Characterizing neural mechanisms of attention-driven speech processing

Fuglsang, Søren

Publication date:
2018

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

CONTRIBUTIONS TO
HEARING RESEARCH

Volume 37

Søren Asp Fuglsang

Characterizing neural mechanisms of attention-driven speech processing

DTU Electrical Engineering
Department of Electrical Engineering

Characterizing neural mechanisms of attention-driven speech processing

PhD thesis by
Søren Asp Fuglsang

Preliminary version: November 13, 2018

Technical University of Denmark

2018

This PhD dissertation is the result of a research project carried out at the Hearing Systems Group, Department of Electrical Engineering, Technical University of Denmark.

The project was partly financed by the EU H2020-ICT grant 644732 (COCOHA) (2/3) and by the Technical University of Denmark (1/3).

Supervisors

Prof. Torsten Dau
Hearing Systems Group, Department of Electrical Engineering
Technical University of Denmark
Kgs. Lyngby, Denmark

Dr. Jens Hjortkjær
Hearing Systems Group, Department of Electrical Engineering
Technical University of Denmark
Kgs. Lyngby, Denmark

Acknowledgments

I would like to express my deepest gratitude to Torsten Dau and Jens Hjortkjær for excellent guidance, endless support, and encouragement. I also want to thank Alain de Cheivegné, Daniel D. E. Wong and Jonatan Marcher-Rørsted for close collaborations and inspiring discussions.

I want to thank the all members of the COCOHA team, including Malcolm Slaney, Thomas Lunner, Carina Graversen, Giovanni Di Liberto, Enea Ceolini, Dorothée Arzounian and Sergi Rotger-Griful, for collaborations. I also want to thank Lars Kai Hansen, Mette V. Carstenssen, Søren Vørnle and Nicolai Pedersen for collaborations, Sarah K. Madsen and Alain de Cheivegné for helpful comments on my dissertation and Oticon for providing EarEEG systems. Finally, I want to thank all colleagues from DTU Hearing Systems for many good moments and stimulating discussions.

Abstract

The ability to selectively attend to speech in adverse listening environments plays a key role in human communication. However, listening to speech in everyday multi-talker environments can be challenging for hearing-impaired listeners, even when wearing modern hearing aids. Studies in normal-hearing listeners have shown that it is possible to decode which of two speakers a listener is attending to from scalp electroencephalography (EEG). This has led to the idea that EEG-based brain-computer interfaces (BCIs) could be integrated into hearing assistive devices to help hearing-impaired listeners by selectively amplifying attended sound sources. To accomplish this, however, single-trial EEG measures of attention to speech must first be investigated in hearing-impaired listeners, in everyday listening scenarios, and under different task demands.

This thesis explored single-trial cortical EEG correlates of selective attention to speech. In the first study, the influence of sensorineural hearing loss on cortical EEG responses to tones and to naturalistic speech was investigated. It was shown that hearing impairment enhances the fidelity of the low-frequency cortical EEG entrainment to envelopes of simple tones and to envelopes of attended speech streams. For loudness-matched competing speech streams, the attended target could be classified from single-trial EEG responses with equally high classification accuracies in normal- and hearing-impaired listeners. The second study explored single-trial EEG correlates of selective auditory attention to speech in reverberant, multi-talker environments. It was shown that the attentional selection of normal-hearing listeners could be decoded from single-trial EEG data with equally high classification accuracies in both anechoic- and reverberant listening environments. The third study analyzed how different constraints on EEG-based stimulus-response models influence the predictive power of the models. The results from this study suggested that stimulus-response model regularization is important for maintaining high classification accuracies with backward decoding models. The fourth study investigated if working memory demands affect single-trial EEG correlates of speech

envelope processing. Using an auditory n-back task, it was found that working memory load can affect spatio-spectral EEG power in the theta and alpha bands and that EEG measures of speech envelope entrainment decrease with high task load. The fifth study investigated whether EEG-based attention decoding can be achieved in a real-time closed-loop BCI system. Here, it was shown that a hearing-impaired listener was able to selectively amplify attended speech using a closed-loop EEG-based attention decoding BCI system. Overall, the work presented in this thesis suggests that EEG-based attention decoding may have relevance for future BCI systems.

Resumé

Evnen til selektivt at lytte til tale i komplekse lydmiljøer spiller en vigtig rolle for menneskets evne til at kommunikere. Selv med høreapparater kan hørehæmmede dog have problemer med at forstå tale når flere taler på samme tid. Nyere studier med normalthørende lyttere har vist, at det er muligt at afkode hvilken lytter man fokuserer på fra elektroencefalografi (EEG). Det har åbnet op for muligheden for at inkorporere EEG-baserede brain-computer interface (BCI) systemer i høreapparater, således at høreapparaterne selektivt kan forstærke den lydkilde, som den hørehæmmede bruger fokuserer på. For at det overhovedet skulle kunne lade sig gøre, bliver man dog nødt til at undersøge, hvordan EEG korrelater af selektiv opmærksomhed på tale påvirkes af høretab, af udfordrende lyttemiljøer samt af lytterens kognitive bearbejdning af tale.

Denne afhandling undersøgte EEG korrelater af selektiv auditiv opmærksomhed på tale. I det første studie blev det undersøgt, hvordan et sensorineuralt høretab påvirker kortikalt EEG målt fra personer, der lytter til enten toner eller til tale. Det blev her vist, at et høretab kan påvirke måden, hvorpå EEG signaler synkroniserer til tone-sekvenser og til envelopes af talesignaler. Det blev vist, at man fra EEG optagelser i scenarier med to simulatane lydstyrke-balancerede talere kunne afkode hvilken taler, lytterne fokuserede på. Tilsvarende høje klassifikationsnøjagtigheder blev fundet i de normalthørende og i de hørehæmmede forsøgspersoner. Det andet studie udforskede EEG korrelater af selektiv auditiv opmærksomhed i lydmiljøer med to eller flere talere og forskellige grader af rumklang. Det blev her vist, at det er muligt at afkode selektiv auditiv opmærksomhed fra EEG med tilsvarende høje klassifikationsnøjagtigheder i simulerede rum med efterklang samt i lydmiljøer med baggrundsstøj. Det tredje studie undersøgte, hvordan forskellige typer af regularisering påvirker EEG-baserede stimulus-respons modellers evne til at prædiktere data. Resultaterne pegede i retning af, at det var vigtigt at regularisere afkodningsmodeller for at opnå høje klassifikationsnøjagtigheder. Det fjerde studie undersøgte, hvorvidt forskellig belastning af arbejdshukommelsen påvirker EEG korrelater af tale-envelope

bearbejdning. Ved brug af en auditiv n-tilbage opgave blev det her vist, at det at bebyrde arbejdshukommelsen kan påvirke EEG effektmål i theta og alpha frekvensbånd, men på samme tid også påvirke måden, hvorpå EEG signaler synkroniserer til tale envelopes. Det femte studie undersøgte, hvorvidt EEG-baseret afkodning af auditiv opmærksomhed kunne opnås i et real-tids closed-loop BCI system. Det blev her vist, at en hørehæmmet bruger var i stand til selektivt at forstærke enkelte lydkilder ved brug af det EEG-baserede closed-loop BCI system. Samlet set indikerer resultaterne fra denne afhandling, at EEG-baseret afkodning kan have relevans for fremtidige BCI systemer.

Related publications

Journal papers

- Fuglsang, S. A.; Marcher-Rørsted, J.; Dau, T.; Hjortkjær, J. H (2018). The influence of a sensorineural hearing loss on cortical auditory EEG entrainment during selective listening *In preparation*
- Wong, D. E.; Fuglsang, S. A.; Ceolini, E.; Hjortkjær, J. H.; Slaney, M.; de Cheveigné, A. (2018). A Comparison of Regularization Methods in Forward and Backward Models for Auditory Attention Decoding *Frontiers in Human Neuroscience*
- de Cheveigné, A; di Liberto, G. M.; Arzounian, D.; Wong, D. E.; Hjortkjær, J. H.; Fuglsang, S. A.; Parra, L. C. (2018). Multiway Canonical Correlation Analysis of Brain Signals *NeuroImage*, Under review.
- Hjortkjær, J. H.; Marcher-Rørsted, J.; Fuglsang, S.A. & Dau, T (2017). Cortical oscillations and entrainment in speech processing during working memory load *European Journal of Neuroscience*
- Fuglsang, S.A., Dau, T. & Hjortkjær, J. H. (2017). Noise-robust cortical tracking of attended speech in real-world acoustic scenes *NeuroImage*.

Published abstracts

- Marcher-Rørsted, J.; Fuglsang, S. A.; Wong, D. E.; Dau, T. D.; Hjortkjær, J. H. (2017). Closed-loop BCI control of auditory feedback using selective attention *International Hearing Aid Research Conference (IHCON), Tahoe City, California, August 2018*
- Fuglsang, S. A.; Marcher-Rørsted, J.; Dau, T. D.; Hjortkjær, J. H. (2017). Cortical EEG entrainment to attended speech in hearing-impaired listen-

ers *International Hearing Aid Research Conference (IHCON), Tahoe City, California, August 2018*

- Wong, D.; Fuglsang, S. A.; Hjortkjær, J.; di Liberto, G. M.; de Cheveigné, A. (2017). Classifying Attended Talker from EEG Using Artificial Neural Networks *Association for Research in Otolaryngology (ARO), Mid-Winter Meeting, San Diego, February 2017*

- Marcher-Rørsted, J.; Fuglsang, S. A.; Dau, T. & Hjortkjær , (2017). Dynamics of cortical oscillations during an auditory N-back task *International Symposium on Auditory and Audiological Research (ISAAR), Nyborg, Denmark, August 2017.*

- Carstensen, M. L. V.; Fuglsang, S. A.; Dau, T. & Hjortkjær , (2017). Sensitivity of encoding models for auditory fMRI *International Symposium on Auditory and Audiological Research (ISAAR), Nyborg, Denmark, August 2017.*

- Fuglsang, S. A.; Dau, T. & Hjortkjær , (2017). Decoding attentional modulations of single-trial EEG responses in real-world acoustic scenes *European Federation of Audiology Societies (EFAS)*

- Wong, D.D.E.; Hjortkjær, Fuglsang, S.A.; de Cheveigné, A, (2017). Classifying Attended Speech from EEG Using Canonical Correlation *Association for Research in Otolaryngology (ARO), Mid-Winter Meeting, Baltimore, MA, February 2017. Volume 40, P. 331*

- Fuglsang, S. A.; Dau, T. & Hjortkjær , (2016). Neural reconstructions of speech in reverberant multi-talker environments *Association for Research in Otolaryngology (ARO), Mid-Winter Meeting, Baltimore, MA, February 2017. Volume 40, P. 328*

Contents

1

General introduction

In everyday multi-talker situations, listeners rely on selective attention to focus on one particular speech stream and ignore irrelevant ones. Auditory attention may enhance the cortical responses to sounds that are attended relative to ignored ones (e.g., Hillyard et al., 1973; Kaya and Elhilali, 2017; Mesgarani and Chang, 2012; Mesgarani et al., 2010; Woldorff et al., 1993). Which of two competing speech streams a listener is attending to may even be decoded from short segments of unaveraged electroencephalogram (EEG) data (Mirkovic et al., 2015; O'Sullivan et al., 2014). This opens the possibility of using EEG in brain-computer-interfaces (BCIs) that decode the attentional selection of listeners in multi-talker scenarios and enhance attended sound sources. Yet, many uncertain aspects of this idea remain to be resolved. Although EEG-based attention decoding may be achieved in relatively simple listening scenarios, e.g. with two competing talkers (Mirkovic et al., 2015; O'Sullivan et al., 2014), it is unclear how robust such decoding is in everyday listening environments where speech signals are corrupted by reverberation and background noise. Normal-hearing listeners are able to navigate almost effortlessly through complex auditory environments, but little is known about the cortical processes underlying this perceptual robustness. Hearing-impaired listeners are commonly challenged in complex listening situations, even when wearing hearing aids, but it is not clear how hearing impairment affects cortical EEG responses to sound mixtures. It is also unclear how cognitive aspects, such as mental effort or working memory load, influence single-trial EEG measures of speech processing. Finally, real-time acoustic feedback that enhances the attended sound stream in a BCI context constitutes a feedback loop between the changed acoustic input and the corresponding EEG correlates of brain processing. The dynamics of such a 'closed-loop' system are not yet understood and it is unclear to what extent a hearing-impaired person can profit from such a system.

The aim of this thesis was to explore these questions. The thesis is divided in seven main chapters, as listed below:

Chapter 2 provides an introduction to the concept of stimulus-response models that can be used to analyze single-trial EEG responses to naturalistic continuous speech.

Chapter 3 examines the effects of hearing-impairment on single-trial EEG measures of auditory attention. Selective speech-listening tasks are considered. Moreover, this chapter considers EEG experiments with synthetic sounds to probe temporal auditory processing abilities. Results from these experiments are compared to behavioral measures of speech intelligibility and basic spectro-temporal processing.

Chapter 4 uses stimulus-response models to explore the single-trial EEG correlates of selective auditory attention to speech in reverberant, multi-talker environments. This is investigated using simulations of real-world rooms presented to a cohort of normal-hearing listeners during EEG recordings and a selective attention task.

Chapter 5 examines how different constraints on the stimulus-response models influence the predictive power of the models. Specifically, it is explored how different regularization techniques influence the ability to map between envelopes of attended speech and continuous EEG responses using stimulus-response models. In this study, the model performance is evaluated both in terms of correlation coefficients between model predictions and target data, but also how robustly the models can discriminate between the attended and unattended speech.

Chapter 6 investigates whether working memory demands affect the EEG correlates of speech envelope processing when listening to speech embedded in noise. An auditory n-back working memory task with spoken digit streams is considered where listeners are asked to report whether each spoken digit was similar to the one presented n items earlier during EEG recordings.

Chapter 7 presents results from a BCI case study that investigates whether EEG-based attention decoding can be achieved in a real-time BCI system. Results from closed-loop and open-loop BCI experiments are presented for one older hearing-impaired listener performing an auditory attention-switching task. The subject is also interviewed to evaluate the subjective experience of closed-loop attention steering.

Chapter 8 summarizes the main findings of the present work and discusses their potential implications for BCI applications.

Appendix A describes challenges associated with the interpretation of results

from encoding- and decoding analyses.

2

Characterizing attention-driven M/EEG entrainment using stimulus-response models

Electroencephalography (EEG) and magnetoencephalography (MEG) have been widely used to study the neuroelectric and neuromagnetic correlates of auditory attention. Various studies have focused on effects of attention on M/EEG responses to discrete, repeated sounds (see e.g., Hillyard et al., 1973; Näätänen, 2018; Woldorff et al., 1993) or to periodically modulated sounds (Bidet-Caulet et al., 2007; Ross et al., 2004).

When using repetitive short stimuli, simple averaging procedures can be used to isolate stimulus-related neural activity that is time-locked to the discrete events. Such a procedure was employed by Hillyard et al. (1973), who recorded EEG activity from subjects listening to two simultaneous streams of tone pips presented dichotically with irregular inter-tone-intervals. The tone pips presented to the left-ear had higher carrier frequencies than those presented to the right-ear and both tone streams included randomly placed tone deviants. When instructing the listeners to count tone deviants in one ear and to ignore sound input from the other ear, the authors found that event-related potentials evoked by the attended tones elicited a larger N1 negativity (at a latency of around 100 ms) than those evoked by ignored tones. Since then, various studies have investigated how attention modulates the cortical processing of discrete, repeated sounds (Näätänen, 2018; Woldorff et al., 1993). However, considering the complexity of the sounds the auditory system is exposed to in natural environments, it is not clear whether effects observed with such discrete and repeated stimuli may generalize to more naturalistic listening scenarios (e.g., Theunissen and Elie, 2014).

More recently, researchers have considered M/EEG responses to complex naturalistic sounds, such as speech (Ahissar et al., 2001; Ding and Simon, 2012a; Luo and Poeppel, 2007; Mesgarani and Chang, 2012; O'Sullivan et al., 2014). One

way of identifying M/EEG response patterns related to the processing of complex sounds is to look for M/EEG response consistencies across repeated stimuli (Dmochowski et al., 2014; Hasson et al., 2004, 2012; Luo and Poeppel, 2007; de Cheveigné et al., 2018b). Another approach is to use statistical modelling techniques to identify stimulus- or task-related activity in M/EEG responses to sounds that are presented only once (Lalor and Foxe, 2010; Lalor et al., 2009b). This chapter reviews such modelling techniques that can provide insights into what is encoded in "single-trial" M/EEG responses to natural sounds. These model-driven approaches differ from other M/EEG analyses that focuses on only M/EEG response properties in the context of well-defined task- or stimulus manipulations. The model-driven approaches described here are typically referred to as *stimulus-response models* as they model the functional relationship between the sound stimulus and the neural response it elicits. These models have been found to be useful tools for studying how task-driven attention affects M/EEG responses to sound mixtures.

A stimulus-response relationship can be analyzed with encoding or decoding models. Whereas an encoding model attempts to predict a neural response from the stimulus input, the decoding model reverses the problem and attempts to infer information about the stimulus (or task) from the neural response. These are related modelling approaches, but with important differences (Haufe et al., 2014; Holdgraf et al., 2017; Kriegeskorte, 2011; Mesgarani et al., 2009; Naselaris et al., 2011; Weichwald et al., 2015). The goal of both modelling approaches is to develop predictive models, i.e. models that are capable of predicting novel data. The following subsections provide a brief introduction to encoding and decoding models, with emphasis on models that provide a map between sound features and neural responses.

2.1 Encoding models

A standard technique in auditory electrophysiology is to use receptive field models to characterize the selectivity of single neurons to spectrotemporal features (e.g., Aertsen and Johannesma, 1981; Theunissen et al., 2001). These models have been used to characterize the responses of auditory neurons to sounds as a linear transformation of the corresponding stimulus spectrogram (Aertsen and Johannesma, 1981; Atencio et al., 2008; Fritz et al., 2003; Woolley et al., 2005). The kernel that describes the mapping from the sound spectrogram to

the neural response has in such cases been referred to as the spectro-temporal receptive field (STRF), and it characterizes how changes in the time-frequency characteristics of the sound spectrogram are expected to affect the neural response.

It has recently been demonstrated that model-based approaches similar to receptive field models are relevant for the analysis of M/EEG data (Burns et al., 2013; Lalor and Foxe, 2010; Lalor et al., 2009b). M/EEG responses evoked (or induced) by repetitive transient sounds typically have distinct shapes with components that can be grouped by their post-onset latencies. This has led to the idea that averaged M/EEG responses to repeated transient sounds may reflect a convolution between standardized "unitary response functions" and stimulus-driven discharge patterns (Dau, 2003; Goldstein Jr and Kiang, 1958; Rønne et al., 2012). This has inspired researches to use systems engineering approaches to characterize how features of continuous synthetic sounds relate to M/EEG responses (Lalor and Foxe, 2010; Lalor et al., 2009b). Intriguingly, it has been shown that similar methods can be applied to more complex stimuli, such as speech, to characterize how speech features relate to M/EEG responses in humans, and how their representations in M/EEG responses are affected by selective attention. The models have in this context been referred to as "encoding models" or "forward models". There exists a variety of different encoding model formulations that provide a mapping from the sound input to the M/EEG output. A linearized M/EEG encoding model, operating on a stimulus feature representation $S(t,n)$, provides a forward map from $S(t,n)$ to an M/EEG response $R(t,n)$ in the following way:

$$\hat{R}(t,n) = \sum_{f}^{F} \sum_{i}^{M} S(t-\tau_i, f) H_n(\tau_i, f), \tag{2.1}$$

where $\hat{R}(t,n)$ is the estimated M/EEG response recorded at channel/sensor $n = 1,...,N$ at time $t = 1,...,T$, $S(t,f)$ is a representation of the time-varying stimulus, $\tau = \{\tau_1, \tau_2, ...\tau_M\}$ are the time-lags of interest and H_n contains the linear filter coefficients (that would correspond to the STRF in the context of receptive field models). Eq. 2.1 predicts the response $R(t,n)$ at each channel/sensor n independently.

2.2 Decoding models

In contrast to encoding models, decoding models directly learn features in the multi-dimensional M/EEG data (e.g. spatial patterns) to infer information about the stimulus (or the task). Depending on the problem at hand, one usually distinguishes between different types of decoding models. When inferring about discrete variables, the decoding problem is a classification problem. Another type of decoding approach that has been widely used in electrophysiology is "stimulus reconstruction", which treats the neural decoding problem as a regression problem rather than a classification problem. In this case, the desired output of the decoder is a continuous variable with one or more dimensions. The goal of stimulus reconstruction is to accurately reconstruct features in the sound input from the neural responses (Mesgarani and Chang, 2012; Mesgarani et al., 2009; Naselaris et al., 2009). As for the encoding models, these models can both be linear or nonlinear. A linearized M/EEG stimulus reconstruction model that attempts to reconstruct a sound feature $S(t,f)$ from a multi-dimensional M/EEG response, $R(t,n)$, to a sound stimulus takes the form:

$$\hat{S}(t,f) = \sum_{n}^{N} \sum_{i}^{M} R(t-\tau_i, n) H_f(\tau_i, n), \tag{2.2}$$

Here, $\hat{S}(t,f)$ is the neural reconstruction that estimates $S(t,f)$ and H_f is the kernel that maps neural response back to the features of the sound input. This model reconstructs $S(t,f)$ independently at each dimension $f = 1,...,F$.

The linearized stimulus reconstruction models offer a useful tool for visualizing, in the stimulus feature domain, how different task contexts may "shape" the neural reconstructions. In a study, Mesgarani and Chang (2012) trained a stimulus reconstruction model that was capable of reconstructing spectrograms of attended speech from electrode responses to single-talker stimuli recorded directly from the surface of the temporal lope of epileptic patients. Next, the authors instructed listeners to listen selectively to one of two competing talkers. It was found that the stimulus reconstruction models trained on the data from single-talker scenarios could robustly reconstruct spectrograms of attended speech in two-talker listening conditions, suggesting a clear and decodable effect of attention on the neural reconstructions. In this way, the reconstruction model could, in the stimulus feature domain, demonstrate how task-driven attention may affect the neural responses.

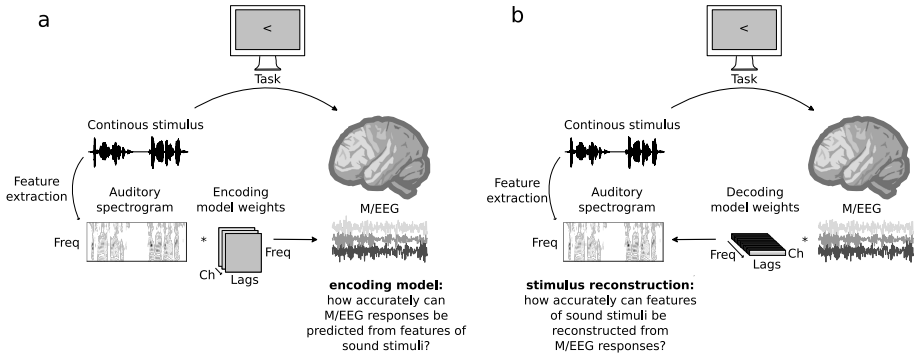

Figure 2.1: Using stimulus-response models to ask questions about what stimulus-related information that is represented in neural data under different listening tasks. The encoding models attempt to explain variance in the neural responses from stimulus or task-related information. In contrast, the stimulus-reconstruction models reverses the problem an attempt to infer information about stimulus from the neural responses.

2.3 Estimating model parameters and model validation

The goal of predictive encoding and decoding models is to find a set of model parameters, H, that provide a robust stimulus-response mapping between $S(t,f)$ and $R(t,n)$ for a given model class that may be linear or nonlinear. To do so, it is necessary to define a loss function that can characterize the lack of fit between the model predictions and the actual target variable (Wu et al., 2006). Since $S(t,f)$ and $R(t,n)$ will have different statistical properties, there may be some loss functions that are better suited for encoding analyses than for decoding analyses and vice versa. A commonly applied loss function is a square loss. During learning, the objective is to find a set of parameters that minimize the loss function. It is, for example, possible to use ordinary least squares (OLS) to estimate a set of filter parameters that minimizes the residual sum of squares between the model predictions and target data over the training set. To see this, it is convenient to express the models in Eq. 2.1 and Eq. 2.2 in matrix form:

$$\begin{aligned} \widehat{Y} &= \tilde{X} W_{\text{encoder}} \\ \widehat{X} &= \tilde{Y} W_{\text{decoder}} \end{aligned} \tag{2.3}$$

where $Y \in \mathbb{R}^{T \times N}$ is a matrix that contains the multi-dimensional neural response, $\tilde{Y} \in \mathbb{R}^{T \times N \cdot M}$ is a matrix that contains Y augmented to include multiple time delays $\{\tau_1, \tau_2, \ldots \tau_M\}$ and W_{decoder} contains the linearized decoding filter

coefficients (Mesgarani et al., 2009). Similarly, $X \in \mathbb{R}^{T \times F}$ is a matrix that contains the stimulus feature, $\tilde{X} \in \mathbb{R}^{T \times F \cdot M}$ is a matrix that contains X augmented to include multiple time lags and W_{encoder} contains the linearized encoding filter coefficients. Henceforth, it will be assumed that the data have been standardized. The OLS estimators in this case takes the form:

$$\begin{aligned} W_{\text{encoder, OLS}} &= \underset{W}{\operatorname{argmin}} \left[\left(Y - \tilde{X} W \right)^T \left(Y - \tilde{X} W \right) \right] \\ &= \left(\tilde{X}^T \tilde{X} \right) \tilde{X}^T Y \\ W_{\text{decoder, OLS}} &= \underset{W}{\operatorname{argmin}} \left[\left(X - \tilde{Y} W \right)^T \left(X - \tilde{Y} W \right) \right] \\ &= \left(\tilde{Y}^T \tilde{Y} \right) \tilde{Y}^T X, \end{aligned} \tag{2.4}$$

Here, it has (for the moment) been assumed that the regressor matrices in both cases have full rank. From these equations, it can be seen that the decoding model takes correlations among M/EEG channels over a finite range of lags into account when estimating X (due to the term $\tilde{Y}^T \tilde{Y}$), whereas the encoding model takes statistical regularities in the sound features into account (due to the term $\tilde{X}^T \tilde{X}$) (Holdgraf et al., 2017). This has consequences for both the stimulus-response mapping and for the interpretation of the model parameters (see Appendix A).

It is often necessary to impose regularization constraints on the model parameters by adding a complexity penalty to the loss function, which can stabilize the regression weights and prevent overfitting. This may be the case for linear regression models in scenarios where the regressor is a high-dimensional, low-rank matrix that has a poorly estimated covariance matrix given limited amounts of training data. How different regularization methods may affect the ability of linearized stimulus-response models to map between envelopes of attended speech and multi-channel EEG responses is explored in more details in Chapter 4 (from Wong et al., 2018).

A straightforward and effective way of seeking models that better generalize to novel data is by imposing an L_2 size constraint on the model coefficients, which yields the so-called Ridge estimator:

$$
\begin{aligned}
W_{\text{encoder, Ridge}} &= \underset{W}{\operatorname{argmin}}\left[\left(Y-\tilde{X}W\right)^T\left(Y-\tilde{X}W\right)+\lambda W^T W\right] \\
&= \left(\tilde{X}^T\tilde{X}+\lambda I\right)\tilde{X}^T Y \\
W_{\text{decoder, Ridge}} &= \underset{W}{\operatorname{argmin}}\left[\left(X-\tilde{Y}W\right)^T\left(X-\tilde{Y}W\right)+\lambda W^T W_d\right] \\
&= \left(\tilde{Y}^T\tilde{Y}+\lambda I\right)\tilde{Y}^T X,
\end{aligned}
\tag{2.5}
$$

where λ is the regularization parameter that defines the extend to which the model coefficients (smoothly) are shrunk to zero and to one another (Hastie et al., 2001). This is here achieved by smoothly shrinking low-variance directions in the regressor. In certain scenarios it can be desirable to promote sparsity in the solution. This can be achieved by imposing an L_1 (Lasso) penalty on the filter coefficients:

$$
\begin{aligned}
W_{\text{encoder, Lasso}} &= \underset{W}{\operatorname{argmin}}\left[\left(Y-\tilde{X}W\right)^T\left(Y-\tilde{X}W\right)+\lambda||W||_1\right] \\
W_{\text{decoder, Lasso}} &= \underset{W}{\operatorname{argmin}}\left[\left(X-\tilde{Y}W\right)^T\left(X-\tilde{Y}W\right)+\lambda||W||_1\right]
\end{aligned}
\tag{2.6}
$$

This is sometimes referred to as the Lasso estimator. Here, λ controls how many of the filter coefficients that are set to zero. The three aforementioned estimators will later be considered in an example (Appendix A).

Whenever the model has additional free parameters (e.g., regularization parameters, such as λ), the additional free parameters can be tuned by cross-validation to seek models with lowest generalization error. One way of tuning the regularization parameters is by splitting the data into a training set, a validation set and a test set, over which the training set is used to fit the models for different regularization parameters and the validation set is used to tune the regularization parameter. Finally, once a set of parameters has been selected, it is important to obtain unbiased estimates of the model performance. This can be done by evaluating how well the model can predict held-out data not used for model fitting or parameter tuning.

2.4 Factorized and parameterized models

To develop predictive stimulus-response models it can be necessary to impose constraints on model parameters. One way of efficiently reducing the space

encoding model parameters is by putting constraints on the model shape via factorizing and parameterizing (Kay et al., 2013; Thorson et al., 2015). For instance, for a given STRF model, $W_{\text{STRF}} \in \mathbb{R}^{F \times M}$, it is possible to approximate STRF as the outer product of a spectral filter, $W_s \in \mathbb{R}^{F \times 1}$, and a temporal filter, $W_t \in \mathbb{R}^{1 \times M}$. This reduces the number of free parameters in the model from $F \times M$ to $F + M$. To reduce the space of the model parameters even more, the spectral and temporal filters can be parameterized. Such a model parameterization will later be considered in an example in Chapter A. In this example, the STRF is approximated as the outer product of a spectral filter, $W_s = \begin{pmatrix} w_{\text{s}}(f_1) & \dots & w_{\text{s}}(f_F) \end{pmatrix}^T$ (parameterized by a single Gaussian function with a center frequency, f_{μ_0}, and a bandwidth, σ_{f_0}) and a temporal filter, $W_t = \begin{pmatrix} w_t(\tau_1) & \dots & w_t(\tau_M) \end{pmatrix}$ (parameterized by three Gaussian functions each centered around time lag τ_{μ_j} and with widths σ_{τ_j}, $j = 1,2,3$):

$$
\begin{aligned}
W_{\text{STRF}} &= W_{\text{s}} W_{\text{t}}, \qquad (2.7)\\
w_{\text{s}}(f_i) &= \frac{1}{\sigma_{f_0}\sqrt{2\pi}} \exp\left(-\frac{\left(f_i - f_{\mu_0}\right)^2}{2{\sigma_{f_0}}^2} \right)\\
w_{\text{t}}(\tau_i) &= \sum_{j=1}^{3} A_j \frac{1}{\sigma_{\tau_j}\sqrt{2\pi}} \exp\left(-\frac{\left(\tau_i - \tau_{\mu_j}\right)^2}{2{\sigma_{\tau_j}}^2} \right)
\end{aligned}
$$

Such model constraint can be viewed as a form of model regularization. The model parameters can for such models be estimated using nonlinear optimization algorithms.

2.5 Hybrid encoding-decoding models

Since $Y \in \mathbb{R}^{T \times N}$ and $X \in \mathbb{R}^{T \times F}$ both can be high dimensional and have complex statistical properties, multivariate models can be effective for learning transformations that can be applied to both Y and X to better reveal statistical stimulus-response dependencies. One way of doing so is by using canonical correlation analysis (CCA) (Biessmann et al., 2011; Dmochowski et al., 2017; de Cheveigné et al., 2018a). Assume for now that X contains a spectrogram representation of a stimulus and that Y contains a multi-channel EEG response. Similarly, assume that $\tilde{X}$ is a matrix that contains X augmented to include multiple time

delays. A multivariate CCA model can here simultaneously learn spectrotemporal filters $W_e = \begin{pmatrix} a_{e_1} & a_{e_2} & ... \end{pmatrix}$ and spatial EEG filters $W_d = \begin{pmatrix} b_{d_1} & b_{d_2} & ... \end{pmatrix}$ such that the correlation between $\tilde{X} a_{e_i}$ and $Y b_{d_i}$ is successively maximized for each set of filters $i = 1, 2, ...$ while simultaneously being uncorrelated with previous projections. This has been referred to as a hybrid encoding-decoding model as it simultaneously learns encoding filters and decoding filters. In the aforementioned example, the model coefficients can be estimated from:

$$\begin{aligned} \{W_{\text{e, CCA}}, W_{\text{d, CCA}}\} &= \underset{W_e, W_d}{\operatorname{argmax}} \left[\operatorname{tr}\left(W_e^T \tilde{X}^T Y W_d \right) \right] \\ \text{s.t. } & W_e^T \tilde{X}^T \tilde{X} W_e = I \\ & W_d^T Y^T Y W_d = I \end{aligned} \tag{2.8}$$

It can be shown that a solution to this problem can be found by solving a generalized eigenvalue problem (Biessmann et al., 2011; Borga, 1998). A simple formulation of CCA is presented in de Cheveigné et al., 2018b, which involves whitening each matrix, $\tilde{X}$ and Y (by applying principal component analysis and scaling each component to be unit norm), concatenating the whitened matrices and applying them to another PCA. One way of constraining the model parameters is to discard PCs with the lowest variance after the initial PCA. This can reduce the risk of overfitting. Such a CCA model will later be considered in an example in Appendix A. Moreover, Chapter 7 considers a CCA model for real-time auditory attention decoding. In this case, a dyadic filterbank is applied to EEG/audio data to put constraints on the spectral representations and reduce the number of model parameters.

2.6 Linearized models and linearization transforms

The relationship between sound pressure waves and neural activity is nonlinear due to the various nonlinear processing stages along the auditory pathway. When considering stimulus-response models that are linear in their coefficients, it is therefore necessary to transform the sound stimulus into a stimulus representation $S(t, f)$ that is assumed to be linearly related to the response (Aertsen and Johannesma, 1981; Wu et al., 2006). When the transformation from the sound stimulus to the stimulus representation is nonlinear, the encoding and de-

coding models described in Eq. 2.1 and 2.2 are both considered to be *linearized* stimulus-responds models (Naselaris et al., 2011). The nonlinear transformations of sound stimuli (or feature extraction schemes) are often motivated by *a priori* knowledge about what sound features that may drive (or be reflected in) the neural responses (Broderick et al., 2018; Di Liberto et al., 2015; Ding and Simon, 2012a), but they can also be learned using data-driven approaches (Kell et al., 2018; Yamins and DiCarlo, 2016).

3

The influence of a sensorineural hearing loss on cortical auditory EEG entrainment during selective listening[a]

Abstract

Listeners with age-related sensorineural hearing loss (presbycusis) often struggle to follow speech in noisy environments, even when reduced audibility has been compensated for, e.g., by a hearing aid. Selectively attending to a speech stream in a competing talker scenario is known to entrain low-frequency cortical activity (<10 Hz) to fluctuations in the attended speech signal in normal-hearing listeners, but it is unclear if and in which way a peripheral hearing loss may affect such entrainment. Here, we used psychoacoustics and electroencephalography (EEG) to examine effects of sensorineural hearing loss on EEG correlates of attention-driven speech envelope entrainment in cortex. Behaviorally, presbycutic hearing-impaired (HI) listeners showed degraded speech-in-noise sentence reception thresholds and a reduced temporal acuity compared to age-matched normal-hearing (NH) controls. During EEG recordings, we used a selective attention task with spatially separated speech streams where both NH and HI listener groups showed equally good speech recognition. Cortical envelope entrainment was found to be enhanced in HI listeners, both for the attended speech stream, but also for speech-in-quiet and periodic tone sequences in passive listening conditions. Suppression of the ignored stream relative to the attended one was found to be similar in the

[a] This chapter is based on: Fuglsang, S. A.; Marcher-Rørsted, J.; Dau, T.; Hjortkjær, J. H (in preparation). "The influence of a sensorineural hearing loss on cortical auditory EEG entrainment during selective listening".

two groups, allowing for the attended talker to be classified from the EEG data with equal high classification accuracy. Despite the robust attention-modulated entrainment, the HI-listeners rated the competing speech task to be more challenging. Overall, while degraded temporal processing and an enhanced envelope representation may degrade selective listening at more challenging signal-to-noise ratios, HI listeners can engage attention to modulate cortical speech envelope entrainment responses when perceptual segregation of the speech signal is intact.

3.1 Introduction

Following the voice of a single speaker among competing speakers can be challenging for listeners with a sensorineural hearing loss (SNHL). It has been argued that a SNHL may interfere with selective auditory attention and cause deficits in the ability to ignore distracting sound sources (Dai et al., 2018; Petersen et al., 2016; Shinn-Cunningham and Best, 2008; Shinn-Cunningham, 2017). Physiological studies have demonstrated that a SNHL may, in fact, enhance subcortical and cortical envelope coding (Kale and Heinz, 2010, 2012; Millman et al., 2017; Zhong et al., 2014). It is conceivable that enhanced neural envelope coding in hearing-impaired (HI) listeners exaggerates perceived envelope fluctuations in amplitude modulated sounds (e.g., Moore et al., 1996; Wiinberg et al., 2018), but may conversely disrupt the segregation of speech from background noise (Millman et al., 2017; Moore and Glasberg, 1993). Although cortical envelope entrainment responses have been found to be modulated by attention in normal-hearing (NH) listeners, it is less clear how attention-modulated cortical auditory entrainment responses are affected by a SNHL.

Enhanced neural envelope coding in the aging auditory system has previously been reported (Goossens et al., 2016; Parthasarathy et al., 2019; Presacco et al., 2016a,b; Walton et al., 2002). SNHL may be more prevalent in older listeners (Agrawal et al., 2008; Cruickshanks et al., 1998; Davis, 1995), and older listeners can show degraded speech perception despite normal audiometric thresholds (Dubno et al., 1984; Füllgrabe and Rosen, 2016). This may be ascribed to age-related changes in central- and cognitive factors (Caspary et al., 2008; Frisina and Frisina, 1997; Füllgrabe and Rosen, 2016; Henry et al., 2017; Krull et al., 2013; Peelle et al., 2009; Pichora-Fuller, 2003; Presacco et al., 2016a,b), or

to age-related peripheral deficits (Sergeyenko et al., 2013). The prevalence of peripheral deficits may increase with age (Sergeyenko et al., 2013), and damages to the auditory periphery may affect sound-evoked cortical responses, but may have little impact on audiometric thresholds (Chambers et al., 2016). Age-related peripheral deficits and age-related changes in central- or cognitive factors can thus be difficult to dissociate in studies on cortical auditory processing (Humes et al., 2012). To date, it is not clear whether a SNHL further contributes to already exaggerated speech-evoked cortical responses in older listeners during selective listening (Presacco et al., 2016a,b).

The present study sought to further elucidate how a SNHL impacts cortical auditory envelope entrainment responses when listening selectively to one of two competing speakers. In an attempt to better separate age-related effects from SNHL-related effects, we recruited age-matched HI and NH listeners. We employed a selective auditory attention EEG experiment with two spatially separated speech streams that were loudness matched. Using stimulus-response models to assess relations between the envelopes of individual speech streams and low-frequency EEG responses, we show that the fidelity of the low-frequency cortical EEG entrainment to envelopes of attended speech is enhanced in the HI listener group, and that both listener groups show attentional modulations of the EEG envelope entrainment responses in two-talker listening conditions. In addition, we show that an SNHL enhances the fidelity of the low-frequency EEG phase-locking to envelopes of tone sequences unfolding on fast (gamma range) and slow (theta range) time scales. The findings provide insights into the effects of SNHL on attention-driven cortical envelope entrainment and suggest that peripheral deficits do not *per se* hinder attentional modulations of cortical auditory entrainment responses.

3.2 Materials and methods

3.2.1 Participants and audiometry

Forty-five subjects participated in this study. Hearing-impaired (HI, N = 22, 9 females, 21 right handed) and normal-hearing (NH, $N = 23$, 16 females, 19 right handed) subjects between 51 and 76 years of age participated. The HI and NH groups were matched in age ($t(41.982) = 1.5947$, $p = 0.1183$; NH: mean age 63.14 ± 7.26; HI: mean age 66.59 ± 7.11). The HI subjects were selected to have a

symmetric steeply sloping high-frequency hearing loss indicating presbyacusis (Bisgaard et al., 2010). For the NH subjects, the inclusion criterion was audiometric thresholds within 20 dB of normal hearing level (HL) at frequencies up to 2 kHz and within 35 dB HL at frequencies above 2 kHz. All selected subjects had symmetric audiograms, defined as less than 15 dB difference between ears at two or more neighbouring frequencies. Bone-conduction thresholds were measured at 0.5, 1 and 2 kHz. Subjects with air-bone gaps above 10 dB at any audiometric frequency were excluded from the analysis. Tympanometry and otoscopy screening was used to assure normal middle- and outer ear function. One of the NH subjects was excluded from the analysis because EEG data could not be obtained.

All subjects provided written informed consent to participate. The experiment was approved by the Science Ethics Committee for the Capital Region of Denmark (protocol no. H-16036391) and was conducted in accordance with the Declaration of Helsinki.

3.2.2 Behavioral hearing profiles

Speech perception in noise

A Danish hearing-in-noise test (DaHINT, Nielsen and Dau, 2009) was used to estimate sentence-reception thresholds (SRT). Listeners were presented with sentences in stationary noise spectrally matched to the speech and asked to repeat the sentence. The audio stimuli were presented diotically using Sennheiser HD650 headphones at a fixed noise level of 65 dB sound pressure level (SPL) in a double-walled sound booth. The level of the speech signal was varied adaptively to identify speech reception thresholds for each subject. SRTs indicate the signal-to-noise ratio (SNR) at which the listeners correctly recognizes 50% of the presented sentences. The speech reception thresholds were averaged across three repetitions of a list with 20 sentences.

Temporal processing abilities

A psychoacoustic tone-in-noise detection test (adapted from Larsby and Arlinger, 1999) was used to assess temporal processing acuity. A pulsating pure tone (500 Hz, 275 msec duration, 2.22 pulses/sec) was presented in different background noise conditions. First, the threshold for tone detection in wideband noise (with a passband corresponding to six equivalent rectangular band-

width (Moore, 1986) around the target tone frequency) was measured. Next, a temporal gap of 50 ms centered around the tone was introduced. The temporal masking release, i.e. the difference in detection thresholds between the no-gap and gap conditions, was then calculated as a measure of listeners abilities to utilize temporal fluctuations in the noise masker for improved detection. The noise was presented at a fixed level of 55 dB SPL and the level of the target tone was varied using a Békésy tracking procedure to identify the thresholds. The subjects performed each condition (no gap, temporal gap) twice for each ear. The stimuli were presented using Sennheizer HDA200 headphones in a double-walled sound booth.

Working memory performance

Speech perception in noise by older listeners may not only depend on their hearing status but also on cognitive abilities (Akeroyd, 2008). Moreover, hearing impairment may itself affect cognitive function (Wingfield and Peelle, 2012). To ensure that the recruited older NH and HI subjects were matched in cognitive abilities, a digit-span test was used to measure working memory performance. In the test, listeners were asked to recall a presented sequence of numbers (between 1 and 9) in reverse order. The digit span score was then calculated as the number of items that could be repeated correctly (Blackburn and Benton, 1957). The auditory stimuli were presented via Sennheiser HD650 headphones at a comfortable level.

Self-evaluated hearing disability

All subjects completed the Speech, Spatial and Qualities of Hearing Scale questionnaire (SSQ, Gatehouse and Noble, 2004)). The SSQ questionnaire consists of 49 questions related to self-rated hearing abilities in everyday situations. Hearing aid users are asked to rate their listening abilities when wearing their hearing aid. The questions address hearing in three domains: ”Speech” (e.g., comprehending speech and selectively attending to a particular talker in everyday listening situations), ”Spatial” (e.g., judging direction, distance and movement of sound sources), ”Qualities” (e.g., segregation of sound sources, clarity and listening effort).

3.2.3 Accounting for reduced audibility

The auditory stimuli in the EEG experiments and DaHINT test were amplified based on audiometric thresholds to account for reduced audibility in the HI listeners. A linear gain was applied per audiometric frequency using the 'Cambridge formula' (Moore and Glasberg, 1998) and was limited to 30 dB SPL gain at any given frequency. The level of the speech stimuli used in the selective attention EEG experiment was 65 dB SPL before the frequency-dependent amplification. In the DaHINT test the level of the speech signal was varied adaptively. For tone stimuli used in the EEG experiments, the stimuli were adjusted to comfortable level per subject.

3.2.4 EEG experiments

The EEG experiments were performed in an electrically shielded double-walled sound booth. In all EEG experiments, the subjects were comfortably seated and instructed to fixate their eye-gaze at a cross hair presented on a computer screen. EEG data were recorded using a BioSemi ActiveTwo system with 64-scalp electrodes positioned according to the 10-20 system. Two additional bipolar electrooculography electrodes were mounted above and below the left eye. The EEG data were digitized at a sampling rate of 512 Hz. For 19 of the 44 subjects (10 NH and 9 HI), EEG was also measured inside the ear canals, but this data were not included in the analysis. The auditory stimuli were presented via ER-3 insert earphones (Etymotic Research).

Envelope-following responses with tone stimulation

Envelope-following responses (EFRs) were recorded from subjects listening passively to multi-scale periodic tone sequences designed to induce activity in the gamma (40 Hz) and theta frequency ranges. Two types of stimulation paradigms were used. In both paradigms, 1 kHz tone pulses (10 ms Hann-shaped ramps) with an inter-pulse-interval of 40 Hz were presented in epochs of 2 s stimulation, alternating with 1 s silence. In the first stimulation paradigm, 0.5-s long 40 Hz tone sequences alternated with 0.5-s long silence intervals. This resulted in a periodic 4 Hz onset/offset pattern to induce 4 Hz theta activity. In the second paradigm, no 4 Hz onset/offset pattern was imposed. In each of the two stimulations, 60 3-s long epochs were presented. The stimuli presented are illustrated in Figure 3.9 in the Supplementary Material.

Event-related potentials

Event related potentials (ERPs) were recorded from subjects listening passively to 1 kHz pure tones. The tone stimuli had duration of 100 ms and were ramped using a 10 ms long Hann windows. The tones were presented at an average inter-tone-interval rate of 1 Hz that was randomly jittered +/-25 ms. Each subject listened to 180 tone repetitions.

