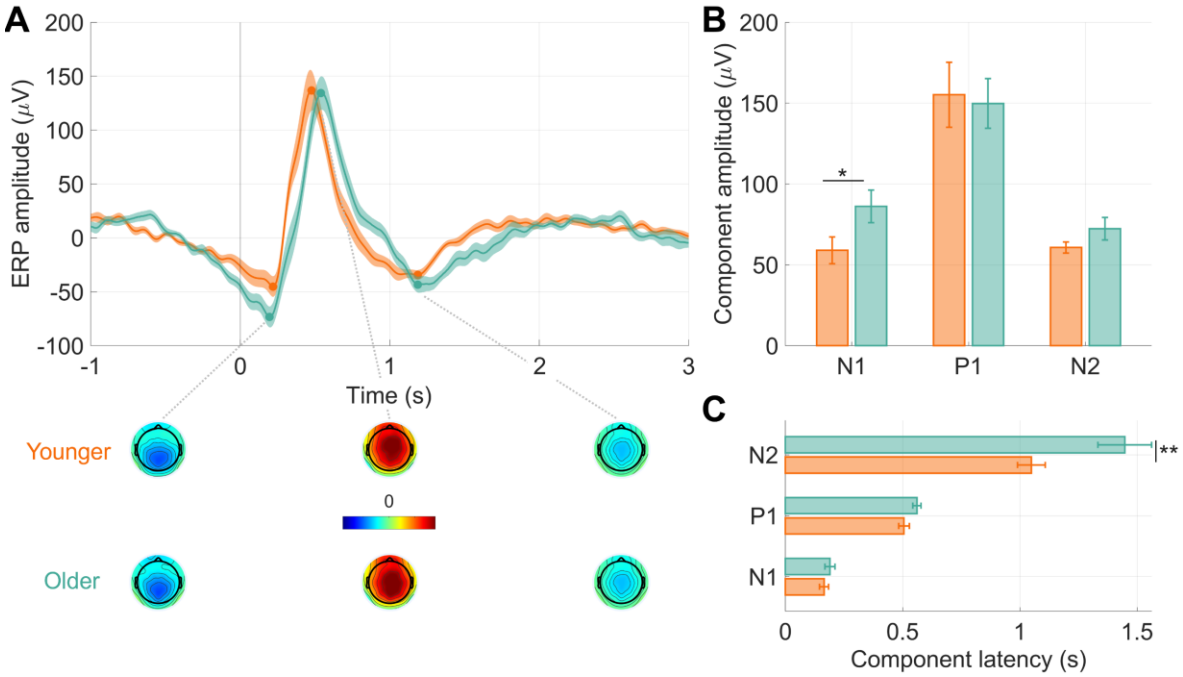


Figure 5.4: Brain activity time-locked to the attention switching cue displays a P300-like response. **(A)** Event-related potential (ERP) time-locked to the switching cue, averaged across younger (orange) and older (green) participants. Shaded areas indicate the standard error of the mean (SEM) across participants. Highlighted dots indicate the main N1, P1 and N2 components of the response. Scalp topographies for the components are reported below, separately for younger and older individuals. **(B)** Amplitude of the N1, P1, and N2 components for the two age-groups. Error bars indicate SEM across participants. **(C)** Time-latency of the components for younger and older participants. As above, error bars correspond to the SEM across participants. Asterisks indicate statistically significant comparisons (independent sample t-tests; *p ≤ 0.05; **p ≤ 0.01; **p ≤ 0.001).

5.3.5 Neural tracking around the switch in younger and older listeners

Dynamic attention reorienting was examined, as in Study 2, using a fine-grained neural tracking analysis around the switching cue. For each participant, encoding mTRF models (fit

on Env, EnvDer and WO features) were trained on the target speech stream and then tested separately on Spk1 and Spk2. EEG signals predicted based on Spk1 and Spk2 were correlated with the true EEG time course using a sliding window, and the resulting correlations were averaged across participants and electrodes, to yield the mean tracking of the two competing streams around the switch. Differently from the previous study and despite using the same reference window length (4 s), in this experiment the tracking estimation appeared noisier, particularly for the older age group (**Figure 5.5 A, right panel**), for which the difference between Spk1 and Spk2 stream did not even reach significance after correcting for multiple comparisons (paired two-tailed t-test of Spk1-Spk2 difference against zero, FDR-corrected for multiple comparisons, $p < 0.05$). For the younger age group (**Figure 5.5 A, left panel**), instead, the neural tracking of Spk1 and Spk2 was shown to follow the switching instruction with a statistically significant different tracking pattern, despite an overall noisier estimate compared to the previous dataset. The same pattern of results was observed also when computing the correlations with a window of 8 s, suggesting that the issue, particularly with older listeners, might lie in serious inter-participant variability, in addition to low neural SNR.

As such, the estimation of the temporal dynamics of attentional disengagement and engagement was also presumably impacted by data quality. As in Study 2, initial and final points of the two processes were compared separately, with the addition of the comparison between older and younger participants in this experiment (**Figure 5.5 C**). A repeated-measure ANOVA was run on the 4s-window EEG prediction correlations, to assess the effect of Process (Disengagement vs. Engagement), and Age (Younger vs. Older).

For the start time of engagement and disengagement, the analysis revealed a marginal difference between the two processes ($F(1,36) = 3.95$, $p = 0.055$, $\eta_p^2 = 0.1$). The medium effect size suggests that disengagement tended to begin earlier than engagement (post-hoc t-test: $t(1,37) = 1.99$, $p = 0.055$, Cohen's $d = 0.43$). No significant difference emerged across age groups ($F(1,36) = 0.09$, $p = 0.77$, $\eta_p^2 = 0.002$), nor in the Age*Process interaction ($F(1,36) = 0.33$, $p = 0.57$, $\eta_p^2 = 0.009$). When jointly considering multiple sliding window lengths (1s, 2s, 4s, 8s, 16s), a significant effect of Process was found ($F(1,36) = 4.14$, $p = 0.049$, $\eta_p^2 = 0.1$), with the disengagement starting earlier than the engagement (post-hoc t-test: $t(1,37) = 2.03$, $p = 0.049$, Cohen's $d = 0.29$), as hinted by the trend found in the 4 s window analysis. Considering the number of levels in the Window factor (5 windows), this repeated-measures ANOVA was corrected with a Greenhouse-Geisser correction for violating the assumption of sphericity. Even after the correction, the Window factor proved to be statistically significant ($F(1,36) = 3.53$, $p = 0.034$, $\eta_p^2 = 0.09$), with longer windows yielding progressively

longer time estimates. Estimates of the start time for disengagement and engagement did not differ significantly across age groups ($F(1,36) = 0.06$, $p = 0.8$, $\eta_p^2 = 0.002$), and none of the interactions of Age, Process, and Window reached significance.

Regarding the end time of engagement and disengagement, a repeated-measures ANOVA including Process (Engagement vs. Disengagement) and Age (Younger vs. Older) found a trending significance for the difference between the two processes (Process: $F(1,36) = 3.87$, $p = 0.057$, $\eta_p^2 = 0.1$), suggesting that the disengagement may finish earlier than the engagement (post-hoc t-test: $t(1,37) = 1.97$, $p = 0.057$, Cohen's $d = 0.47$). No significant difference was observed across age groups ($F(1,36) = 0.39$, $p = 0.54$, $\eta_p^2 = 0.01$) nor in the Age*Process interaction ($F(1,36) = 0.16$, $p = 0.69$, $\eta_p^2 = 0.004$). Considering multiple window lengths, as above, yielded a pattern of results similar to those specific to the 4s window. The repeated-measures ANOVA, with factors Window, Age, Process, was once again corrected for violating sphericity assumptions (Greenhouse-Geisser), and found no statistically significant effect of Process ($F(1,36) = 2.32$, $p = 0.14$, $\eta_p^2 = 0.06$), nor Age ($F(1,36) = 2.14$, $p = 0.15$, $\eta_p^2 = 0.056$). However, as for the start time, a significant effect of Window emerged ($F(1,36) = 11.31$, $p = 2.5e\text{-}5$, $\eta_p^2 = 0.24$), indicating once again that latencies became significantly longer with longer windows. None of the interactions among the main factors resulted in statistical significance.

A key result of Study 2 was the change in alpha-band (8-12Hz) neural dynamics around the switching cue. This result was replicated in the current dataset, using a 4s sliding window, with temporal dynamics ERSP and scalp topographies comparable to those of the previous study (**Figure 5.5 B**). Importantly, also in this dataset the encoding minimum of the alpha ERSP was found to occur significantly after the encoding switch (i.e., the intersection of neural tracking values of Spk1 and Spk2 as the switch unfolds), as revealed by a repeated-measures ANOVA with factors Age (Younger vs. Older), and Metric (Alpha Minimum vs. Encoding Switch) (**Figure 5.5 E**). Indeed, this analysis indicated a significant effect of Metric ($F(1,36) = 14.32$, $p = 5.63e\text{-}4$, $\eta_p^2 = 0.28$), while no significant effect of Age was found ($F(1,36) = 0.05$, $p = 0.82$, $\eta_p^2 = 0.001$), nor an Age*Metric interaction ($F(1,36) = 7.74$, $p = 0.24$, $\eta_p^2 = 0.04$). No difference was found between younger and older listeners in the latency of the minimum of the alpha ERSP (independent-samples t-test: $t(1,36) = 0.79$, $p = 0.43$, Cohen's $d = 0.25$), nor in the amplitude of the minimum (independent-samples t-test: $t(1,36) = 0.19$, $p = 0.85$, Cohen's $d = 0.06$) (**Figure 5.5 D**).

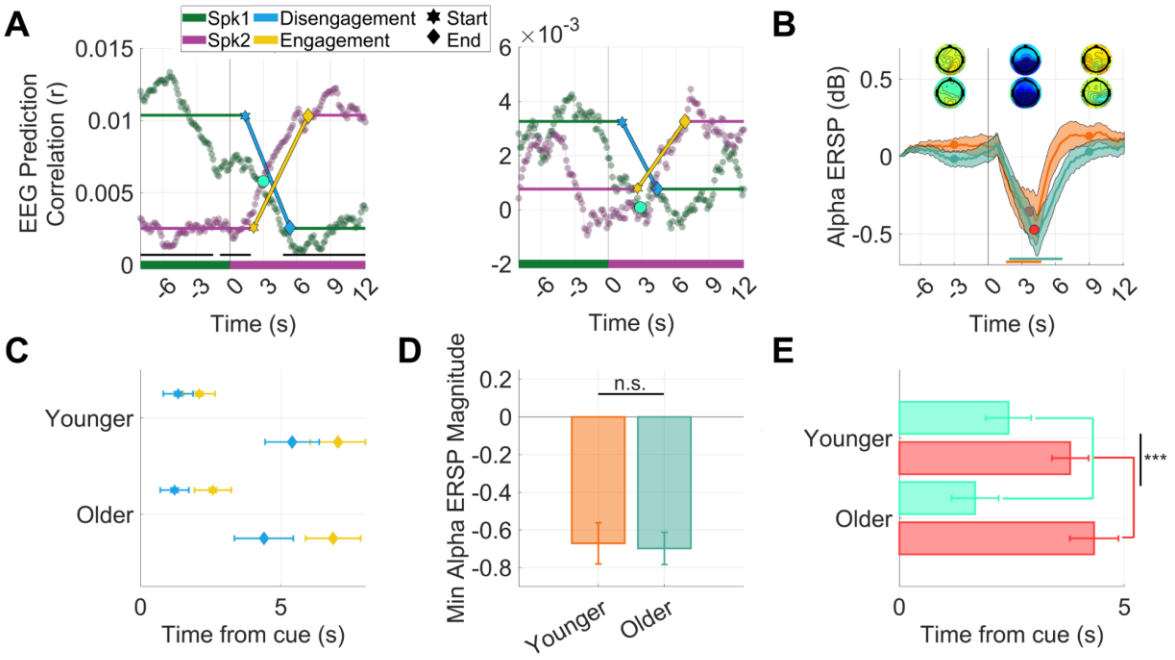


Figure 5.5: Attention switching dynamics in younger and older individuals (A) Sliding-window EEG prediction correlations (average across all channels) for younger (**left**) and older (**right**) participants, obtained from an mTRF encoding model including envelope, envelope derivative and word onset features. Correlations were computed with a window of 4 s, and a step size of 0.125 ms. Coloured bars below indicate attention instructions. The turquoise dot marks the encoding switch between speakers. A piecewise linear fit models disengagement and engagement dynamics, overlaid on bootstrapped averages (2 participants). Hexagrams indicate the onset of disengagement (blue) and engagement (yellow); diamonds mark their respective offsets. **(B)** Alpha-band (8–12 Hz) ERSP around the attention-switching cue, averaged across all channels using a 4s sliding window. Scalp topographies at chosen time points show posterior desynchronization, with a significant drop following the switch instruction (indicated by thick orange and green lines). The red dot marks the average ERSP minimum across participants; shaded areas reflect SEM. **(C)** Temporal dynamics of disengagement and engagement onset and offset for the two age-groups. (**D)** Amplitude of the alpha ERSP minimum in younger and older participants. Error bars reflect SEM across participants. **(E)** Comparison of EEG encoding switch timing (turquoise bar) and alpha ERSP minimum (red bar) using a 4s sliding window. Alpha desynchronization occurs significantly after the encoding shift in younger and older listeners. Asterisks indicate significant differences (paired t-tests; *$p \leq 0.05$, **$p \leq 0.01$).

Discussion

Chapter 4 introduced a novel experimental paradigm to study the disengagement and engagement stages of attention reorienting with continuous speech streams in a complex multi-talker scene. This chapter represents the first application of that methodology to a cohort of participants belonging to an older age group (mean age: ~61 years), whose speech processing abilities are known to evolve as a consequence of age-related hearing and cognitive changes (Herrmann, 2025; Herrmann & Butler, 2021; Irsik et al., 2022), with implications that are still uncertain, especially in the context of dynamic reallocation of attention. Based on the existing literature on single-talker listening and sustained attention in multi-talker scenarios, the main

differences emerging in older listeners may lie in the representation of the speech acoustics, with a potentially enhanced neural encoding of the target stream compared to the younger group (C. Brodbeck et al., 2018; Decruy et al., 2019; Herrmann, 2025), and overall slower switching times, due to their decreased inhibitory control and working memory abilities (Borella et al., 2008; Salthouse, 2010; Stenbäck et al., 2021).

To address this question, we collected a novel EEG dataset of 19 younger and 19 older participants, instructed to switch their attention between two TED talks (pronounced by artificial voices) delivered through two spatially separated loudspeakers located in front of the listener, while also ignoring unintelligible babble noise coming from four loudspeakers in the background. Participants additionally underwent a separate cognitive testing session, where their fluid intelligence, visual working memory and processing speed were assessed with a non-verbal computer-based test battery.

This attention-switching study indicates that 1) older adults show a decreased neural tracking of speech compared to younger adults, with a decay unfolding over the course of each trial; 2) auditory components (N1, P2) occur at similar latencies across the two age groups, while the amplitude of the N1-P2 complex is consistently reduced in older listeners; 3) the switching cue induces a characteristic P300-like ERP response in both younger and older populations 4) disengagement from Spk1 and engagement to Spk2 display similar temporal dynamics between younger and older listeners, however the estimate is inconsistent for older adults; 5) the change in alpha ERSP follows a similar pattern across age groups, and the minimum of the alpha ERSP consistently follows the encoding switch time, in both younger and older participants.

The finding that older listeners display reduced neural tracking of speech and attentional bias compared to younger listeners is at odds with previous studies which instead reported an exaggerated tracking response in older (C. Brodbeck et al., 2018; Decruy et al., 2019; Herrmann, 2025; Karunathilake et al., 2023; Presacco et al., 2016a, 2016b; Zan et al., 2020) and hearing-impaired populations (Lien Decruy et al., 2020; Fuglsang et al., 2020; Gillis et al., 2022; Goossens et al., 2018; Millman et al., 2017; Schmitt et al., 2022). Since this experiment featured an unprecedented attention switching paradigm for this population, the reason for the difference could be attributed to a decay of neural tracking over the course of the trial, due to the frequent switching instructions, which may have particularly impacted older listeners. Our mTRF models fit separately on the first (B1-B4) and second (B6-B9) parts of the trials provided support for this interpretation, with a consistent decrease in neural tracking across age groups, for both target and masker streams. The fact that the two competing streams were both less

tracked in the second part of the trial, regardless of their attentional status, suggests that this phenomenon may be caused by a general withdrawal from the task, potentially due to a heightened listening effort, rather than an increasing distraction by the masker stream. While the trajectory of the neural tracking decay was similar across age groups in the mTRF models fit on first vs. second parts of the trials, the additional analysis with models fit on single sustained attention blocks (B1-B9) revealed that neural tracking of the target stream declined more sharply and rapidly in older compared to younger adults. To confirm that the reduction in neural tracking is indeed due to an increasing cognitive load introduced by the switching task, future analyses of this dataset could leverage the cognitive tests and the metrics of alpha ERSP collected as part of the study and explore their relationship to the rate or magnitude of the decay in neural tracking scores over the course of the trial.

Examining the auditory components identified on the target envelope mTRF weights, age-related differences emerged in the amplitude of the N1-P2 complex, which was found to be larger for younger listeners. This may reflect an overall weaker speech encoding in the older group, consistently with our findings on the neural tracking. Regarding component latencies, the P2 was found to occur at similar latencies in younger and older listeners, both in the model fit on the whole trial length, and in those fit on the first and second parts of the trials. The N1 component, instead, was overall delayed in older listeners (with only a trend towards significance in the whole-trial mTRF model), suggesting additional processing demands, possibly linked to reduced hearing abilities. Interestingly, we also observed an earlier N1 latency in the second part of the trial across both age groups, which may reflect adaptation to the switching task, or changes in the allocation of attention over the trial. The study of amplitude- and latency-related effects of N1 and P2 components across age groups has a relatively long tradition, with many ERP paradigms investigating it by using tones, syllables or simple speech stimuli (Burkhardt et al., 2022; McCullagh & Shinn, 2013; Rufener et al., 2014; Tremblay et al., 2004; Tremblay et al., 2002). More recently, also studies with continuous speech stimuli have looked into how ageing impacts the timing and amplitude of standard auditory components (Gillis et al., 2023; Karunathilake et al., 2023). However, results from those studies are not uniform, likely due to the high variability in experimental paradigms, methods, and stimulus types. As a general pattern, the N1 is often found to be delayed in older compared to younger listeners (Rufener et al., 2014; Tremblay et al., 2004; Tremblay et al., 2002; Tremblay et al., 2003), even though some studies identify it at similar latencies across age groups (Getzmann & Falkenstein, 2011; Getzmann et al., 2016; Karunathilake et al., 2023). Similarly, the P2 is very often found to occur at later latencies in older listeners (Burkhardt et

al., 2022; Getzmann et al., 2016; McCullagh & Shinn, 2013; Rufener et al., 2014; Tremblay et al., 2004; Tremblay et al., 2002; Tremblay et al., 2003). In terms of amplitude, the N1 is found to be larger, or not significantly different in older listeners relative to younger adults, while the P2 is generally not affected by the age factor (Getzmann et al., 2016; Karunathilake et al., 2023; Tremblay et al., 2004; Tremblay et al., 2002; Tremblay et al., 2003) or, in some studies, it is found to be smaller in older listeners (Getzmann & Falkenstein, 2011; Rufener et al., 2014). The results emerging from this study are overall in line with the literature when considering the latencies of N1 and P2. Although their reduced amplitude in older compared to younger listeners contrasts with previous studies, it is consistent with the diminished neural tracking scores observed here.

Reorienting of attention in this experiment requires first recognising that a switching cue has appeared (i.e., the arrow on the screen flips to the other side) and then investing cognitive resources to inhibit the currently attended stream and redirect attention to the competing speech. Computing an ERP time-locked to the switching cue, it was possible to observe a robust response, spanning approximately two seconds after the cue appears, and consisting of a negative deflection, followed by a positive peak and another broader negative component. In the present paradigm, since there is no silent gap after the cue, nor a preparation time for the switch, there is naturally an overlap between the process of detecting the cue and that of reorienting attention. As such, mapping the observed negative and positive deflections to standard ERP components is not straightforward. Based on the task demands, the first negative and positive components may be related to the detection of the switching cue, similarly to a N200-P300 complex (Folstein & Van Petten, 2008; Polich, 2007) which is often observed following a cue in go/no-go paradigms (Harper et al., 2014; Huang et al., 2025; Kopp et al., 2020). The later negative component, instead, may be linked to the Contingent Negative Variation (CNV) (Birbaumer et al., 1990; Jongsma et al., 2013; Walter et al., 1964), a slow ERP usually measured between a warning stimulus (in this case, the switching cue), and a following event requiring a response (here, the shifting of attention between speakers). The return to baseline could then be interpreted as the conclusion of the switching task and thus signal that attention has been reoriented. From this perspective, it is interesting to observe a significant difference in the latency of the CNV between younger and older listeners, which occurs later for the older group.

The temporally-detailed neural tracking around the switch revealed similar engagement to Spk2 and disengagement from Spk1 dynamics across both age groups. Unlike the previous study, in this experiment there was a tendency of the disengagement to precede the engagement

process, across both their start and end time points. However, those results should be interpreted with the caveat of the lack of a proper EEG denoising, which will be discussed in detail below. In terms of alpha ERSP around the switching cue, we observed a similar time course for younger and older listeners, hinting at similar resource allocation to perform the switch. Consistently with Study 2, also here the minimum of the alpha ERSP was found to significantly follow the encoding switch (i.e., the intersection of neural tracking of Spk1 and Spk2 after the switching instruction), with no difference between younger and older listeners.

Despite a similar behavioural performance in comprehension questions, older listeners showed reduced neural tracking of speech and noisier estimates of speech tracking around the switching cue, compared to younger listeners. Even in the younger age group, however, the values of neural tracking around the switch revealed a variable pre-switch baseline. This suggests that, overall, the current EEG dataset presents a lower SNR compared to the dataset in Study 2 and might benefit from tailored denoising methods. Factors that would decrease the SNR in the current experiment are, for instance, an increased difficulty of the experiment due to more frequent attention switches per trial (8 vs. 6 of the previous experiment) and, particularly, the introduction of a frequent change of topic within the competing speech streams, which challenged the reliance on the previous context. Another potential difference might be introduced by the fact that the speech streams here were uttered by AI-generated voices. Future analyses will have to clarify the contribution of these factors to the attention switching ability, to devise experimental paradigms of suitable difficulty for populations of different ages and hearing abilities.

5.4 Summary

In this study, we investigated the neural encoding of competing speech streams in a dynamic attention scenario, comparing younger and older listeners. Our findings highlight an overall weaker cortical tracking of speech in older listeners, with a sharp reduction of encoding accuracy with successive attention switches. In the TF weights to the speech envelope, N1 and P2 responses were compared in amplitude and latency between younger and older listeners, revealing smaller N1-P2 amplitude in the older population and delayed temporal latencies. The switching cue initiates a robust P300-like response in the two populations, with interesting amplitude and latency differences in the main components of the response. As in Study 2, the switch prompts a drop of ERSP in the alpha frequency band, and a pattern of attention reorienting observed in its temporal unfolding through a sliding-window neural tracking

metric. Current findings regarding switching times and its relationship to the alpha metric need to be expanded to account for potential noise in the data, and correlate them to behavioural metrics.

Chapter 6. General Discussion

While the neural underpinnings of speech perception are often studied in ideally controlled laboratory experiments, speech perception in the real world entails challenges to its success. As detailed in Section 2.5, factors hindering speech perception can derive from the internal or external experience of the listener, resulting in diverse kinds of complex listening scenarios. For example, a *Cocktail Party* scenario introduces the need to ignore competing speakers to focus on one particular person of interest. Those challenging situational demands are magnified by factors internal to the listener, for instance, in the presence of hearing impairment or processing changes due to ageing.

Despite the impact of these factors on speech perception, research on speech neurophysiology has focussed on simpler listening paradigms on young normal-hearing populations. Even when modelling complex listening settings, it has overlooked fundamental aspects of how we perceive speech in our daily life, such as our ability to rapidly change our focus of attention, shifting it dynamically between multiple speakers.

This thesis takes steps towards addressing some of those knowledge gaps. One aspect of this work is to extend well-known sustained attention paradigms to participants with hearing impairment, to investigate the neural encoding of levels of speech information unexplored in this population, such as phonology. The other aspect is to move beyond static sustained attention paradigms and study the flexible reorienting of attention, which is fundamental for speech perception in multi-talker listening environments. Ultimately, this work involves various factors impacting on speech perception: hearing status, age, task difficulty and SNR, level of speech information (e.g., acoustic vs. semantic). While many of those dimensions have been studied in experiments of sustained attention, their impact is still unclear in specific populations, such as in listeners with hearing difficulties. Even less is known about how those dimensions and their interplay would translate to a scenario of attention switching.

In this Chapter, I discuss the contribution of my work, elaborating on the challenges that I have encountered, their practical and theoretical implications, and the novel questions that arise from my findings and that may inform future research.

6.1 Neural tracking beyond the envelope to reveal speech processing differences

The neural underpinnings of speech perception in listeners with hearing impairment have been investigated in recent years with studies probing both single-talker and multi-talker scenarios, with continuous speech stimuli (Alickovic et al., 2020; Seifi Ala et al., 2020). This body of work has mostly focussed on the neural tracking of acoustic features and has led to important insights into core neural mechanisms, finding that the speech envelope is often over-represented in hearing impaired listeners (Lien Decruy et al., 2020; Fuglsang et al., 2020; Gillis et al., 2022; Goossens et al., 2018; Herrmann & Butler, 2021; Millman et al., 2017; Schmitt et al., 2022). On the other hand, there are also studies reporting no difference in envelope cortical tracking relative to listeners with normal hearing (Petersen et al., 2017; Presacco et al., 2019) (see Section 2.5). However, that research has been limited in its exclusive focus on acoustic properties of speech, leaving most of the speech hierarchy unexplored. As such, it has been unclear whether that effect emerges specifically for the neural tracking of the speech envelope, or instead it is distributed across different levels of the speech hierarchy, including phonological information.

The first main finding of this thesis offers insights to fill this gap: First, we expand our understanding of attentional filtering in listeners with hearing impairment, demonstrating that there is an enhanced cortical tracking of acoustic-phonological information of the target speech in a *Cocktail Party* scene, similarly to what is observed in listeners with normal hearing (Teoh et al., 2022). Secondly, we find that information specific to phoneme onsets is robustly encoded even for the masker stream, pointing to a potentially different processing strategy due to hearing impairment.

The enhanced acoustic-phonological tracking of the target stream relative to the masker is not a surprising finding in a context of sustained attention instructions, even in the presence of hearing impairment. While a general attentional cortical bias emerges at the level of acoustic-phonological information, we observe that the variance explained by phoneme onsets is enhanced for the masker instead. This suggests an increased sensitivity towards this particular level of speech information, which may be in line with the difficulties in masker suppression reported by listeners with hearing impairment. It remains to be clarified whether this over-representation of the masker's phoneme onsets reflects a specific sensitivity towards this level of information, or instead a generic sensitivity to acoustic edges which has not been

systematically investigated in Study 1. There is also the possibility that the pattern of results observed in relation to phoneme onsets may be an effect mediated by different strategies of word segmentation in participants with hearing impairment, which have not been considered as part of this study. Future work could explore those aspects by including the half-wave rectified envelope derivative and word onsets among the speech regressors, in search of potentially relevant lower- or higher-level sources of the observed differences. Taken together, the current results point to an overall reduced action of the attentional filter on the masker at the level of phoneme onsets, which has interesting implications.

First, it highlights the importance of isolating the separate contributions of different speech features to the neural tracking response. While the speech envelope is robustly encoded in brain responses, relying on envelope tracking alone can result in overlooking feature-specific differences in the neural representation of speech. Delving into multiple levels of the speech hierarchy, instead, may reveal fundamental processing differences between diverse cohorts of listeners, which would be otherwise homogeneous in their envelope tracking response. Furthermore, as described in Chapter 2.2, there is evidence that the speech hierarchy is not serial in nature (Gwilliams et al., 2025; Hamilton et al., 2021; Leonard et al., 2024) , opening the possibility that certain layers of information, despite being nested, may be processed by parallel neural mechanisms with fundamentally different processing strategies. For example, the over-sensitivity to the masker's phoneme onsets may, counterintuitively, aid in the neural suppression of its envelope. As such, when a divergence between two cohorts emerges in terms of behaviour, subjectively-reported experience, or listening effort, it is useful to probe the neural encoding of different speech features, in search for a mechanistic explanation. As explained in Chapter 2.3.3, disentangling the neural encoding of different features from brain activity is a complex task, due to their covariance. In this context, the mTRF approach allowed to model multiple features occurring simultaneously, and to isolate their unique contribution to the neural tracking response through backward elimination (FS-S, F*sh*S-S), revealing a speech processing mechanism that was not apparent from the measure of envelope tracking.

Secondly, this also invites to consider the potential of complex listening scenarios to uncover processing differences that may not emerge in ideal listening conditions. Ultimately, most of our real-world experience with speech involves noise and disruptions of various kinds. As a result, the source of many perceptual difficulties that impact the lives of persons with hearing impairment, among others, may not be understood when only investigating the neurophysiology of speech in-quiet.

While this was the first study, to our knowledge, to investigate phonological encoding of continuous competing speech in listeners with hearing impairment, it did so with speech stimuli in the Danish language. Danish is known for its particularly complex phonology, due to the large number of vowel sounds compared to consonants (Grønnum, 2003) and the fact that consonants are often pronounced as semi-vowels, resulting in the lack of strong phonetic and prosodic cues to help word segmentation (Bleses et al., 2008). This often produces an indistinct pronunciation, especially at word endings. All these features are known to impact language acquisition milestones in Danish infants, and in second-language speakers (Gooskens et al., 2010; Grønnum, 2003; Trecca et al., 2020). In this context, it is possible that phoneme onsets in Danish are more relevant than in other languages, serving as reference cues for segmentation. It remains to be tested whether a similar pattern of results would be observed in other languages with a more transparent phonological structure.

Overall, these results support the idea that attentional focus interacts differently with different levels of the speech processing hierarchy, depending on the listener's characteristics. In this particular study, participants had mild-to-moderate hearing loss, but a similar analysis framework can be extended to different populations, where speech processing difficulties may arise from a variety of causes, not necessarily from deficits of the auditory pathway. For example, focussing attention on a particular speaker in a multi-talker scene may be more complex for second-language speakers, or for listeners with ADHD. However, the neural basis of selective attention to speech remains unclear for many of those cohorts, and this study invites further theoretical and experimental investigations to uncover the source of those differences.

6.2 A novel approach to study dynamic attention

This thesis builds upon extensive work on the neural mechanisms of speech perception in *Cocktail Party* scenarios, which has focussed on paradigms where listeners were instructed to consistently maintain attention on one speaker. As such, there is limited research on how attention can be flexibly and dynamically reoriented between different speakers, despite the real-world relevance of this phenomenon. Recent work has begun to investigate attention switching, showing that the unfolding of the switch can be reliably quantified with neural tracking metrics from EEG responses, and that the power of neural activity within the alpha frequency band is modulated by the switch, along with the pupil dilation response (Haro et al., 2022). However, a number of fundamental questions remain unanswered, especially regarding the detailed temporal dynamics of the switch, the fate of the unattended stream in a context

where attention shifts, and the very meaning of speech separation and selective attention in such a dynamic scenario.

Adopting the perspective of auditory scene analysis, the competing speech stimuli in this flexible listening scenario are separate auditory objects. The distinction between them, beyond the different spectro-temporal characteristics of the two voices (male vs. female), is also aided by spatial cues informing the listener about the location of the sound sources (left vs. right speaker). A classic problem in theories of auditory scene analysis is how to justify that a time-varying auditory stimulus can be consistently perceived as one single auditory object. This problem of *object permanence* is particularly relevant for complex auditory stimuli, such as music or speech, where the unfolding of the stimulus over time, e.g. *streaming*, is key to its meaning (Shinn-Cunningham & Lee, 2017; Shinn-Cunningham, 2008). The concept of streaming becomes even more complex in the presence of multiple sound sources, since sound features must be not only bound over time, but also assigned to the correct source, yielding multiple *streams*. Temporal coherence has been proposed as a fundamental mechanism allowing different features of a sound to be bound into a single stream, based on their correlation in time (Elhilali et al., 2009; Shamma et al., 2013; Shamma et al., 2011). As a result, when selectively focussing attention on one particular speaker in a multi-talker scene, it is sufficient to track one specific sound feature of that speaker, e.g. their pitch, to then being able to extract other temporally correlated features, disentangling it from the mixture and perceiving it as a single stream.

However, I contend there is one aspect that has been overlooked when considering the formation and permanence of separate streams in the case of complex and abstract stimuli such as speech: the semantic coherence provided by the topic. After listening to a few words or sentences of a certain speech stimulus, we are usually able to identify its latent semantics, that is, whether the speaker is talking about their dog, a hiking trip, or what they ate for dinner. This broad semantic representation, usually referred to as *topic*, becomes a relevant feature to identify and track a particular speech stream from a multi-talker scene, as it is expected to remain consistent over relatively long temporal scales. In this view, the permanence of a particular auditory object in a multi-talker scene, and the ability to track it, is the result of temporally coherent sound properties, along with temporally coherent topic information within that stream. Just as differences in pitch make it possible to tease apart two speakers in the spectro-temporal domain, differences in topic also provide a reliable cue for discriminating between competing speakers. Being a coarse and temporally robust representation of the speech content, topic detection and tracking does not even require a precise word-by-word lexical

analysis nor a complete focus of attention on the speech stimulus. Thus, with relatively little investment of higher-level resources, the listener can assign different speech streams to separate points of the scene's semantic space and leverage that broad semantic distance to further separate the competing speakers. While semantic coherence may be a useful cue for the formation and tracking of competing speech streams over time, it could hardly be investigated through current multi-talker paradigms, since they mostly consisted of static attentional scenarios, where the listener was not motivated to process the masker speech beyond a certain level.

The novel attention-switching paradigm introduced in Study 2 and Study 3 (Chapters 3 and 4) provides an opportunity to investigate the neural representation of two streams which are both, alternatively, the focus of attention. In this context, the speakers' topics become a relevant cue for identifying them as separate streams, and for rapidly reorienting attention between them. In Study 2, we were able, for the first time, to design four competing models of how previously attended context may impact the processing of the currently attended stream, and to test their ability to explain EEG data. We observed that a model based only on the current segment outperformed alternative models which took previous context into account. This approach demonstrates that our methods are sensitive to subtle differences in the neural encoding of semantic information, modelled from next-word probabilities in LLMs. However, those semantic models were based on word-level lexical information, such as entropy and surprisal, and did not explicitly model the speaker's topic. It would be interesting to explore this problem from the perspective of a topic change, treating it as a dynamic warping of the neural semantic space to match the shift in the attended speaker's topic. Furthermore, there is the possibility that this result is specific to a particular level of difficulty of the attention switching task, cohort of participants, or topic distance between attention blocks. A potentially different result, for example, may emerge in the older participants examined in Study 3, considering their possible higher reliance on previous context.