Selective speech attention experiment

Continuous EEG were recorded from subjects selectively listening to one of two competing speech streams or to speech in quiet. The speech stimuli consisted of two different audiobooks read aloud by a male and a female speaker. Silent periods in the sound stimuli exceeding 450 ms were truncated to 450 ms. The audio files were split into 50 s long trials. The speech streams were spatially separated at $\pm 90^o$ using non-individualized head related transfer functions (HRTF) provided by (Oreinos and Buchholz, 2013). The audio files were low pass filtered at 12 kHz using a 2nd order Butterworth filter to avoid excessive high-frequency amplification for subjects with low audiometric thresholds. The audio of the two talkers was matched in loudness before spatialization according to the current ITU standard (ITU-R BS.1770-1). The speech stimuli were presented at 65 dB SL.

Figure 3.1 presents a schematic of the trial structure of the selective listening experiment. In 50 s long trials, the subjects listened to either a single talker or two competing talkers. The auditory attention experiment consisted of 48 trials. Each subject listened to two blocks of 12 trials with the male speaker as the target, and two blocks of 12 trials with the female speaker as the target. Each block of 12 trials consisted of four single-talker trials, and eight two-talker trials. At the onset of each trial, the subject was instructed to attend to either the male or the female talker. As an additional cue, the target speech stream was switched on ∼4 seconds before the interfering speech stream. The EEG data recorded in this period were discarded from analysis. The number of left versus right targets trials was balanced over the experiment. After each trial, the subjects were asked to rate how easy or difficult it was to understand the attended speech on an analogue rating scale marked 'easy' and 'difficult' at the extremes. After the rating, subjects were prompted to answer four multiple-choice comprehension questions related to the content of the attended speech stream. The first of the

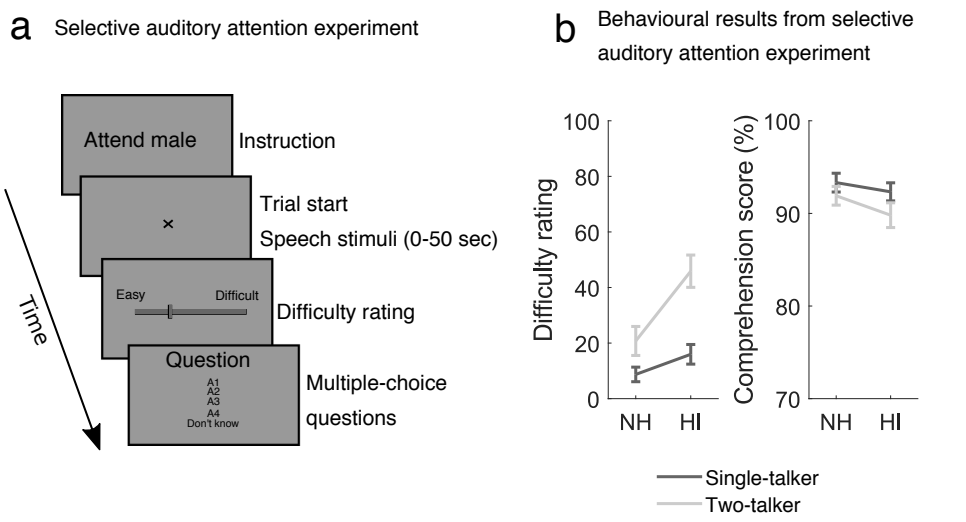

Figure 3.1: Auditory attention EEG experiment. (a) Schematic illustration of the trial sequences of the selective auditory attention EEG experiment. EEG data were recorded from subjects selectively listening to 50 s long excerpts of audiobooks narrated by a male and female speaker. In single-talker trials (i.e., trials with only one speaker), the masker was not presented. After each trial, the subjects were asked to rate task difficulty and respond to multiple-choice questions related to the content of the attended speech stream. Each subject responded to 192 multiple-choice questionnaries. (b) The behavioural results from the selective auditory attention experiment. (b, left): Difficulty rating scores, (b, right): Comprehension scores. Data shown are mean and s.e.m.

four comprehension questions was also shown before the trial started. Subjects were given feedback on their responses.

3.2.5 Data analysis

EFR and ERP data preprocessing

The EEG data were de-trended using a 1^{st} order polynomial. The ERP data and EFR data were high-pass filtered at 0.1 Hz (Rousselet, 2012; VanRullen, 2011), using a windowed sinc type I linear phase finite impulse response filter with delay compensation, and thereafter low-pass filtered at 30 Hz. The EFR data were low-pass filtered at 60 Hz. For both EFR and ERP data, we shifted the data to adjust for filter delays. The data were filtered prior to epoch segmentation. Both data sets were re-referenced relative to the average response of the electrodes TP7 and TP8. Responses from "bad electrodes" were visually identified and interpolated using template-based nearest neighbourhood averaging as implemented in Fieldtrip. On average, 1.7273 ± 1.4685 electrodes were removed. Line noise

was removed via notch filtering. An automized joint decorrelation framework (de Cheveigné and Parra, 2014) was employed to remove electroocular (EOG) artefacts for the analysis of ERPs (see sect. 3.2.5).

ERP analysis

For the ERP data, we examined the mean amplitude of the N1 component (Clayson et al., 2013) in the time window from 75 to 125 ms. The ERP data were averaged over a subset of 14 fronto-central electrodes over which the N1 amplitude was prominent (FC5, FC3, FC1, FCz, Fz, FC2, FC4, FC6, F5, F3, F1, F2, F4, F6; see e.g. Fig. 3.6 in the Supplementary Material for topographical depiction of the N1 mean amplitude). The data from one subject had to be excluded from the analysis of ERPs elicited by the tone pulses due to excessive artefacts.

For the EFR data, we also examined the mean amplitude of the N1 related to the onset of the auditory stimulation. This was done to better understand whether potential differences in the inter-trial phase coherence (ITPC) across NH and HI subjects (see the next subsection) were driven by N1 amplitude differences across the groups (Diepen and Mazaheri, 2018). This analysis was carried out in the same way as for the ERP data. For this analysis, the EFR data were also low-pass filtered at 30 Hz (adjusting for filter delay). The same electrode cluster and the same time averaging window as for the ERP data were here considered.

EFR inter-trial phase coherence

The inter trial phase coherence (ITPC) was computed for EEG responses to EFR stimuli. We performed a time-frequency decomposition of each electrode response by convolving the EEG responses with complex Morlet wavelets with a fixed number of twelve cycles per wavelet (Larsen et al., 2018), as implemented in the Fieldtrip toolbox (Oostenveld et al., 2011). No spatial filtering was performed prior to the analysis. With f_0 representing the bandpass center frequency of each Morlet wavelet, we considered an f_0 range between 1 Hz and 50 Hz with a resolution of 0.01 Hz. The complex output, $F_k(f,t,n)$, for trial $k = 1,...,N$, electrode n, time bin t and center frequency f was then used to compute the

ITPC:

$$\mathrm{ITPC}(f,t,n) = \frac{1}{N}\sum_{k=1}^{N}\frac{F_k(f,t,n)}{|F_k(f,t,n)|} \tag{3.1}$$

The ITPC ranges between 0 and 1, and indicates the degree of phase consistency of EEG responses to the EFR stimuli over trials (where 0 corresponds to no consistency and 1 indicates full consistency).

For the statistical analysis of the EFR data, the average of all scalp electrodes was considered, due to topographical differences in EFR responses at slow- and fast rates (see e.g. Fig. 3.6 in the Supplementary Material).

Speech EEG data

The EEG data from the speech-listening experiment were de-trended using a 1^{st} order polynomial and high pass filtered at 0.5 Hz, using a windowed sinc type I linear phase finite impulse response filter with delay compensation. Line noise was removed via notch filtering. The EEG data were re-referenced to the average response of the two posterior temporal electrodes (TP7 and TP8) that exhibited stable electrode connections over all trial and subjects. "Bad electrodes" were removed according to the procedure described above. The data were then low-pass filtered at 40 Hz with a 10^{th}-order low-pass finite impulse response (FIR) filter and time-shifted to account for the delay introduced by the filter.

A joint decorrelation (JD) framework was employed to remove EOG artifacts (de Cheveigné and Parra, 2014) in similar way as described in Wong et al., 2018. Data segments containing EOG artifacts were detected using the Hilbert envelope of EOG channel responses and responses over three frontal electrodes, Fp1, Fpz and Fp2. The Hilbert envelope of each of these channel responses was bandpass filtered (1-30 Hz) and z-scored. Time points where either of the resulting signals exceeded a threshold of four were considered artefactual. The artefactual segments were extended by 0.1 s on both sides. This implementation was similar to that implemented in the Fieldtrip toolbox (Oostenveld et al., 2011). The labelled segments were then used to compute an artefact biased covariance matrix. The estimated artefact biased covariance matrix and the covariance matrix estimated from the entire dataset were whitened (based on a principal component analysis). Eigenvectors characterizing the maximum variance differences between the two covariance matrices was then computed (Wong et al., 2018; de Cheveigné and Parra, 2014). Eigenvectors with eigenvalues

larger than >80 % of the maximum eigenvalue were subsequently regressed out from the data (Wong et al., 2018). Next, the data were downsampled to 64 Hz. The EEG data were finally filtered between 1 and 9 Hz. This was done by first applying a 64^{th}-order FIR high pass filter with a 1 Hz cut-off and then low-pass filtering the data with a 64^{th}-order FIR filter with a 9 Hz cut-off. The data were shifted to account for the filter delay.

3.2.6 Speech envelope entrainment analyses

Speech envelope extraction

The speech envelopes of the attended and unattended speech streams were extracted using a simplistic functional model of the auditory periphery. Since the audio stimuli were spatially separated, we extracted both the attended and unattended speech signals and collapsed the sum across spatial dimension to obtain audio waveforms. The monaural audio waveforms were passed through a gammatone filterbank (Patterson et al., 1987) consisting of 24 4th-order gammatone bandpass filters with center frequencies on an equivalent rectangular bandwidth scale (ranging between 100 Hz and 4000 Hz) (Glasberg and Moore, 1990) and 0 dB attenuation at their individual center frequencies. The output from each gammatone filter was power-law compressed, $\text{sgn}(x)|x|^c$, with a compressive factor of $c = 0.3$ (Ruggero, 1992). The compressed subband signals were subsequently full-wave rectified, low-pass filtered with a low-order anti-aliasing filter and resampled to match the EEG sampling rate. Next, we averaged the output subband envelopes across gammatone frequency channels to obtain a univariate temporal envelope for each speech stream. Finally, we band-pass filtered the envelope between 1 Hz and 9 Hz. This was done by first applying a 64^{th}-order FIR high pass filter with a 1 Hz cut-off and then low-pass filtering the data with a 64^{th}-order FIR filter with a 9 Hz cut-off. The data were shifted to account for the filter delay and downsampled to 64 Hz. For the HI subjects, we found a high correlation between the temporal envelopes of the amplified (compensated) audio stimuli and those of the uncompensated audio files. We achieved comparable results in the stimulus-responses analyses with both types of envelope representations, and therefore only report results obtained with the envelopes of the uncompensated stimuli.

Stimulus-response analyses

Following a number of previous speech-attention studies (Ding and Simon, 2012a, 2013), we considered two complementary analyses of statistical stimulus-response dependencies between the envelope of the attended and unattended speech streams and the EEG responses. We considered both forward regression models, i.e., encoding models, and backward regression models, i.e., decoding models. The encoding models attempt to predict neural responses to speech stimuli, $R(t,n)$, from fluctuations in the speech envelopes $S(t)$.

$$\hat{R}(t,n) = \sum_{k=1}^{K} S(t-\tau_k) w(\tau_k, n) \tag{3.2}$$

where $\hat{R}(t,n)$ is an estimate of the EEG response at a given electrode, $n = 1,2,...N$. For the encoding analysis, we considered time lags, $\tau_k = \{\tau_1, \tau_2, ...\tau_K\}$, ranging between 0 ms and 500 ms post-stimulus. The regression weights $w(\tau_k, n)$ define a temporal response function (TRF) or temporal kernel that predicts the EEG response from a weighted sum of the time-lagged speech envelope amplitudes.

The backward decoding model integrates information over all EEG electrodes and all time-lags to reconstruct the speech envelope:

$$\hat{S}(t) = \sum_{n=1}^{N} \sum_{k=1}^{K} R(t-\tau_k, n) w(\tau_k, n) \tag{3.3}$$

For the decoding analysis, we considered time-lags $\tau_k = \{\tau_1, \tau_2, ...\tau_K\}$ ranging between −500 ms and 0 ms post-stimulus. For both encoding and decoding models, we only included data from 6 s after trial onset (i.e. after the onset of masking stimulus) to 43 s after trial onset.

The weights of the linear regression models were estimated via ridge regression. Let $\mathbf{X}$ be a standardized matrix and let $\mathbf{X} = \mathbf{UDV}^T$ be the singular-value decomposition of $\mathbf{X}$. Similarly, let $\mathbf{Y}$ be a vector with zero mean and unit standard deviation. The linear regression model can now be formulated as:

$$\hat{\mathbf{Y}} = \mathbf{Xw}, \tag{3.4}$$

where $\hat{\mathbf{Y}}$ is an estimate of $\mathbf{Y}$. The Ridge regression estimator then takes the form:

$$\begin{aligned}\mathbf{w} &= \underset{\mathbf{w}}{\operatorname{argmin}}\left[(\mathbf{Y}-\mathbf{X}\mathbf{w})^T(\mathbf{Y}-\mathbf{X}\mathbf{w})+\lambda\mathbf{w}^T\mathbf{w}\right] \\ &= \left(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{V}\left(\mathbf{D}^2+\lambda\mathbf{I}\right)^{-1}\mathbf{D}\mathbf{U}^T\mathbf{Y},\end{aligned} \tag{3.5}$$

In the case of a forward encoding model, $\mathbf{X}$ is a matrix containing speech envelope, $S(t)$, at multiple time lags and $\mathbf{Y}$ is the EEG response at a given channel. In this case, separate Ridge parameters are estimated for each electrode, each subject and each experimental condition. In the case of a backward model, $\mathbf{X}$ is a matrix containing the multi-channel and time-lagged EEG response and $\mathbf{Y}$ is the speech envelope.

To assess the predictive performance of each model we used a nested cross-validation procedure. The nested cross-validation procedure consisted of an outer 10-fold cross-validation loop and an inner 5-fold cross-validation loop. The data was split 10 times into a training set and a test set, and for each split we further divided the training data randomly into five equal parts to optimize the Ridge λ parameter. In this way, the Ridge parameter was optimized over the training set and the generalization error was evaluated over the held out test set. We considered 50 ridge parameters, λ, logarithmically spaced between 10^{-3} and 10^{6}. During model fitting and evaluation, the data were standardized to the empirical mean and standard deviation of the data used for model fitting. The prediction accuracy was indexed by the Pearson's correlation coefficient between the model prediction $\hat{\mathbf{Y}}$ and the target data, $\mathbf{Y}$. For each outer fold, the Ridge parameter that yielded the highest prediction accuracies over the inner loop was chosen as optimal. Pearson's correlation coefficient was chosen as the metric since it is bound between -1 and $+1$, is invariant to scaling and shift errors in the predictions, and since it has been successfully used in the context of M/EEG-based attention decoding (Ding and Simon, 2012a; O'Sullivan et al., 2014). To make our results readily comparable to previous studies, we did not attempt to normalize the correlation metric to account for the inherent variability in neural responses across repeated sounds. The prediction accuracy was estimated as the average over the 10 initial splits of the data. The performance of the stimulus-response models was in all cases evaluated on data from trials that had not been used for model fitting.

For the statistical analysis of results from encoding analyses, we averaged the encoding accuracies over the same subset of fronto-central electrodes as in the

ERP analysis (FC5, FC3, FC1, FCz, Fz, FC2, FC4, FC6, F5, F3, F1, F2, F4, F6) (see Fig. 3.6 in Supplementary Material). The noise floor was estimated as in Wong et al., 2018 by performing the same analyses but with the target regressors being phase scrambled (phase scrambling EEG responses in the forward modelling case).

We additionally sought to investigate whether backward models could be used to decode attention, i.e. classify who the listener was attending to based on single-trial EEG data. The ability to discriminate between attended and unattended speech envelopes may provide a measure of how robustly the EEG entrainment responses are modulated by attention. For this analysis, we tuned the Ridge parameter to give highest leave-one-trial-out prediction accuracies. Moreover, in this case we optimized the EOG denoising filters based only on the single-talker training data. Once the stimulus reconstruction models had been trained, we used EEG data from the remaining 32 two-talker trials to reconstruct the speech envelopes of the attended talker. We considered decoding segments that were 10 s long (taking into account the 0.5 s long kernel of the stimulus reconstruction models). The decoding segments were non-overlapping, and each decoding segment was shifted by 5 s long time shifts. We computed the Pearson's correlation coefficient between the reconstructed envelopes and the envelopes of the individual speech streams (henceforth denoted r_{attended} and $r_{\text{unattended}}$). We considered a classification to be correct whenever the neural reconstruction was more correlated with the envelope of the actual attended speech stream than with the envelope of the unattended speech stream (i.e., $r_{\text{attended}} > r_{\text{unattended}}$). The chance-level of classification performance was here assessed with a binomial distribution.

3.2.7 Statistical tests

Repeated measures analysis of variances (ANOVAs) were used to analyze results from the stimulus-response analyses in the single-talker and the two-talker conditions in both listener groups. Moreover, repeated measures ANOVAs were used to analyze average ITPC results over short time windows in both groups of subjects. Welch's t-tests were used to compare the results obtained from the NH and HI listener groups. Pearson's correlation coefficients were transformed using the Fisher Z-transformation prior to statistical analyses. Classification scores, speech comprehension scores, ITPC values and difficulty ratings were arcsin transformed prior to statistical analyses. When appropriate, we used

a false discovery rate (Benjamini and Hochberg, 1995) to correct for multiple comparisons. All statistical tests were conducted using R version 3.4.4 (2018-03-15).

3.3 Results

3.3.1 Behavioral tests

Figure 3.2 shows the results of the behavioral tests. In terms of speech-in-noise perception, HI subjects showed significantly higher sentence reception thresholds compared to the age-matched NH controls (monaural DaHINT test: $t(38.859) = 3.4194$, $p = 0.001487$; Fig. 3.2b). This reduced speech-in-noise performance was observed despite the fact that the speech stimuli were amplified to compensate for reduced audibility in the HI subjects. Noticeably, the 50 % speech reception thresholds were negative also for most HI subjects and well below SNRs typically encountered in everyday environments (Billings and Madsen, 2018). HI subjects also exhibited a reduced temporal masking release compared to the NH subjects ($t(-32.739) = -5.5266$, $p < 2 \cdot 10^{-6}$; Fig. 3.2d), suggesting degraded temporal processing abilities. In addition, HI listeners reported greater difficulties with speech listening in everyday listening situations when wearing their own hearing aid, as rated on the SSQ questionnaire. The different SSQ ratings related to spatial hearing, speech perception and sound quality were correlated (Spearman's rank correlations: r(SSQ speech, SSQ spatial)$= 0.79$, r(SSQ speech, SSQ quality)$= 0.82$, r(SSQ spatial, SSQ quality)$= 0.86$) . The ratings averaged across the three subsections of the questionnaire were significantly lower for the HI group compared to the NH listeners ($t(27.185) = -6.5927$, $p < \cdot 10^{-6}$; Fig. 3.2c). The backward digit span test confirmed similar working memory performance in the age-matched normal-hearing and hearing-impaired listeners ($t(37.847) = -0.67855$, $p > 0.5$; Fig. 3.2e).

3.3.2 Behavioural results from selective attention experiment

Figure 3.1b shows the behavioral results obtained in the EEG selective auditory attention experiment. Both normal-hearing and hearing-impaired listeners showed accurate speech comprehension, both in the single-talker condition and in the condition with two loudness matched competing talkers. A repeated measures ANOVA showed no main effect of hearing impairment on speech

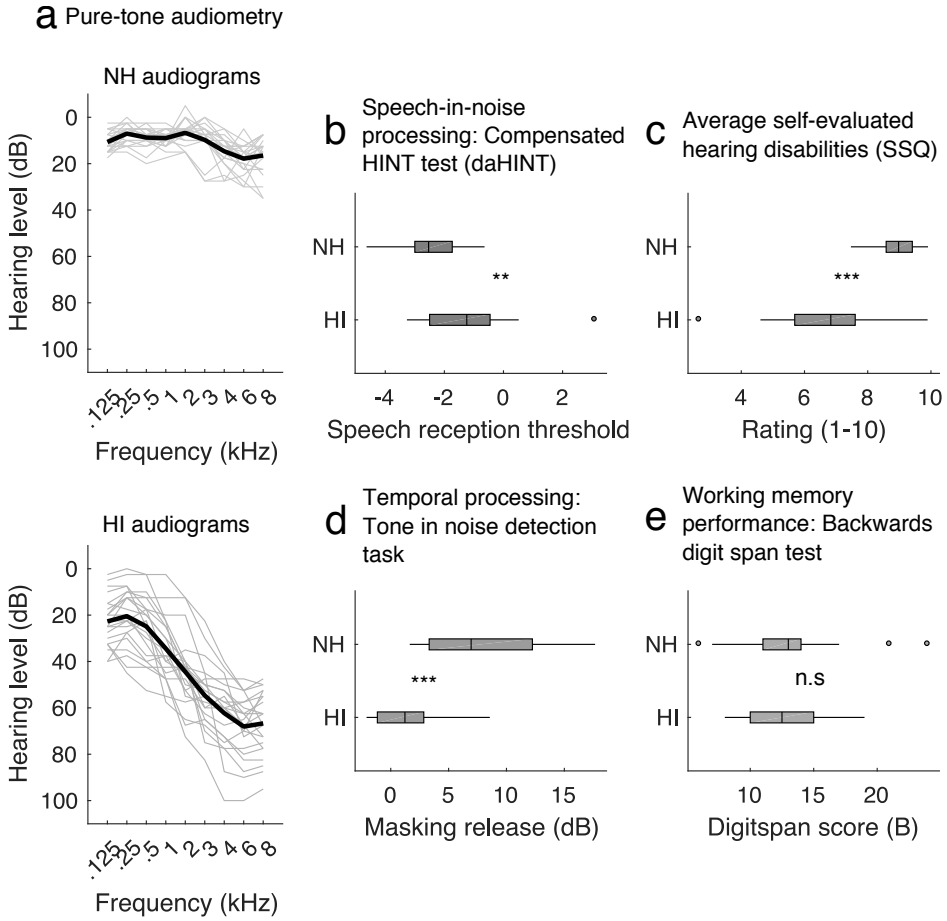

Figure 3.2: Behavioral hearing profiles. (a) Pure-tone audiograms for the normal- hearing (upper panel) and hearing-impaired (lower panel) subjects. Each thin line represents the audiogram averaged over both ears for a single subject. The thick lines represent averages over all subjects in each group. (b) Sentence reception thresholds (SRT) in the two groups measured in a speech-in-noise recognition task. (c) Self-assed hearing disabilities as measured by the Speech, Spatial and Qualities of Hearing Scale questionnaire (SSQ). Lower ratings indicate greater self-rated listening difficulties in everyday acoustic environments. The SSQ scores shown here are averaged over three SSQ subsections (”speech”, ”qualities” and ”spatial”). (d) Tone detection in noise with or without a 50 ms temporal gap. Hearing-impaired listeners showed less temporal release of masking, i.e. they showed a smaller benefit from temporal gaps in the noise masker compared to normal hearing listeners. (e) Working memory performance as measured by an backwards digit span test. No differences in working memory performance were observed.

comprehension scores ($F(1,42) = 1.305$, $p > 0.26$), but a main effect of talker condition (single vs two talkers, $F(1,42) = 8.420$, $p < 0.00589$).

Although the two subject groups answered the comprehension questions with high accuracy, we found that the HI listeners rated the competing speech listening task to be more significantly difficult compared to the NH listeners and compared to the single-talker condition (interaction between normal vs impaired hearing and single vs two talkers: $F(1,42) = 4.126$, $p = 0.0486$) (Fig. 3.1b, left).

3.3.3 Cortical EEG correlates of speech envelope entrainment

Normal-hearing and hearing-impaired subjects listened to speech in quiet or to speech masked by loudness-matched competing speech. To investigate EEG correlates of cortical speech envelope entrainment in the two groups, we used forward and backward stimulus-response models. Forward model prediction accuracies, i.e. the correlation between the low-frequency EEG response and the response predicted by the encoding model, are shown in Fig. 3.3. A repeated-measures ANOVA was used to test the effect of hearing impairment on the fidelity of attended speech entrainment measured by the prediction accuracies. Main effects of hearing status (NH vs HI, $F(1,42) = 7.902$, $p < 0.00747$) and stimulus condition (single-talker vs two-talker, $F(1,42) = 80.507$, $p < 2.57 \cdot 10^{-11}$) were found, but with no significant interaction between the two ($F(1,42) = 0.934$, $p > 0.339$). An enhanced envelope entrainment to attended speech was observed in HI listeners compared to NH controls in both stimulus conditions. We found no effect of hearing status on the accuracies for the unattended speech ($t(42) = 0.28064$, $p > 0.7804$).

Our entrainment analysis assumes a compressed representation of the speech envelope. Age-related hearing loss, however, is typically associated with loss of outer hair cells, which reduces the compressive properties of the inner ear. We performed the same analysis, but with joint encoding models operating on multiple speech envelopes that were power-law compressed with ($c = \{0.1, 0.2, 0.3, ..., 1\}$). An increased prediction accuracy would suggest that other power-law compressive factors lead to improved predictive power. A repeated-measures ANOVA was used to investigate the effect of HI on prediction accuracies obtained with joint models trained on attended speech. This analysis revealed a main effect of HI on prediction accuracies ($F(1,42) = 7.961$, $p < 0.00727$), a main effect of listening condition (single-talker vs two-talker)

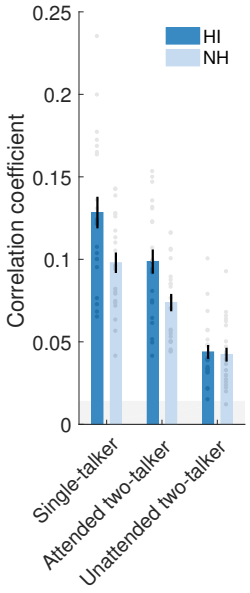

Figure 3.3: Results of the forward model analysis of speech envelope entrainment. Bars indicate the correlation between the speech envelope model and the EEG responses of the normal-hearing (light blue) and hearing-impaired (dark blue) listeners. Data is here averaged over a fronto-central cluster of electrodes. Each point represent data from a single subject. Error bars represents s.e.m. across subjects. The shaded area indicates the estimated noise floor

($F(1,42) = 62.35$, $p < 7.81 \cdot 10^{-10}$) and no significant interaction ($F(1,42) = 0.48$, $p > 0.492$). No effect of hearing impairment on unattended accuracies obtained with this model was found $t(41.999) = 0.44691$, $p > 0.6572$).

The backward model reconstructs the envelope of attended and unattended speech streams from the multi-channel EEG response. The reconstruction accuracies provide a complimentary measure of envelope entrainment that takes all EEG electrodes into account. Results from this analysis are shown in Figure 3.7. Again, we observed a main effect of hearing status ($F(1,42) = 13.18$, $p = 0.000764$), and stimulus condition (single-talker vs two-talker) ($F(1,42) = 73.748$, $p < 8.6 \cdot 10^{-11}$), suggesting again an enhanced fidelity of cortical envelope entrainment in the HI subject group. We did not observe a any effect of hearing impairment on the ability to reconstruct envelopes of unattended speech ($t(39.195) = 1.275$, $p > 0.2098$).

Both our analyses suggested a robust differential envelope entrainment to the attended and unattended speech signals in both groups, similar to what has been reported for normal-hearing listeners (e.g. Ding and Simon, 2012a; Mesgarani and Chang, 2012; O'Sullivan et al., 2014). We next investigated the

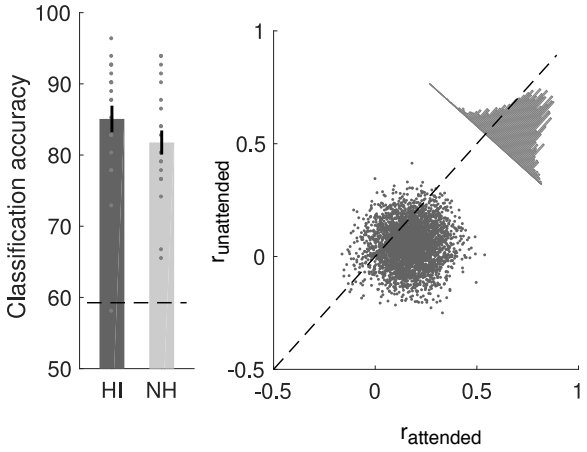

Figure 3.4: Results of the attention decoding analysis. Stimulus reconstruction models trained on EEG responses to single-talker speech stimuli were used to decode envelopes of attended speech from EEG responses to two-talker stimuli in 10 s long decoding segments. For a given segment, correct classification indicates that r_{attended} was higher than $r_{\text{unattended}}$. Left: Attention classification accuracies in normal hearing and hearing-impaired listeners. Each point represent data averaged from a single subject. The dashed line represents chance-level. Errorbars indicate s.e.m. Right: Reconstruction accuracies. Each point reflect accuracies for a given subject and a given decoding segment. Data is here shown for all subjects and all decoding segments.

degree to which this differential response could be used to classify the attentional focus from single-trial EEG responses (O'Sullivan et al., 2014). Here, we focused on backward models trained on data from single-talker trials. A correct classification here does not by itself necessarily suggest that r_{attended} is high, but only that it is higher than $r_{\text{unattended}}$. The classification accuracies were evaluated for 10 s long decoding segments. We found that the reconstruction filters could reliably discriminate between attended and unattended speech in both groups of listeners (Fig. 3.4). Classification accuracies were above chance-level for all subjects except for one ($\alpha < 0.05$, binomial tests). The mean classification accuracy for 10 s long segments of unaveraged EEG data were 85.07% for the HI listeners and 81.76% for the NH listeners (Fig. 3.4). We found no effect of hearing loss on the attention classification accuracies ($t(41.625) = 1.4141$, $p > 0.1648$).

3.3.4 Envelope entrainment to tones during passive stimulation

Our speech experiments with both single and competing talkers suggested an enhanced envelope entrainment in HI listeners, possibly indicating a stimulus-

driven effect. Next, we recorded EEG responses to repeated tone sequences during passive stimulation to obtain measures of cortical envelope entrainment unrelated to attention or listening effort. We used periodic multi-scale tone stimuli designed to evoke activity in the gamma (40 Hz) or theta (4 Hz) range (stimuli illustrated in Fig. 3.9). We computed the inter-trial phase coherence (ITPC) to assess how precisely EEG activity synchronized to the periodic tone sequences. The top row in Figure 3.5 shows the ITPC results for the two types of tone stimuli in non-overlapping analysis windows of 0.5 s. Results have been averaged over all electrodes (see Fig. S1 for topographies of the 4 Hz ITPC and 40 Hz ITPC). For the 4 Hz stimulation (Fig. 3.5, top left), a repeated measures ANOVA on the ITPC revealed a main effect of hearing impairment ($F(1,42) = 4.956$, $p = 0.0314$), a main effect of time ($F(3,126) = 85.458$, $p < 2 \cdot 10^{-16}$) as well as an interaction effect ($F(3,126) = 5.404$, $p = 0.00157$). Post hoc t-tests revealed that the 4 Hz ITPC was significantly higher (after applied FDR corrections, $q = 0.05$) for HI than for NH in the time period 0 s to 1 s post onset (0-0.5 s: $t(41.92) = 2.7799$, $p = 0.008103$; 0.5-1.0 s: $t(42) = 2.7452$, $p = 0.008861$; 1.0-1.5 s: $t(42) = 1.5808$, $p = 0.1214$; 1.5-2.0 s: $t(42) = 0.95423$, $p = 0.3454$). On the other hand, the ITPC across sustained 40 Hz gamma stimulation (Fig. 3.5, top right) showed no main effect of hearing loss ($F(1,42) = 1.531$, $p = 0.223$), or time ($F(3,126) = 1.414$, $p = 0.242$).

3.3.5 Event-related potentials

We measured ERPs to transient 1 kHz tones beeps during passive listening. No effect of hearing impairment on the mean amplitude of the N1 was found ($t(41.928) = -1.1891$, $p = 0.2411$; Figure 3.8 in Supplementary Material). We additionally extracted ERPs elicited by the onset of the periodic tone sequences described above. Again, no effect of hearing impairment on the mean N1 amplitudes was observed, neither for the theta ($t(40.936) = -0.71995$, $p = 0.4756$) nor the gamma stimuli ($t(41.983) = -1.8861$, $p = 0.06621$). This could also indicate that the group difference in the EFRs measured by the ITPC was not driven by amplitudes of transient ERP responses.

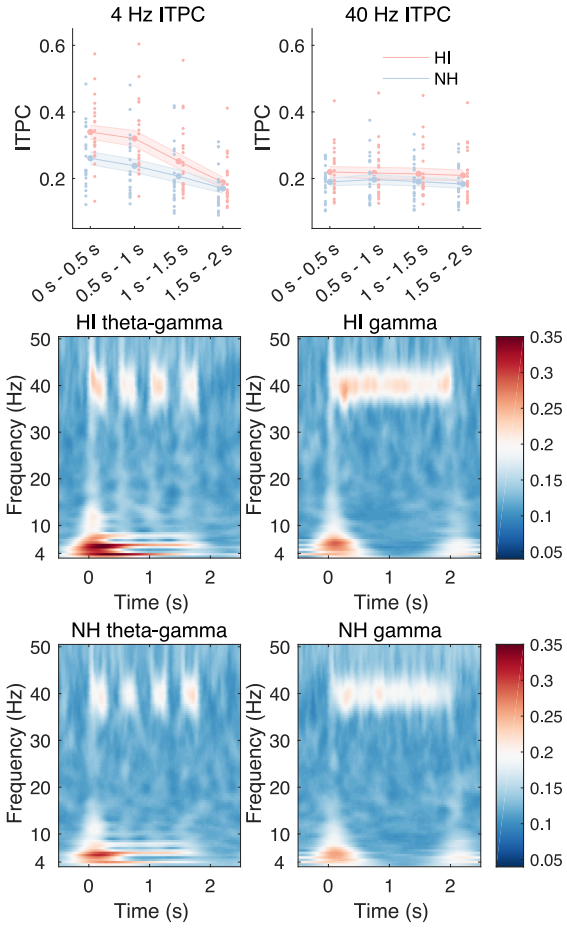

Figure 3.5: EEG responses to periodic tone sequences during passive stimulation. Left column: ITPC for EEG responses to 4 Hz theta tone stimuli. Right column: ITPC for EEG responses to 40 Hz gamma stimuli. Top row: Mean ITPC across subjects for both groups of listeners, averaged over all scalp electrodes in nonoverlapping 0.5 s long time intervals over the 2 s long stimulation periods. Shaded areas indicate s.e.m. The individual red (HI) and blue (NH) points represent data from each subject. Middle and bottom rows: Time-frequency representations of the ITPC for each subject group.

3.4 Discussion

In the present study, we investigated effects of peripheral hearing loss on EEG correlates of attention-driven cortical speech envelope entrainment. Twenty-two HI subjects and twenty-three age-matched NH subjects took part in the experiment. The results from a monaural speech-in-noise test (DaHINT) suggested that the identification of speech in noise was more vulnerable to the level of background noise in the HI subject group than in the NH subject group. The HI listeners demonstrated lower temporal processing abilities in a behavioral tone detection task (FT test) and reported greater everyday-life hearing disabilities (SSQ scores), but performed equally well as the NH listeners in a backwards digit span task. In the auditory attention EEG experiment where the subjects listened to speech in quiet or to speech masked by competing speech, we found that the fidelity of the cortical EEG entrainment to the envelopes of attended speech was enhanced in the HI listeners compared to the NH listeners, but that speech comprehension scores were equally high in the two subject groups. Moreover, it was possible to classify from EEG activity which of the two concurrent speech streams the subjects were attending to. We obtained equally high classification accuracies in the two subject groups, suggesting robust attentional modulations of EEG entrainment responses in both subject groups. Finally, in a passive listening EFR paradigm, we observed an enhanced 4 Hz EEG phase-locking to rhythmic tone sequences in the HI subject group. This suggest that group differences in the fidelity of the cortical EEG envelope entrainment responses may not necessarily be effort-related or speech-specific. The findings support that, even though an SNHL may be associated with exaggerated sound evoked cortical responses, it does not necessarily hamper attentional modulations of cortical auditory entrainment responses.

Exaggerated cortical responses to auditory stimuli have previously been observed in the aging auditory system Alain, 2014; Lister et al., 2011; Presacco et al., 2016a,b; Sörös et al., 2009; Tremblay et al., 2003. Yet, studies on the effects of aging on central auditory processing often compare young normal hearing listeners with older listeners who have non-normal pure tone audiograms (suggestive of deficits to the auditory periphery). The results from this study suggest that such peripheral deficits could be an important contributor to exaggerated sound-evoked cortical entrainment responses.

HI listeners often experience listening difficulties in selective listening tasks

(e.g. when listening to monaural speech embedded in stationary background noise; Fig 3.2a). A degraded peripheral input can affect the perceptual segregation of sounds and interfere with central attention processes (Dai et al., 2018). Yet, the behavioural results from the auditory attention EEG experiment suggested that there was no effect of an SNHL on the ability to recognize speech in quiet or speech masked by competing speech. This was mirrored by results from the stimulus-response analysis that suggested equally robust attentional modulations of cortical auditory entrainment responses in both subject groups. Furthermore, we found that the HI subject group rated selective listening task to be more difficult than the NH subject groups did. This may indicate that HI listeners put more effort into maintaining effective attentional control (Peelle and Wingfield, 2016; Peelle et al., 2011).

The present study extended previous studies (Alain, 2014; Kale and Heinz, 2010; Millman et al., 2017; Zhong et al., 2014) by demonstrating that a HI can enhance the fidelity of the cortical EEG entrainment to rhythmic tone sequences during passive listening, but also to envelopes of attended speech during selective listening. It would be interesting to consider these results in the light of the ”fluctuation-profile coding” framework that was recently proposed (Carney, 2018). This framework proposes that slowly varying amplitude fluctuations across auditory-nerve (AN) channels convey important information about complex sound features. The neural fluctuation-profiles depend, among other things, on inner hair cell transduction properties and on an efferent system that controls the cochlear gain and the sharpness of the fluctuation-profile. A sensorineural hearing loss may, according to this framework, negatively impact the neural fluctuation-profiles in several ways, e.g. as a consequence of changes in the inner hair-cell transduction properties. The resulting effect is that the fluctuation-profile contrasts across AN channels may be reduced, while the fluctuation amplitudes are increased. The enhancement in fluctuation amplitudes may not be perceptually beneficial, as reduced contrasts across AN channels could impair the system. This could potentially explain why we observed enhanced cortical EEG envelope entrainment responses in listeners with a SNHL who otherwise showed a degraded performance in a behavioural temporal processing task and in a speech-in-noise identification task.

The speech streams were spatially separated in the auditory attention EEG experiment in order to support the perceptual segregation of sound streams. The spatial separation of the competing speech streams was not explicitly taken

into account in the stimulus-response analyses, but may have affected the results (Dai et al., 2018; Marrone et al., 2008). However, the gender of the target speaker as well as the spatial lateralization of the target speaker was randomized across trials in an attempt to avoid systematic biases in EEG responses related to the spatial position of the target speaker as well as the gender of the target speaker. The motivation behind spatially separating the speech streams was that HI listeners often experience listening difficulties in monaural two-talker listening conditions (Neher et al., 2009). Reduced speech intelligibility (possibly related to deficits in perceptual sound segregation) in the HI subject group could potentially affect results from stimulus-response analysis (Ding and Simon, 2013; Kong et al., 2015). An alternative strategy to improve speech intelligibility and support perceptual sound segregation would be to present the speech mixtures at positive target-to-masker ratios. However, in this case, the EEG synchronization to the envelopes of attended speech could reflect a cortical encoding of the sound mixture rather than of individual sound streams. Selective responses to attended speech would, in this case, not be dissociable from effects of differences in sound level between streams.

In our stimulus-response analyses, we assumed that the predictive performance of the models would represent the fidelity of the cortical EEG entrainment to target envelopes (Ding and Simon, 2012a). Statistical stimulus-response dependencies between speech envelopes and M/EEG responses may reflect both sensitivity to, and the binding of, different sound features beyond the envelope (Ding and Simon, 2014). Interestingly, we found an effect of SNHL on the fidelity of cortical envelope entrainment responses both to attended speech, but also to rhythmic tone stimuli. This could be interpreted as indicating that enhanced speech envelope entrainment responses observed in HI listeners primarily reflect cortical encoding of low-level acoustic features rather than higher-level phonetic- or linguistic features.

Recent evidence suggest that attentional modulations of tone-evoked cortical EEG responses may increase over short time scales for NH listeners, but not for HI listeners (Dai et al., 2018). In our attention experiment, we considered long trials with speech stimuli where the target speech stream was onset a few seconds before the distracting speech stream. However, it is possible that similar attentional build-up effects to those reported in (Dai et al., 2018) could affect EEG responses to speech mixtures.

Several studies have investigated the effects of SNHL on ERPs in age-matched

older listening groups (Alain, 2014; Bertoli et al., 2005; Harkrider et al., 2006; Tremblay et al., 2003). However, results have been inconsistent regarding N1 amplitudes. Bertoli et al., 2005 found reduced tone-evoked N1 amplitudes in HI listeners, Harkrider et al., 2006 reported enhanced speech-evoked N1 amplitudes in HI listeners, Tremblay et al., 2003 reported enhanced N1 amplitudes in HI listeners for voiceless speech with longer voice onset times, but not for stimuli with short voice onset times, and finally, Alain, 2014 found no significant effect of SNHL on N1m (the magnetic counterpart of N1). Here, we did not observe any effects of HI on the mean amplitudes of the N1 component elicited by tone stimuli. This was the case both for EFR stimuli and for ERP stimuli. One way to interpret our results is that the stimulus-response models and the 4 Hz ITPC measures can provide additional information about the temporal dynamics of band-limited EEG activity. However, this conclusion is premature at this point, particularly considering the relatively low number of epochs used for extracting ERPs.

In the auditory attention EEG experiment, we applied a frequency-dependent amplification scheme to speech stimuli to account for reduced audibility in the HI subject group. This was done to investigate speech processing and effects of attention beyond changes in audibility of the signal. However, how sound amplification might affect the results remains unclear. It is possible that the cortical encoding of some sound features depends on the physical sound pressure level of the acoustic stimulus rather than on their sensation level (Billings et al., 2007; Carney, 2018; Jenkins et al., 2018). This could, in turn, affect results from the M/EEG experiments. Here, different sound amplification strategies were considered for the selective attention speech experiment and for the tone stimulation experiments, but results from both experiments suggested an enhanced cortical envelope entrainment in HI listeners. Finally, it is possible that some of the HI listeners were accustomed to certain hearing aid settings. It is not clear how everyday usage of hearing aids affect behavioural performance in selective attention tasks.

3.5 Supplementary Material

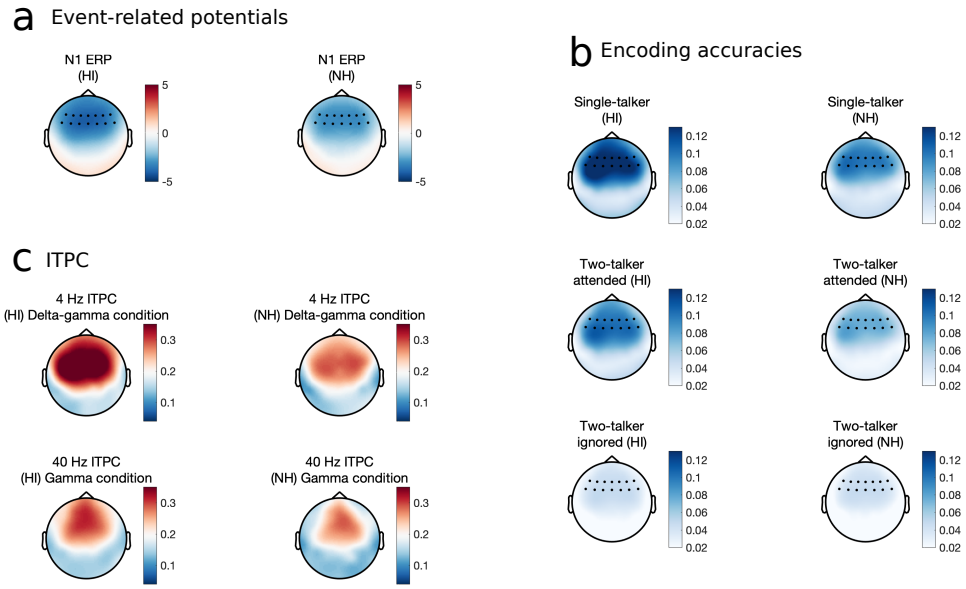

Figure 3.6: Topographies of different analyses. a: Topographies of mean N1 amplitudes in the tone-evoked ERPs. b: encoding accuracies obtained with condition-specific forward models in the single-talker conditions (top-row) and in two-talker conditions (middle row and bottom-row). c: ITPC across EEG responses to multi-scale stimuli. Top-row here depicts 4 Hz ITPC for 4 Hz theta stimulation. Bottom row depicts 40 Hz ITPC for gamma stimulation. The ITPC was here averaged over a time period from 0 s to 2 s for illustration purposes. The highlighted electrodes in a and b were considered for the statistical analyses.

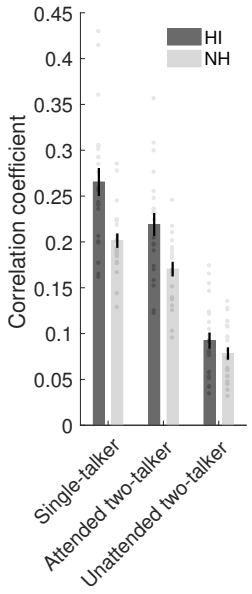

Figure 3.7: The ability to reconstruct envelopes of attended and unattended speech streams from EEG activity in the different listening conditions. Model performance was assessed by nested cross-validation procedures. Reconstruction accuracies reflect Pearson's correlation coefficient between neural reconstruction and target envelope over trials not used for model fitting. Points reflect averaged data for each individual subject. Errorbars represent s.e.m. Bar height reflect group-mean average.

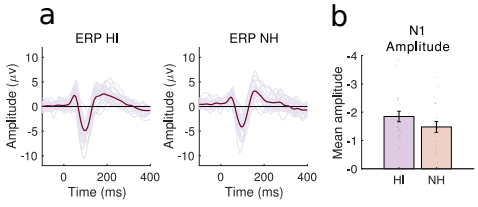

Figure 3.8: Cortical event-related potentials (ERPs) to 1000 Hz ramped tones presented with pseudo-random repetition rates. **a** Individual traces of ERP data averaged over the same fronto-central electrode cluster that was used for the entrainment analysis. Thin lines reflect data from individual subjects. Thick lines reflect group-mean averages. **b**: Mean amplitude of N1 ERPs. Each point reflect data from individual subjects. Errorbars show s.e.m. across subjects.