Still from the auditory scene perspective, sustained attention scenarios can be considered to probe mainly the stages of auditory object formation and selection, which determine the identification of different speech streams. Although neural tracking of a target speech stream can be reliably measured, sustained attention scenarios do not allow to analyse object formation and selection as they unfold, since attention is maintained consistently on one speaker for the whole duration of the trial. In our attention-switching paradigm, instead, we were able to measure the neural consequences of a voluntary interruption of *streaming* following an attention switching cue. Each appearance of the switching cue initiates a new

process of object selection and stream formation, as attention needs to be reoriented. Naturally, as the trial unfolds, the speakers' sound features will become more familiar, along with their topics, at least in Study 2. As such, those processes do not start completely anew, especially for what concerns object formation. Nonetheless, our results from both Study 2 and Study 3 show that switching attention between continuous speakers introduces a cognitive cost, signalled by a drop in brain activity power at the alpha band, and by a certain switching time. Furthermore, in Study 2, we found that the temporal dynamics of engaging to a new speaker are faster than those of disengaging, indicating that, in that paradigm, those processes are even transiently simultaneous. Future studies could explore which stages of attention reorienting underpin the phenomena of engagement and disengagement, and the relative contribution of low-level vs. topic information to the formation and tracking of those flexibly attended auditory streams.

Ultimately, the approach we introduced to study attention switching contributes beyond the possibility to observe a fine-grained temporal evolution of attention reorienting. It also offers a uniquely detailed window onto established processes of auditory scene analysis and potentially enables the investigation of questions regarding the features that determine speech stream formation. For example, 1) What is the impact of abstract semantic information on auditory object formation and selection in a multi-talker listening environment? 2) What is the relationship between lower-level acoustic characteristics of the auditory scene (e.g., target vs. masker SNR or pitch difference) and higher-level semantic features (e.g., topic distance) for effective speakers' separation in a dynamic scenario? 3) Is a *Cocktail Party* setting interpretable as a problem of navigation in a multi-dimensional semantic space?

While all those questions may require specific experimental designs to be fully addressed, a naturalistic experiment of attention reorienting already offers the opportunity to investigate the fundamental temporal dimension inherent to any auditory scene.

6.3 Perspectives on ageing and attention reorienting

Current theories agree that declines in hearing and cognition influence the strategies older adults use to understand speech (Herrmann & Butler, 2021; Herrmann & Johnsrude, 2020; Herrmann et al., 2023; Irsik et al., 2022; Peelle & Wingfield, 2016). However, hearing and cognitive factors interact in complex ways, making it difficult to pinpoint exactly the neurophysiological mechanisms giving rise to the observed differences relative to younger listeners. The problem is twofold: First, it is unclear which of the observed speech processing changes are due to systematic shifts in processing strategies, rather than originating from

increased variance within the older population; secondly, even when a difference can be reliably observed, it remains difficult to uncover its causes, and to relate it to differences observed in other domains. For example, as introduced in Chapter 2.5, there is evidence in favour of an increased tendency of older adults to speak in a less semantically coherent way (Hoffman et al., 2018; Sheldon et al., 2024). This has been related to multiple factors, such as their wide lexical vocabulary originating from their extensive experience with language, and the inability to suppress irrelevant semantic information as a consequence of decreased domain-general inhibitory control (Salthouse, 2010; Wu & Hoffman, 2022). Intuitively, similar lexical and semantic selection mechanisms may also apply to speech comprehension, although, in that domain, it is not straightforward to tease them apart from perceptual changes due to hearing levels. Neurophysiological measures of speech comprehension have been derived (lexical surprisal, entropy, dissimilarity), highlighting delayed or weaker neural tracking of semantics, but not consistently across studies. However, there is not a reference framework to connect these findings. That is, the relationship between semantic coherence in spoken language, inhibitory control, delayed or reduced semantic neural tracking is not defined as part of a unitary phenomenon. Furthermore, these effects were observed with monologue listening or sustained attentions scenarios, and it is unclear how they would be impacted by more dynamic listening requirements, which may recruit cognitive resources in different ways.

Our attention switching experiment described in Study 3 offers the opportunity to study age-related differences in the dynamic reorienting of attention. Compared to Study 2, it further introduces a frequent change of topic after the attentions switching cue, making a stream's topic a less reliable cue for its segregation. The stream's topic may be a more important cue for older compared to younger listeners, when having to reorient in the auditory scene, as lower-level sound features may be less accessible to older participants. Based on our finding of reduced neural tracking of acoustic and word onset features of speech in older participants, which also decays sharply over the course of the trial, this indeed seems to be the case. It remains to be tested whether a similar decrease in neural tracking would be observed for higher level lexical features, and particularly whether a model of semantic topic distance could reveal fundamental differences between younger and older listeners. Unfortunately, the current state of the analysis does not allow us to draw solid conclusions on the comparison of switching time between the two populations, nor on the engagement and disengagement stages. Future analyses will determine whether this pattern of results is due to noise uncorrelated to the EEG data, or instead is the consequence of an excessive difficulty of the current experimental paradigm, potentially linked to the frequent topic changes. Yet, it is interesting to observe that the behavioural

performance to comprehension questions remains similar both across age groups and experiments (around 86% correct answers) suggesting that similar behavioural outcomes can be the result of different neurophysiological and cognitive states. As such, it will also be crucial to correlate the metrics of switching time or alpha power to our cognitive test scores, tapping into working memory, processing of information speed, and fluid intelligence.

Overall, these findings invite to consider neural metrics of attention reorienting between competing speakers as a potentially useful indication of speech processing differences between diverse cohorts of listeners. At the same time, this work also offers the opportunity to discuss theories addressing fundamental changes in the neurophysiological underpinnings of speech processing across the life span.

6.4 Ongoing and Future Work

The research presented in this thesis probed the neural encoding of naturalistic multi-talker scenes of sustained and dynamic attention paradigms through non-invasive EEG recordings. Across the three studies here described, I considered different populations in terms of hearing ability and age, offering novel insights into the neurophysiology of attentional filtering at the level of phonological features, and the rapid reorienting of attention, modelled in its stages of engagement and disengagement between competing speakers. Nonetheless, further research is needed to paint a more complete picture of selective attention mechanisms across different populations and across the life span.

One ongoing line of research that I am pursuing is the modelling of semantic information to quantify neural tracking of speech meaning in an attention switching paradigm. Particular interest lies in uncovering potentially different semantic strategies adopted by older adults. For example, I am testing the hypothesis that, due to reduced semantic inhibition, older listeners may rely on a wider pool of potential next-word candidates while engaging in lexical predictions. By raising the LLM's temperature parameter, which controls how deterministic the model's output is, I am extracting lexical surprisal and entropy based on a flatter next-word probability distribution. This would ideally simulate a general inability to suppress irrelevant lexical alternatives, which may approximate the neural mechanisms of reduced semantic inhibition in the ageing brain. I am also considering to model semantics as the *gist of the story*, and to examine whether the neural distance between two speakers may be impacted by their topic/semantic distance. Regarding the switching and alpha power dynamics observed in Study 2 and Study 3, current work is also investigating the robustness of these neural metrics of

attentional reallocation, and their potential to be extended to different populations, e.g. to listeners with hearing impairment.

Another line of research I am currently investigating is the neural encoding of multi-talker scenes in infants (7 and 11 months of age), as they are exposed to spatially separated or segregated competing speakers. Given the extremely young age of the participants, they cannot be explicitly instructed on where to direct their attention. As such, what is assumed to be the target speaker is a stream of infant-directed speech, while the two maskers consist of two adult-directed audiobooks. The results support the idea that across both spatial configurations, infant-directed speech is more tracked by the infants' EEG responses compared to the audiobook streams, indicating that the ability to focus and maintain attention on a stream of interest develops quite early in life. Further analyses will then compare whether the spatial separation of the speakers impacts the neural encoding of the competing streams, which will also give insights into the development of the integration of spatial cues for stream segregation.

This research work has the potential to contribute to our understanding of the neural mechanisms underlying speech processing in complex listening scenarios, which better mirror the real-world perception of language. It has already provided novel insights into such neural dynamics, across diverse attention requirements and populations. Beyond the main research questions motivating these studies, it may additionally prove to be useful to derive objective neural metrics of the listener's attention and attention reorienting. These metrics may then be used, for example, to automatically steer the directionality of a hearing-device, enhancing the voice of the newly attended speaker, or to detect that a disengagement from the attended speaker has occurred, and loosen the noise reduction or directionality parameters of the hearing-device to allow for an easier search of an alternative stream. While those applications are beyond the scope of my work, these studies offer a theoretical framework to investigate their feasibility.

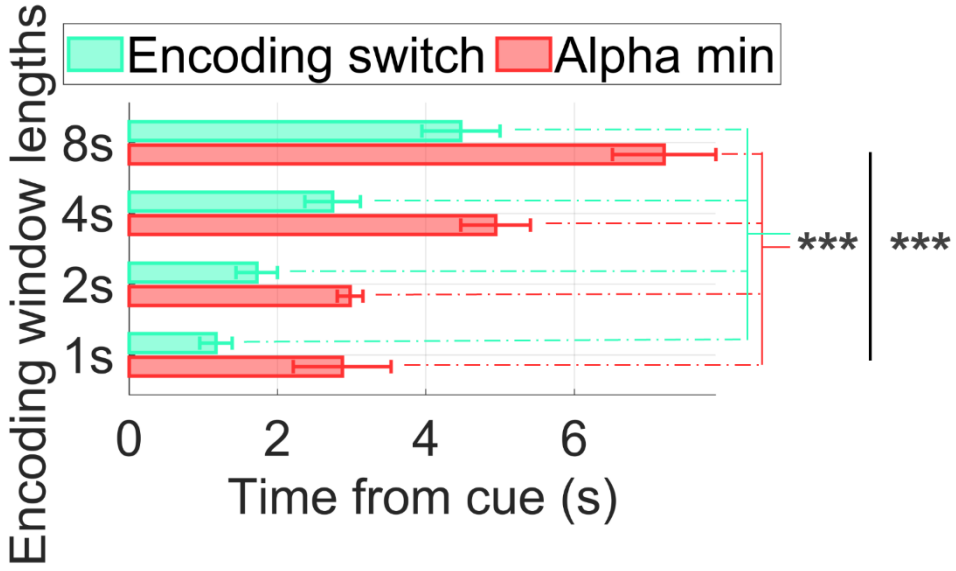
Appendix

Study 1: Investigating phonological processing of competing speakers in hearing impaired listeners

			Hearing thresholds (dB HL)									
Subject	Age (years)	Sex	250	500	750	1000	1500	2000	3000	4000	6000	8000
1	71	M	50	50	60	55	55	55	60	65	75	80
2	73	M	40	45	55	55	55	65	70	65	60	75
3	54	M	60	55	60	70	75	70	65	65	30	10
4	69	M	10	15	15	20	25	40	80	80	75	80
5	59	F	35	40	40	45	35	35	50	55	40	50
6	76	M	25	55	69	55	59	65	65	70	75	70
7	62	M	60	65	65	75	65	75	60	65	65	70
8	68	M	20	10	25	25	40	60	65	65	75	80
9	76	M	30	35	40	40	50	55	65	90	80	95
10	72	M	25	30	40	40	45	65	65	70	65	70
11	76	F	30	35	40	35	40	55	60	65	85	80
12	74	M	10	5	5	10	5	25	45	60	75	80
13	73	M	0	0	10	10	25	45	60	65	80	80
14	80	M	10	30	40	45	55	55	80	80	70	75
15	55	M	25	25	30	40	40	35	35	25	30	20
16	68	M	45	45	50	45	55	60	65	75	60	80
17	78	F	30	40	45	50	50	55	55	50	50	75
18	61	M	35	40	45	40	45	55	45	60	75	90
19	84	M	40	35	40	45	60	55	60	65	65	65
20	81	M	30	35	60	60	60	65	65	75	70	80
21	21	M	35	45	50	60	70	75	75	75	70	60

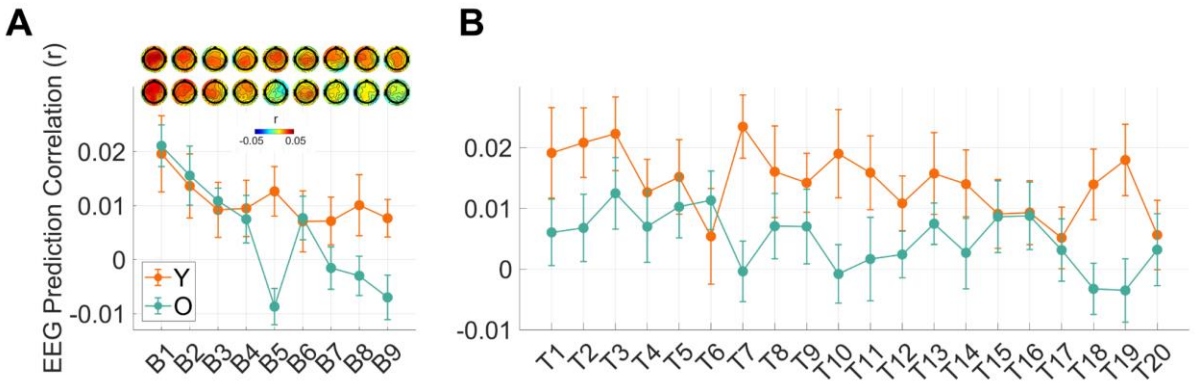
Table 3.1: Participants' information and bilateral air conduction hearing thresholds.

Study 2: Simultaneous cortical tracking of competing speech streams during attention switching



Supplementary Figure 4.2: Time latencies of the encoding switch of EEG prediction correlations 898 (turquoise) and alpha ERSP minimum (red) for four sliding window lengths. Bars represent the average 899 time latencies across participants and electrodes, error bars indicate the SEM across participants. Stars 900 represent statistically significant comparisons (Significance levels: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

Study 3: Investigating the neurophysiology of attention switching in the ageing brain



Supplementary Figure 5.3: Neural encoding of speech features sharply decays with successive attention switches in older but not younger individuals. (A) EEG prediction correlations computed with a separate mTRF fit for each sustained attention block (B1-B9). Error bars indicate the standard error of the mean (SEM) across participants. Scalp topographies show the distribution of prediction correlations over the channels. **(B)** EEG prediction correlations for each trial. The mTRF model including Env, EnvDer and WO features was fit on all sustained attention blocks (B1-B9), and it is here reported for each trial separately. As above, error bars show the SEM across participants.

Bibliography

Agmon, G., Jaeger, M., Magen, E., Pinto, D., Perelmuter, Y., Zion-Golumbic, E., & Bleichner, M. G. (2025). Challenges and Methods in Annotating Natural Speech for Neurolinguistic Research. *Neurobiology of Language*, 1-42. https://doi.org/10.1162/nol.a.12

Agmon, G., Yahav, P. H.-S., Ben-Shachar, M., & Golumbic, E. Z. (2021). Attention to speech: mapping distributed and selective attention systems. *Cerebral Cortex*, *32*(17), 3763-3776. https://doi.org/10.1093/cercor/bhab446

Ahmad, A., Jagdhane, N., Ademmer, K., & Choudhari, K. (2024). Carl Wernicke of the Wernicke Area: A Historical Review. *World Neurosurgery*, *185*, 225-233. https://doi.org/https://doi.org/10.1016/j.wneu.2024.02.103

Aitchison, J. (2003). *Words in the mind: An introduction to the mental lexicon, 3rd ed*. Blackwell Publishing.

Akram, S., Presacco, A., Simon, J. Z., Shamma, S. A., & Babadi, B. (2016). Robust decoding of selective auditory attention from MEG in a competing-speaker environment via state-space modeling. *NeuroImage*, *124*, 906-917. https://doi.org/https://doi.org/10.1016/j.neuroimage.2015.09.048

Alhanbali, S., Dawes, P., Millman, R. E., & Munro, K. J. (2019). Measures of Listening Effort Are Multidimensional. *Ear and Hearing*, *40*(5). https://journals.lww.com/ear-hearing/fulltext/2019/09000/measures_of_listening_effort_are_multidimensional.4.aspx

Alickovic, E., Lunner, T., Gustafsson, F., & Ljung, L. (2019). A Tutorial on Auditory Attention Identification Methods. *Front Neurosci*, *13*, 153. https://doi.org/10.3389/fnins.2019.00153

Alickovic, E., Lunner, T., Wendt, D., Fiedler, L., Hietkamp, R., Ng, E. H. N., & Graversen, C. (2020). Neural Representation Enhanced for Speech and Reduced for Background Noise With a Hearing Aid Noise Reduction Scheme During a Selective Attention Task [Original Research]. *Frontiers in Neuroscience*, *14*. https://doi.org/10.3389/fnins.2020.00846

Alickovic, E., Ng, E. H. N., Fiedler, L., Santurette, S., Innes-Brown, H., & Graversen, C. (2021). Effects of Hearing Aid Noise Reduction on Early and Late Cortical Representations of Competing Talkers in Noise [Original Research]. *Frontiers in Neuroscience*, *15*. https://doi.org/10.3389/fnins.2021.636060

Alink, A., Schwiedrzik, C. M., Kohler, A., Singer, W., & Muckli, L. (2010). Stimulus Predictability Reduces Responses in Primary Visual Cortex. *The Journal of Neuroscience*, *30*(8), 2960-2966. https://doi.org/10.1523/jneurosci.3730-10.2010

Anderson, A. J., Davis, C., & Lalor, E. C. (2024). Deep-learning models reveal how context and listener attention shape electrophysiological correlates of speech-to-language transformation. *PLOS Computational Biology*, *20*(11), e1012537. https://doi.org/10.1371/journal.pcbi.1012537

ANSI, S.-. (1997). Methods for calculation of the speech intelligibility index. *American National Standards Institute, New York*.