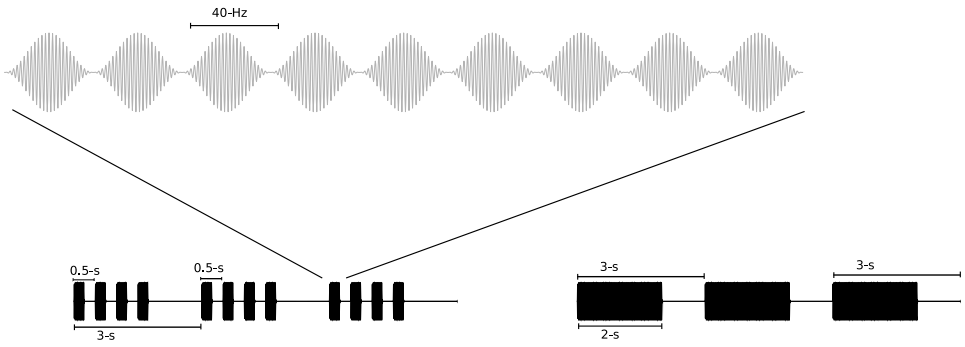

Figure 3.9: Schematic of sound stimuli used for EFR experiments. The carrier frequency of the pure tones was 1 kHz. During sound stimulation, the inter-tone-interval was 40 Hz (top). Left: schematic of delta-gamma stimuli. In this case, 0.5 s long trains of tone beeps alternated with 0.5 sec silence. After three repititions, the stimuli were followed by 1 s of silence. Right: gamma stimuli. In this case the the tone sequences lasted 2 s, followed by 1 s silence.

4

EEG correlates of entrainment to speech envelopes in reverberant, multi-talker environments[a]

Abstract

Selectively attending to one speaker in a multi-speaker scenario is thought to synchronize low-frequency cortical activity to the attended speech signal. In recent studies, reconstruction of speech from single-trial electroencephalogram (EEG) data has been used to decode which talker a listener is attending to in a two-talker situation. It is currently unclear how this generalizes to more complex sound environments. Behaviorally, speech perception is robust to the acoustic distortions that listeners typically encounter in everyday life, but it is unknown whether this is mirrored by a noise-robust neural tracking of attended speech. Here we used advanced acoustic simulations to recreate real-world acoustic scenes in the laboratory. In virtual acoustic realities with varying amounts of reverberation and number of interfering talkers, listeners selectively attended to the speech stream of a particular talker. Across the different listening environments, we found that the attended talker could be accurately decoded from single-trial EEG data irrespective of the different distortions in the acoustic input. For highly reverberant environments, speech envelopes reconstructed from neural responses to the distorted stimuli resembled the original clean signal more than the distorted input. Single-trial attention decoding accuracies based on 40-50s long blocks of data from 64 scalp electrodes were equally high (80-90% correct) in all considered listening environ-

[a] This chapter is based on: Fuglsang, S. A.; Dau, T.; Hjortkjær, J. H (2017). "Noise-robust cortical tracking of attended speech in real-world acoustic scenes". Neuroimage

ments and remained statistically significant using down to 10 scalp electrodes and short (< 30-s) unaveraged EEG segments. In contrast to the robust decoding of the attended talker we found that decoding of the unattended talker deteriorated with the acoustic distortions. These results suggest that cortical activity tracks an attended speech signal in a way that is invariant to acoustic distortions encountered in real-life sound environments. Noise-robust attention decoding additionally suggests a potential utility of stimulus reconstruction techniques in attention-controlled brain-computer interfaces.

4.1 Introduction

Speech communication is remarkably robust to the signal distortions encountered in everyday acoustic environments. Successful speech comprehension in noisy situations relies both on the ability of the auditory system to segregate simultaneous sound sources, but also on the listeners' ability to direct attentional focus to a potentially degraded sound stream while suppressing irrelevant information. Numerous electrophysiological studies in humans have reported a synchronization between the slow (<20 Hz) temporal modulations inherent in speech signals and low-frequency cortical activity in the delta (1-4 Hz) and theta (4-8 Hz) frequency ranges (Ahissar et al., 2001; see Ding and Simon, 2014 for a review). In scenarios with more than one talker, selective attention has been shown to enhance the cortical tracking of the attended speech and to suppress synchronization of the ignored speech (Ding and Simon, 2012a,b; Golumbic et al., 2013; Horton et al., 2013; Kerlin et al., 2010; Mesgarani and Chang, 2012; Power et al., 2012). A number of recent studies have employed 'stimulus reconstruction' techniques (Bialek et al., 1990; Mesgarani et al., 2014a; Rieke et al., 1995) to reconstruct the envelopes of competing speech signals from the EEG response. It has been shown that an enhanced neural reconstruction of the attended speech signal can be used to decode which talker a listener is attending to with less than one minute of unaveraged EEG data (Mirkovic et al., 2015; O'Sullivan et al., 2014). However, successful decoding of attention from single-trial EEG has only been demonstrated in acoustically controlled environments with two competing talkers. It is currently unknown how well these results generalize to real-life acoustic scenarios where the speech signals to be decoded are distorted, e.g. by reverberation or background noise.

In order to extract speech sources from a complex scene, the brain must represent the relevant signal in a way that is robust to noise. Converging evidence from animal electrophysiology suggests that noise-invariant representations of sounds emerge at later stages in the auditory pathway by neuronal adaptation to stimulus statistics (Mesgarani et al., 2014a; Moore et al., 2013; Rabinowitz et al., 2013). Using stimulus reconstruction with population responses in ferrets, (Mesgarani et al., 2014a) demonstrated that speech stimuli with added reverberation or stationary noise reconstructed from auditory cortex resembled the original clean signal more than the distorted signal. Temporal coding of amplitude modulations in the auditory midbrain may even be enhanced in real-world reverberation compared to anechoic conditions (Slama and Delgutte, 2015). In cortex, the entrainment of low-frequency activity to speech envelope fluctuations has been shown to be robust to stationary noise, even when the intensity of the background noise is greater than the speech signal (Ding and Simon, 2013). However, in situations where the 'background' sound stream is itself a potential auditory object of interest (e.g., another speech signal), perception cannot rely solely on bottom-up mechanisms. In this case, top-down attention plays a critical role in ignoring irrelevant information. Yet, it remains unclear how attention may contribute to the formation of noise-invariant representations in human cortex.

In real-world listening environments, speech signals are inevitably distorted, e.g. by sound reflections and reverberation. Unlike background sounds that can form a separate sound stream to be ignored, such 'transmission' distortions degrade the attended speech signal itself. Previous studies have mainly considered how acoustic degradations that reduce speech intelligibility affect envelope entrainment. Noise-vocoded speech with a temporal envelope that resembles that of the original signal but degrades speech intelligibility has been found to reduce cortical speech tracking responses in the theta range (Ding and Simon, 2014; Peelle et al., 2012) and to diminish the differential cortical response between the attended and the unattended talker (Kong et al., 2014; Rimmele et al., 2015). However, such artificial signal manipulations also affect the statistics of natural speech stimuli and degrade the acoustic cues that listeners typically use to maintain stable speech recognition. In real-world reverberant rooms, normal-hearing listeners can maintain robust speech recognition even when the acoustic envelope of a signal has been substantially distorted (Darwin and Hukin, 2000; Ruggles and Shinn-Cunningham, 2011). Robust discrimination of

sound sources has been proposed to rely on the statistical regularities that are found in real-world reverberation (Traer and McDermott, 2016). It is currently unknown whether robust speech perception, despite the envelope distortions imposed by real-world rooms, is mirrored by a cortical entrainment mechanism that is robust to these distortions.

In the current study, we investigated this question using speech stimuli embedded in real-world acoustic scenes. Using acoustic simulations, we created virtual auditory scenes with multiple talkers and different reverberant decay properties. In a two-tiered approach for reproduction of these sound scenes, we first reproduced the scenes over earphones in an electromagnetically shielded environment to control for noise in the EEG measurements. Next, we reproduced the sound scenes using a multi-loudspeaker virtual reality facility providing accurate free-field reconstructions of real-world rooms. This provided listeners with the full range of acoustic cues typically encountered in everyday listening situations. Since reverberation and background talkers can severely disrupt the envelope of an attended speech signal, this approach allowed us to examine the cortical tracking of distorted envelopes with intact speech recognition. We measured ongoing scalp EEG while the subjects selectively attended to a particular talker embedded in these different scenes. Using stimulus reconstruction to derive envelopes of attended speech streams from single-trial EEG, we investigated (i) whether the clean envelope of an attended speech stream could be reconstructed from a distorted input when speech perception is robust, and (ii) whether this might facilitate the decoding of the attended talker in real-world acoustic scenes. We hypothesized that attention promotes noise-robust neural representations of attended sound streams such that attended speech envelopes reconstructed from cortical EEG resemble the clean signals to a similar or even higher degree than the distorted input. Such a cortical robustness would allow attention decoding accuracies in real-world acoustic scenes that are similar to those observed with undistorted signals.

4.2 Material and methods

4.2.1 Participants

Twenty-nine subjects (13 females, 25 right-handed), aged from 19 to 30 years, participated in the experiment. Three subjects were excluded from the analysis

because of missing data from several trials. All participants were students with self-reported normal hearing and no history of neurological disorders. The subjects received financial compensation for their participation in the experiment. The experimental procedure was approved by the Science Ethics Committee for the Capital Region of Denmark, and written informed consents were obtained from all participants before the experiment in accordance with the Declaration of Helsinki.

4.2.2 Stimuli and virtual room simulations

Two hours of speech material were recorded from a male and a female professional story teller narrating fictional stories. The speech material was recorded in an anechoic chamber at the Technical University of Denmark (DTU) and sampled at a frequency of 48 kHz. The naturally spoken stories were subsequently segmented into consecutive 50-second long segments that were each accompanied by multiple-choice questions.

Virtual auditory environments (VAEs) were simulated using the room acoustic modeling software Odeon (version 13.02). To create auditory scenes representative of everyday listening environments, we simulated the acoustics of a mildly reverberant room and a highly reverberant room. A model of a square classroom at DTU (9×7×3 m^3) was used to represent a mildly reverberant environment and a model of the Hagia Irene church (39.000 m^3) represented a highly reverberant listening scenario. Binaural impulse responses were derived for each of the VAEs at two source-receiver positions, with source-receiver distances of 2.4 m and target sources positioned at $\pm 60^o$ along the azimuth with 0^o elevation angles (see Figure 1C). The impulse responses of the simulated rooms and the 50-s long excerpts of the speech material were used to create virtual auditory scenes with the two speakers of different gender talking at the same time from different positions. The two concurrent speech streams were normalized to have the same root-mean-square (RMS) value and were presented at a sound pressure level (SPL) of 65 dB. For the mildly reverberant room, a scenario with additional 6 talkers (3 male, 3 female) positioned uniformly along the azimuth direction 2.4 m from the listener was simulated in addition to the two target talkers positioned at $\pm 60^o$ azimuth. The multi-talker babble of the six additional speakers was presented at a level of 55 dB SPL. The average decay time constant of the mildly- and highly reverberant rooms were T_{30}=0.9 s and T_{30}=4 s, respectively. The clarity, defined as the ratio of the direct 80-ms

sound energy to the remaining energy, ranged between $C_{(80,63Hz)}$=5.7 dB and $C_{(80,4kHz)}$=7.4 dB for the mildly reverberant room and between $C_{(80,63Hz)}$=6.7 dB and $C_{(80,4kHz)}$=9.7 dB for the highly reverberant room. For the data analysis, all reverberant signals and their corresponding anechoic signals were temporally aligned using cross-correlation. For comparison, a condition with the two talkers positioned at $\pm 60^o$ azimuth in an anechoic room with no reverberation was also included.

The speech stimuli were delivered via two different sound reproduction systems. For nine subjects, the VAEs were reproduced with a multi-loudspeaker virtual reality facility. The multi-loudspeaker setup consisted of 64 loudspeakers mounted in an anechoic chamber on a full spherical array with a radius of 2.4 m and four subwoofers. The listener was positioned on a suspended wire floor at the center of the loudspeaker array. To reproduce the sound fields within the loudspeaker array, we used a loudspeaker-based room auralization system (Favrot et al., 2010), which separately processes three parts of binaural room impulse responses (accounting for the direct sound, early reflections, and late reflections) using higher-order Ambisonics (Malham and Myatt, 1995). For the other 20 subjects, the binaural VAEs were reproduced in a soundproof, electrically-shielded listening booth with ER-2 insert earphones (Etymotic Research) to control for electromagnetic interference during the EEG recordings. In order to spatially separate clean speech signals presented with insert earphones, the clean signals were convolved with non-individualized head-related impulse-responses for azimuth angles of $\pm 60^o$ and an elevation angle of 0^o. Listening environments with more than two competing talkers were only simulated for the recordings in the multi-loudspeaker array.

4.2.3 Extraction of acoustic speech features

For stimulus-response analysis, the time-varying patterns of amplitude modulations in the speech signals were derived from a physiologically inspired model of envelope processing in the peripheral auditory system. For each speech mixture, we separately obtained monaural versions of the speech streams from each talker. In listening environments with reverberation and sound reflections, the monaural speech signals were obtained both for the speech streams with reverberation and for their underlying clean signals (without reverberation). Each of the monaural speech signals was passed through a bandpass filterbank (Patterson et al., 1987) to account for the band-pass characteristics of the basilar

membrane. To mimic the nonuniformity of psychophysically derived auditory filter shapes with increasing frequency (Moore and Glasberg, 1983), the centre frequencies of the bandpass filters were uniformly spaced between 150 Hz and 8000 Hz on an equivalent rectangular bandwidth (ERB) rate scale. The envelope of the signal was extracted at the output of each cochlear filter via the analytical signal obtained from the Hilbert transform and raised to the power of 0.3 to account for the compressive response of the inner ear (Plack et al., 2008). Subsequently, the narrowband envelopes were low-pass filtered at 10 Hz using a single zero-phase modulation filter. Finally, the output was averaged across cochlear filters.

4.2.4 Experimental procedure

During 50-s long presentation intervals, the subjects listened selectively to one of two or more simultaneous speech streams in the different simulated acoustic environments. Each subject listened to sixty trials in which they were cued to listen attentively to the target speech stream while ignoring any competing talker. The subjects were instructed to minimize motor activity and fix their eye gaze to a cross hair during stimulus presentation. After each trial, the subjects were required to answer a multiple-choice questionnaire related to the content of the attended speech stream. The answers to the comprehension questions served as indicators of whether the subjects attended the target talker and whether the speech was comprehensible in the different listening conditions. The subjects had to take longer breaks after 10 successive trials. The order of presentation of the different virtual auditory environments (two-talker anechoic, two-talker mild reverberation, 8-talker mild reverberation, two-talker high reverberation) was independently randomized across trials for each participant. Moreover, the position of the target speaker relative to that of the listener ($\pm 60^o$) as well as the gender of the target speaker were randomized across blocks for each subject. Finally, we randomized the order in which the stories were presented across subjects. The randomization procedure was done to avoid the likelihood that the neural responses to attended and unattended speech would be systematically biased by talker identity, pairing of the fictional stories or talker position.

4.2.5 EEG data acquisition

EEG were recorded from 64 electrodes mounted on a head cap (10/20 layout) using a BioSemi ActiveTwo system. The data collected from each electrode were recorded with a sampling rate of 512 Hz. Two additional electrodes were placed on the mastoids as physiological reference signals. For the purpose of removing electroocular activity, we recorded vertical and horizontal electrooculograms (EOGs) using six bipolar electrodes.

4.2.6 Data analysis

EEG preprocessing

The EEG data were analyzed offline using Matlab (MathWorks) and the Fieldtrip toolbox (Oostenveld et al., 2011). The signals picked up by each electrode were band-pass filtered using a zero-phase forward filter with delay compensation with -6dB cutoffs at 1 Hz and 64 Hz. The data for each channel were subsequently downsampled to 128 Hz and re-referenced to the average response of the mastoid electrodes. The time series of each channel were visually inspected and excessively noisy channels were removed from the analysis. On average, there were 1.8 (SD =1.7) channels rejected. Infomax-based independent component analysis (Makeig et al., 1996) was used to identify and remove electrooculographic components from the EEG recordings for each subject. On average, 5.8 (SD = 2.2) components were removed from the data by linear decomposition. Retained components were backprojected to the electrode domain and missing channels were re-interpolated using a spline interpolation algorithm available with Fieldtrip (Oostenveld et al., 2011). Since a number of studies have reported that the neural tracking of speech envelopes measured by EEG or magnetoencephalography (MEG) is most prominent in the delta (1-4 Hz) and theta (4-8 Hz) frequency range (Ding and Simon, 2012a,b; Luo and Poeppel, 2007; O'Sullivan et al., 2014), we finally lowpass-filtered EEG signals at 8 Hz using a zero-phase second-order Butterworth filter.

Reconstructing speech from neural responses

We used a stimulus-reconstruction technique (Mesgarani et al., 2009; O'Sullivan et al., 2014; Pasley et al., 2012) to derive finite impulse response (FIR) filters that linearly map a spatio-temporal neural response to stimulus features (i.e., recon-

struction filters; (Bialek et al., 1990; Mesgarani et al., 2009; Rieke et al., 1995). Here we used the reconstruction method to estimate the temporal envelope of the speech signal, $S(t)$, at time point t from a weighted combination of neural activity $R(t,n)$ picked up by the EEG electrodes across L lags:

$$\hat{S}(t) = \sum_{n=1}^{N} \sum_{l=1}^{L} g(\tau_l, n) R(t - \tau_l, n) \tag{4.1}$$

where $\tau = \{\tau_1, \tau_2, ...\tau_L\}$ represents the time-lags, $n = \{1, 2, ...N\}$ indicate the EEG channels and $\hat{S}(t)$ is the estimated envelope.

To determine the identity of attended speakers from cortical activity, we trained reconstruction filters to estimate the envelope of either the attended speech stream (i.e. attended decoders) or the unattended speech stream (i.e. unattended decoders) (Ding and Simon, 2012a; O'Sullivan et al., 2014). The filter coefficients were estimated using ridge regression (Crosse et al., 2015; Hoerl and Kennard, 1970), trained separately on data from each subject in each acoustic condition (two-talker anechoic, two-talker mild reverberation, two-talker high reverberation and 8-talker cocktail party). Fitting such condition-specific reconstruction filters ensured that the decoders were not biased by the other listening conditions. Attended decoders were fitted using the envelope of the attended speech streams and EEG data from all-but-one trials of a particular condition. The unattended decoders were trained in the same manner, but on the envelope of the unattended speech streams. The decoders were then used to reconstruct envelopes of the individual speech streams from the EEG data in the trial that had not been used for training the models. Reconstruction accuracies were computed as the Pearson's correlation coefficient between the predicted speech envelopes $\hat{S}(t)$ and the envelopes of the target speech stream. We considered a data segment to be successfully decoded, if the reconstructed envelope had a higher correlation with the attended speech envelope than with the unattended speech envelope, or similarly, if the reconstruction of the ignored speech signal was more correlated with the unattended envelope than with the attended one (O'Sullivan et al., 2014). To investigate potential noise-robust cortical representations of the speech envelope, we reconstructed both the temporal envelopes of the speech streams of the individual talkers distorted by reverberation as well as the envelopes of the corresponding original clean signal without reverberation. The 'distorted' signals thus corresponded to the separate speech signals from each talker as they were presented in the simulated

rooms whereas the 'clean' signals corresponded to the underlying anechoic speech signals that were used to generate the room simulations. Both attended and ignored reconstruction filters covered time lags ranging from 0 ms to 500 ms post-stimulus. On average, roughly 13 minutes of data were used for training the decoders in each condition.

To determine an unbiased estimate of the ridge regularization parameter, we recorded additional 500-s of data for each subject with presentation of a single-talker stimulus and empirically estimated the ridge parameter that provided the highest reconstruction accuracy of the envelope on that data. After doing so, the ridge parameter was set to a fixed high value of 2^{14}. We chose this approach to prevent biasing the decoding performance by optimizing the ridge parameter on the same multi-talker data also used for testing as done in previous work (Crosse et al., 2015; Mirkovic et al., 2015). For comparison, we also performed our analyses using the former approach which tunes the ridge parameter for each acoustic condition. We found that the two regularization approaches led to highly similar results.

In listening scenarios with multiple talkers, there is a risk that the subjects' attention might be caught by unattended speech streams if the attended speech stream contains excessively long periods of silences. Moreover, endogenous neural activity during silent periods cannot be accounted for by the temporal fluctuations in the envelope with a linear model. In contrast to previous studies where longer periods of silence in the presented speech stimuli were truncated (Ding and Simon, 2012a,b; Kong et al., 2014, 2015; Koskinen et al., 2013; Mirkovic et al., 2015; O'Sullivan et al., 2014) , we did not shorten silent periods in the presented speech stimuli in order to preserve the natural fluctuations of the speech. Instead, segments of data where either the attended or the unattended speech stream contained periods of silence longer than 500 ms were discarded from the analysis. The same criterion for removal of silences was used for the attended and the unattended speech streams. We finally selected 40-s of data from each trial (excluding the first second of each trial) for the subsequent data analysis, after truncating silent periods in the attended and the unattended speech stream.

Channel selection and stimulus reconstruction

To investigate how accurately the identity of attended speakers can be decoded using a subset of the scalp EEG electrodes, we employed an iterative backward

channel-selection approach (Mirkovic et al., 2015). Specifically, we sought to find spatially sparse models by iteratively removing channels that did not contribute significantly to the prediction of the speech envelope. We did this by first fitting reconstruction filters to the training data using all channels and all post-stimulus time-lags between 0-ms and 500-ms. We then sequentially removed individual channels with the lowest root-mean-square filter weights across all lags. The backward stepwise selection scheme was combined with ridge regression at each iteration, with the ridge parameter still fixed to 2^{14}. Decoding accuracies were evaluated on non-overlapping test blocks of varying duration from the left-out trial ranging between 5-s and 40-s.

Temporal response functions

One shortcoming of the backward model (sections 4.2.6 and 4.2.6) is that the filter weights cannot directly be interpreted as a measure of stimulus-related neural activity (Haufe et al., 2014). The filters can exploit both stimulus-related activity but also activity related to other spatially correlated brain processes to discriminate the attended and unattended signals. Instead, a forward model that maps in the opposite direction from the stimulus input to the EEG responses can be used to characterize the stimulus-evoked cortical response (Di Liberto et al., 2015; Ding and Simon, 2012a,b; Lalor et al., 2009a; Power et al., 2012). We used this approach to find filters, referred to as temporal response functions (TRFs), that map linearly from the time-lagged speech envelope $S(t)$ to the EEG channel responses $R(t,n)$:

$$\hat{R}(t,n) = \sum_{l=1}^{L} h(\tau_L, n) S(t-\tau_l) \tag{4.2}$$

The weights of the temporal response function $h(\tau_l, n)$ can be viewed as an averaged evoked-response elicited by the continuous speech stimulus over a finite range of latencies $\tau = \{\tau_1, \tau_2, ...\tau_L\}$ at electrode n. We estimated TRFs both for the attended and unattended speech as well as for the clean and distorted signals. As for the backward reconstruction models, we estimated the filter coefficients by ridge regression. The ridge parameter was again tuned to give highest cross-validated prediction accuracies (as indexed by Pearson's correlation between estimated and recorded EEG responses) on the single-talker EEG data recorded from each subject. All time-lags ranging from 0 ms to 500 ms post stimulus were considered for this analysis.

Evaluating the significance of decoding accuracies

Permutation tests were used to assess whether the decoding accuracies differed significantly from chance-level (Noirhomme et al., 2014). To generate a null-distribution of random classification accuracies, we randomly permuted the labels of the target talker (attended/unattended), fitted reconstruction filters and computed leave-one-out decoding accuracies. This procedure was repeated $m = 2048$ times for each subject in each listening condition. From the obtained distribution, we estimated p-values as $p = (b+1)/(m+1)$, where b is the number of elements in the null distribution exceeding the actual observed decoding accuracy (Phipson and Smyth, 2010). We considered results to be statistically significant if $p < 0.05$. We employed a permutation test procedure rather than binomial tests (Mirkovic et al., 2015; O'Sullivan et al., 2014) since parametric tests can yield biased estimates of significance for cross-validated classification accuracies (Noirhomme et al., 2014).

Statistical analysis

The accuracies of the EEG attention decoding and the behavioral comprehension data were analyzed using generalized linear mixed models (GLMMs) with logit link functions using the lme4 package in the R environment (Bates et al., 2014). The models all included random effects of subjects on model intercepts and the different listening environments (i.e., two-talker anechoic, two-talker mild reverberation, two-talker high reverberation and 8-talker mild reverberation) as fixed effects. The model reported in section 3.3 included both listening environments and decoder type (attended decoder or unattended decoder) to examine potential interactions between attention and the acoustic conditions. To obtain p-values for the statistical significance of the fixed factors of interest, we compared nested models (models with and without the given factor) using likelihood ratio tests, as sample sizes in all cases were large relative to the number of factors being tested.

4.3 Results

4.3.1 Behavioral results

Overall, the subjects responded accurately to the comprehension questions in all conditions (Figure 4.1B). Importantly, there were no statistical differences

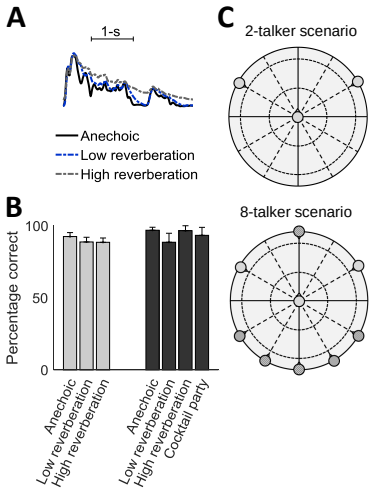

Figure 4.1: (A) Envelope of a single, representative speech signal in the considered listening environments with varying amounts of reverberation. (B) Percentage of correctly answered comprehension questions related to the content of the attended stories in the different listening conditions presented via headphones (gray) or loudspeaker array (black). Error bars represent upper 95%-confidence limits estimated by bootstrap sampling. (C) Schematic of source-receiver configurations in the two-talker and 8-talker listening scenarios. The listener was placed in the center 2.4 meters from the concurrent talkers. The two target talkers were positioned $\pm 60^o$ along the azimuth direction relative to the listener. In the 8-talker scenario, multi-talker babble from six additional talkers (represented here as dashed circles) were presented at a lower sound pressure level.

between the different acoustic conditions in terms of the percentage of correctly answered questions in either group of participants (Bonferroni corrected; headphone listening: $\chi^2(2) = 4.77$, $p > 0.18$; loudspeaker listening: $\chi^2(3) = 7.61$, $p > 0.11$).

4.3.2 Neural decoding of attended speech in reverberant environments

Figure 4.2 shows the degree to which the sound envelope of the attended speech signals could be decoded from single-trial, 64-channel EEG data in the different acoustic environments. The results shown in Figure 4.2 were all obtained using decoders trained on the attended speech streams. The figure shows (A) attention decoding accuracies and (B-C) reconstruction accuracies obtained with decoders trained to reconstruct either the envelope of the speech streams distorted by reverberation (gray bars) or the envelope of their clean versions (white bars). Across the different listening conditions, envelope reconstructions (Figure 4.2B-C) were consistently more correlated with the attended (r_{attended}) relative to

the unattended ($r_{\text{unattended}}$) speech envelopes (paired t-test, $t(145) = 34.86$, $p < 10^{-5}$). This resulted in equally high decoding accuracies (Figure 1A) in the different acoustic scenarios (grand mean: 87.1 % correct, SD=8.8 %). Linear mixed-effects models revealed that there was no main effect of the different acoustic conditions and decoding strategies (decoding on clean or distorted envelopes) on the decoding accuracies, neither for the subjects listening to speech presented over the multi-loudspeaker array ($\chi^2(6) = 9.17$, $p > 0.33$, Bonferroni corrected) nor for the subjects listening to the acoustic simulations through earphones ($\chi^2(4) = 1.91$, $p = 1$, Bonferroni corrected).

In the listening conditions with no or low reverberation and in the condition with multiple competing talkers, the reconstructed envelopes correlated equally well with the envelopes of the original 'clean' speech streams as with those of the speech streams distorted by reverberation. In the highly reverberant condition, however, the neural reconstructions of attended speech streams were more strongly correlated with the envelopes of the clean speech streams than with the envelopes of the reverberant speech streams (paired t-tests, Bonferroni corrected; earphone presentation, $t(359) = -9.82$, $p < 10^{-5}$ loudspeaker presentation $t(113) = -3.89$, $p < 0.0017$).

4.3.3 Effects of attention on noise-robust speech processing

To further investigate potential interactions between selective auditory attention and the different room conditions, we compared the performance of attended and unattended decoders trained at different stimulus-response latencies. Figure 4.3(A-B) shows the scalp topographies of the reconstruction filter weights used to predict envelopes of either the attended (A) or unattended (B) speech streams. At latencies in the 150-200 ms range, the attended reconstruction filters all exhibit positive weights at centro-temporal electrodes bilaterally, consistent with previous findings (Mirkovic et al., 2015; O'Sullivan et al., 2014). The unattended reconstruction filters show the opposite pattern of the attended decoders, with negative weights at the same positions for similar late time lags (cf. O'Sullivan et al., 2014). Panel (C-D) show the classification rates obtained with attended (C) and unattended (D) decoders at individual time-lags. While a peak in decoding performance was observed at latencies near 188 ms in all acoustic conditions for the attended decoder, a similar peak was only observed in the anechoic condition for the unattended decoder.

Fixed-effects analyses of the decoding accuracies at the location of the peak

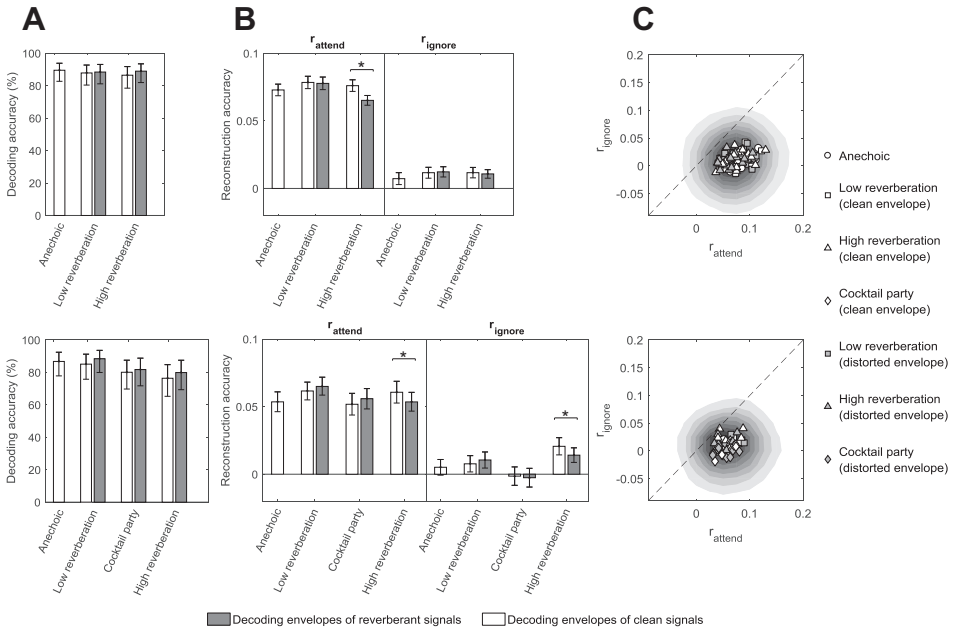

Figure 4.2: Reconstruction of attended speech envelopes from single-trial EEG responses to speech mixtures in different acoustic conditions: anechoic, low reverberation, low reverberation 8-talker cocktail party, and high reverberation. Top row: Results based on headphone presentation. Bottom row: Results based on multi-loudspeaker simulations. (A) Decoding accuracies based on neural reconstructions of the envelopes of the reverberant speech signals (gray) and the envelope of the corresponding clean signals (white). The error bars show 95%-confidence intervals of the mixed effects analyses. (B) Correlations between the neural reconstructions and the envelope of the attended (r_{attended}) and ignored ($r_{\text{unattended}}$) speech streams. The height of each bar represents the average correlation across all subjects. Error bars represent 95%-confidence limits on the mean estimated by bootstrap sampling. Asterisks indicate significant differences ($p < 0.05$, paired t-tests, Bonferroni corrected). (C) Comparison of reconstruction accuracies for each participant for the attended and unattended speech. The contours depict the empirically estimated probability distribution ($r_{\text{attended}}, r_{\text{unattended}}$) across listening conditions. For each trial, the attended speaker was considered to be correctly classified if $r_{\text{attended}} > r_{\text{unattended}}$. The dashed lines depict $r_{\text{attended}} = r_{\text{unattended}}$.

(188 ms) revealed significant interactions between attention (attended/unattended decoder) and acoustic condition (anechoic, mild reverberation, high reverberation), both for decoding based on clean envelopes ($\chi^2(2) = 9.26$, $p < 0.0196$, Bonferroni corrected) and for decoding based on reverberant envelopes ($\chi^2(2) = 11.58$, $p < 0.006$, Bonferroni corrected). Post hoc pairwise comparisons revealed that the decoding accuracies for the unattended decoders were significantly higher in the anechoic conditions compared to the mildly reverberant condition (clean envelopes: $z = 3.62, p < 0.0036$, Bonferroni corrected; reverberant envelopes: $z = 4.27, p < 0.00023$, Bonferroni corrected) and compared to the highly

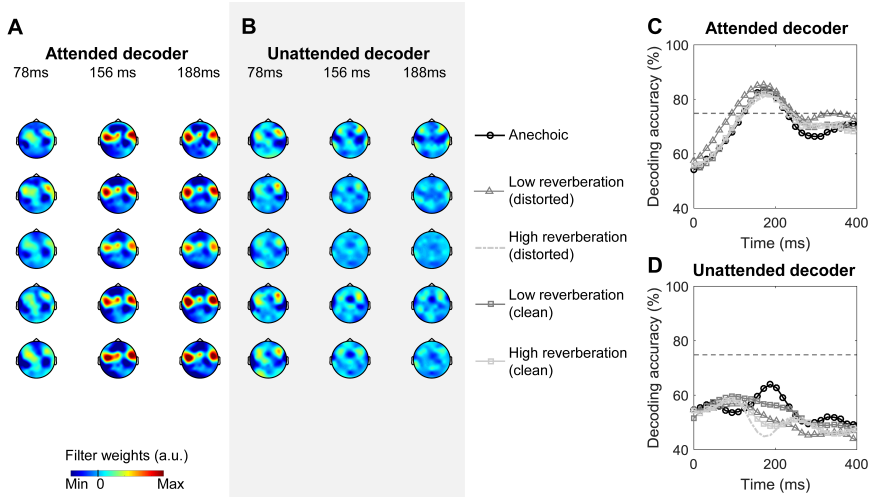

Figure 4.3: Neural decoding of attended and unattended speakers at different time-lags. Topographies of the lag-specific attended (A) and unattended (B) decoder weights averaged across subjects at individual post-stimulus latencies. Average decoding accuracies at individual stimulus-response latencies are shown for the attended (C) and unattended (D) decoders. In contrast to decoding of the attended speech, decoding of the unattended speech was significantly affected by the acoustic environment. The dashed horizontal line indicates the chance-level decoding estimated as the 95% upper limit of the permutation distribution across conditions at 188-ms.

reverberant conditions (clean envelopes; $z = 6.45$, $p < 10^{-5}$, Bonferroni corrected; reverberant envelopes; $z = 7.12$, $p < 10^{-5}$, Bonferroni corrected), while the attended decoders did not differ in accuracy between any of the three listening conditions. However, the decoding accuracies obtained with lag-specific unattended decoders at 188 ms were statistically significant ($p < 0.05$, permutation test) for 10 (of 26) subjects in the anechoic condition, for only 4 subjects in the mildly reverberant conditions, and for 2 subjects in the highly reverberant conditions. In contrast, the lag-specific (188 ms) attended decoders yielded statistically significant classification accuracies ($p < 0.05$, permutation test) for 21 subjects in the anechoic condition, for 24 subjects in the mildly reverberant condition and for 22 in the highly reverberant condition.

4.3.4 Temporal response functions

The envelope reconstruction technique revealed noise-robust cortical representations of the individual speech streams but does not provide a measure of how the individual speech streams are transformed into cortical activity. To characterize the average cortical response to the individual speech streams

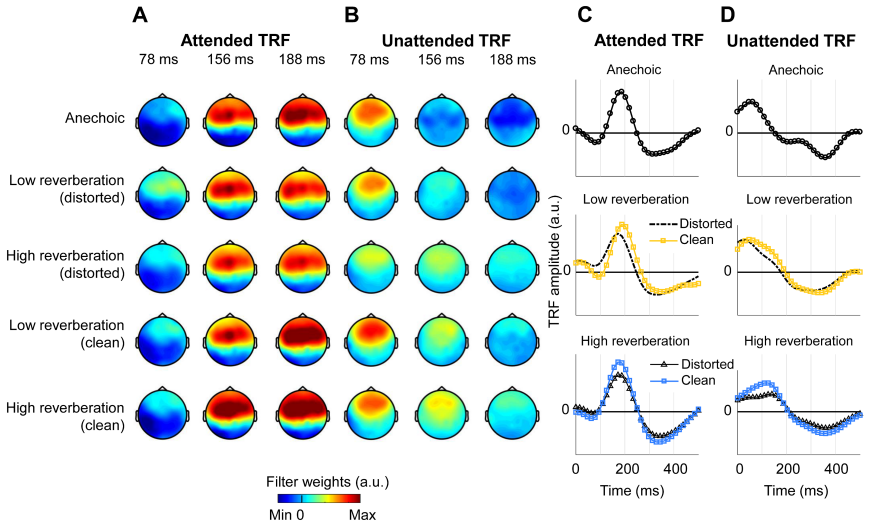

Figure 4.4: Results of the TRF analysis. A-B: Scalp topographies of the grand average TRFs at different response latencies for the attended (A) and unattended (B) speech. C-D: The TRFs averaged over all electrodes for the attended (C) and the unattended (D) speech streams in the different listening conditions. In the reverberant listening scenarios, the TRFs are shown both for the envelopes of the reverberant streams (black solid lines) and for the envelopes of their clean versions (dashed lines).

more directly, we estimated TRFs for the different listening conditions (Figure 4.4). All of the attended TRFs had clear late response peaks at latencies near 188 ms, which were prominent over frontal electrodes. In contrast, the unattended TRFs contained earlier peaks near 78 ms, but no distinct late peak. To assess differences in the cortical response to the clean versus distorted envelope, we analyzed differences in the TRF amplitudes at the position of the late peak. In both reverberant rooms, we found a significantly enhanced response to the clean envelope of the attended speech signal compared to the distorted signal (paired t-tests, Bonferroni corrected; low reverberation, $t(25) = 4.28$, $p < 0.0019$; high reverberation $t(25) = 3.78$, $p < 0.007$). The early peak of the unattended TRF was significantly higher for the clean envelope in the highly reverberant room (paired t-test, Bonferroni corrected; high reverberation $t(25) = 4.73$, $p < 0.0006$) but not in the mildly reverberant room ($t(25) = 2.47$, $p > 0.17$).

4.3.5 Effect of electrode number and trial duration on decoding accuracy

To investigate which sensors contributed to decoding, we iteratively removed electrodes and evaluated the decoder performance with a smaller subset of electrodes and shorter durations of test segments (Mirkovic et al., 2015). For this analysis, we focused on envelope decoding of the actual attended speech streams (with distortion in the case of the reverberant rooms). Figure 4.5A shows the average decoding performance across listening conditions as a function of the number of electrodes and the test segment duration. Fewer electrodes resulted in a slight improvement in performance with a peak in accuracy around 20 electrodes. The decoding accuracies remained statistically significant ($p < 0.05$, permutation testing) for all subjects with only 10 electrodes and 10-s long test segments in the anechoic and mildly reverberant listening conditions. Decoding based on 10 electrodes and 10-s segments was significantly above chance in all except two subjects in the highly reverberant room and for all except one subject in the 8-talker cocktail party condition. Figure 4.5B shows the topography of electrodes retained at different selection steps. The selection procedure suggested strong contributions to attention decoding from temporal electrodes bilaterally, consistent with the spatial patterns of the decoder weights at late stimulus-response latencies (Figure 4.3A).

4.4 Discussion

In the present study, we examined the effects of selective attention on cortical entrainment to competing speech streams in different real-life listening scenarios. Using a stimulus reconstruction approach, it was shown that cortical tracking of an attended speech source was robust to convolutive and additive distortions in the acoustic input. In the considered listening scenarios, speech comprehension was found to be unaffected even in the reverberant scenarios with multiple interfering talkers, which was mirrored by robust cortical envelope tracking and accurate attentional decoding from single-trial EEG data.

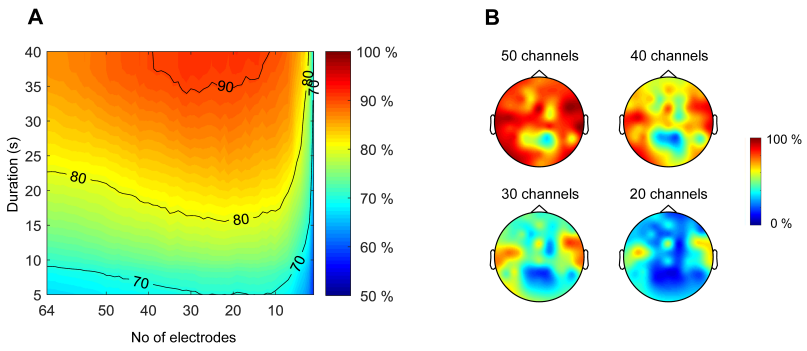

Figure 4.5: Decoding accuracies (percentage correctly classified data segments) as a function of the number of EEG electrodes and duration of the EEG test data segments. (A) Group-mean decoding accuracies across all subjects and all listening conditions. (B) Topographies illustrating the relative number of retained electrodes across all subjects and listening conditions at different iterations of the backward channel-selection. 100 % represents that all of the attended reconstruction filters contained a given electrode at a given iteration. Across subjects, centro-temporal electrodes bilaterally were most frequently retained.

4.4.1 Robust cortical representations of attended speech in different listening environments

In this study, we considered the speech envelope reconstructed from neural responses as a measure of cortical speech tracking. In the highly reverberant environments, we found that the neural response tracked the underlying clean speech signal more accurately than the presented speech stimulus arriving at the ears of the listener (Figure 4.2B). Different reconstruction accuracies for the clean and the distorted signals emerged only in the highly reverberant condition, most likely due to the fact that the clean and distorted envelopes are themselves highly correlated at low or moderate levels of reverberation. An enhanced cortical representation of the clean signal, was supported by the TRF analysis showing significantly higher TRF response amplitudes for the clean signal in both the mildly and highly reverberant room. Accurate neural representation of the original clean speech signal is consistent with stimulus reconstruction results obtained with population responses reported in animal physiology, suggesting that speech signals distorted by reverberation or stationary noise are represented in a noise-invariant manner at the level of the auditory cortex (Mesgarani et al., 2014a). In the current study, simulations of real-life scenes provided acoustic cues for sound source segregation that al-

lowed listeners to maintain equally high speech comprehension levels with the different distorted inputs. A number of previous studies have reported acoustic distortions to reduce cortical speech tracking in humans (Ding and Simon, 2013; Ding et al., 2014; Doelling et al., 2014; Peelle et al., 2012; Vander Ghinst et al., 2016) and to decrease the differential responses between attended and unattended signals (Kong et al., 2015; Rimmele et al., 2015). However, reduced responses in the theta-band were also correlated with reduced speech intelligibility created by these distortions. This has been proposed to indicate that cortical entrainment to speech mainly reflects rhythmic variations in high-level speech features, such as phoneme onsets, that typically co-vary with acoustic envelope fluctuations (Luo and Poeppel, 2007; Zoefel and VanRullen, 2015). In our study, robust tracking of intelligible but distorted signals is consistent with the notion that intelligibility plays an important role for reliable cortical tracking of speech (Peelle et al., 2012). However, although cortical envelope tracking is enhanced during comprehension, envelope tracking has also been observed with unintelligible speech stimuli (Howard and Poeppel, 2010; Millman et al., 2015), non-speech sounds (Lalor et al., 2009a), or with speech sounds in non-human primates (Steinschneider et al., 2013). Reduced speech-tracking responses observed with distorted speech stimuli in previous studies (Kong et al., 2015; Rimmele et al., 2015) may also be attributed to the loss of acoustic cues hindering succesful segregation of speech streams. Stream segregation is, in its own right, a requirement for object-based selective attention and thus for successful speech comprehension in complex auditory scenes. Noise-robust cortical representations of the attended speech signal found in the current study could thus be attributed to successful formation and selection of auditory objects enabled by real-world acoustic cues.

The ability to comprehend the target speech may, however, provide additional cues that facilitate noise-robust tracking of the speech envelope. It has been proposed that cortical speech entrainment mainly reflects listeners' ability to parse high-level speech units at multiple time scales (e.g., phonemes, syllables, phrases) that often co-vary with envelope fluctuations (Ghitza, 2011; Ghitza and Greenberg, 2009; Giraud and Poeppel, 2012; Luo and Poeppel, 2007; Peelle and Davis, 2012; Zoefel and VanRullen, 2015). Although we interpret the observed enhanced responses to the original clean envelope to indicate a noise-robust cortical representation of the acoustic signal, this does not rule out a role of active tracking of higher-level speech features. Intelligible speech

provides important contextual information that allows listeners to anticipate the onset of forthcoming speech events (Golumbic et al., 2013; Peelle et al., 2012). The predictability of an intelligible sound stream has been proposed to enable a phase alignment of neuronal excitability with points in time where the attended signal is expected (Giraud and Poeppel, 2012; Lakatos et al., 2005, 2008; Peelle et al., 2012; Schroeder and Lakatos, 2009). In complex scenarios with competing sound sources, the presence of contextual linguistic information that allows listeners to predict forthcoming events may thus modulate oscillatory activity in auditory cortex to ensure robust encoding of relevant acoustic information (Golumbic et al., 2013; Schroeder and Lakatos, 2009). However, the precise nature of potential top-down contributions of speech comprehension on cortical envelope tracking cannot be determined from the current study.

Although our results suggested cortical representations that resembled the clean signals more than the reverberant input, we did not find significant differences in attention decoding accuracies based on clean or distorted signals. A possible explanation for this may be that sufficient information about the attended speech streams was already captured by the distorted speech envelopes for single-trial attention decoding. Yet, we also found that the unattended signal was only decodable in the anechoic condition, in contrast to the equally high decoding accuracies observed for the attended speech across all acoustic conditions (Figure 4.3D). This might suggest that attention additionally promotes noise-robust envelope tracking. We note that the low reconstruction accuracies reported in figure 4.2B for unattended speech in anechoic conditions were based on decoders trained to identify the attended talker. In contrast, the decoding accuracies reported in Figure 4.3D were obtained with decoders trained to identify the unattended talker. A higher decoding accuracy in the anechoic condition here implies that the unattended decoders could extract the unattended speech stream separately from the attended speech stream. Moreover, the unattended speakers were only decodable in anechoic conditions at later latencies (188 ms, in accord with (O'Sullivan et al., 2014)), but the TRFs suggested that envelope fluctuations in the unattended speech stream elicited early (78 ms) positive responses. This could suggest that the unattended decoders exploit an active suppression mechanism for unattended sounds in multi-talker scenarios occurring at later processing stages (Kong et al., 2014) and that the early positive peaks 78 ms post stimulus do not sufficiently discriminate the competing speech streams.

4.4.2 Decoding of attended speech with single-trial EEG responses

The classification accuracies obtained with the attended decoders in the different acoustic environments were of the same order of magnitude as those previously reported with two-talker stimuli in 'clean' anechoic conditions (Mirkovic et al., 2015; O'Sullivan et al., 2014). Here we achieved a comparable decoder performance with individualized decoders and shorter segments of test data (<40-s, see Figure 5). We observed an increase of the decoding accuracy when iteratively removing electrodes down to about 25 electrodes. This could be attributed to the removal of noisy or "irrelevant" electrodes that did not contribute significantly to predicting the target envelopes. With high-dimensional EEG data, ridge-regularized regression shrinks the size of many filter coefficients (particularly in low-variance directions) to prevent overfitting, rather than enforcing filter weights to be zero. By iteratively removing channels and re-fitting the filters, the effects of single poor channels on model fits will therefore become less prominent. In O'Sullivan et al., 2014, speech mixtures were presented dichotically and listeners were instructed to attend to the speech stream presented in one ear while ignoring the speech stream presented to the other ear. Similarly, in Mirkovic et al., 2015 the subjects were presented with spatially separated ($\pm 30^o$ along the azimuth) speech streams and instructed to selectively direct attention to the same speaker throughout the experiment. The speaker locations were then balanced out across subjects rather than within subjects. Since spatial auditory attention may have distinct spatio-spectral signatures in the EEG (Kerlin et al., 2010) that can be modulated in synchrony with speech modulations (Wöstmann et al., 2016), the fixed spatial position of the attended talker could potentially bias the decoders. In the present study, we randomized both position and gender of the target speakers across trials within subjects to remove potential biases of talker identity and spatial positions on the decoders. Our results suggest that the attended decoders generalize well across characteristics of the individual talker and spatial positions of target speakers.

4.5 Conclusion

We found noise-robust and stable neural representations of attended speech in real-life acoustic scenes where the target speech remained highly intelligible despite signal distortions. Using a stimulus-reconstruction filtering technique,

we found that sufficient information about attended talkers could be extracted from single-trial EEG recordings to decode the attentional selection of listeners in multi-talker scenarios. Our results suggest that the selective cortical tracking of competing speech streams forms neural representations of attended speech that are stable against interfering signals and varying amounts of reverberation.

4.6 Funding

This work was supported by the EU H2020-ICT grant number 644732 (COCOHA: Cognitive Control of a Hearing Aid). J.H. was supported by the Oticon Centre of Excellence for Hearing and Speech Sciences.

5

A Comparison of Regularization Methods in Forward and Backward Models for Auditory Attention Decoding[a]

Abstract

The decoding of selective auditory attention from noninvasive electroencephalogram (EEG) data is of interest in brain computer interface and auditory perception research. The current state-of-the-art approaches for decoding the attentional selection of listeners are based on linear mappings between features of sound streams and EEG responses (forward model), or vice versa (backward model). It has been shown that when the envelope of attended speech and EEG responses are used to derive such mapping functions, the model predictions can be used to discriminate between attended and unattended talkers. However, the predictive performance of the models is dependent on how the model parameters are estimated. There exist a number of model estimation methods that have been published, along with a variety of datasets. It is currently unclear if any of these methods perform better than others, as they have not yet been compared side by side on a single standardized dataset in a controlled fashion. Here, we present a comparative study of the ability of different estimation methods to classify attended speakers from multi-channel EEG data. The performance of the model estimation methods is evaluated using different performance metrics on a set of labeled EEG data from 18 subjects listening to mixtures of two speech streams. We find that when forward models predict the

[a] This chapter is based on: Wong, D. E.; Fuglsang, S. A.; Ceolini, E.; Hjortkjær, J. H.; Slaney, M.; de Cheveigné, A. (2018). "A Comparison of Regularization Methods in Forward and Backward Models for Auditory Attention Decoding". Frontiers in Human Neuroscience

EEG from the attended audio, regularized models do not improve regression or classification accuracies. When backward models decode the attended speech from the EEG, regularization provides higher regression and classification accuracies.

5.1 Introduction

A fundamental goal of auditory neuroscience is to understand the mapping between auditory stimuli and the cortical responses they elicit. In magneto/electroencephalography (M/EEG) studies, this mapping has predominantly been measured by examining the average cortical evoked response potential (ERP) to a succession of repeated short stimuli. More recently, these methods have been extended to continuous stimuli such as speech by using linear system-response models, broadly termed 'temporal response functions' (TRFs), that are estimated using system identification methods. The TRF is a stimulus-response model that characterizes how a unit impulse in an input feature corresponds to a change in the M/EEG data. TRFs can be used to generate continuous predictions about M/EEG responses as opposed to characterizing the response (ERP) to repetitions of the same stimuli. Importantly, it has been demonstrated that the stimulus-response models can be extracted both from EEG responses to artificial sound stimuli (Lalor et al., 2009a; Lalor et al., 2006; Power et al., 2011) but also from EEG responses to naturalistic speech (Lalor and Foxe, 2010). A number of studies have considered mappings between the slowly varying temporal envelope of a speech sound signal (<10 Hz) and the corresponding filtered M/EEG response (Ding and Simon, 2012a,b, 2013, 2014; Lalor and Foxe, 2010). However, TRFs are not just limited to the broadband envelope, but can also be obtained with the speech spectrogram (Ding and Simon, 2012a,b), phonemes (Di Liberto et al., 2015), or semantic features (Broderick et al., 2018). This has opened new avenues of research into cortical responses to speech, advancing the field beyond examining responses to repeated isolated segments of speech.

TRF methods have proven particularly apt for studying how the cortical processing of speech features are modulated by selective auditory attention. A number of studies have considered multi-talker 'cocktail party' scenarios, where a listener attends to one speech source and ignores others. It has been demonstrated that both attended and unattended acoustic features can be linearly mapped to the cortical response (Ding and Simon, 2012a,b; Golumbic

et al., 2013; Power et al., 2012; Puvvada and Simon, 2017).

Conversely, the same linear model, which maps speech features to the cortical response (forward direction), can be adapted to provide a linear mapping from the cortical response to the speech features (backward direction) (Bialek et al., 1990; Ding and Simon, 2012a,b; Fuglsang et al., 2017; Mesgarani and Chang, 2012; Mesgarani et al., 2009; Mirkovic et al., 2015; O'Sullivan et al., 2014; Van Eyndhoven et al., 2017). The mapping from acoustic features to cortical responses is typically referred to as a forward model (or TRF), whereas the mapping from cortical responses to acoustic features is referred to as a backward model (Haufe et al., 2014). The quality of model fit reflects the degree to which cortical activity is driven by stimulation. In a cocktail party scenario, the quality of fit between each of the speech streams and the cortical activity can be used to infer which speech stream is being attended. Differences in the accuracy of forward/backward model-derived estimates between the attended and unattended speech signal can be used to predict or 'decode' to whom a listener is attending based on unaveraged M/EEG data. Single-trial measures of auditory selective attention in turn suggests BCI applications, for instance, for cognitively-steered hearing aids (Das et al., 2016; O'Sullivan et al., 2017b; Van Eyndhoven et al., 2017; Zink et al., 2017).

The ability of forward/backward stimulus-response models to generalize to new data is generally limited by the need to estimate a relatively large number of parameters based on noisy single-trial M/EEG responses. Like many aspects of machine learning, this necessitates regularization techniques that constrain the model coefficients to prevent overfitting (Crosse et al., 2016b; Holdgraf et al., 2017). A number of methods for regularizing the forward/backward stimulus-response models have been presented in various studies (David et al., 2007; Goutte et al., 2000; Machens et al., 2004; Theunissen et al., 2000, 2001; Thorson et al., 2015). Each of these methods attempt to address the challenge of having sufficient data to compute a reliable stimulus-response mapping function. To reduce the data requirement, regularization can be applied in the form of a smoothness and/or sparsity constraint.

To date, little work has been done to compare these methods against each other. A meta-analysis would be difficult as many variables, such as subjects, stimuli and data processing are different between each study. The present paper

uses a standardized publicly available dataset[a] (Fuglsang et al., 2018), based on the attended-versus-unattended talker discrimination task, as well as preprocessing and evaluation procedures to compare these algorithms. In addition, the present paper examines the relationship between different evaluation metrics to highlight their similarities and differences. The methods for computing forward/backward stimulus-response models have been implemented in the publicly available Telluride Decoding Toolbox[b].

5.2 Material and Methods

Temporal response functions can be used to predict the EEG response to a multi-talker stimulus from the attended speech envelope or, alternatively, the equation can be adapted to reconstruct the attended speech envelope from the EEG response. The first case is denoted as a "forward model" (as it maps from speech features to neural data) and the second as a "backward model" (as it maps from neural data back to speech features) (Haufe et al., 2014).

5.2.1 Stimulus-Response Models

The linear stimulus-response models below described below map a matrix $\mathbf{X}$ (stimulus features for a forward model, EEG for a backward model) to a matrix $\mathbf{Y}$ (EEG channels for a forward model, stimulus features for a backward model):

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{W}, \tag{5.1}$$

where $\mathbf{X} = [x_{t,(f,c)}]$ is a multichannel data matrix (channels indexed by c),augmented to include time-lagged versions of the data (lags indexed by f), and $\hat{\mathbf{Y}} = [y_t]$ is the model estimate in the form of a vector indexed by time t. Time lags, limited to a range such as -500 to + 500 ms, allow the model to handle delays and convolutional mismatch between $\mathbf{X}$ and $\mathbf{Y}$. Dimensions c and f are combined when performing matrix multiplications.

In the following subsections we introduce different approaches to estimating the linear model parameters, $\mathbf{W}$. Each method uses different regularization techniques to optimize the generalizability of the mapping functions.

[a] http://doi.org/10.5281/zenodo.1199011
[b] http://www.ine-web.org/software/decoding

Ordinary Least Squares (OLS)

The cost function that is minimized when solving the regression model is:

$$\mathscr{L}(\mathbf{W}) = (\mathbf{Y}-\mathbf{X}\mathbf{W})^T(\mathbf{Y}-\mathbf{X}\mathbf{W}). \tag{5.2}$$

The filter coefficients of this model can be estimated via ordinary least squares:

$$\mathbf{W} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}, \tag{5.3}$$

where $\mathbf{X}^T\mathbf{X}$ is the estimated autocovariance matrix and $\mathbf{X}^T\mathbf{Y}$ is the estimated cross-covariance matrix. The ordinary least-squares solution was here estimated using the Cholesky decomposition method, via the *mldivide* routine in Matlab. One advantage of the OLS estimator is that it has no additional hyperparameters that must be optimized. However, in practice the OLS estimator is often outperformed by the regularized solutions described in the following subsections. This is often the case when the regressor, $\mathbf{X}$, is high-dimensional and has a poorly estimated covariance matrix given limited amounts of training data, or contains auto-correlations and/or cross-channel correlations resulting in a low rank matrix. In other words, the inverse problem is ill-posed. Such is the case when using non-stochastic data for $\mathbf{X}$, such as speech or EEG data.

If $\mathbf{X}$ were white and standardized, the autocovariance matrix would be a multiple of the identity matrix, and the OLS and regularized approaches reduce to a straight-forward cross-correlation, also known as reverse correlation (Ringach and Shapley, 2004).

Ridge

Ridge regression minimizes the residual sum of squares, but puts an $L2$ constraint on the regression coefficients (Broderick et al., 2018; Crosse et al., 2016a; Crosse et al., 2015; Di Liberto et al., 2015; Holdgraf et al., 2016; Machens et al., 2003; O'Sullivan et al., 2017a). An $L2$ constraint smooths the regression weights by penalizing the square of the weights in $\mathbf{W}$ with a regularization constant λ for the ridge regression cost function:

$$\mathscr{L}(\mathbf{W})_\lambda = (\mathbf{Y}-\mathbf{X}\mathbf{W})^T(\mathbf{Y}-\mathbf{X}\mathbf{W}) + \lambda\mathbf{W}^T\mathbf{W} \tag{5.4}$$

(Hastie et al., 2001; Machens et al., 2004). Ridge regression corresponds to

imposing a Gaussian prior on the filter coefficients (Wu et al., 2006). The ridge solution is:

$$\mathbf{W} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{Y}, \tag{5.5}$$

where λ is the regularization parameter that controls the amount of parameter shrinking.

Low-Rank Approximation (LRA)

The LRA-based regression relies on a low-rank approximation of the covariance matrix, $\mathbf{X}^T\mathbf{X}$. This is achieved by employing a singular value decomposition (SVD) of $\mathbf{X}^T\mathbf{X}$:

$$\mathbf{X}^T\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T, \tag{5.6}$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthonormal matrices that contain respectively the left and right singular vectors, and where $\mathbf{S}$ is a diagonal matrix, $\mathbf{S} = \text{diag}(s_1, s_2, ..s_d)$ with sorted diagonal entries. Since $\mathbf{X}^T\mathbf{X}$ is a positive semidefinite matrix we have $\mathbf{U} = \mathbf{V}$. LRA uses a rank-K approximation of $\mathbf{X}^T\mathbf{X}$ by only retaining the first $1 \leq K \leq d$ diagonal elements of $\mathbf{S}$. The cost function is:

$$\mathscr{L}(\mathbf{W})_K = (\mathbf{Y}-\mathbf{X}\mathbf{W})^T(\mathbf{Y}-\mathbf{X}\mathbf{W}) - \mathbf{W}^T\mathbf{V}_{K+1\ldots d}\mathbf{S}_{K+1\ldots d,K+1\ldots d}\mathbf{V}_{K+1\ldots d}^T\mathbf{W}, \tag{5.7}$$

where $\mathbf{V}_{K+1\ldots d}$ are the $K+1\ldots d$ columns of $\mathbf{V}$ and $\mathbf{S}_{K+1\ldots d,K+1\ldots d}$ is the square matrix formed by taking the $K+1\ldots d$ rows and columns of $\mathbf{S}$. By forming $\hat{\mathbf{S}}^{-1} = \text{diag}(1/s_1, 1/s_2, ..., 1/s_K, 0..0, 0, 0)$, the regression coefficients can be estimated from:

$$\mathbf{W} = \left(\mathbf{U}\hat{\mathbf{S}}^{-1}\mathbf{V}^T\right)\mathbf{X}^T\mathbf{Y}. \tag{5.8}$$

The number of diagonal elements, K, to retain are typically chosen such that a diagonal element is retained if the sum of the eigenvalues to be kept cover a fraction λ of the overall sum, or $0 < \dfrac{\sum_{i=1}^{K} s_i}{\sum_{i=1}^{d} s_i} < \lambda \leq 1$. Note that the regularization parameter, λ, here is analogous to λ for Ridge Regression, but that the values are not comparable between the two. LRA is the term used in systems identification (Marconato et al., 2014), however, this type of regression has also been referred

to as normalized reverse correlation (NRC) in auditory neuroscience literature (David et al., 2004, 2007; Mesgarani and Chang, 2012; Mesgarani et al., 2009; Theunissen et al., 2000, 2001).

Shrinkage

Shrinkage (Blankertz et al., 2011; Friedman, 1989) is a method used for biasing the covariance matrix by flattening its eigenvalue spectrum with some tuning parameter, λ. In the context of regression, the Shrinkage cost function is:

$$\mathscr{L}(\mathbf{W})_\lambda = (\mathbf{Y}-\mathbf{X}\mathbf{W})^T(\mathbf{Y}-\mathbf{X}\mathbf{W}) + \lambda \mathbf{W}^T(\nu\mathbf{I}-\mathbf{X}^T\mathbf{X})\mathbf{W}, \tag{5.9}$$

where ν is here defined as the average eigenvalue trace of the covariance matrix $\left(\mathbf{X}^T\mathbf{X}\right)$. The solution for the cost function is:

$$\mathbf{W} = \left((1-\lambda)\mathbf{X}^T\mathbf{X} + \lambda\nu\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{Y}. \tag{5.10}$$

When $\lambda = 0$, it becomes the standard ordinary least squares solution. When $\lambda = 1$, the covariance estimator becomes diagonal (i.e. it becomes spherical), reducing the Shrinkage equation to a cross-correlation (Blankertz et al., 2011).

These regularization schemes are related. Whereas Ridge Regression and Shrinkage both penalize extreme eigenvalues in a smooth way, LRA discards eigenvalues. Ridge and Shrinkage in other words flatten out the eigenvalue trace. Ridge shifts it up, and Shrinkage shrinks it towards an average value ν (Blankertz et al., 2011), whereas LRA cuts if off.

Tikhonov

The scheme which we shall refer to as *Tikhonov regularization*, is a first-derivative type of Tikhonov regularization (Tikhonov, 1963) that takes advantage of the fact that there is usually a strong correlation between adjacent columns of $\mathbf{X}$ when $\mathbf{X}$ includes time shifts, because of the strong serial correlation of the stimulus envelope (for the forward model) or the filtered EEG (for the backward model). In other words, Tikhonov regularization imposes *temporal smoothness* on the model. Tikhonov regularization achieves temporal smoothness by putting a constraint in the derivative of the filter coefficients (Crosse et al., 2016b; Crosse et al., 2015; Goutte et al., 2000; Lalor et al., 2006; Lalor and Foxe, 2010). Here we focus on first order derivatives of the filter coefficients and assume that the first

derivatives can be approximated by $\frac{\partial w_i}{\partial i} \approx (w_{i+1} - w_i)$ for any neighboring filter pairs w_{i+1} and w_i. This type of regularization is more generally referred to as 1st order Tikhonov regularization as it attempts to constrain the first derivative of the filter via central difference approximations. This gives the cost function:

$$\mathscr{L}(\mathbf{W})_\lambda = (\mathbf{Y} - \mathbf{X}\mathbf{W})^T (\mathbf{Y} - \mathbf{X}\mathbf{W}) + \lambda \sum_i (w_i - w_{i+1})^2 \tag{5.11}$$

Tikhonov regularized model filters can, under this approximation, be implemented as:

$$\mathbf{W} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{M}\right)^{-1}\mathbf{X}^T\mathbf{Y}, \tag{5.12}$$

where

$$\mathbf{M} = \begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}.$$

Note that cross-channel leakage can occur whenever the regressor, **X**, reflects data recorded from multiple channels, as is the case with the backward model. This means that filter endpoints can be affected by neighboring channels as a result of the off-diagonal elements in the **M** matrix. Due to the potential for cross-channel leakage, Tikhonov has been primarily used for the forward modeling case (Crosse et al., 2016b). Despite the potential problems associated with cross-channel leakage, we also report results obtained with Tikhonov regularization for the backward model for completeness.

Elastic Net

Whereas the aforementioned regularization techniques often show improvements over the ordinary least regression in terms of generalizability, they tend to preserve all regressors in the models. This can e.g. result in nonzero filter weights assigned to irrelevant features. Lasso regression attempts to overcome this issue by putting an L1-constraint on the regression coefficients (Tibshirani, 1996). This serves to drive unnecessary coefficients in the model towards zero.

Lasso has been found to perform well in many scenarios, although it was empirically demonstrated that it is outperformed by Ridge regression in nonsparse scenarios with highly correlated predictors (Tibshirani, 1996; Zou and Hastie, 2005). In such scenarios, *Elastic Net* regression (Zou and Hastie, 2005) has been found to improve the predictive power of Lasso by combining Lasso with the grouping effect of Ridge regression. The elastic net has two hyperparameters: α controlling the balance between L1 (lasso) and L2 (ridge) penalties, and λ controlling the overall penalty strength. For the purpose of this paper, we use a readily available algorithm, GLMNET (Qian et al., 2013), for efficiently computing the elastic net problem. This is a coordinate descent algorithm for solving the following problem:

$$\underset{\mathbf{W}}{\operatorname{argmin}} \frac{1}{2N}\|\mathbf{Y}-\mathbf{X}\mathbf{W}\|^2 + \lambda\left[(1-\alpha)\|\mathbf{W}\|^2/2 + \alpha\|\mathbf{W}\|\right]. \tag{5.13}$$

We used GLMNET for computing the Elastic Net solution for $\alpha = 0.25$, $\alpha = 0.50$, $\alpha = 0.75$ and $\alpha = 1.00$. We will henceforth refer the last case as the Lasso solution. The GLMNET has previously been used to estimate spectro-temporal receptive models (e.g. (Willmore et al., 2016)).

5.2.2 Evaluating Performance

Characterizing Model Fit

While the objective function of linear models is minimizing the mean-squared-error, the goodness of fit is typically analyzed in terms of Pearson's correlation between estimated and actual values for interpretability. The term *regression accuracy* will henceforth be used to characterize the goodness of fit for models trained and evaluated on attended audio features ($r_{attended}$). For forward models, regression accuracies were measured by the Pearson's correlation between the actual EEG and the EEG predicted by the attended envelope over the test folds. This was done separately for each EEG channel. Similarly, for backward models, regression accuracies were measured by the correlation between the attended envelope and its EEG-based reconstruction. The regression accuracies were computed on test folds, using the nested cross-validation scheme described in section 5.2.2. This procedure ensures that the test data is not used during any part of the training process, including hyperparameter tuning. The regression accuracies were averaged over all test folds. Other metrics for as-

sessing the predictive/reconstructive performance of the models have been previously proposed (Schoppe et al., 2016). However, for simplicity and to be consistent with previous studies (Ding and Simon, 2012a,b; O'Sullivan et al., 2014), this paper characterizes the goodness of the fit using Pearson's correlation coefficients.

In the forward case, the response at multiple EEG channels is predicted by the model. Rather than using multiple correlation coefficients to characterize the regression accuracy in this case, we chose to take the average of the correlation coefficients between the predicted channels and the actual EEG data as a validation score. We used the same metric over the test set to characterize the fit of the model. In the backward case, characterizing the fit is straightforward as the model predicts a single audio envelope that can be correlated with the attended audio envelope.

Decoding Selective Auditory Attention

Performance was also evaluated on a classification task based on the forward/backward stimulus-response model. The task of the classifier was to decide, on the basis of the recorded EEG and the two simultaneous speech streams presented to the listener (see Section 5.2.4), to which stream the subject was attending. The classifier had to make this decision on the basis of a segment of test data, the duration of which was varied as a parameter (1, 3, 5, 7, 10, 15, 20 and 30s), which will be referred to as the decoding segment length. This duration includes the kernel length of the forward/backward model (500 ms). The position of this segment of data was stepped in 1s increments throughout the evaluated data.

As described further in section 5.2.2, a nested cross-validation loop was used to tune the forward/backward stimulus-response model regularization parameter (where applicable) on training/validation data and test the trained classifier on unseen test data.

The classification relied on correlation coefficients between EEG and the attended speech, and between the EEG and the unattended speech. These correlation coefficients were computed over the aforementioned restricted time window. These coefficients were used to classify whether the subject was attending to one stream or the other. For a backward model, classification hinged merely on which correlation coefficient was largest (stream A or stream B). Performance of this classifier was evaluated on the test set. For a forward model, the situation is more complex because there is one model per EEG

channel. For each of the 66 channels a pair of correlation coefficients was calculated (one each for unattended and attended streams), and this set of pairs was used to train a support vector machine (SVM) classifier with a linear kernel and a soft margin constant of 1. SVM classifiers were trained on the correlation coefficient features over the validation set that was used for hyperparameter tuning. The SVM classifier performance was finally evaluated on data from the held out test fold.

The classifier score was averaged over all test folds. In every case, the classifier trained over the entire training/validation set was tested on a short interval of data, the duration of which was varied as a parameter, as explained above. An illustration of this classification task is shown in figure 5.1.

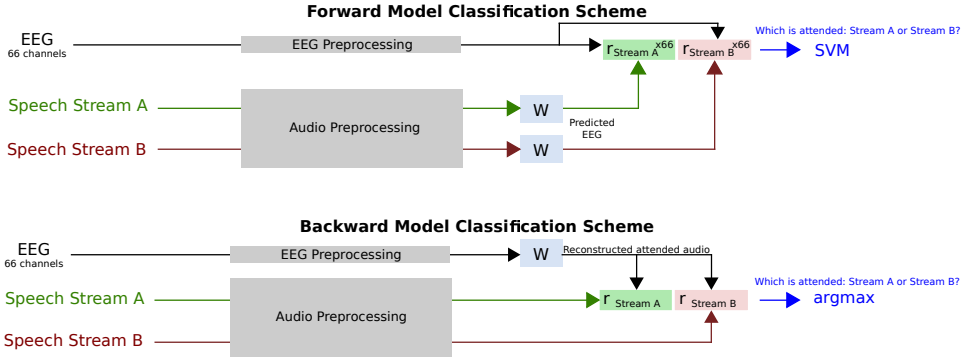

Figure 5.1: Diagram of classification task. For the forward model, 66 EEG channels are predicted from the speech stream A and B envelopes using the same linear mapping function, W. After correlation with the 66 channel EEG data, this results in 66 correlation coefficients for each speech stream, which are used as features for the SVM to distinguish the attended talker. For the backward model, a single attended audio envelope channel is estimated from the EEG data using the linear mapping function, W. After correlation with the speech stream A and B envelopes, a single correlation coefficient for each speech stream is obtained. Classification of the attended talker is performed by determining the larger coefficient.

Classification performance was characterized for different decoding segment durations using the raw classification score, receiver operating characteristic (ROC) curve, and information transfer rate (ITR). The raw classification score measured what proportion of trials were classified correctly. It should be noted that in measuring classification performance, the two classes were balanced. The ROC curve characterizes the true-positive and false-positive rates for decoding segment trials where the classifier discrimination function lies above a given threshold, as the threshold is varied. The classifier decision function is the distance between the classified point and the decision boundary, with the sign indicating the class label. In the case of an SVM classifier for the

forward model, the decision function is a weighted sum of the input features (correlations), plus a bias term. In the case of the argmax function for the backward model, the decision function is the difference of the correlations between the reconstructed attended audio and the two speech streams. Thresholding the classifier discrimination function throughout the range of values it yields in a dataset affects the number of correctly and incorrectly classified trials (above threshold) out of the total number of correctly and incorrectly classified trials, which are the true and false positive rates, respectively.

The ITR metric corresponds to the number of classifications that can be reliably made by the system in a given amount of time. The dependency of ITR on decoding segment length is a tradeoff between two effects. On one hand, longer decoding segments allow more reliable decisions. On the other, short durations allow a larger number of independent decisions. There is thus an optimal decoding segment duration. A number of metrics to compute the ITR have been proposed. The most common is the Wolpaw ITR (Wolpaw and Ramoser, 1998), which is calculated in bits per minute as:

$$ITR_W = V\left[\log_2 N + P\log_2 P + (1-P)\log_2 \frac{1-P}{N-1}\right], \tag{5.14}$$

where V is the speed in trials per minute, N is the number of classes, and P is the classifier accuracy. We also report the Nykopp ITR, which assumes that a classification decision does not need to be made on every trial (Nykopp, 2001). This can be done by first calculating the confusion matrix p for classifier outputs where the classifier decision function magnitude exceeds a given threshold. Typically the larger the classifier decision function magnitude, the more accurate the classifier prediction. As such, raising the threshold on the decision function magnitude results in more accurate classifications at the expense of foregoing a classification decision on more trials. To obtain the Nykopp information transfer rate, the threshold on the classifier decision function magnitude is adjusted to maximize:

$$ITR_N = V\left[\max_{p(x)} \sum_{i=1}^{N}\sum_{j=1}^{M} p(w_i)p(\hat{w}_j|w_i)\log_2 p(\hat{w}_j|w_i) - \sum_{j=1}^{M} p(\hat{w}_j)\log_2 p(\hat{w}_j)\right], \tag{5.15}$$

where $p(w_i)$ is the probability of the actual class being class i, $p(\hat{w}_j|w_i)$ is the probability of the predicted class being class j given the actual class being class

i, and $p(\hat{w}_j)$ is the probability of the predicted class being class j. It is $p(\hat{w}_j|w_i)$ and $p(\hat{w}_j)$ that are affected by decision function magnitude thresholding as this limits the number of trials on which a classification decision is made.

Cross-Validation Procedure

The forward/backward stimulus-response models used in sections 5.2.2 and 5.2.2 were all trained and tested using cross-validation with a 10-fold testing procedure involving nested cross-validation loops. This procedure ensures that the test data used to evaluate the forward/backward model is not used during any part of the training process. During this cross-validation procedure the models were characterized under an N-fold testing framework where the data was divided into 10 folds. In this outer cross-validation loop, one fold was held out for testing (i.e. characterizing model fit and classifying the attended stream), while data from the remaining 9 folds were used to compute the forward/backward models using an inner cross-validation loop. This inner cross-validation loop was used to tune the hyperparameters. The stimulus-response models were in all cases fit to the envelope of the attended sound streams during the training phase. The regularization parameter was swept through a range of values to evaluate its effect on the correlation coefficient between the model prediction/reconstruction and the actual measured data for each inner cross-validation fold. For Ridge and Lasso regularization schemes that allowed a regularization parameter between zero and infinity, a parameter sweep was performed between 10^{-6} and 10^{8} in 54 logarithmically-spaced steps. This was done using the following formula:

$$\lambda_n = \lambda_0 \times 1.848^n, n \in [0,53], \tag{5.16}$$

where $\lambda_0 \equiv 10^{-6}$. For LRA, Elastic Net, and Shrinkage schemes, where the regularization parameter range was between 0 and 1, a parameter sweep was performed between 10^{-6} and 1 using a log-sigmoid transfer function that compresses the values between 0 and 1 using the following iterative formula:

$$\lambda_{n+1} = \text{logsig}(\ln(\lambda_n) - \ln(1-\lambda_n) + 0.475), n \in [0,40]. \tag{5.17}$$

The hyperparameter value that yielded the maximum correlation between the model prediction/reconstruction and actual measured data, averaged across all inner cross-validation folds, was used to evaluate the test set. Using this

hyperparameter value, the weights of the models generated for each inner cross-validation fold were then averaged to generate an overall cross-validated model that could then be applied to the test set. It should be noted that for each test fold, the hyperparameter value was selected independently.

5.2.3 Implementation

The implementations of the forward/backward stimulus-response model algorithms used here are distributed as part of the Telluride Decoding Toolbox[c], specifically in the FindTRF.m function of that toolbox. Data preprocessing, model training, and evaluation were implemented with the COCOHA Matlab Toolbox[d].

5.2.4 Stimuli

A previous report gives a detailed description of the stimuli and data collection procedure (Fuglsang et al., 2017). This dataset is available online (Fuglsang et al., 2018). In brief, a set of speech stimuli were recorded by one male and one female professional Danish speakers speaking different fictional stories. These recordings were performed in an anechoic chamber at the Technical University of Denmark (DTU). The recording sampling rate was 48 kHz. Each recording was divided into 50-s long segments for a total of 65 segments.

5.2.5 Experimental Procedure

The 50-s long speech segments were used to generate auditory scenes comprising a male and a female simultaneously speaking in anechoic or reverberant rooms. The two concurrent speech streams were normalized to have similar root-mean square values. The speech stimuli were delivered to the subjects via ER-2 insert earphones (Etymotic Research). The speech mixtures were presented binaurally to the listeners, with the two speech streams lateralized at respectively -60^o and $+60^o$ along the azimuth direction and a source-receiver distance of 2.4 meters. This was achieved using nonindividualized head-related impulse responses that were simulated using the room acoustic modeling software, Odeon (version 13.02). Each subject undertook sixty trials in which they were presented the 50s-long speech mixtures. Before each trial, the subjects

[c] http://www.ine-web.org/software/decoding
[d] http://doi.org/10.5281/zenodo.1198430

were cued to listen selectively to one speech stream and ignore the other. After each trial, the subjects were asked a comprehension question related to the content of the attended speech stream. The position of the target streams as well as the gender of the target speaker were randomized across trials. Moreover, the type of acoustic room condition (either anechoic, mildly reverberant or highly reverberant) were pseudo-randomized over trials. In the analysis, data recorded from all acoustic conditions were pooled together. The reasons for doing this were twofold. Firstly, it provides sufficient data for the stimulus-response analysis. This is particularly important as insufficient data in worst case can lead to poorer model estimates (Mirkovic et al., 2016). Secondly, by using this approach we get a better idea of how well the models will generalize to different experimental conditions. This is an important practical aspect, as it gives a better estimate of how well a classifier will perform in different listening conditions (rather than just focusing on training on anechoic data and evaluating on anechoic data).

5.2.6 Data Collection

Electroencephalography (EEG) data were recorded from 19 subjects in an electrically shielded room while they were listening to the stimuli described above. Data from one subject were excluded from the analysis due to missing data from several trials. The data were recorded using a Biosemi Active 2 system, with a sampling rate of 512 Hz. Sixty-four channel EEG data (10/20-system) were recorded from the scalp. Six additional electrodes were used for recording the EEG at the mastoids, and vertical and horizontal electrooculogram (V and H-EOG). Approximately 1 hour of EEG data was recorded per subject. This study was carried out in accordance with the recommendations of 'Fundamental and applied hearing research in people with and without hearing difficulties, Videnskabsetiske komitee'. The protocol was approved by the Science Ethics Committee for the Capital Region of Denmark. All subjects gave written informed consent in accordance with the Declaration of Helsinki.

5.2.7 Data Preprocessing

EEG Data

50 Hz line noise and harmonics in the EEG data were filtered out by convolution with a $\frac{512}{50}$ sample square window (the non-integer window size was imple-

mented by interpolation) (de Cheveigné and Arzounian, 2017). The EEG data was then downsampled to 64 Hz using a resampling method based on the Fast Fourier Transform (FFT). To downsample, this method reduces the size of the FFT of the signal by truncating high frequency components. An inverse FFT is then used to restore the signal to the time domain. A 1st order detrend was performed on the EEG data to minimize filter startup artifacts. EEG data were highpassed at 0.1 Hz using a 4th order forward-pass Butterworth filter. The group delay was less than 2 samples above 1 Hz.

The joint decorrelation framework (de Cheveigné and Parra, 2014) was employed to remove eye artifacts in an automated fashion. Let $\mathbf{X} = [x_{tj}]$ be a matrix that contains EEG data from each electrode, j, for each time sample t. In this implementation, a conservative eye artifact time-point detection was first performed by computing a Z-score on 1-30 Hz bandpassed VEOG and HEOG bipolar channels and marking time samples where the absolute Z-score on either channel exceeded 4. This is similar to the eyeblink detection method implemented in the FieldTrip EEG processing toolbox (Oostenveld et al., 2011). This resulted in a subset of time samples, A, indexing the temporal locations of each EOG artifact. An artifact covariance matrix $\mathbf{R}_A = \mathbf{X}_A^T \mathbf{X}_A$ was then computed from the EEG (and EOG) data, $\mathbf{X}_A = [x_{aj}]$, at the artifact time samples $a \in A$. After using principal component analysis to whiten $\mathbf{R}_A$ and $\mathbf{R}$, the generalized eigenvalue problem was then solved for $\mathbf{R}_A\mathbf{v} = \lambda\mathbf{R}\mathbf{v}$, where $\mathbf{R} = \mathbf{X}^T\mathbf{X}$ is the covariance matrix for the entire EEG dataset. The resulting eigenvectors $\mathbf{V}$, sorted by eigenvalue, explain the maximum difference in variance between the artifact and data covariance matrices. Components corresponding to eigenvalues > 80% of the maximum eigenvalue were regressed out of the data. In practice, this 80% threshold is a conservative one, typically resulting in the removal of one or two components. Lastly, the EOG channels were removed from the data, which was then referenced to a common average over all channels.

For the forward/backward model analysis, the EEG was bandpassed between 1-9 Hz using a windowed sync type I linear-phase finite-impulse response (FIR) filter, shifted by its group delay to produce a zero-phase (Widmann et al., 2015) with a conservatively chosen order of 128 in order to minimize ringing effects. This frequency range was selected as it has been shown that cortical responses time-lock to speech envelopes in this range (O'Sullivan et al., 2014). As part of the cross-validation procedure, individual EEG channels were finally centered and standardized (Z-normalized) across the time dimension using the

individual channel mean and standard deviation of the training data. A kernel length of 0.5 s (33 samples) was used when computing the forward/backward models.

Audio Features

The forward/backward stimulus-response model estimation methods used for attention decoding attempt to characterize a relationship between features of attended speech streams and EEG activity. We calculated temporal envelope representations from each of the clean speech streams (i.e. without reverberation). We did not try to derive them from the reverberant or mixed audio data, as explored elsewhere (Aroudi and Doclo, 2017; Fuglsang et al., 2017). In trials with reverberant speech mixtures, we used envelope representations of the underlying clean signals to estimate the models. To derive the envelope representations, we passed monaural versions of both attended and unattended speech streams through a 31-band gammatone filterbank with a frequency range of 80-8000Hz(Patterson et al., 1987). The envelope of each filterbank output was calculated via the analytic signal obtained with the Hilbert transform, raised to the power of 0.3. This rectification and compression step was intended to partially mimic that which is seen in the human auditory system (Plack et al., 2008). The audio envelope was then calculated by summing the rectified and compressed filterbank outputs across channels. The audio envelope data was subsequently downsampled to the same sampling frequency as the EEG (64 Hz) using an FFT-based resampling method. The EEG and envelopes were then temporally aligned using start-trigger events recorded in the EEG. The envelopes were subsequently lowpassed at 9 Hz. As part of the cross-validation procedure, audio envelopes were finally centered and standardized (Z-normalized) across the time dimension using the mean and standard deviation of the attended speech envelope in the training data.

5.2.8 Statistical Analysis

All statistical analyses were calculated using MATLAB. Repeated-measures analysis of variance (ANOVA) tests were used to assess differences between the regression accuracies (section 5.2.2) and classification performances 5.2.2 obtained with the different forward/backward model estimation methods. Regression accuracies and classification performances for individual subjects were

averaged across folds prior to statistical comparison.

Given the non-Gaussian distribution of regression accuracies (range -1 to 1) and classification performance metrics (range 0 to 1), Fisher Z-transforms and arcsine transforms were applied to these measures, respectively, prior to statistical tests and correlations.

5.3 Results

The forward/backward stimulus-response model estimation methods introduced in Section 5.2 were used to decode attended speech envelopes from low-frequency EEG activity. The following sections analyze results with metrics of 1) regression accuracy, 2) classification accuracy, 3) receiver operating characteristic (ROC), and 4) information transfer rate (ITR). Results are shown for each of the regularization schemes, for both forward and backward models. For each regularization scheme, the regularization parameter(s) are tuned to maximize regression accuracy. These parameter values are then used for all regression and classification comparisons. Regression accuracy compares different regularization schemes in predicting/reconstructing test data using the optimal regularization parameter. Classification accuracy uses the regression accuracy values to classify the attended/unattended talker and compares the different regularization schemes in performing this task. The ROC curve visualizes the relationship between the true and false-positive rates for different classifier discrimination function thresholds. Lastly, the ITR describes the impact of decoding segment length on the bit-rate, for different points on the ROC curve.

5.3.1 Regularization Parameter Tuning

The forward/backward model estimation methods, except for the OLS method, use regularization techniques to prevent overfitting and therefore require a selection of the appropriate tuning parameters. Figure 5.7 (in section 5.6) shows the correlation coefficient between estimated (validation set) data and the actual target data (*regression accuracy*) over a range of regularization parameters. In general, there is a broad region where validation regression accuracy is flat, which peaks before quickly falling off with increasing λ. It is also apparent that the regression accuracies obtained with backward models generally are higher than those obtained with forward models.

Figure 5.8 (in section 5.6) shows regression accuracies for forward/backward models with Elastic Net penalties. Unlike the other linear models investigated in the present study the Elastic Net has two hyperparameters. The α parameter adjusts the balance between $L1$ and $L2$ penalties. Similar to the other regularization schemes, for each value of α, there is a broad range of λ values that give good correlation performance.

5.3.2 Regression Accuracy

For each regression method (and each value of α for elastic net), the forward/backward stimulus-response model was estimated and the optimum lambda estimated on the training/validation set. This optimal model was then applied to the test set, and the regression accuracy was compared between regression methods. This is shown in figure 5.2. One might expect that the averaging of prediction-response correlations across channels for the forward model may have resulted in lower regression accuracies compared to the backward model. This was demonstrating using a t-test between the forward and backward models, over all regularization schemes and subjects ($\Delta = 0.083$, $T_{107} = 17.8$, $p = 1.1 \times 10^{-33}$). However, when using maximum correlation across channels, instead of the average, for the forward model, there was still a significant difference ($\Delta = 0.045$, $T_{107} = 9.8$, $p = 9.4 \times 10^{-17}$).

For forward models, a repeated measures ANOVA with regularization method as the factor found no significant effect of regularization method on the average of correlation coefficients, even when using the average of the correlation coefficients of the 5 channels with the largest correlation coefficients for each subject. For the backward models, a similar repeated measures ANOVA, found a significant effect of regularization method on regression accuracy ($F_{(5,85)} = 78.0$, $p < 1.0 \times 10^{-16}$). Tikhonov regularization yielded a regression accuracy that was significantly greater than each of the other schemes, using a Bonferonni correction to account for the family-wise error rate ($p < 0.045$). This is contrary to the expectation that Ridge regression would outperform Tikhonov for the backward model due to the inter-channel leakage introduced by the Tikhonov kernel. Moreover, OLS had a regression accuracy that was significantly smaller than the other schemes (with Bonferonni correction, $p < 1.3 \times 10^{-10}$). This highlights the importance of regularization for the backward models.

For Elastic Net regularization, α values was characterized at 0.25, 0.5, 0.75 and 1 (Lasso) to sample different degrees of sparsity/smoothness. The value

α=0 (Ridge) was not sampled due to sub-optimal solver performance near this point. A repeated measures ANOVA analysis with factors of α and subject, using optimal λ values, showed no significant effect of α for forward models. This means that adjusting the model sparsity had no significant effect on the regression accuracy. However, a significant effect of α was found for backward models ($F_{(3,51)} = 12.4$, $p = 3.3 \times 10^{-6}$). A posthoc paired t-test with a Bonferonni correction revealed that the best regression accuracy was obtained with $\alpha = 0.25$ ($p = 6.2 \times 10^{-4}$). It was, however, noted that the average difference between regression accuracies for $\alpha = 0.25$ and $\alpha = 1$ was only 8×10^{-4}.

To obtain an estimate of the significance of the regression accuracies presented in figure 5.2, we randomized the phase of the audio data passed to the forward models, and the phase of the EEG data passed to the backward models. The goal was to provide an estimate of the correlation noise floor for the models. The models were those trained on unaltered data using each of the regularization schemes. Randomizations were performed 100 times per subject to yield an estimate of the noise floor regression accuracies. The regression accuracies were computed the same way as before. A two-sample Kolmogorov-Smirnov test conducted pairwise showed that, within subjects, the distribution of noise floor correlations were not significantly different between regularization schemes, or channels in the case of the forward model. The within-subject distributions were thus combined, and a two-sample Kolmogorov-Smirnov test was performed pairwise between subjects. No significant difference in distributions was found between subjects. As such, all distributions were combined. The 95% confidence interval of the noise floor correlations was [-0.001, 0.001] for the forward model and [-0.032, 0.032] for the backward model.

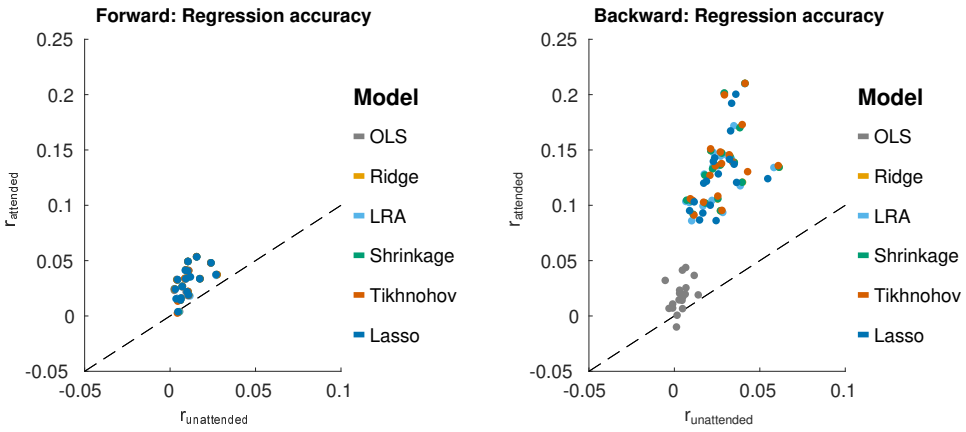

Figure 5.2: Test set regression accuracies (r_{attend}) for each forward/backward model estimation method plotted against $r_{unattend}$. Left: results from the forward modeling approach. Right: results from the backward modeling approach. For each scheme (represented by a color), each point represents average data from one subject. The black line shows $r_{attend} = r_{unattend}$.

5.3.3 Classification Accuracy

We further sought to investigate how the different forward/backward models perform in terms of discriminating between attended and unattended speech on a limited segment of data. The duration of the segment was varied as a parameter (1, 3, 5, 7, 10, 15, 20 and 30s). This was characterized on held-out test data for each TRF method, using the λ value that yielded the maximum regression accuracy in the validation data. The results from this analysis are shown in figure 5.3. A 2-way repeated measures ANOVA with factors of regularization scheme and model (forward or backward), based on 30s decoding segment lengths, found a main significant difference between backward and forward models ($F_{(1,17)} = 17.3$, $p = 6.5 \times 10^{-4}$), with a significant interaction with the effect of regularization scheme ($F_{(5,85)} = 208.9$, $p < 1.0 \times 10^{-16}$). A posthoc paired t-test showed that backward model performs better than the forward model for all regularization schemes excluding the case where ordinary least squares (OLS) was applied ($T_{17} = 9.35$, $p = 4.2 \times 10^{-8}$). For OLS, the forward model outperformed the backward model ($T_{17} = 7.32$, $p = 1.2 \times 10^{-6}$).