Antonello, R., & Huth, A. (2024). Predictive Coding or Just Feature Discovery? An Alternative Account of Why Language Models Fit Brain Data. *Neurobiology of Language*, *5*(1), 64-79. https://doi.org/10.1162/nol_a_00087

Arbogast, T. L., Mason, C. R., & Kidd, G., Jr. (2002). The effect of spatial separation on informational and energetic masking of speech. *J Acoust Soc Am*, *112*(5 Pt 1), 2086-2098. https://doi.org/10.1121/1.1510141

Attaheri, A., Choisdealbha, Á. N., Di Liberto, G. M., Rocha, S., Brusini, P., Mead, N., Olawole-Scott, H., Boutris, P., Gibbon, S., Williams, I., Grey, C., Flanagan, S., & Goswami, U. (2022). Delta- and theta-band cortical tracking and phase-amplitude coupling to sung speech by infants. *NeuroImage*, *247*, 118698. https://doi.org/https://doi.org/10.1016/j.neuroimage.2021.118698

Aydelott, J., Leech, R., & Crinion, J. (2010). Normal adult aging and the contextual influences affecting speech and meaningful sound perception. *Trends Amplif*, *14*(4), 218-232. https://doi.org/10.1177/1084713810393751

Baayen, H. (2007). Storage and computation in the mental lexicon. In (pp. 81–104). https://doi.org/10.1163/9780080548692_006

Bastos, Andre M., Usrey, W. M., Adams, Rick A., Mangun, George R., Fries, P., & Friston, Karl J. (2012). Canonical Microcircuits for Predictive Coding. *Neuron*, *76*(4), 695-711. https://doi.org/https://doi.org/10.1016/j.neuron.2012.10.038

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, *403*(6767), 309-312. https://doi.org/10.1038/35002078

Berger, H. (1929). Über das Elektrenkephalogramm des Menschen. *Archiv für Psychiatrie und Nervenkrankheiten*, *87*(1), 527-570. https://doi.org/10.1007/BF01797193

Billings, C. J., Penman, T. M., McMillan, G. P., & Ellis, E. M. (2015). Electrophysiology and Perception of Speech in Noise in Older Listeners: Effects of Hearing Impairment and Age. *Ear and Hearing*, *36*(6). https://journals.lww.com/ear-hearing/fulltext/2015/11000/electrophysiology_and_perception_of_speech_in.9.aspx

Birbaumer, N., Elbert, T., Canavan, A. G., & Rockstroh, B. (1990). Slow potentials of the cerebral cortex and behavior. *Physiol Rev*, *70*(1), 1-41. https://doi.org/10.1152/physrev.1990.70.1.1

Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P., Madsen, T. O., & Basbøll, H. (2008). Early vocabulary development in Danish and other languages: a CDI-based comparison. *J Child Lang*, *35*(3), 619-650. https://doi.org/10.1017/s0305000908008714

Boersma, P. (2007). Praat : doing phonetics by computer. *http://www.praat.org/*. https://cir.nii.ac.jp/crid/1571417125760875008

Bolt, E., & Giroud, N. (2024). Neural encoding of linguistic speech cues is unaffected by cognitive decline, but decreases with increasing hearing impairment. *Scientific Reports*, *14*(1), 19105. https://doi.org/10.1038/s41598-024-69602-1

Borella, E., Carretti, B., & De Beni, R. (2008). Working memory and inhibition across the adult life-span. *Acta Psychologica*, *128*(1), 33-44. https://doi.org/https://doi.org/10.1016/j.actpsy.2007.09.008

Borges, Heidi B., Zaar, J., Alickovic, E., Christensen, Christian B., & Kidmose, P. (2025). Speech Reception Threshold Estimation via EEG-Based Continuous Speech Envelope Reconstruction. *European Journal of Neuroscience*, *61*(6), e70083. https://doi.org/https://doi.org/10.1111/ejn.70083

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. The MIT Press. https://doi.org/10.7551/mitpress/1486.001.0001

Brennan, J. R., & Hale, J. T. (2019). Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PLOS ONE*, *14*(1), e0207741. https://doi.org/10.1371/journal.pone.0207741

Broadbent, D. E. (1965). Information Processing in the Nervous System. *Science*, *150*(3695), 457-462. https://doi.org/doi:10.1126/science.150.3695.457

Brodbeck, C., Bhattasali, S., Cruz Heredia, A. A. L., Resnik, P., Simon, J. Z., & Lau, E. (2022). Parallel processing in speech perception with local and global representations of linguistic context. *eLife*, *11*, e72056. https://doi.org/10.7554/eLife.72056

Brodbeck, C., Hong, L. E., & Simon, J. Z. (2018). Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech. *Current Biology*, *28*(24), 3976-3983.e3975. https://doi.org/https://doi.org/10.1016/j.cub.2018.10.042

Brodbeck, C., Jiao, A., Hong, L. E., & Simon, J. Z. (2020). Neural speech restoration at the cocktail party: Auditory cortex recovers masked speech of both attended and ignored speakers. *PLOS Biology*, *18*(10), e3000883. https://doi.org/10.1371/journal.pbio.3000883

Brodbeck, C., Presacco, A., Anderson, S., & Simon, J. Z. (2018). Over-representation of speech in older adults originates from early response in higher order auditory cortex. *Acta Acust United Acust*, *104*(5), 774-777. https://doi.org/10.3813/aaa.919221

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech. *Curr Biol*, *28*(5), 803-809.e803. https://doi.org/10.1016/j.cub.2018.01.080

Broderick, M. P., Di Liberto, G. M., Anderson, A. J., Rofes, A., & Lalor, E. C. (2021). Dissociable electrophysiological measures of natural language processing reveal differences in speech comprehension strategy in healthy ageing. *Scientific Reports*, *11*(1), 4963. https://doi.org/10.1038/s41598-021-84597-9

Bronkhorst, A. W. (2015). The cocktail-party problem revisited: early processing and selection of multi-talker speech. *Attention, Perception, & Psychophysics*, *77*(5), 1465-1487. https://doi.org/10.3758/s13414-015-0882-9

Bronkhorst, A. W., & Plomp, R. (1988). The effect of head-induced interaural time and level differences on speech intelligibility in noise. *J Acoust Soc Am*, *83*(4), 1508-1516. https://doi.org/10.1121/1.395906

Brookes, M. J., Leggett, J., Rea, M., Hill, R. M., Holmes, N., Boto, E., & Bowtell, R. (2022). Magnetoencephalography with optically pumped magnetometers (OPM-MEG): the next generation of functional neuroimaging. *Trends in Neurosciences*, *45*(8), 621-634. https://doi.org/10.1016/j.tins.2022.05.008

Burkhardt, P., Müller, V., Meister, H., Weglage, A., Lang-Roth, R., Walger, M., & Sandmann, P. (2022). Age effects on cognitive functions and speech-in-noise processing: An event-related potential study with cochlear-implant users and normal-hearing listeners [Original Research]. *Frontiers in Neuroscience*, *Volume 16 - 2022*. https://doi.org/10.3389/fnins.2022.1005859

Buus, S., & Florentine, M. (2002). Growth of Loudness in Listeners with Cochlear Hearing Losses: Recruitment Reconsidered. *Journal of the Association for Research in Otolaryngology*, *3*(2), 120-139. https://doi.org/10.1007/s101620010084

Cai, J., Hadjinicolaou, A. E., Paulk, A. C., Soper, D. J., Xia, T., Wang, A. F., Rolston, J. D., Richardson, R. M., Williams, Z. M., & Cash, S. S. (2025). Natural language processing models reveal neural dynamics of human conversation. *Nature Communications*, *16*(1), 3376. https://doi.org/10.1038/s41467-025-58620-w

Carta, S., Aličković, E., Zaar, J., Valdés, A. L., & Di Liberto, G. M. (2024). Cortical encoding of phonetic onsets of both attended and ignored speech in hearing impaired individuals. *PLOS ONE*, *19*(11), e0308554. https://doi.org/10.1371/journal.pone.0308554

Caucheteux, C., Gramfort, A., & King, J.-R. (2021). *Disentangling syntax and semantics in the brain with deep networks* Proceedings of the 38th International Conference on Machine Learning, Proceedings of Machine Learning Research. https://proceedings.mlr.press/v139/caucheteux21a.html

Caucheteux, C., Gramfort, A., & King, J.-R. (2022). Deep language algorithms predict semantic comprehension from brain activity. *Scientific Reports*, *12*(1), 16327. https://doi.org/10.1038/s41598-022-20460-9

Caucheteux, C., Gramfort, A., & King, J.-R. (2023). Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-022-01516-2

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*(1), 134. https://doi.org/10.1038/s42003-022-03036-1

Chalas, N., Meyer, L., Lo, C.-W., Park, H., Kluger, D. S., Abbasi, O., Kayser, C., Nitsch, R., & Gross, J. (2024). Dissociating prosodic from syntactic delta activity during natural speech comprehension. *Current Biology*, *34*(15), 3537-3549.e3535. https://doi.org/10.1016/j.cub.2024.06.072

Chalehchaleh, A., Winchester, M. M., & Di Liberto, G. M. (2025). Robust assessment of the cortical encoding of word-level expectations using the temporal response function. *Journal of Neural Engineering*, *22*(1), 016004. https://doi.org/10.1088/1741-2552/ada30a

Cherry, E. C. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America*, *25*(5), 975-979. https://doi.org/10.1121/1.1907229

Chevillet, M., Riesenhuber, M., & Rauschecker, J. P. (2011). Functional Correlates of the Anterolateral Processing Hierarchy in Human Auditory Cortex. *The Journal of Neuroscience*, *31*(25), 9345-9352. https://doi.org/10.1523/jneurosci.1448-11.2011

Choudhari, V., Han, C., Bickel, S., Mehta, A. D., Schevon, C., McKhann, G. M., & Mesgarani, N. (2024). Brain-Controlled Augmented Hearing for Spatially Moving Conversations in Multi-Talker Environments. *Advanced Science*, *11*(41), 2401379. https://doi.org/https://doi.org/10.1002/advs.202401379

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181-204. https://doi.org/10.1017/S0140525X12000477

Cohen, D. (1972). Magnetoencephalography: Detection of the Brain's Electrical Activity with a Superconducting Magnetometer. *Science*, *175*(4022), 664-666. https://doi.org/doi:10.1126/science.175.4022.664

Corsi, P. M., & Michael, P. (1972). Human memory and the medial temporal region of the brain.

Coulson, S., King, J. W., & Kutas, M. (1998). Expect the Unexpected: Event-related Brain Response to Morphosyntactic Violations. *Language and Cognitive Processes*, *13*(1), 21-58. https://doi.org/10.1080/016909698386582

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli [Methods]. *Frontiers in Human Neuroscience*, *10*. https://doi.org/10.3389/fnhum.2016.00604

Crosse, M. J., Zuk, N. J., Di Liberto, G. M., Nidiffer, A. R., Molholm, S., & Lalor, E. C. (2021). Linear Modeling of Neurophysiological Responses to Speech and Other Continuous Stimuli: Methodological Considerations for Applied Research [Review]. *Frontiers in Neuroscience*, *15*. https://doi.org/10.3389/fnins.2021.705621

Cusack, R., Deeks, J., Aikman, G., & Carlyon, R. P. (2004). Effects of location, frequency region, and time course of selective attention on auditory scene analysis. *J Exp Psychol Hum Percept Perform*, *30*(4), 643-656. https://doi.org/10.1037/0096-1523.30.4.643

Dau, T., Maercher Roersted, J., Fuglsang, S., & Hjortkjær, J. (2018). Towards cognitive control of hearing instruments using EEG measures of selective attention. *The Journal of the Acoustical Society of America*, *143*(3_Supplement), 1744-1744. https://doi.org/10.1121/1.5035691

Daube, C., Ince, R. A. A., & Gross, J. (2019). Simple Acoustic Features Can Explain Phoneme-Based Predictions of Cortical Responses to Speech. *Curr Biol*, *29*(12), 1924-1937.e1929. https://doi.org/10.1016/j.cub.2019.04.067

Dave, S., Brothers, T., Hoversten, L. J., Traxler, M. J., & Swaab, T. Y. (2021). Cognitive control mediates age-related changes in flexible anticipatory processing during listening comprehension. *Brain Res*, *1768*, 147573. https://doi.org/10.1016/j.brainres.2021.147573

de Cheveigné, A., Wong, D. D. E., Di Liberto, G. M., Hjortkjær, J., Slaney, M., & Lalor, E. (2018). Decoding the auditory brain with canonical component analysis. *NeuroImage*, *172*, 206-216. https://doi.org/https://doi.org/10.1016/j.neuroimage.2018.01.033

De Clercq, P., Vanthornhout, J., Vandermosten, M., & Francart, T. (2023). Beyond linear neural envelope tracking: a mutual information approach. *Journal of Neural Engineering*, *20*(2), 026007. https://doi.org/10.1088/1741-2552/acbe1d

de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L., & Theunissen, F. E. (2017). The Hierarchical Cortical Organization of Human Speech Processing. *The Journal of Neuroscience*, *37*(27), 6539-6557. https://doi.org/10.1523/jneurosci.3267-16.2017

Decruy, L., Lesenfants, D., Vanthornhout, J., & Francart, T. (2020). Top-down modulation of neural envelope tracking: The interplay with behavioral, self-report and neural measures of listening effort. *Eur J Neurosci*, *52*(5), 3375-3393. https://doi.org/10.1111/ejn.14753

Decruy, L., Vanthornhout, J., & Francart, T. (2019). Evidence for enhanced neural tracking of the speech envelope underlying age-related speech-in-noise difficulties. *Journal of Neurophysiology*, *122*(2), 601-615. https://doi.org/10.1152/jn.00687.2018

Decruy, L., Vanthornhout, J., & Francart, T. (2020). Hearing impairment is associated with enhanced neural tracking of the speech envelope. *Hearing Research*, *393*, 107961. https://doi.org/https://doi.org/10.1016/j.heares.2020.107961

Degano, G., Donhauser, P. W., Gwilliams, L., Merlo, P., & Golestani, N. (2024). Speech prosody enhances the neural processing of syntax. *Communications Biology*, *7*(1), 748. https://doi.org/10.1038/s42003-024-06444-7

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117-1121. https://doi.org/10.1038/nn1504

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J Neurosci Methods*, *134*(1), 9-21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Deutsch, J. A., & Deutsch, D. (1963). Attention: Some theoretical considerations. *Psychological Review*, *70*(1), 80-90. https://doi.org/10.1037/h0039515

DeWitt, I., & Rauschecker, J. P. (2012). Phoneme and word recognition in the auditory ventral stream. *Proceedings of the National Academy of Sciences*, *109*(8), E505-E514. https://doi.org/doi:10.1073/pnas.1113427109

Di Liberto, G. M., Attaheri, A., Cantisani, G., Reilly, R. B., Ní Choisdealbha, Á., Rocha, S., Brusini, P., & Goswami, U. (2023). Emergence of the cortical encoding of phonetic features in the first year of life. *Nature Communications*, *14*(1), 7789. https://doi.org/10.1038/s41467-023-43490-x

Di Liberto, G. M., Nidiffer, A., Crosse, M. J., Zuk, N. J., Haro, S., Cantisani, G., Winchester, M. M., Igoe, A., McCrann, R., Chandra, S., Lalor, E. C., & Baruzzo, G. (2024). A

standardised open science framework for sharing and re-analysing neural data acquired to continuous stimuli. *ArXiv*.

Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. *Curr Biol*, *25*(19), 2457-2465. https://doi.org/10.1016/j.cub.2015.08.030

Di Liberto, G. M. D., Nie, J., Yeaton, J., Khalighinejad, B., Shamma, S. A., & Mesgarani, N. (2021). Neural representation of linguistic feature hierarchy reflects second-language proficiency. *NeuroImage*, *227*, 117586. https://doi.org/https://doi.org/10.1016/j.neuroimage.2020.117586

Dikker, S., & Pylkkänen, L. (2013). Predicting language: MEG evidence for lexical preactivation. *Brain and Language*, *127*(1), 55-64. https://doi.org/https://doi.org/10.1016/j.bandl.2012.08.004

Dimitrijevic, A., Smith, M. L., Kadis, D. S., & Moore, D. R. (2019). Neural indices of listening effort in noisy environments. *Scientific Reports*, *9*(1), 11278. https://doi.org/10.1038/s41598-019-47643-1

Ding, N., & Simon, J. Z. (2012a). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, *109*(29), 11854-11859. https://doi.org/doi:10.1073/pnas.1205381109

Ding, N., & Simon, J. Z. (2012b). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology*, *107*(1), 78-89. https://doi.org/10.1152/jn.00297.2011

Dmochowski, J. P., Bezdek, M. A., Abelson, B. P., Johnson, J. S., Schumacher, E. H., & Parra, L. C. (2014). Audience preferences are predicted by temporal reliability of neural processing. *Nature Communications*, *5*(1), 4567. https://doi.org/10.1038/ncomms5567

Doelling, K. B., Arnal, L. H., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. *NeuroImage*, *85*, 761-768. https://doi.org/https://doi.org/10.1016/j.neuroimage.2013.06.035

Donhauser, P. W., & Baillet, S. (2020). Two Distinct Neural Timescales for Predictive Speech Processing. *Neuron*, *105*(2), 385-393.e389. https://doi.org/10.1016/j.neuron.2019.10.019

Drijvers, L., Small, S. L., & Skipper, J. I. (2025). Language is widely distributed throughout the brain. *Nature Reviews Neuroscience*, *26*(3), 189-189. https://doi.org/10.1038/s41583-024-00903-0

Dronkers, N. F., Plaisant, O., Iba-Zizen, M. T., & Cabanis, E. A. (2007). Paul Broca's historic cases: high resolution MR imaging of the brains of Leborgne and Lelong. *Brain*, *130*(5), 1432-1441. https://doi.org/10.1093/brain/awm042

Durlach, N. I. (1972). Binaural signal detection-Equalization and cancellation theory.