The interaction of the effect of regularization scheme on the classification accuracy of forward and backward models was investigated. A repeated measures ANOVA with factors of regularization scheme, applied only to the forward TRF classification accuracy scores, found no significant effect of regularization scheme on classification accuracy. This is consistent with the lack of signifi-

cant differences being detected in regression accuracies for different forward model regularization schemes, even when limiting the number of channels to 5 with the highest regression accuracies. In this case, the SVM classifier can be viewed as a data-driven approach to select channels that are most relevant to attention classification. For the backward models, however, a significant effect of regularization scheme on classification accuracy was found ($F_{(5,85)} = 229.4$, $p < 1.0 \times 10^{-16}$). A posthoc paired t-test analysis with a Bonferonni correction revealed that the classification accuracy for the OLS scheme was significantly worse than each of the others ($\bar{\Delta} = -29.1$, $p < 7.9 \times 10^{-10}$). Lasso performed significantly worse than each of the remaining schemes ($\bar{\Delta} = -1.2$, $p < 0.040$). In short, regularized backward schemes outperform OLS by a relatively large margin, as seen in figure 5.3.

For Elastic Net regularization, a repeated measures ANOVA with factors of α and subject did not find any significant effect of α on classification accuracy for forward or backward models.

In summary, for the forward model there was no difference between schemes (regularization and OLS), and for the backward model there was no difference between Ridge, Tikhonov, Shrinkage and LRA, but all regression methods were better than OLS.

Relation to regression accuracy

The discrimination between attended and unattended speech streams from EEG data is done in two stages: the computation of regression accuracies, followed by classification. We sought to investigate how the classification accuracies obtained with each model relate to the test set regression accuracies. A plot of this relationship is shown in figure 5.4.

For forward models, the average correlation between regression accuracy and classification performance across decoding segments and over all regularization schemes is 0.69 ($T_{108} = 9.83$, $p = 2.2 \times 10^{-16}$). For backward models, the correlation between the regression accuracy and classification performance is 0.89 ($T_{108} = 22.4$, $p < 1.0 \times 10^{-16}$). This suggests that classification performance varies with regression accuracy. However, as was previously described for the backward models, while Tikhonov regularization achieved a significantly higher regression accuracy compared to all other methods, it did not achieve a significantly higher classification performance compared to Shrinkage, Ridge Regression or LRA. To explain this, we examined the classification feature in terms of

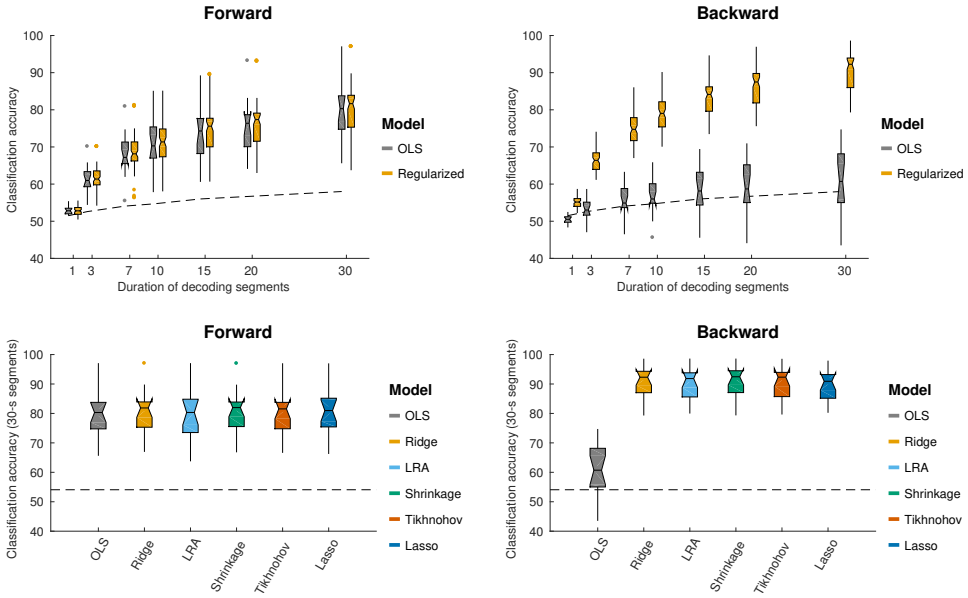

Figure 5.3: Using different forward/backward models to decode selective auditory attention from multi-channel EEG data. Classification performance is shown for different decoding segment lengths (1s, 3s, 7s, 10s, 15s, 20s, 30s). Top-left and -right panels show the classification performance for forward and backward models respectively. Performance is shown for the OLS scheme and an average across regularized schemes. Regularized schemes were averaged to concisely illustrate the higher classification accuracy obtained by these schemes compared to OLS for the backward model, but not the forward model. Bottom-left and -right panels show the classification performance for 30 s long decoding segments. The different regularization schemes are shown in different colors (see legend). Notched boxplots show median, and first and third quartiles. Whiskers show 1.5 × IQR. Dots indicate outliers. The dashed line shows the above-chance significance threshold at $p = 0.05$.

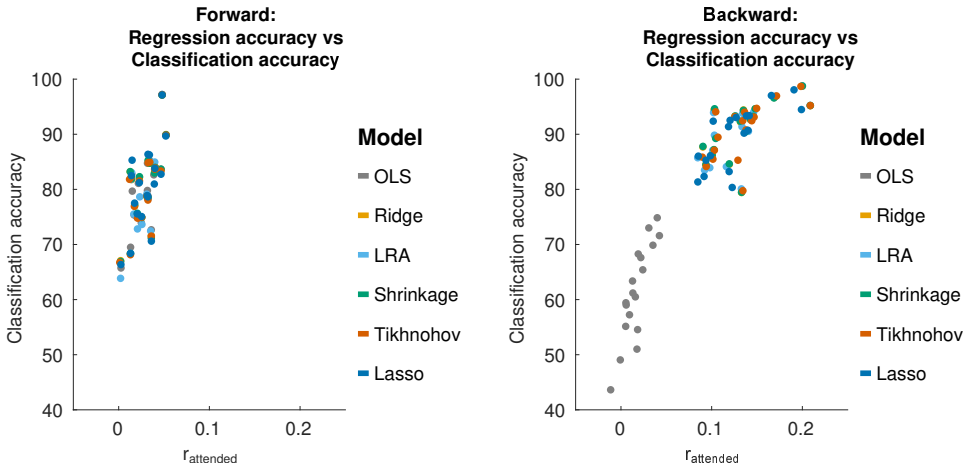

Figure 5.4: Relationship between regression accuracy and classification accuracy, using 30s decoding segment lengths.

the difference between class means ($\bar{r}_{attend} - \bar{r}_{unattend}$) and the within-class variance ($0.5(\sigma^2_{r_{attend}} + \sigma^2_{r_{unattend}})$). Both of these terms affect the separability between classes.

For backward models, Tikhonov regularization had a significantly larger difference between class means compared to Ridge Regression and Shrinkage (Tikhonov>Ridge: $T_{17} = 2.62$, $p = 0.018$), (Tikhonov>Shrinkage: $T_{17} = 2.59$, $p = 0.019$). At the same time, the between-class variance computed over 101 independent decoding segments of 30s each was also significantly larger for Tikhonov regularization (Tikhonov>Ridge: $F_{100,100} = 2.37$, $p = 1.2 \times 10^{-5}$), (Tikhonov>Shrinkage: $F_{100,100} = 2.37$, $p = 1.4 \times 10^{-5}$). This suggests that while Tikhonov regularization yields a better regression accuracy (correlation coefficient), this is offset by an increased variance in the regression accuracy computed over short decoding segments, nullifying any potential gains in classification performance.

5.3.4 Receiver Operating Characteristic

The receiver operating characteristic (ROC) curve, shown in figure 5.5, shows the relationship between the true-positive rate and false-positive rate for decoding segment trials where the classifier discrimination function lies above a given threshold, as the threshold is varied. The classification accuracy score that we report corresponds to the point on the ROC that lies along the line between (0,100) and (100,0). This is also the point at which the Wolpaw information transfer rate (ITR) is estimated, whereas the Nykopp ITR estimation finds a point that lies further left along the ROC curve. The area under the curve is highly correlated with classification accuracy (over all regularization schemes and decoding segment lengths, ($r = 0.99$, $T_{862} = 219.9$, $p < 1.0 \times 10^{-16}$). The Nykopp ITR, on the other hand lies further left along the ROC curve, demonstrating that by avoiding the classification of some trials, it is possible to maximize the ITR.

5.3.5 Information Transfer Rate

The Wolpaw ITR represents the transfer rate when all decoding segments are classified, whereas the Nykopp ITR represents the maximum achievable transfer rate when some classifications are withheld based on classification discrimination function output. Figure 5.6 shows the Wolpaw and Nykopp ITR values as a function of decoding segment duration, based on models computed with

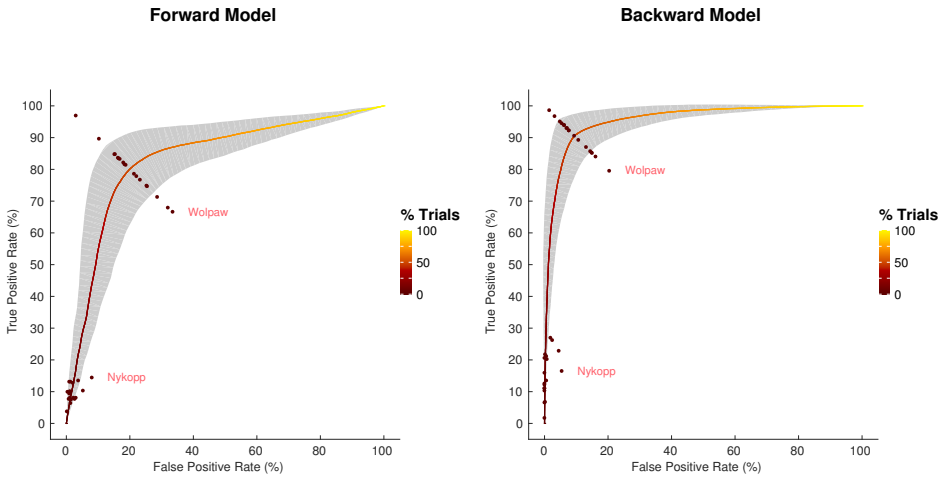

Figure 5.5: Average receiver operating characteristic curve, with standard deviation band, for 30s decoding segments using Tikhonov regularization. Points at which Wolpaw and Nykopp information transfer rates were evaluated for each subject are shown. Color along curve indicates percentage of decoding segment trials evaluated to obtain each point. The gray band indicates the standard deviation boundaries of the curve in both x and y directions.

Tikhonov regularization. Both the Wolpaw and Nykopp ITR show an increase followed by a decrease with increasing decoding segment duration. The plots suggest that for brain computer interface applications with fixed decoding segment lengths, it may be advisable to use decoding segments of 3-5 seconds to maximize the ITR. While the Nykopp measure is an upper-bound, its increase over the Wolpaw ITR value (forward model, 5s: $T_{17} = 13.1$, $p = 2.6 \times 10^{-10}$), (backward model, 5s: $T_{17} = 16.7$, $p = 5.4 \times 10^{-12}$) demonstrates that by adjusting the classifier decision function cutoff, it could be possible to increase the ITR.

5.4 Discussion

In this study, we systematically investigated the effects of forward/backward stimulus-response model estimation methods on the ability to decode and classify attended speech envelopes from single-trial EEG responses to speech mixtures. The performance of stimulus/EEG decoders based on forward models (mapping from attended speech envelopes to multi-channel EEG responses) and backward models (mapping from EEG response back to speech envelopes) were compared. It was found that the backward models outperformed the forward models in terms of regression and classification accuracies. While for-

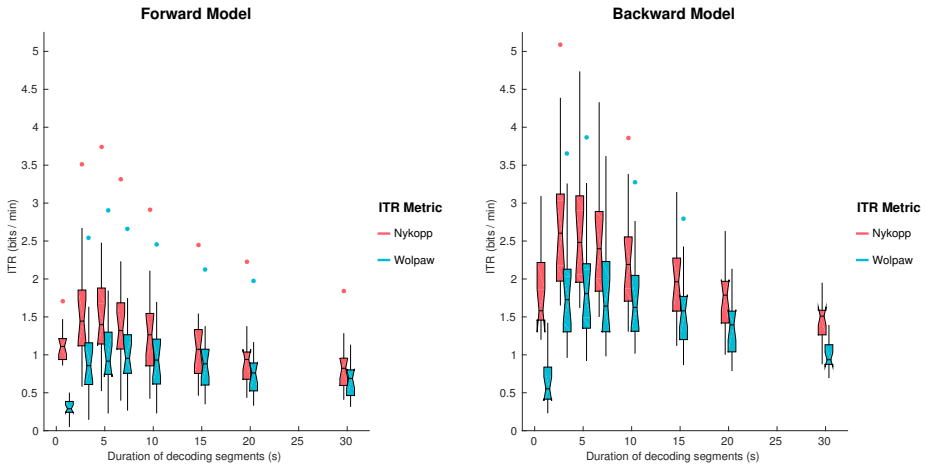

Figure 5.6: Wolpaw and Nykopp information transfer rates (ITR) as a function of decoding segment duration for the forward and backward models, using Tikhonov regularization. Notched boxplots show median, and first and third quartiles. Whiskers show 1.5 × IQR. Dots indicate outliers.

ward models could be expected to have higher regression accuracies due to the averaging of correlation coefficients across channels for forward models, the regression accuracy for the backward model was still higher when compared to the maximum correlation coefficient across channels for the forward model. We hypothesize that the models do a better job of reconstructing audio (the backward model) than predicting EEG data (the forward model) because the EEG data contains a lot of information from other brain functions. It is impossible to predict these signals from the stimulus, hence the limited success of a forward model, but it is possible to filter them out, hence the better performance of a backward model. There are also other fundamental differences between the models, such as statistical and structural properties of the regressor variable, and number of parameters estimated. For instance, the eigenspectrum of the EEG autocovariance matrix in figure 5.9 (in section 5.6) suggests that the matrix is ill-conditioned, particularly compared to that of the speech envelope. Different regularization schemes were not found to significantly affect the forward model classification accuracies. However, for the backward models, the decoding schemes that yielded the best classification accuracy were Ridge Regression, LRA, Shrinkage and Tikhonov. Lasso had a lower classification accuracy by a small but significant margin. Classification accuracy increased monotonically as a function of duration, reflecting the greater amount of discriminative information available in longer segments. ITR however peaked at an intermediate

segment duration, reflecting the tradeoff between the accuracy of individual classification judgments (greater at long durations) and number of judgments (greater at short durations). The optimum was around 3-5s.

For the analysis, we used different linear approaches to decode selective auditory attention from stimulus and EEG data. These analyses all relied on the explicit assumption that the human cortical activity selectively tracks attended and unattended speech envelopes. To fit the models, we made a number of choices based on common practices in literature, and with the goal of being able to compare forward/backward models and regularization schemes. For example, a 500 ms kernel was used as was done by others (Fuglsang et al., 2017). While shorter kernels have been explored as well (O'Sullivan et al., 2014), a longer one tests the ability of the model estimation method to handle a larger dimensionality and allows for a more flexible stimulus-response modeling capturing both early and late attentional modulations of the neural response. Additionally, we chose to focus on 1-9 Hz EEG activity as the attentional modulation of EEG data has been found prominent in this range. It is likely that other neural frequency bands robustly track attended speech (e.g. high gamma power (Pasley et al., 2012)) and that the neural decoders potentially could benefit from having access to other neural frequency bands. This is, however, outside the scope of this paper.

5.4.1 Decoding selective auditory attention with forward and backward models

The forward models performed significantly worse than the backward models in terms of classification accuracies. Single-trial scalp EEG signals are inherently noisy, in part because activity picked up by each electrode reflects a superposition of activity from signals that are not related to the selective speech processing (Blankertz et al., 2011). We refer here to any aspects of the EEG signals that systematically synchronize with the attended speech streams as target signals and anything that does not as noise. To improve the signal-to noise ratio one can efficiently use spatio-temporal filtering techniques. This in part relates to the fact that stimulus-irrelevant neural activity tends to be spatially correlated across electrodes. The spatio-temporal backward models implicitly exploit these redundancies to effectively filter out noise and improve signal-to-noise-ratio. This makes them fairly robust to spatially correlated artifact activity

(e.g. electro-ocular and muscle artifacts) when trained on data from a large number of electrodes. This is also reflected in the high classification accuracies that were obtained with the backward models. However, for the relatively high number of electrodes used in this study, it was found that the spatio-temporal reconstruction filters were effective only when properly regularized.

The forward models, on the other hand, attempt to predict the neural responses of each electrode in a mass-univariate approach. These models do not, therefore, explicitly use cross-channel information to regress out stimulus-irrelevant activity. The relative contribution of the individual channels to the classification accuracies were instead found via an SVM trained on correlation coefficients computed per channel, over short time segments. In short, backward models remove spatial information prior to classification when regressing out non-stimulus-related activity, whereas forward models preserve this information, but do not regress out non-stimulus-related activity. It can therefore be beneficial to apply dimensionality reduction techniques (e.g. independent component analysis (Bell and Sejnowski, 1995) or joint decorrelation (de Cheveigné and Parra, 2014)) to represent the EEG data as a linear combination of fewer latent components prior to fitting the forward models. Alternatively, canonical component analysis can be used to jointly derive spatio-temporal filters for both audio and EEG such that the correlation between the filtered data is maximized (de Cheveigné et al., 2018a).

Regularization

Each regularization scheme makes certain assumptions and simplifications that are therefore adopted by studies employing them. Because these methods have not been previously evaluated side by side, it is unknown how valid these assumptions are.

While no regularization (OLS) was found to work well for forward models in producing classification accuracies roughly in line with regularized models, this method performs relatively poorly when applied to backward models. This is likely reflective of the higher dimensional kernel required for the backward problem. For comparison, a forward model had 33 parameters (per channel) that needed to be fit, whereas a backward model had 2,178 parameters.

We generally found that the reconstruction accuracies (r_{attend}) plateaued over a large range of λ values for linear models (Figure 5.7).

Elastic net regularization permits the adjustment of the balance between L1

and L2 regularization via the α parameter. For the backward model, it was shown that a smaller α value improved the correlation between the reconstructed and attended audio stream by only a narrow margin.

The α value had no significant impact on classification accuracy for either forward or backward models. As such, the higher classification performance of Ridge Regression ($\alpha = 0$), compared to Lasso ($\alpha = 1$) may be a result of differences between the closed form solution used for Ridge Regression and the coordinate descent solution used for the Elastic Net, as well as between the solvers themselves (MATLAB's *mldivide* versus GLMNET (Qian et al., 2013)).

Another coordinate descent method, known as boosting, has been used in several studies (Calabrese et al., 2011; David et al., 2007; Thorson et al., 2015). It has been shown that boosting promotes sparse solutions in the context of spectro-temporal receptive fields with single-unit recordings (David et al., 2007). This method was not explored in the present study because boosting tends to be computationally intractable for backward models due to the high number of parameters, and because it involves a large set of hyperparameters. This makes a direct comparison of the regularization methods difficult. Instead we used the Elastic Net algorithm to investigate how the stimulus-response models could benefit from sparsity.

For the forward model, all regularization schemes yielded regression and classification accuracies that were not significantly different from each other. For the backward model, Tikhonov regularization yielded the best regression accuracy, despite the fact that cross-channel leakage may have lead to a suboptimal solution. However, it was found that the improved regression accuracy did not lead to a better classification accuracy compared to other regression schemes with closed-form solutions (i.e. Ridge, Shrinkage and LRA) due to an associated increased variance in the correlation coefficient computed over short decoding segment lengths. It has been reported that, in practice, the Ridge Regression approach appears to perform better than LRA (Vajargah, 2013); however, no significant difference was found in the present study. LRA removes lower variance components after the eigendecomposition of $\mathbf{X}^T\mathbf{X}$, essentially performing a hard-threshold. In contrast, Ridge Regression is a smooth down-weighting of lower-variance components Blankertz et al., 2011.

5.4.2 Realtime Performance

The information transfer rate results provide insight into how classification performance can be optimized. It is worth noting that the ITR measures represent particular points along the ROC curve, as is illustrated in Figure 5.5. For a binary classification problem, with balanced classes, the Wolpaw ITR corresponds to the point on the ROC curve along the line connecting the corners of the plot at coordinates (100,0) and (0,100). The Nykopp ITR, on the other hand corresponds to the point that maximizes the ITR, essentially trading the number of classified samples for increased classification accuracy. In practice, other considerations besides ITR can influence the choice of the point on the ROC. For instance, if there is a high penalty on incorrect classifications, then the classifier threshold may be adjusted to operate at another point on the ROC curve. In short, the ROC and ITR are useful tools in identifying a suitable balance between sensitivity and specificity.

The ITR results in the present study suggest a 3-5 s decoding segment length to achieve the maximum bit-rate. It should be noted that this assumes that switches in attention can occur frequently, on the order of the decoding segment length, such as in a real-world cognitive control setting where system response latency is an important constraint. In cases, where switches in attention are known to be sparse *a priori*, it may instead be more desirable to increase decoding segment length and sacrifice bit-rate to put more emphasis on accuracy, since the loss in bit rate due to long decoding segments is only evident during attention switches. Such an approach was taken by O'Sullivan and colleagues (O'Sullivan et al., 2017b), where the theoretical performance of a realtime backward model decoding system was characterized for switches in attention every 60 s. In that study, a decoding segment length between 15-20 s was reported as optimal to achieve the best speed-accuracy tradeoff.

5.4.3 Summary

There are many methods that can be used to compute forward/backward stimulus-response models. The present study uses a baseline dataset and procedures for the evaluation of these methods. In consideration of the multiple applications in which forward/backward models are used, primarily dealing with reconstruction accuracies or classification performance, this paper considered multiple metrics of performance. By characterizing the regularization

and performance of the model estimation methods, and the relationship between performance metrics, a more complete understanding of the validity of the assumptions underlying each method is provided, as well as the impact of the assumptions on the end result. While these experiments were done with EEG data, we expect that the results apply equally to magnetoencephalography (MEG) data. The key findings from this study were 1) the importance of regularization for the backward model, 2) the superior performance of Tikhonov regularization in achieving higher regression accuracy although this does not necessarily entail superior classification performance, and 3) optimal ITR can be achieved in the 3-5 s range and by adjusting the classifier discrimination function threshold.

5.5 Author Contributions

DW, SF, JH, EC, MS, and AdC contributed to the code used in the paper. DW, SF, JH, and AdC determined the data analysis procedure. DW created some of the figures, performed statistical analyses, wrote parts of the paper, and was responsible for the overall paper. SF created some of the figures, and wrote parts of the paper. JH, MS, and AdC provided critical feedback on the paper

5.6 Supplementary Material

The effect of regularization parameters on regression accuracy are shown in figures 5.7 and 5.3. The eigenspectra of the audio and EEG autocovariance matrices are shown in figure 5.9.

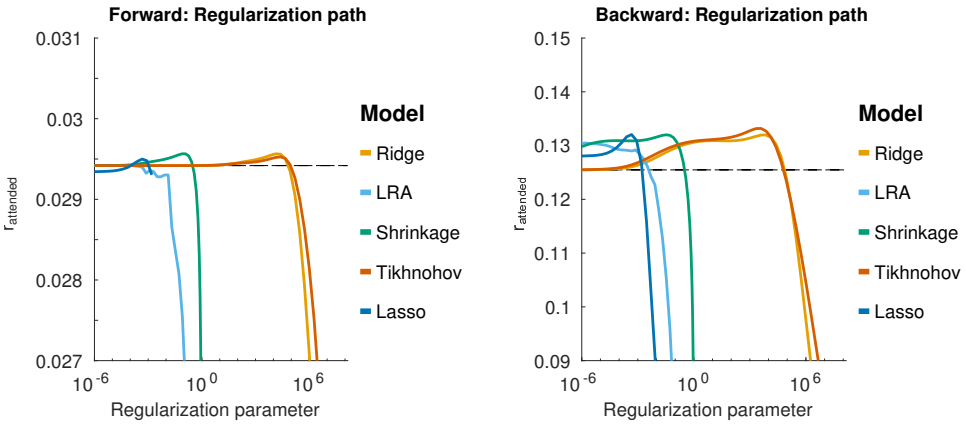

Figure 5.7: Group-mean validation-set regression accuracies obtained with different forward/backward model estimation methods as the regularization parameters λ are varied. The left-hand and right-hand panel present results obtained with forward and backward models, respectively. The x axis shows the strength of the λ regularization parameters. The y axis shows the regression accuracies in terms of Pearson's correlation coefficients between predicted data and target data. The dashed line shows the regression accuracy for OLS.

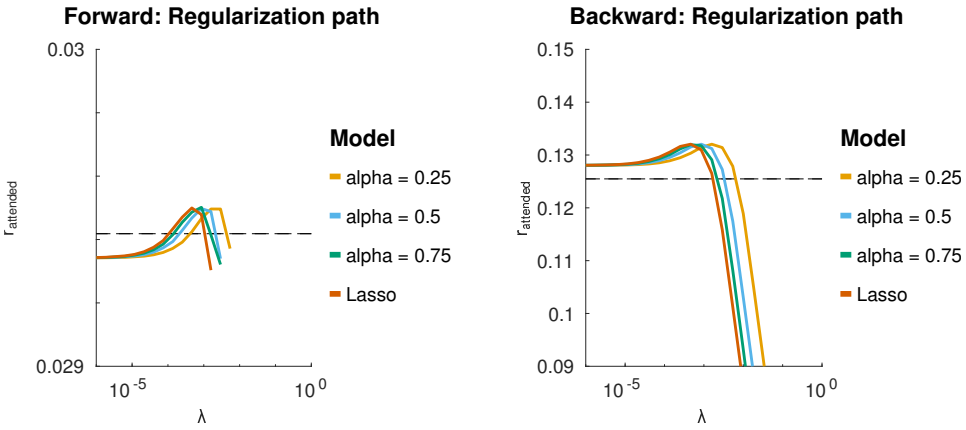

Figure 5.8: Group-mean validation-set regression accuracies obtained from forward/backward models with elastic net penalties. The elastic net has two tuning parameters, λ and α. The two panels show the group-mean validation set regression accuracies cross-validated over a relatively small grid of λ and α values. The prediction accuracies remain stable over a large range of λ values. The dashed line shows the regression accuracy for OLS.

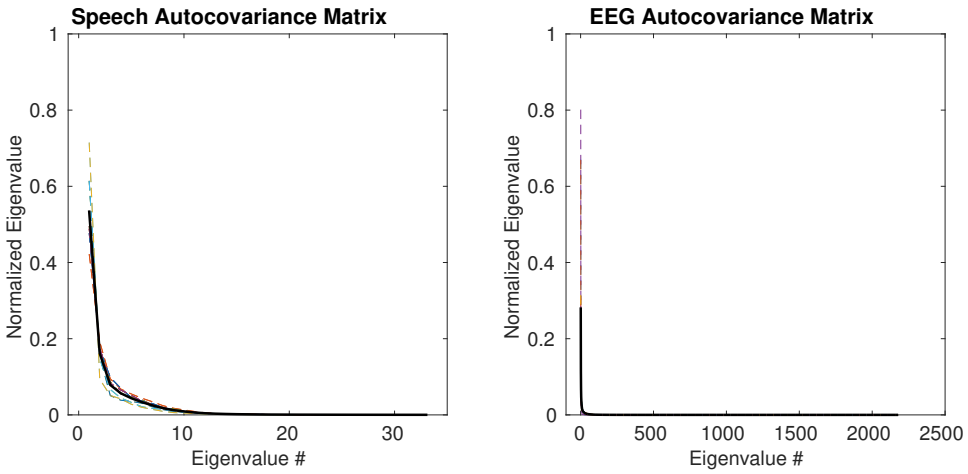

Figure 5.9: Normalized eigenspectra for speech and EEG autocovariance matrices $\left(\mathbf{X}^T\mathbf{X}\right)$. Bold black line is the average across subjects. Individual subjects are shown as thin dashed lines.

6

EEG correlates of working memory load during auditory processing[a]

Abstract

Neuronal oscillations are thought to play an important role in working memory (WM) and speech processing. Listening to speech in real-life situations is often cognitively demanding but it is unknown whether WM load influences how auditory cortical activity synchronizes to speech features. Here we developed an auditory n-back paradigm to investigate cortical entrainment to speech envelope fluctuations under different degrees of WM load. We measured the electroencephalogram (EEG), pupil dilations and behavioural performance from 22 subjects listening to continuous speech with an embedded n-back task. The speech stimuli consisted of long spoken number sequences created to match natural speech in terms of sentence intonation, syllabic rate, and phonetic content. To burden different WM functions during speech processing, listeners performed an n-back task on the speech sequences in different levels of background noise. Increasing WM load at higher n-back levels was associated with a decrease in posterior alpha power as well as increased pupil dilations. Frontal theta power increased at the start of the trial and increased additionally with higher n-back level. The observed alpha-theta power changes are consistent with visual n-back paradigms, suggesting general oscillatory correlates of WM processing load. Speech entrainment was measured as a linear mapping between the envelope of the speech signal and low-frequency cortical activity (<13 Hz). We found that increases in both

[a] This chapter is based on: Hjortkjær, J. H.; Marcher-Rørsted, J.; Fuglsang, S.A. & Dau, T (2017). "Cortical oscillations and entrainment in speech processing during working memory load". European Journal of Neuroscience

types of WM load (background noise and n-back level) decreased cortical speech envelope entrainment. Although entrainment persisted under high load, our results suggest a top-down influence of WM processing on cortical speech entrainment.

6.1 Introduction

Cortical oscillations have been hypothesized to play a functional role in speech processing (Ghitza, 2011; Giraud and Poeppel, 2012). Oscillatory activity, particularly in the delta (1-3 Hz) and theta (4-7 Hz) frequency bands, has been found to entrain to the slow temporal modulations inherent in natural speech signals (Ahissar et al., 2001; Di Liberto et al., 2015; Luo and Poeppel, 2007). Selective attention is known to modulate this response by enhancing the entrainment between low-frequency cortical activity and the speech stream that the listener is attending to, relative to the ignored stream (Ding and Simon, 2012a; Golumbic et al., 2013; O'Sullivan et al., 2014). However, listening to speech in everyday life also involves working memory (WM) to maintain and relate speech content over time or to inhibit irrelevant information. Across modalities, WM tasks have been associated with different oscillatory networks in cortex (Roux and Uhlhaas, 2014), but potential relations to speech processing are unclear. Oscillatory power in higher-order cortical areas are thought to influence speech entrained activity in auditory cortex (Keitel et al., 2017; Park et al., 2015), but it is unclear whether such functional couplings might reflect an interaction between WM processes and auditory processing of the speech stimulus.

The nature of a potential relationship between WM tasks and speech entrainment is not clear. Several scenarios are possible. First, although speech entrainment is known to be shaped by selective attention (Ding and Simon, 2012a; Mesgarani and Chang, 2012; O'Sullivan et al., 2014), theta and alpha signatures of WM demands could reflect general WM processes that do not interact with auditory processing. In this case, attending to a speech stimulus is sufficient to establish an entrained response and additional task demands leave the entrainment response unaffected. Alternatively, higher degrees of WM load may distribute neural resources away from sensory processing of the speech stimulus and towards processing related to the cognitive task. Cortical responses evoked by visual stimuli during WM tasks have consistently been found to be attenuated with increasing cognitive demands (Gevins et al., 1996;

Pratt et al., 2011; Scharinger et al., 2015, 2017; Watter et al., 2001). If this generalizes to speech entrainment, then higher WM load might be associated with a decrease in entrainment. Finally, it is also conceivable that increased task-engagement associated with higher WM load may recruit additional neural resources for the processing of the task-relevant stimulus. In this case, WM load would instead increase the cortical entrainment to the speech signal.

Numerous human EEG/MEG studies have related WM demands to changes in oscillatory power, particularly in the theta and alpha frequency ranges (Klimesch, 1999). Despite the consistent involvement of theta and alpha oscillations, the functional characterization of these oscillations in terms of specific WM functions are still debated. The n-back task is often used to probe WM function (Owen et al., 2005). In an n-back task, subjects are asked to detect whether the presented stimulus in a sequential stream of items matches the one presented n positions back. In visual n-back tasks, increasing WM processing load (higher n) is associated with a frontocentral increase in theta power and a decrease in alpha band power at posterior recording sites (Gevins and Smith, 2000; Gevins et al., 1997; Haegens et al., 2014; Pesonen et al., 2007; Scharinger et al., 2015, 2017). In tasks involving memorization of a number of items (e.g. the Sternberg task), on the other hand, both alpha and theta band power have been found to increase with the number of elements held in memory (Jensen and Tesche, 2002; Jensen et al., 2002; Krause et al., 1996; Leiberg et al., 2006; Obleser et al., 2012; Raghavachari et al., 2001).

Different WM sub-processes are thus associated with different and sometimes opposing alpha-theta changes. In a minimal definition, working memory involves a temporary memory storage (sensory buffers) and attention-related control functions for maintenance and manipulation of WM content ('central executive') (Baddeley, 2003). Executive functions have been further divided into memory updating functions that actively maintain and replace information and WM inhibition that suppress information that is not relevant to the current task (Miyake et al., 2000). The n-back task has been suggested to specifically target WM updating load (Miyake et al., 2000; Scharinger et al., 2015). In visual tasks, inhibitory demands on WM are often manipulated with incongruent items, e.g. in a flanker task. Whereas updating load has been related to decreases in alpha power, inhibitory WM load have been associated with increasing alpha power (Händel et al., 2011; Snyder and Foxe, 2010), consistent with the notion of alpha oscillations as a suppression mechanism (Foxe and Snyder, 2011; Jensen

and Mazaheri, 2010). In auditory tasks, acoustic degradations or noise are a common source of interference and have been shown to increase behavioural WM load (Pichora-Fuller et al., 1995). For spoken or memorised words, acoustic degradations have been associated with increasing alpha power at posterior channels (Obleser et al., 2012; Wöstmann et al., 2017) consistent with an increase in inhibitory WM load. In natural speech processing, however, executive functions related to maintenance of relevant information and inhibition of irrelevant information are typically engaged at the same time. Yet, it is unclear how these WM processes may interact in speech perception. Multiple studies have reported that WM load influences the ability to ignore distracting information, but the nature of this relation appears to be highly dependent on the stimulus type and the type of cognitive task involved (Lavie et al., 2004; SanMiguel et al., 2008; Scharinger et al., 2015; Sörqvist et al., 2012; Vandierendonck, 2014).

Recent studies indicate that speech-entrained activity in the auditory cortex is functionally dependent on oscillatory power in multiple fronto-parietal networks (Keitel et al., 2017; Park et al., 2015). Keitel et al., 2017 recently reported that entrained auditory cortical activity, quantified as the mutual information between the phase of low-frequency activity in auditory cortex and the phase of slow speech envelope modulations, interacted with oscillatory power in distinct cortical networks. In particular, delta entrainment in the auditory cortex was dependent on central alpha and frontal beta power and modulated parietal theta power. This could indicate a top-down influence on speech-entrained activity in auditory cortex by oscillations in a larger cortical network involved in cognitive control or attention. Such a top-down influence could reflect language-specific functions such as semantic memory (Keitel et al., 2017), but could also be related to more general WM functions. To test more directly whether WM processing influence cortical speech entrainment, however, it needs to be demonstrated that imposing a WM processing load in behavioural tasks influences concurrent speech entrainment.

Here, we developed an experimental paradigm to investigate influences of WM load on cortical speech envelope entrainment. We designed a 'number speech' material consisting of sequences of spoken numbers that match important properties of natural continuous speech. During speech listening, participants performed either a 1-back or 2-back task with the speech sequences embedded in either a high or a low level of background noise. This allowed us to examine the individual and combined effects of WM updating (n-back

level) and inhibition (noise level) load during continuous speech processing. We recorded the EEG as well as changes in pupil sizes which are often used as a physiological marker of WM demands (Scharinger et al., 2015; Van Gerven et al., 2004; Wendt et al., 2016; Zekveld et al., 2010). To examine potential differences in speech entrainment during the different load conditions, we used regression techniques to analyse the relationship between ongoing low-frequency cortical activity and envelope fluctuations in the corresponding speech signal (Ding and Simon, 2012a; Lalor et al., 2009b). Using continuous speech, our paradigm also allowed us to study the dynamics of prolonged WM load and load-related measures over longer time segments.

6.2 Materials and Methods

6.2.1 Participants

22 healthy volunteers (six females, ages 19-28, mean age: 24, SD: 3 years) participated with informed consent. Eye-tracking data was recorded in 15 of the participants. All participants reported normal hearing. The experiment was approved by the Science Ethics Committee for the Capital Region of Denmark (protocol no. H-16036391) and conducted in accordance with the Declaration of Helsinki.

6.2.2 Speech stimuli

We created a speech material that could be used to control the WM load imposed on the listener and monitor their task performance during listening. The speech material consisted of spoken number sequences created to match natural continuous speech in terms of syllabic rate, intonation, and sentence rhythm. First, two- or three-digit numbers were read by a male Danish speaker and recorded in an anechoic chamber. For each number, several tokens spoken in rising or falling intonation patterns were recorded. The recorded number tokens were afterwards concatenated into sequences of 'number sentences' consisting of three or four numbers (see Fig. 1). The time interval between numbers was set at random durations ranging between 150-230 ms and the time interval between number sentences was set randomly between 300-700 ms to match the word and sentence rhythm of natural speech. The number sentences were then used to synthesize long sequences of spoken numbers for the experimental

trials. We created 20 trial lists each of 30 spoken numbers (resulting in durations between 45 s and 55 s). Each trial list contained n=1,2,3 back repetition targets, i.e. numbers which were identical to the number presented n numbers previously. We ensured that the n-back targets were equally distributed between the first and second half of the list.

To generate speech-shaped stationary background noise with the same spectral characteristics as the original speech stimuli, we computed the average of a large number of speech waveforms until the signal had no distinct slow envelope modulations. In the experiment, we wanted to impose the noise at a signal-to-noise ratio (SNR) that resulted in maximal interference without disrupting speech intelligibility. For this reason, we measured speech reception thresholds for the number tokens in a separate psychoacoustic test with four normal hearing listeners not participating in the main experiment. The lowest SNR point on the psychometric function that resulted in 100 % correct identification was estimated to 0 dB SNR. In the main experiment, this noise level was defined as the 'high-noise condition'. A noise level 10 dB lower (i.e., 10 dB SNR) was defined as the 'low noise-condition'. Different speech-shaped noise tokens were used in every trial, i.e. the noise was not frozen.

For the analysis of speech entrainment, the temporal amplitude envelopes of the continuous speech signals were extracted using an auditory model of envelope processing in the peripheral auditory system. The audio waveforms were first passed through a gammatone filterbank mimicking the spectral filtering characteristics of the basilar membrane (Patterson et al., 1987). At the output of each filter, the envelope was extracted via the Hilbert transform and raised to the power 0.3 to account for the compressive response of the inner ear (Plack et al., 2008). The spectrally decomposed envelopes were then resampled to match the EEG sampling rate and averaged across frequency channels.

6.2.3 Experimental design

To control WM load during speech listening, participants listened to the continuous speech stimuli while performing an n-back task in different levels of background noise. The conditions formed a 2×2 factorial design consisting of two n-back task levels (1-back, 2-back) and two noise levels (low noise: 10 dB SNR, high noise: 0 dB SNR). In the 1-back condition, participants were asked to detect whenever a number was repeated, and in the 2-back task, they detected whether the presently spoken number was the same as one spoken two times

back (Fig. 1). Note that repeated numbers were not acoustically identical but different speech tokens of the same number within the continuous speech stream. The same speech lists were used in the different n-back conditions such that subjects heard the same speech stimuli in the two different behavioural contexts. Since the occurrences of 1-2-3 back repetitions were equally distributed, the same occurrences acted as either targets or lures (repetitions to be ignored) depending on the n-back task.

Fig. 1 presents a schematic illustration of the trial timeline. Each trial began with a 7.2 s silent resting baseline where subjects fixated on a cross positioned in the middle of a black background screen. 2 s after the onset of the resting baseline, a green screen was shown for 200 ms in order to measure the pupil light reflex, followed by another 5 s of a black screen baseline. Following the black screen baseline, a grey screen was presented for 500 ms before the onset of the sound stimulation. During sound stimulation, the participants maintained eye fixation on a cross on the grey background screen. The sound stimulation started with 1.5 s of the background noise at 0 dB or 10 dB SNR before the onset of the speech stimulus. During the following $\backsim$45-55 s presentation of the speech stimulus, the participants were asked to press a button when an n-back target was detected. The participants were not instructed to use of any particular finger for responding. They were not informed about the noise level prior to the sound presentation. Responses were considered correct when they occurred between the onset of the target number and the onset of the following number plus an additional 200 ms. Responses that did not fall in this time-interval were considered false alarms. After the speech task, the pre-trial baseline and screen flash were repeated. Subjects performed 8 initial training trials during which they received feedback whenever n-back targets occurred in the speech stimulus. During the main experiment, feedback was only provided between trials by showing the average percent correctly identified n-back targets. Each participant performed 10 trials for each of the four experimental conditions. Lists contained either 4 (15 out of 20 lists) or 3 (5 out of 20 lists) n-back targets.

6.2.4 Data acquisition

The experiment was performed in an electrically shielded double-walled sound booth (IAC Acoustics, North Aurora, IL, USA). The subjects were seated 60 cm in front of a presentation screen with dim background lighting that was kept constant for all participants. The auditory stimuli were presented via ER-2 insert

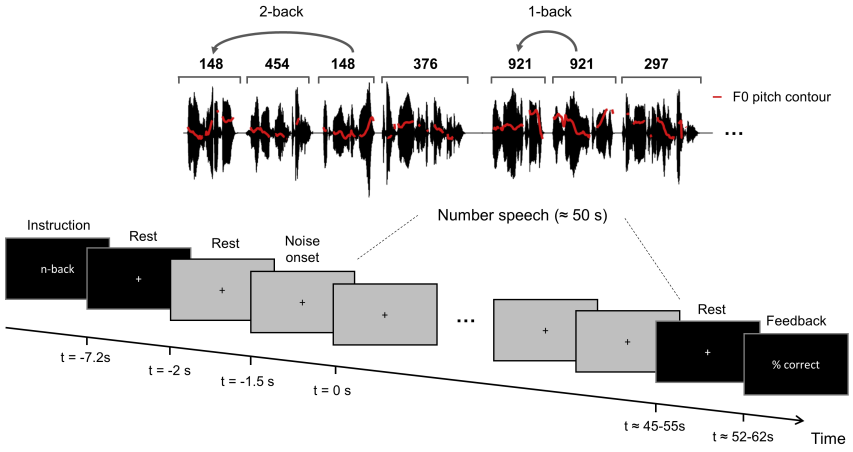

Figure 6.1: Schematic illustration of the trial structure and task. EEG and pupillometry were recorded while subjects listened to continuous speech stimuli consisting of spoken number sequences. Red lines on the waveform represent the pitch contour of the continuous speech signal. In different trials, listeners identified either 1-back or 2-back number targets in different levels of background noise. Please see the Methods section for details.

earphones (Etymotic Research, Elk Grove Village, IL, USA). The speech stimuli were presented at a fixed level of 65 dB SPL. The level of the speech stimuli was kept constant and the level of background noise relative to the speech signal varied across noise conditions.

EEG was recorded continuously at 64 scalp electrodes according to the international 10/20 system using a BioSemi ActiveTwo system (BioSemi, Amsterdam, Netherlands). The sampling rate was 512 Hz. Additional electrodes were placed on the left and right mastoids. Eye movements were detected using six bipolar electrooculographic channels positioned vertically and horizontally around the eyes.

For 15 subjects, pupil sizes were recorded using an Eyelink 1000 desktop system (SR Research Ltd., Ottawa, ON, Canada) with a sampling frequency of 250 Hz. Measurements were conducted on one eye, which varied between subjects. The eye tracking system was calibrated at the beginning of the experiment using a custom calibration routine.

6.3 Data preprocessing and analysis

6.3.1 Behavioral data

A measure of d-prime (d') was used to estimate subjects' sensitivity in the n-back task. This was defined as the difference between the inverse cumulative distribution function (CDF) of correct n-back target detections (hits) and the inverse CDF responses made in the absence of a target (false alarms). To examine performance in the time course of the trial, we also computed the percentage correctly identified n-back targets at their temporal positions in the trial.

6.3.2 EEG data pre-processing

The EEG data were analyzed using MATLAB and the FieldTrip toolbox (Oostenveld et al., 2011). The data were epoched from 5 s before the onset of the speech stimulus to 45 s after the speech onset. The data were high-pass filtered at 0.5 Hz, re-referenced to the average of the two mastoid electrodes, and resampled to 128 Hz. For one subject, the data were re-referenced to the average of all 64 scalp electrodes due to noisy mastoid electrodes. Bad (i.e. noisy) channels were identified visually and removed from the data. On average, 2.4+/—1.9 channels were rejected. The bad channels were interpolated using a nearest neighbour method average.

The logistic infomax independent component analysis (ICA) algorithm (Bell and Sejnowski, 1995; Delorme and Makeig, 2004; Winkler et al., 2015) was used to decompose the re-referenced EEG data from each subject high-pass filtered at 1 Hz. The components were visually inspected and artefactual components were rejected. On average, 4.2+/-1.7 components (range 2-7) were rejected, corresponding to 6.9+/-2.6 % of components. Of the rejected components, 2.4+/-1.3 components were considered electroocular (EOG) artefacts as they were highly correlated with the EOG electrodes, with strong weights at frontal scalp regions. The remaining 1.8+/-1.4 components were identified as either muscle or cardiac-related artefacts that appeared consistently across trials. The ICA-derived mixing matrices were thereafter used to spatially filter out artefactual activity from the original EEG data high-pass filtered at 0.5 Hz (Winkler et al., 2015). Trials were inspected visually for artefacts after ICA cleaning, and remaining bad trials were removed. Additionally, trials in which the subjects detected less than 25 % of the target were rejected from further analysis. On

average, 7.6 +/- 4.2 trials were rejected per subject. Three subjects with more than 50 % of the data rejected in any given condition were removed from further analysis. For the remaining 19 subjects, there were no statistical differences in the number of trials removed between conditions (n-back and noise interaction: $F_{1,18} = 0.9404$, $p = 0.345$, n-back: $F_{1,18} = 0.0705$, $p = 0.7937$, noise: $F_{1,18} = 1.8331$, $p = 0.192$). On average, 8.1 +/-1.4 trials remained in each condition for the remaining subjects.