Egner, T. (2023). Principles of cognitive control over task focus and task switching. *Nat Rev Psychol*, *2*(11), 702-714. https://doi.org/10.1038/s44159-023-00234-4

Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., & Shamma, S. A. (2009). Temporal Coherence in the Perceptual Organization and Cortical Representation of Auditory Scenes. *Neuron*, *61*(2), 317-329. https://doi.org/https://doi.org/10.1016/j.neuron.2008.12.005

Evans, S., McGettigan, C., Agnew, Z. K., Rosen, S., & Scott, S. K. (2016). Getting the Cocktail Party Started: Masking Effects in Speech Perception. *J Cogn Neurosci*, *28*(3), 483-500. https://doi.org/10.1162/jocn_a_00913

Federmeier, K. D., McLennan, D. B., de Ochoa, E., & Kutas, M. (2002). The impact of semantic memory organization and sentence context information on spoken language

processing by younger and older adults: An ERP study. *Psychophysiology*, *39*(2), 133-146. https://doi.org/https://doi.org/10.1111/1469-8986.3920133

Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New Method for fMRI Investigations of Language: Defining ROIs Functionally in Individual Subjects. *Journal of Neurophysiology*, *104*(2), 1177-1194. https://doi.org/10.1152/jn.00032.2010

Fedorenko, E., Ivanova, A. A., & Regev, T. I. (2024). The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, *25*(5), 289-312. https://doi.org/10.1038/s41583-024-00802-4

Fedorenko, E., Piantadosi, S. T., & Gibson, E. A. F. (2024). Language is primarily a tool for communication rather than thought. *Nature*, *630*(8017), 575-586. https://doi.org/10.1038/s41586-024-07522-w

Felix, R. A., Gourévitch, B., & Portfors, C. V. (2018). Subcortical pathways: Towards a better understanding of auditory disorders. *Hearing Research*, *362*, 48-60. https://doi.org/https://doi.org/10.1016/j.heares.2018.01.008

Fiedler, L., Johnsrude, I., & Wendt, D. (2025). Salience-Dependent Disruption of Sustained Auditory Attention Can Be Inferred from Evoked Pupil Responses and Neural Tracking of Task-Irrelevant Sounds. *The Journal of Neuroscience*, *45*(14), e2066232025. https://doi.org/10.1523/jneurosci.2066-23.2025

Fiedler, L., Seifi Ala, T., Graversen, C., Alickovic, E., Lunner, T., & Wendt, D. (2021). Hearing Aid Noise Reduction Lowers the Sustained Listening Effort During Continuous Speech in Noise-A Combined Pupillometry and EEG Study. *Ear Hear*, *42*(6), 1590-1601. https://doi.org/10.1097/aud.0000000000001050

Fiedler, L., Wöstmann, M., Herbst, S. K., & Obleser, J. (2019). Late cortical tracking of ignored speech facilitates neural selectivity in acoustically challenging conditions. *NeuroImage*, *186*, 33-42. https://doi.org/https://doi.org/10.1016/j.neuroimage.2018.10.057

Fischer-Jørgensen, E. (1989). Phonetic Analysis of the Stød in Standard Danish. *Phonetica*, *46*(1-3), 1-59. https://doi.org/doi:10.1159/000261828

Folstein, J. R., & Van Petten, C. (2008). Influence of cognitive control and mismatch on the N2 component of the ERP: a review. *Psychophysiology*, *45*(1), 152-170. https://doi.org/10.1111/j.1469-8986.2007.00602.x

Foxe, J. J., & Snyder, A. C. (2011). The Role of Alpha-Band Brain Oscillations as a Sensory Suppression Mechanism during Selective Attention [Review]. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00154

Francis, A. L., & Love, J. (2020). Listening effort: Are we measuring cognition or affect, or both? *Wiley Interdiscip Rev Cogn Sci*, *11*(1), e1514. https://doi.org/10.1002/wcs.1514

French, N. R., & Steinberg, J. C. (1947). Factors Governing the Intelligibility of Speech Sounds. *The Journal of the Acoustical Society of America*, *19*(1), 90-119. https://doi.org/10.1121/1.1916407

Freyman, R. L., Helfer, K. S., McCall, D. D., & Clifton, R. K. (1999). The role of perceived spatial separation in the unmasking of speech. *J Acoust Soc Am*, *106*(6), 3578-3588. https://doi.org/10.1121/1.428211

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815-836. https://doi.org/doi:10.1098/rstb.2005.1622

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293-301. https://doi.org/https://doi.org/10.1016/j.tics.2009.04.005

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philos Trans R Soc Lond B Biol Sci*, *364*(1521), 1211-1221. https://doi.org/10.1098/rstb.2008.0300

Fuglsang, S. A., Dau, T., & Hjortkjær, J. (2017). Noise-robust cortical tracking of attended speech in real-world acoustic scenes. *NeuroImage*, *156*, 435-444. https://doi.org/https://doi.org/10.1016/j.neuroimage.2017.04.026

Fuglsang, S. A., Märcher-Rørsted, J., Dau, T., & Hjortkjær, J. (2020). Effects of Sensorineural Hearing Loss on Cortical Synchronization to Competing Speech during Selective Attention. *J Neurosci*, *40*(12), 2562-2572. https://doi.org/10.1523/jneurosci.1936-19.2020

Gabhart, K. M., Xiong, Y., & Bastos, A. M. (2025). Predictive coding: a more cognitive process than we thought? *Trends in Cognitive Sciences*, *29*(7), 627-640. https://doi.org/https://doi.org/10.1016/j.tics.2025.01.012

Gagné, J. P., Besser, J., & Lemke, U. (2017). Behavioral Assessment of Listening Effort Using a Dual-Task Paradigm. *Trends Hear*, *21*, 2331216516687287. https://doi.org/10.1177/2331216516687287

Geirnaert, S., Vandecappelle, S., Alickovic, E., Cheveigne, A. d., Lalor, E., Meyer, B. T., Miran, S., Francart, T., & Bertrand, A. (2021). Electroencephalography-Based Auditory Attention Decoding: Toward Neurosteered Hearing Devices. *IEEE Signal Processing Magazine*, *38*(4), 89-102. https://doi.org/10.1109/MSP.2021.3075932

Getzmann, S., & Falkenstein, M. (2011). Understanding of spoken language under challenging listening conditions in younger and older listeners: A combined behavioral and electrophysiological study. *Brain Research*, *1415*, 8-22. https://doi.org/https://doi.org/10.1016/j.brainres.2011.08.001

Getzmann, S., Golob, E. J., & Wascher, E. (2016). Focused and divided attention in a simulated cocktail-party situation: ERP evidence from younger and older adults. *Neurobiology of Aging*, *41*, 138-149. https://doi.org/https://doi.org/10.1016/j.neurobiolaging.2016.02.018

Getzmann, S., Klatt, L.-I., Schneider, D., Begau, A., & Wascher, E. (2020). EEG correlates of spatial shifts of attention in a dynamic multi-talker speech perception scenario in younger and older adults. *Hearing Research*, *398*, 108077. https://doi.org/https://doi.org/10.1016/j.heares.2020.108077

Gillis, M., Decruy, L., Vanthornhout, J., & Francart, T. (2022). Hearing loss is associated with delayed neural responses to continuous speech. *European Journal of Neuroscience*, *55*(6), 1671-1690. https://doi.org/https://doi.org/10.1111/ejn.15644

Gillis, M., Kries, J., Vandermosten, M., & Francart, T. (2023). Neural tracking of linguistic and acoustic speech representations decreases with advancing age. *NeuroImage*, *267*, 119841. https://doi.org/https://doi.org/10.1016/j.neuroimage.2022.119841

Gillis, M., Vanthornhout, J., Simon, J. Z., Francart, T., & Brodbeck, C. (2021). Neural Markers of Speech Comprehension: Measuring EEG Tracking of Linguistic Speech Representations, Controlling the Speech Acoustics. *The Journal of Neuroscience*, *41*(50), 10316. https://doi.org/10.1523/JNEUROSCI.0812-21.2021

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511-517. https://doi.org/10.1038/nn.3063

Goldstein, A., Wang, H., Niekerken, L., Schain, M., Zada, Z., Aubrey, B., Sheffer, T., Nastase, S. A., Gazula, H., Singh, A., Rao, A., Choe, G., Kim, C., Doyle, W., Friedman, D., Devore, S., Dugan, P., Hassidim, A., Brenner, M., . . . Hasson, U. (2025). A unified acoustic-to-speech-to-language embedding space captures the neural basis of natural language processing in everyday conversations. *Nature Human Behaviour*, *9*(5), 1041-1055. https://doi.org/10.1038/s41562-025-02105-9

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto,

C., Fanda, L., Doyle, W., Friedman, D., . . . Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369-380. https://doi.org/10.1038/s41593-022-01026-4

Gooskens, C., van Heuven, V. J., van Bezooijen, R., & Pacilly, J. J. A. (2010). Is spoken Danish less intelligible than Swedish? *Speech Communication*, *52*(11), 1022-1037. https://doi.org/https://doi.org/10.1016/j.specom.2010.06.005

Goossens, T., Vercammen, C., Wouters, J., & van Wieringen, A. (2018). Neural envelope encoding predicts speech perception performance for normal-hearing and hearing-impaired adults. *Hearing Research*, *370*, 189-200. https://doi.org/https://doi.org/10.1016/j.heares.2018.07.012

Grace, J. A., Amin, N., Singh, N. C., & Theunissen, F. E. (2003). Selectivity for Conspecific Song in the Zebra Finch Auditory Forebrain. *Journal of Neurophysiology*, *89*(1), 472-487. https://doi.org/10.1152/jn.00088.2002

Greenwood, D. D. (1961). Auditory Masking and the Critical Band. *The Journal of the Acoustical Society of America*, *33*(4), 484-502. https://doi.org/10.1121/1.1908699

Grønnum, N. (2003). Why are the Danes so hard to understand? In H. G. Jacobsen, D. Bleses, T. O. Madsen, & P. Thomsen (Eds.), (Vol. Take Danish – For instance, pp. 119-130). University Press of Southern Denmark.

Guenther, F. H., & Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *J Acoust Soc Am*, *100*(2 Pt 1), 1111-1121. https://doi.org/10.1121/1.416296

Gwilliams, L., Bhaya-Grossman, I., Zhang, Y., Scott, T., Harper, S., & Levy, D. (2025). Computational Architecture of Speech Comprehension in the Human Brain. *Annual Review of Linguistics*, *11*(Volume 11, 2025), 209-226. https://doi.org/https://doi.org/10.1146/annurev-linguistics-031120-111245

Gwilliams, L., King, J.-R., Marantz, A., & Poeppel, D. (2022). Neural dynamics of phoneme sequences reveal position-invariant code for content and order. *Nature Communications*, *13*(1), 6606. https://doi.org/10.1038/s41467-022-34326-1

Gwilliams, L., Marantz, A., Poeppel, D., & King, J.-R. (2023). Top-down information shapes lexical processing when listening to continuous speech. *Language, Cognition and Neuroscience*, 1-14. https://doi.org/10.1080/23273798.2023.2171072

Hagoort, P. (2019). The neurobiology of language beyond single-word processing. *Science*, *366*(6461), 55-58. https://doi.org/doi:10.1126/science.aax0289

Hamilton, L. S., Edwards, E., & Chang, E. F. (2018). A Spatial Map of Onset and Sustained Responses to Speech in the Human Superior Temporal Gyrus. *Curr Biol*, *28*(12), 1860-1871.e1864. https://doi.org/10.1016/j.cub.2018.04.033

Hamilton, L. S., Oganian, Y., Hall, J., & Chang, E. F. (2021). Parallel and distributed encoding of speech across human auditory cortex. *Cell*, *184*(18), 4626-4639.e4613. https://doi.org/https://doi.org/10.1016/j.cell.2021.07.019

Hansen, J. C., & Hillyard, S. A. (1980). Endogenous brain potentials associated with selective auditory attention. *Electroencephalogr Clin Neurophysiol*, *49*(3-4), 277-290. https://doi.org/10.1016/0013-4694(80)90222-9

Har-shai Yahav, P., & Zion Golumbic, E. (2021). Linguistic processing of task-irrelevant speech at a cocktail party. *eLife*, *10*, e65096. https://doi.org/10.7554/eLife.65096

Haro, S., Rao, H. M., Quatieri, T. F., & Smalt, C. J. (2022). EEG alpha and pupil diameter reflect endogenous auditory attention switching and listening effort. *European Journal of Neuroscience*, *55*(5), 1262-1277. https://doi.org/https://doi.org/10.1111/ejn.15616

Harper, J., Malone, S. M., & Bernat, E. M. (2014). Theta and delta band activity explain N2 and P3 ERP component activity in a go/no-go task. *Clinical Neurophysiology*, *125*(1), 124-132. https://doi.org/https://doi.org/10.1016/j.clinph.2013.06.025

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject Synchronization of Cortical Activity During Natural Vision. *Science*, *303*(5664), 1634-1640. https://doi.org/doi:10.1126/science.1089506

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The Faculty of Language: What Is It, Who Has It, and How Did It Evolve? *Science*, *298*(5598), 1569-1579. https://doi.org/doi:10.1126/science.298.5598.1569

Hausfeld, L., Hamers, I. M. H., & Formisano, E. (2024). FMRI speech tracking in primary and non-primary auditory cortex while listening to noisy scenes. *Communications Biology*, *7*(1), 1217. https://doi.org/10.1038/s42003-024-06913-z

Hausfeld, L., Riecke, L., Valente, G., & Formisano, E. (2018). Cortical tracking of multiple streams outside the focus of attention in naturalistic auditory scenes. *NeuroImage*, *181*, 617-626. https://doi.org/https://doi.org/10.1016/j.neuroimage.2018.07.052

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & de Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, *119*(32), e2201968119. https://doi.org/doi:10.1073/pnas.2201968119

Heilbron, M., & Chait, M. (2018). Great Expectations: Is there Evidence for Predictive Coding in Auditory Cortex? *Neuroscience*, *389*, 54-73. https://doi.org/https://doi.org/10.1016/j.neuroscience.2017.07.061

Herff, C., Krusienski, D. J., & Kubben, P. (2020). The Potential of Stereotactic-EEG for Brain-Computer Interfaces: Current Progress and Future Directions [Mini Review]. *Frontiers in Neuroscience*, *Volume 14 - 2020*. https://doi.org/10.3389/fnins.2020.00123

Herrmann, B. (2025). Age-related differences in the impact of background noise on neural speech tracking. *Neurobiology of Aging*, *153*, 10-20. https://doi.org/https://doi.org/10.1016/j.neurobiolaging.2025.06.001

Herrmann, B., & Butler, B. E. (2021). Hearing loss and brain plasticity: the hyperactivity phenomenon. *Brain Structure and Function*, *226*(7), 2019-2039. https://doi.org/10.1007/s00429-021-02313-9

Herrmann, B., & Johnsrude, I. S. (2020). A model of listening engagement (MoLE). *Hearing Research*, *397*, 108016. https://doi.org/https://doi.org/10.1016/j.heares.2020.108016

Herrmann, B., Maess, B., Henry, M. J., Obleser, J., & Johnsrude, I. S. (2023). Neural signatures of task-related fluctuations in auditory attention and age-related changes. *NeuroImage*, *268*, 119883. https://doi.org/https://doi.org/10.1016/j.neuroimage.2023.119883

Hesselmann, G., Sadaghiani, S., Friston, K. J., & Kleinschmidt, A. (2010). Predictive Coding or Evidence Accumulation? False Inference and Neuronal Fluctuations. *PLOS ONE*, *5*(3), e9926. https://doi.org/10.1371/journal.pone.0009926

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393-402. https://doi.org/10.1038/nrn2113

Hillyard, S. A., & Picton, T. W. (1978). ON and OFF components in the auditory evoked potential. *Perception & Psychophysics*, *24*(5), 391-398. https://doi.org/10.3758/BF03199736

Hink, R. F., & Hillyard, S. A. (1976). Auditory evoked potentials during selective listening to dichotic speech messages. *Perception & Psychophysics*, *20*(4), 236-242. https://doi.org/10.3758/BF03199449

Hjortkjær, J., Wong, D. D. E., Catania, A., Märcher-Rørsted, J., Ceolini, E., Fuglsang, S. A., Kiselev, I., Di Liberto, G., Liu, S.-C., Dau, T., Slaney, M., & de Cheveigné, A. (2025). Real-time control of a hearing instrument with EEG-based attention decoding. *Journal of Neural Engineering*, *22*(1), 016027. https://doi.org/10.1088/1741-2552/ad867c

Hoffman, P. (2019). Divergent effects of healthy ageing on semantic knowledge and control: Evidence from novel comparisons with semantically impaired patients. *J Neuropsychol*, *13*(3), 462-484. https://doi.org/10.1111/jnp.12159

Hoffman, P., Loginova, E., & Russell, A. (2018). Poor coherence in older people's speech is explained by impaired semantic and executive processes. *eLife*, *7*. https://doi.org/10.7554/eLife.38907

Howard, M. F., & Poeppel, D. (2010). Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. *J Neurophysiol*, *104*(5), 2500-2511. https://doi.org/10.1152/jn.00251.2010

Huang, P.-C., Schils, L. A. P., Koch, I., Stephan, D. N., & Hsieh, S. (2025). Age-related Electrophysical Correlates of Cross-modal Attention Switching. *Journal of Cognitive Neuroscience*, *37*(1), 43-62. https://doi.org/10.1162/jocn_a_02248

Huang, S., Belliveau, J. W., Tengshe, C., & Ahveninen, J. (2012). Brain Networks of Novelty-Driven Involuntary and Cued Voluntary Auditory Attention Shifting. *PLOS ONE*, *7*(8), e44062. https://doi.org/10.1371/journal.pone.0044062

Hyde, M. (2009). The N1 Response and Its Applications. *Audiology and Neurotology*, *2*(5), 281-307. https://doi.org/10.1159/000259253

Irani, F., Platek, S. M., Bunce, S., Ruocco, A. C., & Chute, D. (2007). Functional near infrared spectroscopy (fNIRS): an emerging neuroimaging technology with important applications for the study of brain disorders. *Clin Neuropsychol*, *21*(1), 9-37. https://doi.org/10.1080/13854040600910018

Irsik, V. C., Johnsrude, I. S., & Herrmann, B. (2022). Age-related deficits in dip-listening evident for isolated sentences but not for spoken stories. *Scientific Reports*, *12*(1), 5898. https://doi.org/10.1038/s41598-022-09805-6

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., & El Sayed, W. (2023). Mistral 7B. In: arXiv.