We examined relative changes in theta and alpha band power between the experimental conditions. Theta activity was defined in the frequency range from 4 to 7 Hz, and alpha from 8 to 13 Hz. Filtering was performed using high order finite impulse response filters. To compute band power, we calculated the sum of the squared absolute values of the filtered EEG signal for each of the frequency ranges in time windows of 5 s with 90 % overlap. To account for individual differences, the power measures were normalized globally by dividing by the power measures in each trial by the global average in band power across all trials. To further explore oscillatory power changes over a larger frequency range, we examined time-frequency representations (TFRs) of power changes by computing the spectral power as above but in 2-Hz wide frequency analysis windows from 1 to 30 Hz, in steps of 0.5 Hz. The TFRs were normalized per frequency bin to the grand average power across all trials and all frequency bins.

To study whether cortical EEG speech entrainment is modulated by working-memory related processes, we derived temporal response functions (TRFs) (Ding and Simon, 2012a; Lalor et al., 2009b) that map linearly from the envelope of the continuous speech signal $S(t)$ to the EEG responses $R(t,n)$:

$$\hat{R}(t,n) = \sum_{l=1}^{L} h(\tau_l, n) S(t - \tau_l)$$

where $n = 1...N$ denotes the number of electrodes and $\tau = \{\tau_1, \tau_2, ...\tau_L\}$ are the time lags between the stimulus and response. The TRFs, $h(\tau)$, were fitted separately on the data from each subject in each of the four experimental conditions. The TRFs were estimated using regularized regression with a quadratic penalty term (Lalor and Foxe, 2010). The regularization parameter of the were set to a fixed high value that gave the highest group-mean leave-one-out prediction accuracy across all subjects ($lambda = 2^12$). The temporal response functions covered time lags ranging between 0 ms to 400 ms post-stimulus in steps of 7.8 ms (sampling frequency of 128 Hz). The EEG data and

speech envelopes were standardized to have zero mean and unit variance. The TRF models were computed using Matlab code publicly available at *www.ine-web.org/software/decoding*.

For the TRF analyses, the EEG data were filtered between 1 and 13 Hz using high order finite impulse response filters. To quantify changes by either n-back or noise on the TRF, the peak amplitude, as well as the latency of the peak was examined. This was done by extracting the maximum value of the TRF from 100 ms to 300 ms for each subject. The latency was defined as the time at which the peak value of the TRF occurred. A leave-one-trial-out cross-validation procedure was used to estimate model prediction accuracies in each experimental condition. The prediction accuracies were quantified as Pearson's correlation coefficient between the predicted EEG responses and the actual recorded EEG data on the held-out trials. This correlation served as indicator of the degree of speech entrainment, i.e. how tightly the cortical activity was synchronized to the speech envelope. We also examined band-specific entrainment by filtering the EEG data in delta (1-3 Hz), theta (4-7 Hz) and alpha (8-13 Hz) ranges. In the statistical analysis of condition-specific differences, we focused on 12 fronto-temporal electrodes (FC5, FC3, FC1, FC2, FC4, FC6, F5, F3, F1, F2, F4, F6) previously found to be speech-relevant (Di Liberto et al., 2015). To estimate chance level prediction, we used a permutation procedure where we predicted EEG responses based on envelopes of non-matching speech sequences. The 97.5 % percentile of the chance distribution was defined as the noise floor.

6.3.3 Pupil data

Eye blinks were classified as samples in the time series where the absolute value of the pupil diameter exceeded 3 standard deviations of the mean pupil diameter. Blink-corrupted segments were linearly interpolated from 350 ms before to 700 ms after the blink (Wendt et al. 2016). Trials containing more that 20 % of corrupted data were rejected from further analysis. Furthermore, three subjects with more than 50 % of rejected trials were excluded from the analysis. The subjects excluded due to noisy EEG data were not the same as the subjects excluded due to noisy pupilometry data. For the remaining subjects, 2+/-3 trials were rejected. The blink-removed data were smoothed using a 25-point (100 ms) moving average filter. To account for individual differences between subjects, the data were normalized to the pupil diameter averaged over the 200 ms time window directly preceding the noise onset.

6.3.4 Statistical analysis

We used repeated measures analyses of variance (ANOVA) to assess statistical group-level differences between the 2×2 conditions (n-back, noise) on all load-related measures: behavioural performance, average pupil size and maximum pupil dilations, EEG band power, TRF peak amplitudes, TRF peak latencies, and prediction accuracies. All statistical calculations were performed using Matlab. Shapiro-Wilk tests (alpha = 0.05) were used to test for the normality assumptions of the parametric tests. For the analysis of the band-specific oscillatory EEG power, we assessed group-level differences for the time-averaged theta band power over a frontal electrode (AFz) and alpha band power over a posterior electrode (Oz). This restriction was motivated by previous results showing effects of WM load in the theta band at frontal-midline electrodes, as well as effects in the alpha band at posterior electrodes (Gevins and Smith, 2000; Gevins et al., 1997; Scharinger et al., 2015). To further explore differences in the trial-averaged power over all electrodes sites, we performed cluster-based permutation tests (as implemented in the Fieldtrip toolbox, (Oostenveld et al., 2011). We computed the group-level effects of n-back and noise level on theta and alpha band power with a cluster-defining threshold (cluster alpha) of $p<0.01$, a cluster-level threshold of $p<0.01$, and an electrode neighbourhood extent of 40 mm.

6.4 Results

6.4.1 Behavioural performance

Response accuracy in the n-back task (measured in d') was significantly lower in the 2-back condition compared to the 1-back task ($F(1,21)= 203.77$, $p<0.001$) but was not affected by the level of the background noise ($F(1,21)=0.7487$, $p=0.397$) (Fig. 2B). This result was expected since the higher noise level was predetermined to yield the speech fully intelligible. The n-back targets (and lures) were uniformly distributed over the trial duration. This allowed us to inspect potential differences in response accuracy in different parts of the trials. As shown in Fig. 2A, the identification of 1-back targets remained high throughout the trial, whereas the identification of 2-back targets declined as the trial progressed.

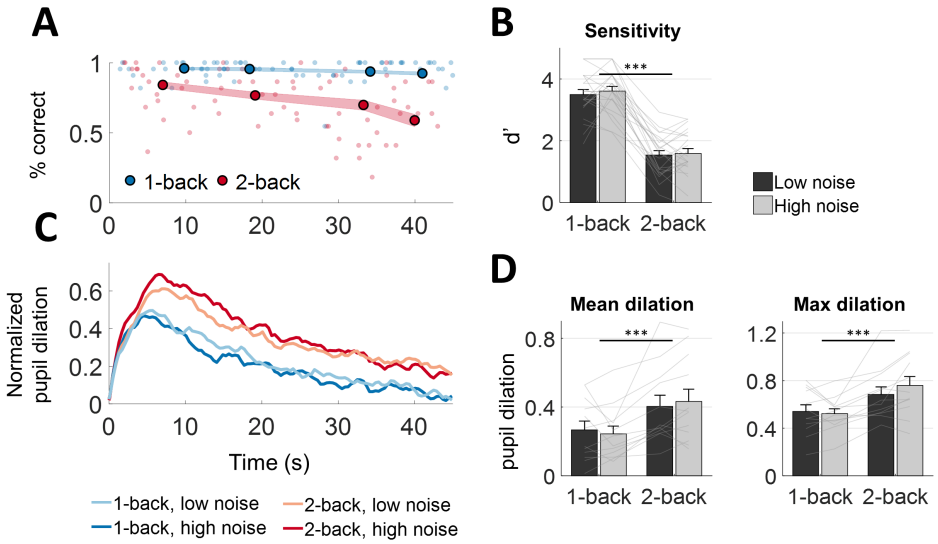

Figure 6.2: Behavioural performance (above) and pupil responses (below). A. Percentage of correctly detected 1-back and 2-back targets during the speech trial. Larger circles represent the average % correct at the average position of the targets. B. Behavioural sensitivity (d-prime) for n-back target detection measured over the trial duration. C. The average trace of the pupil dilations relative to a pre-stimulus baseline. D. Mean and peak pupil dilation over the trial duration. Error bars represent +/- 1 s.e.m. ***p<0.001

6.4.2 Influence of WM load on pupil dilations

All WM task conditions evoked a pupil dilation response with a peak 5-10 s after trial onset, followed by a gradual decrease for the remaining duration of the trial (Fig. 2C). The pupil dilations increased with the n-back task level but did not increase additionally with the level of the background noise. Both the mean and peak pupil dilation were significantly higher for the 2-back task compared to the 1-back task (mean dilation: $F(1,11)= 17.00$, $p=0.0017$; peak dilation: $F(1,11)=20.16$, $p<0.001$). No significant effects of noise level were found on the pupil measures (mean dilation: $F(1,11)=0.31$, $p=0.58$; peak dilation: $F(1,11)= 0.76$, $p=0.40$).

6.4.3 Influence of WM load on alpha and theta power

We first investigated changes in posterior alpha power and frontal theta power previously associated with WM load. As illustrated in Figure 3, increasing WM load in the more difficult 2-back task compared to the 1-back task was associated with a decrease in posterior alpha power. An ANOVA on trial-averaged alpha power at electrode Oz revealed a main effect of n-back level ($F(1,18)=30.15$, $p<0.001$). Frontal theta power increased at the start of the trial and increased additionally during the 2-back task compared to the 1-back task (main effect at electrode Afz: $F(1,18)=10.88$, $p=0.004$). No significant effects of the background noise level were observed on trial-averaged alpha ($F(1,18)=1.90$, $p=0.18$) or theta ($F(1,18)=0.29$, $p=0.60$) power changes.

Influence of WM load on speech envelope entrainment We derived temporal response functions (TRFs, Fig. 4A-B) to analyse how low-frequency cortical activity entrained to fluctuations in the speech envelope. The TRF can be viewed as a speech-evoked response generalized to continuous stimuli (Lalor et al. 2009). In all conditions, we observed a late (170 ms) positive peak in the TRF amplitudes (Fig. 4A-C). Both the amplitude and latency of the late peak was found to be affected by the background noise level (Fig. 4C). For the higher noise level, the peak latency increased ($F(1,18)=20.43$, $p<0.001$) and the peak amplitude decreased ($F(1,18)=12.95$, $p=0.002$). No significant changes in peak amplitude ($F(1,18)=0.80$, $p=0.381$) or latency ($F(1,18)=0.84$, $p=0.371$) were found for the change in n-back level.

To quantify how precisely the cortical activity entrained to the speech envelope in the different WM conditions, we computed the correlation coeffi-

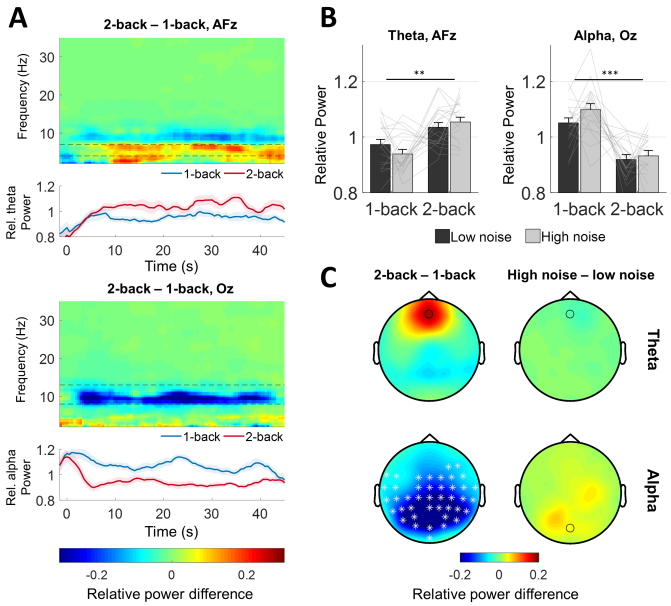

Figure 6.3: Changes in oscillatory power during the n-back speech task. A. Time-frequency representations of the power changes between the 2-back and 1-back tasks at frontal electrode AFz (above) and posterior electrode Oz (below). Stippled lines mark the location of the theta (above) and alpha (below) bands. Traces below the TFRs show the normalized theta and alpha band power in the two n-back tasks. Shaded areas in the traces represent +/- 1 s.e.m. across subjects for each 5 s time-window. B. Trial-mean (5-45 s) power in frontal theta (left) and posterior alpha (right). C. Topographies showing the trial-mean differences in theta (above) and alpha (below) power between the 2-back and 1-back tasks (left) and between high and low noise levels (right). Circles indicate the position of electrodes AFz and Oz. White asterisks indicate electrodes showing significant power differences between the n-back conditions revealed by the cluster analysis ($p<0.01$, corrected). Error bars represent +/- 1 s.e.m. **$p<0.01$, ***$p<0.001$.

cient (Pearson's r) between the responses predicted by the TRF models and the measured EEG (Fig. 4D-E). The TRF models were first used to predict the low-frequency (1-13 Hz) EEG response at 12 frontotemporal electrodes from the speech envelopes. As shown in Fig. 4D, the average prediction correlation across experimental conditions was high over fronto-temporal electrodes, in accordance with previous TRF studies (Di Liberto et al., 2015; Crosse & Lalor, 2014). Analysis of prediction correlations between WM conditions revealed a significant interaction between n-back level and noise level ($F(1,18)=6.02$, $p=0.025$) (Fig. 4E). The prediction values were found to decrease with increasing n-back level (main effect: $F(1,18)=10.68$, $p<0.005$) and increasing noise level (main effect: $F(1,18)=10.54$, $p=0.005$), but the effect of the background noise

was found to be larger in the 1-back condition than in the 2-back condition.

Since previous work has pointed to different functional roles for delta and theta-band entrainment in speech coding (Ding & Simon, 2014), we also investigated the effects of behavioural WM load on speech entrainment separately in different frequency bands (Fig. 4E). This was done by computing the prediction accuracies of TRF models estimated from EEG responses bandpass filtered in the delta, theta and alpha frequency bands. The prediction correlations were only above the noise floor in the delta and theta band, but not in the alpha band. As in the analysis of the broadband signal (1-13 Hz), the speech entrained response in the delta and theta bands was significantly reduced with increases in both types of WM load (Fig. 4E). Increasing the background noise level reduced prediction correlations in the delta-band (main effect: $F(1,18)=16.75$, $p<0.001$), and in the theta-band (main effect: $F(1,18)=4.95$, $p=0.039$). Increased WM load in the n-back task also decreased entrainment in both the theta band (main effect: $F(1,18)=7.10$, $p=0.016$) and in the delta band (main effect: $F(1,18)=5.91$, $p=0.026$).

In our analysis, we focused on entrainment between the envelope of the speech signal and cortical activity. Reduced entrainment with increased background noise levels could potentially reflect cortical entrainment to the presented noisy speech stimulus rather than the underlying speech signal. To investigate whether the cortical activity tracks the actual noisy stimulus envelope rather than the underlying speech envelope, we performed the same TRF analysis but for the envelopes of the noisy speech mixture. Prediction accuracies based on the noisy speech envelopes were significantly lower than for the envelope of the clean signals (paired t-test, $t=4.93$, $p<0.001$), suggesting that the cortical activity mainly entrains to the clean speech signal rather than to the noisy sound mixture.

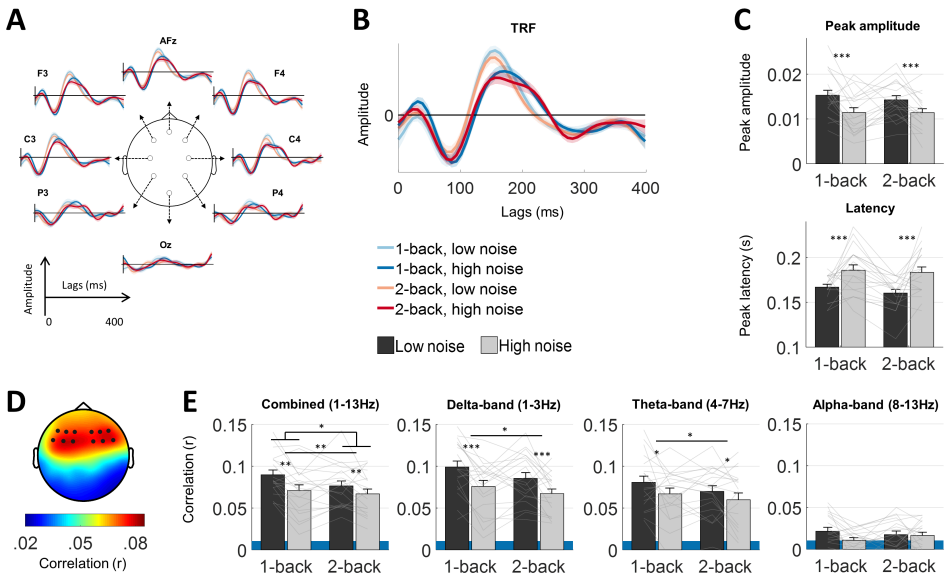

Figure 6.4: EEG responses to speech envelopes in the different WM load conditions. Above (A-C): Temporal response functions (TRFs) derived from linear regression between EEG data and the speech stimulus. Below (D-E): Speech entrainment measured as the correlation between the cortical response predicted by the speech envelope and the EEG. A. TRFs at selected electrode locations to illustrate the responses at different scalp positions. B. TRFs averaged over fronto-central electrodes in the different experimental WM conditions. C. The amplitude (above) and latency (below) of the late positive peak in the average TRF around 170 ms. D. Topographical distribution of the EEG prediction accuracies (Pearson's r) averaged across conditions. The dots indicate the positions of the analysed fronto-central electrodes. E. Average prediction accuracies in different frequency bands. The shaded areas represent chance level prediction. Error bars represent +/- 1 s.e.m. *p<0.05, **p<0.01, ***p<0.001.

6.5 Discussion

We devised an auditory n-back task embedded in continuous speech to investigate interactions between WM load and speech processing. Consistent with previous visual n-back paradigms (Gevins and Smith, 2000; Gevins et al., 1997; Haegens et al., 2014; Pesonen et al., 2007; Scharinger et al., 2015, 2017), increasing load with higher n-back levels was associated with increased frontal theta band power and decreased posterior alpha power. At the same time, cortical entrainment to the speech envelope decreased with increasing WM load. Both increased background noise levels and higher n-back levels decreased speech entrained responses in the delta and theta bands.

6.5.1 Dynamics of alpha and theta power and pupil dilations during WM load

The continuous speech paradigm allowed us to observe the dynamics of load-related measures over prolonged periods of WM load. Load-specific changes in behavioural performance, EEG band power, and pupil size each exhibited different dynamics over the trial duration. The observation of task-evoked pupil dilations in the initial 5-10 s of the trial (Fig. 2C) is consistent with numerous previous pupil studies of WM load or cognitive effort in paradigms with shorter trials (Beatty, 1982) (Koelewijn et al., 2012; Scharinger et al., 2015; Wendt et al., 2016; Zekveld et al., 2010). However, we also observed that this was followed by a similar decrease in pupil sizes for the remaining duration of the trial. During this decrease, the pupil dilations remained sensitive to n-back load (Fig. 2C). Behavioural performance, on the other hand, decreased during the trial but only during the difficult 2-back task (Fig. 2A). This could indicate fatigue. However, a similar pattern specific to the high-load condition was not reflected in either the EEG band power or in the pupil responses. In the EEG theta or alpha power (Fig. 3A), we did not find similar patterns of change throughout the trial but the individual power traces had considerable local variation.

An initial increase in theta power was observed in the beginning of the trials (Fig. 3A). This could reflect the increase of items held in WM when participants were presented with the first numbers of the sequence, consistent with visual WM tasks (Raghavachari et al., 2001). In the remaining parts of the trial, theta power remained high and increased additionally during 2-back task compared to the 1-back task. (Scharinger et al., 2017) recently reported a similar increase in frontal theta emerging in the course of a visual n-back task, but did not observe a similar theta pattern during memorization in WM span tasks. This could suggest that theta is more specifically related to the organization and continuous update of WM items, and less to memory storage of those items. Specifically, our results are consistent with a functional role of theta oscillations for organizing multiple items in a sequential order in short-term memory (Lisman and Jensen, 2013; Raghavachari et al., 2001).

Decreased alpha power for higher n-back levels throughout the trial is also consistent with previous n-back studies (Gevins and Smith, 2000; Pesonen et al., 2007; Scharinger et al., 2015, 2017). Reduced alpha power, however, has also been observed in a number of other complex WM tasks, and may

reflect the complex nature of the n-back task. The n-back task requires subjects to simultaneously update WM information and match stored items with the current input (Watter et al., 2001). Alpha desynchronization has been proposed to reflect the fact that a number of WM processes are simultaneously required for task performance (Klimesch, 1999; Scharinger et al., 2015). Simultaneous involvement of different WM functions in different task strategies may also explain the fact that we observed a considerable variability in alpha patterns between subjects in our data (see Fig. 3B). While some subjects may be able to search WM content before a new number is presented, others may try to match stored items each time a new speech item is heard (Watter, 2001). Different processing strategies that put different demands on the matching subtasks could potentially generate variability in the observed alpha patterns.

6.5.2 WM processes influence speech entrainment?

Speech envelope entrainment was found to decrease with an increase in the two types of WM load examined. In visual n-back tasks, the amplitude of P300 evoked potentials have consistently been found to be attenuated by increasing WM load at higher n-back levels (Gevins et al., 1996; McEvoy et al., 1998; Scharinger et al., 2015; Watter et al., 2001; Wintink et al., 2001). This reduction has been interpreted in terms of a re-distribution of resources between WM processes at higher load levels. Yet, decreased speech entrainment with increasing WM load, as observed in the current study, points to an interaction between WM processing and auditory processing of the speech stimulus. Thus, decreased entrainment with higher load levels may reflect a re-allocation of WM resources at the expense of parsing of the speech stimulus.

A possible explanation for the WM-specific reduction in speech entrainment could be an interaction between WM processing and attention (Gazzaley and Nobre, 2012). Numerous studies have demonstrated that selective attention to a particular talker reduces entrainment to ignored speech streams (Ding and Simon, 2012a; Fuglsang et al., 2017; Golumbic et al., 2013; O'Sullivan et al., 2014; Power et al., 2012). Auditory entrainment to speech has also been reported even in the absence of overt auditory input, e.g. during imagined speech (Deng et al., 2010; Martin et al., 2014). This raises the possibility that attention-driven speech entrainment can operate entirely on internal speech representations. In a continuous updating task such as our current speech n-back paradigm, WM processing may direct attentional focus towards the internal rehearsal of

verbal items in the phonological loop. In this case, new items in the continuous speech stream compete for selective attention with verbal information currently in the phonological loop. Increasing attention towards the phonological loop for higher n-back levels would then explain a decrease in cortical activity entrained to the ongoing speech stimulus. Such a mechanism would need to be examined more closely, for example, by comparing entrainment to matching vs mismatching search targets. We note that the observed reduction in speech entrainment during WM load is relatively small compared to the reduction in entrainment typically reported for ignored speech streams in selective attention tasks.

While higher n-back levels reduced delta-theta entrainment, this was not accompanied by a significant reduction in TRF amplitudes. Increasing background noise levels, on the other hand, significantly attenuated and shifted the latency of the TRF peak. Consistent with this, increasing levels of continuous background noise have previously been found to increase ERP latencies of both N100 and P300 components in a syllable discrimination task (Whiting et al., 1998). Latency shifts and attenuated amplitudes of TRFs with increasing background noise levels have also been reported previously, but only for earlier TRF components (50 ms) observed in MEG component space (Ding and Simon, 2013). Our current TRF method did not reveal any clear early components and the later peak may reflect a compound effect of early and later auditory processing.

Higher WM load levels decreased speech entrainment (Fig. 4) and, at the same time, induced load-specific alpha-theta power changes (Fig. 3). The phase of auditory cortical activity entrained to speech has previously been suggested to be functionally coupled with alpha, theta and beta power in fronto-parietal regions (Keitel et al., 2017; Park et al., 2015), but the functional significance of these couplings has not been clarified. In line with the present results, (Keitel et al., 2017) found that reduced entrainment in the delta band was associated with increases in parietal theta power. The authors proposed that this could reflect working memory involvement to compensate for weaker entrainment. Our results suggest instead that WM load, here induced by the behavioural task, reduces the speech-entrained response. The WM-specific power changes found in the current study (Fig. 3) also point to executive functions that are not specific to speech. However, the functional coupling involved in WM-specific modulation of speech entrainment may be different from those observed in

paradigms without specific WM tasks.

In our study, speech entrainment was defined in terms of a linear mapping between the speech envelope and the EEG signal. A decreased prediction accuracy for increasing WM load indicates that WM load influences how accurately cortical activity tracks acoustic amplitude variations in the speech signal. Such a picture is consistent with the notion of a general oscillatory network that modulates activity in sensory cortices in a top-down manner (Schroeder and Lakatos, 2009). While conceivable, this conclusion may be premature based on the current results in isolation. Delta-theta envelope entrainment has also been reported for non-speech signals or unintelligible speech sounds (Lalor et al., 2009b; Millman et al., 2015; O'Sullivan et al., 2014). In speech signals, however, the amplitude envelope correlates with the quasi-rhythmic variations in higher-level speech features, such as the onsets of phonemes or syllables. Cortical entrainment in speech processing has also been suggested to be related to parsing of such high-level speech units (Di Liberto et al., 2015; Ghitza, 2011; Giraud and Poeppel, 2012; Zoefel and VanRullen, 2016), and WM load could modulate speech processing at any or several different levels of speech processing.

Although we suggest that the effects of background noise on delta-theta entrainment reflect WM load, changes in entrainment could potentially have been related to the acoustic degradation of the sound envelope. To investigate whether a reduction in entrainment might reflect the fact that cortical activity entrains to the noisy signal, we compared entrainment to the clean speech signal (without noise) with entrainment to the noisy sound stimulus actually presented to the listeners. In agreement with previous results (Ding and Simon, 2013; Fuglsang et al., 2017), this suggested that cortical activity predominantly entrained to the underlying speech signal rather than to the noisy sound mixture. While this suggests an effect of WM load induced by the noise interference, our design does not allow us to completely dissociate the effects of acoustic degradation of the sound signal from inhibitory load caused by these degradations. Alternative paradigms that burden WM inhibitory load without affecting the acoustic stimulus, e.g. by presenting incongruent speech features, might further dissociate these effects.

6.6 Limitations

In our study, we used long continuous speech stimuli (∼45-55 s) to investigate auditory entrainment during WM load. However, simultaneously examining WM-dependent effects on speech entrainment and on oscillatory power involves a trade-off in terms of experimental design. TRF methods are generally found to be more robust to EEG artefacts but they require long trials for estimating the stimulus-response mapping at lower frequencies. Although the TRF methods allow neural responses to continuous speech to be examined, longer trials are not optimally suited to track spectral power changes in the EEG. Power estimates are more susceptible to EEG artefacts and activity unrelated to the stimulus or task. In the current study, we observed a substantial individual variability in the considered power measures. It is possible that alternative paradigms using shorter trials and more trial averages would be more sensitive to the oscillatory power changes associated with these WM tasks and could reveal additional effects. Also, the current analyses relied on EEG power estimates in fixed frequency bands, although the spectral characteristics of alpha and theta power may vary considerably between subjects (Haegens et al., 2014). The group analyses of theta-alpha power may thus be susceptible to between-band leakages.

6.7 Acknowledgements

This research was supported by the EU H2020-ICT Grant Number 644732 (COCOHA: Cognitive Control of a Hearing Aid) and the Oticon Centre of Excellence for Hearing and Speech Sciences (CHeSS). The authors would like to thank Ole Edmond Pedersen for assistance with the speech record- ings.

6.8 Conflict of interests

The authors declare no conflict of interests.

6.9 Author contributions

J.H. conceived the study; J.M., S.F. and J.H. designed the experiment; J.M collected the data; J.M., S.F. and J.H. analysed the data and interpreted the results;

and all authors wrote the paper.

6.10 Abbreviations

CDF, cumulative distribution function; EEG, electroencephalogram; EOG, electrooculogram; ERP, event-related potential; MEG, magnetoencephalo- gram; SNR, signal-to-noise ratio; TFR, time–frequency representation; TRF, temporal response function; WM, working memory.

7

Real-time enhancement of attended speech in an EEG-based brain-computer interface system: a case study

Chapter 3, 4 and 5 demonstrated that it is possible to decode the attentional selection of normal-hearing and hearing-impaired listeners in multi-talker environments from offline electroencephalogram (EEG) data. This opens the possibility for using EEG in brain-computer interface systems (BCIs) that allow listeners to adjust the relative level of attended and unattended speech via EEG correlates of auditory attention. However, to date, EEG-based classification of attended speech has only been demonstrated in "offline" listening scenarios where classification decision does not affect acoustic input, and whether attention-based enhancement of attended speech can be achieved in real-time remains unanswered. In particular, it remains unclear how a BCI-controlled dynamical change in acoustic feedback influences EEG correlates of auditory attention, and whether listeners can switch attention to an BCI-attenuated speech streams.

This chapter presents, as a case-study, a real-time BCI experiment in which a hearing-impaired listener performed an auditory attention-switching task in a closed-loop setup. A real-time processing system was designed for this purpose. First, the system synchronizes EEG and two pre-recorded speech audio streams. The system then uses a support vector machine (SVM) on canonical correlations between canonical variate pairs of EEG features and audio features to classify attended speech streams. The classifier decision function is then mapped to a relative gain of the two talkers. Both closed-loop (with acoustic feedback), and open-loop trials (without acoustic feedback) were considered. In 90 s long closed-loop and open-loop trials, the subject was asked to first

This chapter is based on work by Daniel D.E. Wong, Søren A. Fuglsang, Jonatan Marcher-Rørsted, Torsten Dau, Sergi Rotger Griful, Enea Ceoline, Alain de Cheveigné & Jens Hjortkjær.

attend to one speaker, and then cued to switch attention to the other speaker after 45 s. In closed-loop trials, the classifier decision function was mapped to a relative target-to-masker gain ratio of maximum 10 dB. The attention-switching paradigm allowed for an investigation of whether the subject could switch attention to previously ignored speech streams that had been attenuated by the BCI system.

7.1 Experimental details

The subject who took part in the experiment was a 53 year old male with a steeply-sloping symmetric audiometric profile (Fig. 7.1). Speech stimuli (audio-book material) consisting of two talkers (one male, one female) were presented over two loudspeakers at ($\pm 60^o$ relative to the position of the subject). The speech stimuli were normalized to have equal loudness when played without EEG feedback, and amplified to compensate for reduced audibility based on the subject's audiogram (Moore & Glasberg, 1998). 64-channel EEG data were recorded using a BioSemi ActiveTwo system. Additional electrodes were placed at the mastoids as well as below and above the right eye of the subject.

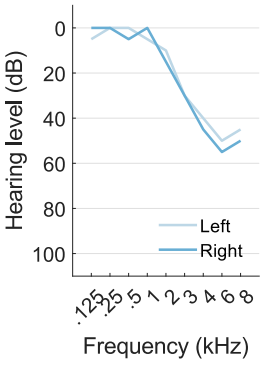

Figure 7.1: Pure-tone audiogram profile of the participant.

The experiment proceeded as follows. In 28 trials of 60 s duration, the subject was instructed to attend to either the male- or female talker. The spatial position of the male and female speaker, as well as the position of the target speaker was randomized. Classification models were trained on data from these trials and used in subsequent open-loop and closed-loop trials. Open-loop and closed-loop trials were grouped in two blocks of 12 trials. Each block of 12 trials

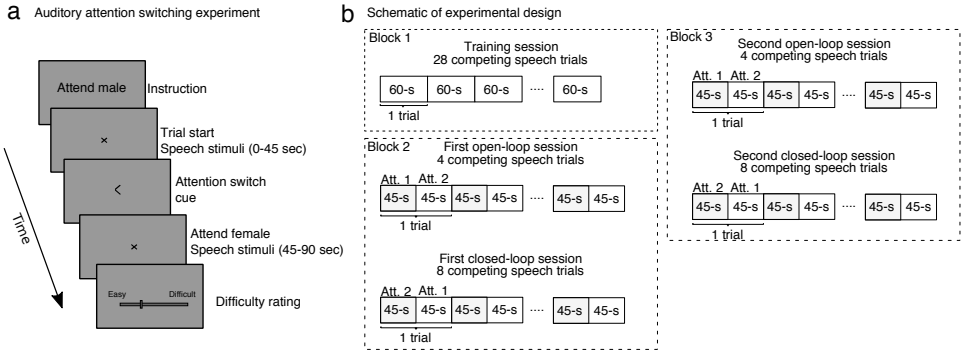

Figure 7.2: a Schematic illustration of the trial sequences in the attention-switching experiment. b Schematic of the experimental design. During the training session, the subject listened selectively to one of two competing speech streams in 60 s long trials. The EEG data recorded from the training session was used to train EEG-based attention classifiers. In 8 open-loop and 12 closed-loop trials, the subject listened to 90 s long competing speech streams and was cued to switch attentional focus to the competing speaker after 45 seconds. In the open-loop trials the classifier output did not affect the relative level difference between the two speech streams. In the closed-loop trials, the output from the classifier changed the relative levels of the individual speakers. The gender of the attended speech as well as the position of the attended speaker was in all cases randomized across trials.

consisted of 4 open-loop trials and 8 closed-loop trials. In the open-loop trials, the output from the classifier did not affect sound level of the two speech streams. In the closed-loop trials, the output of the classifier was mapped to relative gain differences between two speech streams. The subject was instructed to fixate gaze in the middle of the presentation screen during all trials.

7.1.1 Real-time decoding pipeline

A schematic of the real-time decoding pipeline is shown in figure 7.3. The decoding pipeline was implemented in OpenVibe (Renard et al., 2010). For real-time decoding, EEG and audio signals were streamed to OpenVibe (Renard et al., 2010) via Lab Streaming Layer (Kothe, 2013), synchronized in time and thereafter downsampled to 64 Hz. The EEG data were high-pass filtered and re-referenced to a common-average over all scalp EEG electrodes. Electroocular (EOG) artefacts were regressed out from the data via a pretrained spatial filter. The envelopes of the two speech audio streams were extracted and power-law compressed ($c = 0.3$). The speech envelopes and the EEG data were both processed through a dyadic filterbank that was formed via multiple smoothing kernels and first order forward differencing. Whitening and PCA transforms obtained from the training data was applied to whiten the data

and reduce dimensionality. The pre-processed audio features were offset by a fixed time lag relative to the EEG data. The EEG and audio were then transformed using transform matrices obtained with canonical correlation analysis (CCA) (de Cheveigné et al., 2018a). CCA transform matrices were applied separately to features in EEG data and each audio stream. Correlation scores were computed over 8 s long windows for each canonical correlate pair (separately for each speech stream). A pretained SVM classifier was then used to classify attended speech from z-scored correlation scores. The classifier output was nonlinearly transformed using a double-sided Weibull function. The output from the Weibull function was used to control the relative sound pressure level of each of the sound streams. This nonlinear transformation ensured that unreliable classification decisions did not have a strong effect on audio level, while simultaneously mapping "near-certain" classification decisions to maximum (10 dB) sound pressure level enhancements.

The EOG denoising filter, the whitening matrices, the CCA transform matrices and the SVM classifier were trained offline in a Matlab implementation that replicated the real-time decoding pipeline (Wong et al., 2018; de Cheveigné et al., 2018a).

7.2 Results

7.2.1 Results from open-loop experiments

Figure 7.5 shows the results from the open-loop experiment. The data in this figure has been sign-corrected such that the first 45 s of each trial correspond to "attend speaker 1" regardless of the speaker identify or speaker location. The left panel in Figure 7.4 shows the output of the SVM classifier in eight open-loop trials. This panel depicts the classifier output per trial over the entire trial duration. After the attention switch at 45 s, there is a sign change in the average classifier decision function at 55.3 s. The classifier assigned correct labels to each of the two speakers 74 % of the time. When excluding data from each trial in periods from 0 s to 10 s and from 45 s to 55 s, the classifier assigned correct labels to each of the two speakers 84 % of the time.

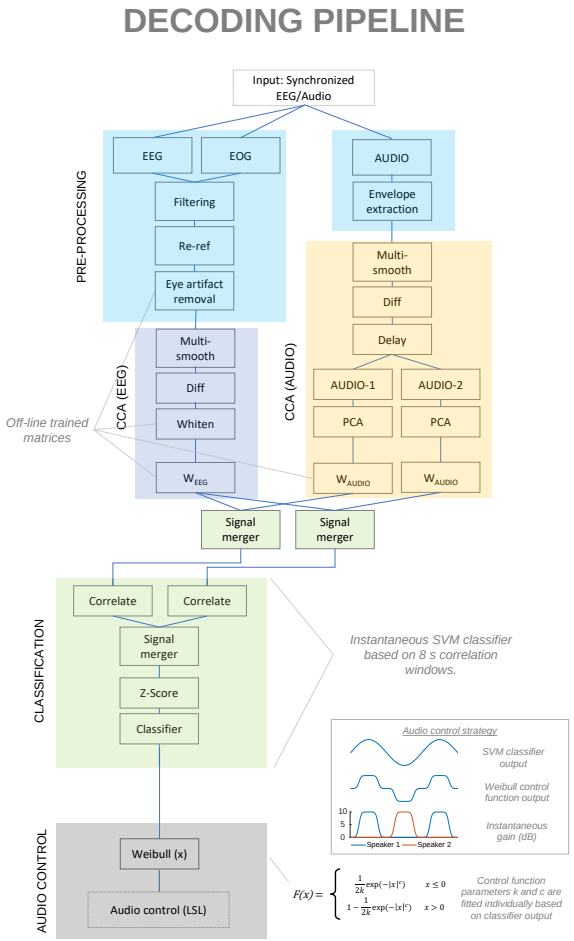

Figure 7.3: Schematic overview over the CCA-based decoding pipeline implemented in OpenVibe. Each box represents a python processing module. The decoder receives synchronized EEG and audio streams. The two audio streams are separately processed (right branch). Envelopes of both audio streams are extracted and downsampled to 64 Hz. Next, the envelopes are filtered with smoothing kernels and each transformed using CCA projection vectors. The EEG data (left branch) are filtered, denoised, passed through a dyadic filterbank and finally transformed according to the CCA mapping. Each of the canonical components is correlated pairwise, such that the EEG components are correlated with the components extracted from both of the audio streams. These correlations can subsequently be mapped into a classification decision using the SVM classifier that had been trained on the offline data recorded from the training session. Finally, the classifier decision function is nonlinearly transformed using a double-sided Weibull function. The output from the Weibull function is used as a control stream that adjusts the relative sound pressure level difference between the two speech streams.

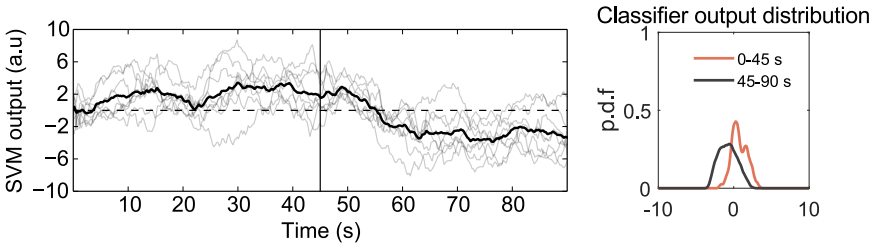

Figure 7.4: Results from open-loop experiments. In this figure, the output from the classifier has been sign-corrected such that output from the classifier over the first 45 s always correspond to "attend speaker 1" regardless of the speaker identify or speaker location. Left panel: output of SVM classifier in each of the 8 open-loop trials. Gray traces here reflect data from individual trials and black line indicates average across trials. Right panel: distribution over the classifier decision function output from all trials in the time period from 0 s to 45 s post trial onset (orange curve) and from 45 s to 90 s post-trial onset (black curve).

7.2.2 Results from closed-loop experiments

Fig 5. shows the results from the 12 closed-loop trials where the classifier decision function mapped out to relative sound enhancements of the two competing audio streams. The data in this figure has been sign-corrected as in Fig. 7.4, such that the first 45 s of each trial correspond to "attend speaker 1" regardless of the speaker identify or speaker location. Fig 7.5a show the classifier decision function for each of the 12 trials. After the attention switch at 45 s, there is a sign change in the average classifier decision function at 48.9 s. In closed-loop trials, the classifier assigned correct labels to each of the two speakers 77 % of the time (Fig. 7.5d). On average, the BCI system produced an relative increase in the sound pressure level of the attended speech of 3.63 dB (Fig 7.5b). When excluding data from each trial in periods from 0 s to 10 s and from 45 s to 55 s, the relative increase in the sound pressure level of the attended speech was 4.24 dB.

7.2.3 Subjective ratings in open-loop and closed-loop trials

At the end of each trial the subject was asked to answer questions about task difficulty and speech understanding. Here, the subject responded to each question with a score between 0 and 10. When asked how much effort it required to follow the attended speaker (0 = no effort, 10 = high effort), the participant obtained a score of 4.9 (moderate effort) in open-loop trials, but 3.6 in closed-loop trials. Similarly, when asked how much effort it required to switch attention from one

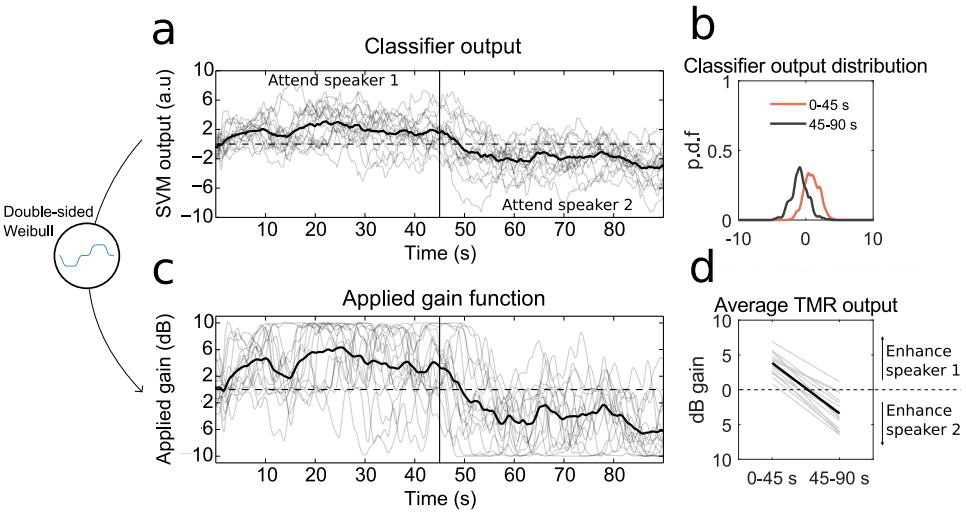

Figure 7.5: Results from closed-loop experiments. a: output of SVM classifier each of the closed-loop trials. b: Distribution of the classifier output over all trials in the time period from 10 s to 45 s post trial onset (orange curve) and from 55 s to 90 s post-trial onset (black curve). c: Applied gain (dB) to each of the two speakers in both trials. c: Average applied gain (in dB) to the two speakers in the first half of each trial (from 0 s to 45 s post trial onset) and in the second half of each trial (from 45 s to 90 s post trial onset).

speaker to the other, the subject scored low in both open-loop trials (2.5) and closed-loop trials (2.45) indicating that it required little effort. When asked to rate speech understanding (0 = nothing understood, 10 = everything understood) the subject scored 6.2 in open-loop trials and 7.1 in closed-loop trials, potentially indicating an improved speech understanding with the closed-loop BCI system.

7.3 Discussion and future work

This BCI case study demonstrates that real-time EEG-based attention decoding can be achieved in a hearing-impaired listener both in a scenario where classifier output does not affect stimulus input (open-loop scenarios), but also in a scenario where the classifier output controls the relative level difference between attended and unattended speech. The latter scenario constitutes a closed-loop system, as classifier decision depends on the audio input.

A crucial aspect of EEG-based attention decoding is how robustly and rapidly the BCI system detects attention switches. The performance and speed of this BCI system is, however, limited by the length of the windows used to compute

canonical correlation coefficients as well as the duration of the EEG/envelope smoothening kernels. In the present study, we chose decision windows and kernel lengths inspired by findings in a recent offline study (Wong et al., 2018). However, although the BCI system was relatively robust at decoding attended speech streams, the speed of the BCI system was rather limited. More work will be needed to investigate whether similar performance can be obtained with shorter decision windows in closed-loop scenarios.

In the present study, the output from the classifier could at most amplify attended speech with 10 dB. This choice was based on preliminary findings from BCI sessions with normal-hearing subjects. If the relative enhancement of attended speech becomes too pronounced, it may be impossible to switch attention to the suppressed speech stream. It is possible that other ways to map from classifier output to relative sound level differences between attended and unattended speech may improve speech recognition and BCI performance.

It is possible that neurofeedback in closed-loop scenarios may enable listeners to rely on other strategies to improve the neurofeedback. The BCI-feedback could, for example, result in increased listening effort or motivation during closed-loop control. This would, in turn, potentially affect both classifier performance, but also speech understanding. It remains to be clarified whether BCI operation is affected by user training during closed-loop experiments.

7.4 Conclusion

The results from this BCI case study demonstrate that real-time EEG-based attention decoding could be achieved in a hearing-impaired listener both in open-loop and in closed-loop scenarios. The classification scheme considered here was based on a stimulus-response model that mapped between envelopes of attended speech and EEG data. This work provides a step further towards a BCI system with real-time EEG-based attentional control over acoustic scene analysis hardware.

8

Overall discussion

8.1 Summary of main findings

Chapter 3 investigated whether a sensorineural hearing impairment influences cortical EEG responses to tones and to speech. A cohort of twenty-two older adults with sensorineural hearing loss and twenty-three age-matched normal-hearing controls took part in the study. Each participant underwent a battery of clinical, psychophysical and electrophysiological tests. Results from psychophysical tests suggested HI-related deficits in the identification of speech in noise even when the speech signal was amplified to compensate for reduced audibility. Similarly, the HI listeners exhibited less temporal release from masking compared to age-matched NH listeners in a tone detection task. In a selective speech-listening EEG experiment with two loudness-matched competing speech streams, comprehension questions suggested equally good speech comprehension in both listener groups. However, the HI listeners rated the two-talker listening task as being more difficult. Results from stimulus-response analyses suggested that the fidelity of the cortical EEG entrainment to envelopes of attended speech was enhanced in the HI listener group compared to the NH listener group. In addition, it was possible to classify attended speakers from EEG responses to speech mixtures with equally high classification accuracies in both listener groups. To further investigate whether HI-related effects on cortical EEG responses would be observed for non-speech sounds, EEG responses were also obtained to tone sequences unfolding on fast (gamma range) and slow (theta range) time scales during passive listening. The results from these experiments suggested a more precise theta-range EEG phase-locking to rhythmic tone stimuli in HI compared to NH listeners. No effect of HI on N1 amplitudes was observed for ERPs evoked by tones. Overall, these results suggested that, even though a sensorineural hearing loss may be associated with exaggerated sound-evoked cortical EEG responses, it does not necessarily hinder attentional modulations of cortical auditory entrainment responses.