Jin, I. K., Lee, D., Jeong, Y., Seo, Y. J., Kong, T. H., Suh, M. J., Cho, W. H., Lee, H. J., Choi, S. J., Cha, D., Park, K. H., & Oh, S. H. (2024). Trends in Distributions of Hearing Threshold Levels by Ages: A Comparison of the ISO 7029 and Newly Available Country-Specific Data. *J Audiol Otol*, *28*(1), 1-9. https://doi.org/10.7874/jao.2023.00626

Jongsma, M. L. A., van Rijn, C. M., Gerrits, N. J. H. M., Eichele, T., Steenbergen, B., Maes, J. H. R., & Quian Quiroga, R. (2013). The learning-oddball paradigm: Data of 24 separate individuals illustrate its potential usefulness as a new clinical tool. *Clinical Neurophysiology*, *124*(3), 514-521. https://doi.org/https://doi.org/10.1016/j.clinph.2012.09.009

Joyal, M., Groleau, C., Bouchard, C., Wilson, M. A., & Fecteau, S. (2020). Semantic Processing in Healthy Aging and Alzheimer's Disease: A Systematic Review of the N400 Differences. *Brain Sci*, *10*(11). https://doi.org/10.3390/brainsci10110770

Jun, J. J., Steinmetz, N. A., Siegle, J. H., Denman, D. J., Bauza, M., Barbarits, B., Lee, A. K., Anastassiou, C. A., Andrei, A., Aydın, Ç., Barbic, M., Blanche, T. J., Bonin, V., Couto, J., Dutta, B., Gratiy, S. L., Gutnisky, D. A., Häusser, M., Karsh, B., . . . Harris, T. D. (2017). Fully integrated silicon probes for high-density recording of neural activity. *Nature*, *551*(7679), 232-236. https://doi.org/10.1038/nature24636

Kahneman, D. (1973). *Attention and Effort*. Prentice-Hall.

Kahneman, D., & Beatty, J. (1966). Pupil Diameter and Load on Memory. *Science*, *154*(3756), 1583-1585. https://doi.org/doi:10.1126/science.154.3756.1583

Karunathilake, I. M. D., Brodbeck, C., Bhattasali, S., Resnik, P., & Simon, J. Z. (2025). Neural Dynamics of the Processing of Speech Features: Evidence for a Progression of Features from Acoustic to Sentential Processing. *The Journal of Neuroscience*, e1143242025. https://doi.org/10.1523/jneurosci.1143-24.2025

Karunathilake, I. M. D., Dunlap, J. L., Perera, J., Presacco, A., Decruy, L., Anderson, S., Kuchinsky, S. E., & Simon, J. Z. (2023). Effects of aging on cortical representations of continuous speech. *Journal of Neurophysiology*, *129*(6), 1359-1377. https://doi.org/10.1152/jn.00356.2022

Kathleen Pichora-Fuller, M. (2008). Use of supportive context by younger and older adult listeners: balancing bottom-up and top-down information processing. *Int J Audiol*, *47 Suppl 2*, S72-82. https://doi.org/10.1080/14992020802307404

Katthi, J. R., Ganapathy, S., Kothinti, S., & Slaney, M. (2020). Deep Canonical Correlation Analysis For Decoding The Auditory Brain. *Annu Int Conf IEEE Eng Med Biol Soc*, *2020*, 3505-3508. https://doi.org/10.1109/embc44109.2020.9176208

Kaufman, M., & Zion Golumbic, E. (2023). Listening to two speakers: Capacity and tradeoffs in neural speech tracking during Selective and Distributed Attention. *NeuroImage*, *270*, 119984. https://doi.org/https://doi.org/10.1016/j.neuroimage.2023.119984

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, *98*(3), 630-644.e616. https://doi.org/10.1016/j.neuron.2018.03.044

Keller, G. B., & Mrsic-Flogel, T. D. (2018). Predictive Processing: A Canonical Cortical Computation. *Neuron*, *100*(2), 424-435. https://doi.org/10.1016/j.neuron.2018.10.003

Khalighinejad, B., Patel, P., Herrero, J. L., Bickel, S., Mehta, A. D., & Mesgarani, N. (2021). Functional characterization of human Heschl's gyrus in response to natural speech. *NeuroImage*, *235*, 118003. https://doi.org/10.1016/j.neuroimage.2021.118003

King, A., Teki, S., & Willmore, B. (2018). Recent advances in understanding the auditory cortex [version 1; peer review: 2 approved]. *F1000Research*, *7*(1555). https://doi.org/10.12688/f1000research.15580.1

King, J.-R., Charton, F., Lopez-Paz, D., & Oquab, M. (2020). Back-to-back regression: Disentangling the influence of correlated factors from multivariate observations. *NeuroImage*, *220*, 117028. https://doi.org/https://doi.org/10.1016/j.neuroimage.2020.117028

Kisler, T., Reichel, U., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, *45*, 326-347. https://doi.org/https://doi.org/10.1016/j.csl.2017.01.005

Kjems, U., & Jensen, J. (2012, 27-31 Aug. 2012). Maximum likelihood based noise covariance matrix estimation for multi-microphone speech enhancement. 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO),

Klem, G. H., Lüders, H. O., Jasper, H. H., & Elger, C. (1999). The ten-twenty electrode system of the International Federation. The International Federation of Clinical Neurophysiology. *Electroencephalogr Clin Neurophysiol Suppl*, *52*, 3-6.

Koerner, T. K., & Zhang, Y. (2015). Effects of background noise on inter-trial phase coherence and auditory N1–P2 responses to speech stimuli. *Hearing Research*, *328*, 113-119. https://doi.org/https://doi.org/10.1016/j.heares.2015.08.002

Koning, R., & Wouters, J. (2012). The potential of onset enhancement for increased speech intelligibility in auditory prostheses. *J Acoust Soc Am*, *132*(4), 2569-2581. https://doi.org/10.1121/1.4748965

Kopp, B., Steinke, A., & Visalli, A. (2020). Cognitive flexibility and N2/P3 event-related brain potentials. *Scientific Reports*, *10*(1), 9859. https://doi.org/10.1038/s41598-020-66781-5

Kryter, K. D. (1962). Methods for the calculation and use of the articulation index. *The Journal of the Acoustical Society of America*, *34*(11), 1689-1697.

Kubanek, J., Brunner, P., Gunduz, A., Poeppel, D., & Schalk, G. (2013). The Tracking of Speech Envelope in the Human Cortex. *PLOS ONE*, *8*(1), e53398. https://doi.org/10.1371/journal.pone.0053398

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annu Rev Psychol*, *62*, 621-647. https://doi.org/10.1146/annurev.psych.093008.131123

Kutas, M., & Hillyard, S. A. (1980). Reading Senseless Sentences: Brain Potentials Reflect Semantic Incongruity. *Science*, *207*(4427), 203-205. https://doi.org/doi:10.1126/science.7350657

Kutas, M., & Hillyard, S. A. (1983). Event-related brain potentials to grammatical errors and semantic anomalies. *Memory & Cognition*, *11*(5), 539-550. https://doi.org/10.3758/BF03196991

Lachaux, J.-P., Rodriguez, E., Martinerie, J., & Varela, F. J. (1999). Measuring phase synchrony in brain signals. *Human Brain Mapping*, *8*(4), 194-208. https://doi.org/https://doi.org/10.1002/(SICI)1097-0193(1999)8:4<194::AID-HBM4>3.0.CO;2-C

Larson, E., & Lee, A. K. (2013). The cortical dynamics underlying effective switching of auditory spatial attention. *NeuroImage*, *64*, 365-370. https://doi.org/10.1016/j.neuroimage.2012.09.006

Lauteslager, T., Sullivan, J. A. O., Reilly, R. B., & Lalor, E. C. (2014, 26-30 Aug. 2014). Decoding of attentional selection in a cocktail party environment from single-trial EEG is robust to task. 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society,

Lawo, V., Fels, J., Oberem, J., & Koch, I. (2014). Intentional attention switching in dichotic listening: Exploring the efficiency of nonspatial and spatial selection. *Quarterly Journal of Experimental Psychology*, *67*(10), 2010-2024. https://doi.org/10.1080/17470218.2014.898079

LCM team, B., L., Duquenne, P.-A., Elbayad, M., Kozhevnikov, A., Alastruey, B., Andrews, P., Coria, M., Couairon, G., Costa-jussà, M. R., Dale, D., Elsahar, H., Heffernan, K., Janeiro, J. M., Tran, T., Ropers, C., Sánchez, E., San Roman, R., Mourachko, A., Saleem, S., & Schwenk, H. (2024). Large concept models: Language modeling in a sentence representation space. In: arXiv.

Lecaignard, F., Bertrand, O., Gimenez, G., Mattout, J., & Caclin, A. (2015). Implicit learning of predictable sound sequences modulates human brain responses at different levels of the auditory hierarchy [Original Research]. *Frontiers in Human Neuroscience*, *Volume 9 - 2015*. https://doi.org/10.3389/fnhum.2015.00505

Lee, A. K. C., Larson, E., Maddox, R. K., & Shinn-Cunningham, B. G. (2014). Using neuroimaging to understand the cortical mechanisms of auditory selective attention. *Hearing Research*, *307*, 111-120. https://doi.org/https://doi.org/10.1016/j.heares.2013.06.010

Leonard, M. K., Baud, M. O., Sjerps, M. J., & Chang, E. F. (2016). Perceptual restoration of masked speech in human cortex. *Nature Communications*, *7*(1), 13619. https://doi.org/10.1038/ncomms13619

Leonard, M. K., Gwilliams, L., Sellers, K. K., Chung, J. E., Xu, D., Mischler, G., Mesgarani, N., Welkenhuysen, M., Dutta, B., & Chang, E. F. (2024). Large-scale single-neuron

speech sound encoding across the depth of human cortex. *Nature*, *626*(7999), 593-602. https://doi.org/10.1038/s41586-023-06839-2

Lewald, J., & Getzmann, S. (2015). Electrophysiological correlates of cocktail-party listening. *Behavioural Brain Research*, *292*, 157-166. https://doi.org/https://doi.org/10.1016/j.bbr.2015.06.025

Liberman, A. M. (1957). Some Results of Research on Speech Perception. *The Journal of the Acoustical Society of America*, *29*(1), 117-123. https://doi.org/10.1121/1.1908635

Liberto, G. M. D., Attaheri, A., Cantisani, G., Reilly, R. B., Choisdealbha, Á. N., Rocha, S., Brusini, P., & Goswami, U. (2022). Emergence of the cortical encoding of phonetic features in the first year of life. *bioRxiv*, 2022.2010.2011.511716. https://doi.org/10.1101/2022.10.11.511716

Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., & Medler, D. A. (2005). Neural Substrates of Phonemic Perception. *Cerebral Cortex*, *15*(10), 1621-1631. https://doi.org/10.1093/cercor/bhi040

Liuzzi, A. G., Aglinskas, A., & Fairhall, S. L. (2020). General and feature-based semantic representations in the semantic network. *Scientific Reports*, *10*(1), 8931. https://doi.org/10.1038/s41598-020-65906-0

Livingston, G., Huntley, J., Sommerlad, A., Ames, D., Ballard, C., Banerjee, S., Brayne, C., Burns, A., Cohen-Mansfield, J., Cooper, C., Costafreda, S. G., Dias, A., Fox, N., Gitlin, L. N., Howard, R., Kales, H. C., Kivimäki, M., Larson, E. B., Ogunniyi, A., . . . Mukadam, N. (2020). Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *The Lancet*, *396*(10248), 413-446. https://doi.org/10.1016/S0140-6736(20)30367-6

Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature*, *453*(7197), 869-878. https://doi.org/10.1038/nature06976

Lorch, M. P. (2008). The merest Logomachy: The 1868 Norwich discussion of aphasia by Hughlings Jackson and Broca. *Brain*, *131*(6), 1658-1670. https://doi.org/10.1093/brain/awn058

Loughrey, D. G., Kelly, M. E., Kelley, G. A., Brennan, S., & Lawlor, B. A. (2018). Association of Age-Related Hearing Loss With Cognitive Function, Cognitive Impairment, and Dementia: A Systematic Review and Meta-analysis. *JAMA Otolaryngology–Head & Neck Surgery*, *144*(2), 115-126. https://doi.org/10.1001/jamaoto.2017.2513

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, J., Ly, A., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Wild, A., Knight, P., Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2019). JASP: Graphical Statistical Software for Common Statistical Designs. *Journal of Statistical Software*, *88*(2), 1 - 17. https://doi.org/10.18637/jss.v088.i02

Lunner, T. (2003). Cognitive function in relation to hearing aid use. *Int J Audiol*, *42 Suppl 1*, S49-58. https://doi.org/10.3109/14992020309074624

Lunner, T., Alickovic, E., Graversen, C., Ng, E. H. N., Wendt, D., & Keidser, G. (2020). Three New Outcome Measures That Tap Into Cognitive Processes Required for Real-Life Communication. *Ear Hear*, *41 Suppl 1*(Suppl 1), 39s-47s. https://doi.org/10.1097/aud.0000000000000941

Luo, H., & Poeppel, D. (2007). Phase Patterns of Neuronal Responses Reliably Discriminate Speech in Human Auditory Cortex. *Neuron*, *54*(6), 1001-1010. https://doi.org/10.1016/j.neuron.2007.06.004

Luo, T., Xu, M., Zheng, Z., & Okazawa, G. (2025). Limitation of switching sensory information flow in flexible perceptual decision making. *Nature Communications*, *16*(1), 172. https://doi.org/10.1038/s41467-024-55686-w

Madsen, J., & Parra, L. C. (2022). Cognitive processing of a common stimulus synchronizes brains, hearts, and eyes. *PNAS Nexus*, *1*(1). https://doi.org/10.1093/pnasnexus/pgac020

Mai, A., Riès, S., Ben-Haim, S., Shih, J. J., & Gentner, T. Q. (2024). Acoustic and language-specific sources for phonemic abstraction from speech. *Nature Communications*, *15*(1), 677. https://doi.org/10.1038/s41467-024-44844-9

Makov, S., Pinto, D., Har-shai Yahav, P., Miller, L. M., & Zion Golumbic, E. (2023). "Unattended, distracting or irrelevant": Theoretical implications of terminological choices in auditory selective attention research. *Cognition*, *231*, 105313. https://doi.org/https://doi.org/10.1016/j.cognition.2022.105313

McCloy, D. R., Lau, B. K., Larson, E., Pratt, K. A. I., & Lee, A. K. C. (2017). Pupillometry shows the effort of auditory attention switchinga). *The Journal of the Acoustical Society of America*, *141*(4), 2440-2451. https://doi.org/10.1121/1.4979340

McCullagh, J., & Shinn, J. (2013). Auditory cortical processing in noise in younger and older adults. *Hearing, Balance and Communication*, *11*. https://doi.org/10.3109/21695717.2013.855374

Meister, H., Wenzel, F., Gehlen, A. K., Kessler, J., & Walger, M. (2020). Static and dynamic cocktail party listening in younger and older adults. *Hearing Research*, *395*, 108020. https://doi.org/https://doi.org/10.1016/j.heares.2020.108020

Menn, K. H., Männel, C., & Meyer, L. (2023). Phonological acquisition depends on the timing of speech sounds: Deconvolution EEG modeling across the first five years. *Science Advances*, *9*(44), eadh2560. https://doi.org/doi:10.1126/sciadv.adh2560

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*(7397), 233-236. https://doi.org/10.1038/nature11020

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, *343*(6174), 1006-1010. https://doi.org/10.1126/science.1245994

Mesik, J., Ray, L., & Wojtczak, M. (2021). Effects of Age on Cortical Tracking of Word-Level Features of Continuous Competing Speech [Original Research]. *Frontiers in Neuroscience*, *15*. https://doi.org/10.3389/fnins.2021.635126

Meyer, L., Sun, Y., & Martin, A. E. (2020). Synchronous, but not entrained: exogenous and endogenous cortical rhythms of speech and language processing. *Language, Cognition and Neuroscience*, *35*(9), 1089-1099. https://doi.org/10.1080/23273798.2019.1693050

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In: arXiV.

Mikulasch, F. A., Rudelt, L., Wibral, M., & Priesemann, V. (2023). Where is the error? Hierarchical predictive coding through dendritic error computation. *Trends in Neurosciences*, *46*(1), 45-59. https://doi.org/10.1016/j.tins.2022.09.007

Millman, R. E., Mattys, S. L., Gouws, A. D., & Prendergast, G. (2017). Magnified Neural Envelope Coding Predicts Deficits in Speech Perception in Noise. *J Neurosci*, *37*(32), 7727-7736. https://doi.org/10.1523/jneurosci.2722-16.2017

Miran, S., Akram, S., Sheikhattar, A., Simon, J. Z., Zhang, T., & Babadi, B. (2018). Real-Time Tracking of Selective Auditory Attention From M/EEG: A Bayesian Filtering Approach [Original Research]. *Frontiers in Neuroscience*, *12*. https://doi.org/10.3389/fnins.2018.00262

Mischler, G., Li, Y. A., Bickel, S., Mehta, A. D., & Mesgarani, N. (2024). Contextual feature extraction hierarchies converge in large language models and the brain. *Nature Machine Intelligence*, *6*(12), 1467-1477. https://doi.org/10.1038/s42256-024-00925-4

Mitchell, J., & Lapata, M. (2010). Composition in Distributional Models of Semantics. *Cognitive Science*, *34*(8), 1388-1429. https://doi.org/https://doi.org/10.1111/j.1551-6709.2010.01106.x

Mondor, T. A., & Zatorre, R. J. (1995). Shifting and focusing auditory spatial attention. *J Exp Psychol Hum Percept Perform*, *21*(2), 387-409. https://doi.org/10.1037//0096-1523.21.2.387

Moro, A., Greco, M., & Cappa, S. F. (2023). Large languages, impossible languages and human brains. *Cortex*, *167*, 82-85. https://doi.org/https://doi.org/10.1016/j.cortex.2023.07.003

Mortensen, L., Meyer, A. S., & Humphreys, G. W. (2006). Age-related effects on speech production: A review. *Language and Cognitive Processes*, *21*(1-3), 238-290. https://doi.org/10.1080/01690960444000278

Murphy, D. R., Daneman, M., & Schneider, B. A. (2006). Why do older adults have difficulty following conversations? *Psychol Aging*, *21*(1), 49-61. https://doi.org/10.1037/0882-7974.21.1.49

Näätänen, R., Gaillard, A. W. K., & Mäntysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. *Acta Psychologica*, *42*(4), 313-329. https://doi.org/https://doi.org/10.1016/0001-6918(78)90006-9

Näätänen, R., & Picton, T. (1987). The N1 Wave of the Human Electric and Magnetic Response to Sound: A Review and an Analysis of the Component Structure. *Psychophysiology*, *24*(4), 375-425. https://doi.org/https://doi.org/10.1111/j.1469-8986.1987.tb00311.x

Nastase, S. A., Gazzola, V., Hasson, U., & Keysers, C. (2019). Measuring shared responses across subjects using intersubject correlation. *Social Cognitive and Affective Neuroscience*, *14*(6), 667-685. https://doi.org/10.1093/scan/nsz037

Nguyen, M., Vanderwal, T., & Hasson, U. (2019). Shared understanding of narratives is correlated with shared neural responses. *NeuroImage*, *184*, 161-170. https://doi.org/https://doi.org/10.1016/j.neuroimage.2018.09.010

Nilsson, M., Soli, S. D., & Sullivan, J. A. (1994). Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. *J Acoust Soc Am*, *95*(2), 1085-1099. https://doi.org/10.1121/1.408469

Noble, S., Curtiss, J., Pessoa, L., & Scheinost, D. (2024). The tip of the iceberg: A call to embrace anti-localizationism in human neuroscience research. *Imaging Neuroscience*, *2*, 1-10. https://doi.org/10.1162/imag_a_00138

Norman-Haignere, S. V., Long, L. K., Devinsky, O., Doyle, W., Irobunda, I., Merricks, E. M., Feldstein, N. A., McKhann, G. M., Schevon, C. A., Flinker, A., & Mesgarani, N. (2022). Multiscale temporal integration organizes hierarchical computation in human auditory cortex. *Nat Hum Behav*, *6*(3), 455-469. https://doi.org/10.1038/s41562-021-01261-y

O'Sullivan, J., Chen, Z., Herrero, J., McKhann, G. M., Sheth, S. A., Mehta, A. D., & Mesgarani, N. (2017). Neural decoding of attentional selection in multi-speaker environments without access to clean sources. *J Neural Eng*, *14*(5), 056001. https://doi.org/10.1088/1741-2552/aa7ab4

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A., & Lalor, E. C. (2015). Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. *Cereb Cortex*, *25*(7), 1697-1706. https://doi.org/10.1093/cercor/bht355

O'Sullivan, J. A., Shamma, S. A., & Lalor, E. C. (2015). Evidence for Neural Computations of Temporal Coherence in an Auditory Scene and Their Enhancement during Active Listening. *J Neurosci*, *35*(18), 7256-7263. https://doi.org/10.1523/jneurosci.4973-14.2015

O'Sullivan, J., Herrero, J., Smith, E., Schevon, C., McKhann, G. M., Sheth, S. A., Mehta, A. D., & Mesgarani, N. (2019). Hierarchical Encoding of Attended Auditory Objects in

Multi-talker Speech Perception. *Neuron*, *104*(6), 1195-1209.e1193. https://doi.org/https://doi.org/10.1016/j.neuron.2019.09.007

Oberem, J., Koch, I., & Fels, J. (2017). Intentional switching in auditory selective attention: Exploring age-related effects in a spatial setup requiring speech perception. *Acta Psychologica*, *177*, 36-43. https://doi.org/https://doi.org/10.1016/j.actpsy.2017.04.008

Obleser, J., Wöstmann, M., Hellbernd, N., Wilsch, A., & Maess, B. (2012). Adverse Listening Conditions and Memory Load Drive a Common Alpha Oscillatory Network. *The Journal of Neuroscience*, *32*(36), 12376-12383. https://doi.org/10.1523/jneurosci.4908-11.2012

Oganian, Y., Bhaya-Grossman, I., Johnson, K., & Chang, E. F. (2023). Vowel and formant representation in the human auditory speech cortex. *Neuron*, *111*(13), 2105-2118.e2104. https://doi.org/10.1016/j.neuron.2023.04.004

Oganian, Y., Kojima, K., Breska, A., Cai, C., Findlay, A., Chang, E., & Nagarajan, S. S. (2023). Phase Alignment of Low-Frequency Neural Activity to the Amplitude Envelope of Speech Reflects Evoked Responses to Acoustic Edges, Not Oscillatory Entrainment. *J Neurosci*, *43*(21), 3909-3921. https://doi.org/10.1523/jneurosci.1663-22.2023

Ogawa, S., Lee, T. M., Kay, A. R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences*, *87*(24), 9868-9872. https://doi.org/doi:10.1073/pnas.87.24.9868

Ohlenforst, B., Zekveld, A. A., Jansma, E. P., Wang, Y., Naylor, G., Lorens, A., Lunner, T., & Kramer, S. E. (2017). Effects of Hearing Impairment and Hearing Aid Amplification on Listening Effort: A Systematic Review. *Ear Hear*, *38*(3), 267-281. https://doi.org/10.1097/aud.0000000000000396

Ohlenforst, B., Zekveld, A. A., Jansma, E. P., Wang, Y., Naylor, G., Lorens, A., Lunner, T., & Kramer, S. E. (2017). Effects of Hearing Impairment and Hearing Aid Amplification on Listening Effort: A Systematic Review. *Ear and Hearing*, *38*(3). https://journals.lww.com/ear-hearing/fulltext/2017/05000/effects_of_hearing_impairment_and_hearing_aid.1.aspx

Oline, S. N., Ashida, G., & Burger, R. M. (2016). Tonotopic Optimization for Temporal Processing in the Cochlear Nucleus. *J Neurosci*, *36*(32), 8500-8515. https://doi.org/10.1523/jneurosci.4449-15.2016

Orf, M., Wöstmann, M., Hannemann, R., & Obleser, J. (2023). Target enhancement but not distractor suppression in auditory neural tracking during continuous speech. *iScience*, *26*(6). https://doi.org/10.1016/j.isci.2023.106849

Park, H., & Gross, J. (2023). Get the gist of the story: Neural map of topic keywords in multi-speaker environment. In: eLife Sciences Publications, Ltd.

Parthasarathy, A., Hancock, K. E., Bennett, K., DeGruttola, V., & Polley, D. B. (2020). Bottom-up and top-down neural signatures of disordered multi-talker speech perception in adults with normal hearing. *eLife*, *9*, e51419. https://doi.org/10.7554/eLife.51419

Parvizi, J., & Kastner, S. (2018). Promises and limitations of human intracranial electroencephalography. *Nat Neurosci*, *21*(4), 474-483. https://doi.org/10.1038/s41593-018-0108-2

Paul, B. T., Chen, J., Le, T., Lin, V., & Dimitrijevic, A. (2021). Cortical alpha oscillations in cochlear implant users reflect subjective listening effort during speech-in-noise perception. *PLOS ONE*, *16*(7), e0254162. https://doi.org/10.1371/journal.pone.0254162

Payne, B. R., & Federmeier, K. D. (2018). Contextual constraints on lexico-semantic processing in aging: Evidence from single-word event-related brain potentials. *Brain Res*, *1687*, 117-128. https://doi.org/10.1016/j.brainres.2018.02.021

Peelle, J. E. (2012). The hemispheric lateralization of speech processing depends on what "speech" is: a hierarchical perspective [Opinion]. *Frontiers in Human Neuroscience*, *Volume 6 - 2012*. https://doi.org/10.3389/fnhum.2012.00309

Peelle, J. E. (2018). Listening Effort: How the Cognitive Consequences of Acoustic Challenge Are Reflected in Brain and Behavior. *Ear and Hearing*, *39*(2), 204-214. https://doi.org/10.1097/aud.0000000000000494

Peelle, J. E., Troiani, V., Wingfield, A., & Grossman, M. (2010). Neural Processing during Older Adults' Comprehension of Spoken Sentences: Age Differences in Resource Allocation and Connectivity. *Cerebral Cortex*, *20*(4), 773-782. https://doi.org/10.1093/cercor/bhp142

Peelle, J. E., & Wingfield, A. (2016). The Neural Consequences of Age-Related Hearing Loss. *Trends in Neurosciences*, *39*(7), 486-497. https://doi.org/10.1016/j.tins.2016.05.001

Penfield, W., & Jasper, H. (1954). *Epilepsy and the functional anatomy of the human brain*. Little, Brown & Co.

Pennington, J., Socher, R., & Manning, C. (2014). *Glove: Global Vectors for Word Representation* (Vol. 14). https://doi.org/10.3115/v1/D14-1162

Petersen, E. B., Wöstmann, M., Obleser, J., & Lunner, T. (2017). Neural tracking of attended versus ignored speech is differentially affected by hearing loss. *J Neurophysiol*, *117*(1), 18-27. https://doi.org/10.1152/jn.00527.2016

Piazza, G., Carta, S., Ip, E. Y. J., Pérez-Navarro, J., Kalashnikova, M., Martin, C. D., & Di Liberto, G. M. (2025). Are you talking to me? How the choice of speech register impacts listeners' hierarchical encoding of speech. *Imaging Neuroscience*, *3*. https://doi.org/10.1162/imag_a_00539

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W. Y., Humes, L. E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C. L., Naylor, G., Phillips, N. A., Richter, M., Rudner, M., Sommers, M. S., Tremblay, K. L., & Wingfield, A. (2016). Hearing Impairment and Cognitive Energy: The Framework for Understanding Effortful Listening (FUEL). *Ear and Hearing*, *37*. https://journals.lww.com/ear-hearing/fulltext/2016/07001/hearing_impairment_and_cognitive_energy__the.2.aspx

Pichora-Fuller, M. K., Schneider, B. A., & Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *The Journal of the Acoustical Society of America*, *97*(1), 593-608. https://doi.org/10.1121/1.412282

Picton, T. (2013). Hearing in time: evoked potential studies of temporal processing. *Ear Hear*, *34*(4), 385-401. https://doi.org/10.1097/AUD.0b013e31827ada02

Picton, T. W., & Hillyard, S. A. (1974). Human auditory evoked potentials. II. Effects of attention. *Electroencephalogr Clin Neurophysiol*, *36*(2), 191-199. https://doi.org/10.1016/0013-4694(74)90156-4

Pinker, S. (1994). *The language instinct*. William Morrow & Co.

Poeppel, D., Emmorey, K., Hickok, G., & Pylkkänen, L. (2012). Towards a New Neurobiology of Language. *The Journal of Neuroscience*, *32*(41), 14125-14131. https://doi.org/10.1523/jneurosci.3244-12.2012

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clin Neurophysiol*, *118*(10), 2128-2148. https://doi.org/10.1016/j.clinph.2007.04.019

Posner, M. I. (1980). Orienting of Attention*. *Quarterly Journal of Experimental Psychology*, *32*(1), 3-25. https://doi.org/10.1080/00335558008248231

Power, A. J., Foxe, J. J., Forde, E.-J., Reilly, R. B., & Lalor, E. C. (2012). At what time is the cocktail party? A late locus of selective attention to natural speech. *European Journal of Neuroscience*, *35*(9), 1497-1503. https://doi.org/https://doi.org/10.1111/j.1460-9568.2012.08060.x

Presacco, A., Simon, J. Z., & Anderson, S. (2016a). Effect of informational content of noise on speech representation in the aging midbrain and cortex. *Journal of Neurophysiology*, *116*(5), 2356-2367. https://doi.org/10.1152/jn.00373.2016

Presacco, A., Simon, J. Z., & Anderson, S. (2016b). Evidence of degraded representation of speech in noise, in the aging midbrain and cortex. *Journal of Neurophysiology*, *116*(5), 2346-2355. https://doi.org/10.1152/jn.00372.2016

Presacco, A., Simon, J. Z., & Anderson, S. (2019). Speech-in-noise representation in the aging midbrain and cortex: Effects of hearing loss. *PLOS ONE*, *14*(3), e0213899. https://doi.org/10.1371/journal.pone.0213899

Pressnitzer, D., Sayles, M., Micheyl, C., & Winter, I. M. (2008). Perceptual Organization of Sound Begins in the Auditory Periphery. *Current Biology*, *18*(15), 1124-1128. https://doi.org/https://doi.org/10.1016/j.cub.2008.06.053

Prinsloo, K. D., & Lalor, E. C. (2022). General Auditory and Speech-Specific Contributions to Cortical Envelope Tracking Revealed Using Auditory Chimeras. *The Journal of Neuroscience*, *42*(41), 7782-7798. https://doi.org/10.1523/jneurosci.2735-20.2022

Pulvermüller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., & Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences*, *103*(20), 7865-7870. https://doi.org/doi:10.1073/pnas.0509989103

Puschmann, S., Regev, M., Baillet, S., & Zatorre, R. J. (2021). MEG Intersubject Phase Locking of Stimulus-Driven Activity during Naturalistic Speech Listening Correlates with Musical Training. *The Journal of Neuroscience*, *41*(12), 2713-2722. https://doi.org/10.1523/jneurosci.0932-20.2020

Puvvada, K. C., & Simon, J. Z. (2017). Cortical Representations of Speech in a Multitalker Auditory Scene. *J Neurosci*, *37*(38), 9189-9196. https://doi.org/10.1523/jneurosci.0938-17.2017

Radford, A., & Narasimhan, K. (2018). Improving Language Understanding by Generative Pre-Training.

Raghavan, V. S., O'Sullivan, J., Bickel, S., Mehta, A. D., & Mesgarani, N. (2023). Distinct neural encoding of glimpsed and masked speech in multitalker situations. *PLoS Biol*, *21*(6), e3002128. https://doi.org/10.1371/journal.pbio.3002128

Raghavan, V. S., O'Sullivan, J., Herrero, J., Bickel, S., Mehta, A. D., & Mesgarani, N. (2024). Improving auditory attention decoding by classifying intracranial responses to glimpsed and masked acoustic events. *Imaging Neuroscience*, *2*, 1-19. https://doi.org/10.1162/imag_a_00148

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79-87. https://doi.org/10.1038/4580

Rauschecker, J. P., & Afsahi, R. K. (2023). Anatomy of the auditory cortex then and now. *Journal of Comparative Neurology*, *531*(18), 1883-1892. https://doi.org/https://doi.org/10.1002/cne.25560

Raven, J., & Raven, J. (2003). Raven Progressive Matrices. In *Handbook of nonverbal assessment.* (pp. 223-237). Kluwer Academic/Plenum Publishers. https://doi.org/10.1007/978-1-4615-0153-4_11

Ravizza, S. M., & Carter, C. S. (2008). Shifting set about task switching: behavioral and neural evidence for distinct forms of cognitive flexibility. *Neuropsychologia*, *46*(12), 2924-2935. https://doi.org/10.1016/j.neuropsychologia.2008.06.006

Regev, T. I., Casto, C., Hosseini, E. A., Adamek, M., Ritaccio, A. L., Willie, J. T., Brunner, P., & Fedorenko, E. (2024). Neural populations in the language network differ in the size

of their temporal receptive windows. *Nature Human Behaviour*, *8*(10), 1924-1942. https://doi.org/10.1038/s41562-024-01944-2

Rhode, W. S., & Smith, P. H. (1986). Encoding timing and intensity in the ventral cochlear nucleus of the cat. *Journal of Neurophysiology*, *56*(2), 261-286. https://doi.org/10.1152/jn.1986.56.2.261

Rivenez, M., Darwin, C. J., & Guillaume, A. (2006). Processing unattended speech. *The Journal of the Acoustical Society of America*, *119*(6), 4027-4040. https://doi.org/10.1121/1.2190162

Rogalsky, C., Basilakos, A., Rorden, C., Pillay, S., LaCroix, A. N., Keator, L., Mickelsen, S., Anderson, S. W., Love, T., Fridriksson, J., Binder, J., & Hickok, G. (2022). The Neuroanatomy of Speech Processing: A Large-scale Lesion Study. *J Cogn Neurosci*, *34*(8), 1355-1375. https://doi.org/10.1162/jocn_a_01876

Rogers, C. S., & Wingfield, A. (2015). Stimulus-independent semantic bias misdirects word recognition in older adults. *J Acoust Soc Am*, *138*(1), El26-30. https://doi.org/10.1121/1.4922363

Rogers, R. D., & Monsell, S. (1995). Costs of a predictible switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, *124*(2), 207-231. https://doi.org/10.1037/0096-3445.124.2.207

Rosenfelder, I., Fruehwald, J., Evanini, K., Seyfarth, S., Gorman, K., Prichard, H., & Yuan, J. (2014). FAVE (Forced Alignment and Vowel Extraction) Suite Version 1.1.3. . https://doi.org/https://doi.org/10.5281/zenodo.9846

Rowland, C. F., Westermann, G., Theakston, A. L., Pine, J. M., Monaghan, P., & Lieven, E. V. M. (2025). Constructing language: a framework for explaining acquisition. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2025.05.015

Ru, X., He, K., Lyu, B., Li, D., Xu, W., Gu, W., Ma, X., Liu, J., Li, C., Li, T., Zheng, F., Yan, X., Yin, Y., Duan, H., Na, S., Wan, S., Qin, J., Sheng, J., & Gao, J.-H. (2022). Multimodal neuroimaging with optically pumped magnetometers: A simultaneous MEG-EEG-fNIRS acquisition system. *NeuroImage*, *259*, 119420. https://doi.org/https://doi.org/10.1016/j.neuroimage.2022.119420