Chapter 4 extended previous investigations on the neural correlates of selective attention to speech (Ding and Simon, 2012a,b; Mesgarani and Chang, 2012; Mirkovic et al., 2015; O'Sullivan et al., 2014), to study EEG correlates of cortical speech envelope entrainment in noisy and reverberant acoustic settings. Single-trial EEG activity was recorded from normal-hearing subjects listening selectively to one of two or more competing speakers in anechoic and reverberant environments. It was demonstrated that spatiotemporal reconstruction filters trained to reconstruct envelopes of attended speech from EEG activity could be used to discriminate between attended and unattended speakers with equal classification accuracy in anechoic and reverberant environments. EEG-based attention classification accuracies were in the same order of magnitude as those reported in previous studies with anechoic speech mixtures (Mirkovic et al., 2015; O'Sullivan et al., 2014). Moreover, the EEG entrainment to envelopes of attended speech was stable against competing speech both in anechoic and reverberant listening environments. These results suggested that the cortical EEG correlates of entrainment to envelopes of attended speech are relatively robust in naturalistic listening environments.

Chapter 5 investigated how different regularization schemes affect stimulus-response mappings between envelopes of competing speech streams and EEG responses. Five different regularization schemes and two different stimulus-response model types (forward/backward models) were considered. Since the stimulus-response models may be used to discriminate between attended and unattended speech, the model performance was evaluated on a classification task where the goal was to infer which of two speech streams a listener attends to (e.g., in terms of attention classification accuracies). Moreover, the regression model performance was evaluated by means of how robustly it could predict unseen data (regression accuracies). It was found that the regularized backward models outperformed regularized forward models in terms of classification accuracies. Whereas regularization did not have an effect on regression or classification accuracies for forward models, it was found to improve regression and classification accuracies for backward models. The results suggested that some regularization methods may lead to better regression accuracies for backward models that have a higher number of free parameters. However, the increased regression accuracies do not necessarily entail better attention classification accuracies.

In chapter 6, the effects of cognitive task load on continuous speech EEG

measures were investigated. Multi-channel EEG were recorded from younger normal-hearing listeners performing an auditory n-back task on speech sequences in different levels of background noise. Two different n-back levels and two different background noise levels were considered. The n-back level was varied to induce a load of working memory (WM) and the noise level was varied to impose increased inhibitory demands on WM. In the 1-back conditions, listeners were instructed to report repeated digits, and had to continuously update digits maintained in WM. In the 2-back conditions, the listeners had to detect whether digits matched those presented two digits earlier. In 2-back conditions, the subjects had to update digits maintained in WM and to shift attentional focus between to-be-ignored and to-be-compared digits (Scharinger et al., 2017). The findings suggested that n-back level and noise-level may affect not only spatio-spectral EEG power, but also the cortical EEG correlates of speech envelope processing. Consistent with results from visual n-back paradigms, the results suggested that posterior alpha band power (8-13 Hz) decreases with n-back level, while frontal theta (4-8 Hz) power increases with n-back level. On the other hand, speech envelope entrainment as measured with EEG forward models was found to decrease with both higher noise level and higher n-back level. These results could support the idea of a task-dependent coupling between alpha or theta power and speech entrainment (Keitel et al., 2017), although such couplings were not investigated quantitatively.

In chapter 7, a real-time attention switching BCI paradigm was discussed. This case study was exploratory in nature and was intended as a proof-of-principle to demonstrate closed-loop steering of acoustic feedback based on online EEG attention decoding. An EEG-based attention decoder with 8-s long classification windows was here trained offline on 28 minutes of EEG data prerecorded from a hearing-impaired subject listening selectively to one of two competing speech streams. The attention decoder was then used to classify attended speech in real-time in trials where the subject performed an attention switching task on spatially separated competing speech streams. The results suggested that the hearing-impaired listener was able to selectively amplify attended speech using closed-loop BCI control.

8.2 Limitations of the approach chosen in this work

The work presented in this thesis focused on EEG responses to continuous speech stimuli. Recently there has been an increased interest in the use of continuous speech stimuli and naturalistic sound stimuli in auditory neuroscientific research (Theunissen and Elie, 2014), including in human EEG studies (Broderick et al., 2018; Crosse and Lalor, 2014; Crosse et al., 2015; Khalighinejad et al., 2017; Mirkovic et al., 2016; O'Sullivan et al., 2017a; O'Sullivan et al., 2014), fMRI studies (Heer et al., 2017; Huth et al., 2016; Kell et al., 2018; Santoro et al., 2014, 2017), ECoG studies (Golumbic et al., 2013; Hullett et al., 2016; Mesgarani and Chang, 2012; Mesgarani et al., 2014b; Pasley et al., 2012) and MEG studies (Ding and Simon, 2012a,b, 2013; Luo and Poeppel, 2007; Puvvada and Simon, 2017). It seems plausible that results from neuroscientific experiments with speech stimuli may better generalize to everyday listening scenarios than results obtained in experiments with synthetic sounds. However, as described in Appendix A, there can be a number of challenges in the interpretation of encoding- and decoding results from experiments with speech stimuli. Auditory processing of speech involves various complex, interrelated hierarchical processing steps. Although encoding models theoretically can provide a full functional description of what task- or stimulus-related features that are represented in the neural data recordings (and how much of the explainable variance in the data they account for; Naselaris et al., 2011), it is not possible to exhaust all possible speech features that are represented in the neural recordings. The stimulus-response models can erroneously signify stimulus representations due to correlated stimulus features that have not been considered in the analysis (see discussion in Appendix A) or due to nonlinear computations that are not captured by the models (Christianson et al., 2008).

In the same vein, there can be conceptual caveats entailed by experiments on task-driven auditory attention (Fritz et al., 2007) and it can be difficult to experimentally monitor the attentional focus of listeners. This is especially relevant for experimental paradigms with prolonged trials and dynamic stimuli, such as the experiments considered in this thesis. The main motivation for considering experiments with prolonged speech stimuli was that they may represent naturalistic listening scenarios better than experiments with synthetic stimuli. In experimental paradigms with shorter trials, it may be more straightforward to assess attentional focus from behavioural data and task de-

sign. However, in such experiments, the stimuli may not be as ”engaging” and may lack the statistical properties of naturalistic sounds. Moreover, it is in these cases not always clear whether the drawn conclusions generalize to even more complex listening scenarios (Churchland et al., 2010). There exists, as often the case, a trade-off between experimental control and task naturalness.

For the analysis of single-trial EEG responses to speech mixtures, the work presented in this thesis focused mainly on slow (<10 Hz) modulations in the speech signals (in line with previous work, Crosse et al., 2015; Ding and Simon, 2012a,b, 2013; Mirkovic et al., 2016; O'Sullivan et al., 2014; Pasley et al., 2012 and previous theoretical models proposed by Ghitza, 2011; Giraud and Poeppel, 2012). Such slow envelope modulations are thought to provide important cues for speech perception (Drullman et al., 1994; Jørgensen et al., 2013). However, M/EEG measures of cortical auditory entrainment to speech envelopes may reflect distinct neural computations, such as neural computations related to parsing syllables, detection of edges in the envelopes or to perceived loudness (Ding and Simon, 2014). M/EEG envelope entrainment may even be influenced by bottom-up adaptive processes, binaural fusion, listening effort or listening context (e.g. listening cues or predictability). What has been referred to as ”auditory EEG envelope entrainment” in this work could therefore be epiphenomenal to other correlated brain processes. Other experimental paradigms would be needed to establish the causal nature of these observations.

Finally, as for any correlational neuroscientific study, the observation of a feature that appears to be represented in the neural data recordings (e.g. manifested in high encoding accuracies) does not necessarily imply that this feature has relevance to the brain/listener (Wit et al., 2016), but may only support speculations about why the feature is encoded in data. Similarly, if the neural data do not represent stimulus-related information, this does not necessarily imply that the underlying neural circuits do not represent such information (e.g. Haynes, 2015).

8.3 Perspectives

In the single-trial EEG experiments considered in this thesis, the main focus was on listening environments with near-intact speech recognition. However, in challenging listening environments with more unfavourable signal-to-noise ratios, it may be impossible to segregate target sounds from listening background.

This may be reflected in cortical auditory M/EEG entrainment responses and can affect EEG-based attention decoding. Future studies will be needed to clarify whether it is possible to classify attended speech with BCI systems in even more challenging listening environments. Moreover, the premise for the work in this thesis was that the attention decoders have access to non-corrupted versions of the attended and unattended speech. In order for the EEG-based stimulus-response models to have relevance for portable BCI systems in everyday life, the system must also address the acoustic scene analysis problem, for instance, by using microphone beamforming techniques to recover the individual sound sources (O'Sullivan et al., 2017b; Van Eyndhoven et al., 2017). Finally, although it has been established that the attentional selection of listeners can be decoded from offline single-trial EEG data, it still remains to be clarified whether similar results can be achieved in a variety of closed-loop scenarios (e.g. Chapter 8) with fewer electrodes or even with portable EEG systems outside laboratory settings (De Vos et al., 2014; Debener et al., 2012). Future studies will be needed to investigate whether other decoding strategies (e.g., (Akram et al., 2016; Miran et al., 2018; de Cheveigné et al., 2018a)) can improve EEG-based auditory attention decoding accuracies.

Bibliography

Aertsen, A. and P. Johannesma (1981). "The spectro-temporal receptive field". In: *Biological cybernetics* 42.2, pp. 133–143.

Agrawal, Y., E. A. Platz, and J. K. Niparko (2008). "Prevalence of hearing loss and differences by demographic characteristics among US adults: data from the National Health and Nutrition Examination Survey, 1999-2004". In: *Archives of internal medicine* 168.14, pp. 1522–1530.

Ahissar, E., S. Nagarajan, M. Ahissar, A. Protopapas, H. Mahncke, and M. M. Merzenich (2001). "Speech comprehension is correlated with temporal response patterns recorded from auditory cortex". In: *Proceedings of the National Academy of Sciences* 98.23, pp. 13367–13372.

Akeroyd, M. A. (2008). "Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults". In: *International journal of audiology* 47.sup2, S53–S71.

Akram, S., A. Presacco, J. Z. Simon, S. A. Shamma, and B. Babadi (2016). "Robust decoding of selective auditory attention from MEG in a competing-speaker environment via state-space modeling". In: *NeuroImage* 124, pp. 906–917.

Alain, C. (2014). "Effects of age-related hearing loss and background noise on neuromagnetic activity from auditory cortex". In: *Frontiers in systems neuroscience* 8, p. 8.

Aroudi, A. and S. Doclo (2017). "EEG-based Auditory Attention Decoding: Impact of Reverberation, Noise and Interference Reduction". In: *39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*.

Atencio, C. A., T. O. Sharpee, and C. E. Schreiner (2008). "Cooperative nonlinearities in auditory cortical neurons". In: *Neuron* 58.6, pp. 956–966.

Baddeley, A. (2003). "Working memory: looking back and looking forward". In: *Nature reviews neuroscience* 4.10, pp. 829–839.

Bates, D., M. Mächler, B. Bolker, and S. Walker (2014). "Fitting linear mixed-effects models using lme4". In: *arXiv preprint arXiv:1406.5823*.

Beatty, J. (1982). "Task-evoked pupillary responses, processing load, and the structure of processing resources." In: *Psychological bulletin* 91.2, p. 276.

Bell, A. J. and T. J. Sejnowski (1995). "An information-maximization approach to blind separation and blind deconvolution". In: *Neural computation* 7.6, pp. 1129–1159.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300.

Bertoli, S., J. Smurzynski, and R. Probst (2005). "Effects of age, age-related hearing loss, and contralateral cafeteria noise on the discrimination of small frequency changes: psychoacoustic and electrophysiological measures". In: *Journal of the Association for Research in Otolaryngology* 6.3, pp. 207–222.

Bialek, W., F. Rieke, R. R.d. R. van Steveninck, and D. Warland (1990). "Reading a neural code". In: *Advances in neural information processing systems*, pp. 36–43.

Bidet-Caulet, A., C. Fischer, J. Besle, P.-E. Aguera, M.-H. Giard, and O. Bertrand (2007). "Effects of selective attention on the electrophysiological representation of concurrent sounds in the human auditory cortex". In: *Journal of Neuroscience* 27.35, pp. 9252–9261.

Biessmann, F., S. Plis, F. C. Meinecke, T. Eichele, and K.-R. Muller (2011). "Analysis of multimodal neuroimaging data". In: *IEEE Reviews in Biomedical Engineering* 4, pp. 26–58.

Billings, C. J. and B. M. Madsen (2018). "A perspective on brain-behavior relationships and effects of age and hearing using speech-in-noise stimuli". In: *Hearing research*.

Billings, C. J., K. L. Tremblay, P. E. Souza, and M. A. Binns (2007). "Effects of hearing aid amplification and stimulus intensity on cortical auditory evoked potentials". In: *Audiology and Neurotology* 12.4, pp. 234–246.

Bisgaard, N., M. S. Vlaming, and M. Dahlquist (2010). "Standard audiograms for the IEC 60118-15 measurement procedure". In: *Trends in amplification* 14.2, pp. 113–120.

Blackburn, H. L. and A. L. Benton (1957). "Revised administration and scoring of the digit span test." In: *Journal of consulting psychology* 21.2, p. 139.

Blankertz, B., S. Lemm, M. Treder, S. Haufe, and K.-R. Müller (2011). "Single-trial analysis and classification of ERP components: a tutorial". In: *NeuroImage* 56.2, pp. 814–825.

Borga, M. (1998). "Learning multidimensional signal processing". PhD thesis. Linköping University Electronic Press.

Broderick, M., A. Anderson, G. Di Liberto, M. Crosse, and E. Lalor (2018). "Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech". In: *Curr. Biol.* 28.5, 803–809.e3.

Burns, M. D., N. Bigdely-Shamlo, N. J. Smith, K. Kreutz-Delgado, and S. Makeig (2013). "Comparison of averaging and regression techniques for estimating event related potentials". In: *Engineering in medicine and biology society (EMBC), 2013 35th annual international conference of the IEEE.* IEEE, pp. 1680–1683.

Calabrese, A., J. Schumacher, D. Schneider, L. Paninski, and S. Woolley (2011). "A generalized linear model for estimating spectrotemporal receptive fields from responses to natural sounds". In: *PLoS One* 6.1, e16104.

Carney, L. H. (2018). "Supra-Threshold Hearing and Fluctuation Profiles: Implications for Sensorineural and Hidden Hearing Loss". In: *Journal of the Association for Research in Otolaryngology*, pp. 1–22.

Caspary, D. M., L. Ling, J. G. Turner, and L. F. Hughes (2008). "Inhibitory neurotransmission, plasticity and aging in the mammalian central auditory system". In: *Journal of Experimental Biology* 211.11, pp. 1781–1791.

Chambers, A. R. et al. (2016). "Central gain restores auditory processing following near-complete cochlear denervation". In: *Neuron* 89.4, pp. 867–879.

Christianson, G. B., M. Sahani, and J. F. Linden (2008). "The consequences of response nonlinearities for interpretation of spectrotemporal receptive fields". In: *Journal of Neuroscience* 28.2, pp. 446–455.

Churchland, M. M. et al. (2010). "Stimulus onset quenches neural variability: a widespread cortical phenomenon". In: *Nature neuroscience* 13.3, p. 369.

Clayson, P. E., S. A. Baldwin, and M. J. Larson (2013). "How does noise affect amplitude and latency measurement of event-related potentials (ERPs)? A methodological critique and simulation study". In: *Psychophysiology* 50.2, pp. 174–186.

Crosse, M., G. Di Liberto, and E. Lalor (2016a). "Eye can hear clearly now: inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration". In: *J. Neurosci.* 36.38, pp. 9888–9895.

Crosse, M., G. Di Liberto, A. Bednar, and E. Lalor (2016b). "The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli". In: *Front. Hum. Neurosci.* 10, p. 604.

Crosse, M. J. and E. C. Lalor (2014). "The cortical representation of the speech envelope is earlier for audiovisual speech than audio speech". In: *Journal of neurophysiology* 111.7, pp. 1400–1408.

Crosse, M. J., J. S. Butler, and E. C. Lalor (2015). "Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions". In: *Journal of Neuroscience* 35.42, pp. 14195–14204.

Cruickshanks, K. J. et al. (1998). "Prevalence of hearing loss in older adults in Beaver Dam, Wisconsin: The epidemiology of hearing loss study". In: *American journal of epidemiology* 148.9, pp. 879–886.

Dai, L., V. Best, and B. G. Shinn-Cunningham (2018). "Sensorineural hearing loss degrades behavioral and physiological measures of human spatial selective auditory attention". In: *Proceedings of the National Academy of Sciences*, p. 201721226.

Darwin, C. and R. Hukin (2000). "Effects of reverberation on spatial, prosodic, and vocal-tract size cues to selective attention". In: *The Journal of the Acoustical Society of America* 108.1, pp. 335–342.

Das, N., S. Van Eyndhoven, T. Francart, and A. Bertrand (2016). "Adaptive attention-driven speech enhancement for EEG-informed hearing prostheses". In: *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2016, pp. 77–80.

Dau, T. (2003). "The importance of cochlear processing for the formation of auditory brainstem and frequency following responses". In: *The Journal of the Acoustical Society of America* 113.2, pp. 936–950.

David, S., W. Vinje, and J. Gallant (2004). "Natural stimulus statistics alter the receptive field structure of v1 neurons". In: *J. Neurosci.* 24.31, pp. 6991–7006.

David, S., N. Mesgarani, and S. Shamma (2007). "Estimating sparse spectro-temporal receptive fields with natural stimuli". In: *Netw. Comput. Neural Syst.* 18.3, pp. 191–212.

David, S. V. and J. L. Gallant (2005). "Predicting neuronal responses during natural vision". In: *Network: Computation in Neural Systems* 16.2-3, pp. 239–260.

Davis, A. (1995). *Hearing in adults: the prevalence and distribution of hearing impairment and reported hearing disability in the MRC Institute of Hearing Research's National Study of Hearing*. Whurr Publishers London.

De Vos, M., K. Gandras, and S. Debener (2014). “Towards a truly mobile auditory brain–computer interface: exploring the P300 to take away”. In: *International journal of psychophysiology* 91.1, pp. 46–53.

Debener, S., F. Minow, R. Emkes, K. Gandras, and M. De Vos (2012). “How about taking a low-cost, small, and wireless EEG for a walk?” In: *Psychophysiology* 49.11, pp. 1617–1621.

Delorme, A. and S. Makeig (2004). “EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis”. In: *Journal of neuroscience methods* 134.1, pp. 9–21.

Deng, S., R. Srinivasan, T. Lappas, and M. D’Zmura (2010). “EEG classification of imagined syllable rhythm using Hilbert spectrum methods”. In: *Journal of neural engineering* 7.4, p. 046006.

Di Liberto, G. M., J. A. O’Sullivan, and E. C. Lalor (2015). “Low-frequency cortical entrainment to speech reflects phoneme-level processing”. In: *Current Biology* 25.19, pp. 2457–2465.

Diepen, R. M. and A. Mazaheri (2018). “The Caveats of observing Inter-Trial Phase-Coherence in Cognitive Neuroscience”. In: *Scientific reports* 8.1, p. 2990.

Ding, N. and J. Z. Simon (2012a). “Emergence of neural encoding of auditory objects while listening to competing speakers”. In: *Proceedings of the National Academy of Sciences* 109.29, pp. 11854–11859.

Ding, N. and J. Z. Simon (2012b). “Neural coding of continuous speech in auditory cortex during monaural and dichotic listening”. In: *Journal of neurophysiology* 107.1, pp. 78–89.

Ding, N. and J. Z. Simon (2013). “Adaptive temporal encoding leads to a background-insensitive cortical representation of speech”. In: *Journal of Neuroscience* 33.13, pp. 5728–5735.

Ding, N. and J. Z. Simon (2014). “Cortical entrainment to continuous speech: functional roles and interpretations”. In: *Frontiers in human neuroscience* 8.

Ding, N., M. Chatterjee, and J. Z. Simon (2014). “Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure”. In: *Neuroimage* 88, pp. 41–46.

Dmochowski, J. P., M. A. Bezdek, B. P. Abelson, J. S. Johnson, E. H. Schumacher, and L. C. Parra (2014). “Audience preferences are predicted by temporal reliability of neural processing”. In: *Nature communications* 5, p. 4567.

Dmochowski, J. P., J. J. Ki, P. DeGuzman, P. Sajda, and L. C. Parra (2017). "Extracting multidimensional stimulus-response correlations using hybrid encoding-decoding of neural activity". In: *NeuroImage*.

Doelling, K. B., L. H. Arnal, O. Ghitza, and D. Poeppel (2014). "Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing". In: *Neuroimage* 85, pp. 761–768.

Drullman, R., J. M. Festen, and R. Plomp (1994). "Effect of reducing slow temporal modulations on speech reception". In: *The Journal of the Acoustical Society of America* 95.5, pp. 2670–2680.

Dubno, J. R., D. D. Dirks, and D. E. Morgan (1984). "Effects of age and mild hearing loss on speech recognition in noise". In: *The Journal of the Acoustical Society of America* 76.1, pp. 87–96.

Favrot, S. E., J. Buchholz, and T. Dau (2010). *A loudspeaker-based room auralization system for auditory research*. Technical University of Denmark. Danmarks Tekniske Universitet, Centre for Applied Hearing ResearchCentre for Applied Hearing Research.

Foxe, J. J. and A. C. Snyder (2011). "The role of alpha-band brain oscillations as a sensory suppression mechanism during selective attention". In: *Frontiers in psychology* 2.

Friedman, J. (1989). "Regularized discriminant analysis". In: *J. Am. Stat. Assoc.* 84.405, pp. 165–175.

Frisina, D. R. and R. D. Frisina (1997). "Speech recognition in noise and presbycusis: relations to possible neural mechanisms". In: *Hearing research* 106.1-2, pp. 95–104.

Friston, K. J. (2009). "Modalities, modes, and models in functional neuroimaging". In: *Science* 326.5951, pp. 399–403.

Fritz, J. B., M. Elhilali, S. V. David, and S. A. Shamma (2007). "Auditory attention focusing the searchlight on sound". In: *Current opinion in neurobiology* 17.4, pp. 437–455.

Fritz, J., S. Shamma, M. Elhilali, and D. Klein (2003). "Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex". In: *Nature neuroscience* 6.11, p. 1216.

Fuglsang, S., D. Wong, and J. Hjortkjær (2018). "Data from: EEG and audio dataset for auditory attention decoding". In: *Zenodo*.

Fuglsang, S. A., T. Dau, and J. Hjortkjær (2017). "Noise-robust cortical tracking of attended speech in real-world acoustic scenes". In: *NeuroImage*.

Füllgrabe, C. and S. Rosen (2016). "On the (un) importance of working memory in speech-in-noise processing for listeners with normal hearing thresholds". In: *Frontiers in psychology* 7, p. 1268.

Gatehouse, S. and W. Noble (2004). "The speech, spatial and qualities of hearing scale (SSQ)". In: *International journal of audiology* 43.2, pp. 85–99.

Gazzaley, A. and A. C. Nobre (2012). "Top-down modulation: bridging selective attention and working memory". In: *Trends in cognitive sciences* 16.2, pp. 129–135.

Gevins, A. and M. E. Smith (2000). "Neurophysiological measures of working memory and individual differences in cognitive ability and cognitive style". In: *Cerebral cortex* 10.9, pp. 829–839.

Gevins, A. et al. (1996). "High resolution evoked potential imaging of the cortical dynamics of human working memory". In: *Electroencephalography and clinical neurophysiology* 98.4, pp. 327–348.

Gevins, A., M. E. Smith, L. McEvoy, and D. Yu (1997). "High-resolution EEG mapping of cortical activation related to working memory: effects of task difficulty, type of processing, and practice." In: *Cerebral cortex (New York, NY: 1991)* 7.4, pp. 374–385.

Ghitza, O. (2011). "Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm". In: *Frontiers in psychology* 2.

Ghitza, O. and S. Greenberg (2009). "On the possible role of brain rhythms in speech perception: intelligibility of time-compressed speech with periodic and aperiodic insertions of silence". In: *Phonetica* 66.1-2, pp. 113–126.

Giraud, A.-L. and D. Poeppel (2012). "Cortical oscillations and speech processing: emerging computational principles and operations". In: *Nature neuroscience* 15.4, pp. 511–517.

Glasberg, B. R. and B. C. Moore (1990). "Derivation of auditory filter shapes from notched-noise data". In: *Hearing research* 47.1-2, pp. 103–138.

Goldstein Jr, M. H. and N. Y.-S. Kiang (1958). "Synchrony of neural activity in electric responses evoked by transient acoustic stimuli". In: *The Journal of the Acoustical Society of America* 30.2, pp. 107–114.

Golumbic, E. M. Z. et al. (2013). "Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party"". In: *Neuron* 77.5, pp. 980–991.

Goossens, T., C. Vercammen, J. Wouters, and A. v. Wieringen (2016). "Aging affects neural synchronization to speech-related acoustic modulations". In: *Frontiers in aging neuroscience* 8, p. 133.

Goutte, C., F. Nielsen, and K. Hansen (2000). "Modeling the hemodynamic response in fMRI using smooth FIR filters". In: *IEEE Trans. Med. Imag.* 19.12, pp. 1188–1201.

Haegens, S., H. Cousijn, G. Wallis, P. J. Harrison, and A. C. Nobre (2014). "Inter- and intra-individual variability in alpha peak frequency". In: *Neuroimage* 92, pp. 46–55.

Händel, B. F., T. Haarmeier, and O. Jensen (2011). "Alpha oscillations correlate with the successful inhibition of unattended stimuli". In: *Journal of cognitive neuroscience* 23.9, pp. 2494–2502.

Harkrider, A. W., P. N. Plyler, and M. S. Hedrick (2006). "Effects of hearing loss and spectral shaping on identification and neural response patterns of stop-consonant stimuli". In: *The Journal of the Acoustical Society of America* 120.2, pp. 915–925.

Hasson, U., Y. Nir, I. Levy, G. Fuhrmann, and R. Malach (2004). "Intersubject synchronization of cortical activity during natural vision". In: *science* 303.5664, pp. 1634–1640.

Hasson, U., A. A. Ghazanfar, B. Galantucci, S. Garrod, and C. Keysers (2012). "Brain-to-brain coupling: a mechanism for creating and sharing a social world". In: *Trends in cognitive sciences* 16.2, pp. 114–121.

Hastie, T., R. Tibshirani, and J. Friedman (2001). "Linear methods for regression". In: *The elements of statistical learning theory*. New York: Springer. Chap. 3, pp. 43–100.

Haufe, S. et al. (2014). "On the interpretation of weight vectors of linear models in multivariate neuroimaging". In: *Neuroimage* 87, pp. 96–110.

Haynes, J.-D. (2015). "A primer on pattern-based approaches to fMRI: principles, pitfalls, and perspectives". In: *Neuron* 87.2, pp. 257–270.

Heer, W. A. de, A. G. Huth, T. L. Griffiths, J. L. Gallant, and F. E. Theunissen (2017). "The hierarchical cortical organization of human speech processing." In: *Journal of Neuroscience*, pp. 3267–16.

Henry, M. J., B. Herrmann, D. Kunke, and J. Obleser (2017). "Aging affects the balance of neural entrainment and top-down neural modulation in the listening brain". In: *Nature communications* 8, p. 15801.

Hillyard, S. A., R. F. Hink, V. L. Schwent, and T. W. Picton (1973). "Electrical signs of selective attention in the human brain". In: *Science* 182.4108, pp. 177–180.

Hoerl, A. E. and R. W. Kennard (1970). "Ridge regression: Biased estimation for nonorthogonal problems". In: *Technometrics* 12.1, pp. 55–67.

Holdgraf, C. et al. (2016). "Rapid tuning shifts in human auditory cortex enhance speech intelligibility". In: *Nat. Commun.* 7, p. 13654.

Holdgraf, C. R., J. W. Rieger, C. Micheli, S. Martin, R. T. Knight, and F. E. Theunissen (2017). "Encoding and decoding models in cognitive electrophysiology". In: *Frontiers in systems neuroscience* 11, p. 61.

Horton, C., M. D'Zmura, and R. Srinivasan (2013). "Suppression of competing speech through entrainment of cortical oscillations". In: *Journal of neurophysiology* 109.12, pp. 3082–3093.

Howard, M. F. and D. Poeppel (2010). "Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension". In: *Journal of neurophysiology* 104.5, pp. 2500–2511.

Hullett, P. W., L. S. Hamilton, N. Mesgarani, C. E. Schreiner, and E. F. Chang (2016). "Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli". In: *Journal of Neuroscience* 36.6, pp. 2014–2026.

Humes, L. E. et al. (2012). "Central presbycusis: a review and evaluation of the evidence". In: *Journal of the American Academy of Audiology* 23.8, pp. 635–666.

Huth, A. G., T. Lee, S. Nishimoto, N. Y. Bilenko, A. T. Vu, and J. L. Gallant (2016). "Decoding the semantic content of natural movies from human brain activity". In: *Frontiers in systems neuroscience* 10, p. 81.

Jenkins, K. A., C. Fodor, A. Presacco, and S. Anderson (2018). "Effects of amplification on neural phase locking, amplitude, and latency to a speech syllable". In: *Ear and hearing* 39.4, pp. 810–824.

Jensen, O. and A. Mazaheri (2010). "Shaping functional architecture by oscillatory alpha activity: gating by inhibition". In: *Frontiers in human neuroscience* 4.

Jensen, O. and C. D. Tesche (2002). "Frontal theta activity in humans increases with memory load in a working memory task". In: *European journal of Neuroscience* 15.8, pp. 1395–1399.

Jensen, O., J. Gelfand, J. Kounios, and J. E. Lisman (2002). "Oscillations in the alpha band (9–12 Hz) increase with memory load during retention in a short-term memory task". In: *Cerebral cortex* 12.8, pp. 877–882.

Jørgensen, S., S. D. Ewert, and T. Dau (2013). "A multi-resolution envelope-power based model for speech intelligibility". In: *The Journal of the Acoustical Society of America* 134.1, pp. 436–446.

Kale, S. and M. G. Heinz (2010). "Envelope coding in auditory nerve fibers following noise-induced hearing loss". In: *Journal of the Association for Research in Otolaryngology* 11.4, pp. 657–673.

Kale, S. and M. G. Heinz (2012). "Temporal modulation transfer functions measured from auditory-nerve responses following sensorineural hearing loss". In: *Hearing research* 286.1-2, pp. 64–75.

Kay, K. N., J. Winawer, A. Rokem, A. Mezer, and B. A. Wandell (2013). "A two-stage cascade model of BOLD responses in human visual cortex". In: *PLoS computational biology* 9.5, e1003079.

Kaya, E. M. and M. Elhilali (2017). "Modelling auditory attention". In: *Phil. Trans. R. Soc. B* 372.1714, p. 20160101.

Keitel, A., R. A. Ince, J. Gross, and C. Kayser (2017). "Auditory cortical delta-entrainment interacts with oscillatory power in multiple fronto-parietal networks". In: *NeuroImage* 147, pp. 32–42.

Kell, A. J., D. L. Yamins, E. N. Shook, S. V. Norman-Haignere, and J. H. McDermott (2018). "A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy". In: *Neuron* 98.3, pp. 630–644.

Kerlin, J. R., A. J. Shahin, and L. M. Miller (2010). "Attentional gain control of ongoing cortical speech representations in a "cocktail party"". In: *Journal of Neuroscience* 30.2, pp. 620–628.

Khalighinejad, B., G. Cruzatto da Silva, and N. Mesgarani (2017). "Dynamic Encoding of Acoustic Features in Neural Responses to Continuous Speech". In: *J. Neurosci.* 37.8, pp. 2176–2185.

Klimesch, W. (1999). "EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis". In: *Brain research reviews* 29.2, pp. 169–195.

Koelewijn, T., A. A. Zekveld, J. M. Festen, and S. E. Kramer (2012). "Pupil dilation uncovers extra listening effort in the presence of a single-talker masker". In: *Ear and Hearing* 33.2, pp. 291–300.

Kong, Y.-Y., A. Mullangi, and N. Ding (2014). “Differential modulation of auditory responses to attended and unattended speech in different listening conditions”. In: *Hearing research* 316, pp. 73–81.

Kong, Y.-Y., A. Somarowthu, and N. Ding (2015). “Effects of spectral degradation on attentional modulation of cortical auditory responses to continuous speech”. In: *Journal of the Association for Research in Otolaryngology* 16.6, pp. 783–796.

Koskinen, M., J. Viinikanoja, M. Kurimo, A. Klami, S. Kaski, and R. Hari (2013). “Identifying fragments of natural speech from the listener’s MEG signals”. In: *Human brain mapping* 34.6, pp. 1477–1489.

Kothe, C. (2013). “Lab streaming layer (LSL).
Available online at: https://github.com/sccn/labstreaminglayer”. In:

Krause, C. M., A. H. Lang, M. Laine, M. Kuusisto, and B. Pörn (1996). “Event-related. EEG desynchronization and synchronization during an auditory memory task”. In: *Electroencephalography and clinical neurophysiology* 98.4, pp. 319–326.

Kriegeskorte, N. (2011). “Pattern-information analysis: from stimulus decoding to computational-model testing”. In: *Neuroimage* 56.2, pp. 411–421.

Krull, V., L. E. Humes, and G. R. Kidd (2013). “Reconstructing wholes from parts: effects of modality, age, and hearing loss on word recognition.” In: *Ear and hearing* 34.2, e14–23.

Lakatos, P., A. S. Shah, K. H. Knuth, I. Ulbert, G. Karmos, and C. E. Schroeder (2005). “An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex”. In: *Journal of neurophysiology* 94.3, pp. 1904–1911.

Lakatos, P., G. Karmos, A. D. Mehta, I. Ulbert, and C. E. Schroeder (2008). “Entrainment of neuronal oscillations as a mechanism of attentional selection”. In: *science* 320.5872, pp. 110–113.

Lalor, E. C., A. J. Power, R. B. Reilly, and J. J. Foxe (2009a). “Resolving precise temporal processing properties of the auditory system using continuous stimuli”. In: *J. Neurophysiol.* 102.1, pp. 349–59.

Lalor, E., B. Pearlmutter, R. Reilly, G. McDarby, and J. Foxe (2006). “The VESPA: a method for the rapid estimation of a visual evoked potential”. In: *Neuroimage* 32.4, pp. 1549–1561.

Lalor, E. C. and J. J. Foxe (2010). “Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution”. In: *European journal of neuroscience* 31.1, pp. 189–193.

Lalor, E. C., A. J. Power, R. B. Reilly, and J. J. Foxe (2009b). “Resolving precise temporal processing properties of the auditory system using continuous stimuli”. In: *Journal of neurophysiology* 102.1, pp. 349–359.

Larsby, B. and S. Arlinger (1999). “Auditory temporal and spectral resolution in normal and impaired hearing”. In: *Journal of the American Academy of Audiology* 10.4, pp. 198–210.

Larsen, K. M. et al. (2018). “Altered auditory processing and effective connectivity in 22q11. 2 deletion syndrome”. In: *Schizophrenia research*.

Lavie, N., A. Hirst, J. W. De Fockert, and E. Viding (2004). “Load theory of selective attention and cognitive control.” In: *Journal of Experimental Psychology: General* 133.3, p. 339.

Leiberg, S., W. Lutzenberger, and J. Kaiser (2006). “Effects of memory load on cortical oscillatory activity during auditory pattern working memory”. In: *Brain research* 1120.1, pp. 131–140.

Lescroart, M. D., D. E. Stansbury, and J. L. Gallant (2015). “Fourier power, subjective distance, and object categories all provide plausible models of BOLD responses in scene-selective visual areas”. In: *Frontiers in computational neuroscience* 9, p. 135.

Lisman, J. E. and O. Jensen (2013). “The theta-gamma neural code”. In: *Neuron* 77.6, pp. 1002–1016.

Lister, J. J., N. D. Maxfield, G. J. Pitt, and V. B. Gonzalez (2011). “Auditory evoked response to gaps in noise: older adults”. In: *International Journal of Audiology* 50.4, pp. 211–225.

Luo, H. and D. Poeppel (2007). “Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex”. In: *Neuron* 54.6, pp. 1001–1010.

Machens, C., M. Wehr, and A. Zador (2004). “Linearity of cortical receptive fields measured with natural sounds”. In: *J. Neurosci.* 24.5, pp. 1089–1100.

Machens, C. K., M. Wehr, and A. M. Zador (2003). “Spectro-temporal receptive fields of subthreshold responses in auditory cortex”. In: *Advances in neural information processing systems*, pp. 149–156.

Makeig, S., A. J. Bell, T.-P. Jung, and T. J. Sejnowski (1996). "Independent component analysis of electroencephalographic data". In: *Advances in neural information processing systems*, pp. 145–151.

Malham, D. G. and A. Myatt (1995). "3-D sound spatialization using ambisonic techniques". In: *Computer music journal* 19.4, pp. 58–70.

Marconato, A., L. Ljung, Y. Rolain, and J. Schoukens (2014). "Linking regularization and low-rank approximation for impulse response modeling". In: *IFAC Proc. Vol.* 47.3, pp. 4999–5004.

Marrone, N., C. R. Mason, and G. Kidd Jr (2008). "The effects of hearing loss and age on the benefit of spatial separation between multiple talkers in reverberant rooms". In: *The Journal of the Acoustical Society of America* 124.5, pp. 3064–3075.

Martin, S. et al. (2014). "Decoding spectrotemporal features of overt and covert speech from the human cortex". In: *Frontiers in neuroengineering* 7.

McEvoy, L. K., M. E. Smith, and A Gevins (1998). "Dynamic cortical networks of verbal and spatial working memory: effects of memory load and task practice." In: *Cerebral cortex (New York, NY: 1991)* 8.7, pp. 563–574.

Mesgarani, N. and E. F. Chang (2012). "Selective cortical representation of attended speaker in multi-talker speech perception". In: *Nature* 485.7397, pp. 233–236.

Mesgarani, N., S. V. David, J. B. Fritz, and S. A. Shamma (2009). "Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex". In: *Journal of neurophysiology* 102.6, pp. 3329–3339.

Mesgarani, N., J. Fritz, and S. Shamma (2010). "A computational model of rapid task-related plasticity of auditory cortical receptive fields". In: *Journal of computational neuroscience* 28.1, pp. 19–27.

Mesgarani, N., S. V. David, J. B. Fritz, and S. A. Shamma (2014a). "Mechanisms of noise robust representation of speech in primary auditory cortex". In: *Proceedings of the National Academy of Sciences* 111.18, pp. 6792–6797.

Mesgarani, N., C. Cheung, K. Johnson, and E. F. Chang (2014b). "Phonetic feature encoding in human superior temporal gyrus". In: *Science* 343.6174, pp. 1006–1010.

Millman, R. E., S. R. Johnson, and G. Prendergast (2015). "The role of phase-locking to the temporal envelope of speech in auditory perception and speech intelligibility". In: *Journal of cognitive neuroscience*.

Millman, R. E., S. L. Mattys, A. D. Gouws, and G. Prendergast (2017). "Magnified neural envelope coding predicts deficits in speech perception in noise". In: *Journal of Neuroscience*, pp. 2722–16.

Miran, S., S. Akram, A. Sheikhattar, J. Z. Simon, T. Zhang, and B. Babadi (2018). "Real-time tracking of selective auditory attention from M/EEG: A bayesian filtering approach". In: *Frontiers in neuroscience* 12.

Mirkovic, B., M. Bleichner, M. De Vos, and S. Debener (2016). "Target speaker detection with concealed EEG around the ear". In: *Front. Neurosci.* 10, p. 349.

Mirkovic, B., S. Debener, M. Jaeger, and M. De Vos (2015). "Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications". In: *Journal of neural engineering* 12.4, p. 046007.

Miyake, A., N. P. Friedman, M. J. Emerson, A. H. Witzki, A. Howerter, and T. D. Wager (2000). "The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis". In: *Cognitive psychology* 41.1, pp. 49–100.

Moore, B. C. (1986). *Frequency selectivity in hearing*. Academic Press.

Moore, B. C. and B. R. Glasberg (1983). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns". In: *The Journal of the Acoustical Society of America* 74.3, pp. 750–753.

Moore, B. C. and B. R. Glasberg (1993). "Simulation of the effects of loudness recruitment and threshold elevation on the intelligibility of speech in quiet and in a background of speech". In: *The Journal of the Acoustical Society of America* 94.4, pp. 2050–2062.

Moore, B. C. and B. R. Glasberg (1998). "Use of a loudness model for hearing-aid fitting. I. Linear hearing aids". In: *British journal of audiology* 32.5, pp. 317–335.

Moore, B. C., M. Wojtczak, and D. A. Vickers (1996). "Effect of loudness recruitment on the perception of amplitude modulation". In: *The Journal of the Acoustical Society of America* 100.1, pp. 481–489.

Moore, R. C., T. Lee, and F. E. Theunissen (2013). "Noise-invariant neurons in the avian auditory cortex: hearing the song in noise". In: *PLoS computational biology* 9.3, e1002942.

Näätänen, R. (2018). *Attention and brain function*. Routledge.

Naselaris, T., R. J. Prenger, K. N. Kay, M. Oliver, and J. L. Gallant (2009). "Bayesian reconstruction of natural images from human brain activity". In: *Neuron* 63.6, pp. 902–915.

Naselaris, T., K. N. Kay, S. Nishimoto, and J. L. Gallant (2011). "Encoding and decoding in fMRI". In: *Neuroimage* 56.2, pp. 400–410.

Neher, T. et al. (2009). "Benefit from spatial separation of multiple talkers in bilateral hearing-aid users: Effects of hearing loss, age, and cognition". In: *International Journal of Audiology* 48.11, pp. 758–774.

Nielsen, J. B. and T. Dau (2009). "Development of a Danish speech intelligibility test". In: *International journal of audiology* 48.10, pp. 729–741.

Noirhomme, Q. et al. (2014). "Biased binomial assessment of cross-validated estimation of classification accuracies illustrated in diagnosis predictions". In: *NeuroImage: Clinical* 4, pp. 687–694.

Nykopp, T. (2001). *Statistical modelling issues for the adaptive brain interface*. Master's thesis.

O'Sullivan, A., M. Crosse, G. Di Liberto, and E. Lalor (2017a). "Visual Cortical Entrainment to Motion and Categorical Speech Features during Silent Lipreading". In: *Front. Hum. Neurosci.* 10, p. 679.

O'Sullivan, J. et al. (2017b). "Neural decoding of attentional selection in multi-speaker environments without access to clean sources". In: *J. Neural Eng.* 14.5, p. 056001.

O'Sullivan, J. A. et al. (2014). "Attentional selection in a cocktail party environment can be decoded from single-trial EEG". In: *Cerebral Cortex* 25.7, pp. 1697–1706.

Obleser, J., M. Wöstmann, N. Hellbernd, A. Wilsch, and B. Maess (2012). "Adverse listening conditions and memory load drive a common alpha oscillatory network". In: *Journal of Neuroscience* 32.36, pp. 12376–12383.

Oostenveld, R., P. Fries, E. Maris, and J.-M. Schoffelen (2011). "FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data". In: *Computational intelligence and neuroscience* 2011, p. 1.

Oreinos, C. and J. M. Buchholz (2013). "Measurement of a full 3D set of HRTFs for in-ear and hearing aid microphones on a head and torso simulator (HATS)". In: *Acta Acustica united with Acustica* 99.5, pp. 836–844.

Owen, A. M., K. M. McMillan, A. R. Laird, and E. Bullmore (2005). "N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies". In: *Human brain mapping* 25.1, pp. 46–59.

Park, H., R. A. Ince, P. G. Schyns, G. Thut, and J. Gross (2015). "Frontal top-down signals increase coupling of auditory low-frequency oscillations to

continuous speech in human listeners". In: *Current Biology* 25.12, pp. 1649–1653.

Parra, L. C., C. D. Spence, A. D. Gerson, and P. Sajda (2005). "Recipes for the linear analysis of EEG". In: *Neuroimage* 28.2, pp. 326–341.

Parthasarathy, A., B. Herrmann, and E. L. Bartlett (2019). "Aging alters envelope representations of speech-like sounds in the inferior colliculus". In: *Neurobiology of aging* 73, pp. 30–40.

Pasley, B. N. et al. (2012). "Reconstructing speech from human auditory cortex". In: *PLoS biology* 10.1, e1001251.

Patterson, R., I. Nimmo-Smith, J. Holdsworth, and P. Rice (1987). "An efficient auditory filterbank based on the gammatone function". In: *a meeting of the IOC Speech Group on Auditory Modelling at RSRE*. Vol. 2. 7.

Peelle, J. E. and M. H. Davis (2012). "Neural oscillations carry speech rhythm through to comprehension". In: *Frontiers in psychology* 3.

Peelle, J. E. and A. Wingfield (2016). "The neural consequences of age-related hearing loss". In: *Trends in neurosciences* 39.7, pp. 486–497.

Peelle, J. E., V. Troiani, A. Wingfield, and M. Grossman (2009). "Neural processing during older adults? comprehension of spoken sentences: age differences in resource allocation and connectivity". In: *Cerebral Cortex* 20.4, pp. 773–782.

Peelle, J. E., V. Troiani, M. Grossman, and A. Wingfield (2011). "Hearing loss in older adults affects neural systems supporting speech comprehension". In: *Journal of Neuroscience* 31.35, pp. 12638–12643.

Peelle, J. E., J. Gross, and M. H. Davis (2012). "Phase-locked responses to speech in human auditory cortex are enhanced during comprehension". In: *Cerebral cortex* 23.6, pp. 1378–1387.

Pesonen, M., H. Hämäläinen, and C. M. Krause (2007). "Brain oscillatory 4–30 Hz responses during a visual n-back memory task with varying memory load". In: *Brain research* 1138, pp. 171–177.

Petersen, E. B., M. Wöstmann, J. Obleser, and T. Lunner (2016). "Neural tracking of attended versus ignored speech is differentially affected by hearing loss". In: *Journal of neurophysiology* 117.1, pp. 18–27.

Phipson, B. and G. K. Smyth (2010). "Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn". In: *Statistical applications in genetics and molecular biology* 9.1.

Pichora-Fuller, M. K. (2003). "Cognitive aging and auditory information processing". In: *International journal of audiology* 42.sup2, pp. 26–32.

Pichora-Fuller, M. K., B. A. Schneider, and M. Daneman (1995). "How young and old adults listen to and remember speech in noise". In: *The Journal of the Acoustical Society of America* 97.1, pp. 593–608.

Plack, C. J., A. J. Oxenham, A. M. Simonson, C. G. O'Hanlon, V. Drga, and D. Arifianto (2008). "Estimates of compression at low and high frequencies using masking additivity in normal and impaired ears". In: *The Journal of the Acoustical Society of America* 123.6, pp. 4321–4330.

Power, A., R. Reilly, and E. Lalor (2011). "Comparing linear and quadratic models of the human auditory system using EEG". In: *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*. IEEE, pp. 4171–4174.