Rufener, K. S., Liem, F., & Meyer, M. (2014). Age-related differences in auditory evoked potentials as a function of task modulation during speech–nonspeech processing. *Brain and Behavior*, *4*(1), 21-28. https://doi.org/https://doi.org/10.1002/brb3.188

Salthouse, T. A. (2010). Selective review of cognitive aging [Review]. *Journal of the International Neuropsychological Society*, *16*(5), 754-760. https://doi.org/10.1017/S1355617710000706

Schiel, F. (1999). *Automatic Phonetic Transcription of Non-Prompted Speech* 14th International Congress of Phonetic Sciences, San Francisco, USA. https://epub.ub.uni-muenchen.de/13682/

Schiel, F. (2015). A statistical model for predicting pronunciation. International Congress of Phonetic Sciences,

Schmitt, L.-M., Erb, J., Tune, S., Rysop, A. U., Hartwigsen, G., & Obleser, J. (2021). Predicting speech from a cortical hierarchy of event-based time scales. *Science Advances*, *7*(49), eabi6070. https://doi.org/doi:10.1126/sciadv.abi6070

Schmitt, R., Meyer, M., & Giroud, N. (2022). Better speech-in-noise comprehension is associated with enhanced neural speech tracking in older adults with hearing impairment. *Cortex*, *151*, 133-146. https://doi.org/https://doi.org/10.1016/j.cortex.2022.02.017

Schneider, B. A., Daneman, M., & Pichora-Fuller, M. K. (2002). Listening in aging adults: from discourse comprehension to psychoacoustics. *Can J Exp Psychol*, *56*(3), 139-152. https://doi.org/10.1037/h0087392

Schönmann, I., Szewczyk, J., de Lange, F. P., & Heilbron, M. (2025). Stimulus dependencies—rather than next-word prediction—can explain pre-onset brain encoding during natural listening. In: eLife Sciences Publications, Ltd.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118. https://doi.org/doi:10.1073/pnas.2105646118

Seifi Ala, T., Graversen, C., Wendt, D., Alickovic, E., Whitmer, W. M., & Lunner, T. (2020). An exploratory Study of EEG Alpha Oscillation and Pupil Dilation in Hearing-Aid Users During Effortful listening to Continuous Speech. *PLOS ONE*, *15*(7), e0235782. https://doi.org/10.1371/journal.pone.0235782

Shahsavari Baboukani, P., Graversen, C., Alickovic, E., & Østergaard, J. (2022). Speech to noise ratio improvement induces nonlinear parietal phase synchrony in hearing aid users [Original Research]. *Frontiers in Neuroscience*, *16*. https://doi.org/10.3389/fnins.2022.932959

Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., & Fedorenko, E. (2020). fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, *138*, 107307. https://doi.org/https://doi.org/10.1016/j.neuropsychologia.2019.107307

Shamma, S., Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., Pressnitzer, D., Yin, P., & Xu, Y. (2013). Temporal coherence and the streaming of complex sounds. *Adv Exp Med Biol*, *787*, 535-543. https://doi.org/10.1007/978-1-4614-1590-9_59

Shamma, S. A., Elhilali, M., & Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. *Trends Neurosci*, *34*(3), 114-123. https://doi.org/10.1016/j.tins.2010.11.002

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*(3), 379-423. https://doi.org/https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Sheldon, S., Sheldon, J., Zhang, S., Setton, R., Turner, G. R., Spreng, R. N., & Grilli, M. D. (2024). Differences in the content and coherence of autobiographical memories between younger and older adults: Insights from text analysis. *Psychol Aging*, *39*(1), 59-71. https://doi.org/10.1037/pag0000769

Shinn-Cunningham, B., Best, V., Kingstone, A., Fawcett, J., & Risko, E. (2015). Auditory selective attention. *The handbook of attention*, *99*.

Shinn-Cunningham, B., & Lee, A. K. C. (2017). Auditory Object Formation and Selection. In (pp. 7-40). https://doi.org/10.1007/978-3-319-51662-2_2

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends Cogn Sci*, *12*(5), 182-186. https://doi.org/10.1016/j.tics.2008.02.003

Shinn-Cunningham, B. G., & Best, V. (2008). Selective attention in normal and impaired hearing. *Trends Amplif*, *12*(4), 283-299. https://doi.org/10.1177/1084713808325306

Shomstein, S., & Yantis, S. (2006). Parietal cortex mediates voluntary control of spatial and nonspatial auditory attention. *J Neurosci*, *26*(2), 435-439. https://doi.org/10.1523/jneurosci.4408-05.2006

Simon, J. Z. (2015). The encoding of auditory objects in auditory cortex: Insights from magnetoencephalography. *International Journal of Psychophysiology*, *95*(2), 184-190. https://doi.org/https://doi.org/10.1016/j.ijpsycho.2014.05.005

Singh, N. C., & Theunissen, F. E. (2003). Modulation spectra of natural sounds and ethological theories of auditory processing. *The Journal of the Acoustical Society of America*, *114*(6), 3394-3411. https://doi.org/10.1121/1.1624067

Slaats, S., & Martin, A. E. (2025). What's Surprising About Surprisal. *Computational Brain & Behavior*, *8*(2), 233-248. https://doi.org/10.1007/s42113-025-00237-9

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, *416*(6876), 87-90. https://doi.org/10.1038/416087a

Sohoglu, E., & Davis, M. H. (2016). Perceptual learning of degraded speech by minimizing prediction error. *Proceedings of the National Academy of Sciences*, *113*(12), E1747-E1756. https://doi.org/doi:10.1073/pnas.1523266113

Spence, C., & Driver, J. (1994). Covert Spatial Orienting in Audition: Exogenous and Endogenous Mechanisms. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 555-574. https://doi.org/10.1037/0096-1523.20.3.555

Steinschneider, M., Nourski, K. V., & Fishman, Y. I. (2013). Representation of speech in human auditory cortex: Is it special? *Hearing Research*, *305*, 57-73. https://doi.org/https://doi.org/10.1016/j.heares.2013.05.013

Stella, M., Citraro, S., Rossetti, G., Marinazzo, D., Kenett, Y. N., & Vitevitch, M. S. (2024). Cognitive modelling of concepts in the mental lexicon with multilayer networks: Insights, advancements, and future challenges. *Psychonomic Bulletin & Review*, *31*(5), 1981-2004. https://doi.org/10.3758/s13423-024-02473-9

Stenbäck, V., Marsja, E., Hällgren, M., Lyxell, B., & Larsby, B. (2021). The Contribution of Age, Working Memory Capacity, and Inhibitory Control on Speech Recognition in Noise in Young and Older Adult Listeners. *Journal of Speech, Language, and Hearing Research*, *64*(11), 4513-4523. https://doi.org/doi:10.1044/2021_JSLHR-20-00251

Stilp, C. E., & Kluender, K. R. (2010). Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility. *Proceedings of the National Academy of Sciences*, *107*(27), 12387-12392. https://doi.org/doi:10.1073/pnas.0913625107

Sweet, R. A., Dorph-Petersen, K.-A., & Lewis, D. A. (2005). Mapping auditory core, lateral belt, and parabelt cortices in the human superior temporal gyrus. *Journal of Comparative Neurology*, *491*(3), 270-289. https://doi.org/https://doi.org/10.1002/cne.20702

Teoh, E. S., Ahmed, F., & Lalor, E. C. (2022). Attention Differentially Affects Acoustic and Phonetic Feature Encoding in a Multispeaker Environment. *The Journal of Neuroscience*, *42*(4), 682-691. https://doi.org/10.1523/jneurosci.1455-20.2021

Teoh, E. S., Cappelloni, M. S., & Lalor, E. C. (2019). Prosodic pitch processing is represented in delta-band EEG and is dissociable from the cortical tracking of other acoustic and phonetic features. *Eur J Neurosci*, *50*(11), 3831-3842. https://doi.org/10.1111/ejn.14510

Teoh, E. S., & Lalor, E. C. (2019). EEG decoding of the target speaker in a cocktail party scenario: considerations regarding dynamic switching of talker location. *J Neural Eng*, *16*(3), 036017. https://doi.org/10.1088/1741-2552/ab0cf1

Thomson, R. S., Auduong, P., Miller, A. T., & Gurgel, R. K. (2017). Hearing loss as a risk factor for dementia: A systematic review. *Laryngoscope Investigative Otolaryngology*, *2*(2), 69-79. https://doi.org/https://doi.org/10.1002/lio2.65

Thornton, M., Mandic, D., & Reichenbach, T. (2022). Robust decoding of the speech envelope from EEG recordings through deep neural networks. *Journal of Neural Engineering*, *19*(4), 046007. https://doi.org/10.1088/1741-2552/ac7976

Tikochinski, R., Goldstein, A., Meiri, Y., Hasson, U., & Reichart, R. (2025). Incremental accumulation of linguistic context in artificial and biological neural networks. *Nature Communications*, *16*(1), 803. https://doi.org/10.1038/s41467-025-56162-9

Trecca, F., Bleses, D., Højen, A., Madsen, T. O., & Christiansen, M. H. (2020). When Too Many Vowels Impede Language Processing: An Eye-Tracking Study of Danish-

Learning Children. *Lang Speech*, *63*(4), 898-918. https://doi.org/10.1177/0023830919893390

Treisman, A. (1964). Monitoring and storage of irrelevant messages in selective attention. *Journal of Verbal Learning and Verbal Behavior*, *3*(6), 449-459. https://doi.org/https://doi.org/10.1016/S0022-5371(64)80015-3

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97-136. https://doi.org/https://doi.org/10.1016/0010-0285(80)90005-5

Tremblay, K. L., Billings, C., & Rohila, N. (2004). Speech evoked cortical potentials: effects of age and stimulus presentation rate. *J Am Acad Audiol*, *15*(3), 226-237; quiz 264. https://doi.org/10.3766/jaaa.15.3.5

Tremblay, K. L., Piskosz, M., & Souza, P. (2002). Aging alters the neural representation of speech cues. *NeuroReport*, *13*(15), 1865-1870. https://doi.org/10.1097/00001756-200210280-00007

Tremblay, K. L., Piskosz, M., & Souza, P. (2003). Effects of age and age-related hearing loss on the neural representation of speech cues. *Clinical Neurophysiology*, *114*(7), 1332-1343. https://doi.org/https://doi.org/10.1016/S1388-2457(03)00114-7

Tune, S., Alavash, M., Fiedler, L., & Obleser, J. (2021). Neural attentional-filter mechanisms of listening success in middle-aged and older individuals. *Nat Commun*, *12*(1), 4533. https://doi.org/10.1038/s41467-021-24771-9

van der Heijden, M., Lorteije, Jeannette A. M., Plauška, A., Roberts, Michael T., Golding, Nace L., & Borst, J. Gerard G. (2013). Directional Hearing by Linear Summation of Binaural Inputs at the Medial Superior Olive. *Neuron*, *78*(5), 936-948. https://doi.org/https://doi.org/10.1016/j.neuron.2013.04.028

Wacongne, C., Changeux, J.-P., & Dehaene, S. (2012). A Neuronal Model of Predictive Coding Accounting for the Mismatch Negativity. *The Journal of Neuroscience*, *32*(11), 3665-3678. https://doi.org/10.1523/jneurosci.5003-11.2012

Wager, T. D., Jonides, J., & Reading, S. (2004). Neuroimaging studies of shifting attention: a meta-analysis. *NeuroImage*, *22*(4), 1679-1693. https://doi.org/https://doi.org/10.1016/j.neuroimage.2004.03.052

Walsh, K. S., McGovern, D. P., Clark, A., & O'Connell, R. G. (2020). Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Annals of the New York Academy of Sciences*, *1464*(1), 242-268. https://doi.org/https://doi.org/10.1111/nyas.14321

Walter, W. G., Cooper, R., Aldridge, V. J., McCallum, W. C., & Winter, A. L. (1964). CONTINGENT NEGATIVE VARIATION: AN ELECTRIC SIGN OF SENSORIMOTOR ASSOCIATION AND EXPECTANCY IN THE HUMAN BRAIN. *Nature*, *203*, 380-384. https://doi.org/10.1038/203380a0

Wang, Y., Zhang, J., Zou, J., Luo, H., & Ding, N. (2018). Prior Knowledge Guides Speech Segregation in Human Auditory Cortex. *Cerebral Cortex*, *29*(4), 1561-1571. https://doi.org/10.1093/cercor/bhy052

Wechsler, D. (1940). THE MEASUREMENT OF ADULT INTELLIGENCE. *The Journal of Nervous and Mental Disease*, *91*(4). https://journals.lww.com/jonmd/fulltext/1940/04000/the_measurement_of_adult_intelligence.75.aspx

Weissbart, H., & Martin, A. E. (2024). The structure and statistics of language jointly shape cross-frequency neural dynamics during spoken language comprehension. *Nature Communications*, *15*(1), 8850. https://doi.org/10.1038/s41467-024-53128-1

Wendel, K., Väisänen, O., Malmivuo, J., Gencer, N. G., Vanrumste, B., Durka, P., Magjarević, R., Supek, S., Pascu, M. L., Fontenelle, H., & Grave de Peralta Menendez, R. (2009).

EEG/MEG source imaging: methods, challenges, and open issues. *Comput Intell Neurosci*, *2009*, 656092. https://doi.org/10.1155/2009/656092

Wendt, D., Hietkamp, R. K., & Lunner, T. (2017). Impact of Noise and Noise Reduction on Processing Effort: A Pupillometry Study. *Ear and Hearing*, *38*(6), 690-700. https://doi.org/10.1097/aud.0000000000000454

Wilroth, J., Alickovic, E., Skoglund, M. A., Signoret, C., Rönnberg, J., & Enqvist, M. (2025). Improving Tracking of Selective Attention in Hearing Aid Users: The Role of Noise Reduction and Nonlinearity Compensation. *eneuro*, *12*(2), ENEURO.0275-0224.2025. https://doi.org/10.1523/eneuro.0275-24.2025

Winer, J. A., & Schreiner, C. E. (2005). *The Inferior Colliculus*. Springer New York, NY. https://doi.org/https://doi.org/10.1007/b138578

Winkler, I., & Czigler, I. (1998). Mismatch negativity: deviance detection or the maintenance of the 'standard'. *NeuroReport*, *9*(17). https://journals.lww.com/neuroreport/fulltext/1998/12010/mismatch_negativity__deviance_detection_or_the.8.aspx

Winn, M. B., Wendt, D., Koelewijn, T., & Kuchinsky, S. E. (2018). Best Practices and Advice for Using Pupillometry to Measure Listening Effort: An Introduction for Those Who Want to Get Started. *Trends in Hearing*, *22*, 2331216518800869. https://doi.org/10.1177/2331216518800869

Wlotko, E. W., Lee, C. L., & Federmeier, K. D. (2010). Language of the aging brain: Event-related potential studies of comprehension in older adults. *Lang Linguist Compass*, *4*(8), 623-638. https://doi.org/10.1111/j.1749-818X.2010.00224.x

Wöstmann, M., Lim, S. J., & Obleser, J. (2017). The Human Neural Alpha Response to Speech is a Proxy of Attentional Control. *Cereb Cortex*, *27*(6), 3307-3317. https://doi.org/10.1093/cercor/bhx074

Wu, P. Z., & Liberman, M. C. (2022). Age-related stereocilia pathology in the human cochlea. *Hear Res*, *422*, 108551. https://doi.org/10.1016/j.heares.2022.108551

Wu, W., & Hoffman, P. (2022). Validated measures of semantic knowledge and semantic control: normative data from young and older adults for more than 300 semantic judgements. *R Soc Open Sci*, *9*(2), 211056. https://doi.org/10.1098/rsos.211056

Young, N. J., & McGarrah, M. (2021). Forced alignment for Nordic languages: Rapidly constructing a high-quality prototype. *Nordic Journal of Linguistics*, 1-27. https://doi.org/10.1017/S033258652100024X

Zaar, J., Simonsen, L. B., Sanchez-Lopez, R., & Laugesen, S. (2024). The Audible Contrast Threshold (ACT) test: A clinical spectro-temporal modulation detection test. *Hearing Research*, *453*, 109103. https://doi.org/https://doi.org/10.1016/j.heares.2024.109103

Zan, P., Presacco, A., Anderson, S., & Simon, J. Z. (2020). Exaggerated cortical representation of speech in older listeners: mutual information analysis. *Journal of Neurophysiology*, *124*(4), 1152-1164. https://doi.org/10.1152/jn.00002.2020

Zatorre, R. J. (2022). Hemispheric asymmetries for music and speech: Spectrotemporal modulations and top-down influences. *Front Neurosci*, *16*, 1075511. https://doi.org/10.3389/fnins.2022.1075511

Zhang, Y., Han, K., Worth, R., & Liu, Z. (2020). Connecting concepts in the brain by mapping cortical representations of semantic relations. *Nature Communications*, *11*(1), 1877. https://doi.org/10.1038/s41467-020-15804-w

Zink, M. E., Zhen, L., McHaney, J. R., Klara, J., Yurasits, K., Cancel, V. E., Flemm, O., Mitchell, C., Datta, J., Chandresekaran, B., & Parthasarathy, A. (2025). Increased listening effort and cochlear neural degeneration underlie speech-in-noise deficits in normal-hearing middle-aged adults. *eLife*, *13*, RP102823. https://doi.org/10.7554/eLife.102823

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., Goodman, R. R., Emerson, R., Mehta, A. D., Simon, J. Z., Poeppel, D., & Schroeder, C. E. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron*, *77*(5), 980-991. https://doi.org/10.1016/j.neuron.2012.12.037

Zoefel, B., & Kösem, A. (2024). Neural tracking of continuous acoustics: properties, speech-specificity and open questions. *European Journal of Neuroscience*, *59*(3), 394-414. https://doi.org/https://doi.org/10.1111/ejn.16221

Zoefel, B., & VanRullen, R. (2016). EEG oscillations entrain their phase to high-level features of speech sound. *NeuroImage*, *124*, 16-23. https://doi.org/https://doi.org/10.1016/j.neuroimage.2015.08.054