Power, A. J., J. J. Foxe, E.-J. Forde, R. B. Reilly, and E. C. Lalor (2012). "At what time is the cocktail party? A late locus of selective attention to natural speech". In: *European Journal of Neuroscience* 35.9, pp. 1497–1503.

Pratt, N., A. Willoughby, and D. Swick (2011). "Effects of working memory load on visual selective attention: behavioral and electrophysiological evidence". In: *Frontiers in human neuroscience* 5.

Presacco, A., J. Z. Simon, and S. Anderson (2016a). "Effect of informational content of noise on speech representation in the aging midbrain and cortex". In: *Journal of neurophysiology* 116.5, pp. 2356–2367.

Presacco, A., J. Z. Simon, and S. Anderson (2016b). "Evidence of degraded representation of speech in noise, in the aging midbrain and cortex". In: *Journal of neurophysiology* 116.5, pp. 2346–2355.

Puvvada, K. and J. Simon (2017). "Cortical representations of speech in a multitalker auditory scene". In: *J. Neurosci.* 37.38, pp. 9189–9196.

Qian, J., T. Hastie, J. Friedman, R. Tibshirani, and N. Simon (2013). *Glmnet for Matlab*.

Rabinowitz, N. C., B. D. Willmore, A. J. King, and J. W. Schnupp (2013). "Constructing noise-invariant representations of sound in the auditory pathway". In: *PLoS biology* 11.11, e1001710.

Raghavachari, S. et al. (2001). "Gating of human theta oscillations by a working memory task". In: *Journal of Neuroscience* 21.9, pp. 3175–3183.

Renard, Y. et al. (2010). "Openvibe: An open-source software platform to design, test, and use brain–computer interfaces in real and virtual environments". In: *Presence: teleoperators and virtual environments* 19.1, pp. 35–53.

Rieke, F, D. Bodnar, and W Bialek (1995). “Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents”. In: *Proceedings of the Royal Society of London B: Biological Sciences* 262.1365, pp. 259–265.

Rimmele, J. M., E. Z. Golumbic, E. Schröger, and D. Poeppel (2015). “The effects of selective attention and speech acoustics on neural speech-tracking in a multi-talker scene”. In: *Cortex* 68, pp. 144–154.

Ringach, D. and R. Shapley (2004). “Reverse correlation in neurophysiology”. In: *Cognitive Science* 28.2, pp. 147–166.

Rønne, F. M., T. Dau, J. Harte, and C. Elberling (2012). “Modeling auditory evoked brainstem responses to transient stimuli”. In: *The Journal of the Acoustical Society of America* 131.5, pp. 3903–3913.

Ross, B, T. Picton, A. Herdman, and C Pantev (2004). “The effect of attention on the auditory steady-state response.” In: *Neurology & clinical neurophysiology: NCN* 2004, pp. 22–22.

Rousselet, G. A. (2012). “Does filtering preclude us from studying ERP time-courses?” In: *Frontiers in psychology* 3, p. 131.

Roux, F. and P. J. Uhlhaas (2014). “Working memory and neural oscillations: alpha–gamma versus theta–gamma codes for distinct WM information?” In: *Trends in cognitive sciences* 18.1, pp. 16–25.

Ruggero, M. A. (1992). “Responses to sound of the basilar membrane of the mammalian cochlea”. In: *Current opinion in neurobiology* 2.4, pp. 449–456.

Ruggles, D. and B. Shinn-Cunningham (2011). “Spatial selective auditory attention in the presence of reverberant energy: individual differences in normal-hearing listeners”. In: *Journal of the Association for Research in Otolaryngology* 12.3, pp. 395–405.

SanMiguel, I., M.-J. Corral, and C. Escera (2008). “When loading working memory reduces distraction: behavioral and electrophysiological evidence from an auditory-visual distraction paradigm”. In: *Journal of Cognitive Neuroscience* 20.7, pp. 1131–1145.

Santoro, R. et al. (2014). “Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex”. In: *PLoS computational biology* 10.1, e1003412.

Santoro, R. et al. (2017). “Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns”. In: *Proceedings of the National Academy of Sciences* 114.18, pp. 4799–4804.

Scharinger, C., A. Soutschek, T. Schubert, and P. Gerjets (2015). "When flanker meets the n-back: What EEG and pupil dilation data reveal about the interplay between the two central-executive working memory functions inhibition and updating". In: *Psychophysiology* 52.10, pp. 1293–1304.

Scharinger, C., A. Soutschek, T. Schubert, and P. Gerjets (2017). "Comparison of the working memory load in n-back and working memory span tasks by means of EEG frequency band power and P300 amplitude". In: *Frontiers in human neuroscience* 11.

Schoppe, O., N. Harper, B. Willmore, A. King, and J. Schnupp (2016). "Measuring the performance of neural models". In: *Front. Comput. Neurosci.* 10.

Schroeder, C. E. and P. Lakatos (2009). "Low-frequency neuronal oscillations as instruments of sensory selection". In: *Trends in neurosciences* 32.1, pp. 9–18.

Sergeyenko, Y., K. Lall, M. C. Liberman, and S. G. Kujawa (2013). "Age-related cochlear synaptopathy: an early-onset contributor to auditory functional decline". In: *Journal of Neuroscience* 33.34, pp. 13686–13694.

Shinn-Cunningham, B. G. and V. Best (2008). "Selective attention in normal and impaired hearing". In: *Trends in amplification* 12.4, pp. 283–299.

Shinn-Cunningham, B. (2017). "Cortical and sensory causes of individual differences in selective attention ability among listeners with normal hearing thresholds". In: *Journal of Speech, Language, and Hearing Research* 60.10, pp. 2976–2988.

Slama, M. C. and B. Delgutte (2015). "Neural coding of sound envelope in reverberant environments". In: *Journal of Neuroscience* 35.10, pp. 4452–4468.

Snyder, A. C. and J. J. Foxe (2010). "Anticipatory attentional suppression of visual features indexed by oscillatory alpha-band power increases: a high-density electrical mapping study". In: *Journal of Neuroscience* 30.11, pp. 4024–4032.

Sörös, P., I. K. Teismann, E. Manemann, and B. Lütkenhöner (2009). "Auditory temporal processing in healthy aging: a magnetoencephalographic study". In: *BMC neuroscience* 10.1, p. 34.

Sörqvist, P., S. Stenfelt, and J. Rönnberg (2012). "Working memory capacity and visual–verbal cognitive load modulate auditory–sensory gating in the brainstem: Toward a unified view of attention". In: *Journal of cognitive neuroscience* 24.11, pp. 2147–2154.

Steinschneider, M., K. V. Nourski, and Y. I. Fishman (2013). "Representation of speech in human auditory cortex: is it special?" In: *Hearing research* 305, pp. 57–73.

Theunissen, F., K. Sen, and A. Doupe (2000). “Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds”. In: *J. Neurosci.* 20.6, pp. 2315–2331.

Theunissen, F., S. David, N. Singh, A. Hsu, W. Vinje, and J. Gallant (2001). “Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli”. In: *Netw. Comput. Neural Syst.* 12.3, pp. 289–316.

Theunissen, F. E. and J. E. Elie (2014). “Neural processing of natural sounds”. In: *Nature Reviews Neuroscience* 15.6, pp. 355–366.

Thorson, I., J. Liénard, and S. David (2015). “The essential complexity of auditory receptive fields”. In: *PLOS Comput. Biol.* 11.12, e1004628.

Tibshirani, R. (1996). “Regression shrinkage and selection via the lasso”. In: *J. Royal Statist. Soc. B* 58.1, pp. 267–288.

Tikhonov, A. N. (1963). “Solution of incorrectly formulated problems and the regularization method”. In: *Soviet Math. Dokl.* 4, pp. 1035–1038.

Traer, J. and J. H. McDermott (2016). “Statistics of natural reverberation enable perceptual separation of sound and space”. In: *Proceedings of the National Academy of Sciences* 113.48, E7856–E7865.

Tremblay, K. L., M. Piskosz, and P. Souza (2003). “Effects of age and age-related hearing loss on the neural representation of speech cues”. In: *Clinical Neurophysiology* 114.7, pp. 1332–1343.

Vajargah, K. (2013). “Comparing ridge regression and principal components regression by Monte Carlo simulation basedon MSE”. In: *Journal of Computer Science and Computational Mathematics* 3.2, pp. 25–29.

Van Eyndhoven, S., T. Francart, and A. Bertrand (2017). “EEG-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses”. In: *IEEE Trans. Biomed. Eng.* 64.5, pp. 1045–1056.

Van Gerven, P. W., F. Paas, J. J. Van Merriënboer, and H. G. Schmidt (2004). “Memory load and the cognitive pupillary response in aging”. In: *Psychophysiology* 41.2, pp. 167–174.

VanRullen, R. (2011). “Four common conceptual fallacies in mapping the time course of recognition”. In: *Frontiers in psychology* 2, p. 365.

Vander Ghinst, M. et al. (2016). “Left superior temporal gyrus is coupled to attended speech in a cocktail-party auditory scene”. In: *Journal of Neuroscience* 36.5, pp. 1596–1606.

Vandierendonck, A. (2014). “Symbiosis of executive and selective attention in working memory”. In: *Frontiers in human neuroscience* 8.

Walton, J. P., H. Simon, and R. D. Frisina (2002). “Age-related alterations in the neural coding of envelope periodicities”. In: *Journal of neurophysiology* 88.2, pp. 565–578.

Watter, S., G. M. Geffen, and L. B. Geffen (2001). “The n-back as a dual-task: P300 morphology under divided attention”. In: *Psychophysiology* 38.6, pp. 998–1003.

Weichwald, S., T. Meyer, O. Özdenizci, B. Schölkopf, T. Ball, and M. Grosse-Wentrup (2015). “Causal interpretation rules for encoding and decoding models in neuroimaging”. In: *NeuroImage* 110, pp. 48–59.

Wendt, D., T. Dau, and J. Hjortkjær (2016). “Impact of background noise and sentence complexity on processing demands during sentence comprehension”. In: *Frontiers in psychology* 7.

Whiting, K. A., B. A. Martin, and D. R. Stapells (1998). “The effects of broadband noise masking on cortical event-related potentials to speech sounds/ba/and/da”. In: *Ear and hearing* 19.3, pp. 218–231.

Widmann, A., E. Schröger, and B. Maess (2015). “Digital filter design for electrophysiological data–a practical approach”. In: *Journal of neuroscience methods* 250, pp. 34–46.

Wiinberg, A., J. Zaar, and T. Dau (2018). “Effects of Expanding Envelope Fluctuations on Consonant Perception in Hearing-Impaired Listeners”. In: *Trends in hearing* 22, p. 2331216518775293.

Willmore, B., O. Schoppe, A. King, J. Schnupp, and N. Harper (2016). “Incorporating midbrain adaptation to mean sound level improves models of auditory cortical processing”. In: *J. Neurosci.* 36.2, pp. 280–289.

Wingfield, A. and J. E. Peelle (2012). “How does hearing loss affect the brain?” In: *Aging health* 8.2, pp. 107–109.

Winkler, I., S. Debener, K.-R. Müller, and M. Tangermann (2015). “On the influence of high-pass filtering on ICA-based artifact reduction in EEG-ERP”. In: *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE, pp. 4101–4105.

Wintink, A. J., S. J. Segalowitz, and L. J. Cudmore (2001). “Task complexity and habituation effects on frontal P300 topography”. In: *Brain and Cognition* 46.1, pp. 307–311.

Wit, L. de, D. Alexander, V. Ekroll, and J. Wagemans (2016). "Is neuroimaging measuring information in the brain?" In: *Psychonomic bulletin and review* 23.5, pp. 1415–1428.

Woldorff, M. G. et al. (1993). "Modulation of early sensory processing in human auditory cortex during auditory selective attention". In: *Proceedings of the National Academy of Sciences* 90.18, pp. 8722–8726.

Wolpaw, J. and H. Ramoser (1998). "EEG-based communication: improved accuracy by response verification". In: *IEEE Trans. Rehabil. Eng.* 6.3, pp. 326–33.

Wong, D. D., S. A. Fuglsang, J. Hjortkjær, E. Ceolini, M. Slaney, and A. de Cheveigné (2018). "A comparison of regularization methods in forward and backward models for auditory attention decoding". In: *Frontiers in neuroscience* 12, p. 531.

Woolley, S. M., T. E. Fremouw, A. Hsu, and F. E. Theunissen (2005). "Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds". In: *Nature neuroscience* 8.10, p. 1371.

Wöstmann, M., B. Herrmann, B. Maess, and J. Obleser (2016). "Spatiotemporal dynamics of auditory attention synchronize with speech". In: *Proceedings of the National Academy of Sciences* 113.14, pp. 3873–3878.

Wöstmann, M., S.-J. Lim, and J. Obleser (2017). "The Human Neural Alpha Response to Speech is a Proxy of Attentional Control". In: *Cerebral Cortex* 27.6, pp. 3307–3317.

Wu, M., S. David, and J. Gallant (2006). "Complete functional characterization of sensory neurons by system identification". In: *Annu. Rev. Neurosci.* 29, pp. 477–505.

Yamins, D. L. and J. J. DiCarlo (2016). "Using goal-driven deep learning models to understand sensory cortex". In: *Nature neuroscience* 19.3, p. 356.

Zekveld, A. A., S. E. Kramer, and J. M. Festen (2010). "Pupil response as an indication of effortful listening: The influence of sentence intelligibility". In: *Ear and hearing* 31.4, pp. 480–490.

Zhong, Z., K. S. Henry, and M. G. Heinz (2014). "Sensorineural hearing loss amplifies neural coding of envelope information in the central auditory system of chinchillas". In: *Hearing research* 309, pp. 55–62.

Zink, R., S. Proesmans, A. Bertrand, S. Van Huffel, and M. De Vos (2017). "Online detection of auditory attention with mobile EEG: closing the loop with neurofeedback". In: *BioRxiv*.

Zoefel, B. and R. VanRullen (2015). “The role of high-level processes for oscillatory phase entrainment to speech sound”. In: *Frontiers in human neuroscience* 9.

Zoefel, B. and R. VanRullen (2016). “EEG oscillations entrain their phase to high-level features of speech sound”. In: *Neuroimage* 124, pp. 16–23.

Zou, H. and T. Hastie (2005). “Regularization and variable selection via the elastic net”. In: *J. R. Stat. Soc. Series B Stat. Methodol.* 67.2, pp. 301–320.

de Cheveigné, A. and D. Arzounian (2017). “Robust detrending, rereferencing, outlier detection, and inpainting for multichannel data”. In: *bioRxiv*, p. 232892.

de Cheveigné, A., D. Wong, G. Di Liberto, J. Hjortkjær, M. Slaney, and E. Lalor (2018a). “Decoding the auditory brain with canonical component analysis”. In: *Neuroimage* 172, pp. 206–216.

de Cheveigné, A. and L. C. Parra (2014). “Joint decorrelation, a versatile tool for multichannel data analysis”. In: *Neuroimage* 98, pp. 487–505.

de Cheveigné, A. and J. Z. Simon (2008). “Denoising based on spatial filtering”. In: *Journal of neuroscience methods* 171.2, pp. 331–339.

de Cheveigné, A. et al. (2018b). “Multiway Canonical Correlation Analysis of Brain Signals”. In: *bioRxiv*, p. 344960.

A

Challenges in the interpretation of results from encoding and decoding analyses

Encoding and decoding models are important tools that can provide insights into how information about sounds is represented in M/EEG activity. However, encoding and decoding models each have their strengths and their weaknesses, and caution should be made when interpreting results from encoding- and decoding analyses. The following sections provide examples that illustrate caveats entailed by encoding and decoding analyses.

A.1 Example 1: Correlations among stimulus features may affect interpretation of decoding weights

It was demonstrated in Chapter 2 section 2.3 that a linearized encoding model can account for correlations among stimulus features when predicting a neural response. The stimulus-reconstruction model in Eq. 2.2 does, however, not take correlations among stimulus features into account when reconstructing $S(t, f)$ (Holdgraf et al., 2017). Figure A.1 illustrates how this may be reflected in stimulus-response mappings in a scenario with simulated neural data. In this simulation, a neural response, $r(t)$, is simulated as a convolution of $s(t, f)$ and a pre-defined STRF plus a distractor signal $d(t)$. The simulated STRF (Fig. A.1b) is tuned to low frequencies and has an early and temporally local excitatory region (red) followed by an inhibitory region (blue). Fig. A.1c,d shows weights of encoding and decoding models that were fit to the data using Ridge regression (Chapter 2, Eq. 2.5). It can be seen that the encoding model (Fig. A.1c) approximates the simulated STRF with high accuracy. The weights of the decoding model (Fig. A.1d) are in contrast "smeared" across frequencies compared to the target STRF. Due to non-orthogonality of the speech spectrogram representation, the decoding model wrongly assigns significant weights to frequency

channels that did not drive the neural response, but were correlated with the ones that do (Holdgraf et al., 2017; Weichwald et al., 2015). This can in the worst case lead to incorrect conclusions about which spectro-temporal attributes of the stimulus feature that drive the response.

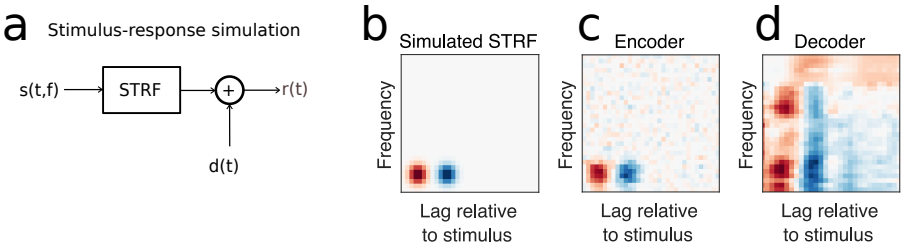

Figure A.1: Visualising spectro-temporal receptive fields estimated with encoding- and decoding models. (a) Neural data, $r(t)$, was simulated as a convolution of $s(t,f)$ and a predefined STRF (b). (c) When using a Ridge-based encoding model to estimate the STRF from $r(t)$ and $s(t,f)$ the model approximates the target STRF well. However, the STRF recovered with the decoding model (d) has artefacts and is "smeared" out compared to the target STRF, due to statistical regularities in $s(t,f)$. Figure was inspired by work in Holdgraf et al., 2017

A.2 Example 2: The influence of stimulus-unrelated activitiy on model interpretation

Encoding and decoding models can offer complementary information about statistical stimulus-response dependencies (Kriegeskorte, 2011; Weichwald et al., 2015). Since the encoding model in Chapter 2 Eq. 2.1 predicts $R(t,n)$ independently for each n, an encoding model can be useful for investigating whether a feature, $S(t,n)$, is encoded in a single-channel response. If this is the case, then the model should be able to predict the given channel response with prediction accuracies that are greater than chance level. In contrast, the decoding model in Eq. 2.2 exploits statistical regularities in R over a range of channels/sensors and time lags to reconstruct $S(t,f)$. Therefore, if a decoding model can reconstruct $S(t,f)$ based on a multi-dimensional M/EEG response, then it instead implies that information about $S(t,f)$ is represented in response-patterns *over* a set of channels/sensors (Friston, 2009).

This is illustrated below using simulated data. Consider two simulation scenarios where data from two channels, $y_1(t)$ and $y_2(t)$, are recorded. Both channels contain a noisy distractor signal, $d(t)$, that is correlated across channels. However, in the first scenario, $y_1(t)$ also contains a time-delayed target signal $s(t)$ and in the second scenario, the channels are switched such that $y_2(t)$ now instead contains a time-delayed target signal $s(t)$. Encoding models can be used to investigate whether each of the two channel responses, $y_1(t)$ and $y_2(t)$, can be predicted from time-lagged versions of $s(t)$. From figure A.2a, it can be seen that the encoding models accurately predict which of the two channels contains $s(t)$ in the two scenarios, but can only account for some fraction of variance in the data due to the noise $d(t)$.

Conversely, the decoding models reverse the problem and learn to reconstruct $s(t)$ from statistical regularities across the time-shifted versions of the two-channel response. In this case, since the distractor signal is correlated across channels, the decoder can simply regress out $d(t)$ (e.g., by assigning weights with opposite signs to each channels) and therefore perfectly reconstruct $s(t)$ in both scenarios. However, although the decoder here was more robust against the distractor $d(t)$, the predictive performance does not provide information about which of the two channels actually contains the signal. Notice that in scenarios where target signal and distractor signals are not linearly separable, the stimulus reconstruction decoder performs less well (Fig. A.3)c).

To further highlight differences between encoding and decoding models, Fig. A.3 shows model weights of three stimulus-response models that were trained on simulated neural data. The same simulation scenario as in Fig. A.2a is considered here (i.e., $y_1(t) = s(t - 125\text{ ms}) + d(t)$ and $y_2(t) = d(t)$). It can be seen in Fig. A.3 that the encoder (estimated using Ridge regression, Chapter 2, Eq. 2.5), only assigns a significant nonzero weight to channel 1 at a time-lag of 125 ms, which is consistent with the actual relation between s and y_1. The encoder weights may thus provide information about how s relates to y_1 and y_2 and how a unitary change in s maps out to the two channels. In contrast, the decoder (Fig. A.3c) assigns significant nonzero filter weights of opposite sign to channel 1 and 2 at a time-lag of 125 ms post-stimulus, even though y_2 does not contain s. The reason for this is that the distractor signal, d, is correlated across channels. The decoder can exploit this to regress out d. The CCA model (Chapter 2, Eq. 2.8) learns a spatial decoding filter (Fig. A.3e), which also assigns significant nonzero weights of opposite sign to both channels, to "filter out" the

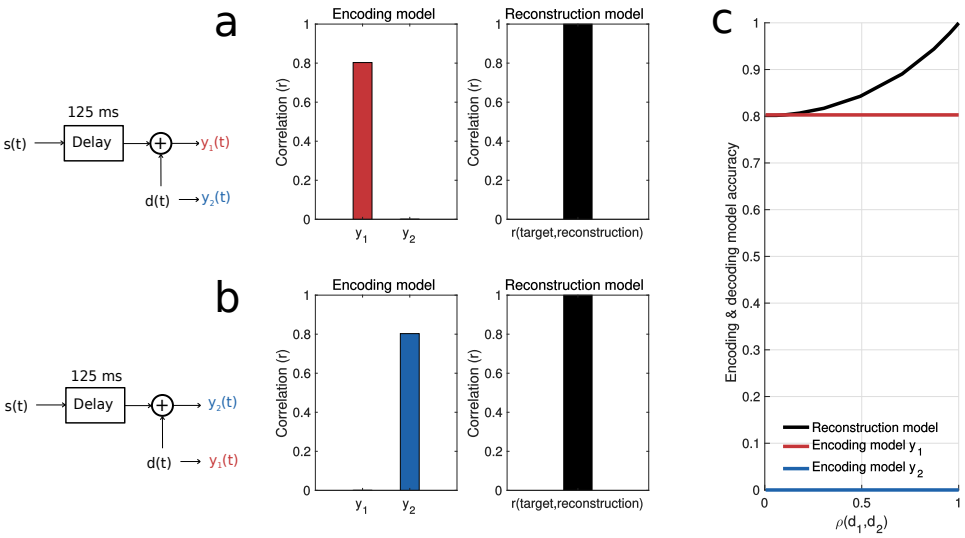

Figure A.2: Using encoding and decoding models to characterize the stimulus-response relationship between a signal, s, and two recording channels y_1 and y_2. Different simulation scenarios are considered. In simulation (a) and (b) it is assumed that only one of the two channels responds to the stimulus, $s(t)$, (after a 125 ms delay), but that both channels contain correlated noise $d(t)$. The encoding analyses accurately detects which of the two channels that responds to s in both cases. The stimulus reconstruction model perfectly manages to reconstruct s because the noise is correlated across channels. However, as formulated here, the stimulus reconstruction model does provide insights about which of the individual channels that contain s. Simulation (c) extends simulation (a), by evaluating encoding and decoding performance as a function of the between-channels correlation $\rho(d_1(t), d_2(t))$ for distractor signals, d_1 and d_2, where $y_1(t) = s(t-125ms) + d_1(t)$ and $y(t) = d_2(t)$. In this case, when d_1 and d_2 are no longer correlated, the decoding model achieve similar accuracies as the encoding model. See text for more details.

noise from the responses. In addition, since the objective function of the CCA model is sign invariant, the CCA encoding filter (Fig. A.3d) has a strong negative activation 125 ms after the stimulus onset. This simulation highlights that only the encoding model parameters can inform about how $s(t)$ relates to the each of the observed responses y_1 and y_2.

Decoding models can be transformed into corresponding encoding models which can be useful for model parameter interpretation (Dmochowski et al., 2017; Haufe et al., 2014; Parra et al., 2005). The idea is that if a decoding model transforms a neural response Y into another representation $V = Y W_b$ then a corresponding forward map can be characterized in a least-squares sense ($W_f = \left(V^T V\right)^{-1} V^T Y$) (Haufe et al., 2014). This idea can be extended to CCA model from Chapter 2, Eq. 2.8 (Dmochowski et al., 2017).

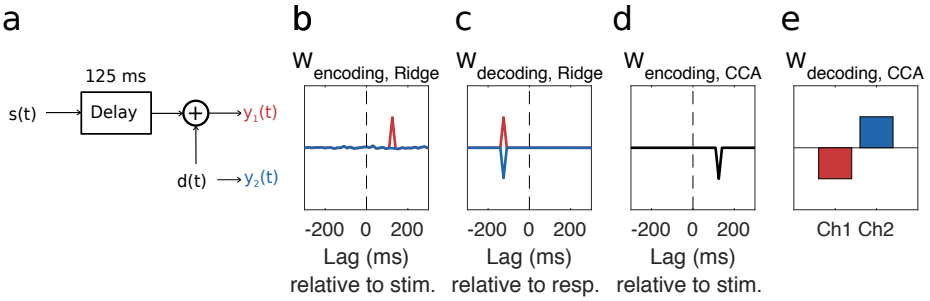

Figure A.3: An illustration of how linear encoding and decoding model weights can differ and how it can be difficult to interpret decoding model weights. This example simulates a neural data recorded from two channels, y_1 and y_2. The data in channel y_1 is assumed to respond to a stimulus signal, s, with a delay of 125 ms plus a distractor signal $y_1(t) = s(t - 125\text{ ms}) + d(t)$. It assumed that the other channel, y_2 only picks up the distractor signal, $y_2 = d$ (see a). See text for details.

.

A.3 Example 3: Regularization methods may affect interpretation of model weights

Linearized encoding models with different model constraints may achieve comparable predictive performance, but have different filter shapes (Thorson et al., 2015). This can affect the interpretation of encoding model weights, since "near-optimal" encoding models may have different "shapes" depending on what type of regularization was used to constrain the model coefficients.

Figure A.4 illustrates how different types of regularization can affect STRF model weights and result in similar test-set prediction accuracies. In this simulation, a neural response to speech, $r(t)$ is simulated as a convolution of a speech spectrogram $s(t, f)$ and a predefined STRF (Fig. A.4b, top left panel) plus additive noise. Here, the regressor is highly autocorrelated. The different regularization methods introduced in Chapter 2 are considered: Ridge regression (Chapter 2, Eq. 2.5), Lasso (Chapter 2, Eq. 2.6) and a model that constrains the model coefficients via factorization and parameterization (Chapter 2, Eq. 2.7). Fig. A.4c-e shows the recovered STRF models with the three different models. The test set prediction accuracies (indexed by Pearson's correlation coefficient between model predictions and target) are shown in the center of each panel for the different models. It can be seen that the Ridge estimator (Fig. A.4c) recovers the target STRF with good accuracy, but that many of its filter weights attain nonzero values (unlike the actual target STRF). From Fig. A.4d it can be

seen that the filter estimated with the Lasso correctly enforces more of the filter weights to zero, but that the excitatory and inhibitory regions of the recovered STRF are non-smooth unlike the target STRF. The factorized and parameterized model (Fig. A.4e) in this simulation almost perfectly recovers the STRF, since the simulated STRF had Gaussian shapes. For other more complex STRF shapes with several excitatory/inhibitory regions, the factorized and parameterized model in Chapter 2, Eq. 2.7 is expected to perform less well. From this simulation it was apparent that, despite differences between model weights, the different regularization methods yielded comparable test-set accuracies. The choice of adequate regularization methods may thus not always be obvious, but can affect interpretation of model weights.

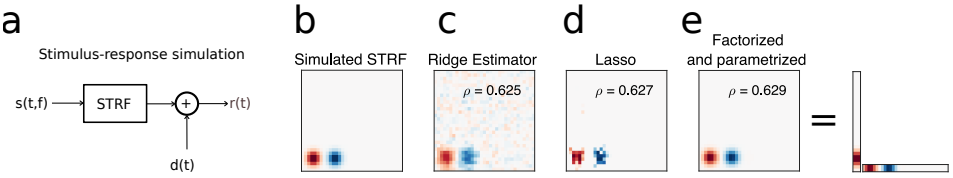

Figure A.4: An illustration of how model constraints affect encoding model weights for a simulated neural response to a speech stimulus. The neural response, $r(t)$ was simulated as a convolution of a speech spectrogram, $s(t,f)$ and a predefined STRF (b, top left panel) plus Gaussian noise $d(t)$. Different systems engineering methods are used to recover the predefined STRF from $r(t)$ and $s(t,f)$. The factorized and parametrized filters were learned using a nonlinear optimization algorithm. c-e shows the recovered STRF obtained with different methods. See text for more details.

A.4 Example 4: Interpretation of model performance and model parameters with little data

The predictive power of encoding or decoding models may depend on how much data that are available for model fitting. This may both affect interpretation of model performance, but also interpretation of model weights. This is illustrated in Fig. A.5 for an STRF model fit to actual EEG data using Ridge regression. The data used for this is example is EEG data recorded from 22 subjects listening to selectively to one of two competing speech streams. The

encoding models were here optimized to predict EEG responses from spectrograms of attended speech. From Fig. A.5, it can be seen that the ability of the encoding models to predict novel data is substantially better for encoders fit to ∼17 min of data compared to encoders fit to ∼4 min of data. Similarly, when comparing the weights of the encoders (inlets, Fig. A.5), it can be seen that the early positive activations are ”smeared out” for encoders trained on ∼4 min of data compared to encoders trained on ∼17 min of data. This underlines the importance of training data and exemplifies that the interpretation of filter weights and predictive performance may be affected by the amount of data used for model fitting.

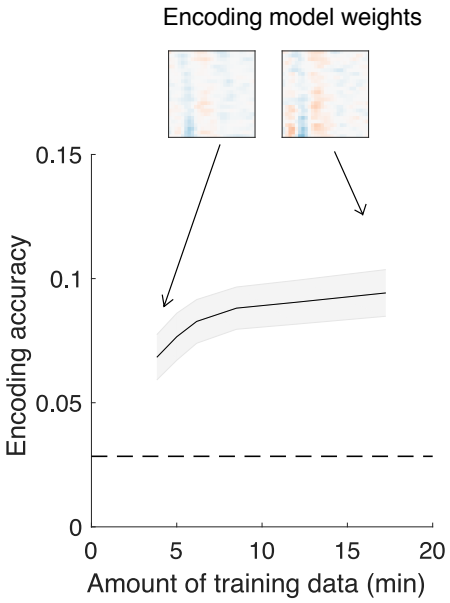

Figure A.5: An example of how amount of results from an encoding analysis can be affected by the amount of data used for model fitting. This example reflects data from 22 older normal hearing listeners listening selectively to one of two competing speech streams. The encoding models were trained and evaluated per subject and operated on cochleograms of attended speech streams. EEG responses are here decomposed into one major response component using denoising source component, (DSS; de Cheveigné and Simon, 2008) trained on ERP data. The DSS extraction filters were polarity normalized such that the N100 DSS component evoked by the repetitive tones always was negative. The encoding accuracy represents the ability of the encoders to predict a held-out test set. The Ridge parameter was tuned on a validation set. Once adequate Ridge parameters were chosen, an ”optimal” Ridge model was fit to the entire training and validation set. Errorbars represent s.e.m. across all 22 subjects. The inlets show grand average encoding model weights (peak normalized) for models fit to ∼4 min of data and ∼17 min of data. Duration of training data here also reflects data for DSS filter extraction.

A.5 Example 5: Influence of filtering on interpretation of model parameters

It is common practice to temporally filter M/EEG and ECoG data. Filter ringing can affect the interpretation of encoding model weights, as filtering alters the time course of the neural data recordings. When filtering neural recordings prior to fitting encoding models, the encoding model weights may appear to be smeared out in time. This is illustrated in Figure A.6 for simulated neural data. Two simulation scenarios are here considered. Scenario 1 simulates a response, $R_1(t)$, as convolution of a signal, $T(t)$, with a temporal response function, $\text{TRF}_{1,\text{target}}$. Similarly, scenario 2 simulates a response, $R_2(t)$, as convolution of a signal, $T(t)$, with a temporal response function, $\text{TRF}_{2,\text{target}}$. In both simulations, the responses are low-pass filtered with an FIR filter with a cut-off frequency of 15 Hz, and subsequently shifted to adjust for filter delay. The left and the right panel in Fig. A.6 show encoding filters recovered with Ridge regression in both simulation scenarios. From these panels, it can be seen that the recovered TRFs are smeared out in time, since the encoding models estimate the convolution of the "true" TRFs with the impulse response of the low-pass filters (see dashed lines in Fig. A.6; the data has here also been shifted to adjust for filter delay). However, the filter distortions are less pronounced in Scenario 1, as $\text{TRF}_{1,\text{target}}$ has a bandpass characteristic with a low-frequency passband. This simulation demonstrates that filtering may affect interpretation of both latencies and peaks of estimated encoding filters.

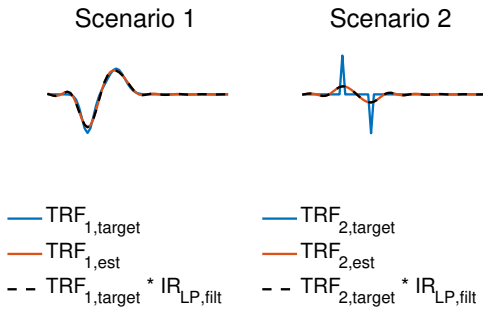

Figure A.6: The effect of filtering on encoding model parameters. Two simulation scenarios are here considered. Scenario 1 simulates a response, $R_1(t)$, as convolution of a signal, $T(t)$, with $\text{TRF}_{1,\text{target}}$. Scenario 2 simulates a response, $R_2(t)$, as convolution of a signal, $T(t)$, with $\text{TRF}_{2,\text{target}}$. In both simulations, the responses are low-pass filtered with an FIR filter with a cut-off frequency of 15 Hz, and subsequently shifted to adjust for filter delay. The two panels compares the actual TRFs with the estimated TRFs. See text for more details.

A.6 Example 6: Model interpretation in scenarios where responses exhibit feature selectivity

A neural response, $R(t)$, may simultaneously represent information about different sound features (Di Liberto et al., 2015; Lescroart et al., 2015; Naselaris et al., 2011). Both encoding and decoding models can be used to investigte whether a neural response exhibits feature selectivity. This question can be addressed with encoding models by fitting separate models on different feature types and evaluating how much of the explainable variance in the neural data that can be predicted uniquely by each feature type (see Heer et al., 2017; Lescroart et al., 2015). Similar model comparisons can also be performed with decoding models (Naselaris et al., 2011), but such model comparisons can be more difficult to interpret when the different stimulus features have different statistical properties.

Figure A.7 illustrates how encoding models can be used to determine the fraction of variance in a neural response that is uniquely explained by different feature spaces (Heer et al., 2017; Lescroart et al., 2015). Three different simulation scenarios are here considered. Scenario 1 (Fig. A.7a) simulates a response as a convolution between a feature E and a temporal response function TRF_E plus another feature O convoled with a temporal response function TRF_O. The O feature responds systematically to onsets in E. Scenario 2 (Fig. A.7b) simulates a response as a convolution between a feature E and a temporal response function TRF_E. Finally, Scenario 3 (Fig. A.7c) simulates a response as a convolution between a feature O and a temporal response function TRF_O.

In all three scenarios it is possible to estimate the unique contribution of the features, O and E, to variance in the response. To do this, separate encoding models are fit to the two features, E and O, and a joint encoding model is fit to the two features concatenated along feature dimension. How much variance in the data each of these three models explain can be estimated by computing Pearson's correlation coefficient, r, between model predictions and target data (over a test set), and squaring the correlation coefficient r^2 (David and Gallant, 2005; Heer et al., 2017; Lescroart et al., 2015). Assume that a joint model based on a union of two feature spaces, $E \cup O$, explains some variance in the data, $r^2_{E\cup O}$ (Lescroart et al., 2015). It is now possible to estimate the unique variance explained by each model, $r^2_E = r^2_{E\cup O} - r^2_O$, and the variance shared between the two models $r^2_{E\cap O} = r^2_E + r^2_O - r^2_{E\cup O}$ (Heer et al., 2017; Lescroart et al., 2015). In

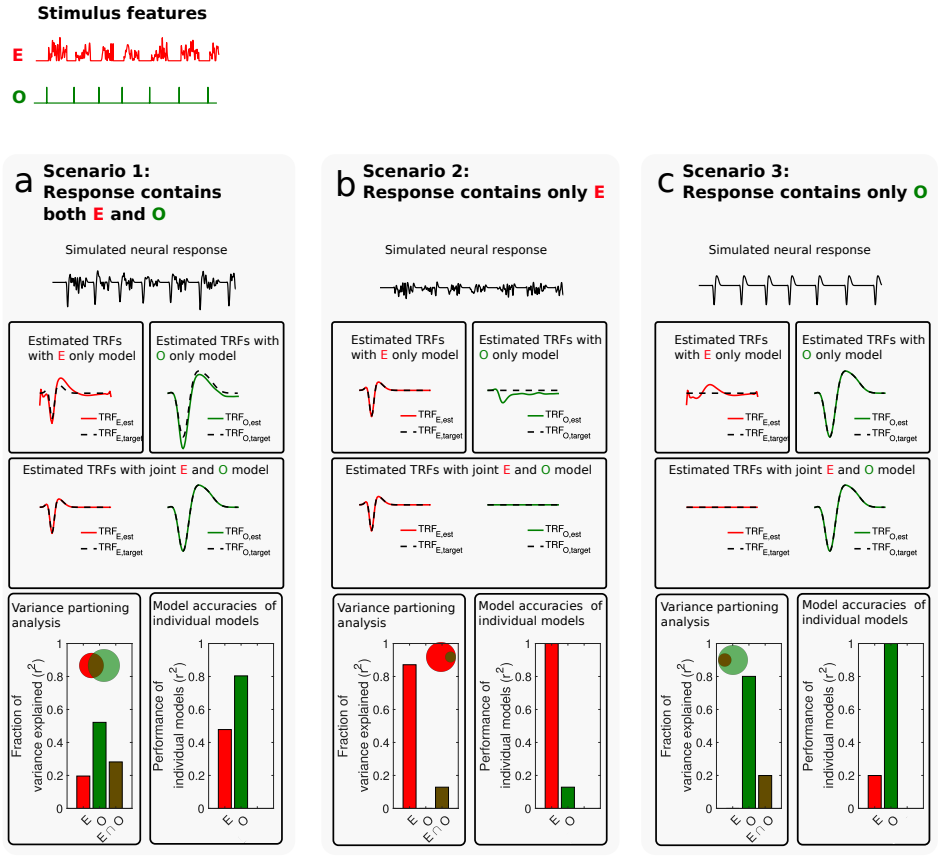

Figure A.7: Using encoding models to estimate how much variance in simulated neural responses that is uniquely explained by different feature spaces. Three simulation scenarios are here considered. In scenario 1, the neural response is simulated as convolution between E and a temporal response function, TRF_E plus O convolved with a temporal response function TRF_O. In scenario 2, the neural response is simulated as convolution between E and TRF_E. In scenario 3, the neural response is simulated as convolution between O and TRF_O. In all of the three simulation scenarios, it is possible to estimate the unique contribution of the two features, E and O to variance in the simulated response (bottom left panels in a, b and c). However, if either of the two features are not included in the analysis, it can affect interpretation of modelling results. For instance, if one only considers O in scenario 1 or 2, then $\text{TRF}_{O,\text{est}}$ deviates from $\text{TRF}_{O,\text{target}}$). Similar effects occur in scenario 1 and 3, when only considering E.

this way, it is possible to assess whether the different models account for unique or shared variance in the simulated response. However, the accuracy of the estimates may be affected by sampling noise (Heer et al., 2017).

It can be seen in Fig. A.7a that both individual models explain some unique variance (r_E^2 and r_O^2) in the simulated neural responses. However, a substantial fraction of variance in the neural data is shared between the two models. Since

the true stimulus-response relation is known, the encoding filters recovered with the different encoding models can be compared to the true temporal response functions, TRF_E and TRF_O. As expected, the joint model $E \cup O$ show complete recovery of TRF_E and TRF_O (Fig. A.7a, bottom), as it can account for regularities across features to predict the response (see e.g. Example 1). Conversely, when fitting individual models, the estimated temporal response functions differ from TRF_E and TRF_O (Fig. A.7a top row), as they do not account for the shared variance explained by both models.

In Fig. A.7b and Fig. A.7c the variance partitioning analyses identify that only one of the two features uniquely contribute to variance in the responses, and that a fraction of the explained variance is shared between model features. Again, in these simulations the joint models show complete recovery of TRF_E and TRF_O. In both scenarios, when the response does not contain a certain feature, the individual models may attain nonzero weights unlike their targets. This could in both cases lead to wrong conclusions about what attributes in the stimulus features that drive the response. Fitting an individual model and not rigorously assessing the fraction of variance explained by other relevant models, can in other words affect interpretation of models weights (top panels, in Fig. A.7a,b,c) and interpretation of model accuracies (bottom right panels, in Fig. A.7a,b,c).

Contributions to Hearing Research

Vol. 1: *Gilles Pigasse*, Deriving cochlear delays in humans using otoacoustic emissions and auditory evoked potentials, 2008.

Vol. 2: *Olaf Strelcyk*, Peripheral auditory processing and speech reception in impaired hearing, 2009.

Vol. 3: *Eric R. Thompson*, Characterizing binaural processing of amplitude-modulated sounds, 2009.

Vol. 4: *Tobias Piechowiak*, Spectro-temporal analysis of complex sounds in the human auditory system, 2009.

Vol. 5: *Jens Bo Nielsen*, Assessment of speech intelligibility in background noise and reverberation, 2009.

Vol. 6: *Helen Connor*, Hearing aid amplification at soft input levels, 2010.

Vol. 7: *Morten Løve Jepsen*, Modeling auditory processing and speech perception in hearing-impaired listeners, 2010.

Vol. 8: *Sarah Verhulst*, Characterizing and modeling dynamic processes in the cochlea using otoacoustic emissions, 2010.

Vol. 9: *Sylvain Favrot*, A loudspeaker-based room auralization system for auditory research, 2010.

Vol. 10: *Sébastien Santurette*, Neural coding and perception of pitch in the normal and impaired human auditory system, 2011.

Vol. 11: *Iris Arweiler*, Processing of spatial sounds in the impaired auditory system, 2011.

Vol. 12: *Filip Munch Rønne*, Modeling auditory evoked potentials to complex stimuli, 2012.

Vol. 13: *Claus Forup Corlin Jespersgaard*, Listening in adverse conditions: Masking release and effects of hearing loss, 2012.

Vol. 14: *Rémi Decorsière*, Spectrogram inversion and potential applications for hearing research, 2013.

Vol. 15: *Søren Jørgensen*, Modeling speech intelligibility based on the signal-to-noise envelope power ration, 2014.

Vol. 16: *Kasper Eskelund*, Electrophysiological assessment of audiovisual integration in speech perception, 2014.

Vol. 17: *Simon Krogholt Christiansen*, The role of temporal coherence in auditory stream segregation, 2014.

Vol. 18: *Márton Marschall*, Capturing and reproducing realistic acoustic scenes for hearing research, 2014.

Vol. 19: *Jasmina Catic*, Human sound externalization in reverberant environments, 2014.

Vol. 20: *Michał Fereczkowski*, Design and evaluation of individualized hearing-aid signal processing and fitting, 2015.

Vol. 21: *Alexandre Chabot-Leclerc*, Computational modeling of speech intelligibility in adverse conditions, 2015.

Vol. 22: *Federica Bianchi*, Pitch representations in the impaired auditory system and implications for music perception, 2016.

Vol. 23: *Johannes Zaar*, Measures and computational models of microscopic speech perception, 2016.

Vol. 24: *Johannes Käsbach*, Characterizing apparent source width perception, 2016.

Vol. 25: *Gusztáv Löcsei*, Lateralized speech perception with normal and impaired hearing, 2016.

Vol. 26: *Suyash Narendra Joshi*, Modelling auditory nerve responses to electrical stimulation, 2017.

Vol. 27: *Henrik Gerd Hassager*, Characterizing perceptual externalization in listeners with normal, impaired and aided-impaired hearing, 2017.

Vol. 28: *Richard Ian McWalter*, Analysis of the auditory system via synthesis of natural sounds, speech and music, 2017.

Vol. 29: *Jens Cubick*, Characterizing the auditory cues for the processing and perception of spatial sounds, 2017.

Vol. 30: *Gerard Encina-Llamas*, Characterizing cochlear hearing impairment using advanced electrophysiological methods, 2017.

Vol. 31: *Christoph Scheidiger*, Assessing speech intelligibility in hearing-impaired listeners, 2018.

Vol. 32: *Alan Wiinberg*, Perceptual effects of non-linear hearing aid amplification strategies, 2018.

Vol. 33: *Thomas Bentsen*, Computational speech segregation inspired by principles of auditory processing, 2018.

Vol. 34: *François Guérit*, Temporal charge interactions in cochlear implant listeners, 2018.

Vol. 35: *Andreu Paredes Gallardo*, Behavioral and objective measures of stream segregation in cochlear implant users, 2018.

Vol. 36: *Borys Kowalewski*, Assessing hearing-aid signal processing based on variations of the Turing test, 2018

The end.

To be continued...

This thesis investigated how selective attention affects single-trial electroencephalography (EEG) responses to speech. Work from EEG studies in normal-hearing (NH) and hearing-impaired (HI) listeners is presented. It is explored how different task demands and listening scenarios affect EEG responses to speech. It is shown that the auditory attentional selection of NH and HI listeners can be decoded from EEG data with reasonably high classification accuracies in different listening environments. Overall, the work presented in this thesis suggests that EEG-based attention decoding may have relevance for future brain-computer interface systems.

DTU Electrical Engineering
Department of Electrical Engineering

Ørsteds Plads
Building 348
DK-2800 Kgs. Lyngby
Denmark
Tel: (+45) 45 25 38 00
Fax: (+45) 45 93 16 34
www.elektro.dtu.dk